INAUGURAL - DISSERTATION
zur
Erlangung der Doktorwürde
der
Naturwissenschaftlich-Mathematischen Gesamtfakultät
der
Ruprecht - Karls - Universität
Heidelberg

vorgelegt von
Diplom-Biologe Nicolas Delhomme
aus Villeneuve sur Lot, Frankreich
Tag der mündlichen Prüfung:

Thema

Integrative and Comparative Analysis of Retinoblastoma and Osteosarcoma

Gutachter: P.D. Dr. Karsten Rippe
Prof. Dr. Peter Lichter

Die vorliegende Arbeit wurde am Deutschen Krebsforschungszentrum in der Abteilung von Prof. Dr. Lichter in der Zeit von 01.04.2004 bis 30.11.2008 unter der wissenchaftlichen Anleitung von Prof. Dr. Peter Lichter ausgeführt.

Eidesstattliche Versicherung

Ich erkläre hiermit, dass ich diese vorgelegte Dissertation selbst verfasst und mich dabei keiner eanderen als den von mir ausdrücklich bezeichneten Quellen und Hilfen bedient habe. Diese Dissertation wurde in dieser or andere Form weder bereits als Prüfubgsarbeit verwendet, noch einer anderen Fakultät als Dissertation vorgelegt. An keiner anderen Stelle ist ein Prüfungsverfahren beantragt.

Heidelberg, den 1. Januar 2013

Nicolas Delhomme

Acknowledgments

- For the data used in this work, published or not, I want to thank Prof. Dr. Dietmar Lohmann, Dr. Sandrine Gratias, Dr. Boris Zielinski, and Dr. Stephan Wolf.

- For offering me the chance to do my PhD and waiting all these years for it to complete, I'm definitely grateful to Prof. Dr. Peter Lichter.

- For the nice working atmosphere, the Chistmas parties, the pre-retreat and the retreat, the whole of the B060 division. Thanks for all the memories.

- For every day work - and more - Frédéric Blond, Dr. Grischa Tödt and of course Dr. Natalia Becker. Natalia, I'm glad you made it in that crowded geek's office! The last member of the SOG, which one can't forget - he will recognize himself - made these days in the office never boring for better or worse. Thanks for that Felix. A special thank to Margit MacLeod and Michael Hain for helping out with the printouts and corrections. Finally, a big thanks to Dr. Felix Kokocinski, who introduced me to the B060s and got me my first publication!

- For the submission process, the deanery and the Helmholtz International Graduate School for Cancer Research, in particular Heidi Costas and Dr. Lindsay Murrells for their help.

I would not have achieved this and be were I'm without an innumerable number of persons, among which I want to thank a few in particular.

- Prof. Ladel who transmitted me his passion for both biological and computational science; **you started it all!**

- Prof. Antoine de Daruvar who entrusted me his tool although I was just fresh from school. His continuous mentoring helped me through all these years. **Merci Antoine.**

- Bertrand Fabre for his accessibility when I was still a student and for all the advices he gave me.

- A number of other ex-LIONs, Christian Marcazzo, Jean-Stéphane Morin, Niels Bojunga, Bernhard Sultzer, who made me trust my work.

- Emilie Fritsch et Sophie Adjaley for their uninterrupted availability to help, even from miles (swedish ones) away.

- Michael Knop for the long collaboration on an exciting project and especially for accepting to chair my defense committee.

- Charles Girardot and Julien Gagneur for our long Heidelberg common history. We're now dispersed, but I really appreciated working with you. It kept me wanting to get better to reach out to your spheres. Let's the future tells us if I managed.
- Hermann Rupp who offered me shelter for a long time. I suppose I was your longest "guest".
- **My family for everything they did for me.**
- My other family for their interest, their constant support and their kindness.
- And Verena, my half. I cannot be thankful enough for all the efforts she conceded, the patience she showed and the support she gave me while I was working on this thesis. Literally, without her, this document would not be in your hands.

This thesis is dedicated to her and to long time friends that probably never believed this chapter of my life will be closing one day. **Verena, Marc, Xavier, Martin, this is for you.**

Zusammenfassung

In den letzten eineinhalb Jahrzehnten hat die breite Nutzung der Hochdurchsatzmethoden in der Molekularbiologie zu großen Mengen von Datensätzen geführt, die eine unerwartete Komplexität der Zellregulation zeigen. Die kürzlich veröffentlichten Ergebnisse des ENCODEProjekts (ENCODE Project Consortium et al., 2012) haben das Ausmaß dieser Mechanismen im menschlichen Genom gezeigt und sicherlich werden in der Zukunft noch weitere entdeckt. Diese Komplexität innerhalb einer einzelnen Zelle, ganz zu schweigen von Zell-Zell-Wechselwirkungen oder dem Einfluss der Mikroumgebung, ist schwer zu erfassen. Dieses Verständnis ist allerdings der Schlüssel zur Entwicklung von individuell angepassten Behandlungen genetischer Krankheiten oder Störungen, darunter auch Krebs.

In der Mathematik geht man solche komplexen Probleme mit Methoden an, die die Komplexität reduzieren, so dass man sie in auflösbarer Weise modellieren kann. In der Biologie haben Forscher diese Methode angewandt, um das Systembiologiekonzept zu entwickeln, welches das Verständnis der Zellregulationsmechanismen vereinfacht. Die meisten veröffentlichten Studien, in denen Hochdurchsatz-Technologien verwendet wurden, waren jedoch nur auf eine einzige Art der Zellregulation ausgerichtet und können daher nicht als solche verwendet werden, um die Regulationswechselwirkungen zu untersuchen. Außerdem ist bei solchen Studien die Unterscheidung zwischen auslösenden Faktoren und Störfaktoren schwierig.

Diese beiden Punkte waren meine ursprüngliche Motivation für die Entwicklung statistischer Methoden, die die integrative und vergleichende Analyse von verschiedenen Arten von Datensätzen ermöglichen. Es wurden drei verschiedene Softwares entwickelt, um dieses Ziel zu erreichen. Erstens, "customCDF": dies ist eine Methode, um die **Custom Definition File** (CDF) des Affymetrix *GeneChip*®s neu zu definieren; dies dient im Wesentlichen dazu, die ständige Aktualisierung der Sequenz des menschlichen Genoms und dessen Annotationen zu erfassen. Zweitens, "aSim": eine Methode, um Microarray-Daten zu simulieren; dies erstellt die notwendigen Daten um die entwickelten Algorithmen auszuwerten. Drittens, eine Reihe von kombinierten statistischen Methoden, die integrative Analysen ermöglichen und die schließlich durch gezielte Modifikationen auch vergleichende Analysen erlauben. "CustomCDF" und "aSim" wurden auf unabhängigen Datensätzen validiert, währen die entwickelten analytischen Methoden auf "aSim" simulierten Dateien und öffentlich verfügbaren Datensätzen validiert wurden.

Die oben beschriebenen Methoden wurden angewandt, um zwei biologische Fragen zu beantworten. Zunächst wurden zwei Retinoblastom-Datensätze benutzt, um die Auswirkung von Genom-Aberrationen auf die Gen-Expressionen zu untersuchen. Dann, motiviert durch die Tatsache, dass Retinoblastom-Patienten im späteren Leben ein höheres Risiko haben ein Osteosarkom zu entwickeln als der Durchschnitt der Bevölkerung, wurden Datensätze von beiden Tumoren vergleichend analysiert, um Ähnlichkeiten und Unterschiede zu identifizieren. Trotz der eher begrenzten Anzahl von Datensätzen waren beide Ansätze dank ihrer hohen Präzision und niedrigen Fehlerrate erfolgreich und haben so die Basis für größere Analysen gebildet.

In der Tat hat die hier angewandte integrative Analyse des Retinoblastoms gezeigt, dass dem Zugewinn des Chromosoms 6 eine wichtige Bedeutung in der Progression der Krankheit zukommt, was wiederum darauf hinweist, dass viele Gene auf diesem Chromosom eine Krebsentwicklung fördern. Im Vergleich zu Microarray-Standardanalysen war diese Analyse darüber hinaus in der Lage, die Interaktion von Regulierungsmechanismen zu entdecken: Beispiele von positivem und negativem Ausgleich der Genexpression in Regionen mit DNA-Zugewinn beziehungsweise -Verlust, sowie Beispiele von Antisense-Transkription, Pseudogen- und snRNA-Regulation wurden in diesem Datensatz identifiziert.

Durch die vergleichende Analyse hingegen konnte gezeigt werden, dass Retinoblastome und Osteosarkome große Ähnlichkeit aufweisen und darüber hinaus, dass beide Vorteil aus ihren jeweiligen Mikroumgebungen ziehen und damit verschiedene Signalwege, *PKC/Calmodulin* in Retinoblastom und *GPCR/RAS* in Osteosarkom, zu nutzen scheinen.

In dieser Arbeit konnte die Bedeutung und der Nutzen der entwickelten Softwares und statistischen Methoden demonstriert werden: durch sie konnten präzise Antworten auf die zwei gestellten biologischen Fragen gefunden werden. Desweiteren konnte dadurch eine Reihe interessanter Hypothesen aufgestellt werden, die weitere Untersuchungen erfordern. Diese Softwares sind nicht auf Microarray-Analysen begrenzt, sondern können auf alle Hochdurchsatz-Daten appliziert werden: mittels der hier entwickelten Methoden kann das Konzept der Systembiologie auf die Erforschung der Karzinogenese angewandt werden.

Abstract

In the last one and a half decades, the generalization of high throughput methods in molecular biology has led to the generation of vast amounts of datasets that unraveled the unfathomed complexity of the cell regulatory mechanisms. The recently published results of the ENCODE project (ENCODE Project Consortium et al., 2012) demonstrated the extend of these in the human genome and certainly more regulation mechanisms will be discovered in the future. Already, this complexity within a single cell - without taking into account cell-cell interaction or micro-environment influences - cannot be abstracted by the human mind. However, understanding it is the key to devise adapted treatments to genetic diseases or disorders, among which is cancer.

In mathematics, such complex problems are addressed using methods that reduce their complexity, so that they can be modeled in a solvable manner. In biology, it led researchers to develop the concept of systems biology as a mean to abstract the complexity of the cell regulatory network. To date, most of the published studies using high throughput technologies only focus on one kind of regulatory mechanism and hence cannot be used as such to investigate the interactions between these. Moreover, distinguishing causative from confounding factors within such studies is difficult.

These were my original motivations to develop analytical and statistical methods that control for confounding factors effects and allow the integrative and comparative analysis of different kinds of datasets. *In fine*, three different tools were developed to achieve this goal. First, "customCDF": a tool to redefine the **Custom Definition File** (CDF) of Affymetrix *GeneChip*®s. It results in the increased sensitivity of downstream analyses as these benefit from the constantly evolving human genome reference and annotations. Second, "aSim": a tool to simulate microarray data, which was required to benchmark the developed algorithms. Third, for the integrative analysis, a set of combined statistical methods and finally for the comparative analysis, a modification of the integrative analysis approach. These were bundled in the "crossChip" R package.

The "customCDF" and "aSim" tools were first validated on independant datasets. The developed analytical methods ("crossChip") were first validated on "aSim" simulated data and publicly available datasets and then used to answer two biological questions. First, using two retinoblastoma datasets, the effect of genomic copy number variations on gene-expression was investigated. Then, motivated by the fact that retinoblastoma patients have a higher chance to develop osteosarcoma later in life than the average population, datasets of both these tumors were comparatively analyzed to assess these tumors similarities and differences.

Despite a rather limited number of samples within the selected datasets, the developed approaches with their higher sensitivity and sensibility were successful and set the ground for larger scale analyses. Indeed, the integrative analysis applied to retinoblastoma revealed the high importance of the chromosome 6 gain at a later stage of the disease, indicating that many genes on that chromosome are beneficial to cancerogenesis. Moreover, in comparison to standard **microarray** analyses, it demonstrated its efficacy at detecting the interplay of regulatory mechanisms: examples of positive and negative compensation of gene expression in lost and gained regions, respectively, as well as examples of antisense transcription, pseudogene and snRNAs regulation were identified in this dataset. The comparative analysis on the other hand revealed the high similarity of the retinoblastoma and osteosarcoma tumors, while at the same time showing that either of them take advantage of their distinct micro-environment and consequently appear to make use of different signaling pathways, *PKC/calmodulin* in retinoblastoma and *GPCR/RAS* in osteosarcoma.

The developed tools and statistical methods have demonstrated their validity and utility by giving sensible answers to the two biological questions addressed. Moreover, they generated a large number of interesting hypotheses that need further investigations. And as they are not limited to microarray analysis but can be applied to analyze any high-throughput generated data, they demonstrated the usefulness of "systems biology" approaches to study **cancerogenesis**.

Contents

List of Figures [1]

[1] Most figures in the **Introduction** are extracted from **The Biology of Cancer** (Weinberg, 2007) and are Copyright 2007 (c) Garland Science.

List of Tables [2]

[2] Most tables in the **Introduction** are extracted from **The Biology of Cancer** (Weinberg, 2007) and are Copyright 2007 (c) Garland Science.

Chapter 1

Introduction

The subject of this work is to assess the potential of integrative and comparative analyses to study genetics in cancer. In the first part of this introduction, our current knowledge about cancer will be detailed as well as possible causes for **tumorigenesis**. I will then further focus on the two types of tumors studied in this work: ***Retinoblastoma*** and ***Osteosarcoma***. Finally, I will detail the analyses performed and especially concentrate on their combination as an example to describe the benefit of such integrative and comparative analyses to extract the relevant, likely causative, events that lead to tumorigenesis.

1.1 Cancer

This work concentrates on two specific types of, mostly pediatric, tumors: **Retinoblastoma** and **Osteosarcoma**. A detailed description of these will be provided in the following sections, preceded by an historical description of cancer research to introduce key findings that delineate our current understanding of the biology of cancer.

1.1.1 A history of research

First description: The description of cancer and cancer cells first occured in the second half of the nineteenth century and a century later not much progress had been made as to explaining why cancer arises. By that time it was known that cancer cells are normal cells, which proliferation went uncontrolled. In addition it appeared that:

- the majority of the tumor mass were clonal cells issued from a unique founder cell that had gone awry
- any cell type of an organism could undergo that process

- tumor could metastasize, the main reason for almost every cancer death

Despite that research century being unable to decipher any of the cancer mechanisms, it provided the research community with a detailed classification of tumors. These are classified into 4 major groups: epithelial, mesenchymal, hematopoietic and neuroectodermal, nowadays split in numerous sub-classifications. This classification is constantly refined by new findings and is maintained by the "World Health Organization" (WHO) in their **International Classification of Diseases** (ICD-10) (World Health Organization, 2010). The major discoveries in molecular biology of the following decades, *e.g.* the **deoxyribonucleotide acid** (DNA) structure, although helpful to explain many biological processes, were still unable to shed light on the mechanism of cancer.

A viral breakthrough: The breakthrough came from an entirely different field of biology, that of **virology**. In the first decade of the twentieth century, a major discovery was made by **Peyton Rous**: in his experiments he was able to transmit the sarcoma tumor from a hen to another just by injecting the "receiver" with the filtrate of the donors' ground tumor (Rous, 1911). The identified causative virus is known as the **Rous Sarcoma Virus** (RSV). Many more viruses were identified that could transform normal cultured cells into tumorigenic ones, indicating the presence of powerful **oncogenes** in these viruses' genome. The final breakthrough occurred when researchers realized that the oncogenes harbored by these viruses were just modified copies of genes present in every vertebrate genome. This was first established when the RSV oncogene *v-src* was identified as the copy of the *c-src* gene, present in a normal cellular genome and lead to the discovery of many more transforming viruses (Stehelin et al., 1976). An additional discovery shed further light on cancer pathogenesis: non-transforming viruses were able to induce cultured cell malignant transformation as well. This was traced back to the integration of viral DNA in front of or within genes of the host cell, thus deregulating or changing the function of so-called **proto-oncogenes**. Viruses are nowadays known to be the causative factor of a fitfh of the human cancers worldwide. Studying these viruses lead to the first anti-cancer vaccine being developed against the causative agent of the cervical cancer: the **Human papillomavirus** (HPV). Harald zur Hausen, professor at the Heidelberg University and former director of the **German Cancer Research Center** (DKFZ), was awarded the Nobel Price of Medicine for his work on HPV in 2008.

Oncogenes' discovery: By the end of the 1970's, it was hypothesized that the pathogenesis of the remaining 80% of cancer could be explained by somatic mutations affecting proto-oncogenes, resulting in an uncontrolled

cell growth. Indeed it had been shown that structural or regulatory changes affecting a gene could lead to cancerogenesis. For example, the chromosomal translocation `t(9;22)(q34;q11)` was shown to result in the hybrid gene *bcr/abl*, a major causative factor of the **Chronic Myelogenous Leukemia** (CML) and **Acute Lymphoblastic Leukemia** (ALL). Despite the newly acquired understanding of the origin of oncogenes and their role in cancer pathogenesis, much remained to be explained. Among these open questions, one received a lot of attention in the early 1980s: how could cancer cells escape the tight proliferation control placed on their normal counterparts?

The implication of cell signaling: A part of the answer to that question came from the study of oncogenes such as *erbB*, which is closely related to the **epidermal growth factor** (EGF) **receptor**. It established the link between **growth factor** signaling and cell transformation and through the years, most signaling pathways have been associated with cancer, see Table 1.1 extracted from Baudot et al. (2010) that lists some examples. The first deciphered pathways act through the phosphorylation of target protein(s) on specific Tyrosine residue(s), an event that leads to the activation of these proteins and through it to the propagation of the signal. The phosphorylating proteins are often the signal initiator as they are either part of, or coupled to, extra-cellular receptors, and are grouped under the term **Receptor Tyrosine Kinase** (RTK). They recruit key signal transduction components: the small GTPase *Ras* and the lipid kinase **phosphatidylinositol-3OH kinase** (PI(3)K) that influence many cellular processes such as growth, differentiation, migration, apoptosis, *etc.*

RTKs are no the only signal transduction mechanism, many more have been described involving for example **Serine/Threonine Kinases** - *e.g.* the **Transforming Growth Factor Beta** (TGF-β) receptor - or **integrins**, receptors that sense the association between the cell and the **extracellular matrix (ECM)**. Overall, these discoveries can be summarized as the establishment of the first **hallmark of cancer**: **the self-sufficiency in growth signals** (McCormick, 1999).

Tumor suppressors: The discovery of the oncogenes and of the implication of the cell signaling circuitry still was not satisfying to explain most cancer arousal. Indeed it required the responsible agent to be a dominant trait, which is the case for viral infections but not for the remainder of cancer where a viral origin could not be established. Hence, in the 1970s and 1980s researchers hypothesized the existence of **tumor suppressor genes**: recessive cancer inducing genes. A very essential finding in the genetics of tumor suppressor genes came from the study of a rare childhood tumor: retinoblastoma. Retinoblastoma, as will be described in more detail in the related section 1.1.2, page 16, can be declined in

	Tumour type	Mutated genes
Cellular pathways		
Adipocytokine signalling pathway (hsa04920)	Colorectal	*ACSL4 ACSL5 AKT1 CAMKK2 CHUK MTOR IRS2 IRS4 JAK1 MAPK9 PRKCQ PTPN11 STK11 TYK2*
	Lung	*AKT1 CAMKK2 CHUK MTOR IKBKB IRS1 IRS4 JAK1 JAK2 JAK3 PPARGC1A PRKCQ PTPN11 RELA STAT3 STK11 TYK2*
CBL-mediated ligand-induced downregulation of EGF receptors (h_cblPathway)	Leukaemia	*CBL CSF1R EGFR MET PDGFRA*
Epithelial cell signalling in *Helicobacter pylori* infection (hsa05120)	Lung	*CHUK CSK CXCR2 EGFR IKBKB MAP2K4 MAP3K14 MAPK11 MAPK14 MAPK8 MET PLCG1 PLCG2 PTPN11 PTPRZ1 RELA SRC*
NF-κB activation by non-typeable *Hemophilus influenzae* (h_nthiPathway)	Lung	*CHUK CREBBP EP300 IKBKB MAP2K6 MAP3K14 MAPK11 MAPK14 RELA SMAD4 TGFBR1 TGFBR2*
GnRH signalling pathway (hsa04912)	Colorectal	*ADCY8 ADCY9 CAMK2A CAMK2B EGFR GNAS HRAS KRAS MAP2K4 MAP2K7 MAP3K2 MAP3K3 MAPK9 MMP2 NRAS PLA2G2A PRKCA PRKCB SRC*
	Lung	*CAMK2G EGFR GNAS HRAS ITPR2 KRAS MAP2K4 MAP2K6 MAP3K2 MAP3K3 MAP3K4 MAPK11 MAPK14 MAPK7 MAPK8 MMP2 NRAS PLCB1 PRKACA PRKACB PRKCB PRKCD PRKX RAF1 SRC*
Regulation of transcriptional activity by PML (h_pmlPathway)	Leukaemia Melanoma	*CREBBP HRAS KRAS NRAS PML RARA RB1 TP53* *CREBBP DAXX HRAS KRAS NRAS RB1 TNF TNFRSF1A TP53*
Trefoil factors initiate mucosal healing (h_tffPathway)	Lung	*AKT1 CSH2 CTNNB1 CYCS EGFR ERBB2 GHR HRAS KRAS MUC2 NRAS PTK2 SHC1 SOS1*
Bioactive peptide induced signalling pathway (h_biopeptidesPathway)	Lung	*CAMK2G CDK5 FYN HRAS JAK2 KRAS MAPK14 MAPK8 NRAS PLCG1 PRKCB RAF1 SHC1 SOS1 STAT5A*
Processes		
Huntington disease (hsa05040)	Leukaemia	*CLTC CLTCL1 CREBBP EP300 HIP1 TP53*
Base-excision repair (GO:0006284)	Lung	*CKN2 MBD4 MSH6 NTHL1 PCNA POLG RAD51L3 TP53*
Double-stranded break repair through non-homologous end joining (GO:0006303, GO:0000726)	Melanoma	*POLA1 PRKDC XRCC4 XRCC6*
Cytokinesis (GO:0000910)	Lung	*ARHGEF11 AURKB AURKC BRCA2 DIAPH2 ESPL1 INCENP MYH9 ROCK1*
Protein import into nucleus, translocation (GO:0000060)	Leukaemia	*ARNT BCL3 BCL6 JAK2*
Protein import into nucleus, translocation (GO:0000060)	Leukaemia	*ARNT BCL3 BCL6 JAK2*
Regulation of protein export from nucleus (GO:0046825)	Lung	*BARD1 GSK3B PTPN11*
Protein domains		
BRK (IPR006576)	Lung	*CHD5 CHD6 CHD9 SMARCA4*
Bromodomain (IPR001487)	Brain	*BRD2 EP300 KAT2B MLL PHIP SMARCA2 SMARCA4 SP100 TAF1 TAF1L TRIM24 TRIM33 ZMYND8*
	Lung	*BAZ1A CREBBP EP300 MLL SMARCA4 TAF1 TAF1L TRIM24 TRIM33*
Helicases (IPR014001, IPR001650, IPR014021)	Breast	*ATRX BRIP1 CHD1 CHD5 CHD7 CHD8 CHD9 DDX10 DDX18 DDX3X DDX59 EIF4A2 ERCC3 ERCC6 FANCM HELQ RAD54L SKIV2L SMARCAD1 SMARCAL1*
	Melanoma	*ATRX CHD6 CHD8 DDX24 DDX28 DDX3X DDX4 DDX54 DICER1 POLQ RAD54L SMARCA1 SMARCA4 SMARCA5 SMARCAD1 SMARCAL1 SNRNP200*
Diacylglycerol kinase (IPR000756, IPR001206)	Melanoma	*CERKL DGKB DGKD DGKG DGKZ*
Laminin G (IPR001791, IPR012680)	Pancreas	*CELSR1 CNTNAP4 COL11A1 COL5A1 FAT1 FAT3 FAT4 LAMA1 LAMA4 LAMA5 NRXN3*

CBL, Casitas B-lineage lymphoma; EGF, epidermal growth factor; GnRH, gonadotropin-releasing hormone; GO, gene ontology; PML, promyelocyclic leukaemia.

Table 1.1 – Example of cellular pathways associated to specific cancer

two forms: **unilateral** and **bilateral**. The bilateral cases are most often a **familial** form, whereas the unilateral ones are mainly sporadic. The **pedigree** of families with bilateral cases shows a **mendelian** inheritance of a recessive allele, in accordance with the presence of a tumor suppressor gene. Knudson (1971) devised from that a - since then well-established - theory: the **two-hit hypothesis**. Based on 48 cases of retinoblastoma of both kinds, he could establish a statistical model explaining that two hits (mutations) are necessary for retinoblastoma to occur, see Figure 1.1. As familial cases already carried a mutation, their likelihood of developing retinoblastoma was increased as well as was its bilateral occurence. This breakthrough consolidated the tumor suppressor gene hypothesis: tumor suppressor genes are recessive genes essential to prevent cancer; *i.e.* both copies of these genes need to be altered for an effect on the cell phenotype to be observed. Since tumor suppressor genes are recessive, it is extremely difficult to prove that a gene is a tumor suppressor gene and not just a silent gene or an unrelated bystander of the cancerogenesis. In other words it is difficult to establish a supposedly tumor suppressor gene as a "cause" and not a "consequence". For that reason the definition of a tumor suppressor gene is complex. Weinberg (2007) describes it as: *a gene whose partial or complete inactivation, occurring in either the germ line or the genome of a somatic cell, leads to an increased likelihood of cancer development* (Weinberg, 2007).

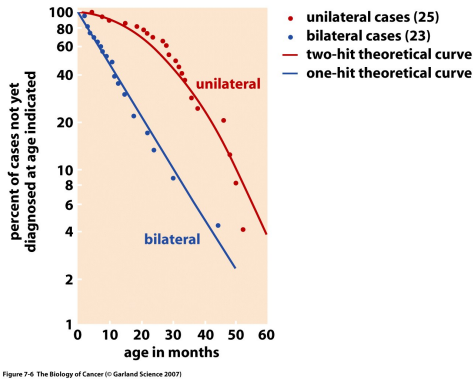


Figure 1.1 – Knudson two-hit hypothesis model - adapted from **The Biology of Cancer** (Weinberg, 2007), Copyright 2007 (c) Garland Science.

pRb and *TP53*: Two major tumor suppressor genes were quickly identified: ***retinoblastoma*** gene (*RB1*) and ***tumor protein p53*** (*TP53*). *RB1* is a key element of the cell cycle: a strict series of events that allows a cell to duplicate its DNA content and divide into two daughter cells. Its deregulation is an essential step to cancerogenesis and especially the proteins involved in the regulation of the cell cycle G1 checkpoint, the **R point**, are affected. The **R point** decides of the cell fate between growth and quiescence. When a normal cell is subjected to enough **cumulative mitogenic** signal it commits to undergo mitosis. One of the *master regulators* of this decision point is *RB1*. After its isolation in 1986, it was shown that its protein phosphorylation state changes around the R point. During most of the G1 phase it is not phosphorylated. Ahead of the R point, it becomes

hypophosphorylated and is progressively getting hyperphosphorylated until the end of the mitosis (Figure 1.2). *RB1* is the R point *gate keeper*; it is

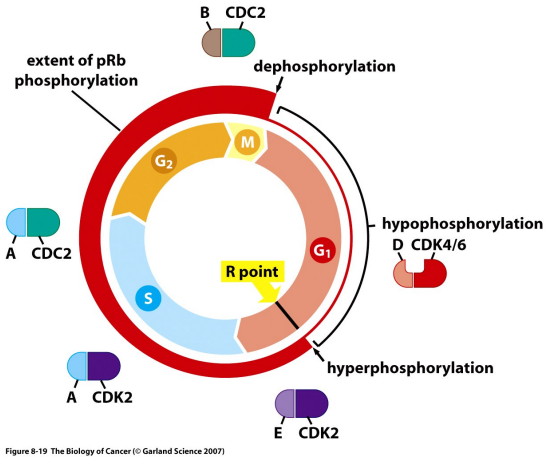


Figure 1.2 – pRb phosphorylation cycle - adapted from **The Biology of Cancer** (Weinberg, 2007), Copyright 2007 (c) Garland Science.

the main receptor of the mitogenic signals. Once this major checkpoint is overcome, the cellular machinery is anyway committed to going through the cell cycle.

But the de-regulation of the cell cycle alone was not sufficient to reproduce cancerogenesis in mouse models. The observed tumor presented a very high level of **apoptosis**. This discovery led to the understanding of the apoptosis role in cancer and of one of its major players: *TP53*. *TP53* was first identified through its interaction with viral oncoproteins, however its classification as an oncogene or tumor suppressor gene was for a long time a matter of discussion. Hence, it does not follow the Knudson model of tumor suppressor genes: *TP53* -/- mouse models are viable and do not, unlike other tumor suppressor gene mouse models, die during embryogenesis. The reason is that *TP53* acts as a **homotetramer** and that the presence of a single mutant allele in a cell is enough to deregulate *TP53* functionalities. More than 50% of all cancers harbor a mutant *TP53* allele; it is often a **dominant negative** trait. With an equal proportion of both alleles, only 1/16 of the complex will

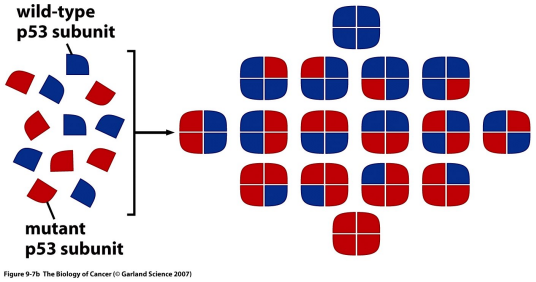


Figure 1.3 – Combinations of wild-type (blue) and mutant (red) *TP53* proteins - adapted from **The Biology of Cancer** (Weinberg, 2007), Copyright 2007 (c) Garland Science.

have their normal ability as shown in Figure 1.3. The remaining wild-type complexes are able to maintain a certain level of the *TP53* normal abilities as most cancers show a concurrent **loss of heterozygosity** (LOH) at the *TP53* locus, either through deletion or duplication of the mutant allele. It is established that in most cases the *TP53* mutation precedes the LOH event. During the early 1990s, most of the activating signals and the *TP53*'s downstream effects were discovered, see Figure 1.4. *TP53* is the **master guardian** of the cell as well as its **executioner**. It monitors the cell and its environment: DNA structure, lack of nucleotides, presence of oxygen, cell cycle status, *etc.* and reacts accordingly to any perturbation by triggering *e.g.* DNA repair, cell cycle arrest, *etc.* to restore an healthy cell state. But in cases where the damages are too important, it resorts to apoptosis. It is surprising that most of the control over the cell fate has been empowered to a single protein, but explains why it is so often mutated in cancer. Moreover, it has been shown that when *TP53* is not directly mutated, its regulators are, *e.g.*

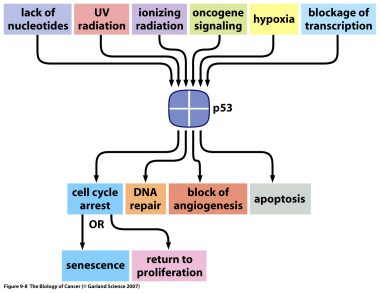


Figure 1.4 – *TP53* activating signals and effects - adapted from **The Biology of Cancer** (Weinberg, 2007), Copyright 2007 (c) Garland Science.

- **mouse double minute 2** (Mdm2) ubiquitinates *TP53* and results in its rapid degradation. An over-expression of Mdm2 results in the constant degradation of *TP53*.
- **Alternative Reading Frame** ($p14^{\mathrm{ARF}}$) targets Mdm2 and results in its translocation to the **nucleolus**, hence avoiding *TP53* degradation. Therefore, a down-regulation of $p14^{\mathrm{ARF}}$ results in a constant degradation of *TP53*

Moreover, the $p14^{ARF}$ locus is overlapping with the $p16^{INK4A}$ locus and in the close vicinity of the $p15^{INK4B}$ one; they are all located in a 40 kb locus. These last two genes encode proteins that are inhibiting the activity of a specific **Cyclin Dependant Kinases** (CDKs): the CyclinD-CDK4/6 complex. CDKs are other cell cycle master regulators; in continuously dividing cells, these proteins are associated with proteins the levels of which have cyclic fluctuations: the **Cyclins**. There are five major Cyclin-CDK complexes: Cyclin D-CDK4/6, Cyclin E-CDK2, Cyclin A-CDK2, Cyclin A-CDC2 and Cyclin B-CDC2, which are only active in a given time-frame during the cell cycle: G1, G1-S transition, S, S-G2 transition and M, respectively. Consequently, a homozygous deletion of the $p14^{ARF}$, $p16^{INK4A}$,

$p15^{INK4B}$ locus affects both the retinoblastoma and *TP53* pathways and leaves the cell control-free. The discovery of the pRb and *TP53* functions and their importance in cancerogenesis established another hallmark of cancer: **the insensitivity to anti-growth signals** (McCormick, 1999).

The other hallmarks of cancer: As mentioned, by the end of the 1980s, it became clear that a hallmark of cancer is the ability of its cells to generate their own mitogenic signals endogenously, *de-facto* bypassing all the constraint established on normal cells. To achieve this, different signaling pathways can be targeted, e.g. Ras, PI(3)K, **Nuclear Factor - Kappa B** (NF-κB), Jak/Stat, Notch, Patched, TGF-β, Wnt/β-catenin. Moreover, the more pathways were analyzed, the more cross-connections were identified, see Figure 1.5 for an example of the Ras effector pathway. Despite this complex intricateness, a still unexplained confounding fact is that a tumor of the same tissue of origin from different patients will often have the same mitogenic pathway affected, e.g. 90% of pancreatic carcinoma patients carry a mutant K-*ras* oncogene. This realization and the ever increasing complexity of cancer biology in the following 20 years led Hanahan and Weinberg (2000) to define **hallmark of cancers**, in an effort to summarize cancer as a whole. These are the traits that describe the capabilities of cancer cells:

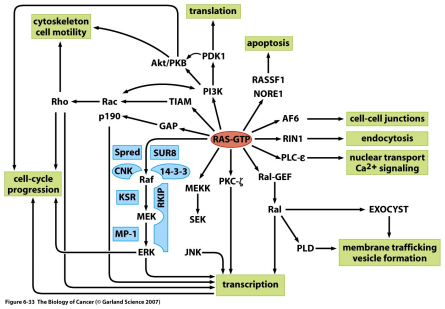


Figure 1.5 – The Ras effector pathway - adapted from **The Biology of Cancer** (Weinberg, 2007), Copyright 2007 (c) Garland Science.

- Self-sufficiency in growth signals
  As introduced previously (paragraph 1.1.1, page 3), cancer cells are able to produce their own mitogenic signals or to stimulate neighboring cells to do so.

- Insensitivity to anti-growth signals
  As introduced, see paragraph 1.1.1, page 5, cancer cells to proliferate efficiently have to weaken the strict controls maintaining tissue **homeostasis**; the main targets being *RB1* and *TP53* for their cell monitoring function.

- Evading apoptosis
  Another *TP53* functionality that cancer cells must disrupt, is its ability to induce apoptosis. *TP53* can up-regulate the expression of the *BCL2-associated X* (*Bax*) gene in response to sensing DNA damage,

which forces mitochondria to release **cytochrome C**, a potent catalyst of apoptosis. Cancer cells have devised other means to avoid apoptosis, *e.g.* by over-expressing pro-survival genes such as those of the PI(3)K-**protein kinase B** (AKT/PKB) pathway or by disabling inhibitors of such pathways as *e.g.* **phosphatase and tensin homolog deleted on chromosome 10** (PTEN).

- Limitless replicative potential
  These three first hallmarks of cancer would appear to be sufficient for a cancer cell to thrive endlessly, but it is not the case. Cells have yet another means of controlling proliferation by a mechanism uncoupled from the cell cycle. After a given number of replications, a cell will stop growing; it has a finite number of doublings independent of the *RB1* and *TP53* induced **senescence**. Double null mutants would eventually reach a **crisis** state characterized by massive cell death, karyotypic disarray showing chromosome end-to-end fusions. In the 1990s, it was realized that **telomeres** are shortening with every replication. Indeed, during the cell cycle the DNA polymerase is unable to completely replicate the 3' ends of chromosomes, resulting in a 50-100bp loss per duplication. Eventually no telomere remains and chromosome ends fuse, which results in the observed karyotypic disarray and cell death. In culture, 1 in 10^7 cells survives that crisis and become truly **immortalized** (Hanahan, 2000). Further research showed that this can be achieved in two ways, either by over-expressing the enzyme responsible for the telomere maintenance: the **telomerase** that adds hexanucleotides (TTAGGG) to the end of the telomere or by an alternative process named **ALT**, where the telomeres are maintained by means of homologous recombination and copy switching (Dunham et al., 2000). In this mechanism, proteins involved in recombination such as RAD51 and RAD52 create a replication loop among or within telomeres taking advantage of the tandem-repeat structure of the telomeres. It results in a dynamic maintenance of the telomeres length, where sequence material can be added, removed or exchanged between telomeres.

- Sustained angiogenesis
  The hallmarks of cancer described so far only focus on the cancer cells. This research approach: **reductionism** got more and more challenged during the 1990s as it could not explain another hallmark of cancer: the sustained angiogenesis of more advanced primary cancer. Additional observations, *e.g.* the fact that most Hodgkins lymphoma' cells (>99%) are not neoplastic, indicated that a tumor is a complex tissue, where **heterotypic** signaling (*i.e.* cell signaling across different types of cells) is essential, see Figure 1.6.

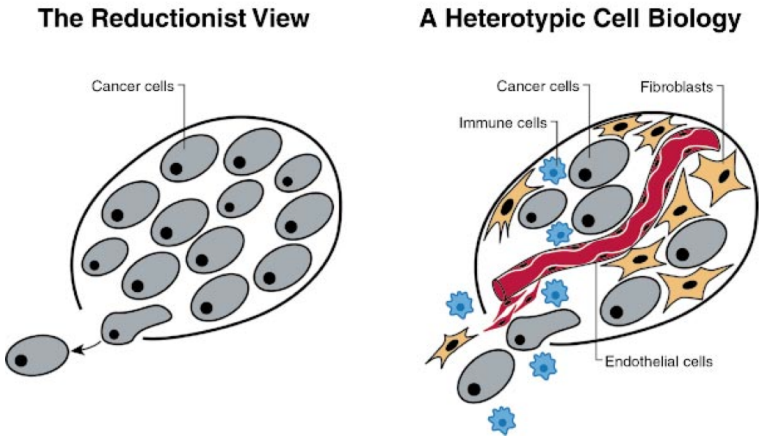


Figure 1.6 – Tumor as complex tissues - from Hanahan and Weinberg (2000)

During cancerogenesis, tumors that have acquired the hallmarks of cancer previously described are seldom able to reach a life threatening size. The rate at which they proliferate is balanced by the rate at which they undergo apoptosis or **necrosis**. As cancer cells replicate, several physical constraints arises:

- the local pressure increases as the available space for expansion is limited
- **hypoxia** appears in the inner tumor part as it is not vascularized; oxygen can only diffuse as far as 0.2 mm within a normal tissue

However, more advanced tumors are able to reshape their tissue boundaries and present a developed vasculature. They often express **platelet-derived growth factor** (PDGF), a growth factor that attracts **stromal**

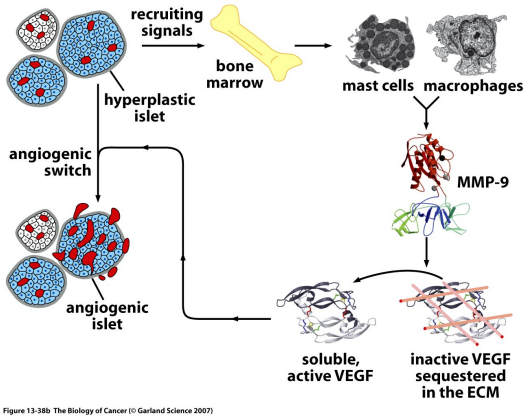


Figure 1.7 – The cell types recruited by tumors and their heterotypic interactions - adapted from **The Biology of Cancer** (Weinberg, 2007), Copyright 2007 (c) Garland Science.

cells. Among these, fibroblasts stimulated by the tumor **micro-environment** secrete **stroma-derived factor-1** (SDF1) to attract endothelial precursor cells, the capillaries building blocks. This eventually results in

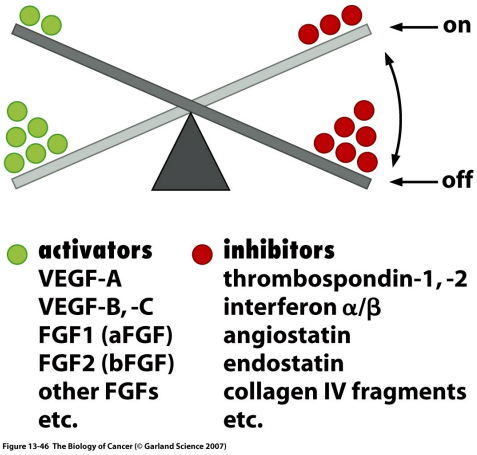


Figure 1.8 – Balancing the angiogenic "switch" - adapted from **The Biology of Cancer** (Weinberg, 2007), Copyright 2007 (c) Garland Science.

the neo-vasculature of the tumor. Fibroblasts are not the only cell types present in the stroma that are induced to support the tumor; among others macrophage and mast cells are attracted and hijacked to release pro-angiogenic factors, see Figure 1.7. This ability to provoke **neo-angiogenesis** that changes a tumor from being contained to a thriving tumor mass was originally refered to as the **angiogenic switch** (Coussens et al., 1999). However, further research showed how the field of angiogenesis, and as a whole the micro-environment of a tumor is a complex one. As of today, 10s of genes have been assigned to be pro or anti-angiogenic, see Figure 1.8. It appears that cancer cells are able to switch it on through mutations, but only with the support of the hijacked stroma.

- Tissue invasion & metastasis

  This hallmark of cancer, as described by Hanahan and Weinberg (2000), is the one that leads to 90% of all deaths by cancer. This process, although known and studied for more than a century, is still only vaguely defined at the molecular level. Both the trait of invasion and the ability to form a **metastasis** have been tied together and unlike earlier theories, are not acquired late in cancerogenesis. This is supported by the **Cancer of Unknown Primary** (CUP), where the patient suffers of metastases, while the primary tumor cannot be found (about 70% will be after autopsies, but 30% remains elusive). Despite these uncertainties, the overall process is well understood and described as the **invasion-metastasis cascade**, see Figure

1.9. During the 1990s and since then, the observation that tumors are

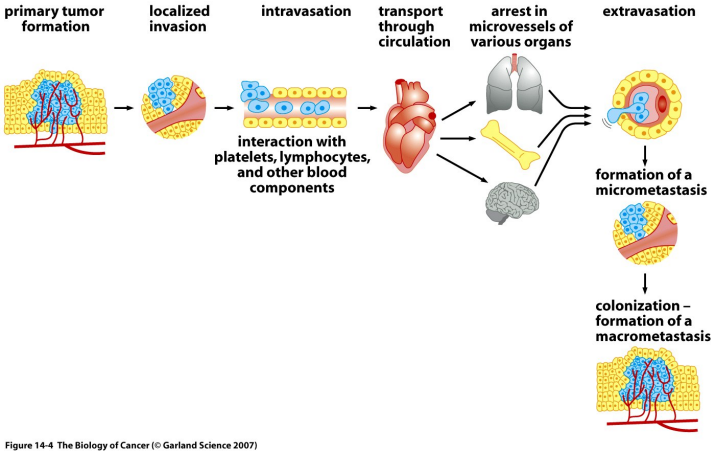


Figure 1.9 – The invasion-metastasis cascade - adapted from **The Biology of Cancer** (Weinberg, 2007), Copyright 2007 (c) Garland Science.

wounds that never heal, shed new light on the mechanism by which a cancer cell could acquire the **intravasation** and **extravasation** abilities. At the periphery of a wound, the epithelial cells undergo a transformation into mesenchymal like cells: the **Epithelial Mesenchymal Transition** (EMT). This gives them a motility ability, which they will use to resorb the wound. Once this is achieved, they undergo the mirror transition: the **Mesenchymal Epithelial Transition** (MET) to reform the epithelial layer. It has been shown that in epithelial cells, which are normally immotile and bound together by **tight junctions**, the motility/immotility balance is regulated through two proteins: the **N-cadherin** and the **E-cadherin**, respectively. This model has got-

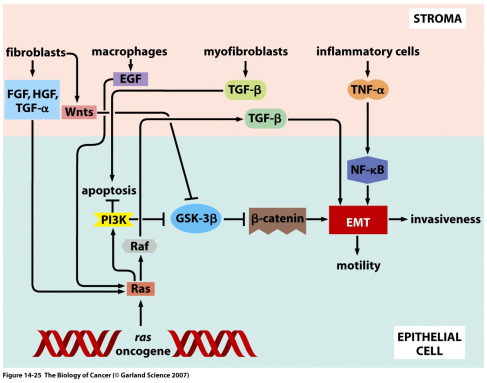


Figure 1.10 – The importance of the stroma - adapted from **The Biology of Cancer** (Weinberg, 2007), Copyright 2007 (c) Garland Science.

ten more and more support over the last decades. It is now known that this process is dependent on the micro-environment, *e.g.* the ex-

perimental removal of TGF-β and **Tumor Necrosis Factor alpha** (TNF-α) expressed by the stroma cells leads to noninvasive tumors, see Figure 1.10. The existence of this latent program in epithelial cells

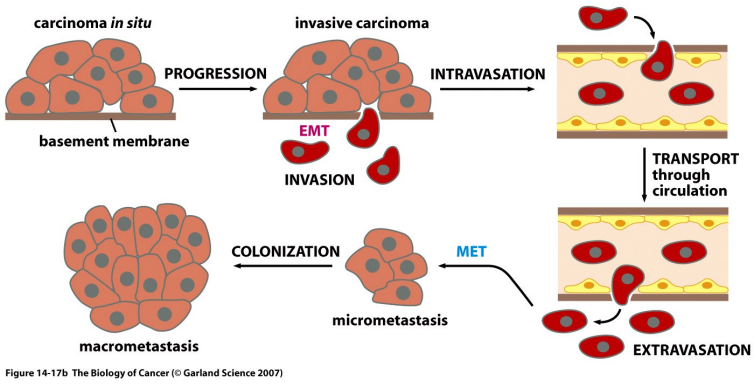


Figure 1.11 – The role of EMT and MET in metastases establishment: it would appear that the EMT is necessary for intravasation. During their transport, cancer cells remain in a mesenchymal state, which is reverted by the MET during the process of extravasation and the establishment of a metastasis - adapted from **The Biology of Cancer** (Weinberg, 2007), Copyright 2007 (c) Garland Science.

led to the proposal that carcinoma cells use it for metastasizing, see Figure 1.11. This heterotypic signaling has to be complemented by gene alterations within the tumor. Many affected genes have been reported among which, especially relevant for osteosarcoma, are the **Matrix Metallo-proteinases** (MMP)**s** and **urokinase Plasminogen Activator** (uPA), all involved in the ECM degradation.

An additional complexity of that hallmark of cancer is the non random, cancer specific sites, at which cancer cells will create metastases, see Figure 1.12. This was observed as early as 1889 by the british pathologist Stephen Paget, who suggested a *seed and soil* hypothesis, *i.e.* that metastases will only thrive in environment similar to that of the primary tumor. It has since then been amended, as it cannot explain for example, the rarity of **contra-lateral** metastases, *e.g.* breast or kidney cancer will rarely metastasize to their contra-lateral organ. The current understanding is that large amounts of cancer cells will evade from the primary tumor and their dispersion depends on the vasculature they reach (this explains $> 65\%$ of all breast metastases). Eventually, some circulating cancer cell will get trapped in a capillary and extravasate, creating a **micrometastasis**. If the microenvironment is benefitial, this micrometastasis can develop and turn into a **macrometastasis**.

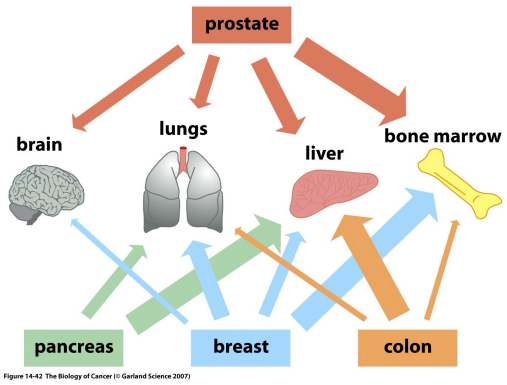


Figure 1.12 – The non random distribution of metastases - adapted from **The Biology of Cancer** (Weinberg, 2007), Copyright 2007 (c) Garland Science.

Otherwise, if the micro-environment is inappropriate, it might stay as a micrometastasis unless it acquires abilities to adapt to its new environment.

An increasingly more complex puzzle: Recent discoveries even extended the number of hallmarks of cancer. First, cancer cells have **the ability to evade the immune response**. Second, what was thought to be a consequence of cancer progression, *i.e.* the cancer cell **genomic instability**, seems to be a more pro-active factor of the tumorigenesis. Indeed, acquiring genomic instability appears to be necessary for cancer cells to thrive. Recently, Stephens et al. (2011) and Rausch et al. (2012) showed a new mechanism by which cancer cells can gain this trait: **chromotripsis**; a one step event resulting in catastrophic DNA rearrangements. Moreover, it is now evident that human cells have evolved many different mechanisms of regulations, some of which have been only recently discovered, for example:

- Johnson et al. (2005) showed that *Ras* is regulated by the **microRNA let-7**. *let-7* targets the untranslated region of H-ras, N-ras and K-ras and its deletion leads to an elevation of the Ras activity.
- Kowalczyk et al. (2012) discovered a new kind of **messenger ribonucleotide acid (RNA)** (mRNA), resulting from the transcription of alternative tissue-specific promoters producing abundant, spliced, **multiexonic poly(A)$^+$ RNA** (meRNA)**s**. What the role of these meRNA might be is still unclear.

This complexity implies that no cancer can be cured by targeting a single gene or its product. Although there are quite some successful drugs on the market, patients treated with these will eventually suffer from a relapse, as the cancer cells endeavor new escape paths, see Figure 1.13. Therefore, it is important to start working at the systems level by integrating the results

of different technical approaches and understanding their combined effects on broader phenotypic traits, *i.e.* to approach cancer "not as a collection of genes" but as "a system of interactions"[1]. Ultimately this understanding

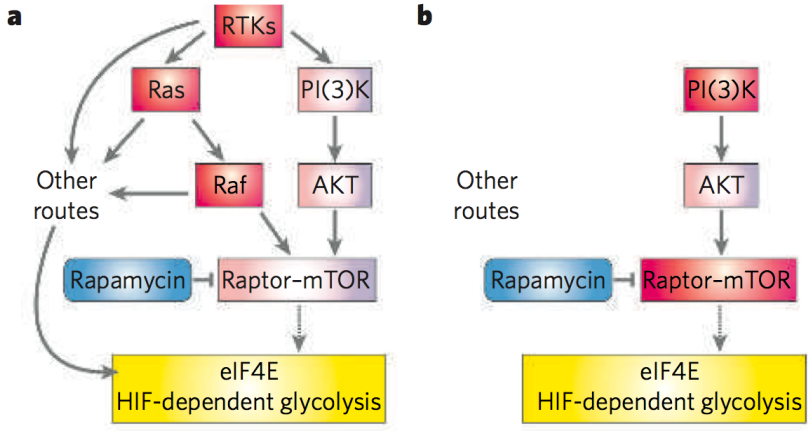


Figure 1.13 – Pathway circuitry influences therapeutic response: the action of the Rapamycin drug, inhibiting the mTOR pathway, is circumvented by Ras mutations. Adapted from Shaw and Cantley (2006).

will help define therapies best adapted to every single cancer patient, a necessity as every cancer is probably as different as different individuals are. In that regard, the unprecedented troughput of genomic technologies have made the bio-medicine research aim to "personalize" medicine reachable in the coming next decade(s). However, in the recent years, the ease with which humongous amounts of data have been generated revealed that the current bioinformatics analytical capabilities are the limiting factors (Park, 2009; Pepke et al., 2009; Koboldt et al., 2010; Scholz et al., 2012) and that consequent efforts have to be invested to implement the necessary tools for **Systems Biology** analyses. Relevant approaches to this issue will be presented in more details (*c.f* section 1.2, page 21) after a short overview of both tumors analyzed in this work.

[1] Adapted from Noble (2008)

1.1.2 Retinoblastoma

Retinoblastoma is an embryonic malignant neoplasm of retinal origin, that occurs most often in early childhood and is often bilateral - i.e. affecting both eyes - in hereditary cases. The ***retinoblastoma*** gene (*RB1*) was the first tumor suppressor gene (see paragraph 1.1.1, page 3) to be cloned in 1986 and it validated more than a decade of theories on how tumor suppressor genes could be involved in tumorigenesis. *RB1* was mapped to the band 14.2 of the q arm of chromosome 13 (13q14.2) and its role elucidated; it is a negative regulator of the cell cycle, binding the **transcription factor** (TF) E2F and repressing the transcription of genes necessary for the cell cycle S phase. Although Zhang et al. (2012) report very few chromosomal aberrations, *RB1* has been associated with **aneuploidy** and **Chromosomal INstability** (CIN). Zielinski et al. (2005) reported frequent aberrations: gains on chromosome arms 1q, 2p, 6p and 13q and losses on chromosome arms 13q and 16q.

Symptoms: The most common retinoblastoma symptoms are leukocoria (a late sign) and strabismus (an early sign), but many other ocular signs have been observed such as pseudohypopyon, elevated intraocular pressure, diffuse intra-ocular seeding - *i.e.* diffuse clusters ot tumor cells-, *etc.*[2]

Diagnosis: Ophthalmoscopic examination often shows a white 'cat's eye' reflex, a sign of a **leukocoria**, *i.e.* an opacity that obscures the retina, see Figure 1.14. This leukocoria signals the presence of a retinal tumor in one or both eyes, usually diagnosed by the age of 3 years.

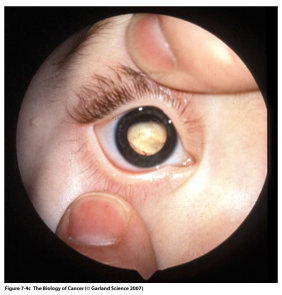


Figure 1.14 – A leukocoria - adapted from **The Biology of Cancer** (Weinberg, 2007), Copyright 2007 (c) Garland Science.

Prevalence: Retinoblastoma is a rare disease; rate between 1 in 23,000 (live birth in the US) and 1 in 200,000 (children under 15 in the US) are found in the literature. This difference in rate is probably due to the existence of inherited **familial** retinoblastoma as opposed to **sporadic** cases.

Inheritance: Familial retinoblastoma is often affecting both eyes: 2/3 of the hereditary cases are **bilateral**. In total, about 35% to 45% of all cases are hereditary. Bilateral tumors are almost never observed in sporadic cases. In most cases, one of the *RB1* alleles is lost by LOH, while the remaining

[2]The original source of information for this and the remaining sections on retinoblastoma is the **Online Mendelian Inheritance in Man** database (OMIM) entry 180200

mutated allele is significantly more frequently of paternal origin. Diverse theories have been proposed to explain that fact, from a difference between spermatogenesis and oogenesis to a lack of DNA repair in the early embryo for paternal chromosomes.

Penetrance: The penetrance of the mutated allele is highly dependent on the mutations affecting *RB1*. There are many possible mutations affecting different regions of the gene. Nonsense mutation usually result in more aggressive phenotypes. Missense mutations or deletions, affecting the protein function, have different degrees of penetrance. The mutant having the least penetrance will often manifest by the development of **retinoma**, a benign neoplastic growth.

Secondary tumors: More than 2/3 of the secondary tumors are of mesenchymal origin. 60% of those are osteosarcomas while others are soft tissue sarcomas. Strikingly, there's a 500 fold increase of osteosarcoma for retinoblastoma hereditary cases. Additionally, some familial cases will develop a **trilateral** tumor, where the **pineal gland** (the so-called third eye, due to its shared tissue of origin) will present a morphologically similar neoplasm.

Treatment: As of today, most of the time, the eventual treatment is the enucleation. Other approaches, irradiation, transscleral cryocoagulation, argon laser photocoagulation of tumors and feeder vessels, combination chemotherapy, *etc.* often fail to cure the disease.

Therapy: Very recently, Zhang et al. (2012) showed by a **Next-Generation Sequencing** (NGS) approach using four retinoblastoma samples and matched germline controls that only *RB1* had mutations. However, by analyzing the epigenetic profile, they observed the induction of the proto-oncogene **Spleen Tyrosine Kinase** (*SYK*) gene. Additional experiments *in-vitro* and *in-vivo* showed that targeting *SYK* with a small-molecule inhibitor resulted in retinoblastoma cell-death. This discovery is an interesting prospect for therapy.

1.1.3 Osteosarcoma

Osteosarcoma is a malignant neoplasm arising from mesenchymal transformed cells showing **osteoblastic** differentation. It is the most common form of primary bone cancer. Unlike other cancers, not much is known about the molecular genetics of osteosarcoma. As described in paragraph 1.1.2, page 16, it is often a secondary tumor associated with retinoblastoma, indicating a possible pre-dominant role of *RB1* in its tumorigenesis. Another interesting fact is that bones are a preferred metastatic site for many of the cancers occurring in the Western world, *e.g.* breast, lung and prostate carcinomas. Bones are a lively tissue, where **osteoblast** and **osteoclast** renew 10% of the skeletal mass per year. Osteoclast demineralizes the bones and then degrades the ECM. Osteoblast reconstructs them. This allows our skeleton to adapt to the possibly changing physical constraints of our bodies. Bones seem to be such a beneficial micro-environment for metastases because the bone ECM is unusually rich in trophic and mitogenic factors. Therefore metastases will thrive there once they deregulate the osteoclast/osteoblast balance toward one or the other, resulting in **osteolytic** metastasis, where bone is dissolved or **osteoblastic** metastasis where bones accumulate in the tumor vicinity.

Osteosarcoma, certainly benefit from that micro-environment, but apart from its **pathology** presented next[3], not much more is known.

Molecular genetics: As discussed, pRb is frequently affected. But unlike for retinoblastoma, it is here not a sufficient condition for a tumor to appear. Additional mutations of *TP53* or **CHEckpoint Kinase 2** (*CHEK2*) have been reported. Recently, Sadikovic et al. (2009) have shown the possible implication of **Runt-related transcription factor 2** (*RUNX2*), **Dedicator Of CytoKinesis 5** (*DOCK5*), **Tumor Necrosis Factor Receptor SuperFamily member 10A** (*TNFRSF10A*) and **Tumor Necrosis Factor Receptor SuperFamily member 10D** (*TNFRSF10D*). *RUNX2* is involved in the cell cycle regulation and *DOCK5*, *TNFRSF10A* and D encode for receptors involved in apoptosis.

Cytogenetics: As expected by the involvement of pRb and *TP53*, LOH of chromosome 13q14 and 17p13 are frequently observed. Additional aberrations have been reported, such as a 18q LOH, and in a recent study, Sadikovic et al. (2009), showed additional aberrations: 1q21.1-q21.3 and 6p21.1-p12.3 gain and 8p21.3-p21.2 deletion.

Diagnosis: Osteosarcoma will often first be misdiagnosed as cysts or muscle problems. Only x-ray or scans, *e.g.* CT-scan, can reveal the existence of

[3]The original source for these information is OMIM, Entry 259500

the tumor.

Symptoms: Patients complain of pain and if the tumor is large a swelling can appear. As the bone structure is affected, pathological fractures can occur.

Inheritance: As previously mentioned, *c.f.* paragraph 1.1.2, page 17, familial inheritance of a mutated *RB1* allele increases the risk of osteosarcoma. Additional inherited conditions predispose to the disease:

- Bone **dysplasias** such as the Paget's disease
- Li-Fraumeni syndrome, where a *TP53* mutated allele is inherited
- Rothmund-Thomson syndrome

Prevalence: The incidence rate is of 5 per million per year in the general population in the US. Almost half of these affected are children under 15, making it the 6^{th} most common childhood cancer.

Penetrance: Variable penetrance has been reported in familial cases. In murine animal models, it varies from very low to a 100% depending on the gene(s) mutated. Based on these observations, one can expect a similar situation for human osteosarcoma.

Treatment: The treatment relies on chemotherapy and surgical resection.

In the next sections, the bioinformatics' methodologies used to analyze the biology of these two cancers will be described.

1.2 Microarray and data analysis

During the last two decades, our abilities to decipher in the laboratory the genetic characteristics of a cell, such as karyotypes, genomic aberrations, epigenetic modifications, gene expressions, protein levels, *etc.* have been extended many folds through the establishment of high throughput technologies. For the analysis of genomic aberrations and gene expression - as performed in this thesis - it first started with the establishment of the microarray technology that offered an unprecedented resolution until the recent advent of NGS instruments. The upcoming of these high throughput technologies led to the development of a new field: **bioinformatics**. Indeed, the amount of generated data could no more be analyzed manually and specific softwares and algorithms had to be developed, requiring the interaction of the biological, chemical, statistical and computational fields. This occurred across a decade and turned microarray into an understood and mature technology. The developed tools gave researchers the possibility to move away from a reductionistic - *i.e.* focusing on a gene - to a systematic approach of the cancer disease. This shift is still occurring slowly, as it challenges both our bioinformatics competencies and resources as well as our human ability to understand such complex data. It is however a necessary step if we want to cure cancer. Very recently, Rausch et al. (2012); Stephens et al. (2011); Zhang et al. (2012) have demonstrated the advantages of such approaches.

1.2.1 Microarray technology overview

DNA microarrays consist of an arrayed series of thousands of microscopic spots of DNA oligonucleotides, originally printed on a coated silicon or glass slide support. Every spot contains picomoles of a unique DNA fragment: a probe (or a feature), to which the complementary DNA or cDNA fragment (the target) will hybridize. The targets have been modified to carry a fluorochrome (see Figure 1.15 [4]) and the amount of fluorescence emitted after excitation is recorded as the raw result. Signal intensities need to be normalized for both technical and biological variations before they can be used for further analyses.

History: The first approach at high throughput gene expression screening by Augenlicht and Kobrin dates back to 1982. The first experiment using an array of distinct DNA sequences with a computer assisted scanning and image processing was performed by Kulesh et al. in 1987. The first miniaturized array use was reported in 1995 by Schena et al. and the first complete **eukaryotic** genome on microarray, that of *Saccharomyces cerevisiae* was reported in 1997 by Lashkari et al. Rapidly, conventional genomics approaches were ported to microarrays *e.g.* the development of a microarray

[4] This image is from www.en.wikipedia.org/wiki/DNA_microarray

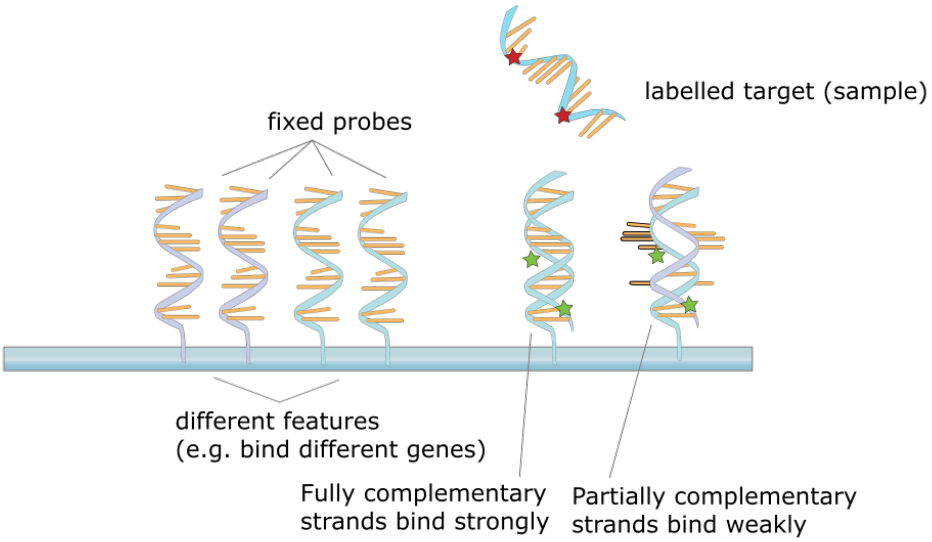


Figure 1.15 – The principle of microarray

based **Comparative Genomic Hybridization** (CGH) by Solinas-Toldo et al. (1997). These successes lead to industrialization and many companies started to produce microarrays, among which Agilent, Affymetrix, Illumina and Nimblegen. But before the microarray became an established everyday technique in most molecular genetics laboratories, many of these were producing their own spotted microarrays.

