GIL TOMÁS

Gene Expression Markers of Proliferation and Differentiation in Cancer

The Extent of Prognostic Signals in the Cancer Transcriptome

UNIVERSITÉ LIBRE DE BRUXELLES

Gene Expression Markers of Proliferation and Differentiation in Cancer &
The Extent of Prognostic Signals in the Cancer Transciptome

Thèse présentée en vue de l'obtention du grade académique de Docteur en Sciences Biomédicales.

Le Jury est constitué de Pierre Bergmann, Président; Vincent Detours, Promoteur; François Fuks et Christos Sotiriou; et par Raphaël Marechal, secrétaire.

Gianluca Bontempi (Département Informatique, Faculté des Sciences, ULB) et Ann Nowé (VUB) siègent avec voix consultative au titre « d'expert extérieur ».

This dissertation was typeset with LaTeX, using the Tufte-LaTeX document class file.

PUBLISHED BY UNIVERSITÉ LIBRE DE BRUXELLES

First printing, June 2016

Gene Expression Markers of Proliferation and Differentiation in Cancer

The Extent of Prognostic Signals in the Cancer Transcriptome

in memory of
Stephen J. Gould
(1941–2002)

dedicated to
THE RAMONES

Tenho que escolher o que detesto—
ou o sonho, que a minha inteligência odeia,
ou a acção, que a minha sensibilidade repugna;
ou a acção, para que não nasci,
ou o sonho, para que ninguém nasceu.
Resulta que, como detesto ambos,
não escolho nenhum;
mas, como hei-de, em certa ocasião,
ou sonhar ou agir,
misturo uma coisa com a outra.

Fernando Pessoa
Livro do Desassossego

Résumé

Le cancer est un groupe de maladies génétiques opérationnellement défini par une prolifération cellulaire incontrôlée, impliquant une défaillance de l'homeostasie de l'organisme. La recherche sur le cancer vise à fournir des outils diagnostics précis et des traitements ajustés pour chacune de ces maladies. La technologie microarray permet la quantification de l'expression de tous les produits de transcription du génome humain et constitue donc un outil pour mieux comprendre la nature polygénique du cancer. La technologie microarray permet à la fois de découvrir de nouvelles classes de cancers et de prédire l'issue de maladie en fonction de profils d'expression préalables. En outre, l'utilisation de signatures d'expression géniques en tant que marqueurs représentatifs de certains processus physiologiques moléculaires permet l'emploi de données microarray pour tester des hypothèses biologiques.

Cette dissertation a deux objectifs: (*a*) établir la mesure dans laquelle des marqueurs d'expression génique de la différenciation et de la prolifération cellulaire peuvent contribuer à la classification des maladies cancéreuses; et (*b*) d'évaluer l'étendue des signaux pronostiques dans les transcriptomes cancéreux.

Nous avons mis au point une méthode objective pour extraire des signatures de différentiation organe-spécifiques à partir de données d'expression génique. Nous avons ensuite démontré qu'une signature génique de différentiation tissu-spécifique est capable de distinguer avec précision entre des sous-types histologiques de difficile classification dans un modèle thyroïdien. Ceci fait preuve du potentiel valeur clinique et diagnostique des signatures de différentiation dans le domaine oncologique.

Nous montrons aussi qu'une fraction non négligeable des transcriptomes cancéreux est capable de prédire l'issue des respectives maladies, à la suite d'une analyse systématique de 114 cohortes de profiles d'expression cancéreux englobant 19 types de cancers différents. Cet observation est probablement liée à une vaste structure de corrélation parmis les profils d'expression cancéreux, partiellement expliquée par des variables techniques et biologiques. Cette evidence met en cause l'utilisation généralisée d'associations statistiques entre des marqueurs d'expression géniques et les issues de chaque maladie parmis plusieurs patients afin d'en déduire l'implication de mécanismes biologiques particuliers dans la progression du cancer.

Summary

Cancer is a group of genetic diseases operationally defined by uncontrolled cellular proliferation, with consequent disruption of the organism homeostasis. Cancer research seeks to provide accurate diagnoses and tailored treatments to this broad spectrum of diseases. Microarray technology allows for the monitoring of the expression of all transcription products in the human genome, and thus presents with a tool to address the polygenic nature of cancer. Microarray technology can assist both to the task of cancer class discovery and cancer outcome prediction based on prior expression profiles. Furthermore, the use of gene expression signatures as surrogate markers for physiological molecular processes enables the possibility of hypothesis testing using microarray data.

The aim of this dissertation is two-fold: (*a*) to ascertain the extent by which gene expression markers of differentiation and proliferation may contribute to disease classification; and (*b*) to assess the extent of prognostic signals in cancer transcriptomes.

We have devised an unbiased method to derive organ-specific differentiation signatures from gene expression data. We then demonstrated that tissue-specific gene expression signatures of differentiation can accurately discriminate between challenging histopathological subtypes of cancer in a thyroid model, therefore showing potential clinical diagnostic value.

We also show that a non-negligible fraction of cancer transcriptomes is associated with disease outcome, on the count of the analysis of 114 cancer cohorts spanning 19 different cancer types. This is likely due to an extensive correlation structure in cancer expression profiles, partly explained by technical and biological variables. Such evidence disavows the widespread use of statistical association between expression markers and cancer patients outcome in order to infer implication of particular biological mechanisms in disease progression.

Acknowledgements

I am indebted to Vincent Detours for the opportunity to realize this dissertation in his research group, for his patience and his stern commitment to individual emancipation. Jacques Dumont was throughout this thesis a vibrant source of motivation and a pillar for intellectual exchange. I am also grateful to Carine Maenhaut and Pierre Roger for supervising the progression of this work at large.

This work could not have been without the wisdom and support of the many colleagues I had the pleasure to meet during these formative years. I am particularly beholden to the the warm companionship of Maxime Tarabichi, Danai Fimereli and David Gacquer. Soazig le Pennec, Tomasz Konopka, David Weiss Solís, Wilma van Staveren, Genevieve Dom, Soetkin Verstheye, Sébastien Floor, Aline Hebrant, Aline Antoniou, Jaime Pita, Raquel Ramos, Robert Opitz, Luca Tiberi, Benjamin Beck, Roxane van Heurck, Gabriele Zoppoli, David Brown, Germain Mazères, Fredérique Savagner and Christophe Trésallet are among the many I had the privilege to work and learn with during this doctorate. Many thanks as well to Danièle Leemans-De Vos, Genevieve Dalle and Joelle Sente for their invaluable assistance. I hope you won't mind if I call you my friends.

The years I spent working on this thesis will always be tied with my memories of Brussels. My dear friends Kenneth, Irina, Andrzej, Ivan, Florent, Ricardo, Seyne, Kate, Yanna, Iris, Mathilde, Chiara, David, Eric, Nereida, Elvis, Steven, Fillipo, and all the lost souls at the Lord Byron and at the Cobra are also part of this.

Portugal is another home of mine. In spite of the distance, the fingerprints of my amigos Luísa, Pedro, Xana, Catarina, Bruno, Cristiana, Jorge e Armando can be found all throughout this document.

I must not forget to mention Ana Sofia Rocha, without whom this thesis would not have begun nor ended, and Karen Haast, who held my hand throughout so much of it.

Finally, my deepest appreciation goes to the unwavering commitment and support of my family. To Celina, Gabriel, and very specially to my mom and dad—muito obrigado por tudo.

Financial Support

This work received the financial support of *l'agence Wallonie-Bruxelles International*, of the *Hoguet Foundation*, and of the *Fonds David & Alice van Buuren*.

Contents

List of Figures

List of Tables

Introduction

Cancer

Cancer is a disruption of the mechanisms evolved to promote the consolidation of the somatic lineage of multicellular organisms. This rupture is caused by a collection of critical failures of the genetic systems evolved to ensure the correct and timely integration of the cellular unit's physiology at the tissue and organism's level.[1] Neoplastic cells are operationally defined by their ability to sustain chronic proliferation, to invade tissues and to set up satellite growths in other organs. When left unchecked, these features can compromise the host organism's ability to uphold homeostatic balance,[2] and eventually lead to its systemic failure—and death.

[1] Maynard Smith and Szathmary, 1997

[2] While the concept of homeostasis emphasises the stability of the internal milieu toward perturbation, perhaps a more accurate formulation could be that of *homeodynamics*—a concept that seeks to account for the diverse behaviour exhibited by biological systems, including all its emergent characteristics, i.e., bistable switches, thresholds, waves, gradients, mutual entrainment, and periodic as well as chaotic behaviour (Lloyd et al., 2001).

The dynamics of cancer

Sixty years ago, Armitage and Doll developed a multistage theory to analyze rates of cancer progression.[3] They reasoned that cancer builds upon a sequence of cellular systems' cumulative failures. Each such failure, for instance the abrogation of a critical DNA repair pathway or the loss of control over cellular death, moves the system one step closer to the onset of disease.[4] Rather than a static physiological condition specified by a unique set of cellular dysfunctions, cancer is depicted as a progression along the course of a dynamic evolutionary history.

[3] Armitage and Doll, 1954

[4] Frank, 2007

Models of neoplasia evolution historically identify the seminal transforming event with an alteration on a single somatic cell that triggers cancer progression. Initiation events contributing to the early stages of neoplastic transition are caused by mutations in specific genes whose output is either enhanced (oncogenes) or repressed (tumour supressor genes).[5] Such genetic mutations can be structural, including nucleotide substitutions or mutations resulting from gene fusion,[6] juxtaposition to enhancer elements,[7] or by amplification. Alterations that imply a discrete change of output in gene expression, such as translocations or other structural mutations, can occur as initiating events[8] or during tumour progression, whereas amplification usually occurs during progression.[9]

[5] Alterations in nearly 500 of such genes have been linked to cancer initiation and progression (Forbes et al., 2008).

[6] Konopka et al., 1985

[7] Tsujimoto et al., 1985

[8] Finger et al., 1986

[9] Croce, 2008

However, a single genetic change is rarely sufficient to trigger the development of a neoplasia. Since the term *neoplasia* is generally used to refer to any new, abnormal growth of tissue,[10] the original oncogenic hit is usually associated with a mutation disrupting the balance between proliferation and cell death. From then on, cancer progression is modeled as a reiterative

[10] Concomitantly, *malignant* neoplasia is defined by the acquired capacity of neoplastic cells to invade locally and metastasize.

process of clonal expansion, with sequential subclonal selection.[11] The dynamics of this evolution are dictated by successive genetic and epigenetic[12] changes in the neoplasm, and are constrained by the ecological context in which the tumour is developing.

The prevailing mode of clonal evolution is through the gradual emergence of selectively advantageous "driver" genetic injuries against a complex background of mostly deleterious and selectively neutral "passenger" lesions. Alternatively, or perhaps concurrently, another mode of tumoural evolution contemplates the possibility of a few drastic events generating multiple lesions at once across the genome. These dramatic punctuated changes can be prompted by an acute insult or a single catastrophic pangenomic event—of which chromothripsis, at the chromosome level, is an example.[13,14]

The time frame of somatic evolution is a function of the mutational rate of neoplasms. While events like chromothripsis and kataegis[15] have the potential to provide nearly instant triggers for the onset of disease, the high frequency of clinically covert pre-maligant lesions[16] suggests that transformation of somatic cells is a far more frequent event than suggested by incidence curves. Furthermore, the fact that the majority of cancers only become clinically relevant at old age is a testament to both the prevalence of cancer-suppressing mechanisms as to the relatively slow rates of mutational accretion in neoplasms. Intriguingly, the rate of epigenetic change has been reported to be orders of magnitude higher than that of genetic change,[17] but its role in clonal evolution is not yet completely understood.

While the evolution of neoplasms is driven by their underlying genetic and epigenetic lability, it is their tissue ecosystems that provide the adaptive landscape for clonal fitness selection.[18] Systemic regulators, such as hormones, growth factors, immune and inflammatory cells as well as cytokines may conspire either to counteract or promote neoplastic growth.[19] Architectural constraints, in the form of physical compartments, basement membranes and confined metastatic niches, restrict the growth of tumoural masses and set a boundary for neoplastic microevolution. But perhaps most striking is the ability of tumours to remodel tissue micro-environments to their competitive advantage—illustrated by the capacity of transformed cells to promote neovascularization in response to anoxia or to incite malignant phenotypes in their adjacent stromal cells.[20]

[11] Nowell, 1976; and Greaves and Maley, 2012

[12] In this text, *epigenetics* will refer to the range of global modifications in gene expression that are not under control of the genetic code itself. The modifiable and reversible nature of certain cancer programs can largely be explained by epigenetics.

[13] Stephens et al., 2011

[14] The argument of gradualism versus punctuated equilibrium (Gould and Eldredge, 1993) is a longstanding debate in species evolution and is another example of how much our current conceptualization of cancer progression owes to the developments of the theory of evolution in the second half of the 20th century.

[15] *Kataegis*, a term derived from the ancient Greek word for "thunder", refers to a pattern of localized hypermutation identified in some cancer genomes (Nik-Zainal et al., 2012).

[16] Sakr et al., 1993

[17] Siegmund et al., 2009

[18] Greaves and Maley, 2012

[19] Bierie and Moses, 2006; and Hanahan and Weinberg, 2011

[20] Lathia et al., 2011

The linear model of cancer progression

At the phenotypic level, the course of neoplastic evolution is tagged along two major defining axis: one concerning increasing proliferation rates and another reporting the loss of morphological and physiological differentiation at the cellular level.[21]

In order to grow and become clinically conspicuous, neoplastic masses must shut down built-in genetic programs acquired during the establishment of multicellularity, to enforce compliance with societal rules. These include programs evolved to eradicate deviant phenotypes, such as apoptosis, senescence and necrosis and a range of cell-to-cell signaling programs designed to control and suppress unwarranted growth. Cancer cells must

[21] Tarabichi et al., 2013

also persistently evade the immune response of the host organism in its diverse ecological contexts.

For neoplasms can also invade surrounding tissues and disseminate in the organism, a feature responsible for most cancer-related deaths. This process is termed the invasion-metastasis cascade[22] and requires an array of well coordinated genetic and epigenetic adaptations. It can be schematized as a sequence of discrete steps that starts with the local invasion of the surrounding tissues. It then follows with the intravasion of cancer cells into nearby blood and lymphatic vessels—and with their extravasion into the parenchyma of distant tissues. Eventually, new micrometastatic lesions are established with the potential to develop into macroscopic tumours.

[22] Valastyan and Weinberg, 2011

To reconcile this increase in phenotypic resilience and plasticity of neoplastic cells with the canonical multistep model of cancer progression,[23] carcinogenesis is envisioned as a linear Darwinian evolutionary process.[24] According to this view,[25] inheritance, environmental factors and spontaneous errors in DNA replication cause mutations or epimutations in critical caretaker genes, entailing genetic and epigenetic instability. This, in turn, promotes mutations or epimutations in specific gatekeeper genes[26] (oncogenes or tumour supressor genes), triggering uncontrolled growth. The ensuing genomic instability creates a feedback loop that increases evolutionary and proliferation rates.[27] This results in heritable variation in the form of clonal diversity, upon which Darwinian selection operates. Clones that progressively acquire the biological hallmark capabilities of cancer[28] gain a competitive edge and become prevalent, promoting a form of neoplastic progression that, at the macroscopic level, is consistent with a multistep development.

[23] Land et al., 1983; and Vogelstein and Kinzler, 1993

[24] Merlo et al., 2006; and Polyak, 2014

[25] Podlaha et al., 2012

[26] Among cancer-susceptibility genes, the distinction between *caretaker* and *gatekeeper* genes is a subtle yet conceptually important one. While the former are responsible for maintaining the integrity of the genome, the latter directly control cellular proliferation. This is epitomized by the breast-cancer-susceptibility genes *BRCA1* and *BRCA2*, that can functionally play both roles at different points of breast cancer progression (Kinzler and Vogelstein, 1997).

[27] Sieber et al., 2003

[28] These include sustaining proliferative signaling; evading growth suppressors; resisting cell death; enabling replicative immortality; inducing angiogenesis; and activating invasion and metastasis (Hanahan and Weinberg, 2011).

This reasoning is supported by the observed stepwise progression of most solid tumours. For instance, the gradual evolution of colon cancer is well documented.[29] Here, the first manifestations of neoplasia occur in the colorectal epithelium, in the form of small benign adenomas. Such tumours are reasonably confined and are almost normal in their intra- and intercellular organisation. With time, adenomas start to grow (proliferation) and become morphologically and physiologically disorganized (dedifferentiation). Eventually the tumour evolves into an aggressive neoplasia (carcinoma), presumably because one of the cells in the adenoma has acquired a sufficient number of mutations to drive the process of invasion and metastasis.

[29] Vogelstein and Kinzler, 1993

The concept of gatekeeper gene is central to this understanding. Consider for instance the *TP53* gene, the first tumour-suppressor gene to be identified in 1979. This gene encodes for the p53 protein, a master transcription factor with an overarching role in the maintenance of the integrity of the genome.[30] Under normal circumstances, p53 is functionally inactive due to its rapid degradation. However, upon the infliction of virtually any form of cellular stress, p53 degradation is halted, and the protein gains full competence in transcriptional activation. The regulatory networks under its control are associated with several critical mechanisms for cancer progression, namely apoptosis, cell-cycle inhibition, genome stability and inhibition of angiogenesis.[31] Unsurprisingly, about 50% of all human cancers have lost p53 expression or express an inactive mutant of the protein.[32]

[30] Efeyan and Serrano, 2007

[31] Vogelstein et al., 2000

[32] Toledo and Wahl, 2006

The evolutionary origins of the *TP53* gene can be inferred from modern-

day descendants of both the single cell choanoflagellates and the early multicellular sea anemone. The function of the homolog to this ancestral gene in the sea anemone is to protect the germ-line gametes from DNA damage.[33] Over the course of the last billion years, this function has not only been conserved, but enhanced through pleiotropy[34] to set *TP53* as a major enforcer of somatic conformity in multicellular organisms.

[33] Belyi et al., 2010

[34] The ability of a single gene to influence multiple, seemingly unrelated phenotypic traits.

Another recurrent oncogene with an established role in cell-cycle progression and apoptosis, *MYC*, has had its evolutionary roots traced back to at least 600 million years ago.[35] The fact that many cancer-susceptibility genes are ancient, highly conserved and may have taken a role in the transition to multicellularity[36] has been interpreted as evidence that they play a pivotal role in regulating the normal physiology of the somatic cell.[37] Their disruption during carcinogenesis could then symbolize the unshackling of the transformed cell from its social bindings.

[35] Hartl et al., 2010

[36] Srivastava et al., 2010

[37] Weinberg, 1983; and Weinberg, 2013

Shortcomings of the linear model of cancer progression

This model describes the neoplastic cell as a unit set free to thrive through uncontrolled replication in a hostile environment. It becomes a metaphor for the unicellular evolutionary stage, with cancer lineages competing with each other and with normal cells for survival.[38] The success of any one lineage is dependent on its step-wise acquisition of cancer hallmarks through Darwinian evolution. However, this formulation does not fully account for a number of observations.[39]

[38] Merlo et al., 2006

[39] Davies and Lineweaver, 2011

First, it falls short to explain the high degree of cooperative organization among cancer cells. Increasingly, tumours are being recognized as ecosystems with complex and dynamic interactions between neoplastic cells and their microenvironment.[40] These heterotypic reciprocations include the stimulation via paracrine signaling of normal stromal cells to produce mitogenic signals; the required signaling to induce neoangiogenesis; and the diverse cell-to-cell interactions during the invasion-metastasis cascade.[41] Thus, in order for transformed cells to prosper, they must acquire a very specific minimal set of communication skills from early on. Rather than conceiving transformed cells evolving such intricate adaptations independently via adaptive mutations, a more prosaic explanation could involve a shift or modulation of the original set of rules somatic cells use to engage with their partners.

[40] Polyak et al., 2009

[41] Axelrod et al., 2006

Furthermore, it doesn't fully integrate the role of genetic instability in cancer progression. Because advantageous mutations are rare, transformed cells are thought to promote genomic instability in order to accelerate evolutionary rates and increase the odds of, via Darwinian selection, produce a fully malignant phenotype.[42] Yet, this mutational arms race among neoplastic cell lineages to reach complete neoplastic competence is riddled by a paradox: too few mutations in the mix and the cell won't escape genetic controls; too many and it dies. Especially troublesome are the pan-genomic mutations that lead to gross structural changes, including aberrant chromosomes and aneuploid cells.[43] It is thought-provoking to note that the neoplastic cells with the most deranged genomes are precisely those with the most competitive phenotypes. Jarle Breivik proposed an elegant solution

[42] Sieber et al., 2003

[43] "If you look at most solid tumours in adults, it looks like someone set off a bomb in the nucleus"—*William C. Hahn*

to this inconsistency.[44] Rather than postulating genomic instability as a prerequisite to cancer progression, he refashioned it as a by-product of the lack of competitive fitness of reparing-phenotypes in the tumour environment.[45]

A tentative proposition to address these inconsistencies has been put forward by Davies and Lineweaver.[46] Instead of modeling the transformed cell as a free-agent requiring the cumulative *acquisition* of enabling properties to obtain malignant status, it seeks to explain hallmark adaptations as the reenactment of "atavic" unicellular genetic systems that were functionally repressed during the transition to multicellularity. According to this view, the disruption of high-order gatekeeper genes drives the neoplastic cell to reconfigure its regulatory networks around a *pre-existing* toolkit of primitive adaptations reminiscent of those evolved during early transition to multicellularity. Although its original formulation has been either ignored[47] or thoroughly dismissed,[48] this idea could provide a basis for a reevaluation of some core features of cancer.

Consider the Warburg effect[49] for instance. The majority of cancer cells favour a metabolism based on anaerobic glycolysis followed by lactic acid fermentation in the cytoplasm. Because glycolysis is far less efficient than oxidative respiration for ATP production, this metabolic shift has been implied to be an essential adaptation to cancer progression—perhaps in response to intermittent hypoxia in malignant lesions.[50] A more economical interpretation could be that transformed cells, without gatekeeper genes to enforce the proper regulation of the high-performing metabolic mode of somatic cells, are now forced to fall back to a more basic, yet dependable, metabolic outlet for energy production. While the former interpretation taxes neoplastic progression with another requirement and begs for a rationale for the selective fitness of anaerobic metabolism, the latter simply conceives the cancer cell as a system seeking a dynamic equilibrium in a novel and unstable environment.

More recently, another key concept in cancer research has been the cancer stem cell. Tissue-specific stem cells have been identified at the top of the differentiation hierarchy of many organs. These stem cells are functionally defined by their long-term self-renewal capacity and their ability to differentiate into one or more tissue lineages. This hierarchical organization of tissue differentiation has been co-opted to explain tumour growth and heterogeneity, with cancer-specific stem cells responsible for the maintenance and growth of tumours. Cancer stem cells (CSCs), or tumour initiating cells, are operationally defined by their ability to re-form the parental tumour on transplantation into immunodeficient mice. They have been isolated from a range of solid tumours, such as breast cancer, brain tumours, colorectal cancer, and others.[51] Fittingly, this model complies with the requirements of clonal evolution, as tumour progression can be explained by derivatives of CSCs bearing different mutational signatures competing with each other. This interpretation suggests that all cancer cells in a tumour share a unique genetic and epigenetic history, in accordance with a linear progression. However, the reported coexistence of multiple genetic clones during acute lymphocytic leukemia progression suggests a more dynamic and modular clonal architecture instead.[52] This evidence, together with the fact that phenotypic conversion also occurs among non-hierarchically organized tumour

[44] Breivik, 2005

[45] "Don't stop for repairs in a war zone" is the metaphor used by Breivik to explain why a mutagenic environment would increase the fitness of the non-repairing phenotype.

[46] Davies and Lineweaver, 2011

[47] At the date of this writing, Google Scholar reported a total of 36 citations of the original article.

[48] Pettit, 2012; and Myers, 2012

[49] Warburg, 1956

[50] Gatenby and Gillies, 2004

[51] Beck and Blanpain, 2013

[52] Anderson et al., 2011; and Notta et al., 2011

cells in melanoma,[53] indicates that the CSC phenotype may be a transient response to the selective pressures of the tumour's milieu and of the stochastic events that govern its internal homeodynamics.[54] The CSC would then be, rather than a deterministic, a dynamically reversible phenotype; and represent, instead of qualitative, a quantitative modulation.[55]

Another example of the phenotypic modularity of the neoplastic cell is the epithelial-mesenchymal transition program (EMT)—and its reverse process, the mesenchymal-epithelial transition (MET). As with the CSC phenotype, this mechanism is borrowed from embryogenesis, where it takes part in gastrulation, neural crest formation, heart valve formation, palatogenesis and myogenesis.[56] This comprehensive genetic program globally shifts the physiology and morphology of the cell between an epithelial phenotype (characterized by a well defined polarity, tight junction of the cells and a their binding to a basal lamina) and a mesenchymal phenotype (characterized by a lack of polarity, spindle-shape morphology and loose cell-to-cell interaction).[57] Under normal physiological circumstances, the dynamics of the EMT-MET program are under strict and orderly genetic control, even if it can be re-enacted in a post tissue differentiation context—for instance in response to injury. In pathological conditions, the unwarranted activation of the EMT-MET program can cause organ fibrosis. But most importantly, it provides cancer cells with an outlet to break free from the primary tumour and embark the invasion-metastasis cascade under the cover of the mesenchymal phenotype. Conversely, the colonization of new micrometastatic niche is facilitated by a transition back to an epithelial phenotype. This constitutes the most thorough illustration of how neoplastic cells can recruit built-in genetic apparatuses to deploy complex adaptations.

[53] Quintana et al., 2010

[54] Visvader and Lindeman, 2012; and Aktipis et al., 2013

[55] Maenhaut et al., 2010; and Tarabichi et al., 2013

[56] Thiery et al., 2009

[57] Thiery and Sleeman, 2006

Cancer epidemiology

CANCER IS a leading cause of death worldwide, accounting for 8.2 million deaths in 2012. From the clinical point of view, cancer is a general designation for a group of more than 100 diseases. Lung, liver, stomach, colorectal and breast cancers are responsible for the majority of cancer deaths each year. The most frequent types of cancer, as well as their incidence, differ between women and men (Figure 1).

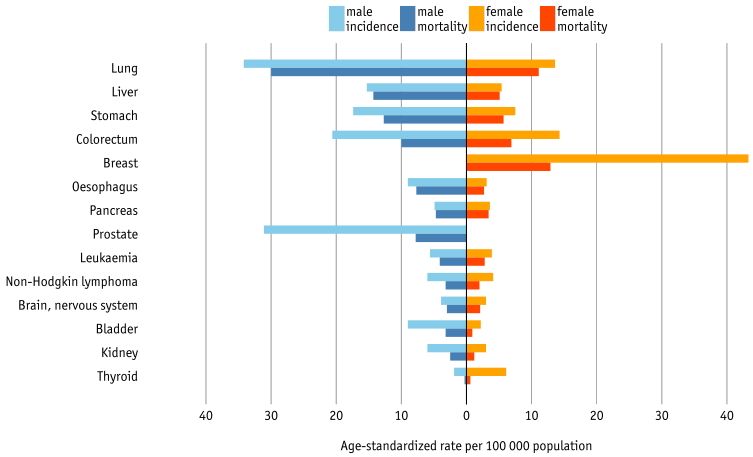

Figure 1: Global estimates of cancer incidence and mortality by sex. Age-standardized rate per 100 000 population (2012). Source: GLOBOCAN (Ferlay et al., 2014).

The list of factors involved in the causation of cancer is wide and diverse. Heritable genetic susceptibility, in the form of highly penetrant, dominant allelic variants, could account for 2 to 5% of fatal cancers. Environmental factors, including smoking, alcohol consumption, dietary habits and infectious agents of viral (e.g., HPV) or bacterial (e.g., *Helicobacter pylori*) origin, are responsible for varying degrees of cancer susceptibility.[58] According to the estimates of the American Cancer Society, approximately 40% of cancer deaths in 1998 were due to tobacco and excessive alcohol use. The 1996 Harvard Report on Cancer Prevention concluded that over 90% of malignant melanoma is attributable to solar radiation. While other exposures, such as radiation and environmental pollutants, could account for up to 5% of the cancer burden, few causal links with other potential carcinogens have been firmly established.

[58] Cassidy et al., 2010

The medical act of assessing the degree of development and spreading of the neoplastic disease is called staging. Correct cancer staging is critical because treatment (in the form of pre-operative therapy and/or adjuvant therapy) and disease prognosis are based on this evaluation. Staging systems are specific for each type of cancer and are usually sanctioned by international organizations, like the UICC and the AJCC.[59] The most prevalent staging system mirrors cancer progression and classifies solid cancers according to their surgical tractability: Stage 0 represents a tumour confined *in situ*; stage I, a localized, still surgically removable tumour; stage II and stage III describe locally advanced cancers (the specifics depending on the type of cancer being staged); and stage IV marks a metastatic cancer, spread to other organs throughout the body.[60] Another staging system, the TNM

[59] Respectively, the Union for International Cancer Control and the American Joint Committee on Cancer.

[60] Greene, 2002

classification of malignant tumours, also applies to the majority of solid cancers. It relies on the size and extension of the primary tumour; on its lymphatic involvement; and on the presence of metastases to classify cancer malignancy and to inform treatment decisions.[61] For breast cancer, the Nottingham modification of the Bloom-Richardson scale is most commonly used. This grading scale classifies each cancer in a scale from 1 to 3—each describing, respectively, low-, intermediate- and high-grade neoplasias.[62]

[61] Denoix, 1946

[62] This classification system positions cancer along a differentiation axis: from low-grade, well differentiated tumours; to high-grade, poorly differentiated ones.

Cancer research

Most of these classification systems have been implemented during the second half of the last century and reflect a compromise between the need to find sensible and universal guidelines for cancer treatment and our unfolding understanding of the disease. Medical cancer research has come a long way since the times when nearly every disease was attributed to the workings of some invisible force such as bile, miasmas or bad humours (Table 1).

It was in 1863 that Rudolph Virchow, through the lens of a microscope, deduced the cellular origin of cancer.[63] At once, cancer was being recasted as the quintessential disease of hyperplasia and rescued back to the somatic realm.[64] Albeit localized in its origin, cancer was nonetheless perceived as a humoural disease and a systemic illness. The supposition that cancer spreads in a centrifugal fashion from the primary tumour to adjacent structures was the foundation for William Halsted to introduce, in 1894, the radical mastectomy for breast cancer.[65]

[63] Virchow, 1863

[64] *Omnis cellula e cellula*—every cell originates from a cell alike—, is the epigram popularized by Virchow, stating a shift from the tenet of spontaneous generation that dominated the 19th century school of thought concerning cancer's origins. Noting that this form of cellular multiplication was fundamentally novel and inexplicable, he coined it *neo*plasia.

[65] This surgical procedure demands that the breast, the underlying chest muscles and the lymph nodes of the axila be removed (Halsted, 1894). From 1895 to the mid-1970's, about 90% of the women treated for breast cancer in the USA underwent radical mastectomy.

During the 19th century, surgery was the only known way to treat cancer. The first example of a cancer cure by surgery happened in 1809 with the removal of an ovarian tumour without anesthesia. Surgery protocols were subsequently enhanced with the use of anesthesia, first reported in 1846,[66] and the introduction of antisepsis, in 1867.[67] Halsted's radical mastectomy advocated the *en bloc* resection of the surrounding tissue to remove all cancer cells. As cancers kept on relapsing locally after surgery, Halsted reasoned that more and more tissue had to be extirpated in order to root out the last of malignant cells. This pushed radical mastectomy into the "super-radical" and then into the "ultra-radical."[68] By the turn of the century, *en bloc* resection became known as "the cancer operation" and turned into the standard approach for the removal of all other cancers.

[66] Warren, 1846

[67] Lister, 1867

[68] This was an extraordinarily morbid, disfiguring procedure in which surgeons removed the breast, the pectoral muscles, the axillary nodes, the chest wall, and occasionally the ribs, parts of the sternum, the clavicle, and the lymph nodes inside the chest (Mukherjee, 2011).

This surgical tradition came to be challenged in 1968 by Bernard Fisher, a surgeon from Philadelphia. Against the prevailing consensus, Fisher conducted a series of clinical trials in the 1960's to compare the performance of radical mastectomy with localized surgery (a "lumpectomy"), supplemented with radiation. The results of these trials showed that *en bloc* surgery was no more effective to treat early-stage breast cancer than the combination of surgical extraction of the tumour mass and radiation therapy.[69] Radical mastectomy was a fallacy that pushed too far when the tumour was localized, and not enough when the cancer had turned metastatic.

[69] Fisher et al., 1985a; and Fisher et al., 1985b

In 1928, Henry Coutard, a radiologist from the Institut Curie in Paris, showed that fractioned radiation treatments could be used to cure head and neck cancers.[70] In those days, the treatment was called Roentgen therapy,

[70] Coutard, 1932

Year	Discovery or Event	Relative Survival Rate
1863	Cellular origin of cancer (Virchow)	
1889	Seed-and-soil hypothesis (Paget)	
1912	Transplantable rodent tumours	
1914	Chromosomal mutations in cancer (Boveri)	
1928	Head and neck cancer cured by fractionated radiotherapy	
1950	Experimental evidence links lung cancer to smoking	
1953	Report on structure of DNA	35%
1961	Breaking of the genetic code	
1967	Proof of principle: drug cures for Hodgkin's disease and childhood leukemia	
1974	Adjuvant chemotherapy for breast cancer	
1976	Cellular origin of retroviral oncogenes	50%
	Link discovered between HPV and cervical cancer	
1979	Epidermal growth factor and receptor	
1981	Suppression of tumour growth by p53	
1984	G proteins and cell signaling	
1985	First effective cancer immunotherapy with interleukin-2	
1986	Retinoblastoma gene	
1990	First decrease in cancer incidence and mortality	
1991	Association between mutation in *APC* gene and colorectal cancer	
1994	Association between *BRCA1* and breast cancer	
1996	Proof of principle: targeted therapy with imatinib for CML	
1998	Tamoxifen reduces breast-cancer incidence	
2000	Sequencing of the human genome	
	FDA approves HPV vaccine to prevent cervical cancer	
2002	Epigenetics in cancer	
	micro RNAs in cancer	
2005	First decrease in total number of deaths from cancer	68%
2006	Tumour-stromal interaction	

Table 1: Some of the landmarks in the last 200 years of cancer research (adapted from DeVita and Rosenberg, 2012).

after Wilhelm Röntgen, a lecturer at the Würzburg Institute in Germany. While working with an electron tube in 1895, Röntgen discovered a form of radiant energy he called x-rays. The discovery of radium in 1898 by Pierre and Marie Curie further opened the door to the use of radiation to kill cancer by "burning" it.[71] All throughout the first half of the 20th century, this use of radiation therapy mirrored the advent of the atomic age, with "cyclotrons", "supervoltage rays", "neutron beams" and "millions of tiny bullets of energy" all harnessed to eradicate what the surgeon's knife could not reach.[72] The pinnacle of this era came in 1968, in Stanford, when Henry Kaplan demonstrated, with what was to become one of the first controlled medical trials in oncology, that Gamma-knife radiosurgery could significantly increase the survival rate of early-stage Hodgkin's disease.[73]

These advances in surgery and radiotherapy were most beneficial to patients with early, localized forms of cancer. Metastatic disease, on the

[71] It was not just cancer cancer cells that were being burned. Marie Curie died of leukemia in July 1934. Emil Grubbe, the first American to use x-rays in the treatment of cancer, had his fingers amputated to remove necrotic and gangrenous bones and his face cut up in repeated operations to remove radiation-induced tumours and pre-malignant warts. He died at the age of eighty-five to metastatic cancer (Mukherjee, 2011).

[72] Mukherjee, 2011

[73] Kaplan, 1968

other hand, requires a systemic cure. Furthermore, not all tumours respond equally to generic treatments, underscoring a fundamental aspect of cancer's biology—its diversity. When, a hundred years ago, the German chemist Paul Ehrlich launched the first systematic attempt to find chemical substances with specific affinity to malignant cells in order to poison them, he was in essence devising a new form of treating cancer. He called it chemotherapy.[74]

To maximize the efficiency of anti-cancer drug screening, two conceptual advances had to be achieved. First, a cancer model was needed in which the impact of therapy could be effectively quantified. Second, a form of circumventing the ethical limitations of testing human patients had to be found. The first issue was addressed by Sydney Farber at the Children's Hospital in Boston when he turned his attention to childhood's leukemia in the 1940's.[75] In cancer medicine, leukemia is a particularly appealing model because it offers the possibility to actually count the number of cancer cells flowing in the blood—and thus measure the response to the treatment. The recruitment of murine models to test drug response in grafted tumours provided the second piece of the puzzle needed to spur the hunt for anti-cancer drugs.[76]

In 1955, a national screening effort for the development and testing of chemotherapeutic drugs was launched in the USA. This lead to the concoction of highly toxic cocktails of drugs aimed at "maximal, intermittent, intensive, upfront" chemotherapy to vanquish the disease.[77] One of such high-dose, life-threatening regimens, known as VAMP,[78] was tested on children with acute lymphoblastic leukemia (AML), in a trial based at the NCI in 1961. The morbidity of the treatment was egregious and only two of the fifteen children subjected to the initial protocol survived it[79]—one of which was still alive in 2008.[80] In spite of all their complications, the VAMP[81] and the MOPP[82] (an equally aggressive regimen to treat advanced Hodgkin's disease) trials proved that cancer could be cured by chemotherapy.

All these experimental drugs were selected from lists of synthetic chemicals, fermentation products and plant derivatives. Their anti-cancer ability was purely deduced on an empirical basis and the only feature they shared was their rather indiscriminate effect as cell cycle inhibitors. The first case of a chemotherapeutic drug being reasoned from its biological underpinnings occurred in 1969, with the confirmation that tamoxifen could bring metastatic breast cancer into remission.[83] Tamoxifen, a molecular mimicker of estrogen, was first synthesized in 1962 in the UK with the goal of being marketed as a hormonal contraceptive. However, tamoxifen turned out to be an estrogen antagonist instead: by binding to the estrogen receptor, it deprives the cell from the necessary signaling to trigger its cell cycle.[84] Adjuvant chemotherapy for the treatment of breast cancer, i.e., the use of chemotherapy after surgical extraction of the primary tumour, was shown to decrease the rate of relapse for the first time, in a trial launched in 1974.[85]

By 1990, the three-pronged approach of surgery, radiotherapy and chemotherapy, complemented with prevention campaigns and improved diagnostic tools for early diagnosis, led to the first reported decrease of cancer incidence and mortality.[86] It was a worthy achievement, but one that fell short of the haughty rhetoric to "cure",[87] "conquer",[88] or win "the war on cancer"[89] that was voiced throughout the 20th century from several corners

[74] "Give up all hope ye who enter"—was the reading on Ehrlich's lab door (DeVita and Chu, 2008).

[75] DeVita and Chu, 2008

[76] Clowes and Baeslack, 1905

[77] Frei, 1985

[78] VAMP is based on a combination of four drugs: vincristine, amethopterin, mercaptopurine and prednisone.

[79] "If we didn't kill the tumour, we killed the patient"—*William Moloney on the early days of chemotherapy* (Moloney et al., 1997).

[80] Mukherjee, 2011

[81] Frei et al., 1965

[82] DeVita et al., 1970

[83] Cole et al., 1971

[84] Jordan, 1977

[85] Bonadonna et al., 1976

[86] DeVita and Rosenberg, 2012

[87] "We have a cure for breast cancer"—*Emil Frei to a colleague*, summer of 1982 (Mukherjee, 2011).

[88] "Why don't we try to conquer cancer by America's 200th birthday? What a holiday that would be!"—advertisement published in the *New York Times* in December 1969.

[89] The National Cancer Act was signed by Richard Nixon on December 23, 1971.

of the cancer research establishment. In fact, the field was starting to come to terms with the evidence that, no matter how aggressive the treatment,[90] our capacity to arrest the progression of cancer had been pushed to its limits. The most expressive sign of this tacit acquiescence was perhaps the rise in prominence of palliative medicine during the 1980's—prolonging life at any cost no longer was the purpose of cancer medicine.

In 1997, John Bailar published a review article in the *New England Journal of Medicine* entitled "Cancer undefeated".[91] He concluded:

> The war against cancer is far from over. Observed changes in mortality due to cancer primarily reflect changing incidence or early detection. The effect of new treatments for cancer on mortality has been largely disappointing. The most promising approach to the control of cancer is a national commitment to prevention, with a concomitant rebalancing of the focus and funding of research.

The call for a more fundamental understanding of the neoplastic cell had been made. But, in order to embrace it, complementary approaches to microscopy, x-ray scans, or mice models, would be needed.

[90] In order to allow for an otherwise intolerably high chemotherapeutic dosage, Emil Frei devised in 1982 a trial for advanced breast cancer treatment contemplating an autologous bone marrow transplantation. The re-implanted frozen bone marrow cells would thus be spared the excessively high drug dosage. This regimen, known as STAMP, became embroiled in controversy during the next twenty years. It was finally put to rest in 2011, when proof was published that it added no discernible benefit to patient's overall survival (Berry et al., 2011).

[91] Bailar and Gornik, 1997

Microarrays

In October 1995, a short report in *Science* magazine caught the eye with a figure showing six grids of colourful stains on a dark background (Figure 2). The colour of each spot captured the fluorescence emitted when a 3.5 mm by 5.5 mm slide of glass was scanned with a laser. Each slide had been previously spotted with an array of microscopic droplets of forty-five clones of cDNA isolated from *Arabidopsis thaliana*—a small flowering plant with the smallest genome of any known higher eukaryote. Before being scanned, the slides, or microarrays, were hybridized with a solution of fluorescently labeled, reverse transcribed cDNAs, synthesized from mRNA templates extracted from the plant.

The vivid readouts from the microarrays were literally illuminating the transcription patterns of the arabidopsis genome.

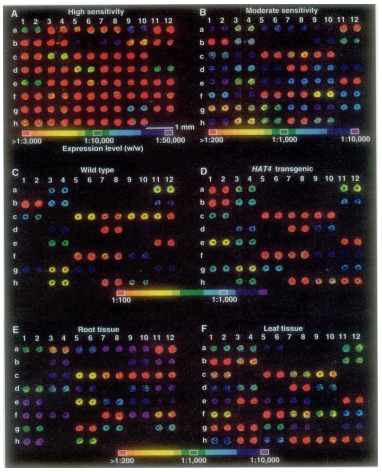

Figure 2: Gene expression of *Arabidopsis thaliana* monitored with cDNA microarrays. **A–F**: each panel shows the hybridization intensity of a mix of fluorescently labeled cDNAs with a collection of forty-five gene-specific probes from arabidopsis, plus three controls, under each stated condition (see text). Adjacent pairs of spots are experimental duplicates. Negative controls were spotted on positions *c*(11, 12) and *h*(11, 12). Positive controls were provided by adding a fixed diluted quantity of mRNA of the human acethylcoline receptor gene to each sample before reverse transcription. cDNA probes of the positive control were printed on positions *a*(1, 2). Probes for the *HAT4* gene were printed on positions *e*(1, 2) (reproduced from Schena et al., 1995).

With this seminal report from Stanford, Mark Schena and Patrick Brown demonstrated three things. First, that microarray technology was *sensitive* and *specific* enough to discriminate between distinct mRNA species in solution (Figure 2, **A–B**). Second, that it could *quantify* levels of mRNA transcripts (Figure 2, **C–D**). Third, that the technology was suited to investigate gene expression patterns in diverse tissue types (Figure 2, **E–F**). As Patrick Brown would state later, microarrays were developed to "enable a new method for relating sequence differences in genes to complex traits in people." And what bigger challenge of complexity could there be but cancer?

Cancer class discovery with microarrays

Four years on, Todd Golub and Eric Lander, at the Broad Institute in Boston, reported the results of the first tackling on cancer using microarrays.[92] As a test case, they took to Farber's acute leukemia, isolating mRNA from 38 biological samples of neoplastic bone marrow and peripheral blood. Their approach was framed as a classification task. Up to then, the problem of classifying cancer subtypes was mostly left to the expertise of the pathologist,[93] and many cancers still lacked molecular markers for their accurate definition. Moreover, the correct distinction between acute lymphoblastic leukemia (ALL, originating from lymphoid precursors) and acute myeloid leukemia (AML, originating from myeloid precursors) was critical for the determination of the chemotherapeutic regimen to be used.

Using a microarray with probes reporting for 6817 human genes, they concluded that the patterns of expression of a subset of these genes could be used to accurately discriminate between AML and ALL. They then sought to use the expression data to blindly *infer* eventual tumour sub-classes among the samples. Using a learning algorithm to build a classifier from the data and then testing it with a cross-validation procedure, they showed that the AML-ALL distinction could be automatically discovered and confirmed without a biological *a priori*.[94] What's more, this class discovery approach could be further refined to automatically detect the distinction between B-cell and T-cell ALL. The use of microarrays could thus enable a molecular

[92] Golub et al., 1999

[93] Clinical practice for cancer classification would involve an experienced pathologist's interpretation of the tumour's morphology, histochemistry, immunophenotyping, and cytogenetic analysis.

