UNIVERSITA' DEGLI STUDI DI PAVIA

FACOLTA' DI INGEGNERIA
DIPARTIMENTO DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

DOTTORATO DI RICERCA IN BIOINGEGNERIA E BIOINFORMATICA
XXXI CICLO - 2018

BIOINFORMATICS TOOLS FOR THE IDENTIFICATION OF NOVEL ONCOLOGICAL TARGETS: THE KINASE CASE

PhD Thesis by
ROBERTA BOSOTTI

Advisor:
Prof. Paolo Magni

PhD Program Chair:
Prof. Riccardo Bellazzi

Science is not only a disciple of reason but, also, one of romance and passion

(*Stephen Hawking*)

Abstract (Italiano)

L'attività di ricerca descritta in questa tesi è stata condotta presso Nerviano Medical Sciences srl (NMS), un'azienda farmaceutica dedicata alla ricerca e allo sviluppo di farmaci antitumorali "mirati", la cosiddetta "target-therapy". La pipeline di NMS attualmente comprende composti che sono attivi contro diverse protein chinasi, in diversi stadi di sviluppo preclinico e clinico.

Con il termine terapia "mirata" si intende un farmaco in grado di bloccare la crescita del tumore interferendo con molecule specifiche, chiamate *target*, che sono coinvolte nello sviluppo e nella crescita tumorale. In genere un *target* è una proteina che viene espressa esclusivamente o almeno preferenzialmente nelle cellule tumorali, ma non nelle cellule normali. I nuovi farmaci non vengono infatti sviluppati per essere genericamente utilizzati contro tutti i tipi di tumore e/o paziente, come succede nel caso dei tradizionali farmaci citotossici, ma per agire in modo molto specifico contro un determinato *target* molecolare, presente solo in uno o più sottotipi tumorali.

Le chinasi, una famiglia di circa 500 enzimi con ruoli chiave in diverse funzioni cellulari, sono state trovate alterate (tipicamente iperattivate) in diversi tipi di tumore. Sono caratterizzate dalla presenza di una tasca, in grado di legare l'ATP, che è altamente conservata. Questa tasca può essere sfruttata per bloccare l'attività catalitica dell'enzima attraverso il legame con piccole molecole chimiche che, legandosi alla tasca dell'ATP, lo spiazzano. Questa peculiarità rende le chinasi un *target* ideale per lo sviluppo di nuovi farmaci "mirati". In genere, nei tumori, le chinasi sono attivate a seguito di una mutazione genica o da una loro overespressione e conseguente attivazione. L'espressione anomala di una chinasi può avvenire come conseguenza di un'alterazione del numero di copie del gene o a causa di riarrangiamenti cromosomici più complessi, che hanno come conseguenza la fusione del dominio catalitico della chinasi con un altro gene. Quest'ultimo gene, tipicamente espresso a livello costitutivo, è il responsabile dell'espressione della chinasi in un tessuto dove normalmente non è presente. Si tratta di eventi rari, che caratterizzano pochi pazienti all'interno di un sottotipo tumorale.

In questa tesi è descritta l'implementazione di un nuovo tool, chiamato KAOS (Kinase Automatic Outliers Search), che è stato specificamente sviluppato allo scopo di identificare *target* chinasici, contro i quali sviluppare nuovi farmaci antitumorali

KAOS è stato pensato per consentire l'identificazione di chinasi che presentano un profilo di espressione anomalo (*outlier*), se paragonato a quello osservato in altri campioni dello stesso tipo di tumore. Il software richiede come input dati di espressione genica, che possono essere stati prodotti sia con la tecnologia microarray, sia mediante Sequenziamento di Nuova Generazione (Next Generation Sequencing, NGS) ed utilizza l'espressione anormalmente alta di una chinasi come indicatore della presenza di un potenziale evento di fusione.

KAOS è stato utilizzato su dati di espressione disponibili pubblicamente su un ampio pannello di linee cellulari provenienti da diversi sottotipi tumorali ed ha permesso di identificare sia fusioni già note, sia nuove, mai riportate in precedenza.

In questa tesi è descritta inoltre l'implementazione di un sistema per la valutazione dell'espressione dell'intero chinoma umano, comprensivo sia di una parte di disegno e implementazione sperimentale, sia di un sistema dedicato all'analisi dei dati prodotti. La piattaforma, chiamata KING-REX (KINase Gene RNA EXpression), consente di analizzare la sola porzione di genoma relativa al chinoma (circa 500 chinasi), con una conseguente riduzione dei tempi di analisi e dei costi. Inoltre permette di identificare potenziali eventi di fusione genica a carico delle chinasi. La piattaforma si basa su un approccio *custom* di RNAseq mirato, il TruSeq Targeted RNA expression kit (TREx) prodotti da Illumina®.

La piattaforma KING-REX è stata ideata allo scopo di profilare il chinoma umano su sequenziatori Illumina® di piccola/media scala, richiedendo in questo modo risorse computazionali ridotte sia in termini di archiviazione dei dati, sia di processamento. KING-REX è quindi un sistema rapido e economicamente vantaggioso per identificare potenziali chinasi bersaglio per lo sviluppo di nuovi farmaci.

In parallelo allo sviluppo di nuovi farmaci, che colpiscono specifiche chinasi riarrangiate, è importante l'implementazione di metodi di screening e di validazione che consentano la selezione dei pazienti, portatori della chinasi bersaglio del farmaco, da sottoporre a trattamento.

Con l'avvento delle nuove tecnologie di sequenziamento, l'identificazione di specifiche fusioni geniche può beneficiare dell'altissima sensibilità degli approcci di RNAseq "mirati", che vengono quindi sempre più proposti anche come metodi diagnostici. Uno di questi sistemi è l'Anchored Multiplex PCR (AMP) (Archer®), un sistema basato sulla tecnologia NGS che consente l'identificazione di riarrangiamenti che coinvolgono una o poche chinasi, senza la necessità di conoscere il partner coinvolto nella fusione. In questa tesi, viene descritto l'utilizzo di questa tecnologia, combinata con un pre-screening basato sull'immunoistochimica (IHC), per l'analisi di campioni clinici di tumore colorettale (CRC). L'utilizzo della tecnologia AMP ha consentito di identificare pazienti affetti da CRC portatori di due nuovi riarrangiamenti genici che coinvolgono le chinasi NTRK1 e ALK e che sono risultati responsivi al trattamento con entrectinib, un farmaco che ha come bersaglio proprio queste chinasi, inizialmente sviluppato nei laboratori di NMS.

Abstract (English)

The research activity described in this thesis has been conducted at Nerviano Medical Sciences srl (NMS), a research-based company dedicated to the discovery and development of innovative target drugs for the treatment of cancer, with a pipeline including a number of kinase target compounds at different stages of development.

A target therapy is a drug able to block cancer growth by interfering with specific molecules (targets), involved in cancer development. Typically, a target is a protein specifically or at least preferentially expressed in cancer, but not in normal cells. These new drugs are not meant to be active generically against all tumors and/or patients, as in the case of traditional cytotoxic drugs, but to act against specific molecular targets expressed in one or more tumor subtypes only.

Kinases are a family of about 500 enzymes involved in several key cellular functions, which have been often found deregulated in cancer. They are characterized by a conserved ATP-binding pocket, which can be exploited for the binding of small molecules, blocking the catalytic activity of the enzyme, thus representing ideal targets for drug development. Kinases in tumors are activated by gene mutations or overexpression, as a consequence of copy number alteration or more complex genomic rearrangements, like gene fusions, which are rare events resulting in the overexpression and activation of the driver kinase.

In this thesis, in order to identify potential new targets for the development of novel drugs, the implementation of a tool, called KAOS (Kinase Automatic Outliers Search) is described.

KAOS was specifically developed for the identification of kinases showing an outlier gene expression profile, when compared to other samples from the same tumor subtype. The tool requires in input gene expression data from either microarray or RNAseq and uses the anomalous overexpression of a kinase as readout of the presence of a gene fusion event. The use of KAOS on publicly available cell line gene expression data allowed for the detection of known and novel kinase gene rearrangements.

In addition, the implementation of a comprehensive whole kinome expression screening, called KING-REX (KINase Gene RNA

EXpression), is also described, enabling the analysis of all human kinases with reduced time and costs. The platform permits investigating kinase expression and identifying potential gene fusion events using a customized Illumina® RNAseq targeted NGS approach (Illumina® TruSeq Targeted RNA expression kit, TREx), together with an ad hoc analysis pipeline. KING-REX has been conceived for the profiling of the human kinome on small/medium scale Illumina® sequencers, requiring reduced computational resources in terms of storage space and data processing, thus representing a rapid and cost effective kinome investigation tool in the field of kinase target identification, for applications in cancer biology.

In parallel, with the development of new drugs targeting specific kinase rearrangements, it is important the development of screening and validation methods, allowing for the selection of the patient population harboring a specific driver gene, for treatment prescription. With the advent of Next Generation Sequencing (NGS), the detection of specific gene fusions can benefit from the high sensitivity of target-RNAseq approaches and has been proposed as diagnostic platforms. One of these methods is the Anchored Multiplex PCR (AMP) (Archer®), an NGS-based system allowing for the detection of rearrangements for a selected number of kinases, without requiring the knowledge of the rearrangement partner. In this thesis the use of AMP technology, combined to immunohistochemistry (IHC) pre-screening, is described for the analysis of colorectal cancer (CRC) clinical specimens. The use of this test permitted the identification of patients harboring novel rearrangements of the kinases NTRK1 and ALK in CRC patients, responsive to the treatment with entrectinib, a drug initially developed at NMS specifically targeting these kinases.

Contents

Chapter 1

Introduction

1.1. Target therapies and personalized medicine

With the determination of the DNA sequence of the entire human genome [1-3] and the increasing knowledge of molecular mechanisms underlying the genesis and progression of tumors, research in the oncology field is geared towards the new paradigm of targeted therapies. Targeted cancer therapies are drugs able to block the growth and proliferation of cancer by interfering with specific molecules (targets), involved in cancer growth and progression [4]. Typically, a target is a protein specifically or at least preferentially expressed in cancer but not in normal cells. In this way, target therapies result generally less toxic than traditional chemotherapy and are well tolerated, being directed vs. tumor cell lines only while sparing normal cell lines.

Recent technological advances, including comprehensive sequencing of different cancer subtypes, have indeed highlighted how genetic alterations are associated with specific hallmarks of cancer [5].

In parallel, clinical practice is evolving from a histology and organ-based classification of tumors to therapeutic choices based on the genetic background of the individual patient and the molecular profile of the individual tumors, the so-called "personalized" therapy, with the aim of delivering drugs targeting specific oncogenic pathways and active in a given tumor. Patients are therefore divided in subgroups,

based on the genomic characteristics, and treated with the most appropriate drug for their disease, to obtain the best response rate [6, 7]. The new drugs in fact are not meant to be active generically against all tumors and/or patients, as in the case of traditional cytotoxic drugs, but to act against specific molecular targets for one or more tumor subtypes, with a significant change in the paradigm of "one drug fits all" vs. the "one drug, one patient" treatment.

The response to a specific therapy can be indeed predicted based on the presence of a functional defect in the tumor of the patient: not all patients will respond to the treatment, but only those with a tumor harboring a target which is constitutively activated as a consequence of a genomic alteration, such as gene mutation, gene overexpression or even more complex genomic rearrangements resulting in new gene fusions.

1.2. Protein kinases as oncological drug targets

Protein kinases constitute one of the largest families of enzymes that share a highly homologous catalytic domain (the kinase domain), which transfers the gamma phosphate from nucleoside triphosphates (ATP) to protein substrates, activating signal cascades and regulating multiple complex cellular processes. In several human diseases, such as cancer, kinases are often deregulated by gene alterations, leading to their anomalous expression and activation [8].

The abnormal oncogenic activation of protein kinases in tumors can occur as consequence of different types of genomics alterations [9, 10]. The most common being somatic point mutations [11], as in the case of the BRAF V600E mutation in melanoma [12] or the RET M918T in thyroid carcinoma [13]. Another type of alteration is gene overexpression as a consequence of gene amplification, an example being HER2 amplification in breast cancer [14], or of more complex genomic rearrangements, involving chromosomal inversions or translocations, and resulting in the generation of new fusion genes.

A fusion gene is a hybrid gene originated from the juxtaposition of the catalytic domain of a kinase with the N-terminal portion of another gene, constitutively expressed. In this way kinase anomalous expression is driven by the promoter of that second gene. Fusion genes

are highly expressed and more active that the wild type ones and are responsible of the tumorigenesis. An example is the chromosomal translocation between chr9 and chr22, generating the Philadelphia chromosome, resulting in the production of an aberrant fusion protein: BCR-ABL (the Abelson tyrosine kinase, ABL), which drives cell proliferation in Chronic Myeloid Leukemia (CML) [15].

Patients harboring tumors driven by gene rearrangements are excellent candidates for targeted kinase inhibitor therapies. Moreover 'druggability' by small molecule inhibitors, binding the conserved ATP-pocket, makes kinases therapeutically very attractive [16]: more than 500 kinases (the kinome) are encoded in the human genome, and kinase inhibitors now account for a quarter of all current drug discovery research and development efforts [17-20].

The clinical success of tyrosine kinase inhibitors has been proven by a number of examples, the first one being imatinib (gleevec) active in BCR–ABL fusion-positive leukemia patients [21]. In this tumor, the BCR-ABL kinase fusion gene is constitutively expressed and activated and becomes the driver of the oncogenicity [22]. The establishment of imatinib as the standard therapy for CML has indeed improved the 10-year survival rates from less than 20% in 1983 to 92% in 2013, with a life expectancy near to the one of healthy individuals [23]. Patient responses to imatinib treatment were so striking that the pillow was nicknamed as the "magical bullet".

These encouraging results have moved the oncology research field forward, versus the search for other oncogenic kinase gene fusions in many different tumor types, representing an interesting opportunity for the discovery of novel therapeutic targets. This strategy has been indeed successfully applied in the treatment of lung adenocarcinoma, where the introduction into the clinical practice of therapeutic approaches based on the specific targeting of activated kinases has dramatically changed the clinical outcome of many patients [24, 25]. Genomics analysis have clearly shown that non-small-cell lung cancer (NSCLC) does not represent a single disease, but it can be classified in several subgroups, based on the presence of different kinase alterations: KRAS missense mutations and EGFR activating mutations representing the most frequent alterations in NSCLC, followed by ALK rearrangements and other genomics alterations, each affecting a smaller fraction of the tumors (around 1-3%). For about 30% of NSCLC the

driver gene has not been identified so far and this "grey area" represents a potential source for the discovery of novel targets [26].

In particular, ALK (Anaplastic Lymphoma Kinase) gene was initially described in 1994 as involved in a translocation with nucleophosmin (NPM) in anaplastic large cell lymphoma [27], while the first EML4–ALK (echinoderm microtubule–associated protein 4) rearrangement in a patient with NSCLC was identified in 2007 [28]. ALK fusions arise from the combination of the catalytic kinase domain at the 3′ of ALK and the 5′ portion of a different gene, which provides its promoter. Multiple different 5′ partners have been identified so far [29]. Crizotinib was the first-in-class ALK kinase inhibitor, followed by other potent inhibitors such as ceritinib, alectinib and brigatinib, which have received approval from the Food and Drug Administration (FDA) and/or the European Medicines Agency (EMA) [30].

Gene fusions involving other kinases, such as ROS1 (ROS Proto-Oncogene 1, Receptor Tyrosine Kinase) and TRKA (Neurotrophic Receptor Tyrosine Kinase 1), have also emerged as driver events in NSCLC and other cancer types [31]. TRKA protein, encoded by NTRK1 gene, is normally not expressed in colon but it becomes overexpressed once the region encoding its catalytic kinase domain is fused with the 5'portion of TPM3 (Tropomyosin 3) gene, as the consequence of an intra-chromosomal inversion, resulting in the TPM3-NTRK1 fusion gene [32]. Other examples include rearrangements involving FGFRs, RET, ROS1, MET or EGFR [33-37].

To specifically target ALK, TRKs and ROS1 gene fusions, NMS has developed entrectinib, a novel, potent, orally available inhibitor active against tumors driven by these activated kinases [38, 39]. NMS initiated entrectinib clinical development, then continued by Ignyta (San Diego, CA, USA), through two Phase I studies and a Phase II potentially registrative study, now almost completed [40]. Ignyta was recently acquired by Roche, who will complete entrectinib development up to commercialization. The compound granted by the FDA the orphan drug designation for the treatment of TRKs, ROS1 or ALK-positive NSCLC, orphan drug designation and rare pediatric disease designation for the treatment of neuroblastoma and for treatment of TRK fusion-positive solid tumors. It also received by the FDA the break-through therapy designation in 2017.

The development of entrectinib prompted the implementation of an adequate screening strategy for the selection of the patients harboring ALK, TRK or ROS1 gene fusions, eligible for this type of treatment.

1.3. Approaches for the detection of specific kinase gene fusions

In parallel with the development of new drugs targeting specific kinase rearrangements, it is getting increasingly important the development of screening and validation methods allowing for the selection of the patient population harboring that specific driver gene, for treatment prescription. Diagnostic tests are indeed an indispensable part of personalized medicine, and several examples of companion diagnostics have already faced the market, like for instance the c-Kit pharmDx™ (DAKO), required for the diagnosis of c-KIT positive gastrointestinal tumors (GIST) and the prescription of treatment with gleevec or the DAKO Herceptest™ for the detection of HER2 positive patients, as an indication for treatment with transtuzumab [41].

Patient screening strategies are often based on the use of immunohistochemistry (IHC) analysis, for the identification of samples harboring an anomalous expression of the target kinase, as a hallmark of the presence of a kinase gene fusion, followed by Fluorescence in situ Hybridization (FISH) analysis, for its confirmation in the IHC positive samples. The identification of rare events of kinase over-expression in tumor samples can be indeed used as readout of underlying genomics rearrangements, leading to the expression of the target gene in a tissue where it is normally not significantly expressed.

With the advent of Next Generation Sequencing (NGS), the detection of specific gene fusions can benefit from the high sensitivity of target-RNAseq approaches and has been proposed as diagnostic platform.

In particular, Anchored Multiplex PCR (AMP) method (Archer®) is an NGS-based system allowing for the detection of rearrangements for a selected number of kinases, without requiring the knowledge of the rearrangement partner. Using this technology, combined to IHC pre-screening, a two-step diagnostic test has been implemented for the identification of NTRK1, NTRK2, NTRK3, ROS1 and ALK rearrangements in formalin-fixed paraffin-embedded (FFPE) clinical

specimens [42]. The use of this diagnostic test for the identification of patients harboring novel NTRKs and ALK rearrangements in CRC will be described in Chapter 2.

1.4. Approaches for the screening of new kinase gene fusion events

In cancer, besides the investigation of individual kinase genetic alterations [8], 'kinomics' approaches are emerging in the definition of co-expressed kinase functional roles in health and disease [43], as well as in integrative 'polypharmacology' approaches exploring synergizing effects of highly promiscuous kinase inhibitors [44]. Moreover, currently approved drugs target a very limited portion of the human kinome, leaving much of the kinase therapeutic potential unexplored.

Based on these considerations, the quest for novel kinase targets effective in oncogene-defined tumor types [19] is strongly encouraged to investigate tumor biology and to identify new candidate targets in specific disease contexts, also through the continuous generation of molecular data and the development of novel methods for kinome screening.

Screening of cancer samples from The Cancer Genome Atlas (TCGA) [45] showed that kinase rearrangements are rare events that can be detected in few tumor samples across a specific tumor type [8], demanding for new computational methods able to specifically detect rare recurring rearrangement events.

Several methods have been reported to detect genes with an outlier expression profile, using different algorithms for cancer outlier profile analysis like COPA (Cancer Outlier Profile Analysis) [46] or Gene Tissue Index (GTI) algorithm [47]. Both methods require transcriptional data from both tumor and normal tissue counterpart and therefore cannot be applied in the cases where normal counterpart is not available or in the profiling of cancer cell lines. Other approaches, such as ZODET [48] or the method proposed by Kothary and colleagues [49], search for abnormalities in the gene expression profile of an individual compared to a reference population. While GTI and COPA search sub-populations of samples for outlier expression levels of a gene, in this case sub-groups of genes (or a single gene) are analyzed in

a single sample and the algorithm searches for the gene(s) showing the highest expression value both in absolute terms and with respect to a reference population.

All these methods use microarray data as input, with the exception of Kothary's method which has been implemented to work with RNASeq data, expressed as RPKM values [50]. Kothary's method has not been implemented as a standalone tool, but it requires R [51] for statistical analysis and an external tool for visualization (GraphPad Prism) [52].

Other bioinformatics methods have been developed that do not rely on gene expression data as an indirect readout of gene rearrangements, but search for unbalanced 5'/3' gene expression or try to locate gene fusion genomics breakpoints. For example, Cancer Gene Census [53] compares expression levels of all the proximal versus distal exons for each exon-exon junction, to predict the existence of transcriptional breakpoints. Other methods try to predict fusion genes at the genomic level by checking genes in their transition regions or analyzing Copy Number Variations (CNVs) [54]. Finally, several tools, such as TopHat, FusionFinder, SOAPfuse, EricScript, deFuse, Chimerascan use RNA-Seq data to predict gene fusions, considering the discordant read pairs aligning to two different genes. Others, such as Comrad, CRAC, FusionMap, IPD-fusion and nFuse, use both RNA-Seq and whole genome sequencing (WGS) data to predict the presence and the mechanism behind the gene fusion, as reviewed in [55] and in [56]. The main limitation of such tools resides on the high number of false positives contaminating the results [57].

In this scenario, the implementation of a computational method aimed at the automatic identification of kinases selectively over-expressed in a very small fraction of samples within a specific tissue will be described in Chapter 3.

These 'omics' analysis approaches can be used for the detection of rearranged kinases, however they produce huge amounts of molecular information that is not necessary if the task is restricted to kinase target identification, while requiring substantial computational power for data storage and management. A comprehensive whole kinome expression screening by next-generation sequencing (NGS) targeted RNA approaches would instead enable the analysis of focused portions of the transcriptome, with reduced time and costs.

For the detection of gene fusions, RNA-targeted approaches using assays spanning all the exons of the gene and all exon-exon boundaries

of widely characterized kinase diagnostic targets are available (Archer® FusionPlex® NGS assays [58]; QuantideX® NGS RNA Lung Cancer Kit [59]; Ovation® Fusion Panel Target Enrichment System V2 [60]; Ion AmpliSeq RNA Fusion Lung Cancer Research Panel [61]; [62, 63]). However, typically both commercial and custom based approaches have been developed for targeting very small panels of kinases, and don't permit a comprehensive whole kinome expression screening.

For these reasons a comprehensive method for whole kinome screening has been implemented. The platform development will be described in Chapter 4 and an example of its application will be shown in Chapter 5.

