Multi-omics Characterisation of Myxoid Liposarcoma and Investigation of Trabectedin Mechanism of Action for Its Treatment

Thesis

Istituto di Ricerche Farmacologiche Mario Negri IRCCS

The Open University

Doctor of Philosophy Program in

School of Life, Health and Chemical Science

Multi-omics characterisation of myxoid liposarcoma and investigation of trabectedin mechanism of action for its treatment

PhD Dissertation of

Laura Mannarino

Personal Identifier:

Supervisors

Dr Maurizio D'Incalci

Dr Florian Markowetz

Director of Studies

Dr Sergio Marchini

JUNE 2021

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University or Institute. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgements.

I declare that Roberta Frapolli was responsible for patient-derived xenograft implants and treatment; Ilaria Craparotta, Sara Ballabio and Nicolò Panini were responsible for wet lab sample preparation, processing and experiments.

This work was performed at the Istituto di Ricerche Farmacologiche Mario Negri IRCCS in Milan, Italy, from 2017 to 2021, under the supervision of Sergio Marchini (director of studies), Maurizio D'Incalci (supervisor), Florian Markowetz (external supervisor) and Luisa Diomede (third party monitor).

Abstract

Myxoid liposarcoma (MLPS) is a rare soft tissue sarcoma of the mesenchymal cells characterised by the expression of the FUS-DDIT3 oncoprotein that impairs the adipocyte differentiation. Among the available therapeutic choices for the treatment of MLPS, trabectedin has shown significant anti-tumour activity. Although prolonged treatment is well-tolerated, recurrence of the resistant tumour invariably occurs even in initially responder patients. With the aim to characterise yet unknown molecular mechanisms of trabectedin mode of action and of resistance in MLPS, we integrated genomic, transcriptomic and protein-DNA binding data from patient-derived MLPS xenograft model, ML017, and its trabectedin-resistant derivate, ML017/ET.

DNA-Seq analysis showed that acquired-resistance to trabectedin may be due to the loss of genetic material in the *4p15.2*, *4p16.3* and *17q21.3* cytobands and to the consequent inhibition of the genes mapping on these regions. Integration of longitudinal RNA-Seq and ChIP-Seq data revealed a two-phases mechanism of action of trabectedin. An early phase, characterised by the cytotoxic action of trabectedin and probably independent from FUS-DDIT3 activity; a second late phase, led by the DNA-binding activity of the chimera. Indeed, trabectedin displaces FUS-DDIT3 from most of its targets and modulates the transcription of the related genes involved in processes that sustain adipocyte differentiation. To note, no such differences were observed in ML017/ET resistant model.

These results shed light on the complex mechanism of action of trabectedin in MLPS. They provide insights into the process leading to drug resistance and can be the basis for the development of novel combinatorial strategies for the treatment of MLPS.

I would like to dedicate this thesis to myself

Acknowledgements

I would like to thank Maurizio D'Incalci and Sergio Marchini for giving me the opportunity to study such an interesting topic in the field of rare cancer diseases and for introducing me to biology and pharmacology.

I would like to thank Florian Markowetz for hosting me in his amazing Lab and for giving me the chance to improve computational competences.

A special thanks goes to Luisa Diomede for her kind support.

This work could not have been possible without two special colleagues, Ilaria Craparotta and Sara Ballabio: I thank them a lot for the time spent together.

I would like to thank Giulio Pavesi for his kind help.

I would like to acknowledge Luca Beltrame for the support given throughout these years, since my first steps in biomedical research.

All my gratitude for both professional support and friendship to Luca Clivio, Tony Travis and Maurizio Callari.

Throughout these years our Lab has hosted a lot people. I have picked something good from each of them, I thank them all.

Researchers' work is improved by people around them: I thank Elda, Silvia and Clara for being there.

I would like to thank Francesca, the hidden starting point of this, now embraced in harmony with lovely Laura.

All my love and gratitude go to Ale and Cami.

I would like to express gratitude to Filippo and his family, they made me stronger.

I am indebted to my family for all the love and support given. I wish I could make you proud and happy.

Publications by the PhD candidate

Ballabio, S., Craparotta, I., Paracchini, L., Mannarino, L., Corso, S., Pezzotta, M.G., Vescio, M., Fruscio, R., Romualdi, C., Dainese, E., Ceppi, L., Calura, E., Pileggi, S., Siravegna, G., Pattini, L., Martini, P., delle Marchette, M., Mangioni, C., Ardizzoia, A., Pellegrino, A., Landoni, F., D'Incalci, M., Beltrame, L., Marchini, S., 2019. Multisite analysis of high-grade serous epithelial ovarian cancers identifies genomic regions of focal and recurrent copy number alteration in 3q26.2 and 8q24.3. International Journal of Cancer 145. https://doi.org/10.1002/ijc.32288

Belgiovine, C., Bello, E., Liguori, M., Craparotta, I., Mannarino, L., Paracchini, L., Beltrame, L., Marchini, S., Galmarini, C.M., Mantovani, A., Frapolli, R., Allavena, P., D'Incalci, M., 2017. Lurbinectedin reduces tumour-associated macrophages and the inflammatory tumour microenvironment in preclinical models. British Journal of Cancer 117. https://doi.org/10.1038/bjc.2017.205

Bello, E., Brich, S., Craparotta, I., Mannarino, L., Ballabio, S., Gatta, R., Marchini, S., Carrassa, L., Matteo, C., Sanfilippo, R., Gronchi, A., Casali, P.G., Pilotti, S., D'Incalci, M., Frapolli, R., 2019. Establishment and characterisation of a new patient-derived model of myxoid liposarcoma with acquired resistance to trabectedin. British Journal of Cancer 121. https://doi.org/10.1038/s41416-019-0550-2

Frapolli, R., Bello, E., Ponzo, M., Craparotta, I., Mannarino, L., Ballabio, S., Marchini, S., Carrassa, L., Ubezio, P., Porcu, L., Brich, S., Sanfilippo, R., Casali, P.G., Gronchi, A., Pilotti, S., D'Incalci, M., 2019. Combination of PPARg agonist pioglitazone and trabectedin induce adipocyte differentiation to overcome trabectedin resistance in myxoid liposarcomas. Clinical Cancer Research 25. https://doi.org/10.1158/1078-0432.CCR-19-0976

Fuso Nerini, I., Roca, E., Mannarino, L., Grosso, F., Frapolli, R., D'Incalci, M., 2020. Is DNA repair a potential target for effective therapies against malignant mesothelioma? Cancer Treatment Reviews 90. https://doi.org/10.1016/j.ctrv.2020.102101

Hiltbrunner, S., Mannarino, L.[1], Kirschner, M.B., Optiz, I., Rigutto, A., Laure, A., Lia, M., Nozza P., Maconi, A., Marchini, S., D'Incalci, M., Curioni-Fontecedro, A., and Grosso, F., 2021, Tumor immune microenvironment and genetic alterations in mesothelioma. Frontiers of Oncology 11:660039. https://doi.org/10.3389/fonc.2021.660039 (accepted for publication)

[1] Co-first author

Libera, L., Craparotta, I., Sahnane, N., Chiaravalli, A.M., Mannarino, L., Cerutti, R., Riva, C., Marchini, S., Furlan, D., 2018. Targeted gene sequencing of Lynch syndrome–related and sporadic endometrial carcinomas. Human Pathology 81. https://doi.org/10.1016/j.humpath.2018.06.029

Liguori, M., Digifico, E., Vacchini, A., Avigni, R., Colombo, F.S., Borroni, E.M., Farina, F.M., Milanesi, S., Castagna, A., Mannarino, L., Craparotta, I., Marchini, S., Erba, E., Panini, N., Tamborini, M., Rimoldi, V., Allavena, P., Belgiovine, C., 2021. The soluble glycoprotein NMB (GPNMB) produced by macrophages induces cancer stemness and metastasis via CD44 and IL-33. Cellular and Molecular Immunology 18. https://doi.org/10.1038/s41423-020-0501-0

Mannarino, L., Paracchini, L., Craparotta, I., Romano, M., Marchini, S., Gatta, R., Erba, E., Clivio, L., Romualdi, C., D'Incalci, M., Beltrame, L., Pattini, L., 2018. A systems biology approach to investigate the mechanism of action of trabectedin in a model of myelomonocytic leukemia. Pharmacogenomics Journal 18. https://doi.org/10.1038/tpj.2016.76

Martini, P., Paracchini, L., Caratti, G., Mello-Grand, M., Fruscio, R., Beltrame, L., Calura, E., Sales, G., Ravaggi, A., Bignotti, E., Odicino, F.E., Sartori, E., Perego, P., Katsaros, D., Craparotta, I., Chiorino, G., Cagnin, S., Mannarino, L., Ceppi, L., Mangioni, C., Ghimenti, C., D'Incalci, M., Marchini, S., Romualdi, C., 2017. LncRNAs as novel indicators of patients' prognosis in stage i epithelial ovarian cancer: A retrospective and multicentric study. Clinical Cancer Research 23. https://doi.org/10.1158/1078-0432.CCR-16-1402

Nastasi, C., Mannarino, L., D'incalci, M., 2020. DNA damage response and immune defense. International Journal of Molecular Sciences 21. https://doi.org/10.3390/ijms21207504

Petrillo, M., Zannoni, G.F., Beltrame, L., Martinelli, E., DiFeo, A., Paracchini, L., Craparotta, I., Mannarino, L., Vizzielli, G., Scambia, G., D'Incalci, M., Romualdi, C., Marchini, S., 2016. Identification of high-grade serous ovarian cancer miRNA species associated with survival and drug response in patients receiving neoadjuvant chemotherapy: A retrospective longitudinal analysis using matched tumor biopsies. Annals of Oncology 27. https://doi.org/10.1093/annonc/mdw007

Romano, M., della Porta, M.G., Gallì, A., Panini, N., Licandro, S.A., Bello, E., Craparotta, I., Rosti, V., Bonetti, E., Tancredi, R., Rossi, M., Mannarino, L., Marchini, S., Porcu, L., Galmarini, C.M., Zambelli, A., Zecca, M., Locatelli, F., Cazzola, M., Biondi, A., Rambaldi, A., Allavena, P., Erba, E., D'Incalci, M., 2017. Antitumour activity of trabectedin in myelodysplastic/myeloproliferative neoplasms. British Journal of Cancer 116. https://doi.org/10.1038/bjc.2016.424

List of abbreviations

ChIP Chromatin Immunoprecipitation

ChIP-Seq Chromatin Immunoprecipitation followed by sequencing

CHOP C/EBP-Homologous Protein

CTRL control/basal

DBPs differentially bound peaks

DBSs double strand breaks

DDIT3 DNA-inducible transcript 3

DEGs differentially expressed genes

DETs differentially expressed transcripts

DMGs differentially mutated genes

DNA-Seq DNA-Sequencing

DOXO 24h after the first dose of doxorubicin

EMA European Medicine Agency

ET trabectedin

ET-15 15 days after the third dose of trabectedin

ET-24 24h after the first dose of trabectedin

ET-72 72h after the first dose of trabectedin

FACS fluorescence-activated cell sorting

FDA Food and Drug Administration

FUS fused in sarcoma

GSEA gene set enrichment analysis

HR homologous recombination

IGF insulin-like growth factor

IgG non-nuclear antigen

MLPS myxoid liposarcoma

NES normalised enrichment score

NGS next generation sequencing

NY-ESO-1 Autoimmunogenic Cancer/Testis Antigen

PDX patient-derived xenograft

PBMNC peripheral blood mononuclear cells

RNA-Seq RNA-sequencing

SCNA somatic copy number alteration

SNV somatic nucleotide variant

TC-NER transcription-coupled nucleotide excision repair

TLS translocated in sarcoma

TME tumour microenvironment

TPM transcripts per million

TSS transcription start site

UTR untranslated region

WES Whole Exome Sequencing

WGS Whole Genome Sequencing

TABLE OF CONTENTS

1 INTRODUCTION

Sequencing technologies have revolutionised biological research. They scaled up the genomic assay and paved the way to the development of a novel field that is known as computational biology or bioinformatics (Markowetz, 2017), (Bourne, 2021). Oncology research received a significant contribution from this innovation. Indeed, computational approaches have broadened our understanding of cancer biology, from tumour characterisation to tumour evolution and heterogeneity. The transformation did not address only basic and pre-clinical research, but it remodelled the approach to the clinic. Sequencing technologies promise to improve clinical intervention by better patient stratification leading to personalised medicine.

This PhD work contributes to this field. The overall data presented in this project are an example of how computational approaches elucidate tumour biology and provide hypotheses for pharmacological strategies. Indeed, in this work the different levels of cancer complexity are exploited and integrated using computational approaches for DNA- and RNA-Sequencing to characterise myxoid liposarcoma (MLPS) and shed light on the mechanisms of sensitivity to and resistance against trabectedin.

MLPS is a mesenchymal tumour that comes from the adipocytic tissue. MLPS cells are blocked in a permanent undifferentiated state caused by the FUS-DDIT3 oncoprotein that originates from the translocation t(12;16)(q13;p11)(Yu et al., 2019). MLPS is sensitive to chemotherapy and radiotherapy in contrast to other soft-tissue sarcomas, however few therapeutic options are available (Mocellin, 2021). In case of refractory tumours, the pharmacological option is represented by trabectedin. This is a marine-derived compound, today produced by chemical synthesis, that has a complex mechanism acting on both cancer cells and the surrounding microenvironment (D'Incalci and Galmarini, 2010).

The work of the PhD candidate will deal with the use of bioinformatics and computational tools to dissect the complex mechanism of action of trabectedin in MLPS and unveil the molecular processes that lead to acquired-resistance.

2 BACKGROUND

2.1 MYXOID LIPOSARCOMA

Myxoid liposarcoma (MLPS) is a rare tumour that originates from mesenchymal cells. It accounts for 20-30% of liposarcomas and 5% of soft tissue sarcomas (Mocellin, 2021). The unique molecular feature that specifically characterises MLPS is the chromosomal translocation t(12;16)(q13;p11) that leads to the expression of the chimeric oncoprotein FUS-DDIT3 present in 95% of all cases (*Figure 2.1*), while the remaining 5% is characterised by the alternative translocation t(12;22)(q13;q12) that generates EWSR1-DDIT3 (Mocellin, 2021), (Yu et al., 2019).

2.1.1 EPIDEMIOLOGY AND PATHOLOGY

MLPS usually presents in the deep soft tissues of the extremities, such as the thighs. It affects mainly young adults in the range of 30-50 years, while it is rare in childhood. No differences have been observed between males and females (Mocellin, 2021), (Lansu et al., 2020).

Clinically MLPS presents as a large painless mass. At histological level, it is composed of oval non-lipogenic cells and various lipoblasts in a myxoid stroma. The capillary vasculature is typically branching, thus the name of “chicken wire”. MLPS usually lacks nuclear pleomorphism and giant tumour cells, the level of mitotic activity is low and necrosis is absent. However, these features characterise the low-grade MLPS; indeed, a subgroup of MLPS shows a high density of round cells with a high nucleus-to-cytoplasm ratio and lack of myxoid stroma. These high-grade tumours are called myxoid round cell liposarcomas and are associated with a worse prognosis (Grosso et al., 2009). The overall recognised prognostic factor for MLPS is the percentage of round cells, indeed when more than 5% it is an index of bad prognosis (Mocellin, 2021). More recently, the GEIS-20 (Spanish Group for Research in Sarcoma) has identified p53 and FAS as markers of bad prognosis in primary MLPS tumours (Assi et al., 2019).

Despite the different histological presentation and related prognosis, myxoid round cells and low cellularity tissues represent an histological continuum occurring as a gradual transition from one another. This is further confirmed by the fact that both subtypes carry the same chromosomal translocation (Mocellin, 2021), (Lansu et al., 2020).

2.1.2 FUS-DDIT3

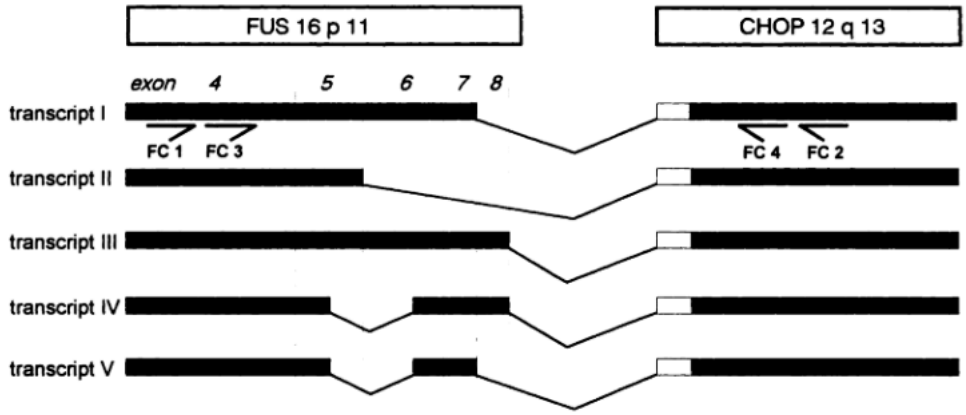

Figure 2.1 Schematic representation of exons involved in FUS-DDIT3 chimera. DDIT3 is here termed as CHOP. Adapted from Willeke et al. 1998.

The detection of the typical FUS-DDIT3 oncoprotein represents the best diagnostic approach for MLPS, especially in the case of morphological similarities with non-adipocytic sarcomas with round cell pattern (Compton and Al-Rohil, 2021) and is usually identified through immunohistochemistry (Scapa et al., 2021).

As depicted in *Figure 2.1*, FUS-DDIT3 derives from a chromosomal translocation that leads to the fusion of *FUS* and *DDIT3*, two genes that have distinct functional roles in the cell. In particular, the *FUS* gene (*i.e.*, fused in sarcoma), also known with the acronym *TLS* (*i.e.*, translocated in sarcomas) belongs to the FET family of RNA-binding protein and has a role in chromatin remodelling and gene expression regulation. The *DDIT3* gene (*i.e.*, DNA damage inducible transcript 3), also known as *CHOP* (*i.e.*, C/EBP homologous protein), belongs to the CCAAT/enhancer-binding protein (C/EBP) family of transcription factors (Crozat et al., 1993). It is a cell-stress marker and it is involved in adipogenesis and erythropoiesis programs. The FUS-DDIT3 oncoprotein exists in twelve known different isoforms with breakpoints always occurring in intronic regions (Antonescu et al., 2000).

All of them involve exons from 2 to 4 of *DDIT3*, an intriguing molecular feature since exon 2 of the *DDIT3* gene is an untranslated region (UTR) in its wild-type form. Otherwise, the exons of the *FUS* gene involved in the chimera determine the type of FUS-DDIT3 isoforms (*Figure 2.1*). The most common isoform is the type II, in which the *FUS* gene provides the chimera with the exons from 1 to 5. In type I isoform the *FUS* gene contributes with exons from 1 to 7, while in the type III isoform the *FUS* gene contributes with exons from 1 to 8 (Willeke et al., 1998). Studies performed so far have demonstrated the absence of clinical significance and no correlation with outcome associated with FUS-DDIT3 variants, therefore the identification of the exact isoform is no longer part of the clinical practice (Mocellin, 2021).

2.1.3 FUS-DDIT3 MOLECULAR MECHANISM

The molecular mechanism of the oncogenic protein FUS-DDIT3 in MLPS has been described in detail by Pérez-Mancera et al. (Pérez-Mancera et al., 2008). According to the current accepted model, the FUS-DDIT3 chimera inhibits the positive loop between the PPARγ and C/EBPα transcription factors thus blocking the process that transforms a mesenchymal cell into a fully differentiated adipocyte (*Figure 2.2*). Although this study showed evidence of the block of the differentiation process in MLPS, FUS-DDIT3 is a DNA-binding oncoprotein and its targets at the genome-wide level have not been identified yet, thus it is likely to be involved in other not already known functions.

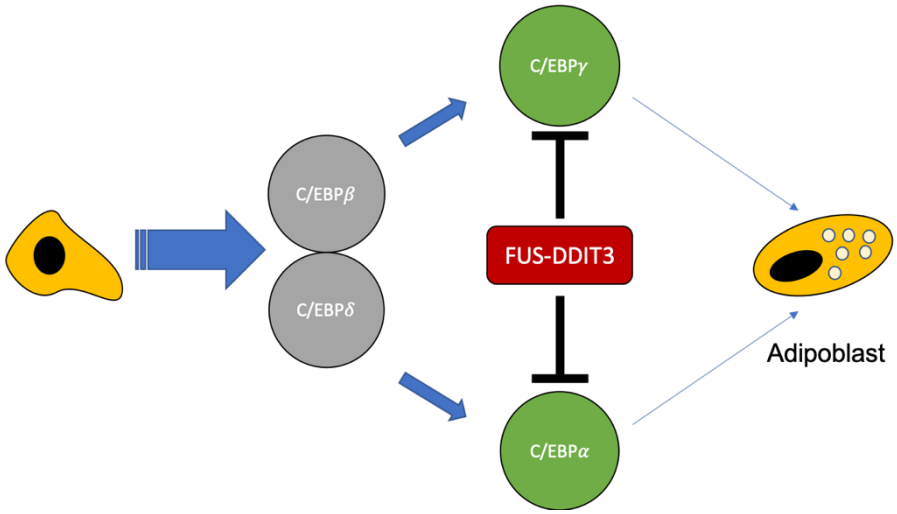

Figure 2.2 FUS-DDIT3 blocks adipocyte differentiation by the inhibition of PPARγ and C/EBPα. Modified and adapted from Pérez-Mancera et al., 2008.

2.1.4 Other molecular features associated with MLPS

Although FUS-DDIT3 remains the only molecular marker that uniquely characterises the genome of MLPS, high-throughput approaches have revealed previously unknown genomic alterations. For example, studies have reported high frequency of mutations in the *PTEN* and *PIK3CA* genes (Mocellin, 2021), (Keung and Somaiah, 2019) as well as overexpression of *RET*, *IGF1R* and *IGF2* have been reported (Assi et al., 2019).

Gains have been reported in chromosomes 1, 5, 8 and in the q arm of 13 (Ohguri et al., 2006), amplifications of *11q22* where the *YAP1* gene maps and *14q22.2* (Kanojia et al., 2015) and losses in chromosomes 6 and 16q have also been reported (Koczkowska et al., 2017). The amplification of the promoter region of TERT, a feature that is associated with tumour aggressiveness in other cancers, has been reported in Hofvander et al. (Hofvander et al., 2018). Overall, different studies agree on the monoclonal nature of MLPS and reported a slow clonal evolutionary pattern associated with a low mutational burden (Antonescu et al., 2000), (Hofvander et al., 2018).

2.1.5 Therapeutic approaches

In comparison to other types of sarcoma, MLPS is particularly responsive to radiotherapy that is usually applied as adjuvant therapy after surgery and has been established as a standard intervention, however new therapeutics are needed (Noguchi et al., 2020). Surgery is usually recommended in low-grade MLPS when the disease is mainly localised. For more advanced tumours surgery is used alone if tumour is less than 5 cm otherwise surgery is followed by adjuvant chemotherapy (Mocellin, 2021).

Patients with advanced tumours or metastatic diseases are treated with anthracyclines, like doxorubicin, that are still the gold-standard in MLPS therapy. Second-line treatment consists in the administration of gemcitabine plus docetaxel and high-dose of ifosfamide. In other cases, in second-line therapy the marine agent trabectedin is usually administered, especially in case of first-line anthracyclines failure (*Figure 2.3*). Trabectedin treatment of MLPS is further discussed in Section 2.3 of this chapter.

Recently, eribulin has been introduced as a second-line therapy for patients with metastatic liposarcoma and has been proven to be effective in a case that did not respond to trabectedin in second-line (Tamiya et al., 2020).

Intervention on the tumour microenvironment (TME) has not been attempted in MLPS, however past studies have shown the high and specific expression of the cancer-testis antigen NY-ESO-1 (Hemminger et al., 2013) that has been proven to be also a specific marker for myxoid round cell liposarcoma (Hemminger and Iwenofu, 2013). This evidence has led to a phase I study testing the safety, tolerability and immunogenicity of a vaccine called CMB305 that acts against NY-ESO-1 antigen (Somaiah et al., 2020).

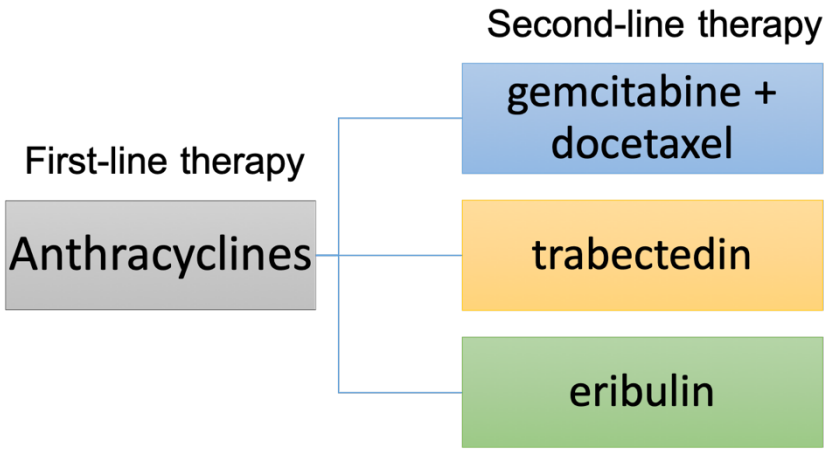

Figure 2.3 First line treatment suggested for advanced unresectable metastatic MLPS and the three available second-line options.

2.2 PATIENT-DERIVED XENOGRAFT

An important gap in the clinical implementation of novel therapeutics is the high discrepancy between drug efficacy assessed in the preclinical phase and the real outcome in phases I/II of clinical trials (Bhimani et al., 2020). Moreover, recent advances in personalised medicine require accurate predictive strategies for pharmacological testing, a point that randomised control trials have failed to some extent especially due to their selective nature (Saturni et al., 2014). Thus, a significant role in pre-clinical but also basic research is held by patient-derived xenografts (PDXs). PDXs are tumour fragments that are explanted from a patient after surgical resection and implanted in immunodeficient mice (Yoshida, 2020), (Xu et al., 2019). They represent a model of the disease and have been especially used in the cancer research field.

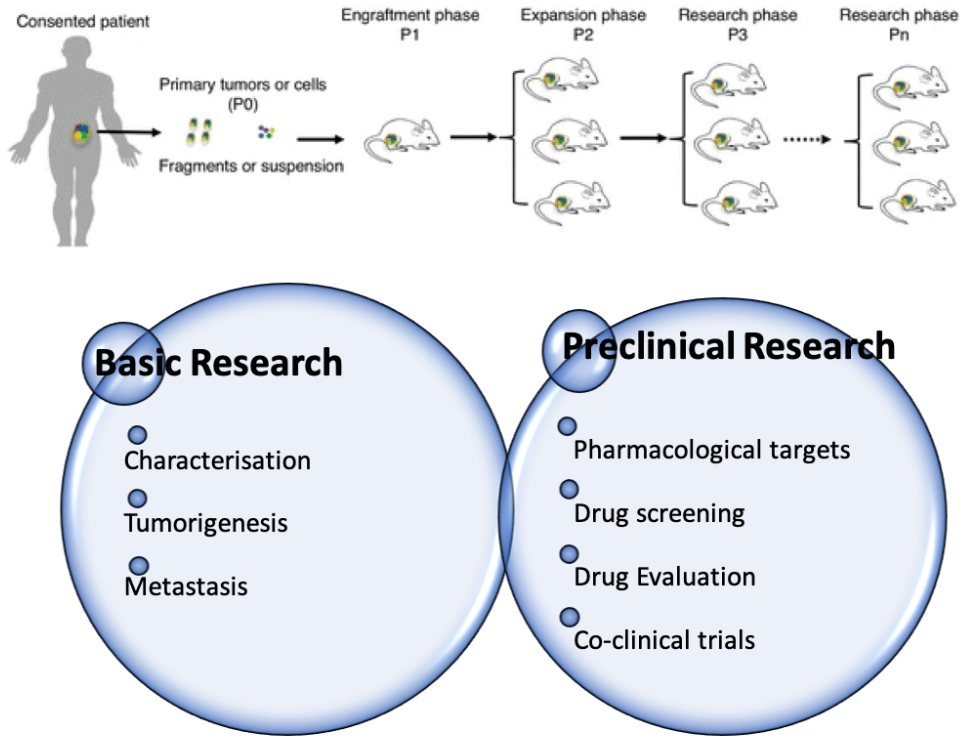

Figure 2.4 Establishment of PDX model from primary human tumours and its applications in basic and pre-clinical research. Modified from Lai et al. 2017.

The advantages of using PDXs models are:

1. they preserve with high accuracy the genomic profile between the explant and the parental tumour they derived from (Yoshida, 2020);
2. the original tumour architecture is fairly persevered (Yoshida, 2020);
3. susceptibility to pharmacological agents is closely correlated to the clinical data the tumours derived from (Lai et al., 2017), (Yoshida, 2020).

Although with some limitations as stated further in this paragraph, these characteristics make PDXs a valid and precious means in both basic and preclinical cancer research as summarised in *Figure 2.4*. In fact, the preservation of the genomic profile of the original tumour allows the characterisation of cancer biology that turns out to be particularly useful in the case of rare cancer diseases for which the availability of tumour tissue from patients is scarce as in the case of MLPS. Moreover, the study of cancer heterogeneity and tumour evolution has been particularly encouraged by PDXs development. Then, tumours implanted in mice have a metastatic potential that mimic that behaviour of the tumours they come from, thus PDXs are extensively used in basic research for the study of metastasis onset and spread.

Notably, the great revolution by PDXs was brought in preclinical cancer research. As stated previously, a high percentage *e.g.* 66% of the therapeutic compounds do not survive clinical trials because of the poor predictive power of preclinical testing (Xu et al., 2019). Otherwise, PDXs have proven to be strong predictive experimental models also for therapeutic response, even more than cell line-derived xenograft or genetically engineered mouse models (Xu et al., 2019). In particular, they have found extensive application in personalised precision medicine (Yoshida, 2020), (Lai et al., 2017). In fact, they can be used for the detection of novel cancer biomarkers and guide selective treatment. For this reason, they are used in the so-called co-clinical trials, where PDXs are developed in concomitance to the progress of the trial. In this way, PDXs can help in patient selection strategies for drug therapy and also predict drug response or development of drug-resistance (Yoshida, 2020), (Xu et al., 2019).

However, PDXs models have some limitations that can be summarised as follows:

1. The human stroma is gradually substituted by that of the mouse, thus the study of the TME is essentially hindered.
2. In nude mice the absence of the immune system represents a substantial limitation to the study of tumours in a real biological context.
3. Usually, aggressive tumours exhibit strong formation rates in mice and easy growing, while early stage or less aggressive tumours grow less. This is a problematic issue when considering patients in clinical trials with initial tumours, indeed they are the ones that might benefit from improved treatment outcomes in comparison to advanced-stages patients for whom PDXs development is easier.
4. The cost and maintenance of the animal enclosure are fairly high.

2.2.1 PDXs MODELS OF MLPS

The rarity of MLPs has pushed the need for novel cancer models. To this aim, a panel of PDXs models of MLPs has been developed from the long-lasting collaboration between Istituto Nazionale dei Tumori of Milano IRCCS and Istituto di Ricerche Farmacologiche Mario Negri IRCCS, and they will be the subject of this work of thesis. To our knowledge there is only one additional PDX established by a group from the Chinese University of Shanghai (Qi et al., 2017). The timeline with the milestones of MLPS PDX models development is reported in *Figure 2.5* and discussed further.

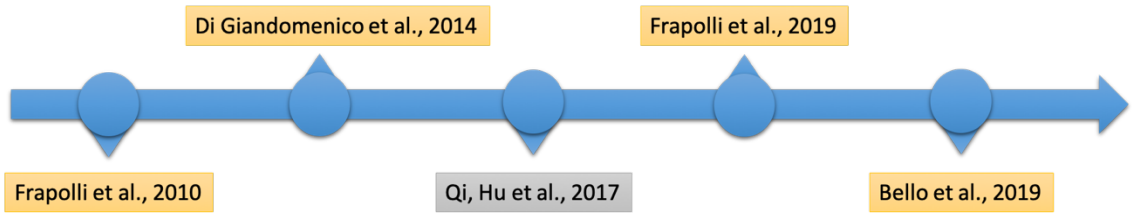

Figure 2.5 Timeline of the studies reporting the establishment and investigation of PDX models of MLPS. Studies from the same research group are in the same colour.

The establishment of these models goes back to 2010 when Frapolli et al. published their first characterisation (Frapolli et al., 2010). The fragments of 17 sarcomas were transplanted in female athymic nude mice. Among them three were classified as usual myxoid liposarcoma and nine as round-cell myxoid liposarcoma, among which those used in this PhD work, namely ML006, ML015 and ML017. These last were selected because

they grew successfully throughout different passages. Specifically, these models retained the constitutive chromosomal translocation of FUS-DDIT3 being of the same isoform of the human tumour (see also Section 2.1.2 of this Chapter): ML017 carried the type I chimera, with a breakpoint at exon 7 of *FUS*; ML015 was a type II chimera, with a breakpoint at exon 5 of *FUS*; ML006 was a type III, with a breakpoint at exon 8 of *FUS* (*Figure 2.1*). Initially, FUS-DDIT3 isoform classification was included in the clinical practice, however the lack of correlation to clinical intervention and clinical outcome led to the abandonment of this molecular classification (Mocellin, 2021). Nevertheless, in the work of Frapolli et al., these models were characterised at both morphologic and molecular level (Frapolli et al., 2010). Specialised pathologists studied the morphology of the tissues cut from PDX through different passages and compared to their matched original tumours. They assessed the high level of overlapping histological features between PDX and tumour specimens. Moreover, in the same study authors used these models as a useful tool to investigate the effects of trabectedin treatment. Interestingly, they showed that the response to the drug was histologically comparable to histological assessment on patients' tissue after treatment, thus further confirming the value of these models to test and exploit pharmacological compounds.

The mechanism of action of trabectedin was further tested in the same models with a specific focus towards the interaction with FUS-DDIT3 chimera (Di Giandomenico et al., 2014). Results from this work suggested a direct role of trabectedin in modifying FUS-DDIT3 interaction with the DNA. In fact, authors showed that trabectedin is able to displace FUS-DDIT3 from the promoters of two target genes, namely *PTX3* and *FN1*, and that this mechanism happens with a different kinetics depending on the type of FUS-DDIT3 isoform.

In 2017, a novel model of MLPS was established by another research group from the Fudan University of Shanghai (Qi et al., 2017). This model was established from a patient affected with myxoid round cell liposarcoma carrying the canonical FUS-DDIT3 chimera and characterised by a mutation in the *PI3KCA* gene. The model was used to investigate the effect of the drug PF-04691502, a PI3K inhibitor.

As discussed previously, in the case of very rare diseases for which research studies are limited due to the lack of statistically required number of samples as in the case of MLPS, the availability of PDX models becomes an extremely helpful means for drug testing. In fact, the same models previously described were studied to investigate the combinatorial effects of trabectedin and pioglitazone, a drug used for the treatment of diabetes (Frapolli et al., 2019). Authors showed that the efficacy of trabectedin in inducing adipocyte differentiation was strongly increased by the co-administration of pioglitazone. Moreover, the combination was effective in overcoming refractory response also in a model that developed resistance to trabectedin. These important evidences were of strong support for the approval of a pilot phase II study testing the combination versus trabectedin treatment alone (NCT04794127, see Background, Section 2.3.4).

Recent developments of these models regard the establishment of a new one called ML017/ET (Bello et al., 2019). This model originated from the previously characterised ML017 (Frapolli et al., 2010) and was established through continuous passages from mouse to mouse in an outstanding process that lasted over two years. Over this period of time, mice were treated with trabectedin till the achievement of resistance to the drug was acquired. This model represents a precious resource to study the mechanisms that lead to acquired-resistance, a big issue that is registered in the clinic.

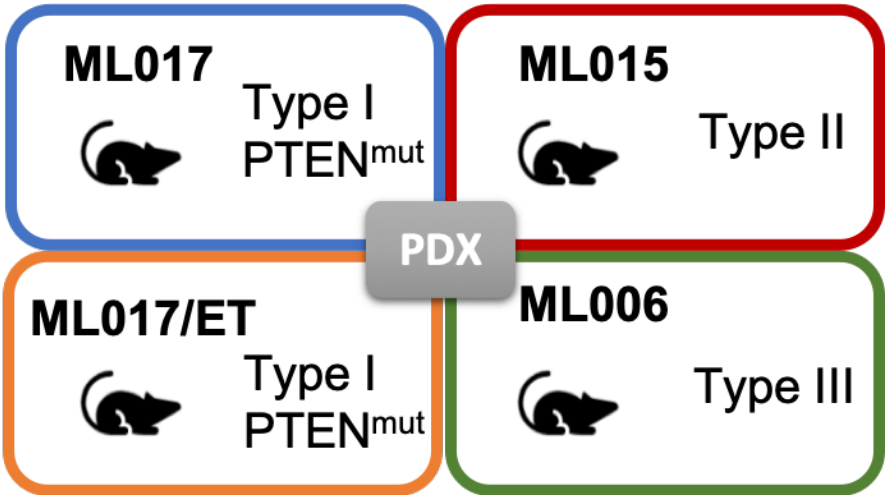

Figure 2.6 PDX models of MLPS developed by Frapolli et al. 2010 and Bello et al. 2019 as discussed in this paragraph. Models are indicated with their identification label. The type of FUS-DDIT3 chimera is reported in each corresponding box.

A summary of all PDX models developed at Istituto di Ricerche Farmacologiche Mario Negri IRCCS is provided in *Figure 2.6*.

Overall, the lack of numerous cohorts for statistical testing due to the rarity of MLPS makes the presented models a precious resource for translational research. As discussed previously, these PDXs have allowed not only a deep tumour characterisation, but also the study of drug effect and mechanism, to test novel combinatorial approaches and to understand the molecular mechanisms leading to drug-resistance (*Figure 2.7*).

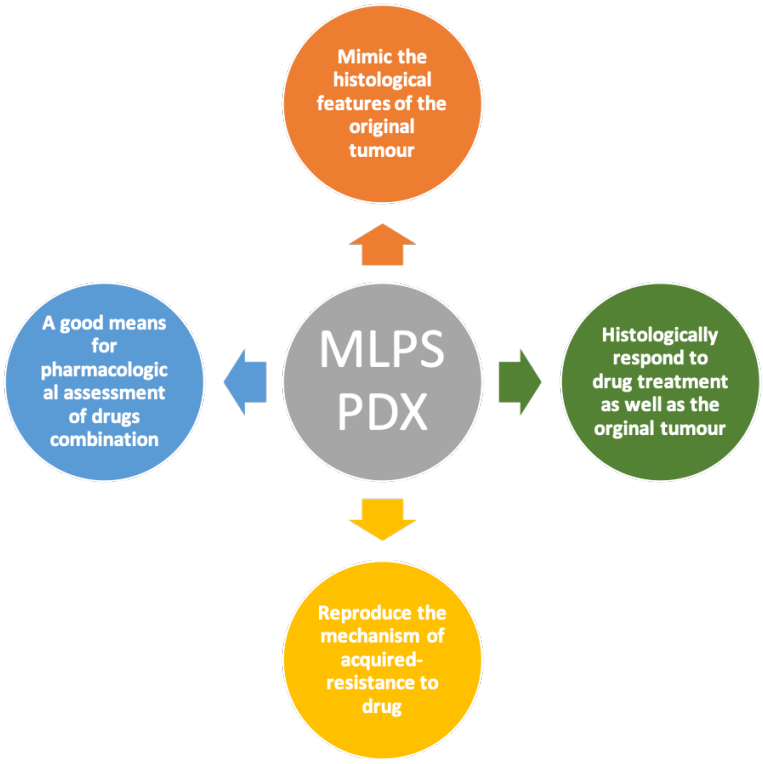

Figure 2.7 Advantages of MLPS PDX.

2.3 TRABECTEDIN

2.3.1 CHEMICAL STRUCTURE AND PROPERTIES

The marine environment has become a rich source of biological active molecules that have found wide use in the clinical treatment of human diseases. These molecules are preferentially found in complex habitats, such as coral reefs, where they have developed specific chemical properties that confer a unique bioactivity (Barreca et al., 2020). Among them, the alkaloid trabectedin, originally known as ET-743, has been widely studied and introduced in clinical practice. It was initially isolated from the Caribbean tunicate *Ecteinascida turbinate*, however the low yield of extraction, *e.g.* 1 g of trabectedin isolated from 1 ton of sea squirt, has prompted the synthetic development of this drug (Barreca et al., 2020).

The chemical formula of trabectedin is $C_{39}H_{43}N_3O_{11}S$ and its chemical structure is shown in *Figure 2.8*. It is composed of three fused tetrahydroisoquinoline rings (A, B and C). A and B rings form a mono-bridged pentacyclic skeleton that is linked to ring C (Barreca et al., 2020), (D'Incalci and Galmarini, 2010). In contrast to traditional DNA major groove alkylating agents that bind to N7 or O6 of guanine, trabectedin binds to the exocyclic N2 amino group of guanines in the DNA minor groove through rings A and B. This DNA adduct is further stabilised by hydrogen bonds and Van der Waals interactions between rings A and B and the neighbouring nucleotides. This structural configuration favours triplets like -TGG, -CGG, -AGC and -GGC with the N2 of middle guanine becoming the sites of alkylation. The binding of the drug in the minor groove induces the formation of DNA adducts and bends the DNA toward the major groove conferring a unique behaviour to this compound. The ring C apparently does not participate in DNA binding, however it protrudes out of the DNA, being able to interact with DNA binding proteins (D'Incalci and Galmarini, 2010).

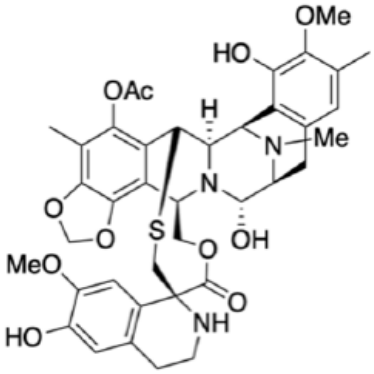

Figure 2.8 Trabectedin chemical structure. Adapted from Barreca et al., 2020.

2.3.2 TRABECTEDIN MECHANISM OF ACTION IN TUMOURS

From its development trabectedin has attracted the attention of researchers due to its complex and pleiotropic mechanism of action. Molecular studies have helped in the understanding of trabectedin activity that can be summarised as depicted in *Figure 2.9* and described in the following:

1. transcriptional regulator;
2. interaction with DNA repair pathways;
3. influence on the tumour microenvironment;
4. antiangiogenic action.

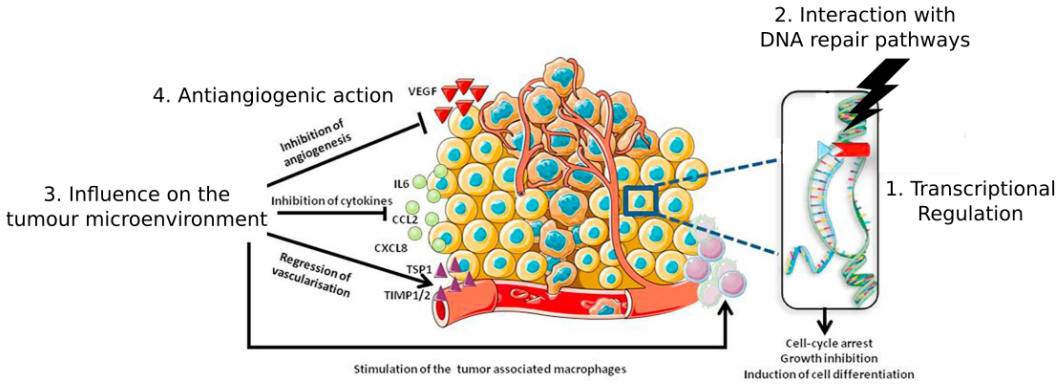

Figure 2.9 Schematic representation of trabectedin mechanism of action. Modified from Assi et al., 2019

It has been demonstrated that the changes induced in the bending of the DNA by trabectedin confer to the drug a role of transcriptional regulator. In fact, modifications in the DNA conformation are an obstacle to DNA binding proteins and can facilitate the binding of other proteins. In particular, at the cell level trabectedin seems to affect only the transcription of activated genes. For example, it effectively blocks HSP-70 and MDR1 (Minuzzo et al., 2000) and it also induces the degradation of RNA polymerase II in cells with a proficient transcription-coupled nucleotide excision repair (TC-NER) (D'Incalci and Galmarini, 2010). Moreover, the regulation of gene transcription has come up to be cell-specific, indeed trabectedin can either up-regulate or down-regulate the same genes in different cell types, thus suggesting its peculiar role as transcriptional regulator (D'Incalci and Galmarini, 2010).

The introduction of trabectedin in the clinical practice has been prompted especially by the response obtained in cells with specific DNA repair deficiencies. Specifically, when defects occur in the homologous recombination (HR) pathway which is involved in the repair of DNA double strand breaks (DSBs) the cells are 100 times more sensitive to trabectedin treatment (D'Incalci and Galmarini, 2010). This evidence was then confirmed in the clinic where BRCA1/2 deficient patients have shown good response to the treatment and led to the European Medicine Agency approval of trabectedin for relapse platinum-sensitive ovarian cancers in 2009 (OVA-301 study, (Monk et al., 2010)). A similar high sensitivity was also found in cells with defective Fanconi anaemia genes, like *FANCA, FANCC, FANCF, FANCD* or *FANCD1* (D'Incalci and Galmarini, 2010).

