Proteogenomics methods for translational cancer research

Julia Casado

Research Program in Systems Oncology, Faculty of Medicine
University of Helsinki
Finland

Academic dissertation

To be publicly discussed, with the permission of
the Faculty of Medicine of the University of Helsinki,
in Biomedicum Helsinki 1, Lecture Hall 2, Haartmaninkatu 8, Helsinki,
on 22nd April at 12:00 noon.

Helsinki 2021

Supervisor

Sampsa Hautaniemi, DTech, Professor
Systems Oncology Research Program, Medicum
Faculty of Medicine, University of Helsinki
Helsinki, Finland

Thesis follow-up committee

Juho Rousu, PhD, Professor
Computer Science Department, Aalto University
Espoo, Finland

Maciej Lalowski, PhD, Adjunct Professor
Department of Biochemistry and Developmental Biology,
University of Helsinki
Helsinki, Finland

Reviewers appointed by the Faculty

Janne Lehtiö, PhD, Professor
SciLifeLab, Karolinska Institutet
Stockholm, Sweden

Jussi Paananen, PhD, Assistant Professor
Institute of Biomedicine, University of Eastern Finland
Kuopio, Finland

Faculty representative appointed by the Faculty

Nina Peitsaro, PhD, Docent
Flow Cytometry Core Facility, University of Helsinki
Helsinki, Finland

Opponent appointed by the Faculty

Emmanuel Barillot, PhD, Head of Department
Department of Computational Systems Biology of Cancer, Institut Curie
Paris, France

No. 23/2021
ISSN 2342-3161
ISBN 978-951-51-7219-8 (paperback)
ISBN 978-951-51-7220-4 (PDF)

http://ethesis.helsinki.fi
Picaset Oy, Helsinki 2021
The Faculty of Medicine uses the Urkund system to examine all doctoral dissertations

If you try to take a cat apart to see how it works,
the first thing you have in your hands is a nonworking cat.
Douglas Adams.

Abstract

Cancer continues to be a major clinical and societal challenge. Globally, the cancer burden rises every year with a new record of 18.1 million new cases and 9.6 million cancer deaths, as reported by the World Health Organization. Despite the increased financial efforts of western countries to cure this disease, it is expected that in the year 2040, over one-third of the population will be diagnosed with cancer.

The gap in translating basic research into clinical benefit requires cross-disciplinary approaches to harness large data from the complex molecular systems and cellular organization within the tumor. The main obstacles in current cancer care are late detection and therapy resistance. While high-throughput and single-cell methodologies have become an advisable tool to analyze molecular profiles, their use for clinical decision-making is still missing.

This thesis aims to propose efficient methodologies to connect molecular and cellular profiling research to cancer therapy outcomes. In the first and second publications, we made available software to rapidly analyze large mass cytometry data with high resolution and interaction-assisted interpretation steps for the analysis. These methods allow the rapid profiling of tumor cell populations and their association with therapy response. As part of the third project, we developed new image analysis methods to identify therapy response predictors from highly multiplexed images. We found that spatial organization within the tumor microenvironment was highly associated with DNA damage genome scarring. In the fourth study, we designed a new method to identify epigenetically reversible drug resistance mechanisms in tumor cells.

The application of novel methodologies contributed to a better understanding of the roles of genomic and proteomic features in the tumor-immune microenvironment in response to modern anti-cancer therapies.

Tiivistelmä

Syöpä on edelleen merkittävä kliininen ja yhteiskunnallinen haaste. Maailmanlaajuisesti syöpätapausten määrä nousee vuosittain, uuden ennätyksen ollessa Maailman terveysjärjestö WHO:n mukaan 18,1 miljoonaa uutta tapausta ja 9,6 miljoonaa syöpäkuolemaa. Huolimatta länsimaiden lisääntyneistä taloudellisista ponnisteluista tämän taudin parantamiseksi, on odotettavissa, että vuonna 2040 yli kolmanneksella väestöstä tullaan diagnosoimaan syöpä.

Aukko perustutkimuksen muuntamisessa kliiniseksi hyödyksi vaatii monialaisia lähestymistapoja, jotta voidaan hyödyntää suuria määriä dataa kasvaimen sisäisistä monimutkaisista molekyylijärjestelmistä ja solurakenteesta. Suurimmat esteet nykyisessä syöpähoidossa ovat myöhäinen havaitseminen ja hoitoresistenssi. Vaikka suuritehoisesta tutkimuksesta, eli automatisoinnin avulla tehostetusta mittauksesta, ja yksittäisen solutason tutkimusmenetelmistä on tullut suositeltu työkalu molekyyliprofiilien analysointiin, niiden käyttö hoitopäätösten teossa puuttuu yhä.

Tämän tutkielman tarkoituksena on esittää tehokkaita menetelmiä molekyylien ja solujen profilointitutkimuksen yhdistämiseksi syöpähoitojen tuloksiin. Ensimmäisessä ja toisessa julkaisussa kehitimme ohjelmiston, jolla voidaan analysoida nopeasti suuren kokoluokan massasytometriadataa korkealla resoluutiolla ja hyödyntää interaktiivisia tulkintavaiheita analyysissä. Nämä menetelmät mahdollistavat kasvainsolukantojen nopean profiloinnin ja niiden yhdistämisen hoitovasteeseen. Kolmannen projektin osana kehitimme uusia kuva-analyysimenetelmiä hoitovastetta ennustavien markkereiden tunnistamiseksi erittäin monikanavaisista kuvista. Havaitsimme, että solujen järjestäytyminen kasvaimen mikroympäristössä oli voimakkaasti yhteydessä DNA-vaurion aiheuttamaan genomin arpeutumiseen. Neljännessä tutkimuksessa suunnittelimme uuden menetelmän epigeneettisesti kumottavien lääkeresistenssimekanismien tunnistamiseksi kasvainsoluissa.

Uusien tutkimusmenetelmien soveltaminen johti parempaan ymmärrykseen genomin ja proteomin ominaisuuksien roolista kasvaimen immuunimikroympäristössä ja niiden merkityksestä nykyaikaisten syöpähoitojen vasteeseen.

Resumen

El cáncer sigue representando un gran desafío clínico y social. A nivel mundial, la incidencia aumenta cada año con un nuevo récord de 18,1 millones de casos nuevos y 9,6 millones de muertes por cáncer, según datos de la Organización Mundial de la Salud. A pesar del aumento del esfuerzo económico para curar esta enfermedad, se espera que en el año 2040, más de un tercio de la población sea diagnosticada con cáncer.

La traslación de los descubrimientos fundamentales en beneficio clínico requiere enfoques interdisciplinarios para sacar provecho de la gran cantidad de información, desde complejos sistemas moleculares, a la organización celular dentro de cada tumor. Los obstáculos principales en el tratamiento actual del cáncer son la detección tardía y la resistencia al tratamiento. Aunque las metodologías de alto rendimiento y resolución unicelular se han convertido en una herramienta recomendable para analizar perfiles moleculares, todavía no hay aplicación para la toma de decisiones clínicas.

Esta tesis tiene como objetivo proponer metodologías eficientes para conectar la investigación molecular y celular con la efectividad de las terapias anti-cáncer. En las publicaciones primera y segunda, desarrollamos software de código abierto para analizar rápidamente datos de citometría de masas de forma interactiva. Los dos métodos permiten analizar las poblaciones de células en el tumor y su asociación con la respuesta a la terapia. Como parte del tercer proyecto, desarrollamos nuevos métodos de análisis de imágenes para identificar predictores de respuesta a la terapia a partir de imágenes multiplexadas de antígenos múltiples. Descubrimos la organización celular dentro del microentorno tumoral asociada con la ruptura y reparación del ADN, un predictor de la respuesta al tratamiento. En el cuarto estudio, diseñamos un nuevo método para identificar mecanismos de resistencia a quimioterapia que son epigenéticamente reversibles.

La aplicación de nuevas metodologías presentadas en esta tesis ha contribuido al conocimiento y comprensión de la función de las características genómicas y proteómicas en el microentorno tumoral e inmune en respuesta a las terapias modernas contra el cáncer.

Contents

Abbreviations

AUC	Area under the curve
BRDi	Bromodomain protein inhibitor
CO2	Carbon dioxide
CSV	Comma sepparated values
CyTOF	Mass cytometry by time-of-flight
DLBCL	Diffuse large B-cell lymphoma
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DNMTi	DNA methyltransferase inhibitor
DSB	Double strand breaks
DSRT	Drug sensitivity and resistance testing
FACS	Fluorescence-activated cell sorting
FCS	Flow cytometry standard
FFPE	Formalin-fixed paraffin-embedded
FPKM	Fragments per kilobase million
HATi	Histone acetyltransferase inhibitor
HDACi	Histone deacetylase inhibitors
HDMi	Histone demethylase inhibitor
HGSOC	High-grade serous ovarian cancer
HMTi	Histone methyltransferase inhibitor
HR	Homologous recombination
HRD	Homologous recombination deficiency
IC50	Concentration for 50% inhibition
IF	Immunofluorescence
LOH	Loss of heterozygosity
MAF	Minor allele frequency
MDS	Multidimensional scaling
MFI	Median fluorescence intensity
mRNA	messenger RNA
MST	Minimum spanning tree
NHEJ	Non-homologous end joining
NRS	Non-redundancy score
PARP1	Poly ADP-ribose polymerase 1
PBMC	Peripheral Blood Mononuclear Cells
PBS	Phosphate-buffered saline
PD-1	Programmed Death 1
PD-L1	Programmed death-ligand 1
PFI	Platinum free interval
PFS	Progression free survival
qSNE	quadratic rate t-SNE optimizer
R-CHOP	Rituximab, Cyclophosphamide, Doxorubicin Hydrochloride (Hydroxydaunomycin), Vincristine Sulfate (Oncovin), and Prednisone
RNA	Ribonucleic acid
RNA-seq	RNA sequencing

tCyCIF	Tissue-based cyclic immunofluorescence
tSNE	t-distributed stochastic neighbor embedding
UMAP	Uniform Manifold Approximation and Projection
VAF	Variant allele frequency
WES	Whole exome sequencing

Publications and author's contributions

Publication I **Casado J**, Lehtonen O, Rantanen V, Kaipio K, Pasquini L, Häkkinen A, Petrucci E, Carpen O, Biffoni M, Farkkila A and Hautaniemi S.
Agile workflow for interactive analysis of mass cytometry data.
Bioinformatics, 2020, doi:10.1093/bioinformatics/btaa946

Publication II Häkkinen A, Koiranen J, **Casado J**, Kaipio K, Lehtonen O, Petrucci E, Hynninen J, Hietanen S, Carpen O, Pasquini L, Biffoni M, Lehtonen R and Hautaniemi S.
qSNE: Quadratic rate t-SNE optimizer with automatic parameter tuning for large data sets.
Bioinformatics, 2020, doi:10.1093/bioinformatics/btaa637

Publication III Farkkila A, Gulhan D **/ Casado J**, Jacobson C, Nguyen H., Koruchupakkal B, Maliga Z, Yapp C, Chen YA, Schapiro D, Zhou Y, Graham J, Dezube B, Munster P, Santagata P, Garcia E, Rodig S, Lako A, Chowdhury D, Shapiro G, Matulonis U, Park P, Hautaniemi S, Sorger P, Swisher E, and D'Andrea AD / Konstantinopoulos P.
Immunogenomic profiling determines responses to combined PARP and PD-1 inhibition in ovarian cancer.
Nature Communications, 2020. doi:10.1038/s41467-020-15315-8

Publication IV Facciotto C **/ Casado J**, Turunen L, Leivonen SK, Tumiatti M, Rantanen V, Kauppi L, Lehtonen R, Wennerberg K, Leppa S and Hautaniemi S.
Drug screening approach combines epigenetic sensitization with immunochemotherapy in cancer.
Clinical Epigenetics 11, 192 (2019). doi:10.1186/s13148-019-0781-3

/ equal contribution.

Author's contributions

Publication I	Conceptualized the methodology and designed the system. Developed the user interface (together with VR), the cytometry analysis pipeline and individual logic components (together with OL), performed the data analysis, and wrote the paper.
Publication II	Designed and implemented data processing steps for the high-grade serous ovarian cancer case study on matched chemotherapy exposed and naive CyTOF measurements.
Publication III	Designed and implemented the workflow to process multiplexed immunofluorescence images, developed the quality assessment, cell type calling, and spatial neighborhood algorithms, and performed the single cell data analysis.
Publication IV	Designed the study, conceptualized and performed the epigenetic drug screening experiments, wrote the paper (all together with CF), and analyzed the DNA-damage immunofluorescence image data and whole exome sequencing data.

1 Introduction

The prevalence or demise of any species depends on how fit it is to thrive and reproduce in its circumstances [1]. The fitness of a species is determined by the fitness of its members and the result of their actions. In turn, the fitness of each individual is affected by the functioning and interplay of the organs and functional parts that make up these individuals; all the way down to the cellular and molecular complexes that build these parts. In a predator-prey scenario, the circumstances test both the fitness of the individuals involved and that of their pack or herd. For instance, previous access to resources helps one grow stronger than the other, and collective behaviors, either cooperative or competitive, also factor in the individual's final fate or even its species. Similarly, human diseases test our fitness, in the form of the ability of our bodies to detect, correct, and recover from any calamity that challenges our health. A cell's fitness can be described by how well a cell performs its function, and the function often entails interacting with other cells or with external stimuli [2]. The proper or flawed functioning of a cell in an environment is determined by the proteins performing its function and structure. Proteins are encoded in the DNA, but their production is regulated by a complex program that we summarize as epigenetics. This program responds to proteins inside and outside of the cell, as well as RNA and other molecules [3].

A disease such as cancer is strongly determined by aberrant DNA, which causes aberrant functioning of the cell [4]. The newly aberrant cell does not follow the same collective behaviors as the other cells, thus challenging the fitness of the whole system. Now, the normal cells' ability to communicate correctly with each other and work against the cancer cells will determine the prevalence or demise of this new aberrant cell. For instance, immune cells can be recruited by cancer cells to help the tumor grow without turning them into cancer cells, but also immune cells are often very successful at identifying and terminating tumor cells. If a tumor cell is more advanced, it can evade the terminating signals. They can rapidly become more advanced due to unlimited replication, which launches an accelerated evolutionary process by which the fittest cells, either by their DNA changes or epigenetic programs, proliferate despite our defense mechanisms. Tumor evolution produces multiple subtypes of tumor cells, making it even more difficult for the immune system and the medical treatment to kill all of them. This heterogeneous mixture of tumor cells makes up the tumor microenvironment together with infiltrating and recruited immune and other normal cells. It has been referred to as a battlefield in popular literature [5].

Fortunately, technological advances in molecular biology in the past two decades have shed light on both sides' inner workings, the tumor cells, and the immune cells,

and provided tools to study their interplay on human cancer specimens straight from the operation theatre [6, 7]. Deep sequencing of thousands of cancer genomes and transcriptomes has made the blueprint of the potential tumor cells' capabilities that lead to many successful drugs being developed. As a result, the overall survival of cancer patients has doubled since 1970 [8]. However, in this process, we have learned that finding general or targeted treatments to kill the tumor cells is not enough [9]. Thus, we need smart drug combinations and strategic scheduling, as well as treatments that take advantage of the body's own defense systems. Research on predictive biomarkers has developed efficient systems to stratify patients by the expected response to treatment [10]. Eradicating this disease will require large team efforts to efficiently share data and collaborate to translate results into patient benefits [11].

While translational research plays two important roles, the first one being placed between basic science and clinical research, and the second being the adoption of findings from clinical research into practice, both parts of this process are necessary [12]. The latter one is in the hands of multiple stakeholders such as policymakers, investors, and citizens. The first one involves large networks of cross-disciplinary international collaborations. This thesis aims to help smooth such collaborations by redesigning the steps where data translation has been one of the bottlenecks.

In this thesis, we use the name *translational steps* as the building blocks of translational cancer research [13]. The keyword *translation* implies that we must be translating from some origin domain to some target domain. In translational research, the first domain consists of basic research findings, while the latter one corresponds to the benefit these findings effect on patients' lives. Hence, translational research is defined as the process of exchanging knowledge between the laboratory bench and clinical setting with the main aim of clinical applications [14]. For example, while several studies were needed to identify endogenous DNA damage as a potential weakness to target cancer cells [15], it took a large body of research to develop compounds that could safely inhibit key elements on DNA repair mechanisms and hinder the cells' ability to repair DNA damage [16, 17].

The original contribution of this thesis work is in the form of effective methodologies to translate between molecular biology, computational science, and clinical setting. The thesis describes the new methods and their application to study drug response in human cancers. Publication I combines single cell cytometry analyses with agile principles to design a generalized pipeline that can answer the most common questions on cytometry experiments such as population abundance, detection of rare cell populations, and expression on different cell types within the same tumor microenvironment. Publication II presents a method that tackles the challenge of visualizing large cytometry datasets. Both, Publication I and II,

show the benefit of this type of analysis in fresh tumor samples and explain how to identify rare but important tumor cells with stem-like expression associated with a short time to progression. Publication III is a large cross-disciplinary effort to translate findings from a clinical trial back into the biomarker discovery channels. The clinical trial assessed the benefit of combining immunotherapy and targeted therapy against DNA repair mechanisms of the cells. This thesis presents novel image analysis methods to resolve the cellular organization of tumor and immune cells in synergy with genomic features and treatment response. In Publication IV, we developed an experimental protocol to test epigenetic reprogramming options with standard laboratory robotics. We demonstrate the utility of this protocol by reverting drug resistance *in vitro* in a set of lymphoma cells. Taken together, we report new cross-disciplinary methods and discoveries in the field of translational cancer research.

2 Background

Cancer is a cell disease characterized by uncontrolled growth. This uncontrolled growth is enabled by a series of mechanisms that allow the cells to malfunction, and to eventually cause fatal failure of vital organs [18]. This chapter starts with a short primer on cancer biology and broad cellular mechanisms that are discussed later in this thesis, followed by the technologies and methodologies utilized to study the cellular and molecular composition of human cancers.

2.1 Introduction to cancer biology

Two decades ago, Doctors Hanahan and Weinberg compiled the body of molecular cancer biology knowledge into a framework of six biological capabilities developed during tumorigenesis [18]. This framework, called the *Hallmarks of Cancer*, consisted of sustained proliferative signaling, evasion of growth suppressing signals, resisting cell death, limitless cell replication, angiogenesis, and activation of invasive and metastasis capabilities. The hallmarks became a guiding beacon for today's cancer biology research.

The second edition [19, 4] incorporated four new hallmarks. Two emerging hallmarks: reprogramming of energy metabolism allows the cells to survive in overcrowded environments, evading immune response by tricking the immune system to not target tumor cells; and two enabling hallmarks: tumor-promoting inflammation, and genomic instability. On one hand, immune cells infiltrate the cancerous or pre-cancerous lesion as they do for wound healing, but their presence inadvertently aids tumor progression by providing molecules that enable new hallmark capabilities. On the other hand, genomic instability is necessary to enable cancer cells to acquire multiple hallmarks capabilities. While epigenetic reprogramming due to environmental factors can enable hallmark capabilities, current knowledge points to genomic instability as the most prominent step that further accelerates the appearance of random genetic changes that can enable further hallmark capabilities. This goes in line with the idea that cancer arises from an unfortunate sequence of mutations and genomic changes [20].

The complexity of this disease stems from multiple levels of heterogeneity. Previously defined as inter- and intra-tumor heterogeneity, were recently redefined to address the different levels at which we can differentiate tumors; first, morphological differences help find subtypes within a cancer type; second, within a tumor type we can observe different clinical responses; third, molecular heterogeneity shows that even within the same subtypes, each tumor has its own set of genetic, epigenetic, and

immune features; and fourth, the tumor cell heterogeneity describes the subclones resulting from darwinian evolution [21].

The discovery of the conflicting role of immune cells to aid and attack tumor cells [22], together with the advancements of single-cell technologies opened a new door into the complexities of the molecular heterogeneity level. To eradicate this disease and cure all patients we must learn how all the moving parts interact as cancer manifests on each patient as its own complex system [9]. The following chapters will explain the mechanisms and interactions tackled in this thesis.

