University of Liège - Faculty of Medicine
GIGA-Research, Unit of Human Genetics

Bioinformatics contribution to the analysis of omics data in the clinical, technical, and molecular domains of human cancer

---

Ir. Stephane Wenric

Supervisors

Prof. Vincent Bours
Department of Genetics
University of Liège

Prof. Guy Jerusalem
Department of Oncology
University of Liège

PhD dissertation
June 2017

Foreword

Some say bioinformatics is not science *per se*. It's merely a tool.

I say *what a wonderful tool, it is, then!*

There is no field that I know of which is more interdisciplinary.

The recent advances in next-generation sequencing have opened the door to a deluge of data. One should be cautious, though, as data is not a synonym of knowledge.

The task which I have considered to be mine during those 4 years was the transformation of data into scientific knowledge. This not an automatic process. There is no algorithm which, given a large set of files, verifies a hypothesis.

The interdisciplinarity of bioinformatics forces oneself to become learned in state-of-the-art methods from computer science or statistics, it pressures one to understand intricacies of biology.

I feel fortunate to have entered this field at what I believe is still its infancy. I have witnessed the development of new methods from scratch, the reuse of existing tools from other fields, and the almost infinite need for new developments.

This abundance of technical and biological data, the inherent interdisciplinarity, and the exponential necessity for innovative analyses are blessings. But they may look like a curse when one is trying to find a unifying thread to link all the developments and results accomplished in a thesis' time.

I hope the reader will be understanding while browsing these pages. My short experience tells me that bioinformatics is messy, but it's a pleasant way of finding answers. My desire is that, in addition to contributing to the advance of the addressed fields, this pleasantness is reflected in the following chapters.

Abstract

Human cancer is a disease of the cell and of the genome. In this work, I apply machine learning, algorithmics, and software engineering methods to genomics and transcriptomics data arising from cancer related questions.

My contributions are divided into three parts. First, this thesis treats a technical question: the detection of copy-number variations (CNVs) in the genome of multiple myeloma patients. Although routinely done with an existing technique (comparative genomic hybridization), the question of the feasibility of this kind of analysis with next-generation sequencing data had already risen. Here, we examine the process of detecting CNVs based on the whole exome sequence of tumoral cells and in the absence of the exome sequence of germinal cells. We show that usual comparisons to assess the accuracy of a CNV profile are insufficient, and we propose a new method to compare CNV profiles. Based on this method, we show that it is possible to accurately detect CNVs in multiple myeloma patients based on exome sequencing data without matched healthy tissue, if one uses a pool of unrelated healthy individuals as control sample.

Second, the diagnostic aspect of breast cancer is considered, as we get into the development of a non-invasive, blood based, diagnostic tool for breast cancer. Based on a cohort of 378 women, we have shown that the levels of circulating microRNAs can be used as biomarkers for the diagnostic of breast cancer. We designed a diagnostic model using 8 microRNAs, whose levels are combined through the use of the Random Forests algorithm. Furthermore, the specificity of our diagnostic model is assessed on patients in remission, gynecologic tumors, and metastatic breast cancers.

In the last part, the use of whole stranded RNA sequencing data on a cohort of 23 ER+/HER2- breast cancer patients yields various results regarding the potential role of antisense long non-coding RNAs and their disruption in tumors. Three different gene extraction methods are presented. For each of these methods, the

corresponding gene set is reviewed and assayed in survival data from an external cohort of a thousand breast cancer patients.

Résumé

Le cancer humain est une maladie de la cellule et du génome. Dans ce travail, j'ai appliqué des méthodes issues de l'apprentissage automatique, de l'algorithmique, et du développement logiciel, afin d'analyser des données génomiques et transcriptomiques relatives à des problématiques liées au cancer.

Mes contributions sont séparées en trois parties. Premièrement, cette thèse aborde une question technique : la détection de variants de nombre de copies (CNV) dans le génome de patients atteints de myélome multiple. Bien que réalisée usuellement à l'aide d'une technique existante (CGH), la question de la faisabilité de ce genre d'analyse à l'aide de techniques dite de *next-generation sequencing* a déjà été abordée. Ici, nous examinons le processus de détection de CNVs sur base de la séquence complète d'exome de la cellule tumorale et en l'absence de la séquence d'exome de celulles germinales. Nous montrons que les comparaisons habituellement utilisées pour déterminer la validité d'un profil de CNVs sont insuffisantes, et nous proposons une nouvelle méthode de comparaison de profils de CNVs. Sur base de cette méthode, nous montrons qu'il est possible de détecter des CNVs chez les patients atteints de myélome multiple sur base de la séquence d'exome et en l'absence de tissu sain correspondant, en utilisant un *pool* d'échantillons sains comme contrôle.

Ensuite, l'aspect diagnostic du cancer du sein est abordé, dans le cadre du développement d'un outil de diagnostic non-invasif du cancer du sein basé sur une prise de sang. En partant d'une cohorte de 378 femmes, nous avons démontré que les niveaux de microARNs circulants peuvent être utilisés comme biomarqueurs pour le diagnostic du cancer du sein. Nous avons conçu un outil diagnostic sur base de 8 microARNS, dont les niveaux sont combinés à l'aide de l'algorithme Random Forests. En outre, la spécificité de notre outil diagnostic est évaluée sur des patients en rémission, des tumeurs gynécologiques, et des cancers du sein métastatiques.

Dans la dernière partie de cette thèse, l'utilisation de données de séquençage ARN brin-spécifique sur une cohorte de 23 patientes atteintes de cancer du sein ER+/HER2- a permis de mettre en évidence une série de résultats relatifs au rôle potentiel des longs ARN non-codants antisens et à leur perturbation dans la tumeur.

Trois méthodes différentes d'extraction de gènes sont présentées. Pour chacune de ces méthodes, la liste de gènes correspondant est passée en revue, et sa pertinence dans le cadre d'une analyse de survie est évaluée sur une cohorte externe d'un millier de patientes atteinte de cancer du sein.

Remerciements

Durant ces 4 années de travail, de nombreuses personnes m'ont aidé et m'ont apporté leur soutien.

J'aimerais tout d'abord remercier Vincent Bours et Guy Jerusalem, pour m'avoir initialement accordé leur confiance et permis d'entamer cette aventure. J'espère m'être montré digne de cette confiance et avoir fait grandir l'expertise de leurs laboratoires. Au-delà de cette confiance, leur compétence dans leurs domaines respectifs fut également une excellente source d'apprentissage. Je me dois également de remercier chaleureusement Wouter Coppieters pour son soutien apporté à un moment particulièrement opportun.

Je remercie bien évidemment Claire Josse, qui m'a accompagné et guidé tout au long de cette thèse. J'ai non seulement pu bénéficier de tes nombreux conseils et de ton expertise, mais tu m'as également permis de progresser et m'affirmer en tant que chercheur en m'accordant l'autonomie qui m'était nécessaire pour que j'explore par moi-même le champ des possibles. Lorsque je décris à de jeunes doctorants la situation idéale, je me réfère à cet équilibre délicat entre autonomie et guidance que j'ai connu grâce à toi.

Le travail de chercheur est intrinsèquement un travail d'équipe, et aucun des projets qui composent cette thèse n'aurait pu se réaliser seul. Cependant, c'est avant tout les amis plutôt que les collaborateurs que je remercie ici.
Pierre, quelle belle aventure nous avons vécu. Si l'étude que nous avons peaufiné ensemble fut un processus relativement long, chaque occasion de mettre en commun nos idées était à la fois constructive et agréable. Travailler avec toi fut un réel plaisir ainsi qu'une fenêtre grande ouverte sur le domaine de la recherche clinique que je connaissais relativement peu. Ce premier projet m'a rapidement convaincu de l'intérêt de la complémentarité et la pluridisciplinarité d'une équipe de recherche.
Tiberio, s'il est vrai que nous avons passé plus de temps à attendre les résultats du reviewing qu'à écrire notre papier, cette expérience n'en fut pas moins plaisante, tout autant que ces quelques années passées dans le même bureau !
Sonia, j'ai apprécié collaborer avec toi et confronter nos idées. Ce projet que nous avons réalisé est un autre bel exemple de l'intérêt d'une équipe pluri-disciplinaire

et complémentaire. Je te souhaite beaucoup de succès dans la suite de ta carrière académique.

Au-delà des collaborations, d'autres personnes croisées chaque jour ont fait de cette expérience un plaisir. Nadège, Marie, Meriem, Corinne, Jérôme, Guillaume, Aurélie, Aurélie, Christophe, Ahmed, merci d'avoir été des collègues et amis agréables au point qu'il nous arrive quelques fois d'oublier que nous étions au labo pour y travailler.

J'aimerais remercier Jean-Hubert, pour ses conseils avisés et ses corrections apportées à ce manuscrit.

Je tiens à remercier également Vincent Botta et Marie Schrynemackers pour leurs conseils et leur relecture du chapitre consacré au machine learning.

Je me dois bien sûr de remercier Marianne, dont la disponibilité, l'aide, et la gentillesse sont sans égales.

Je suis profondément reconnaissant envers mes parents et mon frère, qui m'ont toujours soutenu, ainsi qu'envers mes différents groupes d'amis, qui ont toujours été présents durant ces 4 années.

Contents

1 Introduction

> *I would rather have questions that can't be answered than answers that can't be questioned.*
>
> **— Richard Feynman**

Cancer

Hallmarks of cancer

Cancers are an ensemble of diseases involving cellular abnormalities. Among those abnormalities, ten are considered to form the *hallmarks of cancer* (see. Fig. 1.1). These hallmarks include sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inflammation, inducing angiogenesis, genomic instability, reprogramming the energy metabolism, evading immune destruction, and activating invasion and metastasis [1].

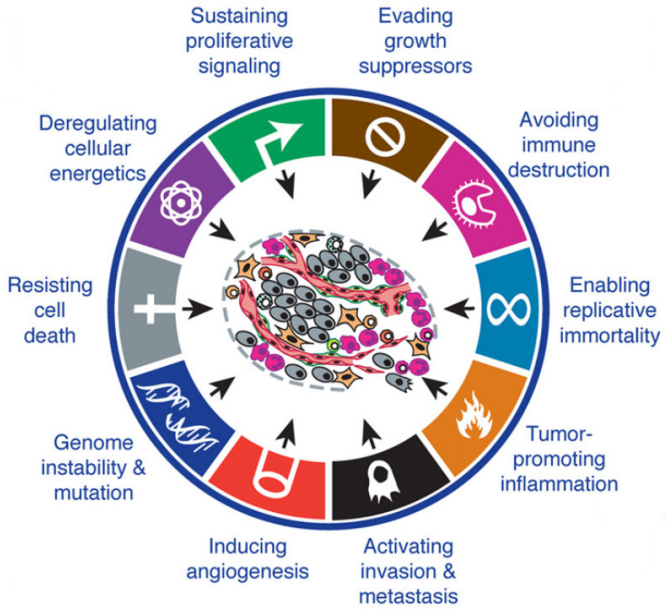


Fig. 1.1.: The hallmarks of cancer. (Hanahan & Weinberg [1])

Oncogenesis

These ten traits are an emergent property of both inflammation and the genomic instability of the cancer cell. Cancer is indeed a genetic disease and the genetic causes of cancer can either be inherited or acquired during one's lifetime.

Even though the abnormally large number of genetic mutations constitutes a characteristic of cancer cells, not all mutations play a role in the cancer progression. Most mutations do not even change the protein product of the mutated gene. Among the non-synonymous mutations, 95% are single-base substitutions. It should be noted that the average number of non-synonymous mutations varies greatly between tumor types (see Fig. 1.2) [2].

Several particular gene types play a role in cancer: (proto-)oncogenes, tumor suppressor genes, DNA repair genes.

Proto-oncogenes are genes which, at their normal state, do not cause cancers, but whose tumorigenic potential can be activated by retroviruses or by genetic alterations (mutations, amplifications in gene copy number, translocations) [3].
For example, *HER2/neu* is a proto-oncogene which codes for a cell surface receptor and whose over-expression plays a role in approximately 15-20% of breast cancers [4].

Tumor suppressor genes are involved in the repression of cancerous growth. They must be inactivated or lost for cancer to develop. Usually, a mutation, a loss, or an epigenetic modification of both alleles of the tumor suppressor gene is required for the cancer to progress.
For instance, *TP53* is a tumor suppressor gene whose protein, *p53*, plays a role in apoptosis, cell cycle arrest and senescence. It is involved in more than 50% of human cancers [5].

Both oncogenes and tumor suppressor genes rarely act alone in the tumorigenesis process. This process is indeed multistep. Tumor development arises from the natural selection of cells, where each additional mutation in an oncogene or a tumor suppressor gene will grant greater proliferation and selective advantage (see Fig. 1.3).

DNA repair genes, whose function is to correct errors in DNA during the cell division process, play an important role in preventing cancer progression. When their protein product is altered or absent, because of mutations or deletions or other type of alterations, the mutation rate is increased and oncogenesis is accelerated.

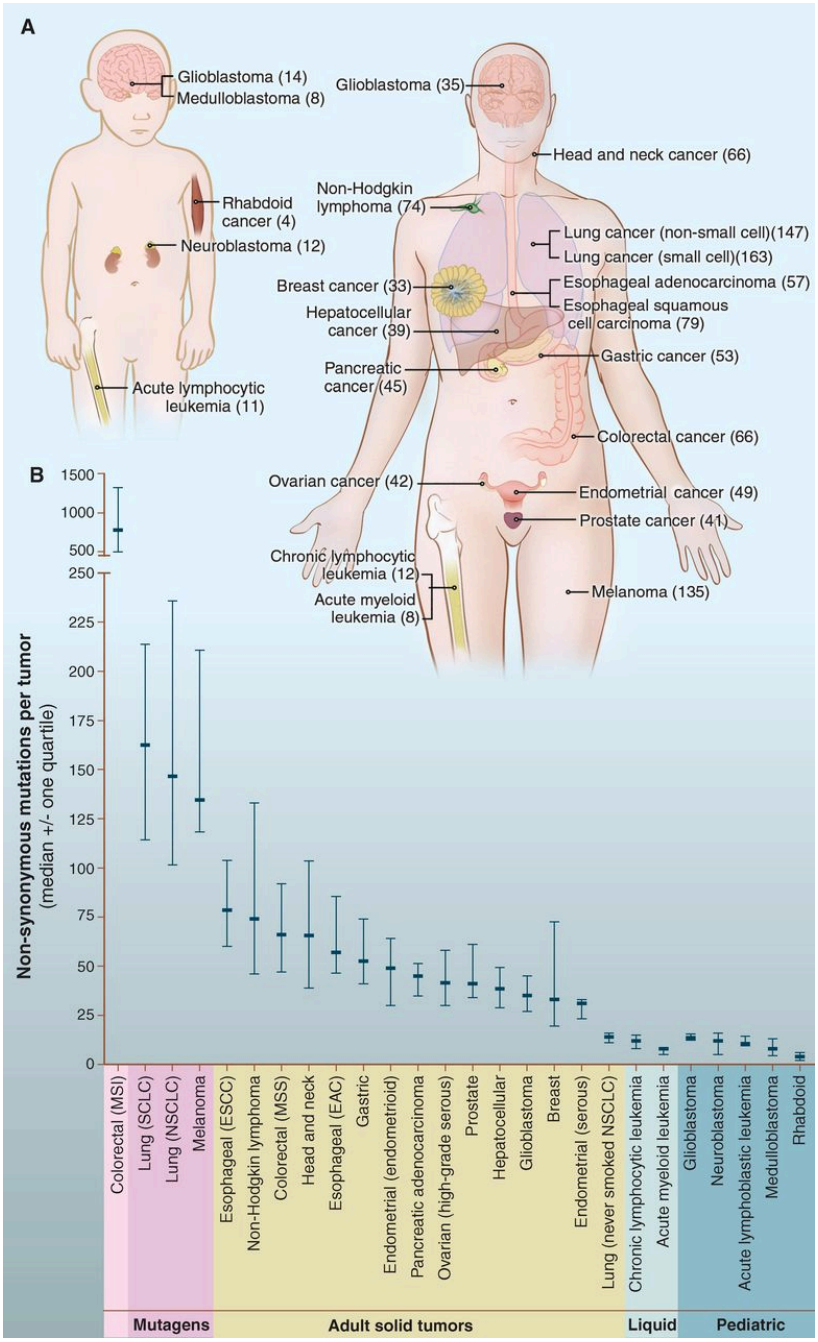


Fig. 1.2.: Median number of somatic non-synonymous mutations per tumor in representative human cancers. (Vogelstein *et al.* [2])

For example, *BRCA1* and *BRCA2* are DNA repair genes which are involved in most cases of hereditary breast and ovarian cancers [7].

The genetic mutations which play a role in cancers are called "**driver mutations**", as opposed to "passenger mutations", whose effect on the tumor cell survival and

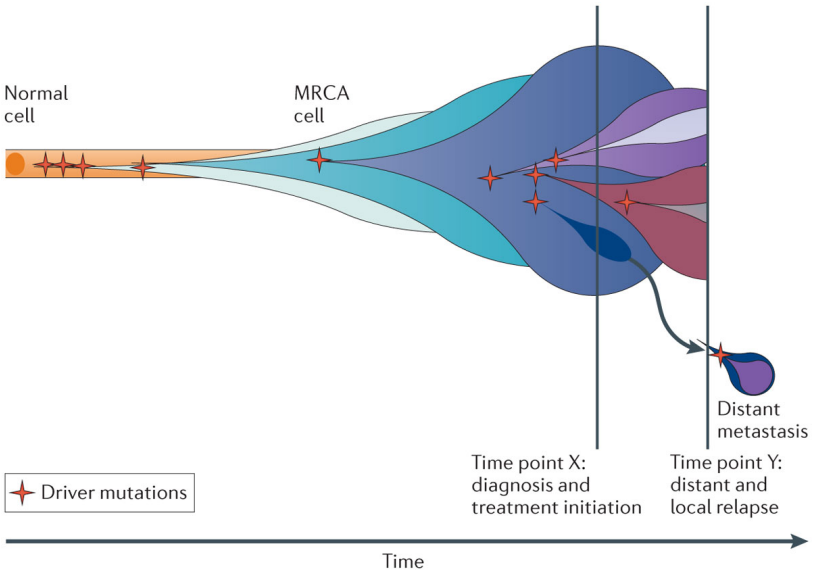


Fig. 1.3.: The multistep tumorigenic process is the consequence of the accumulation of driver mutations. Branching evolution results in competing subclones with diverse effects in terms of disease progression and severity. (Yates & Campbell [6])

proliferation is considered to be non-existent. Passenger mutations arise during the evolution of the cancer cell because of the deficiency in the DNA repair mechanisms [2].

Most genetic mutations related to cancer appear at low frequencies, but they tend to be linked to a small number of pathways, which play a key role in the tumor cell's survival. This low frequency both in the affected genes and the mutations makes it harder to look for statistically significant evidence linking genes and genetic mutations to cancer. Moreover, cancers can rarely be explained by a single mutation or a single affected gene. Genes act together, and inactivating mutations in different genes may show a phenotypic effect only when combined with several other protein-altering mutations [6].

The genetic mutations involved in cancer are not limited to single nucleotide substitutions. Large chromosomal aberrations also play a key role in some cancer types (see 1.2.1), as well as epigenetic modifications.

Epidemiology

More than 14 million new cases of cancer are diagnosed worldwide each year, with lung and breast cancers being the most frequent cancers in men and women respectively. In Belgium, more than 65000 cases of cancer were diagnosed in 2013 and this number is expected to rise to 78000 by 2025 [8, 9].

Cancer remains a leading cause of death worldwide, with more than 8 million deaths each year. In Belgium, more than 26 000 deaths were caused by cancer in 2012 [9, 10].

In Belgium, **breast cancer** is the most frequent cancer type in females, accounting for approximately 35% of all cancers, with more than 10000 cases diagnosed each year. It is also the first cause of cancer death in females (approximately 20% of all cancer deaths). The mean age at the diagnosis for breast cancer in Belgium is approximately 65. In the USA, it is approximately 68. Younger women have a lower incidence than older women [9, 11].

Multiple myeloma accounts for approximately 15% of lymphatohematopoietic cancers (LHC) and 2% of all cancers in the US. In Belgium, with more than 750 cases diagnosed each year, multiple myeloma represents approximately 1% of all cancers. The median age at diagnosis for multiple myeloma is 71 and it rarely affects people younger than 40. The median survival after diagnosis is approximately 3 years. Incidence increases with age, and it varies with ethnicity (higher prevalence in African-Americans compared to European-Americans) [9, 12, 13].

Breast cancer

Breast cancer is a disease involving uncontrolled growth of breast cells. It usually starts in the cells of the lobules or the ducts or, less frequently, in the stromal tissues.

Although all breast cancers are caused by genetic abnormalities, only 5-10% of breast cancers have a hereditary component and approximately 15% of breast cancer patients have a first degree parent (mother, sister, daughter) with breast cancer [14, 15].

Most hereditary breast cancers are associated with heterogeneous mutations in genes *BRCA1* and *BRCA2*, which are DNA repair genes. When there is a mutation in the *BRCA1* gene, the cumulated risk of breast cancer at age 70 is 65%. If a germline mutation affects the *BRCA2* gene, the risk at the same age is 45% [7, 16].

Several other genes playing a role in the DNA repair process are also associated with hereditary breast cancers: *BRIP1*, *RAD51*, *CHEK2*, *ATM*, *PALB2*, *TP53*, *STK11*, *CDH1* [7, 17].

Other risk factors, such as exposure to hormones (estrogen) and environmental factors (alcohol consumption, smoking, lack of exercise) have also been associated with breast cancer [3].

Several subtypes of breast cancer can be defined based on histopathology, molecular pathology, and transcriptomics.

Histologically, the vast majority of breast cancers can be divided into 3 subtypes: invasive ductal cancers (IDC or NOS for not otherwise specified), accounting for approximately 75% of cases; invasive lobular cancers (ILC), accounting for approximately 10% of cases; and rare subtypes (mixed IDC/ILC, medullary breast cancers, etc.).

The IDC and ILC subtype can be additionally subclassified through the use of **molecular pathology** into the following subtypes: hormone receptor positive tumors (ER-positive and PR-positive), based on the expression of the the estrogen receptor alpha and the progesterone receptor; HER2-amplified tumors, where the *HER2/neu* gene is amplified and over-expressed; and triple-negative breast cancer (TNBC), which express neither ER, PR, nor HER2 [3].

Alternatively, breast cancers can be classified into four different groups, based on the analysis of the **expression of genes**. These four subtypes (Luminal A, Luminal B, Basal-like, HER2/ERBB2+) overlap with the histopathological and molecular pathology subclassifications (see. Table 1.1) [18–20].

Multiple Myeloma

Multiple myeloma is a cancerous disease characterized by an unrestrained proliferation of plasmocytes (clonal plasma cells in the bone marrow). It is associated with high levels of serum paraprotein. All MM tumors have numeric and/or structural chromosome abnormalities [21].

Multiple myeloma is almost always predated by a pre-cancerous condition called MGUS (*monoclonal gammopathy of undetermined significance*) which is defined by conditions similar but less serious than multiple myeloma, namely a level of serum paraprotein lower than 30g/L and the presence of less than 10% plasma cells in the bone marrow. MGUS is present in 1% of adults older than 25 and the probability for an MGUS to evolve into multiple myeloma is approximately 1%/year [22].

Luminal A	ER+ HER2- PR+	Ki67 low (< 14%)
Luminal B	ER+ HER2-	Ki67 high (≥ 14%)
	ER+ HER2+	No Ki67 specificity
Basal-like	ER- PR- HER2-	No Ki67 specificity
HER2/ERBB2+	HER2+ ER- PR-	No Ki67 specificity

Tab. 1.1.: Gene expression based subtypes compared with molecular pathology based classification of breast cancer.

Most MM and MGUS tumors present (among other affected genes) a dysregulated and/or over-expression of *CCN D1*, *CCN D2*, or *CCN D3* which can be caused by different chromosomal abnormalities:

- Approximately half of multiple myeloma patients share a hyperdiploid karyotype (49 or more chromosomes, with gains of several of chromosomes 3, 5, 7, 9, 11, 15, 19, 21). Patients with a hyperdiploidy have a better prognosis, but said prognosis aggravates greatly if they acquire a specific chromosomal alteration (loss of chromosome 13, gain in the long arm of chromosome 1). *CCN D1* is dysregulated in a majority of hyperdiploid tumors.

- The other half of patients is split into hypodyploid, pseudodiploid, near-diploid or tetraploid karyotypes. Hypodiploid karyotypes are associated with a shortened survival and lower likelihood of response to therapy [21, 23, 24].

In addition to chromosomal gains and losses, multiple myeloma presents frequent aberrations such as large translocations, copy number alterations and point mutations.

The (complete or partial) monosomy of chromosome 13, which is the most common alteration in multiple myeloma, occurs in approximately 60% of tumors, 72% of non-hyperdiploid tumors and 37% of hyper-diploid multiple myeloma. It involves the deletion of the *RB* tumor-suppressor gene.

The immunoglobulin heavy-chain (IgH) translocation, which is present in approximately 40% of patients (most non-hyperdiploid tumors), is a translocation involving the IgH, located on chromosome 14, and different oncogenes which act as chromosomal partners (see Tab. 1.2).

Partner Oncogenes	Location	Prevalence	Prognosis
MMSET, FGFR3	4p16	15%	Unfavorable
CCN D3	6p21	3%	Favorable/neutral
CCN D1	11q13	15%	Favorable/neutral
c-MAF	16q23	5%	Unfavorable
MAFB	20q11	2%	Unfavorable

Tab. 1.2.: Partner oncogenes of the IgH translocation. (Bergsagel & Kuehl [21])

The 1p monosomy, which is present in approximately 30% of patients, can span variable lengths, but it usually includes cytobands 1p32 and 1p21, where *P53*-interacting genes *CDKN2C* and *CDC14A* are respectively located [25].

One should also mention the 17p13 deletion, associated with poor prognosis, which is present in approximately 10% of MM patients and also in chronic lymphoid leukemia and chronic myeloid leukemia, and the partial or complete 1q amplification which is also associated with an unfavorable prognosis [26, 27].

Hyperdiploidy and the IgH translocation are usually already present at the early stage of the disease, while some other recurrent chromosomal alterations appear only later during the disease evolution (13p and 1p monosomy, 1q, 16q, Xq duplications) [21].

Tumor suppressor genes *TP53*, *FAM46C*, *UTX*, *BIRC2*, *BIRC3* can be affected by deletions, while oncogenes such as *MYC*, *HGF*, *MCL1*, *IL6R* can be affected by amplifications.

The following 10 genes are frequently affected by point mutations: *NRAS*, *KRAS*, *TP53*, *CCND1*, *FAM46C*, *DIS3*, *PNRC1*, *ALOX12B*, *HLA-A*, *MAGED1* [28].

Screening and diagnosis

Breast cancer screening is usually done through mammography. European recommendations advise to carry out a screening mammography every 2 years for women between 50 ans 69 years old [29].

The sensitivity of screening mammography decreases significantly with increasing breast density and in younger women with dense breasts [30].

A large retrospective study encompassing more than 2.5 million screening mammographies, between 1996 and 2002, showed a mean positive predictive value of 4.8% for this test [31].

Moreover, mammography performance is operator-dependent and it exposes patients to ionizing radiations, which can be an additional risk factor for breast cancer. Moreover, since the advent of screening mammography, the detection rate of large tumors fell while it rose for smaller tumors. However, women were more likely to have breast cancer that was overdiagnosed than to have earlier detection of a tumor that was destined to become large. [32–34].

The most common symptoms of **multiple myeloma** are bone problems (pains, fractures), low blood cell counts, high blood levels of calcium, renal insufficiency [13, 35, 36].

When these symptoms are present, a diagnosis can be performed thanks to a bone marrow biopsy, complete blood cell count, serum protein electrophoresis, immunofixation, quantitation of immunoglobulins and measurement of free light chains [37].

Copy Number Variations

Copy number variations (CNVs) constitute variations of the human genome involving gains or losses of at least 50 basepairs to hundreds of kilobases of genomic DNA [38, 39]. Duplications of whole chromosomes were first described in the late 1950s thanks to karyotypes performed on patients suffering from Down syndrome and Klinefelter syndrome [40, 41].

CNVs can be present in phenotypically normal individuals, and large duplications or deletions can encompass genes without involving early onset, highly penetrant genomic disorders. On the other hand, the presence of a CNV can be associated with severe effects such as embryonic lethality. In the same way as single nucleotide polymorphisms, some CNVs can be population-private [39, 42].

A recent meta-analysis combining lists of benign CNVs present in various populations from 55 different studies has shown that 4.8–9.5% of the normal human genome can contribute to CNVs although, on average, the normal human genome has gains spanning only 0.35% of the genome and losses covering less than 0.1%. Moreover, benign CNVs are as rare as they are diverse, as a specific genomic region is affected by a CNV in only 2.6-4.3% of all individuals in the meta-analysis. However, individuals from the same ethnicity tend to share a higher number of CNVs, as these are often population-private. A summary of this human CNV map is presented in Table 1.3 [39, 43].

CNVs are unevenly distributed in the genome and along the chromosomes, as the pericentromeric and subtelomeric regions have a higher rate of CNVs. Moreover, large parts of chromosomes seem to never harbor benign CNVs, as the percentage of a chromosome that is prone to a CNV ranges from 1.1% to 16.4% for gains and from 4.3% to 19.2% for losses. Chromosomes 22, Y, 16, 9, 15 have the highest rate of gains while chromosomes 3 and 18 have the lowest. Chromosomes 19, 22 and Y have the highest rate of losses while chromosomes 5, 8 and 18 have the lowest [39].

Duplications or deletions of parts of chromosomes have been known to play a key role or to be directly responsible for several pathologies such as spinal muscular atrophy, Charcot-Marie tooth disease, DiGeorge syndrome; although one should remain cautious, as some of the CNVs associated with several genetic disorders show incomplete penetrance [42].

	Inclusive count	Stringent count
Parts of the genome susceptible to CNVs (%)	9.5	4.8
Parts of the genome susceptible to CNVs (Mb)	273	136.6
Median CNV length (bp)	981	1237
Mean CNV length (bp)	11362	11647
Number of identified CNV regions	24032	11732
Number of identified gain regions	3132	1169
Number of identified loss regions	23438	11530
Mean prevalence of a CNV (% population)	4.3	2.6

Tab. 1.3.: Summary of the copy number variation map of the human genome, based on a meta-analysis of 55 studies encompassing 2647 individuals. The inclusive threshold counts CNVs present in at least two subjects and one study for each variant. The stringent threshold counts CNVs present in at least two subjects and two studies. Some of the CNVs are counted both in the gains and in the losses as the same genomic region can show both patterns in different samples. (Zarrei *et al.* [39])

Role in cancer

In 2010, a large study using SNP arrays on more than 3000 cancer samples coming from 26 different cancer types showed that, on average, each cancer sample harbored 24 somatic gains and 18 losses. Regions of segmental duplication showed an enriched rate of CNVs. On average, gains spanned 17% of the cancer genome and losses covered 16%. These values differ strinkingly from the average 0.35% and 0.1% of the normal genome contributions to gains and losses, highlighting the key role of CNVs in cancer [43].

Interestingly, most somatic CNVs are either very small, or span exactly one chromosome arm or one whole chrosome (see. Fig 1.4), resulting in approximately 25% of the cancer genome affected by arm-level CNVs and 10% by very small ones (with 2% belonging to both categories). The most frequent very small somatic CNVs involve the amplification of the *MYC* gene and the deletion of *CDKN2A/B*, both present in approximately 14% of all cancer samples. Arm-level somatic CNVs are present in 15-29% of all cancer samples, depending on the affected chromosome.

Germline CNVs can also play a role in cancer, as some of the major cancer genes can be included in rare, non-polymorphic germline CNVs. Although less studied than

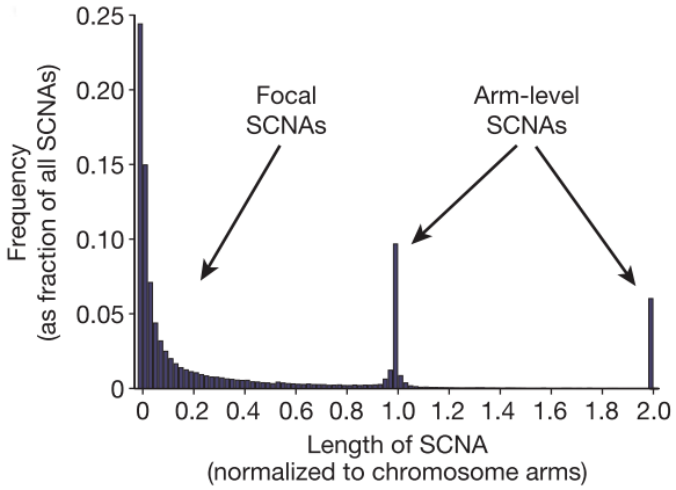


Fig. 1.4.: Distribution of somatic CNV lengths across 3131 cancer samples. The authors use the SCNA (somatic copy number alteration) notation. (Beroukhim *et al.* [43])

somatic CNVs, a few recent publications have listed germline CNVs as associated with susceptibilities for some forms of cancer [44, 45].

As breast cancer can be associated with the over-expression of several genes (see 1.1.4), it is logical to look for CNVs as potential causes of said over-expression.
An array CGH survey of 89 breast tumors has shown that several recurrent CNVs were present. The most frequent alterations are gains on 1q (35%), 8q (35%), 11q (26%), and 16p (14%), and losses on 4q (58%), 5q (54%), 6q (43%), 8p (48%), and 14q (48%) [46].
A more recent study on 773 breast tumors showed that the regions containing *PIK3CA*, *HER1/EGFR*, *FOXA1*, and *HER2/ERBB2* were often amplified and the regions containing *MLL3*, *PTEN*, *RB1* and *MAP2K4* were often deleted. Moreover, many of the CNVs outlined in this study, particularly some arm-length ones, correlated with the gene expression-based breast cancer classification; as the gain of 10p and the loss of 5q in Basal-like breast cancers, and the gain of 1q and/or the loss of 16q in Luminal breast cancers [47].

In multiple myeloma, other than the IgH translocation, several CNVs are also present:

- the monosomy of chr 13 or of 13q is the most frequent CNV in multiple myeloma and it affects approximately 45-50% of patients
- the 17p or 17p13 loss, which is also present in chronic lymphoid leukemia and chronic myeloid leukemia patients, is common in approximately 8-10% of MM patients
- the 1p loss is present in approximately 30% of patients

- the 1q gain is present in approximately 33-35% of patients
- the 5q gain is present in approximately 50% of patients
- the 12p loss is is present in approximately 10% of patients [26, 48–50]

Detection

Over the years, the detection of CNVs has been done with various methods, however no single technique was able to accurately detect all CNV classes, ranging from very small to very large, and including CNVs located in hard to probe regions such as segmental duplications. Table 1.4 shows the differences between existing methods.

Fluorescence in-situ hybridization (FISH) is a visual technique used to detect chromosomal alterations using fluorescent probes. Even if it can only observe a single locus at a time, it is often the only available method to detect certain forms of structural variations along the genome, and it is thus still used routinely.

Several techniques allowing to detect CNVs at a single locus or at a small number of loci, with varying throughput and resolution, should also be mentionned (Southern Blot, PFGE, qPCR, MAPH, MLPA, PRT) [39, 51].