On the other hand, the response to the drug is completely different in cells lacking a proficient TC-NER. This pathway is involved in the recognition and repair of DNA damages especially caused by UV rays. Molecular studies have shown that TC-NER deficient cells are not responsive to trabectedin treatment and that this mechanism can be due to a mutant form of XPG gene (Takebayashi et al., 2001).

Then, one of the peculiarities of this drug consists in the completion of its action not only on cancer cells but also on the surrounding compartment known as TME. In particular,

trabectedin inhibits the transcriptional activity of cytokines and chemokines like CCL2, CXCL2 and IL6. This activity against monocytes and macrophages is mainly related to the rapid activation of caspase 8 by membrane signalling TRAIL receptors, which are more expressed in these cell types than in neutrophils and lymphocytes (D'Incalci et al., 2014), (Germano et al., 2013).

Finally, studies have shown that trabectedin is able to inhibit important extracellular matrix factors like VEGF and PTX3 exerting an important anti-angiogenic action on the tumour-surrounding capillary network that is further enhanced by the up-regulation of matrix metalloproteinases like TIMP-1 and TIMP-2 inhibitors and also of thrombospondin-1 (D'Incalci and Galmarini, 2010), (Assi et al., 2019).

In conclusion, trabectedin shows a complex mechanism involving different targets and pathways, far to be fully understood. Further molecular studies are necessary to clarify these aspects. Modern sequencing technologies have broadened the possibilities to exploit drug action, thus they can be useful and up-to-date approaches to study trabectedin mechanism of action.

2.3.3 DEVELOPMENT OF RESISTANCE TO TRABECTEDIN TREATMENT

The mechanism of resistance against trabectedin treatment has been described in cancer cell line models. In particular, the loss of heterozygosity of 13q33 (the cytoband to which *XPG* belongs to) explains drug resistance of colorectal carcinoma ER5 cell, while transcriptional alterations of genes related to the cytoskeleton architecture where linked to resistance in chondrosarcoma cell line CS1 (Marchini et al., 2005).

In the clinical practice, patients can sustain prolonged treatment since trabectedin is well tolerated and does not cause cumulative toxicity. However, it is a clinical experience that long lasting chemotherapy usually leads to the development of drug resistance. The molecular reasons at the basis of the acquired resistance against trabectedin in MLPS have not been elucidated yet, however a recent work has proposed a possible link with a deficient TC-NER (Bello et al., 2019).

Recently, the presence of M1 infiltrating macrophages has been associated with poorer responses to trabectedin, as well as genes like *CLU4A* and *ERCC1*, involved in DNA repair processes, are highly predictive of trabectedin response (Assi et al., 2019).

2.3.4 TRABECTEDIN TREATMENT IN MYXOID LIPOSARCOMA

The standard first-line therapy for MLPS consists in the administration of DNA-intercalating agents such as anthracyclines. However, the side effects especially cardiomyopathies and the common onset of drug-resistance have accelerated the individuation of novel therapeutic agents. Initially, trabectedin was approved by the European Medicine Agency (EMA) in 2007 for the treatment of advanced soft-tissue sarcoma, among which MLPS, after the failure of first-line therapy with anthracyclines and ifosfamide (Barreca et al., 2020), (Frapolli et al., 2010), but it was only in 2015 that the Food and Drug Administration (FDA) gave its approval for the treatment of advanced pre-treated metastatic liposarcoma and also leyomiosarcoma (Bello et al., 2019), (Assi et al., 2019).

Due to its rarity, MLPS has been included in clinical trials with other subtypes of sarcoma, thus the real effects of trabectedin on MLPS have been probably shadowed (Assi et al., 2019). The first solid data on trabectedin responsiveness in MLPS has been reported in 2007 by Grosso et al. where only MLPS patients were selected and showed an overall survival rate of 51% (Grosso et al., 2007). From that on, the number of studies on trabectedin in MLPS has increased. Trabectedin has shown efficacy mainly in pre-treated metastatic patients, while studies testing its administration as a first-line treatment did not show an advantage in comparison to other drugs (Assi et al., 2019). However, a study where first-line chemotherapy with trabectedin was compared with the standard therapy with doxorubicin, showed that the growth curves diverged after long time e.g. more than 20 months in favour of trabectedin suggesting a long-latency response to trabectedin treatment (Blay et al., 2014). Interestingly, this evidence was confirmed also in other studies and this led the EMA to the approval for continuous administration till progression (Assi et al., 2019).

The high tolerability to trabectedin treatment with few side effects, *e.g.* liver toxicity, has also led to the clinical practice of interrupting the treatment with rechallenge upon progression (Sanfilippo et al., 2015). This particularly safe pharmacological profile has paved the way for the development of combinatorial treatment that can enhance the response. A study by Forni et al. has shown that when trabectedin was administered as neoadjuvant chemotherapy the tumour cells develop into mature lipoblast, confirming molecular evidence of differentiation (Forni et al., 2009). A recent study by Frapolli et al. has demonstrated the ability of trabectedin to promote adipocyte differentiation in PDX models of MLPS (Frapolli et al., 2019). Interestingly, this effect was enhanced when trabectedin was administered in combination with pioglitazone, a PPARγ agonist used for the treatment of diabetes, and it was also able to overcome trabectedin resistance. This has recently led to the approval of a pilot phase II study (NCT04794127) called TRABEPIO that has the primary end point to assess the percentage of responders when trabectedin is administered in combination with pioglitazone in comparison to trabectedin alone (Mario Negri Institute for Pharmacological Research, 2021).

3 AIMS OF THE STUDY

In this work, we made a step further towards MLPS characterisation and the study of the molecular mechanisms that deal with trabectedin treatment.

The main aims of this PhD work are:

1. to characterise PDX models of MLPS and evaluate the extent to which they reproduce the clinical features of MLPS tumours;
2. to study the mechanism of action of the drug trabectedin at the molecular level;
3. to investigate the mechanisms that lead to the acquisition of resistance against trabectedin treatment in MLPS.

To pursue the three main goals of this work different bioinformatics and computational approaches applied to high-throughput sequencing technologies will be used. In details, starting from a cohort of four PDX models of MLPS (see Background Section 2.2), the work will follow four steps as reported in the following.

1. DNA-sequencing allows to detect genomic variations like single nucleotide variants (SNVs), small insertions or deletions, and also larger genomic instabilities like gene-fusions or somatic copy number alterations (SCNAs). In this PhD work, DNA-Seq will be analysed: i. to detect SNVs and SCNAs that constitutively characterise MLPS PDX; ii. to study whether the sensitivity to trabectedin depends on specific genomic variants: iii. to identify features associated with resistance to the drug.
2. Drugs do not only act at the DNA level, but they can have a role in the modulation of the transcription of target genes leading to activation or inhibition of specific pathways. RNA-Sequencing (RNA-Seq) methods allow the study of different aspects of RNA biology, like gene or transcript expression, translation and also RNA structure (Stark et al., 2019). In this work RNA-Seq will be used for differential expression analysis: i. to identify transcriptional features that discriminate between responsive and resistant tumours; ii. to study the transcriptional mechanisms that lead response to the drug over time. Broader applications of

RNA-Seq have unveiled novel aspects of biology, for example the regulation of non-coding RNAs and the identification of unknown transcripts (Stark et al., 2019). In this work, we will use this approach to explore yet unknown transcriptional features related to the mechanism of responsiveness to trabectedin.

3. Among sequencing approaches, chromatin immunoprecipitation followed by sequencing (ChIP-Seq) allows the detection of transcription factor binding sites genome-wide. The study of histone modifications, such as enhancers or the analysis of the chromatin state are possible with ChIP-Seq. This represents an important approach to analyse biological functions associated with epigenetic signatures and to study their response to treatment (Nakato and Sakata, 2021). FUS-DDIT3 is the only molecular feature that distinguishes MLPS from other tumours. Its role in MLPS has been extensively described, however a study on its target sites at the whole-genome level has not been done yet. To our knowledge, none has already studied how drug treatment can affect FUS-DDIT3 mechanism of binding to the DNA. In this PhD work, ChIP-Seq will be used to tackle these points.
4. All these approaches contribute singularly to reveal a specific aspect of tumour biology. However, the single depicted pictures cannot be considered independently. Altogether, they contribute to rendering a comprehensive description of the biological problem. Thus, in this work all these different aspects will be integrated by computational approaches in order to provide as far as possible a whole picture of MLPS and their response to trabectedin.

4 MATERIALS AND METHODS

Sections from 4.1 to 4.6 were provided for the sake of completeness, although they do not represent a direct work of the PhD candidate.

4.1 ANIMALS

Female athymic nude mice (six to eight weeks old) were obtained from Envigo (RMS, Udine, Italy) and were housed in individually ventilated cages. Sterilised food and water *ad libitum* were provided under specific pathogen-free conditions in the Animal Care Facility of Istituto di Ricerche Farmacologiche Mario Negri IRCCS which meets international standards. Health monitoring, animal welfare supervision, experimental protocols and review of procedures were done by a certified veterinarian. Experiments meet the Italian Governing Law (D.lgs 26/2014; Authorisation no.19/2008-A issued March 6, 2008, by Ministry of Health). The Institutional Regulations and Policies of Istituto di Ricerche Farmacologiche Mario Negri IRCCS were provided for internal authorisation for people conducting animal experiments (Quality Management System Certificate—UNI EN ISO 9001:2008—Reg. No. 6121); the NIH Guide for the Care and Use of Laboratory Animals (2011 edition) and EU directives and guidelines (EEC Council Directive 2010/63/UE) and in line with the guidelines for the welfare and use of animals in cancer research (Workman et al., 2010).

4.2 DRUGS

Trabectedin (YondelisR©, ET743) was kindly supplied by PharmaMar, S.A. (Colmenar Viejo, Spain) dissolved in water and further diluted in saline immediately before use. Doxorubicin (SANDOZ clinical formulation) was diluted with water immediately before use.

4.3 TUMOUR MODELS

The ML017, ML015 and ML006 patient-derived xenograft are models of round-cell myxoid liposarcoma with type I, type II and type III FUS-DDIT3 chimera, respectively. They

were maintained through serial transplantation in mice as reported in Frapolli et al. (Frapolli et al., 2010). Tumours from donor mice were cut into small fragments and then engrafted subcutaneously in athymic nude mice under isoflurane anaesthesia. Histological assessment of tumours in mice was made after each passage to assess the clinical relevance. The ML017/ET model was derived from ML017 model and shows acquired resistance to trabectedin. The development and maintenance of this model is as reported in Bello et al. (Bello et al., 2019). When tumour burden reached about 300 to 400 mg, mice bearing xenografts were randomized to receive trabectedin 0.15 mg/kg intravenously, every 7 days for three times (q7dX3) or doxorubicin 8 mg/kg intravenously every 7 days for two times (q7dx2).

4.4 Extraction and sequencing of DNA and RNA

gDNA and total RNA were extracted from tumours with the QIAamp DNA Mini kit and (QIAGEN) and the miRNeasy Mini kit (QIAGEN) respectively, following protocols' instructions and using an automatic nucleic acid purification system (Qiacube, QIAGEN).

Before library preparation, gDNA and RNA concentration were evaluated using Qubit® dsDNA High Sensitivity Assay Kit and Qubit™ RNA High Sensitivity Assay Kit (Invitrogen, Carlsbad, California, USA) respectively, while the quality was established using 4200 Tapestation (Agilent Technologies).

200 ng of DNA were sheared on Bioruptor (Diagenode, Seraing (Ougrée), Belgium) then purified with AMPure XP beads (Beckman Coulter, Brea, California, USA). Following the Sure Select XT protocol (Agilent Technologies), libraries were generated using the Bravo automatic liquid handling station (Agilent Technologies). OneSeq Constitutional Research Panel (Agilent Technologies) was used as capture probes: a detailed description of the panel is provided in Section 4.8.3.1.1. After the last AMPure XP beads purification, samples were examined for quality and quantity and the sequencing run was done on the NextSeq500 sequencer (Illumina San Diego, California, USA).

Following the TruSeq Stranded Total RNA protocol, 500 ng of RNA (RIN value between 6 and 9) was processed for sequencing. Libraries with optimal quality and quantity were run on the NextSeq500 sequencer (Illumina San Diego, California, USA).

4.5 ChIP and ChIP-Seq

Chromatin immunoprecipitation (ChIP) experiments were performed on PDX tumour biopsies previously fixed in 1% formaldehyde and quenched with 125 mM glycine. Tumour specimens were resuspended in a lysis buffer supplemented with protease inhibitors and homogenized using an ultra-turrax (VWR). After centrifugation, chromatin was sheared on Bioruptor (Diagenode) and added to 10 µg of antibody, previously incubated all day at 4°C in a rotation wheel with dynabeads (Thermo-fisher). 1% of lysate was collected, stored at -20°C and then used as an input sample. After overnight incubation in a rotation wheel at 4°C, immunoprecipitated samples (IPs) were washed with a RIPA wash buffer, dynabeads were removed in a magnetic stand and samples were incubated at 65°C overnight in order to remove crosslinks. Recovered material was purified using QIAquick PCR purification kit (Qiagen) and DNA was quantified using Qubit assay and analysed by qPCR. Antibodies used in ChIP experiments are listed in *Table 4.1*. ChIP-Seq experiments were performed adapting the TruSeq ChIP protocol (Illumina): 50 ng of IPs and input were used for library preparation. Since no size selection was performed, samples were analysed by Tapestation in order to verify the peak size.

Antibody	Code	Company
Rabbit Polyclonal DDIT3	15204-1-AP	Proteintech
Rabbit Histone H3K4Me3	V13-39159	Active Motif
Normal Rabbit IGg	2729S	Cell Signaling

Table 4.1 Antibodies used in ChIP experiments.

4.6 FACS

Fluorescence-activated cell sorting (FACS) analysis was done on ML017 and ML017/ET models only. Tumour tissues were mechanically disaggregated by MediMachine (BD Biosciences, Franklin Lakes, New Jersey, USA). Cell suspension was filtered using a 100

µm CellTrics disposable filter (Sysmex Europe, GmbH, Bornbarch, Norderstedt, Germania) and fixed with ethanol (70%). After 4h, 2 × 106 cells were incubated over night at 4 °C with 1 ml of Propidium Iodide 25 µg/ml (Calbiochem, Merck Burlington, Massachusetts, USA) and 12.5 µl of RNAse 1 mg/ml (Calbiochem, Merck Burlington, Massachusetts, USA) in phosphate buffered saline (PBS). A Beckton Dikinson FACSCalibur flow cytometer (BD Biosciences, Franklin Lakes, New Jersey, USA) equipped with blue (488 nm) and red (630 nm) lasers was used to analyse the DNA. To construct each histogram, 10.000–20.000 cells were used. Instrument was calibrated with peripheral blood mononuclear cells (PBMNC). It was also used as internal diploid standard. Aneuploidy was estimated by the DNA index value, calculated as the ratio between the modal channel of the G0/G1 peak of the sample under study and the modal channel of the G0/G1 peak of the reference standard. For a diploid cell population the DNA index is set to 1.00.

4.7 TREATMENT SCHEDULE

As previously stated (see Background, Section 2.1) MLPS is a rare soft tissue sarcoma with a low incidence in the population, therefore it is very difficult to recruit biological material to perform translational studies. To overcome this issue, PDX models of MLPS have been developed in the Department of Oncology of the Istituto di Ricerche Farmacologiche Mario Negri IRCCS, namely ML017, ML017/ET, ML015 and ML006 as described in Background, Section 2.2.1.

Figure 4.1 depicts the experimental design that was applied to the study. The treatment schedule was established based on the previous work done on these preclinical models at the Department of Oncology of the Mario Negri Institute (Frapolli et al., 2010), (Di Giandomenico et al., 2014). In detail: untreated control samples were used as baseline, a time-course approach for trabectedin treatment was used consisting of 24 and 72 hours after the first dose of drug to study the early response to treatment and one point at 15 days after the third dose to study the late effects. The interval between doses was of one week (see also Materials and methods, Section 4.3). A time point at 24 hours after one dose with the standard first-line drug doxorubicin was used as a comparison.

The experimental design did not include untreated control samples taken in correspondence of each time of treatment since:

- mice randomization was done after 30 days from the inoculation: at this time point and after either 24h or 72h after treatment the tumour maintains its characteristics being in a linear point of the growth curves as shown in the previous work from Frapolli et al. and Di Giandomenico et al. ((Frapolli et al., 2010), (Di Giandomenico et al., 2014));
- at the last time point, *e.g.* 15 days after the third dose, the issue deals again with the tumour growth: mice can either not survive till this time point or have a tumour that reaches a weight of 2g or more before this point, being incompatible with mice survival and requiring mice scarification in accordance to the rules of the welfare and use of animals in cancer research (Workman et al., 2010). Otherwise, mice treated with trabectedin can reach this time-point thanks to the treatment that sustains their survival.

From now on, each time point condition will be addressed as follows:

- basal condition as CTRL;
- trabectedin treatment at 24 hours after the first dose as ET-24;
- trabectedin treatment at 72 hours after the first dose as ET-72;
- trabectedin treatment at 15 days after the third dose as ET-15;
- doxorubicin treatment at 24 hours after the first dose as DOXO.

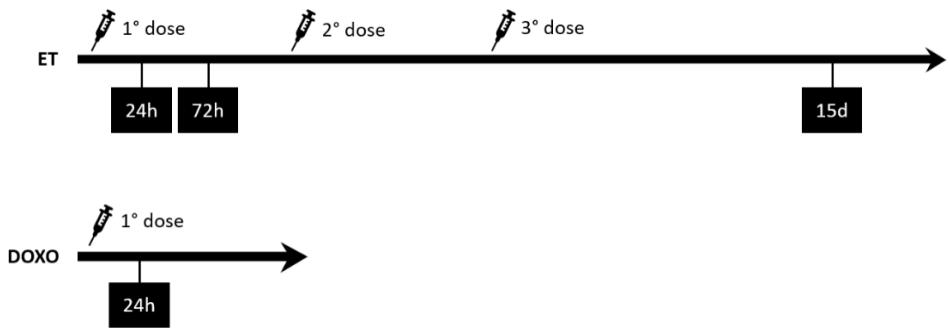

Figure 4.1. Treatment schedule used for PDX models. ET, trabectedin. DOXO, doxorubicin.

As summarised in *Figure 4.2*, in the case of ML017 and ML017/ET models, four replicates per each condition were available and in just two cases out of four replicates were technical since they came from two different tumours of the same mouse (left and right tumours). Thus, a total of 40 samples was analysed for ML017 and ML017/ET, twenty each. Instead, for ML006 and ML015 models replicates were not available, thus only one sample for each condition was analysed.

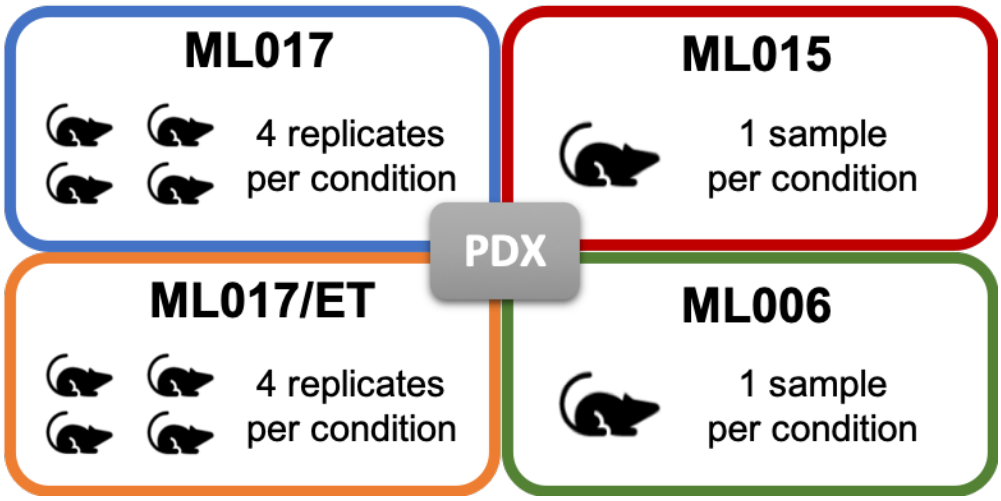

Figure 4.2 Schematic representation of the number of replicates for each PDX model used in this work.

4.8 BIOINFORMATICS DATA ANALYSIS

The implementation of the present work has been done by the complex integration analysis of different types of sequencing data starting from either DNA or RNA (DNA-Seq, RNA-Seq) or chromatin immunoprecipitation (ChIP-Seq). The work of the PhD candidate did not consist only in data analysis oriented to the achievement of the biological aims of the study, but also to test tools, to choose the ones that best applied to the aim, and then organise a specific pipeline of analysis. The methods described here are the final realisation of an accurate and deep research. The world of bioinformatics tools for data analysis is wide and in rapid evolution with a daily birth of novel methods and software. Being aware of the broad range of possibilities available for the development of specific pipelines, the ones presented here are not the only possible choices, however they come from tools already accepted by the scientific community, adapted and customised to the aims of this work as expressly stated and detailed in the following sections.

4.8.1 FIRST STEPS IN DATA ANALYSIS

The first steps in bioinformatics data analysis were applied to all the technologies used in this work and shown in *Figure 4.3*.

First, processed sequencing libraries were loaded in the NextSeq500 Illumina sequencer with a high output flow cell (output of 60 Giga bases, 400 million reads). Although some changes could have been done due to technical reasons, overall experiments and loading were made as reported in *Table 4.2* and described as follows.

- DNA-Seq was done following these parameters: 28 Mb for the genome size, 300 (2x150) cycles for read length, paired-end, single index, 70% of on-target bases, 20% of duplicates and 150x of coverage. Based on these parameters, 16 samples were loaded in each flow cell as calculated by the Illumina Calculator ("Sequencing Support – Coverage Calculator," n.d.);
- RNA-Seq was run in paired-end mode, 300 (2x150) cycles for read length, single index, calculating around 60 million reads per sample, thus 7 samples per flow cell were loaded;

- ChIP-Seq was run in single-end mode, single index, 75 bp of fragment length. 10 samples per flow cell were run, expecting 40 million reads for both transcription factors and histone modification H3K4me3, while for the input samples 4 samples were loaded per flow cell expecting 100 million reads each.

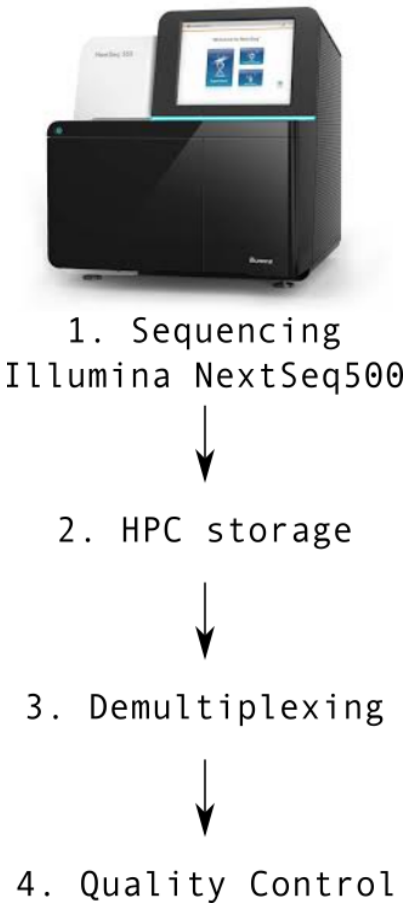

Figure 4.3 First steps in bioinformatics data analysis.

Raw data were stored in a high performance computing (HPC) cluster based at the Istituto di Ricerche Farmacologiche Mario Negri IRCCS. Raw sequencing reads were demultiplexed with Illumina *bcl2fastq* Conversion Software v1.8 ("bcl2fastq Conversion Software," n.d.). In the last step, quality assessment of sequencing data was performed with FastQC ("Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data," n.d.) as further explained.

	Fragment length (bp)	Paired-end	Samples per flow cell	Expected output (reads)
DNA-Seq	150	Yes	16	25 M
RNA-Seq	150	Yes	7	57 M
ChIP-Seq (TF)	75	No	10	40 M
ChIP-Seq (input)	75	No	4	100 M

Table 4.2 Sequencing loading specifications for the sequencing technologies used. bp, base pair; TF, transcription factor; M, million.

4.8.2 QUALITY CONTROL

Quality control is a compulsory step in sequencing data analysis. It allows to check the quality of each sequenced base and that of the overall sample cohort. The aim of this process is to identify samples that do not reach quality standards and to exclude them from further analysis. Data quality assessment is important since post-processing of low quality samples could lead to false positive results and misleading biological findings.

Overall, the quality of a sequenced base is expressed through the Q-score that is defined as in *Equation 4.1*. Q-score information is stored in *fastq* files (raw data) and is expressed through ASCII symbols.

$$Q - score = -10 \log_{10} p$$

Equation 4.1 Q-score definition. P is the probability of a wrong base call.

In this work quality control of sequenced samples was done before alignment through FastQC ("Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data," n.d.) that allows to retrieve quality parameters based on the Q-scores, while post-alignment quality control was done with samtools (Li et al., 2009) and bcbio-nextgen ("Contents — bcbio-nextgen 1.2.4 documentation," n.d.).

In general, the sequenced fragments are expected:

- To have the putative sequence length, e.g. in this work 150 bp for DNA-Seq and 75 bp for ChIP-Seq;

- To have a low percentage of sequence error rate defined from CIGAR strings as in *Equation 4.2* where CIGAR stands for Concise Idiosyncratic Gapped Alignment Report that is a compressed representation of alignments in SAM formats;
- To achieve the desired coverage;
- To have a percentage of mapped reads proximal to 100%.

$$error\ rate = {}^{mismatched\ bases}/_{mapped\ bases}$$

Equation 4.2 Error rate definition based on CIGAR strings.

4.8.2.1 Quality control of DNA sequencing data

ML006 samples are reported in *Table 4.3* with the associated quality parameters on pre- and post-aligned data. For each condition, such as CTRL, ET-24, ET-72, ET-15 and DOXO (see Materials and methods Section 4.7), one sample was sequenced. A commercial reference genome was provided by Agilent Technologies and used as matched control for somatic variant detection as further explained. All samples matched quality standards with a mean sequence depth of 105.05x and more than 90x each, a percentage of duplicate less than 6%, a percentage of error rate less than 1%, and 99.8% of reads mapped of which a mean of 2.68 million were mapped on the mouse genome as further described and excluded from the final aligned samples.

Similarly, ML015 samples matched quality standards as reported in *Table 4.4*. Again, only one sample was sequenced for each condition, while a healthy sample from the original patient the PDX derived from was used as control for variant calling purposes as further explained. All samples matched quality standards with a percentage of duplicated sequences less than 6%, mean sequence depth of 108.66x excluding the healthy sample that was sequenced at lower coverage; the percentage of error rate was less than 1%, and 98.75% of reads were correctly mapped of which a mean of 2.03 million were mapped on the mouse genome as further described and excluded from the final aligned samples.

In relation to the analysis of ML017 and ML017/ET models, four replicates per condition were used as reported in *Table 4.5* and *Table 4.6* (see also Materials and methods, Section 4.7). Samples were referred either as -A or -B since they belonged to two different batches of mice (two different period of time) and numbers like 1- and 2- indicated the replicate number. The same nomenclature was maintained with the other sequencing approaches as reported in the next sections. One healthy sample from the patient ML017 derived from was used as control for variant calling. The two cohorts met sequencing quality standard: a mean of 4.9% and 6.28% of duplicate sequences respectively, less than 2.10% of error rate and more than 99% of mapped reads, of which 3.32 and 2.16 million mapped to the mouse genome in ML017 and ML017/ET, respectively.

Overall, all samples met quality standards both before and after alignment, thus they were all considered for downstream analysis.

	FastQC				bcbio-nextgen			samtools				
	% Dups	% GC	Length bp	% Failed	% Dup	Depth	M disamb. mm10 reads	% Error rate	M Non-Primary	M Reads Mapped	% Mapped	M Total seqs
ML006_CTRL	3.3	46	148	8	5.3	105.31	2.65	0.58	0.2	52.1	99.8	52.2
ML006_DOXO	3.0	45	148	8	4.8	98.19	2.68	0.59	0.2	49.4	99.8	49.5
ML006_ET-24	3.4	45	149	25	5.1	93.32	2.74	0.59	0.2	47.2	99.8	47.3
ML006_ET-72	3.4	45	148	25	5.6	94.53	2.46	0.62	0.2	50.3	99.8	50.4
ML006_ET-15	3.7	45	149	17	5.7	97.5	2.88	0.60	0.1	49.9	99.8	50
REF	3.4	48	148	17	5.2	141.45	-	0.61	0.3	59.9	99.8	60

Table 4.3 Quality control of DNA-sequenced samples from ML006 and the commercial reference (REF). FastQC parameters (derived from down-sampling on chromosome 1): %Dups, percentage of duplicate reads; %GC, percentage of average content of CG bases; Length, average sequence length (bp); %Failed, percentage of modules failed in FastQC report. Bcbio-nextgen parameters (on the post-aligned whole bam files): %Dup, % duplicated reads; Depth, average target read coverage; M disamb. mm10 reads, this metric shows the number of removed reads because ambiguously mapped on mm10. Samtools parameters: % Error rate, percentage of error rate using CIGAR; M Non-primary, Non-primary alignments (millions); M Reads Mapped, reads in the bam file (millions); % Mapped, percentage of mapped reads in the bam file; M Total seqs, total sequences in the bam file (millions). CTRL, control; ET, trabectedin; 24 and 72 refer to 24 and 72 hours after the first dose of drug; 15 refers to 15 days after the third dose of drug; DOXO, doxorubicin; REF, commercial reference genome.

	FastQC				bcbio-nextgen			samtools				
	% Dups	% GC	Length bp	% Failed	% Dup	Depth	M disamb. mm10 reads	% Error rate	M Non-Primary	M Reads Mapped	% Mapped	M Total seqs
ML015_CTRL	3.5	46	148	17	5.9	98.65	0.98	0.60	0.2	50.1	99.8	50.2
ML015_DOXO	2.9	46	148	17	4.5	118.1	1.25	0.58	0.2	58.9	99.8	59
ML015_ET-24	3.3	46	148	17	5.0	108.8	0.83	0.61	0.2	55.8	99.8	55.9
ML015_ET-72	3.7	46	148	8	4.5	123.18	1.57	0.60	0.2	59.5	99.8	59.6
ML015_ET-15	3.6	46	149	17	5.3	94.57	5.52	0.61	0.2	48.7	99.8	48.8
healthy	5.3	51	144	25	5.3	66	-	1.09	0	32	97.7	32.8

Table 4.4 Quality control of DNA-sequenced samples from ML015 and matched healthy sample. FastQC parameters (derived from down-sampling on chromosome 1): %Dups, percentage of duplicate reads; %GC, percentage of average content of CG bases; Length, average sequence length (bp); %Failed, percentage of modules failed in FastQC report. Bcbio-nextgen parameters (on the post-aligned whole bam files): %Dup, % duplicated reads; Depth, average target read coverage; M disamb. mm10 reads, this metric shows the number of removed reads because ambiguously mapped on mm10. Samtools parameters: % Error rate, percentage of error rate using CIGAR; M Non-primary, Non-primary alignments (millions); M Reads Mapped, reads in the bam file (millions); % Mapped, percentage of mapped reads in the bam file; M Total seqs, total sequences in the bam file (millions). CTRL, control; ET, trabectedin; 24 and 72 refer to 24 and 72 hours after the first dose of drug; 15 refers to 15 days after the third dose of drug; DOXO, doxorubicin.

	FastQC				bcbio-nextgen			samtools				
	% Dups	% GC	Length bp	% Failed	% Dup	Depth	M disamb. mm10 reads	% Error rate	M Non-Primary	M Reads Mapped	% Mapped	M Total seqs
ML017-1-CTRL-A	3.0	46	151	9	4.7	98	2.6	0.60	0.2	57.5	99.8	57.6
ML017-1-CTRL-B	3.1	47	151	9	4.3	90	1.4	1.05	0.2	50.9	99.7	51.1
ML017-2-CTRL-A	3.3	47	151	9	4.3	99	2.0	1.05	0.2	55.7	99.7	55.9
ML017-2-CTRL-B	3.6	49	151	0	4.6	116	2.6	1.05	0.2	59.5	99.7	59.7
ML017-1-DOXO-A	2.8	47	150	0	3.8	83	1.4	0.63	0.2	48.1	99.8	48.2
ML017-1-DOXO-B	3.1	47	151	18	3.8	82	1.2	0.93	0.2	48.7	99.7	48.8
ML017-2-DOXO-A	3.0	46	151	9	4.7	83	1.3	1.15	0.1	48.9	99.7	49.1
ML017-2-DOXO-B	3.6	47	151	9	4.5	114	2.2	0.96	0.2	64.3	99.7	64.5
ML017-1-ET-24-A	2.9	46	151	0	4.3	99	2.5	0.58	0.2	57.8	99.8	57.9
ML017-1-ET-24-B	3.1	46	151	18	4.7	129	2.3	0.91	0.2	76.7	99.7	76.9
ML017-2-ET-24-A	3.5	49	151	9	5.5	110	2.3	1.04	0.3	56.9	99.7	57.1
ML017-2-ET-24-B	3.0	47	151	9	4.2	107	3.8	1.01	0.2	61.5	99.7	61.7
ML017-1-ET-72-A	3.2	46	151	9	5.0	113	4.3	0.55	0.2	64.9	99.8	65.1
ML017-1-ET-72-B	3.0	47	151	18	4.2	117	3.4	0.94	0.2	68.6	99.7	68.8
ML017-2-ET-72-A	2.9	46	151	9	4.3	91	2.9	1.02	0.1	51.6	99.7	51.7
ML017-2-ET-72-B	2.9	47	151	9	4.4	100	3.1	0.99	0.2	56.4	99.7	56.6
ML017-1-ET-15-A	3.5	46	151	9	6.6	65	12.6	0.57	0.1	39.7	99.8	39.7
ML017-1-ET-15-B	3.4	47	151	9	3.9	81	5.1	0.97	0.1	46.1	99.7	46.3
ML017-2-ET-15-A	2.6	46	151	9	4.6	122	4.1	1.03	0.2	72.4	99.7	72.6
ML017-2-ET-15-B	2.8	46	151	9	5.0	91	5.4	1.02	0.1	54	99.7	54.2
healthy	4.3	46	134	9	11.5	43	-	1.33	3.1	36.4	99.5	36.6

Table 4.5 Quality control of DNA-sequenced samples from ML017 and matched healthy sample. FastQC parameters (derived from down-sampling on chromosome 1): %Dups, percentage of duplicate reads; %GC, percentage of average content of CG bases; Length, average sequence length (bp); %Failed, percentage of modules failed in FastQC report. Bcbio-nextgen parameters (on the post-aligned whole bam files): %Dup, % duplicated reads; Depth, average target read coverage; M disamb. mm10 reads, this metric shows the number of removed reads because ambiguously mapped on mm10. Samtools parameters: % Error rate, percentage of error rate using CIGAR; M Non-primary, Non-primary alignments (millions); M Reads Mapped, reads in the bam file (millions); % Mapped, percentage of mapped reads in the bam file; M Total seqs, total sequences in the bam file (millions). CTRL, control; ET, trabectedin; 24 and 72 refer to 24 and 72 hours after the first dose of drug; 15 refers to 15 days after the third dose of drug; DOXO, doxorubicin.

	FastQC				bcbio-nextgen			samtools				
	% Dups	% GC	Length bp	% Failed	% Dup	Depth	M disamb. mm10 reads	% Error rate	M Non-Primary	M Reads Mapped	% Mapped	M Total seqs
ML017/ET-1-CTRL-A	3.6	45	149	9	6.7	87	1.9	0.87	0.1	55.3	99.7	55.4
ML017/ET-1-CTRL-B	2.4	45	151	9	5.9	45	1.4	1.97	0.1	31.3	99.5	31.4
ML017/ET-2-CTRL-A	3.0	45	151	0	5.7	73	1.4	1.78	0.2	48.4	99.6	48.6
ML017/ET-2-CTRL-B	3.3	44	151	0	5.7	68	11.4	1.58	0.2	45.4	99.7	45.6
ML017/ET-1-DOXO-A	3.2	45	148	18	6.4	120	3.7	0.80	0.2	70.4	99.7	70.6
ML017/ET-1-DOXO-B	2.7	46	151	9	5.7	114	1.7	1.67	0.3	70.7	99.6	71
ML017/ET-2-DOXO-A	3.2	46	150	9	6.0	105	2.1	1.64	0.4	66.7	99.7	67
ML017/ET-2-DOXO-B	3.8	45	151	0	5.0	90	1.1	1.68	0.2	56.4	99.7	56.5
ML017/ET-1-ET-24-A	4.5	46	147	9	11.6	99	1.8	0.76	0.2	62.5	99.7	62.6
ML017/ET-1-ET-24-B	2.3	44	151	9	5.5	53	0.9	1.94	0.1	36.1	99.5	36.2
ML017/ET-2-ET-24-A	3.2	45	151	9	5.8	76	1.5	1.78	0.2	49.1	99.6	49.3
ML017/ET-2-ET-24-B	2.5	45	151	9	6.0	99	1.9	1.69	0.2	64.8	99.6	65
ML017/ET-1-ET-72-A	3.2	46	148	18	6.0	113	2.9	0.79	0.2	68.4	99.7	68.6
ML017/ET-1-ET-72-B	2.6	44	151	9	6.1	54	1.1	2.01	0.1	37.2	99.5	37.4
ML017/ET-2-ET-72-A	3.5	45	151	0	6.5	90	1.6	1.76	0.2	60.3	99.6	60.5
ML017/ET-2-ET-72-B	3.0	46	150	9	5.5	115	1.9	1.69	0.4	68.7	99.7	68.9
ML017/ET-1-ET-15-A	3.1	46	148	9	7.0	120	1.2	0.83	0.2	72.7	99.7	73
ML017/ET-1-ET-15-B	2.6	45	151	0	5.2	46	0.7	1.84	0.1	29.4	99.5	29.5
ML017/ET-2-ET-15-A	3.4	46	151	9	7.4	76	0.7	1.67	0.2	46.9	99.6	47.1
ML017/ET-2-ET-15-B	3.1	46	151	9	6.0	93	2.3	1.60	0.3	57.7	99.7	57.8

Table 4.6 Quality control of DNA-sequenced samples from ML017/ET. FastQC parameters (derived from down-sampling on chromosome 1): %Dups, percentage of duplicate reads; %GC, percentage of average content of CG bases; Length, average sequence length (bp); %Failed, percentage of modules failed in FastQC report. Bcbio-nextgen parameters (on the post-aligned whole bam files): %Dup, % duplicated reads; Depth, average target read coverage; M disamb. mm10 reads, this metric shows the number of removed reads because ambiguously mapped on mm10. Samtools parameters: % Error rate, percentage of error rate using CIGAR; M Non-primary, Non-primary alignments (millions); M Reads Mapped, reads in the bam file (millions); % Mapped, percentage of mapped reads in the bam file; M Total seqs, total sequences in the bam file (millions). CTRL, control; ET, trabectedin; 24 and 72 refer to 24 and 72 hours after the first dose of drug; 15 refers to 15 days after the third dose of drug; DOXO, doxorubicin.

4.8.2.2 Quality control of RNA sequencing data

RNA-Seq was performed on ML017 and ML017/ET only, due to reasons reported in Results Section 5.2.1. Four biological samples were sequenced for each condition as described previously and reported in *Table 4.7* and *Table 4.8*.

The mean percentage of duplicated reads was 67.8% in ML017 and 60% in ML017/ET. The percentage of identified ribosomal RNA (rRNA) that could have contaminated human RNA was less 0.25% in all samples. The percentage of error rate was less than 0.7% and the mean percentage of reads properly paired were 77.38% and 80.35% in ML017 and ML017/ET, respectively. Overall, each sample obtained more than 100 reads. The mean percentage of mapped reads was 82.40% in ML017 and 84.93% in ML017/ET. Samples ML017/ET-2-CTRL-B, ML017/ET-2-ET-72-B and ML017/ET-2-ET-15-B were marked as warnings due to exceeding insert size such as 3192, 2346, 1700 respectively and an unusual percentage of duplicates (less than 50%) and further downstream analysis confirmed them as outliers as reported in the PCA in Materials and methods, Section 4.8.3.2.1. Thus, they were excluded from the final cohort.

	FastQC				bcbio-nextgen			samtools						
	% Dups	% GC	Length	% Failed	% Dup	Mean IS	% rRNA	% Error rate	M Non-Primary	M Reads Mapped	% Mapped	% Proper Pairs	% MapQ 0 Reads	M Total seqs
ML017-1-CTRL-A	9.4	47	140	18	73.8	984	0	0.64	17.3	137.1	82.7	77.4	0.8	165.8
ML017-1-CTRL-B	9	46	141	9	78.2	848	0	0.6	12.4	147.9	84.6	78.9	0.6	174.8
ML017-2-CTRL-A	7.4	51	140	9	59.2	1548	0	0.63	44.2	134.5	82.1	75.4	1.5	163.8
ML017-2-CTRL-B	8.1	45	145	18	78.7	934	0	0.49	11.9	144	87.3	82.9	0.5	165.1
ML017-1-DOXO-24-A	8.4	48	140	9	77.8	833	0	0.49	26.9	189.9	85.1	81.1	0.7	223.2
ML017-1-DOXO-24-B	7.4	45	143	18	79.4	738	0	0.46	10.9	120.4	84.7	81	0.5	142.1
ML017-2-DOXO-24-A	15.1	52	144	18	54.2	1390	0	0.55	18.8	94.6	83.6	77.1	1.2	113.1
ML017-2-DOXO-24-B	9.1	44	139	9	78.2	667	0	0.52	11.2	102.8	82.1	75.5	0.6	125.2
ML017-1-ET-24-A	8.7	46	137	9	79.9	736	0	0.57	25.6	213.4	84.5	79.8	0.8	252.6
ML017-1-ET-24-B	8.9	50	132	0	67.6	1153	0	0.59	57.8	180	83.6	78.9	1.6	215.2
ML017-2-ET-24-A	9.4	50	139	9	67.4	1194	0	0.61	39.9	180	83.1	77.6	1.2	216.7
ML017-2-ET-24-B	8.4	50	139	9	62	1396	0	0.62	41.7	150.1	83.8	78.7	1.4	179.2
ML017-1-ET-72-A	9.1	48	138	9	74.1	882	0	0.58	24.1	154.7	78.3	73.3	0.9	197.5
ML017-1-ET-72-B	9.1	47	140	9	75	829	0	0.62	16.1	157.2	76.8	72.1	0.8	204.7
ML017-2-ET-72-A	26.7	55	138	0	33.3	2845	0	0.66	107.5	176.3	83.6	77.6	3.1	211
ML017-2-ET-72-B	11.1	46	140	18	75.2	880	0	0.47	13.5	125	83	78.8	0.6	150.7
ML017-1-ET-15-A	9.3	48	138	9	74.5	936	0	0.62	17.9	137.5	81.7	77.1	0.9	168.4
ML017-1-ET-15-B	8.6	48	140	9	71.7	989	0	0.66	20	150.3	77.1	72	1.1	195
ML017-2-ET-15-A	10.8	43	138	18	47.4	745	0	0.39	9.8	109.3	82.1	78.7	0.4	133
ML017-2-ET-15-B	11.6	46	135	9	48.4	849	0	0.54	13.7	96.8	78.3	73.7	0.8	123.7

Table 4.7 Quality control of RNA-sequenced samples from ML017. FastQC parameters (derived from down-sampling on chromosome 1): %Dups, percentage of duplicate reads; %GC, percentage of average content of CG bases; Length, average sequence length (bp); %Failed, percentage of modules failed in FastQC report. Bcbio-nextgen parameters (on the post-aligned whole bam files): %Dup, % duplicated reads; Mean IS, Mean insert size; %rRNA, percentage of aligned reads to ribosomal RNA. Samtools parameters: % Error rate, percentage of error rate using CIGAR; M Non-primary, Non-primary alignments (millions); M Reads Mapped, reads in the bam file (millions); % Mapped, percentage of mapped reads in the bam file; % Proper Pairs, percentage of properly paired reads; % MapQ 0 Reads, percentage reads that are ambiguously placed (MapQ=0); M Total seqs, total sequences in the bam file (millions). CTRL, control; ET, trabectedin; 24 and 72 refer to 24 and 72 hours after the first dose of drug; 15 refers to 15 days after the third dose of drug; DOXO, doxorubicin.

