



HAL
open science

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:

Jean-Philippe Villemin. Approche à large échelle visant à détecter de nouveaux régulateurs de l'épissage alternatif au cours de la transition épithélio-mésenchymateuse.. Médecine humaine et pathologie. Université Montpellier, 2020. Français. NNT : 2020MONTT070 . tel-03329759

HAL Id: tel-03329759

<https://theses.hal.science/tel-03329759v1>

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

Examineur

Co-directeur de Thèse

Directeur de Thèse



UNIVERSITÉ
DE MONTPELLIER

CONTENTS

ACKNOWLEDGMENTS	4
TABLE OF FIGURES	5
LIST OF TABLES	6
ENGLISH SUMMARY	7
FRENCH SUMMARY	9
PUBLICATIONS	11
1. Introduction	12
1.1. Biology of cancer : an overview	13
1.1.1. Epidemiology	13
1.1.2. What is cancer?	13
1.1.3. Cancer is a genetic disease	14
1.1.4. Intratumor heterogeneity.....	16
1.1.5. EMT and metastatic cascade.....	16
1.2. Epithelial-Mesenchymal Transition (EMT)	18
1.2.1. Three different subtypes based on the biological context	20
1.2.1.1. During implantation, embryogenesis and organ development	20
1.2.1.2. Associated with tissue regeneration and organ fibrosis	21
1.2.1.3. Related to cancer progression	21
1.2.2. Resistance to treatment and survival.....	22
1.2.3. EMT regulatory programs.....	23
1.2.3.1. Transcriptional regulation.....	23
1.2.3.2. Post-transcriptional regulation.....	26
1.2.3.2.1. Regulatory network of micro-RNA.....	27
1.2.3.2.2. At the protein level	27
1.2.3.2.3. Emerging layers of regulation	28
1.3. Alternative splicing and tumor progressio	29
1.3.1. Splicing reaction.....	29
1.3.2. Aspects of its regulation.....	32
1.3.3. Definition of alternative (AS) splicing events	35
1.3.4. Alternative splicing and cancer	37
1.3.4.1. Affecting hallmarks of cancer	37
1.3.4.2. During EMT.....	41

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.	Hierachical 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.

FRENCH SUMMARY

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 épithélio-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

Thomas, A., Barriere, S., Broseus, L., Brooke, J., Lorenzi, C., **Villemin, J. P.**, Beurier, G., Sabatier, R., Reynes, C., Mancheron, A., & Ritchie, W. (2019). GECKO is a genetic algorithm to classify and explore high throughput sequencing data. *Commun Biol.* 2019;2:222. Published 2019 Jun 20. doi:10.1038/s42003-019-0456-9

Lorenzi C, Barriere S, **Villemin J.P**, Dejardin Bretones L, Mancheron A, Ritchie W. iMOKA: k-mer based software to analyze large collections of sequencing data. *Genome Biol.* 2020;21:261. Published 2020 Oct 13. doi:10.1186/s13059-020-02165-2

Leventoux, N., Augustus, M., Azar, S., Riquier, S., **Villemin, J. P.**, Guelfi, S., Falha, L., Bauchet, L., Gozé, C., Ritchie, W., Commes, T., Duffau, H., Rigau, V., & Hugnot, J. P. (2020). Transformation Foci in IDH1-mutated Gliomas Show STAT3 Phosphorylation and Downregulate the Metabolic Enzyme ETNPPL, a Negative Regulator of Glioma Growth. *Sci Rep.* 2020;10(1):5504. Published 2020 Mar 26. doi:10.1038/s41598-020-62145-1

IN PROCESS

Jean-Philippe Villemin, 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)

Alexandre Segelle*, Yaiza Núñez-Álvarez*, Kimberly M Webb, **Jean-Philippe Villemin**, Philipp Voigt, Reini F. Luco. Histone marks are drivers of splicing changes necessary for an epithelial-to-mesenchymal transition. (In preparation)

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 heterogenous 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 these 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 occur (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**).

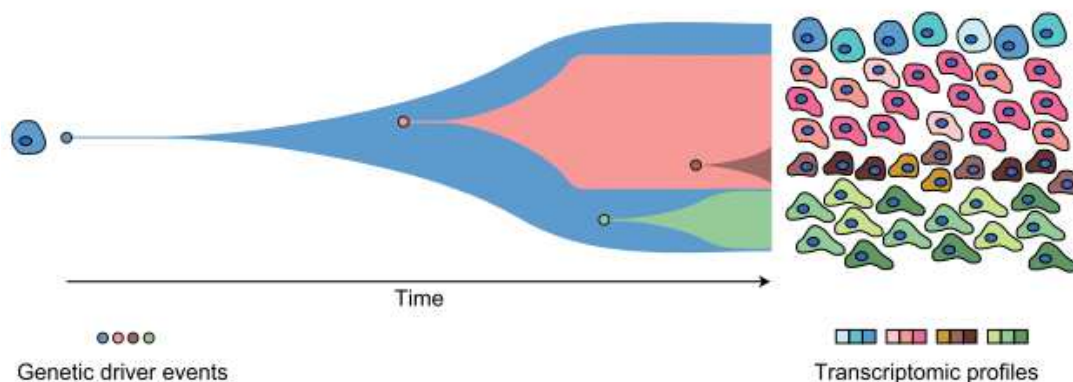


Figure 1-1 Evolutionary Trajectories and Transcriptomic Heterogeneity.

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 tumor 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 tumor 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

(Thiery 2002). 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) (Samatov, Tonevitsky, and Schumacher 2013).

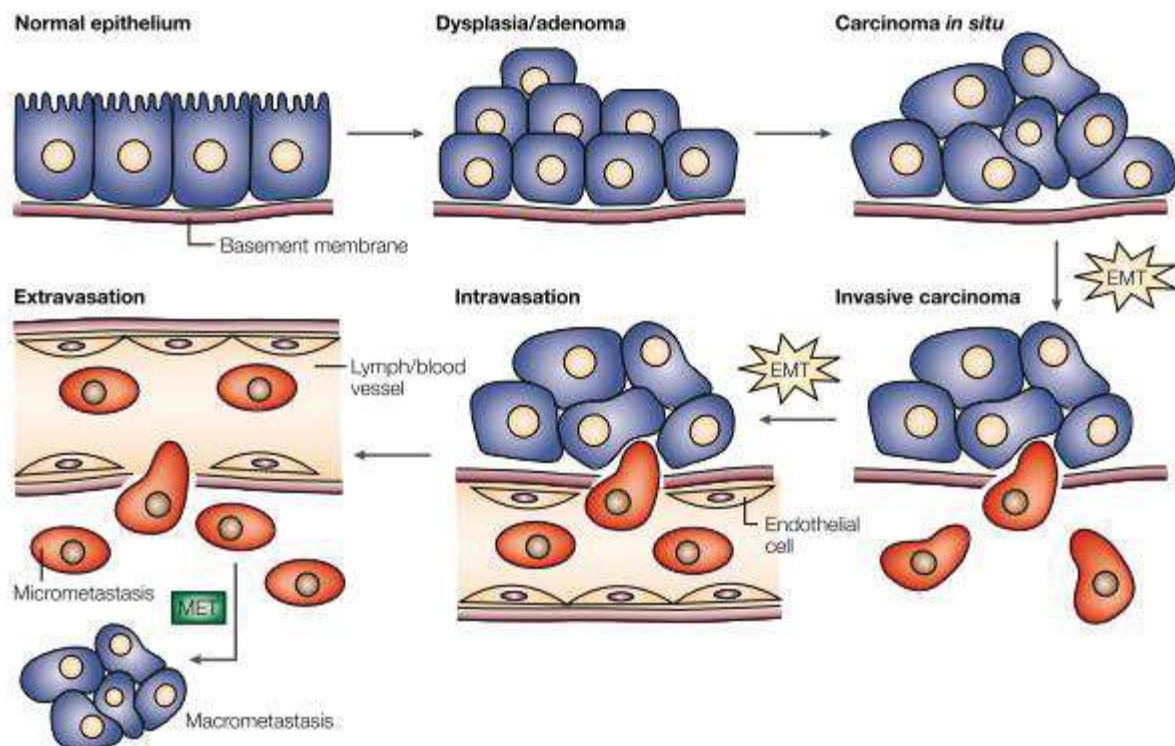


Figure 1-2 Sites of EMT & MET in the emergence and progression of carcinoma.

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 Thiery 2002)

Epithelial-to-Mesenchymal Transition (EMT) is a process that probably plays a role in the migration of cancer cells (Nieto et al. 2016; Pastushenko and Blanpain 2019; T. Brabletz et al. 2018; Lambert, Pattabiraman, and Weinberg 2016). 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 (W. Lu and Kang 2019; Ye and Weinberg 2015). Of note, a reverse process called

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 are 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).

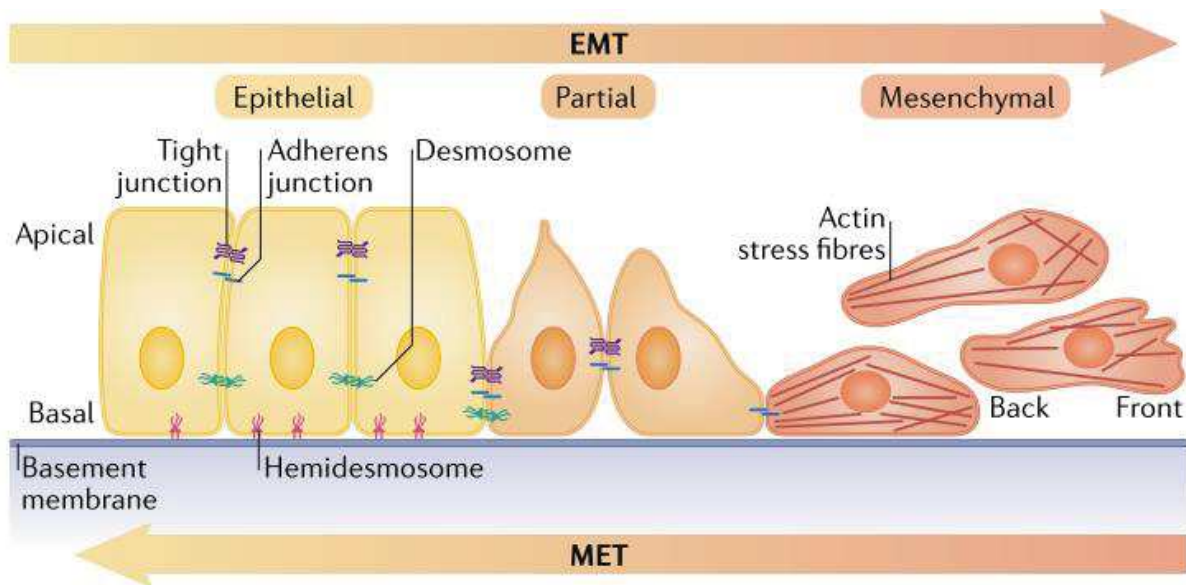


Figure 1-3 Outline of a typical EMT program.

Epithelial cells displaying apical–basal polarity are held together by tight junctions, adherens junctions and desmosomes and attached to the underlying basement membrane by hemidesmosomes. They follow a progressive loss of epithelial features, accompanied by acquisition of a partial set of mesenchymal features with retention of certain epithelial features. During EMT, cells become motile and acquire invasive capacities. Mesenchymal cells display front- to-back polarity and an extensively reorganized cytoskeleton. EMT is a reversible process, and mesenchymal cells can revert to the epithelial state by undergoing mesenchymal–epithelial transition (MET). (Adapted from Dongre and Weinberg 2018)

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.

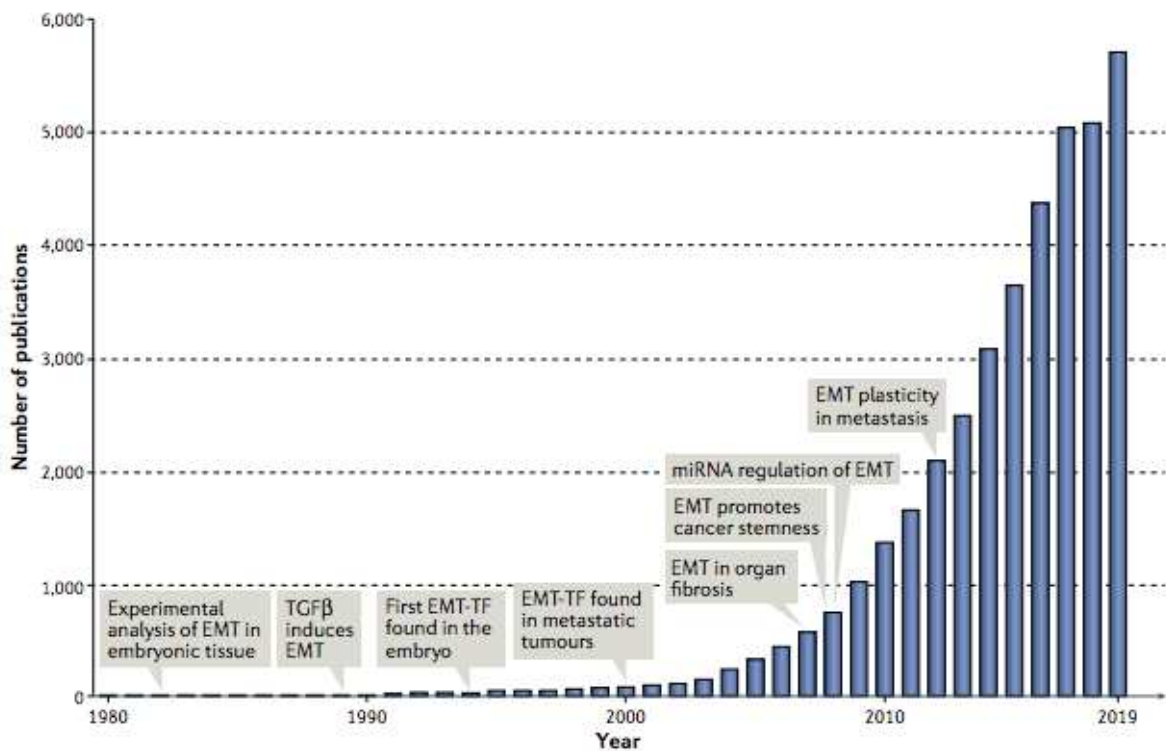


Figure 1-4 Growth of the primary literature in EMT.

The graph indicates primary papers published each year, identified by a search of the Web of Science database. The total numbers of publications in 2019 is 5,700 articles. Growth in the field has been logarithmic since 2003 (Adapted from EMT International Association 2020).

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 specific 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-

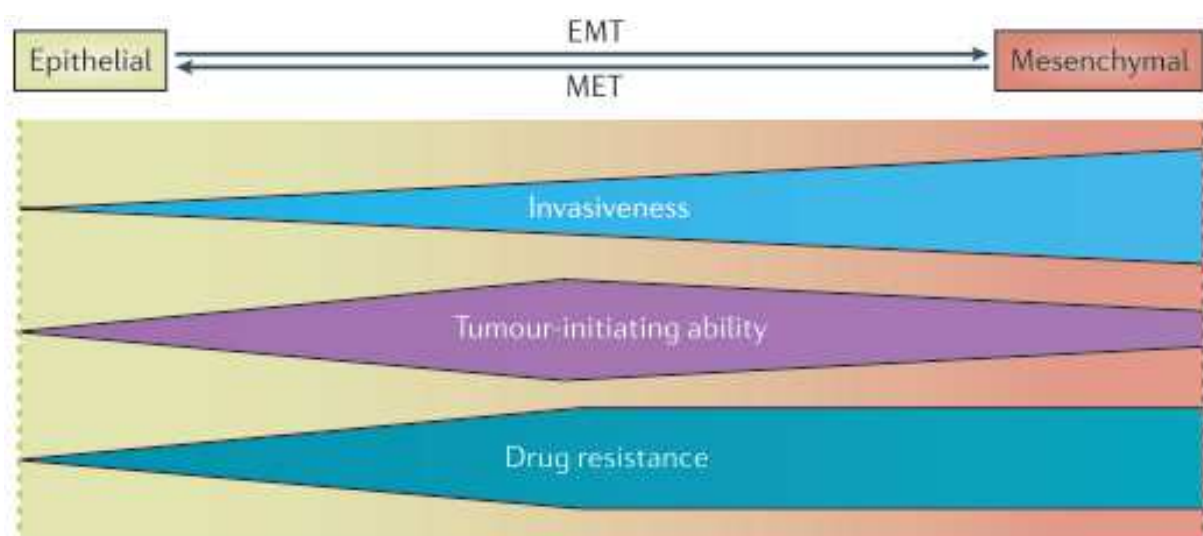


Figure 1-5 Summary of the physiological outcomes of EMT in carcinoma.

This figure shows the extent of invasiveness, the tumour-initiating ability, and degree of drug resistance of carcinoma cells that are thought to change across the spectrum of EMT-programme. (Adapted from Shibue & Weinberg 2017)

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).

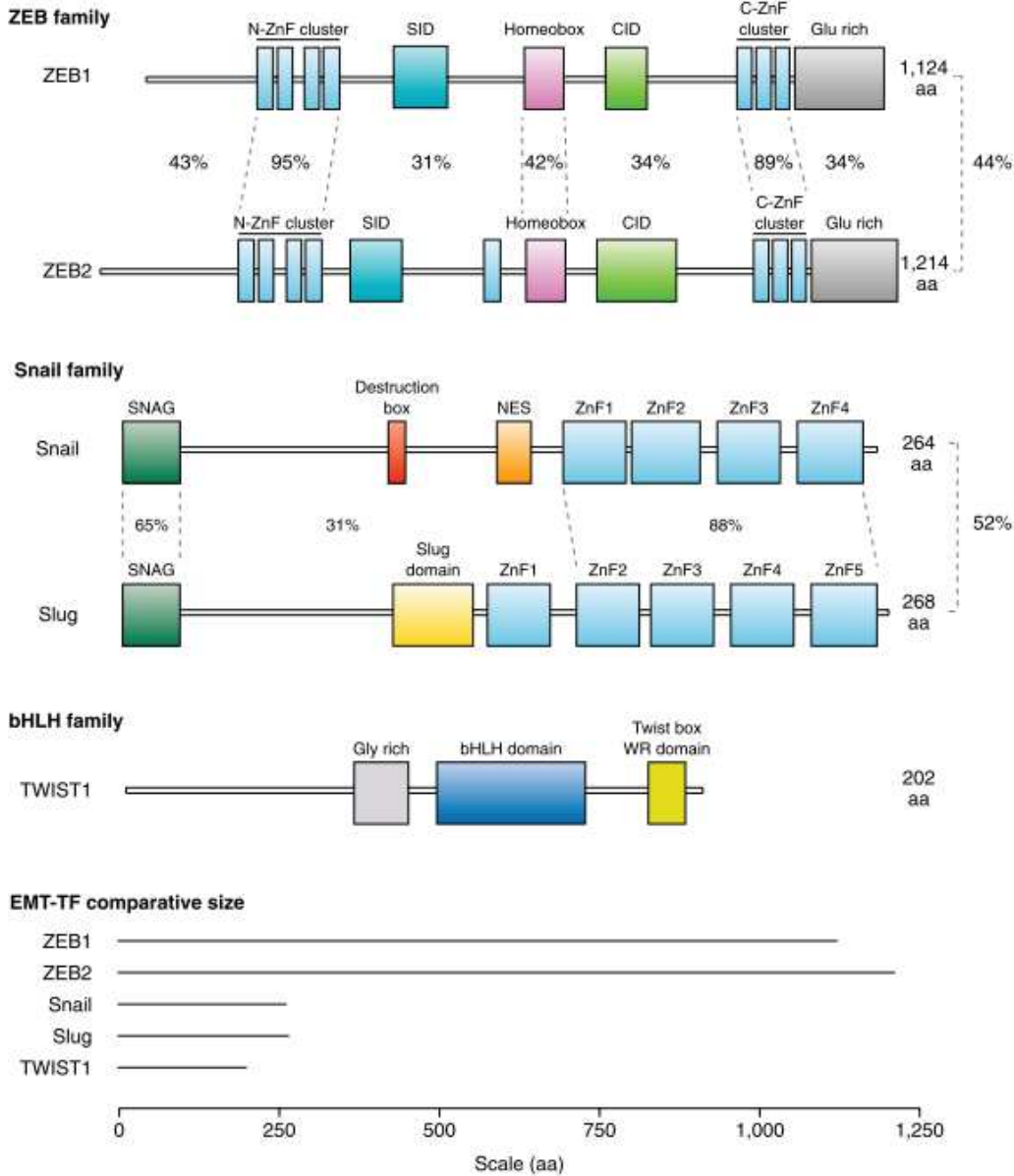


Figure 1-6 Overview of EMT-TF protein structures.

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

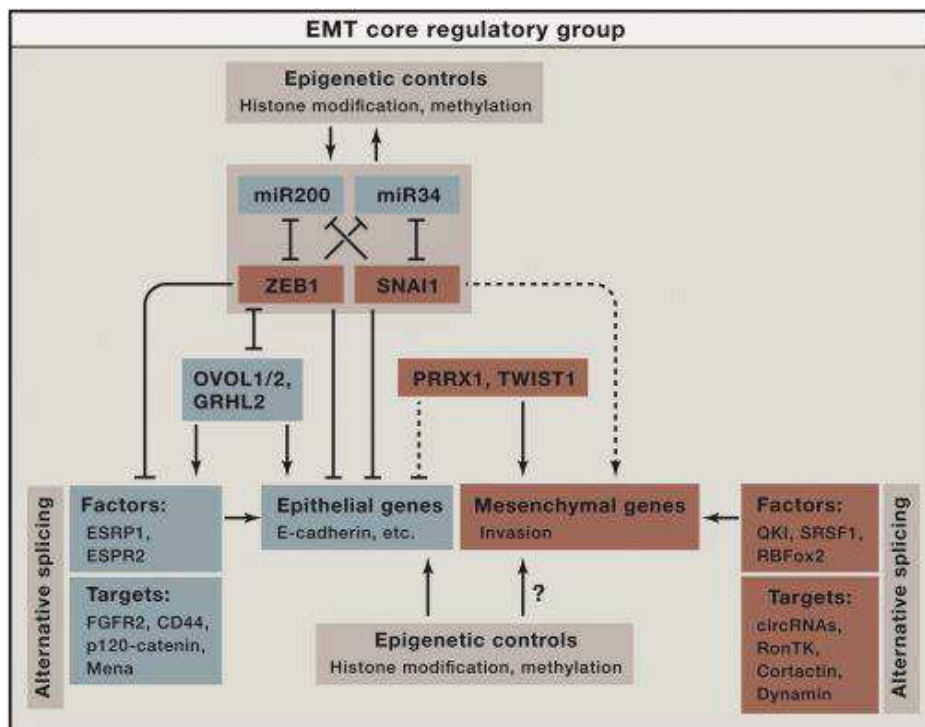


Figure 1-7 Multilayer of regulation during EMT

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 molecule (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-3 β phosphorylates SNAIL at two consecutive motifs that control its ubiquitination (B. P. Zhou et al. 2004). First, GSK-3 β 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-3 β 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.

1.3. ALTERNATIVE SPLICING AND TUMOR PROGRESSIO

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.

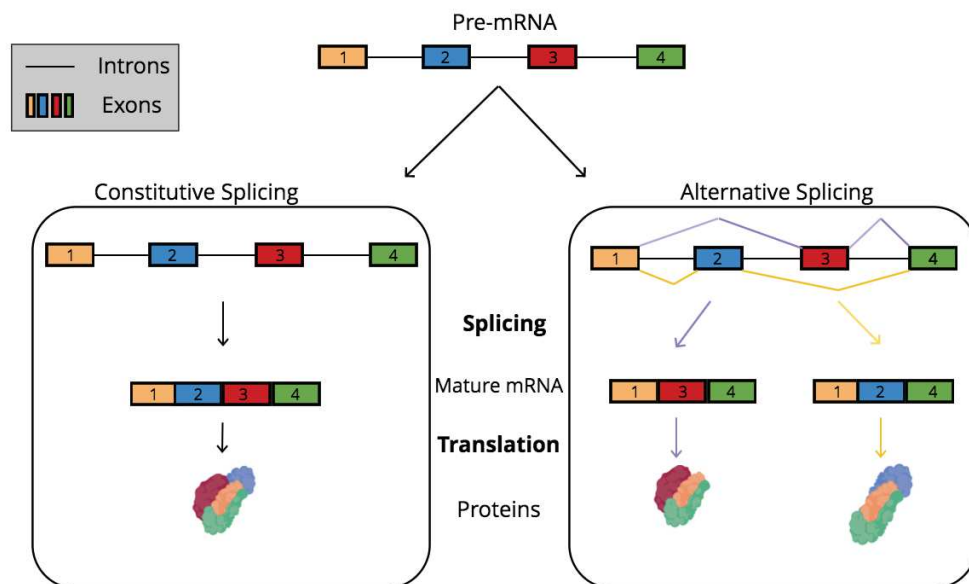


Figure 1-8 Constitutive splicing and alternative splicing.

Constitute exons involves the excision of all introns 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.

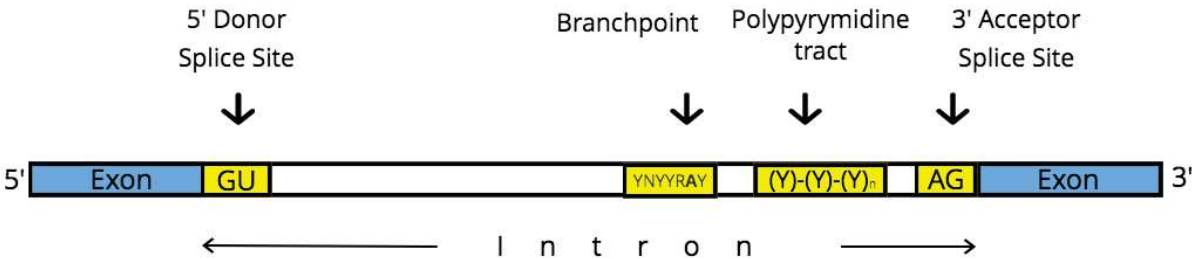


Figure 1-9 Exons and introns in pre-mRNA with their consensus sequences.

Within introns, a donor site (5' end of the intron), a branch site (near the 3' end of the intron) and an acceptor site (3' end of the intron) are required for splicing. The GU dinucleotide and the AG nucleotide, respectively donor and acceptor sites, are highly conserved. The polypyrimidine tract lies between the branch site and 3' intron–exon junction. Further upstream from the polypyrimidine tract is the branchpoint, which includes an adenine nucleotide involved in lariat formation (adapted from Srebrow and Kornblihtt, 2006).

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).

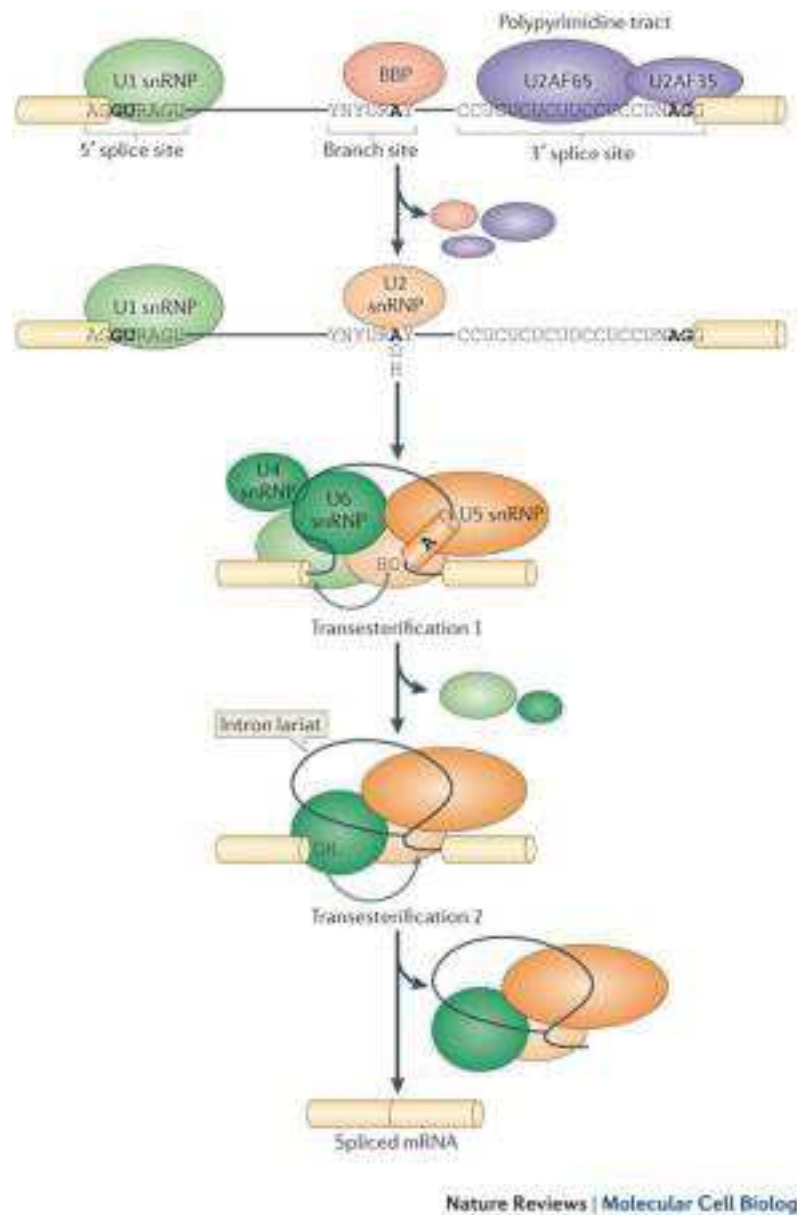


Figure 1-10 The spliceosome mediates a two-step splicing reaction.

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).

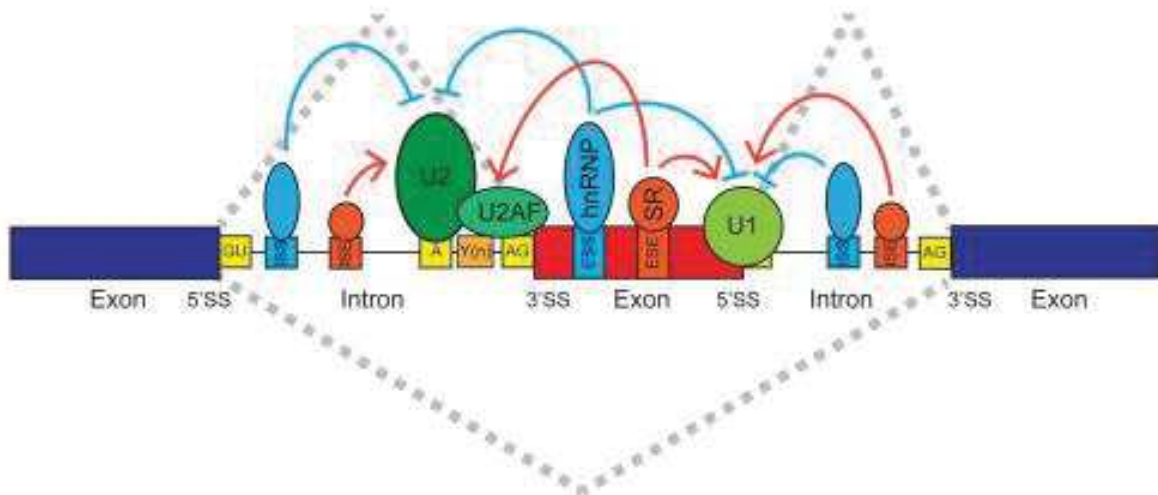


Figure 1-11 Cis and Trans regulation of alternative splicing.

Alternative splicing is regulated by an extensive protein-RNA interaction network involving CIS elements within the pre-mRNA and TRANS-acting factors that bind to these CIS elements. Exonic splicing enhancers (ESEs), exonic splicing silencers (ESSs), intronic splicing enhancers (ISEs), and intronic splicing silencers (ISSs) are pre-mRNA cis regulatory motifs that recruit various RNA-binding proteins (e.g., SR and hnRNP proteins) to regulate alternative splicing (adapted from Park, 2018).

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.

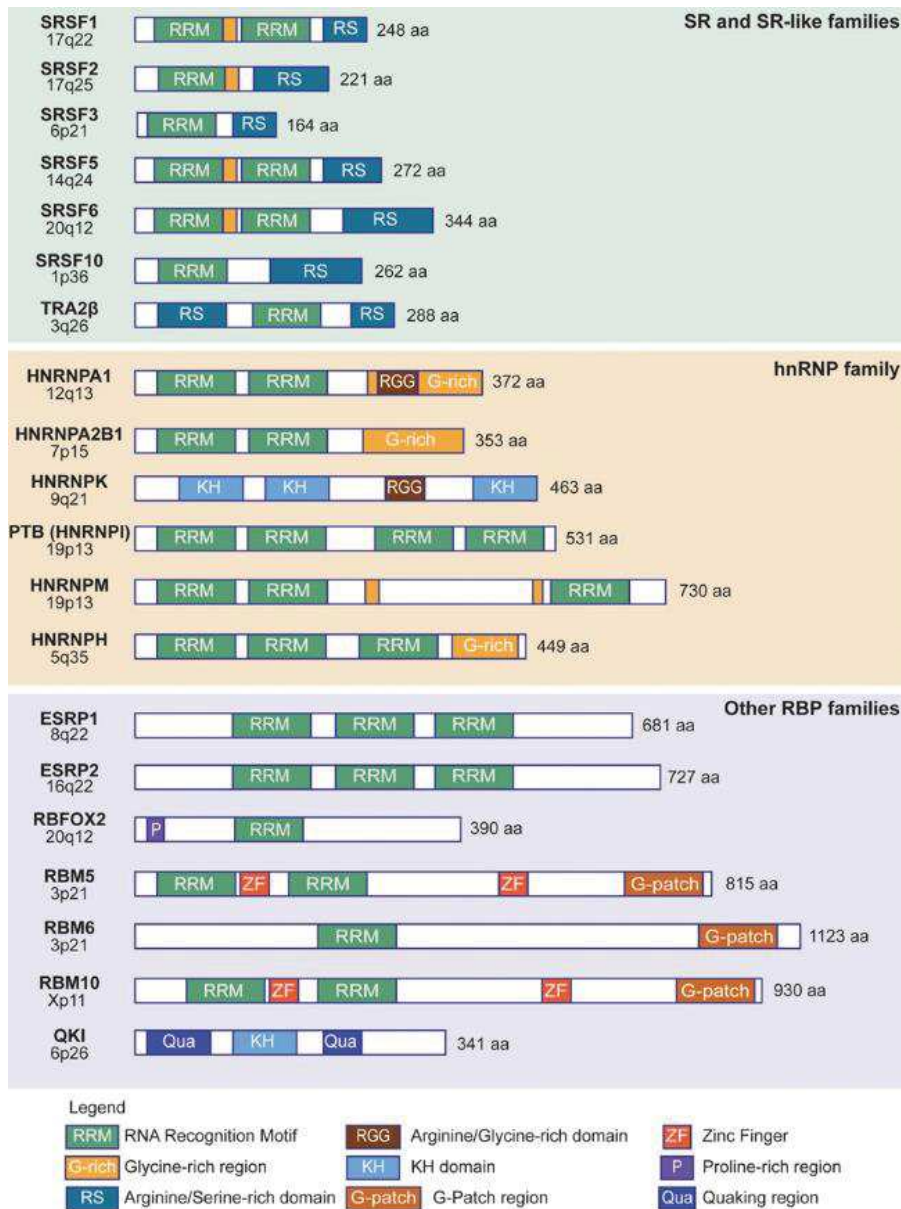


Figure 1-12 Domain structure of splicing factors altered in solid tumors.

For each RNA-binding protein (RBP) representative of the indicated families, the annotated protein domains or regions are shown in the diagrams (see legend for details), along with the size (in amino acids) of the human protein. (adapted from Anczukow & Krainer, 2016)

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-Barríos et al. 2018; L.-Y. Zhu et al. 2018). For example, enrichment of H3K79m2 have been observed in specific AS events

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) SPLICING 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 major splicing events are defined (Figure 1-13) (E. Wang and Aifantis 2020).

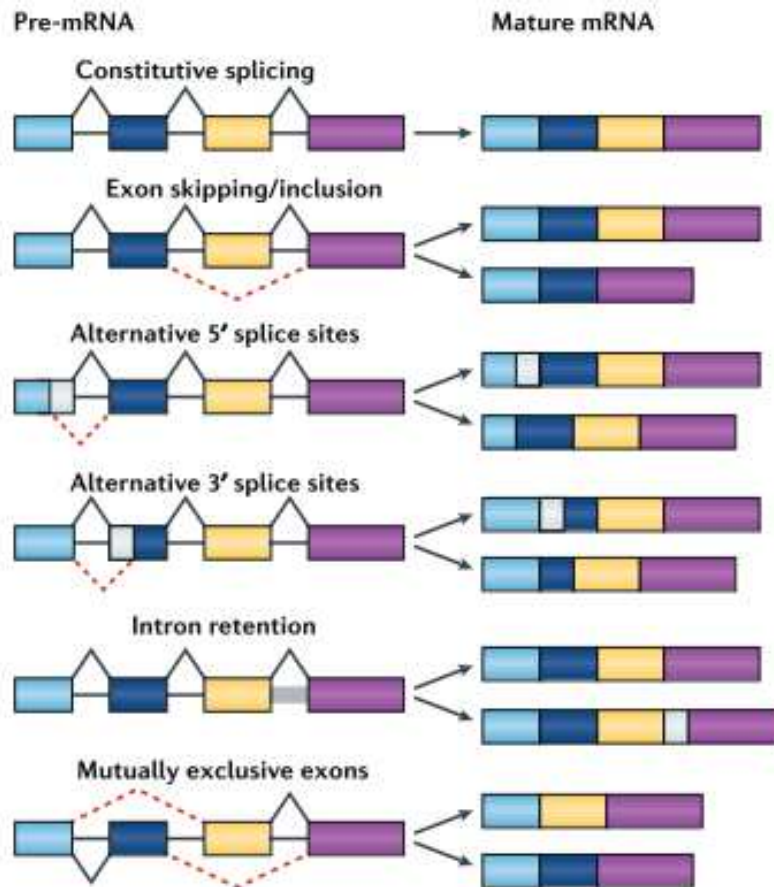


Figure 1-13 Constitutive splicing versus alternative splicing events.

Schematic depicting constitutive splicing, as well as the five common modes of alternative splicing: exon skipping/inclusion, alternative donor, alternative acceptor, intron retention and mutually exclusive exons. Shown on the right are the mature mRNA transcripts derived from each event (adapted from Franki, 2019).

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-apoptotic 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-angio-genic 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) ([Björklund et al. 2017](#)). More recently, different AS patterns in tumors have been demonstrated again using TCGA dataset ([Sebestyén, Zawisza, and Eyraş 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.

Cancer hallmark	Gene name	Splicing event type	Isoform structure and function	Tumor types													Experimental evidence			
				Bladder	Breast	Colorectal	Gynecologic	Head & neck	Hematologic	Kidney	Liver	Lung	Pancreatic	Prostate	Skin	Stomach	Other	Cell lines OE	Cell lines KD	Xenograft
+	BIN1	ES	BIN1 pro-apoptotic			SRSF1, HNRNPA2/B1														
			BIN1+12A anti-apoptotic																	
+	BCL2L1	A5'SS	BCL-xS pro-apoptotic			SRSF1, Sam68, RBM4, RBM25, RBM5, RBM10, HNRNPA1, PTBP1, HNRNPA2B/1														
			BCL-xL anti-apoptotic																	
+	BCL2L11	ES	BIM-EL, -L, -S pro-apoptotic			SRSF1, SRSF6, PTBP1, HNRNPC														
			BIM-y anti-apoptotic																	
+	CASP2	ES	CASP-2L pro-apoptotic			SRSF3, RBM5														
			CASP-2S anti-apoptotic																	
→	CCND1	IR	CCND1-a pro-proliferative																	
			CCND1-b pro-invasive																	
→	CD44	ES	CD44s mesenchymal			SRSF2, TRA2β, ESRP1, ESRP2, HNRNPA1, HNRNPL														
			CD44v epithelial																	
→	ENAH	ES	ENAH-11a epithelial anti-invasive																	
			ENAHΔv6 mesenchymal pro-invasive																	
			ENAH-INV mesenchymal pro-invasive																	
+	FAS	ES	FAS-FL pro-apoptotic			HNRNPA1, TIA1, RBM5, PTBP1, EWS														
			sFAS anti-apoptotic																	
→	FGFR	MXE	FGFR-IIIb tumor-suppressive epithelial			HNRNPH1, HNRNPF, ESRP1, ESRP2														
			FGFR-IIIc pro-proliferative, pro-invasive mesenchymal																	
→	HER2	ES	HER2-FL proliferative			SRSF3, HNRNPH1														
			d16HER2 pro-proliferation, pro-invasive																	
→	HRAS	ES	HRAS-IDX pro-proliferative																	
			HRAS-FL tumor suppressive																	
→	KLF6	A5'SS	KLF6-FL tumor suppressive			SRSF1, TGF-β1, RAS signalling														
			KLF6-SV1 pro-proliferative pro-invasive																	
+	MCL1	ES	MCL-1ES pro-apoptotic			SRSF1														
			MCL-1L anti-apoptotic																	
→	MKNK2	ES	MKNK2-a pro-apoptotic, anti-proliferative			SRSF1, SRSF6														
			MKNK2-b anti-apoptotic, pro-proliferative																	

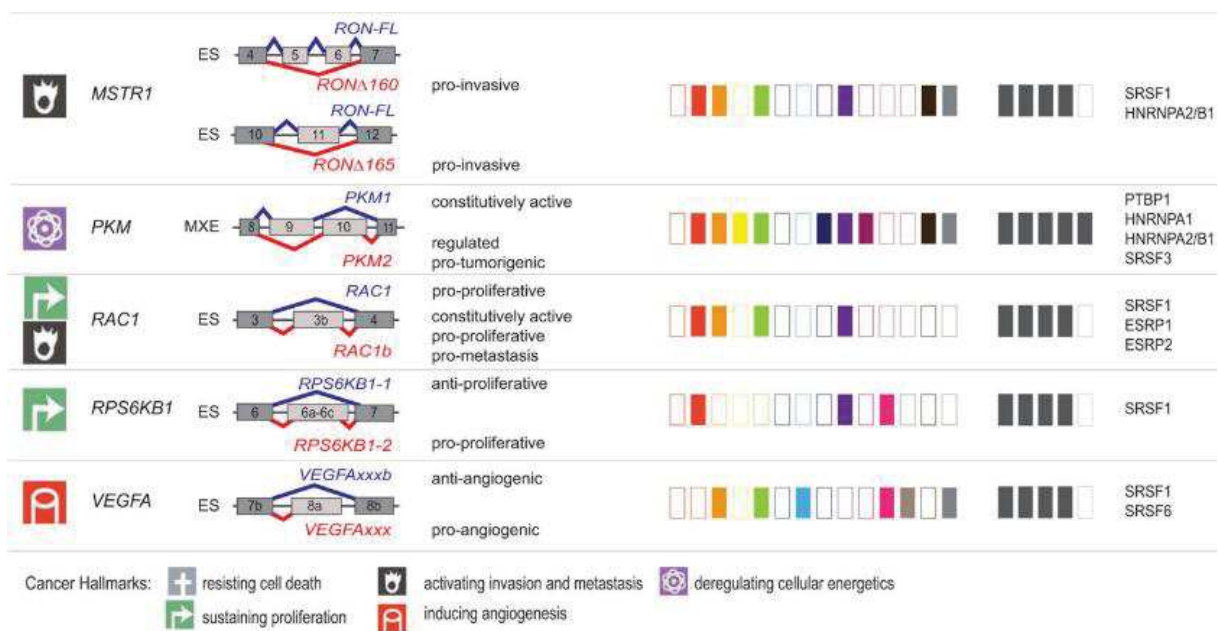


Table 1.1 Tumor-associated isoforms representative of the cancer hallmarks.

Splicing event type, isoform structure, tumor expression, and experimental evidence for selected alternative splicing isoforms detected in human tumors (adapted from Urbanski, 2018).

1.3.4.2. DURING EMT

As I mentioned before, EMT is 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 Durtertre & 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 IIIc, 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 IIIc 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-IIIc 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 it 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).

Gene (other names)	Function	Epithelial splicing	Domain affected/functional difference
FGFR2	Transmembrane receptor tyrosine kinase	Exon IIIb	Confers ligand binding specificity
CTNND1 (p120-catenin)	Delta-catenin; regulator of cell adhesion and signaling	Skipped	Coiled-coil domain; stabilizes interaction with RalA
CD44	Cell-surface glycoprotein involved in cell adhesion and migration	Exons v8-v10	Extra-cellular membrane proximal region; creates a heavily glycosylated stalk
ENAH (Mena)	Regulator of actin dynamics	Included	Ena/Vasp homology domain; contains a phosphorylation site that may disrupt actin binding
NUMB	Complex protein implicated in many roles including cell migration and adhesion	Included	Phosphotyrosine binding domain; the encoded peptide confers localization to the plasma membrane
FLNB	F-actin cross-linking protein	Included	Contributes to the hinge domain; allows for more rigid actin branching
DNM2	GTPase that binds cytoskeletal proteins	Included	Pleckstrin homology domain; affects subcellular localization
TCF7L2 (Tcf4)	Transcription factor involved in Wnt signaling pathway	Included	Differential activation of Wnt/ β -catenin target genes
BAIAP2 (Irsps53)	Cdc42 effector protein involved in lamellipodia and filopodia formation	Included	Pentultimate exon with stop codon; differentially phosphorylated in response to IGF-1
MAP3K7 (Tak1)	Kinase that mediates TGF- β and BMP signal transduction	Included; Skipped	Peptide encoded by downstream exon is required for interaction with Tab2/3
ARHGAP17 (Rich1)	GTPase-activating protein involved in maintenance of the tight junction	Skipped	Part of proline rich domain
MAGI1 (Baiap1)	Scaffolding protein associated with complexes at the inner plasma membrane	Skipped	Encodes peptide between the two WW domains
LRRFIP2	Involved in activation of Wnt signaling	Skipped	Predicted coiled coil domain; encoded peptide may enhance interaction with Dvl3
SCRIB	Scaffolding protein associated with tight junctions and cell polarity	Skipped	Encodes a peptide proximal to the first PDZ domain
EPB41L5 (Ymo1)	A FERM protein that interacts with Crumbs complex to regulate cell architecture	Short isoform	Paxillin-binding domain; enhances focal adhesion complexes
RALGPS2	A guanine nucleotide exchange factor involved in cytoskeleton reorganization	Included	Between a PxxP motif and a pleckstrin homology domain; may influence GEF activity
ITGA6	Alpha subunit of integrin, a laminin receptor	Included	Light chain and cytoplasmic domain; changes C-terminus sequence
SLK	STE20-like kinase with a role in promoting cell motility	Included	Predicted coiled-coil domain; may specify interaction partners
ARHGEF11 (PDZ-RhoGEF)	RhoA-specific guanine nucleotide exchange factor	Skipped	C-terminus; may influence homodimerization or interaction with PAK4 and LARG

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.

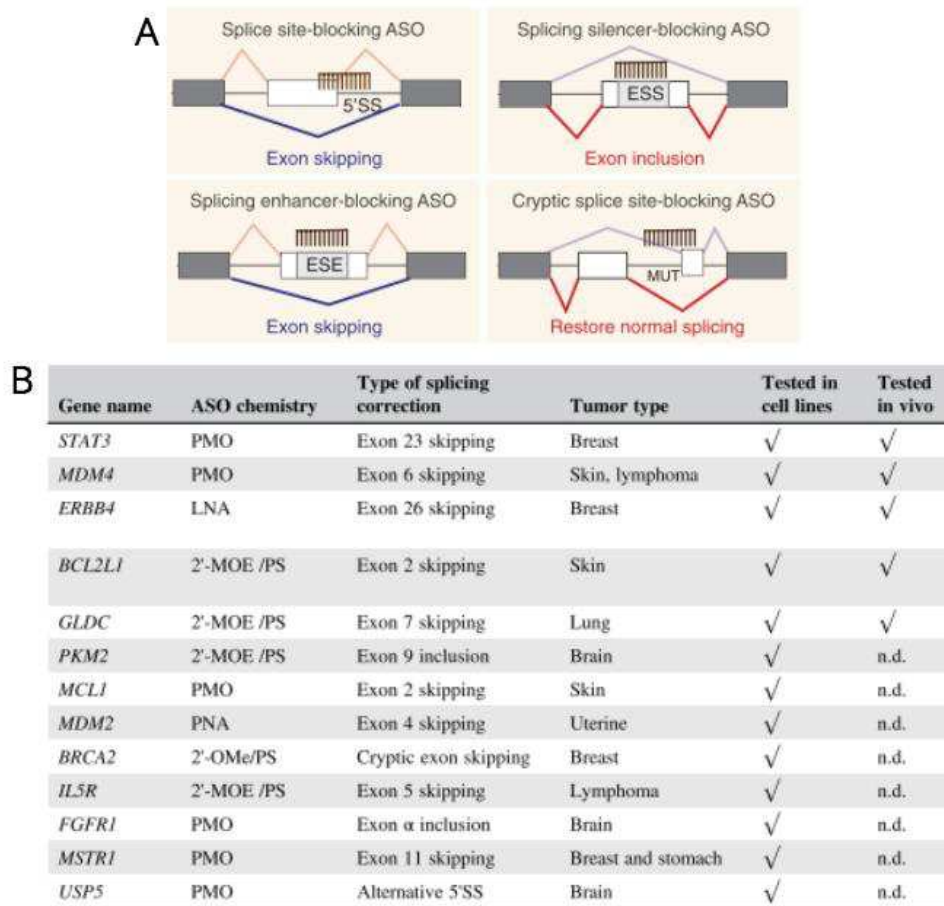


Figure 1-14 Therapeutic strategies to target-splicing alterations in tumors.

(A) Isoform specific inhibition can be achieved by using splice-switching antisense oligonucleotides (ASOs) that bind in a sequence specific manner and modulate the outcome of a specific splicing isoform. ASOs can promote exon skipping or inclusion by blocking the 5'SS, an exonic silencer (ESS), or enhancer element (ESE) or by preventing the usage of a mutant (MUT)/cryptic splice site. (B) List of cancer-associated human isoforms targeted by splice-switching ASO. (adaptated by Urbanski, 2018)

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.

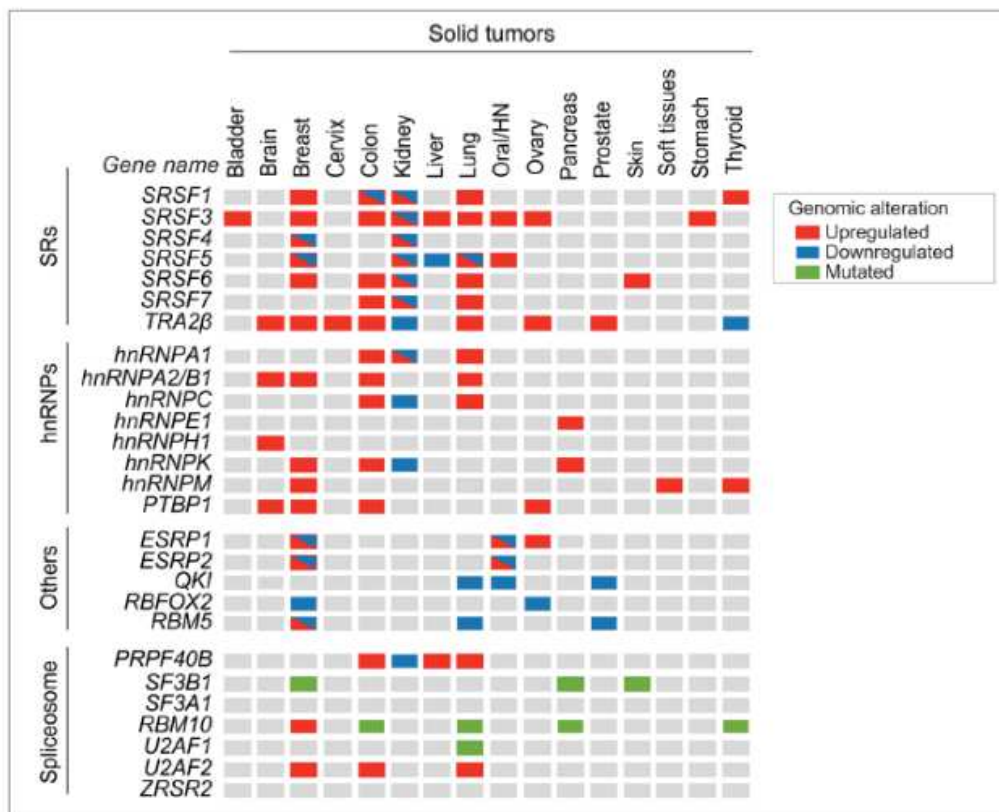


Figure 1-15 Recurrent splicing factor alterations detected in cancer.

