VU VRIJE UNIVERSITEIT AMSTERDAM

VU Research Portal

Analysis of chromosomal copy number aberrations in gastrointestinal cancer

Haan, J.C.

2014

document version
Publisher's PDF, also known as Version of record

Link to publication in VU Research Portal

citation for published version (APA)
Haan, J. C. (2014). *Analysis of chromosomal copy number aberrations in gastrointestinal cancer*.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:
vuresearchportal.ub@vu.nl

Download date: 20. Sep. 2021

Chapter 1

General introduction and outline of this thesis

BACKGROUND ON COLORECTAL CANCER

Colorectal cancer (CRC) is the second leading cause of cancer death in the western world. In the Netherlands 12755 new cases have been reported in 2010 (www.ikcnet.nl), of whom about 40% will not survive this disease. Worldwide, over 1.2 million diagnosed cancer cases and 608.700 deaths are estimated to have occurred in 2008.[1] Colorectal cancer is the third most commonly diagnosed cancer in males, after lung and prostate cancer. In females it is the second most common cancer, after breast cancer.

CRC results from the accumulation of multiple genetic and epigenetic alterations.[2] These alterations occur in the genome of colorectal epithelial cells, resulting in gain of function of oncogenes that promote tumor growth and loss of function of tumor suppressor genes (TSG). Typical tumor driver genes of CRC include oncogenes like KRAS and BRAF and TSGs like APC, TP53, PTEN and SMAD4.

Genomic instability is a crucial step in carcinogenesis, which mainly occurs through either one of two mechanisms: the chromosomal instability pathway (CIN), where instability occurs at the chromosome level resulting in chromosomal gains and losses of large portions or whole chromosomes,[3] and the microsatellite instability pathway (MSI), where instability at the nucleotide level results in accumulation of multiple mutations. In CRC, the majority (85%) is chromosomally unstable[3] and 15% of all tumors are microsatellite unstable. Yet, MSI and CIN are not mutually exclusive. CIN is particularly associated with the progression of premalignant adenomas to invasive cancer.[4] In addition to tumor specific genetic mutations and chromosomal aberrations, gene function can also be disrupted by epigenetic alterations, mostly involving silencing of gene transcription due to promoter hypermethylation.[5] Hypermethylation of tumor suppressor genes overlaps with both MSI and CIN pathways and it seems to occur early in tumor development.[6]

Diagnosis and treatment options are currently based on clinicopathological staging. Stage I/IV are classified according to the TNM classification system, the local spread of tumor (T), lymph node metastasis (N) and distant metastasis (M).[7]

For patients with stage I-III colon cancer, surgery with curative intent is the treatment of choice. Adjuvant chemotherapy significantly improves the overall and disease free survival of stage III CRC patients and possibly also of high risk stage II patients.[8] For patients with metastatic colorectal cancer (mCRC), curative treatment options are currently highly restricted and often not available. Palliative systemic therapy prolongs the median overall survival. 5-fluorouracil plus leucovorin (5-FU/LV) gives an overall survival of approximately 12 months and this has been the standard drug treatment for many years.[9] More recently, oral capecitabine was introduced. This agent is less toxic and at least equally effective compared to 5-FU/LV.[10] Addition of new cytotoxic drugs such as irinotecan and oxaliplatin to the standard treatment of capecatibine has further increased survival times.[11,12]

Targeted therapies have recently been introduced and consist of a different class of agents which interfere with cell signaling pathways that are critical for tumor cells. Three targeted agents for CRC have been approved so far by the US Food and Drug Administration (FDA), namely the monoclonal antibodies cetuximab, panitumumab and bevacizumab. Cetuximab and panitumumab are anti-EGFR antibodies, which bind to the extracellular domain of the epidermal growth factor receptor (EGFR), interrupting with the RAS-RAF-ERK pathway, which results in inhibition of cell growth and proliferation.

Bevacizumab inhibits angiogenesis by binding to vascular endothelial growth factor (VEGF). Binding of VEGF to the VEGF-receptors on endothelial cells is prevented, which in turn reduces growth of vessels that may nurture tumor cells. Besides the three FDA approved agents, two novel targeted agents, aflibercept and regorafenib, showed survival benefits.[13,14] Aflibercept targets VEGF and regorafenib targets a variety of tyrosine kinases involved in angiogenesis and tumor cell growth. Moreover, many other novel targeted agents are currently being explored in clinical trials.[15]

A NEED FOR PROGNOSTIC AND PREDICTIVE BIOMARKERS

Clinical outcome is largely determined by aggressiveness of the tumor, which is classified by the stage of disease. Five-year overall survival declines from greater than 90% to about 19% from stage I to IV.[16,17] Clinical outcome is furthermore determined by the possibility to perform radical resection of metastases and response to drug therapy. Current drug therapy for colorectal cancer is still largely based on shotgun approaches ("one-size-fits-all"). Only a subset of tumors respond to the drugs prescribed, due to the biological heterogeneity of the disease. On top of this, drug toxicity that patients experience independent of response is a major concern.[18] To avoid unnecessary toxicity and reduce healthcare costs, markers to predict whether patients will respond to a given therapy are urgently needed.[19] The success of both classical drug therapies as well as of novel targeted therapies can be improved by matching biomarkers with the right combination of drugs, i.e. " personalized therapy " as opposed to "one-size-fits-all".

Biomarkers correlating with outcome can be either prognostic or predictive. A prognostic biomarker correlates with clinical outcome, independent of therapy. For example CRC patients with MSI have a better prognosis than CIN tumors and CRC that show loss of heterozygosity at the 18q arm (18qLOH) appear to have worse prognosis compared with tumors that do not carry this biomarker.[20,21]

Predictive biomarkers classify patients into subgroups who are most likely to respond to a specific therapy.[22] Predictive biomarkers can be classified in pharmacogenetics, i.e. constitutional genetic features that influence drug metabolism, and pharmacogenomics which includes somatic alterations, i.e. changes in the tumor DNA. Many candidates

for predictive biomarkers have been proposed for chemotherapy treatment of mCRC. However, none of these molecular markers have been implemented in the standard of care for colorectal cancer patients due to divergent and inconsistent results on their predictive value.[23,23,24] The only predictive biomarker for mCRC that has been implemented in routine clinical practice is the KRAS mutation status to predict response to the anti-EGFR agents cetuximab and panitumumab.[25-27] Approximately 40% of colorectal cancer patients have a KRAS mutation.[28,29] Inhibition of EGFR leads to inhibition of the RAS-RAF-ERK pathway. Mutation of KRAS, results in activation of the RAS-RAF-ERK pathway downstream of EGFR, and counteracts activity of anti-EGFR therapies.