Technologies: There are as many technologies as there are microarray manufacturers, however they all share the same common protocol described in Figure 1.16[5]. The main difference is how the microarrays are generated. Independent laboratories use robots to print their own collection of probes, usually on a coated glass support. These microarrays are called "spotted" microarray. Their production is time and cost efficient for small production. The scalability is limited and the quality depends on many environmental conditions such as temperature, humidity, *etc.*. Finally, their storage time is relatively short. Probably for these reasons the four main manufacturers have developed their own technologies:

[5]This image is from www.en.wikipedia.org/wiki/DNA_microarray

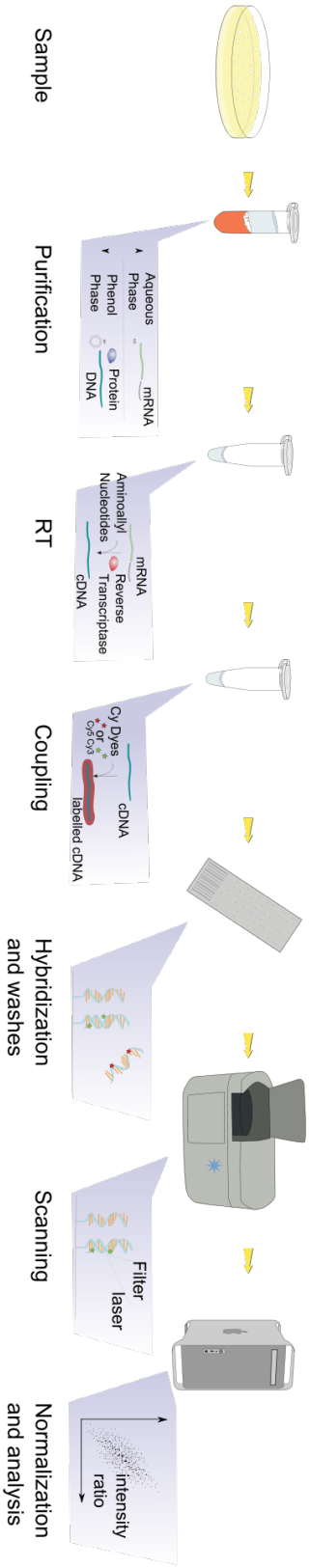


Figure 1.16 – The microarray procedure. First, the DNA or mRNA is extracted from the sample. In the case of mRNA it is converted into cDNA using a retro-transcriptase enzyme. Then the DNA/cDNA is labeled with fluorochromes, usually Cy3 and/or Cy5. The targets are then hybridized onto the microarray and this one is subsequently washed to remove weak, aspecific, bindings. The microarray is then placed into the scanner. This one uses a laser to excite the fluorochromes and record the emitted fluorescence. The recorded fluorescence is the raw data generated by the scanner and is the basis of the bioinformatics pipeline where the data quality will be checked, and if deemed reasonable, followed by the data normalization.

- Affymetrix uses a photolithographic approach. The probe sequences are assembled base per base on a coated support. At every synthesis step, UV light shading masks are used that protect the sequences that should not be extended during that particular step. This process limits the probes size to 25bp. To circumvent this short probe size issue - *i.e.* 25bp probes are likely to map to multiple place in the genome - Affymetrix designed sets of probes per feature: the so-called probe-sets. The very first microarray from Affymetrix was called *GeneChip*® (Lockhart et al., 1996) and for that reason are microarrays commonly known as **chip**.

- Agilent uses a technology based on industrial scale inkjet printing. The probes are longer, between 50 and 70bp long, as is the case for the next two platforms. Particular care is taken during the design to avoid possible probes cross-hybridization or formation of secondary structures that would impair the hybridization of the target.

- Illumina uses beads to which the probes are attached. The beads are dispersed on a microwell chip, each well being able to contain only a single bead. Every bead possesses a unique sequence identifier decrypted during the scanning process. This randomization of the probes' position helps avoiding border effects observed with other kinds of microarrays. In addition, the higher number of probes present on a single bead in comparison to that on a flat printed surface helps reducing the amount of starting material and Illumina claims that there is no need for sample **Polymerase Chain Reaction** (PCR) amplification prior to the hybridization.

- Nimblegen uses a proprietary system: the **Maskless Array Synthesizer** (MAS), similar to that of Affymetrix, but instead of using masks, they use digitally controlled micromirrors to redirect the UV light used to deprotect the sequences that need to be extended.

Applications: First, it is important to mention a crucial difference in the technical application of the afore mentioned technologies: whether they are used as single or two channel microarray. Single channel microarrays are used for quantitative measurements, whereas two channels are used for relative measurements. The advantage of a dual channel platform is that the difference between two samples is readily available with a single hybridization: *e.g.* two samples, one from the tumor and one from healthy tissue, marked with two different fluorochromes are hybridized on the same array. Most commercial manufacturer platforms are single channels, but there are a few two-channel specific products such as the Agilent Dual-Mode platform. Traditionally, *spotted* microarrays are dual channel.

Both can be used for numerous biological applications, an excerpt of which is listed in the following list. Microarrays are commonly used in two domains: for **genomics** and **transcriptomics**.

For transcriptomics:

- **Expression Profiling** (EP): DNA fragments of genes are spotted or synthesized on the array. In the case of Affymetrix *GeneChip*®, the probes are selected mostly from the 3' end of the gene. This was the original design for many microarrays as the 3' end of genes would be least affected by mRNA degradation. Given the relative low density achieved on such a chip, it was as well a way to measure several isoforms at the same time. Newer generations of microarrays, as well from other manufacturers, now have probes disseminated along the genes.

- **Alternative Splicing**: although probes are located along the whole genes for recent EP microarray design, and thus should allow to decipher isoform expression, the actual signal deconvolution is so complex that it has not been successful to date. To circumvent this issue, so-called **exon arrays** were developed, including probes overlapping the exon-exon junctions, offering a direct measurement of the corresponding isoforms.

- **Single Nucleotide Polymorphism** (SNP): Many diseases are known to originate from SNPs. The 1000 Genomes Project Consortium has reported in 2010 approximately 15 million of them. On average, each person carries 250 to 300 loss-of-function variants due to SNPs, **small insertions and deletions** or **structural variants**. The probes on these arrays represent different heterozygous SNPs, aiming at identifying the hybridization difference between allelic variants.

- **Tiling array**: These are very high density arrays, where the probes can even be partially overlapping. Rather than having known features (*e.g.* genes, exons,...) represented on the array, probes are designed that span entire genomic loci. This is useful for characterizing unannotated regions of the genome and to identify transcripts/expressed regions *de-novo*. Using such arrays, Xu et al. (2009) showed the bidirectionality of the transcription.

For genomics:

- CGH: the array based version of the CGH technique that compares the genomic content of two samples. This application developed by Solinas-Toldo et al. (1997); Pinkel et al. (1998), named either **array-CGH** or **matrixCGH**, relies on dual channel microarrays.

- **Copy Number Variation** (CNV): This application can be seen as a matrixCGH follow-up. Recent, very high density single arrays are used to quantify the genomic content of a sample and its variations. The higher resolution offers the possibility to discover smaller aberrations and to pinpoint more precisely chromosomal breakpoints.

- **ChIP-on-chip**: This application combines a chromatin immunoprecipitation with the hybridization of the precipitated DNA sequence on a microarray such as a tiling array. This has been widely used to identify the binding events of TFs and more recently of specific histone modifications.

- **Methylation**: There have been many kinds of microarrays developed to analyze the genome methylation, especially for looking at **CpG Islands**. An implementation using dual channel microarray was proposed by Pfister et al. (2007). Other applications, such as MeDIP-chip use Methyl-DNA immunoprecipitation (similar to ChIP-on-chip) followed by an hybridization on microarray (Weber et al., 2005) or more recently on tiling array (Pälmke et al., 2011).

- **Chromosome Conformation Capture** (3C): This technique used to identify possible intra and inter chromosomal interactions based on PCRs has been modified to be used in conjunction with microarrays. In **Circularized 3C** (4C), one end of the chromosomal interaction is known and microarrays are used to find its mate(s). In **Carbon-Copy 3C** (5C), the generated fragments are amplified using a multiplex ligation-mediated amplification, *i.e.* a technique where multiple targets can be amplified with a single primer pair, before their hybridization to a microarray, allowing to screen for multiple interacting loci at once.

In this thesis, EP and matrixCGH data are analyzed, which were generated using the Affymetrix *GeneChip*®, and in-house produced 6,200 spots microarrays, respectively.

Data Analysis: A microarray data analysis consists of two steps: the **pre-processing** that deals with converting the raw data into normalized data and the **post-processing**, where this data is analyzed in the light of the experimental design to answer a particular biological question.

The pre-processing starts with the image analysis: the obtained intensities along with additional control parameters are used for **Quality Assessment** (QA). If the quality is within acceptable standards, the intensities are normalized and associated with their respective feature either through an annotation file or through specific identification processes. The pre-processing

results in normalized intensity values ready for post-processing. In the following paragraphs, additional details specific to the EP and matrixCGH microarrays used in this thesis will be given.

Image Analysis: In the case of Affymetrix, this process is fully automated within the scanner. In the case of *spotted* microarray, this process requires software able to determine the position and the size of the spots. Spots the shape of which are not circular enough, the fluorescence of which are not homogenous or the signal of which are indiscernible from the surrounding background signal (due to non-specific target hybridization), *etc.* are filtered out. Evaluating the proportion of those is the first QA step.

Quality Assessment: Additional **Quality Controls** are performed that need to be adapted to the particular experimental design at hand. For *spotted* arrays, one would look for example at possible local effects, denoting fluctuating concentration of material during the hybridization process. For Affymetrix arrays, one would look at "spiked-in" controls, *i.e.* controls that should have an expected fluorescence intensity. In addition, the biological design of the experiment might offer additional means of ensuring quality, *e.g.* the results of technical or biological replicates can be compared. Many softwares are available to perform such analyses, an example for the Affymetrix platform is the **Bioconductor** package *arrayQualityMetrics* (Kauffmann et al., 2009).

The next pre-processing steps, annotation, feature identification and normalization are intertwined and their order depends on the platform used. For that reason, they will be introduced in an arbitrary order in the following paragraphs.

Annotation: One of the limitation of microarrays is our incomplete knowledge of the genome. Indeed, an array is designed based on the genome version available at the time of creation. Given the update rate of most genomes, especially since the advent of NGS techniques, the annotation of the probes might change drastically. For example, only about 70% of the probes of a 10 years old Affymetrix *GeneChip*® (HG-U95Av2) map uniquely in the latest version of the human genome This underlines the importance of keeping the probe annotation up-to-date. In the case of Affymetrix, these annotations are stored in a **Custom Definition File** (CDF). This file records the probe genomic position, as well as their probe-set membership. An additional file records the associated gene information. In the case of matrixCGH arrays, the important information is the genomic location, as accurate as possible, of the probe. Indeed, in a matrixCGH experiment, the final results - gains or losses of chromosomal material and their extend - is not a prior knowledge and can only be determined after normalization

using specialized softwares. Keeping the probe information up-to-date is the task of dedicated softwares, such as "customCDF", which was developed for this thesis work; see appendix C, page 197 for the submitted version of the corresponding manuscript.

Feature Identification: This step is important for matrixCGH arrays. After normalization of the data, it is important to identify the features of interest: the CNV and their boundaries. The process used for this is called **segmentation**. Many algorithms have been published involving different implementations ranging from **circular binary segmentation** to the use of **Hidden Markov Model** (HMM). Most of these methods have been reviewed by Lehmussola et al. (2006) and as no method outperforms the others, the choice of a segmentation method depends essentially on the data at hand.

Normalization: This is an essential step of the data pre-processing. Without it, it would be impossible to compare values within and between arrays. Within arrays, local fluorescence variability, possible cross hybridization, GC nucleotide percentage of the probe sequences, *etc.* are parameters that need to be adjusted for. Between arrays, variation due to the local condition of the run (humidity, temperature, *etc.*), as well as technical variation (more potent fluorochrome, variation of the camera position, *etc.*) need to be corrected. Popular Bioconductor packages that implement such normalization for Affymetrix microarray data are: *rma* (Irizarry et al., 2003) and *vsn* (Huber et al., 2002).

Post-processing normalized data: This step is entirely dependent on the experimental design. Commonly for microarray, differences between samples will be investigated, *e.g.* differences of mRNA abundance between tumor and control samples. Numerous analysis tools have been developed and this for every microarray application field. Not to list all these, only the selection of tools used in this work are detailed in the Materials and Methods chapter (chapter 3, page 41).

Although microarrays are restrictive - the observable data is constrained by their design - they are very powerful tools to investigate gene expression, CNV, *etc.* Each of these tools has a targeted application and although it might bring useful insight into a particular problem, its systemic resolution is limited by the biological complexity in general, and of cancer in particular. To alleviate that complexity and help discerning causes from consequences, a solution is to combine several of these tools together and/or with other non-microarray based approaches. This is the topic of the next section.

1.2.2 Integrative analysis

Relatively early in the usage of microarray, researchers realized that the large amount of data generated could not be analyzed by traditional means. Indeed, by lack of better approaches, most studies would focus on a few candidates from the hundreds or thousands identified. The candidates choices were rather arbitrary and a lot of information remained unveiled. The need for integrative techniques of analysis recognized, a large body of studies started to report their implementation: functional enrichment analysis, transcriptional network analysis, interactome analysis, *etc.* florished (reviewed in Rhodes and Chinnaiyan (2005)). At the same time Segal et al. (2005) underlined the importance of looking at gene modules, *i.e.* set of genes involved in the same biological processes rather than at a handful of candidates; an approach that relies on using interactome and **Gene Ontology** (GO) annotation. The drawback is that such information are incomplete: *e.g.* most of the interactions in the interactome databases come from *yeast 2 hybrid* experiments, which due to that method technicalities are mostly about soluble proteins, including only few membrane proteins - as are mitogenic receptors. This is not the only issue as cancer samples usually consist of an heterogenous combination of cell types and present additional source of variability - *e.g.* chromosome instabilities - that introduces confounding effects. In this context, combining different kinds of data obtained from identical samples allow to condense the data to a more relevant subset, as was achieved by Garrett-Mayer et al. (2008); Łastowska et al. (2007) for EP and matrixCGH data.

In parallel to these integrative approaches, *in-silico* procedures to represent gene modules have been undertaken. They offer the possibility to manipulate them, by changing their input state or by refining their circuitry in order to assess the outcome of given mutations. Franke et al. (2008) reported the effect of infection by *H.pylori* on the *c-Met* signal transduction network by performing *in-silico* knock-outs and validating their prediction *in-vivo*.

Integrative analyses have become an essential part of the studies of cancer, but despite successful recent stories, *e.g.* Sadikovic et al. (2009) or Rausch et al. (2012) who used such analyses in a specific context, it is still difficult to integrate data from different approaches into results that are conceivable to our understanding. To achieve this at the scale of a complete dataset in a statistically sound manner is complex and has not been done to date, although an attempt at analysing the effects of CNV gains on gene expression levels was done by Hyman et al. as early as 2002. A decade later, there is still very few literature available on this topic in biology and bioinformatics. To compare discrete data (such as arrayCGH CNVs) with continuous data (such as EP values) is not trivial and of concern in the field of statistics only. However, a number of statistical methods have been

developed to adress that question in the research field of economy. Among these, 5 methods - some of which are common correlation methods - can be applied to biological data:

- Eta: originally developed in the field of economics, it is a measure of the relationship between the statistical dispersion within individual categories and the dispersion across the whole data.
- Weight: the original method described in Hyman et al. (2002). It evaluates whether the mean difference between two categories can be explained by the sum of the variation within these catagories. A large score is indicative of a correlation between the continous and corresponding discrete values. It is however limited to two categories and every category has to have at least 2 members.
- Welch: a modified F-test for unequal variance introduced by Welch in 1951. Since it uses a weighted approach based on a constant, it cannot be applied in cases where the variance is null, nor if any of the categories has no members.
- Pearson: an established parameterized correlation method assuming that the data is normaly distributed and **homoscedastic**.
- Spearman: rank based parameter free Pearsons' equivalent, independent of the normality and equal variance assumptions.

These methods were applied in this thesis work in addition to *de-novo* developed approaches detailed in the Materials and Method chapter (chapter 3, page 41).

The obtained correlation values are in the $[-1, 1]$ range. While positive correlation are expected (*e.g.* a gene CNV gain resulting in its expression increase), anti-correlation would be more surprising (*e.g.* a gene CNV loss resulting in its expression increase). Table 1.2 shows the correlation range and the possible gene expression dosage effect for an arrayCGH - EP integrative analysis.

To summarize, integrative analyses are tools meant for combining datasets obtained from a similar source (*e.g.* a tumor type) in order to increase the detective power of these and limit confounding factors' effects.

arrayCGH/EP	-1	1
-1	• $w \in (0, 1]$ • decrease	• $w \in [-1, 0)$ • positive compensation
1	• $w \in [-1, 0)$ • negative compensation	• $w \in (0, 1]$ • increase

Table 1.2 – Contingency table of the w correlation score ranges for arrayCGH (change in copy number) - EP (change in expression) pairs observed in an integrative analysis and their respective meaning on gene expression.

1.2.3 Comparative analysis

Combining different technical approaches to get a better understanding of a system, as done by integrative analyses, is one way to address the complexity of these diseases. Another approach is the so-called **comparative analysis** that combines data from different conditions, tissues or organisms to get an overview of system-wide processes. Such approaches have been successful for example to identify the subset of cells within a primary tumor, the signature of which is identical to that of the metastases (Ramaswamy et al., 2003) supporting the theory that the ability to form distant metastases is a consequence of alterations in a subset of cells of the primary tumor and not due to a selection pressure promoting the invasion/metastasis phenotype. Another example of such analyses is the use of animal models to create *in-vivo* models of tumor development. Brumby and Richardson (2005) reviewed *Drosophila melanogaster* models of several hallmarks of cancer, including cell growth and proliferation, invasion and metastasis, survival, and the failure to differentiate.

As both the integrative and comparative analyses have been seldom described in the literature and are a major achievment of the present work, they'll be extensively detailed in the Materials and Methods chapter (chapter 3, page 41)

Ideally, combining these two kinds of analyses should give us an even better understanding of complex problems. For this approach, the fact that familial retinoblastoma patients will have in the majority of cases a subsequent osteosarcoma, analyses of secondary tumors is an interesting setup. But prior to these analyses, it was necessary to establish the validity of the algorithms and equations developed for them. It required the development of yet another tool able to accurately simulate various kinds of microarrays data, introduced in the next section.

1.2.4 Microarray simulation

The use of biologically realistic simulated data is the most appropriate way to benchmark newly developed algorithms and test the validity of their assumptions as the *true* outcome is known. Accurate **True Positive Rate** (TPR) (*sensitivy*) and **False Positive Rate** (FPR) ($1 - specificity$) can be deduced. They help correct for type II and type I errors, respectively, and can be easily visualized as **Receiver Operating Characteristic** (ROC) curves.

An important aspect for the validity of this approach is the adequate selection of the data generating models. For example, EP microarrays models shoud be derived from experimental data by re-sampling or by constructing differential equation models, which describe the time courses of gene expression as in Chen et al. (2000). These models are then modified by including

systematic bias and stochastic noise (Cho and Lee, 2004; Dror et al., 2003; Tu et al., 2002). A set of models, obtained by defining a set of parameters adequately describing the experimental data, is then used to build artificial datasets with known characteristics. The comparison of the observed versus the expected outcome of a given data analysis algorithm applied to simulated data, enables the data analyst to evaluate the strengths and restrictions of the algorithm under different conditions (Costa et al., 2008; Willenbrock and Fridlyand, 2005).

Besides the validation and improvement of single data analysis algorithm, microarray data simulators are useful tools for the design of experiments. By simulating entire microarray experiments, it is possible to estimate the sample size (and the amount of arrays) required to test a hypothesis (Gadbury et al., 2004) or to pinpoint possible problems during the data analysis procedure.

As of today, several microarray simulators have been published (Singhal et al., 2003; Balagurunathan et al., 2002; Nykter et al., 2006; Wierling et al., 2002) - most of them focus on measuring the performance of image segmentation software, which is required to calculate the raw data values for each feature of a microarray after image scanning. The simulators from Balagurunathan et al., Nykter et al. and Wierling et al. therefore create in silico images and benchmark the impact of critical parameters for the image analysis like spot distortion, or background signal effects.

The microarray simulator by Singhal et al. (2003) creates feature values as obtained after image segmentation (raw values) and normalization (Huber et al., 2002; Smyth, 2004). It is a useful tool to benchmark data analysis algorithms, which identify differences in RNA expression levels between sample groups. Unfortunately, its design restricts it to single channel EP microarray data as obtained *e.g.* from nylon filters or Affymetrix arrays. To date, there seem to be no microarray simulator available, which generates feature extracted data for different array platforms like those dedicated to EP or arrayCGH.

To address this need, I implemented a generic microarray simulator, which generates data with realistic biological and statistical characteristics, as an R package: `aSim`. `aSim` provides methods to simulate data for a variety of microarray platforms and is not limited to a specific array layout (Affymetrix, Agilent, Illumina, custom-made, etc.). In addition, the simulation parameters are part of the `aSim` output, which allows the exact reproduction of the simulated data and offers the possibility to exchange these parameters within the microarray community.

The implementation and validation of `aSim` will be described in the next chapters. It was an essential tool to ensure the rigorous integrative and comparative analyses of the retinoblastoma and osteosarcoma dataset presented in this thesis.

Bibliography

1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73, Oct 2010. doi: 10.1038/nature09534.

Leonard H Augenlicht and Diane Kobrin. Cloning and screening of sequences expressed in a mouse colon tumor. *Cancer Res*, 42(3):1088–93, Mar 1982.

Yoganand Balagurunathan et al. Simulation of cdna microarrays via a parameterized random signal model. *J Biomed Opt*, 7(3):507–23, Jul 2002.

Anaïs Baudot et al. Mutated genes, pathways and processes in tumours. *EMBO Reports*, 11(10):805–10, Oct 2010. doi: 10.1038/embor.2010.133.

Anthony M Brumby and Helena E Richardson. Using drosophila melanogaster to map human cancer pathways. *Nat Rev Cancer*, 5(8): 626–39, Aug 2005. doi: 10.1038/nrc1671.

Katherine C Chen et al. Kinetic analysis of a molecular model of the budding yeast cell cycle. *Mol Biol Cell*, 11(1):369–91, Jan 2000.

HyungJun Cho and Jae K Lee. Bayesian hierarchical error model for analysis of gene expression data. *Bioinformatics*, 20(13):2016–25, Sep 2004.

Ivan G Costa et al. Inferring differentiation pathways from gene expression. *Bioinformatics*, 24(13):i156–64, Jul 2008.

Lisa M. Coussens, Wilfred W. Raymond, Gabriele Bergers, et al. Inflammatory mast cells up-regulate angiogenesis during squamous epithelial carcinogenesis. *Genes Dev*, 13(11):1382–97, Jun 1999.

Ron O Dror et al. Bayesian estimation of transcript levels using a general model of array measurement noise. *J Comput Biol*, 10(3-4):433–52, Jan 2003.

Melissa A Dunham et al. Telomere maintenance by recombination in human cells. *Nature Genetics*, 26(4):447–50, Dec 2000. doi: 10.1038/82586.

Raimo Franke, Melanie Müller, et al. Host-pathogen systems biology: logical modelling of hepatocyte growth factor and helicobacter pylori induced c-met signal transduction. *BMC Syst Biol*, 2:4, Jan 2008. doi: 10.1186/1752-0509-2-4.

Gary L Gadbury et al. Power and sample size estimation in high dimensional biology. *Statistical Methods in Medical Research*, 13(4):325–338, 2004.

Elizabeth Garrett-Mayer et al. Cross-study validation and combined analysis of gene expression microarray data. *Biostatistics*, 9(2):333–54, Apr 2008. doi: 10.1093/biostatistics/kxm033.

Douglas Hanahan. Benefits of bad telomeres. *Nature*, 406(6796):573–4, Aug 2000. doi: 10.1038/35020662.

Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, Jan 2000.

Wolfgang Huber et al. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:S96–104, Jan 2002.

Elizabeth Hyman et al. Impact of dna amplification on gene expression patterns in breast cancer. *Cancer Res*, 62(21):6240–5, Nov 2002.

Rafael A Irizarry et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–64, Apr 2003. doi: 10.1093/biostatistics/4.2.249.

Steven M Johnson et al. Ras is regulated by the let-7 microrna family. *Cell*, 120(5):635–47, Mar 2005. doi: 10.1016/j.cell.2005.01.014.

Audrey Kauffmann, Robert Gentleman, and Wolfgang Huber. arrayqualitymetrics–a bioconductor package for quality assessment of microarray data. *Bioinformatics*, 25(3):415–6, Feb 2009. doi: 10.1093/bioinformatics/btn647.

Alfred G Knudson. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci USA*, 68(4):820–3, Apr 1971.

Daniel C Koboldt, Li Ding, Elaine R Mardis, and Richard K Wilson. Challenges of sequencing human genomes. *Briefings in Bioinformatics*, 11(5): 484–98, Sep 2010.

Monika S Kowalczyk, Jim R Hughes, et al. Intragenic enhancers act as alternative promoters. *Molecular Cell*, Jan 2012. doi: 10.1016/j.molcel.2011.12.021.

David A Kulesh et al. Identification of interferon-modulated proliferation-related cdna sequences. *Proc Natl Acad Sci USA*, 84(23):8453–7, Dec 1987.

Deval A Lashkari et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc Natl Acad Sci USA*, 94(24):13057–62, Nov 1997.

Maria Łastowska et al. Identification of candidate genes involved in neuroblastoma progression by combining genomic and expression microarrays with survival data. *Oncogene*, 26(53):7432–44, Nov 2007. doi: 10.1038/sj.onc.1210552.

Antti Lehmussola, Pekka Ruusuvuori, and Olli Yli-Harja. Evaluating the performance of microarray segmentation algorithms. *Bioinformatics*, 22 (23):2910–7, Dec 2006. doi: 10.1093/bioinformatics/btl502.

David J Lockhart et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, 14(13):1675–80, Dec 1996. doi: 10.1038/nbt1296-1675.

Frank McCormick. Signalling networks that cause cancer. *Trends Cell Biol*, 9(12):M53–6, Dec 1999.

Denis Noble. *The Music of Life*. OUP Oxford, 2008. ISBN 0199228361.

Matti Nykter et al. Simulation of microarray data with realistic characteristics. *BMC Bioinformatics*, 7:349, Jan 2006.

Nina Pälmke, Diana Santacruz, and Jörn Walter. Comprehensive analysis of dna-methylation in mammalian tissues using medip-chip. *Methods*, 53 (2):175–84, Feb 2011. doi: 10.1016/j.ymeth.2010.07.006.

Peter J Park. Chip-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669, Sep 2009.

Shirley Pepke, Barbara Wold, and Ali Mortazavi. Computation for chip-seq and rna-seq studies. *Nature Methods*, 6(11 Suppl):S22–32, Nov 2009.

Stefan Pfister, Christof Schlaeger, et al. Array-based profiling of reference-independent methylation status (aprimes) identifies frequent promoter methylation and consecutive downregulation of zic2 in pediatric medulloblastoma. *Nucleic Acids Research*, 35(7):e51, Jan 2007. doi: 10.1093/nar/gkm094.

Daniel Pinkel et al. High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, 20(2):207–11, Oct 1998. doi: 10.1038/2524.

Sridhar Ramaswamy et al. A molecular signature of metastasis in primary solid tumors. *Nature Genetics*, 33(1):49–54, Jan 2003. doi: 10.1038/ng1060.

Tobias Rausch, David T W Jones, Marc Zapatka, Adrian M Stütz, et al. Genome sequencing of pediatric medulloblastoma links catastrophic dna rearrangements with tp53 mutations. *Cell*, 148(1-2):59–71, Jan 2012. doi: 10.1016/j.cell.2011.12.013.

Daniel R Rhodes and Arul M Chinnaiyan. Integrative analysis of the cancer transcriptome. *Nature Genetics*, 37 Suppl:S31–7, Jun 2005. doi: 10.1038/ng1570.

Peyton Rous. A sarcoma of the fowl transmissible by an agent separable from the tumor cells. *Journal of Experimental Medicine*, no. 4(13):397411, 1911.

Bekim Sadikovic, Maisa Yoshimoto, et al. Identification of interactive networks of gene expression associated with osteosarcoma oncogenesis by integrated molecular profiling. *Hum Mol Genet*, 18(11):1962–75, Jun 2009. doi: 10.1093/hmg/ddp117.

Mark Schena, Dari Shalon, et al. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–70, Oct 1995.

Matthew B Scholz, Chien-Chi Lo, and Patrick S G Chain. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Curr Opin Biotechnol*, 23(1):9–15, Feb 2012.

Eran Segal et al. From signatures to models: understanding cancer using microarrays. *Nature Genetics*, 37 Suppl:S38–45, Jun 2005. doi: 10.1038/ng1561.

Reuben J Shaw and Lewis C Cantley. Ras, pi(3)k and mtor signalling controls tumour cell growth. *Nature*, 441(7092):424–30, May 2006. doi: 10.1038/nature04869.

Sunil Singhal et al. Microarray data simulator for improved selection of differentially expressed genes. *Cancer Biol Ther*, 2(4):383–91, Jan 2003.

Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article3, Jan 2004.

Sabina Solinas-Toldo et al. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, 20(4):399–407, Dec 1997.

Dominique Stehelin et al. Dna related to the transforming gene(s) of avian sarcoma viruses is present in normal avian dna. *Nature*, 260(5547):170–3, Mar 1976.

Philip J Stephens et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, 144(1):27–40, Jan 2011. doi: 10.1016/j.cell.2010.11.055.

Yuhai Tu, Gustavo Stolovitzky, and Ulf Klein. Quantitative noise analysis for gene expression microarray experiments. *Proc Natl Acad Sci USA*, 99 (22):14031–6, Oct 2002.

Michael Weber et al. Chromosome-wide and promoter-specific analyses identify sites of differential dna methylation in normal and transformed human cells. *Nature Genetics*, 37(8):853–62, Aug 2005. doi: 10.1038/ng1598.

Robert A Weinberg. *The biology of Cancer*. Garland Science, 2007. ISBN 0815340761.

Bruce L Welch. On the comparison of several mean values: An alternative approach. *Biometrika*, 38(3/4):330–336, Dec 1951.

Christoph K Wierling, Matthias Steinfath, Thorsten Elge, Steffen Schulze-Kremer, Pia Aanstad, Matthew Clark, Hans Lehrach, and Ralf Herwig. Simulation of dna array hybridization experiments and evaluation of critical parameters during subsequent image and data analysis. *BMC Bioinformatics*, 3:29, Oct 2002.

Hanni Willenbrock and Jane Fridlyand. A comparison study: applying segmentation to array cgh data for downstream analyses. *Bioinformatics*, 21 (22):4084–91, Nov 2005.

World Health Organization. *International Classification of Diseases*. World Health Organization, Geneva, Switzerland, 2010. URL `http://www.who.int/classifications/icd/en/`. ISBN 9789241545402.

Zhenyu Xu, Wu Wei, Julien Gagneur, et al. Bidirectional promoters generate pervasive transcription in yeast. *Nature*, 457(7232):1033–7, Feb 2009. doi: 10.1038/nature07728.

Jinghui Zhang, Claudia A Benavente, Justina Mcevoy, Jacqueline Flores-Otero, et al. A novel retinoblastoma therapy from genomic and epigenetic analyses. *Nature*, 481(7381):329–34, Jan 2012. doi: 10.1038/nature10733.

Boris Zielinski et al. Detection of chromosomal imbalances in retinoblastoma by matrix-based comparative genomic hybridization. *Genes Chromosomes Cancer*, 43(3):294–301, Jul 2005. doi: 10.1002/gcc.20186.

Chapter 2

Aims of this doctoral work

The first aim of this thesis work was the development and enhancement of bioinformatics approaches to better analyze high throughput data generated using microarrays. In the recent past, these have been used to investigate every step of the "central" dogma of biology: matrixCGH has been used to characterize DNA copy number changes, EP microarrays to monitor gene expression and protein-arrays to determine protein modifications, interactions, *etc.* A large number of datasets, each concentrating on one of these aspects, have been produced and with their analyses came the realization that cellular mechanisms were even more complex than anticipated. However, understanding these is key to develop preventive, diagnostic and curative methods for a large number of genetic diseases. To start to undertand the complexity, two essential pre-requisites had to be adressed: first the effects of confounding factors needed to be - if not removed - controlled for. Here, the quality assessment, the normalization and the use of proper annotation for pre-processing the microarray data played a key role. Second, the data complexity needed to be deconvoluated so that its elements could be efficiently analyzed, implying the development of new statistical and analytical approaches.

These methods are of a broader interest as they are not restricted to microarray data analysis but can be declined to perform any high throughput generated data analysis. Hence, the second aim of this was to validate them by addressing two cancer relevant biological questions: first, the effect of CNV on gene-expression was investigated in an integrative analysis approach using retinoblastoma datasets. Then, based on the observation that retinoblastoma patients have a higher chance than the average population to develop osteosarcoma, the similarities and differences of these two tumors were assessed at the gene expression level.

Finally, the overall aim of this work was to demonstrate the feasibility and efficacy of systems biology approaches to study cancerogenesis, with the ultimate goal to develop personally adjusted therapies.

Chapter 3

Material and Methods

3.1 Material

Seven sets of data have been used in this study. Five Affymetrix *GeneChip*® (Lockhart et al., 1996) EP experiments were retrieved from the **Gene Expression Omnibus** (GEO) (Barrett et al., 2007) database. The sixth, a matrixCGH experiment (Solinas-Toldo et al., 1997; Wessendorf et al., 2002), was generated in-house and shares the same samples as one of the EP sets. The final set, performed in-house as well, consists of 3 additional EP samples.

3.1.1 Biological samples

Only the samples handled in the **Division of Molecular Genetics at the DKFZ** are described here, the remaining are listed in the section 3.1.2, page 42.

matrixCGH samples: These samples, seventeen cases of retinoblastoma, have been described in Zielinski et al. (2005). Shortly, three were hereditary, among which two are bilateral. Among the fourteen non-hereditary, only one was bilateral.

EP samples: These samples, 3 in total, were obtained from different sources. One was provided by Professor Dr. Lohmann from the Institut of Human Genetics at the University Hospital of Duisburg-Essen, Germany. Another one was provided by Dr. Stephan Wolf, in collaboration with the University Hospital of Heidelberg, Germany. The last one was purchased from Clontech Laboratories, Inc. This last one is a sample pooled from 29 male/female caucasians, see Appendix A, page 175 for more details. For the sample description, see Table A.1, page 177 in the same appendix.

3.1.2 *In-silico* data

Five datasets of EP data were retrieved from the GEO database. These were:

- **GSE5222**: a dataset, companion of the matrixCGH dataset from Zielinski et al. (2005), used in the retinoblastoma studies by Grasemann et al. (2005); Gratias et al. (2007). For the sample description, see Table A.1, page 177 in the appendix A. The samples were hybridized to Affymetrix *GeneChip*® HG-U133A.

- **GSE29683** and **GSE29684**: this dataset originates from the study of Mcevoy et al. (2011), who reported that retinoblastoma cells show multiple features of different other cell types of ocular origin. See the Table A.2 and A.3, page 178-180 in the appendix A for more details. The samples were hybridized to Affymetrix *GeneChip*® HG-U133Plus2.

- **GSE14359**: this dataset from the Fritsche-Guenther et al. (2010) study compares osteosarcoma primary tumor with lung metastasis tissue and non-neoplastic osteoblasts. See the Table A.4, page 181 for more details. The samples were hybridized to Affymetrix *GeneChip*® HG-U133a.

- **GSE14827**: this dataset is from the Kobayashi et al. (2010) study, which investigated 27 cases of osteosarcoma with or without pulmonary metastases to clarify the genomic basis of the development of such metastases. The samples were hybridized to Affymetrix *GeneChip*® HG-U133Plus2. See the Table A.5, page 182 for additional information about these samples.

- **GSE5350**: this dataset is from the **MicroArray Quality Control** (MAQC) study (MAQC Consortium et al., 2006) that investigated the reliability and reproducibility of microarray experiments accross microarray platforms and facilities. Here, only the results from the hybridization of the Universal Human Reference RNA (UHRR) sample from Stratagene to Affymetrix *GeneChip*® HG-U133Plus2 are considered; a total of 30 microarrays that consist of 5 replicates done in 6 different facilities. The UHRR sample has been commercialized by Strategene as a *universal control* for microarray experiments. See the Table A.6, page 183 for a summary.

3.2 Microarray methods

3.2.1 Quality Assessment

The QA is performed on all datasets to identify and remove microarrays, the inclusion of which would introduce confounding factors. For the matrixCGH dataset an in-house software - **ChipYard** (Toedt et al., b) - is used for performing the QA. For the EP datasets, the R Bioconductor package `arrayQualityMetrics` (R Development Core Team, 2009; Gentleman et al., 2004; Kauffmann et al., 2009) is used. See appendix B, page 184 for an example of a QA report generated by that package on the DKFZ dataset.

3.2.2 Probe-set annotation

Affymetrix *GeneChip*®: Affymetrix *GeneChip*®s, as introduced (see paragraph 1.2.1, page 27), depend on CDFs to combine their probes into probe-sets. These probe-sets represent one gene-related feature, *e.g.* a transcript, a gene, *etc.*, and the intensity readouts of their individual probes need to be combined into a single value, representing their expression level. Since the introduction of these arrays, more than a decade ago, the human genome has been constantly refined. These updates imply that the CDF files need to be refined as well. Several attempts to create custom CDFs have been conducted (Gautier et al., 2004; Dai et al., 2005; Ferrari et al., 2007; Lu et al., 2007), some very successful, but always using a very stringent approach. For example, these custom CDFs will ignore about 30% of all the probes present on the HG-U95Av2 *GeneChip*®, because they do not map a gene, or map antisense to it, or map several places in the genome. These unused probes can be grouped into interesting probe-sets, and this approach was taken in the division of Molecular Genetics, DKFZ, to develop a new CDF pipeline (Delhomme et al., submitted), see the manuscript in Appendix C, page 197.

Shortly, the probes are aligned against the genome version of choice. Using the corresponding **Ensembl** (Flicek et al., 2011) version, the probes' annotations are retrieved based on their genomic location. The probes are then grouped into probe-sets in a transcript-centric manner whenever possible. If several transcripts of the same gene are identified, this information is stored in the annotation. If all transcripts are mapped, then the annotation becomes gene-centric for that particular probe-set. This process is done using probes uniquely mapping the genome, provided that there are more than 5 probes per probe-set, a criterion based on the statistical analyses performed by Lu et al. (2007) that should achieve appropriate robustness without significant loss of power. This is similar to the other CDF generation pipelines mentioned, except for a few differences:

- their probe-sets are all gene-centric
- Ferrari et al. (2007) enforce the probe-sets to have a number of member probes similar to that of Affymetrix, *i.e.* 11 to 16 probes per sets.

In addition, in this approach the probes having no associated feature are collected and processed to become part of three groups:

- probes antisense to a gene are grouped into sets provided that the resulting set has at least 5 members and that the inter-probe distance is not larger than 1kb.
- probes located in non-genic regions are grouped into probe-sets with the same limitations as that of the antisense probe-sets.
- multiple mapping probes are grouped into probe-sets, transcript-centric or gene-centric as described above, in a manner that minimizes the amount of locations.

After this process, only a minimal number of probes are discarded (1%) and valuable probe-sets added to the CDF.

Inhouse matrixCGH: The inhouse spotted matrixCGH uses a library consisting of 6,200 **Bacterial Artificial Chromosome** (BAC) genomic fragments. 3,200 originate from the Wellcome Trust Sanger Institute and the remaining 3,000 are either from the **RZPD** or Invitrogen (CalTech BAC library). To ensure traceability, these probe's information was first stored in a local database (*CloneBase*) that interacted with the **Laboratory Information Management System** (LIMS) *QuickLIMS* (Kokocinski et al., 2003) established in the lab. This system was replaced by **PIMS** (Probe Information Management System) (Blond and Delhomme, unpublished) that ensures that these probes' annotations are kept up to date by frequently - once every other month - retrieving the lastest Ensembl (Flicek et al., 2011) relevant genomic and genic information. PIMS has been tightly coupled to the other tools developed in the department: ChipYard (Toedt et al., b, unpublished) and the **Flexible Annotation and Correlation Tool** (FACT) (Kokocinski et al., 2005). FACT offers the possibility to manually curate and re-annote the probes' information.

3.2.3 Expression Profiling

An expression profiling data analysis consists of several steps: QA, sample selection, sample normalization, experimental design and if adequate, **Differential Expression** (DE) analysis. Only after these steps the data can be biologically interpreted. In the following paragraphs, the normalization, experimental design and DE methods used in this manuscript are described.

Selecting the normalization method: There are four main methods to analyze Affymetrix *GeneChip*® data. The Affymetrix proprietary `MAS5` method will not be used nor discussed here. The results of the three remaining methods: `rma` (Irizarry et al., 2003), `gcrma` (Wu and Irizarry, 2005) and `vsn` (Huber et al., 2002) - all implemented in packages available from Bioconductor (Gentleman et al., 2004) - are compared using the GSE5222 dataset and the corresponding Ebased CDF. This is achieved using the `customCDF` R package (Delhomme et al., submitted). This package is under review to be added to Bioconductor. The normalized results obtained using the three different methods are then compared pair-wise.

Creating the experimental designs: Creating a proper experimental design is essential for the final result interpretation. For the datasets introduced previously (section 3.1.2, page 42), this is straightforward as they define at most two conditions, *e.g.* tumor vs. control, see Table 3.1.

GEO ID	Condition 1	Condition 2	Experimental Design
GSE5222	Tumor	Control	Tumor - Control
GSE3791	Stem Cell	-	-
GSE29683	Tumor	-	-
GSE29684	Tumor Single Cell	-	-
GSE14359	Tumor	Control	Tumor - Control
GSE14359	Metastasis	Control	Metastasis - Control
GSE14359	Tumor	Metastasis	Tumor - Metastasis
GSE14827	Tumor with or without metastasis	-	-
GSE16088	Tumor	Control	Tumor - Control
GSE16091	Tumor	-	-

Table 3.1 – GEO datasets original experimental design extracted from the GEO entries and relevant publications.

Differential expression analyses: To validate the Ebased CDF, a set of DE analyses were performed, first using the GSE16088 dataset and its original experimental design (see Table 3.1), then using the GSE29683 tumor and GSE5222 control data in the same "Tumor-Control" experimental design. Every mentioned EP dataset was normalized independently in R using the `svn` package (Huber et al., 2002) encapsulated in the `customCDF` package (Delhomme et al., submitted). Then for both DE analyses, the respective "within-array" normalized datasets are normalized again to correct for "between-array" effects using the R `limma` package. Finally, using the linear models approach implemented in the same package, the DE values are

calculated. The obtained p-values are adjusted for multiple testing using the Benjamini and Hochberg (1995) correction and - unless specified otherwise in the results (chapter 4, page 67) - only probe-sets with p-values lower than 0.05 and an absolute **log2** fold change higher than 2 are returned, *i.e.* fold changes smaller than 0.25 or larger than 4.

3.2.4 matrixCGH

Wet laboratory methods: The sample preparation, BAC-clone selection, DNA-extraction and the matrixCGH microarray preparation, labelling and hybridization are described in Zielinski et al. (2005). The dataset was hybridized in two separate batches using a different microarray layout, see Table 3.2.

Chip ID	Sample ID	Sample Sex	Control Sample	Control Sex	Batch
560	M22058	F	male pool	M	2
561	M22590	F	male pool	M	2
562	M22860	F	male pool	M	2
563	M23869	M	female pool	F	2
564	M24733	M	female pool	F	2
565	M24820	M	female pool	F	2
566	M20517	M	female pool	F	1
567	M22067	M	female pool	F	1
568	M22233	M	female pool	F	1
569	M22641	M	female pool	F	1
570	M22731	M	female pool	F	1
571	M23209	M	female pool	F	1
572	M23215	M	female pool	F	1
573	M24430	M	female pool	F	1
574	M24794	M	female pool	F	1
575	M22808	F	male pool	M	2
576	M23449	F	male pool	M	2

Table 3.2 – Details of the two batches of the (Zielinski et al., 2005) dataset. Note that the samples are matched to an opposite sex control pool (of healthy donors' blood).

***In-silico* methods:** The data-acquisition is described in Zielinski et al. (2005), however, the processing and analysis - although following the same procedure - are performed anew to use more recent tools and annotations.

Probe filtering: In addition to the three filters mentioned in Zielinski et al. (2005): *Mean to Median*, *Signal to Noise* and *Replicate SD*, probes the

intensities of which are not above a minimal value were filtered out. This *Minimal Signal* filter is introduced to remove probes which signal was too close (≤ 1.2 times) to that of the whole array noise level to be likely to be informative. This filter is a global signal-to-noise filter in comparison to the above *Signal to Noise* that assesses local signal variations.

Segmentation algorithm: Numerous methods for segmenting array-CGH data - identifying loci showing CNV - have been developed after the original analysis performed by Zielinski et al. (2005), using:

- an HMM approach (Fridlyand et al., 2004)
- a non-parametric change-point method (DNAcopy) (Olshen et al., 2004; Venkatraman and Olshen, 2007)
- a Gaussian model-based approach (GLAD) (Hupé et al., 2004)
- by building hierarchical clustering-style trees along each chromosome (CLAC) (Wang et al., 2005)
- a penalized likelihood criterion to estimate breakpoints (Picard et al., 2005)
- an expectation/maximization-based method (Myers et al., 2004)
- a Bayesian model that uses parameterized prior distributions and a prior-less maximum a posteriori (MAP) technique to estimate the underlying model (Daruwala et al., 2004)
- a wavelet approach (Hsu et al., 2005)
- *etc.*

These methods were assessed inhouse and based on published benchmarking (Willenbrock and Fridlyand, 2005; Lai et al., 2005; Lehmussola et al., 2006), GLAD was deemed the most appropriate for the Zielinski et al. (2005) dataset. A modified implementation, optimized for the inhouse spotted array: *Alterations* (Toedt et al., a, unpublished) available in ChipYard (Toedt et al., b, unpublished) is used.

Missing value imputation: For discretized CNV data - the results of arrayCGH data segmentation - missing values that impede the statistical power of the downstream analyses were imputed whenever possible. Per chromosome, a sliding window involving 5 probes was used to calculate a median CNV value. This value was associated a confidence score - from 0 to 1 - based on the probes' consistency within the window and the quantity of missing values to infer. If that confidence score was above a selected threshold of 0.8, the missing values were imputed.

Breakpoint detection: As for the segmentation algorithm (see paragraph 3.2.4, page 47), many tools had been published since the original analysis. The same procedure led to the selection of GLAD for the the breapoint detection within the Zielinski et al. (2005) dataset. As previously its ChipYard (Toedt et al., b, unpublished) implementation was used: *Alterations* (Toedt et al., a, unpublished).

3.2.5 Integrative analysis

The term integrative analysis describes analysis that combines different kind of data to increase its detection power. In the following, the different integrative analyses performed in this thesis work are listed.

Multidimensional scaling: The R package `MASS` was used to perform a **Multidimensional Scaling** (MDS) on the Zielinski et al. (2005) arrayCGH dataset. The results were analyzed in conjunction with the clinical data.

Hierarchical clustering: Using the clinical data, the arrayCGH samples from the Zielinski et al. (2005) dataset were subjected to a hierarchical clustering using the R package `hclust`. The `euclidean` distance was used as the metric and the `ward` method was used for clustering.

3.3 Microarray simulation

To assess the validity of newly developed techniques, it is important to be able to assess their sensitivity - to estimate the *type II error* rate, *i.e.* to count how many of the true positive features are identified - and specificity - to estimate the *type I error* rate, *i.e.* to count how many true negative features are identified. For doing so, knowing the expected outcome is essential and is achieved through artificially generating data. As no simulator was available that implemented the necessary assumptions about the arrayCGH and EP data, I created one bundled in a R package: `aSim`.

3.3.1 Simulation workflow

The simulation workflow is described in Figure 3.1. The definition of the simulation parameters can be done by the user (manual mode) or automatically. In either modes, the simulation requires groups of probes/features to be defined. The actual grouping method depends on the microarray type. For matrixCGH, physically linked regions are used, while for expression profiling clusters of probes representing genes, usually independent of physical linkage are defined. A feature can only be assigned to a single group. In the automatic mode, the groups are created as follows (Figure 3.1, step a-d):

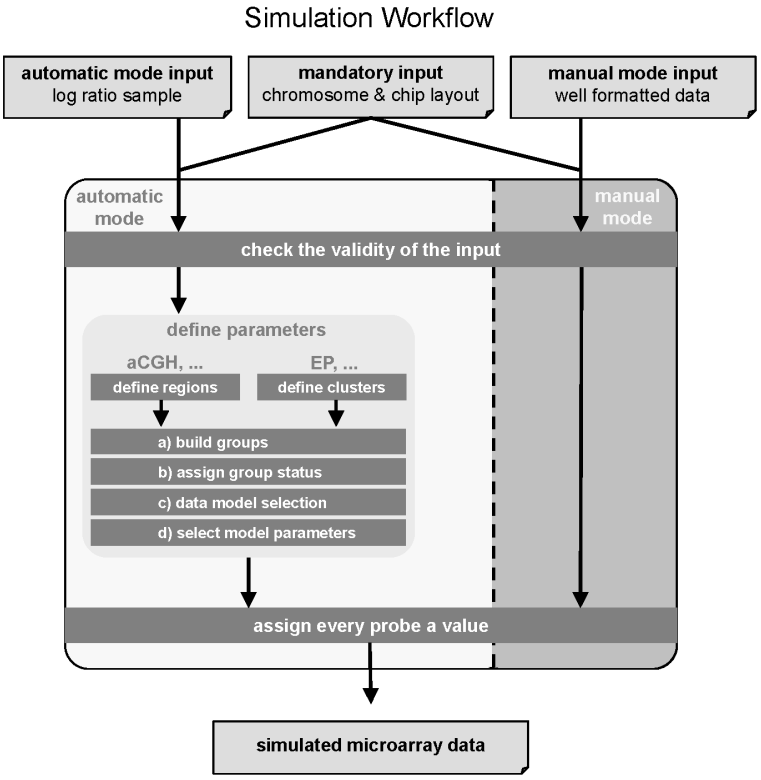


Figure 3.1 – The diagram of the `aSim` workflow.

- Let $G = \{g_1, g_2, \ldots, g_k\}$ be the set of all defined groups
- Let $F = \{f_1, f_2, \ldots, f_n\}$ be the set of all features to be assigned to a group

Every group g_i is defined by:

1. a model M_i, of the data generating process defined by model parameters p_i - *e.g.* for a Gaussian model, p_i is given by:

$$p_i = (\mu_i, \sigma_i^2)$$

   where μ_i is the mean and σ_i^2 is the variance - which is used to model regulatory effects (Figure 3.1, step c-d).

2. a state indicator (Figure 3.1, step b)

$$s_i \in \{-1, 0, 1\}$$

used to define the category of the regulatory effect observed in the data: changes in DNA copy numbers (arrayCGH) or gene expression levels (expression profiling).

- a value of 1 is used for a DNA copy number gain or "high-expressed" gene.
- a value of -1 represents a DNA copy number loss or "low-expressed" gene.
- a value of 0 is synonymous with balanced DNA copy numbers or "average gene expression". Note that this state, albeit discrete, is **Without Loss Of Generality** (WLOG) as it just indicates the direction of the effect observed in the group g_i; the actual parameters of this effect (i.e. its strength, its frequency, etc) are entirely described by the model M_i.

3. a set of features:

$$f_i = f_{i,1}, \ldots, f_{i,j}, \text{ subset of } F$$

Every feature $f_{i,j}$ of a set f_i is assigned a generated data value $v_{i,j}$ drawn from the model M_i distribution and modified by the state indicator s_i:

$$v_{i,j} = \begin{cases} \prod m_{i,j}, s_i & \forall si \in \{-1, 1\} \\ m_{i,j} & \text{otherwise} \end{cases} \tag{3.1}$$

where $m_{i,j}$ is the value drawn from the model M_i distribution.

The output of the simulator contains the list of groups G, describing the data generation parameters used during the simulation and the list of all features F with their associated values. These are the simulated data and a representation of the mixture of the models described by the list G - **mixture model**s have been proposed in the literature to describe microarray data (Ghosh and Chinnaiyan, 2002; Hoyle et al., 2002). In the context of benchmarking data analysis algorithms, the G list describes the reference to which the algorithms outcome has to be compared.

`aSim` can simulate any type of microarray data, which can be abstracted in the described way. For arrayCGH and expression profiling, the simulator can be used in an automatic fashion, given a less detailed input. To achieve this, the necessary simulation parameters were extracted from five datasets of diverse origin. The datasets, parameter extraction and automatic simulation are the topic of the three next sections.

3.3.2 Datasets description

Five datasets, totaling 225 microarrays have been used to develop the `aSim` simulator. Table 3.3 describes these datasets.

Dataset	number of arrays	type	description	availability
Mendrzyk et al. (2006)	88	arrayCGH	ependymoma	GEO GSE3435
Snijders et al. (2001)	15	arrayCGH	12 fibroblast, 2 chorionic villus and 1 lymphoblast cell strains	`GLAD` R package
Sültmann et al. (2005)	74	EP	kidney cancer	`kidpack` R package
Thuerigen et al. (2006)	46	EP	primary breast cancer	GEO GSE4056
Veltman et al. (2003)	2	arrayCGH	bladder cancer	`GLAD` R package

Table 3.3 – Publicly available datasets used for developing the `aSim` simulator.

3.3.3 Parameter extraction

Expression profiling: For simulating microarray data, the following parameters have to be extracted for every group of probes: status, distribution model, expected value and standard deviation. For the expression profiling data, the cluster set G is created by grouping the original probes using a *k-means* clustering. Depending on the total number of probes, the number of clusters is adapted to generate probes group sizes, which are in the range of the KEGG (Kanehisa et al., 2012) pathway size (min = 1, max = 299, median = 31). For the kidney dataset (Sültmann et al., 2005) G consists of 50 clusters, whereas for the breast dataset (Thuerigen et al., 2006) it consists of 300 clusters. This generates a distribution of the probe group sizes similar to the KEGG one (minimum = 2, maximum = 64, median=33). For every g_i, the mean and standard deviation of their members' values is calculated. Their state - *i.e.* "average", "low" or "high-expressed" - is determined according to their mean value as in:

$$s_i = \begin{cases} 0 & \text{if } \bar{\mu} - 2\bar{\sigma} < \mu_i < \bar{\mu} + 2\bar{\sigma} \\ 1 & \text{if } \mu_i \geq \bar{\mu} + 2\bar{\sigma} \\ -1 & \text{if } \mu_i \leq \bar{\mu} - 2\bar{\sigma} \end{cases}$$

where μ_i is the mean of the cluster members' values, $\bar{\mu}$ and $\bar{\sigma}$ the mean and sd of all the values, respectively.

arrayCGH: For the arrayCGH data, GLAD - a data segmentation algorithm that detects chromosomal breakpoints and assigns a genomic status

to every identified chromosomal region - is used and the set of regions G is extracted from its results. For every region g_i, the mean and standard deviation of their members values is calculated and the state derived from these results. Finally, the last parameter set is the distribution model. The data model that would best represent EP and arrayCGH data is highly debated in the community and no consensus has been reached so far (Hoyle et al., 2002; Huber et al., 2002; Li and Yang, 2002; Huber et al., 2003; Steinhoff and Vingron, 2006). Hence, the commonly accepted standard were chosen:

- the Gaussian distribution for the arrayCGH features located in balanced regions and for the EP probes showing no change in expression.
- the log-normal distribution to represent the probes located in the arrayCGH aberrant regions and the differentially expressed EP clusters.

3.3.4 Automatized simulation

These are based on arrayCGH and EP experimental data. First, the key parameters for the simulation are extracted, then the groups are defined and finally the simulation is performed. The following description is based on the default simulator models: **Gaussian** and **log-normal**. Additional distribution models can be integrated in the simulator as additional modules and they do not need to rely on the parameters (i.e. offset and shape) described and used below.

In the first step of the automatic process, the standard deviation to be used for the groups is determined. This is done by calculating the **Median Absolute Deviation** (MAD) of experimental values issued from balanced regions or having an "average" gene expression. The mad estimator is preferred here for its outliers robustness. The calculated value mad_g, describes the overall array variance to be simulated.

The next step is the group definition. For arrayCGH, a group represents a genomic aberration with a defined chromosomal start and end position. For expression profiling, a group defines a probe cluster. Its members represent genes, which can be located on different chromosomes (i.e. a subset of a gene pathway or all genes controlled by a specific transcription factor). These two group-building processes will be detailed in the next two paragraphs.

As described above, every group gets assigned a state s_i and a distribution model M_i, with its offset (e.g. mean) and its shape (e.g. standard deviation). For a group state of 0 (no aberration/no change in expression), the default distribution is *Gaussian*, the offset is set to 0, and the standard deviation is given by the estimate mad_g. For a group state $\neq 0$, the default distribution is *log-normal*, the standard deviation is mad_g and the offset is set to a default offset *od*, equal to 0.46 for arrayCGH and 3.1 for expression profiling. This default value *od* was determined by the analysis of the 225

microarrays described previously in section 3.3.2, page 50. The last step consists of the value assignment for every probe as described in equation (3.1).

arrayCGH groups: These are defined by chromosomal start and end positions. Genomic imbalances identified by arrayCGH are physically linked regions; *i.e.* features are grouped by chromosomal regions having the same status. Therefore $G = g_1, g_2, \ldots, g_K$ is the set of all regions, aberrant or not. It is assumed that the number of chromosomal aberrations for every chromosome arm is defined by a *Poisson* distribution $Po(\lambda = 0.1)$.

$$\text{Let } N = \{n_1, n_2, \ldots, n_s\}$$

be the resulting set of chromosomal arms aberration counts.

$$\text{Let } \forall n_j, Aj = \{a_{n,1}\,, a_{n,2}\,, \ldots, a_{n,j}\,\}$$

be the set of aberrations on a chromosome arm. Every aberration a_j, is defined by:

1. its state $s_j \in \{-1, 1\}$
2. its length l_j in base pairs (bp)
3. its model M_j (as described in equation (3.1))

In cancer, an aberration can vary in size between one bp and the length of a complete chromosome arm. This is modeled by a log-normal distribution with the default parameters $\mu = 15.2$ and $\sigma^2 = 2.12$, therefore:

$$l_j \approx log(\mu = 15.2, \sigma^2 = 4.41)$$

The set of aberration A_j is distributed randomly on its respective chromosome arm and every a_j receives a chromosomal locus. A_j is a subset of G, for the aberrant regions of a given chromosome arm. Remaining loci on this chromosome arm not covered by any aberration are determined. Their occurrence $o_j \in \{0, \ldots, n_j + 1\}$, depends on the locations and sizes of A_j.

$$\text{Let } B_j = \{b_1, b_2, \ldots, b_{n,j}\,\} \text{ for } o_j \neq 0$$

be the set of normal regions. Every b_j has its length derived and is assigned a state value of 0 and a model as described above. G is the union of all the A_j and B_j.

EP groups: The EP groups are probe clusters. Expression levels identified by expression profiling do not need to have any physical linkage, but features having similar expression patterns - *e.g.* features identifying genes regulated by a given transcription factor - may be grouped in clusters. Therefore $G = \{g_1, g_2, \ldots, g_K\}$ is the set of all these clusters. Each cluster g_i is defined by:

1. a state $s_i \in \{-1, 0, 1\}$
2. a set of member features $F_i = \{f_{i,1}, \ldots, f_{i,J}\}$
3. a model M_i (as described in equation (3.1)).

For every cluster $g_i \forall i \in \{1, \ldots, n-1\}$, the amount of member features m_i is determined from an hyperbolic distribution which approximates the observed pathway sizes from KEGG (Kanehisa et al., 2012) -pathway size: min = 1, max = 299, median = 31 - as in:

$$m_i \approx H(\pi = 8, \zeta = 1, \delta = 1, \mu = 1)$$

using the first parameterization of Barndorff-Nielsen and Blaesild (1983). The m_i features are then drawn from F to build f_i. The state s_i is determined from two *Bernoulli* distributions. In the first one, p - the success probability - is the probability of the cluster to show a high or low expression. The second one, determines the direction of that change:

$$si = Bernoulli(p = 0.06) \times ((Bernoulli(p = 0.5) \times 2) - 1)$$

3.3.5 Performance

The simulator performances were assessed by generating a series of microarrays of different sizes from $1,000$ to $100,000$ features.

3.4 Microarray integrative analyses

3.4.1 Selected datasets

For demonstrating the advantages of integrative analyses, the arrayCGH and EP obtained from the Zielinski et al. (2005); Grasemann et al. (2005); Gratias et al. (2007) studies are used.

Different pre-processing steps need to be performed on the EP and arrayCGH data to render them comparable for performing an integrative analysis:

1. combining the EP control samples.
2. defining the EP expression states.

3. rescuing the arrayCGH sex probes.

4. imputing the arrayCGH missing ratios.

5. defining and imputing arrayCGH virtual probes.

Combining the control samples: The EP dataset originally had only one control sample. Two additional samples were hybridized (see section 3.1.1, page 41), the *Clontech* one twice, as technical replicates. The *GeneChip*® HG-U133A being no more available - the one used for the Grasemann et al. (2005); Gratias et al. (2007) studies - the HG-U133Plus2 *GeneChip*® was used instead. To compare the original control sample with the 3 new ones, the probe-sets correspondance between the two *GeneChip*®s had to be determined. The expression values between the samples hybridized on both platforms were then compared pair-wise using the Pearson correlation.

Defining the expression states: The DE of the EP data was calculated as described previously (see section 3.2.3, page 44), using all 4 control samples defined in the former paragraph. The threshold for significance was set to an adjusted p-value of $1e^{-4}$. Probe-sets with a negative significant log2 Fold Change (log2 FC) were attributed a -1 state, whereas those with a positive one were attributed a 1 state. The non-significant probe-sets were given a default 0 state. This state matrix (probe-sets $\times$ samples) was then refined per probe-sets. First, *Z-scores* were calculated per probe-set:

$$z = \frac{x - \mu}{\sigma}$$

with x the log2 FC values for that probe-set, μ the fitted log2 FC value obtained from the DE analysis and finally σ the standard deviation of these probe-set values. These *Z-scores* were used to refine the state of the probe-set within every sample. If its absolute value was smaller than 1.85, no changes were applied. If its value was more extreme, then a value of 1 was subtracted or added for the lower (negative *Z-score*) or higher (positive *Z-score*) extremes, respectively.

Rescuing the sex probes: As the arrayCGH microarrays used had an opposite sex matched design (*i.e.* the microarrays used were dual-channel), the CNV state and log-ratio for the sexual chromosome had to be corrected. For male samples, there should have been a $1X$ loss of the X chromosome and a $1X$ gain of the Y chromosome. The situation was reciprocal for the female. As a consequence the median log2 FC of the reported $1X$ gain or loss was calculated and this value subtracted from the Y chromosome and added to the X chromosomes' probes for male samples and vice-versa for female ones.

Imputing the missing log-ratio: As the arrayCGH experiment used spotted microarrays, missing log-ratio values had to be imputed (as previously described for the states, see the related paragraph in section 3.2.4, page 47). In the present case, the state value (*i.e.* discretized CNV) of the upstream and downstream windows surrounding the value(s) to be imputed were compared. If they were equal (*i.e.* being in a region of common CNV status), then the missing values were imputed using the `aSim` simulator. The necessary simulation parameters (*i.e.* distribution, mean, sd) were "Gaussian" and the median and MAD of the log ratio values calculated for every possible state value.

Defining virtual clones and imputing their values: As the arrayCGH spotted microarray probes achieved only a partial coverage of the genome, virtual probes lying in between actual probes or chromosome physical boundaries (centromere, telomere, heterochromatin regions) were defined. Their values was imputed as described above for the missing values and previously in section 3.2.4, page 47.

3.4.2 Integrative analysis workflow

The following paragraphs describe the different steps performed to compare the two datasets.

Overlay definition: Based on the genomic annotation of the arrayCGH BAC-clones and of the EP probes (contained in the CDF file), the overlap between these two sets of probe was calculated. Probes were paired when their genomic loci overlapped; this information was then stored in an overlay structure on top of both datasets. The overlay was minimized so that it contained the smallest number of the largest possible regions.

Summarization and visualization: Using the overlay, a combined dataset was defined that merged the arrayCGH and EP data; *e.g.* for the arrayCGH discrete data, we obtained a matrix $m \times n$, with m the set of non-overlapping loci and n the samples. The dataset consisted of 4 such matrices, 2 per array kind, one for the discrete and one for the continous values. An additional matrix was generated by combining the discrete value matrices; it summarized the co-occurence of the arrayCGH and EP states and had therefore a size of $m \times n^2$. This matrix was used to calculate a **contingency table** and the R `vcd` package was used to produce an assocation plot (suggested by Cohen (1980) and extended by Friendly (1992)) indicating deviations from the expected independence model, in the present case the residuals were the signed contribution to the Pearson's χ^2.

Correlation: Five methods, as presented in section 1.2.2, page 30 have been implemented:

1. Eta
2. Pearson
3. Spearman
4. Weight
5. Welch

Each of these statistic were calculated per probe-set/gene (*i.e.* the matrix row).

Eta: η^2 is the percent of variance in the dependent variable (the continuous data) explained linearly or nonlinearly by the independent variable (the discrete data, *i.e.* the states). The formula to calculate η^2 is:

$$\eta^2 = \frac{\sum_{i=1}^{n} n_i(\bar{y}_i - \bar{y})^2}{\sum_{i=1}^{n}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2}, \eta^2 \in [0,1] \tag{3.2}$$

where n describes the number of members of every discrete data class (*i.e.* there are five classes: $N = [-2,-1,0,1,2]$ in the discrete data), $\bar{y}_i$ is the mean of the continous value of the class n_i, $\bar{y}$ is the overall mean and y_{ij} are the single continous value of the class n_i.