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.

Splicing factor	Downstream dysregulated isoforms that are functionally linked to cancer
ESRP1 and ESRP2	Promote an epithelial splicing programme to regulate EMT ^{61,62}
hnRNP A1	Contributes to aerobic glycolysis in cancer by promoting the expression of specific isoforms of pyruvate kinase (PKM2 isoform) ^{46,47} and MAX (the delta MAX isoform) ⁴⁸
hnRNP A2	Contributes to aerobic glycolysis in cancer by promoting the expression of a specific isoform of pyruvate kinase (PKM2 isoform) ^{46,47} . Increases breast cancer cell invasion by promoting the expression of a specific isoform of <i>TP53/INP2</i> (REF. 195)
hnRNP A2/B1	Acts as an oncogenic driver in glioblastoma by regulating splicing of several tumour suppressors and oncogenes, including <i>RON</i> ⁴⁹
hnRNP H	Contributes to survival of gliomas and invasion by promoting the expression of specific isoforms of <i>IG20</i> and <i>RON</i> , respectively ¹⁰⁶
hnRNP K	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
hnRNP M	Contributes to EMT in breast cancer and increases metastasis in mice by promoting the expression of a specific isoform of <i>CD44</i> (<i>CD44s</i>) ¹⁹⁷
PRPF6	Promotes cell proliferation in colon cancer by altering splicing of genes associated with growth regulation, including the kinase gene <i>ZAK</i> ¹⁰⁸
PTB (PTBP1)	Contributes to aerobic glycolysis in cancer by promoting the expression of a specific isoform of pyruvate kinase (PKM2 isoform) ^{46,47} . Also promotes the expression of an isoform of the deubiquitylating enzyme-encoding gene <i>USP5</i> (REF. 199) that has been shown to promote glioma cell growth and mobility (although the mechanism underlying this phenotypic association is not resolved)
QKI	Acts as a tumour suppressor by regulating alternative splicing of <i>NUMB</i> in lung cancer cells ⁵¹
RBFOX2	Promotes a mesenchymal splicing programme to regulate EMT ⁶²
RBM4	Acts as a tumour suppressor by promoting the pro-apoptotic isoform BCL-X _s of <i>BCL2L1</i> and opposing the pro-tumorigenic effects of SRSF1 on mTOR activation ⁶⁰
RBM5, RBM6 and RBM10	RBM5 modulates apoptosis by regulating alternative splicing of <i>CASP2</i> (REF. 200) and <i>FAS</i> ²⁰¹ . RBM5 or RBM6 depletion has an opposite effect to RBM10 depletion, as these factors antagonistically regulate the alternative splicing of <i>NUMB</i> ^{58,59}
SRSF1	Promotes an isoform of the kinase MNK2 that promotes eIF4E phosphorylation independently of MAPK signalling ³⁹ . In the context of breast cancer, SRSF1 overexpression promotes alternative splicing of <i>BIM</i> and <i>BIN1</i> (REF. 40) to promote the expression of isoforms that lack pro-apoptotic functions
SRSF3	Regulates alternative splicing of <i>TP53</i> (REF. 43) such that SRSF3 loss promotes expression of p53 β , an isoform of p53 that promotes p53-mediated senescence
SRSF6	Promotes expression of isoforms of the extracellular matrix protein tenascin C that are characteristic of invasive and metastatic skin cancer ⁴⁴ , contributing to epithelial cell hyperplasia
SRSF10	Promotes cell proliferation and colony formation <i>in vitro</i> and increases tumorigenic capacity of colon cancer cells in mice by inducing expression of a specific isoform of <i>BCLAF1</i> (<i>BCLAF1-L</i>) ¹⁰²

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.

1.3.4.4. RELATED TO EMT

As mentioned just earlier, the splicing factor RBFOX2 regulates EMT, and have many splicing targets in breast, pancreatic, and colon tumors ([J. P. Venable 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.

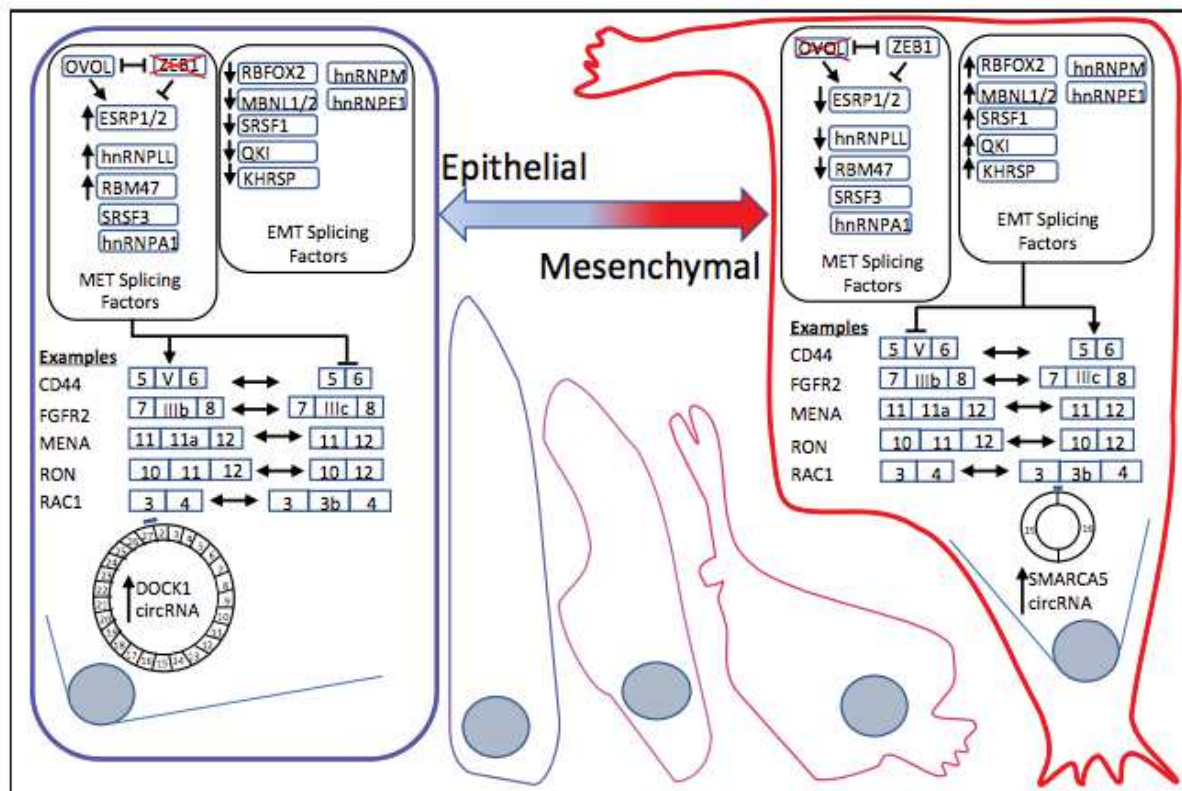


Figure 1-16 Splicing Factors changes and AS events during EMT.

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 RESSOURCES 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.

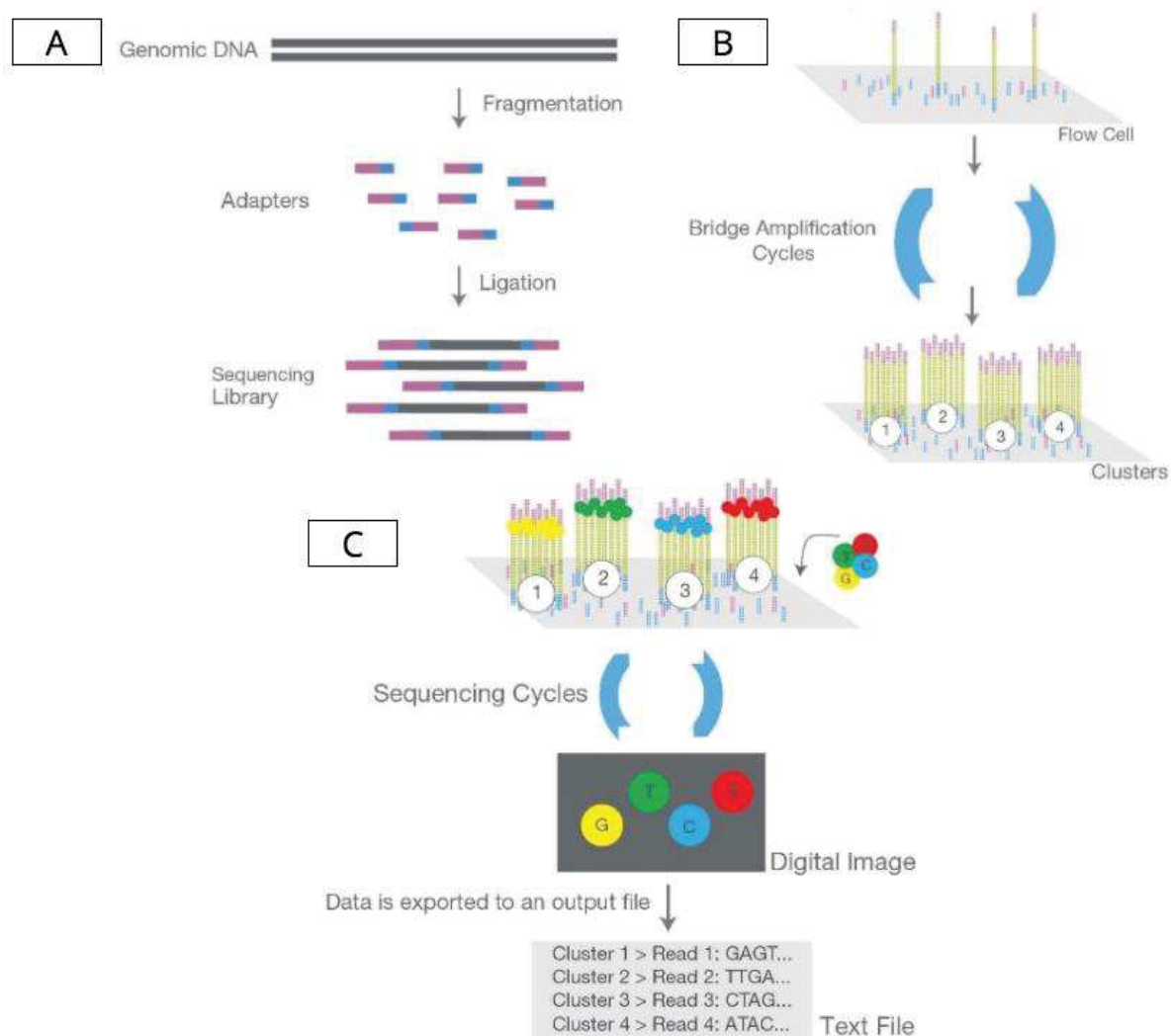


Figure 1-17 Principle and Workflow of Illumina Next-generation Sequencing.

(A) Library preparation: Through ultrasonic fragmentation, the genomic DNA becomes DNA fragment. Then fragments are ligated to adapters. (B) Cluster generation: Library is loaded on a flow cell and the fragments are hybridized to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification. (C) Sequencing: Sequencing reagents, including fluorescently labeled nucleotides, are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated N times to create a read length of N bases.

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**).

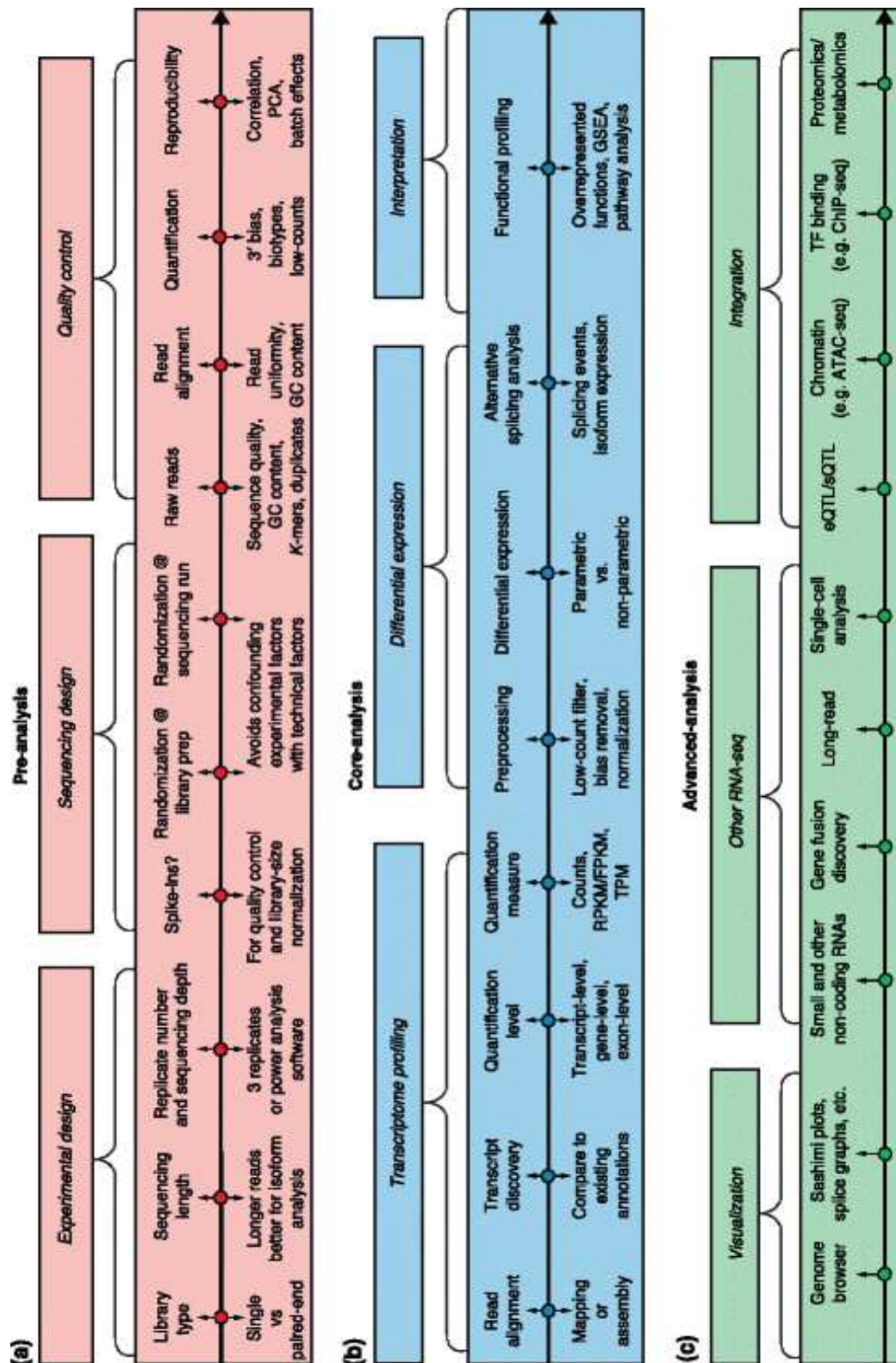


Figure 1-18 RNA-seq computational analyses.

The graphic is divided into three parts: pre-analysis, core analysis and advanced analysis of RNA-seq. With differential expression analyze, alternative splicing analyze are central components of the core analysis of RNA-seq but other advanced analyzes can be performed as Gene Fusion discovery for example (adapted from Conesa & Madrigal 2016).

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.

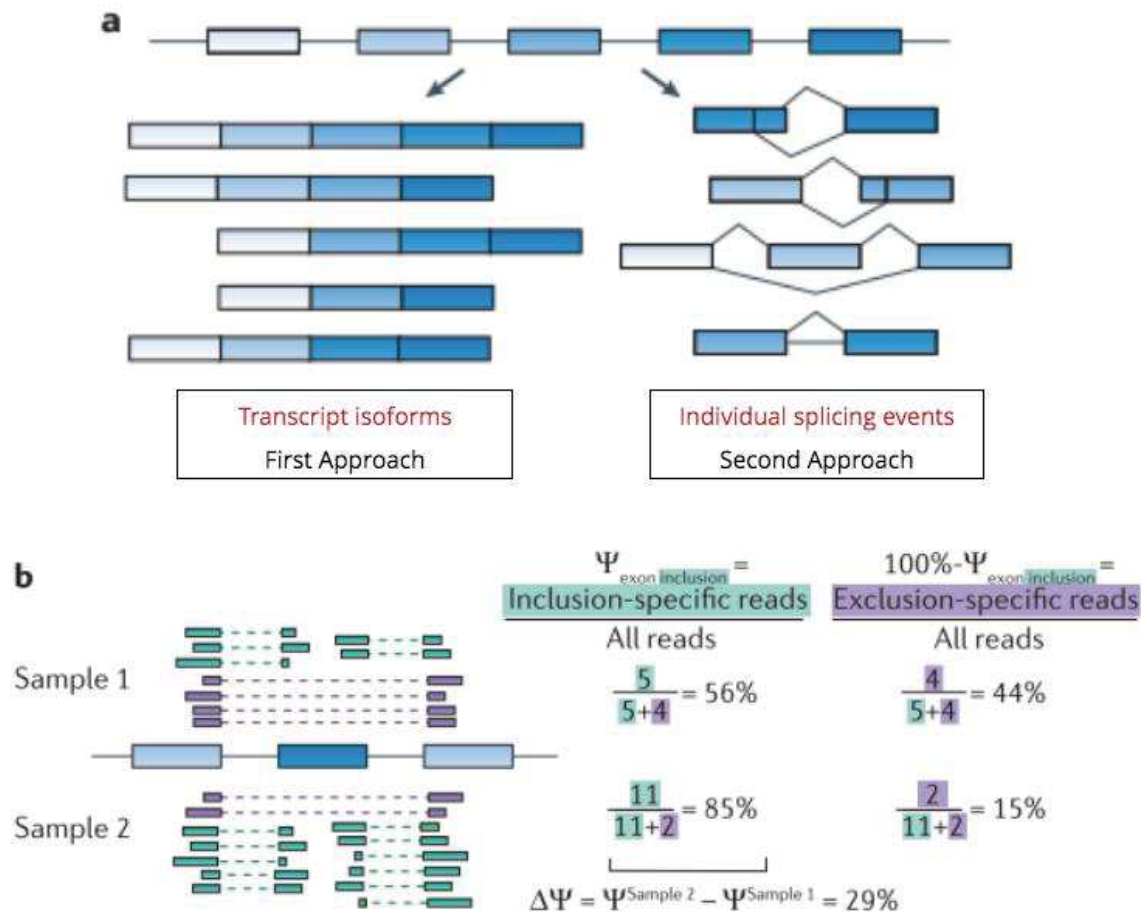


Figure 1-19 Methods to determine differential RNA splicing using RNA-seq data.

(a) Approaches for the study of alternative splicing: Two major approaches can be distinguished. The first one consists in the reconstruction of the transcripts followed by comparison of their expression levels. The second approach focus on a single event, for instance, the inclusion or excision of an exon, in the case of exon skipping. (b) General concept of PSI calculation: Reads that support inclusion of exon are in green. Reads that support exclusion are in purple. For each sample, a ratio called Percentage Spliced-In (PSI) is computed. Then, the difference (delta-psi) is calculated between the two samples and gives an idea of the change in the level of inclusion of an exon between two samples (adapted from Divinge 2016).

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; Kahles 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).

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

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.

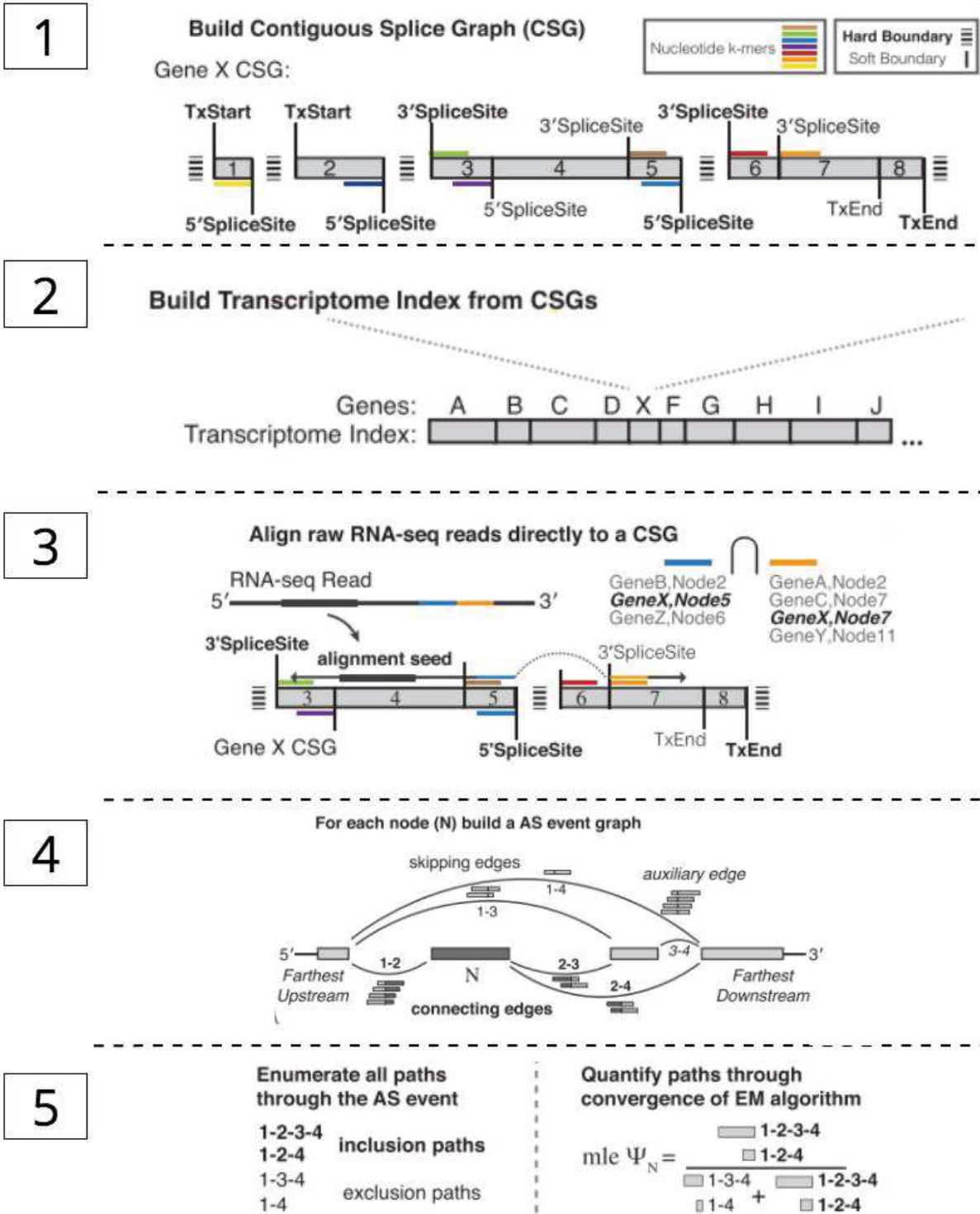


Figure 1-20 Focus on Whippet: global methodology.

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. HIERACHICAL 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.

Endpoint	Definition
OS (overall survival)	It is the period from the date of diagnosis until the date of death from any cause.
DSS (disease-specific survival)	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.
PFI (progression-free interval)	It's the period from the date of diagnosis until the date of the first occurrence of a new tumor event.
DFI (disease-free interval)	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.

Table 1.5 Definition of endpoints in clinical trials.

An endpoint is the primary outcome that is being measured by a clinical trial. However, there is different types of endpoint that can be measured. In this table, we report the distinct endpoints and their definitions as given by TCGA. (adapted from Liu, 2018)

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 in 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 hipoxia, 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 RESSOURCES

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.

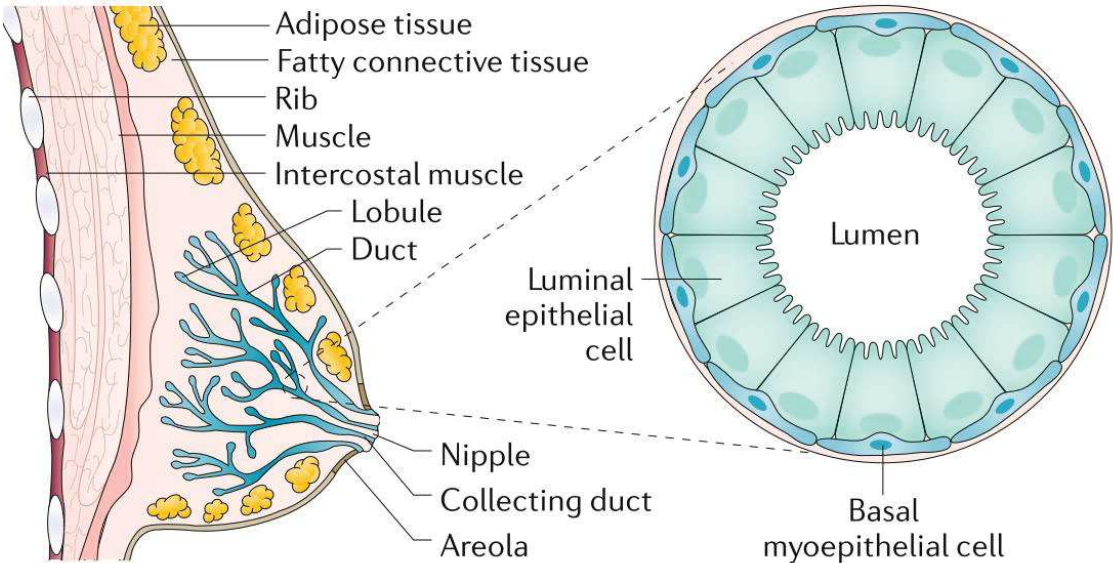


Figure 1-21 Breast anatomy and histology.

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,

[Sørli, 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**).

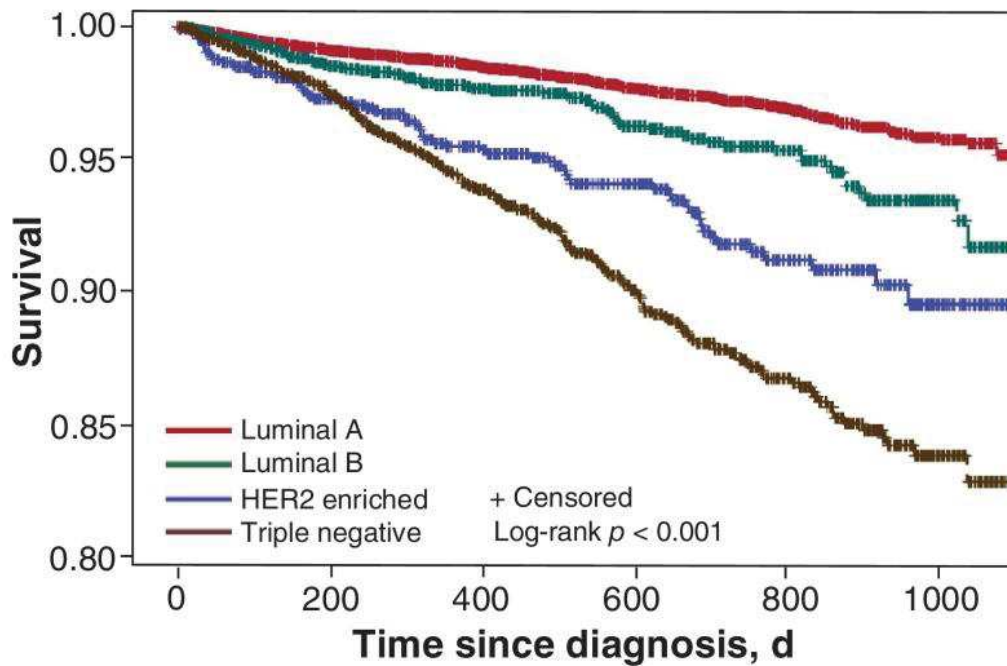


Figure 1-22 Breast cancer survival by molecular subtype.

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).

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.

1.5.5. FOCUS ON BASAL-LIKE BREAST CANCER

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).

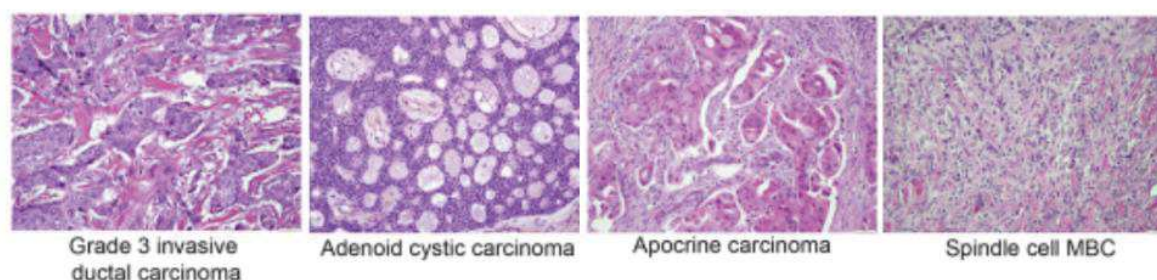


Figure 1-23 Histological heterogeneity of Triple Negative Breast cancer

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, Sørlie, 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

1 **A cell-to-patient machine learning transfer approach uncovers novel basal-like**
2 **breast cancer prognostic markers amongst alternative splice variants**

3
4 Jean-Philippe Villemin¹, Claudio Lorenzi¹, Marie-Sarah Cabrillac¹, Andrew Oldfield¹,
5 William Ritchie^{1*} & Reini F. Luco^{1*}
6

7 1. Institut de Génétique Humaine (IGH), Centre National de la Recherche
8 Scientifique, University of Montpellier, Montpellier, France
9

10 Corresponding authors: william.ritchie@igh.cnrs.fr and reini.luco@igh.cnrs.fr
11

12 **ABSTRACT:**

13 **Background:**

14 Breast cancer is amongst the 10 first causes of death in women worldwide. Around
15 20% of patients are misdiagnosed leading to early metastasis, resistance to treatment
16 and relapse. Many clinical and gene expression profiles have been successfully used
17 to classify breast tumours into 5 major types with different prognosis and sensitivity to
18 specific treatments. Unfortunately, these profiles have failed to subclassify breast
19 tumours into more subtypes to improve diagnostics and survival rate. Alternative
20 splicing is emerging as a new source of highly specific biomarkers to classify tumours
21 in different grades. Taking advantage of extensive public transcriptomics datasets in
22 breast cancer cell lines (CCLE) and breast cancer tumours (TCGA), we have
23 addressed the capacity of alternative splice variants to subclassify highly aggressive
24 breast cancers.

25 **Results:**

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

39 **Conclusions:**

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

48

49 **KEYWORDS** Alternative Splicing, Breast Cancer, Survival, Basal-like, Epithelial-to-
50 Mesenchymal Transition, Machine Learning Classification.

51 **BACKGROUND:**

52 Breast cancer is a heterogenous disease with multiple molecular drivers and
53 disrupted regulatory pathways [1, 2]. The development of large-scale genomics and
54 transcriptomics methods has increased the capacity to identify clinically-relevant
55 tumour subtypes with distinct molecular signatures. These can be used for a better
56 choice of treatment and/or prediction of potential metastasis which can improve
57 survival outcome [3, 4]. However, patients are still facing a high percentage of
58 misdiagnosis in which undetected early metastasis and/or inappropriate choice of
59 treatment can lead to deadly complications with the use of unnecessary severe
60 chemotherapies or the apparition of drug resistance and subsequent tumour relapse
61 [5]. Currently, breast cancer is classified into five major categories (normal-like, luminal
62 A, luminal B, Her2-positive and basal-like) based on expression of three receptors:
63 oestrogen and progesterone hormonal receptors (ER and PR) and the epidermal
64 growth factor receptor ERBB2 (Her2). Basal-like are the most aggressive, and difficult
65 to treat, type of breast cancer tumour. They are usually negative for the three receptors,
66 and thus called triple negative breast cancer (TNBC), which represents 10-20% of all
67 breast cancers. These tumours are usually found in younger patients with a larger size
68 and higher probability of lymph node infiltration and metastasis [2, 6]. Furthermore, the
69 absence of all three receptors reduces the number of targeted therapeutic strategies
70 to be used, leaving nonspecific chemotherapy as the standard treatment of choice,
71 which soon leads to dose-limiting side-effects, resistance to treatment and finally
72 clinical relapse in less than 5 years [6]. A better understanding of the molecular
73 differences in between these tumour categories will improve the choice of treatment
74 and detection of early metastasis, which will significantly impact patient's outcome.
75 There have been many attempts to identify novel therapeutic targets and/or prognostic

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

83 The alternative processing of mRNA precursors enables one gene to produce
84 multiple protein isoforms with different functions, increasing protein diversity and the
85 capacity of a cell to adapt to new environments. An increasing number of splice
86 variants, and their respective splicing regulators, have been shown to confer a
87 selective advantage to tumour cells. For instance, the splicing regulators RBM5, 6 and
88 10 favour tumour cell proliferation and colony formation by regulating the alternative
89 splicing of the membrane-bound protein NUMB [16]. Post-translational activation of the
90 splicing factor SRSF1 (also known as ASF/SF2) confers resistance to apoptosis by
91 inducing inclusion of the anti-apoptotic splice variant in a network of functionally related
92 genes, such as *Bcl-X* and *Mcl1* [17]. Regulation of VEGF splicing is detrimental for
93 stimulation of angiogenesis [18]. A change in the alternative splicing of the pyruvate
94 kinase pre-mRNA can switch tumour cells metabolism to adapt to the increased
95 proliferation [19, 20]. Finally, a list of well-known alternatively spliced variants related
96 to cell adhesion (CTNND1, CD44) and cytoskeleton organisation (ENAH, FLNB) are
97 responsible for the acquisition of migratory and invasive phenotypes necessary for
98 distal metastasis [13, 21–24]. The existence of functionally relevant cancer specific
99 isoforms is therefore a promising new source of highly specific and less toxic

100 therapeutic targets for the development of isoform-specific antibodies and/or splice-
101 switching antisense oligonucleotides [25, 26].

102 By taking advantage of an extensive transcriptomics and anti-tumour compound
103 screening information publicly available in cancer cell lines from the Cancer Cell Line
104 Encyclopedia (CCLE) [27], we identified a splicing signature that can stratify basal
105 breast cancer cell lines into two well-known subtypes, basal A and basal B. In contrast
106 to basal-like breast cancer patients, basal breast cancer cell lines are divided into two
107 subgroups, basal A and basal B, depending on the expression profile of a subset of
108 basal (cytokeratins, integrins), stem cell (CD44, CD24) and mesenchymal markers
109 (Vimentin, fibronectin, MSN, TGFBR2, collagens, proteases) [28–30]. Basal B cell lines
110 are mostly triple negative breast cancer cells that express classical mesenchymal and
111 stem cell markers characteristic of the epithelial-to-mesenchymal transition (EMT), a
112 biological process in which epithelial cells acquire mesenchymal features that are
113 advantageous for the cancer cell, such as increased cell motility to invade distal organs
114 in metastasis, resistance to apoptosis, refractory responses to chemotherapy and
115 immunotherapy, and acquisition of stem cell-like properties like in cancer stem cells
116 [31, 32]. In concordance, basal B cells are morphologically less differentiated, with a
117 mesenchymal-like shape, and a more invasive phenotype in culture assays than basal
118 A and luminal cells [28, 33, 34]. We aimed to transfer this basal A/basal B splicing
119 classification into the clinic by using a semi-supervised machine learning approach.
120 We successfully classified 40% of basal-like breast cancer patients (75/188) from the
121 Cancer Genome Atlas (TCGA) [35] as basal B-like based on a unique 25 spliced gene
122 signature characteristic of cells undergoing EMT. In this signature, we found well-
123 known markers of malignancy, such as ENAH EMT splice variant that promotes lung
124 metastasis [36] or CSF1 variant which promotes macrophage infiltration and distal

125 metastasis [37], together with new promising splicing candidates of tumour progression
126 and invasiveness (PLOD2, CTNND1, SPAG9). Finally, expression of this basal B
127 signature was sufficient to identify triple negative breast cancer tumours with poor
128 survival, highlighting the prognostic value of the newly identified splicing biomarkers to
129 subclassify one of the most heterogenous and difficult to treat type of breast cancer.
130 More studies in cell lines, particularly regarding resistance to treatment and cell
131 invasion will be essential to refine this splicing signature in view of orienting treatment
132 or predicting metastasis sites.

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

144

145 **RESULTS:**

146 **A distinctive Basal B-like breast cancer splicing signature.**

147 Data mining of large-scale genomics and transcriptomics datasets in breast cancer cell
148 lines are a promising source of novel biomarker and therapeutic targets [23, 38, 39].

149 We sought to leverage the wealth of transcriptomics and functional data available in
150 cancer cell lines to better understand different profiles of breast cancer. Hierarchical
151 clustering of changes in alternative splicing of cassette exons and gene expression
152 profile of 80 breast cancer cell lines from two extensive and complementary projects
153 (Additional File 2: Table S1) revealed basal B cell lines as a distinctive group of cells
154 with an expression and splicing profile significantly different from basal A and luminal
155 cancer cells (Additional File 1: Fig.S1). To identify the transcriptional signature
156 characteristic of basal B cells, we repeated the hierarchical clustering in just basal A
157 and basal B cell lines to merge all the differentially expressed and spliced transcripts
158 responsible for the segregation of basal B cell lines (Fig.1). We found 635 genes and
159 217 spliced isoforms with significantly different levels between basal A and basal B
160 cells (Fig.1a,b). In line with published tissue-specific and EMT transcriptomics
161 analyses [40–42], most of the genes differentially spliced were not affected at the
162 expression level, suggesting that two different subsets of genes, and thus regulatory
163 layers, are responsible for the basal B phenotype (Fig.1c). Gene set enrichment
164 analysis (GSEA) [43] between basal B and basal A cells confirmed the EMT and stem
165 cell-like phenotype characteristic of basal B cell lines (Fig.2a,b), which was supported
166 with a higher CD44+/CD24- stem cell score (Fig.2e) [28–30]. DAVID gene ontology
167 analysis of differentially expressed and spliced genes also underlined biological terms
168 that are hallmarks of EMT and cell invasiveness, such as cell-cell junction (Fig.2d) [44].
169 However differentially expressed genes were also enriched in their own unique terms,
170 related to extracellular vesicles/plasma membrane organization. While differentially
171 spliced genes were specifically enriched in terms related to GTPase activity,
172 cytoskeletal protein and cadherin binding, which reinforces the existence of two
173 complementary regulatory pathways (Fig.2d). Finally, another malignant characteristic

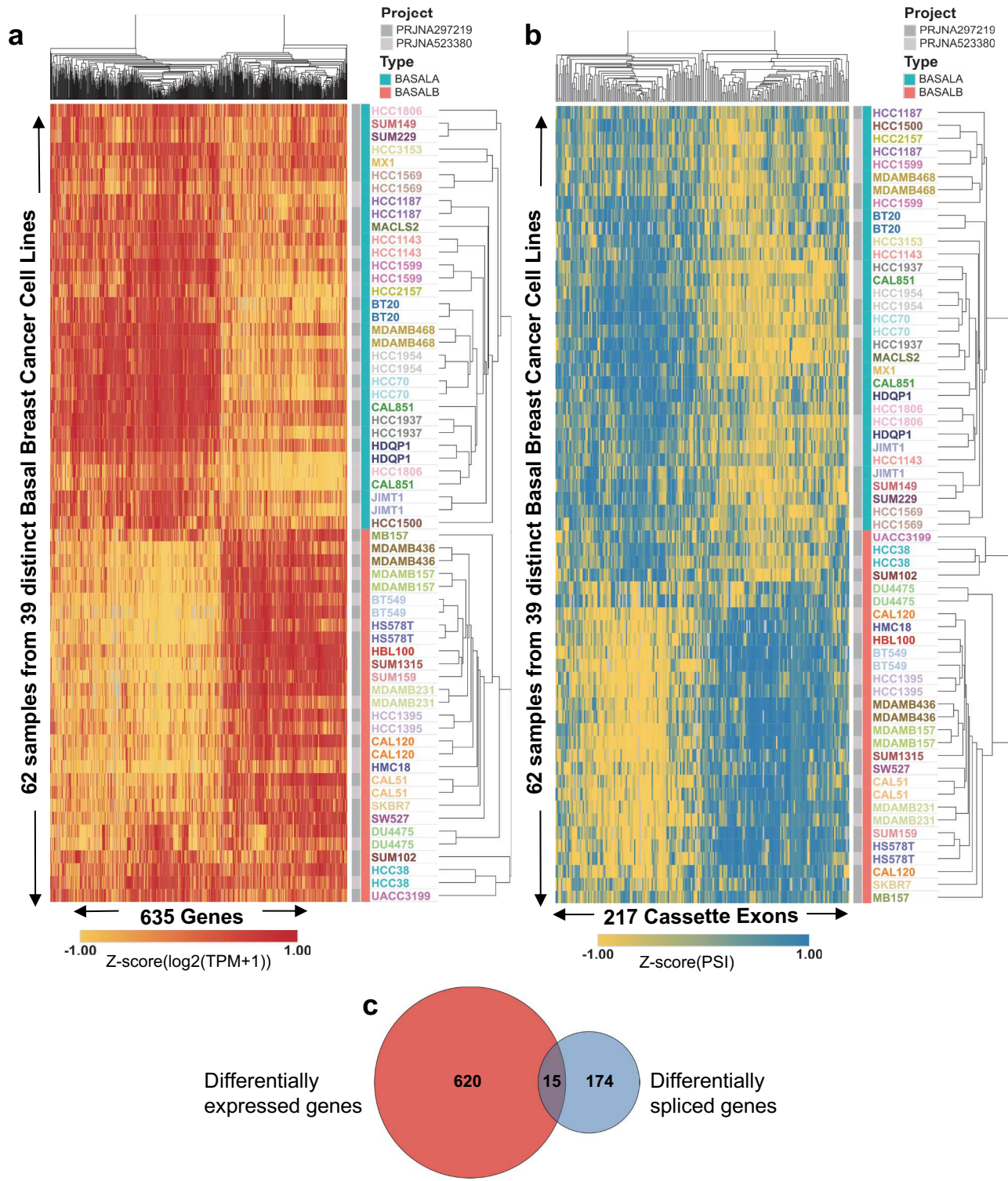


FIGURE 1

174 acquired by cancer cells undergoing EMT is resistance to chemotherapy, which often
175 leads to clinical relapse. Gene set enrichment analysis found upregulation of genes
176 resistant to the Epidermal Growth Factor Receptor (EGFR) inhibitor Gefitinib (Fig.2c),
177 which is an alternative to hormonal therapy in Her2+ breast cancer tumours, but is not
178 efficient in triple negative tumours [45]. Available drug assays from the Genome Drug
179 Sensitivity in Cancer portal (GDSC) [46] confirmed the need of a higher concentration
180 (IC50) of Gefitinib, and other EGFR inhibitors (Erlotinib, Sapitinib), to have the same
181 deleterious effect on basal B compared to basal A cancer cells (Fig.2f). Basal B cell
182 lines also showed a significant resistance to well-known inhibitors of the cell cycle
183 (Irinotecan, Taselisib, 5-Fluorouracil), drug inducers of cell death (AZD5582,
184 AZD5991) and other receptor tyrosine kinase inhibitors, such as Savolitinib which
185 inhibits c-MET to reduce tumour persistence and metastasis [47].

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

192

193 **A semi-supervised machine learning approach to subclassify basal-like breast** 194 **cancer patients.**

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

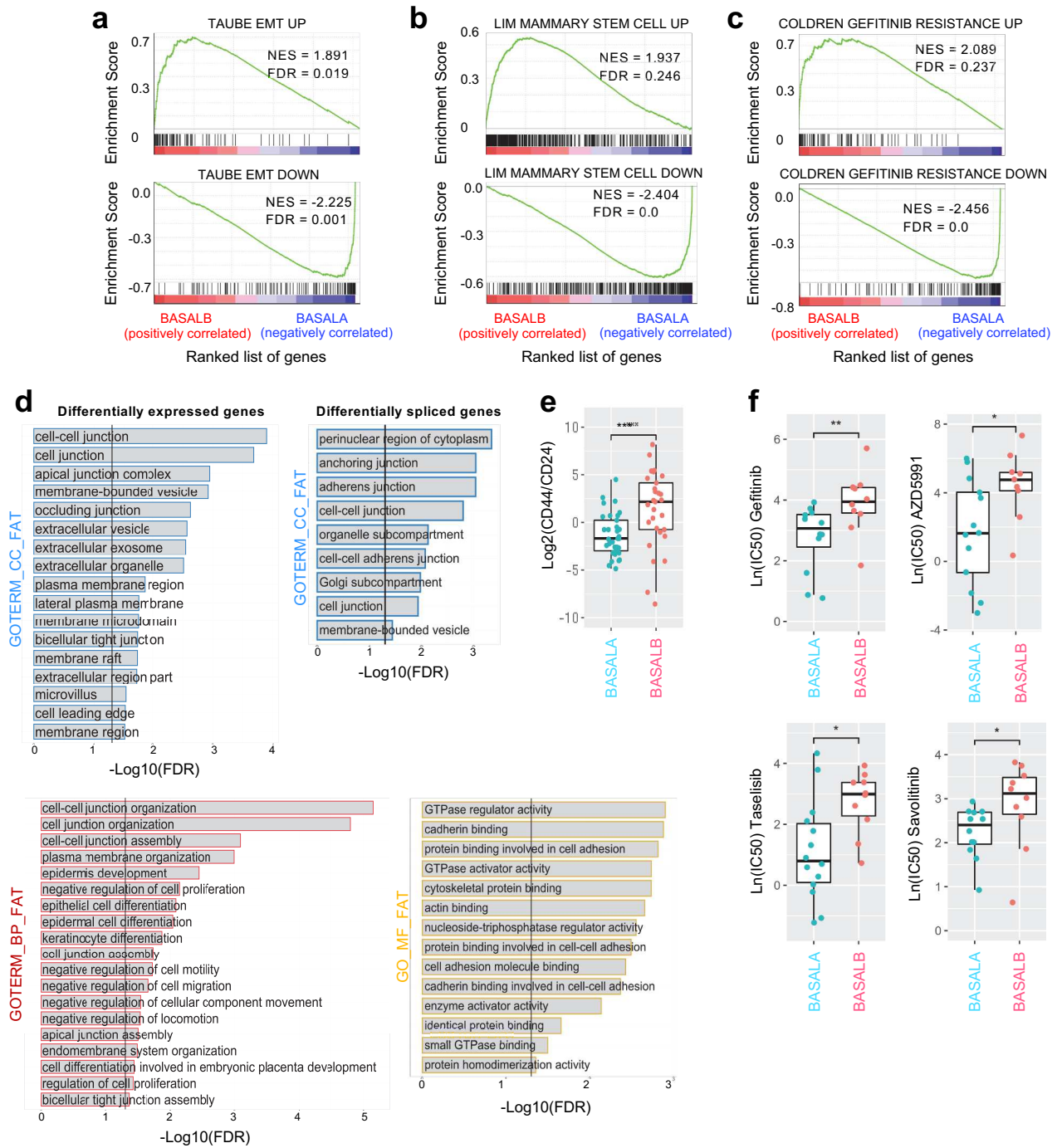


FIGURE 2

199 basal B cell lines (Additional File 1: Fig. S2a,b). Using such method, patients were
200 forced to classify in one of the two groups based on differences in gene expression or
201 splicing patterns. Since basal B cell lines show more invasive, cancer stem cell-like
202 phenotypes, we assessed whether these aggressive characteristics were translated to
203 the “B-like” patient group through differences in disease specific survival (DSS) rates.
204 Kaplan-Meier analysis of DSS did not show significant differences between the two
205 subgroups of basal-like patients (Additional File 1: Fig. S2c,d). However, we did
206 observe a tendency for “B-like” patients to have a poor survival compared to “A-like”
207 when just looking at differences in splicing, contrary to expression levels (p -value=0.09
208 vs 0.57, respectively – Additional File 1: Fig. S2c,d).

