Approche à large échelle visant à détecter de nouveaux régulateurs de l’épissage alternatif au cours de la transition épithélio-mésenchymateuse.

Jean-Philippe Villemin

To cite this version:


HAL Id: tel-03329759
https://tel.archives-ouvertes.fr/tel-03329759
Submitted on 31 Aug 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
THÈSE POUR OBTENIR LE GRADE DE DOCTEUR
DE L’UNIVERSITÉ DE MONTPELLIER

En Biologie Moléculaire et Cellulaire
École doctorale Sciences Chimiques et Biologiques pour la Santé (ED CBS2 168)
Unité de recherche UMR9002 CNRS-UM – Institut de Génétique Humaine (IGH)

Présentée par Jean-Philippe Villemin
Le 24 Novembre 2020

Sous la direction de Reini Fernandez de Luco & William Ritchie

Transcriptome-wide approach to detect novel regulators of alternative splicing during epithelial-to-mesenchymal transition in breast cancer.

Devant le jury composé de

Thérèse Commes, PU/UM, Centre Hospitalier de Montpellier (CHU)
Manuel Irimia, Research Professor et Chef d’équipe, Center for Genomic Regulation (CRG) Barcelone
Cyril Bourgeois, CR, Ecole Normale Supérieure de Lyon (ENS)
Olivier Cuvier, DR et Chef d’équipe, LBME Toulouse
William Ritchie, CR et Chef d’équipe, IGH Montpellier
Reini Fernandez De Luco, CR et Chef d’équipe, IGH Montpellier

Présidente de Jury
Rapporteur
Rapporteur
Examinateur
Co-directeur de Thèse
Directeur de Thèse
# CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>4</td>
</tr>
<tr>
<td>TABLE OF FIGURES</td>
<td>5</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>6</td>
</tr>
<tr>
<td>ENGLISH SUMMARY</td>
<td>7</td>
</tr>
<tr>
<td>FRENCH SUMMARY</td>
<td>9</td>
</tr>
<tr>
<td>PUBLICATIONS</td>
<td>11</td>
</tr>
</tbody>
</table>

## 1. Introduction

### 1.1. Biology of cancer : an overview

1.1.1. Epidemiology
1.1.2. What is cancer?
1.1.3. Cancer is a genetic disease
1.1.4. Intratumor heterogeneity
1.1.5. EMT and metastatic cascade

### 1.2. Epithelial-Mesenchymal Transition (EMT)

1.2.1. Three different subtypes based on the biological context
1.2.1.1. During implantation, embryogenesis and organ development
1.2.1.2. Associated with tissue regeneration and organ fibrosis
1.2.1.3. Related to cancer progression
1.2.2. Resistance to treatment and survival
1.2.3. EMT regulatory programs
1.2.3.1. Transcriptional regulation
1.2.3.2. Post-transcriptional regulation
1.2.3.2.1. Regulatory network of micro-RNA
1.2.3.2.2. At the protein level
1.2.3.2.3. Emerging layers of regulation

### 1.3. Alternative splicing and tumor progression

1.3.1. Splicing reaction
1.3.2. Aspects of its regulation
1.3.3. Definition of alternative (AS) splicing events
1.3.4. Alternative splicing and cancer
1.3.4.1. Affecting hallmarks of cancer
1.3.4.2. During EMT
1.3.4.3. Across cancer types ................................................................. 46
1.3.4.4. Related to EMT................................................................. 50

1.4. High-throughput technology, analysis, models and ressources for cancer research................................................................. 52
1.4.1. High-Throughput technology ................................................................. 52
1.4.1.1. Microarrays ........................................................................ 52
1.4.1.2. High-Throughput sequencing: focus on RNA-SEQ................................. 53
1.4.2. Bioinformatic analysis ......................................................................... 56
1.4.2.1. Differential gene expression ................................................................. 58
1.4.2.2. Differential alternative splicing ................................................................. 59
1.4.2.3. K-mer content ........................................................................ 63
1.4.3. Computational techniques ......................................................................... 64
1.4.3.1. Hierarchical clustering ........................................................................ 64
1.4.3.2. The random forest algorithm ................................................................. 65
1.4.3.3. Survival analysis ........................................................................ 67
1.4.4. Pre-clinical models for cancer ................................................................. 69
1.4.5. Cancer population genomic ressources ................................................................. 71

1.5. A concrete application to breast cancer ......................................................... 73
1.5.1. Epidemiology ........................................................................ 73
1.5.2. Breast anatomy ........................................................................ 73
1.5.3. Clinical characteristics ................................................................. 74
1.5.4. Molecular subtypes of breast cancer ................................................................. 75
1.5.5. Focus on basal-like breast cancer ................................................................. 77
1.5.6. Thesis objectives ........................................................................ 80
1.5.6.1. Identification of AS events associated with poor prognosis in breast cancer .......... 80
1.5.6.2. New insight from k-mers analysis in breast cancer ................................................................. 81

2. RESULTS ................................................................................................. 83

3. GENERAL DISCUSSION AND PERSPECTIVE ......................................................... 86

BIBLIOGRAPHY ................................................................................................. 97

GLOSSARY ................................................................................................. 123

ANNEXES ................................................................................................. 124
ACKNOWLEDGMENTS

First, I would like to thank Dr. Manuel Irimia and Cyril Bourgeois for kindly accepting to be reviewers of my thesis manuscript, Dr. Olivier Cuvier for being an examiner and Pr. Thérèse Commes for accepting to be President of my thesis jury. Thank all of you for accepting to evaluate my thesis work.

I am grateful to Dr. Eduardo Eyras and Dr. Jean Christophe Andrau for accepting to be in my thesis committee during these three years. Thanks for your advices.

Obviously, I am very thankful to Reini Luco who gave me the opportunity to work in her lab during these four years. I also would like to say thanks to William Ritchie for his co-direction and his support.

From the Luco Team, I will miss all its members: Andrew Oldfield, Yaiza Nunez-Alvarez, Alexandre Segelle, Marie-Sarah Cabrillac. I wish you many publications in the years coming.

From the Ritchie Team, I warmly thank Claudio Lorenzi and Sylvain Barriere, wishing them the best for the future.

For all the people who would have helped me and who are disappointed not to be thanked, I am sincerely sorry in advance. However, please know in advance that there is no one I do not wish to thank.
### TABLE OF FIGURES

- **Figure 1-1** Evolutionary Trajectories and Transcriptomic Heterogeneity. .......................... 15
- **Figure 1-2** Sites of EMT & MET in the emergence and progression of carcinoma. ... 17
- **Figure 1-3** Outline of a typical EMT program. ................................................................. 19
- **Figure 1-4** Growth of the primary literature in EMT ....................................................... 20
- **Figure 1-5** Summary of the physiological outcomes of EMT in carcinoma. ............ 23
- **Figure 1-6** Overview of EMT-TF protein structures. ....................................................... 24
- **Figure 1-7** Multilayer of regulation during EMT ............................................................... 26
- **Figure 1-8** Constitutive splicing and alternative splicing. .............................................. 29
- **Figure 1-9** Exons and introns in pre-mRNA with their consensus sequences. ............ 30
- **Figure 1-10** The spliceosome mediates a two-step splicing reaction. ......................... 31
- **Figure 1-11** Cis and Trans regulation of alternative splicing. ........................................ 33
- **Figure 1-12** Domain structure of splicing factors altered in solid tumors ................. 34
- **Figure 1-13** Constitutive splicing versus alternative splicing events. ......................... 36
- **Figure 1-14** Therapeutic strategies to target-splicing alterations in tumors. ............ 46
- **Figure 1-15** Recurrent splicing factor alterations detected in cancer ....................... 47
- **Figure 1-16** Splicing Factors changes and AS events during EMT ............................... 51
- **Figure 1-17** Principle and Workflow of Illumina Next-generation Sequencing ......... 55
- **Figure 1-18** RNA-seq computational analyses. ............................................................... 57
- **Figure 1-19** Methods to determine differential RNA splicing using RNA-seq data .... 60
- **Figure 1-20** Focus on Whippet: global methodology .................................................... 63
- **Figure 1-21** Breast anatomy and histology ................................................................. 74
- **Figure 1-22** Breast cancer survival by molecular subtype ............................................. 77
- **Figure 1-23** Histological heterogeneity of Triple Negative Breast cancer .................. 78
- **Figure 3-1** Transition states occurring during EMT ....................................................... 87
- **Figure 3-2** The Human Tumor Atlas Network (HTAN) ............................................... 95
- **Figure 3-3** LifeTime European Project ................................................................. 96
# LIST OF TABLES

Table 1.1 Tumor-associated isoforms representative of the cancer hallmarks........ 41  
Table 1.2 Examples of genes affected by alternative splicing during EMT. .......... 44  
Table 1.3 Unmutated SFs that function as proto-oncogenes or tumor suppressors.. 48  
Table 1.4 Overview of AS software since 2010............................................. 61  
Table 1.5 Definition of endpoints in clinical trials....................................... 68
Alternative splicing is one of the major mechanisms leading to a diversity in the proteome. It has become very clear that this mechanism is playing a role in many genetic diseases including cancer. During oncogenesis, the cellular content of RNA isoforms is highly altered and this phenomenon seems to be context specific. Even in the same tissue, the pool of transcripts can display specific rearrangements corresponding to different subtypes of the disease.

As we know now, the majority of deaths from solid tumors are caused by metastases. This metastatic cascade might involve the Epithelial-to-Mesenchymal Transition (EMT) which is a complex biological trans-differentiation process that allows epithelial cells to transiently obtain mesenchymal features. During this process, an alternative splicing program is differentially regulated, and increasing number of studies have started to suggest that a simple isoform switching is sufficient to induce or impair an EMT. Stopping the spreading of cancer cells in the human body represents an important challenge in the fight against cancer. In this context, I believe that alternative splicing represents a novel regulatory layer worth exploring to improve cancer diagnosis and identify potential new targets for therapy, which will impact patient's survival and care.

As we are in the era of genomics and transcriptomics, I have taken advantage of the most extensive transcriptomics datasets in breast cancer cell lines (CCLE) and breast cancer patients (TCGA) to identify novel splicing biomarkers of poor prognosis. I identified a 25-gene based splicing signature specific of a subtype of basal-like tumors capable of classifying patients with the worst survival rate. Using several public EMT-induced RNA sequencing projects, I identified this basal-specific splicing signature as a signature characteristic of EMT-induced cells with classical hallmarks of pluripotent stem cells and cell invasion, which are essential for tumor spreading and metastasis.

As a side project, I also got involved in the development of methods of classification using k-mers. I first was involved in a project that tested the ability of k-mer to classify breast cancer subtypes. Secondly, I was focused in the discovery of biological knowledge that k-mers are bringing in the breast cancer stratification.
The results show that alternative splicing or k-mers can be the source of new valuable information to help in the thinner definition of oncogenic subtypes or identification of biological processes in cancer. In a breast cancer subtype that does not benefit from targeted therapy, I demonstrate that alternative splicing relative to an EMT could be used as potential biomarkers to isolate patients where the tumor progresses faster. This work could help to develop new treatments for precision oncology.
L’épissage alternatif est l’un des mécanismes majeurs conduisant à la diversité du protéome. Il est devenu très clair que ce mécanisme joue un rôle dans de nombreuses maladies génétiques, y compris le cancer. Au cours de l’oncogenèse le contenu cellulaire en isoformes d’ARN est fortement altéré et ce phénomène semble être spécifique au contexte. Dans un même tissu, la composition en transcrits peut être différente selon les sous-types de la maladie.

Comme nous le savons maintenant, la majorité des décès dus à des tumeurs solides est causée par des métastases. Cette cascade métastatique pourrait impliquer la transition épithélialo-mésenchymateuse (EMT) qui est un processus biologique complexe de trans-différenciation qui permet aux cellules épithéliales d’obtenir de manière transitoire des caractéristiques mésenchymateuses. Au cours de ce processus, un programme d’épissage alternatif est régulé de manière différentielle, et un nombre croissant d’études commence à suggérer qu’un simple changement d’isoforme pourrait s’avérer suffisant pour amorcer une EMT. Stopper la propagation des cellules cancéreuses dans le corps humain représente un défi important dans la lutte contre le cancer. Dans ce contexte, je pense que l’exploration de l’épissage alternatif pourrait apporter une couche de régulation plus fine pour classer les patients plus précisément, aider à découvrir de nouvelles cibles potentielles pour la thérapie et de ce fait, améliorer la survie et les soins des patients.

Comme nous sommes à l’ère de la génomique et de la transcriptomique, j’ai profité d’un jeu de données exhaustif de lignées cellulaires de cancer du sein (CCLE) et de tumeurs de patients (TCGA) pour identifier de nouveaux biomarqueurs d’épissage alternatif associés à un mauvais pronostic. J’ai identifié une signature d’épissage basée sur 25 gènes spécifiques d’un sous-type de tumeurs basales capable de classer les patients avec le plus mauvais taux de survie. En utilisant plusieurs projets publics de séquençage induisant une EMT dans différents modèles cellulaires, j’ai identifié cette signature basal-spécifique comme une signature caractéristique de l’EMT et de cellules présentant des caractéristiques classiques de cellules souches pluripotentes et invasives, qui sont essentielles pour la propagation de la tumeur et la métastase. En parallèle, je me suis également impliqué dans le développement de méthodes de classification et d’annotation d’événements utilisant des k-mers. J’ai d’abord été
impliqué dans un projet qui teste la capacité des k-mers à classer les sous-types du cancer du sein. Dans un second temps, je me suis focalisé sur la découverte des connaissances biologiques que les k-mers apportent dans la stratification du cancer du sein.

Enfin nos résultats montrent que l’épissage alternatif ou les k-mers peuvent être la source de nouvelles informations précieuses pour aider à la définition plus fine des sous-types oncogènes ou pour permettre l’identification de processus biologiques impliqués dans le cancer. Dans un sous type de cancer du sein qui ne bénéficie pas d’une thérapie ciblée, nous démontrons que l’épissage alternatif en lien avec l’EMT pourrait être utilisé comme biomarqueur potentiel pour isoler les patients ou la tumeur progresse plus rapidement. Ces travaux pourraient aider à développer de nouveaux traitements dans le cadre de l’oncologie de précision.
PUBLICATIONS

ACCEPTED FOR PUBLICATIONS


IN PROCESS

Jean-Philippe Villemín, Caudio Lorenzi, Andrew Oldfield, Marie Sarah Cabrillac, William Ritchie & Reini F. Luco. A cell-to-patient machine learning transfer approach uncovers novel basal-like breast cancer prognostic markers amongst alternative splice variants. (Submitted)

1. INTRODUCTION

In this work, various clinical and biological aspects of cancer are treated. Computational biology facets are described too. This introduction is not intended to be exhaustive. It aims to present the notions and concepts that I will develop in the manuscript. This thesis occurs at the interface of many fields. I hope that this introduction will allow to understand the challenges that fall on each discipline and to exchange with the same vocabulary.

My PhD work was dedicated to the study of alternative splicing in a large cohort of patients harboring a certain type of breast cancer, which is known to be a very heterogeneous disease. I explored the idea that alternative splicing signature, related to an Epithelial-to-Mesenchymal Transition (EMT), a crucial process in tumor progression, could help to better classify patients with different survival outcome and therefore, improve their medical care.

I will first introduce the topic of cancer, with a focus on breast cancer. I will then move to the definition of Epithelial Mesenchymal Transition and its link with tumor progression. As alternative splicing is an important mechanism regulated during EMT, I will recall some definitions and develop several ideas around this subject in the context of cancer.

Next, I will describe current high-throughput technologies and counterpart techniques that are used to analyze the molecular profiles of each tumor sample. I will discuss also the statistical, machine learning and bioinformatics tools that have been used to tackle our initial problematic.

Finally, before reporting our results, I will focus on the basal-like subtype of breast cancer keeping in mind all the concepts that were previously mentioned. I will detail the characteristics of this disease and will explain why the basal-like subtype is a major issue.
1.1. BIOLOGY OF CANCER : AN OVERVIEW

1.1.1. EPIDEMIOLOGY

In 2018, the American Cancer Society estimated the number of new cancer cases at 17 million, and 9 million deaths from cancer worldwide the same year. After cardiovascular diseases, it is the second leading cause of death in developed countries. In males, lung and prostate cancers are the more prevalent disease whereas in women, breast and colon cancer are the most common (American Cancer Society 2018).

By 2040, it's expected to grow to 27.5 million new cancer cases and 16.3 cancer deaths simply due to the growth and aging of the population. Actually, these numbers don’t consider the adoption of lifestyles that are known to increase cancer risk (smoking, unhealthy diet, physical inactivity), which could largely underestimate these predictions (American Cancer Society 2018).

1.1.2. WHAT IS CANCER?

According to Centers for Disease Control and Prevention (CDCP) definition, cancer is a disease in which a subset of cells in the breast grow out of control. It can start any place in the body. Cancer cells usually form a tumor growth that can often be seen on an x-ray or felt as a lump. These cells ignore the normal rules of cell division and thus will have pathological consequences on the human body.

In 2000, Hanahan and Weinberg first summarized how tumors cells differ from normal cells in several aspects (Hanahan and Weinberg 2000). In order to rationalize the complexities of neoplastic diseases, they described six hallmarks of cancer:

- Sustaining proliferative signaling
- Evading growth suppressors
- Resisting cell death
- Enabling Replicative Immortality
• Inducing angiogenesis
• Activating invasion and metastasis.

Acquisition of this functional features arise at various times during the course of tumorigenesis allowing cancer cells to survive, proliferate and disseminate.

Almost ten years after, two emerging hallmarks were added (Hanahan and Weinberg 2011):

• Reprogramming of energy metabolism
• Evading of immune destruction.

Acquisition of these core and new hallmarks was proposed to be the consequence of two other phenomena: (1) genome instability and mutation which generates the genetic diversity that play a role in their acquisition, and (2) inflammation which fosters these multiple functional features. Notably, it has been proposed that aberrant alternative splicing should be added to the growing list of these cancer hallmarks (Ladomery 2013).

1.1.3. CANCER IS A GENETIC DISEASE

During cellular division, the DNA sequence is copied. Errors can be introduced. It can be single nucleotide exchange (mutation) or small insertion and deletion of several bases (indels). Also, modifications of the number of copies of DNA segments can occurs (CNA, copy number alterations).

Normal cells use DNA repair to correct these errors, or apoptosis when repair fails. These processes stop the propagation of the errors in the genetic code that can be responsible for an abnormal cell behavior. Tumor cells are able to bypass these mechanisms, giving them immortality.

Genetic changes can be inherited from our parents, arise after specific environmental exposure, or being the result of spontaneous errors during the cell division.
These modifications can have an impact at the protein level, but it’s not mandatory. Gene harboring changes in its sequence, is transcribed into pre-messenger RNAs. Before being translated into a protein, the transcript needs to be spliced to remove the non-coding intronic sequences. Frequently, pre-mRNAs are also alternatively spliced into different mature RNAs with different subset of exons, including or not the modifications. At the end, the protein produced might not be functional and have an impact on cellular behavior.

Genes have different effects on the cellular phenotype and are not necessarily needed for cancer progression. Some of them have been associated to oncogenesis and several definitions have been settled based on their behavior. Oncogene defines a category of genes which expression promotes tumor progression whereas tumor suppressors are genes that are losing their function during cancer.

Following this same idea, not all mutations contribute equally to cancer progression. Mutations that promotes the occurrence of cancers, are called driver mutations while passenger mutations describe modifications in the sequence that do not have functional impact on the cell. These mutations will occur in different cells from the tumors (Figure 1-1), leading to a patchwork of cellular clones with distinct phenotypes (Hinohara and Polyak 2019).

**Figure 1-1 Evolutionary Trajectories and Transcriptomic Heterogeneity.**
Tumors cells follow different evolutionary trajectories forming genetically distinct sub-clones, some of whom can have advantage during cancer progression due to driver mutations. Each distinct clone can exhibit substantial phenotypic variation due to cellular transcriptomic heterogeneity. (adapted from Hinohara and Polyak 2019)
Recently, it was also suggested that some alternative splicing events could potentially be considered alternative splicing drivers (AS-drivers) leading oncogenic processes by themselves (Climente-González et al. 2017). Also, splicing factors, proteins involved in the RNA splicing, can also act as proto-oncoproteins and tumors suppressors (Dvinge et al. 2016). I will discuss all of these aspects in more depth later in the text.

1.1.4. INTRATUMOR HETEROGENEITY

Intratumor heterogeneity describes the observation that different tumor cells can show distinct morphological and phenotypic profiles as previously illustrated (Figure 1-1). This is one of the greatest challenges in precision cancer therapy (Levitin, Yuan, and Sims 2018). Genomic instability can give a selective advantage to certain cells and promotes their growth. Tumor cells are not homogenous and are represented by several clones (Marusyk and Polyak 2010). Ancestral mutations are acquired at the beginning of the oncogenic process and can be shared by all the tumors cells whereas new events can give new traits with potential benefits to tumor progression (Visvader 2011). For example, a set of somatic mutations can empower cancer cells to disseminate and thereafter proliferate in a distant organ. For example, these mutations can enhance/repress the tumorigenic activity of tumour-initiating cells (TICs) also known as cancer stem cells (CSCs). Interestingly, alternative splicing aberrations could have a similar effect. Nevertheless, the origin of the TICs could have implications for the therapeutic strategies that is used to target them (B. B. S. Zhou et al. 2009). Indeed, there is a huge need to better characterize this heterogeneity to define phenotypic subclasses sharing common features, to better understand resistance to treatment and adapt therapies consecutively.

1.1.5. EMT AND METASTATIC CASCADE

When breast cancer spreads to other parts of the body, through blood vessels and lymph vessels, it is said to have metastasized. Notably, the majority of deaths from solid tumors are caused by metastases (Dillekås, Rogers, and Straume 2019). The model of the metastatic cascade as proposed by Thierry (Figure 1-2) starts with the fact that future metastatic cells have to free themselves from the primary tumor mass.
They enter the blood system and migrate within the whole organism until they find a place to grow again. However, the site of metastasis is dependent on the affinity of the tumor for the given microenvironment, which elegantly explains why some organs (lung, liver, bone marrow) are particularly prone to host metastases while others are not (intestine, skeletal muscle, skin) \cite{Samatov2013}.

Epithelial-to-Mesenchymal Transition (EMT) is a process that probably plays a role in the migration of cancer cells \cite{Nieto2016, Pastushenko2019, TBrabletz2018, Lambert2016}. To acquire an invasive phenotype for metastatic progression in cancer, carcinoma cells exploit EMT to facilitate its dissociation from primary tumor and dissemination into blood circulation \cite{WLu2019, Ye2015}. Of note, a reverse process called

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1-2}
\caption{Sites of EMT & MET in the emergence and progression of carcinoma.}
\end{figure}

Multiple genetic alterations leads to a carcinoma in situ and can induce local dissemination of carcinoma cells, possibly through an epithelial—mesenchymal transition (EMT). The basement membrane becomes fragmented. The cells can intravasate into lymph or blood vessels, allowing their passive transport to distant organs. At secondary sites, solitary carcinoma cells can form a new carcinoma through a mesenchymal—epithelial transition (MET) (adapted from J.P Thierry 2002).
Mesenchymal-to-Epithelial Transition (MET) is thought to play a role in the formation of the new carcinoma and colonization of the new tissue.

Thus, a better understanding of the mechanisms underlying the dissemination of tumor cells into the whole body is necessary to stop the spreading of the disease and is a promising strategy to reduce cancer mortality. In the next section, I will deeply explore the concepts of EMT.

1.2. EPITHELIAL-MESENCHYMAL TRANSITION (EMT)

Epithelial-Mesenchymal Transition (EMT) is a cellular process during which epithelial cells acquire a mesenchymal phenotype and behavior following the downregulation of epithelial features (Derynck and Weinberg 2019; Nieto et al. 2016; Dongre and Weinberg 2018). This is a reversible process. The initial epithelial state of the cell is characterized by stable epithelial cell-cell junctions, apical-basal polarity and interactions with basement membrane. The process of EMT leads to profound phenotypic changes on cells (Figure 1-3). Cytoskeleton and cell-matrix adhesion are remodeled, apical-basal cell polarity is lost and cell-cell adhesion is weakened. An individualization of the cells is observed and in addition, cells gain motility. The modification of the adhesion molecules expressed by the cell allows them to adopt a migratory and invasive behavior. This phenomenon has been observed during the course of development, wound healing, and propagation in cell culture. It’s also thought to be an important mechanism driving malignant progression (Craene and Berx 2013). As I mentioned before, this mechanism is reversible and the reciprocal changes in cellular phenotype that reverse EMT-induced phenotypes are called Mesenchymal-Epithelial Transition (MET) which occurs during cancer and embryonic development (J. Yang and Weinberg 2008).
EMT has long been viewed as a binary process with two distinct cell populations, epithelial and mesenchymal and is often defined by the loss of the epithelial marker E-cadherin and the gain of the expression of the mesenchymal marker vimentin. However, recent studies indicate that EMT occurs in a gradual manner characterized by several cellular states expressing different levels of epithelial and mesenchymal markers and exhibiting intermediate morphological, transcriptional, and epigenetic features, between epithelial and mesenchymal cells. The intermediate states between epithelial and fully mesenchymal states have been referred to as partial, or hybrid EMT states (Nieto et al. 2016; J. Yang et al. 2020; Pastushenko and Blanpain 2019). Moreover, this intermediate state has often been associated with stemness capacities (Mani et al. 2008; Wilson et al. 2020) and has a highly tumorigenic potential (Kröger et al. 2019; Saitoh 2018; Aiello et al. 2018; Jolly 2015; Simeonov et al. 2020).
This research topic has been active for a long time, as testified by the Figure 1-4. In the late 1970s, it was Elizabeth Hay who first described an “epithelial transformation” using a model of chick primitive streak formation. The term Epithelial-mesenchymal transition became the term of use later in 2003. Then in 2007, EMT was proposed to be classified between three different subtypes based on the biological context that I will use to further comment EMT in the following text.

1.2.1. THREE DIFFERENT SUBTYPES BASED ON THE BIOLOGICAL CONTEXT

1.2.1.1. DURING IMPLANTATION, EMBRYOGENESIS AND ORGAN DEVELOPMENT

Just before the first stages of embryogenesis, the implantation of the embryo and the initiation of placenta formation are both associated with an EMT in order to
facilitate mechanism of invasion of a specific layer of cells and the proper anchoring of the placenta (Vicovac and Aplin 1996). After fertilization, the egg undergoes gastrulation, which is a universal process by which the body plan is established, generating three germ layers. A primitive streak generates an intermediate layer of cells, which subsequently separates to form mesoderm and endoderm via an EMT (Nakaya and Sheng 2008). Then, during embryonic development, neural crest cells undergo EMT (Kalcheim 2015) and individual cells migrate before giving rise to different derivatives as for example, the melanocytes that provide pigment to the skin. Another example is during somite formation where a mesenchyme layer undergo MET to form epithelial somites, which then undergo EMT to give rise to the sclerotome (Acloque et al. 2009). EMT is also observed during heart valve formation (Nakajima et al. 2000) and Mullerian duct regression (Klattig and Englert 2007).

### 1.2.1.2. ASSOCIATED WITH TISSUE REGENERATION AND ORGAN FIBROSIS

EMT does not only occur during embryonic development. A similar process to EMT also occurs as a physiological response to injury (Stone et al. 2016). During wound healing, keratinocytes at the border of the wound, release a mix of inflammatory compounds that recapitulate part of the EMT process. In fibrotic tissues, inflammatory cells and fibroblasts release a variety of inflammatory signals as well as components of a complex extracellular matrix. Myofibroblasts accumulate and secrete an excessive amount of collagen that is deposited as fibers, thereby compromising organ function and leading to its failure. The origin of fibrosis had been thought to come from the pathological activation of interstitial fibroblasts that convert to myofibroblasts to form the fibrotic collagen network. An important part of these myofibroblasts might arise from the conversion of epithelial cells through an EMT process (Iwano et al. 2002). That said, the origin of myofibroblasts is still an active source of debate since recent linear tracing studies seems to indicate that few of those epithelial cells contribute to their formation (Humphreys et al. 2010; Lebleu et al. 2013). Nevertheless, the hypothesis of acquisition of partial EMT program seems to agree all points of view (Nieto et al. 2016).

### 1.2.1.3. RELATED TO CANCER PROGRESSION
EMT is thought to be activated in cancer cells, linked to their dissociation from the primary tumor and their intravasation into blood vessels (Dongre and Weinberg 2018; Craene and Berx 2013). During the multistep progression of carcinomas that are initially benign, epithelial cells acquire a few distinctly mesenchymal traits that confer them the ability to invade adjacent tissues and then to disseminate to distant tissues.

Before going into the explanation of the underlying molecular mechanism, it seems important to report aside that if EMT has been so extensively studied in cancer, it is because cells following an EMT appear to be more resistant to treatment. Knowing that recurrent cancer might come back in the host even after chemotherapy, it was pertinent to ask if EMT was not involved in patients where the recurrence was observed.

### 1.2.2. RESISTANCE TO TREATMENT AND SURVIVAL

In a recent review about EMT, the authors state that EMT confers resistance to chemotherapy and immunotherapy (Nieto et al. 2016). This statement should be taken with caution. There is an increasing number of evidence that supports this idea but the mechanisms of resistance might be context and drug dependent. For example, in a very specific manner, EMT has been previously found associated with Doxorubicin resistance in breast cancer (Jin et al. 2019; Q. Q. Li et al. 2009). Interestingly, Dutertre & al, have reported recently that an EMT-related splicing switch is linked to, but does not directly explain, drug resistance (Tanaka et al. 2020). EMT has also been associated with platinum-based chemotherapy in epithelial ovarian cancer (Marchini et al. 2013). Broader spectrum studies have failed to demonstrate that all cancer with mesenchymal features are more resistant to all kind of therapies or have a worse outcome when compared to epithelial carcinomas (Tan et al. 2014). On the other hand, they highlighted some specifics associations. Using an EMT signature based on gene expression, mesenchymal pancreatic cancer, malignant melanoma, renal cancer and liver cancer cell lines were more sensitive to compounds targeting microtubule dynamics, such as Vinblastine and Docetaxel. Mesenchymal breast, lung and uterine cancer cell lines were more resistant to Afatinib and Gefinitib. Association of EMT with resistance to gefitinib (EGFR inhibitor) was also showed elsewhere in non-small cell
lung cancer (Byers et al. 2013). Studies that find a limited contribution of EMT to the establishment of metastases, argue that on the contrary EMT is associated to resistance to treatment, which further supports a functional link between EMT and drug resistance (Fischer et al. 2015; X. Zheng et al. 2015). Nevertheless, since the EMT is not more seen as a binary process (Figure 1-5), several studies started to show an association between partial EMT gene signature and a bad outcome (George et al. 2017; Grosse-Wilde et al. 2015). Finally, even if the link between EMT and resistance to treatment still need to be better understood, underlying molecular mechanisms of EMT have been extensively documented.

![Figure 1-5 Summary of the physiological outcomes of EMT in carcinoma.](image)

1.2.3. EMT REGULATORY PROGRAMS

1.2.3.1. TRANSCRIPTIONAL REGULATION

The EMT is executed in response to pleiotropic signaling factors that induce the expression of specific transcription factors (TFs) (Lamouille, Xu, and Derynck 2014; Puisieux, Brabletz, and Caramel 2014). This core is referred as EMT-TFs (Figure 1-6) and has been found to control cell–cell adhesion, cell migration and ECM degradation, and to play evolutionarily conserved central roles in the execution of EMT in various
biological settings and organisms. Among them, we retrieve transcription factors belonging to the Snail, Twist and Zeb families (Stemmler et al. 2019).

**Snail family**

- **Snail**
  - SnAG
  - Destruction box
  - NES
  - ZnF1
  - ZnF2
  - ZnF3
  - ZnF4

- **Slug**
  - SnAG
  - Slug domain
  - ZnF1
  - ZnF2
  - ZnF3
  - ZnF4

**bHLH family**

- **TWIST1**
  - Gly rich
  - bHLH domain
  - Twist box
  - WR domain

**EMT-TF comparative size**

- ZEB1
- ZEB2
- Snail
- Slug
- TWIST1

---

**Figure 1-6 Overview of EMT-TF protein structures.**

Schematic representation of the protein structures of the core EMT-TFs with depiction of the comparative size of all EMT-TFs at the bottom of the figure. (Adapted from Brabletz 2019)

Snail1 (Snail) and Snail2 (Slug), zinc-finger transcription factors, are composed of a highly conserved carboxy-terminal region containing four to six C2H2-type zinc fingers organized in one cluster, which mediate sequence-specific interactions with DNA promoters containing an E-box sequence (CAGGTG) (Stemmler et al. 2019). In all
vertebrates, we observe also evolutionarily conserved domain (SNAG) in the N-terminal part of the protein which is necessary for the binding of co-repressor complexes of transcription (Y. Wang et al. 2014). Snail represses E-cadherin expression (Batlle et al. 2000), a key marker of epithelial state which is thought to be a metastatic suppressor during tumor progression. It also mediates downregulation of cell adhesion molecules, such as occludins and claudins, and upregulation of matrix metalloproteinases.

Members of the zinc finger-homeodomain transcription factor family, ZEB1 and its paralog ZEB2 are genes which activation occurs frequently upon Snail activation. Both of them also contain C2H2-type zinc fingers, essential for the binding of E-box-like elements in the promoters of their target genes (Stemmler et al. 2019). The two zinc-finger clusters in the ZEB proteins are separated by several hundred amino acids; thus, they have the ability to bind to two, relatively closely spaced E-boxes that are very often present as tandem repeats. Of note, ZEB1/2 is active in some tumors that lack SNAIL1/2 expression and thus the regulation of ZEB1/2 expression should be analyzed independently because the contribution of different EMT inducers is dependent on the cellular context (Vandewalle, Van Roy, and Berx 2009). For instance, ZEB1 expression is important during colon cancer progression (Guo et al. 2017) or pancreatic cancer (Krebs et al. 2017), whereas ZEB2 has been studied in ovarian, gastric, and pancreatic tumors, where it is associated with invasiveness and aggressive behavior (W. Lu and Kang 2019). ZEB1 has also been reported as a well-established transcriptional suppressor of E-cadherin (Eger et al. 2005). It also contributes to the formation of the tumor microenvironment by regulating the levels of various inflammatory cytokines, such as interleukin 6/8 (IL-6/8), which resulted in increased tumor growth in basal-like breast cancer cells (Wu et al. 2020).

TWIST1, a basic helix-loop-helix (bHLH) transcription factor, is a short stretch of basic amino acids that are followed by two amphipathic α-helices separated by a loop of varied length (Stemmler et al. 2019). It binds as dimers and recognize also E-boxes to play its role of transcription factor. In human mammary epithelial cells, TWIST1 upregulated macrophage chemoattractant (CCL2) (Low-Marchelli et al. 2013) and platelet-derived growth factor receptor A (PDGFRA) (Eckert et al. 2011) which
stimulates cell signaling pathways that elicit responses such as cellular growth and differentiation.

These EMT-factors can act differently and cooperate in different manner depending on the tumor context. For example, TWIST1 is upregulated in lung adenocarcinoma, whereas no effect on its regulation is observed in Ewing Sarcoma (Stemmler et al. 2019). ZEB2 loss in melanoma is associated with reduced patient survival and inhibits tumor initiation and metastatic progression in mice (Caramel et al. 2013; Denecker et al. 2014). While ZEB1 expression in melanoma is associated with poor clinical outcome and can instead drive melanoma initiation and malignant progression (Richard et al. 2016; Y. Chen et al. 2017). Finally, it is important to keep in mind that their expression alone is not sufficient to point to an EMT. Several other layers of regulation, described below, are involved in this complex mechanism.

