Finding Regions of Aberrant DNA Copy Number Associated with Tumor Phenotype

Author
Laura Toloşi

Dissertation

zur Erlangung des Grades
des Doktors der Naturwissenschaften (Dr. rer. nat.)
der Naturwissenschaftlich–Technischen Fakultäten
der Universität des Saarlandes

Saarbrücken
2012

Tag des Kolloquiums:	27.09.2012
Dekan:	Prof. Dr. Mark Groves
Vorsitzender des Prüfungsausschusses:	Prof. Dr. Matthias Hein
Berichterstatter:	Prof. Dr. Thomas Lengauer, Ph.D.
	Prof. Dr. Hans-Peter Lenhof
	Prof. Dr. Jörg Rahnenführer
Beisitzer:	Dr. Nico Pfeifer

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt ubernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form in einem Verfahren zur Erlangung eines akademischen Grades vorgelegt.

Saarbrücken, den

Laura Toloşi

Abstract

DNA copy number alterations are a hallmark of cancer. Understanding their role in tumor progression can help improve diagnosis, prognosis and therapy selection for cancer patients and can contribute to the development of personalised therapies. High-resolution, genome-wide measurements of DNA copy number changes for large cohorts of tumors are currently available, owing to the rapid development of technologies like microarray-based array comparative hibridization (arrayCGH). In this manuscript, we introduce a computational pipeline for statistical analysis of tumor cohorts, which can help extract relevant patterns of copy number aberrations and infer their association with various phenotypical indicators. The pipeline makes use of machine learning techniques for classification and feature selection, with emphasis on interpretable models (linear models with penalties, tree-based models).

The main challenges that our methods meet are the high dimensionality of the arrays compared to the small number of tumor samples available, as well as the large correlations between copy number estimates measured at neighboring genomic locations. Consequently, feature selection is unstable, depending strongly on the set of training samples, leading to un-reproducible signatures across different clinical studies. We also show that the feature ranking given by several widely-used methods for feature selection is biased due to the large correlations between features. In order to correct for the bias and instability of the feature ranking, we introduce a dimension reduction step in our pipeline, consisting of multivariate segmentation of the set of arrays. We present three algorithms for multivariate segmentation, which are based on indentifying recurrent DNA breakpoints or DNA regions of constant copy number profile. The multivariate segmentation constitutes the basis for computing a smaller set of super-features, by summarizing the DNA copy number within the segmentation regions. Using the super-features for supervised classification, we improve the interpretability and stability of the models, where the baseline for comparison consists of classification models trained on probe data.

We validated the methods by training models for prediction of the phenotype of breast cancers and neuroblastoma tumors. We show that the multivariate segmentation step affords higher model stability and it does not decrease the accuracy of the prediction. We obtain substantial dimension reduction (up to 200-fold less predictors), which recommends the multivariate segmentation procedures not only for the purpose of phenotype prediction, but also as preprocessing step for downstream integration with other data types.

The interpretability of the models is also improved, revealing important associations between copy number aberrations and phenotype. For example, we show that a very informative predictor that distinguishes between inflammatory and non-inflammatory breast cancers with ERBB2 amplification is the co-amplification of the genomic region located in the immediate vicinity of the ERBB2 gene locus. Therefore, we conclude that the size of the amplicon is associated with the cancer subtype, a hypothesis present elsewhere in the literature. In the case of neuroblastoma tumors, we show that patients belonging to different age subgroups are characterized by distinct copy number patterns, especially when the subgroups are defined as older or younger than 16-18 months. Indeed, considering a large set of age cutoffs, our prediction models are most accurate if the cutoff is around 16-18 months. We thereby confirm the recommendation for a higher age cutoff than 12 months

(current clinical practice) for differential diagnosis of neuroblastoma.

Kurzfassung

Die abnormale Multiplizität bestimmter Segmente der DNS (copy number aberrations) ist eines der hervorstechenden Merkmale von Krebs. Das Verständnis der Rolle dieses Merkmals für das Tumorwachstum könnte maßgeblich zur Verbesserung von Krebsdiagnose, -prognose und -therapie beitragen und somit bei der Auswahl individueller Therapien helfen. Micoroarray-basierte Technologien wie 'Array Comparative Hybridization' (array-CGH) erlauben es, hochauflösende, genomweite Kopiezahl-Karten von Tumorgeweben zu erstellen. Gegenstand dieser Arbeit ist die Entwicklung einer Software-Pipeline für die statistische Analyse von Tumorkohorten, die es ermöglicht, relevante Muster abnormaler Kopiezahlen abzuleiten und diese mit diversen phänotypischen Merkmalen zu assoziieren. Dies geschieht mithilfe maschineller Lernmethoden für Klassifikation und Merkmalselektion mit Fokus auf die Interpretierbarkeit der gelernten Modelle (regularisierte lineare Methoden sowie Entscheidungsbaum-basierte Modelle).

Herausforderungen an die Methoden liegen vor allem in der hohen Dimensionalität der Daten, denen lediglich eine vergleichsweise geringe Anzahl von gemessenen Tumorproben gegenüber steht, sowie der hohen Korrelation zwischen den gemessenen Kopiezahlen in benachbarten genomischen Regionen. Folglich hängen die Resultate der Merkmalselektion stark von der Auswahl des Trainingsdatensatzes ab, was die Reproduzierbarkeit bei unterschiedlichen klinischen Datensätzen stark einschränkt. Diese Arbeit zeigt, dass die von diversen gängigen Methoden bestimmte Rangfolge von Features in Folge hoher Korrelationskoeffizienten einzelner Prädiktoren stark verfälscht ist. Um diesen 'Bias' sowie die Instabilität der Merkmalsrangfolge zu korrigieren, führen wir in unserer Pipeline einen dimensions-reduzierenden Schritt ein, der darin besteht, die Arrays gemeinsam multivariat zu segmentieren. Wir präsentieren drei Algorithmen für diese multivariate Segmentierung, die auf der Identifikation rekurrenter DNA Breakpoints oder genomischer Regionen mit konstanten Kopiezahl-Profilen beruhen. Durch Zusammenfassen der DNA Kopiezahlwerte innerhalb einer Region bildet die multivariate Segmentierung die Grundlage für die Berechnung einer kleineren Menge von 'Super-Merkmalen'. Im Vergleich zu Klassifikationsverfahren, die auf Ebene einzelner Arrayproben beruhen, verbessern wir durch überwachte Klassifikation basierend auf den Super-Merkmalen die Interpretierbarkeit sowie die Stabilität der Modelle.

Wir validieren die Methoden in dieser Arbeit durch das Trainieren von Vorhersagemodellen auf Brustkrebs und Neuroblastoma Datensätzen. Hier zeigen wir, dass der multivariate Segmentierungsschritt eine erhöhte Modellstabilität erzielt, wobei die Vorhersagequalität nicht abnimmt. Die Dimension des Problems wird erheblich reduziert (bis zu 200-fach weniger Merkmale), welches die multivariate Segmentierung nicht nur zu einem probaten Mittel für die Vorhersage von Phänotypen macht.Vielmehr eignet sich das Verfahren darüberhinaus auch als Vorverarbeitungschritt für spätere integrative Analysen mit anderen Datentypen.

Auch die Interpretierbarkeit der Modelle wird verbessert. Dies ermöglicht die Identifikation von wichtigen Relationen zwischen Änderungen der Kopiezahl und Phänotyp. Beispielsweise zeigen wir, dass eine Koamplifikation in direkter Nachbarschaft des ERBB2 Genlokus einen höchst informativen Prädiktor für die Unterscheidung von entzündlichen und nicht-entzündlichen Brustkrebsarten darstellt. Damit bestätigen wir die in der Lite-

ratur gängige Hypothese, dass die Größe eines Amplikons mit dem Krebssubtyp zusammenhängt. Im Fall von Neuroblastoma Tumoren zeigen wir, dass Untergruppen, die durch das Alter des Patienten definiert werden, durch Kopiezahl-Muster charakterisiert werden können. Insbesondere ist dies möglich, wenn ein Altersschwellenwert von 16 bis 18 Monaten zur Definition der Gruppen verwandt wird, bei dem außerdem auch die höchste Vorhersagegenauigkeit vorliegt. Folglich geben wir weitere Evidenz für die Empfehlung, einen höheren Schwellenwert als zwölf Monate für die differentielle Diagnose von Neuroblastoma zu verwenden.

Acknowledgements

I would like to thank my advisor Prof. Thomas Lengauer for his guidance, continuous support and valuable advice throughout the years. Thank you for teaching me to dare aim high in research, but also to accept that some problems are hard and there is value in each attempt at solving them, no matter the result.

I am very grateful to Prof. Jörg Rahnenführer for revealing to me some of the secrets of statistical data analysis. Thank you for teaching me that statistics is not an instrument for distorting reality at will, but a great tool that can bring to light unexpected, complex and useful patterns in data, if interpreted correctly.

I also thank Prof. Hans-Peter Lenhof, who kindly agreed to review this manuscript.

This work would not have been possible without the help of our collaborators: Dr. Jessica Berthold, Dr. med. Barbara Hero, and Prof. Dr. med. Frank Berthold from the Department of Pediatric Oncology and Hematology, University Children's Hospital, Cologne. Also, thanks to Dr. Roman Thomas from MPI Cologne and to all members of the Oncogene Consortium. I am very thankful to André Altmann, Oliver Sander, Konstantin Halachev and Sven-Eric Schelhon for the collaboration and the nice articles we published together. I am also grateful to Adrian Alexa and Yassen Assenov, for sharing helpful tricks about the R language with me.

I thank Barbara Dörr for her great bachelor thesis, and to Violeta Ivanova for good work as master student. I wish you all the best in your future careers!

Thanks to all the former and present members of the Bioinformatics department for extremely fruitful scientific discussions, which helped broaden my knowledge into different fields of research and science, but also for the good times we had together either participating in running events (thanks, Ruth, for organizing most of them!), or having a beer out during the sunny days. Special thanks to my former office mates Adrian Alexa and Jasmina Bogojeska, who had to put up with my complaints whenever the results of my work were not good enough, but who also shared my happiness when they were better. I am also grateful to Achim and Georg for their fast response to any issues related to computers.

I spent great years in Saarbruecken, which I will never forget. I had the incredible chance of making great friends: Rayna, Yassen, Jasmina, Levi, Andre, Evangelia, Adrian, Vitaly, Stefan, Hagen, Sven, Fabian, Ruxandra, Dimitar. Thank you all for the great laughs we had together!

I would like to thank my parents for their continuous support, especially my mother, Emilia and my sister, Roxana.

Above all, I am grateful to Konstantin, for believing in me unconditionally through the years and for always encouraging me to pursue my ideas. Thank you!

Contents

List of Figures

List of Tables

1. Introduction

The decades after the Human Genome Project have changed the perspectives on cancer therapy dramatically. Genome-wide technologies like microarrays and next-generation sequencing have revealed unprecedented insights into the cancer genome, epigenome and transcriptome (in short, cancer omics). It has become clear that cancer is a vastly complex disease, exhibiting a plethora of abnormalities in the function of many genes, which can arise by gradual accumulation of genetic aberrations or by disruption of epigenetic regulation. The research community must now find ways of translating the multidimensional omics data into efficient cancer therapy.

The large number of cancer genomes already available display highly heterogeneous abnormalities, some being essential for tumor progression (driver alterations) and others being spurious, harmless events (passenger alterations). The discrimination between these two categories is essential, but has been recognized as a very difficult problem. In the field of cancer research, the task of identifying key events that have an impact on tumor phenotype is called *biomarker discovery*. The identification of biomarkers is a necessary step towards improved diagnosis and prognosis, as well as towards genotype-informed therapy (or personalized therapy), which aims at administering drugs that target specific aberrations that the patient's cancer cells rely on for proliferation. So far, biomarker discovery has been driven mainly by the goal of finding new drug targets, but more recent studies also propose diagnostic and prognostic biomarkers, which can shed light on the degree of tumor progression, tumor subtype or on the expected survival of the patient.

The computational task of biomarker discovery is based on inferring the statistical dependence between tumor phenotype and tumor genotype. In most of the studies the dependence is measured univariately, e.g. between single gene mutations and the phenotype. Such an approach can detect strong univariate associations, but will not reveal a more complex interplay between multiple genes, which would result for example as a consequence of a disrupted pathway. A far more powerful instrument which allows for multivariate inference is supervised learning, a technique that allows for modeling of the phenotype as a (potentially complex) function of genotype factors. Such models, also called *complex biomarkers*, can help predict the phenotype of new tumors, such as for example response to certain drugs, expected survival, lymph node spread, presence or absence of metastasis and others. However, despite the fact that supervised modeling for phenotype prediction has been proposed in many studies during the past years (mainly based on gene expression data), the gained knowledge has rarely entered clinical practice. One of the major reasons is the lack of biological reasoning in the prediction process, which understandably worries the clinician who has to take responsibility for the outcome of the treatment.

In this context, simple and biologically interpretable models are better received by the medical community than sophisticated and more accurate, not interpretable statistical models. Following these requirements, the work proposed in this thesis presents meth-

ods for supervised classification, which deliver sparse, robust and interpretable models. We restrict our experiments to classification of tumors based on their DNA copy number alterations, which are manifested by gain or loss of genetic material during tumor development and progression. We used high-resolution, microarray-based data such as array-based comparative genomic hybridization data for validation. This type of data has been rarely approached with supervised classification techniques before, mainly due to the increased focus on classification of expression data. As a consequence, the accuracy of phenotype prediction based on copy number alterations is largely unknown in the community.

The challenges that supervised classification faces when applied to copy number data can be summarized into two main aspects:

- the large number of features (genomic loci) compared to the small number of available samples (tumors), well known as *the curse of dimensionality*;
- the large correlations between genomic loci, in particular between probes located closely in the genome sequence.

High dimensionality requires feature selection methods for sparse modeling. We use lasso-penalized logistic regression models, for sparsity and interpretability. We also introduce a novel method for assigning p-values to features in random forest models. By selecting only the features that exceed a significance threshold, we mimic sparsity in random forest (non-linear) models.

We show that correlation between features is dangerous as it can bias feature ranking. We demonstrate that an effective way of dealing with correlation is correlation-based feature grouping. Specifically, features that are highly correlated are grouped together into super-features, which are used for classification. Feature grouping achieves both dimension reduction and removes correlation bias. In our algorithms for feature grouping, we speculate on the local correlation between genomic loci.

In a comprehensive application study to breast cancer and neuroblastoma, we show that prediction of tumor phenotype based on copy number alterations can achieve reasonable accuracy, even when the training set is relatively small. We also show that model interpretability and stability are improved by feature grouping, while preserving prediction accuracy. We also present important biological insights revealed by our models, thereby demonstrating that computational models are capable and ready to assist the biomedical community in improving cancer understanding and treatment.

Outline

In Chapter 2 we begin with an introduction to cancer, in which we describe the current clinical practice, including a short overview on the current treatment options and the standard classification at diagnosis. Then, we continue with presenting the most intensively studied types of molecular changes that are linked to tumorigenesis, including genomic, epigenomic and transcriptomic alterations. We describe in more details the molecular mechanisms that underlie the formation of genomic aberrations that lead to copy number imbalances. Towards the end of Chapter 2, we review the experimental technologies that are currently used for genome-wide measurement of copy-number alterations.

Chapter 3 reviews the existing pipelines that have been proposed for computational analysis of copy number alterations. The general goal of most pipelines is to identify copy number alterations that play a role in tumor onset and progression. To this end, pipelines generally consist of successive steps of reducing the dimension of the data, in which data features are selected in an unsupervised manner, based on assumptions argued biologically. We thereby suggest a novel pipeline, that makes use of minimal assumptions and processing steps in oder to select and rank the most relevant copy number aberrations from a tumor set. The selection is data-driven: a phenotypic indicator of the tumor such as grade, stage, metastasis status, etc. is used as a response variable for a supervised learning model, where the predictors are the copy number aberration profiles. The key step of our pipeline is the consensus segmentation, namely the task of estimating a consensus partition of the tumor genomes into regions of almost constant copy number. In Chapter 3, we only present the steps of the pipeline schematically and argue about its qualitative advantages against traditional pipelines.

In Chapter 4, we define consensus segmentation and introduce three algorithms for this task. The algorithms work by estimating consensus regions of almost constant copy number. The copy-number data of the tumor samples can be represented in the reduced space of the consensus regions without substantial loss of information. We compare the algorithms for consensus segmentation on seven publicly available arrayCGH datasets pertaining to five cancer types. We thereby show that consensus segmentation can achieve a substantial dimension reduction of the genome-wide copy-number data by reducing the correlation between loci, therefore facilitating supervised prediction tasks.

Chapter 5 introduces the supervised learning techniques that we used for prediction of tumor phenotype based on copy number data. We considered only binary classification tasks. We comment on the central problem of classification based on genome-wide omics data, namely the high dimensionality compared to the relatively small sample sizes and the high correlations between the features. We also stress the importance of interpretability of classification models and formulate a set of requirements that a good model should comply with. We show how most widely-used classification models do not comply with these requirements and in this context we define correlation bias – a type of bias that causes large groups of correlated features to receive small weights. We show with simulations that consensus segmentation successfully removes correlation bias.

Chapter 6 contains two applications of our pipeline, to breast cancer and to neuroblastoma. We trained modes for prediction of several phenotypical indicators that are relevant for the respective tumor types. In Chapter 6 we comment on the accuracy of prediction and analyzed the feature ranking that was given by the models. We comment on the biological relevance of our findings and conclude with comments on the strengths and weaknesses of our approaches.

Chapter 7 concludes this thesis and provides an outlook on further directions of improvement.

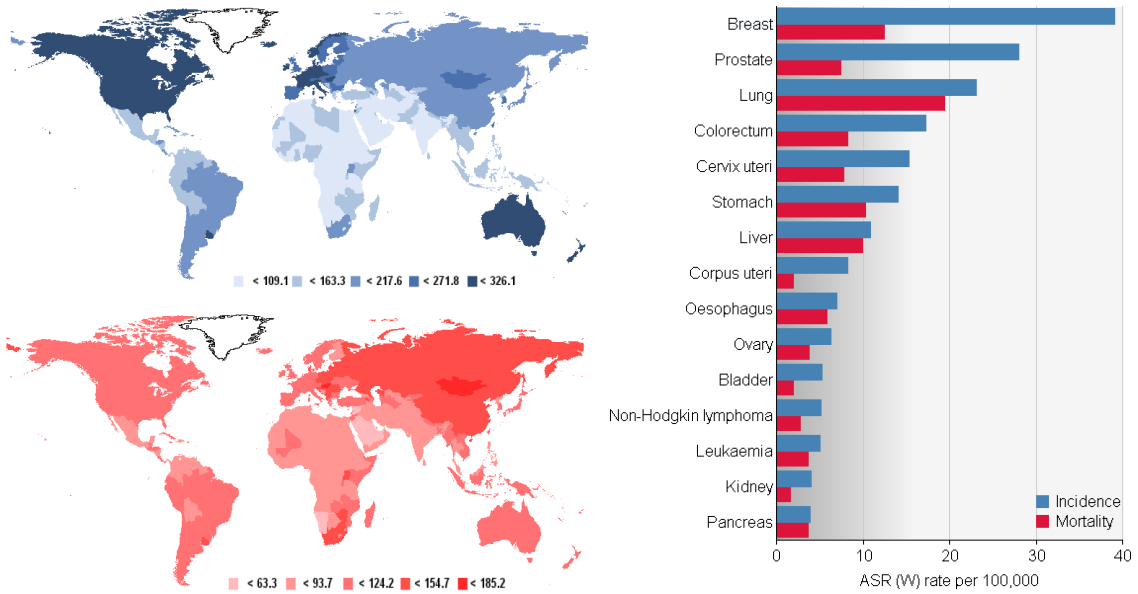
2. DNA Copy Number Aberrations and Cancer

> I have called this principle, by which each slight variation, if useful, is preserved, by the term of Natural Selection.
>
> Charles Darwin

2.1. Cancer overview

Cancer is the generic name given to a wide class of genetic diseases which display an uncontrolled growth of cells. Cancer can affect most organs of the human body (commonly breast, lung, skin, bone, blood, liver, pancreas, ovaries, brain). In order to evade the regulatory mechanisms that prohibit uncontrolled proliferation by inducing apoptosis (programmed cell death), the cancerous cells acquire molecular abnormalities via a process of microevolution similar to the Darwinian evolution (Nowell, 1976). Specifically, during cancer progression, selective pressure acts on cells by promoting molecular changes that disrupt regulatory mechanisms and allow the cells to escape apoptosis. Eventually, the mass of cancerous cells becomes very large, threatening the function of the affected organ, or that of the nearby organs, or can spread to more distant parts of the body via the lymphatic system and bloodstream (process referred to as *metastasis*). Metastases are the main cause of death from cancer.

The causes of cancer are diverse. Heredity plays an important role, studies showing that 5 to 10% of the breast cancers run in families (Colditz et al., 1993). Advanced age is frequently associated with cancer, due to DNA degradation. Exposure to radiation and other environmental pollutants can cause leukemia (Richardson et al., 2009), skin cancer (Narayanan et al., 2010), lung cancer (Cardis et al., 2007; Kreisheimer et al., 2003), among others. Tobacco use accounts for about 80% of lung cancers (Parkin, 2011). Inappropriate diet and obesity has been linked to stomach and colon cancer (Huang and Chen, 2009; Frezza et al., 2006), as has been the lack of physical activity. Viral infections have been indicated as possible causative agents for several cancers including cervical, oral, breast, prostate cancers or lymphoma (Sarid and Gao, 2011) and intense research is being dedicated to tracing viral agents in other tumors. In this context, we have studied the transcriptome of a small set of 14 neuroblastoma tumors and were not able to identify transcripts that could unequivocally be attributed to viral agents and not to human homologs. The methodological work and a positive validation to cervical cancer with HPV are presented in Schelhorn et al. (2012). For many cancers, the causes remain unknown (for example glioblastoma).



(a) (b)

Figure 2.1.: Estimated world-wide age-standardized incidence and mortality rates per 100000 for cancer, according to the WHO and GLOBOCAN (http://globocan.iarc.fr) a) by geographical region and b) by cancer type. Blue shows incidence, red shows mortality rates.

The incidence of cancer worldwide is unevenly distributed, highly developed countries facing higher rates than underdeveloped regions (Figure 2.1a), probably due to the environmental factors (pollutants), diet and increased lifespan. The quality of the health care systems in developed countries however compensates for the higher incidence, resulting in comparatively lower mortality rates (Figure 2.1a). Cancer incidence and mortality also varies with the organ affected (lung, liver, stomach and breast yield highest mortality, see Figure 2.1b), age (risk of cancer generally increases with age), sex, genetic makeup of the patient, etc.

Due to the persistent high incidence and mortality rate, cancer has been and continues to be a very important focus of research worldwide. According to the World Health Organization (WHO), cancer is a leading cause of death. In highly developed countries, mortality due to cancer is second only to heart disease. Despite large investments in cancer research over the past century, development of new treatments has been slow and often unrewarding, owing to the overwhelming heterogeneity of the disease. Remarkable improvements in early diagnosis (e.g. imaging) have allowed for detection of the disease at a stage where treatment is efficient, thus decreasing mortality rates.

2.1.1. Cancer treatment

Several therapeutical options exist currently for cancer patients, the most widely used being *chemotherapy*, *radiation therapy* and *surgery*.

Chemotherapy consists of administering one or more drugs that interfere with tumor development. Older drugs have the general effect of slowing down cell replication in the

body of the patient, for example by interfering with mitosis or promoting apoptosis. Such drugs can damage healthy cells as well, especially those that divide rapidly such as bone marrow or hair, which leads to severe side effects. Modern drugs are designed to target specific protein mutations that occur in cancer cells and thus have fewer side effects.

Drugs are administered either by intravenous injection or taken orally as pills. Through the bloodstream, the drug can reach tumor cells located in various parts of the body, making chemotherapy the main treatment against metastatic cancer.

Radiation therapy uses ionizing radiation directed towards the tumor in order to damage the DNA of tumor cells. When the DNA is damaged beyond repair, the cells die. Radiation is used when the tumor is localized (not against metastasis), sometimes in order to shrink the tumor for follow-up surgery. Although radiation therapy is directed towards the cancerous cells, healthy cells situated close to the tumor can be damaged as well. Modern technologies (such as proton beam radiation therapy, radioimmunotherapy, brachytherapy) are becoming better at targeting only cancerous cells while delivering a higher amount of radiation.

Surgery is the foundation of cancer treatment, being used for multiple purposes during patient care. Surgery is helpful for diagnosis (biopsy), when a part of the tumor or the entire tumor is removed and then studied under a microscope and classified (cancerous, non-cancerous, stage). Surgery is currently the best option for primary treatment of many cancers, especially if the cancer has not spread. When it is not possible to remove the entire tumor, then surgery is an option for removing part of the mass of cells that hinder the functioning of the neighboring organs. Sometimes, surgery is used for relieving symptoms or side effects, for example when the tumor is pressing on a nerve and causes pain. In some cases, surgery can be used for cancer prevention, if a high probability exists that a tumor will develop and the removal of the anatomical part under risk is possible (for example masectomy, prostatectomy).

Other treatment options include immunotherapy (designed to induce the patient's own immune system to recognize and fight the tumor), hormone therapy (in hormone-dependent cancers, blocking of growth hormones can significantly slow tumor development) or inhibiting angiogenesis (prevent the growth of blood vessels that supply the tumor with nutrients). Upon the discovery of viral agents responsible for increased tumor susceptibility, vaccines have been designed for cancer prevention.

2.1.2. Cancer classification

Cancer classification is presently carried out in clinics in order to assign a specific tumor to a known subgroup and thereby allow for selecting the most suitable therapy and prognosis. Classification usually follows several criteria: by site of origin (*histology*), by degree of differentiation of the cells (*grade*) or by degree of progression (*stage*).

Common histological types are:

- *Adenocarcinoma* – originates in glandular tissue;
- *Blastoma* – originates in embryonic tissue of organs;
- *Carcinoma* – originates in epithelial tissue (i.e., tissue that lines organs and tubes);
- *Leukemia* – originates in tissues that form blood cells;

- *Lymphoma* – originates in lymphatic tissue;
- *Myeloma* – originates in bone marrow;
- *Sarcoma* – originates in connective or supportive tissue (e.g., bone, cartilage, muscle).

Depending on the anatomy of the affected organ, histological subtypes may be defined. For example, *ductal* carcinoma is a common histological subtype of breast cancer that forms in the lining of the milk ducts, while *lobular* carcinoma is another subtype that originates in the lobules of the breast, where milk is produced.

Tumor grading requires a biopsy and the examination of tumor tissue under the microscope. Grading is carried out by an experienced pathologist, who evaluates the degree of differentiation of the cells. Poor cell differentiation (anaplasia) and abnormal appearance of the cells is specific to aggressive forms of cancer. The standard system comprises four grades, determined as follows:

- *Grade 1* – cells slightly abnormal and well differentiated;
- *Grade 2* – cells more abnormal and moderately differentiated;
- *Grade 3* – cells highly abnormal and poorly differentiated;
- *Grade 4* – cells highly abnormal and undifferentiated.

The stage (or TNM stage) provides a measure of the extent of the spread of the tumor. Several indicators are typically used, including tumor size (T), spread to lymph nodes (N) and metastasis (M).

Tumor size can take one of the following values:

- *Tx* – tumor cannot be evaluated;
- *T0* – no evidence of tumor;
- *Tis* – carcinoma in situ (limited to surface cells of the organ);
- *T1-4* – increasing tumor size and involvement.

The degree of spread to lymph nodes is one of the following:

- *Nx* – lymph nodes cannot be evaluated;
- *N0* – tumor cells absent from regional lymph nodes;
- *N1* – regional lymph node metastasis present;
- *N2* – tumor spread to an extent between N1 and N3;
- *N3* – tumor spread to more distant or numerous regional lymph nodes.

The indicator of metastasis can be either

- *Mx* – distant metastasis cannot be evaluated;
- *M0* – no distant metastasis;
- *M1* – metastasis to distant organs (beyond regional lymph nodes).

The T, N and M markers are combined to form a tumor stage indicator, as follows:

- *Stage 0* – cancer in situ (limited to surface cells);
- *Stage I* – cancer limited to the tissue of origin, evidence of tumor growth;
- *Stage II* – limited local spread of cancerous cells;
- *Stage III* – extensive local and regional spread;
- *Stage IV* – distant metastasis.

Depending on the organ affected, tumor stage can take different values, necessary for the better description of the particular cancer. For example, neuroblastoma staging involves a subdivision of stage 4 into stage 4 and stage 4S (special), depending on the age of the patient and the location of distant metastases. The staging of ovarian cancer is refined, each main stage I-IV being further indexed with A, B or C, depending on the anatomic characteristics of the spread. A similar subdivision is met in breast cancer staging. Adaptations of the general staging to particular cancer types have the purpose of improving therapy selection and prognosis and is mainly based on empirical observations.

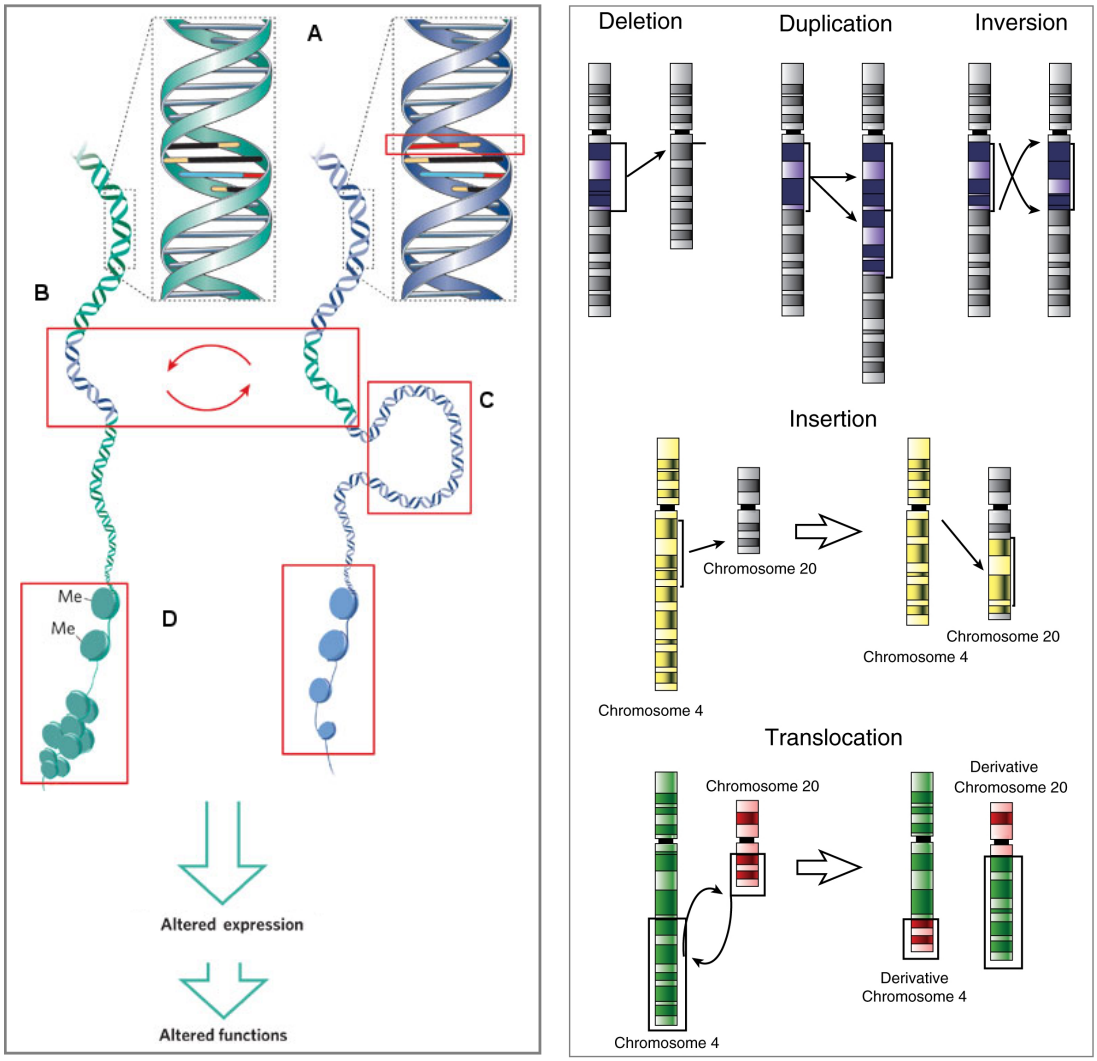
It is obvious that tumor classification according to the above mentioned criteria allows for subjective judgement. Moreover, the need for adapting the definition of the classes to particular cancer subtypes (see staging above) indicates that a very general system is not performant enough (cancer is not only one disease!) and a more specific stratification of patients is needed. In this context, it has become clear that phenotypic indicators such as tumor size, cell differentiation, tumor spread or age of the patient are not specific enough and thus insufficient for good cancer treatment. As a response to these shortcomings, the field of cancer research is currently driven by the discovery of molecular biomarkers – genetic, epigenetic, transcriptomic (and other) aberrations that are mechanistically directly accountable for the response to treatment, patient survival, etc. The discovery of molecular biomarkers and their inclusion in common clinical practice are a necessary step towards personalized cancer therapy, which approaches each tumor as a distinct disease and composes a combination of therapeutical targeting the particular molecular aberrations displayed.

The task of biomarker discovery is difficult, due to the large amount and variety of aberrations occurring in tumors. The next section addresses the most intensely studied types of aberrations frequently occurring in the cancer genome, epigenome or transcriptome and describes some of the known associations between them. The relevance of the aberrations in relation to tumor phenotype is discussed.

In what follows, we will use the short term *cancer omics* for referring to the combined fields of cancer genomics, epigenomics and transcriptomics.

2.1.3. Molecular hallmarks of cancer

Molecular changes that characterize cancer cells can be irreversible, such as genetic aberrations or reversible, for example epigenetic changes (Chin and Gray, 2008). These changes alter the expression and function of genes and regulatory factors (Figure 2.2a). In general, of highest relevance is the gain of function of proto-oncogenes and loss of function of tumor suppressors. *Proto-oncogenes* code for proteins that include the group of growth factors (have the role of inducing cell proliferation), receptor tyrosine kinases (cell surface receptors for growth factors) and others. Gain of function of proto-oncogenes transforms



(a) (b)

Figure 2.2.: a) Cancer omics (adapted from Chin and Gray (2008)): A) somatic point mutations; B) translocation; C) amplification; D) DNA methylation and histones. b) Types of structural aberrations.

these genes into *oncogenes* and contributes to tumorigenesis. *Tumor suppressors* and the proteins they code for have a repressive role in cell proliferation by regulating the cell cycle or by inducing apoptosis. Loss of function of tumor suppressors can trigger carcinogenesis.

The most intensively studied alterations that lead to gain and loss of function are somatic point mutations, deletions, duplications, amplifications, insertions, inversions, translocations, polysomies, histone modifications, DNA methylation.

Point mutations consist of exchanges of a single base nucleotide for another, for example a T-A pair mutates into a C-G pair. If the mutation changes the amino-acid code and the new amino-acid has different properties than the wild type, then the function of the resulting protein can be altered. Mutations of oncogenes are often associated with cancer onset and large efforts are being made for designing drugs that target the mutated trnscripts. For example, a mutation of the EGFR (Epidermal Growth Factor Receptor) gene that leads

to increased cell proliferation is frequently observed in non-small-cell lung cancers (Pao et al., 2004). The drugs gefitinib and erlotinib have been designed to inhibit the EGFR protein product and studies show that patients with EGFR mutations respond better to the treatment with these drugs than the those without mutations (Pao et al., 2004; Lynch et al., 2004). In Sos et al. (2009), we show that mutations of the gene KRAS in non-small-cell lung cancers influence the response to Hsp90 (heat shock protein 90) inhibitors. Another prominent example is the proto-oncogene BRAF that codes for a serine/threonine-protein kinase, which is mutated in 60% of the malignant melanomas, 10% of colorectal cancers and smaller proportion of other cancers (Davies et al., 2002). Much research is currently driven by the goal of designing BRAF inhibitors (King et al., 2006). PTEN (phosphatase and tensin homolog) tumor suppressor gene, involved in cell cycle regulation, is frequently inactivated by mutation in gliomas (Duerr et al., 1998), endometrial cancer (Tashiro et al., 1997), prostate cancer (Gray et al., 1998) and bladder cancers (Cairns et al., 1998).

Deletion occurs when a DNA segment is missing from a chromosome, possibly leading to the loss of function of genes located within this region (Figure 2.2b), via under-expression. Deletions of tumor suppressors are early events in cancer progression (Dong, 2001). For example, the gene CDKN2A (cyclin-dependent kinase-4 inhibitor) coding for the INK4A protein is a tumor suppressor frequently deleted in human cancers such as melanomas, gliomas, lung cancers, leukemias (Nobori et al., 1994), bladder cancer (Williamson et al., 1995) and others. The tumor suppressor gene PTEN is inactivated via deletion in many cancers (as well as by mutations, see above) (Wang et al., 1998; Cairns et al., 1998) and the deletion is usually associated with poor outcome. Frequent deletion of the 1p36 locus is reported in cancer studies dedicated to colorectal cancer, breast cancer, cervical cancer, neuroblastoma, leukemias and lymphomas (Bagchi and Mills, 2008). Although the search for tumor suppressors located within this region continues, several candidates have been proposed (Bagchi and Mills, 2008). Neuroblastomas with 1p36 deletions are reported to have worse outcome than the ones without this deletion (Attiyeh et al., 2005). Gene deletions are difficult to correct by drug therapy, therefore currently they are only used as diagnosis and prognosis markers.

Duplications and *amplifications* consist of the existence of two or more copies of a DNA sequence in the cancer genome (Figure 2.2b), leading to gain of function of dosage-sensitive genes located within the sequence, due to over-expression. Amplification of oncogenes is thought to be one mechanism through which tumors acquire drug resistance. A well known amplification occurring in 30% of the primary breast cancers involves the oncogene ERBB2, usually associated with short survival time, short time to relapse (Slamon et al., 1987) and resistance to tamoxifen, a drug commonly prescribed to estrogen positive tumors. The drug transtuzumab was developed to treat breast tumors with ERBB2 amplification (Pegram and Slamon, 2000). Amplification of the members of the MYC gene family *c*-MYC (8q24.21) or *n*-MYC (2p24.3) which code for transcription factors is another prominent event in solid tumors. *n*-MYC (or MYCN) is amplified in neuroblastomas and is associated with poor outcome (Brodeur et al., 1984). *c*-MYC is amplified in breast (Varley et al., 1987) and lung cancers (Little et al., 1983).

Insertion refers to the displacement of a DNA sequence from one location (chromosome) and its insertion to another location (chromosome) (Figure 2.2b). If translated together with the new surrounding sequence, the insertion can alter the resulting protein and its

function.

Aneuploidy and *polyploidy* occur when the wrong number of chromosomes is present in a cancerous cell. In normal cells, each autosome exists in two copies (homologous chromosomes). Aneuploidy entails the gain or loss of individual chromosomes and polyploidy involves extra copies of the entire genome, often three (*triploidy*) or four (*tetraploidy*). Polyploidy is believed to increase the potential of the cell to generate heterogeneity (Merlo et al., 2010; Risques et al., 2001). In a review article, (Merlo et al., 2010) observe that more than two copies of the DNA ensure the existence of all necessary elements for the viability of the cell, while providing extra copies for mutations and aberrations. Studies have shown that tetraploidy is an early event in tumorigenesis (Olaharski et al., 2006) and it is believed to precede aneuploidy, which arises from a tetraploid state after chromosomal loss (Merlo et al., 2010; Olaharski et al., 2006).

The mechanisms through which aneuploidy and polyploidy relate to tumor progression are difficult to assess, mainly because it is difficult to identify causative genes (Merlo et al., 2010). Aneuploidy is a very common event (Sen, 2000) in solid tumors and it is in general associated with poor patient outcome (Barlogie et al., 1980). More recent studies use high-resolution experiments for establishing a more precise relation between aneuploidy and survival, for example in the case of non-small lung cancer (Choma et al., 2001) or endometrial cancer (Suehiro et al., 2008). However, there are tumor types for which it has been shown that polyploidy and to a lesser extent, aneuploidy are associated with a favorable outcome and increased survival. Such is the case of childhood acute lymphoblastic leukemia (Raimondi et al., 2006) and neuroblastoma (Kaneko and Knudson, 2000; Nakazawa, 1993). In neuroblastoma, the favorable outcome correlates with polyploidy mostly for patients younger than two years and the biological basis of the association is still not understood.

Translocations occur when DNA segments are interchanged between non-homologous chromosomes (Figure 2.2b). A negative effect can occur if the displaced segments are subject to new regulatory mechanisms, which can lead to gain and loss of function. The first translocation linked to cancer was the so called 'Philadelphia chromosome' (Nowell, 2007), consisting of an interchange between chromosome 9 and 22 which leads to the formation of the BCR-ABL gene fusion of oncogenic nature. ABL is a proto-oncogene with role in cell division, which is turned into an oncogene after the juxtaposition to the BCR gene. The fused gene BCR-ABL codes for a tyrosine-kinase that inhibits DNA repair and promotes genomic instability. The Philadelphia chromosome is observed in the majority of the patients with chronic myeloid leukemia (CML) and its transcript is targeted by the drug imatinib mesylate (Druker et al., 2001), which is a tyrosine-kinase inhibitor. Another well studied translocation involves the relocation of the c-MYC proto-oncogene to a locus nearby the promoter of the IGH (immunoglobulin heavy locus) gene, resulting in overexpression of c-MYC in Burkitt's lymphoma (Taub et al., 1982; Kanungo et al., 2005).

DNA methylation is a mechanism for regulating gene expression that typically involves the addition of a methyl group to a cytosine (C) in a CpG dinucleotide context. CpGs in the genome are not randomly distributed, but tend to agglomerate within CpG rich regions called CpG islands. In normal somatic cells, CpG islands are predominantly unmethylated. However, methylation of promoter CpG islands occurs, with the purpose of suppressing gene expression. In cancer cells, *hypermethylation* and *hypomethylation* is observed, as

mechanisms of gene silencing (and loss of function) and of gene activation (and gain of function), respectively. Hypermethylation of the promoter region was reported to activate the tumor-suppressors RB (retinoblastoma) by Sakai et al. (1991), VHL (von Hippel-Lindau tumor suppressor) by Herman et al. (1994), CDKN2A by Herman et al. (1995); Merlo et al. (1995) and BRCA1 by Esteller et al. (2000). Mitotic recombination has been shown to favorably occur in hypomethylated regions, leading to deletions and translocations (Eden et al., 2003). Hypomethylation can also disrupt genomic imprinting (silencing of one allele via methylation), as in the case of the IGF2 gene (insulin-like growth factor), which plays a role in the formation of colorectal cancer (Cui et al., 2003) or Wilms tumor (Feinberg, 1999).

Histones are proteins which are normally found in structural units called nucleosomes, around which the DNA is wound for the purpose of packing it into chromatin. They are grouped into several families: H1, H2A, H2B, H3 and H4, located at different positions in the nucleosome and having different functions. Histones participate in the regulation of gene expression through post-translational *modifications* such as lysine acetylation, arginine and lysine methylation and serine phosphorylation in their tails (Esteller, 2008). Acetylation of histone lysines (K) is generally associated with transcriptional activation. The effect of histone acetylation and methylation depends on the residue (lysine or arginine) and also on the location (e.g. K4, K9, K20). Methylation of H3 at K4 has been shown to activate transcription, whereas methylation of H3 at K9 or K27 or H4 at K20 is generally associated with repressed transcription (Esteller, 2008). It has been observed that hypermethylation of CpG island promoters of tumor suppressor genes is usually associated with deacetylation of H3 and H4, loss of trimethylation at K4 of H3 and gain of H3 trimethylation at K9 and K27(Jones and Baylin, 2007; Ballestar et al., 2003). In the absence of CpG islands, tumor suppressors can be inactivated by hypo-acetylation and hypermethylation of the H3 and H4 histones (Richon et al., 2000). Certain cancers disrupt the normal function of certain histone modifying genes, for example through translocation events and gene fusion as reported in leukemias and sarcomas (Esteller, 2007).

The studies cited above reveal the tight interplay between the various types of alterations occurring in cancers. The activation of an oncogene or the inactivation of a tumor suppressor may be realized by different mechanisms, which may be specific to the anatomic location of the cancer or they may not. Moreover, it is common that upon treatment with an inhibitor, an oncogene would acquire mutations or start producing more oncoproteins via amplification, allowing the tumor to become resistant to treatment (Nardi et al., 2004). Some authors believe that effective cancer treatment should be an iterative process in which first, second, third generation inhibitors should be administered for counteracting the mechanisms of resistance acquired by the tumors (Chin and Gray, 2008). For such an approach to come into practice, biomarkers should be identified for each stage of tumor progression.

Luckily, owing to the rapid development of biotechnology, the field of cancer omics benefits now from a large amount of experimental evidence that is rich in two respects: the measurements are genome-wide and high-resolution and the cohorts of tumors become larger and larger. Consortia such as The Cancer Genome Atlas[1] (TCGA) or the Cancer

[1] http://cancergenome.nih.gov

Genome Project[2] (CGP) have dedicated large resources to harvesting multidimensional omics data that are now public. The bioinformatic analysis of these data in concert will surely allow for the identification of key biomarkers and would be the basis of a successful battle against cancer.

2.2. Molecular mechanisms of formation of copy number alterations

The mechanisms through which cells acquire structural aberrations are in tight connection with the mechanisms of DNA repair, which have the role of counteracting various types of DNA damage (or lesions). Such processes are necessary, because in highly complex genomes like the human genome, DNA damage in cells occurs very often, from 1000 up to a million lesions per cell per day (Lodish et al., 2003). The causes of DNA damage are various, from normal metabolic activities to environmental factors (such as radiation). Some of the lesions are harmless, occurring without disruption of essential function, are passed on to daughter cells and may even lead to DNA mutations beneficial for evolution. Other lesions are harmful, affecting the survival of daughter cells. In such cases, DNA repair processes are activated for restoring the integrity of the DNA and ensuring the normal functioning of the cell. Occasionally, DNA repair mechanisms fail to re-establish the proper functioning of the cell. In such cases, the cell's fate takes one of the following courses: it either enters a state of *senescence*, which means that it loses the ability to replicate but continues to exist because its physical presence is required by the organism for spatial reasons, or undergoes *apoptosis* (programmed cell death). Rarely, DNA repair mechanisms re-construct the DNA incorrectly, which allows the cell to continue to divide and proliferate aberrations that can lead to the formation of tumors. Below we present types of DNA damage and erroneous DNA repair which can lead to structural aberrations and tumorigenesis.

Double-strand breaks (DSB) are lesions of the DNA in which both strands of the double helix are damaged. DBSs are troublesome because neither strand can be used as a template for repair. Several cellular mechanisms can be activated for re-constructing the DNA upon DBSs, among which the best understood are *homologous recombination*(HR), *non-homologous end joining*(NHEJ) and *microhomology-mediated end joining*(MMEJ). These alternative processes are activated in different stages of the cell cycle, generally involve different molecular pathways and thus proteins, have different degrees of repair accuracy and substitute for each other if one or more mechanisms become dysfunctional (Pastink et al., 2001).

HR uses existing homologous sequences situated on the sister chromatid, homologous chromosomes or elsewhere in the genome as templates for restoring the broken DNA site. A simplified illustration of HR is shown in Figure 2.3a (Sharan and Kuznetsov, 2007): as a consequence of damage (B), the DSB locus is first subject to resectioning, which consists of DNA degradation from the 5' end towards the 3' end, producing single-stranded overhangs at the 3' ends (C). The RPA protein then binds the 3' overhangs (Wold, 1997) and together with the Rad51 protein and several other proteins, forms a filament of nucleic acid and

[2]http://www.sanger.ac.uk/genetics/CGP/

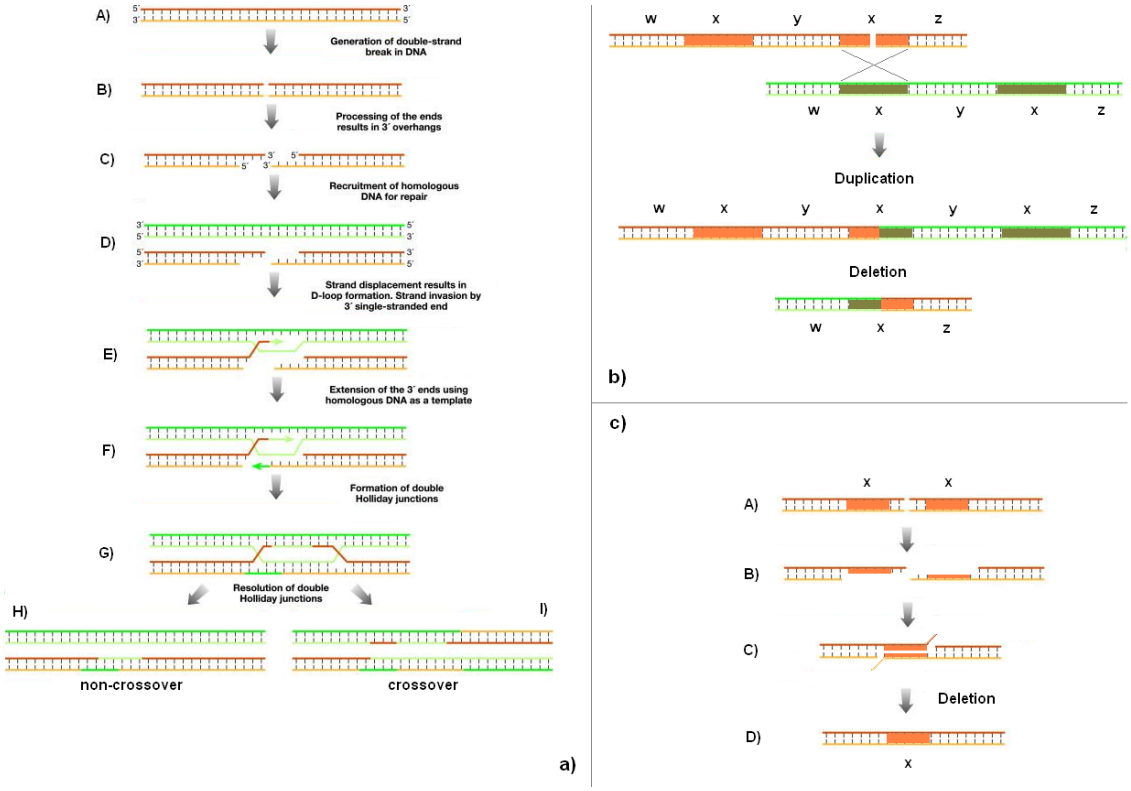


Figure 2.3.: a) Schematic representation of homologous recombination (Sharan and Kuznetsov, 2007). b) Schematic representation of unequal crossover. c) Schematic illustration of the single-strand annealing and of microhomology-mediated end joining.

protein (Sharan and Kuznetsov, 2007). This nucleoprotein filament searches for DNA sequences homologous to that of the 3' overhang. When such a sequence is found (D), the single-stranded nucleoprotein filament invades the homologous DNA in a process called strand invasion (E). The missing DNA sequence is synthesized (F). This process leads to the formation of two cross-structures that hold the two chromosomes together (called Holliday junctions, first introduced in Holliday (1964)) (G). The Holliday junctions are resolved by cleavage, which can result either in non-crossover (H) or crossover products (I).

Usually, HR is a very efficient and accurate mechanism, especially when it relies on the sister chromatid for repair, which offers an identical template. As a consequence, HR works best in the S (synthesis) and G_2 (pre-mitotic) stages of the cell cycle, when the sister chromatid is accessible. If the sister chromatid is not available and a homologous chromosome is used as a template, HR may lead to loss of heterozygozity (LOH)[3] by copying the undamaged allele. LOH is a structural aberration often responsible for tumorigenesis, because it can activate a mutant recessive allele (Pastink et al., 2001). If a direct repeat is used as homologous template – a process called *non-allelic homologous recombination – unequal crossover* can occur. Figure 2.3b illustrates the mechanism of unequal crossover: assume the DSB site is located within a repeat, marked in the figure by x. During DNA

[3]Loss of heterozygosity (LOH) in a cell represents the loss of normal function of one allele of a gene in which the other allele was already inactivated.

repair via homologous recombination, it can happen that the homologous DNA used for repair is misaligned (Figure 2.3b), which results into two crossover products containing a deletion and a duplication, respectively. At the next division of the cell, the two products are transferred to daughter cells, which leads to copy number is changes (Hastings et al., 2009).

It has been shown that HR is vulnerable to sequence repeats, which can mislead homology search (Hastings et al., 2009). For example, an alternative type of HR called *single-strand annealing* (Figure 2.3c) occurs when homologous sequences flank the DSB site – a situation which can easily occur if the DSB lies within repeats. In this case, the two ends anneal after the DNA has been resectioned, which leads to the *deletion* of the region enclosed between the homologous sequences. Figure 2.3c illustrates the steps of SSA: a DSB occurring between two homologous sequences (A) is subject to resectioning (B), exposing the two homologous sequences; the DNA complementary strands anneal to form a double-stranded helix (C), the remaining flaps are removed, the missing regions are filled in by synthesis and ligation completes the process of DNA repair (D). SSA requires at least 30bp homology in order to take place (McVey and Lee, 2008). The SSA pathway in eukaryotes involves the protein RAD52, which is required for the annealing if the single stranded DNA, as well as the RAD59 protein.

A similar process to the SSA is the MMEJ (McVey and Lee, 2008), which relies on different proteins and on much shorter homologous sequences (5-25bp) for re-joining the DNA strands. MMEJ results in deletions in the same way that SSA does (Figure 2.3c).

Many DNA breaks do not present enough homology at the ends, which triggers NHEJ processes which need no homology or very little homology to match the end sequences (0-5bp) (McVey and Lee, 2008). Consequently, the DNA repair process may not be very accurate (full complementarity is not ensured) and may require the deletion or insertion of several bases (1-4bp)(McVey and Lee, 2008; Hastings et al., 2009). NHEJ is therefore not as accurate as HR, but it has the advantage that it can function throughout the cell cycle (HR is limited to stages S or G_2) and in extreme cases when no homology is available (Lieber, 2007).

A special type of DNA damage that results in gene amplification affects the ends of the chromosomes called *telomeres*. Telomeres are 6bp repeat sequences that are associated with special chromatin proteins that protect the ends of the chromosomes. In germline cells, the existence of the telomeres is ensured by the activity of the telomerase protein, which has the role of re-constructing the telomeres if shortened or deleted. In somatic cells, the telomerase is not expressed and with every cell division, the telomeres are shortened and not restored. This mechanism marks the age of the cell and thus allows old cells to be recognized and become senescent or undergo apoptosis. Cancer cells try to avoid cellular death due to telomere loss, which frequently leads to *chromosome fusion* (Lo et al., 2002). Fusion or bridging between two chromosomes lacking a telomere results in the formation of a *dicentric chromosome* (with two centromeres) with telomeres at both ends. During anaphase, the two centromeres are pulled apart and the dicentric chromosome breaks, resulting in another chromosome lacking a telomere. The cycle is repeated, causing amplification. First described by McClintock (1941), this mechanism of amplification is called *breakage-fusion-bridge cycle* (B/F/B). We show a schematic illustration of B/F/B in Figure 2.4a. There is evidence that B/F/B is responsible for low-copy gene amplification

in cancer cells and may be an early step in high-copy gene amplification (Singer et al., 2000; Toledo et al., 1993).

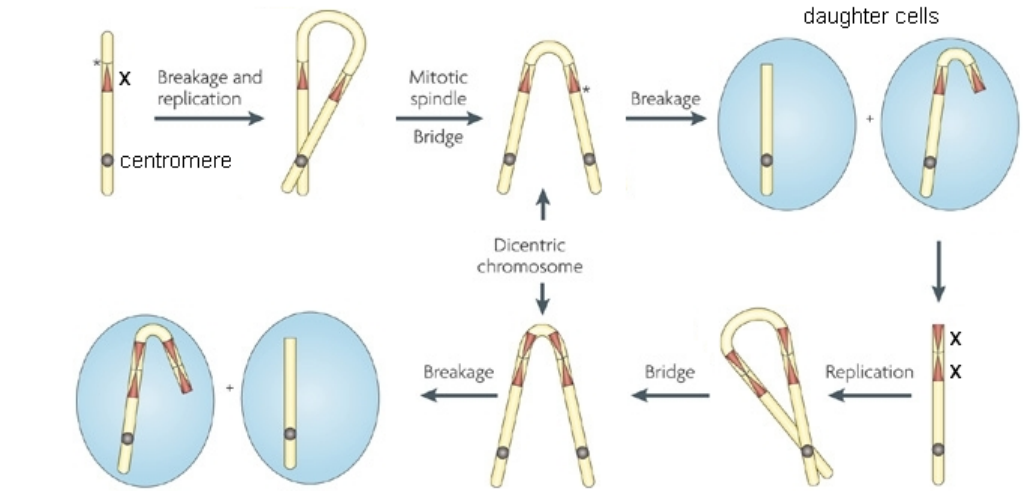
Figure 2.4b depicts the formation of so called *homogeneously staining regions*(HSR), very long regions of amplification which have been frequently observed in tumors. The name reflects their appearance,which is uniform in color after fluorescent staining, due to abundant and homogeneous gene content.

During B/F/B, recombination can occur between the homologous sequences involved in amplification , resulting in *double minute chromosomes* (DM) 2.4. DMs are ring-shaped DNA segments that exist in tumor cells despite not having a centromere or telomeres. They generally consist of a large number of copies of an oncogene, sometimes rearranged, with or without the sequence around it, the existence of which appear to confer selective advantage to the cell. DMs have been first reported to occur in tumors of neuroectodermal origin such as neuroblastoma or gliomas, but have been identified in many other tumor types thereafter (Hahn, 1993). They have been frequently associated with oncogene overexpression and drug resistance. Figure 2.4c shows the MYCN oncogene (red) amplified in four neuroblastoma cell lines via DM and HSR, clearly noticeable by FISH experiments (see Section 2.3 for details on FISH experiments).

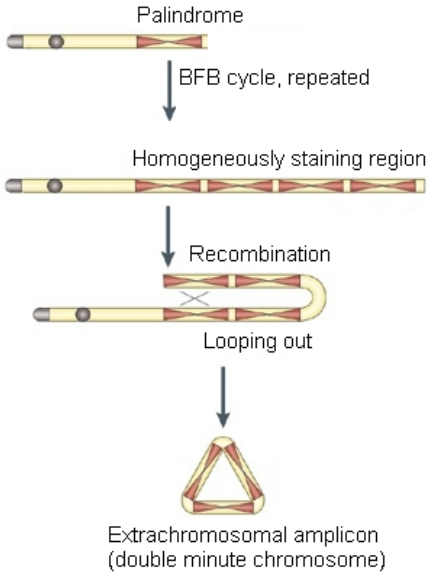
DNA lesions and misfunctioning of the repair machinery are responsible for segmental copy number changes. Whole chromosome gains and losses (aneuploidy), which represent a large part of the common aberrations occurring in cancer, arise via different molecular mechanisms. These mechanisms are consequences of the disfunction of the mitotic cycle, which consists of the following phases: *interphase*, *prophase*, *prometaphase*, *metaphase*, *anaphase*, *telophase* and *cytokinesis*. In the phases preceeding metaphase, the DNA content is duplicated, the chromatin is condensed into chromsomes, the nuclear membrane is disintegrated and a molecular machinery is created which generates the forces necessary to separate correctly the DNA and pull apart the new cells. At metaphase, the *mitotic spindle* is formed: subject to tension towards two opposite poles of the cell, the chromosomes align such that their centromeres lie along the *equatorial plane*, an imaginary line equidistant to the two poles. At anaphase, the sister chromatid is cleaved and the resulting chromosomes start migrating towards the poles of the cell. At telophase, the decondensing chromsomes are encapsulated in the nuclear membrane. Cytokinesis is the last phase of mitosis, during which a *cleavage furrow* forms at the former location of the metaphase equator, separating the two nuclei. The cleavage furrow develops into a cell wall, which separates the two daughter cells and ends the mitosis.

One of the cellular mechanisms which ensure the correct alignment of the chromosomes at the equatorial plane and the correct separation of the DNA content at metaphase is the *mitotic spindle checkpoint* (Pellman, 2007). The protein CENP-E (centromere protein) plays a role in this process and it has been observed that heterozygozity of the corresponding gene (only one working copy) can lead to the failure of the mitotic spindle checkpoint. In Figure 2.5a, bottom row (Pellman, 2007), the incorrect alignment at the equatorial plane is illustrated, with the consequence of aneuploidy in the daughter cells.