	Fiber FISH	**Southern Blot**	**PFGE**	**QPCR**	**MAPH**	**MLPA**	**PRT**	**SNP array**	**Array CGH**	**NGS**
Sample	Cells	2–5 μg DNA	2–5 μg DNA	5–10 ng DNA	0.5–1 μg DNA	100–200 ng DNA	10–20 ng DNA	0.5–1 μg DNA	0.5–1 μg DNA	1–2 μg DNA
Loci	Single	Single	Single	Single	>40	>40	Single	>2 million	>2 million	Genome-wide
Throughput	Low	Low	Low	High	High	High	High	High	High	Moderate
Minimum resolution	>1 kb	>1 kb	0.5–1 kb	100 bp	100 bp	100 bp	100 bp	5–10 kb	5–10 kb	>1 kb
Cost per sample	Low	Low	Low	Low	Low	Low	Low	Moderate	Moderate	High
Time to result	>24 h	2–3 days	2 days	4 h	>24 h	>24 h	4 h	>24 h	>24 h	2–3 days
Labor requirement	High	High	High	Low	Low	Low	Low	Moderate	Moderate	High

Tab. 1.4.: Existing methods to detect CNVs (Cantsilieris *et al.* [51])

Array CGH, which was introduced in 1997, is a microarray-based extension of comparative genomic hybridization (CGH). CGH makes use of the hybridization of a reference DNA in addition to the tested sample, and the fluorescence ratio between both samples is normalized and used to infer the copy-number of the test sample. Several different molecules can be used on the array (genomic DNA clones, cDNA, PCR products or oligonucleotides).

SNP chips or SNP microarrays were originally designed with SNP genotyping in mind, so they don't work by comparing hybridization signals, like CGH. Instead, the

hybridization intensities from a single sample are compared to a set of reference values to extract the copy-number information. Due to the fact that SNP arrays also yield genotype information, it gives an added value when assessing deletions: genotype alone cannot allow the detection of deletions in a single individual (as hemizygosity could be miscalled as homozygosity of the remaining allele). But if a parent-offspring trio is analyzed, regions of losses of heterozygocity can be discovered, yielding additional information regarding potentially deleted regions [51, 52].

Several NGS approaches exist to detect CNVs from whole genome samples:

- the most straighforward one uses the depth of coverage (DOC) or reads depth (the number of sequencing reads mapping at a specified location), with a lower than average DOC being indicative of a loss and a higher than average DOC being indicative of a gain. The DOC is usually computed for sliding windows of varying sizes. The main assumption behind this method is that the coverage should be uniform along the genome, however this is not always exactly the case because of sequencing biases like GC-content.

- paired-end based methods require the use of paired-end sequencing libraries. The idea is that two paired sequencing reads mapping to a copy-neutral region of the genome are separated from each other by a specific distance called the insert size. If paired reads map to the genome with an unexpected insert size between them, this could be indicative of a gain or a loss. This method can also detect translocations, when paired reads map to different chromosomes. The insert size limits the size of CNVs which can be detected with this approach.

- Ratio based methods use a process similar to aCGH, by computing the ratio of mapped read counts between a test sample and a reference sample along the genome (see Fig 1.5). The idea underlying this approach is that it is an evolution of the depth of coverage method, where the use of a reference sample prepared the same way as the test sample should mitigate the sequencing biases [53].

Most methodologies used to detect CNVs from whole genome data have been adapted or at least tested on exome data. The ratio of reads depths analysis seems to be the most robust, but it requires access to a matched reference sample sequenced with the exact same protocol as the test sample, to account for the variability in capture efficiency accross exons [55].

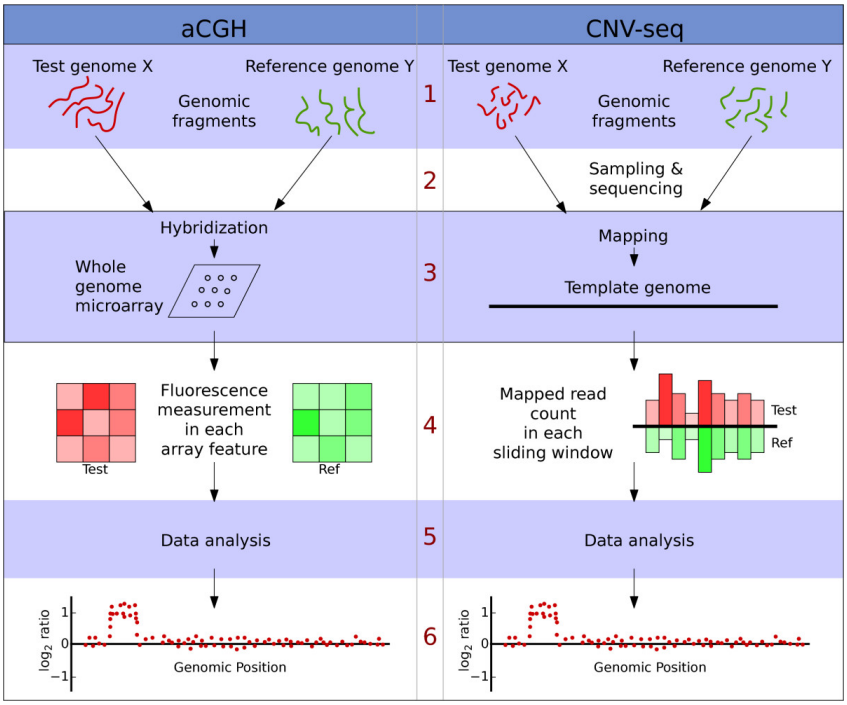


Fig. 1.5.: The read counts ratio approach to CNV detection with whole genome data. (Xie & Tammi [54])

The size distribution of detected CNVs varies, based on the detection method. Whole exome and whole genome sequencing yield smaller variants than array-based methods and are biased towards the detection of deletions. CGH and SNP arrays have a more limited resolution capacity [39].

Comparing CNV profiles

Since different techniques have been developed to detect CNVs, the assessment of said techniques accuracy is an important part of lots of studies. The comparison of CNV profiles is thus critical, since a comparison with an established, well characterized profile obtained with an older technique is often used to validate a new CNV detection method [56].
Moreover, the comparison of profiles can also be useful to indicate potential biases towards specific CNVs (small vs. large ones, gains vs. losses, GC-rich vs. GC-poor, etc.)

Some studies limit themselves to counting the overlap (i.e. the copy-number events in common) between the reference profile and the assessed profile [56].
Such overlaps can have different definitions, as different studies use different percentage thresholds to consider a CNV event to be present in two profiles [57].

Usually, studies only report the rates derived from the confusion matrix (TPR, FPR, TNR, FNR, see. 1.4.2), and in some cases the amplification/deletion ratios and the CNV size distribution.

One of the most comprehensive comparison study computed TPR and FPR values for different CNV lengths and different coverages and copy-numbers. It also computed the *F-score* defined as $F = 2\frac{PR}{P+R}$, where *P* is the *precision* or *positive predictive value* (i.e. $P = \frac{TP}{TP+FP}$, representing the percent of the detected CNV which overlaps with the reference profile) and *R* is the *recall* or *sensitivity* (i.e. $R = \frac{TP}{TP+FN}$, representing the percent of the reference profile CNV that overlaps with the detected CNV) [58].

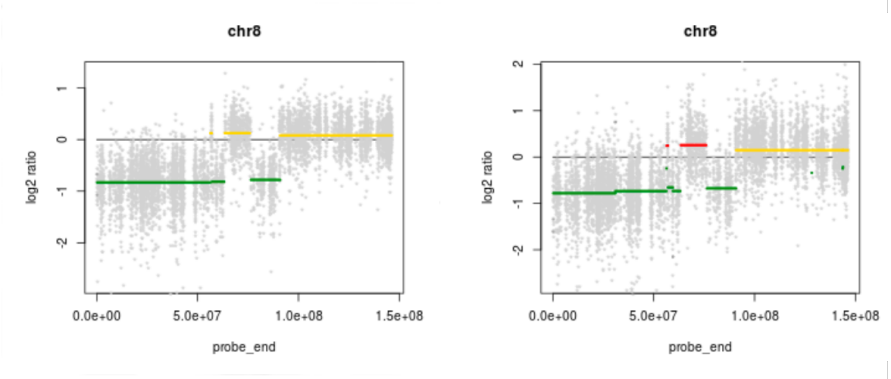


Fig. 1.6.: Two CNV profiles of the same chromosome from the same biological sample, analyzed with two different references. A slight change in *log-ratio* can have an effect on the presence (in red) or absence (in yellow) of a CNV at a specified locus. (Wenric *et al.* [59])

However, all these comparisons making use of confusion matrix derived metrics have to consider CNV events as binary variables (or as successions of binary variables along sliding windows), i.e. CNVs are either present, or absent, based on the *log-ratio* value at a specified locus and the *log-ratio* threshold used to call a CNV. Given said assumptions, slight differences of *log-ratios* between 2 samples can have different consequences if these differences happen below, around, or above the predefined threshold (see Fig. 1.6).

microRNAs

MicroRNAs were first discovered in 1993, when a team led by Ambros discovered that a gene deemed essential for C. elegans (*lin-4*) did not code for a protein but produced short transcripts (later named miRNAs) whose sequences were complementary to the mRNA sequence of another gene (*lin-14*), suggesting a regulation through antisense RNA-RNA interaction [60].

The first human miRNA (*let-7*) was identified in 2001 [61].

miRNAs can be defined as short single-stranded RNA molecules of 20-23 nucleotides which do not code for proteins and are only expressed in eukaryote cells. Their biogenesis is shown in Fig. 1.7. They play a key role in several biological processes where they regulate gene expression at the post-transcriptional stage, by binding to the mRNA, through complementary sequences, and preventing its translation into a protein. miRNA-binding sites are generally positioned in the 3' untranslated region (UTR) of mRNAs. Due to the short binding sequence, one miRNA can target multiple genes, and one gene can be targeted by multiple miRNAs. The biological processess affected by miRNA regulation range from proliferation, differentiation, apoptosis, cell cycle regulation to cell death [62, 63].

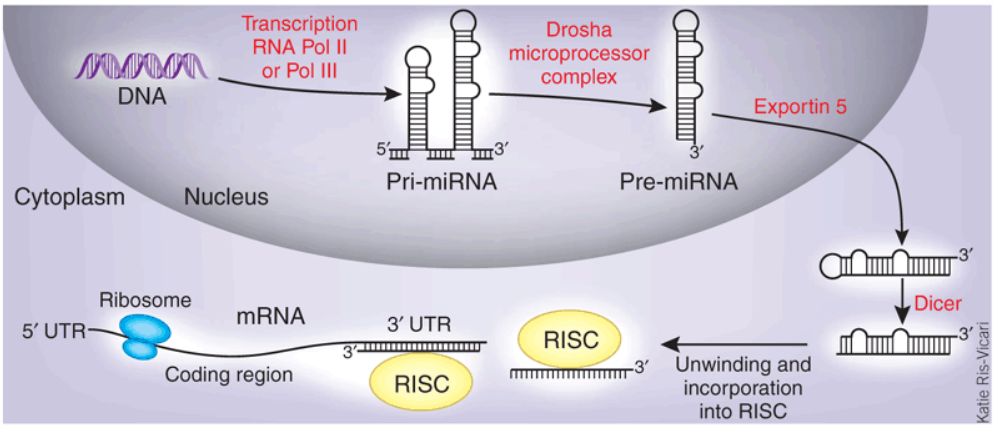


Fig. 1.7.: The key steps of miRNA biogenesis involve several genes and proteins (Drosha, Dicer, AGO1). (Jeffrey [64])

As of today, more than 2500 human miRNAs have been identified. More than 60% of all human protein-coding genes harbor one or several conserved miRNA-binding sites, and, since several non-conserved sites are also present, it is safe to say that a large majority of human protein-coding genes may be regulated by miRNAs [63, 65].

Several miRNAs share parts of their sequence (notably in binding regions) and are considered to be part of the same family. It is common that miRNAs of a family target the same mRNAs and share functional consequences. Studies have shown that, as miRNAs from a same family share a functional role, low levels of one miRNA could be balanced by higher levels of another miRNA from the same family [66].

miRNAs from the same family are indicated by lettered suffixes (e.g. *mir-34a* and *mir-34b*). Each genomic locus produces two mature miRNAs, coming from both strands of the precursor; they are indicated by an additionnal sufix (e.g. *mir-125a-3p* and *mir-125a-5p*). Usually, most of the total expression comes from only one of the two mature miRNAs (96-99% on average). As different miRNAs can be produced from close genomic loci, they are sometimes transcribed together, in the form of clusters [63].

microRNAs in cancer

Altered levels of miRNAs in cancer were first reported in a 2002 study on chronic lymphocytic leukemia, quickly followed by other tumor types. Globally, miRNAs are less expressed in tumors, compared to normal tissue [67, 68].

Different processes can explain the varying levels of miRNAs in cancer:

- miRNA genes are often located in regions of chromosomal instability (gains, losses, translocations). The genes for the *mir-15a/16-1* cluster are often deleted in chronic B-cell lymphocytic leukemia. The genes for the *mir-17-92* cluster are often amplified in lymphoma and translocated in T-cell acute lymphoblastic leukemia. A study on 227 samples of human breast cancers, ovarian cancers, and melanoma showed that a high proportion of genomic loci encompassing miRNA genes were included in copy number variation regions [69].

- Some miRNAs see their expression regulated by transcription factors from tumor-suppressor or oncogenes pathways. *mir-34a* is regulated by *TP53*, *mir-21* is regulated by *RAS*, and the *mir-17-92* cluster is regulated by *MYC*.

- miRNAs are also subject to phenomenons of epigenetic modulation such as hyper- or hypomethylation: different studies have shown that *mir-223* is silenced through CpG methylation or that other miRNAs can have elevated expression levels because of DNA demethylation and histone deacetylase inhibition. In several tumor types, levels of *mir-34a/b/c* can be lowered because of CpG methylation [70, 71].

- Several parts of the miRNA biogenesis pathway can also be altered in tumors through altered expression or mutations. For example, a study has shown that in lung cancer patients, the levels of *Dicer* mRNA and *let-7* miRNA were correlated and that low levels of both were associated with reduced survival [72].

Since miRNAs regulate gene expression, their dysregulation can have an effect on tumorigenesis, in the same way as tumor-suppressor genes or oncogenes (see Fig. 1.8) [66].

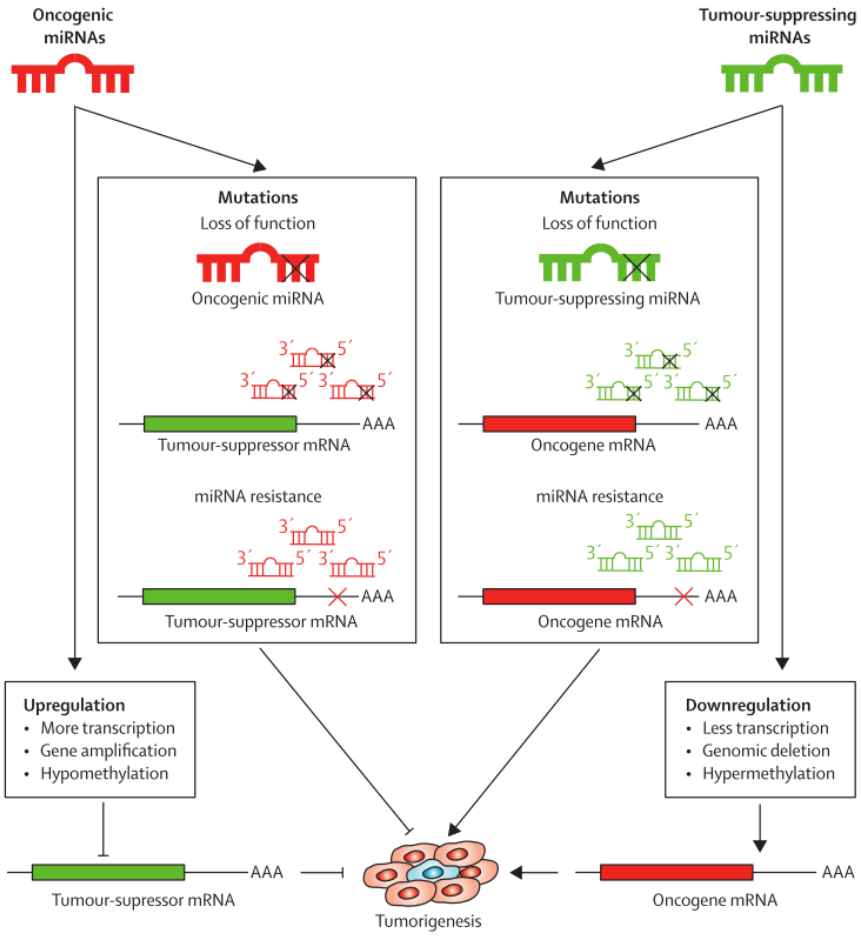


Fig. 1.8.: Regulation of tumorigenesis by miRNAs. An upregulation of oncogenic miRNAs can down-regulate the expression of tumor-suppressor genes, while a downregulation of tumor-suppressor miRNAs can up-regulate the expression of oncogenes. Moreover, mutations can also affect the regulating process in which miRNAs are involved. (Kong *et al.* [73])

Lots of studies have shown patterns of miRNAs dysregulation in several cancers, although one should remain cautious, as an up- or down-regulation of a certain miRNA is not necessarily indicative of a causative role in tumorigenesis. Table 1.5 shows some of the most commonly altered miRNAs in human cancers.

miRNA	Dysregulation	Cancer type
let-7/98 cluster	down	CLL, lymphoma, gastric, lung, prostate, breast, ovarian, colon, leiomyoma, melanoma
mir-15a/16-1 cluster	down	CLL, lymphoma, multiple myeloma, pituitary adenoma, prostate, pancreatic
mir-17-92 cluster	up	Lymphoma, multiple myeloma, lung, colon, medulloblastoma, breast, prostate
mir-21	up	Lymphoma, breast, lung, prostate, gastric, cervical, head and neck, colorectal, glioblastoma
mir-26a	down	Lymphoma, hepatocellular carcinoma, thyroid carcinoma
	up	Glioblastoma
mir-34a/b/c	down	CLL , lymphoma, pancreatic, colon, neuroblastoma, glioblastoma, breast
mir-155	up	Lymphoma (Burkitt's, Hodgkin's, non-Hodgkin's), CLL, breast, lung, colon, pancreatic
mir-141/200a cluster	down	Breast, renal clear cell carcinoma, gastric, bladder
	up/down	Ovarian
mir-205	down	Prostate, bladder, breast, esophageal
	up	Ovarian
mir-206	down	Rhabdomyosarcoma, breast
mir-9	down	Medulloblastoma, ovarian
	up/down	Breast

Tab. 1.5.: Commonly altered miRNAs in human cancer. (Farazi *et al.* [66] and Peurala *et al.* [74])

Other studies have shown that miRNAs could be associated with metastasis (miR-10b, miR-9, miR-31 and miR-335 in breast cancer) or with tumor aggressiveness (miR-210 in breast cancer) [75, 76].

As mentioned earlier, a dysregulation of miRNA levels in tumor samples does not necessarily indicate an active, functional role of said miRNA in the tumorigenic process. However, several studies have indicated the biological role of some miRNAs in breast cancer:

- *mir-21*, which is over-expressed in most tumors (cf. Table 1.5), has several oncogenes and tumor-suppressor genes as putative targets (*PTEN*, *SKI*, *RAB6A*, *RAB6C*, *RHOB*, *TGFBR2*, *RASA1*, *BCL2*, *PDCD4*), and it has been found to be antiapoptotic and to favor cell and tumor growth [73, 77–79].

- *mir-155*, which is over-expressed in several cancer types (cf. Table 1.5) promotes angiogenesis by targeting the *VHL* tumor-suppressor. Moreover, *BRCA1*, which is a tumor-suppressor gene involved in breast cancer (cf. 1.1.2), epigenetically represses miR-155 [80, 81].

- The *mir-17-92* cluster is involved in a feedback system with the E2F proteins (a family of transcription factors which are critical regulators of cell proliferation). An overexpression of this miR cluster, which happens in several cancer types (cf. Table 1.5), disrupts the feedback loop to promote cell proliferation. Moreover, *mir-19* inhibits *PTEN*, leading to the activation of the *AKT* signalling pathway and promoting cancer-cell survival [70, 73].

- *mir-10b*, which is over-expressed in breast cancer stem cells, targets *PTEN* which is an important regulator of the *PI3K/AKT* pathway involved in metastasis, cell survival, and self-renewal. Moreover, an over-expression of *mir-10b* in metastatic breast cancer cells up-regulates *c-Jun* (a transcription factor playing a key role in stimulation of cell proliferation and tumor progression) via the down-regulation of *HOXD10* and *NF1* (two proteins implicated in cytoskeletal dynamics) [82, 83].

- the *mir-15/16* cluster: the over-expression of *mir-16* inhibits progestin-induced breast tumor growth [84].

- the *let-7* family targets several genes such as *KRAS*, *HRAS*, *HMGA2*, which are implicated in cell growth, differentiation, and survival. *let-7b* inhibits breast cancer cell motility through a down-regulation of several genes of the actin cytoskeleton pathway [73, 85].

- *mir-34a*, which is part of the *mir-34* family, has its expression regulated by *TP53* and is down-regulated in tumors (cf. supra). This microRNA inhibits osteoclastogenesis and bone resorption through a down-regulation of *Tgif2*. It inhibits bone metastasis formation in mouse models of breast cancer [71, 86].

- *mir-125b*, which is under-expressed in metastatic breast cancer, targets the *HER2/neu* oncogene [87].

microRNAs as diagnostic biomarkers

Circulating RNA molecules were already observed in 1999, in the plasma of cancer patients [88].

In 2008, two different studies showed that it was possible to detect circulating miRNAs: placental miRNA in maternal plasma and tumoral miRNAs in the serum of patients suffering of diffuse large B-cell lymphoma. Since the profile of serum miRNAs was different between patients and controls, this discovery paved the way to the use of circulating miRNAs as biomarkers. Later studies showed that miRNAs could also be detected in other fluids (saliva, urine, breast milk, tears, amniotic fluid, CSF, pleural fluid) [89–91].

Circulating miRNAs, which are released in the peripheral bloodstream by most cells, both in normal and pathological conditions, remain stable despite RNase because they are protected from degradation by different structures:

- Argonaute-family (AGO) proteins
- Exosomes: extracellular vesicles of 40-100 nm formed through an invagination of the plasma membrane
- Other microvesicles larger than exosomes (100-1000 nm)
- High-density lipoproteins (HDL)

90-95% of circulating miRNAs are associated with AGO proteins [91].

Several methods can be used to determine profiles of miRNAs: RT-qPCR, microarrays, and sequencing. Microarrays yield fold-changes of miRNA levels between samples, while sequencing methods derive the levels of miRNA from the sequencing read counts [66].

	MicroRNA RT-qPCR	**MicroRNA microarray**	**Small RNA-seq**
Principle	PCR amplification	Hybridization	Sequencing
Throughput	Medium to high	High	Ultra high
Costs	Economic	Economic	Comparatively high
Required amount of RNA	10 ng–700 ng	100 ng–10,000 ng	250 ng–10,000 ng
Data generation	1 day	Up to more than 2 days	Up to more than 1 week
Data information	Assumption based; dependent on the number and nature of targeted transcripts	Assumption based; dependent on the number and nature of targeted transcripts	Assumption free, de novo identification of transcripts within the small RNA transcriptome
Data analysis duration	Short	Moderate	Long
Preferential field of application	Relative and absolute quantification	Relative and absolute quantification of miRNA regulation	De novo identification of small RNAs, simultaneous relative quantification of different small RNA species
Common normalization strategies	Invariant-based (e.g., stable reference small non-coding RNAs)	Quantile	Scaling to library or sub-library (e.g., miRNA) size
	Plate normalizing factor	LOESS	Quantile
	Global mean expression	Variance stabilization	Trimmed
		Invariant-based	
		Scaling (e.g., Z-score, mean, median, 75th percentile)	
		Personalized logistic regression model	

Tab. 1.6.: Commonly used methods for the quantification of miRNAs. (Meyer *et al.* [92])

In the RT-qPCR setting, normalization is necessary to account for reverse transcription and PCR reaction efficiencies. Normalization can be based on predefined invariant controls, reference miRNAs, or other RNA molecules. The selection of reference miRNAs can be performed through different methods (stepwise elemination of the least stable miRNA, pairwise correlation between miRNAs, linear mixed-effects models). A 2009 study showed that the mean expression value of all expressed miRNAs could also provide a stable normalization [92, 93].

Recent studies have shown that circulating miRNAs could be used as biomarkers in a wide variety of diseases such as Alzheimer's and other neurodegenerative diseases, cardiovascular diseases, diabetes, obesity, endometriosis, inflammatory diseases, and of course cancers (breast cancer, gastric cancer, colon cancer, hepatocellular carcinoma, prostate cancer, acute myeloid leukemia, neck squamous cell carcinoma, bladder cancer), but also when investigating non-pathological conditions such as the response to exercise and diet [94–101].

As soon as more than one biomarker is involved, which is very often the case with circulating miRNAs, appropriate techniques are needed to efficiently treat the information. These methods are briefly described in the next section of this chapter.

Starting in 2010, seveal studies have shown that breast cancer patients had a dysregulated circulating miRNAs profile and that it was possible to use this difference as the basis for diagnosis [102–109].

Supervised learning

Machine learning describes an ensemble of methods whose goal is to learn a model from data, i.e. to make accurate predictions based on past observations. The main goals of machine learning methods are both the possibility to predict an output for new data and to gain a better understanding of the role of the different variables.

Supervised learning refers to the specific machine learning tasks where the output related to a set of labeled data is already known and used to predict the output for unknown data. The problem is called a **regression** problem if the output of interest is quantitative (i.e. a number), and a **classification** problem if the output is qualitative (i.e. a category, like *case* or *control*). Fig. 1.9 shows an example of a supervised classification problem. Supervised learning has been applied to an extended range of topics ranging from insurance fraud detection to product recommendation and medical diagnosis [110].

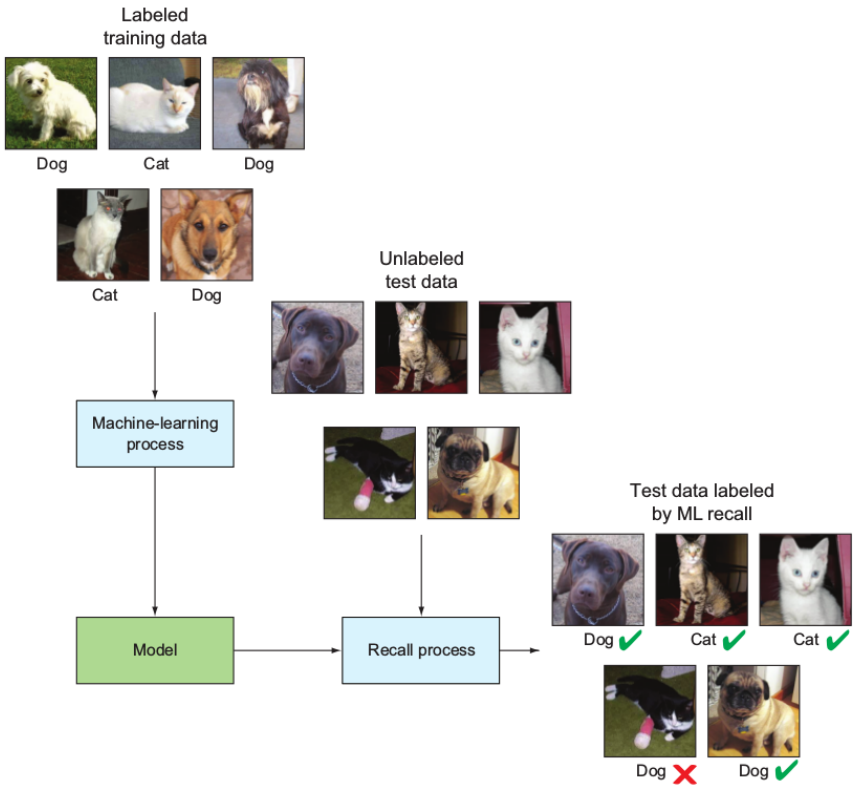


Fig. 1.9.: Example of a supervised classification problem. The variables used and the classification algorithm are not shown. (Brink *et al.* [110])

Many different algorithms have been developed to tackle classification problems. One can cite decision trees, k-nearest neighbors, naive bayes classifiers, support vector machines, neural networks, deep learning. One should also note that regression

algorithms can also be used in classification problems, by the use of threshold values linking the numerical model output and the categories [111].

Random Forests

Decision trees are one of the most straightforward supervised classification algorithm. A decision tree is made of nodes and branches, where nodes test one variable of the dataset, and branches correspond to a specific value or range of values for the variable. Each leaf node is labeled with a predicted class. The advantages of decision trees are the very high interpretability of the model (one can directly see which variables are used) and its non-parametric nature. Fig. 1.10 shows a simple decision tree [112].
A non-parametric model does not make strong assumptions about the model or the number of variables used. It is more flexible than parametric models, but it is more prone to overfitting [113].

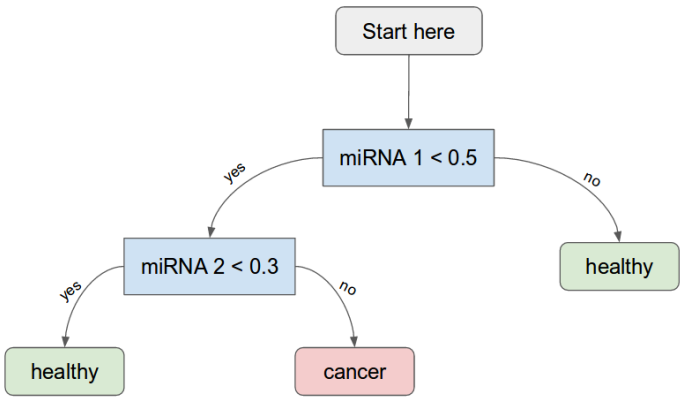


Fig. 1.10.: Example of a simple decision tree. We start at the root of the tree to classify a sample of unknown output.

During the learning phase (i.e. building of the model), the tree is grown in a top-down manner, starting with the most informative variable (i.e. the one which splits the learning set into subsets having the most similar outputs). This process of selecting the most informative variable and splitting the set of samples at a node into 2 subsets is repeated recursively until all samples at a node have the same output (e.g. all *cases* or all *controls*) or until a specificied stopping criterion is met (e.g. a minimum number of samples at each leaf node).
When the stopping criterion is met, each leaf node is labeled with the majority class of the samples present in that node.

One of the pitfalls of decision trees is the high variance or the risk of overfitting. **Variance** is the sensitivity of the model to small fluctuations in the learning set.

Overfitting means that the model will take into account too much of the variability of the samples of the learning set and thus instead of only modeling the "real" relationship between the features of the learning set and the output variable, it also encompasses some of the random noise present in the learning set [114].

One of the families of **ensemble methods** allows to reduce the variance of decision trees by combining the predictions of several models. The variance of each model is thus averaged on a large number of models and the global variance is lower than the single variance of each model [111].

Different ensemble methods aiming to reduce variance have been developed: bagging, random forests, extremely randomized trees, etc.

Another group of ensemble methods have also been developed: boosting type algorithms, which are more suited to reduce bias [115].

Bagging (stands for bootstrap aggregating) starts with the initial learning set of size n, generates m bootstrap samples (i.e. m learning sets of of size n' obtained by sampling with replacement from the initial learning set), builds m models on each boostrap sample, and averages the prediction of each model. The resulting ensemble model has a smaller variance than the original decision tree, and the variance reduction effect is proportional to m [116].

Random forests can be defined as a combination of bagging and random variable selection: the method starts with the initial learning set of size n, generates m bootstrap samples and builds a tree on each bootstrap sample, but instead of building each of these trees the classical way (i.e. by choosing the most informative variable at each node), it is the most informative variable *from a random subset (without replacement) of all the variables* which is selected at each node. While the direct interpretations of the variables are lost since the global model is a "black box", it is still possible to extract a ranking of all the variables from this global model. Fig. 1.11 shows an example of random forests [117].

Different parameters influence the global model: the number of trees, the number of randomly selected variables at each node.
As the number of trees grows, the aggregated variance lowers but the computation time grows. There is thus a trade-off based on the available computational power. An empirical way of chosing an appropriate value for the number of trees to grow is to observe the stabilization of the variable rankings as the number of trees grows gradually (see Fig. 1.12) [62].

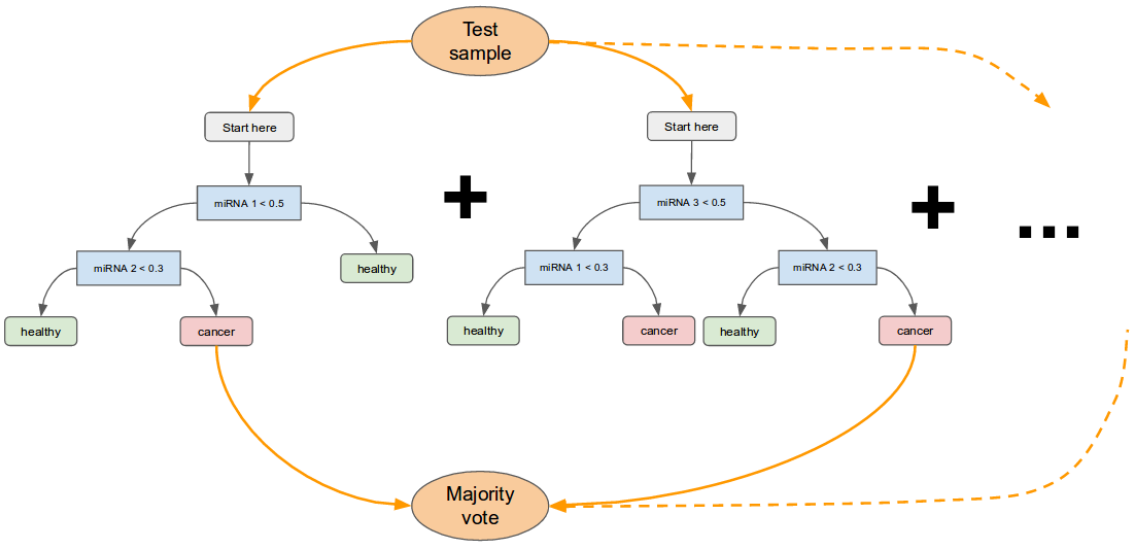


Fig. 1.11.: Random forests used in prediction mode.