Prediction by using genome-wide analysis

Development of new laboratory genomics techniques such as microarrays and massive parallel sequencing (MPS) accompanied by progress in sophisticated data-analysis by biostatisticians and bioinformaticians as described in the sections "The development of genomic laboratory techniques" and "Genome-bioinformatics" has opened the way for genome-wide studies. For example, genome-wide expression profiling revealed breast-cancer expression signatures to predict recurrence risk after surgery. Two signatures, MammaPrint and Oncotype DX, have been approved by FDA and used in clinical practice in the US.[30,31] In addition genome-wide expression profiling revealed two EGFR ligands, namely Amphiregulin (AREG) and Epiregulin (EREG), which have been associated with response to cetuximab treatment.[32]

Other genome-wide studies demonstrated that colorectal cancer patients who do respond to systemic combination therapy with capecitabine and irinotecan have regions located on chromosome 18 frequently deleted.[33]

Biomarkers and metastases

Twenty percent of CRC patients have developed metastases at time of diagnosis and up to 50% of all CRC patients develop metastasis at any time during the course of their disease.[34] The liver is the predominant metastatic site in approximately 80% of mCRC patients. Other metastatic sites include lung, the central nervous system, adrenal glands, ovaries, skeleton and skin.[35] Even though therapies are targeted against metastases and metastasis is the principal event leading to death in mCRC patients, current clinical practice is to use archived material of the primary tumor to analyze with genomics for therapy selection. Consequently predictive biomarkers are based on the biological properties of primary tumors. Biological properties of tumor cells to metastasize are obtained by the accumulation of (epi)genetic alterations. If these alterations occur early in the development of the primary tumor, selecting therapy does not require metastatic tumor tissue. There is accumulating evidence to support the hypothesis that the full genomic program that determines the biological and clinical phenotype of a tumor is already present at the time when a primary cancer arises.[36-40]

THE DEVELOPMENT OF GENOMIC LABORATORY TECHNIQUES

Genomics is the study of all chromosomes, genes, proteins and their function. Techniques have been developed for DNA, RNA and protein measurements on a genome-wide scale.

The knowledge that cells contain 46 chromosomes in 23 pairs, was discovered by karyotyping. Karyotyping was introduced in the late 1950s and subsequently lead to the discovery of gross chromosomal aberrations. Those aberrations were associated to several diseases including cancer.[41] In the 1960s chromosomal aberrations could be studied more precisely by the development of a technique, called chromosomal banding. However, this technique was limited by its low spatial resolution. Based on karyotyping in combination with fuorescent in situ hybridization, comparative genomic hybridization (CGH) was introduced in the 1990s, offering the possibility to genome-wide explore numerical chromosomal aberrations in tumor DNA, even from formalin fixed paraffin embedded material.[42] This method was the predecessor of array based CGH discussed in section "The development of microarrays".

Genomics thus is an old discipline, but the word itself only came into fashion around the turn of the 21st century when it took a new direction with the sequencing of the human genome. In 2001 the first draft of the human DNA sequence was published by the International Human Genome Sequencing Consortium (IHGSC).[43] One day later Celera Genomics published a paper in Science reporting the draft human sequence based on their own data.[44] The IHGSC and Celera Genomics used different approaches for the sequencing. Both were based on Sanger sequencing.[45] IHGSC constructed a physical map by mapping a library of 300,000 bacterial artificial chromosomes (BACs) and subsequently sequencing each individual clone. BAC clones are large clones of 150 to 200kb. Celera Genomics used an alternative approach, namely the whole-genome shotgun sequencing[46] for which no prior knowledge of the physical map was required per se. Random genomic DNA fragments were assembled after sequencing, and ordered based on their overlapping sequence.

The complete human genome sequence radically changed molecular biology research from hypothesis-driven research to more unbiased hypothesis-free data driven research as described in this thesis.

Development of laboratory techniques and tools in Human genomics rapidly commenced after the completion of Human Genome Project (HGP) in 2003.[47] The sequence of 2001 covered 90% of genome, interrupted by 250,000 gaps. By 2004 99.7% of genome was sequenced and only 300 gaps were left.[48] This rapid technical development went hand in hand with an unsurpassed development of data storage, statistics and analysis.

The development of microarrays

Microarrays were the first important genomic technique that was developed based on and in parallel to the HGP. A microarray is a glass slide on which thousands of different pieces of DNA are printed as tiny little dots, called clones or probes. They are available

in different flavors, either to measure mRNA levels (expression arrays) or DNA copy numbers (array comparative genomic hybridization (arrayCGH)). In cancer research microarrays brought about an enormous acceleration of the research and discovery of genes involved in cancer and subsequently the development of diagnostic tools. The first type of array that was developed were expression arrays.[49] To produce these microarrays pieces of cDNA were spotted to measure mRNA levels in the cell.[50] cDNAs are short DNA molecules derived from mRNA. Each probe represents the expression of one single gene, which makes it possible to measure thousands of genes in parallel. Changes in expression patterns in cancer cells can be caused by chromosomal aberrations, but also external factors or circadian rhythm.

Rather than the use of cloned cDNA fragments, synthetically produced oligonucleotides were spotted on the glass slide to produce the arrays in 2005.[51] Nowadays several commercially produced expression arrays are available such as Agilent, Affymetrix, NimbleGen and Illumina. Each company uses different techniques to produce these arrays. For example, Affymetrix uses very short match and mismatch probes. Illumina uses beads with long oligonucleotide probes. Even though different techniques are used, The MicroArray Quality Control (MAQC) project reported that microarray measurements are highly reproducible within and across different microarray platforms by evaluating six different commercially available platforms.[52] A consequence of different microarray platforms and experimental designs is also that data output is generated in various formats. Additionally they are preprocessed in different ways, which makes comparison and integration of these data a bioinformatics challenge. MIAME, the Minimum Information About a Microarray Experiment was proposed to facilitate the organization of microarray data in databases.[53] Gene Expression Omnibus (GEO)[54] is an example of a database which support MIAME-compliant data submissions. GEO is a public repository that archives and freely distributes microarray and other genome-wide data submitted by the scientific community. The data produced and used in this thesis are submitted to GEO.