209 In fact, it was not surprising that the transcript-level and splicing signatures did
210 not translate directly from simplistic cell culture models to much more complex tumour
211 patients with specific cell micro-environments and differences in cell heterogeneity.
212 However, because the patients showed clear “A-like” and “B-like” signatures, we
213 sought to develop a machine learning approach that would allow us to transfer part of
214 the molecular and phenotypic observations found in cell-lines to patient data. Transfer
215 learning is a recent research methodology that focuses on storing the knowledge
216 gained when solving a problem, to apply it to a different, but related, one. Because we
217 wanted to ensure that the newly developed cell-to-patient transfer learning algorithm
218 could create interpretable models, we used a decision tree-based approach called
219 Random Forest. In this cell-to-patient random forest classification method, we started
220 by classifying basal A or basal B cell-lines based on their splicing and/or expression
221 profile (Fig. 3a and Additional File 1: Fig.S3-S4). Then, once the model was trained on
222 cell-lines, we would start integrating patient data gradually into the model. This was
223 done iteratively by integrating at each round of classification the patients best predicted

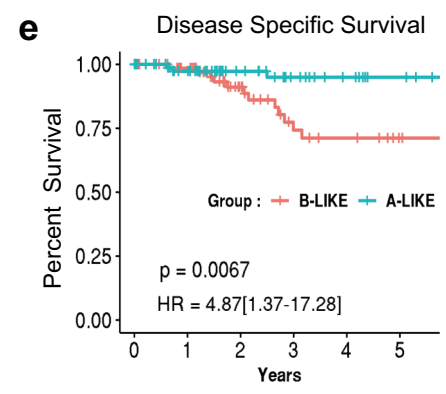
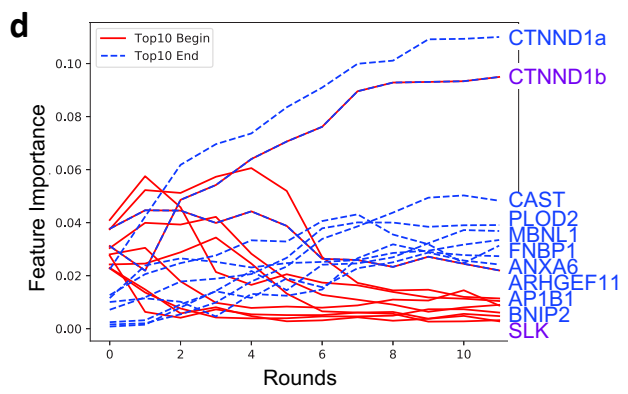
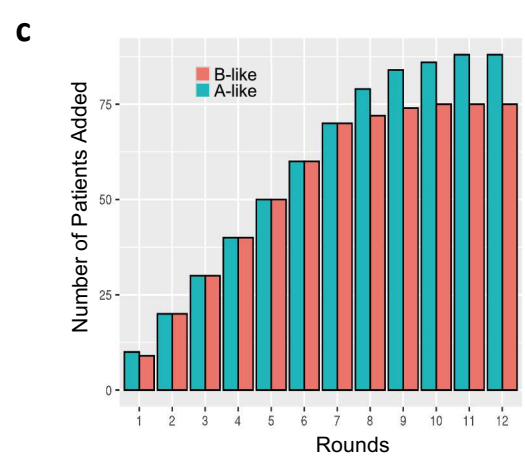
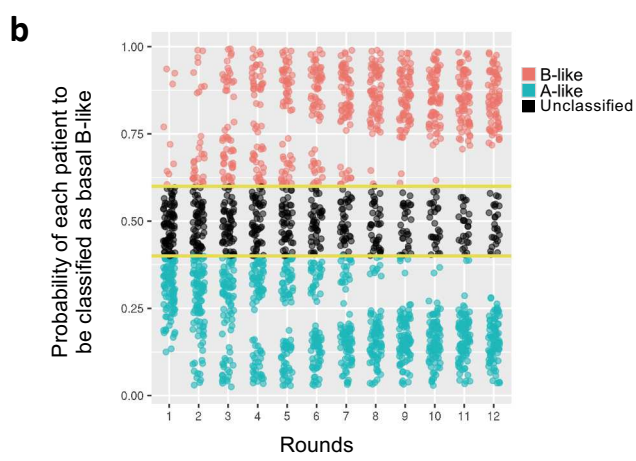
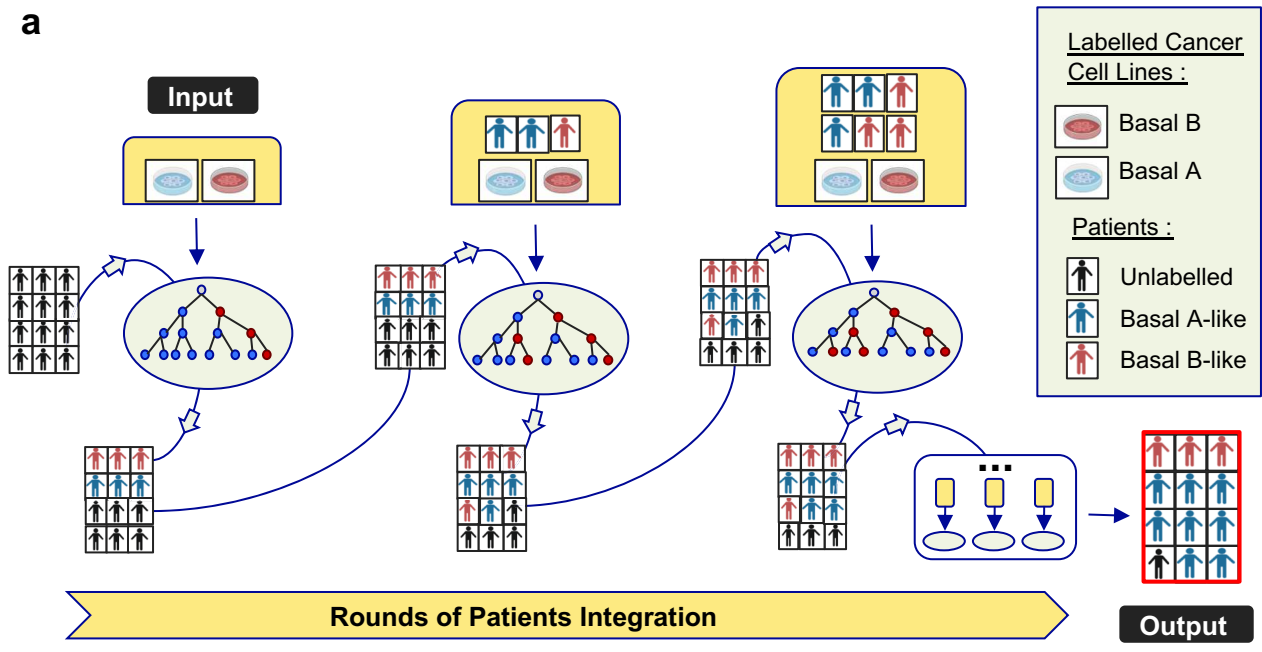


FIGURE 3

224 to be basal A-like and basal B-like, so their added informative value could be used
225 back to train the system and improve the next round of classification (Fig.3a). With this
226 semi-supervised approach, the probability of assigning a patient to a specific subgroup
227 evolves and improves at each round based on the updated information obtained from
228 the best predicted patients, reaching at the end a stable population with the labels
229 'basal A-like', 'basal B-like' or 'unclassified' determined by the algorithm after 10-12
230 rounds (Fig. 3b,c and Additional File 1: Fig.S3b,c-S4b,c). Thanks to the gradual
231 addition of patients at each round of training, there is a progressive increase, or
232 decrease, in the feature importance of the splicing variants used to classify patients
233 (Fig.3d and Additional File 1: Fig.S3d-S4d). Out of the 188 basal-like patients, 75 were
234 classified as basal B-like, 88 as basal A-like and 25 could not be classified based on
235 their splicing signature. Using only expression levels, there was a slight bias towards
236 the basal A-like phenotype, with 56 patients classified as basal B-like, 122 as basal A-
237 like and 10 unclassified (Additional File 1: Fig.S3b-c). Combining differentially spliced
238 and expressed features seemed to be the most performant classifier with 84 patients
239 as basal B-like, 100 as basal A-like and just 4 unclassified (Additional File 1: Fig.S4b-
240 c). Taken together, depending on the features used (splicing patterns, expression
241 levels or both), patients were differently classified in basal A-like or basal B-like.

242

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

244 To address which classifier translates the best to patients the invasive, EMT-
245 like and drug resistant basal B phenotype found in cancer cells, we calculated the 5-
246 year survival rate for each group of basal A-like and basal B-like issued from the three
247 types of classification. Only basal B-like patients classified based on splicing levels had
248 a poor prognosis compared to basal A-like patients (log-rank test $p = 0.0067$, HR =

249 4.87; IC95%: [1.37-17.28] in Kaplan-Meier analysis and univariate Cox regression)
250 (Fig.3e). Basal B-like patients subclassified based on gene expression levels, or gene
251 expression and splicing features, did not show significant differences in disease
252 survival rate (Additional File 1: Fig.S3e-4e), suggesting that splicing biomarkers might
253 be more informative to further subclassify basal-like patients based on prognosis. We
254 thus decided to focus on the role of alternative splicing in identify triple negative basal-
255 like breast cancer with poor prognosis.

256 To extract the most informative splicing features from the cell-to-patient transfer
257 learning classifier, we used the Boruta feature selection method [49]. This allowed us
258 to select the key splicing events responsible for the basal A/B classification without the
259 need to predefine arbitrary thresholds (Fig.4a). Out of the 217 differentially spliced
260 exons between basal A/B cell lines, just 25 were needed to subclassify breast cancer
261 patients in basal A or basal B-like tumours (Fig.4a and Additional File 2: Table S2).
262 Sashimi plots representing the splicing patterns of some of these basal B-specific
263 splicing events, such as the well-known splicing biomarker of cancer metastasis ENAH
264 [26] and the newly identified splicing biomarkers PLOD2, SPAG9 and KIF13a,
265 validated the observed changes in splicing between basal A and basal B-like patients
266 (Fig.4b-c and Additional File 1: Fig. S5a-b). Moreover, the changes in percentage of
267 spliced-in (PSI) of the 25 basal B-specific splicing events between the two subtypes of
268 basal-like patients correlated with the observed splicing changes between basal A/B
269 cell lines (Additional File 1: Fig.S5c-d), further supporting the transfer of knowledge
270 from the laboratory to the clinic. Finally, in the absence of publicly available RNA-seq
271 data on a second cohort of basal-like breast cancer patients, we took advantage of
272 three independent sequencing projects on breast cancer cell lines, different from the
273 ones used for the training of the semi-supervised classifier (Additional File 2: Table

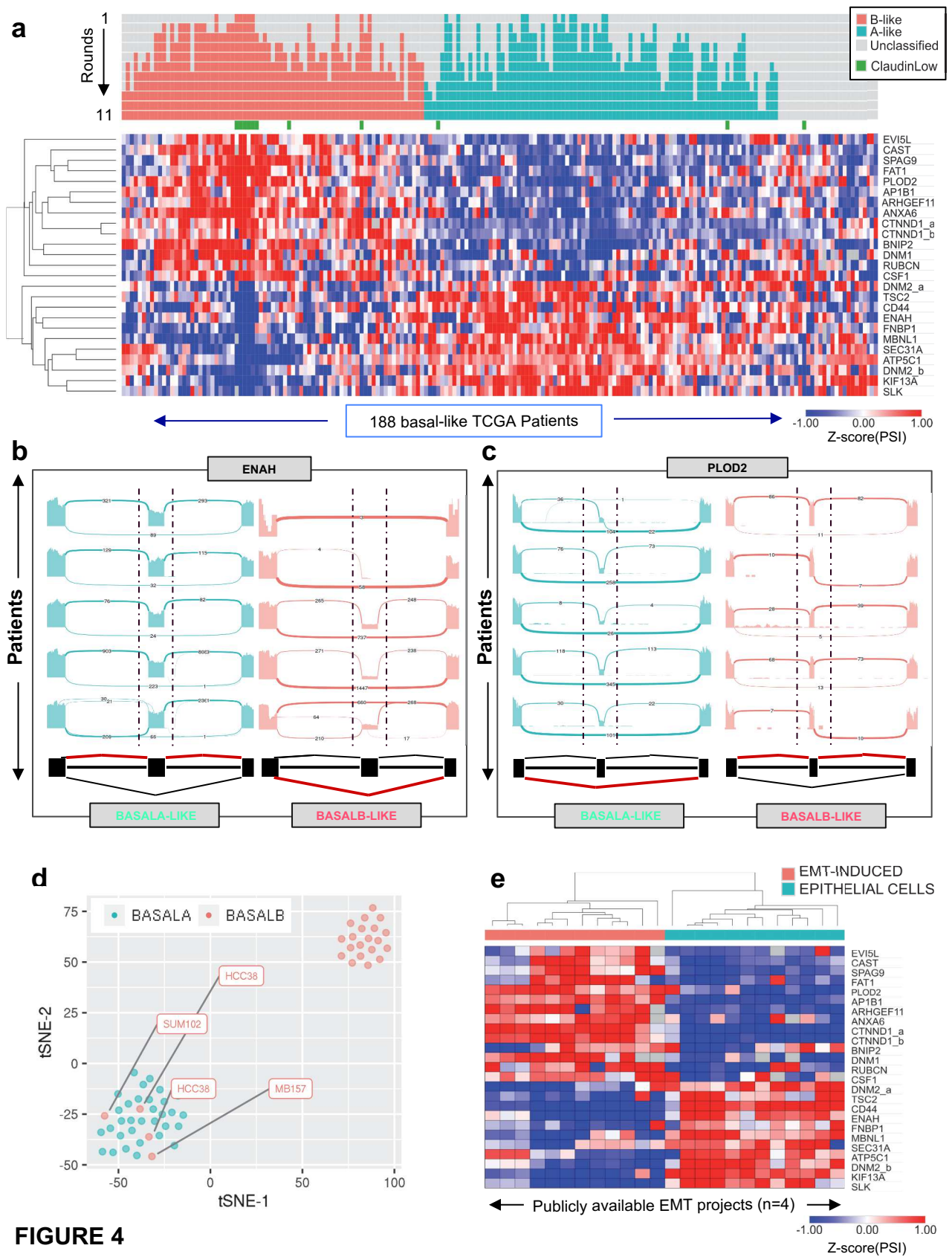


FIGURE 4

274 S1). Distribution of 52 independent breast cancer cell lines showed a 93% accuracy in
275 the spatial segregation (t-SNE) of basal A from basal B cells based on the splicing
276 pattern of the 25 newly identified splicing events (Fig.4d). Just three cell lines were
277 misclassified as basal A (HCC38, SUM102 and MDA-MB-157). It is worth noting that
278 one of these, HCC38, was also labelled as basal A in the DepMap portal
279 (www.depmap.org), which validated our methodology and the specificity of the splicing
280 signature towards a basal B-like phenotype.

281 Consistent with basal B cell lines being more mesenchymal, differences in the
282 alternative splicing of these 25 basal B-specific splicing events in four different cellular
283 models of EMT, coming from different cell types and methods of EMT induction [50–
284 53], successfully clustered epithelial cells from mesenchymal with a pattern of splicing
285 equivalent to basal A and basal B-like patients, respectively (Fig.4e). Of note, another
286 25 gene-based EMT-like splicing signature characteristic of luminal breast cancer
287 tumours has also been identified capable of subclassifying mesenchymal-like breast
288 cancer tumours with poor prognosis [38]. Consistent with a more luminal-specific
289 signature, despite both marking EMT phenotypes, not more than six splicing events
290 were found in common between the two splicing signatures (ATP5C1, CTNND1,
291 KIF13a, PLOD2, SEC31a and SPAG9), which further supports the specificity of our
292 newly identified splicing signature for basal-like triple negative breast cancer. Finally,
293 using one of the first established molecular subtypes of triple negative breast cancer
294 tumours based on gene expression, which is the Lehman classification [54], we found
295 that basal B-like patients are mostly found in the categories associated with
296 Mesenchymal stem-like (MSL) and Immunomodulatory (IM) subtypes (Fig.5a), which
297 goes in line with a gene set enrichment of terms related to inflammatory responses and
298 hallmark of EMT (Fig.5b).

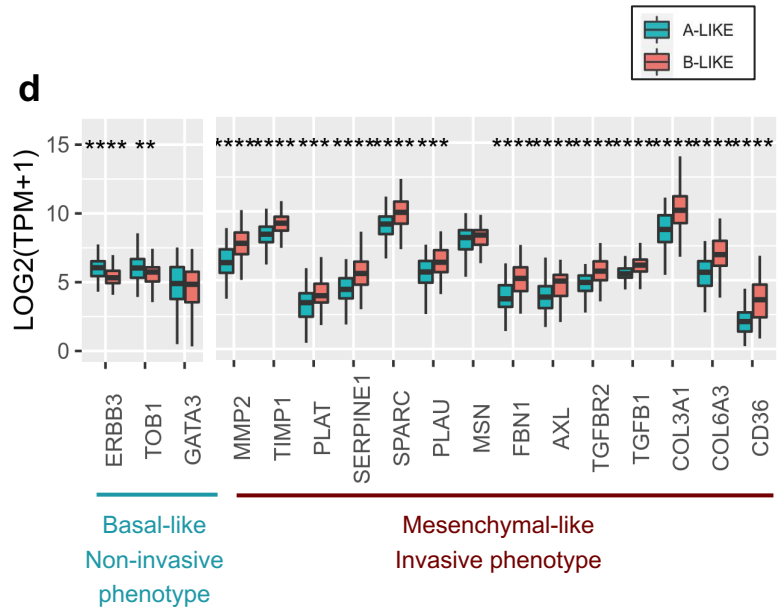
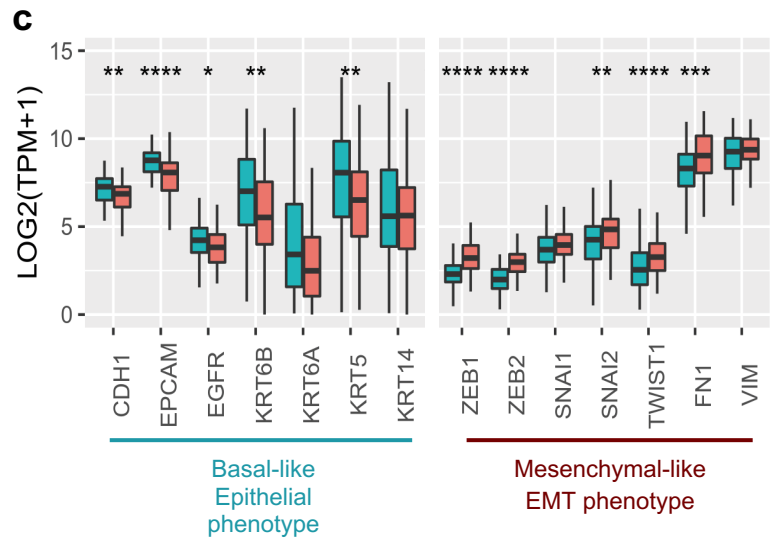
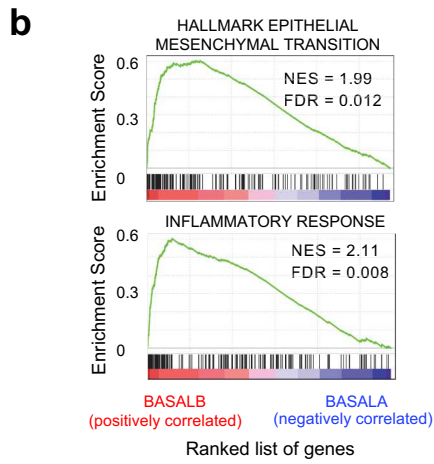
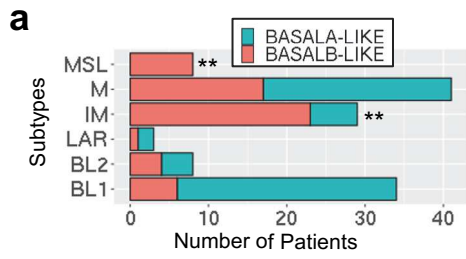


FIGURE 5

299 When looking at the expression of well-known basal and EMT biomarkers in the
300 two subpopulations of basal A/B-like patients, we found that basal A-like patients
301 express classical basal/epithelial markers, such as E-cadherin, EPCAM and
302 cytokeratin KRT5/KRT6/KRT14, together with ERBB3 and TOB1 which are markers of
303 more differentiated, non-invasive cells [2]. On the other hand, basal B-like patients
304 express classical EMT/mesenchymal markers such as Fibronectin, the EMT inducers
305 Twist and Slug, and the Zinc-finger transcriptional regulators Zeb1 and Zeb2 which
306 have recently been shown to confer stemness properties that can increase the
307 plasticity and invasive capacity of the tumour cells [55] (Fig.5c-d). In line with a more
308 aggressive, invasive phenotype, basal B-like patients express cytoskeletal (MSN, FN1)
309 and extracellular matrix signalling proteins (TGFB1, TGFBR2, FBN1, AXL), collagens
310 (COL3A1, COL6A3) and proteases (MMP2, TIMP1, CTSC, PLAU, SERPINE1/2,
311 PLAT), which are necessary for cell's migration and dissemination to distal organs
312 during metastasis [2]. Finally, basal B-like patients overexpress a recently identified
313 new marker of metastasis-initiating cells, the fatty acid receptor CD36 [20]. Clinically,
314 the presence of CD36 positive cells has been correlated with a lower survival rate in
315 many carcinomas, including breast cancer, and inhibition of CD36 impairs metastasis
316 in breast cancer-derived tumours, turning this receptor into an important biomarker of
317 tumour cell dissemination and a potential new target to reduce cell invasion. The fact
318 that basal B-like tumour cells co-express this metastasis-initiating marker further
319 strengthens the aggressive nature of this tumour subclass and the clinical relevance
320 of the basal B-specific splicing signature in tumour progression and relapse.

321 Overall, we have identified a novel splicing signature, specific of triple negative
322 breast cancer tumours, that marks patients with the poorest prognosis. This basal B-
323 like splicing signature is responsible of a stem-like, EMT phenotype that favours tumour

324 growth, invasion of distal organs and increased drug resistance, which eventually leads
325 to tumour relapse and metastasis. Interestingly, some of the genes differentially
326 expressed in this basal B-like patients are well-known markers of metastasis-initiating
327 cells, such as the alternatively spliced CTNND1 and PLOD2 genes or the fatty acid
328 receptor CD36, turning these biomarkers into promising new targets for innovative
329 therapies, such as the use of splicing specific antibodies [6, 26].

330

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

333 Hierarchical clustering of basal A and B cell lines based on the differential
334 expression of RNA-binding proteins highlighted six RNA regulators, ESRP1, ESRP2,
335 RBM47, TMEM63A, KRR1 and RBMS3 (Fig.6a) (Kruskal-Wallis $p < 10^{-9}$). Interestingly,
336 ESRP1/2 and RBM47 are significantly less expressed in basal B-like than basal A-like
337 patients (Fig.6b), consistently with the known inhibitory effect of these three splicing
338 regulators in EMT progression and metastasis [53, 56, 57]. Available transcriptomics
339 data in ESRP1/2 and RBM47 lung carcinoma NCI-H358-depleted cells [53] and
340 RBM47 overexpressing breast cancer metastatic MDA-MB-231 cells [58] showed that
341 19 of the 25 splicing events responsible for the newly identified basal B-specific splicing
342 signature could potentially be regulated by ESRP1/2 and/or RBM47 in breast cancer
343 cells (Fig.6c-d). Importantly, in the cell types analysed, ESRP1/2 and RBM47 induce
344 the epithelial, basal A-like splicing phenotype, suggesting a potential tumour
345 suppressor effect for these splicing regulators (Fig.6e-g, 4e and Additional File 1: S5c-
346 d). Consistently with this observation, low expression of RBM47 in basal-like breast
347 cancer patients was associated with poor overall survival (log rank test $p=0.031$,
348 $HR=3.36$, $IC95\%:[1.05 - 10.79]$ - Fig.6h-i), which supports previous experimental

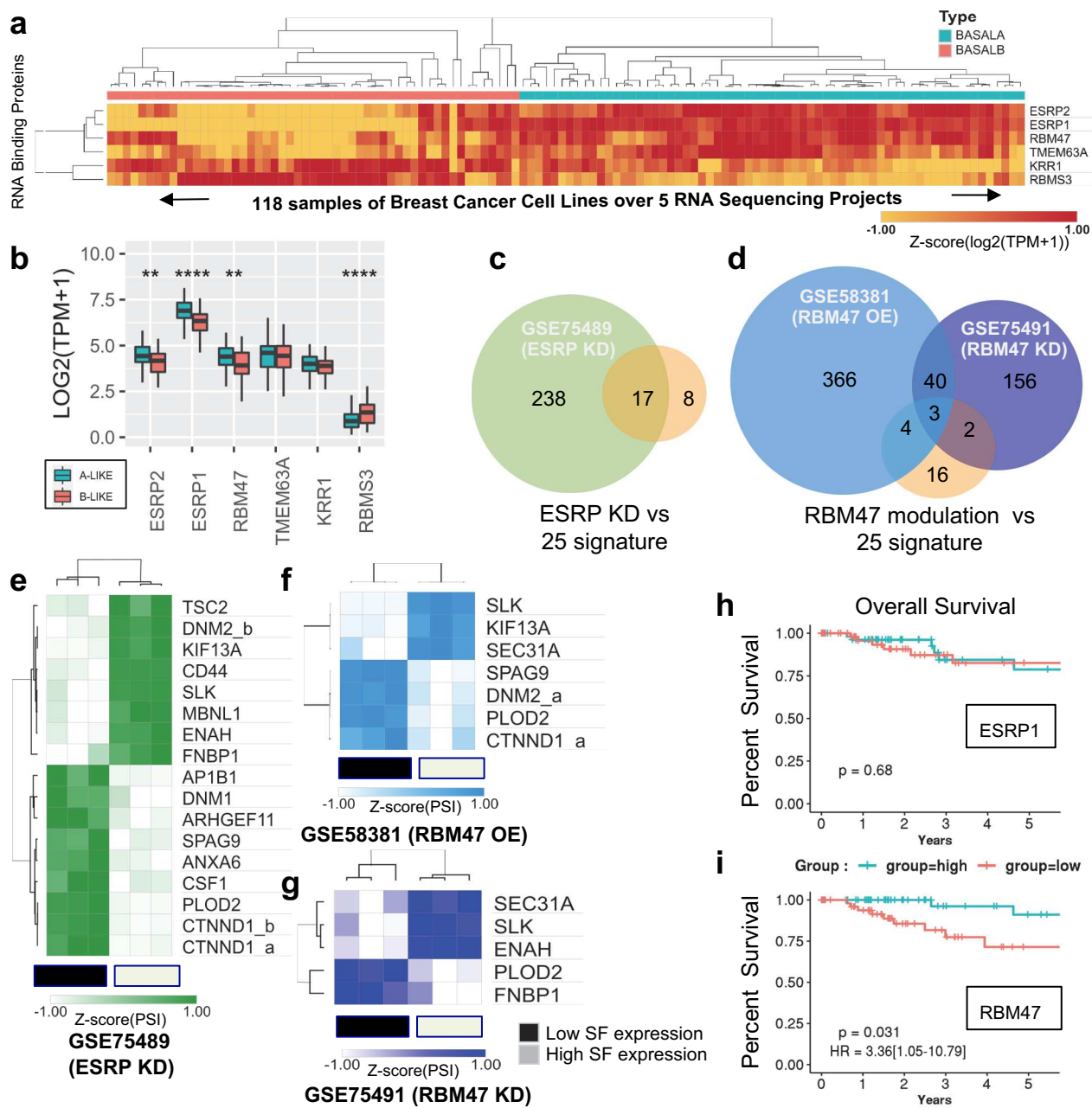


FIGURE 6

349 evidence of a role for RBM47 in suppressing breast cancer metastasis and progression
350 [57]. In fact, RBM47-dependent basal B-specific splicing events were found to be
351 functionally interconnected by physical and/or genetic interactions, which points to the
352 existence of a common basal B-specific regulatory network associated with tumour
353 malignancy (Additional File 1: Fig. S6a). In support, most of RBM47-dependent basal
354 B-specific splicing events play well-known roles in cell-cell adhesion (CTNND1) [59],
355 cytoskeleton organization (ENAH, SLK, FNBP1) [60, 61], endocytosis (KIF13A, DNM2)
356 [62] and association with the extracellular matrix (PLOD2) [63], which are all key
357 processes for gaining the cell motility and invasiveness necessary in tumour
358 metastasis (54-58). Of note, expression of just one of these basal B-specific splice
359 variants, which are CTNND1, ENAH and PLOD2, is sufficient to lower the disease-
360 specific survival rate of basal B-like breast cancer patients compared to basal A-like
361 (Additional File 1: Fig.S6b-g). These splicing events could turn into promising new
362 therapeutic strategies aiming at specific key regulatory genes instead of a pleiotropic
363 splicing regulator that could have unsuspected secondary effects.

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

373

374 **DISCUSSION:**

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

391 Large scale public molecular data sets on genomics (copy number and
392 mutation), epigenomics, transcriptomics, proteomics, *in vitro* and *in vivo* cell
393 invasiveness and response to anti-tumour compounds in a large number of patients
394 (11,000 patients across 33 different tumour types from the Genome Cancer Atlas) and
395 human-derived cell lines (1000 cancer cell lines across 36 tumour types from the Broad
396 Institute's Cancer Cell Line Encyclopaedia) has become an extraordinary toolbox to
397 identify novel prognostic markers of early metastasis and/or resistance to specific
398 drugs, which are the two major reasons for clinical relapse and low survival rate [68–

399 70]. Unfortunately, the translatability of these pre-clinical findings is often limited since
400 culture cells are not representative of the variety of individuals nor the biological reality
401 of the tumour's multicellular environment. Yet, culture procedures are improving with
402 the creation of organoids, and machine learning approaches combined with large-scale
403 data mining are bypassing some of these important caveats. This is the case of our
404 cell-to-patient random forest classifier approach, in which the addition at each round of
405 selection of novel informative features, based on the patients classified in previous
406 rounds, allows an algorithm to make use of the information learned from cell lines.
407 Thanks to this approach, we were able to identify the first splicing signature, composed
408 of 25 alternatively spliced exons, capable of subclassifying basal-like breast cancer
409 patients into two subtypes with different prognoses: basal A- and basal B-like.

410 Actually, this newly identified basal B-like splicing signature underlined a stem-
411 cell like EMT signature, with hallmarks of cell invasiveness and drug resistance. Five
412 of these 25 alternatively spliced genes are well-known to play a role in cancer
413 (ARHGEF11, CD44, CTNND1, ENAH, MBNL1) [74–76]. Six have been indirectly
414 linked to tumour malignancy and are thus new splicing targets to study (CAST, CSF1,
415 PLOD2, SLK, SPAG9, TSC2) [61, 63, 77–80]. The rest are completely unknown for
416 their splicing role in cancer, even though changes in expression of some of them have
417 been shown to play a role in tumour progression, chemosensitivity and metastasis
418 without specifically addressing which splice variant (ATP5C1, BNIP2, FAT1, FNBP1,
419 SEC31A, ANXA6, DNM1, DNM2) [62, 81]. Of special interest are ARHGEF11 and
420 CTNND1 splice variants. Both proteins are involved in cell-cell adhesion and the basal
421 B-specific splice variants promote cell migration and invasiveness in several cancer
422 types, such as breast cancer (13,54,74,67). Moreover, depletion of ARHGEF11 in
423 basal breast cancer cells is sufficient to alter cell morphology, which suppresses the

424 cancer cell growth and survival *in vitro* and *in vivo* [75]. On the other hand, the
425 existence of an isoform-specific antibody for CTNND1 pro-invasive splice variants
426 turns this splicing candidate as a valuable new target to reduce tumour metastasis [82].
427 ENAH and CD44 are amongst the most studied splicing events impacting cancer and
428 are well-known biomarkers of poor prognosis. ENAH's inhibition decreases metastasis
429 by slowing down tumour progression and reducing cell invasion and intravasation [83–
430 85]. While the change to basal B splicing signature of CD44, a transmembrane protein
431 that maintains tissue structure, is sufficient to drive an EMT and to increase cell
432 invasion and plasticity by promoting stem cell characteristics [22, 86]. Interestingly,
433 MBNL1 splicing regulation has also been involved in pluripotent stem cell
434 differentiation [87] and cell viability via inhibition of DNA damage response [88].
435 Promising new splice variants with a potential link with cancer are CSF1, PLOD2, SLK,
436 SPAG9 and TSC2. CSF1 is a macrophage marker which splice variant could correlate
437 with infiltration of tumour-promoting macrophages [77, 89]. Changes in the alternative
438 splicing of the procollagen-lysine PLOD2, which catalyses the deposition and cross-
439 link of collagens in the extracellular matrix, have been intimately linked to EMT
440 progression and cervical, breast, lung, colon and rectal cancer prognosis [40, 90]. Its
441 inhibition reduced proliferation, migration and invasion of cancer cells, while its
442 overexpression promoted cancer stem cell properties and resistance to drugs [63, 91].
443 SLK was identified as a prognostic biomarker in several cancers and is necessary for
444 the induction of cell migration and invasion during EMT [61, 76, 92]. SPAG9 is a
445 scaffold protein that organizes mitogen-activated protein kinases and has been
446 associated with invasion in several types of tumours and prognosis [79, 93, 94]. Finally
447 TSC2 basal B-specific splicing isoform cannot be phosphorylated by AKT, which leads
448 to a continuously activated mTOR pathway and oncogenic autophagy [78]. More

449 functional studies on the impact of each of these cassette exons splice variants in
450 cancer will increase our knowledge on tumour progression and metastasis with the
451 long term goal of improving diagnostics and treatment. Of note, other types of splicing
452 events, different from the studied cassette exons, have also been shown to play
453 important roles in tumorigenesis, such as alternative splice sites and intron retention
454 [71–73]. It is necessary to extend this type of approaches to all types of splicing events
455 and validate them using independent cohorts of patients. The increase of accessible
456 sequencing data in primary tumours will thus be essential to continue with this type of
457 approaches.

458 Finally, it is interesting to note that these 25 alternatively spliced exons are
459 basically dependent on three well-known splicing regulators, ESRP1/2 and RBM47,
460 which are intimately linked to EMT and metastasis. ESRP1 is the major regulator of a
461 newly identified epithelial-specific splicing signature [53]. Its expression in cancer cells
462 promotes tumour growth and a mesenchymal-to-epithelial transition which are
463 essential for the formation of new tumours at distal organs during metastasis [95, 96].
464 RBM47 is a newly identified splicing regulator of EMT that has also been associated
465 with metastasis [57, 97, 98] . Through integrative analysis of clinical breast cancer
466 gene expression datasets, cell line models and mutation data from cancer genome
467 resequencing studies, RBM47 was identified as a suppressor of breast cancer
468 progression and metastasis. It was found mutated in patients with brain metastasis and
469 its expression was necessary to inhibit brain and lung metastatic progression *in vivo*
470 [57]. Interestingly, despite regulating just 9/25 splicing events of the basal B-specific
471 splicing signature, low expression of RBM47, and not ESRP1, correlated with a poor
472 prognosis and lower survival rate in basal-like breast cancer patients, which increases
473 the interest to design new therapies targeting this splicing regulator.

474 In fact, this basal B-specific splicing signature has highlighted a subpopulation
475 of basal-like triple negative breast cancer patients differentially expressing several
476 hallmarks of invasive, EMT-like aggressive cancer, such as the newly identified
477 biomarker of metastasis CD36 [20]. CD36 is a fatty receptor expressed in metastasis-
478 initiating cells. Neutralizing antibodies that block CD36 completely inhibited the
479 formation of metastasis in orthotopic mouse models of human oral cancer, and CD36
480 inhibition impaired metastasis in human melanoma and breast cancer-derived
481 tumours. Interestingly, the fatty acid-binding protein 7 (FABP7) correlates with a higher
482 incidence of brain metastasis and lower survival rate in breast cancer patients, which
483 all together points to a potential connection between fatty acid metabolism and
484 metastasis in our subclass of basal-like breast cancer patients [99]. Furthermore, cells
485 expressing our newly identified basal B-specific splicing signature also showed
486 resistance to several EGFR inhibiting drugs. Therapies targeting EGFR have variable
487 and unpredictable responses in breast cancer [100]. By better subclassifying sensitive
488 from resistant tumour cells, diagnoses could be improved, which will impact the choice
489 of treatment and thus the chances of tumour relapse. Extensive drug screening of cells
490 derived from basal B-like patients combined with machine learning strategies to
491 transfer the splicing knowledge obtained will certainly improve the identification of
492 much more suitable treatments for triple-negative breast cancer cells and reduce
493 tumour relapse, thus improving the survival rate.

494

495

496 **CONCLUSION:**

497 Taking advantage of extensive available experimental data in breast cancer cell
498 lines, we performed a knowledge transfer to clinical data to identify the first splicing

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

515

516

517 **METHODS**

518 **RNA-seq transcriptomics analysis: gene expression and alternative splicing**

519 RNA-seq reads were aligned to the human genome (GRCh38, primary assembly)
520 using STAR [101] version 2.5.2b with standard parameters. Gencode v25 (derivated
521 from Ensembl v85) was used for all analysis requiring annotation.

522 TPMCalculator [102] (v0.0.1) was used to compute Transcripts Per Million (TPM)
523 values and obtain read counts. Q parameter was set to 255 to keep only unique
524 mapped reads and ExonTPM value was used to consider only reads mapped to exons.
525 Whippet-quant from Whippet software (v10.4) was used to compute Percentage
526 Spliced-In (PSI) values for splicing analysis. Conjointly to Kruskal-Wallis testing, the
527 output from Whippet-quant was further filtered to include only events for which the sum
528 of inclusion counts (IC) and skipping counts (SC) was greater or equal to 10 for both
529 sets of samples. Whippet-delta was used to compute differential splicing (deltaPsi) and
530 probability that there is some change in splicing between conditions. Two heuristic
531 filters were applied on splicing events as advised in whippet documentation; $|\text{deltaPsi}|$
532 > 0.1 and $P(|\text{deltaPsi}| > 0.0) \geq 95\%$ were considered reliable parameters to filter
533 biologically relevant AS events.

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

536

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

542

543 Gene Set Enrichment Analysis (GSEA v20.0.5) was carried out on the GenePattern
544 [105] web platform using phenotype for permutation type and 1000 for number of
545 permutations to execute. FDR cutoff of 25% for potential true positive finding was used

546 as documented in the GSEA user guide. Read counts were previously normalized
547 using DESeq2 [106] (v 1.10.1) on the same Platform.

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

549

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

557

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

559

560 **Machine Learning and feature selection:**

561 First, we construct a classifier to distinguish basal B / A cell lines using a Random
562 Forest with 1000 trees. After, we applied this model to the TCGA patients. Based on
563 Gini impurity, we computed the class probability to predict patient labelled as B-like or
564 A-like. Then, mixing initial cell lines with a subset of patients classified with the more
565 reliability (the ones picked up with higher class probability not passing below a
566 threshold of $P=0.6$), we create a new model. Each addition of patients is called a round,
567 during which a new model is created, giving new predictions (probabilities) for the
568 remaining patients. By limiting the number of new patients added at each round ($10 \times$
569 $n_current_round$) (Fig.3c and Additional File 1: Fig.S3c-4c), the model can gradually
570 learn from the patient data and avoid overfitting. With such conditions, we can observe

571 a gradual shifting in feature importance from the ones informative to classify cell lines
572 to the ones informative to classify patients and cell lines (Fig.3d and Additional File 1:
573 Fig.S3d-4d). The algorithm stops when it can no longer incorporate the patients into
574 one or the other group given the cut-off of $P=0.6$. ML analyse was done with Python
575 3.7.3 based on scikit-learn version 0.21.2.

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

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

589

590 **Breast Cancer Annotation**

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

594 Claudin Low status was defined with script downloaded from
595 <https://github.com/clfougner/ClaudinLow/blob/master/Code/TCGA.r> [110] using

596 dataset from http://download.cbioportal.org/brca_tcga_pan_can_atlas_2018.tar.gz
597 [111, 112].

598

599 **Survival Analysis**

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

607

608 **Statistics**

609 Wilcoxon Rank Sum Test were used to assess statistical significance within boxplots

610 They were noted. $P < 0.05$ (*), $P < 0.01$ (**), and $P < 0.001$ (***), $P < 0.0001$ (****).

611 Kruskal-Wallis Test was used to keep differential features for expression (TPM values)
612 or splicing (PSI values) when Luminal, Basal A & B cell lines were compared and
613 displayed in heatmap figures. A threshold of p-value $<10^{-5}$ was used to filter out
614 potential false positive and reduce the number of features in order to apply hierarchical
615 clustering. This threshold was adapted depending on the number of samples in the
616 comparison. For RNA binding proteins, a higher cut off of $p < 10^{-9}$ was used because
617 5 projects were pulled together.

618

619 **Code**

620 Code and annotation files are available here.

621 https://github.com/LucoLab/Villemin_2020.

622

623 **Ethics approval and consent to participate**

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

626 (<https://www.cancer.gov/about-nci/organization/ccg/research/structural->

627 [genomics/tcga/history/policies](https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/history/policies))

628

629 **Consent for publication**

630 All patients gave consent for publication of their personal information.

631

632 **Availability of data and materials**

633 All datasets are available in the Gene Expression Omnibus (GEO): GSE75489,
634 GSE58381, GSE75491, GSE61220, PRJEB25042, GSE74881, GSE75492,
635 PRJNA523380, PRJNA297219, PRJNA210428, PRJNA251383, PRJEB30617
636 (detailed in Additional File 2: Table S1) and The Cancer Genome Atlas (TCGA)

637 repositories upon request ([https://www.ncbi.nlm.nih.gov/projects/gap/cgi-](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000178.v11.p8)
638 [bin/study.cgi?study_id=phs000178.v11.p8](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000178.v11.p8))

639

640 **Competing interests**

641 The authors declare no competing interests.

642

643 **Funding**

644 Luco team is supported by the Agence Nationale de la Recherche [ANRJCJC - 2016 -
645 EpiSplicing] and the Labex EpiGenMed [ANR-10-LABX-12-01]. Ritchie team is

646 supported by the Agence Nationale de la Recherche [ANRJCJC - WIRED], the Labex
647 EpiGenMed [ANR-10-LABX-12-01] and the MUSE initiative [GECKO].

648

649 **Authors' contributions**

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

654

655 **Acknowledgements**

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

657

658 **List of abbreviations**

659 AS: Alternative Splicing

660 CE: cassette exons

661 EMT: Epithelial-to-Mesenchymal Transition

662 CSC: Cancer Stem Cells

663 CTC: Circulating Tumour Cells

664 PSI: Percentage Spliced-In

665 TPM: Transcripts per Million

666 DSS: Disease Specific Survival

667 TCGA: The Cancer Genome Atlas

668 RBPs: RNA binding proteins

669

670

671 **Supplementary Information**

672 **Additional File 1: Figures S1-S6.**

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

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

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

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

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

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

684

685 **Additional File 2: Table S1-S2**

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

687 Table S2 – Name, coordinates (Hg38) and PSI mean value and standard error for the
688 25 exons of the basal B-specific signature in Basal A and Basal B cancer cells and
689 patients. The difference in splicing levels between basal B and basal A is shown as
690 deltaPSI.

691

692 **Figure Legends**

693 **Figure 1. Basal cell lines are divided in two subgroups based on gene expression
694 and splicing patterns. a.** Heatmap of the Transcripts per Million (TPM) values of the
695 635 genes which differential expression can cluster breast cancer cell lines into basal

696 A and basal B (P-value $< 10^{-3}$ by Kruskal-Wallis Test). **b.** Heatmap of the Percentage
697 Spliced-In (PSI) values of the 217 exons which differential splicing can cluster breast
698 cancer cell lines into basal A and basal B (P-value $< 10^{-3}$ by Kruskal-Wallis Test). **c.**
699 Venn Diagram of the genes differentially expressed and/or spliced between basal A
700 and basal B cancer cell lines. The overlap is not higher than expected by Fisher's exact
701 test, two tail ($p=0.098$)

702

703 **Figure 2. Basal B cell lines show mesenchymal, stem-like and resistance to**
704 **treatment characteristics. a,b,c.** Gene Set Enrichment Analysis (GSEA) of
705 differentially expressed genes between basal A and B cell lines for three different
706 signatures: Mammary Stem Cell, EMT and Resistance to Gefitinib. Up-regulated genes
707 in all signatures are enriched in basal B cell lines ($FDR < 0.25$). **d.** Gene ontology
708 analysis bar graphs for differentially expressed (left) and differentially spliced (right)
709 genes between basal A and B cell lines. Gene ontology terms related to Cellular
710 Component (GO_CC_FAT), Molecular Function (GO_MF_FAT) and Biological
711 Process (GO_BP_FA) are shown in the y axis in blue, yellow and red, respectively.
712 Benjamini false discovery rate (FDR, $-\log_{10}$) is shown on the x axis. Vertical lines mark
713 an FDR threshold of $FDR=0.05$ ($-\log_{10}(0.05)=1.3$) for differentially expressed and
714 spliced genes, respectively. **e.** Box plots of the median and 25th percentile of the
715 CD44/CD24 \log_2 expression ratio for basal A and B cell lines. P-value is calculated
716 using the Wilcoxon rank-sum test. **f.** Boxplots comparing IC50 values in basal A and B
717 cell lines upon treatment with different drugs from the Genomics of Drug Sensitivity in
718 Cancer 2 (GDS2) dataset. P-values are calculated using the Wilcoxon rank-sum test.

719

720 **Figure 3. A Random Forest Classifier using knowledge transfer from cell lines**

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

738

739 **Figure 4. The basal B-specific splicing signature is associated to EMT features.**

740 **a.** Heatmap of the Percentage Spliced-In (PSI) values of the 25 cassette exons most
741 informative to classify TCGA basal-like patients into basal B-like (red) or basal A-like
742 (blue). Claudin low tumors are highlighted in green. **b,c.** Sashimi plots displaying ENAH
743 and PLOD2 splicing patterns in randomly selected patients classified as basal A-like
744 and basal B-like. **d.** Changes in alternative splicing of these 25 basal B-specific splicing
745 events is sufficient to properly cluster 55 basal breast cancer cell lines from 3 unrelated

746 sequencing projects into basal B and basal A using t-SNE. Of note, three basal B cell
747 lines, HCC38, MDA-MB-157 and SUM102 were misclassified as Basal A cell lines (red
748 dots). Although HCC38 has also been classified as Basal A in the DepMap portal
749 (www.depmap.org). **e.** Heatmap of the PSI values of the 25 basal B-specific splicing
750 signature in public RNA-seq datasets from four different EMT projects. Basal B-like
751 events have the same splicing patterns as EMT-induced cells.

752

753 **Figure 5. Basal B-like patients express hallmarks of EMT and metastasis that**
754 **leads to a poor prognosis. a.** Lehman classification for basal A- and B-like patients.
755 ** $p < 0.01$ in Fisher's exact test, two tail, comparing basal B to basal A. **b.** Gene Set
756 Enrichment Analysis (GSEA) of the genes differentially expressed between basal A-
757 and B-like patients. Hallmark EMT and inflammatory response signatures are enriched
758 in basal B-like patients. **c.** Box plots of the median and 25th percentile of the expression
759 levels (in TPM) of major epithelial and mesenchymal-like EMT markers in basal A-like
760 (blue) and basal B-like (red) patients. **d.** Box plot of of the mean and 25th percentile of
761 the expression levels (in TPM) of Basal-like non-invasive and mesenchymal-like
762 invasive markers in basal A-like (blue) and basal B-like (red) patients. ** $P < 0.01$, ***
763 $P < 0.001$, **** $P < 0.0001$ in Wilcoxon rank-sum test comparing basal A-like to basal B-
764 like.

765

766 **Figure 6. The basal B-specific splicing signature is co-regulated by ESRP1 and**
767 **RBM47. a.** Heatmap of Transcripts per Million values for RNA Binding Proteins (RBP)
768 differentially expressed in basal A and basal B cell lines (P -value $< 10^{-9}$ by Kruskal-
769 Wallis Test). **b.** Box plots of the mean and 25th percentile of the expression levels (in
770 TPM) of the same RBP as in a, but in basal A-like and basal B-like patients. **c,d.** Venn

771 diagrams of the number of splicing events from the basal B-specific splicing signature
772 dependent on the splicing factors (SF) ESRP1/2 and RBM47 using a cutoff of
773 $|\Delta\text{Psi}| > 0.1$ and a higher probability ≥ 0.95 . **e,f,g.** Heatmaps of the PSI values of
774 the ESRP and RBM47-dependent exons from c and d in ESRP1/2 knock down H358
775 cells, RBM47 overexpressed MDA-MB-231 cells and RBM47 knock down H358
776 cells. **h,i.** Kaplan Meier plots for overall survival in basal-like TCGA patients expressing
777 the highest tercile (blue) or the lowest tercile (red) of ESRP1 and RBM47 expression
778 levels. HR (Hazard Ratio) and Logrank p-values (P) discriminating between groups are
779 shown.

780

781

782 **References**

- 783 1. Sims AH, Howell A, Howell SJ, Clarke RB. Origins of breast cancer subtypes and
784 therapeutic implications. *Nature Clinical Practice Oncology*. 2007.
- 785 2. Toft DJ, Cryns VL. Minireview: Basal-like breast cancer: From molecular profiles to
786 targeted therapies. *Molecular Endocrinology*. 2011.
- 787 3. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene
788 expression patterns of breast carcinomas distinguish tumor subclasses with clinical
789 implications. *Proceedings of the National Academy of Sciences of the United States*
790 *of America*. 2001;98:10869–74.
- 791 4. Dai X, Li T, Bai Z, Yang Y, Liu X, Zhan J, et al. Breast cancer intrinsic subtype
792 classification, clinical use and future trends. *American Journal of Cancer Research*.
793 2015.

- 794 5. Cardoso F, Van't Veer LJ, Bogaerts J, Slaets L, Viale G, Delaloge S, et al. 70-
795 Gene signature as an aid to treatment decisions in early-stage breast cancer. *New*
796 *England Journal of Medicine*. 2016.
- 797 6. Jiang Y-Z, Ma D, Suo C, Shi J, Xue M, Hu X, et al. Genomic and Transcriptomic
798 Landscape of Triple-Negative Breast Cancers: Subtypes and Treatment Strategies.
799 *Cancer Cell*. 2019;;428–40.
- 800 7. Marcotte R, Sayad A, Brown KR, Sanchez-Garcia F, Reimand J, Haider M, et al.
801 Functional Genomic Landscape of Human Breast Cancer Drivers, Vulnerabilities, and
802 Resistance. *Cell*. 2016.
- 803 8. Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, et al. Association
804 analysis identifies 65 new breast cancer risk loci. *Nature*. 2017.
- 805 9. Milne RL, Kuchenbaecker KB, Michailidou K, Beesley J, Kar S, Lindström S, et al.
806 Identification of ten variants associated with risk of estrogen-receptor-negative breast
807 cancer. *Nature Genetics*. 2017.
- 808 10. Garcia-Closas M, Couch FJ, Lindstrom S, Michailidou K, Schmidt MK, Brook MN,
809 et al. Genome-wide association studies identify four ER negative-specific breast
810 cancer risk loci. *Nature Genetics*. 2013.
- 811 11. Karni R, De Stanchina E, Lowe SW, Sinha R, Mu D, Krainer AR. The gene
812 encoding the splicing factor SF2/ASF is a proto-oncogene. *Nature Structural and*
813 *Molecular Biology*. 2007.
- 814 12. Climente-González H, Porta-Pardo E, Godzik A, Eyraes E. The Functional Impact
815 of Alternative Splicing in Cancer. *Cell Reports*. 2017;20:2215–26.

- 816 13. Sebestyén E, Zawisza M, Eyraş E. Detection of recurrent alternative splicing
817 switches in tumor samples reveals novel signatures of cancer. *Nucleic Acids*
818 *Research*. 2015;43:1345–56.
- 819 14. Kahles A, Lehmann K-V, Toussaint NC, Hüser M, Stark SG, Sachsenberg T, et
820 al. Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705
821 Patients. *Cancer Cell*. 2018;0:1–14.
- 822 15. David CJ, Manley JL. Alternative pre-mRNA splicing regulation in cancer:
823 Pathways and programs unhinged. *Genes and Development*. 2010.
- 824 16. Bechara EG, Sebestyén E, Bernardis I, Eyraş E, Valcárcel J. RBM5, 6, and 10
825 differentially regulate NUMB alternative splicing to control cancer cell proliferation.
826 *Molecular Cell*. 2013;52:720–33.
- 827 17. Moore MJ, Wang Q, Kennedy CJ, Silver PA. An alternative splicing network links
828 cell-cycle control to apoptosis. *Cell*. 2010.
- 829 18. Amin EM, Oltean S, Hua J, Gammons MVR, Hamdollah-Zadeh M, Welsh GI, et
830 al. WT1 Mutants Reveal SRPK1 to Be a Downstream Angiogenesis Target by
831 Altering VEGF Splicing. *Cancer Cell*. 2011.
- 832 19. Chen M, Zhang J, Manley JL. Turning on a fuel switch of cancer: hnRNP proteins
833 regulate alternative splicing of pyruvate kinase mRNA. *Cancer Research*. 2010.
- 834 20. Pascual G, Avgustinova A, Mejetta S, Martín M, Castellanos A, Attolini CSO, et
835 al. Targeting metastasis-initiating cells through the fatty acid receptor CD36. *Nature*.
836 2017.

- 837 21. Xu Y, Gao XD, Lee JH, Huang H, Tan H, Ahn J, et al. Cell type-restricted activity
838 of hnRNPM promotes breast cancer metastasis via regulating alternative splicing.
839 *Genes and Development*. 2014;28:1191–203.
- 840 22. Brown RL, Reinke LM, Damerow MS, Perez D, Chodosh LA, Yang J, et al. CD44
841 splice isoform switching in human and mouse epithelium is essential for epithelial-
842 mesenchymal transition and breast cancer progression. *Journal of Clinical*
843 *Investigation*. 2011;121:1064–74.
- 844 23. Li J, Choi PS, Chaffer CL, Labella K, Hwang JH, Giacomelli AO, et al. An
845 alternative splicing switch in FLNB promotes the mesenchymal cell state in human
846 breast cancer. *eLife*. 2018;7:1–28.
- 847 24. Ranieri D, Rosato B, Nanni M, Magenta A, Belleudi F, Torrisi MR. Expression of
848 the FGFR2 mesenchymal splicing variant in epithelial cells drives epithelial-
849 mesenchymal transition. *Oncotarget*. 2016;7:5440–60.
- 850 25. Lee SCW, Abdel-Wahab O. Therapeutic targeting of splicing in cancer. *Nature*
851 *Medicine*. 2016.
- 852 26. Bonomi S, Gallo S, Catillo M, Pignataro D, Biamonti G, Ghigna C. Oncogenic
853 alternative splicing switches: Role in cancer progression and prospects for therapy.
854 *International Journal of Cell Biology*. 2013.
- 855 27. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al.
856 The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug
857 sensitivity. *Nature*. 2012;483:603–7.

- 858 28. Neve RM, Chin K, Fridlyand J, Yeh J, Baehner FL, Fevr T, et al. A collection of
859 breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer*
860 *Cell*. 2006;10:515–27.
- 861 29. Mani SA, Guo W, Liao MJ, Eaton EN, Ayyanan A, Zhou AY, et al. The epithelial-
862 mesenchymal transition generates cells with properties of stem cells. *Cell*.
863 2008;133:704–15.
- 864 30. Hennessy BT, Gonzalez-Angulo A-M, Stemke-Hale K, Gilcrease MZ,
865 Krishnamurthy S, Lee J-S, et al. Characterization of a Naturally Occurring Breast
866 Cancer Subset Enriched in Epithelial-to-Mesenchymal Transition and Stem Cell
867 Characteristics. *Cancer Res*. 2009;69:4116–24.
- 868 31. Thiery JP, Acloque H, Huang RYJ, Nieto MA. Epithelial-Mesenchymal Transitions
869 in Development and Disease. *Cell*. 2009;139:871–90.
- 870 32. Ye X, Tam WL, Shibue T, Kaygusuz Y, Reinhardt F. Distinct EMT programs
871 control normal mammary stem cells and tumour-initiating cells. 2016;525:256–60.
- 872 33. Kao J, Salari K, Bocanegra M, Choi Y La, Girard L, Gandhi J, et al. Molecular
873 profiling of breast cancer cell lines defines relevant tumor models and provides a
874 resource for cancer gene discovery. *PLoS ONE*. 2009;4.
- 875 34. Charafe-Jauffret E, Ginestier C, Monville F, Finetti P, Adélaïde J, Cervera N, et
876 al. Gene expression profiling of breast cell lines identifies potential new basal
877 markers. *Oncogene*. 2006;25:2273–84.

