Aus der Klinik für Gynäkologie und Geburtshilfe
der Heinrich-Heine-Universität Düsseldorf
Direktorin: Professor Dr. med. Tanja Fehm

Endocrine Resistance in Breast Cancer

Utilizing Liquid Biopsy and Next Generation Sequencing in Decentralized Multicenter Trials

Dissertation

zur Erlangung des Grades eines Doktors der Medizin
der Medizinischen Fakultät der Heinrich-Heine-Universität Düsseldorf

vorgelegt von
Jan-Philipp Cieslik
2022

Als Inaugural-dissertation gedruckt mit Genehmigung der
Medizinischen Fakultät der Heinrich-Heine-Universität Düsseldorf

gez.:

Dekan:	Professor Dr. Nikolaj Klöcker
Erstgutachter:	Professor Dr. Hans Neubauer
Zweitgutachter:	Professor Dr. Günter Niegisch

Dedicated to my family and friends, who
have helped me in every possible way.

Parts of this work have been published:

H Asperger☯, **J-P Cieslik☯**, B Alberter, C Köstler, B Polzer, V Müller, K Pantel, S Riethdorf, A Koch, A Hartkopf, L Wiesmüller, W Janni, F Schochter, B Rack, A Franken, D Niederacher, T Fehm & H Neubauer
"ViBiBa: Virtual BioBanking for the DETECT multicenter trial program-decentralized storage and processing." **Translational Oncology** 14.8 (2021): 101132.

☯These authors contributed equally

Summary

English

Introduction

Breast cancer (BC) is the leading type of cancer in women and the biggest contributor to cancer mortality. Analysis of tumor tissue (e.g., through biopsy) allows for valuable insight into the spatial and temporal heterogeneity of the disease. Liquid biopsy enables downstream analysis of tumor cells (or parts of them) through the patient blood. I co-developed the peer-reviewed and published software ViBiBa (Virtual Bio Banking), which helps multicenter trials with decentralized sample storage of special specimen like circulating tumor cells (CTCs) [1]. Additionally I propose an NGS panel (ENDOpanel) for single cell analysis and demonstrate its feasibility. At last, an Estrogen Receptor α (ERα/ESR1) sequencing project aims to find novel ESR1 mutations on single cells to find novel causes of endocrine resistance.

Methods

ViBiBa is a web platform built with PHP in the back end and MySQL as the database language. The front end utilizes the open source bootstrap framework and some additional plugins. Both the ESR1 sequencing project and the NGS panel work on single cells that were identified with the CellSearch technology and isolated using the CellCelector. The ENDOpanel is based on the SureSelectQXT platform and covers the exons of 12 proto-oncogenes and the complete range of the tumor suppressor gene PTEN.

Results

ViBiBa is currently in use for the sample management of the DETECT trial group. The platform automatically processes non-uniform data from multiple laboratories into a structured central database, accessible by all participating laboratories. The ENDOpanel covered all important ranges of the preselected genes and was successfully performed with single cells. Additionally, I successfully conducted the ESR1 sequencing project including CTCs from 25 metastatic breast cancer patients and identified 21 mutant CTCs.

Conclusions

My thesis covers three dimensions of breast cancer research: ViBiBa created new ways for real world collaboration in clinical trials. The ENDOpanel could be established as a novel gene panel in single cell analysis and its feasibility demonstrated. Lastly, the ESR1 sequencing project detected novel variants and the subsequent literature review and in silico analysis offered new hypotheses for endocrine resistance.

German

Einleitung

Brustkrebs ist die häufigste maligne Erkrankung der Frau und verursacht die meisten krebsbedingten Todesfälle. Analysen von Tumorgewebe (z.B. durch Biopsien) erlauben einen wertvollen Einblick in die räumliche und zeitliche Heterogenität der Erkrankung. Ein Teil meiner Arbeit basiert auf der mittlerweile peer-reviewten und publizierten Software ViBiBa (Virtual Bio Banking), welche ich mitentwickelt habe [1]. Sie erlaubt Multicenter-Studien eine dezentrale Lagerung und Verwaltung von raren Bioproben wie z.B. zirkulierenden Tumorzellen (CTCs). Zusätzlich zeige ich ein neues NGS Panel (ENDOpanel) und demonstriere dessen Funktionalität bei der Analyse einzelner Zellen wie CTCs. Zuletzt habe ich ein Sequenzierprojekt auf CTCs für den Estrogenrezeptor α (ERα/ESR1) durchgeführt, um neuartige Mutationen im ESR1-Gen und Hinweise auf neue Wege der endokrinen Resistenz zu finden.

Methoden

ViBiBa ist eine Webplattform, welche serverseitig PHP als Skriptsprache und MySQL als Datenbankdialekt nutzt. Die Nutzeroberfläche basiert auf dem quelloffenen Bootstrap Gerüst und einigen zusätzlichen Plugins. Sowohl das ESR1-Sequenzierprojekt, als auch das ENDOpanel fokussieren sich auf CTCs welche mittels der CellSearch Technologie angereichert und gefärbt sowie mittels CellCelector isoliert wurden. Das ENDOpanel basiert auf der SureSelect QXT Plattform und deckt den kodierenden Bereich von 12 Proto-Onkogenen sowie den gesamten Bereich des Tumorsuppressorgens PTEN ab.

Ergebnisse

ViBiBa wird momentan für das Bioprobenmanagement der DETECT Studiengruppe eingesetzt. Die Plattform prozessiert automatisch nicht uniforme Daten aus mehreren Laboren und führt diese in eine strukturierte, durch alle teilnehmenden Labore erreichbare, Datenbank ein. Das ENDOpanel deckt ausreichende Teile der ausgewählten Gene ab und konnte erfolgreich in Analysen einzelner Zellen angewandt werden. Im ERα-Sequenzierprojekt konnten CTCs von 25 Patientinnen analysiert werden, darunter fanden sich 21 mutierte CTCs.

Fazit

Mit ViBiBa konnte ein reales Problem der Kollaboration in Multicenter-Studien angesprochen werden. Ebenso konnte ich das ENDOpanel etablieren sowie im ERα-Sequenzierprojekt neue Mutationen mit möglicher klinischer Relevanz aufzeigen.