1.2.3.2. POST-TRANSCRIPTIONAL REGULATION

Figure 1-7 Multilayer of regulation during EMT

Although multiple non-coding RNAs control EMT, two regulatory networks have been described that can be considered as the core regulatory machinery: the miR34-SNAI1 and miR200-ZEB1. They not only contribute to the epigenetic control of EMT, but are also targets for epigenetic modifications. Downstream of these axes, regulation of transcript processing would shape the landscape of epithelial and mesenchymal effectors through alternative splicing (Adapted from Nieto & al. 2016)
1.2.3.2.1. REGULATORY NETWORK OF MICRO-RNA

MicroRNA (miRNA) are small non-coding molecules (containing about 22 nucleotides) that act as post-transcriptional regulation of gene expression via base-pairing with complementary sequences of the targeted mRNA molecule. Multiple miRNAs are thought to govern EMT (Abba et al. 2016), but two regulatory networks involving miR-200 (Hill, Browne, and Tulchinsky 2012; S. Brabletz and Brabletz 2010) and miR-34 (Imani et al. 2017; Siemens et al. 2011), together with ZEB1 and SNAI1, have been described (Figure 1-7). These two miR-transcription factor (TF) axes employ a double-negative feedback mechanism in which miR34-SNAI1 and miR200-ZEB1 repress each other (Nieto et al. 2016). Several micro-RNAs within the miR-200 family, miR-200a/b/c, miR-141, and miR-429 were identified to target and inhibit ZEB1 translation to a different extent (S. Brabletz et al. 2011).

1.2.3.2.2. AT THE PROTEIN LEVEL

EMT is also regulated by post-translational modifications as phosphorylation which is known to control Snail1. Phosphorylation is catalyzed by enzymes called Protein Kinases (PK) that catalyze the transfer of γ-phosphate of ATP to serine, threonine or tyrosine residues on target proteins. GSK-3β phosphorylates SNAIL at two consecutive motifs that control its ubiquitination (B. P. Zhou et al. 2004). First, GSK-3β binds to SNAIL and phosphorylates SNAIL at one motif, which induces the nuclear export of SNAIL. Then, the phosphorylation on a second motif promotes the ubiquitin-mediated proteasome degradation of SNAIL by β-Trcp. The inhibition of GSK-3β results in the upregulation of Snail1 and downregulation of E-cadherin that results in the activation of the EMT program. Phosphorylation also affects Snail1 subcellular localization (Domínguez et al. 2003). Of note, Twist1 has also been described as phosphorylated by MAP kinases (J. Hong et al. 2011) and more recently, Fattet et al. reveal a pathway in which extracellular matrix stiffness promotes EPHA2/LYN complex activation, leading to phosphorylation of TWIST1 and its nuclear localization, triggering EMT in breast cancer (Fattet et al. 2020).

SUMOylation is another post-translational modification characterized by the reversible binding of Small Ubiquitin-like MOdifier (SUMO) to the target protein. FoxM1
can promote EMT through its direct binding at the SLUG promoter. FoxM1 is subject to SUMOylation at lysine 463 and this posttranslational modification is required for the full repression of miR-200b/c in breast cancer cells that is another layer of regulation of the EMT process (C. M. Wang et al. 2014).

1.2.3.2.3. EMERGING LAYERS OF REGULATION

Epigenetic control (Histone modifications, methylation) is certainly also an important part of the regulation of EMT (Bedi et al. 2014). For example, the miR-200 family is subjected to epigenetic modifications which is regulated by a histone demethylase, KDM5B (Enkhbaatar et al. 2013). Regarding transcriptions factors involved in EMT, SNAI1, responsible for the E-Cadherin repression is regulated by the recruitment to specific DNA sequences of several proteins (Peinado et al. 2004) as chromatin modifiers, (HDAC1, HDAC2) that will determine the acetylation status of histones. Polycomb repressive complex 2 (PRC2) has also been identified in the repression of E-cadherin (Herranz et al. 2008) as well as several histone methyltransferase (Dong et al. 2013; Y. Lin, Dong, and Zhou 2014). In a model of human mammary epithelial cells where snail was induced, transient and long-lasting chromatin changes that sustains EMT were globally described (Javaid et al. 2013). Millanes-Romero & al, shows also that SNAI1 could also regulate heterochromatin transcription (Millanes-Romero et al. 2013). Finally, the histone deacetylases HDAC1 and HDAC2 are also recruited by ZEB1 to downregulate E-cadherin expression in pancreatic cancer (Aghdassi et al. 2012).

All these layers of regulation illustrate how complex the regulation of EMT can be. During my PhD work, I focused on alternative splicing layer. As we will see, alternative splicing is also an important layer of regulation during EMT. In the next section, I will define the usual splicing mechanism and the process of alternative splicing. I will illustrate the fact that splicing is an important mechanism for the definition of the cell phenotype in the context of EMT and tumor progression.
The central dogma of molecular biology is that the genetic information encoded in DNA is transcribed into RNA and then translated into protein. RNA splicing is a form of RNA processing in which a newly made precursor messenger RNA (pre-mRNA) transcript is transformed into a mature messenger RNA (mRNA). Many eukaryotic genes are interrupted by non-coding intervening sequences, or introns, that will be removed from these precursor gene transcripts before being translated into proteins. The remaining flanking sequences are called exons, and are pasted together giving birth to the mRNA. It is only after this processing that the mRNA will be translocated into the cytoplasm for its translation and protein synthesis.

Alternative splicing (AS) (Figure 1-8) allows for the production of various protein isoforms from one single coding gene. When AS involves the use of alternative donor and acceptor sites, we speak about alternative splicing events that lead to the production of several transcript isoforms with distinct sequence content and potentially different biological functions. Therefore, alternative splicing represents a critical step of gene expression.

![Figure 1-8 Constitutive splicing and alternative splicing.](image)

Constitutive exons involves the excision of all exons to form a mature mRNA containing only the exons. This gene will give one protein. In the case of alternative splicing, exons can be included or excluded. A gene can thus form different mature mRNA and therefore different proteins (adapted from Clara Benoit Pilven 2016)
In order to produce short functional RNA messengers, splicing must be specific and reproducible. Short conserved sequences at the ends of introns — splice sites — are crucial for intron recognition and for the accuracy of the splicing reactions (Mount 1982). Most commonly, introns are flanked by conserved GU dinucleotide at its 5′ end and AG at its 3′ end (Figure 1-9). Upstream the 3′ splice site, there is a region rich in pyrimidines (C and U), the polypyrimidine tract, and the branch point, located anywhere around 30 nucleotides upstream from the 3′ end of an intron. The branch point always contains an adenine but its closest adjacent nucleotides are loosely conserved. A typical sequence is YNYRAY, where Y indicates a pyrimidine, N denotes any nucleotide, R denotes any purine, and A denotes the conserved adenine.

Splicing effectors form the spliceosome (Figure 1-10), a ribonucleoprotein (RNP) complex comprised of five snRNPs (small nuclear ribonucleoprotein particles) and numerous proteins (Hegele et al. 2012). Each of the snRNPs that makes up the spliceosome contains a small nuclear RNA (snRNA) named U1, U2, U4/U6 et U5. Of note, U1 to U5 snRNAs are transcribed by RNA polymerase II and are processed the same way while U6 snRNA is transcribed by RNA polymerase III and has its own processing pathway. At the beginning of the splicing reaction, the branch site is initially recognized by the branchpoint-binding protein (BBP) and small nuclear ribonucleoproteins (snRNPs) with auxiliary factors, including U2AF65 and U2AF35, will recognize the consensus sequences in the pre-mRNA. However, from a biochemical
point of view, the RNA splicing reaction is a relatively simple process that consists of two transesterification reactions (Saldanha et al. 1993).

Figure 1-10 The spliceosome mediates a two-step splicing reaction.

Both steps involve transesterification reactions that occur between RNA nucleotides. This two-step biochemical process is driven by the spliceosome. The first transesterification step consists of the nucleophilic attack by the 2’OH group of a key adenosine in the branch consensus site on the 5’ splice site, resulting in the formation of a branched RNA intermediate known as the intron lariat. In the second transesterification step, the 3’OH group of the upstream exon attacks the 3’ splice site, and this produces the spliced mRNA and the excised intron lariat, which is subsequently degraded. (adapted from Kornblihtt, 2013).
First, the pre-mRNA is cleaved at the 5’ end of the intron following the attachment of a snRNP called U1 to its complementary sequence within the intron. The hydroxyl (OH) group of a specific adenosine at the branch site near the 3’ end of the intron attacks the 5’ splice site. This reaction releases the 5’ exon and leaves the 5’ end of the intron joined by a phosphodiester bond to the branch site adenosine. This first reaction of transesterification forms a looped structure known as a lariat.

Then, the snRNPs U2 and U4/U6 appear to contribute to positioning of the 5’ end and the branch point in proximity. With the participation of U5, the 3’ end of the intron is brought into proximity, cut, and joined to the 5’ end. The second transesterification occurs when another hydroxyl (OH) group of the 5’ exon intermediate attacks the 3’ splice site, producing the released spliced mRNA and lariat-shaped intron product that will be degraded by cellular nucleases (Montemayor et al. 2014).

1.3.2. ASPECTS OF ITS REGULATION

Spliceosomal recognition of these core elements is modulated by a myriad of additional sequence elements in both exons and introns that either activate (exonic splicing enhancer, ESE; intronic splicing enhancer, ISE) or repress (exonic splicing silencer, ESS; intronic splicing silencer, ISS) spliceosome recruitment (Figure 1-11) (Z. Wang and Burge 2008; Blencowe 2006). This cis-regulatory layer is driven by short sequences (~10 nucleotides). A trans-regulation layer is added by the interaction of these regulatory sequences with a variety of splicing factors (SF) (Yoshida, Kenichi Ogawa 2014), including serine-arginine-rich (SR) proteins and heterogeneous nuclear ribonucleoproteins (hnRNPs). SF can be divided into two types depending on their downstream effects on alternative splicing. For example, SR proteins tend to enhance the inclusion of alternative spliced exon, these SF are called activators. When the SF are more subject to play an inhibitory role, as hnRNPs, leading to exon skipping, these SFs are called repressors (E. Park et al. 2018; E. Wang and Aifantis 2020).
The splicing machinery therefore has to select between multiple splice sites in a context-dependent manner, relying on sequence features in cis and trans-acting splicing regulators that either promote or repress splice site recognition and spliceosome assembly. Regulation by SFs can be the playground of complex interaction between them (Koedoot et al. 2019). Until now, thousand splicing factors have been described grouped into different families with high degenerated binding motifs (Ray et al. 2013; Dominguez et al. 2018). Figure 1-12 shows the landscape of domain structure of major splicing factors altered in solid tumors.
Interestingly, novel regulatory layers have recently emerged. Coupling with the transcriptional machinery, chromatin conformation and histone modifications, post-transcriptional RNA modifications and non-coding RNAs have all been shown to play a role in the final splicing outcome (Romero-Barrios et al. 2018; L.-Y. Zhu et al. 2018).

For example, enrichment of H3K79m2 have been observed in specific AS events (Romero-Barrios et al. 2018; L.-Y. Zhu et al. 2018).
across normal and cancer cell types \((T. \, Li \, et \, al. \, 2018)\); H3K36me3 was associated with fate determination in hESC (human embryonic stem cell) \((Yungang \, Xu \, 2017)\); NCAM alternative splicing was shown to be influenced by H3K9 hyper-acetylation restricted to a region surrounding the alternative exon; a mechanism of chromatin-mediated splicing was shown to involve a long noncoding RNA (lncRNA) within the human FGFR2 locus \((I. \, Gonzalez \, et \, al. \, 2015)\); intragenic looping mediated by CTCF was suggested to regulate alternative exon usage \((Ruiz-Velasco \, et \, al. \, 2017)\).

To conclude, there are therefore many mechanisms which influence the regulation of AS. In the next sections, I will introduce the core definition of AS events and discuss further the importance of AS in cancer.

1.3.3. DEFINITION OF ALTERNATIVE (AS) SPICING EVENTS

Alternative splicing ensures the biodiversity of proteins that can be encoded by the genome. Modern analyses of human transcriptome have revealed that >95% of our genes undergo alternative splicing, which permits a limited genome to encode a vast proteomic repertoire \((Pan \, et \, al. \, 2008; \, Black \, 2003; \, Barash \, et \, al. \, 2010)\). The expressed isoforms will depend on the cell type, the differentiation state, the physiological state or the developmental stage. These isoforms are the result of splicing events where a portion of a gene can be kept or removed in the final mature RNA. Five majors splicing events are defined \((Figure \, 1-13)\) \((E. \, Wang \, and \, Aifantis \, 2020)\).
Exon skipping (cassette exons) is thought to be the predominant one. The complete exon sequence can be included in the mature RNA or totally spliced out. We speak about mutually exclusive exons when inclusion of one exon lead to the exclusion of the next one (and vice versa). In this particular case, the two alternative exons are never present at the same time in the mature RNA. When only a part of the exon is present in the final RNA, it is called alternative 5’/3’ splice site depending on which site is chosen to be included or not. Finally, far from being the least interesting, introns can be retained in the final transcript. This kind of event is called Intron Retention but in most of the cases transcripts produced are thought to be quickly degraded via nonsense-mediated decay (NMD), a surveillance mechanism that eliminates aberrant mRNAs. Sometimes Alternative First/Last Exon is also listed as an alternative splicing phenomenon, but it is not strictly an alternative splice variant. It’s important to note that
several of these events can occur together in the same gene, leading to a complex recombination of the sequences which make the splicing analysis even more difficult.

Interestingly, the idea that alternative splicing leads to protein diversity has also been recently questioned by large-scale mass-spectrometry experiments where it was proposed that only a minor fraction of the splice variants detected by transcriptomics profiling were actually translated (Tress, Abascal, and Valencia 2017). This started a great debate where advocates of protein diversity related to alternative splicing responded that numerous studies have shown the link between AS and proteomic complexity, and the poor overlap was due to a technical limitation of the proteomics studies (Blencowe 2017). Moreover, it was recently demonstrated in an elegant manner that at least 75% of human exon-skipping events detected in transcripts using RNA-seq data were also detected in ribosome profiling data, thus indicating a role for AS in modulating translational output (Weatheritt, Sterne-Weiler, and Blencowe 2016).

1.3.4. ALTERNATIVE SPlicing AND Cancer

Nowadays, it is well established that AS is highly associated with numerous genetic diseases (Scotti and Swanson 2016). AS changes are frequently observed in cancer and are starting to be recognized as important signatures for tumor progression and survival. AS has the capacity to radically alter the composition and function of the encoded protein, this is why it represents an interesting research path in order to develop new therapies.

1.3.4.1. Affecting Hallmarks of Cancer

AS is involved in several characteristics of cancer cells described by Hanahan and Weinberg, such as resistance to cell death, angiogenesis, activation of invasion or the formation of metastases (Hanahan and Weinberg 2011). For instance, FAS gene, which encodes for a cell receptor, can produce one isoform with a pro-aptotic function while another isoform has the opposite function (Miura, Fujibuchi, and Unno 2012). When exon 6 of FAS is excluded, the protein stays soluble and the signal of apoptosis
is not transmitted anymore (Izquierdo et al. 2005). BCL-X isoforms, which are the result of alternative 5' splice sites at exon 2, are also a famous case of antagonist isoforms (Boise et al. 1993). BCL-XS has pro-apoptotic functions while BCL-XL has anti-apoptotic functions. In colorectal cancer (CRC), overexpression of the long isoform of SYK significantly suppresses the proliferation and metastasis of CRC cells, while overexpression of the short isoform does not (Ni et al. 2016). The vascular endothelial growth factor A (VEGF-A) gene encode for proteins involved in angiogenesis, and this gene gives rise also to two transcripts with opposite functions (Guyot and Pagès 2015; Bates et al. 2002). Exon 8 is the key determinant of isoform switching between a pro-angiogenic and anti-angiogenic isoform. When exon8a is included, exon 8b is excluded and the corresponding protein is pro-angiogenic. In contrast, when exon 8b is included, exon 8a is excluded and the protein is anti-angiogenic. Human cyclin D1 (CCND1) is expressed as two isoforms derived by alternate RNA splicing, termed D1a and D1b, which differ for the inclusion of intron 4 in the D1b mRNA. Cyclin D1b displays relatively higher oncogenic potential and was involved in the formation of metastases (Augello et al. 2015). An exon inclusion change in NUMB has been shown to promote cell proliferation (Bechara et al. 2013). Similarly, an exon-skipping event in MST1R (RON) has been related to the acquisition of cell motility during cancer cell invasion (Ghigna et al. 2005). Although I have mentioned a lot of examples, this is not an exhaustive list of all AS that are playing a role in oncogenesis.

In 2008, Thorsen & al mark the start of large-scale studies for alternative splicing in cancer (Thorsen et al. 2008). Using 102 normal and cancer tissue samples, from colon, bladder and prostate, they identified several AS cancer specific events in these tissues. The following year, Venables & al followed this path and found 288 AS in ovarian breast cancer and 232 AS in breast cancer compared to normal tissue using high-throughput RT-PCR (Julian P. Venables et al. 2009). One year later, using cancer cell lines, a transcriptome wide study based on Junction Arrays discovered 181 splice events occurring during breast cancer, amongst which some are specific to breast cancer subtype (Lapuk et al. 2010). This was also demonstrated later elsewhere using RNA-SEQ data from The Cancer Genome Atlas (TCGA) (Bjørklund et al. 2017). More recently, different AS patterns in tumors have been demonstrated again using TCGA dataset (Sebestyén, Zawisza, and Eyras 2015; Tsai et al. 2015; Y. Li et al. 2017). Trincado & al highlighted the fact that transcript isoform signatures appear especially
relevant to determine lymph node invasion and metastasis (Trincado, Sebestyén, et al. 2016). In another study, it was showed that AS deregulation in cancer often impacted functional protein domains that are frequently mutated in tumors and potentially affected protein-protein interaction in cancer pathways. They introduced the concept of cancer alternative splicing changes (CASCs) and proposed that these particular events could be oncogenic drivers on their own (Climente-González et al. 2017). In 2018, a comprehensive analysis of alternative splicing across tumors from 8705 patients confirmed tumors have up to 30% more alternative splicing events than normal samples (Kahles et al. 2018). They also concluded that AS in tumors leads to cancer-specific RNA transcripts that are translated into tumor-specific proteins with the potential for Major Histocompatibility Complex (MHC) presentation and, hence, could be a promising target for new immunotherapy treatments.

For the past 10 years, thanks to transcriptome wide study, several abnormal altered transcripts have been discovered in a plethora of cancer. Mutations in splicing factors or direct mutations in splicing site or regulatory elements may be the reason for these changes at the transcriptomic level. However, there is still work to be done in order to characterize their functional impact and relevance to tumorigenesis. Below there is a table (Table 1-1) summarizing some of the known oncogenic AS isoforms and their function in cancer biology.
<table>
<thead>
<tr>
<th>Gene name</th>
<th>Splicing event type</th>
<th>Splicing structure and function</th>
<th>Tumor type</th>
<th>Experimental evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BN1</strong></td>
<td>EB</td>
<td>pro-apoptotic</td>
<td></td>
<td>BN1</td>
</tr>
<tr>
<td><strong>BCL2L1</strong></td>
<td>A5</td>
<td>anti-apoptotic</td>
<td></td>
<td>BN1</td>
</tr>
<tr>
<td><strong>BCL2L11</strong></td>
<td>EB</td>
<td>pro-apoptotic</td>
<td></td>
<td>BN1</td>
</tr>
<tr>
<td><strong>CASP2</strong></td>
<td>EB</td>
<td>anti-apoptotic</td>
<td></td>
<td>BN1</td>
</tr>
<tr>
<td><strong>CCND1</strong></td>
<td>IR</td>
<td>pro-proliferative</td>
<td></td>
<td>BN1</td>
</tr>
<tr>
<td><strong>CDH4</strong></td>
<td>EB</td>
<td>mesenchymal</td>
<td></td>
<td>BN1</td>
</tr>
<tr>
<td><strong>EMA4</strong></td>
<td>EB</td>
<td>epithelial</td>
<td></td>
<td>BN1</td>
</tr>
<tr>
<td><strong>FA5</strong></td>
<td>EB</td>
<td>anti-apoptotic</td>
<td></td>
<td>BN1</td>
</tr>
<tr>
<td><strong>FGFR</strong></td>
<td>MXE</td>
<td>tumor-suppressive, epithelial</td>
<td></td>
<td>BN1</td>
</tr>
<tr>
<td><strong>HER2</strong></td>
<td>EB</td>
<td>pro-proliferative, pro-invasive</td>
<td>mesenchymal</td>
<td>BN1</td>
</tr>
<tr>
<td><strong>HRAS</strong></td>
<td>EB</td>
<td>pro-proliferative, pro-invasive</td>
<td>mesenchymal</td>
<td>BN1</td>
</tr>
<tr>
<td><strong>KLF6</strong></td>
<td>A5/SB</td>
<td>tumor suppressive</td>
<td></td>
<td>BN1</td>
</tr>
<tr>
<td><strong>MCL1</strong></td>
<td>EB</td>
<td>pro-apoptotic</td>
<td></td>
<td>BN1</td>
</tr>
<tr>
<td><strong>MKNK2</strong></td>
<td>EB</td>
<td>anti-apoptotic, anti-proliferative</td>
<td>anti-proliferative</td>
<td>BN1</td>
</tr>
</tbody>
</table>
As I mentioned before, EMT is an important process in the tumor progression. EMT program relies not only on transcriptional modifications but also extensive changes in alternative splicing are observed (Grosso, Martins, and Carmo-Fonseca 2008; Shapiro et al. 2011). For example, the value of considering AS as a marker for the formation of metastases has been showed by Dutertre & al (Dutertre, Vagner, and Auboeuf 2010). They identified expression of alternative splicing exons associated to tumors with different metastatic capabilities. They highlight the fact that some exons were associated with dissemination of primary tumor cells to sites of pulmonary metastasis. Finally, differentially spliced variant transcripts identified in their mouse 4T1 primary mammary tumor model was associated with poor prognosis in a large clinical cohort of patients with breast cancer. Moreover, AS were also associate with EMT elsewhere in large cohort of tumors (Danan-Gotthold et al. 2015) or in large panel of cancer cells.
lines with different invasive properties (Lapuk et al. 2010; Neve et al. 2006; Kao et al. 2009).

Recent studies started to suggest that the switch of an exon can drive a more mesenchymal state and/or leads to cancer progression (Ji Li et al. 2018; Ranieri et al. 2016; Brown et al. 2011). The fact that some AS are sufficient to trigger or impair EMT by themselves, suggest that alternative splicing is a key regulatory mechanism during EMT. Shapiro & al were among the first to identify an EMT-Driven alternative splicing program that occurs in human breast cancer and modulates cellular phenotype (Shapiro et al. 2011). This complex network of interaction occurring during EMT was further described by Yang & al (Y. Yang et al. 2016a). Gradually, a landscape of alternately modified exons began to appear and studies on the specific functions of these isoforms have started to emerge. Several functions of AS have been well described (Table 1-2) as FGFR2, CTNND1, CD44 that will be detailed further.

FGFR2 is a transmembrane receptor tyrosine kinase of the fibroblast growth factor receptor family. The ligands of the fibroblast growth factor family (FGFs) are responsible of its activation (Turner and Grose 2010). Two mutually exclusive alternative exons control this behavior. Exons, IIIb and IIc, encodes for a part of the third extra-cellular immunoglobulin-like domain of FGFR2. Exon IIIb is known to be predominantly included in epithelial cells, whereas exon IIc is limited to mesenchymal cells (Warzecha et al. 2010; Carstens et al. 1997; Carstens, Wagner, and Garcia-Blanco 2000). The mesenchymal splicing variant of the transmembrane receptor FGFR2 can recognize FGF-2 as a ligand whereas the epithelial isoform has less affinity for it, which will affect differentiation, growth and capacity to invade of cells (X. Zhang et al. 2006). During EMT, we can observe a switch between FGFR2-IIIb and FGFR2-IIc isoforms (Gil-Diez De Medina et al. 1999; Savagner et al. 1994). Newly, the specific expression of the FGFR2-IIIc variant was shown to be sufficient to promote cell migration, invasiveness and proliferation in response to FGF-2 (Sanidas et al. 2014).

CTNND1 encodes the p120-catenin (p120) protein which regulates transmembrane cell-cell adhesion receptors called cadherins. For instance, it is known to stabilize E-Cadherin (a well-known marker of epithelial state). Besides cell-cell interactions, p120 regulates the activity of Rho family GTPases and downstream
cytoskeletal dynamics (Davis, Ireton, and Reynolds 2003). Yanagisawa showed that is was able to promote invasion and cell motility (Yanagisawa et al. 2008).

During EMT, there is a switch between two isoforms from a short to long isoform (Y. Zhang et al. 2014). Alternative exons, that are skipped in epithelial cells, are included in the mesenchymal cells. The lack of these exons is responsible to the absence of a coiled-coil domain in the epithelial variant, domain that stabilizes RhoA binding and inhibits RhoA activity, resulting in an increase of migration and cell invasiveness (Epifano, Megias, and Perez-Moreno 2014; Keirsebilck et al. 1998). Interestingly, CTNND1 isoform has recently been involved in a specific signature for basal-like breast cancer, one of the most aggressive and deadly breast cancer subtype with mesenchymal features (Sebestyén, Zawisza, and Eyras 2015).

CD44 gene encodes for a transmembrane glycoprotein involved in many cellular processes such as cell survival, migration and proliferation (Zöller 2011; Prochazka, Tesarik, and Turanek 2014). CD44 has a high structural heterogeneity associated to the presence of ten alternatively spliced exons in its coding sequence, giving rise to a plethora of isoforms. The CD44 transcript is composed of 20 exons in human, including 10 variable exons (v1-v10) and 14 constitutive exons (exons 1-5 and 16-20). The inclusion of the variable exons leads to an increase of the size of the extracellular region of CD44, providing new interaction sites for additional molecules (Bennett et al. 1995). Overexpression of CD44 v6 variant is associated with poor patient prognosis in gastric cancer progression (Fang et al. 2016) whereas expression of CD44 v10 isoform correlates with anti-metastatic properties in pancreatic cancer (Navaglia et al. 2003). During EMT, a switch occurs from an isoform (including v8-v10 exons) to a shortened isoform. It has even been reported that this switch is required to trigger EMT, showing how strong the link between alternative splicing and EMT is (Brown et al. 2011). Initially, the mesenchymal splice variant was associated to the formation of invadopodia, increasing cell migration (C. Chen et al. 2018). Recently, Müller & al describe a new function where the mesenchymal isoform mediates the endocytosis of iron bound hyaluronates in tumors. In this way, iron operates as a metal catalyst to demethylate repressive histone marks that govern the expression of mesenchymal genes and show this mechanism is enhanced during EMT (Müller et al. 2020).
Other alternative splicing have been shown to play an important role in EMT/tumor progression (H. Lu et al. 2013; Tripathi et al. 2019; Itoh et al. 2017; Braeutigam et al. 2014), as ENAH, EXO70, TAK1, ARHGEF11, SEC13A, SLK (among them, some are cited in Table 1-2) but it remains to see if they can drive an EMT, be considered as EMT alternative splicing drivers or Cancer-Associated Splicing Changes (CASCs) as previously defined by Climente-Gonzalez & al. (Climente-González et al. 2017). This idea has already emerged among some. For instance, in 2018, Li & al., showed that an AS splicing switch in FLNB, an actin-binding protein, promotes the mesenchymal cell in human breast cancer (Ji Li et al. 2018).

<table>
<thead>
<tr>
<th>Gene (other names)</th>
<th>Function</th>
<th>Epithelial splicing</th>
<th>Domain affected/functional difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGFR2 (p120-catenin)</td>
<td>Transmembrane receptor tyrosine kinase</td>
<td>Exon IIIb</td>
<td>Confers ligand binding specificity</td>
</tr>
<tr>
<td>CTNND1</td>
<td>Delta-catenin; regulator of cell adhesion and signaling</td>
<td>Skipped</td>
<td>Coiled-coil domain; stabilizes interaction with RaA</td>
</tr>
<tr>
<td>CD44</td>
<td>Cell-surface glycoprotein involved in cell adhesion and migration</td>
<td>Exons v8–v10</td>
<td>Extra-cellular membrane proximal region; creates a heavily glycosylated stalk</td>
</tr>
<tr>
<td>ENA-H (Mena)</td>
<td>Regulator of actin dynamics</td>
<td>Included</td>
<td>Ena/Vasp homology domain; contains a phosphorylation site that may disrupt actin binding</td>
</tr>
<tr>
<td>NUMB</td>
<td>Complex protein implicated in many roles including cell migration and adhesion</td>
<td>Included</td>
<td>Phosphotyrosine binding domain; the encoded peptide confers localization to the plasma membrane</td>
</tr>
<tr>
<td>FLNB</td>
<td>F-actin cross-linking protein</td>
<td>Included</td>
<td>Contributes to the hinge domain; allows for more rigid actin branching</td>
</tr>
<tr>
<td>DNM2</td>
<td>GTPase that binds cytoskeletal proteins</td>
<td>Included</td>
<td>Pleckstrin homology domain; affects subcellular localization</td>
</tr>
<tr>
<td>TCF7L2 (Tcf4)</td>
<td>Transcription factor involved in Wnt signaling pathway</td>
<td>Included</td>
<td>Differential activation of Wnt/β-catenin target genes</td>
</tr>
<tr>
<td>BAIAP2 (Irs53)</td>
<td>Cdc42 effector protein involved in lamellipodia and filopodia formation</td>
<td>Included</td>
<td>Pentultimate exon with stop codon; differentially phosphorylated in response to IGF-1</td>
</tr>
<tr>
<td>MAP3K7 (Tak1)</td>
<td>Kinase that mediates TGF-β and BMP signal transduction</td>
<td>Included: Skipped</td>
<td>Peptide encoded by downstream exon is required for interaction with Tab2/3</td>
</tr>
<tr>
<td>ARHGPAP17 (Rich1)</td>
<td>GTPase-activating protein involved in maintenance of the tight junction</td>
<td>Skipped</td>
<td>Part of proline rich domain</td>
</tr>
<tr>
<td>MAG11 (Baiap1)</td>
<td>Scaffolding protein associated with complexes at the inner plasma membrane</td>
<td>Skipped</td>
<td>Encodes peptide between the two WW domains</td>
</tr>
<tr>
<td>LRRFIP2</td>
<td>Involved in activation of Wnt signaling</td>
<td>Skipped</td>
<td>Predicted coiled coil domain; encoded peptide may enhance interaction with Dvl3</td>
</tr>
<tr>
<td>SCRIB</td>
<td>Scaffolding protein associated with tight junctions and cell polarity</td>
<td>Skipped</td>
<td>Encodes a peptide proximal to the first PDZ domain</td>
</tr>
<tr>
<td>EPB41L5 (Ymo1)</td>
<td>A FERM protein that interacts with Crumbs complex to regulate cell architecture</td>
<td>Short isoform</td>
<td>Paxillin-binding domain; enhances focal adhesion complexes</td>
</tr>
<tr>
<td>RALGPS2</td>
<td>A guanine nucleotide exchange factor involved in cytoskeleton reorganization</td>
<td>Included</td>
<td>Between a PoxP motif and a pleckstrin homology domain; may influence GEF activity</td>
</tr>
<tr>
<td>ITGA6</td>
<td>Alpha subunit of integrin, a lamnin receptor</td>
<td>Included</td>
<td>Light chain and cytoplasmic domain; changes C-terminus sequence</td>
</tr>
<tr>
<td>SLK</td>
<td>STE20-like kinase with a role in promoting cell motility</td>
<td>Included</td>
<td>Predicted coiled-coil domain; may specify interaction partners</td>
</tr>
<tr>
<td>ARHGEF11 (PDZ-RhoGEF)</td>
<td>RhoA-specific guanine nucleotide exchange factor</td>
<td>Skipped</td>
<td>C-terminus; may influence homodimerization or interaction with PAK4 and LARG</td>
</tr>
</tbody>
</table>

**Table 1.2 Examples of genes affected by alternative splicing during EMT.**

Genes are presented with their function, the type of alternative splicing event that occurs, and the domain/function that will be affected (adapted from Carstens & Warzecha, 2012).
Importantly, more and more therapeutic strategies aim to restore normal splicing patterns in cells harboring genetic disorders (Liang et al. 2020; L. M. Urbanski, Leclair, and Anczuków 2018). The first approval from the US Food and Drug Administration for a therapy based on RNA interference (RNAi), with patisiran (a drug targeting a rare condition that can impair heart and nerve function) was only released very recently in 2018. Antisense oligonucleotides (ASOs) represent a compelling therapeutic approach to target exon leading to a pathologic state. Notably, application of antisense oligos (ASOs) are currently in clinical trials for Duchenne muscular dystrophy and spinal muscular atrophy (Havens and Hastings 2016; Pires et al. 2017). Recently, an in vitro cancer cell model, Hong & al identified the effect of an ASO (AZD9150), reducing signal transducer and activator of transcription 3 (STAT3) in lung cancer and lymphomas (D. Hong et al. 2015). ASO (AZD4785) targets the KRAS gene and was showed to diminish its proliferative activity in some cancers (J. C. Lin 2018). ASO-mediated exclusion of MDM4 exon 6 leads to a decrease in MDM4 abundance through the AS-NMD pathway, which enhances the drug sensitivity and apoptosis of melanoma cells (Dewaele et al. 2016). While RNA interference approaches are still in its premise, new successes for splice-switching oligonucleotides (SSO) are emerging. SSO approaches were used to target ERG oncogene (L. Li et al. 2020) or modulates MKNK2 alternative splicing, in prostate cancer and glioblastoma (Mogilevsky et al. 2018), respectively. Identifying and characterizing the function of ASE involved in EMT, and therefore tumor progression, can thus offer new opportunities for modern therapeutic strategies based on RNA.
1.3.4.3. ACROSS CANCER TYPES

Splicing factors are important modulators of RNA processing. Alternative splicing is frequently regulated by these trans-acting splicing factors, which bind to sequence motifs that are associated with the promotion (enhancers) or repression (silencers) of splicing. For instance, SR proteins, hnRNPs can act as both oncoproteins and tumor suppressors (Dvinge et al. 2016). Tumor progression can be boosted by...
different regulation of their expression (L. Urbanski and Leclair 2019) or mutations in their sequences (Yoshida, Kenichi Ogawa 2014) (Figure 1-14) resulting in a drastic change of the underlying transcriptomic programs. Seiler et al. report that 119 splicing factor genes (over 400 SF) carry putative driver mutations over 33 tumor types in TCGA (Seiler et al. 2018). Among all the tumor types analyses, bladder carcinoma and uveal melanoma had significantly higher rates of splicing factor driver mutations than would be expected by chance, suggesting that splicing deregulation is an important hallmark for these tumors. These mutations were associated with deregulation of immune response, cell cycle checkpoint, DNA damage response (DDR), and metabolism.

**Figure 1-15 Recurrent splicing factor alterations detected in cancer.**

Genomic alterations include expression changes and recurrent somatic mutations. Splicing-factor upregulation are depicted in red, downregulation in blue, and somatic mutations in green. Several splicing factors can be found both upregulated and downregulated in tumors of the same tissue, suggesting that distinct splicing-factor genomic alterations are associated with distinct tumor subtypes within the same tissue. (adapted from Urbanski, 2018)
In another study, co-regulated SFs were associated with aggressive breast cancer phenotypes and enhanced metastasis formation (Koedoot et al. 2019). Table 1-3 gives a more complete overview of SFs that are functionally linked to cancer.