- 878 35. Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF,
879 et al. Comprehensive molecular portraits of human breast tumours. *Nature*.
880 2012;490:61–70.
- 881 36. Yae T, Tsuchihashi K, Ishimoto T, Motohara T, Yoshikawa M, Yoshida GJ, et al.
882 Alternative splicing of CD44 mRNA by ESRP1 enhances lung colonization of
883 metastatic cancer cell. *Nature Communications*. 2012;3 May.
- 884 37. De Faria Poloni J, Bonatto D. Influence of transcriptional variants on metastasis.
885 *RNA Biology*. 2018.
- 886 38. Qiu Y, Lyu J, Dunlap M, Harvey SE, Cheng C. A combinatorially regulated RNA
887 splicing signature predicts breast cancer EMT states and patient survival. *Rna*.
888 2020;;rna.074187.119.
- 889 39. Sebestyén E, Singh B, Miñana B, Pagis A, Mateo F, Pujana MA, et al.
890 Large-scale analysis of genome and transcriptome alterations in multiple tumors
891 unveils novel cancer-relevant splicing networks. *Genome Research*. 2016;26:732–
892 44.
- 893 40. Shapiro IM, Cheng AW, Flytzanis NC, Balsamo M, Condeelis JS, Oktay MH, et
894 al. An EMT–Driven Alternative Splicing Program Occurs in Human Breast Cancer
895 and Modulates Cellular Phenotype. *PLoS Genet*. 2011;7.
- 896 41. Warzecha CC, Jiang P, Amirikian K, Dittmar KA, Lu H, Shen S, et al. An ESRP-
897 regulated splicing programme is abrogated during the epithelial-mesenchymal
898 transition. *EMBO J*. 2010;29:3286–300.

- 899 42. Pan Q, Shai O, Misquitta C, Zhang W, Saltzman AL, Mohammad N, et al.
900 Revealing Global Regulatory Features of Mammalian Alternative Splicing Using a
901 Quantitative Microarray Platform. *Molecular Cell*. 2004;16:929–41.
- 902 43. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et
903 al. Gene set enrichment analysis: A knowledge-based approach for interpreting
904 genome-wide expression profiles. *Proceedings of the National Academy of Sciences*.
905 2005;102:15545–50.
- 906 44. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID:
907 Database for Annotation, Visualization, and Integrated Discovery. *Genome biology*.
908 2003.
- 909 45. Dragowska WH, Wepler SA, Qadir MA, Wong LY, Franssen Y, Baker JHE, et al.
910 The combination of gefitinib and RAD001 inhibits growth of HER2 overexpressing
911 breast cancer cells and tumors irrespective of trastuzumab sensitivity. *BMC Cancer*.
912 2011.
- 913 46. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al.
914 Genomics of Drug Sensitivity in Cancer (GDSC): A resource for therapeutic
915 biomarker discovery in cancer cells. *Nucleic Acids Research*. 2013.
- 916 47. Ho-Yen CM, Jones JL, Kermorgant S. The clinical and functional significance of
917 c-Met in breast cancer: A review. *Breast Cancer Research*. 2015.
- 918 48. Berger AC, Korkut A, Kanchi RS, Hegde AM, Lenoir W, Liu W, et al. A
919 Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers.
920 *Cancer Cell*. 2018;33:690-705.e9.

- 921 49. Kursa MB, Rudnicki WR. Feature selection with the boruta package. Journal of
922 Statistical Software. 2010;36:1–13.
- 923 50. Tian B, Li X, Kalita M, Widen SG, Yang J, Bhavnani SK, et al. Analysis of the
924 TGF β -induced program in primary airway epithelial cells shows essential role of NF-
925 KB/RelA signaling network in type II epithelial mesenchymal transition. BMC
926 Genomics. 2015.
- 927 51. Pillman KA, Phillips CA, Roslan S, Toubia J, Dredge BK, Bert AG, et al. miR-
928 200/375 control epithelial plasticity-associated alternative splicing by repressing the
929 RNA -binding protein Quaking . The EMBO Journal. 2018.
- 930 52. Pattabiraman DR, Bieri B, Kober KI, Thiru P, Krall JA, Zill C, et al. Activation of
931 PKA leads to mesenchymal-to-epithelial transition and loss of tumor-initiating ability.
932 Science. 2016.
- 933 53. Yang Y, Park JW, Bebee TW, Warzecha CC, Guo Y, Shang X, et al.
934 Determination of a Comprehensive Alternative Splicing Regulatory Network and
935 Combinatorial Regulation by Key Factors during the Epithelial-to-Mesenchymal
936 Transition. Molecular and Cellular Biology. 2016;36:1704–19.
- 937 54. Lehmann BD, Shyr Y, Pietenpol JA, Lehmann BD, Bauer JA, Chen X, et al.
938 Identification of human triple-negative breast cancer subtypes and preclinical models
939 for selection of targeted therapies Find the latest version : Identification of human
940 triple-negative breast cancer subtypes and preclinical models for selection of targ.
941 2011;121:2750–67.
- 942 55. Caramel J, Ligier M, Puisieux A. Pleiotropic Roles for ZEB1 in Cancer. 2018;78.

- 943 56. Bebee TW, Park JW, Sheridan KI, Warzecha CC, Cieply BW, Rohacek AM, et al.
944 The splicing regulators *Esrp1* and *Esrp2* direct an epithelial splicing program
945 essential for mammalian development. *eLife*. 2015;4:1–27.
- 946 57. Vanharanta S, Marney CB, Shu W, Valiente M, Zou Y, Mele A, et al. Loss of the
947 multifunctional RNA-binding protein RBM47 as a source of selectable metastatic
948 traits in breast cancer. *eLife*. 2014;2014:1–24.
- 949 58. Park SH, Brugiolo M, Akerman M, Das S, Urbanski L, Geier A, et al. Differential
950 Functions of Splicing Factors in Mammary Transformation and Breast Cancer
951 Metastasis. *Cell Reports*. 2019;29:2672-2688.e7.
- 952 59. Hendley AM, Wang YJ, Polireddy K, Alsina J, Ahmed I, Lafaro KJ, et al. p120
953 catenin suppresses basal epithelial cell extrusion in invasive pancreatic neoplasia.
954 *Cancer Research*. 2016.
- 955 60. Braeutigam C, Rago L, Rolke A, Waldmeier L, Christofori G, Winter J. The RNA-
956 binding protein *Rbfox2*: An essential regulator of EMT-driven alternative splicing and
957 a mediator of cellular invasion. *Oncogene*. 2014;33:1082–92.
- 958 61. Roovers K, Wagner S, Storbeck CJ, O'Reilly P, Lo V, Northey JJ, et al. The
959 Ste20-like kinase *SLK* is required for ErbB2-driven breast cancer cell motility.
960 *Oncogene*. 2009.
- 961 62. Meng J. Distinct functions of dynamin isoforms in tumorigenesis and their
962 potential as therapeutic targets in cancer. *Oncotarget*. 2017.

- 963 63. Song Y, Zheng S, Wang J, Long H, Fang L, Wang G, et al. Hypoxia-induced
964 PLOD2 promotes proliferation, migration and invasion via PI3K/Akt signaling in
965 glioma. *Oncotarget*. 2017.
- 966 64. Urbanski LM, Leclair N, Anczuków O. Alternative-splicing defects in cancer:
967 Splicing regulators and their downstream targets, guiding the way to novel cancer
968 therapeutics. *Wiley Interdisciplinary Reviews: RNA*. 2018; January:e1476.
- 969 65. Anczukow O, Krainer AR. Splicing-factor alterations in cancers. *Rna*.
970 2016;22:1285–301.
- 971 66. Pagliarini V, Naro C, Sette C. Splicing regulation: A molecular device to enhance
972 cancer cell adaptation. *BioMed Research International*. 2015.
- 973 67. Di Modugno F, Iapicca P, Boudreau A, Mottolese M, Terrenato I, Perracchio L, et
974 al. Splicing program of human MENA produces a previously undescribed isoform
975 associated with invasive, mesenchymal-like breast tumors. *Proceedings of the*
976 *National Academy of Sciences of the United States of America*. 2012.
- 977 68. Weinstein JN. Cell lines battle cancer. *Nature*. 2012.
- 978 69. Jiang G, Zhang S, Yazdanparast A, Li M, Pawar AV, Liu Y, et al. Comprehensive
979 comparison of molecular portraits between cell lines and tumors in breast cancer.
980 *BMC Genomics*. 2016.
- 981 70. Yu K, Chen B, Aran D, Charalel J, Yau C, Wolf DM, et al. Comprehensive
982 transcriptomic analysis of cell lines as models of primary tumors across 22 tumor
983 types. *Nature Communications*. 2019;10.

- 984 71. Dvinge H, Bradley RK. Widespread intron retention diversifies most cancer
985 transcriptomes. *Genome Medicine*. 2015;7:45.
- 986 72. Jung H, Lee D, Lee J, Park D, Kim YJ, Park WY, et al. Intron retention is a
987 widespread mechanism of tumor-suppressor inactivation. *Nat Genet*. 2015.
- 988 73. Chen J, Weiss WA. Alternative splicing in cancer: implications for biology and
989 therapy. *Oncogene*. 2015;34:1–14.
- 990 74. Warzecha CC, Carstens RP. Complex changes in alternative pre-mRNA splicing
991 play a central role in the epithelial-to-mesenchymal transition (EMT). *Seminars in*
992 *Cancer Biology*. 2012;22:417–27.
- 993 75. Itoh M, Radisky DC, Hashiguchi M, Sugimoto H. The exon 38-containing
994 ARHGEF11 splice isoform is differentially expressed and is required for migration
995 and growth in invasive breast cancer cells. *Oncotarget*. 2017.
- 996 76. Zhao N, Guo M, Wang K, Zhang C, Liu X. Identification of Pan-Cancer Prognostic
997 Biomarkers Through Integration of Multi-Omics Data. *Frontiers in Bioengineering and*
998 *Biotechnology*. 2020.
- 999 77. Wang H, Shao Q, Sun J, Ma C, Gao W, Wang Q, et al. Interactions between
1000 colon cancer cells and tumor-infiltrated macrophages depending on cancer cell-
1001 derived colony stimulating factor 1. *Oncolmmunology*. 2016.
- 1002 78. Chen Y, Lu Y, Ren Y, Yuan J, Zhang N, Kimball H, et al. Starvation-induced
1003 suppression of DAZAP1 by miR-10b integrates splicing control into TSC2-regulated
1004 oncogenic autophagy in esophageal squamous cell carcinoma. *Theranostics*. 2020.

1005 79. Yan Q, Lou G, Qian Y, Qin B, Xu X, Wang Y, et al. SPAG9 is involved in
1006 hepatocarcinoma cell migration and invasion via modulation of ELK1 expression.
1007 OncoTargets and Therapy. 2016.

1008 80. Chen X, Zhao C, Guo B, Zhao Z, Wang H, Fang Z. Systematic Profiling of
1009 Alternative mRNA Splicing Signature for Predicting Glioblastoma Prognosis. Frontiers
1010 in Oncology. 2019.

1011 81. Zhang L, Liu X, Zhang X, Chen R. Identification of important long non-coding
1012 RNAs and highly recurrent aberrant alternative splicing events in hepatocellular
1013 carcinoma through integrative analysis of multiple RNA-Seq datasets. Molecular
1014 Genetics and Genomics. 2016.

1015 82. Venhuizen JH, Sommer S, Span PN, Friedl P, Zegers MM. Differential expression
1016 of p120-catenin 1 and 3 isoforms in epithelial tissues. Scientific Reports. 2019.

1017 83. Roussos ET, Wang Y, Wyckoff JB, Sellers RS, Wang W, Li J, et al. Mena
1018 deficiency delays tumor progression and decreases metastasis in polyoma middle-T
1019 transgenic mouse mammary tumors. Breast Cancer Research. 2010.

1020 84. Philippar U, Roussos ET, Oser M, Yamaguchi H, Kim H Do, Giampieri S, et al. A
1021 Mena Invasion Isoform Potentiates EGF-Induced Carcinoma Cell Invasion and
1022 Metastasis. Developmental Cell. 2008.

1023 85. Li Q, Su YL, Zeng M, Shen WX. Enabled homolog shown to be a potential
1024 biomarker and prognostic indicator for breast cancer by bioinformatics analysis.
1025 Clinical and Investigative Medicine. 2018.

- 1026 86. Zhang H, Brown RL, Wei Y, Zhao P, Liu S, Liu X, et al. CD44 splice isoform
1027 switching determines breast cancer stem cell state. *Genes and Development*.
1028 2019;33:166–79.
- 1029 87. Venables JP, Lapasset L, Gadea G, Fort P, Klinck R, Irimia M, et al. MBNL1 and
1030 RBFOX2 cooperate to establish a splicing programme involved in pluripotent stem
1031 cell differentiation. *Nature Communications*. 2013;4 May:1–10.
- 1032 88. Tabaglio T, Low DHP, Teo WKL, Goy PA, Cywoniuk P, Wollmann H, et al.
1033 MBNL1 alternative splicing isoforms play opposing roles in cancer. *Life Science*
1034 *Alliance*. 2018.
- 1035 89. Soncin I, Sheng J, Chen Q, Foo S, Duan K, Lum J, et al. The tumour
1036 microenvironment creates a niche for the self-renewal of tumour-promoting
1037 macrophages in colon adenoma. *Nature Communications*. 2018.
- 1038 90. Markus MA, Yang YHJ, Morris BJ. Transcriptome-wide targets of alternative
1039 splicing by RBM4 and possible role in cancer. *Genomics*. 2016.
- 1040 91. Sheng X, Li Y, Li Y, Liu W, Lu Z, Zhan J, et al. PLOD2 contributes to drug
1041 resistance in laryngeal cancer by promoting cancer stem cell-like characteristics.
1042 *BMC Cancer*. 2019.
- 1043 92. Conway J, Al-Zahrani KN, Pryce BR, Abou-Hamad J, Sabourin LA. Transforming
1044 growth factor β -induced epithelial to mesenchymal transition requires the Ste20-like
1045 kinase SLK independently of its catalytic activity. *Oncotarget*. 2017.

- 1046 93. de Miguel FJ, Pajares MJ, Martínez-Terroba E, Ajona D, Morales X, Sharma RD,
1047 et al. A large-scale analysis of alternative splicing reveals a key role of QKI in lung
1048 cancer. *Molecular Oncology*. 2016.
- 1049 94. Yang X, Zhou W, Liu S. SPAG9 controls the cell motility, invasion and
1050 angiogenesis of human osteosarcoma cells. *Experimental and Therapeutic Medicine*.
1051 2016.
- 1052 95. Jeong HM, Han J, Lee SH, Park HJ, Lee HJ, Choi JS, et al. ESRP1 is
1053 overexpressed in ovarian cancer and promotes switching from mesenchymal to
1054 epithelial phenotype in ovarian cancer cells. *Oncogenesis*. 2017.
- 1055 96. Hayakawa A, Saitoh M, Miyazawa K. Dual roles for epithelial splicing regulatory
1056 proteins 1 (ESRP1) and 2 (ESRP2) in cancer progression. In: *Advances in*
1057 *Experimental Medicine and Biology*. 2017.
- 1058 97. Sakurai T, Isogaya K, Sakai S, Morikawa M, Morishita Y, Ehata S, et al. RNA-
1059 binding motif protein 47 inhibits Nrf2 activity to suppress tumor growth in lung
1060 adenocarcinoma. *Oncogene*. 2017;36:5083.
- 1061 98. Rokavec M, Kaller M, Horst D, Hermeking H. Pan-cancer EMT-signature
1062 identifies RBM47 down-regulation during colorectal cancer progression. *Scientific*
1063 *Reports*. 2017;7:1–15.
- 1064 99. Cordero A, Kanojia D, Miska J, Panek WK, Xiao A, Han Y, et al. FABP7 is a key
1065 metabolic regulator in HER2+ breast cancer brain metastasis. *Oncogene*. 2019.

- 1088 109. Dai X, Cheng H, Bai Z, Li J. Breast cancer cell line classification and Its
1089 relevance with breast tumor subtyping. *Journal of Cancer*. 2017;8:3131–41.
- 1090 110. Fougner C, Bergholtz H, Norum JH, Sørлие T. Re-definition of claudin-low as a
1091 breast cancer phenotype. *Nature Communications*. 2020;11:756411.
- 1092 111. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio
1093 Cancer Genomics Portal: An open platform for exploring multidimensional cancer
1094 genomics data. *Cancer Discovery*. 2012.
- 1095 112. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al.
1096 Integrative analysis of complex cancer genomics and clinical profiles using the
1097 cBioPortal. *Science Signaling*. 2013.
- 1098 113. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al.
1099 An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality
1100 Survival Outcome Analytics. *Cell*. 2018;173:400-416.e11.

1101

1066 100. Savage P, Blanchet-Cohen A, Revil T, Badescu D, Saleh SMI, Wang YC, et al.
1067 A Targetable EGFR-Dependent Tumor-Initiating Program in Breast Cancer. *Cell*
1068 *Reports*. 2017;21:1140–9.

1069 101. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR:
1070 Ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013.

1071 102. Alvarez RV, Pongor LS, Mariño-Ramírez L, Landsman D. TPMCalculator: One-
1072 step software to quantify mRNA abundance of genomic features. *Bioinformatics*.
1073 2019.

1074 103. Tischler G, Leonard S. Biobambam: Tools for read pair collation based
1075 algorithms on BAM files. *Source Code for Biology and Medicine*. 2014.

1076 104. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of
1077 large gene lists using DAVID bioinformatics resources. *Nature Protocols*. 2009.

1078 105. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P. GenePattern 2.0 - Nature
1079 Genetics. *Nature Genetics*. 2006.

1080 106. Love MI, Huber W, Anders S. Moderated estimation of fold change and
1081 dispersion for RNA-seq data with DESeq2. *Genome Biology*. 2014.

1082 107. Garrido-Martín D, Palumbo E, Guigó R, Breschi A. ggsashimi: Sashimi plot
1083 revised for browser- and annotation-independent splicing visualization. *PLoS*
1084 *Computational Biology*. 2018.

1085 108. Mills GB, Sanchez-Garcia F, Virtanen C, Marcotte R, Pe'er D, Brown KR, et al.
1086 Functional Genomic Landscape of Human Breast Cancer Drivers, Vulnerabilities, and
1087 Resistance. *Cell*. 2016;164:293–309.

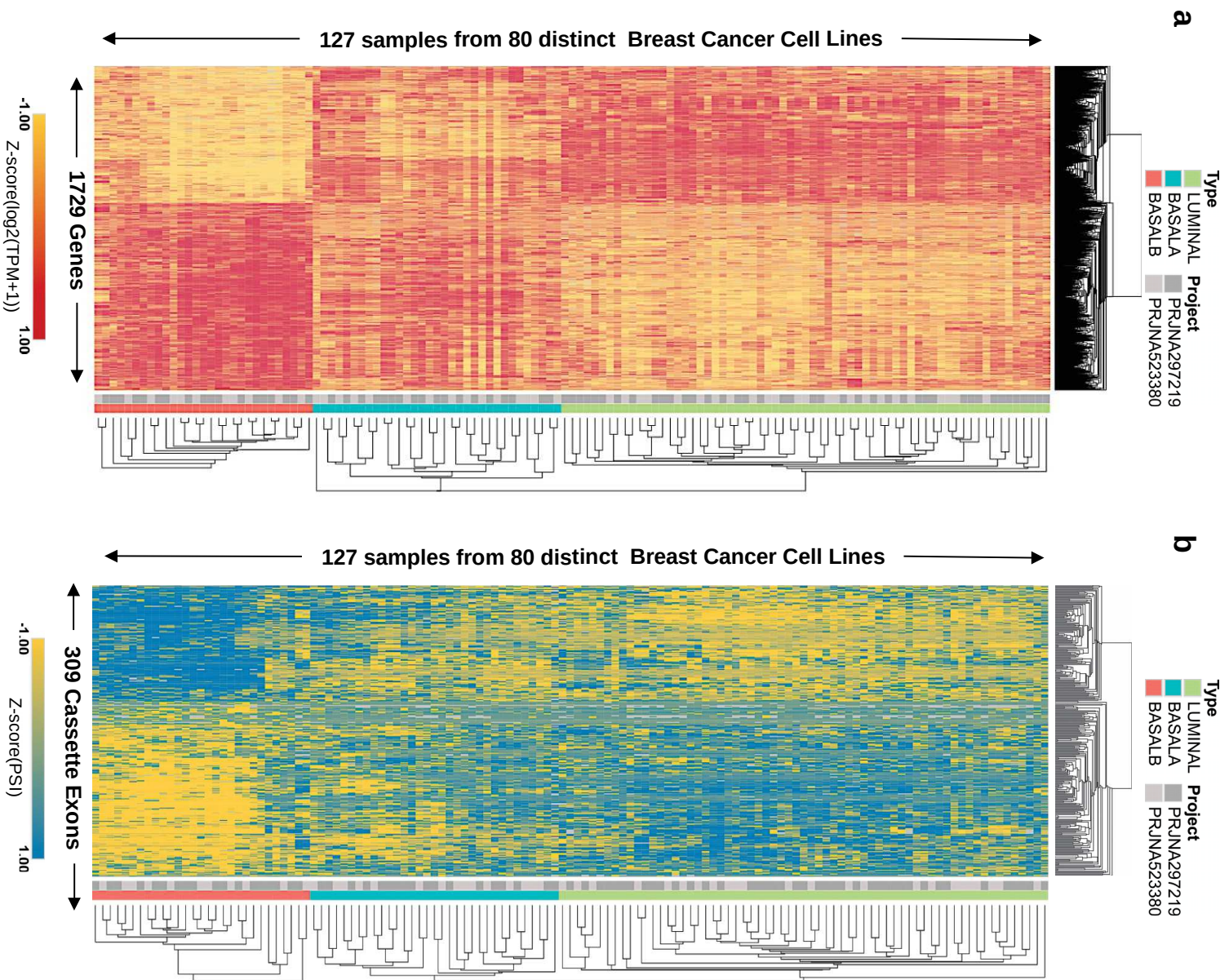


Figure S1

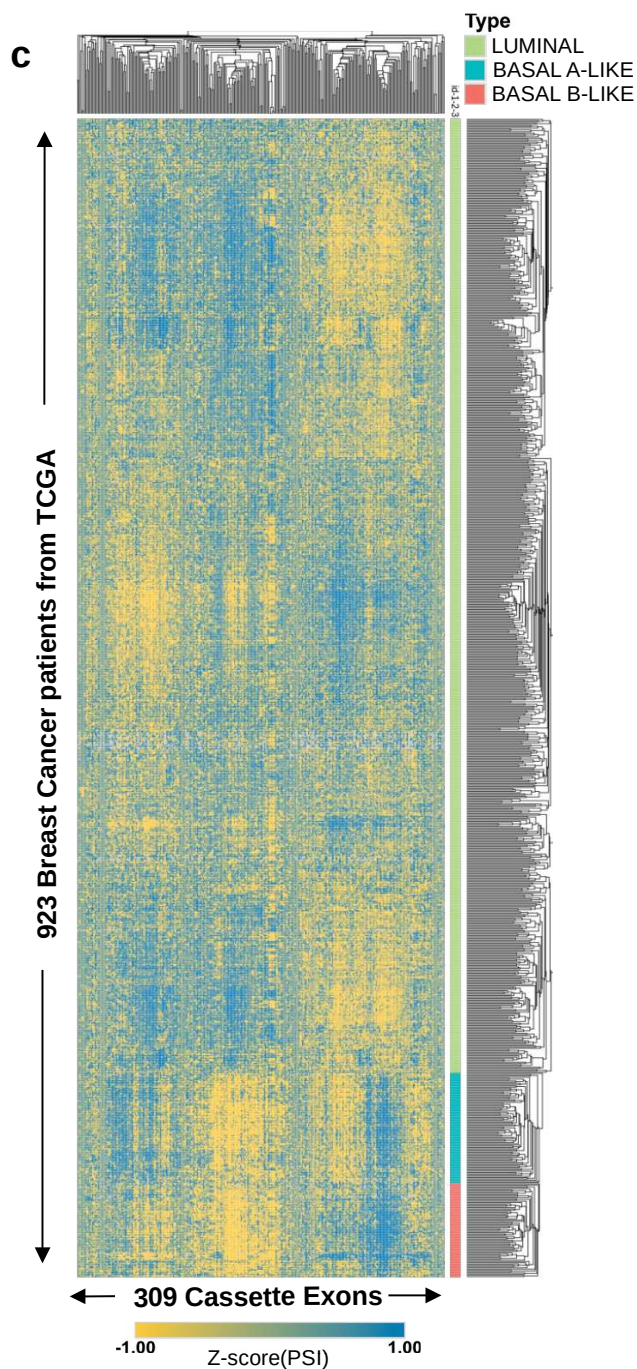


Figure S1. Differential clustering of basal B cell lines based on gene expression and splicing patterns. **a.** Heatmap of gene expression levels, in Transcripts per Million (TPM) values, of 1729 genes differentially regulated between Luminal, basal A and basal B cell lines (P-value < 10^{-5} by Kruskal-Wallis 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^{-5} 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.

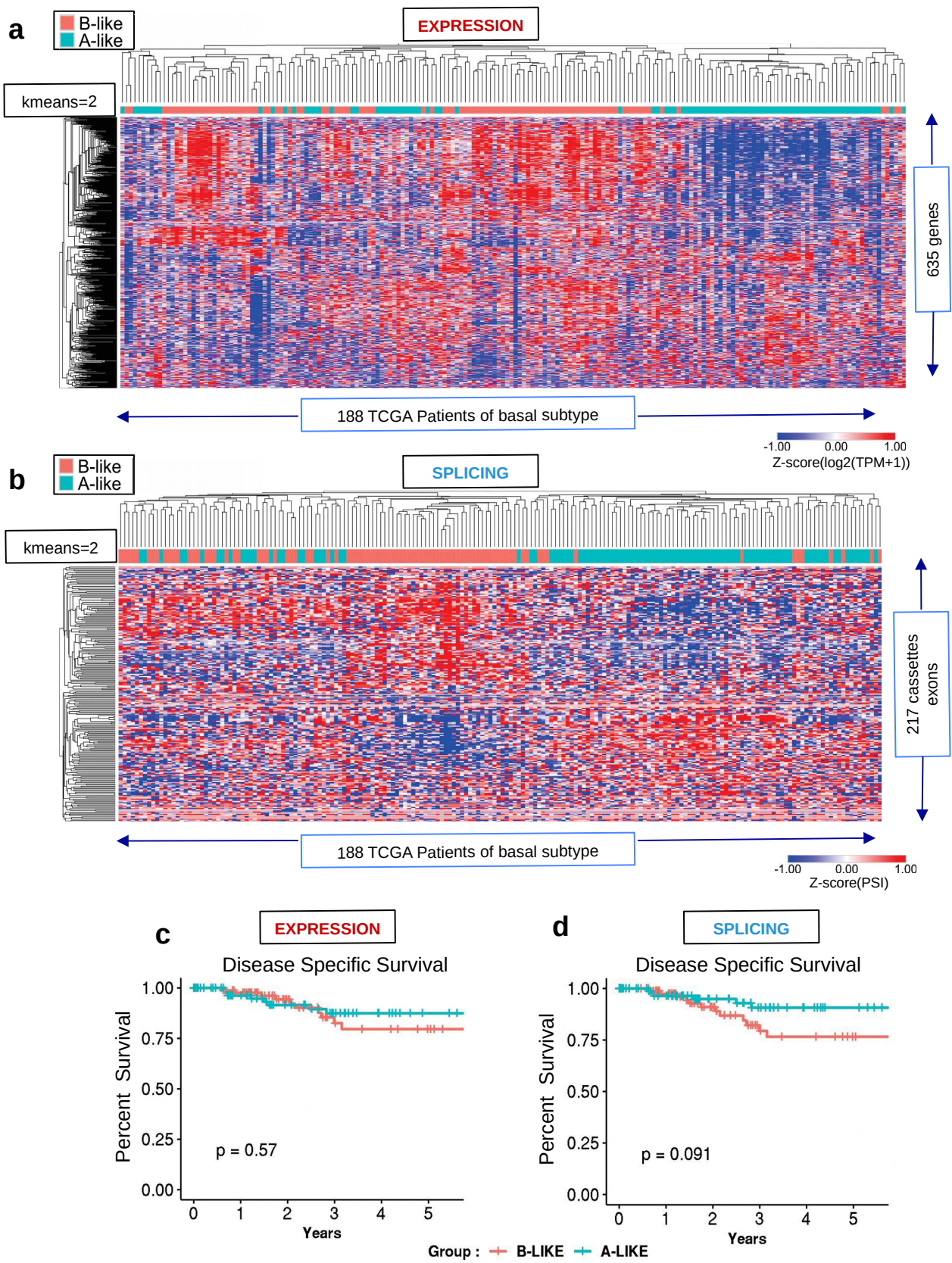


Figure S2

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.

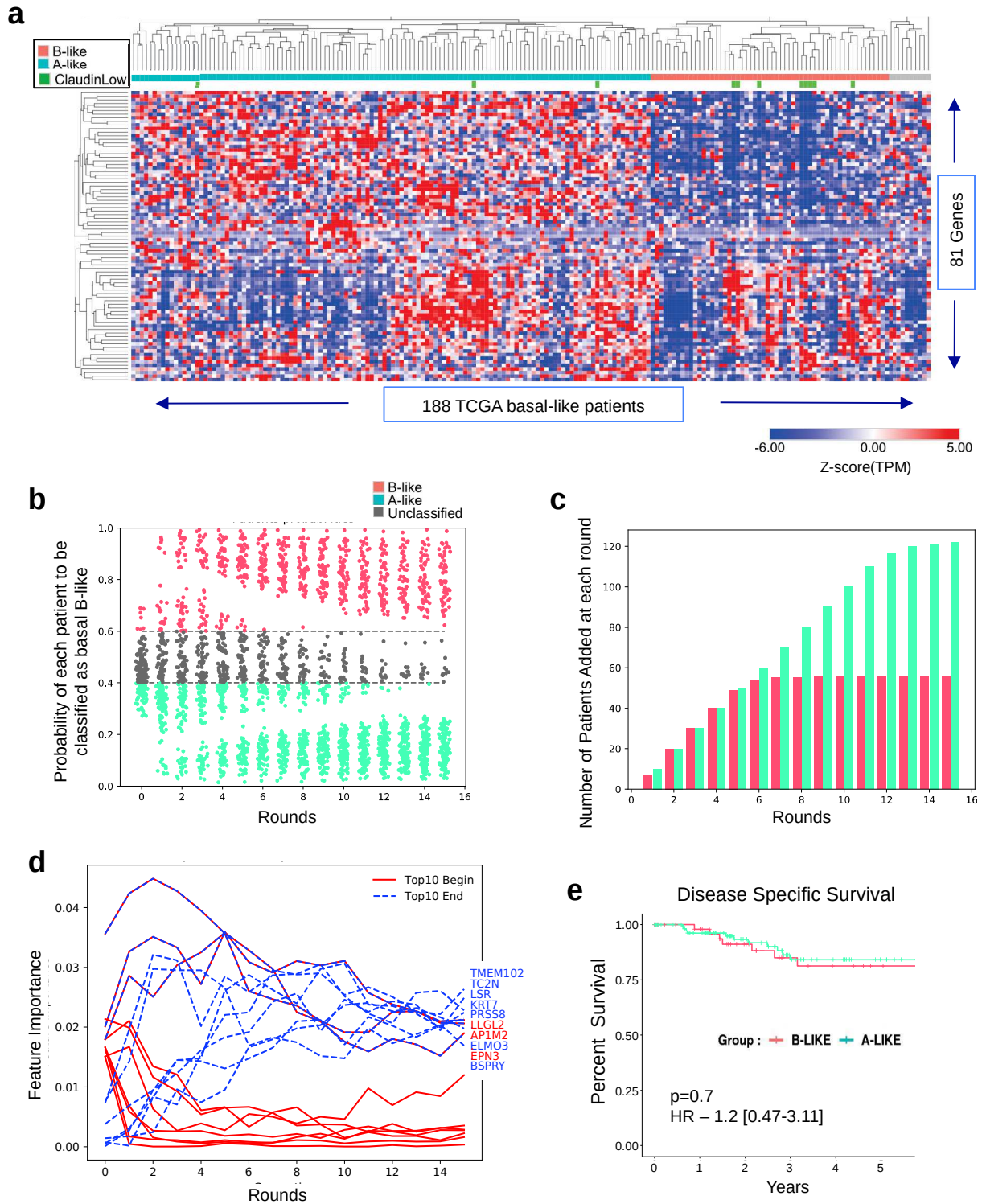


Figure S3

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.

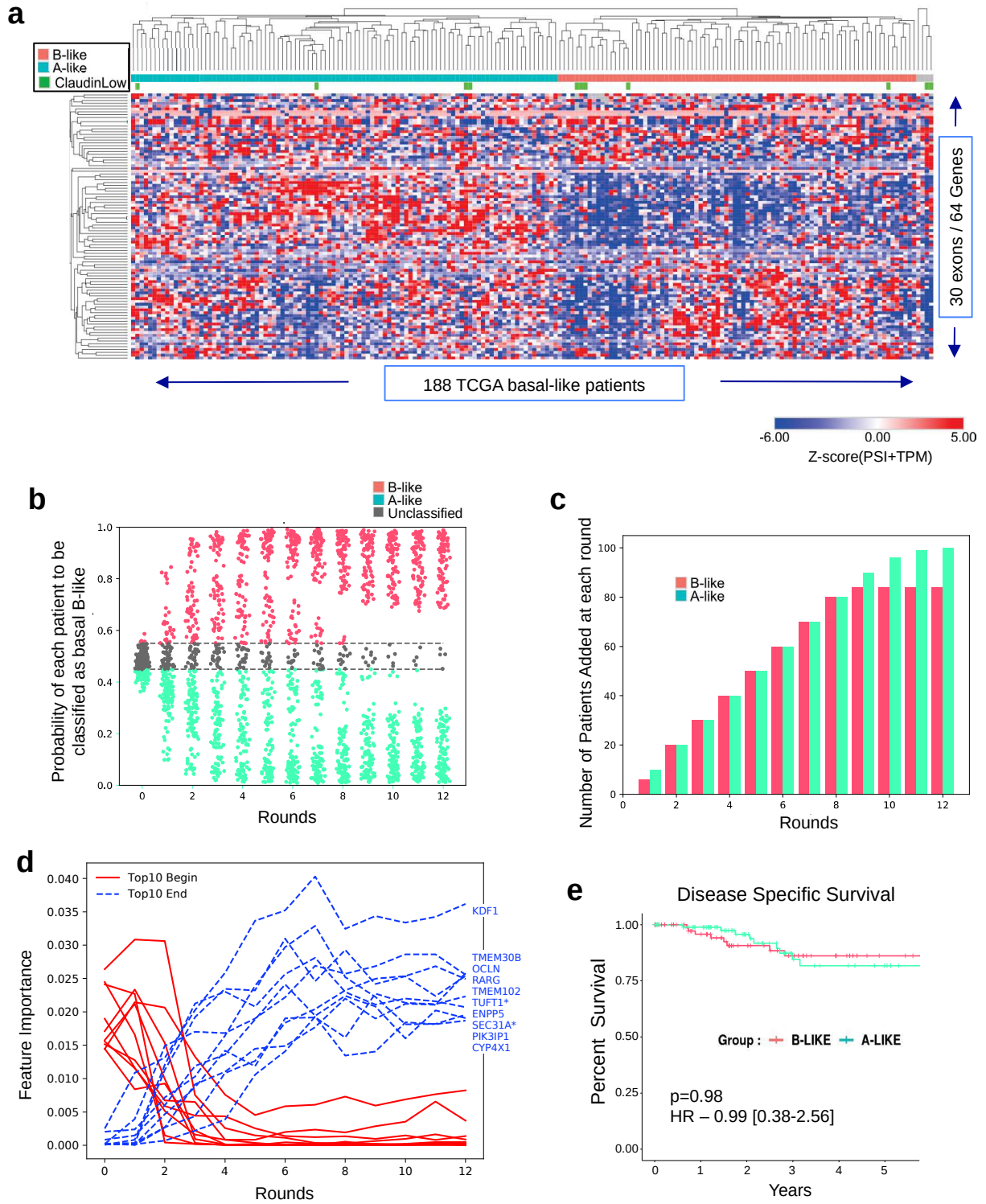


Figure S4

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.

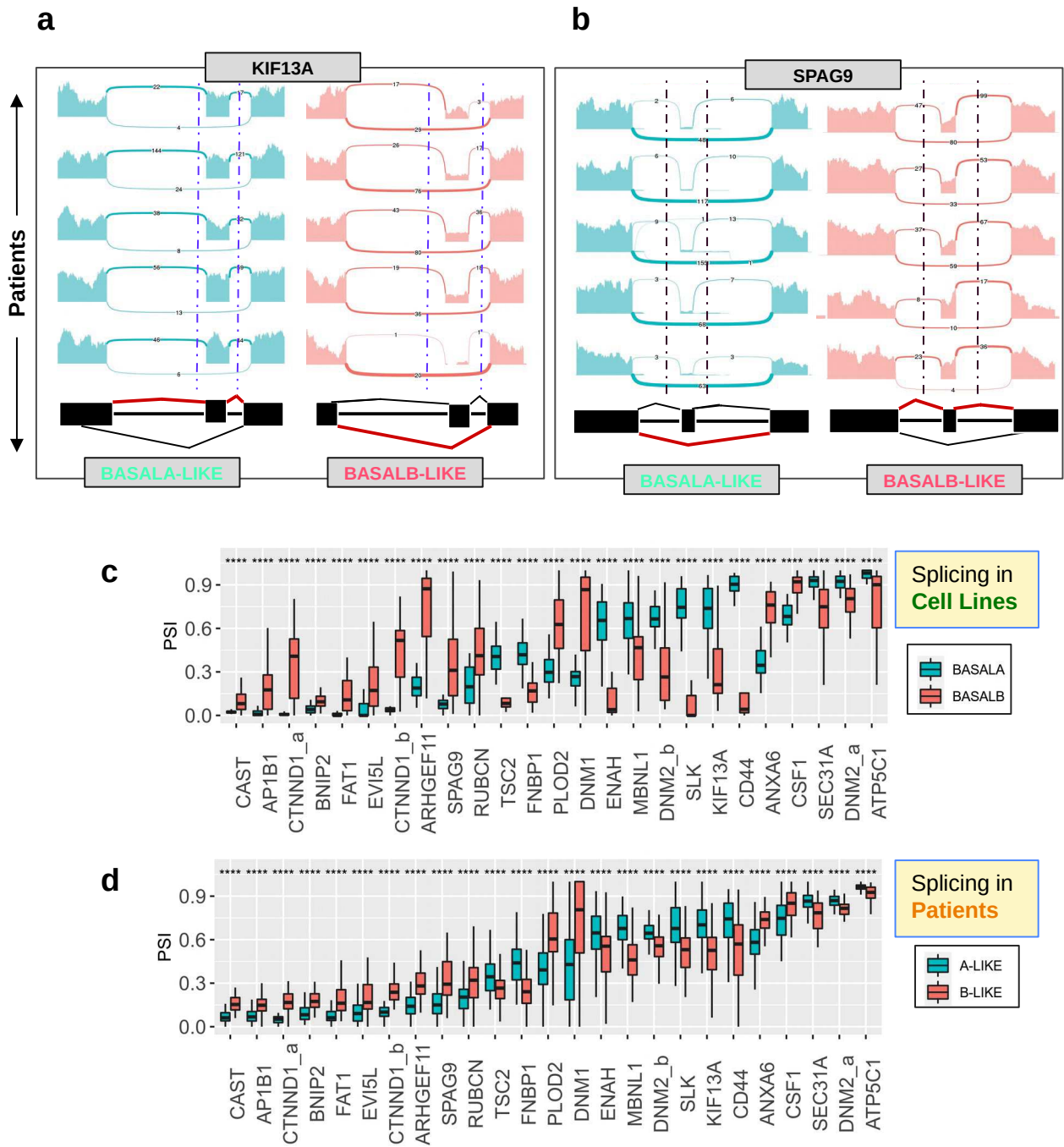


Figure S5. *In silico* validation of basal B splicing signature. a,b. Sashimi plots of KIF13A and SPAG9 patterns of splicing in randomly selected basal A-like and basal B-like patients. **c,d.** Box plots of the median and 25th percentile of the Percent Spliced-In (PSI) values for the 25 cassette exons 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.

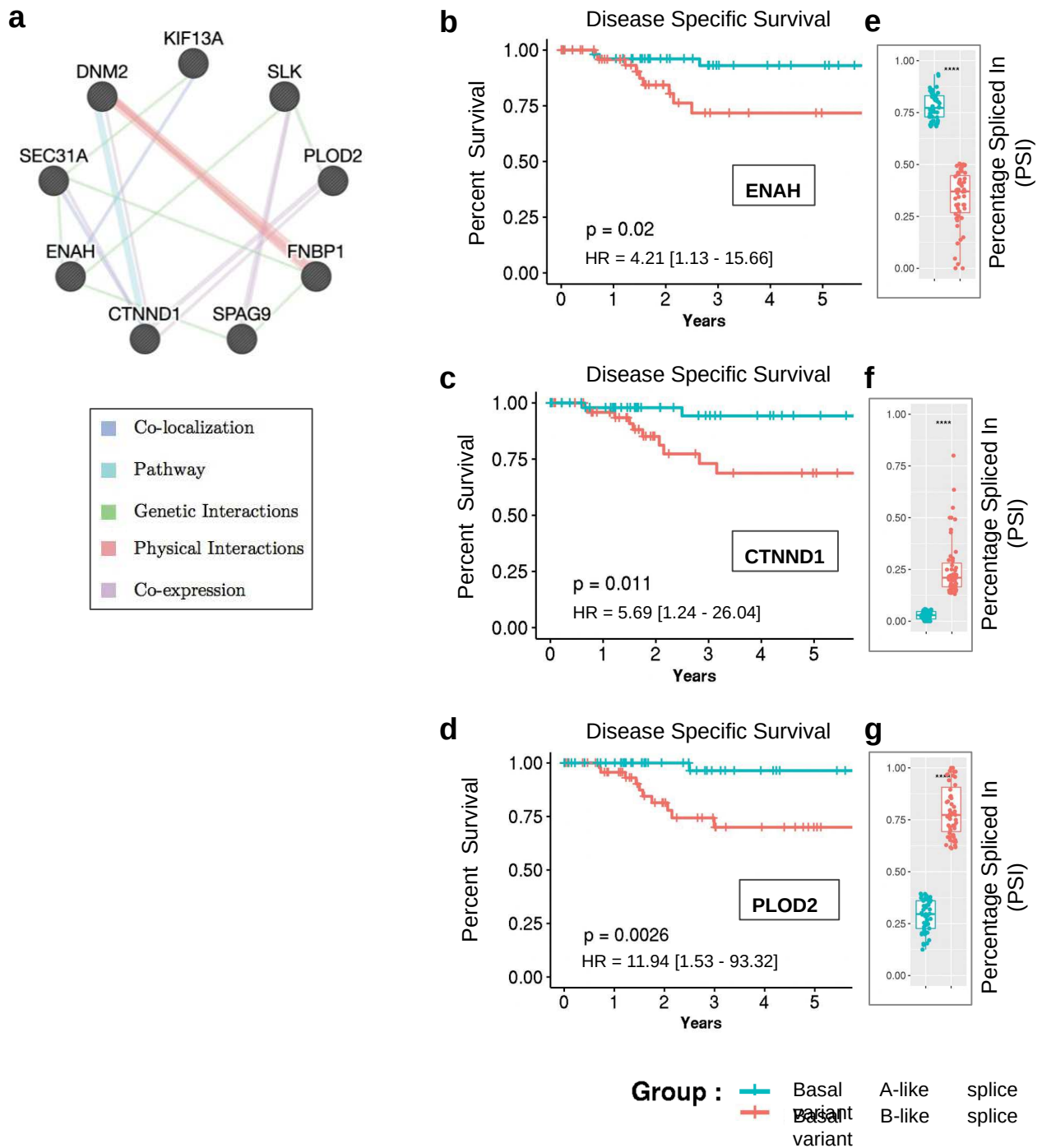


Figure S6. Prognostic value of individual alternatively spliced genes from the basal B-specific signature. **a.** Network of functional association (GeneMania) between RBM47-dependent spliced genes from the 25 basal B-specific splicing signature. **b,c,d.** Kaplan-Meier curves of disease specific survival 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.

USED IN MODEL				FOR VALIDATION				LUMINAL			
PRJNA523380 (1)	PRJNA523380 (1)	PRJNA297219 (2)	PRJNA297219 (2)	PRJNA210428(3)	PRJNA210428(3)	PRJNA251383 (4)	PRJNA251383 (4)	PRJEB30617(5)	PRJEB30617(5)	PRJNA523380 (1)	PRJNA297219 (2)
BT20	BASALA	BT20	BASALA	SUM225	BASALA	BT20	BASALA	BT20	BASALA	AU565	600MPE
CAL851	BASALA	CAL851	BASALA	HCC1143	BASALA	HCC1143	BASALA	HCC1187	BASALA	BT474	AU565
HCC1143	BASALA	HCC1143	BASALA	MX1	BASALA	HCC1187	BASALA	HCC1569	BASALA	BT483	BT474
HCC1187	BASALA	HCC1187	BASALA	HCC70	BASALA	HCC1569	BASALA	HCC1806	BASALA	CAL148	BT483
HCC1500	BASALA	HCC1569	BASALA	HCC1569	BASALA	HCC1599	BASALA	HCC1937	BASALA	CAMA1	CAL148
HCC1569	BASALA	HCC1599	BASALA	HCC1954	BASALA	HCC1937	BASALA	HCC1954	BASALA	EFM19	CAMA1
HCC1599	BASALA	HCC1806	BASALA	HCC3153	BASALA	HCC1954	BASALA	HCC70	BASALA	EFM192A	EFM19
HCC1806	BASALA	HCC1937	BASALA	SUM149	BASALA	HCC70	BASALA	MDAMB468	BASALA	HCC1419	EFM192A
HCC1937	BASALA	HCC1954	BASALA	HCC1937	BASALA	MDAMB468	BASALA	HCC1500	BASALA	HCC1428	EVSAT
HCC1954	BASALA	HCC3153	BASALA	HCC1806	BASALA	SUM149	BASALA	BT549	BASALB	HCC202	HCC1008
HCC2157	BASALA	HCC70	BASALA	HCC1599	BASALA	BT549	BASALB	CAL120	BASALB	HCC2218	HCC1419
HCC70	BASALA	HDQP1	BASALA	JIMT1	BASALA	MDAMB157	BASALB	CAL51	BASALB	KPL1	HCC1428
HDQP1	BASALA	JIMT1	BASALA	SUM229	BASALA	MDAMB231	BASALB	HCC1395	BASALB	MCF7	HCC202
JIMT1	BASALA	MACLS2	BASALA	BT549	BASALB	MDAMB436	BASALB	HS578T	BASALB	MDAMB134VI	HCC2185
MDAMB468	BASALA	MDAMB468	BASALA	MDAMB231	BASALB	SUM102	BASALB	MDAMB157	BASALB	MDAMB175VII	HCC2218
BT549	BASALB	MX1	BASALA	HCC38	BASALB	SUM159	BASALB	MDAMB231	BASALB	MDAMB361	HCC2688
CAL120	BASALB	SUM149	BASALA	HS578T	BASALB	HCC38	BASALB	MDAMB436	BASALB	MDAMB415	HCC712
CAL51	BASALB	SUM229	BASALA	MB157	BASALB			SKBR7	BASALB	MDAMB453	KPL1
DU4475	BASALB	MB157	BASALB	HCC1395	BASALB			HCC38	BASALB	SKBR3	LY2
HCC1395	BASALB	BT549	BASALB	SUM1315	BASALB					T47D	MCF7
HCC38	BASALB	CAL120	BASALB							UACC812	MDAMB134VI
HMC18	BASALB	CAL51	BASALB							UACC893	MDAMB175VII
HS578T	BASALB	DU4475	BASALB							ZR751	MDAMB330
MDAMB157	BASALB	HBL100	BASALB							ZR7530	MDAMB361
MDAMB231	BASALB	HCC1395	BASALB								MDAMB415
MDAMB436	BASALB	HCC38	BASALB								MDAMB453
		HS578T	BASALB								MFM223
		MDAMB157	BASALB								OCUBM
		MDAMB231	BASALB								SKBR3
		MDAMB436	BASALB								SKBR5
		SKBR7	BASALB								SUM185
		SUM102	BASALB								SUM190
		SUM1315	BASALB								SUM225
		SUM159	BASALB								SUM44
		SW527	BASALB								SUM52
		UACC3199	BASALB								T47D
											UACC812
											UACC893
											ZR751
											ZR7530
											ZR75B

Table S1

Table S2

ID	Coords
PLOD2	chr3:146077861-146077924
DNM2_a	chr19:10796060-10796199
CTNND1_b	chr11:57791493-57791673
SPAG9	chr17:50975862-50975901
RUBCN	chr3:197691073-197691148
CTNND1_a	chr11:57789036-57789155
EVI5L	chr19:7857091-7857124
DNM2_b	chr19:10808568-10808580
CAST	chr5:96726793-96726859
FAT1	chr4:186590367-186590403
TSC2	chr16:2077597-2077726
FNBP1	chr9:129915965-129915980
BNIP2	chr15:59668102-59668138
CSF1	chr1:109923165-109923711
ANXA6	chr5:151110626-151110644
ARHGEF11	chr1:156938417-156938513
ENAH	chr1:225504990-225505053
AP1B1	chr22:29329711-29329720
ATP5C1	chr10:7806973-7807010
DNM1	chr9:128247923-128247935
CD44	chr11:35209964-35210054
MBNL1	chr3:152446703-152446757
SEC31A	chr4:82842184-82842481
KIF13A	chr6:177711113-17771218
SLK	chr10:104010815-104010908

3. GENERAL DISCUSSION AND PERSPECTIVE

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 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.

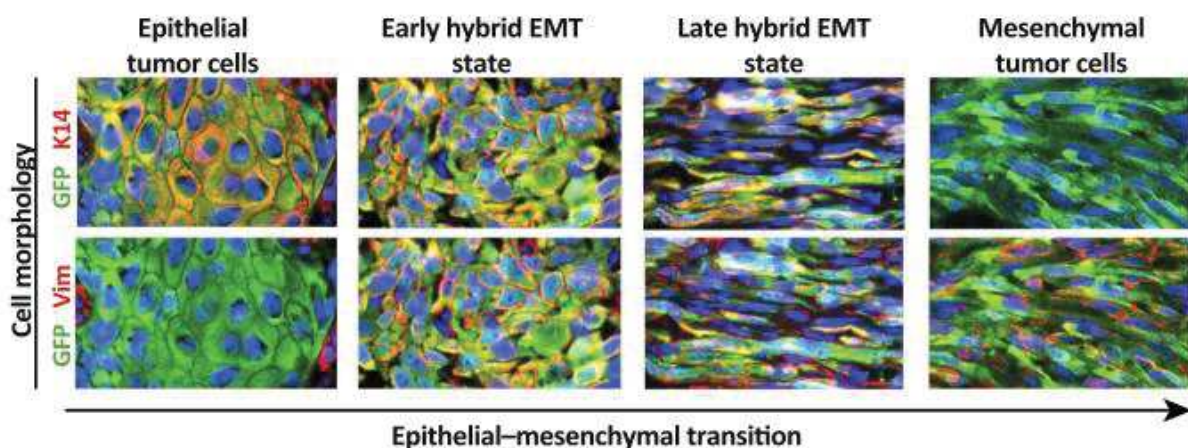


Figure 3-1 Transition states occurring during EMT

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 accessed under specific request guidelines. Now, it's not difficult to deliver a pre-trained model that can be easily ported 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 types 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).