	FastQC				bcbio-nextgen			samtools						
	% Dups	% GC	Length	% Failed	% Dup	Mean IS	% rRNA	% Error rate	M Non-Primary	M Reads Mapped	% Mapped	% Proper Pairs	% MapQ 0 Reads	M Total seqs
ML017/ET-1-CTRL-A	9.4	48	138	9	69.4	1065	0	0.53	30.8	155.2	85.9	81.6	0.9	180.6
ML017/ET-1-CTRL-B	10	48	136	9	67.5	1078	0	0.54	32.8	150.4	84	79.7	1	179.2
ML017/ET-2-CTRL-A	13.9	48	137	18	73.5	915	0	0.62	19.5	132.4	82.3	75.9	0.9	160.9
ML017/ET-2-CTRL-B	21.2	57	132	27	20.9	3192	0.1	0.71	119.9	126.3	76.7	70.7	4.8	164.7
ML017/ET-1-DOXO-24-A	11.4	50	137	9	60.3	1381	0	0.44	39.2	135.5	83.3	79	1	162.8
ML017/ET-1-DOXO-24-B	9.6	50	135	9	60.8	1290	0	0.44	40.4	135	86.4	82.6	0.9	156.3
ML017/ET-2-DOXO-24-A	13.3	49	130	18	68.9	998	0.1	0.47	43.8	177.2	86.7	82.7	1	204.4
ML017/ET-2-DOXO-24-B	13.5	53	128	18	52.3	1618	0.1	0.51	77.6	167.7	84.6	80.2	1.6	198.3
ML017/ET-1-ET-24-A	10.7	49	143	9	68.8	1316	0	0.59	19.5	84.2	86.8	82.1	1.1	97
ML017/ET-1-ET-24-B	10.6	49	134	9	68.2	1157	0	0.53	46.8	188.1	85.9	81.2	1.1	218.9
ML017/ET-2-ET-24-A	17.5	51	133	18	62.6	1432	0.1	0.59	53.4	155.1	85.6	80.4	1.5	181.2
ML017/ET-2-ET-24-B	13.1	51	126	18	60.7	1273	0.2	0.58	74.6	181.3	86.4	82.1	1.6	209.9
ML017/ET-1-ET-72-A	11.1	51	136	9	62.4	1453	0	0.44	45.4	158.3	85.8	80.9	1	184.5
ML017/ET-1-ET-72-B	8.7	50	138	9	63.9	1263	0	0.43	45.4	163.8	89.6	86.4	0.9	182.8
ML017/ET-2-ET-72-A	12.7	50	135	18	69.2	1119	0	0.61	35.4	152.2	84.6	79.6	1.2	179.9
ML017/ET-2-ET-72-B	21.2	54	118	18	39.6	2346	0.2	0.53	118.9	166.5	85.8	80.6	2.7	194.1
ML017/ET-1-ET-15-A	10.7	49	133	9	69.6	1078	0	0.41	35.6	144.3	86.7	82.7	0.8	166.4
ML017/ET-1-ET-15-B	13.2	50	137	9	66.1	1220	0	0.45	20.3	90.5	86.9	82.3	0.8	104.1
ML017/ET-2-ET-15-A	24.2	53	125	9	50.8	1676	0.2	0.47	69.9	145.8	86.2	81.4	1.6	169.1
ML017/ET-2-ET-15-B	15.5	54	127	18	44.8	1700	0.1	0.51	68.7	150	78.4	74.8	1.6	191.4

Table 4.8 Quality control of RNA-sequenced samples from ML017/ET. FastQC parameters (derived from down-sampling on chromosome 1): %Dups, percentage of duplicate reads; %GC, percentage of average content of CG bases; Length, average sequence length (bp); %Failed, percentage of modules failed in FastQC report. Bcbio-nextgen parameters (on the post-aligned whole bam files): %Dup, % duplicated reads; Mean IS, Mean insert size; %rRNA, percentage of aligned reads to ribosomal RNA. Samtools parameters: % Error rate, percentage of error rate using CIGAR; M Non-primary, Non-primary alignments (millions); M Reads Mapped, reads in the bam file (millions); % Mapped, percentage of mapped reads in the bam file; % Proper Pairs, percentage of properly paired reads; % MapQ 0 Reads, percentage reads that are ambiguously placed (MapQ=0); M Total seqs, total sequences in the bam file (millions). CTRL, control; ET, trabectedin; 24 and 72 refer to 24 and 72 hours after the first dose of drug; 15 refers to 15 days after the third dose of drug; DOXO, doxorubicin.

4.8.2.3 Quality control of ChIP sequencing data

ChIP-Seq was performed for the ML017 model only due to technical issues related to the recruitment of biological samples and DNA processing. Immunoprecipitation involved:

- the *DDIT3* transcription factor at basal conditions or immunoprecipitated after treatment with trabectedin at the same time points as reported previously*, e.g.* 24 and 72 hours after the first dose and 15 days after the third dose. For each condition, at least three biological replicates were derived.
- the tri-methylation at the 4^{th} lysine residue of the histone H3 protein (H3K4me3) that is recognised as a marker of activated gene expression. Due to technical reasons, such as unavailability of samples or poor quality in DNA immunoprecipitation, only basal condition and treatment with trabectedin at 24 hours were available. However, biological replicates were available only for trabectedin-treated samples.
- Input, such as fragmented and cross-linked DNA under the same conditions as the immunoprecipitated matched-samples (Landt et al., 2012), that was used as control for peak calling. In order to ensure an homogeneous coverage and accurate peak calling, input samples were sequenced expecting at least 100 reads (see Materials and methods Section 4.8.1).

All samples had the required sample fragment (75 bp). The fraction of reads occurring in enriched regions (reads in peaks, RiP) were assessed expecting more than 1% for transcription factor immunoprecipitation (Landt et al., 2012) and more than 10% for histones. For each sample, the ENCODE quality parameters were assessed, such as:

- PCR bottleneck coefficient 1 (PBC1) and PCR bottleneck coefficient 2 (PBC2) that are a measure of the library complexity defined as in *Equation 4.3* and *Equation 4.4*;
- The non-redundant fraction (NRF) that is the number of distinct uniquely mapping reads over the total number of reads;

- Both PBC and NRF correspond to complexity levels defined as in *Table 4.9* in accordance to the ENCODE definitions ("Terms and Definitions – ENCODE," n.d.).

$$PBC1 = {}^{N1}/_{Nd}$$

Equation 4.3 PCR bottleneck coefficient 1. N1, number of genomic locations to which there is one exact unique mapping; Nd, the number of genomic locations to which there is at least one unique mapping.

$$PBC2 = {}^{N1}/_{N2}$$

Equation 4.4 PCR bottleneck coefficient 2. N1, number of genomic locations to which there is one exact unique mapping; Nd, the number of genomic locations to where two reads map uniquely.

PBC1	PBC2	Bottlenecking level	NRF	Complexity
< 0.5	< 1	Severe	< 0.5	Concerning
0.5 ≤ PBC1 < 0.8	1 ≤ PBC2 < 3	Moderate	0.5 ≤ NRF < 0.8	Acceptable
0.8 ≤ PBC1 < 0.9	3 ≤ PBC2 < 10	Mild	0.8 ≤ NRF < 0.9	Compliant
≥ 0.9	≥ 10	None	> 0.9	Ideal

Table 4.9 Bottlenecking level and complexity associated with PCR bottlenecking coefficients (PBC) and Non Redundant Fraction (NRF) as reported by ENCODE terms and definitions.

	FastQC				bcbio-nextgen (ENCODE)						samtools				
	% Dups	% GC	Length bp	% Failed	RiP	PBC1	PBC2	Bottlenecking level	NRF	ENCODE Complexity	% Error rate	M Reads Mapped	% Mapped	% MapQ 0 Reads	M Total seqs
ML017_DDIT3_I	5	41	75	9	2.99	0.6	2.41	severe	0.59	concerning	1.14	43.3	68.8	2.1	62.8
ML017_DDIT3_II	4.8	41	75	9	41.7	0.67	3.04	severe	0.64	concerning	0.65	78.1	93.7	0.9	83.4
ML017_DDIT3_III	9.1	43	75	18	5.21	0.59	2.33	severe	0.6	concerning	0.57	31.4	76.9	1.5	40.9
ML017_DDIT3_ET24_I	27.4	44	75	18	3.11	0.23	1.7	severe	0.23	concerning	0.61	34.8	78.7	1.9	44.3
ML017_DDIT3_ET24_II	2	41	75	9	0.67	0.81	5.2	moderate	0.78	acceptable	0.86	91.6	86.1	2.3	106.3
ML017_DDIT3_ET24_III	3.6	41	75	9	26.69	0.76	4.19	moderate	0.73	acceptable	0.73	75.3	91.1	1.3	82.7
ML017_DDIT3_ET72_I	8.7	40	75	9	1.74	0.81	5.16	moderate	0.8	acceptable	0.62	34	85.1	2	39.9
ML017_DDIT3_ET72_II	14.9	42	75	18	12.39	0.71	3.36	moderate	0.71	acceptable	0.51	32.4	78.6	1.2	41.2
ML017_DDIT3_ET72_III	15	43	75	9	19.73	0.77	4.33	moderate	0.76	acceptable	0.49	34.1	81.1	1	42
ML017_DDIT3_ET15_I	7.2	41	75	9	0.53	0.51	1.88	severe	0.54	concerning	0.71	50.7	68.1	1.8	74.4
ML017_DDIT3_ET15_II	14.9	41	75	9	0.09	0.42	1.56	severe	0.46	concerning	0.69	25.8	65	1.7	39.6
ML017_DDIT3_ET15_III	7.8	40	75	9	0.2	0.66	2.8	severe	0.67	concerning	0.81	28.4	37.5	1.2	75.9

Table 4.10 Quality control of samples (ChIP-Seq) from ML017 related to immunoprecipitation with DDIT3. FastQC parameters (derived from down-sampling on chromosome 1): %Dups, percentage of duplicate reads; %GC, percentage of average content of CG bases; Length, average sequence length (bp); %Failed, percentage of modules failed in FastQC report. Bcbio-nextgen and ENCODE parameters (on the post-aligned whole bam files): RiP, percentage of reads in peaks; PBC1, PCR Bottlenecking coefficient 1; PBC2, PCR Bottlenecking coefficient 2; Bottlenecking level, ENCODE bottlenecking-level; NRF, non-redundant fraction; ENCODE complexity. Samtools parameters: % Error rate, percentage of error rate using CIGAR; M Reads Mapped, reads in the bam file (millions); % Mapped, percentage of mapped reads in the bam file; % MapQ 0 Reads, percentage reads that are ambiguously placed (MapQ=0); M Total seqs, total sequences in the bam file (millions).

	FastQC				bcbio-nextgen (ENCODE)						samtools				
	% Dups	% GC	Length bp	% Failed	RiP	PBC1	PBC2	Bottlenecking level	NRF	ENCODE Complexity	% Error rate	M Reads Mapped	% Mapped	% MapQ 0 Reads	M Total seqs
ML017_H3K4me3_I	8.5	51	75	18	47.56	0.42	1.71	severe	0.43	concerning	1.15	44.4	81.4	1.6	54.6
ML017_H3K4me3_II	15.5	47	75	18	20.71	0.31	1.39	severe	0.35	concerning	0.61	41.1	66.3	1.6	62
ML017_H3K4me3_ET24_I	6.2	47	75	18	26.81	0.54	2.14	severe	0.54	concerning	1.11	46	72.1	1.7	63.7
ML017_H3K4me3_ET24_II	7.2	66	75	18	55.79	0.66	2.83	severe	0.44	concerning	0.87	27.5	83.8	2.1	32.9
ML017_H3K4me3_ET24_III	6.1	53	75	9	54.51	0.65	3.2	severe	0.6	concerning	0.92	57.5	84.7	1.3	67.9
ML017_H3K4me3_ET24_IV	6.7	48	75	18	21.46	0.7	3.38	moderate	0.69	concerning	0.72	30.8	53.4	1.2	57.6
ML017_H3K4me3_ET24_V	5.8	48	75	18	28.3	0.74	4.17	moderate	0.7	acceptable	0.72	56.3	72.2	1.5	78
ML017_input	2.5	41	75	9	/	0.94	16.52	none	0.91	ideal	0.75	115.9	79	2	146.7
ML017_input_ET24	2.1	41	75	9	/	0.93	13.89	none	0.9	acceptable	0.73	127.1	83.8	2	151.7
ML017_input_ET72	16.3	40	75	9	/	0.9	10.28	moderate	0.87	acceptable	0.62	88.3	72.2	1.5	122.4
ML017_input_ET15	24.9	41	75	9	/	0.86	7.33	moderate	0.85	acceptable	0.6	80.1	69.8	1.3	114.8

Table 4.11 Quality control of samples (ChIP-Seq) from ML017 related to immunoprecipitation for histone mark H3K4me3 and input samples. FastQC parameters (derived from down-sampling on chromosome 1): %Dups, percentage of duplicate reads; %GC, percentage of average content of CG bases; Length, average sequence length (bp); %Failed, percentage of modules failed in FastQC report. Bcbio-nextgen and ENCODE parameters (on the post-aligned whole bam files): RiP, percentage of reads in peaks; PBC1, PCR Bottlenecking coefficient 1; PBC2, PCR Bottlenecking coefficient 2; Bottlenecking level, ENCODE bottlenecking-level; NRF, non-redundant fraction; ENCODE complexity. Samtools parameters: % Error rate, percentage of error rate using CIGAR; M Reads Mapped, reads in the bam file (millions); % Mapped, percentage of mapped reads in the bam file; % MapQ 0 Reads, percentage reads that are ambiguously placed (MapQ=0); M Total seqs, total sequences in the bam file (millions)

4.8.3 *PIPELINES OF ANALYSIS*

The first processing of data analysis of this work was done with bcbio-nextgen ("Contents — bcbio-nextgen 1.2.4 documentation," n.d.) a publicly available python toolkit that provides self-customised pipelines for high-throughput data analysis. bcbio-nextgen was installed and run on the HPC cluster in the Department of Oncology using the slurm workload manager ("Slurm Workload Manager - Documentation," n.d.). The bcbio-nextgen version used for DNA and RNA sequencing was *version 1.1.1*, while *version 1.2.3* was used for ChIP-Seq analysis. This is due to the updates made to pipelines starting from the first year of this PhD work till the last.

According to bcbio-nextgen documentation, pipelines were configured through a comma separated value (*.csv)* file, a *yaml* file and then run with a *batch* file.

The *.csv* is composed by a the samples to be analysed and metadata associated with them. An example is reported in *Table 4.12*.

sample name	description	batch	condition
Sample1	ML017_CTRL_I	ML017	CTRL
Sample2	ML017_ET-24_I	ML017	ET-24
Sample3	ML017ET_CTRL_I	ML017-ET	CTRL
Sample4	ML017ET_ET-24_I	ML017-ET	ET-24

Table 4.12 Example of a .csv file for bcbio-nextgen analysis.

A *yaml* file is a configuration file which stores the required software for the analysis. The "details" section requires the indication of the type of analysis, *e.g.* RNA-Seq, and the genome build, *e.g.* hg19. The sub-section "algorithm" is assigned to the software to be used. Finally, the "upload" section requires the path and the name of the folder to save data. The configuration files of the analysis used in this PhD thesis are available at[2] with the name **-config.yaml*.

A *batch* file is required to run the analysis of the HPC cluster. The directory in which the analysis will be run is indicated together with the configuration file. The number

[2] Figshare: https://doi.org/10.6084/m9.figshare.17181104.v1,
Zenodo: https://doi.org/10.5281/zenodo.5807695

of cores is set and *bcbio_nextgen.py* is called as reported in the *batch* file available at[2] as *analysis.batch*.

bcbio_nextgen creates three folders:

1. *Config*: where configuration files are stored.
2. *Work*: where the analysis is performed and temporary files are stored. This folder is usually removed when analysis is completed.
3. *Final*: where all processed files have been saved.

4.8.3.1 Pipeline for DNA-Seq analysis

4.8.3.1.1 OneSeq hybrid solution

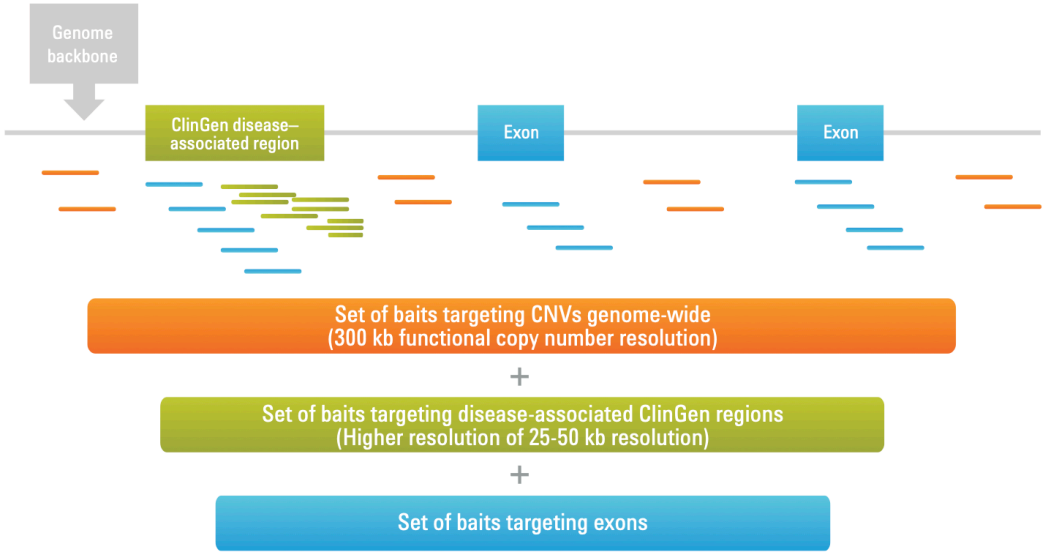

Figure 4.4 OneSeq panel design as reported by Agilent Technologies

DNA-Sequencing (DNA-Seq) has been one the first fully exploited high-throughput sequencing approach. It can be divided into three macro-categories: whole genome sequencing (WGS) that allows the study of the entire genome; whole exome sequencing (WES) that is focused on protein-coding genes only; and targeted-resequencing that consists in a selection of a panel of coding genes, usually in the range of hundred. DNA-Seq allows the identification of genomic events like single nucleotide variants (SNVs), small insertions or deletions, gene-fusions and somatic copy number alterations (SCNAs). In this PhD work, a hybrid DNA-Seq solution was

used, called OneSeq (*Figure 4.4*). The OneSeq panel was developed by Agilent Technologies and allows to simultaneously detect both SNVs and SCNAs in a single experiment. Through this approach it is possible to sequence many samples at lower costs than for WGS or WES, still obtaining broad and informative results on their genomic sequence. It is composed of 5971 genes associated with diseases and regions that are spotted all over the genome that are addressed as “genomic backbone”.

4.8.3.1.2 DNA-Seq data analysis

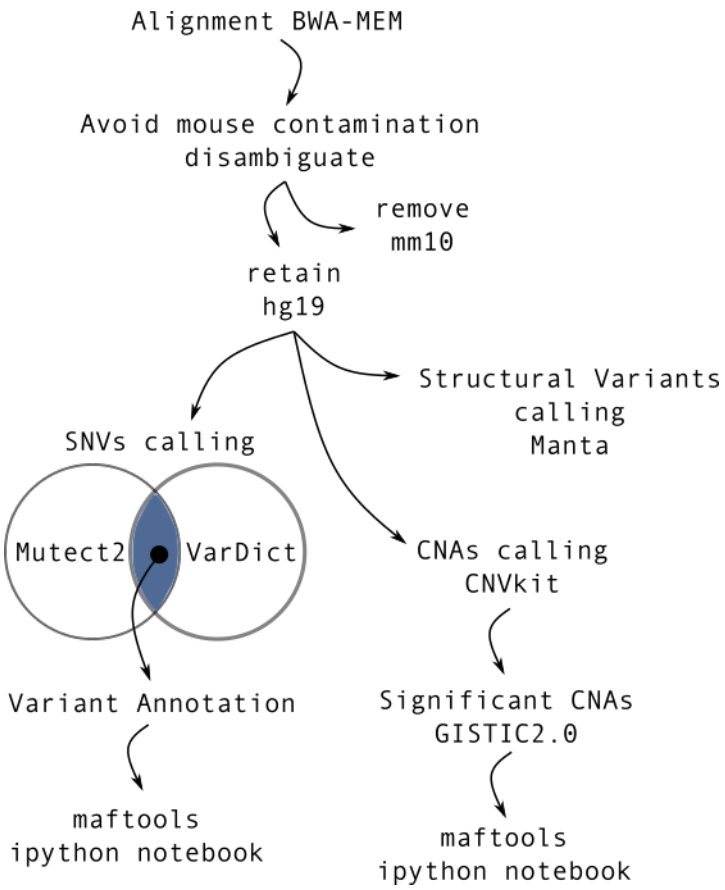

Figure 4.5 Workflow for DNA-Seq analysis.

bcbio-nextgen pipeline was customised in order to follow the workflow reported in *Figure 4.5*.

The DNA extracted from PDX may be contaminated by a genome derived from the mouse. Thus, In order to avoid possible contamination from mouse genome, the *disambiguate* software (Ahdesmäki et al., 2016) was used. Raw reads were aligned on both human *hg19* and *mm10* mouse genomes: reads certainly mapped on this last were excluded from further analysis, while those unambiguously mapped on *hg19* were retained (*Figure 4.6*). Raw reads (*fastq*) were aligned with the BWA-MEM aligner *version 0.7.17* (Li and Durbin, 2009).

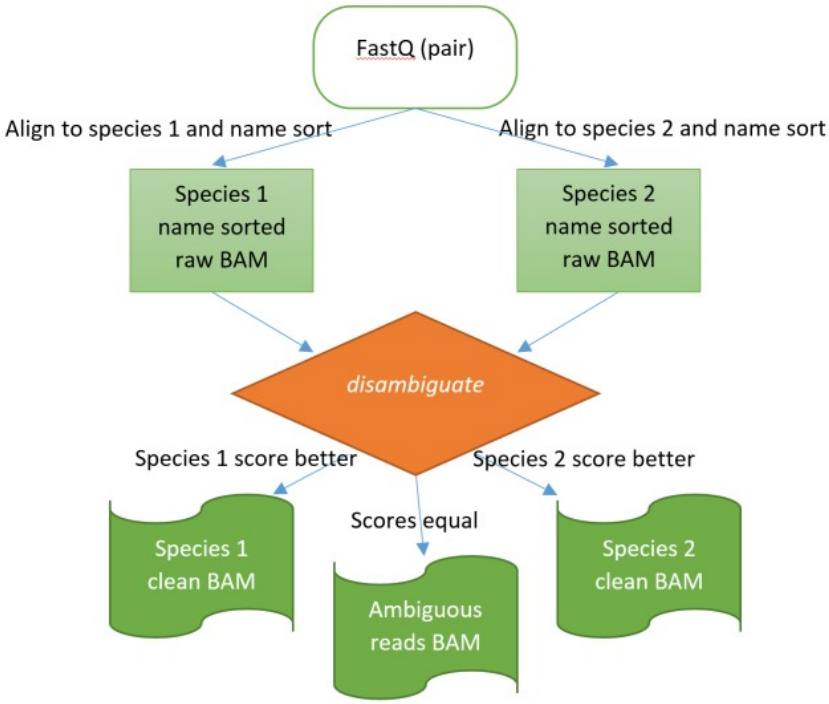

Figure 4.6 The disambiguation process as reported in (Ahdesmäki et al., 2016).

Single nucleotide variant (SNV), insertions and deletions calling was done with two variant callers, such as Mutect2 (Cibulskis et al., 2013) and VarDict (Lai et al., 2016), independently. Variant callers are algorithms developed with the aim to identify SNVs or *indels*. It has been reported a low concordance in variant calling from different callers. This is mainly associated with the algorithm itself and not to bias (O'Rawe et al., 2013). Indeed, the use of different variant callers is recommended. In this work, variants called by both Mutect2 and VarDict were retained. For variant calling detection a *bed* file composed by regions from the Agilent Focused Exome was used. For each variant a threshold of 10X of coverage was set. Variants were annotated with the Variant Effect Predictor (VEP) (McLaren et al., 2016) and variant

impact prediction we done with PolyPhen (Ramensky et al., 2002), SIFT (Ng and Henikoff, 2003) and CADD (Rentzsch et al., 2019). *vcf* (Variant Call Format) files were stored in a GEMINI database (Paila et al., 2013) that was then converted into a MAF (Mutation Annotation Format).

In parallel, in order to identify structural variants like breakends leading to gene fusions, Manta software (X. Chen et al., 2016) was used, while CNVkit (Talevich et al., 2016) *version 0.9.4* was used for SCNAs identification. For both Manta and CNVkit we used a *bed* file of regions derived from merging regions from the Agilent Focused Exome and the regions from the Agilent Backbone.

GISTIC2.0 is a software used for the identification of arm-level or focal SCNAs (Mermel et al., 2011). It is usually applied to large cohorts in order to identify the most frequent SCNAs. In this work it was used to define frequent altered regions in the whole cohort of ML017 and ML017/ET samples. In order to define specific altered regions and avoid bias, significant SCNAs were retained only when called in all four replicates for each condition.

Downstream analysis and graphical representation of SNVs, insertions and deletions was done with Maftools (Mayakonda et al., 2018) using the followings functions:

- *plotmafsummary* to plot the number of variants per sample;
- *oncoplot* to show frequently mutated genes;
- *tcgaCompare* to compare PDX models to cohorts from The Cancer Genome Atlas (TCGA);
- *coOncoplot* and *forestPlot* for identification and representation of differentially mutated genes (DMGs). Both functions help in the visualisation of the results that come from the *mafCompare* function of the same package. This performs a Fisher test between two selected cohorts and identifies DMGs with their associated p-value.

Chromosomal breakends and gene fusions were visualised and represented with Integrative Genomics Viewer (IGV) (Robinson et al., 2017).

Identification of GISTIC2.0 significant regions in all replicates per each condition was done through the IPython notebook (Perez and Granger, 2007) with a combination of the *pandas* (Jeff Reback et al., 2020) and the *numpy* (Harris et al., 2020) packages. Images were created using matplotlib (Hunter, 2007) and seaborn (Michael Waskom et al., 2020) packages, while final styling was done with inkscape ("Home | Inkscape," n.d.).

4.8.3.2 Pipeline for RNA-Seq data analysis

4.8.3.2.1 Gene expression counts

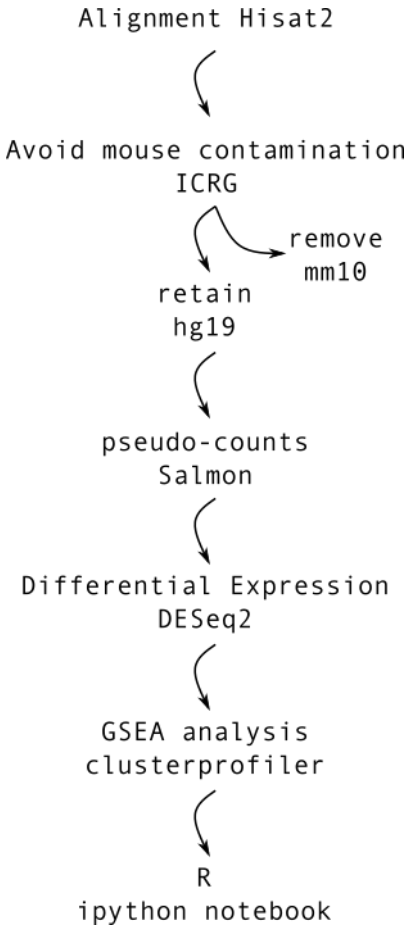

Figure 4.7 Schematic workflow of the gene expression pipeline of analysis.

Figure 4.7 shows the workflow used for RNA-Seq data processing. Again, to avoid mouse genome contamination a customised alignment was performed. In this case, this first step was done outside the bcbio-nextgen pipeline, using the In silico Combined human-mouse Reference Genome (ICRG) (Callari et al., 2018) workflow. The reason to use ICRG instead of disambiguate (Ahdesmäki et al., 2016) as it has been done for DNA-Seq (see Materials and methods, Section 4.8.3.1.2) is due to

the novelty of ICRG that was demonstrated to perform better than disambiguate for RNA-Seq data (Callari et al., 2018). According to ICRG, human *hg19* and mouse *mm10* genomes were merged as described in ("cclab-brca/ICRG," n.d.) and a new index for Hisat2 (Kim et al., 2019) aligner was created. Raw reads were mapped to the new genome with Hisat2 and mouse-aligned sequences were discarded.

Human aligned reads were used for computation of pseudo-counts with the wicked-fast inference algorithm of Salmon (Patro et al., 2017) *version 0.13.1* starting from pre-computed alignments. Salmon was chosen for because it exceeds the accuracy of other methods for the quantification of transcript abundance and also it is even faster. Moreover, It supports the analysis of reads from strand-specific protocols and performs well with paired-end reads (Patro et al., 2017). The choice comes also from the "Advance RNA-Seq and ChIP-Seq data analysis" course attended by the PhD candidate at the EMBL, Hinxton, UK in 2017 where Salmon pseudo-counts algorithm was presented and assessed through practical examples as the most performing in comparison to the others. Counts were imported and read with *tximport* (Soneson et al., 2015) and data analysed with DESeq2 (Love et al., 2014).

Samples ML017/ET-2-CTRL-B, ML017/ET-2-ET-72-B and ML017/ET-2-ET-15-B that were previously marked as warnings (see Materials and methods, Section 4.8.2.2) were confirmed as outliers as reported in *Figure 4.8* where they are indicated with a red star.

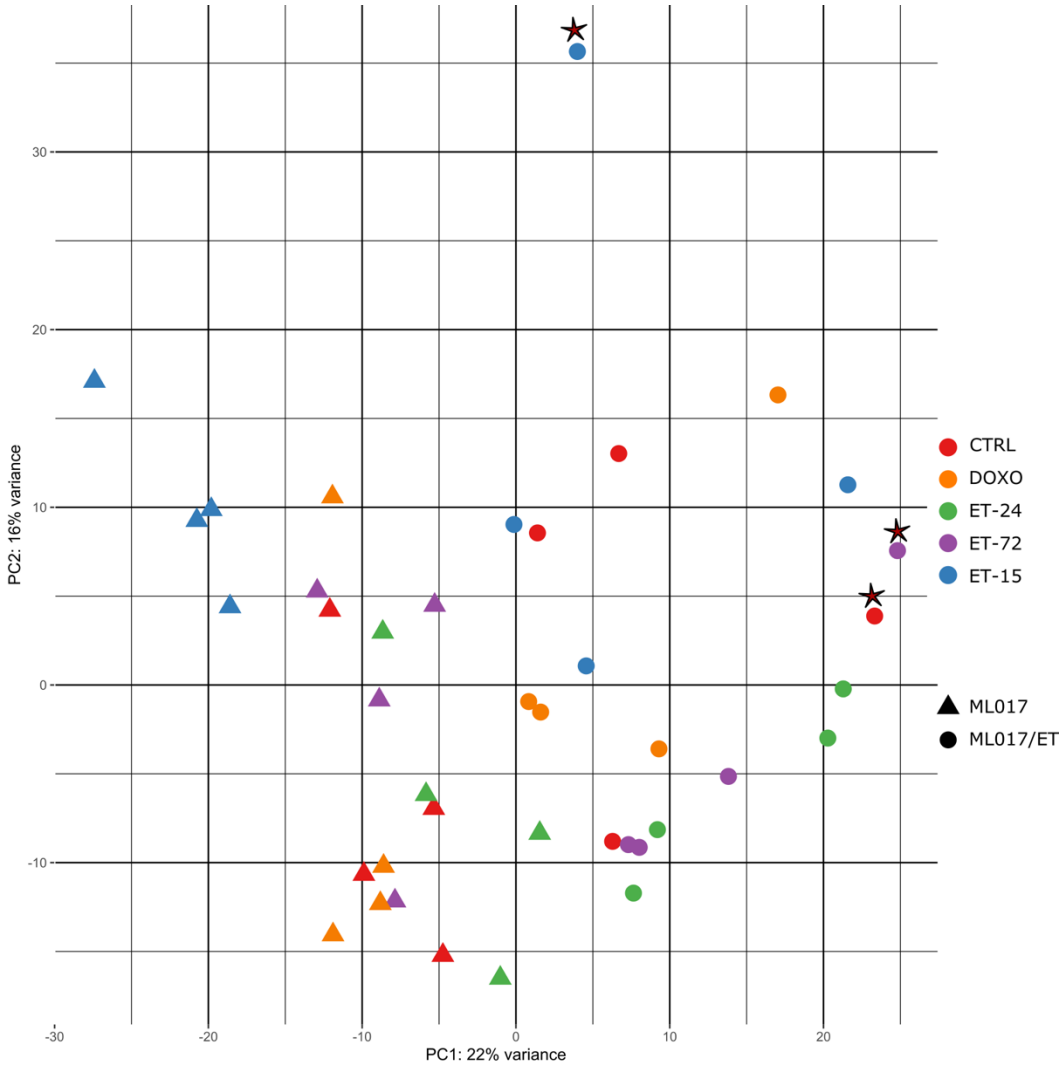

Figure 4.8 PCA of the whole samples. ML017 are depicted as triangles, ML017/ET as circles. Colours indicate conditions as in the legend. PC, principal component.

ML017 and ML017/ET samples were normalised together. There are other packages that allows the processing of RNA-Seq data and differential expression analysis, such as edgeR (Y. Chen et al., 2016) and limma-voom (Law et al., 2014). The choice of using DESeq2 in this work was mainly due to a highly curated and updated documentation associated with this package. Moreover, the bioinformatic community around DESeq2 is very active, thus a wide range of case studies is covered so that it is useful to find examples that can contribute to apply the package properly. Then, DESeq2 is also used in the background of the package DiffBind (Stark and Brown, 2011) that was applied to ChIP-Seq analysis as further explained (see Materials and methods, Section 4.8.3.3), thus the choice was to maintain a coherence in the methods used.

Differential expression analysis was done contrasting each treatment condition to control samples in ML017 and ML017/ET models, independently or contrasting the basal conditions of ML017/ET versus ML017. The test used for differential expression analysis was the Wald test (Love et al., 2014). The whole process consists in three subsequent steps that are made by the DESeq2 functions listed as follows:

1. dds <- estimateSizeFactors(dds): estimation of the size factors that controls for differences in the depth of sequencing;
2. dds <- estimateDispersions(dds): estimation of the dispersion value for each gene;
3. dds <- nbinomWaldTest(dds): fitting to the model;

where the *dds* is the object created at the very beginning of the analysis which contains the matrix with gene-sample counts.

A list of differentially expressed genes (DEGs) was identified with an associated *p-value*. Correction for multiple testing was applied using the False Discovery Rate (FDR) ("Controlling the False Discovery Rate," n.d.). DEGs were retained with a p-adjust of 0.05.

Pathway analysis was done using the clusterProfiler (Yu et al., 2012) package performing the Gene Set Enrichment Analysis (GSEA, (Subramanian et al., 2005)) with the Reactome database (Jassal et al., 2020). Only selected DEGs with a p-adjust of 0.05 were used for the GSEA and they were sorted according to their log2 Fold Change, from the most up-regulated to the most down-regulated. An enrichment score is associated with each pathway representing the degree at which a gene-set is over-represented in the pathway. Enrichment scores are transformed into normalised enrichment score (NES) in order to make results comparable. The NES takes into account the differences in gene set size and in correlations between gene sets and the expression datasets. Finally, p-values associated with each pathway were corrected for multiple testing using the FDR correction ("Controlling the False Discovery Rate," n.d.).

Data visualisation was done either with embedded functions of DESeq2 or with the matplotlib (Hunter, 2007) and the seaborn (Michael Waskom et al., 2020) packages in the IPython notebook (Perez and Granger, 2007), while final styling was done with inkscape ("Home | Inkscape," n.d.).

4.8.3.2.2 Transcripts reconstruction

For transcript-level analysis and identification of novel isoforms, an adapted version of the Tuxedo protocol presented in Pertea et al. (Pertea et al., 2016) was used (*Figure 4.9*). Specifically, raw sequencing reads were aligned with the Hisat2 aligner (Kim et al., 2019), then StringTie (Pertea et al., 2015) was used for transcripts assembly and quantification with *hg19* annotation file. Finally, *gffcompare* (Pertea et al., 2016) allowed to determine the number of transcripts matching the annotation. Next, the *merge* function of StringTie was used to merge transcripts from all samples in a consistent set.

Then, starting from aligned files and previously identified merged transcripts, transcripts quantification was done through the *-e -B* mode of StringTie. The table of transcripts counts was created with *prepDE.py* ("prepDE," n.d.) a Python script provided by the StringTie documentation.

The transcripts count matrix was loaded and analysed with the DESeq2 package (Love et al., 2014). Differential expression analysis was done contrasting each treatment condition with controls in each model, independently. Differentially expressed transcripts (DETs) were filtered with a p-adjust less than 0.05. Transcripts were annotated accordingly to the StringTie annotation file. Then, DETs were divided into already known and unknown transcripts. These last were categorised according to StringTie annotation into probably new isoforms of known transcripts and novel transcripts. Finally, new isoforms of known transcripts were analysed with the Coding-Potential Assessment Tool (CPAT) (Wang et al., 2013) to predict their coding potential.

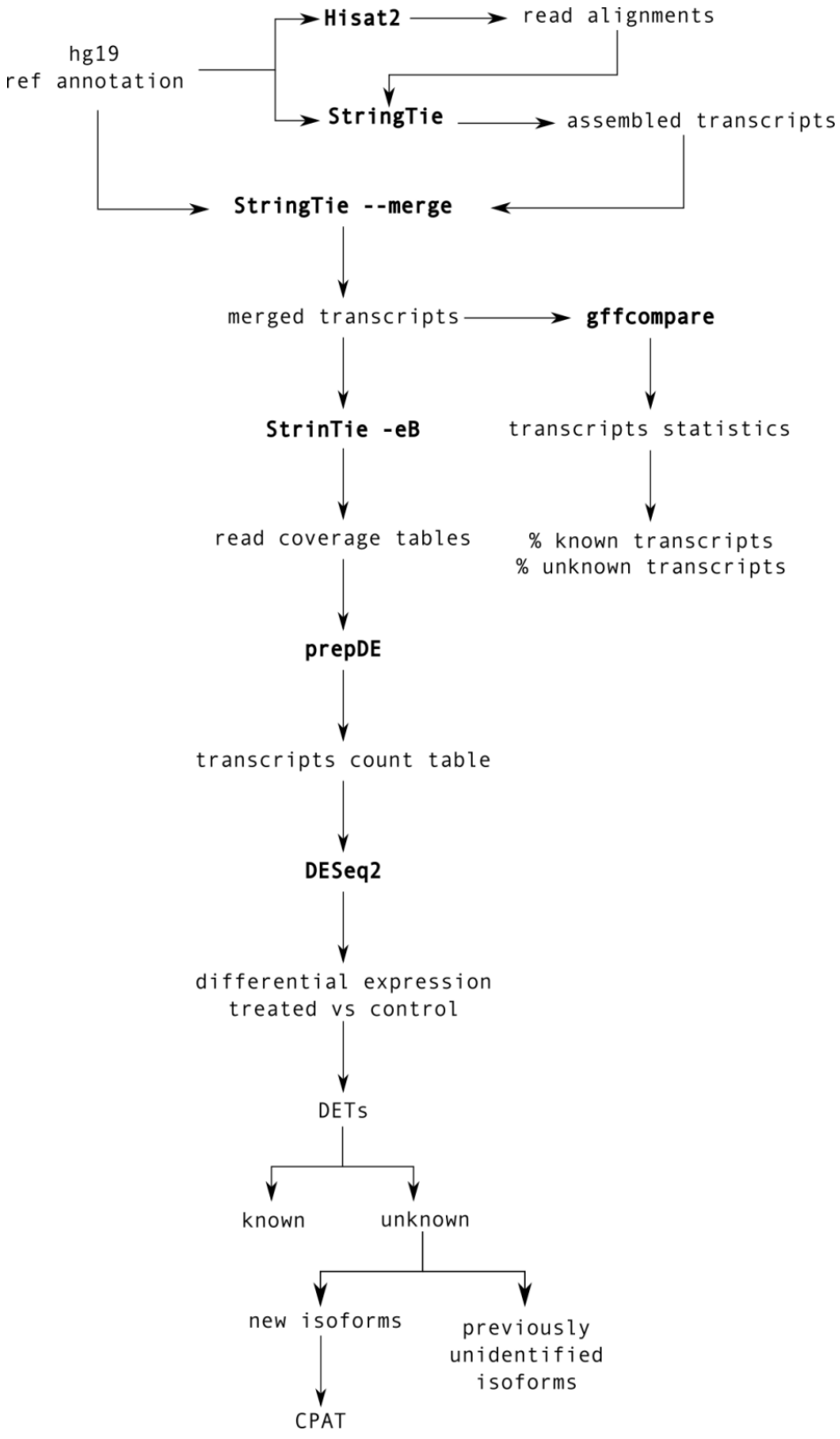

Figure 4.9 workflow used for transcripts reconstruction and analysis.

4.8.3.3 Pipeline for ChIP-Seq data analysis

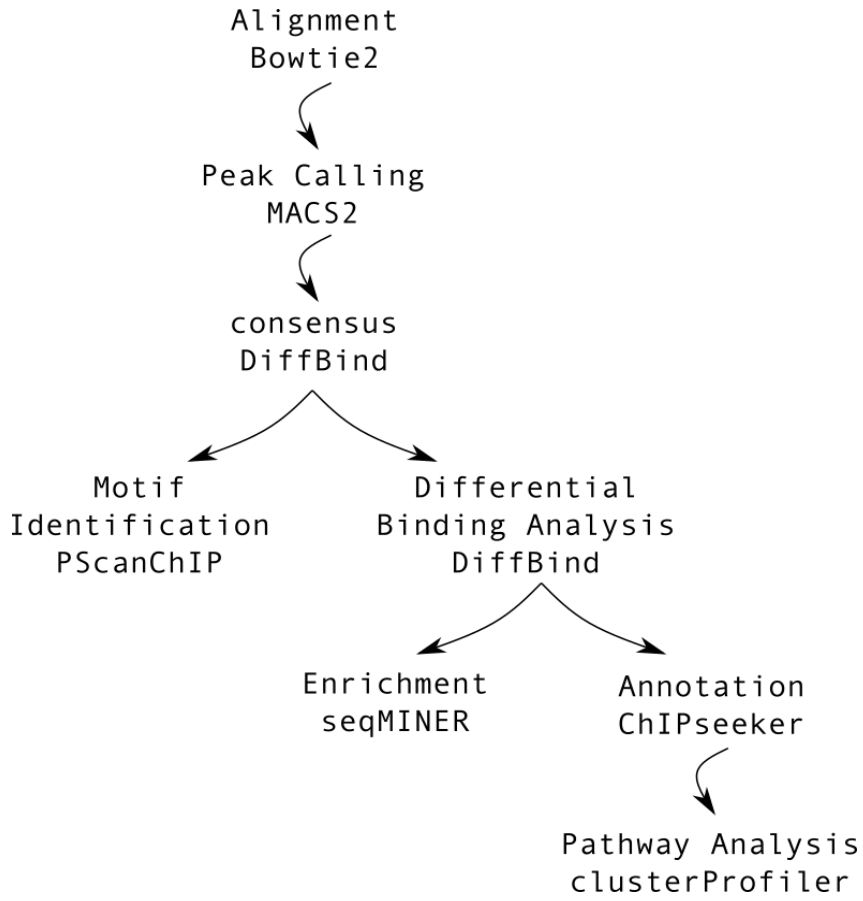

Figure 4.10 Workflow for ChIP-Seq data analysis.

A schematic workflow of ChIP-Seq analysis is reported in *Figure 4.10*. In detail, raw data sequences were aligned with Bowtie2 *version 2.3.5.1* (Langmead and Salzberg, 2012). Both DDIT3 transcription factor and H3K4me3 histone mark were considered as narrow peak binding factors. Peak calling, that is the identification of transcription factor or histone mark binding sites on the DNA, needs to be associated with an appropriate control sample. The ENCODE guidelines suggest one from the *input* or the *mock* sample. The first is DNA isolated from the same origin of the matched transcription factor and fragmented under the same conditions; the latter is derived from a control antibody that reacts with non-nuclear antigen IgG (Landt et al., 2012). In this work we used one *input* sample for each condition. Peaks were called with MACS2 (Zhang et al., 2008) with a q-value of 0.05.

The identification of the most represented binding sequence, e.g. *motif*, was done with PscanChIP (Zambelli et al., 2013) querying the Jaspar database (Fornes et al., 2020).

DiffBind (Ross-Innes et al., 2012), (Stark and Brown, 2011) was used to compute consensus peaks at basal conditions and to identify differentially bound peaks (DBPs), that are peaks with significant changes between conditions. DiffBind defines a "change" based on the differential binding affinity that is measured with the read density. Then, differential analysis is done with either DESeq2 (Love et al., 2014) or edgeR (McCarthy et al., 2012). In this work DBPs were called with DESeq2 in line with the methods used for RNA-Seq analysis (see Materials and methods, Section 4.8.3.2).

The binding of a transcription factor on the DNA can have functional effects on the gene expression levels of neighbouring genes (Jiang and Mortazavi, 2018). The modulation of the gene expression depends also on the location of the binding on the genome. In order to associate peaks to regions of interest, we used the ChIPSeeker package (Yu et al., 2015). We defined promoter regions as 5 kb upstream and 1 kb downstream the transcription start site (TSS), while distal regions were those beyond the TSS till 1000 kb (*Figure 4.11*).

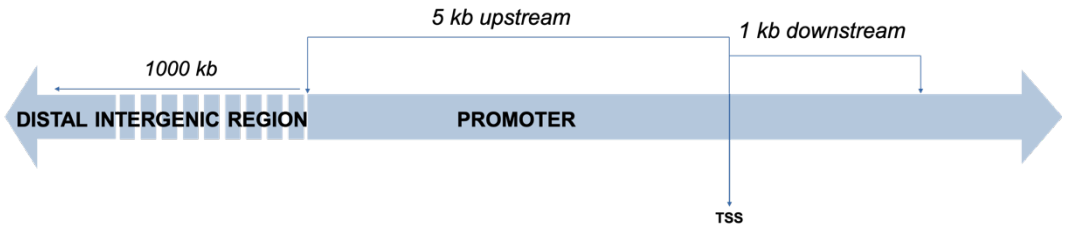

Figure 4.11 Annotation rules for the binding sites used with ChIPSeeker (Yu et al., 2015). TSS, transcription start site.

