UNIVERSITÀ
DI TORINO

University of Turin
Department of Life Sciences and Systems Biology

Dissertation submitted for the degree of Doctor of Philosophy in Complex Systems for Life Sciences
XXXV Cycle

Bioinformatics driven data mining: a fundamental instrument for uncovering novel insights in the fields of cancer and immunology.

Vladimir Nosi

Tutors

Raffaele Calogero

Francesca Cordero

PhD Program Coordinator

Enzo Medico

Turin, 2023

Table of Contents

List of Figures

List of Tables

Acknowledgements

Numerous people have contributed more or less directly to the work present in this thesis, all important in their own right.

I want to extend great gratitude towards prof. Raffaele Adolfo Calogero and prof. Francesca Cordero for supervising me during the period of my PhD, for their advice and for the opportunities they offered me. My thanks as well to all the members of the Bioinformatics and Genomics unit (B&Gu) headed by prof. Calogero and of the quantitive-biology (q-Bio) group headed by prof. Cordero and prof. Marco Beccuti, which provided with a stimulating environment between the start of my PhD and the time when I left for Basel. In particular, thank you dr. Luca Alessandri for his invaluable contributions that guided the works I participated in during the first part of my PhD.

Thanks as well to prof. Ivan Molineris for his guidance during our brief but edifying collaboration.

I'm also very grateful towards prof. Gennaro De Libero for letting me in his laboratory and introducing me to immunology, as well as for his guidance, which I thank prof. Petr Hruz and dr. Lucia Mori for as well. I feel indebted towards the members of the Experimental Immunology laboratory for providing me a home away from home, in particular to dr. Jan Devan for the long talks on immunology and more and for slowly teaching me how to play some proper table tennis, to dr. Andrew Chancellor for having the patience to work with me as I was taking my first steps in this new research environment and to dr. Julian Spagnuolo for welcoming me into the world of immunoinformatics with open arms, but the feeling extends to all the members of the laboratory, past and present during my time here: dr. Giuliano Berloffa, dr. Daniel Constantin, dr. Alessandro Vacchini, dr. Gennaro Prota, dr. Marco Stringhini, Aisha Beshirova, Rodrigo Colombo, Natalie Kehrer, Giulia Montanelli, Pedro Loureiro, Vanshika Rastogi, Verena Schäfer, dr. Corinne De Gregorio and dr. Qinmei Yang.

Finally, thanks to my friends and family back in Turin, who have always made me feel as if I never left whenever I've had the chance to visit and spend time with them.

Abstract

The advent of big data has transformed the landscape of biomedical research, moving from a data-poor to a data-rich environment. The use of high-throughput technologies has not only increased the amount of data produced in single experiments, but also lowered the associated costs, making it accessible to small laboratories. This proliferation of accessible technologies has allowed for the investigation of multiple features spanning across different omics levels. However, appropriate methods and tools to extract knowledge from this vast and diverse resource are not always in place, creating a need for a multi-omics approach to better understand disease processes. During the course of his three-year PhD, the author was involved in various projects, focusing on the development of computational approaches in different research areas. The PhD work was divided into three main topics. The first involved the creation of a neural network solution for the fast screening of exon 14 skipping events in MET from RNA-seq data, which is a common alteration in different cancer types. The second focused on the development of a case study for Laniakea@ReCaS validation, a service that provides cloud resources to be used for the implementation of on-demand instances of Galaxy, a platform for data integration and analysis. The third involved the bioinformatic mining of high-throughput data generated in the Experimental Immunology group, specifically focused on T-cell immune-repertoires. The projects discussed highlight the need for appropriate methods and tools to extract knowledge from the vast and diverse data available in a data-rich environment. The author's work demonstrates the importance of a multi-omics approach in understanding disease processes and the development of computational approaches to improve data analysis and integration.

Thesis Structure

During my three years of PhD I was involved in various projects. The decision to focus on multiple projects came from the need of a mature bioinformatician, involved in data mining, to have a broad knowledge of the use of computation approaches in different research areas. Specifically, at the beginning of the PhD, I was involved in a project in which we try to grab new knowledge by the integration of multi-omics data to understand the mechanisms underlying the resistance to crizotinib in ALK-driven lymphomas. Unfortunately, after few months of work, we realized that the data was not sufficient to fulfill our aim and sadly we had to abandon this project till further data will be available. Because of the limited results I obtained on this project, it is not inserted in the thesis. The rest of my PhD work was divided on three main topics:

- Depicting disease linked RNA isoforms by means of deep learning
    - in Chapter 1 I describe the work conducted during the first year and a half of my PhD, as part of the development of the MET Observatory project. In particular, my work was focused on the creation of a neural network solution for the fast screening from RNA-seq data of exon 14 skipping events in MET, which is a common alteration in different cancer types.
- Development of a case study for Laniakea@ReCaS validation.
    - In Chapter 2 I recap my contribution to the Laniakea@ReCaS project, a service that provides cloud resources to be used for the implementation of on-demand instances of Galaxy, a platform for data integration and analysis with the aims to make computational biology approachable to scientists without computational expertise.
- Mining T-cell immune-repertoires.
    - At the conclusion of my participation in the MET Observatory project, I was given the opportunity to move to the University of Basel as part of a collaboration with the Experimental Immunology group (EXPI), led by Prof. Gennaro De Libero. This period was specifically devoted to the application of

available computational methodologies to bioinformatically mine the high-throughput data generated in EXPI. In Chapter 3, I present the work done within two projects in which I was involved with during my stay at EXPI.

Thus, the thesis is divided in three chapters and, because the application fields are quite different, instead of providing a unique introduction there is a topic specific introduction in each chapter.

Chapter 1

metObservatory

1. Introduction

1.1. Alternative splicing

1.1.1. Mechanisms of alternative splicing

All eukaryotes carry introns in at least some of their genes. These introns are removed from precursor messenger RNAs (pre-mRNAs) in the process of splicing that leads to mature mRNAs, with two transesterification steps catalyzed by the spliceosome, a ribonucleoprotein complex with a highly dynamic composition that allows for accuracy and flexibility [1].

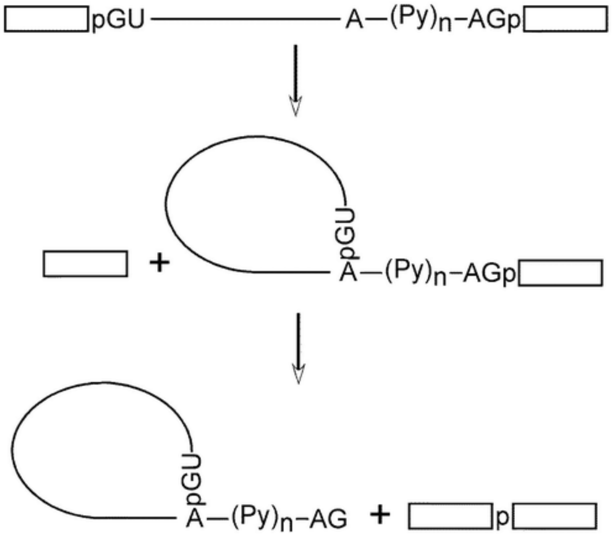


***Figure 1**. Two transesterifications reaction catalyzed by the spliceosome resulting in splicing of an intron in pre-mRNAs. Boxes represent the exons; the intron is represented by the solid line. The phosphodiester bonds that get broken and established during the reaction are represented by the letter p. Figure from [2].*

Two splice sites are present at the junctions between introns and exons, the 5' site (donor site) presents a nearly invariant GU dinucleotide included in a longer less conserved consensus region, while the 3' side of the intron presents a branch site including an adenine, a polypyrimidine tract and a highly conserved AG dinucleotide (acceptor site) (**Figure 1**). In the first step of the reaction the 2'-hydroxyl group of the adenine in the branch site performs a nucleophilic attack on the phosphodiester bond at the 5' splice site, this cleaves the exon from the intron at the 5' end, which then ligates to the 2'-hydroxyl of the adenine in the branch site, forming a lariat attached to the 3' exon. In the second step of the reaction the 3'-hydroxyl of the detached 5' exon carries a second nucleophilic attack towards the phosphodiester bond at the 3'end of the intron, resulting in the ligation of the 5' exon with the 3' exon and the detachment of the intron as the lariat formed in the previous step [2].

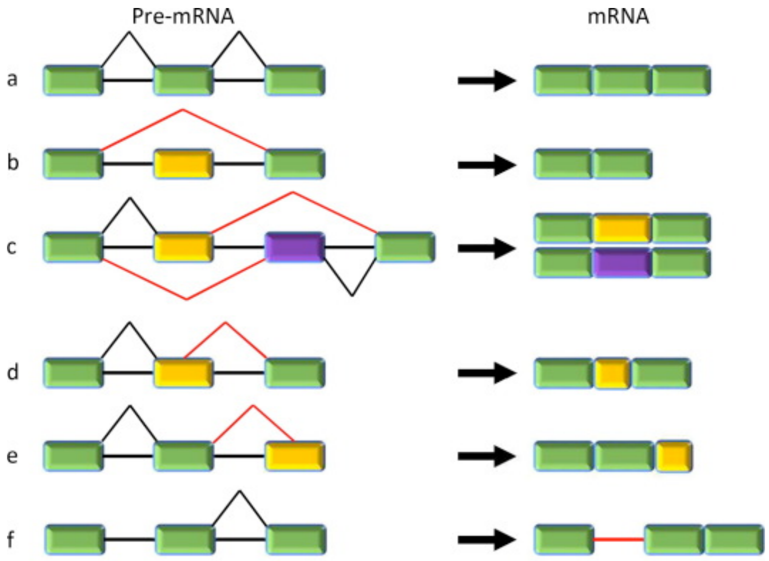


***Figure 2**. Different forms of splicing. (**A**) Constitutive splicing. (**B**) Exon skipping. (**C**) Mutually exclusive exons. (**D, E**) Alternative 5' and 3' splicing sites. (**F**) Retained intron. Figure from [3].*

This form of splicing in which the spliceosome removes introns from a pre-mRNA molecule is called constitutive splicing (**Figure 2A**), and if it was the only form of splicing, only one transcript from a given gene would be produced, containing all exons in order. However, the spliceosome is also involved in the process of alternative splicing, a post-transcriptional modification that from a single gene allows multiple different transcripts to be created [4] together with events that are often discussed in the same context as alternative splicing but are not splicing dependent, such as alternative promoters and alternative polyadenylation. Alternative splicing events are generally divided into 5 main categories [3] depending on what distinguishes the event from a constitutive splicing event: exon skipping events have a loss of an exon, called cassette exon, in the final transcript (**Figure 2B**); mutually exclusive exons events are cases in which different transcripts can only include one of the exons that are mutually exclusive (**Figure 2C**); alternative 5' and 3' splicing sites are similar events in which an exon is shortened or lengthened due to the presence of an additional splicing site in the 5' or 3' ends of an intron (**Figure 2D-E**); retained introns are introns that are not removed during the splicing process and end up being part of the mature mRNA (**Figure 2F**). The regulation of these splicing events, both constitutive and alternative, is dependent on both cis and trans elements, beyond the strength of the splice site consensus sequences [5, 6]. The auxiliary cis elements are represented by enhancer and silencer sequences (**Figure 3**) located on introns (intron splicing enhancers, ISE; intron splicing silencers, ISS) or exons (exon splicing enhancers, ESE; exon splicing silencers, ESS) and the trans elements are the splicing factors that bind these sequences. Splicing factors that bind enhancer sequences, like the Serine-Arginine (SR) proteins binding ESE regions, promote the assembly of the spliceosome and facilitate exon inclusion, while factors binding ISS and ESS regions, like some members of the heterogeneous nuclear ribonucleoproteins (hnRNP) family, repress splicing by blocking the assembly of the spliceosome and recognition of splicing sites [5-8].

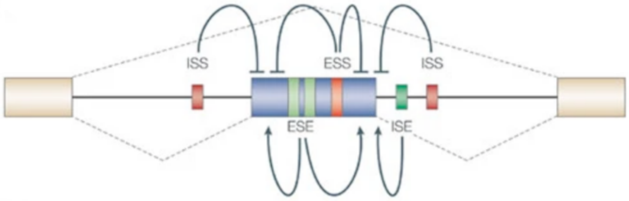


Figure 3*. ESE, ESS, ISS and ISE contribute to the regulation of splicing. Figure from [6].*

The balance between the elements promoting splicing and the ones repressing it influences the likelihood of a splice site being activated. Computational predictions suggest that around a third of alternative splicing events cause frame shifts that give rise to premature termination codons (PTCs) which cause the transcripts to be targeted by the nonsense-mediated decay (NMD) pathway for degradation [9], in a process called alternative splicing coupled to NMD (AS-NMD). On top of ensuring that the PTC-containing transcripts don't give rise to truncated proteins, AS-NMD has been shown to be a gene regulation strategy, as is observed during granulocytic differentiation, where intron retention coupled to NMD plays an important role [10, 11]. Overall, Next-Generation Sequencing (NGS) studies have shown that nearly all multi-exon pre-mRNAs in humans undergo some form of alternative splicing [12, 13], with an estimate of around 100,000 alternative splicing events present with significant frequency in major tissues in humans [13]. While it is not clear how many of these events give rise to functional transcripts [14], alternative splicing is the biggest contributor in clearing the gap between the number of human genes and the number of observed human proteins [4], often operating in a tissue-specific manner [12, 15, 16]. For example, alternative splicing has been shown to be heavily involved in various differentiation processes on top of the already mentioned granulocytic differentiation through AS-NMD gene regulation [10, 11], like the determination of cell fate during cerebral cortex development [17] through the activity of the splicing factors PTBP1 and RBFOX, the differentiation of adipocytes through the action of the splicing factor SRSF10 on lipin1 [18] and the differentiation of terminal erythropoietic cells through different mechanisms, also involving AS-NMD gene regulation [19]. Different isoforms of the same protein can actively contribute to opposite effects, like in the case of KLF6, where the SV2 variant promotes apoptosis and inhibits proliferation while the SV1 variant leads to cell proliferation [20, 21]. In general, alternative splicing affects many different biological processes [22] across human tissues, so ensuring its correct function is crucial for the health of the organism.

1.1.2. Alternative splicing in disease

Due to the ubiquitous nature of alternative splicing and its importance in numerous vital processes, it's not surprising that its dysregulation can lead to severe pathological outcomes. Alterations to either the trans-acting or the cis-acting elements of splicing can lead to disease, with alterations to the trans-acting ones having generally larger effects due to the ability to impact multiple splicing events but also being rarer. This rarity is probably due to the importance of splicing limiting the possibility of widespread mutations to splicing machinery, especially loss-of-function ones, without being lethal already during embryonic development [23, 24]. However, spliceosome core components alterations that don't involve complete loss-of-function can still take place and are linked to some genetic diseases like Retinitis Pigmentosa and Spinal Muscular Atrophy [25, 26], some craniofacial disorders like Cerebro-Costo-Mandibular Syndrome and Nager Syndrome [27] and are drivers in some cancers like the Myelodysplastic Syndromes [25, 28], the most common forms of adult malignancies of the myeloid lineage [29]. The types of cis-acting element alterations that have been found in disease is disparate, covering both exonic and intronic splice regulation sequences, but most involve the splice sites in the introns [30]. In Familial Dysautonomia, a hereditary disease of the sensory and autonomic nervous system, a point mutation in the 5' splice site of intron 20 combined with a normally weak 3' splice site causes the skipping on exon 20 of the IKBKAP gene, leading to abnormal nervous system development [23, 24], and in β-thalassemias some of the possible mutations to β-globin causing the disease can be found in either the invariant dinucleotides at the splice sites, in the surrounding sequences or in cryptic sites both in introns and exons [31]. Hutchinson-Gilford progeria is a rare genetic disorder that confers to the people affected by it premature acquisition of some of the characteristics of aging caused in most cases by a point mutation in what is probably an ESE on the exon 11 of the LMNA, the gene coding for lamin A, causing a cryptic splice site activation that leads to a transcript lacking 150bp of exon 11 and a truncated protein product called progerin [32]. Intron retention has also recently been shown as a potential pathogenic avenue. For example, in the Autoimmune Polyendocrine Syndrome type 1, a rare inherited disorder characterized by multiple autoimmunities, the retention of intron 3 of the autoimmune regulator gene, involved in the negative selection of self-immune T cells, leads to the introduction of a premature stop, truncating the protein product [33]. In a class of neurodegenerative disorders with accumulation of MAPT called Tauopathies, the spatial and temporal regulation of the 6 isoforms of the MAPT gene can be altered as a result of mutations on the exon 10 and the change in fraction of isoforms is causative in for frontotemporal dementia [34] and evidence shows that relative frequency changes between tau isoforms could be related to other neurodegenerative processes [35-37]. Isoform balance disruption can also be found with the WT1 gene in Frasier syndrome [38]. We've already seen that spliceosome alterations are found in some cases of myeloid malignancies, but in general aberrant splicing is a frequent element in tumorigenic processes. In the previous sub-chapter we've seen how the gene KLF6 has isoforms with opposite effects on cell growth, with the KLF6-SV1 variant losing the characteristic zing finger DNA binding domain due to an alternative 5' splice site in exon 2 and acting as a dominant-negative against the inhibition of cell proliferation and migration provided by the tumor suppressing KLF6-SV2 isoform [21, 39]. This kind of dynamic plays out in prostate cancer, where a point mutation that leads to the creation of a splicing factor binding site enhancing KLF6-SV1 expression represents an increase in cancer risk [40] and increases cancer progression, while its knockdown reduces the growth of the tumor [41]. BRCA1 is causative in most hereditary breast and ovarian cancers [42] and mutations in its cis and trans-acting splicing actors are considered to be involved in cancer risk [23], in particular a nonsense mutation in exon 18 that affects an ESE and leads to skipping of the exon and a single mutation that can activate a cryptic 3' splice site that results in a truncated protein product were found in families affected by high risk of these cancer types [43, 44]. Splicing factors are common dysregulation targets in cancers [45], with many SR proteins going through upregulation [46, 47] and levels of SR and hnRNP splicing factors in colon carcinomas have been correlated to levels of alternative splicing events of CD44 which are associated to tumor metastasis, among other processes [48]. Intron retention is also abundant in the transcriptome of many liquid and solid cancers compared to normal tissues, even without mutations affecting the core splicing machinery [49]. Strategies to target aberrant splicing events involve the usage of antisense oligonucleotides and of small molecules. Antisense oligonucleotides are designed to bind to specific cis-acting regulatory elements on RNAs to modulate their splicing or to bind nascent RNAs to promote their degradation. Duchenne muscular dystrophy is a progressive disease of muscles caused by mutations on the gene DMD coding for dystrophin, which lead to myofiber necrosis. The mutations are often frame shifting deletions, but antisense oligonucleotides that induce skipping of exon 51, 44, 45 or 53 have been approved in the USA as ways to restore partial function, although without the ability to provide full recovery [50, 51]. The small molecules approach involves the targeting of splicing factors to modulate them or of RNA sequences to block aberrant splicing factor recruitment. Spinal Muscular Atrophy is caused by the loss of the SMN1 gene encoding for the SMN protein, which can't be properly supplied by the almost identical gene SMN2 which differs by a single nucleotide in exon 7 leading to predominant exclusion of this exon from SMN2-derived transcripts, producing a truncated and unstable form of SMN [52]. Small molecule compounds able to promote exon 7 inclusion in SMN2-derived transcripts have been considered for the treatment of this disease [52, 53], with one achieving in the USA and 30 other countries.

1.2. MET

1.2.1. MET

The MET proto-oncogene was first isolated as an oncogene in MNNG-HOS [54-56], a cell line derived from chemically transforming the human osteosarcoma cell line HOS (CRL-1543) with a carcinogenic nitrosamine (MNNG). In this cell line a rearrangement fuses the MET locus with the Translocated Promoter Region (TPR) locus, leading to the production of the fusion protein TPR-MET which has kinase capacity and is constitutively phosphorylated. The MET proto-oncogene is located on chromosome 7q31 and spans 126,182 bases, with 21 exons separated by 20 introns. The protein product is a Tyrosine kinase receptor [57] with a variety of names: Met, c-Met, Hepatocyte Growth Factor Receptor (HGFR) and Scatter Factor Receptor (SFR). The protein is produced as a single chain precursor that is proteolytically cleaved into an α chain and a β chain, which then get linked with a disulfide bond [58-60].

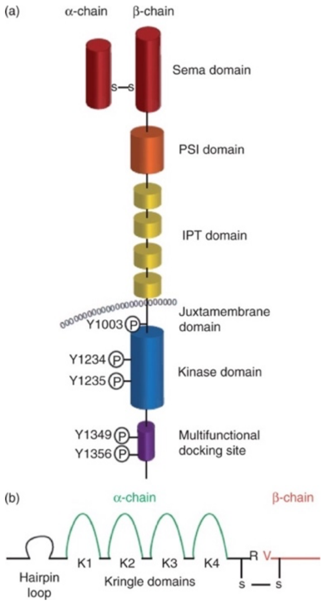


***Figure 4**. Structure of Met and its ligand. (**A**) Structure of the Met Tyrosine kinase receptor protein. (**B**) Structure of the HGF, ligand of the Met protein with the site of cleaving represented by the Arginine (R) and the valine (V) between the two chains. Figure from [58].*

The resulting protein (**Figure 4A**) is composed of an extracellular portion containing the Sema domain, which includes the binding site for the ligand and encompasses the entirety of the α chain and part of the β chain and shares sequence homology with semaphorins, followed by a cysteine rich domain found also in plexins, semaphorins and integrins (thus called PSI) and 4 domains related to immunoglobulin-like domains that are also found in immunoglobulins, plexins and transcription factors (thus called IPT) connecting to the transmembrane helix. The intracellular portion includes a juxtamembrane domain that contains a Tyrosine in position 1003 that if phosphorylated can bind to ubiquitin ligase Cbl leading to ubiquitination of Met, the catalytically active Tyrosine kinase domain with Tyrosines 1234 and 1235 that are phosphorylated during receptor activation and finally a docking site that contains Tyrosines 1349 and 1356, crucial to recruit several downstream adaptors carrying Src homology-2 (SH2) domains [58-60]. The entirety of the first exon is not translated, exon 2 contains the cleaving site for the α chain and the β chain, between exon 2 and 12 is the extracellular domain, exon 13 encompasses the transmembrane domain and the intracellular domain is encoded by the exons from 14 to the last exon 21, with exon 14 encoding for the juxtamembrane domain, exons 15 to 20 covering the catalytic domain and exon 21 the docking site [60]. The ligand of Met was first discovered by two different groups as two different factors, the hepatocyte growth factor (HGF) and the scatter factor (SF), which were later determined to be the same molecule [61-63]. HGF is secreted as a precursor (pro-HGF) that undergoes cleaving by extracellular proteases into an α chain and a β chain, which are linked together through a disulfide bond. The α chain presents a hairpin loop (**Figure 4B**) and 4 doubled looped structures composed of internal disulfide bridges called kringle domains and the β chain carries a domain with homology to Serine proteases but without enzymatic activity [58, 59]. HGF binds to the Sema domain of Met resulting in the homodimerization of the receptor and phosphorylation of Tyrosines 1234 and 1235 in the kinase domain, then Tyrosines 1349 and 1356 in the docking site are phosphorylated and form an SH2 recognition motif that can recruit a number of signaling effectors [58-60, 64, 65], either directly or through the scaffolding protein GAB1. These interactions result into the activation of multiple signal transduction pathways (**Figure 5**), among the major ones the MAPK cascade, the PI3K-Akt pathway and the STAT pathway. The MAPK cascade (**Figure 5A**) is composed of 3 pathways that sequentially activate [66] following two possible mechanisms. In one the docking site of MET interacts with a complex of GRB2 and SOS either directly or through an adaptor protein (SHC), GRB2-SOS then activates Ras [67, 68], in the other the recruitment of p120 is stopped through the dephosphorylation of the p120 binding site on GAB1 by the action of SHP2 [69]. In both methods, Ras activation has multiple results, one is achieved through the Ras-Raf-MEK-ERK pathway which leads to phosphorylation of ERK1 and ERK2 which then translocate to the nucleus and promote the progression of the cell cycle, proliferation and cell motility [70, 71].

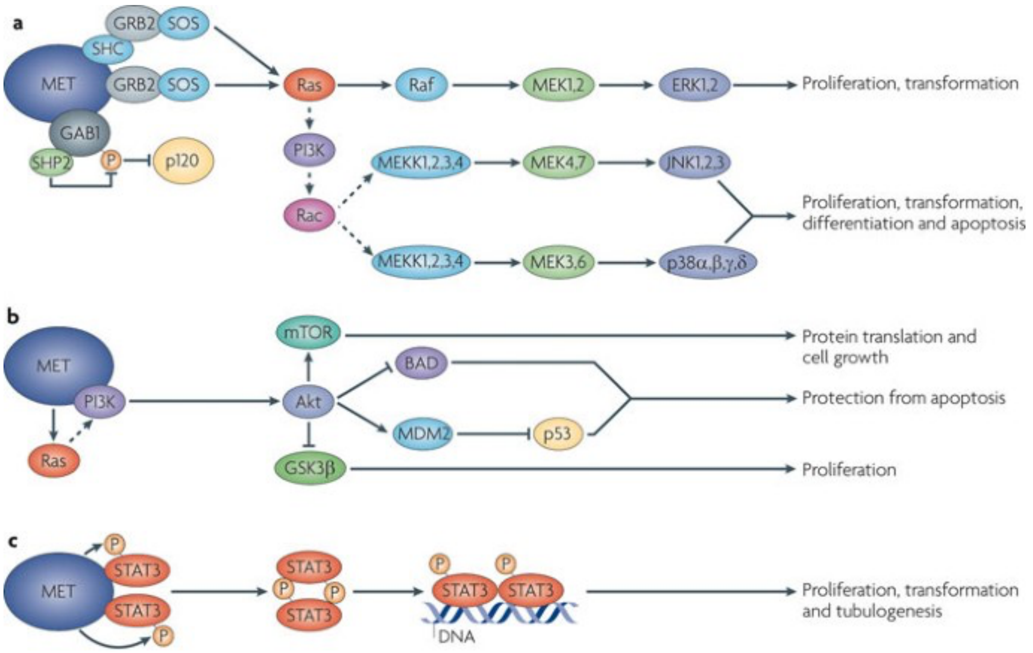


***Figure 5**. Some of the major signalling pathways tied to the Met Tyrosine kinase receptor. (**A**) MAPK cascade. (**B**) PI3K-Akt pathway. (**C**) STAT pathway. Figure from [65].*

Another result of the Ras activation is the activation of Rac through a Ras-PI3K-mediated pathway, causing in turn the activation of JNK1, JNK2, JNK3 by MEK4 and MEK7 and of p38α, p38β, p38γ and p38δ by MEK3 and MEK 6, leading to the control of different cellular processes, among which proliferation, apoptosis, cell motility and cytoskeletal rearrangement [72-75]. The PI3K-Atk pathway (**Figure 5B**) can be activated either by direct interaction of PI3K with the docking domain of MET, by indirect interaction with MET via GAB1 or as part of the MAPK cascade and it leads to the activation of Akt which promotes proliferation by blocking GSK3β, cell growth and protein synthesis by stimulating the activity of mTOR and cell survival by inhibiting the activity of both P53 and BAD [76]. The STAT pathway sees the homodimerization of STAT3 after MET-dependent phosphorylation, allowing it to operate as a transcription factor controlling proliferation and tubulogenesis [65, 77]. The activation of these pathways by MET elucidates some of the mechanisms through which epithelial cells develop an invasive growth phenotype following MET stimulation [65], leveraging the increased motility and the improved cell survival and proliferation to expand into new environments, also contributing to the stimulation of angiogenesis [78]. This invasive phenotype is crucial during development [79] but also in mature individuals in cases of tissue damage in a variety of organs, where MET signalling plays crucial roles [80-83].

1.2.2. MET signalling alterations in cancer

Dysregulation of the HGF-MET pathway is a common element found in tumor cells [59, 65, 84], with many of the resulting cell responses favoring tumor growth, survival and metastatic spread. The mechanisms through which tumors achieve high MET activation can generally be divided into three categories: a) MET overexpression without amplification, b) ligand dependent and c) MET gene alternation dependent. The first case is the most common mechanism of aberrant MET activation and involves an increase in MET signalling as a result of transcriptional overexpression of the protein [85-90]. This happens frequently in tumors due to the ability of the hypoxic microenvironment to induce MET expression [91], but it can also be due to the activity of other oncogenes like Ras and Ret [92] and it results in higher sensitivity to HGF, which is ubiquitously expressed, and to poor prognosis for the affected patients [59, 93]. The ligand dependent mechanisms are the second most common route for aberrant MET activation and involve activation through increased exposure to the ligand HGF, which is physiologically secreted by mesenchymal cells while MET is expressed by epithelial cells. This increased exposure can be due to the establishment of an autocrine loop, which is observed in some mesenchymal cell derived tumors that acquire the ability to produce MET [89, 94, 95] and in ectodermal derived tumors which acquire the ability to produce HGF [96, 97]. However, more frequently the stimulation is paracrine and it’s a result of increase secretion of HGF by tumor stromal cells [98, 99]. The MET gene alteration cases involve multiple different mechanisms due to the different alterations that can affect the gene. MET amplification is a rarer route to achieve MET overexpression compared to the transcriptional route described before and it has been detected in various cancers [87, 100-104]. The gene rearrangement event leading to the formation of the TPR-MET fusion protein that was involved in the first isolation of the MET gene was for a long time the only known human tumors gene rearrangement, being found in gastric carcinomas [105, 106], however more recently more instances have been discovered. Some very rare fusion events were found involving a dimerization motif fusing to the N-terminal of the MET gene in a few cases of some carcinomas [107] while pediatric glioblastoma presents a high frequency of occurrence of MET rearrangement events [108]. Both somatic or inherited mutations can be found is inherited and sporadic renal papillary carcinomas [109-111], impacting the Tyrosine kinase domain and leading to autophosphorylation. As we've seen previously, Tyrosine 1003 in the juxtamembrane domain is involved in MET degradation through ubiquitination, so mutations in this position lead to tumorigenesis [112, 113]. Somatic mutations were also found in glioma, mucinous ovarian carcinoma, head and neck squamous cell carcinoma and childhood hepatocellular carcinoma [114-117] and collections of such mutations are listed in somatic mutation databases like COSMIC [118]. In head and neck squamous cell carcinomas, somatic MET mutations leading to constitutive MET activation have been found in metastases, pointing to a selection of these mutations during progression [114]. The most frequent of these MET activation mechanisms are secondary, usually related to environmental factors leading to higher MET activity due to high availability of HGF or overexpression due to hypoxia or the activity of other oncogenes, making the MET involvement a late tumor development event that enhances tumor survival and confers to it the invasive growth phenotype facilitating tumor dissemination. The interaction between tumor microenvironment hypoxia and MET overexpression and function is a possible explanation for some preclinical models showing promotion of invasion and metastatic spread of tumors subjected to anti-angiogenic therapies, which can be blocked with the addition of MET inhibitors to the treatment [119-121]. MET is also involved in the radiation resistance displayed by glioblastoma stem cells, which can be avoided through MET inhibition [122] and in the resistance to treatment of breast cancer with PARP inhibitors, which can also be overcome through combination therapy with MET inhibitors [123]. While useful in combinatorial approaches like the ones just presented, pre-clinical studies suggest that MET blockade is in itself ineffective in the limitation of tumor growth in tumors characterized by MET activity without MET genetic alterations [124], although it can still affect metastatic spread [125]. On the other hand, there's indications that some tumors with MET genetic lesions, especially amplifications, could be more affected by MET inhibitors, showing MET oncogene addiction [126, 127] and offering a possible therapeutic target, with responsiveness being achieved in some case studies [128, 129].