Measuring chromosomal copy number aberrations by arrays

In this thesis chromosomal copy number measurements on a genome-wide scale have been pivotal. DNA copy number aberrations (CNAs) are a hallmark of cancer. Chromosomal losses, gains and amplifications can change the level of gene expression thereby contributing to tumorigenesis and influence response to treatment and prognosis. One of the most prominent examples is the amplification of ERBB2 which is used as a predictive marker for treatment of breast and gastric cancer patients.[55-57]

In 1992, prior to array-comparative genomic hybridization (arrayCGH), comparative genomic hybridization (CGH) was developed to measure copy number alterations at a genome-wide scale.[58] Tumor and reference DNA were labeled with dyes in two different colors, e.g. Cy3 and Cy5, and hybridized to metaphase chromosomes to be able to calculate relative differences between tumor and reference DNA. The CGH technique was further refined through introduction of arrayCGH.[59,60] Each array contained 2400-5000 clones of the human genome, BAC clones.[61]

Later, high-resolution arrayCGH platforms with oligonucleotides, containing DNA of 50-60 bp long and single nucleotide polymorphism (SNP) genotyping platforms with 25bp probes came commercially available.[62] Nowadays, these arrays are available containing up to 6 million probes.

An important aspect of arrayCGH is that the technique works well with DNA extracted from formalin fixed paraffin embedded (FFPE) archival tissue, which in routine clinical practice often is the only material available. We compared different commercially available platforms for the quality of DNA copy number aberrations detection with DNA isolated from FFPE material.[63]

The ever-improving spatial resolution for the detection of CNAs by arrayCGH has led to the discovery of focal aberrations (defined as aberrations 3Mb or smaller in size) and copy number variations (CNVs). Varambally et al (2008) discovered that decrease of microRNA-101 was caused by a focal deletion of a region on which this microRNA is located.[64] In the same year several papers were published that all made use of high resolution arrays in different tumor types.[65-69] Many aberrations concern large chromosomal regions, which makes it difficult to distinguish driver from passenger genes. Studying focal aberrations overcomes this problem since there are only a few genes located on these small gains and losses, making it easier to pinpoint genes involved in cancer. Focal chromosomal aberrations are described in chapter 2 "Candidate driver genes in focal chromosomal aberrations of stage II colon cancer." In this chapter we used 44K arrays, which had sufficiently high resolution to detect focal chromosomal aberrations. We show that by studying these short chromosomal regions we were able to identify driver genes.

Genome consortia and catalogues

Genomics has accelerated the discovery of cancer causing abnormalities and unraveling the pathways involved in cancer development and metastasis. Thereby new prognostic and predictive biomarkers have been discovered.

Currently, large projects like the The Cancer Genome Atlas (TCGA)[70,71] and the International Cancer Genome Consortium (ICGC)[72] are studying ~500 tumors per cancer type. The primary goals of these consortia are to generate catalogues of genomic abnormalities such as somatic mutations, gene expression and epigenetic modifications in tumors from 50 different cancer types and make these data available to the entire research community as rapidly as possible. The first genome landscape papers that have been published are for glioblastomas, colorectal, pancreas and breast cancer[65-67] and more cancer types will follow.

Also projects to catalogue genes that are mutated in cancer have been established. The Welcome Trust Sanger Institute started the Cancer Gene Census, a project to list genes that have been reported to contribute to oncogenesis.[73] By 2004 291 genes had been collected and currently already 487 genes are listed. A working list in the format of an excel file is available at their website (http://www.sanger.ac.uk/genetics/CGP/Census/) and has been

annotated with information like chromosomal location, tumor type in which mutations are found and the type of alteration (i.e. translocation, mutation, copy number aberration). Other initiatives are the Catalogue Of Somatic Mutations In Cancer (COSMIC),[74] which aims to provide somatic mutation frequencies of the genes in the cancer census gene list. Samples entered include benign and malignant tumors and cancer cell lines.

There are also catalogues available to directly correlate genes to drug response, such as The Genomics of Drug Sensitivity in Cancer project (http://www.cancerrxgene.org).

GENOME-BIOINFORMATICS

A leading branch in bioinformatics is the use of computer methods in studies of genomes. Genome-bioinformatics involves analysis and interpretation of various types of data including DNA and RNA sequences.

Since the genome sequences of several organisms are available, tools are developed for genome-wide approaches and in parallel lots of data are generated. These data need to be stored in databases, which need to be maintained and made accessible for researchers. Human genome sequence data are accessible through for example the UCSC Genome Browser or Ensembl.[75,76] Both visualization as well as analytical tools have been developed and made available through these genome browsers. Development and implementation of tools has been done in several programming languages such as JAVA, C, Python, Perl, Matlab and R. In addition web-based interfaces are developed that provide access to these bioinformatical tools. For example GenePattern (http://www.broadinstitute.org/cancer/software/genepattern/), developed by the Broad Institute, is a genomic analysis platform that provides access to more than 160 tools for gene expression analysis, proteomics, SNP analysis and common data processing tasks. Genepattern was not used to analyze the data described in this thesis since it falls short of DNA copy number data tools. Another means to access and share analysis tools is offered by Bioconductor[77] (www.bioconductor.org). Bioconductor is an open source project that aims to share statistical and graphical methods for the analysis and comprehension of high-throughput genomic data and is primarily based on the R programming language. The functional scope of Bioconductor packages includes the analysis of expression and CGH arrays, sequence and other data. A large part of the data analysis described in this thesis has been performed in R-packages available through Bioconductor.

Also, analysis tools are commercially available and popular in use since they are more user-friendly. An example of user-friendly software for DNA copy number analysis is Nexus Copy Number (http://www.biodiscovery.com/software/nexus-copy-number).

Genomics has become a major factor in translational research (i.e. arriving at better outcome for patients by linking biology underlying a disease to its phenotype), and

consequently also data management and analysis, including genome-bioinformatics. The TraIT (Translational research IT) project, that is running in the context of the Center for Translational Molecular Medicine, is a major exercise to bring translational researchers, genomics and IT experts and bioinformaticians together to further professionalize this field.

Data analysis, statistics and arrayCGH

Preprocessing

Statistical methods for analyzing DNA copy number data are aimed at identifying genomic locations with an aberrant copy number. Prior to the downstream interpretation of the data, microarray experiments need to be preprocessed to make a series of arrays interpretable.

The preprocessing of arrayCGH data is first done for individual experiments, rather than all experiments at the same time. The first step of preprocessing is done by feature extraction software. Feature extraction is the process of converting different intensities into a number that reflects the fluorescence intensity of the probes on the array. Subsequently, log2-ratios are calculated by dividing the intensities of the test sample by the intensities of the reference sample followed by a log2 conversion. In addition the probes are sorted in chromosomal order, which enables to display DNA copy number profile plots with chromosomal positions on the x-axis and the log2-ratio's on the y-axis. Before further analysis some technical artifacts can be removed.[78] For example waves which are mainly caused by variable GC percentage in probes may appear in the profiles. These waves can introduce noise into the data. In this thesis waves from arrayCGH profiles have been removed through calibration on a series of reference profiles.[79] Tumor samples contain a mixture of tumor and normal cells which complicates the analysis and interpretation. A pathologist usually estimates the tumor cell percentages for which the results can be corrected.[80] Normalization is needed to make log2-ratios of samples on different microarrays comparable and to asses the value of a profile corresponding to two copies. Assuming that the majority of probes corresponds to two copies of DNA, mode-subtraction for normalization often works best, however many other ways of normalization have been reported.[78]

After normalization of the profiles, in this thesis, a segmentation algorithm is applied.