List of Abbreviations

Abbreviation	Explanation
AF	activation function
AI	aromatase inhibitor
BC	breast cancer
BM	basement membrane
bp	base pairs
BWA	Burrows-Wheeler Aligner
cfDNA	circulating free DNA
CNV	copy number variant
CTC	circulating tumor cells
ctDNA	circulating tumor DNA
CTMS	clinical trial management systems
DAPI	4',6-diamidino-2-phenylindole
DCIS	ductal carcinoma in situ
DLA	diagnostic leukapheresis
DNA	deoxyribonucleic acid
DTC	disseminated tumor cells
E2	estrogen
EpCAM	epithelial cell adhesion molecule
ER	estrogen receptor
ET	endocrine therapy
FDA	Food and Drug Administration
GnRH	gonadotropin-releasing hormone
HER2	human epidermal growth factor receptor 2
HR	hormone receptor
LB	liquid biopsy
LBD	ligand-binding domain
MBC	metastatic breast cancer
MySQL	structured query language
NGS	next generation sequencing
OS	overall survival
OU	organizational unit
PCR	polymerase chain reaction
PFS	progression-free survival
PHP	PHP: Hypertext Preprocessor
PI3K	phosphoinositide 3-kinase
PR	progesterone receptor
PT	primary tumor

Abbreviation	Explanation
QC	quality control
RT-PCR	real time polymerase chain reaction
SBS	sequencing by synthesis
SERD	selective estrogen receptor degrader
SERM	selective estrogen receptor modulator
SNP	single nucleotide polymorphism
SSO	single sign-on
TNBC	triple-negative breast cancer
UI	user interface

Contents

1 Introduction

1.1 Epidemiology (Germany)

Breast cancer (BC) is the leading type of cancer in women, accounting for 69 thousand new breast cancer cases and 6 thousand cases of in situ breast cancer each year [2]. Additionally, breast cancer causes 17.6% of cancer deaths, making it the greatest cause of death among all cancer subtypes [2]. Nearly 30% of newly diagnosed women are under 55 years old [2]. With advancing treatment options, the mortality rate has been reduced to a 10-year overall survival (OS) of 66% and adjusted (for cancer-related mortality) OS of 82% [2].

1.2 Screening and Etiology (Germany)

The German government is currently running a screening program offering women above the age of 30 regular screening visits with an additional biannual mammography screening for patients between 50 and 69 years [2]. A genetic component like BRCA 1/2 mutations accounts for only 5 to 10% of breast cancers [3]. Other risk factors are mainly associated with elevated hormonal levels, e.g., early menarche, late menopause, childlessness and a prolonged hormone replacement therapy [2].

1.3 Prognostic and Predictive Factors

While prognostic markers describe the prognosis of a patient independently of the future treatment, predictive factors relate to the outcome of the patient dependent on future therapy decisions. Traditional prognostic factors include lymph node invasion, tumor size/grading and hormonal receptor status [4]. Newer biomarkers include gene panels or mRNA real time polymerase chain reactions (RT-PCRs) such as Oncotype DX and liquid biopsy (LB) approaches [4]. While some traditional prognostic markers are also suitable as therapy predictors, e.g., the estrogen receptor (ER) expression status is used to determine if a patient is suitable for endocrine therapy, they do not accurately predict resistance to endocrine therapy [4]. Trials on the predictive value of DNA sequencing (from LB specimens) are still ongoing [4]. Since a few decades, breast cancer is generally divided into multiple subtypes by a set of molecular markers first described by Perou *et al.* and refined in the following years (Table 1). This classification allows distinguishing hormonal positive (luminal) from non-luminal HER2+ and triple negative breast cancers (TNBC). Based on this simple classification, therapy decisions and an initial prognosis can be made.

Table 1: Molecular Subtypes of Breast Cancer

Subtype	ER/PR Status	HER2-Status	Ki-67
Luminal A	positive	negative	low
Luminal B	positive	negative	high
Luminal B	positive	positive	any
Non luminal, HER2+	negative	positive	any
Triple negative	negative	negative	any

Molecular subtyping of breast cancer currently used in the clinic [5].

1.4 Therapy

Like in all malignant diseases, the therapy regime depends heavily on the stage of the breast cancer and the subtype of the primary tumor (PT). The therapeutic pathways in breast cancer can be divided into at least four categories: chemotherapy, radiation, surgery and endocrine therapy. Loco-regionally limited breast cancer is mainly treated with early (breast-conserving) surgery, biopsy of the sentinel lymph node and adjuvant radiation [6]. Higher grade and relapsing breast cancer patients are treated according to their individual risk profile [6].

1.4.1 Endocrine Therapy

Endocrine therapy (ET) in hormone receptor-positive breast cancer patients has become an integral part of the breast cancer therapy pathway. The roots of endocrine therapy in breast cancer date back to the 19^{th} century, when it was discovered that oophorectomy leads to a response in breast cancer patients [7]. After the discovery of the ER, a link between estrogen (E2) levels and cell proliferation in ER expressing cells could be shown in numerous studies [8]. Endocrine treatment can rely on blocking the production of E2 or its action on the estrogen receptor alpha (ERα). Blockage of E2 production can be achieved for instance by aromatase inhibition (AI), surgery (oophorectomy) or gonadotropin-releasing-hormone (GnRH) analogs. On the other hand, the action of E2 on ERα can be interfered with antiestrogens like selective estrogen receptor modulators (SERMs) such as tamoxifen or with selective estrogen receptor degraders (SERDs) such as fulvestrant. New recommendations suggest a need to routinely test hormone receptors not only in invasive breast cancer patients but also in patients with ductal carcinoma in situ (DCIS) [9]. This enables the prediction of endocrine responsiveness based on the expression status of the estrogen receptor [10]. Extended endocrine therapy (e.g., up to 10 years of adjuvant AI treatment) is currently recommended for all ER-positive and node-positive breast cancer patients [11]. The German guidelines currently recommend extending endocrine therapy until progress of the cancer [6]. In case of progression, the endocrine therapy should be switched to another endocrine treatment [6].

1.4.2 Chemotherapy

In case of a predicted higher relapse risk, chemotherapy might be indicated. According to the current German guidelines, one of the following criteria should be fulfilled for adjuvant chemotherapy [6]:

- "HER2+ tumors (from pT1b, N0; pT1a, N0 without additional risk factors: G3, ER/PR neg., Ki-67 high)" [translated]
- Triple-negative tumors
- "Luminal-B-tumors with increased relapse risk (Ki-67 high, G3, high-risk multi-gene assay, young age, lymphatic node involvement)" [translated]

In the metastatic setting chemotherapy is required when a quick tumor reduction is needed as a result of critical organ infiltration (e.g., lung or liver) [6].