<table>
<thead>
<tr>
<th>Splicing factor</th>
<th>Downstream dysregulated isoforms that are functionally linked to cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESRP1 and ESRP2</td>
<td>Promote an epithelial splicing programme to regulate EMT^101</td>
</tr>
<tr>
<td>hnRNP A1</td>
<td>Contributes to aerobic glycolysis in cancer by promoting the expression of specific isoforms of pyruvate kinase (PKM2 isoform^101^ and MAX (the delta MAX isoform)^101^</td>
</tr>
<tr>
<td>hnRNP A2</td>
<td>Contributes to aerobic glycolysis in cancer by promoting the expression of a specific isoform of pyruvate kinase (PKM2 isoform^101^), increases breast cancer cell invasion by promoting the expression of a specific isoform of TPS3BP2 (REF 195)</td>
</tr>
<tr>
<td>hnRNP A2/B1</td>
<td>Acts as an oncogenic driver in glioblastoma by regulating splicing of several tumour suppressors and oncogenes, including RON^101^</td>
</tr>
<tr>
<td>hnRNP H</td>
<td>Contributes to survival of gliomas and invasion by promoting the expression of specific isoforms of IG20 and RON, respectively^101^</td>
</tr>
<tr>
<td>hnRNP K</td>
<td>Serves as a tumour suppressor in leukaemia. Deletion is associated with aberrant p21 and C/EBP expression (although mechanistic links to splicing and gene expression are unclear)^51</td>
</tr>
<tr>
<td>hnRNP M</td>
<td>Contributes to EMT in breast cancer and increases metastasis in mice by promoting the expression of a specific isoform of CD44 (CD44S)^97</td>
</tr>
<tr>
<td>PRPF6</td>
<td>Promotes cell proliferation in colon cancer by altering splicing of genes associated with growth regulation, including the kinase gene ZAK^101^</td>
</tr>
<tr>
<td>PTB (PTBP1)</td>
<td>Contributes to aerobic glycolysis in cancer by promoting the expression of a specific isoform of pyruvate kinase (PKM2 isoform^101^). Also promotes the expression of an isoform of the deubiquitinating enzyme encoding gene USP5 (REF 195), that has been shown to promote glioma cell growth and mobility (although the mechanism underlying this phenotypic association is not resolved)</td>
</tr>
<tr>
<td>QKI</td>
<td>Acts as a tumour suppressor by regulating alternative splicing of NUMB in lung cancer cells^34</td>
</tr>
<tr>
<td>REFOX2</td>
<td>Promotes a mesenchymal splicing programme to regulate EMT^102</td>
</tr>
<tr>
<td>RBM4</td>
<td>Acts as a tumour suppressor by promoting the pro-apoptotic isoform BCL-XL of BCL2L1 and opposing the pro-tumorigenic effects of SRSF1 on mTOR activation^100</td>
</tr>
<tr>
<td>RBM5, RBM6 and RBM10</td>
<td>RBM5 modulates apoptosis by regulating alternative splicing of CASP2 (REF 200) and FAS^101^ . RBM5 or RBM6 depletion has an opposite effect to RBM16 depletion, as these factors antagonistically regulate the alternative splicing of NUMB^101^</td>
</tr>
<tr>
<td>SRSF1</td>
<td>Promotes an isoform of the kinase MK2 that promotes eIF4E phosphorylation independently of MAPK signalling^101^ . In the context of breast cancer, SRSF1 overexpression promotes alternative splicing of BIM and BIM1. (REF 40) to promote the expression of isoforms that lack pro-apoptotic functions</td>
</tr>
<tr>
<td>SRSF3</td>
<td>Regulates alternative splicing of TP53 (REF 43) such that SRSF3 loss promotes expression of p53β, an isoform of p53β that promotes p53-mediated senescence</td>
</tr>
<tr>
<td>SRSF6</td>
<td>Promotes expression of isoforms of the extracellular matrix protein tenascin C that are characteristic of invasive and metastatic skin cancer^101^ , contributing to epithelial cell hyperplasia</td>
</tr>
<tr>
<td>SRSF10</td>
<td>Promotes cell proliferation and colony formation in vitro and increases tumorigenic capacity of colon cancer cells in mice by inducing expression of a specific isoform of BCLAF1 (BCLAF1-L)^101^</td>
</tr>
</tbody>
</table>

**Table 1.3 Unmutated SFs that function as proto-oncogenes or tumor suppressors.**

This table describes unmutated splicing factors with the downstream effect of their dysregulated isoforms highlighting their functional link with cancer. (adapted from Divinge, 2016).
When we look individually to these SFs, SR splicing factor 1 (SRSF1; also known as ASF and SF2) is upregulated in several cancers, including lung, colon and breast cancer (Anczuków et al. 2012; Karni et al. 2007; Anczuków et al. 2015). SRSF1 acts synergistically with MYC, and their co-expression correlates with higher tumor grade and decreased survival in breast and lung cancer patients (Anczuków et al. 2015). Several alternative isoforms regulated by SRSF1 are implicated in cancer-relevant processes as apoptosis (e.g., BCL2L1, BCL2L11, BIN1), cell growth (RPS6KB1), cell survival (MKNK2), or motility (RON) (L. M. Urbanski, Leclair, and Anczuków 2018). Interestingly, the overexpression of one such isoform, exon-9-included CASC4, increased acinar size and proliferation, and decreased apoptosis, highlighting the strong impact a single isoform can have on the phenotype (Anczuków et al. 2015). Another example is QKI, that is thought to play a role of tumor suppressor in lung cancer, in which it is commonly downregulated, in part by regulating the alternative splicing of NUMB (Zong et al. 2014).

In breast, SRSF4, SRSF6 or TRA2b promotes mammary cell proliferation and invasion and it seems that TRA2b, regulated by MYC, plays a role in the formation of metastasis (S. H. Park et al. 2019). Numerous SFs (hnRNPK hnRNPA2/B1, SRSF6, SRSF3) are frequently overexpressed in breast and other tumors (L. M. Urbanski, Leclair, and Anczuków 2018). Notably, hnRNPM promotes breast cancer metastasis by activating the switch of alternative splicing that occurs during epithelial–mesenchymal (Yilin Xu et al. 2014) and has been shown to cooperate with ESRP proteins (S. E. Harvey et al. 2018). In addition, RBFOX2 was shown to be repressed in breast and ovarian cancers and associated with many abnormal alternative splicing events (Julian P. Venables et al. 2009; J. P. Venables et al. 2013). Among others, which I will detail below, the splicing factor RBFOX2 has also been linked with EMT (Danan-Gotthold et al. 2015; Lapuk et al. 2010; Shapiro et al. 2011; Julian P. Venables et al. 2013).

In summary, alternative splicing regulation in tumor progression should not be seen like a mechanism led by a single major actor, but as a complex network of interactions between different players.
As mentioned just earlier, the splicing factor RBFOX2 regulates EMT, and have many splicing targets in breast, pancreatic, and colon tumors (J. P. Venables et al. 2013; Braeutigam et al. 2014; Lapuk et al. 2010; Danan-Gotthold et al. 2015). The loss of RBFOX2 in mesenchymal cells leads to a partial reversion of the epithelial phenotype (Shapiro et al. 2011). Notably, it was involved in alternative splicing of FGFR2 discussed earlier (Hovhannisyan and Carstens 2005) and was implicated in the survival of human embryonic stem cells (Yeo et al. 2009). In the clinical field, ESRP1/RBFOX2 ratio value was linked to a higher risk of metastasis in early breast cancer patients (Fici et al. 2016).

However, the major EMT splicing regulators are the epithelial specific regulatory proteins 1 and 2 (ESRP1/2). These regulators are the most downregulated in multiple models of EMT whereas RBFOX2 is slightly increase (Warzecha et al. 2010). ESRP-targeted transcripts undergo a switch from epithelial to mesenchymal isoforms (Warzecha and Carstens 2012). They affect splicing of target genes involved in EMT, including CD44, ENAH, FGFR2, and RAC1, playing role in cell-cell junction adhesion, cytoskeleton, actin dynamics and extracellular matrix (ECM) (Dittmar et al. 2012; Shapiro et al. 2011; C. Chen et al. 2018; Warzecha et al. 2010; Y. Yang et al. 2016a). They are also very low expressed in claudin low tumors and basal B cells lines, both harboring mesenchymal features, with high invasive properties (Lapuk et al. 2010; Neve et al. 2006; Kao et al. 2009). ESRP1/2 were shown to be upregulated in normal epithelium but downregulated in invasive fronts (Ishii et al. 2014). Interestingly, using single-cell transcriptomics, it was also shown that cells expressing a partial EMT program were spatially localized at the leading edge of primary tumors in head and neck cancer (Puram et al. 2017). More surprisingly, higher expression of ESRP1, which is downregulated during EMT, correlated with a worse prognosis for ER+ breast cancer (Gökmen-Polar et al. 2019) or ovarian cancer (Jeong et al. 2017).

Other splicing factors were suggested to cooperated with ESRP proteins as hnRNPM whose splicing levels of coregulated exons were associated with breast cancer patient survival (S. Harvey et al. 2018). Another example of cooperation, is the RNA binding motif protein 47 (RBM47), which is downregulated during EMT (Y. Yang et al. 2016a). In breast cancer cells, RBM47, via its ability to modulate splicing, has
been demonstrated as a potential metastasis formation inducer (Vanharanta et al. 2014). Its down regulation was also observed during colorectal cancer progression (Rokavec et al. 2017). In a model of lung adenocarcinoma, it was also proposed as a tumor-suppressor (Sakurai et al. 2016). Based on TCGA data, RBM47 appears to be lowly expressed in Claudin-Low and basal-like breast tumors, which are the most aggressive tumors (Vanharanta et al. 2014). Recently, the A-Kinase Anchor Protein (AKAP8) was reported as a splicing regulatory factor that inhibits EMT and breast cancer metastasis (X. Hu et al. 2020). MBNL and CELF proteins have also been implicated in EMT (Shapiro et al. 2011) highlighting the complexity of cancer-associated splicing dysregulation. Figure 1-15 displays a resume of major SFs regulated during EMT, and examples of splicing switches.

![Figure 1-16 Splicing Factors changes and AS events during EMT.](image_url)

Transition of cells between epithelial (blue) and mesenchymal (red) states is associated with shifts in abundance and/or activity of splicing factors. These modifications have an impact on the regulation of an alternative splicing program during EMT. Some alternative splicing events impacted are given as examples (CD44, FGFR2, MENA, RON, RAC1) (adapted from Neumann, 2017).
1.4. HIGH-THROUGHPUT TECHNOLOGY, ANALYSIS, MODELS AND RESOURCES FOR CANCER RESEARCH

At the end of the 20th century, molecular biology knew an unprecedented revolution. In the early 1980s, Kary Mullis invented the polymerase chain reaction (PCR) technology for amplifying DNA which was first published in the journal Science in 1985 (Kaunitz 2015). Then, the 1990s were the witnesses of a wave of applications based on discoveries in the previous decades and the arrival of huge volumes of genomic/transcriptomic data from automated sequencing and DNA microarray technologies that biologists couldn’t treat alone. It was from there that computer science began to enter laboratories in order to help researchers to store and process this incredible amount of data being produced. Two new disciplines, genomics and bioinformatics, were born. In parallel, an initial rough draft of the human genome was released in 2000. Now, technologies are mature enough to ask several questions about changes in DNA sequences, modification of gene expression, epigenetic and proteins variations in different tissues or contexts (normal or tumor cells) at an affordable price and a shortened period of time.

1.4.1. HIGH-THROUGHPUT TECHNOLOGY

1.4.1.1. MICROARRAYS

The microarrays emerged in the late 1990s. This system consists in a simple surface of glass or plastic where a collection of microscopic DNA spots (probes) are attached. The single-strand oligonucleotides probes are chosen to be specific to a DNA region or transcript. They can be used to detect DNA (as in comparative genomic hybridization CGH) or detect RNA (as in cDNA after reverse transcription). The RNA is extracted from a sample, amplified and labeled with a fluorochrome before being hybridized on the chip. After the fluorochrome has been stimulated at the appropriate wave length, the signal intensity of the fluorescence light allows quantifying the expression levels of targets which are attached to the probe. Hybridization-based approaches are relatively inexpensive, but several limitations exist. Their design relies of the knowledge we have of the genome making them
impossible to discover novel transcripts. Due to cross hybridization (hybridization
between sequences that are not strictly complementary) background levels of the
signal are high. Thus, genes that are lowly expressed in a sample could not be
distinguished from background chip level, and overexpressed genes may lead to signal
saturation limiting their exact quantification. Moreover, comparing expression levels
across different experiments is often difficult and can require complicated
normalization methods.

Affymetrix, leader in the market, proposed three kinds of chip to study
transcriptome. Classical microarray, with probes targeting only the 3’ region of a gene
in order to study their expression. This technology (mostly used in large-scale
sequencing project of a population) is inappropriate to study alternative splicing since
it is not detecting splice junctions. In contrast, their two other products can do gene
expression and AS analysis. Exon arrays are designed with probes matching exons,
but the number of probes by exon is low and the results depend very strongly on the
quality of the hybridization and the fluorescent labeling. The last technology, named
Junction arrays, has the advantage of having probes at exon-exon junctions making
the AS analysis more reliable.
Yet, with the apparition of high-throughput sequencing (RNA-seq), these type of
hybridization approaches have become obsolete (see definition in the next section).

1.4.1.2. HIGH-THROUGHPUT SEQUENCING: FOCUS ON RNA-SEQ

Historically, the sequencing method of reference was Sanger sequencing, a
method of DNA sequencing based on the selective incorporation of chain-terminating
dideoxynucleotides by DNA polymerase during in vitro DNA replication. Then the
emergence of technologies called High-Throughput Sequencing (HTS) or Next
generation sequencing (NGS) allowed researchers to sequence DNA and RNA much
faster and cheaper. Since the TCGA RNA-SEQ data analyzed in this thesis was
produced from an Illumina HiSeq platform, I will present only this technology, but it is
worth mentioning that other technologies exist with different chemistry (Roche 454,
SOLiD, IonTorrent).
The Illumina next-generation sequencing (NGS) method is based on sequencing-by-synthesis (SBS), and reversible dye-terminators that enable the identification of single bases as they are introduced into DNA strands. First, the RNA extracted from a sample is broken into small fragments that are converted to DNA through reverse transcription. Once turned into cDNA, the molecules can be sequenced as regular DNA. Short sequences of nucleotides, called adaptors, are attached to both fragments ends. These adaptors will be used to anchor the fragment on one end of the flow cell (Figure 1-17 A). The second step consists in the formation of a cluster of sequences (Figure 1-17 B). The DNA fragments in the sequencing library are fixed to adapters attached to the surface of the flow cell when they pass through it. This fixation is possible only if one of the adapters added at the ends of the DNA fragment of the library match the complementary sequence of the ones attached on the surface of the flow cell. Then, the other end of each fragment is folded over and binds to another adapter on the flow cell surface. The adapters on flow cell are used as a template to initiate synthesis of the complementary strand in a process called Bridge PCR. Multiple rounds of amplification are performed to obtain clusters containing approximately 1000 copies of the original single-stranded DNA molecule. The purpose of this process is to amplify the signal intensity of the base to meet the signal requirements for sequencing. The last step is the actual sequencing based on sequencing-by-synthesis (SBS). DNA polymerase and 4 dNTP with base-specific fluorescent markers are added to the reaction system (Figure 1-17 C). The 3′-OH of these dNTP are protected by chemical methods, which ensures that only one base will be added at a time during the sequencing process. Several cycles are performed during which the 4 fluorescently tagged nucleotides compete for addition to the growing nucleotides chain. After the addition of each nucleotide, unused reactants are washed away and clusters are excited by a laser. Fluorescence signal is recorded by optical equipment and recorded on a computer as a sequence of nucleotide bases. When the fluorescence signal is recorded, a chemical reagent is added to quench the fluorescence signal and remove the dNTP 3′-OH protective group, so that the next round of sequencing reaction can be performed.
Finally, the sequences obtained are small fragments of the genome (reads) that need to be aligned along a reference genome to be analyzed. Several types of sequencing library can be realized. For example, to have a better depth of sequencing, ribosomal RNA, that are the majority of RNA inside a cell, can be removed. This protocol is called Ribo-zero. Another famous protocol is the polyA+ protocol where only...
RNA with PolyA tail are retained, removing majority of non-coding RNA and non-mature RNA.

In RNA Sequencing, the number of reads that fall into a given gene or exon, quantifies its level of expression. Compared to microarrays, RNA sequencing does not need to design any probes since transcripts are directly sequenced. The gene expression level is estimated by counting data rather than fluorescent signals. The estimation is then much more precise without saturation. RNA-Seq enables scientists to perform several kinds of unbiased analyses at the gene, transcript and exon level.

Even if DNA sequencing is widely used in the field of cancer, I will not detail this technology because it was not performed in this work, but the principle of sequencing is the same. I will just mention some of these applications and the global principle of algorithms used. DNA sequencing can be used to discover polymorphisms, mutations and copy number variations (CNV - amplifications, deletions, rearrangements and copy-neutral loss of heterozygosity). After the reads have been mapped along the genome, a software dedicated to CNV detection can identify the reads that are over/under represented on a large portion of the genome or, in the case of short variations, inspect if the content of nucleotides from the reads sequenced differs from the reference sequence where they have been mapped.

1.4.2. BIOINFORMATIC ANALYSIS

Several bioinformatic analysis can be performed based on RNA Sequencing (Conesa et al. 2016). In this section, I will discuss the main investigations which have been carried out during this thesis work and I will not describe exhaustively all procedures that can be done (Figure 1-18).
Figure 1.18 RNA-seq computational analyses.

The graphic is divided into three parts: pre-analysis, core analysis, and advanced analysis. Gene Fusion discovery is an example of the core analysis of RNA-seq, while alternative splicing analyses are performed as part of the advanced analysis.
1.4.2.1. DIFFERENTIAL GENE EXPRESSION

In the field of transcriptomic, the most common application of RNA-seq is to estimate gene and transcript expression. The first level of analysis consists in the mapping of the reads over the genome. This approach will make possible to quantify raw counts of mapped reads. Using a file containing mapped reads, some programs as (HTSEQ-count (Anders, Pyl, and Huber 2015), or featureCounts (L. Yang, Smyth Gordon K, and Wei 2014) are dedicated to this purpose, but modern mapper as STAR (Dobin et al. 2013) have also now directly integrated this functionality in their algorithm. Raw read counts alone are not sufficient to compare expression levels among samples, a normalization process is always necessary.

Historically, simple approaches were first developed to normalize away the sequencing depth which is the most important factor for comparing samples. The measure RPKM (reads per kilobase of exon model per million reads) is a within-sample normalization method that will remove the feature-length and library-size effects (Mortazavi et al. 2008). Correcting for gene length is not necessary when comparing changes in gene expression within the same gene across samples, but it is necessary for correctly ranking gene expression levels within the sample to account for the fact that longer genes accumulate more reads. TPMs (transcripts per million), which effectively normalize for the differences in composition of the transcripts in the denominator rather than simply dividing by the number of reads in the library, are considered more comparable between samples of different origins (Conesa et al. 2016). TPM established itself as a reference and this metric is now the most frequently reported RNA-seq gene expression value.

Differential gene expression (DGE) analysis can be done using this metric with classical statistical test to determine whether the gene expression is statically different between groups. More advanced methods have been developed for DGE (TMM (Robinson and Oshlack 2010), DESeq (Anders and Huber 2010), PoissonSeq (Jun Li et al. 2012) and UpperQuartile (Bullard et al. 2010) which ignore highly variable and/or highly expressed features. Algorithm as voom implemented in Limma package (Ritchie et al. 2015) proposed to apply a linear model to log transformed data and a locally weighted regression (LOWESS) to weight the standard linear model. Two methods
(DESeq2 (Love, Huber, and Anders 2014) & edgeR (Robinson, McCarthy, and Smyth 2009)) have become very popular and use the negative binomial as the reference distribution and likelihood ratio test to assess the significance of the genes. However, several comparison studies point out that no single method is likely to perform favorably for all datasets.

Several methods based on k-mer counting in reads were also born recently (Sailfish (Patro, Mount, and Kingsford 2014), Kallisto (Bray et al. 2016), Salmon (Patro et al. 2017)). It turns out to be faster methods because they ignore the read alignment step. They were quickly accepted by the community due to their speed and the fact that they directly compute TPM values, a measure which is now used as a standard.

Differential gene expression analysis highlights genes differentially expressed between conditions. Nevertheless, it does not inform whether different transcripts are expressed or not. This is where the analysis of alternative splicing comes into play to provide a finer layer of information.

1.4.2.2. DIFFERENTIAL ALTERNATIVE SPlicing

Two major methodologies appear when it comes to detection of alternative splicing (Figure 1-19 a). The first approach, which I will not discuss in details, is based on quantification of the expression of transcript isoforms from the same gene and their comparison. An example that illustrates this case is the Cufflinks/CuffDiff2 algorithm (Trapnell et al. 2012) that estimates isoform expression first and then compares their differences. This kind of method suffers from the difficulty to accurately identify expression at the isoform level due to the intrinsic limitations of short-read sequencing.

The second approach (Figure 1-19 b) is based on specific algorithms focused on specific alternative splicing events. The so-called ‘exon-based’ approach skips the estimation of isoform expression and detects signals of alternative splicing by comparing the distributions of reads on exons and junctions of the genes between the compared samples. The advantage of exon or junction methods is their greater accuracy in identifying individual alternative splicing events.
These methods provide as output PSI (Percent Spliced In) values explained Figure 1-19 b. This index, ranging from 0% to 100%, is an estimation of the fraction of isoforms that include the exon. It indicates the efficiency of splicing for a specific exon into the transcript population of a gene. The dPSI (delta Percentage Spliced In), which is a difference in PSI, is then generally computed to see if there is a change in the level of inclusion of an exon between two conditions. The last ten year, several tools have been developed to measure AS. Their approach can vary based on how to calculate PSI values, the way they take reads into account, depending on whether it counts the
reads spanning the junctions or/and mapping the exon, and especially by the statistical method to estimate the robustness of the dPSI prediction. I provide a short summary of the latest tools (Sterne-Weiler et al. 2018; Trincado, Entizne, et al. 2016; Vaquero-Garcia et al. 2016; Kahlès et al. 2016; Y. Hu et al. 2013; Katz et al. 2010; Anders, Reyes, and Huber 2012; Brooks et al. 2011; Shen et al. 2014; Tapial et al. 2017; Tiberi and Robinson 2020) developed recently but which is not intended to be an exhaustive review (Table 1-4). This has been done elsewhere by Alamancos & al, who made a catalogue of all methods, appeared before 2015, to study splicing from high-throughput RNA sequencing data (Alamancos, Agirre, and Eyras 2014).

<table>
<thead>
<tr>
<th>Year</th>
<th>Name</th>
<th>Method / Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>2020</td>
<td>Bandits</td>
<td>DTU (differential transcript usage) - Bayesian hierarchical model, with a Dirichlet-multinomial structure</td>
</tr>
<tr>
<td>2018</td>
<td>Whippet</td>
<td>Junction-Kmer based – Splice Graph + likelihood function iteratively optimized by EM algorithm</td>
</tr>
<tr>
<td>2018</td>
<td>Suppa2</td>
<td>ΔPSI values as a function of the expression (TPM) of transcripts involved in the event</td>
</tr>
<tr>
<td>2016</td>
<td>MajiQ</td>
<td>Junction - Bayesian Psi modeling, and bootstrapping to report posterior psi and psi distributions for Local Spliced Variation (LSV)</td>
</tr>
<tr>
<td>2016</td>
<td>SpiAdder</td>
<td>Junction - Negative Binomial distribution + Generalized Linear Model (GLM)</td>
</tr>
<tr>
<td>2014</td>
<td>Rmats</td>
<td>Exon/Junction – (unpaired replicates) Binomial distribution for the estimation uncertainty in individual replicates + Normal distribution the variability among replicates + likelihood-ratio test</td>
</tr>
<tr>
<td>2014</td>
<td>Vast-Tools</td>
<td>Junction - Bayesian inference followed by differential analysis of posterior distributions</td>
</tr>
<tr>
<td>2013</td>
<td>Diffs splice</td>
<td>Exon/Junction – Graph-based, Jensen–Shannon divergence (JSD)</td>
</tr>
<tr>
<td>2012</td>
<td>DexSeq</td>
<td>Exon/Junction - Negative binomial distribution + Generalized Linear Model (GLM)</td>
</tr>
<tr>
<td>2011</td>
<td>Juncbase</td>
<td>Junction - Fisher exact test</td>
</tr>
<tr>
<td>2010</td>
<td>Miso</td>
<td>Exon/Junction - Bayes Factor (BF)</td>
</tr>
</tbody>
</table>

**Table 1.4 Overview of AS software since 2010.**

Description of softwares published for the study of alternative splicing since 2010. The table displays year of publication, name of the tool and methods/models used. This table is not an exhaustive review but demonstrate the strong activity of the field.
A drawback of these tools is that it can be greedy in memory and time consuming in order to be executed. This is why the latest tools often highlight in the title of their publications the fact that they are fast or can be operated on a simple computer. This is not a negligible point because the medical community does not always have a computer infrastructure for intensive computing and if we want to integrate AS algorithms in the clinical field, they must be as efficient as possible to give as soon as possible a diagnosis to the patient.

During my thesis work, I had to analyze a very large number of patients for splicing, I looked for a fast and precise AS detection algorithm. Several tools were tested and I found that Whippet (Sterne-Weiler et al. 2018) performed faster and was convenient to use. It also gave us the more reliable results based on a tested dataset with published results on alternative splicing. Whippet accurately and rapidly quantifies simple and complex AS events (Figure 1-20). It works in four main steps: (1) Based on annotation (and as an option, with already mapped reads file), it will collapse gene structure into non-overlapping exon intervals (nodes). It builds a Contiguous Splice Graph (CSG), where each nodes node has two boundaries. All 5’ splice site and 3’ splice site boundaries have k-mer indices (colored lines) that are used latter for spliced read alignment in step (3). (2) A single transcriptome full-text index in minute space (FM-Index) is built from concatenated CSG sequences. It will be used to efficiently find the number of occurrences of a pattern (k-mer) within the compressed text, as well as locate the position of each occurrence. (3) Raw reads of RNA-SEQ will be mapped directly to a CSG using previously indexed structure. K-mers from a simple read will map and join different nodes (GeneX, Node5 and GeneX, Node7). (4) For each node, a repertory of all AS event associated to a node will be built thanks to an AS event graph and each node will be associated to several paths based on this structure. Paths can include, or on the contrary, exclude the node. (5) All paths through the AS event are enumerated and quantified. Finally, Whippet will give a ratio of inclusion paths over all paths for the AS event, the Percentage Splice In (PSI) value.
Figure 1-20 Focus on Whippet: global methodology.

Whippet was used all along the thesis work to profile alternative splicing and compute PSI values. Its algorithm can be divided into five steps that are further explained in the text of the manuscript. (adapted from Sterne-Weiler 2018).
**k-mers** are defined as short subsequences of length k contained within a biological sequence. In DNA/RNA sequencing, the biological sequence considered is a read. As I mentioned before, whether it is the expression or alternative splicing, new k-mer-based methods have appeared in different layers of analysis (Bray et al. 2016; Sterne-Weiler et al. 2018). K-mer counting (Manekar and Sathe 2018) has spread widely in applications for genome and transcriptome assembly, quality control, error correction, multiple sequence alignment, and repeat detection (Pickett, Miller, and Ridge 2017; Mapleson et al. 2017; Durai and Schulz 2016). The big advantage of k-mer-based methods compared to alignment-based methods is the shorter computation times. Bypassing the mapping to a reference genome, make them a great solution to explore huge amount of data in order to classify them and discover new biological events. Recently a published software (Audoux et al. 2017), was able to detect numerous transcription and RNA processing events from RNA-SEQ using a k-mer approach. Another advantage of these approach is that they are not data-specific, and can be apply to a wide range of sequencing experiments as bisulfite sequencing, ChIP-Seq or whole-exome/genome sequencing.

There is therefore an emerging opportunity in terms of computational research and development to extract biological knowledge with these newly designed algorithms.

### 1.4.3. COMPUTATIONAL TECHNIQUES

The use of sequencing technologies brings a huge amount of data. Raw data can usually be pre-processed by bioinformatic tool of the field, but further investigations can be necessary to get out added value from these first analyzes, traduced by big matrices of expression or alternative splicing values for example.

I will describe some machine learning (ML) methods I applied during this thesis work. I will start with a simple case – hierarchical clustering – (Gentle, Kaufman, and Rousseuw 1991) followed by a description of a more advanced algorithm called Random Forest (Ho 1995), that can be used to classify individual in groups with common characteristics. Finally, I will present techniques that are commonly used to relate genomic data to survival in different groups of a population.

#### 1.4.3.1. HIERARCHICAL CLUSTERING
Hierarchical clustering analysis (HCA), is an unsupervised technique, meaning that it can infer patterns from a dataset without known reference, labels or outcomes. HCA groups similar features into clusters. The final result is a set of clusters, where each cluster is distinct from each other, and the features within each cluster are broadly similar to each other. It allows to build tree structures from data similarities in order to observe different classes.

In order to decide which clusters should be combined or split, a measure of dissimilarity between sets of observations is required. This is achieved by use of an appropriate metric (a measure of distance between pairs of observations) that can be for example, Euclidian distance or 1-Pearson Correlation. Then, the linkage criterion specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets. Several methods are available to achieve this goal:

- **Complete-linkage**: the distance between two clusters is defined as the longest distance between two points in each cluster.
- **Single-linkage**: the distance between two clusters is defined as the shortest distance between two points in each cluster. This linkage may be used to detect high values in your dataset which may be outliers as they will be merged at the end.
- **Average-linkage**: the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster.
- **Centroid-linkage**: finds the centroid of cluster 1 and centroid of cluster 2, and then calculates the distance between the two before merging.

Results can be quickly visualized using a dendrogram which is a tree-like diagram that records the sequences of merges or splits. One limitation of HCA is that it cannot handle a huge amount of data. For this reason, it’s worth filtering your data before applying this technique. Another drawback here, is that you cannot reuse knowledge we could have gained from previous datasets. This is where another type of model, called supervised model, can be useful because it will learn from pre-existing labeled data to classify new unlabeled data. This is the case of random forest discussed below.

**1.4.3.2. THE RANDOM FOREST ALGORITHM**
Random Forest (RF) is a recent algorithm which principles were first proposed by Ho in 1995 (Ho 1995). These last years, the use of such algorithms based on information theory has been made possible by the development of machine learning library (as Keras, SkicitLearn or Pytorch) that make these high-level algorithms accessible to a wider community (Paszke et al. 2017; Pedregosa et al. 2011; Chollet 2015). Nowadays, with explosion of biological data, ML techniques are becoming more and more popular in life sciences, including biology and medicine. I will explain the main idea being the algorithm and I will introduce some concepts used in ML.

Random Forest is based on decision trees. Decision trees are used to make prediction following several branches of “if… then…” decision splits - similar to the branches of a tree. At each branch, the feature thresholds that best split the (remaining) samples locally is found. The most common metrics for defining the “best split” are Gini impurity and information gain for classification tasks. As this is the metric used by default in RF, I will just say a few words about the Gini impurity to better understand how RF works.

Gini impurity measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen and it’s used to determine how to split the data into smaller group. While building the decision tree, feature with the least Gini Impurity relative to the root node will be chosen. Of note, Gini Impurity, unlike information gain, isn’t computationally intensive as it doesn’t involve the logarithm function used to calculate entropy in information gain, which is why Gini Impurity is preferred over information gain.

Single decision trees are very easy to visualize and understand because they follow a method of decision-making that is very similar to how humans make decisions: with a chain of simple rules. However, there are not very robust, it’s here RF come into play.

RF makes predictions by combining the results from many individual decision trees. Because it uses multiple learning algorithms to obtain better predictive performance, RF falls into the category of Ensemble learning. One major way for combining the multiple decision trees in a random forest is Bagging, which is also called Bootstrap aggregation, where decision trees are trained on randomly sampled subsets of the data.
A big advantage of bagging over individual trees is that it decreases the variance of the model. Individual trees are very prone to overfitting and are very sensitive to noise in the data. Combining them with bagging will make them more robust.

In addition to randomly sampling instances from our data, RF also uses feature bagging. With feature bagging, at each split in the decision tree, only a random subset of features is considered. This technique helps reducing the impact of very strong predictor variables (i.e. features that have a very strong influence on predicting the target or response variable).

Finally, in the context of our analyzes, I end with a model able to predict the probability of a patient to be classified in a group or other. Important characteristics will also emerge from this model, leaving the researcher the possibility of carrying out in-depth studies on the biological meaning of these characteristics. Also, these newly detected features could then be tested for a link with patient outcome as discussed below.

1.4.3.3. SURVIVAL ANALYSIS

When clinical outcome is available, it becomes possible to assess difference in survival between groups of patients with distinct clinical or genomic features. For instance, in 2018, TCGA released a standardized dataset named the TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR), which includes four major clinical outcome endpoints and usage recommendations for each cancer type (Liu et al. 2018). This provide an unprecedented opportunity for investigating cancer biology and differences in survival for large cohort of patients with cancer.

To apply survival analyzes and in order to distinguish individuals, the values of one specific genomic feature can be divided in two groups using median or quartiles. Otherwise, based on other approach of clustering, we can separate groups sharing several genomic characteristics. Then, the survival or time-to-event analysis can be performed to test difference in survival. It encompasses several methods that are used routinely in the clinical field. I will describe succinctly the purpose of each one without
diving into the mathematics behind so as not to lose the untrained reader. Table 1-5 gives the definition given by TCGA for different endpoints that can be explored by survival analyzes in cancer.

<table>
<thead>
<tr>
<th>Endpoint</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>OS (overall survival)</td>
<td>It is the period from the date of diagnosis until the date of death from any cause.</td>
</tr>
<tr>
<td>DSS (disease-specific survival)</td>
<td>Event is death from the disease, and the event time is from the date of initial diagnosis until the date of death from the disease.</td>
</tr>
<tr>
<td>PFI (progression-free interval)</td>
<td>It’s the period from the date of diagnosis until the date of the first occurrence of a new tumor event.</td>
</tr>
<tr>
<td>DFI (disease-free interval)</td>
<td>It’s defined as the period from the date of diagnosis until the date of the first new tumor progression event subsequent to the determination of a patient’s disease-free status after their initial diagnosis and treatment. For DFI the time interval should start from the time when the patient was first determined to be disease-free, but such information was not available in the TCGA clinical data.</td>
</tr>
</tbody>
</table>

First, the Kaplan-Meier (KM) estimator (Kaplan and Meier 1958) is a non-parametric method used to estimate the survival probability from observed survival times. The survival probability during the follow up time can be graphically presented by KM curves which are an easy way to interpret the outcome of patients. An example of KM curve is shown later in the manuscript (Figure 1-22).

Then, to test the statistical significance of the difference in survival probability, the log rank test can be used to assess the null hypothesis of no difference in survival
between two or more independent groups. The test compares the entire survival experience between groups and can be thought of as a test of whether the survival curves are identical (overlapping) or not. Usually, a log rank test cut-off of 0.05 is considered as reliable.

Finally, Cox regression (or proportional hazards regression) is the method for investigating the effect of one or several variables upon the time a specified event takes to happen by making comparisons between the number of survivors in each group at multiple points in time (Cox 1972). This approach is used to compute the hazard ratio (HR) which is an estimate of the ratio of the hazard rate in one tested group versus the control group. It gives a prognostic value for a specific feature between groups. For instance, HR>1 means the tested group is associated with bad prognosis for the specific feature considered. (Conversely, HR<1 means the tested group is associated with good prognosis for the specific feature considered).