Segmentation algorithms are used to detect the chromosomal breakpoints and divide the genome into neighboring segments. All data points located on one segment are assumed to have the same log2-ratios and thereby the same underlying copy number. For this purpose circular binary segmentation (CBS)[81] was used, the most commonly used segmentation algorithm.[82] The statistics of CBS used to determine breakpoints is similar to Student statistic. A p-value provides information about the strength of the breakpoint and if the p-value is below a certain threshold it is defined as breakpoint. The procedure is applied recursively to find all breakpoints in the data.

After segmentation we apply in our group at the VUmc Cancer Center Amsterdam the in-house developed calling algorithm CGHcall[80] available in Bioconductor[77] to classify segment data into copy number states (>2 copies), loss (< 2 copies), normal (2 copies) or amplification (>= 5 copies). The reason to use a calling algorithm is that it automates data interpretation, which has advantages for down-stream analysis.[83] The start and end of the segments and the log2 ratio's of the probes located on the segments are combined with a mixture model to obtain the most likely classification per segment rather than per individual clone. Posterior probabilities are returned which are transformed into calls by use of Bayes'rule. The final output contains discrete data of the called gains and losses and the posterior probabilities. The latter is used for clustering of called data.

Experiments run on multiple platforms

Datasets from different studies can be reused and merged to answer new research questions, e.g. to increase the size of datasets, to compare different datasets[84] or to test a newly built classifier in an independent dataset.[30] For example publicly available datasets in GEO[54] can be downloaded and combined. However different arrayCGH platforms may have been used varying from BAC arrays, usually containing around 3000 large probes, to high-resolution oligonucleotide arrays containing up to 6 million probes.

Before combining the data, the probes on the arrays need to be matched.[85] In addition log2-ratio's need to be made comparable. To be able to confirm that differences between samples are due to features of the samples rather than the different platforms, some samples can be hybridized on the respective platforms used for calibration purposes as seen in Jong et al.[84] and Haan et al.[86] In chapter 3 "Small bowel adenocarcinoma copy number profiles are more closely related to colorectal than to gastric cancers" we compared small intestinal cancers with gastric and colorectal cancer selected from three independent studies. The experiments were performed on two different platforms namely, BAC arrays and 30K oligonucleotide arrays. To make the different platforms comparable by calibration 6 samples were run on both platforms. The data were matched by resampling the data points to 2000 data points and the log2-ratios of 30K arrays were calibrated to the BAC arrays. This method made it possible to combine different studies for a comparison analysis.

Dimension reduction

A chromosomal aberration can be large, covering a whole chromosomal arm but also a focal aberration. The size of an aberration does not necessarily reflect its biological or clinical relevance. However, in multivariate downstream analysis, like clustering, larger aberrations will have larger effects than smaller aberrations. This can be solved by dimension reduction. After dimension reduction each gain or loss will have the same weight such that a focal aberration would have an equal weight in the cluster analysis as a long "dull" aberration. Moreover p-values obtained with univariate analysis of multiple variables, e.g. arrayCGH data points, need to be corrected for multiple testing as described in the next section "Downstream analysis". A large number of data points may affect

the corrected p-values, leaving no significant results due to too conservative correction methods. For these reasons the R-package CGHregions[87] was developed. This algorithm was designed to convert series of neighboring clones on the chromosome whose arrayCGH-signature, a vector of calls across the samples, is shared by all clones to one datapoint, named a region. Regions capture the essential features of the data and may contain one important aberration containing just a few probes or a whole chromosome arm.

Genomic Identification of Significant Targets

In section "Measuring chromosomal copy number aberrations by arrays" and chapter 2 of this thesis we describe an approach to distinguish driver genes from passenger genes by studying focal aberrations. Genomic Identification of Significant Targets in Cancer (GISTIC)[88] is another approach to identify driver genes. This method identifies regions of the genome that are aberrant more often than would be expected by chance in a whole series of tumors. The regions are determined by a score based on both the amplitude and frequency of copy-number changes for each probe, using permutation testing to determine significance. This method may lead to the discovery of novel drug targets, since genes located on such regions are likely to be involved in carcinogenesis.

In contrast to the preprocessing steps described above in section "Preprocessing" for this method normalized log2-ratios of a group of samples of interest are used as input and preprocessing is done within the GISTIC algorithm. This method was used in chapter 6, "Chromosomal copy number aberrations in colorectal metastases resemble their primary counterparts and differences are typically non-recurrent". In this chapter we aimed to identify metastasis specific regions by comparing a group metastases to their corresponding primary tumor in the same patients. First we subtracted the log2-ratios of the primary tumors from the metastases, resulting in a combined dataset. We applied the GISTIC algorithm to this combined dataset to identify those regions that were significantly more often aberrant in the metastases compared to the primary tumors. We identified two regions that were only amplified in 3 samples, but with a high copy number.

Downstream analysis

After preprocessing the data are combined into a single file containing all the sample identifiers and (dimension reduced) data points. Data can be visualized and explored by plotting frequencies of gains and losses of groups of profiles.

Different kind of downstream analyses are performed in this thesis, depending on the research question. Downstream analysis can be split into two basic approaches, namely supervised and unsupervised analysis. Supervised analysis is used to explore genomic differences between a priori defined (e.g. clinically relevant) categories. Unsupervised analysis is used to discover subgroups independent of any prior knowledge of categories existing within the sample series studied.

Unsupervised analysis

Cancer is a heterogeneous disease containing many subgroups. Existing classifications can be based on e.g. site of origin, histopathological subtypes, stage, outcome or response to treatment.. Unsupervised cluster analysis aims to identify subgroups based on the genomics data alone, and these results can then be correlated to these existing classifications. For example Jong et al.[84] demonstrated that e.g. adenocarcinomas from different sites of origin clustered together based on arrayCGH data, when compared to e.g. tumors of mesenchymal origin. In addition cluster analysis can reveal clonal relations between samples. For example Stange[37] et al used unsupervised hierarchical clustering to show that primary tumors and their corresponding metastases of the same patient were more similar to each other than primary tumors across patients. Clustering algorithms were first used for mRNA array expression profiles, including Pearson or Spearman correlation as distance measurements. The same clustering algorithms have been used for DNA copy number profiles with normalized or segmented data as input. Later dedicated algorithms for handling called data were developed.[78] In this thesis weighted clustering of called aCGH data (WECCA)[89] was used and is described here in more detail. WECCA is a hierarchical clustering method tailor-made for called and regioned arrayCGH data, i.e. data on an ordinal scale consisting of "loss", "normal", "gain" and "amplification".