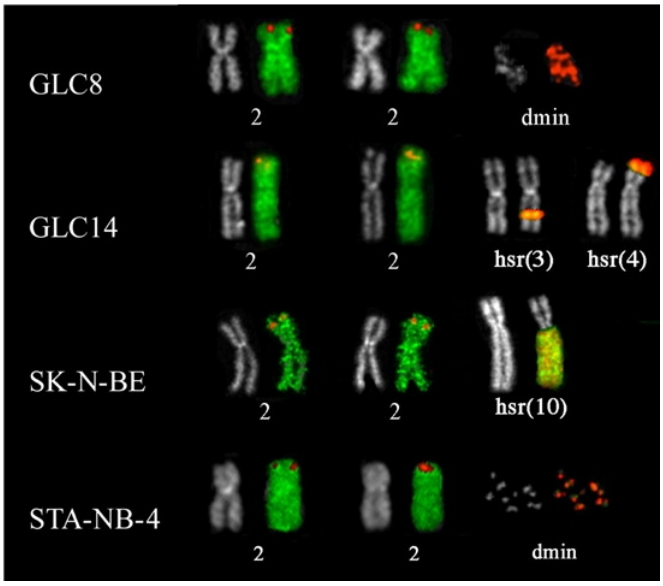
Pellman (2007) describe alternative mechanisms of formation of aneuploidy, caused by overexpression of the MAD2 protein, which is also involved in the mitotic spindle checkpoint. Due to misalignment at the equatorial plane, the cleavage furrow is retracted and the cytokinesis phase is not completed (Figure 2.5b, top row). A tetraploid cell forms and



(a)



(b)



(c)

Figure 2.4.: a) Breakage-fusion-bridge cycle (Tanaka and Yao, 2009). A chromosome break near the telomere is marked by asterisk and a DNA sequence located in the vicinity of the break is indicated by x. After cell replication, the two chromosome copies lacking telomeres fuse, forming dicentric chromosomes. The centromeres are pulled apart at anaphase and an uneven break occurs. Consequently, two asymmetric chromosomes are formed, one laking the sequence x and the other containing a duplication of x. The duplication is palindromic, meaning that the two copies are oriented differently. The chromosome containing the duplication is further replicated in the daughter cell. Because they lack telomeres, the two sister chromosomes fuse again and the cycle continues, leading to amplification. b) Homogeneously staining regions consisting of high amplification of a particular DNA sequence as a result of repeated B/F/B. Recombination can occur between the homologous sequences involved in the amplification (marked with X), leading to the formation of double-minute chromosomes (Tanaka and Yao, 2009). c) FISH experiments showing amplification of the MYCN gene in four neuroblastoma cell lines via DM and HSR (Storlazzi et al., 2010).

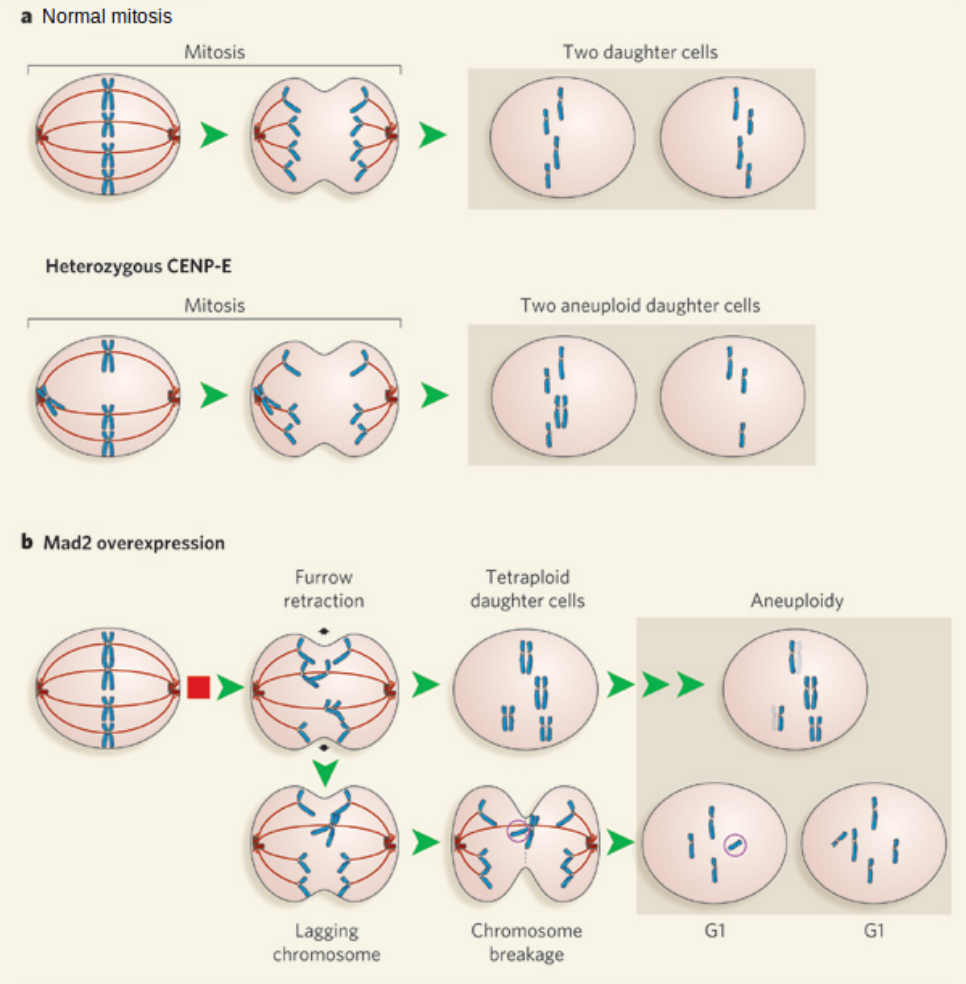
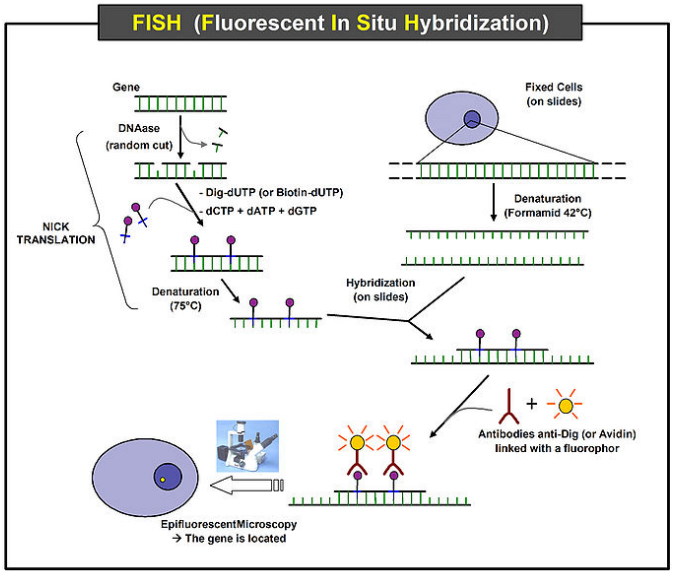
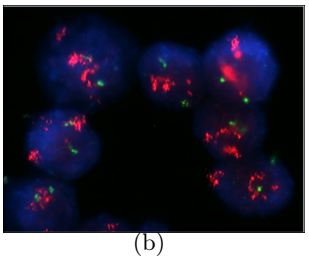


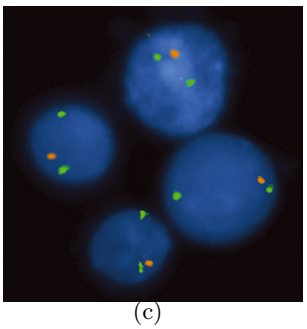
Figure 2.5.: Formation of aneuploidy (Pellman, 2007). a) Normal separation of chromosome copies during mitosis. b) In CENP-E heterozygous cells, incorrect positioning of the chromosomes during the mitotic spindle gives rise to aneuploidy in daughter cells. c) In cells with MAD2 overexpression, cytokinesis failure can occur (top row), meaning that the cell fails to divide and the duplicated DNA content gives rise to tetraploid cells, which later on can lead to aneuploidy by loss of DNA. The bottom row illustrates how anaphase lag can lead to the isolation of a chromosome into a micronucleus, which is lost from the daughter cell and leads to aneuploidy.



(a)



(b)



(c)

Figure 2.6.: a) Schematic illustration of the FISH procedure. Image from Wikipedia (http://en.wikipedia.org/wiki/Fluorescence_in_situ_hybridization). b) Amplification of HER-2 gene detected by FISH. Image from a study by Shigematsu et al. (2011). c)Deletion of 5q detected with FISH. Image from Cerveira et al. (2003).

after subsequent deletions, aneuploid cells emerge. Alternatively, incorrect separation of the chromosomes can occur due to *anaphase lag*, which consists of delayed movement of a chromosome towards the corresponding pole. In such a case, the lagging chromosome is isolated into a micronucleus and lost from the daughter cell, causing aneuploidy.

2.3. Experimental assays for determining DNA copy number aberrations

Below we present the technologies that allow for the estimation of DNA copy number, from the earliest to the most modern approaches. We pay particular attention to the arrayCGH assays, because the bioinformatics methods presented in this thesis have been validated on arrayCGH data. However, the most efficient technology for genome-wide DNA copy number measurements, albeit more expensive at the moment, is the recent next generation sequencing (NGS).

2.3.1. Fluorescence in situ hybridization (FISH)

Fluorescence *in situ* hybridization has been developed in the early 1980's (Langer-Safer et al., 1982). It has been used and continues to be used for detecting the presence or absence of a particular DNA sequence in the genomes of cells. Figure 2.6a presents the main steps of this procedure. The target probes, corresponding for example to a gene of interest, are isolated, purified and amplified, then they are tagged with *fluorochromes*. Fluorochromes are small molecules that form covalent bounds with the DNA fragments and emit light at specific wavelengths, which can be detected by fluorescence microscopy.

The probes of interest are hybridized onto metaphase chromosomes, which are fixed on glass slides and after about 12 hours the partially hybridized or the unhybridized probes are washed away. With image analysis, the abundance of the DNA sequence of interest is detected. For example, in Figure 2.6b, amplification of the HER-2 gene located on chromosome 17 in a population of breast cancer cells is evident by the uneven proportion of red spots (corresponding to the HER-2 sequence) and green spots (marking the centromeres of the pair of chromosomes 17), respectively. Similarly, deletion of chromosome arm 5q is observable in Figure 2.6c, where two signals corresponding to the control loci at 5p15 are visible in green, but only one signal corresponding to a probe located at 5q33-34 locus is present (in orange).

FISH assays are suitable for investigating the copy number state of a known biomarker, for instance in cancer diagnostics. For the purpose of biomarker discovery and in general, genome-wide copy-number analysis, an extension of the FISH technology called Comparative Genomic Hybridization (CGH) has been proposed.

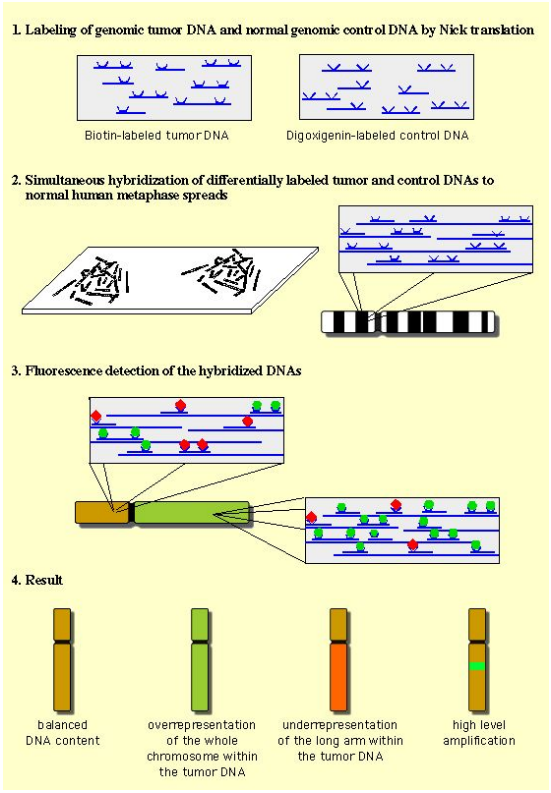
2.3.2. Comparative genomic hybridization (CGH)

The first efficient technique for genome-wide copy number measurement was called CGH and was proposed for the first time in 1992 by Kallioniemi et al. (1992). The experiment consists of the following steps (see also Figure 2.7 for schematic representation): first, DNA from tumor and normal tissue is separately isolated and tagged with different fluorescent labels (for example, red and green in the figure). The labeled DNA is mixed and *hybridized* onto metaphase chromosomes. The core principle of this technique is therefore DNA hybridization, which refers to the property of complementary DNA sequences to pair with each other forming hydrogen bonds. The last step is image analysis, which is used to determine the ratio of hybridization intensities (red/green) along the chromosomes, which ideally mimics the true ratio of DNA copy number between the tumor and control tissue.

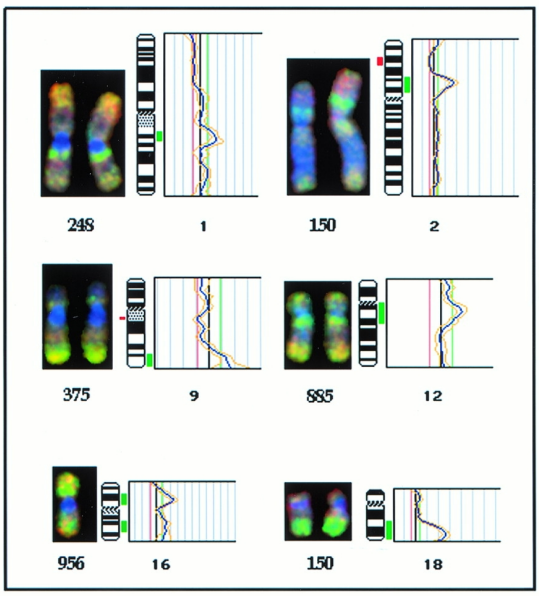
Using metaphase chromosomes was one of the main tools for cytogenetic studies in the early 1990s. However, the chromosomes are highly coiled and therefore the resolution of the copy number measurements is not very high. It is appreciated that high-level amplifications (tens or hundreds of copies) can be detected by CGH if their length is larger than roughly 1Mb, whereas deletions must be at least 5 to 10Mb long in order to be identified. Figure 2.7b shows an example output of a CGH experiment: images of chromosomes give a gross indication of neutral, loss or gain regions.

2.3.3. arrayCGH

Array-based comparative genomic hybridization (arrayCGH or aCGH) uses similar principles as classical CGH, but affords much higher resolution by replacing the metaphase chromosomes with microarray plates as hybridization base. The technique was described first in 1997, by Solinas-Toldo et al. (1997) and Pinkel et al. (1998). The main steps of array-based CGH are schematically shown in Figure 2.8. Generally, tumor and control DNA are labeled with different fluorochromes (usually from the group of cyanines) and are co-hybridized onto DNA microarray plates, then a special scanner measures the intensities of each fluorochrome separately. The ratio of these intensities is ideally proportional to the true copy number.



(a)



(b)

Figure 2.7.: a) Schematic illustration of the CGH. Image from Wikipedia (http://en.wikipedia.org/wiki/Comparative_genomic_hybridization). b) Typical low-resolution signal obtained by CGH experiments. Image from an early study by Rao et al. (1998).

A DNA microarray consists of a large number of DNA spots called *probes* or *targets*, attached to a solid surface. Each spot contains many copies of a specific DNA sequence, which is typically a part of a gene or other sequence of interest. Several types of DNA microarrays exist, depending on how the target sequences are obtained. The initial technique that was proposed by Solinas-Toldo et al. (1997) and Pinkel et al. (1998) makes use of large BAC clones (*bacterial artificial chromosome*), which can be up to a few hundred kilobases long. A BAC is an engineered DNA molecule used to clone DNA sequences in bacterial cells (for example in E. coli). Although high-resolution BAC arrays for many mammalian genomes are being produced, a large effort is required for obtaining enough DNA sequences to make one array (Pinkel and Albertson, 2005) and alternative methods have been proposed. For example, a very popular technique is that based on synthesis of *oligonucleotides* – short nucleic acid polymers of about fifty or fewer bases. Oligonucleotides are assembled base-by-base in an iterative process that is error-prone and sets limits to the length of the polymer. Although they are easier to obtain, the small size of the oligonucleotides and therefore decreased complexity lead to less specific hybridization, meaning that similar sequences located elsewhere in the genome can bind to a specific probe. This drawback is often compensated for by a higher resolution which can be afforded by oligonucleotide arrays.

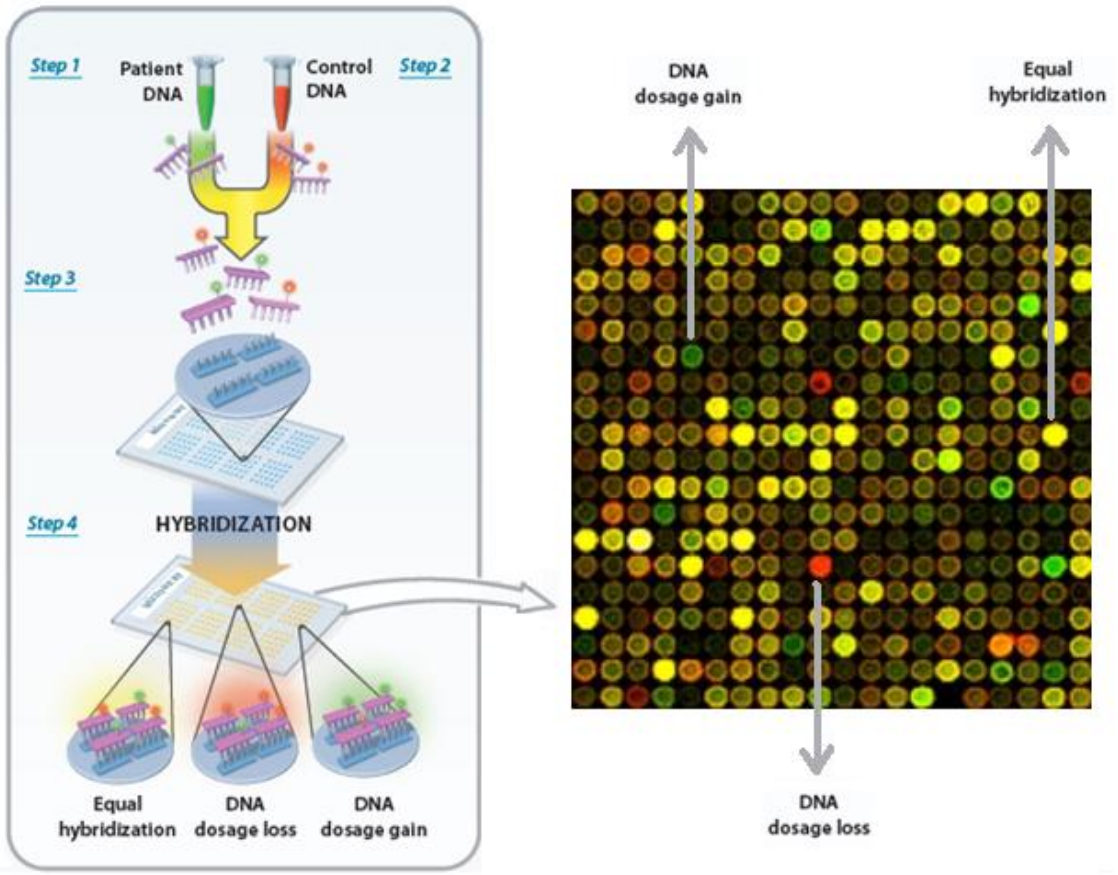


Figure 2.8.: Schematic illustration of the arrayCGH experiment. Image adapted from Theisen (2008).

Several factors influence the success of arrayCGH experiments. Among those, we mention the *heterogeneity of tumor specimens*, which refers to the existence of normal cells or cells in different stages in the tumor tissue. Such mixture introduces biases in the copy number measurements (Pinkel and Albertson, 2005). Under the simplified assumption that the mixture contains only normal and tumor cells in the same stage of progression, a correction is possible if the proportion of normal cells in the tumor specimen can be estimated.

The phenomenon of *signal saturation* is another source of bias in arrayCGH experiments. Saturation affects probes against which a very large number of DNA fragments hybridize, for example those located in highly amplified regions with hundreds of copies in the tumor. In such cases, the scanner truncates the fluorescence signal with important negative consequences for the downstream analysis. Hsiao et al. (2002) propose a method for identification of probes likely to be affected by signal saturation. Other sources of bias are the proportion of *repetitive content* in sequence, which, if high, can lead to unspecific hybridization, or the *reassociation of double-stranded nucleic acids* during hybridization. BAC arrays and oligonucleotide arrays are unequally affected by these biases and thus, for each particular application, the more appropriate arrayCGH technology must be carefully chosen (Pinkel and Albertson, 2005).

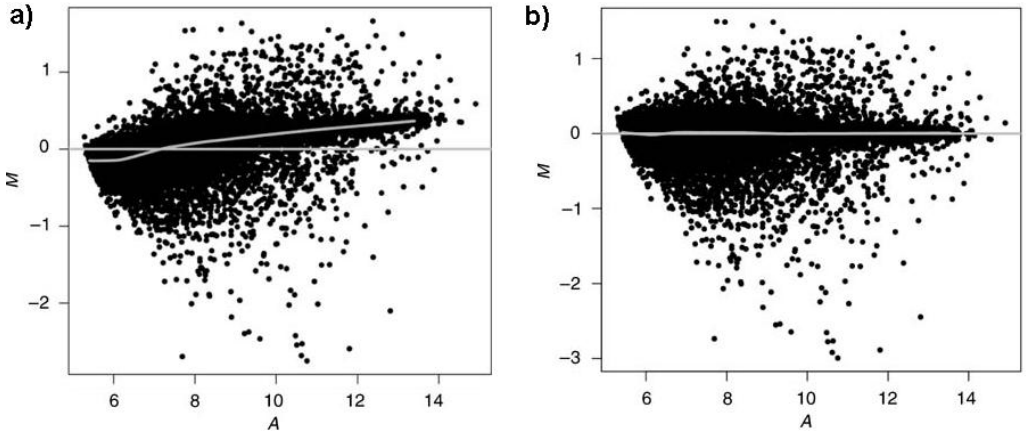


Figure 2.9.: Lowess normalization. Image adapted from Grant et al. (2001). a) M-A plot presenting the dependence between the signal intensity $A = \log_2(R * G)$ and the log-ratio $M = \log_2(R/G)$. The Lowess regression line and the horizontal through zero are shown. b) After normalization, the Lowess regression matches the null horizontal.

Low-level processing of array CGH data

The ratio between the two color intensities – let them be *red* (R) and *green* (G) for convenience – expresses the ratio of abundance of a specific sequence between the tumor and the control DNA. Traditionally, the ratio is transformed by applying the base-2 logarithm and the measurement is called *log-ratio* or *log2ratio*. A log-ratio of value $\log_2 R/G = 0$ corresponds to a target sequence that is equally abundant in tumor and control cells. An number of four copies results in a log-ratio of 1 and the loss of one copy results in a log-ratio of -1. An alternative transformation using the *arsinh* function instead of the logarithm has been proposed for expression data analysis (Huber et al., 2002), motivated by more accurate estimation of weak expression signals, however evidence that such transformation would improve copy number estimates has not been published yet.

Technology-specific biases affect the log-ratios and various normalization methods for bias removal have been proposed. The final purpose of normalization is to make comparison between different experiments more meaningful (Quackenbush, 2002). The most frequently performed normalization consists of centering the log-ratios around zero by subtracting the median value from all log-ratios. This procedure is based on the assumption that the majority of the probes have a neutral copy number. rarely, this assumption does not hold, for example in the case of near diploid or triploid cancer genomes.

Another well studied phenomenon is the dependence between the log-ratio ($log_2(R/G)$) and signal intensity ($\log_2(R * G)$). Specifically, large intensity associates with larger log-ratio and low intensity associates with smaller log-ratio. Such undesired dependence can be removed for example by using Lowess regression (Smyth and Speed, 2003). Lowess (locally weighted least squares regression) is a technique for locally fitting a smooth curve to a set of observations. The normalization based on Lowess regression transforms the data such that the regression curve corresponds to the null horizontal line (see Figure 2.9 for an illustration). Other sources of bias are the labeling scheme, meaning that log-ratio measurements depend on whether the tumor is labeled with green and control with

red or viceversa or spatial biases, which affect probes located next to each-other on the microarray.

In general, the most widely used methods for normalization of arrayCGH data have been borrowed from the field of expression data analysis. It has been observed in articles that this may not be a good practice, due to different distributions of the log-ratios and consequently arrayCGH-specific algorithms have been proposed. Neuvial et al. (2006) present MANOR, an algorithm which extends Lowess regression to accept as input two-dimensional log-ratio data, corresponding to physical locations on the CGH microarray. This way, MANOR can correct for spatial biases. Staaf et al. (2007) use a clustering approach for stratifying the log-ratios into three clusters (likely corresponding to neutral, gained and lost target sequences). Then, they use only the largest group for Lowess normalization, evidently assuming that the majority of the probes are in a normal copy number status. Despite the better performance, application studies have not adopted the normalization methods especially designed for arrayCGH data, but have continued to use the more established methods designed for expression arrays.

3. Computational analysis of DNA Copy Number Aberrations

> Number is the ruler of forms and ideas, and the cause of gods and demons.
>
> Pythagoras

3.1. Introduction

Most types of cancers accumulate during progression a large number of copy number aberrations (CNAs), affecting genomic regions of very small size (a few Kbp) to large regions spanning whole chromosomes or chromosome arms. The high genomic instability and the selective pressure acting on the tumor cells lead to very complex patterns of CNAs, that can be either common to a large subset of samples belonging to a tumor type or subtype (*recurrent aberrations*) or particular to an individual tumor sample or to a very small subset of them. A common assumption in the research field is that the recurrent CNAs are accountable for the tumor phenotype, whereas the spurious CNAs have a lesser role. Most bioinformatics methods for computational analysis of CNAs are driven by the two-fold objective of identifying recurrent CNAs in a collection of samples belonging to a tumor type or subtype and characterizing their influence on tumor phenotype.

Over the last decade, many methods for computational analysis of CNAs were published, advertising their ability of identifying CNA signatures characterizing tumor types or subtypes. However, despite the intense research, very few of these signatures are currently used in clinical practice for diagnosis, prognosis or treatment. The main critique of the studies that propose CNA signatures is their lack of reproducibility (Ein-Dor et al., 2005). Specifically, under variations of the tumor cohort investigated and of the parameters of the methods, the CNA signatures change significantly. Two possible reasons can explain the instability of the CNA signatures: first, the biological diversity of the tumor set, which cannot be corrected and second, the complexity and lack of robustness of the computational pipeline used for identifying CNA signatures. The latter source of variance can be corrected, if more robust computational pipelines are proposed. Existing pipelines apply successive processing steps, at every step making assumptions about the nature of the data. These assumptions, depending on how restrictive or permissive they are, can affect downstream analysis by obstructing relevant information or adding false information. The effects of the processing steps on stability of the signatures under variability of the tumor set are largely uncharacterized in scientific studies.

In this chapter, we first give a detailed overview of the computational pipelines commonly

used for addressing the objective of identifying and characterizing CNAs (Background Section). Then, we introduce a modified pipeline, which makes less restrictive assumptions about the nature of the data. We give a high-level description of the main steps of the pipeline. In the following two chapters (Chapter 4 and 5), we will describe these steps and the corresponding algorithms in detail.

3.2. Background

Up to present, the problem of automatically identifying CNAs and characterizing their impact on phenotype remains a challenge. Despite numerous approaches proposed in the literature, there exists no methodological gold standard, that has been proven to be applicable to a large number of different cancers (Rueda and Díaz-Uriarte, 2010). Statistical modeling becomes very difficult because of the high dimensionality of the measurements compared to the relatively small number of tumor samples available and the notorious heterogeneity of CNAs. Most computational approaches consist of pipelines involving successive steps of data processing and dimension reduction. The general workflow can be summarized by several alternative pipelines, the steps of which are schematically presented in Figure 3.1. Most of the methods adopt a two-stage pipeline, namely *single-array analysis* and *multiple-array analysis*.

The single-array analysis addresses one of the first computational challenges of array-CGH analysis: the identification of all CNAs that a particular tumor harbors, given the normalized log-ratios measured by an arrayCGH experiment.

The multiple-array analysis makes use of large collections of arrays obtained from the same type of tumor in order to identify recurring CNAs and to characterize their impact on tumor phenotype.

Below we present each step of the pipeline in details. We summarize the most popular methods that address their problematics and we comment on their advantages and drawbacks.

3.2.1. Segmentation

As already shown in chapter 2, various types of stochastic noise and bias from experimental and biological sources affect the true copy number ratios, in such a way that the experimentally determined ratios do not take discrete values of the form $n/2$, corresponding to n copies of DNA. Figure 3.2a shows the normalized log-ratios from an arrayCGH experiment. The single-array segmentation algorithms make the first step towards inferring the true copy number from the normalized log-ratios. Specifically, they estimate an optimal partition of the genome into intervals of constant copy number. The boundaries of the intervals mark locations of copy number change, which are called *breakpoints*. Formally, the segmentation algorithm assumes that the true log-ratio depends on the genomic location x via a step function θ as follows:

$$\theta : \mathcal{X} \to \mathbb{R}, \quad \theta(x) = \sum_{k=1}^{p} a_k \mathbb{1}_{I_k}(x) \tag{3.1}$$

where $\mathcal{X}$ denotes the set of all genomic positions and $I_1, ..., I_p$ are disjoint intervals that

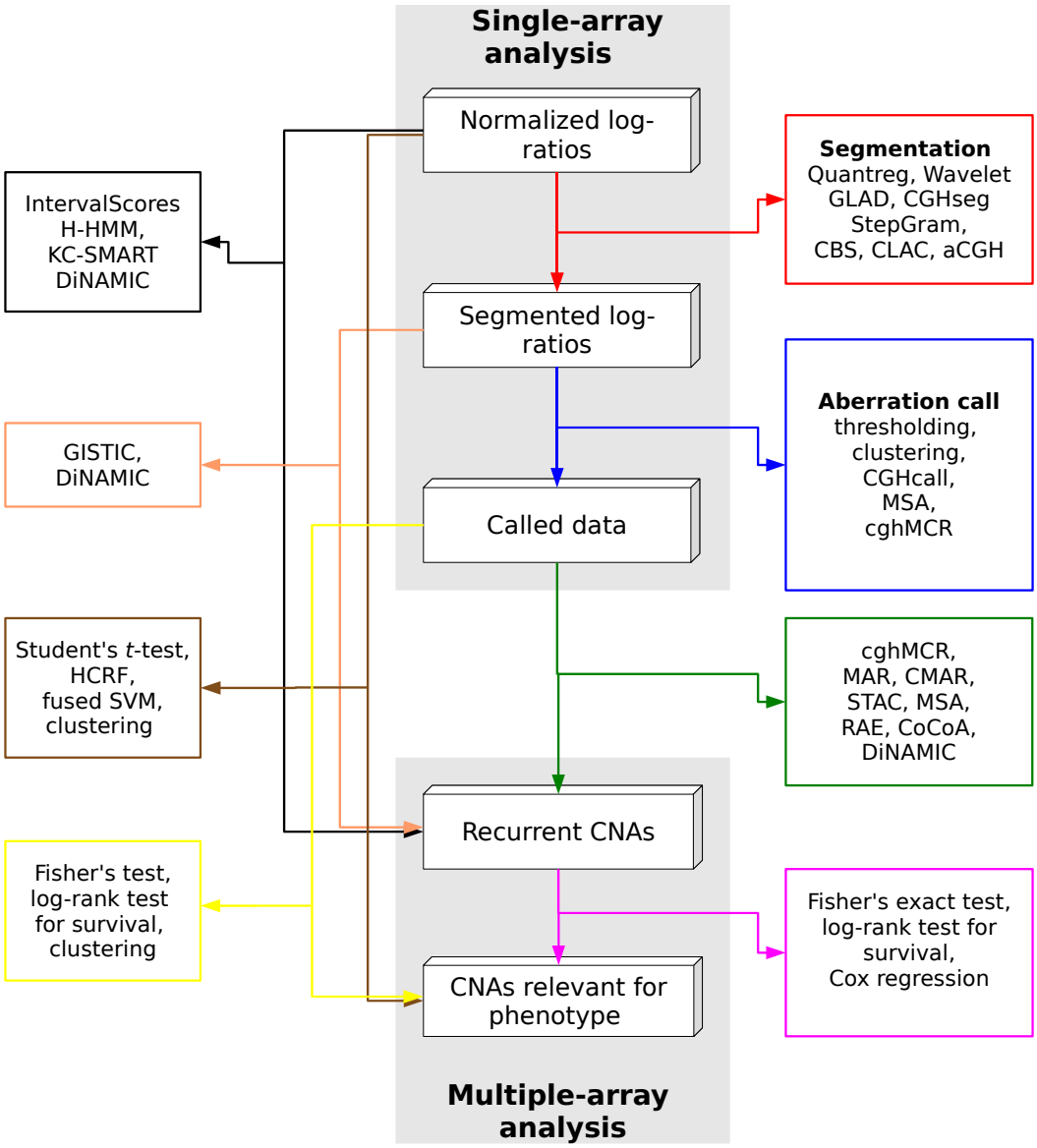


Figure 3.1.: Computational pipelines for characterizing the associations between of CNAs and tumor phenotype. In 3D boxes, the type of input/output data involved in the various steps are shown. In colored boxes, algorithms which perform the corresponding steps are shown.

cover all genome $\mathcal{X} = \bigcup_{k=1}^{p} I_k$. The number of intervals $p \in \mathbb{N}$, the intervals $I_1, ..., I_p$ and the constants $a_1, ...a_p \in \mathbb{R}$ are parameters of the function θ.

Assume that we have given an array α consisting of normalized log-ratio values $\alpha_1, ..., \alpha_d$ measured at d genomic loci $P_1 < ... < P_d$. Under the model assumption given by (3.1), the true log-ratio at P_i is $\theta(P_i)$. The observed log-ratio at P_i is α_i. The purpose of array segmentation is to use the observations $(P_1, \alpha_1), ..., (P_d, \alpha_d)$ for estimating the parameters of the true model θ. Figure 3.2b, illustrates array segmentation by an intuitive example.

In a review article, Lai et al. (2005) summarize the most popular methods for array segmentation and compare their performance on simulations and real cancer data. A large class of nonparametric solutions use appropriate smoothing or denoising techniques in order to obtain a simpler signal that is closer to the true piece-wise constant model θ. The smoothing algorithms minimize some loss function, which evaluates the distance between the observed log-ratios and the estimated log-ratios. In order to avoid oversegmentation, which decreases the loss function but increased model complexity and leads to overtraining, penalties are used to keep the number of segments p small. For example, Eilers and de Menezes (2005) use a quantile smoothing technique which minimizes the penalized L_1 distance between the observations and the estimated model, which is shown to result in sharp boundaries between segments. This approach is called Quantreg. Hsu et al. (2005) propose Wavelet, a method based on denoising by wavelets. Both Wavelet and Quantreg are shown to perform well on arrays with large signal-to-noise ratio (Lai et al., 2005), however they are not easy to interpret because they leave to the user the task of deciding whether a jump in the smoothed signal is large enough to be considered a breakpoint.

A large class of parametric methods model the observed log-ratios as the sum of the true signal, given by the function θ and Gaussian noise. Formally, for all probes $i = 1, ..., d$,

$$\alpha_i = \theta(P_i) + \varepsilon_i,$$

where ε_i are i.i.d $\mathcal{N}(0, \sigma)$. Then, a model $\hat{\theta}$ is estimated that maximizes the likelihood of the observed data. The likelihood function is usually penalized, such that the number of segments is kept small and overfitting is avoided. In Hupé et al. (2004), a weighted likelihood function is maximized in the neighborhood of each genomic position. The weights are iteratively updated, such that they reflect the maximal neighborhood of constant log-ratio. The updating of the weights follows closely the Adaptive Weights Smoothing procedure by Polzehl and Spokoiny (2002). The algorithm is called GLAD. In an evaluation on real data, GLAD has been shown to return segments consisting of single-probe outliers, which are most likely bad probes (Lai et al., 2005). An improved method is CGHseg, presented by Picard et al. (2005), based on a likelihood model with a novel penalty that is chosen such as to avoid overestimation of the number of constant intervals.

Wang et al. (2005) present the algorithm CLAC, based on hierarchical clustering of the observed log-ratios along the chromosomes, using a special distance between consecutive probes. Then, a heuristic is used for selecting clusters corresponding to segments of large or small copy number.

A different approach using Hidden Markov Models (HMM) is presented by Fridlyand et al. (2004). The authors assume that the underlying log-ratios are successive states of an HMM with a certain transition probability. The states are chosen to represent the underlying copy number: one, two, three, etc. copies. An HMM is trained to estimate the state of each probe and the transitions between different states (the breakpoints). However, the method is slow in practice and it does not perform well on real data (Lai et al., 2005).

The most frequently used and cited procedure for array segmentation is presented by Olshen et al. (2004). The authors adapt the binary segmentation procedure by Sen and Srivastava (1975), which can identify a change in mean (breakpoint) in a series of observations, to the task of discovering an arbitrarily large number of breakpoints. Their approach is called Circular Binary Segmentation (CBS). The CBS algorithm uses a statistic, which

tests the hypothesis that the interval between probes i and j has a different log-ratio mean than the rest of the chromosome. The null distribution of the test statistic is obtained via random permutations. In the review article by Lai et al. (2005), CBS is shown to perform consistently well, both on artificial and on real cancer data, the only reported drawback being the relatively slow running time. Later, the authors of CBS introduced a faster version of their method (Venkatraman and Olshen, 2007), which scales to the needs of the high-throughput arrays.

The advantage of performing segmentation as part of the analysis pipeline is two-fold: first, the noise is substantially reduced and the true log-ratio signal is revealed, which is beneficial for downstream analysis. Second, segmentation affords significant dimension reduction, which helps in the analysis of multiple tumors. Specifically, while the probe-data live in a d-dimensional space, the segmented data can be represented in the p-dimensional space of the intervals of constant log-ratio. In the context of the model described by (3.1), the array $(\alpha_1, ..., \alpha_d)$ can be replaced by $(a_1, ..., a_p)$. The latter representation is more meaningful, because it reflects the genomic instability of the tumor investigated, and not the resolution of the microarray technology used.

Segmentation can also lead to information loss, for example if the segmentation procedure fails to detect low-amplitude changes, or intervals consisting of very few probes. In high-throughput arrays however, the probes cover the genome so densely that enough evidence is available for a very accurate segmentation.

3.2.2. Aberration call

Aberration calling is the task of assigning a discrete copy number state to each interval of constant copy number: loss, neutral or gain. Calling aberrations is frequently formulated as a problem of assessing statistical significance. For example, a segmented log-ratio value larger than zero, but small enough, can be still attributed to experimental bias (for example differences in dye incorporation efficiency) and therefore should be classified as neutral. In contrast, segments of log-ratio significantly larger or smaller than zero should be classified as loss or gain.

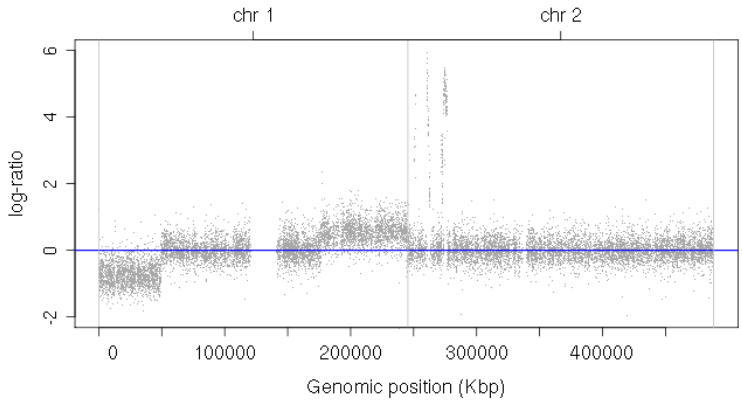
A large class of methods for aberration calling are based on choosing significance thresholds for loss and gain. In the context of the segmentation model given in (3.1), let *gain_ thr* and *loss_ thr* be such thresholds. Then:

$$\text{interval } I_k \text{ is classified as } \begin{cases} \text{gain}, & \text{if } a_k > gain\_thr \\ \text{neutral}, & \text{if } loss\_thr \leq a_k \leq gain\_thr \\ \text{loss}, & \text{if } a_k < loss\_thr \end{cases}, \quad k = 1, ..., P$$

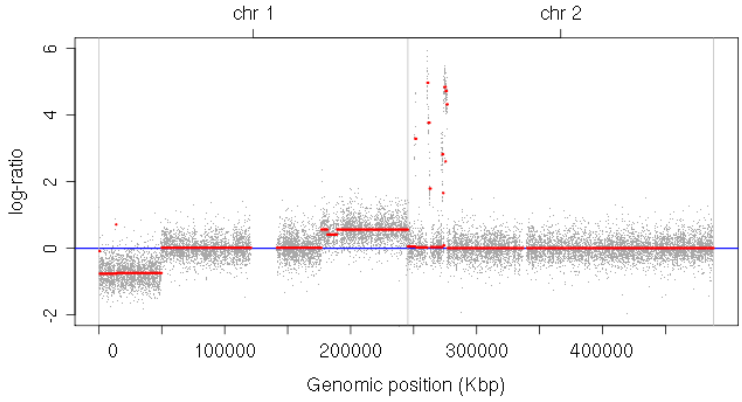
In figure 3.2c, we show an example of aberration thresholds.

The statistical significance is determined by taking into account the background distribution and the variability of the log-ratios in arrayCGH experiments. For example, some authors use normal-to-normal hybridizations [1] in order to isolate the variability introduced by the experiments, in the absence of copy number aberrations (Wang et al., 2005). Veltman et al. (2003) choose the *loss_ thr* and *gain_ thr* such that only a very small

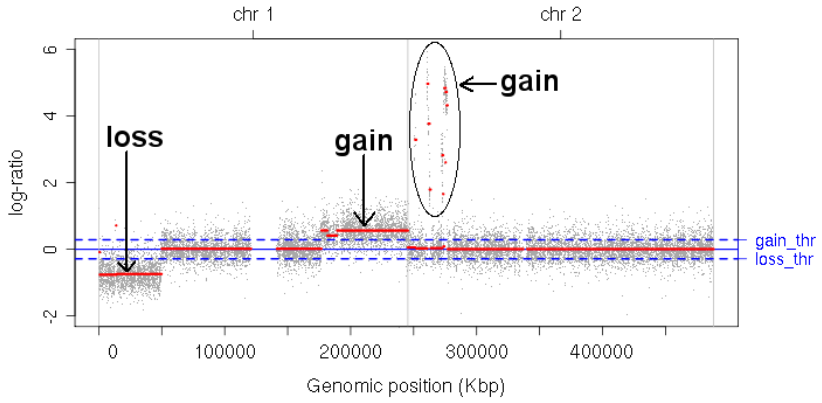
[1] Hybridization of normal tissue versus normal tissue.



(a)



(b)



(c)

Figure 3.2.: ArrayCGH experiment on a neuroblastoma tumor. a) Normalized log-ratios, ordered along the genome. Each grey dot corresponds to a probe on the microarray. The blue horizontal line corresponds to a log-ratio equal to zero and indicates a normal copy number. b) Segmented log-ratios: the red line represents the piecewise constant model fitted to the observed log-ratios. 3) Aberration call: the blue dotted lines mark the gain and loss thresholds.

percentage of the normal-to-normal log-ratios fall outside the interval $[loss\_thr, gain\_thr]$. Other authors estimate the standard deviation SD of normal-to-normal log-ratios and de-

fine the gain threshold *gain_thr* by $3SD$ and the loss threshold *loss_thr* by $-3SD$ (Nakao et al., 2004; Hodgson et al., 2001). However, using normal-to-normal hybridizations require supplementary expenses and identical experimental conditions as the normal-to-tumor hybridizations, which are difficult to ensure.

Hupé et al. (2004) use directly the tumor arrays for estimating the variation of the experimental noise. They propose to use the interquartile range[2] of the differences between successive log-ratios as a robust measure of variance. Then, the authors cluster the segmented log-ratios into three classes corresponding to loss, neutral and gain status.

In the case of most cancers, the three copy number states discussed above are insufficient for capturing the landscape of copy number alterations (Wu et al., 2009), which may range from zero to tens of copies. Therefore, in more recent articles, authors prefer to use multiple copy number levels. Guttman et al. (2007) split the gain class into two subclasses, in order to differentiate between low-amplitude aberrations and high-amplitude aberrations. Van de Wiel et al. (2007) propose the method called CGHcall, which clusters the segmented log-ratios into six classes, two for each type of signal: loss, neutral and gain. In a different publication (van de Wiel and van Wieringen, 2007), the same group of authors also suggest that the segmented log-ratios are grouped into four states: loss, neutral, gain and amplification, where amplification indicates a large copy number, of at least four copies. Willenbrock and Fridlyand (2005) propose the method MergeLevels, which iteratively joins intervals, the log-ratios of which cannot be distinguished based on a Wilcoxon rank sum test[3]. Aguirre et al. (2004) also propose four states (method cghMCR): gain (segmented log-ratio ≥ 4 standard deviations from the median), loss (segmented log-ratio ≤ 4 standard deviations from the median), amplification (segmented log-ratio $\geq 97\%$ quantile) and deletion (segmented log-ratio $\geq 3\%$ quantile).

Regardless of the granularity of the copy number classes, the main weakness of aberration calling is that it relies on hard thresholding. The intervals of log-ratio close to the threshold values have a high risk of being wrongly classified. These errors are propagated through the next steps of the pipeline. Moreover, aberrations of low amplitude are ignored, despite the fact that in many studies, recurrent low-amplitude aberrations across many tumors constitute evidence of positive selection and thus potential impact on phenotype.

3.2.3. Identification of recurrent CNAs in a set of tumors

The most difficult and the most intensely explored problem in the field of DNA copy number analysis is the identification of recurrent CNAs in a set of tumors. Its biologically motivated goal is to single out the driver aberrations, which play an important role in tumor development, from the passenger aberrations, which occur as a consequence of the high genomic instability, but have no impact on tumor progression. The high recurrence (or frequency) of a CNA is generally accepted as evidence of relevance, because it suggests that the respective CNA plays a favorable role in the survival and proliferation of the tumor cell, being consistently selected for during tumor progression. Additionally, recurrent CNAs are most likely to harbor disease-critical genes that can be targeted by therapies.

[2]The interquartile range is a measure of statistical dispersion and is defined as the difference between the third and first quartiles of a distribution.

[3]Wilcoxon rank sum test is a non-parametric statistical hypothesis test, which is used for assessing whether two populations have equally large values.

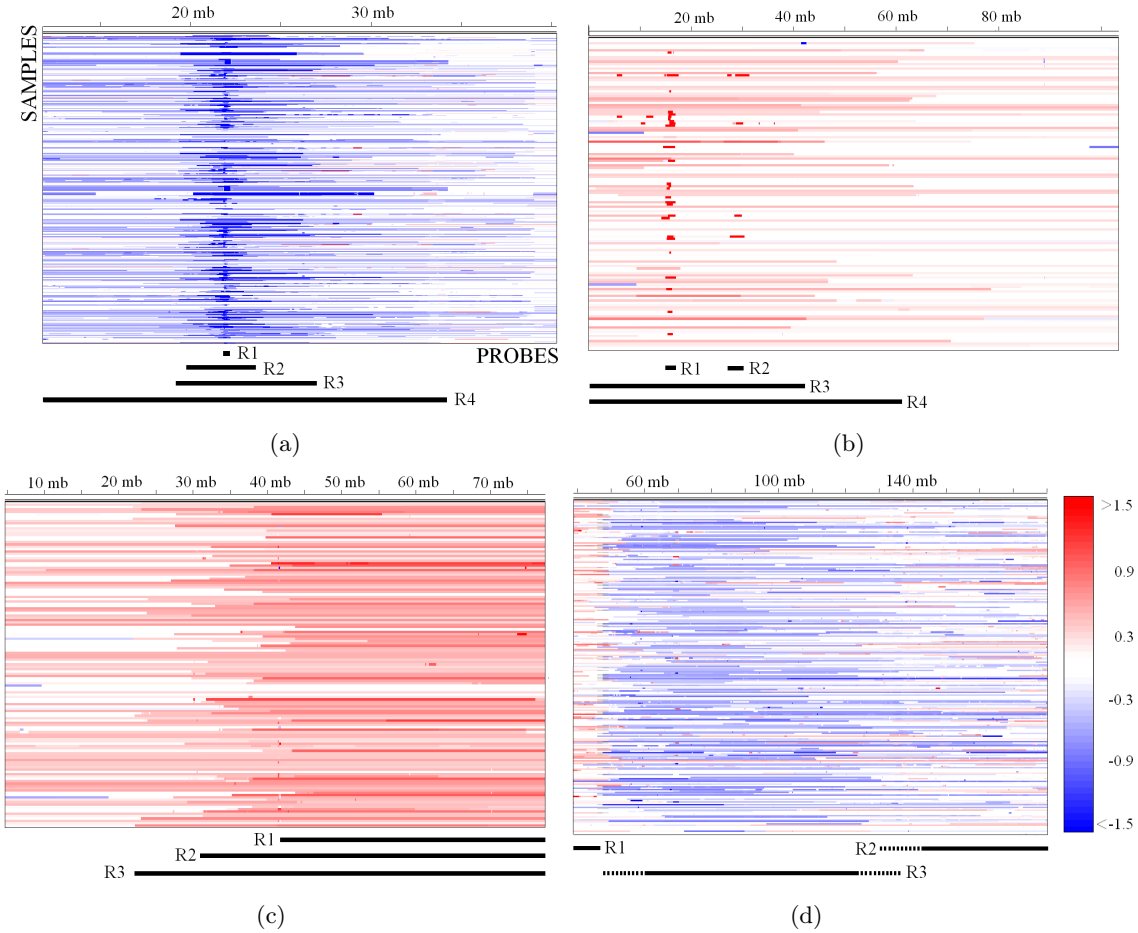


Figure 3.3.: Examples of recurrent CNAs in various types of cancers. The samples are arranged on the vertical axis, the probes are represented along the horizontal axis, ordered according to their genomic position. The color scale encodes the smooth log-ratio values: red stands for gain, blue stands for loss and white represents neutral copy number. With black segments, we mark genomic regions that, based on visual inspection and human judgment, are considered recurrent CNAs. a) Deletions of chromosome 9 segments in Glioblastoma, concentrated around the CDKN2A tumor suppressor gene. b) Amplifications of the MYCN and ALK gene loci in certain tumors and of a larger region in other tumors, in a Neuroblastoma cohort. c) Recurrent amplifications with various boundaries in chromosome 17 from Neuroblastoma. d)Deletion with highly variable boundaries on the right side, overlapping with gain regions, in ovarian cancer.

To this date, automated identification of recurrent CNAs remains very challenging. In a review paper, Rueda and Díaz-Uriarte (2010) observe that the difficulty of the problem stems mainly from the lack of agreement on what constitutes a recurrent CNA region. For an intuitive understanding of the difficulty, we illustrate in Figure 3.3 several types of recurrent CNAs that occur in tumors. In Figure 3.3a, a recurring deletion on chromosome 5 in a Glioblastoma cohort is shown, which can be easily recognized by visual inspection. The frequency of deletion is very high, occurring in more than 50% of the samples. The challenge in this case consists of precisely defining the genomic interval (or region) which

is most likely to contain the genetic factors involved in tumor progression. In Figure 3.3a, one can observe the high variability of the location and length of deletions among samples. Some samples exhibit a very short deletion at the locus marked in the figure by 'R1'. Other samples appear to have, additionally, a larger deletion, overlapping with '$R1$', which can be summarized, for example, by interval '$R2$'. Progressively larger deletions are suggested by '$R3$' and '$R4$'. From a biological perspective, it is not clear which region is most relevant and should be selected as a recurrent CNA. Many authors prefer '$R1$', which is the sub-region with maximal recurrence, meaning that it occurs in a maximal subset of samples. Such a maximally recurrent region is also the shortest of the stack of overlapping, recurrent CNAs, therefore it is often referred to as the 'minimal recurrent region' (Rouveirol et al., 2006; Rueda and Díaz-Uriarte, 2010). The minimal recurrent region is attractive because it narrows down maximally the search for disease-associated genes. Indeed, in the example discussed, $R1$ harbors the tumor suppressor gene CDKN2A. However, tumors with a larger deletion around the CDKN2A locus may have a different phenotype, or may have progressed in a different way, which can shed light on tumor evolution. Therefore, the larger regions must not be discarded.

Figure 3.3b shows gains and amplifications on chromosome 2 from a Neuroblastoma cohort. Regions '$R1$' and '$R2$' approximate the locations of two recurrent amplifications, containing the oncogenes MYCN and ALK, respectively. In this case, very little evidence (4 samples out of 150) supports region $R2$, however the boundaries of the amplifications are in high agreement, which makes it improbable that the aberrations co-localize by chance. Additionally, some samples that do not have amplifications at $R1$ and $R2$ locations harbor longer gains (of lower copy number), suggested by regions '$R3$' and '$R4$'. It is necessary that both the gain and amplification are detected, since they are likely to belong to two different tumor subtypes.

In Figure 3.3c, chromosome 17 from a Neuroblastoma cohort is shown. Disregarding the samples exhibiting whole chromosome gain, a shorter recurrent gain with variable boundaries on the left is apparent. Despite the variability, accumulations of breakpoints around several sites, for example as defined by regions '$R1$', '$R2$' and '$R3$', point to locations in which the DNA is more vulnerable and easy to break. Identifying all these sites, where breakpoints accumulate, can reveal alternative mechanisms of copy number alteration.

Finally, Figure 3.3d presents recurrent deletions and gains on chromosome 5 from a set of ovarian tumors. The deletions marked by region '$R3$' overlap with the gains marked by '$R2$'. Unlike in the previous scenarios discussed above, it is very hard to decide which region is likely to harbor disease-related factors using human judgment (hence the dotted line, marking uncertainty).

In the light of the examples above, it is obvious that giving a formal definition of a recurrent region is a very difficult task. In general, authors formulate their preferred criteria for selecting recurrent regions and propose algorithms that are guided by these criteria. In what follows, we summarize the most frequently cited works in the literature. These approaches, together with their input formats are schematically presented in Figure 3.1.

Aguirre et al. (2004) propose cghMCR, which is one of the first algorithms for identification of recurrent CNAs. The algorithm is applied to called data (i.e. some aberration call method has been applied to the arrays), with a separate handling of low magnitude

gains and losses, and of high amplitude amplifications and deletions. The goal of the algorithm is to identify 'Minimal Common Regions' (MCRs), which are defined by the authors as continuous segments that are altered in at least a fraction of the samples (recurrence threshold), the value of which is defined by the user. However, it is not clear which recurrence threshold results in best performance and no optimization scheme is suggested. The authors also make some ad-hoc decisions, such as joining altered segments that are closer than 500kbp, and MCRs that are separated by only one probe. Rueda and Díaz-Uriarte (2010) observe that the method is very sensitive to changes of these parameters.

Lipson et al. (2006) present the method IntervalScores, which works directly on normalized data. The authors formulate a scoring function that can be applied to any genomic interval I and any set of samples S, which consists of a normalized sum of all log-ratios within I and over all samples in S. Significantly large or small scores indicate recurrent CNAs. The significance of the score is obtained by comparison to a null score, which stems from a parametric model of the log-ratios under the assumption than no recurrent aberration is present. The authors provide an efficient algorithm for identifying the interval of maximum score, then re-iterate the procedure on the left and right side of the optimal interval. Despite the strong theoretical support for the algorithm, the method has limitations, in that the authors do not make qualitative statements with respect to the biological meaning of their scoring schemes.

The algorithm STAC, introduced by Diskin et al. (2006), is applied to called data. It makes use of two statistics: the 'frequency statistic', which is calculated as a simple count of aberrations at each probe, and the 'footprint statistic', which is a score that is large if the boundaries of recurrent aberrations are tightly aligned. The idea of the authors is that a recurrent CNA should be either highly frequent in the pool of samples, or well aligned at the boundaries (for example, region $R2$ from Figure 3.3b). In order to identify regions with significant frequency or footprint statistic, the authors construct a null model, which does not contain recurrent aberrations, by randomly permuting the altered segments of each array. For this reason, the algorithm is slow, in practice. The authors propose a faster approach, called MSA, in Guttman et al. (2007).

Rouveirol et al. (2006) make the first attempt towards a rigorous definition of recurrent CNAs. Based on called data, the authors define 'minimal alteration regions' (MAR). In a simplified view, a MAR is very similar in concept to a bicluster (Mirkin, 1996), consisting of a subset of consecutive probes all of which are altered over a subset of samples. Observing that the number of MARs occurring in a real data set is too large, the authors introduce CMAR (constrained minimal alteration regions), which comprise only those MARs which occur with a particular minimum frequency, have a minimum or maximum size and are aligned well enough. However, the formal framework is too complicated and examples are insufficient, which is probably the reason that the MAR and CMAR methods did not become too popular in the applied research.

One of the most widely used methods is GISTIC (Genomic Identification of Significant Targets in Cancer) (Beroukhim et al., 2007; Weir et al., 2007). The authors of GISTIC use segmented data to compute the 'G-score', a simple statistic that sums the log-ratios over all samples. The G-score captures both the frequency and the amplitude of aberration, in such a way that either few high-magnitude alterations or many low-magnitude alterations can result in a high score. In order to identify regions with significantly high score, the authors

of GISTIC use a permutation scheme for computing null G-scores, then use multiple testing correction for controlling the False Discovery Rate (FDR) (Benjamini and Hochberg, 1995) and report a q-value for significance.

The CoCoA method (Ben-Dor et al., 2007) applies to segmented data and is based on a statistic that, for a specific CNA region and a specific tumor sample, quantifies how likely it is that an amplitude as high is attained by an arbitrary region of the same size. For multiple samples, the probability of observing at least a minimum number of CNAs at a specific location is estimated based on an adjusted binomial distribution.

Shah et al. (2007) introduce the H-HMM (Hierarchical HMM) method, which extends the single-array method for segmentation and aberration calling based on Hidden Markov Models (Fridlyand et al., 2004). A 'master process' is accountable for switching between driver alterations and passenger alterations in the cohort of tumors. An HMM is used for inferring the master process, which uses the normalized data of all samples simultaneously. A direct limitation of the method is that infrequent recurrent aberrations cannot be detected (Shah, 2008). Also, a drawback of this method is that it reports recurrent probes, and not recurrent regions.

RAE (Taylor et al., 2008) is a very interesting approach, because it combines unprecedently many biological assumptions regarding the sources of variations affecting CNAs and statistical methodology. The method is applied to segmented data. Motivated by the difference in biological mechanisms, RAE treats low-level and high-level aberrations differently and handles deletions and amplifications separately. The scoring function used by RAE is very similar to that used by GISTIC (Beroukhim et al., 2007), namely the average over the segmented log-ratios of all samples. The significance is determined in reference to a null case, obtained by an elaborate permutation scheme, which incorporates knowledge on recombination hotspots for a more realistic randomization. Incorporating complex biological assumptions into the method can be truly advantageous, however the assumptions may not hold always, which can result in erroneous results.

Klijn et al. (2008) present the method KC-SMART, which is applied directly to normalized data. At each probe location, a weighted average over all samples of the log-ratios in the neighborhood gives an informative score. The weights are given by a flat-top Gaussian kernel centered at the current probe, which affords a smooth definition of the neighborhood and ensures that probes in the immediate vicinity bring more information to the score than the remote probes. Reference null scores are obtained via a permutation scheme. By varying the width of the kernel, the method is able to detect both large and small CNAs. The main limitation of the method is that it is not able to detect recurrent CNAs that affect only a small fraction of the samples.

DiNAMIC (Walter et al., 2011) is among the most recent methods for identifying recurrent CNAs. It can be applied to any kind of input (normalized, segmented, called). DiNAMIC uses the same scoring function as GISTIC and RAE, namely the sum of log-ratios at each probe, over all samples. The novelty is brought by the permutation scheme that is used for assessing significance: the authors use circular permutations on each sample independently, in order to preserve the spatial relationship between the log-ratios. The drawback of the method is that it cannot detect low-frequency CNAs.

To this date, there exists no gold-standard algorithm for detection of recurrent CNAs, because no comparative study has been published. Two review articles (Shah, 2008; Rueda

and Díaz-Uriarte, 2010) comment on the approaches, however, only qualitatively, stating strengths and possible limitations. We think that a quantitative comparison is difficult, for at least three reasons: first, the lack of benchmark datasets with annotated recurrent CNAs, second, the lack of agreement to what constitutes a recurrent CNA (different approaches have different goals) and third, the code unavailability of many of the methods.

3.2.4. Quantifying the association of a recurrent CNA with the phenotype

The ultimate goal of studies on copy number aberrations is to bring to light mechanisms of tumor progression that can be directly addressed by drug therapies. To this end, two main directions have been undertaken in the literature: *genetic marker discovery* and *tumor subtype discovery* (Kallioniemi, 2008).

In the context of DNA copy number analysis, genetic marker discovery refers to the task of identifying cancer-associated genes, that can be either targeted by drugs, or that can help predict patient survival, or response to a particular therapy, or can suggest the progression status of the tumor. In other words, given a phenotypic label such as survival, treatment response or tumor progression, genomic loci are identified, the copy number of which significantly associate with the respective phenotype. More recently, a secondary topic has emerged, closely related to genetic marker discovery, namely *genetic pattern discovery*. It extends the problem of identification of single-gene markers to sets of probes (or genes) that, combined in a meaningful way, can help predict tumor phenotype. In general, genetic marker discovery consists of a univariate selection of informative probes, whereas pattern discovery requires multivariate data analysis. From a computational perspective, genetic marker and pattern discovery can be formulated as a *supervised feature selection* problem.