1.5. Outline

In this thesis the development and use of different tools specifically implemented for the identification of kinase gene fusions, representing potential targets for new drug development, will be presented:

Chapter 2: In Chapter 2, the application of a small-scale RNAseq-target approach, based on Archer® Anchored Multiplex PCR technology, for the detection of ALK, ROS1 and NTRKs colorectal cancer positive samples will be described. The system permits investigating the expression of few kinases at a time. Its application to clinical samples allowed for the identification of new kinase rearrangements, never previously reported.

Chapter 3: In Chapter 3, the implementation of KAOS, a bioinformatics pipeline for the screening of gene expression data and the identification of kinases selectively over-expressed in a very small fraction of samples within a specific tissue, as readout of the presence of a gene fusion event, will be described. The use of KAOS on publicly available cell line gene expression data allowed for the detection of known and novel kinase gene fusions.

Chapter 4: In Chapter 4, the implementation of KING-REX, a comprehensive kinome RNA targeted custom assay-based panel, will be described. The platform permits investigating kinase expression and identifying potential gene fusion events using a customized Illumina® RNAseq targeted NGS approach together with an ad hoc analysis pipeline.

Chapter 5: In Chapter 5, the use of KING-REX for the characterization of kinome gene expression of a newly established chordoma cell line will be described, as an example of KING-REX application.

Chapter 2

Identification of novel kinase gene fusions in colorectal cancer clinical samples by targeted RNA sequencing[1]

This chapter describes the molecular characterization of two tumor samples from colorectal cancer patients using Archer® Anchored Multiplex PCR, a targeted-RNA sequencing approach for the detection of rearrangements involving a selected number of kinases, without requiring the a priori knowledge of the rearrangement partner.

This approach allowed for the identification of two novel gene fusions, never previously reported. The first one is a new NTRK1 rearrangement (SCYL3-NTRK1) resulting from an inversion within chromosome 1, fusing exons 1–11 of the SCY1 Like Pseudokinase 3

[1] The content of this chapter is published in:

- Amatu A, Somaschini A, Cerea G, Bosotti R, Valtorta E, Buonandi P, Marrapese G, Veronese S, Luo D, Hornby Z, Multani P, Murphy D, Shoemaker R, Lauricella C, Giannetta L, Maiolani M, Vanzulli A, Ardini E, Galvani A, Isacchi A, Sartore-Bianchi A and Siena S. *Novel CAD-ALK gene rearrangement is drugable by entrectinib in colorectal cancer*, Br J Cancer;113(12) (2015)
- Milione M, Ardini E, Christiansen J, Valtorta E, Veronese S, Bosotti R, Pellegrinelli A, Testi A, Pietrantonio F, Fucà G, Wei G, Murphy D, Siena S, Isacchi A and De Braud F. *Identification and characterization of a novel SCYL3-NTRK1 rearrangement in a colorectal cancer patient*, Oncotarget; 8:55353-55360 (2017)

(SCYL3) gene with exons 12–17 of NTRK1 gene. The rearrangement was identified in the colorectal cancer sample of a 61-year-old patient. The second one is an ALK rearrangement (CAD-ALK) resulting from an inversion within chromosome 2, fusing exons 1–35 of the carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase (CAD) gene with exons 20–29 of ALK (C35-A20), detected in a colon adenocarcinoma sample of a 53-year-old patient. Treatment of this patient with entrectinib, an orally available pan-TRK, ROS1 and ALK inhibitor initially developed at NMS resulted in objective tumor response.

These results show how the screening for rearrangements involving specific kinases may help identifying the subset of patients able to derive benefit from treatment with specific targeted inhibitors.

2.1. Clinical background

Colorectal cancer is the third most frequently diagnosed cancer worldwide, with more than one million new cases per year, and represents a major cause of cancer related death [64]. Patients with advanced colorectal cancer are primarily treated with fluoropyrimidine-based systemic chemotherapy with or without irinotecan or oxaliplatin.

Despite intense efforts dedicated to the identification of new and more effective therapies, many patients still have a poor treatment outcome. Although new targeted agents (such as cetuximab, panitumumab, bevacizumab, aflibercept, ramucirumab and regorafenib) have been incorporated into clinical practice during the last decade proving incremental survival gains [65, 66], the identification of novel tumor targets and new targeted therapies is warranted to ensure a better long-term clinical benefit.

Activated ALK gene fusions, found in hematologic and solid malignancies, have been successfully exploited as therapeutic targets in lung and inflammatory myofibroblastic tumors with the ALK kinase inhibitors crizotinib and ceritinib [67, 68]. In colorectal cancer (CRC), ALK rearrangements are recurring events found in 0.4-3% of samples and involve as partner genes either EML4, C2orf44 or PRKAR1A [66]. Although expression of the resulting chimeric transcripts was shown in some patients, at the time of this analysis, the final evidence of ALK

fusion protein expression in CRC has been lacking, hampering the exploitation of these observations also in this therapeutic setting.

Similarly, many different NTRK1 rearrangements have been already identified in a wide range of solid tumors (see [69]), however only a limited number of fusion partners has been detected so far in CRC (TPM3-NTRK1, LMNA-NTRK1 and TPR-NTRK1 fusion genes) [32, 70-72]. All these rearrangements result in the expression of fusion proteins harboring a constitutively activated TRKA kinase domain as consequence of protein dimerization, due to the presence of a coiled-coil domain in the N-terminal sequence of the partner protein. Preclinical data demonstrated that activated TRKA-fusion proteins are associated with proliferation and survival in these subsets of CRC tumors [38, 73] and evidence of clinical benefit achieved with the orally available pan-TRK, ROS1 and ALK inhibitor entrectinib was clearly shown in a CRC patient bearing a LMNA-NTRK1 positive tumor, providing clinical validation of activated TRKA as a target in CRC [71].

2.2. The Anchored Multiplex PCR (AMP) method

In *vitro* and/or in *silico* technologies have been extensively applied to the identification of rearranged kinases and their partner genes. Recently, more sensitive approaches have been developed based on NGS. In particular, Anchored Multiplex PCR (AMP) method (Archer®) [42] is an NGS-based system allowing for the detection of rearrangements for a selected number of kinases, without requiring the knowledge of the rearrangement partner. Library preparation is based on PCR amplification using two primers: a universal primer and a primer designed on the target genes.

A custom panel allows for the determination of gene rearrangements in NTRK1, NTRK2, NTRK3, ROS1, ALK and RET genes (encoding the TRKA, TRKB, TRKC, ROS1, ALK and RET proteins), along with associated housekeeping genes to assess RNA fragmentation. The AMP method allows for the identification of novel fusion partners due to an initial adapter ligation step that facilitates priming without de novo

knowledge of the gene rearrangement partner and limitations from the small fragments generated in FFPE tissues (Figure 2.1).

Fusion detection is performed using the Archer™ Analysis Software [74], starting from FASTQ files. Briefly, upon adapter trimming of the reads and de-duplication of the reads using the random 8-mer molecular barcode, the reads are firstly mapped to the targets defined for the assay, to increase mapping specificity, and then the remaining ones are mapped directly to the human genome. If at least one read is found spanning two separate genes, this is considered a fusion candidate. Each fusion candidate read is grouped together to generate a consensus sequence, which is then used to annotate the two fusion partners by comparison to the human genome with BLAST and the annotations from the RefSeq.

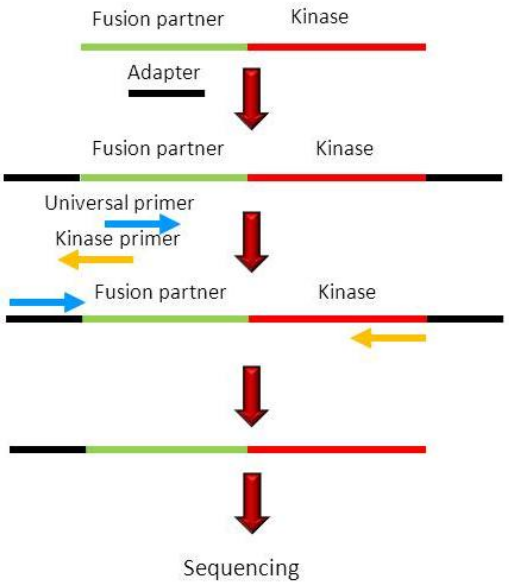

Figure 2.1: **Anchored Multiplex PCR (AMP) method**. Schema of Anchored Multiplex PCR (AMP) fusion detection method.

As part of a wide screening program at Fondazione IRCCS Istituto Nazionale dei Tumori and Niguarda Cancer Center, the AMP

technology was applied to ALK or NTRK immunohistochemistry (IHC) positive samples from colorectal cancer tumors.

2.3. Identification of a novel SCYL3-NTRK1 gene fusion

The patient was a 61-year-old female diagnosed with adenocarcinoma of the right colon, infiltrating the pancreas. The patient progressed early on two standard treatment lines (FOLFOX-panitumumab followed by FOLFIRI-aflibercept). A deeper molecular characterization of the patient's primary tumor was performed. Results indicated that the tumor was wild type for RAS, BRAF, and EGFR with high microsatellite instability (MSI-H) profile. At the time of progression to second line therapy, the patient underwent endoscopic biopsy of the right-sided tumor mass as part of the pre-screening procedures for the enrollment in the entrectinib phase I clinical trial [40]. The tumor was tested by immunohistochemistry for TRKA, ROS1 and ALK proteins, whose expression may indicate the result of a genetic alteration. The IHC analysis revealed strong positivity for TRKA protein with a clear cytoplasmic distribution, suggesting a potential aberrancy of NTRK1 gene. The observed immunoreactivity was uniformly characterized by a basic faint cytoplasmic staining associated with a more intense staining organized in irregular or ovoidal clods, preferentially localized around the nuclei (Figure 2.2 A). To verify if the detected TRKA expression was indeed the result of a genomic rearrangement, fluorescence in situ hybridization (FISH) analysis was performed using a commercial break-apart probe for the NTRK1 gene (Abnova). Results from this analysis showed the presence of break-apart signal, with separate green and orange signals in 90% of analyzed nuclei, confirming the presence of a NTRK1 rearrangement (Figure 2.2 B) and prompting further molecular characterization for the identification of the N-terminal partner gene.

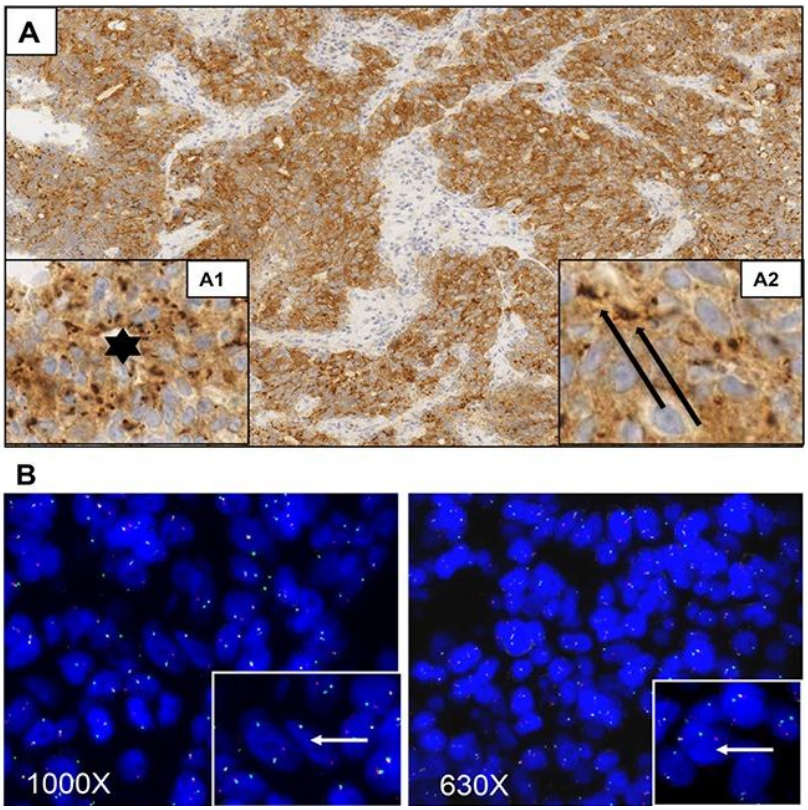

Figure 2.2: **Histologic, immunohistochemical, and fluorescence-in-situ hybridization analyses of the case presented** (from Milione M et al. Oncotarget 2017). Immunohistochemical and Fluorescent-In-Situ Hybridization (FISH) images of poorly differentiated CRC. In immunohistochemical assays Panel (**A**), magnification 100X, whole neoplastic cells are stained by TRKA. The observed staining is characterized by a basic faint uniformly cytoplasmic staining associated to more intense staining organized in irregular round or ovoidal dark bodies (insert A1, star: magnification 400X) preferentially distributed around nucleus. Dark bodies' size is variable ranging from tiny dot spherules-like to bigger bodies with irregular size and shape. The bigger bodies are fused in coarse ovoid structure surrounding nucleus (insert A2, arrow; magnification 400X). In panel (**B**), FISH analysis using the Abnova Break Apart probes showed the presence of one fusion signal along with separate green and orange signals suggesting the presence of a rearrangement of *NTRK1* gene.

The Anchored Multiplex PCR NGS approach was then applied [42] and allowed for the identification of a new NTRK1 rearrangement

resulting from an inversion within chromosome 1 fusing exons 1–11 of the SCY1 Like Pseudokinase 3 (SCYL3) gene with exons 12–17 of NTRK1 gene (Figure 2.3).

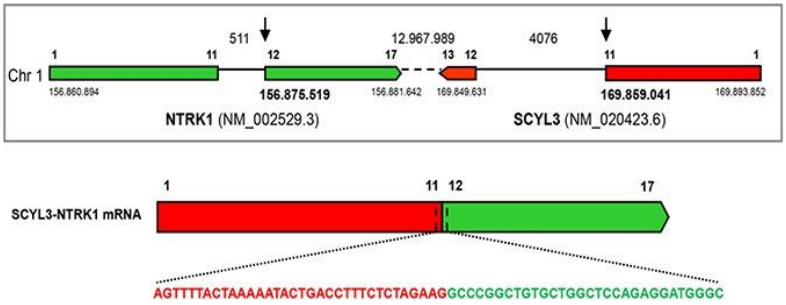

Figure 2.3: **Identification of the SCYL3-NTRK1 gene rearrangement** (from Milione M et al. Oncotarget 2017). Schematic representation of the *SCYL3-NTRK1* rearrangement. Exons involved in the rearrangement are represented by colored boxes: *SCYL3* is reported in green, while *NTRK1* is reported in red. The sequence spanning the junction site is shown in detail.

SCYL3, also known as PACE-1, a gene widely expressed in a variety of tissues, encodes for a protein whose function is still poorly understood. The full-length protein is able to bind to ezrin through a C-terminal domain determining its localization at the level of lamellipodia, a specialized cell structure that facilitates cell advancement across a substrate, where it probably plays a role in cell spreading and motility. In addition, the N-terminal region contains a myristoylation motif that is responsible for its association with the Golgi apparatus. Intriguingly, SCYL3 protein contains a kinase domain but any attempt to demonstrate its intrinsic kinase activity failed so far, suggesting that it should be probably considered a pseudokinase [75].

No other NTRK1 fusion transcripts were observed. As this rearrangement has never been previously described, to formally demonstrate its oncogenic potential, IL3-dependent Ba/F3 cells (a murine pro-B system that does not possess any endogenous human TRKA expression) were transfected with SCYL3-NTRK1 cDNA construct. As consequence of the expression of the corresponding TRKA-containing fusion protein, Ba/F3 cells acquired IL3-independent proliferation capability, demonstrating that SCYL3-NTRK1 is indeed an oncogenic driver. These transformed cells were used to test and

compare the potency of TRK inhibitors [40, 73]. Treatment with three different TRK inhibitors strongly affected the proliferation of Ba/F3-SCYL3-NTRK1 whereas no growth inhibition was observed in the parental Ba/F3 cells or the Ba/F3- SCYL3-NTRK1 cells grown in the presence of IL-3 [76]. Entrectinib was found to be the most potent compound inhibiting the proliferation of TRKA-driven cells with an IC50 value of 1.5 nM. Other TRK inhibitors LOXO-101 and crizotinib were also able to decrease the proliferation of Ba/F3-SCYL3-NTRK1 cells with IC50 values of 11.2 nM and 160 nM, respectively (Figure 2.4 A). The mechanism of action of entrectinib was confirmed by flow cytometer analysis of Propidium Iodide (PI) stained cells 18-hour post entrectinib treatment. Consistent with previous publication [38], entrectinib induced cell cycle arrest at G0/G1 (Figure 2.4 C). Furthermore, caspase 3/7 activities were peaked between 24 to 30 hours upon entrectinib treatment, at as low as 3.7 nM (Figure 2.4 D). Western Blot analysis demonstrated a dose-dependent modulation of TRKA phosphorylation with concomitant inhibition of phosphorylation of the downstream signal transducer, PLCγ (Figure 2.4 B), consistent with prior studies [38]. Crizotinib was also able to modulate TRKA signaling, although at significantly higher doses, in agreement with its calculated IC50.

The in vitro data clearly demonstrate that the *SCYL3-NTRK1* fusion gene is oncogenic and that the resulting fusion protein has constitutive kinase activation. Targeted inhibitors, in particular entrectinib, strongly inhibit the SCYL3-TRKA phosphorylation leading to cell growth inhibition and confirming the role of the expressed fusion protein as driver for proliferation. Preclinical data demonstrate that entrectinib represents an important potential opportunity for the treatment of patients whose tumors harbor this new NTRK1 rearrangement. Unfortunately, the clinical conditions of this patient rapidly degraded because of tumor progression and it was not possible to treat her with entrectinib.

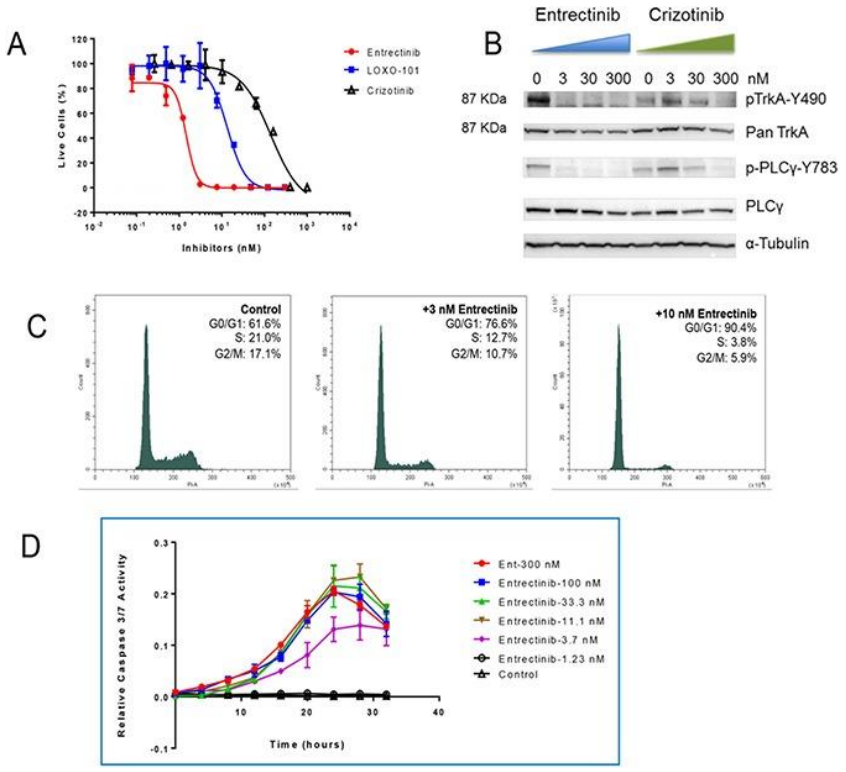

Figure 2.4: **Mechanism of action of entrectinib** (from Milione M et al. Oncotarget 2017). Confirmation that the SCYL3-NTRK1 fusion gene can be targeted by TRK inhibitors. (**A**) Comparison of anti-proliferation activities of 3 tyrosine kinase inhibitors in Ba/F3-SCYL3-NTRK1 cell line. (**B**) Western blot analysis of the changes in phosphorylation levels of TRKA and its downstream transducer PLCγ 2 hours post entrectinib and crizotinib treatment in Ba/F3-SCYL3-NTRK1 cells. (**C**) Cell cycle analysis of Ba/F3-SCYL3-NTRK1 treated with 3 and 10 nM of entrectinib for 18 hours compared to untreated cells. (**D**) Relative caspase 3/7 activities in Ba/F3-SCYL3-NTRK1 cells treated with increasing concentrations of entrectinib over time.

2.4. Identification of a novel CAD-ALK gene fusion

The patient was a 53-year-old woman with metastatic CRC (brain, thoracic lymph nodes and liver) who was in disease progression after

standard therapies, including surgery (right hemicolectomy), external beam radiation therapy to the central nervous system (CNS) and thoracic lymph nodes, and two lines of chemotherapy, both based on oxaliplatin, 5-fluorouracil/leucovorin, and bevacizumab, administered before and after the radiation therapy. The primary tumor was a grade 3 adenocarcinoma of the right colon metastatic to the supraclavicular lymph node. At progression, the patient gave written informed consent to a biopsy of liver metastases. All samples showed histology of primary CRC and CRC metastases (Figure 2.5 A1–C1). High levels of ALK protein were observed by IHC in the primary tumor, thoracic lymph node and liver metastasis (Figure 2.5 A2–C2), and the underlying ALK abnormality consisted of an ALK rearrangement as demonstrated by FISH (Figure 2.5 A3–C3).

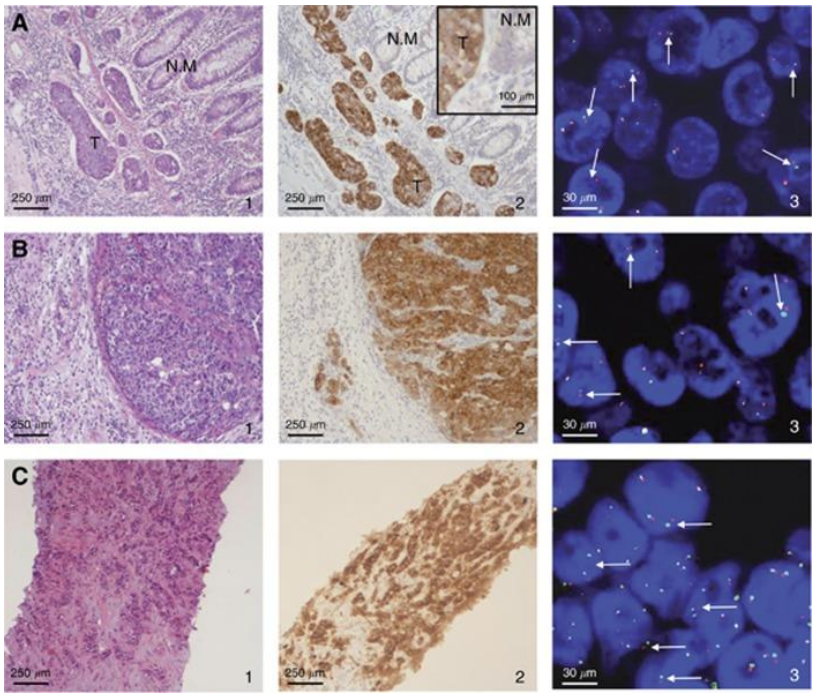

Figure 2.5: Histological, immunohistochemical, and fluorescence in situ hybridisation analyses of the primary right colonic tumor, lymph node, and liver metastasis of the case presented (from Amatu A et al, Br J Cancer 2015)**.** Haematoxylin & eosin, immunohistochemical and FISH images of the primary colon tumor (A 1–3; N.M: normal mucosa, T: tumor), lymph node (B 1–3), and liver metastasis (C 1–3) of the patient presented in this report,

showing malignant tumor (A1, B1, C1), ALK protein overexpression (A2, B2, C2) and ALK gene rearrangements (A3, B3, C3, white arrows). Original magnification of images: × 100 for haematoxylin & eosin and immunohistochemical staining, Insert × 400; × 630 for FISH analysis.