1.5 Endocrine Resistance

Unfortunately, after initial response to endocrine therapy many women with ER+ metastatic breast cancer develop an acquired resistance. The mechanisms that lead to endocrine resistance are currently under intense research and include coregulator proteins, altered metabolism, growth factors (receptors), cell-cycle regulators, autophagy as well as changes in key pathways like the activation of the PI3K/Akt/mTOR pathway [12, 13]. While SERMs, SERDs and AIs lead to the development of endocrine resistance, the mechanisms seem to be different, as patients developing resistance after AI treatment respond to treatment with a SERM/SERD [14]. A target for acquired mutations after ET is the ligand-binding domain (LBD) of ERα, a hotspot for post-ET metastatic breast cancer (MBC) patients [15]. Currently, mutational testing in the clinic is focused on a few pre-established Estrogen Receptor 1 (ESR1) hotspot mutations such as Y537S or D538G. These hotspots have been linked with a worse outcome in MBC patients [16]. Most ESR1 mutations develop as a response to endocrine treatment, which constitutes a selection pressure on the heterogeneous cancer cell population [17]. The already established hotspot mutations then lead to endocrine resistance as they enable ERα to be activated even in the absence of a ligand [15, 18].

1.6 Formation of Metastases and Tumor Cell Dissemination

Tumor metastasis, like tumor proliferation, is a multi-step procedure [19]. While PTs (of all cancer entities) account for only 10% of deaths caused by cancer, metastases lead to 90% of deaths in cancer patients [19]. Cancer cells have to transform in order to leave the PT and spread throughout the body [19]. Often specific cancer entities prefer specific distant sites for metastases (e.g., breast cancer prefers to spread towards bones) [19]. The processes are still not fully understood and are under intense research [19]. As normal

breast tissue and breast cancer cells are epithelial cells, they have to break through the basement membrane (BM) (separating the epithelial cells from the deeper stroma) [19]. Breast cancer that has not broken through the BM yet, like DCIS, is considered to be mostly benign [19]. To enable motility and invasiveness the epithelial cells have to transform from their epithelial phenotype to a mesenchymal one [19]. This process is called epithelial-mesenchymal transition (EMT), which is not cancer specific as it can also occur during wound healing and embryogenesis [19]. After breaking through the BM, cancer cells can spread throughout the bordering stroma and gain access to blood and lymphatic vessels [19]. Next, the cells need to enter these vessels, a process called intravasation, which is still not well understood [19]. Once inside the vessels, the cells travel through the body, during this time they are subject to the physical forces inside the blood stream and at risk of tearing of the cell membrane and other hostile environmental factors [19]. These cells are now called circulating tumor cells (CTCs). If they survive the hostile environment in the blood, the CTCs get trapped in arterioles, e.g., in the lungs [19]. How CTCs reach destinations after flowing through the lung is still debated [19]. Through the process of extravasation the CTCs can invade new tissues and begin forming micrometastases [19]. The cumulative probability of a single cell surviving the whole process from PT to proliferating distant metastasis is extremely low, the literature refers to this as metastatic inefficiency [19]. In contrast to CTCs, tumor cells which disseminated into the bone marrow (DTCs) tend to accumulate [19].

1.7 Liquid Biopsy

The concept of LB tries to solve the problems derived from single time point tissue sampling [20–23]. Typically tissue samples are taken from the PT or in some cases from metastases. While the analysis of those specimens is helpful, it does not depict the broader heterogeneity of the cancer, especially as it changes over time with every line of treatment [24]. LB enables the detection of CTCs or parts of them (e.g., circulating tumor DNA (ctDNA)) from the patient blood providing real-time information for researchers and clinicians. Meanwhile, the enumeration of CTCs via the CellSearch system received FDA approval as a new prognostic marker with a multitude of clinical trials demonstrating CTC count as an effective marker for survival (both overall and progression-free survival (PFS)) in nearly all stages of breast cancer (therapy) [25–28]. In contrast, CTCs have yet to show a valuable predictive characteristic, e.g., in predicting the response/resistance to endocrine therapy [29]. There is a multitude of hurdles that hamper quick progress in CTC research. First of all, CTCs have to be processed in a timely manner, before a degradation of the cells takes place and they can no longer be distinguished from the debris around them. This leaves clinical trials with two options: either they use regular blood collection tubes and process the samples on site or expensive proprietary collection tubes like the CellSave preservation tubes fitted with a fixation agent, and ship the samples to a remote laboratory. CellSave preservation tubes enable processing times of

up to 96 hours from blood withdrawal to CellSearch and downstream analysis [24]. The fixation reagents in CellSave preservation tubes and similar products come at the cost of viability, as the CTCs are fixated and subsequent analysis is limited to enumeration and genomic analysis [24]. As CTC analysis requires very specialized equipment, most trials have to rely on preservation tubes as the blood samples have to be shipped to a laboratory and the specimen spends a significant amount of time in transit [24]. Even with the help of preservation tubes, the equipment for analysis is very sparse. This necessitates decentralized processing and storage of CTCs and its byproducts [24]. New LB assays are constantly being developed and are often based on gene sequencing approaches that are already established in solid tumor biopsies [24]. One example is the MSK-ACCESS panel for LB [30], which is based on the MSK-IMPACT panel [31] designed for solid tumors. Of special interest are changes in CTC count over the course of multiple treatment regimes, new mutations not found in the PT or a switch of protein expression. The DETECT III trial is one current study that explores HER2 targeted therapies in patients with a HER2 negative primary tumor and HER2 positive CTCs [32].

1.8 ENDOpanel

The ENDOpanel (Fig. 1) is a purpose-built next generation sequencing (NGS) library targeting all exons from 12 genes encoding proto-oncogenes (in alphabetical order: AKT1, AKT2, EIF4EBP1, ERBB2, ESR1, INPP4B, MTOR, PDL1, PDL2, PIK3CA, PIK3CB, RPS6KB1) and the complete range of the tumor suppressor gene PTEN. These genes were mainly selected in light of a suspected impact on endocrine resistance. While the ENDOpanel is centered around the PI3K/Akt/mTOR pathway, some additional genes (e.g., PDL1/2) were added to the NGS library to broaden the observed pathways.