2.1.1 Epigenetic reprogramming

The epigenome refers to all the cell's processes that regulate gene expression, and therefore cell identity [23]. These processes are the reason that cells with identical DNA perform very different functions, and are affected by previous cell states, as well as environmental cues [2]. The mode of action of the epigenetic program is intrinsically related to the three-dimensional packaging of the DNA molecules (Figure 1). The accessibility to the DNA depends on how tightly parts of the DNA

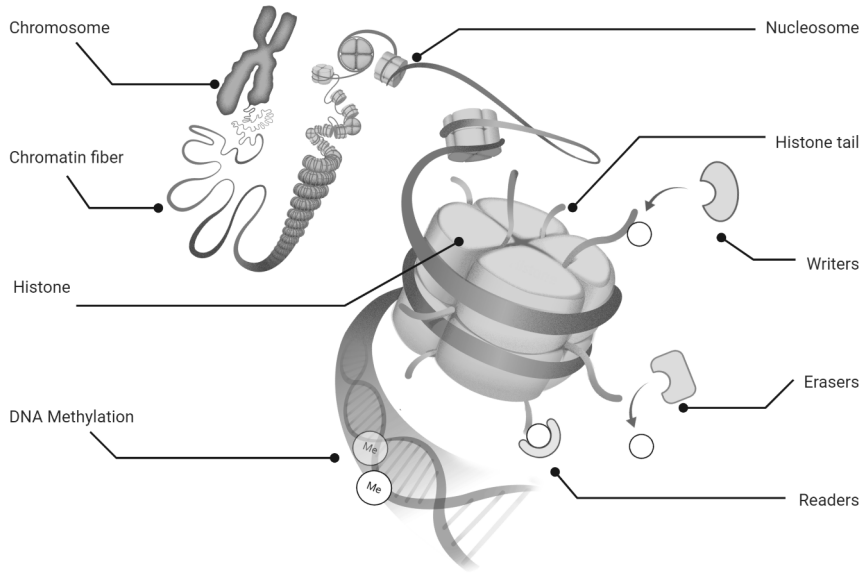

Figure 1: The DNA is wrapped around histone complexes called nucleosomes, the state of the histone complexes as well as the presence of covalently bonded methyl molecules to the DNA regulate how loose or tight the nucleosomes are organized. *Image adapted from zenithepigenetics.com, copyright of Richard E. Ballermann.*

are wrapped around the *histone* proteins. Hence, the epigenome is comprised of molecular tags, called epigenetic marks, that bind to the DNA as well as to the histones, applying pressure to wrap the DNA more loosely or tightly, and thus rendering some regions of the DNA unreadable for the gene expression machinery [24].

The epigenetic marks are set in place and removed by a family of epigenetic enzymes, that can be further subdivided into epigenetic readers, writers and erasers. The dysregulation or reprogramming of the epigenome can thus change the cell identity without genomic changes. In cancer epigenomes, the expression of the enzymes as well as the resulting epigenetic marks are disrupted from their normal function [25].

New pharmacological developments have allowed a plethora of epigenetic inhibitors. That is, drugs that target the epigenetic enzymes so they cannot modify their corresponding epigenetic tags. As the epigenome is still to be fully understood, these inhibitors yield new leads to study causal effect between epigenetic enzymes and measurable gene expression. Epigenetic inhibitors have been tested in clinical trials as monotherapy, however, they have been recently shown to reprogram the cancer cells to a drug-sensitive state [26].

2.1.2 DNA breaks and DNA repair in cancer

Errors in the replication process of DNA molecules can happen as well as damage caused by external events [2]. However, DNA, being the blueprint of the individual, is protected by multiple redundant care-taking systems. DNA integrity is regularly checked, and if a copy turns out to be corrupt, several alarms and protocols are set off [27]. If the damage is harmless for the cell, it might not trigger any further action; if it is deemed to be fixable, the repair mechanisms will try to repair it; and finally, if it is damaged beyond repair, an auto-destruction system, called apoptosis, is started and the cell dies. The correct functioning of these systems keeps the accumulation of mutations and mutant cells to a minimum.

Mutagens can be both endogenous to the host and external, such as radiation, smoke, and viruses. While cells are regularly exposed to mutagens, tumor cells undergo higher mutation rate due to replication stress and evolve quickly avoid apoptosis signaling and maintain a proliferative state despite clear DNA aberrations [28].

DNA damage can occur to a single strand or to both strands of the DNA double helix. The latter one, called double strand breaks (DSB), is critical for the cell [29] and commonly caused by chemotherapy. Hence, repair of DSB directly effects drug response. Figure 2 shows the repair mechanisms that activate when DSB

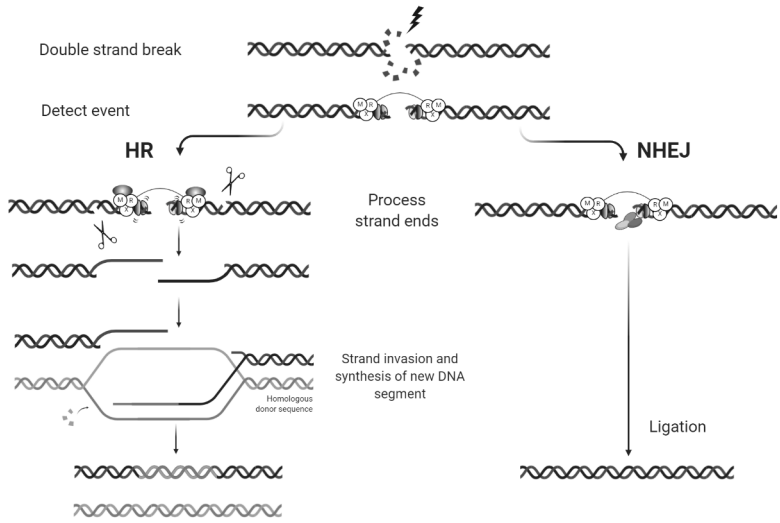

Figure 2: Double DNA strand breaks are detected by *sensor* proteins. When following HR repair pathway, the ends of the 5' strand are cut to leave the 3' strand free. The loose strand then invades an available homologous sequence to synthesise the new DNA segment. When following NHEJ, the strand ends are not cut, instead, the repair machinery will attempt a ligation that may introduce small deletions or insertions to correct for loose nucleotides.

are detected have two main types: homologous recombination (HR) and non-homologous end-joining (NHEJ)[30]. HR driven repair is a mechanism that uses the homologous chromosome as a template to assemble one strand of the missing segment, allowing the broken part to look the same as the template. NHEJ system simply connects the ends of the broken strands, if some nucleotides were lost, this repair mechanism will lead to loss of integrity of the DNA. The DNA repair machinery comprises multiple proteins that work together to restore the original state to the genome, however, if these proteins are (epi)genetically inhibited or their function is compromised by mutations, the quality of the repairs is affected. For example, homologous recombination deficiency (HRD) has been shown to cause loss of heterozigosity (LOH)[31].

When DSB events are detected, protein complexes act as *mediators* to recruit *effector* proteins that will repair the damage. While the cell has multiple redundant mechanisms, the mutations in tumor cells genome often disrupt some of these proteins, forcing the cell to rely on fewer care-taking options. A successful example in the HR pathway is the dependency of having a functional BRCA protein or a functional PARP, however, when both are disrupted it was shown to render the cell unable to repair DSB [17, 16]. This weakness has been successfully exploited by

the prescription of PARP inhibitor therapy to patients with existing *BRCA* mutations [28].

2.1.3 Elements within the tumor microenvironment

In contrast to a traditional cancer-centric view of autonomous mutant cells replicating and invading the patient's organs, recent advances have shown the tumor development to be affected by a multitude of normal cells recruited to the site [32]. The assemblage of cells inside and around the tumor is called *tumor microenvironment* (Figure 3), and the collective function or dysfunction of the cells must be understood to turn key players into an actionable weakness [33]. Over the last decade, it has developed into a large research field on itself.

Advances in preclinical models, microscopy, and cytometry, have identified seemingly normal cells recruited by the tumor and tricked into promoting tumorigenesis [22]. The cells can be classified in three groups, each being actively developed in a

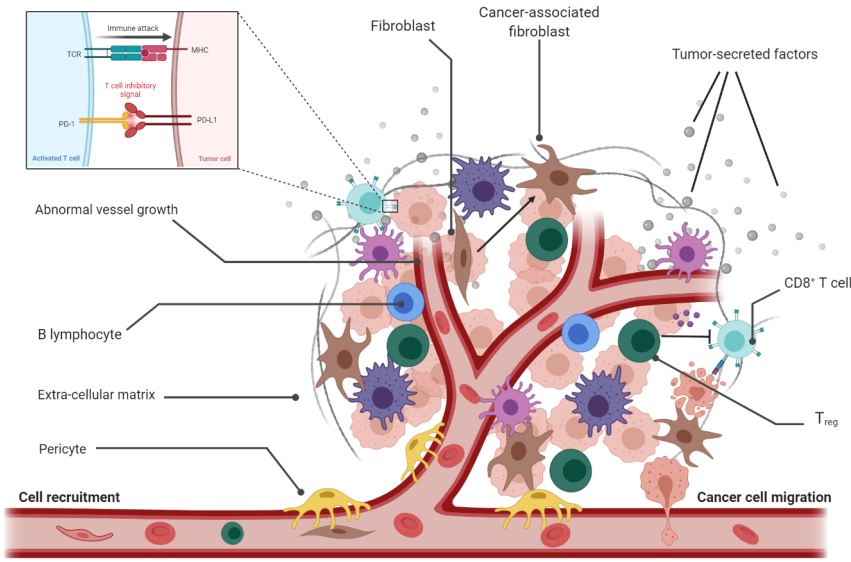

Figure 3: The tumor microenvironment is largely shaped by secreted factors that allow cells to send signals to neighboring cells. The figure shows main three cell states of cancer cells (light pink) as epithelial, apoptotic, and undergoing transition to mesenchymal (EMT); the stromal compartment includes endothelial structure, pericytes (yellow), and fibroblasts (brown); the immune compartment involves an interplay between dendritic cells (pink), macrophages (purple), B cells (blue), and T tells (teal). The close up (top left) shows the interface between a PD-L1+ positive tumor cell and a PD1+ CD8+ T cell.

race to turn them against the tumor. The first and most widely known group, are endothelial cells and pericytes, that are necessary recruits to build nutrient supply routes in a process called angiogenesis. During tumorigenesis, the tumor may secrete growth factors to trick the host into building new blood and lymphatic vessels for the new tumor cells in need of resources. The second group are fibroblasts, also part of the stromal compartment like the previous group, however, when normal fibroblasts are recruited into cancer-associated fibroblasts (CAFs) they can support tumor growth and metastasis [33]. The third group are the infiltrating immune cells; they can be divided into multiple subtypes and new findings about their pro- and anti-tumorigenic functions are being discovered every year [34]. The complexity of this field is exacerbated by the notion that the composition of the microenvironment and the function of these cells varies in function of the anatomical site, types of tumor cells present, and therapeutic interventions [35]. While immunology is a large and specialized field that will yield new opportunities on anti-cancer treatments, in this thesis we focus on the anti-tumor response by T cells as studied by clinical trials.

The promise of immunotherapies, drug compounds that help the immune system better detect and kill cancer cells, has had varying degree of success in different cancer types [36, 37, 38]. However, many different immunotherapeutic approaches exist today, and new studies must now find out when, where, and how they are beneficial [39].

2.2 Diffuse large B-cell lymphoma

B-cell lymphomas are a disease of the blood cells. Lymphomas are commonly classified into Hodgkins and non-Hodgkins lymphomas. In this thesis, we focus on the drug resistant phenotype of the most common subtype of non-Hodgkins lymphoma, diffuse large B-cell lymphoma (DLBCL) [40]. In the past decades the standard of care for DLBCL has improved patient prognosis significantly, to the point that approximately 70% of the patients can be cured with the standard treatment [41, 42]. The standard of care for DLBCL is R-CHOP, a combination of five drugs: rituximab, cyclophosphamide, doxorubicin, vincristine, and prednisone, that is effective in 60% of the cases [43].

Relapse patients have a dire prognosis and depend on the chances of being eligible candidates for clinical trials that are recruiting at the time. Identifying the mechanisms that lymphoma cells utilize to evade the most toxic compounds in R-CHOP is the next step to propose options for the patients that do not respond to R-CHOP [44]. Recent results in drug synergy and preclinical models points out that the R-CHOP combination does not have synergistic effects and points out to

solutions for drug testing [45]. In this line, emerging anti-cancer therapies targeting epigenetic mechanisms [46] rather than direct DNA damage have shown promising results in various lymphomas [47, 48].

2.3 High-grade serous ovarian cancer

Ovarian cancer is the 7^{th} leading cancer type in incidence rankings and the 8^{th} in cancer fatalities among women [49]. High-grade serous ovarian cancer (HGSOC) is the most aggressive subtype of ovarian cancer. While most patients respond positively to a primary therapy with platinum-based chemotherapy (a DNA damaging agent), the 5-year survival rate is still below 50% due to relapse and chemotherapy resistance [50]. The standard of care in ovarian cancer is primary debulking surgery (PDS) followed by platinum-based chemotherapy when the diagnosis is such that debulking is possible. When complete removal of the tumor is not possible, platinum-based chemotherapy is then the first line of treatment [51].

HGSOC is a cancer of epithelial cells that is characterized by genomic instability [52]. Importantly, half of the HGSOC tumors suffer from homologous recombination deficiency, due to mutations or epigenetic silencing of caretaker genes *TP53*, *BRCA1* and *BRCA2*, and other HR pathway members [53]. HR deficiency can now be exploited by poly(ADP-ribose) polymerase (PARP) inhibitors that disables the cells ability to recover from DNA-damage [54].

Previous research identified patients that, after a relapse from platinum, could be treated with new therapy combinations targeting DNA repair pathways, anti-tumor immune response, or tumor heterogeneity. Large international studies yielded data to bring back into the biomarker discovery pipeline [55, 56].

2.4 From preclinical models to clinical trials

Before a discovery can be translated into new treatment opportunities for patients, it must follow a rigorous roadmap [57]. Identifying a druggable target requires understanding of the mechanisms that the target sets in motion, and assessing the schedule strategies as well as possible combinations. The validation process involves *in-vitro* and *in-vivo* assays in preclinical models. A preclinical model is an experimental setting with live cells or organisms that are suitable to test a discovery due to shared traits to human physiology and fast development to produce timely results. Examples of models are mice, fruit flies, zebra fish, and human-derived cells in different culture conditions [58, 59, 60].

Biological samples associated to therapy response information are invaluable to close the bidirectional translational gap in cancer research (Figure 4). Samples

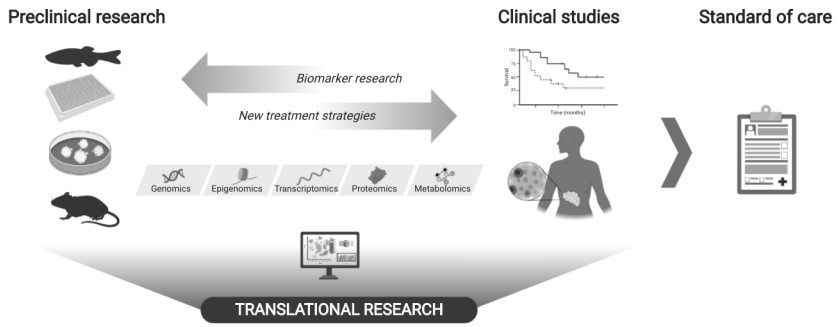

Figure 4: The translational road from *"bed to bench and back"*. Preclinical research findings help guide novel therapeutic options. On the other opposite direction, data from clinical trials, prospective studies, biobanks and other human-centered studies, are indispensable inputs for biomarker discovery.

are obtained with previous consent from clinical trial patients or from standard of care clinical setting, and enable deep qualitative and quantitative research into the mechanisms that explain therapy response. A huge amount of biomedical data in the form of clinical variables and follow up information and clinical samples have an invaluable yield of data in genetic sequences and an uncountable amount of experiments [61, 62, 63]. Furthermore, clinical samples are often deposited in national biobanks, enabling future research on questions we cannot yet consider [64].

Vast amount of data are not manageable with traditional manual analysis and require very technical bioinformatic tools for the analysis and cross-disciplinary teams for the interpretation [65]. These advances have improved overall prognosis in cancer patients, but the remaining pieces of the puzzle to completely cure current drug resistant cancer are still missing [9].

2.5 Biological data acquisition

This chapter describes the technologies that were used to produce the biological data at the core of this thesis.

2.5.1 Genomics and transcriptomics

The study of complete genomes and transcriptomes from biological samples is done from the sequence of nucleotides that make up the DNA and RNA molecules. There

are many available technologies to extract the sequences and quantify molecules [6]. In the research setting, they have become so widely used that caused a bioinformatics revolution [66]. The sample preparation process varies, but the general steps consist of breaking previously isolated DNA or RNA molecules into smaller fragments to then create multiple copies of each molecule that are then sequenced by the sequencer.

The sequencer produces an enormous amount of *reads*, the nucleotide sequences from each fragment. To construct a digital representation of the DNA or RNA we must first align the reads. The field of bioinformatics has produced produced a plethora of tools and methodologies to produce biological insight from omics data as well as tested best practices [67, 68].

Once the reads are aligned we can produce multiple layers of information. By the RNA reads we can characterize current transcriptional state of the cells, or look at RNA splicing events that can disrupt the correct functioning of the cells. By the DNA reads we can characterize chromosomal rearrangements that produce fusion genes, copy number alterations, as well as deletions and insertions of one or more single nucleotides. Information from these extracted features are stored in public databases that help international efforts at associating them to cell mechanisms and therapy responses [69]. The field of genomics was able to associate mutational patterns to known mutagens, hence providing an edge to find exploitable mutational signatures in different cancers [70].

These techniques have offered a new understanding of cancer [71]. However, so far these studies have been performed on bulk samples, and miss the detail of the heterogeneity within the tumor microenvironment [72]. Bioinformatics methods to solve this problem, and technologies such as single cell sequencing, belong to the current efforts of the community towards to study the cell to cell variation [72].

2.5.2 Single cell proteomics in cytometry

Proteins are the ultimate molecules that perform cell function [2]; however, analysis and quantification of protein molecules from individual cells pose challenges different from DNA or RNA. Recent advances point in the direction that we can expect to see soon full single-cell proteomes at the level at which we measure single-cell transcriptomes [73]. At the moment, the most widely used technology is mass cytometry time-of-flight (CyTOF) [74]. CyTOF couples cell sorting microfluidics from flow cytometry with heavy metal tags measured by mass spectrometry.

Figure 5 describes how this technology works. Metal-conjugated antibodies bind to specific proteins inside and outside the cell, cells tagged with antibodies are shot

one by one into a nebulizer that breaks the cell into a particle cloud. The particle clouds are send one by one into a mass spectrometer tunnel, when the metal atoms land on the detector, the time-of-flight is used to calculate the mass of the atom. Because we know the exact mass of each metal used as a tag, we can calculate how much of each antibody was originally present before the cell was fragmented. The final output is the single cell data table, where the rows depict cells that were shot one by one into the machine, and the columns depict the metal-conjugated antibodies.

CyTOF enables scientists to measure intensity of protein expression for each cell as long as we have validated antibodies. This means hundreds of validated antibodies, and millions of cells. The size of the datasets obtained from CyTOF is vertical, i.e. more data points than variables, compared to a genomics or transcriptomics dataset that have horizontal shape with more variables (genes and mutations) than data points (samples). Because of this, many new bioinformatics tools have been developed for this specific problem [75]. While data acquisition standards have been reached to improve reproducibility, data analysis practices are reported in varying degrees of detail and often include human manipulation of the data and specialized users, making this an obstacle towards clinical use [76].