Pathway analysis was done using clusterProfiler package (Yu et al., 2012) with both the Reactome (Jassal et al., 2020) and the Wikipathways (Martens et al., 2021) databases. To this aim, we used the *enrichPathway* and the *enrichWP* functions of clusterProfiler for the Reactome and the Wikipathways databases, respectively.

Both functions perform an enrichment analysis of a set of genes. The enrichment is evaluated by the hypergeometric test defined as in *Equation **4.5***.

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

Equation 4.5. Hypergeometric test equation. N, total number of genes in the background; M, number of genes of the background that are annotated to a pathway of interest; n, size of the list of genes of interest; k, the number of genes within that list annotated to the pathway of interest.

In the *Equation 4.5*, N represents the total number of genes in the background that by default is all the genes that have annotation; M is the number of genes of the background that are annotated to a pathway, n is the size of the list of genes of interest and k is the number of genes within that list which are annotated to that pathway. For each pathway a p-value is assigned and corrected for multiple testing with the FDR correction ("Controlling the False Discovery Rate," n.d.) setting a threshold of 0.05. In the figures reported in the Result, Section 5.4, the "gene ratio" is defined as the ratio between k and n.

5 RESULTS

5.1 STUDY WORKFLOW

To achieve the aims of this PhD thesis (see Aims of the study, Chapter 3), the following *in vivo* models were used (see Background Section 2.2.1):

1. ML017, a FUS-DDIT3 chimera type I bearing model, responsive to trabectedin;
2. ML017/ET which was developed from ML017 and made resistant against trabectedin treatment;
3. ML015, a FUS-DDIT3 chimera type 2 bearing model;
4. ML006, a FUS-DDIT3 chimera type 3 bearing model.

For each model genomic and transcriptomic signatures were generated and further integrated in order to:

1. characterise PDX models and evaluate the extent to which they can reproduce MLPS characteristics;
2. study the mechanism of action of trabectedin in MLPS;
3. elucidate the mechanisms leading to acquired drug-resistance in MLPS.

Figure 5.1 depicts the experimental design that was applied to the study (see also Materials and methods, Section 4.7). ML017, ML015 and ML006 modelled the first responsiveness to trabectedin treatment that is achieved in the clinic, while ML017/ET model mimics the clinical issue of acquired resistance against trabectedin treatment.

To answer all these points in a comprehensive and exhaustive manner, we guided our analysis through different levels, such as:

1. the genomics level to find somatic variants like single nucleotide variations and copy number alterations that could be attributed either to drug response or drug resistance;

2. the transcriptomics level, to follow the transcriptional changes induced by drug treatment;
3. the epigenomics level to define FUS-DDIT3/DNA binding and how they change under drug exposure;
4. the integration of the previously cited levels to describe the pharmacogenomics of trabectedin.

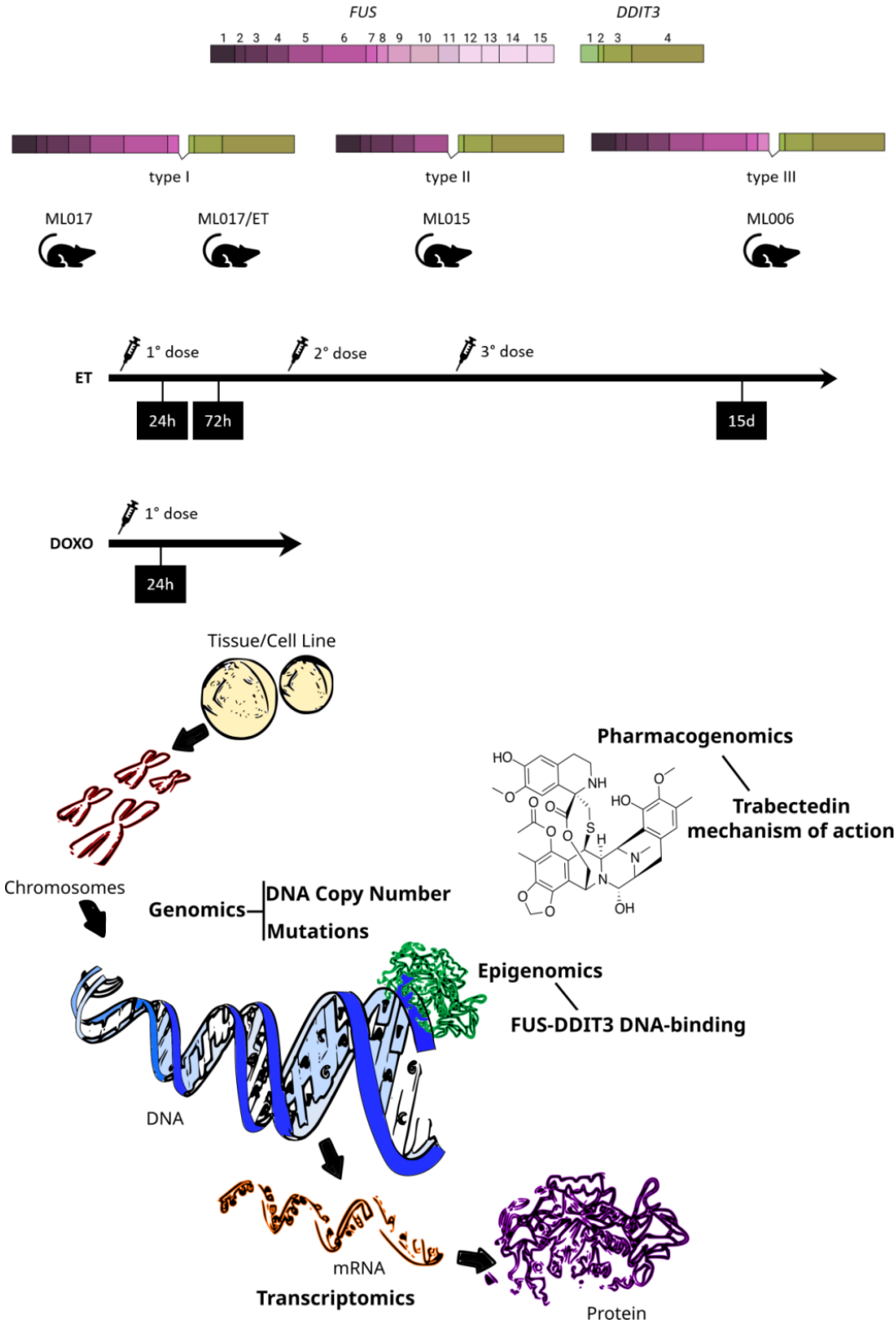

Figure 5.1 Samples, experimental design and schematic workflow of this work. ET, trabectedin; DOXO, doxorubicin.

5.2 THE GENOMIC LANDSCAPE OF MLPS AND THE RESPONSE TO TRABECTEDIN TREATMENT

As a first step in achieving the aims of the PhD program, libraries were generated from genomic DNA purified from PDX models at the baseline and at different times of drug exposure to:

1. characterise the genomic landscape of each tumour model and evaluate the extent to which they mirror the genome of MLPS tumours and the different clinical sensitivity of type ML017, ML015 and ML006 (see Background Section 2.2.1) to trabectedin treatment;
2. determine the genomic impairments that drive the acquisition of resistance against trabectedin in the ML017/ET model;
3. study the effects of trabectedin exposure on the genomics architecture in both responsive and resistant models, in comparison to the standard treatment with doxorubicin.

To pursue these goals, we generated a somatic genomic database by next generation sequencing (NGS) experiments for somatic variant calling, such as single nucleotide variants (SNVs), small insertions and small deletions, as well as somatic copy number alterations (SCNAs).

For both analyses, ML017, ML017/ET and ML015 models were matched to their normal counterpart obtained from the same patient the PDX derived from; differently, since the normal tissue was not available for ML006 model, the analysis was performed against a commercial reference genome (see Materials and methods, Section 4.8.3.1).

The whole cohort of sequencing experiments passed the quality assessment, therefore we did not exclude any sample.

5.2.1 IDENTIFICATION OF FUS-DDIT3 IN PDX MODELS

First, we queried the database to assess the presence of FUS-DDIT3 gene fusion in each model (see Background Section 2.1.2). We identified the characteristic FUS-

DDIT3 gene fusion in ML017, ML017/ET and ML006 models, however it was not detected in ML015, thus this model could not be considered as a reliable model of MLPS and we excluded it from further analysis.

In agreement with pathological analysis the type I isoform of FUS-DDIT3 chimera was correctly called in ML017 and ML017/ET models, while the type III chimera was called in ML006 model (see Background, Section 2.2.1). In all models, breakpoints corresponded to intronic regions as reported in *Table 5.1* and in *Figure 5.2* for ML017 and ML017/ET models.

	Chrom	Breakend	Ref	Alt	Translocation	Variant
ML017 and ML017/ET	chr12	57914179	G	G]chr16:31198193]	FUS-DDIT3	gene fusion and frameshift variant
	chr16	31198192	C	C]chr12:57914180]		
	chr12	57914182	G	[chr16:31198193[G	DDIT3-FUS	gene fusion and frameshift variant
	chr16	31198192	C	[chr12:57914183[C		
ML006	chr12	57912006	C	C]CHR16:31199785]	FUS-DDIT3	gene fusion and frameshift variant
	chr16	31199783	T	T]CHR12:57912008]		
	chr12	57912024	T	[CHR16:31199532[T	DDIT3-FUS	gene fusion and frameshift variant
	chr16	31199530	T	[CHR12:57912026[T		

Table 5.1 FUS-DDIT3 and DDIT3-FUS gene fusions annotations. Chrom, chromosome; Ref, reference base; Alt, alternate bases.

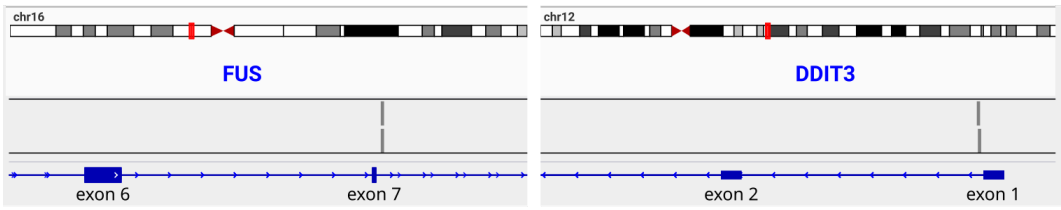

Figure 5.2 Type I chimera FUS-DDIT3 in ML017 and ML017/ET models represented with the Integrative Genome Viewer (IGV). Gray bars correspond to the site of the break point.

As reported in *Table 5.1*, in addition to the canonical chimera FUS-DDIT3, we identified another gene fusion made up of the exons not involved in the canonical fusion gene that we named DDIT3-FUS. The presence of FUS-DDIT3 counterpart was already reported by (Antonescu et al., 2001) and (Braig et al., 2019) and never investigated further since it does not lead to a protein product.

5.2.2 *XENOGRAFT MODELS ARE REPRESENTATIVE OF THE TUMOUR BIOLOGY OF MLPS*

The first aim of this project is to evaluate the extent to which the genomic landscape of selected xenograft models, i.e. ML017, ML017/ET and ML006, represent the clinical picture. This is a fundamental step to generate data on the sensitivity to and resistance against trabectedin that could be further translated into the clinic. Data retrieved from the literature and reported in Background Section 2.1.4 on the genomic landscape of MLPS can be briefly summarised in *Table 5.2* and detailed as follows:

- presence of inactivating mutations in *PTEN* and *PIK3CA* genes;
- in line with the sarcoma cohort (SARC) of The Cancer Genome Atlas (TCGA) MLPS tumour genome is expected to have a low mutational burden with less than two variants per sample *(Figure 5.3*);
- amplifications in cytobands *13q, 8p23.3, 8q23.1* and deletions in the chromosome arm *16q* are hallmarks of MLPS.

Molecular feature	ML017	ML017/ET	ML006	Ref
PTEN	✓	✓	✓	(Assi et al., 2019), (Keung and Somaiah, 2019)
PIK3CA	✓	✓	✓	(Assi et al., 2019), (Keung and Somaiah, 2019)
low mutational burden	✓	✓	X	(Abeshouse et al., 2017)
amp 3q, 8p23.3, 8q23.1	✓	✓	X	(Ohguri et al., 2006)
del 16q	✓	✓	✓	(Koczkowska et al., 2017)

Table 5.2 MLPS specific features as reported in literature.

The synopsis reported in *Table 5.2* and in *Figure 5.3* describes for each model the extent to which their genomic architecture mirrors the above mentioned genomic benchmarks. In particular, while ML017 and ML017/ET genomic features perfectly match the genomic MLPS characteristics, the ML006 model presented a very high mutational burden, even greater than the highest mutated tumour like Skin Cutaneous Melanoma (SKCM) in TCGA. A possible explanation is that for ML006 replicated samples were not available (see Materials and methods, Section 4.8.2.1) and the lack of the matched healthy tissue could have affected the evaluation of both SNVs and SCNAs. These two technical problems overestimated variant calling, introducing false positive results.

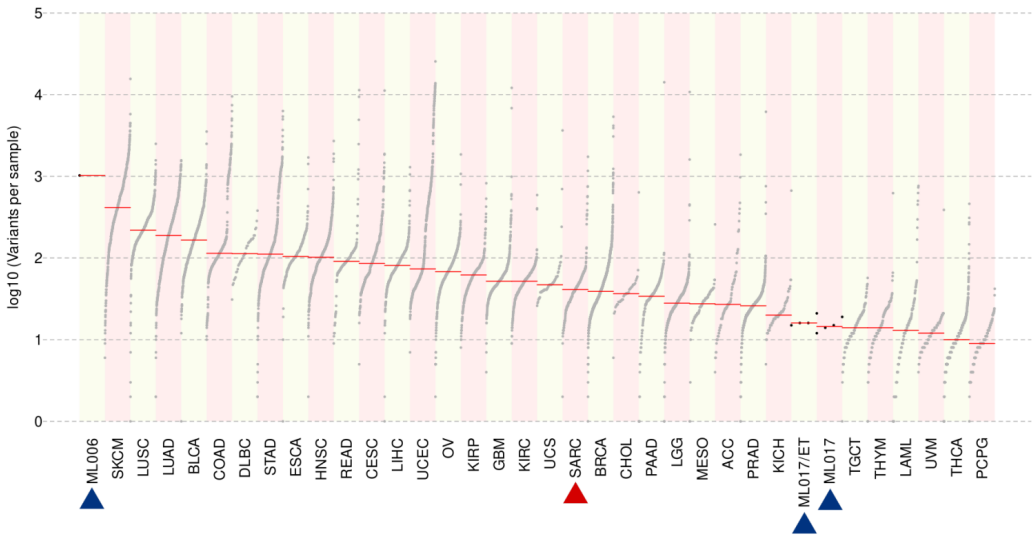

Figure 5.3 Number of variants per sample expressed in logarithmic scale of the whole cohort of TCGA and ML017, ML017/ET, ML006 models indicated by the blue arrows. The red arrow indicates the sarcoma (SARC) cohort.

Given these premises we concluded that ML006 could not be considered as a reliable MLPS model and thus it was excluded from downstream analysis.

For this reason, from this point on our analysis focused on the two models ML017 and ML017/ET that fairly reproduced MPLS features and showed either sensitivity to or resistance against trabectedin, respectively.

A summary of the workflow followed till this point is reported in *Figure 5.4*.

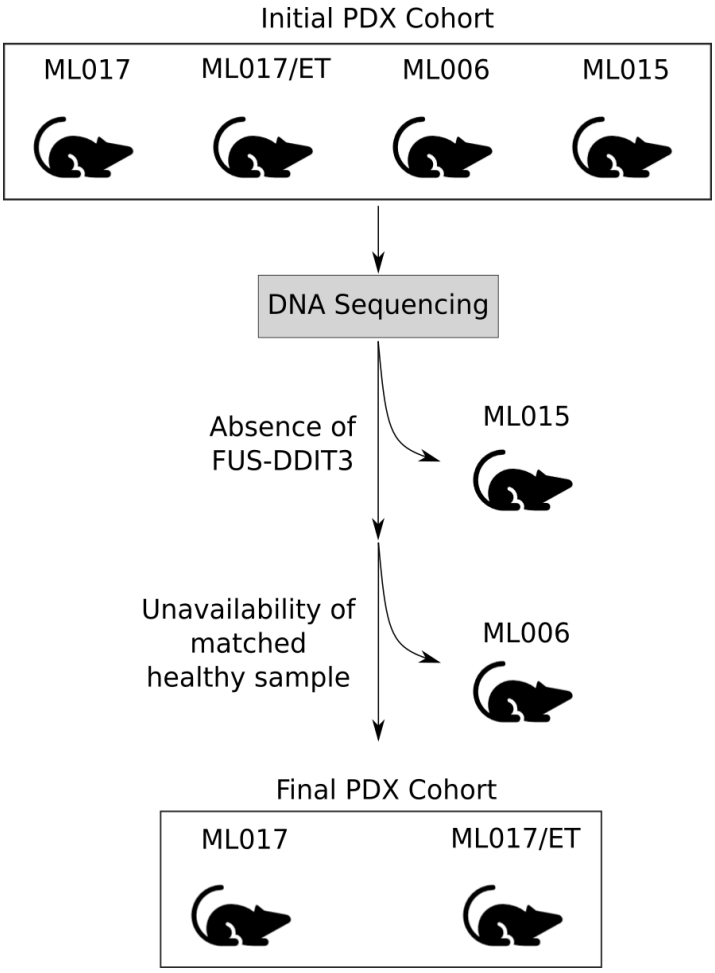

Figure 5.4 Workflow of the analysis of PDX models: from the starting cohort to the final one.

5.2.3 GENOMICS OF ML017 AND ML017/ET MODELS

Once assessed the fair reproducibility of the PDX models with the disease, the next step was to provide a more detailed map of the genomic architecture of ML017 and ML017/ET models and to study the effects of trabectedin exposure. In particular, the aims of this part of the project were:

1. to identify the modifications induced by trabectedin treatment at the genomic level;
2. to identify genomic features that determine the acquisition of resistance against trabectedin.

First, we examined the ploidy of the genome of both ML017 and ML017 models through fluorescence-activated cell sorting (FACS) (*Figure 5.5*). The genomic profile of ML017 and ML017/ET was compared to that of peripheral mononucleated blood cells (PBMC) that is used as a gold standard for diploidy. Both MLPS models showed a diploid genome.

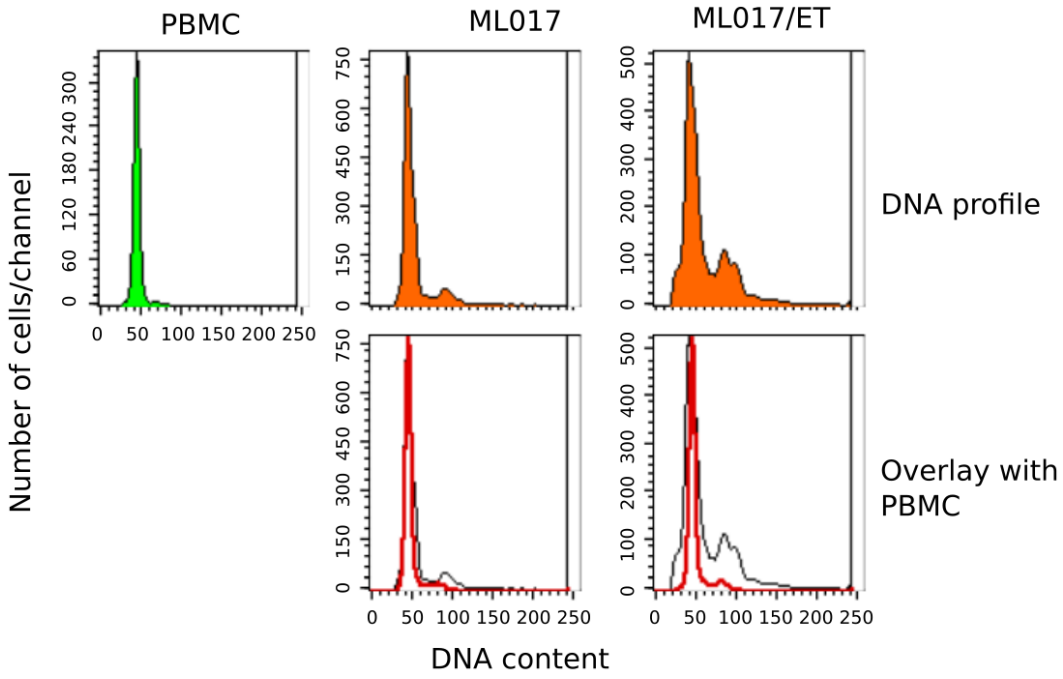

Figure 5.5 FACS analysis of ploidy in ML017 and ML017/ET control samples. The chromosomes set of both ML017 and ML07/ET was compared to peripheral mononucleated blood cells (PBMC).

5.2.4 TRABECTEDIN SELECTS SPECIFIC SCNAS REGIONS THAT COULD BE ASSOCIATED WITH DRUG RESISTANCE

Using a 400 bp bin density resolution across the genome, we found that at basal level the genome of both ML017 and ML017/ET models was widespread permeated by gains (mean length of 600 kbp, *Table 5.3*) and losses (mean length of 2 Mbp, *Table 5.4*) of genomic material. Next, to evaluate whether trabectedin exposure is able to increase the genomic instability, we compared the genomic landscape of ML017 and ML017/ET at different time points of exposure to trabectedin with the genomic landscape observed at baseline. Violin plots reported in *Figure 5.6* show: a difference in the total amount of SCNAs, with a median of 741 and of 1574 in the sensitive and resistant model, respectively (t-test, $p<0.01$),

although it should be noticed the high variability across the samples and under each condition. Both treatment with trabectedin and doxorubicin induced an increase in the number of SCNAs to a level that remained fairly the same under all conditions also in the resistant model.

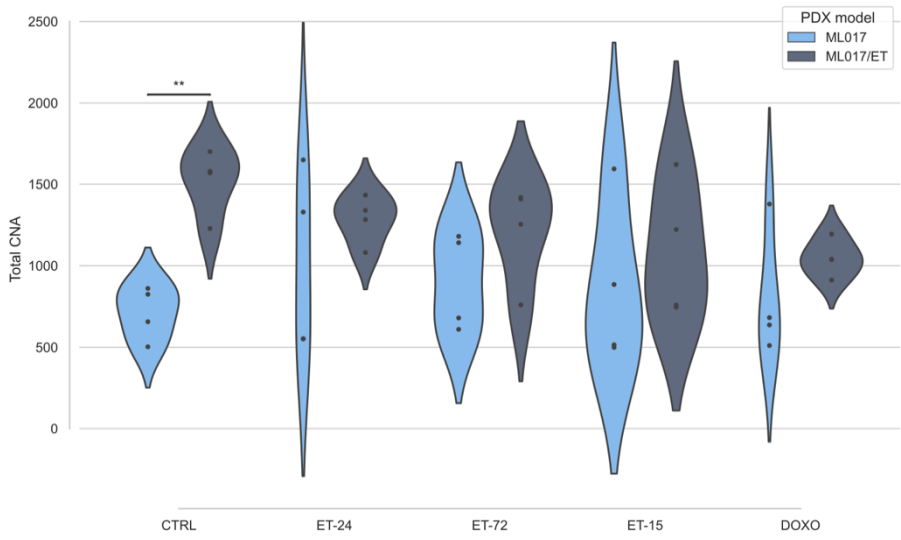

*Figure 5.6 Violin plots of the total number of CNAs in each condition: in light blue ML017, in violet ML017/ET as reported in the legend. ** p-value less than 0.001 in the statistical test.*

Chrom	Cytoband	N Genes	Wide Peak Limits	Peak Length (kbp)	Original q-values
chr1	1p31.3	2	61589474-61914436	324.96	<0.01
	1p34.3	2	37240800-37663887	423.09	<0.01
	1q21.2	6	147077908-147435089	357.18	<0.01
	1q25.3	1	181456957-181763014	306.06	<0.01
	1q42.13	1	228475861-228494372	18.51	<0.01
	1q42.2	3	231753619-232219082	465.46	<0.01
chr2	2p16.3	1	50107915-51255743	1147.83	<0.01
	2q14.2	1	121580216-121775182	194.97	<0.01
	2q31.1	14	176893120-177118668	225.55	<0.01
	2q31.2	1	179566802-179667063	100.26	<0.01
	2q31.2	3	179394648-179489597	94.95	<0.01
	2q33.1	1	200036891-200303102	266.21	<0.01
	2q36.1	1	223088347-223153835	65.49	<0.01
chr3	3p13	1	69808566-70044282	235.72	<0.01
	3p22.2	3	38591819-38889347	297.53	<0.01
	3p25.3	6	10370541-11096137	725.6	<0.01
	3p26.3	2	1-334035	334.03	<0.01
		3	2124598-3151723	1027.13	<0.01
	3q28	2	189286941-189584279	297.34	<0.01
chr6	6p21.1	1	45398115-45505708	107.59	<0.01
	6q16.3	1	101846861-102514980	668.12	<0.01
chr7	7p14.1	1	41892035-42269029	376.99	<0.01
	7q22.1	3	103134587-103958967	824.38	<0.01

	7q31.32	3	121960402-122519345	558.94	<0.01
	7q35	7	145773624-147993317	2219.69	<0.01
	7q36.3	1	159029132-159138663	109.53	<0.01
chr8	8p22	109	4016068-16172561	12156.49	<0.01
	8p23.3	8	315079-1658670	1343.59	<0.01
	8q13.3	1	72106345-72265754	159.41	<0.01
	8q23.1	1	106560444-106815714	255.27	<0.01
	8q23.3	1	116434272-116699126	264.85	<0.01
chr10	10p12.31	3	21104009-21482733	378.72	<0.01
	10q11.23	4	49917805-50266498	348.69	<0.01
	10q23.1	1	83567798-84768079	1200.28	<0.01
	10q23.1	2	87709155-88124016	414.86	<0.01
	10q24.32	6	103104366-103447020	342.65	<0.01
chr11	11p15.2	1	15998696-16210201	211.51	<0.01
	11p15.5	36	1020554-2199566	1179.01	<0.01
chr12	12p13.33	4	2304508-2828831	524.32	<0.01
	12q24.21	2	114607340-114874224	266.88	<0.01
chr13	13q21.32	4	66856843-67804352	947.51	<0.01
	13q22.3	4	78474639-78598534	123.9	<0.01
	13q31.3	5	92052308-95072405	3020.1	<0.01
chr15	15q14	2	33862775-34167146	304.37	<0.01
	15q21.1	7	48044868-48963771	918.9	<0.01
chr16	16p13.3	1	6028377-7800253	1771.88	<0.01
chr18	18p11.21	2	13781861-14009091	227.23	<0.01
	18q21.2	2	52921697-53254735	333.04	<0.01
	18q22.3	2	70451002-70911696	460.69	<0.01
chr20	20p13	4	1-194377	194.38	<0.01
	20q13.32	4	57382883-57457840	74.96	<0.01

Table 5.3 Significant gains as identified by GISTIC2. Chrom, chromosome; N Genes, number of genes in the cytoband. Original q-values refer to q-values as reported by GISTIC2.

Chrom	Cytoband	N Genes	Wide Peak Limits	Peak Length (kbp)	Original q-values
chr1	1p13.2	34	112539459-115315028	2775569	<0.01
	1q21.1	2	144290423-144360167	69744	<0.01
	1q21.2	2	148224797-148362580	137783	<0.01
chr2	2p13.3	17	69472725-70701833	1229108	<0.01
	2p21	5	47366522-47749906	383384	<0.01
	2p22.3	6	31799651-32585413	785762	<0.01
	2q35	16	219029999-219510284	480285	0.01
chr3	3p21.31	29	47050542-48601711	1551169	<0.01
	3q26.1	1	164054699-164909416	854717	<0.01
chr4	4p15.2	23	21944934-28822631	6877697	<0.01
	4p16.3	65	1-3514129	3514128	<0.01
chr5	5q13.2	16	68834623-70680201	1845578	<0.01
	5q23.2	6	125585649-126205190	619541	<0.01
chr6	6p21.33	191	29715511-33412760	3697249	<0.01
	6q23.2	3	134368652-135239520	870868	<0.01
chr7	7p22.1	31	5108588-6728836	1620248	<0.01
	7q11.23	14	74547063-75167292	620229	<0.01
	7q32.1	4	128021110-128318003	296893	<0.01

chr9	9p21.3	20	21304680-26841974	5537294	<0.01
	9q13	50	38615166-71152989	32537823	<0.01
	9q33.3	10	127525655-128201403	675748	<0.01
chr10	10q11.22	25	46159105-48361732	2202627	<0.01
		4	48854748-49365043	510295	<0.01
	10q11.23	10	50962310-51828194	865884	<0.01
	10q21.3	22	69453154-71168965	1715811	<0.01
	10q22.1	44	73552721-76804678	3251957	<0.01
chr11	11p15.5	48	1-846991	846,99	<0.01
	11q24.1	5	122843940-123259019	415079	<0.01
chr12	12p13.31	9	6232487-6602293	369806	<0.01
chr13	13q22.1	3	73329284-74268469	939185	<0.01
chr14	14q11.2	8	23572621-23815799	243178	<0.01
	14q13.2	14	34417868-36008844	1590976	<0.01
	14q21.1	2	39355533-39623557	268024	<0.01
	14q21.3	17	48262302-50789048	2526746	<0.01
	14q24.3	26	73026992-74714617	1687625	<0.01
chr15	15q21.2	10	50506753-51248273	741,52	<0.01
	15q25.2	9	82741209-83164924	423715	<0.01
chr16	16q23.1	24	74228326-76320597	2092271	<0.01
	16q24.3	4	89290812-89628699	337887	<0.01
chr17	17p11.2	18	15718462-16696018	977556	0.07
	17q21.31	61	41375044-43522663	2147619	<0.01

Table 5.4 Significant losses as identified by GISTIC2. Chrom, chromosome; N Genes, number of genes in the cytoband. Original q-values refer to q-values as reported by GISTIC2.

In order to overcome technical biases and the high variability, and to identify high confidence SCNAs, we used very stringent conditions: we selected only SCNAs called as significant by the GISTIC2 algorithm that were present in all replicates per each condition and discarded all the remaining. Using this approach we probably lost information, however this allowed us to select high confidence events. Overall we identified 51 regions in gain and 42 regions in loss that we grouped into groups.

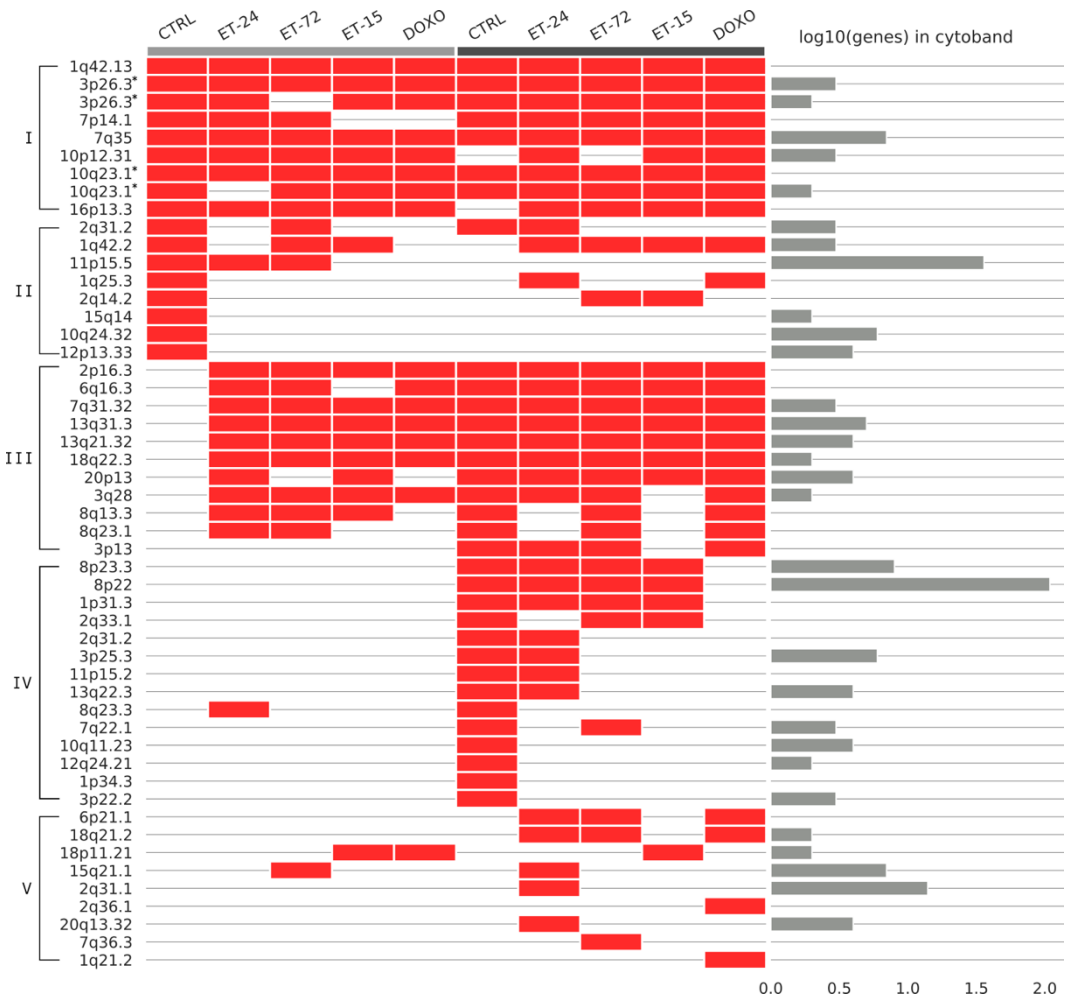

Figure 5.7 Representation of the regions of gain in ML017 (grey bar) and ML017/ET (black bar) models. The number of genes in the cytoband are expressed in logarithmic scale and shown in the right side.

As shown in *Figure 5.7* we defined five groups for gains. Group I includes those that are present under all conditions in both models, thus being representative of the PDX models. Group II and V show gains that are scattered across treatment and models. Group III is composed by genomic regions emerged upon trabectedin or doxorubicin treatment. These regions still characterise ML017/ET in each

condition: they could have been selected through the prolonged (over two-years) treatment with trabectedin, thus they could be associated with the mechanism of acquired resistance. Finally, Group IV is composed of cytobands that are affected in ML017/ET only with a group of them consisting of *8p23.3, 8p22, 1p31.3, 3p25.3* that are specific of trabectedin treatment.

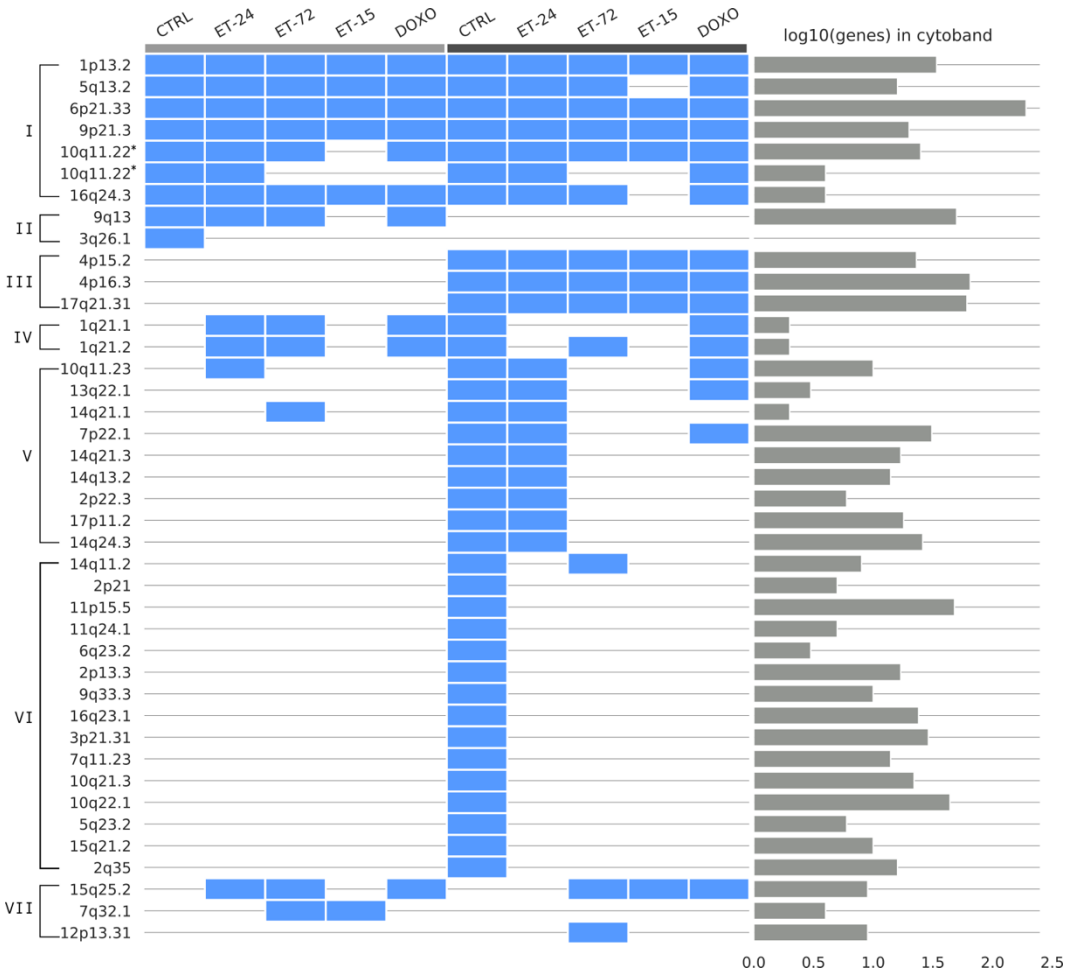

Figure 5.8 Representation of the regions of loss in ML017 (grey bar) and ML017/ET (black bar) models. The number of genes in the cytoband are expressed in logarithmic scale and shown in the right side.

Then, we defined seven groups for the losses *(Figure 5.8)*. Groups I, II, III and IV could be associated with the understanding of the mechanisms of drug response, while the remaining groups V, VI and VII, seem not to be related to any specific condition. Group I represents a specific feature of PDX models since it comprises regions that are common to all conditions in both. Group II shows the cytoband *9q13* that is affected in all conditions in only ML017 apart from ET-15, thus it probably represents a region selected through prolonged treatment. Interestingly,

Group III is composed of regions that are specific to ML017/ET model only, such as *4p15.2, 4p16.3* and *17q21.31* that comprise 149 genes. The exclusivity of these regions in ML017/ET might be associated with acquired resistance against trabectedin. Finally, Group IV is composed of two cytobands, such as *1q21.1* and *1q21.2*, which are lost after treatment regardless of trabectedin or doxorubicin at the early time points, while they are longer present at ET-15, as well as after treatment with trabectedin in ML017/ET.

In conclusion, we identified CNAs that are specific features of the PDX models, like Group I in either gains or losses; Group III (gains) which could be implicated in the process of acquired resistance; and Group III (losses) which may be a specific characteristic of ML017/ET and could be defined as marker of resistance.

5.2.5 *Beyond PTEN and PIK3CA: DSC3, FAT4, IMPDH2 and PAK7 as novel mutational features of MLPS*

Focusing on the SNVs compendia, we initially assessed the percentage of the genome affected by SNV, a parameter known as mutational load in ML017 and ML017/ET. We identified 82 and 142 affected genes, respectively, and a similar median number of variants per sample of 16 and 19 (*Table 5.5*). Data suggests no significant variant selection upon treatment. The different number of total variants, 336 and 415 (*Table 5.5*) in ML017 and ML017/ET, respectively, is mainly due to one replicate of ET-24 condition in ML017/ET that behaves as an outlier (*Figure 5.9*). Most variants in both models were single base changes, while less than 50 were insertions or deletions.

	ML017	**ML017/ET**
Number of samples	20	20
Number of affected genes	82	142
Median number of variants per sample	16	19
Total variants	336	415

Table 5.5 Number of mutated genes and number of variants in ML017 and ML017/ET cohorts.

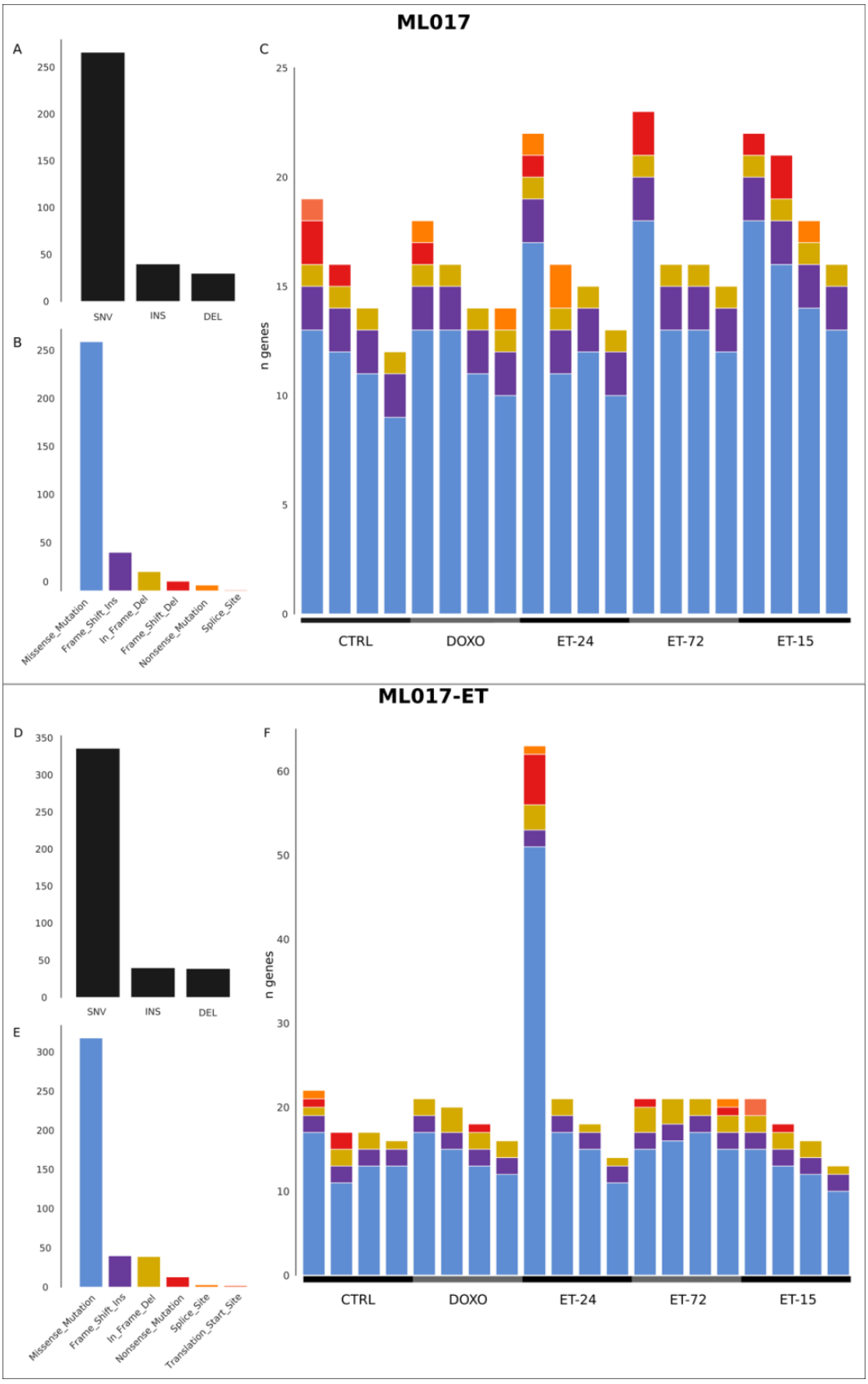

Figure 5.9 Mutational load of ML017 (upper panel) and ML017/ET (lower panel) cohorts. The number of SNV, insertions (INS) and deletions (DEL) are reported.

Then, we went deeper in the characterisation of these variants and identified the most frequent in the two cohorts. As shown in *Figure 5.10*, in addition to the previously cited mutations in *PTEN* and *PIK3CA* genes (described in Table 4.6), we identified a set of genes such as *ANK3, CRYGB, DSC3, FAT4, IMPDH2, KCNA3, MUC7* and *PAK7* (also known as *PAK5*) harbouring mutations that characterised the whole cohort in both models (described in *Table 5.7* and *Table 5.8*). Other genes such as *ABCA13, BRCA1, SNTA1, AOX1, ZFPM2, THBS4, ABCC9, ABCG2* in ML017 and CRYGB, HYDIN, MAP2, DIAPH2, AKAP13, PAX9, DICER1, SRP72 in ML017/ET are specific of the single replicate implying some biological variability within the samples. As reported in *VariantsAllelicFraction.xlsx* available at[3], all samples are characterised by the same mutated locus in each gene with a high allelic fraction close to 50%, suggesting a prevalently monoclonal nature of the tumour cells. Genes in *Table 5.7*, such as *ANK3, CRYGB, FAT4* (variant c.9674C>A), *MUC7*, and *KCNA3,* were reported as benign or well-tolerated, thus with no significant impact on the final protein structure. Differently, genes in *Figure 5.10* and *Table 5.8*, such as *DSC3, FAT4* (c.9678T>A), *IMPDH2* and *PAK7*, carried probably damaging or deleterious variants, thus with a detrimental effect on the protein translation.