Different settings for distance measure and linkage are available in WECCA. In addition WECCA also has an option to give weights to regions so that different chromosomal regions can have variable contribution to the clustering, e.g based on a priori knowledge of the relevance of a given genomic region within a disease or because a given genomic region has a larger gene density in the array design used. Recently a modified version of WECCA has become available that uses call probabilities as generated by CGHcall[80] instead of calls. The use of called probabilities will give a more subtle picture of the similarities and differences between the samples. As an output, WECCA generates a dendrogram and a heatmap of the copy number data.

With hierarchical clustering the number and stability of the clusters are not defined yet. The number of clusters described is usually based on the results and preferences of the researcher. To overcome this problem several solutions have been implemented, one of them is consensus clustering.[90] With consensus clustering the number of the most stable clusters can be determined by running the clustering algorithms many times (e.g. 1000) and leaving out a percentage of the samples each time. Robustness of clusters is determined by counting the number of times pairs of samples appear in the same cluster.

In chapter 3 "Small bowel adenocarcinoma copy number profiles are more closely related to colorectal than to gastric cancers" we aimed to investigate whether copy number profiles of small bowel cancers are more similar to gastric or colorectal cancers. By the use of unsupervised hierarchical clustering we could demonstrate that small bowel cancer is more similar to colorectal cancer than to gastric cancer, since the clustering resulted in two main clusters, one containing the majority of colorectal and small bowel cancers and the other

one containing the majority of gastric cancers. Moreover we could confirm the robustness of these two clusters by the use of consensus clustering.

Supervised analysis

DNA copy number profiles may be correlated to clinical outcome features such as survival, relapse or response to drug therapy. For this kind of research questions supervised analysis can be used. Individual aberrant regions can be associated with outcome, this is called univariate analysis. Multivariate analysis is a procedure when multiple variables, or even entire genomic profiles, are correlated with outcome at the same time.

For univariate analysis of called or regioned data the R-package CGHMultiArray[91] was developed. In this package several tests are available including chi-square test and Wilcoxon signed-rank test corrected for ties to compare frequencies of gains and losses in different groups and the log-rank test to correlate gains and losses to survival. In this package a column-wise permutation for null-distribution is implemented to estimate the p-value. Univariate analysis with a large number of data points has the disadvantage of a high number of false-positives, even after dimension reduction. This statistical dilemma requires correction for multiple testing. Different methods to adjust for multiple testing have been developed. The most well known is the Bonferroni method, which multiplies the p-values by the number of tests. Bonferroni is usually too conservative since arrays measure many data points, many of which are not independent, and usually a limited number of samples. For copy number data analysis a less conservative method is the frequently used Benjamini-Hochberg method to control the False Discovery Rate (FDR).[78] Benjamini and Hochberg first ranked the p-values from smallest to largest. The largest p-value retains its value. The second largest p-value is multiplied by the number of tests divided by its rank and so on for all p-values, resulting in p-values that have been corrected for multiple testing.

In this thesis, univariate testing was performed for several association studies. In chapter 5 “Genomic landscape of metastatic colorectal cancer”, we aimed to correlate gains and losses to response to therapy. For this purpose we performed the univariate log-rank test to individually correlate each region to progression free survival. In chapter 6 “Chromosome 20p11 gains contribute to hepatic-specific metastasis in colorectal cancer patients” we aimed to correlate gains and losses to metastatic sites. For this purpose we performed the univariate chi-square test to compare frequencies of gained and lost regions in hepatic and extrahepatic metastases.

Next to the association studies of individual regions, whole DNA copy number profiles can be correlated to outcome by classification. Classification is a supervised multivariate analysis and gives a different result then unsupervised clustering since prior knowledge of the groups is available. Classification aims to construct a rule (classifier) that assigns objects (tumors) to pre-specified classes (“dead“ vs. “alive“ or “responder“ vs. “non-responder”) on the basis of measurements (copy number profiles). For classification a training set as well as a test set are needed. The training set is used to determine the

classifier and the test set should be independent of the training set and is used to evaluate the performance of the classifier. Instead of using a training set and test set, leave-one-out cross-validation (LOOCV) can be used. Each sample is left out in turn, then the model fit on the remaining N-1 samples, the left out sample is supplied and its class predicted the average of the prediction errors is used to estimate the training error. Although statistically sound, this method is not accepted in medicine since they require validation in an independent dataset.

To identify a set of features to build a classifier different algorithms can be used.

Examples of algorithms are k-nearest neighbour, naive Bayes and support vector machine (SVM) and the R-packages Prediction Analysis for Microarrays (PAM)[92] and Random forest. The latter builds decision trees. For the analysis of copy number data the method "fused SVM" was proposed.[78] The method is a variant of the SVM that incorporates the biological specificities of DNA copy number variations along the genome as prior knowledge.

To study the biological relevance of DNA copy number data, copy number values of genes can be correlated to mRNA expression values of the same sample if available. Genes located on gained or amplified regions are expected to have a higher expression than genes located on normal or lost regions. Genes located on lost regions are expected to have lower gene expression. Since copy number profiles consist of thousands of data points located on inter-and intragenic regions, first the copy number values of the genes need to be determined by mapping. For mapping of data points to genes, copy number information of the segments should be used, rather than the probes. A Wilcoxon-rank test can be performed to test significant differences of expression levels between "loss", "normal", "gain" and "amplification". This method, including different mapping methods, has been implemented in the Rpackage ACE-it (Test for copy number impact on gene expression).[93,94]

We performed a Wilcoxon-rank test to correlate DNA copy number to gene expression in three chapters, by comparing gene expression levels of samples containing either "loss" versus "no loss" or "gain" versus "no gain". In chapter 2 "Candidate driver genes in focal chromosomal aberrations of stage II colon cancer" we used this method to confirm the biological relevance of the genes located on the focal chromosomal aberrations. By performing such a test we were able to demonstrate that genes located on focal losses had lower expression in comparison to genes not located on losses. Genes located on amplifications had higher expression compared to genes not located on amplifications. In chapter 5 "Genomic landscape of metastatic colorectal cancer" and chapter 6, "Chromosome 20p11 gains contribute to hepatic-specific metastasis in colorectal cancer patients" we performed such a test to identify biologically relevant genes located on the statistically significant regions. However no expression profiles of the same samples were available. For this reason we downloaded expression and copy number microarray data of colorectal cancer samples from the TCGA dataset (described in section "Genome consortia and catalogues") as a validation set. In this way we were able to identify potential genes responsible for drug resistance or metastasis to the liver in chapters 5 and 6 respectively.