[94] These class prediction and class discovery tasks illustrate the distinction between *supervised* and *unsupervised* learning. While the former derives a function from labeled training data (thus requiring an *a priori* knowledge of the classes it tries to predict), the latter aims to produce a classification without any prior knowledge of the structure in the data.

classification of cancer, even if this landmark experimental setup failed to find a multigene expression signature to predict response to chemotherapy.

Back in Stanford, the focus was being directed towards breast cancer. With a microarray probing for 8102 human genes, Charles Perou, Patrick Brown and David Botstein studied the variation in gene expression patterns of breast tumours from 42 patients.[95] Using an unsupervised hierarchical clustering algorithm, they were able to identify at least five distinct molecular breast cancer classes.

Two types of epithelial cells are found in the human mammary gland: basal cells (the outer layer of myoepithelial cells in the mammary duct) and luminal cells (at the apical surface of the ducts, with secretory proprieties). In Perou's study, cancers of luminal origin were found to cluster in two previously unrecognized groups, termed luminal A (of lower grade) and luminal B (of higher grade). Cancers with a basal-like phenotype, over-expressing the HER2 receptor, or with a normal-like phenotype were each found to cluster together in their respective group. However, the most robust distinction was observed between the transcriptome of breast cancers expressing the estrogen receptor (ER+) and those that did not (ER–). This pioneering study showed that a new taxonomy of breast cancer could be based on its molecular features—a classification that would be challenged, extended and refined throughout the ensuing decade.[96]

While these findings were being reported, Eric Lander was leading another collaborative effort that would redefine the breadth of microarray technology. On February 2001, a public-funded consortium reported the first draft of the human genome (Figure 3). Prior to this achievement, the estimated number of genes in our genome was around 100 000.[97] As the genome sequence quality and gene finding methods improved, this figure was progressively revised down to an estimated 20 000 to 25 000 human protein coding genes. The prospect of measuring the *entire* transcriptome of the neoplastic cell with microarrays was now within reach—and would soon turn into a reality. If cancer was fundamentally a genetic disease,[98] then the study of the cancer genome with microarrays would bring it into the genomic era.

It was with microarrays probing for most of the then reported human genome that the expression profiles of lung adenocarcinomas[99], hepatocellular carcinomas,[100] and gastric cancers[101] were interrogated. Time and again, unsupervised classification methods were highlighting clinical subtypes that recapitulated morphological categorizations, underlined tumour differentiation stages, or even uncovered tentative progression markers. Each of these portraits revealed a wide diversity in tumour profiles, both at the intra- and inter-patient sampling level, and a relatively minimal variation in normal tissue profiles. While nuanced, the transcriptomes of these different cancers were still remarkably consistent within each disease and largely reminiscent of the expression profiles of the normal tissues from which they were derived.[102]

The use of class discovery algorithms, mostly as descriptive techniques, dominated the early genomic approach to cancer biology.[103] However, the problem of predicting cancer progression, response to treatment or survival time remained an elusive one. In order to provide sound evidence

[95] Perou et al., 2000

Figure 3: Cover of *Nature* magazine of February 15, 2001.

[96] Sørlie et al., 2001; Sørlie et al., 2003; Hu et al., 2006; Pusztai et al., 2006; Rakha et al., 2008; Parker et al., 2009; Gusterson, 2009; Weigelt et al., 2010; and Prat and Perou, 2011

[97] Cox et al., 1994

[98] "The revolution in cancer research can be summed up in a single sentence: cancer is, in essence, a genetic disease"—*Bert Vogelstein* (Vogelstein and Kinzler, 2004).

[99] Garber et al., 2001

[100] Chen et al., 2002

[101] Leung et al., 2002

[102] Botstein, 2003

[103] Matros et al., 2004; and Eschrich and Yeatman, 2004

for predictive genomic markers, an experimental setting with a systematic, long term follow-up of cancer patients was in demand.

Cancer outcome prediction with microarrays

In the early 1980's, the pathologists of the Nederlands Kanker Instituut (NKI) in Amsterdam began a frozen tissue bank of tumours from Dutch women breast cancers. Twenty years on, the mRNA of these samples, along with the patient's clinical histories, would be the subject of a page turning gene expression profiling experiment. A group including NKI's head of molecular pathology, Laura van't Veer, head researcher René Bernards, and Stephen Friend, a Weinberg trainee turned Rosetta Inpharmatics founder in Seattle, analyzed a selection of primary tumours from 98 women younger than 55 who did not develop lymph node metastasis.[104]

[104] van't Veer et al., 2002

When the tumours were originally resected, treatment standards did not require adjuvant chemotherapy after surgery. Thirty-four patients, or roughly one third of the women in the study, would later relapse of of their cancer. According to modern guidelines, approximately 95% of the original patients would have received chemotherapy in the United States, and 85% would have been treated under European norms. This entails that 55% to 65% of the patients would have needlessly undergone an aggressive and debilitating form of chemotherapy. To assess whether gene expression profiles could predict metastatic relapse of disease, the NKI team profiled the tumours on a microarray containing approximately 25 000 genes. Using a supervised iterative learning procedure on a subset of these features, they narrowed down a list of 70 genes whose expression levels correlated with the development of distant metastasis. The robustness of this 70-gene prognosis profile was subsequently validated on a wider cohort of 295 breast cancer tumours with either positive or negative nodal status.[105] In this study, the classifier accurately predicted overall survival and distant metastasis in stratified univariate analyses, and was the strongest predictor of distant metastasis in a multivariate model that included traditional breast cancer predictors.[106] Microarrays were now being used to stratify cancer risk among patients based on the gene expression profile of the tumour.

[105] Van De Vijver et al., 2002

[106] Classical prognosticators for breast cancer include age, tumour size, status of axillary lymph nodes, histological type of the tumour, pathological grade and hormone-receptor status.

The genes included in the predictor were scrutinized for potential "insight into the underlying biological mechanism leading to rapid metastasis." The van't Veer article in *Nature* reports that "genes involved in cell cycle, invasion and metastasis, angiogenesis, and signal transduction are significantly upregulated in the poor prognosis signature." Not only did this work produce the first genomic predictor to inform treatment decisions, it also paved the way for an alternative to infer physiological mechanisms in human cancers. If the tumour transcriptome already contained information regarding disease progression,[107] then querying for biologically motivated collections of genes among the predictive features could make proof of the implication of particular genetic programs in cancer biology.

[107] "Even though you could look under the microscope and they all look the same (…) some have built into them programs to become aggressive"—*Stephen Friend*

It was with a similar reasoning in mind that the Stanford group presented the argument for the link between a wound healing genetic program and cancer progression.[108] The argument begins with the recognition of the

[108] Chang et al., 2004

similarities between the tumour microenvironment and normal wound healing. It then proceeds by characterizing a gene expression profile of fibroblast serum response (which physiologically only occurs in the context of a local injury), in a cell culture model profiled by microarray. Finally, this signature profile is used to test specific hypotheses using publicly available gene expression data from human cancers.

Accordingly, they demonstrated that, in a cohort of 51 breast cancer patients with equal treatment, those with a higher expression of the core serum response signature were significantly more likely to develop metastasis and to die in a 5-year follow-up period. Similar results were obtained by segregating the 295-sample NKI cohort along an axis of expression of the serum response signature. The signature was also shown to be predictive of outcome in a dataset of 62 patients with stage I and stage II lung adenocarcinomas[109] and a dataset of 42 patients with stage III gastric carcinomas.[110] This formulation established a novel framework to infer biological determinants of cancer progression based on gene expression profiles of clinical samples, obviating the need for experimental setups on *in vivo* models.

The rehashing of this strategy would prove exceptionally prolific. In the wake of the Chang et al. publication, links between cancer progression and various biological signature markers were reported—including gene expression programs of stem cell-ness;[111] p53-status;[112] stromal component;[113] response to hypoxia;[114] chromosomal instability;[115] loss of *PTEN* expression;[116] EMT transition;[117] *E2F1* perturbation;[118] among many more.

[109] Garber et al., 2001
[110] Leung et al., 2002

[111] Glinsky et al., 2005; and Ben-Porath et al., 2008
[112] Miller et al., 2005
[113] West et al., 2005
[114] Chi et al., 2006
[115] Carter et al., 2006; and Buffa et al., 2010
[116] Saal et al., 2007
[117] Welm et al., 2007; and Taube et al., 2010
[118] Hallstrom et al., 2008

State of the art of microarray technology

By 2005, ten years after the arabidopsis report, the partnership between microarray technology and cancer research was in full swing (Figure 4). Reflecting the increasing appeal of the technology, at least half a dozen vendors were then marketing whole genome microarrays, each relying on their own specifics (Table 2).

Figure 4: Cover of *The Journal of Clinical Investigation* of June 1st, 2005.

Microarrays can typically be designed in a two-colour (dual-channel) or in a single-colour (single-channel) setup. In dual-channel microarrays, cDNAs prepared from two samples (usually diseased *versus* healthy tissue) are each labeled with its own fluorophore, then mixed and hybridized on the same microarray. The ratios of the measurements of the fluorescence emission in the wavelength of each fluorophore is then used to estimate the relative abundance of individual transcripts in the two samples. Single-channel microarrays measure the hybridization intensities of a single population of cDNAs labeled with a unique fluorophore and, therefore, express the relative abundance of transcript expression across biological samples processed in the same experiment. Oligonucleotide microarrays often carry control probes designed to hybridize with RNA spike-ins. The degree of hybridization between the spike-ins and the control probes is used to normalize the hybridization measurements for the target probes.

Other platform specific attributes include the probe manufacturing process (made *in situ* by photolithographic or ink-jet methods, or by standard oligonucleotide synthesis protocols followed by attachment to various

substrates); the probe substrates (activated glass slides, silicon chips, or membranes); the probe design and location (most probes are derived from the 3' end of the gene coding sequences to accommodate the fact that target labeling usually begins at the 3' end of mRNAs); probe size and number per array (Table 2); and the proper probe annotation (as sequence databases were still in state of flux, probe annotations were constantly being revised and did not necessarily target their designated gene).[119]

[119] Kawasaki, 2006

Vendor	Probe Size	# Probesets	# Probes per array
ABI	60mer	33 000	33 000
Affymetrix	25mer	54 000	1 000 000
Agilent	60mer	44 000	44 000
GE Amersham	30mer	57 000	57 000
Illumina	50mer	46 000	1 500 000
Microarrays, Inc.	70mer	49 000	49 000
NimbleGen	60mer	38 000	380 000
Phalanx Biotech	60mer	30 000	30 000
"Home brew"	50mer–*n*70mer or CDNAS	40 000	40 000

Table 2: Technical attributes of the principal commercial microarray platforms by 2005. A probeset constitutes a collection of probes targeting a specific gene (adapted from Kawasaki, 2006).

In addition, the expression data resulting from a microarray experiment can be influenced by a number of experimental factors, like target CDNA synthesis (linearly amplified RNA may contain biases in the original mRNAs ratio);[120] target labelling (different fluorescent dyes present distinct stabilities, quantum efficiencies and wavelengths for stimulation and emission); hybridization and washing protocols (every commercial platform abides by its own methodology); and the imaging of the arrays (usually done by confocal and non-confocal scanners—yet variables like laser power, pixel sizes or scan time are not standardized).

[120] Nygaard and Hovig, 2006

These sources of technical variation were making comparison of results obtained from different microarray platforms difficult or even impossible. Some studies were reporting poor correlations between expression levels measured with different platforms.[121] In order to improve the reliability and concordance of microarray data, international consortia and technical study groups were assembled to determine a core set of *Minimum Information About a Microarray Experiment* (MIAME) standards[122] and, in 2006, the *MicroArray Quality Control* (MAQC) project was launched.[123] As a result of these concerted efforts, the National Center for Biotechnology Information in the United States created, in 2002, the *Gene Expression Omnibus* (GEO),[124] an online repository for high-throughput gene expression data. In 2003, the European Bioinformatics Institute started *ArrayExpress*,[125] a public database of microarray gene expression data.

[121] Tan et al., 2003; and Shi et al., 2005

[122] Brazma et al., 2001

[123] MAQC Consortium et al., 2006

[124] `http://www.ncbi.nlm.nih.gov/geo/` (Edgar et al., 2002)

[125] `http://www.ebi.ac.uk/arrayexpress` (Brazma et al., 2003)

Gene expression profiling studies are also challenged by biological idiosyncrasies. For one, global gene expression patterns are a function of fluid states in coordinated cellular ecosystems in constant readjustment. Microarray experiments consist of snapshots of such dynamic ranges, which may account for some of the variation across experiments. Distinct synthesis and degradation rates of the probed mRNA transcripts may further nuance expression readings. Even traditional housekeeping genes (fundamental to the basic biology of the cell and thus considered gold standards) have been

shown to differ across cell types and experimental conditions.[126] What is more, biological heterogeneity in the biopsy sample can significantly bias reports of gene expression, as distinct cellular types may be differently represented in distinct samples.

[126] Thorrez et al., 2008

But by and large, the most contentious aspect of the application of microarrays in cancer research concerns the methodological analysis of the experimental data. In 2007, a critical detailed review of forty-two studies for cancer outcome appeared in the *Journal of the National Cancer Institute* by Alain Dupuy and Richard Simon, from the Hôpital Saint-Louis in Paris and the NCI in Maryland.[127] They identified three endemic analytic flaws permeating the reviewed studies.

[127] Dupuy and Simon, 2007

Microarray experiments typically aim for one or more of the following objectives: (*a*) to identify individual genes (transcripts) whose expression is correlated with a phenotypic trait; (*b*) to identify multiple genes interactively involved in regulatory networks and in mediating biological phenomena or disease pathogenesis; (*c*) to discover potential targets for drug development; and (*d*) to identify molecular markers that can be used as tools for disease diagnosis and prognosis or as predictors of clinical outcome.[128] In outcome-related microarray experiments, these aims can be approached with statistical tools addressing three kinds of tasks: finding genes correlated with outcome, class discovery, and supervised prediction.

[128] Kim et al., 2010

For the outcome-related gene finding task, Dupuy and Simon identified a trend for an inadequate, unclear, or unstated method for controlling the number of false-positive differentially expressed genes. Because microarray analysis involves making inferences about each gene whose expression is measured on the array, the large number of hypotheses being tested can yield a higher than desirable proportion of false positives. Statistical procedures to correct for excess of false positives in multiple testing, like the false discovery rate, are thus necessary.[129]

[129] Benjamini and Hochberg, 1995; and Noble, 2009

Concerning the class discovery task, they recognized a tendency to credit expression clusters with biological meaning when the clustering procedure was itself based on genes selected for their correlation with outcome. This causes non-independent evidence that outcome can be predicted based on expression levels—a statistical misconception known as feature selection bias.[130]

[130] Ambroise and McLachlan, 2002

For the supervised prediction task, they flagged analytic lapses causing a biased estimation of the prediction accuracy through incorrect cross-validation procedures. The more common of these experimental design errors involved the violation of the principle of separation of the training and testing sets during the validation of the classifier. The models derived thusly were likely prone to data overfitting.[131]

[131] Overfitting occurs when a classifier describes random error or noise pertaining to the training set instead of the underlying structure of the data.

With the emerging developments on the computational analysis of microarray data[132] and the concomitant accrued sensitivity to its technical specifics, the stream of published microarray studies started to be tempered by a series of reviews questioning the validity, reproducibility and biological significance of the results.[133]

[132] Quackenbush, 2001; and Irizarry et al., 2003a

[133] Michiels et al., 2005; Tinker et al., 2006; Kawasaki, 2006; Cahan et al., 2007; Gusnanto et al., 2007; Mathoulin-Pelissier et al., 2008; and Kim et al., 2010

In iconic fashion, John Ioannidis, now a Professor of Medicine at Stanford, chastised the field with a dire assessment on the lack of reproducibility of some high-profile microarray research findings.[134] Upon indepen-

[134] Ioannidis et al., 2009

dent dissection of the analysis protocols of eighteen studies published in *Nature Genetics* during 2005 and 2006, his team concluded that ten of them could not be reproduced at all. The main reason for failure to reproduce was data unavailability, and discrepancies were mostly due to incomplete data annotation or specification of data processing and analysis.

In the meantime, prognostic gene expression signatures of clinical outcome of breast cancer were accumulating in the literature at a steady rate.[135] Intriguingly, very few genes were showing in common among the distinct prognostic markers. On a thorough review on the correspondence between these biomarkers and the clinicopathological features of breast cancer, Christos Sotiriou and Lajos Putszai, from the Institut Jules Bordet in Brussels and the University of Texas in Houston, sought to form a synthesis of the evidence supporting genomic prognostic signals.[136] Sotiriou and Putszai reasoned that the absence of common markers between signatures could be a feature of complex gene expression systems entertaining a large number of correlated variables. They also stressed the general tendency for prognostic signatures to perform better among ER+ tumours, as they best discriminate low-proliferation luminal A tumours from high-proliferation luminal B tumours, whereas they mostly classify ER– tumours (comprising the basal-like and HER-positive phenotypes) as high-risk. This could be explained by the ability of prognostic signatures to capture molecular features of tumour differentiation and tumour grade, both linked with cancer progression and metastatic spread. They summed up by proposing that models for breast cancer prognostication should include both genomic and clinical variables for best accuracy.

[135] van't Veer et al., 2002; Paik et al., 2004; Ma et al., 2004; Wang et al., 2005; Chang et al., 2005; Miller et al., 2005; Glinsky et al., 2005; Foekens et al., 2006; Naderi et al., 2006; Teschendorff et al., 2006; Sotiriou et al., 2006; and Liu et al., 2007

[136] Sotiriou and Pusztai, 2009

In 2011, David Venet and Vincent Detours, at the Université Libre de Bruxelles, further added to the debate concerning the biological interpretation of prognostic genomic signals in breast cancer.[137] Probing the 295-sample NKI reference cohort, they compared the prognostic ability of fourty-seven published breast cancer outcome signatures with signatures made of random genes. They showed that 60% of them were not significantly better outcome predictors than random signatures of identical size. In addition, more than 90% of random signatures with more than 100 genes were shown to be significant outcome predictors. Interestingly, they observed that adjusting breast cancer expression data for a proliferation marker abrogated most of the outcome association of published and random signatures. By systematically exploring outcome associations in the NKI-295 cohort, Venet and Detours exposed a wider and more pervasive range of prognostic signals than previously anticipated. They obtained similar results by replicating the analysis in expression profiles of an independent cohort of 380 breast cancer patients from another study.[138] A decade worth of biological extrapolations based on the transcription profiles of clinical samples was eventually challenged by a more stringent formulation of experimental controls.

[137] Venet et al., 2011

[138] Loi et al., 2007

Microarrays, once hailed as the "21st century divining rod,"[139] have highlighted exciting new avenues to engage the neoplastic cell. Still, on the battlefield, cancer remained as deceptive a foe as ever. Translational research with direct impact on clinical practice resulting from the microarray boom has proved tentative and timid at best. The most significant contributions

[139] He and Friend, 2001

were made in the context of predicting a patient's prognosis by interpreting a panel of specific tumour-related genes. The first FDA clearance of microarray-based gene profiling reagents was obtained in May 2011.[140] As of 2013,[141] three genomic assays were commercially available for prognostication in early stage breast cancer: Oncotype DX (consisting of a 21-gene profile narrowed down from a list of 250 candidate genes that were analyzed in a total of 447 patients from 3 separate studies); MammaPrint© (based on the 70-gene NKI predictor; FDA approved in 2007); and PAM50 (a 50-gene set used for standardizing subtype classification).

[140] http://investor.affymetrix.com/phoenix.zhtml?c=116408&p=irol-newsArticle&ID=1561100

[141] Kittaneh et al., 2013

The mining of the wealth of data yielded by cancer expression profiling has been as much a source of promising guidance as of humbling reappraisal.

RNA Sequencing Technology

Microarray technology has played a pivotal role in enhancing our understanding of cancer genomics. Nevertheless, recent technological advancements have largely replaced microarrays as the assay of choice to study transcriptomics. RNA sequencing technology, or RNA-Seq, was developed to approach to transcriptome profiling using deep-sequencing technologies.[142] Compared to array based technology, RNA sequencing has the following advantages: *(a)* it allows for unbiased detection of novel transcripts, alternative transcripts, gene fusions, single nucleotide variants, small insertions and deletions, RNA editing, as no species- or transcript-specific probes are required; *(b)* it has a broader dynamic range, for it quantifies discrete, digital sequencing read counts, and thus gene expression measurement is not limited by background at the low end and signal saturation at the high end, as in microarrays; *(c)* it has increased specificity and sensitivity; and *(d)* it has easier detection of rare and low-abundance transcripts, as sequencing coverage depth can be increased to detect rare transcripts, single transcripts per cell, or weakly expressed genes.

[142] Wang et al., 2009

In the work presented in this dissertation, data generated with microarray technology and RNA-seq technology was used.

Motivation & Contributions of this Thesis

The work contributed in this thesis is anchored in the analytic corpus developed during twenty years of gene expression assaying of cancer biospecimens with microarray technology.

Gene expression markers of proliferation and differentiation in cancer

We sought to investigate the potential of gene expression signatures of proliferation and differentiation to chart cancer progression using whole genome microarray profiles of tumour samples. As most molecular classifiers of cancer rely on the variant features of the transformed neoplastic transcriptome, we reasoned that a complementary approach could consist of leveraging the invariant features specified by the unique transcriptional signatures of each tissue of origin.

As a case study, we took to thyroid cancer. Neoplasms of the thyrocyte cell are characterized by a well defined linear progression from benign, fully differentiated tumour types, up to one of the most lethal human cancers, the anaplastic thyroid carcinoma (Figure 5). We aimed to build a genomic marker of thyroid cancer progression based on a gene expression signature of healthy thyrocytes. As a result, we devised a general method to derive robust organ-specific gene expression-based differentiation indices, published in the journal *Oncogene*.[143]

[143] Tomás et al., 2012

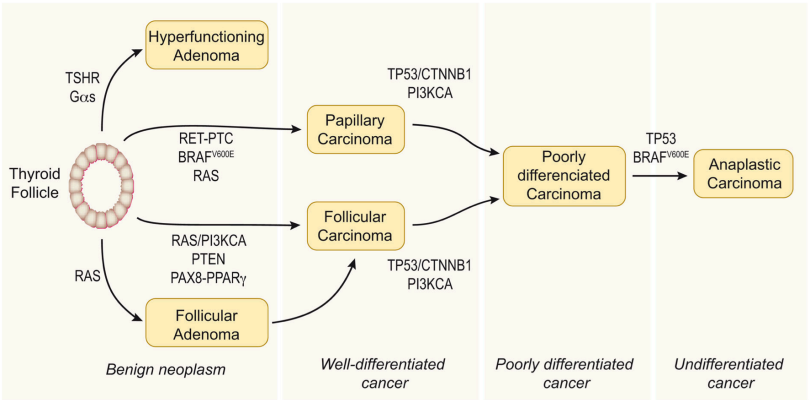

Figure 5: Step model of thyroid carcinogenesis. Thyroid epithelial cells may undergo transformation via alterations in different oncogenes and tumour supressor genes, giving rise to well-differentiated papillary or follicular carcinomas. Additional mutational load can cause progression of a differentiated tumour into a poorly differentiated one, and eventually into an anaplastic carcinoma. Particularly challenging, from the histopathological point of view, is the distinction between follicular adenomas and follicular carcinomas; and between follicular variants of papillary carcinomas and their classical counterpart (reproduced from Sastre-Perona and Santisteban, 2012).

Contributions made in the context of this work include:

- An unbiased procedure to derive organ-specific differentiation markers from gene expression profiles of healthy tissues.
- Proof of concept of the clinical utility of differentiation signatures in cancer diagnosis, featuring thyroid cancer as a test case. Specifically, we demonstrated that, in a panel of expression profiles composed of thyroid cancers of distinct subtypes and normal thyroid samples:

  1. The expression of a thyroid-specific differentiation biomarker, consisting of 15 genes, is inversely correlated with that of a proliferation biomarker, also independently derived from expression profiles of healthy tissues. Conversely, the differentiation biomarker is positively

correlated with the proliferation biomarker in expression profiles of a time-course experiment where thyrocytes in culture were treated with TSH hormone (the thyroid-stimulating hormone, TSH, induces both the metabolic activity and proliferation of thyrocytes). These observations support the independence of the two biomarkers and prove that the differentiation biomarker does capture a transcriptome signature particular to the epithelial thyroid cell.

2. A multidimensional scaling analysis representation of the profiled clinical samples exposes a non-overlapping continuum of thyroid tumours of increasing malignancy.

3. The differentiation biomarker can accurately discriminate between follicular adenomas and follicular carcinomas; and between follicular variants of papillary carcinomas and classical papillary carcinomas—two challenging histopathological diagnosis. Moreover, the accuracy of the differentiation biomarker in this supervised classification task was not significantly different from the accuracy of two supervised machine learning classifiers trained within the whole gene expression space of the tested samples.

The extent of prognostic signals in the cancer transcriptome

Uncontrolled proliferation is not just a hallmark of cancer[144]—but its very own operational definition. Using a proliferation biomarker consisting of 129 genes derived from healthy tissues, Venet and Detours[145] showed that most of the prognostic content found in the reference NKI-295 dataset was linked to a pervasive proliferative signal in the neoplastic transcriptome.

[144] Hanahan and Weinberg, 2011

[145] Venet et al., 2011

We took to a wider collection of 114 distinct outcome-related cancer cohorts, spanning 19 types of cancer to (*a*) evaluate the extent of prognostic signals in human cancers transcriptomes; and (*b*) dissect the potential technical and biological variables linked to prognostic content in different cancer. The results of this work are currently under submission and will be thoroughly detailed in the Results section.

Contributions made in the context of this work include:

- Substantiation of the heterogeneous nature of prognostic signals in cancer transcriptomes.

- Evidence of an extensive correlation structure in cancer transcriptomes as assayed by microarrays. This is concluded on the count that, in 76% of the cancer cohorts analyzed, more than 5% of random gene expression signatures is associated either with patient death or relapse of disease.

- Demonstration that both technical and biological variables are responsible for the heterogeneity of prognostic contents observed in breast cancer. This was determined from a bootstrap analysis of prognostic fractions of breast cancer transcriptomes in the largest publicly available breast cancer cohort, comprised of nearly 2000 breast cancers.

Methods

The contributions of this thesis are based on the analysis of global gene expression profiles of cancer biospecimens with microarray technology. This chapter begins by introducing the technology itself. It then presents the flow of microarray data analysis, including the preprocessing of raw data. Next, it lists and details the analytic methods used to make inferences about gene expression data. Finally, it describes the public microarray datasets used in the analyses reported in the Results section.

Microarray technology

Microarray technology relies on the non-covalent, sequence-specific interaction between complementary strands of nucleic acids[146] to detect and quantify specific populations of mRNA in a solution. A microarray chip consists of a universe of oligonucleotide probes attached to a substrate through covalent bonds. Each such probe is synthesized to specifically match a unique messenger RNA molecule. When the chip is exposed to a solution of fluorecently labeled mRNAs, only those that hybridize with their respective probes will be retained upon washing off non-specific bonding sequences. This allows for the quantification of the fluorescent signals emitted when the chip is scanned with a laser beam of a specific wavelength. The measured signals relay the relative quantity of each mRNA molecule assayed by the microarray, as each spot has a known position on the chip (Figure 6).

[146] Watson and Crick, 1953

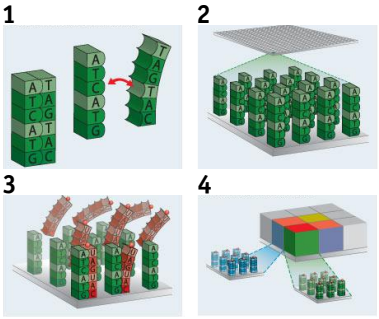

Figure 6: A schematic representation of how microarrays work. **1.** Microarrays rely on a fundamental property of nucleic acids, the monomeric units that polymerize into DNA or RNA strands. Adenine (A) are complementary to thymine (T), and cytosine (C) are complementary to guanine (G). Just one incorrect base can prevent two strands from binding. **2.** A microarray typically contains thousands of squares, or spots. Each spot anchors many copies of a particular sequence of single-stranded DNA, corresponding to a particular gene. **3.** Messenger RNA fragments extracted from a tissue and labeled with different fluorescent dyes are washed over the microarray and hybridize with DNA strands with the complementary sequence. **4.** The dyes are illuminated using fluorescent light. It is then possible to show which RNA fragments were retained in which spots—and hence which genes were being expressed in the tissue from which the RNA was extracted. Source: *The Economist; Affymetrix*.

Microarray data preprocessing

By virtue of their design, microarrays allow for the monitoring of expression levels for thousands of gene transcription products simultaneously. Microarray expression data are thus characterized by high dimensionality and noisiness. This prompts the need for preprocessing methods aiming at removing systematic biases in expression measurements, introduced during experimentation.[147]

The goal of microarray data preprocessing is to convert raw imaging data into meaningful biological data and to enable comparison of results obtained from different arrays. It comprises three steps: (*a*) the transformation of image data into intensity values; (*b*) the assessment of array quality; and (*c*) the removing of technical biases (through background adjustment, normalization and feature filtering and summarization).

The digital imaging of fluorescence signals is typically performed by proprietary software designed by the microarray manufacturer. These

[147] Shakya et al., 2010

software packages assign coordinates to each spot in the array, quantify signal intensity and uniformity of each spot, and compare their signal intensity relative to background.

Quality control is a critical step in the preprocessing of microarray data. In spite of the many efforts to provide standards for the technology, such as the External RNA Control Consortium[148] and the MicroArray Quality Control[149] initiatives, there remains a lack of consensus in both defining and measuring microarray quality.[150] Computational strategies to tackle quality assessment in single-channel and double-channel arrays have been implemented in the Bioconductor package arrayQualityMetrics.[151]

Chips meeting quality standards then undergo background adjustment and normalization. Normalization methods aim to compensate for procedural biases that are independent from biological signal. Early approaches for microarray normalization were based on the assumption that most genes, and in particular so-called housekeeping genes,[152] should have similar expression levels across samples. Housekeeping genes have since been shown to vary in expression by 30% or more across healthy samples, and even more in tumour samples.[153] Data-driven normalization approaches were then developed, such as median correction,[154] variance stabilizing transformation,[155] locally weighted linear regression[156] and spline based methods.[157]

Normalization strategies for double-channel microarrays (spotted oligonucleotide or cDNA arrays[158]) are different from those for single-channel microarrays (*in situ* synthesized high density oligonucleotide arrays,[159] such as *Affymetrix GeneChip*). *Affymetrix GeneChip* arrays use multiple probes per gene and a single-colour detection system, as one sample is hybridized per chip. Spotted oligonucleotide or cDNA arrays use one probe per gene and a two-colour scheme, where two different samples are hybridized on the same array. Consequently, single-channel arrays measure the overall abundance of a probe sequence in a target sample, whereas cDNA arrays measure the relative abundance of a probe sequence in two target samples.

For double-channel arrays chips, normalization methods commonly seek to remove biases within each array with local regression algorithms. Terry Speed's lab, at Berkeley, identified an intensity-dependent dye bias concerning cDNA arrays. In these arrays, the $\log_2$ of the dye intensity ratios shows a systematic dependence on intensity, characterized by a deviation from zero for low-intensity spots. Frequently, under-expressed genes appear up-regulated in the red channel (R), and moderately expressed genes appear up-regulated in the green channel (G). This effect can be visualized by plotting the measured $\log_2(\frac{R_i}{G_i})$ for each feature in the array as a function of the $\log_2(R_iG_i)$ product intensities. This ratio-intensity plot is termed MA plot.[160] This technical bias may be corrected by fitting a locally weighted regression, known as *lowess* smoothing (Figure 7).[161] More specific sources of technical bias, including spatially-dependent bias resulting from the print tips used in the manufacturing process of the array, may also be addressed by a *lowess*-based, within group normalization.

Affymetrix GeneChip are the reference arrays in the single-channel class and consist of several tens of thousands probe-sets. A probe-set is a collection of probe pairs designed to interrogate a specific sequence and contains

[148] Baker et al., 2005

[149] Consortium, 2010

[150] McCall et al., 2011

[151] Kauffmann et al., 2009; and Kauffmann and Huber, 2010

[152] *Housekeeping* genes are genes defined as participating in basic, thus universal, cellular processes.

[153] Lee et al., 2002; and Eisenberg and Levanon, 2003

[154] Cho et al., 1998; and Selinger et al., 2000

[155] Durbin et al., 2002

[156] Yang et al., 2002

[157] Workman et al., 2002

[158] Schena et al., 1995

[159] Lockhart et al., 1996

[160] The name of the plot comes from "minus" and "add", respectively the ratio and product in the logarithmic scale.

[161] Yang et al., 2001

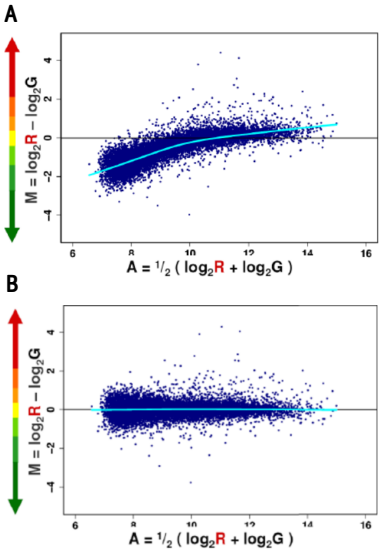

Figure 7: Example of *lowess* normalization. **A:** MA plot showing colour dye dependent bias. **B:** MA plot after correction with *lowess* normalization (Yang et al., 2002).

11 to 20 probe pairs of 25-mer oligonucleotides each. Each probe pair consists of a perfect match probe (PM) and a mismatch probe (MM). The MM probe differs from the PM probe by a single substitution at the center base position, conceived to disturb the binding of the target gene transcript. This design allows for the quantification of background and nonspecific hybridization effects.

Different between-array normalization methods have been proposed in the literature. The MAS5 algorithm[162] normalizes each array independently and sequentially and uses the value of MM probes to compute summarized averages with linear scaling. The RMA (robust multi-array average) algorithm[163] normalizes the value of each probe by quantile normalization[164] in multiple arrays, neglecting information from MM probes. The GCRMA algorithm[165] applies the same normalization and summarization methods as RMA, but uses probe sequence information to estimate and correct for probe affinity to non-specific binding. More recently, the fRMA (frozen RMA) algorithm[166] leverages pre-computed estimates of probe-specific effects and variance from public microarray databases to, in concert with the information from new arrays, normalize and summarize the data.

Several reviews on these and other normalization methods were produced in the specialized literature,[167] and a more detailed technical exposition of the computational implementation of these algorithms can be found in Gentleman et al.,[168] in the context of the Bioconductor project.[169]

The normalized microarray data for p genes and n biological samples are denoted by $X_{n\times p}$ such that x_{ij} represents the expression of gene j of sample i.

Microarray data analysis

In cancer research, analysis of genomic experiments performed with DNA microarray data may address a variety of tasks, including finding gene associations with particular phenotypes, tumour class discovery or tumour class prediction. The work contributed in this dissertation concerns: (*a*) the class prediction of tumour types based on gene expression profiles; and (*b*) the modeling of association of gene expression profiles with the time until a particular outcome is observed (e.g., death of a patient or relapse of disease), using survival analysis. Both these problems are examples of application of supervised learning techniques to the analysis of microarray data.

In machine learning,[170] supervised learning methods seek to infer a function from labeled training data. Conversely, unsupervised learning methods aim to find intrinsic structure from unlabeled data.[171] Unsupervised learning is thus suited to uncover coherent genomic signals from microarray data, whereas supervised learning can be used to find associations between genomic signals and phenotypic classes of samples.

Due to the high dimensionality of expression profiles, methods for dimensionality reduction are required for microarray data analysis. These include *feature transformation* methods and *feature selection* methods.[172]

Feature transformation is an unsupervised approach that consists in reducing the feature space of a microarray gene expression matrix, such that new features retain biological pertinence, maximum information and

[162] Hubbell et al., 2002

[163] Irizarry et al., 2003b

[164] Quantile normalization is a global adjustment method that assumes the statistical distribution of expression values of each sample is the same. Normalization is achieved by imposing the same distribution to all samples, using an average distribution as reference. The average distribution is estimated from the average of each quantile across samples.

[165] Wu et al., 2004

[166] McCall et al., 2010

[167] Ploner et al., 2005; Bolstad et al., 2003; and Harr and Schlötterer, 2006

[168] Gentleman et al., 2006

[169] The Bioconductor project is an open source and open development software project for the analysis and comprehension of genomic data (Gentleman et al., 2004). It is rooted in the open source statistical computing environment R.

[170] Machine learning is a branch of computational science whose purpose is to implement algorithms capable to infer models from training data. Models derived from machine learning methods are then expected to make predictions or to inform decisions on testing data.

[171] Webb, 2003

[172] Haibe-Kains, 2009

generalizability to similar experiments:

$$X_{n\times p} \rightarrow X'_{n\times p'} : p \gg p'. \quad (1)$$

Examples of microarray feature transformation techniques include principal component analysis and clustering methods.

Feature selection techniques, on the other hand, are supervised approaches to reduce data dimensionality, in order to produce simplified and interpretable models.

Gene expression signatures, or metagenes, are collections of genes sharing a combined expression pattern associated with a given phenotype. The range of biological phenotypes confining the characterization of expression signatures include the modulation of signaling pathways,[173] the specification of tumour classes,[174] or the definition of distinct clinical outcomes.[175] In the last two decades, a wealth of microarray studies on perturbed *in vitro* biological systems have generated an extensive number of gene signatures related to various cellular mechanisms.[176] Public repositories, like the Gene Ontology consortium (GO),[177] the Kyoto Encyclopedia of Genes and Genomes (KEGG),[178] GeneSigDB,[179] or the Molecular Signatures Database (MSigDB),[180] have sought to curate and articulate this volume of information in order to guide the interpretation of genome-wide expression profiles. Because biologically motivated gene signatures can act as surrogate markers for the molecular processes they capture, they provide entry points for hypothesis testing in public microarray data.

[173] Itadani et al., 2008
[174] Ramaswamy et al., 2001
[175] van't Veer et al., 2002
[176] Chibon, 2013
[177] Ashburner et al., 2000
[178] Kanehisa and Goto, 2000
[179] Culhane et al., 2012
[180] Subramanian et al., 2005

The use of a common vocabulary to refer to microarray features is thus essential to this goal. Given the wide range of microarray platforms on the market, preprocessing routines often produce genomic expression data with feature annotations that are disjoint, inconsistent, or conflicting. Computational strategies to interface curated annotation databases in order to update and standardize feature nomenclatures are generally poorly discussed in the literature. A solid foundation for *in silico* solutions to bridge the gap between the knowledge of transcript sequence and the knowledge of transcript function is provided in Gentleman et al.[181]

[181] Gentleman et al., 2006

Accordingly, for all datasets used in this dissertation (described in the Microarray datasets section), Bioconductor resources were used to update feature annotations, and referents for HUGO Gene Nomenclature Committee (HGNC) gene symbols[182] were universally retained as feature descriptors. Microarray gene annotation may also be used to perform dimension reduction of the expression feature space. Hence, expression matrices where multiple features addressed the expression of the same gene product were collapsed using a `maxMean` routine.[183] This approach consists of selecting, among all the features measuring the expression of the same gene, the probeset with highest mean expression.

[182] http://www.genenames.org/
[183] Miller et al., 2011

All analyses described in the Results chapter were conducted in the R environment for statistical computing,[184] with extensive use of computational resources from the Bioconductor project.[185]

[184] R Core Team, 2014
[185] Gentleman et al., 2004

This section will proceed with a brief overview of tools for visualization of genomic data, namely heatmaps and multidimensional scaling. It will then discuss the unsupervised learning tools used in this dissertation, including principal component analysis and a summary on machine learning

algorithms. Next, it will cover ROC curves, a tool to evaluate the performance of classifiers. Finally, it will detail survival analysis, the branch of supervised learning that seeks to explain the relationship between a number of measured features (gene expression data) and the time duration until the occurrence of a particular event (survival outcome).

Visualization techniques

The heatmap is the most recognizable visual representation of microarray expression matrices (Figure 8). It translates the $\log_2$ values of expression into a colour-coding scheme that seeks to convey the phenotype-specific transcription patterns of the profiled samples. Because of the high dimensionality of microarray data, and of the nuanced nature of transcription profiles, methods of feature selection are routinely used to narrow down the choice of genes represented in heatmaps. The result is a direct visual representation of the expression matrix, highlighting a subset of genes that maximize the contrast between sampled phenotypes or conditions.

More sophisticated techniques of visualization may require a degree of data transformation. Such is the case of multidimensional scaling (MDS), a visualization technique that seeks to project the n-dimensional feature space of a collection of samples into a two-dimensional space, in such a way that the similarities between samples are preserved as best as possible (Figure 9). The implementation of the MDS algorithm used in this dissertation, known as non-metric multidimensional scaling, requires finding the optimal coordinate matrix whose configuration minimizes a loss function called *stress*, based on a dissimilarity matrix derived from the gene expression space.[186]

Principal component analysis

Principal component analysis, or PCA, is an unsupervised technique of feature transformation, whose purpose is to reduce the dimensionality of a dataset to a few, interpretable, and linear combinations of variables.[187] It proceeds by identifying orthogonal directions, termed principal components, along which the variation of the data is maximal. Principal components are uncorrelated meta-variables, each explaining a decreasing fraction of the partitioned variation within the original data. Similarities and differences between samples are thus best visualized when projected in this new feature space.

Due to the intricate patterns of correlations in their expression space, microarray data are particularly suited for PCA-driven dimensionality reduction.[188] One appealing aspect of PCA analysis of expression profiles is the potential biological interpretability of principal components. For instance, the first principal component of a collection of expression profiles of breast cancers has been shown to explain the separation of samples according to estrogen receptor status (a major predictor of disease progression).[189]

In this dissertation, PCA is used to derive eigengenes from the feature expression space of cancer biospecimens. An eigengene is a meta-variable seizing the modulation of a gene signature in a collection of samples. It can be defined, for instance, by the first principal component of the expression of the metagene in those samples.

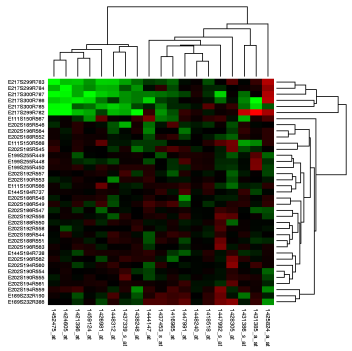

Figure 8: Example of a heatmap generated from an expression matrix, $X_{n\times p}$, issued from a DNA microarray experiment. The expression of a selection of features (columns) is shown for all the profiled samples in the experiment (rows). The traditional colour coding scheme ranges from bright green to bright red, for features highly expressed or repressed between conditions (or regarding a control sample, in double channel arrays), respectively. Features coded in darker shades are not differentially expressed between conditions. Features and samples are hierarchically clustered in dendrograms, to reflect gene co-expression motifs and related expression patterns between samples.

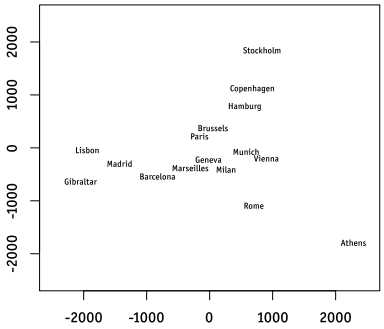

Figure 9: Example of a multidimensional scaling (MDS) based on a distance matrix of road distances, in km, between 16 European cities. In this representation, the data transformation involves the projection of uni-dimensional variables into a two-dimensional space. In microarray data analysis, MDS transformation requires optimally solving the projection of the high dimensional expression space into a plane, so that the between-sample distances are best exposed (see text for details).

[186] Borg and Groenen, 2005

[187] Pearson, 1901

[188] Ringnér, 2008

[189] Saal et al., 2007

Machine learning analysis

Machine learning refers to computational and statistical inference processes employed to create, on the basis of observational data, reusable algorithms for prediction.[190] An essential property of a learning algorithm is generalization. In the context of machine learning, generalization refers to the ability of the trained classifier to perform accurately on new samples. In microarray data analysis, this means identifying expression features related to the underlying biology of the phenotypic classes under analysis.

[190] Gentleman et al., 2006

Several classification and feature selection methods have been co-opted for the identification of differentially expressed genes in microarray data.[191] In this dissertation, we made use of Bioconductor implementations of two supervised machine learning algorithms in order to optimize expression feature selection towards two clinically challenging diagnostic problems in thyroid cancer.