1.2.2.1. MET exon 14 skipping

A relevant MET alteration that we've not discussed until now is the skipping of exon 14 of the MET gene. As detailed previously, exon 14 encodes for the juxtamembrane domain of the receptor, which contains residues crucial for the regulation of MET activity, like Tyrosine 1003 which promotes the degradation of the protein via ubiquitination. Deletion of this domain leads to accumulation of MET receptors [130]. More than 500 genetic lesions resulting in MET exon 14 skipping have been found, including point mutations, insertions and deletions in regions pivotal for splicing, most commonly point mutations at donor sites [131, 132]. MET exon 14 skipping is found in around 3% of non-small cell lung cancers, which is where it was also detected for the first time in human tissues [133], with particularly high frequency in the adenosquamous cell carcinoma and pleomorphic carcinoma subtypes [134]. Compared to other driver mutations, patients carrying exon 14 skipping tend to be older and more of them had a history of smoking [135, 136]. MET exon 14 skipping has also been found in other tumors, such as neuroblastomas and gastric cancer [137-139]. In 2020 capmatinib, the first MET-targeted therapy drug, was approved, targeting non-small cell lung cancers presenting MET exon 14 skipping and in 2021 a second one, tepotinib, was approved as well. Both are Tyrosine kinase inhibitors that interact with the Tyrosine 1230 residue in the kinase domain of activated MET [140-143] but resistance is found in around 33% to 50% of patients initially and the rise of acquired resistance is a near certainty [144-146], as it's the case with other Tyrosine kinase inhibitors therapies. Therapies involving MET antibodies or immune checkpoint inhibitors, possibly in combination therapy with the currently available MET Tyrosine kinase inhibitors, offer possible future perspectives for the treatment of non-small cell lung cancers carrying MET exon 14 skipping alterations [134, 147].

1.3. Neural Networks

1.3.1. Feed forward artificial neural networks

Feed forward artificial neural networks (ANN) are composed of simple connected processors called neurons that form the nodes of an acyclic graph [148]. These neurons are organized in sequential layers formed each of a variable number of neurons, and the simplest layout, a shallow neural network, involves 3 layers, a first input layer, a hidden layer in the middle and a final output layer. Neurons in one layer are connected to neurons in the following layer by a weighted edge. The unidirectionality of this connection is what qualifies these types of neural networks as feed forward as opposed to recurrent neural networks, which won't be part of this work.

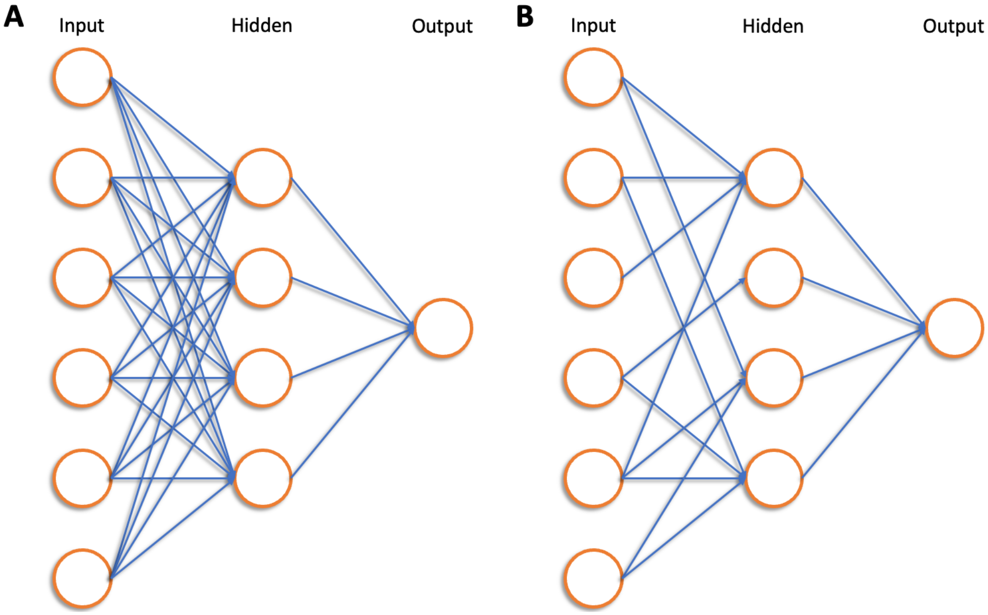


Figure 6*. Simple Feed Forward Artificial Neural Networks. In red the nodes and in blue the edges. (****A****) Fully connected or dense ANN. (****B****) Sparsely connected ANN.*

The connections between two sequential layers can be from each neuron of the first to all the neurons of the second, in which case we are dealing with a fully connected or dense layer (**Figure 6A**), or only connect each neuron on the starting layer to some of the ones in the second layer, giving rise to a sparsely connected layer (**Figure 6B**). Neurons in the input layer receive the raw data and don't perform any operations on it, their only job is to feed it forward to the first hidden layer. Each neuron in the hidden layer calculates the values it receives from the previous layer as follows:

$$z = \sum_{i=1}^{n} x_i w_i + b \qquad \textbf{\textit{Eq. 1}}$$

where n is the number of connections of the current neurons to neurons of the previous layer, x is the value passed by each of the previous layer neurons to the current neuron, w is the weight of the edge connecting each previous layer neuron to the current neuron and b is the bias of the current neuron. An activation function f is applied to the resulting value z to obtain a resulting neuron output y that represents the value that the current neuron will pass to the later layers (**Figure 7**). If y is equal to 0, the neuron is not activated and no value is passed.

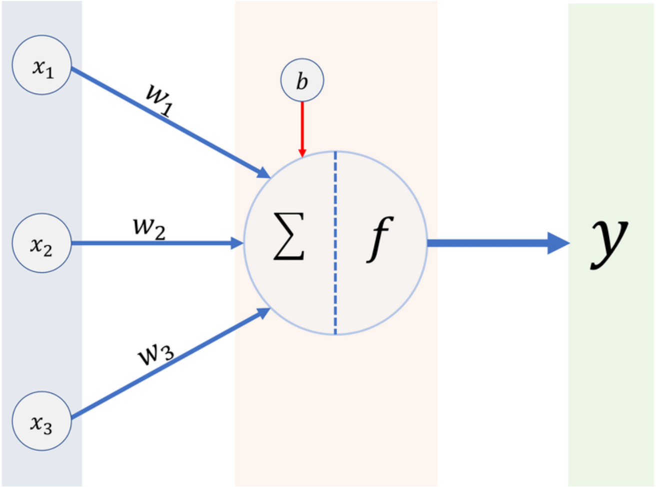


Figure 7*. Computations performed by each neuron in layers other than the input. Inputs x are multiplied by the weights w of the edges connecting them to the neuron, the bias b is added and the result is used in the activating function f to obtain the final output y. Figure from [149].*

Activation functions come in a variety of forms, depending on the purpose they cover. The two simplest activations functions are a linear or identity function and binary step function which are shown in **Eq. 3** and **Eq. 2** respectively.

$$y = f(z) = z \qquad \textbf{\textit{Eq. 2}}$$

$$y = f(z) = \begin{cases} 0, & z < 0 \\ 1, & z \geq 0 \end{cases} \qquad \textbf{\textit{Eq. 3}}$$

The linear function simply uses the value of *z* passes it directly to the next layer, this is equivalent to a linear regression. The binary step function is used in perceptrons, single layer neural network that execute binary classifications that are sufficient for linearly separable sets of data, meaning that the elements can be completely separated by an *n-1* dimensional hyperplane, where *n* is the dimensionality of the dataset. These activation functions are not appropriate for tackling more complex problems, however, which is why non-linear activation functions are widely used. The two most common non-linear activation functions are the logistic or sigmoid or softstep activation function and the Rectified Linear Unit (ReLU) activation function (**Eq. 4** and **Eq. 5**).

$$y = f(z) = \frac{1}{1 + e^{-z}} \qquad \textbf{\textit{Eq. 4}}$$

$$y = f(z) = max(0, z) \qquad \textbf{\textit{Eq. 5}}$$

The sigmoid function takes the input value *z* and transforms it into a value in range *(0, 1)*, so no neuron will ever be deactivated and is generally used when predicting probabilities. The ReLU function output *y* is equal to the value of *z* if it's above 0, otherwise it defaults to 0, making the neuron inactive, contributing to its computational efficiency. In deep neural networks with multiple hidden layers, the output *y* of each active neuron in the first hidden layer would be then passed to the second hidden layer in the same way the input layer passed the raw values, and the second hidden layer would execute the same process we've just described. In the simple 3 layers structure we saw in **Figure 6** all the hidden layer active neurons pass their output *y* to the output layer neuron, which also computes the same steps as a hidden layer neuron and then returns its output *y* to the user. Usually in deep neural networks the activation function used in hidden layer neurons is the ReLU function while the activation function for the output neurons depends on the purpose of the network. For a binary classifier between two classes, for example *cat* and *dog,* we can utilize a single output neuron in which *cat* corresponds to the neuron not activating while *dog* corresponds to the neuron activating, and with the sigmoid activation function we will be provided with a value in the range of *(0, 1)* representing the probability of the neuron activating for a given element, meaning the probability of that element being part of the *dog* class. Multiple class classifiers would need the softmax activation function, which is an extension of the sigmoid function. Weights and biases are the learnable parameters of the neural network, the weights can be viewed as an evaluation of the importance of a given starting neuron input to the output of a neuron, while bias works as a threshold modulator for the neuron by acting as a flat contribution (or detraction) to the weighted sum of all inputs to the neuron (**Figure 8**).

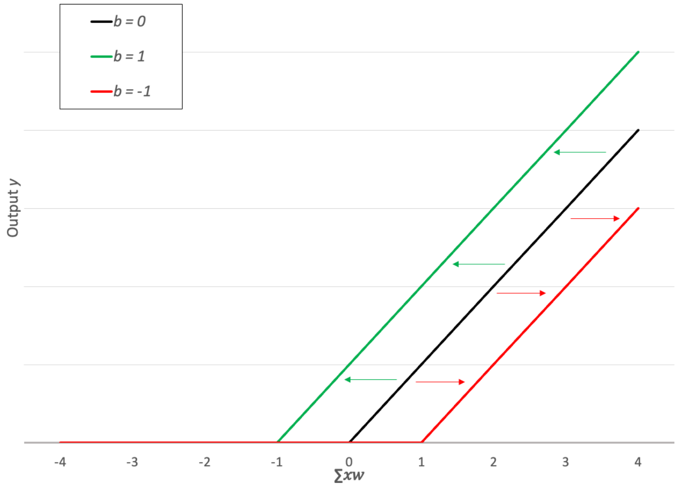


Figure 8*. Effect of the value of the bias b on the required result of the weighted sum of all inputs* $\sum x_i w_i$ *to lead to activation with a ReLU activation function.*

The initial values of weights and biases are assigned usually through random sampling from small number distributions. These starting weights and biases then get altered during the training of the neural network on a labeled training dataset through a process called backpropagation. Backpropagation in the case of traditional feed forward artificial neural networks is a supervised learning method that evaluates the output of the entire network (the output of the output layer neurons) run on the training dataset samples and compares it to the expected results from the labels provided with the same dataset. This is done either after the entire dataset has gone through the network or after a number of samples equal to the batch number which is set on creation of the network. The comparison is executed through a loss or cost function that calculates how well our network performed the intended task. An example of a widely used loss function is the Mean Square Error (**Eq. 6**):

$$J(w) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^r - y_i^p\right)^2 \qquad \textbf{\textit{Eq. 6}}$$

where *J(w)* is the cost function associated with the set of weights and biases *w*, *n* is the number of elements in the dataset (or the batch), $\mathrm{y_i^r}$ represents the real values of each of these elements and $\mathrm{y_i^p}$ represents the predicted values. Based on the resulting cost, the network updates the value of weights and biases based on the gradient descent method to minimize the cost function. The gradient descent method calculates the change to apply to the weights and biases as follows:

$$w_u = w_s - \eta\nabla J(w_s) \qquad \textbf{\textit{Eq. 7}}$$

where $\mathrm{w_u}$ is the updated set of weights, $\mathrm{w_s}$ is the starting set of weights and biases, η is the learning rate and $\nabla\mathrm{J(w_s)}$ is the gradient of the loss function when the set of weights is $\mathrm{w_s}$. The learning rate is a value that is selected on creation and defines how big of a step we want the change of weights to take at each update and the gradient of the loss function is a vector that points in the direction of steepest ascent, so to minimize the loss function we head in the direction opposite of the gradient. After all the samples have gone through the network, the process starts again, and each of these iterations running through the entire dataset is called an epoch, with the number of epochs being also a parameter set on network setup. Getting too close to the real values has its own drawbacks, because in training a neural network our objective is to extract some broader features that allow us to generalize our model and run it on data not included in our training dataset. If we run the network on a training dataset for too many epochs we risk incurring in overfitting, achieving great accuracy for our training dataset at the cost of accuracy when we use the same model on new data, due to our model learning the statistical noise present in our training set. Dropout is a strategy to reduce the risk of overfitting by dropping out random nodes in the input and hidden neurons layers. Another strategy is early stopping, in which a number of samples are kept as a validation set and the network isn't trained on them, however the performance on the network is routinely evaluated for the validation set as well, and if the performance of the model stops improving on the validation set, training is stopped even if the set number of epochs hasn't been reached and the training dataset loss is still decreasing.

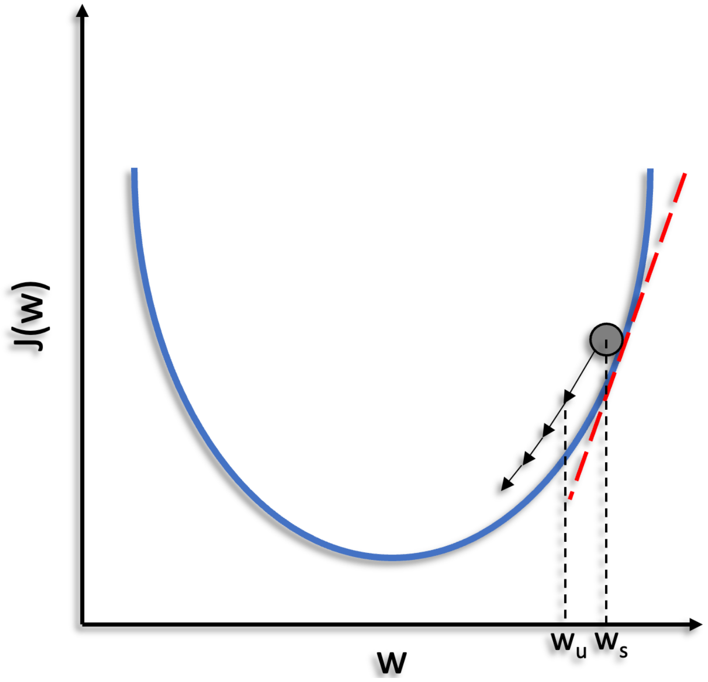


***Figure 9**. Scheme illustrating gradient descent. Blue line: cost function J(w); red dotted line: gradient $\nabla J(w_s)$; black arrows: subsequent updates of the set of weights; w_s: starting weights set. w_u: first updated weights set.*

1.3.1.1. Convolutional neural networks

Convolutional neural networks (CNNs) are a type of feed forward artificial neural network that employ a convolution operation (**Eq. 9**) in at least one of their layers [150]. Convolutional networks are often used on inputs organized in at least 2-dimensional grids, like black and white images (colored images have 3 channels, which act as a third dimension). The convolutional layer is a hidden layer in which a neuron only sees a limited local receptive field, a small localized region of the input (**Figure 10A**). Visually, in the example of a square 2-dimensional input like an image this would be represented by a small $l \times l$ area starting on the top left of the larger $n \times n$ whole image for the first neuron of the convolutional layer. This small $l \times l$ then slides across the $n \times n$ image by shifting right in the image by a number of input neurons equal to the stride parameter s defined during network design, capturing a new $l \times l$ region associated to the second convolutional layer neuron and keeps shifting for subsequent convolutional layer neurons until the right border of the window reaches the right edge of the map, at which point it shifts downwards by *s* input neurons and starts moving leftwards until the edge of the image and continuing in this manner. This produces a convolutional layer that can also be organized as an $m \times m$ grid with

$$m = \frac{n - l + 2p}{s} + 1 \qquad \textbf{\textit{Eq. 8}}$$

where *p* is the amount of optional padding rows or columns that can be added to the input image grid to resolve some issues, like the decreased coverage of corner areas due to being touched by only one local receptive field compared to central areas that are involved in multiple local receptive fields. The connections between the local receptive field neurons and the convolutional layer neuron they are related to are learnable weights that can be arranged in a grid of the same size of the local receptive field ($l \times l$), called kernel or filter. For each convolutional layer neuron, the output *y* is calculated through convolution which is executed as follows:

$$y = f\left(\sum_{i=1}^{l}\sum_{j=1}^{l} w_{i,j}x_{i,j} + b\right) \qquad \textbf{\textit{Eq. 9}}$$

where $\mathrm{x}_{\mathrm{i,j}}$ is the value associated to the input neuron in position *i,j* in the grid of the local receptive field associated to the convolutional layer neuron, $\mathrm{w}_{\mathrm{i,j}}$ is the weight in position *i,j* in the grid of the kernel, *b* is the bias of the convolutional layer neuron and *f* is the activation function. Both the kernel with its weights and the bias are shared for all convolutional layer neurons. The result is that the same feature that is encoded in the kernel weights is searched across the entirety of input image, and when detected the associated convolutional layer neuron will activate. Due to this mechanism, the resulting grid of convolutional layer neurons is called a feature map. However, being able to detect a single feature is not enough for most applications, so convolutional layers in CNNs usually deploy multiple feature maps, each characterized by its own kernel weights and biases shared among all neurons of the feature map and keeping the same kernel size and stride, resulting in the same size of grid neurons. The complete output of the convolutional layer ends up being a 3D grid of neurons and dimensions $m \times m \times Fn$ with m as in **Eq. 8** and Fn being the number of feature maps.

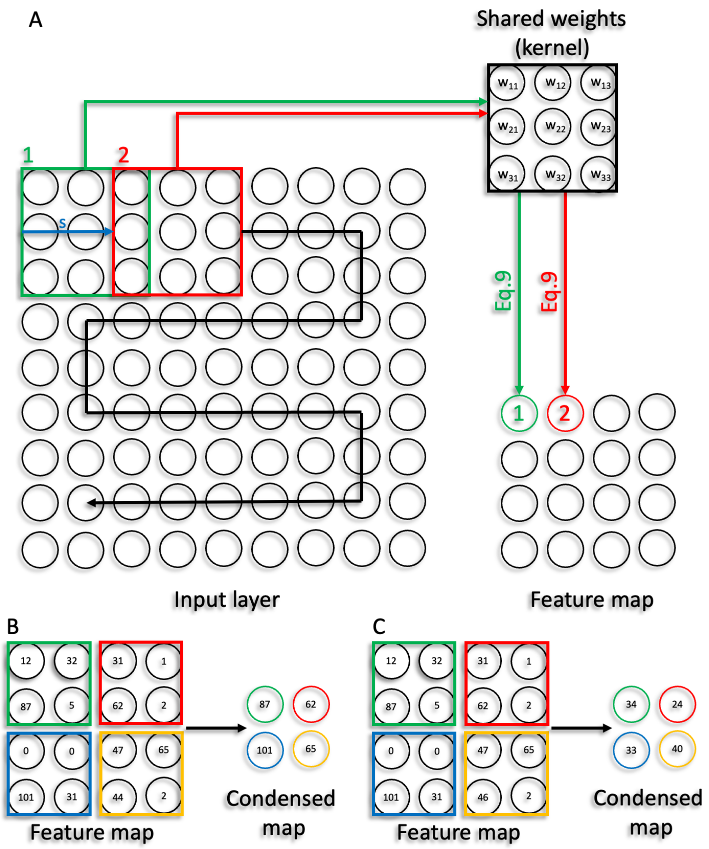


***Figure 10**. Representation of the computations in a CNN. (**A**) Convolutional layer operations on a 9 × 9 input image using a 3 × 3 kernel and a stride s equal to 2 with no padding. Green: local receptive field 1 operations; red: local receptive field 2 operations; blue arrow: stride s; black arrow: path of shifting of the local receptive field with a stride s equal to 2. (**B**) Max pooling on a 4 × 4 feature map with a non-overlapping 2 × 2 filter. Each colored box in the feature map represents the input neurons for the condensed map neurons of the same color. (**C**) Average pooling on a 4 × 4 feature map with a non-overlapping 2 × 2 filter. Each colored box in the feature map represents the input neurons for the condensed map neurons of the same color.*

Often convolutional layers are followed by pooling to simplify the outputs of the neurons in feature maps by computing a condensed version of the feature map. This pooling step also involves a sliding window, usually of size 2 × 2 with a stride of 2 to avoid overlap, although overlapped pooling strategies exist, for example with sliding windows of size 3 × 2 and stride

2. Each step of the sliding window is condensed into a single condensed map neuron through different types of pooling algorithms, for example max pooling which selects the highest value among the ones in the window and average pooling which averages all the values of the window (**Figure 10B-C**). This operation is conducted separately for each feature map produced by the convolutional layer and in the case of a 2 × 2 non-overlapping pooling strategy it reduces the dimensions of the data from $m \times m \times Fn$ to $m/2 \times m/2 \times Fn$. More convolutional layers followed by pooling can be added to the network and the output of the last pooling is then flattened by concatenating all of the outputs of the neurons in all features as a 1-dimensional array which is then used as input for a fully connected neural network. The convolution and pooling steps act as automated extraction of features to provide to the dense neural network that fulfills the task of classification based on the extracted features. Like traditional artificial neural networks, CNNs utilize backpropagation in a supervised manner to update the shared weights and biases associated to each feature map in the convolutional layers. CNNs are widely used on image data in object recognition, image classification and object detection tasks like face recognition and have also been applied in speech recognition [151].

1.3.1.2. Shallow sparsely connected autoencoders

Autoencoders are a type of feed forward artificial neural network that is composed of two parts, an encoder which maps the input to a latent space and a decoder that produces a reconstruction based on the mapping in the latent space [148, 152]. The autoencoder is trained to replicate the input in the output, seeking to optimize through backpropagation a loss function that evaluates the difference between the two. Because the target that the output is compared to is the input data itself and there’s no need for labeling, autoencoders are an example of unsupervised backpropagated learning. Autoencoders are designed to not be able to reproduce the data perfectly, but they are intended to be restricted in ways that force them to learn the most important features of the data instead. The reconstructed output is only used for the training of model, the target is an encoding function with the ability to extract those important features. In undercomplete autoencoders, the restriction imposed is that the latent space, also called bottleneck, is of smaller dimensionality than the input, forcing the model to identify the features that must be conserved for an accurate reconstruction of the input, leading to linear or non-linear dimensionality reduction or feature extraction applications. Another approach to limiting the ability of the model to reach the identity function is found in regularized autoencoders, which change the loss function to push the model to have additional properties. These regularizations can be applied to undercomplete autoencoders or allow autoencoders with latent spaces of the same number of dimensions or even more dimensions than the input (overcomplete autoencoders) to learn something other than the identity function. Sparse autoencoders introduce in the loss function a penalty dependent on the number of activations in the hidden layers, trying to build a model that relies only on a small quantity of active neurons, also creating a bottleneck like in the case of undercomplete autoencoders. Denoising autoencoders are trained on a version of the input that has gone through an addition of noise [153]. The loss function is modified because instead of the loss function comparing the output of the autoencoder to the input of the autoencoder, meaning the noisy dataset, we compare the output to the data before the noise addition. This allows the autoencoder to learn how to remove the noise, maintaining the elements relevant to the original data. Contractive autoencoders, like sparse autoencoders, apply a penalty to the loss function, however in the case of contractive autoencoder what is penalized is the derivative of the activations in the hidden layers. The effect is that small changes of the input will result in the same encoded result [154]. Variational autoencoders [155] are a class of generative autoencoders in which the encoder returns a latent space that is a distribution instead of a vector of values, making it continuous. Autoencoders have been employed in the field of RNAseq analysis for tasks such as dimensionality reduction before clustering single cell data [156] and batch correction [157]. Usually these are deep autoencoders, however shallow sparsely-connected autoencoders (SSCA) have been used recently for the modeling of gene set analysis [158] and to derive functional features from cell clusters in single cell RNAseq datasets [159]. These SSCAs are built with a single sparsely-connected hidden layer in which the connections to the input are hardcoded and represent pre-existing biological knowledge, compared to sparse autoencoders in which the sparsity is a regularization method enforced through a loss function penalization.

1.3.2. Machine learning in MET exon 14 skipping detection

The susceptibility of non-small cell lung cancer patients carrying MET exon 14 skipping mutations to targeted therapy opens up a need for accurate and fast detection of these events. Detection of exon skipping events can be performed in a single-gene manner through techniques like RT-PCR starting from RNA or Sanger sequencing using DNA [160]. However, these present serious issues in the investigation of diseases like non-small cell lung cancer, where we want to screen for multiple biomarkers at the same time and in which the amount of material retrieved from biopsies is usually too small to perform multiple single-gene tests [161]. As a tool to solve these issues, next-generation sequencing technologies can allow for the screening of multiple regions of interest at the same time on the same sample, with a model built on metastatic non-small cell lung cancer patients showing a decrease in both the time required to perform the tests and the cost associated to testing [162]. Sequencing strategies in the field of clinical oncology mostly focus on clinically relevant regions using library prep strategies like amplicons [163], however DNA-based amplicon-mediated targeted sequencing techniques have shown to perform inadequately in the detection of MET exon 14 skipping events [164], suffering from high false negatives due to the high variety and diversity in size of the alterations leading to the exon skipping event. RNA sequencing can take into consideration a more consistent marker, looking only at the exclusion of exon 14, regardless of the which genetic alteration is causing it. As a result, targeted RNA-based technologies have shown to perform better than DNA-based ones [165], although in the clinical setting RNA poses some technical issues, as it degrades much more easily than DNA. The efficient detection of these events is not just a matter of finding the right wet lab methods, but also devising the right algorithms to parse the data and identify cases in which the events are taking place. In particular deep learning and machine learning could provide a big contribution for fast screening of large cohorts. At the time of writing of the work that is the subject of this Chapter (https://pubmed.ncbi.nlm.nih.gov/, accessed on 01 March 2021), our inspection of the available literature regarding MET exon 14 skipping didn’t return any articles using deep learning or machine learning for MET exon 14 skipping detection. We found some generalist exon skipping event prediction algorithms: Zhang et al. used CNNs to classify splice junctions derived from primary RNA-seq data [166]; Du et al.

integrated RNAseq data and genome sequence information in a Rotation Forest algorithm [167]; Jaganathan et al. developed the tool SpliceAI which uses CNNs to predict splicing from pre-mRNA sequences [168]; Zuallaert et al. developed SpliceRover based on CNNs to predict splice sites [169]. Owing to the higher potential for accuracy for a tailor-made tool in calling for the exon skipping events, we decided to investigate the potential of different neural network architectures in providing sensitive and rapid MET exon 14 skipping detection starting from RNAseq data, as it offers the best results. In our investigation we compared 3 different architectures: traditional feed forward neural network (NN), convolutional neural network (CNN) and shallow sparsely-connected autoencoder (SSCA).

2. Results

2.1. Neural Network for the Detection of MET Exon 14 Skipping (METΔ14).

To detect MET exon 14 skipping events, an NN made of six layers was built [170]; **Figure 11A**.

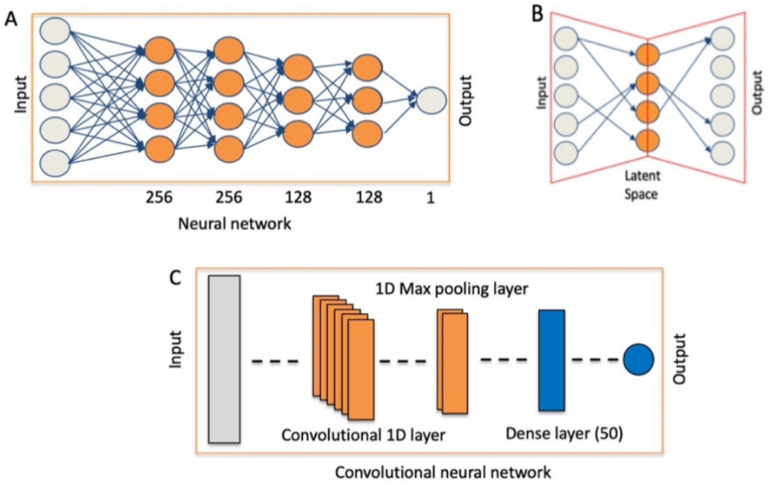


Figure 11. *Neural networks used for the analysis of MET variants.* ***(A)*** *NN for the detection of exon 14 skipping events.* ***(B)*** *Shallow sparsely-connected autoencoders for the detection of MET transcription variants.* ***(C)*** *Convolutional neural network (CNN).*

As a training set for the NN, we used data from amplified WT MET and exon 14 skipping MET (**Table 1**).

Cell Line	Tumor type	Status	RNAseq (Million Reads)	MET (Thousand Reads)
EBC-1	Lung cancer	Amplified MET	113	1447
Hs746T	Gastric cancer	Amplified METΔ14	95	846
A549	Lung cancer	MET	115	109
NCI-H596	Lung cancer	METΔ14	118	114

Table 1. *MET cell line RNAseq data.*

Specifically, we split the MET reads in random non-overlapping subgroups of 1000 reads. Although at 1000 reads, coverage of the detection of METΔ14 becomes a bit blurry—**Figure 12D;** this threshold allows for a generation of large number of MET (1447) and METΔ14 (846) to not overlapping subsamples, thus providing a high numerosity of training data, which is an important element for an efficient training of the NN.