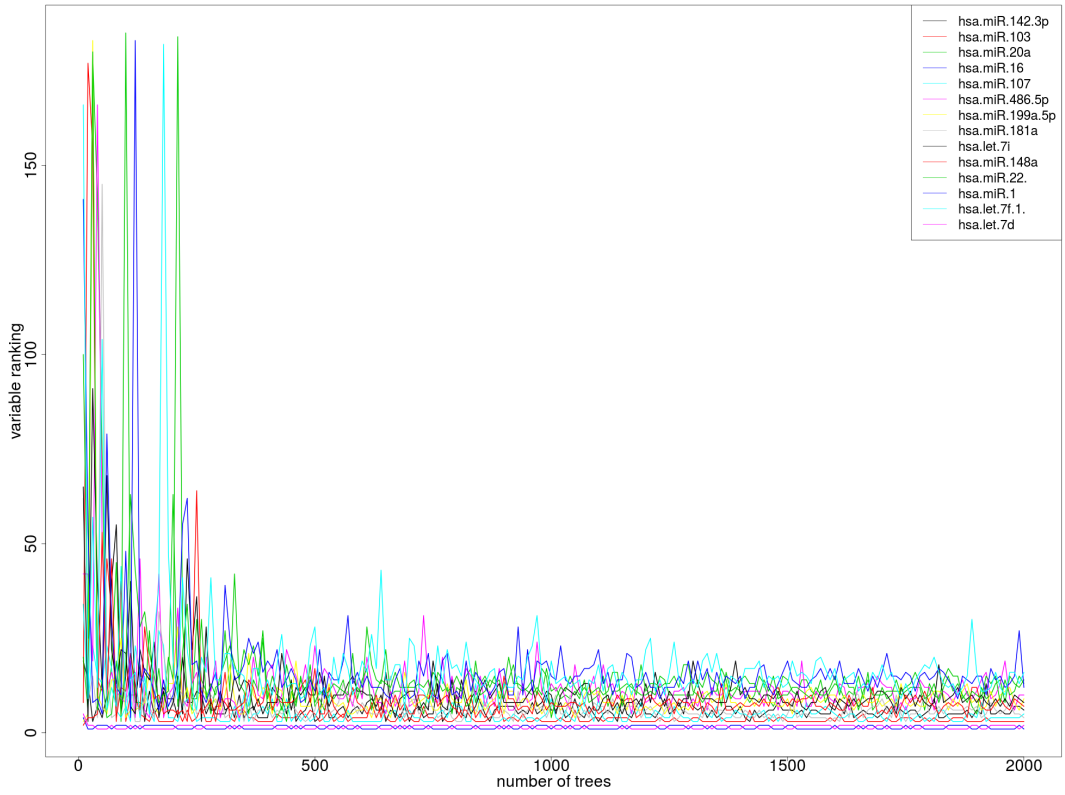


Fig. 1.12.: In this specific case, the variable ranking (based on the mean decrease in accuracy and the mean decrease in Gini) stabilizes after approximately 1000 trees. (Frères *et al.* [62])

For a dataset with d variables, a random subset of size $K = \sqrt{d}$ constitutes a good default number of randomly selected variables at each node, for classification problems [118, 119].

Extremely randomized trees or extra-trees differ from random forests in that there is no bootstrapping of the samples. The variable used at each node splitting is the most informative one among a random subset of all variables, and the cut-off value for this split is also selected randomly [119].

Evaluating predictions

As stated earlier, one of the goals of supervised learning is to make accurate predictions. The systematic assessment of said accuracy requires different techniques.

A straightforward and ideal **evaluation protocol** would involve the availability of a very large test sample (or test set), independent from the learning sample (or learning set). The independency between learning and test sample is important because it allows to evaluate the performance of models on unknown data. This is what comes the closest from "real-life" situations, where the model would be used to classify new samples.

Unfortunately, a large independent test sample is not always available, especially in studies involving human samples. A way to bypass this problem is to use **cross-validation**: it involves splitting the global dataset into k subsets (e.g. $k = 5$ or $k = 10$) of approximately the same size and the same partition of output values; for each of those k sub-samples, a model is built on the rest of the dataset and it is evaluated on the sub-sample; the performance measure is averaged on all k sub-samples, to give an approximation of the performance on the whole dataset.

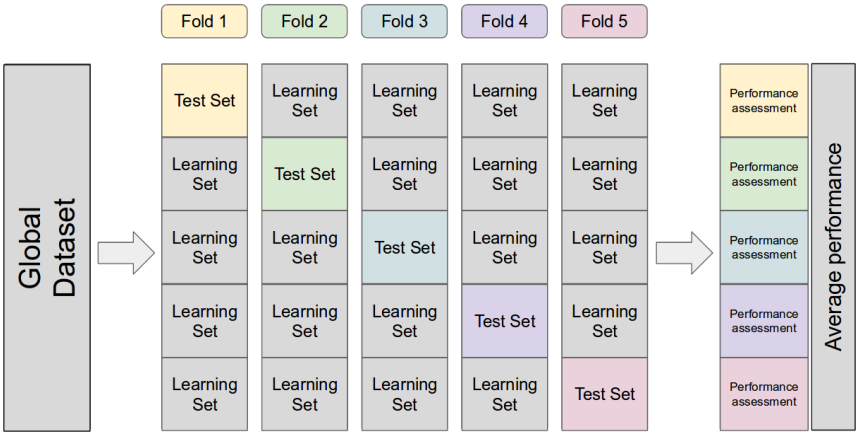


Fig. 1.13.: Illustration of 5-fold cross-validation.

Different **evaluation measures** exist to assess the accuracy and performance of a model on a test set.
For binary classification (e.g. *cancer* vs. *control*), the most obvious idea would be to count the number of correctly classified objects from the test set. But this simple measure does not take into account the potential imbalance towards one of the classes in the test set and the potential bias of the model (e.g. if the model classifies everything as cancer, the classification will be correct for all real cancer cases).

A natural extension of this global error rate is the **confusion matrix** (Table 1.7) :

		actual condition	
		actual positive (P)	actual negative (N)
predicted condition	predicted positive (predP)	true positive (TP)	false positive (FP)
	predicted negative (predN)	false negative (FN)	true negative (TN)

Tab. 1.7.: Confusion matrix.

From this matrix, different metrics can be directly computed, and one can easily see if there is a potential bias towards the misclassification of positives or negatives. These metrics are:

- the true positive rate (or sensitivity): $TPR = \frac{TP}{P}$
- the true negative rate (or specificity): $TNR = \frac{TN}{N}$
- the false positive rate: $FPR = \frac{FP}{N}$
- the false negative rate: $TNR = \frac{FN}{P}$

The confusion matrix requires that the model outputs a well-defined class.
But in binary classification problems, some algorithms are able to output a probability of belonging to one or the other class.
If two objects from the test sample have, say, a probability of 60% and 80% of belonging to the cancer class respectively, one would like to have access to a performance measure which takes into account this difference in terms of confidence of the predictions.

The **ROC curve** (receiver operating characteristic curve) allows this. As shown on Fig. 1.14, it is a graphical plot created by plotting the true positive rate (TPR) against the false positive rate (FPR). To be able to plot the TPR against the FPR, we need ranges of values for these 2 metrics. Such ranges can be obtained by varying the

decision thresholds required when the output of a classifier is a probability.
Let's get back to the previous example, with the two objects and their respective probability of belonging to the cancer class: if the threshold chosen is a probability higher than 55%, then both objects will be classified as cancer, if the threshold chosen is a probability higher than 70%, then only one of them will be classified as cancer. One can directly see how the variation of the threshold will generate a full range of different TPR and FPR values.
Extreme values of the threshold (where either all samples are classified as cancers or they are all classified as controls) will yield extreme values of TPR and FPR (0 and 1).

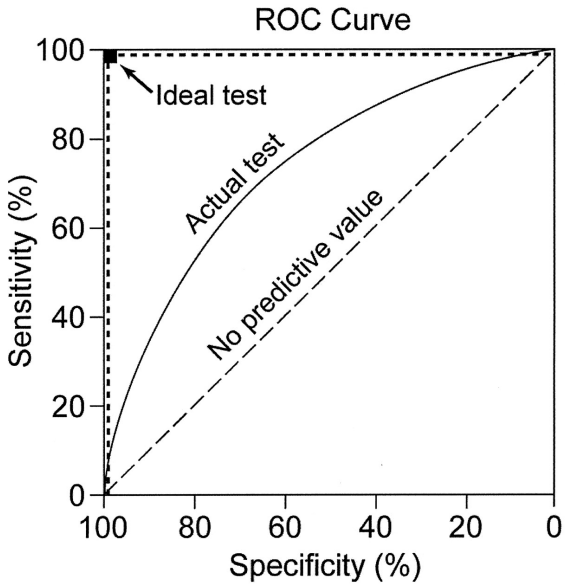


Fig. 1.14.: Receiver Operating Characteristic curve

To be able to compare different models, it is often useful to be able to summarize the ROC curve with a single score.
One can directly see that the **Area Under the ROC Curve (AUC)** constitutes an appropriate metric. The AUC will have a value of 1 for a perfect classifier, and a value of 0.5 for a completely random classifier. The AUC is independent of the threshold and the class distributions of the samples (= the incidence rate).

Random forests in the diagnostic setting

Already in his 2001 paper introducing the concept of random forests, Breiman used a breast cancer dataset (699 samples and 9 features computed from a digitized image of a fine needle aspirate of a breast mass), to test the binary classification performances of his algorithm and classify the samples as either malignant or benign

[117].
Since then, starting from 2003, several studies have been using this classification method in the diagnostic setting, with diverse feature types, ranging initially from proteomics mass spectrometry data to microarray data [120–128].

microRNAs and random forests

The use of miRNAs as features to classify samples with the random forests algorithm was first performed in 2010 on a melanoma dataset, to predict post-recurrence survival [129].
Later studies have shown the usefulness of miRNAs with random forests in glioma biology, colorectal and pancreatic cancers, invasive bladder tumors, prostate cancers, Alzheimer's disease [128, 130–133].

The feature selection challenge

Feature selection in classification or regression problems has become a critical step, given the high dimensionality of many studies making use of recent technologies such as microarrays, next-generation sequencing, or miRNA quantification.
Feature selection indeed constitutes a way of reducing the dimensionality of a classification problem while still retaining the original variables (contrary to principal components analysis) [122, 134, 135].

The motivations behind feature selection are twofold:

- Finding the variables which are the most relevant to the prediction.
- Reducing the number of input variables, to avoid over-fitting, to improve the performance of the model, and to be able to build more cost-effective models.

It should be noted that the additional feature selection step adds a layer of complexity in the search for the optimal solution: instead of looking for the optimal model built with all the initial features, we are now looking for the optimal model among all possible models built with all possible subsets of the initial features.
In the case of a highly dimensional initial classification problem, the *brute-force* exploration of the solution space can thus quickly become computationally intractable.

Different feature selection techniques have thus been developed, and they can be broadly divided into three categories:

- **Filter** methods: the features are ranked based on an univariate score (e.g. p-value of a statistical test performed independently of the classifier), and the k best features are selected. This method is very fast and scalable to highly dimensional datasets, but it does not take into account relationships between features and it might thus miss pairs or groups of features which, albeit seeming insignificant while being considered alone, provide a high amount of classifying information while being considered together.
- **Wrapper** methods build a model to rank each tested feature subset (the ranking is based on a performance measure through cross-validation). The choice of the next feature subset to test can be based on the performance of previously tested subsets. Wrapper methods can be prone to over-fitting and they are more computationally intensive than filter methods, since they have to build a model for each tested feature subset.
- **Embedded** methods perform the feature selection process while building the model, and are usually specific to a given algorithm.

Several feature selection methods have been designed specifically with random forests in mind:

- Iterative feature elimination, where features with the smallest importance metric are iteratively discarded until reaching a minimum out-of-bag (OOB) error [122].
- Sequential feature introduction, works the other way around, by iteratively adding features in candidate models, based on their importance metric computed on a previous complete model, and stopping the addition of features when the model accuracy reaches a maximum [136].

Transcriptomics

The **transcriptome** is the complete set of transcripts (mRNAs, non-coding RNAs, small RNAs) present in a cell. **Transcriptomics** usually refers to both the identification and the quantification of said transcripts. Comparing transcriptomes allows the identification of genes or transcripts that are differentially expressed in distinct tissues, or in response to different treatments [137].

Over time, several technologies have been developed to study transcriptomics.
Until the end of the 1990s, northern blotting was the most widely used technique to quantify RNA.
In 2000, a study introduced reverse transcription polymerase chain reaction (RT-PCR), and its use to quantify gene expression through the creation of complementary DNA (cDNA) transcripts from mRNA and the quantitative measuring of cDNA amplification using fluorescent probes. This study paved the way to the use of RT-PCR as one of the go-to methods for gene quantification, DNA microarrays constituting the other one, being more adapted to the quantification at the whole transcriptome scale. (see 1.2.2) [138, 139].

Microarray technologies, being relatively inexpensive and having a high throughput, rapidly allowed to study a wide range of biological questions including the identification of genes being differentially expressed between diseased and healthy tissues, new insights into developmental processes, pharmacogenomic processes, and the evolution of gene regulation in different species. However, microarrays suffer from several limitations: hybridization and cross-hybridization artifacts, the need for prior knowledge about studied transcripts, a limited dynamic range of detection, the need for complex normalization methods and/or reference sample because of discrepancies in probe hybridization properties [137, 140].

RNA-Seq refers to the use of next-generation sequencing technologies to sequence cDNA derived from RNA molecules, and infer the transcriptome composition.
The whole process can be summarized as follows: from an RNA sample, mRNA is either selected by its polyadenylation tail and fragmented in short reads of hundreds of bases, or ribosomial RNA (representing 25% of all RNA) is removed, to have an enrichment in mRNA. The mRNA is converted into cDNA by random priming and the use of reverse transcriptase. cDNA libraries are then prepared for sequencing. Sequence reads can then either be mapped on a reference genome, or used in de novo quantification pipelines.

In addition to preventing some of the limitations of microarrays, sequencing allows to attain single base resolution, allowing the detection of point mutations, and

sequencing yields gene or transcripts expression levels which are quantifiable rather than relative. Beyond expression levels, RNA-Seq also provides information about chrimeric or fusion transcripts, alternative splicing, allele-specific expression, RNA editing [141].

Standard RNA-Seq libraries do not preserve information about which strand was originally transcribed. The library preparation steps involving the synthesis of randomly primed double-stranded cDNA and the addition of adaptors lead to the loss of the strand information relative to the original mRNA template. Therefore, different methods have been developed to perform **strand-specific RNA-Seq**, the motivation being the capacity to accurately identify and detect antisense transcripts and to resolve the correct levels of expression between overlapping transcripts located on opposite strands [142].

Computational methods in RNA-Seq

Transcriptomics studies require the use of several computational methods related to the different steps of the global analysis pipeline. Among these steps, mapping the sequenced reads, quantifying gene or transcript expression, and analyzing patterns of differential expression constitute fundamental functions belonging to most if not all bioinformatics pipelines.

RNA splicing is the biological process by which the pre-mRNA transcript is edited, with the introns being removed and the exons joined together. In several cases, exons can be skipped and introns can be retained, creating diversity in the splicing process.

Reads mapping or "reads alignment" in the context of RNA-Seq differs from the DNA setting in that there is an added layer of complexity caused by the potential splicing together of non-contiguous exons to create mature transcripts. If DNA mapping tools also have to address challenges such as the presence of mismatches (point mutations), insertions, deletions, sequencing errors, the task of mapping sequencing reads to non-continuous genomic regions joined together from spliced RNA constitutes a specific pain point encountered by RNA-Seq mapping tools [143].

Spliced aligners, developed for this specific task, can be divided into 2 categories, based on the method used to map reads. These methods can be briefly described as follows:

- The "exon-first" method, used by TopHat among other software, is a two-step process where, at first, reads corresponding to a unique exon are mapped.

Second, the remaining unmapped reads are split into short fragments which are then aligned separately. The region surrounding the aligned fragments is searched for splice junctions [144].

- The "seed-extend" method, used by STAR and several other tools, proceeds by splitting all reads into short seeds first. Then, the seeds are iteratively mapped, starting by the first base of the read. If the next seed cannot be mapped contiguously to the previous one (due to a splice junction), another acceptable location (an acceptor splice site) is looked for in the vicinity of the genomic region [143].

Expression quantification is the process of estimating the genes or transcripts expression in samples, based on sequence read counts. A normalizing step is always required at first, to address two main sources of potential variability: RNA fragmentation during library construction which causes longer transcripts to generate more sequencing reads compared to shorter transcripts present at the same level and the inter-sample variability which yields different numbers of reads across samples.

The most common normalization method used for RNA-Seq data is called reads per kilobase of transcript per million mapped reads (RPKM): it normalizes the read counts for a transcript or a gene by both the length of said transcript or gene and by the total number of mapped reads in the sample. Other normalization methods model read counts as following a negative binomial distribution, with specific normalization factors to account for gene-based and sample-based differences [145, 146].

As genes can have multiple isoforms and multiple transcripts, one exon can be shared by different isoforms. The task of assigning a sequencing read to one isoform or the other can thus be complex. Different strategies exist: one can estimate isoform expression by counting only the unambigous reads mapping only to one isoform but this method does not work for genes with most exons shared among isoforms; another technique is to estimate the "most likely" expressed isoform corresponding to the sequencing reads.

Depending on the the goal of the analysis, a choice of counting sequencing reads by genes and not by transcripts or isoforms can also be made, simplifying the process. In this case, a count of all sequencing reads overlapping all the exons of a gene is performed. An important feature of some expression quantification tools is the dismissal of reads mapping ambiguously to multiple genes. This is justified to prevent discrepancies in the downstream analysis process, which is usually the search for differentially expressed genes [145, 147].

Differential expression analysis allows to systematically detect changes of gene or transcript expression across experimental conditions. Several characteristics of the read counts used as input, such as non-normality, a dependence of the variance on the mean, and a usually small number of samples, make the use of specific statistical models required [146].

Due to the often small number of samples, testing each gene separately makes it difficult to account for the uncertainty of the within-group variance (or dispersion). A way to work around this, thanks to the large number of genes assessed, is to make the assumption that the within-group variance of different genes is similar in the same experiment and/or that the variance is dependent on the average expression of the gene.

In most differential expression analysis software, the normalized read counts distributions is modeled by a negative binomial model, which allows the use of exact (non-asymptotic) testing for differential expression [146, 148, 149].

The number of replicates required in an RNA-Seq experiment varies based on the effect of the expected biological variation (i.e. the fold change) and the variability in measurement, which depends on the technical noise. Table 1.8 shows an example of calculations for the probability of detecting differential expression for a single gene at a significance level of 5%, when comparing two groups with a negative binomial model, as computed by the RNASeqPower package [150].

		effect size (fold change)			
		1.25	**1.5**	**1.75**	**2**
number of replicates per group	**3**	0.09	0.21	0.36	0.51
	5	0.13	0.32	0.54	0.72
	10	0.21	0.56	0.83	0.95
	15	0.29	0.73	0.95	0.99

Tab. 1.8.: Relationship between effect size, number of replicates, and statistical power. The within-group variance and the average read depth have been fixed at respective values of 0.4 and 40 for the sake of simplicity.

Long non-coding RNAs

The human genome contains a large number of nonprotein-coding sequences. In fact, up to 80% of the human genome is transcribed, thus encompassing a large number

of non-coding RNAs. Among them, long noncoding RNAs (lncRNAs), are defined as nonprotein-coding transcripts that are longer than 200 nucleotides [151–153].

lncRNAs can be defined by their location relative to nearby protein-coding genes:

- **Antisense** lncRNAs or natural antisense transcripts (NAT) are lncRNAs that initiate inside or at the 3' location of a protein-coding gene, which are transcribed in the opposite direction of protein-coding genes, and which overlap at least one coding exon.
- **Intronic** lncRNAs are lncRNAs that initiate inside of an intron end without overlapping any exon.
- **Bidirectional** lncRNAs are transcripts that initiate in a divergent fashion from the promoter of a protein-coding gene.
- **Intergenic** lncRNAs (or large intervening noncoding RNAs or lincRNAs) are lncRNAs with separate transcriptional units from protein-coding genes [154].

Antisense lncRNAs have been shown to regulate the expression of protein coding genes by affecting transcription and mRNA stability. Many expressed genomic loci produce RNAs from both the sense and antisense DNA strand, and more than 50% of all human RNAs share a partial or complete overlap with an opposite-strand transcript. However, antisense transcripts are generally poorly expressed and have, on average, expression levels varying between two to ten orders of magnitudes lower than their sense counterparts. A recent study noted, on average, opposite strand expression from more than 38% of annotated protein coding genes [155–158].

Antisense lncRNAs can exert an effect on the DNA strand from which they are produced (*cis* effect) or on different strands (*trans* effect). Given the fact that both the sense and antisense transcripts are transcribed from the same genomic region, it is expected that antisense transcripts behave more frequently in *cis* than other ncRNAs that commonly function in *trans* [157].

Antisense expression can affect gene expression through different mechanisms, which happen at different stages of the gene expression process. Table 1.9 shows examples of antisense lncRNAs effects on gene expression and Figures 1.15, 1.16, 1.17. detail 3 of those mechanisms.

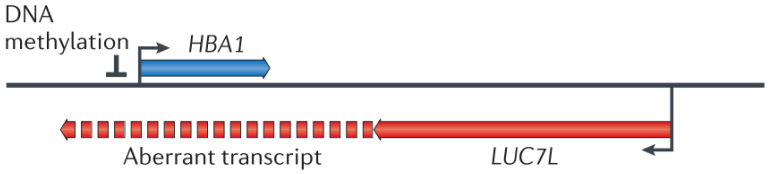


Fig. 1.15.: Abnormal transcriptional extension of the *LUC7L* locus creates an antisense transcript overlapping with *HBA1*, which methylates the *HBA1* promoter and inhibits its expression. (Pelechano & Steinmetz [157])

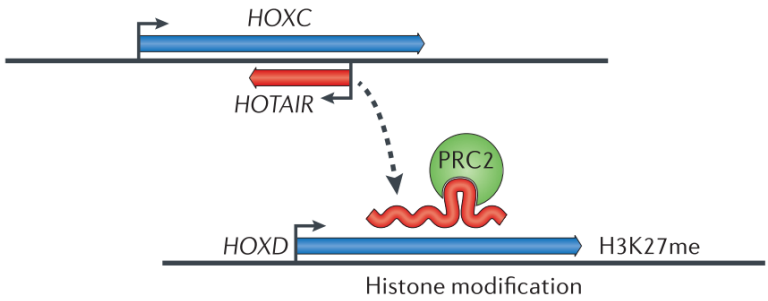


Fig. 1.16.: *HOTAIR* inhibits the homeobox D (*HOXD*) locus in *trans* via Polycomb repressive complex 2 (*PRC2*) recruitment. (Pelechano & Steinmetz [157])

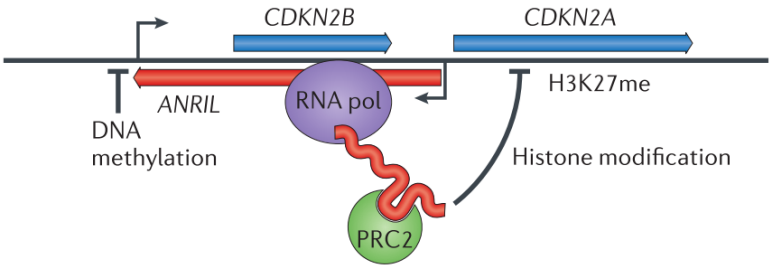


Fig. 1.17.: *ANRIL* recruits *PRC2* in *cis*, which induces histone H3 lysine 27 (*H3K27*) methylation. This represses the transcription of *CDKN2B-CDKN2A*. (Pelechano & Steinmetz [157])

Antisense lncRNA	Mechanism of action	Effect
LUC7L	DNA methylation	Methylates HBA1 promoter CpG island, which represses its expression
XIST	Chromatin modification	Inactivates X chromosome gene expression
ANRIL	Chromatin modification	Represses the tumor-suppressor locus CDKN2B–CDKN2A by both *H3K27* methylation and DNA methylation
BDNF-AS	Chromatin modification	Represses BDNF by histone modification
HOTAIR	Chromatin modification	Silences the HOXD locus in trans by the recruitment of Polycomb proteins
ZEB2-AS	Isoform variation	Induces exon skipping in ZEB2, which produces an alternative isoform with increased translation efficiency
BACE1-AS	RNA stability	Increases stability of BACE1 by masking an miRNA-binding site
WDR83, DHPS	RNA stability	Increase their mutual stability by forming a duplex within their 3' untranslated regions

Tab. 1.9.: Examples of antisense lncRNAs effects on gene expression. (Pelechano & Steinmetz [157])

Antisense lncRNAs in cancer

As shown in Table 1.9, several antisense lncRNAs play a role in cancer.

The oncogenic process involving antisenses can involve tumor supressor genes:

- In 2008, a study had identified antisense lncRNAs for each of 21 well-known tumor suppressor genes, and it showed that tumor suppressor *CDKN2B* was silenced, through the expression of its antisense [159].
- *Wrap53*, a natural antisense transcript of *p53*, regulates endogenous *p53* mRNA levels. Moreover, a knock-down of *Wrap53* leads to a decrease in *p53* levels and to the removal of *p53* induction following DNA damage [160].

These 2 examples already show that antisense lncRNAs can behave both in concordant or discordant regulation, i.e. the antisense lncRNA can either augment or lower the levels of the corresponding mRNA, respectively. The majority of sense/antisense pairs of mammalian transcripts show concordant regulation [161].

Several other lncRNAs have been studied and shown to be associated with cancer:

- *HIF-1*, which is a transcription factor regulating genes involved in angiogenesis, invasion, and tumor progression, is over-expressed in a large part of human cancers and it is correlated with poor prognosis and chemoresistance. Camptothecin (CPT), which is an antitumor DNA topoisomerase I (Top1) inhibitor, increases the levels of two antisense lncRNAs of *HIF-1* [162].
- An over-expression of *NCYM*, which is a an antisense of *MYCN*, is associated with poor prognosis in neuroblastoma via promotion of production of anti-apoptotic protein *Myc-nick* [163].
- *Survivin*, which inhibits apoptosis, is expressed in cancer cells. *EPR-1*, which is an antisense lncrRNA of *Survivin* downregulates its expression, resulting in a decrease in cell proliferation and an increase in apoptosis [164] .
- *HOTAIR* is strongly induced in approximately 25% of breast cancers, and *HOTAIR* expression is highly predictive of metastasis and death. *HOTAIR* overexpression drives breast cancer metastasis in vivo and elevated *HOTAIR* levels are predictive of metastasis or progression in colon and liver cancers, suggesting a general oncogenic trait [154].
- *WT1* is a developmental gene mutated in Wilms' tumor (WT) and acute myeloid leukaemia (AML). Its antisense, *WT1-AS*, is aberrantly spliced in AML and is subject to epigenetic defects in WT [161].
- *AFAP1-AS1*, which is an antisense of *AFAP1*, is extremely hypomethylated and overexpressed in Barrett's esophagus (BE) and esophageal adenocarcinoma

(EAC). When it is silenced by siRNA, there is an inhibition of proliferation and colony-forming ability, induced apoptosis, and reduced EAC cell migration and invasion [165].

- *p15*, which is a cyclin-dependent kinase inhibitor involved in leukaemia, is silenced through the recruitment of PRC2 by the *ANRIL* antisense [166].

These numerous examples highlight the important role played by antisense lncRNAs in human cancers. In some cases, the mechanism by which the antisense lncRNA acts is well understood, while in others, only an over- or under-expression, or a mutation of the antisense is reported. It should be noted that antisense lncRNAs and lncRNAs in general are also involved in several other pathologies such as Alzheimer's disease, spinocerebellar ataxia, Parkinson's disease, schizophrenia, fragile X syndrome, Huntington's disease, psoriasis, etc.

This global presence of regulatory mechanisms involving lncRNAs and antisenses tend to support their role as an added layer of complexity in most biological pathways [167].

2

Exome Copy Number Variation detection: use of a pool of unrelated healthy tissue as reference sample

> *Technology presumes there's just one right way to do things and there never is.*
>
> **— Robert M. Pirsig**

Summary

The detection of copy-number variation (CNV) alterations in the genomic profile of cancer patients constitutes a standard procedure routinely performed in hospitals. Over the years, different techniques have been used to gather the CNV profile of samples, with varying precision and cost. As of today, comparative genomic hybridization microarrays remain the most used technique.

In this context, the use of exome sequencing, which is already applied to detect point mutations, as a tool able to yield CNV profiles had already been considered. However, in a cancer setting, existing exome-based techniques often make use of a healthy reference sample from the same patient. Existing publications had suggested that a pool of unrelated healthy individuals could be used as reference sample, but the idea had never been formally tested.

Here, we validated this proposal by testing it on a small number of multiple myeloma patients. Moreover, faced with the imprecision of existing CNV profiles comparison methods, we introduced a new metric, designed to quantify the difference between CNV profiles irrespectively of the technique used to generate said profiles.

My personal contributions to this research project range from the first bioinformatics steps required by the exome sequencing analysis, to the different comparisons performed between CNV profiles. I also designed and implemented the distance

metric used to compare CNV profiles. Finally, I took part in the writing of the manuscript.

Received: 31 July 2015 | Revised: 9 August 2016 | Accepted: 21 September 2016

DOI: 10.1002/gepi.22019

RESEARCH ARTICLE

WILEY Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

Exome copy number variation detection: Use of a pool of unrelated healthy tissue as reference sample

Stephane Wenric[1]* | Tiberio Sticca[1]* | Jean-Hubert Caberg[3] | Claire Josse[1] | Corinne Fasquelle[1] | Christian Herens[3] | Mauricette Jamar[3] | Stéphanie Max[4] | André Gothot[4] | Jo Caers[2,5] | Vincent Bours[1,3]

[1]Laboratory of Human Genetics, GIGA-Research, University of Liège, Liège, Belgium

[2]Laboratory of Haematology, GIGA-Research, University of Liège, Liège, Belgium

[3]Department of Human Genetics, University Hospital (CHU), Liège, Belgium

[4]Department of Haematology and Immunohaematology, University Hospital (CHU), Liège, Belgium

[5]Department of Clinical Haematology, University Hospital (CHU), Liège, Belgium

Correspondence
Vincent Bours, GIGA-Research, Laboratory of Human Genetics, University of Liège, Domaine Universitaire, CHU, Sart Tilman, 4000 Liège, Belgium.
Email: vbours@ulg.ac.be

*These authors contributed equally to this work.

ABSTRACT

An increasing number of bioinformatic tools designed to detect CNVs (copy number variants) in tumor samples based on paired exome data where a matched healthy tissue constitutes the reference have been published in the recent years. The idea of using a pool of unrelated healthy DNA as reference has previously been formulated but not thoroughly validated. As of today, the gold standard for CNV calling is still aCGH but there is an increasing interest in detecting CNVs by exome sequencing. We propose to design a metric allowing the comparison of two CNV profiles, independently of the technique used and assessed the validity of using a pool of unrelated healthy DNA instead of a matched healthy tissue as reference in exome-based CNV detection. We compared the CNV profiles obtained with three different approaches (aCGH, exome sequencing with a matched healthy tissue as reference, exome sequencing with a pool of eight unrelated healthy tissue as reference) on three multiple myeloma samples. We show that the usual analyses performed to compare CNV profiles (deletion/amplification ratios and CNV size distribution) lack in precision when confronted with low LRR values, as they only consider the binary status of each CNV. We show that the metric-based distance constitutes a more accurate comparison of two CNV profiles. Based on these analyses, we conclude that a reliable picture of CNV alterations in multiple myeloma samples can be obtained from whole-exome sequencing in the absence of a matched healthy sample.

KEYWORDS

NGS, WES, CNV, aCGH, normalization, multiple myeloma, control, read count

1 | INTRODUCTION

Copy number variations (CNVs) are genomic modifications responsible of phenotypic diversity but are also involved in many pathologies like cardiovascular diseases, autoimmune diseases, neurodegenerative diseases, and cancers (Beroukhim et al., 2010; Kim et al., 2013). In cancers chromosomal alterations might lead to several specific genomic profiles which can be linked to prognosis or response to treatment, for example the amplification of the ERBB2 gene in breast cancer leads to its overexpression, and to sensitivity to treatment by trastuzumab (Robert et al., 2006).

Multiple myeloma is a hematological cancer characterized by a high level of CNV, implicating plasma cells. Some of them are linked to an adverse prognosis: del(17)(p), del(1)(p), dup(1)(q), and del(13) (Fonseca et al., 2004; Walker et al., 2010). On the other hand, hyper-diploidies involving odd chromosomes are rather associated with a favorable outcome (Smadja et al., 2001). CNV assessment during treatment course of these malignancies is also essential to evaluate disease progression (Avet-Loiseau et al., 2009; Chung, Mulligan, Fonseca, & Chng, 2013).

Traditionally, CNV detection has been performed with cytogenetic techniques such as fluorescent in situ hybridization (FISH). Comparative genomic hybridization arrays (aCGH) are currently considered as the reference technology to measure genomic alterations. However, next-generation sequencing (NGS) could soon become an essential tool for cancer study as it allows the detection of punctual mutations and insertions/deletions. Moreover, whole genome sequencing (WGS) can also be used for the detection of CNVs and displays a higher resolution than aCGH, down to 40 bp (Xi et al., 2011). However in the clinical field, WGS is too expensive and WES or targeted sequencing is more commonly considered. CNV are more easily computed from WGS data, as the entire genome is theoretically sequenced at constant

coverage and one does not have to take into account the inter-probe coverage variability that arises in WES (Hwang et al., 2015; Liu et al., 2013). That being said, WES focuses on a highly function-enriched subset of the genome and it requires smaller computational resources for processing and storage of the data than WGS. For these reasons, a number of dedicated computational algorithms have been developed to accurately retrieve segmental CNV from WES data (Guo et al., 2013; Tan et al., 2014).

Several factors are responsible for biases in CNV detection: GC rich fragments, variability of the fragmentation process during library preparation, or copy number polymorphisms. Most of the bioinformatic tools set-up for CNV detection in tumor by WES consider these potential biases and try to minimize them (Xi et al., 2011). Some of the algorithms designed to detect CNV on tumor samples also require a matched paired healthy tissue sampled from the same patient, as they use the read depth ratio between tumor and healthy sample to infer the copy number at each locus. This control sample needs to be compiled from the same technological platform. However, such paired reference tissue is very seldom available, especially in large epidemiology studies, and could theoretically be replaced by the use of a pool of unrelated healthy tissues from patients of the same ethnicity (Sathirapongsasuti et al., 2011). However, no data are currently available in the literature to state if this solution would allow the acquisition of comparable CNV results.

To evaluate if the replacement of the matched paired healthy tissue with a pool of unrelated healthy tissue confers the same results, we have compared the performances of these two reference types against results obtained by aCGH, considered as the gold standard.

The whole study was conducted on a multiple myeloma (MM) cohort. Malignant cells population was enriched by positive selection, and analyzed by WES (Nextera, Illumina) and aCGH (SureSelect, Agilent).

2 | MATERIALS

2.1 | Ethical concerns

Ethics approval was obtained from the Institutional Review Board (Ethical Committee of the Faculty of Medicine of the University of Liège) in compliance with the Declaration of Helsinki. All patients signed a written informed consent form. This work consisted of a prospective study and did not lead to any change in the treatment of enrolled patients.

2.2 | Patients and sample preparation

Bone marrow samples of 10 MM patients were obtained from CHU of Liège. CD138 human MicroBeads (Miltenyi Biotec) were used to positively select plasma cells and enrich malignant cell populations. Genomic DNA (gDNA) was extracted from enriched plasma cells using AllPrep DNA extraction kit (Qiagen) following manufacturer's instructions. Normal gDNA for three of these patients was collected and extracted from buccal cells with Gentra Puregene Buccal Cell Kit (Qiagen) following manufacturer's instructions. Eight additional normal DNA were also extracted using the same methodology and separately sequenced to constitute a pool of normal DNA.

2.3 | aCGH and CNV analysis

Plasma cells of the whole MM cohort were analyzed with the SurePrint G3 Human CGH Microarray Kit 8 × 60K (Agilent Technologies) according to manufacturer's instructions, and results were interpreted using the Cytogenomics software (Agilent Technologies). The arrays were scanned with a G2565CA microarray scanner (Agilent Technologies) and the images were extracted and analyzed with CytoGenomics software v2.0 (Agilent Technologies). An ADM-2 algorithm (cut-off 6.0), followed by a filter to select regions with three or more adjacent probes and a minimum average log2 ratio of ±0.25, was used to detect copy number changes. The quality of each experiment was assessed by the measurement of the derivative log ratio spread with CytoGenomics software v2.0. Genomic positions were based on the UCSC human reference sequence (hg19) (NCBI build 37 reference sequence assembly).