Pearson: The Pearsons's r statistic based on a sample of paired data (X_i, Y_i) is:

$$r = \frac{1}{n-1}\sum_{i=1} n(\frac{X_i - \bar{X}}{\sigma_x})(\frac{Y_i - \bar{Y}}{\sigma_y}) \tag{3.3}$$

where $\bar{X}$ and σ_x are the mean and standard deviation, respectively.

$$\frac{X_i - \bar{X}}{\sigma_x}$$

is actually the standard score (*i.e.* the *Z-score*).

Spearman: The Spearman's ρ statistic was calculated by the formula:

$$\rho = 1 - \frac{6\sum D^2}{N(N^2-1)}, \rho \in [-1,1] \tag{3.4}$$

where D is the difference between the ranks of the arrayCGH and EP ratios and N is the number of pairs of values.

Weight: w was modified from Hyman et al. (2002), where they only considered positive CNV, *i.e.* gain of chromosomal material. The original equation:

$$w = \frac{\mu_g - \mu_n}{\sigma_g + \sigma_n}$$

where μ_g and σ_g are the mean and **Standard Deviation** (SD) of the EP log2 FC from amplified regions and μ_n, σ_n their counterpart for normal regions, was extended to:

$$w = \frac{|\mu_g - \mu_n| + |\mu_l - \mu_n|}{\sigma_g + \sigma_l + \sigma_n}$$

where μ_l, σ_l are representative of the EP log2 FC of the lost regions. The numerator represents the distance between every class and can be further abstracted to any number of classes n by the formula:

$$w = \frac{\sum_{i \neq j} |\bar{y_i} - \bar{y_j}|}{n - 1 \sum_i \sigma_i}, w \geq 0 \tag{3.5}$$

for $i, j = 1, \ldots, n$.

Welch: v^2 is a modified F-test for unequal variances (Welch, 1951). It is calculated by:

$$v^2 = \frac{\sum_{i=1}^{k} w_i \frac{y_i - \hat{y}^2}{k-1}}{1 + \frac{2(k-2)}{k^2-1} \sum_{i=1}^{k} \frac{1}{f_i}(1 - \frac{w_i}{\sum w_i})^2} \tag{3.6}$$

where $y_i, i \in [1, \ldots, k]$ are statistical quantities *i.i.d* and f_i the number of degree of freedom:

$$\hat{f}_1 = (k - 1)$$

$$\hat{f}_2 = [\frac{3}{(k^2 - 1)} \sum_{i=1}^{k} \frac{1}{f_i}(1 - \frac{w_i}{\sum w_i})^2]^{-1}$$

Since the weight used in Welch statistic is $w_i = 1/\lambda_i \sigma_i^2$, where λ_i is a constant, one cannot compute the statistic if any one group has a zero standard deviation. Moreover, sample sizes of all groups have to be greater than or equal to zero.

Extreme cases correction: As just described for Welch, the mentioned algorithms have a certain number of limitations, for which corrections can be applied. The evaluated methods were:

1. *compact*

$$c = \frac{C}{N} + \frac{D}{N} \times \sum_{i=1}^{d} |D_{i,1} - D_{i,2}|), c \in [0,1] \tag{3.7}$$

where N is the total number of arrayCGH - EP pairs, C, D the number of concordant and discordant pairs and $D_{i-1} - D_{i-2}$ the distance of the discordant pairs.

2. *gamma* (Goodman and Kruskal, 1972)

$$\gamma = \frac{C - D}{C + D}, \gamma \in [-1,1] \tag{3.8}$$

where C is the number of concordant arrayCGH - EP pairs and D the number of discordant ones.

3. *kappa* is modified from Cohen (1960).

$$\kappa_{new} = \frac{B - C}{1 - C}, \kappa_{new} \in [0,1] \tag{3.9}$$

where B and C are respectively the sum and sum of squares of the arrayCGH - EP pairs contingency matrix diagonal.

4. *none* no correction applied.

5. *percent*

$$p = \frac{C}{N}, p \in [0,1] \tag{3.10}$$

where C and N are as described for equation (3.7).

Not all corrections make sense for every method, *i.e.* Pearson and Spearman do not need any, while Welch can not be corrected by either gamma or kappa.

Benchmarking: To benchmark the five available algorithm families, a dataset of diverse arrayCGH and EP experiments was simulated using the package `aSim` (introduced in section 3.3, page 48). The original parameters used are described in Table 3.4. To be able to create ROC curves and measure the **Area Under the Curve** (AUC), the specificity and sensitivity of every method was measured on a dataset series where the noise was increased stepwise, *i.e.* the σ value was increased and a new data dataset was generated, the FPR and TPR recorded and so on until the different method results converged. By that time, the simulated data had no biological characterisctics anymore. To visualize the effect of noise increase, the AUC was recorded for every method and every condition and plotted against *delta*: the inverse of the noise.

kind	model	μ	σ
arrayCGH gain	log-normal	0.40	0.15
arrayCGH loss	log-normal	-0.48	0.15
arrayCGH no change	normal	0.016	0.08
EP diff. exp.	log-normal	1.24	0.99
EP no exp. change	normal	0.002	0.16

Table 3.4 – `aSim` parameters used to simulate the first benchmarking dataset.

Gene annotation: Candidate gene annotation were retrieved from http://www.genecards.org (Rebhan et al., 1997).

3.5 Gene Ontology analyses

GO term enrichment analysis were performed using `Ontologizer` (Bauer et al., 2008). Enrichment were measured using the *Parent-Child* method (Grossmann et al., 2007) and corrected for multiple testing using the Benjamini-Hochberg correction (Benjamini and Hochberg, 1995). An adjusted p-value cutoff of 10% was used for the integrative analysis analysis and of 1% for the comparative analysis.

The ontologies and gene annotations for *Homo sapiens* were automatically retrieved by `Ontologizer` from the www.geneontology.org website.

For the enrichment analysis, the population was defined as all the genes present on the respective microarray, *e.g.* 12,599 genes were used when the data originated from an experiment performed on Affymetrix *GeneChip*® HG-U133A.

All GO evidence codes were used when the study sets were of small sizes ($\leq$ 200 genes), otherwise the **IEA** (Inferred from Electronic Annotation) code were filtered out, as they are of lower quality (see Table 1 in Rhee et al. (2008))

3.6 Microarray comparative analysis

The comparative analysis performed between retinoblastoma and osteosarcoma made use of the same statistical approaches described previously for the microarray integrative analysis, see section 3.4, page 54 with a few modifications described below.

3.6.1 Sample selection

As the samples were not matched, *e.g.* primary retinoblastoma and related osteosarcoma metastases, samples were selected from the Mcevoy et al. (2011) and Kobayashi et al. (2010) studies (see section 3.1.2, page 42). Both were hybridized to Affymetrix *GeneChip*® HG-U133Plus2, which has twice more probe-sets ($50,000$) than the HG-U133A. After performing the QA, 15 samples from both studies, the most similar ones, were selected and assigned an arbitrary common sample name. Since both these studies do not have a proper set of controls, the appropriate samples from the **GSE5350** dataset (MAQC Consortium et al., 2006) were selected.

3.6.2 Workflow modifications

Due to the fact that every sample originated from EP studies - unlike the integrative analysis performed previously - two steps of the previously detailed workflow had to be adapted.

Defining the expression states: First, the expression states were calculated for both datasets - **GSE29683** (Mcevoy et al., 2011) and **GSE14827** (Kobayashi et al., 2010) - as described in paragraph 3.4.1, page 55.

Defining the overlay: Second, since both datasets share the same platform, there was no need to create an overlay and the data was directly subjected to the **Weight** correlation with a **percent** correction, see equations (3.5) and (3.10).

The rest of the analyses is similar to that described in sections 3.4 and 3.5

Bibliography

Ole Barndorff-Nielsen and Preben Blaesild. Reproductive exponential families. *The Annals of Statistics*, 11(3):770–782, 1983.

Tanya Barrett et al. Ncbi geo: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Research*, 35(Database issue): D760–5, Jan 2007. doi: 10.1093/nar/gkl887.

Sebastian Bauer et al. Ontologizer 2.0–a multifunctional tool for go term enrichment analysis and data exploration. *Bioinformatics*, 24(14):1650–1, Jul 2008.

Yoav Benjamini and Yosef Hochberg. 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, Jan 1995.

Ayala Cohen. On the graphical display of the significant components in a two-way contingency table. *Communications in Statistics-Theory and Methods*, A9:1025–1041, 1980.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Manhong Dai et al. Evolving gene/transcript definitions significantly alter the interpretation of genechip data. *Nucleic Acids Research*, 33(20):e175, Jan 2005. doi: 10.1093/nar/gni179.

Raoul-Sam Daruwala et al. A versatile statistical analysis algorithm to detect genome copy number variation. *Proc Natl Acad Sci USA*, 101(46): 16292–7, Nov 2004.

Nicolas Delhomme et al. Ensembl based custom definition file for affymetrix genechip. *submitted*.

Francesco Ferrari et al. Novel definition files for human genechips based on geneannot. *BMC Bioinformatics*, 8:446, Jan 2007. doi: 10.1186/1471-2105-8-446.

Paul Flicek et al. Ensembl 2011. *Nucleic Acids Research*, 39(Database issue): D800–6, Jan 2011. doi: 10.1093/nar/gkq1064.

Jane Fridlyand et al. Hidden markov models approach to the analysis of array cgh data. *Journal of multivariate analysis*, 90(1):132–153, 2004.

Michaal Friendly. Graphical methods for categorical data. *User Group International Conference Proceedings*, 17:190–200, 1992.

Raphaela Fritsche-Guenther et al. De novo expression of epha2 in osteosarcoma modulates activation of the mitogenic signalling pathway. *Histopathology*, 57(6):836–50, Dec 2010. doi: 10.1111/j.1365-2559.2010.03713.x.

Laurent Gautier et al. Alternative mapping of probes to genes for affymetrix chips. *BMC Bioinformatics*, 5:111, Aug 2004. doi: 10.1186/1471-2105-5-111.

Robert C Gentleman et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology 2010 11:202*, 5(10):R80, Jan 2004.

Debashis Ghosh and Arul M Chinnaiyan. Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, 18(2):275–86, Feb 2002.

Leo Goodman and William Kruskal. Measures of association for cross classifications, iv: Simplification of asymptotic variances. *Journal of the American Statistical Association*, 67(338):415–421, Jun 1972.

Corinna Grasemann et al. Gains and overexpression identify dek and e2f3 as targets of chromosome 6p gains in retinoblastoma. *Oncogene*, 24(42): 6441–9, Sep 2005. doi: 10.1038/sj.onc.1208792.

Sandrine Gratias et al. Allelic loss in a minimal region on chromosome 16q24 is associated with vitreous seeding of retinoblastoma. *Cancer Res*, 67(1): 408–16, Jan 2007. doi: 10.1158/0008-5472.CAN-06-1317.

Steffen Grossmann et al. Improved detection of overrepresentation of gene-ontology annotations with parent child analysis. *Bioinformatics*, 23(22): 3024–31, Nov 2007.

David C Hoyle et al. Making sense of microarray data distributions. *Bioinformatics*, 18(4):576–84, Apr 2002.

Li Hsu et al. Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, 6(2):211–26, Apr 2005.

Wolfgang Huber et al. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:S96–104, Jan 2002.

Wolfgang Huber et al. Parameter estimation for the calibration and variance stabilization of microarray data. *Stat Appl Genet Mol Biol*, 2:Article3, Jan 2003.

Philippe Hupé et al. Analysis of array cgh data: from signal ratio to gain and loss of dna regions. *Bioinformatics*, 20(18):3413–22, Dec 2004.

Elizabeth Hyman et al. Impact of dna amplification on gene expression patterns in breast cancer. *Cancer Res*, 62(21):6240–5, Nov 2002.

Rafael A Irizarry et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–64, Apr 2003. doi: 10.1093/biostatistics/4.2.249.

Minoru Kanehisa et al. Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, 40(Database issue):D109–14, Jan 2012.

Audrey Kauffmann, Robert Gentleman, and Wolfgang Huber. arrayqualitymetrics–a bioconductor package for quality assessment of microarray data. *Bioinformatics*, 25(3):415–6, Feb 2009. doi: 10.1093/bioinformatics/btn647.

Eisuke Kobayashi et al. Reduced argininosuccinate synthetase is a predictive biomarker for the development of pulmonary metastasis in patients with osteosarcoma. *Mol Cancer Ther*, 9(3):535–44, Mar 2010.

Felix Kokocinski, Gunnar Wrobel, et al. Quicklims: facilitating the data management for dna-microarray fabrication. *Bioinformatics*, 19(2):283–4, Jan 2003.

Felix Kokocinski, Nicolas Delhomme, et al. Fact–a framework for the functional interpretation of high-throughput experiments. *BMC Bioinformatics*, 6:161, Jan 2005.

Weil R Lai et al. Comparative analysis of algorithms for identifying amplifications and deletions in array cgh data. *Bioinformatics*, 21(19):3763–70, Oct 2005.

Antti Lehmussola et al. Evaluating the performance of microarray segmentation algorithms. *Bioinformatics*, 22(23):2910–7, Dec 2006.

Wentian Li and Yaning Yang. Zipf's law in importance of genes for cancer classification using microarray data. *J Theor Biol*, 219(4):539–51, Dec 2002.

David J Lockhart et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, 14(13):1675–80, Dec 1996. doi: 10.1038/nbt1296-1675.

Jun Lu et al. Transcript-based redefinition of grouped oligonucleotide probe sets using aceview: high-resolution annotation for microarrays. *BMC Bioinformatics*, 8:108, Jan 2007. doi: 10.1186/1471-2105-8-108.

MAQC Consortium et al. The microarray quality control (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*, 24(9):1151–61, Sep 2006.

Justina Mcevoy, Jacqueline Flores-Otero, et al. Coexpression of normally incompatible developmental pathways in retinoblastoma genesis. *Cancer Cell*, 20(2):260–75, Aug 2011. doi: 10.1016/j.ccr.2011.07.005.

Frank Mendrzyk et al. Identification of gains on 1q and epidermal growth factor receptor overexpression as independent prognostic markers in intracranial ependymoma. *Clin Cancer Res*, 12(7 Pt 1):2070–9, Apr 2006.

Chad L Myers et al. Accurate detection of aneuploidies in array cgh and gene expression microarray data. *Bioinformatics*, 20(18):3533–43, Dec 2004.

Adam B Olshen et al. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5(4):557–72, Oct 2004.

Franck Picard et al. A statistical approach for array cgh data analysis. *BMC Bioinformatics*, 6:27, Jan 2005.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

M Rebhan, V Chalifa-Caspi, J Prilusky, and D Lancet. Genecards: integrating information about genes, proteins and diseases. *Trends Genet*, 13 (4):163, Apr 1997.

Seung Yon Rhee et al. Use and misuse of the gene ontology annotations. *Nature Reviews Genetics*, 9(7):509–15, Jul 2008.

Antoine M Snijders et al. Assembly of microarrays for genome-wide measurement of dna copy number. *Nature Genetics*, 29(3):263–4, Nov 2001.

Sabina Solinas-Toldo et al. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, 20(4):399–407, Dec 1997.

Christine Steinhoff and Martin Vingron. Normalization and quantification of differential expression in gene expression microarrays. *Brief Bioinformatics*, 7(2):166–77, Jun 2006.

Holger Sültmann et al. Gene expression in kidney cancer is associated with cytogenetic abnormalities, metastasis formation, and patient survival. *Clin Cancer Res*, 11(2 Pt 1):646–55, Jan 2005.

Olaf Thuerigen et al. Gene expression signature predicting pathologic complete response with gemcitabine, epirubicin, and docetaxel in primary breast cancer. *J Clin Oncol*, 24(12):1839–45, Apr 2006.

Grischa Toedt et al. Alterations, a mixture model segmentation algorithm. *unpublished*, a.

Grischa Toedt et al. Chipyard, a framework for microarray data analysis. *unpublished*, b.

Joris A Veltman et al. Array-based comparative genomic hybridization for genome-wide screening of dna copy number in bladder tumors. *Cancer Res*, 63(11):2872–80, Jun 2003.

Ennapadam S Venkatraman and Adam B Olshen. A faster circular binary segmentation algorithm for the analysis of array cgh data. *Bioinformatics*, 23(6):657–63, Mar 2007.

Pei Wang et al. A method for calling gains and losses in array cgh data. *Biostatistics*, 6(1):45–58, Jan 2005.

Bruce L Welch. On the comparison of several mean values: An alternative approach. *Biometrika*, 38(3/4):330–336, Dec 1951.

Swen Wessendorf et al. Automated screening for genomic imbalances using matrix-based comparative genomic hybridization. *Lab Invest*, 82(1):47–60, Jan 2002.

Hanni Willenbrock and Jane Fridlyand. A comparison study: applying segmentation to array cgh data for downstream analyses. *Bioinformatics*, 21 (22):4084–91, Nov 2005.

Zhijin Wu and Rafael A Irizarry. Stochastic models inspired by hybridization theory for short oligonucleotide arrays. *J Comput Biol*, 12(6):882–93, Jan 2005. doi: 10.1089/cmb.2005.12.882.

Boris Zielinski et al. Detection of chromosomal imbalances in retinoblastoma by matrix-based comparative genomic hybridization. *Genes Chromosomes Cancer*, 43(3):294–301, Jul 2005. doi: 10.1002/gcc.20186.

Chapter 4

Results

Here, the results of the different methods introduced in the previous chapter are presented; first the EP, then the matrixCGH and finally the integrative and comparative analyses results will be described.

4.1 Expression Profiling analyses

This section describes the pre-processing results as well as the higher level analyses performed on the EP Affymetrix *GeneChip*®s for both tumors: retinoblastoma and osteosarcoma. A total of 6 datasets were retrieved from GEO: 3 retinoblastoma (GSE5222, GSE29683 and GSE29684), 2 osteosarcoma (GSE14359 and GSE14827) and a control set extracted from the GEO GSE5350 MAQC study. For the GSE5222 dataset, an additional set of 3 control samples was generated in the division of Molecular Genetics, DKFZ.

4.1.1 Quality Assurance

For every dataset, a number of microarray samples had to be removed as they did not pass the QA criteria. These, based on the results of the Bioconductor package *arrayQualityMetrics* (Kauffmann et al., 2009), were stringent to ensure the highest possible similarity of the different samples within a dataset. As the integrative and comparative approaches need to combine several of these datasets together, this is necessary to limit the effect of confounding factors. The number of samples removed are listed in the Table 4.1. For an example of a QA report, generated for the GSE5222 dataset, see appendix B (page 184).

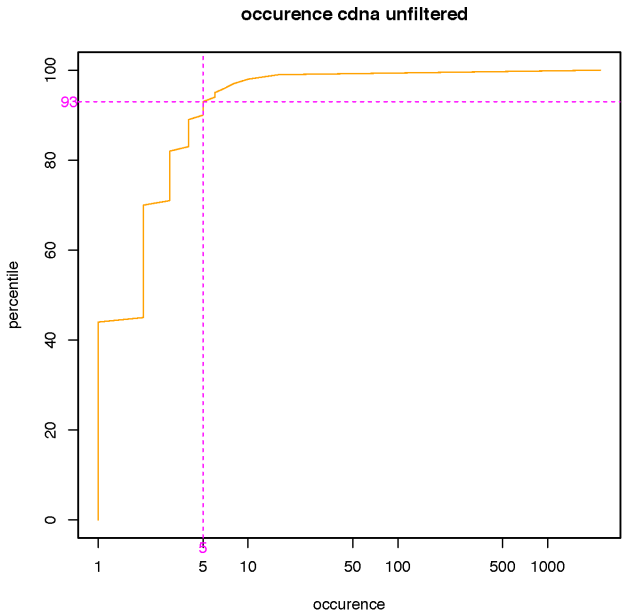


Figure 4.1 – Number of alignments per probe against the *Homo sapiens cdna (Ensembl version 53)* reference for the HG-U133 Plus2 *GeneChip*®. Note that the x axis is on a log scale.

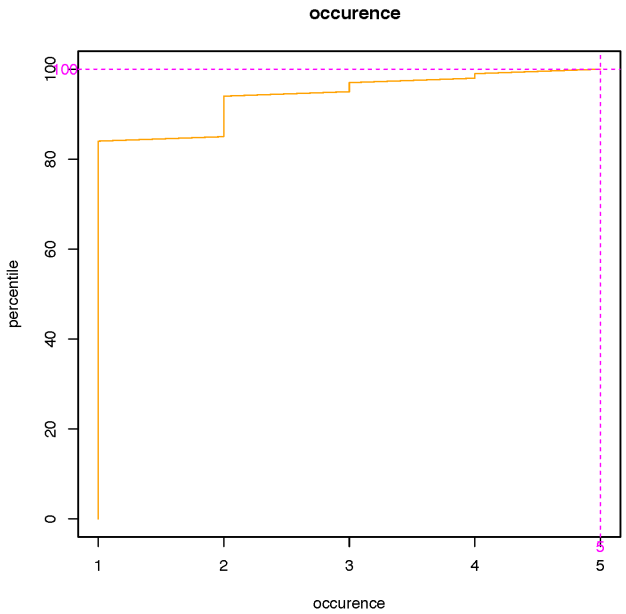


Figure 4.2 – Number of alignments per **selected** probe against the *Homo sapiens cdna and dna (Ensembl version 53)* reference for the HG-U133 Plus2 *GeneChip*®

	GSE 5222	GSE 29683	GSE 29684	GSE 14359	GSE 14827	GSE 5350
Fail	3/21	17/62	4/20	4/20	6/27	5/30
%	14.3	27.4	20	20	22.2	16.7

Table 4.1 – EP dataset QA summary

4.1.2 Probe-set annotation

To analyse *GeneChip*® microarrays, it is important to use accurate probe-set information. As described in section 3.2.2, page 43, a probe-set consists of a variable number of probes (16 on average) identifying the same gene. The quick pace of the human genome re-sequencing necessitates a frequent update of the probe-set containing information file: the **Custom Definition File** (CDF) file. As presented in the manuscript (Delhomme et al., submitted) (see Appendix C, page 197), in comparison to the other CDFs our Ebased CDF offers a higher sensitivity, is more frequently updated and uses as many probes as possible to benefit from all the possible information present on a *GeneChip*® microarray. In addition, it contains additional probe quality information, such as the number of genes mapped by a given probe-set. Applied on the frequently analyzed ALL dataset (Chiaretti et al., 2004, 2005), it unravels three new potential candidate genes, with implications in cancer already shown in other tumors. A few additional key properties of these CDFs, not presented in the manuscript, are introduced here:

1. the maximum number of alignments taken into consideration
2. the different probe-set classes

Maximum number of alignments: Some probes will return a humongous number of alignments. As these probes are consequently uninformative, probes aligning to more than 5 different loci are discarded. The reason for that choice of threshold is that given the probability that a gene is expressed - and detectable using microarrays - in a tissue being $30\% - 40\%$ (Su et al., 2002; Ramsköld et al., 2009), there is still one chance in four that the value observed for a probe-set mapping to 5 different gene loci comes from the expression of a single one of them - *i.e.* the likelihood, given a binomial distribution with a probability value of 0.4, of a single value from a group of 5 to be TRUE is 26%. The figure 4.1 shows the number of alignments per probes against the *Homo sapiens cdna (Ensembl version 53)* (Flicek et al., 2011) reference for the Affymetrix *GeneChip*® HG-U133 Plus2. About 93% of the probes have less than 6 reported alignments. The proportion is a little

higher (95%) for the alignment against the *Homo sapiens dna (Ensembl version 53)* reference (not shown); as detailed previously, only the probes with a maximum of 5 alignments are kept. The *cdna* alignments not spanning any **exon-exon junction** (EEF) are a subset of the *dna* alignments. After combining these common alignments, more than 80% of the probes have a unique alignment, about 10% align to 2 different loci and the remaining 10% to 3 or more (see Figure 4.2).

The different probe-set classes: As introduced - see section 3.2.2, page 43 - the Ebased CDF defines 3 main classes of probe-sets: *transcript* and *gene*-centric and *genomic*. This last class can not be associated with any gene information. These 3 classes can be refined in sub-classes, *e.g.* the *antisense* class that identifies a probe-set antisense to the related transcript or gene. As can be seen in Figure 4.3, showing the 6 different kinds of sub-classification, about 30% only are of the "transcript" class, a fact that underlines the necessity to use **updated** CDFs to decrease the false discovery rate of any downstream analysis.

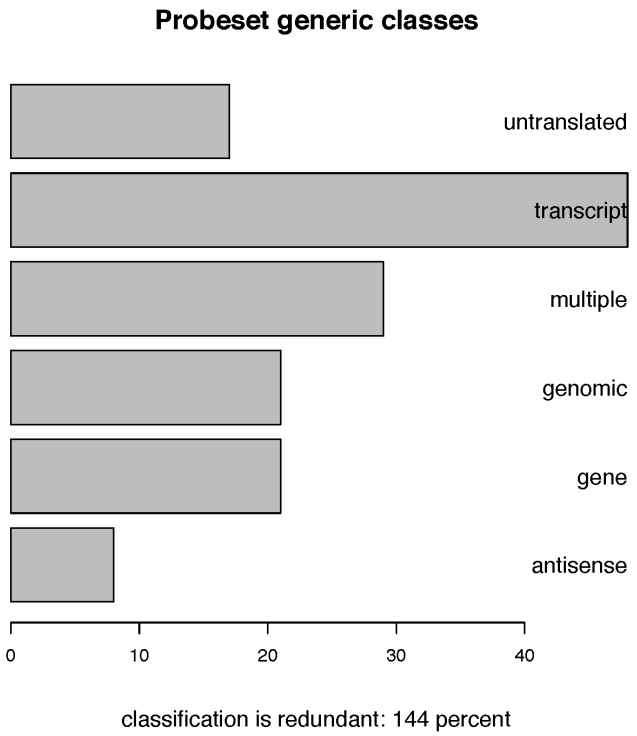


Figure 4.3 – Proportion of the different probe-set classes determined for the HG-U133 Plus2 *GeneChip*®

Note that this representation of the classification is redundant, a probe-set being possibly a member of different sub-classes. When looking at the "transcript" or "gene" centric probe-set non redundant classes (see Figure 4.4 and 4.5), the vast majority of the generated probe-sets match a unique feature. About 17% match untranslated loci, mostly in the 3'UTR region of

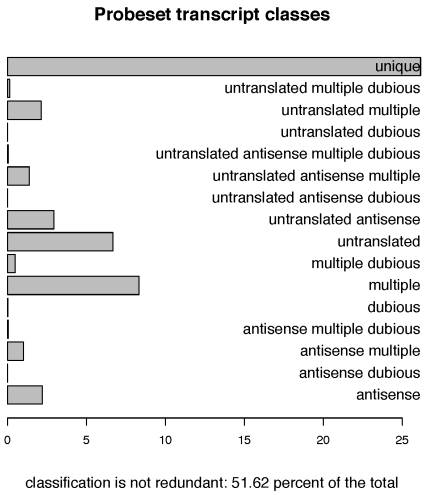


Figure 4.4 – Proportion of the different transcript-centric probe-set classes determined for the HG-U133 Plus2 *GeneChip*®

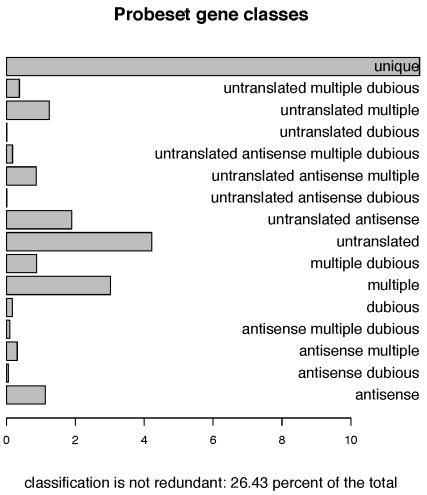


Figure 4.5 – Proportion of the different gene-centric probe-set classes determined for the HG-U133 Plus2 *GeneChip*®

the genes - an expected effect since Affymetrix used **Expressed Sequence Tags** (ESTs) to design the probes at the 3'end of the transcripts. An additional 17% of the probe-sets are associated with multiple features, the proportion of which can be seen in Figure 4.6.

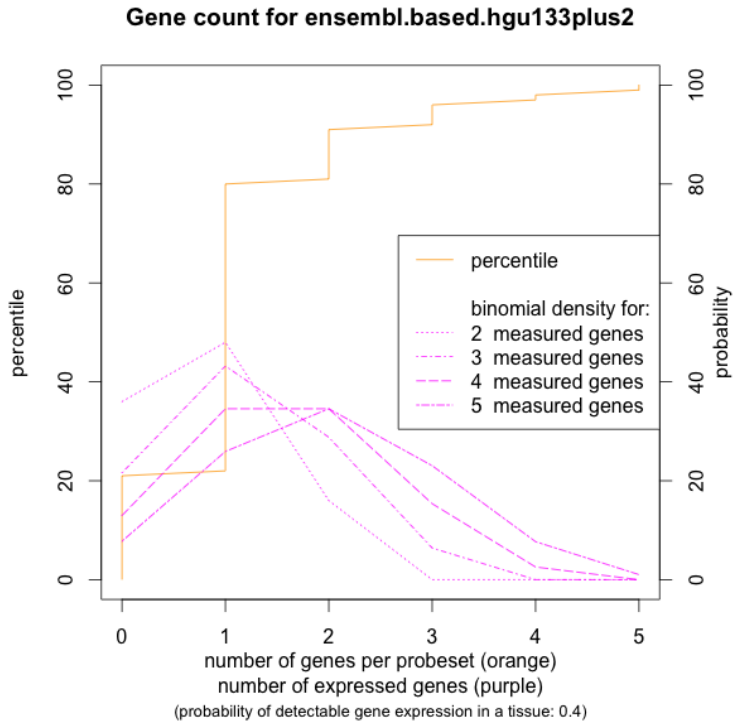


Figure 4.6 – Gene count per probe-set for the HG-U133 Plus2 *GeneChip*®

Half of the probe-sets are associated with two genes whereas the other half is associated with three or more. In the same figure it is interesting to note that 60% of the probe-sets have a unique target while 20% have no identified targets. These are the "non-genic" probe-sets created by the Ebased CDF generation process, see section 3.2.2, page 43. Altogether, using the *Ebased* CDFs adds one third more information that using any other custom CDF.

4.1.3 Data normalization

Selecting a normalization method: As introduced in the previous section, custom CDFs were created to enhance the Affymetrix *GeneChip*® probes' usage. To evaluate the possible effects of the redefined probe-sets on the commonly used Affymetrix normalization methods in R - `rma` (Irizarry et al., 2003), `gcrma` (Wu and Irizarry, 2005) and `vsn` (Huber et al., 2002) - the GSE5222 dataset was normalized using these three methods and the results compared pair-wise. In Figure 4.7, the comparison of the `gcrma` and `rma` normalization methods on the M20517 sample of the GSE5222 dataset is shown. The `gcrma` normalization takes into account the probes' sequence,

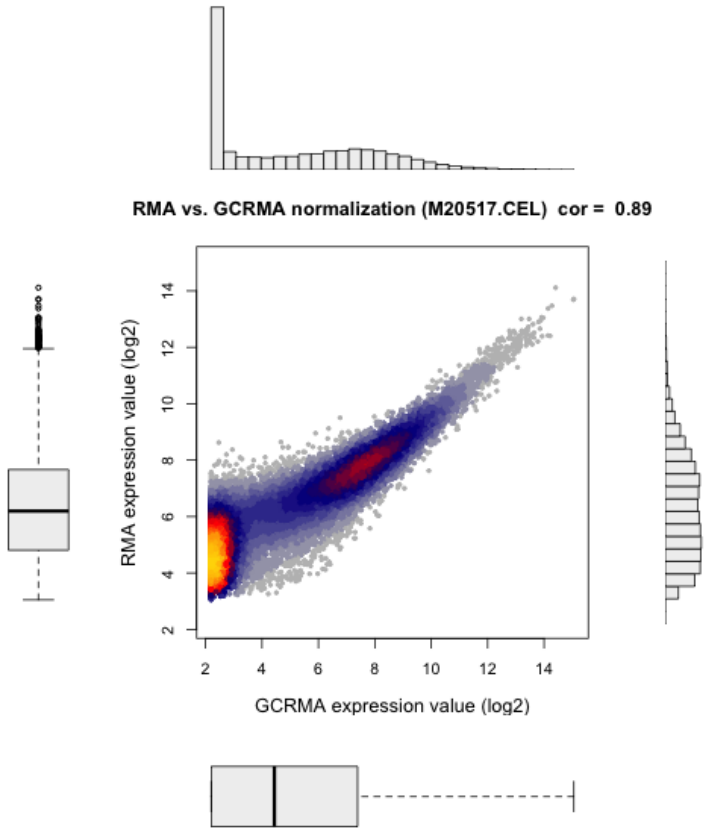


Figure 4.7 – Comparison of the `gcrma` and `rma` normalization methods on the M20517 sample of the GSE5222 dataset. The color scale blue (sparse) - yellow (dense) represents the density of the data points.

structure and affinity to correct the probe-set expression value. As can be seen in the Figure 4.7, this results in the majority of the probe-sets having extremely low log2 expression values with a mean of 2. The detection limit,

commonly accepted on Affymetrix microarray due to its signal-to-noise ratio, is within a log2 value range of 5-6; hence the affinity correction applied here by the `gcrma` approach is innapropriate - determining experimentally the exact probes' affinity values and providing them to `gcrma` would probably correct this - and therefore this normalization method cannot be used.

The comparison of `gcrma` and `vsn` gives the same results and is not shown here.

The comparison of the `rma` and `vsn` methods, shown in Figure 4.8, reveals

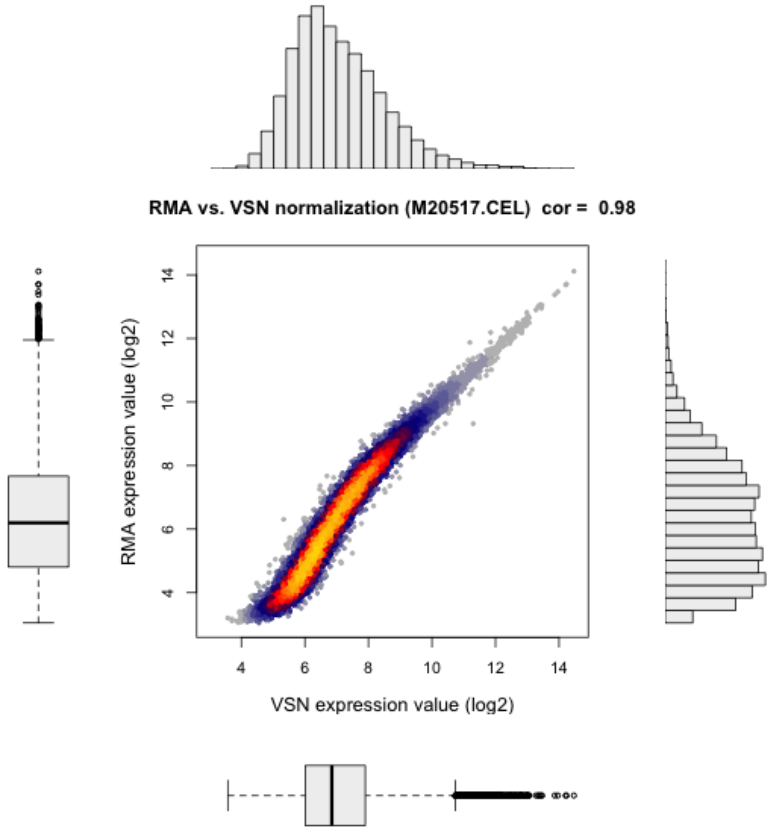


Figure 4.8 – Comparison of the `vns` and `rma` normalization methods on the M20517 sample of the GSE5222 dataset. The color scale blue (sparse) - yellow (dense) represents the density of the data points.

that both yield highly similar results (the Pearson correlation score being very close to one), however the `vsn` package variance stabilization effects are clearly visible: the characteristic "banana-shape" of the scatterplot and the `vsn` expression values distribution close to the expected shape of a log2

distribution. These indicate that the variance was stabilized across the 10 log2 FCs of the distribution: *i.e.* high expression values that tend to have larger variance and low expression values, where the variance tends to be much larger than the actual expression values, have had their variance harmonized, rendering the data homoscedastic, a pre-requisite for downstream analyses assuming a normal or log-normal distribution (*e.g.* the linear models used in the following sections). Consequently, the `vsn` normalization was selected.

Finally, the last step of normalization - the normalization between arrays - is adequate as shown in the Figure 4.9 boxplots for the GSE5222 dataset, where every sample distribution is highly similar. For comparison, the third page of the QA report presented in the appendix page 184 show the corresponding raw data boxplot.

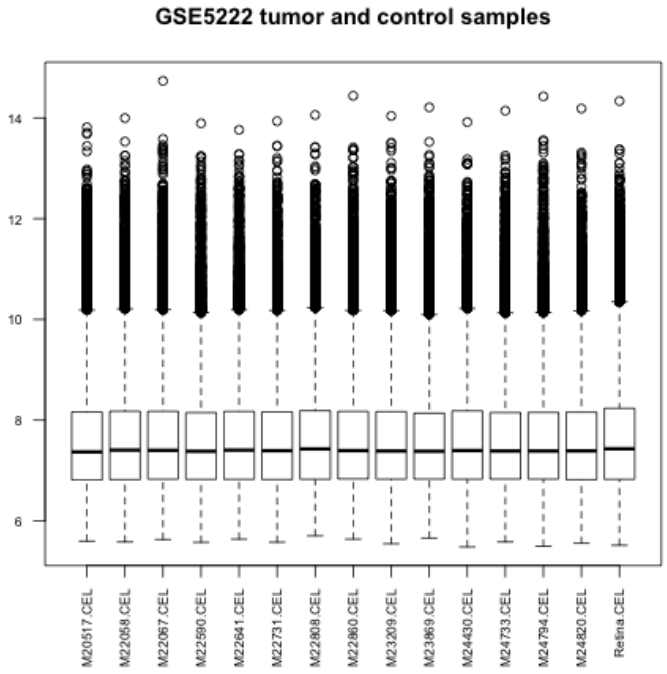


Figure 4.9 – Boxplots showing the normalized distribution of the expression values of all GEO GSE5222 samples.

To conclude, using the redefined CDFs does not have an effect on the data normalization.

4.1.4 Differential Expression

To validate the Ebased CDF, two differential expression analyses are performed, one with retinoblastoma and one with osteosarcoma data. Both follow the same experimental design "Tumor vs. Control".

Retinoblastoma: The tumor samples from the GSE29683 dataset are compared to the GSE5222 control samples using the R `limma` package. Only probe-sets having a p-value lower than 0.05 and a log2 FC higher than 2 are conserved. 988 probe-sets pass these criteria, the 20 first of which are reported in Table 4.2, 11 are up and 9 down-regulated.

Over-expressed genes[1]: The results presented in this table validate the Ebased CDF and indicate that the normalization approach has corrected possible technical bias (such as the *GeneChip*® being prepared and hybridized in different facilities). Indeed, 6 genes identified by the 11 over-expressed probe-sets (from a total of 9 genes) have been associated with cancer: *RRM2* and *ASPM* are implicated in neo-angiogenesis (Zhang et al., 2009; Lin et al., 2008), *TMSB15A* is implicated in cell migration and metastasis (Tang et al., 2011), *TOP2A* is amplified in breast cancer and is the target of therapeutic agents such as trastuzumab and anthracyclines (Arriola et al., 2008), *NUF2* is implicated in a cell cycle checkpoint (DeLuca et al., 2003) and *CDC2* is an essential cell cycle member. *EFHC2, LCORL, EZH2* have not yet been clearly associated with cancer.

Down-regulated genes: All the genes (n=8) identified with the 9 down-regulated probe-sets are present in differentiated tissue, and are retina-specific. *RHO* is a retina pigment, *PDE6A* is a phosphodiesterase expressed in retina rod cells, *SAG* is a major soluble photoreceptor protein, *PPEF2* is involved in the rhodopsin (the product of the *RHO* gene) dephosphorylation (Ramulu et al., 2001), *PTGDS* is involved in the regulation of non-rapid eye movement sleep in mice, binds small non-substrate lipophilic molecules such as retinoic acid and is possibly involved in the development and maintenance of the blood-retina barrier in human, *TTR* mutation may result in vitreous opacities, *GNG13* is an heterotrimeric G protein gamma subunit and is expressed in retina, *RLBP1* is a soluble retinoid carrier essential for the function of both rod and cones photoreceptors.

[1] gene annotation from http://www.genecards.org (Rebhan et al., 1997)

Gene symbol	Gene name	log FC	Avg. expr.	Adj. p-value	Gene count
RRM2	Ribonucleoside-diphosphate reductase subunit M2	6.19	11.53	0.00	1
RRM2	Ribonucleoside-diphosphate reductase subunit M2	5.44	11.38	0.00	2
RHO	Rhodopsin	-5.17	8.06	0.00	1
TMSB15A	NB thymosin beta	5.07	11.47	0.00	1
TOP2A	topoisomerase (DNA) II alpha 170kDa	5.04	10.18	0.00	1
PDE6A	Rod cGMP-specific 3',5'-cyclic phosphodiesterase subunit alpha	-4.84	7.95	0.00	1
PPEF2	Serine/threonine-protein phosphatase with EF-hands 2	-4.83	7.04	0.00	1
SAG	S-arrestin	-4.82	9.23	0.00	1
RRM2	Ribonucleoside-diphosphate reductase subunit M2	4.80	10.68	0.00	2
EFHC2	EF-hand domain-containing family member C2	4.69	10.72	0.00	2
PTGDS	Prostaglandin-H2 D-isomerase Precursor	-4.65	8.95	0.00	1
TTR	Transthyretin Precursor	-4.59	8.54	0.00	1
ASPM	Abnormal spindle-like microcephaly-associated protein	4.55	9.47	0.00	1
NUF2	Kinetochore protein Nuf2	4.46	9.59	0.00	1
CDC2	Cell division control protein 2 homolog	4.39	10.97	0.00	1
GNG13	Guanine nucleotide-binding protein G	-4.38	7.25	0.00	1
LCORL	Ligand-dependent nuclear receptor corepressor-like protein	4.37	9.56	0.00	1
RLBP1	Retinaldehyde-binding protein 1	-4.37	7.77	0.00	1
PTGDS	Prostaglandin-H2 D-isomerase Precursor	-4.36	9.30	0.00	1
EZH2	Polycomb protein EZH2	4.21	10.62	0.00	2

Table 4.2 – retinoblastoma GSE29683 (tumor) vs. GSE5222 (control) differential expression. The top 20 probe-sets are presented.

Genomic probe-sets: Another validation of the Ebased CDF comes from the first identified "genomic" probe-set: "15_38743354_38743629_plus_genomic_at" that maps just downstream of the *CASC5* gene - the "cancer susceptibility candidate 5" gene - probably in its 3' UTR region, see Table 4.3. The UTR regions are difficult to define experimentally in human and have been shown to vary in size depending on the tissue, *i.e.* different termination/poly-adenylation sites exist for single genes and their usage is regulated in a tissue-dependent manner (Zhang et al., 2005).

ID	log FC	Avg. expr.	Adj. p-value
15_38743354_38743629 _plus_genomic_at	3.87	8.79	0.00

Table 4.3 – The first most significant "genomic" probe-set of the retinoblastoma GSE29683 tumor vs. GSE5222 control differential expression analysis.

Osteosarcoma: The GSE14359 tumor and metastasis samples are compared with the non-neoplastic osteoblast samples. The data from the Fritsche-Guenther et al. (2010) study allow four kinds of comparisons:

1. Tumor *vs.* Control
2. Metastasis *vs.* Control
3. Tumor and Metastasis *vs.* Control
4. Tumor *vs.* Metastasis (the Control samples cancelling each-other out)

all of which were performed to determine the significant transcripts involved in osteosarcoma primary and metastasis development, as well as those specific for the metastazation process.

Tumor and Metastasis *vs.* Control: The Table 4.4 contains the results of applying linear models to determine significant changes in gene expression using the same parameters as above (see paragraph 4.1.4, page 76); the 20 most significantly changed probe-sets are shown. 10 of the down-regulated genes ($n = 17$ in total) are precursors of either trophic or mitogenic factors or of protein receptors. This is as expected for healthy bone tissue - see section 1.1.3, page 18 - which ECM acts as a reserve for such factors. The 5 remaining genes: *HSPB6, AHNAK2, PDLIM2, NQO1, PAPB2, PTGIS* can all be related to cancer. *HSPB6*, as well knows as *Hsp20*, has been associated with a proliferation suppression effect - unlike most other heat shock proteins - in hepatocellular carcinoma (Matsushima-Nishiwaki

ID	GeneSymbol	GeneName	logFC	AveExpr	adj.P.Val	GeneCount
ENST00000292896_ENST00000380237_transcript_multiple_at	HBE1	Hemoglobin subunit epsilon	2.73	6.94	0.00	3
ENSG00000006016_gene_at	CRLF1	Cytokine receptor-like factor 1 Precursor	-4.23	6.00	0.00	1
ENSG00000134363_gene_at	FST	Follistatin Precursor	-3.48	6.90	0.00	1
ENST00000314922_transcript_at	PENK	Proenkephalin A Precursor	-3.47	5.87	0.00	1
ENST00000260356_ENST00000397593_transcript_at	THBS1	Thrombospondin-1 Precursor	-2.67	6.03	0.00	1
ENST00000004982_transcript_at	HSPB6	Heat shock protein beta-6	-2.56	6.64	0.00	1
ENSG00000064205_gene_at	WISP2	WNT1-inducible-signaling pathway protein 2 Precursor	-4.03	6.37	0.00	1
ENST00000311330_transcript_at	CD248	Endosialin Precursor	-2.86	6.94	0.00	1
ENST00000301464_transcript_at	IGFBP6	Insulin-like growth factor-binding protein 6 Precursor	-3.80	7.06	0.00	1
ENSG00000162407_gene_at	PPAP2B	Lipid phosphate phosphohydrolase 3	-3.37	7.36	0.00	1
ENST00000260356_transcript_at	THBS1	Thrombospondin-1 Precursor	-2.74	7.20	0.00	1
5_148308172_148308417_plus_genomic_multiple_at			-3.00	7.30	0.00	0
ENST00000320623_ENST00000379046_transcript_at	NQO1	NAD(P)H dehydrogenase, quinone 1	-3.10	5.76	0.00	1
ENSG00000129538_gene_at	RNASE1	Ribonuclease pancreatic Precursor	4.30	9.51	0.00	1
ENST00000333244_transcript_at	AHNAK2	AHNAK nucleoprotein 2	-2.53	5.57	0.00	1
ENST00000374431_transcript_at	LPAR1	Lysophosphatidic acid receptor 1	-2.51	6.11	0.00	1
ENSG00000133048_gene_at	CHI3L1	Chitinase-3-like protein 1 Precursor	-5.29	6.35	0.00	1
ENST00000361970_transcript_antisense_at	CCDC152	Coiled-coil domain-containing protein 152	2.52	10.22	0.00	1
ENST00000244043_transcript_multiple_at	PTGIS	Prostacyclin synthase	-2.98	7.49	0.00	2
ENST00000308354_truncated_6_transcript_at	PDLIM2	PDZ and LIM domain protein 2	-2.24	7.22	0.00	1

Table 4.4 – osteosarcoma GSE14359 tumor vs. control differential expression. The top 20 probe-sets are presented.

et al., 2011). *AHNAK*, which has a high sequence homology to *AHNAK2*, has been shown to be downregulated in cell lines of neuroblastoma, small cell lung carcinoma and Burkitt lymphoma (Amagai, 2004). *PDLIM2* and *NQO1* are both negative regulators of the NF-κB pathway (Tanaka et al., 2007; Jamshidi et al., 2012). *PAPB2* is involved in the poly(A) tail formation (Hurschler et al., 2011), a process often deregulated in cancer (Audic and Hartley, 2004). Finally, *PTGIS* is involved in the prostacyclin pathway. An analog of this lipid molecule has been shown to inhibit non-small cell lung cancer (Tennis et al., 2010). The last down-regulated probe-set - of the "genomic" class - lies in a gene empty region and cannot clearly be associated with any genic feature. Moreover, as some of its probes map multiple locations any analysis without further experimental validation would be speculative.

Concerning the up-regulated genes, two of the probe-sets - "ENST00000292896_ENST00000380237_transcript_multiple_at" and "ENST00000361970_transcript_antisense_at" - need further analyses. The first one, associated with the gene *HBE1*, has probes mapping over genes. A closer look reveals that those genes: *HBD, HBG2, HBG1, HBB* are all Hemoglobin subunits. The second probe-set, antisense to the *CCDC152* actually overlaps the 3' UTR region of the *SEPP1* gene located on the opposite strand. That later gene has often been associated with cancer (Persson-Moschos et al., 2000; Gonzalez-Moreno et al., 2011), but as being down-regulated. Finally the *RNASE1* gene has been associated with cancer (Leland et al., 2001; Barrabés et al., 2007), but its role not clearly elucidated.

Even though the results are sensible, it is difficult to determine which genes are causative and which not, *i.e.* the change in expression of 12 out of the 20 most significantly differentially expressed genes is probably a side effects of the de-differentiation of osteoblasts during cancerogenesis. In the following, other comparisons are performed on this dataset to assess whether it can help resolve this issue.

Tumor only *vs.* Control: To assess whether the metastasis samples are introducing confounding factors that affect the results, the DE analysis was performed next using only the tumor samples. The obtained list of differentially expressed genes is very similar to the former one, see Figure 4.10 for a Venn diagram comparing the identified probe-sets. Only 3 probe-sets differ in the first 30 of both lists, identifying two down-regulated genes, both coding for protein precursors: *N-sulphoglucosamine sulphohydrolase Precursor* and *CD44 antigen Precursor*. These probe-sets became significant in the tumor-only DE analysis because their expression in metastasis is highly variable as exemplified on Figure 4.11 for *CD44*. That gene is involved in many biological processes: cell proliferation, migration, *etc.* and has been reported as being over-expressed in many cancers. The fact that it is down-

regulated here agrees with the hypothesis of the previous paragraph: *i.e.* that these precursors are produced to be stored in the bone ECM. However, focusing on the tumor samples only for the DE analysis did not help to identify more likely causative genes.

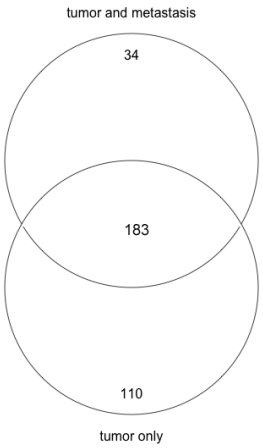


Figure 4.10 – Venn diagram of the probe-sets differentially expressed in the tumor-only and tumor + metastasis analyses

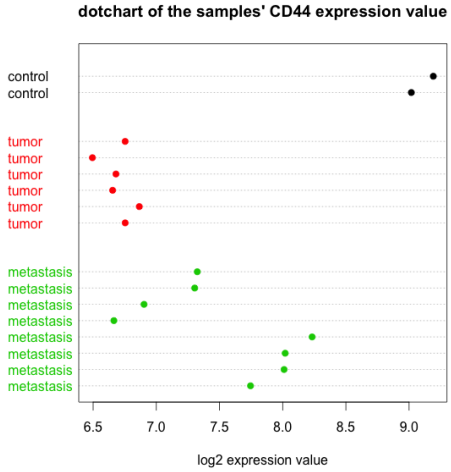


Figure 4.11 – Dotchart of the GEO GSE14359 dataset *CD44* gene expression values. The metastasis samples show a higher variability.

Metastasis only *vs.* Control: As in the previous paragraph, performing a DE expression using only the metastasis samples did not reveal any potentially causative genes, only one differs in the first 40 - the *Cytokine-like protein 1 Precursor* - when comparing the complete and the metastasis-only differentially expressed gene lists with each other, see Figure 4.12. It encodes another protein precursor, which seems not to be expressed in metastasis samples - the microarray detection limit is within the range of 5 to 6 -, see Figure 4.13.

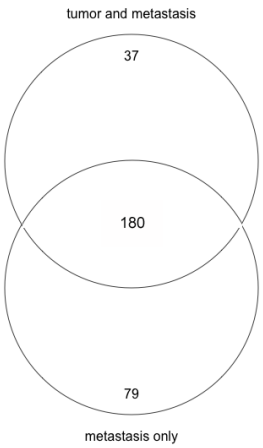


Figure 4.12 – Venn diagram of the probe-sets differentially expressed in the metastasis-only and tumor+metastasis analyses

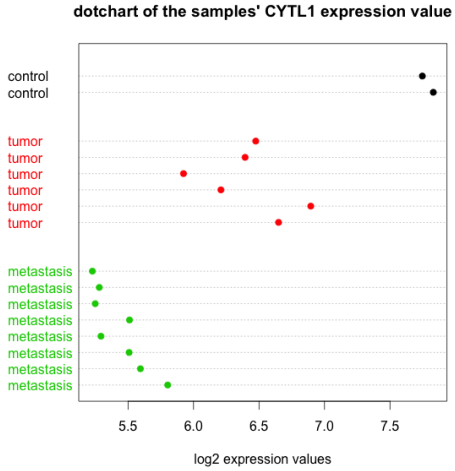


Figure 4.13 – Dotchart of the GEO GSE14359 dataset *CYTL1* gene expression values. The metastasis samples have values very close to the Affymetrix microarrays detection limit.

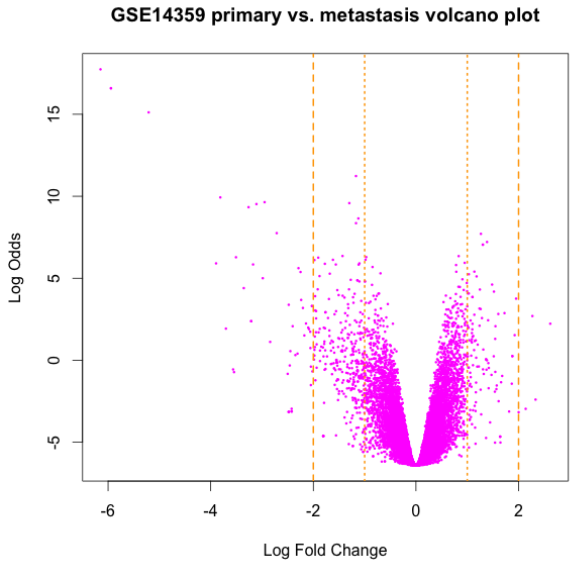


Figure 4.14 – Volcano plot of the GSE14359 primary *vs.* metastasis DE statistic significance. The orange doted lines represent the log2 FC cutoff used while the dashed ones represent the commonly used one in the rest of the analyses.

Osteosarcoma primary vs. metastasis samples: Neither the comparison of the primary tumors nor that of the metastasis with the osteoblast controls yielded evidence as of which genes may be causative of the cancerogenesis. The direct comparison of the primary and metastasis samples offers a possibility to reduce the effect of the confounding factors, *i.e.* those genes that are differentially expressed as a consequence of cancerogenesis. As both kinds of samples are similar, the likelihood to observe differential expression is lower; as a consequence the log2 FC threshold has been reduced to 1 instead of 2, a choice justified by the volcano plot presented in Figure 4.14. The 20 most significantly differentially expressed genes are shown in Table 4.5[2]. In that table, it is evident that other confounding factors appear: *e.g. Pulmonary surfactant-associated proteins*, *SLC34A2* that may have a role in the synthesis of surfactant in the lungs' alveoli, *etc.* However some interesting candidates show up, such as *HMBOX1* that encodes a TF that acts as a transcriptional repressor, *HOPX* that may be a tumor suppressor gene, *MCTS1* that is versatilely associated with cancer and *TACSTD2* that may function as a growth factor receptor.

Limiting confounding factors: To further remove the confounding factors, the 4 candidate lists obtained from the different comparisons of the GSE14539 dataset are intersected, see the Venn diagram in Figure 4.15.

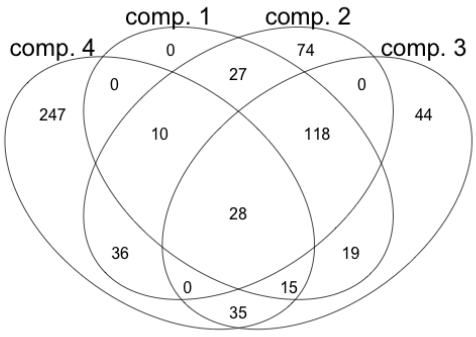


Figure 4.15 – Venn diagram of the four candidate genes lists obtained by comparing the different samples of the GSE14359 dataset. The comparison are numbered as introduced on page 78, *e.g.* "comp. 1" is the first comparison: tumor + metastasis *vs.* control.

[2]gene annotation retrieved from www.genecards.org (Rebhan et al., 1997)

ID	GeneSymbol	GeneName	logFC	AveExpr	adj.P.Val	GeneCount
ENST00000318561_transcript_at	SFTPC	Pulmonary surfactant-associated protein C Precursor	-6.15	9.27	0.00	1
ENST00000372316_transcript_multiple_at	SFTPA2B	Pulmonary surfactant-associated protein A2 Precursor	-5.94	9.00	0.00	4
ENST00000372329_transcript_multiple_at	SFTPA2	Pulmonary surfactant-associated protein A2 Precursor	-5.94	9.00	0.00	4
ENST00000393822_transcript_at	SFTPB	Pulmonary surfactant-associated protein B Precursor	-5.21	8.05	0.00	1
ENSG00000196188_gene_at	CTSE	Cathepsin E Precursor	-1.17	6.33	0.00	1
ENST00000342375_ENST00000409383_transcript_at	SFTPB	Pulmonary surfactant-associated protein B Precursor	-3.81	7.83	0.00	1
ENSG00000197249_gene_at	SERPINA1	Alpha-1-antitrypsin Precursor	-2.95	7.54	0.00	1
ENSG00000181523_gene_multiple_at	SGSH	N-sulphoglucosamine sulphohydrolase Precursor	-1.30	6.87	0.00	3
ENST00000355814_..._transcript_at	SERPINA1	Alpha-1-antitrypsin Precursor	-3.11	8.49	0.00	1
ENST00000382051_transcript_at	SLC34A2	Sodium-dependent phosphate transport protein 2B	-3.26	7.04	0.00	1
ENST00000355231_ENST00000397358_transcript_at	HMBOX1	Homeobox-containing protein 1	-1.12	7.36	0.00	1
ENST00000265368_..._truncated_11_transcript_at	SYNE1	Nesprin-1	-1.17	6.52	0.00	1
ENST00000317745_..._transcript_at	HOPX	Homeodomain-only protein	-2.71	7.18	0.00	1
ENSG00000101898_gene_multiple_at	MCTS1	Malignant T cell amplified sequence 1	1.27	7.75	0.00	2
ENST00000233573_ENST00000397033_transcript_at	ITGA4	Integrin alpha-4 Precursor	1.38	5.32	0.00	1
2_124236879_124237296_plus_genomic_multiple_at			1.30	7.82	0.00	0
ENSG00000115233_gene_multiple_at	PSMD14	26S proteasome non-ATPase regulatory subunit 14	1.30	7.82	0.00	2
ENST00000332160_transcript_at	METTL7A	Methyltransferase-like protein 7A Precursor	-1.43	7.32	0.00	1
ENSG00000170017_gene_at	ALCAM	CD166 antigen Precursor	-1.90	7.33	0.00	1
ENST00000371225_transcript_at	TACSTD2	Tumor-associated calcium signal transducer 2 Precursor	-1.62	6.02	0.00	1

Table 4.5 – GSE14359 primary vs. metastasis differential expression; the top 20 probe-sets are presented. A negative fold change represents an over-expression of the probe-set in metastasis samples while a positive fold change represents an over-expression in the primary tumor samples.

GeneSymbol	GeneName	Description
Over-expressed in metastasis		
CLIC3	Chloride intracellular channel protein 3	may participate in cellular growth control, based on its association with ERK7, a MAP kinase
NPR3	natriuretic peptide receptor C	involved in metabolic and growth processes
CAV1	Caveolin-1	co-activator of the NF-κB and *Wnt* pathways
CSTA	Cystatin-A	has been proposed as prognostic and diagnostic tools for cancer.
TM4SF1	Transmembrane 4 L6 family member 1	plays a role in the regulation of cell development, activation, growth and motility; highly expressed in different carcinomas.
S100A6	Protein S100-A6	Involved in the cycle progression and differentiation. Chromosomal rearrangements and altered expression have been implicated in melanoma.
MEST	Mesoderm-specific transcript homolog protein	Involved in development. It is imprinted, exhibiting preferential expression from the paternal allele. The loss of imprinting has been linked to cancer and may be due to promotor switching.
Over-expressed in primary tumor		
MAD2L1	Mitotic spindle assembly checkpoint protein MAD2A	a component of the mitotic spindle assembly checkpoint. Prevents the onset of anaphase until all chromosomes are properly aligned at the metaphase plate
MNAT1	CDK-activating kinase assembly factor MAT1	Involved in cell cycle control and in RNA transcription by RNA polymerase II
RRM2	Ribonucleoside-diphosphate reductase subunit M2	involned in neo-angiogenesis, see section 4.1.4, page 76
PRAME	Melanoma antigen preferentially expressed in tumors	transcriptional repressor, inhibiting the signaling of retinoic acid (RA) through the RA receptors. Prevents RA-induced cell proliferation arrest, differentiation and apoptosis.

Table 4.6 – Subset of the candidate genes that are differentially expressed between the primary and metastasis samples of the GEO GSE14359 dataset. (source: www.genecards.org)

36 probe-sets are differentially expressed between the tumor *vs.* the controls and metastasis respectively. Half of them are trophic factor precursors (n=15), genomic loci (n=4) or their known annotation irrelevant (n=4). The remaining genes are presented in Table 4.6. Both cell populations are different, supporting the hypothesis that metastasis is not a trait appearing late in tumors (see section 1.1.1, page 11).

Despite how revealing these results are, they are still obscured by confounding factors. It is clear that such approaches solely based on microarray EP, although refining our knowledge of cancer, are not sufficient to grasp the cancerogenesis process. To address this, integrating and comparing the EP results with that of other analyses is necessary. In the next section, I introduce a complementary arrayCGH dataset used for that purpose.

4.2 Array based CGH

The dataset from Zielinski et al. (2005) was actually performed in two batches using different microarray layouts, see paragraph 3.2.4, page 46. The QA and segmentation analysis -including the data filtering and normalization - were performed on both sets independently. It is important to note that the samples have been matched with an opposite sex control as an additional QA. This has the drawback that the sexual chromosome abberations cannot be determined.

4.2.1 Quality Assurance

All seventen samples pass the QA. For an example of a quality report, see appendix B, page 184. The sample presented there is M23215, one of the samples failing to pass the QA in the GSE5222 EP dataset (see section 4.1.1, page 67).

Merging batches: As mentioned, the Zielinski et al. (2005) dataset was performed on two batches. The quality of both batches is similar and not affected by the use of different microarray layouts. The two layouts used contain the same probes in an equal number of replicates, only their positioning differs. This actually offers an additional control for the possible technical artefacts such as "border effects" - *i.e.* when the measured intensity of a spot would be affected by a physical boundary such as a chip edge.

4.2.2 Profile segmentation

The segmentation performed using the *Alterations* (Toedt et al., a, unpublished) tool, an optimized implementation of `GLAD` (Hupé et al., 2004) for ChipYard (Toedt et al., b, unpublished) (Toedt et al., b) gave similar results as that of the original analysis performed by Zielinski et al. (2005). The Figure 4.16 shows the results for the M23215 (Chip ID: 572) sample and the Figure 4.17 shows the CNV results for the whole dataset.