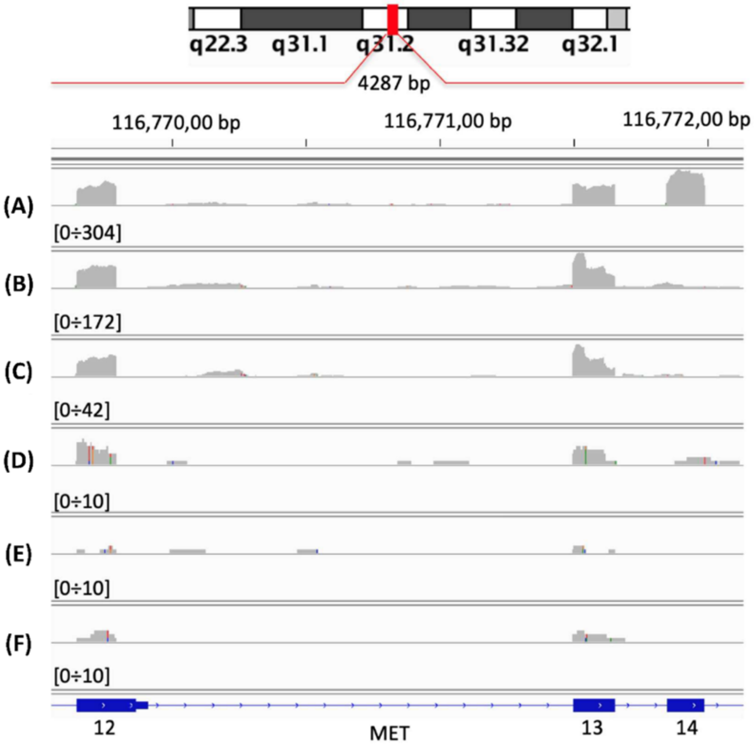


Figure 12. *Expected coverage for exons 13, 14 and 15.* ***(A)*** *WT MET from A549 RNAseq sample (33 million reads), 27,152 reads mapping on MET locus,* ***(B)*** *METΔ14 from NCI-H596 RNAseq*

*sample (27 million reads), 24,850 reads mapping on MET locus, **(C)** 5000 reads randomly selected from **(B)**, **(D)** 1000 reads randomly selected from **(B)**, **(E)** 500 reads randomly selected from **(B)**, **(F)** 250 reads randomly selected from **(B)**.*

Each of the above-mentioned subgroups was converted in 31 and 16 k-mers. MET expression was represented by the amount of each k-mers spanning over MET exons, and these data were used to train the NN. We observed that the learning curve at 16 k-mers was slightly better than the one at 31 k-mers (not shown), and thus we ran the following analyses using the 16 k-mers representation of MET. As training sets, we also used k-mer count frequency [171] for full MET locus, k-mer count frequency for MET exons 13 ÷ 15 and coverage frequency for MET exons 13 ÷ 15.

As test sets, we used subsets of WT and exon 14 skipping MET from cell lines characterized by a physiological MET expression (**Table 1**). NN performance was investigated using as test set: (i) subsets made of random not overlapping subgroups of 500, 1000 and 5000 reads, converted in 16 k-mer counts, (ii) k-mer count frequency on full MET locus, (iii) k-mer count frequency on MET exons 13 ÷ 15 and (iv) coverage frequency for MET exons 13 ÷ 15.

The detection efficiency of METΔ14 using 16 k-mers frequency counts showed best performances at 500 and 1000 reads coverage, **Figure 13A**, **B**, as instead at 5000 reads coverage of all the different test sets performed in the same way, **Figure 13C**.

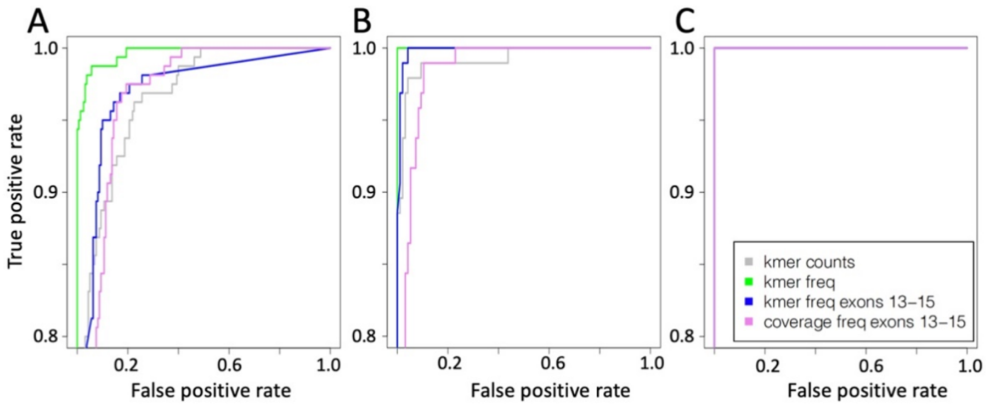


***Figure 13.** ROC curve for NN prediction. **(A)** Training based on 1000 reads coverage for MET locus and test with a coverage of 500 reads. **(B)** Training based on 1000 reads coverage for*

MET locus and test with a coverage of 1000 reads. ***(C)*** *Training based on 1000 reads coverage for MET locus and test with a coverage of 5000 reads. Grey line training and test using k-mer counts, green line training and test using k-mer counts frequency, blue line training and test using k-mer counts frequency only for exons 13 ÷ 15, violet line training and test using coverage counts frequency only for exons 13 ÷ 15.*

2.1.1. Neural Network Validation and Discovery on TCGA Samples

To validate the METΔ14 discovery potential of the above-described NN, we used a set of 690 RNAseq samples from the TCGA bronchus and lung dataset. The 690 samples were manually inspected using the Broad's integrative genomics viewer (IGV) [172] and we detected 17 exon 14 skipping events (2.4%), which is in line with the frequency of the exon 14 skipping events observed in published literature [139, 173]. We tested on this tumor set the NN trained with k-mer counts frequency, which predicted 4 samples out of 17 as METΔ14, but only one was a real exon skipping event (sensitivity 5.88%, specificity 99.5%). The NN trained with exons 13 ÷ 15 MET k-mer counts frequency improved the detection of METΔ14 events, 9 out of 17 (sensitivity 52.9%), but this prediction included a massive increase of false positives, 129 samples, (specificity 81.3%). The best results were obtained using the NN trained using only the coverage frequency for MET exons 13 ÷ 15, which predicted 18 skipping events, including all 17 true skipping events (sensitivity 100%) and one false positive (specificity 99.8%).

Using the NN trained with the coverage frequency for MET exons 13 ÷ 15, we extended the METΔ14 discovery to 2605 TCGA tumor tissues; **Table 2**.

TCGA Tissue	**# Inspected Tissue**	**# Detected METΔ14**	**# Detected False METΔ14**
Adrenal gland	10	0	0
Bladder	280	1	0
Brain	28	0	0
Breast	162	0	0

TCGA Tissue	**# Inspected Tissue**	**# Detected METΔ14**	**# Detected False METΔ14**
Bronchus and lung	690	17	1
Cervix (uterus)	236	0	6
Corpus uteri	109	0	4
Esophagus	165	0	0
Hearth/mediastinum/pleura	78	0	1
Kidney	435	0	3
Pancreas	89	0	0
Skin	288	0	1
Soft tissues	35	0	0

***Table 2**. TCGA samples inspected for the presence of METΔ14.*

We could detect only one METΔ14 in 280 bladder samples. Then, we detected few false METΔ14 in cervix, corpus uteri, heart/mediastinum/pleura, kidney and skin samples, **Table 2**. The six transcripts detected in cervix, **Table 2**, were erroneously detected as METΔ14, because they have a blurry coverage on exons 13 ÷ 15, **Figure 14C**. However, when the full MET locus is observed, it is clear that these METΔ14 false positives are completely different type of transcripts. A shared characteristic of these transcripts is the high accumulation of reads in the second intron (approx. chr7:116,715,690–116,717,329), **Figure 14A**, in the 6th exon, **Figure 14B**, and in the last non-coding MET exon, **Figure 14D**.

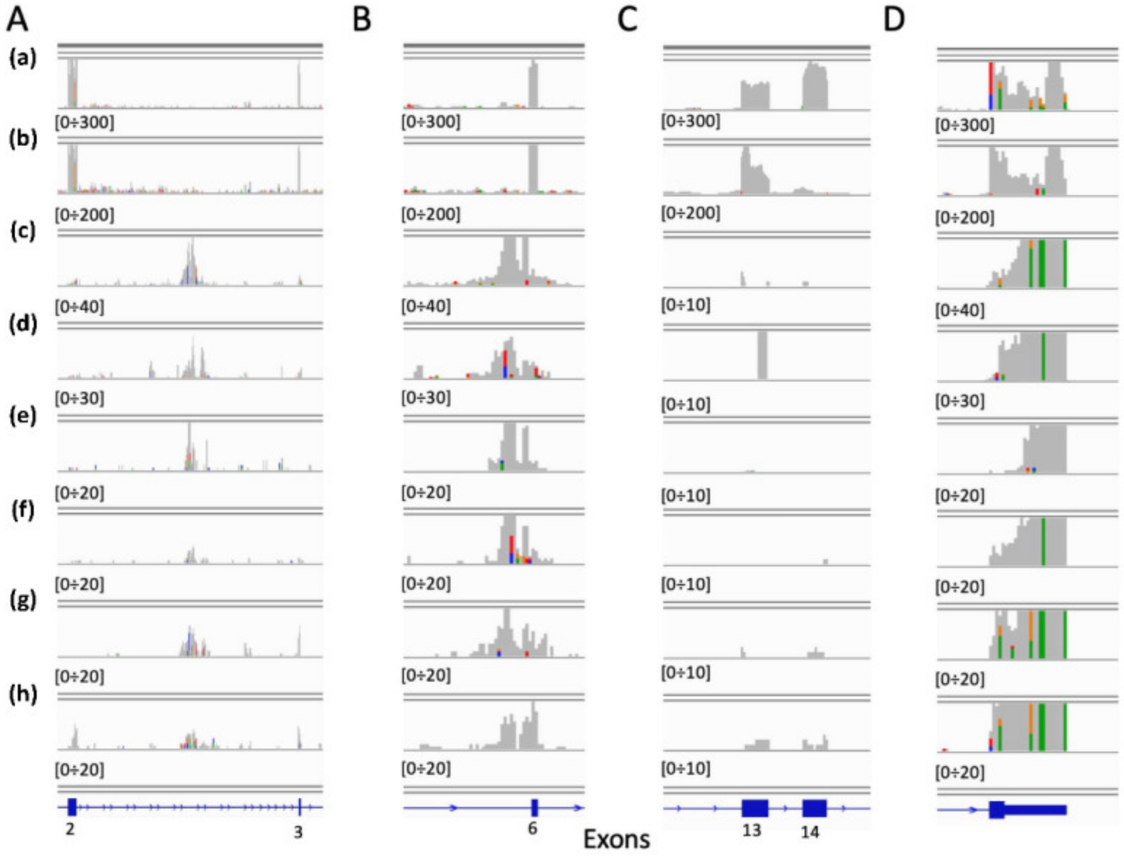


Figure 14*. METΔ14 false positive detected in cervix. (a) WT MET from A549 RNAseq sample (33 million reads), 27,152 reads mapping on MET locus, (b) METΔ14 from NCI-H596 RNAseq sample (27 million reads), 24,850 reads mapping on MET locus, (c–h) False METΔ14. (****A****) Zoom in the 2–3 exons region. (****B****) Zoom in 6th exon region. (****C****) Zoom in 13–15t exons region. (****D****) zoom in last exon region.*

The above observation also applies to the other false METΔ14 detected in corpus uteri, heart/mediastinum/pleura, kidney and skin samples (**Figure 15**).

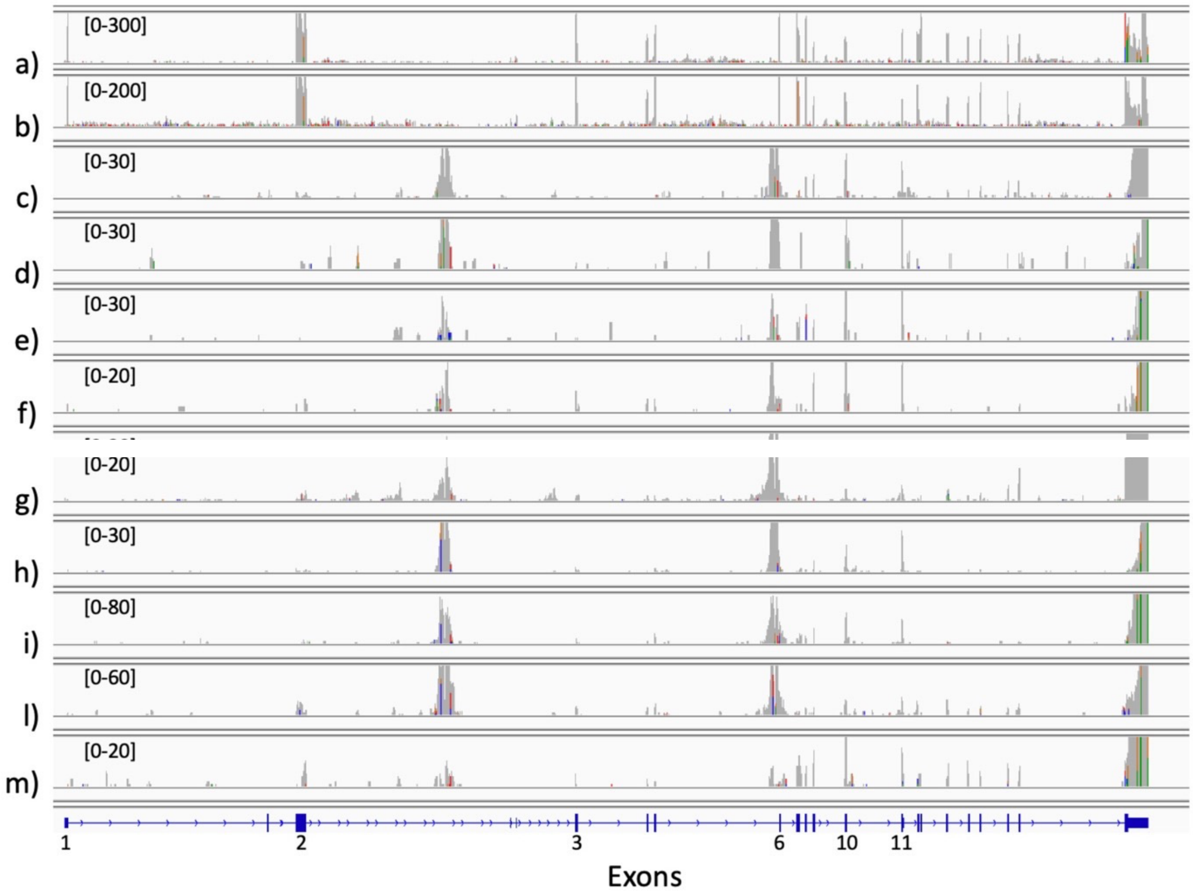


***Figure 15**. METΔ14 false positive detected in corpus uteri. (a) WT MET from A549 RNAseq sample (33 million reads), 27152 reads mapping on MET locus. (b) METΔ14 from NCI-H596 RNAseq sample (27 million reads), 24850 reads mapping on MET locus. (c-f) False METΔ14 in corpus uteri samples. (g) False METΔ14 in heart/mediastinum/pleura samples. (h- l) False METΔ14 in kidney samples. (m) False METΔ14 in skin samples.*

A possible explanation could be that we are observing the transcriptional effect of a LINE1-MET fusion, which was firstly described a few years ago in triple negative breast cancers [174]. We further investigated this point searching for LINE1 alignment, in the subset of MET reads, where only one of the two pair-end reads maps on MET. Indeed, in 10 out of 15 samples, detected as characterized by a transcription peak in MET second intron, we detected LINE1 mapping read. From these we extracted the paired reads associated with MET reads, i.e., only one read of the pair is mapping in MET locus. We blasted [175] these reads on a LINE1 sequence (chr1:62194249–62212928, hg38) and indeed, some of these reads map to LINE1 sequence. On the basis of the MET read position we could identify the putative fusion point with MET, which is mainly located in MET intronic regions and in the last non-coding exon. Unfortunately, we cannot pair the TCGA RNAseq samples to genomics data to further validate the presence of a LINE1 insertion on the basis of genome sequencing data.

2.2. Convolutional Neural Network (CNN) for the Detection of METΔ14

To detect MET exon 14 skipping events, we constructed a CNN made by a 1D convolutional layer, 1D Max pooling layer, a flat fully connected dense layer with 50 nodes and an output layer with one node (**Figure 11C**). The CNN was challenged with the same training and test set used for the previously described feed forward neural network. In this implementation, the convolutional layer included 10 kernels, for more information see Material and Method section. In **Figure 16**, the METΔ14 detection ability of CNN on the basis of different representation of the MET expression data are reported.

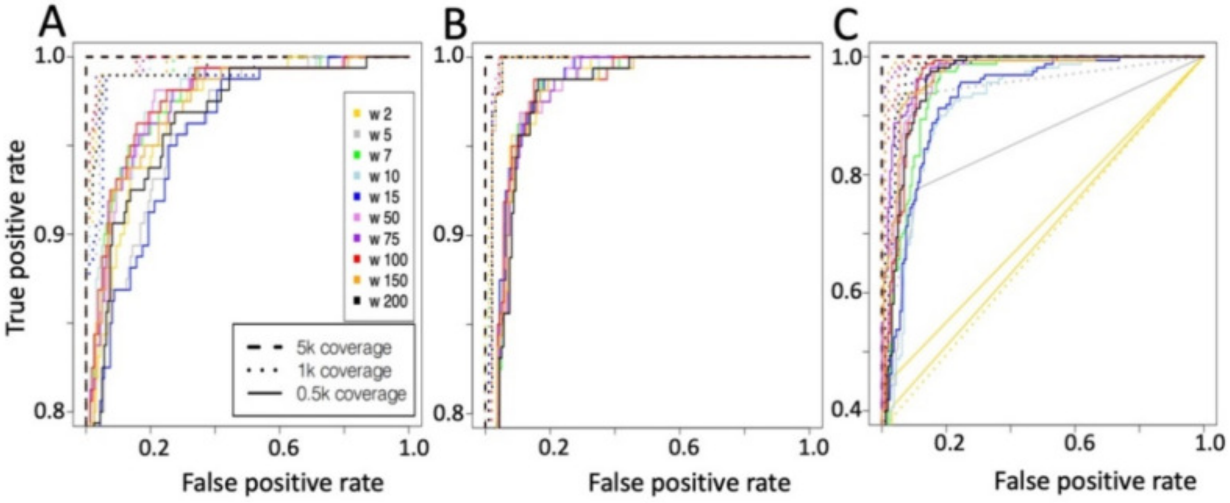


***Figure 16**. ROC curve for CNN prediction. (**A**) 16 k-mer counts frequency for whole MET exons. (**B**) 16 k-mer counts frequency for MET 13 ÷ 15 exons. (**C**) Coverage frequency for MET 13 ÷ 15 exons.*

The results are organized (**Figure 16**) on the basis of the type of input data, i.e., whole MET exons k-mer counts (**Figure 16A**), MET exons 13 ÷ 15 k-mer count frequency (**Figure 16B**) and MET exons 13–15 coverage frequency (**Figure 16C**). The best ratio between true positive and false positive is shown for all kernels using test samples characterized by 5000 reads coverage. As also seen for NN (**Figure 13**), the specificity progressively decreases when the coverage is reduced.

2.2.1. Convolutional Neural Network Validation on Bronchus and Lung Samples

We validated the CNN model using the kernel 100, which is one of the best performing kernels independently by the coverage of the test set (**Figure 16**). The validation was done on the 690 TCGA bronchus and lung sample manually inspected for the presence of METΔ14. We tested on this tumor set the CNN trained with k-mer counts, which predicted 10 samples as METΔ14, but only one was a real exon skipping events (sensitivity 5.88%, specificity 97.6%). The best results were obtained using the CNN trained with exons 13 ÷ 15 MET k-mer counts frequency. All of the 16 samples predicted as METΔ14 belong to the 17 true METΔ14 present in the data set (sensitivity 94.11%, specificity 100%). Finally, the CNN trained using only the coverage frequency for MET exons 13 ÷ 15, predicted 8 skipping events and all of them belong to the true skipping events (sensitivity 47.05%, specificity 100). Since we observed that NN was detecting some false positives in cervix tumor tissues (**Figure 14**), we evaluated if CNN was more specific than NN. CNN trained with exons 13 ÷ 15 MET k-mer counts frequency detects the same false positives detected by NN (**Figure 14**).

2.3. Shallow sparsely Connected Autoencoders (SSCA) to Detect MET Non-Canonical Isoforms

Our group have recently published a paper on the use of SSCA for the identification of hidden functional regulatory elements in single cell RNAseq data [159, 176, 177]. We tested this type of autoencoder to see if we could depict non-canonical isoforms from the analysis of the TCGA samples used in the previous paragraph. The SSCA was designed to take as input k-mer count frequency or coverage frequency of MET exons. The SSCA hidden layer, i.e., latent space, is representing MET exons. Input nodes are only connected to the exon nodes they are associated (**Figure 11B**). We trained the SCA with the 2605 TCGA samples and clustered the latent space data using gridFLOW [178]. To estimate the stability of clusters generated using the SSCA latent space, we compared thousands of pairs of clusters generated by SSCA latent space clustering, as previously described by us [159]. The rationale of this approach is that, if a cluster's organization is conserved, it should be depicted by the multiple comparisons of randomly paired latent space cluster representations [159]. The best results were obtained using normalized [179] MET coverage frequency data (**Figure 16A**). Unfortunately, the stability of the clusters was very poor (**Figure 17A**). Inspection of a randomly selected subset of samples associated with cluster 2 (**Figure 17B**) suggests that at least cluster 2 seems to be made mainly of transcripts recalling the organization of MET-LINE1 fusion, which we have described in previous paragraphs.

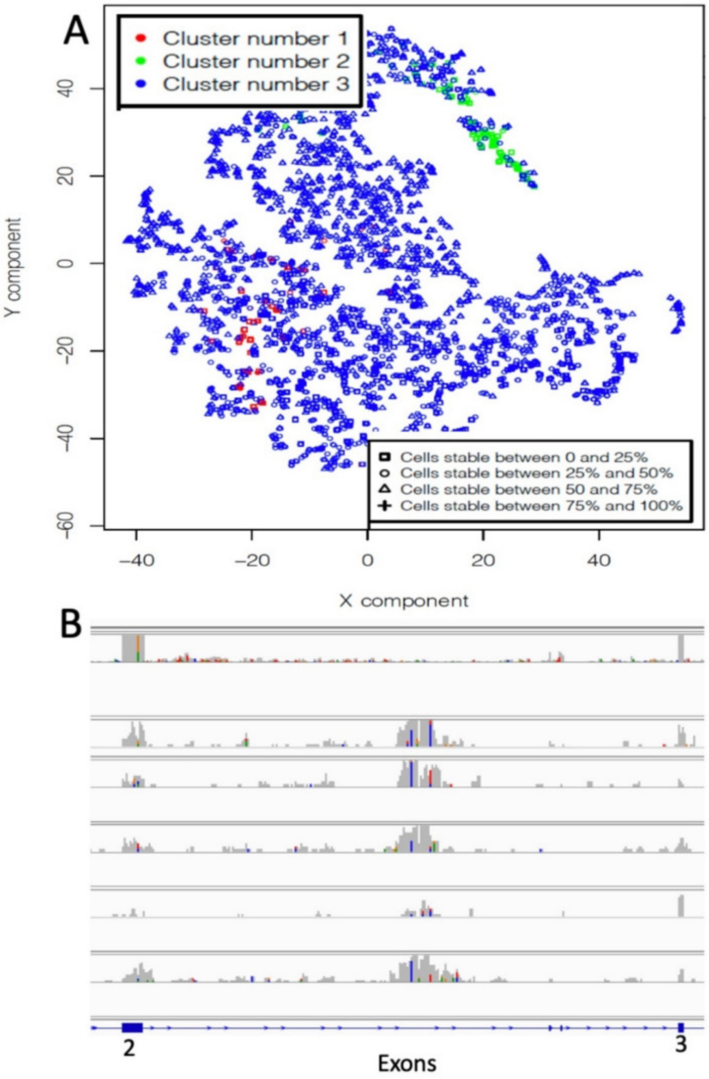


Figure 17. *Autoencoder on saver normalized data. (**A**) Clustering results of the latent space trained with 2605 TCGA samples. (**B**) A limited number of samples (green group in A) is*

characterized the presence of a transcription patter, i.e., the coverage peak in intron 2, which resembles the presence of a LINE1-MET fusion. In B, it is shown a set of samples randomly picked from cluster 2.

3. Conclusions

We used MET exon 14 skipping as a case study for the detection of genetic variants in cancer driver genes through deep learning. In recent years, a lot of evidence has indicated that MET inhibitors have a good anti-tumor effect in patients with MET exon 14 skipping mutation, suggesting that MET exon 14 skipping can be a interesting target for NSCLC patients [147, 180]. Thus, the availability of effective tools for the detection of MET exon 14 skipping are needed for a fast identification of patients suitable for MET targeted therapy.

It is notable that, digging into the published literature, all the found exon skipping tools use nucleotide sequence analysis to infer skipping events, and they are only able to predict skipping events in a generalist way. Since we could not find any tool providing the detection of a unique skipping event in a gene over a large cohort of specimens, we designed specific neural networks for the identification of MET exon 14 skipping, using transcript expression information.

We designed a conventional feed forward neural network (NN) made of four fully connected hidden layers and a convolutional neural network (CNN) made of one 1D convolutional layer, one 1D max pooling layer and a fully connected dense layer. Although we performed an automated optimization of the hyperparameters, the prediction efficacy of our CNN and NN comes from the special attention we put on defining the optimal representation of the data for each architecture, i.e k-mer counts for CNN and coverage from NN.

The NN and the CNN training was done using the RNAseq data of a lung cancer cell line expressing amplified form of the wild type MET (WT, EBC-1), and a gastric cancer cell line expressing exon 14 skipped MET (HS746T). HS746T cell line was selected because, to the best of our knowledge, it is the only cell line displaying amplification of MET exon 14 skipping isoform. MET gene amplification has been observed in about 2–5% of gastroesophageal cancers and represents an oncogenic driver and therapeutic target [181, 182]. MET exon 14 skipping was initially described in NSCLCs (caused by a mutation in the splice donor site in intron 14) and afterwards reported in a variety of tumors, including gastrointestinal cancers, suggesting it as a potential mechanism leading to MET activation [183]. Therefore, HS746T, together with EBC-1, were invaluable instruments to provide a large amount of data for the NN/CNN training. Validation was done instead using RNAseq data from lung cancer cell lines expressing at physiological level MET (A549 expressing WT MET and NCI-H596 expressing exon 14 skipped MET).

Since we could not compare our models with respect to pre-existing methods for MET exon 14 skipping, we manually curated a set of TCGA data to provide an objective evaluation of the performance of our tool. Specifically, we manually curated a cohort of WT and exon 14 skipped samples made of the 690 RNAseq samples belonging to the TCGA (https://www.cancer.gov/tcga, accessed on 1 March 2020) bronchus and lung collection (1310 samples) showing a MET coverage of at least 5000 reads. Given the manual curation of this dataset, i.e., each single sample was inspected on IGV browser for the presence of MET exon 14 skipping, it represents a robust instrument to quantify the predictive performance of our neural network models.

Skewed datasets are not uncommon and the MET exon 14 skipping detection is a typical example. Although skewed datasets are tough to handle, our models, i.e., CNN and NN, seem to handle this issue efficiently, since sensitivity greater than 94% and specificity greater than 99% are reached on an extremely skewed data set such as TCGA bronchus and lung 690 samples with only 17 MET exon skipping events (2.46%). Notably, the high sensitivity is obtained by CNN with a training based on k-mer counts spanning among MET exon 13 and exon 15. Instead, in the case of the NN the optimal sensitivity was obtained with a training based on coverage data encompassing the region among MET exon 13 and exon 15.

Our analysis, using both CNN and NN, on 2605 TCGA tumors (13 primary sites, **Table 2**) highlights that MET exon 14 skipping is a peculiar event of lung specimens. Then, mainly in uterine cancers, we detected a set of MET exon 14 skipping false positives, sharing a common feature: an unexpected peak of coverage in the MET intron 2. This observation brought us to speculate that we were observing a transcriptional signature for a LINE1-MET fusion event [174]. This hypothesis has been supported by the identification of MET paired-end reads, having one read mapping on MET and the other on LINE1 sequence. Notably, transcription of the LINE1-MET fusion was observed in advanced stages of cancer [174, 184], but very little is still known about the effect of the LINE1-MET chimera in cancer.

At the present time, we cannot manage to eliminate LINE1-MET false positives, mainly because we do not have enough data to train a model to detect LINE1-MET fusion, to be implemented in parallel with the MET exon 14 skipping models. However, we are generating large RNAseq data from MCF7, a breast cancer cell line harboring LINE1-MET fusion [174], to build a specific CNN to be integrated with our MET exon 14 skipping models, to improve their specificity.

Having identified more than one artefactual event in MET, we investigated the possibility to discover those anomalous events by the integration of a particular type of deep learning tool, shallow sparsely connected autoencoders [159], with clustering techniques used in multicolor cytometry. Although the actual implementation of the SSCA tool could be further improved in terms of its precision and sensitivity, currently we were able to detect from TCGA specimens a set of tumors sharing the putative LINE1-MET fusion.

Taken together, our results indicate that neural networks can be an effective tool to provide a quick classification of pathological transcription events. However, from the discovery point of view there is still some work to be done to obtain an effective discovery tool using sparsely connected autoencoders.

4. Materials and Methods

4.1. Cell Lines

A549 (lung adenocarcinoma); NCI-H596 (lung adenocarcinoma); Hs746T (gastric adenocarcinoma) cell lines were purchased from ATCC (Rockville, MD, USA); EBC-1 (non-small cell lung cancer) were acquired from HSRRB cell bank (Osaka, Japan). All cells were kept in culture for less than 4 weeks and used between passage 2 and 10. Cells were grown in recommended media (Sigma Aldrich, St. Louis, MO, USA) supplemented with 50 units/mL penicillin (Sigma Aldrich, St. Louis, MO, USA), 50 mg/mL streptomycin (Sigma Aldrich, St. Louis, MO, USA), 2 mM glutamine (Sigma Aldrich, St. Louis, MO, USA) and 10% Foetal Bovine Serum (Lonza Sales Ltd., Basel, Switzerland) as indicated. Cells were maintained at 37 °C in a 5% CO_2 atmosphere.

4.2. Generating the Data for the Neural Network Training and Test Set

We generated RNAseq data from EBC-1 [101], a non-small cell lung cancer (NSCLC) cell line, harboring MET amplification and from Hs746T, a gastric cancer cell line, harboring amplified MET exon 14 skipped isoform (METΔ14) [138]. Furthermore, we have performed RNAseq on human lung adenocarcinoma cell line A549, expressing c-Met [185] and on NCI-H596, derived from an NSCLC, expressing exon 14 skipped MET [130]. Both cell lines express physiological levels of MET.

Total RNA was extracted from cell lines using Trizol reagent (Invitrogen, Carlsbad, CA, USA), following the manufacturer indication. Total RNA was quantified using the Qubit 2.0 fluorimetric Assay (Thermo Fisher Scientific, Waltham, MA, USA) and sample integrity, based on the RIN (RNA integrity number), was assessed using an RNA ScreenTape assay on TapeStation 4200 (Agilent Technologies, Santa Clara, CA, USA).

Libraries were prepared from 400 ng of total RNA using the RNAseq (total RNA Full length) sequencing service (Next Generation Diagnostics srl) which included rRNA-Globin depletion, library preparation, quality assessment and sequencing on a NovaSeq 6000 sequencing system using a paired-end, 300 cycle strategy (2 × 150) (Illumina Inc., San Diego, CA, USA). Read were trimmed to remove adapters sequences using skewer (https://github.com/relipmoc/skewer accessed on 1 March 2021) Read were mapped using STAR on ENSEMBL HG38 human genome assembly.