In conclusion, we identified two genes that are frequently mutated in MLPS, such as *PTEN* and *PIK3CA*, and a group of variants in genes such as *DSC3*, *FAT4* (variant c.9678T>A), *IMPDH2* and *PAK7*, that could impaired protein translation.

[3] Figshare: https://doi.org/10.6084/m9.figshare.17181104.v1,
Zenodo: https://doi.org/10.5281/zenodo.5807695

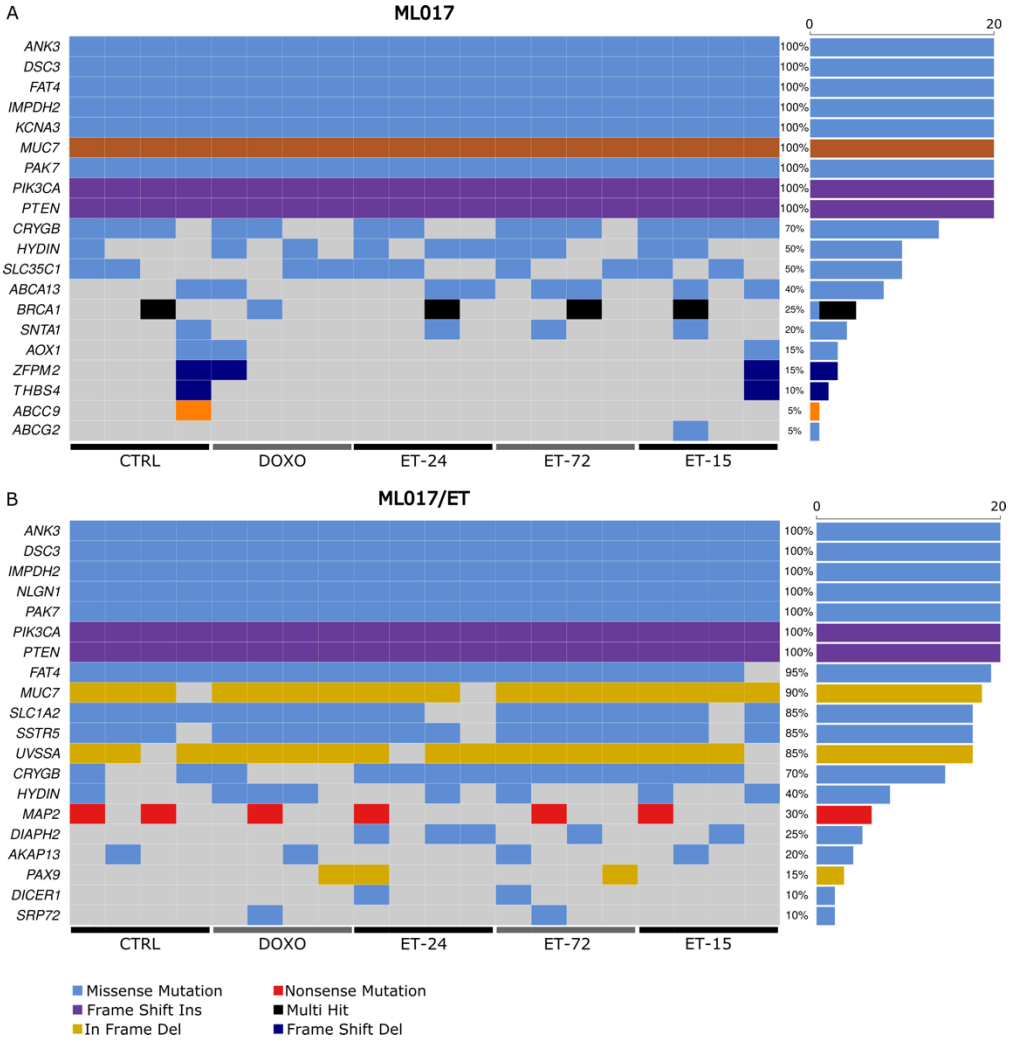

Figure 5.10 Oncoplot of ML017 (upper panel) and ML017/ET (lower panel) cohorts. Genes are reported on the left, the number of samples carrying a mutation in that gene is reported on the right. ET, trabectedin; DOXO, doxorubicin.

Gene Symbol	Entrez ID	Cytoband	start	end	ref	alt	type	subtype	impact
PIK3CA	5290	3q26.32	178952148	178952149	C	CTTTT	indel	ins	frameshift_variant
PTEN	5728	10q23.31	89653831	89653832	G	GGC	indel	ins	frameshift_variant

Gene Symbol	codon_change	aa_change	polyphen	sift	dbsnp	Impact severity	in Cosmic	in 1000 Genome Project	cadd scaled
PIK3CA	c.3205_3206insTTTT	p.Ter1069PhefsTer5	-	-	-	HIGH	No	No	-
PTEN	c.132_133dup	p.Val45AlafsTer10	-	-	-	HIGH	No	No	-

Table 5.6 Annotations for variants in typical MLPS genes PTEN and PIK3CA. ref, reference; alt, alterate; impact, type of variant; aa, aminoacid; sift, SIFT prediction; dbsnp, code of the variant as reported in dbsnp; cadd scaled, impact of the variant as reported by CADD. For more information on annotations see Materials and methods, Chapter 4.

Gene Symbol	Entrez ID	Cytoband	start	end	ref	alt	type	subtype	impact
ANK3	288	10q21.2	61834708	61834709	G	T	snp	tv	missense_variant
CRYGB	1419	2q33.3	209010498	209010499	G	A	snp	ts	missense_variant
FAT4	79633	4q28	126371844	126371845	C	A	snp	tv	missense_variant
MUC7	4589	4q13.3	71347456	71347460	GACT	G	indel	del	inframe_deletion

Gene Symbol	codon_change	aa_change	polyphen	sift	dbsnp	Impact severity	in Cosmic	in 1000 Genome Project	cadd scaled
ANK3	c.5930C>A	p.Pro1977Gln	benign	Tolerated low confidence	-	MED	No	No	8,78
CRYGB	c.251C>T	p.Pro84Leu	benign	tolerated	rs200143566	MED	No	No	5,29
FAT4	c.9674C>A	p.Thr3225Lys	benign	tolerated	-	MED	No	No	-
MUC7	c.1003_1005del	p.Thr335del	-		rs867973893	MED	No	No	-

Table 5.7 Annotations for shared variants in ML017 and ML017/ET. ref, reference; alt, alterate; impact, type of variant; aa, aminoacid; sift, SIFT prediction; dbsnp, code of the variant as reported in dbsnp; cadd scaled, impact of the variant as reported by CADD. For more information on annotations see Materials and methods, Chapter 4.

Gene Symbol	entrezID	Cytoband	start	end	ref	alt	type	subtype	impact
DSC3	1825	18q12.1	28574258	28574259	C	A	snp	tv	missense_variant
FAT4	79633	4q28	126371848	126371849	T	A	snp	tv	missense_variant
IMPDH2	3615	3p21.31	49064004	49064005	A	G	snp	ts	missense_variant
PAK7	57144	20p12.2	9520248	9520249	G	A	snp	ts	missense_variant

Gene Symbol	codon_change	aa_change	polyphen	sift	dbsnp	Impact severity	in Cosmic	in 1000 Genome Project	cadd scaled
DSC3	c.2573G>T	p.Gly858Val	Probably damaging	deleterious		MED	No	No	18,1
FAT4	c.9678T>A	p.Asn3226Lys	Probably damaging	deleterious		MED	No	No	11,73
IMPDH2	c.857T>C	p.Met286Thr	benign	deleterious		MED	No	No	13,69
PAK7	c.2020C>T	p.Arg674Trp	Probably damaging	deleterious	rs575889963	MED	No	Yes	-

Table 5.8 Annotations for shared variants with a predicted damaging impact on the protein structure in ML017 and ML017/ET. ref, reference; alt, alterate; impact, type of variant; aa, aminoacid; sift, SIFT prediction; dbsnp, code of the variant as reported in dbsnp; cadd scaled, impact of the variant as reported by CADD. For more information on annotations see Materials and methods, Chapter 4.

5.2.6 *UVSSA AS A POSSIBLE MARKER OF ACQUIRED RESISTANCE*

The comparison between ML017 and ML017/ET revealed five differentially mutated genes (DMGs) (Fisher-test, p-value <0.01) that characterised each model exclusively as shown in *Figure 5.11*: *KCNA3* was specifically mutated in ML017, while variants in genes like *NLGN1, UVSSA, SSTR5, SLC1A2* were private of ML017/ET. All variants had no impact on the protein structure as shown in *Table 5.9*. However, the *UVSSA* gene, affected by an in-frame deletion of six bases and was also identified in the *4p.16* cytoband specifically lost in ML017/ET (see Section 5.2.4 of this Chapter). Given the diploidy of the genome of ML017 and ML017/ET (*Figure 5.5*), *UVSSA* lost one allele and the other one was affected by an in-frame deletion, thus it is unlikely to be translated into a protein.

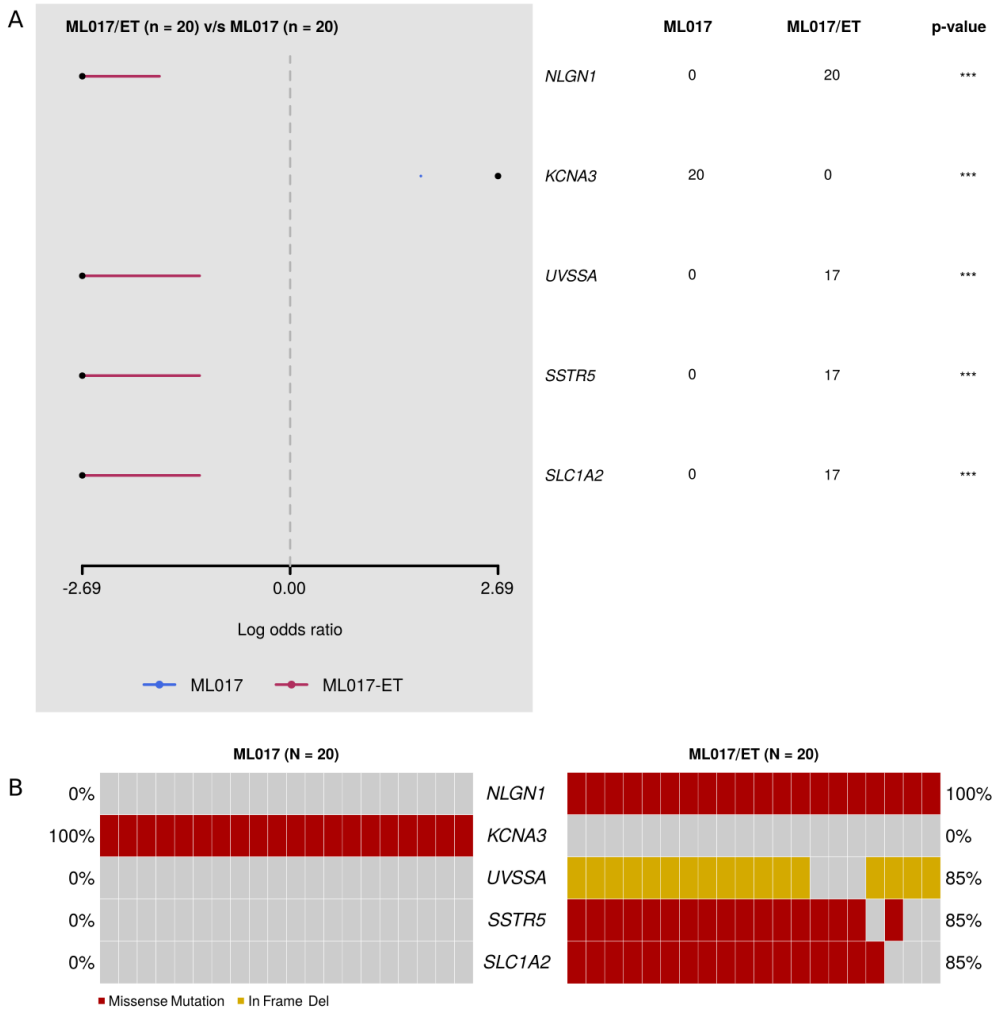

Figure 5.11 Differentially mutated genes (DMGs) between ML017 and ML017/ET. Panel A: forest plot and associated p-values. Panel B: samples carrying variants in DMGs in ML107 and ML017/ET.

Gene Symbol	entrezID	Cytoband	start	end	ref	alt	type	subtype	impact
SLC1A2	6506	11p13	35333893	35333894	G	T	snp	tv	missense_variant
SSTR5	6755	16p13.3	1129828	1129829	C	G	snp	tv	missense_variant
NLGN1	22871	3q26.31	173998554	173998555	C	A	snp	tv	missense_variant
UVSSA	57654	4p16.3	1343466	1343473	AGTTCCT	A	indel	del	inframe_deletion
KCNA3	3738	1p13.3	111215915	111215916	G	C	snp	tv	missense_variant

Gene Symbol	codon_change	aa_change	polyphen	sift	dbsnp	Impact severity	in Cosmic	in 1000 Genome Project	cadd scaled
SLC1A2	c.412C>A	p.Leu138Met	possibly_damaging	tolerated		MED	No	No	-
SSTR5	c.961C>G	p.Leu321Val	benign	tolerated		MED	No	No	14,16
NLGN1	c.1934C>A	p.Pro645His	benign	tolerated		MED	No	No	13,69
UVSSA	c.256_261del	p.Phe86_Leu87del				MED	No	No	-
KCNA3	c.1516C>G	p.Gln506Glu	benign	tolerated		MED	No	No	9,59

Table 5.9 Annotations for variants in DMGs in ML017 and ML017/ET. ref, reference; alt, alterate; impact, type of variant; aa, aminoacid; sift, SIFT prediction; dbsnp, code of the variant as reported in dbsnp; cadd scaled, impact of the variant as reported by CADD. For more information on annotations see Materials and methods, Chapter 4.

5.3 STUDY ON THE EFFECTS OF TRABECTEDIN AT THE TRANSCRIPTIONAL LEVEL

In order to study the transcriptional effects of trabectedin exposure in PDX models, RNA-Seq experiments were performed on RNA purified from tumour biopsies withdrawn at the same treatment schedule using a whole-transcriptome RNA-Seq kit which allows the quantification of both protein-coding and non-coding RNAs (see Materials and methods, Section 4.8.3.2). After quality control three samples, namely ML017/ET-2-CTRL-B, ML017/ET-2-ET-72-B, ML017/ET-2-ET-15-B, were excluded from downstream analysis (see Materials and methods, Section 4.8.2.2).

The compendia of RNA reads were analysed with two different levels of complexity: i) pseudo-counts algorithms were used to investigate the modulation of known annotated genes (see Materials and methods, Section 4.8.3.2.1); ii) transcripts reconstruction algorithms were exploited to identify novel features that could drive yet unknown transcriptional mechanisms (see Materials and methods, Section 4.8.3.2.2).

The aims of this part of the project can be briefly summarised as follows:

- to compare the RNA-Seq data of MLPS models at basal conditions to identify transcriptional markers that could explain acquired resistance against the treatment;
- to study the time course effects of trabectedin in both models;
- to compare the transcriptional effects of trabectedin to the standard treatment with doxorubicin.

First, we evaluated sample distribution with Principal Component Analysis (PCA) in order to identify preferential groups in an unsupervised manner. Data depicted in *Figure 5.12* shows three major groups:

- group A: defined by samples on the right-bottom side of the graph. It is composed by CTRL ML017 samples and those at ET-24 or ET-72 and the vast majority of samples from ML017/ET PDX model. A more detailed insight revealed that group A is heterogeneous as it still possible to distinguish two subgroups: A1 including ML017 sample (indicated by triangles) and A2

including ML017/ET samples (indicated by circles), thus suggesting a slight differences between the two models;

- group B depicted in orange colour, is associated with samples under doxorubicin treatment in both ML017 and ML017/ET. This suggests that treatment with doxorubicin elicits similar responses in both ML017 and ML017/ET models;
- group C represented by all four replicates of ET-15 in ML017, thus suggesting a completely different transcriptional modulation at this time point in the responsive model only in comparison to the previous groups.

Overall, this unsupervised cluster analysis strongly suggests that long term exposure to trabectedin in ML017 PDX models induces a completely different transcriptional program in comparison to early time point that is warranted of detailed investigation.

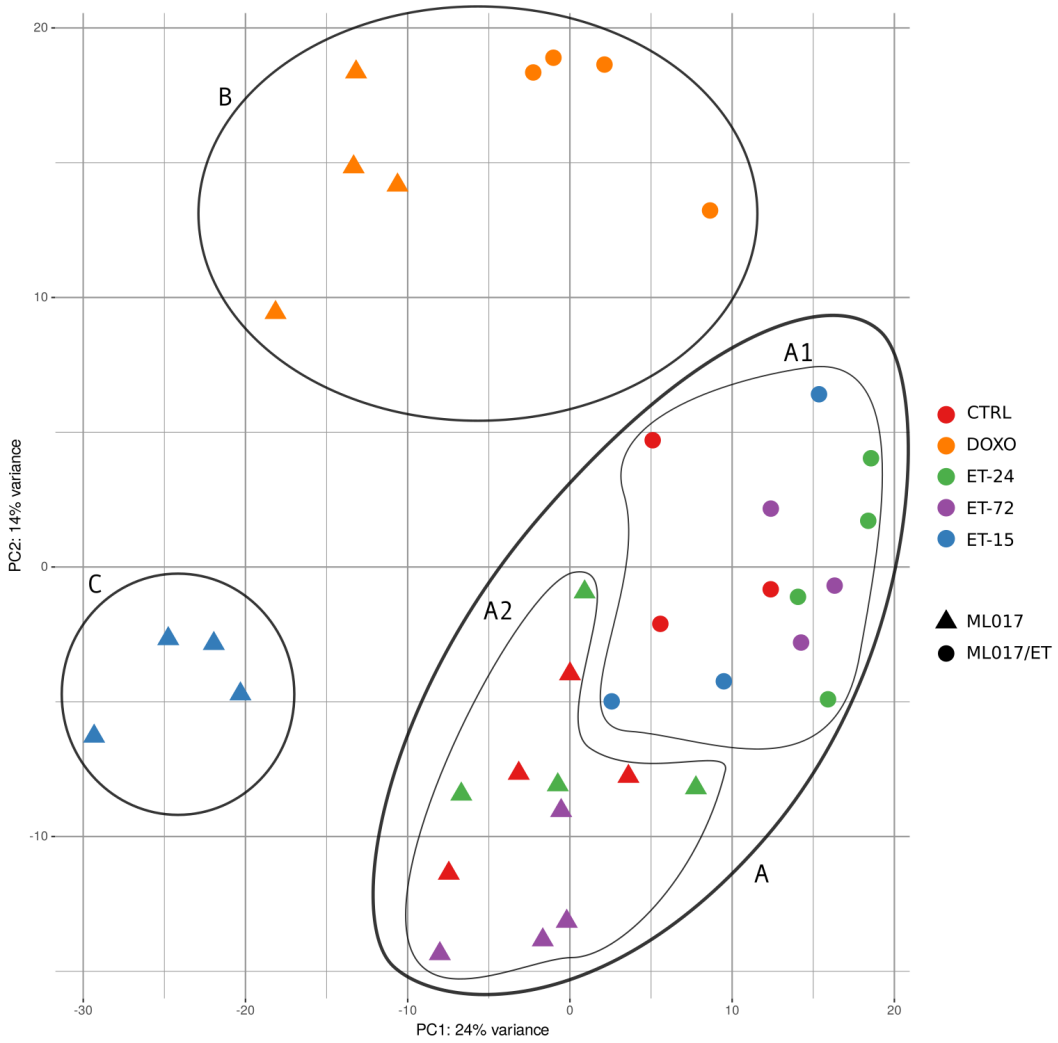

Figure 5.12 PCA of ML017 and ML017/ET samples based on pseudocounts. ML017 are depicted as triangles, ML017/ET as circles. Colours indicate conditions as in the legend. PC, principal component.

5.3.1 TRANSCRIPTIONAL COMPARISON AND GENOMIC DATA INTEGRATION OF *ML017* AND *ML017/ET* AT BASAL CONDITIONS

In order to identify transcriptional markers that could explain acquired resistance to trabectedin treatment, we compared the transcriptional landscape derived from PDX tumours at basal conditions. To this aim, we performed a differential expression analysis between ML017/ET and ML017. We identified 243 differentially expressed genes (DEGs) 74 of which were up-regulated and 169 of which were down-regulated (*Figure 5.13*).

We initially performed two different types of analysis in order to dissect the role of DEGs in driving trabectedin resistance: 1) pathway analysis and 2) genomic mapping analysis. The first revealed that DEGs were not significantly involved in any biological pathway suggesting that the mechanism of resistance does not depend on specific biological function. Instead, when we analysed the genomic regions on which each DEG mapped, we observed that the vast majority of them belonged to the same genomic cytobands. In particular, 63 down-regulated genes mapped on *4p16.3*, *4p15.2* and *17q21.3* cytobands, thus exactly the same regions found in the ML017/ET model (*Figure 5.8*). As shown in *Figure 5.14*, the transcriptional down-regulation and the loss of genetic material are coherent events, thus suggesting a specific inhibition of these genes in ML017/ET, therefore they could be considered as specific discriminant between responsive and resistant models.

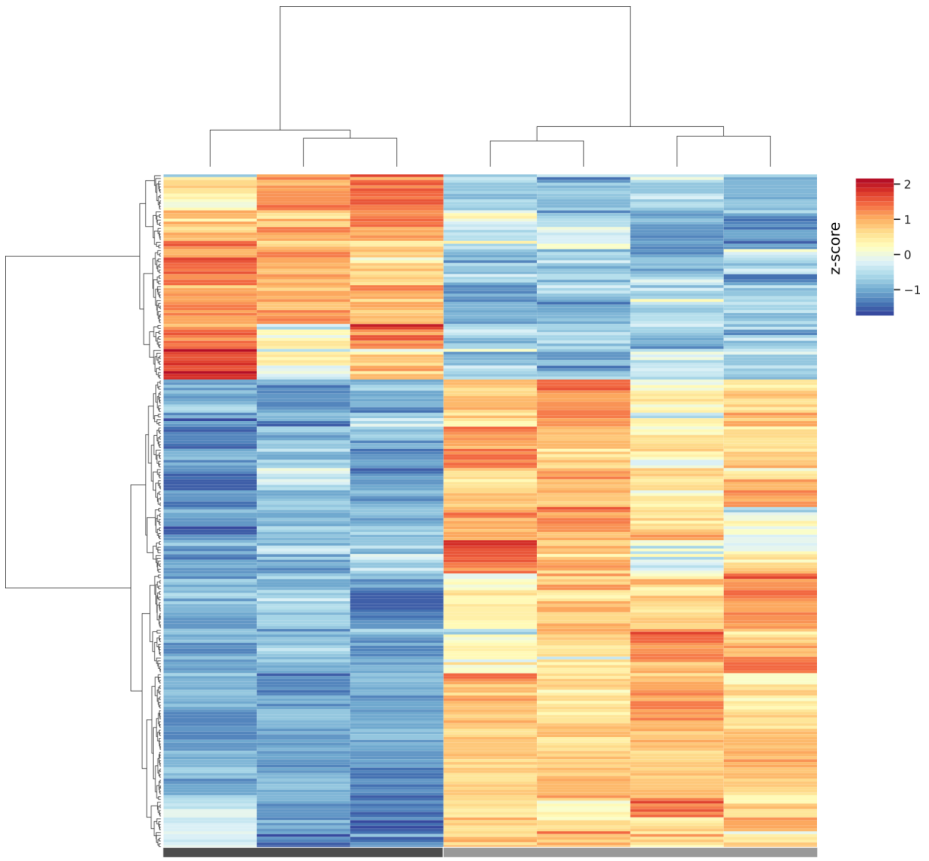

Figure 5.13 Heatmap showing z-scores of normalised value counts of DEGs in the comparison between ML017/ET (black bar) and ML017 (grey bar) control samples.

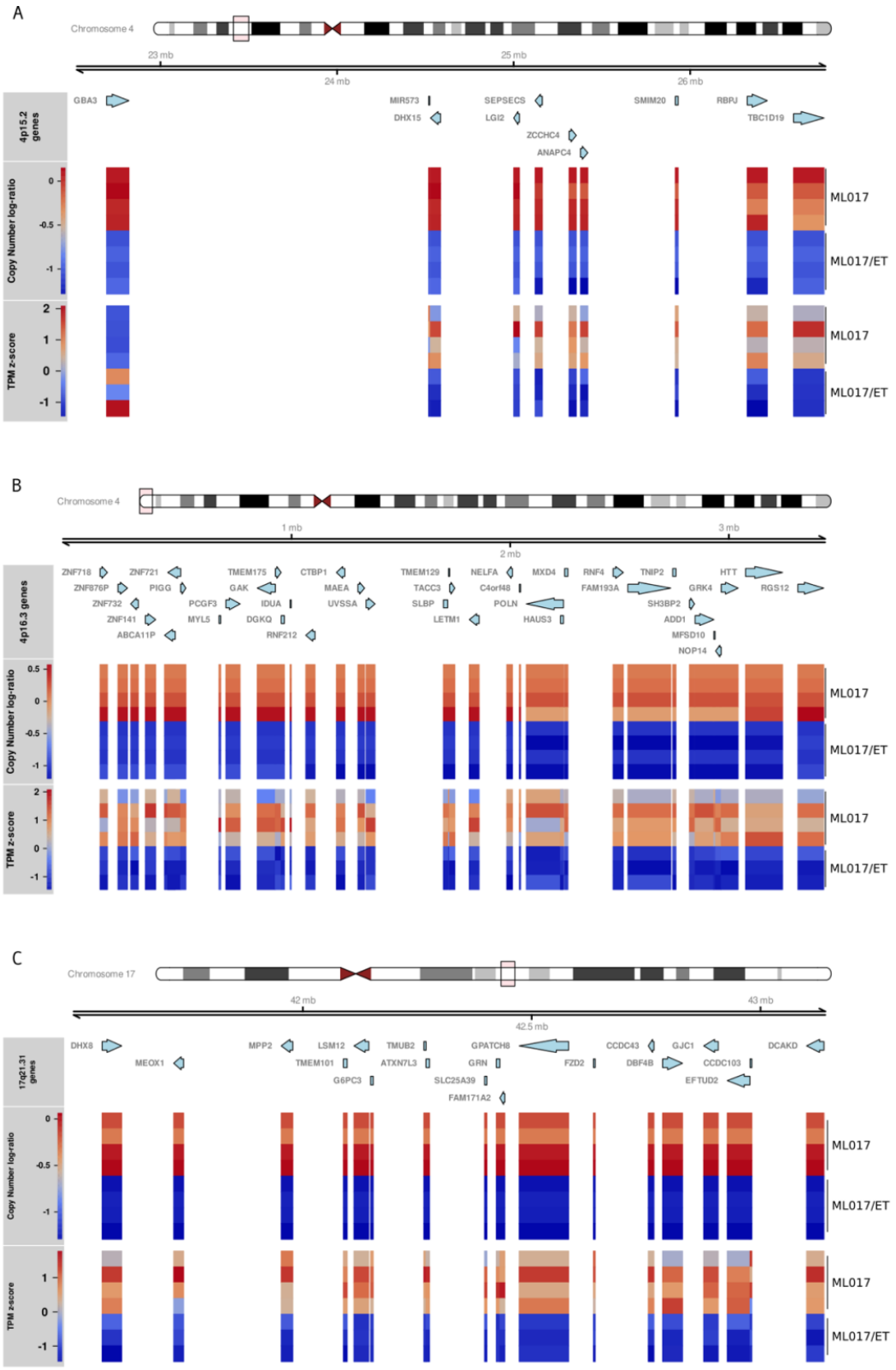

Figure 5.14 Copy number and transcriptional data integration related to DEGs mapping to cytobands in Group I losses. Panel A refers to 4p15.2, panel B to 4p16.3 and panel C to 17q21.31. For each panel, the upper side shows the chromosome ideogram with cytoband of interest highlighted by a rectangle. Lines show the length of the region of interest (mb, mega bases). Gene panel shows genes that map to the region and are also DEGs. Arrow indicates the coding reading direction. Copy number log-ratio panel shows values as reported by GISTIC2. TPM z-score panel depicts the transcript per million (TPM) z-score of gene expression. Colours as reported in the legends on the left.

5.3.2 *TIME-DEPENDENT TRANSCRIPTIONAL EFFECTS OF TRABECTEDIN*

In order to study the time course effects of trabectedin in both models, we studied the transcriptional modulation induced by trabectedin treatment. For each model, analysis was performed by comparing each time point to its own basal conditions as reported and described in Materials and methods, Section 5.3.2.

Upper panel of *Figure 5.15* shows that in ML017 the greatest transcriptional modulation was elicited at ET-15 (4883 DEGs), while 829 and 209 DEGs were found at ET-24 and ET-72, respectively[4]. The lower panel of Figure 5.15 reported a selection of the most significant pathways obtained through gene set enrichment analysis (GSEA)[4]: to provide an indication of the pathway regulation, circles are coloured based on the normalised enrichment score (NES, see Materials and methods, Chapter 4), purple for positive values and green for negative values. At ET-24, pathways were mainly related to the transcriptional regulation of the TP53 gene, the activation of general gene expression and MAPK-related processes. The same pathways, excepting the MAPK, were still up-regulated at 72h, although resulted less enriched. At the same time point we found the downregulation of the pathways related to the metabolism of steroids and lipids. At ET-15 the 4883 DEGs were grouped into 233 pathways, of which 40 were activated and 193 of which were inhibited. The activated pathways were mainly involved in functions like the remodelling of the extracellular matrix organisation, the production of collagen and the regulation of the insulin-like Growth Factor (IGF) suggesting a phenotypic change in MLPS cells. Differently, at this time point functions like the methylation of the DNA, the regulation of rRNAs and the pathway of the RNA polymerase I were inhibited.

[4] Figshare: https://doi.org/10.6084/m9.figshare.17181104.v1,
Zenodo: https://doi.org/10.5281/zenodo.5807695

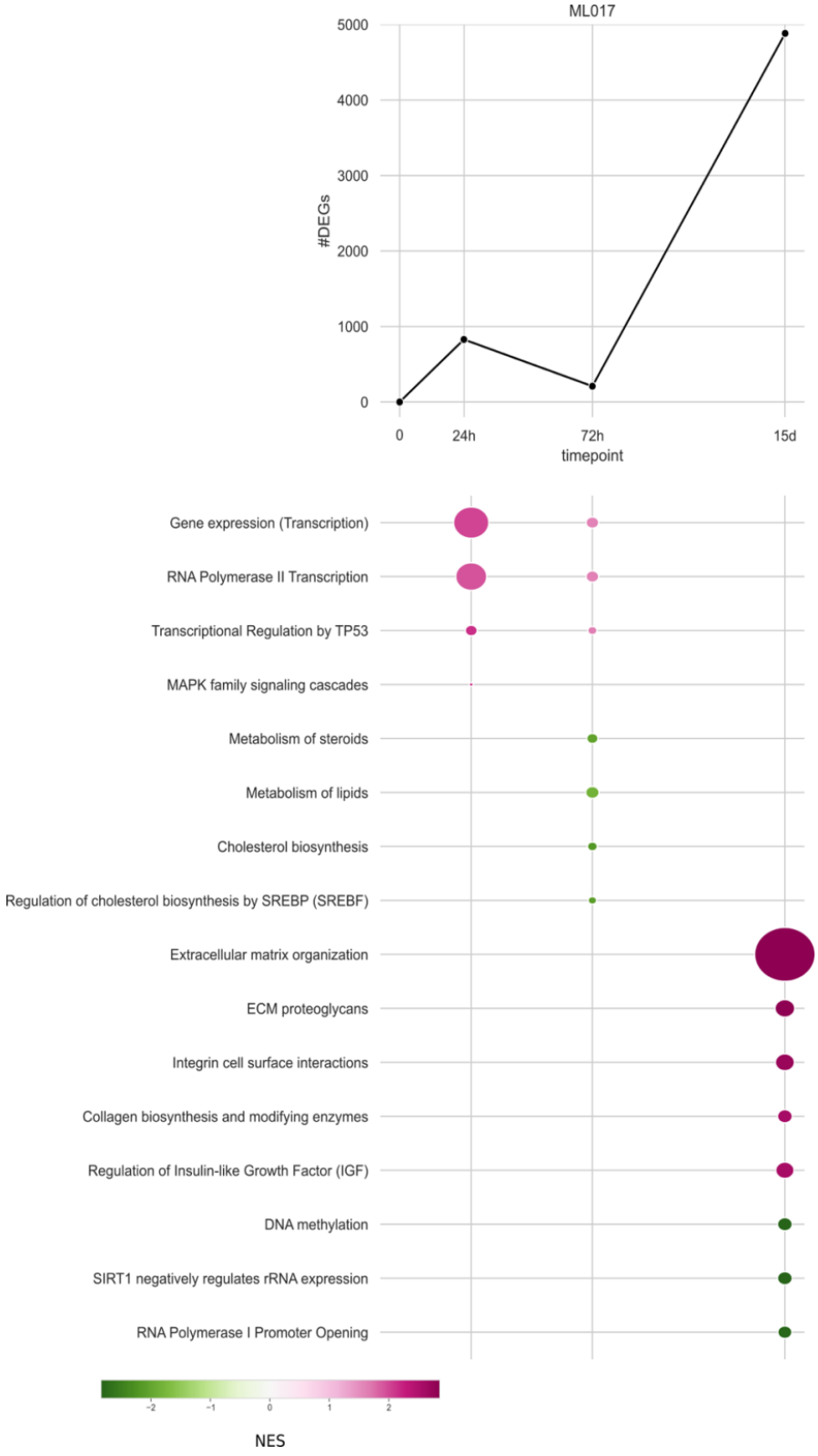

Figure 5.15 Upper panel shows the number of DEGs modulated by trabectedin in ML017 model. Control is set to zero. Lower panel shows most significant pathways from GSEA with Reactome. The colour represents the normalised enrichment score (NES) as indicated in the legend: negative values for negative enrichment (down-regulation), positive values for positive enrichment (up-regulation). Radius of the circles is proportional to the number of genes in the pathways.

In summary, the overall data suggests that trabectedin has a biphasic effect on the sensitive ML017 PDX model: an early effect (24h and 72h) mainly characterised by a-specific cytotoxic response followed by a second delayed transcriptional effect that drives the phenotypic and morphological changes of the tumour cells.

Analysis was performed following the same pipeline in the ML017/ET resistant model. As shown in *Figure 5.16*, trabectedin induced an early effect at ET-24 (1052 DEGs), while it dramatically decreased along with the time of treatment (219 and 3 DEGs, respectively). In correspondence to the first response, we identified the activation of the transcriptional regulation of TP53, the RNA polymerase II transcription and the gene expression pathways that were equally regulated in the responsive model at the same time point. Otherwise, at the same time in ML017/ET we also found the activation of the pathways of the cell cycle, the homology direct repair and the inhibition of the pathway of the GPCR. At 72h, the main effect was related to the activation of the immune system, while at 15 days the 3 DEGs were not associated with any function.

This evidence shows the differences in the transcriptional response to trabectedin between ML017 and ML017/ET models: in the responsive model the drug is able to trigger a time-dependent effect as shown previously, while in the resistant model it induces an early cytotoxic effect. Prolonged treatment does not elicit any other response in ML017/ET thus confirming resistance to trabectedin.

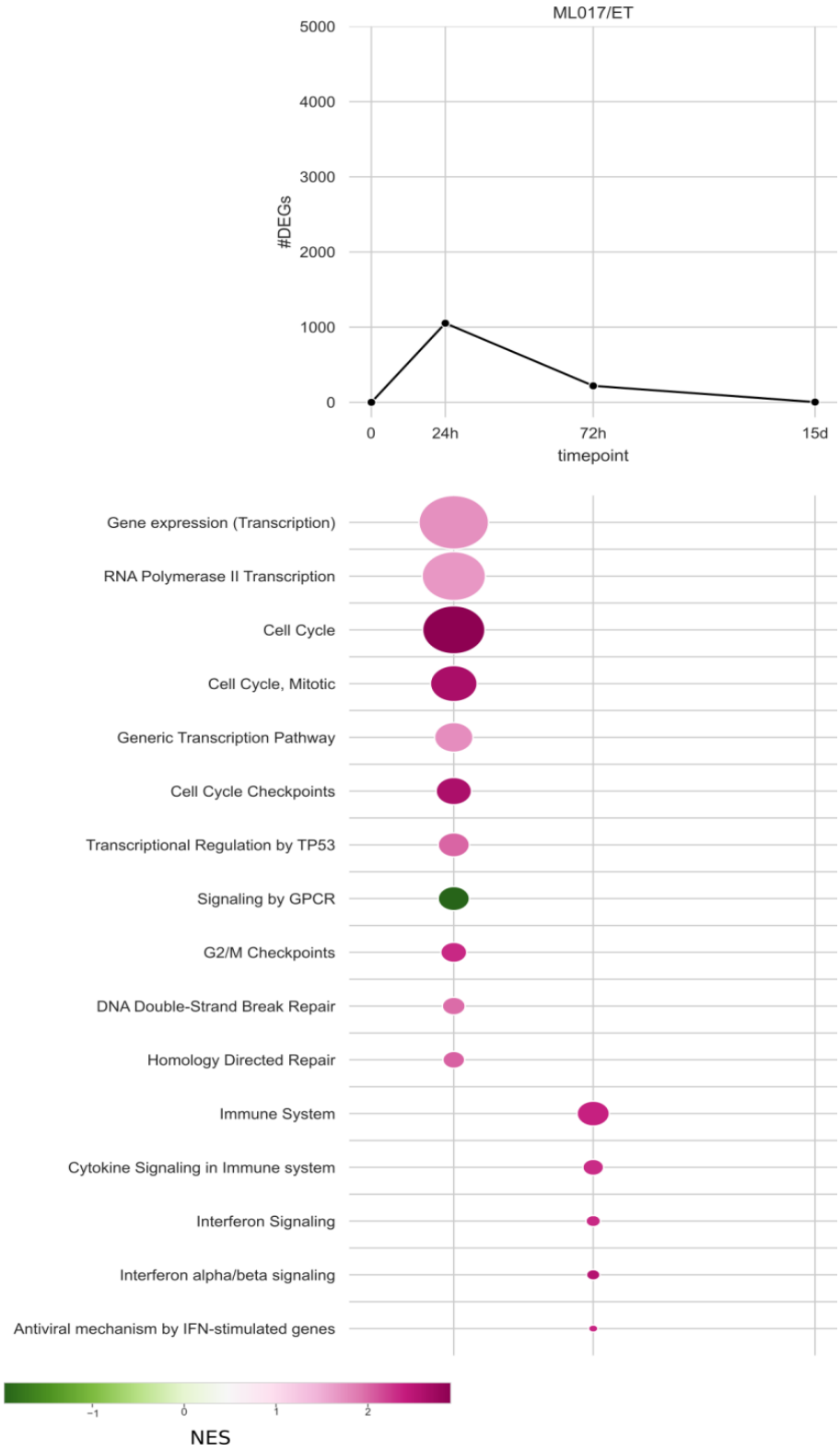

Figure 5.16 Upper panel shows the number of DEGs modulated by trabectedin in ML017/ET model. Control is set to zero. Lower panel shows the most significant pathways from GSEA with Reactome. The colour represents the normalised enrichment score (NES) as indicated in the legend: negative values for negative enrichment (down-regulation), positive values for positive enrichment (up-regulation). Radius of the circles is proportional to the number of genes in the pathways.

Finally, we also investigated the effects of doxorubicin, that is the golden-standard therapy for MLPS.

As shown in *Figure 5.17*, at 24h doxorubicin is able to elicit a transcriptional response in both models, with a greater effect in ML017 (2420 DEGs) than in ML017/ET (484 DEGs). In the responsive model, DEGs were classified in 169 pathways, 7 activated and 162 inhibited. These last were mainly involved in pathways like the Cell Cycle, the DNA Repair, the Rho GTPases and the regulation of cholesterol biosynthesis, while those up-regulated were mainly related to the pathway of TP53 and extracellular matrix remodelling, however with lower enrichments than those obtained with trabectedin treatment.

This evidence shows that at 24h some effects elicited by doxorubicin partially overlap those triggered by trabectedin in a time-dependent manner, even though with different genes and different enrichments. Besides these slight analogies, the two drugs cause overall different transcriptional responses. Finally, as regards to doxorubicin, deregulated genes in ML017/ET participate in the same biological functions as in ML017 suggesting no cross-resistance to this drug.

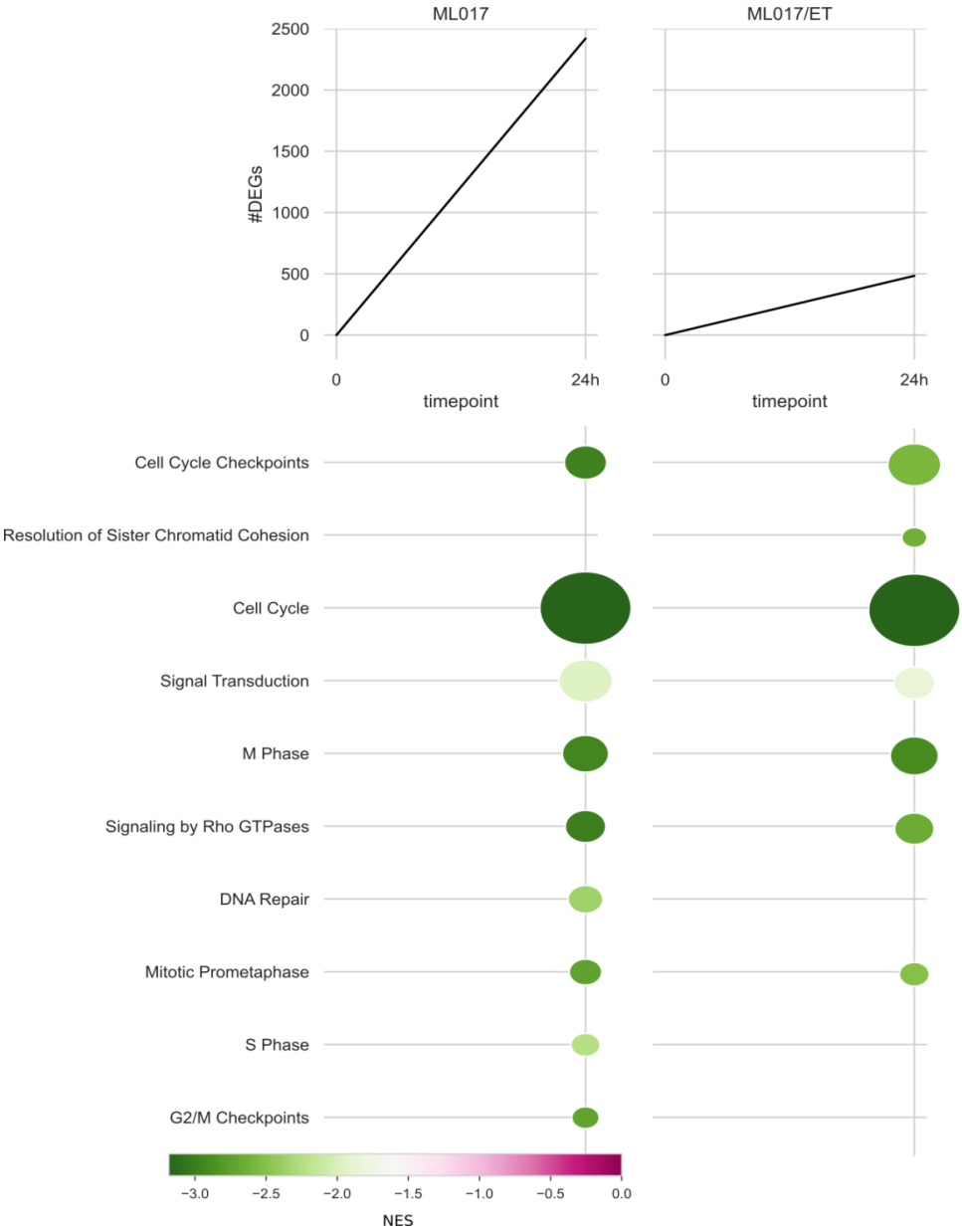

Figure 5.17 Upper panel shows the number of DEGs modulated by doxorubicin in ML017 (on the right) and in ML017/ET (on the left). Control is set to zero. Lower panel shows the most significant pathways from GSEA with Reactome. The colour represents the normalised enrichment score (NES) as indicated in the legend: negative values for negative enrichment (down-regulation), positive values for positive enrichment (up-regulation). Radius of the circles is proportional to the number of genes in the pathways.

5.3.3 DISCOVERING YET UNKNOWN TRANSCRIPT MODULATION

The evidence of the great transcriptional modulation at 15 days after the third dose of trabectedin and the technical availability of a deep read depth allowed us to investigate yet unknown transcriptional mechanisms underlying drug mechanism

of action. To this aim, we customised the Tuxedo protocol described in (Pertea et al., 2016) and further explained in Materials and methods, Chapter 4.