1.9 PI3K/Akt/mTOR Pathway

As mentioned beforehand, the ENDOpanel focuses on the PI3K pathway with downstream targets like the Akt and mTOR proteins. The PI3K pathway is often altered in breast cancer as roughly 18-40% demonstrate PIK3CA hyperactivity while 8% are Akt overexpressed and 20-33% show a mutant or underexpressed PTEN [19]. Phosphatidylinositol 3-kinases (PI3Ks) are a family of proteins. Their name giving feature is the phosphorylation of the 3' hydroxyl residue of an inositol ring on a membrane-bound phosphatidylinositol [19]. The family is divided into three classes, which differ by their structure and function [33]. As shown, the ENDOpanel focuses on the PIK3CA gene which encodes a class I PI3K (p110α). A multitude of upstream pathways is able to activate PI3K (e.g., Ras, tyrosine kinases phosphorylation) [19, 33]. One of the main features of PI3K is the conversion of PIP2 [$PI(4,5)P_2$] into PIP3 [$PI(3,4,5)P_3$] [19]. Phosphatidylinositol (bi-/tri-)phosphates like PIP2 and PIP3 are mainly found as additions to the hydrophilic head groups of the lipid bilayers of the cell membrane [19]. The phosphorylated inositol

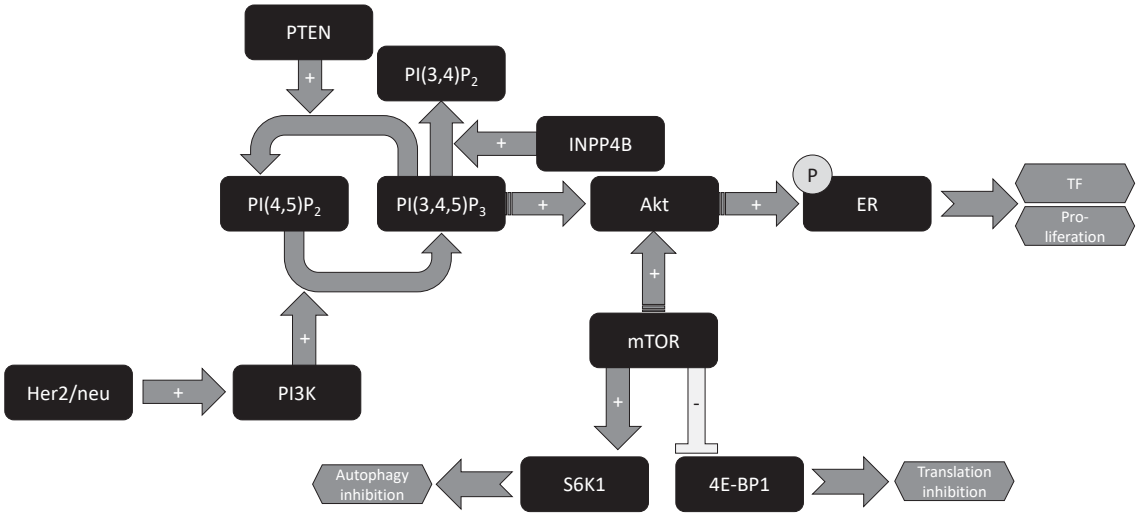


Figure 1: ENDOpanel Schematic

head group PIP3 plays a key role in the subsequent signaling, while tumor suppressor genes like PTEN are able to revert PIP3 into PIP2 thereby reducing the PI3K downstream signaling [19]. Another way to remove PIP3 is to transform it into another form of PIP2, which is phosphorylated on atypical sites ($PI(3,4)P_2$) [34]. Further, PIP3 attracts proteins with a pleckstrin homology domain (like the Akt kinase), which have a high affinity towards PIP3 [19]. After docking to PIP3, Akt is phosphorylated on two sites by the kinases PDK1 and PDK2 which leads to activation of Akt [19]. Activated Akt then inactivates proteins involved in apoptosis invocation and interacts with proteins that are involved in cell cycle regulation and thus cell proliferation [19]. Further, AKT phosphorylates ERα (Ser-167) which leads to increased ERα signaling [35]. Initially, mammalian target of rapamycin (mTOR) was seen as a downstream target of Akt, while new evidence suggests that mTOR is likely an upstream regulator of Akt [19]. Currently, two complexes of mTOR are described: TORC1 or "mTOR-Raptor complex" and TORC2 or "mTOR-Rictor complex". TORC1 controls a variety of genes that are involved in translation through phosphorylation, e.g., 4E-BP1 and S6K1 (both included in the ENDOpanel) [19]. Additionally, S6K1 is known to induce apoptosis and inhibit proliferation and EMT [36].

1.10 Estrogen Receptor

The estrogen receptor (ER) family consists of ERα and ERβ, the corresponding genes (ESR1 and ESR2) were discovered in 1985 and 1996 respectively [37]. While the two proteins are not expressed equally in different tissues, they share a few common features [37]. They are nuclear receptors as estrogen is a steroid and can pass freely through the cell membrane [37]. Additionally, they have three domains: the DNA-binding domain, the ligand-binding domain (LBD) and the N-terminal domain [37]. Further, two

activation function (AF) domains are described: a ligand-independent activation function (AF1) inside the N-terminal domain and a ligand-dependent activation function (AF2) inside the LBD [37]. Estrogen dependent activation of ERα is achieved through a conformation change after ligand binding, as helix 12 shifts into an agonist conformation [38]. Certain mutations in the LBD can lead to a conformation change that leads to helix 12 being in the agonist conformation without a ligand. Currently, this mechanism is described for the amino acids 536 to 538 and 380 [38, 39]. ERα and ERβ share 97% of the amino acid sequence in the DNA-binding domain, which is capable of binding to estrogen-responsive elements on the DNA, acting as a transcription factor [37]. The LBD/AF2, on the other hand, only shares 59% of the amino acid sequence, explaining the different ligand affinities of the receptors [37]. As ERα is overexpressed in breast tumors, it is a viable cancer therapy target, as discussed earlier. As reported in previous studies [17, 40, 41], mutations in the ESR1 gene mainly appear in a few known hotspots coding for the LBD of ERα. One study reports nearly as many patients with an ESR1 hotspot mutation as with a non-hotspot mutation in the genomic DNA of their CTCs [40]. Since the mutations outside of the hotspots seem to be scattered throughout the ESR1 gene, their effect on the receptor's function is hard to predict especially when such mutations are only detected in one or only a few CTCs. We currently lack a way to differentiate non-hotspot ESR1 mutations into variants that confer resistance to ET from variants which are random passenger mutations without effects. Some of the known ESR1 mutations influencing endocrine resistance result in constitutive and ligand-independent activation of ERα. These mutations are most often acquired under ET through selection pressure and correlate with a more aggressive disease making ESR1 mutations a promising biomarker [16, 17, 42]. Since ESR1 mutations are rarely present in the PT [17], the use of ESR1 as a biomarker is limited. Obtaining tissue samples of metastatic sites is often limited due to inaccessibility (e.g., brain metastases), thus limiting research in endocrine resistance and the use of mutations as a predictive marker. With the ongoing development of LB, characterization of the ESR1 gene on CTCs could allow for a timely evaluation of the mutation status in precision medicine tumor boards [20, 21].