4.2.1. 16/31. k-mer Training Set

The training set for the neural networks (NN/CNN) was generated using the cell lines with MET amplification: EBC-1 and Hs746T. EBC-1 and Hs746T reads were organized in subgroups of 1000 reads, randomly selected and not overlapping. This approach generated a large set of samples for the training the NN, i.e., 1447 subsets for EBC-1 and 846 for Hs746T. Subsampled reads were associated to MET exons and converted in 16/31 k-mers using BFcounter.

4.2.2. 16/31. k-mer Test Set

The test set for the neural networks (NN/CNN) was generated using the cell lines with physiological MET expression: A549 and NCI-H596. A549 and NCI-H596 reads were organized in subgroups of 500, 1000 and 5000 reads, randomly selected and not overlapping. Subsampled reads were associated with MET exons and converted in 16/31 k-mers using BFcounter [186].

4.2.3. Coverage Training and Test Set

The training and test set for the neural networks (NN/CNN) was generated using the RNAseq data used for the 16/31 k-mers training and test sets. For the training, reads were organized in subgroups of 1000 reads, randomly selected and not overlapping, for the test set, reads were organized in subgroups of 500, 1000 and 5000 reads. Subsampled reads were used to calculate coverage associated with MET exons 13, 14 and 15.

4.3. TCGA RNAseq Datasets

We registered a TCGA project for the study of MET exon 14 skipping events, to obtain access to TCGA raw sequencing data, i.e., RNAseq BAM files. Since the size of the TCGA transcription data exceeds 200 TB, we progressively downloaded the BAM files on the basis of the cancer tissue locus. Then, from each BAM file, we extracted the reads encompassing MET locus (chr7:116672196–116798377, hg38 human genome assembly). We kept only samples where the MET locus was covered by at least 5000 reads. To define 5000 reads as the minimal coverage for MET, we inspected the expected coverage for exons 13, 14 and 15 in A549 (WT MET cell line), in NCI-H596 (METΔ14 cell line) and in random subsets of 5000, 1000, 500 and 250 reads from NCI-H596. We observed that the detection of exon 14 skipping become blurry below 5000 reads coverage. Together with the MET linked reads we also extracted the MET paired reads, where only one of the two reads maps on MET locus.

4.4. Model Coding and Hyperparameter Selection for NN

We constructed a NN made of 6 layers. The input layer has variable size depending on the type of input (k-mers or coverage). 1st and 2nd hidden layers are made of 256 nodes, 3rd and 4th are made of 128 nodes, all using RELU (rectified linear unit) as activation function and 0.1 as dropout rate. The output layer is made by 1 node, associated with a sigmoid activation function. We implemented the models in python (version 3.7) using TensorFlow package (version 2.0.0), Keras (version 2.3.1), pandas (version 0.25.3), numpy (version 1.17.4), matplotlib (version 3.1.2), sklearn (version 0.22), scipy (version 1.3.3). Optimization was done using Adam (Adaptive moment estimation), with the following parameters lr = 0.01, beta_1 = 0.9, beta_2 = 0.999, epsilon = 1e−08, decay = 0.0, loss = 'mean_squared_error'. Hyperparameter optimization was done using Talos (https://github.com/autonomio/talos, accessed on 01 January 2021), which is an automated tool to define the optimal combination of the hyperparameters. Specifically, Talos takes as input the hyperparameter space to be investigated. Then, Talos performs all possible combinations and selects the optimal configuration of the hyperparameters.

The trained NN is implemented in a docker container together with all tools needed to extract MET reads from fastq data. The NN can be used for the discovery of METΔ14 using conventional RNAseq or MET targeted RNAseq. The tool can be requested for the corresponding author. It is provided free of charge to Accademia and non-profit organizations for research use only.

4.5. Model Coding and Hyperparameter Selection for CNN

We constructed a CNN made of one Convolutional 1D layer, characterized by 64 filters and the following kernel sizes: 2, 5, 7, 10, 15, 50, 75, 100, 150, 200. One MaxPooling 1D layer with pool size of 2, for dimensionality reduction. One dense layer with activation RELU and 50 nodes and one dense layer with activation sigmoid and 1 node. We implemented the models in python (version 3.7) using TensorFlow package (version 2.0.0), Keras (version 2.3.1), pandas (version 0.25.3), numpy (version 1.17.4), matplotlib (version 3.1.2), sklearn (version 0.22), scipy (version 1.3.3). Optimization was done using Adam (Adaptive moment estimation) using the following parameters lr = 0.01, beta_1 = 0.9, beta_2 = 0.999, epsilon = 1e−08, decay = 0.0, loss = 'mean_squared_error'. Hyperparameter optimization was done using Talos as done for NN.

4.6. Model Coding and Hyperparameter Selection for Shallow Sparsely Connected Autoencoders (SSCA)

Autoencoders learning is based on an encoder function that projects input data onto a lower dimensional space. Then, autodecoder function recovers the input data from the low-dimensional projections minimizing the reconstruction. We implemented the models in python (version 3.7) using TensorFlow package (version 2.0.0), Keras (version 2.3.1), pandas (version 0.25.3), numpy (version 1.17.4), matplotlib (version 3.1.2), sklearn (version 0.22), scipy (version 1.3.3). Optimization was done using Adam (Adaptive moment estimation) with the following parameters lr = 0.01, beta_1 = 0.9, beta_2 = 0.999, epsilon = 1e−08, decay = 0.0, loss = 'mean_squared_error'. RELU (rectified linear unit) was used as activation function for the dense layer.

Chapter 2

Laniakea@ReCaS contribution

1. Introduction

1.1. Cloud computing for bioinformatics

The explosion in the amount and types of data produced because of the development of next generation sequencing is still ongoing. The increase in the quantity of data produced is constant and it is outpacing the progress of the hardware necessary to handle it, in particularly the progress of data storage solutions [187, 188]. The hardware requirements to store and analyze very large datasets may become increasingly prohibitive for smaller entities, as the costs to produce data keep diminishing. This shows the inefficiencies of the field of analysis of big genomics datasets, where sequencing laboratories upload their data to sequencing archives from which bioinformaticians download them and analyze them on their own machine [188]. Another issue that can limit the exploration of the available data is the fact that bioinformatics workflows are complex, dependent on the specifics of the dataset or datasets being analyzed. These workflows are composed of various different tools, some of which overlapping in functionality, most of which are not trivial to setup, maintain and employ for users without a computational background [189]. This complexity both in the tools to use and how to set them up also translates into bioinformatics workflows being often difficult to reproduce [190]. Cloud computing applied to bioinformatics offers a route to solve these issues [188]. In cloud computing, computing services are rendered available on-demand over the Internet. Three basic service models define different levels of control that the user can have on the environment of the remote machines: infrastructure as a service (IaaS) or hardware as a service offers full control over the amount of virtualized resources being provided and over what runs on the machines even at the operating system level, allowing for great customizability in the to the user, however at the cost of ease of use, as the issues related to setting up the system and the elements of the workflow would still remain;

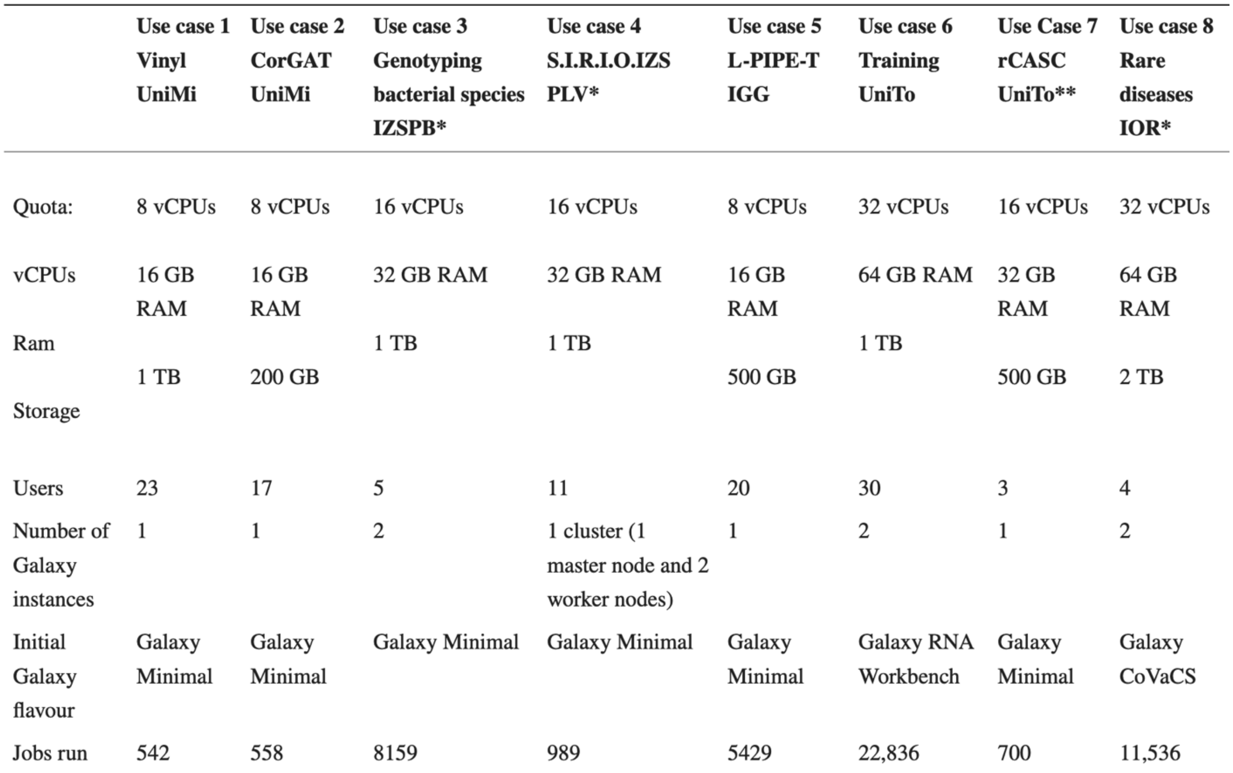
platform as a service (PaaS) removes some of this complexity by allocating resources to the user according to what is needed automatically and dynamically, the control over the software running on the machine is also more limited, not extending to the operating system level but still allows modulation of what software is installed and its configuration; software as a service (SaaS) is the most simplified model, in which specific tools are accessible through specified interfaces like web clients or program interfaces with usually limited configuration settings available. Examples of these models are all available to various extents for bioinformatics purposes [191] both in a specialized manner and in the form of general services like cloud storage services, but one particularly promising candidate for solving the obstacles detailed previously is represented by cloud implementations of the Galaxy platform [192].

1.1.1. Galaxy

Galaxy is a browser-based workflow management system first introduced in 2005 with the objective of allowing experimental biologists with no programming knowledge to explore genomic data by facilitating the access to sequencing data integration, annotation, alignment and querying tools in a user-friendly interface and relying on genome browsers like the UCSC Genome Browser [193] for the visualization of the results on a genome track. Since its inception the platform has expanded outside of genomics, including workflows and related tools for single-cell omics, metagenomics and metabolomics, but also the integration with outside services, including cloud storage and even sees application in works beyond the bioinformatics domain [194]. Galaxy allows the user to setup the environment they need for their analysis, installing the appropriate tools by selecting them from the list of supported ones. Using the tools involves selection through the graphical user interface in the browser with selection of the parameters that the tool needs, avoiding the usage of command line interfaces. The built in history function keeps track of the operations and of the various intermediate stages of the analysis. These histories can be saved as a generalized workflow represented as a flowchart connecting each step of the analysis to the next one with arrows representing the piping of the output of one tool as input for the next, while also saving the specific version of each tool used, allowing for better reproducibility of an analysis on other data. These workflows can be exported and imported by other users that can then reproduce the same environment required to run them by installing the needed tools with the correct versions. Galaxy is available as a public server managed directly by the Galaxy community but can also be ran as an instance on private servers, with hundreds of tools compatible with Galaxy available through the Galaxy ToolShed. The public server has limitations both in the resources available to the user and in the possible concerns related to data privacy, which can be very important when dealing with clinical data. Setting up a private instance is a complicated task, so Laniakea was developed to provide Galaxy on-demand as a cloud-based PaaS [195]. Laniakea@ReCaS is the first Laniakea-based instance and was launched in February 2020 by the Italian node of the ELIXIR initiative [196].

2. Results

I was involved in 2 of the 8 case studies for the Laniakea@ReCaS service [197], **Table 3**.



	Use case 1 Vinyl UniMi	**Use case 2 CorGAT UniMi**	**Use case 3 Genotyping bacterial species IZSPB***	**Use case 4 S.I.R.I.O.IZS PLV***	**Use case 5 L-PIPE-T IGG**	**Use case 6 Training UniTo**	**Use Case 7 rCASC UniTo****	**Use case 8 Rare diseases IOR***
Quota:	8 vCPUs	8 vCPUs	16 vCPUs	16 vCPUs	8 vCPUs	32 vCPUs	16 vCPUs	32 vCPUs
vCPUs	16 GB RAM	16 GB RAM	32 GB RAM	32 GB RAM	16 GB RAM	64 GB RAM	32 GB RAM	64 GB RAM
Ram			1 TB	1 TB		1 TB		
	1 TB	200 GB			500 GB		500 GB	2 TB
Storage								
Users	23	17	5	11	20	30	3	4
Number of Galaxy instances	1	1	2	1 cluster (1 master node and 2 worker nodes)	1	2	1	2
Initial Galaxy flavour	Galaxy Minimal	Galaxy Minimal	Galaxy Minimal	Galaxy Minimal	Galaxy Minimal	Galaxy RNA Workbench	Galaxy Minimal	Galaxy CoVaCS
Jobs run	542	558	8159	989	5429	22,836	700	11,536

Table 3. Summary of the allocated resources and some usage statistics for the Laniakea@ReCaS use cases. (*This instance will be soon updated to increase compute and storage requirements).

2.1. Learning platform for undergraduate bioinformatics students (use case 6)

Teaching bioinformatics to a broad audience requires huge computation infrastructure and a massive teaching effort. With little setup effort by teachers, the Laniakea@ReCaS service provides to students the Galaxy web-based interface, allowing students to practice with bioinformatics concepts and algorithms, avoiding the steep learning curve needed to use UNIX shell, R or Python environments. The platform enabled biology-oriented students (without a specific computer science background) to run complex workflows, analyze real data and learn how to interpret the results in a learning-by-doing environment.

The cloud-based infrastructure of Laniakea@ReCaS proved to be an invaluable tool for teaching that, due to the COVID-19 pandemic. Starting from publicly available databases or custom files shared with students through the Galaxy file-sharing system, students were able to follow laboratory lessons in teams from home, as well as practice alone when they preferred without the need to book computer rooms at the University.

Assessment tests were performed in the same environment, providing the students with real-world data to analyses, evaluating the knowledge acquired, and the competence developed and the skills mastered by students at the end of the course.

The analysis history logbook provided by Galaxy is particularly interesting; indeed, it allowed us to evaluate the progress of each student step by step, promptly identifying points that showed specific difficulties during lectures and exercises, and to check for cheating during exams.

To allow concurrent practical exams for more than 30 students effortlessly, we developed a custom procedure to replicate Galaxy virtual machine images (including user authentication data and shared files). The course contributed to the realization of the objectives of the Biological Sciences Course, providing the students with basic knowledge in the field of bioinformatics. The Laniakea@ReCaS service was used by students and teachers to perform simple tasks (like aligning two protein sequences) as well as entire NGS pipelines as RNA-seq, ChIP-seq, variant-calling, including some downstream analysis like GO enrichment and KEGG pathway analysis.

2.2. Porting rCASC (reproducible classification analysis for single cells) to Galaxy: a complete analysis workflow facilitating single-cell RNAseq data analysis (use case 7)

Single-cell RNA-seq (scRNAseq) is a very powerful instrument to depict the overall cell complexity of healthy and disease tissues [198]. scRNAseq has today many different facets, spanning from full transcripts single-cell sequencing [199] to spatial transcriptomics [200] via droplet-based technology [201]. Different types of scRNAseq methods require dedicated data analysis workflows, which often are not user-friendly enough to be handled by life scientists with limited coding skills. rCASC [159, 202, 203] was developed at the University of Turin to provide a friendly environment to life scientists for the analysis of multiplatform scRNAseq, granting functional and computational reproducibility [204]. rCASC provides a complete set of analysis tools and pipelines allowing: i) conversion of raw data in count table, ii) cells' quality control, iii) preprocessing, iv) normalization, v) clustering, vi) cluster-specific markers detection, vii) biological knowledge extraction [159]. One of the peculiarities of rCASC is the possibility to evaluate clusters' robustness via the cell stability score (CSS) [203]. The 88 tools and functions of the rCASC workflow are currently packaged as Docker containers, while input, output and tools are managed through R scripts. In this use case, we are at work making the whole workflow compatible with Galaxy leveraging the Laniakea@ReCaS service and are currently at one-third of the effort. For example, we have recently finished the porting in Galaxy of the new rCASC data mining instrument based on Sparsely Connected Autoencoders (SCA) [159]. This mining tool allows the identification of elusive players of cell clusters formation, such as transcription factors and miRNAs. CSS and SCA require the execution of multiple clustering jobs, making it difficult to perform such tasks onto conventional laptops. The Galaxy implementation of rCASC, which we are developing at Laniakea@ReCaS, offers at the same time a user-friendly environment and the possibility to customize Galaxy instances optimized for this specific analytical task and the dataset under analysis.

3. Conclusions

Laniakea@ReCaS provides researchers with a ready-to-use Galaxy environment backed by suitable computational and storage resources to handle their data analysis needs. As such, the service represents an example of a straightforward access channel to the computational resources provided by scientific cloud facilities and infrastructures, a channel that conveniently hides the complexity of the underlying software and hardware layers.

One of the defining features of the service, as it emerges from most of the reported use cases, appears to be its customizability, that is the possibility for the user to freely and easily configure, modify, and manage the Galaxy environment. As such, this feature represents one of the most notable differences between a Galaxy on-demand service like Laniakea@ReCaS and a classic Galaxy public instance. Perhaps, the most interesting outcome made possible by this feature, one that we did not fully anticipate when Laniakea@ReCaS was launched, is that the service is being actively used as a platform to quickly develop and make available or more accessible to the community novel Galaxy based services as rCASC (use case 7).

Chapter 3

Experimental Immunology collaboration

These analyses are part of two papers in which I am coauthor (Devan et al. and Chancellor et al.), which are currently under review.

1. Introduction

1.1. Innate and Adaptive Immune Systems

The immune system represents the defense of the body against pathogens, toxins and altered cells like cancer cells [205]. The immune system is split between two aspects, the innate immune system and the adaptive immune system, which interact with each other heavily. The innate immune system is the first layer of defense against pathogens, providing a response during the first hours and days after exposure, and it includes constitutive and inducible mechanisms. Constitutive mechanisms are always active without requiring a stimulus, like the physical and chemical barriers provided by the skin, mucus, saliva and stomach acids. Inducible innate immunity on the other hand is a host of cell-mediated responses that activate when the constitutive mechanisms aren't enough. These mechanisms involve cells like macrophages, neutrophils, eosinophils, basophils, mast cells and dendritic cells, that carry conserved germline-encoded pattern-recognition receptors (PRRs) such as Toll-like receptors, which are able to recognize highly conserved regions in microbial components called pathogen-associated molecular patterns (PAMPs), with different PRRs being specific for different PAMPs [206]. The binding of the PRRs to their target PAMPs leads to the production of cytokines and chemokines that enable the attraction of other immune cells to the site of infection and contribute to the activation of the adaptive immune system [207]. The adaptive system provides a slower line of defense than the innate immune system, requiring weeks to be engaged [205], however it offers high specificity and the ability to conserve a memory of the immunological response. It is mediated by two cell types, B lymphocytes or B cells and T lymphocytes or T cells. Molecules that are recognized by these cells are called antigens and both B cells and T cells are able to recognize a wide range of antigens while also retaining high specificity in their response thanks to their receptors, the BCR and the TCR respectively, which are formed through the somatic recombination of the DNA segments composing them, allowing the antigen-binding site to be unique [205]. B cells generate their unique BCRs during development in the bone marrow, from which derives the B of B cells, through recombination of the immunoglobulin heavy and light chains which compose antibodies, of which the BCR is a membrane bound form of. Each cell produces only antibodies with the same antigen-binding site, in accordance with Burnet's theory of clonal selection, with only one of the alleles being expressed and the other silenced through the process known as allelic exclusion [208-210]. In the bone marrow B cells that present BCRs with specificity for self-antigens are removed from the repertoire of B cells through negative selection. The B cells surviving negative selection migrate through the bloodstream to secondary and tertiary lymphoid organs in which they can encounter their antigens. After experiencing their antigens B cells undergo activation, proliferate and differentiate into either effector cells and eventually plasma cells or memory cells. Effector cells can secrete soluble antibodies and their maturation endpoint is plasma cells, which are large cells that continuously secrete antibodies and lose the ability to proliferate further [208]. The secreted antibodies are able to inactivate viruses and toxins but are also the classical pathway of activation for the complement cascade, which can lead to results like inflammation, lysis of the target or opsonization, tagging the target for phagocytosis [211, 212]. Memory B cells persist after the elimination of the antigen and are able to be activated by their antigen much faster than other B cells, leading to a more rapid and effective response to a second infection.

1.1.1. T cells

T cells derive from multipotent hematopoietic stem cells in the bone marrow, however instead of maturating in the bone marrow, T cell precursors migrate through circulation to the thymus where they mature, which is the reason for the name T cell. During maturation in the thymus, T cells are known as thymocytes. The T cell receptor (TCR) is the mediator of T cell activation and it's a heterodimer either of one α and one β chain in αβ T cells or of one γ and one δ chain in γδ T cells and recognizes antigens in complex with an antigen-presenting cell, which is able to present to the surface immunogenic molecules that are normally localized intracellularly, allowing T cells to detect antigens that are within cells instead of only on the surface like B cells [208]. Like with the antibodies in B cells, T cells generally only produce TCRs with one type of antigen-binding site. The TCR is rearranged in the thymus and the cells undergo a positive selection step based on the ability of their TCRs to recognize self-antigens in complex with the antigen-presenting molecules, however if the reaction is too strong they undergo apoptosis, providing also a negative selection step. Mature Naïve T cells migrate to secondary or tertiary lymphoid organs where they can encounter their antigens associated to the appropriate antigen-presenting molecules on the surface of antigen presenting cells like dendritic cells [213], which starts their activation, leading to clonal expansion and differentiation into effector and memory T cells, in some cases causing the T cell to move to the site of infection to help eliminating the target either through direct killing in the case of cytotoxic T cells or by activating macrophages or B cells in the case of helper T cells [205, 214].

1.1.1.1. T cell receptor

As described before, the T cell receptor is a heterodimer formed of two different chains, either one α and one β chain in αβ T cells or one γ and one δ chain in γδ T cells, linked by a disulfide bond. These chains all share a similar structure, being composed of a constant region C, a variable region V and a joining region J, with the addition in the case of the β and δ chains of a diversity region D (**Figure 18**). In the germline DNA the genes encoding for all the chains carry multiple possible segments for these regions and during somatic recombination a single segment is selected for each region in each chain in every developing T cell, except in the case of the D region of δ chains in γδ T cells, where multiple D segments can be selected. The V segments for each chain are separated in superfamilies that are highly similar within their members, with variability between them being concentrated in two regions called CDR1 and CDR2, which form loops that mainly make contact with the antigen-presenting molecule in the antigen-antigen-presenting molecule complex [215-217].

Specificity	V segments	J segments	D segments
α chain	~70	61	0
β chain	52	13	2

Specificity	V segments	J segments	D segments
γ chain	14	5	0
δ chain	8	4	3

Table 4. *Number of TCR chain gene segments contributing to combinatorial TCR loop diversity for each chain.*

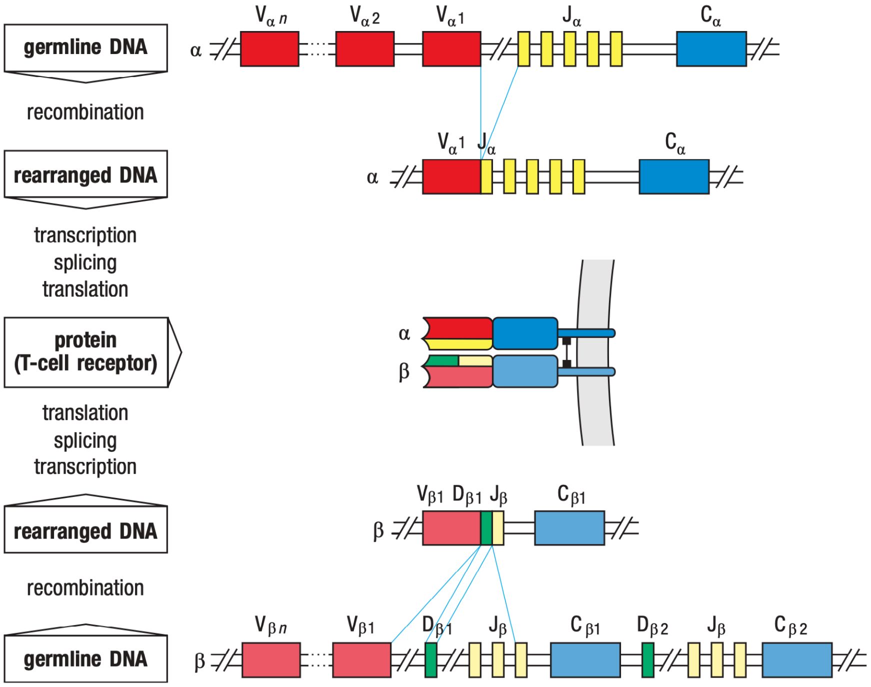


Figure 18. *Rearrangement of the α and β chains into an $\alpha\beta$ TCR. Figure from [205].*

A third CDR region, the CDR3, is hypervariable and the main source of variability of the TCR and is found in the junction between the V and J regions, including the D region between them in the case of β and δ chains. The CDR3 forms another loop in the protein that makes contact with the antigen bound in the antigen-antigen-presenting molecule complex and is a major driver of antigen specificity [217]. The role of the CDR3 loops in recognizing antigens requires them to be highly variable to be able to target a wide variety of antigens with high specificity. The variety afforded by the selection of single segments out of the various possibilities for the two or three regions contributing to the CDR3 is defined combinatorial diversity (**Table 4**), however another CDR3 variability component is found in junctional diversity. Junctional diversity is diversity that arises due to the loss and addition of random nucleotides at the interfaces between the V, J and, when present, D regions. Exonucleases may remove some nucleotides during the joining process, while deoxyribo-nucleotidyl transferases (TdT) can add random nucleotides called N nucleotides and in some rearrangement cases some nucleotides may shift from one DNA strand to the other in the location of the cut between the two DNA strands, causing palindromic sequences called P nucleotides [205, 218]. The rearranged TCRs are tested for their ability to recognize self-peptide-antigen-presenting molecule complexes in the thymus and eliminated if their reaction is too weak or too strong and due to allelic exclusion usually only one of the alleles for the chains is expressed, however cases of double expression for α chains are relatively frequent, estimated at ~30% at the transcript level and ~10% at the surface expression of the protein level [219].

1.1.1.2. MHC-restricted T cells

The vast majority of T cells are $\alpha\beta$ TCR T cells that recognize antigens in complex with Major Histocompatibility Complex (MHC) class I or class II, which are highly polymorphic antigen-presenting molecules [205]. These T cells are defined as conventional T cells, compared to the unconventional T cell subsets which recognize antigens bound to other antigen-presenting molecules. Two main populations of MHC-restricted cells can be defined by considering the expression of the co-receptors CD4 and CD8. The expression of these two co-receptors can be tracked during maturation of the T cells in the thymus, where they start as double-negative thymocytes before TCR rearrangement, and convert in double-positive thymocytes during the rearrangement process and undergo rapid proliferation [205]. During positive selection these cells lose the expression of one of the co-receptors, remaining as single-positive thymocytes which will mature into single-positive Naïve T cells. CD8$^+$ T cells recognize short peptide antigens in complex with MHC class I, which is found on all nucleated cells. These cells are also called cytotoxic T cells because after stimulation they undergo rapid proliferation and start producing cytotoxic granules containing granzymes and perforins, resulting in the death of the targeted cell. CD4$^+$ T cells recognize antigens in complex with MHC class II, which is only found on the surface of cells specialized in antigen presentation, like dendritic cells. The main role of CD4$^+$ T cells is of helper T cells, supporting other immune cell subsets in their duties by releasing cytokines and different subtypes (Th1, Th2, Th9, Th17, Th22 and Treg) of CD4$^+$ T cells can be defined depending on what cytokines they release [220].

1.1.1.2.1. Memory T cells functional maturation

During functional maturation the long-lived memory cells of the classical T cell subsets, both CD8$^+$ and CD4$^+$, can be split into four main populations: Naïve, Central Memory (CM), Effector Memory (EM) and Terminally Differentiated Effector Memory re-expressing CD45RA (TEMRA) [214, 221, 222]. Naïve T cells are antigen-inexperienced cells located in secondary lymphoid organs like lymph nodes and after antigen encounter, they can transition to either Central Memory or Effector Memory depending on cytokine stimulation and evidence suggests that TCR signaling strength also plays a factor, with weaker TCR signaling leading to CM rather than EM and CM being cells that were arrested in an intermediate stage of differentiation due to suboptimal stimulation [221, 223]. CM cells home preferentially to the lymph nodes and they show very low effector capabilities, but compared to Naïve cells they have higher sensitivity to antigen restimulation, after which they show high proliferative capacity and are able to differentiate efficiently to effector cells [221]. EM cells on the other hand migrate to peripheral tissues and, as the name suggests, show immediate effector function, but they are equipped with reduced proliferative capacity [221]. This suggests a model in which EM cells manage the first response to a pathogen while CM cells proliferate and differentiate into a second wave of effector cells [214]. TEMRA cells are a highly differentiated subset of EM cells that have very high effector potential but very low proliferative capabilities due to having short telomeres and approaching a senescent state, suggesting that the EM cells have to undergo repetitive proliferation to reach this stage [224]. These different populations can be tracked by following the surface expression of the marker CC motif chemokine receptor 7 (CCR7), which facilitates homing to secondary lymphoid organs, and of two isoforms of CD45, CD45RA and CD45R0 (or CD45RO). Naïve and CM T cells reside in lymph nodes and are CCR7$^+$, however Naïve cells are CD45RA$^+$, while CM cells are CD45R0$^+$. EM cells are more likely to be found in tissues than in lymphoid organs, so it's not surprising that they are CCR7$^-$ but like CM cells they are CD45R0$^+$. TEMRA cells are CCR7$^-$ like EM cells but they switch back to CD45RA$^+$ (**Figure 19**) [222]. Other markers like

KLRG1, PD1, CD57, CD31, CD27 and CD28 can provide additional insight in the proliferative history and potential of the cells due to the relationship between their expression and the length of the telomeres (**Figure 19**) [222].