AIMS AND OUTLINE OF THIS THESIS

The overall aim of this thesis is to identify chromosomal copy number markers of colorectal and other gastrointestinal cancers as candidate biomarkers for clinical use.

In the first part we aim to investigate whether focal aberrations can be detected with arrayCGH and DNA extracted from FFPE archival tissue, which in routine clinical practice often is the only material available. In the second part of the thesis the aim is to investigate whether copy number profiles may have clinical diagnostic value. In the third part we investigate signatures of CRC primary tumors and their metastases.

The first part contains one chapter with the title "Candidate driver genes in focal chromosomal aberrations of stage II colon cancer". The aim of this study is to identify recurrent focal chromosomal aberrations and their candidate driver genes in a well-defined series of stage II colon cancers. Since the increasing resolution of platforms, focal aberrations became detectable with arrayCGH. These focal aberrations are defined as smaller than 3Mb on which only a few genes are located, which makes it easier to pinpoint the driver genes.

The second part contains three chapters (3-5) in which we investigate the potential of DNA copy number aberrations in gastrointestinal tumors to guide therapy selection. In chapter 3, "Small bowel adenocarcinoma copy number profiles are more closely related to colorectal than to gastric cancers" we give address to the important clinical question whether DNA copy number profiles of small intestinal tumors are more similar to gastric or to colorectal cancers. Small intestinal cancer is a rare disease and for that reason opportunities for running clinical trials to determine the best treatment are limited. In practice these tumors are treated either like gastric or colorectal cancers. Comparing their copy number profiles to gastric and colorectal cancers, may reveal to which of these they most resemble at the genomic level, which in turn may support the choice for either a gastric cancer or colorectal cancer type of therapy.

The aim of chapter 4 "High level copy number gains of established and potential drug target genes in gastric cancer as a potential lead for treatment development and selection", is to identify drug targets in gastric cancer by cataloging genes that are gained at high levels in gastric cancer and evaluate their potential as drug targets. Most current developments of therapies focus on new combination treatment strategies including biological agents that target specific molecules. Copy number gains at the DNA level is an important mechanism of gene overexpression and may serve as a screen for potential candidate drug targets relevant to the individual patient. In chapter 5, "Genomic landscape of metastatic colorectal cancer" the aim is to document the landscape of DNA copy number changes in primary tumors of a defined subset of colorectal cancer, i.e. patients that developed metastatic disease. Patients were selected from two different phase III randomized clinical trials, CAIRO[95] and CAIRO2.[96] For that reason we are able to correlate arrayCGH profiles to response to treatment and produce a landscape of genes based on the results. In

addition, as in chapter 4, we catalogue high level copy number gains of loci carrying genes against the products of which already drugs exist and have been approved for use.

The third part of the thesis contains two chapters (6 and 7) and is about copy number patterns of metastases. In chapter 6, "Chromosome 20p11 gains contribute to hepatic-specific metastasis in colorectal cancer patients" the aim is to identify chromosomal aberrations associated with hepatic versus extra-hepatic metastases in CRC patients.

In chapter 7, "Chromosomal copy number aberrations in colorectal metastases resemble their primary counterparts and differences are typically non-recurrent" the aim is to compare copy number profiles of primary tumors of CRC to their corresponding metastasis in the same patient. In routine clinical practice, molecular diagnostics for guiding therapy selection in patients with metastatic CRC for practical reasons is performed on tissue samples from the primary tumor, under the assumption that the genomic aberrations present are highly similar to those in the metastases. While this assumption has been proven to be true for KRAS mutation status,[97] this remained to be demonstrated for high resolution DNA copy number data.

REFERENCES

1. Jemal A, Bray F, Center MM *et al*. Global cancer statistics. CA Cancer J Clin 2011;61(2):69-90.
2. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell 1990;61(5):759-67.
3. Lengauer C, Kinzler KW, Vogelstein B. Genetic instabilities in human cancers. Nature 1998;396(6712):643-9.
4. Sillars-Hardebol AH, Carvalho B, van EM *et al*. The adenoma hunt in colorectal cancer screening: defining the target. J Pathol 2012;226(1):1-6.
5. Baylin SB, Esteller M, Rountree MR *et al*. Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. Hum Mol Genet 2001;10(7):687-92.
6. Esteller M, Corn PG, Baylin SB *et al*. A gene hypermethylation profile of human cancer. Cancer Res 2001;61(8):3225-9.
7. Sobin LH, Wittekind CH. *TNM Classification of Malignant Tumours*, sixth edition, UICC. 6 ed. New York: Wiley-Liss: 2002.
8. Gill S, Loprinzi CL, Sargent DJ *et al*. Pooled analysis of fluorouracil-based adjuvant therapy for stage II and III colon cancer: who benefits and by how much? J Clin Oncol 2004;22(10): 1797-806.
9. Modulation of fluorouracil by leucovorin in patients with advanced colorectal cancer: evidence in terms of response rate. Advanced Colorectal Cancer Meta-Analysis Project. J Clin Oncol 1992;10(6):896-903.
10. Van Cutsem E, Hoff PM, Harper P *et al*. Oral capecitabine vs intravenous 5-fluorouracil and leucovorin: integrated efficacy data and novel analyses from two large, randomised, phase III trials. Br J Cancer 2004;90(6):1190-7.
11. de Gramont A, Figer A, Seymour M *et al*. Leucovorin and fluorouracil with or without oxaliplatin as first-line treatment in advanced colorectal cancer. J Clin Oncol 2000;18(16):2938-47.
12. Punt CJ. New options and old dilemmas in the treatment of patients with advanced colorectal cancer. Ann Oncol 2004;15(10):1453-9.
13. Van CE, Tabernero J, Lakomy R *et al*. Addition of aflibercept to fluorouracil, leucovorin, and irinotecan improves survival in a phase III randomized trial in patients with metastatic colorectal cancer previously treated with an oxaliplatin-based regimen. J Clin Oncol 2012;30(28):3499-506.
14. Grothey A, Van CE, Sobrero A *et al*. Regorafenib monotherapy for previously treated metastatic colorectal cancer (CORRECT): an international, multicentre, randomised, placebo-controlled, phase 3 trial. Lancet 2013;381(9863):303-12.
15. Chu E. An update on the current and emerging targeted agents in metastatic colorectal cancer. Clin Colorectal Cancer 2012;11(1):1-13.
16. O'Connell JB, Maggard MA, Ko CY. Colon cancer survival rates with the new American Joint Committee on Cancer sixth edition staging. J Natl Cancer Inst 2004;96(19):1420-5.
17. Kopetz S, Chang GJ, Overman MJ *et al*. Improved survival in metastatic colorectal cancer is associated with adoption of hepatic resection and improved chemotherapy. J Clin Oncol 2009;27(22):3677-83.