Further molecular characterization showed wild-type KRAS and NRAS exons 2, 3, and 4 and BRAF exon 15, no amplification of HER2, and no ROS1 or NTRK rearrangements. Moreover, no amplification of the ALK gene was found. Further investigation by Anchored Multiplex PCR NGS approach allowed for the identification of an ALK rearrangement resulting from an inversion within chromosome 2, fusing exons 1–35 of the carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase (CAD) gene with exons 20–29 of ALK (C35-A20). The CAD gene encodes a trifunctional protein that is associated with the enzymatic activities of the first three enzymes in the six-step pathway of pyrimidine biosynthesis: carbamoylphosphate synthetase (CPS II), aspartate transcarbamoylase, and dihydroorotase [77]. CAD is expressed in normal colonic tissue, and its expression is increased in inflammatory bowel disease [78]. CAD is also detected at medium–high levels in CRC by IHC [79]. The chimeric protein comprises the carbamoyl phosphate synthetase large-chain oligomerisation domain fused to the entire ALK kinase domain, potentially resulting in constitutive dimerisation causing ALK kinase activation.

The presence of a CAD-ALK rearrangement was also confirmed by PCR/Sanger sequencing (Figure 2.6 A and B).

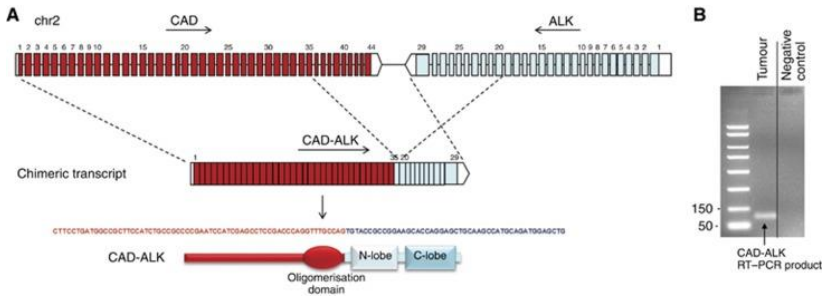

Figure 2.6: **Identification of the *CAD-ALK* gene rearrangement** (from Amatu A et al, Br J Cancer 2015)**.** (**A**) The upper section shows a schematic

representation of the *CAD-ALK* genomic DNA rearrangement and the resulting transcript. The sequence spanning the rearrangement junction is also shown. Exons are represented by colored boxes, and introns are represented by lines: *CAD* in red and *ALK* in light blue. The lower section shows the functional domains conserved in the chimeric CAD-ALK protein. (**B**) Characterization of the *CAD-ALK* transcript by PCR. Agarose gel showing amplification with primers for the rearranged *CAD-ALK* chimeric transcript, spanning *CAD* exon 35 to *ALK* exon 20. The tumor sample was compared with a negative control sample (U138-MG cell line, expressing *ALK* full length).

These findings prompted us to enroll the patient in the entrectinib phase I study at the recommended Phase II dose (RP2D) of 400 mg/m2 once a day. The first-response assessment via Computed Tomography (CT), which was performed 4 weeks after the beginning of treatment, showed a partial response per RECIST [2] v1.1 with a decrease in the sum of the target lesions by 38% (Figure 2.7). Computed tomography performed 4 weeks later confirmed this response. CNS metastases (brain and cerebellum) were stable. No drug-related adverse events were recorded.

The clinically meaningful anti-tumor effect observed upon treatment with entrectinib of the novel CAD-ALK rearranged gene in CRC provides the proof of concept that ALK alterations can act as drivers in CRC, building a new step towards personalized therapy in this clinical setting.

Therefore, although activated NTRK and ALK rearrangements are rare events in CRC [80, 81], a screening strategy based on simple NTRK and ALK assessment by IHC followed by targeted NGS based on anchored multiplex PCR [82] represents a feasible strategy, which will enhance the identification of patients who can benefit from entrectinib treatment in CRC and other histologies.

[2] RECIST (Response Evaluation Criteria in Solid Tumors) is a set of published rules which define when there is an improvement (response), a stable situation (stable disease) or a progression (progression) of the disease condition during a treatment. The version 1.1 is a revised version of the guidelines.

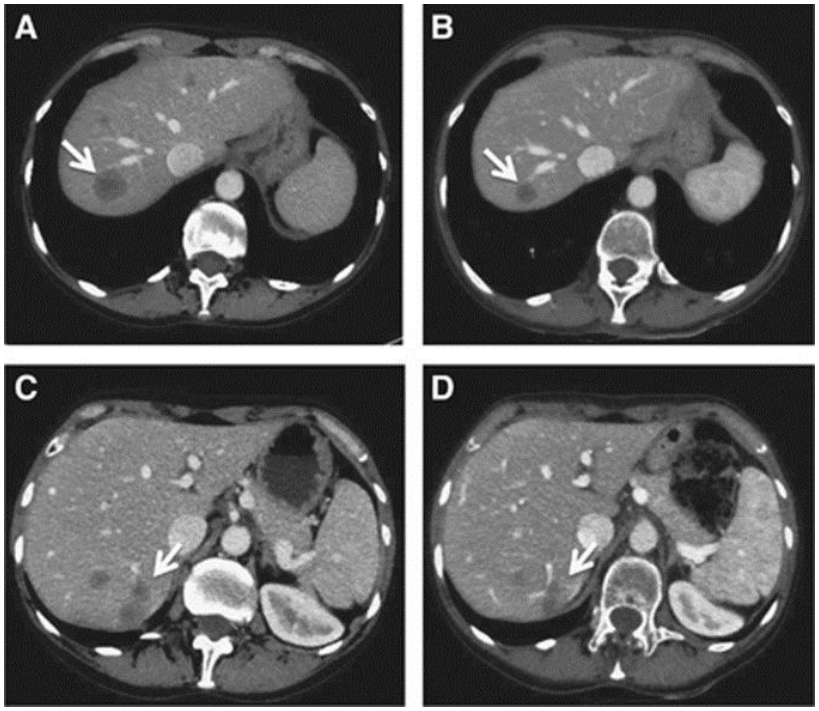

Figure 2.7: **Identification of the *CAD-ALK* gene rearrangement** (from Amatu A et al, Br J Cancer 2015)**.** (**A,C**) The baseline abdominal CT scan demonstrated liver involvement with the two largest lesions both in hepatic segment VII, measuring 27 and 33 mm in longest diameter, respectively (arrows). (**B, D)** At the first-response assessment, 4 weeks after the initiation of treatment, CT showed a RECIST partial response with an overall decrease in the sum of the target lesions of 38%, and lesions in segment VII displaying longest diameters of 15 and 22 mm (arrows).

Chapter 3

KAOS: a tool for the identification of overexpressed kinases, as readout of the presence of gene fusion events[3]

In the previous chapter, it has been shown how a targeted RNA-sequencing approach (Archer® FusionPlex technology) allows for the identification of gene fusions in samples where the presence of a specific kinase rearrangement is suspected on the basis of its overexpression. This system can therefore be used as a potent and sensitive validation method to confirm the presence of a rearrangement, permitting the identification of the partner gene involved in the gene fusion.

However, to identify new oncological targets, a broader screening strategy of cancer cell lines and clinical samples is needed, which is not limited to a small number of kinases but allows for the interrogation of multiple kinases at a time. The search for novel rearrangements of kinases may indeed contribute to understand the biology of cancerogenesis, as well as to lead to the identification of new candidate targets for drug discovery.

[3] The content of this chapter is published in: Nuzzo A, Carapezza G, Di Bella S, Pulvirenti A, Isacchi A and Bosotti R. *KAOS: a new automated computational method for the identification of overexpressed genes*, BMC Bioinformatics 17(Suppl 12):1188 (2016)

Kinase gene fusions are typically rare events, occurring in very small fractions (1-3%) of different tumor subtypes, therefore large datasets need to be queried in order to identify such rare events. The presence of a rearrangement in a tumor sample can be suspected when a kinase is expressed at anomalous higher level, when compared to other samples from the same tumor subtype. Thus, kinase overexpression can be used as readout of the presence of a gene fusion event.

In this chapter, the implementation of a dedicated bioinformatics tool for the identification of kinases selectively over-expressed in a very small fraction of samples within a specific tissue is described. The tool, called KAOS (Kinase Automatic Outliers Search), does not require a healthy counterpart or a reference sample for the analysis and can be therefore applied also to transcriptional data generated from cell lines. It requests as input gene expression data, either obtained from microarray experiments, as well as generated by RNASeq technologies.

The tool enables the automatic execution of iterative searches for the identification of extreme outliers and for the graphical visualization of the results. Filters can be applied to select the most significant outliers. The performance of the tool was evaluated using a synthetic dataset and compared to state-of-the-art tools. KAOS performs particularly well in detecting genes that are overexpressed in few samples or when an extreme outlier stands out on a high variable expression background.

For KAOS validation purposes, publicly available microarrays data form the Cancer Cell Line Encyclopedia (CCLE [83]) were used, showing that the tool is able to detect genes which are known to be overexpressed in certain tissue samples as well as to identify novel ones.

3.1. Tool implementation

Discovering candidate rearrangement in a panel of tumor cell lines based on their gene over-expression could be seen as a multidimensional problem, thus claiming for a systematic and automated approach. While a manual visual inspection of the expression pattern of a specific gene in a cell line is rather trivial, it is more complicated to extend the same analysis on the genome scale and on a high number of tumor samples. This is especially true when

searching for a rare event, such as the detection of the occurrence of an outlier gene expression only in few samples among a tumor tissue type.

The KAOS algorithm was implemented using the R statistical environment [51]. In particular, the R function "boxplot-with-outlier-label" [84] was used to calculate the statistics and the boxplots and the function fastNonDominatedSorting of the "nsga2R" package [85] to compute the rank of each outlier.

The algorithm has then been embedded in a software tool with a graphical user interface and a data interface to a MySQL database.

The software assumes MySQL, Java and R installed on the computer. An executable file automatically creates the database and populates it with an example artificial dataset and a user guide with the instructions on how to run the application and how to prepare the input files. It has been tested on different platforms and can be used on Linux, Windows and Mac OS.

KAOS can be used on different types of gene expression data, uploaded in the database.

More in detail, the tool implements the following strategy:

1. Pre-processing of the dataset: in this first step the tumor samples under investigation are annotated with information on the tissue of origin. This is a manual annotation pre-processing step, which can be achieved using external annotation tool or custom software scripts.
2. Identification of statistical outliers for each tissue-specific distribution: once a group of cell lines or tumor samples is grouped according to the tissue they belong to, a distribution of the gene expression values is computed and plotted together with a box-and-whiskers plot. R statistical environment, which implements the Grubb test for outliers, was used [86]. The result of the test includes all the values statistically considered outliers, independently of the absolute gene expression value. These lead to a high number of outliers since, in principle, the most extreme values of any distribution might be considered outliers (Figure 3.1 A).

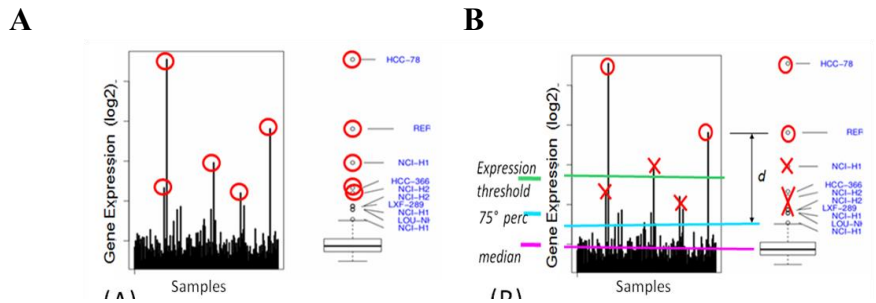

Figure 3.1: **Outlier detection** (from Nuzzo A. et al. BMC Bioinformatics 2015). Outlier detection method is reported. (**A**) Statistical detection: for each kinase, gene expression level in all the analysed samples belonging to a specific tumor type is reported as a histogram (left panel) and as boxplot (right panel). "Rare events" (a kinase over-expressed in one or a few cell lines and low/not expressed in the others) are identified by mean of the Grubb test and reported as a red circle. (**B**) Prioritization and filtering: the most relevant outlier kinases are selected applying specific filter criteria (minimal expression threshold; maximum median level of expression over the tumor type; minimum distance from the 75th percentile of the tissue-specific distribution; proportion of the number of outliers with respect to the whole dataset of outlier occurrences). Samples that do not consistently pass the imposed filters are removed (reported in the figure as red crosses).

3. Application of a chain of filtering criteria to isolate tissue-specific outliers with the objective of identifying samples with an extreme distribution for the examined gene in the tumor tissue under consideration, even when the background expression level is relatively high. Specific filtering criteria are applied to prioritize the more robust outliers. From the set of computed outliers, the KAOS algorithm selects the gene expression values fulfilling the following criteria (see Figure 3.1 B):
   - the expression value is above a minimal expression threshold
   - the gene has a minimum median (med) level of expression in the tissue-specific distribution
   - the value has a minimum distance (d) from the 75th percentile of the tissue-specific distribution
4. Ranking of the outliers to eliminate potential false positives: the remaining outliers are then ranked using the "non-dominated sorting" ranking algorithm, commonly used in multi-objective optimization field

[87], which allows for the selection of the best solution on the basis of two or more metrics. The ranking algorithm iteratively checks if each outlier *i* dominates over any other outlier *j* in the set, such that: if a) all the distances of *i* are greater than or equal to the corresponding distances of *j*, and b) at least one of distances of *i* is strictly greater than the corresponding distance of *j*, then *i* is assigned rank 1. Once all the outliers identified with rank 1 are discarded from the outliers set, the same comparison is performed to assign rank 2. The same iteration is repeated to assign all greater ranks until the set is empty. Figure 3.2 shows a bi-dimensional example on a 2-metrics computation.

To obtain the optimal settings for the detection of the extreme outliers, the following metrics were used:

a. the distances of the gene expression level from the 75th and the 50th percentile of the tissue-specific distribution

b. the proportion of the number of outliers with respect to the whole dataset of outlier occurrences for the given gene (a proportion that should be kept <5%).

5. Provision of a graphical summary of the most relevant outliers and related gene expression distributions.

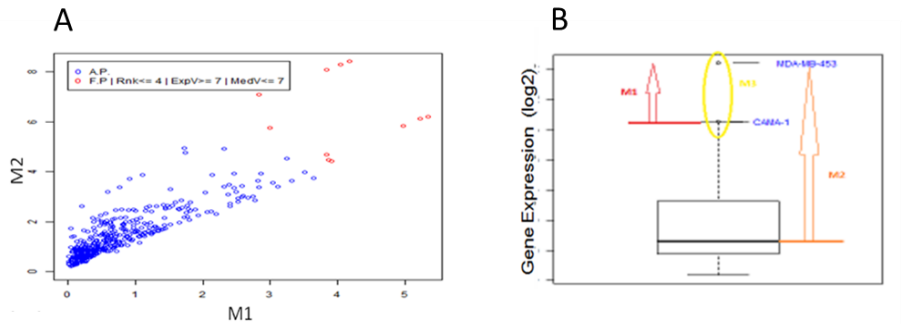

Figure 3.2: **The ranking algorithm** (from Nuzzo A. et al. BMC Bioinformatics 2015). (**A**) 2-d plot of the two measured distance: M1 is the distance from the upper wisker; M2 is the distance from the median. The "best" outliers lie on the top right corner of the graph, that corresponding to a major distance from both upper wisker and median, and are reported as red dots. (**B**) The metrics used for ranking are reported: M1 (red arrow) is the distance from the upper wisker; M2 (orange arrow) is the distance from the median; M3 (yellow circle) is the number of samples in which the gene has an outlier expression value.

3.2. Graphical interface

To permit a user-friendly interaction in parameter setting and visualizing of the results, as well as to enable an interactive use of the search strategies, the method was provided with a graphical interface, developed using the Java programming language, shown in Figure 3.3.

On the left-side, the filters which can be customized by the user for the selection of the outliers, including specification of the tumor tissue type and of the gene name of interest. For the outlier selection, a variety of statistical filter thresholds can be customized, such as gene expression level, median value and upper whisker. The user can also set the maximum threshold value for the rank and for the total number of outliers for the gene of interest in all the other tumor tissues, allowing for the identification of tumor specific, as well as general outliers. An additional function permits selecting the order in which results are sorted (by mean of gene expression values, rank, tissue or gene name). On the top left corner of the interface a menu provides the commands to save and reload the selected search criteria.

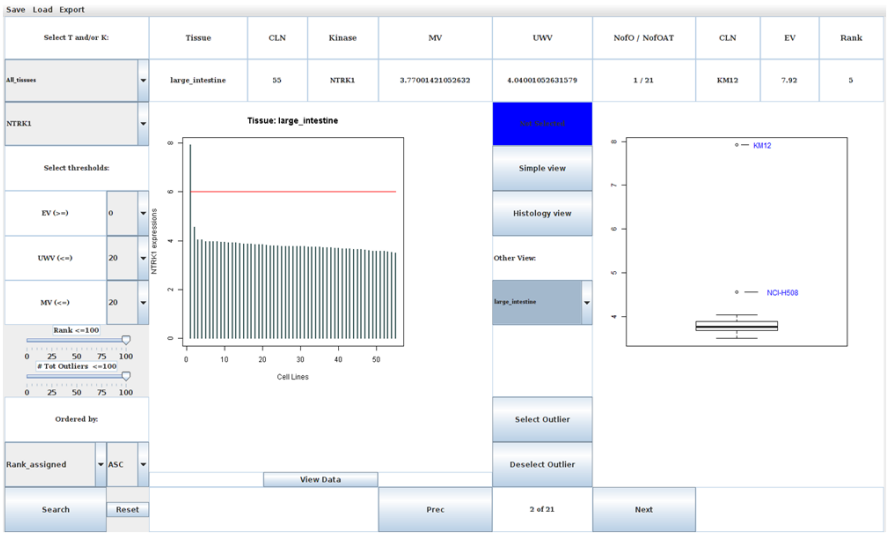

Figure 3.3: **Graphical User Interface** (from Nuzzo A. et al. BMC Bioinformatics 2015). KAOS graphical user interface, developed in Java, is reported. The interface permits both visualizing the information on the

detected outliers (top panel) and graphically representing the results (central panel) at the same time. The interface permits to customize query parameters and to filter the results (left panel).

Expression level of the gene of interest in the selected tissue is shown in the left panel, while the detected outliers are shown in the box plot on the right. Command buttons allow for the visualization of the expression profile of a gene in each single tumor tissue type one by one or all together at the same time. Additional information related to each outlier is reported at the top.

3.3. Performance evaluation on simulated data

By making use of simulated data, KAOS performance were compared with published outlier detection methods such as GTI [47], ZODET [48] and Kothari method [49].

Following the simulation proposed in [47] and [48], an artificial dataset was firstly generated with 1000 genes having an equal number of cancer and normal samples (30 in each class). The expression values of the genes were drawn from a normal distribution having mean 7 and standard deviation 1. Such values reflect the Affymetrix microarrays data analysis standard practice of considering 6-7 as a minimal expression value, as well as the typical average found in TCGA [45] and CCLE [83] datasets. The genes assumed to be differentially expressed, named True Positive (TP), were generated by adding a constant m to their expression in the k samples which have been marked as outliers' samples. The TP rate is 5% (i.e. 50 genes). To find the simulated false positives (FP), in each simulated experiment the genes were ranked according to their score and only the top t (t ranging from 10 to 50) were considered as predicted outliers. Within such a computed list, the correctly predicted TP genes and FP genes, the number of False Negative (FN) and the number of True Negative (TN) were then calculated. The average Precision, Recall and F-Measure was computed by running 50 simulations. The performances of KAOS, GTI, ZODET and the Kothari et al. method were analysed by varying k from

10 to 1 and t from 10 to 50. Since KAOS does not need case/control data it was tested on the 30 cancer cases only.

Table 3.1 gives the measures obtained for k=1 in all the compared tools. The results clearly show that KAOS outperforms the other tested methods in terms of Precision/Recall when the top 10 and 20 outliers are considered. In such a case KAOS seems to be the most robust method. On the other hand, when a higher threshold is applied, ZODET outperforms the other methods. Tables 3.2 and 3.3 give the measures for k=5 and k=10. In these cases, GTI has the best performances in terms of Precision/Recall.

Table 3.1 Tools comparison on simulated data for k = 1 (from Nuzzo A. et al. BMC Bioinformatics 2015)

	k=1, T=10			**k=1, T=20**			**k=1, T=50**		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
KAOS	**0.348**	0.348	**0.348**	**0.267**	**0.267**	**0.267**	0.162	0.162	0.162
Zodet	0.232	**0.232**	0.232	0.243	0.243	0.243	**0.220**	**0.220**	**0.220**
GTI	0.182	0.182	0.182	0.175	0.175	0.175	0.146	0.146	0.146
Khotary et al.	-	0.038	-	0.114	0.032	0.050	0.121	0.012	0.022

The comparison of Kaos performances is based on 50 simulations on a synthetic dataset made of 1000 genes expression values for 30 cases and 30 cancer test samples. The expression values were drawn from a normal distribution with mean 7 and standard deviation 1, where k samples which have been marked as outliers' samples (see Methods section for further details) and T is the top T number of outlier genes found. The table shows average Precision, Recall and F-Measure for k =1 and t ranging from 10 to 50.