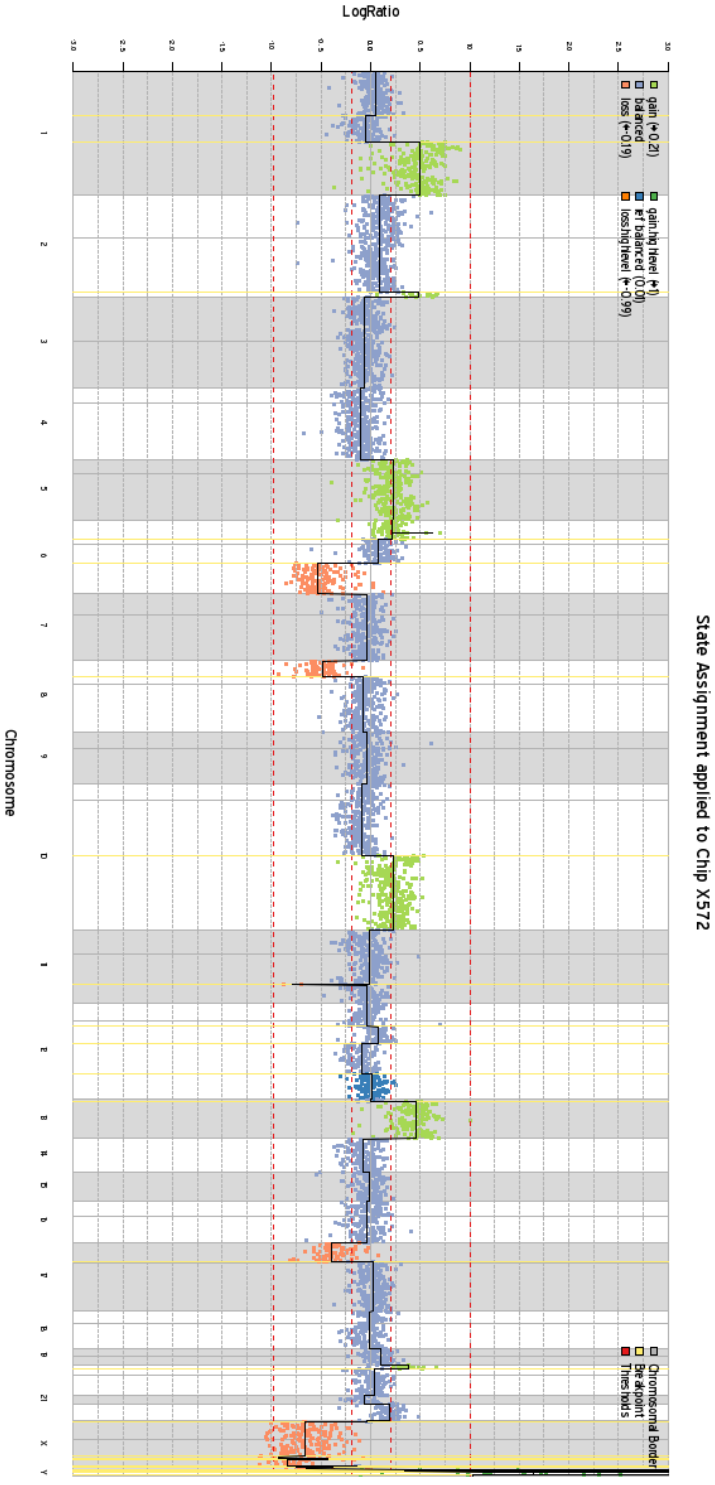


Figure 4.16 – CNV segmentation and state assignment of the M23215 sample. In blue are the non-aberrant loci, in green and in red are represented gains and losses of genomic material, respectively. The dark blue region is the one used as a reference to infer the loci state.

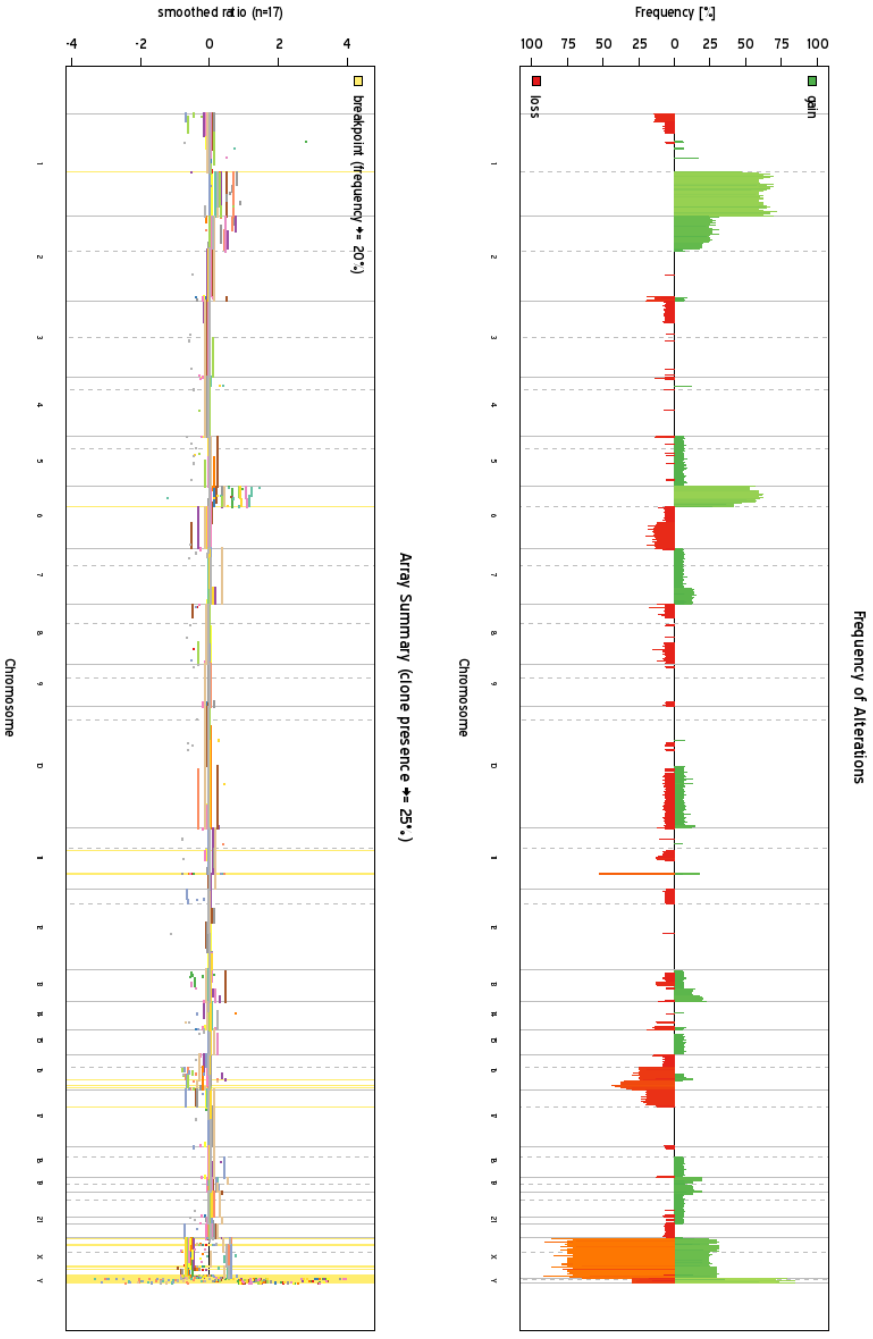


Figure 4.17 – Frequency of alterations in the complete arrayCGH dataset. Shown in green are the gains and in red the losses. The chromosome X and Y are technical artefacts, i.e. the samples were opposite-sex matched. The proportion of male in the analysis - represented by the loss of X and the similar gain of Y - roughly 70%, is as expected (12 males/17samples) validating the overall experiment.

Almost all the abberations are found that have been identified in Zielinski et al. (2005) (case #12, table 1, page 296, n=19). As can be seen in the tables 4.7 and 4.8, the identified alterations encompass the ones previously identified, with only two differences. The chromosome 22q13.2 gain

	locus	type
1	1q21.1-1q44	gain
2	2q37.1-2q37.3	gain
3	5p15.33-5p12	gain
4	5q11.1-5q35.3	gain
5	6p25.3-6p21.1	gain
6	10q24.31-10q26.3	gain
7	13q12.13-13q34	gain
8	20p13-20p13	gain
9	6q22.33-6q27	loss
10	8p23.3-8p12	loss
11	11q22.1-11q22.1	loss
12	17p13.3-17p11.2	loss

Table 4.7 – Aberrations identified in the M23215 sample using the *Alterations* tool

	locus	type
1	1q21-1q44	gain
2	2q37.2q37.3	gain
3	5p15.32p15.31	gain
3	5p13.2p13.2	gain
4	5q11.2q11.2	gain
4	5q13.2q13.2	gain
4	5q23.1q35.3	gain
5	6p21.33p21.1	gain
6	10q24.31q25.1	gain
6	10q25.2q25.2	gain
6	10q25.3q25.3	gain
6	10q26.11q26.12	gain
6	10q26.13q26.3	gain
7	13q12.13q34	gain
8	20p13p13	gain
13	22q13.2q13.2	gain
9	6q22.33q27	loss
10	8p23.3p12	loss
12	17p13.2p11.2	loss

Table 4.8 – Aberrations originally identified in the M23215 sample by Zielinski et al. (2005).

originally observed falls under the detection threshold of the new method, whereas an almost high level loss on chromosome 11q22.1 was originally overseen. Overall, the frequency previously observed and newly calculated are identical as shown in Table 4.9. Numerous additional aberrations are found affecting mostly a few individuals (n = 1-3), but for the chromosome 11q22.1 band that is lost in 60% of the cases and gained in 18 others.

locus	%gain	%loss
Zielinski et al. (2005)		
1q	**71**	
2p	29	
6p	**59**	
13q	12	12
16q		**41**
17p		18
19p	12	
19q	24	
newly identified		
1p		12
2q		12
3p		6
5p	6	
5q	6	
6q		9
7p	6	
7q	9	
8p		6
8q		6
10q	6	6
11q	18	**60**
12p		6
14q	6	12
15q	6	
16q	6	
18q	6	
19p	12	
19q	12	
20p	6	
20q	6	
21q	6	
22q		6

Table 4.9 – Alterations originally reported and newly identified in the Zielinski et al. (2005) dataset.

Genes present on band 11q22.1: Three genes are found in that region, see Table 4.10. Among those, *PGR* could be an interesting candidate as it is involved in cell proliferation and differentiation.

Gene symbols	Gene names
CNTN5	Contactin-5 precursor (Neural recognition molecule NB-2) (hNB-2).
Q96M56_HUMAN	Pseudogene, weakly similar to Homo sapiens oligophrenin 1.
PGR	Progesterone receptor (PR).

Table 4.10 – Genes mapped to the chromosome 11q22.1 band, lost in 60% and gained in 18% of the samples of the Zielinski et al. (2005) dataset.

4.2.3 Integrative analysis using clinical parameters

The availability of clinical data - see Table A.1, page 177 in the Appendix A, page 175 - allowed us to look for correlations between the CNVs and different traits.

Multidimensional Scaling: The first approach using MDS to reduce confounding factors did not reveal any striking correlation between the arrayCGH data and the clinical factors, as can be see on Figure 4.18.

Hierarchical clustering: The second approach, using **hierarchical clustering** revealed that the presence of vitreous seeding correlates with the number of aberrations, see Figure 4.19. Tumors without vitreous seeding have a median of 1 aberration per tumor, whereas their converse have a median of 6 aberrations per tumor, a statistically significant difference (Welch Two Sample t-test p-value of 0.02). A similar finding had been reported in Gratias et al. (2007). It is interesting to note that out of the 4 samples with few aberrations and no vitreous seeding, two are hereditary and bilateral, while a third one - spontaneous - is unilateral but multifocal, indicative of a possible germline mutation, which was experimentally confirmed. As expected from the Knudson two-hit hypothesis, patients with a germline mutation (n = 6 from a total of 14 - 3 values are missing) have less aberrations: a median value of 3 *vs.* 5 respectively (Welch Two Sample t-test p-value of 0.1). However the association between the gain on chromosome 1q and a later onset of the disease, reported in Gratias et al. (2005) could not be verified - median values of 675 *vs.* 550 days with a Welch Two Sample t-test p-value of 0.9.

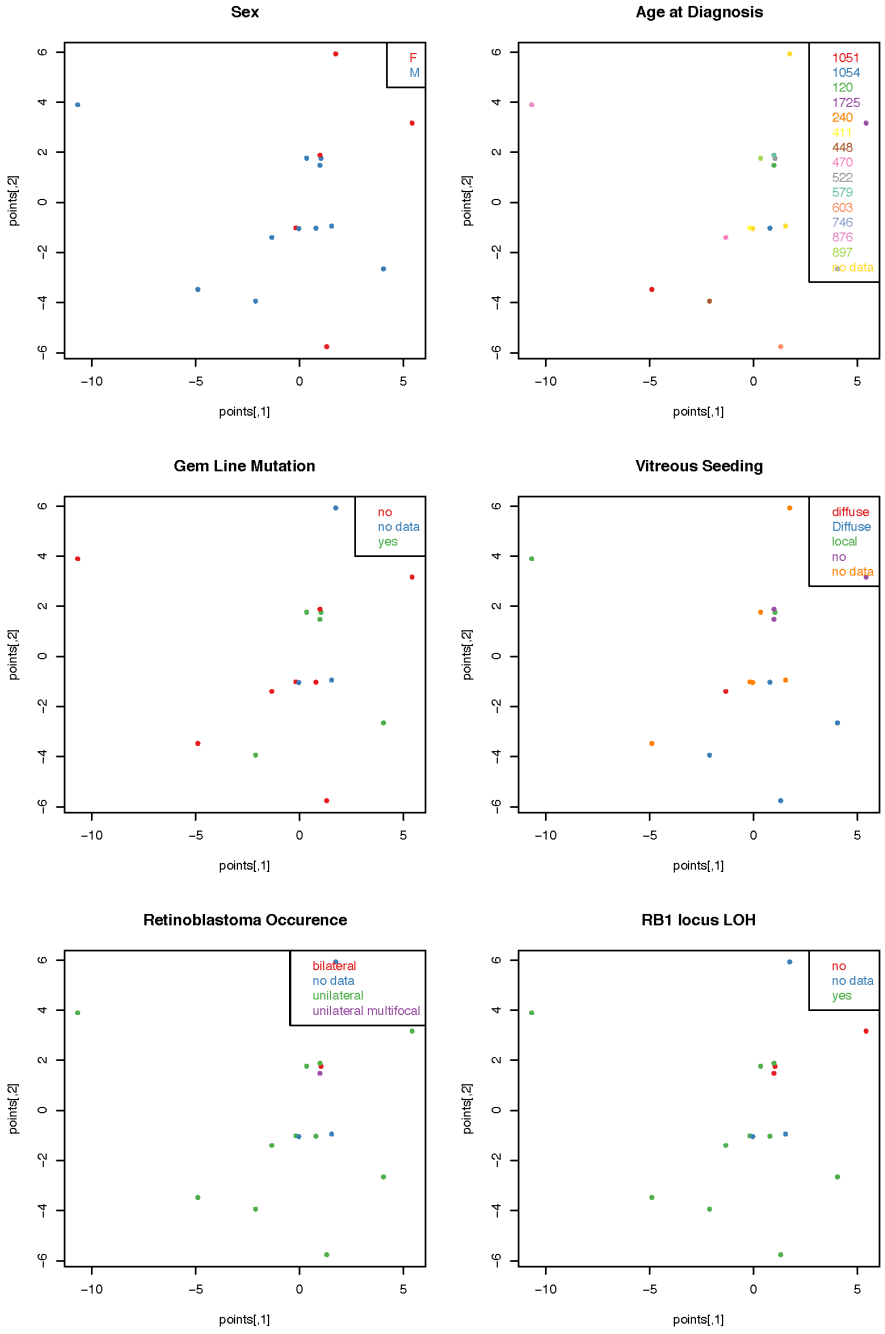


Figure 4.18 – MDS of the Zielinski et al. (2005) dataset. Only the first two dimensions are represented. In every panel, the title describe the trait under investigation and in the upper right corner are the different values color coded. There is no obvious grouping of the data points according to a trait.

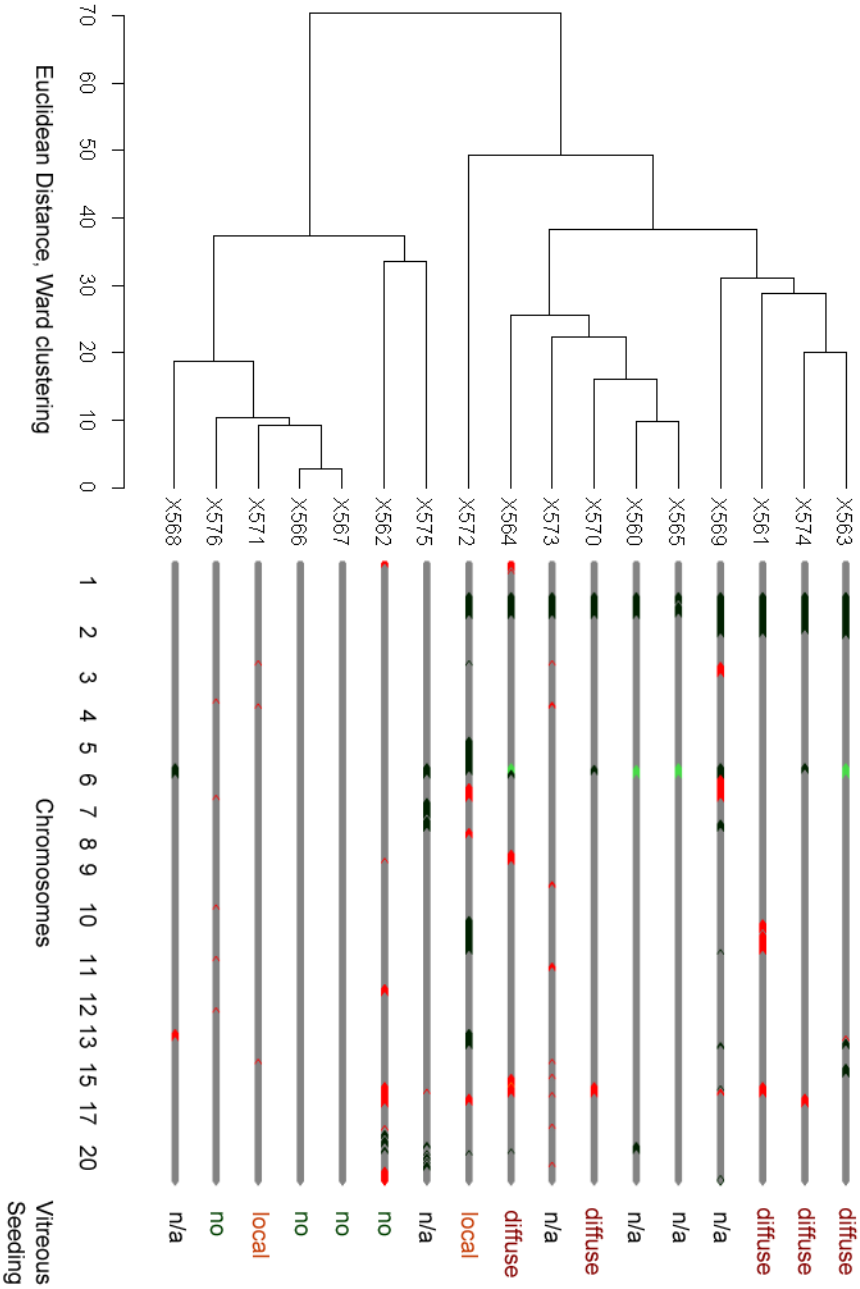


Figure 4.19 – hierarchical clustering of the Zielinski et al. (2005) dataset. The samples are ordered by similarities, using the Euclidean distance and the Ward clustering. The aberrations are reported in green (gain) and red (loss). The vitreous seeding status is indicated on the right of the figure.

These interesting results show the potential of performing integrative analyses, however the rather small amount of arrayCGH samples makes it difficult to draw more significant conclusions. To achieve the later, as presented in the section 4.4, page 103, I have combined the retinoblastoma arrayCGH with the EP data. This required the development of new statistical analysis methods, which in turn needed to be assessed for their sensitivity and specificity. This is best done using simulated data, *i.e.* the outcome (*expected*) is well defined and can be used to score the method results (*observed*). As no simulator could be identified that implemented the necessary data models, I developed a microarray simulator, which assumptions and performances are shortly evaluated in the next section.

4.3 Microarray simulation

As introduced in the section 3.3, page 48, the developed `aSim` package uses mixture models to simulate EP or arrayCGH data. Figure 4.20 presents an example of a mixture model as used in the simulations and analyses introduced further on.

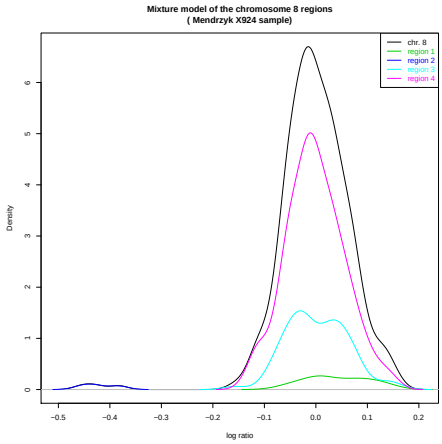


Figure 4.20 – arrayCGH mixture model: plotted are the simulated values from the chromosome 8 of the ependyma sample X924 from the Mendrzyk et al. (2006) dataset. The overall density function (in black) is the result of the convex combination of the density functions of the four groups present - *i.e.* the four regions having different copy numbers.

4.3.1 Simulations setup

In order to generate biologically realistic microarray data with `aSim`, the simulation parameters of 225 arrays issued from 5 publicly available microarray datasets (see Table 3.3, page 51) were extracted. An example of these parameters for the Veltman et al. (2003) "p9" sample chromosome 2 are presented in Table 4.11. Using these parameters, series involving the

chr	start	end	offset	sd	state	model
2	1	90899990	-0.14	0.10	0	normal
2	95694714	173121000	-0.14	0.10	0	normal
2	173121001	212567000	0.41	0.06	1	log.normal
2	212567001	242951149	-1.06	0.09	-1	log.normal

Table 4.11 – Parameter set to simulate the Veltman et al. (2003) arrayCGH "P9" sample chromosome 2. The second half of the q arm presents first a single copy gain, then a complete loss of its telomeric part.

225 chips representative for the datasets were simulated. In total more than 14,000 simulations were performed. For the expression profiling, the parameter extraction process was stringent, to ensure a high similarity between the original and simulated data and prove the simulators principle. For the arrayCGH data, the `GLAD` (Hupé et al., 2004) algorithm was used to extract the simulation parameters, to prove the simulators concept of testing algorithms. For all the simulations, the simulator default distribution models, i.e. *Gaussian* and *log-normal* were used.

4.3.2 Data comparison

Independently of the platform being simulated, the expectations are that the original and simulated data agree and correlate with each other. Spearmans rank correlation coefficient was computed to compare the simulated data with the original data. To assess the similarity between the original and the simulated data, two criteria were computed: the **Limits of Agreement** (LOA) (Bland and Altman, 1999) and the **Total Deviation Index** (TDI) (Lin et al., 2007). For both the arrayCGH and EP simulation method, series of 5 simulations per sample were performed.

Expression profiling simulation: A high similarity between the original and simulated EP data, due to the stringent parameter extraction, is observed, as exemplified in the Figure 4.21.

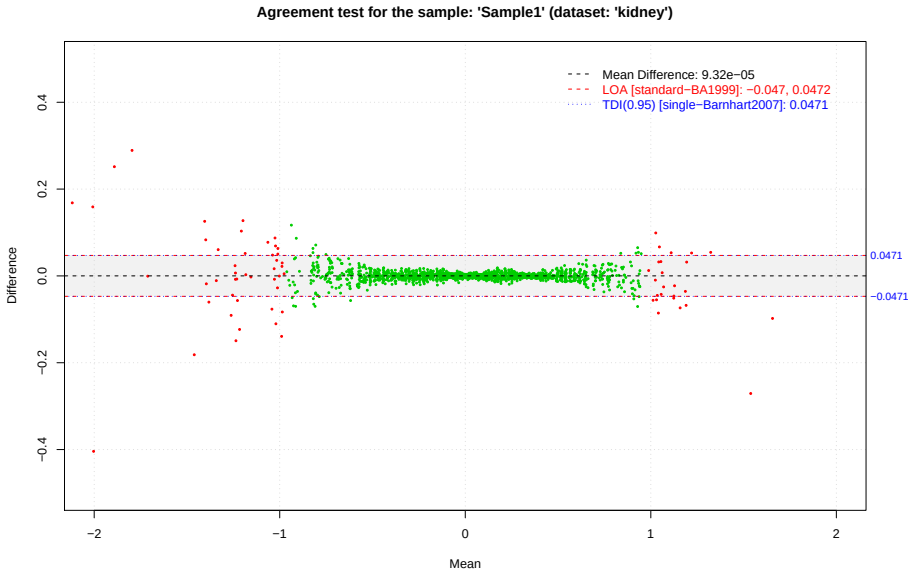


Figure 4.21 – Agreement tests (LOA and TDI performed on the first Sültmann et al. (2005) EP sample. The mean difference is close to zero, both agreement tests yield similar results. The variability increase for larger mean difference is expected, as the variance in de-regulated gene clusters is higher than in unchanged ones.)

All the datasets show a high Spearmans rank correlation coefficient (close to 1), and a small TDI (close to 0). There is no systematic mean shift in the data (close to 0) and, as expected under this condition, the TDI tolerance interval [-TDI, TDI] and the LOA are very similar. This shows that the simulator principle is able to simulate data highly similar to the original one.

Array CGH simulation: For the arrayCGH simulations, the expectations are the same. The TDI is small, there is no systematic mean shift in the data and the TDI tolerance interval [-TDI, TDI] is almost identical to the LOA: the simulated data "agrees" with the original data. However, the Spearmans rank correlation coefficients are very variable: $0.13 - 0.8$. Actually, the coefficients correlate with the chip variability (Pearsons correlation coefficient = 0.84 (95% **Confidence Interval** (CI): $0.77 - 0.9$)), which is a measure of the standard deviation of a chip, based on its number of aberrations and their size as in:

$$\text{variability} = \text{SD}(\mu_1 \times w_1, \mu_2 \times w_2, \ldots, \mu_n \times w_n)$$

with n being all the regions present on the chip, μ their average log expression and w a representation of their size. For example, an arrayCGH profile with a few small aberrations will have a low variability (hence a low correlation coefficient) and one with numerous wide aberrations will have a higher variability (therefore a high coefficient). To explain this correlation, consider the data as having a structure made of two layers: the first one describing the variation introduced by the aberrations and the second one describing the additional variation introduced by the "noise" in these regions. In a chip with a low variability, the overall "noise" effect is greater and results in a smaller correlation coefficient. Since the Spearmans correlation is based on ranks, one could think that they would vary more in a chip with a low variability, where all the data are close to each other, compared to a chip with a higher variability. Using the Pearsons correlation instead, leads to the same conclusion (data not shown). However, we observed that the Pearson correlation coefficient between the original and the replicates of the simulated data has a narrow distribution; it only spans a small range, and it is specific for every chip. This can be explained by the dual structure of the data: the aberrations are responsible for the chip specificity of the correlation coefficient and the "noise" is responsible for the narrow distribution. This property describes the relation between the original and simulated data. If the parameter extraction by `GLAD` and the simulators assumptions are correct, this property should be conserved in a set of simulated data.

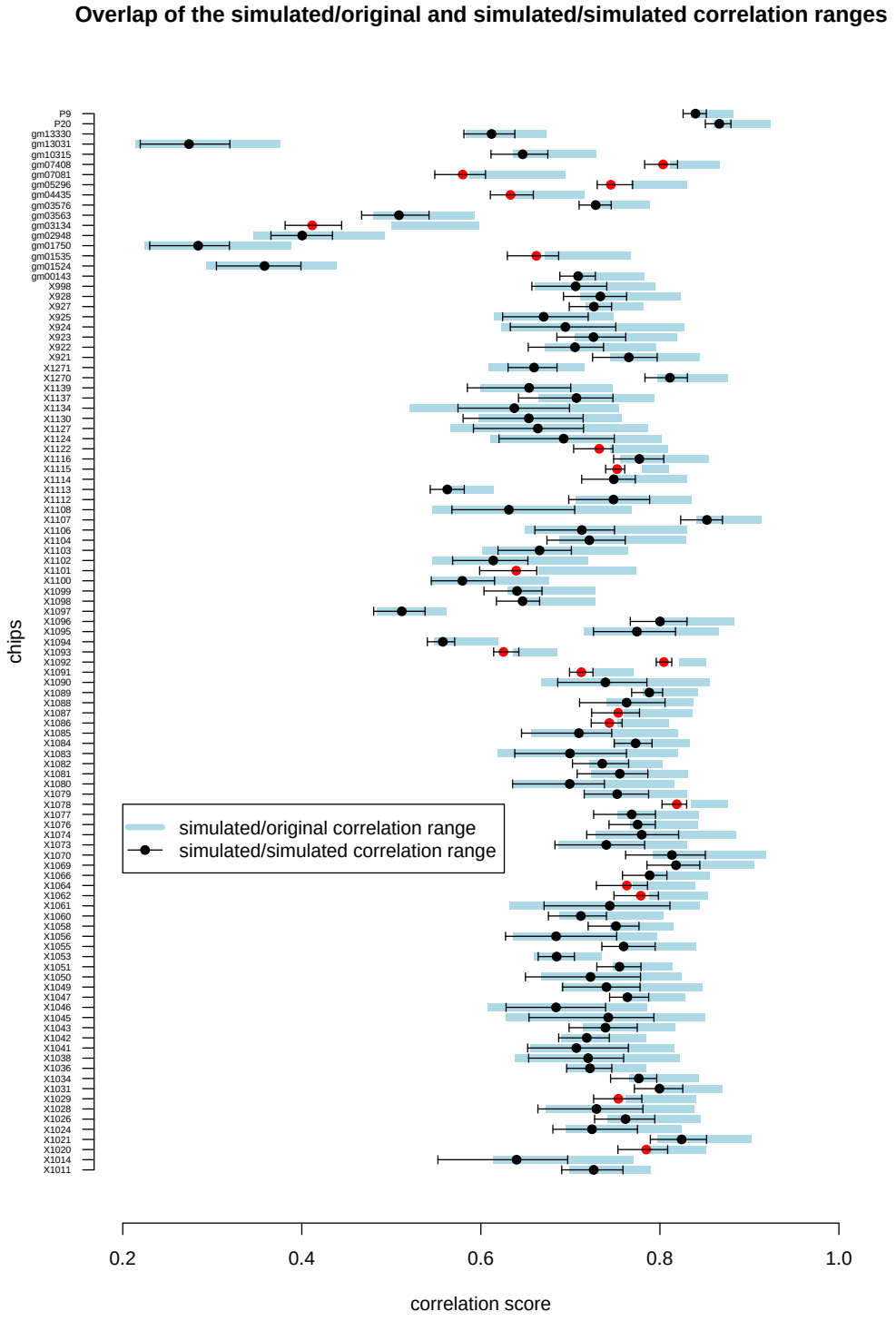


Figure 4.22 – Comparison of the simulated-simulated and simulated-original Pearson correlations of every arrayCGH samples. 16% of the samples. shown in red, fail this test.

The expectation is that the correlation range of independently simulated data (e.g. simulated-1 versus simulated-2) will overlap with the correlation range of these simulated data against the original one (e.g. original versus simulated-1 and original versus simulated-2). A method using this property was implemented, and for every arrayCGH chip a hundred simulations were performed (a total of $10,500$ simulations). The range spanned by the correlation coefficient between the original and simulated data was compared with the one of the correlation coefficient obtained pair-wise between the simulated data. These tests assess whether the data model assumption used during the parameter extraction (implemented by `GLAD`) and the simulation are valid. As shown in Figure 4.22, only 84% of the samples pass the correlation range test, suggesting that either both or one of the `GLAD` or `aSim` assumptions are not optimal for every sample.

4.3.3 Parameter extraction limits

For the arrayCGH data, in every dataset some of the correlation ranges (simulated/simulated versus original/simulated) did not overlap. A deeper analysis done by tuning the `GLAD` parameters to over-fit them to the original data was performed to find out the best simulations parameters. In this context, two `GLAD` parameters are important: "bandwidth" (the maximal bandwidth for the **Adaptive Weight Smoothing** (AWS) number of iterations) and "qlambda" (the scale parameter for the AWS stochastic penalty). Using different values of these parameters for extracting the simulators input for all the affected chips allowed defining which set of the `GLAD` parameters maximizes the correlation coefficient between the original and simulated data as shown on Figure 4.23 for the Snijders et al. (2001) "gm03134" sample. These optimal `GLAD` parameters were then used to extract the simulators input. The correlation range test was performed again for the chips failing at the previous test with the new simulators input. As shown in Figure 4.24 for the fibroblasts gm03134 sample, the simulated/simulated and simulated/original ranges are then overlapping. This is a proof of concept that `aSim` can be used to assess the parameter range of existing or new algorithms and can be used to benchmark them.

4.3.4 Performance

In addition to the ability of reproducing data, the minimal time required for their generation is a critical parameter. To assess the `aSim` performances, expression profiling simulations were done with an increasing number of probes. Times for $1k, 5k, 10k, 20k, 50k$and$100k$ probes simulation were recorded (5 replicates each) on a desktop PC (2.8GHz, 1GB RAM) and on a Quad-Core server (2.7GHz, 8GB RAM), using one CPU. The results shown in Figure 4.25 indicate that `aSim` has a time complexity $O(n^2)$.

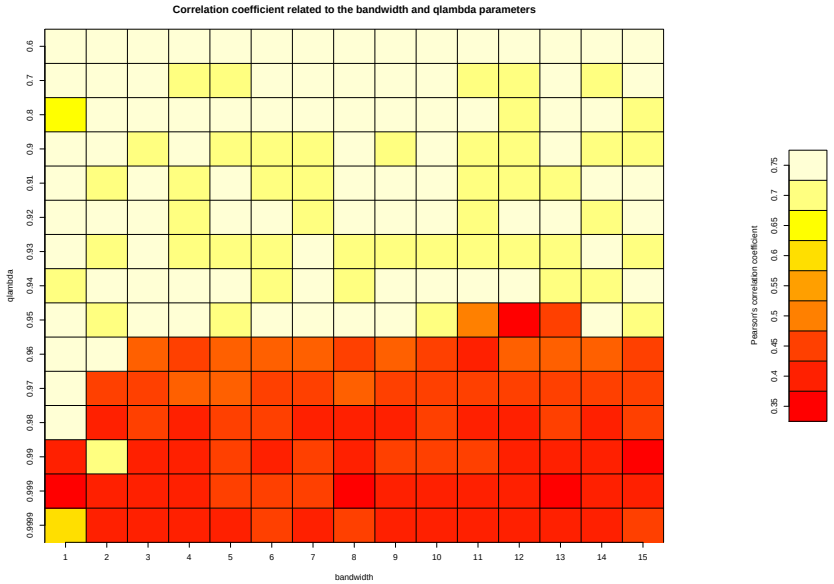


Figure 4.23 – Identification of the "badwidth" and "qlambda" parameters value that maximizes the correlation score for the Snijders et al. (2001) "gm03134" sample. In that example, it is important that the qlambda value is lower than 0.94

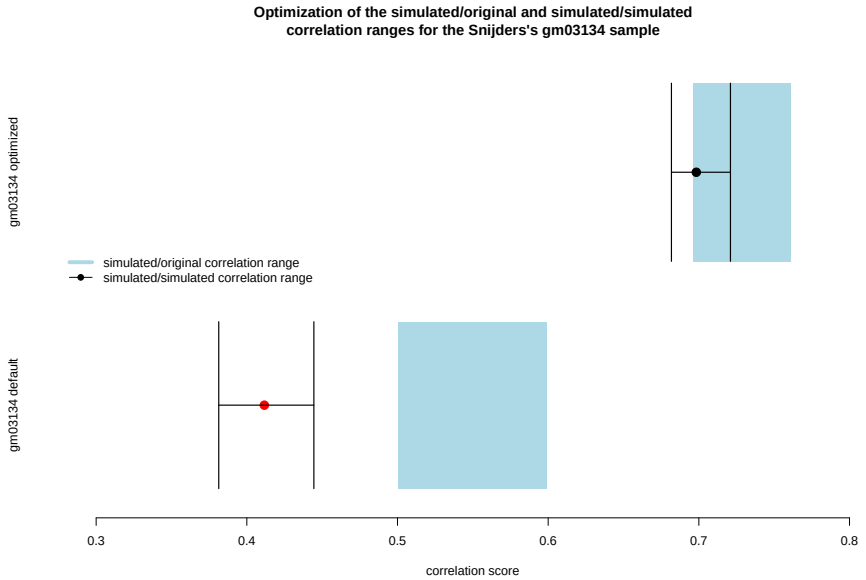


Figure 4.24 – Comparison of the simulated-simulated and simulated-original Pearson correlations for the Snijders et al. (2001) "gm03134" sample simulation using refined parameters

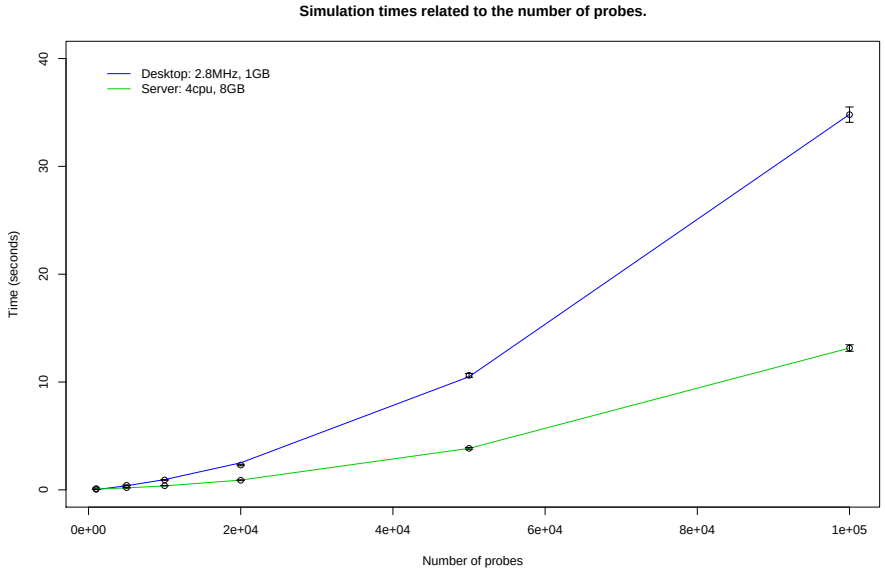


Figure 4.25 – `aSim` performance

4.3.5 Customized simulations

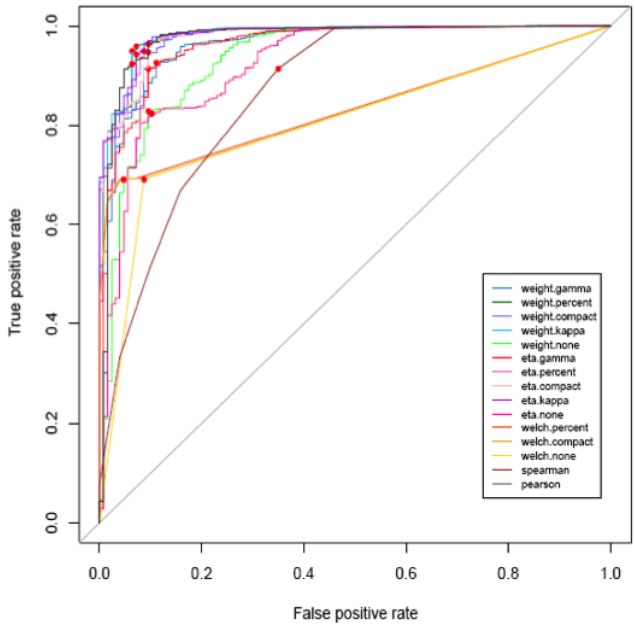
The `aSim` default parameters - *e.g.* number of genomic aberrations - reflect the analysis outcome of the ependymoma (Mendrzyk et al., 2006) and breast cancer (Thuerigen et al., 2006) experimental data. It is possible to customize any level of the simulation process, *i.e.* all the parameters can be modified to alter the data simulation, *e.g.* increasing the number of differentially expressed gene clusters. In addition, the default data models can be replaced by custom ones. This flexibility permits the implementation of virtually any kind of data-model to simulate microarray data, an important point for testing the different statistical models used for the integrative analysis, see section 4.4, page 103.

4.4 Microarray integrative analysis

First, the different developed methods were evaluated using simulated data. Then, after their previous introduction (see section 4.2.3, page 92) and for further demonstrating the advantage of performing integrative analyses, the arrayCGH and EP data from the Zielinski et al. (2005); Grasemann et al. (2005); Gratias et al. (2007) studies were analyzed using the method selected in the first step.

4.4.1 Method selection

Determining the best correlation method: For integrating the array-CGH and EP discrete and continuous data together, five methods have been identified in the literature and extended: *Eta, Pearson, Spearman, Weight and Welch*. To determine, which method was the most sensitive and most specific, they were all benchmarked against a dataset simulated using the package `aSim` (see section 3.3, page 48 and 3.4.2). By increasing the dataset noise stepwise and recording the specificity and sensitivity of the different methods for every of these conditions, ROC curves were constructed, as shown in Figure 4.26. At low noise level, the *Pearson, Spearman* and *Welch* performed poorer than *Eta* and *Weight*. At higher noise level, the results obtained from that first set of methods were actually random (*i.e.* they overlaid the ROC curve diagonal). For the *Eta* and *Weight* group, the *kappa* and *percent* correction functions were the most accurate, while *compact* had an higher FPR and *gamma* had a maximal recall of 75% after which the assignment were random.



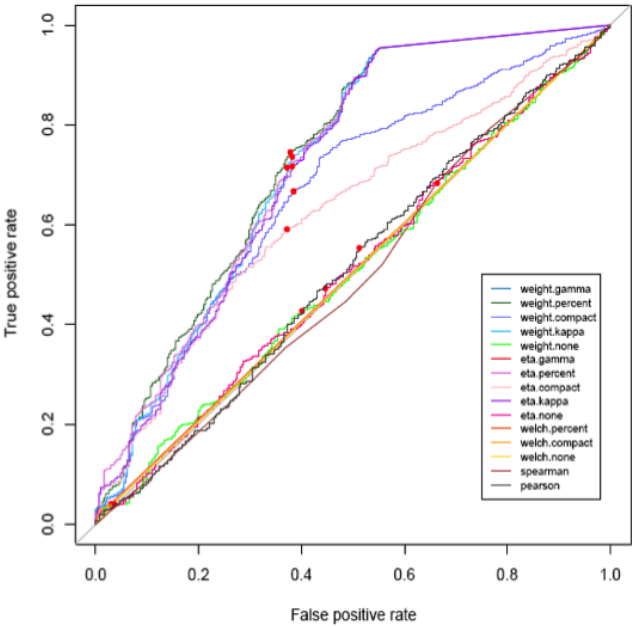


Figure 4.26 – By increasing the noise stepwise in the simulated datasets and recording the TPR and FPR for each conditions, ROC curves were obtained. In the upper panel, the dataset had distinct characteristics whereas these were confused by a high noise level in the lower panel.

To validate these results, the AUC was calculated for every condition and every method and plotted against delta (the inverse of the noise) as shown in Figure 4.27. Based on these results the *Weight* method and *percent*

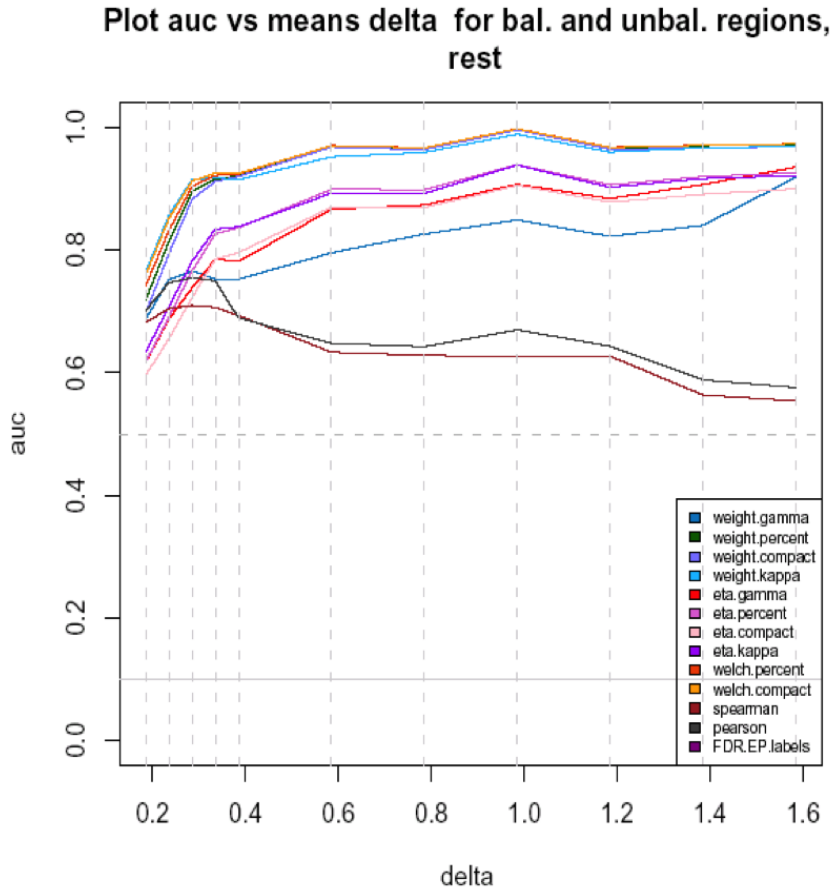


Figure 4.27 – The AUC recorded for every condition for every method was plotted against delta: the inverse of the noise. The weight.percent and weight.kappa had the best recall even in high noise conditions.

correction were chosen for the analysis.

4.4.2 Data pre-processing

Once the method was selected the raw data were pre-processed, requiring different steps for the EP and arrayCGH data, as described in the following paragraphs.

Expression profiling dataset pre-processing: The original EP dataset had the limitation that only a single control experiment was performed. This resulted in a lower detection power as compared to the GSE29683

analysis described previously (see section 4.1.4, page 76), 257 *vs.* 988 significantly differentially expressed probe-sets respectively. 70% of the GSE5222 dataset are common to the GSE29683. The log2 FC for the common probe-sets is significantly enriched for higher absolute values (Student t test p-value = 0.015 and 5.88e-5 for the GSE5222 and GSE29683 respectively) indicating a good agreement between both datasets, but as well the existence of confounding factors. To increase the detection power within the GSE5222 dataset, additional control samples were hybridized, however on HG-U133Plus2 *GeneChip*®s rather than on the HG-U133A as the rest of the dataset.

HG-U133Plus2 and HG-U133A probe-sets concordance: The HG-U133Plus2 *GeneChip*® actually combines updated versions of the HG-U133A and HG-U133B *GeneChip*®s. $22,142 (> 98\%)$ probe-sets of the updated *Ebased* HG-U133A CDF are found in the *Ebased* HG-U133Plus2 CDF. There are 382 probe-sets that are specific to the *Ebased* HG-U133A and are all either of the "multiple" or "dubious" class as shown on Figure 4.28. Altogether they involve only $1,325$ probes (0.5% of the amount of probes present on the *GeneChip*®). The control samples performed on the HGU-

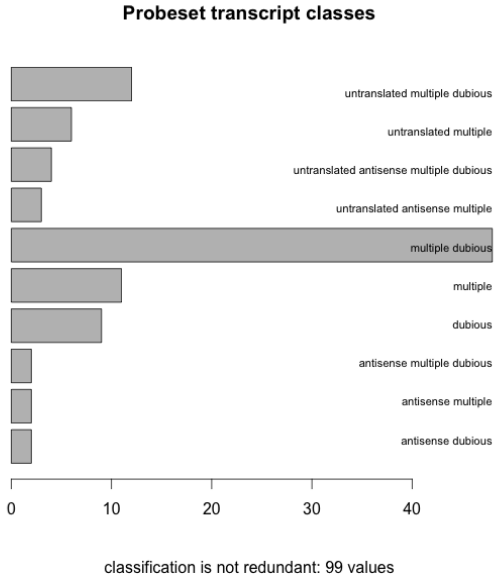


Figure 4.28 – Classification of the HG-U133A specific probe-sets that are not part of the HG-U133Plus2 probe-sets. All of them are either mapping "multiple" loci or "dubious".

133Plus2 platform can therefore easily be restricted to the set common to both platforms.

Control samples comparison: To assess whether the expression observed for the control sample hybridized on the *Ebased* HG-U133A *GeneChip*® was similar to that of the three other samples hybridized to the *Ebased* HG-U133Plus2 *GeneChip*®, their log2 FC was compared using a Pearson comparison, both before and after normalizing the samples between arrays using a *cyclic loess* approach. The results of the normalization can be observed in the Figure4.29, it has only a slight impact on the results of the Pearson

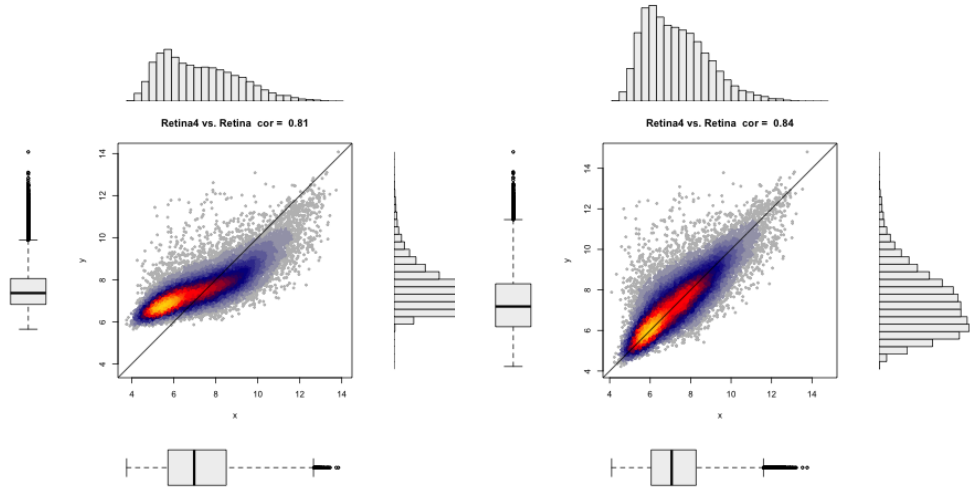


Figure 4.29 – Scatterplot of the log2 FC of the GEO GSE5222 Retina-4 (HG-U133Plus2) *vs.* Retina-1 (HG-U133A) samples. The right panel shows the in-between array normalized data, which slightly increase the Pearson correlation (0.81 to 0.84)

correlation, 0.03 on average. The results of the pair-wise sample correlation is presented in the Table 4.12 and show values expected for biological (≥ 0.8) and technical (≥ 0.9) replicates of good quality. As the 4 samples

	Retina1	Retina2	Retina3	Retina4
Retina1	1			
Retina2	0.94	1		
Retina3	0.94	0.99	1	
Retina4	0.84	0.81	0.81	1

Table 4.12 – Pairwise correlation of the GEO GSE5222 retinoblastoma control samples. The technical replicates (Retina2 and Retina3) show a close to 1 correlation as expected. These show with the other biological replicates (Retina1 and Retina4) excellent correlation too.

correlate nicely, they are all retained for the analyses. All the GSE5222

samples expression values are reduced to the probe-sets list defined previously and normalized between arrays. Performing a DE analysis using a linear model between the "tumor" and "control" samples and comparing the list of significantly differentially expressed genes with the results of the GSE29683 analysis described previously (see section 4.1.4, page 76), showed an increase in the detection power as compared to the similar analysis performed in the first paragraph of the section 4.4.2. 531 probe-sets are found to be differentially expressed, 350 of which are identified by the GSE29683 analysis as well (60%). This time, however, the common subset is extremely significantly enriched in higher log2 FC, p-value $\leq 2.2e-16$.

Defining expression states: The statistics used for the integrative analysis take advantage of both continuous and discrete values. For that reason, expression states - *e.g.* down or up-regulated - were defined using the tumor *vs.* control log2 FC and the corresponding DE p-value. Figure

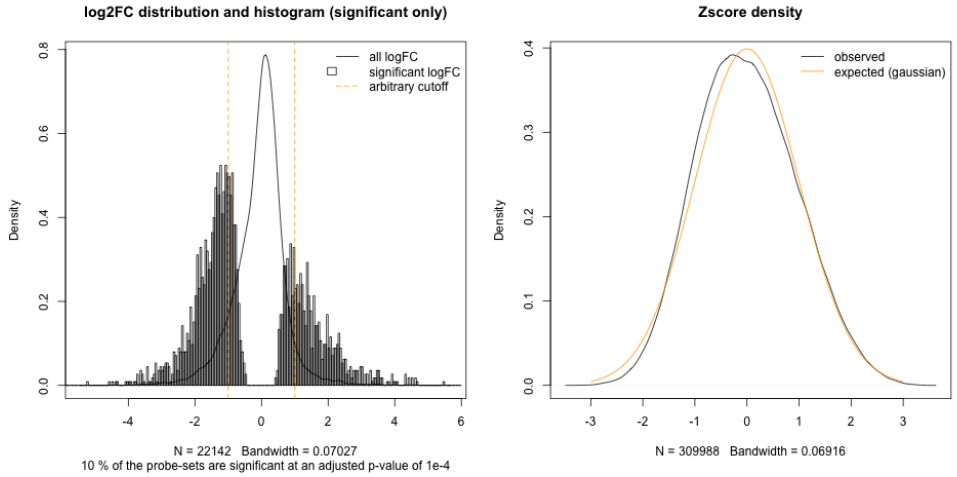


Figure 4.30 – The left pannel shows the density of all the log2 FC (black line), the density of the significantly differentially expressed probe-sets histograms and a set of arbitrary cutoffs (orange line). The right panel shows the observed *Z-score* density and the expected one (standard normal distribution)

4.30 shows in the left panel the log2 FC of the probe-sets significantly differentially expressed at an adjusted p-value cutoff of $1e^{-4}$, roughly 10% of all probe-sets and show the advantage of a selection based on a DE p-value cutoff rather than using an arbitrary log2 FC cutoff (such as the orange lines). Indeed, many smaller log2 FC are kept and some larger ones are discarded (*e.g.* probe-sets where the variance due to outliers is so high that even a large log2 FC is not significant), increasing our detection power. The right panel shows the distribution of the *Z-scores* calculated on the same

values. Another validation of the presented approach is the very close fit of the observed *Z-score* distribution (black) with the expected standard normal distribution (*Gaussian*). The proportion of absolute *Z-score* larger than 2 is expected to be 5%. The implemented procedure is conservative as only 4% of the *Z-score* are above that value, and the threshold has to be reduced to 1.85 to get a 5% proportion. This value corresponds in the normal distribution to a 6.9% proportion, indicating that the designed approach only fails to explain 1.9% of the variability of the data. The proportion of the different states is presented in Table 4.13. The negative states rep-

State	-2	-1	0	1	2
Occurence	274	23,553	265,770	20,158	233

Table 4.13 – GEO GSE5222 expression states occurence

resent a down-regulation observed for the probe-set, and the converge an up-regulation. The $-2, 2$ states represent larger variations in probe-set expression, *i.e.* they are outside the 95% confidence interval defined by the *Z-score*.
The EP dataset is ready for the integrative analysis. In the next paragraph the arrayCGH data preparation is detailed.

ArrayCGH pre-processing: As the microarrays used for the arrayCGH experiments differed from the ones used for EP, additional pre-processing steps were required, namely the sex-probes rescue and the missing data imputation that are described in the following paragraphs.

Sex specific probes' rescue: As the arrayCGH data used an opposite sex matched sample for the second channel - the microarrays used were dual-channel - it is not possible to decipher directly the CNV state based on the log-ratio. However, a male sample is expected to show a $1X$ gain of the Y chromosome and a $1X$ loss of the X chromosome. Based on this assumption (and the reciprocal one for female samples), the log2 FC values were corrected. This allows the identification of a loss on chromosome X for the M22808 sample: dim(Xp11.22q23) as reported in Zielinski et al. (2005) (Case #3 in Table I, p.296). The corrected log2 FC are shown in Figure 4.31.

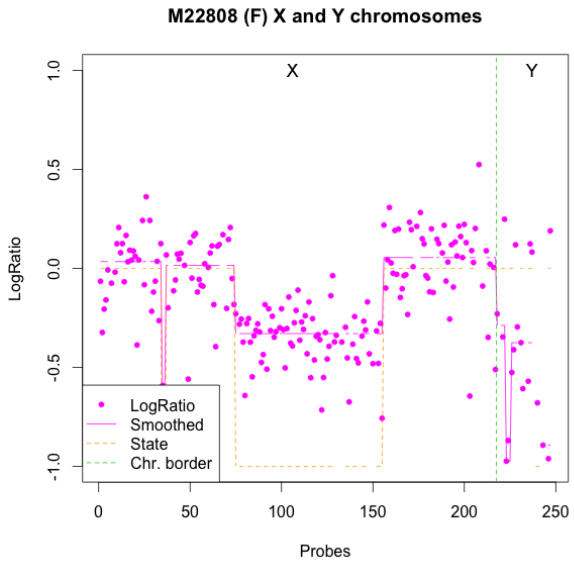


Figure 4.31 – The rescue of the sex chromosome probes allowed the identification of a loss on chromosome X for the sample M22808.

As can be seen from this figure, the values for the chromosome Y are very sparsed and very variable and this independently of the sample gender. This variability is more likely due to the high heterochromatin content - roughly 50% of the $59Mb$- of the Y chromosome and the high similarity of the rest of it with chromosome X (pseudo-autosomal regions) than to real CNV. Moreover, there are only 30 probes covering it: half the number of probes for chromosome 22, which has a similar euchromatin content (n = 75). Finally, there are only 45 known genes reported on chromosome Y. For these reasons, the values for chromosome Y are ignored for the rest of the integrative analysis.

Data imputation: As the arrayCGH experiment was based on spotted microarrays, a variable proportion of spots had been manually flagged by the experimentator as being innapropriate for further analyses (*e.g.* presence of an air bubble, disformed shape, inconsistent fluorescence, *etc.*): between 0.003% and 17% of the 6318 probes present on the microarray. These missing values were imputed and across the 14 arrays only 21 probes could not be imputed, with at most 4 per array. Imputing the data has no effect on the log2 FC distribution (Welch two-sample t-test p-values: $[0.84, 0.99]$).

Virtual probe imputation: As the arrayCGH spotted microarrays only have a partial coverage of the genome (31%), 5458 virtual probes were created spanning the loci between probes and physical barriers such as cen-

tromeres, in order to increase the number of arrayCGH - EP pairs taken into consideration by the integrative analysis. The values for these probes were imputed as described in section 3.4.1, page 56. Between $10 - 14$ virtual probes per sample could not be imputed, a proportion of 0.2%. The imputed value distributions of the virtual and experimental probes are very similar to the respective mean $\mu = 0.02$ and 0.007 and SD $\sigma = 0.18$ and 0.19, although these two distributions have a significantly different mean (Welch t-test p-value of $2.2e^{-16}$, with a 95% confidence interval of: $0.12 \leq$ mean difference ≤ 0.16).
The arrayCGH data are now ready as well for the integrative analysis that will be described in the next section.

4.4.3 Data Analysis

Using the obtained overlayed arrayCGH and EP data (see section 3.4.2, page 56), a pre-analysis is performed using the "discrete" states and the calculated contingency table. Then, the entire set of data is used for more complete statistical analyses.

Contingency table analysis: The contigency table is obtained by summarizing the arrayCGH and EP data (see Table 4.14). No obvious effect can

arrayCGH/EP	-2	-1	0	1	2
-2	0	3	42	3	0
-1	13	1,110	10,903	541	2
0	351	30,663	337,055	25,138	269
1	15	1,771	27,721	2,809	55
2	0	124	2,506	296	6

Table 4.14 – Contingency table of the arrayCGH change in copy number and EP change in expression states.

be identified, *i.e.* almost every arrayCGH-EP pairs has at least one state value of 0. To discover and statistically evaluate the presence of any bias in the data distribution, the Pearson signed χ^2 contribution - a measure of the residuals from a model assuming the independence of the contingency table variables - is calculated. Its result is shown in Figure 4.32.

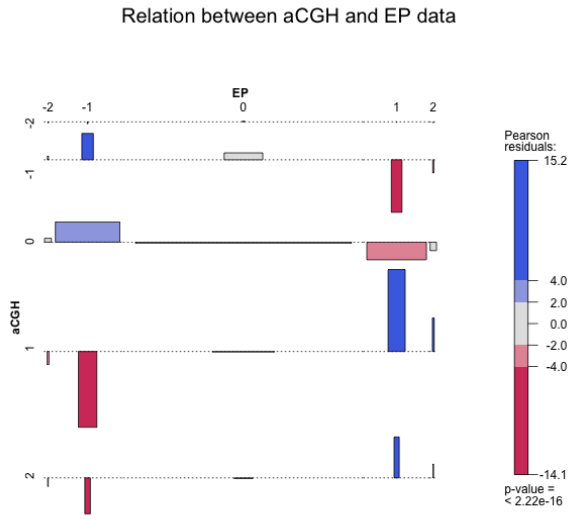


Figure 4.32 – Association plot of the contingency table showing the residuals of the independence model. The area of every rectangle is proportional to the difference between the observed and expected frequencies. The direction of that rectangle to its baseline indicates whether the contribution is higher (above) or lower (below) than expected. In addition the rectangles are colored according to the Pearson signed χ^2 contribution: from red (negative) to blue (positive).

It becomes visible that there is a positive contribution when the two states have the same sign (*e.g.* a copy number gain and an increased expression) and a negative contribution for the reciprocal cases. This is as expected - *i.e.* when a locus is gained, one expects an increase in expression - and validates the integrative analysis approach: the two considered variables are not independant when the states are not equal to 0.

In the next paragraphs, additional analysis are performed to further validate this integrative analysis approach.

Correlation results validation: The results obtained from the selected method - the w correlation score - are first evaluated for their validity.

The w correlation score is sound: The w correlation scores obtained (see equation 3.5, page 58) are compared to their calculated significance (*i.e.* p-value and **False Discovery Rate** (FDR)). The results are as expected: large score (in absolute value) correlate with the best significance score, see Figure 4.33. Moreover, it underlines the importance of calculating the FDR using permutations, as this removes a high number of false posi-

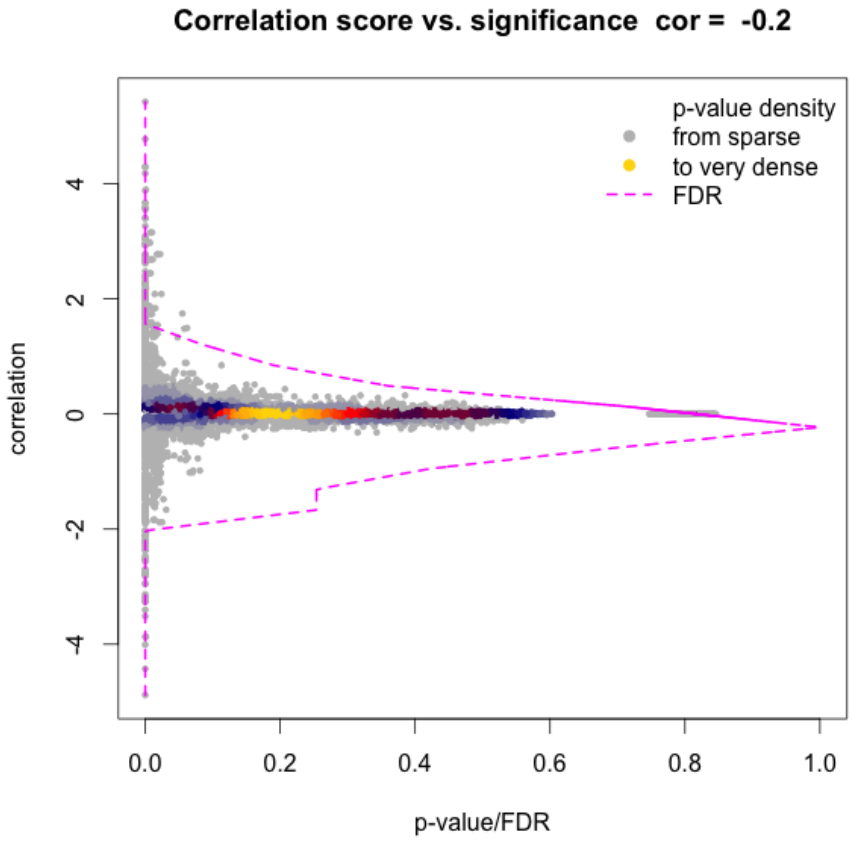


Figure 4.33 – The w correlation score of every arrayCGH - EP pair is compared to its significance values: p-value and FDR. The p-value density is described by the gray (sparse) to gold (very dense) scale. The FDR is represented as a dashed line.

tives that would be kept when using just a p-value based threshold. Given this result, an FDR cutoff of 10% is selected for the rest of the integrative analysis. As mentioned before, chromosome Y values are ignored (see section 4.4.2, page 109).

Overlap with the expression profiling results: At that FDR threshold, 271 significantly correlated pairs are identified involving 210 EP probe-sets. Comparing this set with the significant sets identified in the EP results from the section 4.1.4, page 76 and section 4.4.2, page 105 using the same adjusted p-value revealed a common subset of 55 probe-sets (26%). This indicates that the expression change of the remaining 74% cannot be identified by just an EP approach, *i.e.* the gene variation between control and sample is not sufficient to be deemed significant with the threshold used. The Table

4.15 shows the contingency table of the correlation score. Positive correla-

arrayCGH/EP	-1	1
-1	5	1
1	33	132

Table 4.15 – Contingency table of the w correlation score sign for the array-CGH - EP pairs based on their observed status. The majority (80%) behaves as expected while the remaining show either positive ($n = 1$) or negative ($n = 33$ compensation

tion scores are expected, while negative scores indicate putative regulation of the gene expression, see section 1.2.2, page 30. These will be analysed in more details in the next paragraph.