2.4 | Whole exome sequencing and CNV call

Fifty nanograms of double-stranded gDNA were used to prepare libraries with a Nextera Rapid Capture Expanded Exome Kit (Illumina) according to the manufacturer's instructions. Libraries were checked for integrity using Agilent High Sensitivity DNA Kit (Agilent Technologies) after tagmentation and after the last step of library preparation. Sequencing reactions were performed on a HiSeq2000 sequencer (Illumina).

3 | METHODS AND RESULTS

3.1 | Whole exome sequencing and CNV call

The raw sequencing data were aligned on the Human reference genome (NCBI build 137 hg19) with the BWA software (Li & Durbin, 2009). The resulting alignment BAM files went through several filtering and correcting steps (local realignment, base quality score recalibration, low quality reads filtering, and PCR duplicate reads removal) performed using the Genome Analysis Toolkit (McKenna et al., 2010) and the Picard software package (http://picard.sourceforge.net/).

A slightly modified version of the coverage files generated by the CalculateHsMetrics tool of the Picard software

package (using the *PER_TARGET_COVERAGE* software option) was used as input of the ExomeCNV software (version 1.4).

For three tumor samples for which matched normal tissue was available, two CNV profiles were called using the recommended parameters of ExomeCNV: one with the matched normal sample as control and the other one with a pool of unrelated healthy samples as control.

The ExomeCNV input file representing the pool of eight unrelated healthy samples is generated thanks to a Perl script that averages the *coverage* and *average_coverage* columns of the Exome CNV input file among all unrelated healthy samples.

The Perl scripts used to convert the output files of the CalculateHsMetrics tool to input files suitable for ExomeCNV and to generate the ExomeCNV input file for the pool are available as supplementary data.

3.2 | CNV profiles comparison

Several analyses were performed to compare the CNV profiles obtained through aCGH, Exome CNV with the matched normal sample as control and Exome CNV with the pool of eight unrelated healthy samples as control. Only autosomes were considered in this study.

For the sake of brevity, for each sample S_k, let us note S_kC, S_kM, S_kP, respectively, the CNV profile obtained through arrayCGH, the CNV profile obtained through the Exome CNV software with the matched normal sample as control and the CNV profile obtained through the Exome CNV software with the pool of eight unrelated healthy samples as control.

3.3 | Deletion/Amplification ratio

The deletion/amplification ratio has been determined for each CNV profile, as to detect possible method-specific biases. Amplifications and deletions with a |LRR| (|Log-R-Ratio|) smaller than 0.29 (corresponding to alterations whose copy number is approximately between 1.6 and 2.4) were considered to be inconsistent and were filtered out for all CNV profiles. The ratio is based on the total number of deleted and amplified bases, as this gives a more reliable information than a ratio based on the count of amplifications and deletions. As shown on Figure 1, both Samples 1 and 3 show close deletion/amplification ratios and absolute values for each of the three profiles. No specific bias in favor of amplification or deletion is found in the CNV profiles obtained through Exome CNV software with the pool of eight unrelated healthy samples as control. Interestingly, for Sample 2, the absolute values for the number of deleted bases are very similar, but the number of amplified bases varies. S_2P thus shows a deletion/amplification ratio much more similar to S_2C than S_2M. This is explained by the fact that most of the missing amplifications in S_2M are present but have in fact a low LRR and are filtered out. Due to their low LRR, these amplifications are undistinguishable from false positives.

3.4 | CNV size distribution

To know if the use of a pool as reference had an impact on the size of detected CNV, we determined the CNV size distribution for each profile (see Fig. 2). Amplifications and deletions with a |LRR| smaller than 0.29 were filtered out. Although the absolute count of very small CNVs (< 1 kb) is higher in profiles obtained through the use of a pool as reference, their relative contribution remains unchanged and insignificant (see Additional File 1).

3.5 | Confusion matrix

For both exome-based CNV profiles of each sample (S_kM and S_kP), TPR (true-positive rate), FPR (false-positive rate), TNR (true-negative rate), and FNR (false-negative rate) were determined separately for amplifications and deletions, as shown in Table 1. Amplifications and deletions with a |LRR| smaller than 0.29 were filtered out. Interestingly, CNV profiles obtained through the use of a pool of eight unrelated healthy individuals yield overall slightly better results. The low TPR value for the amplifications of S_2M can again be explained by LRR values not passing the aforementioned threshold.

3.6 | CNV profile distance metric

Each of the previous analyses highlights potential biases, similarities, and/or discrepancies between CNV profiles but, due to methodological specificities, none gives a global picture of the real distance between profiles, as even the outcomes of the confusion matrix do not take into accounts the variation of copy number in amplifications and deletions (e.g., if the reference contains a segment of copy-number 3 and the tested profile contains a segment of copy-number 4 at the same locus, both profiles are considered to contain an amplification, no difference penalty is taken into account and the confusion matrix values are the same as they would be if both segments shared the exact same copy-number).

We propose a new distance metric, designed to compare CNV profiles, which takes into account the exact LRR values, thus giving a more precise insight, independently of the technique used to obtain the profiles.

Each CNV profile is represented as a combination of sequences of LRR segments of fixed size, one sequence for each chromosome.

Let j, k be the indexes of two CNV profiles.

At each segment s, the difference in terms or LRR between the two profiles is noted as

$$\left| LRR_{sj} - LRR_{sk} \right|.$$

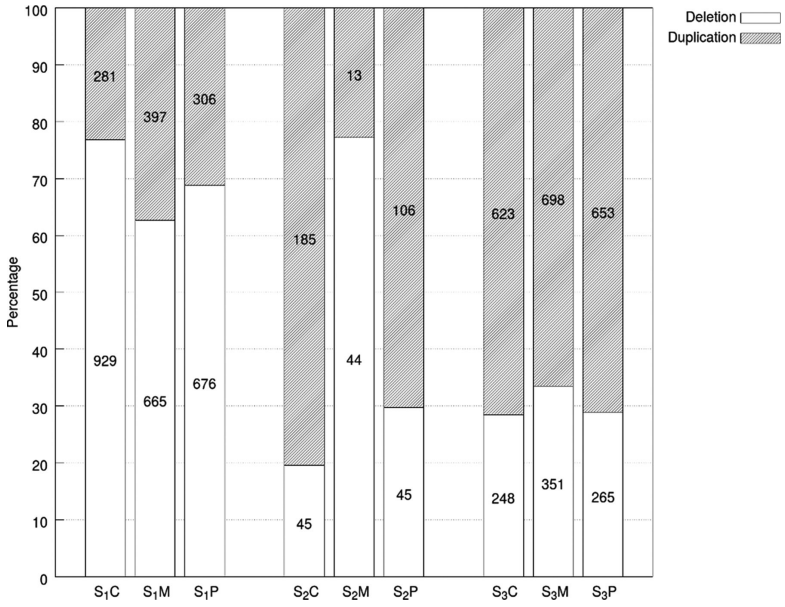


FIGURE 1 Amplification and deletion ratios for each sample. Bar heights represent the percentage of amplified or deleted bases. The total number of amplified or deleted megabases is written inside each bar

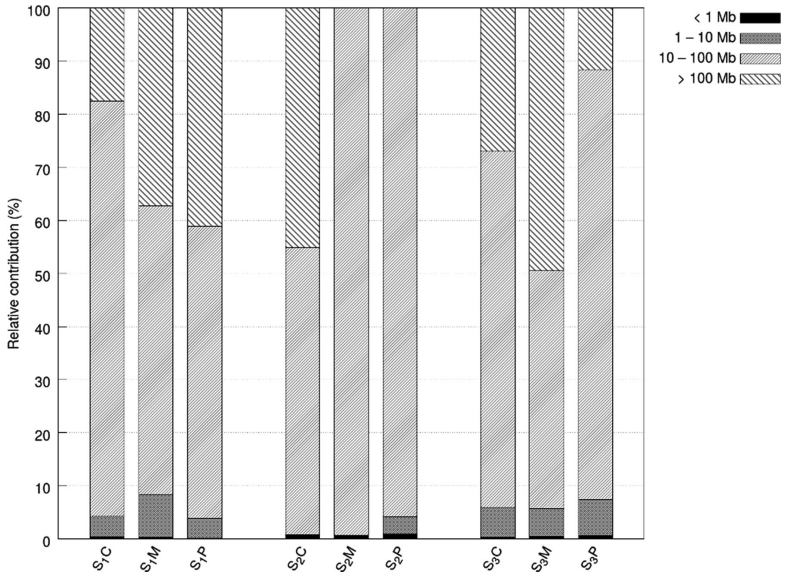


FIGURE 2 CNV size distribution. Bar heights represent the relative contribution of each group

From each LRR value, the corresponding copy-number for autosomes can be derived by

$$CN = 2 \times 2^{LRR}.$$

The difference in terms of copy number at each segment s can thus be expressed as

$$\left|CN_{sj} - CN_{sk}\right| = \left|2 \times \left(2^{LRR_{sj}} - 2^{LRR_{sk}}\right)\right|.$$

We defined the distance metric between two CNV profiles as the sum of distances between all segments divided by the total number of segments.

$$d\,(j,k) = \frac{1}{S} \times \sum_{s}^{S} \left|2 \times \left(2^{LRR_{sj}} - 2^{LRR_{sk}}\right)\right|,$$

where S is the total number of segments.

The relation between the genome size, segment size, and total number of segments is noted:

TABLE 1 Confusion Matrix

		S1_M	S1_P	S2_M	S2_P	S3_M	S3_P
AMP	TPR	97.29	96.58	12.31	99.24	99.96	97.01
	FPR	2.71	3.42	87.69	0.76	0.04	2.99
	TNR	63.83	82.14	99.98	97.79	89.23	92.63
	FNR	36.17	17.86	0.02	2.21	10.77	7.37
DEL	TPR	94.83	96.43	98.01	98.01	96.82	95.88
	FPR	5.17	3.57	1.99	1.99	3.18	4.12
	TNR	97.5	97.59	99.63	97.4	68.06	89.29
	FNR	2.5	2.41	0.37	0.6	31.94	10.71

TABLE 2 Distance Between Each CNV Profile

Sheet 1									
	S1_M	S1_P	S1_C	S2_M	S2_P	S2_C	S3_M	S3_P	S3_C
S1_M	0	1.4	1.9	3.5	3.5	3.5	4.8	4.7	4.1
S1_P	1.4	0	1.3	3.5	3.1	3.3	4.5	4.5	4.2
S1_C	1.9	1.3	0	2.6	2.5	2.3	4.6	4.4	4.2
S2_M	3.5	3.5	2.6	0	0.7	0.4	4	3.7	3.6
S2_P	3.5	3.1	2.5	0.7	0	0.5	3.8	3.6	3.7
S2_C	3.5	3.3	2.3	0.4	0.5	0	4	3.6	3.5
S3_M	4.8	4.5	4.6	4	3.8	4	0	1	2
S3_P	4.7	4.5	4.4	3.7	3.6	3.6	1	0	2
S3_C	4.1	4.2	4.2	3.6	3.7	3.5	2	2	0

$G = S \times L$, where G is the genome size and L is the segment size

A Perl script implementing this distance metric is available as supplementary data. For clarity, all distance values have been multiplied by 10.

Table 2 shows distance values computed for all possible combinations of the nine CNV profiles generated based on our cohort. Several observations can be made based on these results. For each sample, the smallest distance is always found between the two profiles obtained through the use of ExomeCNV. For each sample, the distance between the aCGH profile and the ExomeCNV profile using a pool of eight unrelated healthy individual as control is similar to the distance between the aCGH profile and the ExomeCNV profile using the matched paired healthy tissue as control. The intersample distance, whatever the technique, is always greater than the intrasample distance.

3.7 | Additional validation

The same analyses were performed on 7 MM samples for which no matched normal tissue was available. Here, only S_kP and S_kC (respectively the profile obtained through ExomeCNV with a pool, and the profile obtained through aCGH) were compared.

The proportion of deletion to amplification does not show any specific bias and the amplified and deleted bases counts are highly correlated between S_kP and S_kC (Pearson correlation coefficient of 0.975, see Additional file 2 for the count of deleted and amplified bases).

The confusion matrix values obtained when comparing S_kP to $S_kC$are relatively similar to the previous values obtained. The average values for the true-positive rate and the true negative rate for the amplifications are respectively 89.85% and 92.41%. The corresponding values for the deletions are respectively 97.04% and 74.97% (see Additional file 3 for the complete data).

The distance metric was computed for each pair of the 14 profiles. As previously, for each sample the intrasample distance is always smaller than all intersamples distances involving this sample. The average value for intrasample distance was 1.8 ± 0.2, while the average value for intersample distance was 3.25 ± 0.16. All distance values are shown in additional file 4.

4 | DISCUSSION

To date, several CNV detection tools catered to WES data exist, some of these tools make use of paired healthy DNA as references, while others use different methodologies and do not need such references. Paired methods that use the read depth or read count ratio are often more effective but inadequate for the analysis of a sample without corresponding healthy DNA.

Although ExomeCNV is a method based on read depth using paired healthy DNA as control, its authors suggested that a pool of unrelated healthy individuals could also be used as reference. Based on preliminary results, the authors also emit the hypothesis that the use of such a pool could lead to more reliable results thanks to a reduction in variance of depth-of-coverage (Sathirapongsasuti et al., 2011). No thorough analysis had previously been performed to assess the validity of these claims. Furthermore, we propose a new, better suited, way to compare CNV profiles, independently of the technique used to obtain said profile and tested this method on a small number of multiple myeloma samples.

Research and clinical application of WES for CNV detection are most useful in the cancer field. Indeed, many clinically actionable genetic changes have been described. These changes include CNV (deletions, amplifications) as well as punctual mutations. Their identification has an increasing clinical impact as they define the prognosis and can also predict treatment response or resistance, paving the way toward personalized medicine and the use of specific targeted treatments. The molecular diagnosis remains, however, difficult as it is presently based on limited amounts of DNA (from biopsies) and has to deal with the tumor heterogeneity. Moreover, large retrospective studies based on samples stored in biobanks are needed to validate genetic biomarkers in various cancers.

In our study we explored MM which is characterized by a high genomic instability. Indeed, alterations with clinical impact like monosomy 13 and trisomy of odd chromosomes are easily detected with this method while partial alterations of chromosome 1 and 17 sometimes show some approximation concerning the exact breakpoints. Even if the impact of punctual mutations in this type of cancer is still unclear, a few studies performed by NGS show a high level of mutations implicating genes frequently involved in cancers and coding for therapeutic targets (Chapman et al., 2011; Lohr et al., 2014). However, as it has been demonstrated that MM is characterized by a high level of clonal heterogeneity according to the stage of the disease, WES allows an evaluation of each clonal population proportion at the different stages of the disease (Walker et al., 2014). It could therefore be helpful for the follow-up of patients to evaluate clonal evolution in response to treatment at relapse. A simple method identifying point mutations and CNVs is certainly required for such a clinical application.

In conclusion, our data indicate that a reliable picture of CNV alterations in MM samples could be obtained from WES in the absence of a matched healthy sample. As our data were obtained on a very low number of MM samples, they will need to be confirmed on larger cohorts of other cancer types. If this can be done, it would considerably facilitate genomic studies on biobank material as well as in the clinical setting as the collection, study and data storage for matched normal DNA is expensive and generates cancer-unrelated incidental findings.

ACKNOWLEDGMENTS

We thank the patients. We also thank the GIGA-genotranscriptomic platform. Funding was obtained from the following institutions: National Fund for Scientific Research (PDR, Télévie, and FRIA); Centre-Anti-Cancéreux (ULg, Liège); Fond d'Investissement à la recherche Scientifique (CHU, Liège); Région Wallonne; and University of Liège (Fonds spéciaux pour la recherche).

All the authors declare that they have no conflict of interest.

REFERENCES

Avet-Loiseau, H., Li, C., Magrangeas, F., Gouraud, W., Charbonnel, C., Harousseau, J. L., …, Minvielle, S. (2009). Prognostic significance of copy-number alterations in multiple myeloma. *Journal of Clinical Oncology*, *27*(27), 4585–4590.

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., …, Meyerson, M. (2010). The landscape of somatic copy-number alteration across human cancers. *Nature*, *463*(7283), 899–905.

Chapman, M. a, Lawrence, M. S., Keats, J. J., Cibulskis, K., Sougnez, C., Schinzel, A. C., …, Golub, T. R. (2011). Initial genome sequencing and analysis of multiple myeloma. *Nature*, *471*(7339), 467–472.

Chung, T. H., Mulligan, G., Fonseca, R., & Chng, W. J. (2013). A novel measure of chromosome instability can account for prognostic difference in multiple myeloma. *Plos One*, *8*(6), 1–8.

Fonseca, R., Barlogie, B., Bataille, R., Bastard, C., Bergsagel, P. L., Chesi, M., …, Avet-Loiseau, H. (2004). Genetics and cytogenetics of multiple myeloma: A workshop report. *Cancer Research*, *64*, 1546–1558.

Guo, Y., Sheng, Q., Samuels, D. C., Lehmann, B., Bauer, J. A., Pietenpol, J., & Shyr, Y. (2013). Comparative study of exome copy number variation estimation tools using array comparative genomic hybridization as control. *BioMed Research International*, *2013*, 1–7.

Hwang, M. Y., Moon, S., Heo, L., Kim, Y. J., Oh, J. H., Kim, Y.-J., …, Kim, B.-J. (2015). Combinatorial approach to estimate copy number genotype using whole-exome sequencing data. *Genomics*, *105*(3), 145–149.

Kim, T. M., Xi, R., Luquette, L. J., Park, R. W., Johnson, M. D., & Park, P. J. (2013). Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. *Genome Research*, *23*(2), 217–227.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–1760.

Liu, B., Morrison, C. D., Johnson, C. S., Trump, D. L., Qin, M., Conroy, J. C., …, Liu, S. (2013). Computational methods for detecting copy number variations in cancer genome using next generation sequencing: Principles and challenges. *Oncotarget*, *44*(11), 1868–1881.

Lohr, J., Stojanov, P., Carter, S., Cruz-Gordillo, P., Lawrence, M., Auclair, D., …, Golub, T. R. (2014). Widespread genetic heterogeneity in multiple myeloma: Implications for targeted therapy. *Cancer Cell*, *25*, 91–101.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., …, DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303.

Robert, N., Leyland-Jones, B., Asmar, L., Belt, R., Llegbodu, D., Loesch, D., …, Slamon, D. (2006). Randomized phase III study of trastuzumab, paclitaxel, and carboplatin compared with trastuzumab and paclitaxel in women with HER-2 - Overexpressing metastatic breast cancer. *Journal of Clinical Oncology*, *24*(18), 2786–2792.

Sathirapongsasuti, J. F., Lee, H., Horst, B. A. J., Brunner, G., Cochran, A. J., Binder, S., …, Nelson, S. F. (2011). Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics*, *27*(19), 2648–2654.

Smadja, N. V. (2001). Hypodiploidy is a major prognostic factor in multiple myeloma. *Blood*, *98*(7), 2229–2238.

Tan, R., Wang, Y., Kleinstein, S. E., Liu, Y., Zhu, X., Guo, H., …, Zhu, M. (2014). An evaluation of copy number variation detection tools from whole-exome sequencing data. *Human Mutation*, *35*(7), 899–907.

Walker, B. A., Leone, P. E., Chiecchio, L., Dickens, N. J., Jenner, M. W., Boyd, K. D., …, Morgan, G. J. (2010). A compendium of myeloma-associated chromosomal copy number abnormalities and their prognostic value. *Blood*, *116*, 56–65.

Walker, B. A., Wardell, C., Melchor, L., Brioli, A., Johnson, D., Kaiser, M. F., …, Morgan, G. J. (2014). Intraclonal heterogeneity is a critical early event in the development of myeloma and precedes the development of clinical symptoms. *Leukemia*, *28*, 384–390.

Xi, R., Hadjipanayis, A. G., Luquette, L. J., Kim, T.-M., Lee, E., Zhang, J., …, Tibshirani, R. (2011). Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *PNAS*, *108*, 1128–1136.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

3

Circulating microRNA-based screening tool for breast cancer

> *Treat the patient, not the X-ray.*
>
> **— James M. Hunter**

Summary

In this study, we further advanced the emerging field of non-invasive diagnosis using circulating biomarkers. Through the recruitment of a large cohort of breast cancer patients and age-matched controls, we were able to gather plasma samples and detect the levels of circulating microRNAs (miRNAs) by the use of RT-qPCR.

Basing our work on the existing differences between cases and controls for a large number of miRNAs, we have developed a multivariate diagnostic model able to classify samples into the 2 classes.

We went further than previous studies, by showing that our model, which exhibited fine performance on a substantial validation cohort, was also able to discriminate between breast cancer and ovarian cancer samples, and to correctly classify breast cancers in remission, and metastatic breast cancers.

My personal contributions to this research project were centered on the development of the random forests-based methodology, encompassing the feature selection step, the model building, and the validation of the model on all cohorts.

I also took part in the writing of the manuscript.

Circulating microRNA-based screening tool for breast cancer

Pierre Frères[1,2,*], Stéphane Wenric[2,*], Meriem Boukerroucha[2], Corinne Fasquelle[2], Jérôme Thiry[2], Nicolas Bovy[3], Ingrid Struman[3], Pierre Geurts[4], Joëlle Collignon[1], Hélène Schroeder[1], Frédéric Kridelka[5], Eric Lifrange[6], Véronique Jossa[7], Vincent Bours[2,*], Claire Josse[2,*], Guy Jerusalem[1,*]

[1]University Hospital (CHU), Department of Medical Oncology, Liège, Belgium

[2]University of Liège, GIGA-Research, Laboratory of Human Genetics, Liège, Belgium

[3]University of Liège, GIGA-Research, Laboratory of Molecular Angiogenesis, Liège, Belgium

[4]University of Liège, GIGA-Research, Department of EE and CS, Liège, Belgium

[5]University Hospital (CHU), Department of Gynecology, Liège, Belgium

[6]University Hospital (CHU), Department of Senology, Liège, Belgium

[7]Clinique Saint-Vincent (CHC), Department of Pathology, Liège, Belgium

[*]These authors contributed equally to this work

Correspondence to: Guy Jerusalem, **e-mail:** g.jerusalem@chu.ulg.ac.be

Keywords: *breast cancer, circulating microRNAs, biomarkers, minimally invasive screening*

Received: *June 24, 2015* ***Accepted:*** *December 05, 2015* ***Published:*** *December 29, 2015*

ABSTRACT

Circulating microRNAs (miRNAs) are increasingly recognized as powerful biomarkers in several pathologies, including breast cancer. Here, their plasmatic levels were measured to be used as an alternative screening procedure to mammography for breast cancer diagnosis.

A plasma miRNA profile was determined by RT-qPCR in a cohort of 378 women. A diagnostic model was designed based on the expression of 8 miRNAs measured first in a profiling cohort composed of 41 primary breast cancers and 45 controls, and further validated in diverse cohorts composed of 108 primary breast cancers, 88 controls, 35 breast cancers in remission, 31 metastatic breast cancers and 30 gynecologic tumors.

A receiver operating characteristic curve derived from the 8-miRNA random forest based diagnostic tool exhibited an area under the curve of 0.81. The accuracy of the diagnostic tool remained unchanged considering age and tumor stage. The miRNA signature correctly identified patients with metastatic breast cancer. The use of the classification model on cohorts of patients with breast cancers in remission and with gynecologic cancers yielded prediction distributions similar to that of the control group.

Using a multivariate supervised learning method and a set of 8 circulating miRNAs, we designed an accurate, minimally invasive screening tool for breast cancer.

INTRODUCTION

Breast cancer is the most frequently diagnosed cancer in females worldwide; its rate in Western countries has increased since the 1990s [1]. During the same period, mortality from breast cancer has decreased due to early detection and improved treatments [2].

Currently, mammographic screening, followed by invasive core needle biopsies in cases of suspected malignancy, allows early breast cancer diagnosis. Mammographic screening is an accessible but unpleasant and inaccurate test; in 1000 screened women, 15 of these women are estimated to have a biopsy because of a suspicious abnormality, and the biopsy is estimated to diagnose breast cancer in 4 of these 15 women [3].

MicroRNAs (miRNAs) are approximately 22-nucleotide long RNAs that inhibit gene expression by binding to target messenger RNAs (mRNAs) [4]. Currently, more than 2000 mature human miRNAs have been identified, and these miRNAs may regulate up to

60% of human protein-coding genes [5]. miRNAs are involved in multiple biological processes including cell proliferation, differentiation and apoptosis [6, 7]. Their expression is modified in various cancer subtypes, where these miRNAs act as tumor suppressors or oncogenes and play a key role in tumorigenesis [8].

All cell types release miRNAs in peripheral blood under both normal and pathological conditions. These circulating miRNAs are wrapped in 40-to 100-nm lipoprotein vesicles called exosomes, which are membrane-enclosed cell fragments [9]. These miRNAs appear to be protected from endogenous RNase activity by exosomes and are therefore particularly stable in plasma [10]. Therefore, circulating miRNAs are promising biomarkers for the early and minimally invasive diagnosis of breast cancer [11]. Several studies have already explored miRNAs from that perspective, leading to mixed results in terms of performances [12–29]. Very different diagnostic signatures have been obtained, most likely due to the choice of the sample preparation, the technology used and the study design, such as choice of proper normalization and careful validation.

In the present study, to propose new tools for breast cancer screening, we constructed a diagnostic test based on 8 circulating miRNAs and confirmed its performance in a large cohort of primary breast cancer patients and controls. The diagnostic test was also validated in patients with breast cancer in remission, patients with metastatic breast cancer and patients with gynecologic cancer to test for breast cancer specificity and follow-up. Moreover, particular attention was given to normalization and bioinformatic analysis procedures.

RESULTS

Patients and controls

Patients with treatment-naive primary breast cancer (n = 149, median age = 55 yr, range = 26–87 yr), breast cancer in remission (n = 35, median age = 49 yr, range = 28–79 yr, median time follow-up since remission = 33 months), metastatic breast cancer (n = 31, median age = 59 yr, range = 35–79 yr) and gynecologic cancer (n = 30, median age = 62 yr, range = 38–83 yr) were recruited prospectively at CHU of Liège and Clinic Saint-Vincent (Liège, Belgium) from 7/2011 to 9/2014. Gynecologic tumors consisted of non-metastatic endometrial (n = 16), ovarian (n = 10) and cervical (n = 4) cancers. Controls were obtained from 133 cancer-free females of similar age (median age = 51 yr, range = 40–74 yr) with normal mammograms (n = 72), benign calcifications (n = 30) or simple cysts (n = 31). Controls had no history of cancer in the last 5 years.

In total, 378 patients were included in this study.

All breast cancer patients and tumor characteristics are summarized in Table 1.

Pilot study

A pilot study that consisted of measuring the expression of 742 plasma miRNAs in 18 primary breast cancer patients was first conducted. In total, 188 miRNAs were chosen based on their expression levels (mean quantification cycle (Cq) value < 36) in the pilot experiment. Clinicopathological data for these patients and the list of the 188 selected miRNAs are summarized in Table 1 and Supplementary Table 1, respectively.

Evaluation of hemolysis

We first evaluated the quality of our sample collection and preparation. Hemolysis leads to the contamination of plasma with RNA from red blood cells. Absorbance at 414 nm (ABS_{414}), the maximum absorbance of hemoglobin, correlates with the degree of hemolysis. ABS_{414} was measured for all samples using a NanoDrop. The median ABS_{414} level was 0.19 ± 0.1, with a hemolysis cut-off value fixed at 0.2. Furthermore, the level of a miRNA highly expressed in red blood cells (miR-451) was compared with the level of a miRNA unaffected by hemolysis (miR-23a), with a ΔCq (miR-23a - miR-451) of more than 5 indicating possible erythrocyte miRNA contamination. The median ΔCq (miR-23a - miR-451) was 2.6 ± 1.5 in our cohort (primary breast cancer group = 3 ± 1.5, control group = 2.1 ± 1.2, breast cancer in remission group = 2.5 ± 1.5, metastatic breast cancer group = 2.8 ± 1.2, gynecologic cancer group = 2.3 ± 1.8). Based on these results, no patients were discarded.

miRNA deregulation is observed in primary as well as metastatic breast cancer patients

When comparing the miRNA profiles of newly diagnosed primary breast cancers to control miRNA profiles, 112 miRNAs were found to be significantly deregulated, with a final set of 107 miRNAs after adjusting the P-value for multiple testing. miR-16 and let-7d were the most up- and downregulated miRNAs, respectively. Global upregulation of miRNA expression was observed in primary breast cancer patients compared to controls (1.35-fold change).

In a second analysis, miRNA profiles from the plasma of patients with metastatic breast cancer were compared to those of the controls. Eighty-four miRNAs were found to be significantly deregulated, with a final set of 53 miRNAs after adjusting the P-value for multiple testing. The most significantly upregulated miRNA was miR-148a, and the most significantly downregulated miRNA was miR-15b. As observed in primary breast cancer samples, global upregulation of miRNA expression was observed in metastatic breast cancer patients when compared to healthy subjects (1.1-fold change).

Table 1: Clinicopathological data and tumor characteristics

Characteristics	Primary breast cancers – pilot study (*n* = 18)	Primary breast cancers – principal study (*n* = 149)	Metastatic breast cancers (*n* = 31)	Breast cancers in remission (*n* = 35)
Median age (range) (y)	58 (29–70)	55 (26–87)	59 (35–79)	49 (28–79)
Estrogen receptor [*n* (%)]	12 (67)	117 (79)	28 (90)	22 (63)
Progesterone receptor [*n* (%)]	11 (61)	109 (73)	22 (71)	18 (51)
HER2 [*n* (%)]	6 (33)	30 (20)	6 (19)	18 (51)
Ki67 (median ± SD) (%)	21 ± 20	20 ± 24	27 ± 23	37 ± 23
Initial T staging [*n* (%)]				
NA	0 (0)	1 (< 1)	2 (6)	0 (0)
1	3 (17)	62 (42)	9 (30)	3 (9)
2	10 (55)	58 (39)	12 (19)	19 (54)
3	2 (11)	15 (10)	6 (19)	5 (14)
4	3 (17)	13 (9)	2 (6)	8 (23)
Lymph node involvement [*n* (%)]	11 (61)	70 (47)	17 (55)	24 (69)
Tumor node metastasis (TNM) stage [*n* (%)]				
NA	0 (0)	1 (< 1)	0 (0)	0 (0)
1	2 (11)	45 (30)	0 (0)	0 (0)
2	9 (50)	73 (49)	0 (0)	20 (57)
3	7 (39)	31 (21)	0 (0)	15 (43)
4	0 (0)	0 (0)	31 (100)	0 (0)
Scarff-Bloom-Richardson grade [*n* (%)]				
NA	0 (0)	1 (< 1)	4 (13)	0 (0)
1	0 (0)	7 (5)	4 (13)	0 (0)
2	7 (39)	84 (57)	12 (39)	15 (43)
3	11 (61)	57 (38)	11 (35)	20 (57)
Histologic subtype [*n* (%)]				
NA	0 (0)	0 (0)	2 (6)	0 (0)
IDC	16 (88)	125 (84)	22 (71)	33 (94)
ILC	1 (6)	19 (13)	7 (23)	2 (6)
Others	1 (6)	5 (3)	0 (0)	0 (0)
Lymphovascular invasion [*n* (%)]	6 (33)	27 (21)	12 (39)	9 (26)

NA = not assessed; ER = estrogen receptor; PR = progesterone receptor; HER2 = human epidermal growth factor 2; IDC = invasive ductal carcinoma; ILC = invasive lobular carcinoma.

Statistical analyses were also performed to compare both primary and metastatic breast cancer patient plasma miRNA profiles to controls using the Kruskal-Wallis test. Fifty-six miRNAs were significantly modified in

the same manner among primary and metastatic breast cancer patient profiles. miR-16 and let-7d were the most co-deregulated miRNAs.

The results of the statistical analysis are available in Supplementary Table 1.

Design and validation of a diagnostic miRNA signature-based model

The analysis and computational methods relied on several steps, which made use of the random forest algorithm. The random forest algorithm is a supervised learning method that operates by building a large ensemble of decision trees, where each tree is trained on a bootstrap sample from the training data by randomizing the features that are selected at each tree node [30].

A methodology somewhat similar to the algorithmic solution proposed by Geurts *et al.* [31] was used as shown in Figure 1.

1. Model construction and miRNA signature identification

An initial random forests model was built on the profiling cohort (86 samples = 30% of the whole cohort: 41 individuals with primary breast cancer and 45 controls) with the normalized expression values of all 188 miRNAs as features to determine the 25 more discriminant miRNAs. To identify the best miRNA signature, all combinations of miRNAs that can be defined from these 25 miRNAs (33554431 in total) were then evaluated using ten-fold cross-validation on the same profiling cohort (see Materials and methods).

The best miRNA combination is composed of the following 8 miRNAs: miR-16, let-7d, miR-103, miR-107, miR-148a, let-7i, miR-19b, and miR-22*. Figure 2 summarizes the Mann-Whitney U *P*-values (Figure 2A) and relative expression changes (Figure 2B) for these 8 miRNAs.

An area under the curve (AUC) of 0.85 ± 0.02 was obtained when performing the ten-fold cross-validation in the profiling cohort.

A threshold value of 0.68 was chosen to derive a diagnostic rule from the random forest model. The value of 0.68 corresponded to an acceptable trade-off between high sensitivity (> 0.9) and satisfactory specificity (± 0.5).

2. Model validation

The validation of our model in a larger cohort (196 samples = 70% of the whole cohort: 108 individuals with primary breast cancers and 88 controls) yielded an AUC of 0.81 ± 0.01. Figure 3A represents the receiver operating characteristic (ROC) curve obtained by testing the model in the validation cohort.

With a threshold value of 0.68, a sensitivity value of 0.91 ± 0.01 and a specificity value of 0.49 ± 0.03 were obtained.

The validation of the classification model in the other cancer groups yielded slightly lower values for sensitivity (0.80 ± 0.05 for metastatic breast cancer patients) and specificity (0.40 ± 0.08 for breast cancer patients in remission and 0.41 ± 0.06 for gynecologic cancer patients) (Figure 3B). As shown in Figure 3B, the patients with breast cancer in remission and gynecologic cancer were classified as the control group.

A comparison between the miRNA signature and the established diagnostic methods

Next, we sought to compare the performance of the miRNA signature to mammographic screenings and CA15.3 assays.

The accuracy of mammographic screening is greatly affected by age. Indeed, young women have dense breasts, making the interpretation of mammography more difficult (AUC = 0.69 ± 0.05 for women under the age of 50 yr) [32]. As shown in Figure 4A, the diagnostic accuracy of the miRNA signature does not appear to be affected by age because the AUC remains stable at 0.81 in patients younger than 50 yr.

CA15.3 is the only biomarker of breast cancer, and its accuracy is directly influenced by tumor stage, with an AUC ranging from 0.56 in stage I to 0.80 in stage III breast cancers [33]. Therefore, CA15.3 is only useful for the diagnosis of late stage and metastatic breast cancers. Interestingly, tumor stage does not seem to affect the signature miRNA performance, remaining stable at 0.81 from stages I to III (Figure 4B).

miRNA signature does not correlate with breast cancer clinicopathological features

The correlations between the expression of the 8 miRNAs and the following breast cancer clinico-pathological markers were computed: estrogen and progesterone receptor expression, HER2 overexpression, tumor size, initial lymph node status, Ki67 index, Scarff-Bloom-Richardson grade and lymphovascular invasion. No significant correlation was obtained using Spearman's test for continuous variables, and no significant difference was found using the Mann-Whitney *U* test for binary variables (Supplementary Table 2).