18. Verheul HM, Pinedo HM. Possible molecular mechanisms involved in the toxicity of angiogenesis inhibition. Nat Rev Cancer 2007;7(6):475-85.
19. Meijer GA, Oudejans JJ. Targeted therapies; who detects the target? Cell Oncol 2005;27(3):165-7.
20. Pritchard CC, Grady WM. Colorectal cancer molecular biology moves into clinical practice. Gut 2011;60(1):116-29.
21. Lee JK, Chan AT. Molecular Prognostic and Predictive Markers in Colorectal Cancer: Current Status. Curr Colorectal Cancer Rep 2011;7(2):136-44.
22. Koopman M, Venderbosch S, van TH *et al*. Predictive and prognostic markers for the outcome of chemotherapy in advanced colorectal cancer, a retrospective analysis of the phase III randomised CAIRO study. Eur J Cancer 2009;45(11):1999-2006.
23. Koopman M, Venderbosch S, Nagtegaal ID *et al*. A review on the use of molecular markers of cytotoxic therapy for colorectal cancer, what have we learned? Eur J Cancer 2009;45(11): 1935-49.
24. Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers. J Natl Cancer Inst 2009;101(21):1446-52.
25. De Roock W., Piessevaux H, De SJ *et al*. KRAS wild-type state predicts survival and is associated to early radiological response in metastatic colorectal cancer treated with cetuximab. Ann Oncol 2008;19(3):508-15.
26. Karapetis CS, Khambata-Ford S, Jonker DJ *et al*. K-ras mutations and benefit from cetuximab in advanced colorectal cancer. N Engl J Med 2008;359(17):1757-65.
27. Amado RG, Wolf M, Peeters M *et al*. Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. J Clin Oncol 2008;26(10):1626-34.
28. Andreyev HJ, Norman AR, Cunningham D *et al*. Kirsten ras mutations in patients with colorectal cancer: the multicenter "RASCAL" study. J Natl Cancer Inst 1998;90(9):675-84.
29. Bos JL, Fearon ER, Hamilton SR *et al*. Prevalence of ras gene mutations in human colorectal cancers. Nature 1987;327(6120):293-7.
30. Smeets SJ, Harjes U, van Wieringen WN *et al*. To DNA or not to DNA? That is the question, when it comes to molecular subtyping for the clinic! Clin Cancer Res 2011;17(15):4959-64.
31. van 't Veer LJ, Dai H, van de Vijver MJ *et al*. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415(6871):530-6.
32. Khambata-Ford S, Garrett CR, Meropol NJ *et al*. Expression of epiregulin and amphiregulin and K-ras mutation status predict disease control in metastatic colorectal cancer patients treated with cetuximab. J Clin Oncol 2007;25(22):3230-7.
33. Postma C, Koopman M, Buffart TE *et al*. DNA copy number profiles of primary tumors as predictors of response to chemotherapy in advanced colorectal cancer. Ann Oncol 2009;20(6):1048-56.
34. Kindler HL, Shulman KL. Metastatic colorectal cancer. Curr Treat Options Oncol 2001;2(6):459-71.
35. Hermanek P, Jr., Wiebelt H, Riedl S *et al*. [Long-term results of surgical therapy of colon cancer. Results of the Colorectal Cancer Study Group]. Chirurg 1994;65(4):287-97.

36. Ramaswamy S, Ross KN, Lander ES *et al*. A molecular signature of metastasis in primary solid tumors. Nat Genet 2003;33(1):49-54.

37. Stange DE, Engel F, Longerich T *et al*. Expression of an ASCL2 related stem cell signature and IGF2 in colorectal cancer liver metastases with 11p15.5 gain. Gut 2010;59(9):1236-44.

38. Vakiani E, Janakiraman M, Shen R *et al*. Comparative genomic analysis of primary versus metastatic colorectal carcinomas. J Clin Oncol 2012;30(24):2956-62.

39. Jones S, Chen WD, Parmigiani G *et al*. Comparative lesion sequencing provides insights into tumor evolution. Proc Natl Acad Sci U S A 2008;105(11):4283-8.

40. Kloosterman WP, Hoogstraat M, Paling O *et al*. Chromothripsis is a common mechanism driving genomic rearrangements in primary and metastatic colorectal cancer. Genome Biol 2011;12(10):R103.

41. Oostlander AE, Meijer GA, Ylstra B. Microarray-based comparative genomic hybridization and its applications in human genetics. Clin Genet 2004;66(6):488-95.

42. Weiss MM, Hermsen MA, Meijer GA *et al*. Comparative genomic hybridisation. Mol Pathol 1999;52(5):243-51.

43. McPherson JD, Marra M, Hillier L *et al*. A physical map of the human genome. Nature 2001;409(6822):934-41.

44. Venter JC, Adams MD, Myers EW *et al*. The sequence of the human genome. Science 2001;291(5507):1304-51.

45. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 1977;74(12):5463-7.

46. Venter JC, Adams MD, Sutton GG *et al*. Shotgun sequencing of the human genome. Science 1998;280(5369):1540-2.

47. Finishing the euchromatic sequence of the human genome. Nature 2004;431(7011):931-45.

48. Lander ES. Initial impact of the sequencing of the human genome. Nature 2011;470(7333): 187-97.

49. Schena M, Shalon D, Davis RW *et al*. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995;270(5235):467-70.

50. Pollack JR, Perou CM, Alizadeh AA *et al*. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet 1999;23(1):41-6.

51. Egeland RD, Southern EM. Electrochemically directed synthesis of oligonucleotides for DNA microarray fabrication. Nucleic Acids Res 2005;33(14):e125.

52. Shi L, Reid LH, Jones WD *et al*. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 2006;24(9):1151-61.

53. Brazma A, Hingamp P, Quackenbush J *et al*. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 2001;29(4):365-71.

54. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30(1):207-10.

55. Slamon DJ, Clark GM, Wong SG *et al*. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science 1987;235(4785):177-82.