[191] Pirooznia et al., 2008

Support vector machines, or SVM, are classifiers that operate by mapping input vectors into a high dimensional feature space in a non-linear fashion. Linear decision surfaces, or hyperplanes, are then constructed, which can be used to separate classes of data. The optimal hyperplane is the linear decision function with maximal margin between the vectors of the considered classes. Margins can be functionally described by a subset of the training data, the so-called *support vectors*.[192] Random forests, or RF, are classifiers consisting of a collection of decision trees grown from the features that are most discerning in the training set.[193] They operate under the assumption that, in the feature expression space of microarray data, while each individual classifier is a weak learner, the ensemble of all classifiers taken together produce a strong learner. Both of these algorithms were employed with a feature-selection step. Feature selections and algorithm parameters were nested in an inner cross-validation loop to preclude any possibility of parameters and feature-selection biases.[194]

[192] Cortes and Vapnik, 1995

[193] Breiman, 2001

[194] Johannes et al., 2010

Classifiers were thusly trained on $^{9}/_{10}$ of our samples and used to predict the remaining $^{1}/_{10}$. They were tuned on $^{9}/_{10}$ of the training samples, and optimized by comparing the prediction on the remaining $^{1}/_{10}$ of the training samples. For each sample, our implementations of the algorithms returns, for RF, a percentage of decision trees votes, and for SVM, a real score between −1 and 1. This loop was repeated 10 times and the results were averaged. Those scores were then used as a diagnostic test to compute receiver characteristic operating curves.

Receiver operating characteristic curves

Receiver operating characteristic curves, or ROC curves, are a tool for diagnostic test evaluation that allows for the creation of a sensitivity and specificity report.[195] A typical ROC curve plots the true positive rate (sensitivity), as a function of the false positive rate (1 − specificity), for different cut-off points of a parameter produced by a classifier model. Each point on the ROC curve represents a sensitivity/specificity pair, corresponding to a particular decision threshold (Figure 10).

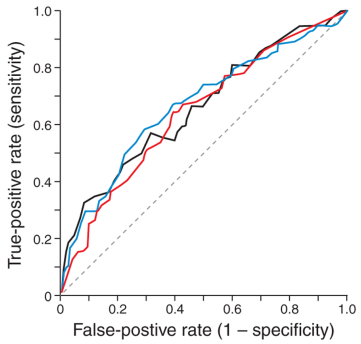

Figure 10: Receiver operating characteristic curves (ROC) are visual representations of the trade-off between sensitivity (the proportion of actual positives which are correctly identified) and specificity (the proportion of true negatives correctly identified) of a diagnostic test, or classifier. The ability to superimpose different AUCs on the same plot allows for direct comparison of different classifiers.

[195] Fawcett, 2006

The area under the curve, or AUC ($0 < \text{AUC} < 1$), can be interpreted as the probability that the predictor of choice will rank a randomly chosen pos-

itive instance higher than a random negative one.[196] The AUC metric allows for direct comparison of classifier performance. A AUC of 50% represents a chance of correct classification no better than a random classifier.

[196] Hanley and McNeil, 1982

Survival analysis

Survival analysis is a collection of statistical procedures for data analysis for which the outcome variable of interest is time until the event occurs.[197] In follow-up studies of cancer patients, survival analysis is used to model association of the expression of genomic markers in cancer biospecimens with the time until a given clinical outcome is observed.

[197] Kleinbaum and Klein, 1996

Clinical outcomes may include death of the patient (overall survival, or OS), death of the patient caused by the cancer (disease-specific survival, or DSS), the finding of new metastases in the patient (distant metastasis free survival, or DMFS) or recurrence of the cancer (disease-free survival, or DFS). *Survival time* refers to the lapse of time since the beginning of the study up to the moment when an event is observed, regardless of the clinical outcome considered. Whenever the information about an individual's survival time is incomplete, that observation is said to be censored (Figure 11). This may be due because the event was not observed by the end of the study (in which case the follow-up time considered for that patient is the entire duration of the study) or because the patient quit or withdrew from the study before its end (in which case the follow-up time considered for that patient is the time up to dropping out).

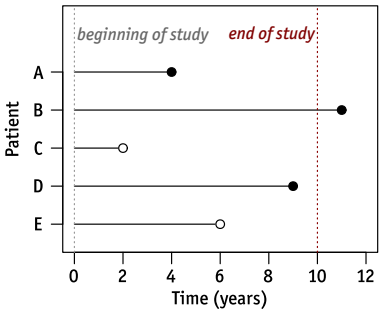

Figure 11: A schematic representation of right-censored survival data. Survival time is said to be *right*-censored when the information regarding the right side of the follow-up period is incomplete. Observed events are denoted by (●). Censored observations are denoted by (○). Notice that patient B is also censored, as no event had been observed by the end of the study (see text for details).

Follow-up studies are not amenable to ordinary regression models because the time to event is typically not normally distributed and these models cannot incorporate censoring data. Instead, survival analysis uses two functions to estimate survival time, the *survival* function and the *hazard* function.

The survival function, $S(t)$, describes the probability of an event occurring later than some specified time t:

$$S(t) = Pr(T > t), \tag{2}$$

where T is a random variable for a patient's survival time.

The hazard function, $h(t)$, describes the instantaneous potential per time unit for the event to occur, given the patient has survived up to time t:

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr\{t \leqslant T < t + \Delta t \mid T \geqslant t\}}{\Delta t}. \tag{3}$$

Both functions are related to each other. However, while the survival function is non-increasing (Figure 12), the hazard function may be modeled by any number of distributions, as it describes a failure rate conditional to the interval of time considered. Building on these functions, survival analysis stipulates a suite of parametric, non-parametric and semi-parametric methods to make inferences over survival time. These methods can then be used to ascertain the relationship between a variable of interest and the time to a clinical outcome.

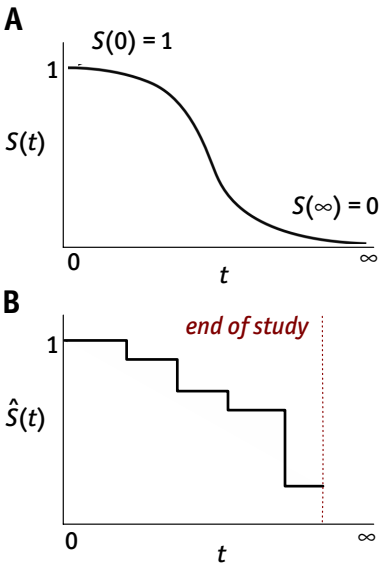

Figure 12: The survival function, $S(t)$ describes the likelihood that a patient will have a lifetime exceeding time t. **A:** The theoretical distribution is non-increasing, and characterized by $S(0) = 1$ and $S(\infty) = 0$. **B:** In practice, the estimated survival function, $\hat{S}(t)$, often takes the shape of a step function. Because study periods are never infinite and there may be competing risks for failure, it is likely that not all patients will experience a clinical outcome by the end of the study.

In the biomedical literature, the most recognizable non-parametric method to estimate the survival function is the Kaplan-Meier estimator.[198] The Kaplan-Meier estimator is defined as the probability of surviving in a

[198] Kaplan and Meier, 1958

given length of time while considering time in many small intervals.[199] It estimates the probability of occurrence of an event at time t by cumulatively multiplying prior probabilities of survival at preceding t_i intervals. The Kaplan-Meier, or product limit estimator, is thus formulated as:

[199] Altman, 1990, pp. 365–93

$$\hat{S}(t) = \prod_{t_i<t} \frac{n_i - d_i}{n_i}, \tag{4}$$

where n_i and d_i are, respectively and for each prior time interval t_i, the number of patients at risk (right-censored observations removed), and the number of patients experiencing an event. The Kaplan-Meier estimator can be used to obtain univariate descriptive statistics for survival data or to compare the survival time for two or more groups of subjects.

The logrank test is a non-parametric hypothesis test to compare the survival distribution of two samples.[200] It challenges the null hypothesis that there is no difference between the survival functions underlying each observed population. To do so, it compares the estimates of the hazard functions of the two groups at each observed event time. The logrank test is based on the same assumptions as the Kaplan-Meier survival curve—namely, that censoring is unrelated to prognosis, the survival probabilities are the same for subjects recruited early and late in the study, and the events happened at the times specified.[201] Importantly, both the Kaplan-Meier estimator and the logrank test are able to incorporate right-censored survival data.

[200] Mantel, 1966; and Peto and Peto, 1972

[201] Bland and Altman, 2004

When modeling the presence of covariates or explanatory variables to explain survival time, fully parametric and semi-parametric approaches are available. Parametric approaches, like the accelerated life class of models, assume that the effect of covariates is proportional with respect to survival time.[202] Under this model,

[202] Kalbfleisch and Prentice, 2011

$$S_1(t) = S_0(t/\gamma). \tag{5}$$

Effectively, this means that the probability that a member of group one will be alive at time t is exactly the same as the probability that a member of group zero will be alive at time t/γ. In parametric models, the estimation of the covariates is conditional to the prior definition of the hazard distributions in both groups.

Alternatively, the Cox proportional hazards family of models assumes that the effects of the covariates is proportional with respect to the hazard.[203] This assumption obviates the need of specifying the underlying hazard functions—the model only seeks to fit the regression parameters, hence being referred to has semi-parametric. In its simplest form, it can be formulated in such a way that the hazard at time t for an individual with covariates X is assumed to be:

[203] Cox, 1972

$$h(t \mid X) = h_0(t) \exp(\beta_1 X_1 + \ldots + \beta_p X_p). \tag{6}$$

This model formulates the hazard of individual i at time t as the product of a baseline hazard function, $h_0(t)$, and of a linear function of a set of p covariates, which is exponentiated. Along with the specification of the model, Sir David Cox developed a maximum partial likelihood method for estimation of its covariates.[204] A logrank statistic can be derived as the

[204] Cox, 1972

score test for the Cox proportional hazards model comparing two groups. The term Cox regression refers to the combination of both the model and the estimation procedure. It has become the tool of choice to estimate the association of the expression of genomic metagenes with differential survival times in cancer follow-up studies.

When considering the association of genomic markers with survival time, the dependent variable is composed of two parts: one consisting of the time to the event (or lack thereof), encoded as a non-negative number; the other registering the event status, encoded as a binary variable (routinely 1 if the event is observed, and 0 if not).

Typically, the predictor is also a binary class, the result of discretizing the patients in groups of good and bad prognosis as a function of the metagene classifier. This requires a method to stratify the cohort with an unsupervised classification procedure. Venet et al.[205] investigated three methods to this goal: (*a*) the standard spliting of the cohort along the two main clusters defined by hierarchical clustering of the metagene's expression data; (*b*) spliting the cohort along the two main clusters defined by applying the *k*-means clustering algorithm to the metagene's expression data; and (*c*) spliting the cohort along the median of the metagene's first principal component. They confirmed that the methods based on the first principal component yielded significantly higher hazard ratios and smaller *p*-values.

[205] Venet et al., 2011

In this dissertation, we chose to model survival time as a function of a single *continuous* predictor, defined by the first principal component of the metagene's expression (or by a single vector of gene expression) throughout the entire cohort. This departure from the standard approach has the disadvantage of making it impossible to calculate a Kaplan-Meier estimator for the model—therefore lacking a visual representation and making it harder to assess the validity of the proportional hazards assumption. However, it has the advantage of using a predictor variable that faithfully mirrors the patterns of expression captured by the metagene and of not requiring the artificial stratification of the cohort in (balanced or not) groups of differential prognosis.

Whenever applicable, association of metagene expression with outcome was thusly computed by fitting a proportional hazards regression model between predictor and dependent variables with the `coxph` function of the Bioconductor survival package.

Microarray datasets

To assess the extent of the prognostic signals in human cancer transcriptomes, a total of 114 public datasets of cancer gene expression profiles spanning 22 types of cancer were downloaded from public repositories, manually curated and pre-processed for downstream analysis. Sources for datasets include the Gene Expression Omnibus,[206] InSilicoDB[207] and the TCGA Research Network site.[208] Individual datasets are described in Table 3. The OS, DFS, DSS and DMFS columns refer to, respectively, the number of patients for which overall survival, disease free survival, disease-specific survival and distant metastasis-free survival were recorded in each dataset.

[206] `http://www.ncbi.nlm.nih.gov/geo/` (Edgar et al., 2002)

[207] Taminau et al., 2011

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

Table 3: Datasets used in this dissertation.

Dataset	Cancer	Platform	Normalization	Patients	Genes	OS	DFS	DSS	DMFS
BLCA	bladder	Agilent G4502A	RMA	122	20501	119	—	—	—
BRCA	breast	Agilent G4502A	RMA	849	20501	574	—	—	—
CESC	cervical	Agilent G4502A	RMA	97	20501	39	—	—	—
COAD	colon	Agilent G4502A	RMA	192	20501	192	—	—	—
GBM	glioblastoma	Agilent G4502A	RMA	169	20501	167	—	—	—
HNSC	head and neck	Agilent G4502A	RMA	303	20501	302	—	—	—
KICH	kidney	Agilent G4502A	RMA	66	20501	65	—	—	—
KIRC	kidney	Agilent G4502A	RMA	480	20501	480	—	—	—
KIRP	kidney	Agilent G4502A	RMA	76	20501	72	—	—	—
LAML	leukemia	Agilent G4502A	RMA	183	20501	172	—	—	—
LGG	glioma	Agilent G4502A	RMA	205	20501	205	—	—	—
LIHC	liver	Agilent G4502A	RMA	34	20501	34	—	—	—
LUAD	lung	Agilent G4502A	RMA	355	20501	329	—	—	—
LUSC	lung	Agilent G4502A	RMA	259	20501	251	—	—	—
OV	ovary	Agilent G4502A	RMA	266	20501	264	—	—	—
PAAD	pancreas	Agilent G4502A	RMA	31	20501	30	—	—	—
PRAD	prostate	Agilent G4502A	RMA	140	20501	126	—	—	—
READ	rectum	Agilent G4502A	RMA	72	20501	72	—	—	—
SKCM	skin	Agilent G4502A	RMA	267	20501	242	—	—	—
STAD	stomach	Agilent G4502A	RMA	57	20501	57	—	—	—
THCA	thyroid	Agilent G4502A	RMA	414	20501	354	—	—	—
UCEC	uterine corpus	Agilent G4502A	RMA	370	20501	369	—	—	—
metabric-discovery-set	breast	Illumina HT-12 V3	quantile normalization	997	19628	980	—	980	—
metabric-validation-set	breast	Illumina HT-12 V3	quantile normalization	995	19628	991	—	991	—
NKI	breast	Agilent HU25K	—	295	12937	295	295	—	295
GSE10846	lymphoma	Affy HG-U133 Plus 2.0	MAS5.0	420	20185	414	—	—	—
GSE2658	myeloma	Affy HG-U133 Plus 2.0	MAS5.0	559	19944	—	—	559	—
GSE39582	colon	Affy HG-U133 Plus 2.0	RMA	566	19945	—	557	—	—
GSE44001	cervical-cancer	Illumina HT-12 V4	quantile normalization	300	20104	—	300	—	—
GSE10645	prostate	Illumina DASL human cancer panel	cyclic loess	596	497	596	—	596	—
GSE7390	breast	Affy HG-U133A	MAS5.0	198	12495	198	198	—	198
GSE16391	breast	Affy HG-U133 Plus 2.0	FRMA	55	19944	—	48	—	—
GSE5327	breast	Affy HG-U133A	MAS5.0	58	12495	—	—	—	58
GSE2034	breast	Affy HG-U133A	MAS5.0	286	12495	—	—	—	286

Continued on next page

Dataset	Cancer	Platform	Normalization	Patients	Genes	OS	DFS	DSS	DMFS
GSE2990	breast	Affy HG-U133A	RMA	189	12495	—	187	—	179
GSE6532-GPL570	breast	Affy HG-U133 Plus 2.0	RMA	87	19944	—	87	—	87
GSE6532-GPL96	breast	Affy HG-U133A	RMA	327	12495	—	306	—	293
GSE6532-GPL97	breast	Affy HG-U133B	RMA	327	9733	—	306	—	293
GSE1456-GPL96	breast	Affy HG-U133A	FRMA	159	12495	159	159	159	—
GSE1456-GPL97	breast	Affy HG-U133B	FRMA	159	9733	159	159	159	—
GSE4922-GPL96	breast	Affy HG-U133A	MAS5.0	289	12495	—	249	—	—
GSE4922-GPL97	breast	Affy HG-U133B	MAS5.0	289	9733	—	249	—	—
GSE9195	breast	Affy HG-U133 Plus 2.0	RMA	77	19944	—	77	—	77
GSE12093	breast	Affy HG-U133A	MAS5.0	136	12495	—	136	—	—
GSE20685	breast	Affy HG-U133 Plus 2.0	quantile normalization	327	18922	—	—	—	327
GSE3494-GPL96	breast	Affy HG-U133A	MAS5.0	251	12495	—	—	236	—
GSE3494-GPL97	breast	Affy HG-U133B	MAS5.0	251	9733	—	—	236	—
GSE1379	breast	Arcturus 22k	median subtraction $\log_2$ ratios gene-wise	60	11558	—	60	—	—
GSE11121	breast	Affy HG-U133A	MAS5.0	200	12495	—	—	—	200
GSE12276	breast	Affy HG-U133 Plus 2.0	MAS5.0	204	19944	—	—	—	204
GSE7378	breast	Affy HT-HG-U133 Plus 2.0	RMA	54	12659	—	54	—	—
GSE1378	breast	Arcturus 22k	median subtraction $\log_2$ ratios gene-wise	60	11558	—	60	—	—
GSE3143	breast	Affy HG-U95av2	MAS5.0	158	7114	158	—	—	—
GSE19615	breast	Affy HG-U133 Plus 2.0	dChip invariant method	115	19944	—	—	—	115
GSE9893	breast	custom array	cross-array median normalization	155	15762	155	—	—	—
GSE19536	breast	Agilent G4112F	quantile normalization	100	19425	98	—	98	—
GSE25307	breast	Swegene H v2.1.1 55k	lowess	577	8428	547	—	—	—
GSE18229-GPL1390	breast	custom array	lowess	199	14457	160	160	—	—
GSE18229-GPL887	breast	custom array	lowess	94	16579	53	53	—	—
GSE12417	leukemia	Affy HG-U133 Plus 2.0	VSN	163	12714	163	—	—	—
GSE12945	colon	Affy HG-U133A	FRMA	62	12714	62	51	—	—
GSE14333	colon	Affy HG-U133 Plus 2.0	FRMA	290	20027	—	226	—	—
GSE14764	ovary	Affy HG-U133A	FRMA	80	12714	80	—	—	—
GSE14814	lung	Affy HG-U133A	FRMA	90	12714	90	—	90	—
GSE16102	bone	Affy HG-U133A	FRMA	57	12714	34	—	—	—
GSE16131	lymphoma	Affy HG-U133A	FRMA	184	12714	180	—	—	—
GSE16581	brain	Affy HG-U133 Plus 2.0	FRMA	68	20027	67	—	—	—

Continued on next page

Dataset	Cancer	Platform	Normalization	Patients	Genes	OS	DFS	DSS	DMFS
GSE17538	colon	Affy HG-U133 Plus 2.0	FRMA	238	20027	232	232	226	—
GSE19188	lung	Affy HG-U133 Plus 2.0	FRMA	91	20027	82	—	—	—
GSE19234	skin	Affy HG-U133 Plus 2.0	FRMA	44	20027	44	—	—	—
GSE19829	ovary	Affy HG-U133 Plus 2.0	FRMA	28	20027	28	—	—	—
GSE22762	leukemia	Affy HG-U133 Plus 2.0	FRMA	107	20027	107	—	—	—
GSE23501	lymphoma	Affy HG-U133 Plus 2.0	FRMA	69	20027	69	69	—	—
GSE23554	ovary	Affy HG-U133A	FRMA	28	12714	28	—	—	—
GSE27020	larynx	Affy HG-U133A	FRMA	109	12714	—	109	—	—
GSE31595	colon	Affy HG-U133 Plus 2.0	FRMA	37	20027	—	37	—	—
GSE31684	bladder	Affy HG-U133 Plus 2.0	GCRMA	93	19851	93	—	93	—
GSE4271	brain	Affy HG-U133A	FRMA	100	12714	77	—	—	—
GSE4412	brain	Affy HG-U133A	FRMA	85	12714	85	—	—	—
GSE4475	lymphoma	Affy HG-U133A	RMA	221	12714	159	—	—	—
GSE7696	brain	Affy HG-U133 Plus 2.0	RMA	80	19851	80	—	—	—
GSE8894	lung	Affy HG-U133 Plus 2.0	FRMA	138	20027	—	138	—	—
GSE9899	ovary	Affy HG-U133 Plus 2.0	FRMA	295	20027	289	286	—	—
GSE31210	lung	Affy HG-U133 Plus 2.0	FRMA	246	20027	226	226	—	—
GSE33113	colon	Affy HG-U133 Plus 2.0	FRMA	90	20027	—	90	—	—
GSE3141	lung	Affy HG-U133 Plus 2.0	FRMA	111	20027	111	—	—	—
GSE13507	bladder	Illumina human-6 v2.0	quantile normalization	165	20045	165	—	165	—
GSE13876	ovary	custom array	quantile normalization	415	13796	—	—	415	—
oberthuer2006	neuroblastoma	custom array	VSM	251	9878	—	251	251	—
GSE3149	ovary	Affy HG-U133A	FRMA	122	12913	—	—	122	—
GSE32062	ovary	Affy HG-U133 Plus 2.0	FRMA	260	19596	260	260	—	—
GSE11318	lymphoma	Affy HG-U133 Plus 2.0	FRMA	203	20026	200	—	—	—
GSE10143	liver	custom array	cubic spline algorithm	80	6090	80	—	—	—
GSE16432-GPL10105	leukemia	custom array	gene mean centering	45	13188	45	—	—	—
GSE16432-GPL10106	leukemia	custom array	gene mean centering	38	13493	38	—	—	—
GSE16432-GPL10107	leukemia	custom array	gene mean centering	82	13084	81	—	—	—
GSE16432-GPL8650	leukemia	custom array	gene mean centering	25	8144	25	—	—	—
GSE16432-GPL8651	leukemia	custom array	gene mean centering	45	12779	45	—	—	—
GSE16432-GPL8652	leukemia	custom array	gene mean centering	50	13070	50	—	—	—
GSE16432-GPL8653	leukemia	custom array	gene mean centering	94	13659	93	—	—	—
GSE16432-GPL8654	leukemia	custom array	gene mean centering	49	12358	49	—	—	—
GSE32894	bladder	Illumina HT-12 V3	quantile normalization	308	15515	—	—	248	—
GSE17674	bone	Affy HG-U133 Plus 2.0	FRMA	44	20027	32	44	—	—
GSE30929	liposarcoma	Affy HG-U133A	FRMA	140	12714	—	—	—	140

Continued on next page

Dataset	Cancer	Platform	Normalization	Patients	Genes	OS	DFS	DSS	DMFS
GSE28026	teratoid/rhabdoid	Affy HG-U133 Plus 2.0	FRMA	18	20027	17	—	—	—
GSE14520	liver	Affy HG-U133A 2.0	FRMA	249	12569	242	242	—	—
GSE16560	prostate	Human 6k gene panel for DASL	cubic spline algorithm	281	6090	281	—	—	—
GSE37745	lung	Affy HG-U133 Plus 2.0	FRMA	196	20184	196	—	—	—
GSE41258	colon	Affy HG-U133A	FRMA	182	12701	—	—	252	—
GSE19915-GPL3883	bladder	Swegene Human 27k RAP UniGene188	lowess	47	10109	—	—	47	—
GSE19915-GPL5186	bladder	Swegene H v3.0.1 35k	lowess	20	14606	—	—	20	—
GSE13041-GPL96	brain	Affy HG-U133A	RMA	191	12496	—	—	191	—
GSE13041-GPL8300	brain	Affy HG-U95Av2	RMA	49	8657	—	—	49	—
GSE13041-GPL570	brain	Affy HG-U133 Plus 2.0	RMA	27	19851	—	—	27	—

Results

This chapter is comprised of three sections. The first consists of a published article, entitled "A general method to derive robust organ-specific gene expression-based differentiation indices: application to thyroid cancer diagnostic."[209] The results exposed in this manuscript are developed in the Discussion chapter, under the section *Differentiation and proliferation signatures in cancer diagnostic*.

[209] Tomás et al., 2012

The second section, entitled "The extent of prognostic signals in the cancer transcriptomes," details the results of a systematic analysis of 114 microarray datasets of cancers affecting 19 different tissue types, aiming at quantifying the range and the nature of the signals linked with differential survival in neoplastic expression profiles. These results are elaborated upon in the corresponding section of the Discussion chapter.

The remaining section reports other publications to which the author contributed during this dissertation.

Differentiation and proliferation signatures in cancer diagnostic

Executive Summary

Proliferation and differentiation are fundamental processes of multicellular life. While quantitative assessments of cellular proliferation in the clinical setting are routinely achieved by methods such as bromodeoxyuridine incorporation, Ki-67 or proliferating cell nuclear antigen (PCNA) immunostaining, no method for objective quantification of tissue differentiation currently exists. This is because cellular differentiation is a complex function of morphological, physiological and molecular criteria. Therefore, assessments of differentiation in the clinical setting remain the object of subjective appreciation of clinical pathologists.

Because cancer progression is concomitant with a loss of tissue differentiation, a quantitative method for assessing differentiation could be useful in cancer diagnostics as a complementary approach to traditional staging systems. Thyroid carcinogenesis is a model particularly suited to test this assertion, as it is characterized by a slow progression from a collection of well-differentiated benign neoplasms to some poorly differentiated aggressive carcinomas, and finally to one of the most aggressive human cancers, the anaplastic thyroid carcinoma. Two particularly challenging histopathological diagnostics in thyroid carcinogenesis are the distinction between follicular adenomas (FAs) and follicular thyroid carcinomas (FTCs); and between follicular variants of papillary thyroid carcinomas (FVPTCs) and

papillary thyroid carcinomas (CPTCs). While the former variants are usually encapsulated and do not require surgical treatment, the latter neoplasias are treated with thyroidectomies and radioactive iodine. A quantitative method for mapping thyroid carcinogenesis dedifferentiation could be of potential use to prevent unnecessary surgery in patients presenting with benign neoplasms, when assayed with fine-needle aspiration cytology.

We present a fully automated, agnostic, and objective method to quantify tissue differentiation from expression profiles of healthy tissue biospecimens. This feature selection method consists of: (*a*) quantifying and ranking transcript abundance in a representative panel of human tissues; and (*b*) retaining, for each tissue in the panel, the genes whose expression is measured above a predefined ranking in the tissue of choice, and beyond a predefined ranking in the remaining tissues profiled in the panel. This method, when applied to a reference collection of 16 tissue types profiled with RNA-seq, has yielded a thyroid-specific biomarker of eight specific genes.

We used this thyroid differentiation biomarker, alongside with a proliferation differentiation signature defined elsewhere,[210] to provide quantitative assessments of differentiation and proliferation in a collection of expression profiles of neoplasias from a thyroid carcinogenesis model. We showed that the differentiation and proliferation biomarkers are negatively correlated in neoplasias of increasing aggressiveness, as consistent with an inverse pattern of decreased differentiation and increased proliferation in cancer pathogenesis. In order to test whether this inverse relationship between the two biomarkers would hold in different physiological conditions, we measured their expression in expression profiles from a TSH stimulation time course of thyrocytes in primary culture. TSH is a known inducer of both differentiation and proliferation in normal thyrocytes and, congruently, both indices showed a positive correlation in this experiment. The two biomarkers thus capture independent biological information.

[210] Venet et al., 2011

Furthermore, we showed that the thyroid differentiation biomarker could accurately differentiate between FAs and FTCs; and between FVPTCs and CPTCs—which sit along two distinct gene expression dedifferentiation continuums from the normal thyrocyte. The discriminatory performance of this diagnostic classifier, measured by its area under the curve, was not statistically different than those of two classifiers trained in the entire expression profile space of each of the pairs of labeled samples. The thyroid differentiation biomarker, devised from expression profiles of healthy tissue types, can thus quantitatively map distinct dedifferentation routes in the thyroid carcinogenesis model; and captures all diagnostic information present in the transcriptomes of challenging histological subtypes of thyroid neoplasias.

The approach here described is a proof-of-concept result. We show that differentiation, a multi-layered feature of multicellular life, can be brought to quantitative terms by selecting for tissue-specific transcriptional features. We show that a thyroid differentiation biomarker thusly defined can be used to track thyrocyte-derived cancer progression in a manner that is independent of proliferation; and that this biomarker can discriminate between clinically challenging pathological diagnoses in a thyroid carcino-

genesis model. The specificity of such differentiation biomarkers is largely dependent of the panel of healthy tissue types used to derive them. The use of resource databases such as the Genotype-Tissue Expression project, GTEx,[211] which reports the results of the transcriptional variation within a biobank of more than 50 human tissue types, could be used to greatly refine the specificity of differentiation biomarkers. The use of such differentiation biomarkers could then be used to dissect expression profiles of more complex cancer models, such as breast cancer, notably through the mining of the large collection of neoplastic expression profiles reported in The Cancer Genome Atlas, TCGA.[212]

[211] Lonsdale et al., 2013

[212] The Cancer Genome Atlas Research Network et al., 2013

Article

See next page.

Oncogene (2012) 31, 4490–4498

www.nature.com/onc

ONCOGENOMICS

A general method to derive robust organ-specific gene expression-based differentiation indices: application to thyroid cancer diagnostic

G Tomás[1], M Tarabichi[1], D Gacquer[1], A Hébrant[1], G Dom[1], JE Dumont[1], X Keutgen[2], TJ Fahey III[2], C Maenhaut[1,3] and V Detours[1]

[1]IRIBHM, Université Libre de Bruxelles (ULB), Campus Erasme, Brussels, Belgium; [2]Department of Surgery, Division of Endocrine Surgery, Weill Cornell Medical College, New York, NY, USA and [3]WelBio, Université Libre de Bruxelles (U.L.B.), Campus Erasme, Brussels, Belgium

Differentiation is central to development, while dedifferentiation is central to cancer progression. Hence, a quantitative assessment of differentiation would be most useful. We propose an unbiased method to derive organ-specific differentiation indices from gene expression data and demonstrate its usefulness in thyroid cancer diagnosis. We derived a list of thyroid-specific genes by selecting automatically those genes that are expressed at higher level in the thyroid than in any other organ in a normal tissue's genome-wide gene expression compendium. The thyroid index of a tissue was defined as the median expression of these thyroid-specific genes in that tissue. As expected, the thyroid index was inversely correlated with meta-PCNA, a proliferation metagene, across a wide range of thyroid tumors. By contrast, the two indices were positively correlated in a time course of thyroid-stimulating hormone (TSH) activation of primary thyrocytes. Thus, the thyroid index captures biological information not integrated by proliferation rates. The differential diagnostic of follicular thyroid adenomas and follicular thyroid carcinoma is a notorious challenge for pathologists. The thyroid index discriminated them as accurately as did machine-learning classifiers trained on the genome-wide cancer data. Hence, although it was established exclusively from normal tissue data, the thyroid index integrates the relevant diagnostic information contained in tumoral transcriptomes. Similar results were obtained for the classification of the follicular vs classical variants of papillary thyroid cancers, that is, tumors dedifferentiating along a different route. The automated procedures demonstrated in the thyroid are applicable to other organs.
Oncogene (2012) **31,** 4490–4498; doi:10.1038/onc.2011.626; published online 23 January 2012

Keywords: gene expression; differentiation; proliferation; cancer; thyroid

Correspondence: Dr V Detours, Institute of Interdisciplinary Research (IRIBHM), Universite Libre de Bruxelles (U.L.B.), 808 Route de Lennik, Brussels, Brussels B-1070, Belgium.
E-mail: vdetours@ulb.ac.be
Received 2 June 2011; revised 4 November 2011; accepted 3 December 2011; published online 23 January 2012

Introduction

Differentiation and proliferation are the fundamental processes of multicellular life. Cell proliferation is routinely measured with objective quantitative methods, such as bromodeoxyuridine incorporation, Ki-67 or proliferating cell nuclear antigen (PCNA) immunostaining. However, the differentiation state of a cell type manifests itself in a wide range of parameters—many of them qualitative—that include cellular- and tissue-level morphologies, molecular state and function. It is unclear how typical organ-specific differentiation markers integrate these phenomena and to what extent they reflect the overall degree of differentiation of cells and tissues. For example, thyroglobulin is a canonical thyroid marker and is essential in thyroid hormone synthesis. FRTL-5 cells express thyroglobulin, trap iodine and grow in response to thyroid stimulation hormone (TSH), but they are depolarized, that is, they lack a basic morphological feature of differentiated thyroid cells. Conversely, FRT cells are polarized, but do not express thyroglobulin and other thyroid-specific properties (Zurzolo *et al*., 1991). Moreover, the evaluation of many differentiation features rest on subjective and qualitative histological observations. Thus, there is currently no method to quantify differentiation objectively. Herein, we propose a procedure to derive quantitative multi-gene gene expression markers for organ-specific differentiation. It is completely automatic and agnostic regarding the biology of organs and which aspects of their differentiation are important.

The thyroid is an interesting organ to assess our method because a single cell type, the follicular thyroid cell, can produce a range of benign and malignant tumors with a range of underlying genetic alterations leading to a range of dedifferentiated phenotypes (Kondo *et al*., 2006); those include hypofunctioning follicular adenomas (FTAs), which are benign encapsulated tumors, and the malignant carcinomas. These can be further subdivided into follicular or papillary carcinomas. They are still partly differentiated, but may both evolve into poorly differentiated carcinoma or into totally dedifferentiated anaplastic carcinoma. Papillary carcinomas further fall in a number of histological subtypes. The most frequent are the classical

and the follicular variants. The latter displays some aspects of the typical follicular morphology of normal thyroid tissues.

The differential diagnosis of FTAs and carcinomas is a major challenge for pathologists—and incidentally for the thyroid-differentiation index derived with our method. Between 4 and 7% of the general population will develop a palpable thyroid nodule (Hegedüs, 2004). Of these, 80% are benign hyperplastic nodules, 10–15% are benign follicular neoplasms and 5% are cancers (Hegedüs, 2004). Fine needle aspirate examination is currently the most direct and sensitive method to detect malignancy. However, it is inconclusive in approximately 20% of nodules, resulting in the need for potentially avoidable diagnostic thyroidectomy. In addition, a second opinion by an independent pathologist leads to a contradictory conclusion in 30–60% of the cases (Baloch *et al*., 2001; Hegedüs, 2004; Clary *et al*., 2005). Microarray studies suggest that there is no clear-cut boundary between FTAs and carcinomas (Lubitz *et al*., 2005).

Because tumor proliferation is in general inversely correlated with differentiation, there is a real possibility that any quantitative measure of differentiation is the trivial mirror image of proliferation measures. However, this inverse correlation does not hold in all biological contexts. For example, the TSH promotes both the proliferation and the differentiation of thyroid cells when combined with insulin. Thus, the thyroid also provides an alternative system to assess the relatedness of our differentiation index with proliferation.

Results

The thyroid index and the meta-PCNA index: unbiased measures of thyroid tissue differentiation and proliferation from gene expression data

We propose an unbiased procedure to compute an organ-specific differentiation index from mRNA expression data, using the thyroid as a test case. It proceeds in three steps. Step #1 selects genes specific of the organ of interest from RNA sequencing (RNA-seq) data. RNA-seq estimates mRNA expression with read counts normalized according to the transcript length, a measure that reflects more accurately the absolute transcription levels than hybridization-based microarray (Wang *et al*., 2009). Step #2, although optional, optimizes the probe selection for these genes in the gene expression platform in which differentiation is to be measured. Finally, step #3 uses the probes/genes derived in previous steps to compute a differentiation index across samples of a data set of interest. An application to the thyroid is described below. Computational details are available in the Materials and methods section.

First, we extracted a set of genes with thyroid-specific expression in BodyMap 2.0 (http://www.ncbi.nlm.nih.gov/geo), a collection of 16 RNA-seq profiles of healthy human organs, that is, these genes were among the 1000 most expressed genes in the thyroid, but not among the 5000 most expressed genes in the non-thyroid organs. Eight genes, listed in Table 1, fulfiled this criterion. Four were canonical thyroid follicular genes: forkhead box protein E1 (FOXE1, a.k.a. TTF2), thyroglobulin, thyroperoxydase and the TSH receptor. Iodotyrosine deiodinase mutations cause congenital hypothyroidism (Moreno *et al*., 2008). Interestingly, solute carrier family 26, member 7 (SLC26A7) and cellular retinoic acid binding protein 1 (CRABP1) were thus far unknown thyroid markers.

The thyroid-specific expression of seven of these genes was confirmed in two independent normal organ gene expression compendia based on massively parallel signature sequencing (Supplementary Figure S1; Jongeneel *et al*., 2005), and Affymetrix arrays (Affymetrix, Santa Clara, CA, USA) (Supplementary Figure S2; Ge *et al*., 2005). The last gene, parathyroid hormone, most probably results from the erroneous inclusion of parathyroid tissues in BodyMap 2.0 thyroid sample. Its high thyroid expression is also observed in the compendium of Jongeneel *et al*. (2005), but not in the compendium of Ge *et al*. (2005).

On most microarray platforms, several probes are available to measure the expression of a given gene. However, because of technical shortcomings, such as annotation errors and non-specificity, not all of them are necessarily accurate. In step #2, we selected, for each of the eight genes, the probe that maximizes the thyroid-specific signal in a normal organ-expression compendium obtained using the platform of interest. Genes with no probe delivering a thyroid-specific signal were eliminated. We applied this procedure to the Affymetrix U95av2 and U133v2 platforms using the compendia of Su *et al*. (2002) and Roth *et al*. (2006), respectively. Parathyroid hormone was eliminated

Table 1 Thyroid index genes

Symbol	*Description*	*Probe u133v2*	*Probe u95av2*
CRABP1	Cellular retinoic acid-binding protein 1	205350_at	543_g_at
FOXE1	Forkhead box E1 (thyroid transcription factor 2)	206912_at	36370_at
IYD	Iodotyrosine deiodinase	231070_at	NA (2)
PTH	Parathyroid hormone	NA (1)	708_at
SLC26A7	Solute carrier family 26, member 7	239006_at	NA (2)
TG	Thyroglobulin	203673_at	39042_at
TPO	Thyroid peroxidase	210342_s_at	35928_at
TSHR	Thyroid-stimulating hormone receptor	210055_at	32471_at

(1) None of the probes targeting the gene is thyroid-specific in platform-specific data. (2) The gene is not profiled in the platform set.

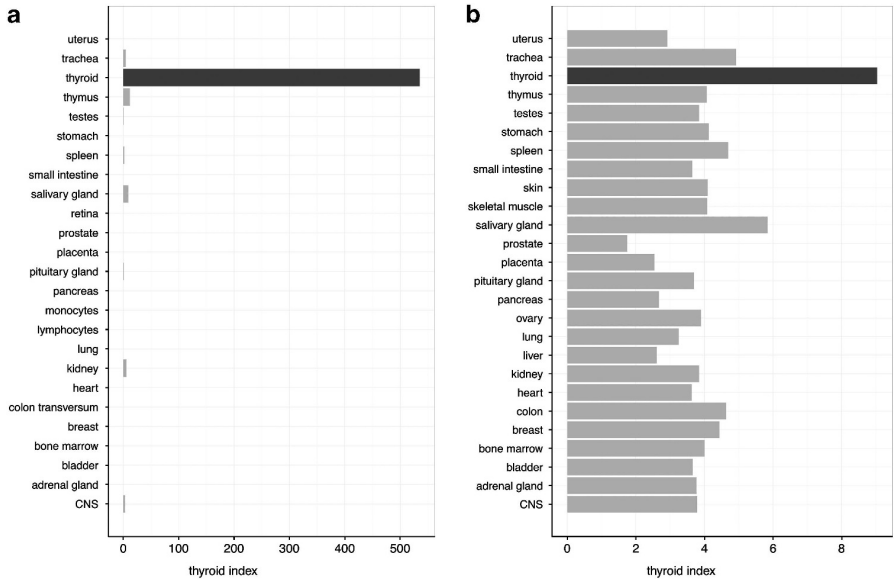

Figure 1 The thyroid index singles out thyroid tissue in two independent normal tissue compendia. The tissues from the compendia of (**a**) Jongeneel *et al*. (2005) and (**b**) Ge *et al*. (2005) were not used to derive the gene list underlying the thyroid index. Organs are aligned along the *y*-axis. The *x*-axis represents the thyroid index ($\log_2$ scale in **b**). The thyroid, depicted with the dark bar, stands out among other organs.

following our optimization with the U133v2 compendia. Thus, in addition to probe optimization, step #2 factored-in extra information that helps refine the list of measurable thyroid genes. However, parathyroid hormone was not eliminated in the U95av2 optimization. We included it in all subsequent analyses conducted on this platform because our goal is to assess an agnostic, automated method.

Finally, given the gene expression profile of a tissue in a given platform we may compute its thyroid index as the median expression of the probes selected in step #2. Figure 1 depicts this index from two normal human organ compendia not used to select the eight genes (Ge *et al*., 2005; Jongeneel *et al*., 2005). The thyroid sample stands out among the other tissues.

We defined meta-PCNA as the 1% of genes that are most positively correlated with PCNA expression across a gene expression compendium of normal human organs (Venet *et al*., 2011). In plain language, meta-PCNA genes are consistently expressed when PCNA is expressed in normal tissues and consistently repressed when PCNA is repressed. Meta-PCNA genes include many canonical proliferation markers, including *MCM2, MKI67, TOP2A* and, of course, *PCNA*. Similar to the thyroid index, we defined the meta-PCNA index of a tissue as the median expression of the meta-PCNA genes.

The thyroid and meta-PCNA indices are negatively correlated in thyroid cancers, but positively correlated in an in vitro *TSH stimulation time course*

We generated full genome expression profiles for 11 anaplastic thyroid cancers (ATC), 49 papillary thyroid cancers (PTC) and 45 healthy tissues adjacent to 45 of the PTCs (see Materials and methods), and compared the thyroid and meta-PCNA indices across them. The thyroid index is lower in ATCs than in PTCs ($P=10^{-9}$) and lower in PTC than in normal tissues ($P=10^{-15}$). Conversely, the meta-PCNA index is higher in ATC than in PTC ($P=10^{-6}$) and higher in PTC than in normal tissues ($P=10^{-4}$). Overall, the two indices are anti-correlated (Spearman's $\rho=-0.71$, $P<10^{-16}$; Figure 2a) in agreement with the general notion that dedifferentiated anaplastic tumors proliferate more and are more aggressive than the more differentiated PTCs—ATCs have a 5-year survival rate $<10\%$ (Kebebew *et al*., 2005) and PTCs $>90\%$ (Kondo *et al*., 2006).

To check that this relation holds in an alternative data set and with tissues displaying less-contrasted dedifferentiation phenotypes, we compiled healthy tissues and differentiated tumor-expression profiles from earlier publications (Aldred *et al*., 2004; Finley *et al*., 2004). A strong anti-correlation of our indices is also observed across this panel ($\rho=-0.52$, $P=10^{-5}$; Figure 2b).

Are the two indices redundant or are they capturing independent biological information? We calculated the thyroid and meta-PCNA indices in expression profiles from a TSH stimulation time course of thyroid cells in primary cultures (van Staveren *et al*., 2006). The indices were positively correlated ($\rho=0.73$, $P=0.01$, Figure 2c), they both increased with stimulation time, in agreement with the fact that TSH promotes both follicular cell proliferation and differentiation.