Distribution of the w correlation score: About 20% of the w correlation scores are negative - *e.g.* showing an expression decrease while that region is frequently gained - with all the occurences but one located on chromosome 6. The exception is on chromosome 16: the gene *SALL1* is over-expressed in a LOH region. *SALL1* is involved in transcription regulation and in *Wnt* signaling among other processes. *SALL1* is identified by the EP performed previously (as the 292 most significant gene), but the present results give it a much higher importance. For the identified genes located on chromosome 6, many are associated with immune response (*AIF1, HLA-A, C, D* and *E, MICA, PSMB8* and *TAPBP*) and signal transduction in general (*e.g. GMPR, RCAN2*). The remaining ones are mostly involved in trancription (*e.g. FOXC1, FOXF2*) and cell cycle regulation (*e.g. IDE4, NEDD9*). Only 40% of these genes were identified by the EP analysis.

The probe-set "multiple" class is over-represented: Unexpectedly, the 210 probe-sets contained a lot of "multiple" mapping, see Figure 4.34. These were manually curated and in most cases, these "multiple" mapping situations resulted from the presence in the reference genome of three different haplotypes of a chromosome 6 region: *6, 6COX and 6QBL.* Additionaly, the few "antisense" and "untranslated" probe-sets were curated. Interestingly, a number of them adresses (processed or not) pseudogenes: *SUCLA2P, PTMAP1* and *PIP5K1P1*, a newly retrotransposed gene: *NUP50P2* and one **small nuclear RNA** (snRNA): U6. In total 203 probe-sets were kept, while for the 7 filtered ones, additional *in-vitro* experiments would be needed to validate them - for example to verify whether or not the pseudogenes are expressed - and determine what their role might be; *e.g.* transcription regulation by processed pseudogenes has been reported in the literature (Tam et al., 2008).

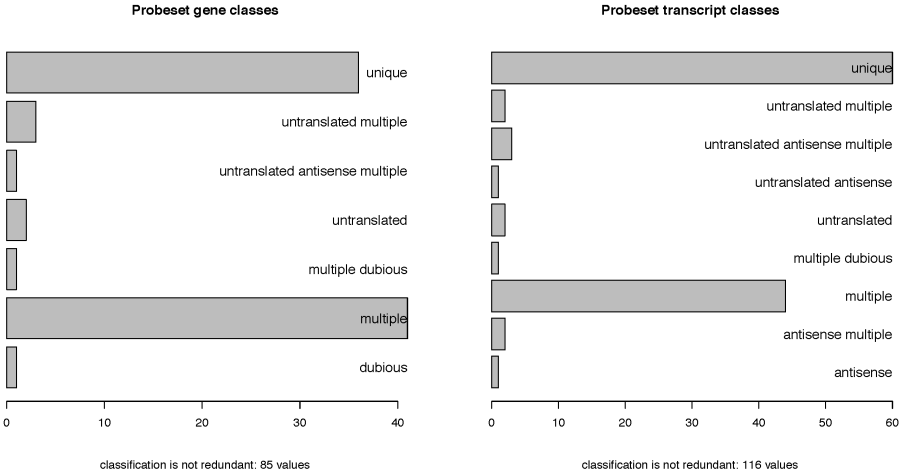


Figure 4.34 – Classification of the 210 selected EP probe-sets. The left panel shows the classification for the gene probe-sets, while the right one shows that of the transcript probe-sets. Both have a high proportion of "multiple".

This set of results display the sensibility of the undertaken integrative analysis approach. In the following paragrahs, the biological significance of the obtained result is presented.

Correlation results analysis: The obtained significant arrayCGH - EP pairs are retrieved and their implication on the biology of retinoblastoma analyzed.

Chromosome 6 is enriched in significant pairs: The analysis revealed that 271 arrayCGH and EP pairs are significant at that FDR cutoff. The chromosomal localisation of the arrayCGH corresponding probes ($n = 125$) is shown in Table 4.16 and agrees with the known reported aber-

chromosome	1	2	6	16
occurence	21	2	93	9

Table 4.16 – Summary of the chromosome associated with a significant correlation of the arrayCGH and EP data at a cutoff of 10% FDR

rations (see Table 4.9, page 91). Moreover, this analysis reveals that the chromosome 6 is the only one among the 4 to be significantly enriched for correlated pairs, possibly implying an higher importance of that chromosome in the tumorigenesis. Indeed, based on the probability derived from the data that 1/10 arrayCGH probe shows a CNV, knowing that 405 probes cover chromosome 6 and assuming that the probes are independant, the

likelihood to record 93 affected probe pairs is $2.1e^{-14}$. The density distributions and the probability for every chromosome is shown in Figure 4.35; chromosome 6 only showing an enrichment in significant correlated pairs.

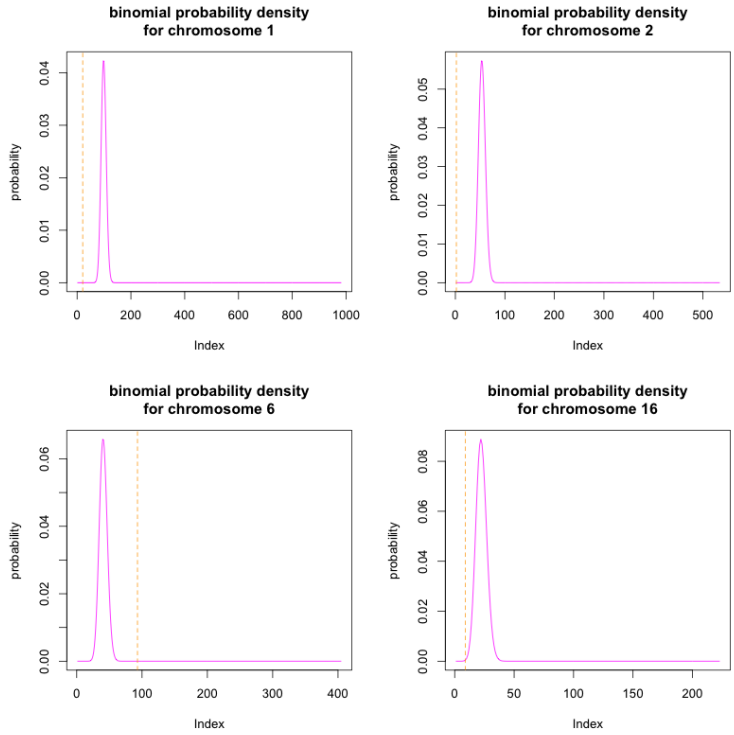


Figure 4.35 – The binomial density of the 4 chromosomes harboring significantly correlated pairs is shown in purple. The dashed orange vertical lines represent the number of occurence recorded for that chromosome (see Table 4.16). Only chromosome 6 shows an enrichment, *i.e.* its number occurence is to the right of the density peak.

The chromosomal regions containing the significant pairs are: *1q21.3-1q44, 2p25.2-2p25.1, 6p25.3-6p11.2* and *16q12.1-16q23.3*, in agreement with the conclusions from Zielinski et al. (2005). Only the chromosome 2 region is smaller (the complete chromosome 2p arm gain was reported) and the chromosome 16q gain does not appear to be of statistical significance here. The minimal overlap region reported on chromosome 2p by Zielinski et al. (2005): *2p24.2-2p24.1* is located slightly downstream of the region identified here.

Known targets: DEK and E2F3 are recalled: Among the identified genes, *E2F3* and *DEK*, reported by Grasemann et al. (2005), are as well identified as having the 5^{th} and 14^{th} highest absolute coefficient in the present study, respectively.

The integrative analysis reveals antisense regulation: Two of the probe-sets map antisense to transcripts 3'UTR, while there's no evidence for transcription on that strand in downstream regions of the gene loci. The genes are *DST* and *EEF1E1*, the first one involved in cell cycle arrest and the second one being a negative regulator of the cell cycle and a positive regulator of apoptosis. In the hypothesis of a negative regulation by antisense transcription, these results are interesting but would need to be validated *in vitro*.

Gene ontology enrichment analyses: The integrative analysis should have removed some of the confusing factors observed in the EP (see section 4.1.4, page 76). To verify this, one approach is to perform GO enrichment analyses on the different gene subsets.

gene.symbol	coef	p.value	fdr.local
KIFC1	5.42	0.00	0.00
RCAN2	-4.88	0.00	0.00
GMNN	4.78	0.00	0.00
HLA-DRA	-4.43	0.00	0.00
E2F3	4.29	0.00	0.00
ELOVL2	4.18	0.00	0.00
HLA-DPA1	-4.01	0.00	0.00
MCM3	3.90	0.00	0.00
ATAT1	3.88	0.00	0.00
FOXF2	-3.87	0.00	0.00
HNRNPL	3.67	0.00	0.00
CDKAL1	3.58	0.00	0.00
PCSK2	3.55	0.00	0.00
DEK	3.41	0.00	0.00
NEDD9	-3.40	0.00	0.00
HIST1H4C	3.39	0.00	0.00
SOX4	3.27	0.00	0.00
DSP	-3.26	0.00	0.00
ID4	-3.25	0.00	0.00
BTN3A3	-3.18	0.00	0.00

Table 4.17 – The 20 most strongly associated arrayCGH - EP pairs, *i.e.* having the highest absolute correlations are listed. *DEK* and *E2F3* have been previously reported by Grasemann et al. (2005). See appendix D, page 206 for the complete gene list.

The 203 curated probe-sets identify a total of 171 genes, 34 (20%) having a negative w score and the remaining 137 a positive one. See Table 4.17 for the most strongly associated pairs and Table D, page 206 for all of them.

The genes identified by several probe-sets show consistant w scores and only the largest w score was kept when reporting the values in these tables. The tight variability of these scores originating from different probe-sets is a further validation of both the approaches at generating the probe-sets (*i.e.* the CDF creation) and performing the integrative analysis.

The GO enrichment analysis performed on the 171 identified genes reveals a clearer picture than that obtained by performing the same analysis on the previously obtained EP data (section 4.1.4, page 76, GO analysis results not shown). As shown in Table 4.18 for all gene ontologies, and in Figure 4.36 for the "biological process" one, only a handful of processes are identified, among which the cell cycle, angiogenesis and immune response - three of the hallmarks of cancer - are represented. Moreover, all the angiogenesis related genes are upregulated, while the immune response ones are negatively regulated.

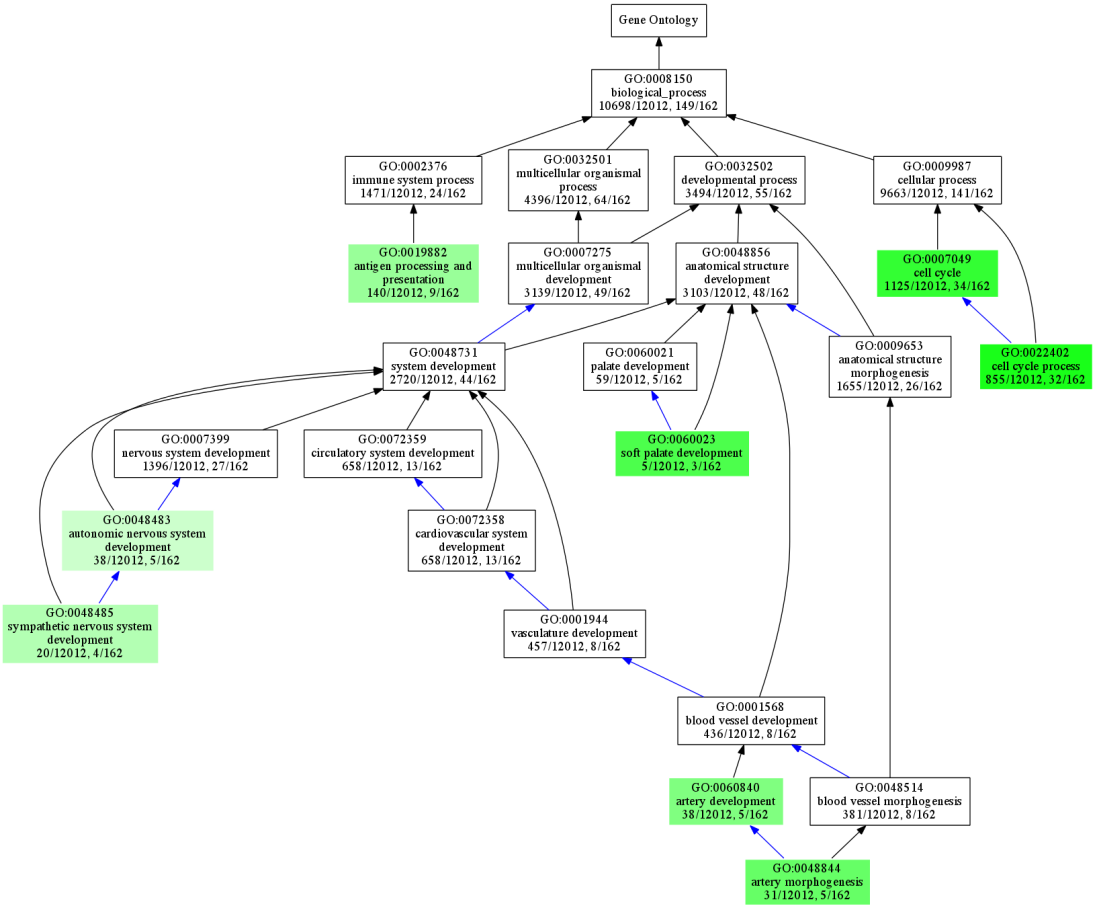


Figure 4.36 – GO enrichment analyses of the genes associated with the integrative analysis significant arrayCGH - EP pairs. Only the "biological process" ontology is shown in the figure, the results for the other ontologies are presented in Table 4.17.

GO term	ontology	population	study
cell cycle process	BP	855	32
MHC protein complex	MF	32	7
cell cycle	BP	1125	34
soft palate development	BP	5	3
organelle	CC	7641	126
endoplasmic reticulum membrane	CC	559	16
alpha DNA polymerase:primase complex	CC	5	3
macromolecular complex	MF	2978	62
artery morphogenesis	BP	31	5
artery development	BP	38	5
antigen processing and presentation	BP	140	9
sympathetic nervous system development	BP	20	4
nucleus	CC	4348	93
autonomic nervous system development	BP	38	5
integral to lumenal side of endoplasmic reticulum membrane	CC	10	4
DNA replication, synthesis of RNA primer	BP	2	2
organelle part	CC	4730	85
nuclear outer membrane-endoplasmic reticulum membrane network	CC	573	16
RNA polymerase II transcription coactivator activity	MF	18	4
intracellular	CC	9003	137

Table 4.18 – GO enriched terms. The population represents the total number of genes present on the microarray that are associated with the term. The study is the number of genes present in the integrative analysis significant subset. Even at an 0.1 adjusted p-value (Benjamini-Hochberg) threshold, some highly generic or very specific terms *e.g.* "intracellular", "alpha DNA polymerase:primase complex" are identified. The terms not falling in either categories are highlighted in grey. The ontology abbreviations are: Biological Process (BP), Cellular Component (CC) and Molecular Function (MF).

In Table 4.18, despite the stringent 0.1 adjusted p-value (Benjamini-Hochberg) cutoff, some very generic terms (*i.e.* the population size is large) - *e.g.* "nucleus" - as well as some very specific terms (*i.e.* the population and study sizes are both small and very similar) - *e.g.* "soft palate development" - are still identified. This is a typical GO analysis bias mainly due to incomplete gene annotations. For clarity, the most relevant terms have been highlighted in grey in the table.

These results conclude the integrative analysis performed on matched retinoblastoma arrayCGH and EP samples. Its displayed sensitivity and the sensibility of its results, encouraged me to conduct the same kind of analysis across tumor kinds, *i.e.* performing a comparative analysis, to further assess the developed methods. This is the topic of the next section.

4.5 Comparative analysis

This section describes the comparative analysis of retinoblastoma and osteosarcoma tumors. The pre-processing is similar to that performed for the retinoblastoma EP dataset used for the integrative analysis above. The selected statistical method and parameters used are identical as well and hence not presented here. The next sections present briefly the results of the data pre-processing before focusing on the biological significance of the obtained results more in details.

4.5.1 Data pre-processing

As there was no suitable control for either the retinoblastoma or osteosarcoma dataset, the UHRR sample was used as a control for both EP analyses. The possible effect on the expression state calculation for both datasets was consequently assessed anew.

Expression profiling *vs.* Stratagene Universal Control: Using the UHRR control sample, 40% and 47% of the $50,513$ probe-sets were significantly differentially expressed for the osteosarcoma and retinoblastoma datasets, respectively, at the 10^{-4} adjusted p-value threshold previously used for the GEO GSE5222 retinoblastoma sample. These results show the high sample similarity within the three datasets used: GSE29683, GSE14827 and GSE5350, as well as their heterogeneity with one another. The very large proportion of significant adjusted p-values indicates that using the UHRR sample as a control is sub-optimal. This however does not affect the rest of the analyses as its effects cancel out when comparing the retinoblastoma and osteosarcoma datasets together. Neither, as shown in the following paragraph, does it have an effect on the Z-score calculation.

Defining expression states: The obtained Z-scores are very close to the expected distribution, see Figure 4.37. The difference is smaller than that obtained previously for the integrative analysis (see paragraph 4.4.2, page 108): only 1.4% of the data variability is not captured. The proportion

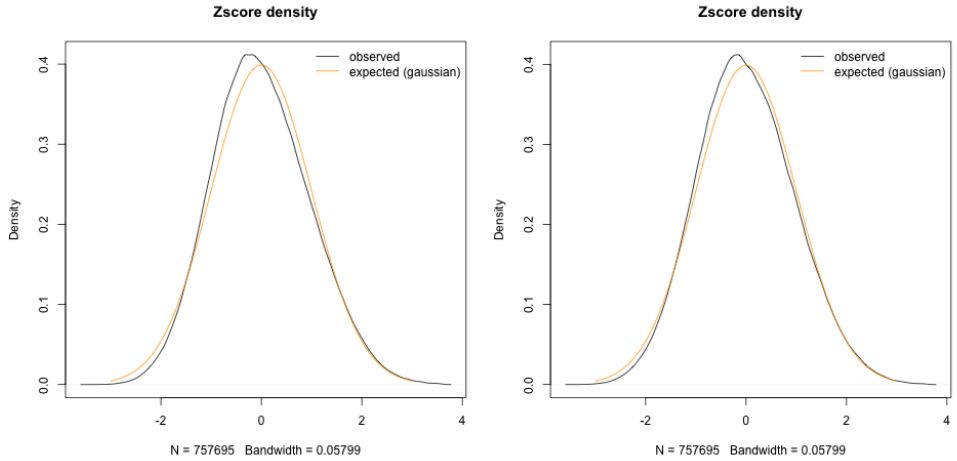


Figure 4.37 – The left and right panel show the observed *Z-score* density and the expected one (standard normal distribution) for the GSE14827 and GSE29683 dataset, respectively .

of the different states for the 15 retinoblastoma samples selected from the GEO GSE29683 dataset are shown in Table 4.19. The 15 selected samples

State	-2	-1	0	1	2
Occurence	0.42	22	52	25	0.72

Table 4.19 – GEO GSE29683 expression states proportion

from the osteosarcoma GEO GSE14827 dataset have the state proportion presented in Table 4.20.

State	-2	-1	0	1	2
Occurence	0.3	18.2	58.3	22.5	0.7

Table 4.20 – GEO GSE14827 expression states proportion

This reveals that the use of the UHRR control did not affect the calculation of the Z-score. Moreover, the proportion of expression states for both tumor types is similar. To further assess their similarity, a contingency table analysis was performed, the resultsof which are described as the first

part of the next section that later on focuses on the biological implications of the comparative analysis results.

4.5.2 Data analysis

Contingency table analysis: The Table 4.21 shows the contingency table of the tumors' expression states. The diagonal appears to be enriched,

osteosarcoma retinoblastoma	-2	-1	0	1	2
-2	26	1,580	1,332	239	8
-1	1,435	82,691	70,300	13,860	381
0	743	44,976	274,801	71,055	2,169
1	92	8,098	92,291	83,195	2,994
2	2	180	2,652	2,478	117

Table 4.21 – Contingency table of the change in expression states for the retinoblastoma and osteosarcoma datasets.

but for a better visualization of the contingency table, the Pearson signed χ^2 contribution - a measure of the residuals from a model assuming the independence of the contingency table variables - is shown in Figure 4.38. The diagonal is significantly enriched, indicating that both tumors are very similar at the gene expression level; an observation reinforced by the fact that every discordant state pairs (*e.g.* a gene being highly expressed in one tumor while lowly expressed in the other one) is strongly decreased.

This results indicate that the use of the UHRR control should not affect the comparative analyses.

Identified probe-sets: $1,171$ significant pairs are identified at a 1% FDR cutoff. As the layout is the same for both datasets in use, this implies that that exact number of probe-sets is significantly correlating between both tumor kinds. In the next paragraphs, these correlations and related probe-sets are described.

Correlation scores contingency table: Unlike for the integrative analysis, both positive and negative correlations are expected, as two different tumor types are compared together. A positive correlation reveals a common feature - *e.g.* a gene similarly deregulated - shared between tumors while negative correlations identify discordant features. The Table 4.22 summarizes the concordant and discordant pairs. Both tumor kinds seems very alike; 93% of the pairs display similar effects. The 18 pairs displaying no change in expression in osteosarcoma were manually checked and none

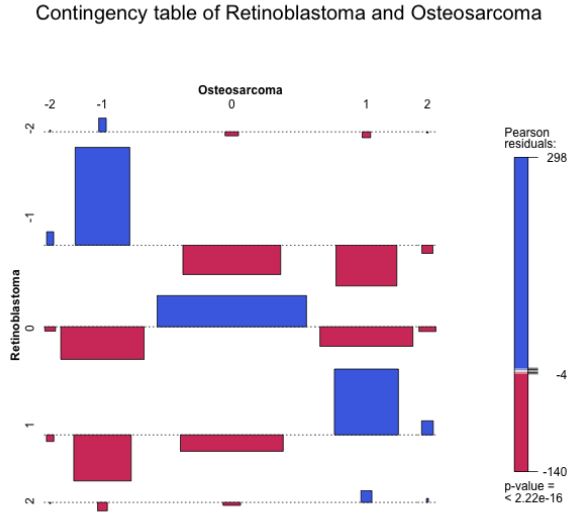


Figure 4.38 – Assocation plot of the contingency table showing the residuals of the independence model. The area of every rectangle is proportional to the difference between the observed and expected frequencies. The direction of that rectangle to its baseline indicates whether the contribution is higher (above) or lower (below) than expected. In addition the rectangles are colored according to the Pearson signed χ^2 contribution: from red (negative) to blue (positive).

retinoblastoma osteosarcoma	-1	1
-1	305	12
0	9	9
1	35	526

Table 4.22 – Contingency table of the w correlation score sign for the retinoblastoma - osteosarcoma pairs based on their observed status. The large majority (93%) agrees between tumor kinds, whereas 65 pairs show discordant tendencies.

has its values close to the threshold selected for determining the expression states; *i.e.* these are genuine observations and not technical artefacts.

Overlap with the expression profiling results: Between 16 and 32% of the 1,000 most significant probe-sets identified by the retinoblastoma and osteosarcoma EP analyses, respectively, are present in the set of significant probe-sets from the comparative analysis. This proportion is similar to that previously obtained for the integrative analysis.

No chromosomal locus enrichment: The distribution of the identified probe-sets does not show a significant enrichment at any genomic locus, unlike previously observed for the integrative analysis.

Probe-sets classification and manual curation: Out of the 1,171 significant probe-sets, the "dubious" and "genomic_multiple" were ignored: 14% of the total. Of the remaining 1,006 probe-sets, 473(47%) were manually curated as they belong to non-exonic classes (*e.g.* untranslated, antisense) or possibly have multiple targets. 77 probe-sets were discarded and 77 had their annotation updated; a total of 30% of the curated probe-sets. *In fine*, 896 probe-sets remain after curation.

Probe-sets associated genes: There are 789 unique genes identified by the curated probe-sets. 526 are commonly up-regulated, 305 commonly down-regulated, 35 are specifically up-regulated in osteosarcoma while the remaining 30 are specific to retinoblastoma, being either more or less expressed in comparison to osteosarcoma.

Evidence of transcript isoform regulation: Out of the 896 probe-sets, 186 redundantly identify 79 genes. As for the integrative analysis, these show almost invariable w correlation scores. Only one gene: *MAP1B*, identified by two probe-sets mapping in its 3'UTR region presents a discrepancy, as indicated in Table 4.23. The second probe-set, furthest away from

	5_71536979_71538579 _plus_genomic_at	5_71540574_71540953 _plus_genomic_at
retinoblastoma	2.90	2.50
osteosarcoma	1.70	0.60

Table 4.23 – *MAP1B* is identified by two probe-sets that map within its 3'UTR region. While the first probe-set, closer to the gene stop codon is over-expressed in both tumors, the second one shows a statistically significant difference (p-value = $1e^{-4}$) between both tumors, as well as within the osteosarcoma tumor.

the gene stop codon, shows a significant expression difference (Welch Two Sample t-test p-value of $1e^{-4}$) between retinoblastoma and osteosarcoma,

which could be explained by a different, tissue-specific, regulation of that gene transcript isoforms. Indeed, *MAP1B* has been reported to be involved in microtubule assembly and suggested to play an important role in the development and function of the nervous system.

Gene ontology analyses: The GO enrichment analyses were performed on 5 subsets of genes from the 789 identified genes by the comparative analysis as well as on the $1,000$ most significant genes retrieved from both retinoblastoma and osteosarcoma EP analyses; as follow:

1. the full comparative analysis set
2. the positively correlated over-expressed subset
3. the positively correlated down-regulated subset
4. the negatively correlated retinoblastoma specific subset
5. the negatively correlated osteosarcoma specific subset
6. the retinoblastoma EP set
7. the osteosarcoma EP set

As shown in Figure 4.39, the full set shows significant enrichment for the *extra-cellular environment, immune process, locomotion, cell activation* and *coagulation*. The analyses performed on the positively correlated subsets (*i.e.* the gene expression changes is identical in both tumors) reveals that the *coagulation* process is down-regulated, see Table 4.24 while the other ones are up-regulated (data not shown).

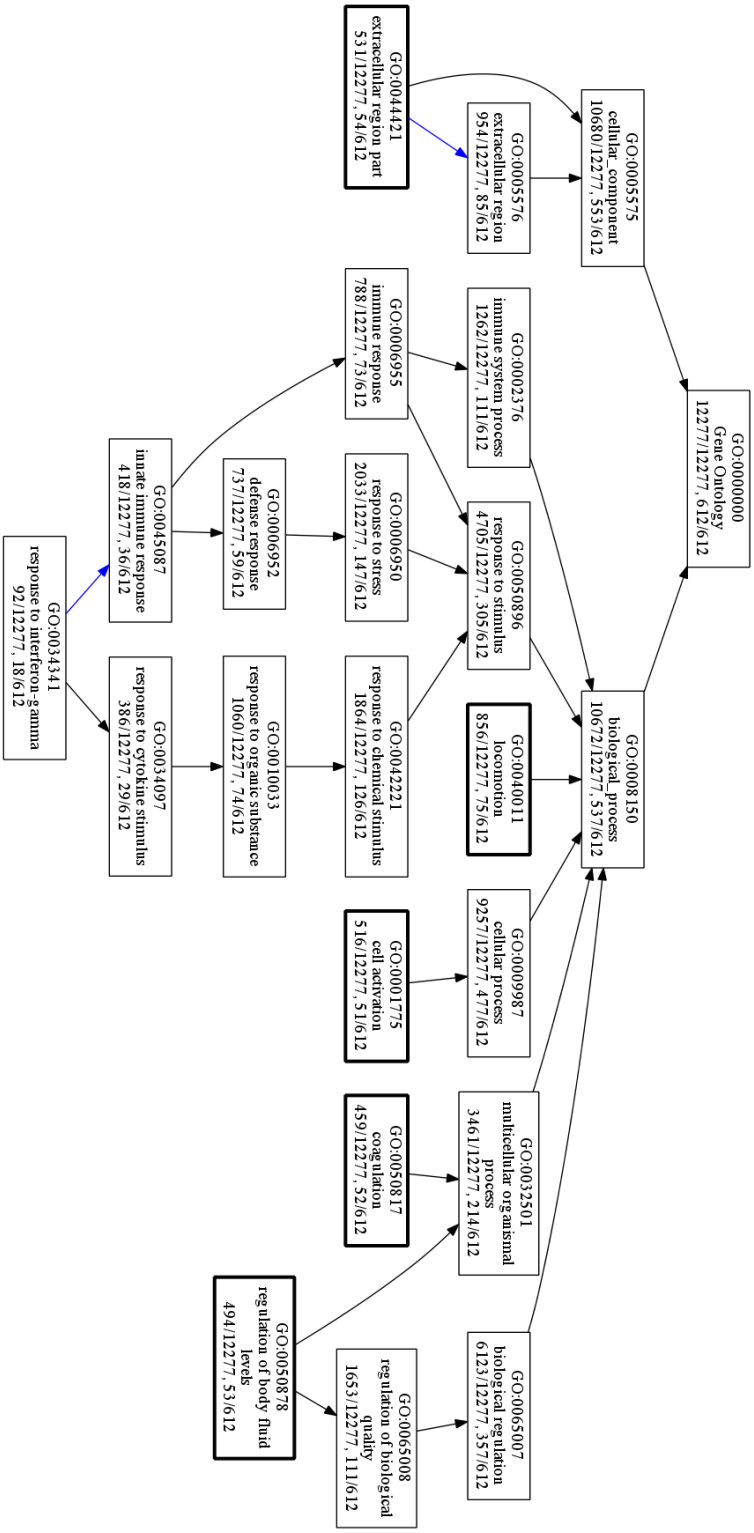


Figure 4.39 – GO enrichment analysis for the set of 789 genes identified by the comparative analysis. Only GO terms with an adjusted p-value higher than 1% are shown.

Name	Adj. p-value	Study Count	Population Count
amine metabolic process	0.004	25 (11.2%)	460 (3.7%)
steroid metabolic process	0.004	17 (7.6%)	194 (1.6%)
aminoglycoside antibiotic metabolic process	0.015	5 (2.2%)	8 (0.1%)
response to jasmonic acid stimulus	0.022	3 (1.3%)	4 (0.0%)
cellular response to jasmonic acid stimulus	0.022	3 (1.3%)	4 (0.0%)
protein-lipid complex	0.024	5 (2.2%)	22 (0.2%)
extracellular region	0.024	36 (16.1%)	954 (7.8%)
glycoside metabolic process	0.024	5 (2.2%)	14 (0.1%)
coagulation	0.024	22 (9.9%)	459 (3.7%)
regulation of body fluid levels	0.042	22 (9.9%)	494 (4.0%)

Table 4.24 – GO over-representation analysis with settings "Parent-Child-Union/Benjamini-Hochberg". For this analysis, a total of 12,277 genes were in the population set, of which a total of 223 genes were in the study set. A cutoff of 5% on the adjusted p-value was used.

While most of the enriched GO terms can be explained in a tumor environment, the fact that the *immune response* - a child of the *response to stimulus* term - is up-regulated is surprising, even though *receptor activity, SMAD binding* and *signal transduction* are GO terms significantly enriched in the positively correlated up-regulated subset. Whereas no GO terms are enriched for the retinoblastoma specific subset, the osteosarcoma one shows enrichment for the *extracellular matrix*, as well as for the *immune response, apoptosis* and *leukocyte migration*. Finally, although the EP subsets recall the results of the comparative analysis, they identify 3 times more terms at a stringent 1% adjusted p-value cutoff, in line with the hypothesis that most of these gene alterations are consequences and not causes of the cancerogenesis.

Retinoblastoma specific genes[3]: The seldom fact that there is no GO term enrichment for the retinoblastoma specific subset is very likely due to the sparse GO annotation of its genes. These are presented in Table 4.25. Grouping them according to their annotations reveals some interesting clusters:

- *FAH, HARS2, RPS21* encode important cellular components (ribosomes, tRNAs) or are involved in essential pathways (tyrosine catabolism).

[3]gene annotation from http://www.genecards.org (Rebhan et al., 1997)

symbol	gene name	Rb	Ots
NFIB	Nuclear factor 1 B-type	1	0
OLFM1	Noelin Precursor	1	0
ANAPC10P1	anaphase promoting complex subunit 10 pseudogene 1	1	0
HTR1A	5-hydroxytryptamine receptor 1A	1	0
MAP1B	microtubule-associated protein 1B	1	0
P4HA1	Prolyl 4-hydroxylase subunit alpha-1 Precursor	-1	0
PRKAR2B	cAMP-dependent protein kinase type II-beta regulatory subunit	1	-1
MT1E	Metallothionein-1E	-1	0
C10orf58	Uncharacterized protein C10orf58 Precursor	1	-1
AP1S2	AP-1 complex subunit sigma-2	1	-1
BEX4	brain expressed, X-linked 4	1	-1
FAH	Fumarylacetoacetase	-1	0
HPCAL1	Hippocalcin-like protein 1	-1	0
SNX9	Sorting nexin-9	-1	0
STOM	Erythrocyte band 7 integral membrane protein	-1	0
TSC22D3	TSC22 domain family protein 3	-1	0
PGRMC2	Membrane-associated progesterone receptor component 2	1	0
SCRN1	Secernin-1	1	-1
TRIM37	Tripartite motif-containing protein 37	1	-1
UBE2B	Ubiquitin-conjugating enzyme E2 B	1	0
PRKAR2B	cAMP-dependent protein kinase type II-beta regulatory subunit	1	-1
AQP3	Aquaporin-3	1	-1
MCM7	DNA replication licensing factor MCM7	1	-1
CLGN	Calmegin Precursor	1	-1
MARCKSL1	MARCKS-related protein	-1	0
PLSCR1	Phospholipid scramblase 1	-1	0
TRIM24	Transcription intermediary factor 1-alpha	1	-1
RPS21	40S ribosomal protein S21	1	0
HARS2	D-tyrosyl-tRNA	1	-1
SUMO1	Small ubiquitin-related modifier 1 Precursor	1	0

Table 4.25 – List of the 30 genes which expression is specific to retinoblastoma. Abbreviations are Rb and Ots for Retinoblastoma and Osteosarcoma, respectively.

- *BEX4, HPCAL1, HTR1A, MAP1B, OLFM1* were localized to the brain or associated with neural processes.
- *ANAPC10, MCM7, NFIB, SUMO1, TRIM24, TRIM37, TSC22D3, UBE2B* are related to transcriptional activity. While *ANAPC10* was not directly identified here, its pseudogene *ANAPC10P1* was.
- *SUMO1, TRIM24, TRIM37, UBE2B* were associated with the transcription regulation achieved through ubiquitination.
- *AP1S2, HTR1A, PGRMC2 , PLSCR1, PRKAR2B , SCRN1 , SNX9* are involved in signaling and protein trafficking.
- *CLGN, HPCAL1, MARCKSL1, PLSCR1, SCRN1, STOM* are involved or related to calcium signaling. *HPCAL1* is a member of a neuron-specific calcium-binding proteins family found in the retina and brain. It may be involved in the calcium-dependent regulation of rhodopsin phosphorylation and may be of relevance for neuronal signalling in the central nervous system. *CLGN* has only been reported to be involved in spermatogenesis so far, but binds calcium. *MARCKSL1* couples the protein kinase C and calmodulin signal transduction systems. *PLSCR1* and *SCRN1* are activated upon calcium binding.
- *AQP3* is a water channel required to promote glycerol permeability and water transport across cell membranes.
- *MT1E, P4HA1* have no relevant annotations.

Although some of the identified genes might be just confounding factors (*e.g. RPS21*), the analysis revealed potential new gene candidates and might have unravelled some signaling processes specific to retinoblastoma.

Osteosarcoma specific genes[4]: In addition to the enriched GO terms they reveal, the 35 genes specific to osteosarcoma present additional interesting characteristics listed in the following. The complete list of genes is presented in Table 4.26.

- involved in signal transduction: *RAB5B, FYN, S100A10, RAB31, IGFBP7, ITGB5, ATP6AP2*
- related to the ECM: *ISCU, TNC, TIMP1, CTHRC1, NID1, POSTN, SPARC, COL1A2, COL1A1, SH3PXD2B, COL5A2*

[4]gene annotation from http://www.genecards.org (Rebhan et al., 1997)

symbol	gene name	Rb	Ots
RAB5B	Ras-related protein Rab-5B	-1	1
COL1A2	Collagen alpha-2	-1	1
TPST2	Protein-tyrosine sulfotransferase 2	-1	1
ISCU	Iron-sulfur cluster assembly enzyme ISCU, mitochondrial Precursor	-1	1
GSN	Gelsolin Precursor	-1	1
SERF2	Small EDRK-rich factor 2	-1	1
FYN	Proto-oncogene tyrosine-protein kinase Fyn	-1	1
TNC	Tenascin Precursor	-1	1
RBM9	RNA-binding protein 9	-1	1
TIMP1	Metalloproteinase inhibitor 1 Precursor	-1	1
NID1	Nidogen-1 Precursor	-1	1
CNN3	Calponin-3	-1	1
SH3BGRL	SH3 domain-binding glutamic acid-rich-like protein	-1	1
POSTN	Periostin Precursor	-1	1
COL1A2	Collagen alpha-2	-1	1
CTHRC1	Collagen triple helix repeat-containing protein 1 Precursor	-1	1
SATB1	DNA-binding protein SATB1	-1	1
S100A10	Protein S100-A10	-1	1
COPZ2	Coatomer subunit zeta-2	-1	1
COL1A1	Collagen alpha-1	-1	1
SPARC	SPARC Precursor	-1	1
SCARF2	Scavenger receptor class F member 2 Precursor	-1	1
IGFBP7	Insulin-like growth factor-binding protein 7 Precursor	-1	1
ITGB5	Integrin beta-5 Precursor	-1	1
GPX8	Probable glutathione peroxidase 8	-1	1
MYO1B	Myosin-Ib	-1	1
RAB31	Ras-related protein Rab-31	-1	1
SH3PXD2B	SH3 and PX domain-containing protein 2B	-1	1
PMP22	Peripheral myelin protein 22	-1	1
SLC7A8	Large neutral amino acids transporter small subunit 2	-1	1
FAM46A	Protein FAM46A	-1	1
LRRC15	Leucine-rich repeat-containing protein 15 Precursor	-1	1
PPT1	Palmitoyl-protein thioesterase 1 Precursor	-1	1
COL5A2	Collagen alpha-2	-1	1
ATP6AP2	Renin receptor Precursor	-1	1

Table 4.26 – List of the 35 genes the expression of which is specific to osteosarcoma. Abbreviations are Rb and Ots for Retinoblastoma and Osteosarcoma, respectively.

- involved in vesicular transport and secretion: *TPST2, COPZ2, MYO1B, SLC7A8*
- binding or activated by calcium: *GSN, CNN3, SPARC*
- related to smooth muscle: *CNN3, IGFBP7, MYO1B, LRRC15, PPT1*
- involved in cell migration: *SH3PXD2B, MYO1B, PPT1*
- involved in transcription regulation: *RBM9*
- linked to metastasis *SATB1*
- associated with cell growth regulation *SPARC, PMP22*
- important for cell adhesion *POSTN, SCARF2, IGFBP7*
- producing or degrading **Reactive Oxygen Species** (ROS) *SH3PXD2B, GPX8*
- with vasodilating effects *ATP6AP2 IGFBP7*

In this section, I have presented the results of the methodological developments to enhance microarray annotations, to simulate realistic microarray data and to perform statistical integrative analysis and comparative analyses. The developed methods are sensitive and the obtained results sensible. In particular, the integrative analysis and comparative analysis approaches are broad hypothesis-generating methods. Some of the hypothesis raised by the described analyses of the retinoblastoma and osteosarcoma tumors will therefore be discussed in the next chapters.

Bibliography

Masayuki Amagai. A mystery of ahnak/desmoyokin still goes on. *J Invest Dermatol*, 123(4):xiv–xv, Oct 2004. doi: 10.1111/j.0022-202X.2004.23432.x.

Edurne Arriola et al. Genomic analysis of the her2/top2a amplicon in breast cancer and breast cancer cell lines. *Lab Invest*, 88(5):491–503, May 2008. doi: 10.1038/labinvest.2008.19.

Yann Audic and Rebecca S Hartley. Post-transcriptional regulation in cancer. *Biol Cell*, 96(7):479–98, Sep 2004. doi: 10.1016/j.biolcel.2004.05.002.

Sílvia Barrabés et al. Glycosylation of serum ribonuclease 1 indicates a major endothelial origin and reveals an increase in core fucosylation in pancreatic cancer. *Glycobiology*, 17(4):388–400, Apr 2007. doi: 10.1093/glycob/cwm002.

J Martin Bland and Douglas G Altman. Measuring agreement in method comparison studies. *Stat Methods Med Res*, 8(2):135–60, Jun 1999.

Sabina Chiaretti et al. Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, 103(7):2771–8, Apr 2004.

Sabina Chiaretti et al. Gene expression profiles of b-lineage adult acute lymphocytic leukemia reveal genetic patterns that identify lineage derivation and distinct mechanisms of transformation. *Clin Cancer Res*, 11(20): 7209–19, Oct 2005.

Nicolas Delhomme et al. Ensembl based custom definition file for affymetrix genechip. *submitted*.

Jennifer G DeLuca et al. Nuf2 and hec1 are required for retention of the checkpoint proteins mad1 and mad2 to kinetochores. *Curr Biol*, 13(23): 2103–9, Dec 2003.

Paul Flicek et al. Ensembl 2011. *Nucleic Acids Research*, 39(Database issue): D800–6, Jan 2011. doi: 10.1093/nar/gkq1064.

Raphaela Fritsche-Guenther et al. De novo expression of epha2 in osteosarcoma modulates activation of the mitogenic signalling pathway. *Histopathology*, 57(6):836–50, Dec 2010. doi: 10.1111/j.1365-2559.2010.03713.x.

Oscar Gonzalez-Moreno et al. Selenoprotein-p is down-regulated in prostate cancer, which results in lack of protection against oxidative damage. *The Prostate*, 71(8):824–34, Jun 2011. doi: 10.1002/pros.21298.

Corinna Grasemann et al. Gains and overexpression identify dek and e2f3 as targets of chromosome 6p gains in retinoblastoma. *Oncogene*, 24(42): 6441–9, Sep 2005. doi: 10.1038/sj.onc.1208792.

Sandrine Gratias et al. Genomic gains on chromosome 1q in retinoblastoma: consequences on gene expression and association with clinical manifestation. *Int J Cancer*, 116(4):555–63, Sep 2005.

Sandrine Gratias et al. Allelic loss in a minimal region on chromosome 16q24 is associated with vitreous seeding of retinoblastoma. *Cancer Res*, 67(1): 408–16, Jan 2007.

Wolfgang Huber et al. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:S96–104, Jan 2002.

Philippe Hupé et al. Analysis of array cgh data: from signal ratio to gain and loss of dna regions. *Bioinformatics*, 20(18):3413–22, Dec 2004.

Benjamin A Hurschler, David T Harris, and Helge Grosshans. The type ii poly(a)-binding protein pabp-2 genetically interacts with the let-7 mirna and elicits heterochronic phenotypes in caenorhabditis elegans. *Nucleic Acids Research*, 39(13):5647–57, Jul 2011. doi: 10.1093/nar/gkr145.

Rafael A Irizarry et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–64, Apr 2003. doi: 10.1093/biostatistics/4.2.249.

Maral Jamshidi et al. Nqo1 expression correlates inversely with nfb activation in human breast cancer. *Breast Cancer Res Treat*, 132(3):955–968, Apr 2012. doi: 10.1007/s10549-011-1629-5.

Audrey Kauffmann, Robert Gentleman, and Wolfgang Huber. arrayqualitymetrics–a bioconductor package for quality assessment of microarray data. *Bioinformatics*, 25(3):415–6, Feb 2009. doi: 10.1093/bioinformatics/btn647.

Peter A Leland et al. Endowing human pancreatic ribonuclease with toxicity for cancer cells. *J Biol Chem*, 276(46):43095–102, Nov 2001. doi: 10.1074/jbc.M106636200.

Lawrence Lin, A S Hedayat, and Wenting Wu. A unified approach for assessing agreement for continuous and categorical data. *J Biopharm Stat*, 17(4):629–52, Jan 2007.

Shih-Yeh Lin et al. Aspm is a novel marker for vascular invasion, early recurrence, and poor prognosis of hepatocellular carcinoma. *Clin Cancer Res*, 14(15):4814–20, Aug 2008. doi: 10.1158/1078-0432.CCR-07-5262.

Rie Matsushima-Nishiwaki et al. Suppression by heat shock protein 20 of hepatocellular carcinoma cell proliferation via inhibition of the mitogen-activated protein kinases and akt pathways. *J. Cell. Biochem.*, 112(11): 3430–9, Nov 2011. doi: 10.1002/jcb.23270.

Frank Mendrzyk et al. Identification of gains on 1q and epidermal growth factor receptor overexpression as independent prognostic markers in intracranial ependymoma. *Clin Cancer Res*, 12(7 Pt 1):2070–9, Apr 2006.

Marie E Persson-Moschos et al. Selenoprotein p in plasma in relation to cancer morbidity in middle-aged swedish men. *Nutr Cancer*, 36(1):19–26, Jan 2000. doi: 10.1207/S15327914NC3601_4.

Daniel Ramsköld et al. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol*, 5(12): e1000598, Dec 2009.

Pradeep Ramulu et al. Normal light response, photoreceptor integrity, and rhodopsin dephosphorylation in mice lacking both protein phosphatases with ef hands (ppef-1 and ppef-2). *Molecular and Cellular Biology*, 21 (24):8605–14, Dec 2001. doi: 10.1128/MCB.21.24.8605-8614.2001.

M Rebhan, V Chalifa-Caspi, J Prilusky, and D Lancet. Genecards: integrating information about genes, proteins and diseases. *Trends Genet*, 13 (4):163, Apr 1997.

Antoine M Snijders et al. Assembly of microarrays for genome-wide measurement of dna copy number. *Nature Genetics*, 29(3):263–4, Nov 2001.

Andrew I Su et al. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci USA*, 99(7):4465–70, Apr 2002.

Holger Sültmann et al. Gene expression in kidney cancer is associated with cytogenetic abnormalities, metastasis formation, and patient survival. *Clin Cancer Res*, 11(2 Pt 1):646–55, Jan 2005.

Oliver H Tam et al. Pseudogene-derived small interfering rnas regulate gene expression in mouse oocytes. *Nature*, 453(7194):534–8, May 2008.

Takashi Tanaka, Michael J Grusby, and Tsuneyasu Kaisho. Pdlim2-mediated termination of transcription factor nf-kappab activation by intranuclear sequestration and degradation of the p65 subunit. *Nature Immunology*, 8(6):584–91, Jun 2007. doi: 10.1038/ni1464.

Mei-Chuan Tang et al. Thymosin beta 4 induces colon cancer cell migration and clinical metastasis via enhancing ilk/iqgap1/rac1 signal transduction pathway. *Cancer Lett*, 308(2):162–71, Sep 2011. doi: 10.1016/j.canlet.2011.05.001.

Meredith A Tennis et al. Prostacyclin inhibits non-small cell lung cancer growth by a frizzled 9-dependent pathway that is blocked by secreted frizzled-related protein 1. *Neoplasia*, 12(3):244–53, Mar 2010.

Olaf Thuerigen et al. Gene expression signature predicting pathologic complete response with gemcitabine, epirubicin, and docetaxel in primary breast cancer. *J Clin Oncol*, 24(12):1839–45, Apr 2006.

Grischa Toedt et al. Alterations, a mixture model segmentation algorithm. *unpublished*, a.

Grischa Toedt et al. Chipyard, a framework for microarray data analysis. *unpublished*, b.

Joris A Veltman et al. Array-based comparative genomic hybridization for genome-wide screening of dna copy number in bladder tumors. *Cancer Res*, 63(11):2872–80, Jun 2003.

Zhijin Wu and Rafael A Irizarry. Stochastic models inspired by hybridization theory for short oligonucleotide arrays. *J Comput Biol*, 12(6):882–93, Jan 2005. doi: 10.1089/cmb.2005.12.882.

Haibo Zhang, Ju Youn Lee, and Bin Tian. Biased alternative polyadenylation in human tissues. *Genome Biology*, 6(12):R100, Jan 2005. doi: 10.1186/gb-2005-6-12-r100.

Keqiang Zhang et al. Overexpression of rrm2 decreases thrombspondin-1 and increases vegf production in human cancer cells in vitro and in vivo: implication of rrm2 in angiogenesis. *Molecular Cancer*, 8:11, Jan 2009. doi: 10.1186/1476-4598-8-11.

Boris Zielinski et al. Detection of chromosomal imbalances in retinoblastoma by matrix-based comparative genomic hybridization. *Genes Chromosomes Cancer*, 43(3):294–301, Jul 2005. doi: 10.1002/gcc.20186.

Chapter 5

Discussion

In the previous chapter, the results of the developed methods have been presented and this raised a number of interesting biological hypothesis related to the retinoblastoma and osteosarcoma cancerogenesis. In this chapter, some of the details of the developed methods will be discussed before focusing on these biological results and their possible implications.

5.1 Developed methods

In the frame of this doctoral work, I have enhanced or newly developed three main method kinds. The first dealt with probe and probe-set annotations, the second with simulating microarray data and the final one with the development and validation of statistical methods for performing integrative analyses and comparative analyses.

5.1.1 Annotation: the Ebased CDFs

Given the fast pace of the human genome re-sequencing and the increasing number of haplotypes discovered by *e.g.* the 1000 Genomes Project Consortium (2010), it is necessary to frequently update the annotation files used for microarray experiments as a whole and for EP in particular. It is even more critical for Affymetrix *GeneChip*®s, due to their design where a gene is identified by a set of probes: a probe-set. The information of the probe appartenance to a probe-set is contained in a CDF. Numerous methods have been published that redefine these CDF, however these are stringent and discard up to 30% of probes that may be valuable. The approach presented in Delhomme et al. (submitted) rescues but for 1% of these probes.

Among these rescued probes are some mapping to genomic loci - *i.e.* probe-sets of the "genomic" class - often in the 3' **UnTranslated Region** (UTR) of genes. This is expected as the probe-sets were originally designed by Affymetrix from EST libraries. As RNA extraction protocols were not

well established and sample degradation common, these were designed in the 3'UTR end of these ESTs and for that reason are the *GeneChip*®s nicknamed "3-prime arrays". The definition of UTR is complex in higher organisms and explains why genome refinements might have shifted probe-sets in and out of gene's 3'UTRs. That such probe-sets are accurately recalled, is one of the validation of the approach at regenerating CDFs.

The same holds true for the "untranslated" probe-set class, where the probe-sets are located within gene's introns. As shown in the results, most will identify alternative transcripts or transcripts that have since then been associated with intron-retention. In addition, some of them identify retrotransposed genes or pseudogenes (processed or not). Finally, although this has not been seen in the analyses presented in this study, these probe-sets have the potential to identify transcripts targeted for **Nonsense-mediated Decay** (NMD).

Another portion of the probe-sets affected by the genome refinements now map, rather unexpectedly, "antisense" to transcripts. In some cases, this results from the overlap of the 3'UTRs region of two genes located on opposing strands and is a failure of the CDF approach to correctly associate the probe-set and its target gene. However, in the remaining cases these probe-sets can - to our current knowledge - only identify a potential antisense transcript.

As shown in Delhomme et al. (submitted), applied on the frequently analyzed ALL dataset (Chiaretti et al., 2004, 2005) the newly generated CDF unravels three new potential candidate genes with implications in other tumors. Such CDFs are therefore enhanced tools to perform Affymetrix microarray analyses on either new or published data and by this means extend our biological knowledge.

This was further demonstrated in the present work: the results of the retinoblastoma EP analysis presented in Table 4.2, page 77 show that 6 of the first 9 most significantly over-expressed genes have been associated with cancer: *ASPM* is implicated in neo-angiogenesis (Zhang et al., 2009; Lin et al., 2008), *RRM2* has been associated with many cancers and described as a possible therapeutic target (Morikawa et al., 2010b,a), *TMSB15A* is implicated in cell migration and metastasis (Tang et al., 2011), *TOP2A* is amplified in breast cancer and is the target of therapeutic agents such as trastuzumab and anthracyclines (Arriola et al., 2008), *NUF2* is implicated in a cell cycle checkpoint (DeLuca et al., 2003) while *CDC2* is an essential cell cycle member.

These results confirm the potential of the refined CDF generation. The additional probe-set classes are great sources for defining new hypotheses to be further analyzed at the bench:

- "untranslated" may identify alternative regulatory mechanism such as NMD or intron-retention, which importance has been reviewed in

Matlin et al. (2005)

- "antisense" might reveal transcript antisense regulation, a mechanism that has been shown to be linked to fine-tuning the expression of some gene class (Xu et al., 2011)
- "genomic" may reveal pseudogene expression (Kalyana-Sundaram et al., 2012)

On the other hand, using the Ebased CDFs requires some caution. Particular attention need to be paid to the new probe-set classes, whether "untranslated", "antisense" or "genomic". For these a manual curation is necessary to avoid mis-assigning gene expression, although in most cases, an *in silico* validation is sufficient. However for every probe-set of the "multiple" class, usually matching a single gene family - as seen for the hemoglobins in the results (see paragraph 4.1.4, page 78), a wet lab validation would be required.

5.1.2 Simulation: the `aSim` package

A globally accepted approach to test newly developed statistical analyses is to use *in silico* generated datasets as a gold standard. This data represent the "truth" against which the results of the methods under evaluation are compared to in order to assess these methods' sensibility and sensitivity. It is essential for these artificial datasets to model realistic technical and biological variation. To evaluate the integrative analysis approach described in this thesis, there was a need for a microarray simulator able to generate pre-processed data. However, current publicly available microarray data simulators (Singhal et al., 2003; Balagurunathan et al., 2002; Nykter et al., 2006; Wierling et al., 2002) mainly focus on generating images to benchmark image analysis software and provide limited capabilities for analyzing post-processing algorithms.

This led me to develop a fast and flexible microarray simulator: `aSim`, which is able to simulate microarray data arising from different platforms on a pre-processed data level. It simulates data, which are gathered after feature extraction and normalization of the raw-data read-out from the scanner. It assumes that proper feature quality controls are applied, as well as the right normalization method, so that the effect of technical noise within the array becomes negligible. The default settings of `aSim` invoke the *Poisson* distribution for the number of altered groups within an array and the *Gaussian* and *log-normal* distribution for the generation of a measurement within a group. To complement these default models, `aSim` offers the possibility to replace them by custom-defined ones, *e.g.* for EP the Zipfs law, power law or the related Pareto (Li and Yang, 2002) distributions, which are distributions frequently mentioned in the literature to model EP data. In a similar

fashion, additional noise parameters can be introduced in the simulation process *e.g.* based on the variance stabilization model proposed by Huber et al. (Huber et al., 2002, 2003); a model to simulate the consequence of the inhomogenous cell population frequently observed in tumor samples could be devised. As seen in the results of the comparative analysis, *c.f.* section 4.5.2, page 130, there's an enrichment in the tumor sample of the "immune response" GO term that is likely due to the presence of many activated macrophages and lymphocytes, attracted by the wounds and inflammation resulting from cancerogenesis.

`aSim`, given its flexibility and extensibility, is a good tool to benchmark new or existing algorithms and to test their model assumptions under differing conditions. As demonstrated for the arrayCGH data segmentation algorithm `GLAD`, discrepancies between the original and simulated data can be explained by specific conditions in the parameter space of the algorithm. A possible origin of these discrepancies could be the simulator itself, *e.g.* if the distribution models used to simulate the data were not describing the original data properly. Indeed, as revealed by Hu et al. (2007) and confirmed in my results neither the fibroblast nor the ependyma or the bladder data are normally distributed, challenging the simulator assumptions. However, changing the distributions used to simulate the fibroblast dataset did not result in a significant change of the correlation range test results, see section 4.3.3, page 100. Testing different parameters for `GLAD` revealed that its default parameters resulted in finding fewer segments than there really are in the data introducing a bias in the simulations and explaining the observed discrepancies. An interesting co-finding from this comparison is that the Pearsons correlation coefficient is not influenced by the technical variability of the data only, but by its biological properties as well . This suggests that the results from Pearsons correlations, often used to compare microarray data in the literature, have to be taken with a greater caution and properly cross-validated. Finally, the modular structure of `aSim` enables a variety of applications, ranging from benchmarking arrayCGH segmentation algorithms up to simulating altered pathways in gene expression profiles - an interesting way to benchmark **Genetic Regulatory Network** (GRN) (Lee et al., 2002; Nykter et al., 2006) analysis tools.

In the general context of algorithm benchmarking, several evaluations of arrayCGH and EP segmentation algorithms have been done (Picard et al., 2005; Willenbrock and Fridlyand, 2005; Lehmussola et al., 2006) using simulated data, but their results are difficult to compare and to reproduce, which is a major drawback of simulation studies in general. In `aSim`, a particular attention has been given to that point - the simulation parameters are part of the results - which gives the researcher the possibility to replicate data locally, to test additional simulation conditions and to share data model assumptions (abstracted as `aSim` parameters) with collaboration partners. Using a standardized simulator ensures reproducibility across the research

community. In addition, the possibility to use data simulated under different data model conditions will safeguard against cases where an algorithm will be favored by the fact that it implements the same assumptions as the simulation process does.

Finally, at the community level, there is a demand for a simulator focusing on generating pre-processed data. `aSim` was designed for microarrays, but its flexibility allows for generating NGS data as well. In this new, quickly expanding research field, more and more algorithms dealing with pre-processed and normalized data are developed, that eventually need to be benchmarked. To easily do so, a simulator needs to be capable of generating any kind of data: EP, CNV, Tissue Microarray, *etc.* for any kind of data model: mixture models for microarrays, **negative binomial** for **RNA-Seq** data, ... So far, to my knowledge, no other simulator is that flexible. To summarize, `aSim` simulations are fast and the results reproducible from one simulation to the next. The simulation process is flexible, computationally cheap and extensible: it can be used to simulate other kinds of data such as NGS data. Finally, `aSim` allows users to share with other researchers, their data model assumptions through a reduced set of parameters.

5.1.3 Data analysis: statistical methods

For performing the integrative analysis and comparative analysis of retinoblastoma and osteosarcoma, a number of statistical methods have been developed and evaluated. Assumptions made during this process are discussed in the following, as are limitations and possible enhancements concerning the arrayCGH data imputation and sex rescue, the EP state definition and the integrative analysis statistics.

ArrayCGH data imputation: Imputing values might introduce some bias in the data (*i.e.* increase in both type I and II errors), however if properly performed, it provides the statistical methods with a significant power gain. As missing values impede the detection power of statistical methods by decreasing the number of observations, replacing them by realistic values drawn from data distributions taking into account these values context - *i.e.* missing value within a gained region will be imputed differently than these in a lost region - is a reasonable way to impute values. On the other hand, imputing successive values located within large chromosomal regions - *i.e.* too many successive missing values - or close to chromosomal breakpoints is certainly less acurate. For that reason, in order not to reduce the gained detective power, such values were not imputed.

In the integrative analysis study, only the arrayCGH microarray data for the Zielinski et al. (2005) dataset had to be imputed. This was performed after the data had been normalized within and between arrays. This offers the advantage that the data (*i.e.* log2 FC) is homoscedastic (*i.e.* the

variance is constant accross arrays) and normally distributed. Based on these robust assumptions, it is possible to use the `aSim` simulator to create log2 FC values using the "normal" distribution and parameterized by the median and MAD values extracted from the data. Using the median and MAD rather than the mean and SD makes this approach robust to outliers that may exist in the raw data. As the imputed data does not affect the overall distribution of the log2 FC, this procedure is therefore unlikely to have introduced a bias affecting the rest of the analyses.

In the same study, virtual arrayCGH probes were defined spanning the chromosomal loci between the probes present on the array. These virtual probes cover 69% of the genome and their median length is 2.3 times larger (377 *vs.* 164kb) than that of the spotted probes. Imputing the arrayCGH virtual probe values resulted in a statistically significant mean difference of the virtual *vs.* experimental log2 value distributions. However this mean difference (0.12) is neglectable in regard to the data spread: $[-1.3, 1.4]$ and $[-3.1, 3.6]$, respectively. It might result in a decrease of sensitivity, but is compensated by the massive increase of data available for the comparison: 1.86 times more probes and a 3.22 fold increase of the genomic coverage.

In the analyses presented in this thesis, values were imputed to increase the detection power and the obtained results are sensible and do not challenge the chosen approach. However, it is certain that imputing these values has increased the analyses error rates. A possibility - not evaluated - to control for such unwanted effects would be to associate the imputed values with a confidence score, *e.g.* a weight on a $[0, 1]$ scale, that would be taken into account to refine the p-values obtained by the statistical analyses.

ArrayCGH sex rescue: The microarrays used have an opposite sex matched design, therefore the log-ratio for the sex chromosome have to be corrected, *e.g.* for male samples, a $1X$ loss of the X chromosome and a $1X$ gain of the Y chromosome should be observed. Based on this assumption (and the reciprocal one for female samples), the log2 FC values were corrected. This approach, although basic, allowed the identification of a loss on chromosome X for the M22808 sample: dim(Xp11.22q23) which was previously reported by Zielinski et al. (2005). However it did not help correct the values for the Y chromosome, mainly for two reasons. First, there are only 30 probes covering this 59Mb large chromosome. Second, 50% consists of highly variable heterochromatin content, while the other half has partial homology to the X chromosome (pseudo-autosomal regions). Clearly, these probe values cannot be retained for any analysis as they are too variable. Even if this might affect the analyses, the consequences should be extremely limited, as the Y chromosome is rather gene-poor: 45 genes only are reported. *In fine*, this concerns only 0.5% of the data used in the integrative analysis. These sex-matched protocols, originally used as an internal control for arrayCGH

experiments, leading to data loss have since then been replaced by other - more appropriate - experimental designs such as hybridizing the tumor sample against the patients own blood or healthy tissue.

Expression Profiling state definition: To define the EP state - *i.e.* the discrete expression values - the results of the **Differential Expression** (DE) of the EP data were binned according to the selected adjusted p-value significance level. This resulted in probe-set centric states, *e.g.* probe-sets with a significant negative log2 FC were attributed a -1 state. As the resulting mono-dimensional vector is not representative of the variability across samples, *Z-scores* were calculated per probe-set using as the expected mean μ the fitted log2 FC value obtained from the DE analysis. The *Z-scores* were then used to refine the probe-set centric states for every sample. A major assumption of this process is that the microarray normalization has transformed the data distribution into a "normal distribution" - *i.e.* a Gaussian distribution of mean μ and variance σ. If that assumption holds, then the *Z-scores* should follow a "standard normal distribution" with parameters $\mu = 0$ and $\sigma = 1$. As observed for both the integrative analysis and comparative analysis, this assumption holds as the observed *Z-scores* distributions do not significantly differ from the expected standard normal distribution. Only 1.9 and 1.4 percent of deviation are observed, respectively, which can originate from several factors including, but not limited to, possible unasserted biological and/or technical variability, innacurate data normalization, innacurate linear models, small sample size, *etc.* That last point might be the reason for the 0.5% difference when this procedure is applied to the retinoblastoma (for the integrative analysis) and osteosarcoma (for the comparative analysis) datasets, as this last one contains twice more samples, *i.e.* the mean and SD can more accurately be determined, hence refining the obtained *Z-scores*. In any case, this slight deviation is not worrysome and on the contrary validates the approach. Moreover, it shows that either combining control samples hybridized on different platforms, as in the retinoblastoma EP study, or using an independent set of controls, such as the UHRR used for the osteosarcoma study, does not affect the calculation of the EP states.

ArrayCGH and expression profiling integrative analysis: The use of integrative analysis in molecular biology is not recent (Glass, 1980), however it has only recently been applied to the field of molecular genetics in large scales - reviewed in Rhodes and Chinnaiyan (2005). Moreover, after the pioneer study from Hyman et al. (2002) comparing CNV and EP - actually only chromosomal gains and gene over-expression - very few has been done for comparing such data until recently (Sadikovic et al., 2009). In that last study, the integrative analysis was only based on comparing

the gene-sets identified by three different techniques - arrayCGH, EP and Me-DIP-chip (methylated DNA immunoprecipitation followed by microarray hybridization to an Affymetrix Human Promoter 1.0R Tilling Array) - using Venn diagrams. Moreover, the significance of these associations was not tested, even though some case examples were validated by additional experiments, such as mass-spectrometry. Moreover, the *"copy number and DNA methylation analysis excluded sex chromosomes to avoid bias in the identification of significant genes/regions owing to sex differences between some tumor samples and male human osteoblast controls"*[1]. In comparison, the introduced approach - correcting for such effects and evaluating the significance of the different data type associations - is more robust. Concerning that last point, whereas the equation used in Hyman et al. (2002) was a mere observation, its implementation - weight w - presented in this study has been generalized and relies on solid statistical assumptions. As shown, it out-performs standard statistical methods (*e.g.* eta η^2, welch) used in other fields (*e.g.* banking) for performing that kind of integrative data comparison. Furthermore, η^2 and w are related: a classic approach for comparing data-sets is to perform an **ANalysis Of VAriance** (ANOVA), which relies on partitioning the total sum of squares SS into components related to the effects used in the selected model. The simplest type is:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(x_{ij}-\bar{x})^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}(x_{ij}-\bar{x_j})^2 + \sum_{j=1}^{n} n_j(\bar{x_j}-\bar{x})^2$$

and can be expressed as $SS = SS_t + SS_e$, where SS_t and SS_e represents the effect and error, respectively.

eta^2 - see equation 3.2, page 57 - relates to an ANOVA as follows:

$$\eta^2 = \frac{SS_e}{SS}, \eta^2 \in [0,1]$$

And likewise w does - see equation 3.5, page 58:

$$w = \frac{\sum_{i \neq j} |\bar{y_i} - \bar{y_j}|}{SS_e}$$

Moreover, since $\sum_{i \neq j} |\bar{y_i} - \bar{y_j}|$ is an estimator of SS, w is related to eta^2. The fact that w uses a distance metric rather than one based on the mean - as η^2 - is probably what makes it more accurate.