1.11 Ligand Tunnels

1.11.1 From the Lock-Key towards the Keyhole-Lock-Key Model

The interaction of ligands with proteins like enzymes and receptors is described in various models [43]. One of the oldest and simpler models is the lock and key model proposed by Fischer in 1894, where a *lock* (e.g., enzyme) and a *key* (ligand) fit perfectly into each other [44, 45]. This model was later modified into the *induced fit* model, which takes into account the flexibility of the 3D protein structure; the ligand induces a conformation change in the protein that leads to a better fit [45]. Afterwards, buried ligand binding sites led to the development of the *keyhole-lock-key* model, as the ligand needs to traverse parts of the protein (*keyhole*) to reach its destination [43, 46]. The *keyhole* itself can dis-

criminate different ligands based on chemical forces and 3D structure, this filter function is thought to be as important as the fit of the ligand inside the LBD [46].

1.11.2 Tunnels in the Real World

Tunnel, channel and *keyhole* are terms that are often used interchangeably, this work will focus on the term tunnel, which can be defined as ”a pathway connecting a protein surface with an internal cavity” [46]. The *keyhole-lock-key* model has been pioneered in the field of enzymes, but the existence of *keyholes* has already been shown in all protein classes [46]. Proteins can have one or more tunnels that link the solvent around it with a buried LBD [46]. Ligand tunnels increase the selectivity of the protein and are linked with evolutional advantages, as new bottlenecks can dramatically change the selectivity of the protein [46]. While the impact of mutations in the LBD can be explained more easily, mutations outside of the LBD can lead to significant changes of the properties of the protein and may be explained by ligand tunnel modifications [46].

1.11.3 Bottlenecks and Other Forms of Ligand Selectivity

Of particular interest are the tightest points of a tunnel (bottlenecks), as they play a major role in ligand selectivity [46]. But not only the 3D structure of the tunnel determines its ligand selectivity, as electrostatic, polar and hydrophobic forces change the accessibility of the tunnel for certain ligands [46].

1.11.4 In Silico Tunnel Analysis

Modern computer technology enables the analysis of 3D protein structures that can be obtained freely from the internet from sources like the protein data bank (rcsb.org). The protein structures are generated from a wide range of methods, e.g., X-ray diffraction, and saved in a standardized file format (*.pdb*). These files often contain additional information like ligands or crystal waters, which need to be removed prior to computational analysis [46]. The position of the ligand can be used as a starting point for tunnel calculations, as it conveniently denotes the buried LBD; otherwise a starting position has to be estimated or manually defined [47]. Tools like Caver Web will output the geometry of all possible tunnels through the protein, as well as their length, average width, bottlenecks and tunnel profile [46, 47]. One of the main drawbacks is the static nature of this method, as the protein is only observed at one time point. While molecular dynamic simulations enable the modeling of protein movement and conformation changes over time, they are much more computationally expensive and are therefore often not suitable for screening applications [46, 47].

1.12 Next Generation Sequencing

1.12.1 Whole Genome Amplification

When working with single cells, some unique challenges arise as the starting DNA is extremely limited. To increase the amount of DNA whole genome amplification (WGA) needs to be performed. WGA is not only useful for single cell analysis but can also be found in the field of prenatal testing or forensics [48]. The main goal of WGA is balanced and high genomic coverage of the complete human genome without loss of one or both copies of a gene [48]. A multitude of WGA methods have been developed and most of them rely on the polymerase chain reaction (PCR) [48]. As PCR is best suited for short regions of amplification (amplicons), the complete human genome cannot be amplified with a standard PCR [48]. Early approaches used primers with random sequences (degenerate oligonucleotide primed PCR), while later methods rely on fragmentation of the DNA and ligation of adaptors with a known sequence (ligation-mediated PCR) [48]. The latter approach comes with the advantage of being deterministic, as no randomness is induced with degenerate primers [48]. An advancement of the ligation-mediated PCR is the single-cell comparative genomic hybridization, which utilizes the MSE1 restriction endonuclease that recognizes the 5'-TTAA-3' pattern, which has an average spacing of 126 base pairs (bp) in the human genome [48, 49]. After the digestion of the DNA, specially designed PCR-adaptor sequences are ligated at the end of the newly formed fragments, their special design is optimized for single cells. The presented work utilizes Ampli1, which is based on single-cell comparative genomic hybridization.

1.12.2 Library Preparation

To prepare the DNA for analysis, the DNA must be broken into fragments. There are multiple ways to achieve this: physical, chemical or enzymatic fragmentation [50]. One example of physical shearing is the Covaris system, which utilizes Adaptive Focused Acoustics to create fragments in the 100 - 1500 bp range [51]. Enzymatic fragmentation, on the other hand, can use restriction endonucleases or transposase based assays [50]. Chemical fragmentation does not play a major role in DNA analysis, as it is mostly used to break long RNA fragments with induced chemical forces [50]. Afterwards, the quality of the DNA (e.g., DNA concentration, fragmentation efficiency) has to be evaluated [50]. If the sample is deemed to be of sufficient quality, end repair is performed and it is purified using AMPure XP beads [50]. Depending on the kit used for library preparation a selection of predefined genomic ranges is performed. One example is the SureSelectQXT workflow, which uses a capture library that creates DNA library hybrids. These hybrids can later be captured with streptavidin-coated magnetic beads. Next, adapters are added through ligation on the 5' and 3' end of the DNA fragments, to guarantee that the fragment can bind to the flow cell in the sequencing step [50]. Additional adapters can be inserted to distinguish multiple samples from each other, this allows sequencing multiple

samples at once [50]. After another purification step, the library needs to be validated (e.g., through the Bioanalyzer platform) to assure its quality [50].