In conclusion, the thyroid and meta-PCNA indices put on a quantitative ground the notion that differ-

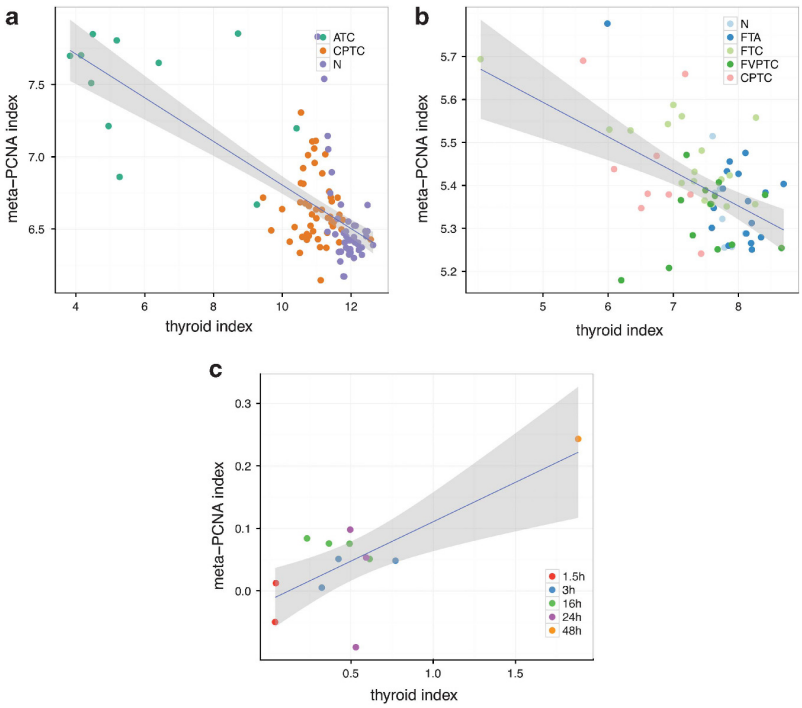

Figure 2 Correlation between the thyroid index and meta-PCNA. Overall, the two indices are anti-correlated in tumors (**a** and **b**), but not in the TSH stimulation time course (**c**). (**a**) Anaplastic thyroid carcinoma (ATC), papillary thyroid carcinomas (CPTC) and normal tissues (N). Note the absence of a clear-cut boundary between PTCs and normal tissues. However, as shown in Supplementary figure S3, tumors have lower thyroid index and higher meta-PCNA index than their patient-matched contralateral normal tissues. (**b**) Data from Aldred *et al*. (2004) and Finley *et al*. (2004), with CPTC and FVPTC, FTA and carcinoma (FTC), and normal tissues (N). (**c**) Thyroid index vs meta-PCNA in the TSH-stimulation time course. Individual points at a given time point represent primary cells extracted from different subjects. Because the platform had only ~4000 genes, the overlap with the 8-gene signature included only 1 gene. Therefore, we used the alternative 72-gene signature of section 'Robustness of the organ-specific gene selection' in panel (**c**). Related data are available in Supplementary tables S1 to S3.

entiation inversely correlates with proliferation in thyroid neoplasms. Yet, these indices do measure biologically independent parameters.

Follicular and papillary thyroid tumors dedifferentiate along distinct routes in the gene expression continuum

A single cell type, the thyroid follicular cell, may give rise to distinct tumoral phenotypes, including PTCs and tumors of the follicular family. The latter include FTA, which may dedifferentiate further into follicular carcinomas (FTC). Most PTCs fall into the follicular variant (FVPTC) category, which retain some follicular morphology, or into the classical variant (CPTC) category, which have lost the follicular morphology. Before assessing the thyroid index in these tumors, we compared their global expression profiles.

We compiled from two studies (Aldred *et al*., 2004; Finley *et al*., 2004) the expression profiles of 7 healthy thyroid tissues, 17 FTAs, 18 FTCs, 13 FVPTCs and 9 CPTCs, and ran a multidimentional scaling analysis (Figure 3). Multidimensional scaling collapses the high-dimensional full-genome gene expression space into two dimensions while preserving the similarity distances between pairs of samples. Hence, samples lying close on the multidimensional scaling plot have similar gene expression profiles.

The samples in Figure 3 are clustered according to tissue types, not the study of origin. Papillary and follicular tumors both radiated from the normal samples, but in opposite directions, that is, they had clearly distinct overall molecular phenotypes. FTAs are closer to normal samples than FTCs. Likewise, FVPTCs are closer to normal tissue than CPTC. As already noted (Lubitz *et al*., 2005), there is an overlap between FTA and FTC suggesting a continuum rather than discrete tumor categories. We show here that the same observation holds for FVPTCs and CPTCs.

Thus, follicular and papillary thyroid tumors dedifferentiate along very different directions in the gene expression continuums. Yet, we show in the next section that a single quantity, the thyroid index, is a useful differentiation marker for both of them.

The thyroid index discriminates follicular adenomas from carcinomas and the classical from the follicular variants of papillary carcinomas

Can the thyroid index, which was derived exclusively from normal tissue data, discriminate tumor samples according to their histological types? We first focused on

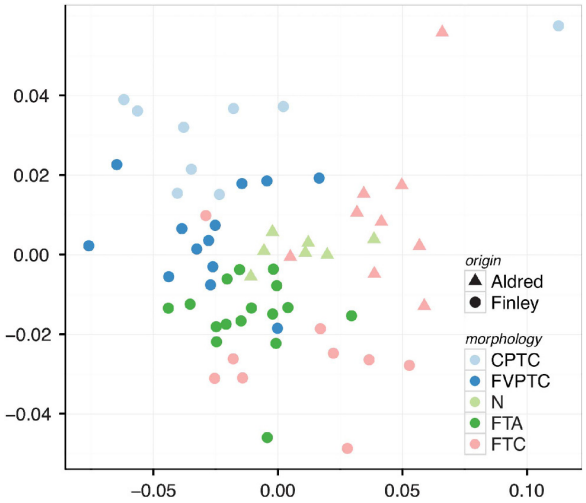

Figure 3 Follicular and papillary tumors dedifferentiate along distinct routes in the gene expression continuums. This multi-dimensional scaling plot reduces the high dimensional full-genome expression data into two dimensions. The 'stress' of the transformation, that is, the average distortion between the pair-wise distances between samples in the reconstructed 2D space and in the actual gene expression space is 11%. Units in the 2D space are arbitrary. Samples cluster by tissue types rather than that of lab of origin. Normal tissues are in the center of the plot with follicular and papillary tumors spread on opposite sides. The more dedifferentiated a tumor, the more distant it tends to be from normal tissues. Given the diagnosis uncertainties with these tumors and limits in the 2D reconstruction, some discrepancies are expected, for example, one FTC stands out between CPTCs and FVPTCs.

the classification of CPTCs and FVPTCs using the data set presented in the previous section. The thyroid index is significantly higher in FVPTCs than in CPTCs ($P=0.003$, Figure 4a), suggesting that FVPTCs are more differentiated than CPTCs. We further assessed its discriminatory performance by computing the area under the receiver operating characteristics (AUC), a statistic that integrates specificity and sensitivity, both shown in Figure 4b. The AUC was 0.86 (Figure 4b). We verified that this AUC was not explained by chance alone, by rerunning the entire AUC calculation on 10 000 indices on which class labels were assigned randomly, yielding $P=0.0003$. The thyroid index may non-specifically capture a diagnostic signal omnipresent in the transcriptome as was observed for most prognostic signatures in breast cancer (Venet *et al*., 2011). To rule out this possibility, we recomputed AUC with 100 000 signatures made of eight probes selected at random on Affymetrix U133v2 arrays, yielding $P=0.005$ (Supplementary Figure S4).

The differential diagnosis of FTAs vs FTCs is a notorious cytopathological challenge. Using the data from the previous section, the thyroid index obtained was higher in FTA than in FTC ($P=0.0008$, Figure 4d) and had an AUC of 0.82 (class label permutations, $P=0.0001$; random probes control, $P=0.02$; Figure 4e; Supplementary Figure S4). Taken together, these results demonstrate that the thyroid index is discriminatory for tumors following distinct dedifferentiation routes.

Does the thyroid index capture all the discriminatory information present in the genome-wide expression profiles? Better classifiers may be discovered by applying supervised machine-learning algorithms directly to the cancer data. Starting with all the genes present on the microarrays, we searched for optimal classifiers with two supervised algorithms, linear kernel support vector machine (SVM) and random forests (RF), both of them combined with a feature-selection step (see Materials and methods). The selections of features and algorithm parameters were nested in an inner cross-validation loop to preclude any possibility of parameters and feature-selection biases. None of the classification strategies produced a better classifier than the thyroid index for the FVPTC vs CPTC and FTA vs FTC classification tasks (Figures 4c and f).

The meta-PCNA index was inversely correlated to the thyroid index in the above panel of tumors ($P=2\times10^{-5}$) and was significantly higher in FTC than FTA ($P=0.01$) and in CPTC than FVPTC ($P=0.05$). Yet, meta-PCNA was not as discriminant as the thyroid index according to the AUC metrics (not shown).

In conclusion, the thyroid index integrates most transcriptional information relevant to the differential diagnosis of thyroid tumors, and does so for tumors arising from different dedifferentiation pathways.

Robustness of the organ-specific gene selection

We defined thyroid-specific genes as among the 1000 most expressed in the thyroid and not among the 5000 most expressed in any of the 15 non-thyroid organs profiled in the BodyMap 2.0. Hence, we choose thyroid genes that have high expression in the thyroid, that is, discarded gene with important thyroid-specific function, but moderate expression. For example, 13 136 genes have higher thyroid expression than the sodium iodide symporter (SLC5A5, a.k.a. NIS) in BodyMap 2.0. In addition, some genes are characteristic of the thyroid, but may also have a role in another organ. For example, paired box gene 8 (*PAX8*) is a transcription factor involved in thyroid and kidney differentiation.

We investigated a radical departure from the above thyroid-gene definition by selecting genes that are among the 5000 most expressed thyroid genes, but not in the 7000 most expressed genes in at least 14 of the 15 non-thyroid organs of the BodyMap 2.0. This resulted in a list of 72 thyroid genes (Supplementary table S4). It included additional known thyroid genes, for example, *NKX2-1* (a.k.a. *TTF1*), *DIO2, DUOX1, DUOX2* and *SLC26A4* (a.k.a. pendrin). But, most of the new genes were, as far as we could tell, not known to be characteristic of this organ. Nevertheless, all the results from the previous sections could be reproduced, with slightly better performances, with a thyroid index computed from this gene list (Supplementary Figures S5 to S7). In particular, the AUC for FTA vs FTC increased from 0.82 to 0.86, and that for FVPTC vs CPTC from 0.86 to 0.94. Thus, our method is robust

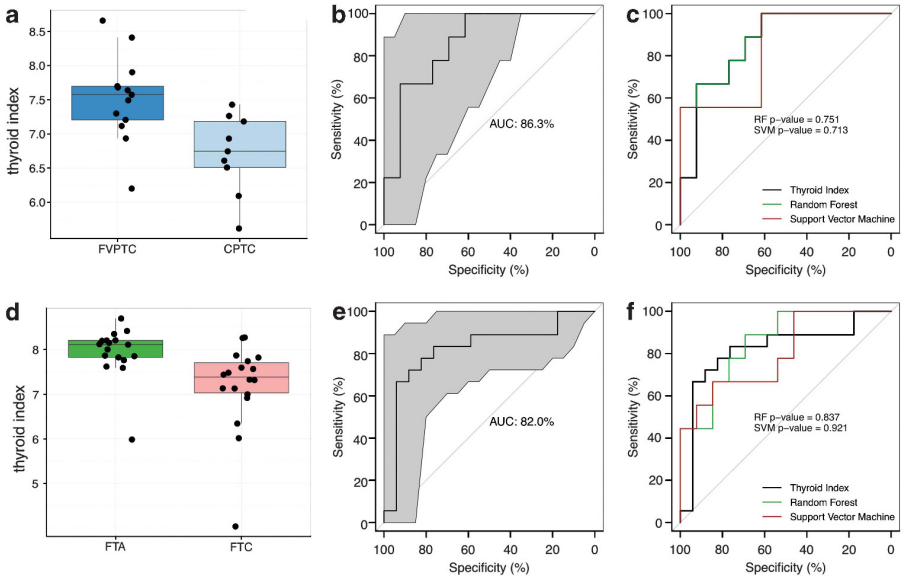

Figure 4 Differential diagnosis. Top panels address the CPTC vs FVPTC classification task, lower panels FTA vs FTC. (**a**) and (**d**) box plots for the thyroid index. (**b**) and (**e**) ROC curves, gray areas represent 95% confidence intervals. (**c**) and (**f**) ROC curves for the thyroid index and two supervised machine-learning algorithms that select optimal classifying genes from the entire set of genes spotted on the microarrays. The *P*-values stand for the difference between the AUCs obtained for the thyroid index and either supervised classifiers. None is significant, that is, the thyroid index, which rests on genes selected in the complete absence of cancer data, is as discriminatory as these classifiers.

with respect to the definition of thyroid-specific expression, and suggests that many genes with thyroid-specific expression await functional characterization.

Discussion

We derived an index of thyroid differentiation exclusively from normal tissue gene expression compendia. We then showed that (1) the thyroid index brings into a quantitative framework the qualitative belief that tumor proliferation and differentiation are inversely related; (2) the thyroid index nevertheless quantifies a process biologically independent of proliferation, as both are positively correlated in a TSH time course experiment; (3) follicular and papillary thyroid cancers dedifferentiate along two distinct gene expression continuums; (4) the thyroid index distinguishes tumor subtypes within each direction and, (5) it does so as accurately as classifiers derived from a whole genome search for genes distinguishing these subtypes.

We could apply the thyroid index and get biologically meaningful results on *in vitro* primary cell data and *in vivo* tumor data generated with Affymetrix U133v2, Affymetrix U95av2 (Aldred *et al*., 2004; Finley *et al*., 2004) and on custom cDNA dual channel arrays (van Staveren *et al*., 2006). Furthermore, the erroneous inclusion of parathyroid hormone in the analysis of U95av2 data did not impair our ability to classify tumors as accurately as possible in this data set. These results support the robustness of the thyroid index across systems and gene expression platforms.

However, all data sets investigated were either generated on single channel arrays, or from sequencing, or from dual channel arrays with mRNA references shared among arrays. Some studies fit none of these setup. The thyroid index could not discriminate FVPTC from CPTC in a data set from our lab (Delys *et al*., 2007; Detours *et al*., 2007) based on dual channel arrays on which individual tumors where hybridized together with patient-matched healthy thyroid tissues (data not shown).

Besides its use as a tool for basic research, the thyroid index could be relevant to a range of clinical problems. It could guide pathologists in the complex differential diagnosis of FTA and FTC, perhaps even preoperatively. In addition, the thyroid index is lower in more aggressive tumors. We observed this when comparing it between FTAs and FTCs, and between PTCs and ATCs. FVPTCs and CPTCs patients have similar overall survival, but the latter have more lymph node metastases and more frequent extrathyroidal involvement (Lang *et al*., 2006; Lin and Bhattacharyya, 2010), in agreement with their lower thyroid index. The thyroid index could also distinguish PTCs from ATCs (Figure 2a). Thus, it could be a useful prognostic marker applicable to several types of thyroid tumors. Measuring the eight genes of the index is manageable in a clinical setup. Large-scale studies are needed to validate the clinical utility of the index.

Rhodes *et al*. (2004) proposed a gene expression signature believed to correlate positively with dedifferentiation in bladder, brain, breast, prostate, lung and ovarian tumors. This signature shares 23 of the 67 genes

of the meta-PCNA index (hypergeometric test yields $P = 5 \times 10^{-30}$), including *PCNA* itself, *TOP2A* and *MCM2* among other proliferation genes. Moreover, the signature of Rhodes *et al*. is positively correlated with meta-PCNA in tumors (Supplementary Figure S8). Thus, it measures dedifferentiation to the extent that it is correlated to proliferation. By contrast, the thyroid index is organ-specific, it correlates negatively with tumor proliferation, dedifferentiation and aggressiveness, and it measures a quantity biologically independent of proliferation as demonstrated by the reversal of the correlation relationship upon TSH treatment of primary thyroid cells. This does not imply of course that the thyroid index of a tumor does not convey information about its proliferation rate.

The thyroid index is tissue specific, but not the method used to derive it. Could this method be applied to other organs? In addition to fibroblasts and endothelial cells, the thyroid is composed of follicular and C cells, but the follicular cells are far more numerous and are the relevant cells in the diseases investigated in this paper. Several other organs have a more complex cell-type composition or the cell type relevant to medical application is not the dominant cell type. For example, the epithelial cells that potentially give rise to breast cancer are few compared with the mass of adipocytes that constitute the normal non-lactating breast. Thus, although profiling of bulk tissues is adequate to select the thyroid index genes, profiling of carefully isolated cell types may be required for other tissues. On the other hand, the cell types in a complex organ may share common transcriptional characteristics. For example, in two normal tissue compendia with a detailed coverage of brain structures, all brain tissues cluster together apart from other non-brain tissues (Jongeneel *et al*., 2005). This opens the possibility to derive both region-specific differentiation indices and a generic brain differentiation index. Their usefulness remains to be investigated in specific applications. The same situation applies to white blood cells. Interestingly, the broad classes of hematological cancers can be traced back to the specific differentiation lineages of the cells they affect. The vast amount of gene expression data sets available for these cancers (Kohlmann *et al*., 2008; Mullighan *et al*., 2008; Mills *et al*., 2009), and the fact that carefully sorted cell types were systematically profiled (Novershtern *et al*., 2011), opens exciting prospects for our method.

Materials and methods

Data

Normal tissue data sets. We obtained the fastq files for the paired-end BodyMap 2.0 data from Illumina (San Diego, CA, USA) (also available from NCBI's GEO http://www.ncbi.nlm.nih.gov/geo, accession number GSE30611). Reads were aligned on the reference human genome hg18 with the Bowtie/Tophat suite (Kim and Salzberg, 2011) in supervised node using the ENSEMBL transcript database (–gtf option) and default parameters otherwise. Expression was then quantified with Cufflink (Roberts *et al*., 2011). We used merged isoform expression. Expression values of mitochondrial genes, pseudogenes and noncoding transcripts were ignored in all subsequent analysis.

Expression profiles from four additional gene atlases profiling healthy human tissues were retrieved as gcrma-normalized (Wu *et al*., 2004), gene-annotated matrices from the InSilico database (insilico.ulb.ac.be, Taminau *et al*., 2011), or from GEO for non-Affymetrix data. These include GSE1747 (Jongeneel *et al*., 2005; 25 tissues profiled by massively parallel signature sequencing), GDS181 (Su *et al*., 2002; 39 tissues profiled on Affymetrix U95a), GSE2361 (Ge *et al*., 2005); 26 tissues profiled Affymetrix U133A) and GSE3526 (Roth *et al*., 2006; 43 tissues profiled on Affymetrix U133v2). The list of tissue names was then manually standardized across studies. For instance, skeletal muscle tissues of distinct origins, otherwise labeled by their anatomical designation, were aggregated under the skeletal.muscle label and central nervous system tissues in GSE3526, representing nearly half of the profiled tissues in that study, were aggregated under the CNS label in order to avoid biases. After standardization, expression values from tissues with the same label were averaged. Diseased and fetal tissues were subsequently discarded.

ATC–PTC–normal thyroid data set. A group of 11 anaplastic thyroid carcinomas (ATCs), together with 49 papillary thyroid carcinomas (PTCs) paired with 45 adjacent tissues (N), were hybridized onto Affymetrix U133v2 arrays. PTCs were obtained from the Chernobyl Tissue Bank (CTB, http://www.chernobyltissuebank.com). ATCs were obtained from the Jules Bordet Institute (Brussels, Belgium) and the Ambroise Paré Hospital (Paris, France). All tissues were immediately dissected, placed on ice, snap-frozen in liquid nitrogen and stored at –80 °C until processing. The ethics committees of the involved institutions approved the protocol. RNAs were prepared using TRIzol Reagent and RNeasy columns (Qiagen, Venlo, Netherlands), followed by a verification of their quality. Amplification and hybridizations were performed following Affymetrix instructions. CEL files were normalized with RMA (Irizarry *et al*., 2003) and gene annotations were reconstructed using the Bioconductor package annotate (Gentleman *et al*., 2004). Data have been deposited in the GEO database (http://www.ncbi.nlm.nih.gov/geo, pending submission).

FTA–FTC–CPTC–FVPTC–normal thyroid data set. Seven normal thyroid samples (N) and nine follicular thyroid carcinomas (FTCs) were profiled by microarray as referenced in Aldred *et al*. (2004). We downloaded them from the authors' web site. A total of 17 follicular thyroid adenomas (FTAs), 9 FTCs, 13 FVPTC and 9 classical papillary thyroid carcinomas (CPTCs) were profiled by microarray as described in Finley *et al*. (2004). These data have been submitted to GEO with accession number GSE29315. CEL files from these 64 samples were normalized altogether with RMA (Irizarry *et al*., 2003) and gene annotations were reconstructed using the Bioconductor (Gentleman *et al*., 2004) package annotate.

TSH-time course data set. This data set, published in van Staveren *et al*. (2006), was downloaded from http://www.ulb.ac.be/medecine/iribhm/microarray/data/ and used without any further processing.

Computational Methods

Software platform. All calculations were performed with the R language for statistics version 2.11.0 (R Development Core Team) and Bioconductor 2.6 (Gentleman *et al*., 2004) software

and annotation packages. Graphical outputs (with the exception of ROC curve analyses) were generated with the R package ggplot2 (Wickham, 2009). All functions were run with default parameters unless specified otherwise.

Gene signatures and indices

Thyroid differentiation signature. To derive a thyroid differentiation signature, we ranked genes according to their FPKM in each tissues and selected genes with rank <1000 in the thyroid and >5000 in all other organs (or <5000 in thyroid, >7000 in 14 non-thyroid tissues for the alternate list).

Platform-specific probe selection. The following was repeated for each gene on both the compendia of Roth *et al*. (Affymetrix U133v2, GSE3526) and Su *et al*. (Affymetrix U95av2, GDS181). We first filtered the probes by retaining those whose median expression in all thyroids profiled ranked in the top 10% of the measurements across all tissues. From those left, if any, we then selected the probe with the highest *t*-statistics when comparing its expression in thyroid vs all other organ expression.

Meta-PCNA signature. The signature can be retrieved from the online supplementary material of Venet *et al*. (2011).

Indices computation. The thyroid and the meta-PCNA indices were computed on a per microarray basis as the median of the expression of the genes comprised, respectively, in the thyroid differentiation signature and in the meta-PCNA signature.

Statistical tests. Non-parametric Mann–Whitney *U* tests, as implemented in the R function wilcox.test, were used to test the null hypothesis that the observed distributions of either the thyroid index or the meta-PCNA index between two morphologically distinct cancer types were drawn from the same distribution.

A test for association of paired samples, based on Spearman's rank correlation (as implemented in the R function cor.test, was used to test the correlations between paired observations of the thyroid index and the meta-PCNA index.

The function roc.test from the pROC package was used to compare ROC curves, implementing the method described in DeLong *et al*. (1988).

Unsupervised analyses

Multidimensional scaling analysis. Multidimensional scaling was performed over a distance matrix of the FTA-FTC-CPTC-FVPTC-N data set obtained by computing the correlation (Pearson method) over all complete pairs of observations between samples. The function isoMDS (with $k=2$, maxit $=1000$, tol $=1e{-}20$) of the MASS R package (http://cran.r-project.org/) was then used to compute the plot of Figure 3.

Supervised analyses

Machine learning. A customized version of the Bioconductor package MCRestimate (Ruschhaupt *et al*., 2004) was used to determine the best possible RF and linear kernel SVM\ classifiers to predict the outcome of the FTA vs FTC and CPTC vs FVPTC classification in the FTA-FTC-CPTC-FVPTC-N data set. This package uses a protocol of repeated inner/outer cross-validation to estimate the expected accuracy of the prediction of each of those classifiers on new data.

RF and SVM classifiers were built for following feature selections: thePreprocessingMethods = varSel.highest.t.stat. Classifiers were optimized over large ranges for their specific parameters. For RF classifiers, the following range of values for each parameter was optimized: var.numbers $\in\{2, 4, 8, 16, 32, 64, 128, 256, 512\}$; nodesize $\in\{1, 3, 5\}$; ntree $\in$ $\{250, 500, 1000\}$. For SVM classifiers, the following range of values for each parameter was optimized: var.numbers $\in\{2, 4, 8, 16, 32, 64, 128, 256, 512\}$; cost $\in\{0.0001, 0.001, 0.01, 0.1, 1\}$.

Classifiers were trained on 9/10 of our samples and used to predict the remaining 1/10. They were tuned on 9/10 of the training samples and optimized by comparing the prediction on the remaining 1/10 of the training samples. For each sample, our version of MCRestimate returns for RF a percentage of decision trees votes and for SVM a real score between -1 and 1. The whole loop was repeated 10 times and the results were averaged. Those scores were then used as a diagnostic test to compute ROC curves.

ROC curve analyses. ROC curves and AUC were computed using the R package pROC. AUC *P*-values were obtained by permuting 10 000 times the class labels and counting the fraction of permutation AUCs greater of equal than those obtained from the original data. We also checked the AUC values against the null hypothesis that any signatures of eight probes is diagnostic (Venet *et al*., 2011) by evaluating AUCs on non-permuted samples from 100 000 indices computed from eight probes selected at random among all probes printed on the microarrays.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

This research was partially funded by the Brussels-Capital IRSIB project ICT-impulse 2006, In Silico Wet Lab. GT is supported by the Wallonie–Bruxelles International grant (7450/AMG/VDL/IN,WBI/doh/2009/21649). MT is supported by a FRIA fellowship from FNRS.

References

Aldred MA, Huang Y, Liyanarachchi S, Pellegata NS, Gimm O, Jhiang S *et al*. (2004). Papillary and follicular thyroid carcinomas show distinctly different microarray expression profiles and can be distinguished by a minimum of five genes. *J Clin Oncol* **22**: 3531–3539.

Baloch ZW, Hendreen S, Gupta PK, LiVolsi VA, Mandel SJ, Weber R *et al*. (2001). Interinstitutional review of thyroid fine-needle aspirations: impact on clinical management of thyroid nodules. *Diagn Cytopathol* **25**: 231–234.

Clary KM, Condel JL, Liu Y, Johnson DR, Grzybicki DM, Raab SS. (2005). Interobserver variability in the fine needle aspiration biopsy diagnosis of follicular lesions of the thyroid gland. *Acta Cytol* **49**: 378–382.

DeLong ER, DeLong DM, Clarke-Pearson DL. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**: 837–845.

Delys L, Detours V, Franc B, Thomas G, Bogdanova T, Tronko M *et al*. (2007). Gene expression and the biological phenotype of papillary thyroid carcinomas. *Oncogene* **26**: 7894–7903.

Detours V, Delys L, Libert F, Weiss Solís D, Bogdanova T, Dumont JE *et al*. (2007). Genome-wide gene expression profiling suggests distinct radiation susceptibilities in sporadic and post-Chernobyl papillary thyroid cancers. *Br J Cancer* **97**: 818–825.

Finley DJ, Zhu B, Barden CB, Fahey TJ. (2004). Discrimination of benign and malignant thyroid nodules by molecular profiling. *Ann Surg* **240**: 425–436; discussion 436–437.

Ge X, Yamamoto S, Tsutsumi S, Midorikawa Y, Ihara S, Wang SM *et al*. (2005). Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics* **86**: 127–141.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S *et al*. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**: R80.

Hegedüs L. (2004). Clinical practice. The thyroid nodule. *N Engl J Med* **351**: 1764–1771.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U *et al*. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**: 249–264.

Jongeneel CV, Delorenzi M, Iseli C, Zhou D, Haudenschild CD, Khrebtukova I *et al*. (2005). An atlas of human gene expression from massively parallel signature sequencing (MPSS). *Genome Res* **15**: 1007–1014.

Kebebew E, Greenspan FS, Clark OH, Woeber KA, McMillan A. (2005). Anaplastic thyroid carcinoma. Treatment outcome and prognostic factors. *Cancer* **103**: 1330–1335.

Kim D, Salzberg SL. (2011). TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol* **12**: R72.

Kohlmann A, Kipps TJ, Rassenti LZ, Downing JR, Shurtleff SA, Mills KI *et al*. (2008). An international standardization programme towards the application of gene expression profiling in routine leukaemia diagnostics: the Microarray Innovations in LEukemia study prephase. *Br J Haematol* **142**: 802–807.

Kondo T, Ezzat S, Asa SL. (2006). Pathogenetic mechanisms in thyroid follicular-cell neoplasia. *Nat Rev Cancer* **6**: 292–306.

Lang BH-H, Lo C-Y, Chan W-F, Lam AK-Y, Wan K-Y. (2006). Classical and follicular variant of papillary thyroid carcinoma: a comparative study on clinicopathologic features and long-term outcome. *World J Surg* **30**: 752–758.

Lin HW, Bhattacharyya N. (2010). Clinical behavior of follicular variant of papillary thyroid carcinoma: presentation and survival. *Laryngoscope* **120**(Suppl 4): S163.

Lubitz CC, Gallagher LA, Finley DJ, Zhu B, Fahey TJ. (2005). Molecular analysis of minimally invasive follicular carcinomas by gene profiling. *Surgery* **138**: 1042–1048; discussion 1048–1049.

Mills KI, Kohlmann A, Williams PM, Wieczorek L, Liu W-min, Li R *et al*. (2009). Microarray-based classifiers and prognosis models identify subgroups with distinct clinical outcomes and high risk of AML transformation of myelodysplastic syndrome. *Blood* **114**: 1063–1072.

Moreno JC, Klootwijk W, van Toor H, Pinto G, D'Alessandro M, Lèger A *et al*. (2008). Mutations in the iodotyrosine deiodinase gene and hypothyroidism. *N Engl J Med* **358**: 1811–1818.

Mullighan CG, Miller CB, Radtke I, Phillips LA, Dalton J, Ma J *et al*. (2008). BCR-ABL1 lymphoblastic leukaemia is characterized by the deletion of Ikaros. *Nature* **453**: 110–114.

Novershtern N, Subramanian A, Lawton LN, Mak RH, Haining WN, McConkey ME *et al*. (2011). Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* **144**: 296–309.

R Development Core Team, R: A Language and Environment for Statistical Computing. 1: ISBN 3-900051-07-0.

Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D *et al*. (2004). Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci USA* **101**: 9309–9314.

Roberts A, Pimentel H, Trapnell C, Pachter L. (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27**: 2325–2329.

Roth RB, Hevezi P, Lee J, Willhite D, Lechner SM, Foster AC *et al*. (2006). Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. *Neurogenetics* **7**: 67–80.

Ruschhaupt M, Huber W, Poustka A, Mansmann U. (2004). A compendium to ensure computational reproducibility in high-dimensional classification tasks. *Stat Appl Genet Mol Biol* **3**: Article37.

van Staveren WCG, Solís DW, Delys L, Venet D, Cappello M, Andry G *et al*. (2006). Gene expression in human thyrocytes and autonomous adenomas reveals suppression of negative feedbacks in tumorigenesis. *Proc Natl Acad Sci USA* **103**: 413–418.

Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T *et al*. (2002). Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci USA* **99**: 4465–4470.

Taminau J, Steenhoff D, Coletta A, Meganck S, Lazar C, de Schaetzen V *et al*. (2011). inSilicoDb: an R/Bioconductor package for accessing human Affymetrix expert-curated datasets from GEO. *Bioinformatics [Internet]*. Available from:http://www.ncbi.nlm.nih.gov/pubmed/21937664.

Venet D, Dumont JE, Detours V. (2011). Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput Biol* **7**: e1002240.

Wang Z, Gerstein M, Snyder M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**: 57–63.

Wickham H. (2009). *ggplot2: Elegant Graphics for Data Analysis* 2nd edn. Springer (http://cran.r-project.org/).

Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F. (2004). A model-based background adjustment for oligonucleotide expression arrays. *J Am Stat Assoc* **99**: 909–917.

Zurzolo C, Gentile R, Mascia A, Garbi C, Polistina C, Aloj L *et al*. (1991). The polarized epithelial phenotype is dominant in hybrids between polarized and unpolarized rat thyroid cell lines. *J Cell Sci* **98**(Part 1): 65–73.

Supplementary Information accompanies the paper on the Oncogene website (http://www.nature.com/onc)

The extent of prognostic signals in the cancer transcriptomes

Executive Summary

Reproducibility of publicly reported results has long been a concern among the scientific community,[213] and in particular in the microarray research field.[214] A common strategy to validate the implication of a particular biological mechanism in cancer progression has consisted of deriving a gene expression biomarker from an *in vitro* system and then establish its association with differential survival outcome in a cohort of cancer patients.[215] Dozens of articles have built upon this approach since 2002 (*cf.* Introduction chapter), which still warrants publication success as late as of 2015.[216]

[213] Ioannidis, 2005

[214] Ioannidis et al., 2009

[215] van't Veer et al., 2002

[216] Voduc et al., 2015; and Bosch et al., 2015

However, this reasoning hinges on the specificity of the association of the biologically-motivated gene expression signature of choice with a particular cancer outcome, on a given cohort of patients. Venet et al. have demonstrated that this assertion does not hold in two reference breast cancer cohorts, as most random gene expression signatures are associated with outcome in these two datasets.[217]

[217] Venet et al., 2011

Here we report the extension of these findings to 114 cohorts of expression profiles spanning 19 human cancers, collected and manually curated from public online databases. To do so, we measured in each of the cancer datasets the fraction of single-genes profiled on each platform; the fraction of MSigDB C2 gene expression signatures, downloaded from the Broad Institute; and the fraction of a randomized collection of gene expression signatures of the same size as MSigDB C2, when associated with disease outcome, at a canonical $p < 0.05$.

The chief finding of this research program is the broad heterogeneous spectrum of prognostic fractions observed across the studied datasets, ranging from 2% to 59% for single-gene markers. A re-sampling experiment of single-gene prognostic fractions on METABRIC, the largest public breast cancer gene expression dataset with follow-up data, has established that prognostic fractions are notably sensitive to sampling variance, size of the cohort, follow-up time and, specifically in the case of breast cancer, to the fraction of ER+ and node positive patients in each sampling of the 1992 expression profiles in the cohort. Furthermore, we identified at least a major technical pre-processing artefact on one of the datasets analyzed, responsible for an artificial inflation of the single-gene prognostic fraction from 19% to 59%.

The main implication of this results is the disavowal of the widespread use of statistical association between expression markers and cancer patients outcome, at canonical *p*-values, in order to infer implication of particular biological mechanisms in cancer progression. This can be illustrated taking the TCGA kidney renal clear cell carcinoma, KIRC, for instance. The multi-gene prognostic fraction for this dataset is 50%. This means than an investigator seeking to validate the prognostic value of any given biologically-motivated expression signature in this cohort has a 50% of a spurious positive result, at $p < 0.05$.

Because we are interested in simulating the process of a scientist screening data for gene expression signatures of outcome, the scope of this study

is strictly epistemological. We sought to reproduce common practices observed by many researchers in the microarray field, quantify the extent of prognostic signals in cancer transcriptomes, and explain the heterogeneous nature of these signals across the datasets analyzed—thus no multiple testing correction for prognostic fractions is warranted. We showed that this heterogeneity is partly explained by a collection of technical, demographic, statistical and biological variables associated to each dataset. Furthermore, this heterogeneity is independent of the technology used to profile the transcriptomes of cancer biospecimens—microarray or RNA-seq—, and of the microarray platform used.

Most importantly, while these results do not challenge the ability of cancer transciptomics to predict disease outcome, the pervasive nature of prognostic signals does call for more stringent controls when seeking to make biological inference from gene expression association to outcome in the cancer setting.

Article

The vast majority of mechanistic cancer studies are performed in animals and/or *in vitro* models for ethical reasons. Proving that a biological mechanism established in these experimental systems contributes to cancer progression in human requires clinical trials, which are costly and take years to perform. Therefore, researchers have relied on weaker correlative evidence to back the relevance of their findings to human diseases. One such evidence is the statistically significant association of a molecular maker of the biological mechanism under investigation with disease outcome in human. Such association can be established with non-interventional clinical studies and has been extensively used in the past decades. Moreover, this approach has recently gained popularity with the availability of web servers[218] that provide, for free, association statistics between any single- or multi-gene markers and cancer outcomes using publicly available cancer transcriptome databases.

[218] Győrffy et al., 2010; Ringnér et al., 2011; and Győrffy et al., 2013

To retain a biological bearing, the reported association must however be specific of the biomarker under analysis and not reflect a global property of the transcriptome. Most studies estimate the association with a log-rank test based on a Cox Proportional Hazards model and use a Kaplan-Meier curve to visualize it. None of these tools control for the possibility that a large fraction of transcriptome could be associated with outcome and that, therefore, the prognostic signal is non-specific.

Disturbingly, this possibility has been proven true in bladder[219] and breast[220] cancers. In particular, our group has shown[221] that more than half of the transcriptome is correlated with proliferation in two breast cancer cohorts and that one out of four single genes and that nine out of ten random signatures comprised of more than 100 genes were associated with overall survival at $p < 0.05$. As a consequence, most published signatures were found to be significantly prognostic, but no more prognostic than random sets of genes.

[219] Lauss et al., 2010

[220] Ein-Dor et al., 2005; Mosley and Keri, 2008; and Venet et al., 2011

[221] Venet et al., 2011

To investigate whether the pervasive association of the transcriptome with outcome extends to other cohorts and other types of cancers we com-

piled 114 published gene-expression datasets with patient follow-up from the GEO[222], InSilicoDB[223] and the TCGA Research Network databases.[224] These included human cancers from 19 organ systems, and a representative range of microarray platforms (Table 3).

[222] Edgar et al., 2002

[223] Coletta et al., 2012

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

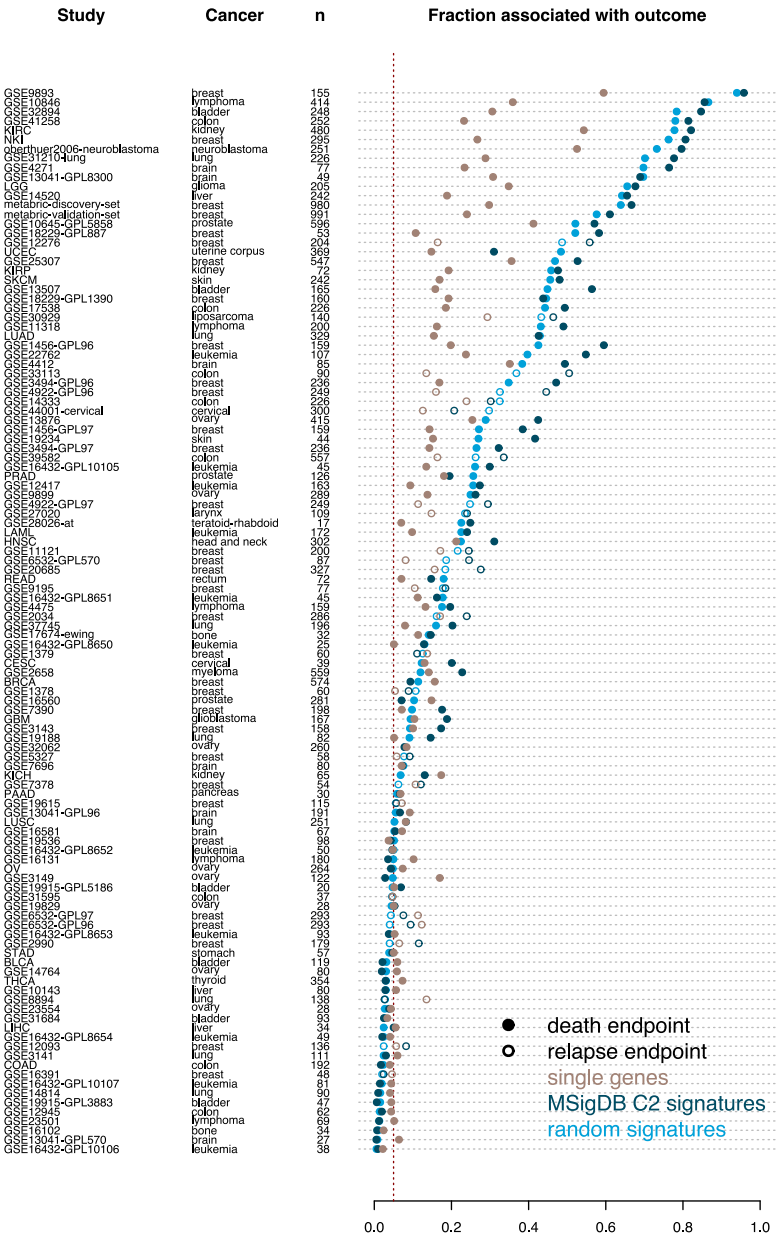

Figure 13: Prognostic content in human cancers. The fraction of markers significantly associated with outcome at logrank $p < 0.05$ is shown for single-gene markers, multi-gene markers from the MSigDB C2 database, and randomly generated multi-genes markers devoid of biological meaning. The latter was used to order the datasets. Fractions were computed from death-related end-points when available (OS or DSS), or relapse otherwise (DFS or DMFS). The dotted red line marks the 0.05 threshold.

For each dataset, we computed the fraction of genes associated with outcome at log-rank $p < 0.05$ (Figure 13). This quantity may be viewed as the probability of observing a significantly prognostic single gene marker by chance alone. We thus refer to these estimates as the baseline prognostic content[225] of cancer transcriptomes. Overall, the median prognostic content across all studies for single-gene markers was 12%. In 100 of the 114 datasets analyzed (88%), more than five percent of single-gene markers

[225] Prognostic content will henceforth refer to the fraction of tested markers found associated with outcome in a given dataset.

showed a significant association with outcome. To take an extreme example, an investigator measuring the association of a gene with outcome in the KIRC datasets (kidney cancer) has a 50% chance to obtain a positive result—a value far above the canonical 5% significance threshold.

To investigate recent multi-gene approaches, we ran a similar calculation for each of the 4722 curated gene sets of the MSigDB C2 database[226] (Figure 13). The prognostic content was larger for multi-gene markers than for single-gene markers, with a median of 19%. The prognostic content for multi-gene markers was larger than 5% in 76% of the datasets (87 out of 114), larger than 20% in 48% (55/114), and larger than 50% in 19% (22/114) of the datasets. To control for possible biases of MSigDB C2 towards oncology-related signatures, we reran the same computation, but replacing each signature by a similarly sized set of randomly selected genes. The overall qualitative result is unchanged (Figure 13).

[226] Liberzon et al., 2011

Intriguingly, single-gene prognostic content was found to be largely heterogeneous across datasets related to the same organ system. For example, it ranged from 4% to 59% among the 33 breast cancer datasets analyzed. To investigate the contributions of potential biological and demographic dataset-specific factors to this effect, we quantified single-marker prognostic fraction for re-samplings of the 1972 transcriptomes of the METABRIC breast cancer cohort, regarding modulations of four variables: sample size, duration of follow-up time, fraction of ER+ patients, and fraction of node positive patients (Figure 14). We chose METABRIC for this analysis because it is one of the largest cohorts of cancer patients with follow-up and extensive clinical annotation data available in the public domain.

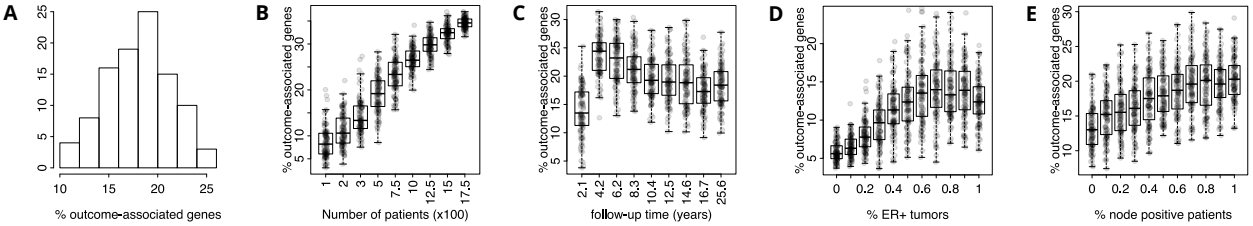

Figure 14: Bootstrapping experiments on the 1972 combined breast cancer transcriptomes of the METABRIC dataset. **A**–Distribution of the prognostic fraction in 100 samplings of 500 expression profiles, out of the total 1972 METABRIC profiles. **B**–Effect of sample size. 100 samplings (grey points) were assessed for each specified sample size. **C**–Effect of follow-up times. For each time *t*, 100 samplings were assessed for which patients beyond time *t* were considered censored at time *t*. **D**–Effect of the fraction of ER+ patients. 100 samplings, each of 300 patients, were assessed with the respective fraction of ER+ samples. **E**–Effect of the fraction of node positive patients. 100 samplings, each of 500 patients, were assessed with the respective fraction of node positive patients.

The estimates of prognostic content were found to be markedly sensitive to sampling variance, as suggested by the dispersal of the distribution of estimates (confidence interval: 12% to 23%), when 100 samplings of 500 random transcriptomes were examined for the fraction of genes associated with outcome (Figure 14*a*). This feature alone is likely to yield a significant contribution to the range of estimates computed in our meta-analysis, as 108 of the 114 datasets analyzed included less than 500 profiled tumours. Moreover, the sensitivity of our estimation procedure is, unsurprisingly, largely dependent on the number of profiles included in the analysis, as shown by a bootstrapping experiment of sample sizes towards the assessment of the prognostic fraction (Figure 14*b*). Provocatively, the estimates of prognostic content do not appear to level off even when the experimental sample size reached 1750—by far the largest in the field. A sampling experiment of sequential truncation of follow-up times was equally shown to impact estimates of prognostic content (Figure 14*c*). Sharply increasing

predictive fractions were observed up to the fifth year of follow-up, followed by a gradual decrease of the estimates for higher follow-up times. This trend suggests that, in breast cancers, prognostic patterns of expression are optimally correlated with short-term forms of progression of the disease, and that long-term forms of progression are less efficiently predicted from primary tumour transcriptomes. The modulation of the fraction of ER+ transcriptomes towards experimental samplings of our estimate has exposed a tendency congruent with the clinical relevance of this receptor in breast cancer pathology (Figure 14*d*). Thus, an increase of ER+ profiles to up to 50% in our samplings leads to a corresponding linear rise in estimates of prognostic content, at which point a further increase in the proportion of ER+ profiles yields little impact on fraction estimates. This observation is in line with the fact that the predictive power of most signatures in breast cancer is mostly confined to ER+ phenotypes.[227] A last sampling experiment with increasing fractions of profiles from node positive patients (Figure 14*e*) also revealed an increasing pattern of prognostic fraction estimates. This trend could be explained by the fact that nodal status is clinically correlated to ER status in breast cancer.