Tumor subtype discovery refers to the problem of discovering cancer subtypes based on specific copy number aberration profiles. This task is unsupervised and it generally implies the usage of some clustering technique.

Further on in this thesis, we will present in great detail the problems of supervised feature selection and unsupervised classification, because our contributed work adopts such techniques (see Chapters 4 and 5). In this section, we will only briefly mention the most highly acknowledged publications dedicated to the task of genetic marker discovery and tumor subtype discovery.

The most widely used tool for marker discovery is statistical hypothesis testing. The typical setting assumes that copy number measurements at a number of loci are given. The measurements can be in various stages of pre-processing: raw (normalized), segmented or called. A phenotype variable is also given, usually binary (0/1), such that a value of 1 stands for a progressed tumor, or a therapy-resistant tumor, etc. For each locus, a statistical test is carried out, which evaluates the difference between the copy number estimate of the tumors labeled with 0 and the ones labeled with 1. If the copy number is given as a continuous value (for example, normalized or segmented data), a t-test (Student, 1908) is usually carried out, for finding whether the difference between the means of the two populations is significant or not (Kresse et al., 2010; Fridlyand et al., 2006; Spitz et al., 2006). If the method is applied to called data, a Fisher's exact test (Fisher, 1922) can be applied to the contingency table between the phenotype and the called copy number data (Joosse et al., 2009; Tagawa et al., 2005). Applying significance tests at many genomic loci simul-

taneously increases the chance of incorrectly rejecting the null hypothesis (type I error). A common statistical practice which keeps the type I error low is multiple testing correction, which can be carried out by means of various methods (Benjamini and Hochberg, 1995; Holm, 1979; Hochberg, 1988). However, most methods for multiple testing correction are based on the assumption that the tests are independent, which does not hold if applied to copy number alteration data. The strong correlations between neighboring loci make the traditional correction methods too conservative. van de Wiel and van Wieringen (2007) notice this problem and propose a dedicated method, however only applicable to called data.

Statistical models that characterize the association between the copy number changes at a certain locus and patient survival are a very popular tool for marker discovery (Kresse et al., 2010; Idbaih et al., 2008; Carrasco et al., 2006; Tagawa et al., 2005). A typical application is the following: the tumors are divided into two groups, according to the presence or absence of a particular CNA. The distributions of patient survival data (Kaplan and Meier, 1958) in these two subgroups are compared by a log-rank test (Mantel, 1966). For the extended purpose of pattern discovery, multivariate regression to survival data can be performed by Cox proportional hazard models (Cox, 1972). For example, Kresse et al. (2010) and Idbaih et al. (2008) report on combinations between CNAs and clinicopathological indicators that are predictive of survival in malignant fibrous histiocystomas and gliomas, respectively.

Pattern discovery is commonly approached by statistical learning methods for classification and feature selection. Specifically, the copy number measurements (normalized, segmented or called) are used as features for predicting a (binary) phenotypical outcome. To this date, only a few publications have proposed classifiers that are tailored to the particularities of copy number data. Liu et al. (2008) describe an SVN classifier (Vapnik, 1998) with a specialized kernel called Raw. The kernel is based on the count of common aberrations between two tumor samples. The drawback of the Raw kernel is that the classification model is not directly interpretable, namely one cannot assess the contribution of particular features to the prediction. The authors propose MIFS (maximum influence feature selection), which is an iterative method for selecting most useful features in a greedy fashion. Here, the main limitation of the feature selection approach is that it cannot guarantee that at least a local minimum has been attained. Rapaport et al. (2008) introduce the fusedSVM, a linear SVN with a Lasso and a fused penalty. Conceptually, the fusedSVN maximizes the separation margin between the outcome classes, with L_1 constraints on the number of features that participate in the model (Lasso penalty) and the difference between the weights of subsequent features (fused penalty). The strength of this approach is that it can be applied directly to normalized data and it can automatically discover regions of (almost) constant copy number that are predictive of phenotype. However, in this thesis we show that the feature selection with fused SVM is biased towards short regions (Toloşi and Lengauer, 2011). Barutcuoglu et al. (2009) construct a special Hidden Conditional Random Field (HCRF) model (Lafferty et al., 2001), in which the observed log-ratios, the underlying copy numbers and the phenotype are the vertices of a network. The topology of the HCRF is chosen to best capture the structure of copy number data. A gradient Lasso (Kim and Kim, 2004) algorithm is used for estimating optimal parameters of the HCRF. Other authors borrow methods tailored to expression analysis and apply them directly to

copy number data: Joosse et al. (2009) use Shrunken Centroids (Tibshirani et al., 2003) for discriminating between two subtypes of breast cancer. Bergamaschi et al. (2006) use Significance Analysis for Microarrays (SAM) (Tusher et al., 2001) to identify significant associations between CNAs and clinico-pathological parameters in breast cancer.

Tumor subtype discovery is usually approached by unsupervised clustering techniques. Specifically, a partitioning of the tumor samples into several groups is performed, based on similarity of their copy number profiles. Fridlyand et al. (2006) and Kresse et al. (2010) use agglomerative hierarchical clustering (Hastie et al., 2003) with Pearson correlation distance between samples and Ward distance (Ward, 1963) between clusters, and conclude that the breast cancer cohort investigated exhibits three main subtypes. van Wieringen et al. (2007) perform hierarchical clustering with a new similarity measure and a new linkage method, that take into account the structure of copy number data. Cheung et al. (2009) use Partitioning Around Medoids (PAM) (Kaufman and Rousseeuw, 1990), with Hamming distance (Hamming, 1950) between a case and a medoid. The authors choose five clusters and argue that they represent distinct subtypes of follicular lymphoma. Shah et al. (2009) introduce a new clustering technique based on Hidden Markov Models (HMM) and estimate the optimal number of clusters via Silhouette values (Rousseeuw, 1987). Carrasco et al. (2006) use Non-negative Matrix Factorization (NMF) (Brunet et al., 2004), a technique for deriving a small number of feature representatives that summarize the original data well, in order to partition a set of multiple myeloma patients into two and four clusters. André et al. (2009) also use NMF, but in addition, the optimal number of clusters was estimated by calculating the cophenetic correlation coefficient of the cluster assignment (Sneath and Sokal, 1962).

3.3. A new pipeline based on supervised selection of CNAs relevant for tumor phenotype

Our main criticism of the traditional pipeline for analysis of CNAs is directed towards the steps: *aberration call* and *identification of recurrent CNAs in a set of tumors*. The aberration call step uses hard thresholds to classify probes into loss, neutral or gain. The probes with log-ratio close to the threshold values have a high risk of being wrongly classified. These errors are propagated and their effect is potentially amplified in the subsequent steps of the pipeline. Considering that aberration call is essential for data interpretation, we believe that it should be one of the last steps of the pipeline. Similarly, the step for identification of recurrent CNAs in a set of tumors involves making implicit or explicit assumptions on their structure, on the minimal frequency of a recurrent CNA, etc. Such assumptions disfavor CNAs characteristic to rare subtypes, or disregard larger aberrations in favor of the minimal recurrent region (as we showed in Figure 3.3a). This type of unsupervised selection of regions, based solely on human judgment, may run the risk of information loss.

In contrast, we believe that array segmentation is beneficial for downstream analysis and should be part of any efficient pipeline, because it eliminates a type of experimental noise that is relatively well understood and thoroughly investigated (for example, via normal-to-normal hybridization). No less importantly, segmentation affords significant dimension

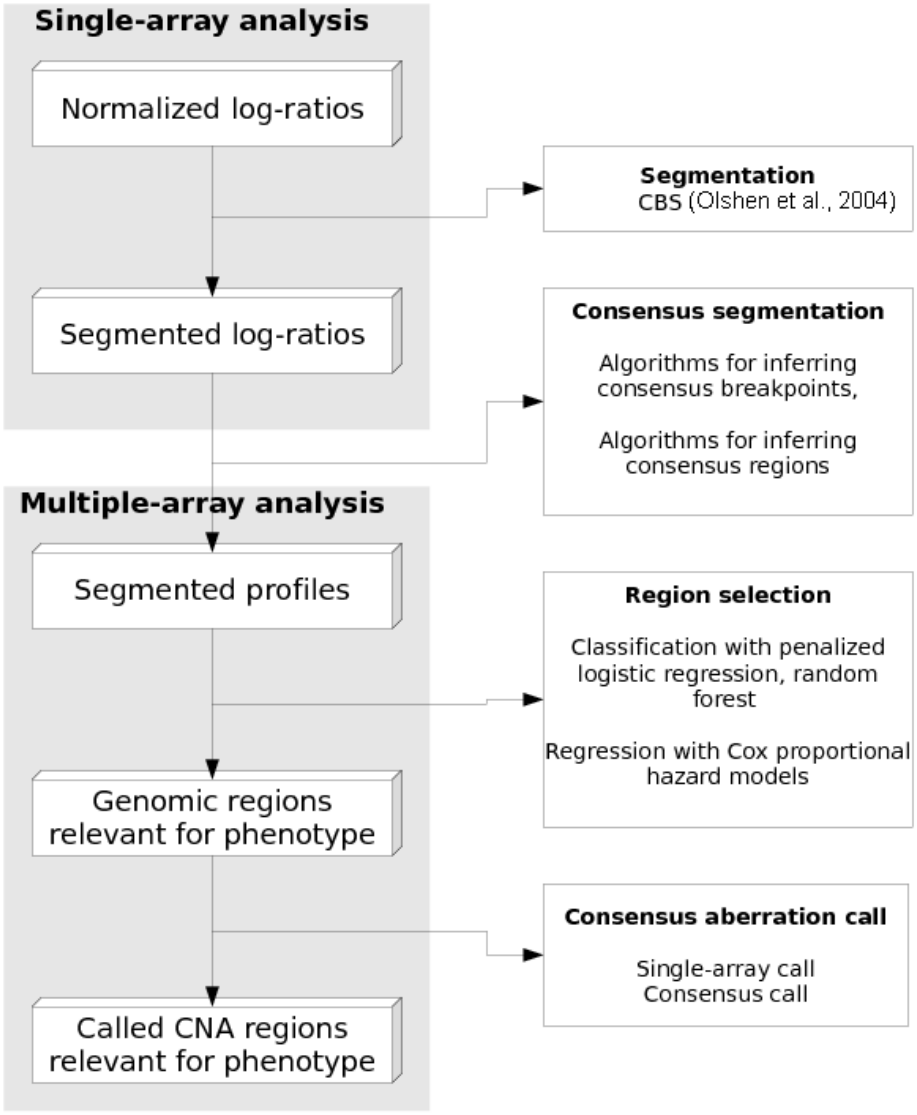


Figure 3.4.: A modified pipeline for inferring CNAs associated with tumor phenotype.

reduction, in the following way: the probes that yield identical smoothed log-ratio in all arrays can be joined into a single interval. Here, we call this procedure *compression*. By compression, the dimension of the data is no longer determined by the resolution of the experimental technology, but by the number of breakpoints characteristic of the tumor investigated.

Following these considerations, we propose a new pipeline for genetic marker and genetic pattern discovery, which performs data driven selection of the most informative genomic regions with respect to a specific phenotype. Figure 3.4 summarizes the steps of the pipeline: 1) *segmentation*, 2) *consensus segmentation*, 3) *region selection* and 4) *aberration call*. The pipeline applies directly to normalized log-ratios. The purpose of the segmentation and

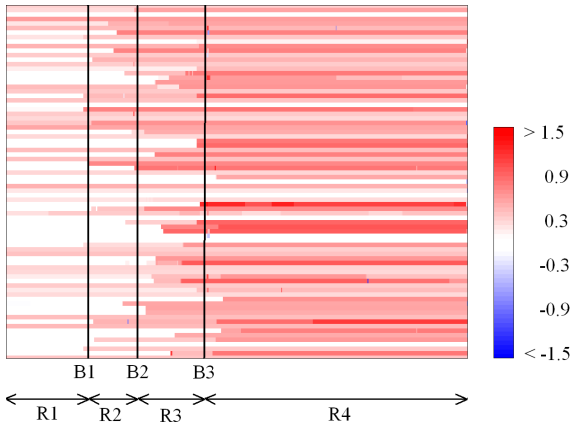


Figure 3.5.: An example of consensus segmentation. Chromosome 17 from a neuroblastoma cohort is partitioned into the following regions of almost constant profile: $R1, R2, R3$ and $R4$.

consensus segmentation steps is to achieve dimension reduction with minimal information loss. We suggest that the segmentation step is carried out by CBS, which is recommended by its top performance (as discussed earlier in this chapter).

The consensus segmentation step is a multivariate generalization of the single-array segmentation. Given a set of arrays, the goal is to partition the genome into a set of regions $R_1, ..., R_m, \quad m \geq 1$ of *almost constant profile*. By region of almost constant profile, we mean a genomic interval R contained in one chromosome, with the property that the true copy number of each array stays constant within interval R. Any observed copy number change within interval R is therefore attributed to noise. For the purpose of dimension reduction, we require that the number of almost constant regions is small.

For example, we reexamine the stack of amplifications with varying boundaries from Figure 3.3c. A possible outcome of the consensus segmentation procedure is given in Figure 3.5. In this case, a possible consensus segmentation consists of four distinct *consensus regions* $R1, R2, R3$ and $R4$, delimited by $B1, B2$ and $B3$, which we call *consensus breakpoints*. The advantage of this procedure over the frequently used minimal common region is that it allows all regions to become marker candidates. Hence, it is possible to evaluate the relevance of larger amplifications, additionally to that of the minimal common region.

Technically, in the consensus segmentation step we use the segmented log-ratios from a set of arrays for estimating an optimal partition into consensus regions. The problem is very challenging, due to the complex structure of the noise, which can be attributed to either misalignment of the breakpoints between regions or passenger alterations, which appear at random in the set of arrays. In Chapter 4, we propose several algorithms for consensus segmentation, which are adapted to the special structure of copy number data. To our knowledge, we are the first to address this problem. The only related method in this respect is CGHregions (van de Wiel and van Wieringen, 2007), which performs consensus segmentation, however on called data. Parenthetically, the authors of CGHregions have introduced the term 'almost constant region'.

The region selection step of the pipeline uses supervised classification and regression models for feature selection, in order to select the most predictive genomic regions for

a particular tumor phenotype. We choose interpretable models such as sparse logistic regression and random forest, in order to be able to quantify the contribution of each feature to the prediction. The main difficulties arise from the high dimensionality of the data and the large correlations between features. In Chapter 5 we present the problems and our solutions in great detail.

The region selection step provides with a list of genomic regions, the copy number of which is informative of the tumor phenotype. For the sake of interpretation, aberration call is necessary at this final stage of the pipeline. We distinguish two subtasks: the *single-array call* and *consensus call*. Given a genomic region of interest R, the single-array call refers to assigning a state of this region – neutral, gain or loss – for each array independently. The consensus call consists of assigning a state – neutral, gain, loss or mixed – to the region itself. The mixed state corresponds to regions that are lost in a subset of arrays and gained in another subset.

The novelty of our pipeline and the main contribution of this thesis is the consensus segmentation step. Throughout this manuscript, we will highlight its benefits for downstream analysis.

4. Methods for Consensus Segmentation

> Truth is ever to be found in the simplicity, and not in the multiplicity and the confusion of things.
>
> Sir Isaac Newton

4.1. Introduction

Array Comparative Genomic Hybridization has been established as a cheap technology for high-resolution measurement of DNA copy number aberrations in large cohorts of tumors. The very high resolution of the experiments ensures an accurate estimation of the location of breakpoints and of copy number changes across the genome, however the dimensionality of the data is so large that basic statistical tools become ineffective. For example, the search for predictive markers meets all the challenges related to clustering in high-dimensional spaces (the curse of dimensionality). One must handle the high correlations between predictors or correct for multiple testing without assuming independence of the tests. Fortunately, the structure of copy number changes facilitates substantial dimension reduction. Specifically, adjacent genomic loci are likely to share the same copy number, unless a breakpoint occurs in between. Single-array segmentation is one of the algorithms that exploits this property for array denoising and dimension reduction (see Chapter 3). Thus, by applying segmentation to every array in a given set, the copy number data can be represented in a reduced space of dimension equal to the number of breakpoints identified in the collection.

In this chapter, we introduce the methodology for a further step for dimension reduction, called *consensus segmentation*, which is a generalization of the single-array segmentation to a set of arrays. Specifically, we segment the genome further into regions of almost constant copy number over the set of arrays. This way, consensus segmentation can be used to delineate regions of consistent alteration from regions with none or very few passenger alterations. Thus, the search for biomarkers can be narrowed down to smaller DNA intervals and the downstream statistical analysis is simplified.

Like the single-array segmentation, consensus segmentation can be approached in two equivalent ways: either by determining regions of almost constant copy number, or by identifying transition locations between regions. These are genomic locations characterized by an enrichment in breakpoints observed in the set of arrays. Such genomic locations are biologically interesting, probably related to cancer phenotype, as they are positively selected for during tumor development.

In this context, we mention the work of van de Wiel and van Wieringen (2007), who introduce method called CGHregions, which uses called data in order to infer regions of

almost constant copy number. The authors search for genomic regions with the property that the L_1 distance between any two loci is smaller than a given constant c, thus ensuring that there are not many breakpoints within a region. Our main criticism of this approach is that it uses called data, which applies hard thresholding for classifying the log-ratios into loss, neutral or gain. In general, called data is a very rough approximation of the raw data, often leading to loss of critical information.

A more recent related approach by Ritz et al. (2011) is dedicated to the identification of recurrent breakpoints in sets of arrayCGH experiments. The method is called Neighborhood Breakpoint Conservation (NBC) and it is applied directly to the raw log-ratios. The authors devise a special single-array segmentation algorithm, which assigns to each pair of adjacent probes a probability of a breakpoint being located between them. Then, locations of recurrent breakpoints are analytically estimated and returned as a list, sorted by significance (p-value). The main shortcoming of the NBC algorithm is its high complexity, the segmentation step using a dynamic program quadratic in the number of array probes, which probably means that NBC is not an attractive solution for high-resolution arrayCGH data or for copy number data from NGS experiments. In comparison, the fastest current approach for segmentation is almost linear (CBS by Venkatraman and Olshen (2007)).

In this Chapter, we will introduce several methods for consensus segmentation, based on identification of either recurrent breakpoints or of regions of almost constant copy number. These methods use segmented data as input. We assume the segmentation to be carried out using the most efficient and scalable approach available (e.g., CBS). We also introduce a measure of evaluating the quality of consensus segmentation on real cancer data and use this measure for comparing the algorithms. Importantly, we present and discuss interesting genetic and epigenetic properties of the recurrent breakpoints and regions identified by our methods.

This Chapter is organized as follows: in the Preliminaries section, we introduce the notations, terminology and model assumptions, which are necessary for presenting the methodological contributions. Sections 4.3, 4.4 and 4.5 present algorithms for consensus segmentation. The Data section introduces the cancer datasets on which we validated the algorithms. The Results section summarizes the performance of the methods and offers insights into the biological relevance of our findings. We conclude the chapter with a Discussion.

4.2. Preliminaries

Assume given a set of N arrays (tumors), consisting of segmented log-ratio measurements at p genomic loci $L_1, ...L_p$. For simplicity, we assume that all loci are located on the same chromosome and ordered. In practice, the methods will be applied to each chromosome independently. We represent the data by a matrix $A \in \mathcal{R}^{N \times p}$, where row a_i corresponds to array i and column a^j contains all segmented log-ratios at genomic locus j. By a_{ij}, we denote the element of A located on row i and column j. Throughout this chapter, the same style of notations for rows, columns and elements of matrices will apply.

Definition We call **region** a set of consecutive loci $\{L_i, L_{i+1}..., L_{i+j}\}$, for some $1 \leq i \leq i + j \leq p$.

Let A be a data matrix as above. We make here a key model assumption, namely that the data have been generated by a stochastic model of the form:

$$a_{ij} = \sum_{k=1}^{m} x_{ik} \mathbb{1}_{R_k}(L_j) + \varepsilon_{ij}, \quad 1 \leq i \leq N,\ 1 \leq j \leq p, \tag{4.1}$$

where $R_1, ..., R_m$ are m non-overlapping regions which cover the whole chromosome, x_{ij} is an element of a matrix $X \in \mathcal{R}^{N \times m}$ and ε_{ij} represents stochastic noise. The number of regions m, the regions $R_1, ..., R_m$ and the constants X are parameters of this model. Intuitively, (4.1) describes a piecewise constant model with noise, which is assumed to be the generator of the observed arrays. In contrast to the piecewise constant model from single-array segmentation (see (3.1)), the model from (4.1) enforces the same partition of the genome to *all* arrays. Also unlike the single-array segmentation, the noise components $\{\varepsilon_{ij}\}_{\substack{1 \leq i \leq N \\ 1 \leq j \leq p}}$ do not follow a particular parametric distribution and it do not have to be i.i.d.

The problem of **consensus segmentation** consists of estimating the parameters of model (4.1) from the observed data matrix A, by optimizing some quality criterion.

Equation (4.1) can be written in an equivalent matrix form as follows:

$$A = X\mathcal{I} + \Xi, \tag{4.2}$$

where $\mathcal{I} \in \{0,1\}^{m \times p}$ is an indicator matrix with entries $\mathcal{I}_{kj} = \mathbb{1}_{R_k}(L_j)$ and $\Xi \in \mathcal{R}^{N \times p}$ is a noise matrix with components ε_{ij}. We define matrix $\hat{A} = X\mathcal{I}$. It follows that $\hat{A}$ is an approximation of the initial data matrix A, determined by the segmentation into m regions $R_1, ..., R_m$. The matrix X is a representation of matrix A in an m dimensional space, with $m \leq p$. Each column x^j of X summarizes the copy number data of a certain region and will be called the **centroid** or **representative** of the region.

Definition Matrix $X \in \mathcal{R}^{N \times m}$ is called the **reduced representation** of the copy number data after consensus segmentation.

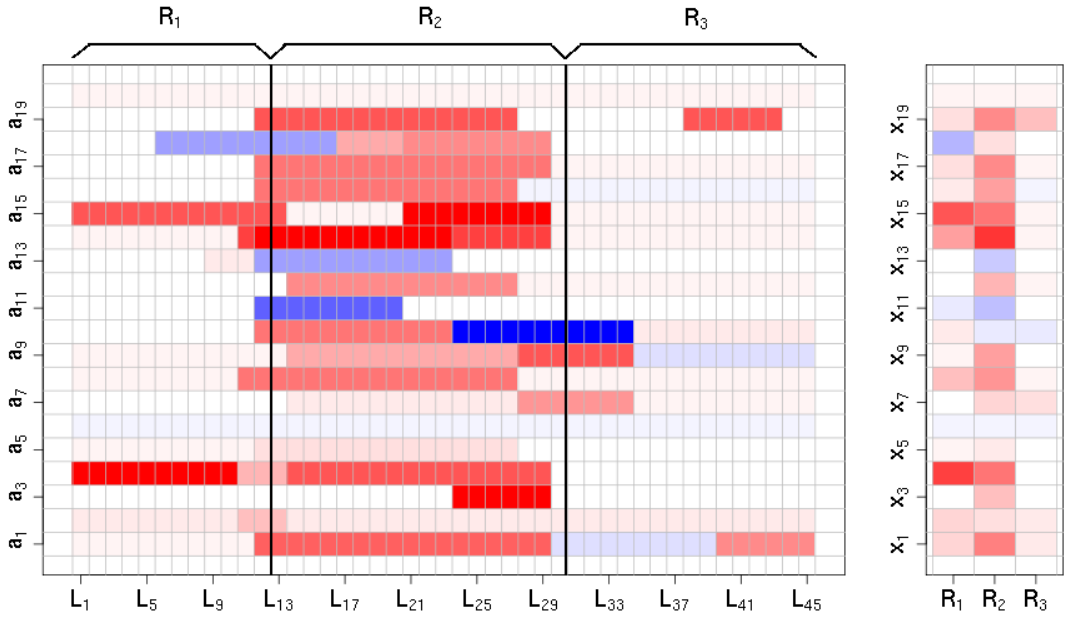
Definition Regions $R_1, ..., R_m$ given by model (4.1) are called **consensus regions**. The consensus regions are assumed to be ordered according to their genomic location.

Definition Let R_k be the k^{th} consensus region, such that $R_k = \{L_i, ..., L_{i+j}\}$, for some $0 \leq i \leq i+j \leq p$. Then L_i is the left boundary of region R_k. We call this genomic location the $\boldsymbol{k^{\text{th}}}$ **consensus breakpoint**, and we denote it by $\boldsymbol{B_k}$.

Observe the duality of the notions of consensus breakpoints and consensus regions: any of the two uniquely determines the other.

Example In Figure 4.1a, we show $N = 20$ arrays consisting of segmented log-ratios at $d = 45$ genomic loci. Increased copy number is consistently observed in the region $R_2 = \{L_{13}, ..., L_{30}\}$, while the other two regions $R_1 = \{L_1, ..., L_{12}\}$ and $R_3 = \{L_{31}, ..., L_{45}\}$ contain very few changes. Therefore, R_1, R_2, R_3 are good candidates for consensus segmentation. The corresponding consensus breakpoints are $B_1 = L_1,\ B_2 = L_{12}$ and $B_3 = L_{30}$.

For the estimation of the parameters of model (4.1), some optimality criterion must be satisfied. For minimal loss of information, the approximation $\hat{A}$ must be as close to



(a) (b)

Figure 4.1.: Dimension reduction via consensus segmentation. a) Example of 15 segmented arrays (y-axis) measured at 45 genomic loci (x-axis). Intense red marks regions of gain, white corresponds to neutral copy number and blue indicates loss. b) Regions $R_1 = \{L_1, ..., L_{12}\}$, $R_2 = \{L_{13}, ..., L_{30}\}$ and $R_3 = \{L_{31}, ..., L_{45}\}$ are consensus regions and B_1, B_2, B_3 are consensus breakpoints. c) Reduced representation after consensus segmentation. The copy number within each region is summarized by its mean value over the region.

A as possible, therefore the expected distance between A and $\hat{A}$, $\mathrm{E}[d(A, \hat{A})]$, must be minimized, for some chosen distance function d. Intuitively, we require that the set of centroids represented by the columns of X approximate the data matrix A well. Clearly, if each locus is considered to be its own region (thus $m = p$), then a perfect fitting is possible, with $d(A, \hat{A}) = 0$. For achieving some dimension and noise reduction however, the number of intervals must be kept small and a trade-off between the quality of approximation and the magnitude of m should be considered.

For simplification, the problem can be broken into two subtasks: first, the number of regions m and the regions $R_1, ..., R_m$ are to be estimated. The second task assumes the regions given and computes the centroids X that provide a good approximation $\hat{A}$ of A. The first problem is notoriously difficult and we approach it in two ways, by *estimating consensus regions* or by *estimating consensus breakpoints*. We call the second task *estimating the reduced representation*. In the rest of this chapter, we introduce algorithms for each of these problems. We start with the simpler task: estimating the reduced representation.

4.3. Methods for estimating the reduced representation

Assume that regions $R_1, ..., R_m$ are given. We search for a set of region representatives $X^{\mathrm{opt}} = (x^1, ..., x^m)$ which approximates best the data matrix A. Let d be a loss function

which defines the distance between two matrices. Then X^{opt} is given by:

$$X^{\text{opt}} = \arg\min_{X} \mathrm{E}[d(A, X\mathcal{I})] \tag{4.3}$$

The target $\mathrm{E}[d(A, X\mathcal{I})]$ depends on the distribution of the stochastic noise Ξ (see (4.2)). Any attempt at solving (4.3) analytically would require parametric assumptions on noise distribution. We believe that such endeavor should be avoided, because there are complex spatial dependencies in the noise which cannot be expressed by any model in a realistic manner. For example, around the genomic locations of consensus breakpoints more variation is expected, since there are transitions between different copy numbers. Another source of variation are the passenger aberrations, which may occur at any location of the genome, probably with higher likelihood around positions where the DNA is more fragile and prone to breakage (Agarwal et al., 2006). False changes, which correspond to false-positive breakpoints reported by the single-array segmentation, can also be considered noise.

In this chapter we do not attempt to characterize the distribution of the noise component. Instead, we propose several heuristic approaches to constructing representatives $X = (x^1, ..., x^m)$, which may not be optimal in terms of accuracy but afford interpretability. We discuss each of them below, as well as their implications from biological perspective.

Let R_k be a fixed region, $R_k = \{L_{i+1}, ..., L_{i+j}\}$. The representative x^k summarizes the set of columns $a^{i+1}, ..., a^{i+j}$.

Summary by mean

$$x^k = (x_{1k}, ..., x_{Nk}), \quad x_{qk} = \text{mean}\{a_{q(i+1)}, ..., a_{q(i+j)}\},\ q = 1, ..., N.$$

The sample mean is the most common choice for summarizing of a set of observations. It has the property that it minimizes the expected L_2 distance to the observations under normality assumptions (maximum likelihood estimator). This means that if d is chosen to be the L_2 distance and the noise Ξ follows a Gaussian distribution, then the mean is the appropriate representative. However, the mean is not appropriate if the log-ratios within region R_k contain outliers, or follow a multimodal distribution. For example, if most log-ratios within region R_k are positive, indicating gain, but there are several negative log-ratios (attributed to noise), the mean value can be close to zero. As a consequence, this region would appear neutral in the reduced representation and important information would be lost.

Summary by median

$$x^k = (x_{1k}, ..., x_{Nk}), \quad x_{qk} = \text{median}\{a_{q(i+1)}, ..., a_{q(i+j)}\},\ q = 1, ..., N.$$

The sample median is a robust summary of the log-ratios within a region. It has the property of minimizing the expected L_1 distance to the observations, under the assumption that the noise Ξ follows a Laplace distribution (maximum likelihood estimator) (Koenker and Bassett, 1978). The median is not influenced by outliers, therefore the negative effects described above would not occur. However, robustness to outliers is only useful if the outliers do not carry important information and need to be discarded. In our particular application, the following may happen: assume that most log-ratios of region R_k are neutral

(close to zero), except for a few larger values, indicating a focal gain. In this case, the median would be close to zero, ignoring the gain. For downstream analysis, a slightly elevated summary value (such as the mean would give) would perhaps be more informative, indicating that the region does contain a small alteration.

Summary by principal component

A popular technique for summarizing multivariate data is principal component analysis (PCA). It consists of exploring the variability of the set of multivariate observations by projecting them onto the first principal components. These are obtained by computing the eigenvalues and eigenvectors of the covariance matrix of the set. In our application, let $v^1, ..., v^j$ be the principal components of the set of observations $a^{i+1}, ..., a^{i+j}$, ordered decreasingly by the magnitude of their corresponding eigenvalues. Then, vector $v^1 \in \mathcal{R}^N$ is the direction which best explains the variability of set of observations. However, principal components have two shortcomings. First, they are known to be influenced by outliers. Methods for robust principal component analysis have been proposed by De la Torre and Black (2001) and Jackson and Chen (2004) and should be preferred. Second, incremental update of the representative x^k as new tumors are added to the input set is difficult. The first principal component of the augmented region R_k needs to be recomputed from scratch. In contrast, the mean and the median can be easily updated by adding the mean or the median of the new observations to the representative.

Summary by medoid

The medoid is defined as the column of region R_k which minimizes the sum of the distances to all other columns of region R_k. The distance is user-defined and does not have to come from a metric space, a dissimilarity matrix suffices. Using the medoid as a summary may improve the interpretability of the reduced representation, because each representative would correspond to the copy number measurements at a particular location on the genome. In contrast, the mean, median, extreme value or the principal component are aggregates. Choosing the medoid as representative can also be dangerous, because it essentially means that one particular genomic position (think gene) from a region with low variability is preferred to the other positions in the region. There is a good chance that the driver gene is discarded by this procedure. Even worse, attention is focused on an 'imposter' – a neighboring gene that is very similar.

We recommend that the choice of the representative is application-specific. For reasons including computational efficiency and interpretability, we prefer the mean and the median. The advantages of one over the other have been the subject of long disputes in the statistical community. We recommend an insightful article by Koenker and Bassett (1978), which comments on the implicit parametric assumptions associated with the mean and median statistics. In our applications we tried both methods. Because the results were not substantially different, we only present the results using the mean.

4.4. Algorithms for estimating consensus breakpoints

Following the definition given in the Preliminaries of this chapter, consensus breakpoints are the genomic locations that mark the start of new consensus regions. Consequently, consensus breakpoints are positions around which a large number of breakpoints occur in the set of arrays. Observe the stochastic nature of a consensus breakpoint: it is generally not exactly determined by a perfect alignment of breakpoints, but it is a summary of a set of breakpoints accumulating around a certain position in a way that is unlikely by chance. In Figure 4.1, breakpoints B_2 and B_3 at locations L_{12} and L_{30} respectively, are consensus breakpoints because breakpoints occur frequently around these locations. Neighboring locations like L_{11} and L_{13} or L_{29} and L_{31} are good alternatives for consensus breakpoints B_2 and B_3, respectively. In contrast, location L_6 is probably not a consensus breakpoint, since evidence shows that only one array contains a breakpoint at this location.

We propose two approaches towards finding consensus breakpoints. First, we present a parametric approach called Consensus Breakpoints via Mixture of Uniform and Gaussians (CB-MUG) and second, a non-parametric approach called Consensus Breakpoints via Kernel Smoothing (CB-KeS).

4.4.1. The CB-MUG algorithm

The parametric approach is based on the following model assumption. Let V be the random variable representing the genomic location of a breakpoint in an arbitrary array. If $B_1, ..., B_m$ are all consensus breakpoints, then we model the distribution of V as a mixture of m Gaussians centered at $B_1, ..., B_m$, with different standard deviations. Also, we include in the mixture a uniform distribution $U(c_{start}, c_{end})$, where c_{start} and c_{end} are the start and end positions of the chromosome. That is to say, a breakpoint is either generated by one of the Gaussians, in which case it contributes to the corresponding consensus breakpoint, or it is generated by the uniform distribution, in which case it is a noisy breakpoint.

Formally, the density of V is given by:

$$g_V(v) = \sum_{k=1}^{m} \pi_k \phi(v; B_k, \sigma_k) + \pi_0 u(v; c_{start}, c_{end}), \tag{4.4}$$

where $\phi(\cdot; B_k, \sigma_k)$ is a Gaussian distribution of mean B_k and standard deviation σ_k, $u(\cdot; c_{start}, c_{end})$ is a uniform distribution over the interval $[c_{start}, c_{end}]$ and $\{\pi_k\}_{k=1}^{m}, \pi_0$ are the mixture probabilities, $\sum_{k=0}^{m} \pi_k = 1$. The means, standard deviations of the Gaussians as well as the mixture probabilities are not known. The number m of Gaussians in the mixture is considered a fixed parameter of the model, however in practice it is also not known and automated methods are needed for computing its value in a data-driven way.

For now, assume m is given and let V be the multiset[1] of breakpoint locations observed in the set of arrays. For simplicity of notation, we used V to denote both the random variable and its observations:

[1] In mathematics, a multiset is a generalization of a set, such that each member of a multiset can be present in multiple instances.

$$V = \bigcup_{i=1}^{N} V_i, \quad \text{where} \quad V_i \text{ is the set of breakpoints of array } a_i.$$

Let $|V| = T$ and the components of V be denoted by $v_1, ..., v_T$.

The purpose of the CB-MUG algorithm is to estimate the parameters of the mixture model (4.4) from the observations V. The algorithm uses the classical Expectation-Maximization (EM) technique.

First introduced with this name in an article by Dempster et al. (1977), the EM algorithm is used to estimate maximum likelihood parameters of statistical models from observed data. Additionally, the models typically depend on unobserved latent variables. The EM procedure iterates between an expectation step (E) and a maximization step (M). In the expectation step the missing data are estimated, given the observed data and the current estimates of the model parameters. In the maximization step, the parameters of the model are updated such that the likelihood function is maximized, by using the estimates of missing data from step E. The likelihood of the model is guaranteed to increase with each iteration and the algorithm converges to a local optimum.

The EM procedure is a very general tool for addressing difficult tasks of maximum likelihood estimation. Here, we describe the particular EM algorithm that can be used to estimated the parameters of model (4.4).

Given an observation $V \sim g_V$, it is not known which of the $m+1$ mixture components has generated V. For this purpose, a *latent* random variable Δ is introduced, with values $\Delta \in \{0, 1, ..., m\}$ at probabilities $\{\pi_0, \pi_1, ..., \pi_m\}$. The role of the latent variable is to indicate which mixture component is responsible for observation V. Specifically, if the value of Δ is $k > 0$, the Gaussian distribution $\phi(\cdot; B_k, \sigma_k)$ is responsible for observation V. If $k = 0$ then V has been generated by the uniform distribution $u(\cdot; c_{start}, c_{end})$. If $X_k \sim \phi(\cdot; B_k, \sigma_k)$ and $X_0 \sim u(\cdot; c_{start}, c_{end})$ are random variables, then V can be written as:

$$V = \delta(\Delta, 0)X_0 + \delta(\Delta, 1)X_1 + ... + \delta(\Delta, m)X_m, \tag{4.5}$$

where δ is Kroneker's delta, with $\delta(i, j) = 1$ if $i = j$ and 0 otherwise. Direct observations on Δ are not available, however estimating the parameters of the mixture model becomes substantially easier if inference on the latent variable is performed.

The parameters of the model are $\theta = (\pi_0; \pi_1, ..., \pi_m; B_1, ..., B_m; \sigma_1, ..., \sigma_m)$. The log-likelihood of θ based on the observations $V = \{v_1, ..., v_T\}$ is :

$$L(\theta; V) = \sum_{i=1}^{T} \log \left[\sum_{k=1}^{m} \pi_k \phi(v_i; B_k, \sigma_k) + \pi_0 u(v_i; c_{start}, c_{end}) \right] \tag{4.6}$$

Estimating θ that maximizes the log-likelihood is difficult because of the sum inside the logarithm. The problem becomes simpler if we assume that the values of the latent variable Δ are known for each of the observations $v_1, ..., v_T$. Indeed, let $\Delta_1, ..., \Delta_T$ be those values. The log-likelihood can be re-written as:

$$L(\theta; V, \Delta) = \sum_{i=1}^{T} \left[\sum_{k=1}^{m} \delta(\Delta_i, k) \log \phi(v_i; B_k, \sigma_k) + \delta(\Delta_i, 0) \log u(v_i; c_{start}, c_{end}) \right] + \sum_{i=1}^{T} \left[\sum_{k=1}^{m} \delta(\Delta_i, k) \log \pi_k + \delta(\Delta_i, 0) \log \pi_0 \right]. \tag{4.7}$$

However, the values $\{\Delta_i\}_{i=1,...,T}$ are not known. In order to evaluate the log-likelihood given by equation (4.7), the values $\delta(\Delta_i, k)$ are needed, for each $i = 1, ..., T$ and $k = 0, ..., m$. In other words, for each observation v_i, the distribution of the mixture which has generated it needs to be found. These quantities cannot be estimated directly, but their expected values can, if the parameters θ are assumed to be known. In order to avoid a cyclic argument – remember that the goal is to estimate θ – let θ^{old} be some already available estimate for θ, with $\theta^{\text{old}} = (\pi_0^{\text{old}}; \pi_1^{\text{old}}, ..., \pi_m^{\text{old}}; B_1^{\text{old}}, ..., B_m^{\text{old}}; \sigma_1^{\text{old}}, ..., \sigma_m^{\text{old}})$. Then the following holds:

$$\gamma_i^k \overset{\text{def}}{=} \mathrm{E}(\delta(\Delta_i, k) | v_i, \theta^{\text{old}}) = \Pr(\Delta_i = k | v_i, \theta^{\text{old}}) \overset{\text{Bayes}}{=} \frac{\overset{\text{prior}}{\Pr(\Delta_i = k)} \Pr(v_i | \Delta_i = k, \theta^{\text{old}})}{\Pr(v_i | \theta^{\text{old}})} = \tag{4.8}$$

$$= \begin{cases} \dfrac{\pi_k^{\text{old}} \phi(v_i; B_k^{\text{old}}, \sigma_k^{\text{old}})}{\sum_{j=1}^{m} \pi_j^{\text{old}} \phi(v_i; B_j^{\text{old}}, \sigma_j^{\text{old}}) + \pi_0^{\text{old}} u(v_i; c_{start}, c_{end})}, & \text{if } k > 1, \\ \dfrac{\pi_0 u(v_i; c_{start}, c_{end})}{\sum_{j=1}^{m} \pi_j^{\text{old}} \phi(v_i; B_j^{\text{old}}, \sigma_j^{\text{old}}) + \pi_0^{\text{old}} u(v_i; c_{start}, c_{end})}, & \text{if } k = 0. \end{cases}$$

In the derivations (4.8), we first used the fact that $\delta(\Delta_i, k)$ is a Bernoulli variable, which takes value 1 with probability π_k and value 0 with probability $1 - \pi_k$. The expectation of a Bernoulli distribution is given by its probability of success. In the subsequent step, we used Bayes' theorem in order to calculate the posterior probability of the distribution to be the k^{th} component of the mixture, given v_i and θ^{old}. The value γ_i^k is called the *responsibility* of component k for observation v_i. Observe that for each observation v_i, the following property holds:

$$\sum_{k=0}^{m} \gamma_i^k = \sum_{k=0}^{m} \Pr(\Delta_i = k | v_i, \theta^{\text{old}}) = 1. \tag{4.9}$$

Given the responsibilities, new parameters θ^{new} can be computed to maximize the conditional expectation of the log-likelihood $\mathrm{E}_{\Delta | V, \theta^{\text{old}}} L(\theta; V, \Delta)$ from equation (4.7):

$$\theta^{\text{new}} = (\pi_0^{\text{new}}; \pi_1^{\text{new}}, ..., \pi_m^{\text{new}}; B_1^{\text{new}}, ..., B_m^{\text{new}}; \sigma_1^{\text{new}}, ..., \sigma_m^{\text{new}}) = \arg\max_{\theta} \mathrm{E}_{\Delta | V, \theta^{\text{old}}} L(\theta; V, \Delta). \tag{4.10}$$

$$\mathrm{E}_{\Delta|V,\theta^{\mathrm{old}}} L(\theta; V, \Delta) = \sum_{i=1}^{T} \left[\sum_{k=1}^{m} \gamma_i^k \left[-\log(\sqrt{2\pi}\sigma_k) - \frac{(v_i - B_k)^2}{2\sigma_k^2} \right] + \gamma_i^0 \log \frac{1}{c_{end} - c_{start}} \right] \tag{4.11}$$

$$+ \sum_{i=1}^{T} \left[\sum_{k=0}^{m} \gamma_i^k \log \pi_k \right]$$

The parameters θ^{new} are computed in a straightforward way, by differentiating the function $\mathrm{E}_{\Delta|V,\theta^{\mathrm{old}}} L(\theta; V, \Delta)$ with respect to all B_k, σ_k and π_k variables and making use of the relation (4.9). For the Gaussians in the mixture, we obtain:

$$B_k^{\mathrm{new}} = \frac{\sum_{i=1}^{T} \gamma_i^k v_i}{\sum_{i=1}^{T} \gamma_i^k},\ k = 1, ..., m;$$

$$\sigma_k^{\mathrm{new}} = \sqrt{\frac{\sum_{i=1}^{T} \gamma_i^k (v_i - B_k^{\mathrm{new}})^2}{\sum_{i=1}^{T} \gamma_i^k}},\ k = 1, ..., m.$$

The mixture probabilities are given by:

$$\pi_k^{\mathrm{new}} = \frac{\sum_{i=1}^{T} \gamma_i^k}{T},\ k = 0, 1, ..., m.$$

In Equation (4.10), the Expectation and Maximization steps are apparent. The algorithm (see Algorithm 1) iterates between the Expectation step, in which the parameters θ of the mixture are assumed known and the responsibilities are computed (4.12) and the Maximization step, in which model parameters are estimated (4.13) so as to maximize the conditional expected log-likelihood.

Convergence of the algorithm

By convergence of the EM algorithm we mean that the likelihood of the model approaches arbitrarily close some finite value after a finite number of iterations. The theoretical aspects involved in the convergence of the EM algorithm are not trivial. We do not reproduce here these arguments, but refer the interested reader to the work of Mclachlan and Krishnan (1996), who include a comprehensive discussion on this topic. A key result in presented by Wu (1983), who show that if the space of model parameters and the likelihood function respect certain regularity conditions, then the EM algorithm is guaranteed to converge to a stationary point of the log-likelihood (which can typically be either local maximum or a point of inflection). Intuitively, under these regularity conditions, it can be proven that both the Expectation and Maximization steps improve the likelihood (meaning, they do not decrease its value), by updating alternatively the latent parameters and the model parameters in the direction of the stationary point.

It is important for this thesis to mention that the regularity conditions by Wu (1983) do not necessarily hold for our particular instance of the EM algorithm. Specifically, the space of model parameters is required to be compact, which does not hold because the space of possible values that the standard deviations of a Gaussian can take is not compact. Indeed,

Algorithm 1 CB-MUG(m)

Require: $T \in \mathbb{N}$, $V = \{v_1, ..., v_T\}$, $c_{start} < c_{end}$

Ensure: Estimated parameters of the mixture density: $\{B_k\}_{k=1}^m$, $\{\sigma_k\}_{k=1}^m$, $\{\pi_k\}_{k=1}^m$, π_0.

1. Take initial guesses for the parameters $\{B_k^{\text{old}}\}_{k=1}^m$, $\{\sigma_k^{\text{old}}\}_{k=1}^m$, $\{\pi_k^{\text{old}}\}_{k=1}^m$, π_0^{old} (see text)
2. *Expectation step:* compute the responsibilities

$$\gamma_i^k = \frac{\pi_k^{\text{old}} \phi(v_i; B_k^{\text{old}}, \sigma_k^{\text{old}})}{\sum_{j=1}^m \pi_j^{\text{old}} \phi(v_i; B_j^{\text{old}}, \sigma_j^{\text{old}}) + \pi_0^{\text{old}} \dfrac{1}{c_{start} - c_{end}}}, \quad i = 1, ..., T, \;\; k = 1, ..., m$$

$$\gamma_i^0 = \frac{\pi_0^{\text{old}} \dfrac{1}{c_{start} - c_{end}}}{\sum_{j=1}^m \pi_j^{\text{old}} \phi(v_i; B_j^{\text{old}}, \sigma_j^{\text{old}}) + \pi_0^{\text{old}} \dfrac{1}{c_{start} - c_{end}}}, \quad i = 1, ..., T \tag{4.12}$$

3. *Maximization step*: compute new weighted means and standard deviances

$$B_k^{\text{new}} = \frac{\sum_{i=1}^T \gamma_i^k v_i}{\sum_{i=1}^T \gamma_i^k}, \quad \sigma_k^{\text{new}} = \sqrt{\frac{\sum_{i=1}^T \gamma_i^k (v_i - B_k^{\text{new}})^2}{\sum_{i=1}^T \gamma_i^k}}, \quad k = 1, ..., m \tag{4.13}$$

$$\pi_k^{\text{new}} = \frac{\sum_{i=1}^T \gamma_i^k}{T}, \quad k = 0, 1, ..., m$$

4. Set $B_k^{\text{old}} = B_k^{\text{new}}$, $\sigma_k^{\text{old}} = \sigma_k^{\text{new}}$, for all $k \geq 1$ and $\pi_k^{\text{old}} = \pi_k^{\text{new}}$, for all $k \geq 0$. Iterate steps 2 and 3 until convergence.
5. Output $B_k = B_k^{\text{new}}$, $\sigma_k = \sigma_k^{\text{new}}$, for all $k \geq 1$ and $\pi_k = \pi_k^{\text{new}}$, for all $k \geq 0$.

the standard deviation σ of a Gaussian $\mathcal{N}(\mu, \sigma)$ takes values in the interval $(0, +\infty)$, which is not compact. It is easy to see that a finite upper limit to σ can be assumed, since the set of training observations is finite and hence has finite support. However, the real problem arises from the constraint $\sigma > 0$. Specifically, assume that at some iteration of the EM algorithm, the mean of a Gaussian, let it be B_k, is equal to one of the observations $B_k = v_j$. Then, as the corresponding standard deviation σ_k tends to *zero*, the likelihood tends to infinity. The EM algorithm would indefinitely improve the likelihood by decreasing the standard deviation σ_k and thus never converge. In practice, this problem can be solved by imposing limits on the standard deviations, such as for all $k \geq 1$, $\sigma_k \geq \epsilon$, for some small positive ϵ, as suggested by Mclachlan and Krishnan (1996). In our implementation, we apply this method and as soon as some $\sigma_k < \epsilon$, at some iteration of the algorithm, we fix σ_k to ϵ.

Another aspect regarding convergence is when to stop the algorithm, because the iterative process can no longer improve the log-likelihood significantly. Usually, a very small constant ξ is chosen and when the increase in log-likelihood is smaller than ξ, the algorithm stops. In cases in which the set of observations is very large (T large), the EM algorithm can be very slow and thus a superior limit on the number of iterations is also necessary.

Initialization of the parameters

The quality of the solution of the EM algorithm depends strongly on the starting values of the parameters (see step 1 of Algorithm 1). For example, if the starting point is close to the global optimum of the log-likelihood, then the algorithm will probably find a solution that is close to optimal. If in contrast, the starting point is close to the boundary of the parameter space, for example if some initial standard deviation of a Gaussian is close to zero, then convergence may not be guaranteed.

A simple approach to initializing parameters for our particular algorithm is to randomly generate m means $B_1, ..., B_m$ in the interval spanned by the training observations $[\min\{v_i | i = 1, ..., T\}, \max\{v_i | i = 1, ..., T\}]$ and then set all standard deviations to the overall sample standard deviance (Hastie et al., 2003). The mixing probabilities can be set all equal. However, we believe that the centers of the Gaussians (consensus breakpoints) are not uniformly distributed over the genome, but tend to accumulate in certain regions.

A data-driven approach is to use k-means clustering for grouping the observations into m clusters. The initial means and standard deviations of the Gaussians are given by the means and standard deviations of the observations in each cluster. The mixing probabilities of the Gaussians are set to be proportional to the cluster sizes and the mixing probability of the uniform component is set to a fixed constant. Using k-means for the initial parameter assignment is supported by the close relation between the EM and the k-means clustering in this case. The k-means procedure is similar to the EM approach, but uses hard assignments of observations to clusters, instead of the soft responsibilities.

In our applications, we use the latter approach. We assign a mixing probability of 5% to the uniform component and divide the rest of 95% among the Gaussians, proportionally to the size of the m-means clusters.

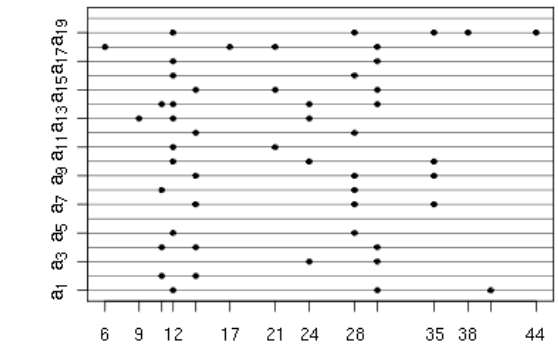
Example In Figure 4.2, we show the result of applying CB-MUG algorithm with $m = 2$ to the artificial data presented in Figure 4.1. The top figure shows the breakpoints occurring in the set of arrays. Below, the univariate histogram of the breakpoints appears bimodal, indicating two locations where breakpoints accumulate. The two Gaussians in the mixture capture the bimodality and the uniform distribution is used to generate breakpoints that are not likely to have been generated by the Gaussians. The bottom plot illustrates the responsibilities of each mixture component for the observed breakpoints. The output of the CB-MUG method in this case consists of two consensus breakpoints at locations $B_1 = L_{12}$ and $B_2 = L_{28}$.

Estimation of the optimal number of Gaussians in the mixture

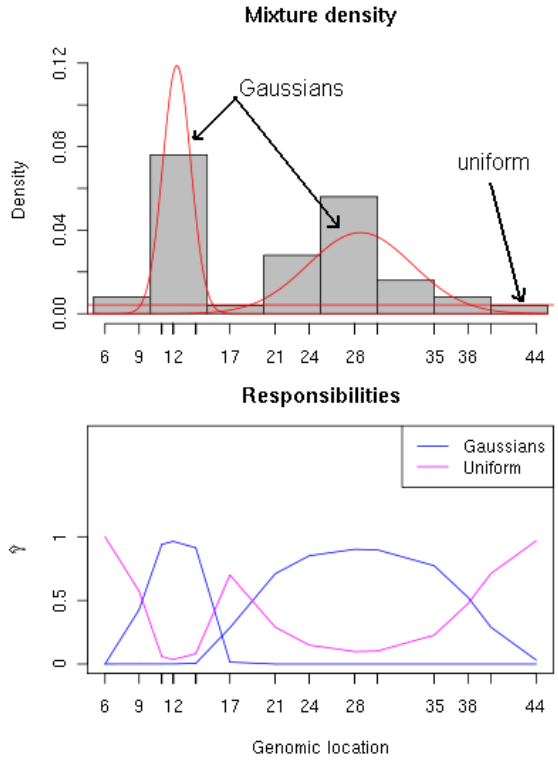
Estimating the optimal number of consensus breakpoints m in a chromosome is a problem of model selection which is discussed in Section 4.6.

4.4.2. Algorithm CB-KeS

The second approach to identifying consensus breakpoints is non-parametric. It uses a kernel smoothing technique for identifying locations around which unexpectedly large accumulations of breakpoints occur. By sliding a location pointer along the genomic sequence, we observe the breakpoints located within the vicinity of the current location and



(a)



(b)

Figure 4.2.: Example illustrating the CB-MUG algorithm. a) The breakpoints in the set of 20 arrays are marked with black dots. b) The outcome of the CB-MUG algorithm with two Gaussians and one uniform distribution. In red, the densities from the mixture model are shown, proportionally scaled by the corresponding mixture probabilities. In the bottom panel, the responsibilities are shown.

estimate the z-score of the observation under a null model. The null model assumes that the locations of the breakpoints of each array are uniformly distributed along the genomic sequence and do not depend on the arrays. The locations at which the estimated z-score is large enough are reported as candidates for consensus breakpoints. In what follows, we give a detailed description of this procedure.

4.4.3. Summarizing breakpoints

Let V be the set of breakpoints observed in the N arrays, represented as triples as follows:

$$V = \{(v_i, s_i, w_i) \mid 1 \leq i \leq T\}, \text{where:} \tag{4.14}$$

v_i is the genomic location of the i^{th} breakpoint ,

s_i is the index of the sample on which the i^{th} breakpoint was observed, $s_i \in \{1, ..., N\}$,

w_i is the magnitude of copy number change at breakpoint v_i and it is called its weight.

The weight component w_i of a breakpoint is computed as the difference between the log-ratio at position v_i and the log-ratio at position $v_i - 1$. It can be positive or negative, depending on whether the copy number decreases (negative weight) or increases (positive weight). The start position of each chromosome is considered a natural breakpoint for each sample and it is a member of V. Because at the start of the chromosome there is no copy number change, the weight of this breakpoint is considered zero.

4.4.4. Scoring genomic locations

Let x be a location on the genome and $\mathcal{K}(\cdot;\, \mu, \sigma)$ be a Gaussian kernel with mean μ and standard deviation $\sigma > 0$. We define the score function Γ as follows:

$$\Gamma(x; \sigma) = \sum_{i=1}^{T} |w_i| \mathcal{K}(v_i;\, x, \sigma) \tag{4.15}$$

The scoring functions Γ quantifies the abundance of breakpoints around location x. The location kernel ensures that the breakpoints in the immediate vicinity of x contribute more to the score than the distant breakpoints. The size of the neighborhood is controlled by the standard deviation of the kernel, σ. The scoring function Γ admits weights. The absolute value of the w_i component of the breakpoint (v_i, s_i, w_i) is used as weight, in order to increase the contribution of very large changes in log-ratio and to reduce the contribution of small changes. Consequently, a high $\Gamma(x, \sigma)$ score is attained either by the contribution of many breakpoints of low-weight around location x, or by few breakpoints of high weight. Additionally, the use of weights helps reduce the influence of small copy-number changes which can be false breakpoints, corresponding to errors of the single-array segmentation procedure (false positives).

We use Γ to score all genomic positions (in practice, only locations $\{v_i \mid 1 \leq i \leq T\}$ are scored) and find local maxima which are significantly large. Such locations are returned as consensus breakpoints. Their significance is assessed by a permutation test which will be described later on in this section.

Example Figure 4.3 illustrates the Γ scoring function on the example data from Figure 4.1, computed for $\sigma = 2$. The score (in blue) yields several local maxima around genomic positions 12, 24 and 29, which point to consensus breakpoints.

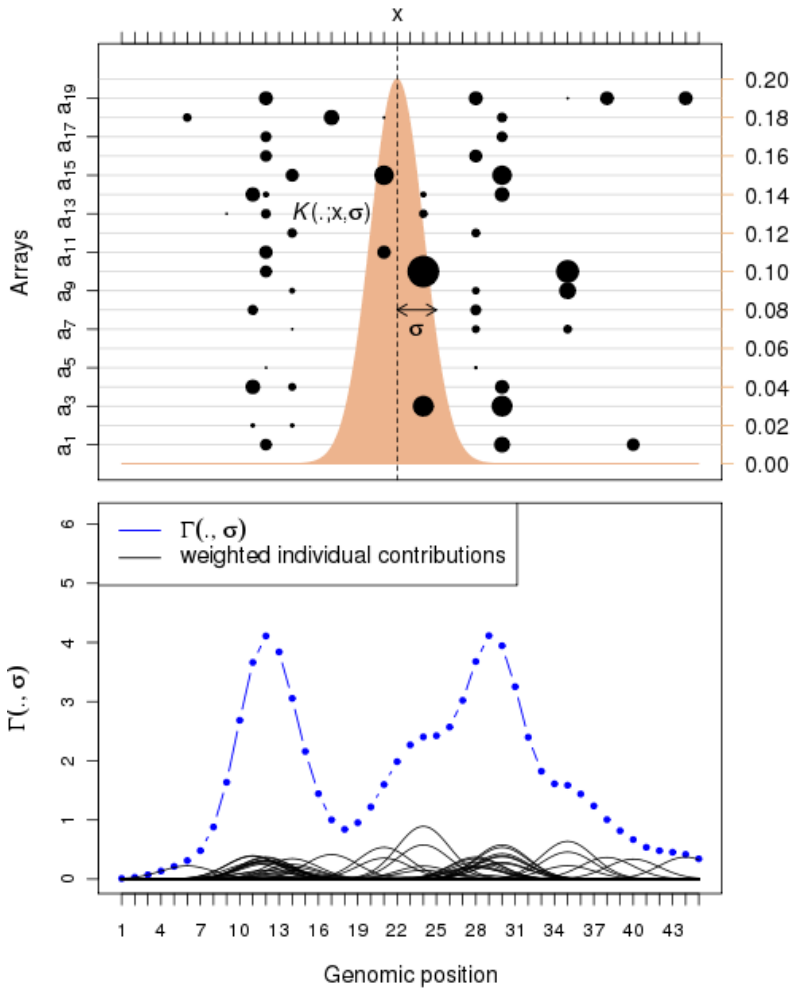


Figure 4.3.: Example illustrating the Γ scoring function. The breakpoints are marked with black dots, the Gaussian kernel of standard deviation σ is shown at arbitrary location x, with brown color. Function Γ is depicted with blue curve. The weights of the breakpoints are marked by the size of the points in the top subfigure.

An obvious problem of the Γ scoring functions is that, in the regions close to either end of the chromosome, fewer breakpoints have non-negligible contribution to the score, due to data censoring on the left (towards the start of the chromosome) and on the right (towards the end of the chromosome), respectively. This effect is well known in the field of signal smoothing with kernel functions. We do not correct the scoring function explicitly, but we consider the effect in the last step of the algorithm, when deciding the significance of the score under a null model.

The scoring function Γ has one parameter, namely the standard deviation σ of the location kernel. We will call this parameter the *kernel width*. The choice of kernel width specifies how 'tightly' the breakpoints should align in order to be considered as aggregated evidence of a consensus breakpoint. In what follows, we will call *diameter of the consensus breakpoint* the standard deviation of the individual breakpoints that form the consensus breakpoint. Based on the observation that in real data breakpoints aggregate in varying

degrees of tightness, we apply the scoring function multiply, with different kernel width values. Specifically, in our experiments kernel widths take values form the set:

$$SD = \{10^3, 10^4, ..., 10^8\}.$$

4.4.5. The null model

Let the kernel width σ be fixed. Clearly, not all local maxima of the scoring function $\Gamma(\cdot;\sigma)$ are candidates for consensus breakpoints. For example, in Figure 4.3 there exists a local maximum at location 43, but the score is too small and probably not significant. In order to infer statistical significance, we compare the observed score to a *null reference*, which is obtained by randomly re-arranging the breakpoints, such that the dependencies between the locations of breakpoints over the set of arrays arrays are destroyed. This way, any accumulation of breakpoints around a certain genomic location can be only due to chance.