1.12.3 Sequencing by Synthesis

Actual sequencing of the DNA can be performed with various technologies. I will focus on the technology of Illumina since my laboratory mainly uses the Illumina MiSeq, which utilizes sequencing by synthesis (SBS) [50]. Afterwards, the library is loaded on to a flow cell [50]. The flow cell contains DNA probes allowing the library to bind through hybridization on to the glass [50, 52]. After hybridization the fragment is amplified, resulting in a clonal cluster [50]. Finally, the sequencing cycles can begin. The flow cell is imaged repeatedly as the single strands are replicated. Since fluorescently labeled nucleotides are used, the emission wavelength and intensity can be used to deduce the current base of every cluster simultaneously [50]. The sequence of DNA bases of a cluster is called a read, this information is then stored in a text file [50].

1.12.4 Computer Analysis

The generated text file used for downstream in silico analysis is saved in the FASTQ format, which stores the base sequences together with quality scores for each base [50]. Since the FASTQ files contain quality scores, we can generate a quality control (QC) report to check for problems in the data before processing and modifying it [50]. If the QC returns satisfactory results, we can proceed to the next stage. Alignment of the reads gives us information about the location of the read in the genome. The demonstrated work uses the *Genome Reference Consortium Human Build 38* (GRCh38 or HG38). Sequence alignment is performed by specialized algorithms that create an index file of the reference genome for a faster alignment process [50, 53]. One such algorithm is the Burrows-Wheeler Aligner (BWA) and its derivative BWA-MEM, which is ideal for sequence reads from the Illumina platform [50, 53]. The alignment algorithm then generates files that store the starting position of each read as well as mismatches with the reference genome and some additional quality information [50]. From this information the sequencing depth and coverage can be calculated. Both terms are only loosely defined and sometimes used interchangeably in the literature [54]. For the purpose of this work, I have defined depth for a specific base as the number of reads that include this base. Accordingly, I have defined coverage as the percentage of bases that have a depth of n. If an exon consists of 100 base pairs and 80 of them have a depth of 20 or more the exon gets a coverage of 80%.

1.13 Aims of my Thesis

The aim of this thesis is to acquire a holistic view of endocrine resistance in BC and present possible ways to predict endocrine resistance. To achieve this, the thesis is built

on three pillars.

The online platform ViBiBa (short for Virtual Bio Banking)[1] tries to solve the problem of low positivity rates in liquid biopsy trials. While leveraging the strength of decentralized processing and storage of CTCs, ViBiBa maintains the advantages of a centralized sample bank.

The ENDOpanel, established for single cell NGS analysis, tries to get a broader view of mutations in liquid biopsy samples that could lead to endocrine resistance. It is centered around the PI3K/AKT/mTOR pathway and aims to cover all exons of 12 genes plus the whole range of PTEN.

The ESR1 gene sequencing project on CTCs is focused on a simpler NGS approach, making it financially feasible for a large-scale project. With preselected CTCs (without hotspot mutations), I search for non-hotspot mutations and analyze them for possible effects on endocrine resistance.

2 Material and Methods

2.1 Used Devices

Table 2: Used Devices

Device	Manufacturer
Bioanalyzer 2100	Agilent, USA
CellCelector™	ALS, Germany
Celltracks Analyzer II®	Menarini Silicon Biosystems, Italy
Celltracks® Autoprep® System	Menarini Silicon Biosystems, Italy
Centrifuge: GS-15	Beckman Coulter, USA
Centrifuge: Megafuge 1.0	Heraeus, Germany
Centrifuge: Rotana	Hettich, Germany
Centrifuge: RotoFix 32 A	Hettich, Germany
Clean Bench	Clean Air Products, USA
CO_2-Incubator Function Line	Heraeus, Germany
Freezing Container: Mr. Frosty	Thermo Fisher Scientific™, USA
Microscope: AxioPlan 2	Zeiss, Germany
Microscope: DM IRB	Leica, Germany
MiSeq	Illumina, USA
Neubauer Counting Chamber	Paul Marienfeld, Germany
Orbital Shaker: Köttermann 4010	Köttermann, Germany
Shaking Water Bath: GFL 1083	GFL®, Germany
Thermocycler: Life ECO	Bioer, China
Thermocycler: T3000	Biometra, Germany

2.2 Used Materials

Table 3: Used Materials

Material	Manufacturer
Ampli1™ WGA Kit	Menarini Silicon Biosystems, Italy
Bioanalyzer: DNA 1000 Kit	Agilent, USA
Bioanalyzer: High Sensitivity DNA Kit	Agilent, USA
Cell Culture Flask: T-25	Sarstedt, Germany
CellSave® Preservation Tubes	Menarini Silicon Biosystems, Italy
CellSearch® Circulating Tumor Cell Kit	Menarini Silicon Biosystems, Italy
Fetal Bovine Serum	Thermo Fisher Scientific™, USA
HEPES	Thermo Fisher Scientific™, USA
KAPA2G Fast Multiplex Mix	Roche, Switzerland
Medium: RPMI 1640	Thermo Fisher Scientific™, USA
Microscope Slide	Paul Marienfeld, Germany
Multiplicom MID Dx	Agilent, USA
Penicillin/Streptomycin	Thermo Fisher Scientific™, USA
Pipette: Eppendorf	Eppendorf, Germany
Pipette: Stripette™	Thermo Fisher Scientific™, USA
Reaction Vessle: Safe-Lock Eppendorf	Eppendorf, Germany
SureSelectQXT	Agilent, USA

2.3 ViBiBa

2.3.1 User Interface

ViBiBa is designed with the CSS framework bootstrap (under MIT license) and some additional plugins, which are also mainly under the MIT license. After the user has logged in, the main application can be accessed. Login is possible through a classic user/password method or through a single sign-on (SSO) module. The UI (user interface) of the main application is divided into multiple sections. On the top, a static header displays the latest notifications and shows the name of the current user. A menu on the left side allows the user to navigate through the application. The site-specific content is displayed on the right-hand side, which is subdivided as well. [1]

2.3.2 Basic Structure

ViBiBa utilizes MySQL as a database server and PHP as the back end processing language. The default deployment method is via multiple docker containers united through a docker-compose configuration. This allows the isolation of services and enhances both maintainability and security [55]. ViBiBa creates a flow of data to transform non-uniform data from a single laboratory into a common pool of standardized data of all available samples (Fig. 3, Fig. 4). The data is processed in three steps. At first, raw data from