However, as observed, even w has limitations, *e.g.* some peculiar cases need to be corrected through the use of sub-methods, a process that is sub-optimal.

An alternative approach could be to use a polyserial correlation: *"it measures the correlation between two continuous variables with a bivariate*

[1] Extracted from Sadikovic et al. (2009)

normal distribution, where one variable is observed directly, and the other is unobserved. Information about the unobserved variable is obtained through an observed ordinal variable that is derived from the unobserved variable by classifying its values into a finite set of discrete, ordered values"[2](Olsson et al., 1982). Presently, in the context of the integrative analysis, the observed variable would be the EP values while the discrete ones would be the arrayCGH segmentation results. A limitation, there, might be the bivariate normality assumption, in particular for studies with few samples or samples with uneven quality.

A better approach, not extensively relying on such assumptions, might be to further refine the w method. As it is performed per gene, the current w has limitations when a class - *e.g.* CNV gain - consists of one observation only or when there's only a single class: *e.g.* only balanced cases are present. In such cases, a sub method, for example "compact" or "percent", has to be used. Another w limitation is that it penalizes correlation pairs with a large SD, *e.g.* if one EP value is highly amplified in the context of no CNV gain. However, these issues can be relatively easily adressed: the later one by multiplying the numerator of w (see equation 3.5, page 58) by the corresponding discrete class value - *i.e.* the arrayCGH label. For the former one, if instead of the distance divided by the SD, the mean value of every class is considered, there is no need for a "correction" sub-method. Finally, the score for every class should be pondered by the number of - if any - missing values and the classes sum returned as the correlation score. If values were imputed - as in this work -, their imputation confidence score could be used to refine that ponderation. The following equation describe this new "weight2" method w_2

$$w_2 = \sum_l \mu_l \times l_l \times w_l \tag{5.1}$$

where μ_l is the mean of the continuous values - *i.e.* EP values - observed for the class label l_l - the arrayCGH values - and their associated weight w_l.

Preliminary results show that w_2 out-performs both w and η^2 but further validations would be required.

A possible caveat: Although the sensitivity and specificity of the methods were assessed and deemed more than satisfactory, a possible caveat remains: the possible effects of the "multiple" class probe-sets. These probe-sets have member probes that map several positions of the genome and possibly record the expression of different genes or different members of the same gene family, as observed for the hemoglobins in osteosarcoma paragraph 4.1.4, page 78. On the brighter side, the use of "tumor *vs.* control" contrasts for the EP analyses performed in this work should be sufficient to

[2]description from www.sas.com

nullify this risk - *i.e.* any non-specific effect of the "tumor" samples will be cancelled out by the same effect from the "control" samples.

As presented, a number of methods have been developed during this doctoral work that enhance the analyses performed on microarrays data. While the CDFs are definitely dedicated to microarrays, both `aSim` and the integrative and comparative analyses (ICA) methods can be applied to generate/study other data. For the ICA, the only constraint is that the data has to be normalized prior to the analysis; *e.g.* for comparing NGS CNV with RNA-Seq data, the CNV data should be segmented - in a similar fashion as for microarrays - while the RNA-Seq data should be summarized by transcripts or gene-models and normalized using a negative binomial model. Such a procedure can already be performed in R using the `fastseg`, `easyRNASeq` and `DESeq` packages (Klambauer et al.; Delhomme et al., 2012; Anders and Huber, 2010); consequently the ICA methods developed in this thesis can readily be applied to publicly available NGS datasets.

The following sections focus on the biological pertinence of the results obtained applying these methods to a series of retinoblastoma and osteosarcoma datasets.

5.2 Data Analyses

It is important to point out that in this section all the results were obtained from *in-silico* methods and would need a wet-lab validation. However, the high confindence results obtained from these analyses, often validated by published work, appear sensible. The previously introduced methods are therefore to be seen as wet-lab "hypothesis-generating" methods and their obtained results discussed as such in the following.

5.2.1 Expression Profiling analyses

All the EP analyses performed came up with results validated by a large body of literature. Indeed, an average 80% of the over-expressed genes presented in the results among the 20 - 40 top candidates for either tumor had been previously associated with cancer. Similarly, the vast majority of the down-regulated genes could be associated to features of the differentiated tissue of origin, *e.g.* in retinoblastoma the gene *RHO* encoding *rhodopsin*, a retina pigment. Moreover, the refined undertaken approach - mainly the redefinition of the CDFs - revealed additional candidates that have so far hardly been investigated: *e.g.* in retinoblastoma the over-expression of *EZH2* (Polycomb protein EZH2) and *LCORL* (Ligand-dependent nuclear receptor corepressor-like protein) that could indicate an implication of repressor protein complexes in the cancerogenesis process. Finally, these CDFs helped identify entirely novel targets. The retinoblastoma EP experiment identified a strongly up-regulated "genomic" probe-set (its log2 FC close to 4) located downstream of the reported 3'UTR of the *CASC5* (cancer suceptibility candidate 5) gene. UTR regions are difficult to define *in-silico* and might vary *in-vivo* depending on the tissue in which that gene is expressed. Moreover, *GeneChip*® microarrays have been originally designed using ESTs libraries to specifically target the 3' end of genes. The probe-set location, next to its 3'UTR makes *CASC5* the most sensible candidate; : a gene that would not be identified by traditional CDFs.

Retinoblastoma: As expected for retinoblastoma tumors, many genes involved in the cell cycle were identified as being significantly de-regulated, *e.g. CDC2, NUF2, ...* Also, as mentioned above, differentiated tissue specific genes were down-regulated indicating that this tumor although appearing in a differentiated tissue consisted of cells that had "regressed" to a less differentiated state or were originally of a less differentiated origin, *i.e.* stem cells. These results, although interesting and significant in themselves, gave few clues about possible causative factors. It would be possible to use them to further investigate a small number of candidate genes, but hardly for investigating cancerogenesis as a "system".

Osteosarcoma: Identical observations were made when looking at the osteosarcoma EP results, even though the setup of the selected experiment (Fritsche-Guenther et al., 2010) consisting of primary and - unrelated - metastasis tumors offered some additional insights. Indeed 4 different contrasting analyses could be performed, see section 4.1.4, page 78, *e.g.* such as evaluating the DE between all tumor samples *vs.* the control samples.

First contrast: all tumors *vs.* control: This analysis recapitulated the observations from the retinoblastoma analysis above. First, tissue-specific genes were down-regulated: in the case of osteosarcoma the most affected were precursors of either trophic or mitogenic factors. Indeed, as introduced - see section 1.1.1, page 11 and 1.1.3, page 18 - the ECM of healthy bone tissue is a reserve for such factors. Then, among the other genes, a number were involved in the regulation of NF-κB, the upregulation of the corresponding signaling pathway having been reported of importance in osteosarcoma and its metastases (Felx et al., 2006; Asai et al., 2002). It was interesting to observe that the two most affected genes within this pathway: *NQO1, PDLIM2* were down-regulated negative regulators. Singling these possible tumor-suppressors as well as other cancer relevant genes out of the large number of non-causative tissue-specific down-regulated genes was not an easy task. Hence, it was hardly achievable to have comprehensive analyses - such as pathway analyses - reach a meaningful detection power.

Nevertheless, these results were interesting for gene centric analyses as well as for generating study hypotheses. In that context, the re-definition of the CDFs brought up among the 20 most up-regulated probe-sets two new, potentially interesting, targets:

- "ENST00000292896_ENST00000380237_transcript_multiple_at"
- "ENST00000361970_transcript_antisense_at"

The first one, associated with the gene *HBE1*, had probes mapping several genes. A closer look revealed that those genes: *HBD, HBG2, HBG1, HBB* are all Hemoglobin subunits. Provided that this was not a technical artifact (see paragraph 5.1.3, page 145), it is unclear why hemoglobin genes would be over-expressed in osteosarcoma. Is it due to the comparison of a tumor mixed cell population against a more unique cell type sample - *i.e.* osteoblasts cells were used as controls - or could it be relevant to the tumor neo-angiogenesis?

The second probe-set, antisense to *CCDC152* actually overlaped the 3' UTR region of the *SEPP1* gene located on the opposite strand. This gene has often been associated with cancer (Persson-Moschos et al., 2000; Gonzalez-Moreno et al., 2011), but as being down-regulated. Without further experimental validation, this candidate could not be validated, *i.e.* is it really the *SEPP1* gene that was expressed or was some antisense regulation of the *CCDC152* gene involved? Answering this question would give

an insight into a potentially new role of either gene. However, as for the retinoblastoma analysis above, these detailed results did not help conceive a more "systemic" view.

Additional insights from the other contrasts: It is interesting that two other contrasts: "primary tumor *vs.* control" and "metastasis *vs.* control" resulted in extremely similar candidate tables: only 3 probe-sets in 30 and 1 in 40, respectively differed from the "all tumor *vs.* control" contrast. Was it the presence of confounding factors that rendered them so similar or was it due to their shared origin? Metastases - at least those that pass the micrometastasis state - are more aggressive than primary tumors and two of the three genes identified by the 4 probe-sets specific to these contrasts supported this:

- *CD44* involved in cell proliferation, migration, *etc.* was over-expressed in most metastasis sample - by 1 log2 FC in 25% and 2 in 50% of all samples.
- *CYKL* involved in chondrogenesis (Kim et al., 2007) in an autocrine fashion was down-regulated in metastasis samples. Whether this was due to a more de-differentiated state of the metastasis compared to the primary tumor, or to a different micro-environment and whether the presence of this cytokine was benefitial or not for the primary tumor remain open questions.

The forth analyzed contrast: "primary tumor *vs.* metastasis" gave some hints to answer the first question. Indeed, despite the presence of confounding factors - *e.g.* numerous *Pulmonary surfactant-associated proteins* coding genes appeared among the most signifiantly differentially expressed candidates, a fact due to the lung localization of the metastasis - a more detailed analysis revealed that the primary tumor and the metastases seemed to have deregulated key cellular mechanisms in different ways. In the primary tumor, more cell cycle components: *MAD2L1, MNAT1, RRM2* were affected. In the metastases, components of the NF-κB, *Wnt* and *MAPK/ERK* pathways were stimulated (*CLIC3, CAV1*), as were *cell growth* (*NPR3, S100A6, TM4SF1*) and *motility* (*TM4SF1*) processes. In addition, the fact - contra-intuitive at first - that *PRAME - Preferentially expressed antigen of melanoma*, a gene that "functions as a transcriptional repressor, inhibiting the signaling of retinoic acid through the retinoic acid receptors RARA, RARB and RARG and prevents retinoic acid-induced cell proliferation arrest, differentiation and apoptosis"[3] - was over-expressed in primary tumors (see Table 4.5, page 84) indicated that this gene was actually down-regulated in the metastases favoring cell-proliferation and apoptosis-escape through an independance to anti-growth signal, in particular retinoic acid (RA).

[3]from http://www.genecards.org (Rebhan et al., 1997)

These results are important for two reasons; first they are in line with the more commonly accepted hypothesis that metastasis is not a trait appearing late in tumors - see section 1.1.1, page 11 - and second they demonstrated that more complex experimental and analytical designs offer the possibility to raise "systemic" hypotheses.

Confounding factors remained: As observed for the four different EP comparisons performed on the GEO GSE14359 dataset above and as seen for the retinoblastoma EP analysis, the down-regulation of a large number of genes appeared to be the consequence of the cancerogenesis rather than a causative factor. As such, these were considered **confounding factors** that diluted the detection power of the true **causative factors** within these analyses. Nevertheless, some interesting hypotheses could be derived, *e.g.* for the osteosarcoma "tumor *vs.* metastasis" comparison, it appeared that the primary tumor might have a more de-regulated cell cycle whereas the metastasis could be self-sufficient with regards to growth-signals.

5.2.2 ArrayCGH analyses

Copy Number Variation analysis: The arrayCGH analysis performed on the Zielinski et al. (2005) dataset gave results identical to those presented in that study, with only minor variations due to the use of improved segmentation algorithms. For example, the original analysis found smaller regions, which might be explained by the missing data imputation performed by the new analysis. The only significant differences were a non-recalled 22q13.2 gain and the identification of an high amplification on chromosome 11q22.1. Concerning the 22q13.2 gain previously reported, it fell below the detection threshold of the new improved method. Given the refined segmentation assumption, this gain probably originated from technical variations although a biological explanation could not be excluded. The 11q22.1 gain was not reported in the original study whereas I could observe it in that study's results. This raised the question whether it has been not reported due to insufficient evidences obtained from the analyses performed or if it was discarded as a technical artifact (*i.e.* based on the knowledge that some of the probes were misannotated or mislabeled). Verifying the probes located in that locus did not help answer that question; the locus was covered by 4 BAC clones totaling 0.72Mb and 3 additional virtual BAC clones spanned the remaining of that locus, covering an additional 1.6Mb. According to the PIMS annotation, there was no evidence of a technical artifact and neither were the probes spanning known repetitive regions. It could still be that some of these clones contained degenerated repetetive elements such as Long Terminal Repeats (LTR), Long INterspersed Elements (LINE), Short INterspersed Elements (SINE), *etc.* , which adds up to about 42% of the human genome (Lander and International Human Genome Sequencing Con-

sortium, 2001). The presence of such elements could explain an artifactual high amplification gain, but it would likely have been reported as such in the annotation. Therefore, wet-lab experiments would be required to validate this locus.

Three genes were found in that region, see Table 4.10, page 92. In the context of cancerogenesis, among those, *PGR* could be a potential candidate as it is involved in cell proliferation and differentiation. The other two genes involvement would be more speculative: *CNTN5* having only been reported during the nervous system development and *Q96M56_HUMAN* being a pseudogene similar to oligophrenin-1: a Rho-GTPase-activating protein, which affects cell migration and cell morphogenesis. Although these annotations could help devise a role for these genes, the evidences are too weak to do so confidently.

Clinical factor integrative analysis: Using the clinical factors, two kinds of integrative analyses were performed: MDS and hierarchical clustering. While the MDS did not reveal any correlations between the data and factors, the hierarchical clustering approach was more fruitful. First, as already reported by Gratias et al. (2007), a significant correlation could be observed between the occurence of vitreous seedings and the number of aberrations per tumor - *i.e.* the more aberrations, the more vitreous seedings arised. Similarly, it was verified that the number of aberrations negatively correlated with the presence of a germline mutation - the presence of the later decreased the likelihood to observe numerous aberrations, as expected from the Knudson two-hit hypothesis (see section 1.1.1, page 3). Also, even though the correlation reported by Gratias et al. (2005) between a 1q gain and a later onset of the disease could not be verified - the observed 0.9 p-value for a Welch Two Sample t-test being certainly due to the smaller dataset size in comparison to the original study: 14 *vs.* 76 samples - a weak correlation could be observed between the number of aberrations and the age at diagnosis (Pearson's product-moment correlation of 0.48 with a p-value of 0.08 and a confidence interval of $[-0.06, 0.8]$. In addition, the age at diagnosis differed between the samples with and without a germline mutation: a mean value of 495 *vs.* 846 days (Welch Two Sample t-test p-value of 0.09). The relationships between age at diagnosis, number of abberations and the presence of a germline mutation can be visualized in Figure 5.1. Taken together - and despite the limited number of samples - these results are in line with the common theory describing cancerogenesis: *i.e.* that a minimum number of events occuring across a "long" period of time are necessary for a cell to acquire a tumorigenic potential. However, from Figure 5.1, it appears that this might not be the only way a cell can acquire such potential: the sample with the largest number of aberrations was indeed diagnosed at an age compatible with a germline mutation. This could

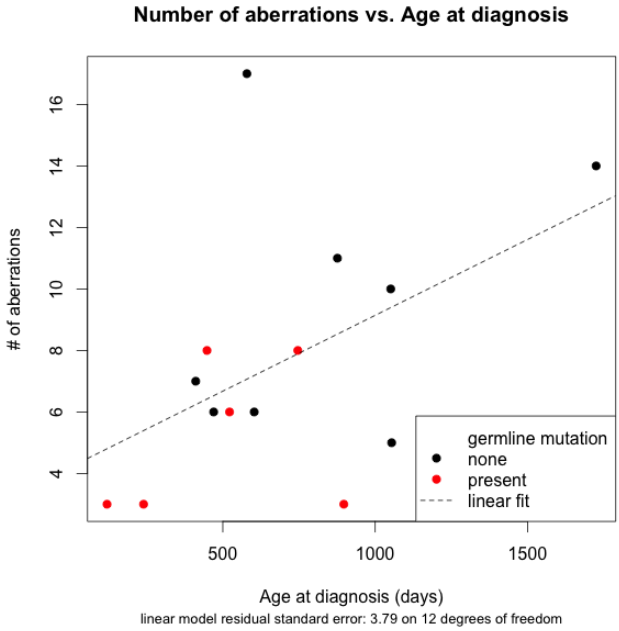


Figure 5.1 – Number of abberrations - identified using the refined segmentation algorithm - *vs.* the number the age at diagnosis. A weak correlation appears between both. On that same picture the effect of the germline mutation on both the number of aberrations and the age at diagnosis is evident.

indicate that a catastophic event such as chromothripsis described recently in different tumors by Stephens et al. (2011) and Rausch et al. (2012) took place instead.

These results' sensibility despite the limited number of sample is certainly due to the refined methods used that have an enhanced sensitivity.

Confounding factors presence: As in the EP analyses discussed previously, it is difficult to determine whether a CNV is a consequence or a cause of the cancerogenesis. For this reason, an integrative analysis approch was undertaken combining these 2 datasets. As shown in Figure 5.2, such approaches help determine the common factors identified by the independent analyses. These factors are likely only a subset of all causative factors, as they can originate from other sources of deregulation than CNV, such as methylation, short RNAs, alternative splicing, alternative poly-adenylation, *etc.* On one hand, concentrating on only 2 types of datasets limited the range of the undertaken integrative analysis approach. On the other hand, it helped validate that very approach and as discussed in the next section, could be extended to take advantages of additional dataset types.

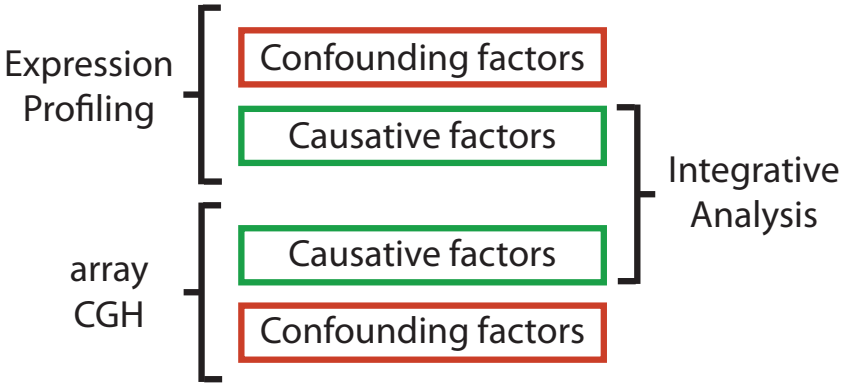


Figure 5.2 – A scheme representing the advantage of performing an integrative analysis over performing separate data analyses.

5.2.3 Retinoblastoma integrative analysis

As discussed previously - see section 5.1.3, page 143 - the statistical methods developed for performing an integrative analysis of EP and arrayCGH datasets were convincingly tested on simulated data. Applied to biological data this analysis revealed a number of interesting facts, unlikely due to confounding factors. These results and their implications are discussed in the following by order of importance to the aim of the present work.

Minor relevance results: These first points, although "minor" in this thesis work context, further validated the sensitivity of the whole process (CDFs re-generation, *etc.*) and brought up some interesting hypotheses.

The w score was almost invariable: The 203 curated probe-sets identify 171 genes, of which 24 have two and 4 have three probe-sets. Their w score were almost invariable: the corresponding *Z-scores* were contained in the $20-80\%$ range of the probability distribution, with the vast majority (24 out of 28) within the $30-60\%$ range. Obtaining such stable w score from different probe-sets was a strong evidence of the method validity. Actually, a statistical advantage could be taken of this probe-set multiplicity by increasing the corresponding genes significance, *e.g.* an additional weight could be added to the w_2 calculation model presented in paragraph 5.1.3, page 143.

Known genes were recalled: A similar approach by Grasemann et al. (2005), using a more basic methodology, reported the genes *E2F3* and *DEK* as being over-expressed and within gained loci. These genes: a major cell cycle TF and a known oncogene were identified as the 5^{th} and 14^{th} candidates in the present work, respectively. This finding from an inde-

pendant study using almost the same dataset was an additional validation of the introduced method.

Known aberrations were recalled: All the significant pairs were located in CNV regions previously reported by Zielinski et al. (2005); yet another validation of the approach. In the arrayCGH results, an high amplification gain was reported for chromosome 11q22.1 - as discussed previously, see paragraph 5.2.2, page 150 - but none of the arrayCGH - EP pair showed a significant correlation at that locus, which could be indicative of a technical artifact.

The "multiple" probe-set class was over-represented: In the 203 curated probe-sets, 99 were of the "multiple" class, in comparison to the 4 to 10% proportion on the whole *GeneChip*® for the gene and transcript probe-sets, respectively; see paragraph 4.1.2, page 70. This was explained by the presence of three different chromosome 6 haplotypes (*6, 6COX, 6QBL*) in the microarray annotation. These probe-sets, once manually curated, all resolved to a common gene set, *i.e.* that set of gene was present 3 times, once for every haplotype. Given the amount of ongoing resequencing project, the amount of haplotypes in the database is likely to increase and these should be differentiated from "real" multiple loci mapping probe-sets; a future improvement to the Ebased CDF generation. At the same time, it is yet another indirect validation of the probe-set generation process.

Evidence of pseudogene involvment? Three of the curated probe-sets were associated with pseudogenes: *SUCLA2P*, *PTMAP1* and *PIP5K1P1*. The last two ones are "processed" pseudogenes - *i.e.* pseudogenes that are transcribed but not translated. As described by Tam et al. (2008) - and others in a growing body of literature - there are evidences of gene regulation by the mean of pseudogenes. One mean of regulation is the creation of *"endogenous small interfering RNAs (endo-siRNAs), which are often processed from doubled stranded RNAs formed by hybridization of spliced transcripts from protein-coding genes to antisense transcripts from homologous pseudogenes. An inverted repeat pseudogene can also generate abundant small RNAs directly"*[4]. These would trigger the RNA interfence pathway and therefore decrease the expression of the corresponding protein-coding genes. In the present case, the corresponding genes were *PTMA*, involved in immune resistance to opportunistic infections and in modulating histone 1 interactions with the chromatin and *PIP5K1* involved in signal transduction through the generation of Phosphatidylinositol 4,5-bisphosphate (PIP2), an important cell second messenger. The *GeneChip*® HG-U133A, used for performing the EP did not contain any probe-set for *PTMA* but for three *PIP5K1*

[4]extracted from Tam et al. (2008)

family members: *A,B,C*. Only *PIP5K1B* probe-sets (n=2) showed a significant (according to the parameter used in this study) down-regulation of 2 log2 FC while the other ones showed a slight non-significant up-regulation (n=3, 0.8 log2 FC). This pseudogene regulation hypothesis would need to be validated, however it further demonstrates the use of integrative analyses.

Regulation through small RNA? One of the curated probe-set identified the U6-snRNA, a key element of the major - canonical - spliceosome (Sheth et al., 2006). As of the date of this manuscript, U6-snRNA has not been reported to be directly associated with cancer although alternative splicing has been (Venables, 2004; Skotheim and Nees, 2007) for a long time. As the original *GeneChip*® - as well as that of other microarray platforms - design relied on mRNA, it is likely that snRNA have been overseen in large scale microarray studies. The recent development of less limited technologies, such as RNA-Seq (Mortazavi et al., 2008) might change this. It might as well be that even if U6-snRNA came up in a EP study, it was discarded as a confounding factor: *i.e.* given the proliferative attribute of tumors, it is understandable that genes involved in processes such as replication, transcription, splicing, *etc.* are considered deregulated as a consequence and not a cause of cancer. The fact revealed here - *i.e.* that its expression correlated with its locus CNV state - might indicate a more important role of that snRNA.

Evidence of antisense regulation? For the genes *DST* and *EE1F1E1*, both involved in the negative regulation of the cell cycle and for the second one in the positive regulation of apoptosis, probe-sets were identified that were significantly over-expressed in conjonction with a chromosomal 6p gain. However, these two probe-sets map uniquely in the genome on the opposite strand of the 3'UTR region of the 2 mentioned genes. For these loci that have no reported haplotypes, it is likely that the probe-sets are recording an antisense expression of their target genes. In addition, the context of retinoblastoma - a tumor known for its cell cycle deregulation - renders these results even more interesting.

Major relevance results:

Only a minority of genes were recalled from EP: Among the 171 genes identified by the 203 curated significant pairs, only 55 (26%) overlaped with the candidate list of the EP analysis performed on the GSE5222 dataset. This was indicative that a large number of the EP candidate genes were confounding and not causative factors. Moreover, it meant that for potential causative factors, their expression variation between samples and controls was not sufficient to be significant. This can be explained by the

smaller sample subset within the dataset displaying that gene de-regulation, *e.g.* only 59% of the samples show a gain on chromosome 6p.

Performing a GO analysis on the commonly identified 55 genes revealed the "cell cycle process" and "cell cycle" terms as the most significant - as expected for retinoblastoma. This displayed the integrative analyses approaches power to remove confouding factors. In addition, it enabled additional analyses by reporting both positively and negatively correlating pairs: about 20% of the reported pairs had a negative correlation, indicative of more complex gene regulation. Among these pairs' associated gene, *SALL1* - involved in transcription and *Wnt* signal transduction - is over-expressed while located in an LOH region. The other genes, all on chromosome 6p - frequently gained - were all down-regulated, see section 4.4.3, page 114. These genes are involved in different processes such as immune response, signal transduction, transcription and cell cycle regulation. These results were suggestive of gene dosage compensation, but whether this is a natural response (*e.g.* resulting from a feedback mechanism) or actively orchestred by the tumor cells remains an open question, although observed facts give more support to the second hypothesis: 9 genes involved in immune response were repressed, as were genes involved in other cell processes: *ID4*, which is a *"basic helix-loop-helix transcription factors which can act as tumor suppressors"*[5] and *NEDD9* that *"plays a central coordinating role for tyrosine-kinase-based signaling related to cell adhesion"*[6]. For other genes, such as the two member of the *Forkhead box* (FOX) family: *FOXC1, FOXF2*, a down-regulation is more difficult to explain as they are member of a *"family of transcription factors that play important roles in regulating the expression of genes involved in cell growth, proliferation, differentiation, and longevity. Many FOX proteins are important to embryonic development."* *FOXC1* *"has been shown to play a role in the regulation of embryonic and ocular development"* and *FOXF2* *"has been shown to transcriptionally activate several lung-specific genes"*[7]. It appears that our knowledge about these genes is still too sparse to postulate any hypothesis concerning their role.

Chromosome 6 was enriched in significant pairs: A very interesting finding that shed some light on retinoblastoma cancerogenesis was the fact that the chromosome 6p arm was enriched in significant pairs: 149 out of the 171 genes present in significant pairs were located on chromosome 6p. This could not be explained as an artifact introduced by the pre-dominance of this CNV in retinoblastoma (59% of the samples used present that aberration) as chromosome 1q gain and 16q loss although very frequent (71% and 41%, respectively) did not show a similar enrichment.

[5]however it lacks DNA binding activity and consequently, the activity of the encoded protein depends on the protein binding partner"

[6]from http://www.genecards.org (Rebhan et al., 1997)

[7]from http://www.genecards.org (Rebhan et al., 1997)

The importance of chromosome 6p was already reported by Grasemann et al. (2005) - implicating the *DEK* and *E2F3* genes, see paragraph 5.2.3, page 153 - for a smaller region of the chromosome: 6p21-pter and by the Zielinski et al. (2005) study that originally reported a minimal deleted region located on 6p21.33-p21.31. A GO analysis of these 149 genes revealed - at the 10% FDR threshold selected in the present work - two terms: "artery morphogenesis" and "artery development". Similar GO analyses performed on the different mentioned subsets: 6p11.2-p12.3, 6p21.33-p21.31 and 6p21-pter ($13, 57, 128$ genes from significant pairs, respectively) did not reveal any functional enrichment. However, the same GO analysis performed on the significant pairs not on chromosome 6p ($n = 22$) showed an enrichment for "cell cycle" and "cell cycle process". This hinted that the "limitless replication" hallmark of cancer is acquired prior to the neo-angiogenesis trait.

In addition, as reported by Grasemann et al. (2005), the chromosome 6p gain is more frequently observed in tumor first diagnosed at a later age: 890 days *vs.* 305 for those without.

These results taken together seem to indicate that the chromosome 6p gain is an event occuring late in the tumor development and that it might benefit the tumor by providing a large number of slight advantages. Indeed, despite the rather consequent number of genes, no particular pathway enrichment appeared, but for "artery morphogenesis" and "artery development". This could possibly due to a lack of GO annotation for these genes, but is unlikely given the significant results obtained for the 22 genes not on chromosome 6. It is therefore reasonable to postulate that the chromosome 6p gain is not one of the principal causative events of retinoblastoma cancerogenesis but that it brings a large palette of enhancements to an already established tumor, among which at least one hallmark of cancer: neo-angiogenesis.

Four hallmarks of cancer were identified: As introduced in the previous paragraphs and conforted by the GO analyses, the results of the integrative analysis identified a number of hallmarks of cancer: neo-angiogenesis, self sufficiency in growth signal, insensitivity to anti-growth signals - here to the immune response - and limitless replicative potential - here through cell cycle processes.

These results are sensible but could be enriched by a different pathway analysis approach, indeed GO annotation are far from being complete and a lot of complementary information is available, *e.g.* from KEGG, Uniprot, *etc.* Moreover, as done in this doctoral work, ignoring *Inferred Electronic Annotation* (IEA) - the actual vast majority of the GO annotation, see (Rhee et al., 2008) for their significance - results in less powerful but more sensitive GO analyses. For example, the integrative analysis results revealed a number

of genes involved in DNA repair (*MSH5, MDC1, MCM3*) but this term is not significantly recalled. This is a known issue that does not challenge the approach undertaken in this work but it lessen its outcome. Nevertheless, the obtained results are sensible and more easily interpretable than those obtained separately for the EP or arrayCGH analyses alone.

The integrative analysis as presented here is a first step in our understanding of a tumor as a whole and the analysis performed underlined the difficulty of the challenges that remain to be addressed. Toward that goal, it is important to note that the integrative analysis developed in the present work is not limited to the datasets used and that it can be used for other technologies - especially those emerging from the rapid raise of NGS *e.g.* RNA-Seq, MEDIP-Seq, ChIP-Seq, etc - or for different analyses as discussed in the next section.

5.2.4 Retinoblastoma - Osteosarcoma comparative analysis

The observation that motivated such a comparative analysis is that patient with retinoblastoma have an higher chance to develop osteosarcoma further in life that the average population. As the microarray databases - such as GEO and ArrayExpress - content has been expanding exponantially a number of publicly available datasets resulting from the study of either of these tumors were readily available However, such a database oriented approach of retrieving data has several disadvantages that decreased the analysis detection power. First of all, the different tumor samples - although carefully selected - were not matched. A much better dataset would be one consisting of both tumors originating from the same patients group, but that's - even just logistically - very difficult to achieve. Second, none of the datasets identified as eligible from either microarray databases had a proper set of associated control samples. For that reason, the Universal Human Reference RNA (UHRR) control used in the **MicroArray Quality Control** (MAQC) study (MAQC Consortium et al., 2006) was selected. This again was suboptimal, as it consisted of a mix of different tumors and healthy tissues and had consequently a RNA content possibly vastly different from the tumors of interest here - as strongly suggested by the very significant score obtained for most of the genes while performing EP analyses on either tumors (section 4.5.1, page 121). Finally, the probe-sets had to be - as discussed for the integrative analysis - manually curated. About 10% of them had their annotation updated (*e.g.* probe-set of the antisense or untranslated class) and an additional 10% were discarded as only wet-lab evidence would decide of their validity. This in itself was a rather minor hindrance in comparison to the analysis power gain obtained by using the refined probe-sets. Moreover, the large number of samples of the different studies ($n = 91$ in total) helped rescue part of the statistical power of the comparative analysis - challenged

by the first two points raised above, as presented in the following.

Minor relevance results:

No evidence of locus enrichment: Unlike for the integrative analysis above and somewhat expected, no chromosomal loci enrichment could be observed between the tumors.

Known tumor-specific genes were recalled: Known tumor-specific genes were recalled: *e.g.* the genes *RUNX2, DOCK4, TNFRSF17* osteosarcoma-specific (see section 1.1.3, page 18) were identified with an FDR of 0%. These results validated the comparative analysis specificity.

Evidence of isoform differential expression: As discussed at the beginning of this chapter, regenerating the CDF file in a transcript-centric manner offered the possibility to measure transcript specific expression rather than gene expression. Out of the 896 probe-sets identified for the comparative analysis, 186 were associated with 79 genes, possibly recording differential isoform expression. As for the integrative analysis, the obtained w score were very consistant between probe-sets associated with the same gene but for *MAP1B* that showed a differential expression between retinoblastoma and osteosarcoma. The probe-sets that allowed for detecting this differential isoform expression between tumor types were both located downstream of the reported 3'UTR region of that gene. The difference of expression could hence be the consequence of a different, tissue-specific, poly-adenylation site usage rather than a consequence of cancerogenesis. This example showed the abilities of the comparative analysis approach to look for isoform differential expression. However, doing so is not trivial, as the data deconvolution might be complex for genes with more than 2 isoforms. Looking at the genes identified by the retinoblastoma integrative analysis, 20 genes were reported that have 2 isoforms. Out of these, 3 (see Table 5.1) had probe-sets that identify possibly different transcripts or one transcript and the gene as a whole (*i.e.* that records the expression - in theory - of both transcripts):

- *RANBP9* showed no difference in expression between the gene and transcript probe-sets, which is explained by either a similar level of expression or no expression of the second isoform.
- *CENPF* had its 'untranslated' probe-set mapping in the 3' region of one of the isoform or in an intronic region of the other isoform, 8-10kb away from its 3'UTR in a 60kb long gene. The original design of the Affymetrix *GeneChip*® using EST to identify 3' gene regions increases the likelihood that both probe-sets identified the different isoforms, in

probe-set	gene	symbol	mean exp.
gene	ENSG00000112308	C6orf62	-0.10
transcript	ENSG00000112308	C6orf62	0.58
gene	ENSG00000010017	RANBP9	0.73
transcript	ENSG00000010017	RANBP9	0.52
transcript	ENSG00000117724	CENPF	3.50
transcript_untranslated	ENSG00000117724	CENPF	0.24

Table 5.1 – Among these three genes having two reported isoforms, the available probe-sets expression might indicate a differential isoform expression for *C6orf62* and *CENPF*

which case they were significantly differentially expressed (Welch Two Sample t-test p-value of $2e^{-9}$).

- *C6orf62* had its 'gene' probe-set records a significantly lower expression that the 'transcript' probe-set (Welch Two Sample t-test p-value of $5e^{-4}$). This indicated a differential expression of both isoforms, *i.e.* the higher expression of the recorded isoform had to be negated by a decreased expression of the non recorded one to result in an overall lower expression of the gene.

These results further demonstrate the strength of the undertaken approach to create new biological hypotheses, to be taken up in the lab. They might explain why *CENPF*, which *"encodes a protein that associates with the centromere-kinetochore complex"*[8] and appears to be involved in chromosome segregation during mitosis is over-expressed[9]. And they could help associate a role and give a name to the, as of yet, "unidentified" *C6orf62* gene.

Major relevance results:

Both tumor kinds appeared very similar: As observed for the integrative analysis, the overlap of the probe-sets identified by the comparative analysis approach and these identified by the retinoblastoma and osteosarcoma EP was low: 16 and 32%, respectively. As discussed previously, this is likely due to the presence of confounding factors, *i.e.* the Sanger Cancer

[8] from http://www.genecards.org (Rebhan et al., 1997)

[9] *CENPF "is a component of the nuclear matrix during the G2 phase of interphase. In late G2 the protein associates with the kinetochore and maintains this association through early anaphase. It localizes to the spindle midzone and the intracellular bridge in late anaphase and telophase, respectively, and is thought to be subsequently degraded. The localization of this protein suggests that it may play a role in chromosome segregation during mitotis"*

Gene Census (Futreal et al., 2004) database reports only 487 genes (as of December 2012) that may be causative of cancerogenesis.
The comparative approach, which potentially removed a number of these confounding factors - a fact difficult to validate due to our ignorance of the true set of causative genes - revealed 789 genes significantly correlating between both tumors. Among these, 90.5% showed similar behaviors, 37% of which is a down-regulation. Among the almost 10% that showed negative correlations, 30 genes were specifically expressed in retinoblastoma and 35 in osteosarcoma. This indicates that these two tumors uses very similar mechanism. It is however unclear how much of this similarity would be shared with other tumors. It is known that different tumors affect different regulatory mechanisms: *e.g.* colorectal cancer is often associated with a *Wnt* pathway deregulation, whereas basal cell carcinoma is often reported associated with a *Hedgehog* pathway deregulation. While it is evident that different pathways have different affectors, the effectors are probably very similar since cancerogenesis results in the same phenotype: essentially an uncontrolled cell growth. It would be interesting to evaluate this by comparing either of the retinoblastoma or osteosarcoma tumors with an entirely unrelated tumor type.

GO analyses revealed common hallmarks of cancer: The same hallmarks of cancer as these identified by the integrative analysis were reported by the comparative analysis - cell activation, neo-angiogenesis, and (in)sensitivity to stimuli - as well as a new one: cell motility. But unlike previously, while most of the enriched GO terms could be explained in a tumor environment, the fact that the *immune response* was up-regulated is surprising, although *receptor activity, SMAD binding* and *signal transduction* were terms significantly enriched in the positively correlated and up-regulated subset. It is known that most tumors - recognized by the organism as "non-self" - are infiltrated by numerous lymphocityc cells. Hence, some of the results observed here might be caused by the "contamination" of the selected tumor samples by a variety of other non-cancerous cell types (*e.g.* fibroblasts), as theorized by (Hanahan and Weinberg, 2000), see the introductory paragraph 1.1.1, page 9. Obtaining the expression profile of only the cancer cells subset from the tumor cell population is a challenge now faced by the community. A number of technologies such as cell sorting, microfluidics and single-cell sequencing, which are being developed, should help tackle it.

Despite the likely heterogeneity of the samples analyzed here, a number of trait specific to either or both tumors appeared. First, both tumors - or their environment - had genes involved in the coagulation process down-regulation, which is in agreement with the necessary neo-angiogenesis of tumors. In parallel, both had their cell-cycle and cellular processes up-

regulated as a whole. Moreover, genes involved in cell motility were as well up-regulated. This is understandable for the osteosarcoma dataset that contains a number of metastasis sample, but the presence of these genes in the retinoblastoma dataset indicates that the "invasion" hallmark of cancer process was probably acquired early during the cancerogenesis process. Finally, as discussed above, both tumor appeared to have developed a self-sufficiency in growth signals.

GO analyses revealed tumor specificities: Although the tumors seemed extremely similar, the GO analysis showed that whereas no GO terms were enriched for the retinoblastoma specific subset, the osteosarcoma was enriched for terms such as *ECM*, *immune response, apoptosis* and *leukocyte migration*. The enrichment observed for the ECM is expected as a large number of trophic factors are naturally stored in the bone ECM and these are likely to be beneficial to the tumor cells. The presence of the *leukocyte migration* GO term further validated the observation made earlier that the analyzed tumor samples are certainly heterotypic.
The absence of any significant term for the retinoblastoma subset is surprising, but can be attributed in part to the rather poor GO annotation of the observed genes but might as well be due to the high similarity of the tumors - *i.e.* the pathways involved in retinoblastoma were a subset of those present in osteosarcoma.

Detailed analysis revealed additional specificities: The lack of GO annotation for the retinoblastoma and osteosarcoma specific subsets of genes was compensated by a detailed analysis of their respective 30 and 35 specific genes.

- retinoblastoma[10]
  4 genes were associated with neural processes, 2 of them being associated with neural specific signaling pathway, which seems relevant as the retina is of ectodermal origin and retinoblastoma might take advantage of these signaling pathways. A total of 9 genes were related to signaling pathways. Another 10 genes were involved in transcriptional activity and regulation, indicative of differences in the tumor cell processes. But more importantly, 5 genes specific to retinoblastoma were involved in calcium signalling and a sixth gene: *aquaporin* involved in osmotic regulation. This lead to hypothesize that the *protein kinase C / calmodulin* pathway might be an important player in the retinoblastoma "self-sufficiency in growth factors" hallmark of cancer.

[10]gene annotation from http://www.genecards.org (Rebhan et al., 1997)

- osteosarcoma[11]
  12 genes were related to the ECM, another indication that tumors take advantage of their direct micro-environment. 2 had vaso-dilatating effect, interesting in the context of neo-angiogenesis. 7 were related to cell migration and metastasis, in agreement with the fact that the osteosarcoma dataset contains samples from tumors at a more advanced state than the retinoblastoma dataset, as well as metastasis samples. Finally, 15 genes were involved in signal transduction and at least three of them (*RAB5B, RAB31, ATP6AP2*) were involved in the *GPCR/RAS* signaling pathway.

The observed facts and devised hypotheses demonstrated the advantage of conducting comparative analyses. Better designed studies than the one presented here are likely to result in more strongly supported hypotheses to be taken up in the wet-lab and help us get a better understanding of cancerogenesis and devise tumor specific targets for therapies.

[11] gene annotation from http://www.genecards.org (Rebhan et al., 1997)

5.3 Concluding remarks

The methods and algorithms developed during this doctoral work have increased the detection power of the various analyses performed and resulted in a number of interesting hypotheses related to retinoblastoma and osteosarcoma cancerogenesis and this at different levels, from very detailed to system-wide processes. These results are summarized in Figure 5.3.

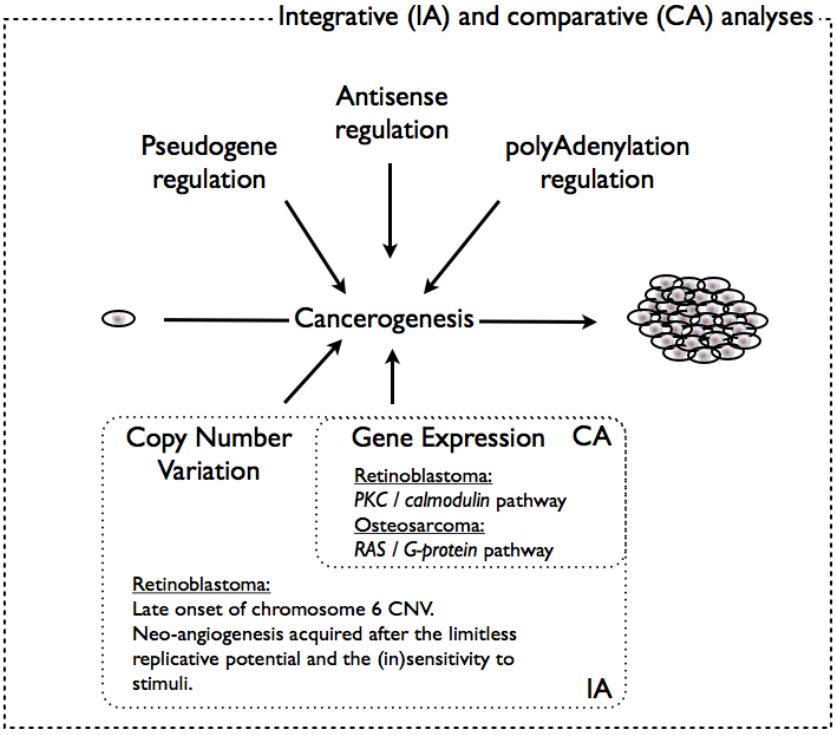


Figure 5.3 – A scheme representing the different hypotheses raised from the findings of the different analyses performed on the retinoblastoma and osteosarcoma datasets.

The "take-home" message of this set of analyses comparing retinoblastoma and osteosarcoma is that although both tumors appeared similar at the molecular level, either of them take advantage of their micro-environment and in addition, due to their different embryonal origin "mis-use" tissue specific pathways resulting in a unique tumor-specific signature. Discovering and understanding these tumor specificities is an essential step to *in fine* be able to develop tumor-specific therapies.

Bibliography

1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73, Oct 2010. doi: 10.1038/nature09534.

Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology 2010 11:202*, 11(10):R106, Oct 2010.

Edurne Arriola et al. Genomic analysis of the her2/top2a amplicon in breast cancer and breast cancer cell lines. *Lab Invest*, 88(5):491–503, May 2008. doi: 10.1038/labinvest.2008.19.

Tatsuya Asai et al. Vcp (p97) regulates nfkb signaling pathway, which is important for metastasis of osteosarcoma cell line. *Cancer Science*, 93(3): 296–304, Mar 2002. doi: 10.1111/j.1349-7006.2002.tb02172.x.

Yoganand Balagurunathan et al. Simulation of cdna microarrays via a parameterized random signal model. *J Biomed Opt*, 7(3):507–23, Jul 2002.

Sabina Chiaretti et al. Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, 103(7):2771–8, Apr 2004.

Sabina Chiaretti et al. Gene expression profiles of b-lineage adult acute lymphocytic leukemia reveal genetic patterns that identify lineage derivation and distinct mechanisms of transformation. *Clin Cancer Res*, 11(20): 7209–19, Oct 2005.

Nicolas Delhomme, Ismaël Padioleau, Eileen E Furlong, and Larsm Steinmetz. easyrnaseq: a bioconductor package for processing rna-seq data. *Bioinformatics*, Jul 2012.

Nicolas Delhomme et al. Ensembl based custom definition file for affymetrix genechip. *submitted*.

Jennifer G DeLuca et al. Nuf2 and hec1 are required for retention of the checkpoint proteins mad1 and mad2 to kinetochores. *Curr Biol*, 13(23): 2103–9, Dec 2003.

Mélanie Felx et al. Endothelin-1 (et-1) promotes mmp-2 and mmp-9 induction involving the transcription factor nf-kappab in human osteosarcoma. *Clin Sci*, 110(6):645–54, Jun 2006.

Raphaela Fritsche-Guenther et al. De novo expression of epha2 in osteosarcoma modulates activation of the mitogenic signalling pathway. *Histopathology*, 57(6):836–50, Dec 2010. doi: 10.1111/j.1365-2559.2010.03713.x.

P Andrew Futreal et al. A census of human cancer genes. *Nature Reviews Cancer*, 4(3):177–83, Mar 2004. doi: 10.1038/nrc1299.

Gene V Glass. Integration of research studies: Meta-analysis of research. methods of integrative analysis; final report. *Nat. Inst. Ed.*, pages 1–329, Aug 1980.

Oscar Gonzalez-Moreno et al. Selenoprotein-p is down-regulated in prostate cancer, which results in lack of protection against oxidative damage. *The Prostate*, 71(8):824–34, Jun 2011. doi: 10.1002/pros.21298.

Corinna Grasemann et al. Gains and overexpression identify dek and e2f3 as targets of chromosome 6p gains in retinoblastoma. *Oncogene*, 24(42): 6441–9, Sep 2005. doi: 10.1038/sj.onc.1208792.

Sandrine Gratias et al. Genomic gains on chromosome 1q in retinoblastoma: consequences on gene expression and association with clinical manifestation. *Int J Cancer*, 116(4):555–63, Sep 2005.

Sandrine Gratias et al. Allelic loss in a minimal region on chromosome 16q24 is associated with vitreous seeding of retinoblastoma. *Cancer Res*, 67(1): 408–16, Jan 2007.

Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, Jan 2000.

Jing Hu et al. Exploiting noise in array cgh data to improve detection of dna copy number change. *Nucleic Acids Res*, 35(5):e35, Jan 2007.

Wolfgang Huber et al. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:S96–104, Jan 2002.

Wolfgang Huber et al. Parameter estimation for the calibration and variance stabilization of microarray data. *Stat Appl Genet Mol Biol*, 2:Article3, Jan 2003.

Elizabeth Hyman et al. Impact of dna amplification on gene expression patterns in breast cancer. *Cancer Res*, 62(21):6240–5, Nov 2002.

Shanker Kalyana-Sundaram et al. Expressed pseudogenes in the transcriptional landscape of human cancers. *Cell*, 149(7):1622–34, Jun 2012.

Jae-Sung Kim, Zae Young Ryoo, and Jang-Soo Chun. Cytokine-like 1 (cytl1) regulates the chondrogenesis of mesenchymal cells. *J Biol Chem*, 282(40): 29359–67, Oct 2007.

Guenter Klambauer et al. fastseg: a fast segmentation algorithm. *submitted*.

Eric S Lander and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822): 860–921, Feb 2001.

Tong Ihn Lee et al. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(5594):799–804, Oct 2002.

Antti Lehmussola, Pekka Ruusuvuori, and Olli Yli-Harja. Evaluating the performance of microarray segmentation algorithms. *Bioinformatics*, 22 (23):2910–7, Dec 2006.

Wentian Li and Yaning Yang. Zipf's law in importance of genes for cancer classification using microarray data. *J Theor Biol*, 219(4):539–51, Dec 2002.

Shih-Yeh Lin et al. Aspm is a novel marker for vascular invasion, early recurrence, and poor prognosis of hepatocellular carcinoma. *Clin Cancer Res*, 14(15):4814–20, Aug 2008. doi: 10.1158/1078-0432.CCR-07-5262.

MAQC Consortium et al. The microarray quality control (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*, 24(9):1151–61, Sep 2006.

Arianne J Matlin, Francis Clark, and Christopher W J Smith. Understanding alternative splicing: towards a cellular code. *Nature Reviews Molecular Cell Biology*, 6(5):386–98, May 2005.

Teppei Morikawa et al. Ribonucleotide reductase m2 subunit is a novel diagnostic marker and a potential therapeutic target in bladder cancer. *Histopathology*, 57(6):885–92, Dec 2010a.

Teppei Morikawa et al. Expression of ribonucleotide reductase m2 subunit in gastric cancer and effects of rrm2 inhibition in vitro. *Hum Pathol*, 41 (12):1742–8, Dec 2010b.

Ali Mortazavi, Brian A Williams, Kenneth Mccue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature Methods*, 5(7):621–8, Jul 2008.

Matti Nykter et al. Simulation of microarray data with realistic characteristics. *BMC Bioinformatics*, 7:349, Jan 2006.

Ulf Olsson, Fritz Drasgow, and Neil J Dorans. The polyserial correlation coefficient. *Psychometrika*, 47(3):337–347, Sep 1982.

Marie E Persson-Moschos et al. Selenoprotein p in plasma in relation to cancer morbidity in middle-aged swedish men. *Nutr Cancer*, 36(1):19–26, Jan 2000. doi: 10.1207/S15327914NC3601_4.

Franck Picard et al. A statistical approach for array cgh data analysis. *BMC Bioinformatics*, 6:27, Jan 2005.

Tobias Rausch, David T W Jones, Marc Zapatka, Adrian M Stütz, et al. Genome sequencing of pediatric medulloblastoma links catastrophic dna rearrangements with tp53 mutations. *Cell*, 148(1-2):59–71, Jan 2012. doi: 10.1016/j.cell.2011.12.013.

M Rebhan, V Chalifa-Caspi, J Prilusky, and D Lancet. Genecards: integrating information about genes, proteins and diseases. *Trends Genet*, 13 (4):163, Apr 1997.

Seung Yon Rhee et al. Use and misuse of the gene ontology annotations. *Nature Reviews Genetics*, 9(7):509–15, Jul 2008.

Daniel R Rhodes and Arul M Chinnaiyan. Integrative analysis of the cancer transcriptome. *Nature Genetics*, 37 Suppl:S31–7, Jun 2005. doi: 10.1038/ng1570.

Bekim Sadikovic, Maisa Yoshimoto, et al. Identification of interactive networks of gene expression associated with osteosarcoma oncogenesis by integrated molecular profiling. *Hum Mol Genet*, 18(11):1962–75, Jun 2009. doi: 10.1093/hmg/ddp117.

Nihar Sheth et al. Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res*, 34(14):3955–67, Jan 2006.

Sunil Singhal et al. Microarray data simulator for improved selection of differentially expressed genes. *Cancer Biol Ther*, 2(4):383–91, Jan 2003.

Rolf I. Skotheim and Matthias Nees. Alternative splicing in cancer: Noise, functional, or systematic? *The International Journal of Biochemistry & Cell Biology*, 39(78):1432 – 1449, 2007. ISSN 1357-2725. doi: 10.1016/j.biocel.2007.02.016.

Philip J Stephens et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, 144(1):27–40, Jan 2011. doi: 10.1016/j.cell.2010.11.055.

Oliver H Tam et al. Pseudogene-derived small interfering rnas regulate gene expression in mouse oocytes. *Nature*, 453(7194):534–8, May 2008.

Mei-Chuan Tang et al. Thymosin beta 4 induces colon cancer cell migration and clinical metastasis via enhancing ilk/iqgap1/rac1 signal transduction pathway. *Cancer Lett*, 308(2):162–71, Sep 2011. doi: 10.1016/j.canlet.2011.05.001.

Julian P Venables. Aberrant and alternative splicing in cancer. *Cancer Research*, 64(21):7647–54, Nov 2004.

Christoph K Wierling, Matthias Steinfath, Thorsten Elge, Steffen Schulze-Kremer, Pia Aanstad, Matthew Clark, Hans Lehrach, and Ralf Herwig. Simulation of dna array hybridization experiments and evaluation of critical parameters during subsequent image and data analysis. *BMC Bioinformatics*, 3:29, Oct 2002.

Hanni Willenbrock and Jane Fridlyand. A comparison study: applying segmentation to array cgh data for downstream analyses. *Bioinformatics*, 21 (22):4084–91, Nov 2005.

Zhenyu Xu, Wu Wei, Julien Gagneur, et al. Antisense expression increases gene expression variability and locus interdependency. *Mol Syst Biol*, 7: 468, Feb 2011.

Keqiang Zhang et al. Overexpression of rrm2 decreases thrombspondin-1 and increases vegf production in human cancer cells in vitro and in vivo: implication of rrm2 in angiogenesis. *Molecular Cancer*, 8:11, Jan 2009. doi: 10.1186/1476-4598-8-11.

Boris Zielinski et al. Detection of chromosomal imbalances in retinoblastoma by matrix-based comparative genomic hybridization. *Genes Chromosomes Cancer*, 43(3):294–301, Jul 2005. doi: 10.1002/gcc.20186.

Chapter 6

Conclusion and Outlook

The main aim of this thesis work was the development and enhancement of bioinformatics approaches to better analyze high throughput data generated using microarrays. Another aim was to investigate the effect of CNV on gene-expression in an integrative analysis approach and finally to assess how similar retinoblastoma and osteosarcoma tumors are, as the later are the prefered site of relapse of the former.

Four different tools have been developed to achieve these goals:

1. "customCDF": a tool to redefine the **Custom Definition File** (CDF) of Affymetrix *GeneChip*®s. Essential to take advantage of the constantly evolving human genome reference and annotations.

2. "aSim": a tool to simulate microarray data. Critical to benchmark the developed algorithms.

3. integrative analysis: a set of statistical methods combined in a pipeline to address the second goal.

4. comparative analysis: a modification of the integrative analysis to adress the final goal.

The first tool, by rescuing as many information as possible for microarray analyses - a process that has not been done extensively so far - was critical to this study: it raised the discovery power of the downstream analyses. The integrative and comparative approaches revealed themselses as highly hypothesis-generating. These relatively high confidence hypotheses - generated *in-silico* - can then be transfered to the wet-lab to be further scrutinized. That meaningful hypothesis - at a fine grained or at the systems scale - can be raised shows the success of the chosen approaches and the potential of the developed methods and tools.

Concerning the biological aim of this thesis, the integrative analysis applied to retinoblastoma revealed the high importance of the chromosome 6

gain, indicating that many genes on that chromosome helps cancerogenesis and this at a later stage of the disease. Moreover, it showed the existence of positive and negative compensation of gene expression in lost and gained regions, showing the complexity of the cancerogenesis mechanism and empaphizing the need to use systemic approaches. This last statement is further supported by the *in-silico* evidence of antisense, pseudogene and snRNA regulation shown in this work. Finally, the integrative analysis revealed that out of all the hallmarks of cancer: deregulation of cell cycle (well known in retinoblastoma) and angiogenesis, as well as the inactivation of the immune response were most prominent, which might help developing more targeted therapies.

In parallel, the comparative analysis revealed the high similarity of the retinoblastoma and osteosarcoma tumors, while at the same time showing that either of them take advantage of their distinct micro-environment and consequently appear to make use of different signaling pathways: the *PKC/calmodulin* pathway for retinoblastoma and the *GPCR/RAS* for osteosarcoma.

These results first need to be validated by wet lab experiments. One limitation of the presented approach was the use of sub-optimal publicly available data; studies specifically designed for integrative and/or comparative analyses should raise even stronger biologically relevant hypotheses. On the other hand, this "limitation" showed that mining the data present in public microarray repositories can be an easy way to define work hypotheses to be tested in the lab. On the mid term, such analyses should be performed routinely and help us get a better understanding of the complex, heterotypic cancer system, and hopefully on the long term give raise to personalized - a thought supported by the current exponential development of sequencing and microfluidics techniques - anti-cancer therapies.

Bibliography

Sebastian Barbus, Björn Tews, Daniela Karra, Meinhard Hahn, Bernhard Radlwimmer, **Nicolas Delhomme**, Christian Hartmann, Jörg Felsberg, Dietmar Krex, Gabriele Schackert, Ramon Martinez, Guido Reifenberger, and Peter Lichter. Differential retinoic acid signaling in tumors of long- and short-term glioblastoma survivors. *J Natl Cancer Inst*, 103(7):598–606, Apr 2011.

ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012.

Robert C Gentleman et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology 2010 11:202*, 5(10):R80, Jan 2004.

Daniel Haag, Petra Zipper, Viola Westrich, Daniela Karra, Karin Pfleger, Grischa Toedt, Frederik Blond, **Nicolas Delhomme**, Meinhard Hahn, Julia Reifenberger, Guido Reifenberger, and Peter Lichter. Nos2 inactivation promotes the development of medulloblastoma in ptch1(+/-) mice by deregulation of gap43-dependent granule cell precursor migration. *PLoS Genet*, 8(3):e1002572, Mar 2012.

Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, Jan 2000.

Felix Kokocinski, **Nicolas Delhomme**, Gunnar Wrobel, Lars Hummerich, Grischa Toedt, and Peter Lichter. Fact–a framework for the functional interpretation of high-throughput experiments. *BMC Bioinformatics*, 6: 161, Jan 2005.

David J Lockhart et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, 14(13):1675–80, Dec 1996. doi: 10.1038/nbt1296-1675.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Daniel E Stange, Felix Engel, Thomas Longerich, B K Koo, Martin Koch, **Nicolas Delhomme**, Martina Aigner, Grischa Toedt, Peter Schirmacher, Peter Lichter, Jürgen Weitz, and Bernhard Radlwimmer. Expression of an ascl2 related stem cell signature and igf2 in colorectal cancer liver metastases with 11p15.5 gain. *Gut*, May 2010.

Robert A Weinberg. *The biology of Cancer*. Garland Science, 2007. ISBN 0815340761.

Boris Zielinski et al. Detection of chromosomal imbalances in retinoblastoma by matrix-based comparative genomic hybridization. *Genes Chromosomes Cancer*, 43(3):294–301, Jul 2005. doi: 10.1002/gcc.20186.

Appendix A

Samples and Datasets

In this appendix are described the different samples and datasets used.

- On page 176 is the description of the retina sample purchased from Clontech.
- On page 177 is the description of the matrixCGH and EP dataset generated at the DKFZ and at the University of Duisburg-Essen.
- On page 178 is the GEO GSE29684 dataset samples description.
- On page 180 is the GEO GSE29683 dataset samples description.
- On page 181 is the GEO GSE14359 dataset samples description.
- On page 182 is the GEO GSE14827 dataset samples description.
- On page 183 is the GEO GSE5350 dataset samples description.

PRODUCT: Human Retina Total RNA

CATALOG No. 636579

LOT No. 6010100

CONCENTRATION: 1 µg/µl

FORM
Suspension of total RNA in DEPC-treated H_2O

STORAGE CONDITIONS: –70°C

SHELF LIFE
1 year from date of receipt under proper storage conditions.

SHIPPING CONDITIONS
Dry ice (–70°C)

FOR RESEARCH USE ONLY

DESCRIPTION
Total RNA isolated by a modified guanidinium thiocyanate method (1).

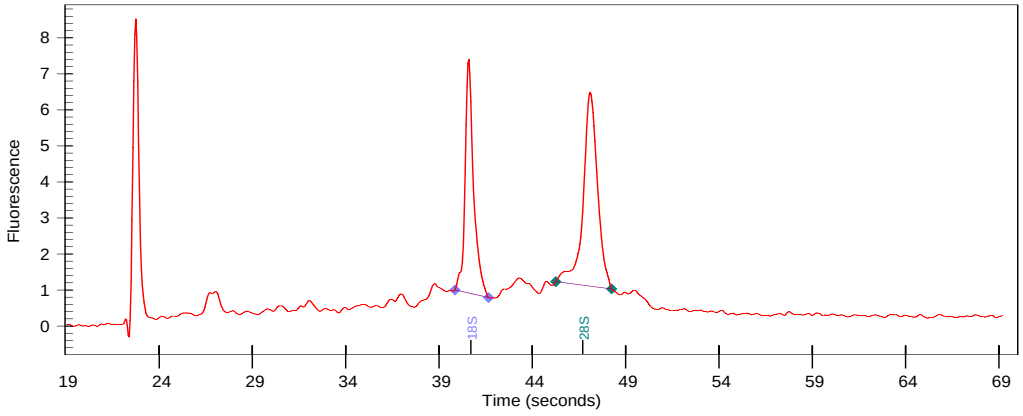
PACKAGE CONTENTS
25 µg Total RNA from the tissues/cells specified below

TOTAL RNA SOURCE
Normal human retina pooled from 29 male/female Caucasians, ages: 20-60; cause of death: sudden death and trauma
No further RNA source information is available.

IMPORTANT NOTE
To prevent contamination by RNases, always wear gloves when handling RNA and exercise care to avoid potential sources of RNase contamination. Avoid multiple freeze/thaw cycles.

QUALITY CONTROL DATA
This lot of total RNA was analyzed by capillary electrophoresis (CE) using an Agilent 2100 Bioanalyzer. The actual electropherogram trace for this RNA is provided below. RNA concentration and purity were evaluated by UV spectrophotometry.
Both the area ratio of the 28S/18S rRNA peaks, and the proportion (relative percentage) of these two peak areas to the total area under the electropherogram provide reliable quantitative estimates of RNA integrity. For both of these criteria, this sample meets or exceeds Clontech standards for high-quality total RNA.

Peak Areas: 28S: 15.5% 18S: 12.7% Ratio 28S/18S: 1.2 Ratio A_{260}/A_{280}: 2.0



REFERENCE
1. Chomczynski, P. & Sacchi, N. (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem.* **162**:156–159.

APPROVED BY:

	Chip ID	Sample ID	Sex	GEO ID	Sample type	Sample Origin	Age at diagnosis	RB1 mutation	Germline mutation	Pass QA
1	560	M22058	F		unilateral	Lohmann	411	Hypermethylation,LOH	FALSE	TRUE
2	561	M22590	F	GSM118317	sporadic unilateral	GEO	603	g.161996G>C,LOH	FALSE	TRUE
3	562	M22860	F	GSM118319	sporadic unilateral	GEO	1725	g.59683C>T,g77051T>C	FALSE	TRUE
4	563	M23869	M	GSM118326	sporadic unilateral	GEO	448	g.64348C>T,LOH	TRUE	TRUE
5	564	M24733	M		unilateral	Lohmann	746	g.73774G>T,LOH	TRUE	TRUE
6	565	M24820	M		unilateral	Lohmann				TRUE
7	566	M20517	M	GSM118320	bilateral(*)	GEO	240	g.77000G>C,LOH	TRUE	TRUE
8	567	M22067	M	GSM118321	sporadic unilateral multifocal	GEO	120	g150098insA	TRUE	TRUE
9	568	M22233	M	GSM118322	unilateral(*)	GEO	897	g.70329C>T,LOH	TRUE	FALSE
10	569	M22641	M	GSM118318	sporadic unilateral	GEO	1051	g.59683C>T,LOH	FALSE	TRUE
11	570	M22731	M		unilateral	Lohmann	1054	g.70330G>A,LOH	FALSE	TRUE
12	571	M23209	M	GSM118323	bilateral(*)	GEO	522	g.64328A>G	TRUE	TRUE
13	572	M23215	M		unilateral	Lohmann	876	g.39562delG,LOH	FALSE	FALSE
14	573	M24430	M		unilateral	Lohmann				TRUE
15	574	M24794	M		unilateral	Lohmann	470	g.78225del7,LOH	FALSE	TRUE
16	575	M22808	F		unilateral	Lohmann				TRUE
17	576	M23449	F	GSM118324	sporadic unilateral	GEO	579	g.78238C>T,LOH	FALSE	FALSE
18		Retina				Lohmann				TRUE
19		Retina02				Clontech				TRUE
20		Retina03				Clontech				TRUE
21		Retina 04				Wolf				TRUE

Table A.1 – DKFZ EP and matrixCGH dataset. In gray are highlighted the samples that did not pass the QA. The asterisks (*) indicates hereditary cases.