The random re-arrangement is carried out as follows: for each array independently, its breakpoints are re-located to random genomic positions which are generated by an uniform distribution over the entire chromosome. Their weights are preserved. After all arrays have been processed, a *null instance* of the consensus breakpoint detection problem is generated, of the same size as the initial problem. Let $V^0 = \{(v_i^0, s_i, w_i) \mid 1 \leq i \leq T\}$ be the null instance. The $\Gamma(\cdot;\sigma)$ score is computed at the null locations $\{v_i^0 \mid 1 \leq i \leq T\}$. This procedure is repeated P times (in our experiments, P=50), enough for obtaining a large population of null scores covering the chromosome. After convenient sorting w.r.t. genomic position and re-indexing, let $v_1^0, ..., v_{PT}^0$ be the genomic locations at which the null scores $\gamma_1^0, ..., \gamma_{PT}^0$ are estimated.

The significance of the observed score at location x for kernel width σ is given by a z-score as follows:

$$z\text{-score}(x,\sigma) = \frac{\Gamma(x;\sigma) - \text{mean}\{\gamma_{i+1}^0, ..., \gamma_{i+j}^0 \mid v_{i+1}^0, ..., v_{i+j}^0 \text{ are } j \text{ nearest neighbors of } x\}}{\text{sd}\{\gamma_{i+1}^0, ..., \gamma_{i+j}^0 | v_{i+1}^0, ..., v_{i+j}^0 \text{ are } j \text{ nearest neighbors of } x\}}$$

The z-score indicates how large is the observed Γ score at a certain location, measured in standard deviations from the mean of the null scores. The local maxima of the scoring function Γ with z-score smaller than some positive threshold $\zeta > 0$ are considered not significant. Therefore, for each fixed σ, the CB-KeS algorithm returns a list of local maxima with positive z-scores, sorted in decreasing order by z-score.

Example In Figure 4.4, we show how the null model performs on our running example. The kernel width is fixed to the value 2. The z-scores yield four local maxima, however only two of them are strictly positive. These are candidates for consensus breakpoints. The corresponding locations are 30 (z-score 3.59) and location 12 (z-score 3.76). Additionally, note how the mean null score drops towards the ends of the chromosome, due to data censoring. This phenomenon affects therefore both the null model and the true model, which ensures a fair local comparison.

4.4.6. Summarizing the output of all kernel widths

For each particular value of the kernel width, significance z-scores are available at all breakpoint locations $(v_i, s_i, w_i) \in V$. For all kernel widths, we retrieve the local maxima

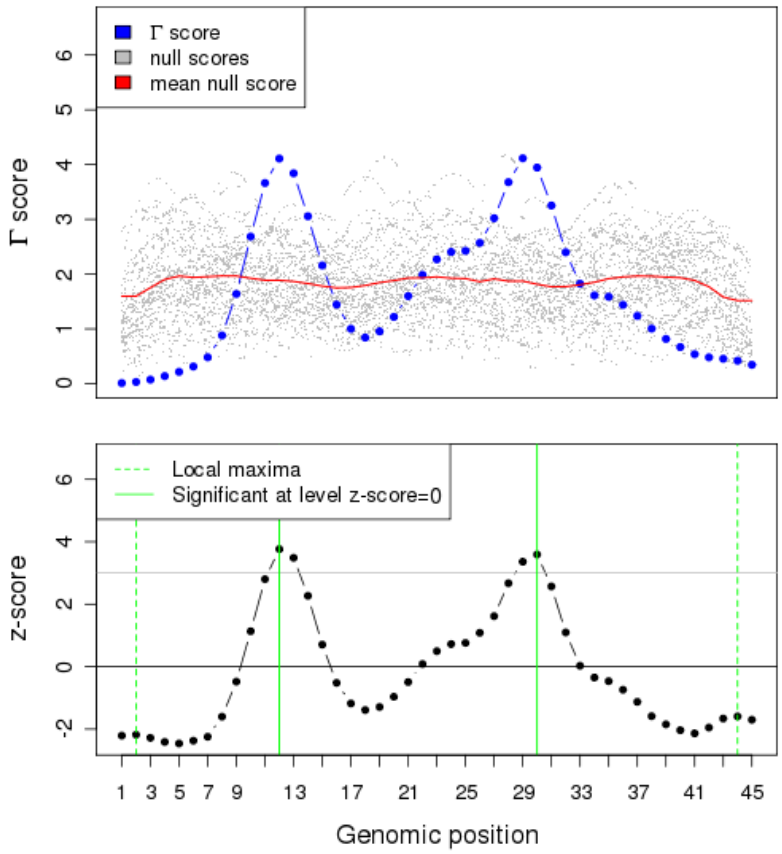


Figure 4.4.: Procedure for assessing significance of the Γ score. In the upper panel, gray dots represent the null Γ scores and the red line estimates the average null score by a moving average. The true scores are represented in blue. In the lower panel, the z-scores associated to each location are presented. The local maxima of the z-scores are marked with vertical green lines. These are consensus breakpoint candidates. The dotted lines mark candidates which do not exceed the significance threshold. The final consensus breakpoints for the particular kernel width are marked with solid green lines.

which yield a z-score larger than some positive cutoff. This procedure gives the list of consensus breakpoints, to which significance z-scores are associated. The situation can occur that a certain genomic location is reported as significant consensus breakpoint by two kernel widths. In this case, the larger z-score is considered.

For example, in Figure 4.5 we show the list of consensus breakpoints with positive z-scores resulting from applying the CB-KeS algorithm to our running example with kernel widths $\sigma = 1.5$, $\sigma = 2$ and $\sigma = 4$. There are seven local maxima of the corresponding z-scores which are positive, five after removing duplicates. In the bottom panel of Figure 4.5, the breakpoints are shown, at locations 12 (z-score 4.18), 13(z-score 1.80), 24(z-score 0.95), 29(z-score 2.96) and 30(z-score 3.59).

The final output of the CB-KeS algorithm is a list of consensus breakpoint candidates of various width, sorted decreasingly by z-score. Choosing an optimal top-k consensus breakpoints is discussed in Section 4.6.

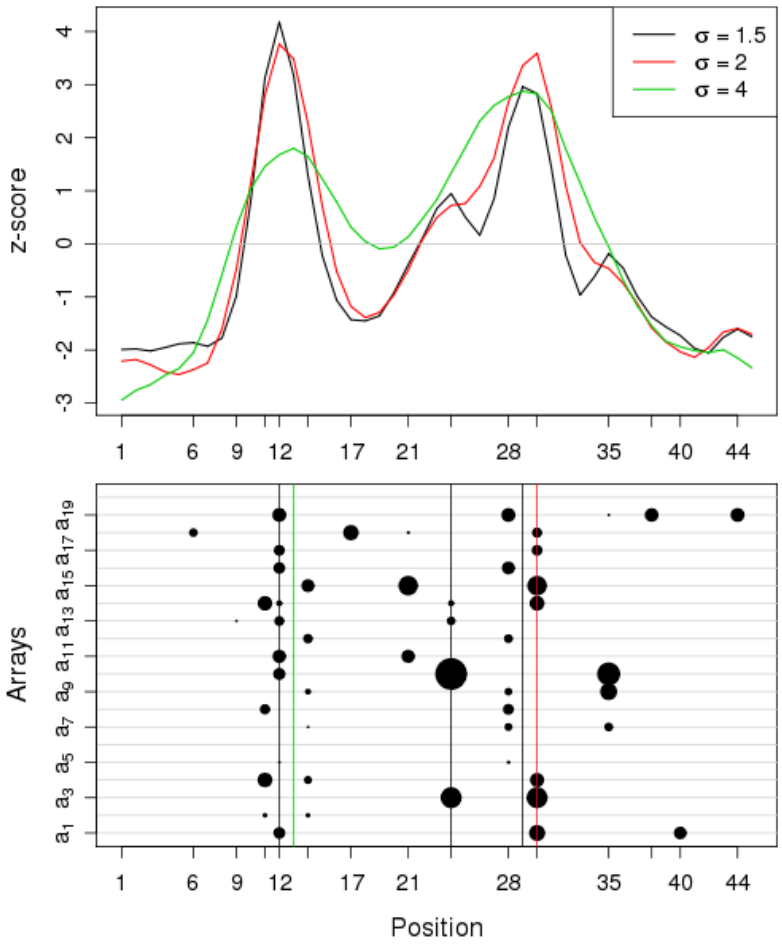


Figure 4.5.: Summarizing the output of all kernel widths with Cb-KeS algorithm. In the upper panel, the z-scores are shown for three choices of kernel width σ. In the bottom panel, the positive local maxima of the z-scores are marked with vertical lines, the color of which correspond to the kernel width that yields the largest z-score.

4.4.7. The algorithm

Algorithm 2 summarizes the steps of the CB-KeS algorithm.

4.5. Algorithms for finding consensus regions

The goal is to find groups of neighboring probes (consensus regions), the copy number of which is almost constant over all arrays. This task can be addressed by clustering the probes represented in the space of the arrays, where the distance between two probes is small if the copy number profile across the arrays is very similar.

However, clustering the probes along the chromosome does not guarantee that each cluster is composed of a sequence of consecutive probes (a region, according to the terminology from this manuscript). In fact, a cluster can be broken into several regions. A simple solution for obtaining a partition into regions is to run a normal clustering procedure (using any clustering approach) and then express each cluster C_k as a union of regions $C_k = \bigcup_i R_k^i$.

Algorithm 2 CB-KeS

Require: The input breakpoints $V = \{(v_i, s_i, w_i) \mid 1 \leq i \leq T\}$; the number of null instances to generate P; a set of kernel widths SD.

Ensure: A set of consensus breakpoints $B_1, ..., B_m$.

1. **For** each kernel width σ in SD **do**
   1.1. Compute the scores $\Gamma(v_i; \sigma)$, for all $1 \leq i \leq T$.
   1.2. Set $null.locations :=$ NULL and $null.scores :=$ NULL
   1.3. **Repeat** P times
      1.3.1. Generate null problem setting: $V^0 = \{(v_i^0, s_i, w_i) \mid 1 \leq i \leq T\}$.
      1.3.2. Set $null.locations := null.locations \bigcup \{v_i^0 \mid 1 \leq i \leq T\}$.
      1.3.3. Compute null scores $\Gamma(v_i^0; \sigma)$, for all $1 \leq i \leq T$.
      1.3.4. Set $null.scores := null.scores \bigcup \{\Gamma(v_i^0; \sigma) \mid 1 \leq i \leq T\}$.
   1.4. **For** each location v_i **do**
      1.4.1. Compute a significance z-score $z_i(\sigma)$ by comparing $\Gamma(v_i; \sigma)$ to the $null.scores$ at $null.locations$ which are close to v_i (for example, 500 nearest $null.locations$ to v_i).
   1.5. Set $z(\sigma) := \{z_1(\sigma), ..., z_T(\sigma)\}$ and define the set of candidate consensus breakpoints $C(\sigma) := \{i \mid z_i(\sigma)$ is a local maximum of sequence $z(\sigma)$ and $z_i(\sigma) > 0\}$.
2. Set $C := \bigcup_{\sigma \in \text{SD}} C(\sigma)$. Denote the elements of C by $B_1, ..., B_m$.
3. Output $B_1, ..., B_m$ as consensus breakpoints.

The final set of regions is given by the set $\{R_k^i\}_{i,k}$.

In order to perform clustering, a distance function between the probes needs to be defined. For some clustering methods, a distance matrix (dissimilarity matrix) between any pair of probes suffices. The choice of the distance influences the quality of the clustering, namely the shape of the clusters and the optimal number of clusters. In this manuscript we compute the distance between two probes by means of the *Euclidean* distance (L_2 norm).

Among the most widely used clustering methods we mention the K-means, K-medoids or PAM (Partitioning Around Medoids, Kaufman and Rousseeuw (1990)) and hierarchical clustering.

The goal of the K-means algorithm is to find a set of K clusters, such that the sum of square distances from the observations to the means of the clusters is minimized. The problem is NP-hard, therefore K-means adopts an iterative approach which guarantees convergence to a local minimum. The K-medoids algorithm is similar to K-means in the sense that it returns K clusters with the property that each observation belongs to the cluster with the nearest centroid. In the case of K-medoids, the centroid of a cluster is not the mean, but the observation that minimizes the sum of dissimilarities within the cluster. Hence, K-medoids requires only the dissimilarity matrix of the observations as input. In terms of shape, both algorithms return spherical clusters. From the point

of view of computational efficiency, both K-means and K-medoids are slow, especially since in practice a large set of values of K need to be evaluated. For large datasets, a faster algorithm for K-medoids called **clara** (Kaufman and Rousseeuw, 1990) is available, however it is slower than hierarchical clustering (discussed below).

In contrast to K-means and K-medoids, which are based on optimizing a loss function, hierarchical clustering is a heuristic approach for building a hierarchy of clusters in the shape of a binary tree. At level K of the tree, the observations are clustered into K groups. The leaves are singleton sets, each containing one observation. Each parent node is the cluster given by the union of its two children. In order to compute the hierarchical clustering, a bottom-up approach is typically used. The algorithms starts with each cluster being a singleton and proceeds by merging at each step the two closest clusters. As a consequence, apart from the distance between observations, hierarchical clustering requires a distance between clusters to be defined (also called *linkage*). We present the most commonly used linkage distances. *Complete-linkage* is defined as the maximum distance between two observations belonging to the two clusters. It tends to determine compact clusters of similar size and it is sensitive to outliers. *Average-linkage* is given by the average distance between all pairs of observations in the two clusters. It is fairly robust. *Single-linkage* is the distance between the two closest observations of the two clusters. It is more suitable than the other methods for cases in which clusters are not spherical or elliptical in shape (Everitt et al., 2001). The function **hclust** from the **R** package **stats** can be used for performing all variants of hierarchical clustering. Hierarchical clustering is also attractive because of its computational efficiency – all clusterings with number of clusters ranging from one to the number of observations being accessible in constant time.

In order to choose a suitable combination of clustering algorithm and cluster validity measure, we visually investigated the spatial structure of the probes in cancer datasets. Figure 4.6 shows 2-dimensional representations of the probes in the space of samples, via multidimensional scaling. The data are from a collection of arrays on breast cancer and we selected four chromosomes for illustration. It is apparent that the probes do not form spherical or elliptical-shaped groups, but chain-like structures. This appearance is easy to explain: neighboring probes are very similar due to the local constancy of DNA copy number, but far-away probes are distinct through the accumulation of many subtle local changes.

At first glance, the chain-like structure suggests the single-linkage approach of hierarchical clustering. However, we argue in what follows that centroid-based clustering algorithms is more appropriate. Indeed, our ultimate goal is to find a set of representatives that approximate well the data and achieve dimension reduction. If we choose single-linkage, each chain-like cluster will have to be represented by its centroid, which will probably lie very far away from the ends of the chain. We believe that several representatives will summarize better the information contained within chain-like clusters. A centroid-based clustering would probably break the cluster into smaller elliptical-shaped groups, which serves well our goal.

Following the above reasoning and also considering computational efficiency, we chose hierarchical clustering with complete linkage as clustering method. The resulting approach for finding consensus regions will be called CR-FC in this manuscript (Consensus Regions via Feature Clustering).

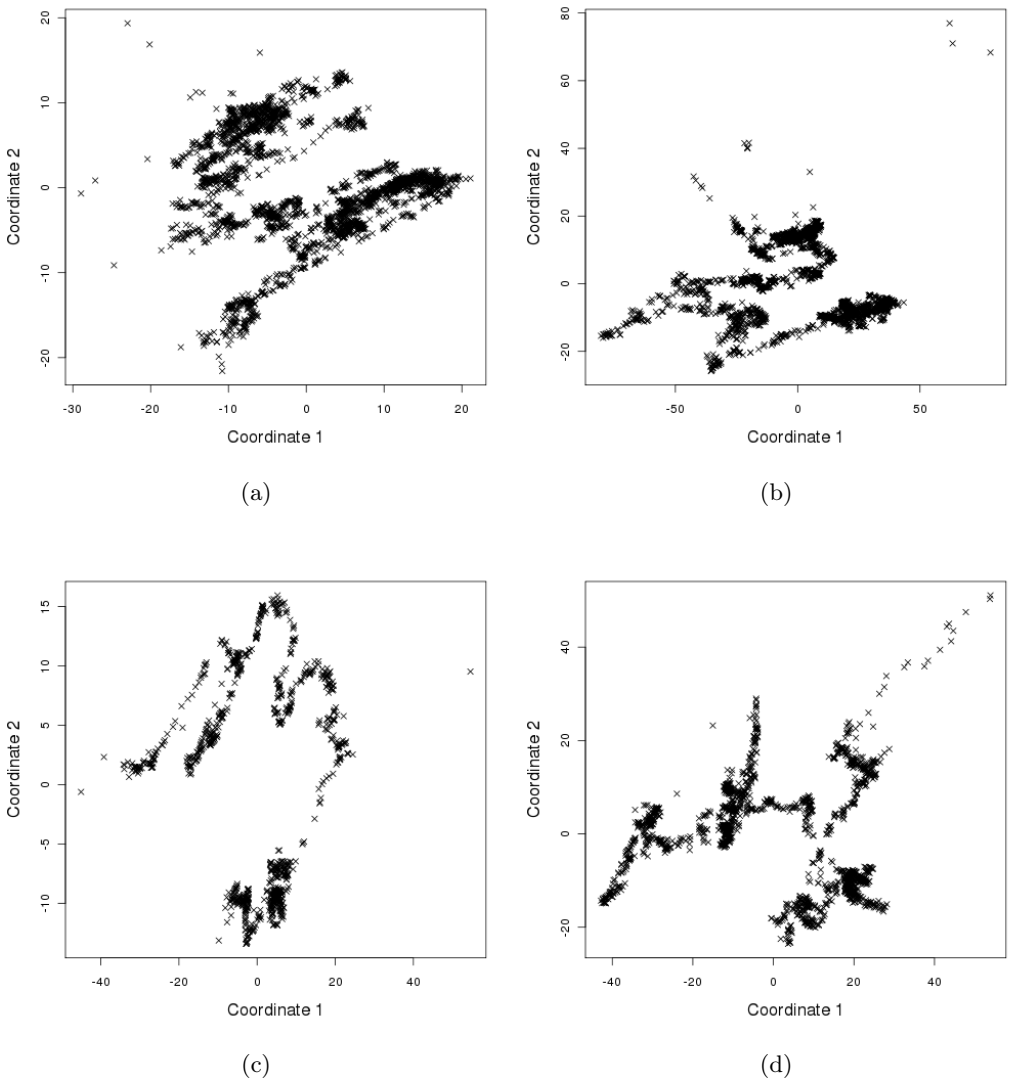


Figure 4.6.: Multidimensional scaling of probes represented in the space of samples. Spatial structures in shape of chains are apparent. Data are from a collection of arrayCGH experiments on breast cancer. a) chromosome 2, b) chromosome 11, c) chromosome 13, d) chromosome 17.

For the problem of selecting the optimal number of clusters, see the discussion on model selection from Section 4.6.

4.6. Model selection: evaluating the quality of consensus segmentation

In the following, we discuss the problem of selecting the optimal number of consensus regions m. We distinguish two classes of methods: supervised and unsupervised.

The supervised approach consists of fitting a consensus segmentation model (with any algorithm) for each value of m from a range $\{m_{\min}, ..., m_{\max}\}$. For each consensus segmen-

tation, the reduced representation is computed, consisting of the set of representatives. Then, the representatives are used as meta-features in a supervised learning task, where the predicted variable is any indicator of tumor phenotype available. The accuracy of prediction is used as a criterion for selecting the best value of m. Two observations are important here: first, note that the selection of the optimal number of regions depends on the phenotype used and on the prediction model chosen. Therefore, the optimal set of consensus regions may not be useful for other unrelated tasks. Second, if cross validation is used for estimating the prediction accuracy, then a different consensus segmentation must be performed for each fold separately, only on the training data. Hence, the computational effort is larger. In Chapter 5, we present methods based on supervised learning which include the selection of the optimal number of consensus regions.

In this chapter, we introduce an approach to unsupervised model selection. To this end, we regard the problem of selecting the optimal number of regions as similar to the task of selecting the optimal number of clusters in a classical clustering setting. The latter task is often called *assessment of cluster validity*. It is notoriously difficult and it has been approached by a multitude of methods over the last decades. *Dunn's index* (Dunn, 1974) and the *Davies-Bouldin index* (Davies and Bouldin, 1979) are widely-used measures for cluster validity which express the trade-off between within-cluster and between-cluster distances. Bezdek and Pal (1998) present a generalization of Dunn's index which outperforms the baseline on spherical-shaped clusters. Rousseeuw (1987) proposed the so-called *silhouette values*, which provides a standardized measure of how well each observation fits the current cluster assignment. The number of clusters is chosen such that the average silhouette over all observations is maximized. Tibshirani et al. (2001) introduce the *gap statistic*, which compares the average distance between the observations and the cluster centroids to a null clustering in which the observations are uniformly distributed within a bounding box. The number of clusters resulting in the largest difference between the observed and null distance is optimal. The gap statistic is computationally expensive. Jung et al. (2003) introduce *clustering balance*, a measure that combines intra-cluster distance and inter-cluster distance in such a way that allows for an efficient computation over the levels of a hierarchical clustering dendogram.

Most of the cluster-validity measures appreciate positively clusters that are compact and well separated. As we showed in Figure 4.6, in our application the probes do not form well separated clusters, but rather elongated chains with smooth transitions. In such cases, our experiments show that the very popular *silhouette values*, but also Dunn's index and the Davies-Bouldin index increase monotonically until each probe forms its own cluster. For the sake of dimension reduction, we aim at finding a small number of clusters and representatives. To this end, we used and adapted the *clustering balance* measure of Jung et al. (2003) as explained below.

4.6.1. Weighted clustering balance

Assume a clustering into m clusters given. The original *clustering balance* measure (Jung et al., 2003) involves two quantities: the *intra-cluster error* ($\Lambda(m)$) and the *inter-cluster error* ($\Gamma(m)$). The intra-cluster error is the sum of the Euclidean distances between the observations and the centroids of the clusters to which they belong. The intra-cluster

error decreases monotonically as the number of clusters increases. The intra-cluster error is computed as the sum of distances between the centroids of the clusters to a global centroid, which is given by the mean of the cluster centroids. The inter-cluster error increases monotonically as the number of cluster increases and it can be interpreted as a penalty on the number of clusters (or model complexity, if the model is given by the centroids of the clusters). The authors Jung et al. (2003) claim that a good clustering is minimizing a weighted sum of the two errors:

$$\Omega\alpha(m) = (1-\alpha)\Lambda(m) + \alpha\Gamma(m),$$

$$m_{opt} = \arg\min \Omega_\alpha(m)$$

for some choice of $\alpha \in [0,1]$. Jung et al. (2003) choose $\alpha = 0.5$ in their experiments but we conducted a simulation study which showed that a larger value of α is more suitable for dealing with structures similar to those shown in Figure 4.6. Specifically, we show that a very large penalty on the number of clusters (or indirectly Γ) can retrieve a more meaningful clustering.

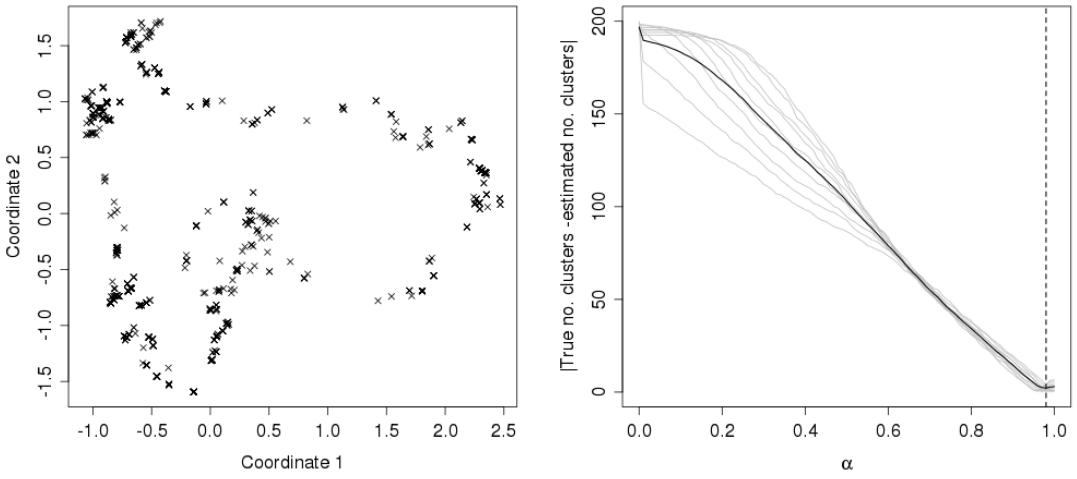
We simulated datasets consisting of $N = 50$ samples and $p = 1000$ probes along an artificial chromosome. For various values of m, $m \in \{1, ..., 10\}$, we randomly selected m independent locations (values between 1 and p), which are set to be consensus breakpoints. We generate piecewise constant log-ratios for each of the N samples independently, with changes occurring around the consensus breakpoints. The precise locations of the changes are obtained by randomly sampling from Gaussians centered at the locations of the consensus breakpoints and with standard deviation randomly selected from the set $\{5, ..., 30\}$. A probability of change is also associated to each consensus breakpoint, randomly selected from the set $\{10\%, ..., 70\%\}$. In this manner, we simulate different recurrence frequencies of copy number change at the consensus breakpoints.

To make the data more realistic, random copy number changes were also added to the artificial tumor samples. For each sample, a number from the set $\{0, 1, 2, 3\}$ is generated at random and as many random locations are selected on the artificial chromosome to yield copy number changes.

Figure 4.7a shows a two-dimensional embedding of one of the artificial datasets, using the same multidimensional scaling method as in 4.6. One can easily notice the same chain-like structure, which provides a minimal visual assurance that the artificial data resembles the real data.

For each value of m, we simulated 100 artificial chromosomes with the procedure above. For a fixed m, the true consensus segmentation consists of $m+1$ regions (clusters), starting at the beginning of the chromosome and with each of the consensus breakpoints, respectively. We applied hierarchical clustering to each of the datasets, then transformed the clustering into consensus segmentations and then selected the number of regions minimizing the weighted clustering balance Ω. The value of the penalty α was set to range from 0 to 1 at increments of 0.01. Figure 4.7b shows the absolute error of the estimator of number of regions as a function of α. The value $\alpha = 0.98$ yields the smallest error, on average.

In our experiments on real data, we use the weighted clustering balance measure with $\alpha = 0.98$ ($\Omega_{0.98}$) for estimating the number of regions of consensus segmentation. Specifically, we minimize $\Omega_{0.98}$ in order to select the optimal number of Gaussians in the case of CB-



(a) (b)

Figure 4.7.: a)Multidimensional scaling of the probes of one artificially generated chromosome with the procedure described in Section 4.6.1. b) Relation between the penalty parameter α and the accuracy of estimation of the true number of clusters. Each of the the gray lines indicates the performance w.r.t. the one particular value of the true number of clusters m. The solid black line averages the performance over all values of m. A minimal error corresponds to the value $\alpha = 0.98$.

MUG, in order to select the top-k consensus breakpoints with CB-KeS and in order to select the optimal number of clusters with CR-FC.

4.7. Results

4.7.1. Datasets

We have applied the consensus segmentation algorithms on several datasets, consisting of arrayCGH experiments performed on various types of cancer tissue (see Table 4.1 for a summary). Below we describe the datasets.

Breast cancer datasets

According to recent world-wide statistics published by the International Agency for Research on Cancer (IARC[1]), breast cancer has the largest incidence among cancers (accounting for 22.9% of all cancers in women) and the highest mortality rate (13.7% of all deaths caused by cancer). As a consequence, breast cancers are being intensely investigated, and so far progress has been made in the direction of subtype identification. The subtypes are characterized by specific clinicopathological and molecular signatures, such as patient's age, tumor grade, tumor stage, estrogen or progesterone receptor status, patterns of altered gene expression, patterns of copy number aberrations (Sørlie, 2004; Fan et al.,

[1] http://globocan.iarc.fr

2006; Kapp et al., 2006; Yang et al., 2007; Lund et al., 2010; Dawood et al., 2011). Large efforts are made to find suitable treatments targeting specific subtypes, however several subtypes remain poorly understood and the mortality rate in the corresponding groups stays high.

André et al. (2009) performed a large study on the impact of chromosomal aberrations on breast cancer. The authors find that specific patters of aberrations characterize different tumor subtypes and also correlate with levels of gene expression, which may point to driver genes and possible drug targets. The study concludes that the analysis of copy number aberrations in breast cancer can be an important complement to the frequently approached expression analysis.

We have obtained from the Gene Expression Omnibus (GEO[2]) three cohorts of DNA copy number experiments on breast cancer, all using the same microarray platform, namely Agilent Human Genome CGH Microarray 244A. The microarrays consist of 236 000 60-mer oligonucleotide probes, covering coding and non-coding sequences with high resolution.

The first cohort consists of 54 breast tumors with amplification at the ERBB2 gene locus (17q12-q21). The experiments and a preliminary analysis have been described in Sircoulomb et al. (2010). The particular interest in the ERBB2-amplified cancers is motivated by their poor prognosis and molecular heterogeneity, as well as by their frequent resistance to targeted treatment. Available phenotypes for this dataset are: age, estrogen receptor (ER) status, progesterone receptor (PR) status, tumor grade, tumor subtype: IBC (inflammatory breast cancer) or NIBC (non-inflammatory breast cancer). In this manuscript, we call this dataset **breast54**.

Another set of arrayCGH experiments comprises 173 breast tumors, out of which 49 are inflammatory breast cancers and 124 are non-inflammatory breast cancers. The experimental data were published by Bekhouche et al. (2011). Inflammatory breast cancer is a very aggressive form of breast cancer, almost always lethal because of its high potential to metastasize. Moreover, they are poorly characterized from a molecular perspective. We call this dataset **breast173**. Annotated phenotypes to this cohort are age, ER status and PR status.

The third set of array CGH experiments consists of 167 breast tumors, published by Russnes et al. (2010). The dataset also contains annotation with age at diagnosis, ER status, PR status, tumor stage, grade, lymph node status and histological subtype. We will call this dataset **breast167**.

Colon cancer dataset

Colon cancer or colorectal cancer occurs mostly in older patients (commonly after the age of 60) and has a higher incidence in developed countries (60% of the cases occur in developed regions, according to the IARC[1]). As a consequence, environmental factors, sedentary lifestyle and nutrition rich in processed food is considered to increase the risk of colorectal cancer. Despite continuous improvement of therapy, the mortality rate remains high at 40%.

The role of copy number aberrations in the prognosis of the disease has been a subject of

[2]http://www.ncbi.nlm.nih.gov/geo/
[1]http://globocan.iarc.fr

Dataset	No. samples	Clinico-histopathological information
colon	98	no information available
breast54	54	**Age(years)**: 31 (min), 48 (median), 80 (max) **Type**: 21 (IBC), 30 (NIBC[4]) **ER status**: 25 (negative), 22 (positive) **PR status**: 28 (negative), 19 (positive)
breast173	173	**Age(years)**: 24 (min), 53 (median), 84 (max) **ER status**: 59 (negative), 114 (positive) **PR status**: 71 (negative), 102 (positive)
breast167	167	**Age**: 28 (min), 64 (median), 90 (max) **Grade**: 11 (Grade 1), 109 (Grade 2), 43 (Grade 3) **Stage**: 57 (Stage 1), 86 (Stage 2), 12 (Stage 3), 6 (Stage 4) **ER status**: 65 (negative), 86 (positive) **PR status**: 66 (negative), 98 (positive) **Histology**: 108 (Ductal), 40 (Lobular) **Lymph node status**: 73 (negative), 70 (positive)
ovarian	290	no information available
glioblastoma	539	no information available
neuroblastoma	162	**Age(months)**: 0 (min), 13 (median), 299 (max) **Stage**: 28(Stage 1), 19(Stage 2), 29(Stage 3), 57(Stage 4), 29 (Stage 4S)

Table 4.1.: Array CGH datasets for validation of consensus segmentation methods.

scientific dispute. Studies either report clear correlation between specific aberration patterns and tumor development (Tsafrir et al., 2006), or fail to identify any such associations (Nakao et al., 2004). In a meta-analysis by Walther et al. (2008), a large number of studies have been compared with the conclusion that there exists a clear association between chromosomal imbalances (aneuploidy or polyploidy) and clinico-pathological characteristics of the tumors.

Veeriah et al. (2010) have investigated a cohort of 98 colon tumors in a study that followed the goal of characterizing somatic mutations of the gene PARK2 in a variety of human malignancies. The experiments based on the Agilent Human Genome CGH Microarray 244A microarrays (as the breast cancer datasets presented above) have been made public via the GEO. The resolution of the arrays is sufficiently high to allow the detection of breakpoints with an average precision of 10 Kb. In this manuscript, we refer to this dataset as **colon**.

Ovarian cancer dataset

Ovarian cancer accounts for 3.7% of all female cancers (according to the IARC). It is difficult to diagnose and consequently, most tumors are already in an advanced stage at the beginning of therapy. Survival in such cases is poor, only 30% of women being expected to survive for five years (Cho and Shih, 2009). Copy number alterations have been associated with clinico-pathological features of the tumors (Cho and Shih, 2009). Specifically, sub-chromosomal (meaning short) amplifications and deletions are usually a sign of poor prognosis.

A comprehensive collection of arrayCGH experiments on 290 ovarian cancer samples

is publicly available via The Cancer Genome Atlas (TCGA[3]). TCGA is a large consortium involving universities and laboratories in the USA that have as a common goal the systematic investigation of molecular changes that characterize a variety of cancer types. Experimental assays on copy number aberrations represent an important part of the data collected at TCGA. The particular set of ovarian tumors have been analyzed using the Agilent Human CGH 1×1M G4447A arrays, with approximately one million probes covering the genome. We call this dataset **ovarian**.

Glioblastoma dataset

Glioblastoma is the most common type of primary brain tumor, although it occurs in only $2 - 3$ cases per $100,000$ people in Europe and North America (source: Wikipedia). It is a very aggressive tumor and. despite the recent treatment improvements, the median survival of the glioblastoma patient is only 14 months (Van Meir et al., 2010). In fighting glioblastoma, surgery is the most efficient option, often achieving a reduction of up to 99% of all tumor cells. Chemotherapy is limited due to the blood-brain barrier and radiotherapy is often dangerous because it can damage healthy brain cells, which have very limited capacity of repairing themselves. The causes of glioblastoma are not fully understood, although recent results point to a viral agent (Vilchez et al., 2003).

Copy number aberrations are present in large number in glioblastoma tumors. Verhaak et al. (2010) identified four different tumor subtypes based on transcriptional patterns and, for each subtype, found characteristic copy number alterations. The authors conclude that there exist key copy number aberrations which lead to altered transcription and therefore play a role in tumor development. We have obtained of 539 glioblastoma arrays publicly available via the TCGA, analyzed using Agilent Human Genome CGH 244A microarray experiments. We refer to this dataset as **glioblastoma**.

Neuroblastoma dataset

Neuroblastoma is a tumor that affects the sympathetic nervous system. It occurs predominantly in children of young age and infants, at an incidence rate of 1 out of $100,000$. For the purpose of therapy selection several factors are currently considered, including classical staging and age at diagnosis. Neuroblastoma staging assigns the tumor to one of five subgroups, defined by taking into account the histological features of the tumor, spread to nearby organs and lymph node and importantly, age of the patient. The stages are 1, 2, 3 (good prognosis, in general localized tumors), 4S (metastasized, with spread limited to liver, skin, or bone marrow and patient younger than one year) and 4 (metastasized, with spread other than defined by 4S). Stage 4 neuroblastoma has a poor prognosis, in general, despite aggressive chemotherapy, whereas stage 4S neuroblastoma has a good prognosis, in general, regressing spontaneously. Age plays an important role in neuroblastoma, most infants (younger than one year) having a very good prognosis, with regression under minimal treatment. In contrast, older patients have poor prognosis, if the tumor is in an advanced stage. Recent studies have suggested that a larger age cutoff (18 months) may be more appropriate for neuroblastoma classification (London et al., 2005), which if applied in clinical practice, could spare many young children from unnecessary aggressive chemotherapy.

[3]http://cancergenome.nih.gov/

Copy number aberrations have been shown to influence the outcome of patients with neuroblastoma. Polyploidy is generally indicative of good outcome and is characteristic to tumors occurring in infants. Amplification of the MYCN gene located on chromosome 2 is associated with poor outcome. However, this is not the only feature responsible for aggressive phenotype, as it is not observed in all aggressive tumors. Deletion of 11p and 1p, amplification of 17q are also indicative of poor prognosis (Ambros et al., 2009).

Copy number aberrations are important for the diagnosis and treatment of neuroblastoma patients. The amplification of the $MYCN$ oncogene or the loss of parts of chromosomes 1 and 11 have already been introduced in clinical practice as genetic markers.

Through a cooperation with the research lab of Prof. Frank Berthold from the Department of Pediatric Oncology and Hematology from Köln University Clinic, we have investigated a cohort of 162 neuroblastoma tumors. The experimental assays use several arrayCGH microarray platforms form Agilent (44k and 100k resolution). A subset of the tumors have been introduced and analyzed by Spitz et al. (2006). The neuroblastoma dataset will be called **neuroblastoma** in this manuscript.

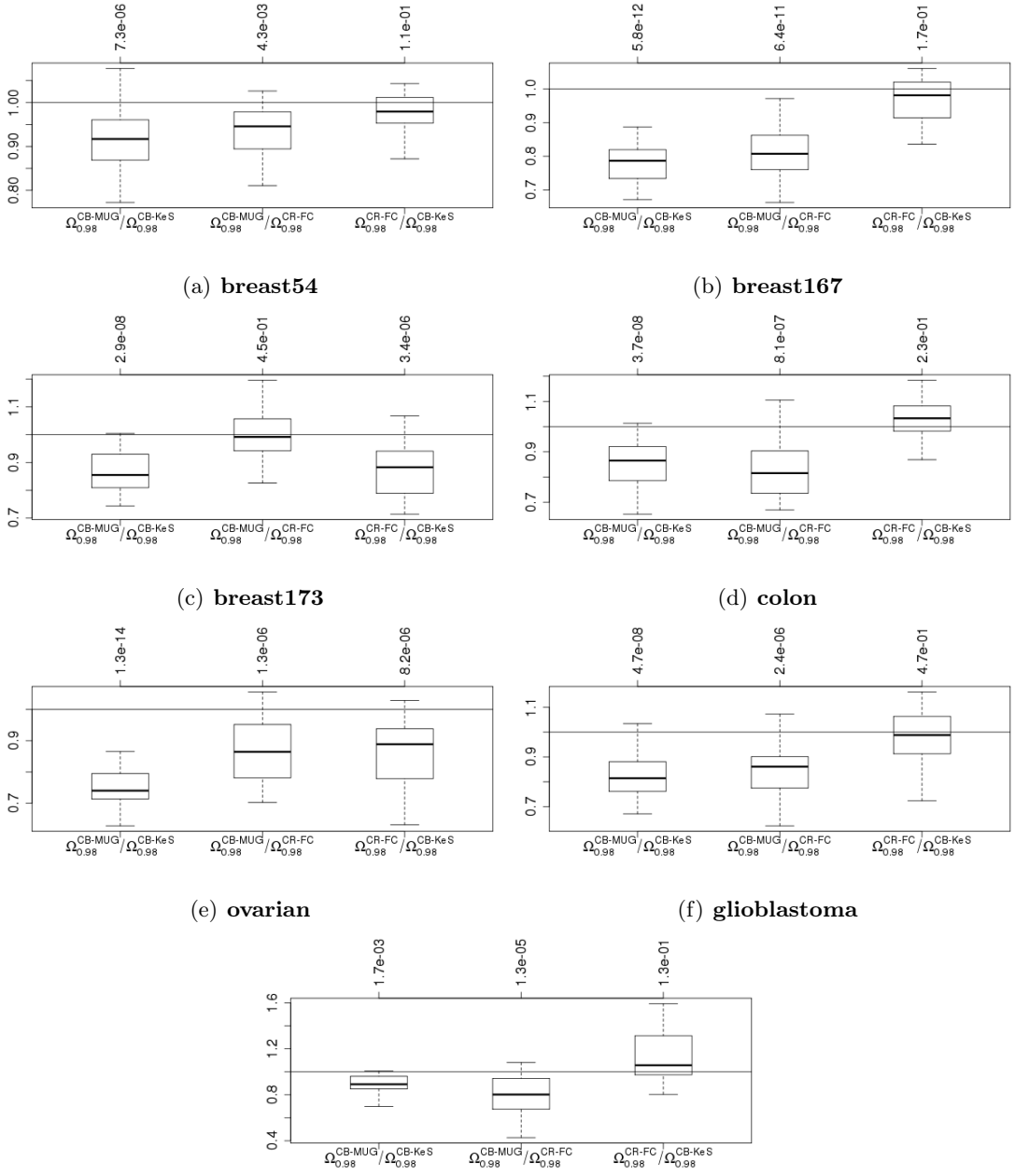
4.7.2. Evaluation of the algorithms for identification of consensus breakpoints

	No. breakpoints		**No. consensus breakpoints**		
	per tumor	total	CB-MUG	CB-KeS	CR-FC
neuroblastoma	54 ± 24	4047	$70 \pm 2(57)$	$62 \pm 7(65)$	$119 \pm 8(34)$
colon	168 ± 62	16985	$223 \pm 6(76)$	$173 \pm 11(98)$	$372 \pm 21(45)$
glioblastoma	261 ± 72	64729	$329 \pm 7(196)$	$492 \pm 35(132)$	$818 \pm 39(79)$
breast173	339 ± 115	49266	$425 \pm 5(116)$	$320 \pm 7(154)$	$1029 \pm 34(48)$
breast54	394 ± 85	20093	$429 \pm 15(47)$	$327 \pm 8(62)$	$679 \pm 44(30)$
breast167	461 ± 182	37055	$517 \pm 8(72)$	$503 \pm 11(73)$	$977 \pm 59(38)$
ovarian	806 ± 201	125000	$605 \pm 8(207)$	$662 \pm 8(189)$	$2061 \pm 48(61)$

Table 4.2.: Number of breakpoints and consensus breakpoints identified in seven cancer datasets – genome-wide statistics. The first column contains the average number of breakpoints per tumor, with indication of standard deviation. The second column shows the total number of distinct breakpoints in the cohorts. Columns three, four and five contain the number of consensus breakpoints identified in the datasets by three consensus segmentation algorithms. Standard deviation is computed based on cross validation. In brackets, the magnitude of the dimension reduction by consensus segmentation is indicated.

We applied the CB-MUG, CB-KeS and CR-FC algorithms on all cancer datasets presented in the Data section. For estimating the optimal number of regions, with each algorithm independently we optimized the $\Omega_{0.98}$ measure (see Section 4.6.1). We compared the algorithms first quantitatively, based on three criteria: the quality of segmentation (given by the minimum value of $\Omega_{0.98}$), the stability of segmentation with respect to perturbations of the set of samples and the magnitude of dimension reduction as a result of segmentation (ratio between initial number of probes and resulting number of regions).

Second, we analyzed the resulting consensus breakpoints qualitatively, by investigating the genetic and epigenetic properties which characterize the respective DNA locations.



(a) **breast54** (b) **breast167**

(c) **breast173** (d) **colon**

(e) **ovarian** (f) **glioblastoma**

(g) **neuroblastoma**

Figure 4.8.: Comparative evaluation of the CB-MUG, CB-KeS and CR-FC algorithms: quality of consensus segmentation. Each box-plot summarizes 24 chromosome-wise ratios between the $\Omega_{0.98}$ scores of two of the three algorithms. The subfigures correspond to different datasets. The top axis shows p-values resulting from t-tests which compare the mean of the population of ratios to one.

Figure 4.8 shows a comparison of the algorithms based on the quality of consensus segmentation. We considered each chromosome of each dataset a separate instance of consensus segmentation and we considered the optimal $\Omega_{0.98}$ value of each algorithm as a measure of performance. Smaller values indicating better performance, we conclude from our results that the CB-MUG approach is significantly superior to CB-KeS on all datasets.

The significance p-value is annotated on the upper horizontal axis and is the result of a t-test that compares the mean of the population of ratios between the $\Omega_{0.98}$ values of two algorithms with one. The CB-MUG algorithm is also superior to CR-FC, significantly in all applications, except for the **breast173** dataset. The comparison between CR-FC and CB-KeS favors to the latter in five out of eight applications, out of which only three yield significant p-values. On the **neuroblastoma** and **colon** datasets, CR-FC appears slightly better, but not significantly better.

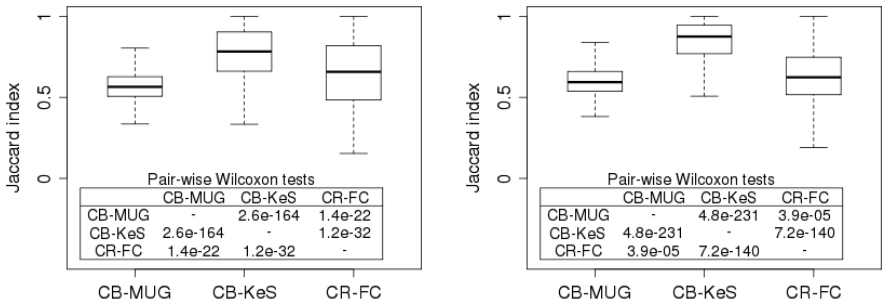
Table 4.2 shows several statistics on the number of breakpoints observed in the tumors investigated and the number of consensus breakpoints identified by our algorithms. The first column indicates the mean and standard deviation of the number of breakpoints per tumor in each cohort, which is a good indicator of the respective degree of genomic instability. Neuroblastoma stands out from the rest of the cancer types with very few breakpoints per tumor, a striking difference probably related to the fact that neuroblastoma appears early during infancy and DNA aberrations do not accumulate over a long time as in the case of all adult tumors. Moreover, segmental gains or losses are not characteristic to neuroblastoma, more frequent events being polyploidy or aneuploidy. These are not delimited by breakpoints located strictly within chromosomes, hence the small counts.

The second column of Table 4.2 contains the total number of distinct breakpoints occurring in the tumor cohorts. This is the number of dimensions necessary for representing the datasets, respectively, independently of the resolution of the microarrays. We compare this number with the number of consensus breakpoints resulting from each algorithm (columns three, four and five) for evaluating the dimension reduction by consensus segmentation. As shown in brackets in Table 4.2, the dimension reduction achieved by CB-MUG and CB-KeS is of similar magnitude and is substantial. The number of consensus breakpoints identified by CB-MUG is more stable than CB-KeS with respect to variations of the input set, obtained via 10-fold cross validation. CR-FC outputs significantly more consensus breakpoints than the CB-MUG and CB-KeS and there is also proportionally higher variance among cross-validation folds.

Stability is an important issue in cluster analysis. Therefore, we discuss in the context of the closely related problem of consensus segmentation. More precisely, it is desired that the segmentation does not vary much when the set of observations (tumor samples) is perturbed. Analysis of stability is usually carried out by sampling techniques such as bootstrap or cross validation. We chose cross validation because it is also necessary for downstream analysis (see Chapter 5). We estimate the stability of consensus segmentation via the following procedure. First we split the samples into ten non-overlapping bins. Then, ten times in a row we hold out one bin and we run the consensus segmentation on the data corresponding to the remaining nine bins. This way, we obtain perturbed input sets with 89% overlap. We compare the resulting consensus segmentations by using the *Jaccard index* (Jaccard, 1912), a similarity measure between two partitions of the same dataset. Given two consensus segmentations of the same dataset (chromosome) $C_1 = \{R_1^1, ..., R_{m_1}^1\}$ and $C_2 = \{R_1^2, ..., R_{m_2}^2\}$, the Jaccard index $J(C_1, C_2)$ is computed as:

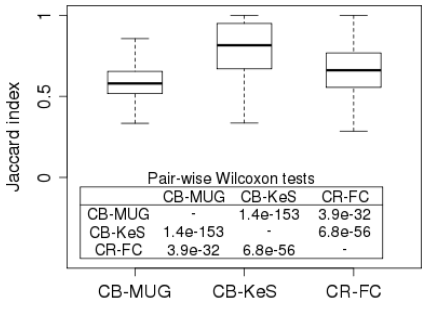
$$J(C_1, C_2) = \frac{a}{a + b + c},$$

where a denotes the number of pairs of probes that belong to the same region from C_1 and

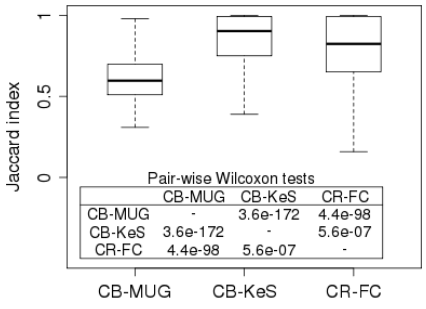


(a) **breast54**

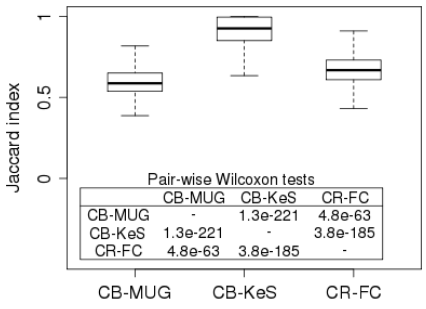
(b) **breast167**



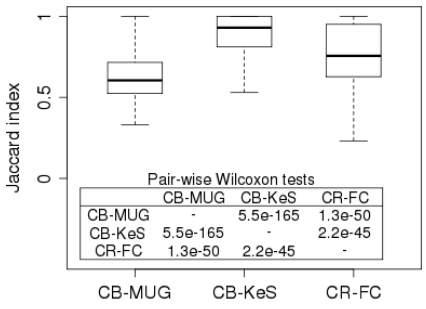
(c) **breast173**



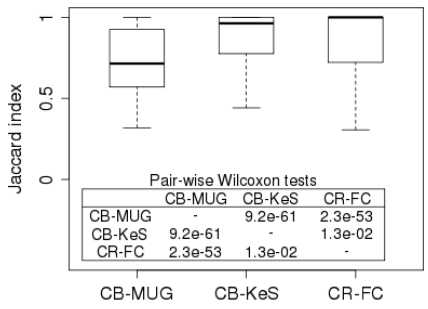
(d) **colon**



(e) **ovarian**



(f) **glioblastoma**



(g) **neuroblastoma**

Figure 4.9.: Comparative evaluation of CB-MUG, CB-KeS and CR-FC: stability of consensus segmentation. Each box-plot summarizes Jaccard indices comparing segmentations between different cross-validation training sets, for each of the 24 chromosomes separately. The tables show p-values resulting from Wilcoxon tests which compare the mean of the Jaccard indices between pairs of algorithms.

also belong to the same region in C_2, b represents the number of pairs of probes assigned to the same region in C_1 but belong to different regions in C_2 and c is the number of pairs of probes assigned to the same region in C_2 but belong to different regions in C_1. The Jaccard index yields a value between 0 and 1, with larger values for similar segmentations.

Given the outputs of the consensus segmentation on each 10 folds, we compute pairwise similarities between them using the Jaccard index, resulting in $\frac{9*10}{2} = 45$ scores. We repeat this procedure for each chromosome separately and summarize the scores for each consensus-segmentation algorithm. Figure 4.9 shows boxplots resulting from the procedure described above. It is clear that the CB-KeS algorithm is most stable, on all datasets investigated. CR-FC ranks second and CB-MUG ranks last with respect to stability, the differences between the average Jaccard index being significant between all pairs of algorithms. We performed Wilcoxon tests and show the significance p-values in the tables annotated in Figure 4.9.

4.7.3. Genomic and epigenomic characteristics of consensus breakpoint locations

We have investigated the genomic and epigenomic properties of the genomic locations at which consensus breakpoints are reported. For this purpose, we used EpiExplorer [2], an in-house web-based application for interactive exploration of sets of genomic regions. EpiExplorer requires as input a set of genomic regions, which we defined as follows. From each consensus breakpoint, we constructed a genomic region which is likely to cover the mass of individual breakpoints that form the consensus breakpoint. We only considered consensus breakpoints given by the CB-KeS algorithm, because CB-KeS is fast, non-parametric and outputs a ranked list of breakpoints, sorted by significance.

We define a consensus breakpoint region as the genomic interval centered at the location of the consensus breakpoint with a width which is twice the kernel width that gave as output the consensus breakpoint. Additionally, the regions are supported by a significance z-score, which is the z-score associated with the consensus breakpoint.

EpiExplorer facilitates for a fast summary of the properties of the set of consensus breakpoints, which include: DNA sequence patterns, overlap with genes and gene elements, overlap with CpG islands, overlap with conserved DNA regions, overlap with known histone marks in various tissue types, etc. Two regions are considered overlapping if they have at least one basepair in common.

For meaningful conclusions, EpiExplorer also facilitates a direct comparison with a set of control regions, which can be randomly generated from the genome, for example. We generated a set of reference genomic regions by randomly selecting genomic intervals, with the constraint that the lengths of the intervals should be the same as the lengths of the consensus breakpoint regions. We have generated only one reference set, containing as many regions as the union of all consensus breakpoint regions from all cancer datasets.

Figure 4.10 summarizes our findings. We observe an enrichment of genes and gene promoters, which indicates that the DNA breakpoints target functional elements and probably disrupt their function. We also notice an enrichment of overlap with CpG islands, which has been reported previously (Tsai et al., 2008; Abeysinghe et al., 2003; Gordon et al., 2007;

[2] http://epiexplorer.mpi-inf.mpg.de/

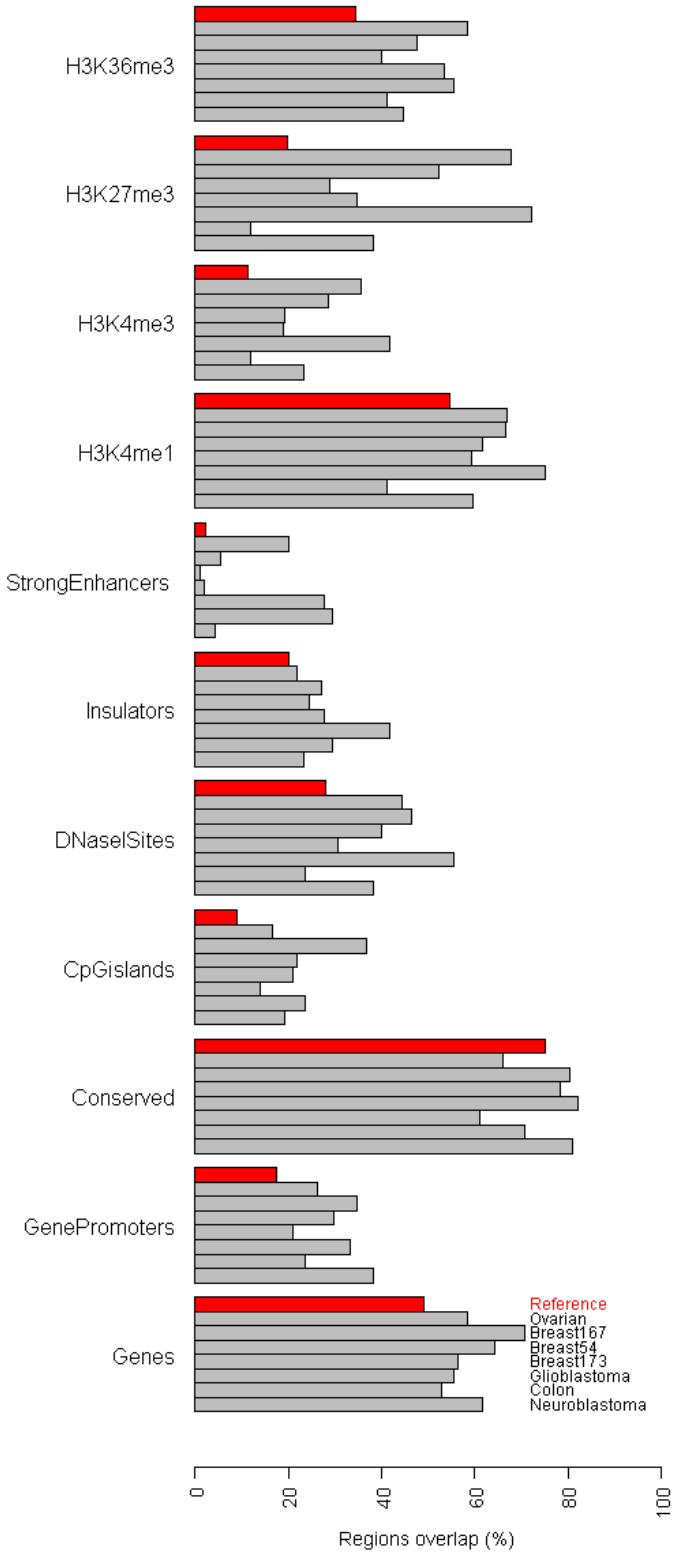


Figure 4.10.: Genetic and epigenetic properties of consensus breakpoint regions by EpiExplorer. On the x-axis, we show the percentage of consensus breakpoints overlapping with some genetic or epigenetic element. On the y-axis, various genetic and epigenetic properties are listed. In gray, the consensus breakpoints from each dataset is represented. In red, the random reference is shown.

Lemaitre et al., 2009). Because the GC-rich content can be attributed to the increased overlap with genes, we excluded the regions overlapping with genes and re-evaluated the overlap with CpG islands (result not showed here). Interestingly, the CpG enrichment persisted. Several studies use evolutionary arguments for explaining the enrichment in CpG islands around breakpoint locations in the genome (Gordon et al., 2007; Lemaitre et al., 2009). Specifically, breakpoints and rearrangements in CpG-rich regions are unlikely to have catastrophic consequences for the cell, given their repetitive patterns. Thus, the cells that acquired such breakpoints survive. At the same time, breakpoints provide a mechanism of functional change affecting the genes and promoters located nearby (Gordon et al., 2007) and contribute to genomic evolution. Lemaitre et al. (2009) suggests that the frequency of breakpoints may be related to the more accessible chromatin in CpG-rich regions, especially in unmethylated CpG islands. Methylation measurements corresponding to the breakpoint hotspots that our algorithms report can be used to confirm this hypothesis. For some of the datasets we use in this thesis, namely the **ovarian** and the **glioblastoma** datasets, public methylation data are available and can be the basis of an interesting follow-up study.

We observe an enrichment of DNA hypersensitive sites (DNaseI Sites in Figure 4.10). These are genomic locations with very high sensitivity to enzyme cleavage and they have been associated before with breakpoint clusters and translocation events in cancer (Zhang and Rowley, 2006). Our results also show slight enrichment of insulator regions overlapping with consensus breakpoints. Insulators are proteins that bind to DNA in specific locations in order to establish transcription boundaries. A famous example is the CTCF insulator protein, the function of which is often disrupted in cancer, for example by hypermethylation (Feinberg and Tycko, 2004). In several datasets, we also detect an enrichment of enhancer regions. These are genomic regions to which proteins bind, with the role of promoting transcription of a cluster of genes. Disregulation of enhancers via copy number aberrations has been reported before (Ford et al., 1985). Finally, we discuss the enrichment of histone marks (H3K4me1, H3K4me3, H3K27me3 and H3K36me3). Histones are proteins that regulate the access to chromatin and thus modulate gene transcription. DNA breaks located at positions regulated by histones probably disrupt the function of the histones and consequently that of the co-localizing genes.

In summary, the properties of the consensus breakpoint locations indicate not surprisingly that the DNA breakpoints target functional genomic regions. We analyzed in the same way the set of *all* DNA breakpoints (not only those located in breakpoint hotspots) and we could not identify the same enrichment of functional elements. Clearly, recurrent breakpoints target specific cellular function in order to render the cell immortal. The results presented in this section constitute a qualitative validation of the CB-KeS consensus segmentation method.

4.8. Applicability to high-throughput sequencing data

The next generation sequencing (NGS) techniques afford more accurate (fewer false positives) and more precise (w.r.t. genomic location) identification of breakpoints in individual tumors. However, the diversity of CNAs among tumors is a biological reality, therefore consensus analysis still meets the challenge of discriminating between driver regions and

passenger alterations. From this perspective, the task of consensus segmentation applies also to high-throughput sequencing data.

Segmentation of individual arrays is a common task for copy number analysis with NGS data, which ensures the input data format that our algorithms need. Efficient algorithms like CBS (Venkatraman and Olshen, 2007) require close to linear running time and have been successfully adapted to NGS data analysis (Campbell et al., 2008). The complexity of the CB-MUG and CB-KeS algorithms depends only on the number of breakpoints identified in the collection of tumors, which is not expected to be much larger than the number of breakpoints given by high-resolution arrays. Therefore, our methods are expected to scale to NGS data.

4.9. Discussion and conclusions

In this chapter, we defined the task of consensus segmentation of a set of DNA copy number experiments and we presented several algorithms that address this task. The CB-MUG and CB-KeS algorithms are targeted towards identifying consensus breakpoints, which are genomic locations characterized by an enrichment in breakpoints. The CB-MUG fits a mixture of Gaussians and one uniform to the population of breakpoints and takes the number of components in the mixture as parameter. The CB-KeS uses Gaussian kernel smoothing with various kernel widths for finding peaks in the population of breakpoints. The CB-KeS uses the amplitude of copy number change as weight for each breakpoint. The qualitative advantages of CB-MUG over the CB-KeS approach are that it can adapt to consensus breakpoints of various diameter and that is it more robust. The advantages of CB-KeS algorithm are its robustness, the fact that it can admit weights and it gives a z-score for each consensus breakpoint.