For each sample we aligned sequencing reads with HISAT2 and assembled transcripts with StringTie (*hg19* human genome reference). Then, we compared assemblies to the reference annotation and counted the number of known transcripts, e.g. those perfectly matching genome annotation, and totally unknown transcripts, e.g. mapping to intergenic regions in comparison to the annotation. As reported in *Table 5.10* the number of known transcripts was fairly the same under each condition in both ML017 (mean of 82705.3) and ML017/ET (mean of 83630.47), while as regards to the number of unknown transcripts, the mean number in ET-15 condition in ML017 is 90856.75, greater than the mean of 73558.81 in other conditions of ML017 and 61747.12 in ML017/ET. Thus, M017 responsive model at ET-15 showed the greatest modulation occurring at novel transcripts level.

In order to investigate further the effects of trabectedin treatment, we computed a transcript count table that allowed differential expression analysis. To this aim, we compared each treatment condition to control samples in ML017 and ML017/ET, independently.

First, we normalised transcript counts (*Figure 5.18*) and then we computed unsupervised analysis with the PCA (*Figure 5.19*). We found that, even if with a low variance, ML017 and ML017/ET samples formed two distinct groups suggesting a different transcriptional modulation. However, at the transcript level we could not identify defined distinct groups as in the gene-level analysis, maybe due to bias introduced by this kind of analysis where the number of novel transcripts could have introduced a sort of background noise.

	Sample id	Transcripts matching annotation	Novel transcripts
ML017	1-CTRL-A	86777	70039
	1-CTRL-B	89609	77285
	2-CTRL-A	82243	59002
	2-CTRL-B	89717	71263
		87086,5 ± 3504,18	69397,25 ± 7619,51
	1-ET-24-A	89494	98666
	1-ET-24-B	83593	84152
	2-ET-24-A	91192	68628
	2-ET-24-B	83257	60233
		86884 ± 4056,14	77919,75 ± 17013,77
	1-ET-72-A	87005	75630
	1-ET-72-B	90164	82551
	2-ET-72-A	73950	43170
	2-ET-72-B	83094	68824
		83553,25 ± 7024,95	67543,75 ± 17188,40
	1-ET-15-A	84357	89429
	1-ET-15-B	86517	94074
	2-ET-15-A	69153	89487
	2-ET-15-B	67232	90437
		76814,75 ± 10025,81	90856,75 ± 2194,05
	1-DOXO-24-A	88701	105714
	1-DOXO-24-B	82484	80969
	2-DOXO-24-A	70330	51998
	2-DOXO-24-B	75237	78817
		79188 ± 8071,25	79374,5 ± 21955,22
ML017/ET	1-CTRL-A	86450	67212
	1-CTRL-B	85428	65177
	2-CTRL-A	85906	71693
		85928 ± 511,35	68027,33 ± 3333,64
	1-ET-24-A	77791	36686
	1-ET-24-B	88123	78929
	2-ET-24-A	83950	63999
	2-ET-24-B	85411	66436
		83818,75 ± 4374,66	61512,5 ± 17796,19
	1-ET-72-A	88147	65539
	2-ET-72-A	86897	70353
	2-ET-72-B	74379	53551
		83141 ± 7613,81	63147,67 ± 8652,49
	1-ET-15-A	84341	67288
	1-ET-15-B	79444	46928
	2-ET-15-A	81775	48797
		81853,33 ± 2449,44	54337,67 ± 11254,18
	1-DOXO-24-A	83049	55071
	1-DOXO-24-B	80645	56581
	2-DOXO-24-A	87270	78988
	2-DOXO-24-B	82712	56473
		83419 ± 2778,61	61778,25 ± 11493,76
	merged	194188	3340

Table 5.10 Table shows the number of transcripts matching annotation and the number of novel transcripts identified in ML017 and in ML017/ET.

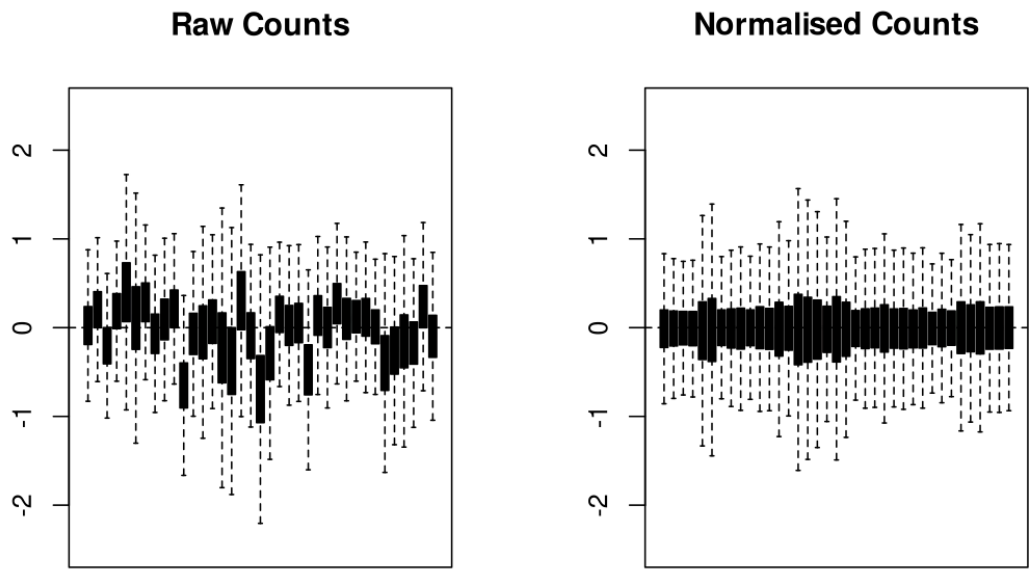

Figure 5.18 Transcript counts before (raw counts) and after normalisation (normalised counts) in the whole cohorts.

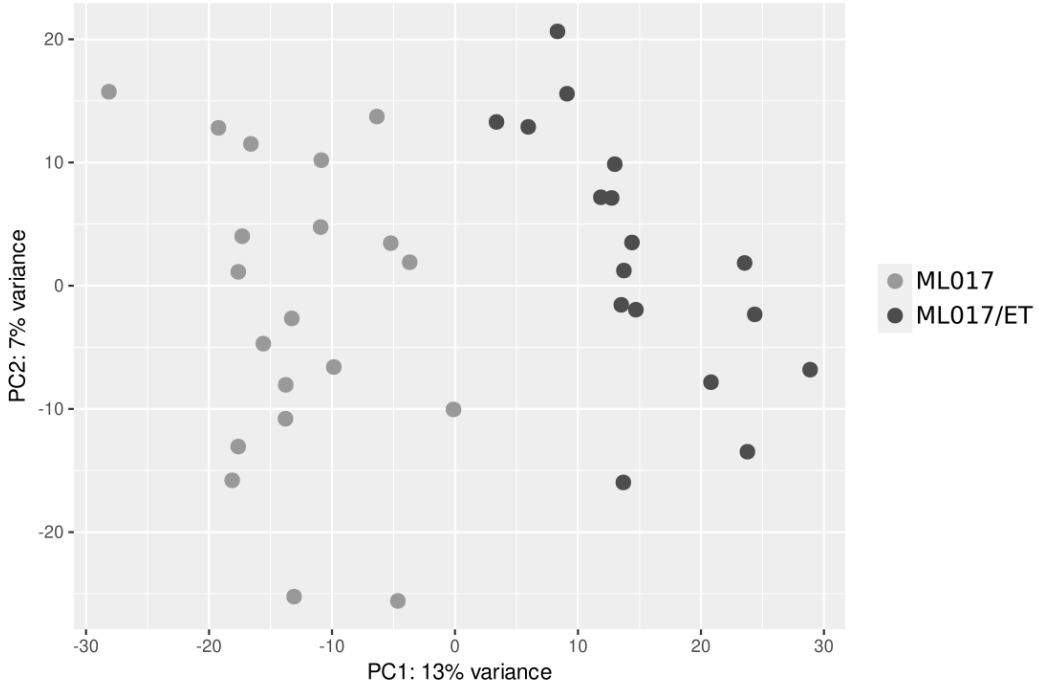

Figure 5.19 PCA of transcripts counts for each sample of ML017 (in grey) and ML017/ET (in black). PC, principal component.

The number of identified differentially expressed transcripts (DETs) is shown in *Figure 5.20*. The number of DETs[5] followed the same trend as at the gene level

[5] Figshare: https://doi.org/10.6084/m9.figshare.17181104.v1,
Zenodo: https://doi.org/10.5281/zenodo.5807695

analysis, with the strongest transcriptional modulation at 15 days after the third dose of trabectedin and while lower responses were present in ML017/ET. The percentage of novel transcripts went from 35.9% in ET-72 to 47.82% in ET-15 in ML017, and from 37% to 53% in both ET-24 and ET-72 in ML017/ET. As expected, known DETs correspond to DEGs that we found modulated with the gene-level analysis. However, as the major transcriptional modulation was again identified at 15 days in ML017, we investigated the possible role of unknown transcripts in the biological functions regulated at this time point.

We compared novel DETs among conditions: we identified and removed from further analysis 48 transcripts since they were coherently modulated by both trabectedin at each time-point and doxorubicin. We retained 2719 DETs specifically related to ET-15, of which 2354 were described as possible new isoforms of already known genes. Most of them were single transcripts related to one gene (*Figure 5.21*, panel A). The remaining 365 novel transcripts were mostly up-regulated (n=280) and equally distributed across chromosomes (*Figure 5.21*, panel B).

The group of 2354 DETs identified as new possible isoforms were analysed with CPAT (see Materials and methods Section 4.8.3.2.2) to assess the probability of being translated into a protein product. Of these, 990 had a high probability to be translated into functional protein products (see CPAT.xlsx[6]).

[6]Figshare: https://doi.org/10.6084/m9.figshare.17181104.v1,
Zenodo: https://doi.org/10.5281/zenodo.5807695

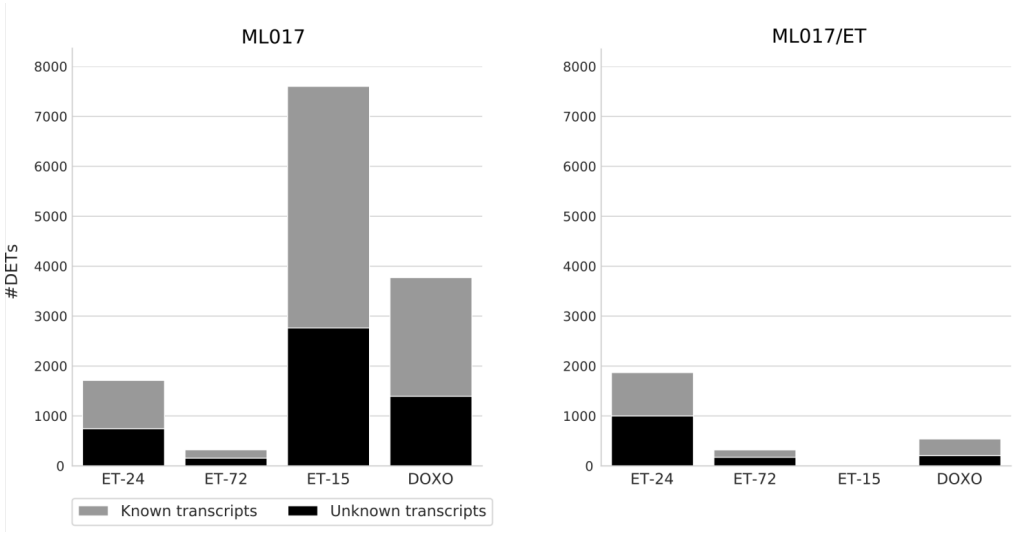

Figure 5.20 Number of DETs in ML017 (left panel) and in ML017/ET (right panel). In grey known transcripts, in black unknown transcripts.

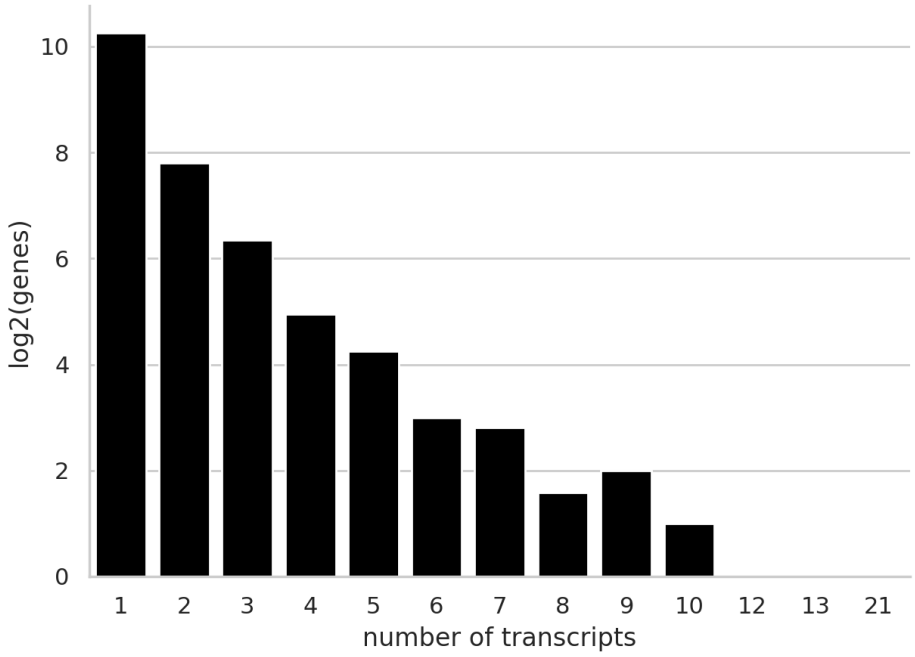

Figure 5.21 Number of transcripts identified for each gene expressed in logarithmic scale .

5.4 STUDY ON THE EFFECTS OF TRABECTEDIN ON FUS-DDIT3 BINDING TO THE DNA

FUS-DDIT3 represents the etiopathological event of MLPS and is considered the leading cause of the block of adipocyte differentiation (see Background Section 2.1.3). According to the results presented so far, trabectedin is able to restore the differentiation process after prolonged treatment. However, this evidence was derived from transcriptional data through which the role of FUS-DDIT3 was not considered.

Therefore, with the aims:

- to characterise the binding sites of FUS-DDIT3 and identify its putative target genes;
- to investigate the role of FUS-DDIT3 in response to trabectedin treatment;

We performed ChIP of DDIT3 followed by sequencing as described in Materials and methods, Chapter 4.

5.4.1 FUS-DDIT3 BINDING SITES AND TARGET GENES

In order to answer the first aim, we selected samples at basal conditions and derived a consensus from the three replicates applying the strict condition of considering a peak as real when present in all replicates. Following this rule, we obtained 30516 binding sites.

In order to assess the reliability of the model, we first made a motif analysis to identify the sequence of the binding site of FUS-DDIT3. As reported in *Table 5.11*, the most over-represented motif was exactly *DDIT3* as reported in the Jaspar database with the ID MA0019.1. The sequence logo is shown in *Figure 5.22*.

The motif was identified in the 97% (n=29449) of the binding regions from the consensus. It is composed of twelve letters of which the most conserved are from the 4^{th} to the 9^{th} position. The position of *DDIT3* motif was almost perfectly centred as shown in *Figure 5.23*. This evidence further confirmed the good quality of the

chromatin immunoprecipitation and the specificity of the antibody used together with the reliability of the analysed binding sites.

Motif name	ID (Jaspar database)	Number of regions	Global p-value	Position	Position p-value
Ddit3::Cebpa	MA0019.1	29449	< 0.001 - Over	[-4,6]	< 0.001
CEBPD	MA0836.1	29449	< 0.001 - Over	[-10,0]	< 0.001
CEBPG	MA0838.1	29449	< 0.001 - Over	[-10,0]	< 0.001
CEBPB	MA0466.2	29449	< 0.001 - Over	[-10,0]	< 0.001
CEBPE	MA0837.1	29449	< 0.001 - Over	[-10,0]	< 0.001
CEBPA	MA0102.3	29449	< 0.001 - Over	[-10,0]	< 0.001

Table 5.11 The six most significant motif results from consensus binding sites in control samples from PscanChIP. Motif name as reported in the Jaspar database (see Materials and methods, Chapter 4).

Figure 5.22 DDIT3::Cebpa sequence logo as reported in the Jaspar database (see Materials and methods, Chapter 4).

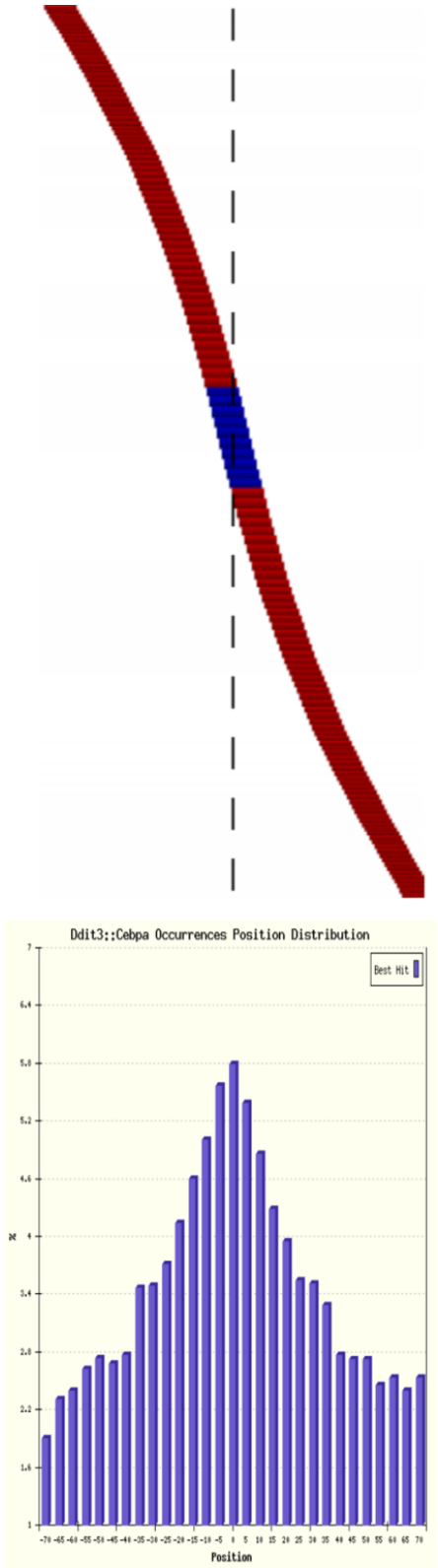

Figure 5.23 Position and histogram of the best motif occurrences of the DDIT3:Cepba regions in ML017 control samples as reported by PScanChIP (see Materials and methods, Chapter 4).

As shown in *Table 5.11*, among the most significant represented motifs we identified also those of the CCAAT/enhancer-binding protein (C/EBP) family of transcription factors, the protein family *DDIT3* belongs to. Interestingly, they were identified in exactly all 29449 regions where DDIT3 motif was detected, however they were not centred, otherwise they were located beside *DDIT3* as reported in the column "Position" of *Table 5.11* and as shown in *Figure 5.23*. These results suggest that the C/EBP family might act as a co-factor of FUS-DDIT3.

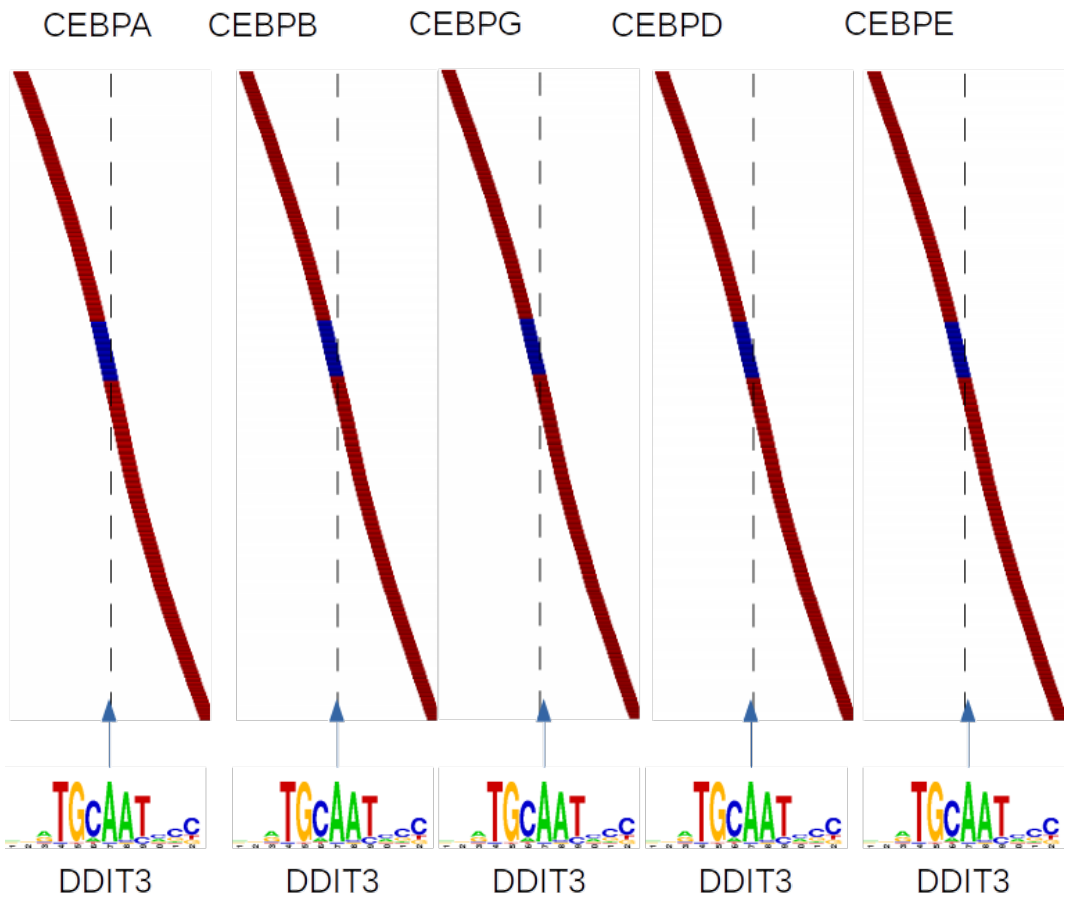

Figure 5.24 Positions of the best occurrences of the C/EBP family of transcription factors, from α to ε in comparison to the centred position of the DDIT3 motif.

Then, in order to investigate the precise genome location of FUS-DDIT3 binding sites and their associated genes, we annotated the binding regions as described in Materials and methods, Chapter 4. As reported in the upper panel of *Figure 5.25*, most of the 30516 binding regions (44%) were annotated to distal regions, e.g. regions that are more than 15 kb far from the nearest transcription start site (TSS);

11% mapped to promoter regions, while the remaining were located in untranslated (UTR) or intronic regions. Coherently, most peaks mapped far more than 10 kb from the TSS as in the lower panel of *Figure 5.25*.

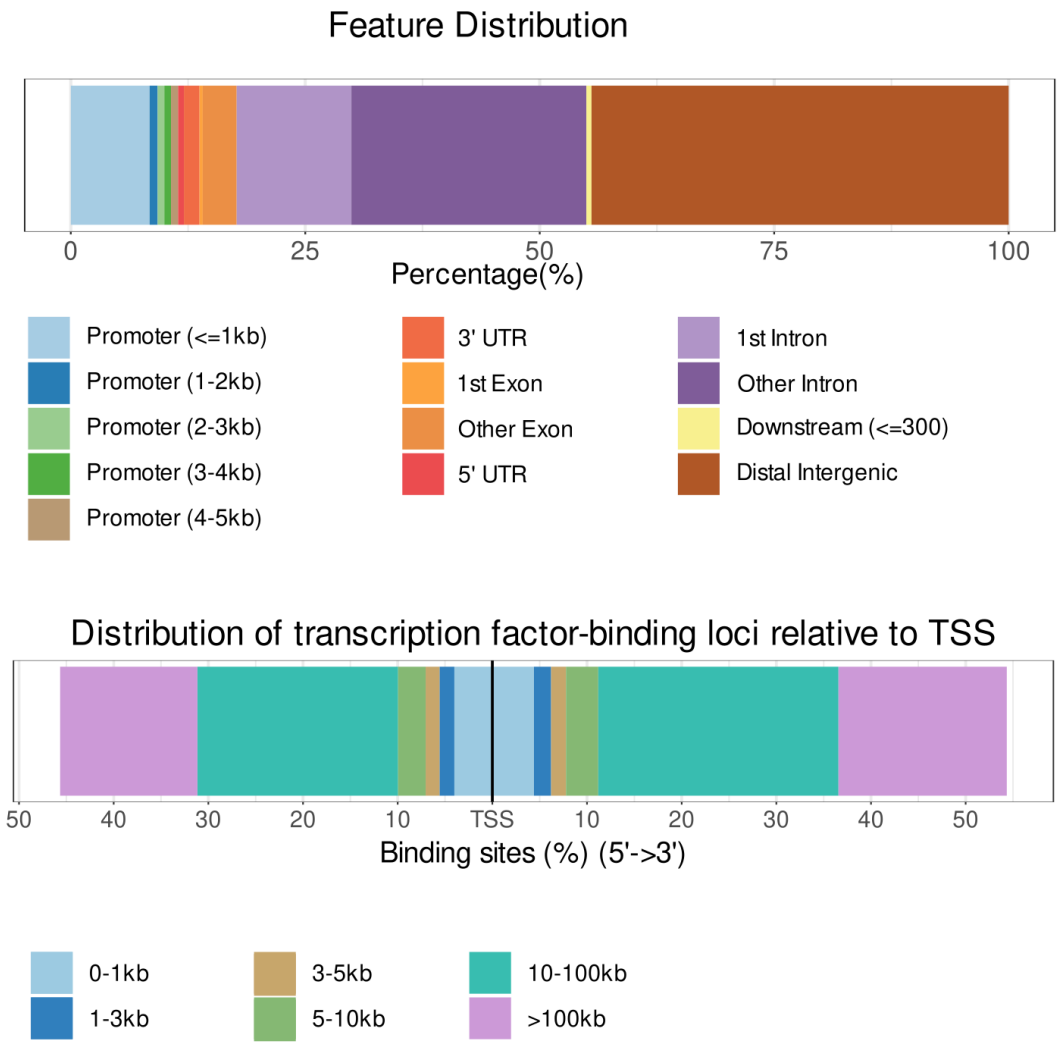

Figure 5.25 Percentage distribution on the genome (upper panel) and distance from the transcription starting site (TSS) of 30516 binding regions in ML017 controls. Colours as in the legends.

Then, we selected the 2981 genes associated with promoter regions and performed an enrichment analysis with the aim to identify biological functions directly related to FUS-DDIT3 targets.

As shown in *Figure 5.26*, FUS-DDIT3 target genes were mainly involved in the biological functions that suggest adipocyte differentiation like adipogenesis, regulation of lipid metabolism by peroxisome, PPARα activates gene expression, white fat cell differentiation, and transcriptional regulation of white adipocyte

differentiation. Besides, we also found the involvement of SMAD2/SMAD3/SMAD4 transcriptional regulation and EGF/EGFR signalling pathway.

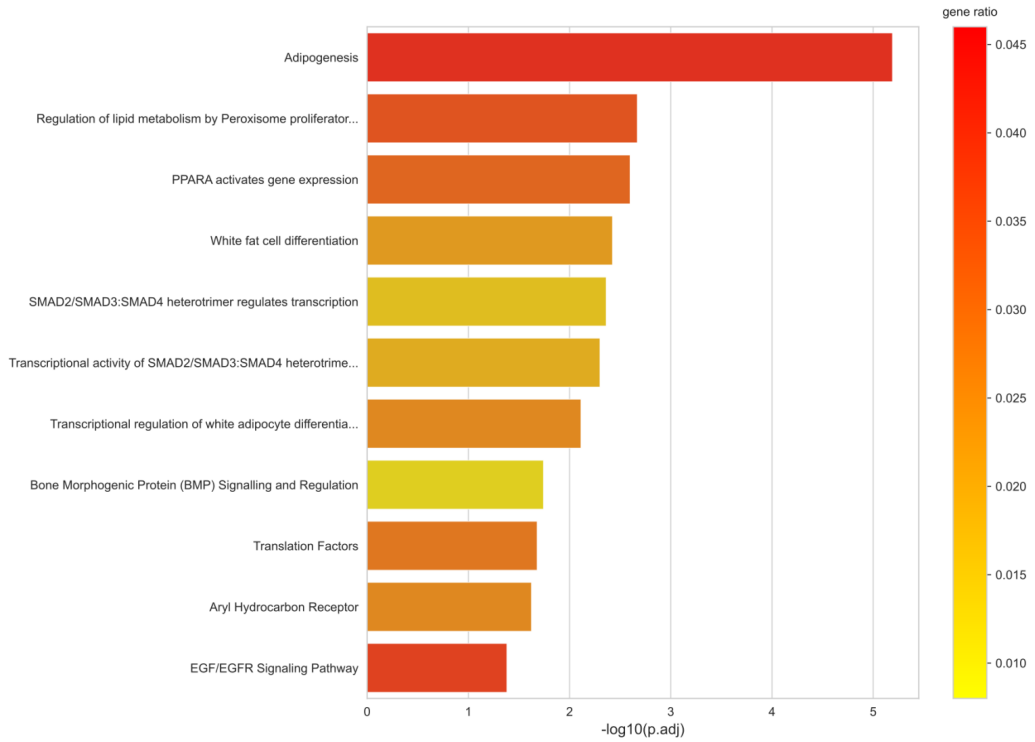

Figure 5.26 Functional enrichment of genes associated with binding sites located on promoter regions based on the Reactome database. On x-axis -log10 of adjust p-values. Colour represents gene ratio as in the legend (see Materials and methods, Chapter 4).

As reported in *Table 5.12*, among the genes involved in these pathways, we identified PPARγ, a key transcription factor in the process of differentiation of mesenchymal cells into completely differentiated adipocytes that is already known to be inhibited by FUS-DDIT3 (Pérez-Mancera et al., 2008). Then, genes like LPL, PLIN1, PLIN2 and ADIPOQ are typical markers of fully differentiated adipocytes, thus their role as target if the chimera suggests a critical role of FUS-DDIT3 towards their transcriptional regulation.

ID	Description	Count	geneID
WP236	Adipogenesis	48	ZMPSTE24, LMNA, MEF2D, RXRG, AGT, KLF6, CISD1, NR1H3, CELF1, STAT6, DDIT3, IGF1, NCOR2, GTF3A, KLF5, HIF1A, NDN, SMAD3, PLIN1, SREBF1, RARA, GADD45B, RETN, EPAS1, CYP26B1, STAT1, KLF7, CREB1, IRS1, TRIB3, PCK1, NRIP1, PPARA, PPARG, KLF15, WWTR1, MBNL1, ADIPOQ, PRLR, MEF2C, NR3C1, EBF1, TWIST1, FZD1, SERPINE1, NAMPT, LPL, PLIN2
WP4149	White fat cell differentiation	16	NR1H3, DDIT3, KLF5, TLE3, NR2F2, SREBF1, RARA, KLF2, CREB1, PPARG, KLF15, MECOM, CTNNA1, NR3C1, EBF1, IRF4
WP1425	Bone Morphogenic Protein (BMP) Signalling and Regulation	8	BMPR1A, TOB1, SMAD4, BMPR2, TOB2, BMPR1B, SMAD1, SMURF1
WP107	Translation Factors	20	EIF4G3, EIF3I, EIF4EBP2, EIF3A, EIF4B, EIF2S1, EIF5, EEF2K, EIF1, EIF3G, EIF2B4, EIF2S2, EIF3D, EIF4G1, EIF4E, PAIP1, ETF1, EEF1A1, EIF3B, EIF2S3
WP2586	Aryl Hydrocarbon Receptor	18	ARNT, PTGS2, KLF6, NCOR2, MAP2K1, NQO1, NFE2L2, NRIP1, NFKB1, VEGFA, GCLC, NCOA7, PLAGL1, ESR1, EGFR, CD36, LPL, MYC
WP437	EGF/EGFR Signaling Pathway	43	ERRFI1, JAK1, VAV3, MEF2D, PIK3C2B, INPPL1, GAB2, EPS8, ATXN2, PEBP1, SPRY2, SOS2, FOS, NEDD4, MAP2K1, MAP2K5, MAPK7, AP2B1, MAP3K3, AP2S1, STAT1, CREB1, NCOA3, CBLB, NCK1, PLSCR1, PLD1, AP2M1, TNK2, PIK3R1, MEF2C, MAPK14, GJA1, MAP3K4, TWIST1, RALA, EGFR, PTPN12, ASAP1, JAK2, STXBP1, ABL1, SH3KBP1
R-HSA-400206	Regulation of lipid metabolism by Peroxisome proliferator-activated receptor alpha (PPARalpha)	37	NFYC, ARNT, G0S2, AGT, ARNTL, NR1H3, GLIPR1, MED13L, NCOR2, CDK8, ARNT2, CHD9, SREBF1, NR1D1, MED13, ACOX1, ANGPTL4, MED26, SIN3B, SLC27A1, MED25, NPAS2, TRIB3, NCOA6, NCOA3, PPARA, PPARG, TBL1XR1, HMGCR, PPARGC1B, CCNC, TIAM2, CD36, ABCB4, FDFT1, PLIN2, TBL1X
R-HSA-1989781	PPARA activates gene expression	36	NFYC, ARNT, G0S2, AGT, ARNTL, NR1H3, GLIPR1, MED13L, NCOR2, CDK8, ARNT2, CHD9, SREBF1, NR1D1, MED13, ACOX1, ANGPTL4, MED26, SLC27A1, MED25, NPAS2, TRIB3, NCOA6, NCOA3, PPARA, PPARG, TBL1XR1, HMGCR, PPARGC1B, CCNC, TIAM2, CD36, ABCB4, FDFT1, PLIN2, TBL1X
R-HSA-2173796	SMAD2/SMAD3:SMAD4 heterotrimer regulates transcription	15	CDK8, SNW1, RNF111, SMAD3, UBB, TGIF1, SMAD7, SMAD4, UBA52, TGIF2, WWTR1, CCNC, SERPINE1, MYC, CDKN2B
R-HSA-2173793	Transcriptional activity of SMAD2/SMAD3:SMAD4 heterotrimer	18	NCOR2, CDK8, SNW1, RNF111, SMAD3, UBB, TGIF1, SMAD7, SMAD4, NEDD4L, UBA52, TGIF2, WWTR1, SKIL, CCNC, SERPINE1, MYC, CDKN2B
R-HSA-381340	Transcriptional regulation of white adipocyte differentiation	27	CDK4, MED13L, NCOR2, CDK8, KLF5, PLIN1, NR2F2, CHD9, SREBF1, MED13, ANGPTL4, MED26, MED25, NCOA6, NCOA3, PCK1, PPARA, PPARG, TBL1XR1, ADIPOQ, NFKB1, EBF1, CCNC, CD36, LPL, FABP4, TBL1X

Table 5.12 Significant pathways and genes in pathways for enrichment analysis of promoter regions referred to control samples of ML017 based on Reactome and Wikipathways databases.

The transcriptional modulation of genes is not strictly related to transcription factors bound to promoter regions, indeed they can play their regulatory role even by attaching to regions of the DNA very far from the TSS. Therefore we considered the whole set of 9097 putative target genes and made an enrichment analysis.

As shown in *Figure 5.27* and *Figure 5.28*, apart from the previously discussed pathways associated with promoters, we found also the involvement of tyrosine kinases, the signalling of VEGF, the extracellular matrix organisation and the TGF-β pathways.

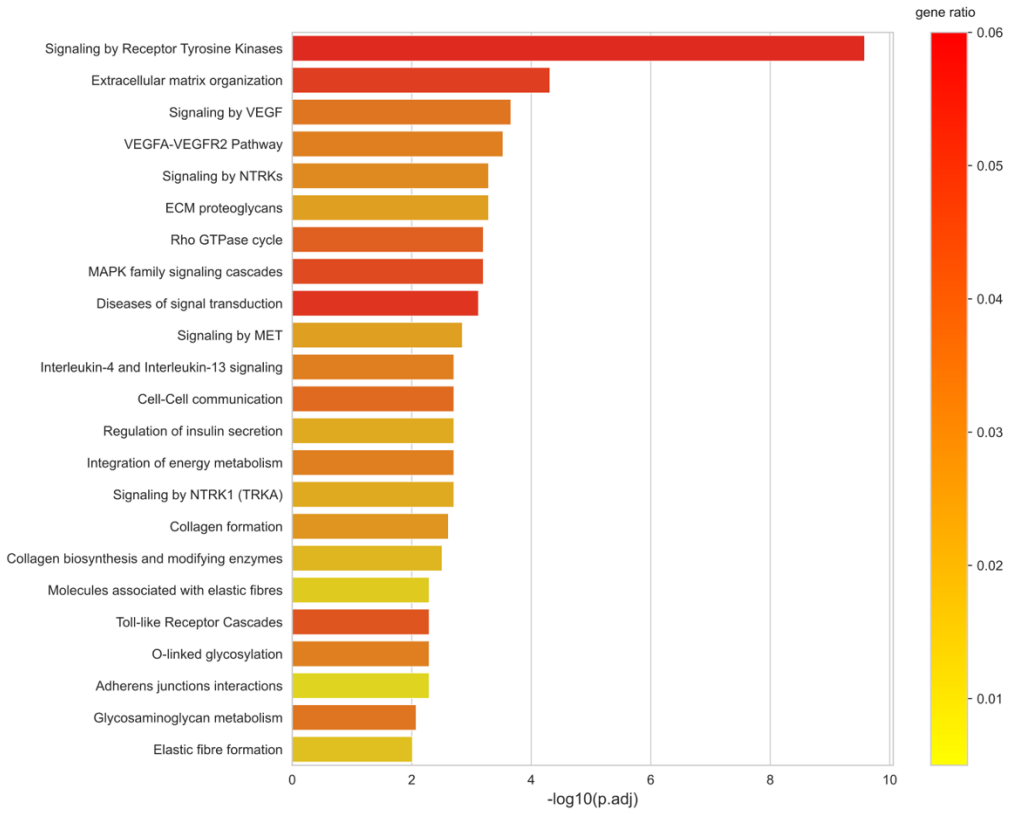

Figure 5.27 Functional enrichment of genes associated with binding sites of FUS-DDIT3 in consensus regions of controls samples of ML017 based on the Reactome database. On x-axis -log10 of adjust p-values. Colour represents gene ratio as in the legend (see Materials and methods, Chapter 4).

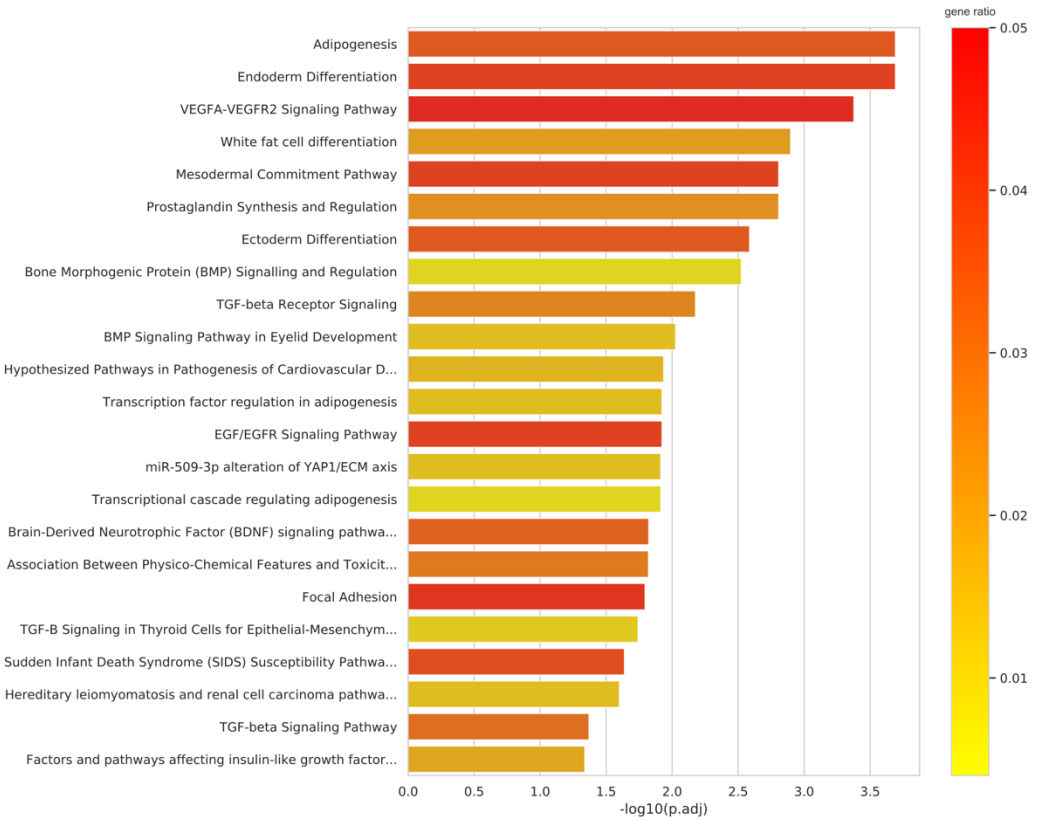

Figure 5.28 Functional enrichment of genes associated with binding sites of FUS-DDIT3 in consensus regions of controls samples of ML017 based on the Wikipathways database. On x-axis -log10 of adjust p-values. represents gene ratio as in the legend (see Materials and methods, Chapter 4).

In conclusion, by targeting *DDIT3* through a ChIP-Seq approach:

1. we identified FUS-DDIT3 binding sites at the genome-wide level and its putative target genes;
2. we confirmed FUS-DDIT3 binding sites correspond to the canonical DDIT3 binding motif;
3. we showed that members of the (C/EBP) family of transcription factors might act as cofactors of FUS-DDIT3;
4. we demonstrated that FUS-DDIT3 target genes are involved in extracellular matrix remodelling and differentiation pathways.

5.4.2 *STUDY OF THE INTERACTION OF FUS-DDIT3 WITH TRABECTEDIN*

In order to understand the role of trabectedin treatment on FUS-DDIT3 in MLPS, we performed ChIP-Seq at 24h and 72h after the first dose of treatment and 15 days after the third dose of treatment with trabectedin. Since ChIP-Seq experiments represent a photograph of the genome-wide binding events of a specific transcription factor, they may be affected by technical biases like non-specific binding or low efficacy of the antibody, we used biological replicates for each condition to determine relevant biological events. To note, the presence of FUS-DDIT3 binding at all time points, *e.g.* 24h and 72h after the first dose and 15 days after the third dose, was first confirmed by western blot done in the Laboratory of Oncology of the Mario Negri Institute (data not shown here).

First, we computed the binding motif for samples under each condition and confirmed the DDIT3 binding motif as in *Figure 5.22*. Then, we evaluated sample distribution through a PCA as shown in *Figure 5.29*: control samples (basal) were similar to those at ET-72, samples at ET-24 showed a high variability, while samples at ET-15 formed an independent group. This first insight suggests a different binding pattern at the late time point of treatment.

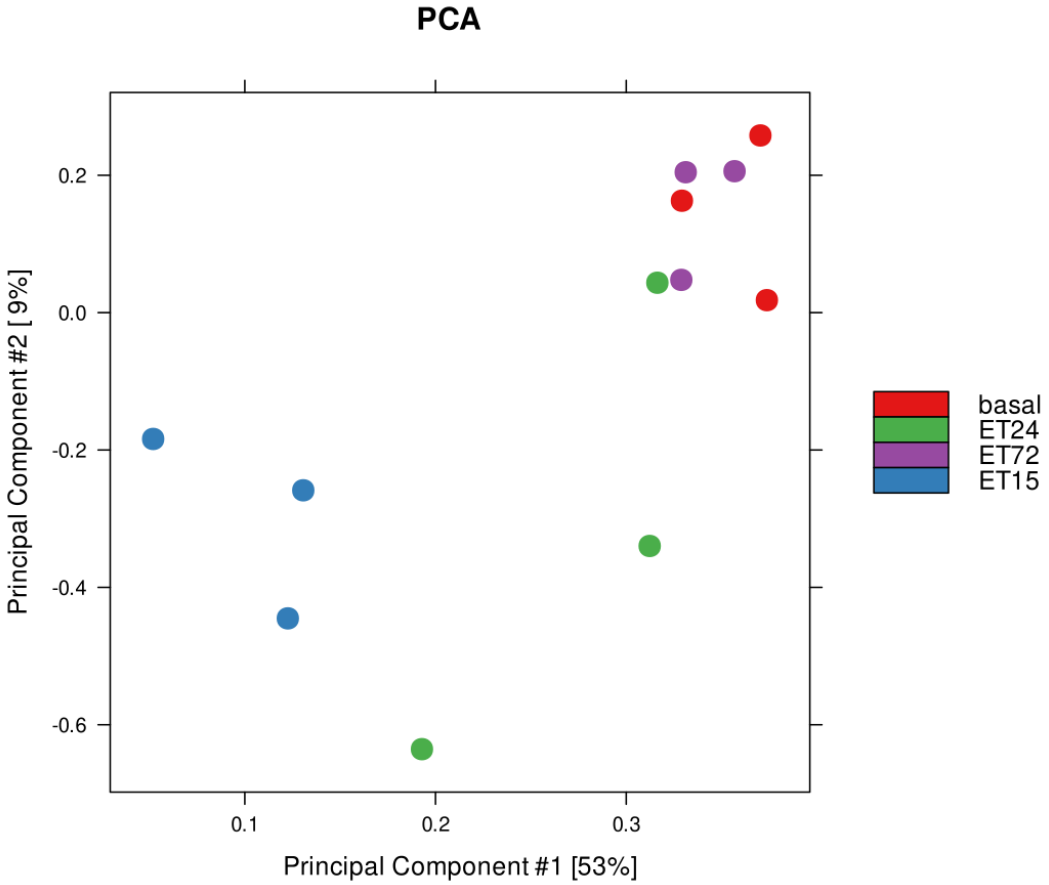

Figure 5.29 PCA of ChIP-Seq ML017 samples after derivation of the consensus among binding sites. Colours as in the legend.