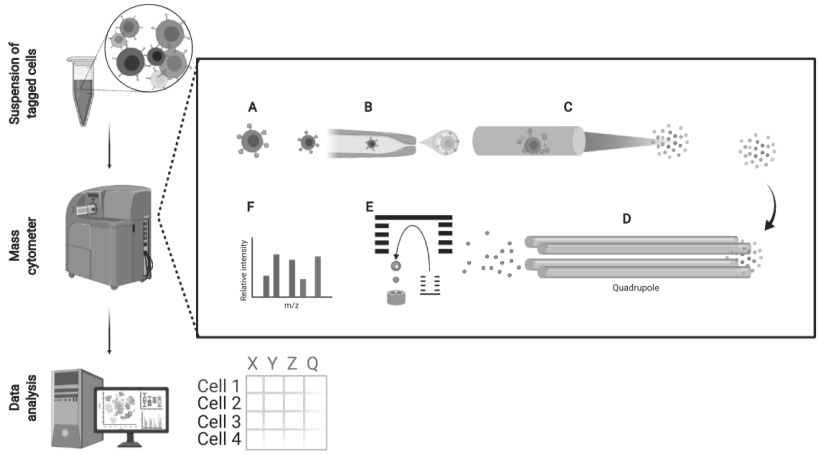

Figure 5: Mass cytometry by time of flight (CyTOF). (A) Cells are stained with antibodies previously coupled with metal tags. Using a fluidics system (B), cells are processed one by one, and shot into the nebulizer (C) to fragment the cells into particle clouds. Each cloud goes into a time-of-flight tunnel (D) to measure the time each atom takes to reach the detector (E). The mass spectra (F) are converted into single cell data with the intensity for each antibody.

Immunofluorescence imaging techniques

Before antibodies were coupled with heavy metals, scientists used fluorophore-conjugated antibodies for decades to measure protein expression intensity on cells directly on their physical location in the tissue or plate [77]. Fluorophores are molecules that are excited by specific light wavelengths. A fluorescence-based microscope excites the sample with a laser of predefined wavelength at the same time that the scanner takes the picture. This way, we see light where the excited molecules were. The microscope then takes one image for each of the lasers (Figure 6.1-2). For large samples, multiple pictures are taken, and they need to be aligned the edges of each plane to compose the bigger picture. This process is called stitching. After stitching, we could analyse each channel image separately, or merge them depending on the research goal behind the image. Often we wish to quantify the cells in the image and measure the intensity that each channel, i.e. antibody, and for that we must define which pixel belongs to each cell via segmentation. Segmentation is a computer vision technique to segment the image into smaller regions, many techniques exist to segment round cells that are separated from each other and have similar size, however tissue images have dense cell distribution with overlapping cells of different sizes and shapes [78]. A regular IF microscope can measure 4-7

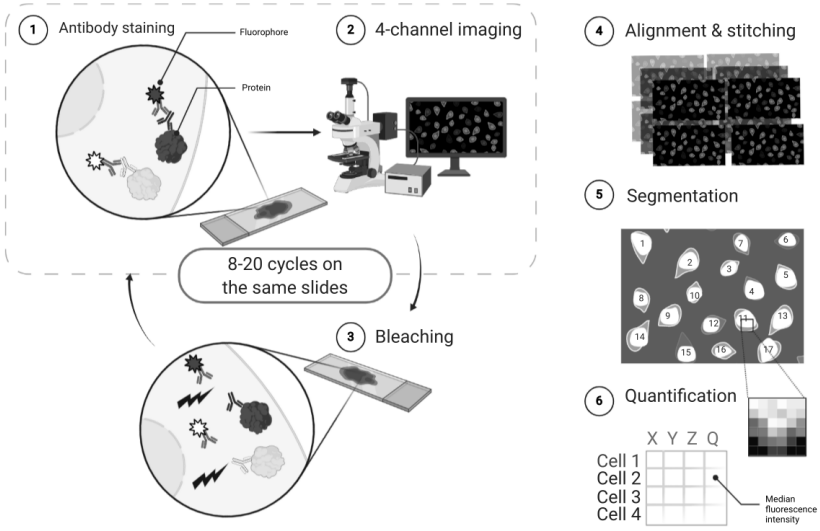

Figure 6: tCyCIF analysis. In each cycle, whole tissue slides are stained, scanned and bleached to produce images for 4 antibodies, for up to 20 cycles depending on tissue integrity. The dashed line shows classical immunofluorescence via a single cycle of staining and scanning. The images are then stitched and aligned across cycles prior to regular image processing to quantify the channel intensity for each cell.

fluorophores for each cell. This thesis also used images from a highly-multiplexed technique called cyclic tissue-based immunofluorescence (tCyCIF) [79] described in Figure 6. Here, at each cycle we scan a different set of antibodies with the same fluorophores. Therefore we can measure up to 50 proteins from a whole slide of tissue in the microscope. Other multiplexed techniques have been developed in the last decade creating a need for bioinformatic analysis methodologies to effectively translate these valuable images into knowledge [80, 81, 82].

2.5.3 High-throughput drug screening

Among other pre-clinical models for drug testing, *in vitro* drug testing (see Figure 7) takes advantage of patient-derived cell lines [83]. By dispensing the same amount of cells in each well of the plate, we can test how they respond to different drugs and doses. *Dose-response analysis* translates experimental read outs like cell inhibition or viability into a relationship between drug dosage and effect. Traditional measurements, such as the dose that kills 50% of the cells (IC50) or the area under the dose-response curve (AUC) are then used to compare and rank drugs by effect [84].

Drug screening technologies are also useful to evaluate the effects of combinations of drugs [85]. The concept of drug synergy is based on the idea that drugs inhibiting pathways of the same cell function will be more cytotoxic due to leaving the cells with fewer options to maintain a given cell function. Drugs that target independent cell functions are also effective but allow the cell to use back up pathways to survive [86]. By testing all combinations of doses between two treatment options we can

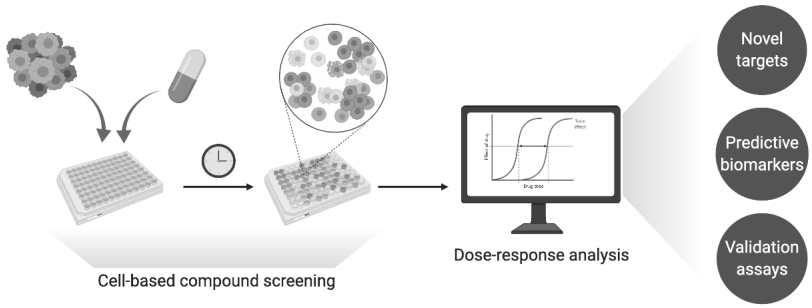

Figure 7: Patient derived cell lines, either primary or immortalized, are dispensed on micro plates coated with increasing doses of the drugs. With a luminance scanner and a reagent of choice we measure the cells alive on each well. Bioinformatics analysis translates these findings into the next translational step.

find how the cells respond to the combined effect of the drugs and which are optimal doses. Further improvements of drug screening methods have been proposed for personalized medicine settings where a patient is not treated until the results from a multidrug screening are at the hands of the clinician [87].

The number of wells and the number of plates can be increased as much as our infrastructure allows. Research institutes often invest on high throughput facilities that use highly specialized robotics to handle the plates and dispense accurate amounts of cells and compounds [88]. This technology has matured to a point where software to analyse and interpret drug resistance and sensitivity tests (DSRT) is usable for non-technical users [89].

2.6 Reproducible data analysis and visualization

Once we have produced and processed data from multiple techniques, we need to translate it into useful information and new leads for future research. The way to achieve this from such complex datasets is through computational methods [90]. With the aid of computers, we can look at data from different points of view. While formal classification of machine learning and data mining methods exists, in computational biology the two fields merge and its clearer to describe them by the applications of groups of methods, even if some methods are hierarchically part of another family of methods. The key computational methods applied in this research are:

Unsupervised clustering. A flexible tool for explorative analysis, clustering algorithms help highlight patterns and find subsets of the data that share similarities.

Supervised classification and prediction. It is a family of supervised algorithms that instead of highlighting the most clear patterns in the data, they try to find features that would best predict a class label (e.g. response to a new therapy) or a continuous value (e.g. gene expression).

Summary descriptive statistic. Although statistics are an inherent part of many supervised and unsupervised learning methods, we often use them on their own to extract features that describe groups.

Inferential statistic. Although statistics offers a rich variety of models to help answer questions from data, only a few of them are commonly applied in bioinformatics. For example, even if visually we are able to assess if a variable is clearly different in different subsets of the data, statistics helps give a numeric value to the confidence that that difference is not due to chance.

Dimensionality reduction. Lastly, this is a family of methods that transforms a set of multidimensional data points onto a lower-dimension space that maintains feature of interest, e.g. variance or similarity. The most popular dimensionality reduction methods in single cell data analysis are t-distributed stochastic neighbor embedding (tSNE) [91] and uniform manifold approximation and projection (UMAP) [92], where the feature that we aim to maintain is the distribution of the distances between points that are similar to each other.

After performing an analysis step, the researchers involved in an experiment need to interpret the results and find useful conclusions to pursue their primary question. This often means designing new analysis or new experiments. While intermediate conclusions with high confidence are shared with the scientific community, in the process of answering a question we often find more questions. Such is the iterative nature of research [93].

One of the challenges to ensure confidence on a scientific result is the reproducibility of results [94, 95]. The reproducibility of a computational result in the analysis depends solely on having enough details of the methods that yielded that iteration and the original data [96]. A strategy that some bioinformatics tools took was to produce log and execution files that accompany each intermediate result [97]. Researchers do not have to figure out what details are necessary to reproduce that result. Instead, these files are a technical sheet that guides reproducing the analysis either by a human or by the same tool [98].

Effective visualization of a result is critical for the interpretation into useful information [99]. However, complex visualization techniques require specialized data scientists to create visualizations for each iteration of the research. Interactive visualization of data helps speed up this process and allows rapid browsing of highly dimensional or complex data [100].

3 Aims of the study

The main goal of this thesis project is to propose efficient methodologies to connect efforts between molecular and cellular profiling research to cancer therapy outcome, therefore the focus of the individual aims builds upon previous advances in translational cancer biology.

The intermediate goals, together with the Publications where the goal is achieved, are the following:

1. Develop computational cytometry methodologies to improve single cell profiling of clinical tumor samples before and after chemotherapy (Publications I, II, III)
2. Identify cellular profiles associated with immunotherapy response on tumors with acquired chemotherapy resistance (Publication III)
3. Design a pre-clinical high-throughput methodology to screen dozens of epigenetic treatments as a way to revert drug resistance (Publication IV)

4 Materials and methods

This chapter briefly describes the methods applied and developed for the key results. Further details are available in the original publications, here denoted with Roman numbers.

4.1 Biological sample data

The data analyzed through this thesis were collected using genomic, proteomics, and drug screening measurement techniques. All the data are of human origin, and appropriate informed consent was reported from each study.

Publication	Samples	Technology	Source
I	19 fresh tumor and ascites samples from 15 patients with stage III and IV ovarian cancer. 9 samples taken at diagnosis time, 6 samples after three cycles of platinum treatment, and 3 samples taken when the disease relapsed	CyTOF	HERCULES Consortium
	14 Control PBMC samples from healthy (n=7) children and adult (n=7) donors	CyTOF	Flow Repository [101]
II	Matched primary and interval ascites samples from an HGSOC patient	CyTOF	HERCULES Consortium
III	26 archival samples from HGSOC patients enrolled in the Topacio clinical trial with acquired resistance to platinum therapy	Cyclic multiplexed immunofluorescence imaging (tCyCIF)	Topacio clinical trial [102]
IV	4 DLBCL cell lines: Su-Dhl-4 belongs to GCB subtype, Oci-Ly-3, and Riva-I to the ABC subtype, and Oci-Ly-19 is unclassified	Epigenetic reprogramming screening	Kindly provided by Dr. Karen Dybkær
		IF imaging, WES, and RNA-seq	

Table 1: Sample materials utilized in this thesis work. Publication II and Publication IV included other data not included in this thesis.

4.1.1 Mass cytometry data (I-II)

We used two cohorts of mass cytometry data samples for testing the feature of our software, *Cyto*. The first one was downloaded from the publicly available samples by Van Unen et al. 2016. We used all 14 FCS files corresponding to CyTOF data from healthy donors. The data set contained 48,611,486 cells, the

Cyto parameters, including a random subsampling of 300,000 cells, and arcsinh transformation. Outliers were determined based on Multidimensional Scaling (MDS) and non-redundancy scores (NRS) visualization and comparing Simpson's diversity index data exported from Cyto.

The second dataset is part of an in-house set of HGSOC samples acquired as part of the HERCULES Consortia. Fifteen tumor and ascites samples from 15 patients were prepared immediately after the surgical intervention and sent to our partner lab to perform CyTOF measurements. Data were processed with CyTOF software version 6.7.1014 to minimize variation due to instrument performance and FlowJo software to export bead-normalized single-viable cells based on channels 191/193Iridium and 103Rhodium. All of these samples were analyzed on Cyto.

Additionally, in Publication II, we analysed two samples from this set that originated from the same patient (PFI 2 months) before and after the first chemotherapy cycle. Single-cell data from this patient were log-transformed and normalized with Z-score to limit the batch effect on the high-resolution analysis with qSNE.

4.1.2 Cyclic immunofluorescence images (III)

Twenty six Formalin-fixed paraffin-embedded (FFPE) tumor samples were collected from patients enrolled in the TOPACIO study [102]. Twelve were from time of diagnosis and 14 were collected after 1 to 5 cycles of platinum therapy. The samples were stained following the tCyCIF protocol [103, 79] using an antibody panel to detect common immune cells and key signaling markers in epithelial cancer cells. Traditional image analysis methods do not work out of the box on tCyCIF data due to the complexities of cyclic staining and the high-dimensionality of the measurements. For the work presented in Publication III, we optimized the choice and order of analysis steps shown in Figure 8, as well as some of the specific steps, like cell type assignment and spatial analysis methodologies.

After performing shading effect correction with the BaSiC [104] tool, we aligned the images from different cycles using the DNA channel (Hoechst dye). The main differences between this workflow and traditional multiplexed imaging, such as the data used in Publication IV, are: (i) tile stitching errors would propagate and affect cycle alignment quality, which in turn is critical for optimal cell segmentation; (ii) segmentation for tumor tissues, which is more challenging than for dissociated cultured cells, requires comprehensive quality filter due to cell loss between cycles; (iii) and cell type-based analysis must handle over 40 channels vs. the traditional 4 channels.

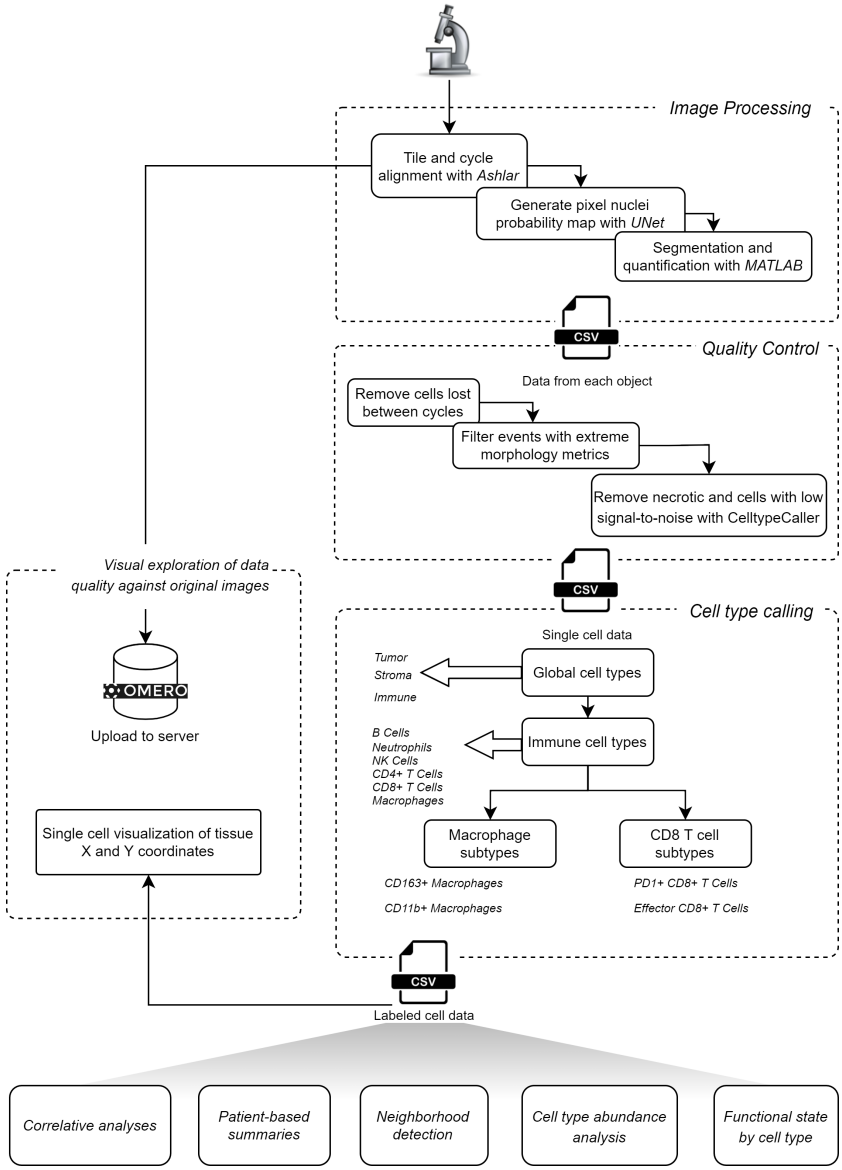

Figure 8: Computational analysis pipeline to process and analyse tCyCIF images. Image processing involves optimized stitching and segmentation to be able to quantify MFI by cell. Quality control steps are tailored to cyclic imaging techniques. The cell type calling module shows the subpopulations used on each iteration. Outside of the regular modules we included the downstream analyses used in Publication III.

4.1.3 Epigenetic inhibitor collection (IV)

A key component to accomplish the results of Publication IV was a carefully curated collection of epigenetic inhibitors (Figure 16). We searched for compounds to inhibit epigenetic enzymes with previous evidence of potential benefits in cancer. The collection contains compounds targeting HDAC (n = 21), DNMT (n = 7), HAT (n = 1), HMT (n = 15), HDM (n = 3), and BRD (n = 13). The compounds were bought from FIMM High-throughput Biology Core facility.

4.1.4 DNA repair assay images of epigenetically treated cell lines (IV)

Cells were incubated in T-25 flasks and treated every 3 days with belinostat, entinostat, vorinostat, or tazemetostat. Untreated cells were used as a control to measure endogenous DNA damage and DNA repair in these cell lines. We used doxorubicin to induce DNA damage (detected via gH2Ax) and measured the activity of DNA repair via homologous recombination (RAD51) and non-homologous end joining (53BP1), and apoptosis (cleaved-Casp3). Quantification was performed using the Anima framework [105], followed by a statistical summaries in R.

4.2 Modular implementation of cytometry workflow

We built Cyto using Anduril 2 [106], an open-source workflow framework that allows bioinformaticians to build analysis pipelines with multiple programming languages. At the same time, it also automatizes parallelization and systematic logging of the analysis steps. Cyto is composed of three modules shown in Figure 9: the graphical user interface as the *data importer*, the interactive *results browser*, and, at the heart of Cyto, the Anduril *cytometry pipeline*.