We argued that the problem of determining the optimal number of regions for consensus segmentation is difficult due to the chain-like structures that probes form if mapped into a space of samples. Most traditional methods address the identification of spherical groups, which means they are not applicable to our data. We propose to use a more flexible measure for estimating the number of regions, which consists of a weighted sum of the intra-cluster and inter-cluster distances. We tune the weight such that the clustering model fits best chain-shaped clusters.

We introduced seven public datasets from arrayCGH experiments on five cancer types. We applied our consensus segmentation algorithms to these datasets and we compared them based on several criteria: the quality of segmentation, given by the $\Omega_{0.98}$ measure, the magnitude of dimension reduction and the stability of segmentation with respect to variations of the sample set. The CB-MUG is superior w.r.t. the $\Omega_{0.98}$ measure, CB-KeS achieves the largest dimension reduction and CB-KeS is the most stable of the algorithms. From the point of view of computational complexity, CB-MUG is the slowest algorithm, whereas the CB-KeS and CR-FC are as fast in practice.

For the purpose of identifying recurrent breakpoints, we recommend CB-KeS over CB-MUG because it is faster, non-parametric and returns a list of breakpoints sorted by significance. In practice, detailed investigation can be performed on a convenient top k set of recurrent breakpoints. In contrast, CB-MUG is slow and requires the estimation of the optimal number of consensus breakpoints. Moreover, if CB-MUG is applied with m and

$m+1$ mixture components, respectively, there is no hierarchical relation between the two resulting sets of consensus breakpoints (in the sense, the latter includes the former). This means that as the number of components increases, there is no incremental improvement of the set of consensus breakpoints towards the optimal m, but rather a set of wrong models until the optimal model is attained. Therefore, finding a good estimate of the number of breakpoints appears more critical for CB-MUG than for CB-KeS.

Based solely on the study presented in this chapter, we cannot state that CB-KeS is superior to CR-FC or CR-FC is superior to CB-KeS. In Chapter 5, we show that if used together with methods for tumor classification, CR-FC slightly outperforms CB-KeS.

We used EpiExplorer to investigate several genetic and epigenetic properties of the consensus breakpoint locations, identified by the CB-KeS algorithm in the cancer datasets. We found that consensus breakpoints are enriched in functional elements such as genes, gene promoters, enhancers, CpG islands, DNA hypersensitive sites and several available histone modifications. Some of these properties have been noted previously in the literature. Therefore, our approach for identification of consensus breakpoints can help discover relevant biological properties of DNA break hotspots.

5. Methods for Identifying Relevant CNAs

> The purpose of models is not to fit
> the data but to sharpen the questions.
>
> Samuel Karlin

5.1. Introduction

The accelerated development of microarrays and, more recently, of high-throughput sequencing techniques affords genome-wide measurements of molecular changes in the cell that have an impact on cancer onset and progression. High-resolution experiments targeting gene expression, DNA copy number or DNA methylation in tumors can be the basis for discovering patterns informative of diagnosis, prognosis and therapy selection (Hicks et al., 2006; Mikeska et al., 2007; van't Veer et al., 2002). Machine learning techniques for classification and feature selection are often used for automated identification of variables associated with particular tumor phenotypes. In this Chapter, we are concerned with two widely discussed aspects of microarray classification: handling high dimensionality and ill-conditioning.

The high dimensionality of microarray-based experiments contrasting with the small number of samples easily leads to overfitting. Regularized linear models such as logistic regression with ridge (Hastie et al., 2003) or Lasso penalty (Tibshirani, 1996) are popular solutions to fitting sparse models in which only a small subset of features plays a role. More sophisticated penalties for sparse model selection are discussed by Zou and Li (2008).

The problem of ill-conditioning refers to the existence of groups of highly correlated features. The high correlations often have a biological basis, for example if the correlated features relate to the same molecular pathway (co-regulated genes in expression data), are in close proximity in the genome sequence (neighboring genes in copy number data) or share similar methylation profile (consecutive CpG dinucleotides in CpG islands). Methods using simple penalties like Lasso typically discard most of the correlated features: only one or a few arbitrary representatives from every group of correlated features enter the model, provided they are relevant for the outcome. As a consequence, the models become unstable: small changes in the training set result in dramatic changes in the selected subset of features. If the purpose of feature selection includes biological interpretation of the model, then stability must be ensured. A successful approach used in many recent articles is that of selection of groups of features. For example, the group Lasso model (Meier et al., 2008) consists of Lasso selection of predefined groups of features. The fused SVM (Rapaport et al., 2008) combines a Lasso and a fused penalty for enforcing similar weights on correlated features, this way performing group discovery and group selection simultaneously. Another approach to group selection adopted in a large class of methods

uses clustering procedures to discover feature groups, compute super-features to summarize every cluster and apply feature selection on the set of super-features. For example, in Park et al. (2007), the features are grouped with a hierarchical clustering procedure and the cluster centroids are used for training linear models. The Metagene method (Huang et al., 2003a,b) consists of k-means clustering of the features, followed by computing the principal components of the clusters, called *metagenes*, which are used for model training. Jäger and Sengupta (2003) use fuzzy clustering to determine groups of features and then select a limited number of representatives from each cluster for training SVM models. Yu et al. (2008) search for dense groups of features by kernel density estimation. The *pelora* method (Dettling, 2004) performs supervised grouping of features, by iteratively updating the groups such that the accuracy of a penalized logistic regression model is increased.

A nonparametric model often used in microarray classification is the random forest (Breiman, 2001; Díaz-Uriarte and Alvares de Andrés, 2006; Pang and Zhao, 2008). In a recent study, Strobl et al. (2008) observe that correlated variables are used interchangeably in the decision trees of the random forest models. The authors analyze the consequence of this phenomenon by simulating artificial datasets containing few correlated variables with different predictive values. They notice that the less relevant variables often replace the predictive ones (due to correlation) and thus receive undeserved, boosted importance. Strobl et al. (2008) introduce a new variable importance measure that better reflects the predictive power of each feature within a correlated group. In contrast to the study by Strobl et al. (2008), we assume that the correlated features in a group share the same predictive value (due to a common underlying biological event) and we investigate how correlation affects the feature importance given by random forest.

Another study showed that, if predictors are categorical, both feature importance measures are biased in favor of variables assuming values from larger sets of categories (Strobl et al., 2007). The authors of the article ascribe the bias to the use of bootstrap sampling and Gini split criterion for training CART trees (Breiman et al., 1984). In the literature, the bias induced by the Gini coefficient has been reported for years (Bourguignon, 1979; Pyatt et al., 1980), and it affects not only categorical variables but also grouped variables (i.e. values of the variable cluster into well separated groups – e.g. multimodal Gaussian distributions), in general. In biology, predictors often have categorical or grouped values (e.g. microarrays, sequence mutations). In the particular case of DNA copy number aberrations, the distribution of log-ratios at each locus is expected to be multimodal, corresponding to loss, normal, gain, amplification, etc.

The first contribution of this Chapter is to raise awareness of a specific effect involving feature correlation in several of the methods mentioned above that can misguide model interpretation. We observed that the Lasso penalized logistic regression, the group Lasso, the fused SVM and the random forest report feature weights which are affected by a type of bias which we call *correlation bias*. Specifically, the features which belong to larger groups of correlated features receive smaller weights, proportional to the group size, due to a shared responsibility in the model. Therefore, if the group is large enough, all features may appear irrelevant, even if they yield high correlations with the outcome. This effect is expected in the case of the sparse Lasso logistic regression, but is surprising in the case of group Lasso and fused SVM, which are specifically designed to afford selection at the group level and improved model interpretation. Moreover, such bias has not been reported for

random forest models previously. We show using simulations that correlation bias exists and affects several widely used classification models for microarray data. We also show that group selection based on consensus segmentation methods can be successfully used for removing the correlation bias. We test and compare the methods investigated on two biological datasets.

The second contribution of this chapter is an algorithm for estimating an unbiased significance of features in random forest models. The algorithm is called PIMP and is based on a permutation test involving re-shuffling of the outcome variable. We show with simulations that PIMP can help reduce the bias of random forest towards variables with more categories (if categorical) or multimodal features (if continuous). We also show how PIMP can be used for variable selection in random forest models.

5.2. Preliminaries: supervised feature selection methods

Given is a classical supervised learning problem: $(x_i, y_i), i = 1, ..., N$ are N i.i.d. observations of a p-dimensional vector $x_i \in \mathcal{R}^p$ and a binary response variable $y_i \in \{0, 1\}$. Denote by $X = (x_1, ..., x_N) \in \mathcal{R}^{N \times p}$ the input matrix and $y \in \{0, 1\}^N$ the binary outcome. The goal of supervised feature selection methods is to select a subset of features which can predict well the outcome variable, or to rank the variables with respect to their relevance for the outcome prediction. Guyon and Elisseeff (2003) present an overview of the methods for supervised feature selection, including the following: *wrapper methods*, *filter methods* and *embedded* methods, etc. Wrapper methods use prediction models as black-box tools for assessing the predictive power of subsets of features. These methods tend to be slow in practice, because many subsets of features need to be evaluated. Filter methods are based on a preprocessing step which eliminates useless features prior to model training. Popular filters are based on univariate assessment of variable relevance (such as correlation with the outcome), which can sometimes lead to the elimination of features which are useful only in combination with other features. Embedded methods perform feature selection during model training, for example by optimizing a penalized loss function. In this manuscript we will use embedded methods almost exclusively, and one filter method based on the mutual information measure of relevance. The reasons for this choice will become clear later on.

Notations: in this chapter we will use small letters to refer to samples $x_1, ..., x_N$ and capital letters to refer to features $X_1, ..., X_p$ of the input matrix X. Also, we will use the notion *feature importance* to refer to the measures of feature relevance commonly used for model interpretation, such as feature weights in linear models or variable importance in random forest.

Lasso-penalized logistic regression

Logistic regression is a popular method for classification of biological data. It models the logarithm of the posterior probabilities of the classes as linear functions of the input features. The parameters $w \in \mathcal{R}^p$ of the model are estimated by maximizing the log-likelihood $L(w; X, y)$ over the observations in the training set. Model sparsity is obtained by adding a Lasso penalty λ (see Equation 5.1), which can be optimized with cross validation. Feature importance is given by the model weights w_{LLR}:

$$w_{\mathrm{LLR}} = \arg\max_{w} L(w; X, y) - \lambda \sum_{j=1}^{p} |w_j| \tag{5.1}$$

In this chapter, we will call this model Lasso Logistic Regression (LLR). In our experiments, we used the **R** package *glmnet* (Friedman et al., 2010) for training LLR models.

Group Lasso

(Meier et al., 2008) uses the logistic regression model with a more specialized penalty, which takes into account some natural grouping of the features. Assume there are G groups of predictors and each group must in its entirely be included in the model by receiving non-zero weights, or be discarded as irrelevant. The group penalty is a combination of a Lasso penalty acting at the group level and a ridge penalty on the predictors within each group. If I_g is the index of the features belonging to group g, then the weights of the logistic group Lasso (GL) model are given by:

$$w_{\mathrm{GL}} = \arg\max_{w} L(w; X, y) - \lambda \sum_{g=1}^{G} \|w_{I_g}\|_2 \tag{5.2}$$

We use the **R** package *grplasso* (by Lukas Meier) for training GL models and cross validation for estimating the optimum penalty λ.

Fused Lasso Support Vector Machines

The *fused SVM* (Rapaport et al., 2008) has been proposed for the special case that the features can be ordered such that neighboring features are expected to be correlated. This is the case with data on copy number aberrations, where the features are genomic sites ordered by position in the genome. Fused SVM (FSVM) is a linear SVM model with two supplementary penalties: a Lasso penalty for model sparsity and a fused penalty, which acts as a smoother of the weights, in such a way that weights of neighboring features are forced to be similar. The weights of the model w_{FSVM} are obtained by minimizing a penalized hinge loss, as follows (see Rapaport et al. (2008) for details):

$$w_{\mathrm{FSVM}} = \arg\min_{w} \sum_{i=1}^{N} [1 - y_i w^T x_i]_+ + \lambda \sum_{i=1}^{p} |w_i| + \mu \sum_{i=2}^{p} |w_i - w_{i-1}| \tag{5.3}$$

The optimization problem given by System 5.3 can be solved by a linear program. We implemented this method using Matlab and the CVX optimization toolbox (Grant and S., 2008). Cross-validation for both penalty parameters λ and μ is necessary, which makes fitting an FSVM model slower than fitting the other methods.

Random forest

Random forest (RF) models (Breiman, 2001) are nonparametric and non-linear models, attractive due to their interpretability. They use bagging (bootstrap aggregating) of decision trees in order to reduce variance of single trees and thus improve prediction accuracy. Typically, a collection of T decision trees using CART methodology Breiman et al. (1984)

are trained on T bootstrap samples of the data, respectively. At each node of each tree, a random subset of fixed size is selected from the features and the one yielding the maximum decrease in Gini index is chosen for the split. The trees are fully grown and left unpruned. The class of a new sample is determined by the majority of the votes of all trees in the random forest. This aggregate model has lower variance and is less susceptible to overfitting than a single decision tree. The test error of random forest models is estimated on the out-of-bag (OOB) data, as follows: after each tree has been grown, the inputs that did not participate in the training bootstrap sample are used as test set, then averaging over all trees gives the test error estimate.

Breiman (2001) proposes two measures for feature importance, the *Variable Importance* (VI) and the *Gini Importance* (GI). The VI of a feature is computed as the average decrease in model accuracy on the OOB samples when the values of the respective feature are randomly permuted. The GI uses the decrease of Gini index (impurity) after a node split as a measure of feature relevance. In general, the larger the decrease of impurity after a certain split, the more informative the corresponding input variable. The average decrease in Gini index over all trees in the random forest defines the GI. It should be observed that the Gini index is closely related to the entropy, both being measures of impurity. In this manuscript, we will analyze mainly the GI measure. The VI was shown to be highly correlated with the GI (Strobl et al., 2007).

We use the **R** package *randomForest* (Liaw and Wiener, 2002) for training RF models. There are two parameters that influence the performance of RF: the number *ntree* of trees in the collection and the number *mtry* of variables considered for each tree split. In our experiments we use the recommended value $mtry = \sqrt{number\ of\ features}$ and we select the optimal value for *ntree* via cross validation. Díaz-Uriarte and Alvares de Andrés (2006) evaluate the performance of RF models for various parameter settings in ten real-world learning instances. Their results suggest that the default value of *mtry* affords either optimal or close to optimal performance.

Mutual Information

Mutual information (MI) originates from information theory and measures how much a random variable X is informative about another random variable Y. It is closely related to the concept of entropy. The entropy of a random variable X, denoted traditionally by $H(X)$, measures the level of uncertainty in variable X. It is computed as:

$$H(X) = -\sum_{x} P_X(x) \log P_X(x) \tag{5.4}$$

where $P_X(x)$ is the probability distribution of X. The conditional entropy $H(X|Y)$ measures the average of the uncertainty in X given the observed variable Y. Then the mutual information $\mathrm{MI}(X, Y)$ is defined as the decrease in uncertainty about X after observing Y:

$$\mathrm{MI}(X, Y) = H(X) - H(X|Y) \tag{5.5}$$

Low, close to zero, MI means that the variables are close to independent. The larger the MI, the larger the reduction of uncertainty in X when Y is known. Mutual information is often used for a quick search of relevant features, when training statistical learning models

requires too much computational effort due to the large number of features, e.g. in the case of artificial neural networks (Battiti, 1994). Typically, the MI between each feature and the outcome is computed and a ranking of the inputs results.

For estimating the MI of two vectors, we use the following formula, which is an immediate equivalent transformation of Equation 5.5:

$$\mathrm{MI}(X,Y) = H(X) + H(Y) - H(X,Y) \tag{5.6}$$

Since the probability distributions of X and Y are unknown, in general, we compute frequency-based estimators (Guyon and Elisseeff, 2003).

5.3. Correlation bias and correction methods

The research presented in this section has been published in Toloşi and Lengauer (2011).

For classification of high-dimensional data containing (large) groups of correlated features, the requirements of model sparsity and of retrieving of all predictive features are in direct competition. In applications in which assessment of feature importance is the main objective, models that give priority to the latter requirement should be preferred. In what follows, we formulate three key properties that we believe a classification model should meet in order to be a good instrument for assessment of feature importance.

Assume that two independent biological events P_1 and P_2 (e.g. the deletion of a chromosome arm can be an event) influence the binary phenotype Y (e.g. tumor stage) and let us denote the magnitude of their effects on Y with $E(P_1)$ and $E(P_2)$, respectively and assume that $E(P_1) > E(P_2)$. Assume that, by means of an experimental technology, variables associated with each of the two events are measured (e.g. all genes located within a deleted chromosome arm). Let us denote with $U_1, ..., U_q$ the variables associated with P_1 and with $V_1, ..., V_p$, the variables associated with P_2, $p, q \geq 1$. Consequently, $\{U_i\}_{1 \leq i \leq q}$ and $\{V_j\}_{1 \leq j \leq p}$ form two groups of correlated variables. Assume a classification model $\mathcal{M}$ is used to predict Y from a set of N observations on features $U_1, ..., U_q, V_1, ..., V_p$ and this model assigns importance values to features: $w_1^1, ..., w_q^1, w_1^2, ..., w_p^2$. Without losing generality, assume all the importance values are positive and a larger value indicates a more predictive feature. The following three properties should hold:

1. The importance values of the correlated features are similar: $w_1^1 \approx w_2^1 \approx ... \approx w_q^1$ and $w_1^2 \approx w_2^2 \approx ... \approx w_p^2$.

2. The importance of the variables reflect the magnitude of the effect of the corresponding process on the outcome: $w_i^1 \geq w_j^2, \forall i = 1..q, \forall j = 1..p$.

3. The importance of the variables $\{w_i^1\}_{1 \leq i \leq q}$ and $\{w_i^2\}_{1 \leq i \leq p}$ does not depend on the corresponding group sizes, namely q and p, respectively.

We require that property 1) holds because, in absence of a true model, it is wise to give fair chances to all correlated variables for being considered as causative for the phenotype. In this case, supplementary evidence from other sources should be used for identifying the causative variable from a correlated group. Property 2) is based on the assumption that $E(P_1) > E(P_2)$ and hence any of the features $\{U_i\}_{i \leq q}$ contributes more to the outcome

than any of the features $\{V_j\}_{j\leq p}$. Thus, the property ensures a fair ranking of the variables, which is important in applications because often only a few top ranking groups are considered for further investigation. Property 3) demands that the importance of the features does not change as more evidence (more variables) about the corresponding events is added to the data.

In this chapter we show that in classification problems with groups of correlated features, the assignment of feature importance by LLR, RF, GL and FSVM does not meet requirements 2) and 3). Specifically, the reported feature importance varies with the sizes of the correlated groups of features and results in biased feature ranking. In the context of the example above, the feature weights $\{w_i^1\}_{1\leq i\leq q}$ and $\{w_i^2\}_{1\leq i\leq p}$ depend on the values of q and p, respectively, in a way that larger group size leads to smaller importance values. As a consequence, if q is much larger than p, variables $\{U_i\}_{1\leq i\leq q}$ can falsely appear less predictive than variables $\{V_j\}_{1\leq j\leq p}$ and P_2 is considered more relevant than P_1.

In sparse models like LLR, correlated features are generally discarded in favor of a single representative. Instability of feature importance is a known issue in such models (Park et al., 2007; Jäger and Sengupta, 2003), and it is easy to observe that the larger the group, the smaller the chance of each particular variable within the group is to be selected by the model. Therefore, under repeated perturbations of the training set, the average weights of the features decrease as the size of the group increases. In the case of FSVM, the weights of correlated features are forced to be equal (or similar). Consequently, if the group of correlated features becomes larger, the common weights need to be decreased, in order to accommodate all features in the model and not violate the Lasso penalty. This rescaling of the weights is possible without decreasing the accuracy of the model, since correlated features provide only redundant information. In the Section 5.3.1 we show how the interaction between the two penalties of FSVM can cause correlation bias. A similar effect can be observed in GL models. In RF, the correlation bias is caused by the bootstrap sampling of the observations and by the sampling of the features at each node of the trees, which causes correlated features to be used interchangeably in the tree components.

In this chapter, we say that models that do not meet requirements 2) and 3) are affected by *correlation bias*.

5.3.1. Example of correlation bias

In what follows, we demonstrate the phenomenon of correlation bias on a simplified version of FSVM. For convenience, we will consider a simple linear regression with Lasso and fused penalties, similar to model (5.3). Assume $U \in \mathcal{R}^p$ and $V \in \mathcal{R}^p$ are two independent standardized random variables, $\mathrm{E}(U) = \mathrm{E}(V) = 0$, $\mathrm{Var}(U) = \mathrm{Var}(V) = 1$. Consider now the variable Y given by the linear model:

$$Y = aU + bV + \epsilon, \quad a, b \in \mathcal{R}, \quad \epsilon \sim N(0, \sigma) \tag{5.7}$$

Consider variables $U_1, ..., U_q$, for some integer $q \geq 1$, mutually correlated and with U. Assume they are also standardized. Formally:

$$\begin{gathered} \mathrm{E}(U_i) = \mathrm{E}(U) = 0, \quad \mathrm{Var}(U_i) = \mathrm{Var}(U) = 1, \quad \forall i \in \{1, ..., q\} \\ \mathrm{Cor}(U, U_i) \approx 1, \forall i \in \{1, ..., q\}, \quad \mathrm{Cor}(U_i, U_j) \approx 1, \forall i, j \in \{1, ..., q\} \end{gathered} \tag{5.8}$$

Similarly, let $V_1, ..., V_r$ be r ($r \geq 1$) standardized variables, mutually correlated and with V:

$$\begin{array}{c} \mathrm{E}(V_i) = \mathrm{E}(V) = 0, \quad \mathrm{Var}(V_i) = \mathrm{Var}(V) = 1, \quad \forall i \in \{1, ..., r\} \\ \mathrm{Cor}(V, V_i) \approx 1, \forall i \in \{1, ..., r\}, \quad \mathrm{Cor}(V_i, V_j) \approx 1, \forall i, j \in \{1, ..., r\} \end{array} \tag{5.9}$$

Assume that any pair U_i, V_j is not correlated, i.e.

$$\mathrm{Cor}(U_i, V_j) \approx 0, \quad \forall i \in \{1, ..., q\}, \quad \forall j \in \{1, ..., r\}$$

We formulate a regression problem with two groups of correlated variables, $G_U = \{U_1, ..., U_q\}$ and $G_V = \{V_1, ..., V_r\}$ and response Y. Linear regression with fused and Lasso penalties for this particular problem minimizes the following expected penalized residual sum of squares (PRSS):

$$\begin{array}{rcl} w_{opt} & = & \arg\min_w \mathrm{PRSS}(w; \lambda, \mu), \\ \mathrm{PRSS}(w; \lambda, \mu) & = & \mathrm{E}\left((Y - \sum_{i=1}^{q} w_i U_i - \sum_{i=1}^{r} w_{q+i} V_i)^2 \right) + \\ & & +\lambda \sum_{i=1}^{q+r} |w_i| + \mu \sum_{i=2}^{q+r} |w_i - w_{i-1}| \end{array} \tag{5.10}$$

and $\lambda > 0, \mu > 0$. Solving the fused regression given by Equation 5.10 will ideally yield optimum weights $\tilde{w}$ of the form $\tilde{w}_1 = ... = \tilde{w}_q = \alpha$ and $\tilde{w}_{q+1} = ... = \tilde{w}_{q+r} = \beta$. This corresponds to a model that correctly identifies two independent groups of features and assigns identical weights to correlated features, which provide same information to the model. Assume that for some $\lambda > 0$ and $\mu > 0$, the model has the ideal form specified above. Below we perform successive transformations on Equation 5.10:

Denote with ERSS the expected residual sum of squares. Then

$$\mathrm{ERSS}(w) = \mathrm{E}\left((Y - \sum_{i=1}^{q} w_i U_i - \sum_{i=1}^{r} w_{q+i} V_i)^2 \right)$$

Then it follows, under the model assumptions given by Equation 5.7:

$$\begin{array}{rcl} \mathrm{ERSS}(\tilde{w}) & = & \mathrm{E}\left(\left(aU + bV + \epsilon - \sum_{i=1}^{q} \tilde{w}_i U_i - \sum_{i=1}^{r} \tilde{w}_{q+i} V_i \right)^2 \right) \\ & = & \mathrm{E}\left(\left(\underbrace{(aU - \alpha \sum_{i=1}^{q} U_i)}_{A} + \underbrace{(bV - \beta \sum_{i=1}^{r} V_i)}_{B} + \epsilon \right)^2 \right) \\ & = & \mathrm{E}(A^2) + \mathrm{E}(B^2) + 2\mathrm{E}(AB) \end{array}$$

We further transform the ERSS, using the assumptions (5.8) and (5.9) and the fact that the correlation and covariance notions are equivalent when applied to standardized variables:

$$\begin{aligned}
\mathrm{E}(A^2) &= \mathrm{E}\left(\left(\sum_{i=1}^{q}(\frac{a}{q}U-\alpha U_i)\right)^2\right) \\
&= \sum_{1\le i,j\le q}\mathrm{E}\left((\frac{a}{q}U-\alpha U_i)(\frac{a}{q}U-\alpha U_j)\right) \\
&\approx q^2(\frac{a}{q}-\alpha)^2
\end{aligned}$$

$$\mathrm{E}(B^2) \approx r^2(\frac{b}{r}-\beta)^2 \qquad \text{(as above)}$$

$$\begin{aligned}
\mathrm{E}(AB) &= \mathrm{E}(abUV - a\beta\sum_{i=1}^{r}UV_i - \alpha b\sum_{i=1}^{q}U_iV - \alpha\beta\sum_{i=1}^{q}\sum_{j=1}^{r}U_iV_j) \\
&= ab\mathrm{E}(UV) - a\beta\sum_{i=1}^{r}\mathrm{E}(UV_i) - \alpha b\sum_{i=1}^{q}\mathrm{E}(U_iV) - \alpha\beta\sum_{i=1}^{q}\sum_{j=1}^{r}\mathrm{E}(U_iV_j) \\
&= ab\mathrm{Cov}(UV) - a\beta\sum_{i=1}^{r}\mathrm{Cov}(UV_i) - \alpha b\sum_{i=1}^{q}\mathrm{Cov}(U_iV) - \alpha\beta\sum_{i=1}^{q}\sum_{j=1}^{r}\mathrm{Cov}(U_iV_j) \\
&= 0.
\end{aligned}$$

This gives the final formula:

$$\mathrm{ERSS}(\tilde{w}) \approx q^2(\frac{a}{q}-\alpha)^2 + r^2(\frac{b}{r}-\beta)^2$$

Evidently, the penalty terms can be written as

$$\lambda\sum_{i=1}^{q+r}|\tilde{w}_i| = \lambda(q|\alpha| + r|\beta|)$$

$$\mu\sum_{i=2}^{q+r}|\tilde{w}_i - \tilde{w}_{i-1}| = \mu|\alpha-\beta|$$

Therefore, the optimization target can be re-written as:

$$\mathrm{PRSS}(\tilde{w};\lambda,\mu) \approx q^2(\frac{a}{q}-\alpha)^2 + r^2(\frac{b}{r}-\beta)^2 + \lambda(q|\alpha| + r|\beta|) + \mu|\alpha-\beta| \tag{5.11}$$

Consider two weight vectors w' and w''

$$w'_1 = ... = w'_q = a, \quad w'_{q+1} = ... = w'_{q+r} = b$$

$$w''_1 = ... = w''_q = \frac{a}{q}, \quad w''_{q+1} = ... = w''_{q+r} = \frac{b}{r}$$

Then:

$$\mathrm{PRSS}(w';\lambda,\mu) = q^2(\frac{a}{q}-a)^2 + r^2(\frac{b}{r}-b)^2 + \lambda(q|a| + r|b|) + \mu|a-b| \quad \text{and}$$

$$\mathrm{PRSS}(w'';\lambda,\mu) = \lambda(|a| + |b|) + \mu\left|\frac{a}{q} - \frac{b}{r}\right|$$

It follows that if a, b, q, r are chosen such that

$$a > b > 0, \quad \frac{a}{q} < \frac{b}{r} \quad \text{and} \quad \left|\frac{a}{q} - \frac{b}{r}\right| < |a - b| \tag{5.12}$$

then $\text{PRSS}(w'; \lambda, \mu) > \text{PRSS}(w''; \lambda, \mu)$, for any arbitrary fixed $\lambda > 0$ and $\mu > 0$. The set of conditions (5.12) are feasible and yield an infinite number of solutions: for any positive $b \in \mathcal{R}$ and positive integer r, any choice of $q > r$ and $a \in \left[\frac{q(r+1)}{r(q+1)}b, \frac{q}{r}b\right]$ satisfy all conditions. In conclusion, w'' yields a smaller value of the objective function and therefore w' can never be a solution of the optimization problem (5.10). In contrast, the objective function is improved by shrinking the weights proportionally to the group size, which corresponds to reversed ranking of the two groups of features, if the conditions (5.12) hold. We use simulation experiments to prove that under similar conditions, the weights given by FSVM are indeed such that the relative importance of the two groups of correlated features is incorrect.

5.3.2. Methods for reducing correlation bias based on feature grouping

Intuitively, a good strategy for reducing the correlation bias is to group the correlated features prior to model fitting and derive corresponding *feature representatives* as a summary of each group. The importance of the original features can be defined as the importance of the corresponding representatives. To this end, the consensus segmentation methods proposed in Chapter 4 can be used. In a cross-validation framework, the methods we propose follow the following steps:

1. Assign the samples randomly to 10 cross-validation bins.
2. For i from 1 to 10, repeat:
   a) Use as training set all samples that are not in the i^{th} bin.
   b) Perform consensus segmentation on the training set and obtain a set of feature representatives.
   c) Fit a model (LLR or RF) on the training set using the feature representatives.
   d) Test the model on the remaining samples.
3. Based on the CV data, perform model selection by estimating optimal Lasso penalty for LLR or optimal number of trees for RF. Optionally, the optimal number of regions for consensus segmentation can be also estimated. Report test accuracy of the optimal model.
4. With the optimal parameters, fit a model on all samples and report feature importance (model weigths for LLR, variable importance for RF).

We show in this chapter that the procedure above can help eliminate correlation bias. The key step is feature grouping with consensus segmentation, which can be carried out by either of the three methods proposed in Chapter 4: CB-MUG, CB-KeS, CR-FC. In order to demonstrate that the correlation bias has been removed, in this chapter we only used

CR-FC, which is the fastest method in practice. In Chapter 6, we used and compared all three algorithms.

Through the rest of this thesis, we will use the following nomenclature for our models:

Model - CS Method - CS Selection Method,

where **Model** can be either 'LLR' or 'RF', **CS Method** is one of the consensus segmentation methods 'MUG', 'KeS' or 'FC' and **CS Selection Method** can be either 'Sup' or 'Unsup', depending on whether the number of regions of consensus segmentation is selected by cross validation or unsupervised, by $\Omega_{0.98}$ criterion (see Chapter 4).

5.3.3. Model evaluation

We compare the performance of classification methods on training sets with (large) groups of correlated features. In particular, we analyze the measures of feature relevance provided by the models investigated and seek for evidence of correlation bias. Additionally, we report and discuss the prediction accuracy of the models (estimated via 10-fold cross validation) and the stability of the respective feature importance measures.

The stability of feature importance is defined as the variability of feature weights under perturbations of the training set. When the goal of classification is to select the most relevant features, small modifications in the training set should not lead to considerable changes in the set of important covariates. When the true distribution of the training set is not known, stability can be inferred via repeated sampling from the available training observations. In Kalousis et al. (2006) classical 10-fold cross validation is used in order to create 10 overlapping training sets and model stability is estimated by comparing the 10 resulting feature weightings. For this purpose, the authors propose several measures of similarity between two vectors of feature weights. For our purposes, the Pearson correlation coefficient is most suitable. Overall model stability is given by the average of all pair-wise Pearson correlations between feature weight vectors provided by the models fitted on the 10 variations of the training set. The stability score has a value between -1 and $+1$, with higher values for more stable models.

5.3.4. Validation data sets

We introduce below two artificial datasets which we use for demostrating the correlation bias. In both settings, we construct groups of correlated features of different cardinalities and compare their importance with respect to the classification models introduced in this chapter. For each dataset we constructed a binary classification instance based on an underlying linear model with noise. The datasets differ by complexity, in the sense that the second dataset has significantly more features and groups of features.

	G_1	G_2	R	y
G_1	0.86	0	-0.02	0.63
G_2	0	0.86	0	0.42
R	-0.02	0	0.87	0

Table 5.1.: Simulation A. Pairwise Pearson correlation between features, averaged within and between groups. The last column shows the average correlation between each feature group and the outcome y.

Simulated data

Simulation A. We generated datasets with $N = 100$ samples and $p = 250$ features. The features are divided into three groups: G_1, G_2 and R. Group R has 50 features and the cardinality of group G_2 is $200 - |G_1|$, for different values of $|G_1| \in \{100, 120, 140, 160, 180\}$. The features in each group are mutually correlated and any two features belonging to different groups are independent. The features in group G_1 are generated from the prototype vector U, which is sampled from a mixture of two Gaussians with equal probabilities: $g_0 = \mathcal{N}(0, 0.2)$ and $g_1 = \mathcal{N}(1, 0.3)$. The particular choice for a Gaussian mixture stems from gene copy number data, where g_0 corresponds to those samples with normal copy number and g_1 indicates aberrations (copy number gains, in this case). We generate features $U_1, ..., U_{|G_1|}$ with the following procedure: randomly select 20% of the components of U and alter them by adding Gaussian noise $\mathcal{N}(0, 0.5)$, then repeat $|G_1|$ times. The features generated this way are correlated with U and with each other and resemble segmented copy number data, which are piecewise constant with occasional changes. Using the same algorithm we generate a prototype vector V independent from U and corresponding features $V_1, ..., V_{|G_2|}$, which form group G_2, and then repeat the procedure to simulate group R. Last, we generate a binary outcome y with the following linear classification rule:

$$y = \begin{cases} 1, & \text{if } 5U + 4V - (\overline{5U + 4V}) + \varepsilon > 0, \\ 0, & \text{otherwise.} \end{cases}$$

where $\varepsilon \sim \mathcal{N}(0, 0.1)$ and $\overline{5U + 4V}$ denotes the average of $5U + 4V$. From the simulation parameters, it follows that features in group G_1 are most relevant to the outcome (being correlated to U), features in group G_2 are less predictive and features in group R are irrelevant. Table 5.1 shows the within group and between group average correlations, summarized over 100 simulated datasets.

Simulation B. We generated more complex artificial datasets by considering 10 groups of predictive features $G_1, ..., G_{10}$ and 20 groups of irrelevant features, $R_1, ..., R_{20}$. The number of samples is $N = 100$. Each of the the groups G_2 to G_{10} and R_1 to R_{20} contains 10 correlated features and the cardinality of group G_1 takes, in turn, one of the values $\{10, 50, 100, 200\}$. The groups of correlated variables $G_1, ..., G_{10}, R_1, ..., R_{20}$ are generated from the prototype variables $U_1, ..., U_{10}, V_1, ..., V_{20}$, respectively. The simulation procedure is similar to that of Simulation A, with dif-

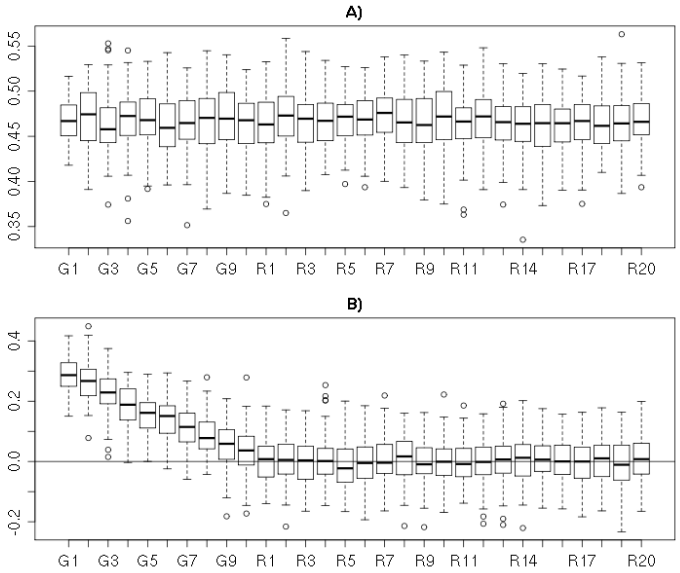


Figure 5.1.: A) Boxplot showing the average correlations between features within each group in Simulation B, summarized over 100 simulations. B) Boxplot showing the average correlation of features within each group with the outcome in Simulation B, summarized over 100 simulations.

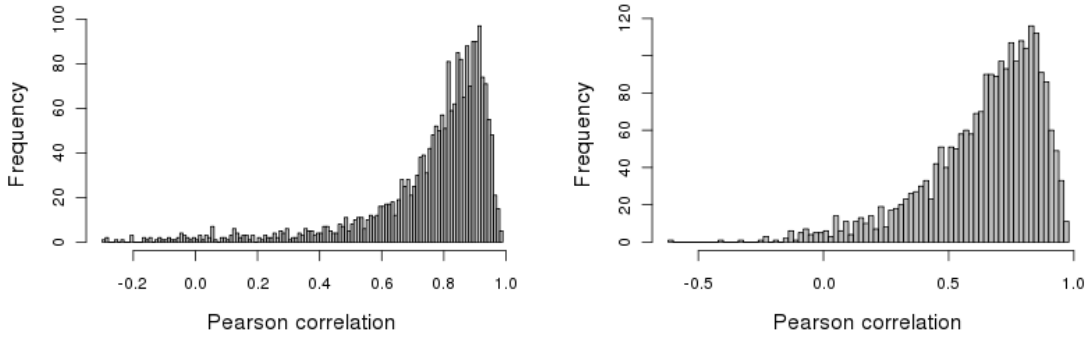
ferent parameters: we alter *all* components of the corresponding prototype vector by adding Gaussian noise $\mathcal{N}(0, 0.6)$. The binary outcome y is given by the linear rule:

$$y = \begin{cases} 1, & \text{if } 10U_1 + 9U_2 + ... + 1U_{10} - (\overline{10U_1 + 9U_2 + ... + 1U_{10}}) + \varepsilon > 0, \\ 0, & \text{otherwise.} \end{cases}$$

The groups G_1 to G_{10} are ordered decreasingly by their relevance to the outcome. In Simulation B, the within-group correlations are smaller than in Simulation A and the number of features is larger, which makes the identification of the groups of features more difficult. In Figure 5.1, we show the average correlations between features and the average correlations between features and the outcome variable, for each group, summarized over 100 simulations. The correlation between pairs of features within each group is positive and large, with mean value 0.47 (Figure 5.1A). The average correlation between the outcome variable and the features belonging to each group decreases, as the index of the group increases. This means that in a univariate sense, features from group G_1 are expected to be most informative of the outcome, followed by features from G_2, etc.

Real data

Bladder tumors. We tested our methods on a set of 98 CGH arrays measuring copy number aberrations in bladder tumors. The experimental settings and data have been described in Blaveri et al. (2005). DNA copy number has been measured for 2142 probes distributed over all autosomes. The correlation between adjacent



(a) (b)

Figure 5.2.: Histograms show high correlations between neighboring probes (w.r.t. genomic position) in the a) bladder dataset and b) breast dataset.

probes is very high (median 0.82), see Figure 5.2a. We considered two binary classification problems, by tumor grade and by tumor stage. For grade classification we used 19 samples with low grade (Grade 1) and 77 samples with high grade (Grade 2 or 3). For stage classification, 84 samples were grouped into two classes: stage Ta (29 samples) and stage T2+ (55 samples). We excluded the intermediary stage T1 (as in Rapaport et al. (2008)). For each classification scenario, we train RF, LLR, FSVM and GL models. We also fit RF and LLR in combination with FC and FC-Sup. For each model we report accuracy (using 10-fold cross validation), area under the curve (AUC) and feature importance.

Breast tumors. In Climent et al. (2007), 185 early-stage breast tumors were analyzed using arrayCGH technology (UCSF Hum Array 2.0). Copy number aberrations are measured for 2369 BAC probes (chromosomes X and Y excluded). High correlations between neighboring probes are observed, with median value 0.69 (see Figure 5.2b). The authors of the study use statistical tests and report significant associations between certain genetic alterations and ER status (oestrogen receptor) and PR status (progesterone receptor) of the tumors. Using the methodology introduced here, we identify genetic lesions which help discriminate between ER positive and ER negative tumors, and PR positive and PR negative tumors, respectively. In the cohort, there are 60 ER negative and 101 ER positive tumors, and 65 PR negative and 96 PR positive tumors. For all models considered, we report classification accuracy and AUC (using 10-fold cross validation) and feature importance.

5.4. Bias of the Gini Importance measure with random forest and correction

The research presented in this section is the result of a collaboration with André Altmann and was published in Altmann et al. (2010).

With respect to random forest classifiers, Strobl et al. (2007) showed that when predictors are categorical, both GI nand VI are biased in favor of variables taking more categories. The bias is caused by the Gini split criterion for training CART trees (Breiman et al., 1984). More generally, the Gini coefficient itself is biased in the same way (Bourguignon, 1979; Pyatt et al., 1980), and the bias affects not only categorical variables but also grouped variables (for example variables with values following some multimodal Gaussian distributions).

Here we introduce a heuristic for correcting biased measures of feature importance, called *permutation importance* (PIMP). The method normalizes the biased measure based on a permutation test and returns significance p-values for each feature. In order to preserve the relations between features, we use permutations of the outcome. We show that this method can be used to correct for the bias of feature importance computed with random forest. Moreover, our method can be used together with any learning method that assesses feature relevance, providing significance p-values for each predictor variable. In the particular case of random forest, our method is very useful because it can provide with an automated model selection procedure. Indeed, despite their interpretability, random forest can only provide with a ranking of the features, but not with a significance cutoff for the ranks. Our method assignes p-values to the features and thus allows for a straight-forward significance assessment.

5.4.1. Permutation Importance

The permutation importance (PIMP) is a heuristic for correcting for the bias of the GI of random forest models.

In a general setting, assume given an algorithm that assesses the relevance of a set of features with respect to a response vector. The PIMP algorithm permutes the response vector s times. For each permutation of the response vector the variable importance for all predictor variables is assessed. This leads to a vector of s importance measures for every variable, which we call the *null importances*. The PIMP algorithm fits a probabilty distribution to the population of null importances, which the user can choose from the following: Gaussian, lognormal, or gamma. Maximum likelihood estimators of the parameters of the selected distribution are computed. Given the fitted distribution, the probability of observing a variable importance of v or higher using the true response vector, can be computed (PIMP p-value). If the user does not know which distribution is most suitable for his or her problem, the PIMP algorithm uses Kolmogorov-Smirnov (KS) tests in order to automatically identify the most appropriate distribution. However, if the tests show little resemblence to any of the three proposed distributions, a non-parametric estimation of the PIMP p-values is used, simply by determining the fraction of null importances that are more extreme than the true importance v (see Algorithm 3 for an illustration of the PIMP method with Gaussian distribution).

In practical applications the variance of the null importances may be very small and therefore small deviations from the mean lead to artificially boosted variable importances. In order to prevent this artifact, we apply a simple heuristic: variances

that are smaller than the mean variance of all variable importances are set to the mean variance.

Permuting the response vector has several advantages. First, the dependence between predictor variables remains unchanged. Second, the number of permutations (s) can be much smaller than the number of predictor variables (p). Third, the approach is general, it can be used together with any method that generates measures for variable importance (biased or unbiased). In this study we demonstrate that PIMP is effective if used with the GI of random forest.

Algorithm 3 Permutation importance (PIMP)

Require: $P \in \mathbf{R}^{n \times p}$ (matrix of predictors), p (the number of features), l (response vector), $VarImp$ (function to calculate variable importance), s (number of permutations)

Ensure: β, a vector of p-values corresponding to the features

1. $\overrightarrow{\alpha} = VarImp(l, P), \quad R \in \mathbf{R}^{s \times p}$
2. for $(i = 1; i \leq s; ++\, i)$
   $l' = permute(l)$
   $R_{i,*} = VarImp(l', P)$
3. $\overrightarrow{\mu} \in \mathbf{R}^p, \overrightarrow{\sigma} \in \mathbf{R}^p$
4. for $(j = 1; j \leq p; ++\, j)$
   $\mu_j = mean(R_{*,j})$
   $\sigma_j = sd(R_{*,j})$
5. $\sigma_\mu = mean(\overrightarrow{\sigma}), \ \overrightarrow{\beta} \in \mathbf{R}^p$
6. for $(j = 1; j \leq p; ++\, j)\{$
   $\sigma' = \max\{\sigma_j, \sigma_\mu\}$
   $\beta_j = pnorm(\alpha_j, \mu_j, \sigma')$
7. return$(\overrightarrow{\beta})$

Note: the notation $R_{i,*}$ and $R_{*,j}$ refers to the i^{th} row and j^{th} column of the matrix R, respectively. The variable p denotes the number of different features (the number of columns of matrix P). The function *pnorm* refers to the **R** function that computes the probability of observing an importance of α_j or larger given a Gaussian distribution with mean μ_j and standard deviation σ'.

5.4.2. Corrected RandomForest models

The CART methodology uses the Gini index as a criterion for choosing best splits during tree construction and thus the resulting model incorporates the bias of this measure. As a consequence, both the CART and the random forest models are biased themselves, not only their derived feature importance measures. Here, we propose a method for improving the random forest models that uses the PIMP algorithm. The method has the following steps: 1) training a classical random forest model on the training data; 2) computing the PIMP scores of the covariates; 3) training a new model with the classical random forest but now using only the significant variables (w.r.t. PIMP scores), by applying for example the classical 0.05 significance threshold. We will call the improved model *PIMP-RandomForest*.

The idea of using the most predictive features for re-training random forest model in order to reduce variance and improve accuracy has been proposed previously. For instance, Díaz-Uriarte and Alvares de Andrés (2006) investigate its benefits on several real-world datasets. The authors show that in some of the instances, the procedure gives good results. However, it may also occur that the random forest models built on the reduced set of features exhibit a slightly decreased performance compared to full random forest model.

In order to assess the improvement in prediction accuracy of the PIMP-RandomForest model, we use an independent test set and we compute the corresponding error rates. Since the PIMP-RandomForest model uses fewer features than the initial random forest, an increase in accuracy can be solely due to decrease of model variance. Thus, we also compare PIMP-RandomForest with classical random forest models trained using only the top ranking features of the initial method (biased) as well as with the (corrected) cforest model proposed by Strobl et al. (2007) on all features.

5.4.3. Validation data sets

Simulated data

Simulation C. For demonstrating the degree of bias in the established measures of importance a dataset comprising 1000 instances was simulated. The predictor variables consist of 31 categorical variables with 2 to 32 categories. We chose categorical variables for simplicity. A similar scenario with continuous variables could be realised, for example by drawing from multimodal distributions with k modes, $2 \leq k \leq 32$. The response is a binary variable. Predictor variables and response were independently sampled from a uniform distribution. Since input and output were randomly generated, no predictor variable is informative. Given an unbiased measure of variable importance all variables should receive equally low values. For verification, the GI was computed for each variable. Then, the PIMP of all measures was computed using $s = 100$. The simulation was repeated 100 times.

Simulation D. The second simulation was targeted at the question of how efficiently predictive variables can be recovered among a large set of non-predictive variables. We generated an artificial dataset with a large number of predictors (p) and a small number of samples (n), with $p = 500$ and $n = 100$. The variables had 1 to 21 categories. The number of categories for every variable was randomly determined, and variables with few categories were more likely than positions with many amino acids. Precisely, a variable with m categories had likelihood $1/m \cdot C^{-1}$, with $C = \sum_{i=1}^{21} 1/i$. Moreover, for every variable the categories were not equally likely, but were sampled from a randomly generated distribution as follows: for each category $j \in \{1, ..., m\}$ of a variable, an integer x_j between 1 and 100 was uniformly sampled. Then the probability of category j of that variable was set to $x_j / \sum_{k=1}^{m} x_k$. The output vector comprises two classes that are randomly sampled with probability 0.5. In order to challenge the ability of the feature importance methods to discover the relevant covariates, a number of relevant variables with a small number of cat-

egories were intermixed among the non-informative positions as follows: the first 12 variables comprised the same two categories and were conditionally dependent (to different degrees) on the binary response variable. Precisely, if the outcome was positive (negative) the category "a" was sampled with probability $0.5 + r$ $(0.5 - r)$ and category "b" was sampled with probability of $0.5 - r$ $(0.5 + r)$, where r varied from 0.24 to 0.02 in steps of 0.02. Apart from the first 12 variables all variables were ordered increasingly with respect to the number of corresponding categories. GI and PIMP scores with $s \in \{10, 50, 100, 500, 1000\}$ were applied for generating feature rankings. An optimal feature ranking method would rediscover all 12 variables that were associated with the outcome. However, since the relation of some variables with the outcome was very weak, these variables were likely to be ranked too low. The simulation was repeated 100 times.

5.5. Results I - Correlation bias

5.5.1. Simulated data

Simulation A.

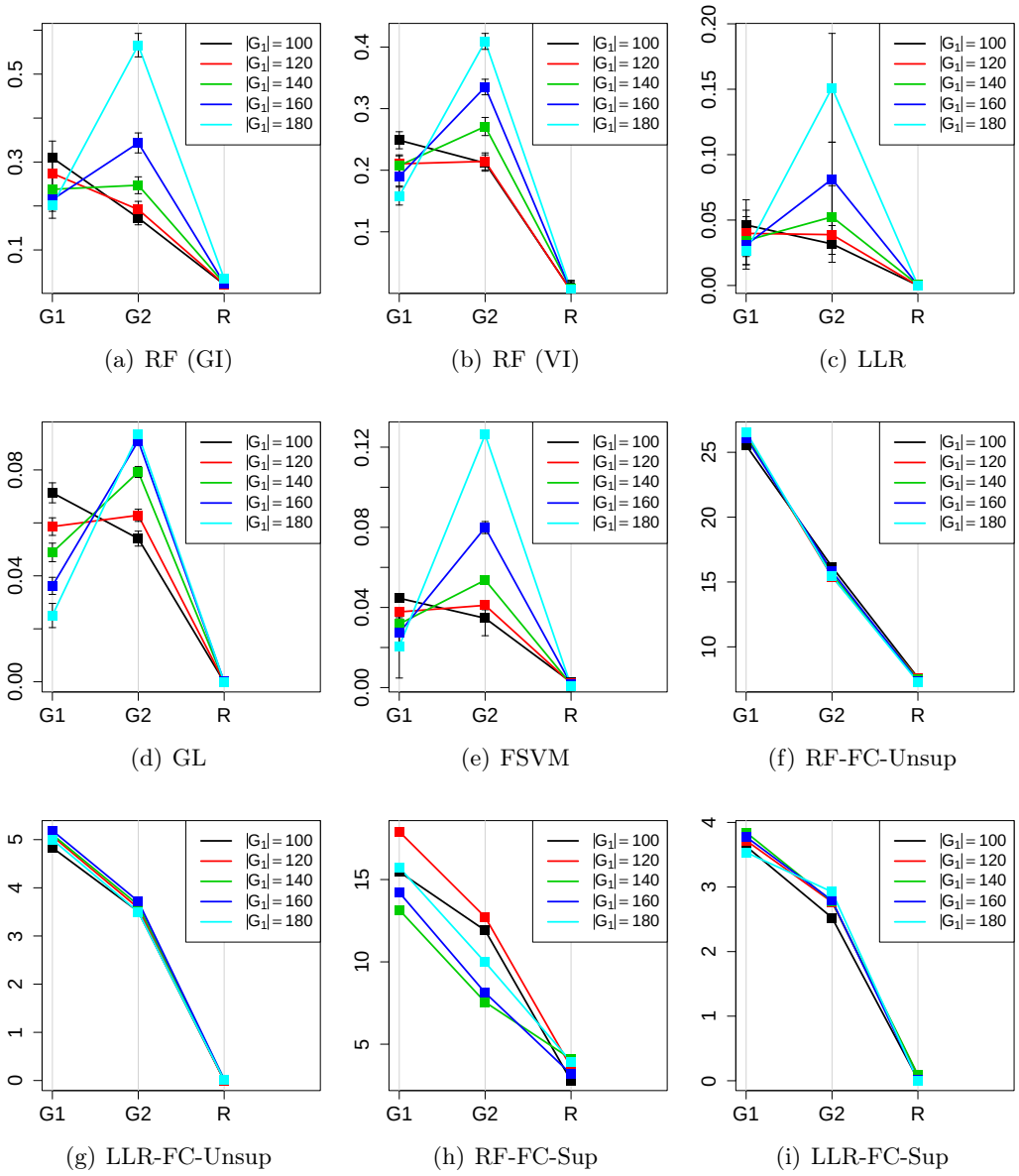
$\lvert G_1 \rvert$	100	120	140	160	180
RF	0.913±0.03	0.906±0.02	0.911±0.02	0.912±0.02	0.901±0.03
RF-FC-Unsup	0.915±0.03	0.921±0.03	0.916±0.03	0.921±0.02	0.915±0.02
RF-FC-Sup	0.928±0.03	0.930±0.02	0.925±0.02	0.924±0.02	0.922±0.02
LLR	0.940±0.03	0.941±0.02	0.939±0.02	0.940±0.02	0.938±0.02
LLR-FC-Unsup	0.966±0.01	0.966±0.02	0.987±0.02	0.970±0.02	0.964±0.02
LLR-FC-Sup	0.972±0.01	0.972±0.02	0.973±0.02	0.973±0.01	0.969±0.01
FSVM	0.967±0.02	0.966±0.02	0.963±0.02	0.965±0.02	0.951±0.02
GL	0.970±0.01	0.965±0.02	0.959±0.02	0.941±0.02	0.884±0.03

Table 5.2.: Accuracy of classification models on data from Simulation A. The values are averaged over 100 simulations.

$\lvert G_1 \rvert$	100	120	140	160	180
RF	0.56±0.14	0.52±0.17	0.55±0.16	0.55±0.16	0.55±0.18
RF-FC-Unsup	0.99±0.01	0.99±0.01	0.99±0.01	0.99±0.01	1.00±0.01
RF-FC-Sup	0.88±0.14	0.89±0.14	0.86±0.15	0.88±0.15	0.90±0.14
LLR	0.72±0.08	0.72±0.08	0.71±0.08	0.73±0.08	0.75±0.07
LLR-FC-Unsup	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
LLR-FC-Sup	0.87±0.21	0.86±0.20	0.91±0.17	0.91±0.18	0.93±0.14
FSVM	0.96±0.03	0.98±0.02	0.96±0.07	0.95±0.04	0.95±0.04
GL	0.95±0.02	0.94±0.02	0.94±0.01	0.94±0.01	0.91±0.02

Table 5.3.: Stability of feature importance of classification models on data from Simulation A. The scores are averaged over 100 simulations.

We evaluated the performance of RF, LLR, FSVM and GL with respect to the criteria described in Section 2.4. Figure 5.3 summarizes the importance values assigned to features from the three groups G_1, G_2 and R. The average feature weights

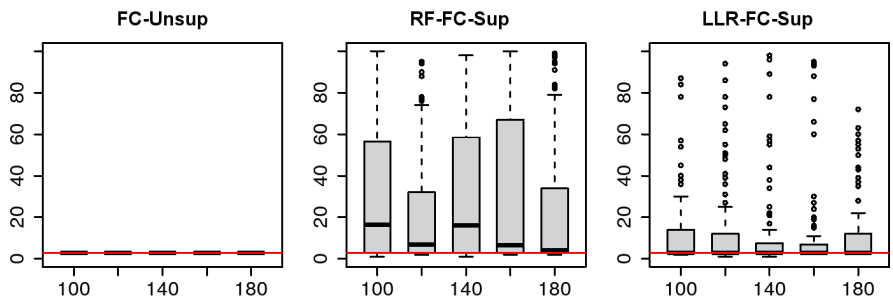


(a) RF (GI) (b) RF (VI) (c) LLR

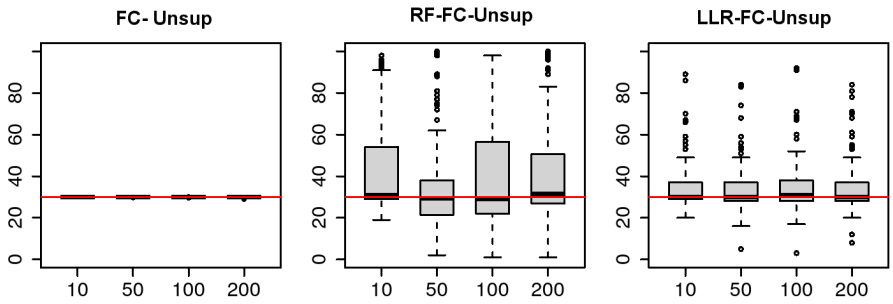
(d) GL (e) FSVM (f) RF-FC-Unsup

(g) LLR-FC-Unsup (h) RF-FC-Sup (i) LLR-FC-Sup

Figure 5.3.: Average importance of features for classification of data from Simulation A. The importance is averaged over groups G_1, G_2 and R.



(a)



(b)

Figure 5.4.: Boxplot summarizing the number of feature groups selected by FC-Unsup and FC-Sup (in combination with RF and LLR) for a) Simulation A and b) Simulation B. The red horizontal line shows the true number of groups. On the x-axis, the cardinality of G_1 is given.

over 100 simulations are shown for each chosen cardinality of group G_1, with indication of standard deviation added. The correlation bias is clearly demonstrated by the decreasing importance of group G_1 as its cardinality increases and conversely, the increasing relevance of G_2 as its cardinality decreases.

In the case of RF, when the number of features in group G_1 is larger than 140 ($|G_1|/|G_2| > 2.3$), the ranking of the groups given by GI and VI is incorrect in that features in G_2 falsely appear most relevant for the model (Figures 5.3a and b). On average, the same effect is observed in LLR models (depicted in Figure 5.3c).

In the case of FSVM and GL, the correlation bias is noticeable even if $G_1 = 120$ ($|G_1|/|G2| = 1.5$) (Figure 5.3e). In the context of the formal example from the Supplementary material (see Example of correlation bias, Supplementary material), note that our experimental results agree with the set of conditions (6) ($a = 5$, $b = 4$, $q = 120$, $r = 80$).

We trained our models based on feature grouping LLR-FC-Sup, LLR-FC-Unsup, RF-FC-Sup, RF-FC-Unsup. The unsupervised selection of the number of regions of consensus segmentation (FC-Unsup FC-Unsup) almost always lead to the correct number of groups of features (three). In contrast, if the number of regions is selected via cross validation, more often than not the number of groups are overestimated. Moreover, the RF models select a larger number of groups than the LLR models. Figure 5.4a shows a summary of the selected number of feature groups by FC-Sup and FC-Unsup. Importantly, none of the methods LLR-FC-Sup, LLR-FC-Unsup, RF-FC-Sup, RF-FC-Unsup is affected by correlation bias, which is evident from Figure 5.3f-i.

The prediction accuracy of all models is summarized in Table 5.2. All linear models outperform RF, which is expected because the simulations are based on a linear model. With FC-Unsup, both baseline methods RF and LLR achieve higher accuracy. FC-Sup always outperforms FC-Unsup, which is surprising, given that FC discovers the true number of feature groups and FC-Sup does not. Most probably, the classification is improved if each group is further split into several subgroups. This is possible because the feature groups are not spherical, but rather elongated, thus a single centroid is not the best representation of the group. As the cardinality ratio $|G_1|/|G_2|$ increases, the accuracy of FSVM and GL decreases and thus the LLR-FC-Unsup and LLR-FC-Sup become significantly better than FSVM and GL.

Table 5.3 shows the stability estimates for the various models. The RF are most unstable, probably due to their increased complexity and thus tendency to overfitting. As expected, the LLR models are most unstable among the linear models. FC improves dramatically the stability of RF and LLR. FC-Sup is always more stable than the baseline methods. Interestingly, FC-Unsup is more stable than FC-Sup. This is the case because the grouping of the features by FC-Unsup is driven only by the features themselves, while in the case of FC-Sup, the outcome also plays a role. The results show that FSVM and GL are also very stable models.

Simulation B.

Figure A.1 clearly demonstrates the correlation bias affecting a) RF (Gini Index), b) RF (Variable Importance), c) LLR, d) GL and e) FSVM models. Most dramatically, in the case of FSVM, as the cardinality of group G_1 increases to 200 features, the features in G_1 appear almost irrelevant. When the size of G_1 exceeds 100 features, the GL model selects only group G_1 and disregards all other predictive groups. As in the case of Simulation A, FC-Unsup and FC-Sup succeed to remove the correlation bias (Figures A.1f-i). FC-Unsup in general finds the true number of groups (thirty), but as the number of correlated features in G_1 increases, the number of groups is sometimes underestimated (Figure 5.4b). FC-Sup often selects a larger number of feature groups.

The accuracy of the models is given in Table 5.4. The models show similar relative performance as seen in Simulation A: FC-Unsup and FC-Sup always outperform the baseline models and FC-Sup slightly outperforms FC-Sup. LLR always outperforms RF, probably due to the underlying linear model. The FSVM and GL lose accuracy as the size of the group G_1 increases.