Table 3.2 Tools comparison on simulated data for k=5 (from Nuzzo A. et al. BMC Bioinformatics 2015)

	k=5, T=10			**k=5, T=20**			**k=5, T=50**		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
KAOS	0.526	0.526	0.526	0.374	0.374	0.374	0.244	0.244	0.244
Zodet	0.828	0.828	0.828	0.699	0.699	0.699	0.516	0.516	0.516
GTI	**0.862**	**0.862**	**0.862**	**0.773**	**0.773**	**0.773**	**0.548**	**0.548**	**0.548**
Khotary et al.	0.454	0.246	0.319	0.450	0.124	0.194	-	0.041	-

The table shows the same simulation results as Table 3.1 when k=5.

Table 3.3 Tools comparison on simulated data for k=10 (from Nuzzo A. et al. BMC Bioinformatics 2015)

	k=10, T=10			k=10, T=20			k=10, T=50		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
KAOS	0.268	0.268	0.268	0.188	0.188	0.188	0.109	0.109	0.109
Zodet	0.986	0.986	0.986	0.948	0.948	0.948	0.765	0.765	0.765
GTI	**0.998**	**0.998**	**0.998**	**0.984**	**0.984**	**0.984**	**0.802**	**0.802**	**0.802**
Khotary et al.	0.837	0.776	0.805	0.754	0.389	0.513	0.767	0.151	0.252

The table shows the same simulation results as Table 3.1 when k=10.

The obtained results confirm the value of KAOS in detecting the most extreme outliers, since the algorithm was designed and optimized with the aim of searching for very rare rearrangements and for extreme outliers in a high variability expression context. That is indeed what resembles most likely real cases, when gene rearrangements are expected to affect 1-3% only of the investigated samples. The simulation for k=1, indeed, represents the condition of a rare case. When the number of outlier samples increases (ie. k=5, 10), other tools show better performance, as they have been designed for such broader search purposes.

3.4. Application to experimental data

To validate the tool on a real dataset, we tested the algorithm on about 500 kinase genes from Cancer Cell Line Encyclopedia (CCLE) gene expression dataset [83]. In this dataset 917 cell lines, belonging to 24 different tumor types, were profiled by microarrays and probe set intensities were calculated using the Robust Multi-array Average (RMA) and normalized by the quantiles method [88]. In this way, per each gene, outlier identification was performed within the same dataset and the same platform.

The method was able to correctly identify several kinases known to be overexpressed in specific cell lines among a tumor tissue type. Indeed, NTRK1 was correctly identified as highly expressed in KM12 colorectal cancer cell line (Figure 3.4 A).

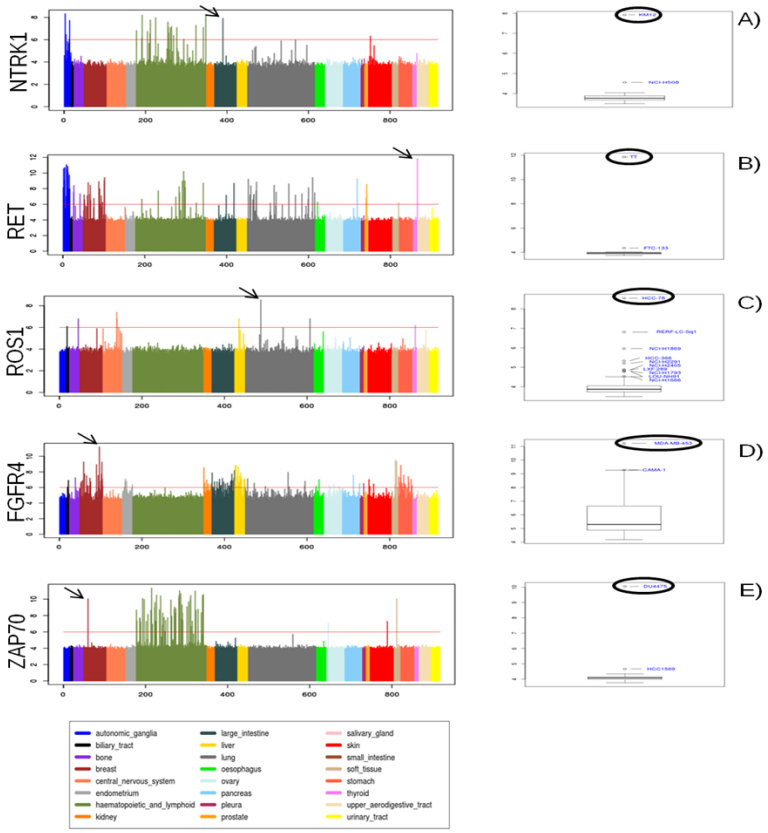

Figure 3.4: **Identification of known and new overexpressed kinases** (from Nuzzo A. et al. BMC Bioinformatics 2015). Left panel shows gene expression level of a selected kinase in 917 cancer cell lines belonging to 24 different tumor types (CCLE data) as histogram. Tumor types are reported in different colors. The boxplot of the tissue-specific distribution of the kinase is reported in the right panel. Outlier samples are reported as black circle. (**A**) NTRK1 is generally expressed in hematopoietic and lymphoid and autonomic ganglia. No expression is observed in large intestine (colon), apart in KM12 colorectal cancer cell line, highlited as outlier in this tissue; (**B**) RET tyrosine kinase is generally expressed in tissues such as autonomic ganglia, haematopoietic tissues, but no expression is observed in thyroid tumors. In this tissue a dramatic expression of RET can be detected in TT papillary tumor cell line only, assigned as outlier by the tool; (**C**) ROS1 tyrosine kinase is typically poorly expressed apart in colon where HCC-78 lung cancer cell line stands out as a clear outlier; (**D**) FGFR4 is highly expressed in few breast cancer cell lines, among those MDA-MB-453 breast cancer cell line appear as highly overexpressed. (**E**) ZAP-70 tyrosine kinase can be observed in haematopoietic and lymphoid tissues only. No expression in breast cancer cell

lines can be appreciated, with the exception of a significant overexpression of the gene in DU4475 breast cancer cell line.

As it has been already described in Chapter 2, NTRK1 is a tyrosine kinase typically not expressed in colorectal cancer tissue, however it become expressed and activated as consequence of a genomic rearrangement involving the C-terminal kinase catalytic domain of NTRK1, which is fused with a fusion partner driving its expression. NTRK1 is known to rearrange with TPM3, a ubiquitously expressed protein, in KM12 colorectal cancer cell line and its over-expression is the driver event of tumorigenesis and renders tumors sensitive to NTRK1 kinase inhibitors in preclinical models [32].

Similarly, an overexpression of the ROS1 tyrosine kinase could be detected in HCC-78 lung cancer cell line only, within lung tumor cancer cells (Figure 3.4 C). ROS1 is indeed overexpressed as consequence of a genomic translocation leading to the expression of a chimeric FIG-ROS gene [89]. Also, the system allowed for the identification of a significant overexpression of RET in the TT cell line, among thyroid papillary tumor cell lines (Figure 3.4 B). In this case RET overexpression and activation is not the result of a rearrangement, but it is a consequence of a mutation event that leads to the expression and activation of the kinase [90]. Moreover, FGFR4 overexpression was observed in MDA-MB-453 breast cancer cell line, among breast cancer cells (Figure 3.4 D). Also in this case the anomalous activation of the kinase is a consequence of the presence of a Y367C oncogenic mutation [91].

The analysis also permitted highlighting unexpected kinase over-expression. This is the case of ZAP-70, a tyrosine kinase normally expressed in T lymphocytes, where it plays a role in initiation and amplification of T-cell receptor signaling. ZAP-70 is indeed selectively expressed in tissues of lymphoid origin and is not typically observed in solid tumors [92]. The analysis showed the expected homogeneous expression distribution of ZAP-70 in lymphoid tissue with no outlier detection in the hematopoietic and lymphoid tissues, as no extreme outlier value stands out over the high variability context (Figure 3.4 E). On the other hand, using KAOS, an anomalous expression of the gene in a single breast cancer cell line, the DU4475 (Figure 3.4 E), was detected as an extreme outlier out of 56 breast cancer analyzed

samples. ZAP-70 expression was further investigated in the DU4475 cell line by western blot, using a specific antibody (sc-1526) against the C-terminal domain of the protein, confirming the high level of ZAP-70 expression also at protein level (see Figure 3.5). No significant expression of ZAP-70 could be appreciated in MCF7 breast cancer cell line, used as control.

Although the functional relevance of ZAP-70 overexpression in DU4475 breast cancer cell line needs further investigation, the application of KAOS to cell line models show how the system can easily detect a number of putative kinase fusion events, representing potential new drug targets. Being based on kinase overexpression as readout of the presence of a gene rearrangement, validation of the results by other experimental methods, such as PCR/sequencing, is needed.

The tool requires Java 1.6 or higher, MySQL 6 or higher, R 3.0 or higher and is freely available at:

www.nervianoms.com/downloads/Kaos_to_distribute.zip

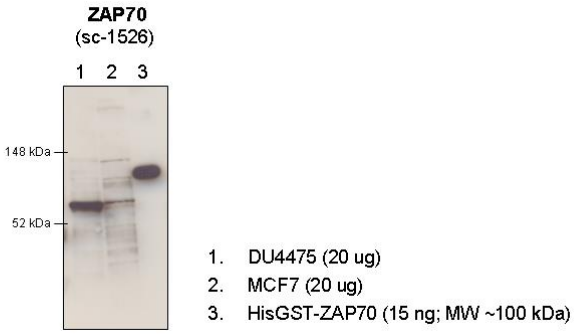

Figure 3.5: **Characterization by Western Blot analysis of ZAP-70 protein** (from Nuzzo A. et al. BMC Bioinformatics 2015)**.** Total cell lysates were subjected to Western Blot analysis using anti-ZAP-70 (sc-1526) goat polyclonal antibody raised against a peptide mapping at the C-terminus of ZAP-70. 1) DU4475; MCF7; HisGST-ZAP-70 recombinant protein (positive control)

Chapter 4

KING-REX: a combined targeted NGS approach for comprehensive kinome expression profiling[4]

The KAOS tool, described in the previous chapter, was conceived to detect kinase overexpression as readout of the presence of kinase gene fusions using whole transcriptome data.

While 'omics' analysis approaches produce huge amounts of molecular information, requiring substantial computational power for data storage and management, NGS targeted RNAseq approaches enable the analysis of focused portions of the transcriptome, with advantages over whole transcriptome sequencing in terms of reduced costs and simplicity of execution.

As it has been shown also in Chapter 2, typically commercial and custom based approaches have been developed for targeting very small kinase panels, but do not allow for whole kinome expression analysis so far.

For this reason, a comprehensive method for whole kinome screening has been implemented and will be described in the present chapter.

The system was called KING-REX (KINase Gene RNA EXpression). The approach consists of an experimental part, which was implemented by customizing the Illumina® TruSeq Targeted RNA expression kit

[4] The content of this chapter is in proceeding to be submitted to BMC Genomics as Carapezza G, Cusi C, Rizzo E, Raddrizzani L, Di Bella S, Somaschini A, Lupi R, Mutarelli M, Nigro V, Di Bernardo D, Magni P, Isacchi A, Bosotti R. *KING-REX: a custom end-to-end solution for kinome gene expression profiling.*

(TREx, [93]), followed by a bioinformatics analysis pipeline, which was implemented for the detection of kinase gene expression and for the identification of potential kinase fusion events.

KING-REX has been conceived for the profiling of the human kinome on small/medium scale Illumina® sequencers, requiring reduced computational resources in terms of storage space and data processing. For 319 kinases, paired assays and custom analysis pipeline features allow for the evaluation of 3'- and 5'-end transcript imbalances as readout for the prediction of gene rearrangements.

Validation tests on cell line models harboring known gene fusions demonstrated a comparable accuracy of KING-REX gene expression assessment as in whole transcriptome analyses, together with a robust detection of transcript portion imbalances in rearranged kinases, even in complex RNA mixtures or in degraded RNA.

These results support the use of KING-REX as a rapid and cost effective kinome investigation tool in the field of kinase target identification for applications in cancer biology and other human diseases.

4.1. Design of the KING-REX panel

NGS technologies currently offer the possibility to design panels of custom based assays for user defined sequences of interest. The focus of this work was the custom design of a targeted RNA procedure for the comprehensive gene expression analysis of the entire human kinome, intended for small/medium scale sequencers. Based on the Illumina® TruSeq Targeted RNA Expression (TREx) approach [93], enabling a custom definition of up to 1,000 assay panels, the KING-REX panel was assembled by selecting pre-designed assays with a specific targeting strategy, to combine the maximum capacity of the custom panel composition with the highest possible kinome coverage.

A comprehensive list of human protein kinases was compiled by integrating information from the currently available kinase resources [94]. For 514 unique genes, clearly annotated as protein kinases, genomic coordinates for 2230 kinase isoforms were retrieved from the UCSC database, providing an acceptable confidence level for transcript annotation [95]. Kinase domain coordinates were then obtained for

1716 protein kinase isoforms harboring the catalytic domain, as reported in the Superfamily database ([96], superfamily ID number 56112, containing the 'Protein kinases, catalytic subunit' subfamily of interest), and directly mapped onto UCSC transcripts. This information was visualized via the Integrated Genome Viewer (IGV) [97] to drive the assembly of a panel of 876 pre-designed amplicon-based assays from the TruSeq Targeted RNA Expression (TREx) kit (Illumina®, San Diego, CA, USA), selected to specifically target 512 human kinases, according to the schema depicted in Figure 4.1.

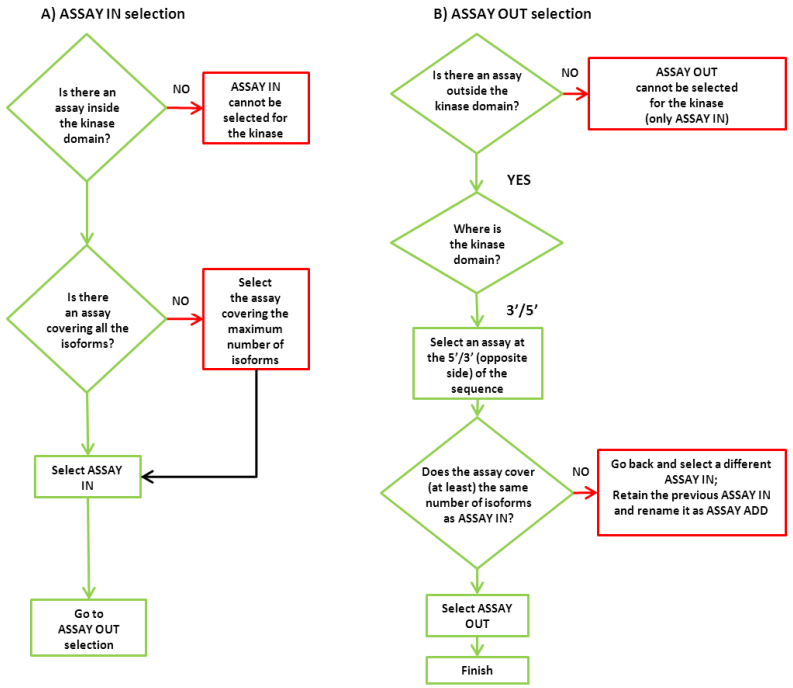

Figure 4.1 Flowchart of kinome assay selection (from Carapezza G. et al. submitted). (**A**) Selection process of ASSAY IN, targeting the kinase domain; (**B**) Selection process of ASSAY OUT, outside the kinase domain; definition of ASSAY ADD upon re-selection of ASSAY IN. All the selected pre-designed assays were derived from Illumina® DesignStudio Custom Assay Design Tool for use with the TruSeq Targeted RNA Expression (TREx) kit (Illumina®, San Diego, CA, USA).

Briefly, in the panel design a first assay for each kinase (ASSAY IN) was selected, targeting the kinase domain common to most of the

reported kinase isoforms, thus prioritizing the expressed druggable portion of the target sequences (Figure 4.1 A). Only for CDK3 and TNNI3K no pre-designed TREx assays were available within their respective kinase domains, so they were excluded from the panel. A second assay (ASSAY OUT) was then selected at the maximum sequence distance from ASSAY IN, targeting a region outside the kinase domain and covering the same isoforms encompassed by ASSAY IN (Figure 4.1 B). An ASSAY OUT fulfilling these criteria could be initially identified for 274 kinases. For other 45 kinases, it was not possible to cover all the isoforms targeted by ASSAY IN with a unique ASSAY OUT: for these, the ASSAY IN was redesigned based on a restricted number of isoforms, in order to allow for the selection of an ASSAY OUT according to the above criteria. In these latter cases, the initial assay encompassing the kinase domain covering the maximum number of isoforms was retained (and renamed as ASSAY ADD) in addition to the restricted ASSAY IN and respective ASSAY OUT (Figure 4.1 B). In this way, isoform coverage between ASSAY IN and ASSAY OUT could be balanced without penalizing the number of isoforms detectable for gene expression (ASSAY ADD). In total, an ASSAY OUT could be included for 319 kinases.

This experimental design enables a more robust evaluation of gene expression for those kinases that are probed with more than one independent assay; at the same time, the detection of uneven expression of kinases through the ASSAY IN and the corresponding ASSAY OUT levels can be exploited as readout to suggest the presence of potential gene rearrangements.

4.2. Implementation of a pipeline for the detection of kinase fusion events

Raw Count (RC) quantification was performed independently for ASSAY IN and ASSAY OUT using Bedtools Coverage tool (v. 2.22.0) [98]. A first normalization was performed using RLE [99] with default parameters. A further normalization step was applied to balance ASSAY IN and ASSAY OUT expression detection differences due to technical artifacts (ie. primer efficiency, degradation, RNA reverse transcription effects). In this step, the normalized counts (NC) of

ASSAY OUT are corrected with a scaling factor, calculated as the median value of the ratio between 'ASSAY IN' NC and the 'ASSAY OUT' NC along all samples n:

$$\text{Scaling factor} = \text{Median}(NC_{\text{ASSAY IN}n})/NC_{\text{ASSAY OUT}n})$$

EdgeR (v. 3.14.0) [100] was applied for the detection of differential expression between the ASSAY IN and ASSAY OUT for each kinase.

4.3. Evaluation of KING-REX performance in detecting kinase expression

The performance of KING-REX in gene expression quantification was evaluated by sequencing a panel of 10 colorectal cancer (CRC) cell lines, using KING-REX and whole transcriptome. The same bioinformatics data analysis pipeline was used for processing both KING-REX and transcriptome datasets.

Fastq files were aligned to the human reference genome (hg19) using STAR (v. 2.5.1b) [101]. Raw Count (RC) quantification was performed using RSEM tool (v. 1.2.30) [102]. Normalization was performed using RLE [99] with default parameters and Log2 transformed. Gene expression levels were reported as Log2 of Normalized Count (NC) for each kinase.

The R squared correlation between kinase expression levels measured in the two different protocols was calculated in all the analyzed cell lines and an extremely good concordance of KING-REX with whole transcriptome measurements was observed (average $R^2>0.8$).

Sensitivity, specificity and accuracy of KING-REX in detecting kinase signals with varying gene expression level thresholds was then evaluated, considering in-house whole transcriptome data on the same CRC cell lines as reference. Individual kinase raw counts for each cell line were obtained from: i) KING-REX data; ii) in-house transcriptome data, and iii) Cancer Cell Line Encyclopedia (CCLE [83]) transcriptome data.

Data were normalized using Upper Quartile Normalization (setting the 75th percentile to 1000) [103]. For each kinase in each cell line (cl), the

presence (P) or absence (N) of the kinase in the reference data was calculated with variable thresholds (thrs), ranging from 0.5 to 7:

$Kinase_{(reference;cl)} > thrs \Rightarrow P_{(cl)}$;
$Kinase_{(reference;cl)} \leq thrs \Rightarrow N_{(cl)}$.

For each kinase in each cell line the concordance of KING-REX with reference control was calculated as follows:

$Kinase_{(reference;cl)} > thrs$ and $Kinase_{(KING\text{-}REX;cl)} > thrs \Rightarrow TP_{(cl)}$;
$Kinase_{(reference;cl)} \leq thrs$ and $Kinase_{(KING\text{-}REX;cl)} \leq thrs \Rightarrow TN_{(cl)}$;
$Kinase_{(reference;cl)} \leq thrs$ and $Kinase_{(KING\text{-}REX;cl)} > thrs \Rightarrow FP_{(cl)}$;
$Kinase_{(reference;cl)} > thrs$ and $Kinase_{(KING\text{-}REX;cl)} \leq thrs \Rightarrow FN_{(cl)}$.

From these data, the total number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) values in all tested cell lines was calculated for each kinase. Sensitivity, specificity and accuracy were calculated for each kinase as follows:

$Sensitivity_{kinase} = TP/P$;
$Specificity_{kinase} = TN/N$;
$Accuracy_{kinase} = (TP+TN)/(P+N)$

Sensitivity and specificity values ranged from about 80% to 90%, with an overall very high accuracy, regardless of the selected threshold. A slight decrease in specificity could be observed at the lowest threshold only (Table 4.1, Threshold < 1). The same analysis was also performed using Cancer Cell Line Encyclopedia (CCLE) RNAseq data as reference [83], obtaining comparable results (Table 4.1).

Table 4.1 Sensitivity, specificity and accuracy of KING-REX on a panel of CRC cell lines (from Carapezza G. et al. submitted)

	KING-REX vs. in-house transcriptome data			KING-REX vs. CCLE transcriptome data		
Threshold	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
0.5	92.7	64.6	92.8	90.0	78.1	92.0
1	92.5	71.3	93.6	90.3	83.0	93.1
2	92.5	81.0	93.6	90.8	83.6	93.0
3	91.0	86.3	92.9	90.7	90.4	93.0
4	90.9	89.5	92.8	89.3	89.2	92.7
5	89.2	88.1	91.6	88.7	89.0	91.6
6	86.7	82.5	89.4	86.4	85.8	89.5
7	85.7	79.2	86.9	81.5	83.0	86.1

Sensitivity, specificity and accuracy of KING-REX vs. in-house or published CCLE [83] transcriptome data on a panel of colorectal cancer cell lines, with varying detection thresholds.

Next, the comparative analysis was extended to a more heterogeneous sample panel by selecting cell lines of different tissue origins (BT-474, breast; HPAC, pancreas; K-562, leukemia; KARPAS 299, lymphoma; NCI-H716, large intestine; SNU-1079, biliary tract; U-118 MG, nervous system) for KING-REX analysis and using publicly available CCLE whole transcriptome RNAseq data as reference [83]. Sensitivity, specificity and accuracy values were again comparable to the ones observed with the CRC cell line panel (Table 4.2).

Table 4.2 Sensitivity, specificity and accuracy of KING-REX on a heterogeneous panel of cancer cell lines (from Carapezza G. et al. submitted)

	KING-REX vs. CCLE transcriptome data		
Threshold	Sensitivity	Specificity	Accuracy
0.5	90.1	73.9	91.1
1	89.4	76.6	90.9
2	89.3	85.1	91.3
3	89.0	87.4	91.0
4	87.0	90.6	90.7
5	86.8	91.0	90.5
6	84.6	89.7	88.9
7	80.4	87.7	86.3

Sensitivity, specificity and accuracy of KING-REX vs. CCLE transcriptome data calculated on a heterogeneous panel of cancer cell lines, with varying detection thresholds.