In order to investigate the differences in the binding activity of FUS-DDIT3 under trabectedin treatment we performed a differential binding analysis as described In Materials and methods, Chapter 4.

First, we computed a consensus among all samples identifying 117643 consensus binding sites, then we made a differential binding analysis by comparing all conditions in pairs. As shown in *Table 5.13*, we identified differentially bound peaks (DBPs) only between ET-15 *versus* control or *versus* ET-72. Since ET-72 did not showed differences with controls, and the 31778 DBPs of ET-15 *versus* control were overlapping with those of ET-15 *versus* ET-72, we focused directly on these 39581 DBPs (*Figure 5.30*).

	CTRL	ET-24	ET-72	ET-15
CTRL	-	n.s.	n.s.	31778
ET-24		-	n.s.	n.s
ET-72			-	39581
ET-15				-

Table 5.13 Number of DBPs per each couple of compared conditions as calculated with DiffBind.

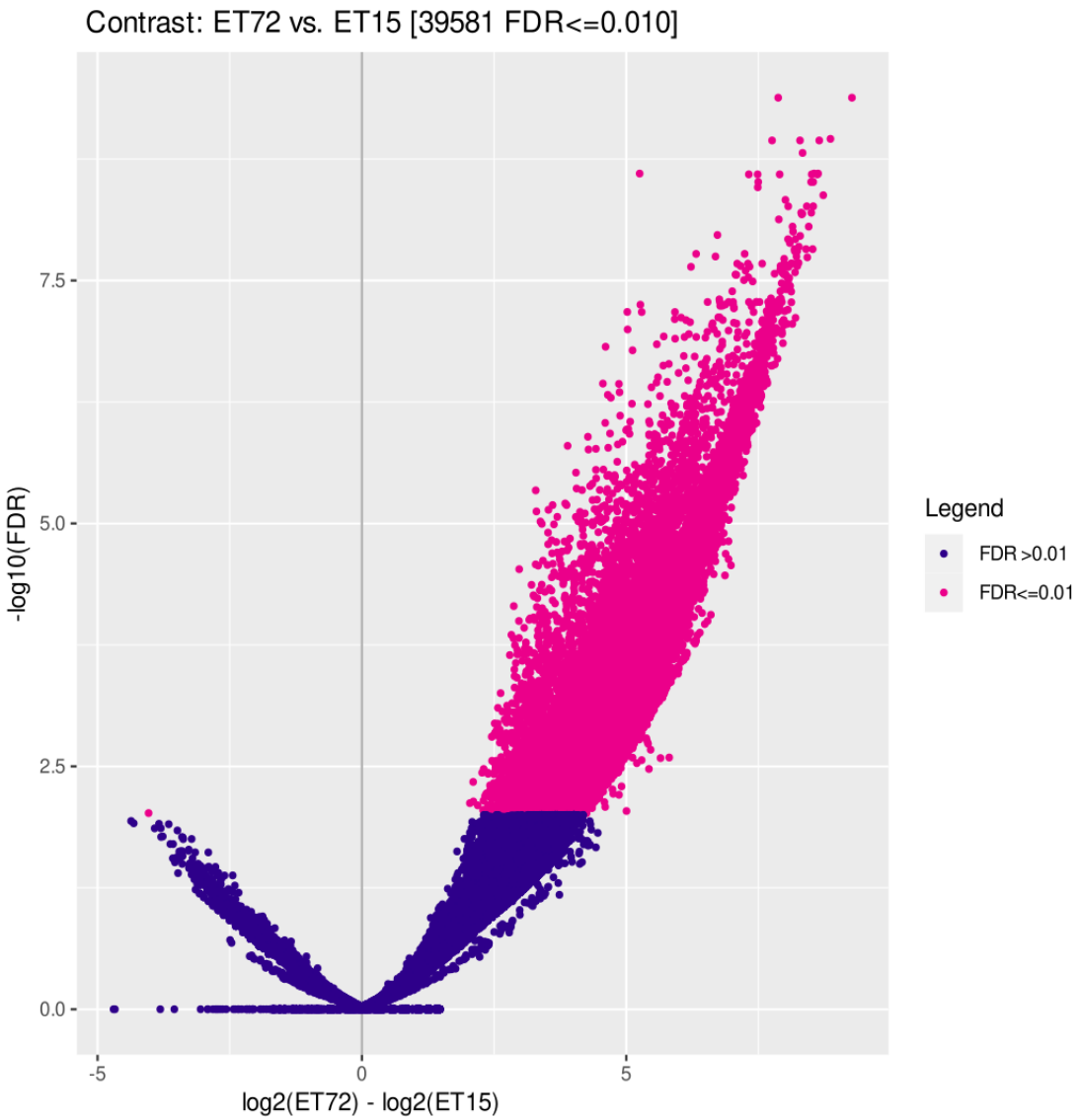

Figure 5.30 Volcano plot showing DBPs in the comparison between ET-72 and ET-15. FDR, False Discovery Rate. In blue non-significant DBPs (FDR>0.01), in pink significant DBPs (FDR<=0.01).

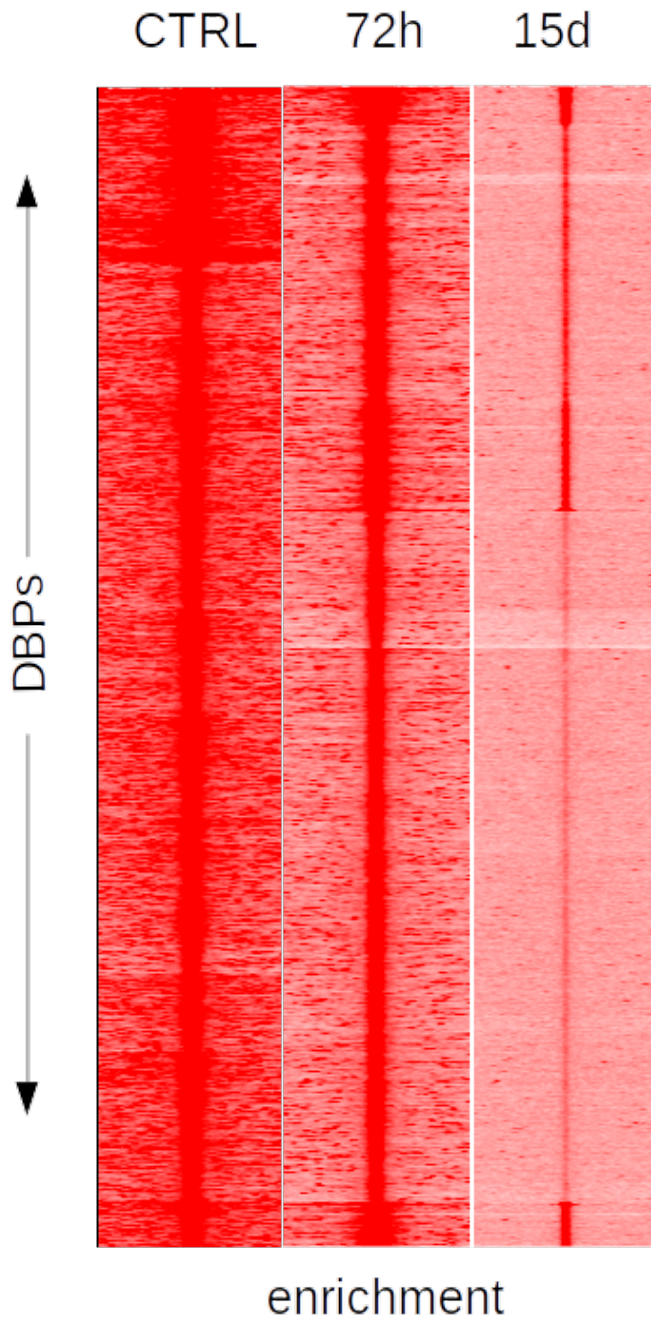

Figure 5.31 Read enrichment mapping onto DBPs evaluated through SeqMINER on merged sample replicates at basal conditions (CTRL), 72h and 15 days

Among the 39581 DBPs, 39580 were either no longer bound or less bound to the DNA at 15 days after the third dose of trabectedin, meaning that at late time points trabectedin was able to displace FUS-DDIT3 from its binding sites. The lower or absent enrichment in these regions is evident in *Figure 5.31*. Therefore, we investigated the nature of these DBPs. First, we annotated them and showed that

almost 43% mapped to distal intergenic regions and far from the TSS, while less than 10% mapped to promoter regions as shown in *Figure 5.32*.

The total 39581 differential regions were associated with 9856 unique genes. Then, we investigated the biological functions in which these genes were implicated. Again, we first selected genes in promoter regions.

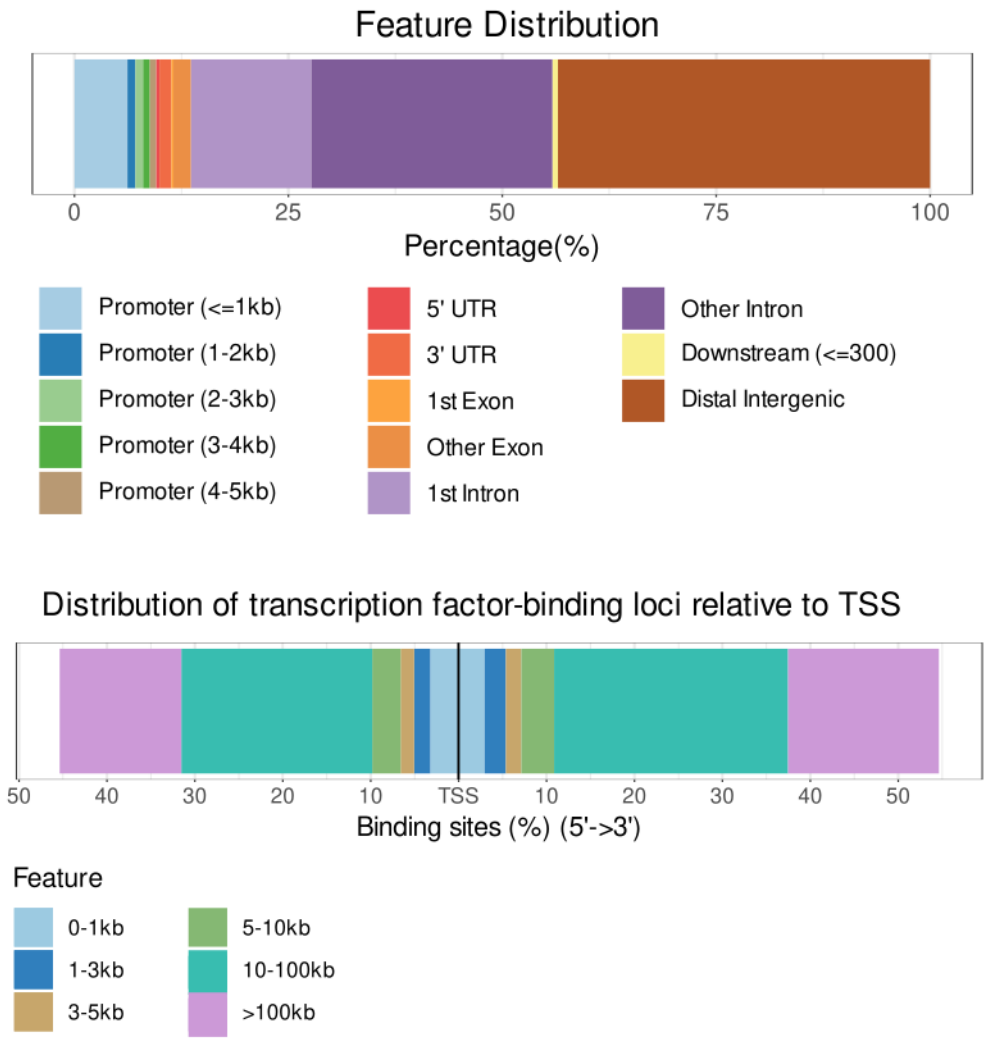

Figure 5.32 Percentage distribution on the genome (upper panel) and distance from the transcription starting site (TSS) of DBPs in the comparison between ET-15 and ET-72. Colours as in the legends.

Among DBPs associated with promoter regions, we found those that were implicated in the adipogenesis and differentiation pathways as shown in *Figure 5.33* and that were previously observed at basal conditions (see Section 5.4.1 of this Chapter).

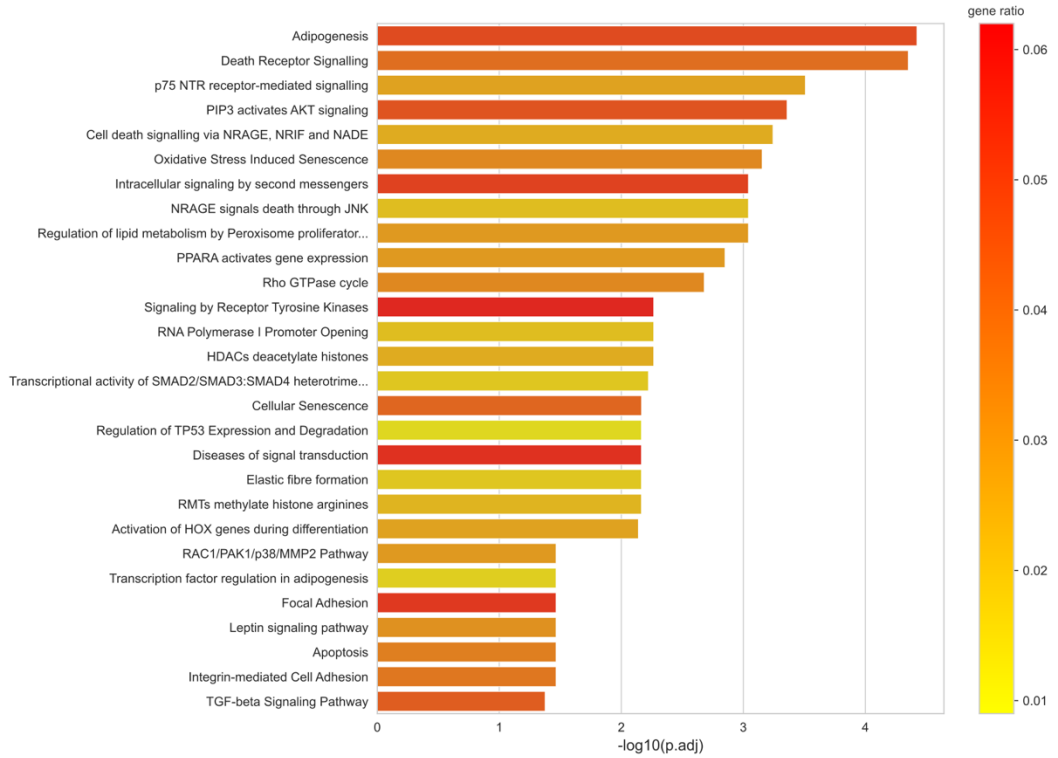

Figure 5.33 Functional enrichment of genes associated with DBPs in the comparison between ET-15 and ET-72 mapping on promoter regions of ML017 based on the Reactome database. On x-axis -log10 of adjust p-values. Colour represents gene ratio as in the legend. (see Materials and methods, Chapter 4).

Then, a further functional analysis on the whole DBPs set (*Figure 5.34*, *Figure 5.35*) showed that most of the target genes affected by the treatment with trabectedin were involved in the main functional processes shown at basal condition (see Section 5.4.1 of this Chapter), like extracellular matrix remodelling and the signalling of VEGF.

In conclusion, we showed that the effect of trabectedin on FUS-DDIT3 target genes is explicated at 15 days, the late time point at which we showed that the drug is able to displace or decrease the binding of FUS-DDIT3 on the DNA. Interestingly, the event is related to key pathways like extracellular matrix remodelling and adipocyte differentiation, thus suggesting a role of trabectedin in restoring the differentiation program through the direct interaction with the chimera.

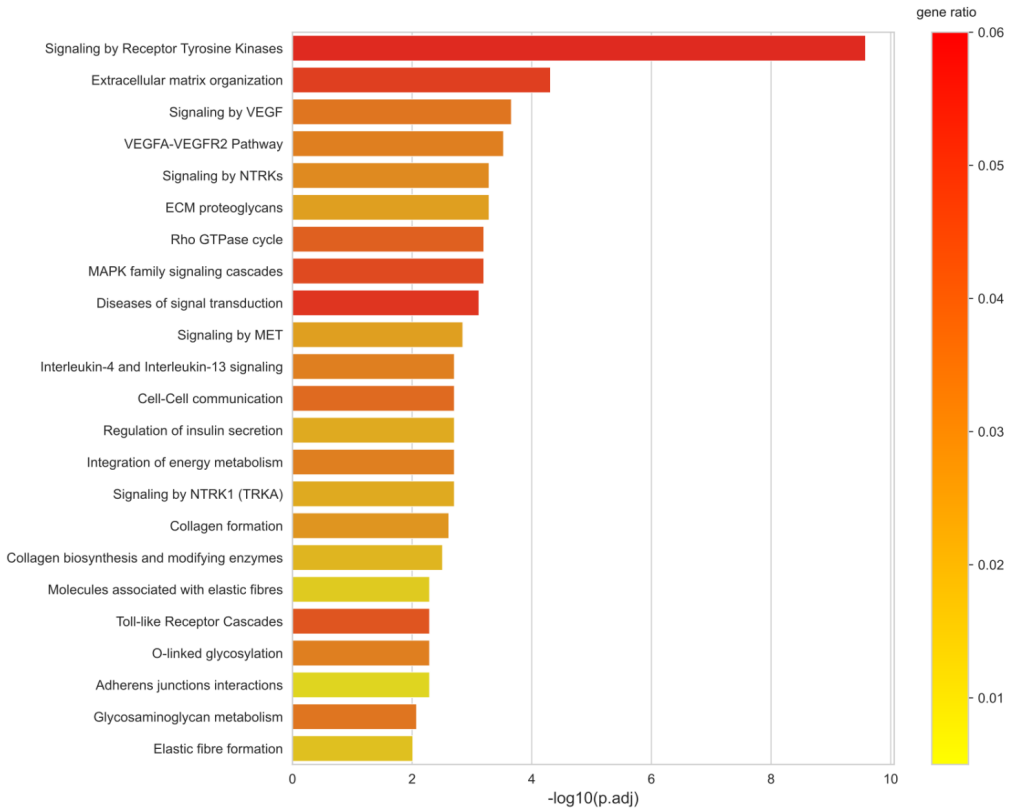

Figure 5.34 Functional enrichment of genes associated with DBPs in the comparison between ET-15 and ET-72 of ML017 based on the Reactome database. On x-axis -log10 of adjust p-values. Colour represents gene ratio as in the legend (see Materials and methods, Chapter 4).

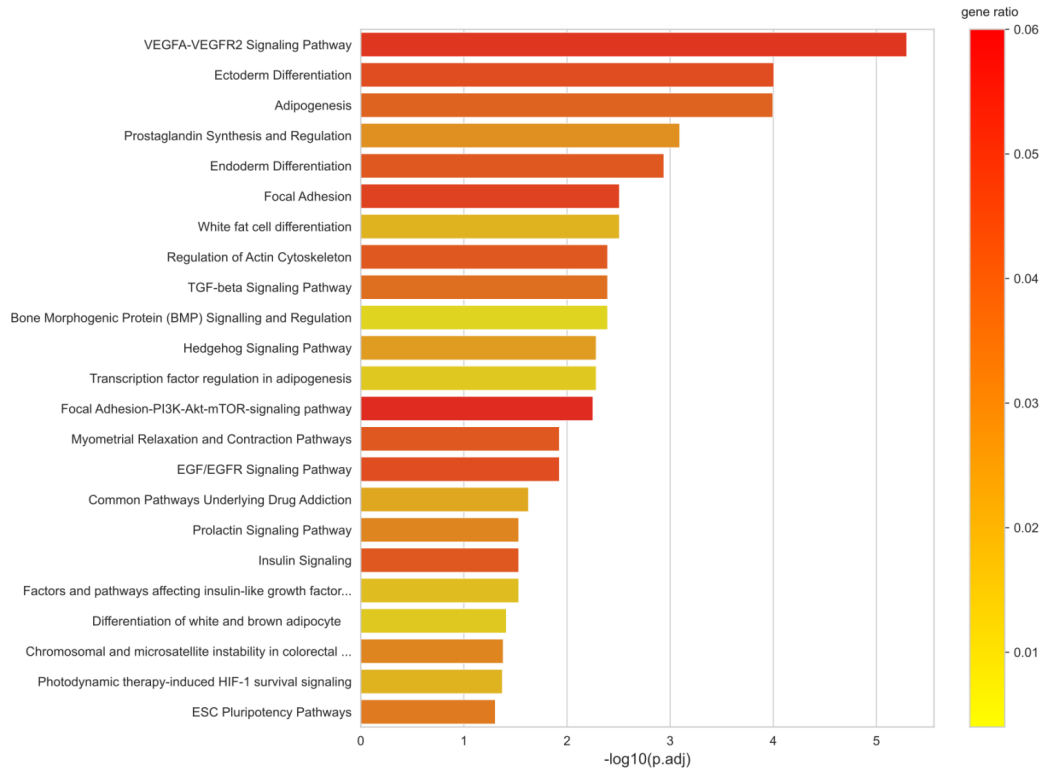

Figure 5.35 Functional enrichment of genes associated with DBPs in the comparison between ET-15 and ET-72 of ML017 based on the Wikipathways database. On x-axis -log10 of adjust p-values. Colour represents gene ratio as in the legend (see Materials and methods, Chapter 4).

5.4.3 Early response to trabectedin is cytotoxic and independent on FUS-DDIT3

ChIP-Seq is a useful tool to investigate the activity of transcription factors starting from the identification of their binding sites to the prediction of their target genes, however it gives no information on the transcriptional regulation. Thus, to fill in this gap, many approaches have been developed among which data integration with transcriptional expression data (e.g. RNA-Seq) or chromatin immunoprecipitation of histone modifications. The first provides a direct indication of the gene regulation when two or more conditions are compared, the latter represents a biological indication of the transcriptional activity. Indeed, H3K4me1 and H3K4me3 correspond to sites of active transcription, while H3K27Ac is a marker of active enhancer. The identification of genomic regions characterised by histone modifications together with specific sites of binding for transcription factors represent a valid indication of the transcriptional activity.

At this point, our main aim was to determine at what extent trabectedin treatment might influence gene transcription through the interaction with FUS-DDIT3 chimera. To pursue this issue, we used an integrated approach to combine transcription factor binding sites, histone marks and transcriptional data.

In our study, we decided to use the histone mark H3K4me3 in order to identify active regions of transcription. However, the unavailability of biological replicates of H3K4me3 mark at basal and further conditions did not allow us to investigate the transcriptional activity at CTRL, ET-72 and ET-15 conditions of ML017. Instead, biological replicates at 24 hours allowed us to identify genes associated with region of active transcriptional activity after one exposure to trabectedin.

First, we derived a consensus peakset from the four available replicates of H3K4me3 at ET-24 identifying 15576 regions. We annotated those regions and, as expected, most of them (84%) mapped on promoters (*Figure 5.36*). Then, we made an enrichment analysis on the 11819 genes associated with these promoter regions in order to identify biological activated functions.

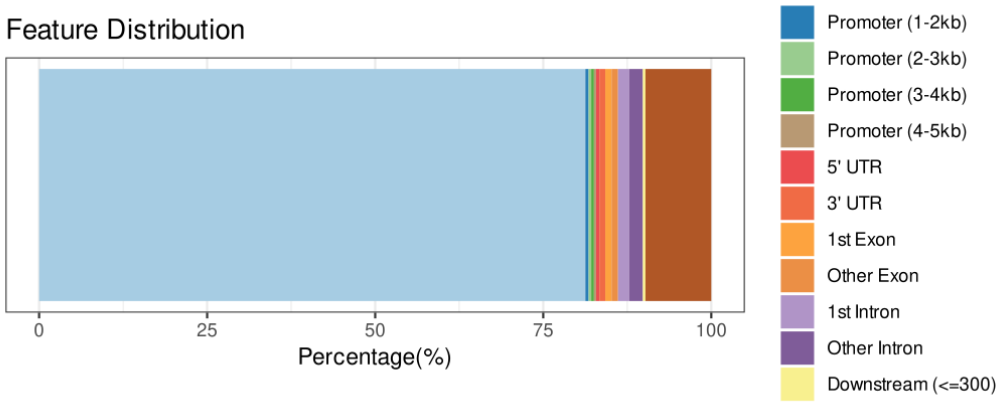

Figure 5.36 Percentage distribution on the genome of H3K4me3 binding regions in ML017 at 24h. Colours as reported in the legend.

A selection of the most significant pathways from the Reactome database is shown in *Figure 5.37*. We found the transcriptional regulation by TP53, the activation of the cell cycle, the mechanisms of apoptosis and DNA repair. We also found the

regulation of PTEN, a gene that was found impaired in MLPS as described in 5.2.3. Then, in order to achieve a complete overview of the transcriptional regulation in tumour basal conditions of MLPS, we compared these results to those from RNA-Seq analysis in Results, paragraph 5.3.2. By overlapping the pathways from H3K4me3 marks and from DEGs at 24h, we confirmed the early cytotoxic response to trabectedin with the activation of regulation of TP53, the involvement of DNA replication and cell cycle, the apoptotic response as shown in *Table 5.14*. Instead, the state of activation or inhibition of Antigen processing: Ubiquitination & Proteasome degradation and Class I MHC mediated antigen processing & presentation are not clear, since they resulted inhibited from RNA-Seq analysis.

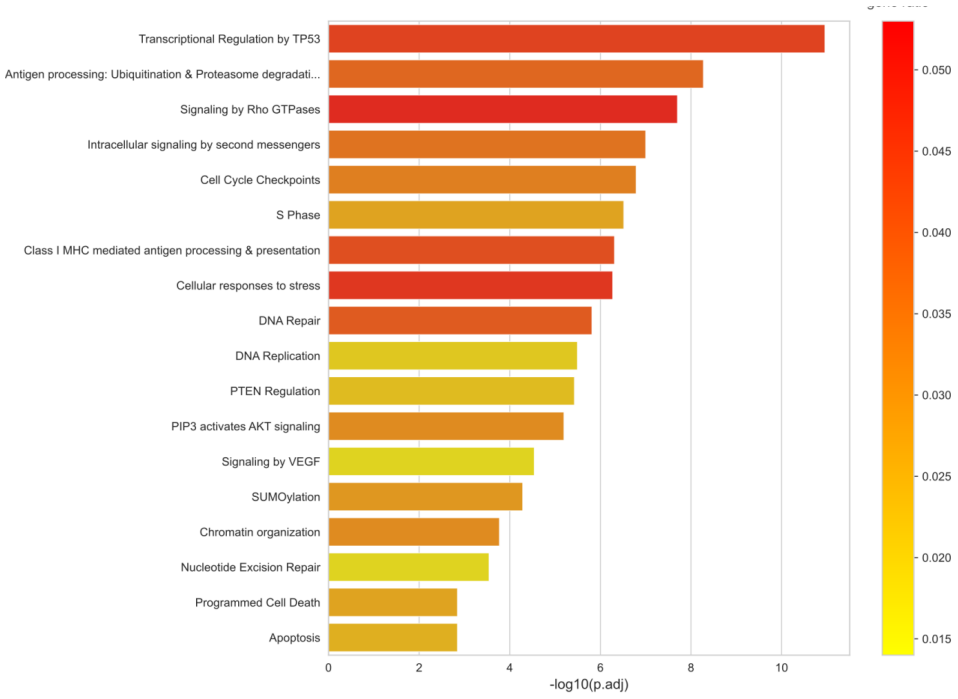

Figure 5.37 Functional enrichment of 11819 genes associated with promoter regions from H3K4me3 ChIP-Seq based on the Reactome database. On x-axis -log10 of adjust p-values. Colour represents gene ratio as in the legend (see Materials and methods, Chapter 4).

ID	Description	N Genes	geneID
R-HSA-3700989	Transcriptional Regulation by TP53	14	ARID3A,CDKN1A,TNFRSF10C,TNFRSF10D,ZNF385A,PIDD1,E2F1,TP73,PLK3,MDM2,GADD45A,E2F7,PMAIP1,DDB2
R-HSA-69306	DNA Replication	6	POLA2,LIG1,CDT1,E2F1,CDC6,E2F2

ID	Description	N Genes	geneID
R-HSA-5663202	Diseases of signal transduction	6	CDKN1A,FGF2,MDM2,DLL4,HBEGF,MAMLD1
R-HSA-69242	S Phase	6	POLA2,LIG1,CDT1,CDKN1A,E2F1,CDC6
R-HSA-69002	DNA Replication Pre-Initiation	5	POLA2,CDT1,E2F1,CDC6,E2F2
R-HSA-1257604	PIP3 activates AKT signaling	4	MDM2,HBEGF,CDKN1A,FGF2
R-HSA-9006925	Intracellular signaling by second messengers	4	MDM2,HBEGF,CDKN1A,FGF2
R-HSA-5357801	Programmed Cell Death	3	TP73,E2F1,PMAIP1
R-HSA-109581	Apoptosis	3	TP73,E2F1,PMAIP1
R-HSA-983169	Class I MHC mediated antigen processing & presentation	1	CDC20
R-HSA-983168	Antigen processing: Ubiquitination & Proteasome degradation	1	CDC20

Table 5.14 Common pathways and genes between RNA-Seq analysis on DEGs at 24h and ChIP-Seq of H3K4me3 at the same time point according to the Reactome database.

As shown from pathways in *Figure 5.37* and in *Table 5.14*, none of them is related to FUS-DDIT3 binding sites identified at CTRL (see Section 5.4.1 of this Chapter), thus suggesting that the first response to drug treatment is cytotoxic and independent from FUS-DDIT3.

5.4.4 TRABECTEDIN DISPLACES FUS-DDIT3 FROM THE DNA AND RESTORES ADIPOCYTE DIFFERENTIATION

As previously explained, the unavailability of histone marks at different times of treatment did not allow us to investigate transcriptionally active regions at different time of treatment. However, since we found that most FUS-DDIT3 binding sites changed at 15 days after trabectedin treatment (see Results Section 5.4.2) and we had transcriptional data at the same time point, we integrated ChIP-Seq and RNA-Seq in order to study the transcriptional modulation of FUS-DDIT3 target genes.

First, we compared the DEGs found at 15 days after the third dose (N=4883, see Results, Section 5.3.2) with the genes associated with the DBPs (N=9856, see Results, Section 5.4.2). As shown in the Venn diagram in *Figure **5.38***, we found 2292 genes in common.

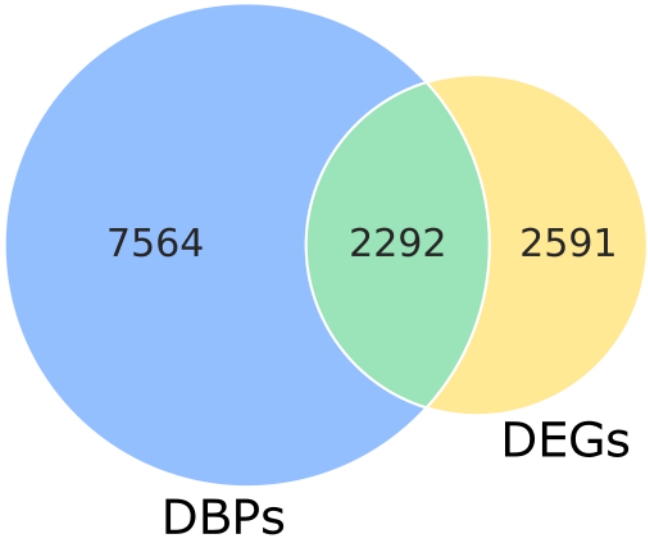

Figure 5.38 Venn diagram showing the number of common and private genes in the comparison between DEGs at 15 days after the third dose from RNA-Seq analysis and the DBPs from the ChIP-Seq analysis.

Then, we compared pathways identified by the transcriptional analysis (see Results, Section 5.3.2) with those associated with DBPs at 15 days (see Results, Section 5.4.2). As shown in *Table 5.15*, we identified 11 overlapping pathways and genes, among which extracellular matrix organisation, collagen and elastic fibre formation and signalling by receptor tyrosine kinases.

Since these pathways were related to target genes of FUS-DDIT3 and they were also transcriptionally modulated, this data might suggest a possible role of trabectedin in displacing FUS-DDIT3 chimera from its targets with a consequent modulation of their transcriptional activity that is directly related to differentiation programs.

Pathway ID	Description	N	Common Genes
R-HSA-1474244	Extracellular matrix organization	65	PHYKPL,FBLN5,COL6A6,ADAMTS5,EFEMP1,COL6A2,MMP15,HTRA1,TNC,COL5A1,TGFB3,FGF2,COL15A1,COLGALT2,LAMA4,LOXL2,P4HA3,ITGA2,COL6A1,LUM,JAM2,ADAM12,MFAP5,ITGA9,ASPN,COL14A1,ITGA8,COL18A1,CAPN13,SERPINE1,PRKCA,MMP14,FBN2,ITGB6,FMOD,ELN,ADAMTS14,LRP4,COL27A1,FBN1,COL12A1,SPARC,NTN4,COL1A2,ITGA11,NID1,LAMA3,LOX,COL16A1,P4HA2,BMP4,COL4A3,COL21A1,ADAMTS3,MMP2,DMD,THBS1,HSPG2,LTBP2,MFAP4,ADAMTS2,CTSS,LTBP1,DCN,FBLN2

R-HSA-1630316	Glycosaminoglycan metabolism	25	SLC9A1,GPC3,XYLT1,CHST7,EXT2,HEXB,LUM,CHST9,UST,VCAN,ST3GAL4,SDC3,FMOD,KERA,EXT1,CHST15,CHST2,CEMIP,CHST11,CD44,ARSB,HSPG2,GNS,B4GALT1,DCN
R-HSA-9006934	Signaling by Receptor Tyrosine Kinases	23	VEGFD,FGF10,COL6A6,COL5A1,FGF2,LAMA4,PDGFD,ITGA2,ELMO1,SPRY1,FGFR1,PRKCA,PIK3R1,COL27A1,THBS2,COL1A2,LAMA3,COL4A3,THBS1,FLRT2,SHC3,FGF1,GRIN2B
R-HSA-3000178	ECM proteoglycans	21	COL6A6,COL6A2,TGFB3,TNC,COL5A1,LAMA4,ITGA2,LUM,ITGA9,ASPN,ITGA8,SERPINE1,ITGB6,FMOD,LRP4,SPARC,COL1A2,LAMA3,COL4A3,HSPG2,DCN
R-HSA-5173105	O-linked glycosylation	17	GALNT5,THSD1,ADAMTS5,ADAMTS14,GCNT1,GALNT16,ADAMTS15,ADAMTS3,SPON2,THBS2,GALNT10,THBS1,ADAMTS2,ADAMTSL4,ST6GALNAC2,B3GLCT,GALNT17
R-HSA-1474290	Collagen formation	17	P4HA2,COL6A6,ADAMTS14,COL4A3,COL5A1,COL21A1,ADAMTS3,COL14A1,COL15A1,COL12A1,COLGALT2,ADAMTS2,P4HA3,COL1A2,LOX,LAMA3,COL16A1
R-HSA-1566948	Elastic fibre formation	16	BMP4,FBLN5,MFAP5,TGFB3,EFEMP1,FBN1,ITGA8,LTBP2,MFAP4,FBN2,LOXL2,LTBP1,LOX,FBLN2,ITGB6,ELN
R-HSA-1650814	Collagen biosynthesis and modifying enzymes	15	P4HA2,COL6A6,ADAMTS14,COL4A3,COL5A1,COL21A1,ADAMTS3,COL14A1,COL15A1,COL12A1,COLGALT2,ADAMTS2,P4HA3,COL1A2,COL16A1
R-HSA-2129379	Molecules associated with elastic fibres	13	BMP4,FBLN5,MFAP5,TGFB3,EFEMP1,FBN1,ITGA8,LTBP2,LTBP1,FBN2,FBLN2,ITGB6,ELN
R-HSA-6806834	Signaling by MET	8	COL5A1,COL27A1,SH3GL3,LAMA4,PIK3R1,COL1A2,ITGA2,LAMA3
R-HSA-194840	Rho GTPase cycle	3	ARHGAP26,ARHGAP20,ARHGEF18

Table 5.15 Common pathways and genes between RNA-Seq analysis on DEGs at 15 days and DBPs at the same time point according to the Reactome database.

6 DISCUSSION AND CONCLUSIONS

In this work for the PhD program, we integrated different types of high-throughput sequencing data from both the DNA and the RNA of MLPS PDX models with the main aim to evaluate the molecular features responsible for the sensitivity to and the resistance against trabectedin. As pharmacological features can be sustained by complex genomic and transcriptomic networks, we used bioinformatics and computational approaches organised in specific pipelines customised to fit the biological questions.

The results achieved through this work shed light on novel biological aspects related to MLPS and their response to trabectedin treatment. The main results of this analysis can be summarised into the following four main points.

1. We provide a measure of the extent to which PDX overlaps clinical features of MLPS tumours.
2. We showed that loss of genomic material in *4p15.2, 4p16.3* and *17q21.31* cytobands is a distinctive feature of acquired-resistance to trabectedin, even if further investigation is needed, and that genes mapping on them are coherently downregulated in the comparison between resistant and responsive controls.
3. We demonstrated that trabectedin elicits its action mainly at the transcriptional level by restoring the process of adipocyte differentiation after prolonged treatments.
4. We hypothesised a direct role of the drug in competing with the FUS-DDIT3 oncoprotein binding to the DNA causing a re-activation of the transcriptional processes.

These results will be discussed accordingly in the following sections.

6.1 THE GENOMIC LANDSCAPE OF PDX OVERLAPS THAT OF MLPS TUMOURS AND AFFECTS THE SENSITIVITY TO TRABECTEDIN

We started this PhD work from the genomic analysis of the previously established PDX models of MLPS namely ML006, ML015, ML017 and ML017/ET (Frapolli et al., 2010), (Bello et al., 2019). However, the absence of FUS-DDIT3 oncoprotein in ML015 and the unavailability of a healthy matched sample for ML006 forced us to analyse only ML017- and ML017/ET-bearing mice in our study.

We showed that ML017 and ML017/ET fairly reproduce the main genomic features of MLPS tumours. Beside FUS-DDIT3 translocation, MLPS PDX carried SNVs in the *PTEN* and *PIK3CA* genes, two of the most frequently affected genes in MLPS tumour biopsies (Assi et al., 2019), (Keung and Somaiah, 2019). In recent years, many genomic studies have been done on liposarcomas, however only few of them considered MLPS in their cohorts (Keung and Somaiah, 2019), (Hofvander et al., 2018), (Ohguri et al., 2006). Copy number features have been described in MLPS, such as gains in chromosome 8, 13q or loss of chromosome 6 and 16q11, however a gold-standard on the genomic architecture of MLPS does not exist yet, even if most studies agree on the diploidy of the genome and the monoclonal nature of MLPS (Antonescu et al., 2000), (Hofvander et al., 2018), a feature that we confirmed in both ML017 and ML017/ET. Moreover, sarcomas are generally considered tumours with a low mutational burden as reported in the sarcoma (SARC) cohort of The Cancer Genome Atlas (TCGA) (Abeshouse et al., 2017), a feature that we confirmed in ML017 and ML017/ET. Given this evidence, we showed that our models are in line with previously reported MLPS characteristics, thus they can be considered representative of the disease and they also give a comprehensive view of the genomic landscape of MLPS.

SCNAs are responsible for carcinogenesis and tumour progression, affecting large portions of the genome in many cancer types. Amplificated regions usually host oncogenes while genetic loss frequently affects tumour suppressor genes. SCNAs

have also a critical role in response to drug therapy (Zack et al., 2013), (Beroukhim et al., 2010). We showed that cytobands *4p15.2*, *4p16.3* and *17q21.31* were specifically lost in the resistant model. Since ML017/ET was established through different and continuous passages in mice following a strict schedule of trabectedin treatment, we can hypothesise that these three regions have been lost upon drug pressure and they could be associated with acquired resistance against trabectedin. However, it should be considered that in this work only one model was used to mimic the resistance against trabectedin, thus results need to be confirmed in other models to strengthen this evidence, and moreover it should have confirmation in patients that have shown resistance to treatment. This point will be pursued as a future plan for this project.

UVSSA (UV-stimulated scaffold protein A) is one of the most interesting genes mapping on *4p16.3* cytoband. *UVSSA* carried an in-frame deletion and lost one allele in ML017/ET model only. This gene is involved in the Transcription-Coupled Nucleotide Excision Repair (TC-NER), a mechanism that removes helix-distorting lesions from the genome (van der Weegen et al., 2020). The role of UVSSA protein in the TC-NER is crucial since it recruits the TFIIH (transcription factor IIH) complex after the stalling of the RNA Pol II (RNA polymerase II) at the site of damage (van der Weegen et al., 2020). The involvement of *UVSSA* in the TC-NER acquired more importance in the light of the previously identified interactions of trabectedin with this pathway. Indeed, one of the main mechanisms of trabectedin is the inhibition of active transcription through the blockade of RNA Pol II. The block of transcription is usually rescued by the degradation of RNA Pol II that depends on a proficient TC-NER. Studies have shown that cells with defective TC-NER due to mutations in XPC, XPD, XPA, XPG are resistant to trabectedin treatment (Assi et al., 2019), (Larsen et al., 2016). The impairment of the *UVSSA* gene in ML017/ET was confirmed also at the protein level with orthogonal experiments and this result was published by Bello et al. (Bello et al., 2019).

6.2 TRABECTEDIN ACTS AS A TRANSCRIPTIONAL REGULATOR AND INDUCES A PHENOTYPIC CHANGE AFTER PROLONGED TREATMENT

In a previous work, we demonstrated that treatment with trabectedin is able to mediate the transcriptional regulation of myxoid liposarcoma cell lines leading to morphological changes indicated by the formation of lipids and further confirmed by the presence of lipoblast *in vivo* (Forni et al., 2009). In order to clarify the molecular mechanisms behind trabectedin mechanism of action, here we studied the transcriptional modulation induced by the drug in both ML017 and ML017/ET models by comparing each treated condition to controls as explained in Materials and methods, Sections 4.7 and 4.8.3.2. Our results suggest that the greatest transcriptional effect was elicited at the latest time point, e.g. 15 days after the third dose of trabectedin. We identified the activation of functions that are related to the remodelling of the extracellular matrix, the production of collagen, the formation of elastic fibres and the regulation of the insulin-like growth factor. This evidence sustains that trabectedin is able to induce a phenotypic change in MLPS that drives the reactivation of adipocytic differentiation also at the molecular level and further confirms the morphological and phenotypical changes toward the adipocyte state assessed by independent techniques. Since early time points suggest an a-specific response to the drug, *e.g.* TP53 activation, the activation of the differentiation process requires prolonged and continuous treatments.

On the other hand, in the resistant ML017/ET model, trabectedin elicits an early cytotoxic response at 24 hours, while no other transcriptional evidence is registered at later time points, in accordance to the feature of acquired-resistance observed with the growth-curves of the animal models (Bello et al., 2019). This data suggests that prolonged treatment with trabectedin leads to acquired resistance against the drug mimicking the clinical outcome. Nonetheless, we showed that ML017 and ML017/ET-bearing models had different transcriptional signatures already at basal condition. Indeed, 243 genes were inhibited in the resistant model and mapped exactly on the same regions that were specifically lost in ML017/ET.

Thus, this result highlights that from a loss of genetic material derives coherent loss of transcriptional activity. This is further justified by the already cited UVSSA protein inactivation.

The use of anthracyclines as doxorubicin as first-line therapy in MLPS has prompted us to compare the effects of this drug to trabectedin. Doxorubicin causes cumulative toxicity, thus prolonged treatment is not allowed (Frapolli et al., 2019). We showed that some effects seen with trabectedin at 15 days after the third dose can be reproduced by doxorubicin already at 24h, even though with different genes involved. However, beside this slight overlap, overall the two drugs trigger different transcriptional responses. Finally, doxorubicin deregulated genes in ML017/ET participate in the same biological functions as in ML017 suggesting no cross-resistance to this drug.

In a recent work, we showed that the differentiation program induced by trabectedin can be further increased by the combination with pioglitazone (Frapolli et al., 2019). Pioglitazone is a drug commonly used for the treatment of type II diabetes which acts as an agonist of PPARγ. The synergistic effect of trabectedin and pioglitazone in boosting the differentiation of tumour cells in animal models, the evidence that the combination could overcome resistance to trabectedin alone (Frapolli et al., 2019) and the availability of pioglitazone as a generic drug, have eased the path to the opening of a phase II pilot study on the combination versus trabectedin alone, called TRABEPIO (NCT04794127, (Mario Negri Institute for Pharmacological Research, 2021)). Despite strong evidence of an enhanced effect on differentiation, the molecular mechanism that underlies this interaction is not known. In the light of the present study that improved the knowledge on trabectedin mechanism of action, it would be interesting to further investigate its interaction with pioglitazone with a similar approach used here.