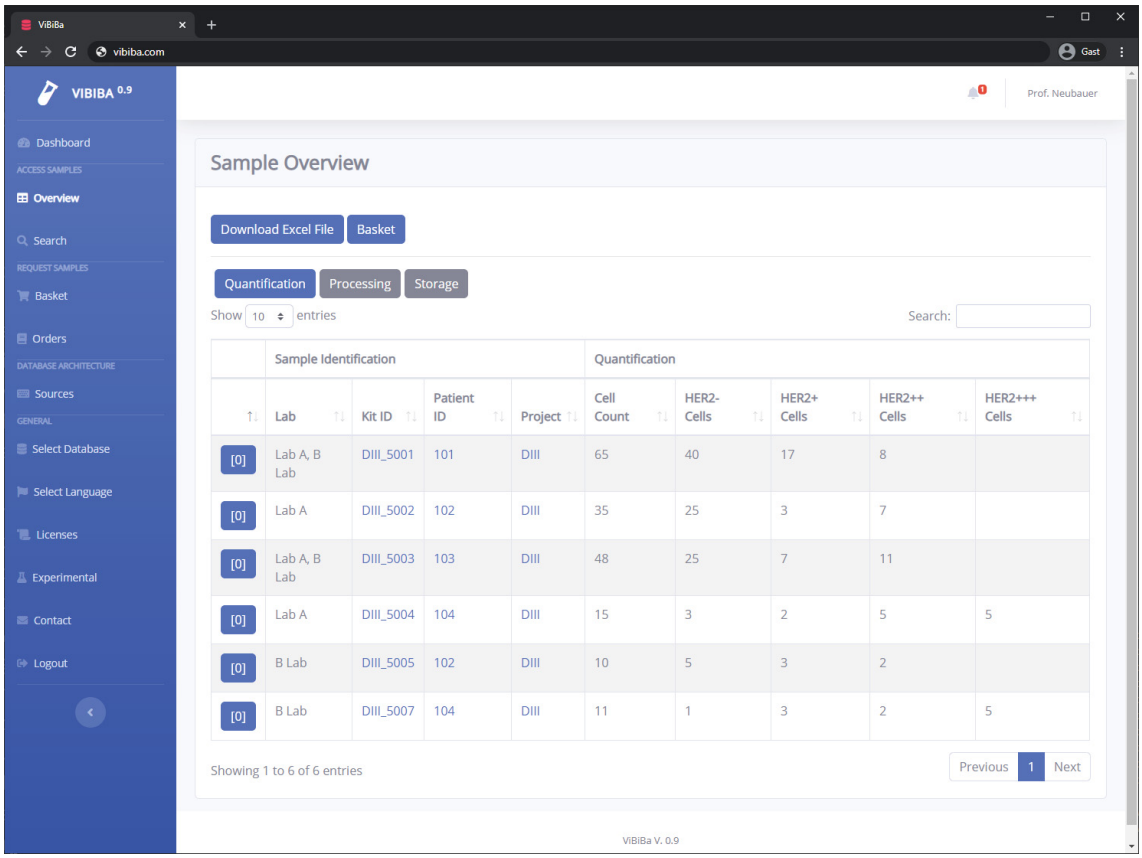


Figure 2: ViBiBa: Screenshot with Sample Data
Taken from Asperger, Cieslik *et al.* [1]

a laboratory is uploaded through the web interface. The uploaded information is called ”input data” and forms a ”source”. Each source can be customized with a plugin script to allow the laboratory to keep the original formatting of its data, thus increasing autonomy of participating laboratories. After processing the ”input data” it is saved into a separate uniform ”source table” with a predefined list of fields (Table 4). Multiple sources then get merged into a single ”summary table”, which now contains multiple entries per sample, each contributed by a specific laboratory. Those entries then get merged again to form a ”condensed summary table”, which only contains a single entry per sample. [1]

2.3.3 Organizational Units

Every user is assigned to an organizational unit (OU). An OU can be a laboratory or any other entity, this naming scheme gives ViBiBa more flexibility when including non-laboratory entities. The permissions are defined based on the OU membership. [1]

2.3.4 Sources and Inputs

ViBiBa only allows user manipulation at the first layer of data, the source layer. The upload permissions are defined on a per OU basis, while all OUs can read the content of every source. File upload is supported as .csv or .xlsx files with the option for excel files

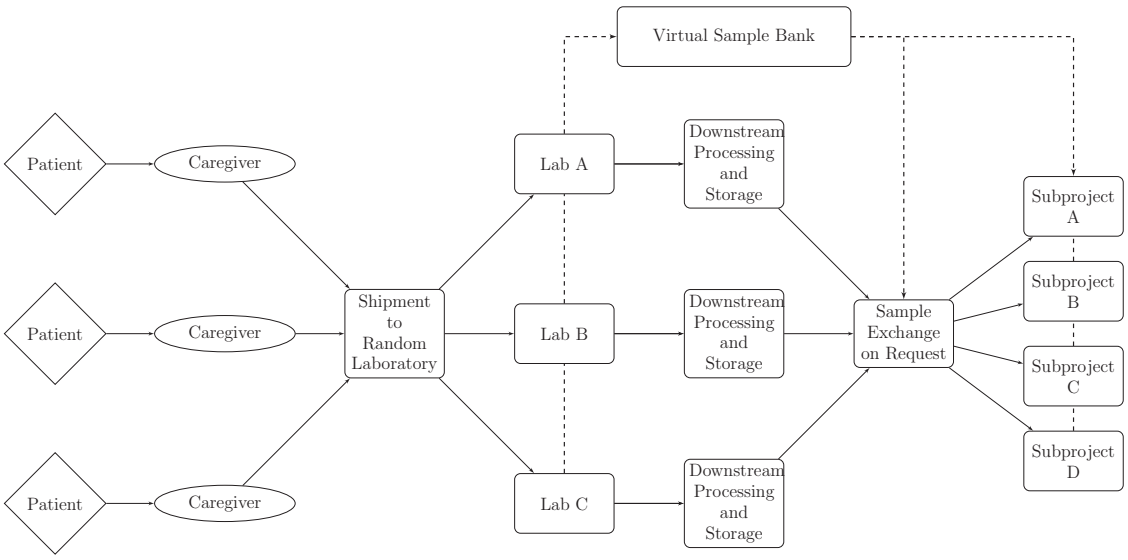


Figure 3: ViBiBa: Logistics in the DETECT trial program

Schematic of the basic (virtual) logistics between the participating DETECT laboratories. Taken from Asperger, Cieslik *et al.* [1]