To resume, the survival analysis results can be graphically presented by the Kaplan-Meier (KM) plot with hazard ratio (HR) and log-rank \( p \) value.

**1.4.4. PRE-CLINICAL MODELS FOR CANCER**

The previous techniques and analysis can be applied on different organic systems where sequencing has been done. I discuss earlier computational aspects, but it’s important to keep in mind that sequencing data can come from distinct contexts (*in vivo* or *in vitro* models) that bring its advantages and drawbacks.

A biological hypothesis cannot be always tested on living human beings. To solve this problem, two kind of in vivo models are used to investigate different facets of cancer biology.

*In vivo* models use intensively mouse as an alternative organism of study due to its relative phylogenetic closeness and physiological similarity to our specie. When the function of a cancer gene is modified to cause the development of a specific cancer,
we speak about Genetically Modified Mice (GEMs). These models are well suited for studying tumor initiation and progression (Cheon and Orsulic 2011).

Another in vivo system, can be the direct transplantation of tumors cells into immunodeficient mice called xenograft tumor model (Marangoni et al. 2007). This model conserves morphology architecture, vasculature, peripheral growth and molecular features of the original tumor from the patient. In this particular setting, this system represents an exciting opportunity to study response to treatment. New technologies are also arising. Organoids are 3D multicellular in vitro tissue construct that mimics its corresponding in vivo organ, such that it can be used to study aspects of that organ in the tissue culture dish (Weeber et al. 2017; H. Xu et al. 2018). This recent technology is a breakthrough that will facilitate drug testing and guides personalized therapy (Kim, Koo, and Knoblich 2020).

Another alternative when a whole living organism is not available is in vitro models, when tumor cells are cultured on a bench in a synthetic environment composed of nutrients. It’s a more convenient way to study the behavior of cancer cells compare to mouse xenograft which are time consuming and engraftment is not guaranteed to be successful. This is the most widely used in oncology because of its ease of use. The only limitation of this system is that it does not provide the context of a true tumor and so the interactions that can occurs with the microenvironment are lost.

The Cancer Cell Line Encyclopedia (CCLE) project is an effort to conduct a detailed genetic characterization of a large panel of human cancer cell lines (Ghandi et al. 2019; Barretina et al. 2012). It actually contains 1457 different cell lines which has been characterized at the level of the genome and transcriptome using high throughput sequencing technologies. Then, these cells are used as the ground of large drug screening initiatives (Tsherniak et al. 2017; Basu et al. 2013; Corsello et al. 2017). These initiatives are using genome-wide RNAi and CRISPR loss-of-function screens to systematically identify essential genes across hundreds of human cancers. In parallel, they are progressively establishing a comprehensive resource for drug sensitivity that are then freely available for the research community.
In vitro models are well suited to induce an EMT in epithelial cells and study features that are impacted during this process. Our lab uses an inducible cell reprogramming system based on normal human mammary epithelial cells (MCF10a). This system stably expresses a tamoxifen inducible form of the EMT transcriptional regulator Snail (MCF10a-Snail-ER). Upon tamoxifen treatment, Snail enters the nucleus and this event will silence key epithelial markers and leads to a rapid reprogramming into mesenchymal cells. However, several settings have been utilized to study EMT, using different cell types and inducers. To give a quick overview I will mention a few examples. In Human non-small cell lung cancer cell lines (H358), the induction of EMT has also be done by doxycycline which will induce Zeb1 another master transcriptional regulator of EMT. TGF-Beta as also been used to active EMT in immortalized human mammary epithelial cells (HMLE). Finally, under certain conditions, such as low cell confluence or hypoxia, cells can undergo a spontaneous EMT without a given inducer. In this case, the new cells freshly obtained with mesenchymal traits will be sorted, harvested and cultured in order to be studied.

1.4.5. CANCER POPULATION GENOMIC RESOURCES

In the previous section, I discussed how to study fundamental cancer biology aspects using cell lines or alternative model organism. Due to the increasing feasibility of sequencing genomes, huge number of primary tumors can be sequenced thanks to vast cohort of patients. Large-scale initiatives have emerged to explore the underlying biology of cancer, based on genomics and basic clinical information at the same time. The first goal of these projects was to catalog and discover major genomic alterations causing cancer, for a better understanding of the disease and in order to improve patient care. The datasets produced turned out to be useful resources to dive into the genetics of cancer and are widely used by researchers to test scientific hypothesis and develop new therapeutic strategies.

In the early 2000s, METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) released a collection of over 2000 clinically annotated primary fresh-frozen breast cancer specimens from tumor banks in the UK and Canada (Curtis et al. 2012). Integrated genomic/transcriptomic analysis of breast cancers with
long-term clinical outcomes were made available for researchers. mRNA for expression was measured with the Illumina HT-12v3 platform. CNA (copy number aberrations) and SNPs were detected with the Affymetrix SNP 6.0 array. And so, for now, due to the technical platforms used, it was not yet possible to study alternative splicing on a large scale.

In 2006, The Cancer Genome Atlas (TCGA), started to deliver a huge amount of sequencing data from human tumors (Hutter and Zenklusen 2018). Over the next dozen years, TCGA generated over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data. The aim of this project was to molecularly characterized over 20,000 primary cancers and matched normal samples spanning 33 cancer types. Thanks to the use of Illumina HiSeq platform, changes in alternative splicing between tumor samples and patients can be analyzed.

In the next part, I will discuss all the aspects I introduced earlier related to cancer biology, EMT and alternative splicing, in the context of a breast cancer cohort with high-throughput sequencing data available from the TCGA. I will introduce the research I pursued during my PhD work focused on the basal-like breast cancer subtype.
1.5. A CONCRETE APPLICATION TO BREAST CANCER

1.5.1. EPIDEMIOLOGY

In women, breast cancer is the most frequent malignancy worldwide with 2.1 new cases and 0.6 million deaths in 2018. In France, it is the most frequent cancer in women followed by colorectal and lung cancer. Like most cancer, aging of the population and other factors as physical inactivity, smoking and alcohol consumption, increase the cancer risk. Postmenopausal hormone use, long menstrual history, family history of breast or ovarian cancer are also specific risk factors for breast cancer in women (American Cancer Society 2018). In about 5 to 10% of cases there is a genetic predisposition to breast cancer due to two breast cancer genes (Tao et al. 2015). These are BRCA1/2 genes that produce tumor suppressor proteins which help repairing damaged DNA. It was estimated that about 72% of women who inherit a harmful BRCA1 mutation and about 69% of women who inherit a harmful BRCA2 mutation will develop breast cancer by the age of 80 (Kotsopoulos 2018).

This disease is curable in ~70–80% of patients with early-stage, non-metastatic disease. However, advanced breast cancer with distant organ metastases is still considered incurable with currently available therapies (Harbeck et al. 2019). It is a real public health issue and better ways of diagnosis and treatment are needed.

1.5.2. BREAST ANATOMY

The breast is an exocrine gland composed of a mass, an areola and a nipple. The nipple is located in the middle of the areola, which is the darker area surrounding the nipple (Figure 1-21). The mammary gland consists of an epithelial bilayer made of cuboidal cells surrounded by myoepithelial cells contained within adipose (fatty) tissue supported by a dense fibrous connective tissue. Embedded in the breast’s fatty and fibrous tissue are 15 to 20 glands called lobes, each of which has many smaller lobules, or sacs, that produce milk (Pandya and Moore 2011). Ducts are thin tubes that carry milk to the nipple. Breast development and function depend on hormones produced by the ovaries, namely estrogen and progesterone. Each breast also contains blood
vessels and lymph vessels that transport a fluid that travels through a network of channels called the lymphatic system and carries cells that help the body to fight infections. The lymph vessels lead to the lymph nodes which are small glands part of the lymphatic system that plays an integral role in the immune functions of the body. Breast cancers can form in the ducts and the lobes. If a cancer has reached these lymph nodes, it may mean that cancer cells have spread to other parts of the body.

![Breast anatomy and histology.](image)

Figure 1-21 Breast anatomy and histology.

On the left, schematic illustration of a breast section. All breast cancers arise in the terminal duct lobular units of the collecting duct. On the right, the lobule section let us see two layers of cells. Inner layer of myoepithelial cells provides structural support to the lobules and assist milk ejection during lactation. Outer layer of luminal epithelial cells produces milk during lactation. (adapted from Harbeck, 2019)

1.5.3. CLINICAL CHARACTERISTICS

To ensure the best possible care for the patient, breast tumors are classified by health professionals according to certain clinical items. This gives an idea of the type of disease and its progress. Three major items are given: the stage, the grade and the histological type. I will give a quick overview to explain what they stand for.
The first classification of breast cancer is based on histological type. Most breast cancers are invasive, meaning that they spread around the surrounding breast tissues but there are different types of invasive breast cancer. The two most common are invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC). IDC represents 80% of the invasive breast cancer and starts in the cells that line a milk duct in the breast. About 1 in 10 invasive breast cancers is an invasive lobular carcinoma (ILC). ILC starts in the milk-producing glands (lobules). Other invasive cases exist and are histological variants, each of which accounts for no more than 5% of all invasive cases (Philipps and Li 2010).

The stage of a cancer is a measurement of the extent of the tumor and its spread. The standard staging system for breast cancer uses a system known as TNM. The T category (T0, T1, T2, T3, or T4) is based on the size of the tumor and higher T numbers mean a larger tumor. The N category (N0, N1, N2, or N3) indicates whether the cancer has spread to lymph nodes near the breast and, if so, how many lymph nodes are affected. Higher numbers after the N indicate more lymph node involvement by cancer. The M category (M0, M1) details if the cancer has spread to distant sites. These categories are combined to give the cancer an overall stage. Stages are expressed in Roman numerals from stage I (the least advanced stage) to stage IV (the most advanced stage).

The grade is based on how much the cancer cells look like normal cells. It is based on the appearance of cancer cells, the shape of the nucleus and the number of cells in division. For each of this feature, a score is given. Then they are added, which gives a number between 3 and 9 that is used to get a grade of 1, 2, or 3. For example, Grade 1 (score 3, 4, or 5) means cancer cells look more like normal breast tissue whereas in Grade 3 (score 8, 9) cancer cells look very different from normal cells and will probably grow and spread faster.

1.5.4. MOLECULAR SUBTYPES OF BREAST CANCER

As I said before, breast cancer is not a single disease. In early 2000, molecular portraits of human breast tumors started to be defined by Perou and Sorlie (Perou,
Sùrlie, et al. 2000; Sorlie et al. 2001). They were based on a 50-gene expression signature named PAM50. Since, several classifications has been proposed (Ali et al. 2014) but this classification stays the reference in the clinical field. Using Pam50, four clinically relevant molecular subtypes were described (luminal A, luminal B, HER2-enriched and basal-like) mostly corresponding to hormone receptor and HER2 status. These intrinsic group are distinct in terms of clinical presentation (lymph nodes invasion, local and regional recurrence, localization of metastases). An additional intrinsic subtype of breast cancer, known as claudin-low, has recently been identified, showing several common features with basal-like tumors and reflecting the diversity of tumors with a low luminal differentiation. Claudin-low are highly enriched in mesenchymal traits and stem cell features and are therefore considered as the most primitive breast cancers (Pommier et al. 2020; Prat et al. 2010). Of note, a normal breast-like group was also initially defined but is thought to be an artefact due to low tumor cellularity so this is why I am not going to detail it.

Luminal A breast cancers are hormone-receptor positive, low-grade, tend to grow slowly and have the best prognosis. This category is the most frequent breast cancer with 60-70% incidence rate. Luminal B breast cancers are hormone-receptor positive and generally grow slightly faster than luminal A cancers and their prognosis is slightly worse. HER2-enriched breast cancer is hormone-receptor negative (estrogen-receptor and progesterone-receptor negative) and HER2 (Human Epidermal Growth Factor Receptor-2) positive; They are characterized by an amplification and overexpression of HER2 tyrosine kinase receptor gene. HER2-enriched cancers tend to grow faster than luminal cancers and can have a worse prognosis. Triple-negative/basal-like breast cancer is hormone-receptor negative (estrogen-receptor and progesterone-receptor negative) and HER2 negative. They display a high rate of recurrence, and have a poor prognosis. They are most commonly high-grade at diagnosis. With HER2+, they are the most aggressive breast tumors (Figure 1-22).
The incidence of each intrinsic subtype is different with the highest incidence related to the luminals (70-80%), followed by the basal-like (10-15%) and the HER2+ (<15%) (Harbeck et al. 2019). Both luminal subtypes are particularly sensitive to targeted hormonal therapy and are associated with a good prognosis. Patients with Her2 enriched subtype, appear receptive to neoadjuvant chemotherapy together with anti-HER2 therapy (trastuzumab and pertuzumab) which has become the standard of care for this subtype. Further, new molecules like T-DM1, dramatically help to have a better outcome.

Nowadays, Triple-negative/basal-like breast cancer is the only cancer subtype that remains without hormonal therapy nor targeted therapy. Thus, there is an urgent medical need to identify therapeutic targets and develop more effective stratified medicine for the treatment of this subtype.

Figure 1-22 Breast cancer survival by molecular subtype.

Kaplan-Meier plot of overall breast cancer survival by molecular subtype, Ontario, 2010-2012. The poorest survival was observed among patients with the triple-negative subtype (adapted from Fallahpour, 2017).
Being one of the most aggressive breast cancer subtypes, basal-like tumors are known for their great heterogeneity (biological, histological and clinical features) and their pattern of relapse that is characterized by frequent and early relapses with poor prognosis. Triple-negative and basal-like breast cancer are terminologies that are often used interchangeably although a small distinction remains (Alluri and Newman 2014). Triple-negative is an immunohistochemical definition who considers the fact that these tumors lack expression of hormone (estrogen and progesterone) receptors and are also characterized by the absence of HER2 receptor. From a histological point of view, most of these tumors are classified as invasive ductal carcinomas. Several rare histologic groups have also been characterized (Figure 1-23) and represent less than 1% of all cases of TNBCs (secretory carcinoma, typical medullary carcinoma, atypical medullary carcinoma, apocrine carcinoma, adenoid cystic carcinoma, spindle-cell metaplastic carcinomas and adenosquamous carcinoma) (Geyer et al. 2017).

Figure 1-23 Histological heterogeneity of Triple Negative Breast cancer

Examples of distinct histologic types of triple-negative breast carcinomas. From left to right: Invasive Ductal, Apocrine, Adenoid cystic, Metaplastic breast carcinomas. Of note, Invasive ductal carcinoma represents 95% of cases. (Adapted from Geyer, Pajer, Weigelt, 2017)

Basal-like term was given by Perou & al because these tumors show some characteristics of myoepithelial cells from the outer layer of duct breast as the expression of cytokeratin CK5/6, CK14, CK17 and EGFR (Epidermal Growth Factor Receptor) (Perou, Sørile, et al. 2000). Basal-like represents the most frequent subtype of TNBC (70-80%) (Prat et al. 2013).

They are also associated with an additional intrinsic subtype of breast cancer known as claudin-low that extends through all the intrinsic subtype but is mostly observed in basal-like subtype. Claudin-low are distinguished by low genomic
instability, mutational burden and proliferation levels, and high levels of immune and stromal cell infiltration. They expressed a low level of critical cell–cell adhesion molecules, including claudins 3, 4, and 7, occludin, and E-cadherin. They were characterized by a low expression of luminal markers and a high expression of mesenchymal marker. Claudin-low tumors displayed the least differentiated phenotype along the mammary epithelial differentiation hierarchy showing enrichment for gene expression signatures derived from human tumor-initiating cells (TICs) and mammary stem cells (Fougner et al. 2020). They have been associated with poor prognosis but not in all cases. This claudin-low phenotype is a further example of the genetic heterogeneity that can be found within the basal-like subtype.

There have been several attempts in the clinical field to better classify TNBCs (Jézéquel et al. 2019; D. Y. Wang et al. 2019; Ignatiadis et al. 2018; Jiang et al. 2019). Lehmann et al were among the first to publish a study trying to better dissect the TNBC specific heterogeneity (Lehmann et al. 2011). Their study proposed 6 molecular subtypes of TNBC: two basal-like-related subgroups (basal-like 1 (BL1) and 2 (BL2)), two mesenchymal-related subgroups (mesenchymal (M) and mesenchymal stem-like (MSL)), one immunomodulatory subgroup (IM) and one luminal androgen receptor group (LAR). Each of these subtypes has specific molecular abnormalities. The BL1 and BL2 subgroups are both enriched in proliferation genes. BL1s also express genes involved in DNA repair whereas the BL2 subgroup expresses genes involved in growth signaling pathways. The M subgroup is enriched with genes involved in cell mobility and the epithelial-mesenchymal transition. The MSL subgroup has an expression profile close to the M subgroup and is enriched in genes involved in angiogenesis and in some immune response signaling pathways. The IM subgroup is enriched with genes involved in the immune response and lymphocyte infiltration. Finally, the LAR subgroup that represents about 10% of the TNBCs, expresses the androgen receptor (AR) in the presence of a luminal-like expression signature and thus, might be treated with agents that target AR. Recently, the clinical relevance of this classification was evaluated in a retrospective analysis of 125 patients with TNBC treated with chemotherapy before surgery (Santonja et al. 2018). The authors show that different responses can be observed according to their TNBCs subtypes. Patients with BL1 tumors achieve the highest pathological complete response rate and patients with tumors classified as BL2, LAR and MSL have the lowest response rates. A more
recent and partially overlapping classification segregated TNBC into (Burstein et al. 2015) into 4 main groups: LAR, mesenchymal (MES), Basal-like immune-suppressed (BLIS) and Basal-like immune-activated (BLIA). They show that BLIS and BLIA tumors have the worst and best prognoses, respectively, compared to the other subtypes. In 2017, Milioli & al, proposed a signature that supports the existence of at least two subgroups of basal-like breast cancers with distinct disease outcome (Milioli et al. 2017). Later, in 2019 Jezequel & al identified three molecular cluster in TNBCs: one molecular apocrine (C1) and two basal-like enriched (C2 and C3). C2 presented pro-tumorigenic immune response (immune suppressive), high neurogenesis (nerve infiltration), and high biological aggressiveness. In contrast, C3 exhibited adaptive immune response associated with complete B cell differentiation that occurs in tertiary lymphoid structures, and immune checkpoint upregulation (Jézéquel et al. 2019). The same year a genomic and transcriptomic analysis of a cohort of 465 Chinese primary triple-negative breast cancer (TNBC) defines a luminal androgen receptor (LAR) subtype (23%) characterized by androgen receptor signaling; (2) an immunomodulatory (IM) subtype (comprising 24% of tumors) with high immune cell signaling and cytokine signaling gene expression; (3) a basal-like and immune-suppressed (BLIS) (39%) subtype characterized by upregulation of cell cycle, activation of DNA repair, and downregulation of immune response genes; and (4) a mesenchymal-like (MES) subtype (15%) enriched in mammary stem cell pathways (Jiang et al. 2019).

This catalog of the various researches carried out with the aim of deciphering the complexity of these aggressive tumors aims to illustrate the importance that requires a better classification inside this tumor subtype in order to improve the care of patients by doctors.

1.5.6. THESIS OBJECTIVES

1.5.6.1. IDENTIFICATION OF AS EVENTS ASSOCIATED WITH POOR PROGNOSIS IN BREAST CANCER.

The main goal of my PhD work was to explore alternative splicing events that could potentially have an impact on patient survival in a specific, aggressive and deadly subtype amongst all breast cancers, the Basal-like breast cancer subtype. A growing body of evidence suggests a central role of EMT in metastasis and tumor progression.
This clinical relevance in combination with increasing evidence for the importance of alternative splicing in EMT was the core of my initial hypothesis. I used the idea that Basal B cell lines, according to literature, were described as the most invasive cell lines, displaying a high number of mesenchymal features. I looked if basal B specific signature could be used to classify basal-like tumors. Using basal-like breast cancer cell lines, I developed a custom random forest method to transfer knowledge from cell lines to tumors from patients where I had clinical follow-up. Once I had isolated this signature in patients, I characterized it and identified an association with an EMT signature by using GSEA analysis and by looking into RNA-seq of EMT-induced public projects. Then, I looked further for potential splicing factors (SFs) in basal cell lines that could drive this AS program. Taking advantage of RNA-seq data upon modulation of expression levels of the candidate SFs ESRP1 and RBM47, I explored to what extent the newly identified basal B-specific signature is regulated by common SFs. Finally, I investigated the association of the expression of these SFs with survival in TCGA patients.

1.5.6.2. NEW INSIGHT FROM K-MERS ANALYSIS IN BREAST CANCER

To a lesser extent, I was also involved in the development of k-mer based approach to classify patients and extract the biological knowledge hidden by k-mers of importance. I was involved in two publications related to k-mers during my PhD. The first article demonstrates that kmers are a powerful tool to classify labeled biological samples compared to classical methods. I was mainly taking part to help in having access to resources annotation of breast cancer subtype, specific publications related to the field, computation of gene expression and alternative splicing, and finally discussion around the use of the random forest classifier. The second paper delivers a software solution to study k-mers in several samples, and highlight the fact that k-mers can lead to detection of novel biological events to better understand mechanisms involved in a specific cellular phenotype, or in order to detect new targets for therapy. Notably, during the benchmark of iMOKA software, we showed that amongst the best k-mers that lead to an accurate classification of breast cancer subtypes, 4 splicing isoforms (MYO6, TPD52, IQCG and ACOX2) were found and already reported as to be amongst the 5 most important isoforms differentially expressed between ER+HER2- and ER-HER2 primary breast tumors. This helped to validate the
consistency of the method. These two publications are in the annexes section at the end of this manuscript.
2. RESULTS
A cell-to-patient machine learning transfer approach uncovers novel basal-like breast cancer prognostic markers amongst alternative splice variants

Jean-Philippe Villemin¹, Claudio Lorenzi¹, Marie-Sarah Cabrillac¹, Andrew Oldfield¹, William Ritchie¹* & Reini F. Luco¹*

1. Institut de Génétique Humaine (IGH), Centre National de la Recherche Scientifique, University of Montpellier, Montpellier, France

Corresponding authors: william.ritchie@igh.cnrs.fr and reini.luco@igh.cnrs.fr

ABSTRACT:

Background:
Breast cancer is amongst the 10 first causes of death in women worldwide. Around 20% of patients are misdiagnosed leading to early metastasis, resistance to treatment and relapse. Many clinical and gene expression profiles have been successfully used to classify breast tumours into 5 major types with different prognosis and sensitivity to specific treatments. Unfortunately, these profiles have failed to subclassify breast tumours into more subtypes to improve diagnostics and survival rate. Alternative splicing is emerging as a new source of highly specific biomarkers to classify tumours in different grades. Taking advantage of extensive public transcriptomics datasets in breast cancer cell lines (CCLE) and breast cancer tumours (TCGA), we have addressed the capacity of alternative splice variants to subclassify highly aggressive breast cancers.

Results:
Transcriptomics analysis of alternative splicing events between luminal, basal A and basal B breast cancer cell lines identified a unique splicing signature for a subtype of tumours, the basal B, whose classification is not in use in the clinic yet. Basal B cell lines, in contrast with luminal and basal A, are highly metastatic and express epithelial-to-mesenchymal (EMT) markers, which are hallmarks of cell invasion and resistance to drugs. By developing a semi-supervised machine learning approach, we transferred the molecular knowledge gained from these cell lines into patients to subclassify basal-like triple negative tumours into basal A- and basal B-like categories. Changes in splicing of 25 alternative exons, intimately related to EMT and cell invasion such as ENAH, CD44 and CTNND1, were sufficient to identify the basal-like patients with the worst prognosis. Moreover, patients expressing this basal B-specific splicing signature also expressed newly identified biomarkers of metastasis-initiating cells, like CD36, supporting a more invasive phenotype for this basal B-like breast cancer subtype.

Conclusions:
Using a novel machine learning approach, we have identified an EMT-related splicing signature capable of subclassifying the most aggressive type of breast cancer, which are basal-like triple negative tumours. This proof-of-concept demonstrates that the biological knowledge acquired from cell lines can be transferred to patients data for further clinical investigation. More studies, particularly in 3D culture and organoids, will increase the accuracy of this transfer of knowledge, which will open new perspectives into the development of novel therapeutic strategies and the further identification of specific biomarkers for drug resistance and cancer relapse.

BACKGROUND:

Breast cancer is a heterogeneous disease with multiple molecular drivers and disrupted regulatory pathways [1, 2]. The development of large-scale genomics and transcriptomics methods has increased the capacity to identify clinically-relevant tumour subtypes with distinct molecular signatures. These can be used for a better choice of treatment and/or prediction of potential metastasis which can improve survival outcome [3, 4]. However, patients are still facing a high percentage of misdiagnosis in which undetected early metastasis and/or inappropriate choice of treatment can lead to deadly complications with the use of unnecessary severe chemotherapies or the apparition of drug resistance and subsequent tumour relapse [5].

Currently, breast cancer is classified into five major categories (normal-like, luminal A, luminal B, Her2-positive and basal-like) based on expression of three receptors: oestrogen and progesterone hormonal receptors (ER and PR) and the epidermal growth factor receptor ERBB2 (Her2). Basal-like are the most aggressive, and difficult to treat, type of breast cancer tumour. They are usually negative for the three receptors, and thus called triple negative breast cancer (TNBC), which represents 10-20% of all breast cancers. These tumours are usually found in younger patients with a larger size and higher probability of lymph node infiltration and metastasis [2, 6]. Furthermore, the absence of all three receptors reduces the number of targeted therapeutic strategies to be used, leaving nonspecific chemotherapy as the standard treatment of choice, which soon leads to dose-limiting side-effects, resistance to treatment and finally clinical relapse in less than 5 years [6]. A better understanding of the molecular differences in between these tumour categories will improve the choice of treatment and detection of early metastasis, which will significantly impact patient’s outcome.

There have been many attempts to identify novel therapeutic targets and/or prognostic
biomarkers to better subclassify breast cancer tumours [7]. Over 170 independent breast cancer susceptibility genomic variants have been identified. Many of which have been associated with a specific tumour category, such as ER positiveness or Her2 amplification. However no clear subcategories exist despite tumour heterogeneity and differences in clinical response to treatment and tumour relapse within the same category [8–10]. Interestingly, alternative splicing is an emerging source of new biomarkers and therapeutic targets in cancer [11–15].

The alternative processing of mRNA precursors enables one gene to produce multiple protein isoforms with different functions, increasing protein diversity and the capacity of a cell to adapt to new environments. An increasing number of splice variants, and their respective splicing regulators, have been shown to confer a selective advantage to tumour cells. For instance, the splicing regulators RBM5, 6 and 10 favour tumour cell proliferation and colony formation by regulating the alternative splicing of the membrane-bound protein NUMB [16]. Post-translational activation of the splicing factor SRSF1 (also known as ASF/SF2) confers resistance to apoptosis by inducing inclusion of the anti-apoptotic splice variant in a network of functionally related genes, such as Bcl-X and Mcl1 [17]. Regulation of VEGF splicing is detrimental for stimulation of angiogenesis [18]. A change in the alternative splicing of the pyruvate kinase pre-mRNA can switch tumour cells metabolism to adapt to the increased proliferation [19, 20]. Finally, a list of well-known alternatively spliced variants related to cell adhesion (CTNND1, CD44) and cytoskeleton organisation (ENAH, FLNB) are responsible for the acquisition of migratory and invasive phenotypes necessary for distal metastasis [13, 21–24]. The existence of functionally relevant cancer specific isoforms is therefore a promising new source of highly specific and less toxic
therapeutic targets for the development of isoform-specific antibodies and/or splice-switching antisense oligonucleotides [25, 26].

By taking advantage of an extensive transcriptomics and anti-tumour compound screening information publicly available in cancer cell lines from the Cancer Cell Line Encyclopedia (CCLE) [27], we identified a splicing signature that can stratify basal breast cancer cell lines into two well-known subtypes, basal A and basal B. In contrast to basal-like breast cancer patients, basal breast cancer cell lines are divided into two subgroups, basal A and basal B, depending on the expression profile of a subset of basal (cytokeratins, integrins), stem cell (CD44, CD24) and mesenchymal markers (Vimentin, fibronectin, MSN, TGFBR2, collagens, proteases) [28–30]. Basal B cell lines are mostly triple negative breast cancer cells that express classical mesenchymal and stem cell markers characteristic of the epithelial-to-mesenchymal transition (EMT), a biological process in which epithelial cells acquire mesenchymal features that are advantageous for the cancer cell, such as increased cell motility to invade distal organs in metastasis, resistance to apoptosis, refractory responses to chemotherapy and immunotherapy, and acquisition of stem cell-like properties like in cancer stem cells [31, 32]. In concordance, basal B cells are morphologically less differentiated, with a mesenchymal-like shape, and a more invasive phenotype in culture assays than basal A and luminal cells [28, 33, 34]. We aimed to transfer this basal A/basal B splicing classification into the clinic by using a semi-supervised machine learning approach. We successfully classified 40% of basal-like breast cancer patients (75/188) from the Cancer Genome Atlas (TCGA) [35] as basal B-like based on a unique 25 spliced gene signature characteristic of cells undergoing EMT. In this signature, we found well-known markers of malignancy, such as ENAH EMT splice variant that promotes lung metastasis [36] or CSF1 variant which promotes macrophage infiltration and distal
metastasis [37], together with new promising splicing candidates of tumour progression and invasiveness (PLOD2, CTNND1, SPAG9). Finally, expression of this basal B signature was sufficient to identify triple negative breast cancer tumours with poor survival, highlighting the prognostic value of the newly identified splicing biomarkers to subclassify one of the most heterogeneous and difficult to treat type of breast cancer. More studies in cell lines, particularly regarding resistance to treatment and cell invasion will be essential to refine this splicing signature in view of orienting treatment or predicting metastasis sites.

In conclusion, by adapting a machine learning approach, we were able to transfer the molecular knowledge obtained in experimental cell lines to identify novel biomarkers of poor prognosis and metastasis amongst triple negative breast cancers in patients. Furthermore, the study of the regulatory pathway involved in this specific splicing signature pointed to RBM47 as one of the splicing regulators responsible for the basal B-specific splicing signature, and for which differential expression levels also correlate with distinct prognostic values, turning this splicing factor a promising novel therapeutic target. Further clinical and functional validation of the 25 splicing events proposed in our basal B-specific splicing signature will open new perspectives in the understanding of triple negative breast cancers and the improvement of currently available therapeutic strategies and survival outcome.

RESULTS:

A distinctive Basal B-like breast cancer splicing signature.

Data mining of large-scale genomics and transcriptomics datasets in breast cancer cell lines are a promising source of novel biomarker and therapeutic targets [23, 38, 39].
We sought to leverage the wealth of transcriptomics and functional data available in cancer cell lines to better understand different profiles of breast cancer. Hierarchical clustering of changes in alternative splicing of cassette exons and gene expression profile of 80 breast cancer cell lines from two extensive and complementary projects (Additional File 2: Table S1) revealed basal B cell lines as a distinctive group of cells with an expression and splicing profile significantly different from basal A and luminal cancer cells (Additional File 1: Fig.S1). To identify the transcriptional signature characteristic of basal B cells, we repeated the hierarchical clustering in just basal A and basal B cell lines to merge all the differentially expressed and spliced transcripts responsible for the segregation of basal B cell lines (Fig.1). We found 635 genes and 217 spliced isoforms with significantly different levels between basal A and basal B cells (Fig.1a,b). In line with published tissue-specific and EMT transcriptomics analyses [40–42], most of the genes differentially spliced were not affected at the expression level, suggesting that two different subsets of genes, and thus regulatory layers, are responsible for the basal B phenotype (Fig.1c). Gene set enrichment analysis (GSEA) [43] between basal B and basal A cells confirmed the EMT and stem cell-like phenotype characteristic of basal B cell lines (Fig.2a,b), which was supported with a higher CD44+/CD24- stem cell score (Fig.2e) [28–30]. DAVID gene ontology analysis of differentially expressed and spliced genes also underlined biological terms that are hallmarks of EMT and cell invasiveness, such as cell-cell junction (Fig.2d) [44]. However differentially expressed genes were also enriched in their own unique terms, related to extracellular vesicles/plasma membrane organization. While differentially spliced genes were specifically enriched in terms related to GTPase activity, cytoskeletal protein and cadherin binding, which reinforces the existence of two complementary regulatory pathways (Fig.2d). Finally, another malignant characteristic
FIGURE 1
acquired by cancer cells undergoing EMT is resistance to chemotherapy, which often leads to clinical relapse. Gene set enrichment analysis found upregulation of genes resistant to the Epidermal Growth Factor Receptor (EGFR) inhibitor Gefitinib (Fig.2c), which is an alternative to hormonal therapy in Her2+ breast cancer tumours, but is not efficient in triple negative tumours [45]. Available drug assays from the Genome Drug Sensitivity in Cancer portal (GDSC) [46] confirmed the need of a higher concentration (IC50) of Gefitinib, and other EGFR inhibitors (Erlotinib, Sapitinib), to have the same deleterious effect on basal B compared to basal A cancer cells (Fig.2f). Basal B cell lines also showed a significant resistance to well-known inhibitors of the cell cycle (Irinotecan, Taselisib, 5-Fluorouracil), drug inducers of cell death (AZD5582, AZD5991) and other receptor tyrosine kinase inhibitors, such as Savolitinib which inhibits c-MET to reduce tumour persistence and metastasis [47].

In summary, we have identified two distinct transcriptional and splicing signatures, specific of basal B cell lines, that underline an EMT phenotype with molecular characteristics related to cell invasion, stemness and resistance to chemotherapy. We next sought to investigate whether this basal B-specific splicing signature could also be used to subclassify basal-like/triple negative breast cancer patients.

A semi-supervised machine learning approach to subclassify basal-like breast cancer patients.

As a first and simple approach, we performed a hierarchical clustering followed by a k-means clustering (k=2 for “A-like” and “B-like”) of the 188 patients, annotated as basal-like in The Cancer Genome Atlas Program (TCGA), using the 635 differentially expressed or 217 differentially spliced cassette exons characteristic of
**FIGURE 2**

- **Panel a:** TAUBE EMT UP and EMT DOWN ranked list of genes.
- **Panel b:** LIM MAMMARY STEM CELL UP and STEM DOWN ranked list of genes.
- **Panel c:** COLDREN GEFITINIB RESISTANCE UP and RESISTANCE DOWN ranked list of genes.

**Differentially expressed genes:**
- **GOTERM_CC_FM**
  - Cell-cell junction organization
  - Cell junction organization
  - Plasma membrane organization
- **GOTERM_BP_FM**
  - Negative regulation of cell proliferation
  - Epithelial cell differentiation
  - Keratinocyte differentiation
- **GOTERM_CC_BP**
  - Cell-cell junction assembly
  - Endocytic system organization
  - Cell differentiation involved in embryonic placenta development
- **GOTERM_BP_BP**
  - Negative regulation of cellular component movement
  - Regulation of cell proliferation
  - Junction assembly

**Differentially spliced genes:**
- **GOTERM_CC_FM**
  - Perinuclear region of cytoplasm
  - Anchoring junction
- **GOTERM_BP_FM**
  - Adhesion junction
  - Cell-cell adherence junction
  - Golgi subcompartment
- **GOTERM_CC_BP**
  - Extracellular vesicle
  - Extracellular organelle
  - Membrane-enclosed vesicle
- **GOTERM_BP_BP**
  - Extracellular region part
  - Cell leading edge
  - Membrane region
basal B cell lines (Additional File 1: Fig. S2a,b). Using such method, patients were forced to classify in one of the two groups based on differences in gene expression or splicing patterns. Since basal B cell lines show more invasive, cancer stem cell-like phenotypes, we assessed whether these aggressive characteristics were translated to the “B-like” patient group through differences in disease specific survival (DSS) rates. Kaplan-Meier analysis of DSS did not show significant differences between the two subgroups of basal-like patients (Additional File 1: Fig. S2c,d). However, we did observe a tendency for “B-like” patients to have a poor survival compared to “A-like” when just looking at differences in splicing, contrary to expression levels (p-value=0.09 vs 0.57, respectively – Additional File 1: Fig. S2c,d).