DISCUSSION

Early breast cancer diagnosis is currently possible using mammographic screenings. However, mammographic screening has the following weaknesses: (i) the risk of false positives, with an overdiagnosis rate of up to 19%, exposing women to harmful anti-cancer therapies and affecting their quality of life; (ii) the risk of false negatives, with mammograms missing breast

cancer in 17% of cases and in more than 30% of cases for women with dense breasts and for women under hormone replacement therapy; (iii) X-ray radiation from mammograms may be one of the factors that can actually trigger breast cancer in high-risk women, e.g., young women carrying a mutation in the BRCA genes, who require early follow up beginning at 30 years, an age where mammography is less effective, and (iv) mammography performance is operator dependent (34–36).

Thus, a diagnostic test using a blood sample could add useful information. CA15.3, which is the only available biomarker for breast cancer, lacks sensitivity in the case of primary breast tumors [33].

Based on 8 circulating miRNAs, we designed a classification model using a decision tree-based ensemble method, which allows primary breast cancers to be screened with greater accuracy than mammography. Consequently, our 8 circulating miRNA signature may be

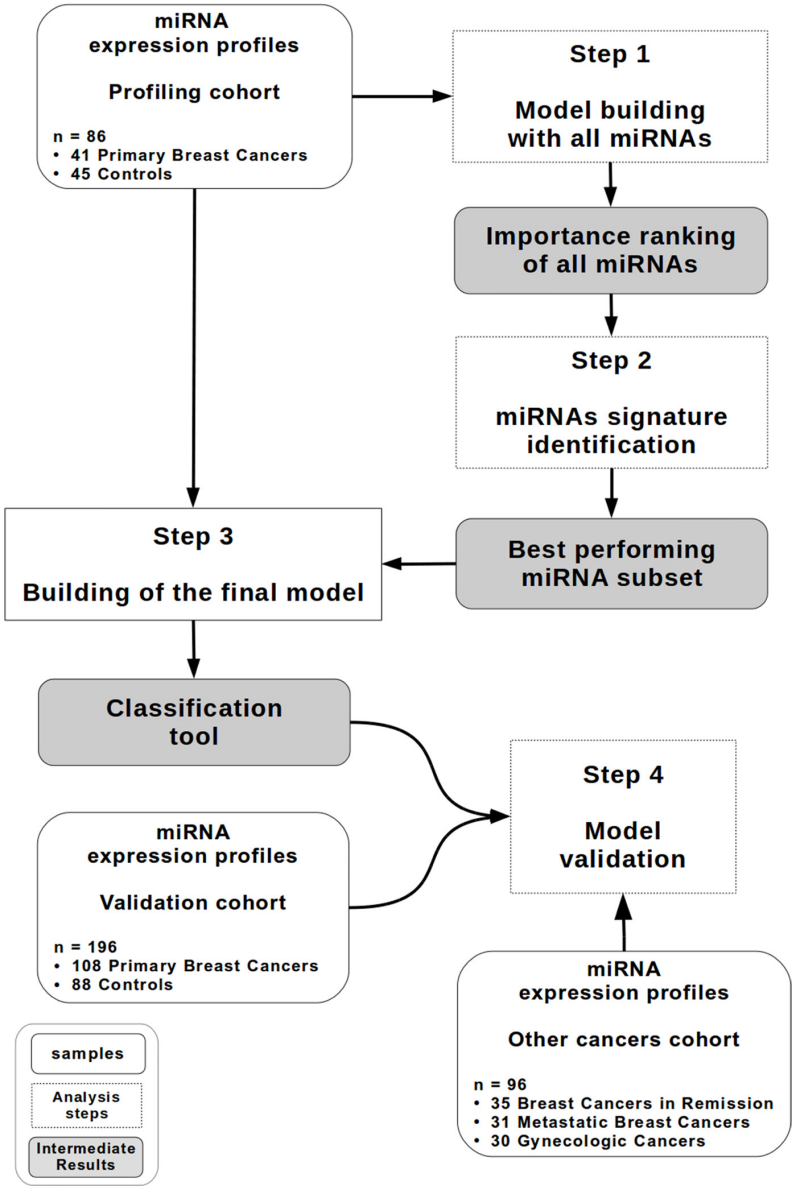


Figure 1: Study design. A diagram describing the random forest-based methodology. The profiling cohort (n = 86) contains 41 patients with primary breast cancer and 45 controls. The validation cohort (n = 196) contains 108 patients with primary breast cancer and 88 controls. The other cancer cohort (n = 96) contains 35 patients with breast cancer in remission, 31 patients with metastatic breast cancer and 30 patients with gynecologic cancer.

extremely useful to help clinicians to identify patients with a high probability of breast cancer without using invasive procedures.

The 8 miRNA-based diagnostic model shows the following interesting characteristics for clinical application: (i) this diagnostic test is not affected by age and may be useful for monitoring young women at high risk for breast cancer, in which mammography is not only less effective but also harmful because of irradiation; (ii) unlike CA15.3, this diagnostic model is effective regardless of tumor stage, which allows for detection at an early stage; (iii) this model can detect metastatic breast cancers and shows approximately the same class prediction distribution for breast cancers in remission and for controls (see Figure 3), offering a potential utility for monitoring patients; (iv) this study is the first to validate the robustness of such a classifier tool with respect to gynecologic cancers. Plasma from patients suffering of other prevalent cancers in women (cervix, endometrial and ovary cancers) [1] were used to check if the diagnostic tool could avoid the detection of other types of cancers. Indeed, the test specificity on gynecologic cancers is similar to the specificity of the control group (see Figure 3).

These aspects were overlooked in previous studies that have shown the potential of circulating miRNAs as diagnostic tools for breast cancer detection [12–29].

A

	Controls vs Primary Breast Cancers		Controls vs Metastatic Breast Cancers		Controls vs Gynecologic Cancers		Controls vs Breast Cancers in Remission	
miRNAs	*P*	Fold change	*P*	Fold change	*P*	Fold change	*P*	Fold change
miR-16	<.0001	1.7	<.05	1.2	>.05	1.5	<.05	1.4
let-7d	<.0001	0.7	<.01	0.8	>.05	0.9	<.001	1.2
miR-103	<.0001	0.8	>.05	1	>.05	1	>.05	1
miR-107	<.0001	0.8	<.001	0.8	>.05	0.9	>.05	1
miR-148a	<.0001	1.4	<.0001	1.6	<.0001	1.9	<.0001	0.5
let-7i	<.001	0.9	<.05	1.1	>.05	1.1	>.05	0.9
miR-19b	<.0001	1.2	>.05	0.9	>.05	1.1	<.0001	0.8
miR-22*	>.05	1	>.05	1.1	<.01	1.5	<.01	1.4

B

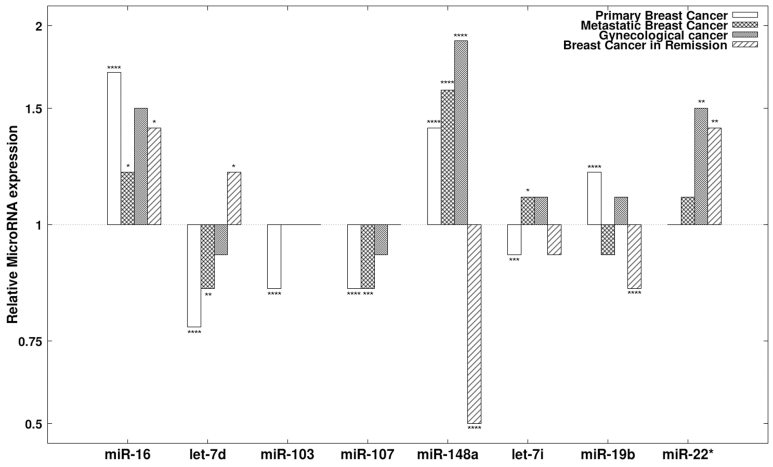


Figure 2: The 8 miRNAs present in the diagnostic signature. (A) The results of statistical analyses comparing the expression of the 8 miRNAs present in the diagnostic signature between different groups. The 8 diagnostic miRNAs were compared between primary breast cancer patients, breast cancer patients in remission, metastatic breast cancer patients, gynecologic cancer patients and the controls. *P*-values and Benjamini-Hochberg adjusted *P*-values were obtained using the Mann-Whitney *U* test. **(B)** The relative expression (mean fold change) of the 8 diagnostic miRNAs in patients with primary breast cancer, patients with breast cancer in remission, patients with metastatic breast cancer and patients with gynecologic cancer compared to controls.

The signatures that these studies have defined differed greatly from one study to another. These discrepancies can be explained by the use of different analysis methods, sample processing and normalization procedures. In the present paper, we show that the appropriate use of a subset of miRNAs combined with a specific normalization method and classification algorithm yields satisfactory results in multiple cohorts. Although decision tree ensemble methods have been proven to be efficient for the classification of biological samples based on various biomarkers [31], to our knowledge, few studies, and never in the field of breast cancer, have used random forest models with miRNA expression values as input features. Two similar studies have nevertheless shown that random forest perform better than other supervised learning methods using miRNA expression values [37, 38].

A second important concern is the normalization choice because the results of the relative quantification obtained by qPCR are entirely dependent on this process. Most of these studies used miR-16 expression alone as a reference gene [20, 25, 28, 29]. However, miR-16, which is predominantly derived from erythrocytes, has been shown to be prone to artificial elevation by hemolysis [18]. The use of blood cell-derived miRNAs as housekeeping RNA for normalization may be more problematic in cases of anemia, a condition often occurring in breast cancer

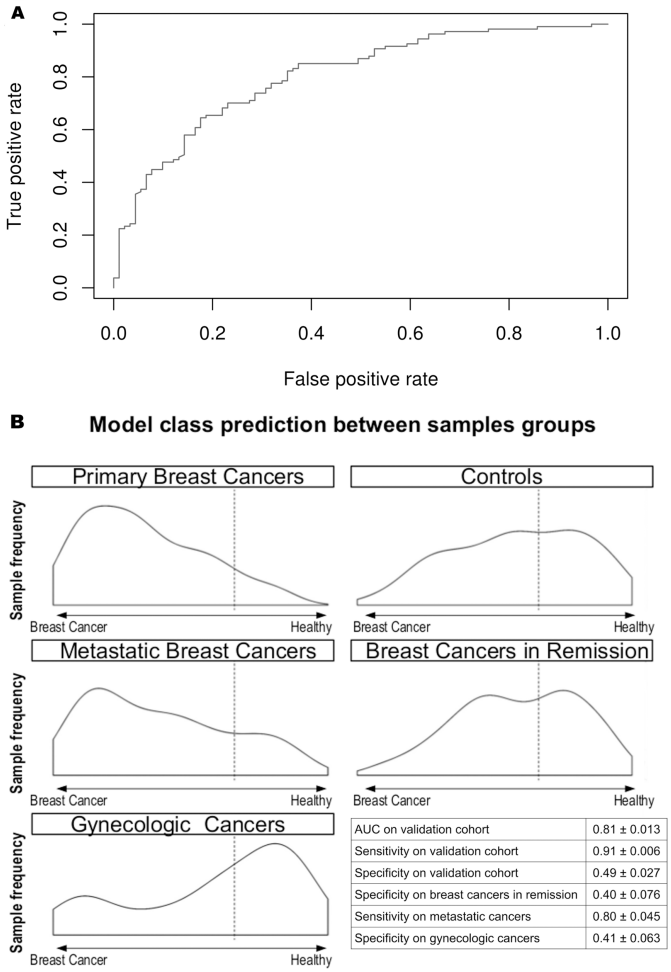


Figure 3: Circulating miRNA-based diagnostic tool performance in the validating cohort. (A) The ROC curve of the diagnostic miRNA model applied to the validating cohort. The AUC obtained is 0.81. **(B)** Model outcome distributions for the primary breast cancers, controls, metastatic breast cancers, breast cancers in remission, and gynecologic cancers. The x-axis corresponds to the model predictions. The dashed line represents the chosen threshold used to compute the sensitivity and specificity values for each cohort. The table reports the AUC, sensitivity and specificity in the validation cohort and the sensitivity and specificity in the other cancer cohort. The true positive count for the metastatic breast cancers is 25. The true negative count for breast cancers in remission and gynecologic cancers is 14.

patients. Meanwhile, global normalization methods have been described to best fit with qPCR analysis [39] but to lead to poor performances in discriminating healthy and cancer patients [17]. In this study, we compared different normalization methods, revealing that the mean of the 50 most expressed plasma miRNAs is more stable than many other normalization methods and allows for good discriminating performances. Interestingly, using this method, our analyses revealed that miR-16 and miR-103, which have been used in other studies as endogenous control genes, are differentially expressed in the plasma from healthy samples and cancer patients [12, 21].

Most of the 8 miRNAs that are part of the diagnostic signature are related to well-described cancer deregulation and were demonstrated to be differentially expressed in breast cancer tumoral tissues [40–44]. However, circulating miRNAs rarely show correlated levels with their tumoral expression [26]. In consequence, the miRNA composition of the diagnostic signature does not allow any conclusion on their biological functions.

Aside from the 8 miRNAs selected for our signature, several other combinations, most of which were composed of more than 8 miRNAs, yielded comparable performances. This finding can be explained by the fact that several miRNAs are often deregulated in the same manner under certain conditions, thus allowing one miRNA to be replaced by another miRNA in a specific signature. Regarding independent validation, it can be noted that, among these alternative combinations, one in particular was made of 11 miRNAs, which were measured

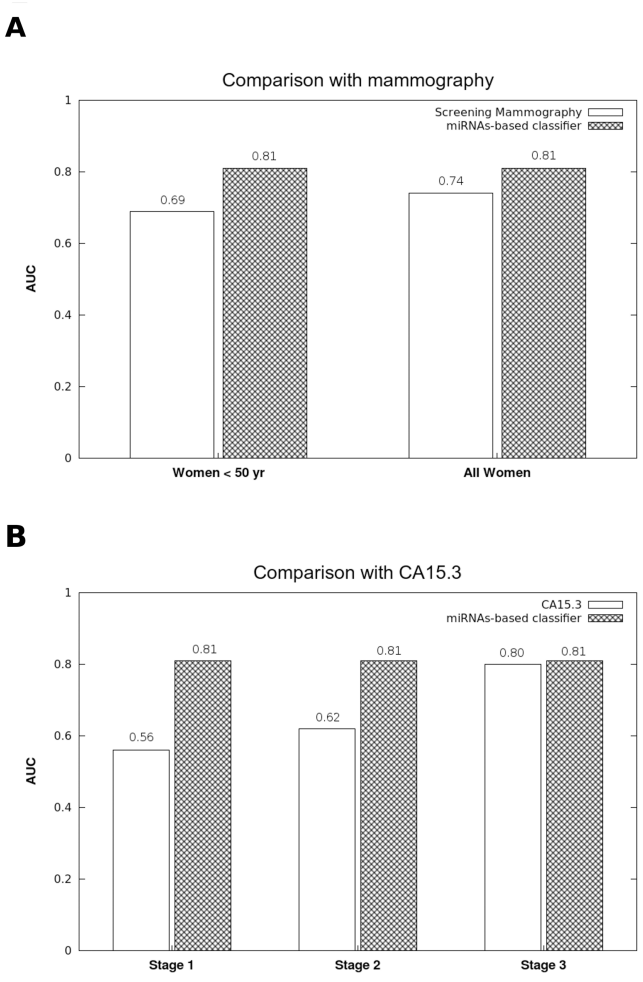


Figure 4: Comparison of the accuracy between the diagnostic miRNA signature, mammographic screenings and CA15.3 assays. (A) While the diagnostic performance of mammographic screenings is weaker in women under 50 yr (32), the AUC of the 8 miRNA-based diagnostic model was stable for women both under and over 50 yr. **(B)** The CA15.3 assay is not useful for the early diagnosis of breast cancer. While the CA15.3 AUC increases proportionally to the tumor stage (33), our model performance was stable regardless of the tumor stage.

in the serum of 54 individuals in another independent study [12]. The performance of a diagnostic model built using this alternative combination has been assessed using both our data (plasma) and the dataset GSE42128 from Chan *et al.* (serum), yielding close results (respective AUCs of 0.80 ± 0.02 and 0.77 ± 0.07, see Supplementary Table 3). Unfortunately, one of the miRNAs present in our original signature is absent from the data from Chan *et al.*, preventing us from testing the original signature.

Regarding the potential prognostic value of the 8-miRNA signature, the available follow-up of the cohorts is insufficient to determine whether the expression of the miRNAs can be correlated with progression-free or overall survival. Since there is no correlation between the expression of the 8 diagnostic miRNAs and the currently used clinicopathological factors of breast cancer, the prognostic role of the miRNA signature cannot be established on that base.

In conclusion, we established an accurate miRNA-based model for the non-invasive screening of primary breast cancer. This model also allows the identification of metastatic breast cancer and the classification of breast cancer patients in remission in the healthy group and therefore may be useful for monitoring patients. Moreover, the performance of this test is not affected by the age of the patient or by the tumor stage.

MATERIALS AND METHODS

Ethical concerns

Ethic approval was obtained from the Institutional Review Board (Ethical Committee of the Faculty of Medicine of the University of Liège) in compliance with the Declaration of Helsinki. All patients signed a written informed consent form. This work consisted of a prospective study and did not lead to any changes in the treatments of enrolled patients.

Plasma samples

Blood samples were withdrawn in 9 ml EDTA tubes. Plasma was prepared within 1 h by retaining the supernatant after double centrifugation at 4°C (10 min at 815 × *g* and 10 min at 2500 × *g*) and was stored at –80°C. The absorbance at 414 nm (ABS_{414}) was measured for all samples using a NanoDrop to evaluate the degree of hemolysis.

RNA extraction and miRNA qRT-PCR

The essential MIQE guidelines were followed during specimen preparation [45].

Circulating miRNAs were purified from 100 µl of whole-plasma using a miRNeasy Mini Kit (Qiagen, Germany) according to the manufacturer's instructions. The standard protocol was modified based on Kroh's recommendations [46]. MS2 (Roche, Belgium) was added to the samples as a carrier, and cel-miR-39 and cel-miR-238 were added as spike-ins. RNA was eluted in 50 µl of RNase-free water at the end of the procedure.

Reverse transcription was performed using a miRCURY LNA™ Universal RT microRNA PCR, Polyadenylation and cDNA Synthesis Kit (Exiqon, Denmark). Quantitative PCR was performed according to the manufacturer's instructions on custom panels of 188 selected miRNAs (Pick-&-Mix microRNA PCR Panels, Exiqon). Controls included the reference genes described in the text, inter-plate calibrators in triplicate (Sp3) and negative controls.

All PCR reactions were performed using an Applied Biosystems 7900HT Real-Time PCR System (Applied Biosystems, USA). miRNAs with Cq values < 36 were considered for analysis.

Data analysis

Analyses were conducted using the $2^{-\Delta Cq}$ method ($\Delta Cq = Cq_{sample} - Cq_{reference\ gene}$) for each sample to obtain a normalized expression value [47].

The data were normalized using the ΔCq method as recommended by Mestdagh *et al.* [39]. The mean Cq of the 50 miRNAs with the highest mean expression as determined in all the patients from all the cohorts was used for normalization because it was the most stable reference gene according to the GeNorm software. The list of the 50 miRNAs and the results of the GeNorm analysis are available in Supplementary Table 4. The whole processes of miRNA signature identification and decision tree building were also conducted on datasets normalized by 12 alternative methods. The best performances were obtained with the normalization by the mean Cq of the 50 most expressed miRNAs. The alternative normalization were: raw data, mean Cq of the 10, 20, 30 or 40 miRNAs with the highest mean expression, the mean Cq of the 50 miRNAs with the highest mean expression minus the four miRNAs that are present in the signature; the mean Cq of the spike-cel-miR-39 and the U6 small RNA; the mean Cq of miR-15b* and miR-125b (the most stable combination according to NormFinder); the global mean Cq; miR-16; the mean Cq of miR-103 and miR-191; and miR-93.

Furthermore, the delta Cq (miR-23a - miR-451) was determined for each sample to evaluate the risk of hemolysis as recommended by Blondal *et al.* [48].

Finally, data homogeneity was tested to detect outliers. Patients presenting extreme values (mean ± 3 sigma) were discarded. This operation led to the elimination of one patient from the analysis.

Statistical analyses were performed with *R version 3.0.1* (R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3–900051–07–0,

URL: http://www.R-project.org/). To compare miRNA expression levels, two-sided Mann-Whitney *U* tests and Kruskal-Wallis one-way tests were used. To correlate the expression of the 8 diagnostic miRNAs and the clinicopathological markers in primary breast cancer patients, Spearman's tests were used for continuous variables. Statistical significance was established as $*P < 0.05$ $**P < 0.01$ $***P < 0.001$ or $****P < 0.0001$. All represented values were adjusted for multiple testing using the Benjamini-Hochberg procedure [49]. The results of the statistical analyses for selected miRNAs are summarized in Supplementary Tables 1 and 2.

Study design

For all steps of the method, an *R* implementation of Breiman's original random forest algorithm, which was provided in the *R* package *randomForest*, was used [50]. A methodology somewhat similar to the algorithmic solution proposed by Geurts *et al.* was used [31] as shown in Figure 1. The different steps are described in detail below.

1. Model building with all miRNAs

An initial random forests model was built on the profiling cohort (86 samples: 41 individuals with primary breast cancer and 45 controls) with the normalized expression values of all 188 miRNAs as features. A conservative value of 3000 for n_{tree} (number of trees in the random forest) was chosen for all steps of the construction of random forest models using our methodology. Because no significant performance change was observed for incremental values of m_{try} (number of variables randomly sampled as candidates at each split), a default value of $m_{try} = \sqrt{number\ of\ miRNAs}$ was chosen for all steps of the construction of random forest models using our methodology. A combined ranking for all 188 miRNAs based on the model importance metrics MDA (Mean Decrease in Accuracy) and MDG (Mean Decrease in Gini) was obtained through the construction of this first model.

2. miRNA signature identification

Variable selection in classification or regression methods constitutes a classical problem related to 2 distinct objectives: (i) Finding relevant variables linked to the classifier output, for interpretation purposes (in this case, finding an ensemble of miRNAs related to breast cancer), (ii) Finding a sufficiently small number of variables as to avoid over-fitting, improve model performance, and provide more cost-effective models (both in terms of computation and implementation) [51, 52]. These 2 objectives may often be contradictory, since the first one will be directed to highlighting all important variables, even if these variables are redundant, while the second one aims to limit the number of variables in the final model. We are aiming for the second objective. One variable selection method for random forests, specifically targeting the second objective, is iterative variable elimination [38, 53], where variables with the smallest importance metric are iteratively discarded until reaching a minimum out-of-bag (OOB) error. Based on the definition of MDA provided earlier and the *R* implementation of the random forests algorithm, this feature selection method is roughly equivalent to the iterative elimination of variables with the lowest MDA metric. Another variable selection methodology works the other way round, by iteratively adding variables in candidate models, based on their importance metric, computed on a previous complete model, and stopping the addition of variables when the model accuracy reaches a maximum [31, 54]. Here, we use a more exhaustive wrapper approach, where a large subset of *m* variables is first selected based on the two variable importance metrics (the OOB-related importance metric MDA, but also the Gini coefficient related importance metric MDG) provided by the *R* implementation of the random forests algorithm, and secondly all *c* possible combinations of 1 to *m* variables from this subset are considered as possible features of a potential classifier, where $c = 2^m - 1$ combinations. This approach thus differs in the fact that it constitutes an exhaustive method, which will test a very large number of combinations. Each of these potential classifiers is cross-validated (with ten folds) to determine the variables combination (also called "signature") yielding the best performing model (where model performance is measured by the AUC). Since the goal of this study is the design of a usable and affordable diagnostic tool, a limited value of $m = 25$ has been chosen (leading to $c = 33554431$). This number corresponds to threshold values of 0.001 and 1 respectively for variable importance metrics MDA and MDG. This limited value of $m = 25$ constitutes a trade-off between an exhaustive testing of the solution space and the time and computational limitations related to a diagnostic test.

3. Building the final model

A random forest model was built on the profiling cohort using the best performing miRNA subset. This classification tool constituted the final diagnostic model. The number of trees chosen to build each model was determined as in step 1, and a default value of $m_{try} = \sqrt{number\ of\ miRNAs\ in\ the\ combination}$ was chosen (i.e. $m_{try} = 3$).

The prediction of the random forest algorithm for a sample is a numerical value representing the probability for this sample to be part of a specific class (case or control). To derive a binary diagnostic rule from this numerical score, a specific threshold was picked to separate the 2 classes, and the specificity and sensitivity values of the corresponding rule were computed.

4. Model validation

Then, the classification tool was validated in a larger cohort with similar cases – controls ratio as in the profiling

cohort. The total number of samples was 2.3 times greater than profiling cohort (196 samples: 108 individuals with primary breast cancers and 88 controls).

An AUC was obtained through this validation. Sensitivity and specificity values were computed using the threshold defined using the profiling cohort.

The classification tool was also validated in a separate cohort consisting of 35 individuals with breast cancer in remission, 31 patients with metastatic breast cancer and 30 patients with gynecologic cancers.

List of abbreviations

3′-UTR = 3′-untranslated region
ABS_{414} = absorbance at 414 nm
AUC = area under the curve
Cq = quantification cycle
DNA = deoxyribonucleic acid
gDNA = genomic DNA
HER2 = human epidermal growth factor 2
LNA = locked nucleic acid
MDA = mean decrease accuracy
MDG = mean decrease Gini
miRNAs = microRNAs
mRNAs = messenger RNAs
NA = not assessed
Ns = non-significant
OOB = out-of-bag
RNA = ribonucleic acid
ROC = receiver operating characteristic

ACKNOWLEDGMENTS

We thank Olivier Dengis, Tiberio Sticca, Sonia El Guendi, Bouchra Boujemla, the GIGA-imagery-platform, the GIGA-immunohistology-platform, the team of medical oncologists and the Biothèque of CHU Liège.

GRANT SUPPORT

PF is a F.R.S.-FNRS PhD fellow. SW is a P.D.R.-FNRS PhD fellow. This work was supported by the French Community of Belgium, the Belgian Funds for Scientific Research (F.R.S.-FNRS), the F.R.S.-FNRS-Televie, CHU Liege (F.I.R.S) and the Region Wallone (Secance, BRAMIR).

CONFLICTS OF INTEREST

None.

REFERENCES

1. Jemal A, Bray F, Center, Melissa M., Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin. 2011; 61:69–90.
2. Bleyer A, Welch HG. Effect of Three Decades of Screening Mammography on Breast-Cancer Incidence. N Engl J Med. 2012; 367:1998–2005.
3. Kohn L, Mambourg F, Robays J, Albertijn M, Janssens J, Hoefnagels K, Ronsmans M, Jonckheer P. Informed choice on breast cancer screening: messages to support informed decision. Good Clinical Practice (GCP) Brussels: Belgian Health Care Knowledge Centre (KCE). 2014. KCE Reports 216. D/2014/10.273/03.
4. Eulalio A, Huntzinger E, Izaurralde E. Getting to the Root of miRNA-Mediated Gene Silencing. Cell. 2008; 132:9–14.
5. Friedman RC, Farh KK-H, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. 2009; 19:92–105.
6. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004; 116:281–97.
7. Ambros V. The functions of animal microRNAs. Nature. 2004; 431:350–5.
8. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, et al. MicroRNA expression profiles classify human cancers. Nature. 2005; 435:834–838.
9. Mathivanan S, Ji H, Simpson RJ. Exosomes: Extracellular organelles important in intercellular communication. Journal of Proteomics. 2010; 73:1907–20.
10. Kosaka N, Iguchi H, Ochiya T. Circulating microRNA in body fluid: a new potential biomarker for cancer diagnosis and prognosis. Cancer Science. 2010; 101:2087–92.
11. Mitchell PS, Parkin RK, Kroh EM, Fritz BR, Wyman SK, Pogosova-Agadjanyan EL, Peterson A, Noteboom J, O'Briant KC, Allen A, Lin DW, Urban N, Drescher CW, et al. Circulating microRNAs as stable blood-based markers for cancer detection. Proc Natl Acad Sci USA. 2008; 105:10513–8.
12. Chan M, Liaw CS, Ji SM, Tan HH, Wong CY, Thike AA, Tan PH, Ho GH, Lee AS-G. Identification of circulating microRNA signatures for breast cancer detection. Clin Cancer Res. 2013; 19:4477–87.
13. Cuk K, Zucknick M, Heil J, Madhavan D, Schott S, Turchinovich A, Arlt D, Rath M, Sohn C, Benner A, Junkermann H, Schneeweiss A, Burwinkel B. Circulating microRNAs in plasma as early detection markers for breast cancer. Int J Cancer. 2013; 132:1602–12.
14. Guo L, Zhao Y, Yang S, Cai M, Wu Q, Chen F. Genome-wide screen for aberrantly expressed miRNAs reveals miRNA profile signature in breast cancer. Mol Biol Rep. 2013; 40:2175–86.
15. Heneghan HM, Miller N, Kelly R, Newell J, Kerin MJ. Systemic miRNA-195 Differentiates Breast Cancer from Other Malignancies and Is a Potential Biomarker for Detecting Noninvasive and Early Stage Disease. The Oncologist. 2010; 15:673–82.

16. Hu Z, Dong J, Wang L-E, Ma H, Liu J, Zhao Y, Tang J, Chen X, Dai J, Wei Q, Zhang C, Shen H. Serum microRNA profiling and breast cancer risk: the use of miR-484/191 as endogenous controls. Carcinogenesis. 2012; 33:828–34.
17. Kodahl AR, Lyng MB, Binder H, Cold S, Gravgaard K, Knoop AS, Ditzel HJ. Novel circulating microRNA signature as a potential non-invasive multi-marker test in ER- positive early-stage breast cancer: A case control study. Mol Oncol. 2014; 8:874–83.
18. Leidner RS, Li L, Thompson CL. Dampening enthusiasm for circulating microRNA in breast cancer. PLoS One. 2013; 8:e57841.
19. Ng EKO, Li R, Shin VY, Jin HC, Leung CPH, Ma ESK, Pang R, Chua D, Chu KM, Law WL, Law SYK, Poon RTP, Kwang A. Circulating microRNAs as Specific Biomarkers for Breast Cancer Detection. PLoS One. 2013; 8:e53141.
20. Roth C, Rack B, Müller V, Janni W, Pantel K, Schwarzenbach H. Circulating microRNAs as blood-based markers for patients with primary and metastatic breast cancer. Breast Cancer Res. 2010; 12:R90.
21. Shen J, Hu Q, Schrauder M, Yan L, Wang D, Medico L, Guo Y, Yao S, Zhu Q, Liu B, Qin M, Beckmann MW, Fasching PA, et al. Circulating miR-148b and miR-133a as biomarkers for breast cancer detection. Oncotarget. 2014; 5:5284. doi:10.18632/oncotarget.2014.
22. Wang F, Zheng Z, Guo J, Ding X. Correlation and quantitation of microRNA aberrant expression in tissues and sera from patients with breast tumor. Gynecologic Oncology. 2010; 119:586–93.
23. Wu Q, Wang C, Lu Z, Guo L, Ge Q. Analysis of serum genome-wide microRNAs for breast cancer detection. Clinica Chimica Acta. 2012; 413:1058–65.
24. Zearo S, Kim E, Zhu Y, Zhao JT, Sidhu SB, Robinson BG, Soon PSH. MicroRNA-484 is more highly expressed in serum of early breast cancer patients compared to healthy volunteers. BMC Cancer. 2014; 14:200.
25. Zhao H, Shen J, Medico L, Wang D, Ambrosone CB, Liu S. A pilot study of circulating miRNAs as potential biomarkers of early stage breast cancer. PLoS One. 2010; 5:e13735.
26. Zhu J, Zheng Z, Wang J, Sun J, Wang P, Cheng X, Fu L, Zhang L, Wang Z, Li Z. Different miRNA expression profiles between human breast cancer tumors and serum. Front Genet. 2014; 5:149.
27. Stückrath I, Rack B, Janni W, Jäger B, Pantel K, Schwarzenbach H. Aberrant plasma levels of circulating miR-16, miR-107, miR-130a and miR-146a are associated with lymph node metastasis and receptor status of breast cancer patients. Oncotarget. 2015; 6:13387. doi:10.18632/oncotarget.3874.
28. Schrauder MG, Strick R, Schulz-Wendtland R. Circulating micro-RNAs as potential blood-based markers for early stage breast cancer detection. PLoS One. 2012; 7:e29770.
29. Cookson VJ, Bentley MA, Hogan B V, Horgan K, Hayward BE, Hazelwood LD, Hughes TA. Circulating microRNA profiles reflect the presence of breast tumours but not the profiles of microRNAs within the tumours. Cell Oncol. 2012; 35:301–8.
30. Breiman L. Random Forests. Machine Learning. 2001; 45: 5–32.
31. Geurts P, Fillet M, de Seny D, Meuwis M-A, Malaise M, Merville M-P, Wehenkel L. Proteomic mass spectra classification using decision tree based ensemble methods. Bioinformatics. 2005; 21:3138–45.
32. Pisano ED, Gatsonis C, Ph D, Hendrick E, Yaffe M, Baum JK, Acharyya S, Conant EF, Fajardo LL, Bassett L, Orsi CD, Jong R, Rebner M. Diagnostic Performance of Digital versus Film Mammography for Breast-Cancer Screening. N Engl J Med. 2005; 353:1773–83.
33. Gion M, Mione R, Leon AE, Lüftner D, Molina R, Possinger K, Robertson JF. CA27.29: a valuable marker for breast cancer management. A confirmatory multicentric study on 603 cases. Eur J Cancer. 2001; 37:355–363.
34. Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M. The benefits and harms of breast cancer screening: an independent review. British Journal of Cancer. 2013; 108:2205–40.
35. Yankaskas BC, Taplin SH, Ichikawa L, Geller BM, Rosenberg RD, Carney PA, Kerlikowske K, Ballard-Barbash R, Cutter GR, Barlow WE. Association between mammography timing and measures of screening performance in the United States. Radiology. 2005; 234:363–73.
36. Théberge I, Chang S-L, Vandal N, Daigle J-M, Guertin M-H, Pelletier E, Brisson J. Radiologist interpretive volume and breast cancer screening accuracy in a Canadian organized screening program. J Natl Cancer Inst. 2014; 106:djt461.
37. Cheng L, Doecke JD, Sharples RA, Villemagne VL, Fowler CJ, Rembach a, Martins RN, Rowe CC, Macaulay SL, Masters CL, Hill AF. Prognostic serum miRNA biomarkers associated with Alzheimer's disease shows concordance with neuropsychological and neuroimaging assessment. Mol Psychiatry. 2014:1–9.
38. Hemphill E, Lindsay J, Lee C, Măndoiu II. Feature selection and classifier performance on diverse bio-logical datasets. BMC Bioinformatics. 2014; 15:S4.
39. Mestdagh P, Van Vlierberghe P, De Weer A, Muth D, Westermann F, Speleman F, Vandesompele J. A novel and universal method for microRNA RT-qPCR data normalization. Genome Biol. 2009; 10:R64.
40. Janaki Ramaiah M, Lavanya A, Honarpisheh M, Zarea M, Bhadra U, Bhadra MP. MiR-15/16 complex targets p70S6 kinase 1 and controls cell proliferation in MDA- MB-231 breast cancer cells. Gene. 2014; 552:255–64.
41. Büssing I, Slack FJ, Großhans H. let-7 microRNAs in development, stem cells and cancer. Trends in Molecular Medicine. 2008; 14:400–9.