[227] Weigelt et al., 2012

Finally, dataset-specific processing details may also distort prognostic estimates. Consider, for instance, dataset GSE9893, which exhibits the highest prognostic fractions measured in our study (Figure 13). A thorough reanalysis of this dataset, detailed in the following section, reveals that its normalization was performed in two batches and induced massive spurious correlations between global values of expression and survival outcome. Accordingly, a proper single-batch normalization of the raw expression data restores a signal-to-noise metrics to comparable values with other datasets, and decreases measurements of prognostic content from 59% to 19% for GSE9893.

We have shown that the fraction of prognostic single- and multi-gene biomarkers is greater that 5% in the majority of publicly available transcriptome datasets. Furthermore, we have demonstrated that the probability of a significant single- or multi-gene marker association with outcome depends on cohorts' demographics, but also to a large extent on technical factors that include sampling effects, cohort size, patient follow-up protocols, protocol randomization and possibly other factors not addressed here. These findings call for the reappraisal of conclusions made by previous studies pertaining to the implication of biological mechanisms to human cancer based on associations of biomarkers with outcome—including low-throughput PCR-based studies. They also call for study-specific controls akin to those presented in Figure 13 in future studies.

However, a biomarker does not need to convey relevant biological information regarding the course of disease in order to be useful in the clinic. Therefore, our results have no bearing on the clinical utility of published biomarker associations.

Re-analysis of dataset GSE9893

To illustrate how procedural biases in the preprocessing of expression profiles may impact estimates of association to outcome, we present here a

detailed re-analysis of dataset GSE9893 (Table 3).

GSE9893 is comprised of 155 samples of tamoxifen-treated primary breast cancers. These samples were hybridized on a homemade 70-mer chip containing 22 680 probes, mapping to 21 329 human specific genes. The original experiment was carried out to look for a gene expression signature to predict the recurrence of tamoxifen-treated primary breast cancer.[228]

The data-set was downloaded from GEO with the Bioconducor GEOquery package, with original normalization. The expression matrix was then feature collapsed using a maxMean routine and median polished.

Among the 114 studies considered in our analysis, GSE9893 shows the highest fraction of genes associated with outcome at $p < 0.05$ (59%). Interestingly, nearly all MSigDB C2 signatures appear associated with outcome in this dataset (Table 4).

[228] Chanrion et al., 2008

Dataset	Fraction of significant tests	Event
GSE9893-breast	0.958	OS
GSE10846-lymphoma	0.856	OS
GSE32894-bladder	0.847	DSS
GSE31210-lung-adenocarcinoma	0.838	DFS
KIRC	0.821	OS
GSE41258-colon	0.814	DSS

Table 4: Top six studies with highest fraction of MSigDB C2 signatures associated with outcome. Detailed information regarding each dataset can be found on to Table 3.

A closer inspection of the metadata associated with the expression profiles reveals that the arrays were scanned in two discrete time intervals during 2005 and 2006, separated by eight months (Figure 15). Surprisingly, patients whose tumours were hybridized in 2006 show a poorer prognosis than those hybridized in 2005 (Figure 16).

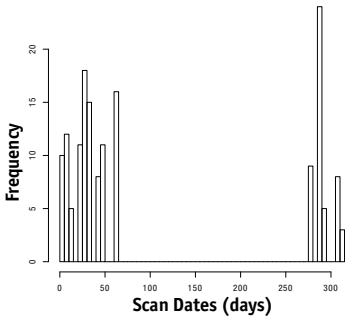

Figure 15: Frequency distribution of hybridization dates of GSE9893 samples relative to the first hybridization date. Information regarding date of hybridization of each of the 155 arrays was parsed from the gpr files downloaded from GEO. The dataset is composed by a batch of samples hybridized during May to July 2005 and a second batch of samples hybridized during February and March 2006, roughly 200 days apart.

This observation can be explained by two facts. First, we discovered a normalization artifact in this dataset related to the 2005 and 2006 batches of samples, as shown by the distribution of expression values and respective batch associations with the first and second principal components of the global expression matrix (Figure 17, left panels). Second, there is an enrichment in the 2006 batch of observed death events compared with the 2005 batch (Figure 18). Because the majority of observed events are linked with a subset of samples whose global expression patterns were distorted due to the normalization artifact, 59% of genes in original matrix appear artificially associated with overall survival in this dataset.

In order to correct for this bias, we downloaded the raw data gpr files and proceeded to re-normalize them with the Bioconductor limma package. As a result, the distribution of values of expression no longer showed a correlation between batches of samples (Figure 17, right panels). In addition, a signal-to-noise quality metric based on gene-gene correlations across expression profiles[229] suggests that the re-normalization of the raw-data has significantly improved the data quality of GSE9893 (Figure 19). As a result, the fraction of genes associated with outcome in this study is reduced from 58% to 19%, and only 74 out of the original 4556 (2%) MSigDB C2 remain associated with overall survival.

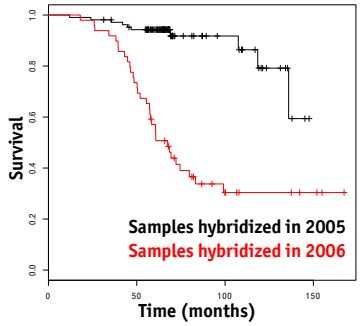

Figure 16: A Kaplan Meier visualization of differential overall survival, between patients included in GSE9893 whose samples were hybridized in 2005 (in black) and those whose samples were hybridized in 2006 (in red). Logrank test: $p = 1.76^{-10}$.

[229] Venet et al., 2012

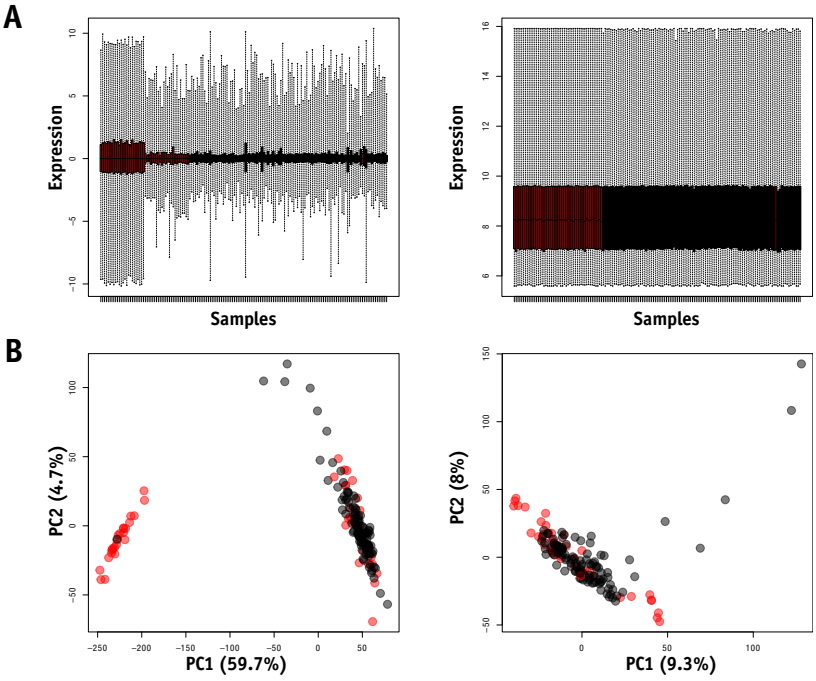

Figure 17: **A**–Gene expression distributions of each of the 155 samples in GSE9893. **B**–GSE9893 samples projected in the space of the first two principal components of their expression matrix. *Left,* original normalization; *right,* quantile re-normalization on the original gpr files. Samples in black are from the 2005 batch and samples in red are from the 2006 batch (See text for details).

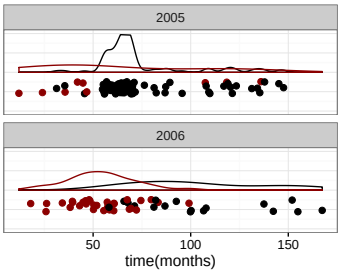

Figure 18: Density distribution of overall survival events in GSE9893. Censored observations are denoted in black and observed death events are denoted in red. Out of the 116 patients in the 2005 batch, only 10 died during the course of the study; whereas out of the 49 patients whose samples where hybridized in 2006, 32 were observed events (χ^2 test: $p = 1.42^{-12}$).

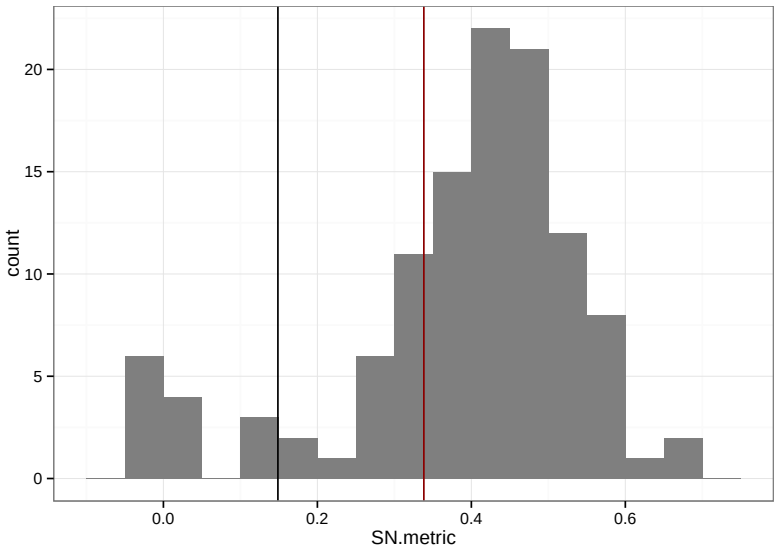

Figure 19: Distribution of signal-to-noise quality metrics across the 114 human cancer datasets included in our study. The signal-to-noise-ratios (SNR) were computed with the Bioconductor SNAGEE package. The black vertical line marks the value computed for GSE9893. The red vertical line shows the value computed for the re-normalized expression matrix of GSE9893. The SNR of a study is based on the correlation between its gene-gene correlation matrix and the expected matrix, and is thus a number between −1 and 1. Practically, numbers near or below 0 are symptomatic of seriously problematic studies (e.g. gene annotation problems, serious normalization issues). Numbers around 20–30% are average, depending on the platform.

Other Contributions

This sections reports published results to which the author contributed at large during this dissertation. For each work, a short synopsis of the main findings as well as the specific contributions of the author will be presented.

Role of Epac and protein kinase A in thyrotropin-induced gene expression in primary thyrocytes

Wilma C.G. van Staveren[a], Sandrine Beeckman[a], Gil Tomás[a], Geneviève Dom[a], Aline Hébrant[a], Laurent Delys[a], Marjolein J. Vliem[b], Christophe Trésallet[c], Guy Andry[d], Brigitte Franc[e], Frédérick Libert[a], Jacques E. Dumont[a], Carine Maenhaut[a,]*

[a]*Institute of Interdisciplinary Research (IRIBHM), Université Libre de Bruxelles, 808 Route de Lennik, B-1070 Brussels, Belgium*
[b]*Molecular Cancer Research, Centre for Biomedical Genetics and Cancer Genomics Centre, University Medical Center Utrecht, 3584 CG Utrecht, The Netherlands*
[c]*Hôpital de La Pitié- Salpêtrière, 47-83, Boulevard de l'Hôpital, 75013 Paris, France*
[d]*Department of Surgery, Jules Bordet Institute, Brussels, Belgium*
[e]*Université de Versailles Saint-Quentin-en-Yvelines (UVSQ), Department of Pathology, Ambroise-Paré Hospital (APHP), Boulogne-Billancourt, France*

ARTICLE INFORMATION

Article Chronology:
Received 27 January 2011
Revised version received
28 November 2011
Accepted 26 December 2011
Available online 4 January 2012

Keywords:
Epac
Rap
RapGAP
Thyroid
Cancer
Microarrays

ABSTRACT

cAMP pathway activation by thyrotropin (TSH) induces differentiation and gene expression in thyrocytes. We investigated which partners of the cAMP cascade regulate gene expression modulations: protein kinase A and/or the exchange proteins directly activated by cAMP (Epac). Human primary cultured thyrocytes were analysed by microarrays after treatment with the adenylate cyclase activator forskolin, the protein kinase A (PKA) activator 6-MB-cAMP and the Epac-selective cAMP analog 8-pCPT-2′-O-Me-cAMP (007) alone or combined with 6-MB-cAMP. Profiles were compared to those of TSH. Cultures treated with the adenylate cyclase- or the PKA activator alone or the latter combined with 007 had profiles similar to those induced by TSH. mRNA profiles of 007-treated cultures were highly distinct from TSH-treated cells, suggesting that TSH-modulated gene expressions are mainly modulated by cAMP and PKA and not through Epac in cultured human thyroid cells. To investigate whether the Epac-Rap-RapGAP pathway could play a potential role in thyroid tumorigenesis, the mRNA expressions of its constituent proteins were investigated in two malignant thyroid tumor types. Modulations of this pathway suggest an increased Rap pathway activity in these cancers independent from cAMP activation.

This work sought to clarify which partners of the cAMP cascade regulate the TSH-induced gene expression modulation in thyrocytes: protein kinase A and/or the EPAC proteins.[230] Contingent to this objective was the characterization of the potential role of the Epac-Rap-RapGAP pathway in thyroid tumorigenesis. The author contributed with data analysis, figure generation and results discussion.

[230] van Staveren et al., 2012

5-Aza-2'-Deoxycytidine has minor effects on differentiation in human thyroid cancer cell lines, but modulates genes that are involved in adaptation in vitro

Geneviève Dom,[1,*] Vanessa Chico Galdo,[1,*] Maxime Tarabichi,[1] Gil Tomás,[1] Aline Hébrant,[1] Guy Andry,[2] Viviane De Martelar,[1] Frédérick Libert,[1] Emmanuelle Leteurtre,[3] Jacques E. Dumont,[1] Carine Maenhaut,[1,4] and Wilma C.G. van Staveren[1]

Background: In thyroid cancer, the lack of response to specific treatment, for example, radioactive iodine, can be caused by a loss of differentiation characteristics of tumor cells. It is hypothesized that this loss is due to epigenetic modifications. Therefore, drugs releasing epigenetic repression have been proposed to reverse this silencing.
Methods: We investigated which genes were reinduced in dedifferentiated human thyroid cancer cell lines when treated with the demethylating agent 5-aza-2′-deoxycytidine (5-AzadC) and the histone deacetylase inhibitors trichostatin A (TSA) and suberoylanilide hydroxamic acid, by using reverse transcriptase–polymerase chain reaction and microarrays. These results were compared to the expression patterns in *in vitro* human differentiated thyrocytes and in *in vivo* dedifferentiated thyroid cancers. In addition, the effects of 5-AzadC on DNA quantities and cell viability were investigated.
Results: Among the canonical thyroid differentiation markers, most were not, or only to a minor extent, reexpressed by 5-AzadC, whether or not combined with TSA or forskolin, an inducer of differentiation in normal thyrocytes. Furthermore, 5-AzadC–modulated overall mRNA expression profiles showed only few commonly regulated genes compared to differentiated cultured primary thyrocytes. In addition, most of the commonly strongly 5-AzadC–induced genes in cell lines were either not regulated or upregulated in anaplastic thyroid carcinomas. Further analysis of which genes were induced by 5-AzadC showed that they were involved in pathways such as apoptosis, antigen presentation, defense response, and cell migration. A number of these genes had similar expression responses in 5-AzadC–treated nonthyroid cell lines.
Conclusions: Our results suggest that 5-AzadC is not a strong inducer of differentiation in thyroid cancer cell lines. Under the studied conditions and with the model used, 5-AzadC treatment does not appear to be a potential redifferentiation treatment for dedifferentiated thyroid cancer. However, this may reflect primarily the inadequacy of the model rather than that of the treatment. Moreover, the observation that 5-AzadC negatively affected cell viability in cell lines could still suggest a therapeutic opportunity. Some of the genes that were modulated by 5-AzadC were also induced in nonthyroid cancer cell lines, which might be explained by an epigenetic modification resulting in the adaptation of the cell lines to their culture conditions.

This work aimed at investigating the extent to which 5-aza-2'-deoxycytidine, a DNA demethylation agent, is able to reactivate the expression of differentiation markers potentially repressed by epigenetic modifications in thyroid cancer cell lines.[231] The author contributed with data analysis, figure generation and results discussion.

[231] Dom et al., 2013

Intratumor heterogeneity and clonal evolution in an aggressive papillary thyroid cancer and matched metastases

Soazig Le Pennec[1], Tomasz Konopka[1], David Gacquer[1], Danai Fimereli[1], Maxime Tarabichi[1], Gil Tomás[1], Frédérique Savagner[3,4], Myriam Decaussin-Petrucci[5], Christophe Trésallet[6], Guy Andry[7], Denis Larsimont[7], Vincent Detours[1,*] and Carine Maenhaut[1,2,*]

[1]IRIBHM, [2]WELBIO, Université libre de Bruxelles (ULB), Campus Erasme, 808 Route de Lennik, 1070 Brussels, Belgium
[3]CHU d'Angers, Bâtiment IRIS, 4 rue Larrey, Angers F-49033, France
[4]EA 3143, Université d'Angers, F-49033 Angers, France
[5]Service d'Anatomie et Cytologie Pathologiques, Centre de Biologie Sud – Bâtiment 3D, Centre Hospitalier Lyon Sud, 69495 Pierre Bénite Cedex, France
[6]Hôpital Pitié-Salpêtrière, Université Pierre et Marie Curie, 47 Boulevard de l'Hôpital, 75013 Paris, France
[7]Institut Jules Bordet, 121 Boulevard de Waterloo, 1000 Brussels, Belgium
[*](V Detours and C Maenhaut contributed equally to this work)

Abstract

The contribution of intratumor heterogeneity to thyroid metastatic cancers is still unknown. The clonal relationships between the primary thyroid tumors and lymph nodes (LN) or distant metastases are also poorly understood. The objective of this study was to determine the phylogenetic relationships between matched primary thyroid tumors and metastases. We searched for non-synonymous single-nucleotide variants (nsSNVs), gene fusions, alternative transcripts, and loss of heterozygosity (LOH) by paired-end massively parallel sequencing of cDNA (RNA-Seq) in a patient diagnosed with an aggressive papillary thyroid cancer (PTC). Seven tumor samples from a stage IVc PTC patient were analyzed by RNA-Seq: two areas from the primary tumor, four areas from two LN metastases, and one area from a pleural metastasis (PLM). A large panel of other thyroid tumors was used for Sanger sequencing screening. We identified seven new nsSNVs. Some of these were early events clonally present in both the primary PTC and the three matched metastases. Other nsSNVs were private to the primary tumor, the LN metastases and/or the PLM. Three new gene fusions were identified. A novel cancer-specific *KAZN* alternative transcript was detected in this aggressive PTC and in dozens of additional thyroid tumors. The PLM harbored an exclusive whole-chromosome 19 LOH. We have presented the first, to our knowledge, deep sequencing study comparing the mutational spectra in a PTC and both LN and distant metastases. This study has yielded novel findings concerning intra-tumor heterogeneity, clonal evolution and metastases dissemination in thyroid cancer.

This study sought to characterize the intratmoural heterogeneity of a specimen of aggressive papillary thyroid carcinoma, and the clonal relationships between the primary tumour and their corresponding lymph node and distant metastases.[232] The author contributed with experimental design input and results discussion.

[232] Pennec et al., 2015

Discussion

Microarrays

Microarray technology, through the simultaneous assessment of the expression of thousands of genes, became the first molecular biology tool capable of addressing the complex polygenic nature of cancer.[233] Genomic perturbations drive cancer progression by disturbing mechanisms for cell cycle control, differentiation, DNA repair, apoptosis, tumour vascularization, and metabolism. The monitoring of gene expression signatures—as surrogates of biological processes—in molecular profiles of cancer biospecimens, can be used to investigate how these mechanisms are impacted during cancer progression.

[233] Grant et al., 2004

In this dissertation, we present the result of two analyses relating to the use of gene expression signatures as biomarkers in cancer research. The first concerns the use of differentiation and proliferation signatures in cancer diagnostic; the second regards the extent of prognostic signals in cancer transcriptomes. Here we offer a discussion of these contributions within the broader context of the use of microarray technology in clinical oncology. This section will then conclude with some remarks on the challenges in the analysis and interpretation of microarray data.

Differentiation and proliferation signatures in cancer diagnostic

Molecular classification of cancer is a common diagnostic problem in clinical oncology.[234] The problem consists in assigning tumours to known taxonomic classes based on their expression profiles and is framed as a supervised learning prediction task. The conventional approach involves training a moleclular classifier in a group of labeled samples and then assess its performance on an independent set of unlabeled samples.[235]

[234] Golub et al., 1999; Alizadeh et al., 2000; and Bullinger et al., 2004

[235] Golub et al., 1999

This approach rests on the assumption that the core molecular features that specify tumour classes are tractable by direct comparison of their expression profiles. While this is an established evidence in some cancer models,[236] in other case studies, molecular diagnostics of clinical sub-types is less consensual.[237] This may be explained in part by technical variance, e.g., data overfitting or different investigators using different experimental methodologies.[238] Additionally, biological variance, in the form of noise due to sampling heterogeneity or erratic patterns of tumour evolution, may also condition the stability of classifiers derived from inter-class comparisons.

[236] Haibe-Kains et al., 2012; and Markert et al., 2011

[237] Travis et al., 2013; and Nikiforov and Nikiforova, 2011

[238] Weigelt et al., 2012

We sought to approach this classification problem from a different perspective. Instead of relying on the *intrinsic*, *variant* features that appear contrasted between samples of different classes, we addressed the task by enrolling the *extrinsic*, *invariant* features of biomarkers for two core processes of multicellular life: differentiation and proliferation. As cancer progression is defined by an increase in proliferation rates and a concomitant decrease in tissue differentiation (Figure 20), we reasoned that mapping the expression profiles of distinct tumour sub-types along these two continuums would result in a classification procedure that is more resistant to technical and biological idiosyncrasies.

To test this idea, two requirements had to be met. First, we needed a cancer model characterized by a well defined linear progression, from benign, differentiated tumour types, to aggressive, anaplastic ones—along which taxonomic tumour classes could be sensibly represented. Second, we needed a robust method to define molecular differentiation signatures from expression profiles of healthy tissues, and a dependable proliferation metagene.

As a case study, we took to thyroid cancer. Thyrocyte-derived carcinomas are broadly divided into well-differentiated, poorly differentiated and undifferentiated types on the basis of histological and clinical parameters (Figure 5).[239] Among the well differentiated thyroid carcinomas are the papillary and follicular types. The anaplastic thyroid carcinoma, at the other extreme of the dedifferentiation continuum, is a highly aggressive and lethal tumour (Figure 21).

Especially fitting to test our classification procedure is the distinction between follicular adenomas and follicular carcinomas; and the distinction between follicular variants of papillary carcinomas and their classical counterpart. These challenging pathological diagnostics[240] are critical from the prognostic point of view. In each case, while the former types behave in an indolent manner and have a good prognosis, the latter are defined as poorly differentiated thyroid carcinomas, and may evolve to develop a malignant phenotype.

Several methods exist to quantitatively measure cell proliferation in biological samples, such as bromodeoxyuridine incorporation, Ki-67 or proliferating cell nuclear antigen (PCNA) immunostaining. Conversely, the differentiation state of a cell is commonly defined by a range of qualitative morphological and physiological parameters. Underlying these phenotypic traits, at the molecular level, are tissue specific expression patterns that define the degree of structural and functional specialization of their cellular types.

To investigate these expression patterns, we devised an agnostic method to select for genes that are consistently highly expressed on a cell type of choice, but among the least expressed in other tissue types. This was formulated by selecting for genes among the 1000 most expressed in the tissue of choice, that were not among the top 5000 most expressed in an assortment of other tissue types. This algorithm was applied to a dataset of sixteen RNA sequencing profiles of healthy human organs.[241] Compared to microarray expression profiling, RNA sequencing technology estimates mRNA expression with read counts normalized by transcript length, therefore reflecting

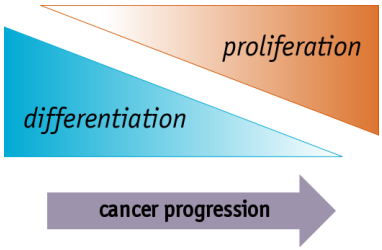

Figure 20: A schematic representation of the inverse relationship between tissue differentiation and proliferation in cancer progression (see text for details).

[239] Kondo et al., 2006

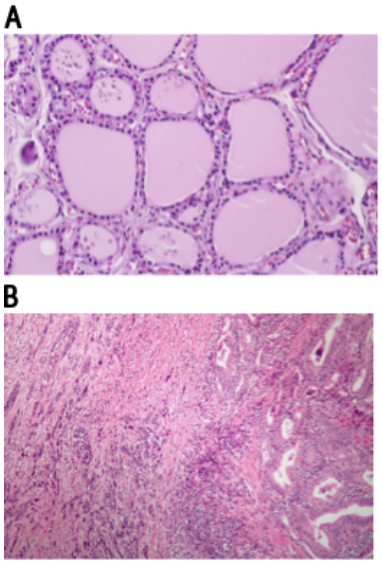

Figure 21: In clinical pathology, the loss of tissue differentiation and increase in proliferation is captured by the concept of neoplastic grading. While cancers with fair prognosis are said to be differentiated, cancers with poor prognosis are referred to as anaplastic. **A:** Micrograph of a low magnification thyroid tissue. The functional units of the thyroid gland are the thyroid follicles, lined by an epithelium of thyrocytes. Thyrocytes delimit the follicular lumen, where the colloid serves as a reservoir for thyroglobulin. **B:** Micrograph of an anaplastic thyroid carcinoma, a stage IV thyroid tumour. These tumours have a high mitotic rate and are among the human tumours with the poorest prognosis. Notice the degree of structural tissular disorganization compared with the tissue of origin.

[240] Lubitz et al., 2005

[241] BodyMap, 2012

more accurately absolute transcription levels.[242] This simple learning algorithm yielded a list of eight thyrocyte specific genes (Table 5).[243]

[242] Wang et al., 2009

[243] The list includes the genes *CRABP1*, *FOXE1*, *IYD*, *PTH*, *SLC26A7*, *TG*, *TPO*, and *TSHR*.

The thyroid differentiation biomarker was then projected in the feature space of two *Affymetrix* microarray platforms, U95Av2 and U133v2. Because *Affymetrix* platforms often bear several probesets targeting for a specific gene, we selected, in two reference compendia of healthy tissues profiled with each platform, the probeset that maximizes the thyroid-specific signal for each gene in our signature. A thyroid differentiation index, dubbed *t-index*, could then be derived from biological samples profiled in any of these platforms, by computing the median expression of the respective selected probesets.

Human BodyMap 2.0 tissue	Number of tissue-specific genes
adipocytes	0
adrenal gland	0
blood	19
brain	96
breast	5
colon	5
heart	13
kidney	18
liver	101
lung	14
lymph nodes	0
ovary	5
prostate	5
skeletal	11
testes	68
thyroid	8

Table 5: Size of tissue differentiation signatures. Tissue-specific differentiation signatures were derived by selecting for genes that are among the most expressed in the tissue of choice, and among the least expressed in the remaining 15 tissue types (see text for details). The size of the signatures reflects the degree of structural and functional specialization of that organ. Tissues for which no gene met the selection criteria are likely to have a less particular metabolism (adipocytes) or to represent a mix of different cell types (adrenal gland and lymph nodes).

A biomarker of proliferation was similarly derived from expression profiles of healthy tissues. The proliferating cell nuclear antigen (PCNA) is a cofactor of DNA polymerase δ and an essential motif for cell replication. A metagene, called meta-PCNA, was obtained by selecting the 1% genes most positively correlated with the *PCNA* gene in a compendium of expression profiles of normal tissues.[244] This metagene consists of 129 genes featuring many significant cell-cycle related genes, like *AURKA*, *MKI67*, *TOP2A*, or *MCM2*. A meta-PCNA index can similarly be derived from expression profiles of biological samples by computing the median expression of the genes of this proliferation biomarker.

[244] Venet et al., 2011

We then evaluated the expression of these two biomarkers in three datasets of normal and neoplastic thyroid expression profiles. The first, hybridized on an *Affymetrix* U133v2 chip (GSE29265), comprised a selection of 49 samples, including anaplastic thyroid carcinomas (ATCs) and papillary thyroid carcinomas (PTCs), paired with their respective adjacent normal tissues. The second, hybridized on *Affymetrix* U95Av2 (GSE29315), included a total of 71 samples spanning normal thyroids, follicular thyroid adenomas and follicular thyroid carcinomas (FTAs and FTCs); altogether with classical papillary thyroid carcinomas and follicular variants of papillary thyroid carcinomas (CPTCs and FVPTCs). The third dataset[245] comes from a kinetic time course study profiling primary cultured thyrocytes with

[245] van Staveren et al., 2006

thyroid-stimulating hormone (TSH), and was hybridized on a home made platform interrogating nearly 4000 genes. A well characterized response of thyrocytes in culture to TSH stimulation is an increase both in metabolic activity (differentiation), as well as in mitotic activity (proliferation).

By quantifying the *t-index* and the meta-PCNA index in these three datasets, we were able to establish that, (*a*) the two indices are negatively correlated in a range of thyrocyte-derived tumours of increasing aggressiveness, yet positively correlated in a time course experiment of TSH stimulation of thyrocytes; (*b*) the *t-index* can accurately discriminate between expression profiles of FTAs when compared with FTCs; and between expression profiles of FVPTCs when compared with CPTCs; and (*c*) the performance of this differential diagnosis classifier is as robust as a classifier derived by training a SVM learning algorithm (validated with a repeated inner/outer cross-validation procedure) on the whole expression space of the labeled samples.

Because defects in cell-cycle regulation are the defining feature of neoplastic pathogenesis, genes participating in proliferation are often found highly expressed in tumour microarrays when compared with normal samples. In spite of the many discordant proliferation gene lists proposed in the literature,[246] increased expression of most of these biomarkers has often been linked with poor clinical prognosis.[247] Proliferation is a universal and conserved theme of cancer transcriptomes,[248] and frequently accounts for most of the power driving the performance of prognostic signatures.[249] Proliferation biomarkers are thus a major component of genomic-based clinical diagnostics for cancer patients.

[246] Whitfield et al., 2006

[247] Dai et al., 2005a; Paik et al., 2004; Rosenwald et al., 2003; and Sørlie et al., 2001

[248] Rhodes et al., 2004

[249] Solé et al., 2009; Venet et al., 2011; and Wirapati et al., 2008

The identification of other potential tumour class-specific genes, related to the particular biology of each taxonomic group, is performed by training learning algorithms on labeled expression profiles. The validity of the selected features, along with the mathematical function used to predict tumour class based on their vector of expression, is then assessed by testing the accuracy of the model in unlabeled samples.[250] This methodology can however be hindered by a number of issues.[251] The classifier may reflect the inherent technical and biological biases specific to the training set, rather than capturing the modulations underpinning class specification—the model is then said to overfit the training set. Moreover, even when models are trained in large enough datasets and proper care is taken to ensure independent validation, the unstable nature of cancer genomes may itself obscure class-specific biological signals by adding random variance to expression measurements.

[250] Simon, 2003

[251] Brenton et al., 2005

Here, we investigate the possibility of enlisting another classifier, external to cancer biology, to address the problem of thyrocyte-derived tumour class prediction. By identifying genes whose expression rank highly exclusively in thyroid compared to other tissues, we defined a molecular differentiation marker that takes no *a priori* concerning the biology of the thyrocyte. Unsurprisingly, this marker contains important genes in thyroid physiology, including genes coding for: thyroglobulin (a dimeric protein used to produce thyroid hormone); thyroid peroxidase (a membrane-bound glycoprotein that takes part in the iodination of the precursors of thyroid hormone); iodotyrosine deiodinase (an enzyme that facilitates the iodide salvage post-

thyroid hormone synthesis); the thyrotropin receptor (a receptor activated by TSH); and the TTF2 transcription factor (active in the developing thyroid, with a role in controlling the onset of its differentiation[252]). The signature also comprises genes coding for the parathyroid hormone (parathyroid cells were likely to be present in the resected profiled biospecimens); cellular retinoic acid binding protein 1 and a member of the solute carrier family 26 (both proteins for which no known role in thyroid has been identified to date).

[252] Zannini et al., 1997

Noticeably, in diagnostic immunochemistry, antigens for thyroglobulin and thyroid peroxidase are already routinely used to provide a quantitative assessment of the malignancy of neoplastic thyroid lesions.[253] Our biomarker of thyrocyte differentiation, the *t-index*, was used as a classifier to accurately disentangle two challenging pathological diagnoses, FTA *vs* FTC and FVPTC *vs* CPTC, based on their expression profiles (*cf.* Figure 4*a*,*d* of the article in the Results chapter). The *t-index* can also be used to chart a linear progression between thyroid neoplasias of increasing malignancy (*cf.* Figure 2*a*,*b* of the article in the Results chapter). Papillary and follicular carcinomas, the two main taxonomic classes of thyroid tumours, are shown to depart from thyrocytes along two distinct axes of gene expression (*cf.* Figure 3 of the article in the Results chapter), suggesting that our biomarker correlates with the molecular courses of thyroid cancer progression.

[253] Tanaka et al., 1996; and Gérard et al., 2003

Furthermore, we show that the *t-index* and the meta-PCNA index are negatively correlated in expression profiles of a panel of thyroid cancers (*cf.* Figure 2*a*,*b* of the article in the Results chapter), yet positively correlated in an *in vitro* experiment of physiological and mitotic thyrocyte activation (*cf.* Figure 2*c* of the article in the Results chapter). This provides supporting evidence for the independence of our classifier from proliferation, unlike potential classifiers derived from the expression space of labeled tumour samples. Nonetheless, class prediction of thyroid tumours based on the *t-index* is as accurate as a state-of-the art machine learning algorithm selecting for the optimal classifying genes from the entire set of features spotted in the arrays (*cf.* Figure 4*c*,*f* of the article in the Results chapter).

This proof of concept brings into a quantitative framework the formulation that neoplastic progression can be mapped along an axis of molecular dedifferentiation. Thyroid cancer is especially suited to test this hypothesis, as it comprises a family of mostly indolent tumours that evolve slowly (Figure 5), providing ample opportunity for sampling malignancies at different stages of progression. In addition, biospecimens of thyrocyte-derived cancers, while populated by a range of distinct cellular types (fibroblasts, C cells, stromal cells), are mostly dominated by thyroid epithelial cells—an essential condition for the biological bearing of our differentiation signature in these expression profiles.

The generalization of this approach to other cancer families may depend on both the feasibility to derive stable markers of differentiation for the cell type of origin, as well as on the tractability of those markers in respective cancer biospecimens. Breast cancer, for instance, is typically derived from epithelial cells of the mammary gland, yet the expression profiles of most breast cancer biospecimens are largely dominated by adipocytes. On the

other hand, cancers derived from the hematopoietic lineage could prove particularly amenable to our strategy, as expression profiles of purified populations of human hematopoietic cells at different stages of maturation are readily available in the literature.[254]

While the use of biologically motivated molecular markers to dissect cancer genomes is not a novelty,[255] to our knowledge, ours is the first study to enroll biomarkers of differentiation to guide the interpretation of cancer expression profiles. We devised a simple method to catalogue genes that are only significantly expressed in a particular tissue type, potentially seizing fundamental tissue-specific expression patterns. We then showed that a thyroid-specific biomarker can be reliably used to map the molecular progression of two distinct taxonomies of thyrocyte-derived neoplasias, and is able to solve two challenging pathological diagnoses. These results raise the possibility that the process of molecular dedifferentiation, a collateral of cancer progression, could be accurately quantified and modeled for prospective use in clinical oncology.

Our methodology is not without shortcomings. Not all genes in the thyroid differentiation signature are pertinent to the class prediction problem. For instance, the expression of the TSH receptor gene is usually preserved during cancer progression,[256] and the expression of the parathyroid hormone gene is irrelevant to the thyrocyte's biology. More sophisticated learning methods, along with more stringent techniques of cell type isolation prior to profiling, might help to alleviate these imprecisions. Additionally, for each of the undertaken classification tasks, we do not provide a proper cut-off point to translate the quantitative predictive *t-index* into a predicted class label. This is because our experimental setting did not include a sufficient number of samples of each class to estimate a sensible cut-off point for the classifier. In both cases, the accuracy of the classifier was estimated by stipulating all possible cut-off points and computing the respective area under the ROC curve (*cf.* Figure 4*b*,*e* of the article in the Results chapter).

Still, when addressing tumour class prediction, classifiers based on differentiation signatures have the advantage of being stable and extraneous to the biology of the disease, and thus virtually immune to biological variance motivated by cancer genome instability or overfitting biases. Their potential to assist diagnostic tasks in cancer oncology remains to be fully explored.

[254] Novershtern et al., 2011

[255] Sund and Kalluri, 2009; and Dave et al., 2004

[256] D'Agostino et al., 2014

The extent of prognostic signals in the cancer transcriptomes

Disease outcome prediction is another important challenge in clinical oncology.[257] The problem consists in uncovering biomarkers whose expression in cancer biospecimens have a predictive ability regarding disease outcome. It can also be conceived as a supervised learning prediction task, with a similar strategy to the class prediction problem. A molecular classifier is derived from a training set of samples with distinct survival times with respect to a clinical event; its predictive ability is then validated with survival analysis on an independent set of samples with associated clinical follow-up data (Figure 22).

This methodology has identified cancer prognostic and predictive signatures with superior performance to conventional histopathological or

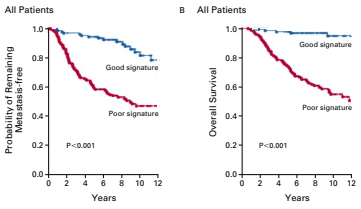

Figure 22: Validation of a genomic marker's predictive ability. A predictive 70-gene signature was derived from a prospective cohort of 98 expression profiles of breast cancer with known survival times (van't Veer et al., 2002). This prognosis-classifier was subsequently used to segregate good and bad prognosis groups in an independent cohort of 295 breast carcinomas. Differential outcome between each group, regarding likelihood of developing metastasis (panel **A**), or likelihood of dying (panel **B**), was then established with Kaplan-Meier analysis (*adapted from* Van De Vijver et al., 2002).

[257] Van De Vijver et al., 2002; Vasselli et al., 2003; and Sanchez-Carbayo et al., 2006

clinical parameters.[258] Additionally, it offers a framework to test for the implication of particular biological processes in cancer progression, through the quantification of the association of their surrogate transcriptional markers with clinical outcome.[259] This formulation assumes that the expression patterns prompting cancer progression are universal motifs, yet specific enough to be modeled by biologically motivated molecular markers.

[258] Solé et al., 2009

[259] Chang et al., 2004

We sought to test this assumption by characterizing the range and the nature of prognostic transcriptional signals in a representative selection of human cancer expression profiles. In order to do so, we compiled 114 expression profile studies of cancers afflicting 19 organ systems (Table 6), with clinical follow-up data. The transcriptomes of these nearly 22 000 neoplastic samples were profiled with over 30 different commercial and custom microarray platforms (Table 3).

Tissue of origin	Number of cancer studies
breast	33
blood cells	12
central nervous system	11
colon	9
lung	8
ovary	8
bladder	6
lymphatic system	5
kidney	3
liver	3
prostate	3
uterus	3
bone	2
squamous cells (mouth, nose or throat)	2
skin	2
adipocytes	1
pancreas	1
stomach	1
thyroid	1

Table 6: Nearly 22 000 expression profiles of cancer biospecimens, across 114 studies, were compiled from the public domain to quantify the extent of prognostic signals in cancer transcriptomes. Among these, experiments profiling breast tumours, with 33 studies, were the most represented. This reflects the prevalence of breast cancer in the population (Figure 1), and its pivotal role as the original model for genomic outcome prediction analyses. Next, in terms of representation, are studies profiling cancers originating from the hematopoietic lineage (12) and central nervous system (11)—perhaps a reflex of the ability to isolate relatively uncontaminated populations of tumoural cells from these cancer types—, followed by colon (9) and lung (8)—two of the most prevalent cancers worldwide. See Table 3 for a thorough description of each dataset used in this meta-analysis.

To gauge the extent of prognostic signals in each of these datasets, we tested association with outcome of each of the 4722 biologically motivated gene expression signatures in the C2 collection of the Molecular Signatures Database (MSigDB C2), curated by the Broad Institute. This collection includes gene sets collected from various sources such as online pathway databases, publications in PubMed, and knowledge of domain experts.

Estimating association between a genomic marker and clinical outcome requires the formulation of an outcome predictor from multi-gene values of expression. Traditionally, this is achieved by a function that stratifies the cohort in good and bad prognosis groups according to a cut-off point for the expression of the gene classifier (Figure 22). Instead, we chose to formulate association with outcome by modeling survival time as a function of the first principal component of each expression signature. The logrank p-value of the corresponding Cox proportional model was used to assert the prognostic value of the biomarker.

To control for an over-representation of cancer-related themes within the MSigDB C2 collection, we replicated the analysis with a synthetic collection

of signatures of the same size, but made up of genes randomly selected from the human genome. Finally, we tested association of single features in every dataset with outcome, by modeling survival times as a function of individual vectors of expression in the expression matrices.

Figure 13 reports the proportion of tested MSigDB C2 signatures, randomized signatures, and single features found associated with outcome in each dataset at a logrank $p < 0.05$. These quantities capture the fraction of each transcriptome bearing information regarding differential patient survival. From the statistical point of view, they can be interpreted as likelihoods of finding a significant genomic association with outcome by chance alone. We thus refer to these estimates as the baseline prognostic content[260] of cancer transcriptomes.

[260] Prognostic content will henceforth refer to the fraction of tested markers found associated with outcome in a given dataset.

Critically, in 100 of the 114 datasets analyzed (88%), more than five percent of single-gene markers show a significant association with outcome. When considering multi-gene markers, up to 87 datasets (76%) registered more than five percent of random signatures associated with outcome. Overall, the median prognostic content across all studies for single-gene markers and random multi-gene markers was, respectively, 12% and 16%. These observations suggest that, in most cancer transcriptomes, prognostic signals are disseminated throughout a nontrivial fraction of their expression features; that the likelihood of finding random associations with outcome is globally non-negligible; and that the nature of prognostic signals is not marker-specific.

A compelling result of this analysis is the considerable heterogeneity of prognostic fractions observed across surveyed cohorts (Figure 13). This observation is not cancer-specific. Among breast cancers, for instance, single-marker prognostic content alone ranged from 4% to 60%. To examine the contributions of potential biological and demographic dataset-specific factors to this effect, we quantified single-marker prognostic fraction for resamplings of the 1972 transcriptomes of the two METABRIC breast cancer cohorts, regarding modulations of four variables: sample size, duration of follow-up time, fraction of ER+ patients, and fraction of node positive patients (Figure 14).

The estimates of prognostic content are markedly sensitive to sampling variance, as suggested by the dispersal of the distribution of estimates (confidence interval: 12% to 23%), when 100 samplings of 500 random transcriptomes were examined for the fraction of genes associated with outcome (Figure 14*a*). This feature alone is likely to yield a significant contribution to the range of estimates computed in our meta-analysis, as 108 of the 114 datasets analyzed included less than 500 profiled tumours. Moreover, the sensitivity of our estimation procedure is, as expected, largely dependent on the number of profiles included in the analysis, as shown by a bootstrapping experiment of sample sizes towards the assessment of the prognostic fraction (Figure 14*b*). Provocatively, the estimates of prognostic content do not appear to level off even when the experimental sample size reached 1750—by far the largest in the field. A sampling experiment of sequential truncation of follow-up times was equally shown to impact estimates of prognostic content (Figure 14*c*). Sharply increasing predictive fractions were observed up to the fifth year of follow-up, followed by a gradual decrease of the esti-

mates for higher follow-up times. This trend suggests that, in breast cancers, prognostic patterns of expression are optimally correlated with short-term forms of progression of the disease, and that long-term forms of progression are less efficiently predicted from primary tumour transcriptomes. The modulation of the fraction of ER+ transcriptomes towards experimental samplings of our estimate has exposed a tendency congruent with the clinical relevance of this receptor in breast cancer pathology (Figure 14*d*). Thus, an increase of ER+ profiles to up to 50% in our samplings leads to a corresponding linear rise in estimates of prognostic content, at which point a further increase in the proportion of ER+ profiles yields little impact on fraction estimates. This observation is in line with the fact that the predictive power of most signatures in breast cancer is mostly confined to ER+ phenotypes.[261] Finally, a sampling experiment with dosed fractions of profiles from node positive patients (Figure 14*e*), has also uncovered a pattern of prognostic fraction estimates that can be largely explained by the fact that nodal status is clinically correlated to ER status in breast cancer.