In fact, it was not surprising that the transcript-level and splicing signatures did not translate directly from simplistic cell culture models to much more complex tumour patients with specific cell micro-environments and differences in cell heterogeneity. However, because the patients showed clear “A-like” and “B-like” signatures, we sought to develop a machine learning approach that would allow us to transfer part of the molecular and phenotypic observations found in cell-lines to patient data. Transfer learning is a recent research methodology that focuses on storing the knowledge gained when solving a problem, to apply it to a different, but related, one. Because we wanted to ensure that the newly developed cell-to-patient transfer learning algorithm could create interpretable models, we used a decision tree-based approach called Random Forest. In this cell-to-patient random forest classification method, we started by classifying basal A or basal B cell-lines based on their splicing and/or expression profile (Fig. 3a and Additional File 1: Fig.S3-S4). Then, once the model was trained on cell-lines, we would start integrating patient data gradually into the model. This was done iteratively by integrating at each round of classification the patients best predicted
FIGURE 3
to be basal A-like and basal B-like, so their added informative value could be used back to train the system and improve the next round of classification (Fig.3a). With this semi-supervised approach, the probability of assigning a patient to a specific subgroup evolves and improves at each round based on the updated information obtained from the best predicted patients, reaching at the end a stable population with the labels ‘basal A-like’, ‘basal B-like’ or ‘unclassified’ determined by the algorithm after 10-12 rounds (Fig. 3b,c and Additional File 1: Fig.S3b,c-S4b,c). Thanks to the gradual addition of patients at each round of training, there is a progressive increase, or decrease, in the feature importance of the splicing variants used to classify patients (Fig.3d and Additional File 1: Fig.S3d-S4d). Out of the 188 basal-like patients, 75 were classified as basal B-like, 88 as basal A-like and 25 could not be classified based on their splicing signature. Using only expression levels, there was a slight biased towards the basal A-like phenotype, with 56 patients classified as basal B-like, 122 as basal A-like and 10 unclassified (Additional File 1: Fig.S3b-c). Combining differentially spliced and expressed features seemed to be the most performant classifier with 84 patients as basal B-like, 100 as basal A-like and just 4 unclassified (Additional File 1: Fig.S4b-c). Taken together, depending on the features used (splicing patterns, expression levels or both), patients were differently classified in basal A-like or basal B-like.

An EMT-related basal B-specific splicing signature that marks poor prognosis.

To address which classifier translates the best to patients the invasive, EMT-like and drug resistant basal B phenotype found in cancer cells, we calculated the 5-year survival rate for each group of basal A-like and basal B-like issued from the three types of classification. Only basal B-like patients classified based on splicing levels had a poor prognosis compared to basal A-like patients (log-rank test p = 0.0067, HR =
4.87; IC95%: [1.37-17.28] in Kaplan-Meier analysis and univariate Cox regression)
(Fig.3e). Basal B-like patients subclassified based on gene expression levels, or gene
expression and splicing features, did not show significant differences in disease
survival rate (Additional File 1: Fig.S3e-4e), suggesting that splicing biomarkers might
be more informative to further subclassify basal-like patients based on prognosis. We
thus decided to focus on the role of alternative splicing in identify triple negative basal-
like breast cancer with poor prognosis.

To extract the most informative splicing features from the cell-to-patient transfer
learning classifier, we used the Boruta feature selection method [49]. This allowed us
to select the key splicing events responsible for the basal A/B classification without the
need to predefine arbitrary thresholds (Fig.4a). Out of the 217 differentially spliced
exons between basal A/B cell lines, just 25 were needed to subclassify breast cancer
patients in basal A or basal B-like tumours (Fig.4a and Additional File 2: Table S2).
Sashimi plots representing the splicing patterns of some of these basal B-specific
splicing events, such as the well-known splicing biomarker of cancer metastasis ENAH
[26] and the newly identified splicing biomarkers PLOD2, SPAG9 and KIF13a,
validated the observed changes in splicing between basal A and basal B-like patients
(Fig.4b-c and Additional File 1: Fig. S5a-b). Moreover, the changes in percentage of
spliced-in (PSI) of the 25 basal B-specific splicing events between the two subtypes of
basal-like patients correlated with the observed splicing changes between basal A/B
cell lines (Additional File 1: Fig.S5c-d), further supporting the transfer of knowledge
from the laboratory to the clinic. Finally, in the absence of publicly available RNA-seq
data on a second cohort of basal-like breast cancer patients, we took advantage of
three independent sequencing projects on breast cancer cell lines, different from the
ones used for the training of the semi-supervised classifier (Additional File 2: Table
FIGURE 4
Distribution of 52 independent breast cancer cell lines showed a 93% accuracy in the spatial segregation (t-SNE) of basal A from basal B cells based on the splicing pattern of the 25 newly identified splicing events (Fig.4d). Just three cell lines were misclassified as basal A (HCC38, SUM102 and MDA-MB-157). It is worth noting that one of these, HCC38, was also labelled as basal A in the DepMap portal (www.depmap.org), which validated our methodology and the specificity of the splicing signature towards a basal B-like phenotype.

Consistent with basal B cell lines being more mesenchymal, differences in the alternative splicing of these 25 basal B-specific splicing events in four different cellular models of EMT, coming from different cell types and methods of EMT induction [50–53], successfully clustered epithelial cells from mesenchymal with a pattern of splicing equivalent to basal A and basal B-like patients, respectively (Fig.4e). Of note, another 25 gene-based EMT-like splicing signature characteristic of luminal breast cancer tumours has also been identified capable of subclassifying mesenchymal-like breast cancer tumours with poor prognosis [38]. Consistent with a more luminal-specific signature, despite both marking EMT phenotypes, not more than six splicing events were found in common between the two splicing signatures (ATP5C1, CTNND1, KIF13a, PLOD2, SEC31a and SPAG9), which further supports the specificity of our newly identified splicing signature for basal-like triple negative breast cancer. Finally, using one of the first established molecular subtypes of triple negative breast cancer tumours based on gene expression, which is the Lehman classification [54], we found that basal B-like patients are mostly found in the categories associated with Mesenchymal stem-like (MSL) and Immunomodulatory (IM) subtypes (Fig.5a), which goes in line with a gene set enrichment of terms related to inflammatory responses and hallmark of EMT (Fig.5b).
FIGURE 5
When looking at the expression of well-known basal and EMT biomarkers in the two subpopulations of basal A/B-like patients, we found that basal A-like patients express classical basal/epithelial markers, such as E-cadherin, EPCAM and cytokeratin KRT5/KRT6/KRT14, together with ERBB3 and TOB1 which are markers of more differentiated, non-invasive cells [2]. On the other hand, basal B-like patients express classical EMT/mesenchymal markers such as Fibronectin, the EMT inducers Twist and Slug, and the Zinc-finger transcriptional regulators Zeb1 and Zeb2 which have recently been shown to confer stemness properties that can increase the plasticity and invasive capacity of the tumour cells [55] (Fig.5c-d). In line with a more aggressive, invasive phenotype, basal B-like patients express cytoskeletal (MSN, FN1) and extracellular matrix signalling proteins (TGFB1, TGFB2, FBN1, AXL), collagens (COL3A1, COL6A3) and proteases (MMP2, TIMP1, CTSC, PLA2, SERPINE1/2, PLAT), which are necessary for cell's migration and dissemination to distal organs during metastasis [2]. Finally, basal B-like patients overexpress a recently identified new marker of metastasis-initiating cells, the fatty acid receptor CD36 [20]. Clinically, the presence of CD36 positive cells has been correlated with a lower survival rate in many carcinomas, including breast cancer, and inhibition of CD36 impairs metastasis in breast cancer-derived tumours, turning this receptor into an important biomarker of tumour cell dissemination and a potential new target to reduce cell invasion. The fact that basal B-like tumour cells co-express this metastasis-initiating marker further strengthens the aggressive nature of this tumour subclass and the clinical relevance of the basal B-specific splicing signature in tumour progression and relapse.

Overall, we have identified a novel splicing signature, specific of triple negative breast cancer tumours, that marks patients with the poorest prognosis. This basal B-like splicing signature is responsible of a stem-like, EMT phenotype that favours tumour
growth, invasion of distal organs and increased drug resistance, which eventually leads to tumour relapse and metastasis. Interestingly, some of the genes differentially expressed in this basal B-like patients are well-known markers of metastasis-initiating cells, such as the alternatively spliced CTNND1 and PLOD2 genes or the fatty acid receptor CD36, turning these biomarkers into promising new targets for innovative therapies, such as the use of splicing specific antibodies [6, 26].

**A metastasis-related common regulatory pathway for the basal B-specific splicing signature.**

Hierarchical clustering of basal A and B cell lines based on the differential expression of RNA-binding proteins highlighted six RNA regulators, ESRP1, ESRP2, RBM47, TMEM63A, KRR1 and RBMS3 (Fig.6a) (Kruskal-Wallis p < 10^{-9}). Interestingly, ESRP1/2 and RBM47 are significantly less expressed in basal B-like than basal A-like patients (Fig.6b), consistently with the known inhibitory effect of these three splicing regulators in EMT progression and metastasis [53, 56, 57]. Available transcriptomics data in ESRP1/2 and RBM47 lung carcinoma NCI-H358-depleted cells [53] and RBM47 overexpressing breast cancer metastatic MDA-MB-231 cells [58] showed that 19 of the 25 splicing events responsible for the newly identified basal B-specific splicing signature could potentially be regulated by ESRP1/2 and/or RBM47 in breast cancer cells (Fig.6c-d). Importantly, in the cell types analysed, ESRP1/2 and RBM47 induce the epithelial, basal A-like splicing phenotype, suggesting a potential tumour suppressor effect for these splicing regulators (Fig.6e-g, 4e and Additional File 1: S5c-d). Consistently with this observation, low expression of RBM47 in basal-like breast cancer patients was associated with poor overall survival (log rank test p=0.031, HR=3.36, IC95%:[1.05 - 10.79] - Fig.6h-i), which supports previous experimental
**FIGURE 6**

**Part a:** Heatmap showing RNA binding proteins across 118 breast cancer cell lines from 5 RNA sequencing projects. The heatmap is color-coded to indicate the expression levels with red representing high expression and blue representing low expression.

**Part b:** A volcano plot illustrating the differential expression of genes. The x-axis represents the log2(TPM+1) values, while the y-axis represents the Z-score (log2(TPM+1)). The plot includes a significance threshold indicated by the significance levels (*** for p < 0.001 and **** for p < 0.0001).

**Part c:** Venn diagram comparing the overlap of genes between two conditions: ESRP KD vs 25 signature and RBM47 modulation vs 25 signature.

**Part d:** Venn diagram showing the overlap of genes between three conditions: GSE58381 (RBM47 OE), GSE75489 (ESRP KD), and GSE75491 (RBM47 KD).

**Part e:** Heatmap of expression levels for genes such as TSC2, DNM2, KIF13A, CD44, SLK, FNNP1, DN1, ARHGEF11, SPAG9, ANX6, CSF1, PLOD2, and CTNND1.

**Part f:** Heatmap of expression levels for genes such as SLK, KIF13A, SEC31A, SPAG9, DNM2, PLOD2, and CTNND1.

**Part g:** Heatmap of expression levels for genes such as SEC31A, SLK, FNNP1, and ENAH.

**Part h:** Graph showing the overall survival with ESRP1 as a marker.

**Part i:** Graph showing the overall survival with RBM47 as a marker.
evidence of a role for RBM47 in suppressing breast cancer metastasis and progression [57]. In fact, RBM47-dependent basal B-specific splicing events were found to be functionally interconnected by physical and/or genetic interactions, which points to the existence of a common basal B-specific regulatory network associated with tumour malignancy (Additional File 1: Fig. S6a). In support, most of RBM47-dependent basal B-specific splicing events play well-known roles in cell-cell adhesion (CTNND1) [59], cytoskeleton organization (ENAH, SLK, FNBP1) [60, 61], endocytosis (KIF13A, DNM2) [62] and association with the extracellular matrix (PLOD2) [63], which are all key processes for gaining the cell motility and invasiveness necessary in tumour metastasis (54-58). Of note, expression of just one of these basal B-specific splice variants, which are CTNND1, ENAH and PLOD2, is sufficient to lower the disease-specific survival rate of basal B-like breast cancer patients compared to basal A-like (Additional File 1: Fig. S6b-g). These splicing events could turn into promising new therapeutic strategies aiming at specific key regulatory genes instead of a pleiotropic splicing regulator that could have unsuspected secondary effects.

In summary, by taking advantage of extensive large-scale transcriptomics data from breast cancer cell lines and patients, we identified the first splicing signature capable of subclassifying basal-like tumours based on their aggressiveness and drug resistance. Importantly, novel splicing biomarkers of poor prognosis were identified that should be further studied in more functional assays to test their capacity to inhibit tumour invasion and metastasis. Results from these assays will open new perspectives in the development of improved target therapies and more accurate diagnostic profiles to identify the basal-like triple negative breast cancer patients with a higher chance of relapse.
DISCUSSION:

Cancer-specific dysregulation of alternative splicing is a promising source of cancer biomarkers and therapeutic targets to improve diagnostics and thus overall survival rate [64]. An increasing number of mutations at core spliceosome components, such as S3FB1 and U2AF1, or upregulation of specific splicing factors, such as SRSF1 and other members of the SR protein family, which are now considered oncogenes, have been intimately linked to tumour progression and malignancy [65]. Furthermore, an increasing number of alternatively spliced events, like CD44, ENAH, CTNND1 and FLNB, have been shown to impact cell invasion and metastasis on their own, making them promising new targets for more specific therapeutic strategies compared to the inhibition of splicing regulators [22, 23, 66, 67]. Effectively, splicing regulators are not only responsible for the regulation of splicing of a subset of genes, but they are also responsible for other RNA related functions such as translation, mRNA export and nonsense-mediated mRNA decay [57, 65], which can have numerous downstream deleterious effects when inhibited in a targeted therapy. By specifically targeting a key downstream splicing event, as in splicing-specific immunotherapy, a more cancer-specific and direct impact on the cell phenotype might be achieved (134, 135).

Large scale public molecular data sets on genomics (copy number and mutation), epigenomics, transcriptomics, proteomics, *in vitro* and *in vivo* cell invasiveness and response to anti-tumour compounds in a large number of patients (11,000 patients across 33 different tumour types from the Genome Cancer Atlas) and human-derived cell lines (1000 cancer cell lines across 36 tumour types from the Broad Institute’s Cancer Cell Line Encyclopaedia) has become an extraordinary toolbox to identify novel prognostic markers of early metastasis and/or resistance to specific drugs, which are the two major reasons for clinical relapse and low survival rate [68–
Unfortunately, the translatability of these pre-clinical findings is often limited since culture cells are not representative of the variety of individuals nor the biological reality of the tumour's multicellular environment. Yet, culture procedures are improving with the creation of organoids, and machine learning approaches combined with large-scale data mining are bypassing some of these important caveats. This is the case of our cell-to-patient random forest classifier approach, in which the addition at each round of selection of novel informative features, based on the patients classified in previous rounds, allows an algorithm to make use of the information learned from cell lines. Thanks to this approach, we were able to identify the first splicing signature, composed of 25 alternatively spliced exons, capable of subclassifying basal-like breast cancer patients into two subtypes with different prognoses: basal A- and basal B-like.

Actually, this newly identified basal B-like splicing signature underlined a stem-cell like EMT signature, with hallmarks of cell invasiveness and drug resistance. Five of these 25 alternatively spliced genes are well-known to play a role in cancer (ARHGEF11, CD44, CTNND1, ENAH, MBNL1) [74–76]. Six have been indirectly linked to tumour malignancy and are thus new splicing targets to study (CAST, CSF1, PLOD2, SLK, SPAG9, TSC2) [61, 63, 77–80]. The rest are completely unknown for their splicing role in cancer, even though changes in expression of some of them have been shown to play a role in tumour progression, chemosensitivity and metastasis without specifically addressing which splice variant (ATP5C1, BNIP2, FAT1, FNB1, SEC31A, ANXA6, DNM1, DNM2) [62, 81]. Of special interest are ARHGEF11 and CTNND1 splice variants. Both proteins are involved in cell-cell adhesion and the basal B-specific splice variants promote cell migration and invasiveness in several cancer types, such as breast cancer (13,54,74,67). Moreover, depletion of ARHGEF11 in basal breast cancer cells is sufficient to alter cell morphology, which suppresses the
cancer cell growth and survival in vitro and in vivo [75]. On the other hand, the existence of an isoform-specific antibody for CTNND1 pro-invasive splice variants turns this splicing candidate as a valuable new target to reduce tumour metastasis [82]. ENAH and CD44 are amongst the most studied splicing events impacting cancer and are well-known biomarkers of poor prognosis. ENAH’s inhibition decreases metastasis by slowing down tumour progression and reducing cell invasion and intravasation [83–85]. While the change to basal B splicing signature of CD44, a transmembrane protein that maintains tissue structure, is sufficient to drive an EMT and to increase cell invasion and plasticity by promoting stem cell characteristics [22, 86]. Interestingly, MBNL1 splicing regulation has also been involved in pluripotent stem cell differentiation [87] and cell viability via inhibition of DNA damage response [88]. Promising new splice variants with a potential link with cancer are CSF1, PLOD2, SLK, SPAG9 and TSC2. CSF1 is a macrophage marker which splice variant could correlate with infiltration of tumour-promoting macrophages [77, 89]. Changes in the alternative splicing of the procollagen-lysine PLOD2, which catalyses the deposition and cross-link of collagens in the extracellular matrix, have been intimately linked to EMT progression and cervical, breast, lung, colon and rectal cancer prognosis [40, 90]. Its inhibition reduced proliferation, migration and invasion of cancer cells, while its overexpression promoted cancer stem cell properties and resistance to drugs [63, 91]. SLK was identified as a prognostic biomarker in several cancers and is necessary for the induction of cell migration and invasion during EMT [61, 76, 92]. SPAG9 is a scaffold protein that organizes mitogen-activated protein kinases and has been associated with invasion in several types of tumours and prognosis [79, 93, 94]. Finally TSC2 basal B-specific splicing isoform cannot be phosphorylated by AKT, which leads to a continuously activated mTOR pathway and oncogenic autophagy [78]. More
functional studies on the impact of each of these cassette exons splice variants in cancer will increase our knowledge on tumour progression and metastasis with the long term goal of improving diagnostics and treatment. Of note, other types of splicing events, different from the studied cassette exons, have also been shown to play important roles in tumorigenesis, such as alternative splice sites and intron retention [71–73]. It is necessary to extend this type of approaches to all types of splicing events and validate them using independent cohorts of patients. The increase of accessible sequencing data in primary tumours will thus be essential to continue with this type of approaches.

Finally, it is interesting to note that these 25 alternatively spliced exons are basically dependent on three well-known splicing regulators, ESRP1/2 and RBM47, which are intimately linked to EMT and metastasis. ESRP1 is the major regulator of a newly identified epithelial-specific splicing signature [53]. Its expression in cancer cells promotes tumour growth and a mesenchymal-to-epithelial transition which are essential for the formation of new tumours at distal organs during metastasis [95, 96]. RBM47 is a newly identified splicing regulator of EMT that has also been associated with metastasis [57, 97, 98]. Through integrative analysis of clinical breast cancer gene expression datasets, cell line models and mutation data from cancer genome resequencing studies, RBM47 was identified as a suppressor of breast cancer progression and metastasis. It was found mutated in patients with brain metastasis and its expression was necessary to inhibit brain and lung metastatic progression in vivo [57]. Interestingly, despite regulating just 9/25 splicing events of the basal B-specific splicing signature, low expression of RBM47, and not ESRP1, correlated with a poor prognosis and lower survival rate in basal-like breast cancer patients, which increases the interest to design new therapies targeting this splicing regulator.
In fact, this basal B-specific splicing signature has highlighted a subpopulation of basal-like triple negative breast cancer patients differentially expressing several hallmarks of invasive, EMT-like aggressive cancer, such as the newly identified biomarker of metastasis CD36 [20]. CD36 is a fatty receptor expressed in metastasis-initiating cells. Neutralizing antibodies that block CD36 completely inhibited the formation of metastasis in orthotopic mouse models of human oral cancer, and CD36 inhibition impaired metastasis in human melanoma and breast cancer-derived tumours. Interestingly, the fatty acid-binding protein 7 (FABP7) correlates with a higher incidence of brain metastasis and lower survival rate in breast cancer patients, which all together points to a potential connection between fatty acid metabolism and metastasis in our subclass of basal-like breast cancer patients [99]. Furthermore, cells expressing our newly identified basal B-specific splicing signature also showed resistance to several EGFR inhibiting drugs. Therapies targeting EGFR have variable and unpredictable responses in breast cancer [100]. By better subclassifying sensitive from resistant tumour cells, diagnoses could be improved, which will impact the choice of treatment and thus the chances of tumour relapse. Extensive drug screening of cells derived from basal B-like patients combined with machine learning strategies to transfer the splicing knowledge obtained will certainly improve the identification of much more suitable treatments for triple-negative breast cancer cells and reduce tumour relapse, thus improving the survival rate.

CONCLUSION:

Taking advantage of extensive available experimental data in breast cancer cell lines, we performed a knowledge transfer to clinical data to identify the first splicing
signature capable of subcategorizing the most aggressive and difficult to treat type of breast cancer, which is basal-like triple negative breast cancer. Based on the pattern of splicing of 25 splicing biomarkers, we could identify two new subclasses of clinically relevant basal-like tumours, basal A and basal B-like, with different sensitivity to drugs and capacity to invade distal organs, which has a direct impact on prognosis. We propose that by testing all basal-like patients with this novel signature, patients with increased chances of creating early metastasis or tumour relapse could be closely monitored to improve their chances of survival. Similarly, by correlating alternative splicing patterns with drug resistance in cancer cell lines, or even cancer cells isolated from patients, more specific splicing biomarkers could be identified for the most adequate and personalized choice of treatment, which is one of the major challenges in triple negative breast cancer. Finally, the newly identified basal B-specific splice variants underline a stem cell-like, highly invasive EMT phenotype, with increased drug resistance, that could be used as novel therapeutic targets to reduce cancer metastasis and relapse, opening new perspectives into the development of improved and more specific treatments for triple negative breast cancer tumours.

METHODS

RNA-seq transcriptomics analysis: gene expression and alternative splicing

RNA-seq reads were aligned to the human genome (GRCh38, primary assembly) using STAR [101] version 2.5.2b with standard parameters. Gencode v25 (derived from Ensembl v85) was used for all analysis requiring annotation.
TPMCalculator [102] (v0.0.1) was used to compute Transcripts Per Million (TPM) values and obtain read counts. Q parameter was set to 255 to keep only unique mapped reads and ExonTPM value was used to consider only reads mapped to exons. Whippet-quant from Whippet software (v10.4) was used to compute Percentage Spliced-In (PSI) values for splicing analysis. Conjointly to Kruskal-Wallis testing, the output from Whippet-quant was further filtered to include only events for which the sum of inclusion counts (IC) and skipping counts (SC) was greater or equal to 10 for both sets of samples. Whippet-delta was used to compute differential splicing (deltaPsi) and probability that there is some change in splicing between conditions. Two heuristic filters were applied on splicing events as advised in whippet documentation; |deltaPsi| > 0.1 and P(|deltaPsi| > 0.0) >= 95 % were considered reliable parameters to filter biologically relevant AS events.

When necessary, Biobambam2 [103] (v 2.0.87) was used to transform bam files into fastq in order to be processed by Whippet.

Gene ontology (GO) analysis was done using the DAVID (v 6.8) [104] functional annotation tool (https://david.ncifcrf.gov/home.jsp) using Benjamini-Hochberg adjusted P-value cutoff of 0.05 to define a term as enriched. Go terms enrichment was restricted to GOTERM BP-FAT, GOTERM MF-FAT, and GOTERM CC-FAT, KEGG_PATHWAY and REACTOME_PATHWAY.

Gene Set Enrichment Analysis (GSEA v20.0.5) was carried out on the GenePattern [105] web platform using phenotype for permutation type and 1000 for number of permutations to execute. FDR cutoff of 25% for potential true positive finding was used
as documented in the GSEA user guide. Read counts were previously normalized using DESeq2 [106] (v 1.10.1) on the same Platform.

R version 3.6.2 was used all along this study excepted for GSEA.

All heatmaps were done online using Morpheus https://software.broadinstitute.org/morpheus/. Values were adjusted by Z-score (subtract mean and divide by standard deviation). Hierarchical clustering was done in Morpheus. We selected "Metric One minus pearson correlation" as a measure of distance between pairs of observation and "Average" as the linkage method. The clusters were done using rows and columns together. Columns were grouped by cancer subtypes.

Sashimi plots to look cassette exons events were done using ggsashimi tool [107].

**Machine Learning and feature selection:**

First, we construct a classifier to distinguish basal B / A cell lines using a Random Forest with 1000 trees. After, we applied this model to the TCGA patients. Based on Gini impurity, we computed the class probability to predict patient labelled as B-like or A-like. Then, mixing initial cell lines with a subset of patients classified with the more reliability (the ones picked up with higher class probability not passing below a threshold of P=0.6), we create a new model. Each addition of patients is called a round, during which a new model is created, giving new predictions (probabilities) for the remaining patients. By limiting the number of new patients added at each round (10 x n_current_round) (Fig.3c and Additional File 1: Fig.S3c-4c), the model can gradually learn from the patient data and avoid overfitting. With such conditions, we can observe
a gradual shifting in feature importance from the ones informative to classify cell lines
to the ones informative to classify patients and cell lines (Fig.3d and Additional File 1: Fig.S3d-4d). The algorithm stops when it can no longer incorporate the patients into one or the other group given the cut-off of P=0.6. ML analyse was done with Python 3.7.3 based on scikit-learn version 0.21.2.

To select the more efficient features that were able to separate B-like from A-like patients, we used Boruta package (0.3) implemented in python. We ran it 10 times with different random states, on the 217 features related to splicing and kept the ones that were present at least 7 times on 10. We ended with 25 AS features. Considering only these 25 AS features, we applied TSNE function from manifold package (with perplexity=20) to 3 other datasets of basal cell lines (n=56) to check the features were sufficient to distinguish spatially these cell lines according to their labels.

For the classification using only differentially expressed genes (Additional File 1: Fig.S3) or a mix of differentially spliced and expressed features (Additional File 1: Fig.S4), we applied the same strategy using the information from the 635 differentially expressed genes and the 217 differentially spliced exons scaling independently the values from the cell lines and patients with sklearn’s StandardScaler. We also had to reduce the probability threshold to 0.55 in the mixed model.

**Breast Cancer Annotation**

Basal B & A cells were labelled according to literature: Neve & al [28], Kao & al [33], Marcotte & al [108], Dai & al [109]. PAM50 intrinsic subtype were retrieved from https://www.cell.com/cancer-cell/fulltext/S1535-6108(18)30119-3 [48].

Claudin Low status was defined with script downloaded from https://github.com/clfougner/ClaudinLow/blob/master/Code/TCGA.r [110] using
dataset from http://download.cbiportal.org/brca_tcgapan_can_atlas_2018.tar.gz [111, 112].

**Survival Analysis**

Log-rank tests were performed using the functions `surv` and `survfit` from R package (survival v3.1.8). A different survival was considered significative if log rank test p-value was <0.05. Coxph function was also used for univariate Cox regression analysis in order to compute Hazard Ratio and 95% Interval of confidence. Kaplan–Meier curve were plotted using function `ggsurvplot` from R package `survminer` (0.4.6) Plots were truncated at 5 years, but the analyses were conducted using all of the data. All endpoints used for survival analysis in this study were retrieved from this study [113].

**Statistics**

Wilcoxon Rank Sum Test were used to assess statistical significance within boxplots. They were noted. P<0.05 (*), P<0.01 (**), and P< 0.001 (***) , P< 0.0001 (****).

Kruskal-Wallis Test was used to keep differential features for expression (TPM values) or splicing (PSI values) when Luminal, Basal A & B cell lines were compared and displayed in heatmap figures. A threshold of p-value <10-5 was used to filter out potential false positive and reduce the number of features in order to apply hierarchical clustering. This threshold was adapted depending on the number of samples in the comparison. For RNA binding proteins, a higher cut off of p< 10-9 was used because 5 projects were pulled together.

**Code**

Code and annotation files are available here.
Ethics approval and consent to participate

Patients data was obtained from The Cancer Genome Atlas upon agreement of TCGA ethics and policies

Consent for publication

All patients gave consent for publication of their personal information.

Availability of data and materials

All datasets are available in the Gene Expression Omnibus (GEO): GSE75489, GSE58381, GSE75491, GSE61220, PRJEB25042, GSE74881, GSE75492, PRJNA523380, PRJNA297219, PRJNA210428, PRJNA251383, PRJEB30617 (detailed in Additional File 2: Table S1) and The Cancer Genome Atlas (TCGA) repositories upon request (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000178.v11.p8)

Competing interests

The authors declare no competing interests.

Funding

Luco team is supported by the Agence Nationale de la Recherche [ANRJCJC - 2016 - EpiSplicing] and the Labex EpiGenMed [ANR-10-LABX-12-01]. Ritchie team is
supported by the Agence Nationale de la Recherche [ANRJCJC - WIRED], the Labex EpiGenMed [ANR-10-LABX-12-01] and the MUSE initiative [GECKO].

Authors’ contributions

JPV performed all the analyses. CL helped with the development of the semi-supervised classifier. MSC and AO helped with the discussion and writing of the manuscript. JPV, RL and WR designed the study and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Yaiza Nuñez-Alvarez and Sylvain Barrière for discussions.

List of abbreviations

AS: Alternative Splicing
CE: cassette exons
EMT: Epithelial-to-Mesenchymal Transition
CSC: Cancer Stem Cells
CTC: Circulating Tumour Cells
PSI: Percentage Spliced-In
TPM: Transcripts per Million
DSS: Disease Specific Survival
TCGA: The Cancer Genome Atlas
RBP: RNA binding proteins
Supplementary Information

Additional File 1: Figures S1-S6.

Fig. S1 - Allele-specific alternative splicing and its functional genetic variants in human tissues.

Fig. S2 - Hierarchical clustering and k-means of patients based on differential gene expression and splicing.

Fig. S3 - Semi-supervised Random Forest Classifier to transfer cell lines knowledge to patients using expression levels.

Fig. S4 - Semi-supervised Random Forest Classifier to transfer cell lines knowledge to patients using alternative splicing and expression levels.

Fig. S5 - In silico validation of basal B splicing signature.

Fig. S6 - Prognostic value of individual alternatively spliced genes from the basal B-specific signature

Additional File 2: Table S1-S2

Table S1 – GEO accession numbers for all the datasets analysed.

Table S2 – Name, coordinates (Hg38) and PSI mean value and standard error for the 25 exons of the basal B-specific signature in Basal A and Basal B cancer cells and patients. The difference in splicing levels between basal B and basal A is shown as deltaPSI.

Figure Legends

Figure 1. Basal cell lines are divided in two subgroups based on gene expression and splicing patterns. a. Heatmap of the Transcripts per Million (TPM) values of the 635 genes which differential expression can cluster breast cancer cell lines into basal
A and basal B (P-value < 10^{-3} by Kruskal-Wallis Test). b. Heatmap of the Percentage Spliced-In (PSI) values of the 217 exons which differential splicing can cluster breast cancer cell lines into basal A and basal B (P-value <10^{-3} by Kruskal-Wallis Test). c. Venn Diagram of the genes differentially expressed and/or spliced between basal A and basal B cancer cell lines. The overlap is not higher than expected by Fisher’s exact test, two tail (p=0.098)

Figure 2. Basal B cell lines show mesenchymal, stem-like and resistance to treatment characteristics. a,b,c. Gene Set Enrichment Analysis (GSEA) of differentially expressed genes between basal A and B cell lines for three different signatures: Mammary Stem Cell, EMT and Resistance to Gefitinib. Up-regulated genes in all signatures are enriched in basal B cell lines (FDR<0.25). d. Gene ontology analysis bar graphs for differentially expressed (left) and differentially spliced (right) genes between basal A and B cell lines. Gene ontology terms related to Cellular Component (GO_CC_FAT), Molecular Function (GO_MF_FAT) and Biological Process (GO_BP_FA) are shown in the y axis in blue, yellow and red, respectively. Benjamini false discovery rate (FDR, -log10) is shown on the x axis. Vertical lines mark an FDR threshold of FDR=0.05 (-log10(0.05)=1.3) for differentially expressed and spliced genes, respectively. e. Box plots of the median and 25th percentile of the CD44/CD24 log2 expression ratio for basal A and B cell lines. P-value is calculated using the Wilcoxon rank-sum test. f. Boxplots comparing IC50 values in basal A and B cell lines upon treatment with different drugs from the Genomics of Drug Sensitivity in Cancer 2 (GDS2) dataset. P-values are calculated using the Wilcoxon rank-sum test.

Figure 3. A Random Forest Classifier using knowledge transfer from cell lines
to patients. a. Workflow scheme: a random forest (RF) model is built using cell lines labelled as Basal B (red) or Basal A (blue). It is then run iteratively, integrating at each round patients whose probability to be classified in one group or the other is amongst the ten highest. The classifier stops when no more patients can be classified. b. Probability of a basal-like patient to be classified as basal B-like, basal A-like or unclassified over each round. Yellow lines indicate thresholds used to classify a patient as basal B-like (>0.6) or basal A-like (<0.4). c. Bar plot of the number of patients added at each round. Patients with the highest probability to be classified are sequentially incorporated to the input cell lines in order to create a new classifier for the next round of integration. d. Evolution of the feature importance at each round of iterative training. In red are the 10 splicing variants (features) most informative at the beginning of the transfer learning process. In blue are the 10 splicing variants most informative at the end. Only two exons remained informative from the beginning to the end (in blue and red). The name of the top 10 final most informative spliced genes are written in blue and in sequential order. e. Kaplan-Meier plots of disease specific survival in basal A-like (blue) and basal B-like patients (red). Hazard ratio (HR) and logrank p-value (P) discriminating the two groups are shown.

Figure 4. The basal B-specific splicing signature is associated to EMT features. a. Heatmap of the Percentage Spliced-In (PSI) values of the 25 cassette exons most informative to classify TCGA basal-like patients into basal B-like (red) or basal A-like (blue). Claudin low tumors are highlighted in green. b,c. Sashimi plots displaying ENAH and PLOD2 splicing patterns in randomly selected patients classified as basal A-like and basal B-like. d. Changes in alternative splicing of these 25 basal B-specific splicing events is sufficient to properly cluster 55 basal breast cancer cell lines from 3 unrelated
sequencing projects into basal B and basal A using t-SNE. Of note, three basal B cell
texts, HCC38, MDA-MB-157 and SUM102 were misclassified as Basal A cell lines (red
dots). Although HCC38 has also been classified as Basal A in the DepMap portal
(www.depmap.org). e. Heatmap of the PSI values of the 25 basal B-specific splicing
signature in public RNA-seq datasets from four different EMT projects. Basal B-like
events have the same splicing patterns as EMT-induced cells.