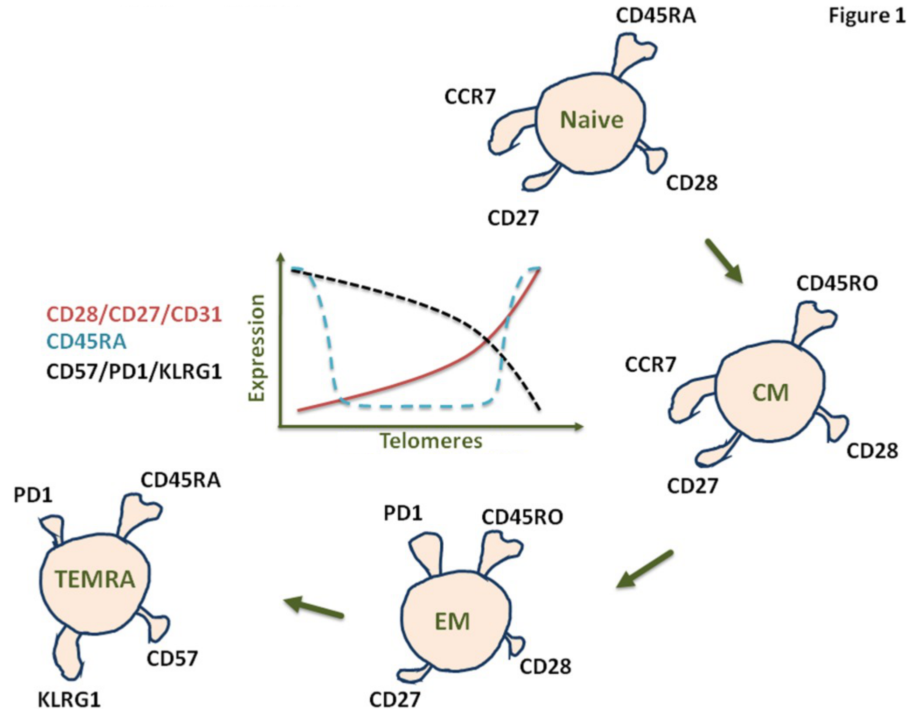


Figure 19. *Expression of surface markers during memory T cell functional maturation and telomeres length. Figure from [222].*

1.1.1.3. CD1-restricted T cells

CD1-restricted T cells recognize lipid antigens in complex with the CD1 family of MHC class-I-like molecules. CD1 is a family of molecules composed of 5 members which are invariant or display limited polymorphism [225]. Of the 5 members of this family, 4 are antigen-presenting molecules divided in two groups based on sequence homology: group 1 is made of CD1a, CD1 and CD1c and group 2 is made of only CD1d [226]. The fifth member of this family, CD1e, is also part of group 1 but it is not expressed on the cell surface and it's considered to be involved in the processing of the lipid antigens and in the loading of the antigens on the antigen-presenting members of the family [227, 228]. Members of the group 1 of CD1 molecules are expressed only on thymocytes and cells specialized in antigen presentation like dendritic cells, while CD1d is more widely expressed, even on nonhematopoietic cells [229]. Group 1-restricted T can be CD4$^+$, CD8$^+$ or double-negative and are known mainly for binding microbial lipid antigens derived from mycobacteria like *Mycobacterium tuberculosis* and *Mycobacterium leprae*, but can also commonly recognize self-lipids and have been implicated in autoimmune disorders [230, 231]. After antigen stimulation, group 1-restricted cells undergo clonal expansion and show cytotoxic effector functions [226]. CD1d-restricted cells are also CD4$^+$, CD8$^+$ or double-negative, they are known as Natural Killer T (NKT) cells due to being often CD161$^+$ and are classified into two types. Type I NKT cells are called invariant NKT (iNKT) cells due to always having an invariant TCR α chain made of TRAV24 and TRAJ18 paired to only a few possible β chain combinations [226]. These cells are considered innate-like T cells as they represent a hybrid between adaptive and innate immunity due to their limited TCR repertoire pushing them closer to the germline-encoded innate PRRs, due to presenting immediate effector capacity without establishment of immunological memory and due to the ability to respond to innate signals independently of TCR stimulation [227, 232]. The first antigen known for iNKT cells was αGalCer, produced by *Agelas mauritianus*, a marine sponge [233]. The involvement of iNKTs in the immune response shows both protective and detrimental effects, for example in autoimmune diseases and numerical or functional deficiencies of iNKT cells have been reported in some tumors [234]. Type II NKT cells are much more heterogeneous in their TCR composition than type I and less is known about them, including specific antigens capable of leading to their identification. Type II NKT cells are not stimulated by αGalCer and in tumor models they have been suggested to be implicated in the suppression of tumor immunosurveillance [235, 236].

1.1.1.4. γδ TCR$^+$ T cells

T cells carrying a $\gamma\delta$ TCR compose between 0.2 and 20% of adult T cells and are a heterogenous group that displays reactivity to a diverse set of antigens associated with various antigen-presenting molecules [237, 238]. They are considered by some to be a cross between innate and adaptive immune systems or outright part of the innate immune system because of the rapidity and strength of their response to stimulation and of their limited combinatorial TCR diversity due to the low number of possible chain segments (**Table 4**), with only some of the V segments of the γ chain encoding for functional TCRs and the presence of preferential pairings [239-241]. The TCR V gene selection on the δ chain and the pairing between the V genes expressed in the γ and the δ chains can provide a method to differentiate between different subsets. Cells carrying the pairing of TRGV9 and TRDV2 (Vγ9Vδ2) represent the main subset of γδ T cells in peripheral blood, recognize phosphorylated metabolites bound to butyrophilin 3A proteins [238, 240] and have been shown to possess anti-tumor activity [242, 243]. Cells expressing TRDV3 (Vδ3) represent a smaller subset found in liver and gut epithelium that are expanded in cases of B cell chronic lymphocytic leukemia and include some cells that are CD1d-restricted or MHC class I-restricted [244, 245]. The subset of cells expressing TRDV1 (Vδ1) is the most prominent in tissues, especially the mucosa [239, 246]. Vδ1 cells have a broad range of antigens they show reactivity to, presented by different antigen-presenting molecules like CD1d, CD1c and other MHC-like molecules [238]. They are also a prominent population within the γδ T cells infiltrating solid tumors, however their influence on tumors is controversial, as they've showed both tumor-suppressing capacity and pro-tumor effects, for example through IL-17 production leading to a tumor-promoting microenvironment [241, 247-249].

1.1.1.5. MR1-restricted T cells

MHC class I-related protein 1 (MR1) is a monomorphic antigen-presenting molecule similar to MHC class I that is transcribed ubiquitously but with low surface expression due to needing a ligand to refold to a stable form [250, 251]. Ligands of MR1 include bacterial metabolites, in particular vitamin B-related antigens like 6-formylpterin (6-FP) and 5-(2-oxopropylideneamino)-6-D-ribitylaminouracil (5-OP-RU) [252, 253], but also tumor-associated antigens [254]. There are two populations of MR1-restricted T cells: Mucosal Associated Invariant T (MAIT) cells and MR1T cells. MAIT cells are semi-invariant T cells similarly to iNKT cells, with an α chain with a fixed V segment (TRAV1-2) and limited J segments (TRAJ12, TRAJ20 or TRAJ33) coupled with limited β chain combinations [254]. They represent around 1-10% of circulating T cells in healthy individuals and are enriched in tissues like mucosa, gut lamina propria, liver, lungs and skin and the majority are single-positive CD8$^+$ (~80%), with double-negative as the second most frequent (~15%) and few single-positive CD4$^+$ and double-positive cells [254]. They are considered innate-like T cells due to sharing characteristics with innate cells in the same way iNKT cells do, carrying semi-invariant TCRs, displaying high levels of CD161 and being characterized in adults by an EM phenotype, responding to antigen restimulation quickly with production of granzyme and perforins that endow them with cytotoxic capacity [227, 255, 256]. MAIT cells recognize some of the bacterial vitamin B-related antigens presented by MR1, 5-OP-RU being a canonical potent stimulator, which, coupled with the potent effector response after stimulation, suggests a role in combating microbial infections. Of particular importance in maintain activity of MAIT cells in the recognition of 5-OP-RU is the conservation of the length of the α chain CDR3 and of a Tyrosine in position 95 of the α chain [254]. The amount of MAIT cells in circulation is lowered in individuals affected by infections, both bacterial and viral, by autoimmune and metabolic disorders and by a number of cancers, indicating their migration towards the sites of inflammation in peripheral tissues [254]. MR1T cells, the second type of MR1-restricted T cells, are a heterogenous population occurring with a much lower frequency than MAIT cells in circulation (~0.04%) and are stimulated by self-antigens bound to MR1 presented on the surface of multiple cancer cell lines in vitro and in vivo and not by bacterial antigens or healthy cells [254, 257, 258]. Currently a lot is unknown about MR1T, including the specific self-antigens involved in the activation and the frequency of these cells in cancer patients, but they represent a great potential avenue for T cell therapy to combat cancer [259].

1.2. Crohn's disease

Crohn's disease is one of the two categories of inflammatory bowel disease (IBD), together with ulcerative colitis (UC). It manifests as transmural chronic inflammation that can affect any part of the gastrointestinal tract. Endoscopic hallmarks of CD are represented by longitudinal ulcers, cobblestone appearance of the mucosa, fissures and thickening of the wall with narrowing of the lumen and presence of abscesses (**Figure 20**) [260-262]. Symptoms involve abdominal pain, diarrhea, blockage, perianal lesions and can extend to extraintestinal effects in ~20-50% of cases that can involve joints, eyes and skin [260, 262]. Prevalence of the disease has been increasing since the 2000 and the incidence is affected by variables such as geographic location. Genetic, epigenetic and environmental elements are all involved, but the specifics of the etiology have yet to be fully elucidated [262].

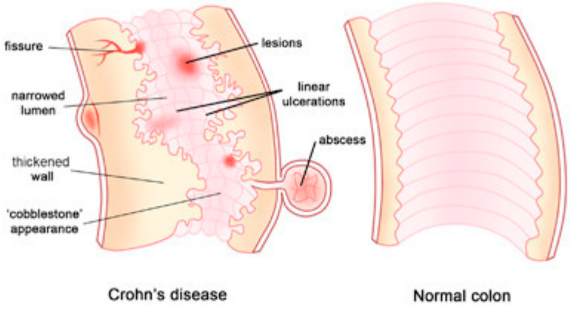


Figure 20*. Endoscopic hallmarks of CD. Figure from [260].*

Risk factors include alterations that affect the function of the intestinal barrier and of the microbiome, in particular defects in the barrier can allow bacteria to penetrate into the tissue and trigger a strong immune response, which can further permeabilize the barrier and perpetuate the inflammatory state further as it happens with neutrophils [263]. CD4$^+$ T cells accumulate in lesions of CD patients, with CD being considered for a long time a Th1-driven condition while UC was characterized by Th2 cells, but the discovery of Th17 cells changed that view to include both Th1 and Th17 as important cells mediating the inflammation [263-265]. While most of the attention was focused on CD4$^+$ T cells, some studies have started exploring the heterogeneity of CD8$^+$ cells in CD patients, with a study showing a population of CD103$^+$ CD8$^+$ T cells with a strongly modified gene expression profile closer to Th17 CD4$^+$ cells and a study demonstrating the possibility of predicting CD patient prognosis based on the gene expression profile of CD8$^+$ cells, implying their importance in the disease [266-268]. Also in need of further investigation, the participation of $\gamma\delta$ T cells in the progression of CD is currently providing mixed results, with studies concerning their quantification in both circulation and in the mucosa of the intestine providing contradicting results [269].

2. Crohn’s Disease (CD) clustering analysis

2.1. Results

2.1.1. Biopsies characterization and selection

We were interested to evaluate the differences in T-cell composition among CD patients in active disease condition and remission condition with respect to healthy individuals. We performed our analysis by Flow Cytometry using a panel of markers for T cells (**Table 5**) (see below for a more detailed description of the markers used for clustering analysis). Biopsies of both inflamed and adjacent not inflamed tissue from the gut were collected from 24 CD patients, 17 gut biopsies were collected from CD patients in remission and 13 gut biopsies were collected from individuals not affected by CD. Clustering analysis was performed to define cell type distribution in biopsies of inflamed tissue from CD patients compared to healthy biopsies primarily and secondarily compared to not inflamed tissue from CD patients and to biopsies of individuals in remission. The analysis was performed on CD4$^+$ T cells and Vδ1 TCR$^+$ T cells separately. The number of cells per biopsy was characterised by very high variability, with the Vδ1 TCR$^+$ T cell counts being overall much lower in active CD biopsies compared to healthy individuals (**Figure 21A**), while the distribution was more even for CD4$^+$ T cell counts (**Figure 21B**).

In the CD4$^+$ T cells case, we selected the 10 biopsies with the higher number of cells for each condition and we used 1000 cells for each biopsy. In the Vδ1 TCR$^+$ T cells case the number of cells was much lower, but this wasn’t always tied to the biopsies of origin being smaller, since often the Vδ1 TCR$^+$ T cell counts represented a much smaller fraction of total cells for CD patient biopsies. Due to the possibility that these biopsies with lower Vδ1 TCR$^+$ T cell fractions held relevant information, we decided to use total cell counts for the selection of patients instead of the Vδ1 TCR$^+$ T cell counts directly. We selected 61 biopsies encompassing at least 4000 cells, and we used at most 500 cells for each biopsy.

Specificity	Clone	Fluorochrome	Ref.no	Manufacturer
CCR7	GO43H7	AF647	353218	Biolegend

Specificity	Clone	Fluorochrome	Ref.no	Manufacturer
CD127	eBioRDR5	PE-CY5.5	35-1278-42	TermoFisher
CD14	63D3	PE-Fire640	367154	Biolegend
CD155	SKII.4	PE- Dazzle 594	337616	Biolegend
CD161	DX1	BUV661	750382	BD Biosciences
CD162	KPL-1	PE-CY7	328816	Biolegend
CD19	HIB19	PE-CY5	302218	Biolegend
CD226	11A8	BV510	338330	Biolegend
CD25	CD25-3G10	PE-AF700	MHCD2524	TermoFisher
CD27	M-T271	PE	356406	Biolegend
CD3	UCHT1	AF700	300424	Biolegend
CD357	108-17	BV421	371208	Biolegend
CD38	HIT2	APC-Fire 810	303550	Biolegend
CD39	A1	BV711	328228	Biolegend
CD4	SK3	Spark Blue	344656	Biolegend
CD45RA	HI-100	Spark NIR 685	304168	Biolegend
CD45R0	UCHL1	BV750	304262	Biolegend
CD49d	9F10	BV785	304344	Biolegend
CD49e	11A1	BUV737	741849	BD Biosciences
CD56	NCAM16.2	BUV563	612928	BD Biosciences
CD57	HNK-1	PerCP-Cy5.5	359622	Biolegend
CD71	CY164	BV650	334116	Biolegend
CD73	AD2	BUV805	748584	BD Biosciences
CD8	RPA-T8	BUV495	612943	BD Biosciences
KLRG1	13F12F2	PerCP-eFluor 710	46-9488-42	TermoFisher
live/dead		Zombie NIR	423106	Biolegend

Specificity	Clone	Fluorochrome	Ref.no	Manufacturer
PD1	MIH4	BUV395	745619	BD Biosciences
TIGIT	A15153G	BV605	372712	Biolegend
Vα7.2 TCR	OF-5A12	BV480	749493	BD Biosciences
Vδ1TCR	TS-1	FITC	TCR2055	TermoFisher
Vδ2 TCR	B6	Pacific Blue	331414	Biolegend
γδ TCR	11F2	BUV395	745681	BD Biosciences

***Table 5**. Monoclonal antibodies for the analysis of T cell populations of biopsies of CD patients inflamed and not inflamed tissue, CD patients in remission and healthy individuals.*

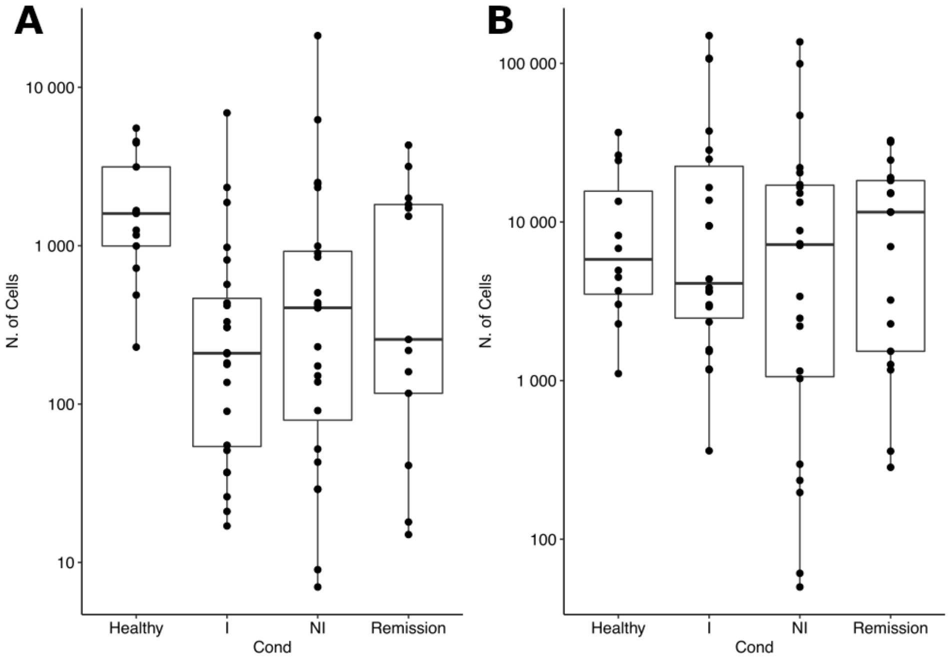


***Figure 21**. Number of cells per donor in 13 gut biopsies from individuals not affected by CD (Healthy), 24 biopsies of inflamed gut tissue of CD patients (I), 24 biopsies of non-inflamed gut tissue of CD patients (NI) and 17 gut biopsies were collected from CD patients in remission (Remission). (**A**) Number of Vδ1 TCR⁺ cells per donor. (**B**) Number of CD4⁺ cells per donor.*

2.1.2. CD4$^+$ T cells clustering

The clustering for the subset of CD4$^+$ cells was executed using the R implementation of Phenograph [270] based on 9 markers, among which markers relevant for CD4$^+$ T cells functional maturation (CD45RA, CD45R0, CCR7), activation (CD38, CD25) or markers that were evaluated as having potential for relevant information in analyses preceding the clustering (CD161, CD73, CD39, CD71). We defined 10 clusters that were enriched in healthy donors compared to every other condition (**Figure 22A**). The maturation markers CD45RA, CD45R0 and CCR7 place most of these clusters (6 out of 10) between the EM and the TEMRA stages identified by co-expression of CD45RA and CD45R0, with 2 clusters showing an EM pattern with only CD45R0 being expressed and the final cluster having a Naïve phenotype co-expressing CD45RA and CCR7 but not CD45R0. For inflamed enriched clusters, we found 8 clusters enriched only compared to healthy, 1 enriched compared to healthy and remission, 2 enriched compared to healthy and not inflamed tissue and 3 enriched compared to all 3 conditions other than inflamed. Among these clusters the main maturation marker pattern that was represented in the healthy clusters with CD45RA and CD45R0 and no CCR7 was completely absent, and aside from 2 clusters showing the pattern of Naïve T cells however lacking in CD73 compared to the Naïve cluster found in the healthy-enriched subset of cells, the majority of clusters was CD45R0$^+$. Of these, 6 clusters showed the hallmark of CM T cells (CD45R0 and CCR7) and 6 clusters shared the EM pattern already found in 2 clusters in the healthy-enriched category, however they differed from the healthy-enriched cases by being all CD73$^-$ and mostly CD71$^+$ while both healthy-enriched clusters presented the opposite pattern. Focusing on the 6 clusters enriched in inflamed compared to more than just healthy, one of them belonged to the CM group of clusters, in particular presenting the lack of other positive markers outside of the ones defining it as CM, while all the remaining ones were EM clusters and were CD71$^+$ (**Figure 22B**).

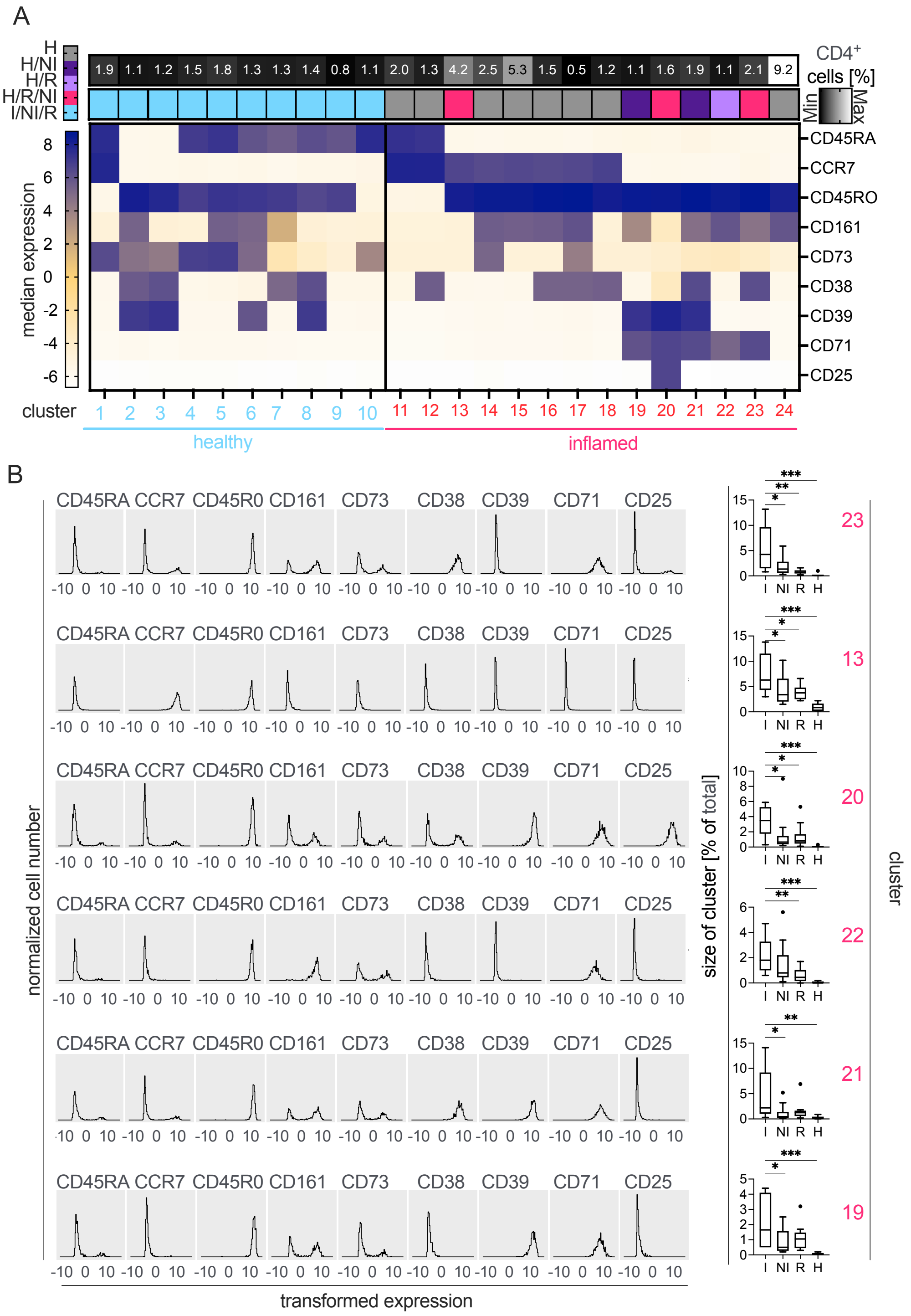


***Figure 22**. Representative iteration of CD4⁺ T cells clustering based on the hyperbolic arcsine transformed expression of 9 markers: CD45RA, CCR7, CD45R0, CD161, CD73, CD38, CD39, CD71, CD25 (I – Inflamed, NI – Not inflamed, R – Remission, H – Healthy). (**A**) Heatmap of the median expression of the 9 clustering markers for each cluster categorized as enriched in healthy (left,*

*cluster numbers in light blue) or inflamed (right, cluster numbers in red) in all 5 clustering iterations. On top of the heatmap are reported the size of clusters as percentage of all CD4⁺ cells and against which conditions the cluster is significantly enriched in healthy for the healthy enriched cluster and in inflamed for the inflamed enriched clusters. (**B**) Histograms of the 9 clustering markers hyperbolic arcsine transformed expression and boxplots showing the fraction of each condition for each cluster enriched in inflamed compared to every other condition. Figure from Devan et al., under review.*

2.1.3. Vδ1 TCR⁺ T cells clustering

The clustering for the subset of Vδ1 TCR⁺ cells was built using 6 markers relevant for Vδ1 TCR⁺ T cells functional maturation (CD45RA, CD45R0, CCR7, CD27, KLRG1, CD57) (**Figure 23**). We defined 2 clusters enriched in healthy individuals compared to the inflamed tissue individuals with active CD, with one cluster being also different in the not inflamed tissue of CD patients. The cluster that was different in both inflamed and not inflamed tissue of CD patients showed a maturation marker pattern placing it into the EM stage and representing 45.1% of all Vδ1 TCR⁺ T cells, while the other cluster covers approximately 4% of cells sharing a CM phenotype. Both clusters showed high levels of both CD38 and CD39 in combination with CD45R0 and absence of CD45RA. In the 7 clusters, enriched in inflamed tissue, this pattern was never present. Of these 7 clusters, 5 were significantly different in inflamed compared not only to healthy individual tissues, but also compared to not inflamed tissue of CD patients and to patients in remission tissue. Most of these clusters (6 out of 7) were CD45RA⁺ and (5 out of 7) KLRG1⁺, which was absent in the healthy enriched clusters, leaning towards the more terminally differentiated stage of TEMRA cells.

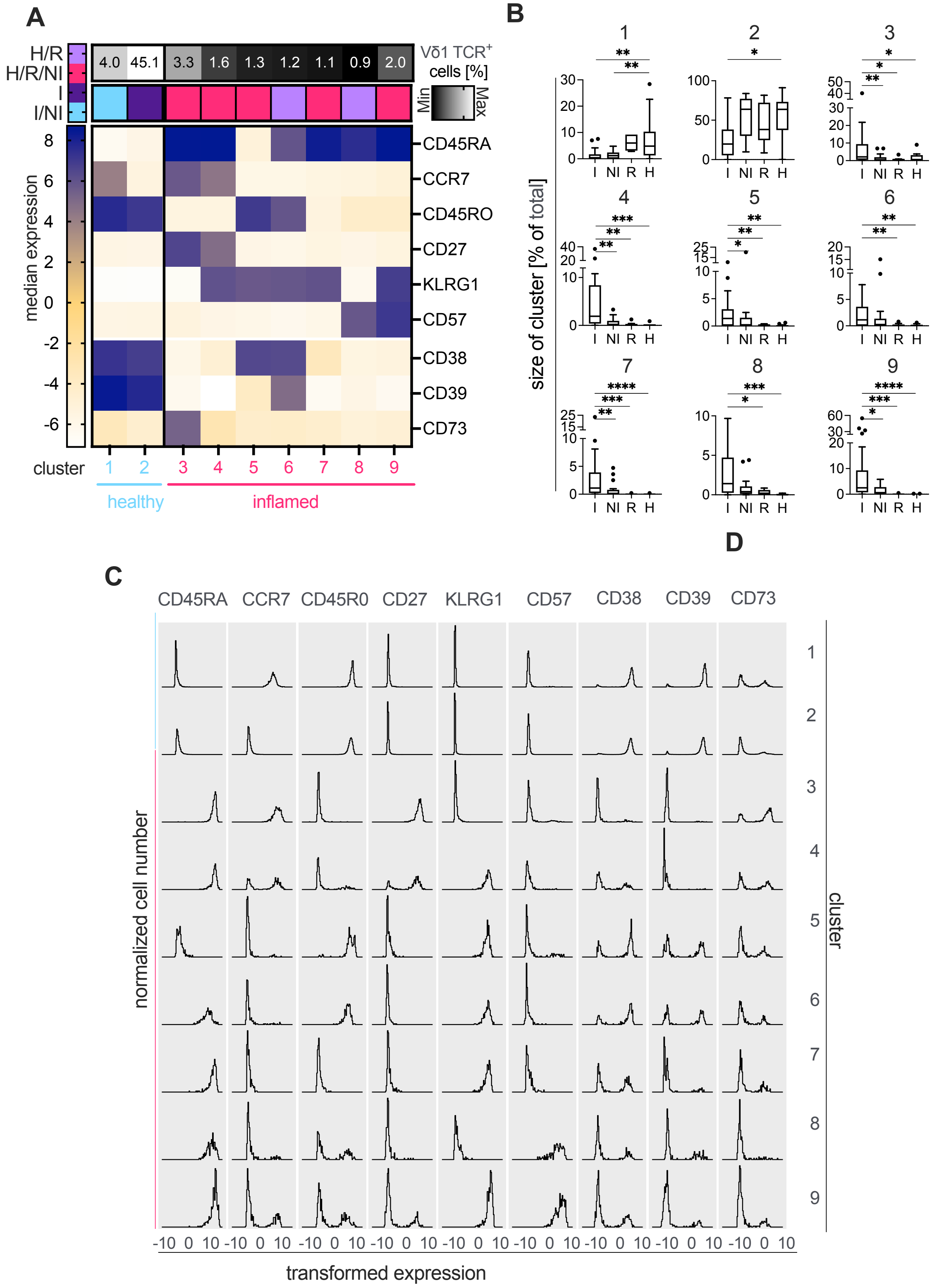


***Figure 23**. Representative iteration of Vδ1 TCR ⁺ T cells clustering based on the hyperbolic arcsine transformed expression of 6 markers: CD45RA, CCR7, CD45R0, CD27, KLRG1, CD57,*

*CD38, CD39, CD73 (I – Inflamed, NI – Not inflamed, R – Remission, H – Healthy). (**A**) Heatmap of the median expression of the 6 clustering markers and 3 more markers of interest for each cluster categorized as enriched in healthy (left, cluster numbers in light blue) or inflamed (right, cluster numbers in red) in all 5 clustering iterations. On top of the heatmap are reported the size of clusters as percentage of all Vδ1 TCR^{+} cells and against which conditions then cluster is significantly enriched in healthy for the healthy enriched cluster and in inflamed for the inflamed enriched clusters. (**B**) Boxplots showing the fraction of each condition for each cluster enriched in either healthy or inflamed. (**C**) Histograms of the 9 clustering markers hyperbolic arcsine transformed expression for each cluster enriched in either healthy or inflamed. Figure from Devan et al., under review.*

2.1.4. PBMC characterization and selection

We also checked if the gut homing potential of circulating CD4^{+} or Vδ1 TCR^{+} T cells was different between individuals affected by CD and healthy individuals. As we did for the biopsies, we performed our analysis by Flow Cytometry using a panel of markers for T cells (**Table 6**) (see below for a more detailed description of the markers used for clustering analysis). We utilized PBMCs from 15 healthy individuals (H) and 20 CD patients (I). The cells amount was, as expected, on average much larger than the that obtained in biopsies (**Figure 24**). Thus, we decided to use 1200 cells per donor in the Vδ1 TCR^{+} T cells clustering and 5000 cells per donor in the CD4^{+} T cells clustering. We also decided to increase the number of iterations to 50 for the Vδ1 TCR^{+} T cells and 30 for the CD4^{+} T cells to better cover the large number of available cells.