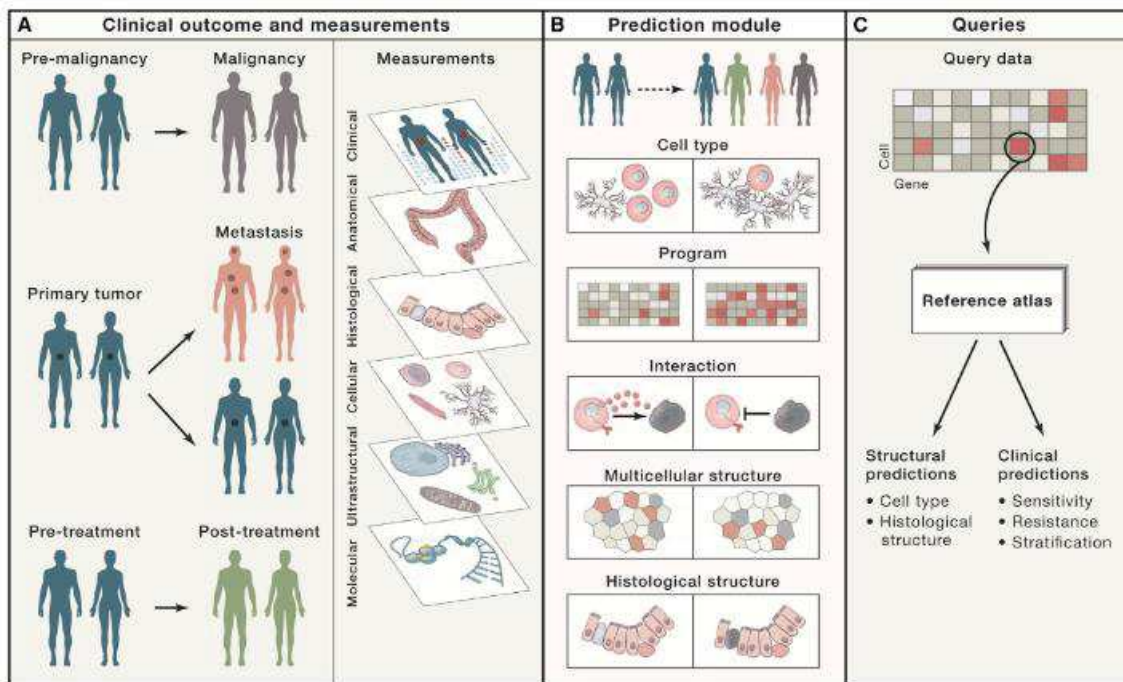


Figure 3-2 The Human Tumor Atlas Network (HTAN)

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)

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.

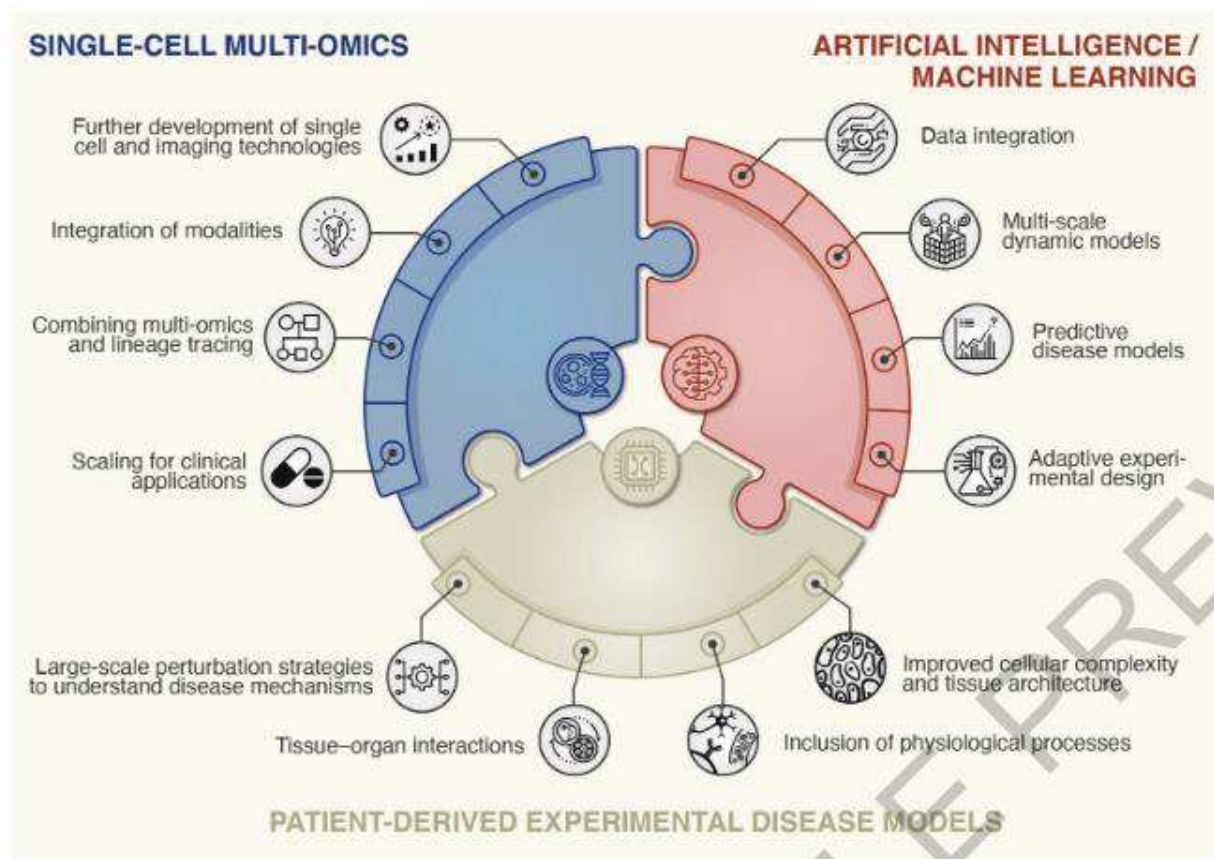


Figure 3-3 LifeTime European Project

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)

BIBLIOGRAPHY

- Abba, Mohammed, Nitin Patil, Jörg Leupold, and Heike Allgayer. 2016. “MicroRNA Regulation of Epithelial to Mesenchymal Transition.” *Journal of Clinical Medicine*. <https://doi.org/10.3390/jcm5010008>.
- Acloque, Hervé, Meghan S. Adams, Katherine Fishwick, Marianne Bronner-Fraser, and M. Angela Nieto. 2009. “Epithelial-Mesenchymal Transitions: The Importance of Changing Cell State in Development and Disease.” *Journal of Clinical Investigation*. <https://doi.org/10.1172/JCI38019>.
- Aghdassi, Ali, Matthias Sendler, Annett Guenther, Julia Mayerle, Claas Olsen Behn, Claus Dieter Heidecke, Helmut Friess, et al. 2012. “Recruitment of Histone Deacetylases HDAC1 and HDAC2 by the Transcriptional Repressor ZEB1 Downregulates E-Cadherin Expression in Pancreatic Cancer.” *Gut*. <https://doi.org/10.1136/gutjnl-2011-300060>.
- Aiello, Nicole M., Ravikanth Maddipati, Robert J. Norgard, David Balli, Jinyang Li, Salina Yuan, Taiji Yamazoe, et al. 2018. “EMT Subtype Influences Epithelial Plasticity and Mode of Cell Migration.” *Developmental Cell* 45 (6): 681-695.e4. <https://doi.org/10.1016/j.devcel.2018.05.027>.
- Alamancos, Gael P., Eneritz Agirre, and Eduardo Eyra. 2014. “Methods to Study Splicing from High-Throughput RNA Sequencing Data.” *Methods in Molecular Biology*. https://doi.org/10.1007/978-1-62703-980-2_26.
- Ali, H. Raza, Oscar M. Rueda, Suet Feung Chin, Christina Curtis, Mark J. Dunning, Samuel A.J.R. Aparicio, and Carlos Caldas. 2014. “Genome-Driven Integrated Classification of Breast Cancer Validated in over 7,500 Samples.” *Genome Biology* 15 (8): 1–14. <https://doi.org/10.1186/s13059-014-0431-1>.
- Alluri, Prasanna, and Lisa A. Newman. 2014. “Basal-like and Triple-Negative Breast Cancers. Searching for Positives among Many Negatives.” *Surgical Oncology Clinics of North America*. <https://doi.org/10.1016/j.soc.2014.03.003>.
- American Cancer Society. 2018. “Estimated Number of New Cancer Cases, All EU Countries, 2018.” https://doi.org/10.1787/health_glance_eur-2018-graph47-en.
- Anczuków, Olga, Martin Akerman, Antoine Cléry, Jie Wu, Chen Shen, Nitin H. Shirole, Amanda Raimer, et al. 2015. “SRSF1-Regulated Alternative Splicing in Breast Cancer.” *Molecular Cell* 60 (1): 105–17. <https://doi.org/10.1016/j.molcel.2015.09.005>.
- Anczuków, Olga, Avi Z. Rosenberg, Martin Akerman, Shipra Das, Lixing Zhan, Rotem Karni, Senthil K. Muthuswamy, and Adrian R. Krainer. 2012. “The Splicing Factor SRSF1 Regulates Apoptosis and Proliferation to Promote Mammary Epithelial Cell Transformation.” *Nature Structural and Molecular Biology*. <https://doi.org/10.1038/nsmb.2207>.
- Anders, Simon, and Wolfgang Huber. 2010. “Differential Expression Analysis for Sequence Count Data.” *Genome Biology*. <https://doi.org/10.1186/gb-2010-11-10-r106>.
- Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber. 2015. “HTSeq-A Python Framework to Work with High-Throughput Sequencing Data.” *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btu638>.
- Anders, Simon, Alejandro Reyes, and Wolfgang Huber. 2012. “Detecting Differential Usage of Exons from RNA-Seq Data.” *Genome Research*.

<https://doi.org/10.1101/gr.133744.111>.

- Audoux, Jérôme, Nicolas Philippe, Rayan Chikhi, Mikaël Salson, Mélina Gallopin, Marc Gabriel, Jérémy Le Coz, Emilie Drouineau, Thérèse Commes, and Daniel Gautheret. 2017. “DE-Kupl: Exhaustive Capture of Biological Variation in RNA-Seq Data through k-Mer Decomposition.” *Genome Biology*. <https://doi.org/10.1186/s13059-017-1372-2>.
- Augello, Michael A, Lisa D Berman-Booty, Richard Carr, Akihiro Yoshida, Jeffrey L Dean, Matthew J Schiewer, Felix Y Feng, et al. 2015. “Consequence of the Tumor-associated Conversion to Cyclin D1b.” *EMBO Molecular Medicine*. <https://doi.org/10.15252/emmm.201404242>.
- Avsec, Žiga, Roman Kreuzhuber, Johnny Israeli, Nancy Xu, Jun Cheng, Avanti Shrikumar, Abhimanyu Banerjee, et al. 2019. “The Kipoi Repository Accelerates Community Exchange and Reuse of Predictive Models for Genomics.” *Nature Biotechnology*. <https://doi.org/10.1038/s41587-019-0140-0>.
- Barash, Yoseph, John A. Calarco, Weijun Gao, Qun Pan, Xinchun Wang, Ofer Shai, Benjamin J. Blencowe, and Brendan J. Frey. 2010. “Deciphering the Splicing Code.” *Nature*. <https://doi.org/10.1038/nature09000>.
- Barrallo-Gimeno, Alejandro, and M. Angela Nieto. 2005. “The Snail Genes as Inducers of Cell Movement and Survival: Implications in Development and Cancer.” *Development*. <https://doi.org/10.1242/dev.01907>.
- Barretina, Jordi, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A. Margolin, Sungjoon Kim, Christopher J. Wilson, et al. 2012. “The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity.” *Nature* 483 (7391): 603–7. <https://doi.org/10.1038/nature11003>.
- Basu, Amrita, Nicole E. Bodycombe, Jaime H. Cheah, Edmund V. Price, Ke Liu, Giannina I. Schaefer, Richard Y. Ebright, et al. 2013. “An Interactive Resource to Identify Cancer Genetic and Lineage Dependencies Targeted by Small Molecules.” *Cell*. <https://doi.org/10.1016/j.cell.2013.08.003>.
- Bates, David O., Tai Gen Cui, Joanne M. Doughty, Matthias Winkler, Marto Sugiono, Jacqueline D. Shields, Danielle Peat, David Gillatt, and Steven J. Harper. 2002. “VEGF165b, an Inhibitory Splice Variant of Vascular Endothelial Growth Factor, Is down-Regulated in Renal Cell Carcinoma.” *Cancer Research*.
- Battle, Eduard, Elena Sancho, Clara Francí, David Domínguez, Mercè Monfar, Josep Baulida, and Antonio García De Herrerros. 2000. “The Transcription Factor Snail Is a Repressor of E-Cadherin Gene Expression in Epithelial Tumour Cells.” *Nature Cell Biology*. <https://doi.org/10.1038/35000034>.
- Bechara, Elias G., Endre Sebestyén, Isabella Bernardis, Eduardo Eyra, and Juan Valcárcel. 2013. “RBM5, 6, and 10 Differentially Regulate NUMB Alternative Splicing to Control Cancer Cell Proliferation.” *Molecular Cell* 52 (5): 720–33. <https://doi.org/10.1016/j.molcel.2013.11.010>.
- Bedi, Upasana, Vivek Kumar Mishra, David Wasilewski, Christina Scheel, and Steven A. Johnsen. 2014. “Epigenetic Plasticity: A Central Regulator of Epithelial-Tomesenchymal Transition in Cancer.” *Oncotarget*. <https://doi.org/10.18632/oncotarget.1875>.
- 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>.
- Bertero, Michela G, Christoph Bock, Annelien L Bredenoord, Giacomo Cavalli, Susanna Chiocca, Hans Clevers, Bart De Strooper, et al. 2020. “LifeTime and Improving European Healthcare through Cell-Based Interceptive Medicine.” *Nature*. <https://doi.org/10.1038/s41586-020-2715-9>.
- Bjørklund, Sunniva Stordal, Anshuman Panda, Surendra Kumar, Michael Seiler, Doug Robinson, Jinesh Gheeya, Ming Yao, et al. 2017. “Widespread Alternative Exon Usage in Clinically Distinct Subtypes of Invasive Ductal Carcinoma.” *Scientific Reports* 7 (1): 1–15. <https://doi.org/10.1038/s41598-017-05537-0>.
- Black, Douglas L. 2003. “Mechanisms of Alternative Pre-Messenger RNA Splicing.” *Annual Review of Biochemistry*. <https://doi.org/10.1146/annurev.biochem.72.121801.161720>.
- Blencowe, Benjamin J. 2006. “Alternative Splicing: New Insights from Global Analyses.” *Cell*. <https://doi.org/10.1016/j.cell.2006.06.023>.
- . 2017. “The Relationship between Alternative Splicing and Proteomic Complexity.” *Trends in Biochemical Sciences*. <https://doi.org/10.1016/j.tibs.2017.04.001>.
- Boise, Lawrence H., Maribel González-García, Christina E. Postema, Liyun Ding, Tullia Lindsten, Laurence A. Turka, Xiaohong Mao, Gabriel Nuñez, and Craig B. Thompson. 1993. “Bcl-x, a Bcl-2-Related Gene That Functions as a Dominant Regulator of Apoptotic Cell Death.” *Cell*. [https://doi.org/10.1016/0092-8674\(93\)90508-N](https://doi.org/10.1016/0092-8674(93)90508-N).
- Brabletz, Simone, Karolina Bajdak, Simone Meidhof, Ulrike Burk, Gabriele Niedermann, Elke Firat, Ulrich Wellner, et al. 2011. “The ZEB1/MiR-200 Feedback Loop Controls Notch Signalling in Cancer Cells.” *EMBO Journal*. <https://doi.org/10.1038/emboj.2010.349>.
- Brabletz, Simone, and Thomas Brabletz. 2010. “The ZEB/MiR-200 Feedback Loop—a Motor of Cellular Plasticity in Development and Cancer?” *EMBO Reports*. <https://doi.org/10.1038/embor.2010.117>.
- Brabletz, Thomas, Raghu Kalluri, M. Angela Nieto, and Robert A. Weinberg. 2018. “EMT in Cancer.” *Nature Reviews Cancer* 18 (2): 128–34. <https://doi.org/10.1038/nrc.2017.118>.
- Braeutigam, C., L. Rago, A. Rolke, L. Waldmeier, G. Christofori, and J. Winter. 2014. “The RNA-Binding Protein Rbfox2: An Essential Regulator of EMT-Driven Alternative Splicing and a Mediator of Cellular Invasion.” *Oncogene* 33 (9): 1082–92. <https://doi.org/10.1038/onc.2013.50>.
- Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. “Near-Optimal Probabilistic RNA-Seq Quantification.” *Nature Biotechnology*. <https://doi.org/10.1038/nbt.3519>.
- Brooks, Angela N., Li Yang, Michael O. Duff, Kasper D. Hansen, Jung W. Park, Sandrine Dudoit, Steven E. Brenner, and Brenton R. Graveley. 2011. “Conservation of an RNA Regulatory Map between *Drosophila* and Mammals.” *Genome Research*. <https://doi.org/10.1101/gr.108662.110>.
- Brown, Rhonda L., Lauren M. Reinke, Marin S. Damerow, Denise Perez, Lewis A. Chodosh, Jing Yang, and Chonghui Cheng. 2011. “CD44 Splice Isoform Switching in Human and Mouse Epithelium Is Essential for Epithelial-Mesenchymal Transition and Breast Cancer Progression.” *Journal of Clinical Investigation* 121 (3): 1064–74. <https://doi.org/10.1172/JCI44540>.

- Bullard, James H., Elizabeth Purdom, Kasper D. Hansen, and Sandrine Dudoit. 2010. "Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments." *BMC Bioinformatics*. <https://doi.org/10.1186/1471-2105-11-94>.
- Burstein, Matthew D., Anna Tsimelzon, Graham M. Poage, Kyle R. Covington, Alejandro Contreras, Suzanne A.W. Fuqua, Michelle I. Savage, et al. 2015. "Comprehensive Genomic Analysis Identifies Novel Subtypes and Targets of Triple-Negative Breast Cancer." *Clinical Cancer Research* 21 (7): 1688–98. <https://doi.org/10.1158/1078-0432.CCR-14-0432>.
- Byers, Lauren Averett, Lixia Diao, Jing Wang, Pierre Saintigny, Luc Girard, Michael Peyton, Li Shen, et al. 2013. "An Epithelial-Mesenchymal Transition Gene Signature Predicts Resistance to EGFR and PI3K Inhibitors and Identifies Axl as a Therapeutic Target for Overcoming EGFR Inhibitor Resistance." *Clinical Cancer Research*. <https://doi.org/10.1158/1078-0432.CCR-12-1558>.
- Caramel, Julie, Eftychios Papadogeorgakis, Louise Hill, Gareth J. Browne, Geoffrey Richard, Anne Wierinckx, Gerald Saldanha, et al. 2013. "A Switch in the Expression of Embryonic EMT-Inducers Drives the Development of Malignant Melanoma." *Cancer Cell*. <https://doi.org/10.1016/j.ccr.2013.08.018>.
- Carstens, Russ P., James V. Eaton, Hannah R. Krigman, Philip J. Walther, and Mariano A. Garcia-Blanco. 1997. "Alternative Splicing of Fibroblast Growth Factor Receptor 2 (FGF-R2) in Human Prostate Cancer." *Oncogene*. <https://doi.org/10.1038/sj.onc.1201498>.
- Carstens, Russ P., Eric J. Wagner, and Mariano A. Garcia-Blanco. 2000. "An Intronic Splicing Silencer Causes Skipping of the IIIb Exon of Fibroblast Growth Factor Receptor 2 through Involvement of Polypyrimidine Tract Binding Protein." *Molecular and Cellular Biology*. <https://doi.org/10.1128/mcb.20.19.7388-7400.2000>.
- Cerami, Ethan, Jianjiong Gao, Ugur Dogrusoz, Benjamin E. Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, et al. 2012. "The CBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data." *Cancer Discovery*. <https://doi.org/10.1158/2159-8290.CD-12-0095>.
- Chen, Chen, Shujie Zhao, Anand Karnad, and James W. Freeman. 2018. "The Biology and Role of CD44 in Cancer Progression: Therapeutic Implications." *Journal of Hematology and Oncology*. <https://doi.org/10.1186/s13045-018-0605-5>.
- Chen, Xueran, Chenggang Zhao, Bing Guo, Zhiyang Zhao, Hongzhi Wang, and Zhiyou Fang. 2019. "Systematic Profiling of Alternative mRNA Splicing Signature for Predicting Glioblastoma Prognosis." *Frontiers in Oncology*. <https://doi.org/10.3389/fonc.2019.00928>.
- Chen, Yao, Xiaoqin 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/s41598-017-00079-x>.
- Cheon, Dong Joo, and Sandra Orsulic. 2011. "Mouse Models of Cancer." *Annual Review of Pathology: Mechanisms of Disease*. <https://doi.org/10.1146/annurev.pathol.3.121806.154244>.
- Ching, Travers, Xun Zhu, and Lana X. Garmire. 2018. "Cox-Nnet: An Artificial Neural

- Network Method for Prognosis Prediction of High-Throughput Omics Data.” *PLoS Computational Biology*. <https://doi.org/10.1371/journal.pcbi.1006076>.
- Chollet, François. 2015. “Keras Documentation.” Keras.Io. 2015.
- Climente-González, Héctor, Eduard Porta-Pardo, Adam Godzik, and Eduardo Eyras. 2017. “The Functional Impact of Alternative Splicing in Cancer.” *Cell Reports* 20 (9): 2215–26. <https://doi.org/10.1016/j.celrep.2017.08.012>.
- Conesa, Ana, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michal Wojciech Szczesniak, et al. 2016. “A Survey of Best Practices for RNA-Seq Data Analysis.” *Genome Biology*. <https://doi.org/10.1186/s13059-016-0881-8>.
- Corsello, Steven M., Joshua A. Bittker, Zihan Liu, Joshua Gould, Patrick McCarren, Jodi E. Hirschman, Stephen E. Johnston, et al. 2017. “The Drug Repurposing Hub: A next-Generation Drug Library and Information Resource.” *Nature Medicine*. <https://doi.org/10.1038/nm.4306>.
- Cox, D. R. 1972. “Regression Models and Life-Tables.” *Journal of the Royal Statistical Society: Series B (Methodological)*. <https://doi.org/10.1111/j.2517-6161.1972.tb00899.x>.
- Craene, Bram De, and Geert Berx. 2013. “Regulatory Networks Defining EMT during Cancer Initiation and Progression.” *Nature Reviews Cancer* 13 (2): 97–110. <https://doi.org/10.1038/nrc3447>.
- Curtis, Christina, Sohrab P. Shah, Suet Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, et al. 2012. “The Genomic and Transcriptomic Architecture of 2,000 Breast Tumours Reveals Novel Subgroups.” *Nature* 486 (7403): 346–52. <https://doi.org/10.1038/nature10983>.
- Danan-Gotthold, Miri, Regina Golan-Gerstl, Eli Eisenberg, Keren Meir, Rotem Karni, and Erez Y. Levanon. 2015. “Identification of Recurrent Regulated Alternative Splicing Events across Human Solid Tumors.” *Nucleic Acids Research* 43 (10): 5130–44. <https://doi.org/10.1093/nar/gkv210>.
- Davis, Michael A., Renee C. Ireton, and Albert B. Reynolds. 2003. “A Core Function for P120-Catenin in Cadherin Turnover.” *Journal of Cell Biology*. <https://doi.org/10.1083/jcb.200307111>.
- Denecker, G., N. Vandamme, Ö Akay, D. Koludrovic, J. Taminau, K. Lemeire, A. Gheldof, et al. 2014. “Identification of a ZEB2-MITF-ZEB1 Transcriptional Network That Controls Melanogenesis and Melanoma Progression.” *Cell Death and Differentiation*. <https://doi.org/10.1038/cdd.2014.44>.
- Derynck, Rik, and Robert A. Weinberg. 2019. “EMT and Cancer: More Than Meets the Eye.” *Developmental Cell* 49 (3): 313–16. <https://doi.org/10.1016/j.devcel.2019.04.026>.
- Dewaele, Michael, Tommaso Tabaglio, Karen Willekens, Marco Bezzi, Shun Xie Teo, Diana H.P. Low, Cheryl M. Koh, et al. 2016. “Antisense Oligonucleotide-Mediated MDM4 Exon 6 Skipping Impairs Tumor Growth.” *Journal of Clinical Investigation*. <https://doi.org/10.1172/JCI82534>.
- Dillekås, Hanna, Michael S. Rogers, and Oddbjørn Straume. 2019. “Are 90% of Deaths from Cancer Caused by Metastases?” *Cancer Medicine* 8 (12): 5574–76. <https://doi.org/10.1002/cam4.2474>.
- Dittmar, K. A., P. Jiang, J. W. Park, K. Amirikian, J. Wan, S. Shen, Y. Xing, and R. P.

- Carstens. 2012. "Genome-Wide Determination of a Broad ESRP-Regulated Posttranscriptional Network by High-Throughput Sequencing." *Molecular and Cellular Biology* 32 (8): 1468–82. <https://doi.org/10.1128/mcb.06536-11>.
- Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/bts635>.
- Dominguez, Daniel, Peter Freese, Maria S. Alexis, Amanda Su, Myles Hochman, Tsultrim Palden, Cassandra Bazile, et al. 2018. "Sequence, Structure, and Context Preferences of Human RNA Binding Proteins." *Molecular Cell* 70 (5): 854-867.e9. <https://doi.org/10.1016/j.molcel.2018.05.001>.
- Domínguez, David, Bàrbara Montserrat-Sentís, Ariadna Virgós-Soler, Sandra Guaita, Judit Grueso, Montserrat Porta, Isabel Puig, Josep Baulida, Clara Francí, and Antonio García de Herreros. 2003. "Phosphorylation Regulates the Subcellular Location and Activity of the Snail Transcriptional Repressor." *Molecular and Cellular Biology*. <https://doi.org/10.1128/mcb.23.14.5078-5089.2003>.
- Dong, C., Y. Wu, Y. Wang, C. Wang, T. Kang, P. G. Rychahou, Y. I. Chi, B. M. Evers, and B. P. Zhou. 2013. "Interaction with Suv39H1 Is Critical for Snail-Mediated E-Cadherin Repression in Breast Cancer." *Oncogene*. <https://doi.org/10.1038/onc.2012.169>.
- Dongre, Anushka, and Robert A. Weinberg. 2018. "New Insights into the Mechanisms of Epithelial–Mesenchymal Transition and Implications for Cancer." *Nature Reviews Molecular Cell Biology*, 1. <https://doi.org/10.1038/s41580-018-0080-4>.
- Du, Hongzhi, Mao Pang, Xiaoying Hou, Shengtao Yuan, and Li Sun. 2017. "PLOD2 in Cancer Research." *Biomedicine and Pharmacotherapy*. <https://doi.org/10.1016/j.biopha.2017.04.023>.
- Durai, Dilip A., and Marcel H. Schulz. 2016. "Informed Kmer Selection for de Novo Transcriptome Assembly." *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btw217>.
- Dutertre, Martin, Stephan Vagner, and Didier Auboeuf. 2010. "Alternative Splicing and Breast Cancer." *RNA Biology* 7 (4): 403–11. <https://doi.org/10.4161/rna.7.4.12152>.
- Dvinge, Heidi, Eunhee Kim, Omar Abdel-Wahab, and Robert K. Bradley. 2016. "RNA Splicing Factors as Oncoproteins and Tumour Suppressors." *Nature Reviews Cancer* 16 (7): 413–30. <https://doi.org/10.1038/nrc.2016.51>.
- Eckert, Mark A., Thinzar M. Lwin, Andrew T. Chang, Jihoon Kim, Etienne Danis, Lucila Ohno-Machado, and Jing Yang. 2011. "Twist1-Induced Invadopodia Formation Promotes Tumor Metastasis." *Cancer Cell*. <https://doi.org/10.1016/j.ccr.2011.01.036>.
- Eger, Andreas, Kirsten Aigner, Stefan Sonderegger, Brigitta Dampier, Susanne Oehler, Martin Schreiber, Geert Berx, Amparo Cano, Hartmut Beug, and Roland Foisner. 2005. "DeltaEF1 Is a Transcriptional Repressor of E-Cadherin and Regulates Epithelial Plasticity in Breast Cancer Cells." *Oncogene*. <https://doi.org/10.1038/sj.onc.1208429>.
- Enkhbaatar, Zanabazar, Minoru Terashima, Dulamsuren Oktyabri, Shoichiro Tange, Akihiko Ishimura, Seiji Yano, and Takeshi Suzuki. 2013. "KDM5B Histone Demethylase Controls Epithelial-Mesenchymal Transition of Cancer Cells by Regulating the Expression of the MicroRNA-200 Family." *Cell Cycle*. <https://doi.org/10.4161/cc.25142>.
- Epifano, Carolina, Diego Megias, and Mirna Perez-Moreno. 2014. "P120-Catenin

- Differentially Regulates Cell Migration by Rho-Dependent Intracellular and Secreted Signals.” *EMBO Reports*. <https://doi.org/10.1002/embr.201337868>.
- Eraslan, Gökcen, Žiga Avsec, Julien Gagneur, and Fabian J. Theis. 2019. “Deep Learning: New Computational Modelling Techniques for Genomics.” *Nature Reviews Genetics* 20 (7): 389–403. <https://doi.org/10.1038/s41576-019-0122-6>.
- Fang, Min, Junrong Wu, Xin Lai, Huaying Ai, Yifeng Tao, Bo Zhu, and Lingsha Huang. 2016. “CD44 and CD44v6 Are Correlated with Gastric Cancer Progression and Poor Patient Prognosis: Evidence from 42 Studies.” *Cellular Physiology and Biochemistry*. <https://doi.org/10.1159/000452570>.
- Fattet, Laurent, Hae Yun Jung, Mike W. Matsumoto, Brandon E. Aubol, Aditya Kumar, Joseph A. Adams, Albert C. Chen, et al. 2020. “Matrix Rigidity Controls Epithelial-Mesenchymal Plasticity and Tumor Metastasis via a Mechanoresponsive EPHA2/LYN Complex.” *Developmental Cell* 54 (3): 302-316.e7. <https://doi.org/10.1016/j.devcel.2020.05.031>.
- Fici, Pietro, Giulia Gallerani, Anne-Pierre Morel, Laura Mercatali, Toni Ibrahim, Emanuela Scarpi, Dino Amadori, Alain Puisieux, Michel Rigaud, and Francesco Fabbri. 2016. “Splicing Factor Ratio as an Index of Epithelial-Mesenchymal Transition and Tumor Aggressiveness in Breast Cancer.” *Oncotarget* 8 (2): 2423–36. <https://doi.org/10.18632/oncotarget.13682>.
- Fischer, Kari R., Anna Durrans, Sharrell Lee, Jianting Sheng, Fuhai Li, Stephen T.C. Wong, Hyejin Choi, et al. 2015. “Epithelial-to-Mesenchymal Transition Is Not Required for Lung Metastasis but Contributes to Chemoresistance.” *Nature*. <https://doi.org/10.1038/nature15748>.
- Foroutan, Momeneh, Dharmesh D. Bhuvu, Ruqian Lyu, Kristy Horan, Joseph Cursons, and Melissa J. Davis. 2018. “Single Sample Scoring of Molecular Phenotypes.” *BMC Bioinformatics* 19 (1): 1–10. <https://doi.org/10.1186/s12859-018-2435-4>.
- Fougner, Christian, Helga Bergholtz, Jens Henrik Norum, and Therese Sørli. 2020. “Re-Definition of Claudin-Low as a Breast Cancer Phenotype.” *Nature Communications* 11 (1): 756411. <https://doi.org/10.1038/s41467-020-15574-5>.
- Gentle, J. E., L. Kaufman, and P. J. Rousseuw. 1991. “Finding Groups in Data: An Introduction to Cluster Analysis.” *Biometrics*. <https://doi.org/10.2307/2532178>.
- George, Jason T., Mohit Kumar Jolly, Shengnan Xu, Jason A. Somarelli, and Herbert Levine. 2017. “Survival Outcomes in Cancer Patients Predicted by a Partial EMT Gene Expression Scoring Metric.” *Cancer Research* 77 (22): 6415–28. <https://doi.org/10.1158/0008-5472.CAN-16-3521>.
- Geyer, Felipe C., Fresia Pareja, Britta Weigelt, Emad Rakha, Ian O. Ellis, Stuart J. Schnitt, and Jorge S. Reis-Filho. 2017. “The Spectrum of Triple-Negative Breast Disease: High- and Low-Grade Lesions.” *American Journal of Pathology*. <https://doi.org/10.1016/j.ajpath.2017.03.016>.
- Ghandi, Mahmoud, Franklin W. Huang, Judit Jané-Valbuena, Gregory V. Kryukov, Christopher C. Lo, E. Robert McDonald, Jordi Barretina, et al. 2019. “Next-Generation Characterization of the Cancer Cell Line Encyclopedia.” *Nature*. <https://doi.org/10.1038/s41586-019-1186-3>.
- Ghigna, Claudia, Silvia Giordano, Haihong Shen, Federica Benvenuto, Fabio Castiglioni, Paolo Maria Comoglio, Michael R. Green, Silvano Riva, and Giuseppe Biamonti. 2005.

- “Cell Motility Is Controlled by SF2/ASF through Alternative Splicing of the Ron Protooncogene.” *Molecular Cell*. <https://doi.org/10.1016/j.molcel.2005.10.026>.
- Gil-Diez De Medina, Sixtina, Zivco Popov, Dominique K. Chopin, Jennifer Southgate, Gordon C. Tucker, Annie Delouvé, Jean Paul Thiery, and François Radvanyi. 1999. “Relationship between E-Cadherin and Fibroblast Growth Factor Receptor 2b Expression in Bladder Carcinomas.” *Oncogene*. <https://doi.org/10.1038/sj.onc.1202958>.
- Goecks, Jeremy, Vahid Jalili, Laura M. Heiser, and Joe W. Gray. 2020. “How Machine Learning Will Transform Biomedicine.” *Cell* 181 (1): 92–101. <https://doi.org/10.1016/j.cell.2020.03.022>.
- Gökmen-Polar, Yesim, Yaseswini Neelamraju, Chirayu P Goswami, Yuan Gu, Xiaoping Gu, Gouthami Nallamotheu, Edyta Vieth, Sarath C Janga, Michael Ryan, and Sunil S Badve. 2019. “Splicing Factor ESRP 1 Controls ER -positive Breast Cancer by Altering Metabolic Pathways.” *EMBO Reports*. <https://doi.org/10.15252/embr.201846078>.
- Gonzalez, Inma, Roberto Munita, Eneritz Agirre, Travis A. Dittmer, Katia Gysling, Tom Misteli, and Reini F. Luco. 2015. “A LncRNA Regulates Alternative Splicing via Establishment of a Splicing-Specific Chromatin Signature.” *Nature Structural and Molecular Biology*. <https://doi.org/10.1038/nsmb.3005>.
- Gonzalez, Veronica D., Nikolay Samusik, Tiffany J. Chen, Erica S. Savig, Nima Aghaeepour, David A. Quigley, Ying Wen Huang, et al. 2018. “Commonly Occurring Cell Subsets in High-Grade Serous Ovarian Tumors Identified by Single-Cell Mass Cytometry.” *Cell Reports*. <https://doi.org/10.1016/j.celrep.2018.01.053>.
- Gotoda, Takuji, Yasuhiro Matsumura, Hitoshi Kondo, Daizo Saitoh, Yasuhiro Shimada, Tomoo Kosuge, Yae Kanai, and Tadao Kakizoe. 1998. “Expression of CD44 Variants and Its Association with Survival in Pancreatic Cancer.” *Japanese Journal of Cancer Research*. <https://doi.org/10.1111/j.1349-7006.1998.tb00493.x>.
- Grosse-Wilde, Anne, Aymeric Fouquier D’Hérouël, Ellie McIntosh, Gökhan Ertaylan, Alexander Skupin, Rolf E. Kuestner, Antonio Del Sol, Kathie Anne Walters, and Sui Huang. 2015. “Stemness of the Hybrid Epithelial/Mesenchymal State in Breast Cancer and Its Association with Poor Survival.” *PLoS ONE* 10 (5): 1–28. <https://doi.org/10.1371/journal.pone.0126522>.
- Grosso, Ana Rita, Sandra Martins, and Maria Carmo-Fonseca. 2008. “The Emerging Role of Splicing Factors in Cancer.” *EMBO Reports*. <https://doi.org/10.1038/embor.2008.189>.
- Guo, Cao, Junli Ma, Ganlu Deng, Yanlin Qu, Ling Yin, Yiyi Li, Ying Han, Changjing Cai, Hong Shen, and Shan Zeng. 2017. “ZEB1 Promotes Oxaliplatin Resistance through the Induction of Epithelial - Mesenchymal Transition in Colon Cancer Cells.” *Journal of Cancer*. <https://doi.org/10.7150/jca.20952>.
- Guyot, Mélanie, and Gilles Pagès. 2015. “VEGF Splicing and the Role of VEGF Splice Variants: From Physiological-Pathological Conditions to Specific Pre-mRNA Splicing.” *Methods in Molecular Biology*. https://doi.org/10.1007/978-1-4939-2917-7_1.
- Hanahan, Douglas, and R. A. Weinberg. 2000. “The Hallmarks of Cancer.” *Cell* 100: 57–70.
- Hanahan, Douglas, and Robert A. Weinberg. 2011. “Hallmarks of Cancer: The next Generation.” *Cell* 144 (5): 646–74. <https://doi.org/10.1016/j.cell.2011.02.013>.
- Harbeck, Nadia, Frédérique Penault-Llorca, Javier Cortes, Michael Gnant, Nehmat Houssami, Philip Poortmans, Kathryn Ruddy, Janice Tsang, and Fatima Cardoso. 2019. “Breast Cancer.” *Nature Reviews Disease Primers* 5 (1). <https://doi.org/10.1038/s41572-019->

0111-2.

- Harvey, Samuel E., Yilin Xu, Xiaodan Lin, Xin D. Gao, Yushan Qiu, Jaegyo Ahn, Xinshu Xiao, and Chonghui Cheng. 2018. "Coregulation of Alternative Splicing by HnRNPM and ESRP1 during EMT." *Rna* 24 (10): 1326–38. <https://doi.org/10.1261/rna.066712.118>.
- Harvey, Samuel, Yilin Xu, Xiaodan Lin, Xin D Gao, Yushan Qiu, Jaegyo Ahn, Xinshu Xiao, and Chonghui Cheng. 2018. "Co-Regulation of Alternative Splicing by HnRNPM and ESRP1 during EMT." *BioRxiv*, 301267. <https://doi.org/10.1101/301267>.
- Havens, Mallory A., and Michelle L. Hastings. 2016. "Splice-Switching Antisense Oligonucleotides as Therapeutic Drugs." *Nucleic Acids Research* 44 (14): 6549–63. <https://doi.org/10.1093/nar/gkw533>.
- Hegele, Anna, Atanas Kamburov, Arndt Grossmann, Chrysovalantis Sourlis, Sylvia Wowro, Mareike Weimann, Cindy L. Will, Vlad Pena, Reinhard Lührmann, and Ulrich Stelzl. 2012. "Dynamic Protein-Protein Interaction Wiring of the Human Spliceosome." *Molecular Cell*. <https://doi.org/10.1016/j.molcel.2011.12.034>.
- Herranz, Nicolás, Diego Pasini, Víctor M. Díaz, Clara Francí, Arantxa Gutierrez, Natàlia Dave, Maria Escrivà, et al. 2008. "Polycomb Complex 2 Is Required for E-Cadherin Repression by the Snail1 Transcription Factor." *Molecular and Cellular Biology*. <https://doi.org/10.1128/mcb.00323-08>.
- Hill, Louise, Gareth Browne, and Eugene Tulchinsky. 2012. "ZEB/MiR-200 Feedback Loop: At the Crossroads of Signal Transduction in Cancer." *International Journal of Cancer*. <https://doi.org/10.1002/ijc.27708>.
- Hinohara, Kunihiro, and Kornelia Polyak. 2019. "Intratumoral Heterogeneity: More Than Just Mutations." *Trends in Cell Biology* xx (xx): 1–11. <https://doi.org/10.1016/j.tcb.2019.03.003>.
- Ho, Tin Kam. 1995. "Random Decision Forests." In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*. <https://doi.org/10.1109/ICDAR.1995.598994>.
- Hong, David, Razelle Kurzrock, Youngsoo Kim, Richard Woessner, Anas Younes, John Nemunaitis, Nathan Fowler, et al. 2015. "AZD9150, a next-Generation Antisense Oligonucleotide Inhibitor of STAT3 with Early Evidence of Clinical Activity in Lymphoma and Lung Cancer." *Science Translational Medicine*. <https://doi.org/10.1126/scitranslmed.aac5272>.
- Hong, Jun, Jian Zhou, Junjiang Fu, Tao He, Jun Qin, Li Wang, Lan Liao, and Jianming Xu. 2011. "Phosphorylation of Serine 68 of Twist1 by MAPKs Stabilizes Twist1 Protein and Promotes Breast Cancer Cell Invasiveness." *Cancer Research*. <https://doi.org/10.1158/0008-5472.CAN-10-2914>.
- Hovhannisyán, Ruben H., and Russ P. Carstens. 2005. "A Novel Intronic Cis Element, ISE/ISS-3, Regulates Rat Fibroblast Growth Factor Receptor 2 Splicing through Activation of an Upstream Exon and Repression of a Downstream Exon Containing a Noncanonical Branch Point Sequence." *Molecular and Cellular Biology*. <https://doi.org/10.1128/mcb.25.1.250-263.2005>.
- Hu, Xiaohui, Samuel E Harvey, Rong Zheng, Jingyi Lyu, Caitlin L Grzeskowiak, Emily Powell, Helen Piwnica-worms, Kenneth L Scott, and Chonghui Cheng. 2020. "The RNA-Binding Protein AKAP8 Suppresses Tumor Metastasis by Antagonizing EMT-

- Associated Alternative Splicing.” *Nature Communications*.
<https://doi.org/10.1038/s41467-020-14304-1>.
- Hu, Yin, Yan Huang, Ying Du, Christian F. Orellana, Darshan Singh, Amy R. Johnson, Anaïs Monroy, et al. 2013. “DiffSplice: The Genome-Wide Detection of Differential Splicing Events with RNA-Seq.” *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gks1026>.
- Humphreys, Benjamin D., Shuei Liong Lin, Akio Kobayashi, Thomas E. Hudson, Brian T. Nowlin, Joseph V. Bonventre, M. Todd Valerius, Andrew P. McMahon, and Jeremy S. Duffield. 2010. “Fate Tracing Reveals the Pericyte and Not Epithelial Origin of Myofibroblasts in Kidney Fibrosis.” *American Journal of Pathology*.
<https://doi.org/10.2353/ajpath.2010.090517>.
- Hutter, Carolyn, and Jean Claude Zenklusen. 2018. “The Cancer Genome Atlas: Creating Lasting Value beyond Its Data.” *Cell*. <https://doi.org/10.1016/j.cell.2018.03.042>.
- Ignatiadis, M, P Aftimos, F Rothe, D Venet, M Piccart, C Sotiriou, and Y Bareche. 2018. “Unravelling Triple-Negative Breast Cancer Molecular Heterogeneity Using an Integrative Multiomic Analysis.” *Annals of Oncology* 29 (4): 895–902.
<https://doi.org/10.1093/annonc/mdy024>.
- Imani, Saber, Chunli Wei, Jingliang Cheng, Md Asaduzzaman Khan, Shangyi Fu, Luquan Yang, Mousumi Tania, et al. 2017. “MicroRNA-34a Targets Epithelial to Mesenchymal Transition-inducing Transcription Factors (EMT-TFs) and Inhibits Breast Cancer Cell Migration and Invasion.” *Oncotarget*. <https://doi.org/10.18632/oncotarget.15214>.
- Ishii, Hiroki, Masao Saitoh, Kei Sakamoto, Tetsuo Kondo, Ryohei Katoh, Shota Tanaka, Mitsuyoshi Motizuki, Keisuke Masuyama, and Keiji Miyazawa. 2014. “Epithelial Splicing Regulatory Proteins 1 (ESRP1) and 2 (ESRP2) Suppress Cancer Cell Motility via Different Mechanisms.” *Journal of Biological Chemistry* 289 (40): 27386–99.
<https://doi.org/10.1074/jbc.M114.589432>.
- Itoh, Masahiko, Derek C. Radisky, Masaaki Hashiguchi, and Hiroyuki Sugimoto. 2017. “The Exon 38-Containing ARHGEF11 Splice Isoform Is Differentially Expressed and Is Required for Migration and Growth in Invasive Breast Cancer Cells.” *Oncotarget*.
<https://doi.org/10.18632/oncotarget.20985>.
- Iwano, Masayuki, David Plieth, Theodore M. Danoff, Chengsen Xue, Hirokazu Okada, and Eric G. Neilson. 2002. “Evidence That Fibroblasts Derive from Epithelium during Tissue Fibrosis.” *Journal of Clinical Investigation*. <https://doi.org/10.1172/JCI0215518>.
- Iwatsuki, Masaaki, Koshi Mimori, Takehiko Yokobori, Hideshi Ishi, Toru Beppu, Shoji Nakamori, Hideo Baba, and Masaki Mori. 2010. “Epithelial-Mesenchymal Transition in Cancer Development and Its Clinical Significance.” *Cancer Science*.
<https://doi.org/10.1111/j.1349-7006.2009.01419.x>.
- Izquierdo, José María, Nuria Majós, Sophie Bonnal, Concepción Martínez, Robert Castelo, Roderic Guigó, Daniel Bilbao, and Juan Valcárcel. 2005. “Regulation of Fas Alternative Splicing by Antagonistic Effects of TIA-1 and PTB on Exon Definition.” *Molecular Cell*. <https://doi.org/10.1016/j.molcel.2005.06.015>.
- Jackson, Hartland W., Jana R. Fischer, Vito R.T. Zanotelli, H. Raza Ali, Robert Mechera, Savas D. Soysal, Holger Moch, et al. 2020. “The Single-Cell Pathology Landscape of Breast Cancer.” *Nature*, no. October 2018. <https://doi.org/10.1038/s41586-019-1876-x>.
- Javid, Sarah, Jianmin Zhang, Endre Anderssen, Josh C. Black, Ben S. Wittner, Ken Tajima, David T. Ting, et al. 2013. “Dynamic Chromatin Modification Sustains Epithelial-

- Mesenchymal Transition Following Inducible Expression of Snail-1.” *Cell Reports* 5 (6): 1679–89. <https://doi.org/10.1016/j.celrep.2013.11.034>.
- Jeong, H. M., J. Han, S. H. Lee, H. J. Park, H. J. Lee, J. S. Choi, Y. M. Lee, Y. L. Choi, Y. K. Shin, and M. J. Kwon. 2017. “ESRP1 Is Overexpressed in Ovarian Cancer and Promotes Switching from Mesenchymal to Epithelial Phenotype in Ovarian Cancer Cells.” *Oncogenesis*. <https://doi.org/10.1038/oncsis.2017.87>.
- Jézéquel, Pascal, Olivier Kerdraon, Hubert Hondermarck, Catherine Guérin-Charbonnel, Hamza Lasla, Wilfried Gouraud, Jean Luc Canon, et al. 2019. “Identification of Three Subtypes of Triple-Negative Breast Cancer with Potential Therapeutic Implications.” *Breast Cancer Research* 21 (1): 1–14. <https://doi.org/10.1186/s13058-019-1148-6>.
- Jiang, Yi Zhou, Ding Ma, Chen Suo, Jinxiu Shi, Mengzhu Xue, Xin Hu, Yi Xiao, et al. 2019. “Genomic and Transcriptomic Landscape of Triple-Negative Breast Cancers: Subtypes and Treatment Strategies.” *Cancer Cell* 35 (3): 428-440.e5. <https://doi.org/10.1016/j.ccell.2019.02.001>.
- Jin, Xiaoxia, Yingze Wei, Yushan Liu, Xiaoyun Lu, Fei Ding, Jiatai Wang, and Shuyun Yang. 2019. “Resveratrol Promotes Sensitization to Doxorubicin by Inhibiting Epithelial-Mesenchymal Transition and Modulating SIRT1/ β -Catenin Signaling Pathway in Breast Cancer.” *Cancer Medicine*. <https://doi.org/10.1002/cam4.1993>.
- Joglekar, Anoushka, Andrey Prjibelski, Ahmed Mahfouz, Paul Collier, Susan Lin, Anna Katharina, Jordan Marrocco, et al. 2020. “Cell-Type, Single-Cell, and Spatial Signatures of Brain-Region Specific Splicing in Postnatal Development.” *BioRxiv*.
- Jolly, Mohit Kumar. 2015. “Implications of the Hybrid Epithelial/Mesenchymal Phenotype in Metastasis.” *Frontiers in Oncology* 5 (July): 1–19. <https://doi.org/10.3389/fonc.2015.00155>.
- Jolly, Mohit Kumar, Kathryn E. Ware, Shivee Gilja, Jason A. Somarelli, and Herbert Levine. 2017. “EMT and MET: Necessary or Permissive for Metastasis?” *Molecular Oncology* 11 (7): 755–69. <https://doi.org/10.1002/1878-0261.12083>.
- Kahles, André, Kjong-Van Lehmann, Nora C. Toussaint, Matthias Hüser, Stefan G. Stark, Timo Sachsenberg, Oliver Stegle, et al. 2018. “Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients.” *Cancer Cell* 0 (0): 1–14. <https://doi.org/10.1016/J.CCELL.2018.07.001>.
- Kahles, André, Cheng Soon Ong, Yi Zhong, and Gunnar Rätsch. 2016. “SplAdder: Identification, Quantification and Testing of Alternative Splicing Events from RNA-Seq Data.” *Bioinformatics* 32 (12): 1840–47. <https://doi.org/10.1093/bioinformatics/btw076>.
- Kalcheim, Chaya. 2015. “Epithelial–Mesenchymal Transitions during Neural Crest and Somite Development.” *Journal of Clinical Medicine*. <https://doi.org/10.3390/jcm5010001>.
- Kao, Jessica, Keyan Salari, Melanie Bocanegra, Yoon La Choi, Luc Girard, Jeet Gandhi, Kevin A. Kwei, et al. 2009. “Molecular Profiling of Breast Cancer Cell Lines Defines Relevant Tumor Models and Provides a Resource for Cancer Gene Discovery.” *PLoS ONE* 4 (7). <https://doi.org/10.1371/journal.pone.0006146>.
- Kaplan, E. L., and Paul Meier. 1958. “Nonparametric Estimation from Incomplete Observations.” *Journal of the American Statistical Association*. <https://doi.org/10.1080/01621459.1958.10501452>.
- Karni, Rotem, Elisa De Stanchina, Scott W. Lowe, Rahul Sinha, David Mu, and Adrian R.