42. Martello G, Rosato A, Ferrari F, Manfrin A, Cordenonsi M, Dupont S, Enzo E, Guzzardo V, Rondina M, Spruce T, Parenti AR, Daidone MG, Bicciato S, et al. A MicroRNA targeting dicer for metastasis control. Cell. 2010; 141: 1195–207.

43. Olive V, Sabio E, Bennett MJ, De Jong CS, Biton A, McGann JC, Greaney SK, Sodir NM, Zhou AY, Balakrishnan A, Foth M, Luftig MA, Goga A, et al. A component of the mir-17–92 polycistronic oncomir promotes oncogene-dependent apoptosis. Elife. 2013; 2:e00822.

44. Tao S, He H, Chen Q, Yue W. GPER mediated estradiol reduces miR-148a to promote HLA-G expression in breast cancer. Biochem Biophys Res Commun. 2014; 451:74–8.

45. Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, Mueller R, Nolan T, Pfaffl MW, Shipley GL, Vandesompele J, Wittwer CT. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin Chem. 2009; 55:611–22.

46. Kroh EM, Parkin RK, Mitchell PS, Tewari M. Analysis of circulating microRNA biomarkers in plasma and serum using quantitative reverse transcription-PCR (qRT-PCR). Methods. 2010; 50:298–301.

47. Schmittgen TD, Livak KJ. Analyzing real-time PCR data by the comparative CT method. Nature Protocols. 2008; 3: 1101–8.

48. Blondal T, Jensby Nielsen S, Baker A, Andreasen D, Mouritzen P, Wrang Teilum M, Dahlsveen IK. Assessing sample and miRNA profile quality in serum and plasma or other biofluids. Methods. 2013; 59:S1–6.

49. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B. 1995; 57:289–300.

50. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002; 2:18–22.

51. Genuer R, Poggi J-M, Tuleau-Malot C. Variable selection using random forests. Pattern Recognition Letters. 2010; 31:2225–36.

52. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007; 23:2507–17.

53. Díaz-Uriarte R, de Andrés SA. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006; 7:3.

54. Ghattas B, Ben Ishak A. Sélection de variables pour la classification binaire en grande dimension: comparaisons et application aux données de biopuces. Journal de la société française de statistiques. 2008; 149:43–66.

4

Transcriptome wide analysis of natural antisense transcripts shows their potential role in breast cancer

> *Contraria sunt complementa.*
> *(Opposites are complementary.)*
>
> **— Niels Bohr**

Summary

In this study, we investigated the role of antisense lncRNAs, also called natural antisense transcripts (NATs) in breast cancer. NATs, which are one of several non-coding RNAs, are RNA sequences which are complementary and overlapping to those of protein-coding transcripts (PCT). Based on a cohort of 23 female ER+/HER2- breast cancer patients, RNA and DNA was extracted both from the tumor and from adjacent healthy tissue, allowing to use this pairing in the study design. We generated a list of 'putative' antisenses, including not only already known and studied antisenses, but also other non-coding transcripts overlapping protein-coding genes. Stranded RNA-Seq was performed, to be able to quantify the expression values of genes and transcripts (both protein-coding and antisenses).

Next, we evaluated how the balance between protein-coding genes and antisenses is disrupted in breast cancer tumors, both at the global scale, but also while considering pairs of protein-coding genes and antisenses sharing the same genomic location.

To identify specific antisenses playing a key role in the oncogenic process, three gene selection methods have been used, yielding lists of antisenses (and their overlapping protein-coding genes).

We assessed the enrichment of survival-associated genes in these lists, by using an external dataset of more than 1000 RNA-Seq samples of female breast cancer (TCGA).

Our results indicate not only the global disruption involving antisenses in the breast tumor pathology, but they also highlight a subset of protein-coding genes and antisenses whose active role should be further investigated and which might become therapeutic targets in the near future.

My personal contributions to this research project span across several domains: I took part in the study design as I was involved in all aspects of the project from its beginning, but I also performed most of the different analytic steps needed as soon as the sequencing reads were available on computing platforms.

I generated the list of pairs of protein-coding genes and antisenses, I set up the RNA-Seq computational analysis pipeline, by comparing and configuring all of the software used for the following tasks: sequencing reads mapping, gene quantification and annotation, differential expression analysis, differential correlation analysis, survival analysis. I implemented the varRatio analysis and defined the threshold used to select genes of interest in that particular method.

Finally, I took part in the writing of the manuscript.

Title of the article

Transcriptome wide analysis of natural antisense transcripts shows their potential role in breast cancer.

Authors and affiliations

Stephane Wenric[1,2,*], Sonia ElGuendi[1,*], Jean-Hubert Caberg[3], Warda Bezzaou[1], Corinne Fasquelle[1], Benoit Charloteaux[4], Latifa Karim[4], Benoit Hennuy[4], Pierre Frères[2], Joëlle Collignon[2], Meriem Boukerroucha[2], Hélène Schroeder[2], Fabrice Olivier[2], Véronique Jossa[5], Guy Jerusalem[2], Claire Josse[1,2,#] and Vincent Bours[1,3,#].

*Stephane Wenric & Sonia ElGuendi contributed equally to this work.

Guy Jerusalem, Claire Josse and Vincent Bours shared the senior authorship.

Institutional addresses:

1 : University of Liège, GIGA-Research, Laboratory of Human Genetics, Liège, Belgium

2 : University Hospital (CHU), Department of Medical Oncology, Liège, Belgium

3 : University Hospital (CHU), Center of Genetics, Liège , Belgium

4 : University of Liège, GIGA-Genomics Platform, Liège, Belgium

5: Clinique Saint-Vincent (CHC), Department of Pathology, Liège, Belgium

Running Title

Natural Antisens Transcripts in Breast Cancer

Keywords

Breast cancer; strand-specific RNA sequencing; lncRNA ; Natural Antisens Transcripts; antisense transcription.

Additional information.

Financial support

This work was supported by the French Community of Belgium, the Belgian Funds for Scientific Research (F.R.S.-FNRS), the F.R.S.-FNRS-Televie, CHU Liege (F.I.R.S) and the Region Wallonne [XSPRELTRIN]. SW is a P.D.R.-FNRS PhD fellow; SE is a F.R.I.A-FNRS PhD fellow. Computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Belgian Funds for Scientific Research (F.R.S.-FNRS) under Grant No. 2.5020.11

Corresponding author

Name : Claire Josse.

Mailing address: GIGA-Research, Laboratory of Human Genetics, Domaine Universitaire du Sart Tilman, University of Liège, 4000 Liège, Belgium.

Phone: +32 4 366 24 74 ; Fax: +32 4 366 81 46

E-mail : c.josse@chu.ulg.ac.be

Conflict of interest disclosure statement

The authors declare that they have no competing interests.

Notes

Word count : abstract = 298 words; manuscript text = 3888 words (.doc file) ; 3 figures (.png files) ; 4 tables (Table 1, 3 and 4 are included in the manuscript .doc file, Table 2 is supplied as a .xls file) ; 1 supplemental methods file (.doc file) ; 2 supplemental data files (.xls files).

Abstract

Non-coding RNAs (ncRNA) represent at least 1/5 of the mammalian transcript amount, and about 90% of the genome length is actively transcribed. Many ncRNAs have been demonstrated to play a role in cancer. Among them, natural antisense transcripts (NAT) are RNA sequences which are complementary and overlapping to those of protein-coding transcripts (PCT). NATs were punctually described as regulating gene expression, and are expected to act more frequently in *cis* than other ncRNAs that commonly function in *trans*. In this work, 22 breast cancers expressing estrogen receptors and their paired healthy tissues were analyzed by strand-specific RNA sequencing. To highlight the potential role of NATs in gene regulations occurring in breast cancer, three different gene extraction methods were used: differential expression analysis of NATs between tumor and healthy tissues, differential correlation analysis of paired NAT/PCT between tumor and healthy tissues, and NAT/PCT read count ratio variation between tumor and healthy tissues. Each of these methods yielded lists of NAT/PCT pairs that were demonstrated to be enriched in survival-associated genes on an independent cohort (TCGA). This work allows to highlight NAT lists that display a strong potential to affect the expression of genes involved in the breast cancer pathology.

Introduction

Over the past decade, RNA sequencing technology allowed to discover that the non-coding part of the genome represents around 1/5 of all transcript amount (1, 2). These non-coding RNAs (ncRNA) are less conserved between species than protein coding genes, but more than introns and random intergenic regions (3, 4). It is therefore likely that these non-coding transcripts have biological roles, which are being progressively deciphered, but still remain largely unknown. Around 30-50% of the protein coding gene loci are additionally expressing ncRNA in an opposite direction of the protein coding gene (4, 5). These naturally occurring antisense transcripts, called NATs, have been less studied than the other classes of ncRNA for technical reasons because their detection and quantification require to preserve information about the transcript originating strand along the sequencing process. Indeed, standard RNA sequencing and expression micro-array techniques require double-stranded cDNA synthesis, which erases RNA strand information, leading to an expression quantification that is the sum of the expressions of the coding RNA and its corresponding NAT. Commercial kits allowing to gather this information have only been made available recently, paving the way to high-throughput studies of stranded RNA sequencing.

NAT expression is subjected to the same expression regulation than other genes, but NATs accumulate preferentially into the nucleus - associated to chromatin - unlike coding mRNAs which are located into the cytoplasm. NATs are also found in other cellular compartments such as mitochondria (6, 7). NAT expression is described in many punctual examples to affect the activity of their sense or neighboring genes in biological events like cell differentiation and carcinogenesis, distinct molecular mechanisms being involved (8–11). NATs can regulate gene expression in *trans* or in *cis*. Given the fact that both the sense and antisense transcripts are transcribed from the same genomic region, it is expected that antisense transcripts behave more frequently in *cis* than other ncRNAs that commonly function in *trans* (11).This last feature means that NATs may regulate their protein coding gene counterpart at the same locus, which is of great interest from the therapeutic point of view: NATs may thus provide a unique entry point for therapeutic intervention on targeted genes by the use of ASO (antisense oligonucleotides) that are drugs already FDA-approved for several diseases (8, 12–14).

To date, a few studies have been performed at the whole transcriptome scale to investigate the role of NATs in the context of breast cancers. These studies have demonstrated that pairs of NAT/PCT are globally deregulated in this pathology (15–17). However, none of those studies compared the whole transcriptome of paired tumorous and healthy tissues of the same patients, with a technology that keeps the strand information of the transcripts. Yet, such an experimental design would be needed to explore if NAT tumor deregulations are cancer-specific, in order to define if they could play a role in the pathology. Here, we describe the results of such an experimental design, in a cohort of 22 ER+ breast cancer patients whose paired healthy and tumorous tissues were analyzed by stranded RNA sequencing.

This work allows clarifying the role played by NATs to regulate their protein coding gene counterpart on the same locus in the breast cancer pathology and to quantify to what extent this phenomenon is occurring. We first defined 3 lists of NAT/PCT pairs that are both deregulated between healthy/tumorous tissues and related to NAT-specific regulations. Next, we demonstrated that those lists are enriched with survival-associated genes. Finally, we established a list of breast cancer-related genes probably regulated by their NATs that could be targeted by ASO in a therapeutic objective.

Material and Methods

Additional information about methods used can be found in the Supplemental Methods 1

Ethical Statement

Tissues were obtained from the Liege University Biobank (N=12) and from the St Vincent Clinic of Rocourt (N=11), Belgium. This study was approved by the local institutional ethical board (2010/229). All aspects of the study comply with the Declaration of Helsinki. Patients of the Liege University Hospital were recruited on the basis of an opt-out methodology. Patients of the St Vincent Clinic of Rocourt were informed of the research work and provided written informed consent.

Patients and samples

This retrospective study was performed on 23 cryopreserved cancerous and adjacent healthy tissues from 23 women suffering from estrogen receptor expressing breast cancer. Samples were collected from 2010 to 2014. One patient was excluded because of poor strand-specificity of the RNA-Seq. The clinical and pathological parameters of the patients included in the final analysis were recorded and summarized in Table 1.

A summary of the experimental design is depicted in Figure 1.

Stranded RNA sequencing

RNA-Sequencing libraries for 22 breast tumors and paired adjacent tissues were constructed from 500 ng of total RNA, using the TruSeq® Stranded Total RNA kit and Ribo-Zero rRNA Removal kit. 2x100bp paired-end stranded RNA sequencing was performed on an Illumina HiSeq2000 apparatus, with a mean of 8.26E+09 bases sequenced for each sample (4 samples/line). Kits and apparatus were from Illumina, The Netherlands.

CGH array

Array comparative genomic hybridization was performed in the healthy and tumorous tissues of the 22 patients using the Agilent 60K microarray platform (G4827A-031746; Agilent Technologies, Santa Clara, CA, USA) according to the manufacturer's instructions.

Gene expression quantification by RNA-Sequencing

A quality control of the sequenced reads has been performed with the FastQC software (v. 0.11.2; https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Sequencing reads were mapped to the Human Genome hg19 GRCh37-75 (Ensembl) using the Star 2.4.1c software (18). Mapping quality was assessed with the Picard RnaSeqMetrics tool of the Picard software suite (v. 1.127; http://broadinstitute.github.io/picard/) using default parameters. The results are available in S1-Supplemental File 1. Read counts assignment was performed with the htseq-count tool of the HTSeq software suite (v. 0.6.1) (19). Data quality assessment was performed by computing the strand specificity (ratio of sequencing reads mapping to the incorrect strands) of all samples with htseq-count, leading to the removal of one patient with aberrant strand specificity. The DESeq2 software (v. 1.10.1) was used to normalize read counts, estimate dispersion, perform variance stabilizing transformation, and perform independent filtering by using the mean of normalized counts as filter statistic, thereby adjusting the filtering threshold at 33%, following the standard workflow (20). Variance-stabilization performance was assessed by producing MA-plots of log2 fold-change versus mean expression with DESeq2. A search for outliers was performed by computing Cook's distances for every gene and every sample with DESeq2 (S2/S3/S4/S5/S6-Supplemental File 1). A principal component plot was performed to assess the appropriate separation between the 2 sample classes (S7-Supplemental File 1). All aforementioned quality and performance measures yielded acceptable results for all remaining samples.

External dataset used for RNASeq gene expression comparison.

Gene expression variations were retrieved from de GEO Dataset GSE65216 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65216), which is an expression study by micro-array (Affymetrix) of the Maire's breast cancer cohort (21).

Definition of the protein-coding/antisense pairs

The list of pairs of protein-coding genes and their corresponding antisense has been generated based on the human genome assembly and gene annotation GRCh37 (release 75) from Ensembl (22). To be included in the list, pairs of genes have to fulfill the 3 following conditions: overlapping coordinates;

opposite strands; one of the two genes has to have the protein_coding biotype, while the other can have any of the following biotypes: *3prime_overlapping_ncrna*, *antisense*, *IG_C_pseudogene*, *IG_V_pseudogene*, *lincRNA*, *misc_RNA*, *polymorphic_pseudogene*, *processed_transcript*, *pseudogene*, *sense_intronic*, *sense_overlapping*, *snoRNA*, *snRNA*. The reasoning behind including all non-protein_coding biotypes as putative antisenses is that the Ensembl annotation of antisenses is limited to already validated antisenses, and it might thus miss out previously unknown antisenses.

NAT/PCT Gene list selection methods

1) DiffCor list: Differential correlation analysis between pairs of protein-coding and antisense transcripts was performed with the DGCA software (v. 1.0.1) (23). Pairs of protein coding/antisense genes whose correlation is significantly different between normal and tumor samples (adjusted p-value < 0.05) and whose correlation class differs between tumor and normal samples (ie. we removed the 0/0, +/+, -/- classes) have been selected.

2) NATDiffExp list: Differential expression analysis between all tumor and healthy samples was performed with the DESeq2 software (v. 1.10.1) for all genes, following the standard multi-factor workflow for paired samples. Pairs of protein coding/antisense genes where the antisense was significantly differentially expressed (adjusted p-value < 0.05) between normal and tumor samples have been selected.

3) varRatio list: The read counts ratio variation analysis had been performed as follows: let us define the variation of read counts ratio (varRatio) for each pair of NAT/PCT genes as

$$varRatio = \frac{tumoral\ read\ counts\ ratio}{normal\ read\ counts\ ratio}$$ where

$$tumoral\ read\ counts\ ratio = \frac{\sum tumor\ read\ counts_{antisense}}{\sum tumor\ read\ counts_{protein\ coding}}$$ and

$$normal\ read\ counts\ ratio = \frac{\sum normal\ read\ counts_{antisense}}{\sum normal\ read\ counts_{protein\ coding}}$$

Pairs of NAT/PCT genes corresponding to extreme values of the varRatio distribution have been selected by applying a threshold (mean ± standard deviation) to the log-transformed distribution of the

varRatios (S8-Supplemental File 1).

For all three gene list selection methods, pairs of genes where either the protein-coding or the antisense was expressed in less than 7 tumor samples or 7 healthy samples have been discarded.

Survival analysis

All protein-coding genes of the 3 gene lists have been tested for association with survival on an external dataset of 1066 RNA-Seq samples from the tumors of female breast cancer patients (Package R TCGA Biolinks; (24)). Association with survival was recorded when the p-value of a log-rank test was inferior to 0.05. The ratio of genes associated with survival in each list has been compared with the same ratio computed with a list of randomly selected protein-coding genes.

Results

The role played by NATs on the expression regulation of their corresponding coding transcripts is still largely unknown, as well as if this potential regulation could play a role in the ER+ breast cancer pathology. Our study experimental design used to answer this question at the whole genome scale is depicted in Figure 1. Twenty-two tumorous tissues of ER+ breast cancer patients, as well as their paired adjacent healthy tissues, were subjected to strand-specific RNA Sequencing and DNA copy number analysis by CGH array. The patient characteristics are summarized in Table 1. The cohort contains only tumors larger than 20mm, and is equally divided between luminal A and B sub-types, and between highly (Ki67>19%) and moderately (Ki67<19%) proliferating tumors. Most of them present a Bloom grade of 2 and 3.

RNA-Seq Validation

The combined analysis of the DNA copy number variations and modifications of RNA transcripts expression levels between tumor and healthy tissues validates our RNA-Seq analysis: the overall expression levels of coding gene transcripts inside genomic amplification or deletion newly acquired in the tumor were respectively increased and decreased, as expected (Figure 2 A).

Moreover, the gene expression changes between healthy and tumor tissues obtained in our RNA-Seq dataset were compared with those obtained in an independent dataset (GSE65216). Gene expression variations between 10 mammary healthy tissues and 22 ER+ tumors (11 luminal A and 11 luminal B) were extracted using Geo2R (25). The expression fold-change of genes that were found to be differentially expressed with an adjusted p-value <0.05 between healthy and tumoral tissues in both our and the GSE65216 datasets were compared, and present an average Spearman correlation coefficient of 0.613 (p-value<0.001). Moreover, 76.6% of those genes were differentially modulated in the same direction (Figure 2 B). At a smaller scale, some RT-qPCR experiments were performed on the RNA samples that were used for our RNA-Seq study to confirm variations of several transcripts expression between tumor and healthy tissues. Among others, the downregulation in tumors of the ADAMTS9 tumor suppressor and its NAT, ADAMTS9-AS2, were confirmed by RT-qPCR (Figure 2 C).

NAT expression accounts for 17% of their coding counterpart in healthy tissues and increases to 26% in tumors.

We next defined the pairs of protein-coding genes and their corresponding antisenses as detailed in the material and methods section. This list can be found in Supplemental File 2 and contains 9632 NAT/PCT pairs where at least one patient has a non-null expression for PC or AS, either in normal tissue or in tumor. As 19846 coding transcripts were expressed in mammary tissues, 49% of coding transcripts have a concomitant corresponding NAT expression. Globally, NAT read counts represent 17% of their coding counterparts in healthy tissues and 26% in tumors (Table 2), suggesting a global increase of the expression levels of NATs in mammary tumors. Moreover, the average read counts ratio between PCT/NAT gene pairs expressed simultaneously by a locus is 1544 in healthy tissues and 1013 in tumors (Supplemental File 2).

Unexpectedly, we observe that the fold change distributions of NATs present both in genomic amplifications and deletions are shifted towards higher fold-changes than the corresponding distributions based on protein-coding genes (Figure 3 A).

Based on the levels of expression in the different tissue types, we chose to focus on NAT/PCT pairs where both the PC and the AS were expressed in at least 7 out of the 22 patients, both in the tumor and the healthy tissue. This represents more than 60% of the total read counts of NAT/PCT pairs (Table 2). In this group of 4884 genes pairs, NAT expression is greatly increased and accounts for 31% of their coding counterpart in healthy tissues and 47.8% in tumor. This gene sub-group contains PC genes that display the stronger potential of being regulated by their NAT counterparts.

Positive correlation of expression between NAT and their corresponding PCT are created in tumorous tissues.

In order to highlight newly appearing or disappearing correlations between NATs and their corresponding PCTs in the tumor, differential correlation analysis between all pairs of PCTs and NATs was performed with the DGCA software (v. 1.0.1) (23). Complete results can be found in Supplemental File 2, showing a global positive correlation of expression between NATs and their corresponding PCTs: the mean Spearman correlation coefficients are 0.431 and 0.533 respectively in

healthy tissues and tumors when a significant correlation is observed (p-value < 0.05), namely in 20% of the NAT/PCT pairs. The number of significantly correlated NAT/PCT pairs does not differ in healthy and tumorous tissues. A positive mean z-score (0.460) is observed in case of significant differential correlation of NAT/PCT between tumor and healthy tissues (p-value < 0.05), meaning that globally, in the 11% of NAT/PCT pair correlations that are deregulated in tumors when compared to healthy tissues (567/4884 pairs), the correlations become more positive. The proportion of the different classes of differential correlations between tumors and healthy tissues is depicted in Figure 3 B, highlighting the fact that mainly positive correlations of expression between NATs and their corresponding PCTs are created, or lost in tumorous tissues. Very few inversions of correlation were observed. Some examples of NAT/PCT pairs presenting deregulated correlation of expression in the tumor tissue are presented in Table 3 showing that genes that are well known in the breast cancer field are displaying deregulated correlation of expression with their antisense transcripts.

Protein coding genes exhibiting deregulated corresponding NAT expression in tumors are preferentially related to survival of breast cancer patients

Three different gene selection methods were used to extract NAT/PCT pairs potentially related to the breast cancer pathology out of the 4884 pairs.

Firstly, the previously described DiffCor is based on the differential correlation of NAT/PCT read counts between healthy and tumors tissues. A list of NAT/PCT pairs whose correlation is significantly different between normal and tumor tissues (p-value < 0.05) and whose correlation class differs between tumor and normal tissues (ie. 0/0, +/+, -/- classes are removed) has been selected and contains 441 NAT/PCT pairs.

The second method is based on the differential expression of the NATs between tumors and healthy tissues. A list of 738 NAT/PCT pairs where the NATs were significantly differentially expressed (adjusted p-value < 0.05) between normal and tumor tissues has been determined.

The third method is based on the variation of the NAT/PCT ratio between healthy and tumor tissues, and allows to define a third list, called VarRatio, which contains NAT/PCT pairs with extreme values on the distribution of the VarRatio (S8 - Supplemental File 1). This VarRatio list can be subdivided in

leftmost and rightmost parts. Leftmost, 610 NAT/PCT pairs have a PCT/NAT ratio that decreases in the tumor either because of a down-regulation of the PCT expression, or an up-regulation of the NAT expression or both; and the reverse is observed for the 540 NAT/PCT pairs on the rightmost part of the distribution.

The 3 lists of genes can be found in the Supplemental File 1 (S9 to S12 – Supplemental File 1) and as expected, many NAT/PCT pairs appear in several of those lists, which contains a total of 1784 unique NAT/PCT pairs deregulated in breast cancers.

To ascertain if the protein coding genes of the DiffCor, NATDiffExp, and VarRatio lists are implicated in the breast cancer pathology, their association with survival was computed based on the RNA-Seq samples from the TCGA dataset. Each of these three lists present a proportion of genes associated with survival in the TCGA dataset greater than the proportion obtained in a list of randomly chosen protein coding genes (Table 4), meaning that PCT exhibiting deregulated corresponding NAT expression in tumor are enriched in genes related to survival of breast cancer patients. A Pearson's chi-squared test yielded statistically significant p-values for each of the 3 lists when compared to a list of randomly chosen protein coding genes.

72 cancer genes present a deregulated profile of NAT expression in breast cancer samples.

When the Cancer Gene Census list of genes from the COSMIC database (http://cancer.sanger.ac.uk/census) was compared with the content of our 3 lists of genes that are probably regulated by their NATs and implicated in the breast cancer pathology, 72 genes were found in common (S13 - Supplemental File 1). This list contains cancer genes that could be targeted by ASO, designed to interact with the corresponding NATs of those genes, in order to specifically regulate their expression.

Discussion

Breast cancer constitutes a public health problem: around 1out of 8 women will suffer from it during their lifetime in industrialized countries. The most frequent subtype is the estrogen receptor expressing breast cancer (ER+/HER2-), with 75% of occurrences. In case of primary disease, most patients are treated by surgery with or without radiotherapy and endocrine therapy. However, a large number of those cancers will suffer from a relapse and develop metastases - a major life-threatening event which is strongly associated with poor outcome - and require chemotherapy in case of symptomatic visceral disease. New therapies are thus searched as well as biomarkers that would give a better prediction of the risk of relapse. Our study explores the still new field of antisense transcription to define cancer gene lists, and will lead to further works to define predictive markers and/or tailor targeted treatment by antisense oligonucleotides (ASO) (14).

This is the first time that a whole transcriptome strand-specific RNA-Seq study focusing on the antisense transcription is performed in paired tumor and healthy mammary tissues. This experimental design allows to detect deregulation of NAT expression that occurs in cancer tissues, and to statistically connect them with the changes of the corresponding coding transcript expression. In particular, we revealed that many positive correlations between NATs and their PCT counterpart were appearing or fading in the tumor, suggesting newly acquired or lost regulations of the protein-coding transcripts in the cancerous tissues. Further functional molecular studies will however be needed to confirm the existence of such regulations of the PCTs by their NATs in the list of cancer gene pairs that were highlighted in this work.

The difference in fold-change between NATs located in genomic alterations and their coding counterparts, with NATs showing higher fold-changes, as well as the significant increase in NAT expression in tumors tend to indicate that NATs may be subject to a particular activating mechanism specific to tumors (Figure 3 A, Table 2).

Moreover, the association of these NATs with survival, which was evaluated through the use of their protein-coding counterparts as proxy in a large independent cohort, shows that the dysregulation observed within the landscape of NATs is not merely a random byproduct of the tumoral process.

Analyses were conducted to explore the relationship of the PC and AS genes with known prognostic factors, but no significant results were found. In the same way, enrichment analyses in pathways genes were conducted without any noticeable results.

Several studies have already explored the role of antisense transcription in breast cancer (15–17). Grinchuk *et al*. analyzed NAT/PCT pairs that are deregulated in breast cancer in order to define pathways in which they are particularly involved, and they defined NAT/PCT-based prognosis signatures. However Affymetrix microarray datasets were the support of this work, and this technique is not intrinsically strand-specific (15). Moreover, mammary normal and tumoral analyzed tissues were not matched. Balbin *et al*. performed a large scale, genome wide, stranded RNA-Seq study on 376 cancers samples with, among them, 60 primary breast cancers (16). But as in Grinchuck's study, tumorous tissues were not matched with healthy ones and in consequence, these studies did not explore, patient by patient, if the NAT/PC expression correlations were already present in the normal tissue, or if they were newly acquired in the tumor. This particularity in our experimental design did allow highlighting the fact that NAT expression is increased in tumorous tissues when compared to their coding counterpart. Indeed, the proportion of NAT reads counts in NAT/PCT pairs is globally raised in tumors when compared to healthy tissues.

As Balbin *et al*. have stated before, at any locus where PCT and NAT are simultaneously transcribed, the PCT is expressed around 1000 times more than the NAT, but we have additionally observed that this difference of expression is lower in tumors than in healthy tissues. We also measured that globally, 10% of the transcripts were coming from the antisense strand in healthy tissues and that this proportion is increased to 13% in tumors (8% were described by Balbin *et al*.). However, some patients present a much higher increase of NAT/PCT proportion in the tumor than others. This heterogeneity in NAT expression deregulation across patients could be used to stratify patients into subgroups of different prognosis. One limitation of our study is the small size and the short follow-up time of our cohort, which did not allow performing such type of analysis.

Our results also confirm the observation by Grinchuk *et al* that the expression correlations between NAT and PCT were different in tumors when compared to unrelated healthy tissues. We refined this

observation by using paired tissues of the same patient, and showed that globally these correlations become more significant and more positive in the tumors. Moreover, we highlighted the gene pairs where potential new PCT/NAT expression regulation occurs in cancerous tissues. After having performed a survival analysis with gene expression data from an external cohort (TCGA), it appears that these NAT/PCT gene pairs were also enriched in survival-associated genes, suggesting that the opposite strand transcription regulation might play a role in the breast cancer disease.

Therefore, our report opens a new field of investigation in cancer and indicates that NAT expression is often increased in cancer samples as compared to matched normal tissues. The relevance of this observation for coding gene expression, cancer biology, prognosis and treatment will need to be determined in specific and large cohorts of paired samples.

Acknowledgments

We thank the GIGA-genotranscriptomic-platform with special attention to Wouter Coppieters, the team of medical oncologists and data managers of the Medical Oncology Department, and the Biothèque of CHU Liège. We also thank Jérome Thiry, Bouchra Boujemla and Christophe Poulet.

References

1. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, et al.: RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **2007**; 316:1484–8.

2. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **2008**; 5:621–628.

3. Diederichs S: The four dimensions of noncoding RNA conservation. *Trends Genet* **2014**; 30:121–123.

4. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, et al.: Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **2009**; 458:223–7.

5. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, et al.: Antisense transcription in the mammalian transcriptome. *Science (80-)* **2005**; 309:1564–6.

6. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al.: The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* **2012**; 22:1775–89.

7. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al.: Landscape of transcription in human cells. *Nature* **2012**; 489:101–108.

8. Hung T, Wang Y, Lin MF, Koegel AK, Kotake Y, Grant GD, et al.: Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nat Genet* **2011**; 43:621–629.

9. Nishizawa M, Ikeya Y, Okumura T, Kimura T: Post-transcriptional inducible gene regulation by natural antisense RNA. *Front Biosci (Landmark Ed* **2015**; 20:1–36.

10. Khorkova O, Myers AJ, Hsiao J, Wahlestedt C: Natural antisense transcripts. *Hum Mol Genet* **2014**; 23:R54-63.

11. Pelechano V, Steinmetz LM: Gene regulation by antisense transcription. *Nat Rev Genet* **2013**; 14:880–93.

12. McGowan MP, Tardif J-C, Ceska R, Burgess LJ, Soran H, Gouni-Berthold I, et al.: Randomized, placebo-controlled trial of mipomersen in patients with severe hypercholesterolemia receiving maximally tolerated lipid-lowering therapy. *PLoS One* **2012**; 7:e49006.

13. Coelho T, Adams D, Silva A, Lozeron P, Hawkins PN, Mant T, et al.: Safety and efficacy of RNAi therapy for transthyretin amyloidosis. *N Engl J Med* **2013**; 369:819–29.

14. Wahlestedt C: Targeting long non-coding RNA to therapeutically upregulate gene expression. *Nat Rev Drug Discov* **2013**; 12:433–446.

15. Grinchuk O V, Motakis E, Yenamandra SP, Ow GS, Jenjaroenpun P, Tang Z, et al.: Sense-antisense gene-pairs in breast cancer and associated pathological pathways. *Oncotarget* **2015**; 6:42197–221.

16. Balbin OA, Malik R, Dhanasekaran SM, Prensner JR, Cao X, Wu Y-M, et al.: The landscape of antisense gene expression in human cancers. *Genome Res* **2015**; 25:1068–79.

17. Grigoriadis A, Oliver GR, Tanney A, Kendrick H, Smalley MJ, Jat P, et al.: Identification of differentially expressed sense and antisense transcript pairs in breast epithelial tissues. *BMC Genomics* **2009**; 10:324.

18. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al.: STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**; 29:15–21.

19. Anders S, Pyl PT, Huber W: HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **2015**; 31:166–9.

20. Love MI, Huber W, Anders S: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **2014**; 15:550.

21. Maire V, Nemati F, Richardson M, Vincent-Salomon A, Tesson B, Rigaill G, et al.: Polo-like kinase 1: a potential therapeutic option in combination with conventional chemotherapy for the management of patients with triple-negative breast cancer. *Cancer Res* **2013**; 73:813–823.

22. Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, et al.: The Ensembl gene annotation system. *Database (Oxford)* **2016**; 2016.

23. McKenzie AT, Katsyv I, Song W-M, Wang M, Zhang B: DGCA: A comprehensive R package for Differential Gene Correlation Analysis. *BMC Syst Biol* **2016**; 10:106.

24. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al.: TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* **2016**; 44:e71.

25. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al.: NCBI GEO:

archive for functional genomics data sets--update. *Nucleic Acids Res* **2013**; 41(Database issue):D991-5.

Figure Legends

Figure 1: Study workflow.

RNA and DNA were simultaneously extracted from 23 breast cancer ER+/HER2- tumors and their paired adjacent healthy tissues. Strand-specific paired-end RNA sequencing and comparative genomic hybridization (CGH) were performed. Quality control steps and RNA-Seq validation were performed and lead to the elimination of one patient because of a poor strand specificity of this sample. This strategy allowed to study differential expression of NATs and PCTs between tumors and healthy tissues, and to perform differential correlation analysis of NAT/PCT pairs. Three lists of genes with deregulated NAT expression in the tumors that could potentially affect their corresponding PC expression were extracted, and the coding genes they contain were subjected to survival analysis with an external cohort (TCGA).

Figure 2: Validation of RNA-Seq.

A. Fold change distributions of genes, as determined by RNA-Seq, which are located in somatic copy-number alterations (amplifications or deletions), as determined by CGH. The distinct curves show a clear effect of the copy-number alterations on the gene expression (fold-changes). As expected, genes located in genomic amplified regions in the tumor showed increased expression, and conversely. **B.** Gene expression fold-changes between tumor and healthy tissues obtained in the current RNA-Seq study were compared to those described in an external Affymetrix micro-array dataset GSE65216. This comparison showed a global concordance of the results, with a Spearman correlation coefficient of 0.613 (p-value < 0.001). **C.** The relative expression of the protein coding ADAMTS9 and its NAT, ADAMTS9-AS2, in tumors and healthy tissues obtained by RNA sequencing and by RT-PCR were compared. The RT-qPCR values were normalized by the expression of the endogenous control gene B2M. [p-value <0.001 (***)].

Figure 3: A. Fold-change distributions of NAT present in genomic amplifications and deletions. NAT expression values are shifted towards higher fold-changes than the corresponding distributions of their protein coding genes. **B. Schematic representation of the proportion of the different classes of differential correlations between tumors and healthy tissues.** Mainly positive correlations of

expression between NAT and their corresponding PCT are created, or lost in tumorous tissues. The numbers indicated in the graph are the numbers of NAT/PCT pairs in this category; +/+ = significant positive correlation between NAT and PCT exists in the adjacent healthy tissue and is conserved in the tumor; +/- = significant positive correlation between NAT and PCT exists in the adjacent healthy tissue and becomes negative in the tumor; +/0 = significant positive correlation between NAT and PCT exists in the adjacent healthy tissue and is lost in the tumor; -/- = significant negative correlation between NAT and PCT exists in the adjacent healthy tissue and is conserved in the tumor; -/+ = significant negative correlation between NAT and PCT exists in the adjacent healthy tissue and becomes positive in the tumor; -/0 = significant negative correlation between NAT and PCT exists in the adjacent healthy tissue and is lost in the tumor; 0/0 = no significant correlation between NAT and PCT exist, nor in the adjacent healthy tissue nor in the tumor.