Specificity	Clone	Fluorochrome	Ref.no	Manufacturer
CCR7	G043H7	Spark NIR 685	353258	Biolegend
CCR9	L053E8	APC	358908	Biolegend
CD103	Ber-ACT8	APC-Cy7	350228	Biolegend
CD127	eBioRDR5	PE-CY5.5	35-1278-42	TermoFisher
CD14	63D3	PE-Fire640	367154	Biolegend
CD155	SKII.4	PE- Dazzle 594	337616	Biolegend

Specificity	Clone	Fluorochrome	Ref.no	Manufacturer
CD161	DX12	BUV661	750382	BD Biosciences
CD162	KPL-1	PE-CY7	328816	Biolegend
CD19	HIB19	Pe-Fire 640	302274	Biolegend
CD226	11A8	BV510	338330	Biolegend
CD25	CD25-3G10	PE-AF700	MHCD2524	TermoFisher
CD3	UCHT1	AF700	300424	Biolegend
CD357	108-17	BV421	371208	Biolegend
CD38	HIT2	APC-Fire810	303550	Biolegend
CD39	A1	BV711	328228	Biolegend
CD4	SK3	Spark blue	344656	Biolegend
CD45RA	HI-100	Spark NIR 685	304168	Biolegend
CD45R0	UCHL1	BV750	304262	Biolegend
CD49d	9F10	BV785	304344	Biolegend
CD49e	11A1	BUV737	741849	BD Biosciences
CD56	NCAM16.2	BUV563	612928	BD Biosciences
CD57	HNK-1	PerCP-Cy5.5	359622	Biolegend
CD71	CY164	BV650	334116	Biolegend
CD73	AD2	BUV805	748584	BD Biosciences
CD8	RPA-T8	BUV495	612943	BD Biosciences
Integrin β7	FIB504 FIB27	PE	321204	Biolegend
KLRG1	13F12F2	PerCP-eFluor710	46-9488-42	TermoFisher
live/dead		Zombie NIR	423106	Biolegend
TIGIT	A15153G	BV605	372712	Biolegend
Vα7.2 TCR	OF-5A12	BV480	749493	BD Biosciences

Specificity	Clone	Fluorochrome	Ref.no	Manufacturer
Vδ1TCR	TS-1	FITC	TCR2055	TermoFisher
Vδ2 TCR	B6	Pacific Blue	331414	Biolegend
γδ TCR	11F2	BUV395	745681	BD Biosciences

***Table 6**. Monoclonal antibodies for the analysis of T cell populations of PBMCs of CD patients and healthy individuals.*

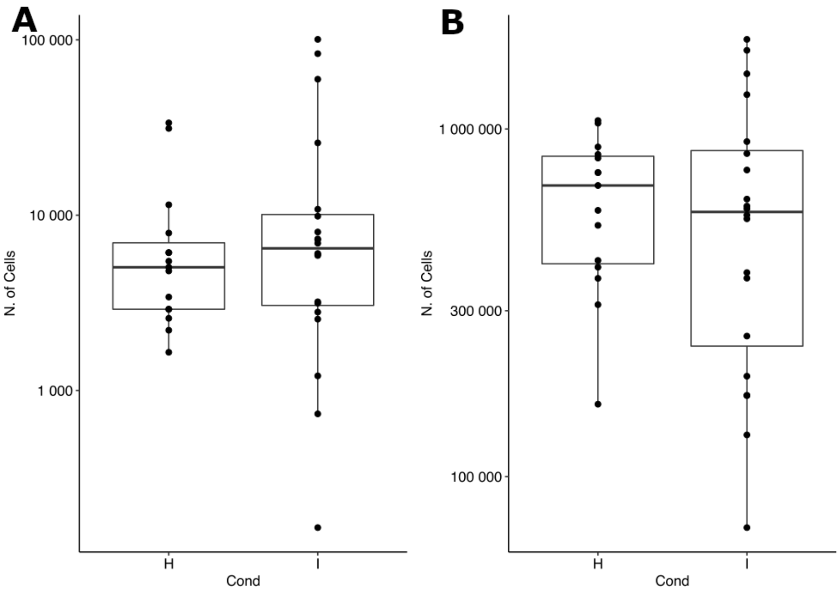


***Figure 24**. Number of cells per donor in PBMCs of 15 healthy donors (H) and 20 CD patients (I). (**A**) Number of Vδ1 TCR$^+$ cells per donor. (**B**) Number of CD4$^+$ cells per donor.*

2.1.5. PBMC CD4$^+$ and Vδ1 TCR$^+$ T cells clustering

For both clustering instances for the PBMCs, the same panel of gut homing markers was used as starting point: CD71, CD49e, CD49d, β7 integrin, CCR9, CD103. In the CD4$^+$ subset only one cluster was enriched in inflamed compared to healthy and it showed expression of CD71, CD49e, CD49d and β7 integrin (**Figure 25A-B**, **D**), suggesting gut homing potential. Looking at other markers expressed not used for the clustering process (**Figure 25C**), these cells are mainly CD45R0$^+$ and KLRG1$^-$, with some of the cells presenting CCR7 and some not presenting it, placing them between CM and EM. In the Vδ1 TCR$^+$ subset we identified 3 clusters, all showing expression of CD49e and CD49d, with only cluster 1 showing the same pattern of gut homing markers as the one found in CD4$^+$ (**Figure 25E-F**, **H**), leading to the consideration that they are also gut homing. For the other 2 clusters the marker expression doesn't give the same amount of certainty about their gut homing capabilities. Differently from the identified cluster in the CD4$^+$ subset, all these clusters present higher levels of KLRG1 and low levels of CD45R0, suggesting a TEMRA maturation stage for most of them (**Figure 25G**).

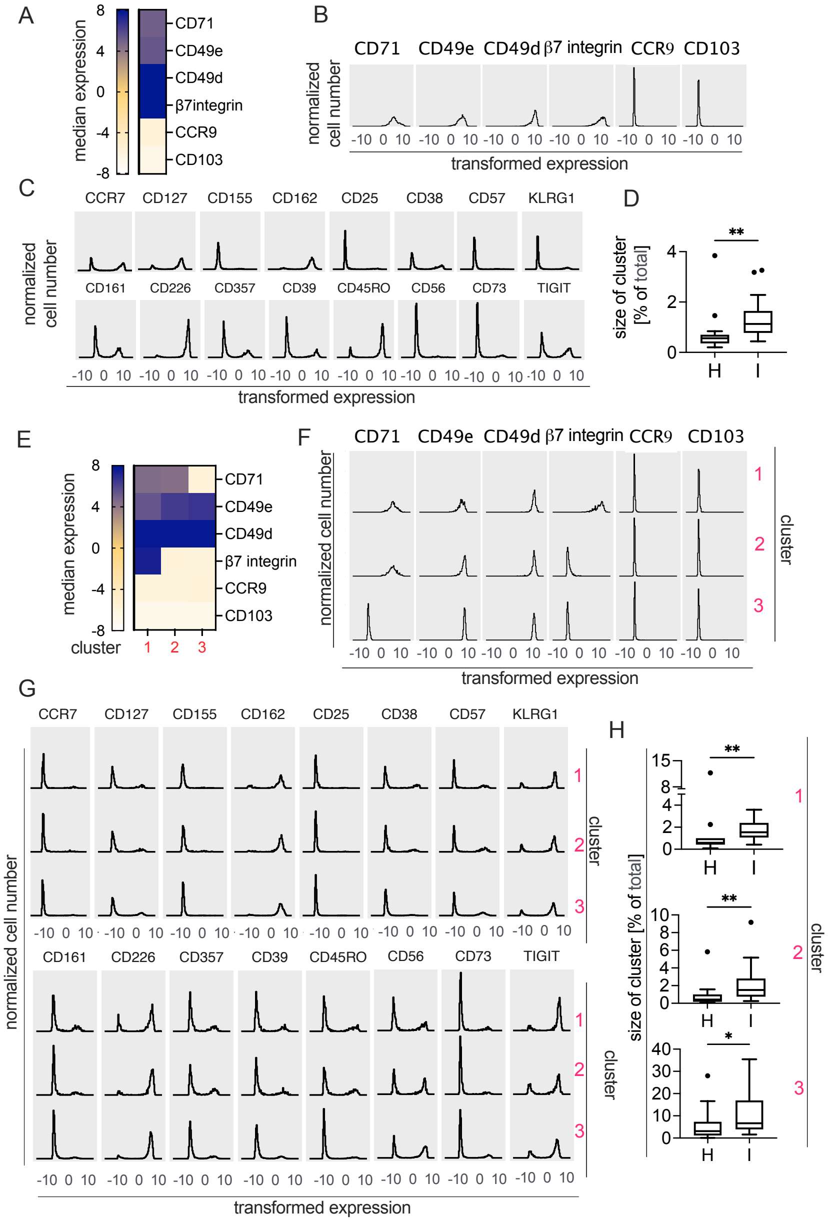


***Figure 25**. Representative iterations of CD4$^+$ T cells and Vδ1 TCR$^+$ T cells separated clusterings based on hyperbolic arcsine transformed expression of 6 markers (CD71, CD49e, CD49d, β7*

*integrin, CCR9, CD103) in 5000 CD4$^+$ cells from each of 15 Healthy and 20 Inflamed donors and 1200 Vδ1 TCR$^+$ cells from each of 15 Healthy and 18 Inflamed donors. (**A**) Heatmap of the median expression of the 6 clustering markers in the cluster categorized as enriched in inflamed in at least 95% of the 30 clustering iterations for the CD4$^+$ cells. (**B**) Histograms of the 6 markers used for clustering hyperbolic arcsine transformed expression for the cluster enriched in inflamed in the CD4$^+$ cells. (**C**) Histograms of 16 additional markers not used for clustering hyperbolic arcsine transformed expression for the cluster enriched in inflamed in the CD4$^+$ cells. (**D**) Boxplots showing the fraction of each condition for the cluster enriched in inflamed in the CD4$^+$ cells. (**E**) Heatmap of the median expression of the 6 clustering markers in the cluster categorized as enriched in inflamed in at least 95% of the 50 clustering iterations for the Vδ1 TCR$^+$ cells. (**F**) Histograms of the 6 markers used for clustering hyperbolic arcsine transformed expression for the clusters enriched in inflamed in the Vδ1 TCR$^+$ cells. (**G**) Histograms of 16 additional markers not used for clustering hyperbolic arcsine transformed expression for the clusters enriched in inflamed in the Vδ1 TCR$^+$ cells. (**H**) Boxplots showing the fraction of each condition for the clusters enriched in inflamed in the Vδ1 TCR$^+$ cells. Figure from Devan et al., under review.*

2.2. Conclusion

The purpose of the above-described work was to explore the differences between the gut T cell populations in individuals affected by Crohn's Disease compared to healthy individuals and individuals in remission. This analysis represents only a portion of the work done to identify markers of importance and populations of importance. In particular, the implemented workflow was aimed to manage cases in which few cells and high variability are characterizing the samples under analysis. Through the above-described workflow I identified differences in two subsets of T cells within the gut of Crohn's Disease patients compared to healthy. Thus, providing to the immunologists specific marker combinations to select cell subsets for additional experiments. The investigation of circulating T cells also showed populations of cells enriched in Crohn's Disease patients with gut homing potential in both CD4$^+$ and Vδ1 TCR$^+$ T cells.

2.3. Materials and Methods

2.3.1. Patient samples (from Devan et al., under review)

Patient samples were collected during regular colonoscopies at the Division of Gastroenterology and Hepatology at the University Hospital Basel (Basel, Switzerland). Crohn's disease patient samples and samples from non-CD donors, isolated from people suffering from diarrhoea predominant irritable bowel syndrome were collected. 8 biopsies were obtained from each patient from inflamed or not inflamed segments, kept in physiological solution for transfer to laboratory and immediately processed. The project was approved by ethical committee of the north-western part of Switzerland (EKNZ PB 2016-02242). All patients involved provided written informed consent.

2.3.2. Intestinal tissue processing (from Devan et al., under review)

Intestinal tissue biopsies were cut into pieces of ± 0.5 mm in diameter using sterile scalpel and incubated in 4 ml Gibco Roswell Park Memorial Institute 1640 media (Bioconcept) supplemented with 25mM HEPES (Cambrex), 200U collagenase IV (Sigma Aldrich), 0.5 mg/ml of DNAse I (Roche), 2,5 µg/ml of Amphotericin B (Life technologies), 5 µg/ml of Vancomycin (Teva Pharma Ag), 30 µg/ml of Piperacillin/Tazobactam (mass ratio 8:1; Sandoz) and 10 µg/ml of Ciprofloxacin (Bayer) at 37°C for 2 hours. Digested tissue was disrupted by pipetting, filtered through 40 µm nylon strainer, washed 2x in PBS (Bioconcept) and viably cryopreserved in solution containing 90% of heat inactivated foetal calf serum (FCS) and 10% of dimethyl sulfoxide (DMSO). Samples were stored at -70°C for up to 3 days and then transferred to liquid nitrogen for long term storage.

2.3.3. Blood samples processing (from Devan et al., under review)

Peripheral blood mononuclear cells (PBMCs) were isolated by density gradient centrifugation using Lymphoprep (Stemcell). Isolated PBMCs were washed 2x in phosphate buffer saline (PBS) (Bioconcept) and immediately used for experiments or viably cryopreserved in solution containing 90% of heat inactivated fetal calf serum (FCS) and 10% of dimethyl sulfoxide (DMSO). PBMCs were stored at -70°C for up to 3 days and then transferred to liquid nitrogen for long term storage.

2.3.4. Flow cytometry analysis of surface markers (from Devan et al., under review)

Cryopreserved cells were thawed and cultivated in RPMI media supplemented with 10 μg/ml of DNAse I, 10% heat inactivated FCS, 100U/ml Kanamycin, 2mM stable glutamine, 1% of minimal essential medium (MEM) nonessential amino acids and 1mM sodium pyruvate (all from Bioconcept) for 2 hours.

2.3.5. Biopsy data clustering

The initial analysis of flow cytometry data was performed using the FlowJo software (TreeStar). Positive cells for each marker were defined based on the combination of fluorescence minus one controls and isotype controls performed on cells pooled from 5 intestinal biopsies, except markers with bimodal expression and clearly separated populations (CD3, CD19, CD14, Vδ1 TCR, Vδ2 TCR, Vα7.2 TCR) in which case a threshold was defined by an expert. Data pre-gated as $CD4^+$ and Vδ1 TCR^+ was imported in R, the expression was centered for each marker around the threshold values previously defined by subtracting the threshold value from all data points, and we performed hyperbolic arcsine transformation. Clustering was performed on the hyperbolic arcsine transformed data for the $CD4^+$ and Vδ1 TCR cell populations separately, using the R implementation of Phenograph [270] using 9 markers for the $CD4^+$ subset (CD45RA, CCR7, CD45R0, CD161, CD73, CD38, CD39, CD71, CD25) and 6 markers for the Vδ1 TCR^+ subset (CD45RA, CCR7, CD45R0, CD27, KLRG1, CD57). The clustering was performed 5 times for each subset and clusters were classified as significantly enriched by comparing the distributions of the fractions of cells for each donor in each condition through a one sided Mann–Whitney U test [271], performing multiple testing correction through the Benjamini-Hochberg method [272].

2.3.6. PBMC data clustering

The analysis process followed the same steps as the biopsy data clustering, however in both $CD4^+$ and Vδ1 TCR^+ T cell clustering processes we used the same 6 markers (CD71, CD49e, CD49d, β7 integrin, CCR9, CD103) and the number of iterations was increased to 50 for Vδ1 TCR^+ T cells and 30 for $CD4^+$ T cells. The clusters of interest definition was also relaxed by taking as enriched all clusters that after multiple testing correction showed an adjusted p-value lower than 0.1 in at least 95% of iterations.

3. Self-Reactive MAITs

3.1. Results

3.1.1. Donor sequence numbers

Researchers from the Basel University group, where I spent the last part of my PhD, discovered that some MAIT cells present sterile reactivity to MR1 tetramers, meaning reactivity to MR1 without the presence of the bacterial antigen. We conducted bulk TCR sequencing of MAIT cells selected from PBMCs of 4 blood bank donors. Specifically, we analyzed MAIT cells lacking (CTV+) or presenting (CTV-) sterile reactivity to MR1 tetramers to define what differentiated them at the TCR level. Due to the semi-invariant nature of MAIT α chains and the impossibility to grab information of paired α and β chain, we decide to focus our analysis on β chain CDR3 sequences. After the alignment and clonotype calling with MiXCR and the filtering of the raw reads, the extracted TCR sequences for all 4 donors showed that only donor 1 (**Table 7**) had a sufficient number of unique β chain CDR3 sequences for further analysis. Thus, all subsequent steps were carried out on this single donor.

Donor	β chain CTV-	β chain CTV+	α chain CTV-	α chains CTV+
1	566	2862	209	640
2	69	2090	50	509
3	75	1612	40	413
4	9	215	20	53

Table 7*. Number of unique TCR sequences for each donor in the two different conditions after all the filtering steps for both α chains and β chains.*

3.1.2. Broad differential analysis between CTV- and CTV+ β chain TCRs

To identify the elements differentiating CTV- β chains from CTV+ β chains, we compared the counts of TCR sequences carrying each TRBV gene by family and single genes. We discovered that CTV- β chains presented a significantly higher amount of TCRs with TRBV genes belonging to the TRBV6 family (**Figure 26A**), with the difference being driven by the TRBV6-3 and TRBV6-6 (**Figure 26B**) genes. We also compared the counts of sequences carrying each TRBJ and TRBD genes between the two conditions, but no differences were found for these two features (**Figure 26C**, **D**). Furthermore, we determined that there wasn't any significant difference in the CDR3 lengths (**Figure 26E**).

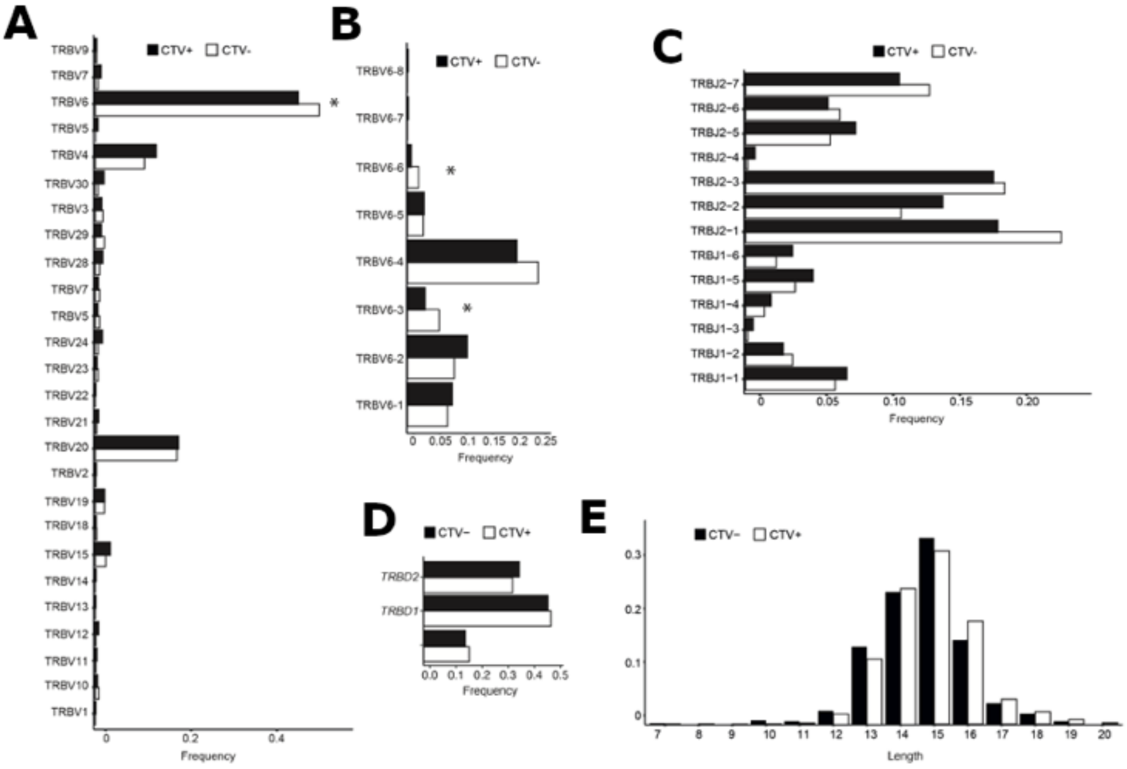


Figure 26*. Differential evaluation between CTV- and CTV+. (****A****) TRBV gene usage of CTV+ or CTV- MAIT cells. (****B****) TRBV6 gene usage of CTV+ or CTV- MAIT cells. (****C****) TRBJ gene usage of CTV+ and CTV- MAIT cells. (****D****) TRBD gene usage of CTV+ and CTV- MAIT cells. (****E****) Distribution of CDR3 lengths within either CTV+ or CTV- populations. Figure from Chancellor et al., under review.*

Unfortunately, bulk sequencing did not allow to estimate the clonality of the two populations.

3.1.3. E8 TCR and the E8-like motif

Motifs identification was difficult to execute, since basic TCR features provided us with only one significantly different feature between CTV$^-$ and CTV$^+$ and the overall number of sequences was quite small. Thus, we decided to perform an analysis derived by the work performed by a group of collaborators at Immunocore Ltd. They managed to isolate a sterile-reacting TCR called E8 through a phage library generated using one canonical MAIT TCR α chain coupled with random TCR β chains. E8 is a canonical MAIT TCR expressing TRAV1-2 and TRAJ33, coupled with a chimera of the genes TRBV6-1 and TRBV6-5. To understand the molecular basis for this reactivity, the crystal structure of E8 in complex with MR1 loaded with 5-OP-RU was solved and aligned with the crystal structure of a classical MAIT TCR lacking sterile reactivity called AF-7 (**Figure 27A**). The two TCRs positioned similarly, with very few differences in the CDR loops. In particular, in both cases the Tyrosine in position 95 of the α chain made contact with the antigen directly (**Figure 27B**), as expected in classical MAIT α chains. However, one crucial difference was found between the two interaction networks, with an Arginine in position 96 of the E8 β chain making salt bridges with two MR1 residues, E76 and E149 (**Figure 27C**). Crystal structures of E8 in complex with empty MR1 and with MR1 carrying other ligands maintained a very similar network of interactions. Molecular dynamics simulations and binding free energy simulations confirmed the importance of the CDR3 β chain loop in acting as an anchor locking the positioning of the E8 TCR independent of being bound to a ligand. By looking at the features of the E8 β chain and at the crystal structure and the network of interactions, we determined an E8-like motif composed of a TRBV gene belonging to the TRBV6 family, an Arginine in position 96 of the β chain and a CDR3β of length 13, to ensure the correct positioning of the Arginine, as differences in CDR3 length would impact the conformation of the loop.

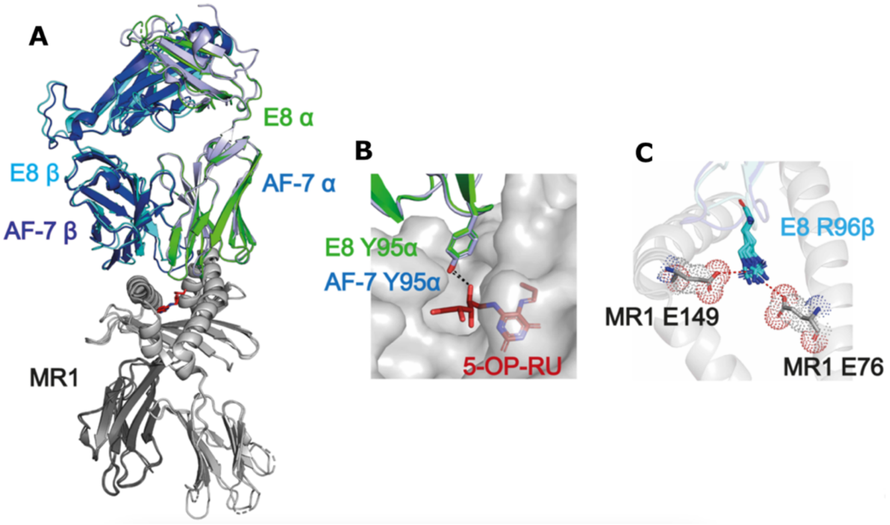


Figure 27*. Comparison between the sterile reacting TCR E8 and a classical MAIT TCR AF-7. (****A****) Crystal structure comparison between E8 and AF-7 when bound to MR1-5-OP-RU. (****B****) Y95α in the CDR3 of the α chain interacting with the antigen in both E8 and AF-7. (****C****) R96β in the β chain of E8 being pinched between two MR1 residues. Image from Chancellor et al., under review.*

3.1.4. E8-like motif detection in CTV- versus CTV+

After identifying the crucial features for the E8 TCR sterile reactivity, we checked if the features were significantly different in the CTV- population compared to the CTV+ population. To conduct this analysis, we first looked at the features singularly, therefore looking if an Arginine was present in position 96 of the CDR3 more often in CTV- compared to CTV+, and this was not the case (**Figure 28A**). The other two singular features we had already checked in the broad analysis showed that the frequency of the TRBV6 gene resulted to be significantly different between the two populations (**Figure 28B**) as instead the CDR3 of length 13 was not (**Figure 28C**). We then looked if the combination of two of the features together was different in the two populations (Arginine in position 6 combined with CDR3 of length 13, **Figure 28D**; Arginine in position 6 combined with a TRBV6 gene, **Figure 28E**; TRBV6 gene combined with CDR3 of length 13, **Figure 28F**), but none of them resulted to be differentially abundant. Overall, most of the singular features and all the paired combinations didn't result different between the two conditions, however when we examined all the three features being present at the same time in the full E8-like motif, we found that the frequency was higher in CTV- compared to CTV+ (**Figure 28G**).

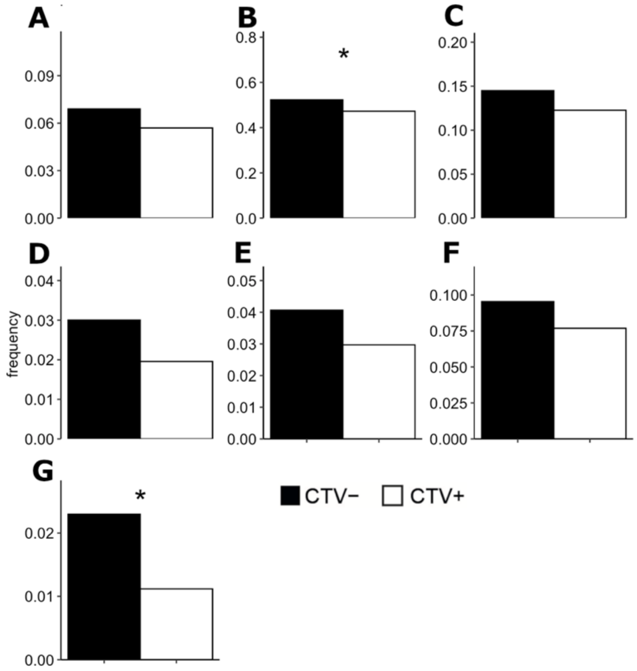


***Figure 28**. Differential evaluation between CTV- and CTV+ for the E8-like motif features. (**A**) Arginine in position 96. (**B**) TRBV6 gene. (**C**) CDR3 of length 13. (**D**) Arginine in position 96 and CDR3 of length 13. (**E**) Arginine in position 96 and TRBV6 gene. (**F**) TRBV6 gene and CDR3 of length 13. (**G**) Full E8-like motif.*

3.1.5. Functional validation

From the E8-like motif search, we found 13 TCRs carrying it in the CTV- population (**Table 8**). TCRdist [273] was used to select the CDR3β with the highest similarity to the E8 CDR3β (β chain 9 in **Table 8**). Functional validation was conducted by co-expressing this TCR β chain with a canonical MAIT α chain in J.RT3-T3.5 cells, with this TCR being called 393. Expression of this TCR produced both canonical reactivity to 5-OP-RU loaded MR1 and sterile reactivity to empty wild-type MR1 and K43A MR1 expressed in A375b cells (**Figure 29**). This reactivity was blocked by anti-MR1mAb.

β chain	CDR3	TRBV gene	TRBD gene	TRBJ gene
1	CASSNRAQSGQYF	TRBV6-3	TRBD2	TRBJ2-7
2	CASNDRESYEQYF	TRBV6-1		TRBJ2-7
3	CASIDRENSPLHF	TRBV6-2	TRBD1	TRBJ1-6
4	CASSDRGTGELFF	TRBV6-4	TRBD2	TRBJ2-2
5	CASSERGTDTQYF	TRBV6-4	TRBD2	TRBJ2-3
6	CASTKRDTDTQYF	TRBV6-2	TRBD2	TRBJ2-3
7	CASSDRATDTQYF	TRBV6-4	TRBD2	TRBJ2-3
8	CATRDRDTGELFF	TRBV6-4		TRBJ2-2
9	CASSDREADTQYF	TRBV6-4		TRBJ2-3
10	CASSDRETGEQFF	TRBV6-4	TRBD2	TRBJ2-1
11	CASSPREVETQYF	TRBV6-6		TRBJ2-5
12	CASSPRETDTQYF	TRBV6-5	TRBD1	TRBJ2-3
13	CASSDRDTGELFF	TRBV6-4		TRBJ2-2

Table 8. *List of the 13 sterile reacting TCR β chains carrying the E8-like motif.*

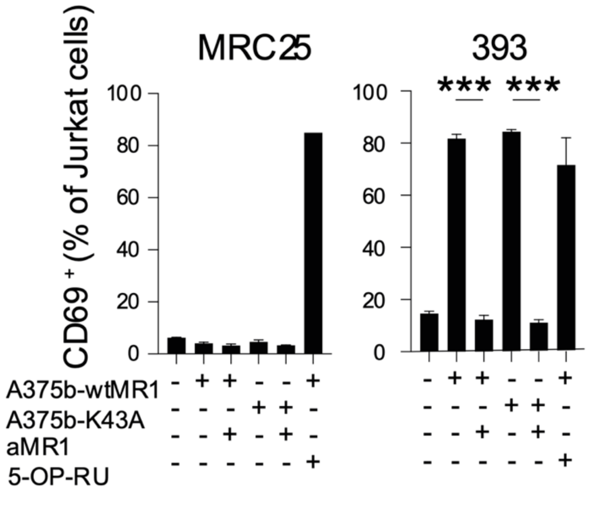


Figure 29. *Functional validation of the ability of the 393 TCR to confer J.RT3-T3.5 cells sterile reactivity to A375b-wtMR1 and A375b-K43A, blocked by anti-MR1 mAb. Activation was evaluated as percentage of cells positive for CD69 and compared to the control MRC25 β chain being used instead of the 393 β chain carrying the E8-like motif. THP-1 cells pulsed with 5-OP-*

RU were used as positive control. Data representative of 3 individual experiments each performed in triplicate. Figure adapted from Chancellor et al., under review.

3.1.6. Motif search in MAIT dataset

For an indication of the frequency of the motif in circulating MAIT cells, we explored a dataset of TCR β chains from PBMCs of 7 healthy donors containing TRAV1-2⁺/TRBV6⁺ both MAIT (CD161⁺) and non-MAIT (CD161⁻). We found that the motif was expressed with a higher frequency in MAIT cells compared to non-MAIT TRAV1-2⁺/TRBV6⁺ cells (**Figure 30**).