- Krainer. 2007. "The Gene Encoding the Splicing Factor SF2/ASF Is a Proto-Oncogene." *Nature Structural and Molecular Biology*. <https://doi.org/10.1038/nsmb1209>.
- Katz, Yarden, Eric T. Wang, Edoardo M. Airoidi, and Christopher B. Burge. 2010. "Analysis and Design of RNA Sequencing Experiments for Identifying Isoform Regulation." *Nature Methods*. <https://doi.org/10.1038/nmeth.1528>.
- Kaunitz, Jonathan D. 2015. "The Discovery of PCR: ProCuRement of Divine Power." *Digestive Diseases and Sciences*. <https://doi.org/10.1007/s10620-015-3747-0>.
- Keirsebilck, Annick, Stefan Bonn , Katrien Staes, Jolanda Van Hengel, Friedel Nollet, Albert Reynolds, and Frans Van Roy. 1998. "Molecular Cloning of the Human P120(Ctn) Catenin Gene (CTNND1): Expression of Multiple Alternatively Spliced Isoforms." *Genomics*. <https://doi.org/10.1006/geno.1998.5325>.
- Kim, Jihoon, Bon Kyoung Koo, and Juergen A. Knoblich. 2020. "Human Organoids: Model Systems for Human Biology and Medicine." *Nature Reviews Molecular Cell Biology*. <https://doi.org/10.1038/s41580-020-0259-3>.
- Klattig, J., and C. Englert. 2007. "The M llerian Duct: Recent Insights into Its Development and Regression." *Sexual Development*. <https://doi.org/10.1159/000108929>.
- Koedoot, Esmee, Marcel Smid, John A. Foekens, John W.M. Martens, Sylvia E. Le D v dec, and Bob van de Water. 2019. "Co-Regulated Gene Expression of Splicing Factors as Drivers of Cancer Progression." *Scientific Reports* 9 (1): 1–13. <https://doi.org/10.1038/s41598-019-40759-4>.
- Kotsopoulos, Joanne. 2018. "BRCA Mutations and Breast Cancer Prevention." *Cancers*. <https://doi.org/10.3390/cancers10120524>.
- Krebs, Angela M., Julia Mitschke, Mar a Lasierra Losada, Otto Schmalhofer, Melanie Boerries, Hauke Busch, Martin Boettcher, et al. 2017. "The EMT-Activator Zeb1 Is a Key Factor for Cell Plasticity and Promotes Metastasis in Pancreatic Cancer." *Nature Cell Biology* 19 (5): 518–29. <https://doi.org/10.1038/ncb3513>.
- Kr ger, Cornelia, Alexander Afeyan, Jasmin Mraz, Elinor Ng Eaton, Ferenc Reinhardt, Yevgenia L. Khodor, Prathapan Thiru, et al. 2019. "Acquisition of a Hybrid E/M State Is Essential for Tumorigenicity of Basal Breast Cancer Cells." *Proceedings of the National Academy of Sciences* 116 (15): 7353–62. <https://doi.org/10.1073/pnas.1812876116>.
- Ladomery, Michael. 2013. "Aberrant Alternative Splicing Is Another Hallmark of Cancer." *International Journal of Cell Biology*. <https://doi.org/10.1155/2013/463786>.
- Lambert, Arthur W, Diwakar R Pattabiraman, and Robert A Weinberg. 2016. "Review Emerging Biological Principles of Metastasis." *Cell* 168 (4): 670–91. <https://doi.org/10.1016/j.cell.2016.11.037>.
- Lamouille, Samy, Jian Xu, and Rik Derynck. 2014. "Molecular Mechanisms of Epithelial-Mesenchymal Transition." *Nature Reviews Molecular Cell Biology*. <https://doi.org/10.1038/nrm3758>.
- Lapuk, Anna, Henry Marr, Lakshmi Jakkula, Helder Pedro, Sanchita Bhattacharya, Elizabeth Purdom, Zhi Hu, et al. 2010. "Exon-Level Microarray Analyses Identify Alternative Splicing Programs in Breast Cancer." *Molecular Cancer Research* 8 (7): 961–74. <https://doi.org/10.1158/1541-7786.MCR-09-0528>.
- Lebleu, Valerie S., Gangadhar Taduri, Joyce O'Connell, Yingqi Teng, Vesselina G. Cooke, Craig Woda, Hikaru Sugimoto, and Raghu Kalluri. 2013. "Origin and Function of

- Myofibroblasts in Kidney Fibrosis.” *Nature Medicine*. <https://doi.org/10.1038/nm.3218>.
- Lehmann, Brian D., Bojana Jovanović, Xi Chen, Monica V. Estrada, Kimberly N. Johnson, Yu Shyr, Harold L. Moses, Melinda E. Sanders, and Jennifer A. Pietenpol. 2016. “Refinement of Triple-Negative Breast Cancer Molecular Subtypes: Implications for Neoadjuvant Chemotherapy Selection.” *PloS One* 11 (6): e0157368. <https://doi.org/10.1371/journal.pone.0157368>.
- Lehmann, Brian D, Yu Shyr, Jennifer A Pietenpol, Brian D Lehmann, Joshua A Bauer, Xi Chen, Melinda E Sanders, A Bapsi Chakravarthy, Yu Shyr, and Jennifer A Pietenpol. 2011. “Identification of Human Triple-Negative Breast Cancer Subtypes and Preclinical Models for Selection of Targeted Therapies Find the Latest Version : Identification of Human Triple-Negative Breast Cancer Subtypes and Preclinical Models for Selection of Targ” 121 (7): 2750–67. <https://doi.org/10.1172/JCI45014.2750>.
- Levitin, Hanna Mendes, Jinzhou Yuan, and Peter A. Sims. 2018. “Single-Cell Transcriptomic Analysis of Tumor Heterogeneity.” *Trends in Cancer* 4 (4): 264–68. <https://doi.org/10.1016/j.trecan.2018.02.003>.
- Li, Ji, Peter S Choi, Christine L Chaffer, Katherine Labella, Justin H Hwang, Andrew O Giacomelli, Jong Wook Kim, et al. 2018. “An Alternative Splicing Switch in FLNB Promotes the Mesenchymal Cell State in Human Breast Cancer.” *ELife* 7: 1–28. <https://doi.org/10.7554/eLife.37184>.
- Li, Jun, Daniela M. Witten, Iain M. Johnstone, and Robert Tibshirani. 2012. “Normalization, Testing, and False Discovery Rate Estimation for RNA-Sequencing Data.” *Biostatistics*. <https://doi.org/10.1093/biostatistics/kxr031>.
- Li, Ling, Lisa Hobson, Laura Perry, Bethany Clark, Susan Heavey, Aiman Haider, Ashwin Sridhar, et al. 2020. “Targeting the ERG Oncogene with Splice-Switching Oligonucleotides as a Novel Therapeutic Strategy in Prostate Cancer.” *British Journal of Cancer*, no. June: 1–9. <https://doi.org/10.1038/s41416-020-0951-2>.
- Li, Qing Quan, Jing Da Xu, Wen Juan Wang, Xi Xi Cao, Qi Chen, Feng Tang, Zhong Qing Chen, Xiu Ping Liu, and Zu De Xu. 2009. “Twist1-Mediated Adriamycin-Induced Epithelial-Mesenchymal Transition Relates to Multidrug Resistance and Invasive Potential in Breast Cancer Cells.” *Clinical Cancer Research*. <https://doi.org/10.1158/1078-0432.CCR-08-2372>.
- Li, Tianbao, Qi Liu, Nick Garza, Steven Kornblau, and Victor X Jin. 2018. “Integrative Analysis Reveals Functional and Regulatory Roles of H3K79me2 in Mediating Alternative Splicing.” *Genome Medicine*, 1–11. <https://doi.org/10.1186/s13073-018-0538-1>.
- Li, Yongsheng, Nidhi Sahni, Rita Pancsa, Daniel J. McGrail, Juan Xu, Xu Hua, Jasmin Coulombe-Huntington, et al. 2017. “Revealing the Determinants of Widespread Alternative Splicing Perturbation in Cancer.” *Cell Reports* 21 (3): 798–812. <https://doi.org/10.1016/j.celrep.2017.09.071>.
- Liang, Xiangping, Dongpei Li, Shuilong Leng, and Xiao Zhu. 2020. “RNA-Based Pharmacotherapy for Tumors: From Bench to Clinic and Back.” *Biomedicine and Pharmacotherapy* 125 (December 2019): 109997. <https://doi.org/10.1016/j.biopha.2020.109997>.
- Lin, Jung Chun. 2018. “Therapeutic Applications of Targeted Alternative Splicing to Cancer Treatment.” *International Journal of Molecular Sciences* 19 (1).

- <https://doi.org/10.3390/ijms19010075>.
- Lin, Yiwei, Chenfang Dong, and Binhua Zhou. 2014. “Epigenetic Regulation of EMT: The Snail Story.” *Current Pharmaceutical Design*.
<https://doi.org/10.2174/13816128113199990512>.
- Liu, Jianfang, Tara Lichtenberg, Katherine A. Hoadley, Laila M. Poisson, Alexander J. Lazar, Andrew D. Cherniack, Albert J. Kovatich, et al. 2018. “An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics.” *Cell*.
<https://doi.org/10.1016/j.cell.2018.02.052>.
- Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology*.
<https://doi.org/10.1186/s13059-014-0550-8>.
- Low-Marchelli, Janine M., Veronica C. Ardi, Edward A. Vizcarra, Nico Van Rooijen, James P. Quigley, and Jing Yang. 2013. “Twist1 Induces CCL2 and Recruits Macrophages to Promote Angiogenesis.” *Cancer Research*. <https://doi.org/10.1158/0008-5472.CAN-12-0653>.
- Lu, Hezhe, Jianglan Liu, Shujing Liu, Jingwen Zeng, Deqiang Ding, Russ P. Carstens, Yusheng Cong, Xiaowei Xu, and Wei Guo. 2013. “Exo70 Isoform Switching upon Epithelial-Mesenchymal Transition Mediates Cancer Cell Invasion.” *Developmental Cell* 27 (5): 560–73. <https://doi.org/10.1016/j.devcel.2013.10.020>.
- Lu, Wei, and Yibin Kang. 2019. “Epithelial-Mesenchymal Plasticity in Cancer Progression and Metastasis.” *Developmental Cell* 49 (3): 361–74.
<https://doi.org/10.1016/j.devcel.2019.04.010>.
- Mak, Milena P., Pan Tong, Lixia Diao, Robert J. Cardnell, Don L. Gibbons, William N. William, Ferdinandos Skoulidis, et al. 2016. “A Patient-Derived, Pan-Cancer EMT Signature Identifies Global Molecular Alterations and Immune Target Enrichment Following Epithelial-to-Mesenchymal Transition.” *Clinical Cancer Research* 22 (3): 609–20. <https://doi.org/10.1158/1078-0432.CCR-15-0876>.
- Manekar, Swati C., and Shailesh R. Sathe. 2018. “A Benchmark Study of K-Mer Counting Methods for High-Throughput Sequencing.” *GigaScience*.
<https://doi.org/10.1093/gigascience/giy125>.
- Mani, Sendurai A., Wenjun Guo, Mai Jing Liao, Elinor Ng Eaton, Ayyakkannu Ayyanan, Alicia Y. Zhou, Mary Brooks, et al. 2008. “The Epithelial-Mesenchymal Transition Generates Cells with Properties of Stem Cells.” *Cell* 133 (4): 704–15.
<https://doi.org/10.1016/j.cell.2008.03.027>.
- Mapleson, Daniel, Gonzalo Garcia Accinelli, George Kettleborough, Jonathan Wright, and Bernardo J. Clavijo. 2017. “KAT: A K-Mer Analysis Toolkit to Quality Control NGS Datasets and Genome Assemblies.” *Bioinformatics*.
<https://doi.org/10.1093/bioinformatics/btw663>.
- Marangoni, Elisabetta, Anne Vincent-Salomon, Nathalie Auger, Armelle Degeorges, Franck Assayag, Patricia De Cremoux, Ludmilla De Plater, et al. 2007. “A New Model of Patient Tumor-Derived Breast Cancer Xenografts for Preclinical Assays.” *Clinical Cancer Research*. <https://doi.org/10.1158/1078-0432.CCR-07-0078>.
- Marchini, Sergio, Robert Fruscio, Luca Clivio, Luca Beltrame, Luca Porcu, Ilaria Fusco Nerini, Duccio Cavalieri, et al. 2013. “Resistance to Platinum-Based Chemotherapy Is Associated with Epithelial to Mesenchymal Transition in Epithelial Ovarian Cancer.”

- European Journal of Cancer*. <https://doi.org/10.1016/j.ejca.2012.06.026>.
- Markolovic, Suzana, Sarah E. Wilkins, and Christopher J. Schofield. 2015. "Protein Hydroxylation Catalyzed by 2-Oxoglutarate-Dependent Oxygenases." *Journal of Biological Chemistry*. <https://doi.org/10.1074/jbc.R115.662627>.
- Marusyk, Andriy, and Kornelia Polyak. 2010. "Tumor Heterogeneity: Causes and Consequences." *Biochimica et Biophysica Acta - Reviews on Cancer* 1805 (1): 105–17. <https://doi.org/10.1016/j.bbcan.2009.11.002>.
- McCorry, Amy M.B., Maurice B. Loughrey, Daniel B. Longley, Mark Lawler, and Philip D. Dunne. 2018. "Epithelial-to-Mesenchymal Transition Signature Assessment in Colorectal Cancer Quantifies Tumour Stromal Content Rather than True Transition." *Journal of Pathology*. <https://doi.org/10.1002/path.5155>.
- Milioli, Heloisa H., Inna Tishchenko, Carlos Riveros, Regina Berretta, and Pablo Moscato. 2017. "Basal-like Breast Cancer: Molecular Profiles, Clinical Features and Survival Outcomes." *BMC Medical Genomics* 10 (1): 1–17. <https://doi.org/10.1186/s12920-017-0250-9>.
- Millanes-Romero, Alba, Nicolás Herranz, Valentina Perrera, Ane Iturbide, Jordina Loubat-Casanovas, Jesús Gil, Thomas Jenuwein, Antonio García de Herreros, and Sandra Peiró. 2013. "Regulation of Heterochromatin Transcription by Snail1/LOXL2 during Epithelial-to-Mesenchymal Transition." *Molecular Cell*. <https://doi.org/10.1016/j.molcel.2013.10.015>.
- Miura, K., W. Fujibuchi, and M. Unno. 2012. "Splice Variants in Apoptotic Pathway." *Experimental Oncology*.
- Mogilevsky, Maxim, Odelia Shimshon, Saran Kumar, Adi Mogilevsky, Eli Keshet, Eylon Yavin, Florian Heyd, and Rotem Karni. 2018. "Modulation of MKNK2 Alternative Splicing by Splice-Switching Oligonucleotides as a Novel Approach for Glioblastoma Treatment." *Nucleic Acids Research* 46 (21): 11396–404. <https://doi.org/10.1093/nar/gky921>.
- Montemayor, Eric J., Adam Katolik, Nathaniel E. Clark, Alexander B. Taylor, Jonathan P. Schuermann, D. Joshua Combs, Richard Johnsson, et al. 2014. "Structural Basis of Lariat RNA Recognition by the Intron Debranching Enzyme Dbr1." *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gku725>.
- Mortazavi, Ali, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq." *Nature Methods*. <https://doi.org/10.1038/nmeth.1226>.
- Mount, Stephen M. 1982. "A Catalogue of Splice Junction Sequences." *Nucleic Acids Research*. <https://doi.org/10.1093/nar/10.2.459>.
- Müller, Sebastian, Fabien Sindikubwabo, Tatiana Cañeque, Anne Lafon, Antoine Versini, Bérangère Lombard, Damarys Loew, et al. 2020. "CD44 Regulates Epigenetic Plasticity by Mediating Iron Endocytosis." *Nature Chemistry*. <https://doi.org/10.1038/s41557-020-0513-5>.
- Nakajima, Yuji, Toshiyuki Yamagishi, Shigeru Hokari, and Hiroaki Nakamura. 2000. "Mechanisms Involved in Valvuloseptal Endocardial Cushion Formation in Early Cardiogenesis: Roles of Transforming Growth Factor (TGF)- β and Bone Morphogenetic Protein (BMP)." *Anatomical Record* 258 (2): 119–27. [https://doi.org/10.1002/\(SICI\)1097-0185\(20000201\)258:2<119::AID-AR1>3.0.CO;2-U](https://doi.org/10.1002/(SICI)1097-0185(20000201)258:2<119::AID-AR1>3.0.CO;2-U).

- Nakaya, Yukiko, and Guojun Sheng. 2008. "Epithelial to Mesenchymal Transition during Gastrulation: An Embryological View." *Development Growth and Differentiation*. <https://doi.org/10.1111/j.1440-169X.2008.01070.x>.
- Navaglia, F., P. Fogar, E. Greco, D. Basso, A. L. Stefani, S. Mazza, C. F. Zambon, et al. 2003. "CD44v10: An Antimetastatic Membrane Glycoprotein for Pancreatic Cancer." *International Journal of Biological Markers*. <https://doi.org/10.5301/JBM.2008.4459>.
- Nellore, Abhinav, Andrew E. Jaffe, Jean Philippe Fortin, Jos Alquicira-Hernandez, Leonardo Collado-Torres, Siruo Wang, Robert A. Phillips, et al. 2016. "Human Splicing Diversity and the Extent of Unannotated Splice Junctions across Human RNA-Seq Samples on the Sequence Read Archive." *Genome Biology* 17 (1): 1–14. <https://doi.org/10.1186/s13059-016-1118-6>.
- Neve, Richard M., Koei Chin, Jane Fridlyand, Jennifer Yeh, Frederick L. Baehner, Tea Fevr, Laura Clark, et al. 2006. "A Collection of Breast Cancer Cell Lines for the Study of Functionally Distinct Cancer Subtypes." *Cancer Cell* 10 (6): 515–27. <https://doi.org/10.1016/j.ccr.2006.10.008>.
- Ni, Beibei, Jun Hu, Dianke Chen, Li Li, Daici Chen, Jianping Wang, and Lei Wang. 2016. "Alternative Splicing of Spleen Tyrosine Kinase Differentially Regulates Colorectal Cancer Progression." *Oncology Letters*. <https://doi.org/10.3892/ol.2016.4858>.
- Nieto, M. Angela, Ruby Yun Y.J. Huang, Rebecca A A. Jackson, and Jean Paul P. Thiery. 2016. "Emt: 2016." *Cell* 166 (1): 21–45. <https://doi.org/10.1016/j.cell.2016.06.028>.
- Pan, Qun, Ofer Shai, Leo J. Lee, Brendan J. Frey, and Benjamin J. Blencowe. 2008. "Deep Surveying of Alternative Splicing Complexity in the Human Transcriptome by High-Throughput Sequencing." *Nature Genetics*. <https://doi.org/10.1038/ng.259>.
- Pandya, Sonali, and Richard G. Moore. 2011. "Breast Development and Anatomy." *Clinical Obstetrics and Gynecology*. <https://doi.org/10.1097/GRF.0b013e318207ffe9>.
- Park, Eddie, Zhicheng Pan, Zijun Zhang, Lan Lin, and Yi Xing. 2018. "The Expanding Landscape of Alternative Splicing Variation in Human Populations." <https://doi.org/10.1016/j.ajhg.2017.11.002>.
- Park, Sung Hee, Mattia Brugiolo, Martin Akerman, Shipra Das, Laura Urbanski, Adam Geier, Anil K. Kesarwani, et al. 2019. "Differential Functions of Splicing Factors in Mammary Transformation and Breast Cancer Metastasis." *Cell Reports* 29 (9): 2672-2688.e7. <https://doi.org/10.1016/j.celrep.2019.10.110>.
- Pastushenko, Ievgenia, and Cédric Blanpain. 2019. "EMT Transition States during Tumor Progression and Metastasis." *Trends in Cell Biology* 29 (3): 212–26. <https://doi.org/10.1016/j.tcb.2018.12.001>.
- Pastushenko, Ievgenia, Audrey Brisebarre, Alejandro Sifrim, Marco Fioramonti, Tatiana Revenco, Soufiane Boumahdi, Alexandra Van Keymeulen, et al. 2018. "Identification of the Tumour Transition States Occurring during EMT." *Nature*. <https://doi.org/10.1038/s41586-018-0040-3>.
- Paszke, Adam, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary Devito Facebook, A I Research, et al. 2017. "Automatic Differentiation in PyTorch." In *Advances in Neural Information Processing Systems*.
- Patro, Rob, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. 2017. "Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression." *Nature Methods*. <https://doi.org/10.1038/nmeth.4197>.

- Patro, Rob, Stephen M. Mount, and Carl Kingsford. 2014. "Sailfish Enables Alignment-Free Isoform Quantification from RNA-Seq Reads Using Lightweight Algorithms." *Nature Biotechnology*. <https://doi.org/10.1038/nbt.2862>.
- Pedregosa, Fabian, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research*.
- Peinado, Hector, Esteban Ballestar, Manel Esteller, and Amparo Cano. 2004. "Snail Mediates E-Cadherin Repression by the Recruitment of the Sin3A/Histone Deacetylase 1 (HDAC1)/HDAC2 Complex." *Molecular and Cellular Biology*. <https://doi.org/10.1128/mcb.24.1.306-319.2004>.
- Pereira, Carla, Daniel Ferreira, Nuno Mendes, Pedro L. Granja, Gabriela M. Almeida, and Carla Oliveira. 2020. "Expression of CD44V6-Containing Isoforms Influences Cisplatin Response in Gastric Cancer Cells." *Cancers*. <https://doi.org/10.3390/cancers12040858>.
- Perou, Charles M., Therese Sørile, Michael B. Eisen, Matt Van De Rijn, Stefanie S. Jeffrey, Christian A. Rees, Jonathan R. Pollack, et al. 2000. "Molecular Portraits of Human Breast Tumours." *Nature*. <https://doi.org/10.1038/35021093>.
- Perou, Charles M., Therese Sørilie, Michael B Eisen, Matt Van De Rijn, Stefanie S Jeffrey, Christian A Rees, Jonathan R Pollack, et al. 2000. "Molecular Portraits of Human Breast Tumours" 533 (May).
- Philipps, Amanda I., and Christopher I. Li. 2010. "Breast Cancer Epidemiology and Clinical Characteristics." *Breast Cancer Epidemiology*, 1–417. <https://doi.org/10.1007/978-1-4419-0685-4>.
- Pickett, Brandon D., Justin B. Miller, and Perry G. Ridge. 2017. "Kmer-SSR: A Fast and Exhaustive SSR Search Algorithm." *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btx538>.
- Pires, Vanessa Borges, Ricardo Simões, Kamel Mamchaoui, Célia Carvalho, and Maria Carmo-Fonseca. 2017. "Short (16-Mer) Locked Nucleic Acid Splice-Switching Oligonucleotides Restore Dystrophin Production in Duchenne Muscular Dystrophy Myotubes." *PLoS ONE* 12 (7): 1–13. <https://doi.org/10.1371/journal.pone.0181065>.
- Pommier, Roxane M, Amélien Sanlaville, Laurie Tonon, Janice Kielbassa, Emilie Thomas, Anthony Ferrari, Anne-sophie Sertier, et al. 2020. "Comprehensive Characterization of Claudin-Low Breast Tumors Reflects the Impact of the Cell-of-Origin on Cancer Evolution." *Nature Communications*, no. 2020: 1–12. <https://doi.org/10.1038/s41467-020-17249-7>.
- Prat, Aleix, Barbara Adamo, Maggie C.U. Cheang, Carey K. Anders, Lisa A. Carey, and Charles M. Perou. 2013. "Molecular Characterization of Basal-Like and Non-Basal-Like Triple-Negative Breast Cancer." *The Oncologist* 18 (2): 123–33. <https://doi.org/10.1634/theoncologist.2012-0397>.
- Prat, Aleix, Joel S. Parker, Olga Karginova, Cheng Fan, Chad Livasy, Jason I. Herschkowitz, Xiaping He, and Charles M. Perou. 2010. "Phenotypic and Molecular Characterization of the Claudin-Low Intrinsic Subtype of Breast Cancer." *Breast Cancer Research* 12 (5). <https://doi.org/10.1186/bcr2635>.
- Prochazka, Lubomir, Radek Tesarik, and Jaroslav Turanek. 2014. "Regulation of Alternative Splicing of CD44 in Cancer." *Cellular Signalling* 26 (10): 2234–39. <https://doi.org/10.1016/j.cellsig.2014.07.011>.

- Puisieux, Alain, Thomas Brabletz, and Julie Caramel. 2014. "Oncogenic Roles of EMT-Inducing Transcription Factors." *Nature Cell Biology* 16 (6): 488–94. <https://doi.org/10.1038/ncb2976>.
- Puram, Sidharth V., Itay Tirosh, Anuraag S. Parikh, Anoop P. Patel, Keren Yizhak, Shawn Gillespie, Christopher Rodman, et al. 2017. "Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer." *Cell* 171 (7): 1611-1624.e24. <https://doi.org/10.1016/j.cell.2017.10.044>.
- Qiu, Yushan, Jingyi Lyu, Mikayla Dunlap, Samuel E. Harvey, and Chonghui Cheng. 2020. "A Combinatorially Regulated RNA Splicing Signature Predicts Breast Cancer EMT States and Patient Survival." *Rna*, rna.074187.119. <https://doi.org/10.1261/rna.074187.119>.
- Ranieri, Danilo, Benedetta Rosato, Monica Nanni, Alessandra Magenta, Francesca Belleudi, and Maria Rosaria Torrisi. 2016. "Expression of the FGFR2 Mesenchymal Splicing Variant in Epithelial Cells Drives Epithelial-Mesenchymal Transition." *Oncotarget* 7 (5): 5440–60. <https://doi.org/10.18632/oncotarget.6706>.
- Ray, Debashish, Hilal Kazan, Kate B. Cook, Matthew T. Weirauch, Hamed S. Najafabadi, Xiao Li, Serge Gueroussov, et al. 2013. "A Compendium of RNA-Binding Motifs for Decoding Gene Regulation." *Nature*. <https://doi.org/10.1038/nature12311>.
- Richard, Geoffrey, Stéphane Dalle, Marie-Ambre Monet, Maud Ligier, Amélie Boespflug, Roxane M Pommier, Arnaud Fouchardière, et al. 2016. "ZEB 1-mediated Melanoma Cell Plasticity Enhances Resistance to MAPK Inhibitors ." *EMBO Molecular Medicine*. <https://doi.org/10.15252/emmm.201505971>.
- Rieke, Nicola, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, et al. 2020. "The Future of Digital Health with Federated Learning." *Npj Digital Medicine*. <https://doi.org/10.1038/s41746-020-00323-1>.
- Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. "Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies." *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkv007>.
- Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2009. "EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btp616>.
- Robinson, Mark D., and Alicia Oshlack. 2010. "A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data." *Genome Biology*. <https://doi.org/10.1186/gb-2010-11-3-r25>.
- Rokavec, Matjaz, Markus Kaller, David Horst, and Heiko Hermeking. 2017. "Pan-Cancer EMT-Signature Identifies RBM47 down-Regulation during Colorectal Cancer Progression." *Scientific Reports* 7 (1): 1–15. <https://doi.org/10.1038/s41598-017-04234-2>.
- Romeo, Elisabetta, Carmelo Antonio Caserta, Cristiano Rumio, and Fabrizio Marcucci. 2019. "The Vicious Cross-Talk between Tumor Cells with an EMT Phenotype and Cells of the Immune System." *Cells*. <https://doi.org/10.3390/cells8050460>.
- Romero-Barrios, Natali, Maria Florencia Legascue, Moussa Benhamed, Federico Ariel, and Martin Crespi. 2018. "Splicing Regulation by Long Noncoding RNAs." *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gky095>.

- Rozenblatt-Rosen, Orit, Aviv Regev, Philipp Oberdoerffer, Tal Nawy, Anna Hupalowska, Jennifer E. Rood, Orr Ashenberg, et al. 2020. “The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution.” *Cell* 181 (2): 236–49. <https://doi.org/10.1016/j.cell.2020.03.053>.
- Ruiz-Velasco, Mariana, Manjeet Kumar, Mang Ching Lai, Pooja Bhat, Ana Belen Solis-Pinson, Alejandro Reyes, Stefan Kleinsorg, Kyung Min Noh, Toby J. Gibson, and Judith B. Zaugg. 2017. “CTCF-Mediated Chromatin Loops between Promoter and Gene Body Regulate Alternative Splicing across Individuals.” *Cell Systems*. <https://doi.org/10.1016/j.cels.2017.10.018>.
- Sabatier, Renaud, Pascal Finetti, Arnaud Guille, José Adelaide, Max Chaffanet, Patrice Viens, Daniel Birnbaum, and François Bertucci. 2014. “Claudin-Low Breast Cancers: Clinical, Pathological, Molecular and Prognostic Characterization.” *Molecular Cancer* 13 (1): 1–14. <https://doi.org/10.1186/1476-4598-13-228>.
- Saitoh, Masao. 2018. “Involvement of Partial EMT in Cancer Progression.” *Journal of Biochemistry* 0 (June): 1–8. <https://doi.org/10.1093/jb/mvy047>.
- Sakurai, T., K. Isogaya, S. Sakai, M. Morikawa, Y. Morishita, S. Ehata, K. Miyazono, and D. Koinuma. 2016. “RNA-Binding Motif Protein 47 Inhibits Nrf2 Activity to Suppress Tumor Growth in Lung Adenocarcinoma.” *Oncogene* 35 (38): 5000–5009. <https://doi.org/10.1038/onc.2016.35>.
- Saldanha, R, G Mohr, M Belfort, and A M Lambowitz. 1993. “Group I and Group II Introns.” *The FASEB Journal*. <https://doi.org/10.1096/fasebj.7.1.8422962>.
- Samatov, Timur R., Alexander G. Tonevitsky, and Udo Schumacher. 2013. “Epithelial-Mesenchymal Transition: Focus on Metastatic Cascade, Alternative Splicing, Non-Coding RNAs and Modulating Compounds.” *Molecular Cancer* 12 (1): 1. <https://doi.org/10.1186/1476-4598-12-107>.
- Sanidas, Ioannis, Christos Polytarchou, Maria Hatzia Apostolou, Scott A. Ezell, Filippos Kottakis, Lan Hu, Ailan Guo, et al. 2014. “Phosphoproteomics Screen Reveals Akt Isoform-Specific Signals Linking RNA Processing to Lung Cancer.” *Molecular Cell*. <https://doi.org/10.1016/j.molcel.2013.12.018>.
- Santonja, Angela, Alfonso Sánchez-Muñoz, Ana Lluch, Maria Rosario Chica-Parrado, Joan Albanell, José Ignacio Chacón, Silvia Antolín, et al. 2018. “Triple Negative Breast Cancer Subtypes and Pathologic Complete Response Rate to Neoadjuvant Chemotherapy.” *Oncotarget* 9 (41): 26406–16. <https://doi.org/10.18632/oncotarget.25413>.
- Santos, Felipe R C, Gabriela D A Guardia, and Filipe F Santos. 2020. “Reboot : A Straightforward Approach to Identify Genes and Splicing Isoforms Associated with Cancer Patient Prognosis.” *BioRxiv*, 1–29.
- Saraiva-Agostinho, Nuno, and Nuno L. Barbosa-Morais. 2019. “Psichomics: Graphical Application for Alternative Splicing Quantification and Analysis.” *Nucleic Acids Research* 47 (2). <https://doi.org/10.1093/nar/gky888>.
- Savagner, Pierre, Ana M. Vallés, Jacqueline Jouanneau, Kenneth M. Yamada, and Jean Paul Thiery. 1994. “Alternative Splicing in Fibroblast Growth Factor Receptor 2 Is Associated with Induced Epithelial-Mesenchymal Transition in Rat Bladder Carcinoma Cells.” *Molecular Biology of the Cell*. <https://doi.org/10.1091/mbc.5.8.851>.
- Scotti, Marina M., and Maurice S. Swanson. 2016. “RNA Mis-Splicing in Disease.” *Nature*

- Reviews Genetics*. <https://doi.org/10.1038/nrg.2015.3>.
- Sebestyén, Endre, Michał Zawisza, and Eduardo Eyras. 2015. “Detection of Recurrent Alternative Splicing Switches in Tumor Samples Reveals Novel Signatures of Cancer.” *Nucleic Acids Research* 43 (3): 1345–56. <https://doi.org/10.1093/nar/gku1392>.
- Seiler, Michael, Shouyong Peng, Anant A Agrawal, James Palacino, Teng Teng, Ping Zhu, Peter G Smith, et al. 2018. “Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types.” *Cell Reports* 23 (1): 282–296.e4. <https://doi.org/10.1016/j.celrep.2018.01.088>.
- Shapiro, Irina M, Albert W Cheng, Nicholas C Flytzanis, Michele Balsamo, John S Condeelis, Maja H Oktay, Christopher B Burge, and Frank B Gertler. 2011. “An EMT–Driven Alternative Splicing Program Occurs in Human Breast Cancer and Modulates Cellular Phenotype.” *PLoS Genet* 7 (8). <https://doi.org/10.1371/journal.pgen.1002218>.
- Sheller, Micah J., Brandon Edwards, G. Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, et al. 2020. “Federated Learning in Medicine: Facilitating Multi-Institutional Collaborations without Sharing Patient Data.” *Scientific Reports*. <https://doi.org/10.1038/s41598-020-69250-1>.
- Shen, Shihao, Juwon Park, Zhi-xiang Lu, Lan Lin, Michael D. Henry, Ying Nian Wu, Qing Zhou, and Yi Xing. 2014. “RMATS: Robust and Flexible Detection of Differential Alternative Splicing from Replicate RNA-Seq Data.” *Proceedings of the National Academy of Sciences* 111 (51): E5593–5601. <https://doi.org/10.1073/pnas.1419161111>.
- Sheng, Xiaoli, Yunxian Li, Yixuan Li, Wenlin Liu, Zhongming Lu, Jiandong Zhan, Mimi Xu, et al. 2019. “PLOD2 Contributes to Drug Resistance in Laryngeal Cancer by Promoting Cancer Stem Cell-like Characteristics.” *BMC Cancer*. <https://doi.org/10.1186/s12885-019-6029-y>.
- Siemens, Helge, Rene Jackstadt, Sabine Hüntgen, Markus Kaller, Antje Menssen, Ursula Götz, and Heiko Hermeking. 2011. “MiR-34 and SNAIL Form a Double-Negative Feedback Loop to Regulate Epithelial-Mesenchymal Transitions.” *Cell Cycle*. <https://doi.org/10.4161/cc.10.24.18552>.
- Simeonov, Kamen P, China N Byrns, Megan L Clark, Robert J Norgard, Beth Martin, Ben Z Stanger, Aaron McKenna, Jay Shendure, and Christopher J Lengner. 2020. “Single-Cell Lineage and Transcriptome Reconstruction of Metastatic Cancer Reveals Selection of Aggressive Hybrid EMT States.” *BioRxiv*. <https://doi.org/10.1101/2020.08.11.245787>.
- Sorlie, T, C M Perou, R Tibshirani, T Aas, S Geisler, H Johnsen, T Hastie, et al. 2001. “Gene Expression Patterns of Breast Carcinomas Distinguish Tumor Subclasses with Clinical Implications.” *Proceedings of the National Academy of Sciences of the United States of America* 98 (19): 10869–74. <https://doi.org/10.1073/pnas.191367098>.
- Stemmler, Marc P, Rebecca L Eccles, Simone Brabletz, and Thomas Brabletz. 2019. “Non-Redundant Functions of EMT-TFs.” *Nature Cell Biology* 21 (January): 102–12. <https://doi.org/10.1038/s41556-018-0196-y>.
- Sterne-Weiler, Timothy, Robert J. Weatheritt, Andrew J. Best, Kevin C.H. Ha, and Benjamin J. Blencowe. 2018. “Efficient and Accurate Quantitative Profiling of Alternative Splicing Patterns of Any Complexity on a Laptop.” *Molecular Cell* 72 (1): 187–200.e6. <https://doi.org/10.1016/j.molcel.2018.08.018>.
- Stone, Rivka C., Irena Pastar, Nkemcho Ojeh, Vivien Chen, Sophia Liu, Karen I. Garzon, and Marjana Tomic-Canic. 2016. “Epithelial-Mesenchymal Transition in Tissue Repair and

- Fibrosis.” *Cell and Tissue Research*. <https://doi.org/10.1007/s00441-016-2464-0>.
- Tan, Tuan Zea, Qing Hao Miow, Yoshio Miki, Tetsuo Noda, Seiichi Mori, Ruby Yun-Ju Huang, and Jean Paul Thiery. 2014. “Epithelial-mesenchymal Transition Spectrum Quantification and Its Efficacy in Deciphering Survival and Drug Responses of Cancer Patients.” *EMBO Molecular Medicine* 6 (10): 1279–93. <https://doi.org/10.15252/emmm.201404208>.
- Tanaka, Iris, Alina Chakraborty, Olivier Saulnier, Clara Benoit-Pilven, Sophie Vacher, Dalila Labiod, Eric W F Lam, et al. 2020. “ZRANB2 and SYF2-Mediated Splicing Programs Converging on ECT2 Are Involved in Breast Cancer Cell Resistance to Doxorubicin.” *Nucleic Acids Research*, 1–18. <https://doi.org/10.1093/nar/gkz1213>.
- Tao, Zi Qi, Aimin Shi, Cuntao Lu, Tao Song, Zhengguo Zhang, and Jing Zhao. 2015. “Breast Cancer: Epidemiology and Etiology.” *Cell Biochemistry and Biophysics*. <https://doi.org/10.1007/s12013-014-0459-6>.
- Tapial, Javier, Kevin C.H. Ha, Timothy Sterne-Weiler, André Gohr, Ulrich Braunschweig, Antonio Hermoso-Pulido, Mathieu Quesnel-Vallières, et al. 2017. “An Atlas of Alternative Splicing Profiles and Functional Associations Reveals New Regulatory Programs and Genes That Simultaneously Express Multiple Major Isoforms.” *Genome Research*. <https://doi.org/10.1101/gr.220962.117>.
- Thiery, J. P. 2002. “Epithelial-Mesenchymal Transitions in Tumor Progression.” *Nature Reviews Cancer* 2 (6): 442–54. <https://doi.org/10.1038/nrc822>.
- Thorsen, Kasper, Karina D. Sørensen, Anne Sofie Brems-Eskildsen, Charlotte Modin, Mette Gaustadnes, Anne-Mette K. Hein, Mogens Kruhøffer, et al. 2008. “Alternative Splicing in Colon, Bladder, and Prostate Cancer Identified by Exon Array Analysis.” *Molecular & Cellular Proteomics* 7 (7): 1214–24. <https://doi.org/10.1074/mcp.M700590-MCP200>.
- Tiberi, Simone, and Mark D. Robinson. 2020. “BANDITS: Bayesian Differential Splicing Accounting for Sample-to-Sample Variability and Mapping Uncertainty.” *Genome Biology*. <https://doi.org/10.1186/s13059-020-01967-8>.
- Tranchevent, Léon-Charles, Fabien Aubé, Louis Dulaurier, Clara Benoit-Pilven, Amandine Rey, Arnaud Poret, Emilie Chautard, et al. 2017. “Identification of Protein Features Encoded by Alternative Exons Using Exon Ontology.” *Genome Research*. <https://doi.org/10.1101/gr.212696.116>.
- Trapnell, Cole, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R. Kelley, Harold Pimentel, Steven L. Salzberg, John L. Rinn, and Lior Pachter. 2012. “Differential Gene and Transcript Expression Analysis of RNA-Seq Experiments with TopHat and Cufflinks.” *Nature Protocols*. <https://doi.org/10.1038/nprot.2012.016>.
- Tress, Michael L., Federico Abascal, and Alfonso Valencia. 2017. “Alternative Splicing May Not Be the Key to Proteome Complexity.” *Trends in Biochemical Sciences*. <https://doi.org/10.1016/j.tibs.2016.08.008>.
- Trincado, Juan L., E. Sebestyén, A. Pagés, and E. Eyra. 2016. “The Prognostic Potential of Alternative Transcript Isoforms across Human Tumors.” *Genome Medicine* 8 (1): 1–14. <https://doi.org/10.1186/s13073-016-0339-3>.
- Trincado, Juan L, Juan C Entizne, Gerald Hysenaj, Babita Singh, Miha Skalic, David J Elliott, and Eduardo Eyra. 2016. “SUPPA2 Provides Fast, Accurate, and Uncertainty-Aware Differential Splicing Analysis across Multiple Conditions.” <https://doi.org/10.1101/086876>.

- Tripathi, Veenu, Jee Hye Shin, Christina H. Stuelten, and Ying E. Zhang. 2019. “TGF- β -Induced Alternative Splicing of TAK1 Promotes EMT and Drug Resistance.” *Oncogene* 38 (17): 3185–3200. <https://doi.org/10.1038/s41388-018-0655-8>.
- Tsai, Yihuan S., Daniel Dominguez, Shawn M. Gomez, and Zefeng Wang. 2015. “Transcriptome-Wide Identification and Study of Cancer-Specific Splicing Events across Multiple Tumors.” *Oncotarget* 6 (9): 6825–39. <https://doi.org/10.18632/oncotarget.3145>.
- Tsherniak, Aviad, Francisca Vazquez, Phil G. Montgomery, Barbara A. Weir, Gregory Kryukov, Glenn S. Cowley, Stanley Gill, et al. 2017. “Defining a Cancer Dependency Map.” *Cell* 170 (3): 564-576.e16. <https://doi.org/10.1016/j.cell.2017.06.010>.
- Turner, Nicholas, and Richard Grose. 2010. “Fibroblast Growth Factor Signalling: From Development to Cancer.” *Nature Reviews Cancer*. <https://doi.org/10.1038/nrc2780>.
- Urbanski, Laura, and Nathan Leclair. 2019. “Alternative-Splicing Defects in Cancer: Splicing Regulators and Their Downstream Targets, Guiding the Way to Novel Cancer Therapeutics.” 9 (4): 1–56. <https://doi.org/10.1002/wrna.1476.Alternative-splicing>.
- Urbanski, Laura M., Nathan Leclair, and Olga Anczuków. 2018. “Alternative-Splicing Defects in Cancer: Splicing Regulators and Their Downstream Targets, Guiding the Way to Novel Cancer Therapeutics.” *Wiley Interdisciplinary Reviews: RNA*, no. January: e1476. <https://doi.org/10.1002/wrna.1476>.
- Vandewalle, C., F. Van Roy, and G. Berx. 2009. “The Role of the ZEB Family of Transcription Factors in Development and Disease.” *Cellular and Molecular Life Sciences*. <https://doi.org/10.1007/s00018-008-8465-8>.
- Vanharanta, Sakari, Christina B. Marney, Weiping Shu, Manuel Valiente, Yilong Zou, Aldo Mele, Robert B. Darnell, and Joan Massagué. 2014. “Loss of the Multifunctional RNA-Binding Protein RBM47 as a Source of Selectable Metastatic Traits in Breast Cancer.” *ELife* 2014 (3): 1–24. <https://doi.org/10.7554/eLife.02734.001>.
- Vaquero-Garcia, Jorge, Alejandro Barrera, Matthew R. Gazzara, Juan Gonzalez-Vallinas, Nicholas F. Lahens, John B. Hogenesch, Kristen W. Lynch, and Yoseph Barash. 2016. “A New View of Transcriptome Complexity and Regulation through the Lens of Local Splicing Variations.” *ELife*. <https://doi.org/10.7554/eLife.11752>.
- Venables, J. P., J.-P. Brosseau, G. Gadea, R. Klinck, P. Prinos, J.-F. Beaulieu, E. Lapointe, et al. 2013. “RBFOX2 Is an Important Regulator of Mesenchymal Tissue-Specific Splicing in Both Normal and Cancer Tissues.” *Molecular and Cellular Biology* 33 (2): 396–405. <https://doi.org/10.1128/MCB.01174-12>.
- Venables, Julian P., Roscoe Klinck, Chushin Koh, Julien Gervais-Bird, Anne Bramard, Lyna Inkel, Mathieu Durand, et al. 2009. “Cancer-Associated Regulation of Alternative Splicing.” *Nature Structural and Molecular Biology* 16 (6): 670–76. <https://doi.org/10.1038/nsmb.1608>.
- Venables, Julian P., Laure Lapasset, Gilles Gadea, Philippe Fort, Roscoe Klinck, Manuel Irimia, Emmanuel Vignal, et al. 2013. “MBNL1 and RBFOX2 Cooperate to Establish a Splicing Programme Involved in Pluripotent Stem Cell Differentiation.” *Nature Communications* 4 (May): 1–10. <https://doi.org/10.1038/ncomms3480>.
- Vicovac, L., and J.D. Aplin. 1996. “Epithelial-Mesenchymal Transition during Trophoblast Differentiation.”
- Visvader, Jane E. 2011. “Cells of Origin in Cancer.” *Nature* 469 (7330): 314–22. <https://doi.org/10.1038/nature09781>.

- Wang, Chiung Min, Runhua Liu, Lizhong Wang, Leticia Nascimento, Victoria C. Brennan, and Wei Hsiung Yang. 2014. "SUMOylation of FOXM1B Alters Its Transcriptional Activity on Regulation of MiR-200 Family and JNK1 in MCF7 Human Breast Cancer Cells." *International Journal of Molecular Sciences*. <https://doi.org/10.3390/ijms150610233>.
- Wang, Dong Yu, Zhe Jiang, Yaacov Ben-David, James R. Woodgett, and Eldad Zacksenhaus. 2019. "Molecular Stratification within Triple-Negative Breast Cancer Subtypes." *Scientific Reports* 9 (1): 1–10. <https://doi.org/10.1038/s41598-019-55710-w>.
- Wang, Eric, and Iannis Aifantis. 2020. "RNA Splicing and Cancer." *Trends in Cancer*, 1–14. <https://doi.org/10.1016/j.trecan.2020.04.011>.
- Wang, Yifan, Jian Shi, Kequn Chai, Xuhua Ying, and Binhua Zhou. 2014. "The Role of Snail in EMT and Tumorigenesis." *Current Cancer Drug Targets*. <https://doi.org/10.2174/15680096113136660102>.
- Wang, Zefeng, and Christopher B. Burge. 2008. "Splicing Regulation: From a Parts List of Regulatory Elements to an Integrated Splicing Code." *RNA*. <https://doi.org/10.1261/rna.876308>.
- Warzecha, Claude C., and Russ P. Carstens. 2012. "Complex Changes in Alternative Pre-mRNA Splicing Play a Central Role in the Epithelial-to-Mesenchymal Transition (EMT)." *Seminars in Cancer Biology* 22 (5–6): 417–27. <https://doi.org/10.1016/j.semcancer.2012.04.003>.
- Warzecha, Claude C, Trey K Sato, Behnam Nabet, John B Hogenesch, and P Russ. 2010. "ESRP1 and ESRP2 Are Epithelial Cell Type-Specific Regulators of FGFR2 Splicing" 33 (5): 591–601. <https://doi.org/10.1016/j.molcel.2009.01.025>.ESRP1.
- Weatheritt, Robert J., Timothy Sterne-Weiler, and Benjamin J. Blencowe. 2016. "The Ribosome-Engaged Landscape of Alternative Splicing." *Nature Structural and Molecular Biology*. <https://doi.org/10.1038/nsmb.3317>.
- Weeber, Fleur, Salo N. Ooft, Krijn K. Dijkstra, and Emile E. Voest. 2017. "Tumor Organoids as a Pre-Clinical Cancer Model for Drug Discovery." *Cell Chemical Biology*. <https://doi.org/10.1016/j.chembiol.2017.06.012>.
- Williams, Elizabeth D., Dingcheng Gao, Andrew Redfern, and Erik W. Thompson. 2019. "Controversies around Epithelial–Mesenchymal Plasticity in Cancer Metastasis." *Nature Reviews Cancer*. <https://doi.org/10.1038/s41568-019-0213-x>.
- Wilson, Molly M., Robert A. Weinberg, Jacqueline A. Lees, and Vincent J. Guen. 2020. "Emerging Mechanisms by Which EMT Programs Control Stemness." *Trends in Cancer*, 1–6. <https://doi.org/10.1016/j.trecan.2020.03.011>.
- Wu, Hua Tao, Hui Ting Zhong, Guan Wu Li, Jia Xin Shen, Qian Qian Ye, Man Li Zhang, and Jing Liu. 2020. "Oncogenic Functions of the EMT-Related Transcription Factor ZEB1 in Breast Cancer." *Journal of Translational Medicine*. <https://doi.org/10.1186/s12967-020-02240-z>.
- Xu, Hanxiao, Xiaodong Lyu, Ming Yi, Weiheng Zhao, Yongping Song, and Kongming Wu. 2018. "Organoid Technology and Applications in Cancer Research." *Journal of Hematology & Oncology*. <https://doi.org/10.1186/s13045-018-0662-9>.
- 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*.

<https://doi.org/10.18632/oncotarget.15581>.

- Xu, Yilin, Xin D. Gao, Jae Hyung Lee, Huilin Huang, Haiyan Tan, Jaegyeon Ahn, Lauren M. Reinke, et al. 2014. “Cell Type-Restricted Activity of HnRNPM Promotes Breast Cancer Metastasis via Regulating Alternative Splicing.” *Genes and Development* 28 (11): 1191–1203. <https://doi.org/10.1101/gad.241968.114>.
- Xu, Yungang. 2017. “Alternative Splicing Links Histone Modifications to Stem Cell Fate Decision,” 1–21. <https://doi.org/10.1186/s13059-018-1512-3>.
- Y., Lou, Diao L., Cuentas E.R.P., Denning W.L., Chen L., Fan Y.H., Byers L.A., et al. 2016. “Epithelial-Mesenchymal Transition Is Associated with a Distinct Tumor Microenvironment Including Elevation of Inflammatory Signals and Multiple Immune Checkpoints in Lung Adenocarcinoma.” *Clinical Cancer Research*. <https://doi.org/10.1158/1078-0432.CCR-15-1434> LK - <http://limo.libis.be/resolver?&sid=EMBASE&issn=15573265&id=doi:10.1158%2F1078-0432.CCR-15-1434&atitle=Epithelial-mesenchymal+transition+is+associated+with+a+distinct+tumor+microenvironment+including+elevation+of+inflammatory+signals+and+multiple+immune+checkpoints+in+lun+g+adenocarcinoma&stitle=Clin.+Cancer+Res.&title=Clinical+Cancer+Research&volume=22&issue=14&spage=3630&epage=3642&aualast=Lou&aufirst=Yanyan&aunit=Y.&aufull=Lou+Y.&coden=CCREF&isbn=&pages=3630->
- Yanagisawa, Masahiro, Deborah Huveltdt, Pamela Kreinest, Christine M. Lohse, John C. Cheville, Alexander S. Parker, John A. Copland, and Panos Z. Anastasiadis. 2008. “A P120 Catenin Isoform Switch Affects Rho Activity, Induces Tumor Cell Invasion, and Predicts Metastatic Disease.” *Journal of Biological Chemistry*. <https://doi.org/10.1074/jbc.M801192200>.
- Yang, Jing, Parker Antin, Geert Berx, Cédric Blanpain, Thomas Brabletz, Marianne Bronner, Kyra Campbell, et al. 2020. “Guidelines and Definitions for Research on Epithelial–Mesenchymal Transition.” *Nature Reviews Molecular Cell Biology* 21 (6): 341–52. <https://doi.org/10.1038/s41580-020-0237-9>.
- Yang, Jing, and Robert A. Weinberg. 2008. “Epithelial-Mesenchymal Transition: At the Crossroads of Development and Tumor Metastasis.” *Developmental Cell* 14 (6): 818–29. <https://doi.org/10.1016/j.devcel.2008.05.009>.
- Yang, Liao, Smyth Gordon K, and Shi Wei. 2014. “FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features.” *Bioinformatics*.
- Yang, Yueqin, Juwon Park, Thomas W. Bebee, Claude C. Warzecha, Yang Guo, Xuequn Shang, Yi Xing, and Russ P. Carstens. 2016a. “Determination of a Comprehensive Alternative Splicing Regulatory Network and Combinatorial Regulation by Key Factors during the Epithelial-to-Mesenchymal Transition.” *Molecular and Cellular Biology*. <https://doi.org/10.1128/mcb.00019-16>.
- . 2016b. “Determination of a Comprehensive Alternative Splicing Regulatory Network and Combinatorial Regulation by Key Factors during the Epithelial-to-Mesenchymal Transition.” *Molecular and Cellular Biology* 36 (11): 1704–19. <https://doi.org/10.1128/MCB.00019-16>.
- Ye, Xin, and Robert A. Weinberg. 2015. “Epithelial-Mesenchymal Plasticity: A Central Regulator of Cancer Progression.” *Trends in Cell Biology* 25 (11): 675–86. <https://doi.org/10.1016/j.tcb.2015.07.012>.