Figure 5. Basal B-like patients express hallmarks of EMT and metastasis that
leads to a poor prognosis. a. Lehman classification for basal A- and B-like patients.
**p<0.01 in Fisher's exact test, two tail, comparing basal B to basal A. b. Gene Set
Enrichment Analysis (GSEA) of the genes differentially expressed between basal A-
and B-like patients. Hallmark EMT and inflammatory response signatures are enriched
in basal B-like patients. c. Box plots of the median and 25th percentile of the expression
levels (in TPM) of major epithelial and mesenchymal-like EMT markers in basal A-like
(blue) and basal B-like (red) patients. d. Box plot of the mean and 25th percentile of
the expression levels (in TPM) of Basal-like non-invasive and mesenchymal-like
invasive markers in basal A-like (blue) and basal B-like (red) patients. ** P <0.01, ***
P <0.001, **** P <0.0001 in Wilcoxon rank-sum test comparing basal A-like to basal B-
like.

Figure 6. The basal B-specific splicing signature is co-regulated by ESRP1 and
RBM47. a. Heatmap of Transcripts per Million values for RNA Binding Proteins (RBP)
 diferentes in basal A and basal B cell lines (P-value <10^-9 by Kruskal-
Wallis Test). b. Box plots of the mean and 25th percentile of the expression levels (in
TPM) of the same RBP as in a, but in basal A-like and basal B-like patients. c,d. Venn
diagrams of the number of splicing events from the basal B-specific splicing signature
dependent on the splicing factors (SF) ESRP1/2 and RBM47 using a cutoff of
\(|\Delta \text{Psi}| > 0.1\) and a higher probability \(\geq 0.95\). e,f,g. Heatmaps of the PSI values of
the ESRP and RBM47-dependent exons from c and d in ESRP1/2 knock downed H358
cells, RBM47 overexpressed MDA-MB-231 cells and RBM47 knock downed H358
cells. h,i. Kaplan Meier plots for overall survival in basal-like TCGA patients expressing
the highest tercile (blue) or the lowest tercile (red) of ESRP1 and RBM47 expression
levels. HR (Hazard Ratio) and Logrank p-values (P) discriminating between groups are
shown.

References
1. Sims AH, Howell A, Howell SJ, Clarke RB. Origins of breast cancer subtypes and
2. Toft DJ, Cryns VL. Minireview: Basal-like breast cancer: From molecular profiles to
expression patterns of breast carcinomas distinguish tumor subclasses with clinical
implications. Proceedings of the National Academy of Sciences of the United States
2015.


1090 110. Fougner C, Bergholtz H, Norum JH, Sørlie T. Re-definition of claudin-low as a
1093 Cancer Genomics Portal: An open platform for exploring multidimensional cancer
1096 Integrative analysis of complex cancer genomics and clinical profiles using the
1099 An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality


Figure S1. Differential clustering of basal B cell lines based on gene expression and splicing patterns. a. Heatmap of gene expression levels, in Transcripts per Million (TPM) values, of 1729 genes differentially regulated between luminal, basal A and basal B cell lines (P-value < 10⁻⁶ by Kruskal-Wallis Test). b. Heatmap of exon inclusion levels, using Percentage Spliced-In (PSI), of 309 exons differentially spliced between luminal, basal A and basal B cell lines (P-value < 10⁻⁶ by Kruskal-Wallis Test). c. Heatmap of exon inclusion levels, using Percentage Spliced-In (PSI), of the 309 exons differentially spliced between cell lines in the 923 luminal and basal-like breast cancer patients available from the TCGA. We separate the basal-like patients in basal A-like and basal B-like based on the signature found in the cell lines.
Figure S2
Figure S2. Hierarchical clustering and k-means of patients based on differential gene expression and splicing. a-b. Using 188 TCGA patients classified as basal-like breast cancer, we applied hierarchical clustering followed by a k-means (n=2) on expression (a) or splicing values (b) characteristic of basal B cell lines. Each time, K-means distinguished two groups we named “B-like” (red) and “A-like” (blue). In a, k-means was applied to TPM expression values for the 635 genes differentially expressed between basal A and B cell lines, which were displayed in the heatmap annotated Expression. In b, k-means was applied to PSI values of the 217 differentially spliced exons between basal A and basal B cell lines, which were displayed in the heatmap annotated Splicing. c,d. Kaplan-Meier plots of disease specific survival (DSS) of basal-like breast cancer patients previously separated in two groups by the k-means algorithm (k=2) for expression and splicing. Logrank test p-values (P) between “B-like” (red line) and “A-like” (blue line) patient groups are shown.
Figure S3
Figure S3. Semi-supervised Random Forest Classifier to transfer cell lines knowledge to patients using expression levels. a. Heatmap of 81 genes TPM values for TCGA basal-like patients predicted as basal B-like (red) or basal A-like (blue) by the semi-supervised random forest classifier based on gene expression levels. Claudin low tumors are highlighted in green. Only the best features are represented. b. For all patients, we plot their probabilities to be classified as basal B-like, basal A-like or unclassified at each round. Dotted lines indicate thresholds used to classify a patient as basal B-like (>0.6) or basal A-like (<0.4). c. Bar plot showing the number of patients added at each round. Patients with the highest probability to be classified are sequentially incorporated to the input cell lines in order to create a new classifier for the next round of integration. d. Evolution of the feature importance at each round of iterative training. In red are the 10 splicing variants (features) most informative at the beginning of the transfer learning process. In blue are the 10 splicing variants most informative at the end. Only three exons remained informative from the beginning to the end (in blue and red). The name of the top 10 final most informative spliced genes are presented in sequential order. e. Kaplan-Meier plots of disease specific survival in patients classified as basal A-like (blue) and basal B-like (red) based on gene expression patterns. Hazard ratio (HR) and logrank p-value (P) discriminating the two groups are shown.
Figure S4
Figure S4. Semi-supervised Random Forest Classifier to transfer cell lines knowledge to patients using splicing and expression levels. a. Heatmap of 30 exons PSI values and 64 genes TPM values for TCGA basal-like patients predicted as basal B-like (red) or basal A-like (blue) by the semi-supervised random forest classifier based on differential splicing and gene expression levels. Claudin low tumors are highlighted in green. Only the best features are represented. b. For all patients, we plot their probabilities to be classified as basal B-like, basal A-like or unclassified at each round. Dotted lines indicate thresholds used to classify a patient as basal B-like (>0.55) or basal A-like (<0.4). c. Bar plot showing the number of patients added at each round. Patients with the highest probability to be classified are sequentially incorporated to the initial model in order to create a new classifier for the next round of integration. d. Evolution of the feature importance at each round of iterative training. In red are the 10 splicing variants (features) most informative at the beginning of the transfer learning process. In blue are the 10 splicing variants most informative at the end. The name of the top 10 final most informative spliced genes are presented in sequential order. With an asterisk we indicate the features that correspond to splicing events e. Kaplan-Meier plots of disease specific survival in patients classified as basal A-like (blue) and basal B-like (red) based on gene expression patterns. Hazard ratio (HR) and logrank p-value (P) discriminating the two groups are shown.
Figure S5. In silico validation of basal B splicing signature. a,b. Sashimi plots of KIF13A and SPAG9 patterns of splicing in randomly selected basal A-like and basal B-like patients. c,d. Box plots of the median and 25th percentile of the Percent Spliced-In (PSI) values for the 25 cassette exons in basal A/B cell lines and basal A-like/B-like patients. **** P <0.0001 in Wilcoxon rank-sum test comparing A to B.
Figure S6. Prognostic value of individual alternatively spliced genes from the basal B-specific signature. a. Network of functional association (GeneMania) between RBM47-dependent spliced genes from the 25 basal B-specific splicing signature. b,c,d. Kaplan-Meier curves of disease specific survival in patients expressing basal A-like (blue) or basal B-like (red) ENAH, CTNND1 and PLOD2 splice variants grouped by PSI terciles. Hazard ratio (HR) and respective logrank p-values (P) discriminating groups are shown. e,f,g. Box plots of the median and 25th percentile of the PSI values of the patients used in the survival curves.
<table>
<thead>
<tr>
<th>Used in Model</th>
<th>PRJNA2523380 (1)</th>
<th>PRJNA2523380 (1)</th>
<th>PRJNA25297219 (2)</th>
<th>PRJNA25297219 (2)</th>
<th>PRJNA2511383 (4)</th>
<th>PRJNA2511383 (4)</th>
<th>PRJEB30617(b)</th>
<th>PRJEB30617(b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRJNA210426(b)</td>
<td>SUM225 BASALA</td>
<td>BASALA</td>
<td>BT20 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1143 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA210426(b)</td>
<td>HCC1143 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1187 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1806 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA210426(b)</td>
<td>HCC1187 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1197 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1937 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA210426(b)</td>
<td>HCC1569 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1954 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA210426(b)</td>
<td>HCC1594 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1500 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA210426(b)</td>
<td>HCC70 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1500 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA2523380 (1)</td>
<td>HCC70 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1500 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA2523380 (1)</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1500 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA2523380 (1)</td>
<td>HCC1937 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1500 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA2523380 (1)</td>
<td>HCC1954 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1500 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA2523380 (1)</td>
<td>HCC1599 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1500 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA2523380 (1)</td>
<td>HCC1806 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1500 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA2523380 (1)</td>
<td>HCC1952 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1500 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA2523380 (1)</td>
<td>HCC1187 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1500 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA2523380 (1)</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1500 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA2523380 (1)</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1500 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA2523380 (1)</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1500 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA2523380 (1)</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1500 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA2523380 (1)</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1500 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA2523380 (1)</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1500 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA2523380 (1)</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1500 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA2523380 (1)</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1500 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA2523380 (1)</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1500 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA2523380 (1)</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1500 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA2523380 (1)</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1500 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA2523380 (1)</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1500 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA2523380 (1)</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1500 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>PRJNA2523380 (1)</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>MDAMB468 BASALA</td>
<td>BASALA</td>
<td>BASALA</td>
<td>HCC1500 BASALA</td>
<td>BASALA</td>
</tr>
<tr>
<td>ID</td>
<td>Coords</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>----------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLOD2</td>
<td>chr3:146077861-146077924</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNM2_a</td>
<td>chr19:10796060-10796199</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTNND1_b</td>
<td>chr11:57791493-57791673</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPAG9</td>
<td>chr17:50975862-50975901</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RUBCN</td>
<td>chr3:197691073-197691148</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTNND1_a</td>
<td>chr11:57789036-57789155</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EVI5L</td>
<td>chr19:7857091-7857124</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNM2_b</td>
<td>chr19:10808568-10808580</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAST</td>
<td>chr5:96726793-96726859</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAT1</td>
<td>chr4:186590367-186590403</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TSC2</td>
<td>chr16:2077597-2077726</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FNBP1</td>
<td>chr9:129915965-129915980</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BNIP2</td>
<td>chr15:59668102-59668138</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSF1</td>
<td>chr1:109923165-109923711</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANXA6</td>
<td>chr5:151110626-151110644</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARHGEF11</td>
<td>chr1:156938417-156938513</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENAH</td>
<td>chr1:225504990-225505053</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AP1B1</td>
<td>chr22:29329711-29329720</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP5C1</td>
<td>chr10:7806973-7807010</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNM1</td>
<td>chr9:128247923-128247935</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD44</td>
<td>chr11:35209964-35210054</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MBNL1</td>
<td>chr3:152446703-152446757</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEC31A</td>
<td>chr4:82842184-82842481</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KIF13A</td>
<td>chr6:17771113-17771218</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLK</td>
<td>chr10:104010815-104010908</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Alternative splicing plays a key role in protein diversity in healthy organisms. When the oncogenesis process is activated, this regulatory layer is disrupted, leading to a modification of the isoforms content of the cell and more widely of the tissue. It’s a key mechanism whose impact on tumor progression no longer needs to be demonstrated. Still the individual functions of these isoforms need to be elucidated, but also it is necessary to understand the extent to which biological processes are deregulated. The expression of a single gene can have drastic effects depending on its context, and depending on the isoform it expresses. When we refer to the transcriptome, we first think of the expression of genes. Due to its greater ease of interpretation, gene expression has been widely studied, but now with the arrival of new technologies and the development of machine learning methods, I hope that it will become possible to focus on splicing isoforms to decipher more subtle mechanisms.

Here, I will discuss the added value of AS in the EMT, a process that is renewed even several decades after its first description. I will comment our main results and discuss further analysis that could be performed. Finally, we will highlight how the field is evolving rapidly and I am going to underline some recent advances in, machine learning, cancer and medicine.

**Interesting contribution of AS in the evolving field of EMT**

During metastatic cascade, it’s obvious that specific mechanisms are deregulated in different proportions depending on the tissue, the microenvironment, and the cell-of-origin that led to the cancer. Mutations, large rearrangements of the genome play an important part in the establishment of the primary tumor. Genetic events in genes driving AS programs, or genes affected by AS is not a mandatory feature to associate AS with cancer and can be the result of a dysregulated biological process (Grosso, Martins, and Carmo-Fonseca 2008). It was suggested a long time ago that one of these deregulated processes might be the EMT.
EMT has long been viewed as a binary process with two cell populations, epithelial and mesenchymal, and is often defined by the loss of the epithelial marker E-cadherin and the gain of the expression of the mesenchymal marker vimentin (Pastushenko and Blanpain 2019). Now, this oversimplified definition has been redesigned, it leaves the door open to a different vision on what is has been established about this process in cell lines or even tumors. At the same time, a growing body of evidence suggest that an important alternative splicing program occurs during EMT and modulates cellular phenotype (Shapiro et al. 2011). From this observation, many studies report that the switch of one isoform can trigger an EMT (Brown et al. 2011; Ranieri et al. 2016; Tripathi et al. 2019). So, if only an isoform switch can have such an effect, taking part in a larger program, it must be possible to identify several markers that are regulated in a coordinated manner. Recently, a consensus statement published recently argue that EMT status cannot be assessed on the basis of one or a small number of molecular markers (J. Yang et al. 2020). Thus, identifying several isoforms that are part of a larger program makes sense. After all these facts, I considered that it seemed interesting to study if changes of isoforms could be observed and related to this EMT continuum (Figure 3-1), with the purpose of discovering new biomarkers to fight cancer progression.

![Figure 3-1 Transition states occurring during EMT](image)

**Figure 3-1 Transition states occurring during EMT**

Immunostaining for keratin 14 (K14) and vimentin (Vim) showing changes in their expression and in the morphology of skin tumor cells during EMT. Epithelial tumor cells have round shape and remain closely attached one to another, express K14, and are negative for Vim. Cells in early hybrid EMT state co-express K14 and Vim, are more elongated, but still cohesive. Cells in late hybrid EMT co-express K14 and Vim and are further elongated, acquiring fibroblast-like appearance. Mesenchymal tumor cells lost the expression of K14 while are uniformly expressing Vim, have fibroblast-like shape, and do not form cell–cell junctions (adapted from Pastushenko, 2018)
Of note, in a freshly published study, Qiu & al (Qiu et al. 2020) presents a twenty-five events signature of AS that were sufficient to classify epithelial and mesenchymal states of the tumors. Unlike our study, they isolated directly splicing events from breast cancer, stratifying by an EMT scoring function based on gene expression, and validated their prediction amongst cell lines. They analyzed the whole set of breast tumors without distinction between subtypes. A drawback of this study is that they use two genes (VIM and CDH1) to define epithelial and mesenchymal groups. As mentioned in Guidelines and definitions for research on epithelial–mesenchymal transition (J. Yang et al. 2020), the complex phenotype of EMT cannot only require a few molecular markers such as E-cadherin and Vimentin to be characterized. Another argument against this, is that VIM is expressed in MCF10A basal normal cells whereas these cells display an epithelial phenotype. On the other hand, what was really interesting is that they try to identify global signature unlike other studies which end up presenting a single candidate as a major contributor to the EMT (Ji Li et al. 2018). These references highlight the fact that AS in the EMT is an active field of research, paving the way to deeper research.

Our main achievements

During this PhD work, I explored the idea that an aggressive and deadly breast cancer can hide an EMT program that can lead to a faster extension of the disease. I focused on alternative splicing because its importance in EMT is not longer to be demonstrated, and it could open the road to new therapeutic strategies to fight cancer. Glioblastoma is not treated the same way melanoma is treated. This concept is also true in cancer from the same tissue. Different subtypes with specific morphological, transcriptional and epigenetic features exist within the same cancer type. This is why I have chosen to have a rational approach by focusing on a very particular subtype which heterogeneity, even with its own subtype, has been demonstrated (Lehmann et al. 2016, 2011). Based on the bibliography of breast cancer lines, I have chosen to use the observations made on a group of cell lines presenting mesenchymal and invasive characteristics. I wondered if I could transfer this to tumors of patients. By hacking random forest methodology, I transferred knowledge from cancer cell lines to human breast cancer tumors. I found that splicing events related to an EMT can distinguish two populations in the same subtype of breast cancer. I found that these two
populations have different outcome, where mesenchymal features seem to lead a shortened survival. It is questionable in what proportion this process is completed but the real interest lies in the fact that the AS brings a new layer of data to observe these changes and to refine what can be observed only with the change of expression. To a lesser extent, I also participated in the analysis of k-mer in breast cancer, letting the door open to the discovery of other biological events that need to be further investigated. So potentially, thanks to this work, new therapeutic targets can be explored to improve patient care.

**Application in the clinical field**

The method I developed let us discover a signature composed of a few genes that can easily be tested on a biopsy by RT-PCR at an affordable price, in a short period of time that could fit in a clinical environment. This signature could serve to monitor the evolution of a patient tumor toward a potential metastasis and that means clinicians could adapt treatments accordingly. Thanks to RNA-Seq from tumors of TCGA, I could test my classification technique but it must be mentioned that these raw data have restricted access and need to be access under specific request guidelines. Now, it’s not difficult to deliver a pre-trained model that can be easily port to production and reused. Some steps have already been taken in fundamental research where predictive models for genomics are shared among community via a centralized public repository (Avsec et al. 2019).

Also, it’s worth asking what would happen if clinical institutions had to apply and share the same approach based on knowledge from cell lines, on their own data from patients, in different type of cancers. Still when it comes to patients, privacy and security concerns always arise. Training data cannot be shared easily. Nevertheless, some federated learning approaches, where model-learning leverages all available data without sharing data between institutions, are emerging (Sheller et al. 2020; Rieke et al. 2020). This solution will make it possible to apply models trained on datasets of unprecedented size, to reach a better reliability and accuracy. Finally, this kind of initiative must emerge from a joint decision taken by clinicians, statisticians, IT (Information Technology), and bioinformatics teams to lead this digital health transformation. Nowadays, due to the fast evolution of data infrastructure, teams and organizations, it really seems possible.
Perspectives

One of the first things to test is the actual impact of these splicing events in EMT. Using CRISPR/dCas13 methods, the lab is now capable of inducing a splicing switch at a specific exon. I would thus test the effect in epithelial cells of inducing a switch towards the mesenchymal isoform. Priority will be given to the splicing events shown to have a prognostic value on breast cancer. Surprisingly, not all the newly identified basal-specific splicing targets are significantly associated with survival, raising the hypothesis that some of them must have a stronger impact on tumor progression and thus outcome. In particular, mesenchymal PLOD2 (Procollagen-Lysine,2-Oxoglutarate 5-Dioxygenase 2) isoform, had a strong link with a bad survival. High PLOD2 expression was associated with poor prognosis in glioblastoma (Yangyang Xu et al. 2017) and contributes to drug resistance in laryngeal cancer by promoting cancer stem cell-like characteristics (Sheng et al. 2019). Previous studies revealed that 2-oxoglutarate and the iron-dependent dioxygenases superfamily function as a hydroxylase/demethylase and that they hydroxylate or demethylate molecules such as transcription factor, histones, and DNA as substrates. Indeed, it has been reported that these enzymes play various roles in cell cycle and gene expression and control of invasion/metastasis of cancer cells in multiple cell lines via modified molecules (Markolovic, Wilkins, and Schofield 2015). Moreover, PLOD2 was described as an enzyme catalyzing collagen cross-linking and thus playing a role in migration and invasion (Du et al. 2017). CD44 mesenchymal isoform was not associated with survival when I looked at it individually but its role with tumor progression and poor prognosis has been widely described elsewhere (Gotoda et al. 1998; Fang et al. 2016; Pereira et al. 2020; C. Chen et al. 2018). Interestingly, a recent study revealed that CD44 takes part in an alternative iron-uptake mechanism that prevails in the mesenchymal state of cells (Müller et al. 2020). This mechanism is enhanced during EMT transition, in which iron operates as a metal catalyst to demethylate repressive histone marks that govern the expression of mesenchymal genes. All taken together, PLOD2 seems an interesting candidate to study its implication in the role of iron in cancer development and EMT.

A downside of our work is that I was unable to test our result in a different cohort of patients other than TCGA. Importantly, our methodology could be applied to other
cell lines of a different type but this statement should be taken with caution as not all tissues have a large number of cell lines available, and especially the classification in distinct groups has not been as documented as in the case of breast cancer. Nonetheless, it’s worth mentioning that there are techniques based on gene expression that can calculate the mesenchymal nature of a cell line (Foroutan et al. 2018; Tan et al. 2014). Another idea could be to use induced EMT RNA-seq projects, specific to the tissue I want to explore, to extract a signature of splicing. Then, this signature could be applied to cancer cell lines in order to stratify them, before exploring the real tumor of the corresponding tissue as I did. Finally, I could attribute a score based on alternative splicing and I think both splicing and expression should be considered together to give a more accurate EMT scoring.

During this work, I was surprised to find that few tools exist to study the impact of expressing different splicing events in survival, while there are plenty of web applications to explore the prognostic value of gene expression, such as Kaplan-Meier Plotter (H. Zheng et al. 2020). Even the well known BioPortal for Cancer Genomics (Cerami et al. 2012), that provide access to multiple types of genomic and survival data, does not offer this function. For example, Saraiva-Agostinho developed recently Psichomics (Saraiva-Agostinho and Barbosa-Morais 2019), a tool to interactively performs survival, dimensionality reduction and median- and variance-based differential splicing and gene expression analyses that benefit from the incorporation of clinical and molecular sample-associated features (such as tumor stage or survival). It's currently packaged in R so it can hardly be used by someone with no informatic background. Even if many survival analyses of alternative splicing events emerge (J. Zhu, Chen, and Yong 2017; D. Zhang et al. 2019; X. Chen et al. 2019), there is no user friendly web based interface to make fast queries on a centralized resource. So, there is a niche to develop web applications for users who want to quickly explore the link between their AS of interest and prognosis in cancer.

Also, during this work, it’s worth mentioning that I tried to tackle our problem using semi-supervised approach where I took benefit from the knowledge we had from cancer cell lines. But we could also have explored models that directly classify patients based on survival as some new tools do. Cox-nnet is an artificial neural network method for prognosis prediction of high-throughput omics data. It was developed using
gene expression but certainly can be extended to splicing (Ching, Zhu, and Garmire 2018). Reboot is another approach to identify genes and splicing isoforms associated with cancer patient prognosis (Santos, Guardia, and Santos 2020). It uses a multivariate strategy with penalized Cox regression (LASSO method) combined with a bootstrap approach, to find gene or transcript signatures (not PSI) relevant to patient prognosis. One advantage of our approach is that the model was designed using knowledge from cell lines. So, it is easier to go back to the bench and study the function of these splicing events in cancer.

As I mentioned in the introduction, several attempts were performed to divide basal-like breast cancer subtype into smaller subgroups based on gene expression (Burstein et al. 2015; Jiang et al. 2019; Lehmann et al. 2016). First, I would like to explore the k-mer content of each of these subgroups to see if singular events could be found. More interestingly, I would like to apply our k-mer classification and annotation methods, over the groups I found with different prognosis using our custom random forest approach based on AS. That way I could potentially discover more therapeutic targets.

For now, a drawback of k-mer is that you always need to go back to something biologically meaningful as gene expression or alternative splicing, in order to describe a biological event. So, there is need to develop large resources of k-mer, in different tissue, cell lines, conditions and diseases. In this way, it will be possible to test directly if a k-mer list is enriched in a disease or specific conditions like GSEA do with gene expression for biological pathways. The same problem is true for lists of alternative splicing exons when one wants to know if a specific set of exons is enriched in a signature of stem cells or apoptosis. Signatures are mostly described with gene expression and, to our knowledge, majority of existing tools for pathways enrichment, are gene expression based. To overcome this, Tranchevent & al, proposed an approach based on exon-ontology focusing on exon-encoded protein features, instead of gene level functional annotations, to discover protein features enriched by list of AS (Tranchevent et al. 2017). Also, for the study of splicing, some resources started to propose large repositories of data from tissues that can be manually annotated by all the users to retrieve AS functionality (Tapial et al. 2017).
Thanks to high performance computing and new computational techniques, we could imagine to process an incredible amount of data in order to construct a large repository of k-mer. Then it would become possible to interrogate directly in which condition or pathology a k-mer is enriched. Still taking isoforms as a parallel example, other study have aligned 21,504 Illumina-sequenced human RNA-seq samples from the Sequence Read Archive (SRA) to the human genome and compared the detected exon-exon junctions with known junctions (Nellore et al. 2016) to further study transcriptome complexity. This illustrates perfectly the fact that intensive computation on big data in bioinformatic can be real and applied to create a large resource of k-mer annotated.

**Challenges and concluding remarks**

Even if the occurrence of EMT during in vitro models is well documented, the role of EMT in patient outcomes remains controversial due to the complex content of a tumor (Iwatsuki et al. 2010; Jolly et al. 2017). Multiple groups have linked gene expression of EMT-associated gene signatures to increased inflammatory immune response in multiple cancer types (Mak et al. 2016; Y. et al. 2016; Romeo et al. 2019). It is often unclear whether clinical EMT signatures originate from mesenchymal malignant cells as opposed to tumor stromal cells (e.g., fibroblasts), which express EMT canonical markers (McCorry et al. 2018; Williams et al. 2019). For example, claudin low breast tumors, which show enrichment for EMT markers also overexpress genes associated with immune response and stroma (Sabatier et al. 2014; Prat et al. 2010). However, novel studies demonstrated recently their true existence (Pommier et al. 2020; Fougner et al. 2020), without neglecting the fact that non-tumor cell infiltration is undoubtedly an important feature of the claudin-low tumor microenvironment, and may even be the feature that induces EMT in this subtype. Anyway, the precise mechanisms by which microenvironment influence cell fate decision during EMT are still unknown.

However, researchers seems to agree on the fact that the display of mixed epithelial and mesenchymal traits by individual cells appears to be the norm rather than exception (Derynck and Weinberg 2019; J. Yang et al. 2020). For example, single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and
neck reveals a partial EMT program regulated by the microenvironment occurring at the leading edge of primary tumors (Puram et al. 2017). By contrast to some lineage-tracing experiments that failed to identify cells in the metastatic site that have undergone EMT (Fischer et al. 2015; X. Zheng et al. 2015), dynamic changes in epithelial and mesenchymal composition of circulating breast tumor cells have been described (Yu et al. 2013). In parallel, it was found in another model, that most tumors lose their epithelial phenotype through an alternative program involving protein internalization rather than transcriptional repression. It results in a partial EMT phenotype, used by carcinoma cells to migrate as clusters (Aiello et al. 2018). Recent studies have attempted to better define EMT states using single-cell approaches; Pastushenko & al demonstrated the existence of partial EMT states in mammary and skin cancer by examining a large number of surface markers with flow cytometry and single-cell RNA-sequencing (Pastushenko et al. 2018). Partial EMT states were identified also in ovarian cancer specimens with mass cytometry (V. D. Gonzalez et al. 2018). In the context of these observations, there are still many questions unanswered that I hope will be able to answer in a near future due to evolving (wet and dry) techniques of analyze.

The combination of computational approaches (Goecks et al. 2020; Eraslan et al. 2019) and novel technologies such as single-cell sequencing (Jackson et al. 2020), chromatin profiling, or in vivo intravital microscopy (Zhao et al. 2016), should help to better understand the dynamics and the molecular mechanisms controlling EMT related cancer heterogeneity. Nowadays, the fields of single-cell, long-read sequencing, and spatial transcriptomics, are evolving at an incredible rate. Tilgner & al, recently produced an analyze of brain-regions specific splicing at an incredible resolution (Joglekar et al. 2020) using all the technologies mentioned above. They provide a robust means of quantifying isoform expression with cell-type and spatial resolution that could benefit to the study of isoforms in tumor and its microenvironment. Following the initiative of TCGA, the Human Tumor Atlas Network (HTAN), part of the National Cancer Institute (NCI) Cancer Moonshot Initiative, will establish a clinical, experimental, computational, and organizational framework to generate informative and accessible three-dimensional atlases of cancer transitions for a diverse set of tumor types (Rozenblatt-Rosen et al. 2020).
The European counterpart, so called Lifetime project (Bertero et al. 2020), will track human cells during the onset and progression of complex diseases, not only cancers. This huge project aims to integrate single-cell multi-omics and imaging, artificial intelligence and patient-derived experimental disease models during progression from health to disease. The way we do science is going to be completely transformed with profound changes in the way data is handled and the techniques used to interpret it. As Goecks & al discuss in their “perspective” article, machine learning will have a central role to play in solving problems related to genetic heterogeneity and cellular mechanisms underlying diseases (Goecks et al. 2020). The predicted deluge of biological data will certainly give birth to unprecedented discoveries in the field of cancer, and alternative splicing will certainly not be left at the doorstep.

**Figure 3-2 The Human Tumor Atlas Network (HTAN)**

HTAN centers will take measures at multiple scales of resolution (molecular, ultrastructural, cellular, histological, anatomical and clinical. Most centers will use both molecular and spatial profiling methods to interrogate cell-type composition, cell-cell interactions, and spatial structures. It’s a massive effort to facilitate clinical and structural predictions. (Adapted from Rozenblatt-Rosen, 2020)
Figure 3-3 LifeTime European Project

Key technologies envisioned by the LifeTime initiative. Integration and analyze of large, longitudinal multi-omics and imaging datasets will require the development of new pipelines and machine learning tools. (Adapted from Bertero, 2020)


Bennett, Kelly L., Brett Modrell, Brad Greenfield, Armando Bartolazzi, Ivan Stamenkovic, Robert Peach, David G. Jackson, Frances Spring, and Alejandro Aruffo. 1995. “Regulation of CD44 Binding to Hyaluronan by Glycosylation of Variably Spliced
Exons.” *Journal of Cell Biology*. https://doi.org/10.1083/jcb.131.6.1623.


Chen, Yao, Xiaqin Lu, Diego E. Montoya-Durango, Yu Hua Liu, Kevin C. Dean, Douglas S. Darling, Henry J. Kaplan, Douglas C. Dean, Ling Gao, and Yongqing Liu. 2017. “ZEB1 Regulates Multiple Oncogenic Components Involved in Uveal Melanoma Progression.” *Scientific Reports.* https://doi.org/10.1038/s41598-017-0079-x.


Epifano, Carolina, Diego Megias, and Mirna Perez-Moreno. 2014. “P120-Catenin


Ghigna, Claudia, Silvia Giordano, Haihong Shen, Federica Benvenuto, Fabio Castiglioni, Paolo Maria Comoglio, Michael R. Green, Silvano Riva, and Giuseppe Biamonti. 2005.


Associated Alternative Splicing.” *Nature Communications.* https://doi.org/10.1038/s41467-020-14304-1.


Fibrosis.” *Cell and Tissue Research.* https://doi.org/10.1007/s00441-016-2464-0.


Xu, Yangyang, Lin Zhang, Yuzhen Wei, Xin Zhang, Ran Xu, Mingzhi Han, Bing Huang, et al. 2017. “Procollagen-Lysine 2-Oxoglutarate 5-Dioxygenase 2 Promotes Hypoxia-Induced Glioma Migration and Invasion.” *Oncotarget*. 119


Yoshida, Kenichi Ogawa, Seishi. 2014. “Splicing Factor Mutations and Cancer.”


GLOSSARY

Antisense oligonucleotides (ASOs)
Alternative splicing (AS)
Cancer Alternative Splicing Changes (CASCs)
Cancer Cell Line Encyclopedia (CCLE)
Centers for Disease Control and Prevention (CDCP)
Contiguous Splice Graph (CSG)
Copy Number Alterations (CNA)
Copy number variation (CNV)
Colorectal cancer (CRC)
Cancer Stem Cells (CSC)
Differential Gene Expression (DEF)
Estrogen Receptor (ER)
Genetically Modified Mice (GEMs)
Hazard Ratio (HR)
HER2 (Human Epidermal Growth Factor Receptor-2)
Hierarchical Clustering Analysis (HCA)
High-Throughput Sequencing (HTS)
Invasive Ductal Carcinoma (IDC)
Invasive Lobular Carcinoma (ILC).
Kaplan-Meier (KM)
Machine Learning (ML)
METABRIC (Molecular Taxonomy of Breast Cancer International Consortium)
Next generation sequencing (NGS)
Nonsense-mediated decay (NMD)
Progesterone Receptor (PR)
Random Forest (RF)
SNP (Single Nuclear Polymorphism)
Splicing factor (SF)
TCGA (The Cancer Genome Atlas)
TNBC (Triple Negative Breast Cancer)
Tumor-Initiating Cell (TIC)
ANNEXES
GECKO is a genetic algorithm to classify and explore high throughput sequencing data

Aubin Thomas¹,⁶, Sylvain Barriere¹,⁶, Lucile Broseus¹, Julie Brooke¹, Claudio Lorenzi¹, Jean-Philippe Villemin¹, Gregory Beurier², Robert Sabatier³, Christelle Reynes³, Alban Mancheron⁴,⁵ & William Ritchie¹

Comparative analysis of high throughput sequencing data between multiple conditions often involves mapping of sequencing reads to a reference and downstream bioinformatics analyses. Both of these steps may introduce heavy bias and potential data loss. This is especially true in studies where patient transcriptomes or genomes may vary from their references, such as in cancer. Here we describe a novel approach and associated software that makes use of advances in genetic algorithms and feature selection to comprehensively explore massive volumes of sequencing data to classify and discover new sequences of interest without a mapping step and without intensive use of specialized bioinformatics pipelines. We demonstrate that our approach called GECKO for GEnetic Classification using k-mer Optimization is effective at classifying and extracting meaningful sequences from multiple types of sequencing approaches including mRNA, microRNA, and DNA methylome data.

¹Institute of Human Genetics, CNRS UPR1142, Machine learning and gene regulation, University of Montpellier, Montpellier, France. ²AGAP, Univ Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France. ³IGF, Centre National de la Recherche Scientifique, INSERM U1191, University of Montpellier, Montpellier, France. ⁴LIRMM, Université de Montpellier, CNRS, UMR5506, Montpellier, France. ⁵Institut Biologie Computationnelle, Montpellier, France. ⁶These authors contributed equally: Aubin Thomas, Sylvain Barriere. Correspondence and requests for materials should be addressed to W.R. (email: william.ritchie@igh.cnrs.fr)
Studies of variation in gene expression, initially through probe-based technology and more recently high throughput sequencing (HTS), have considerably advanced knowledge of disease etiology and classification\(^1\). The recent promotion of HTS across a wide spectrum of diseases has generated a wealth of data that measure gene expression and transcript diversity but also explore its putative genetic and epigenetic regulators. Still, despite more than a decade of development, computational analysis and integration of these data presents a major challenge. Each type of HTS experiment is compartmentalized to a set of computational pipelines and statistical approaches that often require a full-time bioinformatics specialist. In addition, most of these pipelines rely on a reference genome or transcriptome and thus cannot inherently account for the diversity in non-reference transcripts or individual variations\(^4\). To remove the requirement of a reference, recent methodologies use \(k\)-mer representation; they directly compare the counts of nucleotide sequences of length \(k\) between samples\(^3\). These approaches have been successful at detecting novel transcripts but only on a very small subset of RNA sequencing data\(^4\) and would be impossible to implement for the classification of large patient cohorts using the entire transcriptome. In the field of metagenomics, numerous algorithms have been developed to discover unique \(k\)-mers or \(k\)-mer signatures to classify organisms\(^5,6\). However, these were developed for organisms with smaller genomes that do not have billions of different \(k\)-mers. In addition, they were designed for inter-species studies where unique \(k\)-mers can be attributed to the genomes of different taxonomic identities.