[261] Weigelt et al., 2012

A large number of studies has interpreted the statistical association of transcriptional markers with clinical outcomes as evidence that their underlying biological mechanisms are involved in cancer progression. This approach has gained widespread appeal because of the increasing availability in the public domain of expression profiles of cancer biospecimens with associated survival data; and because it obviates the need for time-consuming and costly experimental setups on *in vivo* models.

In order for a molecular marker to be recognized as a prognostic factor, it has to meet three successive criteria: its *specific* association with clinical outcomes has to be asserted; its prognostic performance and accuracy has to be validated in an independent group of patients; and the independence of its prognostic value from other potential factors has to be established with a multivariate analysis.[262]

[262] Chibon, 2013

The assumption that prognostic signatures reported in the literature are specific conveyors of biological signals pertaining to cancer progression was previously challenged by our research group and others.[263] For instance, Venet et al. have shown that, in two reference breast cancer cohorts, most published signatures are significantly prognostic, yet no more prognostic than random sets of genes.[264] Here, we provide a global assessment of the pervasiveness of prognostic signals in human cancer transcriptomes.

[263] Venet et al., 2011; Lauss et al., 2010; Ein-Dor et al., 2005; and Mosley and Keri, 2008

[264] Venet et al., 2011

The chief observation issued from our analysis is the wide heterogeneity of prognostic contents measured across the examined cancer cohorts. When estimated by the fraction of single gene-markers associated with outcome, our assessment of transcriptomic prognostic content ranged from 2% to 60%; whereas a fraction of 4722 random multi-gene markers ranging from 0.1% to 94% was shown to inform patient differential survival across investigated datasets. Interestingly, in most datasets, the fraction of randomized signatures associated with outcome was only marginally inferior to the fraction of biologically motivated prognostic MSigDB C2 signatures (median difference: 2%; maximum difference: 17%). In 37 out of the 114 analysed datasets, the fraction of prognostic random signatures was higher than the fraction of biologically motivated ones.

Our estimate of prognostic content is notoriously sensitive to sampling

variance, an observation consistent with previous reports.[265] This variance may be accounted for in part by a range of dataset-specific demographic factors—such as duration of follow-up times or cohort size—, as well as by cancer-specific biological covariates. Furthermore, in at least one dataset (GSE9893, single-gene marker prognostic content: 59%), we have identified a critical normalization artifact responsible for an artificial over-estimation of the prognostic fraction. Upon re-normalization of the original arrays, we have re-evaluated our estimate of prognostic content down to 19%; conversely, a metric of signal-to-noise ratio,[266] surged from 15% to 34% (a score now within the inter-quartile range of the 114 datasets analysed). This precedent suggests that artefactual spurious correlations in expression matrices may account for significant portions of prognostic content in our observations.

[265] Ein-Dor et al., 2005

[266] Venet et al., 2012

The scope of this analysis remains strictly epistemological. While we have observed that, in the thirteen breast cancer datasets with OS survival data, the single-features with highest scores of association with OS are largely overlapping (excluding GSE9893), the validation of cancer-specific, biologically pertinent molecular markers from microarray data is still a delicate task. In any event, a more stringent formulation of experimental controls is essential to the validation of the predictive ability of biologically motivated biomarkers.[267] Feature transformation techniques might further be required to address spurious structures of correlation and assist to the dissection of additional biological components in genomic prognostic signals.

[267] Beck et al., 2013

Taken together, these observations attest for the pan-transcriptomic nature of prognostic signals in cancer expression profiles—as quantified by association with outcome of single- and multi-gene markers with a Cox proportional hazards model, at logrank $p < 0.05$. This conclusion is not cancer- or platform-specific. Therefore, with current significance thresholds, transcriptomic prognostic signals cannot be convened to infer specific biological mechanisms driving cancer progression. While molecular prognostic signals remain independent, and thus complementary, to clinicopathological parameters, their ubiquity is in agreement with the multitude of incongruous predictive signatures reported in the literature.[268] In addition, the landscape of prognostic signals uncovered by our analysis may account for the complexity of some challenges in clinical oncology—including the difficulty in finding stable molecular predictors of response to specific systemic treatments in breast cancer, for instance.[269] We envisage the phenomena here reported to be of significance for the interpretation of global expression patterns of neoplastic samples; and to have a bearing in the development of next generation genomic predictors.

[268] Gevaert and De Moor, 2009; and Chibon, 2013

[269] Reis-Filho and Pusztai, 2011

Microarray data analysis and interpretation

As microarray technology matured from the original double channel, cDNA-based chips,[270] to the commercial single-channel, *in situ* synthesized and high density oligonucleotide spotted arrays,[271] so did their associated analytic methodologies, along with their data interpretation.

[270] Schena et al., 1995

[271] Lockhart et al., 1996

Prior to any analytic processing of microarray data lies the issue of probe

annotation. While initial cDNA microarrays used reverse-transcribed copies of isolated gene fragments as probes, high density spotted arrays derived their probes from genomic or EST sequences.[272] Each subsequent revision of the human reference genome carries with it the reassignment of a subset of probes in the array to new gene products, therefore potentially upsetting previous conclusions based on former gene annotations. For instance, Bioconductor software annotation packages for most commercially available microarray chips are still updated every six months, following reviewing cycles of the human reference assembly genome.

[272] Expressed sequence tags, or ESTs, are transcribed mRNAs from a tissue type used to produce genome assemblies.

More complexity arises from the fact that not all probes in an array have a designated gene associated to them; other probes may recognize multiple target sequences, known as promiscuous probes; and some genes may have several probes assigned to them, each of which interrogating a unique mRNA transcript. Several strategies exist to collapse the expression of distinct probes targeting the same gene into a single value. The most common include computing the arithmetic mean of redundant probes or retaining the one exhibiting the highest value of expression (or, conservatively, the lowest). Miller et al. have reviewed a collection of strategies for aggregating gene expression data when applied to different genomic applications.[273]

[273] Miller et al., 2011

Specific to the *Affymetrix GeneChip* platform is the issue of summarizing probe expression into probesets. The set of rules used to perform this transformation is defined by a Chip Description File, or CDF, containing the mappings of which probes are to be aggregated into a single probeset. Alternative custom CDF files have been shown to provide both better precision and accuracy in probeset estimates when compared to the original Affymetrix definitions.[274]

[274] Gautier et al., 2004; Carter et al., 2005; Dai et al., 2005b; Sandberg and Larsson, 2007; and Upton et al., 2009

Microarray analysis methodologies are notoriously challenged by *the curse of dimensionality*,[275] a feature of highly-dimensional data where the number of measured variables largely exceeds the number of samples. One of the consequences of the curse of dimensionality is an increased likelihood of detecting false positive models due to chance alone. This impacts both the task of finding differentially expressed genes between phenotypic classes and the task of building predictive models.

[275] Bellman and Bellman, 1961

For the gene selection task, initial assessments of differential expression based on fold-change—defined as the ratio in expression means between two groups—, were rapidly supplemented by more robust models based on *t*-statistics. Instead of just relying on within-gene comparisons, these methods attempted to exploit the between-gene information in the array by weighting the *t*-statistic with global variance estimates—a procedure known as variance shrinkage. Example of such methods include the popular significance analysis of microarrays, or SAM,[276] and more sophisticated approaches that seek to model between-gene relationships with empirical Bayes methods.[277]

[276] Tusher et al., 2001

[277] Newton et al., 2001

To control for the excess of false positive models resulting from the highly-dimensional experimental design, multiple-testing correction methodologies are required. Among these are restrictive family-wise error rate (FWER) correction methods, which proceed by strictly limiting the probability of producing type I errors to less than a pre-determined significance level α across the entire experiment. This initial class of correction

procedures was later replaced by the more flexible false-discovery rate,[278] which aims to control the expected proportion of incorrectly rejected null hypothesis in a list of validated models.

[278] Benjamini and Hochberg, 1995

Nevertheless, strategies for gene selection between classes are further compounded by coarse definitions of biological phenotypes that translate ineffectively at the molecular level;[279] by a poor understanding of the dynamic ranges of gene expression in physiological settings;[280] and by the unlikely assumption that expression measurements are atomic quantities behaving independently of each other.[281] Indeed, it was the acknowledgment of the intricate nature of expression profiles, with groups of genes coordinately expressed as pathway components, that paved the way to more sophisticated analytic methods for microarray data analysis.

[279] Piatetsky-Shapiro and Tamayo, 2003

[280] Nadimpally and Zaki, 2003

[281] Piatetsky-Shapiro and Tamayo, 2003

The analysis of perturbations in the expression of co-regulated genes between conditions is denominated knowledge base-driven pathway analysis. It introduces a framework to interpret the underlying biology of differentially expressed transcripts, and can contribute to mitigate the curse of dimensionality by reducing the volume of inferences made in the gene expression space.

Khatri et al. discuss three generations of pathway-based analytic approaches to dissect microarray data.[282] First generation methods are based on simple over-representation analysis, and focus on determining the fraction of genes in a particular pathway found among the set of significantly perturbed genes issued from the microarray. Inferences made by over-representation analysis use statistics based on the hypergeometric distribution, binomial distribution, or the chi-square distribution.

[282] Khatri et al., 2012

Rather than making inferences on individual genes meeting a criteria for global differential expression, second generation methods, collectively termed functional class scoring (FCS) methods, aim to uncover weaker but coordinated changes in sets of functionally related genes. An example of this class of methods is the gene set enrichment analysis, or GSEA, whose procedure can be generalized in three steps. First, a gene-level statistic is calculated from the microarray experiment; second, gene-level statistics for all genes in a pathway are aggregated into a single pathway-level statistic; and, third, the statistical significance of pathway-level statistics is assessed. When compared to over-representation analysis, FCS methods introduce refinements at three levels. First, they do not require the arbitrary stipulation of a class of differentially expressed genes; second, they integrate gene measurements to infer patterns of coordinated expression; and, third, they account for the non-independent nature of gene expression, by relying on pathway-level statistics.

Third generation functional methods seek to include topological elements concerning the nature (e.g., inhibition, activation), or the cellular localization (e.g., nucleus, cytoplasm), of the gene product interactions in a given pathway. Such methods, like the signaling pathway impact analysis algorithm[283] for instance, aim for a more realistic modeling of gene co-expression between phenotypes.

[283] Tarca et al., 2009

The consolidation of functional analysis methods shifted the focus of microarray data interpretation from the individual gene to the biological pathway level. To that effect, gene expression signatures, defined as col-

lections of genes whose combined expression is associated with a given phenotype, became the operative units for statistical inference concerning diagnostic, prognostic, and predictive tasks,[284] from expression profiles of cancer biospecimens.

The potential for biologically motivated gene signatures to guide the interpretation of microarray experiments is however bounded by important annotation, methodological and conceptual challenges. The low resolution of current pathway knowledge bases, such as the Gene Ontology, has quickly become a major bottleneck to the elucidation of genomic experiments depicting increasingly refined and complex molecular landscapes. This issue is illustrated, for instance, by our incipient understanding of alternative splicing patterns of gene transcripts, which can result in gene products with related, distinct, or even opposing functionality.[285] Furthermore, advances in genomic technologies have not been met by a corresponding increase in granularity of genomic annotation databases. In fact, a large number of protein-coding genes are still affected by low quality or inaccurate annotations—and some are "inferred from electronic annotations," or yet to be annotated at all.[286] In addition, static annotations are likely to misrepresent the contextual information essential to model the dynamics of cellular physiology, effectively neglecting the highly integrated nature of biological systems.

Ultimately, microarray data analysis proposes the rendition of these layers of biological intricacy through the projection of expression measurements into a linear, non-redundant namespace of features, such as Entrez Gene IDs or HGNC gene symbols—which then provide the starting point for statistical inference.

Taken together, these considerations significantly condition the scope of microarray-based discovery of biological knowledge. It follows that special caution is advised when translating findings issued from the analysis of expression profiles of neoplastic biospecimens into clinical cancer research.

[284] Some imprecision exists in the literature regarding the definitions of the terms "prognostic" and "predictive." Antoine Italiano proposed that a prognostic factor is "a clinical or biologic characteristic that is objectively measurable and that provides information on the likely outcome of the cancer disease in an untreated individual." In contrast, a predictive factor is "a clinical or biologic characteristic that provides information on the likely benefit from treatment (either in terms of tumor shrinkage or survival)" (Italiano, 2011).

[285] Wang et al., 2008

[286] Khatri et al., 2012

Cancer

Cancer research aims to improve the diagnosis and treatment through better disease classification and patient stratification. This allows for the design of therapies that are better targeted to specific cancer subtypes and improve the effectiveness of existing regimens, while reducing their morbidity.

In this dissertation, we present two case studies for the application of gene expression signatures to address supervised tasks of class prediction, and outcome prediction, based on publicly available transcriptomes of neoplastic samples. Each analysis addresses a problem related to the clinical management of oncological pathologies, framed by the analytic corpus developed for microarray data research, and grounded by the current understanding of the biology of cancer.

Molecular classification of cancer

Our first contribution concerns the improvement of molecular diagnostics for tumour classification based on expression profiles of clinical samples. Traditionally, tumour classification rests on morphological characterization and immunohistochemical assessment of tissue-specific antigens. Histological types thusly defined denote distinct clinical behaviours and responses to treatment. Molecular characterization of tumour subtypes aims to increase the accuracy of tumour classification based on features that escape histomorphological assessment alone.

Following the proof-of-principle molecular classification of leukemia by Golub et al. in 1999,[287] molecular cancer classification was generalized to solid tumours by Ramaswamy et al. and Su et al. in 2001.[288] Both studies present an experimental design where an initial number of primary carcinomas (respectively, 144 and 100), spanning most human solid cancers, was used to train a multiclass predictor to determine the anatomical origin of a random test sample, given its expression profile. Reported accuracies of these classifiers ranged from 78% to 83%. Ramaswamy et al. remarked that poorly differentiated tumours are less amenable to molecular classification. They interpreted this observation as evidence for a distinct molecular pathogenesis of poorly differentiated tumours when compared to their well differentiated counterparts. Su et al. noted that, for eleven of the tumour classes they analyzed, a minimal core set of genes is sufficient to accurately discriminate between classes. This suggests that expression of genes particular to the basal physiology and morphology of the respective tissues of origin might be actionable for molecular classification.

[287] Golub et al., 1999

[288] Ramaswamy et al., 2001; and Su et al., 2001

We reasoned that these findings may well be explained by the projection of cancer progression along axes of molecular dedifferentiation and increased proliferation. We then demonstrated that two biomarkers, one of thyroid differentiation and another of proliferation, both derived from healthy tissues, could be used to accurately classify tumour subtypes of thyroid origin.[289]

[289] Tomás et al., 2012

In spite of a promising decade for the translation of these research findings into relevant diagnostics with an impact on clinical management, a

review of the field by Schnabel and Erlander[290] cites only three commercial molecular cancer classifiers currently available for clinical use (Table 7).

[290] Schnabel and Erlander, 2012

	bioTheranostics Cancer TYPE ID®	Rosetta Genomics miRview met 2™	PathworkDiagnostics® Tissue of Origin
Tissue specimen	formalin-fixed and paraffin-embedded	formalin-fixed and paraffin-embedded	formalin-fixed and paraffin-embedded
Assay technology (biomolecule)	RT-PCR (mRNA)	microarray (miRNA)	microarray (mRNA)
Biomarkers assessed	92	64	2000
Tumour types classified	54	42	15
Accuracy	87%	85%	89%
Tumour sub-classification	Yes	No	No
FDA clearance	No	No	Yes

Table 7: Commercial molecular cancer classifiers currently available for clinical use (adapted from Schnabel and Erlander, 2012).

Of the three, only one attempts to class molecular tumour subtypes; the other two are designed to locate the primary tumour in patients that present with cancers of uncertain origin in the metastatic setting. Only the PathworkDiagnostics® test has met standards for FDA clearance, yet none of these assays has properly challenged traditional classification methods. At best, they have found their way into clinical management as molecular correlates to histopathological findings.

In their expert opinion piece, Schnabel and Erlander suggest that clinical adoption of gene expression-based classifiers has been curbed by the lack of clinical gold standards; and by the potential confounding effect of tumour heterogeneity in molecular evaluations of mixed populations of cells. Molecular classifiers can, however, contribute to a clinical decision whenever pathological assessment is uncertain or conflicting; or even rule out unlikely tumour classes, facilitating the decision process.

Interestingly, they also raise the possibility that recent paradigm shifts in neoplastic disease conceptualization may also contribute for the lack of reach of molecular classifiers in the clinical setting. Namely, the recent increased attention given to the complex nature of heterotypic signalings in the tumour microenvironment,[291] a feature that lies beyond the scope of microarray technology resolution, may well account for the struggle of traditional molecular classifiers to recapitulate neoplastic disease in its essence.

[291] Weigelt et al., 2014

Molecular prognostication of cancer

Cancer prognostication is one of the most challenging tasks in oncology. Here, the consensus is to adhere to formal anatomical staging systems, such as the the TNM staging system,[292] which provide a basis for prediction of survival, choice of initial treatment, and stratification of patients in clinical trials.[293] The discovery of novel subdivisions of traditional tumour classes—defined by exclusive biomarkers, and presenting different clinical behaviours and therapeutic responses—motivated the search for molecular-based models for cancer prognostication.

[292] Sobin, 2003

[293] Ludwig and Weinstein, 2005

Tangible success was achieved when, in early 2007, the FDA cleared

MammaPrint©, the first microarray-based commercial molecular prognostic test for breast cancer.[294] This assay, for node-negative women under 61 years of age with tumors less than 5 cm in diameter, is based on the 70-gene prognosis profile of van't Veer et al.[295]

[294] FDA, 2007

[295] van't Veer et al., 2002

Breast cancer is, arguably, one of the most thoroughly characterized human cancers from the molecular point of view. In the breast cancer model, genomics classification methods have uncovered at least four intrinsic subtypes: the basal-like subtype, which is estrogen receptor negative (ER–) and HER2–; the HER2 subtype, characterized by increased expression of HER2 and of genes mapping to the HER2 amplicon; and two luminal ER+ subtypes—namely the luminal A subtype, characterized by high levels of ER and ER-related genes; and the luminal B subtype, characterized by lower ER levels and high expression of genes implicated in the proliferation process.[296] Evidence based on prognostic gene signatures empirically derived to discriminate between good- and poor-prognosis cancers has established that good prognosis ER+ tumours (luminal A tumours) derive little, if any, benefit from adjuvant chemotherapy. Conversely, other subtypes show a greater sensitivity to multidrug chemotherapy regimens.[297]

[296] Arpino et al., 2013

[297] Weigelt et al., 2012

While great hope was placed on first generation prognostic signatures to replace clinicopathological parameters in therapy decision making, it was eventually shown that these molecular classifiers largely complement the prognostic ability of tumour size and nodal status—two of the metrics encapsulated by the TNM staging system.[298] Furthermore, the prognostic ability of these signatures is mostly restricted to ER+ breast cancer, with negligible prognostic value reported for patients with ER– disease.

[298] Sotiriou and Pusztai, 2009; and Reis-Filho et al., 2010

Microarray-based markers of predictive response to breast cancer chemotherapy have also shown limited success in producing clinically serviceable tests. Weigelt et al.[299] discuss possible causes for this poor translational output. Among them, they cite challenges related to the confounding impact of molecular heterogeneity of disease when deriving prognostic and predictive markers; the lack of resolution of microarray technology at the post-transcriptional level; and the under-explored potential role for dynamic-response markers in predicting response to treatment.

[299] Weigelt et al., 2012

Recent advances in the field include the prospective validation of a 21-gene prospective assay to refine adjuvant chemotherapy withdrawal using the results of Oncotype DX, in hormone receptor positive, HER2-negative and node-negative breast cancer patients;[300] supporting evidence for the role of carboplatin to improve progression-free survival in patients with *BRCA*-mutated tumours when used as first-line therapy for metastatic disease, and demonstration that palbocilib can improve progression-free survival of patients with metastatic disease in both the first-line and second-line settings.[301]

[300] Sparano et al., 2015

[301] Piccart and Gingras, 2016

A large number of breast cancer prognostic signatures has been reported in the last decade.[302] While their overlap is limited, their prognostic ability has been linked with two main components of breast-derived neoplasias' biology: their proliferative activity, and their ER signaling. Nevertheless, the deceptive biological specificity of these biomarkers has inspired a burgeoning literature purporting to report the implication of biological phenomena in breast cancer progression. This was routinely achieved by validating the

[302] Liu et al., 2014

prognostic ability of a biologically motivated gene signature in a cohort of breast cancer expression profiles.[303] Venet et al. have challenged this reasoning by showing that most random gene expression signatures are associated with breast cancer outcome.[304] In this dissertation, we have extended these findings by characterizing the extent of prognostic signals in 114 cohorts of human cancers afflicting nineteen organ systems.[305] We have shown that, in most cancer transcriptomes, the nature of prognostic signals is notoriously pervasive and their assessment is highly sensitive to sampling variance. We describe a breast cancer case study where a sizable fraction of the prognostic signals quantified could be ascribed to spurious correlations related to a critical normalization artifact. We further established that the extent of prognostic signals in transcriptomes of breast cancer are related to the fraction of ER+ neoplastic transcriptomes prognosticated and, to a lesser extent, to the fraction of node positive patients considered in the analysis.

[303] Chang et al., 2004

[304] Their argument was illustrated, via *reductio ad absurdum*, with the demonstration that expression signatures of postprandial laughter and of mice social defeat can both predict overall survival in the 295-sample NKI reference cohort (Venet et al., 2011).

[305] Under revision.

Biomarker	Official gene name	Clinical use[a]	Cancer type	Source type
α-fetoprotein (AFP)	*AFP*	Stg	Nonseminomatous testicular	Serum
Human chorionic gonadotropin (hGC)	*CGB*	Stg	Testicular	Serum
Carbohydrate antigen 19-9 (CA19-9)		Mnt	Pancreatic	Serum
Carbohydrate antigen 125 (CA125)	*MUC16*	Mnt	Ovarian	Serum
Carcinoembryonic antigen (CEA)	*PSG2*	Mnt	Colorectal	Tissue
Epidermal growth factor receptor (EGFR)	*EGFR*	Prd	Colorectal	Tissue
v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog (KIT)	*KIT*	Prd	Gastrointestinal	Tissue
Thyroglobulin	*TG*	Mnt	Thyroid	Serum
Prostate specific antigen (PSA)	*KLK3*	Scn, Mnt	Prostate	Serum
Carbohydrate antigen 15.3 (CA 15.3)	*MUC1*	Mnt	Breast	Serum
Carbohydrate antigen 27.29 (CA27.29)	*MUC1*	Mnt	Breast	Serum
Estrogen receptor (ER)	*ESR1*	Prg, Prd	Breast	Tissue
Progesterone receptor (PR)	*PGR*	Prg, Prd	Breast	Tissue
v-erb-b2 erythroblastic leukemia viral oncogene homolog 2 (HER2-neu)	*ERBB2*	Prg, Prd	Breast	Tissue
Nuclear matrix protein 22 (NMP-22)		Scn, Mnt	Bladder	Urine
Fibrin/fibrinogen degradation products (FDP)		Mnt	Bladder	Urine
Bladder tumor antigen (BTA)		Mnt	Bladder	Urine
High molecular CEA and mucin		Mnt	Bladder	Urine

[a] Stg: staging; Mnt: monitoring; Prd: prediction; Prg: prognosis; Scn: screening.

Table 8: List of the eighteen FDA-cleared protein cancer biomarkers issued from genomic analyses currently available for clinical use (adapted from Pavlou et al., 2013).

Conclusions & Perspectives

In the last twenty years, cancer medicine has been revolutionized by cancer genomics. By the end of the 20th century, microarrays were the first technology to provide a global depiction of the moleclular portraits of cancer. Microarrays are used to measure the relative concentration of nucleic acid sequences in solution. Only ten years later would RNA-seq become affordable and widespread enough to become the assay of choice in cancer transcriptomics.

Here we report two analyses concerning the making of biological inference based on global expression profiles of healthy and neoplastic tissues. The first study describes a proof-of-concept approach about the use of a thyrocyte differentiation biomarker, devised from profiles of a panel of healthy human tissues, in order to dissect cancer progression in a thyroid carcinogenesis model. We showed that the thyroid differentiation biomarker is independent of proliferation and that it can discriminate between clinically challenging diagnoses in thyroid pathology.

The second study is an epistemological take on the use of survival analysis to validate implication of particular biological mechanisms in cancer progression. We showed that, in a selection of 114 cancer datasets with survival follow-up data, prognostic fractions for multi-gene markers is larger than 20% in roughly half of the cohorts analyzed, a figure suggestive of the pervasive nature of prognostic signals in the cancer transcriptome. Furthermore, we demonstrated that prognostic signals are highly heterogeneous, and that technical, demographic, statistical and biological variables are linked with this disparity.

These analyses seek to investigate the extent to which the cancer transcriptome can be dissected in order to (*a*) address the biological underpinnings of the disease; and (*b*) to improve diagnostics and treatment decisions in the clinical setting. Since the cancer transcriptome is the result of a progressive deregulation of global physiological gene expression networks, special care has to be taken when seeking to make biological inferences based on patterns of gene co-expression in cancer expression profiles. By exposing the heterogeneous nature of prognostic signals in a collection of 114 cancer transcriptomes with survival follow-up data, we challenge the still widespread perception that association with outcome of biologically-motivated biomarkers can be used to implicate causation of the underlying biological processes in cancer progression. While this deductive reasoning has largely been denounced among data analysis research circles,[306] we believe it still has a large bearing among wet-lab research domains,[307] and thus still warrants careful discussion from the epistemological point of

[306] Karagiannis et al., 2016

[307] Győrffy et al., 2013

view. To that end, our meta-analysis on the extent of prognostic signals in the cancer transcriptome has had two main goals. First, to bridge the gap in understanding of the scope of association with outcome of biologically motivated gene expression signatures and increase awareness of abusive conclusions drawn from the analysis of genomic data. Second, to shed light on the collection of variables linked with the heterogeneous nature of prognostic signals in cancer transcriptomes assayed by different high-throughput technologies, in order to inform controlled experiments designed towards an accurate validation of biological phenomena in cancer progression using survival analysis.

The complex nature of cancer transcriptomes, riddled by technical and biological sources of noise, also poses a challenge to learning algorithms seeking to capture patterns of disease progression. To address this limitation, we sought to chart this progression using an extrinsic, invariant property of multicellular systems—differentiation. In order to provide a quantifiable account of differentiation, we devised a simple feature selection algorithm to identify gene tissue-specific transcripts based on panels of healthy tissues. We then showed that this simple approach could yield a marker of cancer progression that, in a thyroid carcinogenesis model, was successful at discriminating between challenging pathological thyrocyte-derived neoplasias. The idea that tracking patterns of molecular differentiation could be useful in molecular pathology has been discussed elsewhere,[308] but remains to be fully exploited, both at the level of the accurate definition of tissue- and cell-type-specific markers, and at the selection of models under which it could be useful.

[308] Tarabichi et al., 2013

Looking forward, novel technological developments could be exploited to push these findings further. For instance, the use of single cell isolation methods could help define biomarkers at the cell type level, as opposed to the tissue level presented in this dissertation. Several methods exist currently to isolate single cell populations; these include micromanipulation, fluorescence activated cell-sorting, laser-capture microdissection or microfluidics.[309] Quantitative single-cell transcriptomics, coupled with tissue-type specific markers of expression, could be used to address questions such as the source of origin of metastatic outgrowths, lineage tracing in cancer genomics, or the analysis of rare cell types such as adult stem cells. Quantification of single cell types in bulk cancer tissues, together with the characterization of their transcriptional profiles, could eventually shed light on which cellular types are associated with particular disease outcomes in survival analyses. The degree of resolution offered by single-cell transcriptional approaches is however hindered by the principle of biological uncertainty.[310] This concept states that, in order to characterize the "cellular momentum" of individual cells in an organism, one must interfere with the behaviour of the cell and hence compromise on precisely knowing the state of the cell.

[309] Shapiro et al., 2013

[310] Shapiro et al., 2013

In the studies here presented, the use of single cell profiling technologies would prevent the inclusion of the parathyroid hormone gene (PTH) in the thyroid differentiation biomarker. The profiling of single cell types, in cancer survival analysis, could help resolving particular cell types correlations with differential outcome, such as stromal components in breast cancer or

tumoural infiltration of immune response cells.

Novel resource databases, such as the Genotype-Tissue Expression project, GTEx,[311] which reports the results of the transcriptional variation within a biobank of more than 50 human tissue types, or The Cancer Genome Atlas, TCGA,[312] which extensively profiled more than 30 human cancer types, could also be used to extend and improve the findings here presented. The GTEx database aims to establish a biobank to study the relationship between genetic variation and gene expression variation in multiple reference tissues. The availability of array- and RNA-seq-based detailed expression profiles of several human tissue sites at high resolution shows great promise for the optimization of tissue-specific biomarkers presented in this dissertation. This is also because GTEx provides with subject-wise profiles, which means inter-individual variability can be factored in potential novel tissue-specific biomarkers. These biomarkers could then be used to dissect cancer progression in distinct TCGA datasets, with the possibility of multivariate analysis of these biomarkers with samples clinical variables. The potential of these differentiation biomarkers is not restricted to the cancer setting; they could also be further extended to the analysis of expression profiles of stem cell or of *in vitro* models of organoid differentiation.

[311] Lonsdale et al., 2013

[312] The Cancer Genome Atlas Research Network et al., 2013

The first molecular cancer biomarkers translated into the clinical space sought to provides information on the likely course of the cancer disease in an untreated individual. Such prognostic markers, like Oncotype DX, MammaPrint© and PAM50 in the breast cancer setting, are usually multigene assays validated in large control cohorts of cancer patients with survival analysis. More recently, interest has been drawn to predictive biomarkers, defined as markers which can be used to identify subpopulations of patients who are most likely to respond to a given therapy. In the breast cancer setting, for instance, expression of the estrogen and progesterone receptors are both examples of biomarkers to predict sensitivity to endocrine therapy and HER2 receptor expression to predict sensitivity to Herceptin treatment. Novel molecular biomarkers for response to therapy for breast cancer,[313] lung cancer,[314] and colorectal cancer,[315] among others, have been used to refine clinical decisions regarding individualized or tailor-made treatment. The prospective validation of predictive multi-gene biomarkers in large cohorts of patients is nevertheless contrived by the same issues raised by our meta-analysis of prognostic signals, e.g., are these predictors independent of the fundamental drivers of cancer progression? For instance, treatment-resistant tumours could also present more aggressive outlooks masked by biological correlates like genomic instability.

[313] Galanina et al., 2011; and Sparano et al., 2015

[314] Andrews et al., 2011

[315] Dienstmann et al., 2011

In this dissertation, we discussed applications of molecular biomarkers for cancer diagnostics and prognosis, based on high-throughput technologies. We conclude that the use of gene expression signatures in cancer research has shown great promise (e.g., with the use of differentiation and proliferation signatures to assist cancer diagnosis), but may have also promoted unfunded expectations (e.g., by mis-interpreting the nature of pervasive prognostic signals in cancer transcriptomes). The research programs here detailed were designed to exploit the vast compendia of cancer expression profiles available in the public domain, and their results must,

therefore, first be contextualized within the technical and analytic frame of the technology that enabled them.

While microarray technology allows for the quantification of mRNA products of biospecimens in solution, it only lends itself to the detection of gene products already mapped and printed in the chip of the platform of choice. Furthermore, eventual mutations (either nucleotide changes or structural variants) in actively transcribed genes may impair the correct estimation of their transcription levels due to the adulteration of their mRNA sequences. Even assuming correct estimates of the transcriptomic load of a given biospecimen, tissue cellularity and heterogeneity may complicate the interpretation of bulk expression profiles. Finally, as microarray technology only probes the Central Dogma of Biology[316] at the transcriptional level, it provides no insight on upstream modulatory effects on gene expression (e.g., epigenetic determinants), as well as on downstream modulatory effects on gene products (e.g., post-transcriptional modifications).

Some of these limitations are overcome by the advent of next-generation sequencing technologies, whose rapid decrease in cost[317] has made them the assay of choice for gene expression profiling. Sequencing allows for a direct quantification of a given type of sequences present in solution, so that their counting is linearly related to their concentration. It also allows to probe sequence diversity without a prior knowledge of which nucleic acids may be present, unlike microarrays. Sequencing is also enables the independent detection of closely related gene sequences, novel splice forms, or RNA editing that may be missed due to cross hybridization on DNA microarrays. It is thus likely that DNA arrays will be fully replaced by sequencing methods in cancer genomics in the near future.[318]

[316] The central dogma of molecular biology, postulated by Francis Crick in 1958 and reasserted in 1970 (Crick, 1958, 1970), pertains to the rules that govern the sequential flow of genetic information between DNA, RNA and proteins. It can be summarized as "DNA makes RNA makes protein," which provides the template for the enactment of hereditary information for all living organisms, and frames the scope of evolutionary forces on genetic systems.

[317] http://www.genome.gov/sequencingcosts/

[318] Bumgarner, 2013

Bibliography

C. Athena Aktipis, Amy M. Boddy, Robert A. Gatenby, Joel S. Brown, and Carlo C. Maley. Life history trade-offs in cancer evolution. *Nature Reviews Cancer*, 13(12):883–892, 2013. URL http://www.nature.com/articles/nrc3606.

A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, February 2000. ISSN 0028-0836. DOI: 10.1038/35000501.

Douglas G. Altman. *Practical Statistics for Medical Research*. CRC Press, November 1990. ISBN 978-0-412-27630-9.

Christophe Ambroise and Geoffrey J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences*, 99(10):6562–6566, May 2002. ISSN 0027-8424, 1091-6490. DOI: 10.1073/pnas.102102699. URL http://www.pnas.org/content/99/10/6562.

Kristina Anderson, Christoph Lutz, Frederik W. Van Delft, Caroline M. Bateman, Yanping Guo, Susan M. Colman, Helena Kempski, Anthony V. Moorman, Ian Titley, John Swansbury, and others. Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature*, 469(7330):356–361, 2011. URL http://www.nature.com/nature/journal/v469/n7330/full/nature09650.html.

Jenny Andrews, Paul Yeh, William Pao, and Leora Horn. Molecular predictors of response to chemotherapy in non-small cell lung cancer. *Cancer Journal (Sudbury, Mass.)*, 17(2):104–113, April 2011. ISSN 1540-336X. DOI: 10.1097/PPO.0b013e318213f3cf.

P. Armitage and R. Doll. The Age Distribution of Cancer and a Multi-stage Theory of Carcinogenesis. *British Journal of Cancer*, 8(1):1–12, March 1954. ISSN 0007-0920. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2007940/.

Grazia Arpino, Daniele Generali, Anna Sapino, Lucia Del Matro, Antonio Frassoldati, Michelino de Laurentis, Paolo Pronzato, Giorgio

Mustacchi, Marina Cazzaniga, Sabino De Placido, Pierfranco Conte, Mariarosa Cappelletti, Vanessa Zanoni, Andrea Antonelli, Mario Martinotti, Fabio Puglisi, Alfredo Berruti, Alberto Bottini, and Luigi Dogliotti. Gene expression profiling in breast cancer: A clinical perspective. *The Breast*, 22(2):109–120, April 2013. ISSN 0960-9776. DOI: 10.1016/j.breast.2013.01.016. URL http://www.sciencedirect.com/science/article/pii/S0960977613000180.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25 (1):25–29, May 2000. ISSN 1061-4036. DOI: 10.1038/75556. URL http://www.nature.com/ng/journal/v25/n1/abs/ng0500_25.html.

Robert Axelrod, David E. Axelrod, and Kenneth J. Pienta. Evolution of cooperation among tumor cells. *Proceedings of the National Academy of Sciences*, 103(36):13474–13479, September 2006. ISSN 0027-8424, 1091-6490. DOI: 10.1073/pnas.0606053103. URL http://www.pnas.org/content/103/36/13474.

John C. Bailar and Heather L. Gornik. Cancer Undefeated. *New England Journal of Medicine*, 336(22):1569–1574, May 1997. ISSN 0028-4793. DOI: 10.1056/NEJM199705293362206. URL http://www.nejm.org/doi/full/10.1056/NEJM199705293362206.

Shawn C. Baker, Steven R. Bauer, Richard P. Beyer, James D. Brenton, Bud Bromley, John Burrill, Helen Causton, Michael P. Conley, Rosalie Elespuru, Michael Fero, Carole Foy, James Fuscoe, Xiaolian Gao, David Lee Gerhold, Patrick Gilles, Federico Goodsaid, Xu Guo, Joe Hackett, Richard D. Hockett, Pranvera Ikonomi, Rafael A. Irizarry, Ernest S. Kawasaki, Tamma Kaysser-Kranich, Kathleen Kerr, Gretchen Kiser, Walter H. Koch, Kathy Y. Lee, Chunmei Liu, Z. Lewis Liu, Anne Lucas, Chitra F. Manohar, Garry Miyada, Zora Modrusan, Helen Parkes, Raj K. Puri, Laura Reid, Thomas B. Ryder, Marc Salit, Raymond R. Samaha, Uwe Scherf, Timothy J. Sendera, Robert A. Setterquist, Leming Shi, Richard Shippy, Jesus V. Soriano, Elizabeth A. Wagar, Janet A. Warrington, Mickey Williams, Frederike Wilmer, Mike Wilson, Paul K. Wolber, Xiaoning Wu, and Renata Zadro. The External RNA Controls Consortium: a progress report. *Nature Methods*, 2(10):731–734, October 2005. ISSN 1548-7091. DOI: 10.1038/nmeth1005-731. URL http://www.nature.com/nmeth/journal/v2/n10/full/nmeth1005-731.html.

Andrew H. Beck, Nicholas W. Knoblauch, Marco M. Hefti, Jennifer Kaplan, Stuart J. Schnitt, Aedin C. Culhane, Markus S. Schroeder, Thomas Risch, John Quackenbush, and Benjamin Haibe-Kains. Significance Analysis of Prognostic Signatures. *PLoS Comput Biol*, 9(1): e1002875, January 2013. DOI: 10.1371/journal.pcbi.1002875. URL http://dx.doi.org/10.1371/journal.pcbi.1002875.

Benjamin Beck and Cédric Blanpain. Unravelling cancer stem cell potential. *Nature Reviews Cancer*, 13(10):727–738, 2013. URL http://www.nature.com/nrc/journal/v13/n10/abs/nrc3597.html.

Richard Bellman and Richard Ernest Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961.

Vladimir A. Belyi, Prashanth Ak, Elke Markert, Haijian Wang, Wenwei Hu, Anna Puzio-Kuter, and Arnold J. Levine. The Origins and Evolution of the p53 Family of Genes. *Cold Spring Harbor Perspectives in Biology*, 2(6), June 2010. ISSN 1943-0264. DOI: 10.1101/cshperspect.a001198. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2869528/.

Ittai Ben-Porath, Matthew W. Thomson, Vincent J. Carey, Ruping Ge, George W. Bell, Aviv Regev, and Robert A. Weinberg. An embryonic stem cell–like gene expression signature in poorly differentiated aggressive human tumors. *Nature genetics*, 40(5):499–507, 2008. URL http://www.nature.com/ng/journal/v40/n5/abs/ng.127.html.

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)*, pages 289–300, 1995. URL http://www.jstor.org/stable/2346101.

Donald A. Berry, Naoto T. Ueno, Marcella M. Johnson, Xiudong Lei, Jean Caputo, Dori A. Smith, Linda J. Yancey, Michael Crump, Edward A. Stadtmauer, Pierre Biron, John P. Crown, Peter Schmid, Jean-Pierre Lotz, Giovanni Rosti, Marco Bregni, and Taner Demirer. High-Dose Chemotherapy With Autologous Hematopoietic Stem-Cell Transplantation in Metastatic Breast Cancer: Overview of Six Randomized Trials. *Journal of Clinical Oncology*, 29(24):3224–3231, August 2011. ISSN 0732-183X, 1527-7755. DOI: 10.1200/JCO.2010.32.5936. URL http://jco.ascopubs.org/content/29/24/3224.

Brian Bierie and Harold L. Moses. Tumour microenvironment: TGFbeta: the molecular Jekyll and Hyde of cancer. *Nature Reviews. Cancer*, 6(7): 506–520, July 2006. ISSN 1474-175X. DOI: 10.1038/nrc1926.

J Martin Bland and Douglas G Altman. The logrank test. *BMJ : British Medical Journal*, 328(7447):1073, May 2004. ISSN 0959-8138. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC403858/.

Human BodyMap. 2.0 data from Illumina (2011) http://www. ensembl. info/blog/2011/05/24/human-bodymap-2-0-data-from-illumina. *Accessed on*, 3, 2012.

Benjamin M. Bolstad, Rafael A. Irizarry, Magnus \AAstrand, and Terence P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19 (2):185–193, 2003. URL http://bioinformatics.oxfordjournals.org/content/19/2/185.short.

Gianni Bonadonna, Ercole Brusamolino, Pinuccia Valagussa, Anna Rossi, Luisa Brugnatelli, Cristina Brambilla, Mario De Lena, Gabriele Tancini,

Emilio Bajetta, Renato Musumeci, and others. Combination chemotherapy as an adjuvant treatment in operable breast cancer. *New England Journal of Medicine*, 294(8):405–410, 1976. URL http://www.nejm.org/doi/full/10.1056/NEJM197602192940801.

I. Borg and P. J. F. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer Science & Business Media, August 2005. ISBN 978-0-387-25150-9.

Ana Bosch, Zhiqiang Li, Anna Bergamaschi, Haley Ellis, Eneda Toska, Aleix Prat, Jessica J. Tao, Daniel E. Spratt, Nerissa T. Viola-Villegas, Pau Castel, Gerard Minuesa, Natasha Morse, Jordi Rodón, Yasir Ibrahim, Javier Cortes, Jose Perez-Garcia, Patricia Galvan, Judit Grueso, Marta Guzman, John A. Katzenellenbogen, Michael Kharas, Jason S. Lewis, Maura Dickler, Violeta Serra, Neal Rosen, Sarat Chandarlapaty, Maurizio Scaltriti, and José Baselga. PI3k inhibition results in enhanced estrogen receptor function and dependence in hormone receptor–positive breast cancer. *Science Translational Medicine*, 7(283):283ra51–283ra51, April 2015. ISSN 1946-6234, 1946-6242. DOI: 10.1126/scitranslmed.aaa4442. URL http://stm.sciencemag.org/content/7/283/283ra51.

D. Botstein. Genomic perspective and cancer. *Cold Spring Harbor Symposia on Quantitative Biology*, 68:417–424, 2003. ISSN 0091-7451.

A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics*, 29(4):365–371, December 2001. ISSN 1061-4036. DOI: 10.1038/ng1201-365.

Alvis Brazma, Helen Parkinson, Ugis Sarkans, Mohammadreza Shojatalab, Jaak Vilo, Niran Abeygunawardena, Ele Holloway, Misha Kapushesky, Patrick Kemmeren, Gonzalo Garcia Lara, Ahmet Oezcimen, Philippe Rocca-Serra, and Susanna-Assunta Sansone. ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, 31(1):68–71, January 2003. ISSN 1362-4962.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. URL http://link.springer.com/article/10.1023/A:1010933404324.

Jarle Breivik. The evolutionary origin of genetic instability in cancer development. *Seminars in Cancer Biology*, 15(1):51–60, February 2005. ISSN 1044-579X. DOI: 10.1016/j.semcancer.2004.09.008.

James D. Brenton, Lisa A. Carey, Ahmed Ashour Ahmed, and Carlos Caldas. Molecular classification and molecular forecasting of breast cancer: ready for clinical application? *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 23(29):7350–7360, October 2005. ISSN 0732-183X. DOI: 10.1200/JCO.2005.03.3845.

F. M. Buffa, A. L. Harris, C. M. West, and C. J. Miller. Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. *British journal of cancer*, 102(2):428–435, 2010. URL http://www.nature.com/bjc/journal/v102/n2/abs/6605450a.html.

Lars Bullinger, Konstanze Döhner, Eric Bair, Stefan Fröhling, Richard F. Schlenk, Robert Tibshirani, Hartmut Döhner, and Jonathan R. Pollack. Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *The New England Journal of Medicine*, 350(16): 1605–1616, April 2004. ISSN 1533-4406. DOI: 10.1056/NEJMoa031046.