4.4. Evaluation of KING-REX performance in predicting kinase fusions

In the KING-REX design, the paired ASSAY IN and corresponding ASSAY OUT, located in opposite 5' and 3' transcript ends, support the identification of potential gene rearrangements for 319 kinases. For this specific purpose, the pipeline described in paragraph 4.2, evaluating imbalances between the kinase ASSAY IN and ASSAY OUT signals, was applied. The ability of KING-REX to identify imbalanced 5' vs. 3' signals was tested in five human cancer cell lines, harboring well known kinase gene fusions: KARPAS 299, KM-12, LC-2/ad, U-118 MG and NCI-H716. KARPAS 299 is a T-cell lymphoma cell line carrying the NPM-ALK gene fusion [104]; KM12 is a colorectal cancer cell line expressing the chimeric TPM3-TRKA protein [38]; LC-2/ad is a lung adenocarcinoma cell line harboring a CCD6-RET fusion [105]; U-118 MG is a glioblastoma cell line characterized by the presence of a

FIG(GOPC)-ROS1 rearrangement [106]. NCI-H716 colorectal cancer cell line was included as a control, harboring an amplified full length FGFR2 kinase, whose hyper-activation is due to gene amplification and not to the presence of a concomitant fusion at the FGFR2 C-terminal with COL14A1 gene, conserving an intact kinase sequence [107]. All these cell lines were sequenced in duplicate using KING-REX and analyzed with the kinase fusion event detection pipeline using ASSAY IN and ASSAY OUT measurements (Figure 4.2). In the pipeline, after the standard data normalization step (Figure 4.2 A), results are further processed with a second normalization step, introduced to balance for possible technical differences between ASSAY IN and ASSAY OUT signals, related to primer efficiency, differential end degradation and/or RNA reverse transcription performance. This resulted in the expected clustering of IN and OUT assays belonging to the same cell line (Figure 4.2 B).

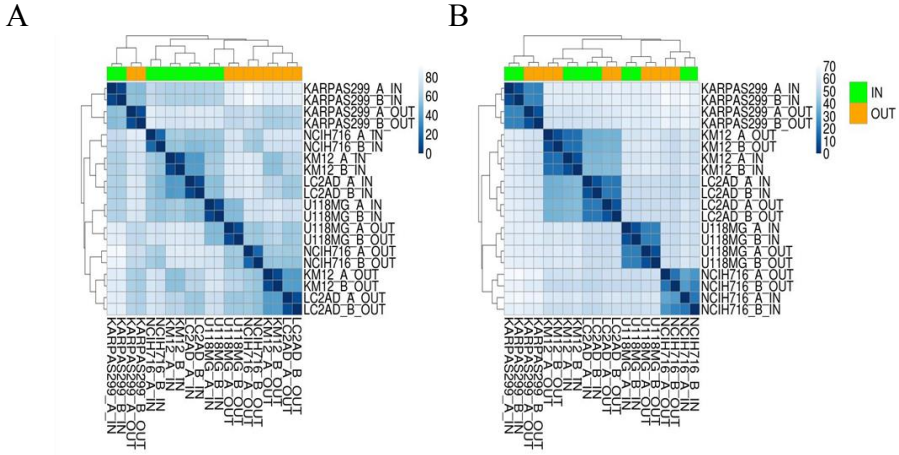

Figure 4.2 Clustering of cancer cell lines before and after the normalization step (from Carapezza G. et al. submitted)**.** Cluster analysis of a heterogeneous panel of 5 cancer cell lines, tested in duplicate with KING-REX, before (**A**) and after (**B**) the second normalization step in the pipeline for potential kinase fusion event detection. ASSAY_IN expression values for each sample are annotated in green, while ASSAY_OUT in orange. The blue shading indicates the Euclidean distance between the expression values of two samples (cell lines), ranging from dark blue (high similarity) to light blue (low similarity).

Imbalanced ASSAY IN and ASSAY OUT signals in all the kinases involved in known gene rearrangements were clearly detected by differential gene expression analysis (Table 4.3). No imbalanced signals could be detected in the negative control cell line NCI-H716, harboring a full length FGFR2 kinase, covered by both ASSAY IN and OUT for FGFR2 (Table 4.3).

Table 4.3 Prediction of kinase fusion events (from Carapezza G. et al. submitted)

Sample	Kinase	FC	p-value	IN_A	IN_B	OUT_A	OUT_B
KARPAS 299	ALK	15.83	2.00E-308	12.8	12.9	0.0	0.0
KM-12	NTRK1	10.58	3.94E-236	12.4	12.1	2.3	1.6
LC-2/ad	RET	10.79	6.23E-219	11.8	11.7	1.6	1.6
U-118 MG	ROS1	13.85	4.58E-174	10.9	10.7	0.0	0.0
NCI-H716	FGFR2	-	-				

Imbalanced kinases detected by KING-REX analysis in a panel of cancer cell lines harboring known kinase gene fusions. (FC= log2 fold change between ASSAY IN and ASSAY OUT; IN = ASSAY IN expression value for duplicates A and B, OUT = ASSAY OUT expression value for duplicates A and B). For differential expression below EdgeR p-value default threshold, no data are reported in the table.

4.5. KING-REX limits of detection in gene expression evaluation

The limits of detection of KING-REX, both in terms of gene expression measurement and of gene fusion prediction, were then explored using 7 samples derived from KARPAS 299 and U-118 MG, both harboring a kinase gene fusion (NPM-ALK in KARPAS 299 and FIG(GOPC)-ROS1 in U-118 MG). The RNA from the two cell lines was mixed in different proportions: 100%-0%, 87.5%-12.5%; 75%-25%, 50%-50%, 25%-75%, 12.5%-87.5%, 0%-100%, to simulate different tissue heterogeneity levels as found in clinical cancer samples. Duplicate samples for each mix were then subjected to KING-REX analysis; sequencing results for technical duplicates clearly clustered according to the relative dilution proportions (Figure 4.3).

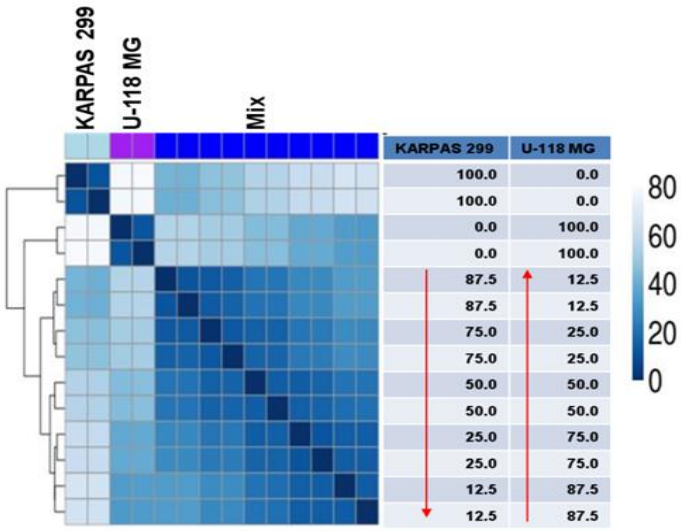

Figure 4.3 Distance matrix analysis of KING-REX mixed samples (from Carapezza G. et al. submitted)**.** Distance matrix of KING-REX analysis results for technical duplicates of RNA from two cell lines (KARPAS 299 and U-118 MG), mixed in different percent dilution proportions as indicated on the right side of the graph. The blue shading indicates the Euclidean distance between the expression values of two samples (cell line mixtures), ranging from dark blue (high similarity) to light blue (low similarity).

Kinases expressed exclusively in U-118 MG or KARPAS 299, i.e. not detected in the other cell line (average gene expression value <1), were then investigated. The disappearance of gene expression signals with the variable increments of KARPAS 299 or U-118MG background respectively was evaluated. A clear signal could be observed after incremental dilutions for all the U-118 MG or KARPAS 299 unique kinases, down to the lowest tested concentration (12.5%), regardless of their basal gene expression level (Table 4.4).

Table 4.4 KING-REX gene expression values of U-118 MG or KARPAS 299 unique kinases in serially diluted cell line RNA mixtures (from Carapezza G. et al. submitted)

	U-118MG												
	100.0%	100.0%	87.5%	87.5%	75.0%	75.0%	50.0%	50.0%	25.0%	25.0%	12.5%	12.5%	0.0%	0.0%
FGFR1	13,1	13,0	13,9	13,9	13,4	13,5	12,8	12,8	11,5	11,8	10,7	10,8	0,0	0,0
KIT	13,1	13,2	12,4	12,3	12,2	12,1	11,5	11,5	10,6	10,4	9,3	9,3	0,0	0,9
ROR1	12,9	12,8	12,9	12,8	12,5	12,5	11,8	11,9	10,7	10,8	9,9	9,7	0,0	0,0
CDK14	12,6	12,4	12,3	12,3	11,9	11,9	11,3	11,3	10,2	10,3	9,2	9,3	0,8	0,9
EPHA5	11,3	11,5	10,1	10,3	10,0	9,8	9,4	9,3	8,2	8,3	7,2	7,1	0,0	0,0
ACVR2A	11,0	10,7	10,5	10,3	10,3	10,2	9,7	9,5	8,5	8,0	7,3	7,3	1,9	0,0
DCLK2	10,8	10,7	10,2	10,1	9,9	9,9	9,1	9,4	8,1	8,2	7,2	7,3	0,0	0,9
LRRK2	10,2	10,1	9,2	9,1	9,1	8,9	8,4	8,3	7,3	7,4	6,4	6,4	0,8	0,0
EPHA4	10,0	10,0	10,5	10,5	10,2	9,9	9,5	9,4	8,4	8,6	7,6	7,6	0,8	0,0
PKDCC	9,8	10,0	9,9	9,9	9,3	9,7	9,0	9,0	7,8	7,9	6,9	6,5	0,0	0,0
CDKL5	9,8	10,0	9,9	9,8	9,3	9,5	8,8	8,8	7,6	7,9	6,4	6,9	0,8	0,9
EPHA3	9,7	9,8	9,0	9,0	8,5	8,4	8,1	8,0	7,0	7,2	6,1	6,4	0,8	0,0
PRKG1	9,4	9,2	9,6	9,6	9,1	9,4	8,7	8,5	7,3	7,4	6,0	6,5	0,8	0,0
NUAK1	8,6	8,0	8,0	7,7	7,4	7,6	7,1	7,0	5,9	5,9	5,5	5,3	0,0	0,9
SPEG	7,8	8,0	8,4	8,1	7,6	7,8	7,0	7,2	5,5	6,1	4,9	5,4	0,0	0,0
DCLK1	7,7	7,7	8,0	8,0	7,6	7,6	6,8	7,1	5,6	5,6	5,1	5,6	0,0	0,9
NPR1	7,3	7,2	7,2	6,9	6,6	6,4	5,9	5,6	5,4	4,0	2,9	3,4	0,0	0,0
PDK4	7,3	7,4	7,0	7,0	6,1	6,3	6,3	6,2	4,7	4,3	3,9	4,7	0,0	0,0
MAP2K6	7,1	7,2	6,1	6,6	6,3	6,1	6,0	6,0	4,1	5,6	4,7	4,1	1,2	0,0
PRKD1	6,1	6,5	6,2	5,9	5,8	6,0	5,4	4,9	4,8	3,7	3,5	3,5	0,8	0,0
NUAK2	5,4	5,3	4,3	4,4	4,0	4,6	3,1	4,4	3,2	0,0	2,1	2,2	0,0	0,9
EPHA7	5,2	5,9	5,1	4,5	4,4	5,3	3,9	2,9	3,2	3,3	2,7	3,2	0,0	0,0
ACVR1B	5,1	5,3	4,7	4,5	4,1	4,4	3,3	4,2	2,8	2,3	2,1	2,2	0,0	0,0
ACVR1C	5,1	5,2	5,8	5,3	5,6	5,9	4,9	4,7	4,7	4,1	3,5	4,2	0,0	1,4
STK32B	4,8	5,6	5,1	5,3	4,9	4,4	4,5	4,3	3,7	2,8	3,1	1,7	0,0	0,0

	KARPAS 299													
	100.0%	100.0%	87.5%	87.5%	75.0%	75.0%	50.0%	50.0%	25.0%	25.0%	12.5%	12.5%	0.0%	0.0%
HCK	13,5	13,6	13,6	13,7	13,6	13,5	13,0	13,0	12,2	12,2	11,2	11,2	0,0	0,0
ROR2	11,0	11,0	10,8	10,8	10,6	10,5	10,0	10,1	9,2	9,0	8,3	8,2	0,0	0,0
PRKCQ	10,9	10,9	10,6	10,5	10,2	10,4	9,8	9,7	8,8	8,8	7,9	8,1	1,9	0,0
EPHA10	6,7	6,7	6,0	6,3	5,9	5,5	4,7	5,2	5,1	4,0	4,1	2,5	0,9	0,0
MAK	5,9	6,4	5,3	5,4	5,2	5,3	4,9	5,0	4,7	2,8	3,8	3,5	0,9	0,0
ITK	5,0	5,4	5,2	5,1	4,7	5,1	3,7	3,8	2,1	3,8	2,8	2,5	1,9	0,0

List of kinases above a mean signal of 5 (Log2NC) in only one of the two tested cell lines (U-118 MG, left panel; or KARPAS 299, right panel) and below a mean signal of 1 (Log2NC) in the other one, with respective gene expression values measured in technical duplicates of RNA mixtures from the two cell lines (KARPAS 299 and U-118 MG) in different proportions (100%-0%; 87.5%-12,5%; 75%-25%; 50%-50%; 25%-75%; 12.5%-87.5%; 0%-100%). The blue shading reflects variations of gene expression values (Log2NC), ranging from 0 (white) to 14 (dark blue).

To evaluate the linearity of kinase gene expression variation within the serially diluted sample set, the measured expression values for each kinase in all the mixed samples was compared versus a 'theoretical' value calculated for each dilution. Gene expression values measured for each kinase in the 100% KARPAS 299 or U-118 MG samples were used to infer the 'theoretical' expression values expected in each dilution mix, according to the following formula:

$$\text{Log2}[NC_{\text{KARPAS 299}}*P + NC_{\text{U-118 MG}}*(1-P)]$$

where NC is the Normalized Counts for each kinase and P corresponds to the serial dilution factor (0.875, 0.75, 0.5, 0.25, or 0.125, respectively). A scatter plot between the measured and 'theoretical' sets of data was generated, and R^2 correlation coefficient was calculated.

Despite the complexity of the assay panel composition, theoretical and measured kinase gene expression levels showed high concordance at all dilution levels, as demonstrated by the observed linearity of the reported scatter plot, with an R^2=0.98 (Figure 4.4), supporting the robustness of the KING-REX kinase expression profiling approach.

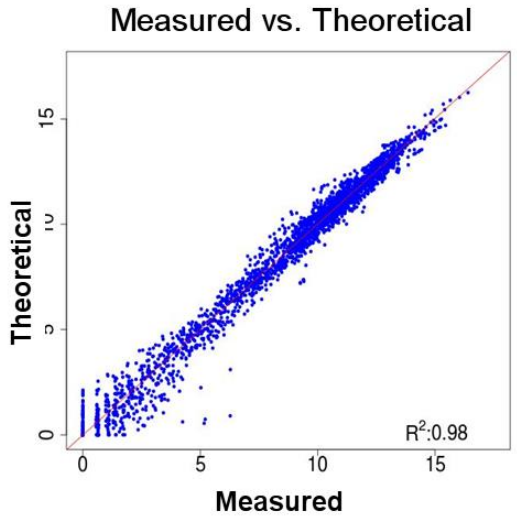

Figure 4.4 Comparison between measured and theoretical kinase gene expression values (from Carapezza G. et al. submitted)**.** Kinase gene expression values measured with KING-REX within a serially diluted sample set of KARPAS 299 and U-118 MG RNAs, mixed in different proportions (87.5%-12.5%; 75%-25%, 50%-50%, 25%-75%, 12.5%-87.5%,), plotted vs. 'theoretical' expression values, calculated for each kinase and for each dilution factor; the respective correlation R2 value is displayed.

4.6. KING-REX limits of detection in gene fusion prediction

Next the limits of detection of fusion events was investigated, by evaluating the ability of KING-REX to correctly predict gene fusions in the serially diluted samples described above, based on ASSAY IN and ASSAY OUT imbalances for the rearranged kinases present in the two cell lines. In U-118 MG, the 3' portion of the ROS1 kinase is detected at high level due to the presence of the FIG-ROS1 gene fusion, while

both 5' and 3' ROS1 signals are undetectable in KARPAS 299, thus representing an ideal background to determine ROS1 detection limits without confounding factors (Figure 4.5 A). The ability of KING-REX to detect the ROS1 5' vs. 3' imbalance in a null background is maintained throughout all the dilutions, down to the experimentally tested limit of 12.5% (Table 4.5).

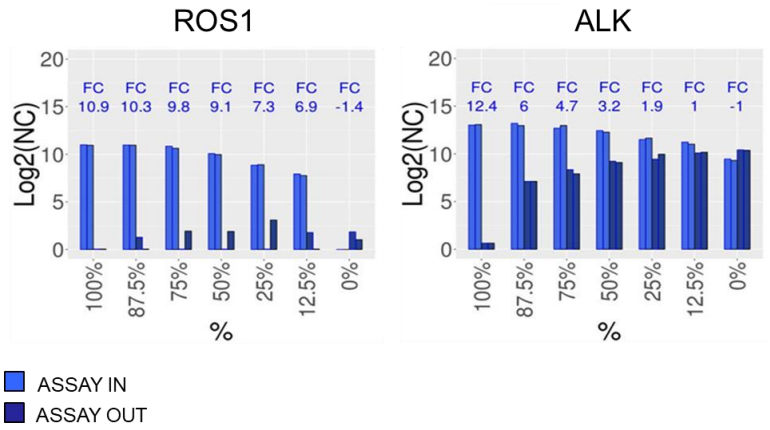

Figure 4.5 Detection of gene fusion events in diluted samples (from Carapezza G. et al. submitted)**. (A)** KING-REX log2(NC) expression values for ROS1 ASSAY IN (light blue) and ASSAY OUT (dark blue) in U-118 MG, diluted in different percentages of KARPAS 299 sample background; (**B**) KING-REX log2(NC) expression values for ALK ASSAY IN (light blue) and ASSAY OUT (dark blue) in KARPAS 299, diluted in different percentages of U-118 MG sample background. Samples are in duplicate.

Table 4.5 Prediction of kinase fusion events at different dilution levels (from Carapezza G. et al. submitted)**.**

KARPAS 299 (ALK)	U-118 MG (ROS1)	Kinase	FC	p-value	IN_A	IN_B	OUT_A	OUT_B	Class
100.0	0.0	ALK	15.8	2.00E-308	12.8	12.9	0.0	0.0	TP
87.5	12.5	ALK	6.5	0	13.0	12.8	6.4	6.3	TP
87.5	12.5	ROS1	7.4	3.46E-078	7.7	7.6	1.6	0.0	TP
75.0	25.0	ALK	5.3	3.11E-124	12.6	12.8	7.5	7.2	TP
75.0	25.0	ROS1	6.8	1.60E-075	8.7	8.7	0.0	3.0	TP
50.0	50.0	ROS1	9.6	3.09E-194	9.8	9.8	0.0	1.6	TP
50.0	50.0	ALK	3.7	5.46E-105	12.2	12.1	8.5	8.3	TP
25.0	75.0	ROS1	10.3	6.40E-167	10.6	10.3	0.0	1.6	TP
25.0	75.0	ALK	2.35	6.75E-028	11.25	11.36	8.68	9.13	FN
12.5	87.5	ROS1	11.3	2.49E-152	10.7	10.7	1.0	0.0	TP
12.5	87.5	ALK	1.40	2.17E-015	10.95	10.77	9.31	9.34	FN
0.0	100.0	ROS1	13.9	4.58E-174	10.9	10.7	0.0	0.0	TP

Imbalanced 5' and 3' kinase signals detected by KING-REX analysis in different dilution mixtures of KARPAS 299 and U-118 MG cell lines. (FC= log2 fold change between ASSAY IN and ASSAY OUT; IN = ASSAY IN expression value for duplicates A and B, OUT = ASSAY OUT expression value for duplicates A and B).

KARPAS 299, expressing an ALK gene fusion (NPM-ALK), was diluted in the U-118 MG cell line background, in this case expressing a full length ALK, thus representing a more frequent real-life scenario, simulating the case of a tumor mixed with normal adjacent tissue and/or infiltrating lymphocytes, which might express full length ALK (Figure 4.5 B). In this case, KING-REX could clearly detect the presence of imbalanced ALK gene expression in KARPAS 299 mixtures above 50% proportion (Table 4.5).

The described cases are only two of many possible experimental scenarios, indicating that the limit of detection of the system is higher when the fusion gene is expressed at lower level or when the background full length kinase is highly expressed. Along with these experimental results, the 'theoretical' lower limits of detection of the system was simulated at different expression levels of a kinase gene fusion (GE_1) in a background with varying expression of the full length kinase (GE_2), by generating a synthetic dataset with different combinations of expression levels. A 'theoretical' dilution matrix was created with virtual kinase gene expression levels (GE), to simulate variable tissue mixture conditions in which cells containing a fusion

kinase (sample 1, S1) are diluted, or 'contaminated', with variable proportions of cells containing a full length kinase (sample 2, S2).

It was assumed that, for fusion kinase genes in S1:

$ASSAY\_IN_{S1} = GE_1$
$ASSAY\_OUT_{S1} = 0$

where GE_1 = gene expression of virtual gene fusion in S_1, ranging from 15 to 7;

and for full length kinase genes in S2:

$ASSAY\_IN_{S2} = GE_2$
$ASSAY\_OUT_{S2} = GE_2$

where GE_2 =gene expression of the full length gene in S_2 ranging from GE_1 value to 0. The two artificial S1 and S2 datasets were subjected to KING-REX analysis for potential kinase fusion detection, after generating 'virtually mixed' samples with different dilution proportions of S1 and S2, using the following formula:

$$Log_2[NC_{S1}*P + NC_{S2}*(1-P)]$$

where:

P is the percentage of the sample S_1 and 1-P is the percentage of sample S_2; P ranges from 100% to 0% in steps of 6.25%;

NC is the Normalized Counts for each kinase ASSAY IN and ASSAY OUT (including the virtual kinase) obtained with the KING-REX pipeline for potential kinase fusion event detection. For each combination of fixed GE1 and GE2 values and for variable P, P_{n-1} was established as the minimum allowed P, if at P_n the gene fusion was not detected by KING-REX pipeline analysis. The matrix in Table 4.6 shows the minimum allowed P of sample S1, containing the fusion kinase gene, required in a S1/S2 sample mixture for successful gene fusion detection by KING-REX analysis for each combination of virtual GE1 and GE2 values.