Tables.

Table 1. Patient clinicopathological characteristics

Clinical Features	**Criteria**	**Patients**
Age (years)	Median	62.9 years
	Range	43-83 years
Tumor size (mm)	> 20	N=22
	< 20	N=0
Ki 67 (%)	< 19	N=11
	≥ 19	N=10
	Unknown	N=1
Histology	IDC + DCIS	N=7
	IDC	N=15
Bloom grade	I	N=4
	II	N=9
	III	N=9
T (x to 4)	1c	N=10
	2	N=11
	3	N=1
N (x to 3)	0	N=13
	1a	N=5
	1c	N=1
	2a	N=2
	3a	N=1
M (0 or 1)	0	N=22
	1	N=0
Molecular subtype	ER+/HER2-	N=22
	Luminal A	N=11
	Luminal B	N=11
Meantime follow-up	Month	43.36

Table 2: Distribution of the relative expression intensities of NAT and their corresponding PC among the 9632 NAT/PC pairs.

This study was focused on NAT/PCT pairs where both the PCT and the AS were expressed in at least 7 out of the 22 patients, both in the tumor and the healthy tissue. This group of 4884 gene pairs contains 60% of the total reads counts, and the NAT/PCT ratio expression is increased in tumors.

Table 3 : Examples of spearman correlations between NAT and cancer-associated coding genes that are altered in tumors when compared to adjacent healthy tissues.

Coding Gene Name	NAT_ID_Ensembl	NAT-Name	NAT/PCT spearman correlation adj healthy tissue	NAT/PCT cor. p-val adj healthy tissue	NAT/PCT spearman correlation tumor tissue	NAT/PCT cor. p-val tumor tissue	Classes
CAMTA1	ENSG00000225126	RP4-549F15.1	-0,09	6,82E-01	0,70	3,06E-04	0/+
CDKN2A	ENSG00000240498	CDKN2B-AS1	0,33	1,32E-01	0,75	5,17E-05	0/+
CTCF	ENSG00000237718	AC009095.4	-0,25	2,63E-01	0,58	4,96E-03	0/+
GNA11	ENSG00000267139	AC005262.3	-0,14	5,49E-01	0,48	2,28E-02	0/+
ACSL6	ENSG00000234758	AC034228.4	-0,35	1,07E-01	0,44	4,01E-02	0/+
HOXC11	ENSG00000228630	HOTAIR	0,08	7,29E-01	0,62	1,87E-03	0/+
ZFHX3	ENSG00000259901	RP5-991G20.4	-0,32	1,44E-01	0,50	1,71E-02	0/+
PPP6C	ENSG00000232630	PRPS1P2	-0,19	4,05E-01	0,45	3,41E-02	0/+
RAP1GDS1	ENSG00000214559	RP11-323J4.1	0,16	4,78E-01	0,91	3,99E-09	0/+
UBR5	ENSG00000246263	KB-431C1.4	0,16	4,65E-01	0,81	4,16E-06	0/+
UBR5	ENSG00000272037	KB-431C1.5	0,24	2,91E-01	0,73	9,92E-05	0/+
WT1	ENSG00000183242	WT1-AS	0,39	7,09E-02	0,89	4,46E-08	0/+
HNF1A	ENSG00000241388	HNF1A-AS1	0,02	9,18E-01	0,67	6,87E-04	0/+
AFF1	ENSG00000235043	TECRP1	0,76	3,74E-05	0,25	2,69E-01	+/0
BRCA1	ENSG00000240828	RPL21P4	0,63	1,87E-03	-0,04	8,57E-01	+/0
CAMTA1	ENSG00000269978	RP11-338N10.3	0,82	3,41E-06	0,14	5,36E-01	+/0
CEBPA	ENSG00000267727	CTD-2540B15.7	0,79	1,31E-05	-0,01	9,55E-01	+/0
MYH11	ENSG00000263065	AF001548.6	0,66	7,94E-04	0,03	8,95E-01	+/0
MSH6	ENSG00000224058	AC006509.7	0,57	5,89E-03	-0,10	6,58E-01	+/0
MSH2	ENSG00000236558	AC138655.6	0,76	3,74E-05	0,16	4,88E-01	+/0
HLF	ENSG00000263096	RP11-515O17.2	-0,49	1,95E-02	0,14	5,29E-01	-/0

MAP2K1	ENSG00000269999	CTD-3185P2.2	-0,52	1,26E-02	0,10	6,62E-01	-/0
GNAS	ENSG00000235590	GNAS-AS1	-0,60	3,22E-03	0,26	2,41E-01	-/0

1

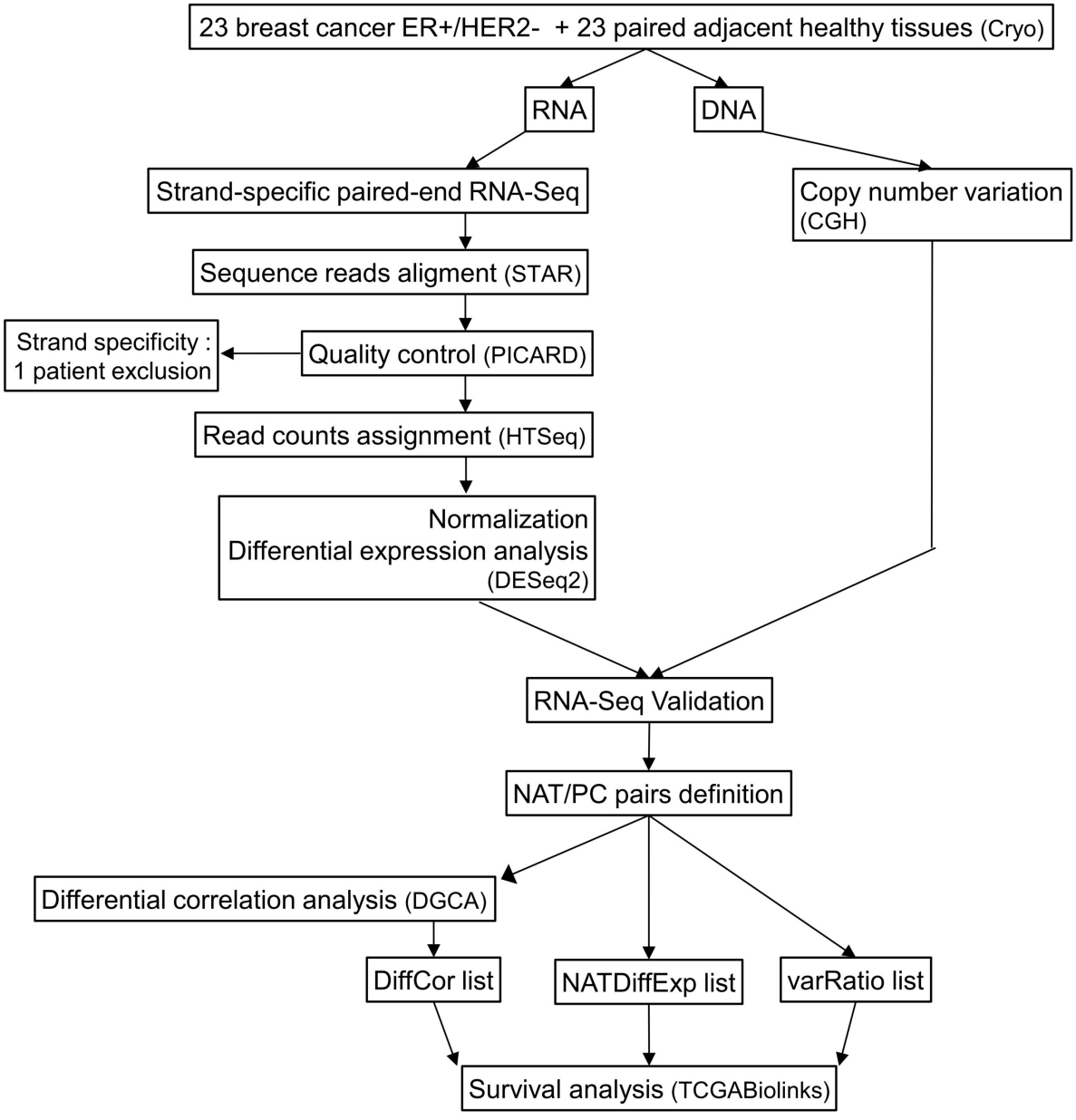
2

25

Table 4: PC genes exhibiting deregulated corresponding NAT expression in tumor are preferentially related to survival of breast cancer patients

Protein coding genes of the 3 gene lists DiffCor, DiffExp and VarRatio (left and right) were tested for association with survival by means of a TCGA RNA-Seq dataset of breast cancers. The percentage of genes associated with survival in each list has been compared with randomly selected protein coding genes.

PC Gene list	DiffCor	DiffExp	VaRatio Left	Varatio Right	Random
Nb genes in list	441	738	610	540	582
Nb genes also present in TCGA dataset	440	729	604	533	582
Nb genes w/ log-rank p-val <= 0.05	71	118	96	84	56.4 +- 7.5
Genes % w/ log-rank p-val <= 0.05	16,1%	16,2%	15,9%	15,8%	11.3 ± 1.6 %
Average log-rank p-val	0,392	0,387	0,388	0,391	0.489 ± 0.016


23 breast cancer ER+/HER2- + 23 paired adjacent healthy tissues (Cryo)
RNA
DNA
Strand-specific paired-end RNA-Seq
Copy number variation (CGH)
Sequence reads aligment (STAR)
Strand specificity : 1 patient exclusion
Quality control (PICARD)
Read counts assignment (HTSeq)
Normalization
Differential expression analysis (DESeq2)
RNA-Seq Validation
NAT/PC pairs definition
Differential correlation analysis (DGCA)
DiffCor list
NATDiffExp list
varRatio list
Survival analysis (TCGABiolinks)

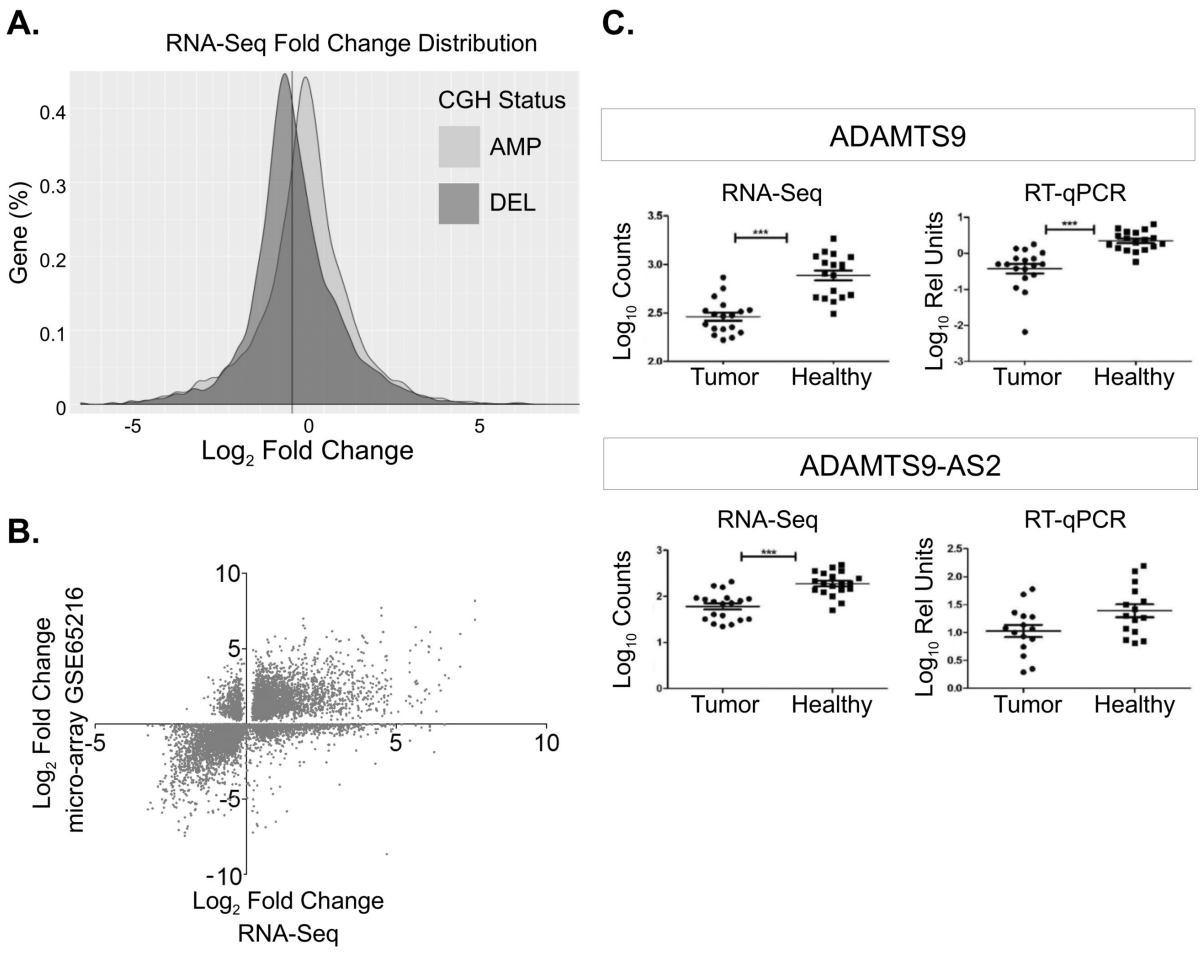
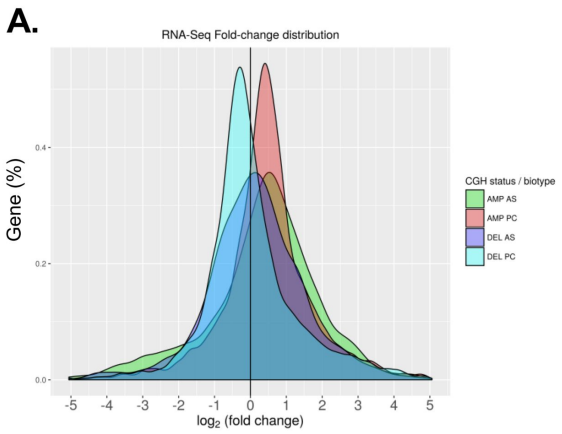
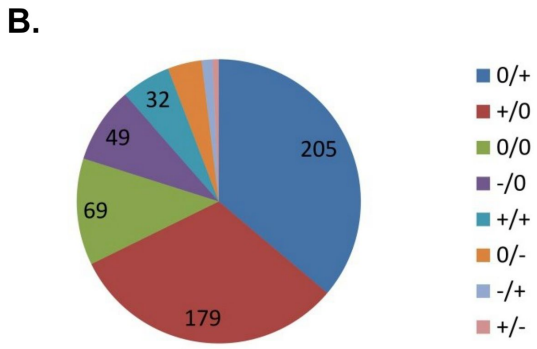

A.
RNA-Seq Fold Change Distribution
CGH Status
AMP
DEL
Gene (%)
0.4
0.3
0.2
0.1
0
-5
0
5
Log2 Fold Change
B.
Log2 Fold Change micro-array GSE65216
Log2 Fold Change RNA-Seq
10
5
-5
-10
C.
ADAMTS9
RNA-Seq
RT-qPCR
Log10 Counts
Log10 Rel Units
***
Tumor
Healthy
ADAMTS9-AS2
RNA-Seq
RT-qPCR
Log10 Counts
Log10 Rel Units
Tumor
Healthy

Table 2 : **Distribution of the relative expression intensities of NAT and their corresponding PC among the 9632 NAT/PC pairs.**

Nber NAT/PC pairs	transcript non-null expression in more than 7/22 patients				sum of read counts			
	PC norm tissue	PC tum	NAT norm tissue	NAT tum	PC norm tissue	PC tum	NAT norm tissue	NAT tum
4884	yes	yes	yes	yes	5.56E+07	4.73E+07	1.73E+07	2.26E+07
3282	yes	yes	no	no	3.51E+07	2.89E+07	1.62E+04	2.49E+04
944	yes	yes	no	yes	1.08E+07	1.02E+07	1.03E+04	2.12E+04
149	yes	yes	yes	no	2.06E+06	1.17E+06	2.61E+03	2.06E+03
149	no	no	yes	yes	4.81E+02	9.21E+02	7.11E+04	6.73E+04
89	no	no	no	no	2.51E+02	3.85E+02	5.65E+02	8.20E+02
50	no	yes	yes	yes	9.32E+02	1.77E+03	2.27E+04	4.29E+04
29	no	yes	no	no	5.16E+02	9.14E+02	1.35E+02	2.42E+02
28	no	no	no	yes	9.85E+01	1.64E+02	3.23E+02	5.69E+02
11	yes	no	yes	yes	2.42E+02	2.38E+02	2.02E+03	2.32E+03
7	no	yes	no	yes	9.24E+01	1.93E+02	9.25E+01	1.88E+02
5	yes	no	no	no	1.50E+02	1.43E+02	1.61E+01	3.80E+01
3	no	no	yes	no	3.26E+01	2.27E+01	4.59E+01	8.58E+01
1	yes	no	yes	no	5.67E+01	1.53E+01	1.00E+01	1.48E+01
1	no	yes	yes	no	4.84E+00	8.32E+00	1.53E+01	2.86E+01
0	yes	no	no	yes	0.00E+00	0.00E+00	0.00E+00	0.00E+00
Total 9632	NAT/PC pairs where at least one patient has a non-null expression in one of those four transcript types				1.04E+08	8.75E+07	1.74E+07	2.28E+07
					2.31E+08			


A.
RNA-Seq Fold-change distribution
Gene (%)
0.4
0.2
0.0
-5
-4
-3
-2
-1
0
1
2
3
4
5
log2 (fold change)
CGH status / biotype
AMP AS
AMP PC
DEL AS
DEL PC


B.
205
179
69
49
32
0/+
+/0
0/0
-/0
+/+
0/-
-/+
+/-

5 Discussion

> *Ideas are like rabbits. You get a couple and learn how to handle them, and pretty soon you have a dozen.*
>
> **— John Steinbeck**

Human cancer may be the disease of the decade, if not of the century.
Therefore, through its complexity, there is a large diversity of entry points by which we can chose to study it, a tally of aspects on which there is still a lot of work to be done.
Picking only a single aspect and focusing one's work on it might yield impressive results. Or it might not.
On the other hand, opting to focus on different problems might lead to the potential reuse and swapping of methods between related domains, it might lead to the emergence of a whole which is greater than the sum of its parts. Or it might not.

I chose to commit myself to multiple projects, hoping that by doing so I could disseminate techniques and procedures from various fields into associated domains.

Reflecting on this choice, one should be able to ask oneself:
"Have I been able to build bridges?"
"Have I used or developed methods in one field which might be useful in another one?"
"Is my work merely the addition of unrelated projects, or am I able to extract something more out of this enumeration?"

Normalization

There are obvious similarities between these projects.

At first, the need to normalize the data.

In **chapter 2**, the sequencing reads coverage (and its ratio) is used as a proxy for the genomic copy-number. We used the hypothesis that such coverage is constant along an exon (*ie.* exons are either in or out of a CNV, but no breakpoint intersects an exon), and we thus started from the coverage values for each of the 201 030 sequenced exons, instead of individual values for each of the 3 billion base pairs of the human genome.

In **chapters 3 and 4**, the unit considered is the "gene" (in a broad sense, as it includes non-protein coding transcribed units) and its expression, but not the single base either.

For these 3 studies, different normalization approaches are considered.

In **chapter 2**, instead of performing an intra-sample normalization, the coverage of each exon of one sample is normalized by the corresponding value from a reference sample. The hypothesis here is that the potential biases which would require a normalization are better addressed through the use of a reference sample which has been through the same protocol and which should thus suffer from exactly the same biases, essentially originating from the exome library preparation kit and its target capture protocol and the sequencing process.
In [168], Sathirapongsasuti et al. still divide the raw read counts by the total number of reads in the samples though, to mitigate the potential effect that an overall increase in local counts might have, due to the increase in total depth-of-coverage. This read count ratio is then adjusted so that the exome-wide median is 1, which is what we might expect if most of the chromosome locations of the tested sample do not harbor a CNV. One could wonder if such an adjustment is wise in the case of samples showing a large number of duplications.

In **chapter 3**, due to the nature of the variable measured, the very low variation between samples, and the goal of the study, the choice of normalization method is critical. That's the reason why 12 alternative normalizing methods were tested. Even though the normalization performed with the 50 miRNAs with the highest mean expression was the most stable one, it should be noted that some of the alternatives were very close in terms of performance, probably due to the fact that the variations measured are very small, both for the whole dataset and for the miRNAs used as normalizers. To be certain about our choice of normalizing method, the 12 other methods were not only compared with regards to their stability as a reference set, but we also performed the whole processes of miRNA signature identification and model building up to obtaining final AUC values.

One could argue that by testing only 13 normalizing methods, we did not explore all the solution space for this particular step of our study. However, given the

performance obtained with the alternative methods, and considering that these alternative methods present a good sampling of what one would expect to be the best performing normalizers, we are very confident about our choice of normalizing method.
Of course, given enough computational power and time, the best thing to do would be to sweep the whole solution space. However, as of today, and considering that this would add an additional layer of complexity to the already computationally intensive steps of our method, such an endeavour remains intractable.

In **chapter 4**, the DESeq2 software package was used for several steps of our analysis process, including normalizing read counts from our RNA-Seq dataset. However, although the data is not shown, we compared the final list of differentially expressed genes given by the DESeq2 pipeline with lists obtained with a different normalization method, namely reads per kilobase of transcript (RPKM), and a different statistical test (Mann-Whitney U). This comparison showed that 70% of differentially expressed genes were found in common.

The theoretical assumptions underlying these different methods tend to favor the DESeq2 normalization method and statistical test, as they integrate inter-sample relationships through the use of size factors and gene-specific normalization factors, through the use of a negative binomial distribution to model the read counts.

Unfortunately, it is the very large difference in terms of variables sampled in **chapter 3** (188 miRNAs) and **chapter 4** (> 60 000 "genes") which prevents the switching of normalizing methods between these studies. However, it would be interesting to see how normalizing methods from one study fare with data having the value range from the other study, as read counts in RNA-Seq datasets typicaly have extremely large ranges, while miRNAs levels assayed by RT-qPCR have a very small range.

Selection

The issue of feature selection, in a broad sense, is integral to both **chapter 3** and **chapter 4**.
In **chapter 3**, we needed to select a handful of miRNAs to build a "simple" model, while in **chapter 4**, the product of our gene selection process constitutes a result in itself.

These two settings are thus very different, however nothing theoretically prevents us from using the random forests based feature selection method developed in **chapter 3** to extract a list of genes of interest as we did in **chapter 4**, since our RNA-Seq

dataset is also a case/control dataset, albeit with far more features and less samples than our circulating miRNAs dataset; but since random forests are notoriously well suited for such datasets, it would be worth the try.

In the other way round, we might conjecture about the use of the gene selection methods used in **chapter 4** for the feature selection step of **chapter 3**. However, we can directly dismiss the gene selection methods involving pairs of genes (*ie.* the differential correlation based method, and the varRatio method), as our miRNA dataset doesn't have paired features.

Future developments

A potential improvement to [168], which might have an effect on the results shown in **chapter 2**, would be to perform the normalization and adjustment steps mentioned earlier chromosome by chromosome instead of doing them on the whole exome scale.

Additionally, since [168] performs circular binary segmentation, instead of starting from the coverage values for exons in chapter 2, we could start from coverage values for each single position covered or for a sliding windows of 10-20 bp, although it would add some time to the computation of the read count ratios.

These two additional developments could potentially improve the results obtained. However, the most interesting future development would probably be to verify if CNV profiles obtained with whole genome data perform better than profiles obtained with exome data, in terms of their distance metric computed with CGH-based profiles. Another compelling information would be to see if whole genome-based CNV profiles are closer to exome-based ones or to CGH-based ones: *ie.* do the differences between profiles arise because of the technique used to obtain said profile (sequencing vs. hybridization) or because of the percentage of the genome assayed (only the exons vs. complete chromosomes)?

In **chapter 3**, due to the characteristics of the data involved, the feature selection process could probably be fine-tuned for miRNAs, as they show high informational redundancy, since there is a significant number of miRNAs which are either highly correlated and/or which share the same target.

One could envision a method where features can be labelled (*eg.* *mir-34a* and *mir-34b* are part of the same family, *mir-141* and *mir-200a* are part of the same cluster, *mir-125b* and *mir-145* target genes which are part of the same pathway,

etc.); these labels would then be used to "guide" the internal variable selection steps happening in the random forests algorithm.
This extension of the feature selection process, to include biologically relevant information, if proven efficient, could be applied to other datasets making use of feature selection for machine learning.

Another, even more obvious, additional development would be to test other classification algorithms.
Even though random forests are theoretically well suited for our dataset, and even if previous comparison studies have been published saying that random forests give the best performance for a miRNA dataset with the same size as ours [135], testing other methods could yield useful results.

Another improvement which comes quickly to mind to improve **chapter 3** would be to add another kind of variable to the dataset. For example, we could look at cell-free DNA (cfDNA) to search for the presence or absence of punctual mutations or CNVs, which are known to be present in tumoral DNA; or we could measure different circulating metabolites by using mass spectrometry. This could add potentially complementary information while staying in a non-invasive blood/plasma based setting.

Finally, the same study design could be applied to other phenotypes. Another study is currently underway, trying to ascertain the possibility of predicting the response to neoadjuvant chemotherapy in breast cancer patients, based on the levels of their circulating miRNAs before treatment and at different time points.

We are also investigating the performance of our diagnostic model on a cohort of Rwandan breast cancer patients, and looking for potential ancestry-based differences in the levels of circulating miRNAs.

The continuity of **chapter 4** will soon be underway to validate the effect that several of the antisenses of our lists have on their protein-coding counterpart:

- With a cell line already showing an expression of the antisense transcript, we can artificially block this expression through the use of antisense oligonucleotides, thereby preventing its hybridization with the sense transcript. Then we would look at the expression of the coding transcript (with RT-qPCR or a Western blot).
- With a cell line which does not express the antisense transcript, we can transfect the cells with a plasmid allowing the expression of the antisense, and then look for changes of expression of the coding transcript.

Moreover, the diversity of possible direct extensions to **chapter 4** is large, as the information provided by RNA-Seq allows to look for far more than just gene expression.

Although our sample size is too small to perform an eQTL analysis, we do have access to the sequence of transcribed RNA from the tumor and from adjacent tissue, including single mutations. We could thus investigate these single mutations and look for the affected genes.

Recent developments in RNA-Seq bioinformatics also allow for the detection of gene fusion events. Gene fusions have been known for a long time to be key events in hematological malignancies, as they occur in 90% of all lymphomas and half of all leukemias but few fusion events have been associated to breast cancer to date [169–171]. We could thus investigate gene fusions in our RNA-Seq dataset.

Another domain of interest which can be explored through RNA-Seq is alternative splicing, where a single gene can yield different proteins based on the inclusion of the different exons, and which plays a role in breast cancer [172, 173].

Finally, alternative transcription start sites could also be explored.

Clinical Perspectives

In **chapter 2**, the detection of CNVs is an integral part of both the diagnosis and the prognosis for multiple myeloma patients. The technological disruptions arising in the field of next-generation sequencing may rapidly render existing techniques obsolete. However, the need to obtain accurate CNV profiles remains. Along with the constant technological evolution, there is thus a requirement for re-evaluation of the techniques used. Exome sequencing is one of such techniques envisioned in the clinical setting.

We showed that, with proper adjustments, it was possible to replace CGH by exome sequencing. Such a change would be beneficial as it will allow to capture both the CNV profile and the point mutations present in the tumoral cells of MM patients. However, we have to keep in mind that, even with adjustments and fine-tuning of complex methods, exome sequencing might rapidly lose the favor of researchers and clinicians alike, because of the decreasing cost of whole genome sequencing and the additional information and precision that it provides.

Our contributions are thus useful, but we have to keep in mind that the benefit they bring might be temporary.

In **chapter 3**, we directly challenged the status quo of mammography, by developing a non-invasive diagnostic test for breast cancer. The improvement that our test could bring varies with the age of the patients, as the performance of mammography increases with the age of the patients, while our test remains independent regarding this parameter. We have seen that several steps such as the normalization, or the reproducibility are critical when dealing with this kind of data, especially in a clinical setting.

Given this, our test could be especially useful for young patients, at risk for severe forms of breast cancer, as the performance of mammography is poor in young women. Even if the performances of our tests, in terms of sensibility and specificity, are not perfect, using it instead of mammography for young patients, or in conjunction with mammography for older patients would surely reduce the need for useless biopsies.

Additional work still has to be done to improve the performance of non-invasive tests (*eg.* by adding other variables to the dataset, as mentioned earlier), and to be able to challenge widely used screening methods such as mammography.

In **chapter 4**, through the exploration of the transcriptome world, our work has led to new fundamental findings which, at first sight, might not seem to directly impact the patient's environment, but it has also led to the identification of new potential therapeutic targets for breast cancer.

However, this work has to be refined, and these potential targets have to be further explored and validated through the use of alternative techniques, as mentioned earlier. These future developments might allow to improve the treatment of breast cancer patients.

Concluding remarks

In conclusion, the present work highlights how different bioinformatics techniques and methods have been developed and applied to three different problematics arising at the junction of oncology and the omics world. Despite their dissimilarities, these studies share some fundamentals, not only in their medical end goal, but also in the technical means to achieve it.

List of Figures

List of Tables

References

1. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144,** 646–674 (2011) (cit. on p. 1).
2. Vogelstein, B., Papadopoulos, N., Velculescu, V. E., *et al.* Cancer Genome Landscapes. *Science* **339,** 1546–1558 (2013) (cit. on pp. 2–4).
3. Mendelsohn, J., Gray, J. W., Howley, P. M., Israel, M. A. & Thompson, C. (B. *The molecular basis of cancer* (Elsevier, 2016) (cit. on pp. 2, 6).
4. Moasser, M. M. The oncogene HER2: its signaling and transforming functions and its role in human cancer pathogenesis. *Oncogene* **26,** 6469–87 (Oct. 2007) (cit. on p. 2).
5. Hainaut, P. & Hollstein, M. p53 and human cancer: the first ten thousand mutations. *Advances in cancer research* **77,** 81–137 (2000) (cit. on p. 2).
6. Yates, L. R. & Campbell, P. J. Evolution of the cancer genome. *Nature Reviews Genetics* **28,** 155–163 (Oct. 2012) (cit. on p. 4).
7. Walsh, T. & King, M.-C. Ten Genes for Inherited Breast Cancer. *Cancer Cell* **11,** 103–105 (Feb. 2007) (cit. on pp. 3, 5).
8. Institute, N. .-.-. N. C. Statistics at a Glance: The Burden of Cancer Worldwide. *Cancer Statistics (updated March 2016)* (2016) (cit. on p. 4).
9. Du Cancer – Stiftung Krebsregister, S. K. .-. F. R. Cancer Burden in Belgium. *Belgian Cancer Registry (revised edition April 2016)* (2015) (cit. on pp. 4, 5).
10. Ferlay, J., Soerjomataram, I., Dikshit, R., *et al.* Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer* **136,** E359–E386 (Mar. 2015) (cit. on p. 5).
11. N, H., AM, N., M, K., *et al.* SEER Cancer Statistics Review, 1975-2013. *http://seer.cancer.gov/csr/1975_2013/, based on November 2015 SEER data submission, posted to the SEER web site, April 2016.* (2015) (cit. on p. 5).
12. Alexander, D. D., Mink, P. J., Adami, H.-O., *et al.* Multiple myeloma: A review of the epidemiologic literature. *International Journal of Cancer* **120,** 40–61 (2007) (cit. on p. 5).

13. Kyle, R. A. & Rajkumar, S. V. Multiple Myeloma. *New England Journal of Medicine* **351,** 1860–1873 (Oct. 2004) (cit. on pp. 5, 9).

14. Clark, A. S. & Domchek, S. M. Clinical Management of Hereditary Breast Cancer Syndromes. *Journal of Mammary Gland Biology and Neoplasia* **16,** 17–25 (Apr. 2011) (cit. on p. 5).

15. Collaborative Group on Hormonal Factors in Breast Cancer. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58 209 women with breast cancer and 101 986 women without the disease. *The Lancet* **358,** 1389–1399 (Oct. 2001) (cit. on p. 5).

16. Antoniou, A., Pharoah, P. D. P., Narod, S., *et al.* Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. *American journal of human genetics* **72,** 1117–30 (May 2003) (cit. on p. 5).

17. Couch, F. J., Nathanson, K. L. & Offit, K. Two Decades After BRCA: Setting Paradigms in Personalized Cancer Care and Prevention. *Science* **343,** 1466–1470 (Mar. 2014) (cit. on p. 5).

18. Perou, C. M., Sørlie, T., Eisen, M. B., *et al.* Molecular portraits of human breast tumours. *Nature* **406,** 747–752 (Aug. 2000) (cit. on p. 6).

19. Weigelt, B., Mackay, A., A'hern, R., *et al.* Breast cancer molecular profiling with single sample predictors: a retrospective analysis. *The Lancet Oncology* **11,** 339–349 (Apr. 2010) (cit. on p. 6).

20. Wang, J., Sang, D., Xu, B., *et al.* Value of Breast Cancer Molecular Subtypes and Ki67 Expression for the Prediction of Efficacy and Prognosis of Neoadjuvant Chemotherapy in a Chinese Population. *Medicine* **95,** e3518 (May 2016) (cit. on p. 6).

21. Bergsagel, P. L. & Kuehl, W. M. Molecular pathogenesis and a consequent classification of multiple myeloma. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **23,** 6333–8 (Sept. 2005) (cit. on pp. 6–8).

22. Landgren, O., Kyle, R. A., Pfeiffer, R. M., *et al.* Monoclonal gammopathy of undetermined significance (MGUS) consistently precedes multiple myeloma: a prospective study. *Blood* **113,** 5412–5417 (May 2009) (cit. on p. 6).

23. Fonseca, R., Debes-Marun, C. S., Picken, E. B., *et al.* The recurrent IgH translocations are highly associated with nonhyperdiploid variant multiple myeloma. *Blood* **102** (2003) (cit. on p. 7).

24. Fonseca, R., Bergsagel, P. L., Drach, J., *et al.* International Myeloma Working Group molecular classification of multiple myeloma: spotlight review. *Leukemia* **23,** 2210–2221 (Dec. 2009) (cit. on p. 7).

25. Boyd, K. D., Ross, F. M., Walker, B. A., *et al.* Mapping of Chromosome 1p Deletions in Myeloma Identifies FAM46C at 1p12 and CDKN2C at 1p32.3 as Being Genes in Regions Associated with Adverse Survival. *Clinical Cancer Research* **17,** 7776–7784 (Dec. 2011) (cit. on p. 8).

26. Munshi, N. C. & Avet-Loiseau, H. Genomics in Multiple Myeloma. *Clinical Cancer Research* **17,** 1234–1242 (Mar. 2011) (cit. on pp. 8, 14).

27. Rajkumar, S. V. Multiple myeloma: 2016 update on diagnosis, risk-stratification, and management. *American Journal of Hematology* **91,** 719–734 (July 2016) (cit. on p. 8).

28. Lohr, J. G., Stojanov, P., Carter, S. L., *et al.* Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. *Cancer cell* **25,** 91–101 (Jan. 2014) (cit. on p. 8).

29. Perry, N., Broeders, M., de Wolf, C., *et al.* European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition–summary document. *Annals of Oncology* **19,** 614–622 (Oct. 2007) (cit. on p. 8).