	GEO ID	Sample ID	Sample type	Pass QA
1	GSM736290	mad767	single retinoblastoma tumor cell	TRUE
2	GSM736291	mad768	single retinoblastoma tumor cell	TRUE
3	GSM736292	mad769	single retinoblastoma tumor cell	TRUE
4	GSM736293	mad770	single retinoblastoma tumor cell	TRUE
5	GSM736294	mad771	single retinoblastoma tumor cell	TRUE
6	GSM736295	mad772	single retinoblastoma tumor cell	TRUE
7	GSM736296	mad773	single retinoblastoma tumor cell	FALSE
8	GSM736297	mad774	single retinoblastoma tumor cell	TRUE
9	GSM736298	mad775	single retinoblastoma tumor cell	TRUE
10	GSM736299	mad776	single retinoblastoma tumor cell	FALSE
11	GSM736300	mad777	single retinoblastoma tumor cell	TRUE
12	GSM736301	mad778	single retinoblastoma tumor cell	TRUE
13	GSM736302	mad779	single retinoblastoma tumor cell	TRUE
14	GSM736303	mad780	single retinoblastoma tumor cell	FALSE
15	GSM736304	mad781	single retinoblastoma tumor cell	FALSE
16	GSM736305	mad782	single retinoblastoma tumor cell	TRUE
17	GSM736306	mad783	single retinoblastoma tumor cell	TRUE
18	GSM736307	mad784	single retinoblastoma tumor cell	TRUE
19	GSM736308	mad785	single retinoblastoma tumor cell	TRUE
20	GSM736309	mad786	single retinoblastoma tumor cell	TRUE

Table A.2 – GEO GSE29684 dataset. In gray are highlighted the samples that did not pass the QA.

	GEO ID	Sample ID	Sample type	Pass QA
1	GSM736228	mad353	cell line Weril	TRUE
2	GSM736229	mad355	cell line Y79	TRUE
3	GSM736230	mad357	primary tumor	FALSE
4	GSM736231	mad358	primary tumor	TRUE
5	GSM736232	mad359	primary tumor	TRUE
6	GSM736233	mad360	primary tumor	TRUE
7	GSM736234	mad361	primary tumor	TRUE
8	GSM736235	mad362	primary tumor	TRUE
9	GSM736236	mad363	primary tumor	TRUE
10	GSM736237	mad364	primary tumor	FALSE
11	GSM736238	mad365	primary tumor	TRUE
12	GSM736239	mad366	primary tumor	TRUE
13	GSM736240	mad367	primary tumor	TRUE
14	GSM736241	mad368	primary tumor	TRUE
15	GSM736242	mad369	primary tumor	TRUE
16	GSM736243	mad370	primary tumor	TRUE
17	GSM736244	mad371	primary tumor	FALSE
18	GSM736245	mad372	primary tumor	TRUE
19	GSM736246	mad373	cell line RB1 13	TRUE
20	GSM736247	mad374	cell line RB355	TRUE
21	GSM736248	mad375	primary tumor	TRUE
22	GSM736249	mad382	primary tumor	TRUE
23	GSM736250	mad383	primary tumor	TRUE
24	GSM736251	mad384	primary tumor	TRUE
25	GSM736252	mad385	primary tumor	TRUE
26	GSM736253	mad386	primary tumor	TRUE
27	GSM736254	mad387	primary tumor	TRUE
28	GSM736255	mad388	primary tumor	TRUE
29	GSM736256	mad389	primary tumor	TRUE
30	GSM736257	mad390	primary tumor	TRUE
31	GSM736258	mad391	primary tumor	TRUE
32	GSM736259	mad392	primary tumor	TRUE
33	GSM736260	mad393	primary tumor	TRUE
34	GSM736261	mad394	primary tumor	TRUE
35	GSM736262	mad395	primary tumor	TRUE
36	GSM736263	mad402	primary tumor	TRUE
37	GSM736264	mad403	primary tumor	TRUE
38	GSM736265	mad404	primary tumor	FALSE

continued on next page

continued from previous page

	GEO ID	Sample ID	Sample type	Pass QA
39	GSM736266	mad405	primary tumor	TRUE
40	GSM736267	mad406	primary tumor	FALSE
41	GSM736268	mad407	primary tumor	FALSE
42	GSM736269	mad408	primary tumor	FALSE
43	GSM736270	mad409	primary tumor	TRUE
44	GSM736271	mad410	primary tumor	TRUE
45	GSM736272	mad411	primary tumor	TRUE
46	GSM736273	mad542	primary tumor	FALSE
47	GSM736274	mad543	primary tumor	TRUE
48	GSM736275	mad544	primary tumor	FALSE
49	GSM736276	mad617	primary tumor	FALSE
50	GSM736277	mad618	primary tumor	TRUE
51	GSM736278	mad619	primary tumor	FALSE
52	GSM736279	mad620	primary tumor	TRUE
53	GSM736280	mad621	primary tumor	TRUE
54	GSM736281	mad681	primary tumor	FALSE
55	GSM736282	mad686	primary tumor	FALSE
56	GSM736283	mad687	primary tumor	FALSE
57	GSM736284	mad688	primary tumor	FALSE
58	GSM736285	mad707	primary tumor	FALSE
59	GSM736286	mad708	primary tumor	FALSE
60	GSM736287	mad709	xenograft-passaged	TRUE
61	GSM736288	mad710	xenograft-passaged	TRUE
62	GSM736289	mad714	xenograft-passaged	TRUE

Table A.3 – GEO GSE29683 dataset. In gray are highlighted the samples that did not pass the QA.

Sample ID	GEO.ID	Sample description	Pass QA
HOBc Replicate 1	GSM359137	Non-neoplastic primary osteoblast cells with limited live span in vitro	TRUE
HOBc Replicate 2	GSM359138	Non-neoplastic primary osteoblast cells with limited live span in vitro	TRUE
OS1 Replicate 1	GSM359139	conventional osteosarcoma tissue, female, 15 years, femur, grade 3	FALSE
OS1 Replicate 2	GSM359140	conventional osteosarcoma tissue, female, 15 years, femur, grade 3	FALSE
OS11 Replicate 1	GSM359141	osteosarcoma lung metastasis tissue, female, 45 years, lung, grade 1	TRUE
OS11 Replicate 2	GSM359142	osteosarcoma lung metastasis tissue, female, 45 years, lung, grade 1	TRUE
OS15 Replicate 1	GSM359143	conventional osteosarcoma tissue, female, 74 years, femur, grade 2	TRUE
OS15 Replicate 2	GSM359144	conventional osteosarcoma tissue, female, 74 years, femur, grade 2	TRUE
OS16 Replicate 1	GSM359145	osteosarcoma lung metastasis tissue, female, 37 years, lung, grade 2	TRUE
OS16 Replicate 2	GSM359146	osteosarcoma lung metastasis tissue, female, 37 years, lung, grade 2	TRUE
OS18 Replicate 1	GSM359147	conventional osteosarcoma tissue, male, 7 years, femur, grade 2	TRUE
OS18 Replicate 2	GSM359148	conventional osteosarcoma tissue, male, 7 years, femur, grade 2	FALSE
OS24 Replicate 1	GSM359149	conventional osteosarcoma tissue, male, 17 years, femur, grade 3	TRUE
OS24 Replicate 2	GSM359150	conventional osteosarcoma tissue, male, 17 years, femur, grade 3	FALSE
OS4 Replicate 1	GSM359151	osteosarcoma lung metastasis tissue, male, 40 years, lung, grade 3	TRUE
OS4 Replicate 2	GSM359152	osteosarcoma lung metastasis tissue, male, 40 years, lung, grade 3	TRUE
OS6 Replicate 1	GSM359153	osteosarcoma lung metastasis tissue, female, 21 years, lung, grade 3	TRUE
OS6 Replicate 2	GSM359154	osteosarcoma lung metastasis tissue, female, 21 years, lung, grade 3	TRUE
OS9 Replicate 1	GSM359155	conventional osteosarcoma tissue, male, 23 years, tibia, grade 3	TRUE
OS9 Replicate 2	GSM359156	conventional osteosarcoma tissue, male, 23 years, tibia, grade 3	TRUE

Table A.4 – GEO GSE14359 dataset. In gray are highlighted the samples that did not pass the QA.

	Sample ID	GEO ID	age	gender	type	site1	site2	chemotherapeutic response	metastasis	pass QA
1	OSR07	GSM371114	19	male	Osteoblastic	Tibia	Distal	good	Yes	FALSE
2	OSR13	GSM371115	9	female	Teleangiectatic	Femur	Proximal	poor	Yes	TRUE
3	OSR17	GSM371116	13	female	Osteoblastic	Tibia	Proximal	poor	Yes	FALSE
4	OSR18	GSM371117	14	male	Osteoblastic	Tibia	Proximal	poor	No	TRUE
5	OSR20	GSM371118	12	male	Osteoblastic	Femur	Distal	poor	No	TRUE
6	OSR24	GSM371119	14	female	Osteoblastic	Femur	Distal	good	No	TRUE
7	OSR25	GSM371120	13	male	Osteoblastic	Femur	Distal	good	Yes	TRUE
8	OSR26	GSM371121	19	male	Osteoblastic	Femur	Distal	poor	Yes	TRUE
9	OSR27	GSM371122	15	female	Osteoblastic	Femur	Distal	good	No	TRUE
10	OSR28	GSM371123	9	female	Osteoblastic	Tibia	Proximal	good	No	TRUE
11	OSR32	GSM371124	19	male	Osteoblastic	Femur	Distal	poor	No	TRUE
12	OSR35	GSM371125	19	female	Fibroblastic	Tibia	Distal	good	No	TRUE
13	OSR39	GSM371126	18	male	Osteoblastic	Tibia	Distal	good	No	FALSE
14	OSR45	GSM371127	21	male	Osteoblastic	Femur	Distal	poor	Yes	TRUE
15	OSR46	GSM371128	18	male	Chondroblastic	Femur	Distal	poor	No	TRUE
16	OSR47	GSM371129	14	male	Chondroblastic	Femur	Proximal	poor	No	FALSE
17	OSR48	GSM371130	8	male	Osteoblastic	Humerus	Proximal	good	No	TRUE
18	OSR49	GSM371131	24	male	Osteoblastic	Tibia	Proximal	poor	Yes	TRUE
19	OSR50	GSM371132	8	female	Osteoblastic	Tibia	Proximal	poor	No	TRUE
20	OSE01	GSM371133	14	male	Osteoblastic	Femur	Distal	poor	No	FALSE
21	OSE02	GSM371134	38	female	Osteoblastic	Femur	Distal	good	No	FALSE
22	OSE05	GSM371135	12	male	Chondroblastic	Femur	Distal	poor	Yes	TRUE
23	OSE08	GSM371136	20	female	Osteoblastic	Femur	Proximal	poor	No	TRUE
24	OSE09	GSM371137	16	male	Osteoblastic	Tibia	Proximal	poor	No	TRUE
25	OSE14	GSM371138	25	male	Osteoblastic	Humerus	Distal	good	No	TRUE
26	OSE15	GSM371139	15	female	Telangiectatic	Femur	Distal	good	No	TRUE
27	OSE16	GSM371140	13	male	Osteoblastic	Tibia	Proximal	poor	Yes	TRUE

Table A.5 – GEO GSE14827 dataset. In gray are highlighted the samples that did not pass the QA.

	MAQC.ID	GEO.ID	Sample	QA.pass
1	AFX_1_A1	GSM122774	UHRR	FALSE
2	AFX_1_A2	GSM122775	UHRR	TRUE
3	AFX_1_A3	GSM122776	UHRR	TRUE
4	AFX_1_A4	GSM122777	UHRR	TRUE
5	AFX_1_A5	GSM122778	UHRR	TRUE
6	AFX_2_A1	GSM122794	UHRR	TRUE
7	AFX_2_A2	GSM122795	UHRR	TRUE
8	AFX_2_A3	GSM122796	UHRR	TRUE
9	AFX_2_A4	GSM122797	UHRR	TRUE
10	AFX_2_A5	GSM122798	UHRR	TRUE
11	AFX_3_A1	GSM122814	UHRR	TRUE
12	AFX_3_A2	GSM122815	UHRR	TRUE
13	AFX_3_A3	GSM122816	UHRR	TRUE
14	AFX_3_A4	GSM122817	UHRR	TRUE
15	AFX_3_A5	GSM122818	UHRR	TRUE
16	AFX_4_A1	GSM122834	UHRR	TRUE
17	AFX_4_A2	GSM122835	UHRR	FALSE
18	AFX_4_A3	GSM122836	UHRR	TRUE
19	AFX_4_A4	GSM122837	UHRR	FALSE
20	AFX_4_A5	GSM122838	UHRR	FALSE
21	AFX_5_A1	GSM122854	UHRR	TRUE
22	AFX_5_A2	GSM122855	UHRR	TRUE
23	AFX_5_A3	GSM122856	UHRR	TRUE
24	AFX_5_A4	GSM122857	UHRR	TRUE
25	AFX_5_A5	GSM122858	UHRR	FALSE
26	AFX_6_A1	GSM122874	UHRR	TRUE
27	AFX_6_A2	GSM122875	UHRR	TRUE
28	AFX_6_A3	GSM122876	UHRR	TRUE
29	AFX_6_A4	GSM122877	UHRR	TRUE
30	AFX_6_A5	GSM122878	UHRR	TRUE

Table A.6 – GEO GSE5350 dataset. In gray are highlighted the samples that did not pass the QA.

Appendix B

QA

In this appendix are presented selected QA reports generated for every dataset.

EP: The first example is for the DKFZ Affymetrix *GeneChip*® dataset. It clearly demonstrate that the sample M23125 does not pass the QA.

arrayCGH: The second example, starting page 193 shows the QA of the same sample used for arrayCGH. On the first page is shown the overall QA of the first arrayCGH batch from the Zielinski et al. (2005) dataset (see paragraph 3.2.4, page 46). The graphs - top to bottom and left to right - show:

1. The number of probes filtered per ChIP and per filter (see paragraph 3.2.4, page 46). The abbreviations are:
   - M2M: Mean to Median
   - S2N: Signal to Noise
   - MinS: Minimal Signal
   - Csd: Replicate Std Deviation
2. The raw intensities of the red and green channels, *i.e.* emitted from the Cy5 and Cy3 dyes respectively as there is no dye-swap in this dataset.
3. The density plots of the normalized intensities of the red and green channels.
4. The density plot of the intensities of the spots flagged out by the M2M filter.
5. The same as above for the S2N filter.

6. As above for the Csd filter.

All the chips have similar performances and pass the QA. Additional QA are performed per chip and are described in the subsequent pages.

arrayQualityMetrics report for affyBatch(obj)

Browser compatibility

This report uses recent features of HTML 5. Functionality has been tested on these browsers: Firefox 10, Chrome 17, Safari 5.1.2

- Array metadata and outlier detection overview

	array	sampleNames	*1	*2	*3	*4	*5	*6	sample	ScanDate
☐	1	M20517.CEL							1	09/30/03 12:49:19
☐	2	M22058.CEL							2	09/26/03 09:49:57
☐	3	M22067.CEL							3	09/30/03 09:39:25
☐	4	M22233.CEL							4	09/26/03 10:00:25
☐	5	M22590.CEL							5	09/26/03 11:22:23
☐	6	M22641.CEL							6	09/30/03 12:59:56
☐	7	M22731.CEL							7	09/02/03 11:34:12
☐	8	M22808.CEL							8	09/02/03 11:24:01
☐	9	M22860.CEL							9	09/26/03 11:32:45
☐	10	M23209.CEL							10	09/26/03 11:43:08
☑	11	M23215.CEL	x	x	x	x	x		11	09/30/03 10:10:45
☑	12	M23449.CEL					x		12	09/30/03 10:00:21
☐	13	M23869.CEL							13	09/30/03 13:20:40
☐	14	M24430.CEL							14	01/09/04 13:19:37
☐	15	M24733.CEL							15	01/09/04 11:48:38
☐	16	M24794.CEL							16	01/09/04 11:58:58
☐	17	M24820.CEL							17	01/09/04 11:25:50
☐	18	Retina.CEL							18	01/09/04 13:09:17

The columns named *1, *2, ... indicate the calls from the different outlier detection methods:

1. outlier detection by Distances between arrays
2. outlier detection by Boxplots
3. outlier detection by Relative Log Expression (RLE)
4. outlier detection by Normalized Unscaled Standard Error (NUSE)
5. outlier detection by MA plots
6. outlier detection by Spatial distribution of M

The outlier detection criteria are explained below in the respective sections. Arrays that were called outliers by at least one criterion are marked by checkbox selection in this table, and are indicated by highlighted lines or points in some of the plots below. By clicking the checkboxes in the table, or on the corresponding points/lines in the plots, you can modify the selection. To reset the selection, reload the HTML page in your browser.

At the scope covered by this software, outlier detection is a poorly defined question, and there is no 'right' or 'wrong' answer. These are hints which are intended to be followed up manually. If you want to automate outlier detection, you need to limit the scope to a particular platform and experimental design, and then choose and calibrate the metrics used.

Section 1: Between array comparison

- Figure 1: Distances between arrays.

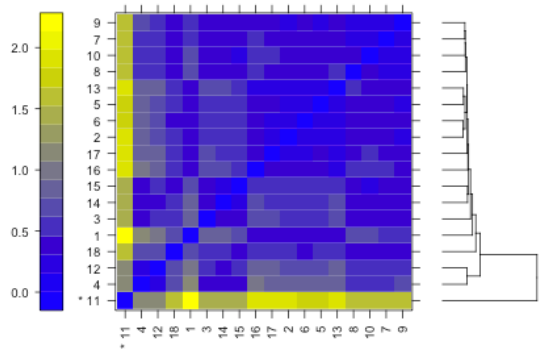


Figure 1 (PDF file) shows a false color heatmap of the distances between arrays. The color scale is chosen to cover the range of distances encountered in the dataset. Patterns in this plot can indicate clustering of the arrays either because of intended biological or unintended experimental factors (batch effects). The distance d_{ab} between two arrays a and b is computed as the mean absolute difference (L_1-distance) between the data of the arrays (using the data from all probes without filtering). In formula, d_{ab} = mean | M_{ai} - M_{bi} |, where M_{ai} is the value of the i-th probe on the a-th array. Outlier detection was performed by looking for arrays for which the sum of the distances to all other arrays, $S_a = \Sigma_b\, d_{ab}$ was exceptionally large. One such array was detected, and it is marked by an asterisk, *.

- Figure 2: Outlier detection for Distances between arrays.

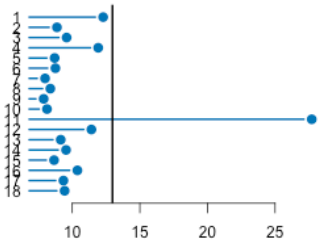


Figure 2 (PDF file) shows a bar chart of the sum of distances to other arrays S_a, the outlier detection criterion from the previous figure. The bars are shown in the original order of the arrays. Based on the distribution of the values across all arrays, a threshold of 13 was determined, which is indicated by the vertical line. One array exceeded the threshold and was considered an outlier.

- Figure 3: Principal Component Analysis.

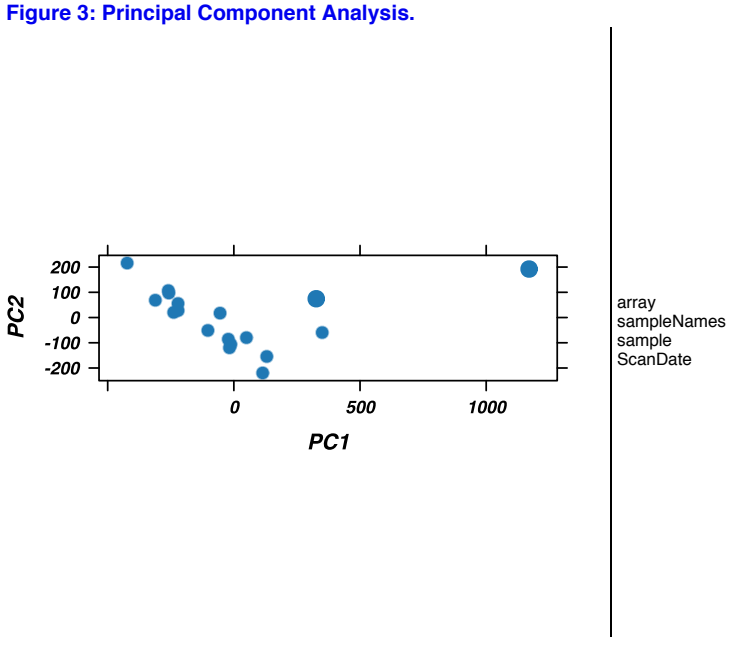


Figure 3 (PDF file) shows a scatterplot of the arrays along the first two principal components. You can use this plot to explore if the arrays cluster, and whether this is according to an intended experimental factor (you can indicate such a factor by color using the 'intgroup' argument), or according to unintended causes such as batch effects. Move the mouse over the points to see the sample names.
Principal component analysis is a dimension reduction and visualisation technique that is here used to project the multivariate data vector of each array into a two-dimensional plot, such that the spatial arrangement of the points in the plot reflects the overall data (dis)similarity

between the arrays.

Section 2: Array intensity distributions

- Figure 4: Boxplots.

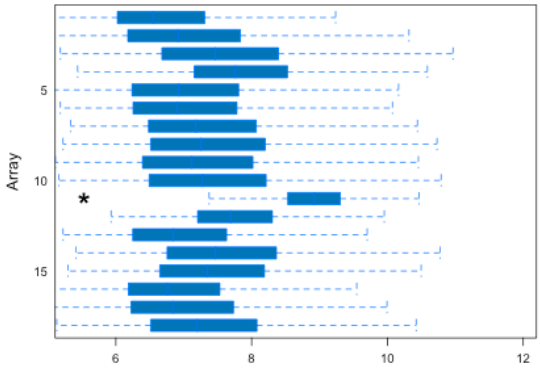


Figure 4 (PDF file) shows boxplots representing summaries of the signal intensity distributions of the arrays. Each box corresponds to one array. Typically, one expects the boxes to have similar positions and widths. If the distribution of an array is very different from the others, this may indicate an experimental problem. Outlier detection was performed by computing the Kolmogorov-Smirnov statistic K_a between each array's distribution and the distribution of the pooled data.

- Figure 5: Outlier detection for Boxplots.

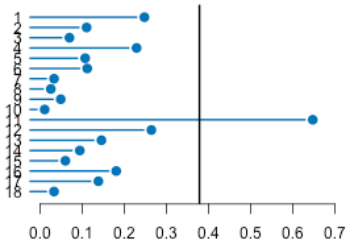


Figure 5 (PDF file) shows a bar chart of the Kolmogorov-Smirnov statistic K_a, the outlier detection criterion from the previous figure. The bars are shown in the original order of the arrays. Based on the distribution of the values across all arrays, a threshold of 0.379 was determined, which is indicated by the vertical line. One array exceeded the threshold and was considered an outlier.

- Figure 6: Density plots.

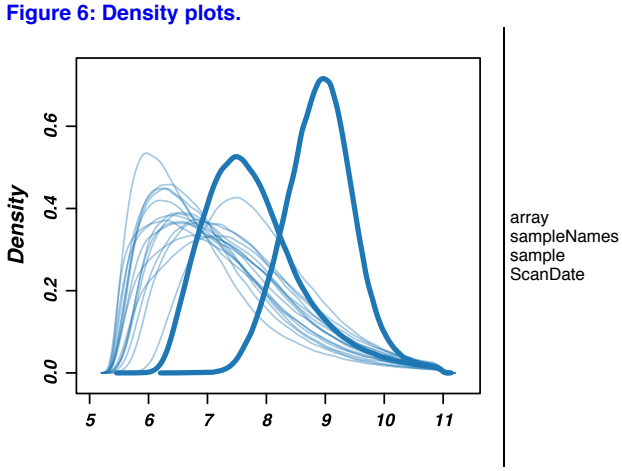


Figure 6 (PDF file) shows density estimates (smoothed histograms) of the data. Typically, the distributions of the arrays should have similar shapes and ranges. Arrays whose distributions are very different from the others should be considered for possible problems. Various features of the distributions can be indicative of quality related phenomena. For instance, high levels of background will shift an array's distribution to the right. Lack of signal diminishes its right right tail. A bulge at the upper end of the intensity range often indicates signal saturation.

Section 3: Variance mean dependence

- Figure 7: Standard deviation versus rank of the mean.

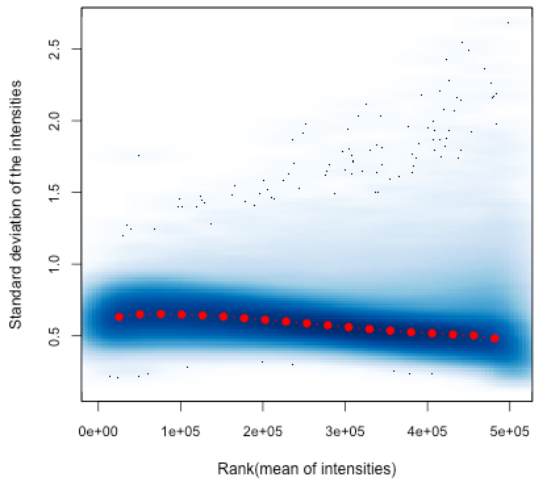


Figure 7 (PDF file) shows a density plot of the standard deviation of the intensities across arrays on the *y*-axis versus the rank of their mean on the *x*-axis. The red dots, connected by lines, show the running median of the standard deviation. After normalisation and transformation to a logarithm(-like) scale, one typically expects the red line to be approximately horizontal, that is, show no substantial trend. In some cases, a hump on the right hand of the x-axis can be observed and is symptomatic of a saturation of the intensities.

Section 4: Affymetrix specific plots

- Figure 8: Relative Log Expression (RLE).

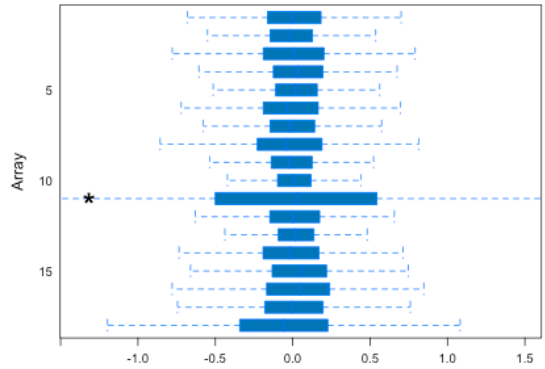


Figure 8 (PDF file) shows the *Relative Log Expression (RLE)* plot. Arrays whose boxes are centered away from 0 and/or are more spread out are potentially problematic. Outlier detection was performed by computing the Kolmogorov-Smirnov statistic R_a between each array's RLE values and the pooled, overall distribution of RLE values.

- Figure 9: Outlier detection for Relative Log Expression (RLE).

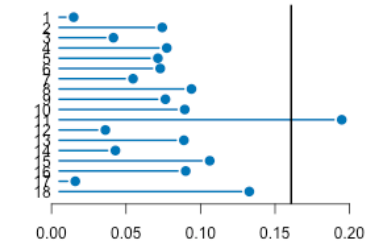


Figure 9 (PDF file) shows a bar chart of the Kolmogorov-Smirnov statistic R_a of the RLE values, the outlier detection criterion from the

previous figure. The bars are shown in the original order of the arrays. Based on the distribution of the values across all arrays, a threshold of 0.161 was determined, which is indicated by the vertical line. One array exceeded the threshold and was considered an outlier.

- Figure 10: Normalized Unscaled Standard Error (NUSE).

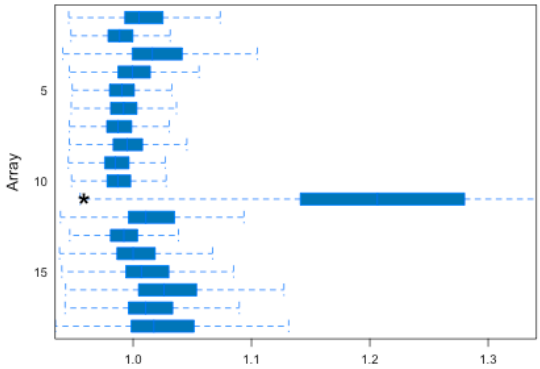


Figure 10 (PDF file) shows the *Normalized Unscaled Standard Error (NUSE)* plot. For each array, the boxes should be centered around 1. An array were the values are elevated relative to the other arrays is typically of lower quality. Outlier detection was performed by computing the 75% quantile N_a of each array's NUSE values and looking for arrays with large N_a.

- Figure 11: Outlier detection for Normalized Unscaled Standard Error (NUSE).

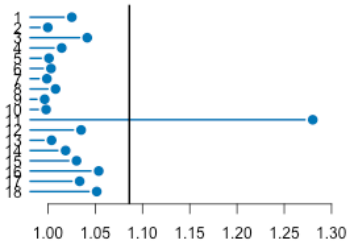


Figure 11 (PDF file) shows a bar chart of the N_a, the outlier detection criterion from the previous figure. The bars are shown in the original order of the arrays. Based on the distribution of the values across all arrays, a threshold of 1.09 was determined, which is indicated by the vertical line. One array exceeded the threshold and was considered an outlier.

- Figure 12: RNA digestion plot.

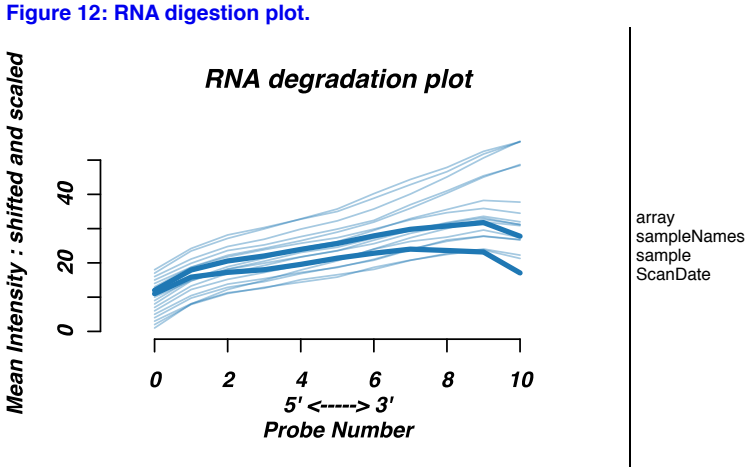


Figure 12 (PDF file) shows the *RNA digestion* plot. The shown values are computed from the preprocessed data (after background correction and quantile normalisation). Each array is represented by a single line; move the mouse over the lines to see their corresponding sample names. The plot can be used to identify array(s) that have a slope very different from the others. This could indicate that the RNA used for that array has been handled differently from what was done for the other arrays.

- Figure 13: Perfect matches and mismatches.

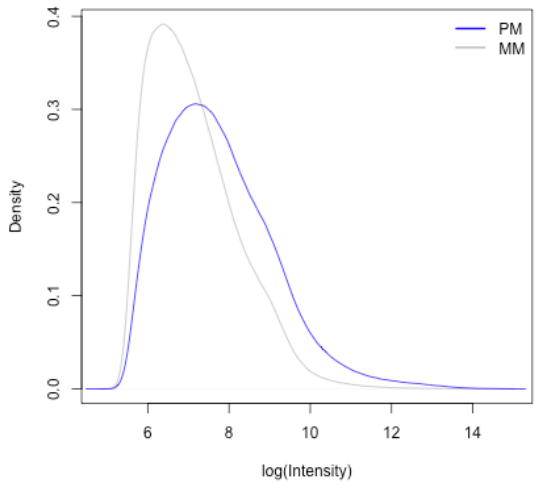


Figure shows the density distributions of the $\log_2$ intensities grouped by the matching type of the probes. The blue line shows a density estimate (smoothed histogram) from intensities of perfect match probes (PM), the grey line, one from the mismatch probes (MM). We expect that MM probes have poorer hybridization than PM probes, and thus that the PM curve be to the right of the MM curve.

Section 5: Individual array quality

- Figure 14: MA plots.

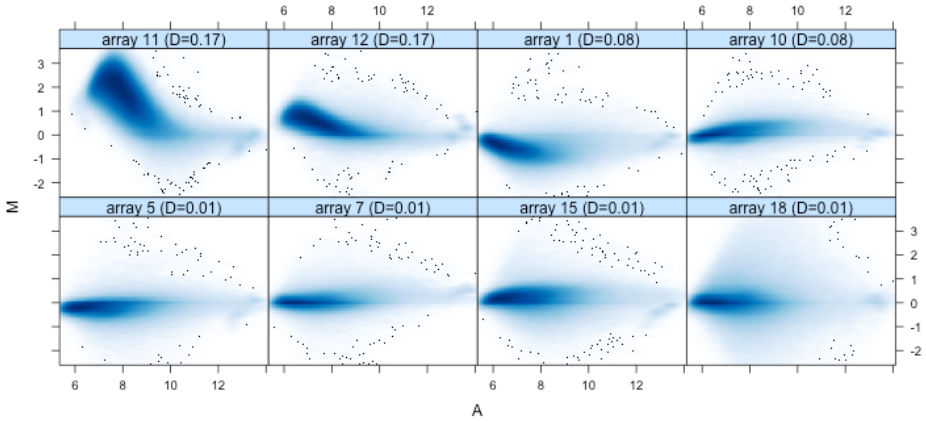


Figure 14 (PDF file) shows MA plots. M and A are defined as:
$M = \log_2(I_1) - \log_2(I_2)$
$A = 1/2\ (\log_2(I_1)+\log_2(I_2))$,
where I_1 is the intensity of the array studied, and I_2 is the intensity of a "pseudo"-array that consists of the median across arrays. Typically, we expect the mass of the distribution in an MA plot to be concentrated along the M = 0 axis, and there should be no trend in M as a function of A. If there is a trend in the lower range of A, this often indicates that the arrays have different background intensities; this may be addressed by background correction. A trend in the upper range of A can indicate saturation of the measurements; in mild cases, this may be addressed by non-linear normalisation (e.g. quantile normalisation).
Outlier detection was performed by computing Hoeffding's statistic D_a on the joint distribution of A and M for each array. Shown are the 4 arrays with the highest value of D_a (top row), and the 4 arrays with the lowest value (bottom row). The value of D_a is shown in the panel headings. 2 arrays had D_a>0.15 and were marked as outliers. For more information on Hoeffing's D-statistic, please see the manual page of the function `hoeffd` in the `Hmisc` package.

- Figure 15: Outlier detection for MA plots.

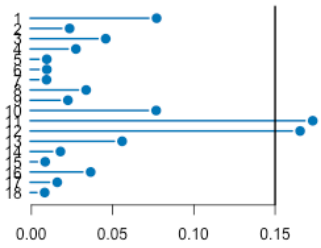


Figure 15 (PDF file) shows a bar chart of the Hoeffding's statistic D_a, the outlier detection criterion from the previous figure. The bars are shown in the original order of the arrays. A threshold of 0.15 was used, which is indicated by the vertical line. 2 arrays exceeded the threshold and were considered outliers.

- Figure 16: Spatial distribution of M.

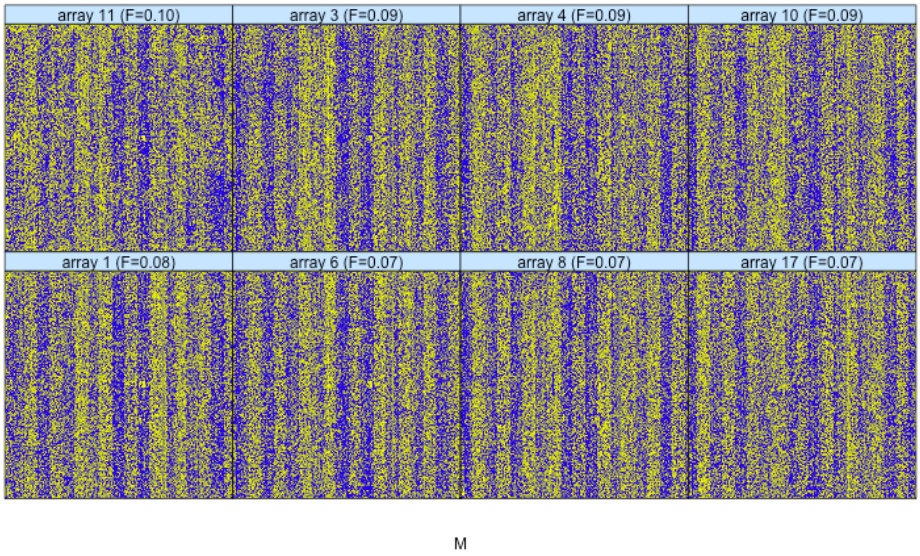


Figure 16 (PDF file) shows false color representations of the arrays' spatial distributions of feature intensities (M). Normally, when the features are distributed randomly on the arrays, one expects to see a uniform distribution; control features with particularly high or low intensities may stand out. The color scale is proportional to the ranks of the probe intensities. Note that the rank scale has the potential to amplify patterns that are small in amplitude but systematic within an array. It is possible to switch off the rank scaling by modifying the argument `scale` in the call of the `aqm.spatial` function.
Outlier detection was performed by computing F_a , the sum of the absolutes value of low frequency Fourier coefficients, as a measure of large scale spatial structures. Shown are the 4 arrays with the highest value of S (top row), and the 4 arrays with the lowest value (bottom row). The value of F_a is shown in the panel headings.

- Figure 17: Outlier detection for Spatial distribution of M.

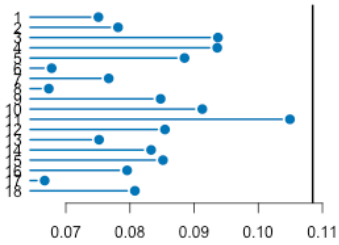
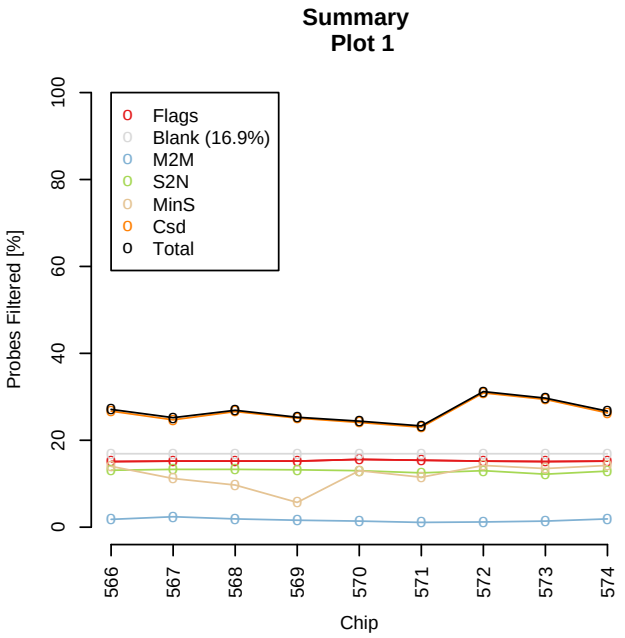
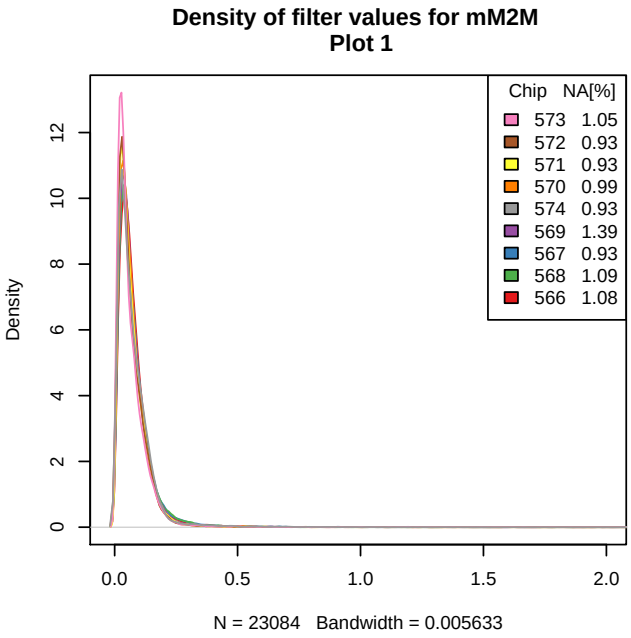


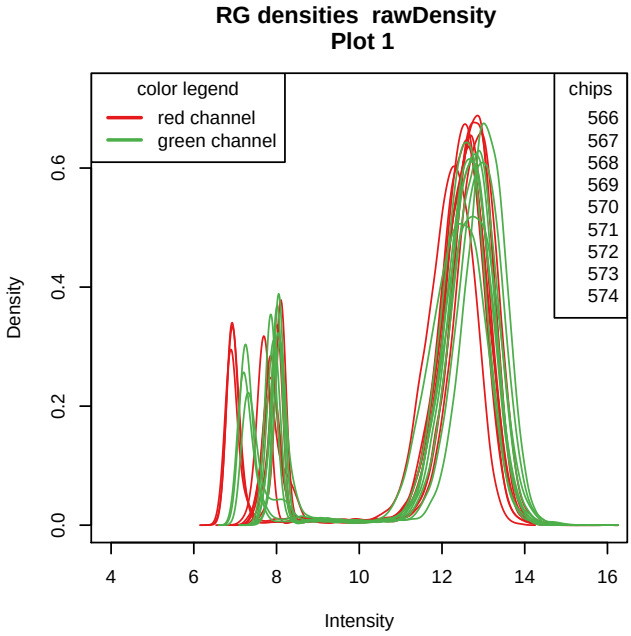
Figure 17 (PDF file) shows a bar chart of the F_a, the outlier detection criterion from the previous figure. The bars are shown in the original order of the arrays. Based on the distribution of the values across all arrays, a threshold of 0.108 was determined, which is indicated by the vertical line. None of the arrays exceeded the threshold and was considered an outlier.

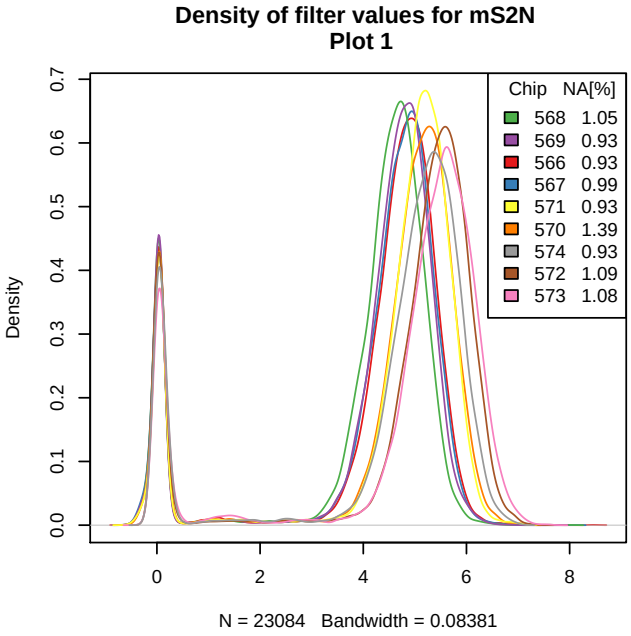
This report has been created with arrayQualityMetrics 3.11.7 under R version 2.14.1 Patched (2011-12-23 r57982).

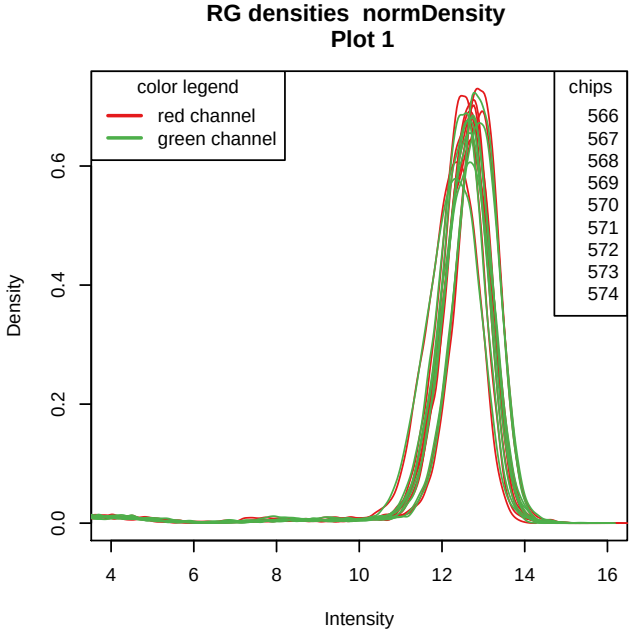
(Page generated on Thu Mar 22 23:24:41 2012 by hwriter)

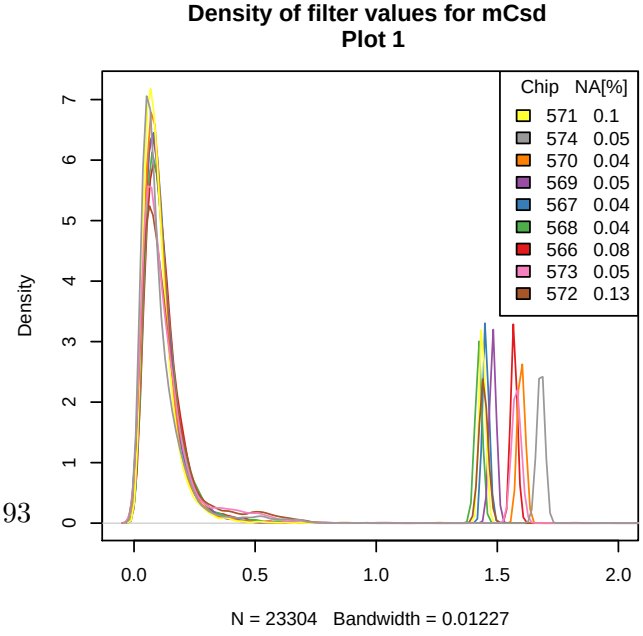

Summary
Plot 1
Flags
Blank (16.9%)
M2M
S2N
MinS
Csd
Total
Probes Filtered [%]
Chip


Density of filter values for mM2M
Plot 1
Chip NA[%]
573 1.05
572 0.93
571 0.93
570 0.99
574 0.93
569 1.39
567 0.93
568 1.09
566 1.08
Density
N = 23084 Bandwidth = 0.005633


RG densities rawDensity
Plot 1
color legend
red channel
green channel
chips
566
567
568
569
570
571
572
573
574
Density
Intensity


Density of filter values for mS2N
Plot 1
Chip NA[%]
568 1.05
569 0.93
566 0.93
567 0.99
571 0.93
570 1.39
574 0.93
572 1.09
573 1.08
Density
N = 23084 Bandwidth = 0.08381


RG densities normDensity
Plot 1
color legend
red channel
green channel
chips
566
567
568
569
570
571
572
573
574
Density
Intensity


Density of filter values for mCsd
Plot 1
Chip NA[%]
571 0.1
574 0.05
570 0.04
569 0.05
567 0.04
568 0.04
566 0.08
573 0.05
572 0.13
Density
N = 23304 Bandwidth = 0.01227

B.0.1 Fluorescence gradient

To ensure that the chip were homogeneously hybridized, the raw intensity of the red and green channel are projected onto the chip layout: Figure B.1 and B.2. In addition, the ratio of both intensities is calculated and displayed in Figure B.3. The green channel shows a clear technical artefact

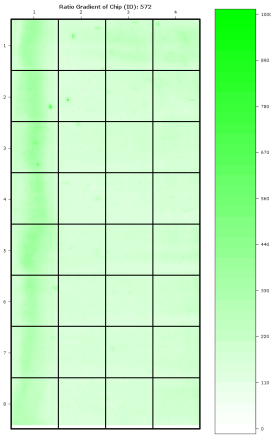


Figure B.1 – M23215 Cy3 gradient

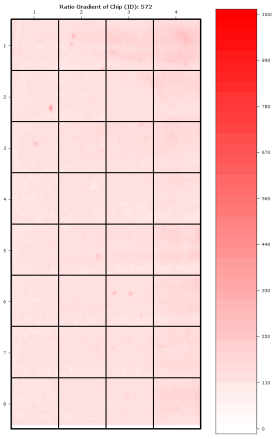


Figure B.2 – M23215 Cy5 gradient

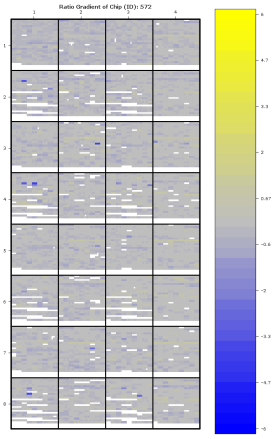


Figure B.3 – M23215 Cy3/Cy5 ratio gradient

in the left-most column. The red channel presents some localized artefacts. The ratio plot shows only very local high or low ratios, unrelated to the technical artefacts observed in the separate channels.

B.0.2 Intensity distribution

As already shown in the previous QA, the intensity differs between the channels: Figure B.4. This variation is successfully controlled for and normalized, see Figure B.5. In addition, a normalization for the print order is as well applied (QA not shown).

B.0.3 Scatterplot

The final QA consist of a comparison of the intensities for the M23125 sample with that of a control - an opposite sex-matched healthy donors' blood sample, female in the present example. In the Figure B.6, the raw intensities are displayed and colored according to their filter status; the M23215 is on the y axis, while the control is on the x-axis. In Figure B.7, the normalized intensities are displayed. On neither figure can a technical artefact be identified. Finally in Figure B.8, the density of the normalized intensities is displayed - from dark blue (low density) to dark red (high

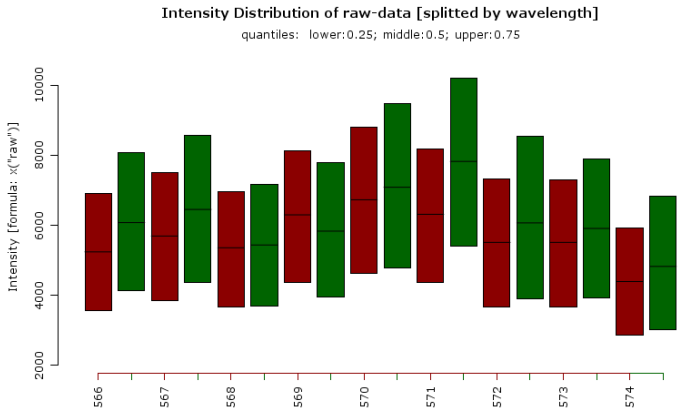


Figure B.4 – Cy3 and Cy5 raw intensities distribution

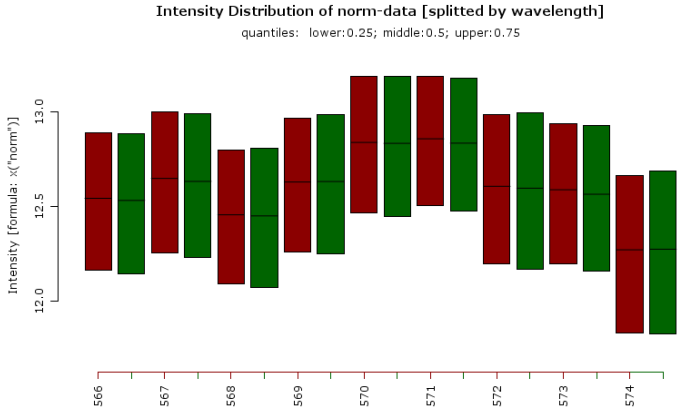


Figure B.5 – Cy3 and Cy5 normalized intensities distribution

density). As expected, the sample having a few CNVs, most of the data lies on the diagonal, *i.e.* the sample and the control having mostly the same genomic copy numbers.

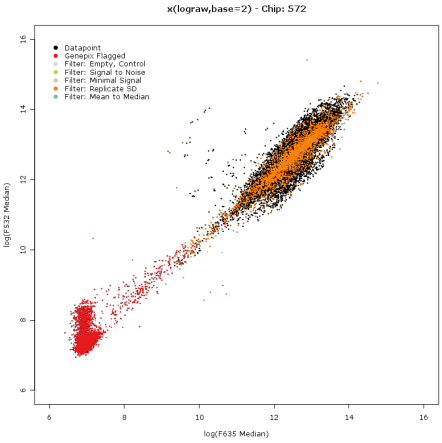


Figure B.6 – M23215 (y-axis) *vs.* control (x-axis) raw intensities. The dots are colored according to their filter status.

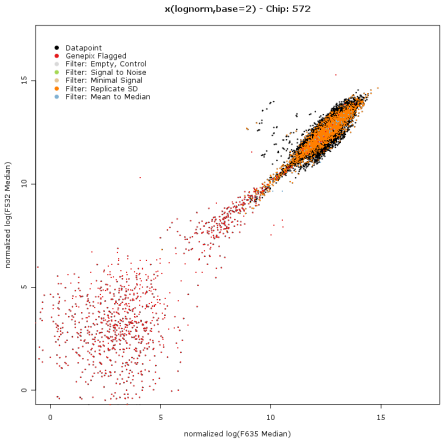


Figure B.7 – M23215 (y-axis) *vs.* control (x-axis) normalized intensities. The dots are colored according to their filter status.

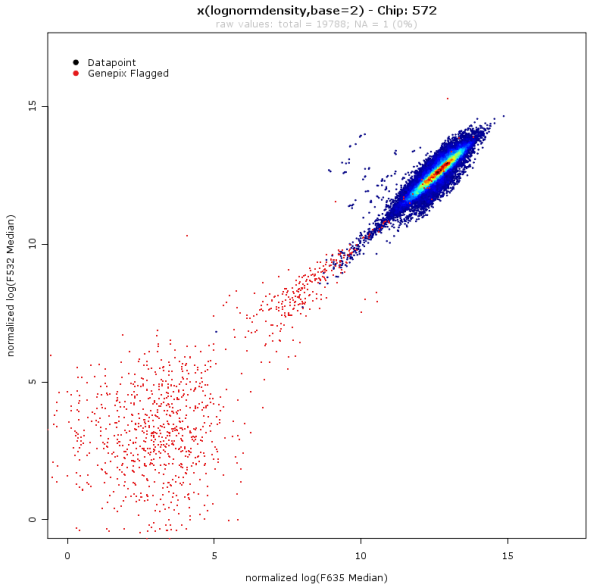


Figure B.8 – M23215 (y-axis) *vs.* control (x-axis) normalized intensities. The dots density is represented by a gradient dark blue (low) - dark red(high). The spot flagged out during the original image analysis (Genepix flagged) are in red and ignored in the density gradient.

Appendix C

Ebased custom CDF

This appendix shows the manuscript: **Ensembl based Custom Definition File for Affymetrix GeneChip** as submitted for peer-review at the journal **Bioinformatics**.

BIOINFORMATICS

Vol. 00 no. 00 2012
Pages 1–8

Ensembl based Custom Definition File for Affymetrix GeneChip

Nicolas Delhomme [1,2]*, Frédéric Blond [1], Natalia Becker [1,3], Michael Hain [1], Peter Lichter [1] and Grischa Toedt [1,4]

[1]Division Molecular Genetics, German Cancer Research Center, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany
[2]Genome Biology Computational Support, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany (current address)
[3]Division Biostatistics, German Cancer Research Center, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany
[4]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany (current address)

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: In addition to the chip definition file (CDF) provided by Affymetrix, there are a number of custom CDFs available, which redefine the aggregation of probes into probe-sets representative of a gene. Despite being more accurate, the usage of these CDF's is limited, due to infrequent updates and a conservative generation process that ignores up to 30% of the probes. To address these issues, we present a custom CDF, retaining all map-able probes, which is automatically updated every other month, in concordance with the Ensembl database.

Results: We compared our custom Ensembl based CDF (Ebased) with the publicly available custom and Affymetrix CDFs. The Affymetrix probes' assignation differs by 30% from every other CDF. Custom CDFs are at least 90% identical. Applied on experimental data, our Ebased CDF is more efficient than the other custom ones and unravels considerably more information due to its extended probes' assignation that adds 30% more probe-sets. It provides the most up-to-date annotation and is a valuable tool to mine the Affymetrix microarray data available from GEO and ArrayExpress.

Availability: The Ebased CDFs and the `customCDF` R package, Bioconductor compliant, are available from the web address: http://www.dkfz.de/en/genetics/pages/projects/bioinformatics/Custom_Chip_Definition_File.html.

Contact: delhomme@embl.de

1 INTRODUCTION

The human genome assembly is not as stable as expected after the completion of the Human Genome Project in 2003. The development of new technologies, such as Next-Generation Sequencing, is completing, correcting and even extending the so-called reference genome at an unprecedented pace. This represents an issue for all the probe-based techniques like human DNA microarrays, as it affects the probe's annotation. It seems particularly critical for Affymetrix microarrays, since the technology relies on a collection of probes identifying an individual transcript, the so-called probe-sets, to evaluate gene expression. The probes' assignment into a probe-set is defined in a Chip Definition File (*CDF*), which is updated, for homo sapiens chips, only once per genome build, *i.e.* early 2006 for the genome build 36 and end 2009 for the genome build 37. Due to the human genome and gene build changes since then, many probe-sets are known to hybridize different and sometimes multiple genes. These wrong annotations greatly affect the experimental data calculated from the respective microarrays. This issue has already been addressed by different groups (Gautier et al., 2004; Dai et al., 2005; Lu et al., 2007; Ferrari et al., 2007), which created custom CDFs. Among these, the ones from Ferrari et al. (2007) (GeneAnnot (*GA*)) and from Dai et al. (2005) (*MBNI*) have been integrated into the Bioconductor package suite (Gentleman et al., 2004). However, their update frequency is rather low (Dai -12 month, Ferrari - 6 month) and does not keep up with the genome updates. In addition, both of them have a conservative approach and discard 20-30% of the probes as being uninformative. Hence, we introduce a new Ensembl based custom CDF (*Ebased*), which retains as many probes as possible and is updated as frequently (every other month) as the Ensembl (Flicek et al., 2011) database. Similarly to existing CDFs, it defines probe-sets of variable size with a minimum of five members as suggested by Lu et al. (2007), to be robust to outliers and have a decent statistical power when summarizing intensity values. To track the expression of transcript variants, we create transcript-centric probe-sets (Lu and Zhang, 2006; Sandberg and Larsson, 2007; Stalteri and Harrison, 2007) and in cases where this is not applicable, we revert to gene-centric probe-sets. In addition, probes left un-annotated by this process are combined into genomic loci probe-sets, which potentially identify unknown or modified transcripts. Finally, multi-mapping probes, unlike any

*to whom correspondence should be addressed

other custom CDFs, are not discarded but assembled into multi-gene probe-sets. The last decade has seen the generation of millions of array based expression profiles. *GEO*, the Gene Expression Omnibus (Edgar et al., 2002; Barrett et al., 2007), and ArrayExpress (Parkinson et al., 2009), the major gene expression / molecular abundance repositories, hold data from more than 735,000 samples. Alone, the most commonly used Affymetrix GeneChip® platform, the Affymetrix Human Genome U133 Plus 2.0 Array, has been used in almost 2400 studies totaling more than 65,200 samples. This wealth of data has been generated and analyzed to answer very specific biological questions, whereas it could be used to answer unrelated questions without having to repeat experiments or be mined in a genome wide fashion as pioneered recently by Lukk et al. (2010). A possible weakness of such approaches is that most of the available data has neither had its annotations updated nor its expression values recalculated, taking into consideration the increased knowledge about the human genome. Using up to date CDFs and extended annotations has the potential to shed light on splice variants differential expression or reveal un-annotated expressed loci like these SUTs (stable un-annotated transcripts) and CUTs (cryptic unstable transcripts) identified in yeast (Xu et al., 2009) from either published or newly generated data. To support such approaches, we are providing optimized Ensembl based custom CDFs, together with their respective gene and probe annotations. We demonstrate that these files perform better than the original Affymetrix annotations and other custom CDFs and that they reveal previously undetectable information.

2 METHODS

2.1 Aligning the probe sequences to the reference genome

For a given GeneChip®, the probe sequences are retrieved from the corresponding Bioconductor probe package. For any maintained Ensembl release, all the sequences are aligned to a set of two nucleotide databases addressing different "genomic" levels. The first one, "cdna", contains the sequences of all Ensembl transcripts, including all known splice variants, as well as non-coding RNAs. The second one - "dna" - is the complete reference genome used by Ensembl for the given release. The alignments are performed using the short read aligner: "bowtie" (Langmead et al., 2009), with the following parameters: *-v 2 -y -a*. A valid alignment therefore has a maximum of 2 mismatches along the whole probe length (25nt), as suggested by He et al. (2005).

2.2 Getting the gene annotation

The `biomaRt` R package (Durinck et al., 2005) is used to connect the Ensembl mart database to retrieve the genic information: its chromosomal mapping, and potential cross-references with Ensembl, Entrez Gene, Locuslink, and UniProt.

2.3 CDF and additional packages generation

Using the alignment information, probes are grouped into probe-sets, transcript-centric whenever possible, ensuring that they contain at least *5* probes, a requirement for unbiased downstream analysis (Lu et al., 2007). If that fails, multi-transcript or gene-centric probe-sets are created. Finally, probes that fail to map any gene and are separated by a maximum of 1kb from each other are grouped into probe-sets. All the probe-sets are bundled into a "CDF" R package and, in addition, the probe (*e.g.* nucleotide sequence, grid position, etc.) and gene annotation (*e.g.* genomic locus, Ensembl gene ID, EntrezGene ID, etc.) packages are generated.

2.4 Other CDFs, annotation and probe R packages

The original Affymetrix (http://www.affymetrix.com) and GeneAnnot CDFs were downloaded from Bioconductor (version 2.3). The MBNI ones were retrieved from the Brainarray website (http://brainarray.mbni.med.umich.edu /Brainarray; version 11). The corresponding annotation and probe packages were retrieved from the same sources. These CDFs will be referred to as *Affymetrix, GA* and *MBNI* respectively. Our custom CDF will be referred to as *Ebased*.

2.5 Comparison of CDFs

CDFs from the different providers are first compared for the probe assignment into probe-sets. Every probe-set from a CDF (the query) is compared with every probe-set of a second CDF (the reference). Similarities are assessed for every possible combination of query and reference. Then, for those probe-sets that are identical, the Entrez Gene IDs are retrieved from the annotation packages and compared for every possible combination of query and reference.

2.6 CDF update rate evaluation

The comparison of the annotation and probes packages generated using Ensembl version *50, 51* and *52* (the packages version *1.0.3, 1.0.4* and *1.0.5*, respectively) allows the assessment of the changes occurring between updates of the human genome. As above, the probes assignment into probe-sets and the Entrez Gene ID annotations are compared sequentially between versions.

2.7 Benchmarking dataset: acute lymphoblastic leukemia

The previously published adult acute lymphoblastic leukemia (*ALL*) dataset (Chiaretti et al., 2004, 2005) has been frequently used for benchmarking and comparing algorithms (Jiang and Gentleman, 2007; Oron et al., 2008) and is publicly available at Bioconductor. It consists of 128 samples hybridized to the hgu95av2 Affymetrix GeneChip® microarray. In this study, the dataset is restricted to the 79 B-cell samples having either a t(9;22)(q34;q11) translocation resulting in the *BCR/ABL* gene fusion or being negative for that genotypic trait. These genotypes are identified as *BCR/ABL* and *NEG* further on. The dataset CEL files (raw data) are processed in R and their quality is assessed using the `arrayQualityMetrics` package (Kauffmann et al., 2009). All the arrays that do not fail any of the QA tests (n=66) are kept. This batch is then background-corrected, normalized and summarized using `RMA` (Irizarry et al., 2003; Bolstad et al., 2003). This last step is performed with every CDF, *i.e.* the original and the three custom ones (GA, MBNI, Ebased) resulting in four different expression matrices.

2.8 Probe-set size and expression variability depending on the CDF

A critical question to the ALL dataset is the identification of differential expression between the *BCR/ABL* and the NEG genotypes, while a critical question to this manuscript is to determine the CDFs' effect on calling differential expression. To assess the effect of the probe-set size and of the variability of its probes' expression values, every expression matrix is first filtered for non-specific probe-sets using the `genefilter` R package. As in Jiang and Gentleman (2007), non-informative probe-sets are filtered out. A probe-set is considered non-informative, when its expression values across all samples is almost invariable; *i.e.* when the Inter Quartile Range (IQR) of these values is smaller than a cutoff value of *0.5*. For the Ebased matrix, an additional filtering step is performed and two sets are generated that retains the probe-sets associated with a maximum of one or two genes, respectively. Then, for every probe-set, the raw expression values of its member probes are retrieved and their standard deviation calculated. At the same time, the probe-set size is retrieved. Finally, the comparison between CDFs is performed using either all filtered probe-sets; *i.e.* the four sets have different sizes, or using only those probe-sets common to all four CDFs,

based on their gene mapping; *i.e.* all four sets are a subset of the original ones and all have the same size.