Exploring a large set of \(k\)-mers to classify samples can be framed as a global optimization problem for which many recent approaches have been published and compared\(^8\). Amongst these is a class of nature-inspired algorithms termed Genetic Algorithm which are based on the processes of mutation, crossing over and natural selection. These have appealing properties that could apply to the exploration of a large set of \(k\)-mers. They have low memory requirements because they explore only part of the data at each stage and they can produce multiple solutions that fit well with biological interpretation of data. However, despite these properties, genetic algorithms are rarely used to optimize problems with relatively small sample sizes and such a large number of parameters, in this case billions of \(k\)-mers.

We have created a novel approach and associated software called GECKO for genetic classification using \(k\)-mer optimization that is especially designed for HTS data. GECKO is based on \(k\)-mer decomposition coupled with an adaptive genetic algorithm that explores HTS data from two or more input conditions. This algorithm searches for groups of \(k\)-mers that, combined together are highly informative; they are able to classify the input categories with high accuracy. Because GECKO uses \(k\)-mer counts, it can theoretically be applied to any type of HTS experiment and does not rely on a reference genome or transcriptome. Here, we successfully apply GECKO to a variety of biological problems and sequencing data. These include microRNA (miRNA) sequencing to classify normal blood cells, mRNA sequencing to classify subtypes of breast cancer and to predict response to chemotherapy, and bisulphite sequencing (BS-seq) on normal versus chronic lymphocytic leukemia (CLL) samples. Regardless of the type of data, GECKO finds small, accurate signatures that classify these samples and could thus be used as diagnostic and prognostic markers. In addition, by visualizing how the genetic algorithm evolves to find solutions, GECKO can be used to explore novel sequences or groups of functionally related sequences associated with normal biology and disease.

**Results**

GECKO is designed around two main steps; these are a \(k\)-mer matrix preparation step and an adaptive genetic algorithm (Fig. 1).

The \(k\)-mer matrix preparation, uses an input sequencing file (.bam or .fastq) to create a matrix of \(k\)-mer counts; that is the number of times a sequence of length \(k\) appears in each sample (\(k = 30\) by default). This matrix is filtered for \(k\)-mers with low counts and non-informative or redundant \(k\)-mers (see the section “Methods”). Then, during the second step an adaptive genetic algorithm will explore the matrix to discover combinations of \(k\)-mers that can accurately classify input samples. The adaptive genetic algorithm starts by creating thousands of digital individuals; these are groups of randomly selected \(k\)-mers. The set of individuals is called a population. This population will then go through phases of mutation, where individuals replace one of their \(k\)-mers with another randomly selected \(k\)-mer; a phase of crossing-over where individuals exchange a portion of their \(k\)-mers with each other and selection, where individuals that do not classify the input samples well enough will be removed from the population and replaced. Mutation allows GECKO to explore local solutions similar to the individual to be mutated; crossing-over, allows GECKO to explore a broader set of solutions and reduces the chances of getting stuck in a local minimum (see the section “Methods”). Each cycle of mutation, crossing-over, and selection is called a generation. By default, GECKO will iterate through 20,000 generations or stop when the number of new solutions discovered throughout generations slows down (see stopping criteria in the section “Methods”). This algorithm is called adaptive because the mutation and crossing-over rates depend on how well individuals in the population perform. Individuals that perform well have lower rates to prevent them from changing drastically and thus enabling them to converge.
faster to a solution; individuals that do not perform well will have higher rates to enable wider exploration of solutions.

In the analyses presented in this study and by default in the software, GECKO’s performance is systematically tested on 1/6th of the data that is randomly selected and set aside before running the algorithm (see the section “Methods”). This test set allows us to evaluate the accuracy and overfitting for each run; it measures whether the algorithm fits too closely to the training set and thus will not correctly predict future input samples. GECKO is thus run on the remaining 5/6th of the data with cross-validation at each generation of the algorithm.

Classifying miRNA sequencing data of blood cells. We first tested GECKO’s performance on a miRNA expression data of seven types of blood cells sorted from 43 healthy patients for a total of 413 samples. We ran GECKO on this dataset using 20-mers (k-mer size of 20; miRNAs generally vary in size from 20 to 23) to find a set of k-mers that could correctly classify the seven blood-cell types.

After 6000 generations (15 h on 15 cores; see Supplementary Table 1 for parameters and Supplementary Fig. 1 for runtimes and memory usage) GECKO discovered an individual composed of only three k-mers (ACCCGTAGAACCGACCTTGC, CCCCA GGTGTGATTCTGATA, AGTGCATGACAGAACTTGGG) that could distinguish the groups with 0.96 accuracy (Fig. 2a, b and Supplementary Data 1 and 2).

In the initial study, the authors described a signature of 136 cell-type-specific miRNAs. These 136 miRNAs could classify the groups with 0.97 accuracy. Thus, we found a much smaller signature that could classify the seven blood-cell types with similar accuracy without the use of a miRNA-dedicated bioinformatics pipeline.

We then aligned the three k-mers discovered by GECKO to a database of known miRNAs\(^\text{10}\). Two of these mapped perfectly to miRNAs 152-3p and 99b-5p, which were annotated in the original study as specific to NK cells and T helper cells, respectively. The third mapped to miRNA 361-3p which was not found to be specific to any of the seven cell types and was thus ignored in the initial study. Separately, the first two k-mers could classify one cell-type each and the third would have been overlooked. Together these three k-mers classify all seven groups with high accuracy because of their contrasting expression between each cell types (Fig. 2c).

Classifying breast cancer subtypes using mRNA sequencing data. Breast cancer is a heterogeneous disease in regards to response to treatment and its transcriptional background. Defining the subtypes luminal A (LumA), luminal B (LumB), HER2-enriched (HER2) and basal-like are crucial for prognosis and predicting outcome of breast cancer. These subtypes were initially defined through unsupervised clustering of gene expression and are currently identified using a standard qPCR assay of 50 genes called the PAM50\(^\text{11,12}\). To assess whether GECKO could identify k-mers that classify breast cancer subtypes, we used a dataset of 1087 mRNA-Seq breast cancer samples from the Cancer Genome Atlas Pan-Gyn cohort\(^\text{13}\) (patients per class: Basal 175, Her2 73, LumA 513, LumB 185). We ran GECKO for 20,000 generations (75 h on 15 cores; see Supplementary Table 1 for parameters and Supplementary Fig. 1 for runtimes and memory usage) and extracted the highest scoring individual at its term (Supplementary Table 2). We then tested how well these k-mers classified the four cancer subtypes compared to PAM50 expression values calculated as transcript per million (TPM). Both the k-mer counts and PAM50 TPMs were trained using a linear support vector machine (see the section “Methods”) with identical training data and evaluated on the same test set. The 10 k-mers had higher accuracy rates compared to the PAM50 on all four classes (Fig. 3 and Table 1).

We then further inspected the 10 k-mers discovered by GECKO by mapping them to the human genome. We found

---

**Fig. 2** GECKO can accurately classify miRNA data from seven types of blood cells using three k-mers. a GECKO output showing the separation of the seven blood-cell types at each generation (G) of GECKO analysis using t-SNE visualization applied to k-mer counts. b GECKO output showing the accuracy of separation for the training and test set across 6000 generations. c variance stabilized counts of the three miRNAs that correspond to the three k-mers discovered by GECKO across the seven blood-cell types (n = 43 biologically independent donors).
that four of the k-mers mapped to genes from the PAM50 list (FOXC1, ESR1, KRT14, KRT17). Three others mapped to genes NISCH, TPX2, and ATF3, the first of which is linked to breast cancer aggressiveness\(^3\) and the two latter both affect cell viability in breast cancer cells\(^4,11\). The three last k-mers mapped to three genes KLHL6, KANSL2, and PHF10 shown to be involved in tumorigenesis but not in breast cancer\(^{16-18}\). Of the 10 k-mers, 3 map to coding regions and 7 map to 3 untranslated regions for which multiple isoforms exist. k-mer counting can thus integrate alternative transcription to classify mRNA-Seq samples.

Classifying response to chemotherapy of triple negative breast cancer on small sample sizes of mRNA-Seq. We then tested GECKO on a dataset with more heterogeneous cell populations and smaller sample sizes. We used a cohort of triple-negative breast cancer patients, an aggressive, heterogeneous subtype of breast cancer with poor outcomes. This cohort taken from the Breast Cancer Genome Guided Therapy (BEAUTY) study\(^{19,20}\), was divided into 19 patients that had a complete response to chemotherapy and 20 patients that did not. In such cases of small sample size and high heterogeneity, we recommend using GECKO’s voting mode (Fig. 4a).

This mode compensates for bias that may be introduced when splitting a small number of samples between training and test datasets and may thus accentuate batch effects. The voting mode will run 10 instances of the genetic algorithm for 10,000 generations. At their term, it will select k-mers from the top individuals across the 10 instances and run a final genetic algorithm on this subset of k-mers for another 10,000 generations. Running multiple genetic algorithms and aggregating their results prevents overfitting on a specific split of the data between the training and test set. In addition, the voting mode introduces Gaussian noise by default into the data to further prevent overfitting. This option is recommended for experiments with <30 samples per condition.

Using the voting mode (83 h using 15 cores; see Supplementary Table 1 for parameters and Supplementary Fig. 1 for runtimes and memory usage), we found an individual that was able to classify patients with 0.93 accuracy (Fig. 4b) with only five k-mers of length 30 (Supplementary Table 3). As expected three of these k-mers mapped to genes that had clear roles in resistance to chemotherapy; JAK3 is involved in chemotherapy resistance in triple-negative breast cancer\(^1\), BOP1 reduces chemotherapy resistance\(^2\) and VTCN1 is associated with poor clinical outcomes in numerous cancers including breast cancer\(^22\).

Classifying BS-seq data. We then wanted to see if GECKO could accurately classify samples using epigenetic sequencing data, such as BS-seq generated to investigate DNA methylation. BS-seq requires extensive bioinformatics processing to discover changes

---

**Table 1** Confusion matrices of breast cancer subtype classification using the frequency of k-mers discovered by GECKO and the transcript per million values of the PAM50 gene set.

<table>
<thead>
<tr>
<th>Classification with GECKO k-mers</th>
<th>Predicted class</th>
<th>Basal</th>
<th>Her2</th>
<th>LumA</th>
<th>LumB</th>
<th>True class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Basal</td>
<td>97.7</td>
<td>2.2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Her2</td>
<td>2.2</td>
<td>67.5</td>
<td>6.2</td>
<td>4.2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>LumA</td>
<td>1.5</td>
<td>1.5</td>
<td>92.3</td>
<td>4.6</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>LumB</td>
<td>0</td>
<td>3.4</td>
<td>18.8</td>
<td>77.8</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Classification with PAM50 TPM values</th>
<th>Predicted class</th>
<th>Basal</th>
<th>Her2</th>
<th>LumA</th>
<th>LumB</th>
<th>True class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Basal</td>
<td>86</td>
<td>5.2</td>
<td>5.5</td>
<td>3.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Her2</td>
<td>15.3</td>
<td>60.6</td>
<td>3.6</td>
<td>20.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LumA</td>
<td>15.3</td>
<td>2.2</td>
<td>88.1</td>
<td>8.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LumB</td>
<td>5.9</td>
<td>15.4</td>
<td>36.5</td>
<td>42.2</td>
<td></td>
</tr>
</tbody>
</table>

---

**Fig. 3** GECKO discovers 10 30-mers that classify breast cancer subtypes. Comparison of breast cancer subtype classification using the frequency of k-mers discovered by GECKO and the transcript per million values of the PAM50 gene. Panels show the t-SNE separation of the four classes

**Fig. 4** GECKO voting mode for small sample sizes. a) GECKO’s voting mode will run 10 separate genetic algorithms with added Gaussian noise. The best solutions of these runs will be fed into a final genetic algorithm to produce a final solution. b) GECKO output showing the t-SNE separation of patients with complete response to chemotherapy from those that did not using five k-mers from the winning individual. Triangles correspond to the test dataset that was excluded from GECKO training can thus be used to estimate overfitting.
sequences were mapped using the Bismark software and the k-mers found three individuals for classification. We ran GECKO for 20,000 generations (39 hr; see Supplementary Table 1 for parameters and Supplementary Fig. 1 for runtime and memory usage) and found a winning individual that was able to classify normal from CLL samples with an accuracy of 1 using 20 k-mers (Fig. 3a; Supplementary Table 4). In addition to this final classification, GECKO plots the evolution of winning organisms across the 20,000 generations (Fig. 3b). This graph can be used to identify individual k-mers that are essential for classification and thus worth investigating. Here we found three k-mers that were most frequently used by winning individuals for classification (Supplementary Table 5).

We verified the methylation status of the loci where these k-mers were mapped using the Bismerk software and found that all three of them displayed dramatic changes in DNA methylation between normal and CLL samples (Fig. 5a). Interestingly the two k-mers that were finally selected after 20,000 generations, K107977 and K90528 overlapped binding sites for CTCF and GATA3, both of which are affected by DNA methylation status. K107977 overlaps a CTCF-binding site for the ATP6V1G1 gene, which codes for a proton pump responsible for acidification of the cell, a hallmark of cancer promotion. K90528 overlaps a GATA3-binding site for the SULF2 gene that has already been identified as a diagnostic and prognostic marker in multiple cancers.

**Discussion**

HTS data analysis often requires extensive data transformations through tailored bioinformatics pipelines to organize the sequences in a manner that is coherent with our understanding of biology. Mapping to a reference, using ad hoc statistical thresholds and grouping sequences by functional elements, such as transcripts are common steps in most bioinformatics pipelines. We designed GECKO with the aim of creating a classifier that could explore HTS data without a reference genome or transcriptome and without the need of bioinformatics pipelines dedicated to a specific library preparation or technology. The approach we describe here can in theory explore any type of sequencing data. Because GECKO considers groups of k-mers for classification, it can make use of co-dependencies between sequences to find smaller and more accurate classifiers. Thus, GECKO is capable of better classification than the commonly used approach that consists of selecting genes for which the expression is statistically significant between conditions to build a classifier (Supplementary Fig. 2). In the miRNA analysis of blood cells for example, one of the k-mers that participated in making an excellent classifier was not statistically significant by itself and would have been overlooked.

Using k-mer counts removes the requirement of a mapping step and makes GECKO applicable to numerous types of sequencing experiments. In addition, we found that using k-mers instead of other metrics, such as fragments per kilobase million (FPKM) or read counts resulted in higher predictive power even when run with the same genetic algorithm (Supplementary Fig. 3). This can be explained by the fact that k-mers can measure changes in transcription, isoform abundance, and sequence simultaneously. When applied to bisulfite converted data, each epigenetic change can potentially lead to the appearance of a novel k-mer in samples where the modification is present. These sample-specific k-mers allow GECKO to make very efficient classifications and to pinpoint the exact location of the modification.

Unlike regression analysis our approach provides multiple solutions (Supplementary Fig. 4). For research purposes this allows us to investigate why different groups of solutions work well together, explore co-dependencies between sequences and functional pathways that allow a good separation of input samples. In a clinical setting, providing multiple good solutions allows more flexibility for selecting diagnostic or prognostic targets. Importantly, the k-mers used for classification are not biased towards higher expressed genes (Supplementary Fig. 5) and mostly map to unique locations in the genome or transcriptome (Supplementary Fig. 6). Thus, GECKO can make use of unique transcriptional elements across a large spectrum of expression.

**Fig. 5** GECKO can accurately classify normal and CLL patients using k-mers from bisulfite sequencing data. a GECKO output showing the t-SNE separation of CLL and normal samples using 20 k-mers from the winning individual. b GECKO output of K-mer exploration across 20,000 generations; k-mers that are frequently found in winning organisms are displayed as horizontal lines across generations; dots represent k-mers that were selected in one generation but eliminated in the following generation often due to a decrease in fitness of the model. c IGV screenshots showing the methylation status of normal and CLL samples of regions corresponding to three most frequently used k-mers in winning organisms determined by the Bismark software.
GECKO’s ability to work across multiple types of data without the need of dedicated bioinformatics tools could make it invaluable for cross-platform large-scale analyses but also for individual researchers and clinicians who would be able to compare HTS data between cohorts of patients with no bioinformatics training. It is worth noting that the longest and computationally intensive part of our procedure is obtaining the k-mer matrix. This step need be performed only once per dataset however and providing a k-mer matrix for online datasets along with sequencing files could result in widespread use of non-biased approaches such as GECKO. In addition, k-mer-based approaches, such as GECKO have the advantage of being portable; k-mer sequences will not change with new versions of the genome.

**Methods**

**Data preparation.** The k-mer decomposition into a matrix of k-mer counts is performed using Jellyfish 2.3.1. This step can be preceded by a filtering of sequencing adaptors by Trim Galore (bioinformatics.babraham.ac.uk/projects/trim_galore/) if the user selects this option in GECKO. GECKO will then eliminate k-mers for which the count is below a noise threshold, k-mers that are uninformative for the given study and k-mers that are redundant (i.e. that share the same information as another k-mer).

The noise threshold is determined empirically from the input samples and is calculated for each separate run of GECKO. To do this, we count the number of times a k-mer count appears in one sample with null values in all other samples from the same group for the same k-mer. Starting at a k-mer count of 1, we search how many times the value 1 appears for a k-mer in one sample with 0 in every other sample for the same k-mer. We then iterate this process for k-mer counts 2, 3, etc. We define this process as determining the slope of frequency counts (determined by calculating the derivative at each point), we consider that we are above background and set the threshold as the k-mer count just before the greatest inflection of the slope (Supplementary Fig. 7).

To determine redundant k-mers, this algorithm is unsupervised and determines the existence and number of separate levels in continuous data. If there are no clear categories, the discretization will output a vector of 1’s. Following this discretization, if there is not a minimum of 10% of samples with a different level, then this k-mer is considered uninformative. By default, this minimum number is set at 10% of the size of the input condition with the least replicates. For example, if the condition with the least replicates has 30 samples, then at least three samples must have a different discretized level to the other samples.

To eliminate redundant k-mers we use symmetric uncertainty (SU) between pairs of k-mers. Instead of comparing each k-mer to all other k-mers, we first split the k-mers into buckets of equal size and perform pairwise comparisons within a bucket. To determine which k-mers will be bucketed together, we calculate the sum of the counts across samples, k-mers with a similar sum across samples are put together; k-mers within a bucket have a higher chance of being redundant than if they were randomly bucketed. When all k-mers within buckets have been compared and redundant k-mers filtered, this process of bucketing by sum and filtering is repeated. This process of bucketing the k-mers by sum to lead to 10 times faster filtering process on smaller samples and larger gains with larger matrices.

The SU between two k-mers A and B is given by the formula:

$$SU(A, B) = 2 \times \left( |H(A) + H(B) - H(A, B)| + |H(A) + H(B)| \right)$$

where H(A) and H(B) are the entropies of the two k-mers along the samples and H (A, B) is the entropy of the combined k-mers counts A and B along the samples. The Entropy is given by the formula:

$$H(A) = - \sum \frac{Mi}{N \times 2^{log2}(Mi/N)}$$

where G is the total number of k-mer frequencies given by the discretization step, Mi is the number of samples at the given discretization level N is the total number of samples. In our analysis, we empirically set the limit of SU at 0.7, above which two k-mers were considered as redundant.

GECKO keeps a record of all k-mers eliminated due to redundancy along with the ID of the k-mer that caused it to be eliminated. Thus, when the genetic algorithm finds a solution, GECKO can provide all the redundant k-mers that would have provided a similar solution.

All code for the data preparation was implemented in C++. The adaptive genetic algorithm. The algorithm begins by splitting the input data into a training and test set. The test set is created by randomly selecting a number of samples from each input category. By default the number of samples selected is 1/6th of the category with the smallest amount of samples. The test set is used to establish a final test score that will have no impact on the genetic algorithm’s evolution but allows us to estimate how well GECKO performs on a given dataset.

**Training.** At each generation of the AG, all individuals are scored based on their ability to classify the input samples using a machine learning algorithm. In this study, the algorithm used was a Linear Support Vector Classification (LibSVC). This method combines excellent results on small datasets and unbalanced groups with a good generalization potential, for a small computational resource cost. LibSVC is implemented in GECKO via the Scikit-learn package. GECKO can also be used with a random forest model or neural networks, however these have higher computational costs and require dedicated hardware to be implemented within reasonable time-frames.

To calculate the fitness score of an individual at each generation we randomly split the training set into two. 2/3 of the training set becomes the inner training set and the remaining 1/3 becomes the inner test set. We contrast the inner test set, which is used to score individuals at each generation of the adaptive genetic algorithm with the test set which is not used to train the adaptive genetic algorithm but instead is used to estimate the performance of our model. The inner split on the training data is random and is performed five times. The score of each individual is an average of these five iterations trained on the inner training sets and tested on the inner test sets. This rotation of the training data avoids sample batch effect biases at each generation.

**Natural selection:** After testing the fitness of each individual of our population we delete individuals with lower fitness scores. By default, this is 30% of the population. We call this process natural selection.

We sort the individuals by ascending rank and then apply the following probabilistic rule:

$$P = \frac{1}{N} \sum_{i=1}^{N} P - value = 1$$

where X is the individual rank and the following conditions are satisfied:

$$P - value = aX + b$$

where a, b are scalar values, N is the size of the population, and N/2 and N/2 are, respectively, the probability for the individual rank N/2 and rank N/2 to be deleted.

**Mutation and crossing over rates:** GECKO makes use of three different types of Genetic Algorithm. These adapt the mutation and cross-over probabilities depending on the homogeneity and the performances of the population in order to converge faster and more accurately.

The three algorithms are:

A simple adaptive genetic algorithm. This algorithm has a fixed factor for individuals for which the fitness is inferior to the average and a decreasing linear function for the better performing half of individuals.

Another improved adaptive genetic algorithm that, similar to the simple adaptive genetic algorithm, has a crossover probability fixed above the average fitness, but uses exponential instead of the linear function for fitness values below the average.

An improved adaptive genetic algorithm that models the probabilities with two linear functions, with a breakpoint for the individuals that have a fitness equal to the average fitness.

We recommend using the last model as it shows better exploration and higher convergence rates for the kind of data used for GECKO. This approach aims to maintain the population’s diversity while protecting good individuals from modifications. The mutation and cross-over probabilities are decreased when the individual’s fitness is high compared to the average and increased if it is low. Similarly, the probabilities are decreased when the population is heterogeneous and increased when the population is homogeneous to favor exploration of novel solutions. These probabilities are modeled by two linear functions depending on whether the individual is above the average fitness of the population or below it and is given by the formula below.

$$P_{m} = \begin{cases} \frac{k_{1}(f_{m} - f_{avg}) + k_{2}(f_{avg} - f_{min})}{k_{1} + k_{2}}, & f < f_{avg} \\ \frac{k_{1}(f_{avg} - f_{min}) + k_{2}(f_{max} - f_{avg})}{k_{1} + k_{2}}, & f \geq f_{avg} \end{cases}$$

where $f_{m}$ is the individual’s fitness, $f_{min}$ is the fitness of the worst individual, $f_{avg}$ is the population’s average fitness and $f_{max}$ is the fitness of the population’s best individual. $k_{1}$ is the rate applied when $f < f_{avg}$ $k_{2}$ when $f \geq f_{avg}$ and $k_{1}$ when $f = f_{avg}$

**Stopping criteria:** By default, GECKO will run for an input number of generations. The user may however choose to make use of a stopping criteria that will stop the algorithm prematurely. The stopping criteria is checked after at least
5000 generations of the genetic algorithm. At this moment, the number of occurrences of each k-mer in the population is calculated across bins of 500 generations from the start of the algorithm to the current generation. The top 1% of most frequent k-mers in each bin are selected. We then estimate the difference in k-mer composition between the current bin and all previous ones using a Hamming distance. This distance measures the quantity of highest scoring k-mers that are changing across generations. When the slope of Hamming distance across generations drops below 1%, the stopping criteria is triggered.

Adding Gaussian noise: The user may add Gaussian noise to the model to prevent overfitting. The characteristics of this noise are determined for each k-mer separately. They are a mean of 0 and a standard deviation equal to the standard deviation of the k-mer in the training set. The user can modify the level of noise by changing noisefactor which multiplies the standard deviation by the input value. This noise is generated at each training of machine-learning model and for each individual.

tSNE visualization: t-SNE plots are generated using scikit-learn with the default parameters but initialization with PCA. This initialization option allows for better reproducibility of t-SNE graphs. Below is the corresponding command-line: manifold.TSNE (n_components = 2, init = ‘pca’, random_state = 0, perplexity = 30.0, early_exaggeration = 12.0, learning_rate = 200.0, n_iter = 1000, n_iter_without_progress = 300).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
The data that support the findings of this study are available from NCBI Gene Expression Omnibus under the accession numbers GSE101467 and GSE58889; the Cancer Genome Atlas under the Pan-Gyn cohort name; the database of Genotypes and Phenotypes under the accession numbers phs001345.v2.p1 and phs001050.v1.p1 but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available by submitting a request to these repositories.

Code availability
GECKO is available at https://github.com/RitchieLabIGH/GECKO under the CeCILL license.

Received: 18 December 2018 Accepted: 8 May 2019
Published online: 20 June 2019

References
3. Martín-Martín, N. et al. Strati

Acknowledgements
We wish to acknowledge the Genotoul platform (genotoul.fr) for providing us with calculation time on their servers. We thank Jerome Audoux, Jean-Philippe Villemen and Giacomo Cavalli for their advice. We wish to acknowledge the Agence Nationale de la Recherche (ANR/CJC-WiRED), the Labex EpigenMed and the MUSE initiative for their financial support.
Author contributions

Additional information
Supplementary information accompanies this paper at https://doi.org/10.1038/s42003-019-0456-9.

Competing interests:
The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
Supplementary Figure 1: Memory and time usage for 100 generations of the IAGA in GECKO for different numbers of k-mers per individual. The runtimes were performed in ideal conditions with no other users on the calculation node. The runtimes in the manuscript were recorded in real conditions with other users sharing the server.
Supplementary Figure 2: comparison of classification accuracy between GECKO and classifiers based on an increasing number of differentially expressed genes

For each experiment in the manuscript, we first determine differentially expressed genes between conditions. We then use an increasing number of these genes to form a classifier (using the same SVM model as GECKO) starting with the genes that have the strongest p-values. For each number of genes used in the classifier, we ran the classifier 10 times using cross-validation and built a boxplot from these 10 replicates. These are compared to the median of 10 GECKO runs indicated by the red horizontal dashed bar.

microRNA levels were calculated using the nf-core smRNA-Seq pipeline v1.5 (github.com/nf-core/smrnaseq). Gene counts were downloaded from the TCGA website directly. We used DESeq2 to perform standard analysis as described in document “Analyzing RNA-seq data with DESeq2” that is available on Bioconductor website. For each dataset, we performed a one-versus-all analysis for each group. As described in¹ we took the genes with the best adjusted p-values as the most differentially expressed genes.

Supplementary Figure 3: comparison of classification accuracy between GECKO applied to k-mers (dark boxplots) and GECKO applied to transcript quantification values (light boxplots) from the same samples (n=10 separate runs)

GECKO was run 10 times for each experiment using either k-mers or FPKM values. microRNA levels were calculated using the nf-core smRNA-Seq pipeline v1.5 (github.com/nf-core/smmaseq). Gene counts were downloaded from the TCGA website directly. For the breast cancer classification, we also added a voting mode to demonstrate that the k-mers had not been as extensively utilized as the FPKMs and thus adding a voting step increased the classification power for k-mers more than it did for FPKMs.
iMOKA: k-mer based software to analyze large collections of sequencing data

Claudio Lorenzi, Sylvain Barriere, Jean-Philippe Villemin, Laureline Dejardin Bretones, Alban Mancheron and William Ritchie

Abstract

iMOKA (interactive multi-objective k-mer analysis) is a software that enables comprehensive analysis of sequencing data from large cohorts to generate robust classification models or explore specific genetic elements associated with disease etiology. iMOKA uses a fast and accurate feature reduction step that combines a Naïve Bayes classifier augmented by an adaptive entropy filter and a graph-based filter to rapidly reduce the search space. By using a flexible file format and distributed indexing, iMOKA can easily integrate data from multiple experiments and also reduces disk space requirements and identifies changes in transcript levels and single nucleotide variants. iMOKA is available at https://github.com/RitchieLabIGH/iMOKA and Zenodo https://doi.org/10.5281/zenodo.4008947.

Keywords: k-mer, NGS analysis, Personalized medicine, Bioinformatics software, Data reduction, Machine learning

Background

Studies of variation in gene expression have considerably advanced knowledge of disease etiology and classification [1–3]. To capitalize on genomic data generated from numerous clinical studies, recent initiatives have aggregated high-throughput sequencing (HTS) experiments from multiple cohorts that measure gene expression, RNA isoform usage, and genome variation. For example, the Genomic Data Commons program controls access to over 84,000 cases [4]. Still, despite these efforts to aggregate and provide data from multiple studies, their computational analysis and integration presents a major challenge; each type of HTS data requires specific bioinformatics pipelines that need to be implemented by a bioinformatics specialist. In addition, most of these approaches require reference genomes or transcriptomes and thus cannot inherently account for the diversity in non-reference transcripts or individual variations [5]. To alleviate the requirement of a reference, recent methodologies use k-mer representation; they directly compare the counts of nucleotide sequences of length k between samples [6]. These k-mer based approaches have been core to the field of metagenomics, where they are used to discover unique k-mers or k-mer signatures to classify organisms [7, 8]. However, when translated to mammalian genomes, k-mer
representation results in a $k$-mer count matrix with as many columns as there are samples and as many rows as there are $k$-mers, generally billions. Exploring such large matrices to find biologically relevant $k$-mers is intractable unless the analysis focuses only on a very small subset of the sequencing data [5] or by using metaheuristics that provide partial solutions [9].

Here we present iMOKA (interactive multi-objective $k$-mer analysis), a novel approach and software that allows non-specialists to make use of $k$-mers to explore large amounts of mammalian sequencing data. This approach is agnostic of the type of sequencing data used, is not biased towards annotated genetic elements, and can analyze transcript levels and single nucleotide variations in one pass. Importantly, iMOKA is interactive; it allows the user to import and merge samples from different studies and tailor their exploration of $k$-mers to specific genomic elements of interest such as splicing events, mutations, or global gene expression. We tested iMOKA on four clinical datasets: the classification of breast cancer subtypes and response to chemotherapy of breast, ovarian cancer, and diffuse large B cell lymphoma (DLBCL). We find that iMOKA found features that are more accurate than classical bioinformatics approaches, takes up less space, uses less memory, has faster runtimes, and can be run on a computer cluster or on a laptop.

Results

iMOKA design

iMOKA imports sequencing files in FASTQ, FASTA, BAM format, or SRR identifiers via its user interface. It then counts the occurrences of all sequences of given length $k$ (default 31) [9] using the KMC3 software [10] in each sample (Fig. 1). It then extracts labels from the sequencing metadata so that the user can define groups they wish to compare. Importantly, each sample is stored as a sorted vector of $k$-mer counts in a dedicated binary file using a custom prefix-suffix structure that drastically reduces the disk space requirements (“Methods” section). For each sample, a JSON file is created that contains metadata and a rescaling factor for $k$-mer count normalization that allows the user to remove or add samples without having to recalculate an entire $k$-mer matrix. It then uses our feature reduction step that combines a Bayes classifier augmented by an adaptive entropy filter to rapidly remove non-relevant $k$-mers (Fig. S1). The aim of this filter is to evaluate each $k$-mer individually by combining the accuracy of the Bayes classifier with the speed of calculating Shannon’s entropy. This evaluation is performed using a Monte Carlo cross validation with a high number of iterations and an early break (“Methods” section) that efficiently reduces overfitting and generates predictions that overcome batch effects. In order to reduce the number of features evaluated, the entropy filter works simultaneously and, learning from the entropies of the $k$-mers that successfully passed the accuracy filter, discards $k$-mers with low entropy. Following this filtering, $k$-mers for which the sequences overlap are assembled into graph structures. These are used to aggregate the $k$-mers that are likely to have been generated from the same biological sequence and are used to eliminate false positive $k$-mers that are mainly singletons (1 $k$-mer) or very short branches in the graph structure. Bifurcations or bubbles in these graphs generally arise from the existence of multiple sequence isoforms that differ by point mutations or alternative splicing events [11].
combining this graph assembly with the relatively permissive Bayesian filter, we are able to generate a list of informative \( k \)-mers in a manner that is fast and accurate.

iMOKA allows the user to align the \( k \)-mer graphs to a reference genome to annotate them with known genomic features such as known RNA transcripts, point mutations, or mRNA splicing events. iMOKA provides a random forest classifier that uses filtered \( k \)-mer graphs as features (Supplementary methods) and provides the user with a
classification model and a sorted list of \( k \)-mer graphs that were most used in the tree models and that are thus of higher interest (Fig. 1). The user may even build classification models based solely on specific genomic features such as point mutations or gene expression for example. Finally, iMOKA uses self-organizing map clustering on the \( k \)-mer graphs to enable users to identify subgroups or outliers amongst their input samples.

**Benchmarking datasets and algorithms**

iMOKA uses a \( k \)-mer based analysis to detect sequence features and create classification models from large cohorts of mammalian RNA sequencing data. To test its performance, we selected four studies that were distinct in their data structures, classification objectives, and sizes. The first was a non-binary classification of 1038 patients aiming to define 4 subtypes of breast cancer which were luminal A (LumA), luminal B (LumB), HER2-enriched (HER2), and basal-like. The second was a cohort of 240 ovarian cancer patients where the objective was to predict response to chemotherapy. The third was a smaller cohort of 118 breast cancer patients where the objective was also to predict response to chemotherapy. The last was an even smaller cohort of 17 DLBCL patients divided according to their responsiveness to the chemotherapy.

In our benchmark, we included methods based on four different types of features which were \( k \)-mer counts, percentage-spliced-in (PSI), transcripts per kilobase million (TPM), and sequencing counts. The two latter were measured and tested across annotated genes and transcripts separately. The algorithms we benchmarked were DESeq2 [12], edgeR [13], and limmaVoom [14] for TPM and sequencing counts; iMOKA for \( k \)-mer counts; and Whippet [15] for alternative splice site usage. We excluded four other \( k \)-mer based methods HAWK [16], KOVER [17], Kissplice [11], and GECKO [9] because they were respectively impossible to run on such big datasets due to segmentation fault errors, were unable to find \( k \)-mers that could classify the input samples or, for the last two methods, were killed after 2 weeks of runtime on our computer cluster.

In our benchmark, we compared the list of features output by each algorithm by using them in a random forest classifier and determining their out of bag scores (OOB score). The out of bag score tests how well each classifier performs without having to set aside a portion of the data specifically as a test set. It is as reliable as using a test set [18, 19] without having to set aside part of the data. We chose the random forest classifier because it is a non-parametric approach and because the importance of each input feature is easy to evaluate.