Results in Table 4.6 show the minimum percentage of sample containing a kinase gene fusion ('sample 1') theoretically required by

the KING-REX system to successfully detect a gene fusion event with varying background levels of full length kinase in a contaminating sample ('sample 2'). This simulated dataset was then compared to the available experimental results to verify the theoretical predictions, at least for the tested conditions.

Indeed, in the case of FIG-ROS1, expressed in U-118 MG cell line and diluted in a null ROS1 background (KARPAS 299), the simulation showed that a 'theoretical' limit of detection of 6.25% might be reached (Table 4.6, highlighted in bold), i.e. below the lowest experimentally tested 12.5% dilution factor (Tab 4.5). Similarly, in the case of KARPAS 299, expressing the NPM-ALK gene fusion diluted in the ALK full length U-118 MG background, a 'theoretical' detection limit of 31.25% might be reached (Table 4.6, highlighted in bold), i.e. between the experimentally tested 50% (positive for fusion detection) and 25% (negative for fusion detection) dilution factors (Table 4.5), supporting experimental validation conclusions and suggesting an even higher sensitivity of the system.

Table 4.6 Theoretical limit of detection for kinase gene fusion events (from Carapezza G. et al. submitted)**.**

		Gene expression value of kinase gene fusion in sample 1 (GE_1)								
		GE_1=15	GE_1=14	GE_1=13	GE_1=12	GE_1=11	GE_1=10	GE_1=9	GE_1=8	GE_1=7
Gene expression value of FL kinase in sample 2 (GE2)	GE_1+0	81.25	81.25	81.25	81.25	87.50	87.50	87.50	87.50	93.75
	GE_1-1	62.50	62.50	62.50	62.50	68.75	75.00	75.00	75.00	93.75
	GE_1-2	50.00	50.00	43.75	43.75	50.00	56.25	62.50	68.75	87.50
	GE_1-3	31.25	31.25	**31.252**	37.50	37.50	37.50	43.75	62.50	81.25
	GE_1-4	18.75	18.75	18.75	25.00	31.25	31.25	37.50	50.00	75.00
	GE_1-5	12.50	12.50	12.50	18.75	18.75	25.00	31.25	43.75	68.75
	GE_1-6	6.25	6.25	12.50	12.50	12.50	18.75	31.25	43.75	68.75
	GE_1-7	6.25	6.25	6.25	12.50	12.50	18.75	25.00	37.50	62.50
	GE_1-8	6.25	6.25	6.25	6.25	12.50	12.50	25.00	37.50	
	GE_1-9	6.25	6.25	6.25	6.25	12.50	12.50	25.00		
	GE_1-10	6.25	6.25	6.25	6.25	6.25	12.50			
	GE_1-11	6.25	6.25	6.25	6.25	**6.251**				
	GE_1-12	6.25	6.25	6.25	6.25					
	GE_1-13	6.25	6.25	6.25						
	GE_1-14	6.25	6.25							
	GE_1-15	6.25								

Theoretical limit of detection (expressed as the percentage of sample containing the gene fusion vs. the background full length (FL) sample) for the

detection of kinases involved in gene fusions with varying expression levels of the fused kinase domain and of the full length WT kinase (background), calculated using -ASSAY IN and ASSAY OUT values as described in M&M. GE_1= gene expression level of the fused kinase in sample1; GE_2= gene expression level of the full length kinase (background) in sample 2.

[1] theoretical limit of detection calculated in the same case of ROS1 kinase in U-118 MG (GE_1 = 11) diluted in KARPAS 299 not expressing ROS1 (GE_2 = 0)

[2] theoretical limit of detection calculated in the same case of ALK kinase in KARPAS 299 (GE_1 = 13) diluted in U-118 MG expressing full length ALK (GE_2 = 10).

4.7. KING-REX performance with degraded RNA

The experimental performance of KING-REX was then evaluated on degraded RNA. RNA was extracted from KM-12 cell line and artificially heat degraded by treating the RNA in aqueous solution at 90°C for different times (0, 5, 10, 20, 40, 60 minutes or 5 hours). RNA quality and integrity was evaluated by assessing RIN and DV200% parameters for all samples with an Agilent Bioanalyzer. As expected, RNA quality was inversely proportional to the time of exposure to high temperature (Table 4.7). Heat degraded samples obtained at four representative time points were selected (time = 0, 10, 20 and 60 minutes) and subjected to KING-REX library preparation in duplicate, followed by sequencing on Illumina MiSeq.

Library performance was evaluated in relation to the RNA degradation level. A trend in the reduction of the library concentration with the decrease in RIN and DV200% parameters was observed. KING-REX sequencing performance appeared acceptable in samples with low RIN but still high DV200% values, since an adequate number of reads was still obtained (PF, passing filter reads in Table 4.7); sequencing analysis of the 60-minute sample, characterized by a DV200% of 10%, indicating extreme degradation, yielded a poor number of PF, in line with Illumina® recommendations for the TREx protocol, suggesting to process samples preferably with a DV200%>30% [108].

In terms of gene fusion detection, an imbalanced ASSAY IN/OUT NTRK1 signal in the KM-12 cell line could be appreciated with significant p-value at the early time points (time = 0, 10 and 20 minutes), while this was not possible in sample duplicates with 60 minutes exposure to heating conditions, due to the generally poor quality of these highly degraded samples (DV200%<10%). These results demonstrated that KING-REX analysis can be applied also to partially degraded samples.

Table 4.7 KING-REX performance with heat degraded RNA (from Carapezza G. et al. submitted)

Time (minutes)	RNA quality		Library preparation and sequencing		Gene expression profiling and fusion detection	
	RIN	DV200% (if RIN<8)	Conc (ng/µl)	PF Reads (10^6)	R^2 vs. T=0	p-value
0	7.6	94	4.2 6.6	2.14 2.00	1.00	2.97E-144
10	3	90	5.5 6.9	1.87 1.87	0.99	3.28E-227
20	2.2	69	2.8 0.9	1.80 0.70	0.97	2.81E-081
60	2.6	10	0.4 0.4	0.04 0.03	0.78	5.96E-003

RNA parameters, library concentration and number of reads obtained from the sequencing of heat degraded KM-12 RNA samples at different time points. Correlation of kinase gene expression profiles at different time point vs. time=0 (T=0) and p-value relative to the detection of the TPM3-NTRK1 kinase gene fusion are reported (PF=reads passing filter).

KING-REX therefore represents a custom targeted RNA approach suitable for the gene expression screening of a comprehensive human kinome panel on Illumina® MiSeq or NextSeq platforms, retaining the possibility to infer the presence of gene fusions for a wide number of kinases (319), using paired assays located within and outside the catalytic domain. KING-REX can correctly and robustly detect kinase rearrangements even in complex background mixtures or in heat-degraded RNA, based on the evaluation of the imbalanced measured expression ratio for paired kinase assays. It is currently the largest

targeted approach for expression analysis of kinases, which could be used as a rapid and cost-effective investigation tool in cancer biology. Thus KING-REX represents a useful setup for the comprehensive analysis of the kinome in cancer or other diseases, for applications in the field of the identification of novel, putative kinase targets.

Chapter 5

Application of KING-REX to the genomic characterization of a new chordoma cancer cell line[5]

This chapter describes the application of the KING-REX to the genomic characterization of a new cell line, called Chor-IN-1, established at NMS in collaboration with Istituto Nazionale dei Tumori of Milano (INT) from a recurrent sacral chordoma tumor.

Chordomas are rare, slowly growing bone tumors with high medical need and no approved therapy. The need for effective therapies is extremely high, but the paucity of established chordoma cell lines has limited preclinical research. These tumors express several activated

[5] The content of this chapter is published in:

- Bosotti R, Magnaghi P, Di Bella S, Cozzi L, Cusi C, Bozzi F, Beltrami N, Carapezza G, Ballinari D, Amboldi N, Lupi R, Somaschini A, Raddrizzani L, Salom B, Galvani A, Stacchiotti S, Tamborini E, Isacchi A. *Establishment and genomic characterization of the new chordoma cell line Chor-IN-1*, Sci Rep.;7(1):9226 (2017)
- Magnaghi P, Salom B, Cozzi L, Amboldi N, Ballinari D, Tamborini E, Gasparri F, Montagnoli A, Raddrizzani L, Somaschini A, Bosotti R, Orrenius C, Bozzi F, Pilotti S, Galvani A, Sommer J, Stacchiotti S, Isacchi A. *Afatinib Is a New Therapeutic Approach in Chordoma with a Unique Ability to Target EGFR and Brachyury.* Mol Cancer Ther.17(3):603-613 (2018)

tyrosine kinase receptors, which prompted attempts to treat patients with tyrosine kinase inhibitors. Although clinical benefit was observed in Phase II clinical trials with imatinib and sorafenib, and sporadically also with EGFR inhibitors, therapies evaluated to date have shown modest activity.

The kinome profile of Chor-IN-1 was performed using the KING-REX approach and compared to the one of other available chordoma cell lines, with the aim of searching for overexpressed kinases, which could represent targets for new drugs or drugs already approved in other clinical settings and therefore with immediate therapeutic potential for chordoma patients.

5.1. Clinical background

Chordomas are malignant bone tumors that arise along any region of the axial skeleton, more frequently at sacro-coccygeal or skull-base ends [109, 110]. They are rare tumors, accounting for 1–4% of bone cancers, that occur more frequently in males than females and have a peak of incidence at around 60 years of age, although adolescents and children can also be affected [111]. Despite being slow progressing, chordomas are locally aggressive and invasive, and can also spread distally generating metastases in soft tissues, liver, lung, lymph-nodes and skin [109, 110]. No therapeutic options have so far revealed to be efficacious for this indication, which is highly resistant to conventional chemotherapy, and consequently no approved standard of care exists. Extensive surgery and/or high dose radiotherapy is currently used to treat the disease, but tumor location, frequently close to cranial or lombo-sacral nerves and blood vessels, makes the achievement of negative surgical margins very challenging [111-113]. In most cases local relapses occur and represent the main cause of death for patients [114, 115]. Chordomas therefore retain a strong unmet medical need, and new efficacious therapies are urgently needed. Chordomas are thought to originate from remnants of the notochord, a transient structure required for patterning the surrounding tissues in all chordate embryos, that disappears later in development and is absent in adults. Among other developmental signals, notochordal cells express the transcription factor brachyury, a key molecule for mesoderm

specification that is silenced in adult tissues [116]. Brachyury re-expression in notochord remnants is believed to play a major role in chordoma onset and maintenance and its expression is considered the main distinctive molecular marker of chordomas [117]. Brachyury expression is therefore considered a mandatory feature for the validation of chordoma cell lines [118]. Chordomas consistently express also other molecular markers, such as epithelial membrane antigen (EMA), vimentin, cytokeratin 19, CD24v and CAM5.23 [116-119]. Moreover, variable expression and activation of several receptor tyrosine kinases (RTKs) and downstream signaling molecules have been reported, with MET, PDGFRβ and EGFR as the most widely expressed, and HER2, KIT and VEGFR also expressed [120-124], therefore the detection of abundantly/differentially expressed kinases in chordoma cell lines might represent a convenient strategy for the identification of potential new pharmacological targets in this disease.

Very few bona fide chordoma cell lines have been available until recently [125-127], limiting the identification of relevant targets and the development of effective drugs. For this reason, the establishment, genomic characterization and validation of a new chordoma cell line represents an important step towards a better comprehension of this rare disease.

5.2. Case report

The surgical sample was obtained from a patient initially diagnosed with sacral chordoma. The patient refused surgery and received imatinib and radiotherapy. After three years, the patient experienced local subcutaneous progression at lombo-sacral level, received again imatinib and, upon further progression, also metformin. A tumor biopsy was performed when the patient finally underwent surgery. A sacral nodule of 2 cm of diameter, invading macroscopically the surrounding soft tissues, was surgically excised. All methods were performed in accordance with the relevant guidelines and regulations.

The patient gave his informed consent to study his tumor, followed by the approval of the Fondazione IRCCS Istituto Nazionale dei Tumori (Milan, Italy) Ethical Committee for the genotypic and phenotypic

characterization of the cell line to confirm the identity with original tumor.

5.3. Chor-IN-1 cell line establishment

The original surgical sample was obtained from a 55-year-old man diagnosed with a locally advanced sacral chordoma. A sacral nodule of 2 cm of diameter, macroscopically invading the surrounding soft tissues, was surgically excised. Histological diagnosis of chordoma was made according to WHO classification (2013). Morphologically, the tumor recapitulated the features of conventional chordoma, exhibiting lobulated growth of round epithelioid cells separated by fibrous septa. The cells, arranged in ribbons and nests, showed eosinophilic and/or vacuolated cytoplasms (physaliferous morphology) and were embedded into abundant extracellular matrix. Immuno-phenotyping revealed expression of vimentin, S100, brachyury and EMA (Figure 5.1 A).

The new cell line was established by mechanical and enzymatic disaggregation of the fresh aseptic surgical chordoma sample followed by seeding of the resulting cell suspension in collagen-coated tissue culture plates. Once stabilized in culture, the Chor-IN-1 cell line was subjected to detailed characterization. The morphology was repeatedly monitored over passages, confirming that most of the cells displayed the typical physaliferous phenotype (Figure 5.1 B) of chordoma cells. The cell line was confirmed to express brachyury and EMA by immunocytochemistry (Figure 5.1 B) and show the typical slow growing curve of chordomas (Figure 5.1 C).

The karyotype of the cell line was analyzed using Cytovision® software and found to be: 45, XY, add(1)(p13), −2, add(3)(q21), +7, del(9)(p21), −13, −22, +mar[cp10], a pseudo-diploid karyotype with chromosomal numerical and structural abnormalities typically found in this tumor (namely: del 1p, monosomy 2, deletion 3q, trisomy 7, monosomy 13 and 22) (Figure 5.1 D) [117].

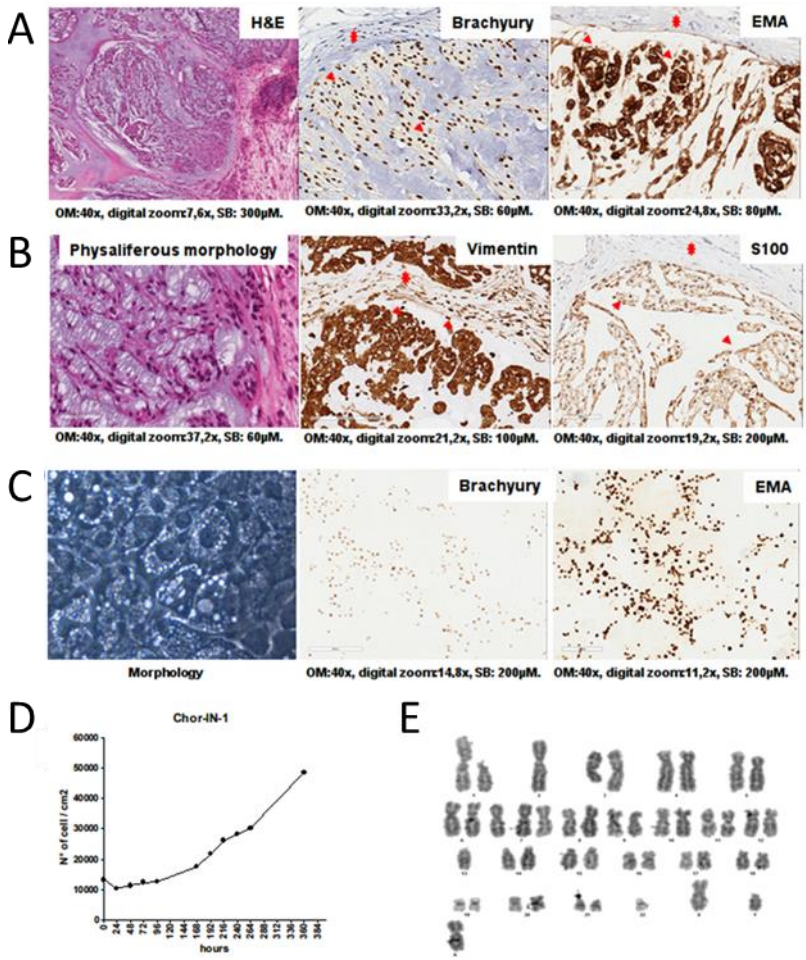

Figure 5.1 Tumor and Chor-IN-1 cell line characterization by H&E and IHC Gene Fusion detection in dilution experiment (from Bosotti R. et al., Sci Rep 2017). (**A)** The original tumor sample and (**B)** the derived Chor-IN-1 cell line were characterized by H&E, revealing the typical physaliferous cells, and by IHC, showing positivity for brachyury and for other chordoma typical biomarkers, as indicated. Arrows and stars indicate examples of tumor and fibroblast cells, respectively. (**C)** Growth curve: the doubling time of Chor-IN-1 cell line was calculated as reported and found to be of about seven days. (**D**) Karyotype analysis of Chor-IN-1 cell line.

5.4. Chor-IN-1 cell line validation and molecular characterization

The Chor-IN-1 cell line was then characterized for the expression of key RTKs and downstream signaling molecules, in comparison to the other chordoma cell lines U-CH1, U-CH2, MUG-Chor1 and JHC7. Western Blot analysis revealed that Chor-IN-1 expresses significant levels of EGFR, PDGFRβ and c-Met proteins, with activation of STAT3 and a very weak P-AKT (Figure 5.2 A).

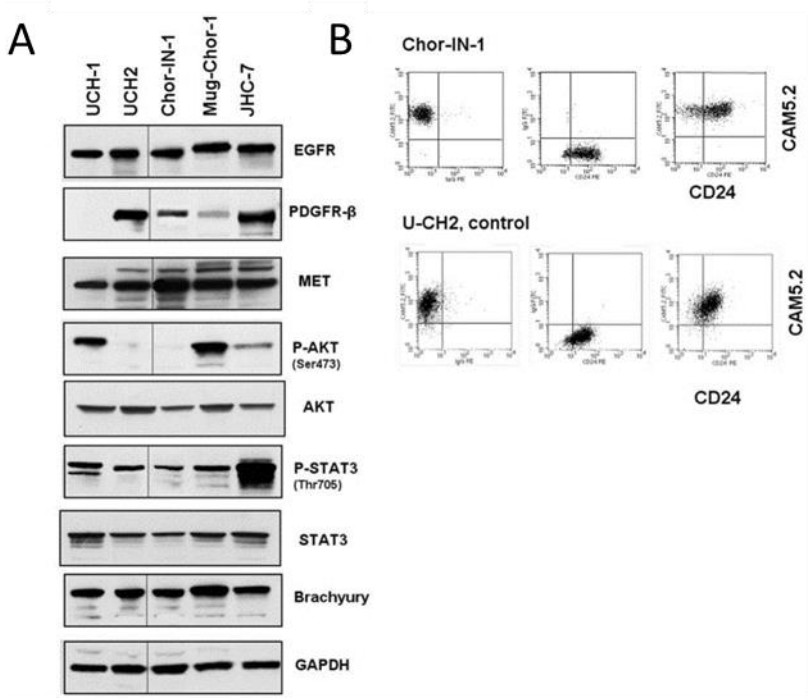

Figure 5.2 Molecular Characterization of Chor-IN-1 cell line (from Bosotti R. et al., Sci Rep 2017). (**A**) Immunoblot analysis of key tyrosine kinase receptors and other signaling molecules: Chor-IN-1, in parallel with U-CH1, U-CH2, MUG-Chor1 and JHC7 cells were seeded and collected at 70% confluence. Protein cell extracts were resolved by SDS-PAGE and filters probed with the respective antibodies. (**B**) Flow Cytometry analysis of chordoma typical membrane proteins: the expression of CD24 and CAM5.2 membrane antigens in Chor-IN-1 was confirmed to be comparable to that of the U-CH2 cell line used as reference. Upper cytograms refer to Chor-IN-1

cells and lower cytograms refer to UCH-2 cells. Left = FITC CAM5.2 *vs*. PE isotype ctrl: 100% of both cell lines express the marker, in the absence of non-specific signals. Middle = PE-CD24 *vs*. FITC isotype ctrl: more than 90% of both cell lines express the marker, in the absence of non-specific signals. Right = FITC CAM5.2 *vs*. PE-CD24: both cell lines are more than 90% CD24/CAM5.2 double positive, confirming both markers are expressed at levels comparable to that of UCH-2 reference cell line.

Brachyury protein was expressed at comparable levels in all cell lines (Figure 5.2 A). Interestingly, RT-qPCR analysis revealed that Chor-IN-1 expresses mRNA levels of the T gene, encoding brachyury, comparable to U-CH1 and U-CH2. Conversely, the MUG-Chor1 and JHC7 cell lines showed higher mRNA expression levels which do not however translate into higher protein levels, likely indicating a strict cellular control on the protein levels of this transcription factor (Figure 5.3).

Cytokeratin 19 and vimentin were also confirmed to be expressed by RT-qPCR analysis, as required for comprehensive chordoma cell line validation (Figure 5.3). Finally, FACS analysis confirmed the expression of CD24 and CAM5.2 membrane antigens in the Chor-IN-1 cell line, similar to the U-CH2 used as control cell line (Figure 5.2 B).

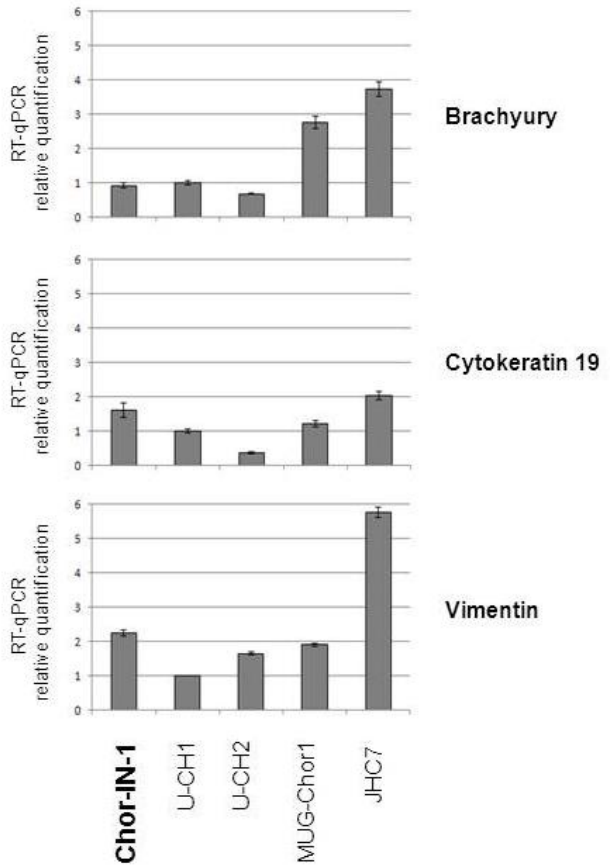

Figure 5.3 RT-qPCR analysis of brachyury, cytokeratin 19 and vimentin gene expression in Chor-IN-1 and in a panel of chordoma cell lines (from Bosotti R. et al., Sci Rep 2017)**.** Histogram bars represent RT-qPCR relative quantification results normalized using U-CH1 as a reference sample.