Table 5.5 shows the stability scores of all models. FC-Unsup and FC-Sup improve the stability of the baseline models, with FC-Unsup being more stable than FC-Sup.

$\|G_1\|$	10	50	100	200
RF	0.741±0.05	0.744±0.03	0.724±0.04	0.714±0.04
RF-FC-Unsup	0.754±0.05	0.758±0.05	0.758±0.05	0.753±0.05
RF-FC-Sup	0.756±0.05	0.758±0.05	0.758±0.05	0.761±0.05
LLR	0.777±0.06	0.787±0.05	0.783±0.05	0.788±0.05
LLR-FC-Unsup	0.857±0.03	0.874±0.03	0.868±0.04	0.863±0.03
LLR-FC-Sup	0.870±0.03	0.890±0.03	0.883±0.03	0.880±0.03
FSVM	0.894±0.03	0.898±0.03	0.891±0.03	0.889±0.04
GL	0.828±0.05	0.739±0.04	0.693±0.05	0.690±0.05

Table 5.4.: Accuracy of classification models on data from Simulation B. The values are averaged over 100 simulations.

$\|G_1\|$	10	50	100	200
RF	0.79 ± 0.05	0.76 ± 0.05	0.67 ± 0.10	0.62 ± 0.10
RF-FC-Unsup	0.93 ± 0.03	0.95 ± 0.03	0.96 ± 0.03	0.97 ± 0.03
RF-FC-Sup	0.90 ± 0.05	0.87 ± 0.14	0.88 ± 0.12	0.92 ± 0.07
LLR	0.78 ± 0.07	0.75 ± 0.07	0.75 ± 0.07	0.73 ± 0.07
LLR-FC-Unsup	0.97 ± 0.01	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.01
LLR-FC-Sup	0.93 ± 0.05	0.93 ± 0.08	0.93 ± 0.09	0.93 ± 0.10
FSVM	0.97 ± 0.04	0.94 ± 0.05	0.93 ± 0.06	0.93 ± 0.05
GL	0.92 ± 0.03	0.89 ± 0.05	0.94 ± 0.03	0.93 ± 0.03

Table 5.5.: Stability scores of classification models on data from Simulation B. The scores are averaged over 100 simulations.

5.5.2. Real data

Bladder tumors.

By applying FC-Unsup to the bladder data, we obtained 113 ± 11 feature groups with grade labelling and 140 ± 31 groups with stage labeling. The varying number of groups corresponds to the 10 training sets of the cross-validation procedure. When applied with RF, FC-Sup method finds 107 groups (with grades labeling) and 20 groups (with stages labeling), respectively. When applied with LLR, FC-Sup partitions the set of features into smaller number of groups: 24 (with grade labeling) and 19 (with stage labeling).

In both classification scenarios (with grades and stages labeling), the LLR with FC-Sup yields best accuracy. RF with FC-Unsup and FC-Sup improve the baseline model in the case of tumor grade prediction but decrease it slightly when tumor stage is predicted (Table 5.6).

The stability scores of the prediction models (Table 5.7) lead to the same conclusions as the experiments on the simulated data: the RF and LLR models are most unstable, however using FC-Unsup and FC-Sup results in significant improvements. FC-Sup is more unstable than FC-Unsup.

Figure 5.5 shows the feature importance reported by some of the methods investigated. For comparison purposes, to the set of prediction methods analyzed, we added a univariate measure of feature relevance, consisting of t-test p-values (log-transformed) (Figure 5.5e). A t-test was applied to each feature independently, in order to evaluate the significance of the difference between the means of the two classes. We do not perform multiple testing correction because we use the log-transformed p-values as scores, and not as indicators of relevance. Concerning interpretability, FC-Unsup and FC-Sup with LLR and RF or FSVM are the better models, reporting clear groups of features with identical weights. GL is also suitable for finding relevant groups, however there is high variance among the weights within groups.

In the absence of a true model, it is difficult to show how correlation bias affects classification models. However, in the case of classification with stage labeling, we speculate that correlation bias is observable in the feature importance given by FSVM. A large group of 175 correlated features on chromosome 7 is ranked fifth (w.r.t. absolute value of the weights) by FSVM (Figure 5.5a). However, a large subgroup of this group of features located towards the short arm of the chromosome is indicated as most relevant by RF-FC-Unsup, LLR-FC-Unsup, as well as by univariate t-tests (Figures 5.5b-d). It is possible that the lower rank of this group of features in the FSVM model can be caused by the correlation bias. The FSVM includes the entire group of correlated features, at the price of lower average weights. In order to verify this hypothesis, we constructed a new dataset, by assigning one feature representative to each group with identical weights in the FSVM model. The representatives are computed by averaging over the corresponding group. All features with null weights were excluded. This procedure essentially uses FSVM for discovering groups of correlated features and computes the centroid of each group

selected by the FSVM model. The resulting reduced dataset has 11 features. We trained and evaluated a FSVM model on the new dataset, this time without fused penalty ($\mu = 0$ in (5.3)), since most probably there are no further groups to be discovered. We call the new model *reduced FSVM*. In Figure 5.5e, the weights of the features in the original FSVM model are represented (in absolute value and only the regions with non-zero weights). Figure 5.5f shows the weights of the reduced FSVM (in absolute values), extended for convenience so as to be aligned to the original weights. The representative feature corresponding to the group located on chromosome 7 receives highest absolute weight, which could indicate that the correlation bias has been removed.

Breast tumors.

FC-Unsup identifies 130±37 groups of features with ER labeling and 127±32 groups of features, with PR labeling. FC-Sup in combination with RF selects an optimal partitioning into 195 groups (with ER labeling) and 163 groups (with PR labeling), respectively. Table 5.8 summarizes the accuracy of the different algorithms on ER and PR classification. FC-Unsup improves the accuracy of the RF in both cases and of LLR in the case of PR labeling, but decreases slightly the accuracy of the LLR model when ER status is predicted. In the case of ER classification, all RF and LLR models outperform FSVM and GL, however by a small margin. In the case of the PR classification, FSVM performs best.

The models investigated have similar stability scores as in the case of bladder tumors: RF and LLR are most unstable, FC with LLR and RF have increased stability, comparable to that of GL and FSVM models and FC-Sup is less stable than FC-Unsup (Table 5.9).

The feature importance reported by the various methods investigated in general confirms the findings reported in the original study (Climent et al., 2007). An interesting aspect is shown in Figure 5.6: FC-Unsup and FC-Sup in combination with RF and LLR select a group of features in chromosome 13 as highly relevant for classification of PR status. In the original study (Climent et al., 2007), none of these features were reported significant, based on univariate association with the outcome (corrected t-test p-value). Allelic loss at chromosome 13 is known to be associated with poor prognosis in breast cancer, due to the loss of the tumor suppressor gene BRCA2, located in this region. Associations with low progesterone content have been reported previously in the literature (Eiriksdottir et al., 1998), which we confirm in our study.

5.6. Results II - Correction of the GI importance of random forest with PIMP

Simulation C demonstrated clearly that random forest GI is biased such that variables with a large number of categories receive a higher variable importance (Figure 5.7, left). In contrast, the PIMP scores (p-values) computed using a gamma distri-

	Grades		Stages	
	Acc	AUC	Acc	AUC
RF	0.792±0.02	0.827	0.833±0.03	0.882
RF-FC-Unsup	0.833±0.01	0.878	0.810±0.02	0.882
RF-FC-Sup	0.833±0.01	0.885	0.810±0.03	0.884
LLR	0.823±0.01	0.800	0.798±0.01	0.821
LLR-FC-Unsup	0.854±0.02	0.838	0.774±0.01	0.757
LLR-FC-Sup	0.865±0.01	0.771	0.845±0.02	0.873
FSVM	0.813±0.02	0.642	0.810±0.05	0.780
GL	0.833±0.02	0.775	0.833±0.04	0.780

Table 5.6.: Performance of different classifiers on the bladder data.

	Grades	Stages
RF	0.55±0.03	0.60±0.03
RF-FC-Unsup	0.80±0.04	0.83±0.04
RF-FC-Sup	0.78±0.06	0.69±0.12
LLR	0.61±0.12	0.66±0.11
LLR-FC-Unsup	0.86±0.04	0.87±0.08
LLR-FC-Sup	0.72±0.11	0.66±0.18
FSVM	0.75±0.16	0.88±0.05
GL	0.72±0.13	0.87±0.09

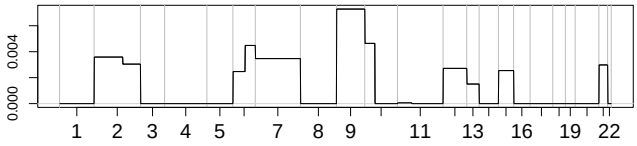
Table 5.7.: Stability of feature importance of classification models on the bladder data.

	ER status		PR status	
	Acc	AUC	Acc	AUC
RF	0.658±0.01	0.664	0.677±0.02	0.673
RF-FC-Unsup	0.670±0.06	0.635	0.682±0.08	0.660
RF-FC-Sup	0.665±0.02	0.663	0.671±0.02	0.667
LLR	0.683±0.03	0.692	0.683±0.03	0.733
LLR-FC-Unsup	0.671±0.02	0.718	0.689±0.04	0.723
LLR-FC-Sup	0.696±0.02	0.676	0.714±0.01	0.691
FSVM	0.658±0.02	0.660	0.745±0.02	0.800
GL	0.658±0.01	0.692	0.702±0.02	0.698

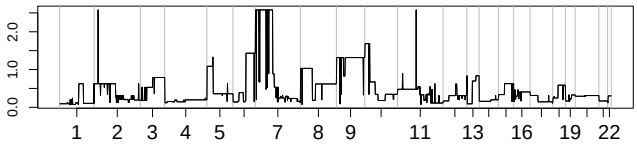
Table 5.8.: Performance of different classifiers on the breast data.

	ER status	PR status
RF	0.53 ± 0.04	0.49 ± 0.03
RF-FC-Unsup	0.70 ± 0.07	0.74 ± 0.08
RF-FC-Sup	0.78 ± 0.05	0.76 ± 0.07
LLR	0.65 ± 0.07	0.68 ± 0.06
LLR-FC-Unsup	0.89 ± 0.05	0.71 ± 0.14
LLR-FC-Sup	0.68 ± 0.08	0.61 ± 0.12
FSVM	0.77 ± 0.07	0.85 ± 0.03
GL	0.54 ± 0.14	0.72 ± 0.13

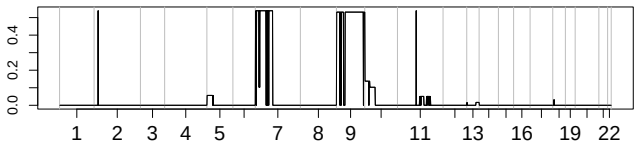
Table 5.9.: Stability of classification models evaluated on the breast data.



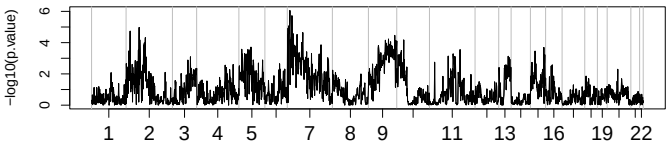
(a) FSVM



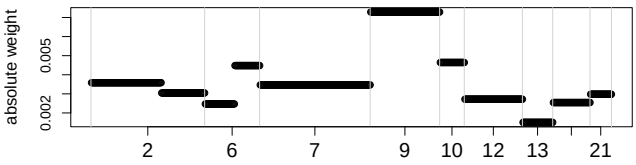
(b) RF-FC



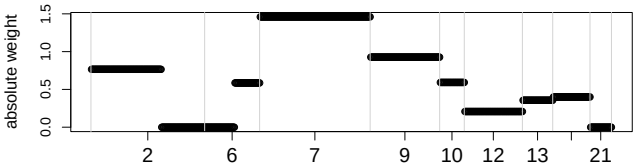
(c) LLR-FC-Unsup



(d) *t*-test

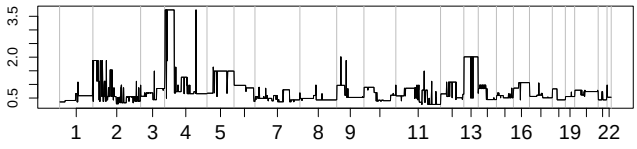


(e) FSVM(only relevant groups)

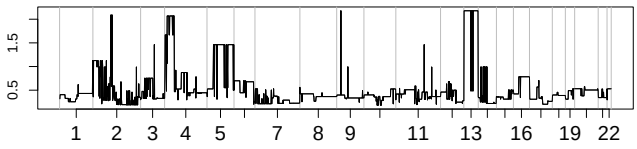


(f) reduced FSVM

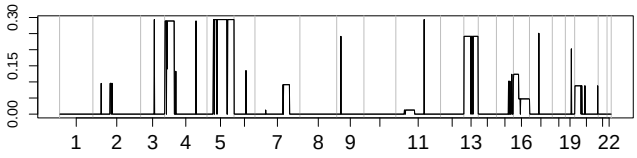
Figure 5.5.: Feature relevance (in absolute value) by different classification models on the bladder dataset with stage labeling. The features are sorted according to genomic position and the chromosomes are shown along the x-axis.



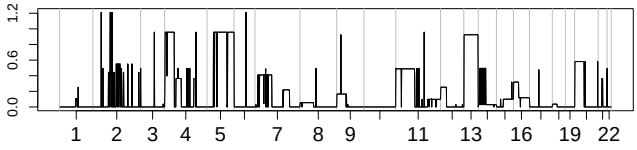
(a) RF-FC-Unsup



(b) RF-FC-Sup



(c) LLR-FC-Unsup



(d) LLR-FC-Sup

Figure 5.6.: Feature relevance (in absolute value) by different classification models on the breast dataset with PR labeling. The features are sorted according to genomic position and the chromosomes are shown along the x-axis.

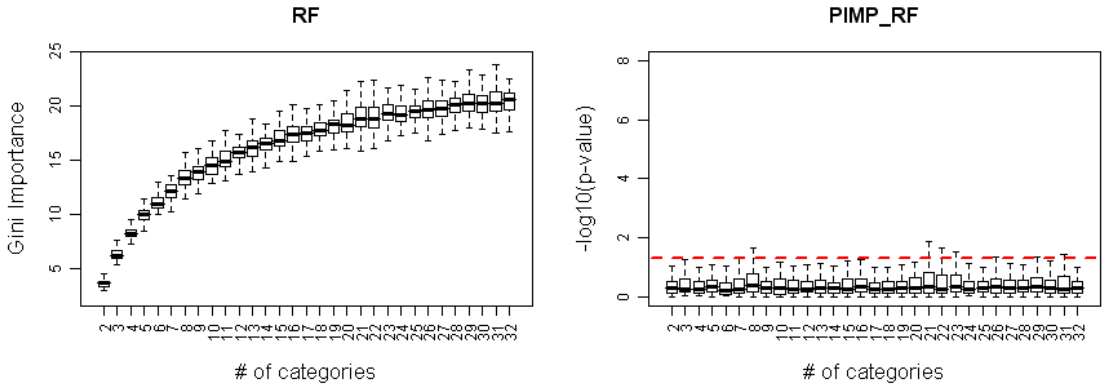


Figure 5.7.: Simulation C: GI variable importance in dependence of number of categories.

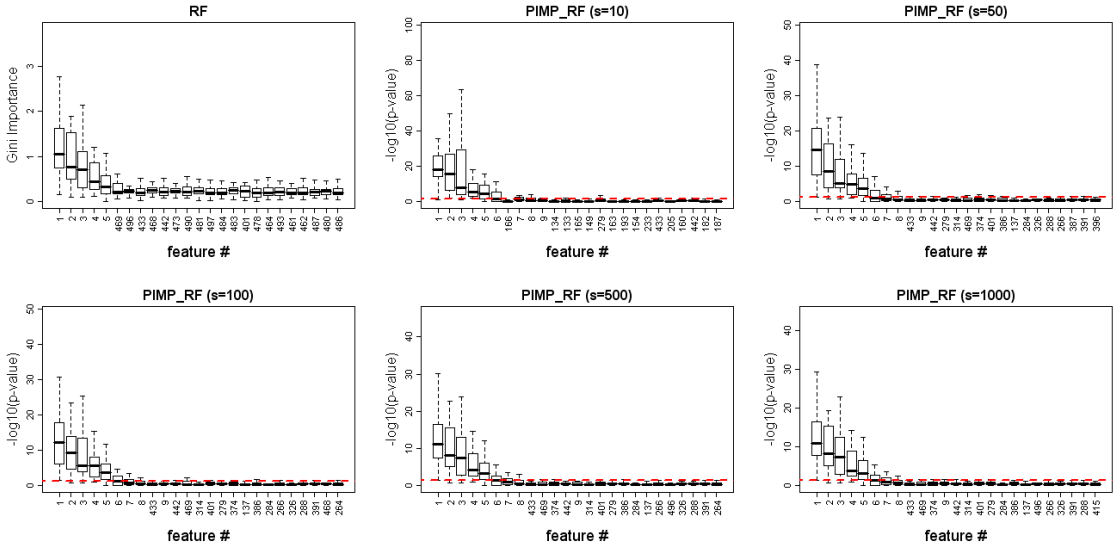


Figure 5.8.: Discovery of relevant features in simulation scenario D.

response variable at a 5% threshold (dashed red line).

Figure 5.8 shows box plots of the RF variable importance computed in the simulation scenario D. The features were ranked with respect to their mean importance in all simulations. For the sake of visualization only the top 25 of the 500 features were displayed. In the first setting, the first 12 variables were selected to be predictive. Using GI (top left), only the first five variables ($r = 0.24$ to $r = 0.16$) were recovered perfectly. By comparison, using the PIMP (gamma distribution) of GI with $s = 10$, the first six variables ($r = 0.24$ to $r = 0.14$) were recovered perfectly and variables seven to nine were ranked eighth to tenth. Larger values of s led to perfect recovery of the first eight variables ($r = 0.24$ to $r = 0.10$) and the ninth feature ($r = 0.08$) is always among the top 13.

RF baseline	PIMP-RF	RF Top 1%	RF Top 5%	RF Top 10%	cforest
0.35 ± 0.06	0.25 ± 0.05	0.27 ± 0.08	0.26 ± 0.06	0.28 ± 0.06	0.32 ± 0.09

Table 5.10.: Performance of different random forest models. The baseline is the classical RandomForest (RF). For comparison, the average error rates are shown for PIMP-RandomForest(PIMP-RF), for RandomForest models trained on the top ranking 1%, 5% and 10% features and for the cforest algorithm.

5.6.1. Model improvement

We used simulation D to validate our improved PIMP-RandomForest model. We ran 100 simulations and we compared the accuracy of RandomForest, PIMP-RandomForest, RandomForest re-trained only using the top ranking features and the cforest model. The error rates were computed on an independent test set. Table 5.10 shows the improvements of accuracy of different methods over the classical RandomForest. The PIMP-RandomForest model performs significantly better than the RandomForest, with an average decrease of OOB error rate of 10%. The RandomForest trained on the top-ranking 1%, 5% and 10% of the features also yields better models, due to the decrease in variance. Choosing the top 5% results in a model with accuracy comparable (although still inferior) to the PIMP-RandomForest. However, it is not clear *a priori* how many top ranking features should be selected for a refined model. With the p-values provided by the PIMP algorithm, one can simply use the classical 0.05 significance threshold for selecting the most relevant variables. Notably, the cforest algorithm is superior to the classical RandomForest, but the average decrease of error rate is significantly smaller than the one achieved by PIMP-RandomForest.

5.7. Discussion and conclusions

We have shown that several widely-used classification algorithms can generate misleading feature rankings when the training datasets contain large groups of correlated features. This can confound model interpretation, since large groups of predictive features can be masked and falsely appear irrelevant. Such an effect is likely to occur because variables relating to a biological process or genomic location of high interest (w.r.t. a phenotype) are over-represented in the probes set of microarray-based experiments. In this article, we have described the correlation bias and have shown that it affects random forest, Lasso logistic regression, group Lasso and fused SVM models. We used two artificial datasets based on linear models to show that the expected importance of the features in a correlated group decreases as the size of the group increases. We also illustrated the correlation bias caused by the combination of fused and Lasso penalties by means of a theoretical example, which considers the particular case of two groups of correlated features. We showed that correlation bias can be reduced using a feature grouping prior to model fitting. For this purpose, we used consensus segmentationbased on feature clustering, with two methods for selecting the number of regions: supervised (by cross validation)and unsupervised (by minimizing $\Omega_{0.98}$).

We showed using simulated data experiments that FC-Unsup and FC-Sup successfully remove the correlation bias, improve the stability of feature importance and increase the accuracy of the baseline methods. FC-Sup outperforms FC-Unsup in terms of accuracy, but FC-Sup is faster and has higher stability. The classification of the real data shows that FC-Unsup dramatically increases the model interpretability and stability of feature importance. Moreover, in five out of eight classification tasks, FC-Unsup improved the accuracy of the baseline models. FC-Sup improves the accuracy of the baseline models in six out of eight classification tasks. FC-Sup used in combination with Lasso logistic regression yields highest accuracy in three out of four cases.

We also proposed an algorithm for correcting for the bias of variable importance given by random forest. The method permutes the response vector for estimating the random importance of a feature. Under the assumption that the random importance of a feature follows a normal distribution, the likelihood of the measured importance on the unpermuted outcome vector can be assessed. The resulting p-value can serve as an unbiased measure of variable relevance. We showed how this method can successfully adjust the feature importance computed with the classical RandomForest algorithm and how it can be used for feature selection with random forest. We also introduced an improved random forest model that is computed based on the most significant features determined with the PIMP algorithm, which validates the feature selection method.

Simulation C demonstrated that the Gini importance of the RandomForest favor features with large number of categories and showed how our algorithm alleviates the bias. Simulation D demonstrated the usefulness of the algorithm for generating a correct feature ranking. For all methods, the feature ranking based on the unprocessed importance measures could be improved.

We proposed a corrected random forest model based on the PIMP scores of the features and we demonstrated that it is superior in accuracy to the cforest model. The major drawback of the method is the requirement of time-consuming permutations of the response vector and subsequent computation of variable importance. However, our simulations showed that already a small number of permutations (e.g. 10) provided improvements over a biased base method. For stability of the results any number from 50 to 100 permutations is recommended. The algorithm can easily be parallelized, since computations of the random variable importance for every permutation are independent, and therefore allow for an even better scalability with respect to available computational resources. With parallelization, the running time of our algorithm is only a few times longer than the running time of a classical RandomForest, which is very fast even for large instances.

We argue that the PIMP algorithm can also be used as a post-processing step with other learning methods that provide (unbiased) measures of feature relevance, such as linear models, logistic regression, SVM, etc. The raw scores given by these models provide with a feature ranking, but usually it is difficult to choose a significance threshold. The PIMP p-values are easier to interpret and provide a common measure that can be used to compare feature relevance among different models. In this work,

we use the PIMP procedure only for selecting features with random forest.

6. Applications: Prediction of Tumor Phenotype for Breast Cancer and Neuroblastoma

> Nature even in chaos cannot proceed otherwise than regularly and according to order.
>
> Immanuel Kant

In this chapter we discuss in detail two interesting applications of our methodological pipeline to the analysis of breast cancer and neuroblastoma tumors. The datasets have been introduced in Chapter 4.

6.1. Prediction of breast cancer phenotype

In Chapter 4 we have introduced three cohorts of breast tumors corresponding to different interesting subtypes of the disease. In this chapter, we present extensive phenotype prediction based on each tumor cohort. The phenotype is always a binary variable. The predictors are LLR and RF as baseline models and combinations with CB-MUG, CB-KeS and CR-FC. We present prediction accuracies for each prediction problem and comment on the feature importance resulting from the best models.

To our knowledge, there are not many studies that attempt supervised prediction of tumor phenotype based on copy number aberrations. The main reason is that expression profiles are thought to reflect the characteristics of a tumor more closely than copy number aberrations. As a consequence, it is largely unknown how informative CNA profiles are of phenotype. In this chapter, we shed some light on this question.

We would also like to clarify that the supervised modeling of a phenotypical indicator based on copy number aberrations does not necessarily aim at the automatic prediction of new cases. Take for example the prediction of the ER (estrogen receptor) status of a breast tumor: it is the expression of the ESR1 binding factor in the breast that determines the status of the tumor and it would be pointless to derive a complex and potentially inaccurate diagnostic marker based on CNAs to evaluate it. However, the groups of ER positive and ER negative tumors are very different overall and probably have undergone distinct progression patterns, which should be understood. Therefore, while being an important indicator of how different the groups are, the prediction accuracy is not the only focus of this section. We will

look at the feature importance to gain insight into the relation between CNAs and phenotype.

Another important comment on our experiments refers to the difficulty of integrating two or more datasets for a more comprehensive analysis. For example, we could combine the three datasets on breast cancer and thus obtain a richer sample set for prediction and better performance. However, our experiments showed the contrary: the performance decreases, due to systematic differences between the cohorts (batch effect). The reproducibility of microarray results has long been an issue of debate in the community (Ein-Dor et al., 2005). We adopted the following strategy in our investigations on breast cancer: we fit different models to each dataset and then we compare the models qualitatively. We comment on these aspects in the discussion section at the end of the chapter.

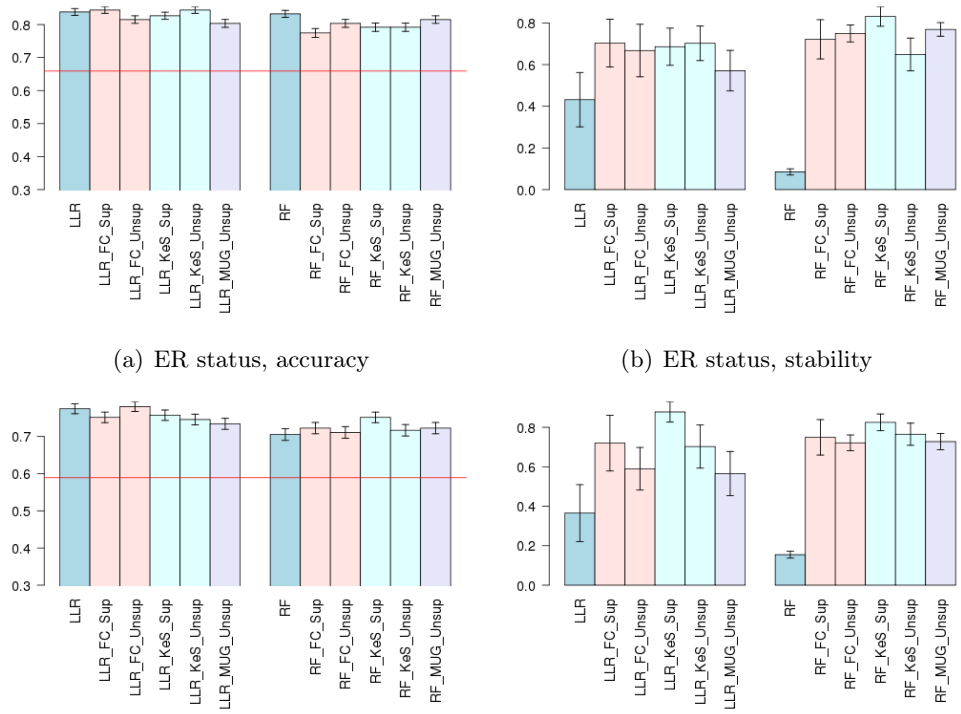
6.1.1. Analysis of breast cancers from dataset breast173

The arrayCGH dataset **breast173** was first introduced by Bekhouche et al. (2011). The original analysis was focused on discriminating between copy number profiles of IBC (inflammatory breast cancer) and NIBC (non-inflammatory breast cancer) subtypes. However, the public repository does not contain the corresponding annotation, so we were not able to repeat the analysis. Instead, we used the available annotations in our experiments, consisting of estrogen receptor and progesterone receptor status, which can be either positive (ER+, PR+) or negative (ER-, PR-).

Figure 6.1 shows the accuracy of our models (for model and axis annotations, revisit Section 5.3.2). The accuracy of all models is slightly larger that 80% in the case of ER status prediction and around 70% in the case of PR status prediction, and significantly larger than the Bayes error (marked with red horizontal line in the figure), indicating a non-negligible difference between copy number profiles of the two groups.

The LLR-based models and RF-based models yield comparable accuracy, with LLR slightly outperforming RF in the case of PR status prediction. The accuracy of the LLR-based models remains similar after consensus segmentation, with worst performance given by the CB-MUG algorithm. The baseline RF model outperforms the models based on consensus segmentation in the case of ER status prediction, but is outperformed in the case of PR status prediction. However, the differences in performance are small (up to 3%). A much larger and convincing difference can be noted in the stability plots (Figure 6.1). The LLR-based models with consensus segmentation yield 1.5-fold larger stability than the baseline, while the RF-based models with consensus segmentation are up to four times more stable.

It has been indicated before that ER positive and ER negative tumors have different molecular characteristics, mostly by studies based on differential gene expression: Gruvberger et al. (2001) and West et al. (2001) report very high prediction accuracy of ER status based on gene expression signatures, of 100% and 90%, respectively and point out that a long list of about a hundred genes is significantly associated with ER status and equivalent models relying on different subgroups can achieve



(a) ER status, accuracy

(b) ER status, stability

(c) PR status, accuracy

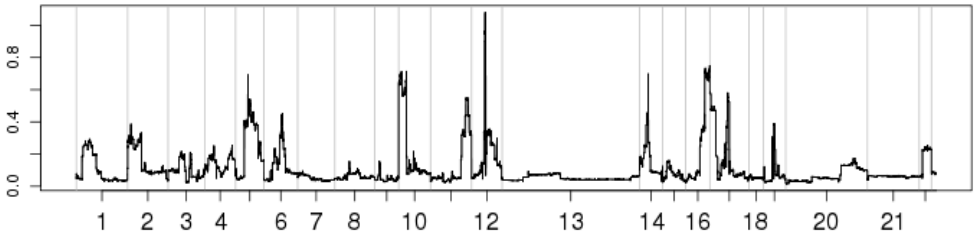
(d) PR status, stability

Figure 6.1.: Accuracy of prediction models trained on the **breast173** dataset and stability of the feature importance. The vertical bars show accuracy, with indication of variance. Colors associate with the specific consensus segmentation method used. The two groups of bars correspond to methods based on LLR (left) and RF (right) of each panel. The red horizontal red line marks the Bayes error.

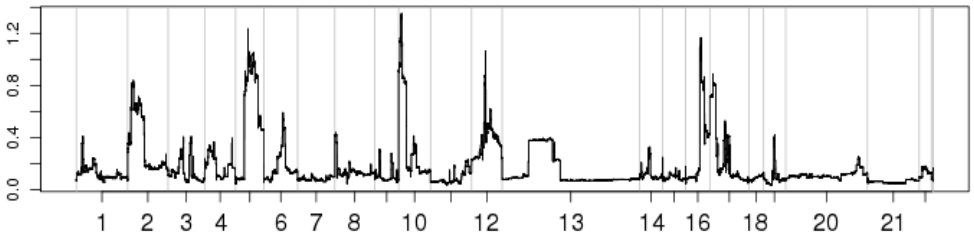
similar performance. Most of the studies that investigate the association between copy number aberrations and ER status use simple t-tests for deciding whether a probe is significantly associated or not with the hormone receptor status (Loo et al., 2004; Fridlyand et al., 2006). Chin et al. (2007) and Horlings et al. (2010) are among the very few articles that attempt supervised classification of tumor samples based on copy number aberrations. Chin et al. (2007) report modest accuracies, of up to 64%, while Horlings et al. (2010) only mention the features that were found discriminative.

Figure 6.2a shows a consensus of the feature importance values of all methods based on consensus segmentation with ER status as outcome. We combined the various models because there was no method clearly outperforming the rest. The combination is linear, with weights proportional to the accuracy of the corresponding model. We find the following aberrations characteristic to ER negative tumors: 2p22-25 gain, 5q loss, 10p gain. The following were aberrations more commonly associated with ER positive tumors: 1p deletion, 5q gain, 6q loss, 11q loss, 11q13 amplification, 16q loss.

The most discriminative features (according to most models) were regions located on chromosome 10 and on chromosome 16. Figure A.2a from the Appendix shows a

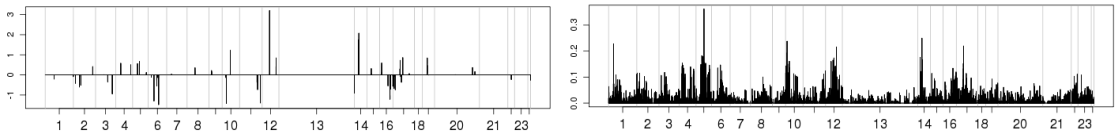


(a)



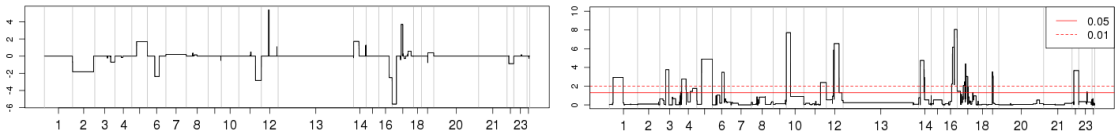
(b)

Figure 6.2.: Consensus feature importance of prediction models on breast173 dataset with a) ER status and b) PR status as outcome.



(a) LLR

(b) RF



(c) LLR-FC-Sup

(d) RF-FC-Sup

Figure 6.3.: Consensus segmentation improves model interpretation. Figure shows feature relevance given by a) LLR coefficients, b) RF importance measure, c) LLR-FC-Sup coefficients and d) RF-FC-Sup PIMP p-values. The x-axis orders the features according to chromosome and genomic position. In red, significance thresholds are shown for the RF-FC-Sup p-values.

detailed view of chromosome 10: the panel above contains the copy number data and the panel below shows the feature importance values if our models (green for LLR-based models, black for RF-based models). RF-based models assign a very large feature importance to the 10p gain event. Figure A.2b from the Appendix shows the deletion of 16q, highly associated with ER status as indicated by all models. Our findings are supported by other authors (Horlings et al., 2010; Hungermann et al., 2011).

Figure 6.3 demonstrates the benefits of consensus segmentation. In Figure 6.3a and b, we show the baseline models LLR and RF. A couple of irrelevant genes from chromosome 16 were selected by the LLR model (their function is unrelated to carcinogenesis), while RF assign moderate constant importance values to the entire chromosome arm 16q. After consensus segmentation, the relevance of chromosome 16q is greatly emphasized, in fact becoming the most predictive feature, for example by FC-Sup (see Figure 6.3c and d). This is an indication that the correlation bias affects the baseline models.

Figure 6.2b shows the consensus importance for the prediction of PR status. There is a significant association between the ER and PR status in breast tumors, therefore it is expected that the models are similar. It appears that the association between the copy number at chromosome 10q is stronger with the PR status than with ER status, as well as with the gain at 2q.

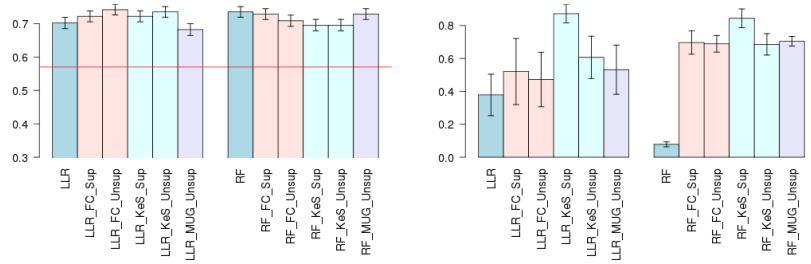
6.1.2. Analysis of breast cancers from dataset breast167

The cohort of 167 breast cancer samples investigated by (Russnes et al., 2010) contains a large variety of tumors, of different histological subtypes, stages, grades and hormone receptor status. The prediction accuracy and the stability of feature importance is shown in Figure 6.4.

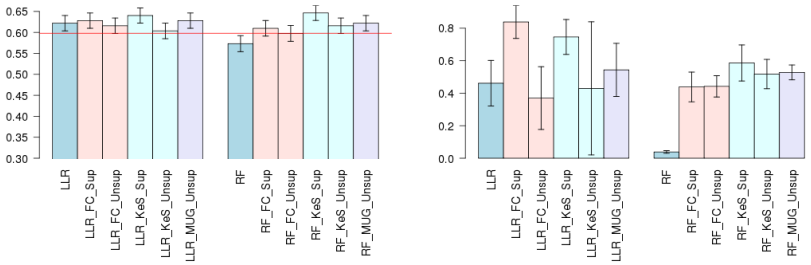
The accuracy after consensus segmentation is comparable to that of the baseline models, being significantly worse only in rare cases. However, the overall prediction performance is poor, the accuracy being often comparable to that of the Bayes error. In the case of ER status and lymph node status prediction, a marginal improvement is observed. The stability of feature importance increases dramatically after grouping the correlated features in the case of the RF-based models, as observed in the previous analysis (on dataset **breast173**).

We trained models with the ER status outcome and we obtained the consensus feature importance shown in Figure 6.5a. Chromosome 16 stands out as one of the most predictive regions, however it is not the same genomic region as given by the similar models trained on **breast173**. The aberration indicated by the present models is gain of 16p, which is more present in ER positive tumors (see Figure A.3a from the Appendix), while the model trained on **breast173** points to loss of 16q as more predictive. Gain of 16p has been associated before with ER positiveness (Rennstam et al., 2003; Fang et al., 2011), as well as loss of 16q. Possible bias in the selection of the patients for the two cohorts may be responsible for the different representation of the two markers. Chromosome 6 is also indicated as predictive, with loss of 6q more often occurring in ER positive tumors, as also shown by Fang et al. (2011) (see A.3b from the Appendix). Almost all the top ranking regions have been shown to discriminate between the ER positive and negative tumors in previous studies (Fang et al., 2011).

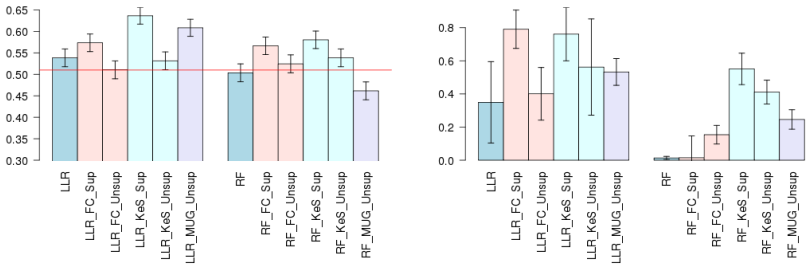
The most informative regions of PR status are located on chromosomes 6 and 7 (see Figure 6.5b). Gain of 16p (top ranking region) is characteristic to PR positive tumors, as also noted by Rennstam et al. (2003). Also, regions located on chromo-



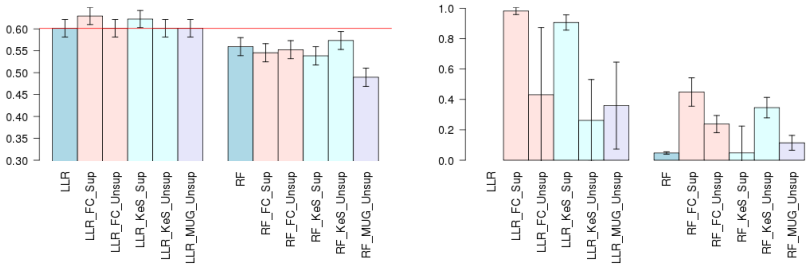
(a) ER status, accuracy (b) ER status, stability



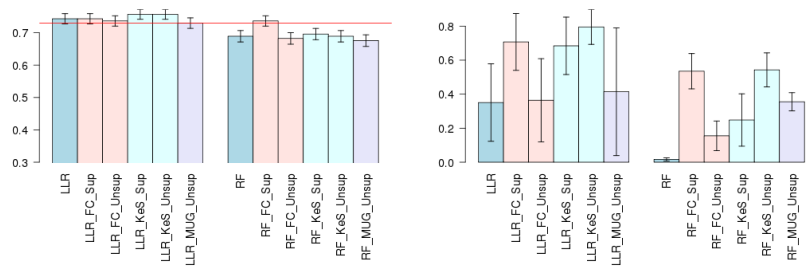
(c) PR status, accuracy (d) PR status, stability



(e) Lymph node status, accuracy (f) Lymph node status, stability



(g) Stage, accuracy (h) Stage, stability



(i) Histological subtype, accuracy (j) Histological subtype, stability

Figure 6.4.: Accuracy of prediction and stability of feature importance models trained on the **breast167** dataset.

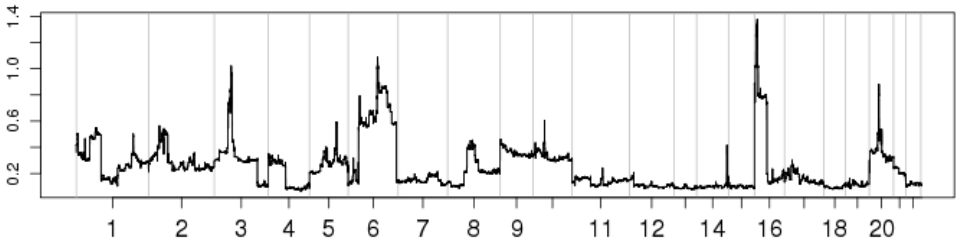
some 8 appear significant, especially the 8q arm. In general, the feature importance is in discordance with the values obtained with **breast173** as training dataset.

We discuss also the most predictive features for lymph node status, despite the low prediction accuracy. Figure 6.5c illustrates the consensus feature importance, computed as before. The methods agree less with each other, which results in ragged consensus feature importance curve. The highest peak is observed at chromosome 9, corresponding to amplification at 9q13. This region contains the oncogenes P16 and PTC and has been associated with lymph node status and metastasis before in the literature (Minobe et al., 1998). Figure A.4a from the Appendix illustrates this region. The next highest peak according to our models is located on chromosome 3. Associations between CNAs at 3q gain is a very strong marker for aggressive tumors and relapse (Janssen et al., 2003). Another very interesting feature is the amplification of 11q13.1 (see Figure A.4b from the Appendix). Amplification of 11q13 is a known hallmark of breast cancers, harboring two oncogene candidates CCND1 (coding for cell cycle regulatory gene cyclin D1) and EMS1 (coding for the filamentous actin binding protein and c-Src substrate cortactin). Ormandy et al. (2003) have described an extended locus of amplification consisting of four 'cores' at 11q13. Our peak is located at the first core (11q13.1), to which candidate oncogenes are still to be assigned.

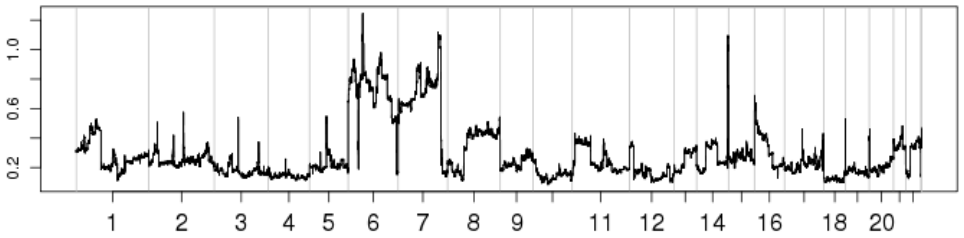
6.1.3. Analysis of breast cancers from breast54 dataset

An interesting study presented by Sircoulomb et al. (2010) is focused on investigating the genetic and transcriptomic properties of breast cancers presenting amplification of the ERBB2 gene locus. The ERBB2 gene (erythroblastic leukemia viral oncogene homolog 2), also known as HER-2/neu codes for an epidermal growth factor receptor. This focal amplification occurs in around 20% of the breast cancers, is often associated with overexpression and importantly, poor outcome of the patient. A common therapy option for ERBB2-amplified tumors consists of administering the humanized monoclonal antibody *transtuzumab* or the kinase-inhibitor *lapatinib*. Many tumors however develop resistance to transtuzumab, through molecular mechanisms that are not fully understood. Sircoulomb et al. (2010) try to tackle the apparent heterogeneity of ERBB2-amplified breast cancers by discovering subtypes and patterns of genetic and transcriptomic aberrations that associate with the subtypes. Although hierarchical clustering of samples based on DNA copy number suggests two subtypes, the authors report that the clinicopathological indicators do not appear to associate significantly with these subtypes.

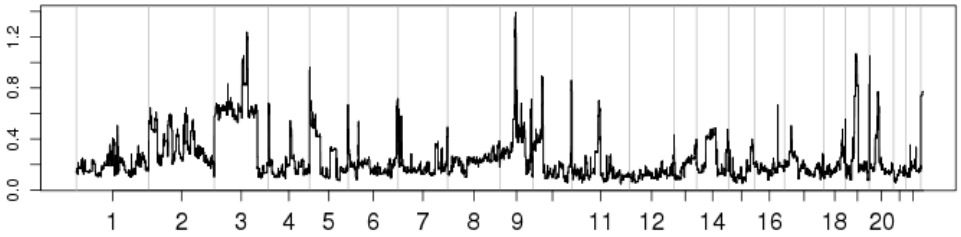
We applied our methodology for prediction of various tumor indicators with the hope that a supervised approach would reveal associations that Sircoulomb et al. (2010) have failed to discover. We conducted the following analyses: prediction of ER (estrogen receptor) status, of PR (progesterone receptor) status and tumor type that can be either inflammatory (IBC) or non-inflammatory breast cancer (NIBC). The number of samples available for each prediction case are: 47 samples for ER and PR prediction and 51 samples for subtype prediction.



(a)



(b)



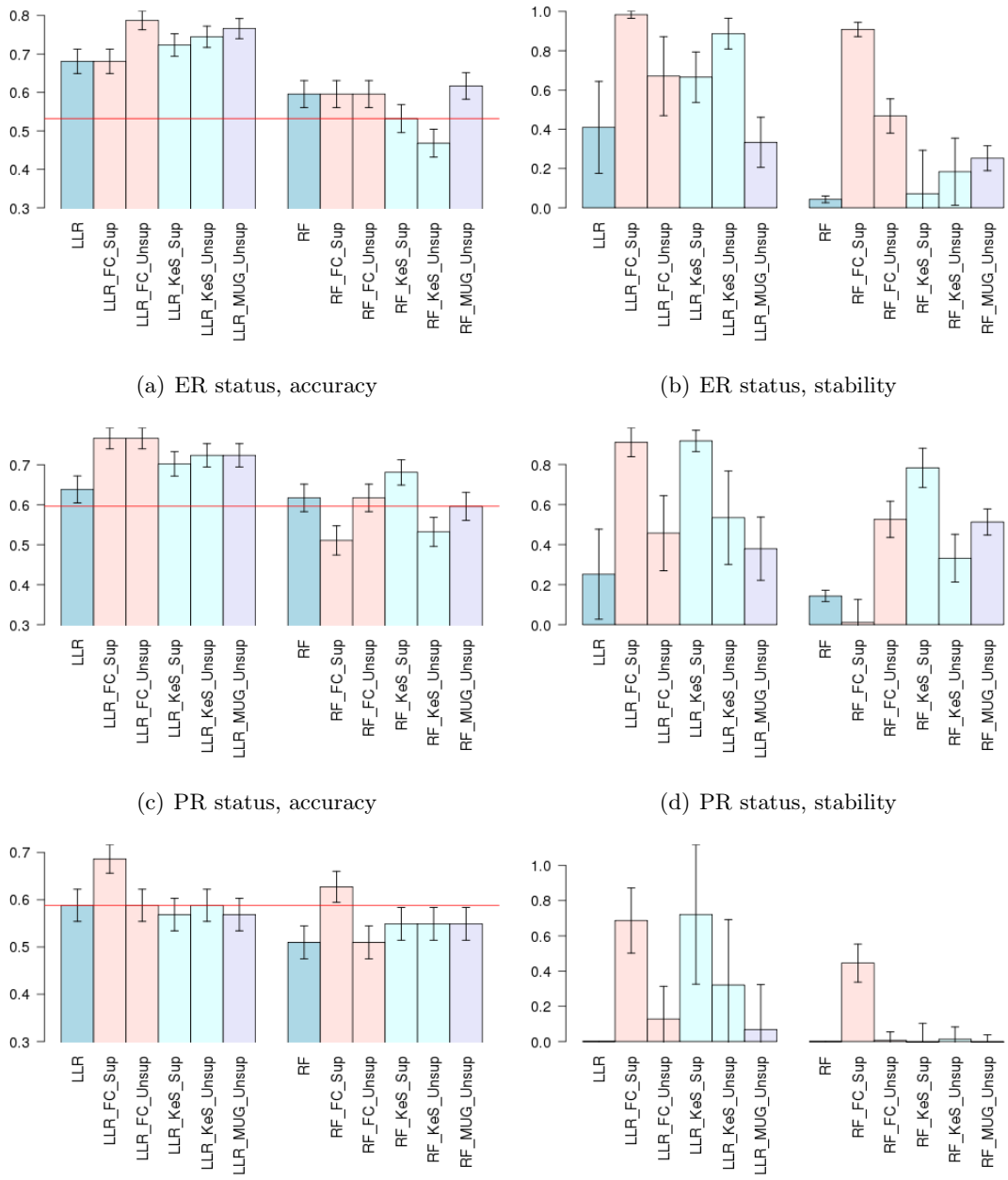
(c)

Figure 6.5.: Consensus feature importance of prediction models on **breast167** dataset with a) ER status, b) PR status and c) Lymph node status.

Figure 6.6 shows the prediction accuracy of all prediction models, as well as the stability of feature importance, as measures of quality of the prediction. It is immediate to observe that LLR models are more accurate than the RF models, which frequently fall frequently below the Bayes error threshold. Given the very small number of samples available, we believe that the poor performance of the RF is due to overfitting.

As seen in the previous two studies (**breast173** and **breast167**), the prediction accuracy is significantly larger than the Bayes error in the case of ER and PR status prediction, confirming that copy number profiles tend to be dissimilar between the two groups of tumors. In contrast, the discrimination between the IBC and NIBC tumors is not very accurate.

In most of the cases, the prediction accuracy of the models with consensus segmentation is comparable with that of the baseline models. Moreover, generally the accuracy of LLR-based models improves after consensus segmentation. The stability



(a) ER status, accuracy

(b) ER status, stability

(c) PR status, accuracy

(d) PR status, stability

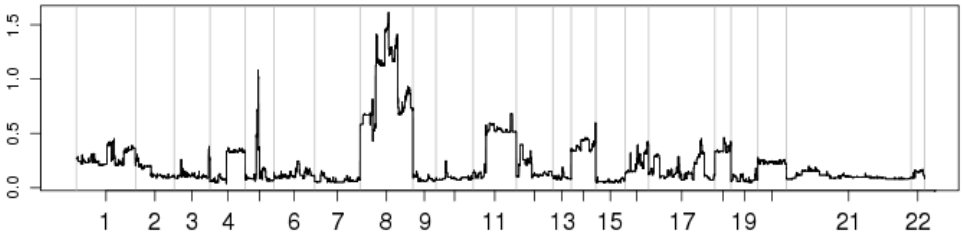
(e) Type (IBC / NIBC), accuracy

(f) Type (IBC / NIBC), stability

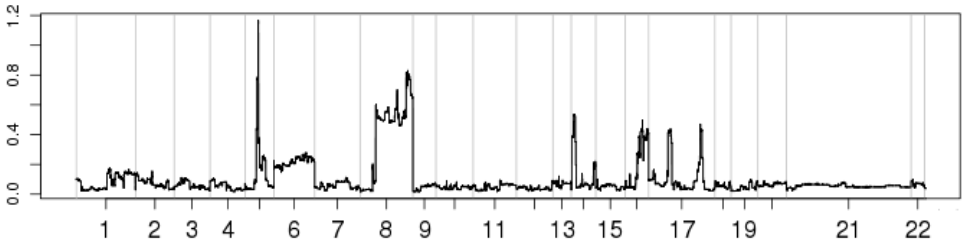
Figure 6.6.: Accuracy of prediction and stability of the feature importance for models trained on the **breast45** dataset. Colors associate with the specific consensus segmentation method used. The two groups of bars correspond to methods based on LLR (left) and RF (right). The red horizontal line on accuracy plots marks the Bayes error.

of feature importance is at least as large as that of the baseline models in all but one model (RF-FC-Sup).

The consensus feature importance assigns the highest value to regions in chromosome 8, specifically 8q gain is more present in ER positive than in ER negative tumors (see Figure A.5 from the appendix for a detailed view). Although indications of relevance of chromosome 8 were present in the **breast167** cohort, it was not a

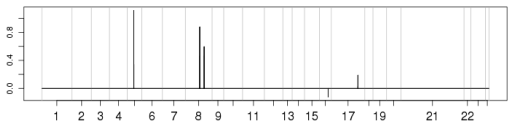


(a)

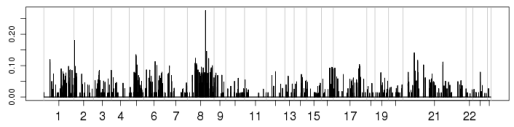


(b)

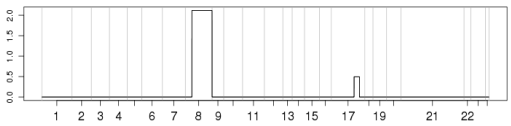
Figure 6.7.: Consensus feature importance of prediction models on **breast54** dataset with a) ER status and b) PR status as outcome.



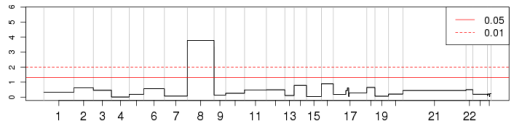
(a) LLR



(b) RF



(c) LLR -FC-Sup



(d) RF-FC-Sup

Figure 6.8.: Consensus segmentation improves model interpretation. Figure shows feature relevance given by a) LLR coefficients, b) RF importance measure, c) LLR-FC-Sup coefficients and d) RF-FC-Sup PIMPp-values. The x-axis orders the features according to chromosome and genomic position. In red, significance thresholds are shown for the RF-FC-Sup p-values.

leading predictor. It is possible that for the particular subtype of ERBB2-amplified cancers, the amplification of 8q is highly predictive, but not in general.

We stress again the advantage of using consensus segmentation for the purpose of model interpretation. In Figure 6.8a, we show the baseline LLR model, which chooses an arbitrary combination of several features for achieving an accuracy of 0.64. For example, the largest absolute relevance is assigned to a probe corresponding to the *CA1* gene (carbon anhydrase, at 8q21.2), which probably plays no direct

role in breast cancer progression and is merely a hitch-hiker. The beneficial effect of consensus segmentation is depicted in Figure 6.8c, in which we show the feature importance given by LLR-FC-Sup. The entire $8q$ arm is selected as relevant, with an increased prediction accuracy to 0.74. The latter predictor is more robust and does not mislead the interpretation by wrongly pointing to one irrelevant marker.

In contrast to LLR, the feature importance given by the RF baseline (Figure 6.8b) is not sparse, neighborhoods sharing similar relevance as a consequence of correlation. Features on chromosome 8 do not stand out as more relevant than the rest. The PIMP-corrected feature importance given by RF-FC-Sup selects the whole chromosome 8 as most predictive and the model achieves the same accuracy as the baseline. Importantly, after consensus segmentation, the LLR and RF models agree.

The LLR-MUG selects a supplementary region located on chromosome 17 (17q23.3 cytoband). Association between ER status and amplification at 17q23.3 has been reported previously in the literature (Han et al., 2006).

There exists a large overlap between the ER and PR status of tumors. Three tumors are ER positive and PR negative, all the rest being either both ER and PR positive or ER and PR negative. In consequence, the predictors are not very different. Amplification at 8q is the most predictive region. Also, all RF-based models point to the deletion of the long arm of chromosome 16 as useful for the prediction (see Figure A.6a).

The most interesting regions indicated as relevant for the prediction of tumor type are located around the ERBB2 amplicon in chromosome 17q. Figure A.6b from the Appendix shows the feature importance given by all models at this location. The majority of methods select the region to the 'left' of ERBB2 locus as predictive, indicating that the the extension of the amplicon over the genes downstream ERBB2 discriminates between the two subtypes. Interestingly, this region is assigned a negative weight by LLR models, which suggests that in fact *not* having a large amplicon makes type IBC more likely. This observation has been made also in the original study (Sircoulomb et al., 2010), but based on visual inspection. The ability of our models to assign importance to the 'length of the aberration' is a direct consequence of the consensus segmentation principle. Indeed, if several consensus breakpoints exist around an aberration, our method breaks the aberration into subregions, which represent essentially the minimal common aberration and several extensions. If some extension is relevant for prediction according to a classification model, then we can draw the conclusion that a larger aberration is associated to the phenotype. In contrast, approaches that only select minimal common regions cannot reveal such associations.

6.2. Prediction of neuroblastoma phenotype

The **neuroblastoma** dataset was introduced in Chapter 4. The 162 samples belong to five stage groups, as follows: 28 Stage 1, 19 Stage 2, 29 Stage 3, 57 Stage 4 and 29 Stage 4S (special). Stages 1 to 3 are localized cancers with generally good prognosis. Stage 4 corresponds to aggressive tumors occurring in children older than one year,

with poor prognosis. Stage 4S is specifically assigned to neuroblastoma tumors that have distant spread (metastasis) to liver, skin or bone marrow and are younger than one year. Stage 4S neuroblastoma is known to have good prognosis.

We used our methodology in order to model tumor phenotype based on copy number profiles, as follows. We considered Stages 1, 2 and 3 as one group of good prognosis neuroblastomas, Stage 4 as the second group and stage 4S as third group. We trained models to discriminate between pairs of groups: Stage 1-3 vs. Stage 4, Stage 1-3 vs. Stage 4S and Stage 4 vs Stage 4S.

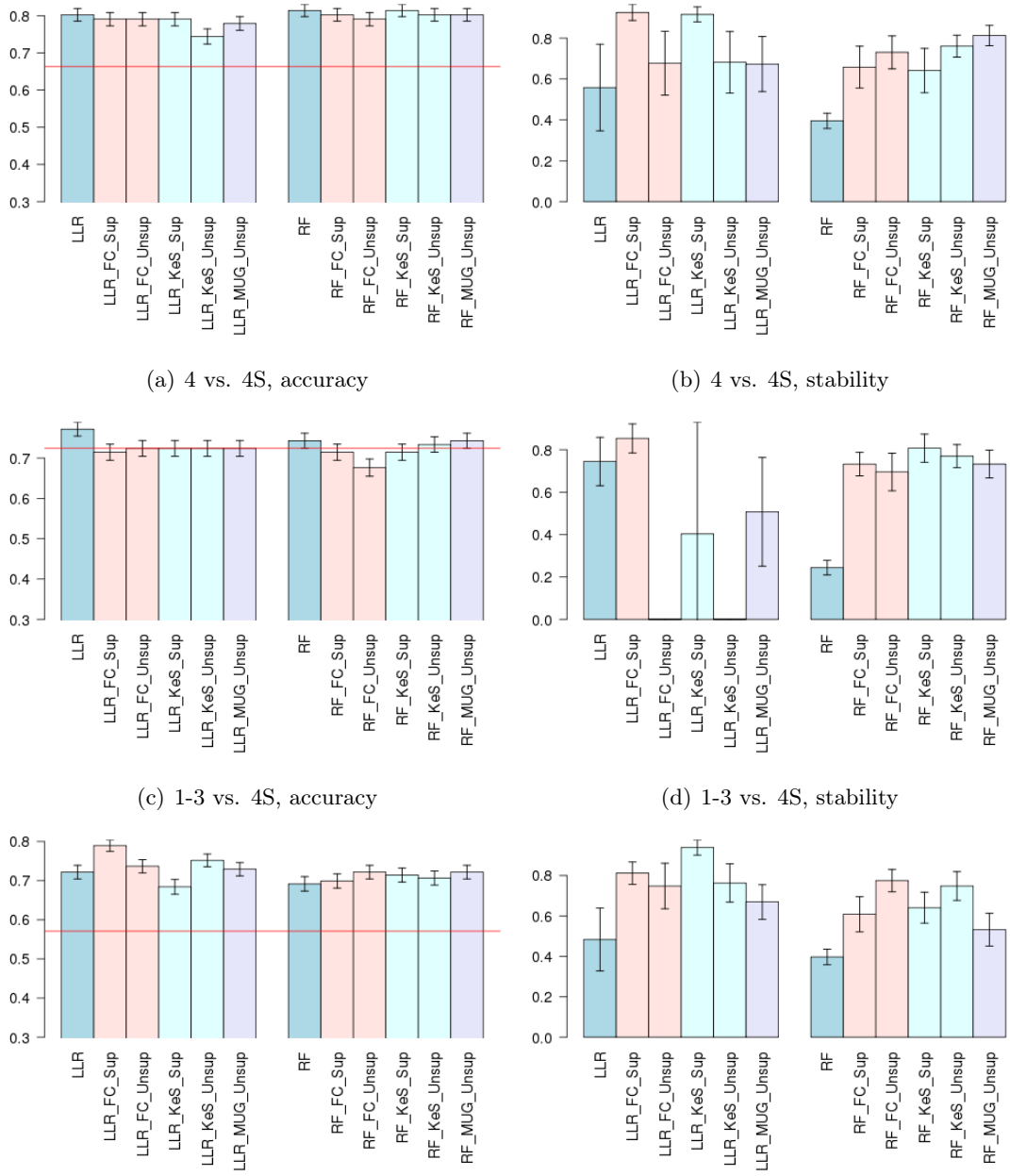
As age is an important prognostic factor for neuroblastoma, we were interested to what extent is age difference represented in the genome of the tumors. In order to coerce the continuous age indicator into a binary response variable, we selected a series of successive cutoffs at 2, 4, 6, ..., 30 and 32 months and for each cutoff, we obtained one binary labeling corresponding to younger and older patients. The age indicator being the age at diagnosis, we admit that there may exist imprecisions in the response variable, due to late detection of the disease, for example.