Finally, for the largest dataset, the molecular classification of breast cancer, we performed a 5-fold cross validation of the entire iMOKA procedure and all other benchmarked algorithms, using 4/5 of the dataset for data reduction and creation of a random forest model and 1/5 of the dataset as the test set.

**Classification of breast cancer subtypes**

Breast cancer is a transcriptionally heterogeneous disease with multiple subtypes that determine prognosis, treatment, and patient outcome. Although breast cancer classification is constantly being updated, a broadly accepted stratification defines four groups
which are luminal A (LumA), luminal B (LumB), HER2-enriched (HER2), and basal-like [20]. We benchmarked iMOKA on a dataset of 1038 mRNA-Seq breast cancer samples from the Cancer Genome Atlas (TCGA) Pan-Gyn cohort [21] (patients per class: basal 190, Her2 82, LumA 559, LumB 207) and tested how well the outputs of each approach could accurately predict the four classes. We found that the list of k-mers output by iMOKA (Additional file 1, Fig. S5) was above all other methods in their ability to classify the four types of breast cancer (Fig. 2a). The worst performing features were the splice site usage statistics given by Whippet. This could be expected because the breast cancer stratifications were originally created using gene expression profiles, not splicing events.

Fig. 2 iMOKA accurately predicts breast cancer subtypes. a The features output by all benchmarked approaches are evaluated for their capacity to classify breast cancer subtypes using Random forest’s oob score plotted as a function of the number of the best features output by each approach. b Screenshot of the iMOKA output with each k-mer sequence, their rank in the classification of breast cancer subtypes, and where these sequences map to on the genome. c Screenshot of the iMOKA display showing k-mer counts of the 3 highest ranking k-mers across the 4 subtypes. d Gene ontology of the genes overlapping the k-mers selected by iMOKA.
We additionally performed a 5-fold cross validation of the entire iMOKA procedure and all other benchmarked algorithms including feature reduction and model generation. The accuracies of the final models (Fig. S2) show a consistent behavior to the oob scores in Fig. 2a.

iMOKA identified 3002 k-mers overlapping different types of events (Table S1 and Additional file 1). Using iMOKA’s interface, we were able to explore the genes to which these k-mers mapped (Fig. 2b). As expected, within the best ranking k-mers, iMOKA found overlaps with genes that have been extensively linked to breast cancer subtypes and are already used in the clinic such as estrogen receptor 1 (ESR1) [22], Forkhead Box A1 (FOXA1) [23], Forkhead Box C1 (FOXC1) [24], xenopus kinesin-like protein 2 (TPX2) [25], and Melanophilin (MLPH) [26]. By clicking on the k-mer sequence in the iMOKA interface, we can visualize the representation of each k-mer in the 4 classes (Fig. 2c). The top three k-mers, whose gene expression is shown in Fig. S3, have representation profiles that clearly explain iMOKA’s high classification accuracy with a small number of k-mers.

It is worth noting that iMOKA picked up 120 potential alternative splicing events. Amongst these were 4 extensively studied splicing isoforms (MYO6, TPD52, IQCG, and ACOX2) [27] identified to be amongst the 5 most important isoforms differentially expressed between ER+HER2− and ER-HER2 primary breast tumors (Fig. S4).

Finally, we used DAVID [28] to perform a functional annotation of the genes overlapping the k-mer selected by iMOKA. The gene list is strongly enriched for breast cancer-associated genes and of genes associated with the function commonly dysregulated in cancer cells, such as cell cycle, cell division, and motility (Fig. 2d and Additional file 4).

**iMOKA identifies events associated with the response to treatment in ovarian cancer patients**

Our second benchmark was performed on a dataset of high-grade serous ovarian cancers taken from the TCGA_OV cohort [29]. We included patients having an annotated [30] response to a first-line treatment to the combination platinum and taxane chemotherapy (patients per class: 174 responsive, 66 non-responsive). iMOKA identified 138 k-mers with individual accuracy between 65 and 75% (Table S1 and Additional file 2). Again, the k-mers found by iMOKA gave the most accurate oob scores for response to chemotherapy (Fig. 3a). The gain compared to other methods is much higher than for the previous breast cancer classification. This can be explained by the fact that most of the methods we benchmark against only make use of gene or transcript expression or splicing sites. Breast cancer stratification is mainly based on gene expression, and therefore, these methods compare well with iMOKA. However, in the case of response to chemotherapy in ovarian cancer, iMOKA is able to also make use of single nucleotide variants (SNVs) and splice site usage to make its predictions (Fig. 3b). Via the iMOKA interface, we can visualize the SNVs with the highest feature importance. Thus, we can observe that iMOKA detected a known nonsense mutation (SNP id: rs10794537) in the alpha-L-iduronidase (IDUA) gene. IDUA is responsible for the degradation of the mucopolysaccharides, heparan sulfate, and dermatan sulfate.
which modulate angiogenesis, cell invasion, metastasis, and inflammation [26] and importantly are ligand receptors for polynuclear platinum anticancer agents [27]. In agreement with this, the gene ontology (Fig. 3d) analysis shows a functional enrichment of small molecule binding proteins.

**iMOKA identifies events associated with the response to neoadjuvant chemotherapy in breast cancer patients**

The third test dataset was taken from the Breast Cancer Genome Guided Therapy (BEAUTY) study [31] and consisted of patients with all 4 types of breast cancer for which we tested the response to neoadjuvant chemotherapy with paclitaxel and anthracycline. This allowed us to test the binary classification of more heterogeneous cell populations on smaller sample sizes: 36 patients that had a complete response to chemotherapy and 82 that did not. It is worth noting that this dataset presented a
significant batch effect, detected using the R package DASC [32], associated with the load date of the samples (Fig. S5). Despite this, iMOKA identified 1248 k-mers with an individual accuracy between 70 and 83.8% (Table S1 and Additional file 3). Again, the k-mers discovered by iMOKA give the highest oob scores for the response to chemotherapy (Fig. 4a).

Our method can identify multiple events on the same gene that are useful for classification. For example, as shown in Fig. 4b for the highest scored k-mers overlapping the gene TBC1D9, iMOKA discovers that the gene as a whole is differentially expressed between conditions but also discovers alternatively expressed introns (Fig. 4c) that were confirmed as being a retained intron using a dedicated algorithm, IRFinder [33].
The gene ontology analysis of the genes overlapping the $k$-mers selected by iMOKA reveals a strong relationship with microtubules and cilia, components influenced by paclitaxel [34, 35], an anti-microtubule agent of the taxane family used as part of the therapy on all the patients in the study. Although the study included heterogeneous cancer types and an unbalanced dataset, iMOKA was able to detect features useful for classification.

**iMOKA identify DE genes associated with DLBCL chemoresistance**

In the last dataset, we tested iMOKA in a frequent scenario where differential representation of transcripts is assessed in a very small cohort. To this end, we considered 17 DLBCL patients [36], 10 responsive to an anthracycline-based regimen R-CHOP (rituximab, cyclophosphamide, doxorubicin, vincristine, and prednisone) and 7 non-responsive. The RNA-seq used for this dataset is targeted, making it impossible to evaluate the PSI values, so only the abundance of the genes and transcripts were considered in the benchmark (Fig. 5 and Fig. S7). iMOKA identified 1928 $k$-mers having an individual accuracy over 80% and five with 100% accuracy. They corresponded to the genes AKT1, BTBD9, ZBTB45, ZBTB17, and BHLHE40. Amongst those, AKT1 is known to play a role in DLBCL chemosensitivity [37] but was not detected as differentially expressed in the original publication [36].

This study highlights another advantage of using $k$-mers; they are agnostic to transcript annotation. For example, the $k$-mer overlapping ZBTB17, a gene involved in B cell development and differentiation [38], is located on the splicing site at position chr1:15,947,123-15,948,295 and is part of Refseq transcript NM_001242884. However, this transcript was not annotated in the GENCODE annotation (Fig. 5b) and thus not detected by salmon.

**iMOKA runtimes and disk space**

iMOKA was designed to be scalable; the user can control the number of threads used and the dedicated RAM, allowing the software to run not only on HPC clusters, but also on a laptop. In Fig. 6, we report the times to analyze three experiments described in the previous sections on a computer with 8-cores and 32 GiB of RAM. Importantly, the higher the number of samples in the cohort, the bigger iMOKA’s gains are.

iMOKA’s most intensive task is the generation of informative $k$-mers, where a large amount of data is filtered and aggregated, while the other benchmarked approaches handle data that are already filtered (reads are already mapped to annotated regions). Finally, most methods that calculate differential expression are designed for relatively small cohorts and do not scale well in memory with large cohorts: DESeq2 and edgeR for example required additional RAM in order to analyze the differential expressed transcripts in the TCGA BRCA (TCGA_BC) analysis (61 GiB and 46 GiB, respectively) (Fig. 6).

**Discussion**

Recent efforts to aggregate and annotate patient HTS data should facilitate our understanding of health trajectories through multiple molecular mechanisms. In theory,
combining gene expression, isoform usage and single nucleotide variation should allow for more nuanced stratification and prediction of disease etiology. However, HTS data analysis often requires extensive data transformations that are often performed with little transverse coherence; each type of analysis produces lists of features that pass a

Fig. 5 IMOKA identifies DE genes and transcripts between chemorefractory and chemosensitive DLBCL patients. 

a The features output by all benchmarked approaches are evaluated for their capacity to classify response to chemotherapy. RF’s oob scores are plotted as a function of the number of the best features output by each approach. IMOKA reach the highest score thanks to the five k-mers with 100% of individual accuracy. b IMOKA GUI screenshot showing the detail of the splicing site of the gene ZBTB17, where the isoform NM_001242884 is detected as an event of interest by IMOKA, present in the RefSeq but not in GENCODE. c Five k-mers able to separate the responsive patients from the chemorefractory ones. k-mer normalized counts and the respective gene counts available in Fig. S6

Fig. 6 IMOKA is faster and scales better with large cohorts. Comparison of the running times between the benchmarked methods. Solid bars represent the time dedicated to the generation of the features (transcript abundance, PSI evaluation, and k-mer count), and the lighter bars represent the time dedicated to the analysis of the features (differential expression, differential splicing, and the machine learning-based filters)
given test and these are then analyzed separately. Mapping to a reference, using ad hoc statistical thresholds for each type of analysis, and grouping sequences by functional elements are common steps in bioinformatics pipelines that may not reflect the complex interaction between each of the processes that make up an individual’s transcriptome.

We designed iMOKA with the aim of analyzing HTS data in the reverse manner; we wished to first discover all sequences that were informative, group them according to how well they could classify the input samples, and then break them down into the different components of gene expression, isoform representation, and SNV presence. In doing so, we created a classifier that could explore HTS data without a reference genome or transcriptome and without the need of dedicated bioinformatics pipelines for each type of transcriptional event.

Using k-mer counts removes the requirement of a mapping step and allows iMOKA to explore and combine multiple transcriptional events to make more accurate predictions and to explore all these events simultaneously without having to apply multiple pipelines. k-mers can measure changes in transcription, isoform abundance, and sequence simultaneously and were thus able to create better predictive models than other metrics such as transcripts per million (TPM), read counts, or splice site usage.

By creating a reliable, cross-platform user interface, iMOKA allows non-specialists to leverage the predictive power of our approach in a manner that is fast and accurate. In addition, iMOKA uses a flexible data structure that allows the easy integration of new samples and uses only a fraction of the disk space required for stocking compressed sequencing files. In addition, k-mer based approaches such as iMOKA have the advantage of being portable; k-mer sequences will not change with new versions of the genome. This is crucial for the integration of omics data with other clinical data such as imaging or patient file records.

**Methods**

**Preprocessing**

The input data can be given as SRR identifier, BAM, FASTA, or FASTQ files. In the first and second cases, the corresponding FASTQ files are automatically generated using sra-tools’ fastq-dump [39] and SAMtools [40], respectively. If the data is stranded paired end sequencing, the user can reverse complement one or both the files using SeqKit [41]. In order to assert the quality of the FASTQ files, the user can use FASTQC [42] by adding the flag “-q”.

For each sample, KMC3 [9] is used to count the k-mers of the length chosen by the user (default $k = 31$). Its output is converted into a sorted binary file optimized for the following steps of iMOKA and a JSON file containing the metadata information.

The binary file is divided into two parts: a suffix portion, containing the nucleotidic sequence and the relative count, and a prefix portion, which contains the prefixes and the positions of the respective suffixes.

The length of the prefix is defined using the following formula, an adaptation from [43]:

\[
p = 0.5 \times \log_2(t) - 0.5 \times \log_2(\log_2(t))
\]
where $p$ is the prefix size and $t$ is the total number of different $k$-mers for the current sample.

**Matrix generation**

The input to the feature reduction step is a JSON file containing the name, group, and localization of the sorted binary $k$-mer count file of each sample in the analysis. The JSON file also stores the sum of all the $k$-mer counts that will be used as a normalization factor:

$$N_{ij} = \frac{C_{ij} \times RF}{T_j}$$

where

- $N_{ij}$ is the normalized count of the $i$th $k$-mer of the sample $j$
- $C_{ij}$ is the raw count of the $i$th $k$-mer of the sample $j$
- $T_j$ is the sum of the counts of all the $k$-mers of the sample $j$
- $RF$ is a rescaling factor, used to increase the value of all the normalized values and avoid computational problems related to precision. By default, $RF = 1e9$

Each thread starts the creation of the matrix and the reduction step in parallel, using an OpenMP implementation, at a different point of the matrix according to the number of threads available using the following formula:

$$K_t = \frac{4^k - 1}{T} \times t$$

where

- $T$ is the total number of threads available
- $K_t$ is the first $k$-mer analyzed by the thread $t$ (from 0 to $T$ excluded) considering all the possible ordered combination from 0 to $4^k$
- $k$ is the length of the $k$-mers (default 31)

The last $k$-mer analyzed by each thread is $K_{t+1} - 1$. For example, with 2 threads ($T = 2$) and $k = 31$, the first $k$-mers for each threads will be:

- $K_0 = \frac{4^{31} - 1}{2} \times 0 = 0 = AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA$  
- $K_1 = \frac{4^{31} - 1}{2} \times 1 = 2305843009213693952 \times 1 = GAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA$

Finally, the buffer size reserved for each sample is dependent on the number of parallel processes, the number of total samples, and the available memory reserved:

$$buff = \frac{\text{RAM}_{\text{avail}}}{a \times N \times T}$$

where

- $buff$ is the length of the buffer
- $\text{RAM}_{\text{avail}}$ is the available RAM in GiB, defined by the user using the environmental variable "IMOKA_MAX_MEM_GB"
- $N$ is the number of samples in the matrix
$T$ is the total number of threads available  
$\alpha$ is a factor representing the GiB occupied by 1000 $k$-mers, approximated to 0.011

**Bayesian classifier $k$-mer accuracy assessment**

The accuracy of each $k$-mer is calculated using the NaiveBayesClassifier method implemented in the library mlpack [45]. For each $k$-mer, the samples are randomly divided into test and training sets, with an equal number of samples for each group scaled to the smallest one:

\[
\begin{align*}
    n_{\text{test}} &= \text{round}(n_{\text{min}} \times p_{\text{test}}) \\
    n_{\text{train}} &= n_{\text{min}} - n_{\text{test}}
\end{align*}
\]

where:

- $n_{\text{min}}$ is the dimension of the smallest group
- $n_{\text{test}}$ and $n_{\text{train}}$ are respectively the dimension of the test and training sets
- $p_{\text{test}}$ is the test fraction, 0.25 by default

Using one feature ($k$-mer count) $x_k$ at a time, the NaiveBayesClassifier class computes for each label $y_j$:

\[
\begin{align*}
    P(X = x_k \land Y = y_j) \\
    P(Y = y_j)
\end{align*}
\]

Given that we use a pairwise comparison with a constant number of training samples amongst the labels, all the $N_{\text{labels}}$ have the same probability

\[
P(Y = y_j) = P\left(Y = y_{j+1}\right) = \frac{1}{N_{\text{labels}}}
\]

The label prediction of a sample $i$ based on the $k$-mer count $x_k$ is then given by:

\[
y_i = \text{argmax}(P(Y = y))
\]

The accuracy of the $k$-mer $k$ is computed considering only the samples part of the test set:

\[
\text{acc}_k = \frac{T}{n_{\text{test}}} \times 100
\]

where

- $\text{acc}_k$ is the accuracy of the $k$-mer $k$
- $T$ is the number of correct labels assigned in the test set

Because the accuracies depend on the random division of the training and test sets, we use a Monte Carlo cross validation [46] with a given number of iterations (-c argument, default 100). This cross validation can be ended by a conditional break that is triggered when the standard error across iterations drops beneath a given threshold (-s argument, default 0.5).

The $k$-mers that achieve an accuracy higher than the accuracy threshold (-a argument, default 65) in at least one of the pairwise comparisons are saved in a text file, along with the accuracy values.
Entropy filter booster

In order to speed up the process of accuracy estimation, we introduced an additional filter based on the Shannon entropy \[47\] of the counts of each \(k\)-mer that runs in parallel to the Bayesian filter (BF).

For a given \(k\)-mer \(k\) and its counts in the different samples \(C_k = (c_{k0}, c_{k1}, ... c_{kn})\), we compute its entropy value \(H_k\) as follows:

\[
H_k = -\sum_{i=0}^{n} f_{ki} \times \log_2(f_{ki})
\]

\[
f_{ki} = \frac{c_{ki}}{\sum_{j=0}^{n} c_{kj}}
\]

The filter uses an adaptive threshold, \(H_{\text{thr}}\), tuned according to the lowest entropy detected in the previous batch of \(k\)-mers that passed the accuracy filter \(H_{\text{min}}\).

Initially \(H_{\text{thr}} = 0\), so all the \(k\)-mers in the first batch are evaluated by the BF and the lowest entropy is saved as \(H_{\text{min}}\). During the analysis, \(H_{\text{thr}}\) is updated when more than \(E_{\text{up}}\) (initially equal to 30) passes the BF. The first assignment is always:

\[
H_{\text{thr}} = H_{\text{min}} - (H_{\text{min}} \times a_1 \times 2)
\]

Subsequently:

\[\text{IF}(H_{\text{thr}} > H_{\text{min}} - (H_{\text{min}} \times a_1)) :\]

\[
H_{\text{thr}} = H_{\text{min}} - (H_{\text{min}} \times a_1)
\]

\[\text{ELSE} :\]

\[
H_{\text{thr}} = H_{\text{min}} + (H_{\text{min}} \times a_2)
\]

The adjustment parameters \(a_1 > a_2\) ensure that the new threshold is not set too close to the minimum \(H_{\text{min}}\).

The number of \(k\)-mers required to update the threshold \(E_{\text{up}}\) increases by 30 at each update in order to reduce the number of computations and reduce the fluctuations of the threshold. Figure S1 shows the entropy in function of the BF estimated accuracy of a sample of \(k\)-mers from the previously defined datasets showing that the number of \(k\)-mer would have been rejected by the entropy filter but would have had an accuracy higher than 60% are rare and that the adaptive threshold is able to find a mild cutoff that can save more than 50% of the computation, like in TCGA BC, or can let the BF evaluate most of the \(k\)-mers in case of difficult datasets, like in BEAUTY.

\(k\)-mer graph generation

The \(k\)-mers that successfully passed the reduction are used as nodes in a graph. A link between two nodes is created if they overlap by a minimum number of nucleotides defined by parameter \(w\) (default = 1). This parameter can be increased if the user notices multiple small sequences in the final result, caused usually by \(k\)-mers with accuracy close to the given threshold arguments -T and -t, respectively the minimum accuracy required to consider a \(k\)-mer in the graph construction and the minimum accuracy required to generate a sequence from a graph.
iMOKA then prunes short bifurcations in the graph where there is only one node following the bifurcation. If there are multiple sequential bifurcations, then the branch with the lowest accuracy is removed.

The accuracy values are then rescaled from 0 to 100 for each pairwise comparison in order to normalize the accuracy values and favor the features that are able to classify pairs of classes that are more difficult to separate.

Since each bifurcation could correspond to a biological event such as a point mutation or splicing isoform, each separate path that results from a bifurcation will be kept as a separate sequence for downstream analysis using a depth-first graph traversal approach. When the traversal meets a bifurcation, the branch having the most similar accuracies values to the bifurcating node is kept in the current sequence and others will generate new sequences. Furthermore, to maintain the context of the bifurcations, three $k$-mers preceding the bifurcation are added to each of those new sequences.

**Graph mapping and annotation**

The sequences generated from the graphs can be aligned to a reference genome. Currently, iMOKA supports any aligner that provides an output in SAM or pslx format and uses the information given in the JSON configuration file “mapper-config” (-m argument) to align and to retrieve the annotation file, in GTF format. In this manuscript, we used gmap v. 2019-05-12 with the human genome GRCh38 and the GENCODE annotation v29, excluding from the file the entries with the transcript type "retained_intron".

Once the $k$-mer graphs are aligned, iMOKA identifies the following “alignment derived features” (ADF):

- Mutations, insertions, deletions, and clipping are identified by the letters "M", “I”, “D” and “S,” respectively, in the alignment’s CIGAR string.
- Alternative splice sites are identified when a $k$-mer graph is split across exons.
- Differential expression (DE) is identified if 50% (set by parameter d) of an annotated transcript is covered by the $k$-mer graphs. Since regions with sequence variations not associated with the classes generate holes in the graphs reducing the portion of the transcripts that generate useful $k$-mers, a higher threshold might result in classifying DE event as general "gene" event, that is, the best $k$-mer in a gene.
- Alternative intronic events are identified if 50% (set by parameter d) of an annotated intron is covered by the $k$-mer graphs.
- Intergenic events are identified if the $k$-mer graph maps to the genome but not to any annotated transcript.
- Unmapped or multimapped events are created for those $k$-mer graphs that have no mapping or map to multiple sites.

iMOKA will preserve one $k$-mer per event, the one with the highest accuracy score. Table S2 contains the list of events with a detailed description.
iMOKA implementation
The feature reduction component of iMOKA is implemented in C++ using the following libraries: MLpack [45], armadillo [48], cephes [49], cxxopts [50], and nlohmann/json [51]. The self-organizing map and the random forest are implemented in python 3 using the following libraries: numpy [52], pandas [53], sklearn [54], and SimpSOM [55]. The whole software is included in a ready-to-use Docker and Singularity [56] image and is released under the Open Source CeCILL license.

Benchmark
Transcript abundance was computed using Salmon [57] version 1.1.0 using the index built on the reference transcriptome GENCODE v29 (hg38). The PSI values were computed using Whippet [15] version v0.10.4. We processed the samples in parallel in 4 processes allowing 2 threads and a maximum of 8 GiB of RAM each. The differential expression analysis was performed between each pair of classes in R v3.6.3 using the parameters and functions described in a recent benchmark [58] for the methods DESeq2 [12], edgeR [13], and limmaVoom [14]. Significantly different PSI values between two subsets were detected using whippet-delta.jl, included in the Whippet package.

Random Forest classifier feature selection and oob score comparison
In order to compare the same number of features extracted by each pipeline, we used the sklearn method SelectFromModel to select 20 features using a decision tree classifier (DTC) trained with all the samples and all the features in order to identify twenty features that, in combination, can be good classifiers. Using an increasing number of features, from 2 to 20, we trained multiple RandomForestClassifier to retrieve the out of the box scores.
We also performed a 5-fold cross validation of the largest and better characterized dataset, TCGA BRCA, to evaluate the accuracy of a model on unseen data. For each fold, we performed the feature reduction using only the training in each method. The final list of features is reduced similarly as for the oob score determination and the balanced accuracy score is estimated for the test set.

Supplementary information
Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-020-02165-2.

Acknowledgements
We wish to acknowledge the Genotoul platform (genotoul.fr) for providing us with calculation time on their servers. The results published here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Review history
The review history is available as Additional file 8.
Peer review information
Yixin Yao was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors’ contributions
C.L., W.R., and A.M. designed the algorithm; C.L. coded the software; S.B. designed and coded the SOM; C.L., W.R., A.M., and J.P.V. designed the experiments; L.D.B. contributed to the binary data structure optimization during her internship; W.R. and C.L. wrote the article. The authors read and approved the final manuscript.

Funding
We wish to acknowledge the Agence Nationale de la Recherche (ANRJCJC - WIRED), the Labex EpiGenMed, and the MUSE initiative for their financial support.

Availability of data and materials
The data used in this manuscript are available from the Cancer Genome Atlas under the project ID TCGA-BRCA [21] and TCGA-OV [29] with dbGaP study accession identifier phs000178.v11.p1 [59]; the BEAUTY dataset [31] is available under the dbGaP study accession identifier phs000178.v11.p1 [59]; the DLBCL targeted RNA-seq data [36] are publicly available in the EMBL-EBI ArrayExpress with the accession number E-MTAB-6597 [60]. iMOKA is available at https://github.com/RitchieLabIGH/iMOKA [61] under the Open Source CeCILL license. The copy of the scripts used for the benchmark is available under the subfolder https://github.com/RitchieLabIGH/iMOKA/paper_codes.

The DOI for the source version used in this article is https://doi.org/10.5281/zenodo.4008947 [62].

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors have no competing interests to declare.

Author details
1 IGH, Centre National de la Recherche Scientifique, University of Montpellier, Montpellier, France. 2 LIRMM, Université de Montpellier, CNRS, Montpellier, France.

Received: 6 May 2020 Accepted: 10 September 2020
Published online: 13 October 2020

References

**Publisher’s Note**
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
iMOKA is a $k$-mer based software to analyze large collections of sequencing data: Supplementary Material.

The data files for each of the 4 benchmark experiments described in the paper are available as Supplementary data in .json format and can be explored with the iMOKA software.

**Supplementary Methods**

**$k$-mer list and genome browser visualization**

A graphical user interface (GUI) allows the user to visualize the information for each $k$-mer, and where the $k$-mer graphs map using the javascript implementation of IGV\textsuperscript{1}. All the graphs present in the software are generated using the javascript implementation of Plotly\textsuperscript{2}.

The GUI is implemented in Electron\textsuperscript{3} and in Angular\textsuperscript{4}, making it available on multiple platforms (Linux, Windows and MacOS). iMOKA can be run from the interface, in local or on a SLURM\textsuperscript{5} cluster, with the only dependency of Singularity\textsuperscript{6}.

**$k$-mer clustering for samples visualization and clustering**

To visualize the differences and similarities in $k$-mer expression between the samples, iMOKA uses a self-organizing map (SOM) to cluster the filtered features.

In the SOM space the features are grouped in nodes by similarity of expression across the samples and similar node are closer in the map space, thanks to the somatotopic capacities of SOM network. We then project a sum of expression for each sample, that is used to visualize the behaviour of the given categories, extracts outliers, or subgroups.
In practice, networks of different size (-n argument) are trained for 1000 iterations (-i argument) to group k-mers with similar count across the samples. The projections of the k-mer counts for each sample is used as new set of reduced features, whose importances are evaluated using an extra tree classifier with 2500 trees and their ability to classify is evaluated by cross validation using a linear support vector machine model (-ct argument, the user can choose among 8 different Machine Learning models). Finally, the software uses the aggregated features in a SOM to perform an unsupervised clusterization of the samples (-cs argument to indicate one or more cluster size). The colour coded representation of the SOM projections of each sample and of the averages for each given group can be displayed in the iMOKA interface, thanks to the JSON output files, and also in a standalone HTML pages, making this module an independent tool that can be used on any type of feature.

**Random Forest Classifier model generator**

iMOKA uses a random forest classifier (RFC) to assign importance to each feature. It also uses a RFC to produce prediction models. Feature importance is estimated with a random forest with 1000 estimators (-n argument) and with min_samples_split (minimum number of samples required to split a node in a tree) of 0.05 (5% of the total number of samples). In order to identify a subset of synergic features, a decision tree classifier (DTC) is trained with all the samples and all the features. The 10 features (-m argument) with the highest feature importance in the DTC are used to produce the final RFC model. If the DTC has less than 10 features, other DTC are created with different seeds. The RFC parameters are chosen using a cross validated grid search on the following values:

- n_estimators: 10, 100 or 500
- min_samples_split: 0.05, 0.10 or 0.15

These parameters can be modified in the random_forest.py script.
All the metrics are evaluated with Monte Carlo cross validation in a similar procedure as described in the reduction step.
Supplementary Figures

A

B

C

D

E

F

G

H
**Supplementary Figure 1:** \(k\)-mer count entropy in function of the accuracy estimated by the Bayesian classifier. Both the entropy and the accuracy were computed for samples of 50000, 1000000, 200000 and 1000000 \(k\)-mers respectively in TCGA BC (A-B), TCGA OV (C-D), BEAUTY (E-F) and DLBCL (G-H) datasets. This graph feature is available in iMOKA_core reduce step using the “-v” argument. **TP:** \(k\)-mers that passed both the entropy filter and the accuracy filter. **TN:** \(k\)-mers discarded by both filters. **FP:** \(k\)-mers that passed the entropy filter but are discarded by the accuracy filter. **FN:** \(k\)-mers discarded by the entropy filter but passed the accuracy filter. The vertical black bars correspond to the values of the adaptive threshold. In DLBCL the first, second and third quartile of the thresholds are represented for clarity. The accuracy thresholds were set to 65 for every dataset except TCGA_OV where it’s of 60 due to the scarcity of positive \(k\)-mers.
**Supplementary Figure 2:** 5 fold cross validation of the entire iMOKA pipeline on the TCGA BRCA dataset. For each fold, a test set is put aside at the start, before the feature reduction step.
**Supplementary Figure 3:** Read counts of the genes FOXC1 (A), TPX2 (B) and ESR1 (C) in TCGA BRCA, whose corresponding overlapping $k$-mers abundances are visible in Figure 2.
Supplementary Figure 4: IGV genome browser visualization (integrated in the iMOKA interface) of the $k$-mers that map to splicing sites and exons known to be involved in different breast cancer
molecular subtypes and the abundances of the representative \(k\)-mers, validated in a recent study\(^7\). A) The inclusion of the exon chr6:75898373-75898410 (hg38) in the gene MYO6. B) The first exon in position chr8:80080315-80080830, that is included in 5 out of 12 possible transcripts of the gene TPDS2 (GENCODE V.24). C) The last four exons forming a transcript with an intronic start site in the gene IQCG. D) The last six exons forming a transcript with an intronic start site in the gene ACOX2.
Supplementary Figure 5: identification of a hidden batch effect in BEAUTY dataset associated with the Load Date using DASC \(^8\).
A. ZBTB45 padj: 0.16

B. Normalized count for DNA sequence GCACACGCGCCCACTGCTGCCGCTTCCTCCC

C. AKT1 padj: 0.18

D. Normalized count for DNA sequence CATCAAGCACTGCGCGCCGCTTACCTCCTC

E. BTBD9 padj: 0.053

F. Normalized count for DNA sequence AGGGCGGAAGGGCGCTGTGTAATCATCTTT
Supplementary Figure 6: DESeq2 normalized gene counts and the respective k-mer normalized abundances identified by iMOKAfor the genes ZBTB45 (A-B), AKT1 (C-D), BTBD9 (E-F), ZBTB17 (G-H) and BHLHE40 (I-J). The genes are not detected as differentially expressed using DESeq2 (adjusted p-value on the top of each boxplot).
Supplementary Figure 7: Biological process gene ontology of the genes overlapped by the $k$-mers found by iMOKA in DLBCL.
## Supplementary tables

<table>
<thead>
<tr>
<th>Study</th>
<th>Event</th>
<th>Number of k-mers</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCGA_BC</td>
<td>Gene</td>
<td>742</td>
</tr>
<tr>
<td></td>
<td>DE</td>
<td>580</td>
</tr>
<tr>
<td></td>
<td>unmapped</td>
<td>523</td>
</tr>
<tr>
<td></td>
<td>spliceBorders</td>
<td>479</td>
</tr>
<tr>
<td></td>
<td>intergenic</td>
<td>427</td>
</tr>
<tr>
<td></td>
<td>intron</td>
<td>162</td>
</tr>
<tr>
<td></td>
<td>multiplesplicejunctions</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td>mutationBorders</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td>insertionBorders</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>deletionBorders</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>misalign</td>
<td>1</td>
</tr>
<tr>
<td>TCGA_OV</td>
<td>gene</td>
<td>105</td>
</tr>
<tr>
<td></td>
<td>unmapped</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>mutationBorders</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>intergenic</td>
<td>8</td>
</tr>
<tr>
<td>BEAUTY</td>
<td>gene</td>
<td>659</td>
</tr>
<tr>
<td></td>
<td>DE</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>unmapped</td>
<td>138</td>
</tr>
<tr>
<td></td>
<td>spliceBorders</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>intergenic</td>
<td>377</td>
</tr>
<tr>
<td></td>
<td>Intron</td>
<td>25</td>
</tr>
<tr>
<td>Event name</td>
<td>Condition</td>
<td>Description</td>
</tr>
<tr>
<td>------------------</td>
<td>---------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>insertionBorders</td>
<td>k-mer overlaps an insertion/deletion/mutation and has a higher accuracy score than other k-mers within the same kmer graph.</td>
<td>This event overlaps a known variation event.</td>
</tr>
<tr>
<td>deletionBorders</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mutationBorders</td>
<td></td>
<td></td>
</tr>
<tr>
<td>spliceBorders</td>
<td>A splicing site where the overlapping k-mers have a higher accuracy than the others.</td>
<td>This event might overlap with a splicing site involved with alternative splicing</td>
</tr>
</tbody>
</table>

**Supplementary Table 1:** Number of k-mers found in each study and the genetic events (one k-mer can have multiple events) they are associated to. The results are not additionally filtered, an interactive view of the k-mers is available using the supplementary data and iMOKA GUI. A detailed description of the events can be found in Supplementary Table 2.
<table>
<thead>
<tr>
<th><strong>DE</strong></th>
<th>An annotated transcript is covered by filtered $k$-mers across more than 50% of its length (by default).</th>
<th>Generally a differentially expressed transcript.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intron</strong></td>
<td>An annotated intron is covered by filtered $k$-mers across more than 50% of its length (by default).</td>
<td>This event might overlap with a retained intron, an alternative starting site or an intronic transcript.</td>
</tr>
<tr>
<td><strong>multiple_splice_junction</strong></td>
<td>This event is called when there are two splicing sites with one common acceptor or donor site.</td>
<td>This splicing site can indicate two differentially expressed transcripts or a different transcript usage.</td>
</tr>
<tr>
<td><strong>Intergenic</strong></td>
<td>This event represents a sequence that mapped correctly but doesn’t overlap with any feature of the given annotation file.</td>
<td>These events might represent novel or unannotated transcripts.</td>
</tr>
<tr>
<td><strong>gene</strong></td>
<td>This event is generic and represents $k$-mers that map on a gene without any other event associated with it.</td>
<td>This event is generally associated with very small transcripts or overlapping genes.</td>
</tr>
<tr>
<td><strong>unmapped</strong></td>
<td>Sequences that didn’t map on the given reference genome.</td>
<td>These can be due to contamination, repetitive elements excluded by the aligner or chimeric transcripts.</td>
</tr>
<tr>
<td><strong>multimap</strong></td>
<td>Sequences that mapped to multiple positions of the reference genome with the same score</td>
<td>These can be repetitive elements, portions of pseudogenes or similar cases.</td>
</tr>
<tr>
<td>misalign</td>
<td>The sequence generated by the graph was mapped correctly, but the best $k$-mer is completely in the clipped region.</td>
<td>Those events should be rare and require a manual investigation.</td>
</tr>
</tbody>
</table>

**Supplementary Table 2**: Description of the events in which iMOKA categorize the group of $k$-mers during the aggregation step.

**Supplementary references:**