Cyto is distributed as an already built Docker container. Docker is a platform that uses virtualization of operating systems to encapsulate the environment of a software or service into packages called *containers*. The data importer is a Flask application server that handles the loading and saving of the data between the local machine and the container and coordinates the launch of the cytometry pipeline and the results browser. The results browser is a separate web application using Python Dash elements. Dash is a powerful Python library that integrates interactive data visualization elements from other libraries. When the user clicks on the "Run Analysis" button, Cyto will launch the cytometry pipeline with Anduril. The order of the Anduril components (Figure 10) has been tested to include the most popular methods in the appropriate order to produce trustworthy results. Each module produces a file output that can be exported and processed on its own if needed. The

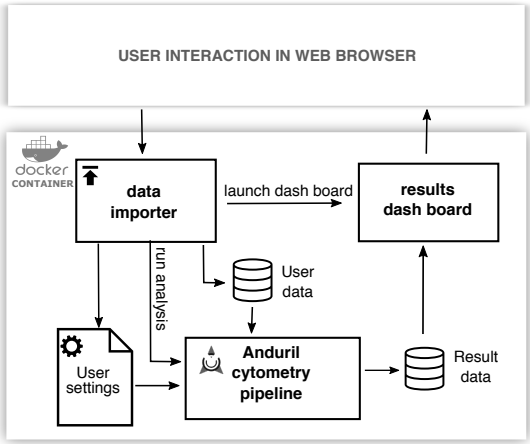

Figure 9: Software architecture to support Cyto's features. All modules are included in a Docker container that acts as a web server. The data importer and the results dash board host the user interaction, and the cytometry pipeline built with Anduril is responsible for processing the inputs, settings, and preparing the results packages.

common steps of a cytometry analysis comprise data preprocessing, unsupervised clustering, and dimensionality reduction. However, the choice of methods and parameters in each of these categories considerably affects the results.

4.3 Cell type-based analysis of single-cell imaging data

For Publication III, we devised and applied a new methodology to assign cell type labels to the single-cell data proceeding from the CyCIF experiment. The traditional gating technique on biaxial plots added high human-to-human variation, a large number of gates per sample and marker, and low confidence on the resulting cell type labels. We overcome this challenge by automatically annotating cell clusters produced by self-organized maps [107, 108]. Due to the variety of distributions across markers and across cell types, our cell type calling method tries to assign only a few labels at a time that are expected to be very dissimilar from each other. In the first iteration, we label tumor, immune, and stromal cells. In the second iteration, we label the immune cells into T cells, macrophages, B cells, NK cells, and neutrophils. The third and following iterations go deeper into subtypes of T cells and macrophages. An expert evaluated the labels by visual inspection with the raw images compared to the assigned labels on a side by side visualization.

The cell type labels and the coordinates of each cell were used for spatial analysis.

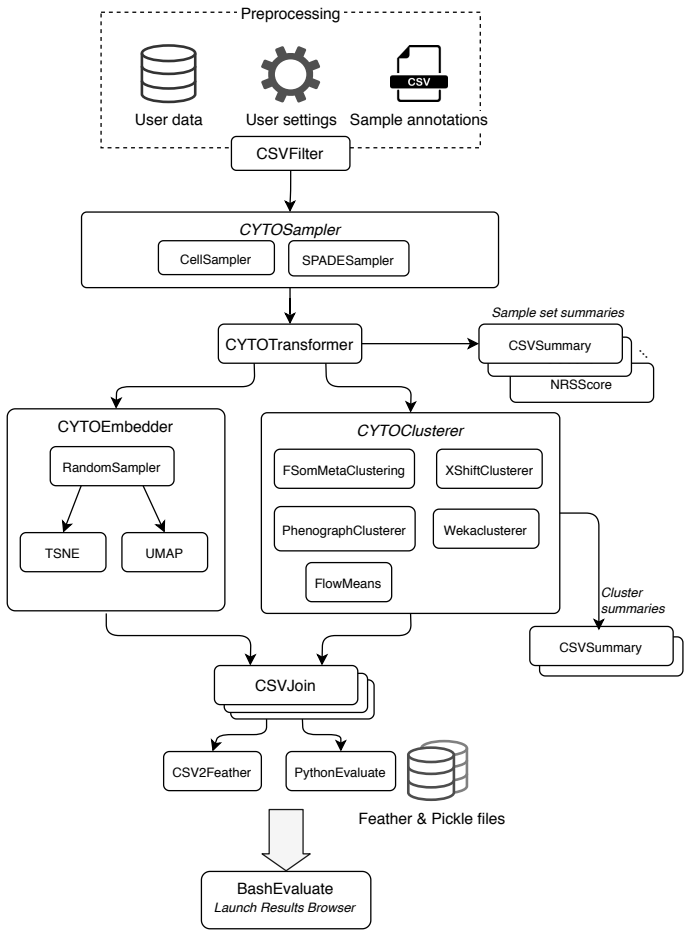

Figure 10: The analysis is composed of Anduril components and functions. The output of each component is an intermediate result that can be accessed at the end of the analysis run. The intermediate results are stored as CSV files and as feather and pickle objects compatible with R and Python.

Two cells were considered neighbors if the distance between their centers was less than 28 pixels, two times the average cell diameter. Two types of spatial analysis were used. First, a permutation test consisting of 1000 random permutations of the cell labels was used to determine which pairs of cell types would *attract* or *avoid* each other significantly ($p < 0.05$). Second, the abundance of PD-L1+ cells in the neighborhood of PD1+ CD8+ T cells was compared by defining PD-L1 thresholds for each sample.

4.4 Epigenetic drug screening

The screening procedure starts with microplates previously coated with a Labcyte Echo 550 acoustic dispenser. The drug concentrations were calculated as 10-fold dilutions from the recommended concentration advised by the provider. Positive (benzethonium chloride) and negative (DMSO) controls were randomized across the plate to detect potential plate effects (coefficient of variation). Cells were seeded using a BioMek MultiFlo TX Random Access Dispenser. The initial cell counts were optimized through titration assay to 3000-4000 cells per well. Plates undergoing 9 days of epigenetic treatment were covered with Labcyte microclimate lid. All plates were incubated at 37'C and 5% CO_2.

In-plate cell passaging was achieved with a Beckman Coulter BioMek FXp pipetting robot to dispense 384 wells simultaneously. The BioMek FXp program we made is included as a supplementary file in Publication IV. After the designated pretreatment time (1, 3, or 9 days), the treatment plates were treated with rituximab and doxorubicin, while the control plates were treated with PBS. The fixed-dose of rituximab and doxorubicin were optimized to achieve minor toxicity (IC20 - IC40) in the combination of the two. The objective measurement was cell viability with CellTiter-Glo reagent luminance readout. Cell viability for each dose was used for dose-response analysis using the following reprogramming score:

$$\text{Score} = \begin{cases} 0 & \text{if } \max_i(x_t^i - x_c^i) < 30\% \quad (1) \\ 0 & \text{if } x_t^a < 30\% \text{ where } a = \operatorname{argmax}_i(x_t^i - x_c^i) \quad (2) \\ 0 & \text{if } a \leq b \text{ where } a = \operatorname{argmax}_i(x_t^i - x_c^i) \text{ and } b = \operatorname{argmin}_i(x_t^i - x_c^i) \quad (3) \\ 0 & \text{if the difference } x_t^i - x_c^i \text{ is not positive for at least two consecutive } i \quad (4) \\ \max_i(x_t^i - x_c^i) & \text{otherwise} \quad (5) \end{cases}$$

where i represents the dose of epigenetic inhibitor. $\mathbf{x_t}$ represent the observed inhibition in the plate treated with rituximab and doxorubicin after pretreatment, and $\mathbf{x_c}$ represents the inhibition on the plate with only pretreatment (control plate). The score is computed as the highest difference between treatment and control dose-response curves, $\max_i(x_t^i - x_c^i)$. Additional cases were added to ensure that a positive score is caused by two differentially effective dose-response curves and not the result of artifacts in the measurement.

4.5 Whole exome sequencing and genomic profiling of DLBCL cell lines

We used NucleoSpin Tissue (Macherey-Nagel) for DNA extraction and SureSelect Human Exome V5 kits for whole-exome sequencing target enrichment and library

preparation. Samples were sequenced by BGI Genomics Co., Ltd. (Hong Kong) using Illumina HiSeq4000 sequencer. The key parameters were paired-end sequencing, 100bp read length, and 50x coverage. A custom analysis pipeline built with Anduril 2 was developed to perform sequencing processing and analysis. The steps for genomic profiling included (i) quality filtering, (ii) read alignment to reference genome, (iii), and variant filtering with public databases. A summary of the steps and methods is described in Table 2.

The first filtering module keeps only splicing and exonic variants with VAF > 20% in at least one cell line. Then, variants with CADD > 10 and either missing SNPdb information or an updated COSMIC [109] annotation. Variants within genes with low expression in our RNA-seq data (FPKM < 1) were discarded. The second filtering module used Minor Allele Frequency (MAF) and somatic type annotations from COSMIC to classify the variants by the likelihood of being *somatic* or *germline*. Germline variants were further filtered based on potential relevance to the epigenetic inhibitors, rituximab, and doxorubicin. All variants were annotated and selected based on annotations from the following databases: CIVIC [110], DGIdb [111], DrugBank [112], PharmGKB [113], LOVD3 [114], and IARC [115].

Step	**Tool**	**Source**
Quality control of raw reads and mates	FastQC	[116]
Trimming read ends	Trimmomatic	[117]
Read alignment to reference genome hg19	Burrows-Wheeler Aligner	[118]
BAM file sorting	Picard tools	[119]
Variant calling and filtering	GATK and Annovar	[67, 120]

Table 2: Whole Exome sequence processing and analysis steps and tools.

5 Results

The following sections present the results of this work. First, the computational methods developed and applied to cytometry data analyses (Pub I-III), followed by tumor microenvironment findings from the clinical trial Topacio (Pub III), and then the methodological and biomedical contributions to preclinical screening (Pub IV).

Description	Category	Publication
Agile workflow for cytometry analyses	Methodological	I
Automatic cell type assignment	Methodological	III
Tumor composition in high grade serous ovarian cancer samples	Biomedical	I, II, III
Spatial organization of cells with PD-L1 and PD1 expression	Biomedical	III
High-throughput screening of non-simultaneous drug combinations	Methodological	IV
HDAC inhibitors shift DNA repair ability on DLBCL cell lines	Biomedical	IV

Table 3: Summary of main contributions from this work.

5.1 An agile-based workflow for mass cytometry analyses

To help interpreting results from highly dimensional data while maintaining reproducible analyses, We designed a method that supports the iterative nature of data analysis and agile methodology principles. It is presented in the form of a Docker-based tool called Cyto (Figure 11).

Agile principles focus on the individuals instead of the processes, and the main goal is to produce value in each iteration. While it was a methodology designed for software development, there are many views on whether it applies to data science and in which form. We aligned Agile principles to the scientific principles of reproducibility, transparency, and systematic evaluation of results in this work. While a primary research question drives a research project, many intermediary or secondary questions relate to designing experiments and determining analysis steps. These questions often require running exploratory analysis on data, discussing the results to conclude that helps formulate the next question. Each of these iterations should aim towards the primary goal of the project.

It does sound simple. However, publications often describe only the last bioinformatic analysis that yielded the results they include in the manuscript. In practice, source code changes from iteration to iteration, and there is a lack of systems to keep the result of each analysis iteration organized.

Data analysis is iterative in nature. Each iteration must answer an important question that leads to the next question and, therefore, the next iteration of the analysis. In Publications I and II, we have questions about differences between sample groups at the cell population level. However, we first needed to ask earlier questions about sample and feature selection to ask such questions. Experimental design must go together with analysis design. However, biological assays do not always turn as we expect, so we must ask questions regarding the quality of the data, or the feasibility of the next questions. Unfortunately, quality assessment often goes unnoticed.

The common pipeline followed for mass cytometry data is included as the core of the Cyto software. Among the most common steps of cytometry data preprocessing is the systematic sampling of cells, which is discussed in Publication II; however, to use state-of-the-art cytometry tools that need smaller datasets, Cyto includes two strategies for cell sampling: random and density-based. After this step, CyTOF data is transformed and normalized to prepare it for downstream analysis. The pipeline includes comprehensive metric calculation and statistic summaries from the global data to interpret inter-sample and inter-marker variability.

The two key steps in all cytometry analysis are dimensionality reduction by tSNE or UMAP and unsupervised high-dimensional clustering. Methods and algorithms in both of these areas are constantly developed. Users tend to choose methods based on their ease of use within their favorite environment rather than by how well a

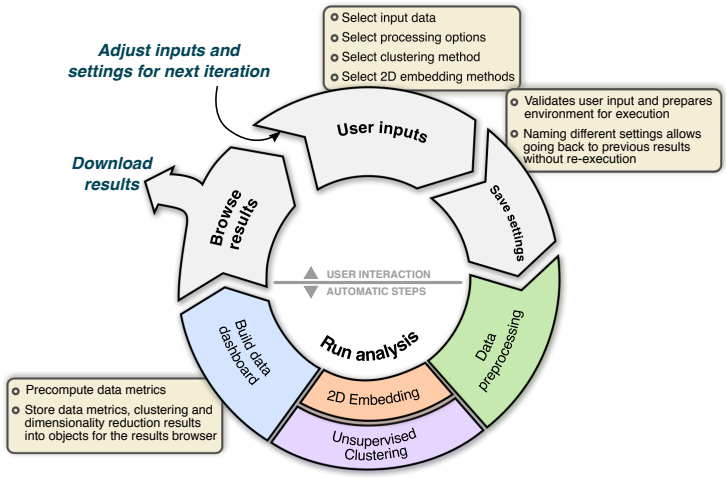

Figure 11: The iterative life cycle of cytometry analysis projects with Cyto. The upper part of the cycle shows the user settings, that must be designed based on data and research questions. The lower part of the cycle shows the automatic steps by the analysis pipeline, this part can run unattended in a computing cluster or locally, in as many samples as needed.

method is targeting their kind of data and research question. In CyTOF, we lack ground truth, so we depend mostly on prior knowledge to decide if the output is informative and useful.

Furthermore, CyTOF analyses in the literature, when they include the source code, it is a collection of R scripts. Here, all the CSV transformations, filterings, and concatenations are done with Anduril components that follow a regular unit testing strategy. Figure 11 shows the steps of the workflow and the architecture that supports this methodology. Interactive visualization supports live analysis during project discussions among collaborating scientists while importing and exporting Cyto configuration files means systematic reporting and out-of-the-box reproducibility of the results.

5.1.1 Interactive outlier detection and cell-type identification

The cytometry field is such a well established field that we take for granted the quality of the experimental protocols. However, while measuring several replicates helps alleviate this problem, it is also important to have control samples to use as a baseline and to be able to assess the quality of these data.

The first iteration of Cyto on the Control samples from Van Unen et al. [101], showed that two samples (52_CtrlAdult5_PBMC and 53_CtrlAdult6_PBMC) are significantly distinct from the rest (Figure 12). Further inspection of the browser shows enrichment of a myeloid population in sample 53 and a generalized low signal in sample 52. The second iteration, including the rest of the samples, shows a clear recapitulation of the known cell types present in peripheral blood samples.

5.1.2 Integration of clinical data to cellular composition and expression profiles

In this study, we used two separate iterations to answer two different questions. First, we analyzed the general cell population abundance with random cell subsampling (Figure 13A). Second, we used density-based downsampling to capture as many tumor cells as possible despite the low purity of ascites samples (Figure 13B).

The first iteration results were used to interrogate the composition of the tumor microenvironment (Figure 13C). The myeloid and CD8- T cell populations are the largest in relation to CD8+ T cells, tumor, and stromal compartments. When comparing solid tumor composition against ascites, we observed more tumor cells and less myeloid cells, but no clear distinction on T cell abundance by tissue type was observed.

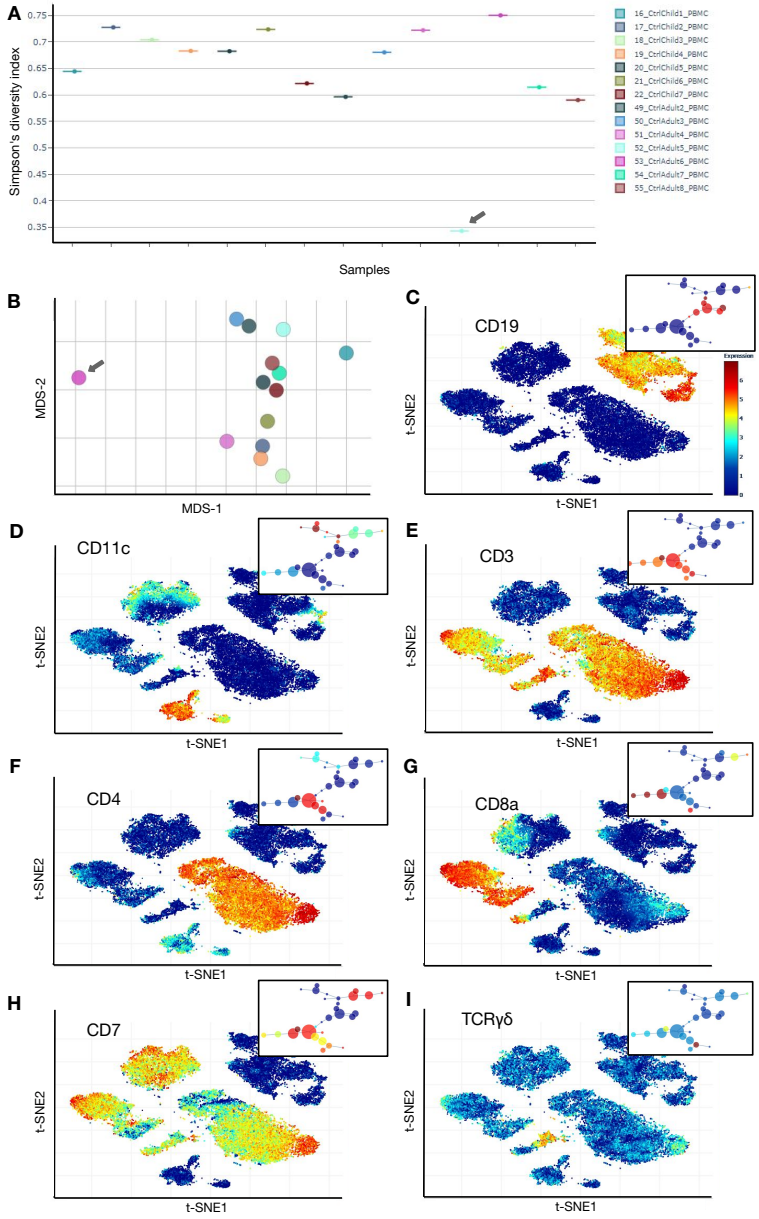

Figure 12: Simpson's diversity index and multidimensional scaling visualization of Peripheral blood mononuclear cell control samples highlights 52_CtrlAdult5_PBMC and 53_CtrlAdult4_PBMC as outliers. After removing the two samples, interactive tSNE and MST visualization recapitulates the cell subpopulations identified by the authors of the data [101].

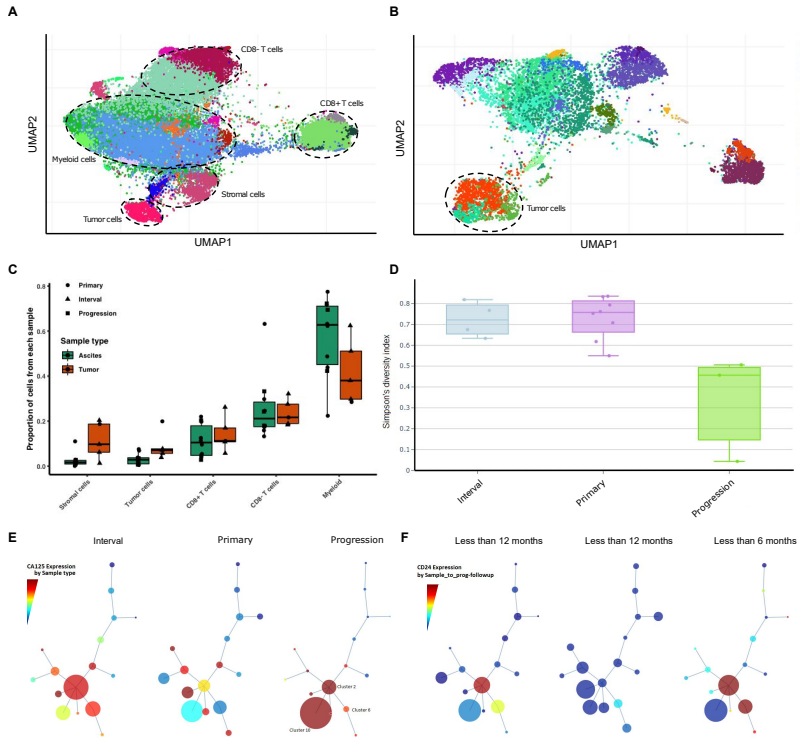

Figure 13: High-grade serous ovarian cancer samples with random (A) and density-biased (B). Random sampling helps detect differences in population sizes (C), while density-biased sampling helps isolate small populations from the rest of the cells and use them for further analysis. Panel D shows different abundance and expression of the tumor cell subpopulations by the sample time (D) and by the *"time from sample to next progression"*.