- Yeo, Gene W., Nicole G. Coufal, Tiffany Y. Liang, Grace E. Peng, Xiang Dong Fu, and Fred H. Gage. 2009. "An RNA Code for the FOX2 Splicing Regulator Revealed by Mapping RNA-Protein Interactions in Stem Cells." *Nature Structural and Molecular Biology*. <https://doi.org/10.1038/nsmb.1545>.
- Yoshida, Kenichi Ogawa, Seishi. 2014. "Splicing Factor Mutations and Cancer."
- Yu, Min, Aditya Bardia, Ben S. Wittner, Shannon L. Stott, Malgorzata E. Smas, David T. Ting, Steven J. Isakoff, et al. 2013. "Circulating Breast Tumor Cells Exhibit Dynamic Changes in Epithelial and Mesenchymal Composition." *Science* 339 (6119): 580–84. <https://doi.org/10.1126/science.1228522>.
- Zhang, Dong, Yi Duan, Jinjing Cun, and Qifeng Yang. 2019. "Identification of Prognostic Alternative Splicing Signature in Breast Carcinoma." *Frontiers in Genetics* 10 (March): 1–20. <https://doi.org/10.3389/fgene.2019.00278>.
- Zhang, Xiuqin, Omar A. Ibrahimi, Shaun K. Olsen, Hisashi Umemori, Moosa Mohammadi, and David M. Ornitz. 2006. "Receptor Specificity of the Fibroblast Growth Factor Family: The Complete Mammalian FGF Family." *Journal of Biological Chemistry*. <https://doi.org/10.1074/jbc.M601252200>.
- Zhang, Yijun, Yue Zhao, Guiyang Jiang, Xiupeng Zhang, Huanyu Zhao, Junhua Wu, Ke Xu, and Enhua Wang. 2014. "Impact of P120-Catenin Isoforms 1A and 3A on Epithelial Mesenchymal Transition of Lung Cancer Cells Expressing e-Cadherin in Different Subcellular Locations." *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0088064>.
- Zhao, Zhen, Xiaoping Zhu, Kemi Cui, James Mancuso, Richard Federley, Kari Fischer, Gao Jun Teng, et al. 2016. "In Vivo Visualization and Characterization of Epithelial-Mesenchymal Transition in Breast Tumors." *Cancer Research*. <https://doi.org/10.1158/0008-5472.CAN-15-2662>.
- Zheng, Hong, Guosen Zhang, Lu Zhang, Qiang Wang, Huimin Li, Yali Han, Longxiang Xie, et al. 2020. "Comprehensive Review of Web Servers and Bioinformatics Tools for Cancer Prognosis Analysis." *Frontiers in Oncology*. <https://doi.org/10.3389/fonc.2020.00068>.
- Zheng, Xiaofeng, Julianne L. Carstens, Jiha Kim, Matthew Scheible, Judith Kaye, Hikaru Sugimoto, Chia Chin Wu, Valerie S. Lebleu, and Raghu Kalluri. 2015. "Epithelial-to-Mesenchymal Transition Is Dispensable for Metastasis but Induces Chemoresistance in Pancreatic Cancer." *Nature*. <https://doi.org/10.1038/nature16064>.
- Zhou, Bin Bing S., Haiying Zhang, Marc Damelin, Kenneth G. Geles, Justin C. Grindley, and Peter B. Dirks. 2009. "Tumour-Initiating Cells: Challenges and Opportunities for Anticancer Drug Discovery." *Nature Reviews Drug Discovery*. <https://doi.org/10.1038/nrd2137>.
- Zhou, Binhua P., Jiong Deng, Weiya Xia, Jihong Xu, Yan M. Li, Mehmet Gunduz, and Mien Chie Hung. 2004. "Dual Regulation of Snail by GSK-3 β -Mediated Phosphorylation in Control of Epithelial-Mesenchymal Transition." *Nature Cell Biology*. <https://doi.org/10.1038/ncb1173>.
- Zhu, Junyong, Zuhua Chen, and Lei Yong. 2017. "Systematic Profiling of Alternative Splicing Signature Reveals Prognostic Predictor for Ovarian Cancer." *Gynecologic Oncology* 148 (2): 368–74. <https://doi.org/10.1016/j.ygyno.2017.11.028>.
- Zhu, Li-Yuan, Yi-Ran Zhu, Dong-Jun Dai, Xian Wang, and Hong-Chuan Jin. 2018. "Epigenetic Regulation of Alternative Splicing." *American Journal of Cancer Research*.

Zöller, Margot. 2011. "CD44: Can a Cancer-Initiating Cell Profit from an Abundantly Expressed Molecule?" *Nature Reviews Cancer*. <https://doi.org/10.1038/nrc3023>.

Zong, Feng Yang, Xing Fu, Wen Juan Wei, Ya Ge Luo, Monika Heiner, Li Juan Cao, Zhaoyuan Fang, et al. 2014. "The RNA-Binding Protein QKI Suppresses Cancer-Associated Aberrant Splicing." *PLoS Genetics*. <https://doi.org/10.1371/journal.pgen.1004289>.

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 Ration (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)

ARTICLE

<https://doi.org/10.1038/s42003-019-0456-9>

OPEN

GECKO is a genetic algorithm to classify and explore high throughput sequencing data

Aubin Thomas^{1,6}, Sylvain Barriere^{1,6}, Lucile Broseus¹, Julie Brooke¹, Claudio Lorenzi¹, Jean-Philippe Villemin¹, Gregory Beurier ², Robert Sabatier³, Christelle Reynes³, Alban Mancheron^{4,5} & 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^{1–3}. 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 variations⁴. 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 samples⁵. These approaches have been successful at detecting novel transcripts but only on a very small subset of RNA sequencing data⁴ 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^{6,7}. 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 compared⁸. 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 bisulfite 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

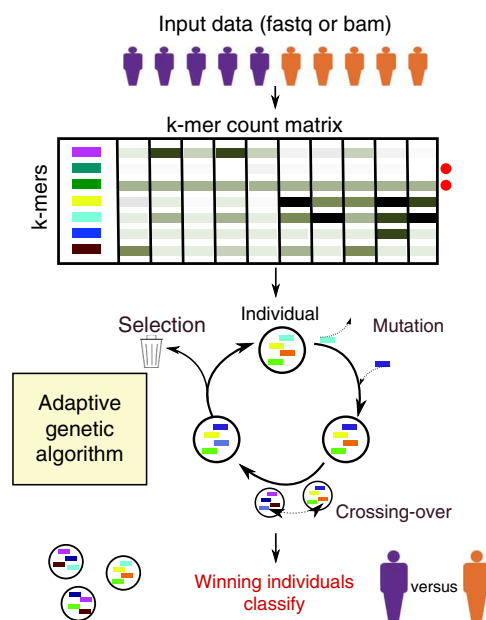


Fig. 1 Overview of the GECKO algorithm. Input fastq or bam files from two or more conditions are transformed into a matrix of *k-mer* counts across all samples. The *k-mers* for which the counts are below a noise threshold or that do not vary across samples are removed (red dots on the right of the *k-mer* matrix). The adaptive genetic algorithm randomly selects groups of *k-mers* from the *k-mer* matrix to form individuals. These individuals will go through rounds of mutation, crossing-over and selection to discover individuals capable of classifying the input samples with high accuracy

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 samples⁹. 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, CCCCCA 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 miRNAs¹⁰. 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^{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 cohort¹³ (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

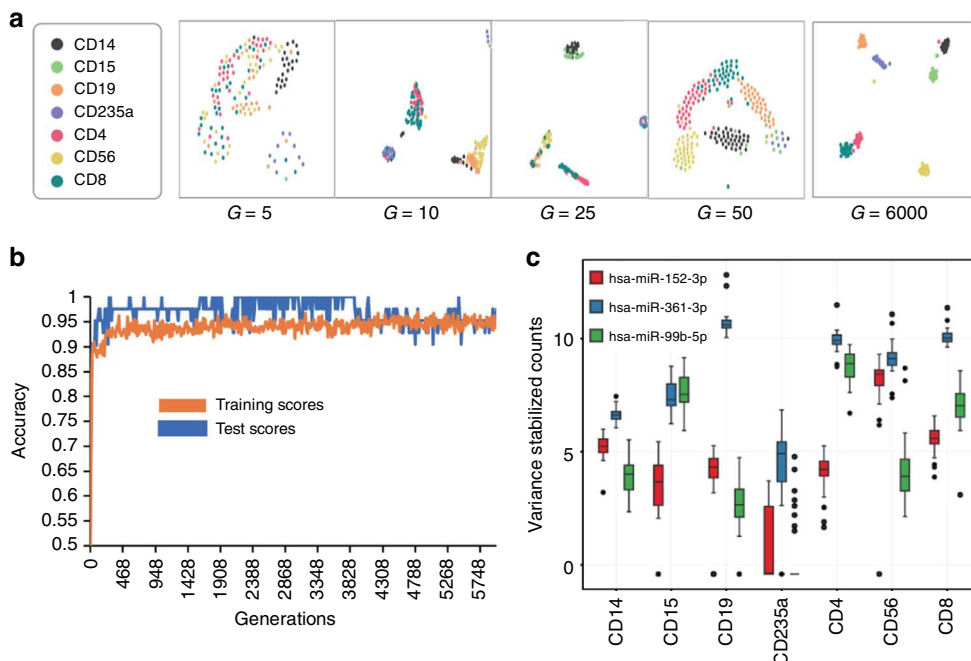


Fig. 2 GECKO can accurately classify miRNA data from seven types of blood cells using three *k-mers*. **a** GECKO output showing the separation of the seven blood-cell types at each generation (G) of GECKO analysis using t-SNE visualization applied to *k-mer* counts. **b** GECKO output showing the accuracy 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 aggressiveness¹³ and the two latter both affect cell viability in breast cancer cells^{14,15}. 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 cancer⁸, *BOPI* reduces chemotherapy resistance²¹ and *VTCN1* is associated with poor clinical outcomes in numerous cancers including breast cancer²².

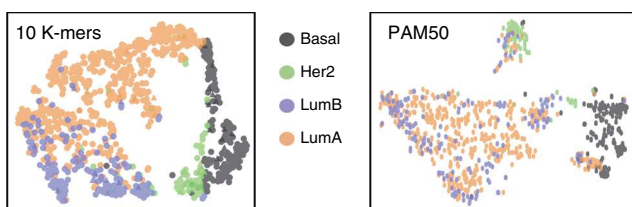


Fig. 3 GECKO discovers 10 30-mers that classify breast cancer subtypes. Comparison of breast cancer subtype classification using the frequency of *k*-mers discovered by GECKO and the transcript per million values of the PAM50 gene. Panels show the t-SNE separation of the four classes

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

Classification with GECKO <i>k</i> -mers					Classification with PAM50 TPM values						
Predicted class	Basal	97.7	2.2	0	0	Predicted class	Basal	86	5.2	5.5	3.3
	Her2	2	87.5	6.2	4.2		Her2	15.3	60.6	3.6	20.6
	LumA	1.5	1.5	92.3	4.6		LumA	15.3	2.2	88.1	8.6
	LumB	0	3.4	18.8	77.8		LumB	5.9	15.4	36.5	42.2
	Basal	Her2	LumA	LumB			Basal	Her2	LumA	LumB	
	True class						True class				

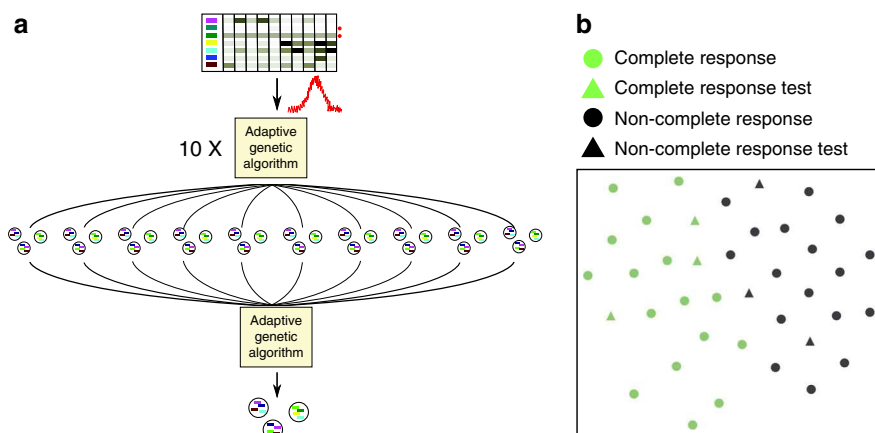


Fig. 4 GECKO voting mode for small sample sizes. **a** GECKO's voting mode will run 10 separate genetic algorithms with added Gaussian noise. The best solutions of these runs will be fed into a final genetic algorithm to produce a final solution. **b** GECKO output showing the t-SNE separation of patients 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

in methylation and thus, a method that could directly classify BS-seq samples could be of great interest. To test GECKO on BS-seq we downloaded raw sequencing files from a study on methylome diversity in 104 primary CLLs samples compared with 26 normal B cell samples²³. Although global hypomethylation has been well described in cancer, these alterations are highly variable between CLL samples²³ and thus present a challenge for classification.

We ran GECKO for 20,000 generations (39 h; see Supplementary Table 1 for parameters and Supplementary Fig. 1 for runtimes 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. 5a; Supplementary Table 4). In addition to this final classification, GECKO plots the evolution of winning organisms across the 20,000 generations (Fig. 5b). 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*-mer sequences were mapped using the Bismark software²⁴ and found that all three of them displayed dramatic changes in DNA methylation between normal and CLL samples (Fig. 5c). 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^{25,26}. K107977 overlaps a CTCF-binding site for the ATP6V1G1 gene²⁷, 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^{28–30}.

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.

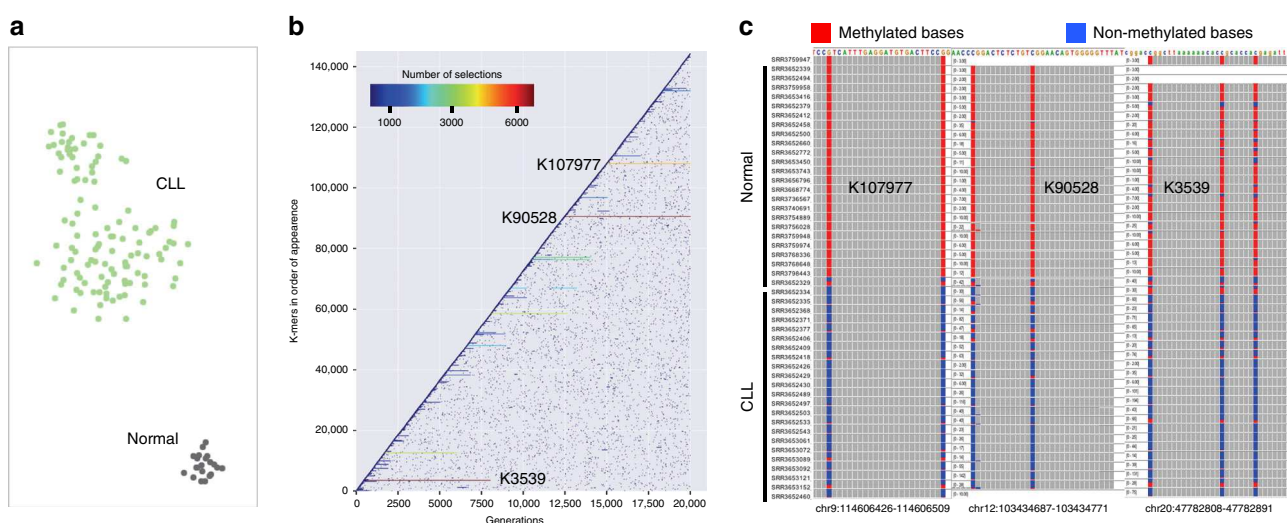


Fig. 5 GECKO can accurately classify normal and CLL patients using *k*-mers from bisulfite sequencing data. **a** GECKO output showing the t-SNE separation of CLL and normal samples using 20 *k*-mers from the winning individual. **b** GECKO output of *K*-mer exploration across 20,000 generations; *k*-mers 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 Jellyfish²³. 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. When this frequency drops dramatically as determined by 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 uninformative k -mers, that is k -mers that do not vary across input samples, we first discretize the k -mer counts using a chi-square statistic that determines the minimum number of discrete intervals with minimum loss of class attribute interdependence³². 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 their 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 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:

$$SU(A, B) = 2 \times ((H(A) + H(B) - H(A, B)) \div (H(A) + H(B)))$$

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 -mer counts A and B along the samples.

The Entropy is given by the formula:

$$H(A) = - \sum_i^G Mi/N^* \log_2(Mi/N)$$

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 (LinSVC). This method combines excellent results on smalls datasets and unbalanced groups with a good generalization potential, for a small computational resource cost. LinSVC is implemented in GECKO via the Scikit-learn package³³. 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 up 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:

$$P - \text{value} = \alpha X + \beta$$

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

$$\sum_n^N P - \text{value} = 1$$

$$P - \text{value} = \frac{N/2}{N} \frac{P - \text{value}}{2} \frac{N}{P - \text{value}}$$

where α , β are scalar values, N is the size of the population, and $\frac{N}{N}$ and $\frac{N/2}{P - \text{value}}$ are, respectively, the probability for the individual rank N 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 adaptative genetic algorithm³⁴. 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 algorithm³⁵ 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 algorithm³⁶ 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.

$$Pm = \begin{cases} \frac{k_1(f_{avg}-f)+k_3(f-f_{min})}{f_{avg}-f_{min}}, & f < f_{avg} \\ \frac{k_2(f_{max}-f)+k_3(f-f_{avg})}{f_{max}-f_{avg}}, & f \geq f_{avg} \end{cases}$$

Here f is the individual's fitness, f_{min} is the fitness of the population's 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 $f=f_{min}$, k_2 when $f=f_{avg}$, and k_3 when $f=f_{max}$.

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 GSE100467 and GSE58889; the Cancer Genome Atlas under the Pan-Gyn cohort name; the database of Genotypes and Phenotypes under the accession numbers phs000435.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

- Learn, C. A. et al. Resistance to tyrosine kinase inhibition by mutant epidermal growth factor receptor variant III contributes to the neoplastic phenotype of glioblastoma multiforme. *Clin. Cancer Res.* **10**, 3216–3224 (2004).
- Zhang, Z.-M. et al. Pygo2 activates MDR1 expression and mediates chemoresistance in breast cancer via the Wnt/ β -catenin pathway. *Oncogene* **35**, 4787–4797 (2016).
- Martin-Martín, N. et al. Stratification and therapeutic potential of PML in metastatic breast cancer. *Nat. Commun.* **7**, 12595 (2016).
- Audoux, J. et al. DE-kupl: exhaustive capture of biological variation in RNA-seq data through k-mer decomposition. *Genome Biol.* **18**, 243 (2017).
- Kirk, J. M. et al. Functional classification of long non-coding RNAs by k-mer content. *Nat. Genet.* **1**, <https://doi.org/10.1038/s41588-018-0207-8> (2018).
- Uunit, R., Wanamaker, S., Close, T. J. & Lonardi, S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mer0s. *BMC Genom.* **16**, 236 (2015).
- Breitwieser, F. P., Baker, D. N. & Salzberg, S. L. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biol.* **19**, 198 (2018).
- Sergeyev, Y. D., Kvasov, D. E. & Mukhametzhonov, M. S. On the efficiency of nature-inspired metaheuristics in expensive global optimization with limited budget. *Sci. Rep.* **8**, 453 (2018).
- Juzenas, S. et al. A comprehensive, cell specific microRNA catalogue of human peripheral blood. *Nucleic Acids Res.* **45**, 9290–9301 (2017).
- Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42**, D68–D73 (2014).
- Perou, C. M. et al. Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
- Hoadley, K. A. et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173**, 291–304. e6 (2018).
- Maziveyi, M. & Alahari, S. K. Breast cancer tumor suppressors: a special emphasis on novel protein nischarin. *Cancer Res.* **75**, 4252–4259 (2015).
- Hasim, M. S., Nessim, C., Villeneuve, P. J., Vanderhyden, B. C. & Dimitroulakos, J. Activating transcription factor 3 as a novel regulator of chemotherapy response in breast cancer. *Transl. Oncol.* **11**, 988–998 (2018).
- Gijn, S. E. van et al. TPX2/Aurora kinase A signaling as a potential therapeutic target in genomically unstable cancer cells. *Oncogene* **1**, <https://doi.org/10.1038/s41388-018-0470-2> (2018).
- Choi, J. et al. Loss of KLHL6 promotes diffuse large B-cell lymphoma growth and survival by stabilizing the mRNA decay factor roquin2. *Nat. Cell Biol.* **20**, 586–596 (2018).
- Solari, N. E. F. et al. The NSL chromatin-modifying complex subunit KANSL2 regulates cancer stem-like properties in glioblastoma that contribute to tumorigenesis. *Cancer Res.* **76**, 5383–5394 (2016).
- Tatarskiy, V. V. et al. Stability of the PHF10 subunit of PBAF signature module is regulated by phosphorylation: role of β -TrCP. *Sci. Rep.* **7**, 5645 (2017).
- Goetz, M. P. et al. Tumor sequencing and patient-derived xenografts in the neoadjuvant treatment of breast cancer. *J. Natl. Cancer Inst.* **109**, 7 (2017).
- Thomas, S. J., Snowden, J. A., Zeidler, M. P. & Danson, S. J. The role of JAK/STAT signalling in the pathogenesis, prognosis and treatment of solid tumours. *Br. J. Cancer* **113**, 365–371 (2015).
- Sapio, R. T. et al. Inhibition of post-transcriptional steps in ribosome biogenesis confers cytoprotection against chemotherapeutic agents in a p53-dependent manner. *Sci. Rep.* **7**, 9041 (2017).
- Podojil, J. R. & Miller, S. D. Potential targeting of B7-H4 for the treatment of cancer. *Immunol. Rev.* **276**, 40–51 (2017).
- Landau, D. A. et al. Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. *Cancer Cell* **26**, 813–825 (2014).
- Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
- Wang, H. et al. Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.* **22**, 1680–1688 (2012).
- Fleischer, T. et al. DNA methylation at enhancers identifies distinct breast cancer lineages. *Nat. Commun.* **8**, 1379 (2017).
- Lesurf, R. et al. ORegAnno 3.0: a community-driven resource for curated regulatory annotation. *Nucleic Acids Res.* **44**, D126–D132 (2016).
- Alhasan, S. F. et al. Sulfatase-2: a prognostic biomarker and candidate therapeutic target in patients with pancreatic ductal adenocarcinoma. *Br. J. Cancer* **115**, 797–804 (2016).
- Rosen, S. D. & Lemjabbar-Alaoui, H. Sulf-2: an extracellular modulator of cell signaling and a cancer target candidate. *Expert Opin. Ther. Targets* **14**, 935–949 (2010).
- Lui, N. S. et al. SULF2 expression is a potential diagnostic and prognostic marker in lung cancer. *PLoS ONE* **11**, e0148911 (2016).
- Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
- Gonzalez-Abril, L., Cuberos, F. J., Velasco, F. & Ortega, J. A. Ameva: an autonomous discretization algorithm. *Expert Syst. Appl.* **36**, 5327–5332 (2009).
- Pedregosa, F. et al. Scikit-learn: machine learning in python. *ArXiv12010490 Cs* (2012).
- Zhang, J., Chung, H. S. H. & Hu, B. J. Adaptive probabilities of crossover and mutation in genetic algorithms based on clustering technique. In *Proc. 2004 Congress on Evolutionary Computation* (ed Greenwood, G. W.) (IEEE Cat. No. 04TH8753), Vol. 2, 2280–2287 (IEEE Portland, OR, USA, USA, 2004).
- Ravindran, S., Jambek, A. B., Muthusamy, H. & Neoh, S.-C. A novel clinical decision support system using improved adaptive genetic algorithm for the assessment of fetal well-being. *Comput. Math. Methods Med.* **2015**, 283532 (2015). <https://doi.org/10.1155/2015/283532>.
- Yan, M. et al. Improved adaptive genetic algorithm with sparsity constraint applied to thermal neutron CT reconstruction of two-phase flow. *Meas. Sci. Technol.* **29**, 55404 (2018).

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 Villemin and Giacomo Cavalli for their advice. We wish to acknowledge the Agence Nationale de la Recherche (ANR/JC-WIRED), the Labex EpiGenMed and the MUSE initiative for their financial support.

Author contributions

W.R., A.T., S.B., J.B., C.R., R.S., G.B., L.B. designed the algorithm; A.M., C.L. designed part of the *k-mer* filtering step; A.T. and S.B. coded the software; W.R., A.T., S.B., J.-P.V. designed the experiments. W.R., A.T., S.B. wrote the article.

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.

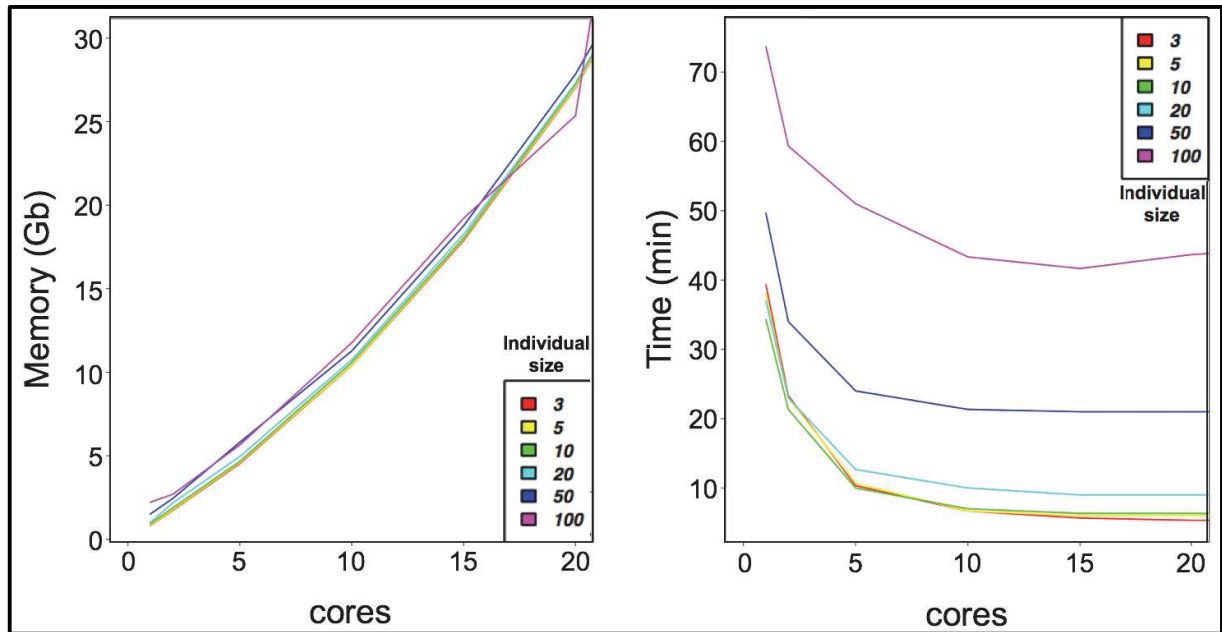
Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

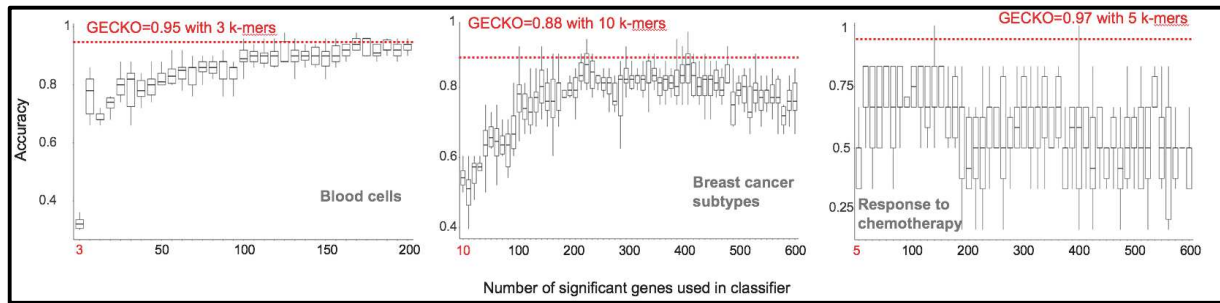


Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019



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.

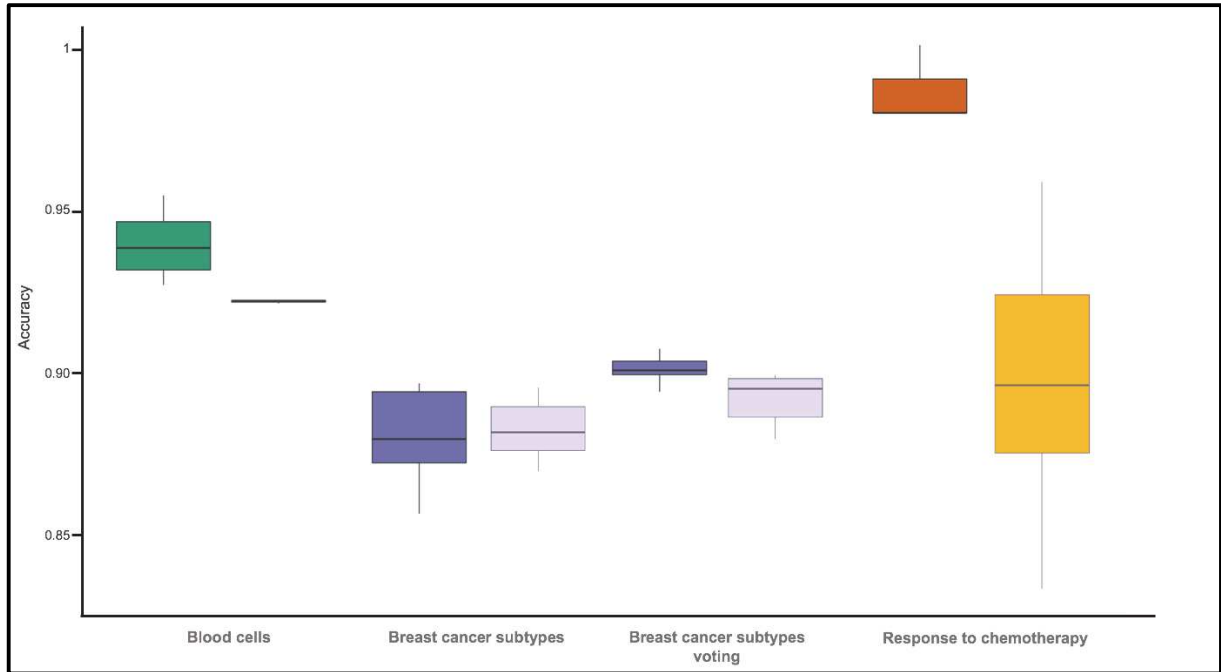


Supplementary Figure 2: comparison of classification accuracy between GECKO and classifiers based on an increasing number of differentially expressed genes

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.

1. Peng, L. *et al.* Large-scale RNA-Seq Transcriptome Analysis of 4043 Cancers and 548 Normal Tissue Controls across 12 TCGA Cancer Types. *Sci. Rep.* **5**, (2015).



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/smrnaseq). 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.

METHOD

Open Access

iMOKA: *k*-mer based software to analyze large collections of sequencing data



Claudio Lorenzi¹, Sylvain Barriere¹, Jean-Philippe Villemin¹, Laureline Dejardin Bretonnes¹, Alban Mancheron² and William Ritchie^{1*}

* Correspondence: william.ritchie@igh.cnrs.fr

¹IGH, Centre National de la Recherche Scientifique, University of Montpellier, Montpellier, France
Full list of author information is available at the end of the article

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



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

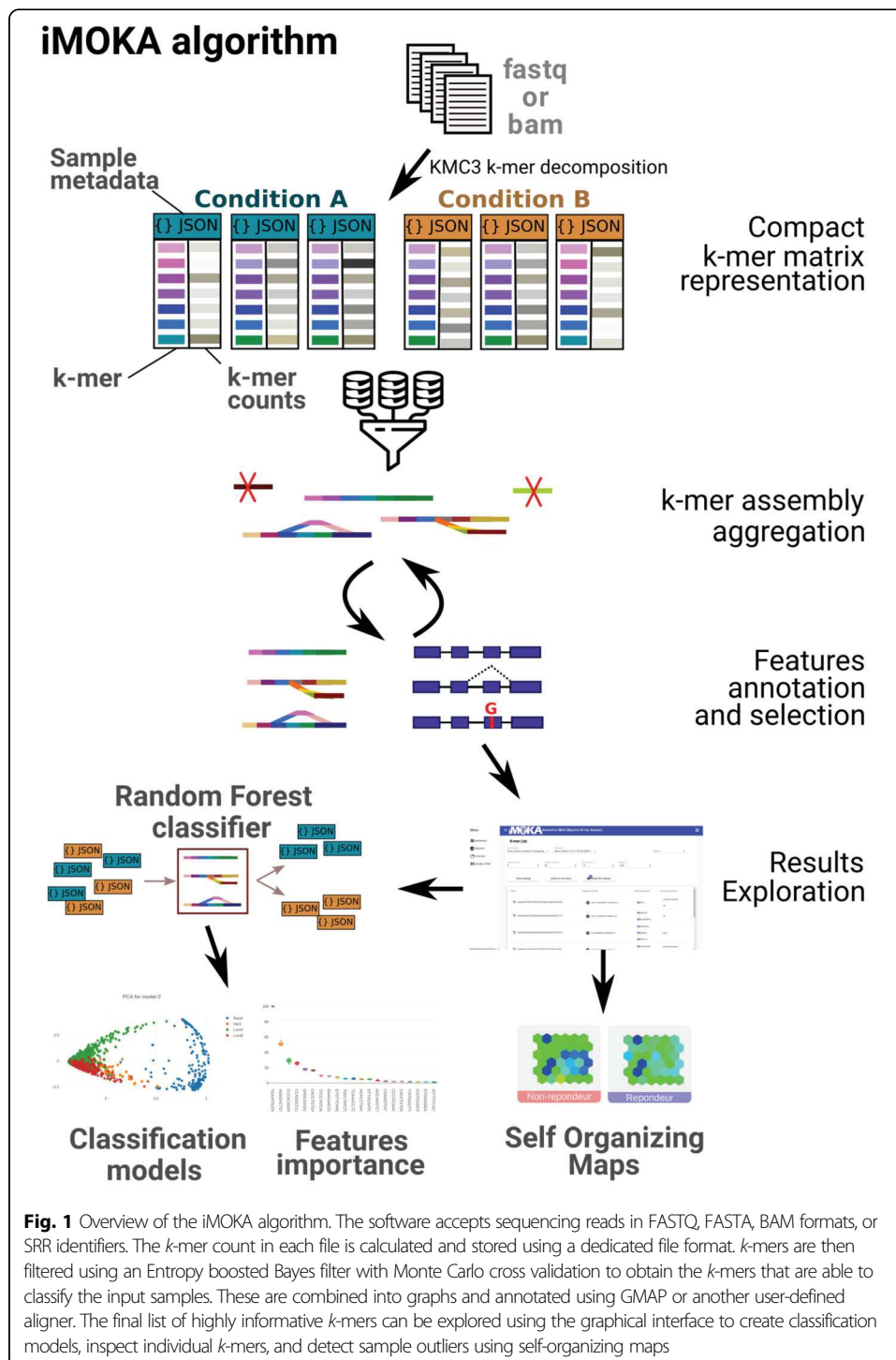
representation results in a k -mer count matrix with as many columns as there are samples and as many rows as there are k -mers, generally billions. Exploring such large matrices to find biologically relevant k -mers is intractable unless the analysis focuses only on a very small subset of the sequencing 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]. By



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.

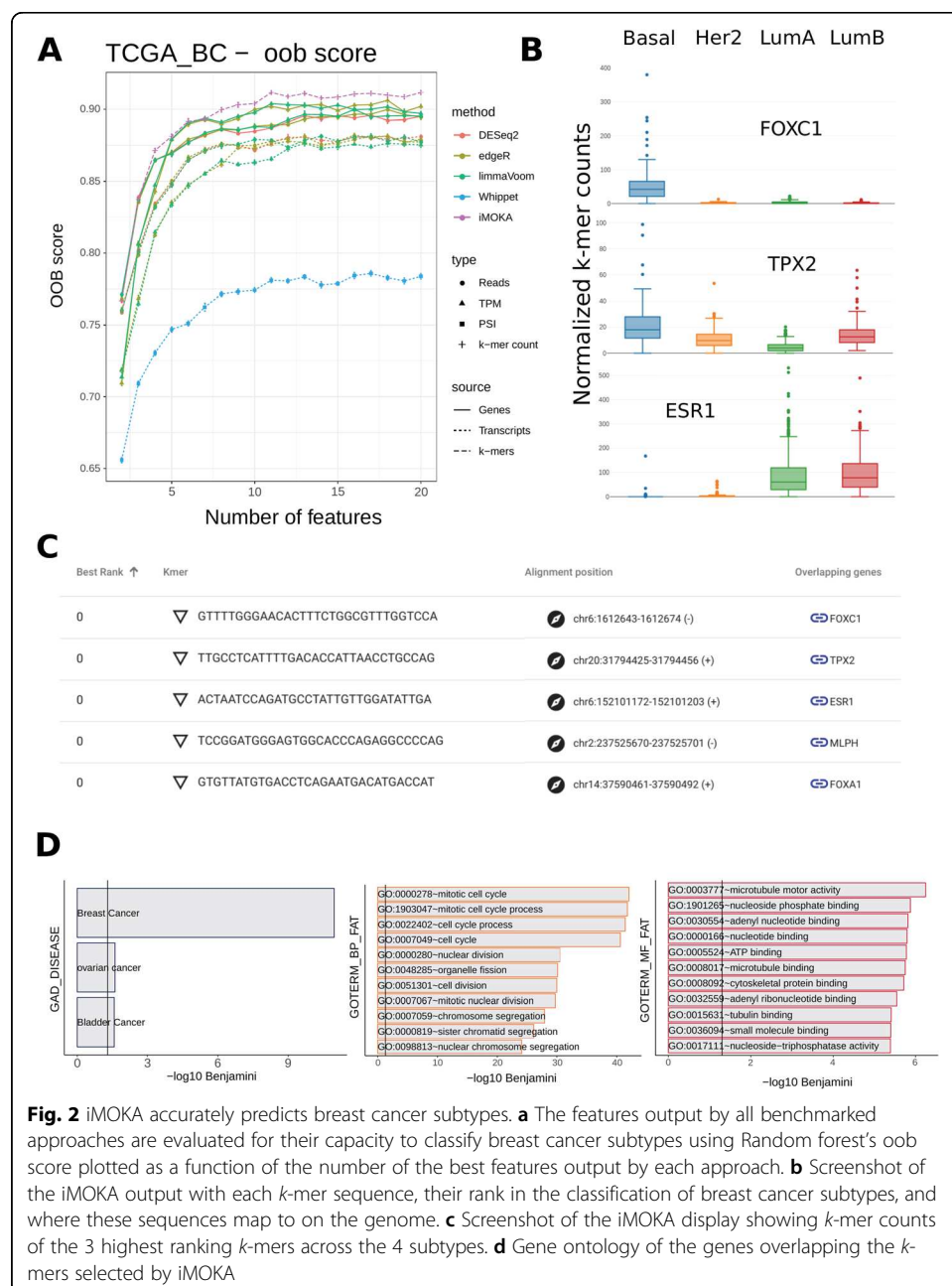


Fig. 2 iMOKA accurately predicts breast cancer subtypes. **a** The features output by all benchmarked approaches are evaluated for their capacity to classify breast cancer subtypes using Random forest's oob score plotted as a function of the number of the best features output by each approach. **b** 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

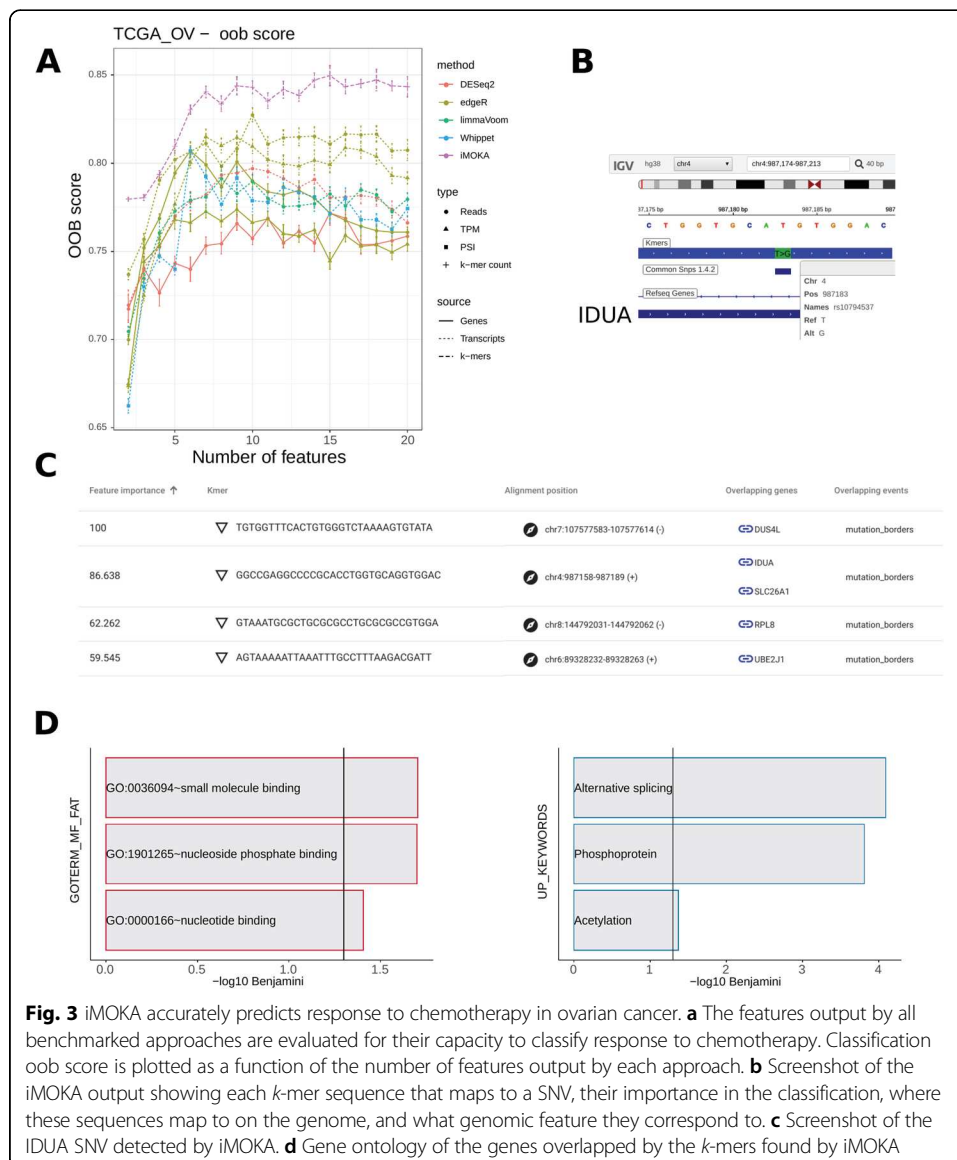


Fig. 3 iMOKA accurately predicts response to chemotherapy in ovarian cancer. **a** The features output by all benchmarked approaches are evaluated for their capacity to classify response to chemotherapy. Classification oob score is plotted as a function of the number of features output by each approach. **b** Screenshot of the iMOKA output showing each *k*-mer sequence that maps to a SNV, their importance in the classification, where these sequences map to on the genome, and what genomic feature they correspond to. **c** Screenshot of the IDUA SNV detected by iMOKA. **d** Gene ontology of the genes overlapped by the *k*-mers found by iMOKA

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].

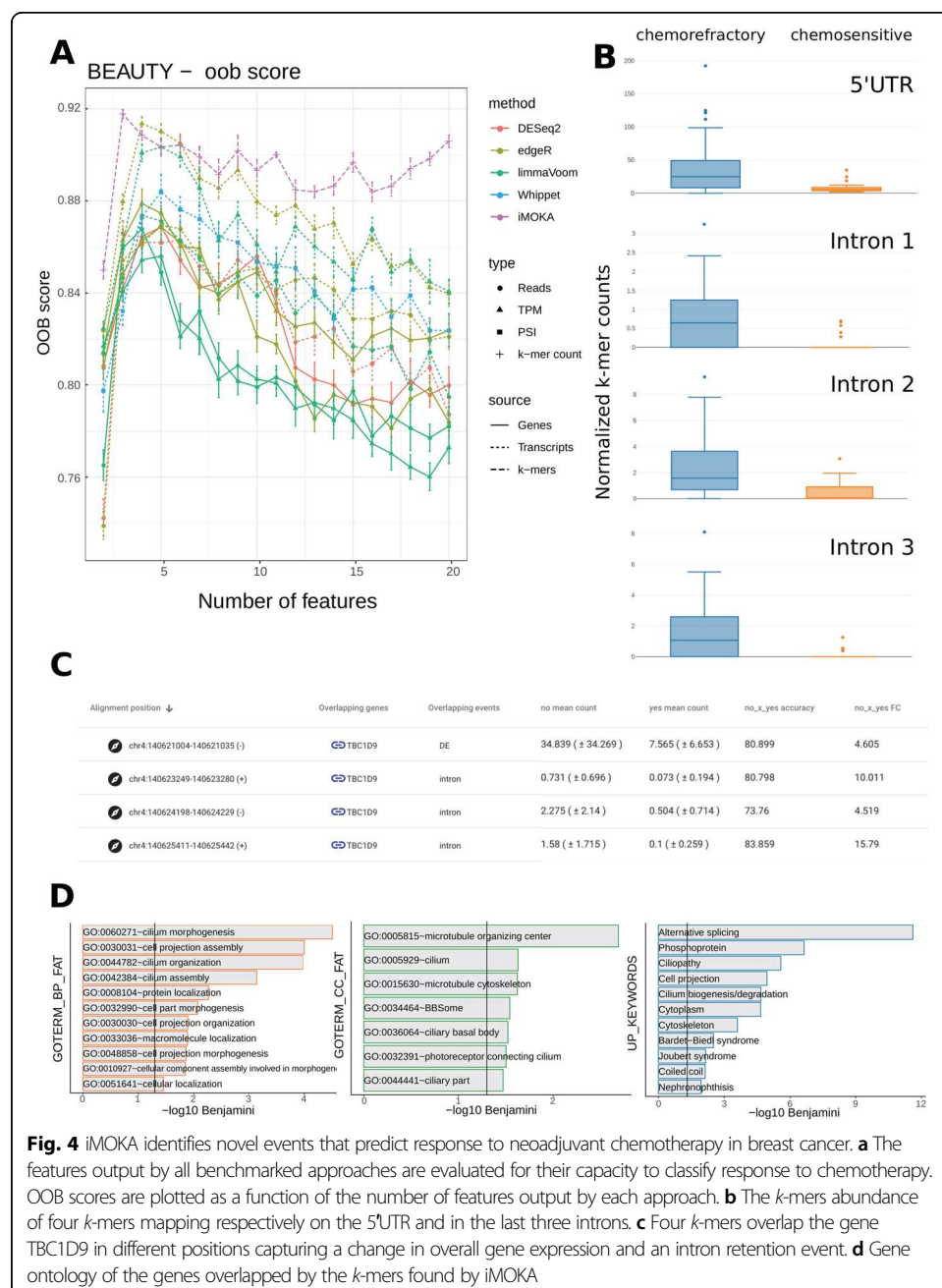


Fig. 4 iMOKA identifies novel events that predict response to neoadjuvant chemotherapy in breast cancer. **a** The features output by all benchmarked approaches are evaluated for their capacity to classify response to chemotherapy. OOB scores are plotted as a function of the number of features output by each approach. **b** The *k*-mers abundance of four *k*-mers mapping respectively on the 5'UTR and in the last three introns. **c** Four *k*-mers overlap the gene TBC1D9 in different positions capturing a change in overall gene expression and an intron retention event. **d** Gene ontology of the genes overlapped by the *k*-mers found by iMOKA

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.

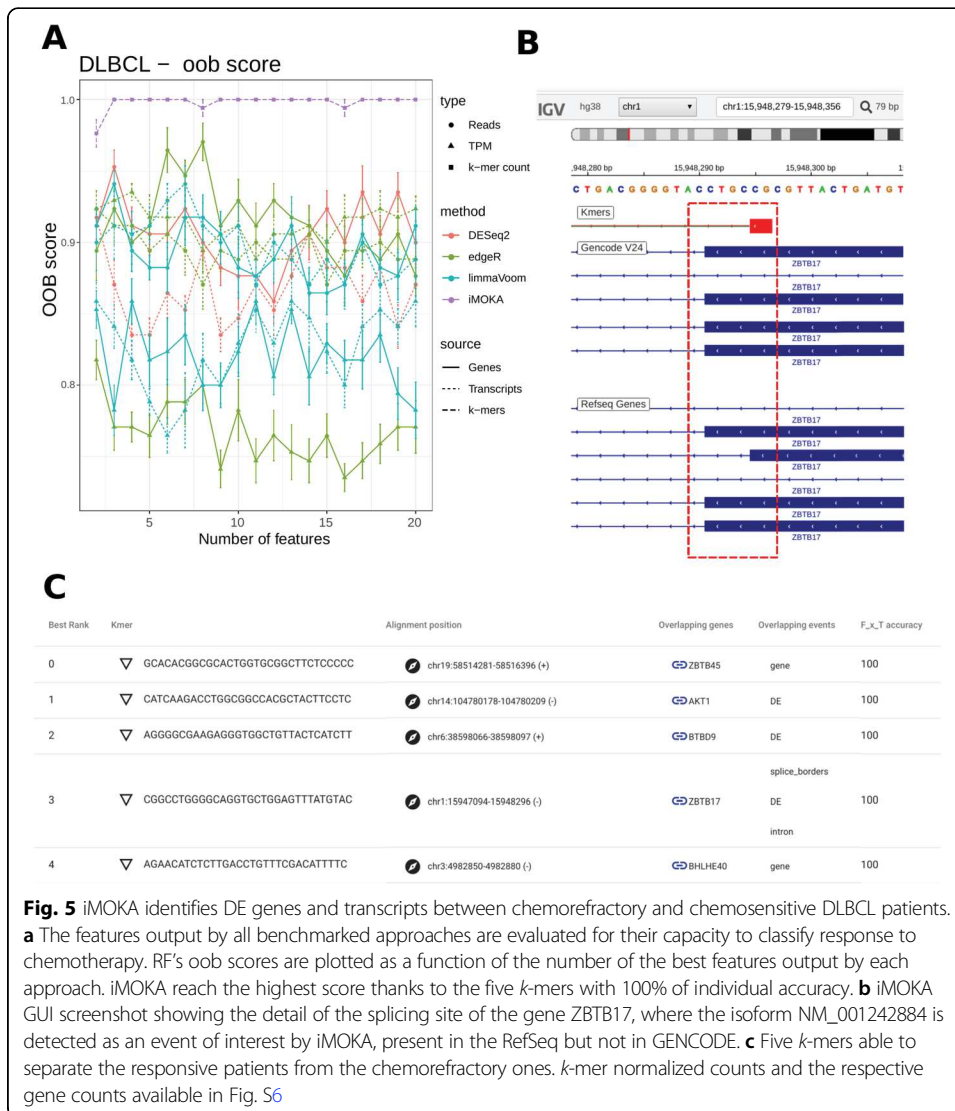
iMOKA runtimes and disk space

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.

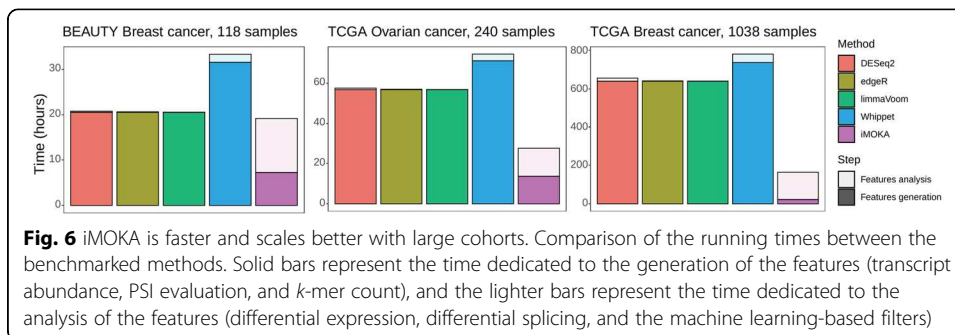
iMOKA's most intensive task is the generation of informative k -mers, where a large amount of data is filtered and aggregated, while the other benchmarked approaches handle data that are already filtered (reads are already mapped to annotated regions). Finally, most methods that calculate differential 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 differentially 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



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 storing 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 nucleotide 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:

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

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 [44] 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:

$$K_0 = \frac{4^{31} - 1}{2} \times 0 = 0 = \text{AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA}$$

$$K_1 = \frac{4^{31} - 1}{2} \times 1 = 2305843009213693952 = \text{GAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA}$$

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:

$$buff = \frac{RAM_{avail}}{\alpha \times N \times T}$$

where

$Buff$ is the length of the buffer

RAM_{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

α 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:

$$n_{\text{test}} = \text{round}(n_{\text{min}} * p_{\text{test}})$$

$$n_{\text{train}} = n_{\text{min}} - n_{\text{test}}$$

where:

n_{min} is the dimension of the smallest group

n_{test} and n_{train} are respectively the dimension of the test and training sets

p_{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 :

$$P(X = x_k \vee Y = y_j)$$

$$P(Y = y_j)$$

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

$$P(Y = y_i) = P(Y = y_{j+1}) = \frac{1}{N_{\text{labels}}}$$

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:

$$acc_k = \frac{T}{n_{\text{test}}} \times 100$$

where

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 $C_k = (c_{k0}, c_{k1}, \dots, c_{kn})$, 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_{thr} , tuned according to the lowest entropy detected in the previous batch of k -mers that passed the accuracy filter (H_{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_{min} . During the analysis, H_{thr} is updated when more than E_{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:

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

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

ELSE :

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

The adjustment parameters $a_1 \gg a_2$ ensure that the new threshold is not set too close to the minimum H_{min} .