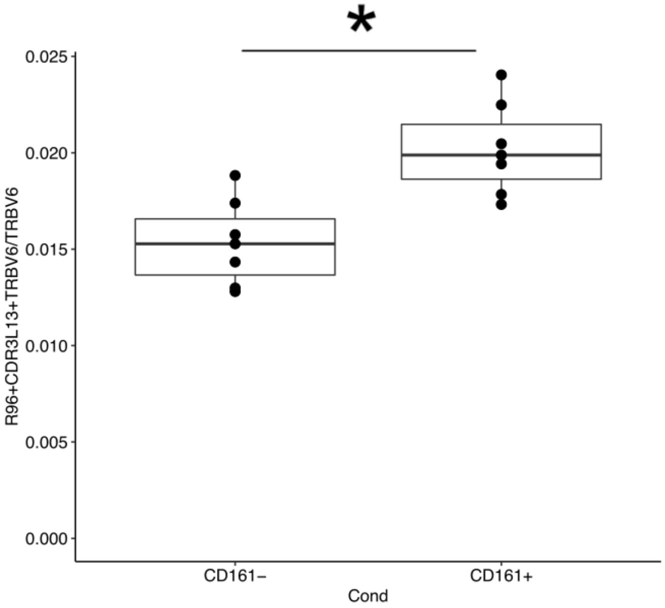


***Figure 30**. Frequency of the E8-like motif in β chains within ex vivo MAIT cells (TRAV1-2⁺/TRBV6⁺ and CD161⁺) compared to non-MAIT (TRAV1-2⁺/TRBV6⁺ and CD161⁻) in PBMCs of 7 healthy donors. Figure adapted from Chancellor et al., pre-print.*

3.2. Conclusions

The work presented in this section is just a portion of a larger project investigating a rare subset of MAIT cells presenting sterile recognition of MR1. The aim of my participation in the project was to analyze the TCR sequencing data, searching for elements differentiating the sterile reacting subset of MAIT cells to the subset of MAIT cells not presenting this sterile reactivity to MR1. The broad characterization of the sterile reacting TCRs didn't lead us to find specific strong markers for explaining this sterile reacting capability and the number of events at our disposal was too low for machine learning-driven motif detection, so we decided to proceed in a more supervised manner, utilizing a crystal produced by collaborators of our laboratory at Immunocore Ltd. Through this crystal we could define some structural requirements for one method of sterile recognition that were enriched in our sterile reacting population, however this was only a small number, leading to think that this is only one possible mechanism for this sterile reactivity. Furthermore, we were able to detect the motif in circulating MAIT cells of healthy individuals, which opens new questions about what their role could be in homeostasis, cancer surveillance or in inflammatory and autoimmune diseases.

3.3. Materials and Methods

3.3.1. Primary T cells preparation for TCR sequencing (from Chancellor et al., under review)

Primary cell lines and clones used in the study were isolated from PBMCs obtained using Lymphoprep (Stemcell Technologies) from blood of 4 blood bank donors and were maintained in culture as described in [274]. MAIT cell lines were generated by magnetic bead enrichment using biotinylated anti-Vα7.2 mAb (Clone 3C10, Biolegend) or specific expansion using 5-OP-RU. Enriched MAIT cells were prelabelled with Cell Trace Violet according to manufacturer instructions and then cultured with irradiated A375b-wtMR1 cells for the indicated number of days in a 1:1 ratio. Human rIL-2 (5 IU/ml, Peprotech) was added at 5 days and thereafter every two days. Cells were washed and rechallenged as indicated (ratio 2:1) in the presence or absence of purified anti-MR1 mAb (20 μg/ml, Ultra-LEAF™ Purified Clone 26.5, Biolegend). From these lines, self-reactive MAIT clones were derived by limiting dilution in the presence of PHA (1 μg/ml), human rIL-2 (100 μg/ml) and irradiate PBMC (5×10^5 cells/ml), and screened for reactivity toward A375b-wtMR1.

3.3.2. Flow cytometry (from Chancellor et al., under review)

When staining with MR1 tetramers (20 μg/ml) or anti-human Vα7.2 (2.5μg/ml Clone 3C10, Biolegend), the cells were pre-treated for 30 min at 37°C with 50 nM dasatinib (Sigma) in

PBS. All mAb for staining were titrated on appropriate cells before use. Tetramers were added first for 20 min at RT and anti-human mAb were added for a further 20 min in PBS with dasatinib: mAb specific for CD3 (Clone UCHT1), CD4 (Clone OKT4), CD8 (Clone RPA-T8), CD161 (Clone HP-3G19) and for activation markers CD137 (Clone 4B4-1), CD69 (Clone FN50), CD25 (Clone BC96), ICOS (Clone DX29), all from Biolegend. DAPI was used to exclude dead cells. Doublets were excluded by FSC-A, FSC-W, SSC-A and SSC-H.

3.3.3. TCR sequencing (from Chancellor et al., under review)

MAIT T cells ($CD3^+$ $CD161^+$ TRAV1-2^+ $V\delta2^-$) were sorted by flow cytometry (BD FACS Aria) and expanded on PHA, IL-2 and irradiated allogenic PBMCs to establish a T cell line from which clones were subsequently generated by limiting dilution. Individual clones were assessed for CD161, TRAV1-2 and CD137 expression by flow cytometry following overnight co-culture with 5-OP-RU loaded THP-1 cells. Positive clones were selected for TCR genes sequencing. Briefly, this involves first-strand cDNA generation and universal amplification using SmartSeq2 chemistry [199], followed by targeted amplification of TCR chains and MiSeq Next Generation Sequencing. Sequencing was paired-end and performed in bulk with no Unique Molecular Identifiers (UMIs). For all reads 3 random nucleotides and barcodes composed of 5 nucleotides unique for each donor and CTV positivity or negativity combination were added on both paired-end reads. Two different sets of 4 nucleotides were added after the 5 nucleotide barcodes on reads starting in the constant region of the TCR to represent the alpha and beta chains. The data was provided in one R1 and one R2 FASTQ file containing the information for all donors, both conditions and both chains, without fixed forward and reverse read between R1 and R2.

3.3.4. Raw data pre-processing

We used FASTX-Toolkit (version 0.0.14) [275] to remove the 3 random nucleotides at the start of the reads, then we employed Cutadapt (version 3.5) [276] to demultiplex the reads according to their 5 + 4 nucleotides barcodes with no allowed mismatches or indels. Because the reads were not orientated so that the forward reads would all be in either the R1 or the R2 file, we applied Cutadapt looking for the barcodes only in the R1 file, extracting from it all reads starting from the constant region of the TCR and the associated mate reads from the R2 file. Then we ran Cutadapt again looking only at the R2 file and inverted the read order, so that all reads would have the same orientation.

3.3.5. TCR alignment and data filtering

Alignment of the TCR alpha and beta chain data was performed utilizing the MiXCR (version 3.0.13) [277]. We selected MiXCR due to its ability to perform correction for both amplification and sequencing errors in the clonotype assignment. The resulting MiXCR tables were imported in R and processed through custom scripts. Because of the lack of UMIs and with the sequencing methodology being bulk, all sequences carrying the same CDR3 sequence were merged into single clonotypes to diminish the possibility of false sequences being added to the data. CDR3 sequences below 7 amino acids or above 20 amino acids in length were excluded from our dataset. Sequences that had only one read corroborating them were removed from the dataset, unless they were shared between CTV+ and CTV- population of a given donor, in which case they were kept in the dataset if the sequence presented at least two corroborating reads in one of the two conditions. Then all sequences that were shared between CTV+ and CTV- populations were removed from the CTV+ population and kept only in the negative one as they were capable of proliferation in sterile conditions.

3.3.6. Differential analysis between CTV- and CTV+

Differences between CTV+ and CTV- populations were evaluated through custom R scripts by comparing the counts of clonotypes with or without given features between the two conditions through Fisher Exact Tests [278], performing multiple testing correction through the Benjamini-Hochberg method [272].

3.3.7. Motif search in independent PBMCs dataset

The search for the E8-like motif in PBMCs was performed through R scripts. The comparison between distributions was performed using the Wilcoxon Signed-Rank Test [279].

3.3.8. Construct design, protein expression and purification (from Chancellor et al., under review)

Wild type MR1, MR1-K43A, B2M and TCR chains were cloned into the pGMT7 vector. TCR constructs were designed to include the variable and constant domains of both α and β chains with an engineered inter-chain disulfide bond as previously described [280]. The proteins were expressed in the BL21 (DE3) Rosetta pLysS strain (Novagen), refolded from inclusion bodies and purified as previously described [280, 281]. For SPR measurements a C-terminal AVI-tag was added to the wtMR1 and MR1-K43A constructs and biotinylated after purification using the Avidity Bir A Biotinylation kit, then purified again using a size exclusion column to remove the biotin and Bir A.

3.3.9. TCR gene transfer (from Chancellor et al., under review)

Total RNA was extracted from snap-frozen cell pellets from each clone. SMARTer RACE 5'/3' kit (Takara) was used for cDNA synthesis and generation of TCR transcripts. Functional TCRα and β chains were identified by sequencing and analysis using the ImMunoGeneTics information system (http://www.imgt.org). The TCRα and β sequences were either synthesized at Integrated DNA Technologies (TCR 393) or amplified from cDNA with gene specific primers (TCRs BC75B31, BC75B38, MRC25) containing cloning adaptors. In both cases the insert was cloned by In-Fusion HD (Takara) to a lentiviral vector for co-transfection of HEK 293 T LX cells. The endotoxin-free vectors were co-transfected together with the lentivirus packaging plasmids pMD2.G, pMDLg/pRRE and pRSV-REV (all from Addgene) to HEK 293 T LX cells with Metafectene PRO reagent from Biontex. Lentiviral supernatants of the corresponding TCRα and β sequences were combined and used to transduce J.RT3-T3.5 cells overnight. TCR-expressing J.RT3-T3.5 cells were sorted for CD3 expression before functional analysis.

3.3.10. Activation assay validation (from Chancellor et al., under review)

T cell clones (5×10^4 cells/well unless otherwise indicated) were co-cultured with indicated target cells (5×10^4 cells /well) in 130 µl total volume in triplicates for 18 h. In some experiments, anti-MR1 mAb (20 µg/ml, Ultra-LEAFTM Purified Clone 26.5, Biolegend) was added and incubated for 30 min prior to the addition of T cells. In other experiments, APCs were pulsed for 2 h at 37°C with indicated concentrations of Ags or freshly-prepared 5-OP-RU as described in [282]. J.RT3-T3.5 activation assays were performed in a 1:1 ratio with the indicated APC for 18 h. Cells were then either harvested and stained for surface CD69 upregulation or luciferase was measured using Bio-Glo (Promega).

References

1. Will, C.L. and R. Luhrmann, *Spliceosome structure and function.* Cold Spring Harb Perspect Biol, 2011. **3**(7).

2. Black, D.L., *Mechanisms of alternative pre-messenger RNA splicing.* Annu Rev Biochem, 2003. **72**: p. 291-336.

3. Jiang, W. and L. Chen, *Alternative splicing: Human disease and quantitative analysis from high-throughput sequencing.* Comput Struct Biotechnol J, 2021. **19**: p. 183-195.

4. Graveley, B.R., *Alternative splicing: increasing diversity in the proteomic world.* Trends Genet, 2001. **17**(2): p. 100-7.

5. Matlin, A.J., F. Clark, and C.W. Smith, *Understanding alternative splicing: towards a cellular code.* Nat Rev Mol Cell Biol, 2005. **6**(5): p. 386-98.

6. da Costa, P.J., J. Menezes, and L. Romao, *The role of alternative splicing coupled to nonsense-mediated mRNA decay in human disease.* Int J Biochem Cell Biol, 2017. **91**(Pt B): p. 168-175.

7. Long, J.C. and J.F. Caceres, *The SR protein family of splicing factors: master regulators of gene expression.* Biochem J, 2009. **417**(1): p. 15-27.

8. Martinez-Contreras, R., et al., *hnRNP proteins and splicing control.* Adv Exp Med Biol, 2007. **623**: p. 123-47.

9. Lewis, B.P., R.E. Green, and S.E. Brenner, *Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans.* Proc Natl Acad Sci U S A, 2003. **100**(1): p. 189-92.

10. Ge, Y. and B.T. Porse, *The functional consequences of intron retention: alternative splicing coupled to NMD as a regulator of gene expression.* Bioessays, 2014. **36**(3): p. 236-43.

11. Wong, J.J., et al., *Orchestrated intron retention regulates normal granulocyte differentiation.* Cell, 2013. **154**(3): p. 583-95.

12. Wang, E.T., et al., *Alternative isoform regulation in human tissue transcriptomes.* Nature, 2008. **456**(7221): p. 470-6.

13. Pan, Q., et al., *Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.* Nat Genet, 2008. **40**(12): p. 1413-5.

14. Pickrell, J.K., et al., *Noisy splicing drives mRNA isoform diversity in human cells.* PLoS Genet, 2010. **6**(12): p. e1001236.

15. Yeo, G., et al., *Variation in alternative splicing across human tissues.* Genome Biol, 2004. **5**(10): p. R74.

16. Castle, J.C., et al., *Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines.* Nat Genet, 2008. **40**(12): p. 1416-25.

17. Zhang, X., et al., *Cell-Type-Specific Alternative Splicing Governs Cell Fate in the Developing Cerebral Cortex.* Cell, 2016. **166**(5): p. 1147-1162 e15.

18. Li, H., et al., *SRSF10 regulates alternative splicing and is required for adipocyte differentiation.* Mol Cell Biol, 2014. **34**(12): p. 2198-207.

19. Pimentel, H., et al., *A dynamic alternative splicing program regulates gene expression during terminal erythropoiesis.* Nucleic Acids Res, 2014. **42**(6): p. 4031-42.

20. Hanoun, N., et al., *The SV2 variant of KLF6 is down-regulated in hepatocellular carcinoma and displays anti-proliferative and pro-apoptotic functions.* J Hepatol, 2010. **53**(5): p. 880-8.

21. Narla, G., et al., *Targeted inhibition of the KLF6 splice variant, KLF6 SV1, suppresses prostate cancer cell growth and spread.* Cancer Res, 2005. **65**(13): p. 5761-8.

22. Kelemen, O., et al., *Function of alternative splicing.* Gene, 2013. **514**(1): p. 1-30.

23. Tazi, J., N. Bakkour, and S. Stamm, *Alternative splicing and disease.* Biochim Biophys Acta, 2009. **1792**(1): p. 14-26.

24. Ward, A.J. and T.A. Cooper, *The pathobiology of splicing.* J Pathol, 2010. **220**(2): p. 152-63.

25. El Marabti, E. and I. Younis, *The Cancer Spliceome: Reprograming of Alternative Splicing in Cancer.* Front Mol Biosci, 2018. **5**: p. 80.

26. Mordes, D., et al., *Pre-mRNA splicing and retinitis pigmentosa.* Mol Vis, 2006. **12**: p. 1259-71.

27. Lehalle, D., et al., *A review of craniofacial disorders caused by spliceosomal defects.* Clin Genet, 2015. **88**(5): p. 405-15.

28. Armstrong, R.N., et al., *Splicing factor mutations in the myelodysplastic syndromes: target genes and therapeutic approaches.* Adv Biol Regul, 2018. **67**: p. 13-29.

29. Scotti, M.M. and M.S. Swanson, *RNA mis-splicing in disease.* Nat Rev Genet, 2016. **17**(1): p. 19-32.

30. Krawczak, M., et al., *Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing.* Hum Mutat, 2007. **28**(2): p. 150-8.

31. Birgens, H. and R. Ljung, *The thalassaemia syndromes.* Scand J Clin Lab Invest, 2007. **67**(1): p. 11-25.

32. De Sandre-Giovannoli, A. and N. Levy, *Altered splicing in prelamin A-associated premature aging phenotypes.* Prog Mol Subcell Biol, 2006. **44**: p. 199-232.

33. Zhang, J., et al., *A functional alternative splicing mutation in AIRE gene causes autoimmune polyendocrine syndrome type 1.* PLoS One, 2013. **8**(1): p. e53981.

34. Andreadis, A., *Misregulation of tau alternative splicing in neurodegeneration and dementia.* Prog Mol Subcell Biol, 2006. **44**: p. 89-107.

35. Fernandez-Nogales, M., et al., *Huntington's disease is a four-repeat tauopathy with tau nuclear rods.* Nat Med, 2014. **20**(8): p. 881-5.

36. Wolfe, M.S., *The role of tau in neurodegenerative diseases and its potential as a therapeutic target.* Scientifica (Cairo), 2012. **2012**: p. 796024.

37. Spillantini, M.G. and M. Goedert, *Tau protein pathology in neurodegenerative diseases.* Trends Neurosci, 1998. **21**(10): p. 428-33.

38. Klamt, B., et al., *Frasier syndrome is caused by defective alternative splicing of WT1 leading to an altered ratio of WT1 +/-KTS splice isoforms.* Hum Mol Genet, 1998. **7**(4): p. 709-14.

39. DiFeo, A., J.A. Martignetti, and G. Narla, *The role of KLF6 and its splice variants in cancer therapy.* Drug Resist Updat, 2009. **12**(1-2): p. 1-7.

40. Narla, G., et al., *A germline DNA polymorphism enhances alternative splicing of the KLF6 tumor suppressor gene and is associated with increased prostate cancer risk.* Cancer Res, 2005. **65**(4): p. 1213-22.

41. Narla, G., et al., *KLF6-SV1 overexpression accelerates human and mouse prostate cancer progression and metastasis.* J Clin Invest, 2008. **118**(8): p. 2711-21.

42. Miki, Y., et al., *A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1.* Science, 1994. **266**(5182): p. 66-71.

43. Hoffman, J.D., et al., *Implications of a novel cryptic splice site in the BRCA1 gene.* Am J Med Genet, 1998. **80**(2): p. 140-4.

44. Liu, H.X., et al., *A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes.* Nat Genet, 2001. **27**(1): p. 55-8.

45. Dvinge, H., et al., *RNA splicing factors as oncoproteins and tumour suppressors.* Nature Reviews Cancer, 2016. **16**(7): p. 413-430.

46. Fischer, D.C., et al., *Expression of splicing factors in human ovarian cancer.* Oncology Reports, 2004. **11**(5): p. 1085-1090.

47. Karni, R., et al., *The gene encoding the splicing factor SF2/ASF is a proto-oncogene.* Nature Structural & Molecular Biology, 2007. **14**(3): p. 185-193.

48. Ghigna, C., et al., *Altered expression of heterogeneous nuclear ribonucleoproteins and SR factors in human colon adenocarcinomas.* Cancer Research, 1998. **58**(24): p. 5818-5824.

49. Dvinge, H. and R.K. Bradley, *Widespread intron retention diversifies most cancer transcriptomes.* Genome Medicine, 2015. **7**.

50. Kole, R. and A.M. Krieg, *Exon skipping therapy for Duchenne muscular dystrophy.* Advanced Drug Delivery Reviews, 2015. **87**: p. 104-107.

51. Takeda, S., P.R. Clemens, and E.P. Hoffman, *Exon-Skipping in Duchenne Muscular Dystrophy.* J Neuromuscul Dis, 2021. **8**(s2): p. S343-S358.

52. Zhao, X., et al., *Pharmacokinetics, pharmacodynamics, and efficacy of a small-molecule SMN2 splicing modifier in mouse models of spinal muscular atrophy.* Hum Mol Genet, 2016. **25**(10): p. 1885-1899.

53. Naryshkin, N.A., et al., *Motor neuron disease. SMN2 splicing modifiers improve motor function and longevity in mice with spinal muscular atrophy.* Science, 2014. **345**(6197): p. 688-93.

54. Cooper, C.S., et al., *Molecular cloning of a new transforming gene from a chemically transformed human cell line.* Nature, 1984. **311**(5981): p. 29-33.

55. Park, M., et al., *Two rearranged MET alleles in MNNG-HOS cells reveal the orientation of MET on chromosome 7 to other markers tightly linked to the cystic fibrosis locus.* Proc Natl Acad Sci U S A, 1988. **85**(8): p. 2667-71.

56. Park, M., et al., *Mechanism of met oncogene activation.* Cell, 1986. **45**(6): p. 895-904.

57. Giordano, S., et al., *Tyrosine kinase receptor indistinguishable from the c-met protein.* Nature, 1989. **339**(6220): p. 155-6.

58. Organ, S.L. and M.S. Tsao, *An overview of the c-MET signaling pathway.* Ther Adv Med Oncol, 2011. **3**(1 Suppl): p. S7-S19.

59. Gentile, A., L. Trusolino, and P.M. Comoglio, *The Met tyrosine kinase receptor in development and cancer.* Cancer Metastasis Rev, 2008. **27**(1): p. 85-94.

60. Petrini, I., *Biology of MET: a double life between normal tissue repair and tumor progression.* Ann Transl Med, 2015. **3**(6): p. 82.

61. Nakamura, T., et al., *Molecular cloning and expression of human hepatocyte growth factor.* Nature, 1989. **342**(6248): p. 440-3.

62. Stoker, M., et al., *Scatter factor is a fibroblast-derived modulator of epithelial cell mobility.* Nature, 1987. **327**(6119): p. 239-42.

63. Weidner, K.M., et al., *Evidence for the identity of human scatter factor and human hepatocyte growth factor.* Proc Natl Acad Sci U S A, 1991. **88**(16): p. 7001-5.

64. Furge, K.A., Y.W. Zhang, and G.F. Vande Woude, *Met receptor tyrosine kinase: enhanced signaling through adapter proteins.* Oncogene, 2000. **19**(49): p. 5582-9.

65. Trusolino, L., A. Bertotti, and P.M. Comoglio, *MET signalling: principles and functions in development, organ regeneration and cancer.* Nat Rev Mol Cell Biol, 2010. **11**(12): p. 834-48.

66. Johnson, G.L. and R. Lapadat, *Mitogen-activated protein kinase pathways mediated by ERK, JNK, and p38 protein kinases.* Science, 2002. **298**(5600): p. 1911-2.

67. Ponzetto, C., et al., *A multifunctional docking site mediates signaling and transformation by the hepatocyte growth factor/scatter factor receptor family.* Cell, 1994. **77**(2): p. 261-71.

68. Pelicci, G., et al., *The motogenic and mitogenic responses to HGF are amplified by the Shc adaptor protein.* Oncogene, 1995. **10**(8): p. 1631-8.

69. Schaeper, U., et al., *Coupling of Gab1 to c-Met, Grb2, and Shp2 mediates biological responses.* J Cell Biol, 2000. **149**(7): p. 1419-32.

70. Tanimura, S. and K. Takeda, *ERK signalling as a regulator of cell motility.* J Biochem, 2017. **162**(3): p. 145-154.

71. Paumelle, R., et al., *Hepatocyte growth factor/scatter factor activates the ETS1 transcription factor by a RAS-RAF-MEK-ERK signaling pathway.* Oncogene, 2002. **21**(15): p. 2309-19.

72. Recio, J.A. and G. Merlino, *Hepatocyte growth factor/scatter factor activates proliferation in melanoma cells through p38 MAPK, ATF-2 and cyclin D1.* Oncogene, 2002. **21**(7): p. 1000-8.

73. Ridley, A.J., P.M. Comoglio, and A. Hall, *Regulation of scatter factor/hepatocyte growth factor responses by Ras, Rac, and Rho in MDCK cells.* Mol Cell Biol, 1995. **15**(2): p. 1110-22.

74. Royal, I., et al., *Activation of cdc42, rac, PAK, and rho-kinase in response to hepatocyte growth factor differentially regulates epithelial cell colony spreading and dissociation.* Mol Biol Cell, 2000. **11**(5): p. 1709-25.

75. Coltella, N., et al., *p38 MAPK turns hepatocyte growth factor to a death signal that commits ovarian cancer cells to chemotherapy-induced apoptosis.* Int J Cancer, 2006. **118**(12): p. 2981-90.

76. Xiao, G.H., et al., *Anti-apoptotic signaling by hepatocyte growth factor/Met via the phosphatidylinositol 3-kinase/Akt and mitogen-activated protein kinase pathways.* Proc Natl Acad Sci U S A, 2001. **98**(1): p. 247-52.

77. Boccaccio, C., et al., *Induction of epithelial tubules by growth factor HGF depends on the STAT pathway.* Nature, 1998. **391**(6664): p. 285-8.

78. Bussolino, F., et al., *Hepatocyte growth factor is a potent angiogenic factor which stimulates endothelial cell motility and growth.* J Cell Biol, 1992. **119**(3): p. 629-41.

79. Birchmeier, C. and E. Gherardi, *Developmental roles of HGF/SF and its receptor, the c-Met tyrosine kinase.* Trends Cell Biol, 1998. **8**(10): p. 404-10.

80. Zhou, D., et al., *Activation of hepatocyte growth factor receptor, c-met, in renal tubules is required for renoprotection after acute kidney injury.* Kidney Int, 2013. **84**(3): p. 509-20.

81. Nakamura, T., et al., *Myocardial protection from ischemia/reperfusion injury by endogenous and exogenous HGF.* J Clin Invest, 2000. **106**(12): p. 1511-9.

82. Borowiak, M., et al., *Met provides essential signals for liver regeneration.* Proc Natl Acad Sci U S A, 2004. **101**(29): p. 10608-13.

83. Watanabe, M., et al., *Hepatocyte growth factor gene transfer to alveolar septa for effective suppression of lung fibrosis.* Mol Ther, 2005. **12**(1): p. 58-67.

84. Boccaccio, C. and P.M. Comoglio, *Invasive growth: a MET-driven genetic programme for cancer and stem cells.* Nat Rev Cancer, 2006. **6**(8): p. 637-45.

85. Lengyel, E., et al., *C-Met overexpression in node-positive breast cancer identifies patients with poor clinical outcome independent of Her2/neu.* Int J Cancer, 2005. **113**(4): p. 678-82.

86. Tsao, M.S., et al., *Differential expression of Met/hepatocyte growth factor receptor in subtypes of non-small cell lung cancers.* Lung Cancer, 1998. **20**(1): p. 1-16.

87. Di Renzo, M.F., et al., *Overexpression of the Met/HGF receptor in ovarian cancer.* Int J Cancer, 1994. **58**(5): p. 658-62.

88. Liu, C., M. Park, and M.S. Tsao, *Overexpression of c-met proto-oncogene but not epidermal growth factor receptor or c-erbB-2 in primary human colorectal carcinomas.* Oncogene, 1992. **7**(1): p. 181-5.

89. Ferracini, R., et al., *The Met/HGF receptor is over-expressed in human osteosarcomas and is activated by either a paracrine or an autocrine circuit.* Oncogene, 1995. **10**(4): p. 739-49.

90. Morello, S., et al., *MET receptor is overexpressed but not mutated in oral squamous cell carcinomas.* J Cell Physiol, 2001. **189**(3): p. 285-90.

91. Pennacchietti, S., et al., *Hypoxia promotes invasive growth by transcriptional activation of the met protooncogene.* Cancer Cell, 2003. **3**(4): p. 347-61.

92. Ivan, M., et al., *Activated ras and ret oncogenes induce over-expression of c-met (hepatocyte growth factor receptor) in human thyroid epithelial cells.* Oncogene, 1997. **14**(20): p. 2417-23.

93. Birchmeier, C., et al., *Met, metastasis, motility and more.* Nat Rev Mol Cell Biol, 2003. **4**(12): p. 915-25.

94. Ferracini, R., et al., *Retrogenic expression of the MET proto-oncogene correlates with the invasive phenotype of human rhabdomyosarcomas.* Oncogene, 1996. **12**(8): p. 1697-705.

95. Scotlandi, K., et al., *Expression of Met/hepatocyte growth factor receptor gene and malignant behavior of musculoskeletal tumors.* Am J Pathol, 1996. **149**(4): p. 1209-19.

96. Koochekpour, S., et al., *Met and hepatocyte growth factor/scatter factor expression in human gliomas.* Cancer Res, 1997. **57**(23): p. 5391-8.

97. Tuck, A.B., et al., *Coexpression of hepatocyte growth factor and receptor (Met) in human breast carcinoma.* Am J Pathol, 1996. **148**(1): p. 225-32.

98. Bhowmick, N.A., E.G. Neilson, and H.L. Moses, *Stromal fibroblasts in cancer initiation and progression.* Nature, 2004. **432**(7015): p. 332-7.

99. Matsumoto, K. and T. Nakamura, *Hepatocyte growth factor and the Met system as a mediator of tumor-stromal interactions.* Int J Cancer, 2006. **119**(3): p. 477-83.

100. Kijima, Y., et al., *Amplification and overexpression of c-met gene in Epstein-Barr virus-associated gastric carcinomas.* Oncology, 2002. **62**(1): p. 60-5.

101. Lutterbach, B., et al., *Lung cancer cell lines harboring MET gene amplification are dependent on Met for growth and survival.* Cancer Res, 2007. **67**(5): p. 2081-8.

102. Nakazawa, K., et al., *Amplification and overexpression of c-erbB-2, epidermal growth factor receptor, and c-met in biliary tract cancers.* J Pathol, 2005. **206**(3): p. 356-65.

103. Tong, C.Y., et al., *Detection of oncogene amplifications in medulloblastomas by comparative genomic hybridization and array-based comparative genomic hybridization.* J Neurosurg, 2004. **100**(2 Suppl Pediatrics): p. 187-93.

104. Houldsworth, J., et al., *Gene amplification in gastric and esophageal adenocarcinomas.* Cancer Res, 1990. **50**(19): p. 6417-22.

105. Soman, N.R., et al., *The TPR-MET oncogenic rearrangement is present and expressed in human gastric carcinoma and precursor lesions.* Proc Natl Acad Sci U S A, 1991. **88**(11): p. 4892-6.

106. Soman, N.R., G.N. Wogan, and J.S. Rhim, *TPR-MET oncogenic rearrangement: detection by polymerase chain reaction amplification of the transcript and expression in human tumor cell lines.* Proc Natl Acad Sci U S A, 1990. **87**(2): p. 738-42.

107. Stransky, N., et al., *The landscape of kinase fusions in cancer.* Nat Commun, 2014. **5**: p. 4846.

108. International Cancer Genome Consortium PedBrain Tumor, P., *Recurrent MET fusion genes represent a drug target in pediatric glioblastoma.* Nat Med, 2016. **22**(11): p. 1314-1320.

109. Olivero, M., et al., *Novel mutation in the ATP-binding site of the MET oncogene tyrosine kinase in a HPRCC family.* Int J Cancer, 1999. **82**(5): p. 640-3.

110. Schmidt, L., et al., *Novel mutations of the MET proto-oncogene in papillary renal carcinomas.* Oncogene, 1999. **18**(14): p. 2343-50.

111. Schmidt, L., et al., *Germline and somatic mutations in the tyrosine kinase domain of the MET proto-oncogene in papillary renal carcinomas.* Nat Genet, 1997. **16**(1): p. 68-73.

112. Weidner, K.M., et al., *Mutation of juxtamembrane tyrosine residue 1001 suppresses loss-of-function mutations of the met receptor in epithelial cells.* Proc Natl Acad Sci U S A, 1995. **92**(7): p. 2597-601.

113. Peschard, P., et al., *Mutation of the c-Cbl TKB domain binding site on the Met receptor tyrosine kinase converts it into a transforming protein.* Molecular Cell, 2001. **8**(5): p. 995-1004.