The second Cyto analysis consisted of using only the tumor cells as the input data. In this case, we demonstrate the integration of sample annotations within Cyto. The most interesting finding was that Simpson's diversity index of tumor cells shows reduced tumor-cell heterogeneity in relapse samples but not at the time right after chemotherapy. Cyto summarizes clustering results as minimum spanning trees with abundance and expression data across different sample annotations. Figure 13E shows tumor cell populations grouped by time point (i.e., Primary, Interval, and Progression), where Cluster-10 is the most dominant population in Progression samples, that have the lowest abundance of Cluster-6. Expression analysis shows that Cluster-6 cells have high Ki67 expression; on the other hand, Cluster-10 cells express MUC1 and CD147 and lack Ki67 and ERK1-2 expression. When selecting the *"time from sample to next progression"* annotation to group the cells, we see

a clear enrichment of stemness marker CD24 in samples with a shorter time to the next progression. CD24 has recently been reported as a potential biomarker of aggressive ovarian cancer, but further validation studies are needed [121].

5.2 High-resolution analysis of fresh HGSOC ascites samples before and after platinum-based therapy

qSNE is an optimizer that improves upon the popular tSNE algorithm. It incorporates a new optimization function and automatic perplexity estimation. These features translate into the option of analyzing large datasets without the need for sub-sampling, thus not missing rare but important information in the data in addition to highly improved performance speed. We chose ascites samples because they are often hard to work with due to low tumor purity. Samples from different patients have high inter-patient variation, while the same patient samples are so similar that analyses would biased to batch or experimental artifacts. The selected samples were acquired before and after one cycle of chemotherapy, then processed and stained immediately after the surgery.

These data were selected as an exclusive application of qSNE's feature of not needing to downscale the data. Figure 14A shows the key populations that are present in either of the two time points. To run the tSNE algorithm, we must reduce the number of cells, which means we may miss important but rare populations of cells.

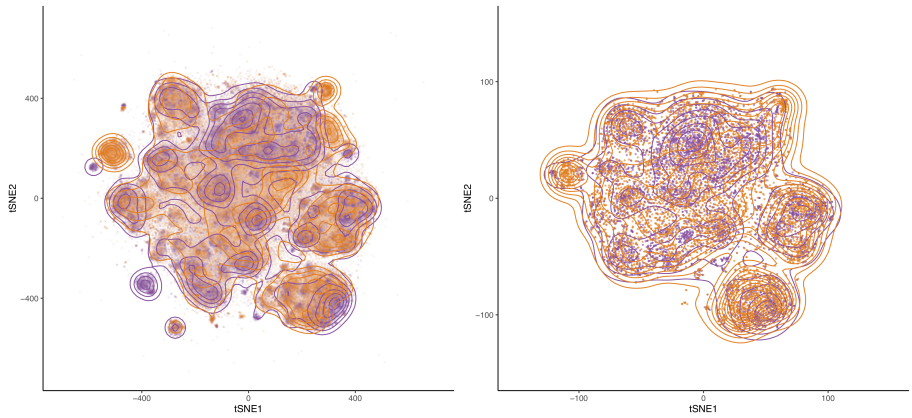

Figure 14: tSNE visualization of cells from the same patient before (purple) and after (orange) chemotherapy. (A) Coordinates produced by qSNE for the full size data (173,374 cells); (B) Coordinates produced by Rtsne package for a subsample of the data (10,000 cells).

5.3 Tumor-immune microenvironment profiles and response to PD1 and PARP1 inhibitors

The main contributions of this thesis to high-dimensional microscopy image analysis are part of Publication III. Cyclic multiplexed immunofluorescence is a modern methodology in which we can scan multiple microscopy images with different staining markers from the same slide. After careful processing methods, these images can be compiled into a large dataset of single-cell mean intensity measurements (MFI) from up to 50 antibody markers. In Publication III, we produced tCyCIF data from 26 whole tissue slides from HGSOC patients enrolled in the Topacio clinical trial [102] Figure 15A. Each sample was stained and scanned following the tCyCIF protocol for 12 cycles. The antibody panel covered antibodies to detect the tumor, stromal, and seven different immune cell types, in addition to functional markers related to DNA damage, interferon signaling, and immune checkpoint. Each sample produced an average of 500,000 cells.

5.3.1 Automatic cell type calling characterizes potential mechanisms of response

We devised an algorithm to assign a cell type label to each cell before cell type based analyses. The expected noise-to-signal ratio from tCyCIF data differs from that of CyTOF or mRNA sequencing in that overlapping cells and imperfect segmentation capture signal from neighboring cells. Using multiple channels helps determine cell identity. Our cell-type caller uses the self-organized maps clustering method because it is fast and can find clusters of different sizes. However, the dissimilarity among immune cell types is less than that among global cell types. We solved this by applying our cell type caller on different levels of the data separately. First, all cells are classified into level 1 categories (Figure 15B): *immune*, *stromal*, and *tumor* cells. The immune cells are then treated as a separate dataset to call level 2 of immune cell types (Figure 15C): CD4 T cells, CD8 T cells, B cells, macrophages, NK cells, neutrophils, and antigen-presenting cells. Some of these cells' subsets are further classified to detect level 3, which comprises macrophage types and T cell subtypes.

Cell type abundance was not enough to confidently associate cell proportions to therapy response on this dataset. However, we observed a larger immune compartment in chemo-exposed samples, in agreement with a positive immune score from the nanostring analysis described in Publication III. Chemo-exposed samples also had significantly more antigen-presenting cells and neutrophils. While cell type proportions were not significantly associated with the response, interferon signaling marker pSTAT1 was upregulated in PD1 CD8 T cells from samples with

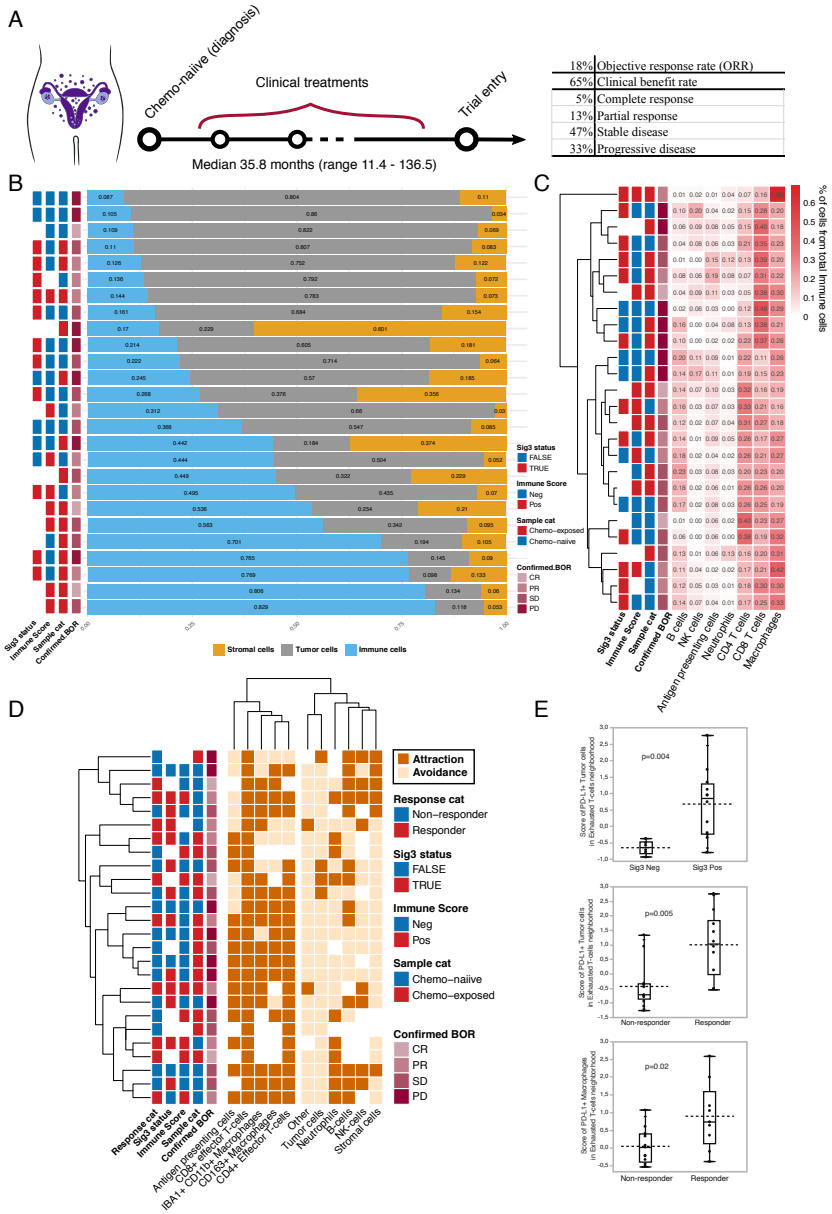

Figure 15: (A) Topacio clinical trial samples were acquired either at diagnosis time or after chemotherapy. Cell type composition of the tumor microenvironment in each sample by global cell types (B) and by immune cell types (C). (D) Cell types with significant attraction or avoidance to PD1+ CD8+ T cells as a result of permutation testing. (E) Correlation analysis of fraction of PD-L1 positive neighbor cells around PD1+ CD8+ T cells with therapy response and genomic mutation biomarker.

the highest objective response to this combination treatment.

5.3.2 Spatial cellular organization associated with clinical data

Overall, cell-type expression of PD-L1 ligand did not show the striking association one would expect in response to PD1 inhibitors. However, PD-L1 positive cells in close physical proximity to PD1+ CD8+ T cells did show a significant correlation. Here, we developed the hypothesis that the cell type variable is not independent of the type of neighboring cells. We created new features in the form of the fraction of neighbors of each level 3 cell type in the sample.

Neighborhood analysis showed that both the fraction of PD-L1 positive tumor cells and PD-L1 positive macrophages neighboring PD1+ CD8+ T cells were higher on the *responders*' group regardless if the sample was or not chemo-exposed. The strongest correlation observed in this analysis was that samples with mutational signature 3, associated with DNA damage, had the highest fraction of PD-L1 positive tumor cells around PD1+ CD8+ T cells. This finding shows in human samples the hypothesis about the synergistic mechanism between PD1 checkpoint and PARP1 inhibitors is due to an interplay between DNA damage and interferon-primed CD8 T cells.

5.4 High-throughput screening of compounds as pre-treatment for resistant DLBCL

Drug sensitivity and resistance testing have been demonstrated as a pre-clinical screening method for personalized medicine [122]. In Publication IV, we built on this methodology by designing, through systematic testing, an assay to screen non-simultaneous drug combinations. The need for this development stems from the mechanism of action of epigenetic inhibitors, a promising family of drugs that promise few side effects and a wide range of reprogramming outcomes on the cells. By taking advantage of microplate handling robots available at high-throughput core facilities of most institutes with personalized medicine programs. Screening methods have the problem of keeping the cells alive long enough to observe the effects of slow drug treatments.

Figure 16A shows the key steps of this method. In brief, cells are seeded on precoated plates and passaged every day *in situ* with epigenetic inhibitor in the culture media. In Publication IV, we showed two applications of this methodology: tailoring the plate design first to measure single drug dose-response curves and, second, computing dose-response combination matrices of two drugs to study

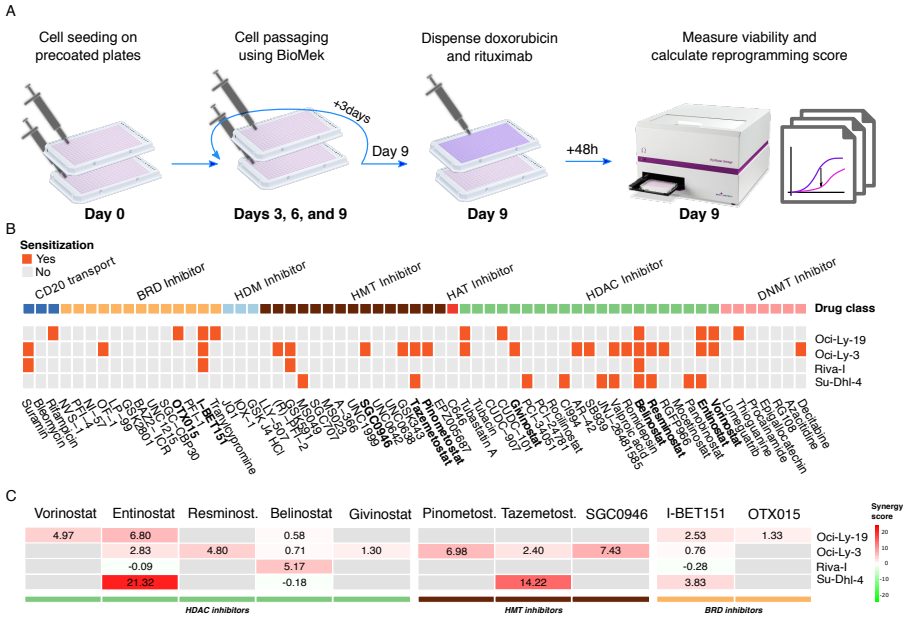

Figure 16: (A) Non-simultaneous drug screening with in-plate passaging, the outcome is the maximum difference between the pretreated cells with standard treatment (purple) and the pretreated cells without added treatment (pink). (B) Complete compound collection sorted by compound class; successful reprogramming at 9 days of pretreatment is marked in orange. (C) Top drug candidates are further validated for synergistic reprogramming to ensure the observed effect is not only the result of cytotoxic doses.

synergistic effects between the combination of epigenetic inhibitors and standard treatment.

We computed the reprogramming score from single-drug dose-response curves to classify them as sensitizing or not. Figure 16B shows that all cell lines were successfully sensitized with various inhibitors. Riva-I turned out to be the most reprogramming resistant, while Oci-Ly-3, the most rituximab- and doxorubicin-resistant, had the most hits. Drug-drug synergy analysis was used to validate the effects on the most promising hits (Figure 16C); entinostat and tazemetostat had the highest synergy, particularly in Su-Dhl-4. Importantly, all the tested combinations had either additive or synergistic effects, but no antagonistic effects were detected. Again, Oci-Ly-3 was sensitized by multiple inhibitors, while Riva-I was only successfully reprogrammed by the pan-HDAC inhibitor belinostat. Taken together, this result shows that rituximab and doxorubicin resistance can be epigenetically reverted.

5.4.1 Epigenetic reprogramming of DNA repair mechanisms reverts doxorubicin resistance

Doxorubicin cytotoxicity is caused by double-strand DNA breakage when doxorubicin molecules intercalate with the DNA. We hypothesized that part of the sensitization effect observed in the reprogramming screening (Figure 16) was due to changes in DNA repair ability. We repeated the same setting of 9 days of epigenetic pre-treatment every 3 days. At the end of the treatment, we used immunofluorescence microscopy to identify double-strand break repair activation in cells affected by doxorubicin.

Figure 17 shows the quantitative result for this experiment as proportions of cells positive for each antibody marker scanned. The cells without induced DNA damage with doxorubicin show their own levels of endogenous DNA damage and repair on the left. The entinostat dose was too high for Su-Dhl-4 cells, which means that a dose-tuning step is necessary to avoid cytotoxic effects. All other cell lines and drug pairs showed no additional DNA damage (gH2Ax) or cell death (cCasp3) than the baseline for each cell line. All HDACi treated cells were unable to activate HR (RAD51) as much as the untreated cells, which shows greater DNA damage and apoptosis levels.

5.4.2 Genotyping cell lines by drug response

We used Whole Exome sequencing and RNA-seq to investigate genetic factors in these cell lines in the context of epigenetic reprogramming in addition to the effects on pathway regulation before and after epigenetic treatment. RNA-seq analysis showed agreement with the image analysis on the dysregulation of DNA repair. Hence, we used functional genomic variant prediction for pathways resulting from RNA-seq and database annotation for previously known targets of the drugs used in our experiment. Table 4 summarizes the findings on each of these groups.

Among epigenetic genes, *EZH2* presented a missense variant in the cell line that responded the most to the *EZH2* inhibitor tazemetostat suggesting that R-CHOP resistant patients with *EZH2* mutations may benefit from tazemetostat pretreatment to revert resistance. Genomic variants in genes from DNA repair machinery and known drug targets may create new research that leads to potential biomarkers. Oci-Ly-3, the most doxorubicin and rituximab resistant cell line, was successfully sensitized but had fewer mutations in the regions analyzed. This result lead to the hypothesis that polymorphisms in the *XRCC3* gene, within DNA repair machinery, could be a key difference between Riva-I, the most epigenetically resistant cell line.

Taken together, careful functional genomic characterization of the cell lines and

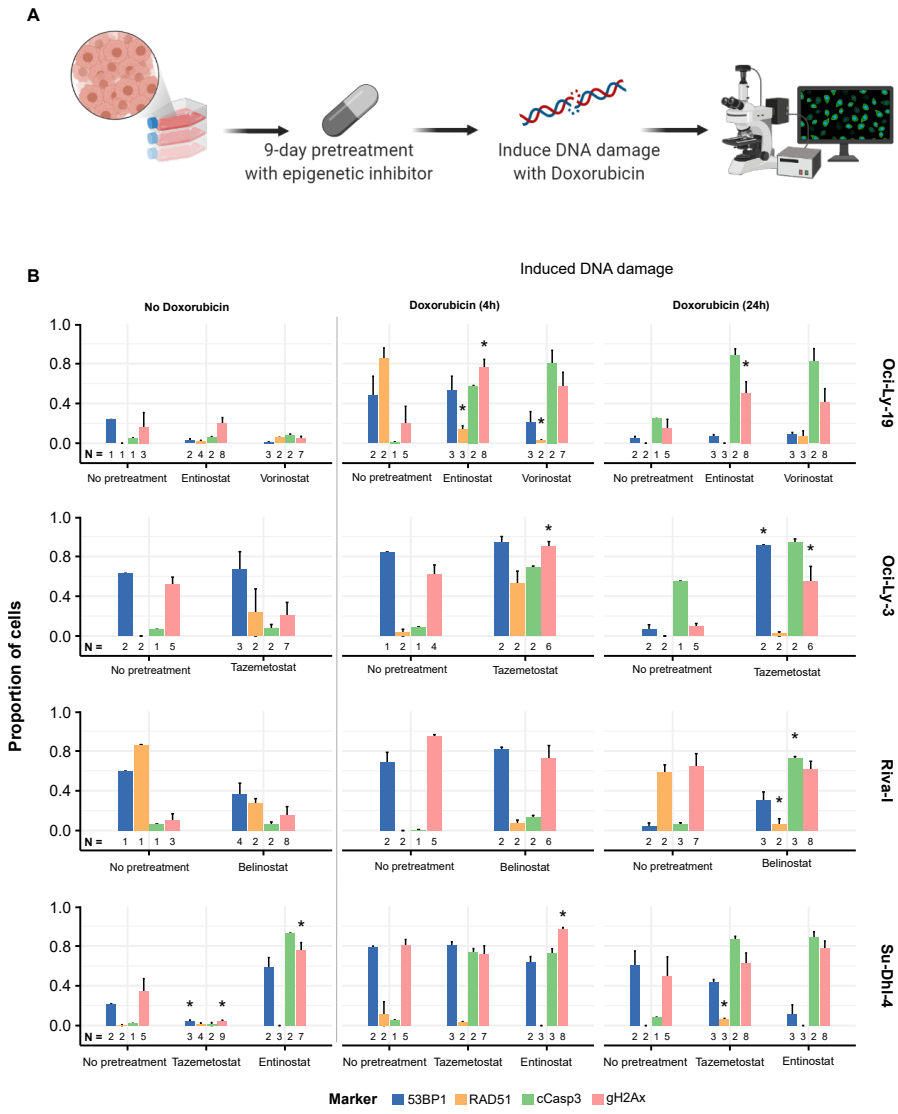

Figure 17: Quantitative results from 4-plex immunofluorescence imaging of the effect of synergistic combinations on the cells response to induced DNA damage via doxorubicin. We quantified proportion of cells positive for NHEJ (represented by 53BP1 expression in blue), HR repair activation (RAD51 in yellow), apoptosis (cCasp3 in green), and DSB (gH2Ax in pink).

their response to epigenetic reprogramming highlighted genes that are commonly mutated in clinical samples from DLBCL patients (18% of patients show *MYD88*, 10% *CREBBP*, *ARID1A*, *TP53*, and 6% *EZH2*) [123].