The number of k -mers required to update the threshold (E_{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 psix 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>.

Additional file 1. TCGA_BC_aggregated.json - iMOKA results for the dataset TCGA_BC.

Additional file 2. TCGA_OV_aggregated.json - iMOKA results for the dataset TCGA_OV.

Additional file 3. BEAUTY_aggregated.json - iMOKA results for the dataset BEAUTY.

Additional file 4. DLBCL_aggregated.json - iMOKA results for the dataset DLBCL.

Additional file 5. GO - folder containing the DAVID gene ontology result for each dataset.

Additional file 6. iMOKA_supplementary.docx - Supplementary materials.

Additional file 7. Supplementary Figures S1-S7.

Additional file 8. Review history.

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., S.B., 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.p8 [59]; the BEAUTY dataset [31] is available under the dbGaP study accession identifier phs001050.v1.p1 [59]. Restrictions apply to the availability of these data, which were used under license for those studies, and so are not publicly available. Data are however available by submitting a request to the respective repositories.

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

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

Received: 6 May 2020 Accepted: 10 September 2020

Published online: 13 October 2020

References

- Learn CA, et al. Resistance to tyrosine kinase inhibition by mutant epidermal growth factor receptor variant III contributes to the neoplastic phenotype of glioblastoma multiforme. *Clin. Cancer Res.* 2004;10:3216–24.
- Zhang Z-M, et al. Pygo2 activates MDR1 expression and mediates chemoresistance in breast cancer via the Wnt/ β -catenin pathway. *Oncogene.* 2016;35:4787–97.
- Martín-Martín N, et al. Stratification and therapeutic potential of PML in metastatic breast cancer. *Nat Commun.* 2016;7:12595.
- Grossman RL, et al. Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* 2016;375:1109–12.
- Audoux J, et al. DE-kupl: exhaustive capture of biological variation in RNA-seq data through k-mer decomposition. *Genome Biol.* 2017;18:243.
- Kirk, J. M. et al. Functional classification of long non-coding RNAs by k-mer content. *Nat. Genet.* 50, 1474–1482 (2018).
- Ounit R, Wanamaker S, Close TJ, Lonardi S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics.* 2015;16:236.
- Breitwieser FP, Baker DN, Salzberg SL. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biol.* 2018;19:198.
- Thomas A, et al. GECKO is a genetic algorithm to classify and explore high throughput sequencing data. *Commun. Biol.* 2019;2:222.
- Kokot M, Dlugosz M, Deorowicz S. KMC 3: counting and manipulating k-mer statistics. *Bioinforma. Oxf. Engl.* 2017;33:2759–61.
- Sacomoto GAT, et al. KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. *BMC Bioinformatics.* 2012;13(Suppl 6):S5.
- Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15:550.
- Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* 2010;26:139–40.
- Ritchie ME, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015;43:e47.
- Sterne-Weiler T, Weatheritt RJ, Best AJ, Ha KCH, Blencowe BJ. Efficient and accurate quantitative profiling of alternative splicing patterns of any complexity on a laptop. *Mol. Cell.* 2018;72:187–200.e6.
- Rahman A, Hallgrímsson I, Eisen M, Pachter L. Association mapping from sequencing reads using k-mers. *eLife* 2018;7:e32920.

17. Drouin A, et al. Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC Genomics*. 2016;17:754.
18. Hastie T, Tibshirani R, Friedman J. *Elements of statistical learning* second edition. Math Intell. 2017;27:83–5.
19. Breiman, L. Out-of-bag estimation. in (1996).
20. Bastien RRL, et al. PAM50 breast cancer subtyping by RT-qPCR and concordance with standard clinical molecular markers. *BMC Med Genomics*. 2012;5:44.
21. Hoadley KA, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*. 2018;173:291–304.e6.
22. Jeannot E, et al. A single droplet digital PCR for ESR1 activating mutations detection in plasma. *Oncogene*. 2020;39:2987–95.
23. Ciriello G, et al. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*. 2015;163:506–19.
24. Han B, et al. FOXC1: an emerging marker and therapeutic target for cancer. *Oncogene*. 2017;36:3957–63.
25. Yang Y, et al. TPX2 promotes migration and invasion of human breast cancer cells. *Asian Pac J. Trop. Med*. 2015;8:1064–70.
26. Thakkar A, et al. High expression of three-gene signature improves prediction of relapse-free survival in estrogen receptor-positive and node-positive breast tumors. *Biomark. Insights*. 2015;10:103–12.
27. Bjørklund SS, et al. Widespread alternative exon usage in clinically distinct subtypes of invasive ductal carcinoma. *Sci. Rep*. 2017;7:5568.
28. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc*. 2009;4:44–57.
29. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*. 2011;474:609–15.
30. Villalobos VM, Wang YC, Sikic BI. Reannotation and analysis of clinical and chemotherapy outcomes in the ovarian data set from the Cancer Genome Atlas. *JCO Clin. Cancer Inform*. 2018;2:1–16.
31. Goetz M. P. et al. Tumor sequencing and patient-derived xenografts in the neoadjuvant treatment of breast cancer. *J Natl Cancer Inst*. 2017;109(7):djw306. <https://doi.org/10.1093/jnci/djw306>.
32. Yi H, Raman AT, Zhang H, Allen GI, Liu Z. Detecting hidden batch factors through data-adaptive adjustment for biological effects. *Bioinforma. Oxf. Engl*. 2018;34:1141–7.
33. Middleton R, et al. IRFinder: assessing the impact of intron retention on mammalian gene expression. *Genome Biol*. 2017;18:51.
34. Shi X, Sun X. Regulation of paclitaxel activity by microtubule-associated proteins in cancer chemotherapy. *Cancer Chemother. Pharmacol*. 2017;80:909–17.
35. Buljan VA, et al. Calcium-axonemal microtubuli interactions underlie mechanism(s) of primary cilia morphological changes. *J. Biol. Phys*. 2018;44:53–80.
36. Fornecker L-M, et al. Multi-omics dataset to decipher the complexity of drug resistance in diffuse large B-cell lymphoma. *Sci. Rep*. 2019;9.
37. Agarwal NK, et al. Transcriptional regulation of serine/threonine protein kinase (AKT) genes by glioma-associated oncogene homolog 1. *J. Biol. Chem*. 2013;288:15390–401.
38. Zhu C, Chen G, Zhao Y, Gao X-M, Wang J. Regulation of the development and function of B cells by ZBTB transcription factors. *Front. Immunol*. 2018;9.
39. *ncbi/sra-tools*. (NCBI - National Center for Biotechnology Information/NLM/NIH, 2020) <https://github.com/ncbi/sra-tools>.
40. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9. <https://doi.org/10.1093/bioinformatics/btp352>.
41. Shen W, Le S, Li Y, Hu F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One*. 2016;11(10):e0163962. Published 2016 Oct 5. <https://doi.org/10.1371/journal.pone.0163962>.
42. FastQC: a quality control tool for high throughput sequence data – <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
43. Park G, Hwang H-K, Nicodème P, Szpankowski W. Profiles of tries. *SIAM J. Comput*. 2009;38:1821–80.
44. L. Dagum and R. Menon. "OpenMP: an industry standard API for shared-memory programming," in *IEEE Computational Science and Engineering*. 1998;5(1):46–55. <https://doi.org/10.1109/99.660313>.
45. Curtin R, et al. mlpack 3: a fast, flexible machine learning library. *J. Open Source Softw*. 2018;3:726.
46. Dzubitzky, W., Granzow, M. & Berrar, D. P. *Fundamentals of data mining in genomics and proteomics*. (Springer Science & Business Media, 2007).
47. Shannon, C. E. The mathematical theory of communication. 1963. *MD Comput. Comput. Med. Pract*. 14, 306–317 (1997).
48. Sanderson C, Curtin R. Armadillo: a template-based C++ library for linear algebra. *J. Open Source Softw*. 2016;1:26.
49. CEPHES Mathematical function library. <http://www.netlib.org/cephes/>.
50. Lightweight C++ command line option parser. jarro2783/cxxopts. 2020. <https://github.com/jarro2783/cxxopts>.
51. JSON for Modern C++, N. nlohmann/json. 2020. <https://github.com/nlohmann/json>.
52. van der Walt S, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng*. 2011. <https://doi.org/10.1109/MCSE.2011.37>.
53. McKinney, W. Data structures for statistical computing in Python. *Proc. 9th Python Sci. Conf.* (2010).
54. Pedregosa F, et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res*. 2011;12:2825–30.
55. Federico Comitani. fcomitani/SimpSOM: v1.3.4. (Zenodo, 2019). <https://doi.org/10.5281/zenodo.2621560>.
56. Kurtzer GM, Sochat V, Bauer MW. Singularity: scientific containers for mobility of compute. *PLOS ONE*. 2017;12:e0177459.
57. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*. 2017;14:417–9.
58. Williams CR, Baccarella A, Parrish JZ, Kim CC. Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq. *BMC Bioinformatics*. 2017;18:38.
59. dbGap/database of genotypes and phenotypes/ National Center for Biotechnology Information, National Library of Medicine (NCBI/NLM) <https://www.ncbi.nlm.nih.gov/gap>.
60. Athar A. et al., 2019. ArrayExpress update - from bulk to single-cell expression data. *Nucleic Acids Res*, <https://doi.org/10.1093/nar/gky964>, PubMed ID 30357387.

61. Lorenzi, C. et al. iMOKA: k-mer based software to analyze large collections of sequencing data. (GitHub, 2020). <https://github.com/RitchieLabIGH/iMOKA>.
62. Lorenzi, C. et al. iMOKA: k-mer based software to analyze large collections of sequencing data. (Zenodo, 2020). <https://doi.org/10.5281/zenodo.4008947>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



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 IGV¹. All the graphs present in the software are generated using the javascript implementation of Plotly².

The GUI is implemented in Electron³ and in Angular⁴, making it available on multiple platforms (Linux, Windows and MacOS). iMOKA can be run from the interface, in local or on a SLURM⁵ cluster, with the only dependency of Singularity⁶.

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

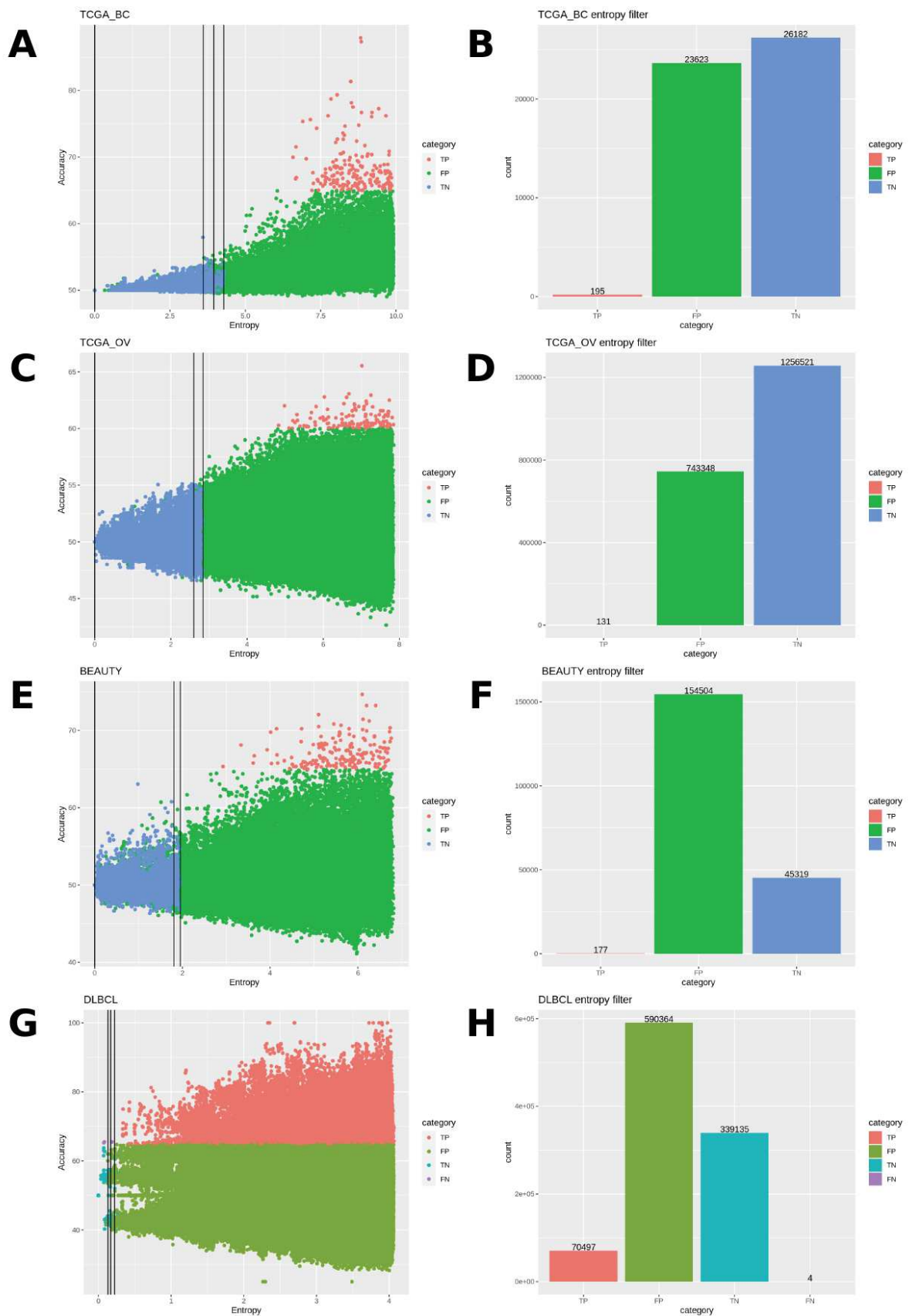
iMOKA uses a random forest classifier (RFC) to assign importance to each feature. It also uses a RFC to produce prediction models. Feature importance is estimated with a random forest with 1000 estimators (-n argument) and with min_samples_split (minimum number of samples required to split a node in 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.

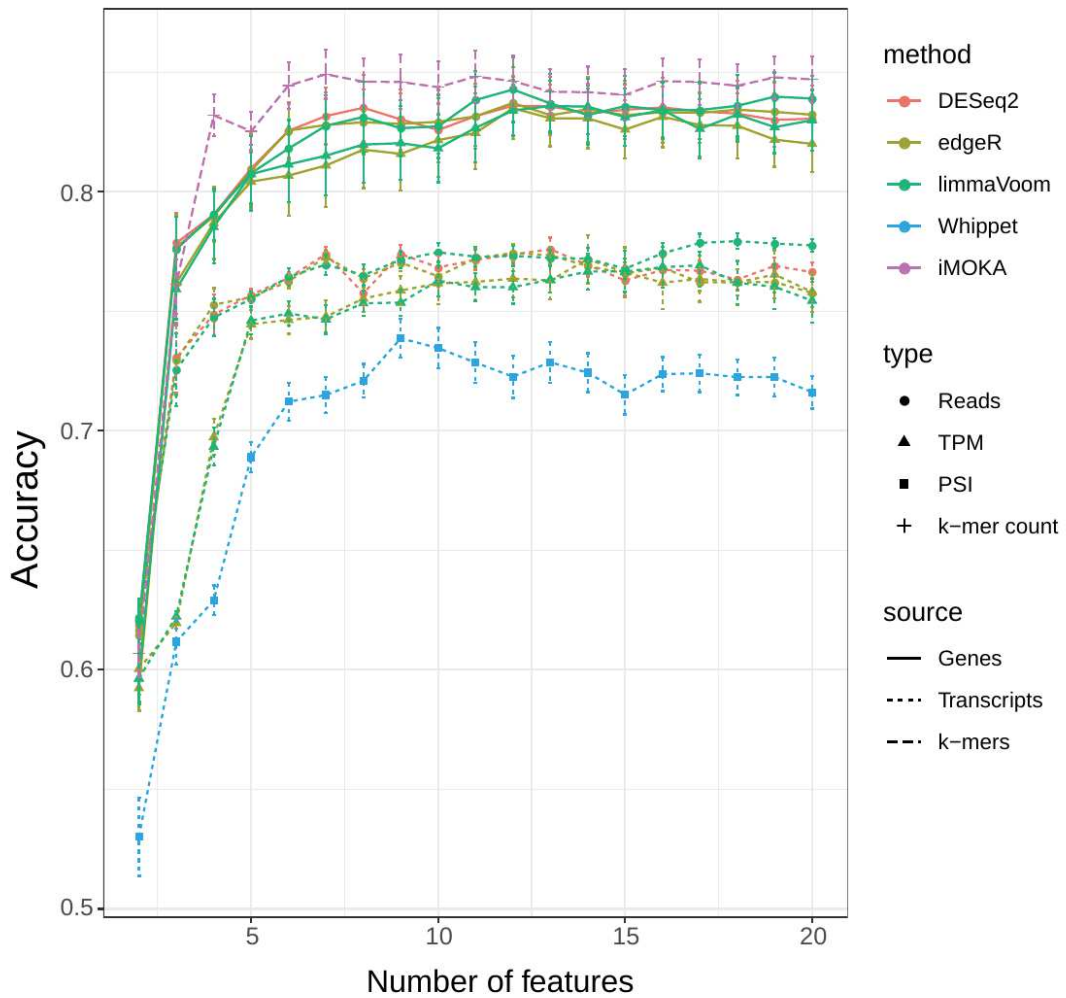
All the metrics are evaluated with Monte Carlo cross validation in a similar procedure as described in the reduction step.

Supplementary Figures

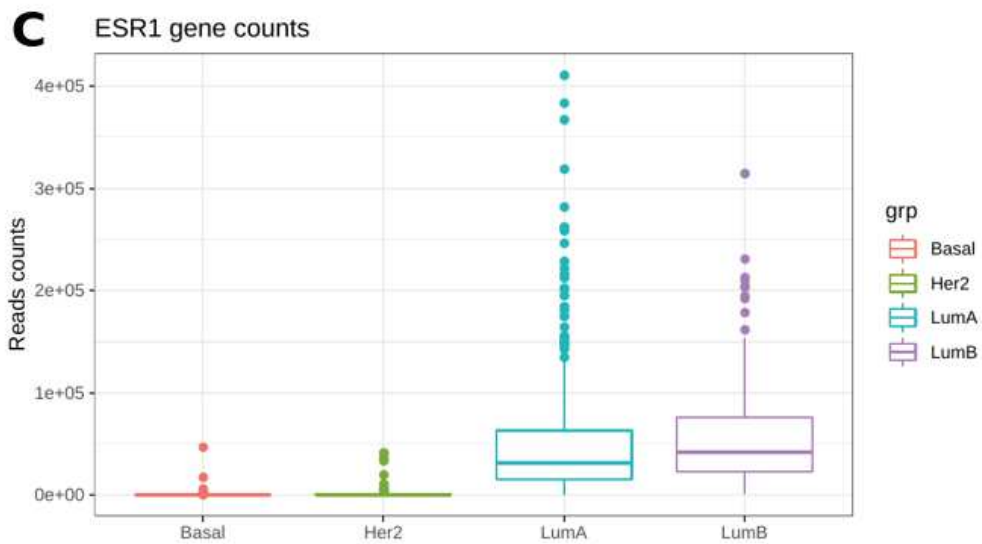
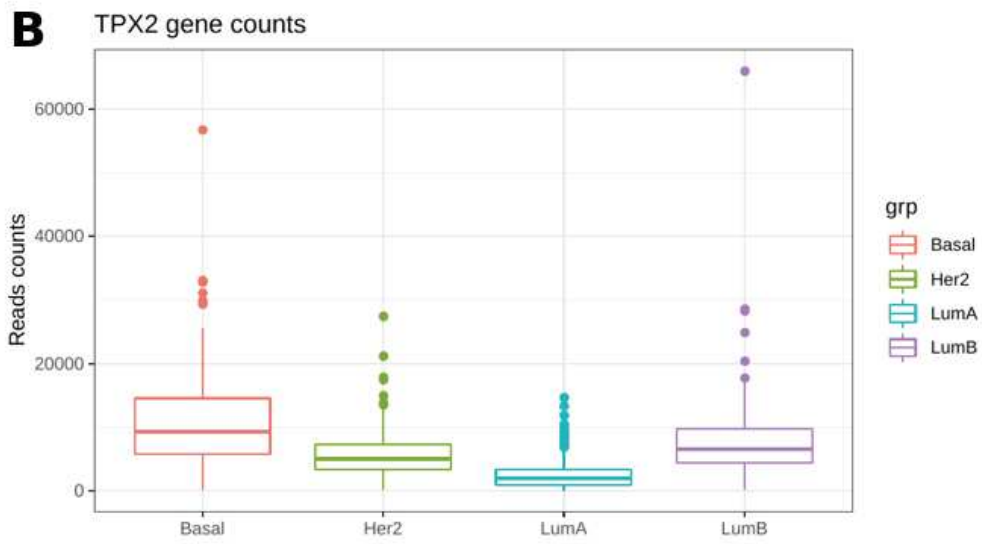
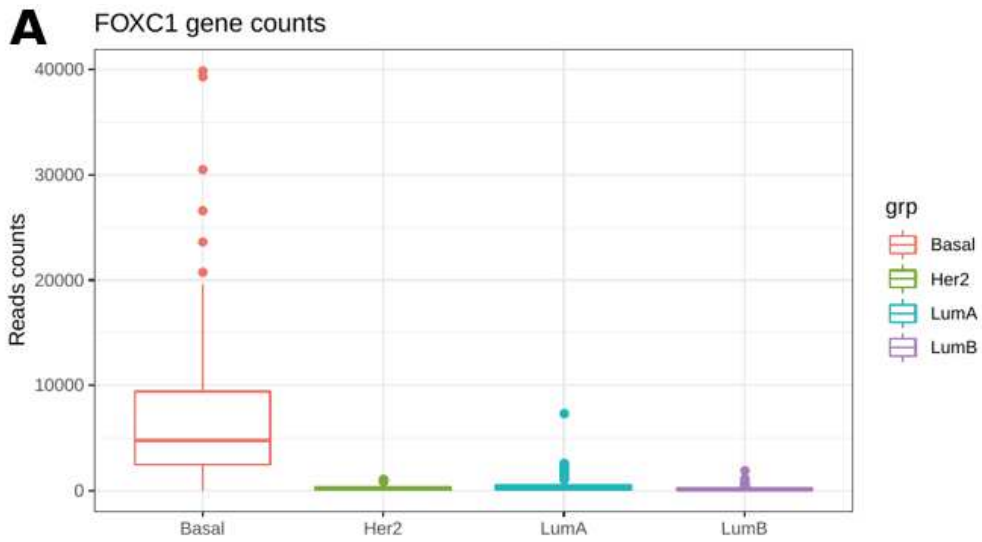


Supplementary Figure 1: k -mer count entropy in function of the accuracy estimated by the Bayesian classifier. Both the entropy and the accuracy were computed for samples of 50000, 1000000, 200000 and 1000000 k -mers respectively in TCGA BC (**A-B**), TCGA OV (**C-D**), BEAUTY (**E-F**) and DLBCL (**G-H**) 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.

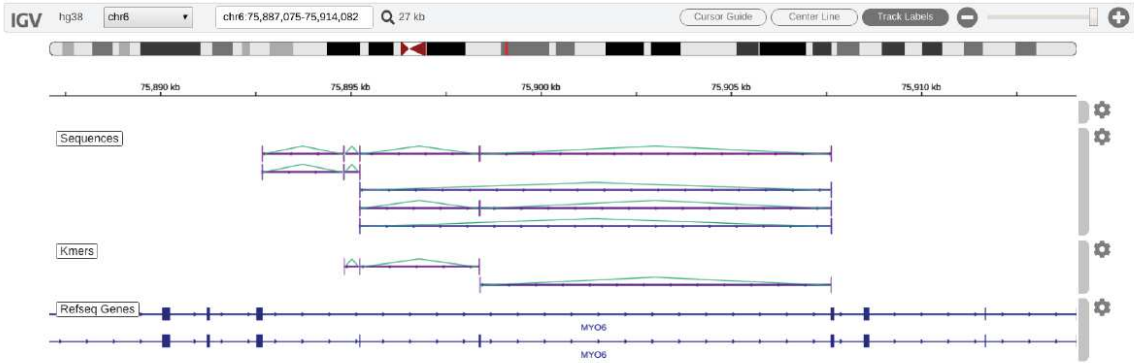
TCGA BC - 5 fold cross validation test accuracy



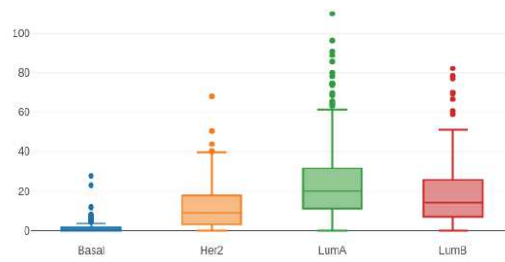
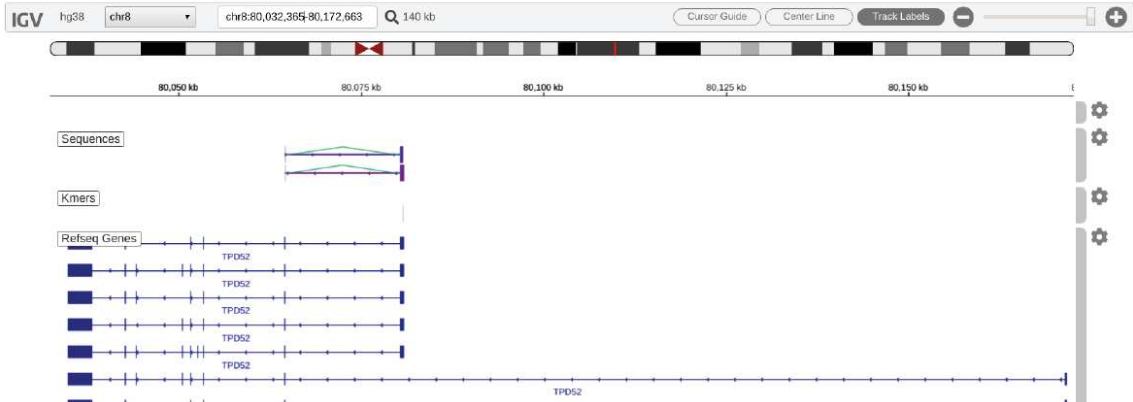
Supplementary Figure 2: 5 fold cross validation of the entire iMOKA pipeline on the TCGA BRCA dataset. For each fold, a test set is put aside at the start, before the feature reduction step.



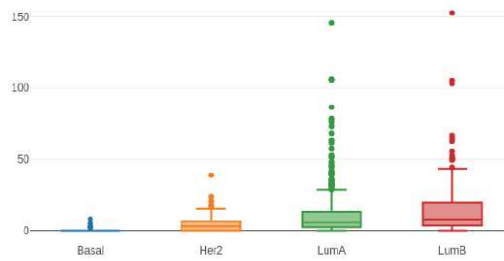
Supplementary Figure 3: Read counts of the genes FOXC1 (A), TPX2 (B) and ESR1 (C) in TCGA BRCA, whose corresponding overlapping k -mers abundances are visible in Figure 2.

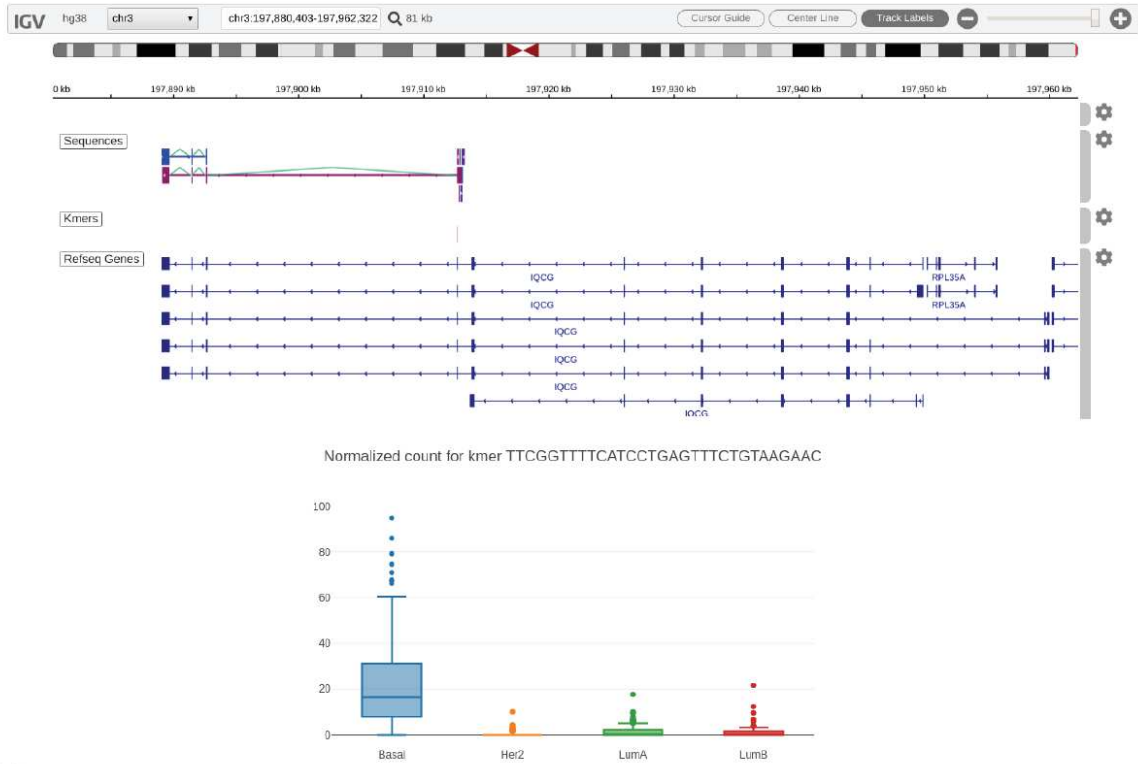
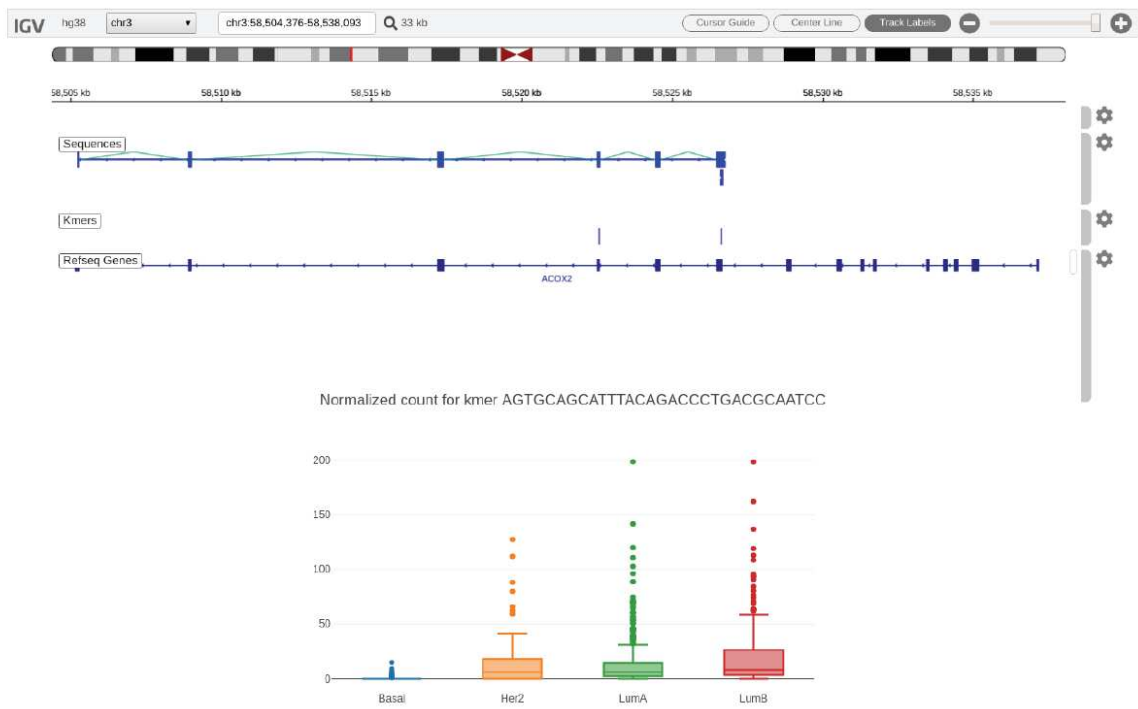
A

Normalized count for kmer TGGGTCTTGTCCATCATTTTAGAAGTTAC

**B**

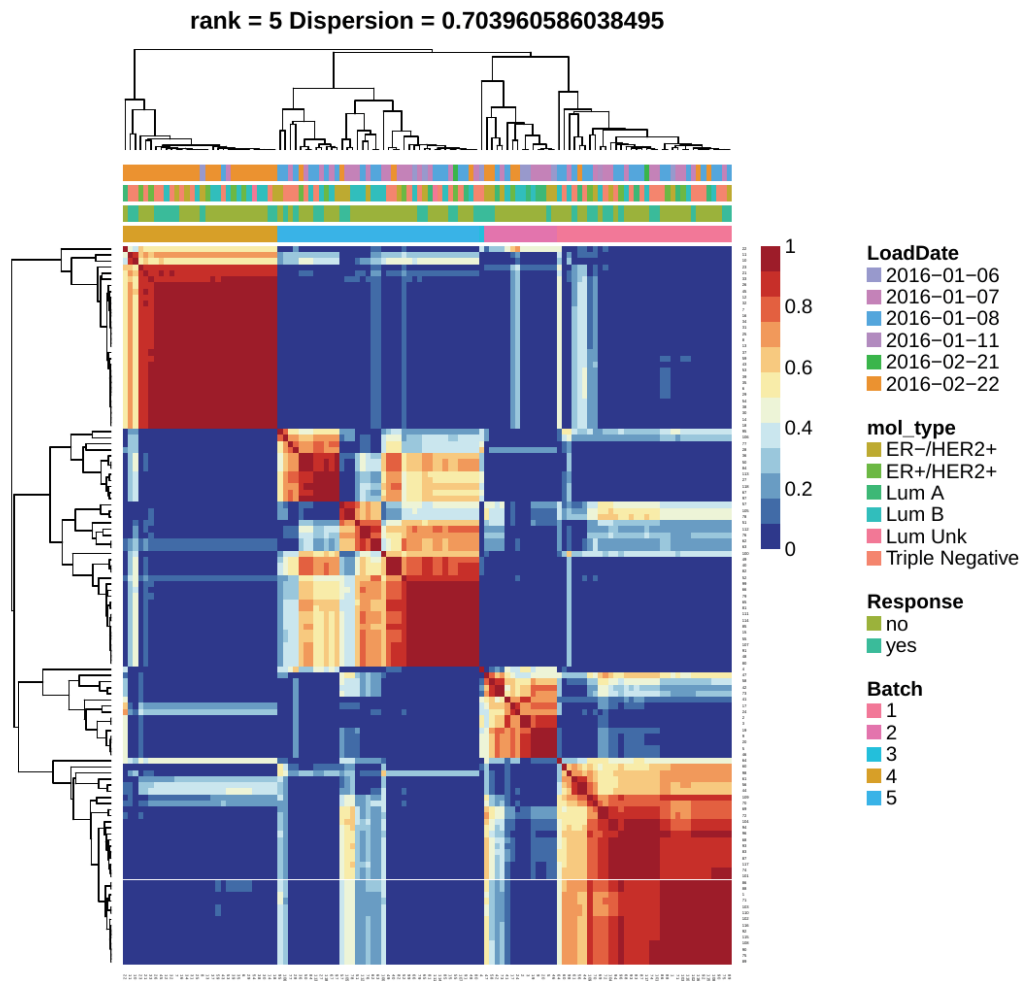
Normalized count for kmer CATATTGCAGAACCCCTGCCCTTCTTTGTGA



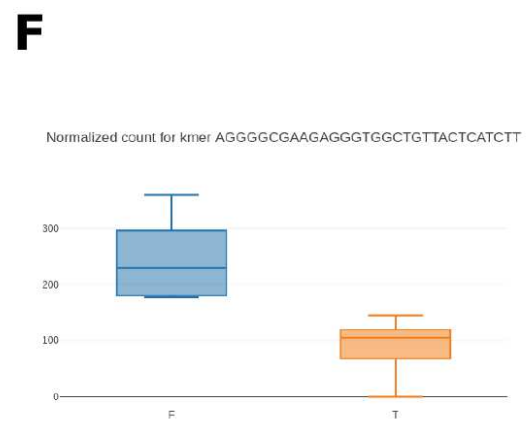
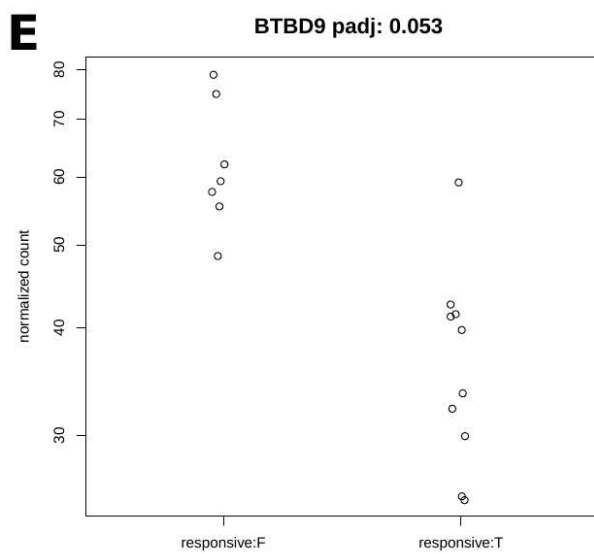
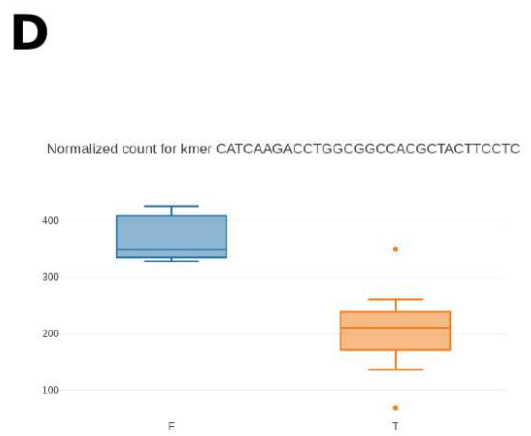
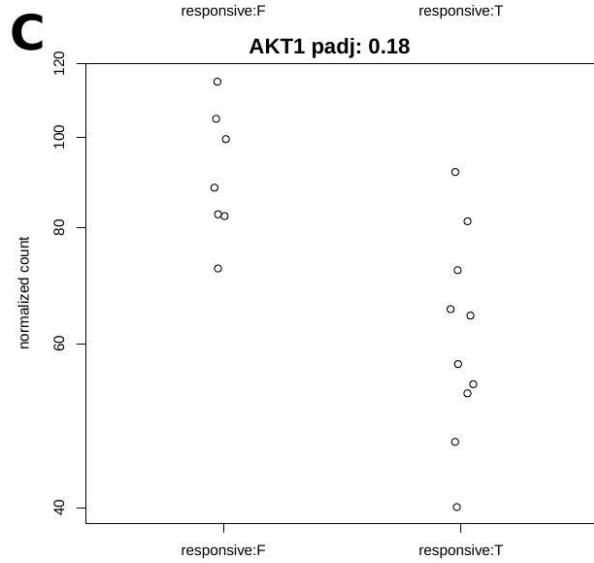
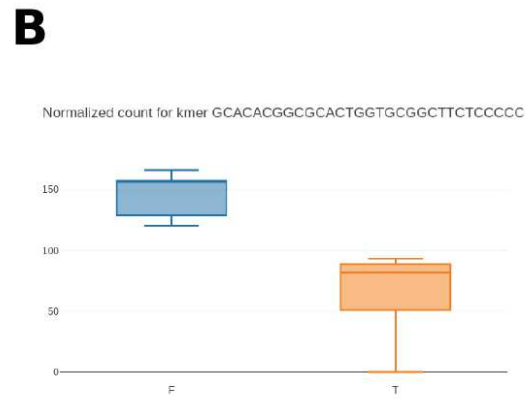
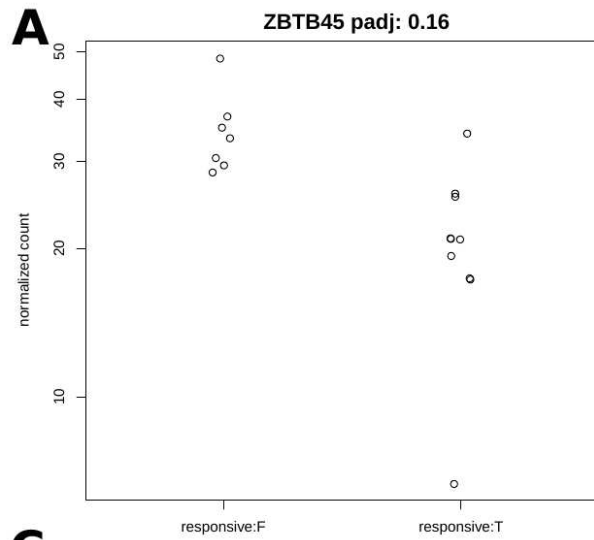
C**D**

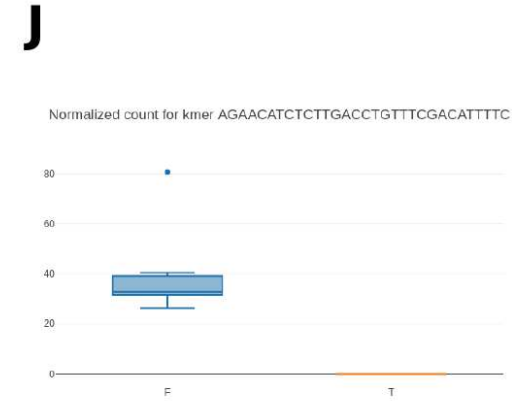
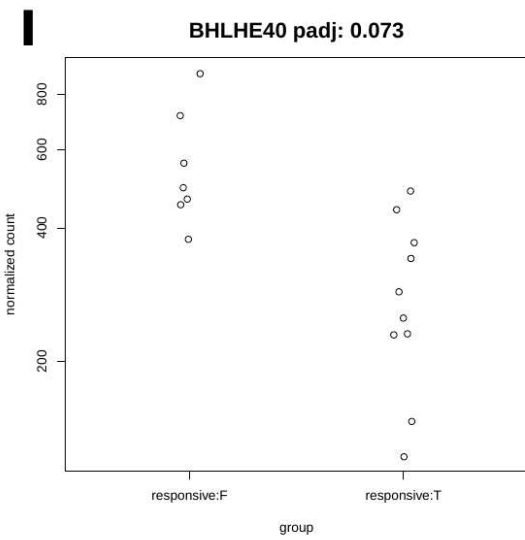
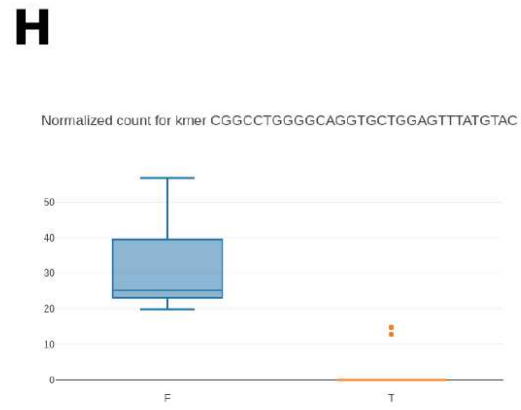
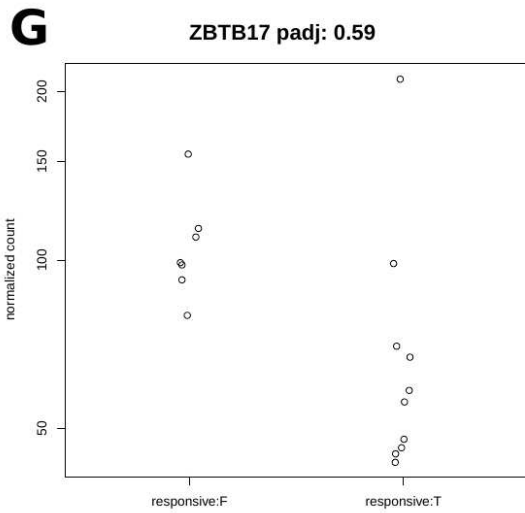
Supplementary Figure 4: IGV genome browser visualization (integrated in the iMOKA interface) of the *k*-mers that map to splicing sites and exons known to be involved in different breast cancer

molecular subtypes and the abundances of the representative *k*-mers, validated in a recent study⁷. **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 TPD52 (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.

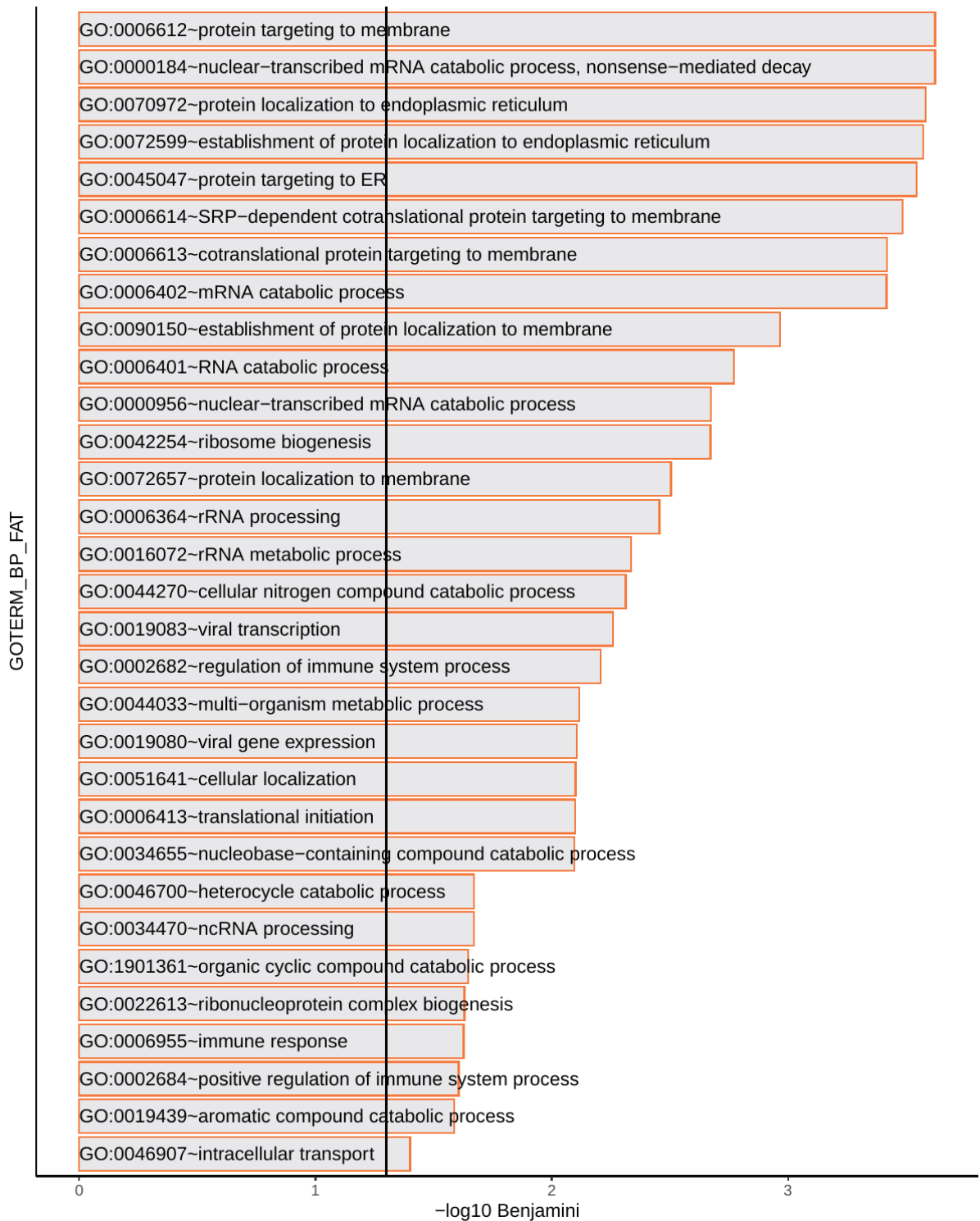


Supplementary Figure 5: identification of a hidden batch effect in BEAUTY dataset associated with the Load Date using DASC ⁸.





Supplementary Figure 6: DESeq2 normalized gene counts and the respective *k*-mer normalized abundances identified by iMOKA for 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

Study	Event	Number of <i>k</i> -mers
TCGA_BC 3002 <i>k</i> -mers	Gene	742
	DE	580
	unmapped	523
	splice_borders	479
	intergenic	427
	intron	162
	multiple_splice_junctions	120
	mutation_borders	86
	insertion_borders	35
	deletion_borders	12
	misalign	1
TCGA_OV 138 <i>k</i> -mers	gene	105
	unmapped	14
	mutation_borders	11
	intergenic	8
BEAUTY 1248 <i>k</i> -mers	gene	659
	DE	20
	unmapped	138
	splice_borders	13
	intergenic	377
	Intron	25

	multiple_splice_junctions	1
	mutation_borders	13
	insertion_borders	5
DLBCL 1915 <i>k</i> -mers	splice_borders	588
	DE	429
	intron	74
	multiple_splice_junction	12
	gene	688
	unmapped	168
	intergenic	3
	multimap	108

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.

Event name	Condition	Description
insertion_borders deletion_borders mutation_borders	<i>k</i> -mer overlaps an insertion/deletion/mutation and has a higher accuracy score than other <i>k</i> -mers within the same kmer graph.	This event overlaps a known variation event.
splice_borders	A splicing site where the overlapping <i>k</i> -mers have a higher accuracy than the others.	This event might overlap with a splicing site involved with alternative splicing

		derived isoforms or intron retention events.
DE	An annotated transcript is covered by filtered <i>k</i> -mers across more than 50% of its length (by default).	Generally a differentially expressed transcript.
Intron	An annotated intron is covered by filtered <i>k</i> -mers across more than 50% of its length (by default).	This event might overlap with a retained intron, an alternative starting site or an intronic transcript.
multiple_splice_junction	This event is called when there are two splicing sites with one common acceptor or donor site.	This splicing site can indicate two differentially expressed transcripts or a different transcript usage.
Intergenic	This event represents a sequence that mapped correctly but doesn't overlap with any feature of the given annotation file.	These events might represent novel or unannotated transcripts.
gene	This event is generic and represents <i>k</i> -mers that map on a gene without any other event associated with it.	This event is generally associated with very small transcripts or overlapping genes.
unmapped	Sequences that didn't map on the given reference genome.	These can be due to contamination, repetitive elements excluded by the aligner or chimeric transcripts.
multimap	Sequences that mapped to multiple positions of the reference genome with the same score	These can be repetitive elements, portions of pseudogenes or similar cases.

misalign	The sequence generated by the graph was mapped correctly, but the best k -mer is completely in the clipped region.	Those events should be rare and require a manual investigation.
----------	--	---

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

Supplementary references:

1. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
2. Plotly, T. Plotly: Collaborative data science. <https://plot.ly>. <https://plot.ly> (2015).
3. Electron- <https://www.electronjs.org/>.
4. Angular - <https://angular.io/>.
5. Yoo, A. B., Jette, M. A. & Grondona, M. SLURM: Simple Linux Utility for Resource Management. in *Job Scheduling Strategies for Parallel Processing* (eds. Feitelson, D., Rudolph, L. & Schwiegelshohn, U.) 44–60 (Springer, 2003). doi:10.1007/10968987_3.
6. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
7. Bjørklund, S. S. *et al.* Widespread alternative exon usage in clinically distinct subtypes of Invasive Ductal Carcinoma. *Sci. Rep.* **7**, 5568 (2017).