Roger Bumgarner. Overview of DNA microarrays: types, applications, and their future. *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]*, Chapter 22:Unit 22.1., January 2013. ISSN 1934-3647. DOI: 10.1002/0471142727.mb2201s101.

Patrick Cahan, Felicia Rovegno, Denise Mooney, John C. Newman, Georges St. Laurent III, and Timothy A. McCaffrey. Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization. *Gene*, 401(1–2):12–18, October 2007. ISSN 0378-1119. DOI: 10.1016/j.gene.2007.06.016. URL http://www.sciencedirect.com/science/article/pii/S0378111907003289.

Scott L. Carter, Aron C. Eklund, Brigham H. Mecham, Isaac S. Kohane, and Zoltan Szallasi. Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements. *BMC Bioinformatics*, 6:107, 2005. ISSN 1471-2105. DOI: 10.1186/1471-2105-6-107. URL http://dx.doi.org/10.1186/1471-2105-6-107.

Scott L. Carter, Aron C. Eklund, Isaac S. Kohane, Lyndsay N. Harris, and Zoltan Szallasi. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nature genetics*, 38(9):1043–1048, 2006. URL http://www.nature.com/ng/journal/v38/n9/abs/ng1861.html.

Jim Cassidy, Donald Bissett, Roy Spence, and Miranda Payne, editors. *Oxford Handbook of Oncology*. Oxford University Press, 2010. ISBN 978-0-19-956313-5. URL http://oxfordmedicine.com/view/10.1093/med/9780199563135.001.1/med-9780199563135.

Howard Y Chang, Julie B Sneddon, Ash A Alizadeh, Ruchira Sood, Rob B West, Kelli Montgomery, Jen-Tsan Chi, Matt van de Rijn, David Botstein, and Patrick O Brown. Gene Expression Signature of Fibroblast Serum Response Predicts Human Cancer Progression: Similarities between Tumors and Wounds. *PLoS Biol*, 2(2):e7, January 2004. DOI: 10.1371/journal.pbio.0020007. URL http://dx.doi.org/10.1371/journal.pbio.0020007.

Howard Y. Chang, Dimitry SA Nuyten, Julie B. Sneddon, Trevor Hastie, Robert Tibshirani, Therese Sørlie, Hongyue Dai, Yudong D. He, Laura J. van't Veer, Harry Bartelink, and others. Robustness, scalability, and

integration of a wound-response gene expression signature in predicting breast cancer survival. *Proceedings of the National Academy of Sciences of the United States of America*, 102(10):3738–3743, 2005. URL http://www.pnas.org/content/102/10/3738.short.

Maïa Chanrion, Vincent Negre, Hélène Fontaine, Nicolas Salvetat, Frédéric Bibeau, Gaëtan Mac Grogan, Louis Mauriac, Dionyssios Katsaros, Franck Molina, Charles Theillet, and Jean-Marie Darbon. A gene expression signature that can predict the recurrence of tamoxifen-treated primary breast cancer. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 14(6):1744–1752, March 2008. ISSN 1078-0432. DOI: 10.1158/1078-0432.CCR-07-1833.

Xin Chen, Siu Tim Cheung, Samuel So, Sheung Tat Fan, Christopher Barry, John Higgins, Kin-Man Lai, Jiafu Ji, Sandrine Dudoit, Irene O. L. Ng, Matt Van De Rijn, David Botstein, and Patrick O. Brown. Gene expression patterns in human liver cancers. *Molecular Biology of the Cell*, 13(6):1929–1939, June 2002. ISSN 1059-1524. DOI: 10.1091/mbc.02-02-0023.

Jen-Tsan Chi, Zhen Wang, Dimitry SA Nuyten, Edwin H. Rodriguez, Marci E. Schaner, Ali Salim, Yun Wang, Gunnar B. Kristensen, \AAslaug Helland, Anne-Lise Børresen-Dale, and others. Gene expression programs in response to hypoxia: cell type specificity and prognostic significance in human cancers. *PLoS medicine*, 3(3):e47, 2006. URL http://dx.plos.org/10.1371/journal.pmed.0030047.

Frederic Chibon. Cancer gene expression signatures – The rise and fall? *European Journal of Cancer*, 49(8):2000–2009, May 2013. ISSN 0959-8049. DOI: 10.1016/j.ejca.2013.02.021. URL http://www.sciencedirect.com/science/article/pii/S0959804913001536.

R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2(1):65–73, July 1998. ISSN 1097-2765.

G. H. A. Clowes and F. W. Baeslack. Further evidence of immunity against cancer in mice after spontaneous recovery. *Medical News*, 87:968–971, 1905.

M. P. Cole, C. T. A. Jones, and I. D. H. Todd. A New Anti-oestrogenic Agent in Late Breast Cancer: An Early Clinical Appraisal of ICI46474. *British Journal of Cancer*, 25(2):270–275, June 1971. ISSN 0007-0920. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2008453/.

Alain Coletta, Colin Molter, Robin Duqué, David Steenhoff, Jonatan Taminau, Virginie De Schaetzen, Stijn Meganck, Cosmin Lazar, David Venet, Vincent Detours, and others. InSilico DB genomic datasets hub: an efficient starting point for analyzing genome-wide studies in GenePattern, Integrative Genomics Viewer, and R/Bioconductor. *Genome Biol*, 13(11): R104, 2012. URL http://www.biomedcentral.com/content/pdf/gb-2012-13-11-r104.pdf.

Maqc Consortium. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology*, 28(8):827–838, August 2010. ISSN 1087-0156. DOI: 10.1038/nbt.1665. URL http://www.nature.com/nbt/journal/v28/n8/full/nbt.1665.html.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995. ISSN 0885-6125, 1573-0565. DOI: 10.1007/BF00994018. URL http://link.springer.com/article/10.1007/BF00994018.

Henri Coutard. Roentgen therapy of epitheliomas of the tonsillar region, hypopharynx and larynx from 1920 to 1926. *Am J Roentgenol*, 28:313–31, 1932.

D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, January 1972. ISSN 0035-9246. URL http://www.jstor.org/stable/2985181.

David R. Cox, Eric D. Green, Eric S. Lander, Daniel Cohen, and Richard M. Myers. Assessing mapping progress in the Human Genome Project. *SCIENCE-NEW YORK THEN WASHINGTON-*, pages 2031–2031, 1994.

F H Crick. On protein synthesis. *Symposia of the Society for Experimental Biology*, 12:138–163, 1958. ISSN 0081-1386.

Francis Crick. Central Dogma of Molecular Biology. *Nature*, 227(5258):561–563, August 1970. DOI: 10.1038/227561a0. URL http://www.nature.com/nature/journal/v227/n5258/abs/227561a0.html.

Carlo M. Croce. Oncogenes and Cancer. *New England Journal of Medicine*, 358(5):502–511, January 2008. ISSN 0028-4793. DOI: 10.1056/NEJMra072367. URL http://www.nejm.org/doi/full/10.1056/NEJMra072367.

Aedín C. Culhane, Markus S. Schröder, Razvan Sultana, Shaita C. Picard, Enzo N. Martinelli, Caroline Kelly, Benjamin Haibe-Kains, Misha Kapushesky, Anne-Alyssa St Pierre, William Flahive, Kermshlise C. Picard, Daniel Gusenleitner, Gerald Papenhausen, Niall O'Connor, Mick Correll, and John Quackenbush. GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. *Nucleic Acids Research*, 40(D1):D1060–D1066, January 2012. ISSN 0305-1048, 1362-4962. DOI: 10.1093/nar/gkr901. URL http://nar.oxfordjournals.org/content/40/D1/D1060.

Maria D'Agostino, Marialuisa Sponziello, Cinzia Puppin, Marilena Celano, Valentina Maggisano, Federica Baldan, Marco Biffoni, Stefania Bulotta, Cosimo Durante, Sebastiano Filetti, Giuseppe Damante, and Diego Russo. Different expression of TSH receptor and NIS genes in thyroid cancer: role of epigenetics. *Journal of Molecular Endocrinology*, 52(2):121–131, April 2014. ISSN 1479-6813. DOI: 10.1530/JME-13-0160.

Hongyue Dai, Laura van't Veer, John Lamb, Yudong D. He, Mao Mao, Bernard M. Fine, Rene Bernards, Marc van de Vijver, Paul Deutsch, Alan Sachs, Roland Stoughton, and Stephen Friend. A cell proliferation signature is a marker of extremely poor outcome in a subpopulation of breast cancer patients. *Cancer Research*, 65(10):4059–4066, May 2005a. ISSN 0008-5472. DOI: 10.1158/0008-5472.CAN-04-3953.

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, 2005b. ISSN 1362-4962. DOI: 10.1093/nar/gni179.

Sandeep S. Dave, George Wright, Bruce Tan, Andreas Rosenwald, Randy D. Gascoyne, Wing C. Chan, Richard I. Fisher, Rita M. Braziel, Lisa M. Rimsza, Thomas M. Grogan, Thomas P. Miller, Michael LeBlanc, Timothy C. Greiner, Dennis D. Weisenburger, James C. Lynch, Julie Vose, James O. Armitage, Erlend B. Smeland, Stein Kvaloy, Harald Holte, Jan Delabie, Joseph M. Connors, Peter M. Lansdorp, Qin Ouyang, T. Andrew Lister, Andrew J. Davies, Andrew J. Norton, H. Konrad Muller-Hermelink, German Ott, Elias Campo, Emilio Montserrat, Wyndham H. Wilson, Elaine S. Jaffe, Richard Simon, Liming Yang, John Powell, Hong Zhao, Neta Goldschmidt, Michael Chiorazzi, and Louis M. Staudt. Prediction of Survival in Follicular Lymphoma Based on Molecular Features of Tumor-Infiltrating Immune Cells. *New England Journal of Medicine*, 351(21): 2159–2169, November 2004. ISSN 0028-4793. DOI: 10.1056/NEJMoa041869. URL http://dx.doi.org/10.1056/NEJMoa041869.

P. C. W. Davies and C. H. Lineweaver. Cancer tumors as Metazoa 1.0: tapping genes of ancient ancestors. *Physical Biology*, 8(1):015001, February 2011. ISSN 1478-3975. DOI: 10.1088/1478-3975/8/1/015001. URL http://iopscience.iop.org/1478-3975/8/1/015001.

P. F. Denoix. Enquete permanent dans les centres anticancereaux. *Bull Inst Natl Hyg*, 1(1):70–75, 1946.

Vincent T. DeVita and Edward Chu. A History of Cancer Chemotherapy. *Cancer Research*, 68(21):8643–8653, November 2008. ISSN 0008-5472, 1538-7445. DOI: 10.1158/0008-5472.CAN-07-6611. URL http://cancerres.aacrjournals.org/content/68/21/8643.

Vincent T. DeVita and Steven A. Rosenberg. Two hundred years of cancer research. *The New England Journal of Medicine*, 366(23):2207–2214, June 2012. ISSN 1533-4406. DOI: 10.1056/NEJMra1204479.

Vincent T. DeVita, ARTHUR A. SERPICK, and PAUL P. CARBONE. Combination chemotherapy in the treatment of advanced Hodgkin's disease. *Annals of Internal Medicine*, 73(6):881–895, 1970. URL http://annals.org/article.aspx?articleid=684972.

Rodrigo Dienstmann, Eduardo Vilar, and Josep Tabernero. Molecular predictors of response to chemotherapy in colorectal cancer. *Cancer*

Journal (Sudbury, Mass.), 17(2):114–126, April 2011. ISSN 1540-336X. DOI: 10.1097/PPO.0b013e318212f844.

Geneviève Dom, Vanessa Chico Galdo, Maxime Tarabichi, Gil Tomás, Aline Hébrant, Guy Andry, Viviane De Martelar, Frédérick Libert, Emmanuelle Leteurtre, Jacques E. Dumont, Carine Maenhaut, and Wilma C.G. van Staveren. 5-Aza-2⊠-Deoxycytidine Has Minor Effects on Differentiation in Human Thyroid Cancer Cell Lines, But Modulates Genes That Are Involved in Adaptation In Vitro. *Thyroid*, 23(3):317–328, March 2013. ISSN 1050-7256. DOI: 10.1089/thy.2012.0388. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3593687/.

Alain Dupuy and Richard M. Simon. Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting. *Journal of the National Cancer Institute*, 99(2):147–157, January 2007. ISSN 0027-8874, 1460-2105. DOI: 10.1093/jnci/djk018. URL http://jnci.oxfordjournals.org/content/99/2/147.

B. P. Durbin, J. S. Hardin, D. M. Hawkins, and D. M. Rocke. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18(suppl 1):S105–S110, July 2002. ISSN 1367-4803, 1460-2059.

Ron Edgar, Michael Domrachev, and Alex E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, January 2002. ISSN 0305-1048, 1362-4962. DOI: 10.1093/nar/30.1.207. URL http://nar.oxfordjournals.org/content/30/1/207.

Alejo Efeyan and Manuel Serrano. p53: guardian of the genome and policeman of the oncogenes. *Cell Cycle (Georgetown, Tex.)*, 6(9):1006–1010, May 2007. ISSN 1551-4005.

Liat Ein-Dor, Itai Kela, Gad Getz, David Givol, and Eytan Domany. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–178, 2005. URL http://bioinformatics.oxfordjournals.org/content/21/2/171.short.

Eli Eisenberg and Erez Y. Levanon. Human housekeeping genes are compact. *Trends in genetics: TIG*, 19(7):362–365, July 2003. ISSN 0168-9525. DOI: 10.1016/S0168-9525(03)00140-9.

Steven Eschrich and Timothy J. Yeatman. DNA microarrays and data analysis: an overview. *Surgery*, 136(3):500–503, September 2004. ISSN 0039-6060. DOI: 10.1016/j.surg.2004.05.038.

Tom Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006. URL http://www.sciencedirect.com/science/article/pii/S016786550500303X.

FDA. 2007 - FDA Clears Breast Cancer Specific Molecular Prognostic Test, 2007. URL http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/2007/ucm108836.htm.

J. Ferlay, I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, and others. *GLOBOCAN 2012 v1. 0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. Lyon, France: International Agency for Research on Cancer*. 2014.

L. R. Finger, R. C. Harvey, R. C. Moore, L. C. Showe, and C. M. Croce. A common mechanism of chromosomal translocation in T- and B-cell neoplasia. *Science*, 234(4779):982–985, November 1986. ISSN 0036-8075, 1095-9203. DOI: 10.1126/science.3490692. URL http://www.sciencemag.org/content/234/4779/982.

Bernard Fisher, Madeline Bauer, Richard Margolese, Roger Poisson, Yosef Pilch, Carol Redmond, Edwin Fisher, Norman Wolmark, Melvin Deutsch, Eleanor Montague, and others. Five-year results of a randomized clinical trial comparing total mastectomy and segmental mastectomy with or without radiation in the treatment of breast cancer. *New England Journal of Medicine*, 312(11):665–673, 1985a. URL http://www.nejm.org/doi/pdf/10.1056/NEJM198503143121101.

Bernard Fisher, Carol Redmond, Edwin R. Fisher, Madeline Bauer, Norman Wolmark, D. Lawrence Wickerham, Melvin Deutsch, Eleanor Montague, Richard Margolese, and Roger Foster. Ten-year results of a randomized clinical trial comparing radical mastectomy and total mastectomy with or without radiation. *New England Journal of Medicine*, 312 (11):674–681, 1985b. URL http://www.nejm.org/doi/pdf/10.1056/NEJM198503143121102.

John A. Foekens, David Atkins, Yi Zhang, Fred CGJ Sweep, Nadia Harbeck, Angelo Paradiso, Tanja Cufer, Anieta M. Sieuwerts, Dmitri Talantov, Paul N. Span, and others. Multicenter validation of a gene expression–based prognostic signature in lymph node–negative primary breast cancer. *Journal of clinical oncology*, 24(11):1665–1671, 2006. URL http://jco.ascopubs.org/content/24/11/1665.short.

S. A. Forbes, G. Bhamra, S. Bamford, E. Dawson, C. Kok, J. Clements, A. Menzies, J. W. Teague, P. A. Futreal, and M. R. Stratton. The Catalogue of Somatic Mutations in Cancer (COSMIC). *Current Protocols in Human Genetics / Editorial Board, Jonathan L. Haines ... [et Al.]*, Chapter 10:Unit 10.11, April 2008. ISSN 1934-8258. DOI: 10.1002/0471142905.hg1011s57.

Steven A. Frank. *Dynamics of Cancer: Incidence, Inheritance, and Evolution*. Princeton University Press, Princeton (NJ), 2007. ISBN 978-0-691-13366-9. URL http://www.ncbi.nlm.nih.gov/books/NBK1568/.

E. Frei, M. Karon, R. H. Levin, E. J. Freireich, R. J. Taylor, J. Hananian, O. Selawry, J. F. Holland, B. Hoogstraten, I. J. Wolman, E. Abir, A. Sawitsky, S. Lee, S. D. Mills, E. O. Burgert, C. L. Spurr, R. B. Patterson, F. G. Ebaugh, G. W. James, and J. H. Moon. The effectiveness of combinations of antileukemic agents in inducing and maintaining remission in children with acute leukemia. *Blood*, 26(5):642–656, November 1965. ISSN 0006-4971.

Emil Frei. Curative cancer chemotherapy. *Cancer research*, 45(12 Part 1):6523–6537, 1985. URL http://cancerres.aacrjournals.org/content/45/12_Part_1/6523.short.

Natalie Galanina, Veerle Bossuyt, and Lyndsay N. Harris. Molecular predictors of response to therapy for breast cancer. *Cancer Journal (Sudbury, Mass.)*, 17(2):96–103, April 2011. ISSN 1540-336X. DOI: 10.1097/PPO.0b013e318212dee3.

M. E. Garber, O. G. Troyanskaya, K. Schluens, S. Petersen, Z. Thaesler, M. Pacyna-Gengelbach, M. van de Rijn, G. D. Rosen, C. M. Perou, R. I. Whyte, R. B. Altman, P. O. Brown, D. Botstein, and I. Petersen. Diversity of gene expression in adenocarcinoma of the lung. *Proceedings of the National Academy of Sciences of the United States of America*, 98(24):13784–13789, November 2001. ISSN 0027-8424. DOI: 10.1073/pnas.241500798.

Robert A. Gatenby and Robert J. Gillies. Why do cancers have high aerobic glycolysis? *Nature Reviews Cancer*, 4(11):891–899, November 2004. ISSN 1474-175X. DOI: 10.1038/nrc1478. URL http://www.nature.com/nrc/journal/v4/n11/full/nrc1478.html.

Laurent Gautier, Morten Møller, Lennart Friis-Hansen, and Steen Knudsen. Alternative mapping of probes to genes for Affymetrix chips. *BMC Bioinformatics*, 5:111, 2004. ISSN 1471-2105. DOI: 10.1186/1471-2105-5-111. URL http://dx.doi.org/10.1186/1471-2105-5-111.

Robert Gentleman, Vincent Carey, Wolfgang Huber, Rafael Irizarry, and Sandrine Dudoit. *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer Science & Business Media, 2006. URL https://books.google.com/books?hl=en&lr=&id=4cuyyv0Vu74C&oi=fnd&pg=PA3&dq=Bioinformatics+and+Computational+Biology+Solutions+Using+R+and+Bioconductor&ots=lWklIYQN5n&sig=qagQD4TQUmkDT1aK_qnDeJCO-Ks.

Robert C. Gentleman, Vincent J. Carey, Douglas M. Bates, and others. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004. URL http://genomebiology.com/2004/5/10/R80.

Olivier Gevaert and Bart De Moor. Prediction of cancer outcome using DNA microarray technology: past, present and future. *Expert Opinion on Medical Diagnostics*, 3(2):157–165, March 2009. ISSN 1753-0059. DOI: 10.1517/17530050802680172. URL http://informahealthcare.com/doi/abs/10.1517/17530050802680172.

Gennadi V. Glinsky, Olga Berezovska, Anna B. Glinskii, and others. Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. *Journal of Clinical Investigation*, 115(6):1503–1521, 2005. URL http://www.jci.org/cgi/content/abstract/115/6/1503.

Todd R. Golub, Donna K. Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P. Mesirov, Hilary Coller, Mignon L. Loh, James R. Downing, Mark A. Caligiuri, and others. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537, 1999. URL http://www.sciencemag.org/content/286/5439/531.short.

Stephan jay Gould and Niles Eldredge. Punctuated equilibrium comes of age. *Nature*, 366(6452):223–227, November 1993. DOI: 10.1038/366223a0. URL http://www.nature.com/nature/journal/v366/n6452/abs/366223a0.html.

Geraldine M. Grant, Amanda Fortney, Francesco Gorreta, Michael Estep, Luca Del Giacco, Amy Van Meter, Alan Christensen, Lakshmi Appalla, Chahla Naouar, Curtis Jamison, Ali Al-Timimi, Jean Donovan, James Cooper, Carleton Garrett, and Vikas Chandhoke. Microarrays in Cancer Research. *Anticancer Research*, 24(2A):441–448, March 2004. ISSN 0250-7005, 1791-7530. URL http://ar.iiarjournals.org/content/24/2A/441.

Mel Greaves and Carlo C. Maley. Clonal evolution in cancer. *Nature*, 481(7381):306–313, January 2012. ISSN 1476-4687. DOI: 10.1038/nature10762.

Frederick L. Greene. *AJCC cancer staging manual*, volume 1. Springer, 2002.

Arief Gusnanto, Stefano Calza, and Yudi Pawitan. Identification of differentially expressed genes and false discovery rate in microarray studies. *Current Opinion in Lipidology*, 18(2):187–193, April 2007. ISSN 0957-9672. DOI: 10.1097/MOL.0b013e3280895d6f.

Barry Gusterson. Do 'basal-like' breast cancers really exist? *Nature Reviews. Cancer*, 9(2):128–134, February 2009. ISSN 1474-1768. DOI: 10.1038/nrc2571.

Balazs Győrffy, Andras Lanczky, Aron C. Eklund, Carsten Denkert, Jan Budczies, Qiyuan Li, and Zoltan Szallasi. An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Research and Treatment*, 123(3):725–731, October 2010. ISSN 0167-6806, 1573-7217. DOI: 10.1007/s10549-009-0674-9. URL http://link.springer.com/10.1007/s10549-009-0674-9.

Balázs Győrffy, Pawel Surowiak, Jan Budczies, and András Lánczky. Online Survival Analysis Software to Assess the Prognostic Value of Biomarkers Using Transcriptomic Data in Non-Small-Cell Lung Cancer. *PLoS ONE*, 8(12):e82241, December 2013. DOI: 10.1371/journal.pone.0082241. URL http://dx.doi.org/10.1371/journal.pone.0082241.

A.-C. Gérard, C. Daumerie, C. Mestdagh, S. Gohy, C. de Burbure, S. Costagliola, F. Miot, M.-C. Nollevaux, J.-F. Denef, J. Rahier, B. Franc,

J. J. M. De Vijlder, I. M. Colin, and M.-C. Many. Correlation between the Loss of Thyroglobulin Iodination and the Expression of Thyroid-Specific Proteins Involved in Iodine Metabolism in Thyroid Carcinomas. *The Journal of Clinical Endocrinology & Metabolism*, 88(10):4977–4983, October 2003. ISSN 0021-972X. DOI: 10.1210/jc.2003-030586. URL http://press.endocrine.org/doi/abs/10.1210/jc.2003-030586.

Benjamin Haibe-Kains. *Identification and assessment of gene signatures in human breast cancer*. PhD thesis, Ph. D. thesis, University Libre de Bruxelles, Bioinformatics Department, 2009. URL http://www.ulb.ac.be/di/map/bhaibeka/phdthesis/haibekains2009phdthesis.pdf.

Benjamin Haibe-Kains, Christine Desmedt, Sherene Loi, Aedin C. Culhane, Gianluca Bontempi, John Quackenbush, and Christos Sotiriou. A three-gene model to robustly identify breast cancer molecular subtypes. *Journal of the National Cancer Institute*, 104(4):311–325, February 2012. ISSN 1460-2105. DOI: 10.1093/jnci/djr545.

Timothy C. Hallstrom, Seiichi Mori, and Joseph R. Nevins. An E2f1-dependent gene expression program that determines the balance between proliferation and cell death. *Cancer cell*, 13(1):11–22, 2008. URL http://www.sciencedirect.com/science/article/pii/S1535610807003704.

William S. Halsted. I. The results of operations for the cure of cancer of the breast performed at the Johns Hopkins Hospital from June, 1889, to January, 1894. *Annals of surgery*, 20(5):497, 1894. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1493925/.

Douglas Hanahan and Robert A. Weinberg. Hallmarks of Cancer: The Next Generation. *Cell*, 144(5):646–674, April 2011. ISSN 0092-8674. DOI: 10.1016/j.cell.2011.02.013. URL http://www.cell.com/article/S0092867411001279/abstract.

J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, April 1982. ISSN 0033-8419. DOI: 10.1148/radiology.143.1.7063747.

Bettina Harr and Christian Schlötterer. Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons. *Nucleic Acids Research*, 34(2):e8–e8, 2006. URL http://nar.oxfordjournals.org/content/34/2/e8.short.

Markus Hartl, Anna-Maria Mitterstiller, Taras Valovka, Kathrin Breuker, Bert Hobmayer, and Klaus Bister. Stem cell-specific activation of an ancestral myc protooncogene with conserved basic functions in the early metazoan Hydra. *Proceedings of the National Academy of Sciences of the United States of America*, 107(9):4051–4056, March 2010. ISSN 1091-6490. DOI: 10.1073/pnas.0911060107.

Yudong D. He and Stephen H. Friend. Microarrays—the 21st century divining rod? *Nature medicine*, 7(6):658–659, 2001. URL http://www.nature.com/nm/journal/v7/n6/abs/nm0601_658.html.

Zhiyuan Hu, Cheng Fan, Daniel S. Oh, J. S. Marron, Xiaping He, Bahjat F. Qaqish, Chad Livasy, Lisa A. Carey, Evangeline Reynolds, Lynn Dressler, Andrew Nobel, Joel Parker, Matthew G. Ewend, Lynda R. Sawyer, Junyuan Wu, Yudong Liu, Rita Nanda, Maria Tretiakova, Alejandra Ruiz Orrico, Donna Dreher, Juan P. Palazzo, Laurent Perreard, Edward Nelson, Mary Mone, Heidi Hansen, Michael Mullins, John F. Quackenbush, Matthew J. Ellis, Olufunmilayo I. Olopade, Philip S. Bernard, and Charles M. Perou. The molecular portraits of breast tumors are conserved across microarray platforms. *BMC genomics*, 7:96, 2006. ISSN 1471-2164. DOI: 10.1186/1471-2164-7-96.

Earl Hubbell, Wei-Min Liu, and Rui Mei. Robust estimators for expression analysis. *Bioinformatics*, 18(12):1585–1592, December 2002. ISSN 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/18.12.1585. URL http://bioinformatics.oxfordjournals.org/content/18/12/1585.

John P. A. Ioannidis. Why Most Published Research Findings Are False. *PLoS Med*, 2(8):e124, August 2005. DOI: 10.1371/journal.pmed.0020124. URL http://dx.doi.org/10.1371/journal.pmed.0020124.

John P. A. Ioannidis, David B. Allison, Catherine A. Ball, Issa Coulibaly, Xiangqin Cui, Aedín C. Culhane, Mario Falchi, Cesare Furlanello, Laurence Game, Giuseppe Jurman, Jon Mangion, Tapan Mehta, Michael Nitzberg, Grier P. Page, Enrico Petretto, and Vera van Noort. Repeatability of published microarray gene expression analyses. *Nature Genetics*, 41(2): 149–155, February 2009. ISSN 1546-1718. DOI: 10.1038/ng.295.

Rafael A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs, and Terence P. Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31(4):e15, February 2003a. ISSN 1362-4962.

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):249–264, April 2003b. ISSN 1465-4644, 1468-4357. DOI: 10.1093/biostatistics/4.2.249. URL http://biostatistics.oxfordjournals.org/content/4/2/249.

Hiraku Itadani, Shinji Mizuarai, and Hidehito Kotani. Can systems biology understand pathway activation? Gene expression signatures as surrogate markers for understanding the complexity of pathway activation. *Current Genomics*, 9(5):349–360, 2008. ISSN 1389-2029. DOI: 10.2174/138920208785133235.

Antoine Italiano. Prognostic or Predictive? It's Time to Get Back to Definitions! *Journal of Clinical Oncology*, 29(35):4718–4718, December 2011. ISSN 0732-183X, 1527-7755. DOI: 10.1200/JCO.2011.38.3729. URL http://jco.ascopubs.org/content/29/35/4718.1.

Marc Johannes, Markus Ruschhaupt, Froehlich Holger, Mansmann Ulrich, Andreas Buness, Patrick Warnat, Wolfgang Huber, Axel Benner, and Tim

Beissbarth. MCRestimate: Misclassification error estimation with cross-validation, 2010.

V. C. Jordan. Effects of tamoxifen in relation to breast cancer. *British medical journal*, 1(6075):1534, 1977. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1607295/.

John D. Kalbfleisch and Ross L. Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011. URL https://books.google.be/books?hl=en&lr=&id=BR4Kq-a1MIMC&oi=fnd&pg=PR7&dq=The+Statistical+Analysis+of+Failure+Time+Data&ots=xClk5HSS7X&sig=K_FQ4eGkhPisGgfzIXMIWWoL0rU.

M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, January 2000. ISSN 0305-1048.

E. L. Kaplan and Paul Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282): 457–481, June 1958. ISSN 0162-1459. DOI: 10.2307/2281868. URL http://www.jstor.org/stable/2281868.

H. S. Kaplan. Clinical evaluation and radiotherapeutic management of Hodgkin's disease and the malignant lymphomas. *The New England journal of medicine*, 278(16):892–899, 1968. URL http://europepmc.org/abstract/MED/4170945.

George S. Karagiannis, Sumanta Goswami, Joan G. Jones, Maja H. Oktay, and John S. Condeelis. Signatures of breast cancer metastasis at a glance. *J Cell Sci*, 129(9):1751–1758, 2016. URL http://jcs.biologists.org/content/129/9/1751.abstract.

Audrey Kauffmann and Wolfgang Huber. Microarray data quality control improves the detection of differentially expressed genes. *Genomics*, 95(3):138–142, March 2010. ISSN 0888-7543. DOI: 10.1016/j.ygeno.2010.01.003. URL http://www.sciencedirect.com/science/article/pii/S0888754310000042.

Audrey Kauffmann, Robert Gentleman, and Wolfgang Huber. arrayQualityMetrics–a bioconductor package for quality assessment of microarray data. *Bioinformatics (Oxford, England)*, 25(3):415–416, February 2009. ISSN 1367-4811. DOI: 10.1093/bioinformatics/btn647.

Ernest S. Kawasaki. The end of the microarray Tower of Babel: will universal standards lead the way? *Journal of biomolecular techniques: JBT*, 17(3): 200–206, July 2006. ISSN 1524-0215.

Purvesh Khatri, Marina Sirota, and Atul J. Butte. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Comput Biol*, 8(2):e1002375, February 2012. DOI: 10.1371/journal.pcbi.1002375. URL http://dx.doi.org/10.1371/journal.pcbi.1002375.

Kyoungmi Kim, Stanislav O. Zakharkin, and David B. Allison. Expectations, validity, and reality in gene expression profiling. *Journal of Clinical*

Epidemiology, 63(9):950–959, September 2010. ISSN 0895-4356. DOI: 10.1016/j.jclinepi.2010.02.018. URL http://www.sciencedirect.com/science/article/pii/S0895435610001368.

Kenneth W. Kinzler and Bert Vogelstein. Gatekeepers and caretakers. *Nature*, 386(6627):761–763, April 1997. DOI: 10.1038/386761a0. URL http://www.nature.com/nature/journal/v386/n6627/abs/386761a0.html.

Muaiad Kittaneh, Alberto J. Montero, and Stefan Gluck. Molecular Profiling for Breast Cancer: A Comprehensive Review. *Biomarkers in Cancer*, 5:61–70, October 2013. ISSN 1179-299X. DOI: 10.4137/BIC.S9455. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3825646/.

David G. Kleinbaum and Mitchel Klein. *Survival analysis*. Springer, 1996. URL http://link.springer.com/content/pdf/10.1007/978-1-4419-6646-9.pdf.

Tetsuo Kondo, Shereen Ezzat, and Sylvia L. Asa. Pathogenetic mechanisms in thyroid follicular-cell neoplasia. *Nature Reviews. Cancer*, 6(4):292–306, April 2006. ISSN 1474-175X. DOI: 10.1038/nrc1836.

J. B. Konopka, S. M. Watanabe, J. W. Singer, S. J. Collins, and O. N. Witte. Cell lines and clinical isolates derived from Ph1-positive chronic myelogenous leukemia patients express c-abl proteins with a common structural alteration. *Proceedings of the National Academy of Sciences of the United States of America*, 82(6):1810–1814, March 1985. ISSN 0027-8424.

H. Land, L. F. Parada, and R. A. Weinberg. Cellular oncogenes and multistep carcinogenesis. *Science*, 222(4625):771–778, November 1983. ISSN 0036-8075, 1095-9203. DOI: 10.1126/science.6356358. URL http://www.sciencemag.org/content/222/4625/771.

Justin D. Lathia, John M. Heddleston, Monica Venere, and Jeremy N. Rich. Deadly Teamwork: Neural Cancer Stem Cells and the Tumor Microenvironment. *Cell Stem Cell*, 8(5):482–485, May 2011. ISSN 1934-5909. DOI: 10.1016/j.stem.2011.04.013. URL http://www.sciencedirect.com/science/article/pii/S1934590911001780.

Martin Lauss, Markus Ringnér, and Mattias Höglund. Prediction of stage, grade, and survival in bladder cancer using genome-wide expression data: a validation study. *Clinical Cancer Research*, 16(17):4421–4433, 2010. URL https://clincancerres.aacrjournals.org/content/16/17/4421.full.

Peter D. Lee, Robert Sladek, Celia M. T. Greenwood, and Thomas J. Hudson. Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies. *Genome Research*, 12(2):292–297, February 2002. ISSN 1088-9051. DOI: 10.1101/gr.217802.

Suet Y. Leung, Xin Chen, Kent M. Chu, Siu T. Yuen, Jonathan Mathy, Jiafu Ji, Annie S. Y. Chan, Rui Li, Simon Law, Olga G. Troyanskaya, I.-Ping Tu, John Wong, Samuel So, David Botstein, and Patrick O. Brown.

Phospholipase A2 group IIA expression in gastric adenocarcinoma is associated with prolonged survival and less frequent metastasis. *Proceedings of the National Academy of Sciences of the United States of America*, 99(25):16203–16208, December 2002. ISSN 0027-8424. DOI: 10.1073/pnas.212646299.

Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P. Mesirov. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011. URL http://bioinformatics.oxfordjournals.org/content/27/12/1739.short.

Joseph Lister. On the antiseptic principle in the practice of surgery. *The Lancet*, 90(2299):353–356, 1867. URL http://www.sciencedirect.com/science/article/pii/S0140673602518274.

Rui Liu, Xinhao Wang, Grace Y. Chen, Piero Dalerba, Austin Gurney, Timothy Hoey, Gavin Sherlock, John Lewicki, Kerby Shedden, and Michael F. Clarke. The prognostic role of a gene signature from tumorigenic breast-cancer cells. *New England Journal of Medicine*, 356 (3):217–226, 2007. URL http://www.nejm.org/doi/full/10.1056/NEJMoa063994.

Zhaoqi Liu, Xiang-Sun Zhang, and Shihua Zhang. Breast tumor subgroups reveal diverse clinical prognostic power. *Scientific Reports*, 4, February 2014. ISSN 2045-2322. DOI: 10.1038/srep04002. URL http://www.nature.com/doifinder/10.1038/srep04002.

David Lloyd, Miguel A. Aon, and Sonia Cortassa. Why Homeodynamics, Not Homeostasis? *The Scientific World Journal*, 1:133–145, 2001. DOI: 10.1100/tsw.2001.20. URL http://www.hindawi.com/journals/tswj/2001/918917/abs/.

David J. Lockhart, Helin Dong, Michael C. Byrne, Maximillian T. Follettie, Michael V. Gallo, Mark S. Chee, Michael Mittmann, Chunwei Wang, Michiko Kobayashi, Heidi Norton, and others. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature biotechnology*, 14(13):1675–1680, 1996. URL http://www.nature.com/nbt/journal/v14/n13/abs/nbt1296-1675.html.

Sherene Loi, Benjamin Haibe-Kains, Christine Desmedt, Françoise Lallemand, Andrew M. Tutt, Cheryl Gillet, Paul Ellis, Adrian Harris, Jonas Bergh, John A. Foekens, Jan G. M. Klijn, Denis Larsimont, Marc Buyse, Gianluca Bontempi, Mauro Delorenzi, Martine J. Piccart, and Christos Sotiriou. Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 25(10):1239–1246, April 2007. ISSN 1527-7755. DOI: 10.1200/JCO.2006.07.1522.

John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, Barbara Foster, Mike Moser, Ellen Karasik, Bryan Gillard, Kimberley Ramsey, Susan Sullivan, Jason Bridge, Harold Magazine, John

Syron, Johnelle Fleming, Laura Siminoff, Heather Traino, Maghboeba Mosavel, Laura Barker, Scott Jewell, Dan Rohrer, Dan Maxim, Dana Filkins, Philip Harbach, Eddie Cortadillo, Bree Berghuis, Lisa Turner, Eric Hudson, Kristin Feenstra, Leslie Sobin, James Robb, Phillip Branton, Greg Korzeniewski, Charles Shive, David Tabor, Liqun Qi, Kevin Groch, Sreenath Nampally, Steve Buia, Angela Zimmerman, Anna Smith, Robin Burges, Karna Robinson, Kim Valentino, Deborah Bradbury, Mark Cosentino, Norma Diaz-Mayoral, Mary Kennedy, Theresa Engel, Penelope Williams, Kenyon Erickson, Kristin Ardlie, Wendy Winckler, Gad Getz, David DeLuca, Daniel MacArthur, Manolis Kellis, Alexander Thomson, Taylor Young, Ellen Gelfand, Molly Donovan, Yan Meng, George Grant, Deborah Mash, Yvonne Marcus, Margaret Basile, Jun Liu, Jun Zhu, Zhidong Tu, Nancy J. Cox, Dan L. Nicolae, Eric R. Gamazon, Hae Kyung Im, Anuar Konkashbaev, Jonathan Pritchard, Matthew Stevens, Timothèe Flutre, Xiaoquan Wen, Emmanouil T. Dermitzakis, Tuuli Lappalainen, Roderic Guigo, Jean Monlong, Michael Sammeth, Daphne Koller, Alexis Battle, Sara Mostafavi, Mark McCarthy, Manual Rivas, Julian Maller, Ivan Rusyn, Andrew Nobel, Fred Wright, Andrey Shabalin, Mike Feolo, Nataliya Sharopova, Anne Sturcke, Justin Paschal, James M. Anderson, Elizabeth L. Wilder, Leslie K. Derr, Eric D. Green, Jeffery P. Struewing, Gary Temple, Simona Volpi, Joy T. Boyer, Elizabeth J. Thomson, Mark S. Guyer, Cathy Ng, Assya Abdallah, Deborah Colantuoni, Thomas R. Insel, Susan E. Koester, A. Roger Little, Patrick K. Bender, Thomas Lehner, Yin Yao, Carolyn C. Compton, Jimmie B. Vaught, Sherilyn Sawyer, Nicole C. Lockhart, Joanne Demchok, and Helen F. Moore. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6): 580–585, June 2013. ISSN 1061-4036. DOI: 10.1038/ng.2653. URL http://www.nature.com/ng/journal/v45/n6/full/ng.2653.html.

Carrie C. Lubitz, Lisa A. Gallagher, David J. Finley, Baixin Zhu, and Thomas J. Fahey. Molecular analysis of minimally invasive follicular carcinomas by gene profiling. *Surgery*, 138(6):1042–1048; discussion 1048–1049, December 2005. ISSN 0039-6060. DOI: 10.1016/j.surg.2005.09.009.

Joseph A. Ludwig and John N. Weinstein. Biomarkers in Cancer Staging, Prognosis and Treatment Selection. *Nature Reviews Cancer*, 5(11):845–856, November 2005. ISSN 1474-175X. DOI: 10.1038/nrc1739. URL http://www.nature.com/nrc/journal/v5/n11/full/nrc1739.html.

Xiao-Jun Ma, Zuncai Wang, Paula D. Ryan, Steven J. Isakoff, Anne Barmettler, Andrew Fuller, Beth Muir, Gayatry Mohapatra, Ranelle Salunga, J. Todd Tuggle, and others. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer cell*, 5 (6):607–616, 2004. URL http://www.sciencedirect.com/science/article/pii/S1535610804001412.

C. Maenhaut, J. E. Dumont, P. P. Roger, and W. C. G. van Staveren. Cancer stem cells: a reality, a myth, a fuzzy concept or a misnomer? An analysis. *Carcinogenesis*, 31(2):149–158, February 2010. ISSN 1460-2180. DOI: 10.1093/carcin/bgp259.

N. Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports. Part 1*, 50(3): 163–170, March 1966. ISSN 0069-0112.

MAQC Consortium, Leming Shi, Laura H. Reid, Wendell D. Jones, Richard Shippy, Janet A. Warrington, Shawn C. Baker, Patrick J. Collins, Francoise de Longueville, Ernest S. Kawasaki, Kathleen Y. Lee, Yuling Luo, Yongming Andrew Sun, James C. Willey, Robert A. Setterquist, Gavin M. Fischer, Weida Tong, Yvonne P. Dragan, David J. Dix, Felix W. Frueh, Frederico M. Goodsaid, Damir Herman, Roderick V. Jensen, Charles D. Johnson, Edward K. Lobenhofer, Raj K. Puri, Uwe Schrf, Jean Thierry-Mieg, Charles Wang, Mike Wilson, Paul K. Wolber, Lu Zhang, Shashi Amur, Wenjun Bao, Catalin C. Barbacioru, Anne Bergstrom Lucas, Vincent Bertholet, Cecilie Boysen, Bud Bromley, Donna Brown, Alan Brunner, Roger Canales, Xiaoxi Megan Cao, Thomas A. Cebula, James J. Chen, Jing Cheng, Tzu-Ming Chu, Eugene Chudin, John Corson, J. Christopher Corton, Lisa J. Croner, Christopher Davies, Timothy S. Davison, Glenda Delenstarr, Xutao Deng, David Dorris, Aron C. Eklund, Xiao-hui Fan, Hong Fang, Stephanie Fulmer-Smentek, James C. Fuscoe, Kathryn Gallagher, Weigong Ge, Lei Guo, Xu Guo, Janet Hager, Paul K. Haje, Jing Han, Tao Han, Heather C. Harbottle, Stephen C. Harris, Eli Hatchwell, Craig A. Hauser, Susan Hester, Huixiao Hong, Patrick Hurban, Scott A. Jackson, Hanlee Ji, Charles R. Knight, Winston P. Kuo, J. Eugene LeClerc, Shawn Levy, Quan-Zhen Li, Chunmei Liu, Ying Liu, Michael J. Lombardi, Yunqing Ma, Scott R. Magnuson, Botoul Maqsodi, Tim McDaniel, Nan Mei, Ola Myklebost, Baitang Ning, Natalia Novoradovskaya, Michael S. Orr, Terry W. Osborn, Adam Papallo, Tucker A. Patterson, Roger G. Perkins, Elizabeth H. Peters, Ron Peterson, Kenneth L. Philips, P. Scott Pine, Lajos Pusztai, Feng Qian, Hongzu Ren, Mitch Rosen, Barry A. Rosenzweig, Raymond R. Samaha, Mark Schena, Gary P. Schroth, Svetlana Shchegrova, Dave D. Smith, Frank Staedtler, Zhenqiang Su, Hongmei Sun, Zoltan Szallasi, Zivana Tezak, Danielle Thierry-Mieg, Karol L. Thompson, Irina Tikhonova, Yaron Turpaz, Beena Vallanat, Christophe Van, Stephen J. Walker, Sue Jane Wang, Yonghong Wang, Russ Wolfinger, Alex Wong, Jie Wu, Chunlin Xiao, Qian Xie, Jun Xu, Wen Yang, Liang Zhang, Sheng Zhong, Yaping Zong, and William Slikker. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24(9):1151–1161, September 2006. ISSN 1087-0156. DOI: 10.1038/nbt1239.

Elke K. Markert, Hideaki Mizuno, Alexei Vazquez, and Arnold J. Levine. Molecular classification of prostate cancer using curated expression signatures. *Proceedings of the National Academy of Sciences*, 108(52): 21276–21281, December 2011. ISSN 0027-8424, 1091-6490. DOI: 10.1073/pnas.1117029108. URL `http://www.pnas.org/content/108/52/21276`.