30. Kolb, T. M., Lichy, J. & Newhouse, J. H. Comparison of the Performance of Screening Mammography, Physical Examination, and Breast US and Evaluation of Factors that Influence Them: An Analysis of 27,825 Patient Evaluations. *Radiology* **225,** 165–175 (Oct. 2002) (cit. on p. 9).

31. Rosenberg, R. D., Yankaskas, B. C., Abraham, L. A., *et al.* Performance Benchmarks for Screening Mammography. *Radiology* **241,** 55–66 (Oct. 2006) (cit. on p. 9).

32. Theberge, I., Chang, S.-L., Vandal, N., *et al.* Radiologist Interpretive Volume and Breast Cancer Screening Accuracy in a Canadian Organized Screening Program. *JNCI Journal of the National Cancer Institute* **106,** djt461–djt461 (Mar. 2014) (cit. on p. 9).

33. Yaffe, M. J. & Mainprize, J. G. Risk of Radiation-induced Breast Cancer from Mammographic Screening. *Radiology* **258,** 98–105 (Jan. 2011) (cit. on p. 9).

34. Welch, H. G., Prorok, P. C., O'Malley, A. J. & Kramer, B. S. Breast-Cancer Tumor Size, Overdiagnosis, and Mammography Screening Effectiveness. *New England Journal of Medicine* **375,** 1438–1447 (Oct. 2016) (cit. on p. 9).

35. Melton, L. J., Kyle, R. A., Achenbach, S. J., Oberg, A. L. & Rajkumar, S. V. Fracture Risk With Multiple Myeloma: A Population-Based Study. *Journal of Bone and Mineral Research* **20,** 487–493 (Nov. 2004) (cit. on p. 9).

36. Goldschmidt, H., Lannert, H., Bommer, J. & Ho, A. D. Multiple myeloma and renal failure. *Nephrology Dialysis Transplantation* **15,** 301–304 (Mar. 2000) (cit. on p. 9).

37. Kyle, R. A. & Rajkumar, S. V. Criteria for diagnosis, staging, risk stratification and response assessment of multiple myeloma. *Leukemia* **23,** 3–9 (Jan. 2009) (cit. on p. 9).

38. Iafrate, A. J., Feuk, L., Rivera, M. N., *et al.* Detection of large-scale variation in the human genome. *Nature Genetics* **36,** 949–951 (Sept. 2004) (cit. on p. 11).

39. Zarrei, M., MacDonald, J. R., Merico, D. & Scherer, S. W. A copy number variation map of the human genome. *Nature Reviews Genetics* **16,** 172–183 (Feb. 2015) (cit. on pp. 11, 12, 14, 16).

40. Lejeune, J., Turpin, R. & Gautier, M. Mongolism; a chromosomal disease (trisomy). *Bulletin de l'Academie nationale de medecine* **143,** 256–65 (cit. on p. 11).

41. Ford, C. E., Jones, K. W., Miller, O. J., *et al.* The chromosomes in a patient showing both mongolism and the Klinefelter syndrome. *Lancet* **1,** 709–10 (Apr. 1959) (cit. on p. 11).

42. Freeman, J. L., Perry, G. H., Feuk, L., *et al.* Copy number variation: new insights in genome diversity. *Genome research* **16,** 949–61 (Aug. 2006) (cit. on p. 11).

43. Beroukhim, R., Mermel, C. H., Porter, D., *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463,** 899–905 (Feb. 2010) (cit. on pp. 11–13).

44. Kuiper, R. P., Ligtenberg, M. J., Hoogerbrugge, N. & Geurts van Kessel, A. Germline copy number variation and cancer risk. *Current Opinion in Genetics & Development* **20,** 282–289 (June 2010) (cit. on p. 13).

45. Krepischi, A. C. V., Pearson, P. L. & Rosenberg, C. Germline copy number variations and cancer predisposition. *Future Oncology* **8,** 441–450 (Apr. 2012) (cit. on p. 13).

46. Bergamaschi, A., Kim, Y. H., Wang, P., *et al.* Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes, Chromosomes and Cancer* **45,** 1033–1040 (Nov. 2006) (cit. on p. 13).

47. Cancer Genome Atlas Network, T. C. G. A. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (Oct. 2012) (cit. on p. 13).

48. Fonseca, R., Blood, E., Rue, M., *et al.* Clinical and biologic implications of recurrent genomic aberrations in myeloma. *Blood* **101** (2003) (cit. on p. 14).

49. Avet-Loiseau, H., Attal, M., Moreau, P., *et al.* Genetic abnormalities and survival in multiple myeloma: the experience of the Intergroupe Francophone du Myelome. *Blood* **109,** 3489–3495 (Apr. 2007) (cit. on p. 14).

50. Neben, K., Jauch, A., Hielscher, T., *et al.* Progression in smoldering myeloma is independently determined by the chromosomal abnormalities del(17p), t(4;14), gain 1q, hyperdiploidy, and tumor load. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **31,** 4325–32 (Dec. 2013) (cit. on p. 14).

51. Cantsilieris, S., Baird, P. N. & White, S. J. Molecular methods for genotyping complex copy number polymorphisms. *Genomics* **101,** 86–93 (2013) (cit. on pp. 14, 15).

52. Carson, A. R., Feuk, L., Mohammed, M. & Scherer, S. W. Strategies for the detection of copy number and other structural variants in the human genome. *Human genomics* **2,** 403–14 (June 2006) (cit. on p. 15).

53. Teo, S. M., Pawitan, Y., Ku, C. S., Chia, K. S. & Salim, A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics* **28,** 2711–2718 (Nov. 2012) (cit. on p. 15).

54. Xie, C. & Tammi, M. T. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* **10,** 80 (Mar. 2009) (cit. on p. 16).

55. Plagnol, V., Curtis, J., Epstein, M., *et al.* A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* **28,** 2747–2754 (Nov. 2012) (cit. on p. 15).

56. Winchester, L., Yau, C. & Ragoussis, J. Comparing CNV detection methods for SNP arrays. *Briefings in Functional Genomics and Proteomics* **8,** 353–366 (2009) (cit. on p. 16).

57. Chen, K., Wallis, J. W., McLellan, M. D., *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods* **6,** 677–81 (Sept. 2009) (cit. on p. 16).

58. Duan, J., Zhang, J.-G., Deng, H.-W. & Wang, Y.-P. Comparative studies of copy number variation detection methods for next-generation sequencing technologies. *PloS one* **8,** e59128 (2013) (cit. on p. 17).

59. Wenric, S., Sticca, T., Caberg, J. H., *et al.* Exome copy number variation detection: Use of a pool of unrelated healthy tissue as reference sample. *Genetic Epidemiology,* 35–40 (2016) (cit. on p. 17).

60. Lee, R. C., Feinbaum, R. L. & Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75,** 843–54 (Dec. 1993) (cit. on p. 19).

61. Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. Identification of Novel Genes Coding for Small Expressed RNAs. *Science* **294** (2001) (cit. on p. 19).

62. Frères, P., Wenric, S., Boukerroucha, M., *et al.* Circulating microRNA-based screening tool for breast cancer. *Oncotarget* **7,** 5416–28 (Feb. 2016) (cit. on pp. 19, 29, 30).

63. Ha, M. & Kim, V. N. Regulation of microRNA biogenesis. *Nature Reviews Molecular Cell Biology* **15,** 509–524 (July 2014) (cit. on pp. 19, 20).

64. Jeffrey, S. S. Cancer biomarker profiling with microRNAs. *Nature Biotechnology* **26,** 400–401 (Apr. 2008) (cit. on p. 19).

65. Griffiths-Jones, S. The microRNA Registry. *Nucleic Acids Research* **32,** 109D–111 (Jan. 2004) (cit. on p. 19).

66. Farazi, T. A., Hoell, J. I., Morozov, P. & Tuschl, T. in *MicroRNA Cancer Regulation* 1–20 (Springer Netherlands, 2013) (cit. on pp. 20–22, 24).

67. Calin, G. A., Dumitru, C. D., Shimizu, M., *et al.* Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences of the United States of America* **99,** 15524–9 (Nov. 2002) (cit. on p. 20).

68. Lu, J., Getz, G., Miska, E. A., *et al.* MicroRNA expression profiles classify human cancers. *Nature* **435,** 834–838 (June 2005) (cit. on p. 20).

69. Zhang, L., Huang, J., Yang, N., *et al.* microRNAs exhibit high frequency genomic alterations in human cancer. *Proceedings of the National Academy of Sciences of the United States of America* **103,** 9136–41 (June 2006) (cit. on p. 20).

70. Peng, Y. & Croce, C. M. The role of MicroRNAs in human cancer. *Signal Transduction and Targeted Therapy* **1,** 15004 (2016) (cit. on pp. 20, 23).

71. Hermeking, H. The miR-34 family in cancer and apoptosis. *Cell Death and Differentiation* **17,** 193–199 (Feb. 2010) (cit. on pp. 20, 23).

72. Karube, Y., Tanaka, H., Osada, H., *et al.* Reduced expression of Dicer associated with poor prognosis in lung cancer patients. *Cancer Science* **96,** 111–115 (Feb. 2005) (cit. on p. 21).

73. Kong, Y. W., Ferland-McCollough, D., Jackson, T. J. & Bushell, M. MicroRNAs in cancer management. *The Lancet Oncology* **13,** e249–e258 (2012) (cit. on pp. 21–23).

74. Peurala, H., Greco, D., Heikkinen, T., *et al.* MiR-34a Expression Has an Effect for Lower Risk of Metastasis and Associates with Expression Patterns Predicting Clinical Outcome in Breast Cancer. *PLoS ONE* **6** (ed Yeudall, A.) e26122 (Nov. 2011) (cit. on p. 22).

75. Hurst, D. R., Edmonds, M. D. & Welch, D. R. Metastamir: The Field of Metastasis-Regulatory microRNA Is Spreading. *Cancer Research* **69,** 7495–7498 (Oct. 2009) (cit. on p. 22).

76. Camps, C., Buffa, F. M., Colella, S., *et al.* hsa-miR-210 Is Induced by Hypoxia and Is an Independent Prognostic Factor in Breast Cancer. *Clinical Cancer Research* **14,** 1340–1348 (Mar. 2008) (cit. on p. 22).

77. Yan, L.-X., Huang, X.-F., Shao, Q., *et al.* MicroRNA miR-21 overexpression in human breast cancer is associated with advanced clinical stage, lymph node metastasis and patient poor prognosis. *RNA (New York, N.Y.)* **14,** 2348–60 (Nov. 2008) (cit. on p. 22).

78. Chuang, J. C. & Jones, P. A. Epigenetics and MicroRNAs. *Pediatric Research* **61,** 24R–29R (May 2007) (cit. on p. 22).

79. Si, M.-L., Zhu, S., Wu, H., *et al.* miR-21-mediated tumor growth. *Oncogene* **26,** 2799–2803 (Apr. 2007) (cit. on p. 22).

80. Kong, W., He, L., Richards, E. J., *et al.* Upregulation of miRNA-155 promotes tumour angiogenesis by targeting VHL and is associated with poor prognosis and triple-negative breast cancer. *Oncogene* **33,** 679–689 (Feb. 2014) (cit. on p. 23).

81. Chang, S., Wang, R.-H., Akagi, K., *et al.* Tumor suppressor BRCA1 epigenetically controls oncogenic microRNA-155. *Nature Medicine* **17,** 1275–1282 (Sept. 2011) (cit. on p. 23).

82. Bahena-Ocampo, I., Espinosa, M., Ceballos-Cancino, G., *et al.* miR-10b expression in breast cancer stem cells supports self-renewal through negative PTEN regulation and sustained AKT activation. *EMBO reports* **17,** 648–658 (May 2016) (cit. on p. 23).

83. Knirsh, R., Ben-Dror, I., Modai, S., *et al.* MicroRNA 10b promotes abnormal expression of the proto-oncogene c-Jun in metastatic breast cancer cells. *Oncotarget* **7,** 59932–59944 (Sept. 2016) (cit. on p. 23).

84. Rivas, M. A., Venturutti, L., Huang, Y.-W., *et al.* Downregulation of the tumor-suppressor miR-16 via progestin-mediated oncogenic signaling contributes to breast cancer development. *Breast cancer research : BCR* **14,** R77 (2012) (cit. on p. 23).

85. Hu, X., Guo, J., Zheng, L., *et al.* The heterochronic microRNA let-7 inhibits cell motility by regulating the genes in the actin cytoskeleton pathway in breast cancer. *Molecular cancer research : MCR* **11,** 240–50 (2013) (cit. on p. 23).

86. Krzeszinski, J. Y., Wei, W., Huynh, H., *et al.* miR-34a blocks osteoporosis and bone metastasis by inhibiting osteoclastogenesis and Tgif2. *Nature* **512,** 431–435 (June 2014) (cit. on p. 23).

87. Ferracin, M., Bassi, C., Pedriali, M., *et al.* miR-125b targets erythropoietin and its receptor and their expression correlates with metastatic potential and ERBB2 / HER2 expression, 1–10 (2013) (cit. on p. 23).

88. Tsang, J. C., Dennis Lo, Y., Tang, K., *et al.* Circulating nucleic acids in plasma/serum. *Pathology* **39,** 197–207 (Apr. 2007) (cit. on p. 23).

89. Lawrie, C. H., Gal, S., Dunlop, H. M., *et al.* Detection of elevated levels of tumour-associated microRNAs in serum of patients with diffuse large B-cell lymphoma. *British Journal of Haematology* **141,** 672–675 (June 2008) (cit. on p. 24).

90. Chim, S. S., Shing, T. K., Hung, E. C., *et al.* Detection and Characterization of Placental MicroRNAs in Maternal Plasma. *Clinical Chemistry* **54,** 482–490 (Mar. 2008) (cit. on p. 24).

91. Turchinovich, A., Weiz, L. & Burwinkel, B. Extracellular miRNAs: the mystery of their origin and function. *Trends in Biochemical Sciences* **37,** 460–465 (2012) (cit. on p. 24).

92. Meyer, S. U., Pfaffl, M. W. & Ulbrich, S. E. Normalization strategies for microRNA profiling experiments: A 'normal' way to a hidden layer of complexity? *Biotechnology Letters* **32,** 1777–1788 (2010) (cit. on pp. 24, 25).

93. Mestdagh, P., Van Vlierberghe, P., De Weer, A., *et al.* A novel and universal method for microRNA RT-qPCR data normalization. *Genome Biol* **10,** R64 (2009) (cit. on p. 25).

94. Motawi, T. K., Rizk, S. M., Ibrahim, T. M. & Ibrahim, I. A.-R. Circulating microRNAs, miR-92a, miR-100 and miR-143, as non-invasive biomarkers for bladder cancer diagnosis. *Cell Biochemistry and Function* **34,** 142–148 (Apr. 2016) (cit. on p. 25).

95. Wang, W.-T., Zhao, Y.-N., Han, B.-W., Hong, S.-J. & Chen, Y.-Q. Circulating MicroRNAs Identified in a Genome-Wide Serum MicroRNA Expression Analysis as Noninvasive Biomarkers for Endometriosis. *The Journal of Clinical Endocrinology & Metabolism* **98,** 281–289 (Jan. 2013) (cit. on p. 25).

96. Mi, S., Zhang, J., Zhang, W. & Huang, R. S. Circulating MicroRNAs as Biomarkers for Inflammatory Diseases. *MicroRNA* **2,** 64–72 (May 2013) (cit. on p. 25).

97. Polakovičová, M., Musil, P., Laczo, E., Hamar, D. & Kyselovicˇ, J. Circulating MicroRNAs as Potential Biomarkers of Exercise Response. *International Journal of Molecular Sciences* **17,** 1553 (Oct. 2016) (cit. on p. 25).

98. Parr, E. B., Camera, D. M., Burke, L. M., *et al.* Circulating MicroRNA Responses between 'High' and 'Low' Responders to a 16-Wk Diet and Exercise Weight Loss Intervention. *PLOS ONE* **11** (ed Song, Y.) e0152545 (Apr. 2016) (cit. on p. 25).

99. Zi, Y., Yin, Z., Xiao, W., *et al.* Circulating MicroRNA as Potential Source for Neurodegenerative Diseases Biomarkers. *Molecular Neurobiology* **52,** 1494–1503 (Dec. 2015) (cit. on p. 25).

100. Obulkasim, A., Katsman-Kuipers, J. E., Verboon, L., *et al.* Classification of pediatric acute myeloid leukemia based on miRNA expression profiles. *Oncotarget* (Mar. 2017) (cit. on p. 25).

101. Hayes, J., Peruzzi, P. P. & Lawler, S. MicroRNAs in cancer: biomarkers, functions and therapy. *Trends in Molecular Medicine* **20,** 460–469 (2014) (cit. on p. 25).

102. Heneghan, H. M., Miller, N., Kelly, R., Newell, J. & Kerin, M. J. Systemic miRNA-195 Differentiates Breast Cancer from Other Malignancies and Is a Potential Biomarker for Detecting Noninvasive and Early Stage Disease. *The Oncologist* **15,** 673–682 (July 2010) (cit. on p. 25).

103. Roth, C., Rack, B., Müller, V., *et al.* Circulating microRNAs as blood-based markers for patients with primary and metastatic breast cancer. *Breast Cancer Research* **12,** R90 (Dec. 2010) (cit. on p. 25).

104. Wang, F., Zheng, Z., Guo, J. & Ding, X. Correlation and quantitation of microRNA aberrant expression in tissues and sera from patients with breast tumor. *Gynecologic Oncology* **119,** 586–593 (Dec. 2010) (cit. on p. 25).

105. Hu, Z., Dong, J., Wang, L.-E., *et al.* Serum microRNA profiling and breast cancer risk: the use of miR-484/191 as endogenous controls. *Carcinogenesis* **33,** 828–834 (Apr. 2012) (cit. on p. 25).

106. Zhao, H., Shen, J., Medico, L., *et al.* A Pilot Study of Circulating miRNAs as Potential Biomarkers of Early Stage Breast Cancer. *PLoS ONE* **5** (ed Creighton, C.) e13735 (Oct. 2010) (cit. on p. 25).

107. Chan, M., Liaw, C. S., Ji, S. M., *et al.* Identification of Circulating MicroRNA Signatures for Breast Cancer Detection. *Clinical Cancer Research* **19** (2013) (cit. on p. 25).

108. Kodahl, A. R., Lyng, M. B., Binder, H., *et al.* Novel circulating microRNA signature as a potential non-invasive multi-marker test in ER-positive early-stage breast cancer: A case control study. *Molecular Oncology* **8,** 874–883 (2014) (cit. on p. 25).

109. Lyng, M. B., Kodahl, A. R., Binder, H. & Ditzel, H. J. Prospective validation of a blood-based 9-miRNA profile for early detection of breast cancer in a cohort of women examined by clinical mammography. *Molecular Oncology* **10,** 1621–1626 (2016) (cit. on p. 25).

110. Brink, H., Richards, J. W. & Fetherolf, M. *Real-world machine learning* 242 (2016) (cit. on p. 27).

111. Hastie, T., Tibshirani, R. & Friedman, J. H. (H. *The elements of statistical learning : data mining, inference, and prediction* 745 (2013) (cit. on pp. 28, 29).

112. Breiman, L. *Classification and regression trees* 358 (Chapman & Hall/CRC, 1998) (cit. on p. 28).

113. James, G. (M. *An introduction to statistical learning : with applications in R* (2016) (cit. on p. 28).

114. Geurts, P. *Contributions to decision tree induction: bias/variance tradeoff and time series classification* PhD thesis (University of Liege, Belgium, May 2002) (cit. on p. 29).

115. Schapire, R. E. The boosting approach to machine learning: an overview. *Nonlinear Estimation and Classification* **171,** 149–171 (2003) (cit. on p. 29).

116. Breiman, L. Bagging predictors. *Machine Learning* **24,** 123–140 (Aug. 1996) (cit. on p. 29).

117. Breiman, L. Random Forests. *Machine learning* **45,** 5–32 (2001) (cit. on pp. 29, 34).

118. Liaw, a. & Wiener, M. Classification and Regression by randomForest. *R news* **2,** 18–22 (2002) (cit. on p. 31).

119. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Machine Learning* **63,** 3–42 (2006) (cit. on p. 31).

120. Wu, B., Abbott, T., Fishman, D., *et al.* Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics (Oxford, England)* **19,** 1636–43 (Sept. 2003) (cit. on p. 34).

121. Geurts, P., Fillet, M., de Seny, D., *et al.* Proteomic mass spectra classification using decision tree based ensemble methods. *Bioinformatics* **21,** 3138–3145 (July 2005) (cit. on p. 34).

122. Díaz-Uriarte, R. & Alvarez de Andrés, S. Gene selection and classification of microarray data using random forest. *BMC bioinformatics* **7,** 3 (2006) (cit. on pp. 34, 35).

123. Özçift, A. Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis. *Computers in Biology and Medicine* **41,** 265–271 (2011) (cit. on p. 34).

124. Machado, R. F., Laskowski, D., Deffenderfer, O., *et al.* Detection of Lung Cancer by Sensor Array Analyses of Exhaled Breath. *American Journal of Respiratory and Critical Care Medicine* **171,** 1286–1291 (June 2005) (cit. on p. 34).

125. Hsieh, C.-H., Lu, R.-H., Lee, N.-H., *et al.* Novel solutions for an old disease: Diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks. *Surgery* **149,** 87–93 (2011) (cit. on p. 34).

126. Mihailescu, D. M., Gui, V., Toma, C. I., Popescu, a. & Sporea, I. Computer aided diagnosis method for steatosis rating in ultrasound images using random forests. *Medical Ultrasonography* **15,** 184–190 (2013) (cit. on p. 34).

127. Alickovic, E. & Subasi, A. Medical Decision Support System for Diagnosis of Heart Arrhythmia using DWT and Random Forests Classifier. *Journal of Medical Systems* **40,** 108 (Apr. 2016) (cit. on p. 34).

128. Cima, I., Schiess, R., Wild, P., *et al.* Cancer genetics-guided discovery of serum biomarker signatures for diagnosis and prognosis of prostate cancer. *Proceedings of the National Academy of Sciences of the United States of America* **108,** 3342–7 (2011) (cit. on p. 34).

129. Segura, M. F., Belitskaya-Lévy, I., Rose, A. E., *et al.* Melanoma MicroRNA Signature Predicts Post-Recurrence Survival. *Clinical Cancer Research* **16** (2010) (cit. on p. 34).

130. Wuchty, S., Arjona, D., Li, A., *et al.* Prediction of Associations between microRNAs and Gene Expression in Glioma Biology. *PLoS ONE* **6** (ed Rogers, S.) e14681 (Feb. 2011) (cit. on p. 34).

131. Piepoli, A., Tavano, F., Copetti, M., *et al.* Mirna Expression Profiles Identify Drivers in Colorectal and Pancreatic Cancers. *PLoS ONE* **7** (ed Navarro, A.) e33663 (Mar. 2012) (cit. on p. 34).

132. Wszolek, M. F., Rieger-Christ, K. M., Kenney, P. A., *et al.* A MicroRNA expression profile defining the invasive bladder tumor phenotype. *Urologic oncology* **29,** 794–801.e1 (2007) (cit. on p. 34).

133. Cheng, L., Doecke, J., Sharples, R., *et al.* Prognostic serum miRNA biomarkers associated with Alzheimer's disease shows concordance with neuropsychological and neuroimaging assessment. *Molecular psychiatry* **20,** 1–9 (2014) (cit. on p. 34).

134. Saeys, Y., Inza, I. & Larranaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **23,** 2507–2517 (Oct. 2007) (cit. on p. 34).

135. Hemphill, E., Lindsay, J., Lee, C., Măndoiu, I. I. & Nelson, C. E. Feature selection and classifier performance on diverse bio- logical datasets. *BMC bioinformatics* **15 Suppl 1,** S4 (2014) (cit. on pp. 34, 105).

136. Ghattas, B. & Ishak, A. B. E. N. Sélection de variables pour la classification binaire en grande dimension: comparaisons et application aux données de biopuces. *J. de la SFdS,* 43–66 (2008) (cit. on p. 35).

137. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics* **10,** 57–63 (2009) (cit. on p. 37).

138. Bustin, S. A. Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *Journal of molecular endocrinology* **25,** 169–93 (Oct. 2000) (cit. on p. 37).

139. Costa, C., Giménez-Capitán, A., Karachaliou, N. & Rosell, R. Comprehensive molecular screening: from the RT-PCR to the RNA-seq. *Translational lung cancer research* **2,** 87–91 (2013) (cit. on p. 37).

140. Zhao, S., Fung-Leung, W.-P., Bittner, A., *et al.* Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLoS ONE* **9** (ed Zhang, S.-D.) e78644 (Jan. 2014) (cit. on p. 37).

141. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5,** 621–628 (2008) (cit. on p. 38).

142. Levin, J. Z., Yassour, M., Adiconis, X., *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature methods* **7,** 709–15 (2010) (cit. on p. 38).

143. Dobin, A., Davis, C. A., Schlesinger, F., *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (Jan. 2013) (cit. on pp. 38, 39).

144. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)* **25,** 1105–11 (May 2009) (cit. on p. 39).

145. Garber, M., Grabherr, M. G., Guttman, M. & Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods* **8,** 469–477 (2011) (cit. on p. 39).

146. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15,** 550 (2014) (cit. on pp. 39, 40).

147. Anders, S., Pyl, P. T. & Huber, W. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166–169 (Jan. 2015) (cit. on p. 39).

148. Robinson, M. D. & Smyth, G. K. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* **9,** 321–332 (July 2007) (cit. on p. 40).

149. Anders, S., McCarthy, D. J., Chen, Y., *et al.* Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. en. *Nature protocols* **8,** 1765–86 (Sept. 2013) (cit. on p. 40).

150. Hart, S. N., Therneau, T. M., Zhang, Y., Poland, G. A. & Kocher, J.-P. Calculating sample size estimates for RNA sequencing data. *Journal of computational biology : a journal of computational molecular cell biology* **20,** 970–8 (Dec. 2013) (cit. on p. 40).

151. Kapranov, P., Cheng, J., Dike, S., *et al.* RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription. *Science* **316** (2007) (cit. on p. 41).

152. Zhang, F., Zhang, L. & Zhang, C. Long noncoding RNAs and tumorigenesis: genetic associations, molecular mechanisms, and therapeutic strategies. *Tumor Biology,* 1–13 (2015) (cit. on p. 41).

153. Carninci, P., Kasukawa, T., Katayama, S., *et al.* The Transcriptional Landscape of the Mammalian Genome. *Science* **309** (2005) (cit. on p. 41).

154. Rinn, J. L. & Chang, H. Y. Genome regulation by long noncoding RNAs. *Annual review of biochemistry* **81,** 145–166 (2012) (cit. on pp. 41, 43).

155. Schein, A., Zucchelli, S., Kauppinen, S., Gustincich, S. & Carninci, P. Identification of antisense long noncoding RNAs that function as SINEUPs in human cells. *Scientific Reports* **6,** 33605 (Dec. 2016) (cit. on p. 41).

156. Derrien, T., Johnson, R., Bussotti, G., *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome research* **22,** 1775–89 (Sept. 2012) (cit. on p. 41).

157. Pelechano, V. & Steinmetz, L. M. Gene regulation by antisense transcription. *Nature reviews. Genetics* **14,** 880–93 (2013) (cit. on pp. 41, 42).

158. Balbin, O. A., Malik, R., Dhanasekaran, S. M., *et al.* The landscape of antisense gene expression in human cancers. *Genome research* **25,** 1068–79 (July 2015) (cit. on p. 41).

159. Yu, W., Gius, D., Onyango, P., *et al.* Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. *Nature* **451,** 202–206 (Jan. 2008) (cit. on p. 43).

160. Mahmoudi, S., Henriksson, S., Corcoran, M., *et al.* Wrap53, a Natural p53 Antisense Transcript Required for p53 Induction upon DNA Damage. *Molecular Cell* **33,** 462–471 (Feb. 2009) (cit. on p. 43).

161. Dallosso, A. R., Hancock, A. L., Malik, S., *et al.* Alternately spliced WT1 antisense transcripts interact with WT1 sense RNA and show epigenetic and splicing defects in cancer. *RNA (New York, N.Y.)* **13,** 2287–99 (Dec. 2007) (cit. on p. 43).

162. Bertozzi, D., Iurlaro, R., Sordet, O., *et al.* Characterization of novel antisense HIF-1a transcripts in human cancers. *Cell Cycle* **10,** 3189–3197 (2011) (cit. on p. 43).

163. Shoji, W., Suenaga, Y., Kaneko, Y., *et al.* NCYM promotes calpain-mediated Myc-nick production in human MYCN-amplified neuroblastoma cells. *Biochemical and Biophysical Research Communications* **461,** 501–506 (June 2015) (cit. on p. 43).

164. Yamamoto, T., Manome, Y., Nakamura, M. & Tanigawa, N. Downregulation of survivin expression by induction of the effector cell protease receptor-1 reduces tumor growth potential and results in an increased sensitivity to anticancer agents in human colon cancer. *European journal of cancer (Oxford, England : 1990)* **38,** 2316–24 (Nov. 2002) (cit. on p. 43).

165. Wu, W., Bhagat, T. D., Yang, X., *et al.* Hypomethylation of noncoding DNA regions and overexpression of the long noncoding RNA, AFAP1-AS1, in Barrett's esophagus and esophageal adenocarcinoma. *Gastroenterology* **144,** 956–966.e4 (May 2013) (cit. on p. 44).

166. Nie, L., Wu, H.-J., Hsu, J.-M., *et al.* Long non-coding RNAs: versatile master regulators of gene expression and crucial players in cancer. *American journal of translational research* **4,** 127–50 (2012) (cit. on p. 44).

167. Di Gesualdo, F., Capaccioli, S. & Lulli, M. A pathophysiological view of the long non-coding RNA world. *Oncotarget* **5,** 10976–96 (2014) (cit. on p. 44).

168. Sathirapongsasuti, J. F., Lee, H., Horst, B. A. J., *et al.* Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* **27,** 2648–2654 (Oct. 2011) (cit. on pp. 102, 104).

169. Latysheva, N. S. & Babu, M. M. Discovering and understanding oncogenic gene fusions through data intensive computational approaches. *Nucleic acids research* **44,** 4487–503 (June 2016) (cit. on p. 106).

170. Edgren, H., Murumagi, A., Kangaspeska, S., *et al.* Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biology* **12,** R6 (2011) (cit. on p. 106).

171. Kim, J., Kim, S., Ko, S., *et al.* Recurrent fusion transcripts detected by whole-transcriptome sequencing of 120 primary breast cancer samples. *Genes, Chromosomes and Cancer* **54,** 681–691 (Nov. 2015) (cit. on p. 106).

172. Venables, J. P., Klinck, R., Bramard, A., *et al.* Identification of Alternative Splicing Markers for Breast Cancer. *Cancer Research* **68,** 9525–9531 (Nov. 2008) (cit. on p. 106).

173. Dago, D. N., Scafoglio, C., Rinaldi, A., *et al.* Estrogen receptor beta impacts hormone-induced alternative mRNA splicing in breast cancer cells. *BMC genomics* **16,** 367 (May 2015) (cit. on p. 106).

Appendices

A

List of publications

Peer-reviewed journal publications

- **Exome Copy Number Variation detection: use of a pool of unrelated healthy tissue as reference sample.**
  Stephane Wenric, Tiberio Sticca, Jean-Hubert Caberg, Claire Josse, Corinne Fasquelle, Christian Herens, Mauricette Jamar, Stéphanie Max, André Gothot, Jo Caers, Vincent Bours.
  Genetic Epidemiology. 2016

- **Genomic Studies of Multiple Myeloma Reveal an Association between X Chromosome Alterations and Genomic Profile Complexity.**
  Tiberio Sticca, Jean-Hubert Caberg, **Stephane Wenric**, Christophe Poulet, Christian Herens, Mauricette Jamar, Claire Josse, Sonia El Guendi, Stéphanie Max, Yves Beguin, André Gothot, Jo Caers, Vincent Bours.
  Genes, Chromosomes and Cancer. 2016

- **Circulating microRNA-based screening tool for breast cancer.**
  Pierre Frères*, **Stéphane Wenric***, Meriem Boukerroucha, Corinne Fasquelle, Jérôme Thiry, Nicolas Bovy, Ingrid Struman, Pierre Geurts, Joëlle Collignon, Hélène Schroeder, Frédéric Kridelka, Eric Lifrange, Véronique Jossa, Vincent Bours, Claire Josse, Guy Jerusalem *: as co-first authors
  Oncotarget. 2015

- **Evaluation of BRCA1-related molecular features and microRNAs as prognostic factors for triple negative breast cancers.**
  Meriem Boukerroucha, Claire Josse, Sonia El Guendi, Bouchra Boujemla, Pierre Frères, Raphaël Marée, **Stephane Wenric**, Karin Segers, Joelle Collignon, Guy Jerusalem, Vincent Bours.
  BMC Cancer. 2015

Submitted journal publications

- **Variations of circulating biomarkers during and after anthracycline-containing chemotherapy in breast cancer patients.**
  Pierre Frères, Nassim Bouznad, Laurence Servais, Claire Josse, **Stéphane Wenric**, Aurélie Poncin, Jérôme Thiry, Marie Moonen, Cécile Oury, Patrizio Lancellotti, Vincent Bours, Guy Jerusalem.
  submitted to **BMC Cancer**. 2016

- **Transcriptome wide analysis of natural antisense transcripts shows their potential role in breast cancer.**
  Stephane Wenric, Sonia El Guendi, Jean-Hubert Caberg, Warda Bezzaou, Corinne Fasquelle, Benoit Charloteaux, Latifa Karim, Benoit Hennuy, Pierre Frères, Joëlle Collignon, Meriem Boukerroucha, Hélène Schroeder, Fabrice Olivier, Véronique Jossa, Guy Jerusalem, Vincent Bours and Claire Josse.
  submitted to **Cancer Research**. 2017

Updated list of thesis publications

Boukerroucha, M., Josse, C., ElGuendi, S., Boujemla, B., Frères, P., Marée, R., ... & Bours, V. (2015). Evaluation of BRCA1-related molecular features and microRNAs as prognostic factors for triple negative breast cancers. BMC cancer, 15(1), 755.

Frères, P., Wenric, S., Boukerroucha, M., Fasquelle, C., Thiry, J., Bovy, N., ... & Kridelka, F. (2016). Circulating microRNA-based screening tool for breast cancer. Oncotarget, 7(5), 5416.

Wenric, S., ElGuendi, S., Caberg, J. H., Bezzaou, W., Fasquelle, C., Charloteaux, B., ... & Boukerroucha, M. (2017). Transcriptome-wide analysis of natural antisense transcripts shows their potential role in breast cancer. Scientific reports, 7(1), 17452.

Sticca, T., Caberg, J. H., Wenric, S., Poulet, C., Herens, C., Jamar, M., ... & Gothot, A. (2017). Genomic studies of multiple myeloma reveal an association between X chromosome alterations and genomic profile complexity. Genes, Chromosomes and Cancer, 56(1), 18-27.

Wenric, S., Sticca, T., Caberg, J. H., Josse, C., Fasquelle, C., Herens, C., ... & Bours, V. (2017). Exome copy number variation detection: Use of a pool of unrelated healthy tissue as reference sample. Genetic epidemiology, 41(1), 35-40.

Frères, P., Bouznad, N., Servais, L., Josse, C., Wenric, S., Poncin, A., ... & Bours, V. (2018). Variations of circulating cardiac biomarkers during and after anthracycline-containing chemotherapy in breast cancer patients. BMC cancer, 18(1), 102.