By studying the transcriptional modulation induced by drug treatment, we focused only on the entire repertoire of already annotated genes. However, the development of sequencing technologies has allowed the identification of novel isoforms of already known genes and/or of novel transcripts. They have become an

attractive topic since new variants of oncogenes have been discovered (Huang et al., 2021) and have been also associated with drug resistance (Lian et al., 2019). Thus, we investigated unknown transcripts that could underlie drug mechanism of action. Although this level of analysis did not show clear distinct groups as in the gene-level analysis probably due a higher background noise in the data, we showed that the highest number of unknown transcripts was identified at 15 days after the third dose of trabectedin in ML017 in correspondence to the highest gene modulation described so far. Here, we identified 990 transcripts with a high probability to be translated into protein, thus they could be implicated in specific regulatory functions warranted further investigation.

6.3 TRABECTEDIN COULD BE RESPONSIBLE FOR THE DISPLACEMENT OF THE FUS-DDIT3 ONCOPROTEIN FROM ITS BINDING SITES

In this work we provide a genome-wide landscape of the FUS-DDIT3 binding sites in MLPS. Previous works have described single targets of the FUS-DDIT3 oncoprotein. Indeed, Di Giandomenico et al. studied the *PTX3* and *FN1* genes as targets of FUS-DDIT3 (Di Giandomenico et al., 2014), Pérez-Mancera showed that FUS-DDIT3 directly blocks the PPARγ and C/EBPα transcription factors (Pérez-Mancera et al., 2008) causing the inhibition of the adipogenesis process, while the only genome-wide analysis of FUS-DDIT3 was done by Chen et al. in a cell line model (Chen et al., 2019). In this work we made a step further by assessing all the binding sites of FUS-DDIT3 on the DNA of human tumours implanted in mice. We confirmed *PTX3*, *FN1*, PPARγ and C/EBPα as targets of the chimera, however we showed that FUS-DDIT3 could affect other factors that are notoriously involved in adipogenesis like the *PPARα, IGF1, ADIPOQ, PLIN2* and *LPL* genes. We found that most genes were associated with pathways like white fat cell differentiation, signalling by VEGF and extracellular matrix remodelling that suggested a role of the oncoprotein in regulation of the expression of these pathways.

Moreover, we showed that the CCAAT/enhancer-binding protein (C/EBP) family of transcription factors is likely to act as a co-factor of FUS-DDIT3. Indeed, C/EBP

binding motif was found beside all bindings sites of FUS-DDIT3. *DDIT3* itself belongs to the C/EBP transcription factor family. It is known to form heterodimers with other C/EBP members and co-participate in transcriptional regulation programs (Wang et al., 2019). This evidence suggests that FUS-DDIT3 may cooperate with the C/EBP family, a mechanism that warranted further investigation.

We exploited further the mechanism of FUS-DDIT3 in MLPS and studied its behaviour under trabectedin treatment. We showed that at a later time point, *e.g.* 15 days after the third dose of trabectedin, most of the binding sites were lost suggesting that the drug could be able to displace FUS-DDIT3 from its target sites or at least to compete with the drug in the binding to the DNA. It is worth noting that at this time point we see a late response after the last administration of the drug, however we cannot exclude that some drug is still attached to the DNA due to the long lasting half-life time of trabectedin. Also, a fraction of cell still carry the chimera FUS-DDIT3 as demonstrated by western blot (not shown here) and proven by the ChIP-Seq that otherwise could have not been done. However, integration with transcriptional data confirmed that the modulation of FUS-DDIT3 binding is followed by a transcriptional modulation of these target genes.

Even if further investigation is needed at both the molecular and pharmacological levels, overall the image depicted by ChIP-Seq and RNA-Seq data suggests a double mechanism of trabectedin in MLPS: a first early effect that is independent from FUS-DDIT3 and is mainly associated with the cytotoxicity of the drug; a second late effect that is exerted in the responsive model with a direct action on FUS-DDIT3, the displacement of which restores the normal differentiation process of the cells.

6.4 LIMITATIONS OF THE STUDY

There are some limitations in this study that could be addressed in the ongoing plan of the research.

First, the study focused on one PDX model, ML017, that represents a limitation to the generalisation of the findings to MLPS. We initially started from the analysis of other two models, ML006 and ML015, however due to unpredictable technical issues they could not be considered for further analysis. In order to overcome this issue, we focused on the biological reproducibility of the data, in fact for each sequencing experiment we set at least three or four samples per condition. In line with this, the study of the resistance against trabectedin in ML017/ET model alone needs further assessment in other models or in clinical-derived data.

The other limitation mainly concerns ChIP-Seq data. Due to sample unavailability and the long working time that this technique required, we could not complete the analysis of the histone modifications in the responsive model and could not analyse the ML017/ET model. Moreover, due to the complex pharmacokinetics of the drug, further and complementary experiments need to be performed to better clarify the mechanisms underlying trabectedin mechanism of action.

6.5 CONCLUSIONS

We started this PhD work with the aim to elucidate the mechanism of responsiveness to and resistance against trabectedin in MLPS.

We presented a comprehensive study of MLPS-bearing PDX models that we exploited with bioinformatics and computational approaches by integrated different levels: genomic, transcriptomic and DNA-protein interactions.

This PhD work offers an unprecedented broad molecular picture on MLPS and their response to drug treatment. It conveys a methodological workflow for pharmacogenomic studies. Then, it presents these PDX models as a valid tool for the study of MLPS that could overcome the need of statistical numerosity for drug testing.

Lastly, the mechanistic insights into the efficacy of trabectedin described here might help to design novel therapeutic strategies to enhance tumour response or to overcome drug resistance mechanisms.

7 REFERENCES

Abeshouse, A., Adebamowo, C., Adebamowo, S.N., Akbani, R., Akeredolu, T., Ally, A., Anderson, M.L., Anur, P., Appelbaum, E.L., Armenia, J., Auman, J.T., Bailey, M.H., Baker, L., Balasundaram, M., Balu, S., Barthel, F.P., Bartlett, J., Baylin, S.B., Behera, M., Belyaev, D., Bennett, J., Benz, C., Beroukhim, R., Birrer, M., Bocklage, T., Bodenheimer, T., Boice, L., Bootwalla, M.S., Bowen, J., Bowlby, R., Boyd, J., Brohl, A.S., Brooks, D., Byers, L., Carlsen, R., Castro, P., Chen, H.-W., Cherniack, A.D., Chibon, F., Chin, L., Cho, J., Chuah, E., Chudamani, S., Cibulskis, C., Cooper, L.A., Cope, L., Cordes, M.G., Crain, D., Curley, E., Danilova, L., Dao, F., Davis, I.J., Davis, L.E., Defreitas, T., Delman, K., Demchok, J.A., Demetri, G.D., Demicco, E.G., Dhalla, N., Diao, L., Ding, L., DiSaia, P., Dottino, P., Doyle, L.A., Drill, E., Dubina, M., Eschbacher, J., Fedosenko, K., Felau, I., Ferguson, M.L., Frazer, S., Fronick, C.C., Fulidou, V., Fulton, L.A., Fulton, R.S., Gabriel, S.B., Gao, J., Gao, Q., Gardner, J., Gastier-Foster, J.M., Gay, C.M., Gehlenborg, N., Gerken, M., Getz, G., Godwin, A.K., Godwin, E.M., Gordienko, E., Grilley-Olson, J.E., Gutman, D.A., Gutmann, D.H., Hayes, D.N., Hegde, A.M., Heiman, D.I., Heins, Z., Helsel, C., Hepperla, A.J., Higgins, K., Hoadley, K.A., Hobensack, S., Holt, R.A., Hoon, D.B., Hornick, J.L., Hoyle, A.P., Hu, X., Huang, M., Hutter, C.M., Iacocca, M., Ingram, D.R., Ittmann, M., Iype, L., Jefferys, S.R., Jones, K.B., Jones, C.D., Jones, S.J., Kalir, T., Karlan, B.Y., Karseladze, A., Kasaian, K., Kim, J., Kundra, R., Kuo, H., Ladanyi, M., Lai, P.H., Laird, P.W., Larsson, E., Lawrence, M.S., Lazar, A.J., Lee, S., Lee, D., Lehmann, K.-V., Leraas, K.M., Lester, J., Levine, D.A., Li, I., Lichtenberg, T.M., Lin, P., Liu, J., Liu, W., Liu, E.M., Lolla, L., Lu, Y., Ma, Y., Madan, R., Maglinte, D.T., Magliocco, A., Maki, R.G., Mallery, D., Manikhas, G., Mardis, E.R., Mariamidze, A., Marra, M.A., Martignetti, J.A., Martinez, C., Mayo, M., McLellan, M.D., Meier, S., Meng, S., Meyerson, M., Mieczkowski, P.A., Miller, C.A., Mills, G.B., Moore, R.A., Morris, S., Mose, L.E., Mozgovoy, E., Mungall, A.J., Mungall, K., Nalisnik, M., Naresh, R., Newton, Y., Noble, M.S., Novak, J.E., Ochoa, A., Olvera, N., Owonikoko, T.K., Paklina, O., Parfitt, J., Parker, J.S., Pastore, A., Paulauskis, J., Penny, R., Pereira, E., Perou, C.M., Perou, A.H., Pihl, T., Pollock, R.E., Potapova, O., Radenbaugh, A.J., Ramalingam, S.S., Ramirez, N.C., Rathmell, W.K., Raut, C.P., Riedel, R.F., Reilly, C., Reynolds, S.M., Roach, J., Robertson, A.G., Roszik,

J., Rubin, B.P., Sadeghi, S., Saksena, G., Salner, A., Sanchez-Vega, F., Sander, C., Schein, J.E., Schmidt, H.K., Schultz, N., Schumacher, S.E., Sekhon, H., Senbabaoglu, Y., Setdikova, G., Shelton, C., Shelton, T., Shen, R., Shi, Y., Shih, J., Shmulevich, I., Sica, G.L., Simons, J.V., Singer, S., Sipahimalani, P., Skelly, T., Socci, N., Sofia, H.J., Soloway, M.G., Spellman, P., Sun, Q., Swanson, P., Tam, A., Tan, D., Tarnuzzer, R., Thiessen, N., Thompson, E., Thorne, L.B., Tong, P., Torres, K.E., van de Rijn, M., Van Den Berg, D.J., Van Tine, B.A., Veluvolu, U., Verhaak, R., Voet, D., Voronina, O., Wan, Y., Wang, Z., Wang, J., Weinstein, J.N., Weisenberger, D.J., Wilkerson, M.D., Wilson, R.K., Wise, L., Wong, T., Wong, W., Wrangle, J., Wu, Y., Wyczalkowski, M., Yang, L., Yau, C., Yellapantula, V., Zenklusen, J.C., Zhang, J. (Julia), Zhang, Hailei, Zhang, Hongxin, Zmuda, E., 2017. COMPREHENSIVE AND INTEGRATED GENOMIC CHARACTERIZATION OF ADULT SOFT TISSUE SARCOMAS. Cell 171, 950-965.e28. https://doi.org/10.1016/j.cell.2017.10.014

Ahdesmäki, M.J., Gray, S.R., Johnson, J.H., Lai, Z., 2016. Disambiguate: An open-source application for disambiguating two species in next generation sequencing data from grafted samples. F1000Research 5, 2741. https://doi.org/10.12688/f1000research.10082.2

Antonescu, C.R., Elahi, A., Healey, J.H., Brennan, M.F., Lui, M.Y., Lewis, J., Jhanwar, S.C., Woodruff, J.M., Ladanyi, M., 2000. Monoclonality of Multifocal Myxoid Liposarcoma: Confirmation by Analysis of TLS-CHOP or EWS-CHOP Rearrangements. Clin. Cancer Res. 6, 2788–2793.

Antonescu, C.R., Tschernyavsky, S.J., Decuseara, R., Leung, D.H., Woodruff, J.M., Brennan, M.F., Bridge, J.A., Neff, J.R., Goldblum, J.R., Ladanyi, M., 2001. Prognostic impact of P53 status, TLS-CHOP fusion transcript structure, and histological grade in myxoid liposarcoma: a molecular and clinicopathologic study of 82 cases. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. 7, 3977–3987.

Assi, T., Kattan, J., El Rassy, E., Honore, C., Dumont, S., Mir, O., Le Cesne, A., 2019. A comprehensive review of the current evidence for trabectedin in advanced myxoid liposarcoma. Cancer Treat. Rev. 72, 37–44. https://doi.org/10.1016/j.ctrv.2018.11.003

Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data [WWW Document], n.d. URL https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed 10.20.20).

Barreca, M., Spanò, V., Montalbano, A., Cueto, M., Díaz Marrero, A.R., Deniz, I., Erdoğan, A., Lukić Bilela, L., Moulin, C., Taffin-de-Givenchy, E., Spriano, F., Perale, G., Mehiri, M., Rotter, A., P Thomas, O., Barraja, P., Gaudêncio, S.P., Bertoni, F., 2020. Marine Anticancer Agents: An Overview with a Particular Focus on Their Chemical Classes. Mar. Drugs 18. https://doi.org/10.3390/md18120619

bcl2fastq Conversion Software [WWW Document], n.d. URL https://emea.support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html (accessed 10.20.20).

Bello, E., Brich, S., Craparotta, I., Mannarino, L., Ballabio, S., Gatta, R., Marchini, S., Carrassa, L., Matteo, C., Sanfilippo, R., Gronchi, A., Casali, P.G., Pilotti, S., D'Incalci, M., Frapolli, R., 2019. Establishment and characterisation of a new patient-derived model of myxoid liposarcoma with acquired resistance to trabectedin. Br. J. Cancer 121, 464–473. https://doi.org/10.1038/s41416-019-0550-2

Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J.S., Dobson, J., Urashima, M., Mc Henry, K.T., Pinchback, R.M., Ligon, A.H., Cho, Y.-J., Haery, L., Greulich, H., Reich, M., Winckler, W., Lawrence, M.S., Weir, B.A., Tanaka, K.E., Chiang, D.Y., Bass, A.J., Loo, A., Hoffman, C., Prensner, J., Liefeld, T., Gao, Q., Yecies, D., Signoretti, S., Maher, E., Kaye, F.J., Sasaki, H., Tepper, J.E., Fletcher, J.A., Tabernero, J., Baselga, J., Tsao, M.-S., DeMichelis, F., Rubin, M.A., Janne, P.A., Daly, M.J., Nucera, C., Levine, R.L., Ebert, B.L., Gabriel, S., Rustgi, A.K., Antonescu, C.R., Ladanyi, M., Letai, A., Garraway, L.A., Loda, M., Beer, D.G., True, L.D., Okamoto, A., Pomeroy, S.L., Singer, S., Golub, T.R., Lander, E.S., Getz, G., Sellers, W.R., Meyerson, M., 2010. The landscape of somatic copy-number alteration across human cancers. Nature 463, 899–905. https://doi.org/10.1038/nature08822

Bhimani, J., Ball, K., Stebbing, J., 2020. Patient-derived xenograft models—the

future of personalised cancer treatment. Br. J. Cancer 122, 601–602. https://doi.org/10.1038/s41416-019-0678-0

Blay, J.-Y., Leahy, M.G., Nguyen, B.B., Patel, S.R., Hohenberger, P., Santoro, A., Staddon, A.P., Penel, N., Piperno-Neumann, S., Hendifar, A., Lardelli, P., Nieto, A., Alfaro, V., Chawla, S.P., 2014. Randomised phase III trial of trabectedin versus doxorubicin-based chemotherapy as first-line therapy in translocation-related sarcomas. Eur. J. Cancer 50, 1137–1147. https://doi.org/10.1016/j.ejca.2014.01.012

Bourne, P.E., 2021. Is "bioinformatics" dead? PLOS Biol. 19, e3001165. https://doi.org/10.1371/journal.pbio.3001165

Braig, D., Becherer, C., Bickert, C., Braig, M., Claus, R., Eisenhardt, A.E., Heinz, J., Scholber, J., Herget, G.W., Bronsert, P., Fricke, A., Follo, M., Stark, G.B., Bannasch, H., Eisenhardt, S.U., 2019. Genotyping of circulating cell-free DNA enables noninvasive tumor detection in myxoid liposarcomas. Int. J. Cancer 145, 1148–1161. https://doi.org/10.1002/ijc.32216

Callari, M., Batra, A.S., Batra, R.N., Sammut, S.-J., Greenwood, W., Clifford, H., Hercus, C., Chin, S.-F., Bruna, A., Rueda, O.M., Caldas, C., 2018. Computational approach to discriminate human and mouse sequences in patient-derived tumour xenografts. BMC Genomics 19, 19. https://doi.org/10.1186/s12864-017-4414-y

cclab-brca/ICRG [WWW Document], n.d. . GitHub. URL https://github.com/cclab-brca/ICRG (accessed 2.16.21).

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., Saunders, C.T., 2016. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 32, 1220–1222. https://doi.org/10.1093/bioinformatics/btv710

Chen, Y., Lun, A.T.L., Smyth, G.K., 2016. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. https://doi.org/10.12688/f1000research.8987.2

Chen, Y., Xu, L., Mayakonda, A., Huang, M.-L., Kanojia, D., Tan, T.Z., Dakle, P., Lin, R.Y.-T., Ke, X.-Y., Said, J.W., Chen, J., Gery, S., Ding, L.-W., Jiang, Y.-Y., Pang, A., Puhaindran, M.E., Goh, B.C., Koeffler, H.P., 2019. Bromodomain and extraterminal proteins foster the core transcriptional regulatory programs and confer vulnerability in liposarcoma. Nat. Commun. 10, 1353. https://doi.org/10.1038/s41467-019-09257-z

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., Getz, G., 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. 31, 213–219. https://doi.org/10.1038/nbt.2514

Compton, M.L., Al-Rohil, R.N., 2021. The Utility of Perilipin in Liposarcomas: PLIN1 Differentiates Round Cell Liposarcoma From Other Round Cell Sarcomas. Appl. Immunohistochem. Mol. Morphol. 29, 152–157. https://doi.org/10.1097/PAI.0000000000000851

Contents — bcbio-nextgen 1.2.4 documentation [WWW Document], n.d. URL https://bcbio-nextgen.readthedocs.io/en/latest/ (accessed 10.21.20).

Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing on JSTOR [WWW Document], n.d. URL https://www.jstor.org/stable/2346101 (accessed 12.29.21).

Crozat, A., Åman, P., Mandahl, N., Ron, D., 1993. Fusion of CHOP to a novel RNA-binding protein in human myxoid liposarcoma. Nature 363, 640–644. https://doi.org/10.1038/363640a0

Di Giandomenico, S., Frapolli, R., Bello, E., Uboldi, S., Licandro, S.A., Marchini, S., Beltrame, L., Brich, S., Mauro, V., Tamborini, E., Pilotti, S., Casali, P.G., Grosso, F., Sanfilippo, R., Gronchi, A., Mantovani, R., Gatta, R., Galmarini, C.M., Sousa-Faro, J.M.F., D'Incalci, M., 2014. Mode of action of trabectedin in myxoid liposarcomas. Oncogene 33, 5201–5210. https://doi.org/10.1038/onc.2013.462

D'Incalci, M., Badri, N., Galmarini, C.M., Allavena, P., 2014. Trabectedin, a drug

acting on both cancer cells and the tumour microenvironment. Br. J. Cancer 111, 646–650. https://doi.org/10.1038/bjc.2014.149

D'Incalci, M., Galmarini, C.M., 2010. A review of trabectedin (ET-743): a unique mechanism of action. Mol. Cancer Ther. 9, 2157–2163. https://doi.org/10.1158/1535-7163.MCT-10-0263

Fornes, O., Castro-Mondragon, J.A., Khan, A., van der Lee, R., Zhang, X., Richmond, P.A., Modi, B.P., Correard, S., Gheorghe, M., Baranašić, D., Santana-Garcia, W., Tan, G., Chèneby, J., Ballester, B., Parcy, F., Sandelin, A., Lenhard, B., Wasserman, W.W., Mathelier, A., 2020. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 48, D87–D92. https://doi.org/10.1093/nar/gkz1001

Forni, C., Minuzzo, M., Virdis, E., Tamborini, E., Simone, M., Tavecchio, M., Erba, E., Grosso, F., Gronchi, A., Aman, P., Casali, P., D'Incalci, M., Pilotti, S., Mantovani, R., 2009. Trabectedin (ET-743) promotes differentiation in myxoid liposarcoma tumors. Mol. Cancer Ther. 8, 449–457. https://doi.org/10.1158/1535-7163.MCT-08-0848

Frapolli, R., Bello, E., Ponzo, M., Craparotta, I., Mannarino, L., Ballabio, S., Marchini, S., Carrassa, L., Ubezio, P., Porcu, L., Brich, S., Sanfilippo, R., Casali, P.G., Gronchi, A., Pilotti, S., D'Incalci, M., 2019. Combination of PPARγ Agonist Pioglitazone and Trabectedin Induce Adipocyte Differentiation to Overcome Trabectedin Resistance in Myxoid Liposarcomas. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. 25, 7565–7575. https://doi.org/10.1158/1078-0432.CCR-19-0976

Frapolli, R., Tamborini, E., Virdis, E., Bello, E., Tarantino, E., Marchini, S., Grosso, F., Sanfilippo, R., Gronchi, A., Tercero, J.C., Peloso, G., Casali, P., Pilotti, S., D'Incalci, M., 2010. Novel Models of Myxoid Liposarcoma Xenografts Mimicking the Biological and Pharmacologic Features of Human Tumors. Clin. Cancer Res. 16, 4958–4967. https://doi.org/10.1158/1078-0432.CCR-10-0317

Germano, G., Frapolli, R., Belgiovine, C., Anselmo, A., Pesce, S., Liguori, M., Erba, E., Uboldi, S., Zucchetti, M., Pasqualini, F., Nebuloni, M., van Rooijen, N., Mortarini, R.,

Beltrame, L., Marchini, S., Fuso Nerini, I., Sanfilippo, R., Casali, P.G., Pilotti, S., Galmarini, C.M., Anichini, A., Mantovani, A., D'Incalci, M., Allavena, P., 2013. Role of macrophage targeting in the antitumor activity of trabectedin. Cancer Cell 23, 249–262. https://doi.org/10.1016/j.ccr.2013.01.008

Grosso, F., Jones, R.L., Demetri, G.D., Judson, I.R., Blay, J.-Y., Le Cesne, A., Sanfilippo, R., Casieri, P., Collini, P., Dileo, P., Spreafico, C., Stacchiotti, S., Tamborini, E., Tercero, J.C., Jimeno, J., D'Incalci, M., Gronchi, A., Fletcher, J.A., Pilotti, S., Casali, P.G., 2007. Efficacy of trabectedin (ecteinascidin-743) in advanced pretreated myxoid liposarcomas: a retrospective study. Lancet Oncol. 8, 595–602. https://doi.org/10.1016/S1470-2045(07)70175-4

Grosso, F., Sanfilippo, R., Virdis, E., Piovesan, C., Collini, P., Dileo, P., Morosi, C., Tercero, J.C., Jimeno, J., D'Incalci, M., Gronchi, A., Pilotti, S., Casali, P.G., 2009. Trabectedin in myxoid liposarcomas (MLS): a long-term analysis of a single-institution series. Ann. Oncol. 20, 1439–1444. https://doi.org/10.1093/annonc/mdp004

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hemminger, J.A., Ewart Toland, A., Scharschmidt, T.J., Mayerson, J.L., Kraybill, W.G., Guttridge, D.C., Iwenofu, O.H., 2013. The cancer-testis antigen NY-ESO-1 is highly expressed in myxoid and round cell subset of liposarcomas. Mod. Pathol. Off. J. U. S. Can. Acad. Pathol. Inc 26, 282–288. https://doi.org/10.1038/modpathol.2012.133

Hemminger, J.A., Iwenofu, O.H., 2013. NY-ESO-1 is a sensitive and specific immunohistochemical marker for myxoid and round cell liposarcomas among related mesenchymal myxoid neoplasms. Mod. Pathol. Off. J. U. S. Can. Acad.

Pathol. Inc 26, 1204–1210. https://doi.org/10.1038/modpathol.2013.65

Hofvander, J., Viklund, B., Isaksson, A., Brosjö, O., Vult von Steyern, F., Rissler, P., Mandahl, N., Mertens, F., 2018. Different patterns of clonal evolution among different sarcoma subtypes followed for up to 25 years. Nat. Commun. 9, 3662. https://doi.org/10.1038/s41467-018-06098-0

Home | Inkscape [WWW Document], n.d. URL https://inkscape.org/it/ (accessed 2.16.21).

Huang, K.K., Huang, J., Wu, J.K.L., Lee, M., Tay, S.T., Kumar, V., Ramnarayanan, K., Padmanabhan, N., Xu, C., Tan, A.L.K., Chan, C., Kappei, D., Göke, J., Tan, P., 2021. Long-read transcriptome sequencing reveals abundant promoter diversity in distinct molecular subtypes of gastric cancer. Genome Biol. 22, 44. https://doi.org/10.1186/s13059-021-02261-x

Hunter, J.D., 2007. Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. 9, 90–95. https://doi.org/10.1109/MCSE.2007.55

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., Loney, F., May, B., Milacic, M., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Weiser, J., Wu, G., Stein, L., Hermjakob, H., D'Eustachio, P., 2020. The reactome pathway knowledgebase. Nucleic Acids Res. 48, D498–D503. https://doi.org/10.1093/nar/gkz1031

Jeff Reback, Wes McKinney, jbrockmendel, Joris Van den Bossche, Tom Augspurger, Phillip Cloud, gfyoung, Sinhrks, Adam Klein, Matthew Roeschke, Simon Hawkins, Jeff Tratner, Chang She, William Ayd, Terji Petersen, Marc Garcia, Jeremy Schendel, Andy Hayden, MomIsBestFriend, Vytautas Jancauskas, Pietro Battiston, Skipper Seabold, chris-b1, h-vetinari, Stephan Hoyer, Wouter Overmeire, alimcmaster1, Kaiqi Dong, Christopher Whelan, Mortada Mehyar, 2020. pandas-dev/pandas: Pandas 1.0.3. Zenodo. https://doi.org/10.5281/zenodo.3715232

Jiang, S., Mortazavi, A., 2018. Integrating ChIP-seq with other functional genomics data. Brief. Funct. Genomics 17, 104–115. https://doi.org/10.1093/bfgp/ely002

Kanojia, D., Nagata, Y., Garg, M., Lee, D.H., Sato, A., Yoshida, K., Sato, Y., Sanada, M., Mayakonda, A., Bartenhagen, C., Klein, H.-U., Doan, N.B., Said, J.W., Mohith, S., Gunasekar, S., Shiraishi, Y., Chiba, K., Tanaka, H., Miyano, S., Myklebost, O., Yang, H., Dugas, M., Meza-Zepeda, L.A., Silberman, A.W., Forscher, C., Tyner, J.W., Ogawa, S., Koeffler, H.P., 2015. Genomic landscape of liposarcoma. Oncotarget 6, 42429–42444. https://doi.org/10.18632/oncotarget.6464

Keung, E.Z., Somaiah, N., 2019. Overview of liposarcomas and their genomic landscape. J. Transl. Genet. Genomics 3. https://doi.org/10.20517/jtgg.2019.03

Kim, D., Paggi, J.M., Park, C., Bennett, C., Salzberg, S.L., 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. 37, 907–915. https://doi.org/10.1038/s41587-019-0201-4

Koczkowska, M., Lipska-Ziętkiewicz, B.S., Iliszko, M., Ryś, J., Miettinen, M., Lasota, J., Biernat, W., Harazin-Lechowska, A., Kruczak, A., Limon, J., 2017. Application of high-resolution genomic profiling in the differential diagnosis of liposarcoma. Mol. Cytogenet. 10, 7. https://doi.org/10.1186/s13039-017-0309-5

Lai, Y., Wei, X., Lin, S., Qin, L., Cheng, L., Li, P., 2017. Current status and perspectives of patient-derived xenograft models in cancer research. J. Hematol. Oncol.J Hematol Oncol 10, 106. https://doi.org/10.1186/s13045-017-0470-7

Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R., Johnson, J., Dougherty, B., Barrett, J.C., Dry, J.R., 2016. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 44, e108. https://doi.org/10.1093/nar/gkw227

Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K.I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A.J., Hoffman, M.M., Iyer, V.R., Jung, Y.L., Karmakar, S., Kellis, M., Kharchenko, P.V., Li, Q., Liu, T., Liu, X.S., Ma, L., Milosavljevic, A., Myers, R.M., Park, P.J., Pazin, M.J., Perry, M.D., Raha, D., Reddy, T.E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J.A., Tolstorukov, M.Y., White, K.P., Xi, S., Farnham, P.J., Lieb,

J.D., Wold, B.J., Snyder, M., 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 22, 1813–1831. https://doi.org/10.1101/gr.136184.111

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923

Lansu, J., Van Houdt, W.J., Schaapveld, M., Walraven, I., Van de Sande, M.A.J., Ho, V.K.Y., Haas, R.L., 2020. Time Trends and Prognostic Factors for Overall Survival in Myxoid Liposarcomas: A Population-Based Study. Sarcoma 2020, 2437850. https://doi.org/10.1155/2020/2437850

Larsen, A.K., Galmarini, C.M., D'Incalci, M., 2016. Unique features of trabectedin mechanism of action. Cancer Chemother. Pharmacol. 77, 663–671. https://doi.org/10.1007/s00280-015-2918-1

Law, C.W., Chen, Y., Shi, W., Smyth, G.K., 2014. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 15, 1–17. https://doi.org/10.1186/gb-2014-15-2-r29

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma. Oxf. Engl. 25, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Subgroup, 1000 Genome Project Data Processing, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lian, B., Hu, X., Shao, Z., 2019. Unveiling novel targets of paclitaxel resistance by single molecule long-read RNA sequencing in breast cancer. Sci. Rep. 9, 6032. https://doi.org/10.1038/s41598-019-42184-z

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550. https://doi.org/10.1186/s13059-014-0550-8

Marchini, S., Marrazzo, E., Bonomi, R., Chiorino, G., Zaffaroni, M., Weissbach, L., Hornicek, F.J., Broggini, M., Faircloth, G.T., D'Incalci, M., 2005. Molecular characterisation of two human cancer cell lines selected in vitro for their chemotherapeutic drug resistance to ET-743. Eur. J. Cancer Oxf. Engl. 1990 41, 323–333. https://doi.org/10.1016/j.ejca.2004.10.021

Mario Negri Institute for Pharmacological Research, 2021. A Phase II Study on Trabectedin in Combination With PPARg Agonist Pioglitazone in Patients With Round Cell Myxoid Liposarcomas or Dedifferentiated G1 and G2 Liposarcomas With Stable Disease After a Monotherapy With Trabectedin. (TRABEPIO) (Clinical trial registration No. NCT04794127). clinicaltrials.gov.

Markowetz, F., 2017. All biology is computational biology. PLOS Biol. 15, e2002050. https://doi.org/10.1371/journal.pbio.2002050

Martens, M., Ammar, A., Riutta, A., Waagmeester, A., Slenter, D.N., Hanspers, K., A Miller, R., Digles, D., Lopes, E.N., Ehrhart, F., Dupuis, L.J., Winckers, L.A., Coort, S.L., Willighagen, E.L., Evelo, C.T., Pico, A.R., Kutmon, M., 2021. WikiPathways: connecting communities. Nucleic Acids Res. 49, D613–D621. https://doi.org/10.1093/nar/gkaa1024

Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C., Koeffler, H.P., 2018. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. 28, 1747–1756. https://doi.org/10.1101/gr.239244.118

McCarthy, D.J., Chen, Y., Smyth, G.K., 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 40, 4288–4297. https://doi.org/10.1093/nar/gks042

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., Cunningham, F., 2016. The Ensembl Variant Effect Predictor. Genome Biol. 17, 122. https://doi.org/10.1186/s13059-016-0974-4

Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., Getz, G., 2011. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal

somatic copy-number alteration in human cancers. Genome Biol. 12, R41. https://doi.org/10.1186/gb-2011-12-4-r41

Michael Waskom, Maoz Gelbart, Olga Botvinnik, Joel Ostblom, Paul Hobson, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, Jordi Warmenhoven, John B. Cole, Julian de Ruiter, Jake Vanderplas, Stephan Hoyer, Cameron Pye, Alistair Miles, Corban Swain, Kyle Meyer, Marcel Martin, Pete Bachant, Eric Quintero, Gero Kunter, Santi Villalba, Brian, Clark Fitzgerald, Constantine Evans, Mike Lee Williams, Drew O'Kane, Tal Yarkoni, Thomas Brunner, 2020. mwaskom/seaborn: v0.11.1 (December 2020). Zenodo. https://doi.org/10.5281/zenodo.4379347

Minuzzo, M., Marchini, S., Broggini, M., Faircloth, G., D'Incalci, M., Mantovani, R., 2000. Interference of transcriptional activation by the antineoplastic drug ecteinascidin-743. Proc. Natl. Acad. Sci. 97, 6780–6784. https://doi.org/10.1073/pnas.97.12.6780

Mocellin, S., 2021. Myxoid/Round Cell Liposarcoma, in: Mocellin, S. (Ed.), Soft Tissue Tumors : A Practical and Comprehensive Guide to Sarcomas and Benign Neoplasms. Springer International Publishing, Cham, pp. 581–584. https://doi.org/10.1007/978-3-030-58710-9_182

Monk, B.J., Herzog, T.J., Kaye, S.B., Krasner, C.N., Vermorken, J.B., Muggia, F.M., Pujade-Lauraine, E., Lisyanskaya, A.S., Makhson, A.N., Rolski, J., Gorbounova, V.A., Ghatage, P., Bidzinski, M., Shen, K., Ngan, H.Y.-S., Vergote, I.B., Nam, J.-H., Park, Y.C., Lebedinsky, C.A., Poveda, A.M., 2010. Trabectedin plus pegylated liposomal Doxorubicin in recurrent ovarian cancer. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. 28, 3107–3114. https://doi.org/10.1200/JCO.2009.25.4037

Nakato, R., Sakata, T., 2021. Methods for ChIP-seq analysis: A practical workflow and advanced applications. Methods, Advance Epigenetics Methods in Biomedicine 187, 44–53. https://doi.org/10.1016/j.ymeth.2020.03.005

Ng, P.C., Henikoff, S., 2003. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 31, 3812–3814.

Noguchi, R., Yoshimatsu, Y., Sei, A., Hirabayashi, K., Ozawa, I., Kikuta, K., Kondo, T., 2020. Establishment and characterization of NCC-MLPS1-C1: a novel patient-derived cell line of myxoid liposarcoma. Hum. Cell. https://doi.org/10.1007/s13577-020-00454-3

Ohguri, T., Hisaoka, M., Kawauchi, S., Sasaki, K., Aoki, T., Kanemitsu, S., Matsuyama, A., Korogi, Y., Hashimoto, H., 2006. Cytogenetic analysis of myxoid liposarcoma and myxofibrosarcoma by array-based comparative genomic hybridisation. J. Clin. Pathol. 59, 978–983. https://doi.org/10.1136/jcp.2005.034942

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W.E., Wei, Z., Wang, K., Lyon, G.J., 2013. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med. 5, 28. https://doi.org/10.1186/gm432

Paila, U., Chapman, B.A., Kirchner, R., Quinlan, A.R., 2013. GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations. PLOS Comput. Biol. 9, e1003153. https://doi.org/10.1371/journal.pcbi.1003153

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., Kingsford, C., 2017. Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods 14, 417–419. https://doi.org/10.1038/nmeth.4197

Perez, F., Granger, B.E., 2007. IPython: A System for Interactive Scientific Computing. Comput. Sci. Eng. 9, 21–29. https://doi.org/10.1109/MCSE.2007.53

Pérez-Mancera, P.A., Bermejo-Rodríguez, C., Sánchez-Martín, M., Abollo-Jiménez, F., Pintado, B., Sánchez-García, I., 2008. FUS-DDIT3 prevents the development of adipocytic precursors in liposarcoma by repressing PPARgamma and C/EBPalpha and activating eIF4E. PloS One 3, e2569. https://doi.org/10.1371/journal.pone.0002569

Pertea, M., Kim, D., Pertea, G.M., Leek, J.T., Salzberg, S.L., 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat. Protoc. 11, 1650–1667. https://doi.org/10.1038/nprot.2016.095

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., Salzberg, S.L., 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33, 290–295. https://doi.org/10.1038/nbt.3122

prepDE [WWW Document], n.d. URL http://ccb.jhu.edu/software/stringtie/dl/prepDE.py (accessed 10.21.20).

Qi, Y., Hu, Y., Yang, H., Zhuang, R., Hou, Y., Tong, H., Feng, Y., Huang, Y., Jiang, Q., Ji, Q., Gu, Q., Zhang, Z., Tang, X., Lu, W., Zhou, Y., 2017. Establishing a patient-derived xenograft model of human myxoid and round-cell liposarcoma. Oncotarget 8, 54320–54330. https://doi.org/10.18632/oncotarget.17352

Ramensky, V., Bork, P., Sunyaev, S., 2002. Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 30, 3894–3900. https://doi.org/10.1093/nar/gkf493

Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., Kircher, M., 2019. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 47, D886–D894. https://doi.org/10.1093/nar/gky1016

Robinson, J.T., Thorvaldsdóttir, H., Wenger, A.M., Zehir, A., Mesirov, J.P., 2017. Variant Review with the Integrative Genomics Viewer. Cancer Res. 77, e31–e34. https://doi.org/10.1158/0008-5472.CAN-17-0337

Ross-Innes, C.S., Stark, R., Teschendorff, A.E., Holmes, K.A., Ali, H.R., Dunning, M.J., Brown, G.D., Gojis, O., Ellis, I.O., Green, A.R., Ali, S., Chin, S.-F., Palmieri, C., Caldas, C., Carroll, J.S., 2012. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature 481, 389–393. https://doi.org/10.1038/nature10730

Sanfilippo, R., Colia, V., Fumagalli, E., Bertulli, R., Morosi, C., Messina, A., Pilotti, S., Dei Tos, A.P., Gronchi, A., Casali, P.G., 2015. Time to secondary resistance (TTSR) after rechallenge with trabectedin (T) in myxoid round cell liposarcoma (MRCLPS) patients. J. Clin. Oncol. 33, 10566–10566. https://doi.org/10.1200/jco.2015.33.15_suppl.10566

Saturni, S., Bellini, F., Braido, F., Paggiaro, P., Sanduzzi, A., Scichilone, N., Santus,

P.A., Morandi, L., Papi, A., 2014. Randomized Controlled Trials and real life studies. Approaches and methodologies: a clinical point of view. Pulm. Pharmacol. Ther. 27, 129–138. https://doi.org/10.1016/j.pupt.2014.01.005

Scapa, J.V., Cloutier, J.M., Raghavan, S.S., Peters-Schulze, G., Varma, S., Charville, G.W., 2021. DDIT3 Immunohistochemistry Is a Useful Tool for the Diagnosis of Myxoid Liposarcoma. Am. J. Surg. Pathol. 45, 230–239. https://doi.org/10.1097/PAS.0000000000001564

Sequencing Support – Coverage Calculator [WWW Document], n.d. URL https://emea.support.illumina.com/downloads/sequencing_coverage_calculator.html (accessed 2.12.21).

Slurm Workload Manager - Documentation [WWW Document], n.d. URL https://slurm.schedmd.com/documentation.html (accessed 2.15.21).

Somaiah, N., Chawla, S.P., Block, M.S., Morris, J.C., Do, K., Kim, J.W., Druta, M., Sankhala, K.K., Hwu, P., Jones, R.L., Gnjatic, S., Kim-Schulze, S., Tuballes, K., Yishak, M., Lu, H., Yakovich, A., Ter Meulen, J., Chen, M., Kenney, R.T., Bohac, C., Pollack, S.M., 2020. A Phase 1b Study Evaluating the Safety, Tolerability, and Immunogenicity of CMB305, a Lentiviral-Based Prime-Boost Vaccine Regimen, in Patients with Locally Advanced, Relapsed, or Metastatic Cancer Expressing NY-ESO-1. Oncoimmunology 9. https://doi.org/10.1080/2162402X.2020.1847846

Soneson, C., Love, M.I., Robinson, M.D., 2015. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research 4, 1521. https://doi.org/10.12688/f1000research.7563.1

Stark, R., Brown, G., 2011. DiffBind: differential binding analysis of ChIP-Seq peak data [WWW Document]. URL http://bioconductor.org/packages/release/bioc/vignettes/DiffBind/inst/doc/DiffBind.pdf

Stark, R., Grzelak, M., Hadfield, J., 2019. RNA sequencing: the teenage years. Nat. Rev. Genet. 20, 631–656. https://doi.org/10.1038/s41576-019-0150-2

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. 102, 15545–15550. https://doi.org/10.1073/pnas.0506580102

Takebayashi, Y., Pourquier, P., Zimonjic, D.B., Nakayama, K., Emmert, S., Ueda, T., Urasaki, Y., Kanzaki, A., Akiyama, S., Popescu, N., Kraemer, K.H., Pommier, Y., 2001. Antiproliferative activity of ecteinascidin 743 is dependent upon transcription-coupled nucleotide-excision repair. Nat. Med. 7, 961–966. https://doi.org/10.1038/91008

Talevich, E., Shain, A.H., Botton, T., Bastian, B.C., 2016. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLoS Comput. Biol. 12, e1004873. https://doi.org/10.1371/journal.pcbi.1004873

Tamiya, H., Sabe, H., Yamashita, K., Imura, Y., Wakamatsu, T., Takenaka, S., 2020. Eribulin Provides a Remarkable Effect in Trabectedin-Resistant Myxoid Liposarcoma. Case Rep. Orthop. 2020, 8873185. https://doi.org/10.1155/2020/8873185

Terms and Definitions – ENCODE [WWW Document], n.d. URL https://www.encodeproject.org/data-standards/terms/ (accessed 3.5.21).

van der Weegen, Y., Golan-Berman, H., Mevissen, T.E.T., Apelt, K., González-Prieto, R., Goedhart, J., Heilbrun, E.E., Vertegaal, A.C.O., van den Heuvel, D., Walter, J.C., Adar, S., Luijsterburg, M.S., 2020. The cooperative action of CSB, CSA, and UVSSA target TFIIH to DNA damage-stalled RNA polymerase II. Nat. Commun. 11, 2104. https://doi.org/10.1038/s41467-020-15903-8

Wang, L., Park, H.J., Dasari, S., Wang, S., Kocher, J.-P., Li, W., 2013. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res. 41, e74–e74. https://doi.org/10.1093/nar/gkt006

Wang, W., Xia, X., Mao, L., Wang, S., 2019. The CCAAT/Enhancer-Binding Protein

Family: Its Roles in MDSC Expansion and Function. Front. Immunol. 10. https://doi.org/10.3389/fimmu.2019.01804

Willeke, F., Ridder, R., Mechtersheimer, G., Schwarzbach, M., Duwe, A., Weitz, J., Lehnert, T., Herfarth, C., von Knebel Doeberitz, M., 1998. Analysis of FUS-CHOP fusion transcripts in different types of soft tissue liposarcoma and their diagnostic implications. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. 4, 1779–1784.

Workman, P., Aboagye, E.O., Balkwill, F., Balmain, A., Bruder, G., Chaplin, D.J., Double, J.A., Everitt, J., Farningham, D. a. H., Glennie, M.J., Kelland, L.R., Robinson, V., Stratford, I.J., Tozer, G.M., Watson, S., Wedge, S.R., Eccles, S.A., Committee of the National Cancer Research Institute, 2010. Guidelines for the welfare and use of animals in cancer research. Br. J. Cancer 102, 1555–1577. https://doi.org/10.1038/sj.bjc.6605642

Xu, C., Li, X., Liu, P., Li, M., Luo, F., 2019. Patient-derived xenograft mouse models: A high fidelity tool for individualized medicine. Oncol. Lett. 17, 3–10. https://doi.org/10.3892/ol.2018.9583

Yoshida, G.J., 2020. Applications of patient-derived tumor xenograft models and tumor organoids. J. Hematol. Oncol.J Hematol Oncol 13, 4. https://doi.org/10.1186/s13045-019-0829-z

Yu, G., Wang, L.-G., Han, Y., He, Q.-Y., 2012. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS J. Integr. Biol. 16, 284–287. https://doi.org/10.1089/omi.2011.0118

Yu, G., Wang, L.-G., He, Q.-Y., 2015. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics 31, 2382–2383. https://doi.org/10.1093/bioinformatics/btv145

Yu, J.S.E., Shane Colborne, Christopher S. Hughes, Gregg B. Morin, Torsten O. Nielsen, 2019. The FUS-DDIT3 Interactome in Myxoid Liposarcoma. Neoplasia N. Y. N 21, 740–751. https://doi.org/10.1016/j.neo.2019.05.004

Zack, T.I., Schumacher, S.E., Carter, S.L., Cherniack, A.D., Saksena, G., Tabak, B.,

Lawrence, M.S., Zhang, C.-Z., Wala, J., Mermel, C.H., Sougnez, C., Gabriel, S.B., Hernandez, B., Shen, H., Laird, P.W., Getz, G., Meyerson, M., Beroukhim, R., 2013. Pan-cancer patterns of somatic copy number alteration. Nat. Genet. 45, 1134–1140. https://doi.org/10.1038/ng.2760

Zambelli, F., Pesole, G., Pavesi, G., 2013. PscanChIP: finding over-represented transcription factor-binding site motifs and their correlations in sequences from ChIP-Seq experiments. Nucleic Acids Res. 41, W535–W543. https://doi.org/10.1093/nar/gkt448

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., Liu, X.S., 2008. Model-based Analysis of ChIP-Seq (MACS). Genome Biol. 9, R137. https://doi.org/10.1186/gb-2008-9-9-r137