56. Campone M, Berton-Rigaud D, Bourbouloux E *et al*. [Her2 positive breast cancer: practices]. Bull Cancer 2011;98(2):154-63.

57. De Vita F., Giuliani F, Silvestris N *et al*. Human epidermal growth factor receptor 2 (HER2) in gastric cancer: a new therapeutic target. Cancer Treat Rev 2010;36 Suppl 3:S11-S15.

58. Kallioniemi A, Kallioniemi OP, Sudar D *et al*. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 1992;258(5083):818-21.

59. Snijders AM, Nowak N, Segraves R *et al*. Assembly of microarrays for genome-wide measurement of DNA copy number. Nat Genet 2001;29(3):263-4.

60. Pinkel D, Segraves R, Sudar D *et al*. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 1998;20(2):207-11.

61. Pinkel D, Albertson DG. Array comparative genomic hybridization and its applications in cancer. Nat Genet 2005;37 Suppl:S11-S17.

62. Ylstra B, van den Ijssel P, Carvalho B *et al*. BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). Nucleic Acids Res 2006;34(2):445-50.

63. Krijgsman O, Israeli D, Haan JC *et al*. CGH arrays compared for DNA isolated from formalin-fixed, paraffin-embedded material. Genes Chromosomes Cancer 2012;51(4):344-52.

64. Varambally S, Cao Q, Mani RS *et al*. Genomic loss of microRNA-101 leads to overexpression of histone methyltransferase EZH2 in cancer. Science 2008;322(5908):1695-9.

65. Jones S, Zhang X, Parsons DW *et al*. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science 2008;321(5897):1801-6.

66. Leary RJ, Lin JC, Cummins J *et al*. Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. Proc Natl Acad Sci U S A 2008;105(42):16224-9.

67. Parsons DW, Jones S, Zhang X *et al*. An integrated genomic analysis of human glioblastoma multiforme. Science 2008;321(5897):1807-12.

68. Weir BA, Woo MS, Getz G *et al*. Characterizing the cancer genome in lung adenocarcinoma. Nature 2007;450(7171):893-8.

69. Beroukhim R, Mermel CH, Porter D *et al*. The landscape of somatic copy-number alteration across human cancers. Nature 2010;463(7283):899-905.

70. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487(7407):330-7.

71. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008;455(7216):1061-8.

72. Hudson TJ, Anderson W, Artez A *et al*. International network of cancer genome projects. Nature 2010;464(7291):993-8.

73. Futreal PA, Coin L, Marshall M *et al*. A census of human cancer genes. Nat Rev Cancer 2004;4(3):177-83.

74. Forbes SA, Bhamra G, Bamford S *et al*. The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet 2008;Chapter 10:Unit.

75. Kent WJ, Sugnet CW, Furey TS *et al*. The human genome browser at UCSC. Genome Res 2002;12(6):996-1006.

76. Flicek P, Amode MR, Barrell D *et al*. Ensembl 2012. Nucleic Acids Res 2012;40(Database issue):D84-D90.

77. Gentleman RC, Carey VJ, Bates DM *et al*. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004;5(10):R80.

78. van de Wiel MA, Picard F, van Wieringen WN *et al*. Preprocessing and downstream analysis of microarray DNA copy number profiles. Brief Bioinform 2011;12(1):10-21.

79. van de Wiel MA, Brosens R, Eilers PH *et al*. Smoothing waves in array CGH tumor profiles. Bioinformatics 2009;25(9):1099-104.

80. van de Wiel MA, Kim KI, Vosse SJ *et al*. CGHcall: calling aberrations for array CGH tumor profiles. Bioinformatics 2007;23(7):892-4.

81. Olshen AB, Venkatraman ES, Lucito R *et al*. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 2004;5(4):557-72.

82. Willenbrock H, Fridlyand J. A comparison study: applying segmentation to array CGH data for downstream analyses. Bioinformatics 2005;21(22):4084-91.

83. van Wieringen WN, van de Wiel MA, Ylstra B. Normalized, Segmented or Called aCGH Data? Cancer Informatics 2007;3:331-7.

84. Jong K, Marchiori E, van der Vaart A *et al*. Cross-platform array comparative genomic hybridization meta-analysis separates hematopoietic and mesenchymal from epithelial tumors. Oncogene 2007;26(10):1499-506.

85. van Wieringen WN, Unger K, Leday GG *et al*. Matching of array CGH and gene expression microarray features for the purpose of integrative genomic analyses. BMC Bioinformatics 2012;13(1):80.

86. Haan JC, Buffart TE, Eijk PP *et al*. Small bowel adenocarcinoma copy number profiles are more closely related to colorectal than to gastric cancers. Ann Oncol 2012;23(2):367-74.

87. van de Wiel MA, van Wieringen WN. CGHregions: dimension reduction for array cgh data with minimal information loss. Cancer Informatics 2007;2:55-63.

88. Beroukhim R, Getz G, Nghiemphu L *et al*. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proc Natl Acad Sci U S A 2007;104(50):20007-12.

89. van Wieringen WN, van de Wiel MA, Ylstra B. Weighted clustering of called array CGH data. Biostatistics 2008;9(3):484-500.

90. Monti S, Tamayo P, Mesirov J *et al*. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning 2003;52(1-2):91-118.

91. van de Wiel MA, Smeets SJ, Brakenhoff RH *et al*. CGHMultiArray: exact P-values for multi-array comparative genomic hybridization data. Bioinformatics 2005;21(14):3193-4.

92. Tibshirani R, Hastie T, Narasimhan B *et al*. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A 2002;99(10):6567-72.

93. van Wieringen WN, Belien JA, Vosse SJ *et al*. ACE-it: a tool for genome-wide integration of gene dosage and RNA expression data. Bioinformatics 2006;22(15):1919-20.

94. van Wieringen WN, van de Wiel MA. Nonparametric testing for DNA copy number induced differential mRNA gene expression. Biometrics 2009;65(1):19-29.

95. Koopman M, Antonini NF, Douma J *et al*. Sequential versus combination chemotherapy with capecitabine, irinotecan, and oxaliplatin in advanced colorectal cancer (CAIRO): a phase III randomised controlled trial. Lancet 2007;370(9582):135-42.

96. Tol J, Koopman M, Cats A *et al*. Chemotherapy, bevacizumab, and cetuximab in metastatic colorectal cancer. N Engl J Med 2009;360(6):563-72.

97. Knijn N, Mekenkamp LJ, Klomp M *et al*. KRAS mutation analysis: a comparison between primary tumours and matched liver metastases in 305 colorectal cancer patients. Br J Cancer 2011;104(6):1020-6.