Figure 6.9 summarized the prediction accuracy and the stability of the feature importance for the cases in which the Stage groups are predicted. From the overall accuracy values, it is immediate that the models can discriminate between Stage 1-3 and Stage 4 tumors with up to 80% accuracy (Figure 6.9c) and between Stage 4 and 4S tumors with up to 80% accuracy (Figure 6.9a). Stage 1-3 tumors and Stage 4S tumors are not easy to dichotomize by our models, the accuracy being comparable to the Bayes error. In this case, LLR-FC-Unsup and LLR-KeS-Unsup are constant, hence the lack of information on stability, which is computed based on Pearson correlation.

In the cases of successful prediction, the models using consensus segmentation yield comparable accuracy (all but two cases, which are outperformed by the baseline). RF-based models achieve comparable accuracy with the LLR-based models. The stability of feature importance increases, as in the previous studies.

Figure 6.10a shows the consensus feature importance for the models that discriminate between Stage 1-3 and Stage 4 tumors. Polyloidy appears to be characteristic of lower Stage tumors, while Stage 4 tumors exibit segmental amplifications and deletions. Figure A.7 shows the gain of chromosome 6, gain of chromosome 22, loss of chromosome 4, gain of chromosome 17 and loss of chromosome 19 as main evidence for abnormal number of chromosome copies characteristic to Stage 1-3 tumors. In contrast, Stage 4 tumors have segmental gains of 6p, loss of 4p, gains of 2p, and amplification of 17q. Our models are in agreement with the study of Spitz et al. (2006).

Figure 6.10b summarizes the features that discriminate between the Stage 4 and Stage 4S tumors. Many of the the highest peaks are also good predictors of the models Stage 1-3 vs. Stage 4 discussed above. Based on the fact that Stage 4S and Stage 1-3 tumors could not be distinguished well, we speculate that they are together very similar and substantially different from Stage 4 tumors. Gains of whole chromosome 4, 6, 17 and losses of whole chromosome 4 and 11 associate with stage 4S and better prognosis. Segmental deletions of 1p, 4p, 11q and amplification of 1q



(a) 4 vs. 4S, accuracy

(b) 4 vs. 4S, stability

(c) 1-3 vs. 4S, accuracy

(d) 1-3 vs. 4S, stability

(e) 1-3 vs. 4, accuracy

(f) 1-3 vs. 4, stability

Figure 6.9.: Accuracy of prediction models trained on the **neuroblastoma** dataset. The vertical bars show accuracy, with confidence intervals resulting from binomial tests. Colors associate with the specific consensus segmentation method used. The two groups of bars correspond to methods based on LLR (left) and RF (right). The red horizontal line marks the Bayes error.

and 17q are associated with stage 4 and worse prognosis (see detailed views of the copy number aberrations in Figure A.8 from the Appendix). All these aberrations and associations have been noted elsewhere (Ambros et al., 2009).

Most our findings suggest that DNA polyloidy plays a role in discriminating between aggressive and non-aggressive forms of neuroblastoma. It has been repeatedly

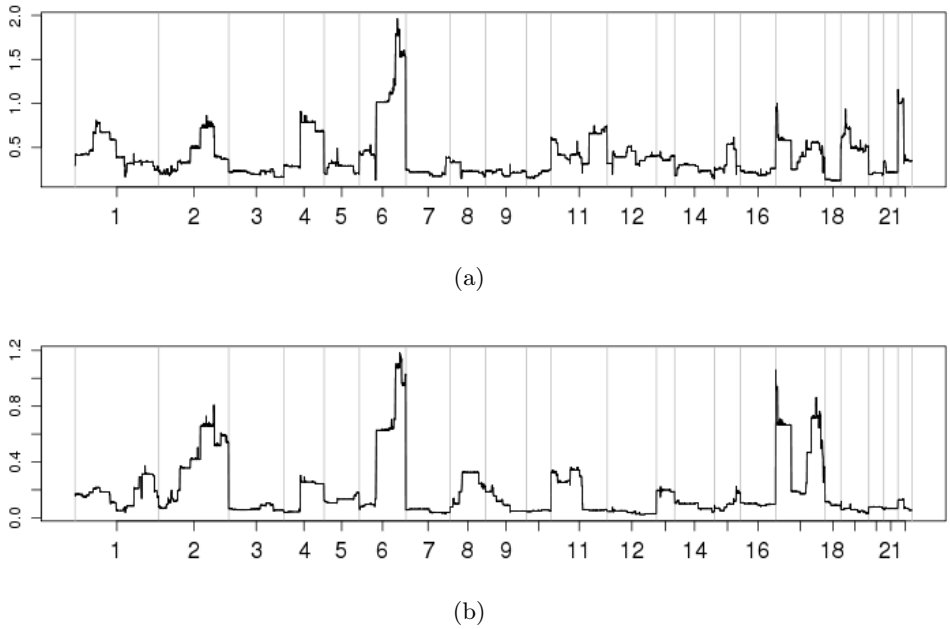


Figure 6.10.: Consensus feature importance of prediction models on **neuroblastoma** dataset with response: a) Stage 1-3 vs Stage 4 and b) Stage 4 vs. Stage 4S 4.

observed that hyperdiploid genomes (with more than two pairs of chromosomes) have a better prognosis and it has been suggested that different mechanisms of progression underlie the formation of the hyperdiploid and diploid tumors. Presently, DNA polyloidy is already used in some clinics for diagnosis and prognosis of neuroblastoma (Ambros et al., 2009).

A very important aspect of model interpretation arises from the examples above. For easy illustration, let us consider the example of chromosome 17 and the task of discriminating between Stage 4 and Stage 4S tumors. Figure 6.11 illustrates the copy number data and the feature importance given by the models that we trained: LLR-based (green) and RF-based (black). Very large importance is given to the copy number of 17p (left in the figure). It is clear from the figure that the lack of amplification or gain at 17p suggests a Stage 4 tumor and thus a worse outcome. Therefore, it is the *lack* of an aberration that associates with a more progressive status, which is unusual in cancer genomics studies. Moreover, searching for causative genes in the discriminative region 17p would be clearly pointless. Even more worrying is the task of classifying a new case: assume the extreme case that there are no copy number aberrations in the genome of this new case, for example corresponding to a healthy tissue. From the point of view of the features located on chromosome 17q, this new case is lacking aberrations and therefore it will be assigned to the more aggressive subtype. Of course, the model was not presented with a control case (healthy), so it cannot know what are its characteristics, however it is unexpected that it is assigned to the more aggressive subtype. This problem is due to the wrong set of features included in the model: the better choice would include the status of the entire chromosome 17 and the status of 17p as two overlapping

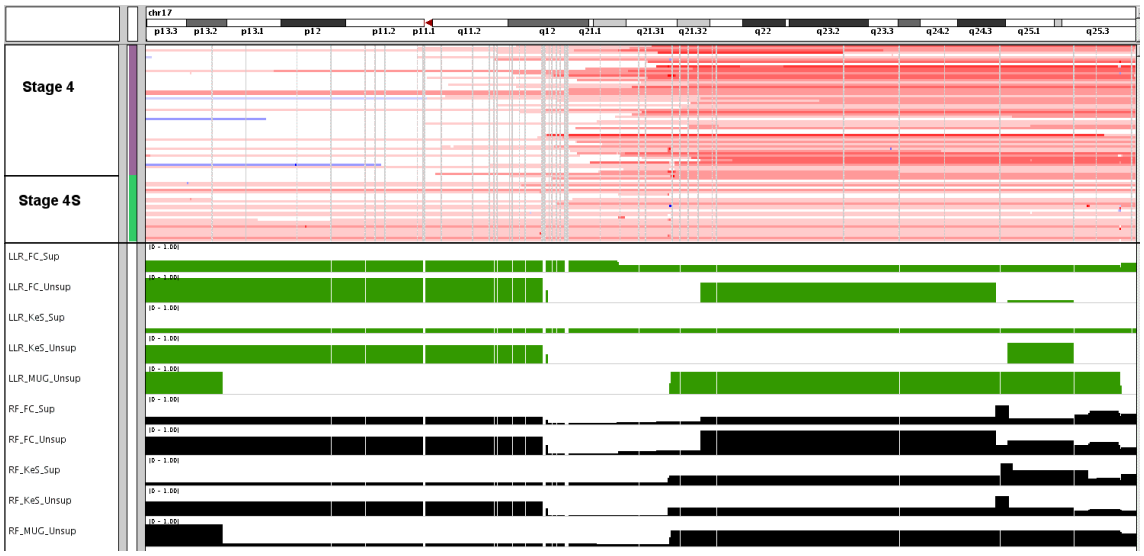


Figure 6.11.: Chromosome 17 of the neuroblastoma cohort, Stages 4 and 4S.

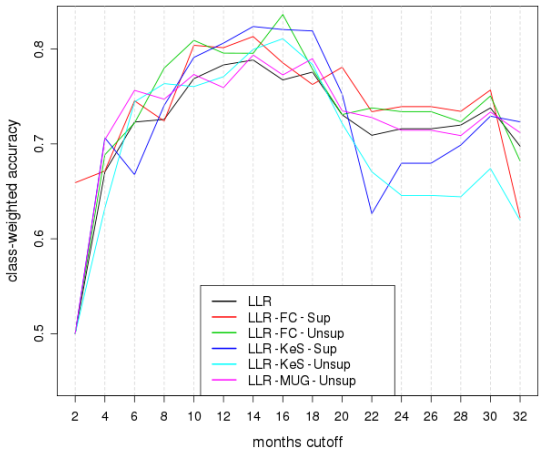
features, motivated by the underlying biology. However, a general and automated selection of features that make sense biologically is not possible presently, unless supplementary knowledge is included in the model. Therefore, we believe that careful model interpretation is required before computer-generated models can be used in the clinics for diagnosis or treatment selection.

The association between age and disease aggressivity is known and used for classification of neuroblastoma. In clinical practice, the age of one year is used for grouping the patients into low risk/good prognosis (younger than one year) and high risk/poor prognosis (older than one year). In Figure 6.12a and 6.12b we show the prediction accuracy of models in dependence of the age cutoff, which is changing from 2 to 32 months. In order to avoid biases due to changing class cardinalities, we computed a weighted accuracy, in which the class-specific accuracies are weighed according to their cardinalities, following the formula:

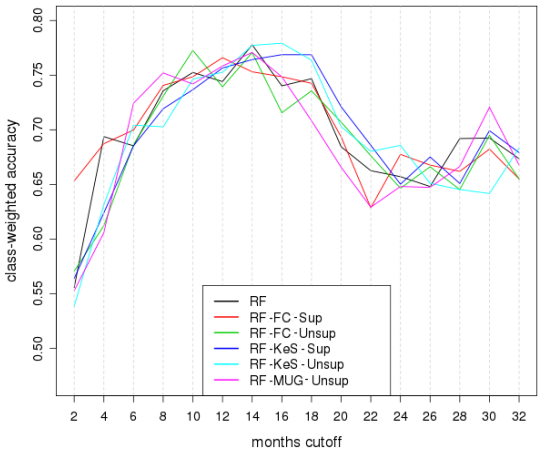
$$\frac{1}{2}\left(\frac{|\{x \in C_1 | \hat{f}(x) = 1\}|}{|C_1|} + \frac{|\{x \in C_0 | \hat{f}(x) = 0\}|}{|C_0|}\right),$$

where C_0 and C_1 are the two response classes and $\hat{f}$ is the prediction function. The weighted accuracy has a concave shape, with a maximum around 14-16 months (Figure 6.12c). This shows that the maximal contrast between two age subgroups is achieved for a cutoff of 14-16 months and not 12 months. Other publications have noted that 18 months may be a more appropriate choice for age cutoff for neuroblastoma classification, motivated for different prognosis (London et al., 2005). Our investigation is the first that supports a higher age cutoff based on maximal genomic differences.

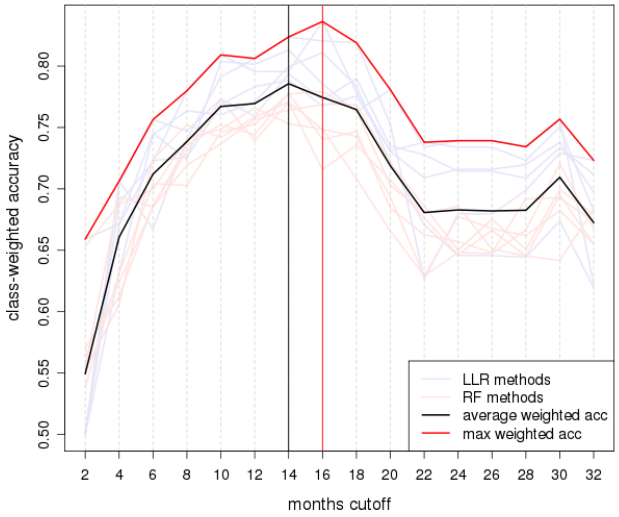
The features that discriminate between patients belonging to different age groups are shown in Figure 6.13. Feature importance is depicted as a function of the age cutoff. We show two LLR-based models and two RF-based models. The smooth



(a) LLR



(b) RF



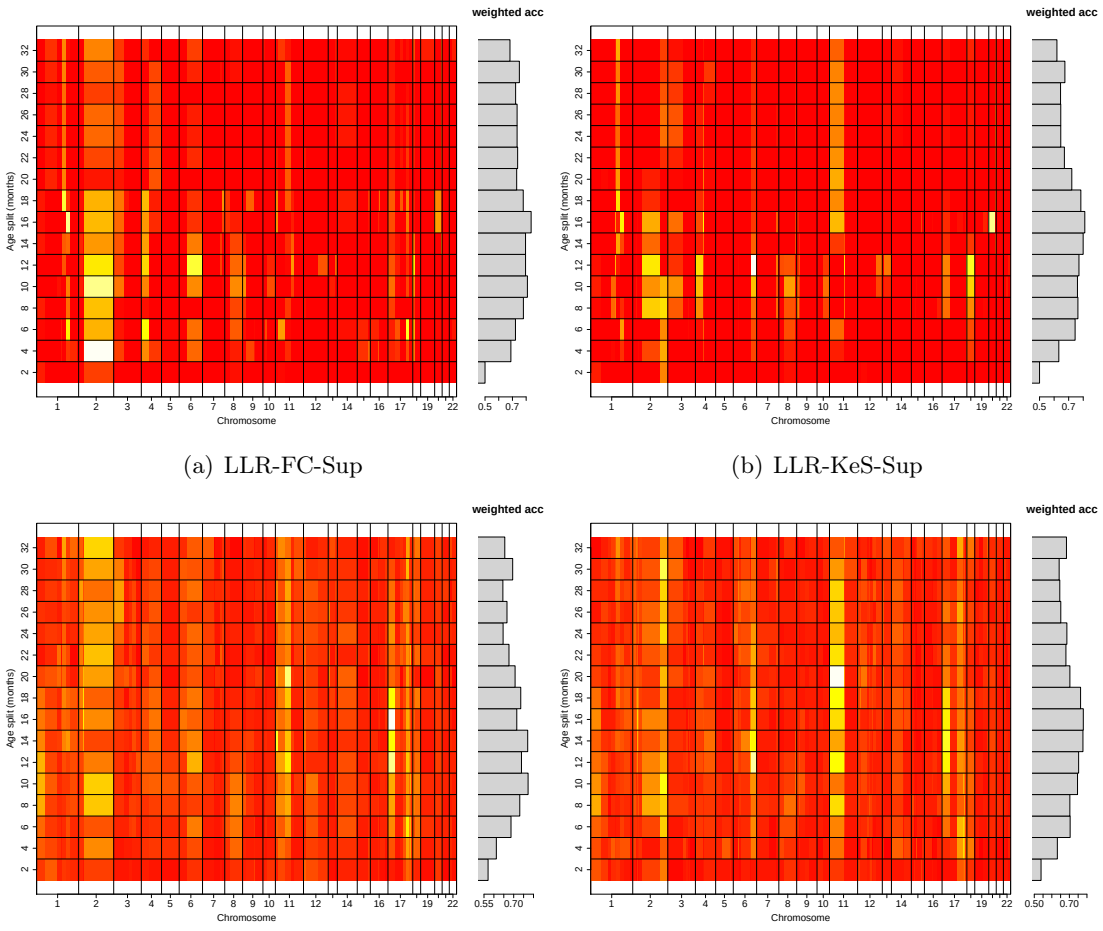
(c) Age-optimal cutoff

Figure 6.12.: Prediction accuracy between various age subgroups. a) LLR-based models, b) RF-based models

transitions between feature importance between consecutive age cutoffs is more evident in RF-based models: features 'enter' the model, then get progressively higher importance and then they 'disappear'. In contrast, the sparse LLR-based models show occasional sharp changes, as for example around the cutoff of 18 months, after which much less features are selected.

Overall, the models agree on the most useful features for discriminating between age subgroups: 1p and 1q, 2q, 6q, 11p, 17p and 17q. The 'predictivity' of these regions peaks at different age cutoffs. For example, chromosome 2q gain discriminates best between patients younger and older than 8–10 months, whereas gain at 17p becomes useful later on, when discriminating between patients younger and older than 12–16 months. The copy number status of 11p is relevant for the separation between patients younger and older than 20 months. Figure A.9 from the Appendix visualizes the copy number arrays sorted by age at diagnosis, with focus on the

interesting chromosomes.



(a) LLR-FC-Sup

(b) LLR-KeS-Sup

(c) RF-FC-Sup

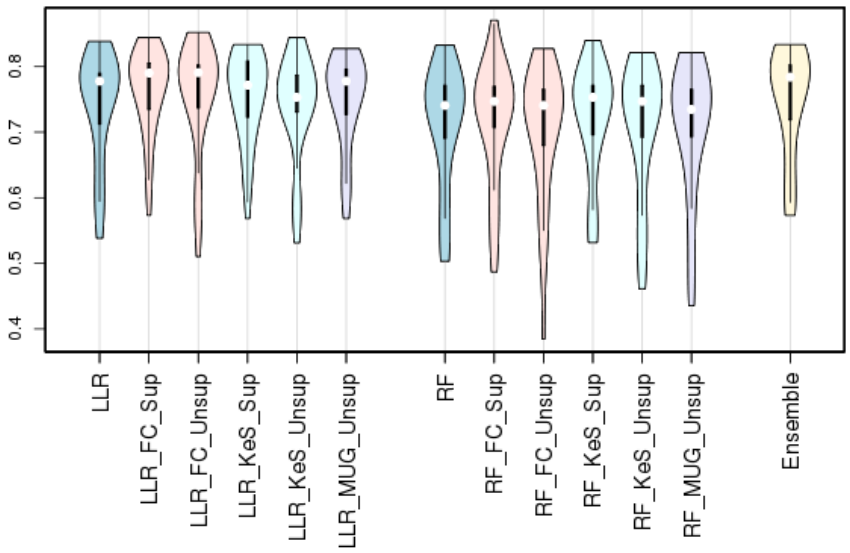
(d) RF-KeS-Sup

Figure 6.13.: Features that discriminate between age subgroups. Dark red corresponds to less significant features, yellow to medium significant and white to very significant features.

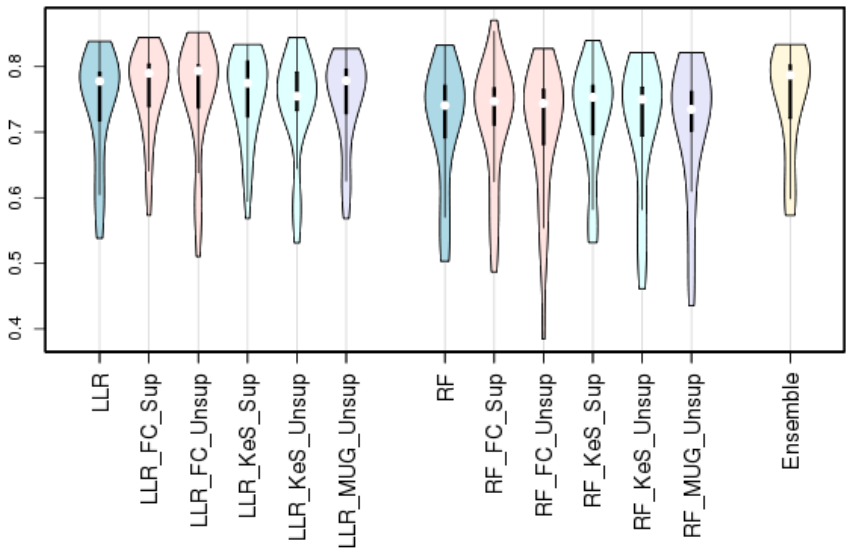
6.3. Discussion

We have presented the results of applying our analysis pipeline to four public arrayCGH datasets with various phenotype annotations (about 15 prediction cases). These results suggest that copy number changes are associated with the phenotypical characteristics of the tumors, some associations being stronger than others. In the cases in which we fail to predict the response variable, it is not clear whether the biological dependence does not exist at all or it is so subtle or complicated that our simple models fail to capture it. However, our predictions were based on relatively small sample sizes and on patient cohorts that could be subject to selection biases, therefore all results should be carefully validated further, for confirmation.

We were not able to find a clear winner among the different consensus segmentation methods, from the point of view of prediction accuracy. In Figure 6.14, we show a summary of all prediction models that we trained, on breast and neurob-



(a) Accuracy



(b) AUC

Figure 6.14.: Comparison of a) accuracy and b) AUC of all models over all datasets (violin plots).

lastoma datasets with various phenotypes. Violin boxplots show that there is no substantial difference between the consensus segmentation approaches. A small but significant increase in accuracy favors the CR-FC method with supervised selection of the number of regions (significantly outperforms LLR, LLR-KeS-Unsup, LLR-MUG-Unsup and all RF-based methods, according to paired Wilcoxon tests). It is interesting that the CB-MUG algorithm, which has best performance with respect to the indicators that we used in Chapter 4 does not help improve classification

accuracy more than the other two consensus segmentation methods. We believe that CB-MUG fails to identify small regions, because the start and end breakpoints of small CNAs do not form clearly two Gaussians. Being too close, they resemble more one Gaussian. Consequently, MUG predicts a breakpoint in the middle of the small region. This type of behavior is not penalized by the measure of segmentation quality $\Omega_{0.98}$, which is robust to outliers (small number of probes that are incorrectly assigned to regions). On the other hand, small regions are sometimes very relevant for prediction, for example focal amplifications. Missing such small regions results in loss of prediction accuracy. The other two segmentation procedures, especially CR-FC, are not as robust as CB-MUG to outliers and identify small regions.

Our experiments show that LLR-based models are significantly more accurate than RF-based models (Figure 6.14). Moreover, when the cohort is really small (54 samples in **breast54**), RF perform really poor. For this reason we believe that RF tend to overfit, which contrasts with the famous claim of the author (Breiman, 2001), that they cannot overfit. Another study by Segal (2004) supports our claim.

In Figure 6.14 we also show the accuracy and AUC of an 'ensemble' model, which returns the majority vote of all participating models. The ensemble model has similar accuracy and AUC as the LLR-based models.

The experiments show that consensus segmentation is beneficial: it affords dimension reduction, higher stability of feature importance and improved interpretability while preserving model accuracy. The LLR-based models are in general different than the RF-based models, even after consensus segmentation. However, some agreement with respect to the feature importance peaks exists.

We commented on the most predictive features of each model and showed that they are confirmed by existing literature. In the case of neuroblastoma, we showed that different age subgroups are characterized by distinct copy number patterns, especially when the groups are defined by a cutoff of 14–16 months, thus confirming that a higher age cutoff than 12 months is more appropriate for diagnosis. We also underlined interesting aspects on how different copy number aberrations are distinct between various age subgroups.

In the case of breast cancer, we fitted models for predicting clinical indicators including ER and PR status. Conflicting feature importance among these models can be explained either by experimental biases or by cohort selection biases.

As supervised learning models based on copy number aberrations are rare in the literature, we believe we are among the first to report on the accuracy of prediction that is achievable based on copy number data only, in the case of breast cancer and neuroblastoma. Predictors based on expression data gives generally good accuracy, a natural next step is to integrate the two data types for improved performance. Our substantial dimension reduction via consensus segmentation sets the premises for data integration.

7. Conclusions and Outlook

This thesis introduced a methodological pipeline for automated prediction of tumor phenotype based on DNA copy number aberrations. The purpose of the pipeline is twofold: first, it provides with a means of automated prediction of the phenotype of new cases based on genotypic information, and second, it can reveal insights into the biological association between the copy number status and phenotype. We discuss here to what extent we succeeded to achieve each of these goals and what is the relevance of our contribution in the context of cancer therapy.

The pipeline that we introduced provides models for supervised classification of tumor phenotype based on genome-wide copy number aberration data. We developed and validated the pipeline on microarray-based data, but from a methodical perspective, the algorithms can be applied to other data types, such as the next generation sequencing data. The pipeline generates classification models which can be considered complex biomarkers. They can be used for automated prediction of important tumor and patient indicators such as lymph node invasion, presence of metastasis, response to a specific treatment type, expected patient survival, etc. Clearly, high prediction accuracy is an imperative requirement for the models in order to enter clinical practice. Our applications to neuroblastoma and breast cancer show that the prediction accuracy is good in general(up to 85%), but not sufficiently high in order to recommend the prediction models for immediate practical use. Below we explain why we think the accuracy is not good enough.

First, the current knowledge on the molecular mechanisms of cancer (partly summarized in this thesis in Chapter 2) indicates that numerical aberrations are only one player in the process of cancer progression, and other players such as epigenetic changes, structural aberrations or mutations are tightly interconnected. It is therefore safe to assume that we base our prediction on incomplete information, which could account for a fraction of the misclassification rate. Second, our models may not have succeeded in capturing existing associations between the copy number profiles and the phenotype. Indeed, for the purpose of interpretability and in order to avoid overfitting, we were forced to choose rather simple models such as linear logistic regression, and thus sacrifice the expressiveness of the models. We attempted to compensate with more complex but interpretable models like the random forest, but experiments indicate that they tend to overfit and were in general outperformed by logistic regression. Third, there is compelling evidence that microarray batch effects have a negative influence on the prediction, demonstrated in this thesis by the large differences in accuracy of progesterone receptor status in different breast cancer cohorts. A related problem is the lack of careful planning in the selection of the patients participating in different public data cohorts: often, the most clinically

'interesting' cases are selected, which constitute statistical outliers and make classification difficult. A carefully planned clinical trial is expected to result in higher accuracy.

A notable achievement is the usage of prediction accuracy for selecting an optimal age cutoff for the diagnosis of neuroblastoma patients. Based on current clinical practice, childred younger than one year are assigned to a good prognostic group and get well with minimal therapy. Patients older than one year are assigned to a poor prognostic group and receive agressive chemotherapy. Recent clinical studies show that even patients up to 16-18 months of age tend to have recessive tumors, thus a higher cutoff of about 16-18 months would spare the evere side effects of chemotherapy for many children. We used the prediction accuracy of our models as an indicator of genomic difference and showed that groups separated by an age cutoff of 14-16 months are most distinct. We thereby provide with the first evidence supporting a different age grouping based solely on copy number data.

Our models are among the very few fully supervised prediction models based on copy number aberrations. Some models exist on breast cancer but to the best of our knowledge, ours are the first models on neuroblastoma data. The accuracies of our models have therefore a state-of-the-art value, challenging the bioinformatics community to propose better methods.

We were more successful in achieving the second goal of the computational pipeline. Namely, we showed on simulated and read data that the interpretability and robustness of our models are superior to that of baseline models. For this purpose we used feature grouping, an efficient method for dimension reduction and for reducing correlation between features, thus improving feature ranking and the stability of the feature importance. Our contribution is important, because the instability of feature ranking against variation of the training set is a well known problem and a reason of concern in the field of microarray classification, leading to lack of reproducibility and mistrust from the medical community. Moreover, since the instability is caused by the high dimensionality and the biological heterogeneity of the tumor samples, it is likely that the problem will persist in the analysis of DNA high-throughput sequencing data.

The key concept of this thesis, namely feature grouping, is not our innovation, as it has been applied in other forms to different data types. We adopt the main principle, that of grouping features based on some criterion and then constructing representatives or super-features that can be used for more meaningful modeling. Technically, the feature grouping methods that we presented in the thesis are tailored to the specific structure of DNA copy number profiles, specifically we speculate on the property of the log-ratios of being locally constant. We called this approach 'consensus segmentation'. We introduced two approaches for consensus segmentation, first based on identification of breakpoint hotspots (consensus breakpoints) and one based on identifying regions of almost constant copy number (consensus regions). Three corresponding algorithms were described: CB-MUG (consensus breakpoints via mixture of Gaussians and one Uniform), CB-KeS (consensus breakpoints via kernel smoothing) and CR-FC (consensus regions via feature clustering). A comparison

of these approaches favors slightly the CB-FC method, which generally improves the accuracy of the baseline model if used together with lasso-penalized logistic regression or random forest.

In Chapter 6, we commented on the most relevant aberrations that play an important role in predicting various tumor phenotypes of neuroblastomas and breast cancers. We gained important insight into the biological mechanisms of copy number change that associate with various phenotype indicators. Two very interesting examples demonstrate the strengths of our approach. First, the model for prediction of tumor subtype based on the breast cancer **breast54** dataset indicated that the size of the amplicon around ERBB2 gene is an important predictor and it did so by selecting the features located on extensions of the amplicon region in the neighborhood and not at the gene itself. Most of the methods for copy number analysis focus on the minimal common region of aberration and thus would fail to discover the importance of their elongations. The consensus segmentation typically considers the minimal common region and its elongations as different regions, for more expressive modeling. The second example refers to the classification of neuroblastomas, in which the ploidy status plays an important role. In neuroblastoma, chromosome 17q amplification is associated with poor prognosis and whole chromosome 17 gain is associated with good prognosis. In consequence, our models select 17p as a very relevant region for classification. It is unlikely that this region contains oncogenes that are involved in neuroblastoma progression. Rather, it is the different molecular mechanisms causing polyploidy and aplification that explain the distinct phenotype. Thus, in contrast to our approach, most methods that are driven by the search for causative genes or for segmental aberrations will fail to discover a very predictive feature.

A different perspective on the above mentioned model on neuroblastoma is that, even though the model indicates that 17p is a discriminative feature, it takes careful interpretation and expertise in order to understand that it is a consequence of the 17q amplification/17 polyploidy. We thereby conclude that model interpretation is of great importance and should be a priority not only to the biomedical community, but to the bioinformatics community as well.

Outlook

Our applications were limited by the availability of phenotypic annotation in public datasets. For this reason, we were not able to carry out prediction of drug response, treatment outcome or patient survival, which are of major importance for improved cancer therapy. Also, a necessary step forward is to apply the automated pipeline to as many cancer types as possible, both for the purpose of validation, as well as for gaining possibly novel biological insight into the association between numerical aberrations and phenotype. Clearly, larger datasets are also necessary.

We payed special attention to the interpretability and stability of our prediction models because these are properties that can make computational models attractive to the biomedical community. However, an interesting approach would be to train

very powerful and complex classification models, in order to compare their prediction accuracy with that of our models.

Following the greater goal of introducing computational models into clinical practice, integration of copy number aberration data with other data types such as somatic mutations, epigenetic modifications and transcriptomic data is needed. Methodologically, integration of omics data is still a challenging problem, facing the high dimensionality of each data type and the heterogeneity of the omics data. We believe that the consensus segmentation of copy number profiles that we proposed ensure substantial dimension reduction with no loss of important information, which suggests this procedure as preprocessing step to data integration.

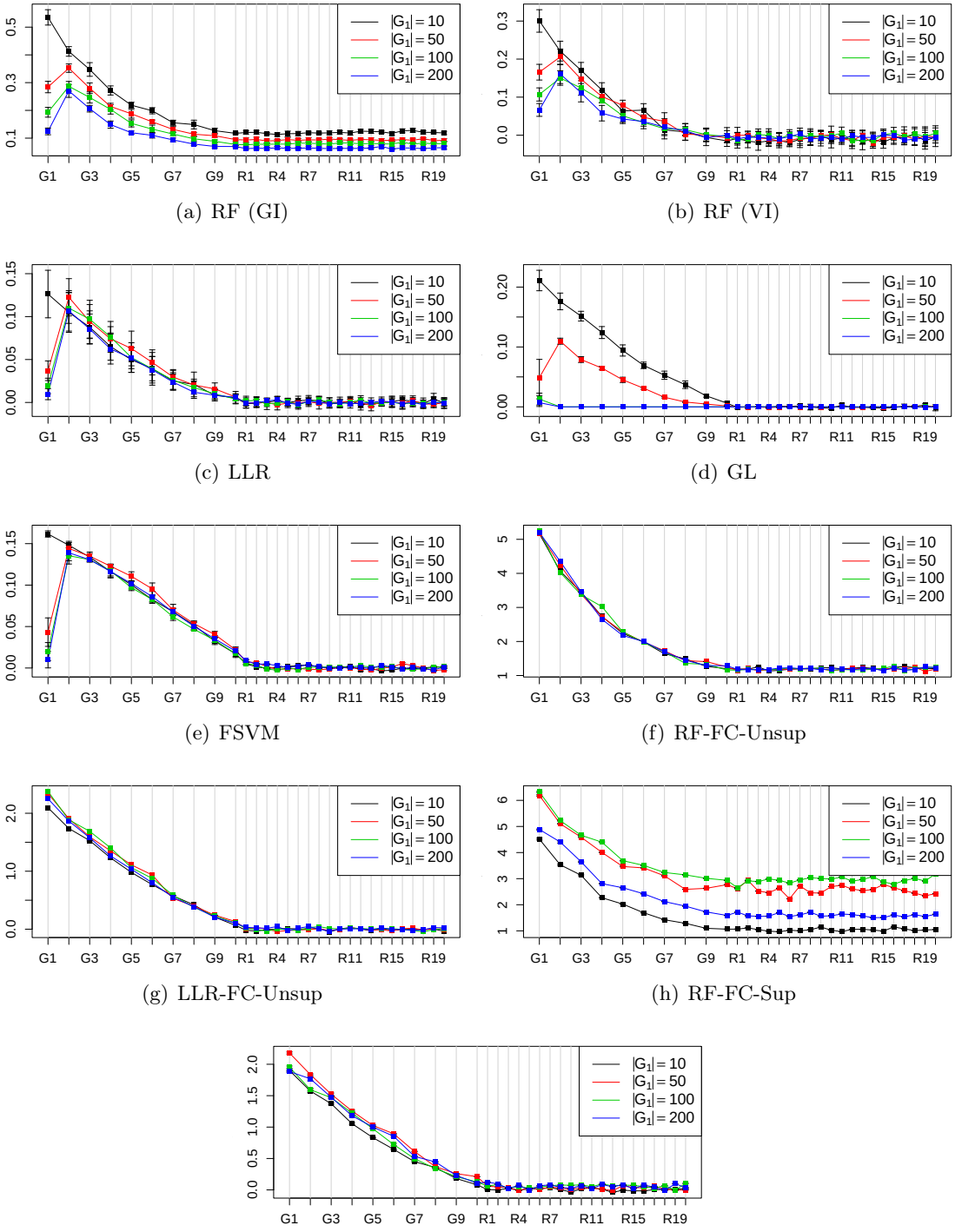
Introducing complex biomarkers – like the classification models presented in this thesis – into clinical practice is a difficult endeavor. The lack of standards, the poor validation or lack of reproducibility have been used as arguments against such models. Interpretability remains a major requirement, as the biomedical community appears more willing to accept a transparent, rule-based decision process, than a black box oracle, regardless of how accurate. We believe the work presented in this thesis has succeeded, at least conceptually, to get closer to these requirements.

In the future, the computational community needs to search for better ways of presenting complex statistical models to the biomedical community and the biomedical community needs to accept the help of non-intuitive but powerful computational instruments, for the benefit of the patient.

A. Appendix

A.0.1. Supplementary figures and tables

In the following plots, we used the IGV application (Robinson et al., 2011) for visualizing the copy number data and the feature importance simultaneously.



(a) RF (GI)

(b) RF (VI)

(c) LLR

(d) GL

(e) FSVM

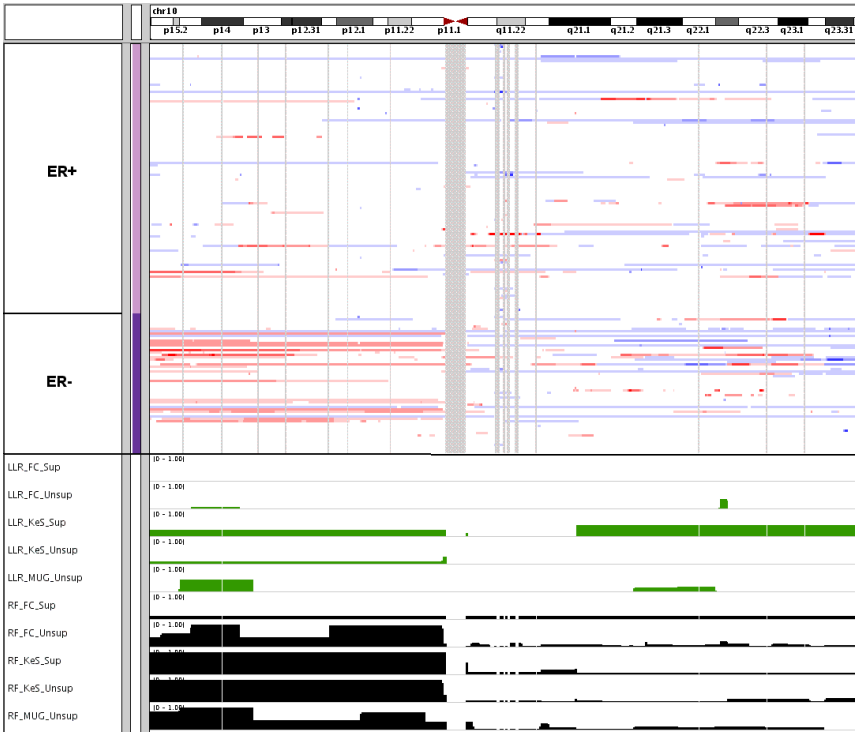
(f) RF-FC-Unsup

(g) LLR-FC-Unsup

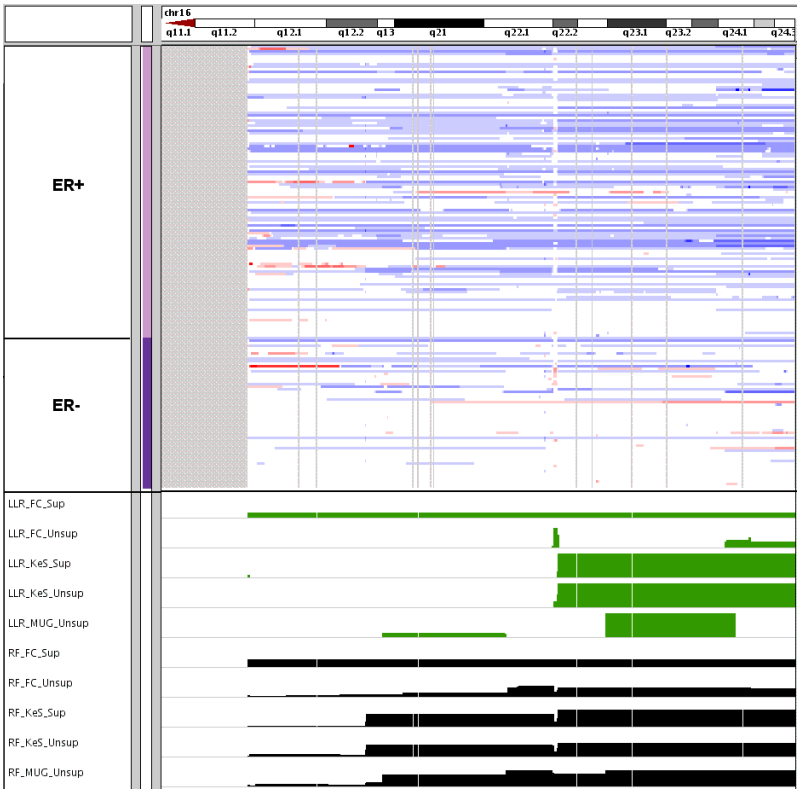
(h) RF-FC-Sup

(i) LLR-FC-Sup

Figure A.1.: Average importance of features for classification of data from Simulation B. The importance is averaged over groups $G_1, ..., G_{10}, R_1, ..., R_{20}$.

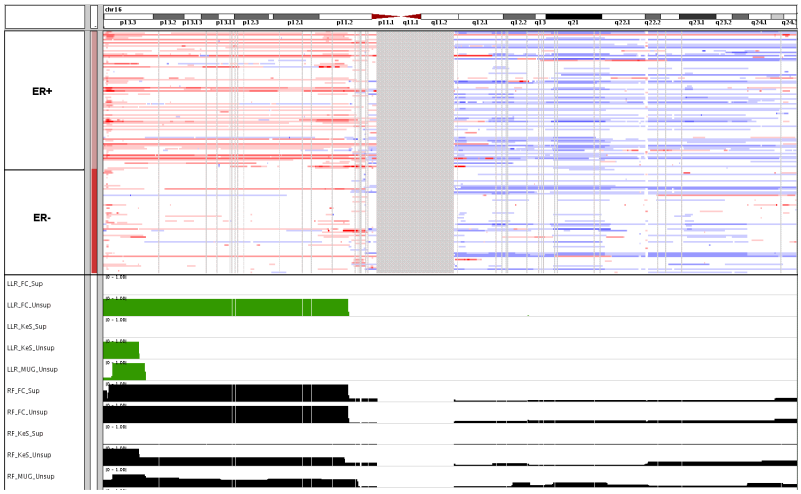


(a) ER status

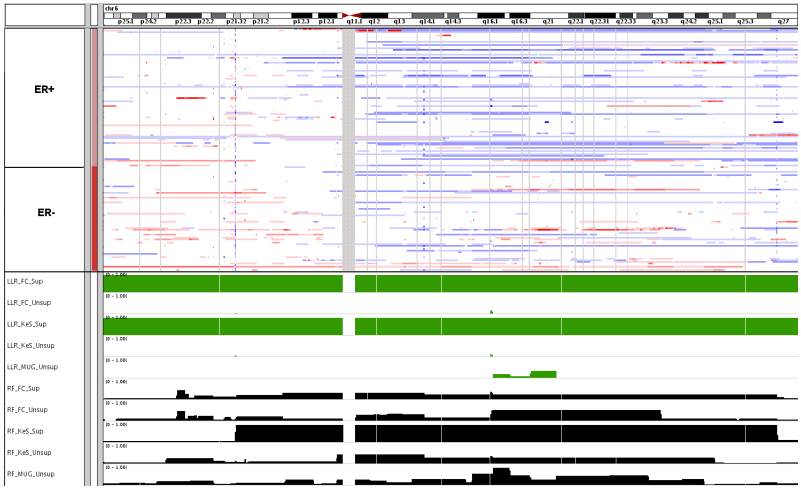


(b) PR status

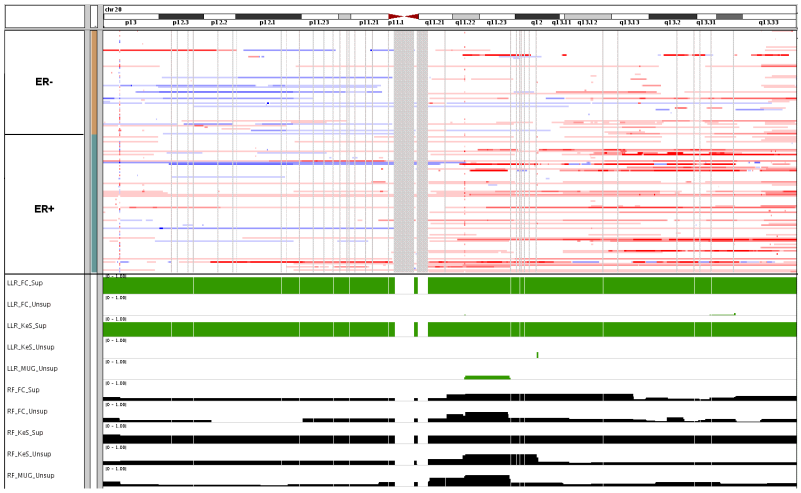
Figure A.2.: Feature importance of prediction models on breast173 dataset with ER status outcome: a) chromsome 10, b) chromosome 16.



(a)

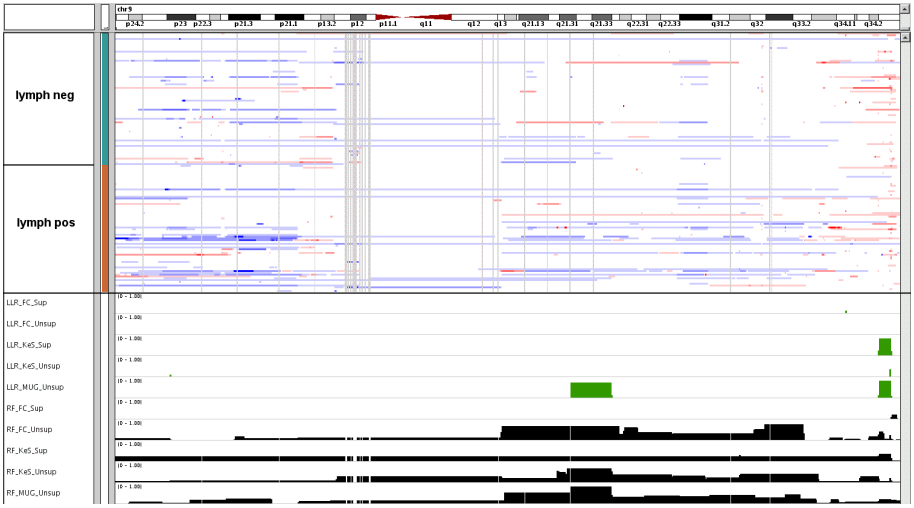


(b)

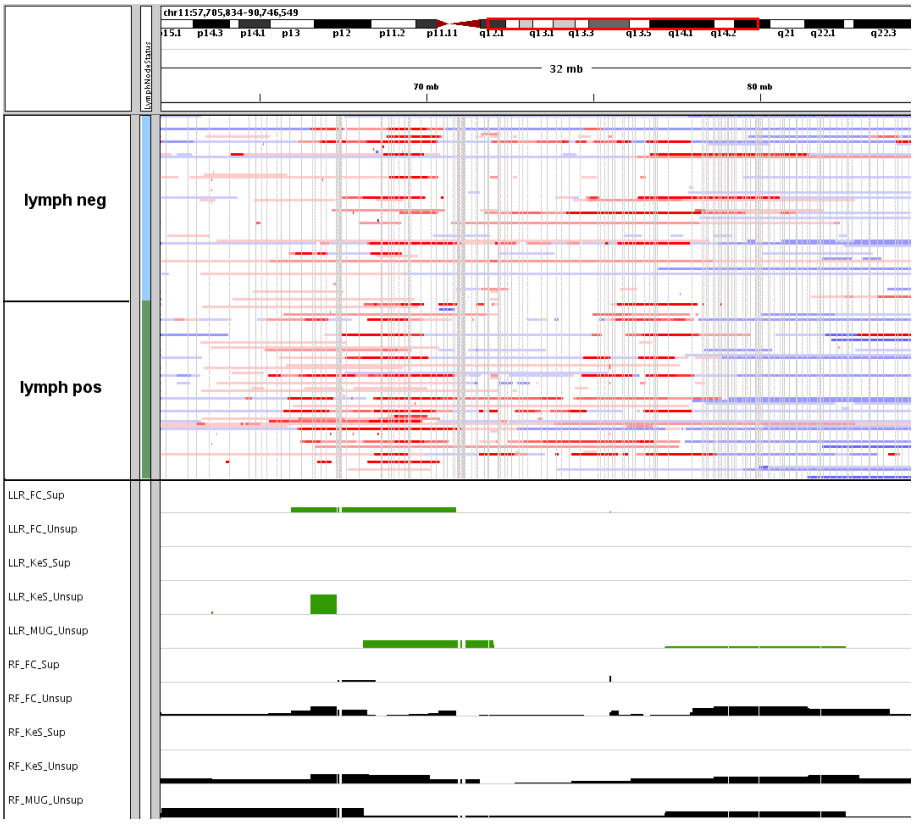


(c)

Figure A.3.: Feature importance given by models trained on **breast167** dataset with ER outcome: a) chromosome 16, b) chromosome 6 and c) chromosome 20.

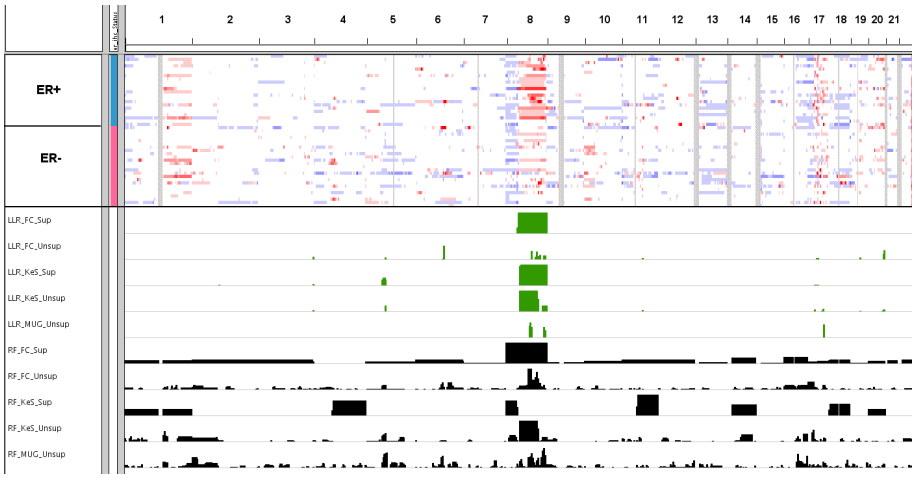


(a)

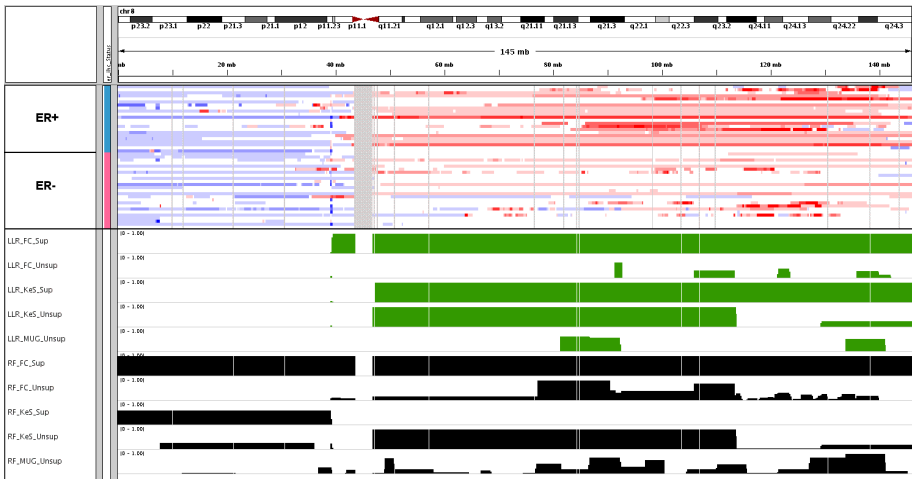


(b)

Figure A.4.: Feature importance given by models trained on **breast167** dataset with lymph status outcome: chromosome 11.

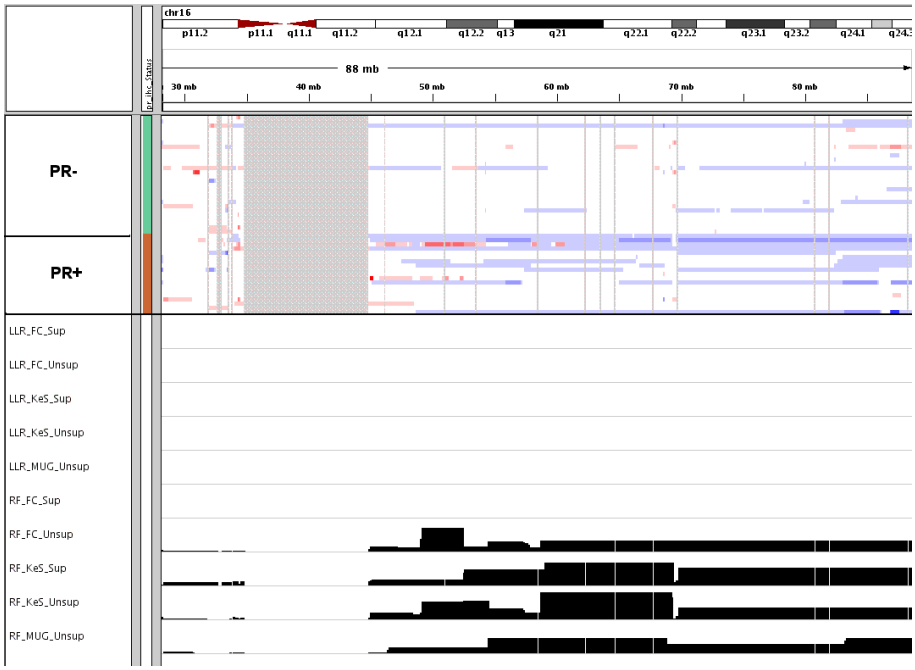


(a)

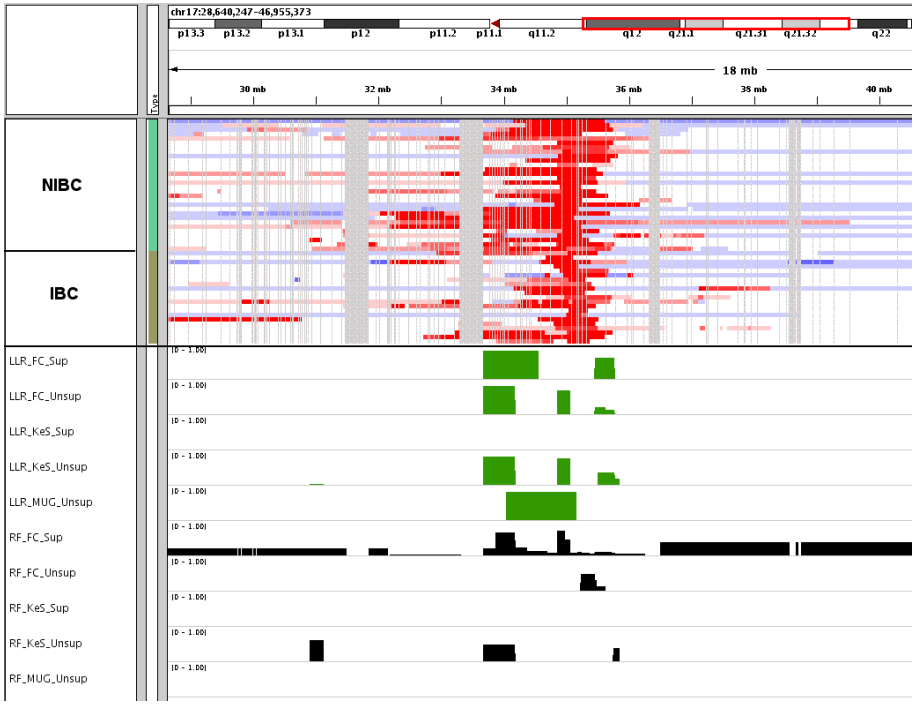


(b)

Figure A.5.: IGV view of feature importance given by models trained on the breast54 dataset with ER status outcome: a) genome-wide view; b) focus on chromosome 8. On the top, the log-ratios are represented, with intense red corresponding to amplifications and dark blue for deletions.



(a)



(b)

Figure A.6.: Predictors of a) PR status on chromosome 16 and b) of tumor type on chromsome 17 from the **breast54** dataset.

(a) (b)

(c) (d)

(e) (f)

Figure A.7.: Regions that discriminate between Stages 1-3 and Stage 4 neuroblastoma: a) chromosome 6, b) chromosome 22, c) chromosome 4, d) chromosome 2, e) chromosome 17 and f) chromosome 19.

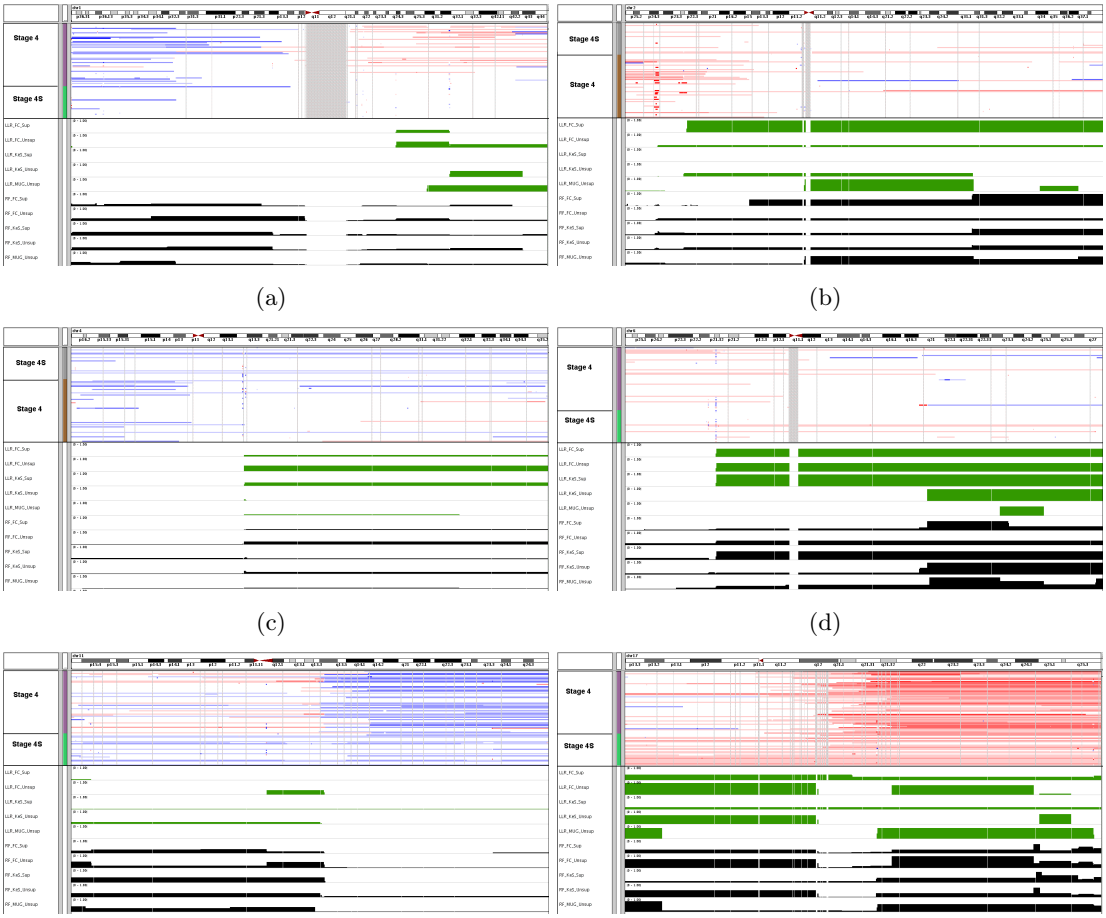


Figure A.8.: Regions that discriminate between Stage 4 and Stage 4S neuroblastoma a) chromosome 1, b) chromosome 2, c) chromosome 4, d) chromosome 6, e) chromosome 11 and f) chromosome 17.