114. Di Renzo, M.F., et al., *Somatic mutations of the MET oncogene are selected during metastatic spread of human HNSC carcinomas.* Oncogene, 2000. **19**(12): p. 1547-55.

115. Tanyi, J., et al., *Evaluation of the tyrosine kinase domain of the Met proto-oncogene in sporadic ovarian carcinomas*.* Pathol Oncol Res, 1999. **5**(3): p. 187-91.

116. Moon, Y.W., et al., *Missense mutation of the MET gene detected in human glioma.* Mod Pathol, 2000. **13**(9): p. 973-7.

117. Park, W.S., et al., *Somatic mutations in the kinase domain of the Met/hepatocyte growth factor receptor gene in childhood hepatocellular carcinomas.* Cancer Res, 1999. **59**(2): p. 307-10.

118. Tate, J.G., et al., *COSMIC: the Catalogue Of Somatic Mutations In Cancer.* Nucleic Acids Res, 2019. **47**(D1): p. D941-D947.

119. Ebos, J.M., et al., *Accelerated metastasis after short-term treatment with a potent inhibitor of tumor angiogenesis.* Cancer Cell, 2009. **15**(3): p. 232-9.

120. Paez-Ribes, M., et al., *Antiangiogenic therapy elicits malignant progression of tumors to increased local invasion and distant metastasis.* Cancer Cell, 2009. **15**(3): p. 220-31.

121. Sennino, B., et al., *Suppression of tumor invasion and metastasis by concurrent inhibition of c-Met and VEGF signaling in pancreatic neuroendocrine tumors.* Cancer Discov, 2012. **2**(3): p. 270-87.

122. De Bacco, F., et al., *MET inhibition overcomes radiation resistance of glioblastoma stem-like cells.* EMBO Mol Med, 2016. **8**(5): p. 550-68.

123. Du, Y., et al., *Blocking c-Met-mediated PARP1 phosphorylation enhances anti-tumor effects of PARP inhibitors.* Nat Med, 2016. **22**(2): p. 194-201.

124. Benvenuti, S., et al., *An 'in-cell trial' to assess the efficacy of a monovalent anti-MET antibody as monotherapy and in association with standard cytotoxics.* Mol Oncol, 2014. **8**(2): p. 378-88.

125. Michieli, P., et al., *Targeting the tumor and its microenvironment by a dual-function decoy Met receptor.* Cancer Cell, 2004. **6**(1): p. 61-73.

126. McDermott, U., et al., *Identification of genotype-correlated sensitivity to selective kinase inhibitors by using high-throughput tumor cell line profiling.* Proc Natl Acad Sci U S A, 2007. **104**(50): p. 19936-41.

127. Bardelli, A., et al., *Amplification of the MET receptor drives resistance to anti-EGFR therapies in colorectal cancer.* Cancer Discov, 2013. **3**(6): p. 658-73.

128. Lennerz, J.K., et al., *MET amplification identifies a small and aggressive subgroup of esophagogastric adenocarcinoma with evidence of responsiveness to crizotinib.* J Clin Oncol, 2011. **29**(36): p. 4803-10.

129. Bahcall, M., et al., *Acquired METD1228V Mutation and Resistance to MET Inhibition in Lung Cancer.* Cancer Discov, 2016. **6**(12): p. 1334-1341.

130. Kong-Beltran, M., et al., *Somatic mutations lead to an oncogenic deletion of met in lung cancer.* Cancer Res, 2006. **66**(1): p. 283-9.

131. Onozato, R., et al., *Activation of MET by gene amplification or by splice mutations deleting the juxtamembrane domain in primary resected lung cancers.* J Thorac Oncol, 2009. **4**(1): p. 5-11.

132. Awad, M.M., et al., *Characterization of 1,387 NSCLCs with MET exon 14 (METex14) skipping alterations (SA) and potential acquired resistance (AR) mechanisms.* Journal of Clinical Oncology, 2020. **38**(15).

133. Ma, P.C., et al., *c-MET mutational analysis in small cell lung cancer: Novel juxtamembrane domain mutations regulating cytoskeletal functions.* Cancer Research, 2003. **63**(19): p. 6272-6281.

134. Fujino, T., K. Suda, and T. Mitsudomi, *Lung Cancer with MET exon 14 Skipping Mutation: Genetic Feature, Current Treatments, and Future Challenges.* Lung Cancer (Auckl), 2021. **12**: p. 35-50.

135. Awad, M.M., et al., *MET Exon 14 Mutations in Non-Small-Cell Lung Cancer Are Associated With Advanced Age and Stage-Dependent MET Genomic Amplification and c-Met Overexpression.* Journal of Clinical Oncology, 2016. **34**(7): p. 721-+.

136. Schrock, A.B., et al., *Characterization of 298 Patients with Lung Cancer Harboring MET Exon 14 Skipping Alterations.* Journal of Thoracic Oncology, 2016. **11**(9): p. 1493-1502.

137. Yan, B., et al., *Identification of MET genomic amplification, protein expression and alternative splice isoforms in neuroblastomas.* J Clin Pathol, 2013. **66**(11): p. 985-91.

138. Asaoka, Y., et al., *Gastric cancer cell line Hs746T harbors a splice site mutation of c-Met causing juxtamembrane domain deletion.* Biochem Biophys Res Commun, 2010. **394**(4): p. 1042-6.

139. Frampton, G.M., et al., *Activation of MET via diverse exon 14 splicing alterations occurs in multiple tumor types and confers clinical sensitivity to MET inhibitors.* Cancer Discov, 2015. **5**(8): p. 850-9.

140. Bladt, F., et al., *EMD 1214063 and EMD 1204831 constitute a new class of potent and highly selective c-Met inhibitors.* Clin Cancer Res, 2013. **19**(11): p. 2941-51.

141. Fujino, T., et al., *Sensitivity and Resistance of MET Exon 14 Mutations in Lung Cancer to Eight MET Tyrosine Kinase Inhibitors In Vitro.* J Thorac Oncol, 2019. **14**(10): p. 1753-1765.

142. Baltschukat, S., et al., *Capmatinib (INC280) Is Active Against Models of Non-Small Cell Lung Cancer and Other Cancer Types with Defined Mechanisms of MET Activation.* Clinical Cancer Research, 2019. **25**(10): p. 3164-3175.

143. Liu, X.D., et al., *A Novel Kinase Inhibitor, INCB28060, Blocks c-MET-Dependent Signaling, Neoplastic Activities, and Cross-Talk with EGFR and HER-3.* Clinical Cancer Research, 2011. **17**(22): p. 7127-7138.

144. Wolf, J., et al., *Capmatinib in MET Exon 14-Mutated or MET-Amplified Non-Small-Cell Lung Cancer.* N Engl J Med, 2020. **383**(10): p. 944-957.

145. Paik, P.K., et al., *Tepotinib in Non-Small-Cell Lung Cancer with MET Exon 14 Skipping Mutations.* N Engl J Med, 2020. **383**(10): p. 931-943.

146. Recondo, G., et al., *Molecular Mechanisms of Acquired Resistance to MET Tyrosine Kinase Inhibitors in Patients with MET Exon 14-Mutant NSCLC.* Clin Cancer Res, 2020. **26**(11): p. 2615-2625.

147. Cerqua, M., et al., *METΔ14 promotes a ligand-dependent, AKT-driven invasive growth.* Life Sci Alliance, 2022. **5**(10).

148. Goodfellow, I., Y. Bengio, and A. Courville, *Deep Learning.* Deep Learning, 2016: p. 1-775.

149. Ahmad, A.B. and T. Tsuji, *Traffic Monitoring System Based on Deep Learning and Seismometer Data.* Applied Sciences-Basel, 2021. **11**(10).

150. Nielsen, M., *Neural Networks and Deep Learning.* 2015.

151. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* Nature, 2015. **521**(7553): p. 436-44.

152. Bank, D., N. Koenigstein, and R. Giryes, *Autoencoders.* arXiv, 2020.

153. Vincent, P., et al., *Extracting and composing robust features with denoising autoencoders.* Proceedings of the 25th International Conference on Machine Learning, 2008: p. 7.

154. Rifai, S., et al., *Contractive auto-encoders: explicit invariance during feature extraction*, in *Proceedings of the 28th International Conference on International Conference on Machine Learning*. 2011, Omnipress: Bellevue, Washington, USA. p. 833–840.

155. Kingma, D.P. and M. Welling, *Auto-Encoding Variational Bayes.* arXiv, 2013.

156. Geddes, T.A., et al., *Autoencoder-based cluster ensembles for single-cell RNA-seq data analysis.* BMC Bioinformatics, 2019. **20**(Suppl 19): p. 660.

157. Wang, T., et al., *BERMUDA: a novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes.* Genome Biol, 2019. **20**(1): p. 165.

158. Gold, M.P., A. LeNail, and E. Fraenkel, *Shallow Sparsely-Connected Autoencoders for Gene Set Projection.* Pac Symp Biocomput, 2019. **24**: p. 374-385.

159. Alessandri, L., et al., *Sparsely-connected autoencoder (SCA) for single cell RNAseq data mining.* NPJ Syst Biol Appl, 2021. **7**(1): p. 1.

160. Reungwetwattana, T., et al., *The race to target MET exon 14 skipping alterations in non-small cell lung cancer: The Why, the How, the Who, the Unknown, and the Inevitable.* Lung Cancer, 2017. **103**: p. 27-37.

161. Yu, T.M., et al., *Multiple Biomarker Testing Tissue Consumption and Completion Rates With Single-gene Tests and Investigational Use of Oncomine Dx Target Test for Advanced Non-Small-cell Lung Cancer: A Single-center Analysis.* Clin Lung Cancer, 2019. **20**(1): p. 20-29 e8.

162. Pennell, N.A., et al., *Economic Impact of Next-Generation Sequencing Versus Single-Gene Testing to Detect Genomic Alterations in Metastatic Non-Small-Cell Lung Cancer Using a Decision Analytic Model.* JCO Precis Oncol, 2019. **3**: p. 1-9.

163. Jennings, L.J., et al., *Guidelines for Validation of Next-Generation Sequencing-Based Oncology Panels: A Joint Consensus Recommendation of the Association for Molecular Pathology and College of American Pathologists.* J Mol Diagn, 2017. **19**(3): p. 341-365.

164. Poirot, B., et al., *MET Exon 14 Alterations and New Resistance Mutations to Tyrosine Kinase Inhibitors: Risk of Inadequate Detection with Current Amplicon-Based NGS Panels.* J Thorac Oncol, 2017. **12**(10): p. 1582-1587.

165. Davies, K.D., et al., *DNA-Based versus RNA-Based Detection of MET Exon 14 Skipping Events in Lung Cancer.* J Thorac Oncol, 2019. **14**(4): p. 737-741.

166. Zhang, Y., et al., *Discerning novel splice junctions derived from RNA-seq alignment: a deep learning approach.* BMC Genomics, 2018. **19**(1): p. 971.

167. Du, X., et al., *Analysis and Prediction of Exon Skipping Events from RNA-Seq with Sequence Information Using Rotation Forest.* Int J Mol Sci, 2017. **18**(12).

168. Jaganathan, K., et al., *Predicting Splicing from Primary Sequence with Deep Learning.* Cell, 2019. **176**(3): p. 535-548 e24.

169. Zuallaert, J., et al., *SpliceRover: interpretable convolutional neural networks for improved splice site prediction.* Bioinformatics, 2018. **34**(24): p. 4180-4188.

170. Nosi, V., et al., *MET Exon 14 Skipping: A Case Study for the Detection of Genetic Variants in Cancer Driver Genes by Deep Learning.* Int J Mol Sci, 2021. **22**(8).

171. Wen, J., et al., *A classification model for lncRNA and mRNA based on k-mers and a convolutional neural network.* BMC Bioinformatics, 2019. **20**(1): p. 469.

172. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotechnol, 2011. **29**(1): p. 24-6.

173. Champagnac, A., et al., *Frequency of MET exon 14 skipping mutations in non-small cell lung cancer according to technical approach in routine diagnosis: results from a real-life cohort of 2,369 patients.* J Thorac Dis, 2020. **12**(5): p. 2172-2178.

174. Miglio, U., et al., *The expression of LINE1-MET chimeric transcript identifies a subgroup of aggressive breast cancers.* Int J Cancer, 2018. **143**(11): p. 2838-2848.

175. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

176. Alessandri, L. and R.A. Calogero, *Functional-Feature-Based Data Reduction Using Sparsely Connected Autoencoders.* Methods Mol Biol, 2023. **2584**: p. 231-240.

177. Alessandri, L., et al., *Sparsely Connected Autoencoders: A Multi-Purpose Tool for Single Cell omics Analysis.* Int J Mol Sci, 2021. **22**(23).

178. Ye, X. and J.W.K. Ho, *Ultrafast clustering of single-cell flow cytometry data using FlowGrid.* BMC Syst Biol, 2019. **13**(Suppl 2): p. 35.

179. Huang, M., et al., *SAVER: gene expression recovery for single-cell RNA sequencing.* Nat Methods, 2018. **15**(7): p. 539-542.

180. Huang, C., et al., *Management of Non-small Cell Lung Cancer Patients with MET Exon 14 Skipping Mutations.* Curr Treat Options Oncol, 2020. **21**(4): p. 33.

181. Apicella, M., et al., *Dual MET/EGFR therapy leads to complete response and resistance prevention in a MET-amplified gastroesophageal xenopatient cohort.* Oncogene, 2017. **36**(9): p. 1200-1210.

182. Lee, J., P. Tran, and S.J. Klempner, *Targeting the MET Pathway in Gastric and Oesophageal Cancers: Refining the Optimal Approach.* Clin Oncol (R Coll Radiol), 2016. **28**(8): p. e35-44.

183. Lee, J., et al., *Gastrointestinal malignancies harbor actionable MET exon 14 deletions.* Oncotarget, 2015. **6**(29): p. 28211-22.

184. Hur, K., et al., *Hypomethylation of long interspersed nuclear element-1 (LINE-1) leads to activation of proto-oncogenes in human colorectal cancer metastasis.* Gut, 2014. **63**(4): p. 635-46.

185. Li, B., et al., *Higher levels of c-Met expression and phosphorylation identify cell lines with increased sensitivity to AMG-458, a novel selective c-Met inhibitor with radiosensitizing effects.* Int J Radiat Oncol Biol Phys, 2012. **84**(4): p. e525-31.

186. Melsted, P. and J.K. Pritchard, *Efficient counting of k-mers in DNA sequences using a bloom filter.* BMC Bioinformatics, 2011. **12**: p. 333.

187. Papageorgiou, L., et al., *Genomic big data hitting the storage bottleneck.* EMBnet J, 2018. **24**.

188. Stein, L.D., *The case for cloud computing in genome informatics.* Genome Biol, 2010. **11**(5): p. 207.

189. Ohta, T., T. Tanjo, and O. Ogasawara, *Accumulating computational resource usage of genomic data analysis workflow to optimize cloud computing instance selection.* Gigascience, 2019. **8**(4).

190. Piccolo, S.R. and M.B. Frampton, *Tools and techniques for computational reproducibility.* Gigascience, 2016. **5**(1): p. 30.

191. Dai, L., et al., *Bioinformatics clouds for big data manipulation.* Biol Direct, 2012. **7**: p. 43; discussion 43.

192. Afgan, E., et al., *Harnessing cloud computing with Galaxy Cloud.* Nat Biotechnol, 2011. **29**(11): p. 972-4.

193. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Research, 2002. **12**(6): p. 996-1006.

194. Jalili, V., et al., *The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update.* Nucleic Acids Res, 2020. **48**(W1): p. W395-W402.

195. Tangaro, M.A., et al., *Laniakea: an open solution to provide Galaxy "on-demand" instances over heterogeneous cloud infrastructures.* Gigascience, 2020. **9**(4).

196. Crosswell, L.C. and J.M. Thornton, *ELIXIR: a distributed infrastructure for European biological data.* Trends Biotechnol, 2012. **30**(5): p. 241-2.

197. Tangaro, M.A., et al., *Laniakea@ReCaS: exploring the potential of customisable Galaxy on-demand instances as a cloud-based service.* BMC Bioinformatics, 2021. **22**(Suppl 15): p. 544.

198. Andrews, T.S. and M. Hemberg, *Identifying cell populations with scRNASeq.* Mol Aspects Med, 2018. **59**: p. 114-122.

199. Picelli, S., et al., *Full-length RNA-seq from single cells using Smart-seq2.* Nat Protoc, 2014. **9**(1): p. 171-81.

200. Stahl, P.L., et al., *Visualization and analysis of gene expression in tissue sections by spatial transcriptomics.* Science, 2016. **353**(6294): p. 78-82.

201. Zhang, X., et al., *Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems.* Mol Cell, 2019. **73**(1): p. 130-142 e5.

202. Alessandri, L., M. Arigoni, and R. Calogero, *Differential Expression Analysis in Single-Cell Transcriptomics.* Methods Mol Biol, 2019. **1979**: p. 425-432.

203. Alessandri, L., et al., *rCASC: reproducible classification analysis of single-cell sequencing data.* Gigascience, 2019. **8**(9).

204. Kulkarni, N., et al., *Reproducible bioinformatics project: a community for reproducible bioinformatics analysis pipelines.* BMC Bioinformatics, 2018. **19**(Suppl 10): p. 349.

205. Murphy, K. and C. Weaver, *Janeway's Immunobiology, 9th Edition.* Janeway's Immunobiology, 9th Edition, 2017: p. 1-904.

206. Akira, S., S. Uematsu, and O. Takeuchi, *Pathogen recognition and innate immunity.* Cell, 2006. **124**(4): p. 783-801.

207. Janeway, C.A., Jr. and R. Medzhitov, *Innate immune recognition.* Annu Rev Immunol, 2002. **20**: p. 197-216.

208. Alberts, B., et al., *Molecular Biology of the Cell, Sixth Edition.* Molecular Biology of the Cell, Sixth Edition, 2015: p. 1-1342.

209. Burnet, F.M., *The clonal selection theory of acquired immunity ; the Abraham Flexner lectures of Vanderbilt University*. 1959: Cambridge University Press.

210. Pernis, B., et al., *Cellular localization of immunoglobulins with different allotypic specificities in rabbit lymphoid tissues.* J Exp Med, 1965. **122**(5): p. 853-76.

211. Brandtzaeg, P., *Role of secretory antibodies in the defence against infections.* Int J Med Microbiol, 2003. **293**(1): p. 3-15.

212. Dunkelberger, J.R. and W.C. Song, *Complement and its role in innate and adaptive immune responses.* Cell Res, 2010. **20**(1): p. 34-50.

213. Bousso, P., *T-cell activation by dendritic cells in the lymph node: lessons from the movies.* Nat Rev Immunol, 2008. **8**(9): p. 675-84.

214. Pennock, N.D., et al., *T cell responses: naive to memory and everything in between.* Advances in Physiology Education, 2013. **37**(4): p. 273-283.

215. Glusman, G., et al., *Comparative genomics of the human and mouse T cell receptor loci.* Immunity, 2001. **15**(3): p. 337-349.

216. Wong, W.K., J. Leem, and C.M. Deane, *Comparative Analysis of the CDR Loops of Antigen Receptors.* Front Immunol, 2019. **10**: p. 2454.

217. Davis, M.M. and P.J. Bjorkman, *T-Cell Antigen Receptor Genes and T-Cell Recognition.* Nature, 1988. **334**(6181): p. 395-402.

218. Srivastava, S.K. and H.S. Robins, *Palindromic Nucleotide Analysis in Human T Cell Receptor Rearrangements.* Plos One, 2012. **7**(12).

219. Schuldt, N.J. and B.A. Binstadt, *Dual TCR T Cells: Identity Crisis or Multitaskers?* J Immunol, 2019. **202**(3): p. 637-644.

220. Luckheeram, R.V., et al., *CD4(+)T cells: differentiation and functions.* Clin Dev Immunol, 2012. **2012**: p. 925135.

221. Sallusto, F., J. Geginat, and A. Lanzavecchia, *Central memory and effector memory T cell subsets: function, generation, and maintenance.* Annu Rev Immunol, 2004. **22**: p. 745-63.

222. Larbi, A. and T. Fulop, *From "Truly Naive" to "Exhausted Senescent" T Cells: When Markers Predict Functionality.* Cytometry Part A, 2014. **85a**(1): p. 25-35.

223. Lanzavecchia, A. and F. Sallusto, *Dynamics of T lymphocyte responses: intermediates, effectors, and memory cells.* Science, 2000. **290**(5489): p. 92-7.

224. Hamann, D., et al., *Evidence that human CD8(+)CD45RA(+)CD27(-) cells are induced by antigen and evolve through extensive rounds of division.* International Immunology, 1999. **11**(7): p. 1027-1033.

225. Han, M., et al., *Polymorphism of human CD1 genes.* Tissue Antigens, 1999. **54**(2): p. 122-127.

226. Schonrich, G. and M.J. Raftery, *CD1-Restricted T Cells During Persistent Virus Infections: "Sympathy for the Devil".* Front Immunol, 2018. **9**: p. 545.

227. Mori, L., M. Lepore, and G. De Libero, *The Immunology of CD1- and MR1-Restricted T Cells.* Annu Rev Immunol, 2016. **34**: p. 479-510.

228. Facciotti, F., et al., *Fine tuning by human CD1e of lipid-specific immune responses.* Proc Natl Acad Sci U S A, 2011. **108**(34): p. 14228-33.

229. Dougan, S.K., A. Kaser, and R.S. Blumberg, *CD1 expression on antigen-presenting cells.* Curr Top Microbiol Immunol, 2007. **314**: p. 113-41.

230. Siddiqui, S., L. Visvabharathy, and C.R. Wang, *Role of Group 1 CD1-Restricted T Cells in Infectious Disease.* Front Immunol, 2015. **6**: p. 337.

231. Lepore, M., L. Mori, and G. De Libero, *The Conventional Nature of Non-MHC-Restricted T Cells.* Front Immunol, 2018. **9**: p. 1365.

232. Van Kaer, L., V.V. Parekh, and L. Wu, *Invariant natural killer T cells: bridging innate and adaptive immunity.* Cell Tissue Res, 2011. **343**(1): p. 43-55.

233. Kawano, T., et al., *CD1d-restricted and TCR-mediated activation of valpha14 NKT cells by glycosylceramides.* Science, 1997. **278**(5343): p. 1626-9.

234. Godfrey, D.I. and M. Kronenberg, *Going both ways: immune regulation via CD1d-dependent NKT cells.* Journal of Clinical Investigation, 2004. **114**(10): p. 1379-1388.

235. Terabe, M., et al., *NKT cell-mediated repression of tumor immunosurveillance by IL-13 and the IL-4R-STAT6 pathway.* Nat Immunol, 2000. **1**(6): p. 515-20.

236. Terabe, M., et al., *Transforming growth factor-beta production and myeloid cells are an effector mechanism through which CD1d-restricted T cells block cytotoxic T lymphocyte-mediated tumor immunosurveillance: abrogation prevents tumor recurrence.* J Exp Med, 2003. **198**(11): p. 1741-52.

237. Hayday, A.C., *Gammadelta T cells and the lymphoid stress-surveillance response.* Immunity, 2009. **31**(2): p. 184-96.

238. De Libero, G., S.Y. Lau, and L. Mori, *Phosphoantigen Presentation to TCR gammadelta Cells, a Conundrum Getting Less Gray Zones.* Front Immunol, 2014. **5**: p. 679.

239. Zheng, J., et al., *gammadelta-T cells: an unpolished sword in human anti-infection immunity.* Cell Mol Immunol, 2013. **10**(1): p. 50-7.

240. Zhao, Y., C. Niu, and J. Cui, *Gamma-delta (gammadelta) T cells: friend or foe in cancer development?* J Transl Med, 2018. **16**(1): p. 3.

241. Li, Y., et al., *The Dual Roles of Human gammadelta T Cells: Anti-Tumor or Tumor-Promoting.* Front Immunol, 2020. **11**: p. 619954.

242. Braza, M.S. and B. Klein, *Anti-tumour immunotherapy with Vgamma9Vdelta2 T lymphocytes: from the bench to the bedside.* Br J Haematol, 2013. **160**(2): p. 123-32.

243. Di Carlo, E., et al., *Mechanisms of the antitumor activity of human Vgamma9Vdelta2 T cells in combination with zoledronic acid in a preclinical model of neuroblastoma.* Mol Ther, 2013. **21**(5): p. 1034-43.

244. Mangan, B.A., et al., *Cutting edge: CD1d restriction and Th1/Th2/Th17 cytokine secretion by human Vdelta3 T cells.* J Immunol, 2013. **191**(1): p. 30-4.

245. Petrasca, A., et al., *Human Vdelta3(+) gammadelta T cells induce maturation and IgM secretion by B cells.* Immunol Lett, 2018. **196**: p. 126-134.

246. Deusch, K., et al., *A major fraction of human intraepithelial lymphocytes simultaneously expresses the gamma/delta T cell receptor, the CD8 accessory molecule and preferentially uses the V delta 1 gene segment.* Eur J Immunol, 1991. **21**(4): p. 1053-9.

247. Kimura, Y., et al., *IL-17A-producing CD30(+) Vdelta1 T cells drive inflammation-induced cancer progression.* Cancer Sci, 2016. **107**(9): p. 1206-14.

248. Park, J.H. and H.K. Lee, *Function of gamma delta T cells in tumor immunology and their application to cancer therapy.* Experimental and Molecular Medicine, 2021. **53**(3): p. 318-327.

249. Schonefeldt, S., et al., *The Diverse Roles of gamma delta T Cells in Cancer: From Rapid Immunity to Aggressive Lymphoma.* Cancers, 2021. **13**(24).

250. Yamaguchi, H., et al., *A highly conserved major histocompatibility complex class I-related gene in mammals.* Biochem Biophys Res Commun, 1997. **238**(3): p. 697-702.

251. Riegert, P., V. Wanner, and S. Bahram, *Genomics, isoforms, expression, and phylogeny of the MHC class I-related MR1 gene.* J Immunol, 1998. **161**(8): p. 4066-77.

252. Corbett, A.J., et al., *T-cell activation by transitory neo-antigens derived from distinct microbial pathways.* Nature, 2014. **509**(7500): p. 361-5.

253. Kjer-Nielsen, L., et al., *MR1 presents microbial vitamin B metabolites to MAIT cells.* Nature, 2012. **491**(7426): p. 717-23.

254. Vacchini, A., et al., *MR1-Restricted T Cells Are Unprecedented Cancer Fighters.* Front Immunol, 2020. **11**: p. 751.

255. Dusseaux, M., et al., *Human MAIT cells are xenobiotic-resistant, tissue-targeted, CD161hi IL-17-secreting T cells.* Blood, 2011. **117**(4): p. 1250-9.

256. Kurioka, A., et al., *MAIT cells are licensed through granzyme exchange to kill bacterially sensitized targets.* Mucosal Immunology, 2015. **8**(2): p. 429-440.

257. Lepore, M., et al., *Functionally diverse human T cells recognize non-microbial antigens presented by MR1.* Elife, 2017. **6**.

258. Crowther, M.D., et al., *Genome-wide CRISPR-Cas9 screening reveals ubiquitous T cell cancer targeting via the monomorphic MHC class I-related protein MR1 (vol 12, pg 125, 2020).* Nature Immunology, 2020. **21**(6): p. 695-695.

259. Mori, L. and G. De Libero, *'Bohemian Rhapsody' of MR1T cells.* Nature Immunology, 2020. **21**(2): p. 108-110.

260. Wagnerova, A. and R. Gardlik, *In vivo reprogramming in inflammatory bowel disease.* Gene Ther, 2013. **20**(12): p. 1111-8.

261. Lee, J.M. and K.M. Lee, *Endoscopic Diagnosis and Differentiation of Inflammatory Bowel Disease.* Clin Endosc, 2016. **49**(4): p. 370-5.

262. Roda, G., et al., *Crohn's disease.* Nat Rev Dis Primers, 2020. **6**(1): p. 22.

263. de Souza, H.S. and C. Fiocchi, *Immunopathogenesis of IBD: current state of the art.* Nat Rev Gastroenterol Hepatol, 2016. **13**(1): p. 13-27.

264. Imam, T., et al., *Effector T Helper Cell Subsets in Inflammatory Bowel Diseases.* Front Immunol, 2018. **9**: p. 1212.

265. Brand, S., *Crohn's disease: Th1, Th17 or both? The change of a paradigm: new immunological and genetic insights implicate Th17 cells in the pathogenesis of Crohn's disease.* Gut, 2009. **58**(8): p. 1152-1167.

266. Casalegno Garduno, R. and J. Dabritz, *New Insights on CD8(+) T Cells in Inflammatory Bowel Disease and Therapeutic Approaches.* Front Immunol, 2021. **12**: p. 738762.

267. Bottois, H., et al., *KLRG1 and CD103 Expressions Define Distinct Intestinal Tissue-Resident Memory CD8 T Cell Subsets Modulated in Crohn's Disease.* Front Immunol, 2020. **11**: p. 896.

268. Lee, J.C., et al., *Gene expression profiling of CD8+ T cells predicts prognosis in patients with Crohn disease and ulcerative colitis.* J Clin Invest, 2011. **121**(10): p. 4170-9.

269. Catalan-Serra, I., et al., *Gammadelta T Cells in Crohn's Disease: A New Player in the Disease Pathogenesis?* J Crohns Colitis, 2017. **11**(9): p. 1135-1145.

270. Levine, J.H., et al., *Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis.* Cell, 2015. **162**(1): p. 184-97.

271. Mann, H.B. and D.R. Whitney, *On a Test of Whether One of 2 Random Variables Is Stochastically Larger Than the Other.* Annals of Mathematical Statistics, 1947. **18**(1): p. 50-60.

272. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society Series B-Statistical Methodology, 1995. **57**(1): p. 289-300.

273. Dash, P., et al., *Quantifiable predictive features define epitope-specific T cell receptor repertoires.* Nature, 2017. **547**(7661): p. 89-93.

274. Lepore, M., et al., *Parallel T-cell cloning and deep sequencing of human MAIT cells reveal stable oligoclonal TCRbeta repertoire.* Nat Commun, 2014. **5**: p. 3866.

275. Hannon, G.J. *FASTX-Toolkit.* 2010.

276. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* 2011, 2011. **17**(1): p. 3.

277. Bolotin, D.A., et al., *MiXCR: software for comprehensive adaptive immunity profiling.* Nat Methods, 2015. **12**(5): p. 380-1.

278. Freeman, G.H. and J.H. Halton, *Note on an exact treatment of contingency, goodness of fit and other problems of significance.* Biometrika, 1951. **38**(1-2): p. 141-9.

279. Wilcoxon, F., *Individual comparisons of grouped data by ranking methods.* J Econ Entomol, 1946. **39**: p. 269.

280. Boulter, J.M., et al., *Stable, soluble T-cell receptor molecules for crystallization and therapeutics.* Protein Eng, 2003. **16**(9): p. 707-11.

281. Reantragoon, R., et al., *Antigen-loaded MR1 tetramers define T cell receptor heterogeneity in mucosal-associated invariant T cells.* J Exp Med, 2013. **210**(11): p. 2305-20.

282. Schmaler, M., et al., *Modulation of bacterial metabolism by the microenvironment controls MAIT cell stimulation.* Mucosal Immunol, 2018. **11**(4): p. 1060-1070.