Gene Classification		Oci-Ly-19	Oci-Ly-3	Riva-I	Su-Dhl-4
Drug targets or reported association	Doxorubicin	*DPYD**, *NQO1**	-	*AKT1*, *TP53*	*DPYD**, *FGFR4**, *NQO1**, *TP53*
	Rituximab	*DPYD**	-	*CREBBP*	*DPYD**
	Tazemetostat	-	-	-	*EZH2*
	Histone deacetylase inhibitors (HDACi)	-	*MYD88*	*TP53*	*EZH2*, *TP53*
Epigenetic	Histone acetyltransferases (HATs)	-	-	*CREBBP*	-
	Histone methyltransferases (HMTs)	-	-	-	*EZH2*
	Epigenetic regulation	*DPYD**	-	-	*DPYD**
Chemoresponse associated cell mechanisms	Cell cycle	-	-	*CREBBP*, *STAG2*, *TP53*	*TP53*
	DNA repair	*XRCC3**	*XRCC3**	*ERCC4**, *TP53*	*TP53*, *XRCC3**
	Transcription factors	*CIC*	*BCL6*	*ARID1A*, *CIC*	*RCOR1*

Table 4: Summary of genomic variants. Marked with an asterisk the potentially germline variants.

5.4.3 Interactive browser of epigenetic reprogramming results

To navigate and visualize data resulting from different types of assays, we built an interactive web application (Figure 18). The results browser is built with the R library Shiny coupled with Plotly visualizations and Heatmaply.

The data proceeding from the reprogramming screening are shown with the raw curves to easily explore whether the reprogramming scores were showing pointing to actual curve differences. The synergy assay can be explored as a matrix or a 3-dimensional visualization of the types of landscapes that each combination produces. The analysis and preprocessing of the dose-response data was done by developing new Anduril components. RNA-seq data can be explored as cell line-centric visualizations with different normalization levels of the read counts. Genomic variant results are shown with the VAF, but hovering with the mouse will let the user see the variants detected.

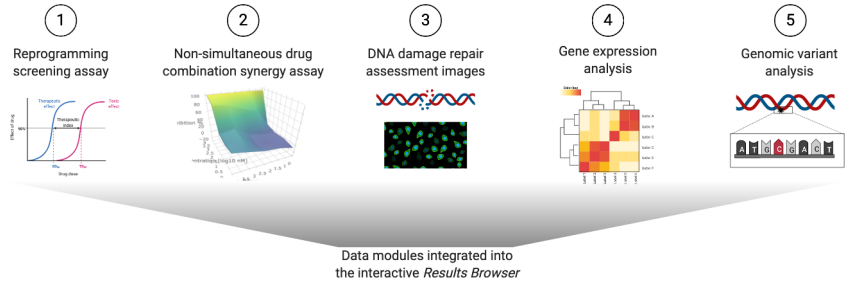

Figure 18: Integration of multi-omics data into an interactive results browser to help during and after the project to share results between the authors and with the community.

6 Discussion

The work throughout this thesis aimed at placing new techniques and reinforcing existing translational steps in the fight against cancer. The first two methodologies describe efficient and effective solutions to interpret mass cytometry data and translate it into biologist-friendly information. The third publication results from an international collaboration, where the novel methodologies described here served as a bridge between the high-throughput experimental side and the doctors seeking answers from a valuable clinical trial for modern treatments of chemotherapy-resistant cancer. The fourth study is the culmination of a human experiment on becoming an interdisciplinary scientist, which resulted in an innovative pre-clinical protocol that highlighted promising findings on epigenetic reprogramming of drug resistance.

6.1 Advances in cytometry data analysis in cancer

Recent technological advances in cytometry produce large amounts of single-cell data to discover new insights into the tumor microenvironment, but they pose new challenges for computational scientists [124]. Additionally, the complexity of single cell analyses also poses challenges to reproducibility. While standard procedures have been set in place to improve the reproducibility of the experiment, the computational analyses still depend on single bioinformaticians modifying scripts on each step of the analysis. Workflow paradigm, containerized applications, and interactive visualizations have been proposed to share results and reproducible steps [106, 98, 100].

Publications I and II provide high and low-level methodologies respectively to tackle limitations in this field and produce new insights on the effects of chemotherapy on HGSOC cellular composition. In Publication I, we published Cyto, an interactive open-source Docker container that enables researchers to run large cytometry analysis pipelines while producing the log and settings files needed to reproduce and share their results through multiple analysis iterations. Cyto integrates state-of-the-art tools tailored to single-cell measurements; however, it also highlighted the limitations on the current standard practice of downsampling large datasets due to limited computational resources. Publication II identifies such limitations and offers a rapid optimization algorithm to compute tSNE coordinates on large datasets.

The application of both methods contributes to the efforts of making use of ascites samples from ovarian cancer patients [56]. Both publications demonstrate how to

visualize tumor cell populations in ascites samples despite the low tumor purity of this type of specimen. Our analysis showed reduced heterogeneity at the time of relapse and stem-like populations in samples with shorter prognosis. Identifying chemotherapy-resistant tumor cells is a critical step to be able to challenge them.

6.2 Image-based interrogation of the tumor-immune microenvironment

New treatment strategies combining PARPi and checkpoint inhibitors showed promising synergy through interferon pathway activation [125, 126]. Publication III reports two novel biomarkers of response to this drug combination; the presence of mutational signature associated with DNA damage scars [127], and interferon response in T cells as measured with nanostring pathway analysis [128]. Further interrogation of the tumor microenvironment with highly multiplexed imaging [79] identified spatial organization patterns linked to therapy response and mutational signatures.

Spatial analysis of high dimensional images is a field under development [129]. While in this study, we designed a tailored analysis to evaluate cell states in function of neighbor's identity. This set of features opens the possibility of countless further analysis of the tumor microenvironment's spatial organization. Taken together, we suggested a model for the synergistic combination of PARPi and PD1 checkpoint blockade in HGSOC, the most aggressive subtype of ovarian cancer.

6.3 The role of epigenetic reprogramming in preclinical models

A major challenge in cancer research is to find options for relapse and refractory cancers. Epigenetic inhibitors are a promising option to revert drug resistance in these cases [130]. Publication IV reports a novel experimental protocol to identify epigenetic mechanisms able to revert resistance to anti-cancer therapeutics.

We demonstrate its value by screening 60 epigenetic inhibitors to revert resistance on four cell lines representing DLBCL subtypes. While all cell lines were successfully (re)sensitized to the combination of doxorubicin and rituximab, the reprogramming effects were different. Immunofluorescence, transcriptomics, and genomics analyses pointed to DNA repair and cell cycle dysregulation as the main reprogrammed mechanisms via HDACi and HMTi. Careful genomic analysis of the cell lines in association with the response to epigenetic reprogramming highlighted mutations commonly detected in DLBCL patients [113]. This opens new leads for validation of potentially predictive biomarkers of epigenetic reprogramming.

Future applications of this method could, however, overcome explicit limitations on this study. The main limitation is an imbalanced number of drugs for each epigenetic inhibitor class due to availability when making the collection, including new drugs, which may shed light on HDM and HAT compounds. Additionally, while rituximab has multiple mechanisms of action, our study with cell lines cannot measure the impact of the tumor microenvironment. Recent and future advances in *in vitro* models may enable measuring reprogramming effects on rituximab's tumor immunity effects [131, 59]. These limitations notwithstanding, this is the first high-throughput method to test non-simultaneous drug combinations and promising area for treatment development.

6.4 Conclusion and future directions

Cancer can be considered a battle inside our body [5]. Rogue cells lose control, stop performing their function, and hinder the ability of the rest of the cells to keep the whole system alive. The cancer cells replicate and evolve fast. It is very diverse that, although they have commonalities, we now know that cancer is not one disease and we will need multiple strategies to cure them all. In the same manner that interactions between cells and their collective behaviors shape the battlefield, collaborative efforts between multiple disciplines are needed to identify weaknesses within the highly complex system that makes the tumor microenvironment. The space between the bench and the bedside contains hectic movement of new technologies, data, hypotheses, clinical trials, and policy makers too [132]. This thesis proposes methodologies both, *in silico* and *in vitro*, to aid traffic control in this space by solving roadblocks and building bridges.

In summary, this work presents two computational cytometry solutions to support iterative analysis of clinical samples, even in cases where tumor purity would be a challenge. These solutions identify stem-like cell abundance as a potential culprit of drug resistance in HGSOC. Future Cyto analysis of many matched pre-, on-, and post-chemotherapy laparoscopy samples paired with an antibody panel representing current knowledge of drug resistance mechanisms, would be necessary to identify actionable cell populations that remain active after chemotherapy. Additionally, ascites samples are readily available and, biomarkers stemming from mass cytometry analysis can be easily transferred to flow cytometry, hence applicable to clinical settings.

Cytometry analysis is further expanded with the application to highly multiplexed microscopy images. Genomic and immune profiling highlight potential biomarkers of response in clinical trials, and spatial analysis decipher the cellular organization associated with response to the combination of PARP and PD1 inhibitors. Future

analysis of cell signaling associated with immune evasion and cell adhesion pathways can shed light on druggable targets to disrupt communication between malignant cells. The fourth study shows a novel pre-clinical protocol to screen for epigenetically reprogrammable mechanisms of drug resistance. Interestingly, homologous recombination repair, modulated in the third study through PARP inhibitors in ovarian cancer, is indirectly regulated via epigenetic reprogramming in lymphoma cells. Our screening approach takes advantage of the plasticity of the epigenome to treat with non-cytotoxic drug doses. Future screening of non-simultaneous drug combinations applied to recent organoid models would be necessary to link epigenetic reprogramming to cellular interactions.

The road towards eradicating cancer is paved with small incremental steps as well as surprising turns. We propose valuable steps to characterize the actors on the battlefield and their weaknesses. I wish to conclude this thesis by seeding the idea that sharing specialized knowledge from one discipline and making it usable for other disciplines will be the backbone of this decade's translational research.

Acknowledgements

First of all, I wish to thank professor Sampsa Hautaniemi for accepting that nervous MSc student for an internship at a time when I did not know that a PhD could be a possibility. The opportunities that I enjoyed since 2012 have all been thanks to you. This work was carried out at the Genome Scale Biology research program at the Faculty of Medicine during 2014-2020. This work was financially supported by Hautaniemi lab thanks to Sigrid Juselius Foundation, Finnish Cancer Foundation (Syöpäsaatio), and the European Union's Horizon 2020 Research and Innovation Programme. I am thankful for the generous grants from the Finnish Cancer Foundation and Instrumentarium Foundation, and to the Doctoral Programme in Biomedicine and Chancellor's grants.

A special thanks to doctor Anniina Färkkilä for making my research visit in Boston in 2018 an exciting and unforgettable learning experience, but most of all for offering me a postdoc full of exciting international collaborations. I enjoy working with you and look forward to future endeavors at Färkkilä Lab!

Thanks to doctor Maciej Lalowski and professor Juho Rousu for seeing my thesis through and for their help in the annual thesis committee meetings. To doctor Jussi Paananen and professor Janne Lehtiö for pre-examining this thesis. And to doctor Nina Peitsaro for agreeing to represent the Faculty in the grading committee. A special thanks to doctor Emmanuel Barillot for accepting to be the opponent, I am looking forward to our discussion and wish you could visit Helsinki for this event.

In this time, I have had the pleasure to work with brilliant collaborators, without whom this thesis would not have been the same. Professors Sirpa Leppä and Krister Wennerberg, and their respective groups were instrumental in the epigenetic reprogramming project. Doctors Luca Pasquini, Eleonora Petrucci, and Mauro Biffoni, together with doctor Katja Kaipio, proved to be excellent remote collaborators before it became mainstream, I appreciate your hard work with sample preparation and CyTOF experiments. I am extremely grateful to professor Peter Sorger for inviting me to the Laboratory of Systems Pharmacology at the Harvard Medical School. I felt welcomed and integrated from the very first day thanks to India Dittemore and Zolta Maliga. Everyone at LSP were helpful and fun to work and party with, a special thanks to Connor, Yu-An, Clarence, Shu, Jeremy, Meri, Jenny, Denis and Klas.

Hautaniemi lab has always been full of great lab mates willing to help. As such, a special mention must go to doctor Ville Rantanen for being a mentor and a friend. To him, Antti, Oskari and Juha for their collaboration on cytometry algorithms and analysis. To the lab mates who shared so many hours, Anduril days, and Coding

camps with me: Ville, Emilia, Riku, Erkka, Ale, Antti, Oskari, Kaiyang, Chiara, Kristian, Kat, Jaana, Juha, Amjad, Ping, Chengyu, and many others who although shared only a short time with me made the lab a better place. I also want to thank doctor Rainer Lehtonen, for teaching me about variant filtering and for hosting the annual Salmari competition. To Chiara, for sharing the ups and downs of wet-lab learning through weekends and holidays on our first years. To the students of the courses I have taught, because I learned from you more than you think, and I enjoyed seeing some of you grow into expert bioinformaticians. Also, a quick thank you to professor Juha Klefström and the organising team of the Cancer Biology Summer School, it was a rewarding experience to work with so many motivated students and PIs over the years. To the ScienceSLAM Helsinki team: Mervi, Erkka, Linda, and Tatiana. Working with you I believe we could accomplish anything and I will always remember and cherish those events.

I want to send a long distance thank you to professor Carlos Linares, professor Jose Daniel Garcia, and doctor David Díez from Universidad Carlos III of Madrid, although they may not read this. Now I can appreciate the efforts you make to improve Spanish science, and if one day I am able to continue that work in Spain it will be thanks to scientists like you who keep pushing against the elements and budget cuts.

All the teaching and support was complemented by a large support network of wonderful people that put up with me week after week, and year after year. If not for the supportive friends participating in all kinds of sports events and therefore helping keep my sanity. Be it bouldering, rowing or training for an obstacle course race, know that it mattered! A special thanks to our team captain and overall life saver in and out of the lab Tiia Pelkonen, and my gym buddy Erdogan Peckan Erkan. To all my friends in no order of relevance: Elsa, Tini, Ale, Emilia, Sampo, Aki, Mikko, Gaurav, Marco, Evisa, Karen, Elyem, and Gaja.

I wish to send a warm thank you to two talented musicians; Antti Siltanen and doctor Tapani Vitala, who welcomed me into their band when I could barely keep a beat. Making noise with friends was the best form of therapy I could wish for. Also thanks to old and new music pals for the fun rehearsals: Hans, Lauri, Julian, Joonas, Prima, Kul.

This book was written in the middle of the COVID pandemic. The PhDForum created by Dr Donna Peach provided some resemblance of a routine and library environment necessary for writing. Thanks to this I connected with a beautiful bunch of people that I would have never met otherwise. They help me keep working while focusing on wellbeing and laughing often. A special thanks to the doctors and pre-docs: Jo, Asma, Niamh, Colin, Grace, Noma, Ruth, Bernadette, Sónia. We

will all get there one frog at a time!

To Elza and Kari, for being the best neighbors I have ever had. When I first started the PhD I also moved to the apartment next to yours. You are the heart of the community and are always willing to help everyone around. Thank you for walking Coco when I had long days in the lab, for delivering pullat and pancakes on those long days, for fixing what he broke, and always being there for me.

To Mervi and Esko, for welcoming me into their family and offering their help all the time, especially with Coco and Blanca even with very short notice.

To my parents Javier and Titina, and my sister Alba, thank you from the bottom of my heart. Thank you for supporting all my crazy ideas, for the constant encouragement to tackle challenges head on, for never judging me and for giving me a safe home. I feel privileged to have this family. You did this. Also a warm thanks to all my family, who keep me connected to Spain through weird and funny memes in the family chat.

Finally, my deepest most heart-felt thanks goes to Hans, for your love and for knowing when I needed a coffee or a glass of wine, and for too many reasons to put into words, but in short for helping make sense of the cat that is this life without taking it apart.

TL;DR: We did it! And everyone who is still around deserves a cookie.

Julia Casado Cuervo
Helsinki, 2021

Page(s)

References

1 [1] Chave, J. (2004) Neutral theory and community ecology. *Ecology Letters* **7**, 241–253.

1, 5, 6, 12 [2] Lodish, H, Berk, A, Zipursky, S. L, Matsudaira, P, Baltimore, D, & Darnell, J. (2000) *Molecular Cell Biology*. (W. H. Freeman), 4th edition.

1 [3] Berger, S. L, Kouzarides, T, Shiekhattar, R, & Shilatifard, A. (2009) An operational definition of epigenetics. *Genes & Development* **23**, 781–783. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

1, 4 [4] Weinberg, R. A. (2014) *The biology of cancer*. (Garland Science, Taylor & Francis Group, New York), Second edition edition.

1, 43 [5] Mukherjee, S. (2011) *The Emperor of All Maladies: A Biography of Cancer*. (Scribner).

2, 12 [6] Levy, S. E & Myers, R. M. (2016) Advancements in Next-Generation Sequencing. *Annual Review of Genomics and Human Genetics* **17**, 95–115. _eprint: https://doi.org/10.1146/annurev-genom-083115-022413.

2 [7] Peng, A, Mao, X, Zhong, J, Fan, S, & Hu, Y. (2020) Single-Cell Multi-Omics and Its Prospective Application in Cancer Biology. *PROTEOMICS* **20**, 1900271. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pmic.201900271.

2 [8] Smittenaar, C. R, Petersen, K. A, Stewart, K, & Moitt, N. (2016) Cancer incidence and mortality projections in the UK until 2035. *British Journal of Cancer* **115**, 1147–1155.

2, 5, 11 [9] Holohan, C, Van Schaeybroeck, S, Longley, D. B, & Johnston, P. G. (2013) Cancer drug resistance: an evolving paradigm. *Nature Reviews Cancer* **13**, 714–726. Number: 10 Publisher: Nature Publishing Group.

2 [10] Hristova, V. A & Chan, D. W. (2019) Cancer biomarker discovery and translation: proteomics and beyond. *Expert review of proteomics* **16**, 93–103.

2 [11] Prensner, J. R, Chinnaiyan, A. M, & Srivastava, S. (2012) Systematic, Evidence-Based Discovery of Biomarkers at the NCI. *Clinical & experimental metastasis* **29**, 645–652.

2 [12] van der Laan, A. L & Boenink, M. (2015) Beyond Bench and Bedside: Disentangling the Concept of Translational Research. *Health Care Analysis* **23**, 32–49.

2 [13] Surkis, A, Hogle, J. A, DiazGranados, D, Hunt, J. D, Mazmanian, P. E, Connors, E, Westaby, K, Whipple, E. C, Adamus, T, Mueller, M, & Aphinyanaphongs, Y. (2016) Classifying publications from the clinical and translational science award program along the translational research spectrum: a machine learning approach. *Journal of Translational Medicine* **14**, 235.