2.9 Gene Set Enrichment Analysis using a linear model (regression) framework

Gene Set Enrichment Analysis (GSEA) is performed as suggested by Oron et al. (2008). Every expression matrix is filtered for non-specific probe-sets using the `genefilter` R package. Probe-sets are kept, which have an IQR greater or equal to *0.5* within all samples and a single Entrez Gene ID assigned. Using the `Category` R package, these genes are mapped to their chromosomal locations. Chromosomal band and sub-bands containing more than 5 genes are kept for the downstream analysis and their probe-set members' residuals calculated.

2.10 Comparison of the "Gene Set Residuals" obtained for the different CDFs

Every obtained residual set is subdivided per genotype (*BCR/ABL* or NEG). Subsets belonging to the same genotype are compared. The median of every sample residuals distribution within a subset are computed and used to cluster the CDFs using the `pvclust` R package (Suzuki and Shimodaira, 2006). The distance method used is Euclidean and the linkage method is complete.

2.11 "Gene Set Residuals for the Y chromosome

As above, every residuals set is subdivided per genotype (*BCR/ABL* or NEG). The residuals of probe-sets located on the non-autosomal part of the Y chromosome are extracted and their distribution compared between CDFs by the mean of an unsupervised classification using an Expectation-Maximization (EM) iterative method (`mclust` R package).

2.12 Genotype comparison using a linear model

For every expression matrix, we apply a linear model using the R `limma` package (Smyth, 2004) to find the genes differentially expressed between the two different disease's genotypes of interest (*BCR/ABL* vs. NEG). Only the genes having an adjusted p-value (Benjamini and Hochberg, 1995) lower than *0.05* are kept. Finally, the four resulting gene lists are compared pair-wise.

3 IMPLEMENTATION

The CDF generation has been implemented in an R (R Development Core Team, 2009) package: `customCDF` that generates not only the CDFs but as well the probe and annotation packages, as exemplified on the webpage: http://www.dkfz.de/en/genetics/pages/projects/bioinformatics/Custom_Chip_Definition_File.html. The `customCDF` package is available from Bioconductor: http://bioconductor.org/packages/devel/bioc/html/customCDF.html to retrieve, use or generate these CDFs.

4 RESULTS AND DISCUSSION

For the last decade, array based expression profiling has generated a huge amount of data that has often been only superficially analyzed. There are several reasons for this: researchers focused on a rather small set of genes, either by interest or because the gene interaction knowledge was still too sparse; the technology and analyses procedures were still being developed and optimized; *etc.* Nowadays, this technique is mature, well understood and analysis pipelines have been standardized, as has the format for storing and retrieving results (MIAME: Brazma et al. (2001)). Databases have been developed toward this purpose: *i.e.* GEO and ArrayExpress that hold hundreds of thousands of microarray-based experiments. In addition, new technologies have unraveled gene networks (Genomic Regulatory Network (*GRN*), Davidson (2001)) and their dynamics (Cheong et al., 2008); de-novo sequencing has corrected and/or completed many genome assemblies and extended their gene annotations. These developments make it realistic to mine the data stored in the array databases, aiming at networks of genes rather than at a few genes only, as recently pioneered by Lukk et al. (2010). A pre-requisite for this type of analysis is to have up-to-date probes' information: a moving target due to the frequent genomic annotations' variations and a situation to which Affymetrix microarrays are more sensitive by design. Several groups are already providing custom mapping of the Affymetrix probes into probe-sets; however, their update frequency is quite low and their generation process very stringent, leaving out on average a third of the probes present on the arrays. To rescue these probes, we group probes spanning intergenic loci and we allow probes to be part of different probe-sets, creating multi-gene probe-sets. Most of these target two genes only, and as a given tissue expresses only a minor part of its genetic repertoire (Su et al., 2002), there is a high likelihood that such a probe-set measures a single gene effect. Moreover, many of these probe-sets' targets are gene families or gene duplication, the biological importance of which has been shown in numerous studies (Bailey and Eichler, 2006; Marques-Bonet et al., 2009).

4.1 One to many probes to probe-sets mapping

Re-using probes to define new probe-sets extends the number of features being monitored by *20%*, *30%* and *40%* for the hgu95av2, hgu133plus2 and hgu133a chips, respectively; at the cost of measuring the expression of different genes as shown in Figure 1. It is evident that not all of these probe-sets are of interest,

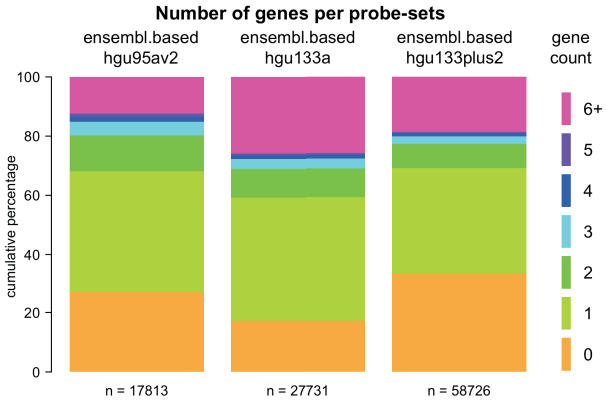


Fig. 1. Number of genes per probe-sets for three common GeneChips

however it is worth noticing that an average of *8%* of the features identify two genes. Considering that a typical tissue expresses *30-40%* of its genetic repertoire, at a level detectable by microarray (Su et al., 2002; Ramsköld et al., 2009) there is a *75%* chance that a signal coming from one of these features is the result of the expression of a single gene. If such features turned up as

Table 1. Pair-wise comparison of the CDF probe - gene mapping

Query	Reference	Query only	Ref. only	Intersection	Union
Affymetrix	Ebased	36,705	2,367	154,926	193,998
Affymetrix	GA	35,850	2,536	155,781	194,167
Affymetrix	MBNI	52,777	2,143	138,854	193,774
MBNI	GA	6,953	24,273	134,044	165,270
MBNI	Ebased	7,210	23,506	133,787	164,503
GA	Ebased	15,737	14,713	142,580	173,030

potential candidates in a study, they can be validated by RQ-PCR. Moreover, analyzing the features for the hgu95av2 chip in more detail, shows that about *70%* of them map to the same gene at different level (*e.g.* one is mapped at the "transcript", the other one at the "gene" level), which increases the probability that such features monitor only a single gene. For the remaining *30%*, the vast majority consists of "transcript" pairs. Among these pairs, *29%* identify genes of the same family, as for example the probe-set "ENSG00000047634_transcript_dubious_multiple_transcript_at" that maps the *SCML1* and *SCML2* genes (*Sex comb on midleg-like protein 1 and 2*, member of the SCM family that holds the polycomb group (PcG) proteins). The rest mainly consists of pairs of a known gene (having a HUGO symbol) with a gene the function of which is most certainly not known (*i.e.* solely annotated with an Open Reading Frame (*ORF*) description or a GENBANK accession number). As a consequence, allowing a probe to be part of several probe-sets extends reliably the information that can be recovered from an Affymetrix GeneChip®.

4.2 Generating probe-sets at gene empty loci

Depending on the GeneChip® version, *15-30%* of the probes on the array cannot be mapped to genes (Figure 1). These probes are clustered together, provided that the probe's inter-distance is smaller or equal to one kb. This offers the possibility to identify loci, which are translated, without being described as such, *e.g.* unknown ORFs. As shown in the following (see paragraph 4.8), three of these loci are significantly differentially expressed in acute lymphoblastic leukemia (ALL), between the *BCR/ABL* gene-fusion and the NEG genotypes.

4.3 Comparison with existing CDFs

To validate and benchmark our Ebased CDF, we compared it to the other aforementioned CDFs. First, to assess if the different CDFs are comparable, the probes to probe-sets assignment was investigated between all possible combinations of CDFs. The CDF pair-wise comparison of their probe-gene mapping for the Affymetrix GeneChip® hgu95av2 (*201,800* probes) is shown in Table 1. To ensure fair comparisons, the Affymetrix and Ebased CDFs were subset to keep their probe-sets mapping to single gene only, in order to match the GA and MBNI CDFs' content. Indeed, the GA and MBNI CDFs use less probes; probes that are for *76%* rescued by the Ebased CDF through multi-gene probe-sets (Supplementary Table 1). For the probe-sets mapping to single gene only, between *19* and *27%* of the Affymetrix probes are rejected by the custom CDFs. These all have *1%* of their probes being assigned to genes not represented in the Affymetrix CDF. These differences are due to the custom probe-set reconstruction. The custom CDFs are comparable as they overlap pair-wise by about *81%*; their differences most probably being due to the different alignment procedures and to the different annotation databases used. Second, as the CDFs showed a sufficient agreement, we compared the raw intensity variability of probe-sets among the ALL dataset samples, in order to measure the effect of using different CDFs on biological data. For every informative probe-set (see paragraph 2.8), the member probes are retrieved and the standard deviation calculated. Their distribution is shown in Supplementary Figure S1. The Ebased CDF shows an overall lower variability; the p-value of a One Way Analysis of Variance (ANOVA) being highly significant ($Pr(>F)$ = *8.47e-35*). Furthermore, the variability for the Ebased CDF probe-sets that are mapping no gene is close to the one of those mapping one gene only, as shown in Supplementary Figure S2. Overall, this demonstrates that the Ebased CDF probes aggregation results in probe-sets that provide more robust measurement. Ultimately, this result is not an artifact of the number of probes assigned per probe-sets, *i.e.* the Ebased CDF probe-sets have a similar number of probes as those of the other CDFs for the filtered (Figure 2a) and the identical probe-sets (Figure 2b). In addition, as shown is

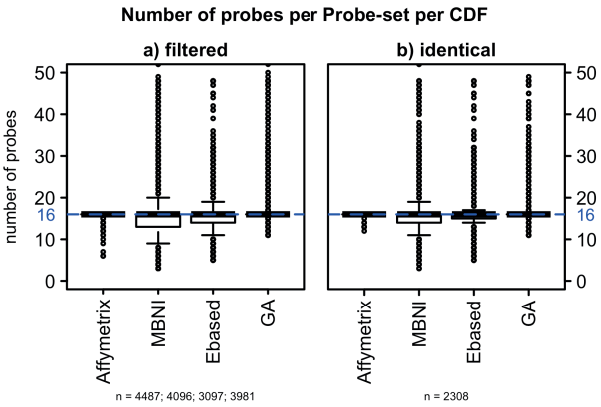


Fig. 2. Number of probes per probe-set per CDF

Supplementary Figure S3, the distribution of probes per probe-set for the Ebased CDF is not affected by probe-sets mapping one or two genes and confirm that the lower variability observed for the Ebased CDF probe-sets is not an artifact of the probes' aggregation process. The different performance of the various custom CDFs is most probably the outcome of the different gene-build process being used: the MBNI being based on EntrezGene (Maglott et al., 2005), GA being based on GeneCards (Rebhan et al., 1997) and the Ebased CDF using the data generated by the Ensembl gene-build process (Curwen et al., 2004). These processes are run regularly to keep genome annotations up to date. Estimating the effect of these updates on custom CDFs is important to assess how often they need to be updated.

4.4 Comparison of different Ebased CDF versions

All custom CDFs have a similar probe to probe-set assignment, however, they differ much by their update frequency: from every other month for the Ebased CDF, to up to one year for the MBNI. To assess whether the frequent human genome updates have a significant effect on the CDFs and their annotations, we analyzed the sequential changes occurring for the hgu95av2 Ebased packages, generated using Ensembl version *50, 51* and *52* (the packages version *1.0.3, 1.0.4* and *1.0.5*, respectively). The probe-sets' annotation variation is summarized in Table 2. The update

Table 2. Evolution of the Ebased CDF

From	To	Identical	Re-annotated	New	Discarded
v.1.0.3	v.1.0.4	95.80%	1.20%	3%	2%
v.1.0.4	v.1.0.5	80.19%	0.81%	19%	1%

performed by Ensembl from version *50* to version *51* did not include a new Gene Build in contrary to the update from version *51* to version *52*. An update encompassing a new Gene Build results in four times more changes (a fifth of all genes are affected) than a "maintenance" update does. However, as can be seen in Supplementary Figure S4 and S5, the variations between versions affect on average one gene or one probe per probe-set, suggesting that the human genome assembly is becoming increasingly precise. Interestingly, as can be seen in Supplementary Table S2 and S3, the number of probes that can be mapped to the genome increases with the Ensembl versions for 3 different Affymetrix Gene Chips (95av2, 133a and 133plus2), due to recent large-scale projects. In particular, RNA-Seq experiments have shown that many UTR regions are much longer than expected (Mortazavi et al., 2008). Extending them increases the chance to map Affymetrix GeneChip® probes, designed to map the last exon and the 3' end of transcripts. It is, therefore, a clear advantage for a CDF to be updated as soon as new annotations are available, to benefit from these refinements.

4.5 Gene Set residuals compared between CDFs

To test the impact of different CDFs on downstream analysis, we first perform a Gene Set Enrichment Analysis (*GSEA*) using linear models, as described in Oron et al. (2008), applied to the same data set they used: *79* acute lymphoblastic leukemia samples having either a *BCR/ABL* translocation (*BCR/ABL*) or not (*NEG*). Gene-sets (*GS*) are created per chromosome band, their residuals calculated per sample and summarized. This is performed iteratively for every CDF and the results compared per genotype, as shown in Figure 3 and Supplementary Figure S6, for the NEG and *BCR/ABL* genotypes, respectively. As shown in Figure 3a, clustering the GS residuals results in two groups: Affymetrix & Ebased and GA & MBNI. This separation and the similarity between the results obtained using the Ebased and Affymetrix CDFs, are due to the Ebased CDF keeping all possible information; *i.e.* probe-sets mapping several genes that are discarded by the other custom CDFs. Nevertheless, as shown in b), where the samples are sorted by the median of their GS residuals' distribution and linked through by red (changed order) or green (identical position) lines, all CDFs present similar rankings (ANOVA $Pr(>F) = 0.9948$). The same is observed

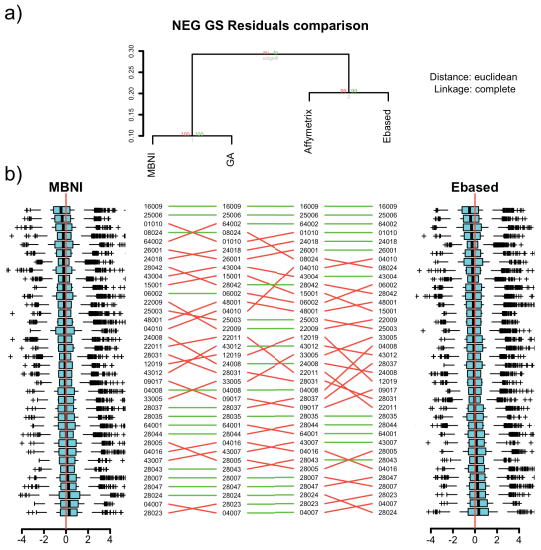


Fig. 3. Relationship of the different CDF Gene Set Residuals

in Supplementary Figure S6 for the *BCR/ABL* genotype; all CDFs perform comparably.

4.6 Chromosome Y GS residuals

When used for a GSEA analysis, every expression matrix generated by the different CDFs performs similarly. In Oron et al. (2008) the result of this analysis is used for a QC validation: since sex is associated with chromosome-level expression differences, they analyzed the ALL samples and identified those that were wrongly annotated; *i.e.* a male sample identified as female and vice-versa. After complementary verification, they showed that two male and one female samples were wrongly annotated. Here, we apply the same analysis to compare if any of the custom CDFs can render this distinction clearer; *i.e.* the obtained distributions should be less scattered and have fewer outliers. The chromosome Y residuals for the different CDFs are extracted and classified using the original sample annotations: male samples should show positive residuals and female samples negative ones. Due to the afore-mentioned wrong annotation, this is not the case, as shown in Figure 4. The MBNI and the Ebased CDFs show the expected results: *3* outliers corresponding to the *3* misannotated samples, whereas the Affymetrix and GA display more outliers. To find out which of the CDFs separates the male from female samples best, we perform an unsupervised classification using an expectation - maximization (EM) iterative method on the chromosome Y GS residuals. The Ebased CDF results in the most confident classification, as shown in Supplementary Figure S7, with all uncertainty scores lower than *0.005*. This indicates that the Ebased CDF is the least affected by the technical measurement artifacts inherent to the micro-array technology and therefore the less prone to identify false positives.

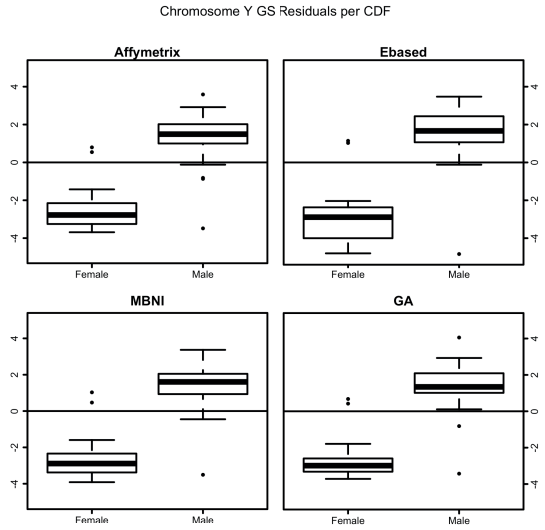


Fig. 4. Distribution of the chromosome Y Gene Set Residuals per sex and CDF

4.7 Differential Expression between the NEG and *BCR/ABL* genotypes

The results obtained from the expression matrices generated using the Ebased CDF are more accurate than those obtained using the other CDFs. But does the Ebased probes rescue deliver additional valuable information, *i.e.* do the multi-gene and intergenic probe-sets reveal otherwise ignored loci? To investigate this, we performed a differential gene expression analysis using linear models between the NEG and the *BCR/ABL* samples for every expression matrix. The differentially expressed genes identified were then compared pair-wise between CDFs. Across all CDFs, *41* genes are differentially expressed (Figure 5). *35* are detected by using the original Affymetrix CDF, but if updated gene annotations (custom CDFs) are used, between *4-12* genes of this gene set are considered as false positives, originating from a wrong probe to probe-set assignment (Supplementary Table S4). All the genes identified using the MBNI (n = 23) or GA (n = 25) CDF are present in the Affymetrix CDF related results. Applying the Ebased CDF recalls *90%* of the genes found by using the MBNI or GA, with the exception of *2* (*paternally expressed 10, LYN*) and *3* (*paternally expressed 10, LYN, ZEB1*) genes, respectively, for a total of *31* genes identified. The magnitude of the difference between the custom CDFs is similar to the one observed for the probe to probe-set assignment described in Table 1 and suggest that it could be the consequence of the different gene/transcript build being used and of translation errors between databases' cross-references, a recurrent issue in gene annotation, as described previously (Drăghici et al., 2006). This hypothesis is supported by the example of the gene *paternally expressed 10* identified by the MBNI and GA CDFs, which is actually found as well by the Ebased CDF through its HUGO name *PEG10* (Paternally expressed gene 10). As for the other genes (*LYN, ZEB1*), both the MBNI and GA identify the *LYN* gene that encodes a protein kinase, as being *60%* more expressed in *BCR/ABL* samples than in NEG samples. When using the Ebased CDF, the corresponding probe-set does not present any significant variation and is filtered out as being invariant. The log2 ratios determined by the three custom CDFs are very similar and the fold change is close to the microarray resolution limit, under which the statistical confidence is impeded. Without using a different technology (*e.g.* qRT-PCR, RNA-Seq) it is impossible to tell, which probe-set is the most accurate and whether that gene is differentially expressed or not.

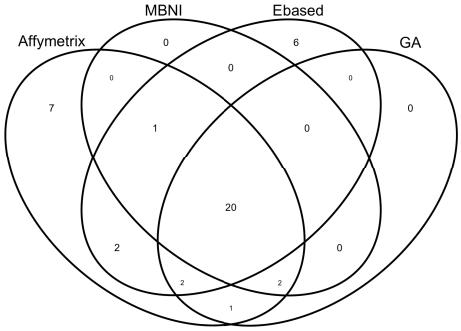


Fig. 5. Venn diagram of the differentially expressed genes identified by the different CDFs

4.8 Genes identified specifically by the Ebased CDF

Five probe-sets are uniquely identified when using the Ebased CDF, among which, only one has probes that map multiple loci (Table 3). Two probe-sets map annotated genes (*FSCN1* and *CD99*) and

Table 3. Genes identified by the Ebased CDF only

Probe-sets	Chr	Start	End	Strand	Gene
1	5	82,912,772	82,912,987	+	*VCAN**
2	10	31,857,709	31,858,087	+	*ZEB1**
3	15	32,576,159	32,576,542	+	*FSCN1**
4	7	5,598,980	5,621,811	+	*FSCN1*
5	X	2,619,228	2,669,348	+	*CD99*

1: 5_82912772_82912987_plus_genomic_at
2: 10_31857509_31858087_plus_genomic_at
3: 15_32576159_32576542_plus_genomic_multiple_at
4: ENSG00000075618_transcript_at
5: ENST00000381192_transcript_at
* probe-sets within 1kb downstream of the gene 3' UTR

three are located in intergenic regions, within 1kb downstream of the 3' UTR of genes reported by Ensembl (*FSCN1, VCAN* and *ZEB1*; marked by an asterisk in Table 3). The probe-set mapping multiple loci (third row in Table 3) is originally annotated as mapping an intergenic location on chromosome 15 and hence is a possible artifact. However, all the probes (n = 8) of that probe-set map solely to another locus on chromosome 7, located next to the 3'UTR region of the already identified *FSCN1* gene. This reinforces the evidence that the *FSCN1* gene is differentially expressed and displays the potential of using such probe-sets. It is interesting to note that the *ZEB1* gene is the third gene specifically identified when using the GA CDF. Therefore, with the exception of the *LYN* gene, the Ebased CDF recalls all the genes identified by the two other custom CDFs, as well as three additional ones: *VCAN, FSCN1* and *CD99*. These genes are *50%* more expressed in the *BCR/ABL* genotype than in the NEG one. *VCAN* (versican) is involved in cell adhesion, migration, and proliferation and an increase of expression is often observed in various tumors' growth. It has been shown to strongly enhance LLC (Lewis lung carcinoma) metastatic growth (Kim et al., 2009). *CD99*, coding for the CD99 antigen, is involved in T-cell adhesion processes. It is involved in spontaneous rosette formation with erythrocytes, a process reported to occur in Burkitt Lymphoma biopsies (Gross et al., 1975). Finally, *FSCN1* is coding for the Fascin protein, which organizes filamentous actin into actin/fascin bundles. It has been shown (Minn et al., 2005) to mark and mediate breast cancer metastasis to the lungs and is patented as a specific marker for pancreatic cancer (Patent number WO2004055519-A2). In conclusion, the five additional probe-sets identified by the Ebased CDF identify genes or transcription products the deregulation of which could have oncogenic consequences.

5 CONCLUSION

Even as the development of new technologies to measure RNA expression, especially Next-Generation Sequencing, is expending, many studies are still array-based and this trend will hold until these new technologies have matured and become cost-competitive for every lab. In addition, the last decade has generated a huge amount of array-based expression profiling data that, until recently (Lukk et al., 2010), have often been only superficially analyzed. To analyze these data, whether published and available from the GEO and ArrayExpress resources, or new requires up-to-date probes' annotation and in the case of Affymetrix microarray a CDF describing the correct probes to probe-sets mapping. In comparison to the other CDFs, our Ebased CDFs offers a higher sensitivity, is more frequently updated and uses as many probes as possible to benefit from all the possible information present on an array. In addition, it contains additional probe quality information, such as the number of genes mapped by a given probe-sets. Applied on the frequently analyzed ALL dataset, it unravels three new potential candidate genes, which implication in cancer has been shown in other tumors. Our CDF is an enhanced tool to perform Affymetrix microarray analyses on either new or published data and by this mean extend our biological knowledge.

AUTHORS' CONTRIBUTIONS

ND implemented the CDF generation, the `customCDF` R package and performed the analyses. FB established the processing pipeline. NB was involved in specific implementation and analysis issues. MH designed, populated and is maintaining the website. GT and PL helped with the project design and planning and coordinated the work. ND, GT and PL wrote the manuscript. All authors read and approved the final version of the manuscript.

ACKNOWLEDGEMENT

The authors want to thank Felix Engel for constructive discussions during this project, Verena Fleig, Charles Girardot and Robert Weatheritt for excellent comments on the manuscript and Marc Zapatka for technical support. This work was supported by grants from the Bundesministerium fr Bildung und Wissenschaft within the National Genome Research Network (NGFN2: 01GS0460, 01GR0417 and 01GR0418) and the Medical Genome Research Program (NGFN-plus: Brain Tumor Network plus; 01GS0883).

REFERENCES

Jeffrey A Bailey and Evan E Eichler. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet*, 7(7):552–64, Jul 2006.

Tanya Barrett et al. Ncbi geo: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Research*, 35 (Database issue):D760–5, Jan 2007.

Yoav Benjamini and Yosef Hochberg. 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, Jan 1995.

B M Bolstad, R A Irizarry, M Astrand, and T P Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–93, Jan 2003.

A Brazma et al. Minimum information about a microarray experiment (miame)-toward standards for microarray data. *Nat Genet*, 29(4):365–71, Dec 2001.

Raymond Cheong, Alexander Hoffmann, and Andre Levchenko. Understanding nf-kappab signaling via mathematical modeling. *Mol Syst Biol*, 4:192, Jan 2008.

Sabina Chiaretti et al. Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, 103(7):2771–8, Apr 2004.

Sabina Chiaretti et al. Gene expression profiles of b-lineage adult acute lymphocytic leukemia reveal genetic patterns that identify lineage derivation and distinct mechanisms of transformation. *Clin Cancer Res*, 11(20):7209–19, Oct 2005.

Val Curwen, Eduardo Eyras, T Daniel Andrews, Laura Clarke, Emmanuel Mongin, Steven M J Searle, and Michele Clamp. The ensembl automatic gene annotation system. *Genome Res*, 14(5): 942–50, May 2004.

Manhong Dai, Pinglang Wang, Andrew D Boyd, Georgi Kostov, Brian Athey, Edward G Jones, William E Bunney, Richard M Myers, Terry P Speed, Huda Akil, Stanley J Watson, and Fan Meng. Evolving gene/transcript definitions significantly alter the

interpretation of genechip data. *Nucleic Acids Research*, 33(20): e175, Jan 2005.

Eric H Davidson. Genomic regulatory systems: development and evolution. *Academic Press*, Jan 2001.

Sorin Drăghici, Sivakumar Sellamuthu, and Purvesh Khatri. Babel's tower revisited: a universal resource for cross-referencing across annotation databases. *Bioinformatics*, 22(23):2934–9, Dec 2006.

Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber. Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–40, Aug 2005.

Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–10, Jan 2002.

Francesco Ferrari, Stefania Bortoluzzi, Alessandro Coppe, Alexandra Sirota, Marilyn Safran, Michael Shmoish, Sergio Ferrari, Doron Lancet, Gian Antonio Danieli, and Silvio Bicciato. Novel definition files for human genechips based on geneannot. *BMC Bioinformatics*, 8:446, Jan 2007.

Paul Flicek et al. Ensembl 2011. *Nucleic Acids Research*, 39 (Database issue):D800–6, Jan 2011.

Laurent Gautier, Morten Møller, Lennart Friis-Hansen, and Steen Knudsen. Alternative mapping of probes to genes for affymetrix chips. *BMC Bioinformatics*, 5:111, Aug 2004.

Robert C Gentleman et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology 2010 11:202*, 5(10):R80, Jan 2004.

R L Gross, C M Steel, A G Levin, S Singh, and G Brubaker. In vitro immunological studies on east african cancer patients. iii. spontaneous rosette formation by cells from burkitt lymphoma biopsies. *Int J Cancer*, 15(1):139–43, Jan 1975.

Zhili He, Liyou Wu, Xingyuan Li, Matthew W Fields, and Jizhong Zhou. Empirical establishment of oligonucleotide probe design criteria. *Appl Environ Microbiol*, 71(7):3753–60, Jul 2005.

Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4 (2):249–64, Apr 2003.

Zhen Jiang and Robert Gentleman. Extensions to gene set enrichment. *Bioinformatics*, 23(3):306–13, Feb 2007.

Audrey Kauffmann, Robert Gentleman, and Wolfgang Huber. arrayqualitymetrics–a bioconductor package for quality assessment of microarray data. *Bioinformatics*, 25(3):415–6, Feb 2009.

Sunhwa Kim et al. Carcinoma-produced factors activate myeloid cells through tlr2 to stimulate metastasis. *Nature*, 457(7225): 102–6, Jan 2009.

Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biology 2010 11:202*, 10(3):R25, Jan 2009.

Jun Lu, Joseph C Lee, Marc L Salit, and Margaret C Cam. Transcript-based redefinition of grouped oligonucleotide probe sets using aceview: high-resolution annotation for microarrays. *BMC Bioinformatics*, 8:108, Jan 2007.

Xuesong Lu and Xuegong Zhang. The effect of genechip gene definitions on the microarray study of cancers. *Bioessays*, 28 (7):739–46, Jul 2006.

Margus Lukk, Misha Kapushesky, Janne Nikkilä, Helen Parkinson, Angela Goncalves, Wolfgang Huber, Esko Ukkonen, and Alvis Brazma. A global map of human gene expression. *Nature Biotechnology*, 28(4):322–4, Apr 2010. Important to show the use of integrating data.

Donna Maglott, Jim Ostell, Kim D Pruitt, and Tatiana Tatusova. Entrez gene: gene-centered information at ncbi. *Nucleic Acids Research*, 33(Database issue):D54–8, Jan 2005.

Tomas Marques-Bonet, Santhosh Girirajan, and Evan E Eichler. The origins and impact of primate segmental duplications. *Trends Genet*, 25(10):443–54, Oct 2009.

Andy J Minn et al. Genes that mediate breast cancer metastasis to lung. *Nature*, 436(7050):518–24, Jul 2005.

Ali Mortazavi, Brian A Williams, Kenneth Mccue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature Methods*, 5(7): 621–8, Jul 2008.

A. P Oron, Z Jiang, and R Gentleman. Gene set enrichment analysis using linear models and diagnostics. *Bioinformatics*, 24(22): 2586–2591, Sep 2008.

Helen Parkinson et al. Arrayexpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Research*, 37(Database issue):D868–72, Jan 2009.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Daniel Ramsköld, Eric T Wang, Christopher B Burge, and Rickard Sandberg. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol*, 5(12):e1000598, Dec 2009.

M Rebhan, V Chalifa-Caspi, J Prilusky, and D Lancet. Genecards: integrating information about genes, proteins and diseases. *Trends Genet*, 13(4):163, Apr 1997.

Rickard Sandberg and Ola Larsson. Improved precision and accuracy for microarrays using updated probe set definitions. *BMC Bioinformatics*, 8:48, Jan 2007.

Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article3, Jan 2004. doi: 10.2202/1544-6115.1027.

Maria A Stalteri and Andrew P Harrison. Interpretation of multiple probe sets mapping to the same gene in affymetrix genechips. *BMC Bioinformatics*, 8:13, Jan 2007.

Andrew I Su et al. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci USA*, 99(7):4465–70, Apr 2002.

Ryota Suzuki and Hidetoshi Shimodaira. Pvclust: an r package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12):1540–2, Jun 2006.

Zhenyu Xu, Wu Wei, Julien Gagneur, Fabiana Perocchi, Sandra Clauder-Münster, Jurgi Camblong, Elisa Guffanti, Françoise Stutz, Wolfgang Huber, and Lars M Steinmetz. Bidirectional promoters generate pervasive transcription in yeast. *Nature*, 457 (7232):1033–7, Feb 2009.

Appendix D

Analyses supplements

This appendix contains additional results related to the section 4.4, page 103 and section 4.5, page 121.

D.1 arrayCGH - EP pairs correlation

The Table D.1 lists all the arrayCGH - EP pairs associated with an FDR $\leq 10\%$.

gene.symbol	coef	p.value	fdr.local	chromosome
KIFC1	5.42	0.00	0.00	6
RCAN2	-4.88	0.00	0.00	6
GMNN	4.78	0.00	0.00	6
HLA-DRA	-4.43	0.00	0.00	6
E2F3	4.29	0.00	0.00	6
ELOVL2	4.18	0.00	0.00	6
HLA-DPA1	-4.01	0.00	0.00	6
MCM3	3.90	0.00	0.00	6
ATAT1	3.88	0.00	0.00	6
FOXF2	-3.87	0.00	0.00	6
HNRNPL	3.67	0.00	0.00	6
CDKAL1	3.58	0.00	0.00	6
PCSK2	3.55	0.00	0.00	6
DEK	3.41	0.00	0.00	6
NEDD9	-3.40	0.00	0.00	6
HIST1H4C	3.39	0.00	0.00	6
SOX4	3.27	0.00	0.00	6
DSP	-3.26	0.00	0.00	6
ID4	-3.25	0.00	0.00	6

continued on next page

continued from previous page

gene.symbol	coef	p.value	fdr.local	chromosome
BTN3A3	-3.18	0.00	0.00	6
DTL	3.15	0.01	0.00	1
HLA-C	-3.13	0.00	0.00	6
LSM2	3.09	0.00	0.00	6
TMEM151B	3.05	0.00	0.00	6
PIM1	3.04	0.00	0.00	6
HMGA1	3.03	0.00	0.00	6
ASPM	2.98	0.01	0.00	1
JARID2	2.96	0.00	0.00	6
FOXC1	-2.96	0.00	0.00	6
TFAP2A	-2.94	0.00	0.00	6
HLA-DPB1	-2.82	0.00	0.00	6
BTN3A2	-2.79	0.00	0.00	6
MAK	2.78	0.00	0.00	6
HLA-E	-2.77	0.00	0.00	6
F13A1	-2.77	0.00	0.00	6
CUTA	2.75	0.00	0.00	6
CYP39A1	-2.75	0.00	0.00	6
MSH5	2.73	0.00	0.00	6
MICA	-2.70	0.00	0.00	6
NEK2	2.69	0.01	0.00	1
NUP153	2.66	0.00	0.00	6
BRD2	2.65	0.00	0.00	6
ZNF193	2.64	0.00	0.00	6
SRPK1	2.64	0.00	0.00	6
ZSCAN16	2.62	0.00	0.00	6
FAM65B	-2.57	0.00	0.00	6
CAP2	-2.54	0.00	0.00	6
CCHCR1	2.48	0.00	0.00	6
SALL1	-2.47	0.00	0.00	16
CENPF	2.44	0.01	0.00	1
BAT2	2.42	0.00	0.00	6
PRR3	2.40	0.00	0.00	6
GMPR	-2.37	0.00	0.00	6
TUBB	2.37	0.00	0.00	6
IER3	-2.36	0.00	0.00	6
TRIM27	2.31	0.00	0.00	6
BMP5	-2.29	0.00	0.00	6
LYRM4	2.29	0.00	0.00	6
CLIC5	-2.28	0.00	0.00	6
BAT3	2.27	0.00	0.00	6

continued on next page

continued from previous page

gene.symbol	coef	p.value	fdr.local	chromosome
TUBB2B	2.26	0.00	0.00	6
RNF8	2.25	0.00	0.00	6
COL21A1	-2.24	0.00	0.00	6
SLC39A7	2.21	0.00	0.00	6
TFAP2B	-2.21	0.00	0.00	6
BTN3A1	-2.21	0.00	0.00	6
TAPBP	-2.20	0.00	0.00	6
NFYA	2.15	0.00	0.00	6
EFHC1	2.14	0.00	0.00	6
ICK	-2.12	0.00	0.00	6
MYLIP	-2.12	0.00	0.00	6
AIF1	-2.11	0.00	0.00	6
SNRPC	2.11	0.00	0.00	6
PIP5K1P1	2.10	0.00	0.00	6
BAT1	2.10	0.00	0.00	6
RXRG	2.09	0.02	0.00	1
GNL1	2.08	0.00	0.00	6
USP49	2.07	0.00	0.00	6
EEF1E1	2.07	0.00	0.00	6
PTMAP1	2.06	0.00	0.00	6
FKBPL	2.06	0.00	0.00	6
CSNK2B	2.05	0.00	0.00	6
HLA-A	-2.04	0.00	0.00	6
NUDT3	2.04	0.00	0.00	6
PSMB8	-2.02	0.00	0.01	6
STK38	2.02	0.00	0.00	6
ZSCAN12	2.01	0.00	0.00	6
OR2B6	1.99	0.00	0.00	6
RGL2	1.98	0.00	0.00	6
FTS	1.97	0.00	0.00	16
PTK7	1.96	0.00	0.00	6
RPP21	1.96	0.00	0.00	6
PBX2	1.92	0.00	0.00	6
PFDN6	1.91	0.00	0.00	6
RBL2	1.91	0.00	0.00	16
CDC5L	1.88	0.00	0.00	6
KLHDC3	1.88	0.00	0.00	6
KIF14	1.88	0.01	0.00	1
HIST1H2BD	1.88	0.00	0.00	6
NCR2	1.88	0.00	0.00	6
CD2AP	1.86	0.00	0.00	6

continued on next page

continued from previous page

gene.symbol	coef	p.value	fdr.local	chromosome
GTF2H4	1.85	0.00	0.00	6
CUL7	1.85	0.00	0.00	6
FTSJD2	1.84	0.00	0.00	6
GNG4	1.84	0.01	0.00	1
PRIM2	1.84	0.00	0.00	6
AGPAT1	1.83	0.00	0.00	6
IQCB2P	1.81	0.00	0.00	6
EHMT2	1.81	0.00	0.00	6
COL11A2	1.81	0.00	0.00	6
CUL9	1.80	0.00	0.00	6
RANBP9	1.79	0.00	0.00	6
PPP1R10	1.78	0.00	0.00	6
ITPR3	1.78	0.00	0.00	6
DDAH2	1.78	0.00	0.00	6
ZNF184	1.77	0.00	0.00	6
NUP50P2	1.75	0.00	0.00	6
RRM2	1.75	0.06	0.00	2
VARS	1.75	0.00	0.00	6
TPR	1.74	0.01	0.00	1
RDBP	1.74	0.00	0.00	6
SLC29A1	1.69	0.00	0.00	6
DLK2	1.68	0.00	0.00	6
PPT2	1.66	0.00	0.00	6
MDC1	1.65	0.00	0.00	6
ATF6B	1.65	0.00	0.00	6
LBR	1.60	0.01	0.00	1
C6orf106	1.58	0.00	0.00	6
ZNF318	1.57	0.00	0.00	6
MEA1	1.56	0.00	0.00	6
PHF1	1.55	0.00	0.00	6
U6 (RFAM)	1.54	0.01	0.00	1
PBX1	1.54	0.01	0.00	1
GRM4	1.54	0.00	0.00	6
ABCF1	1.54	0.00	0.01	6
TBCC	1.54	0.00	0.01	6
MRPS18A	1.53	0.00	0.01	6
PRL	1.53	0.00	0.01	6
FLOT1	1.52	0.00	0.01	6
POLR1C	1.49	0.00	0.02	6
HCG26	1.49	0.00	0.02	6
DOM3Z	1.47	0.00	0.02	6

continued on next page

continued from previous page

gene.symbol	coef	p.value	fdr.local	chromosome
PACSIN1	1.46	0.00	0.02	6
MRPL2	1.46	0.00	0.03	6
PRPF4B	1.44	0.00	0.03	6
GTPBP2	1.44	0.00	0.03	6
MAPK14	1.44	0.00	0.03	6
DST	1.42	0.00	0.03	6
STK19	1.36	0.00	0.05	6
VEGFA	1.36	0.00	0.05	6
SUCLA2P	1.34	0.00	0.05	6
ENAH	1.34	0.01	0.05	1
SOX11	1.33	0.05	0.05	2
ZNF76	1.32	0.00	0.06	6
ZNF451	1.32	0.00	0.06	6
HSD17B2	1.31	0.00	0.06	16
ZNF187	1.30	0.00	0.06	6
NCR3	1.30	0.00	0.06	6
C6orf130	1.27	0.00	0.07	6
C6orf62	1.26	0.00	0.07	6
MMP2	1.25	0.00	0.07	16
STK19P	1.21	0.00	0.08	6
PRIM1	1.20	0.01	0.09	1
RXRB	1.20	0.00	0.09	6
LRRC16A	1.20	0.00	0.09	6
RNF5	1.18	0.00	0.09	6
PIP5K1A	1.17	0.00	0.09	6
FTO	1.17	0.00	0.10	16
C1orf56	1.16	0.01	0.10	1
TBX1	1.16	0.00	0.10	6
FKBP5	-0.72	0.04	0.60	6

Table D.1 – Strongly associated arrayCGH - EP pairs, *i.e.* having an FDR $\leq 10\%$.

Appendix E

Author publications

This appendix lists the publications, which I was involved in, related to this thesis work.

1. In Kokocinski et al. (2005), I was involved in the software design and implementation of the **Flexible Annotation and Correlation Tool** (FACT)

2. In Stange et al. (2010), I was involved in the analyses, especially manually curating annotation using FACT.

3. In Barbus et al. (2011), I was involved in redefining the probes' mapping and annotation used for that study. The probes were provided by Operon (http://www.eurofinsdna.com/home.html) and spotted inhouse on microarray.

4. In Haag et al. (2012), I performed the same probe's curation as in Barbus et al. (2011).

The publication abstracts are listed below in the same order:

1. **FACT–a framework for the functional interpretation of high-throughput experiments.**
   Felix Kokocinski, ***Nicolas Delhomme****, Gunnar Wrobel, Lars Hummerich, Grischa Toedt, Peter Lichter.*
   FACT serves as a highly flexible framework for the explorative analysis of large genomic and proteomic result sets. The program can be used online; open source code and supplementary information are available at http://www.factweb.de. - BMC Bioinformatics (2005) vol. 6 pp. 161

2. **Expression of an ASCL2 related stem cell signature and IGF2 in colorectal cancer liver metastases with 11p15.5 gain.**
D E Stange, F Engel, T Longerich, B K Koo, M Koch, ***N Delhomme****, M Aigner, G Toedt, P Schirmacher, P Lichter, J Weitz, B Radlwimmer.*

Background and aims Liver metastases are the leading cause of death in colorectal cancer. To gain better insight into the biology of metastasis and possibly identify new therapeutic targets we systematically investigated liver-metastasis-specific molecular aberrations.

Methods Primary colorectal cancer (pCRC) and matched liver metastases (LMs) from the same patients were analysed by microarray-based comparative genomic hybridisation in 21 pairs and gene expression profiling in 18 pairs. Publicly available databases were used to confirm findings in independent datasets.

Results Chromosome aberration patterns and expression profiles of pCRC and matched LMs were strikingly similar. Unsupervised cluster analysis of genomic data showed that 20/21 pairs were more similar to each other than to any other analysed tumour. A median of only 11 aberrations per patient was found to be different between pCRC and LM, and expression of only 16 genes was overall changed upon metastasis. One region on chromosome band 11p15.5 showed a characteristic gain in LMs in 6/21 patients. This gain could be confirmed in an independent dataset of LMs (n=50). Localised within this region, the growth factor IGF2 (p=0.003) and the intestinal stem cell specific transcription factor ASCL2 (p=0.029) were found to be over-expressed in affected LM. Several ASCL2 target genes were upregulated in this subgroup of LM, including the intestinal stem cell marker OLFM4 (p=0.013). The correlation between ASCL2 expression and four known direct transcriptional targets (LGR5, EPHB3, ETS2 and SOX9) could be confirmed in an independent expression dataset (n=50).

Conclusions With unprecedented resolution a striking conservation of genomic alterations was demonstrated in liver metastases, suggesting that metastasis typically occurs after the pCRC has fully matured. In addition, we characterised a subset of liver metastases with an ASCL2-related stem-cell signature likely to affect metastatic behaviour of tumour cells. - Gut (2010) pp.

3. **Differential retinoic acid signaling in tumors of long- and short-term glioblastoma survivors.**
Sebastian Barbus, Bjrn Tews, Daniela Karra, Meinhard Hahn, Bernhard Radlwimmer, ***Nicolas Delhomme****, Christian Hartmann, Jrg Felsberg, Dietmar Krex, Gabriele Schackert, Ramon Martinez, Guido Reifenberger, Peter Lichter.*
Although the prognosis of most glioblastoma patients is poor, 3%-5% patients show long-term survival of 36 months or longer after diagnosis. To study the differences in activation of biochemical pathways, we performed mRNA and protein expression analyses of primary glioblastoma tissues from 11 long-term survivors (LTS; overall survival $\geq$ 36 months) and 12 short-term survivors (STS; overall survival $\leq$ 6 months). The mRNA expression ratio of the retinoic acid transporters fatty acid-binding protein 5 (FABP5) and cellular retinoic acid-binding protein 2 (CRABP2), which regulate the differential delivery of retinoic acid to either antioncogenic retinoic acid receptors or prooncogenic nuclear receptor peroxisome proliferator-activated receptor delta, was statistically significantly higher in the tumor tissues of STS than those of LTS (median ratio in STS tumors = 3.64, 10th-90th percentile = 1.43-4.54 vs median ratio in LTS tumors = 1.42, 10th-90th percentile = -0.98 to 2.59; P $<$ .001). High FABP5 protein expression in STS tumors was associated with highly proliferating tumor cells and activation of 3-phosphoinositide-dependent protein kinase-1 and v-akt murine thymoma viral oncogene homolog. The data suggest that retinoic acid signaling activates different targets in glioblastomas from LTS and STS. All statistical tests were two-sided. - J Natl Cancer Inst (2011) vol. 103 (7) pp. 598-606

4. **Nos2 inactivation promotes the development of medulloblastoma in Ptch1(+/-) mice by deregulation of Gap43-dependent granule cell precursor migration.**
Daniel Haag, Petra Zipper, Viola Westrich, Daniela Karra, Karin Pfleger, Grischa Toedt, Frederik Blond, ***Nicolas Delhomme****, Meinhard Hahn, Julia Reifenberger, Guido Reifenberger, Peter Lichter.*
Medulloblastoma is the most common malignant brain tumor in children. A subset of medulloblastoma originates from granule cell precursors (GCPs) of the developing cerebellum and demonstrates aberrant hedgehog signaling, typically due to inactivating mutations in the receptor PTCH1, a pathomechanism recapitulated in Ptch1(+/-) mice. As nitric oxide may regulate GCP proliferation and differentiation, we crossed Ptch1(+/-) mice with mice lacking inducible nitric oxide synthase (Nos2) to investigate a possible influence on tumorigenesis. We observed a two-fold higher medulloblastoma rate in Ptch1(+/-) Nos2(-/-) mice compared to Ptch1(+/-) Nos2(+/+) mice. To iden-

tify the molecular mechanisms underlying this finding, we performed gene expression profiling of medulloblastomas from both genotypes, as well as normal cerebellar tissue samples of different developmental stages and genotypes. Downregulation of hedgehog target genes was observed in postnatal cerebellum from Ptch1(+/+) Nos2(-/-) mice but not from Ptch1(+/-) Nos2(-/-) mice. The most consistent effect of Nos2 deficiency was downregulation of growth-associated protein 43 (Gap43). Functional studies in neuronal progenitor cells demonstrated nitric oxide dependence of Gap43 expression and impaired migration upon Gap43 knock-down. Both effects were confirmed in situ by immunofluorescence analyses on tissue sections of the developing cerebellum. Finally, the number of proliferating GCPs at the cerebellar periphery was decreased in Ptch1(+/+) Nos2(-/-) mice but increased in Ptch1(+/-) Nos2(-/) (-) mice relative to Ptch1(+/-) Nos2(+/+) mice. Taken together, these results indicate that Nos2 deficiency promotes medulloblastoma development in Ptch1(+/-) mice through retention of proliferating GCPs in the external granular layer due to reduced Gap43 expression. This study illustrates a new role of nitric oxide signaling in cerebellar development and demonstrates that the localization of pre-neoplastic cells during morphogenesis is crucial for their malignant progression. - PLoS Genet (2012) vol. 8 (3) pp. e1002572

Glossary

acute lymphoblastic leukemia a form of leukemia - cancer of the white blood cells - characterized by an excess of lymphoblasts. 3, 69, 138

analysis of variance a collection of statistical models for determining the effects of different source of variation on the variance of a variable. 144

aneuploidy a type of chromosomal abnormality showing a non normal number of chromosomes. 16

apoptosis programmed cell death, resulting in the destruction of the cell and its phagocytosis. 6–10, 18

arrayCGH an high throughput array based CGH, synonym of matrixCGH. 25, 29–31, 33, 47, 48, 51–57, 59, 60, 86, 89, 92, 95–97, 99, 103, 105, 109–111, 113–115, 117, 118, 120, 140–142, 144, 145, 150, 153, 154, 158, 184, 206, 210, 218

bacterial artificial chromosome a DNA construct based on a plasmid, used for transforming and cloning in bacteria. 44, 46, 56, 150

Bioconductor Bioconductor (Gentleman et al., 2004) is a software archive written for the R langague (R Development Core Team, 2009) and dedicated to the analyses of high throughput, high dimension datasets. 27, 28, 43, 45, 67

cancer of unknown primary cancer type which primary origin cannot be determined. 11

cancerogenesis the conversion of normal cells to neoplastic cells and their further development into a tumor. 3, 5, 6, 8, 10, 40, 80, 83, 86, 128, 137, 140, 147, 150–152, 156, 157, 159, 161–164, 172

chip a commonly used synonym for microarray. 24, 25

ChIP-on-chip a technique associating a chromatin immunoprecipitation with an hybridization on a microarray. 26

ChipYard a microarray analysis software, developed by Grischa Toedt, in the division of Molecular Genetics, DKFZ. 43, 44, 47, 48, 87

chromosomal instability an increased tendency to acquire chromosomal aberrations. 16

chromosome conformation capture technique used to analyze the organization of the chromosome in a cell. 26

chromothripsis Greek; chromo from chromosome; thripsis, for shattering into pieces; process by which ten to hundreds of chromosomal rearrangements occur in a single step cellular crisis. 14, 152

comparative genomic hybridization technique developed to compare the genomic content of 2 different samples. 22, 25, 215, 218

complementary deoxyribonucleotide acid a DNA strand complementary to its RNA template obtained by using a retro-transcriptase enzyme. 23, 220

contingency table In statistics, a matrix that displays the (multivariate) frequency distribution of the variables. Used to record and analyze the relation between two or more categorical variables. 56, 111

CpG Islands sequence rich in CG dinucleotides, where the cytosine is often methylated, involved in gene regulation. 26

cyclin dependent kinase cell cycle Serine/Threonine kinases, which activity depends on their association with cyclin. 7

deoxyribonucleotide acid (DNA) a long polymer of nucleotides that contains the genetic information of an organism. 2, 5, 14, 21, 23, 26, 39, 46, 216, 219, 220

division of Molecular Genetics, DKFZ Division of Molecular Genetics, at the German Cancer Research Center (DKFZ), lead by Prof. Dr. Peter Lichter. 41, 43, 67, 216

dysplasia enlargment of an organ or tissue by the proliferation of abnormal cells. 19

E-cadherin protein involved in the cells tight junctions, resulting in their anchorage to the ECM and to neighboring cells. 12

Ensembl A database hosted at the **European Bioinformatics Institute** (EBI) that contains gene information for every common model organisms. 43

epithelial mesenchymal transition process by which an epithelial cell transforms into a mesenchymal one, acquiring motility among other traits. 12, 13, 218

exon-exon junction adjacent position in a transcript sequence that originate from different exon, distantly located on the genome. 70

expressed sequence tag a sequence uniquely identifying a transcript, determined from a cDNA library obtained by a retro-transcriptase reaction. 71, 137, 138, 147, 159

expression profiling microarray application, which goal is to measure the amount of mRNA, a proxy of the expression of genes. 25–27, 29–33, 39, 41–43, 45, 48, 51, 52, 54–61, 67, 69, 86, 87, 95–97, 103, 105, 109, 111, 113–115, 117, 118, 120, 121, 125, 126, 128, 137–141, 143–145, 147, 148, 150, 152–155, 158, 160, 175, 177, 184, 206, 210

extracellular matrix extracellular mesh of secreted proteins surronding most tissue cells. 3, 13, 18, 78, 81, 130, 148, 162, 163, 216, 218, 221

extravasation process by which a cancer cell exit from a blood capillary. 12, 13

Gene Expression Omnibus a database maintained by the NCBI storing microarray data and metadata. 41, 42, 45, 51, 67, 75, 81, 82, 85, 107, 109, 121, 122, 150, 158, 175, 178, 180–183

gene ontology a representation of a gene within a directed acyclic graph consisting of three main branches: molecular function, cellular component and biological process. 29, 60, 117–120, 126–128, 130, 140, 156, 157, 161, 162

GeneChip an Affymetrix microarray technology (Lockhart et al., 1996). 24–27, 41–43, 45, 55, 60, 61, 67–73, 76, 106, 107, 137, 138, 147, 154, 155, 159, 171, 184

genomic regulatory network a network description of a set of genes, proteins, small molecules and their mutual regulatory interactions. 140

growth factor protein able to stimulate the growth or proliferation of the cell upon binding a specific cell surface receptor of that cell. 3, 10

hallmark of cancer a trait shared by every cancer, see Hanahan and Weinberg (2000). 3, 8–11, 13, 14, 32, 118, 157, 161, 162, 172

heterotypic refers to interaction between two or more different cell types, the contrary of homotypic. 9, 10, 13

hierarchical clustering a statistical method that uses the data intrinsic properties to identify groups. 48, 92, 94, 151

homeostatis the relatively stable equilibrium between interdependent cells maintained by physiological processes. 8, 218

homoscedastic synonym of homogeneity of variance. Characteristic of a distribution where all random variables have the same finite variance. 30, 75

hypoxia describes a state where cells are subjected to a lower than normal oxygen tension. 10

indel small insertion and deletion. 25

integrative analysis A kind of data analyses that integrates results from different methods as a mean to increase the discovery power by filtering confounding factors. 29–32, 39, 48, 54, 56, 60, 61, 92, 95, 102, 103, 108–113, 115, 117–123, 125, 132, 137, 139, 141–143, 145, 151–153, 155–161, 171, 172

integrin transmembrane protein exposed by a cell that binds the extracellular matrix and promotes cell quiescence and results in tissue . 3

intravasation process by which a cancer cell would enter the blood circulation. 12, 13

laboratory information management system a software to support modern laboratory technics, such as controlling a robot. 44

leukocoria abnormal white reflection from the retina of the eye. 16

loss of heterozygosity the result of a process where only one copy of an heterozygous allele is conserved. 7, 16, 18, 114, 156

macrometastasis thriving metastasis, which size is bigger than 0.2 mm in diameter. 13

matrix metallo-proteinase proteins able to degrade the ECM. 13

matrixCGH an high throughput array based CGH, synonym of array-CGH. 25–29, 39, 41–44, 46, 48, 67, 175, 177

mendelian relating to Mendel's theory of heredity. 5

mesenchymal epithelial transition the converse of the EMT. 12, 13

metastasis tumor forming at one site of the body from cell derived from a primary tumor located at another site of the body. 11, 13, 14, 18, 32, 42, 45, 61, 76, 78, 80–86, 132, 138, 148–150, 162, 163, 182, 218, 219

micro-environment the local environment of a tissue or tumor, suceptible to be influenced by that tumor or tissue. 11–14, 18, 149, 163, 164, 172

microarray a collection of feature spatially arranged in a grid. 21–29, 32, 33, 39, 42, 43, 46, 48, 50, 51, 53–56, 60, 67, 69, 82, 86, 87, 96, 103, 109, 110, 119, 132, 137, 139–144, 146, 147, 154, 155, 158, 171, 172, 211, 215–217

micrometastasis "dormant" metastasis, whose size is less than 0.2 mm in diameter. 13, 14, 149

mitogenic commonly a signal that provokes cell proliferation. 5, 6, 8, 29

mixture model a probabilistic model for describing the presence of sub-populations within an overall population. 50, 96

multidimensional scaling technique used in information visualization to explore (dis)similarities in data. 48, 92, 93, 151

N-cadherin protein involved in the motility of the cells. 12

necrosis process of cell death through different steps distinct from those of apoptosis. 10

neo-angiogenesis process by which novel blood vessels are formed. 11, 76, 85, 138, 148, 157, 161, 163

Next-Generation Sequencing The second generation of DNA sequencers, generating millions of short read (25-200bp) sequences. 17, 21, 27, 141, 146, 158

non mediated decay mechanism of mRNA surveillance, where wrongly spliced pre-mature RNA containing nonsense codon will be marked for degradation. This mechanism has been shown to be a mean of regulating certain transcripts expression.. 138

nucleolus nuclear structure largely devoted to manufacturing ribosomal subunits. 7

oncogene originally a gene that can transform cells, commonly a gene which has a tumorigenic potential. 2, 3, 6, 8

osteoblast cell type responsible for bone regeneration. 18

osteoclast cell type responsible for bone degradation. 18

osteosarcoma a malignant tumor of the bone, as well known as **osteogenic sarcoma**. 1, 13, 17–19, 32, 33, 39, 42, 60, 61, 67, 76, 78, 79, 121–126, 128–132, 137, 141, 143, 145, 146, 148, 150, 158–164, 171, 172

pathology the typical behavior of a disease. 18

pedigree a description of a family tree, with squares representing male and circles female individuals, often used in human genetics and medicine to visualize a disease penetrance. 5

polymerase chain reaction technique developed to amplify DNA fragments in an exponential manner using a polymerase enzyme. 24, 26

proto-oncogene cellular gene that altered through DNA damage acquires the capabilities of an oncogene. 2, 17

pseudohypopyon a purulent collection of fluid within the anterior chamber of the eye. 16

quality assessment process of evaluating the quality of a data set. 26, 27, 43, 44, 67, 69, 75, 87, 184, 220

quality control synonym of QA. 27, 43, 61, 67, 87, 177, 178, 180–185, 194

receiver operating characteristics a graphical representation of the performance of a binary classifier as its discrimination threshold is varied. It is created by plotting the TPR *vs.* the FPR at various threshold settings. 32, 59, 103, 104

receptor tyrosine kinase plasma membrane protein that possesses an extracellular ligand binding side and an intra-cellular kinase activated upon ligand binding. 3

reductionism a scientific research strategy that focuses on analyzing simple components of a complex system rather than the system as a whole. 9

retinoblastoma tumor of the oligopotential stem cells of the retina. 1, 3, 5, 8, 16–18, 32, 33, 39, 41, 42, 60, 61, 67, 76–78, 95, 107, 115, 120–126, 128–132, 137, 138, 141, 143, 146–150, 155–164, 171, 172, 220

retinoma spontaneous regression of a retinoblastoma or a benign manifestation of retinoblastoma. 17

retro-transcriptase an enzyme capable of converting an RNA template in a cDNA/RNA duplex. 23, 216, 217

RZPD Deutsches Ressourcezentrum fuer Genomforschung GmbH, Berlin, Germany; see http://www.rzpd-ia.de. 44

senescence a non-growing state of cell in which cells can remain viable for a long time but display specific phenotypic trait, including the incapacity to proliferate again. 9

stroma the mesemchymal components of epithelial and hematopoietic tissues, constituted of, among others, fibroblasts, endothelial cells, *etc.* and of the ECM. 10, 11

structural variant large insertion or deletion, inversion, tandem-repeat occurence,*etc.* . 25

Systems Biology scientific approach aiming at modeling biological systems as a whole rather than studying their single components independently. 15

telomere end of a chromosome arm, constituted of thousands of 6bp (in human) repeats. 9

transcription factor DNA binding protein responsible for the transcription of its target genes. 16, 26, 83, 153

tumor suppressor gene a gene whose partial or complete inactivation leads to an increased likelihood of cancer development. 3, 5, 6, 83

tumorigenesis the process of a tumor formation. 1, 14, 16, 18, 115

untranslated region flanking regions of a mRNA that are not translated into a protein. 137, 138, 147, 148, 155, 159

Acronyms

3C Chromosome Conformation Capture. 26

4C Circularized 3C. 26

5C Carbon-Copy 3C. 26

AKT/PKB protein kinase B. 9

ALL Acute Lymphoblastic Leukemia. 3, 69

ANOVA ANalysis Of VAriance. 144

AUC Area Under the Curve. 59, 105

AWS Adaptive Weight Smoothing. 100

BAC Bacterial Artificial Chromosome. 44

Bax *BCL2-associated X*. 8

BP Biological Process. 119

CC Cellular Component. 119

CDF Custom Definition File. 27, 43–45, 56, 69, 70, 72, 73, 75, 76, 78, 106, 118, 137–139, 146, 147, 153, 154, 171

CDK Cyclin Dependent Kinase. 7

CGH Comparative Genomic Hybridization. 22

CHEK2 **CHEckpoint Kinase 2**. 18

CI Confidence Interval. 98

CIN Chromosomal INstability. 16

CML Chronic Myelogenous Leukemia. 3

CNV Copy Number Variation. 26, 28–30, 39, 47, 55, 56, 58, 87, 88, 92, 109, 110, 115, 141, 143, 145, 146, 152, 154–156, 171, 195

CUP Cancer of Unknown Primary. 11

cytochrome C cytochrome C. 9

DE Differential Expression. 44, 45, 55, 80–82, 108, 143, 148

DKFZ German Cancer Research Center. 2, 41, 43, 184, 216

DNA deoxyribonucleotide acid. 2

***DOCK5* Dedicator Of CytoKinesis 5**. 18

EBI European Bioinformatics Institute. 216

ECM extracellular matrix. 3

EGF epidermal growth factor. 3

EMT Epithelial Mesenchymal Transition. 12

EP Expression Profiling. 25

FACT Flexible Annotation and Correlation Tool. 44, 211

FDR False Discovery Rate. 112, 113, 115, 157, 159, 206, 210

FPR False Positive Rate. 32, 59, 103, 104, 220

GEO Gene Expression Omnibus. 41

GO Gene Ontology. 29

GRN Genetic Regulatory Network. 140

HMM Hidden Markov Model. 28, 47

HPV Human papillomavirus. 2

ICD-10 International Classification of Diseases. 2

LIMS Laboratory Information Management System. 44

LINE Long INterspersed Elements. 150

LOA Limits of Agreement. 97, 98

log2 FC log2 Fold Change. 55, 58, 75, 76, 82, 83, 106–110, 141–143, 147, 149, 155

LOH loss of heterozygosity. 7

LTR Long Terminal Repeats. 150

MAD Median Absolute Deviation. 52, 56, 142

MAQC MicroArray Quality Control. 42, 67, 158

MAS Maskless Array Synthesizer. 24

Mdm2 mouse double minute 2. 7

MDS Multidimensional Scaling. 48, 151

meRNA multiexonic poly(A)$^+$ RNA. 14

MET Mesenchymal Epithelial Transition. 12

MF Molecular Function. 119

MMP Matrix Metallo-proteinases. 13

mRNA messenger RNA. 14, 23, 25, 28, 155, 217, 219, 221

NF-κB Nuclear Factor - Kappa B. 8, 80, 85, 148, 149

NGS Next-Generation Sequencing. 17, 21

NMD Nonsense-mediated Decay. 138

OMIM the **Online Mendelian Inheritance in Man** database. 16, 18

p14$^{\mathbf{ARF}}$ Alternative Reading Frame. 7

p15$^{\mathbf{INK4B}}$ p15$^{\mathrm{INK4B}}$. 7, 8

p16$^{\mathbf{INK4A}}$ p16$^{\mathrm{INK4A}}$. 7

PCR Polymerase Chain Reaction. 24

PDGF platelet-derived growth factor. 10

PI(3)K phosphatidylinositol-3OH kinase. 3, 8, 9

pRb retinoblastoma protein. 6, 8, 18

PTEN phosphatase and tensin homolog deleted on chromosome 10. 9

QA Quality Assessment. 26, 43, 67

RB1 retinoblastoma gene. 5, 6, 8, 9, 16–19

RNA ribonucleotide acid. 14, 114, 137, 154, 155, 158, 216, 220, 224, 225

ROC Receiver Operating Characteristic. 32

ROS Reactive Oxygen Species. 132

RSV Rous Sarcoma Virus. 2

RTK Receptor Tyrosine Kinase. 3

RUNX2 **Runt-related transcription factor 2**. 18

SD Standard Deviation. 46, 58, 98, 111, 142, 143, 145

SDF1 stroma-derived factor-1. 11

SINE Short INterspersed Elements. 150

SNP Single Nucleotide Polymorphism. 25

snRNA small nuclear RNA. 114, 155, 172

SYK **Spleen Tyrosine Kinase**. 17

TDI Total Deviation Index. 97, 98

TF transcription factor. 16

TGF-β Transforming Growth Factor Beta. 3, 8, 13

TNF-α Tumor Necrosis Factor alpha. 13

TNFRSF10A **Tumor Necrosis Factor Receptor SuperFamily member 10A**. 18

TNFRSF10D **Tumor Necrosis Factor Receptor SuperFamily member 10D**. 18

TP53 tumor protein p53. 5–9, 18, 19

TPR True Positive Rate. 32, 59, 104, 220

UHRR Universal Human Reference RNA. 42, 121–123, 143, 158, 183

uPA urokinase Plasminogen Activator. 13

UTR UnTranslated Region. 137

WLOG Without Loss Of Generality. 50