Simone Mathoulin-Pelissier, Sophie Gourgou-Bourgade, Franck Bonnetain, and Andrew Kramar. Survival end point reporting in randomized cancer clinical trials: a review of major journals. *Journal of*

Clinical Oncology: Official Journal of the American Society of Clinical Oncology, 26(22):3721–3726, August 2008. ISSN 1527-7755. DOI: 10.1200/JCO.2007.14.1192.

Evan Matros, Zhigang C. Wang, Andrea L. Richardson, and James D. Iglehart. Genomic approaches in cancer biology. *Surgery*, 136(3):511–518, September 2004. ISSN 0039-6060. DOI: 10.1016/j.surg.2004.05.040.

John Maynard Smith and Eors Szathmary. *The Major Transitions in Evolution*. OUP Oxford, October 1997. ISBN 978-0-19-850294-4.

Matthew N. McCall, Benjamin M. Bolstad, and Rafael A. Irizarry. Frozen robust multiarray analysis (fRMA). *Biostatistics*, 11(2):242–253, April 2010. ISSN 1465-4644, 1468-4357. DOI: 10.1093/biostatistics/kxp059. URL http://biostatistics.oxfordjournals.org/content/11/2/242.

Matthew N McCall, Peter N Murakami, Margus Lukk, Wolfgang Huber, and Rafael A Irizarry. Assessing affymetrix GeneChip microarray quality. *BMC Bioinformatics*, 12:137, May 2011. ISSN 1471-2105. DOI: 10.1186/1471-2105-12-137. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3097162/.

Lauren M. F. Merlo, John W. Pepper, Brian J. Reid, and Carlo C. Maley. Cancer as an evolutionary and ecological process. *Nature Reviews Cancer*, 6(12):924–935, December 2006. ISSN 1474-175X. DOI: 10.1038/nrc2013. URL http://www.nature.com/nrc/journal/v6/n12/abs/nrc2013.html.

Stefan Michiels, Serge Koscielny, and Catherine Hill. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, 365(9458):488–492, February 2005. ISSN 1474-547X. DOI: 10.1016/S0140-6736(05)17866-0.

Jeremy A. Miller, Chaochao Cai, Peter Langfelder, Daniel H. Geschwind, Sunil M. Kurian, Daniel R. Salomon, and Steve Horvath. Strategies for aggregating gene expression data: the collapseRows R function. *BMC bioinformatics*, 12(1):322, 2011. URL http://www.biomedcentral.com/1471-2105/12/322/.

Lance D. Miller, Johanna Smeds, Joshy George, Vinsensius B. Vega, Liza Vergara, Alexander Ploner, Yudi Pawitan, Per Hall, Sigrid Klaar, Edison T. Liu, and Jonas Bergh. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences of the United States of America*, 102(38):13550–13555, September 2005. ISSN 0027-8424. DOI: 10.1073/pnas.0506230102.

William C. Moloney, Sharon Johnson, and Francis A. Countway Library of Medicine. *Pioneering hematology: the research and treatment of malignant blood disorders– reflections on a life's work*. Francis A. Countway Library of Medicine, September 1997. ISBN 978-0-88135-195-8.

Jonathan D. Mosley and Ruth A. Keri. Cell cycle correlated genes dictate the prognostic power of breast cancer gene lists. *BMC medical genomics*, 1 (1):11, 2008. URL http://www.biomedcentral.com/1755-8794/1/11.

Siddhartha Mukherjee. *The Emperor of All Maladies: A Biography of Cancer*. Scribner, New York, reprint edition edition, August 2011. ISBN 978-1-4391-7091-5.

PZ Myers. Aaargh! Physicists! Again!, December 2012. URL http://scienceblogs.com/pharyngula/2012/11/20/aaargh-physicists-again/.

A. Naderi, A. E. Teschendorff, N. L. Barbosa-Morais, S. E. Pinder, A. R. Green, D. G. Powe, J. F. R. Robertson, S. Aparicio, I. O. Ellis, J. D. Brenton, and others. A gene-expression signature to predict survival in breast cancer across independent data sets. *Oncogene*, 26(10):1507–1516, 2006. URL http://www.nature.com/onc/journal/v26/n10/abs/1209920a.html.

Vinay Nadimpally and Mohammed J. Zaki. A Novel Approach to Determine Normal Variation in Gene Expression Data. *SIGKDD Explor. Newsl.*, 5(2):6–15, December 2003. ISSN 1931-0145. DOI: 10.1145/980972.980975. URL http://doi.acm.org/10.1145/980972.980975.

Michael A. Newton, Christina M. Kendziorski, Craig S. Richmond, Frederick R. Blattner, and Kam-Wah Tsui. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of computational biology*, 8(1): 37–52, 2001. URL http://online.liebertpub.com/doi/abs/10.1089/106652701300099074.

Serena Nik-Zainal, Ludmil B. Alexandrov, David C. Wedge, Peter Van Loo, Christopher D. Greenman, Keiran Raine, David Jones, Jonathan Hinton, John Marshall, Lucy A. Stebbings, Andrew Menzies, Sancha Martin, Kenric Leung, Lina Chen, Catherine Leroy, Manasa Ramakrishna, Richard Rance, King Wai Lau, Laura J. Mudie, Ignacio Varela, David J. McBride, Graham R. Bignell, Susanna L. Cooke, Adam Shlien, John Gamble, Ian Whitmore, Mark Maddison, Patrick S. Tarpey, Helen R. Davies, Elli Papaemmanuil, Philip J. Stephens, Stuart McLaren, Adam P. Butler, Jon W. Teague, Göran Jönsson, Judy E. Garber, Daniel Silver, Penelope Miron, Aquila Fatima, Sandrine Boyault, Anita Langerød, Andrew Tutt, John W. M. Martens, Samuel A. J. R. Aparicio, Åke Borg, Anne Vincent Salomon, Gilles Thomas, Anne-Lise Børresen-Dale, Andrea L. Richardson, Michael S. Neuberger, P. Andrew Futreal, Peter J. Campbell, and Michael R. Stratton. Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell*, 149(5):979–993, May 2012. ISSN 0092-8674. DOI: 10.1016/j.cell.2012.04.024. URL http://www.sciencedirect.com/science/article/pii/S0092867412005284.

Yuri E. Nikiforov and Marina N. Nikiforova. Molecular genetics and diagnosis of thyroid cancer. *Nature Reviews. Endocrinology*, 7(10):569–580, October 2011. ISSN 1759-5037. DOI: 10.1038/nrendo.2011.142.

William S. Noble. How does multiple testing correction work? *Nature Biotechnology*, 27(12):1135–1137, December 2009. ISSN 1546-1696. DOI: 10.1038/nbt1209-1135.

Faiyaz Notta, Charles G. Mullighan, Jean CY Wang, Armando Poeppl, Sergei Doulatov, Letha A. Phillips, Jing Ma, Mark D. Minden, James R. Downing, and John E. Dick. Evolution of human BCR-ABL1 lymphoblastic leukaemia-initiating cells. *Nature*, 469(7330):362–367, 2011. URL http://www.nature.com/nature/journal/v469/n7330/abs/nature09733.html.

Noa Novershtern, Aravind Subramanian, Lee N. Lawton, Raymond H. Mak, W. Nicholas Haining, Marie E. McConkey, Naomi Habib, Nir Yosef, Cindy Y. Chang, Tal Shay, and others. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell*, 144 (2):296–309, 2011. URL http://www.sciencedirect.com/science/article/pii/S0092867411000055.

P. C. Nowell. The clonal evolution of tumor cell populations. *Science (New York, N.Y.)*, 194(4260):23–28, October 1976. ISSN 0036-8075.

Vigdis Nygaard and Eivind Hovig. Options available for profiling small samples: a review of sample amplification technology when combined with microarray profiling. *Nucleic Acids Research*, 34(3):996–1014, 2006. ISSN 0305-1048. DOI: 10.1093/nar/gkj499. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1363777/.

Soonmyung Paik, Steven Shak, Gong Tang, Chungyeul Kim, Joffre Baker, Maureen Cronin, Frederick L. Baehner, Michael G. Walker, Drew Watson, Taesung Park, and others. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine*, 351(27):2817–2826, 2004. URL http://www.nejm.org/doi/full/10.1056/NEJMoa041588.

Joel S. Parker, Michael Mullins, Maggie C. U. Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, John F. Quackenbush, Inge J. Stijleman, Juan Palazzo, J. S. Marron, Andrew B. Nobel, Elaine Mardis, Torsten O. Nielsen, Matthew J. Ellis, Charles M. Perou, and Philip S. Bernard. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 27(8):1160–1167, March 2009. ISSN 1527-7755. DOI: 10.1200/JCO.2008.18.1370.

Maria P. Pavlou, Eleftherios P. Diamandis, and Ivan M. Blasutig. The Long Journey of Cancer Biomarkers from the Bench to the Clinic. *Clinical Chemistry*, 59(1):147–157, January 2013. ISSN 0009-9147, 1530-8561. DOI: 10.1373/clinchem.2012.184614. URL http://www.clinchem.org/content/59/1/147.

Karl Pearson. LIII. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, 2(11):559–572, November 1901. ISSN 1941-5982. DOI: 10.1080/14786440109462720. URL http://dx.doi.org/10.1080/14786440109462720.

Soazig Le Pennec, Tomasz Konopka, David Gacquer, Danai Fimereli, Maxime Tarabichi, Gil Tomás, Frédérique Savagner, Myriam Decaussin-Petrucci, Christophe Trésallet, Guy Andry, Denis Larsimont, Vincent Detours, and Carine Maenhaut. Intratumor heterogeneity and clonal evolution in an aggressive papillary thyroid cancer and matched metastases. *Endocrine-Related Cancer*, 22(2):205–216, April 2015. ISSN 1351-0088, 1479-6821. DOI: 10.1530/ERC-14-0351. URL http://erc.endocrinology-journals.org/content/22/2/205.

C. M. Perou, T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lønning, A. L. Børresen-Dale, P. O. Brown, and D. Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, August 2000. ISSN 0028-0836. DOI: 10.1038/35021093.

Richard Peto and Julian Peto. Asymptotically Efficient Rank Invariant Test Procedures. *Journal of the Royal Statistical Society. Series A (General)*, 135 (2):185–207, January 1972. ISSN 0035-9238. DOI: 10.2307/2344317. URL http://www.jstor.org/stable/2344317.

Jonathan Pettit. Cancer does not give us a view of a bygone biological age, November 2012. URL http://genotripe.wordpress.com/2012/11/19/cancer-does-not-give-us-view-of-a-bygone-biological-age/.

Gregory Piatetsky-Shapiro and Pablo Tamayo. Microarray Data Mining: Facing the Challenges. *SIGKDD Explor. Newsl.*, 5(2):1–5, December 2003. ISSN 1931-0145. DOI: 10.1145/980972.980974. URL http://doi.acm.org/10.1145/980972.980974.

Martine J. Piccart and Isabelle Gingras. Breast cancer in 2015: Academic research sheds light on issues that matter to patients. *Nature Reviews Clinical Oncology*, 13(2):67–68, February 2016. ISSN 1759-4774. DOI: 10.1038/nrclinonc.2015.236. URL http://www.nature.com/nrclinonc/journal/v13/n2/full/nrclinonc.2015.236.html.

Mehdi Pirooznia, Jack Y. Yang, Mary Qu Yang, and Youping Deng. A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics*, 9(Suppl 1):S13, March 2008. ISSN 1471-2164. DOI: 10.1186/1471-2164-9-S1-S13. URL http://www.biomedcentral.com/1471-2164/9/S1/S13/abstract.

Alexander Ploner, Lance D. Miller, Per Hall, Jonas Bergh, and Yudi Pawitan. Correlation test to assess low-level processing of high-density oligonucleotide microarray data. *BMC bioinformatics*, 6(1):80, 2005. URL http://www.biomedcentral.com/1471-2105/6/80.

Ondrej Podlaha, Markus Riester, Subhajyoti De, and Franziska Michor. Evolution of the cancer genome. *Trends in genetics: TIG*, 28(4):155–163, April 2012. ISSN 0168-9525. DOI: 10.1016/j.tig.2012.01.003.

Kornelia Polyak. Tumor heterogeneity confounds and illuminates: a case for Darwinian tumor evolution. *Nature Medicine*, 20(4):344–346, April 2014. ISSN 1546-170X. DOI: 10.1038/nm.3518.

Kornelia Polyak, Izhak Haviv, and Ian G. Campbell. Co-evolution of tumor cells and their microenvironment. *Trends in genetics: TIG*, 25(1): 30–38, January 2009. ISSN 0168-9525. DOI: 10.1016/j.tig.2008.10.012.

Aleix Prat and Charles M. Perou. Deconstructing the molecular portraits of breast cancer. *Molecular Oncology*, 5(1):5–23, February 2011. ISSN 1878-0261. DOI: 10.1016/j.molonc.2010.11.003.

Lajos Pusztai, Chafika Mazouni, Keith Anderson, Yun Wu, and W. Fraser Symmans. Molecular classification of breast cancer: limitations and potential. *The Oncologist*, 11(8):868–877, September 2006. ISSN 1083-7159. DOI: 10.1634/theoncologist.11-8-868.

John Quackenbush. Computational analysis of microarray data. *Nature Reviews Genetics*, 2(6):418–427, June 2001. ISSN 1471-0056. DOI: 10.1038/35076576. URL http://www.nature.com/nrg/journal/v2/n6/abs/nrg0601_418a.html.

Elsa Quintana, Mark Shackleton, Hannah R. Foster, Douglas R. Fullen, Michael S. Sabel, Timothy M. Johnson, and Sean J. Morrison. Phenotypic heterogeneity among tumorigenic melanoma cells from patients that is reversible and not hierarchically organized. *Cancer Cell*, 18(5):510–523, November 2010. ISSN 1878-3686. DOI: 10.1016/j.ccr.2010.10.012.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org/.

Emad A. Rakha, Jorge S. Reis-Filho, and Ian O. Ellis. Basal-like breast cancer: a critical review. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 26(15):2568–2581, May 2008. ISSN 1527-7755. DOI: 10.1200/JCO.2007.13.1748.

S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, T. Poggio, W. Gerald, M. Loda, E. S. Lander, and T. R. Golub. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences of the United States of America*, 98(26):15149–15154, December 2001. ISSN 0027-8424. DOI: 10.1073/pnas.211566398.

Jorge S Reis-Filho and Lajos Pusztai. Gene expression profiling in breast cancer: classification, prognostication, and prediction. *The Lancet*, 378(9805):1812–1823, November 2011. ISSN 0140-6736. DOI: 10.1016/S0140-6736(11)61539-0. URL http://www.sciencedirect.com/science/article/pii/S0140673611615390.

Jorge S. Reis-Filho, Britta Weigelt, Debora Fumagalli, and Christos Sotiriou. Molecular profiling: moving away from tumor philately. *Science translational medicine*, 2(47):47ps43–47ps43, 2010. URL http://stm.sciencemag.org/content/2/47/47ps43.short.

Daniel R. Rhodes, Jianjun Yu, K. Shanker, Nandan Deshpande, Radhika Varambally, Debashis Ghosh, Terrence Barrette, Akhilesh Pandey, and

Arul M. Chinnaiyan. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proceedings of the National Academy of Sciences of the United States of America*, 101(25):9309–9314, June 2004. ISSN 0027-8424, 1091-6490. DOI: 10.1073/pnas.0401994101. URL http://www.pnas.org/content/101/25/9309.

Markus Ringnér. What is principal component analysis? *Nature Biotechnology*, 26(3):303–304, March 2008. ISSN 1546-1696. DOI: 10.1038/nbt0308-303.

Markus Ringnér, Erik Fredlund, Jari Hakkinen, Ake Borg, and Johan Staaf. GOBO: gene expression-based outcome for breast cancer online. *PloS one*, 6(3):e17911, 2011. URL http://dx.plos.org/10.1371/journal.pone.0017911.

Andreas Rosenwald, George Wright, Adrian Wiestner, Wing C. Chan, Joseph M. Connors, Elias Campo, Randy D. Gascoyne, Thomas M. Grogan, H. Konrad Muller-Hermelink, Erlend B. Smeland, Michael Chiorazzi, Jena M. Giltnane, Elaine M. Hurt, Hong Zhao, Lauren Averett, Sarah Henrickson, Liming Yang, John Powell, Wyndham H. Wilson, Elaine S. Jaffe, Richard Simon, Richard D. Klausner, Emilio Montserrat, Francesc Bosch, Timothy C. Greiner, Dennis D. Weisenburger, Warren G. Sanger, Bhavana J. Dave, James C. Lynch, Julie Vose, James O. Armitage, Richard I. Fisher, Thomas P. Miller, Michael LeBlanc, German Ott, Stein Kvaloy, Harald Holte, Jan Delabie, and Louis M. Staudt. The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell*, 3(2):185–197, February 2003. ISSN 1535-6108.

Lao H. Saal, Peter Johansson, Karolina Holm, Sofia K. Gruvberger-Saal, Qing-Bai She, Matthew Maurer, Susan Koujak, Adolfo A. Ferrando, Per Malmström, Lorenzo Memeo, and others. Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. *Proceedings of the National Academy of Sciences*, 104(18):7564–7569, 2007. URL http://www.pnas.org/content/104/18/7564.short.

W. A. Sakr, G. P. Haas, B. F. Cassin, J. E. Pontes, and J. D. Crissman. The frequency of carcinoma and intraepithelial neoplasia of the prostate in young male patients. *The Journal of Urology*, 150(2 Pt 1):379–385, August 1993. ISSN 0022-5347.

Marta Sanchez-Carbayo, Nicholas D. Socci, Juanjo Lozano, Fabien Saint, and Carlos Cordon-Cardo. Defining Molecular Profiles of Poor Outcome in Patients With Invasive Bladder Cancer Using Oligonucleotide Microarrays. *Journal of Clinical Oncology*, 24(5):778–789, February 2006. ISSN 0732-183X, 1527-7755. DOI: 10.1200/JCO.2005.03.2375. URL http://jco.ascopubs.org/content/24/5/778.

Rickard Sandberg and Ola Larsson. Improved precision and accuracy for microarrays using updated probe set definitions. *BMC Bioinformatics*, 8

(1):48, February 2007. ISSN 1471-2105. DOI: 10.1186/1471-2105-8-48. URL http://www.biomedcentral.com/1471-2105/8/48/abstract.

Ana Sastre-Perona and Pilar Santisteban. Role of the Wnt pathway in thyroid cancer. *Cancer Endocrinology*, 3:31, 2012. DOI: 10.3389/fendo.2012.00031. URL http://journal.frontiersin.org/Journal/10.3389/fendo.2012.00031/full.

Mark Schena, Dari Shalon, Ronald W. Davis, and Patrick O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, 1995. URL http://www.sciencemag.org/content/270/5235/467.short.

Catherine A Schnabel and Mark G Erlander. Gene expression-based diagnostics for molecular cancer classification of difficult to diagnose tumors. *Expert Opinion on Medical Diagnostics*, 6(5):407–419, July 2012. ISSN 1753-0059. DOI: 10.1517/17530059.2012.704363. URL http://informahealthcare.com/doi/abs/10.1517/17530059.2012.704363.

D. W. Selinger, K. J. Cheung, R. Mei, E. M. Johansson, C. S. Richmond, F. R. Blattner, D. J. Lockhart, and G. M. Church. RNA expression analysis using a 30 base pair resolution Escherichia coli genome array. *Nature Biotechnology*, 18(12):1262–1268, December 2000. ISSN 1087-0156. DOI: 10.1038/82367.

K. Shakya, H. J. Ruskin, G. Kerr, M. Crane, and J. Becker. Comparison of Microarray Preprocessing Methods. In Hamid R. Arabnia, editor, *Advances in Computational Biology*, number 680 in Advances in Experimental Medicine and Biology, pages 139–147. Springer New York, 2010. ISBN 978-1-4419-5912-6 978-1-4419-5913-3. URL http://link.springer.com/chapter/10.1007/978-1-4419-5913-3_16.

Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, 14(9):618–630, September 2013. ISSN 1471-0056. DOI: 10.1038/nrg3542. URL http://www.nature.com/nrg/journal/v14/n9/abs/nrg3542.html.

Leming Shi, Weida Tong, Hong Fang, Uwe Scherf, Jing Han, Raj K. Puri, Felix W. Frueh, Federico M. Goodsaid, Lei Guo, Zhenqiang Su, Tao Han, James C. Fuscoe, Z. Alex Xu, Tucker A. Patterson, Huixiao Hong, Qian Xie, Roger G. Perkins, James J. Chen, and Daniel A. Casciano. Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential. *BMC bioinformatics*, 6 Suppl 2:S12, July 2005. ISSN 1471-2105. DOI: 10.1186/1471-2105-6-S2-S12.

Oliver M. Sieber, Karl Heinimann, and Ian P. M. Tomlinson. Genomic instability–the engine of tumorigenesis? *Nature Reviews. Cancer*, 3(9):701–708, September 2003. ISSN 1474-175X. DOI: 10.1038/nrc1170.

Kimberly D. Siegmund, Paul Marjoram, Yen-Jung Woo, Simon Tavaré, and Darryl Shibata. Inferring clonal expansion and cancer stem cell dynamics

from DNA methylation patterns in colorectal cancers. *Proceedings of the National Academy of Sciences*, 106(12):4828–4833, March 2009. ISSN 0027-8424, 1091-6490. DOI: 10.1073/pnas.0810276106. URL http://www.pnas.org/content/106/12/4828.

R. Simon. Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. *British Journal of Cancer*, 89(9):1599–1604, November 2003. ISSN 0007-0920. DOI: 10.1038/sj.bjc.6601326.

Leslie H. Sobin. TNM: evolution and relation to other prognostic factors. *Seminars in Surgical Oncology*, 21(1):3–7, 2003. ISSN 8756-0437. DOI: 10.1002/ssu.10014.

Xavier Solé, Núria Bonifaci, Núria López-Bigas, Antoni Berenguer, Pilar Hernández, Oscar Reina, Christopher A. Maxwell, Helena Aguilar, Ander Urruticoechea, Silvia de Sanjosé, Francesc Comellas, Gabriel Capellá, Víctor Moreno, and Miguel Angel Pujana. Biological convergence of cancer signatures. *PloS One*, 4(2):e4544, 2009. ISSN 1932-6203. DOI: 10.1371/journal.pone.0004544.

Christos Sotiriou and Lajos Pusztai. Gene-expression signatures in breast cancer. *New England Journal of Medicine*, 360(8):790–800, 2009. URL http://www.nejm.org/doi/full/10.1056/NEJMra0801289.

Christos Sotiriou, Pratyaksha Wirapati, Sherene Loi, Adrian Harris, Steve Fox, Johanna Smeds, Hans Nordgren, Pierre Farmer, Viviane Praz, Benjamin Haibe-Kains, and others. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*, 98(4):262–272, 2006. URL http://jnci.oxfordjournals.org/content/98/4/262.short.

Joseph A. Sparano, Robert J. Gray, Della F. Makower, Kathleen I. Pritchard, Kathy S. Albain, Daniel F. Hayes, Charles E. Jr. Geyer, Elizabeth C. Dees, Edith A. Perez, John A. Jr. Olson, JoAnne Zujewski, Tracy Lively, Sunil S. Badve, Thomas J. Saphner, Lynne I. Wagner, Timothy J. Whelan, Matthew J. Ellis, Soonmyung Paik, William C. Wood, Peter Ravdin, Maccon M. Keane, Henry L. Gomez Moreno, Pavan S. Reddy, Timothy F. Goggins, Ingrid A. Mayer, Adam M. Brufsky, Deborah L. Toppmeyer, Virginia G. Kaklamani, James N. Atkins, Jeffrey L. Berenberg, and George W. Sledge. Prospective Validation of a 21-Gene Expression Assay in Breast Cancer. *New England Journal of Medicine*, 373(21):2005–2014, November 2015. ISSN 0028-4793. DOI: 10.1056/NEJMoa1510764. URL http://dx.doi.org/10.1056/NEJMoa1510764.

Mansi Srivastava, Oleg Simakov, Jarrod Chapman, Bryony Fahey, Marie E. A. Gauthier, Therese Mitros, Gemma S. Richards, Cecilia Conaco, Michael Dacre, Uffe Hellsten, Claire Larroux, Nicholas H. Putnam, Mario Stanke, Maja Adamska, Aaron Darling, Sandie M. Degnan, Todd H. Oakley, David C. Plachetzki, Yufeng Zhai, Marcin Adamski, Andrew Calcino, Scott F. Cummins, David M. Goodstein, Christina Harris, Daniel J. Jackson, Sally P. Leys, Shengqiang Shu, Ben J. Woodcroft, Michel Vervoort,

Kenneth S. Kosik, Gerard Manning, Bernard M. Degnan, and Daniel S. Rokhsar. The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature*, 466(7307):720–726, August 2010. ISSN 1476-4687. DOI: 10.1038/nature09201.

Philip J. Stephens, Chris D. Greenman, Beiyuan Fu, Fengtang Yang, Graham R. Bignell, Laura J. Mudie, Erin D. Pleasance, King Wai Lau, David Beare, Lucy A. Stebbings, Stuart McLaren, Meng-Lay Lin, David J. McBride, Ignacio Varela, Serena Nik-Zainal, Catherine Leroy, Mingming Jia, Andrew Menzies, Adam P. Butler, Jon W. Teague, Michael A. Quail, John Burton, Harold Swerdlow, Nigel P. Carter, Laura A. Morsberger, Christine Iacobuzio-Donahue, George A. Follows, Anthony R. Green, Adrienne M. Flanagan, Michael R. Stratton, P. Andrew Futreal, and Peter J. Campbell. Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. *Cell*, 144(1):27–40, July 2011. ISSN 0092-8674. DOI: 10.1016/j.cell.2010.11.055. URL http://www.cell.com/article/S0092867410013772/abstract.

Andrew I. Su, John B. Welsh, Lisa M. Sapinoso, Suzanne G. Kern, Petre Dimitrov, Hilmar Lapp, Peter G. Schultz, Steven M. Powell, Christopher A. Moskaluk, Henry F. Frierson, and Garret M. Hampton. Molecular Classification of Human Carcinomas by Use of Gene Expression Signatures. *Cancer Research*, 61(20):7388–7393, October 2001. ISSN 0008-5472, 1538-7445. URL http://cancerres.aacrjournals.org/content/61/20/7388.

Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, October 2005. ISSN 0027-8424, 1091-6490. DOI: 10.1073/pnas.0506580102. URL http://www.pnas.org/content/102/43/15545.

Malin Sund and Raghu Kalluri. Tumor stroma derived biomarkers in cancer. *Cancer and Metastasis Reviews*, 28(1-2):177–183, March 2009. ISSN 0167-7659, 1573-7233. DOI: 10.1007/s10555-008-9175-2. URL http://link.springer.com/article/10.1007/s10555-008-9175-2.

Therese Sørlie, Charles M. Perou, Robert Tibshirani, Turid Aas, Stephanie Geisler, Hilde Johnsen, Trevor Hastie, Michael B. Eisen, Matt van de Rijn, Stefanie S. Jeffrey, Thor Thorsen, Hanne Quist, John C. Matese, Patrick O. Brown, David Botstein, Per Eystein Lønning, and Anne-Lise Børresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874, September 2001. ISSN 0027-8424, 1091-6490. DOI: 10.1073/pnas.191367098. URL http://www.pnas.org/content/98/19/10869.

Therese Sørlie, Robert Tibshirani, Joel Parker, Trevor Hastie, J. S. Marron, Andrew Nobel, Shibing Deng, Hilde Johnsen, Robert Pesich,

Stephanie Geisler, and others. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences*, 100(14):8418–8423, 2003. URL http://www.pnas.org/content/100/14/8418.short.

Jonatan Taminau, David Steenhoff, Alain Coletta, Stijn Meganck, Cosmin Lazar, Virginie de Schaetzen, Robin Duque, Colin Molter, Hugues Bersini, Ann Nowé, and others. inSilicoDb: an R/Bioconductor package for accessing human Affymetrix expert-curated datasets from GEO. *Bioinformatics*, 27(22):3204–3205, 2011. URL http://bioinformatics.oxfordjournals.org/content/27/22/3204.short.

Paul K. Tan, Thomas J. Downey, Edward L. Spitznagel, Pin Xu, Dadin Fu, Dimiter S. Dimitrov, Richard A. Lempicki, Bruce M. Raaka, and Margaret C. Cam. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Research*, 31(19):5676–5684, October 2003. ISSN 1362-4962.

T. Tanaka, K. Umeki, I. Yamamoto, S. Sugiyama, S. Noguchi, and S. Ohtaki. Immunohistochemical loss of thyroid peroxidase in papillary thyroid carcinoma: strong suppression of peroxidase gene expression. *The Journal of Pathology*, 179(1):89–94, May 1996. ISSN 0022-3417. DOI: 10.1002/(SICI)1096-9896(199605)179:1<89::AID-PATH546>3.0.CO;2-R.

M. Tarabichi, A. Antoniou, M. Saiselet, J. M. Pita, G. Andry, J. E. Dumont, V. Detours, and C. Maenhaut. Systems biology of cancer: entropy, disorder, and selection-driven evolution to independence, invasion and "swarm intelligence". *Cancer Metastasis Reviews*, 32(3-4):403–421, December 2013. ISSN 1573-7233. DOI: 10.1007/s10555-013-9431-y.

Adi Laurentiu Tarca, Sorin Draghici, Purvesh Khatri, Sonia S. Hassan, Pooja Mittal, Jung-sun Kim, Chong Jai Kim, Juan Pedro Kusanovic, and Roberto Romero. A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75–82, January 2009. ISSN 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btn577. URL http://bioinformatics.oxfordjournals.org/content/25/1/75.

Joseph H. Taube, Jason I. Herschkowitz, Kakajan Komurov, Alicia Y. Zhou, Supriya Gupta, Jing Yang, Kimberly Hartwell, Tamer T. Onder, Piyush B. Gupta, Kurt W. Evans, and others. Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. *Proceedings of the National Academy of Sciences*, 107(35):15449–15454, 2010. URL http://www.pnas.org/content/107/35/15449.short.

Andrew E. Teschendorff, Ali Naderi, Nuno L. Barbosa-Morais, Sarah E. Pinder, Ian O. Ellis, Sam Aparicio, James D. Brenton, and Carlos Caldas. A consensus prognostic gene expression classifier for ER positive breast cancer. *Genome biology*, 7(10):R101, 2006. URL http://www.biomedcentral.com/1465-6906/7/R101.

The Cancer Genome Atlas Research Network, John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger,

Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45 (10):1113–1120, October 2013. ISSN 1061-4036. DOI: 10.1038/ng.2764. URL http://www.nature.com/ng/journal/v45/n10/full/ng.2764.html.

Jean Paul Thiery and Jonathan P. Sleeman. Complex networks orchestrate epithelial–mesenchymal transitions. *Nature Reviews Molecular Cell Biology*, 7(2):131–142, February 2006. ISSN 1471-0072. DOI: 10.1038/nrm1835. URL http://www.nature.com/nrm/journal/v7/n2/full/nrm1835.html.

Jean Paul Thiery, Hervé Acloque, Ruby Y. J. Huang, and M. Angela Nieto. Epithelial-Mesenchymal Transitions in Development and Disease. *Cell*, 139(5):871–890, November 2009. ISSN 0092-8674. DOI: 10.1016/j.cell.2009.11.007. URL http://www.sciencedirect.com/science/article/pii/S0092867409014196.

Lieven Thorrez, Katrijn Van Deun, Léon-Charles Tranchevent, Leentje Van Lommel, Kristof Engelen, Kathleen Marchal, Yves Moreau, Iven Van Mechelen, and Frans Schuit. Using ribosomal protein genes as reference: a tale of caution. *PloS One*, 3(3):e1854, 2008. ISSN 1932-6203. DOI: 10.1371/journal.pone.0001854.

Anna V. Tinker, Alex Boussioutas, and David D. L. Bowtell. The challenges of gene expression microarrays for the study of human cancer. *Cancer Cell*, 9(5):333–339, May 2006. ISSN 1535-6108. DOI: 10.1016/j.ccr.2006.05.001.

Franck Toledo and Geoffrey M. Wahl. Regulating the p53 pathway: in vitro hypotheses, in vivo veritas. *Nature Reviews. Cancer*, 6(12):909–923, December 2006. ISSN 1474-175X. DOI: 10.1038/nrc2012.

G. Tomás, M. Tarabichi, D. Gacquer, A. Hébrant, G. Dom, J. E. Dumont, X. Keutgen, T. J. Fahey, C. Maenhaut, and V. Detours. A general method to derive robust organ-specific gene expression-based differentiation indices: application to thyroid cancer diagnostic. *Oncogene*, 31(41):4490–4498, October 2012. ISSN 1476-5594. DOI: 10.1038/onc.2011.626.

William D. Travis, Elisabeth Brambilla, and Gregory J. Riely. New Pathologic Classification of Lung Cancer: Relevance for Clinical Practice and Clinical Trials. *Journal of Clinical Oncology*, 31(8):992–1001, March 2013. ISSN 0732-183X, 1527-7755. DOI: 10.1200/JCO.2012.46.9270. URL http://jco.ascopubs.org/content/31/8/992.

Y. Tsujimoto, J. Gorham, J. Cossman, E. Jaffe, and C. M. Croce. The t(14;18) chromosome translocations involved in B-cell neoplasms result from mistakes in VDJ joining. *Science (New York, N.Y.)*, 229(4720):1390–1393, September 1985. ISSN 0036-8075.

Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001. URL http://www.pnas.org/content/98/9/5116.short.

Graham J. G. Upton, Olivia Sanchez-Graillet, Joanna Rowsell, Jose M. Arteaga-Salas, Neil S. Graham, Maria A. Stalteri, Farhat N. Memon, Sean T. May, and Andrew P. Harrison. On the causes of outliers in Affymetrix GeneChip data. *Briefings in Functional Genomics & Proteomics*, 8(3):199–212, May 2009. ISSN 2041-2649, 2041-2657. DOI: 10.1093/bfgp/elp027. URL http://bfg.oxfordjournals.org/content/8/3/199.

Scott Valastyan and Robert A. Weinberg. Tumor Metastasis: Molecular Insights and Evolving Paradigms. *Cell*, 147(2):275–292, October 2011. ISSN 0092-8674. DOI: 10.1016/j.cell.2011.09.024. URL http://www.cell.com/article/S0092867411010853/abstract.

Marc J. Van De Vijver, Yudong D. He, Laura J. van't Veer, Hongyue Dai, Augustinus AM Hart, Dorien W. Voskuil, George J. Schreiber, Johannes L. Peterse, Chris Roberts, Matthew J. Marton, and others. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009, 2002. URL http://www.nejm.org/doi/full/10.1056/nejmoa021967.

Wilma C. G. van Staveren, Sandrine Beeckman, Gil Tomás, Geneviève Dom, Aline Hébrant, Laurent Delys, Marjolein J. Vliem, Christophe Trésallet, Guy Andry, Brigitte Franc, Frédérick Libert, Jacques E. Dumont, and Carine Maenhaut. Role of Epac and protein kinase A in thyrotropin-induced gene expression in primary thyrocytes. *Experimental Cell Research*, 318(5):444–452, March 2012. ISSN 0014-4827. DOI: 10.1016/j.yexcr.2011.12.022. URL http://www.sciencedirect.com/science/article/pii/S0014482711005039.

Wilma CG van Staveren, David Weiss Solís, Laurent Delys, David Venet, Matteo Cappello, Guy Andry, Jacques E. Dumont, Frédérick Libert, Vincent Detours, and Carine Maenhaut. Gene expression in human thyrocytes and autonomous adenomas reveals suppression of negative feedbacks in tumorigenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 103(2):413–418, 2006. URL http://www.pnas.org/content/103/2/413.short.

Laura J. van't Veer, Hongyue Dai, Marc J. Van De Vijver, Yudong D. He, Augustinus AM Hart, Mao Mao, Hans L. Peterse, Karin van der Kooy, Matthew J. Marton, Anke T. Witteveen, and others. Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871):530–536, 2002. URL http://www.nature.com/nature/journal/v415/n6871/abs/415530a.html.

James R. Vasselli, Joanna H. Shih, Shuba R. Iyengar, Jodi Maranchie, Joseph Riss, Robert Worrell, Carlos Torres-Cabala, Ray Tabios, Andra Mariotti, Robert Stearman, Maria Merino, McClellan M. Walther, Richard Simon, Richard D. Klausner, and W. Marston Linehan. Predicting survival in patients with metastatic kidney cancer by gene-expression profiling in the primary tumor. *Proceedings of the National Academy of Sciences*, 100(12):6958–6963, June 2003. ISSN 0027-8424, 1091-6490. DOI: 10.1073/pnas.1131754100. URL http://www.pnas.org/content/100/12/6958.

David Venet, Jacques E. Dumont, and Vincent Detours. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS computational biology*, 7(10):e1002240, October 2011. ISSN 1553-7358. DOI: 10.1371/journal.pcbi.1002240.

David Venet, Vincent Detours, and Hugues Bersini. A measure of the signal-to-noise ratio of microarray samples and studies using gene correlations. *PloS One*, 7(12):e51013, 2012. ISSN 1932-6203. DOI: 10.1371/journal.pone.0051013.

R. Virchow. Cellular pathology: as based upon physiological and pathological histology. translated by F. *Chance, JB Lippincott Philadelphia*, 1863.

Jane E. Visvader and Geoffrey J. Lindeman. Cancer Stem Cells: Current Status and Evolving Complexities. *Cell Stem Cell*, 10(6):717–728, June 2012. ISSN 1934-5909. DOI: 10.1016/j.stem.2012.05.007. URL http://www.cell.com/article/S1934590912002408/abstract.

K David Voduc, Torsten O Nielsen, Charles M Perou, J Chuck Harrell, Cheng Fan, Hagen Kennecke, Andy J Minn, Vincent L Cryns, and Maggie C U Cheang. αB-crystallin expression in breast cancer is associated with brain metastasis. *npj Breast Cancer*, 1:15014, October 2015. ISSN 2374-4677. DOI: 10.1038/npjbcancer.2015.14. URL http://www.nature.com/articles/npjbcancer201514.

B. Vogelstein and K. W. Kinzler. The multistep nature of cancer. *Trends in genetics: TIG*, 9(4):138–141, April 1993. ISSN 0168-9525.

B. Vogelstein, D. Lane, and A. J. Levine. Surfing the p53 network. *Nature*, 408(6810):307–310, November 2000. ISSN 0028-0836. DOI: 10.1038/35042675.

Bert Vogelstein and Kenneth W. Kinzler. Cancer genes and the pathways they control. *Nature medicine*, 10(8):789–799, 2004. URL http://www.nature.com/nm/journal/v10/n8/abs/nm1087.html.

Eric T. Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, November 2008. ISSN 1476-4687. DOI: 10.1038/nature07509.

Yixin Wang, Jan GM Klijn, Yi Zhang, Anieta M. Sieuwerts, Maxime P. Look, Fei Yang, Dmitri Talantov, Mieke Timmermans, Marion E. Meijer-van Gelder, Jack Yu, and others. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, 365 (9460):671–679, 2005. URL http://www.sciencedirect.com/science/article/pii/S0140673605179471.

Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, 10(1):57–63, January 2009. ISSN 1471-0064. DOI: 10.1038/nrg2484.

Otto Warburg. On the Origin of Cancer Cells. *Science*, 123(3191):309–314, February 1956. ISSN 0036-8075, 1095-9203. DOI: 10.1126/science.123.3191.309. URL http://www.sciencemag.org/content/123/3191/309.

John C. Warren. Inhalation of ethereal vapor for the prevention of pain in surgical operations. *The Boston Medical and Surgical Journal*, 35 (19):375–379, 1846. URL http://www.nejm.org/doi/pdf/10.1056/NEJM184612090351902.

James D. Watson and Francis HC Crick. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953. URL http://www.nature.com/physics/looking-back/crick/.

Andrew R. Webb. *Statistical pattern recognition*. John Wiley & Sons, 2003. URL https://books.google.com/books?hl=en&lr=&id=ivMBWCe_f0gC&oi=fnd&pg=PR7&dq=statistical+pattern+recognition&ots=HG_owdml4B&sig=Dcsar28SVqNTZ_NDkqamtTMRHbE.

Britta Weigelt, Frederick L Baehner, and Jorge S Reis-Filho. The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. *The Journal of Pathology*, 220(2):263–280, January 2010. ISSN 1096-9896. DOI: 10.1002/path.2648. URL http://onlinelibrary.wiley.com/doi/10.1002/path.2648/abstract.

Britta Weigelt, Lajos Pusztai, Alan Ashworth, and Jorge S. Reis-Filho. Challenges translating breast cancer gene signatures into the clinic. *Nature Reviews Clinical Oncology*, 9(1):58–64, January 2012. ISSN 1759-4774. DOI: 10.1038/nrclinonc.2011.125. URL http://www.nature.com/nrclinonc/journal/v9/n1/abs/nrclinonc.2011.125.html.

Britta Weigelt, Cyrus M. Ghajar, and Mina J. Bissell. The need for complex 3d culture models to unravel novel pathways and identify accurate biomarkers in breast cancer. *Advanced Drug Delivery Reviews*, 69–70:42–51, April 2014. ISSN 0169-409X. DOI: 10.1016/j.addr.2014.01.001. URL http://www.sciencedirect.com/science/article/pii/S0169409X14000027.

R. A. Weinberg. Oncogenes and the molecular biology of cancer. *The Journal of Cell Biology*, 97(6):1661–1662, December 1983. ISSN 0021-9525.

Robert A. Weinberg. *The Biology of Cancer*. Garland Science, New York, NY, US, 2 edition edition, June 2013. ISBN 978-0-8153-4220-5.

Alana L. Welm, Julie B. Sneddon, Carmen Taylor, Dimitry SA Nuyten, Marc J. van de Vijver, Bruce H. Hasegawa, and J. Michael Bishop. The macrophage-stimulating protein pathway promotes metastasis in a mouse model for breast cancer and predicts poor prognosis in humans. *Proceedings of the National Academy of Sciences*, 104(18):7570–7575, 2007. URL http://www.pnas.org/content/104/18/7570.short.

Robert B. West, Dimitry SA Nuyten, Subbaya Subramanian, Torsten O. Nielsen, Christopher L. Corless, Brian P. Rubin, Kelli Montgomery, Shirley Zhu, Rajiv Patel, Tina Hernandez-Boussard, and others. Determination of stromal signatures in breast carcinoma. *PLoS biology*, 3(6):e187, 2005. URL http://dx.plos.org/10.1371/journal.pbio.0030187.

Michael L. Whitfield, Lacy K. George, Gavin D. Grant, and Charles M. Perou. Common markers of proliferation. *Nature Reviews. Cancer*, 6(2): 99–106, February 2006. ISSN 1474-175X. DOI: 10.1038/nrc1802.

Pratyaksha Wirapati, Christos Sotiriou, Susanne Kunkel, Pierre Farmer, Sylvain Pradervand, Benjamin Haibe-Kains, Christine Desmedt, Michail Ignatiadis, Thierry Sengstag, Frédéric Schütz, Darlene R. Goldstein, Martine Piccart, and Mauro Delorenzi. Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast cancer research: BCR*, 10(4): R65, 2008. ISSN 1465-542X. DOI: 10.1186/bcr2124.

Christopher Workman, Lars Juhl Jensen, Hanne Jarmer, Randy Berka, Laurent Gautier, Henrik Bjørn Nielser, Hans-Henrik Saxild, Claus Nielsen, Søren Brunak, and Steen Knudsen. A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biology*, 3(9):research0048.1–research0048.16, 2002. ISSN 1465-6906. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC126873/.

Zhijin Wu, Rafael A Irizarry, Robert Gentleman, Francisco Martinez-Murillo, and Forrest Spencer. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association*, 99(468):909–917, December 2004. ISSN 0162-1459. DOI: 10.1198/016214504000000683. URL http://amstat.tandfonline.com/doi/abs/10.1198/016214504000000683.

Yee H. Yang, Sandrine Dudoit, Percy Luu, and Terence P. Speed. Normalization for cDNA microarry data. In *BiOS 2001 The International Symposium on Biomedical Optics*, pages 141–152. International Society for Optics and Photonics, 2001. URL http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=901069.

Yee Hwa Yang, Sandrine Dudoit, Percy Luu, David M. Lin, Vivian Peng, John Ngai, and Terence P. Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic acids research*, 30(4):e15–e15, 2002. URL http://nar.oxfordjournals.org/content/30/4/e15.short.

M Zannini, V Avantaggiato, E Biffali, M I Arnone, K Sato, M Pischetola, B A Taylor, S J Phillips, A Simeone, and R Di Lauro. TTF-2, a new forkhead protein, shows a temporal expression in the developing thyroid which is consistent with a role in controlling the onset of differentiation. *The EMBO Journal*, 16(11):3185–3197, June 1997. ISSN 0261-4189. DOI: 10.1093/emboj/16.11.3185. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1169936/.