The Chor-IN-1 cell line was authenticated by Short Tandem Repeat (STR) analysis in parallel with the parental tumor sample. The identity of the other chordoma cell lines was also confirmed by comparison with the published STR profiles (Figure 5.4).

	AMEL_1	AMEL_2	CSF1PO_1	CSF1PO_2	D13S317_1	D13S317_2	D16S539_1	D16S539_2	D18S51_1	D18S51_2	D21S11_1	D21S11_2	D3S1358_1	D3S1358_2	D5S818_1	D5S818_2	D7S820_1	D7S820_2	D8S1179_1	D8S1179_2	FGA_1	FGA_2	TH01_1	TH01_2	TPOX_1	TPOX_2	vWA_1	vWA_2	PentaD_1	PentaD_2	PentaE_1	PentaE_2
Chor-IN-1 cell line (established in house)	X	Y	10	12	11	12	13	13	15	17	31	31.2	17	18	9	12	8	9	12	15	21	22	9.3	9.3	7	8	18	18	13	13	12	21
Chor-IN-1 original tumor	X	Y	10	12	11	12	13	13	15	17	31	31.2	17	18	9	12	8	9	12	15	21	22	9.3	9.3	7	8	18	18	13	13	12	21
UCH-1 cell line (Chordoma Foundation)	X	Y	10	11	11	13	12	13	15	15	28	29	15	15	11	12	9	12	10	15	20	21	7	7	8	11	17	17	11	11	7	10
reference profile (DSMZ_DB)	X	Y	10	11	11	13	12	13							11	12	9	12					7	7	8	11	17	17				
UCH-2 cell line (Chordoma Foundation)	X	X	11	12	11	11	12	12	12	18	29	30	17	17	10	11	8	12	13	13	21	22.2	6	6	8	8	17	17	12	13	12	15
reference profile (COSMIC_DB)	X	X	11	12	11	11	12	12							10	11	8	12					6	6	8	8	17	17				
MUG-Chor-1 cell line (ATCC)	X	X	11	11	11	11	11	14	17	23	29	33.2	14	17	11	12	8	11	11	12	21	26	9.3	9.3	8	8	15	15	13	13	5	12
reference profile (DSMZ_DB)	X	X	11	11	11	11	11	14							11	12	8	11					9.3	9.3	8	8	15	15				
JHC7 cell line (Chordoma Foundation)	X	X	11	11	11	11	11	11	12	12	27	31.2	17	17	13	13	7	10	13	14	21	23	6	8	10	11	17	17	6	11	15	15
reference profile (ATCC_DB)	X	X	11	11	11	11	11	11							13	13	7	10					6	8	10	11	17	17				

Figure 5.4 Characterization of chordoma cell lines by STR profiling (from Bosotti R. et al., Sci Rep 2017**).** Chordoma cell line STR profiles are compared against the corresponding public reference STR profile (dark grey). The profile obtained for the newly established Chor-IN-1 cell line and Chor-IN-1 original tumor is highlighted in light grey.

5.5. Kinase gene expression analysis of Chor-IN-1 by KING-REX

The detection of abundantly/differentially expressed kinases in the chordoma cell lines might allow for the identification of potential new pharmacological targets in this disease. For this reason, we analyzed the global kinase gene expression of Chor-IN-1 in parallel with the other cell lines applying the custom targeted RNA sequencing approach KING-REX, described in Chapter 4, to investigate the kinome gene expression in chordoma.

Results were displayed using Kohonen maps, a data analysis and visualization neural network-based technique for multidimensional quantitative and qualitative data comparison [128]. This analysis showed a very consistent kinase expression profile among chordoma cell lines, which differs from that of control placenta tissue (Figure 5.5). As shown in Figure 5.5 A, in the chordoma panel, about 75% of kinases are expressed, more than half at high levels, while 25% are expressed at very low levels or not expressed.

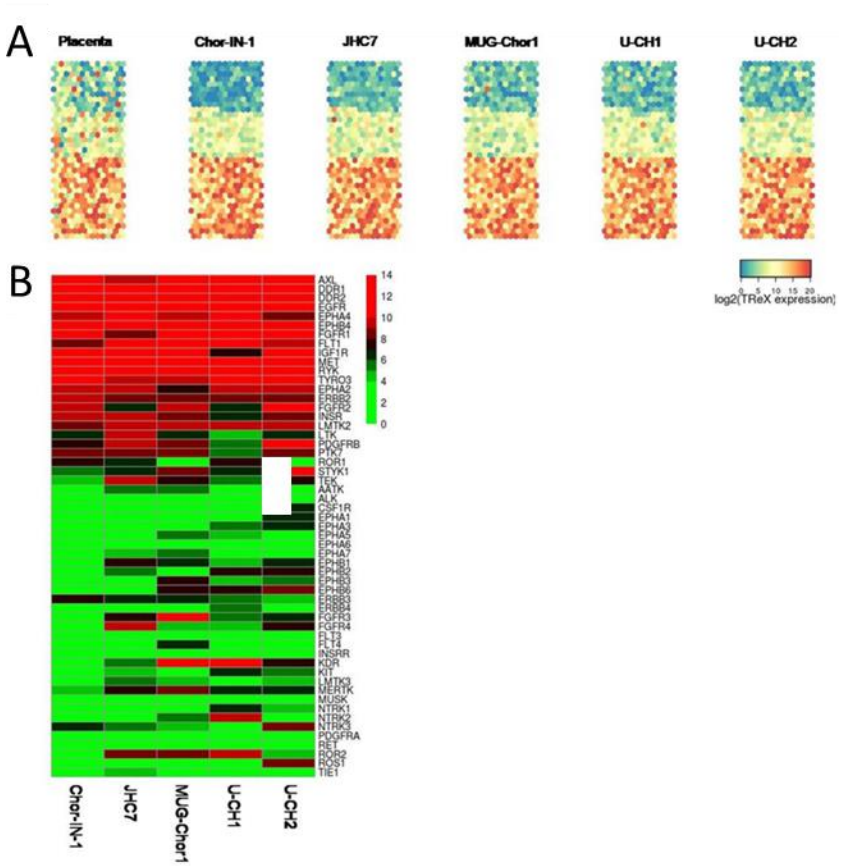

Figure 5.5 Kinase gene expression levels in chordoma cell lines (from Bosotti R. et al., Sci Rep 2017)**.** (**A**) Kohonen maps reporting gene expression levels of 487 kinases in the different chordoma cell lines. Placenta was used as control sample. Each dot represents an individual kinase whose position is kept constant in the map. Colormap gradient from blue to red reflects increasing gene expression level, expressed in log2 scale. (**B**) Heatmap reporting gene expression levels of RTKs (log2 scale) sorted from high to low levels. Colors reflect gene expression levels according to reported scale.

In particular, the analysis was focused on RTKs, which are involved in cell regulation processes and frequently dysregulated in tumors. Out of 56 RTKs, 15 were found expressed at high levels in all cell lines. These include EGFR and MET, as expected, while PDGFR-β was expressed in all cell lines but not in the U-CH1 (Figure 5.5 B). A few kinases showed a cell-line distinctive profile. Therefore, the kinases specifically expressed in the newly established Chor-IN-1 cell line were

analyzed by gene expression differential analysis. The heatmap depicted in Figure 5.6 A shows the differentially expressed kinases in Chor-IN-1 vs. all the other chordoma cell lines (pV<0.05, FC>|2|). ULK4, NPR1 and CDKL4 were the top most expressed kinases in the Chor-IN-1 cell line, while FGFR3, KDR and WNK2 were less expressed or absent in Chor-IN-1 cells as compared to the other chordoma cell lines. The differential expression of these kinases in Chor-IN-1 was confirmed by RT-qPCR (Figure 5.6 B).

The Chor-IN-1 cell line expresses high levels of ULK4, a member of the unc-51-like serine/threonine kinase (STK) family. Although little is known about ULK4 function, its role in chordoma deserves further investigation since the other members of the family have been implicated in autophagic pathways. CDKL4 (Cyclin Dependent Kinase Like 4), a member of the CDK family which includes CDK4 and CDK6, is also highly expressed. Interestingly, the use of CDK4/6-specific inhibitor palbociclib was reported to efficiently inhibit tumor cell growth in vitro in chordoma cell line models [129], providing the rationale for clinical trials evaluating the efficacy of palbociclib in chordoma.

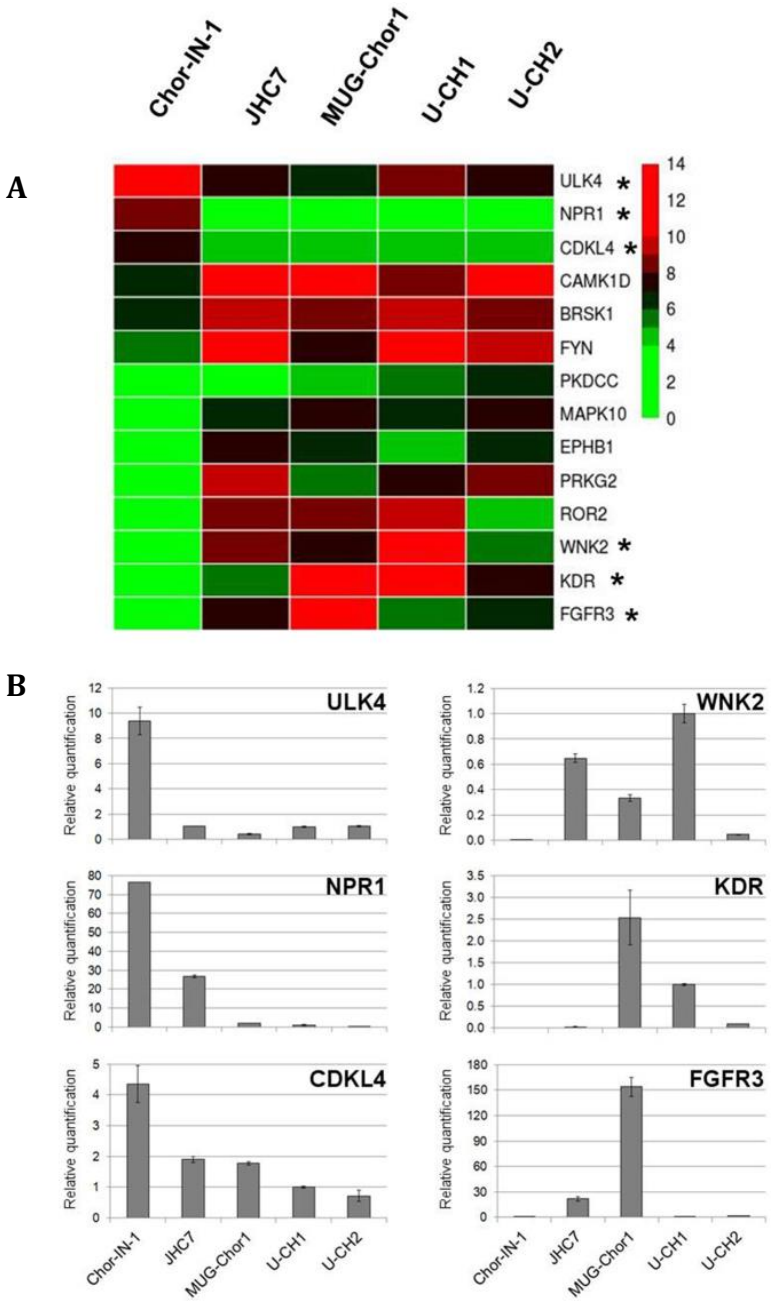

Figure 5.6 Identification of kinases differentially expressed in Chor-IN-1 cell line vs. a panel of chordoma cell lines (from Bosotti R. et al., Sci Rep 2017**).** (**A**) Heatmap representing the expression levels of the most differentially expressed kinases in Chor-IN-1 cell line, as obtained by NGS

analysis using TREx, Illumina, San Diego, CA, USA. Colormap shades reflect relative low to high expression levels, ranging from green (log2 =0) to red (log2=14). (**B**) RT-qPCR validation of selected genes (indicated with a star in the heatmap). Histogram bars represent RT-qPCR relative quantification results, normalized using U-CH1 as a reference sample.

5.6. Identification of sensitivity biomarkers to EGFR inhibitors treatment by KING-REX

Data generated on kinome gene expression in the different chordoma cell lines can be further exploited to investigate the molecular basis for the sensitivity of each cell line to different kinase inhibitors already on the market. Several among the most widely expressed kinases are indeed inhibited by drugs currently undergoing clinical development, which may represent a new therapeutic option also in chordoma.

In particular, the role of EGFR inhibitors was investigated in depth. Anecdotal responses to EGFR inhibitors have been indeed reported, suggesting that EGFR inhibitors might have therapeutic potential for these tumors [111, 114-117]. To evaluate the importance of EGFR for the growth of chordoma cells, we tested the chordoma cell line panel with clinically approved EGFR inhibitor drugs erlotinib, gefitinib, afatinib, and lapatinib, which possess different potency and selectivity within the EGFR/HER2 family [111, 114, 115, 117, 130].

Interestingly a peculiar sensitivity to EGFR inhibitors was observed in U-CH1 and UM-Chor1 versus the other sacral chordoma cell lines (Table 5.1).

Table 5.1: Antiproliferative activity of EGFR inhibitors in chordoma cell lines (Magnaghi P. et al. Mol Cancer Ther.2018)**.**

IC50 (μM) (StdDev)									
Cell Line	**U-CH1**	**UM-Chor1**	**MUG-Chor1**	**U-CH2 (Ch. F.)**	**U-CH2 (ATCC)**	**Chor-IN-1**	**JHC7**	**A431 ctrl**	**A2780 ctrl**
Afatinib	0.014	0.023	0.258	0.494	0.531	0.668	1.346	0.026	1.915
	(0.005)	(0.007)	(0.072)	(0.409)	(0.203)	(0.351)	(0.394)	(0.009)	(0.594)
Erlotinib	0.144	0.617	3.006	8.042	7.776	2.329	2.281	0.346	3.919
	(0.049)	(0.069)	(0.977)	-1.714	-1.953	(0.774)	(0.848)	(0.033)	(0.898)
Lapatinib	0.656	0.516	>10	>10	>10	>10	>10	0.562	3.578
	(0.257)	(0.080)	(-)	(-)	(-)	(-)	(-)	(0.061)	(0.771)
Gefitinib	0.791	0.751	6.241	6.259	5.936	9.040	7.010	0.333	4.762
	(0.446)	(0.055)	-1.390	-2.502	-2.115	-1.389	(0.856)	(0.069)	(0.911)

In Bold: Registered EGFR inhibitors. The average inhibitory activity IC50 (μM) was calculated by comparing treated versus control data using a sigmoidal fitting algorithm. All values were derived from technical duplicates and confirmed in multiple (n > 6) independent biological experiments.

KING-REX analysis identified STK33 as the only kinase with undetectable expression in cell lines sensitive to EGFR inhibitors and higher expression in the other cell lines. STK33 differential expression was further confirmed by RTqPCR analysis (Figure 5.7), suggesting STK33 might represent an interesting biomarker for patient population selection eligible to the treatment with EGFR inhibitors.

The generation and extensive characterization of the new Chor-IN-1 chordoma cell line represents a valuable contribution to the preclinical research in the field, in view of the paucity of current cell models and of the heterogeneity of chordoma tumors. Moreover, the generated kinome data can be exploited for the identification of biomarkers of drug sensitivity, and for the identification of novel pharmacological targets in chordoma. STK33 represents an interesting biomarker of sensitivity to EGFR inhibitors, although it deserves further investigations

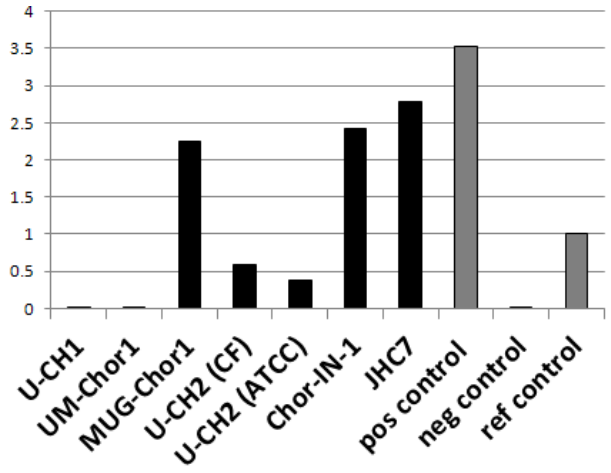

Figure 5.7 Expression of STK33 in the different chordoma cell lines (Magnaghi P. et al. Mol Cancer Ther.2018)**.** RTqPCR analysis of STK33 mRNA expression normalized to reference controls. Cell lines are ordered based on afatinib sensitivity ranking.

.

Chapter 6

Overall conclusions

The identification of oncogenic fusions resulting from chromosomal rearrangements in many different tumor types represents an important opportunity for the use of novel therapeutic strategies. This is definitely true for lung adenocarcinoma, as the introduction into the clinical practice of therapeutic approaches based on the specific targeting of activated kinases has dramatically changed the clinical outcome of many patients [24, 25]. By contrast, in other high unmet medical need pathologies such as CRC, the era of targeted therapy coupled with positive predictive biomarkers, is still in its infancy. For many years molecular analyses performed in metastatic CRC patients have been aimed mainly to identify RAS or BRAF mutations that are associated with the lack of response to anti-EGFR treatment [131]. The identification of chromosomal rearrangements involving NTRK1 or ALK as a low frequency event in a subset of CRC patients is a relatively recent finding [132, 133].

During the screening for patient enrollment into clinical trials with entrectinib, a new potent and selective ALK, NTRKs and ROS1 inhibitor, the presence of rearrangements involving these kinases has been routinely evaluated by IHC in all CRC tumors. This screening performed by the use of a targeted RNAseq approach, specifically developed for this purpose, resulted in the identification of two new oncogenic fusions, SCYL3-TRK and CAD-ALK, which were fully characterized. Despite the significant number of different fusion partners that has already been identified for NTRK1 and ALK in solid tumors, this is the first time that the SCYL3 gene is found to be

involved in a rearrangement with NTRK1 and that the CAD gene is found fused with ALK kinase.

Patients harboring these types of rearrangements derive clinical benefit from the treatment with entrectinib [here and in [40]]. Thus, the screening of CRC tumors for the detection of gene rearrangements specifically involving the NTRKs, ALK and ROS1 genes has the ability to identify a subset of patients able to derive benefit from treatment with entrectinib or other targeted inhibitors.

Conversely, the discovery of new candidate rearrangements for the development of novel drugs require a more systematic and automated approach, both in terms of computational tools and experimental approaches. Indeed, their relevance in different diseases and their druggability makes kinases a very attractive class of pharmacological targets. However, despite the about 40 kinase inhibitor drugs approved in Oncology [19], and the hundreds of compounds in clinical development [18], much of the kinase therapeutic potential remains untapped.

Search for new kinase targets requires the ability to mine existing genomics databases for the rapid and efficient detection of rare events of kinase overexpression in specific tissues, as hallmark of the presence of a rare kinase fusion event, representing the ideal target for novel drugs.

For this propose the tool KAOS was developed and applied to publicly available trascriptomics data from tumor cell lines. KAOS requires gene expression data from either microarray or RNAseq experiments to identify those kinases showing an outlier gene expression profile within samples belonging to the same tissue type. This anomalous gene expression can be used as read out of the presence of a kinase gene fusion. The tool performs particularly well in detecting extreme outliers that stands out on a high variable expression background, being able to identify known and novel kinase fusions. It therefore represents a concrete example of how the increasing overwhelming availability of genomic knowledge bases, which are still growing over time, can be exploited for new target discovery.

While KAOS can be freely downloaded and used on already available whole transcriptomics data, the implementation of a strategy specifically focused on kinase expression evaluation and fusion detection represents an important resource for its use in the clinics, where the tumor sample are typically scarce and of poor RNA quality,

limiting the applicability of whole trancriptome approaches. Indeed, while 'omics' analysis approaches produce huge amounts of molecular information, requiring substantial computational power for data storage and management, NGS targeted RNA approaches enable the analysis of focused portions of the transcriptome, with advantages over whole transcriptome sequencing in terms of reduced costs and simplicity of execution.

KING-REX, a custom targeted RNA kinome NGS approach, was implemented for new kinase target discovery with an applicability in clinical settings, which can exploit selected 'omics' observations to enable more rapid, cost-effective and focused molecular research screens or diagnostic approaches. In general, in the design of RNAseq targeted panels for the detection of kinase gene expression and/or gene fusions, a compromise must be reached between optimal assay performance and limitations imposed by small scale approaches, by maximizing either sequence coverage or target number. KING-REX has been designed in a way that results suitable for the gene expression screening of a comprehensive human kinome panel on Illumina® MiSeq or NextSeq platforms. The system was intended to maximize kinome coverage (512 kinases) by minimizing the number of per-kinase assays, while retaining the possibility to infer the presence of gene fusions for a wide number of kinases (319), including all receptor and non-receptor tyrosine kinases, using paired assays located within and outside the catalytic domain. The application of a similar strategy, based on measuring the imbalanced expression between 5' and 3' transcript ends, has been restricted so far to the analysis of a limited number of kinases [134]. Together with the design and experimental setting of the approach, an ad hoc data analysis pipeline was implemented to streamline both the gene expression analysis workflow and the detection of kinase imbalances, as readout for potential gene fusion events. This was reached by introducing a scaling factor to balance for possible different performances of IN and OUT assays, deriving from technical artifacts, such as primer efficiency, RNA degradation and/or reverse transcription effects. KING-REX demonstrated a high detection sensitivity and a very good overlap ($R2>0.8$) with whole transcriptome results, performed in parallel on the same cell lines. KING-REX gene expression detection accuracy was maintained even in heterogeneous RNA mixtures, mimicking the condition of tumor clinical samples, where contamination with adjacent

normal/stromal tissue or inflammatory infiltration represents a common scenario. In addition, using cell line models harboring known gene fusions, kinase rearrangements can be correctly and robustly detected by the system, distinct from full length sequences, even in complex background mixtures or in heat-degraded RNA, based on the evaluation of the imbalanced measured expression ratio for paired kinase assays.

KING-REX is currently the largest targeted approach for expression analysis of kinases which could be used as a rapid and cost-effective investigation tool in cancer biology and the profiling of rare and poorly characterized tumors, such as chordomas. It represents a useful setup for the comprehensive analysis of the kinome in cancer or other diseases, for applications in the field of the identification of novel, putative kinase targets.