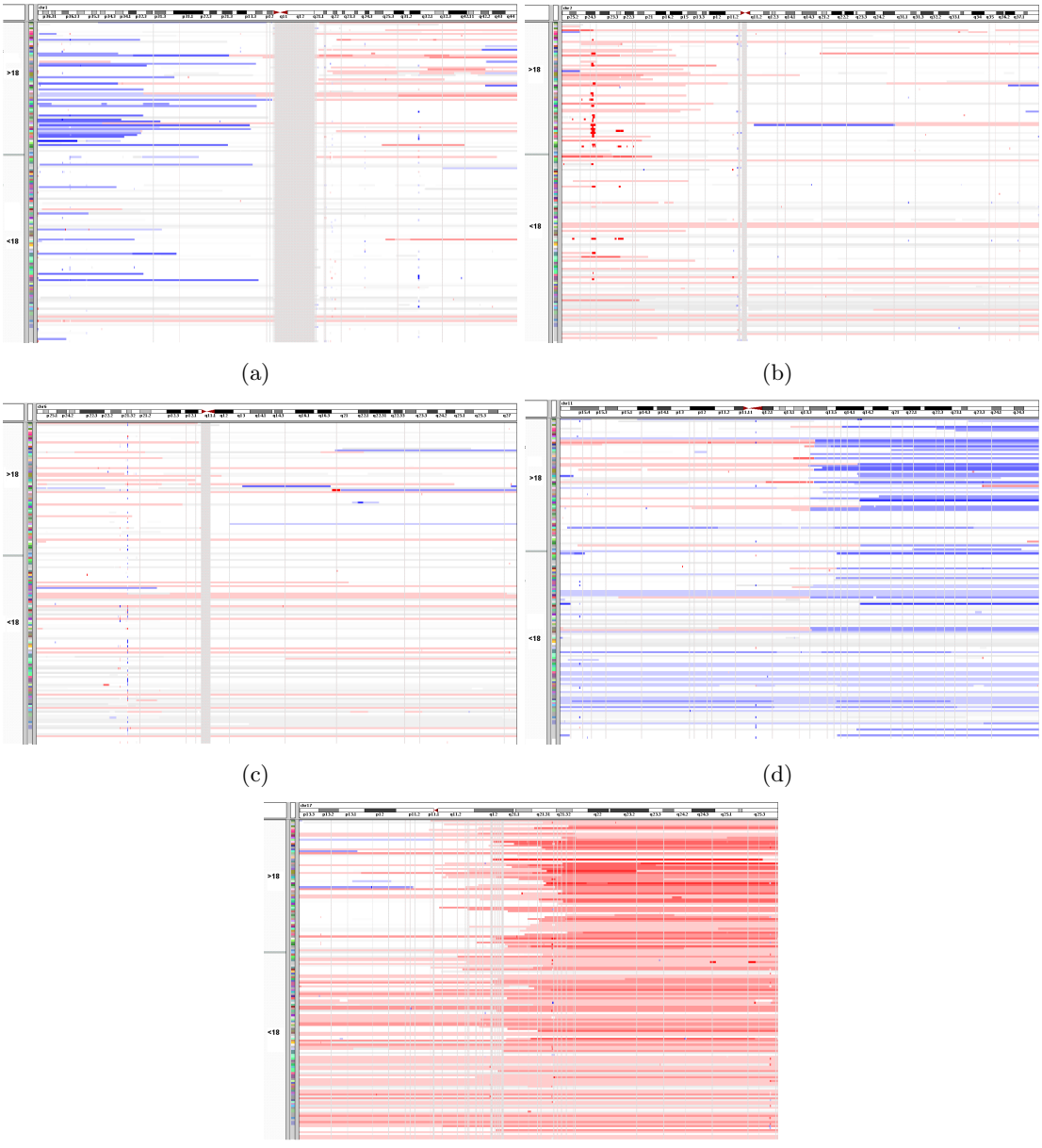


Figure A.9.: Copy number arrays sorted by age of the patient at diagnosis: a) chromosome 1, b) chromosome 2, c) chromosome 6, d) chromosome 11, e) chromosome 17.

Bibliography

Abeysinghe, S., Chuzhanova, N., Krawczak, M., Ball, E., and Cooper, D. (2003). Translocation and gross deletion breakpoints in human inherited disease and cancer I: Nucleotide composition and recombination-associated motifs. *Hum Mutat*, 22(3):229–244.

Achard, S., Pham, D., and Jutten, C. (2005). Criteria based on mutual information minimization for blind source separation in post nonlinear mixtures. *Signal Process.*, 85:965–974.

Agarwal, S., Tafel, A., and Kanaar, R. (2006). Dna double-strand break repair and chromosome translocations. *DNA Repair (Amst)*.

Aguirre, A., Brennan, C., Bailey, G., Sinha, R., Feng, B., Leo, C., Zhang, Y., Zhang, J., Gans, J., Bardeesy, N., Cauwels, C., Cordon-Cardo, C., Redston, M., Depinho, R., and Chin, L. (2004). High-resolution characterization of the pancreatic adenocarcinoma genome. *Proceedings of the National Academy of Sciences of the United States of America*, 101(24):9067–9072.

Allison, D., Cui, X., Page, G., and Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, 7(1):55–65.

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347.

Ambros, P., Ambros, I., Brodeur, G., Haber, M., Khan, J., Nakagawara, A., Schleiermacher, G., Speleman, F., Spitz, R., London, W., Cohn, S., Pearson, A., and Maris, J. (2009). International consensus for neuroblastoma molecular diagnostics: report from the international neuroblastoma risk group (INRG) biology committee. *Br J Cancer*, 100(9):1471–1482.

André, F., Job, B., Dessen, P., Tordai, A., Michiels, S., Liedtke, C., Richon, C., Yan, K., Wang, B., Vassal, G., Delaloge, S., Hortobagyi, G. N., Symmans, W. F., Lazar, V., and Pusztai, L. (2009). Molecular characterization of breast cancer with high-resolution oligonucleotide comparative genomic hybridization array. *Clinical Cancer Research*, 15(2):441–451.

Attiyeh, E., London, W., Mossï£¡, Y., Wang, Q., Winter, C., Khazi, D., McGrady, P., Seeger, R., Look, A., Shimada, H., Brodeur, G., Cohn, S., Matthay, K., and Maris, J. (2005). Chromosome 1p and 11q deletions and outcome in neuroblastoma. *N Engl J Med*, 353(21):2243–53.

Bach, F. R. (2008). Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th international conference on Machine Learning*, ICML '08, pages 33–40. ACM.

Bagchi, A. and Mills, A. (2008). The quest for the 1p36 tumor suppressor. *Cancer Res*, 68(8):2551–6.

Ballestar, E., Paz, M., Valle, L., Wei, S., Fraga, M., Espada, J., Cigudosa, J., Huang, T., and Esteller, M. (2003). Methyl-cpg binding proteins identify novel sites of epigenetic inactivation in human cancer. *EMBO J*, 22(23):6335–45.

Balmain, A. (2001). Cancer genetics: from boveri and mendel to microarrays. *Nat Rev Cancer*, 1(1):77–82.

Barlogie, B., Drewinko, B., Schumann, J., Góhde, W., Dosik, G., Latreille, J., Johnston, D., and Freireich, E. (1980). Cellular dna content as a marker of neoplasia in man. *The American Journal of Medicine*, 69(2):195 – 203.

Barutcuoglu, Z., Airoldi, E., Dumeaux, V., Schapire, R., and Troyanskaya, O. (2009). Aneuploidy prediction and tumor classification with heterogeneous hidden conditional random fields. *Bioinformatics*, 25(10):1307–1313.

Battiti, R. (1994). Using Mutual Information for Selecting Features in Supervised Neural Net Learning. *IEEE Trans. Neural Networks*, 5(4):537–550.

Bekhouche, I., Finetti, P., Adelaide, J., Ferrari, A., Tarpin, C., Charafe-Jauffret, E., Charpin, C., Houvenaeghel, G., Jacquemier, J., Bidaut, G., Birnbaum, D., Viens, P., Chaffanet, M., and Bertucci, F. (2011). High-resolution comparative genomic hybridization of inflammatory breast cancer and identification of candidate genes. *PLoS ONE*, 6(2).

Ben-Dor, A., Lipson, D., Tsalenko, A., Reimers, M., Baumbusch, L., Barrett, M., Weinstein, J., Børresen-Dale, A.-L., and Yakhini, Z. (2007). Framework for identifying common aberrations in dna copy number data. In *Proceedings of the 11th annual international conference on Research in computational molecular biology*, RECOMB'07, pages 122–136.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Bergamaschi, A., Kim, Y., Kwei, K., Choi, Y., Bocanegra, M., Langerød, A., Han, W., Noh, D., Huntsman, D., Jeffrey, S., Børresen-Dale, A., and Pollack, J. (2008). Camk1d amplification implicated in epithelialâĂŞmesenchymal transition in basal-like breast cancer. *Molecular Oncology*, 2(4):327 – 339.

Bergamaschi, A., Kim, Y., Wang, P., Sørlie, T., Hernandez-Boussard, T., Lonning, P., Tibshirani, R., Børresen-Dale, A.-L., and Pollack, J. (2006). Distinct patterns of DNA copy number alteration are associated with different clinicopathological

features and gene-expression subtypes of breast cancer. *Genes, Chromosomes and Cancer*, 45(11):1033–1040.

Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., Vivanco, I., Lee, J., Huang, J., Alexander, S., Du, J., Kau, T., Thomas, R., Shah, K., Soto, H., Perner, S., Prensner, J.and Debiasi, R. M., Demichelis, F., Hatton, C., Rubin, M., Garraway, L., Nelson, S., Liau, L., Mischel, P., Cloughesy, T., Meyerson, M., Golub, T., Lander, E., Mellinghoff, I., and Sellers, W. (2007). Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proceedings of the National Academy of Sciences*, 104(50).

Bezdek, J. C. and Pal, N. R. (1998). Some new indexes of cluster validity. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, 28(3):301–315.

Blaveri, E., Brewer, J., Roydasgupta, R., Fridlyand, J., DeVries, S., Koppie, T., Pejavar, S., Mehta, K., Carroll, P., Simko, J., and Waldman, F. (2005). Bladder cancer stage and outcome by array-based comparative genomic hybridization. *Clinical Cancer Research*, 11(19):7012–7022.

Bourguignon, F. (1979). Decomposable income inequality measures. *Econometrica*, 47(4).

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C., and Olshen, R. (1984). *Classification and Regression Trees*. Chapman & Hall/CRC.

Brodeur, G., Seeger, R., Schwab, M., Varmus, H., and Bishop, J. (1984). Amplification of n-myc in untreated human neuroblastomas correlates with advanced disease stage. *Science*, 224(4653):1121–4.

Brunet, J.-P., Tamayo, P., Golub, T., and Mesirov, J. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences of the United States of America*, 101(12):4164–4169.

Cairns, P., Evron, E., Okami, K., Halachmi, N., Esteller, M., Herman, J., Bose, S., Wang, S., Parsons, R., and Sidransky, D. (1998). Point mutation and homozygous deletion of pten/mmac1 in primary bladder cancers. *Oncogene*, 16(24):3215–8.

Campbell, P., Stephens, P., Pleasance, E., O'Meara, S., Li, H., Santarius, T., Stebbings, L., Leroy, C., Edkins, S., Hardy, C., Teague, J., Menzies, A., Goodhead, I., Turner, D., Clee, C., Quail, M., Cox, A., Brown, C., Durbin, R., Hurles, M., Edwards, P., Bignell, G., Stratton, M., and Futreal, P. (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genetics*, 40(6):722–729.

Cardis, E., Vrijheid, M., Blettner, M., Gilbert, E., Hakama, M., Hill, C., Howe, G., Kaldor, J., Muirhead, C., Schubauer-Berigan, M., Yoshimura, T., Bermann, F.,

Cowper, G., Fix, J., Hacker, C., Heinmiller, B., Marshall, M., Thierry-Chef, I., Utterback, D., Ahn, Y., Amoros, E., Ashmore, P., Auvinen, A., Bae, J., Bernar, J., Biau, A., Combalot, E., Deboodt, P., Sacristan, A., Eklef, M., Engels, H., Engholm, G., Gulis, G., Habib, R., Holan, K., Hyvonen, H., Kerekes, A., Kurtinaitis, J., Malker, H., Martuzzi, M., Mastauskas, A., Monnet, A., Moser, M., Pearce, M., Richardson, D., Rodriguez-Artalejo, F., Rogel, A., Tardy, H., Telle-Lamberton, M., Turai, I., Usel, M., and Veress, K. (2007). The 15-country collaborative study of cancer risk among radiation workers in the nuclear industry: Estimates of radiation-related cancer risks. *Radiat Res*, 167(4):396–416.

Carrasco, D. R., Tonon, G., Huang, Y., Zhang, Y., Sinha, R., Feng, B., Stewart, J. P., Zhan, F., Khatry, D., Protopopova, M., Protopopov, A., Sukhdeo, K., Hanamura, I., Stephens, O., Barlogie, B., Anderson, K. C., Chin, L., Shaughnessy, J. D., Brennan, C., and Depinho, R. A. (2006). High-resolution genomic profiles define distinct clinico-pathogenetic subgroups of multiple myeloma patients. *Cancer Cell*, 9(4):313–325.

Cerveira, N., Correia, C., Dória, S., Bizarro, S., Rocha, P., Gomes, P., Torres, L., Norton, L., Borges, B., Castedo, S., and Teixeira, M. (2003). Frequency of NUP98-NSD1 fusion transcript in childhood acute myeloid leukaemia. *Leukemia*, 17(11):2244–7.

Cheung, K., Shah, S. P., Steidl, C., Johnson, N., Relander, T., Telenius, A., Lai, B., Murphy, K., Lam, W., Al-Tourah, A., Connors, J., Ng, R., Gascoyne, R., and Horsman, D. (2009). Genome-wide profiling of follicular lymphoma by array comparative genomic hybridization reveals prognostically significant DNA copy number imbalances. *Blood*, 113(1):137–148.

Chin, L. and Gray, J. W. (2008). Translating insights from the cancer genome into clinical practice. *Nature*, 452(7187):553–563.

Chin, S., Wang, Y., Thorne, N., Teschendorff, A., Pinder, S., Vias, M., Naderi, A., Roberts, I., Barbosa-Morais, N., Garcia, M., Iyer, N., Kranjac, T., Robertson, J., Aparicio, S., Tavarï£¡, S., Ellis, I., Brenton, J., and Caldas, C. (2007). Using array-comparative genomic hybridization to define molecular portraits of primary breast cancers. *Oncogene*, 26:1959ï£¡–1970.

Cho, K. and Shih, I. (2009). Ovarian cancer. *Annu Rev Pathol.*, (4):287–313.

Choma, D., Daurés, J., Quantin, X., and Pujol, J. (2001). Aneuploidy and prognosis of non-small-cell lung cancer: a meta-analysis of published data. *Br J Cancer*, 85(1):14–22.

Climent, J., Dimitrow, P., Fridlyand, J., Palacios, J., Siebert, R., Albertson, D., Gray, J., Pinkel, D., Lluch, A., and Martinez-Climent, J. (2007). Deletion of chromosome 11q predicts response to anthracycline-based chemotherapy in early breast cancer. *Cancer Research*, 67(2):818–826.

Colditz, G., Willett, W., Hunter, D., Stampfer, M., Manson, J., Hennekens, C., and Rosner, B. (1993). Family history, age, and risk of breast cancer. prospective data from the nurses' health study. *JAMA*, 270(3):338–43.

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Cui, H., Cruz-Correa, M., Giardiello, F., Hutcheon, D., Kafonek, D., Brandenburg, S., Wu, Y., He, X., Powe, N., and Feinberg, A. (2003). Loss of igf2 imprinting: a potential marker of colorectal cancer risk. *Science*, 299(5613):1753–5.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2):224–227.

Davies, H., Bignell, G., Cox, C., Stephens, P., Edkins, S., Clegg, S., Teague, J., Woffendin, H., Garnett, M., Bottomley, W., Davis, N., Dicks, E., Ewing, R., Floyd, Y., Gray, K., Hall, S., Hawes, R., Hughes, J., Kosmidou, V., Menzies, A., Mould, C., Parker, A., Stevens, C., Watt, S., Hooper, S., Wilson, R., Jayatilake, H., Gusterson, B., Cooper, C., Shipley, J., Hargrave, D., Pritchard-Jones, K., Maitland, N., Chenevix-Trench, G., Riggins, G., Bigner, D., Palmieri, G., Cossu, A., Flanagan, A., Nicholson, A., Ho, J., Leung, S., Yuen, S., Weber, B., Seigler, H., Darrow, T., Paterson, H., Marais, R., Marshall, C., Wooster, R., Stratton, M., and Futreal, P. (2002). Mutations of the braf gene in human cancer. *Nature*, 417(6892):949–54.

Dawood, S., Hu, R., Homes, M., Collins, L., Schnitt, S., Connolly, J., Colditz, G., and Tamimi, R. (2011). Defining breast cancer prognosis based on molecular phenotypes: results from a large cohort study. *Breast Cancer Research and Treatment*, 126:185–192.

De la Torre, F. and Black, M. J. (2001). Robust principal component analysis for computer vision. In *Eight International Conference on Computer Vision (ICCV'01)*, volume 1, pages 362–369.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Dettling, M. (2004). Finding predictive gene groups from microarray data. *Journal of Multivariate Analysis*, 90(1):106–131.

Díaz-Uriarte, R. and Alvares de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1).

Diskin, S., Eck, T., Greshock, J., Mosse, Y., Naylor, T., Stoeckert, C., Weber, B., Maris, J., and Grant, G. (2006). STAC: A method for testing the significance of dna copy number aberrations across multiple array-CGH experiments. *Genome Research*, 16(9):1149–1158.

Dong, J. (2001). Chromosomal deletions and tumor suppressor genes in prostate cancer. *Cancer Metastasis Rev*, 20(3-4):173–93.

Druker, B., Sawyers, C., Kantarjian, H., Resta, D., Reese, S., Ford, J., Capdeville, R., and Talpaz, M. (2001). Activity of a specific inhibitor of the bcr-abl tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the philadelphia chromosome. *N Engl J Med*, 344(14):1038–42.

Duerr, E., Rollbrocker, B., Hayashi, Y., Peters, N., Meyer-Puttlitz, B., Louis, D., Schramm, J., Wiestler, O., Parsons, R., Eng, C., and von Deimling, A. (1998). Pten mutations in gliomas and glioneuronal tumors. *Oncogene*, 16(17):2259–64.

Dunn, J. C. (1974). Well separated clusters and optimal fuzzy-partitions. *Journal of Cybernetics*, 4:95–104.

Eckschlager, T., Pilát, D., Kodet, R., Dahbiová, R., Stanková, J., Jasinská, J., and Hrusák, O. (1996). Dna ploidy in neuroblastoma. *Neoplasma*, 43(1):23–6.

Eden, A., Gaudet, F., Waghmare, A., and Jaenisch, R. (2003). Chromosomal instability and tumors promoted by dna hypomethylation. *Science*, 300(5618):455.

Eilers, P. and de Menezes, R. (2005). Quantile smoothing of array CGH data. *Bioinformatics*, 21:1146–1153.

Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E. (2005). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–178.

Eiriksdottir, G., Johannesdottir, G., Ingvarsson, S., Björnsdottir, I., Jonasson, J., Agnarsson, B., Hallgrimsson, J., Gudmundsson, J., Egilsson, V., Sigurdsson, H., and Barkardottir, R. (1998). Mapping loss of heterozygosity at chromosome 13q: loss at 13q12-q13 is associated with breast tumour progression and poor prognosis. *European Journal of Cancer*, 34(13):2076–2081.

Esteller, M. (2007). Cancer epigenomics: Dna methylomes and histone-modification maps. *Nat Rev Genet*.

Esteller, M. (2008). Epigenetics in cancer. *N Engl J Med*, 358(11):1148–59.

Esteller, M., Silva, J., Dominguez, G., Bonilla, F., Matias-Guiu, X., Lerma, E., Bussaglia, E., Prat, J., Harkes, I., Repasky, E., Gabrielson, E., Schutte, M., Baylin, S., and Herman, J. (2000). Promoter hypermethylation and brca1 inactivation in sporadic breast and ovarian tumors. *J Natl Cancer Inst*, 92(7):564–9.

Everitt, B., Landau, S., and Leese, M. (2001). *Cluster Analysis*. Wiley.

Fan, C., Oh, D. S., Wessels, L., Weigelt, B., Nuyten, D. S., Nobel, A. B., van't Veer, L. J., and Perou, C. M. (2006). Concordance among gene-expression-based predictors for breast cancer. *N Engl J Med*, 355(6):560–569.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fang, M., Toher, J., Morgan, M., Davison, J., Tannenbaum, S., and Claffey, K. (2011). Genomic differences between estrogen receptor (er)-positive and er-negative human breast carcinoma identified by single nucleotide polymorphism array comparative genome hybridization analysis. *Cancer*, 117(10):2024–2034.

Feinberg, A. (1999). Imprinting of a genomic domain of 11p15 and loss of imprinting in cancer: an introduction. *Cancer Res*, 59(7 Suppl):1743s–1746s.

Feinberg, A. and Tycko, B. (2004). Timeline: The history of cancer epigenetics. *Nat Rev Cancer*, 4(2):1–11.

Fisher, R. (1922). On the interpretation of $\mathcal{X}^2$ from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, 85(1):87–94.

Ford, M., Davies, B., Griffiths, M., Wilson, J., and Fried, M. (1985). Isolation of a gene enhancer within an amplified inverted duplication after "expression selection". *Proc Natl Acad Sci U S A*, 82(10):3370–4.

Fran cois, D., Wertz, V., and Verleysen, M. (2006). The permutation test for feature selection by mutual information. In *ESANN 2006, European Symposium on Artificial Neural Networks*, pages 239–244.

Frezza, E. E., Wachtel, M. S., and Chiriva-Internati, M. (2006). Influence of obesity on the risk of developing colon cancer. *Gut*, 55(2):285–291.

Fridlyand, J., Snijders, A., Pinkel, D., Albertson, D., and Jain, A. (2004). Hidden markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis*, 90(1):132–153.

Fridlyand, J., Snijders, A. M., Ylstra, B., Li, H., Olshen, A., Segraves, R., Dairkee, S., Tokuyasu, T., Ljung, B. M., Jain, A. N., McLennan, J., Ziegler, J., Chin, K., Devries, S., Feiler, H., Gray, J. W., Waldman, F., Pinkel, D., and Albertson, D. G. (2006). Tumor copy number aberration phenotypes and genomic instability. *BMC Cancer*, 6.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Gisselsson, D., Jin, Y., Lindgren, D., Persson, J., Gisselsson, L., Hanks, S., Sehic, D., Mengelbier, L., Øra, I., Rahman, N., Mertens, F., Mitelman, F., and Mandahl, N. (2010). Generation of trisomies in cancer cells by multipolar mitosis and incomplete cytokinesis. *Proceedings of the National Academy of Sciences*, 107(47):20489–20493.

Gordon, L., Yang, S., Tran-Gyamfi, M., Baggott, D., Christensen, M., Hamilton, A., Crooijmans, R., Groenen, M., Lucas, S., Ovcharenko, I., and Stubbs, L. (2007). Comparative analysis of chicken chromosome 28 provides new clues to the evolutionary fragility of gene-rich vertebrate regions. *Genome Research*, 17(11):1603–1613.

Grant, G., Manduchi, E., and Stoeckert, C. (2001). *Analysis and Management of Microarray Gene Expression Data*. John Wiley & Sons, Inc.

Grant, M. and S., B. (2008). Graph implementations for nonsmooth convex programs. In *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited.

Gray, I., Stewart, L., Phillips, S., Hamilton, J., Gray, N., Watson, G., Spurr, N., and Snary, D. (1998). Mutation and expression analysis of the putative prostate tumour-suppressor gene pten. *Br J Cancer*, 78(10):1296–300.

Gruvberger, S., Ringnï£¡r, M., Chen, Y., Panavally, S., Saal, L., Borg A, ., Fernï£¡, M., Peterson, C., and Meltzer, P. (2001). Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res*, 61(16):5979–5984.

Guttman, M., Mies, C., Dudycz-Sulicz, K., Diskin, S., Baldwin, D., Stoeckert, C., and Grant, G. (2007). Assessing the significance of conserved genomic aberrations using high resolution genomic microarrays. *PLoS Genetics*, 3(8).

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182.

Hahn, P. (1993). Molecular biology of double-minute chromosomes. *Bioessays*, 15(7):477–484.

Halachev, K., Bast, H., Lengauer, T., and Bock, C. (2012, manuscript). EpiExplorer website: http://epiexplorer.mpi-inf.mpg.de/index.php.

Hamming, R. (1950). Error detecting and error correcting codes. *Bell System Technical Journal*, 26(2):147–160.

Han, W., Han, M., Kang, J., Bae, J., Lee, J., Bae, Y., Lee, J., Shin, H., Hwang, K., Hwang, S., Kim, S., and Noh, D. (2006). Genomic alterations identified by array comparative genomic hybridization as prognostic markers in tamoxifen-treated estrogen receptor-positive breast cancer. *BMC Cancer*, 6(1).

Hastie, T., Tibshirani, R., and Friedman, J. H. (2003). *The Elements of Statistical Learning*. Springer.

Hastings, P., Lupski, J., Rosenberg, S., and Ira, G. (2009). Mechanisms of change in gene copy number. *Nature Reviews Genetics*, 10(8):551–564.

Herman, J., Latif, F., Weng, Y., Lerman, M., Zbar, B., Liu, S., Samid, D., Duan, D., Gnarra, J., and Linehan, W. (1994). Silencing of the vhl tumor-suppressor gene by dna methylation in renal carcinoma. *Proc Natl Acad Sci U S A*, 91(21):9700–4.

Herman, J., Merlo, A., Mao, L., Lapidus, R., Issa, J., Davidson, N., Sidransky, D., and Baylin, S. (1995). Inactivation of the cdkn2/p16/mts1 gene is frequently associated with aberrant dna methylation in all common human cancers. *Cancer Res*, 55(20):4525–30.

Hicks, J., Krasnitz, A., Lakshmi, B., Navin, N., Riggs, M., Leibu, E., Esposito, D., Alexander, J., Troge, J., Grubor, V., Yoon, S., Wigler, M., Ye, K., Børresen-Dale, A.-L., Naume, B., Schlicting, E., Norton, L., Hägerström, T., Skoog, L., Auer, G., Månér, S., Lundin, P., and Zetterberg, A. (2006). Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome research*, 16(12):1465–1479.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802.

Hodgson, G., Hager, J., Volik, S., Hariono, S., Wernick, M., Moore, D., Nowak, N., Albertson, D., Pinkel, D., Collins, C., Hanahan, D., and Gray, J. (2001). Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nat Genet*, 29(4):459–464.

Holliday, R. (1964). A mechanism for gene conversion in fungi. *Genetics Research*, 5(2):282–304.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Horlings, H., Lai, C., Nuyten, D., Halfwerk, H., Kristel, P., van Beers, E., Joosse, S., Klijn, C., Nederlof, P., Reinders, M., Wessels, L., and van de Vijver, M. (2010). Integration of dna copy number alterations and prognostic gene expression signatures in breast cancer patients. *Clin Cancer Res*, 16(2):651–63.

Hothorn, T., Hornik, K., and Zeileis, A. (2005). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674.

Hsiao, L. L., Jensen, R. V., Yoshida, T., Clark, K. E., Blumenstock, J. E., and Gullans, S. R. (2002). Correcting for signal saturation errors in the analysis of microarray data. *BioTechniques*, 32(2).

Hsu, L., Self, S. G., Grove, D., Randolph, T., Wang, K., Delrow, J. J., Loo, L., and Porter, P. (2005). Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, 6:211–226.

Huang, E., Cheng, S., Dressman, H., Pittman, J., Tsou, M., Horng, C., Bild, A., Iversen, E., Liao, M., Chen, C., West, M., Nevins, J., and Huang, A. (2003a). Gene expression predictors of breast cancer outcomes. *Lancet*, 361(9369):1590–1596.

Huang, E., Ishida, S., Pittman, J., Dressman, H., Bild, A., Kloos, M., D'Amico, M., Pestell, R., West, M., and Nevins, J. (2003b). Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nature genetics*, 34(2):226–230.

Huang, X.-F. and Chen, J.-Z. (2009). Obesity, the pi3k/akt signal pathway and colon cancer. *Obesity Reviews*, 10(6):610–616.

Huber, W., Von Heydebreck, A., Sültmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1.

Hungermann, D., Schmidt, H., Natrajan, R., Tidow, N., Poos, K., Reis-Filho, J., Brandt, B., Buerger, H., and Korsching, E. (2011). Influence of whole arm loss of chromosome 16q on gene expression patterns in oestrogen receptor-positive, invasive breast cancer. *J Pathol*, 48(4):351–365.

Hupé, P., Stransky, N., Thiery, J.-P., Radvanyi, F., and Barillot, E. (2004). Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, 20:3413–3422.

Idbaih, A., Marie, Y., Lucchesi, C., Pierron, G., Manié, E., Raynal, V., Mosseri, V., Hoang-Xuan, K., Kujas, M., Brito, I., Mokhtari, K., Sanson, M., Barillot, E., Aurias, A., Delattre, J.-Y., and Delattre, O. (2008). BAC array CGH distinguishes mutually exclusive alterations that define clinicogenetic subtypes of gliomas. *International journal of cancer*, 122(8):1778–1786.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50.

Jackson, D. and Chen, Y. (2004). Robust principal component analysis and outlier detection with ecological data. *Environmetrics*, 15(2):129–139.

Jäger, J. and Sengupta, R. (2003). Improved gene selection for classification of microarrays. In *Pacific Symposium on Biocomputing*, volume 8, pages 53–64.

Janssen, E., Baak, J., Guervós, M., van Diest, P., Jiwa, M., and Hermsen, M. (2003). In lymph node-negative invasive breast carcinomas, specific chromosomal aberrations are strongly associated with high mitotic activity and predict outcome more accurately than grade, tumour diameter, and oestrogen receptor. *J Pathol*, 201(4):555–561.

Jones, P. and Baylin, S. (2007). The epigenomics of cancer. *Cell*, 128(4):683–692.

Joosse, S., van Beers, E., Tielen, I., Horlings, H., Peterse, J., Hoogerbrugge, N., Ligtenberg, M., Wessels, L., Axwijk, P., Verhoef, S., Hogervorst, F., and Nederlof,

P. (2009). Prediction of BRCA1-association in hereditary non-BRCA1/2 breast carcinomas with array-CGH. *Breast Cancer Research and Treatment*, 116:479–489.

Jung, Y., Park, H., Du, D., and Drake, B. (2003). A decision criterion for the optimal number of clusters in hierarchical clustering. *J. of Global Optimization*, 25(1):91–111.

Kallioniemi, A. (2008). Cgh microarrays and cancer. *Current Opinion in Biotechnology*, 19(1):36–40.

Kallioniemi, A., Kallioniemi, O. P., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, F., and Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, 258(5083):818–821.

Kalousis, A., Prados, J., and Hilario, M. (2006). Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and Information Systems*, 12(1):95–116.

Kaneko, Y. and Knudson, A. (2000). Mechanism and relevance of ploidy in neuroblastoma. *Genes Chromosomes Cancer*, 29(2):89–95.

Kanungo, A., Medeiros, L., Abruzzo, L., and Lin, P. (2005). Lymphoid neoplasms associated with concurrent t(14;18) and 8q24/c-myc translocation generally have a poor prognosis. *Mod Pathol*.

Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Kapp, A., Jeffrey, S., Langerød, A., Børresen-Dale, A.-L., Han, W., Noh, D.-Y., Bukholm, I., Nicolau, M., Brown, P., and Tibshirani, R. (2006). Discovery and validation of breast cancer subtypes. *BMC genomics*, 7.

Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience.

Kim, Y. and Kim, J. (2004). Gradient LASSO for feature selection. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*. ACM.

King, A., Patrick, D., Batorsky, R., Ho, M., Do, H., Zhang, S., Kumar, R., Rusnak, D., Takle, A., Wilson, D., Hugger, E., Wang, L., Karreth, F., Lougheed, J., Lee, J., Chau, D., Stout, T., May, E., Rominger, C., Schaber, M., Luo, L., Lakdawala, A., Adams, J., Contractor, R., Smalley, K., Herlyn, M., Morrissey, M., Tuveson, D., and Huang, P. (2006). Demonstration of a genetic therapeutic index for tumors expressing oncogenic braf by the kinase inhibitor sb-590885. *Cancer Res*, 66(23):11100–5.

Klijn, C., Holstege, H., Ridder, J., Liu, X., Reinders, M., Jonkers, J., and Wessels, L. (2008). Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array CGH data. *Nucl. Acids Res.*, 36(2).

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Kreisheimer, M., Sokolnikov, M., Koshurnikova, N., Khokhryakov, V., Romanow, S., Shilnikova, N., Okatenko, P., Nekolla, E., and Kellerer, A. (2003). Lung cancer mortality among nuclear workers of the mayak facilities in the former soviet union. an updated analysis considering smoking as the main confounding factor. *Radiat Environ Biophys*, 42(2):129–35.

Kresse, S., Ohnstad, H., Bjerkehagen, B., Myklebost, O., and Meza-Zepeda, L. (2010). DNA copy number changes in human malignant fibrous histiocytomas by array comparative genomic hybridisation. *PLoS ONE*, 5(11).

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the* 18^{th} *International Conference on Machine Learning*, pages 282–289.

Lai, W. R., Johnson, M. D., Kucherlapati, R., and Park, P. J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, 21(19):3763–3770.

Langer-Safer, P., Levine, M., and Ward, D. (1982). Immunological method for mapping genes on Drosophila polytene chromosomes. *PNAS*, 79(14):4381ï£¡–4385.

Lemaitre, C., Zaghloul, L., Sagot, M., Gautier, C., Arneodo, A., Tannier, E., and Audit, B. (2009). Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relation to genome organisation. *BMC Genomics*, 10(1):335+.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3):18–22.

Lieber, M. (2007). The mechanism of human nonhomologous DNA end joining. *J Biol Chem*, 283:1–5.

Lipson, D., Aumann, Y., Ben-Dor, A., Linial, N., and Yakhini, Z. (2006). Efficient calculation of interval scores for dna copy number data analysis. *Journal of Computational Biology*, 13(2):215–228.

Little, C., Nau, M., Carney, D., Gazdar, A., and Minna, J. (1983). Amplification and expression of the c-myc oncogene in human lung cancer cell lines. *Nature*, 306(5939):194–6.

Liu, J., Ranka, S., and Kahveci, T. (2008). Classification and feature selection algorithms for multi-class CGH data. *Bioinformatics*, 24(13).

Lo, A., Sabatier, L., Fouladi, B., Pottier, G., Ricoul, M., and Murnane, J. (2002). DNA amplification by breakage/fusion/bridge cycles initiated by spontaneous telomere loss in a human cancer cell line. *Neoplasia*, 4(6):531–538.

Lodish, H., Berk, A., Matsudaira, P., Kaiser, C. A., Krieger, M., Scott, M. P., Zipursky, L., and Darnell, J. (2003). *Molecular Cell Biology*. W. H. Freeman, fifth edition edition.

London, W., Castleberry, R., Matthay, K., Look, A., Seeger, R., Shimada, H., Thorner, P., Brodeur, G., Maris, J., Reynolds, C., and Cohn, S. (2005). Evidence for an age cutoff greater than 365 days for neuroblastoma risk group stratification in the children's oncology group. *J Clin Oncol*, 23(27):6459–6465.

Loo, L., Grove, D., Williams, E., Neal, C., Cousens, L., Schubert, E., Holcomb, I., Massa, H., Glogovac, J., Li, C., Malone, K., Daling, J., Delrow, J., Trask, B., Hsu, L., and Porter, P. (2004). Array comparative genomic hybridization analysis of genomic alterations in breast cancer subtypes. *Cancer Res*, 64:8541ï£¡–8549.

Lund, M., Butler, E., Hair, B., Ward, K. C., Andrews, J., Oprea-Ilies, G., Bayakly, A., O'Regan, R., Vertino, P., and Eley, J. W. (2010). Age/race differences in HER2 testing and in incidence rates for breast cancer triple subtypes. *Cancer*, 116(11):2549–2559.

Lynch, T., Bell, D., Sordella, R., Gurubhagavatula, S., Okimoto, R., Brannigan, B., Harris, P., Haserlat, S., Supko, J., Haluska, F., Louis, D., Christiani, D., Settleman, J., and Haber, D. (2004). Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N Engl J Med*, 350(21):2129–39.

M., G. and Boyd, S. (2011). CVX: Matlab software for disciplined convex programming, version 1.21. http://cvxr.com/cvx.

Ma, S., Song, X., and Huang, J. (2007). Supervised group lasso with applications to microarray data analysis. *BMC Bioinformatics*, 8(1).

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports.*, 50(3):163–170.

McClintock, B. (1941). The stability of broken ends of chromosomes in zea mays. *Genetics*, 26(2):234–282.

Mclachlan, G. and Krishnan, T. (1996). *The EM Algorithm and Extensions*. John Wiley and Sons, New York.

McVey, M. and Lee, S. (2008). MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends Genet*, 24:529–538.

Meier, L., van de Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71.

Merlo, A., Herman, J., Mao, L., Lee, D., Gabrielson, E., Burger, P., Baylin, S., and Sidransky, D. (1995). 5' cpg island methylation is associated with transcriptional silencing of the tumour suppressor p16/cdkn2/mts1 in human cancers. *Nat Med*, 1(7):686–92.

Merlo, L., Wang, L., Pepper, J., Rabinovitch, P., and Maley, C. (2010). Polyploidy, aneuploidy and the evolution of cancer. *Adv Exp Med Biol*, 676.

Mikeska, T., Bock, C., El-Maarri, O., Hübner, A., Ehrentraut, D., Schramm, J., Felsberg, J., Kahl, P., Büttner, R., Pietsch, T., and Waha, A. (2007). Optimization of quantitative MGMT promoter methylation analysis using pyrosequencing and combined bisulfite restriction analysis. *The Journal of Molecular Diagnostics*, 9(3):368–381.

Minobe, K., Onda, M., Iida, A., Kasumi, F., Sakamoto, G., Nakamura, Y., and Emi, M. (1998). Allelic loss on chromosome 9q is associated with lymph node metastasis of primary breast cancer. *Jpn J Cancer Res*, 89(9):916–922.

Mirkin, B. (1996). *Mathematical classification and clustering*. Nonconvex optimization and its applications. Kluwer Academic Pubishers.

Nakao, K., Mehta, K., Fridlyand, J., Moore, D., Jain, A., Lafuente, A., Wiencke, J., Terdiman, J., and Waldman, F. (2004). High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. *Carcinogenesis*, 25(8):1345–1357.

Nakazawa, M. (1993). The prognostic significance of dna ploidy for neuroblastoma. *Surgery Today*, 23:215–219.

Narayanan, D. L., Saladi, R. N., and Fox, J. L. (2010). Review: Ultraviolet radiation and skin cancer. *International Journal of Dermatology*, 49(9):978–986.

Nardi, V., Azam, M., and Daley, G. (2004). Mechanisms and implications of imatinib resistance mutations in bcr-abl. *Curr Opin Hematol*, 11(1):35–43.

Neuvial, P., Hupe, P., Brito, I., Liva, S., Manie, E., Brennetot, C., Radvanyi, F., Aurias, A., and Barillot, E. (2006). Spatial normalization of array-CGH data. *BMC Bioinformatics*, 7(1).

Ngoma, T. (2006). World Health Organization cancer priorities in developing countries. *Ann Oncol*, 17 Suppl 8.

Nobori, T., Miura, K., Wu, D., Lois, A., Takabayashi, K., and Carson, D. (1994). Deletions of the cyclin-dependent kinase-4 inhibitor gene in multiple human cancers. *Nature*, 368(6473):753–6.

Nowell, P. (1976). The clonal evolution of tumor cell populations. *Science*, 194(4260):23–8.

Nowell, P. (2007). Discovery of the philadelphia chromosome: a personal perspective. *J Clin Invest*, 117(8):2033–5.

Olaharski, A., Sotelo, R., Solorza-Luna, G., Gonsebatt, M., Guzman, P., Mohar, A., and Eastmond, D. (2006). Tetraploidy and chromosomal instability are early events during cervical carcinogenesis. *Carcinogenesis*, 27(2):337–343.

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572.

Ormandy, C., Musgrove, E., Hui, R., Daly, R., and Sutherland, R. (2003). Cyclin D1, EMS1 and 11q13 amplification in breast cancer. *Breast Cancer Res Treat*, 78(3):323–335.

Pang, H. and Zhao, H. (2008). Building pathway clusters from random forests classification using class votes. *BMC Bioinformatics*, 9(1).

Pao, W., Miller, V., Zakowski, M., Doherty, J., Politi, K., Sarkaria, I., Singh, B., Heelan, R., Rusch, V., Fulton, L., Mardis, E., Kupfer, D., Wilson, R., Kris, M., and Varmus, H. (2004). Egf receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proc Natl Acad Sci U S A*, 101(36):13306–11.

Park, M., Hastie, T., and Tibshirani, R. (2007). Averaged gene expressions for regression. *Biostatistics (Oxford, England)*, 8(2):212–227.

Parkin, D. M. (2011). Tobacco-attributable cancer burden in the UK in 2010. *British Journal of Cancer*, 105(S2):S6–S13.

Pastink, A., Eeken, J. C., and Lohman, P. H. (2001). Genomic integrity and the repair of double-strand DNA breaks. *Mutation Research*, 480-481:37–50.

Pegram, M. and Slamon, D. (2000). Biological rationale for her2/neu (c-erbb2) as a target for monoclonal antibody therapy. *Semin Oncol*, 27(5 Suppl 9):13–9.

Pellman, D. (2007). Cell biology: aneuploidy and cancer. *Nature*, 446(7131):38–9.

Picard, F., Robin, S., Lavielle, M., Vaisse, C., and Daudin, J.-J. (2005). A statistical approach for array CGH data analysis. *BMC Bioinformatics*, 6(1):27–31.

Pierce, B. (2007). *Genetics: A Conceptual Approach*. W. H. Freeman, third edition edition.

Pinkel, D. and Albertson, D. (2005). Array comparative genomic hybridization and its applications in cancer. *Nature genetics*, 37 Suppl.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W., Chen, C., Zhai, Y., Dairkee, S., Ljung, B., Gray, J., and Albertson, D. (1998). High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, 20(2):207–211.

Polzehl, J. and Spokoiny, V. (2002). Local likelihood modeling by adaptive weights smoothing. *WIAS-Preprint 787*.

Pyatt, G., Chen, C., and Fei, J. (1980). The distribution of income by factor components. *The Quarterly Journal of Economics*, 95(3):451–73.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature Genetics*, 32 Suppl:496–501.

Raimondi, S., Zhou, Y., Shurtleff, S., Rubnitz, J., Pui, C., and Behm, F. (2006). Near-triploidy and near-tetraploidy in childhood acute lymphoblastic leukemia: association with b-lineage blast cells carrying the etv6-runx1 fusion, t-lineage immunophenotype, and favorable outcome. *Cancer Genet Cytogenet*, 169(1):50–7.

Rao, P. H., Houldsworth, J., Dyomina, K., Parsa, N. Z., Cigudosa, J. C., Louie, D. C., Popplewell, L., Offit, K., Jhanwar, S. C., and Chaganti, R. S. (1998). Chromosomal and gene amplification in diffuse large B-cell lymphoma. *Blood*, 92(1):234–240.

Rapaport, F., Barillot, E., and Vert, J.-P. (2008). Classification of arrayCGH data using fused svm. *Bioinformatics*, 24(13):i375–i382.

Rennstam, K., Ahlstedt-Soini, M., Baldetorp, B., Bendahl, P., Borg, A., Karhu, R., Tanner, M., Tirkkonen, M., and Isola, J. (2003). Patterns of chromosomal imbalances defines subgroups of breast cancer with distinct clinical features and prognosis. a study of 305 tumors by comparative genomic hybridization. *Cancer Res*, 63(24):8861–8868.

Richardson, D., Sugiyama, H., Nishi, N., Sakata, R., Shimizu, Y., Grant, E., Soda, M., Hsu, W., Suyama, A., Kodama, K., and Kasagi, F. (2009). Ionizing radiation and leukemia mortality among japanese atomic bomb survivors, 1950-2000. *Radiat Res*, 172(3):368–82.

Richon, V., Sandhoff, T., Rifkind, R., and Marks, P. (2000). Histone deacetylase inhibitor selectively induces p21waf1 expression and gene-associated histone acetylation. *Proc Natl Acad Sci U S A*, 97(18):10014–9.

Risques, R., Moreno, V., Marcuello, E., Petriz, J., Cancelas, J., Sancho, F., Torregrosa, A., Capella, G., and Peinado, M. (2001). Redefining the significance of aneuploidy in the prognostic assessment of colorectal cancer. *Lab Invest*, 81(3):307–15.

Ritz, A., Paris, P., Ittmann, M., Collins, C., and Raphael, B. (2011). Detection of recurrent rearrangement breakpoints from copy number data. *BMC Bioinformatics*, 12(1):114+.

Robinson, J., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E., Getz, G., and Mesirov, J. (2011). Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26.

Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65.

Rouveirol, C., Stransky, N., Hupé, P., La Rosa, P., Viara, E., Barillot, E., and Radvanyi, F. (2006). Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics*, 22(7):849–856.

Rueda, O. and Díaz-Uriarte, R. (2010). Finding recurrent copy number alteration regions: A review of methods. *Current Bioinformatics*, 5(1):1–17.

Russnes, H., Vollan, H., Lingjaerde, O., Krasnitz, A., Lundin, P., Naume, B., Sørlie, T., Borgen, E., Rye, I., Langerød, A., Chin, S.-F., Teschendorff, A., Stephens, P., Månér, S., Schlichting, E., Baumbusch, L. O., Kåresen, R., Stratton, M., Wigler, M., Caldas, C., Zetterberg, A., Hicks, J., and Børresen-Dale, A.-L. (2010). Genomic architecture characterizes tumor progression paths and fate in breast cancer patients. *Science translational medicine*, 2(38).

Sakai, T., Toguchida, J., Ohtani, N., Yandell, D., Rapaport, J., and Dryja, T. (1991). Allele-specific hypermethylation of the retinoblastoma tumor-suppressor gene. *Am J Hum Genet*, 48(5):880–8.

Sarid, R. and Gao, S. (2011). Viruses and human cancer: From detection to causality. *Cancer Letters*, 305(2).

Schelhorn, S., Fischer, M., Tolo¸si, L., Lengauer, T., and Berthold, F. (2012). No evidence for viral replication in deep transcriptome sequencing data of metastatic neuroblastoma with progressive stage 4 and regressive stage 4S.

Scherer, S. W., Lee, C., Birney, E., Altshuler, D. M., Eichler, E. E., Carter, N. P., Hurles, M. E., and Feuk, L. (2007). Challenges and standards in integrating surveys of structural variation. *Nature Genetics*, 39(7 Suppl):S7–S15.

Sebo, T., Cheville, J., Riehle, D., Lohse, C., Pankratz, V., Myers, R., Blute, M., and Zincke, H. (2001). Predicting prostate carcinoma volume and stage at radical prostatectomy by assessing needle biopsy specimens for percent surface area and cores positive for carcinoma, perineural invasion, gleason score, dna ploidy and proliferation, and preoperative serum prostate specific antigen: a report of 454 cases. *Cancer*, 91(11):2196–204.

Segal, M. (2004). Machine learning benchmarks and Random Forest regression. *Technical Report, Center for Bioinformatics & Molecular Biostatistics, University of California, San Francisco.*

Sen, A. and Srivastava, M. (1975). On tests for detecting change in mean. *The Annals of Statistics*, 3(1):98–108.

Sen, S. (2000). Aneuploidy and cancer. *Curr Opin Oncol*, 12(1):82–8.

Shah, S., Cheung, K.-J., Johnson, N., Alain, G., Gascoyne, R., Horsman, D., Ng, R., and Murphy, K. (2009). Model-based clustering of array CGH data. *Bioinformatics*, 25(12):i30–38.

Shah, S. P. (2008). Computational methods for identification of recurrent copy number alteration patterns by array CGH. *Cytogenetic and genome research*, 123(1-4):343–351.

Shah, S. P., Lam, W. L., Ng, R. T., and Murphy, K. P. (2007). Modeling recurrent DNA copy number alterations in array CGH data. *Bioinformatics*, 23(13):i450–i458.

Sharan, S. and Kuznetsov, S. (2007). Resolving RAD51C function in late stages of homologous recombination. *Cell Div*, 2(1):15.

Shigematsu, H., Kadoya, T., Kobayashi, Y., Kajitani, K., Sasada, T., Emi, A., Masumoto, N., Haruta, R., Kataoka, T., Oda, M., Arihiro, K., and Okada, M. (2011). A case of HER-2-positive recurrent breast cancer showing a clinically complete response to trastuzumab-containing chemotherapy after primary treatment of triple-negative breast cancer. *World J Surg Oncol*, 9(1):146.

Singer, M., Mesner, L., Friedman, C., Trask, B., and Hamlin, J. (2000). Amplification of the human dihydrofolate reductase gene via double minutes is initiated by chromosome breaks. *Proc Natl Acad Sci USA*, 97(14):7921–7926.

Sircoulomb, F., Bekhouche, I., Finetti, P., Adelaide, J., Hamida, A., Bonansea, J., Raynaud, S., Innocenti, C., Charafe-Jauffret, E., Tarpin, C., Ayed, F., Viens, P., Jacquemier, J., Bertucci, F., Birnbaum, D., and Chaffanet, M. (2010). Genome profiling of ERBB2-amplified breast cancers. *BMC Cancer*, 10.

Slamon, D., Clark, G., Wong, S., Levin, W., Ullrich, A., and McGuire, W. (1987). Human breast cancer: correlation of relapse and survival with amplification of the her-2/neu oncogene. *Science*, 235(4785):177–82.

Smyth, G. and Speed, T. (2003). Normalization of cdna microarray data. *Methods (San Diego, Calif.)*, 31(4):265–273.

Sneath, P. and Sokal, R. (1962). Numerical taxonomy. *Nature*, 193:855–860.

Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Döhner, H., Cremer, T., and Lichter, P. (1997). Matrix-based comparative genomic hybridization: Biochips to screen for genomic imbalances. *Genes, Chromosomes and Cancer*, 20(4):399–407.

Sørlie, T. (2004). Molecular portraits of breast cancer: tumour subtypes as distinct disease entities. *European journal of cancer*, 40(18):2667–2675.

Sos, M., Michel, K., Zander, T., Weiss, J., Frommolt, P., Peifer, M., Li, D., Ullrich, R., Koker, M., Fischer, F., Shimamura, T., Rauh, D., Mermel, C., Fischer, S., Stückrath, I., Heynck, S., Beroukhim, R., Lin, W., Winckler, W., Shah, K., Laframboise, T., Moriarty, W., Hanna, M., Tolo¸si, L., Rahnenführer, J., Verhaak, R., Chiang, D., Getz, G., Hellmich, M., Wolf, J., Girard, L., Peyton, M., Weir, B., Chen, T., Greulich, H., Barretina, J., Shapiro, G., Garraway, L., Gazdar, A., Minna, J., Meyerson, M., Wong, K., and Thomas, R. (2009). Predicting drug susceptibility of non-small cell lung cancers based on genetic lesions. *J Clin Invest*.

Spitz, R., Oberthuer, A., Zapatka, M., Brors, B., Hero, B., Ernestus, K., Oestreich, J., Fischer, M., Simon, T., and Berthold, F. (2006). Oligonucleotide array-based comparative genomic hybridization (aCGH) of 90 neuroblastomas reveals aberration patterns closely associated with relapse pattern and outcome. *Genes, Chromosomes and Cancer*, 45(12):1130–1142.

Staaf, J., Jonsson, G., Ringner, M., and Christersson, J. (2007). Normalization of array-CGH data: influence of copy number imbalances. *BMC Genomics*, 8(1).

Storchova, Z. and Pellman, D. (2004). From polyploidy to aneuploidy, genome instability and cancer. *Nat Rev Mol Cell Biol*, 5(1):45–54.

Storlazzi, C., Lonoce, A., Guastadisegni, M., Trombetta, D., D'Addabbo, P., Daniele, G., L'Abbate, A., Macchia, G., Surace, C., Kok, K., Ullmann, R., Purgato, S., Palumbo, O., Carella, M., Ambros, P., and Rocchi, M. (2010). Gene amplification as double minutes or homogeneously staining regions in solid tumors: origin and structure. *Genome Res*, 20(9):1198–1206.

Strobl, C., Boulesteix, A., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1).

Strobl, C., Boulesteix, A., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1).

Student (1908). The probable error of a mean. *Biometrika*, 6:1–25.

Suehiro, Y., Okada, T., Okada, T., Anno, K., Okayama, N., Ueno, K., Hiura, M., Nakamura, M., Kondo, T., Oga, A., Kawauchi, S., Hirabayashi, K., Numa, F., Ito, T., Saito, T., Sasaki, K., and Hinoda, Y. (2008). Aneuploidy predicts outcome in patients with endometrial carcinoma and is related to lack of cdh13 hypermethylation. *Clin Cancer Res*, 14(11):3354–3361.

Tagawa, H., Suguro, M., Tsuzuki, S., Matsuo, K., Karnan, S., Ohshima, K., Okamoto, M., Morishima, Y., Nakamura, S., and Seto, M. (2005). Comparison

of genome profiles for identification of distinct subgroups of diffuse large B-cell lymphoma. *Blood*, 106:1770–1777.

Tanaka, H. and Yao, M. (2009). Palindromic gene amplification - an evolutionarily conserved role for dna inverted repeats in the genome. *Nat Rev Cancer*, 9:215–224.

Tashiro, H., Blazes, M., Wu, R., Cho, K., Bose, S., Wang, S., Li, J., Parsons, R., and Ellenson, L. (1997). Mutations in pten are frequent in endometrial carcinoma but rare in other common gynecological malignancies. *Cancer Res*, 57(18):3935–40.

Taub, R., Kirsch, I., Morton, C., Lenoir, G., Swan, D., Tronick, S., Aaronson, S., and Leder, P. (1982). Translocation of the c-myc gene into the immunoglobulin heavy chain locus in human burkitt lymphoma and murine plasmacytoma cells. *Proc Natl Acad Sci U S A*, 79(24):7837–41.

Taylor, B., Barretina, J., Socci, N., DeCarolis, P., Ladanyi, M., Meyerson, M., Singer, S., and Sander, C. (2008). Functional copy-number alterations in cancer. *PLoS ONE*, 3(9).

Theisen, A. (2008). Microarray-based comparative genomic hybridization (aCGH). *Nature Education*, 1(1).

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science*, 18(1):104–117.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Toledo, F., Buttin, G., and Debatisse, M. (1993). The origin of chromosome rearrangements at early stages of AMPD2 gene amplification in chinese hamster cells. *Curr Biol*, 3(5):255–264.

Toloşi, L. and Lengauer, T. (2011). Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, 27(14):1986–1994.

Torosantucci, L., De Santis Puzzonia, M., Cenciarelli, C., Rens, W., and Degrassi, F. (2009). Aneuploidy in mitosis of PtK1 cells is generated by random loss and nondisjunction of individual chromosomes. *J Cell Sci*, 122:3455–3461.

Tsafrir, D., Bacolod, M., Selvanayagam, Z., Tsafrir, I., Shia, J., Zeng, Z., Liu, H., Krier, C., Stengel, R., Barany, F., Gerald, W., Paty, P., Domany, E., and Notterman, D. (2006). Relationship of gene expression and chromosomal abnormalities in colorectal cancer. *Cancer Res*, 66(4):2129–2137.

Tsai, A., Lu, H., Raghavan, S., Muschen, M., Hsieh, C., and Lieber, M. (2008). Human chromosomal translocations at cpg sites and a theoretical basis for their lineage and stage specificity. *Cell*, 135(6):1130–1142.

Tusher, V., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121.

Van de Wiel, M., Kim, K., Vosse, S., van Wieringen, W., Wilting, S., and Ylstra, B. (2007). CGHcall: calling aberrations for array CGH tumor profiles. *Bioinformatics*, 23.

van de Wiel, M. and van Wieringen, W. (2007). CGHregions: Dimension reduction for array CGH data with minimal information loss. *Cancer informatics*, 3:55–63.

Van Meir, E., Hadjipanayis, C., Norden, A., Shu, H.-K., Wen, P., and Olson, J. (2010). Exciting new advances in neuro-oncology: The avenue to a cure for malignant glioma. *CA Cancer J Clin*, 60(3):166–193.

van Wieringen, W., Van De Wiel, M., and Ylstra, B. (2007). Weighted clustering of called array CGH data. *Biostatistics*, pages 484–500.

van't Veer, L., Dai, H., van de Vijver, M., He, Y., Hart, A., Mao, M., Peterse, H., van der Kooy, K., Marton, M., Witteveen, A., Schreiber, G., Kerkhoven, R., Roberts, C., Linsley, P., Bernards, R., and Friend, S. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley-Interscience.

Varley, J., Swallow, J., Brammar, W., Whittaker, J., and Walker, R. (1987). Alterations to either c-erbb-2(neu) or c-myc proto-oncogenes in breast carcinomas correlate with poor short-term prognosis. *Oncogene*, 1(4):423–30.

Veeriah, S., Taylor, B., Meng, S., Fang, F., Yilmaz, E., Vivanco, I., Janakiraman, M., Schultz, N., Hanrahan, A., Pao, W., Ladanyi, M., Sander, C., Heguy, A., Holland, E., Paty, P., Mischel, P., Liau, L., Cloughesy, T., Mellinghoff, I., Solit, D., and Chan, T. (2010). Somatic mutations of the parkinson's disease-associated gene PARK2 in glioblastoma and other human malignancies. *Nature genetics*, 42(1):77–82.

Veltman, J., Fridlyand, J., Pejavar, S., Olshen, A., Korkola, J., DeVries, Y., Carroll, P., Kuo, W.-L., Pinkel, D., Albertson, D., Cordon-Cardo, C., Jain, A., and Waldman, F. (2003). Array-based comparative genomic hybridization for genome-wide screening of DNA copy number in bladder tumors. *Cancer Res*, 63:2872–2880.

Venkatraman, E. S. and Olshen, A. B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23(6):657–663.

Verhaak, R. W., Hoadley, K., Purdom, E., Wang, V., Qi, Y., Wilkerson, M., Miller, C., Ding, L., Golub, T., and Mesirov, J. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, 17(1):98–110.

Vilchez, R., Kozinetz, C., Arrington, A., Madden, C., and Butel, J. (2003). Simian virus 40 in human cancers. *The American Journal of Medicine*, 114(8):675–684.

Walter, V., Nobel, A., and Wright, F. (2011). DiNAMIC: a method to identify recurrent DNA copy number aberrations in tumors. *Bioinformatics.*, 5(27):678–685.

Walther, A., Houlston, R., and Tomlinson, I. (2008). Association between chromosomal instability and prognosis in colorectal cancer: a meta-analysis. *Gut*, 57(7):941–950.

Wang, P., Kim, Y., Pollack, J., Narasimhan, B., and Tibshirani, R. (2005). A method for calling gains and losses in array CGH data. *Biostatistics*, 6(1):45–58.

Wang, S., Parsons, R., and Ittmann, M. (1998). Homozygous deletion of the pten tumor suppressor gene in a subset of prostate adenocarcinomas. *Clin Cancer Res*, 4(3):811–5.

Ward, J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

Weir, B., Woo, M., Getz, G., Perner, S., Ding, L., Beroukhim, R., Lin, W., Province, M., Kraja, A., Johnson, L., Shah, K., Sato, M., Thomas, R., Barletta, J., Borecki, I., Broderick, S., Chang, A., Chiang, D., Chirieac, L., Cho, J., Fujii, Y., Gazdar, A., Giordano, T., Greulich, H., Hanna, M., Johnson, B., Kris, M., Lash, A., Lin, L., Lindeman, N., Mardis, E., McPherson, J., Minna, J., Morgan, M., Nadel, M., Orringer, M., Osborne, J., Ozenberger, B., Ramos, A., Robinson, J., Roth, J., Rusch, V., Sasaki, H., Shepherd, F., Sougnez, C., Spitz, M., Tsao, M.-S., Twomey, D., Verhaak, R., Weinstock, G., Wheeler, D., Winckler, W., Yoshizawa, A., Yu, S., Zakowski, M., Zhang, Q., Beer, D., Wistuba, I., Watson, M., Garraway, L. A., Ladanyi, M., Travis, W., Pao, W., Rubin, M., Gabriel, S., Gibbs, R., Varmus, H., Wilson, R., Lander, E., and Meyerson, M. (2007). Characterizing the cancer genome in lung adenocarcinoma. *Nature*, 450(7171):893–898.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J., Marks, J., and Nevins, J. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci USA*, 98(20):11462–11467.

Willenbrock, H. and Fridlyand, J. (2005). A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, 21(22):4084–4091.

Williamson, M., Elder, P., Shaw, M., Devlin, J., and Knowles, M. (1995). p16 (cdkn2) is a major deletion target at 9p21 in bladder cancer. *Hum Mol Genet*, 4(9):1569–77.

Wold, M. S. (1997). Replication protein a:a heterotrimeric, single-stranded dna-binding protein required for eukaryotic dna metabolism. *Annual Review of Biochemistry*, 66(1):61–92.

Wu, C. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103.

Wu, L., Chipman, H., Bull, S., Briollais, L., and Wang, K. (2009). A bayesian segmentation approach to ascertain copy number variations at the population level. *Bioinformatics.*, 25(13):1669–1679.

Yang, X., Sherman, M., Rimm, D., Lissowska, J., Brinton, L., Peplonska, B., Hewitt, S., Anderson, W., Szeszenia-Dabrowska, N., Bardin-Mikolajczak, A., Zatonski, W., Cartun, R., Mandich, D., Rymkiewicz, G., Ligaj, M., Lukaszek, S., Kordek, R., and Garcia-Closas, M. (2007). Differences in risk factors for breast cancer molecular subtypes in a population-based study. *Cancer Epidemiol Biomarkers Prev*, 16(3):439–443.

Ylipää, A., Nykter, M., Kivinen, V., Hu, L., Cogdell, D., Hunt, K., Zhang, W., and Yli-Harja, O. (2008). Finding common aberrations in array CGH data. In *ISCCSP 2008*, volume 114.

Yu, L., Ding, C., and Loscalzo, S. (2008). Stable feature selection via dense feature groups. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 803–811. ACM.

Zhang, Y. and Rowley, J. (2006). Chromatin structural elements and chromosomal translocations in leukemia. *DNA Repair (Amst)*, 5(9-10):1282–1297.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36(4):1509–1533.