[14] Broder, S & Cushing, M. (1993) Trends in Program Project Grant Funding at the National Cancer Institute. *Cancer Research* **53**, 477–484. Publisher: American Association for Cancer Research Section: Public Issues. 2

[15] Lindahl, T. (1993) Instability and decay of the primary structure of DNA. *Nature* **362**, 709–715. Number: 6422 Publisher: Nature Publishing Group. 2

[16] Amé, J.-C, Spenlehauer, C, & Murcia, G. d. (2004) The PARP superfamily. *BioEssays* **26**, 882–893. 2, 7

[17] Ashworth, A. (2016) A Synthetic Lethal Therapeutic Approach: Poly(ADP) Ribose Polymerase Inhibitors for the Treatment of Cancers Deficient in DNA Double-Strand Break Repair. *Journal of Clinical Oncology*. Publisher: American Society of Clinical Oncology. 2, 7

[18] Hanahan, D & Weinberg, R. A. (2000) The hallmarks of cancer. *Cell* **100**, 57–70. Publisher: Springer Berlin / Heidelberg. 4

[19] Hanahan, D & Weinberg, R. A. (2011) Hallmarks of cancer: the next generation. *Cell* **144**, 646–674. Publisher: Elsevier Inc. 4

[20] Alexandrov, L. B, Nik-Zainal, S, Wedge, D. C, Aparicio, S. A. J. R, Behjati, S, Biankin, A. V, Bignell, G. R, Bolli, N, Borg, A, Børresen-Dale, A.-L, Boyault, S, Burkhardt, B, Butler, A. P, Caldas, C, Davies, H. R, Desmedt, C, Eils, R, Eyfjörd, J. E, Foekens, J. A, Greaves, M, Hosoda, F, Hutter, B, Ilicic, T, Imbeaud, S, Imielinski, M, Jäger, N, Jones, D. T. W, Jones, D, Knappskog, S, Kool, M, Lakhani, S. R, López-Otín, C, Martin, S, Munshi, N. C, Nakamura, H, Northcott, P. A, Pajic, M, Papaemmanuil, E, Paradiso, A, Pearson, J. V, Puente, X. S, Raine, K, Ramakrishna, M, Richardson, A. L, Richter, J, Rosenstiel, P, Schlesner, M, Schumacher, T. N, Span, P. N, Teague, J. W, Totoki, Y, Tutt, A. N. J, Valdés-Mas, R, van Buuren, M. M, van 't Veer, L, Vincent-Salomon, A, Waddell, N, Yates, L. R, Zucman-Rossi, J, Andrew Futreal, P, McDermott, U, Lichter, P, Meyerson, M, Grimmond, S. M, Siebert, R, Campo, E, Shibata, T, Pfister, S. M, Campbell, P. J, & Stratton, M. R. (2013) Signatures of mutational processes in human cancer. *Nature* **500**, 415–421. Number: 7463 Publisher: Nature Publishing Group. 4

[21] Barillot, E, Calzone, L, Hupe, P, Vert, J.-P, & Zinovyev, A. (2012) *Computational Systems Biology of Cancer*, Chapman & Hall / Crc Mathematical and Computational Biology. (CRC Press). 5

[22] Hanahan, D & Coussens, L. M. (2012) Accessories to the Crime: Functions of Cells Recruited to the Tumor Microenvironment. *Cancer Cell* **21**, 309–322. Publisher: Cell Press. 5, 8

[23] Rivera, C. M & Ren, B. (2013) Mapping Human Epigenomes. *Cell* **155**, 39–55. Publisher: Elsevier. 5

[24] Biswas, S & Rao, C. M. (2017) Epigenetics in cancer: Fundamentals and Beyond. *Pharmacology & Therapeutics* **173**, 118–134. 6

[25] Easwaran, H, Tsai, H.-C, & Baylin, S. B. (2014) Cancer epigenetics: Tumor Heterogeneity, Plasticity of Stem-like States, and Drug Resistance. *Molecular cell* **54**, 716–727. 6

[26] Ahuja, N, Sharma, A. R, & Baylin, S. B. (2016) Epigenetic Therapeutics: A New Weapon in the War Against Cancer. *Annual review of medicine* **67**, 73–89. 6

[27] Branzei, D & Foiani, M. (2008) Regulation of DNA repair throughout the cell cycle. *Nature Reviews. Molecular Cell Biology* **9**, 297–308. 6

[28] Trenner, A & Sartori, A. A. (2019) Harnessing DNA Double-Strand Break Repair for Cancer Treatment. *Frontiers in Oncology* **9**. Publisher: Frontiers. 6, 8

[29] Chatterjee, N & Walker, G. C. (2017) Mechanisms of DNA damage, repair and mutagenesis. *Environmental and molecular mutagenesis* **58**, 235–263. 6

[30] Moore, J. K & Haber, J. E. (1996) Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in Saccharomyces cerevisiae. *Molecular and Cellular Biology* **16**, 2164–2173. 7

[31] Abkevich, V, Timms, K. M, Hennessy, B. T, Potter, J, Carey, M. S, Meyer, L. A, Smith-McCune, K, Broaddus, R, Lu, K. H, Chen, J, Tran, T. V, Williams, D, Iliev, D, Jammulapati, S, FitzGerald, L. M, Krivak, T, DeLoia, J. A, Gutin, A, Mills, G. B, & Lanchbury, J. S. (2012) Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *British Journal of Cancer* **107**, 1776–1782. 7

[32] Pietras, K & Ostman, A. (2010) Hallmarks of cancer: interactions with the tumor stroma. *Experimental Cell Research* **316**, 1324–1331. 8

[33] Turley, S. J, Cremasco, V, & Astarita, J. L. (2015) Immunological hallmarks of stromal cells in the tumour microenvironment. *Nature Reviews. Immunology* **15**, 669–682. 8, 9

[34] Neophytou, C. M, Pierides, C, Christodoulou, M. I, Costeas, P, Kyriakou, T. C, & Papageorgis, P. (2020) The Role of Tumor-Associated Myeloid Cells in Modulating Cancer Therapy. *Frontiers in Oncology* **10**. Publisher: Frontiers Media S.A. 9

[35] Ostman, A. (2012) The tumor microenvironment controls drug sensitivity. *Nature Medicine* **18**, 1332–1334. 9

[36] Noh, J.-Y, Seo, H, Lee, J, & Jung, H. (2020) Immunotherapy in Hematologic Malignancies: Emerging Therapies and Novel Approaches. *International Journal of Molecular Sciences* **21**. 9

[37] Huyghe, N, Baldin, P, & Van den Eynde, M. (2020) Immunotherapy with immune checkpoint inhibitors in colorectal cancer: what is the future beyond deficient mismatch-repair tumours? *Gastroenterology Report* **8**, 11–24. 9

[38] Ackermann, C. J, Adderley, H, Ortega-Franco, A, Khan, A, Reck, M, & Califano, R. (2020) First-Line Immune Checkpoint Inhibition for Advanced Non-Small-Cell Lung Cancer: State of the Art and Future Directions. *Drugs* **80**, 1783–1797. 9

[39] Seliger, B. (2019) Combinatorial Approaches With Checkpoint Inhibitors to Enhance Anti-tumor Immunity. *Frontiers in Immunology* **10**, 999. 9

[40] Friedberg, J. W. (2011) Relapsed/refractory diffuse large B-cell lymphoma. *Hematology. American Society of Hematology. Education Program* **2011**, 498–505. 9

[41] Pfreundschuh, M, Trümper, L, Osterborg, A, Pettengell, R, Trneny, M, Imrie, K, Ma, D, Gill, D, Walewski, J, Zinzani, P.-L, Stahel, R, Kvaloy, S, Shpilberg, O, Jaeger, U, Hansen, M, Lehtinen, T, López-Guillermo, A, Corrado, C, Scheliga, A, Milpied, N, Mendila, M, Rashford, M, Kuhnt, E, Loeffler, M, & MabThera International Trial Group. (2006) CHOP-like chemotherapy plus rituximab versus CHOP-like chemotherapy alone in young patients with good-prognosis diffuse large-B-cell lymphoma: a randomised controlled trial by the MabThera International Trial (MInT) Group. *The Lancet. Oncology* **7**, 379–391. 9

[42] Coiffier, B, Lepage, E, Briere, J, Herbrecht, R, Tilly, H, Bouabdallah, R, Morel, P, Van Den Neste, E, Salles, G, Gaulard, P, Reyes, F, Lederlin, P, & Gisselbrecht, C. (2002) CHOP chemotherapy plus rituximab compared with CHOP alone in elderly patients with diffuse large-B-cell lymphoma. *The New England Journal of Medicine* **346**, 235–242. 9

[43] Roschewski, M, Staudt, L. M, & Wilson, W. H. (2014) Diffuse large B-cell lymphoma-treatment approaches in the molecular era. *Nature Reviews. Clinical Oncology* **11**, 12–23. 9

[44] Coiffier, B & Sarkozy, C. (2016) Diffuse large B-cell lymphoma: R-CHOP failure—what to do? *Hematology* **2016**, 366–378. Publisher: American Society of Hematology. 9

[45] Palmer, A. C, Chidley, C, & Sorger, P. K. (2019) A curative combination cancer therapy achieves high fractional cell killing through low cross-resistance and drug additivity. *eLife* **8**, e50036. Publisher: eLife Sciences Publications, Ltd. 10

[46] Sermer, D, Pasqualucci, L, Wendel, H.-G, Melnick, A, & Younes, A. (2019) Emerging epigenetic-modulating therapies in lymphoma. *Nature reviews. Clinical oncology* **16**, 494–507. 10

[47] Batlevi, C. L, Kasamon, Y, Bociek, R. G, Lee, P, Gore, L, Copeland, A, Sorensen, R, Ordentlich, P, Cruickshank, S, Kunkel, L, Buglio, D, Hernandez-Ilizaliturri, F, & Younes, A. (2016) ENGAGE- 501: phase II study of entinostat (SNDX-275) in relapsed and refractory Hodgkin lymphoma. *Haematologica* **101**, 968–975. 10

[48] Ribrag, V, Morschhauser, F, McKay, P, Salles, G. A, Batlevi, C. L, Schmitt, A, Tilly, H, Cartron, G, Thieblemont, C, Fruchart, C, Gribben, J. G, Lamy, T, Le Gouill, S, Bouabdallah, R, Dickinson, M, Opat, S, Adib, D, Blakemore, S. J, Larus, J, & Johnson, P. (2018) Interim Results from an Ongoing Phase 2 Multicenter Study of Tazemetostat, an EZH2 Inhibitor, in Patients with Relapsed or Refractory (R/R) Diffuse Large B-Cell Lymphoma (DLBCL). *Blood* **132**, 4196–4196. Publisher: American Society of Hematology. 10

[49] Klint, A, Tryggvadóttir, L, Bray, F, Gislum, M, Hakulinen, T, Storm, H. H, & Engholm, G. (2010) Trends in the survival of patients diagnosed with cancer in female genital organs in the Nordic countries 1964-2003 followed up to the end of 2006. *Acta Oncologica (Stockholm, Sweden)* **49**, 632–643. 10

[50] Lindemann, K, Gao, B, Mapagu, C, Fereday, S, Emmanuel, C, Alsop, K, Traficante, N, Australian Ovarian Cancer Study Group, Harnett, P. R, Bowtell, D. D. L, & deFazio, A. (2018) Response rates to second-line platinum-based therapy in ovarian cancer patients challenge the clinical definition of platinum resistance. *Gynecologic Oncology* **150**, 239–246. 10

[51] Bowtell, D. D, Böhm, S, Ahmed, A. A, Aspuria, P.-J, Bast, R. C, Beral, V, Berek, J. S, Birrer, M. J, Blagden, S, Bookman, M. A, Brenton, J. D, Chiappinelli, K. B, Martins, F. C, Coukos, G, Drapkin, R, Edmondson, R, Fotopoulou, C, Gabra, H, Galon, J, Gourley, C, Heong, V, Huntsman, D. G, Iwanicki, M, Karlan, B. Y, Kaye, A, Lengyel, E, Levine, D. A, Lu, K. H, McNeish, I. A, Menon, U, Narod, S. A, Nelson, B. H, Nephew, K. P, Pharoah, P, Powell, D. J, Ramos, P, Romero, I. L, Scott, C. L, Sood, A. K, Stronach, E. A, & Balkwill, F. R. (2015) Rethinking ovarian cancer II: reducing mortality from high-grade serous ovarian cancer. *Nature Reviews. Cancer* **15**, 668–679. 10

[52] Bowtell, D. D. L. (2010) The genesis and evolution of high-grade serous ovarian cancer. *Nature Reviews. Cancer* **10**, 803–808. 10

[53] The Cancer Genome Atlas Research Network. (2011) Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615. 10

[54] Fong, P. C, Boss, D. S, Yap, T. A, Tutt, A, Wu, P, Mergui-Roelvink, M, Mortimer, P, Swaisland, H, Lau, A, O'Connor, M. J, Ashworth, A, Carmichael, J, Kaye, S. B, Schellens, J. H. M, & de Bono, J. S. (2009) Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. *The New England Journal of Medicine* **361**, 123–134. 10

[55] Verhaak, R. G. W, Tamayo, P, Yang, J.-Y, Hubbard, D, Zhang, H, Creighton, C. J, Fereday, S, Lawrence, M, Carter, S. L, Mermel, C. H, Kostic, A. D, Etemadmoghadam, D, Saksena, G, Cibulskis, K, Duraisamy, S, Levanon, K, Sougnez, C, Tsherniak, A, Gomez, S, Onofrio, R, Gabriel, S, Chin, L, Zhang, N, Spellman, P. T, Zhang, Y, Akbani, R, Hoadley, K. A, Kahn, A, Köbel, M, Huntsman, D, Soslow, R. A, Defazio, A, Birrer, M. J, Gray, J. W, Weinstein, J. N, Bowtell, D. D, Drapkin, R, Mesirov, J. P, Getz, G, Levine, D. A, Meyerson, M, & Cancer Genome Atlas Research Network. (2013) Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. *The Journal of Clinical Investigation* **123**, 517–525. 10

[56] Matte, I, Bessette, P, & Piché, A. (2017) in *Ascites - Physiopathology, Treatment, Complications and Prognosis*. (InTech). 10, 41

[57] Burrows, V. (2008) FDA and Clinical Drug Trials: A Short History, Technical report. 10

[58] Barretina, J, Caponigro, G, Stransky, N, Venkatesan, K, Margolin, A. A, Kim, S, Wilson, C. J, Lehár, J, Kryukov, G. V, Sonkin, D, Reddy, A, Liu, M, Murray, L, Berger, M. F, Monahan, J. E, Morais, P, Meltzer, J, Korejwa, A, Jané-Valbuena, J, Mapa, F. A, Thibault, J, Bric-Furlong, E, Raman, P, Shipway, A, Engels, I. H, Cheng, J, Yu, G. K, Yu, J, Aspesi, P, de Silva, M, Jagtap, K, Jones, M. D, Wang, L, Hatton, C, Palescandolo, E, Gupta, S, Mahan, S, Sougnez, C, Onofrio, R. C, Liefeld, T, MacConaill, L, Winckler, W, Reich, M, Li, N, Mesirov, J. P, Gabriel, S. B, Getz, G, Ardlie, K, Chan, V, Myer, V. E, Weber, B. L, Porter, J, Warmuth, M, Finan, P, Harris, J. L, Meyerson, M, Golub, T. R, Morrissey, M. P, Sellers, W. R, Schlegel, R, & Garraway, L. A. (2012) The Cancer Cell Line Encyclopedia enables predictive modeling of anticancer drug sensitivity. *Nature* **483**, 603–607. 10

[59] Kim, J, Koo, B.-K, & Knoblich, J. A. (2020) Human organoids: model systems for human biology and medicine. *Nature Reviews Molecular Cell Biology* **21**, 571–584. Number: 10 Publisher: Nature Publishing Group. 10, 43

[60] Zhang, W, Moore, L, & Ji, P. (2011) Mouse models for cancer research. *Chinese Journal of Cancer* **30**, 149–152. 10

[61] Weinstein, J. N, Collisson, E. A, Mills, G. B, Shaw, K. M, Ozenberger, B. A, Ellrott, K, Shmulevich, I, Sander, C, & Stuart, J. M. (2013) The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nature genetics* **45**, 1113–1120. 11

[62] The ENCODE Project Consortium. (2012) An integrated encyclopedia of {{dna}} elements in the human genome. *Nature* **489**, 57–73. 11

[63] Zehir, A, Benayed, R, Shah, R. H, Syed, A, Middha, S, Kim, H. R, Srinivasan, P, Gao, J, Chakravarty, D, Devlin, S. M, Hellmann, M. D, Barron, D. A, Schram, A. M, Hameed, M, Dogan, S, Ross, D. S, Hechtman, J. F, DeLair, D. F, Yao, J, Mandelker, D. L, Cheng, D. T, Chandramohan, R, Mohanty, A. S, Ptashkin, R. N, Jayakumaran, G, Prasad, M, Syed, M. H, Rema, A. B, Liu, Z. Y, Nafa, K, Borsu, L, Sadowska, J, Casanova, J, Bacares, R, Kiecka, I. J, Razumova, A, Son, J. B, Stewart, L, Baldi, T, Mullaney, K. A, Al-Ahmadie, H, Vakiani, E, Abeshouse, A. A, Penson, A. V, Jonsson, P, Camacho, N, Chang, M. T, Won, H. H, Gross, B. E, Kundra, R, Heins, Z. J, Chen, H.-W, Phillips, S, Zhang, H, Wang, J, Ochoa, A, Wills, J, Eubank, M, Thomas, S. B, Gardos, S. M, Reales, D. N, Galle, J, Durany, R, Cambria, R, Abida, W, Cercek, A, Feldman, D. R, Gounder, M. M, Hakimi, A. A, Harding, J. J, Iyer, G, Janjigian, Y. Y, Jordan, E. J, Kelly, C. M, Lowery, M. A, Morris, L. G, Omuro, A. M, Raj, N, Razavi, P, Shoushtari, A. N, Shukla, N, Soumerai, T. E, Varghese, A. M, Yaeger, R, Coleman, J, Bochner, B, Riely, G. J, Saltz, L. B, Scher, H. I, Sabbatini, P. J, Robson, M. E, Klimstra, D. S, Taylor, B. S, Baselga, J, Schultz, N, Hyman, D. M, Arcila, M. E, Solit, D. B, Ladanyi, M, & Berger, M. F. (2017) Mutational Landscape of Metastatic Cancer Revealed from Prospective Clinical Sequencing of 10,000 Patients. *Nature medicine* **23**, 703–713. 11

[64] Pukkala, E, Andersen, A, Berglund, G, Gislefoss, R, Gudnason, V, Hallmans, G, Jellum, E, Jousilahti, P, Knekt, P, Koskela, P, Kyyrönen, P. P, Lenner, P, Luostarinen, T, Löve, A, Ogmundsdóttir, H, Stattin, P, Tenkanen, L, Tryggvadóttir, L, Virtamo, J, Wadell, G, Widell, A, Lehtinen, M, & Dillner, J. (2007) Nordic biological specimen banks as basis for studies of cancer causes and control–more than 2 million sample donors, 25 million person years and 100,000 prospective cancers. *Acta Oncologica (Stockholm, Sweden)* **46**, 286–307. 11

[65] Creighton, C. J. (2018) Making use of cancer genomic databases. *Current protocols in molecular biology* **121**, 19.14.1–19.14.13. 11

[66] Muir, P, Li, S, Lou, S, Wang, D, Spakowicz, D. J, Salichos, L, Zhang, J, Weinstock, G. M, Isaacs, F, Rozowsky, J, & Gerstein, M. (2016) The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biology* **17**, 53. 12

[67] Van der Auwera, G. A, Carneiro, M. O, Hartl, C, Poplin, R, Del Angel, G, Levy-Moonshine, A, Jordan, T, Shakir, K, Roazen, D, Thibault, J, Banks, E, Garimella, K. V, Altshuler, D, Gabriel, S, & DePristo, M. A. (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics* **43**, 11.10.1–11.10.33. 12, 26