References

1. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
2. International Human Genome Sequencing, C., *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-45.
3. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
4. NIH. *Targeted Cancer Therapies*. 2018; Available from: https://www.cancer.gov/about-cancer/treatment/types/targeted-therapies/targeted-therapies-fact-sheet.
5. Kamps, R., et al., *Next-Generation Sequencing in Oncology: Genetic Diagnosis, Risk Prediction and Cancer Classification.* Int J Mol Sci, 2017. **18**(2).
6. Reungwetwattana, T. and G.K. Dy, *Targeted therapies in development for non-small cell lung cancer.* J Carcinog, 2013. **12**: p. 22.
7. Tulbah, A., et al., *The journey toward personalized cancer therapy.* Adv Anat Pathol, 2014. **21**(1): p. 36-43.
8. Stransky, N., et al., *The landscape of kinase fusions in cancer.* Nat Commun, 2014. **5**: p. 4846.
9. Garraway, L.A. and E.S. Lander, *Lessons from the cancer genome.* Cell, 2013. **153**(1): p. 17-37.
10. Vogelstein, B., et al., *Cancer genome landscapes.* Science, 2013. **339**(6127): p. 1546-58.
11. Torkamani, A., G. Verkhivker, and N.J. Schork, *Cancer driver mutations in protein kinase genes.* Cancer Lett, 2009. **281**(2): p. 117-27.
12. Dankner, M., et al., *Classifying BRAF alterations in cancer: new rational therapeutic strategies for actionable mutations.* Oncogene, 2018. **37**(24): p. 3183-3199.
13. Ball, D.W., *Management of medullary thyroid cancer.* Minerva Endocrinol, 2011. **36**(1): p. 87-98.
14. Iancu, G., et al., *"Triple positive" breast cancer - a novel category?* Rom J Morphol Embryol, 2017. **58**(1): p. 21-26.

15. Sawyers, C.L., *Molecular consequences of the BCR-ABL translocation in chronic myelogenous leukemia.* Leuk Lymphoma, 1993. **11 Suppl 2**: p. 101-3.
16. Zhang, J., P.L. Yang, and N.S. Gray, *Targeting cancer with small molecule kinase inhibitors.* Nat Rev Cancer, 2009. **9**(1): p. 28-39.
17. Bhullar, K.S., et al., *Kinase-targeted cancer therapies: progress, challenges and future directions.* Mol Cancer, 2018. **17**(1): p. 48.
18. Klaeger, S., et al., *The target landscape of clinical kinase drugs.* Science, 2017. **358**(6367).
19. Ferguson, F.M. and N.S. Gray, *Kinase inhibitors: the road ahead.* Nat Rev Drug Discov, 2018. **17**(5): p. 353-377.
20. Duong-Ly, K.C. and J.R. Peterson, *The Human Kinome and Kinase Inhibition as a therapeutic strategy.* Curr Protoc Pharmacol., 2013. 0 2:Unit2.9.
21. Jabbour, E.J., J.E. Cortes, and H.M. Kantarjian, *Tyrosine kinase inhibition: a therapeutic target for the management of chronic-phase chronic myeloid leukemia.* Expert Rev Anticancer Ther, 2013. **13**(12): p. 1433-52.
22. Moore, F.R., C.B. Rempfer, and R.D. Press, *Quantitative BCR-ABL1 RQ-PCR fusion transcript monitoring in chronic myelogenous leukemia.* Methods Mol Biol, 2013. **999**: p. 1-23.
23. Mughal, T.I., et al., *Chronic myeloid leukemia: reminiscences and dreams.* Haematologica, 2016. **101**(5): p. 541-58.
24. Hirsch, F.R., et al., *New and emerging targeted treatments in advanced non-small-cell lung cancer.* Lancet, 2016. **388**(10048): p. 1012-24.
25. Katayama, R., C.M. Lovly, and A.T. Shaw, *Therapeutic targeting of anaplastic lymphoma kinase in lung cancer: a paradigm for precision cancer medicine.* Clin Cancer Res, 2015. **21**(10): p. 2227-35.
26. Tsao, A.S., et al., *Scientific Advances in Lung Cancer 2015.* J Thorac Oncol, 2016. **11**(5): p. 613-38.
27. Morris, S.W., et al., *Fusion of a kinase gene, ALK, to a nucleolar protein gene, NPM, in non-Hodgkin's lymphoma.* Science, 1994. **263**(5151): p. 1281-4.
28. Soda, M., et al., *Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer.* Nature, 2007. **448**(7153): p. 561-6.
29. Holla, V.R., et al., *ALK: a tyrosine kinase target for cancer therapy.* Cold Spring Harb Mol Case Stud, 2017. **3**(1): p. a001115.

30. Ziogas, D.C., et al., *Treating ALK-positive non-small cell lung cancer.* Ann Transl Med, 2018. **6**(8): p. 141.
31. Zer, A. and N. Leighl, *Promising Targets and Current Clinical Trials in Metastatic Non-Squamous NSCLC.* Front Oncol, 2014. **4**: p. 329.
32. Ardini, E., et al., *The TPM3-NTRK1 rearrangement is a recurring event in colorectal carcinoma and is associated with tumor sensitivity to TRKA kinase inhibition.* Mol Oncol, 2014. **8**(8): p. 1495-507.
33. Drilon, A., et al., *Response to Cabozantinib in patients with RET fusion-positive lung adenocarcinomas.* Cancer Discov, 2013. **3**(6): p. 630-5.
34. Gainor, J.F. and A.T. Shaw, *Novel targets in non-small cell lung cancer: ROS1 and RET fusions.* Oncologist, 2013. **18**(7): p. 865-75.
35. Kohno, T., et al., *RET fusion gene: translation to personalized lung cancer therapy.* Cancer Sci, 2013. **104**(11): p. 1396-400.
36. Parker, B.C. and W. Zhang, *Fusion genes in solid tumors: an emerging target for cancer diagnosis and treatment.* Chin J Cancer, 2013. **32**(11): p. 594-603.
37. Shaw, A.T., et al., *Tyrosine kinase gene rearrangements in epithelial malignancies.* Nat Rev Cancer, 2013. **13**(11): p. 772-87.
38. Ardini, E., et al., *Entrectinib, a Pan-TRK, ROS1, and ALK Inhibitor with Activity in Multiple Molecularly Defined Cancer Indications.* Mol Cancer Ther, 2016. **15**(4): p. 628-39.
39. Menichincheri, M., et al., *Discovery of Entrectinib: A New 3-Aminoindazole As a Potent Anaplastic Lymphoma Kinase (ALK), c-ros Oncogene 1 Kinase (ROS1), and Pan-Tropomyosin Receptor Kinases (Pan-TRKs) inhibitor.* J Med Chem, 2016. **59**(7): p. 3392-408.
40. Drilon, A., et al., *Safety and Antitumor Activity of the Multitargeted Pan-TRK, ROS1, and ALK Inhibitor Entrectinib: Combined Results from Two Phase I Trials (ALKA-372-001 and STARTRK-1).* Cancer Discov, 2017. **7**(4): p. 400-409.
41. Agarwal, A., D. Ressler, and G. Snyder, *The current and future state of companion diagnostics.* Pharmgenomics Pers Med, 2015. **8**: p. 99-110.
42. Murphy, D.A., et al., *Detecting Gene Rearrangements in Patient Populations Through a 2-Step Diagnostic Test Comprised of Rapid IHC Enrichment Followed by Sensitive Next-Generation Sequencing.* Appl Immunohistochem Mol Morphol, 2017. **25**(7): p. 513-523.
43. Kilpinen, S., K. Ojala, and O. Kallioniemi, *Analysis of kinase gene expression patterns across 5681 human tissue samples reveals*

functional genomic taxonomy of the kinome. PLoS One, 2010. **5**(12): p. e15068.

44. Ursu, O., et al., *Network modeling of kinase inhibitor polypharmacology reveals pathways targeted in chemical screens.* PLoS One, 2017. **12**(10): p. e0185650.

45. NIH. *The Cancer Genome Atlas*. 2018; Available from: https://cancergenome.nih.gov/.

46. Tomlins, S.A., et al., *Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer.* Science, 2005. **310**(5748): p. 644-8.

47. Mpindi, J.P., et al., *GTI: a novel algorithm for identifying outlier gene expression profiles from integrated microarray datasets.* PLoS One, 2011. **6**(2): p. e17259.

48. Roden, D.L., et al., *ZODET: software for the identification, analysis and visualisation of outlier genes in microarray expression data.* PLoS One, 2014. **9**(1): p. e81123.

49. Kothari, V., et al., *Outlier kinase expression by RNA sequencing as targets for precision therapy.* Cancer Discov, 2013. **3**(3): p. 280-93.

50. Wagner, G.P., K. Kin, and V.J. Lynch, *Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples.* Theory Biosci, 2012. **131**(4): p. 281-5.

51. *The R Project for Statistical Computing*. 2018; Available from: http://www.r-project.org/.

52. *GraphPad Prism 7*. 2018; Available from: http://www.graphpad.com/.

53. Giacomini, C.P., et al., *Breakpoint analysis of transcriptional and genomic profiles uncovers novel gene fusions spanning multiple human cancer types.* PLoS Genet, 2013. **9**(4): p. e1003464.

54. Thieme, S. and P. Groth, *Genome Fusion Detection: a novel method to detect fusion genes from SNP-array data.* Bioinformatics, 2013. **29**(6): p. 671-7.

55. Carrara, M., et al., *State of art fusion-finder algorithms are suitable to detect transcription-induced chimeras in normal tissues?* BMC Bioinformatics, 2013. **14 Suppl 7**: p. S2.

56. Kumar, S., et al., *Identifying fusion transcripts using next generation sequencing.* Wiley Interdiscip Rev RNA, 2016. **7**(6): p. 811-823.

57. Carrara, M., et al., *State-of-the-art fusion-finder algorithms sensitivity and specificity.* Biomed Res Int, 2013. **2013**: p. 340620.

58. *Archer FusionPlex NGS assays*. 2018; Available from: http://archerdx.com/fusionplex-assays/.
59. *QuantideX NGS RNA Lung Cancer Kit*. 2018; Available from: https://asuragen.com/portfolio/oncology/quantidex-ngs-rna-lung-cancer-kit/.
60. *Ovation Fusion Panel Target Enrichment System V2*. 2018; Available from: https://www.nugen.com/products/ovation-fusion-panel-target-enrichment-system-v2.
61. *Ion AmpliSeq RNA Fusion Lung Cancer Research Panel*. 2018; Available from: https://www.thermofisher.com/it/en/home/life-science/cancer-research/cancer-genomics/targeted-sequencing-cancer-mutation-detection/rna-fusion-detection.html.
62. Reeser, J.W., et al., *Validation of a Targeted RNA Sequencing Assay for Kinase Fusion Detection in Solid Tumors.* J Mol Diagn, 2017. **19**(5): p. 682-696.
63. Rogers, T.M., et al., *Multiplexed transcriptome analysis to detect ALK, ROS1 and RET rearrangements in lung cancer.* Sci Rep, 2017. **7**: p. 42259.
64. *The Globocan Project*. 2018; Available from: http://globocan.iarc.fr.
65. Moriarity, A., et al., *Current targeted therapies in the treatment of advanced colorectal cancer: a review.* Ther Adv Med Oncol, 2016. **8**(4): p. 276-93.
66. Seeber, A. and G. Gastl, *Targeted Therapy of Colorectal Cancer.* Oncol Res Treat, 2016. **39**(12): p. 796-802.
67. Grande, E., M.V. Bolos, and E. Arriola, *Targeting oncogenic ALK: a promising strategy for cancer treatment.* Mol Cancer Ther, 2011. **10**(4): p. 569-79.
68. Awad, M.M. and A.T. Shaw, *ALK inhibitors in non-small cell lung cancer: crizotinib and beyond.* Clin Adv Hematol Oncol, 2014. **12**(7): p. 429-39.
69. *NTRK gene fusions: novel targets of cancer therapy*. 2018; Available from: www.ntrkfusions.com.
70. Creancier, L., et al., *Chromosomal rearrangements involving the NTRK1 gene in colorectal carcinoma.* Cancer Lett, 2015. **365**(1): p. 107-11.
71. Sartore-Bianchi, A., et al., *Sensitivity to Entrectinib Associated With a Novel LMNA-NTRK1 Gene Fusion in Metastatic Colorectal Cancer.* J Natl Cancer Inst, 2016. **108**(1).

72. Lee, S.J., et al., *NTRK1 rearrangement in colorectal cancer patients: evidence for actionable target using patient-derived tumor cell line.* Oncotarget, 2015. **6**(36): p. 39028-35.
73. Park, D.Y., et al., *NTRK1 fusions for the therapeutic intervention of Korean patients with colon cancer.* Oncotarget, 2016. **7**(7): p. 8399-412.
74. *Fusion Detection in Archer Analysis Software*. 2018; Available from: https://cdn2.hubspot.net/hubfs/4445440/Tech%20notes/Tech%20Note-Fusion-detection-in-Archer-Analysis-Software.pdf?submissionGuid=eade47fa-4d3f-41af-9765-e23290154ce4.
75. Sullivan, A., et al., *PACE-1, a novel protein that interacts with the C-terminal domain of ezrin.* Experimental Cell Research, 2003. **284**(2): p. 224-238.
76. Warmuth, M., et al., *Ba/F3 cells and their use in kinase drug discovery.* Current Opinion in Oncology, 2007. **19**(1): p. 55-60.
77. Grande-Garcia, A., et al., *Structure, functional characterization, and evolution of the dihydroorotase domain of human CAD.* Structure, 2014. **22**(2): p. 185-98.
78. Richmond, A.L., et al., *The nucleotide synthesis enzyme CAD inhibits NOD2 antibacterial function in human intestinal epithelial cells.* Gastroenterology, 2012. **142**(7): p. 1483-92 e6.
79. Uhlen, M., et al., *Proteomics. Tissue-based map of the human proteome.* Science, 2015. **347**(6220): p. 1260419.
80. Aisner, D.L., et al., *ROS1 and ALK fusions in colorectal cancer, with evidence of intratumoral heterogeneity for molecular drivers.* Mol Cancer Res, 2014. **12**(1): p. 111-8.
81. Medico, E., et al., *The molecular landscape of colorectal cancer cell lines unveils clinically actionable kinase targets.* Nat Commun, 2015. **6**: p. 7002.
82. Zheng, Z., et al., *Anchored multiplex PCR for targeted next-generation sequencing.* Nat Med, 2014. **20**(12): p. 1479-84.
83. Barretina, J., et al., *The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.* Nature, 2012. **483**(7391): p. 603-7.
84. *How to label all the outliers in a boxplot*. 2011; Available from: https://www.r-statistics.com/2011/01/.

85. *nsga2R: Elitist Non-dominated Sorting Genetic Algorithm based on R.* 2013; Available from: https://cran.r-project.org/web/packages/nsga2R/index.html.
86. Komsta, L. *R Package 'outliers'*. 2015; Available from: https://cran.r-project.org/web/packages/outliers/outliers.pdf.
87. Deb, K., et al., *A fast and elitist multiobjective genetic algorithm: NSGA-II.* Ieee Transactions on Evolutionary Computation, 2002. **6**(2): p. 182-197.
88. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 2003. **4**(2): p. 249-64.
89. Charest, A., et al., *Fusion of FIG to the receptor tyrosine kinase ROS in a glioblastoma with an interstitial del(6)(q21q21).* Genes Chromosomes Cancer, 2003. **37**(1): p. 58-71.
90. Santoro, M., et al., *Ret oncogene activation in human thyroid neoplasms is restricted to the papillary cancer subtype.* J Clin Invest, 1992. **89**(5): p. 1517-22.
91. Roidl, A., et al., *The FGFR4 Y367C mutant is a dominant oncogene in MDA-MB453 breast cancer cells.* Oncogene, 2010. **29**(10): p. 1543-1552.
92. Wang, H.P., et al., *ZAP-70: An Essential Kinase in T-cell Signaling.* Cold Spring Harbor Perspectives in Biology, 2010. **2**(5).
93. Illumina. *TruSeq Targeted RNA Expression Library Prep Kits*. 2018; Available from: https://www.illumina.com/products/by-type/sequencing-kits/library-prep-kits/truseq-targeted-rna.html.
94. Manning, G., et al., *The protein kinase complement of the human genome.* Science, 2002. **298**(5600): p. 1912-+.
95. Zhao, S.R. and B.H. Zhang, *A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification.* Bmc Genomics, 2015. **16**.
96. *Superfamily. HMM library and genome assignments server*. 2017; Available from: http://supfam.org/SUPERFAMILY/.
97. Thorvaldsdottir, H., J.T. Robinson, and J.P. Mesirov, *Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.* Briefings in Bioinformatics, 2013. **14**(2): p. 178-192.
98. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-842.

99. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biology, 2014. **15**(12).

100. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-140.

101. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.

102. Li, B. and C.N. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* Bmc Bioinformatics, 2011. **12**.

103. Bullard, J.H., et al., *Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.* BMC Bioinformatics, 2010. **11**: p. 94.

104. Hubinger, G., et al., *The tyrosine kinase NPM-ALK, associated with anaplastic large cell lymphoma, binds the intracellular domain of the surface receptor CD30 but is not activated by CD30 stimulation.* Exp Hematol, 1999. **27**(12): p. 1796-805.

105. Matsubara, D., et al., *Identification of CCDC6-RET fusion in the human lung adenocarcinoma cell line, LC-2/ad.* J Thorac Oncol, 2012. **7**(12): p. 1872-1876.

106. Davies, K.D. and R.C. Doebele, *Molecular pathways: ROS1 fusion proteins in cancer.* Clin Cancer Res, 2013. **19**(15): p. 4040-5.

107. D, N. *Novel FGFR2 fusion genes in NCI-H716 colorectal cancer cell line*. 2014; Available from: https://figshare.com/articles/Novel_FGFR2_fusion_gene_in_NCI_H716_colorectal_cancer_cell_line/1125933.

108. Illumina. *Expression Analysis of FFPE Samples*. 2018; Available from: https://www.illumina.com/documents/products/technotes/technote-expression-analysis-ffpe-samples.pdf.

109. George, B., et al., *Chordomas: A Review.* Neurosurg Clin N Am, 2015. **26**(3): p. 437-52.

110. Yakkioui, Y., et al., *Chordoma: the entity.* Biochim Biophys Acta, 2014. **1846**(2): p. 655-69.

111. Walcott, B.P., et al., *Chordoma: current concepts, management, and future directions.* Lancet Oncol, 2012. **13**(2): p. e69-76.

112. Stacchiotti, S., et al., *Best practices for the management of local-regional recurrent chordoma: a position paper by the Chordoma Global Consensus Group.* Annals of Oncology, 2017. **28**(6): p. 1230-1242.

113. Di Maio, S., et al., *Novel targeted therapies in chordoma: an update.* Therapeutics and Clinical Risk Management, 2015. **11**.

114. Stacchiotti, S., J. Sommer, and G. Chordoma Global Consensus, *Building a global consensus approach to chordoma: a position paper from the medical and patient community.* Lancet Oncol, 2015. **16**(2): p. e71-83.

115. Yamada, Y., M. Gounder, and I. Laufer, *Multidisciplinary management of recurrent chordomas.* Curr Treat Options Oncol, 2013. **14**(3): p. 442-53.

116. Nibu, Y., D.S. Jose-Edwards, and A. Di Gregorio, *From notochord formation to hereditary chordoma: the many roles of Brachyury.* Biomed Res Int, 2013. **2013**: p. 826435.

117. Vujovic, S., et al., *Brachyury, a crucial regulator of notochordal development, is a novel biomarker for chordomas.* Journal of Pathology, 2006. **209**(2): p. 157-165.

118. Yang, X.H.R., et al., *T (brachyury) gene duplication confers major susceptibility to familial chordoma.* Nature Genetics, 2009. **41**(11): p. 1176-1178.

119. Presneau, N., et al., *Role of the transcription factor T (brachyury) in the pathogenesis of sporadic chordoma: a genetic and functional-based study.* J Pathol, 2011. **223**(3): p. 327-35.

120. Tamborini, E., et al., *Molecular and biochemical analyses of platelet-derived growth factor receptor (PDGFR) B, PDGFRA, and KIT receptors in chordomas.* Clin Cancer Res, 2006. **12**(23): p. 6920-8.

121. Akhavan-Sigari, R., et al., *Expression of PDGFR-alpha, EGFR and c-MET in spinal chordoma: a series of 52 patients.* Anticancer Res, 2014. **34**(2): p. 623-30.

122. Dewaele, B., et al., *Frequent activation of EGFR in advanced chordomas.* Clin Sarcoma Res, 2011. **1**(1): p. 4.

123. Tamborini, E., et al., *Analysis of receptor tyrosine kinases (RTKs) and downstream pathways in chordomas.* Neuro Oncol, 2010. **12**(8): p. 776-89.

124. de Castro, C.V., et al., *Tyrosine kinase receptor expression in chordomas: phosphorylated AKT correlates inversely with outcome.* Hum Pathol, 2013. **44**(9): p. 1747-55.

125. Bruderlein, S., et al., *Molecular characterization of putative chordoma cell lines.* Sarcoma, 2010. **2010**: p. 630129.

126. Rinner, B., et al., *Establishment and detailed functional and molecular genetic characterisation of a novel sacral chordoma cell line, MUG-Chor1.* Int J Oncol, 2012. **40**(2): p. 443-51.

127. Hsu, W., et al., *Generation of chordoma cell line JHC7 and the identification of Brachyury as a novel molecular target Laboratory investigation.* Journal of Neurosurgery, 2011. **115**(4): p. 760-769.

128. *kohonen: Supervised and Unsupervised Self-Organising Maps.* 2018.

129. von Witzleben, A., et al., *Preclinical Characterization of Novel Chordoma Cell Systems and Their Targeting by Pharmocological Inhibitors of the CDK4/6 Cell-Cycle Pathway.* Cancer Research, 2015. **75**(18): p. 3823-3831.

130. Davis, M.I., et al., *Comprehensive analysis of kinase inhibitor selectivity.* Nat Biotechnol, 2011. **29**(11): p. 1046-51.

131. Pietrantonio, F., et al., *Predictive role of BRAF mutations in patients with advanced colorectal cancer receiving cetuximab and panitumumab: a meta-analysis.* Eur J Cancer, 2015. **51**(5): p. 587-94.

132. Amatu, A., A. Sartore-Bianchi, and S. Siena, *NTRK gene fusions as novel targets of cancer therapy across multiple tumour types.* Esmo Open, 2016. **1**(2).

133. Pietrantonio, F., et al., *ALK, ROS1, and NTRK Rearrangements in Metastatic Colorectal Cancer.* J Natl Cancer Inst, 2017. **109**(12).

134. Beadling, C., et al., *A Multiplexed Amplicon Approach for Detecting Gene Fusions by Next-Generation Sequencing.* J Mol Diagn, 2016. **18**(2): p. 165-75.