[68] Conesa, A, Madrigal, P, Tarazona, S, Gomez-Cabrero, D, Cervera, A, McPherson, A, Szcześniak, M. W, Gaffney, D. J, Elo, L. L, Zhang, X, & Mortazavi, A. (2016) A survey of best practices for RNA-seq data analysis. *Genome Biology* **17**, 13. 12

[69] Gagan, J & Van Allen, E. M. (2015) Next-generation sequencing to guide cancer therapy. *Genome Medicine* **7**, 80. 12

[70] Alexandrov, L, Nik-Zainal, S, Wedge, D, Campbell, P, & Stratton, M. (2013) Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports* **3**, 246–259. 12

[71] Meyerson, M, Gabriel, S, & Getz, G. (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics* **11**, 685–696. 12

[72] Merlo, L. M. F, Pepper, J. W, Reid, B. J, & Maley, C. C. (2006) Cancer as an evolutionary and ecological process. *Nature Reviews. Cancer* **6**, 924–935. 12

[73] Kelly, R. T. (2020) Single-Cell Proteomics: Progress and Prospects. *Molecular & Cellular Proteomics*. Publisher: American Society for Biochemistry and Molecular Biology Section: Review. 12

[74] Bendall, S. C, Simonds, E. F, Qiu, P, Amir, E.-a. D, Krutzik, P. O, Finck, R, Bruggner, R. V, Melamed, R, Trejo, A, Ornatsky, O. I, Balderas, R. S, Plevritis, S. K, Sachs, K, Pe'er, D, Tanner, S. D, & Nolan, G. P. (2011) Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. *Science (New York, N.y.)* **332**, 687–696. 12

[75] Rybakowska, P, Alarcón-Riquelme, M. E, & Marañón, C. (2020) Key steps and methods in the experimental design and data analysis of highly multi-parametric flow and mass cytometry. *Computational and Structural Biotechnology Journal* **18**, 874–886. 13

[76] Bagwell, C. B, Hunsberger, B, Hill, B, Herbert, D, Bray, C, Selvanantham, T, Li, S, Villasboas, J. C, Pavelko, K, Strausbauch, M, Rahman, A, Kelly, G, Asgharzadeh, S, Gomez-Cabrero, A, Behbehani, G, Chang, H, Lyberger, J, Montgomery, R, Zhao, Y, Inokuma, M, Goldberger, O, & Stelzer, G. (2020) Multi-site reproducibility of a human immunophenotyping assay in whole blood and peripheral blood mononuclear cells preparations using CyTOF technology coupled with Maxpar Pathsetter, an automated data analysis system. *Cytometry. Part B, Clinical Cytometry* **98**, 146–160. 13

[77] Renz, M. (2013) Fluorescence microscopy-a historical and technical perspective. *Cytometry. Part A: The Journal of the International Society for Analytical Cytology* **83**, 767–779. 14

[78] Caicedo, J. C, Goodman, A, Karhohs, K. W, Cimini, B. A, Ackerman, J, Haghighi, M, Heng, C, Becker, T, Doan, M, McQuin, C, Rohban, M, Singh, S, & Carpenter, A. E. (2019) Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. *Nature Methods* **16**, 1247–1253. 14

[79] Lin, J.-R, Izar, B, Wang, S, Yapp, C, Mei, S, Shah, P. M, Santagata, S, & Sorger, P. K. (2018) Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes. *eLife* **7**, e31657. Publisher: eLife Sciences Publications, Ltd. 15, 20, 42

[80] Goltsev, Y, Samusik, N, Kennedy-Darling, J, Bhate, S, Hale, M, Vazquez, G, Black, S, & Nolan, G. P. (2018) Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell* **174**, 968. Publisher: Elsevier. 15

[81] Angelo, M, Bendall, S. C, Finck, R, Hale, M. B, Hitzman, C, Borowsky, A. D, Levenson, R. M, Lowe, J. B, Liu, S. D, Zhao, S, Natkunam, Y, & Nolan, G. P. (2014) Multiplexed ion beam imaging (MIBI) of human breast tumors. *Nature medicine* **20**, 436–442. 15

[82] Giesen, C, Wang, H. A. O, Schapiro, D, Zivanovic, N, Jacobs, A, Hattendorf, B, Schüffler, P. J, Grolimund, D, Buhmann, J. M, Brandt, S, Varga, Z, Wild, P. J, Günther, D, & Bodenmiller, B. (2014) Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature Methods* **11**, 417–422. 15

[83] Kitaeva, K. V, Rutland, C. S, Rizvanov, A. A, & Solovyeva, V. V. (2020) Cell Culture Based in vitro Test Systems for Anticancer Drug Screening. *Frontiers in Bioengineering and Biotechnology* **8**. Publisher: Frontiers. 15

[84] Nierode, G, Kwon, P. S, Dordick, J. S, & Kwon, S.-J. (2016) Cell-Based Assay Design for High-Content Screening of Drug Candidates. *Journal of microbiology and biotechnology* **26**, 213–225. 15

[85] Borisy, A. A, Elliott, P. J, Hurst, N. W, Lee, M. S, Lehar, J, Price, E. R, Serbedzija, G, Zimmermann, G. R, Foley, M. A, Stockwell, B. R, & Keith, C. T. (2003) Systematic discovery of multicomponent therapeutics. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 7977–7982. 15

[86] Tang, J, Wennerberg, K, & Aittokallio, T. (2015) What is synergy? The Saariselkä agreement revisited. *Frontiers in Pharmacology* **6**, 181. 15

[87] Lehár, J, Krueger, A. S, Avery, W, Heilbut, A. M, Johansen, L. M, Price, E. R, Rickles, R. J, Short Iii, G. F, Staunton, J. E, Jin, X, Lee, M. S, Zimmermann, G. R, & Borisy, A. A. (2009) Synergistic drug combinations tend to improve therapeutically relevant selectivity. *Nature Biotechnology* **27**, 659–666. Number: 7 Publisher: Nature Publishing Group. 16

[88] Liu, B, Li, S, & Hu, J. (2004) Technological advances in high-throughput screening. *American Journal of Pharmacogenomics: Genomics-Related Research in Drug Development and Clinical Practice* **4**, 263–276. 16

[89] Potdar, S, Ianevski, A, Mpindi, J.-P, Bychkov, D, Fiere, C, Ianevski, P, Yadav, B, Wennerberg, K, Aittokallio, T, Kallioniemi, O, Saarela, J, & Östling, P. (2020) Breeze: an integrated quality control and data analysis application for high-throughput drug screening. *Bioinformatics (Oxford, England)* **36**, 3602–3604. 16

[90] Searls, D. B. (2010) The Roots of Bioinformatics. *PLOS Computational Biology* **6**, e1000809. Publisher: Public Library of Science. 16

[91] Van Der Maaten, L & Hinton, G. (2008) Visualizing Data using t-SNE, Technical report. Publication Title: Journal of Machine Learning Research Volume: 9. 17

[92] McInnes, L, Healy, J, & Melville, J. (2018) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 17

[93] Assmus, H. E, Herwig, R, Cho, K.-H, & Wolkenhauer, O. (2006) Dynamics of biological systems: role of systems biology in medical research. *Expert Review of Molecular Diagnostics* **6**, 891–902. Publisher: Taylor & Francis _eprint: https://doi.org/10.1586/14737159.6.6.891. 17

[94] Baker, M. (2016) 1,500 scientists lift the lid on reproducibility. *Nature News* **533**, 452. Section: News Feature. 17

[95] Lithgow, G. J, Driscoll, M, & Phillips, P. (2017) A long journey to reproducible results. *Nature News* **548**, 387. Section: Comment. 17

[96] Leprevost, F. d. V, Barbosa, V. C, Francisco, E. L, Perez-Riverol, Y, & Carvalho, P. C. (2014) On best practices in the development of bioinformatics software. *Frontiers in Genetics* **5**. Publisher: Frontiers. 17

[97] Papin, J. A, Mac Gabhann, F, Sauro, H. M, Nickerson, D, & Rampadarath, A. (2020) Improving reproducibility in computational biology research. *PLoS Computational Biology* **16**. 17

[98] Menegidio, F. B, Jabes, D. L, Costa de Oliveira, R, & Nunes, L. R. (2018) Dugong: a Docker image, based on Ubuntu Linux, focused on reproducibility and replicability for bioinformatics analyses. *Bioinformatics (Oxford, England)* **34**, 514–515. 17, 41

[99] O'Donoghue, S. I, Gavin, A.-C, Gehlenborg, N, Goodsell, D. S, Hériché, J.-K, Nielsen, C. B, North, C, Olson, A. J, Procter, J. B, Shattuck, D. W, Walter, T, & Wong, B. (2010) Visualizing biological data—now and in the future. *Nature Methods* **7**, S2–S4. Number: 3 Publisher: Nature Publishing Group. 17

[100] Almeida, J. S. (2010) Computational ecosystems for data-driven medical genomics. *Genome Medicine* **2**, 67–67. 17, 41

[101] Van Unen, V, Li, N, Molendijk, I, Van Bergen, J, Lelieveldt, B. P. F, & Correspondence, K. (2016) Mass Cytometry of the Human Mucosal Immune System Identifies Tissue- and Disease-Associated Immune Subsets. *Cell Press*. 19, 29, 30

[102] Konstantinopoulos, P. A, Waggoner, S, Vidal, G. A, Mita, M, Moroney, J. W, Holloway, R, Van Le, L, Sachdev, J. C, Chapman-Davis, E, Colon-Otero, G, Penson, R. T, Matulonis, U. A, Kim, Y. B, Moore, K. N, Swisher, E. M, Färkkilä, A, D'Andrea, A, Stringer-Reasor, E, Wang, J, Buerstatte, N, Arora, S, Graham, J. R, Bobilev, D, Dezube, B. J, & Munster, P. (2019) Single-Arm Phases 1 and 2 Trial of Niraparib in Combination With Pembrolizumab in Patients With Recurrent Platinum-Resistant Ovarian Carcinoma. *JAMA oncology*. 19, 20, 33

[103] Lin, J.-R, Fallahi-Sichani, M, & Sorger, P. K. (2015) Highly multiplexed imaging of single cells using a high-throughput cyclic immunofluorescence method. *Nature Communications* **6**, 8390. Number: 1 Publisher: Nature Publishing Group. 20

[104] Peng, T, Thorn, K, Schroeder, T, Wang, L, Theis, F. J, Marr, C, & Navab, N. (2017) A BaSiC tool for background and shading correction of optical microscopy images. *Nature Communications* **8**, 14836. Number: 1 Publisher: Nature Publishing Group. 20

[105] Rantanen, V, Valori, M, & Hautaniemi, S. (2014) Anima: Modular Workflow System for Comprehensive Image Data Analysis. *Frontiers in Bioengineering and Biotechnology* **2**. 22

[106] Cervera, A, Rantanen, V, Ovaska, K, Laakso, M, Nuñez-Fontarnau, J, Alkodsi, A, Casado, J, Facciotto, C, Häkkinen, A, Louhimo, R, Karinen, S, Zhang, K, Lavikka, K, Lyly, L, Pal Singh, M, & Hautaniemi, S. (2019) Anduril 2: upgraded large-scale data integration framework. *Bioinformatics (Oxford, England)* **35**, 3815–3817. 22, 41

[107] Kohonen, T. (1990) The self-organizing map. *Proceedings of the IEEE* **78**, 1464–1480. Conference Name: Proceedings of the IEEE. 23

[108] Van Gassen, S, Callebaut, B, Van Helden, M. J, Lambrecht, B. N, Demeester, P, Dhaene, T, & Saeys, Y. (2015) FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry. Part A: The Journal of the International Society for Analytical Cytology* **87**, 636–645. 23

[109] Forbes, S. A, Bhamra, G, Bamford, S, Dawson, E, Kok, C, Clements, J, Menzies, A, Teague, J. W, Futreal, P. A, & Stratton, M. R. (2008) The Catalogue of Somatic Mutations in Cancer (COSMIC). *Current Protocols in Human Genetics* **Chapter 10**, Unit 10.11. 26

[110] Griffith, M, Spies, N. C, Krysiak, K, McMichael, J. F, Coffman, A. C, Danos, A. M, Ainscough, B. J, Ramirez, C. A, Rieke, D. T, Kujan, L, Barnell, E. K, Wagner, A. H, Skidmore, Z. L, Wollam, A, Liu, C. J, Jones, M. R, Bilski, R. L, Lesurf, R, Feng, Y.-Y, Shah, N. M, Bonakdar, M, Trani, L, Matlock, M, Ramu, A, Campbell, K. M, Spies, G. C, Graubert, A. P, Gangavarapu, K, Eldred, J. M, Larson, D. E, Walker, J. R, Good, B. M, Wu, C, Su, A. I, Dienstmann, R, Margolin, A. A, Tamborero, D, Lopez-Bigas, N, Jones, S. J. M, Bose, R, Spencer, D. H, Wartman, L. D, Wilson, R. K, Mardis, E. R, & Griffith, O. L. (2017) CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature Genetics* **49**, 170–174. Number: 2 Publisher: Nature Publishing Group. 26

[111] Cotto, K. C, Wagner, A. H, Feng, Y.-Y, Kiwala, S, Coffman, A. C, Spies, G, Wollam, A, Spies, N. C, Griffith, O. L, & Griffith, M. (2018) DGIdb 3.0: a redesign and expansion of the drug-gene interaction database. *Nucleic Acids Research* **46**, D1068–D1073. 26

[112] Wishart, D. S, Feunang, Y. D, Guo, A. C, Lo, E. J, Marcu, A, Grant, J. R, Sajed, T, Johnson, D, Li, C, Sayeeda, Z, Assempour, N, Iynkkaran, I, Liu, Y, Maciejewski, A, Gale, N, Wilson, A, Chin, L, Cummings, R, Le, D, Pon, A, Knox, C, & Wilson, M. (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* **46**, D1074–D1082. 26

[113] Whirl-Carrillo, M, McDonagh, E. M, Hebert, J. M, Gong, L, Sangkuhl, K, Thorn, C. F, Altman, R. B, & Klein, T. E. (2012) Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology and Therapeutics* **92**, 414–417. 26, 42

[114] Fokkema, I. F. A. C, Taschner, P. E. M, Schaafsma, G. C. P, Celli, J, Laros, J. F. J, & Dunnen, J. T. d. (2011) LOVD v.2.0: the next generation in gene variant databases. *Human Mutation* **32**, 557–563. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/humu.21438. 26

[115] Olivier, M, Eeles, R, Hollstein, M, Khan, M. A, Harris, C. C, & Hainaut, P. (2002) The IARC TP53 database: new online mutation analysis and recommendations to users. *Human Mutation* **19**, 607–614. 26

[116] Andrews, S, Krueger, F, Segonds-Pichon, A, Biggins, L, Krueger, C, & Wingett, S. (2012) FastQC. Published: Babraham Institute. 26

[117] Bolger, A. M, Lohse, M, & Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120. Publisher: Oxford Academic. 26

[118] Li, H & Durbin, R. (2010) Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595. Publisher: Oxford Academic. 26

[119] (2018) Picard toolkit. Publication Title: Broad Institute, GitHub repository. 26

[120] Wang, K, Li, M, & Hakonarson, H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38**, e164. 26

[121] Nakamura, K, Terai, Y, Tanabe, A, Ono, Y. J, Hayashi, M, Maeda, K, Fujiwara, S, Ashihara, K, Nakamura, M, Tanaka, Y, Tanaka, T, Tsunetoh, S, Sasaki, H, & Ohmichi, M. (2017) CD24 expression is a marker for predicting clinical outcome and regulates the epithelial-mesenchymal transition in ovarian cancer via both the Akt and ERK pathways. *Oncology Reports* **37**, 3189–3200. 32

[122] Gorshkov, K, Chen, C. Z, Marshall, R. E, Mihatov, N, Choi, Y, Nguyen, D.-T, Southall, N, Chen, K. G, Park, J. K, & Zheng, W. (2019) Advancing precision medicine with personalized drug screening. *Drug Discovery Today* **24**, 272–278. 35

[123] Reddy, A, Zhang, J, Davis, N. S, Moffitt, A. B, Love, C. L, Waldrop, A, Leppa, S, Pasanen, A, Meriranta, L, Karjalainen-Lindsberg, M.-L, Nørgaard, P, Pedersen, M, Gang, A. O, Høgdall, E, Heavican, T. B, Lone, W, Iqbal, J, Qin, Q, Li, G, Kim, S. Y, Healy, J, Richards, K. L, Fedoriw, Y, Bernal-Mizrachi, L, Koff, J. L, Staton, A. D, Flowers, C. R, Paltiel, O, Goldschmidt, N, Calaminici, M, Clear, A, Gribben, J, Nguyen, E, Czader, M. B, Ondrejka, S. L, Collie, A, Hsi, E. D, Tse, E, Au-Yeung, R. K. H, Kwong, Y.-L, Srivastava, G, Choi, W. W. L, Evens, A. M, Pilichowska, M, Sengar, M, Reddy, N, Li, S, Chadburn, A, Gordon, L. I, Jaffe, E. S, Levy, S, Rempel, R, Tzeng, T, Happ, L. E, Dave, T, Rajagopalan, D, Datta, J, Dunson, D. B, & Dave, S. S. (2017) Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma. *Cell* **171**, 481–494.e15. 38

[124] Newell, E. W & Cheng, Y. (2016) Mass cytometry: blessed with the curse of dimensionality. *Nature Immunology* **17**, 890–895. 41

[125] Wang, Z, Sun, K, Xiao, Y, Feng, B, Mikule, K, Ma, X, Feng, N, Vellano, C. P, Federico, L, Marszalek, J. R, Mills, G. B, Hanke, J, Ramaswamy, S, & Wang, J. (2019) Niraparib activates interferon signaling and potentiates anti-PD-1 antibody efficacy in tumor models. *Scientific Reports* **9**, 1853. 42

[126] Shen, J, Zhao, W, Ju, Z, Wang, L, Peng, Y, Labrie, M, Yap, T. A, Mills, G. B, & Peng, G. (2019) PARPi Triggers the STING-Dependent Immune Response and Enhances the Therapeutic Efficacy of Immune Checkpoint Blockade Independent of BRCAness. *Cancer Research* **79**, 311–319. 42

[127] Gulhan, D. C, Lee, J. J.-K, Melloni, G. E. M, Cortés-Ciriano, I, & Park, P. J. (2019) Detecting the mutational signature of homologous recombination deficiency in clinical samples. *Nature Genetics* **51**, 912–919. 42

[128] Tomfohr, J, Lu, J, & Kepler, T. B. (2005) Pathway level analysis of gene expression using singular value decomposition. *BMC bioinformatics* **6**, 225. 42

[129] Schapiro, D, Jackson, H. W, Raghuraman, S, Fischer, J. R, Zanotelli, V. R. T, Schulz, D, Giesen, C, Catena, R, Varga, Z, & Bodenmiller, B. (2017) histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nature Methods* **14**, 873–876. 42

[130] Oronsky, B, Oronsky, N, Knox, S, Fanger, G, & Scicinski, J. (2014) Episensitization: therapeutic tumor resensitization by epigenetic agents: a review and reassessment. *Anti-*
42 *Cancer Agents in Medicinal Chemistry* **14**, 1121–1127.

[131] Decaup, E, Rossi, C, Gravelle, P, Laurent, C, Bordenave, J, Tosolini, M, Tourette, A, Perrial, E, Dumontet, C, Poupot, M, Klein, C, Savina, A, Fournié, J.-J, & Bezombes, C. (2019) A Tridimensional Model for NK Cell-Mediated ADCC of Follicular Lymphoma. *Frontiers in*
43 *Immunology* **10**, 1943.

[132] Drolet, B. C & Lorenzi, N. M. (2011) Translational research: Understanding the continuum
43 from bench to bedside. *Translational Research* **157**, 1–5. Publisher: Mosby Inc.