Table 4: ViBiBa: Exemplary Database Fields

Identification	CellSearch	Kit Shipment	Single Cell Isolation	Iso. Cells (Count)	Storage
Lab	Entry Date	CS: Shipment	Date of Isolation	CD45-/EpCAM+	Serum Aliquots (ml)
Kit ID	Determination Done	CS: Date	CellCelector	CD45-/EpCAM-	Serumbank (µl)
Patient ID	Reason if no Determ.	CS: Destination	DEPArray	CD45+/EpCAM-	Box Position
Study Arm	Determ. Date	CS2: Shipment	FACS	No Cell Control	Comment
[Origin]	Time Till Determ.	CS2: Date	Manual Isolation	Other (single cells)	
	Blood Volume (ml)	CS2: Destination	Isolated Cell Count		
	Cell Count	EDTA: Shipment	Buffer Water		
	HER2-negativ cells	EDTA: Date	Buffer PBS		
	HER2+ Cells +	EDTA: Destination	Deposition Format		
	HER2+ Cells ++	Cartr.: Shipment	Count Deposition		
	HER2+ Cells +++	Cartr.: Date			
	Tumorcell count	Cartr.: Destination			

Excerpt of the database fields currently used in the DETECT production version of ViBiBa. The fields are user defined and have to match a column type from Table 5. CS = Cellsave Preservative Tubes; Cartr. = Cartridge; Iso. = Isolated; Taken from Asperger, Cieslik *et al.* [1]

with multiple worksheets. If a plugin is enabled and configured, it transforms the data before saving it into a ”source table”. Otherwise the data is saved directly into the ”source table”. [1]

2.3.5 Summary Table

The next layer in the data flow is the summary table. Multiple source tables are merged to create a singular summary table. ViBiBa only allows for a single data entry per OU per sample, so potentially conflicting data has to be handled. To achieve this ViBiBa assigns a priority value to each source such that a higher priority source overwrites lower priority sources. Afterwards a dataset with multiple entries per sample (one per OU) is saved in the summary table. [1]

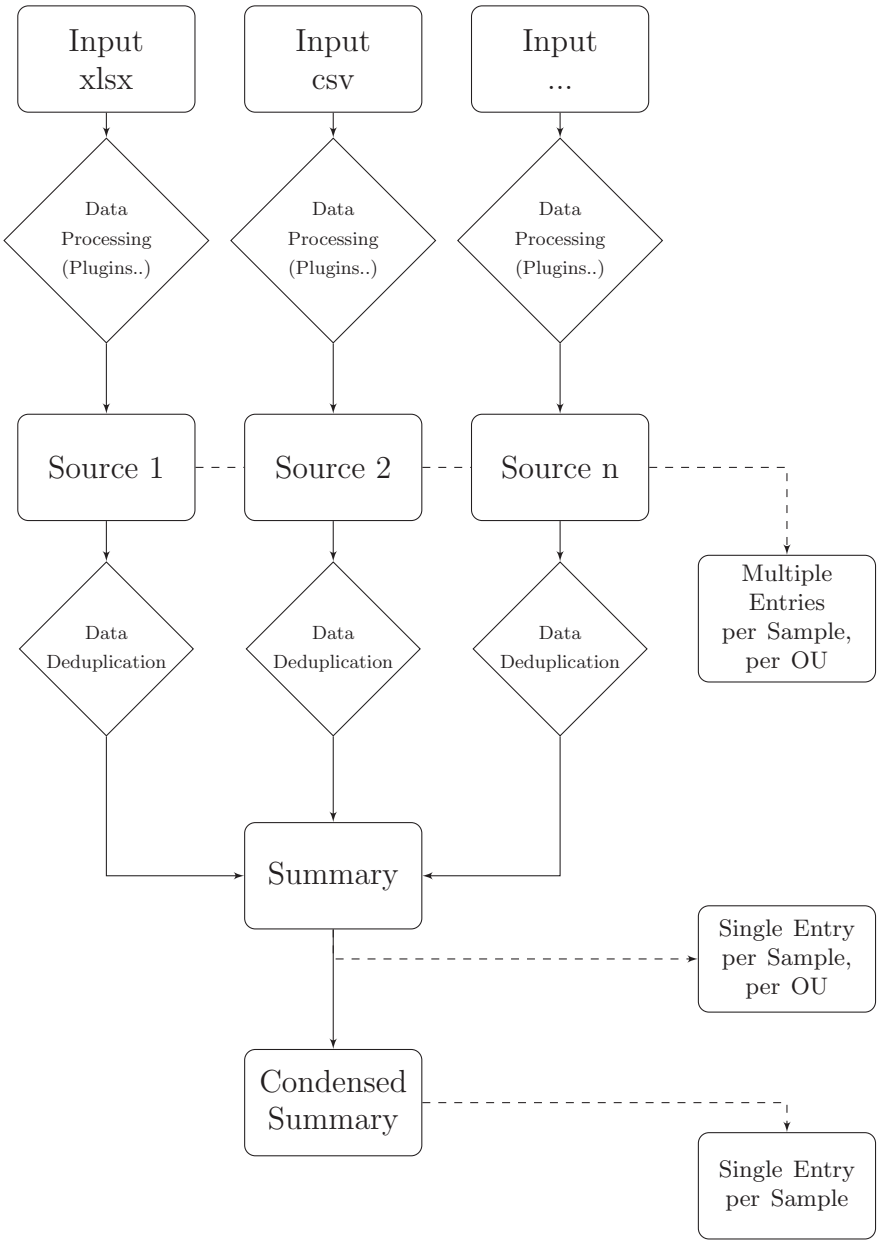


Figure 4: ViBiBa: Architecture

Schematic of the internal data structure and processing. Taken from Asperger, Cieslik *et al.* [1]

2.3.6 Condensed Summary Table

Lastly, the highest layer in the data flow is the condensed summary table. This table is mainly used when displaying data to the user. It is generated by merging all entries for one sample from the summary table. This data transformation differs from the rest, because data has to be consolidated into a single entry. To achieve this, ViBiBa follows certain rules depending on the field type (Table 5). [1]

Table 5: ViBiBa: Column Data Types

Type	Behaviour on Condensation	Example			
		OU 1	OU 2	OU 3	Condensed
Numeric	Addition	2	5	1	8
Numeric [multi-dimensional]	Addition of the matching dimensions, afterwards concatenated to string	2/7/1	0/0/3	1/1/5	3/8/9
Boolean	”TRUE” if at least one entry states ”TRUE”	FALSE	TRUE	FALSE	TRUE
String/Text	Concatenation of strings [except identical entries]	”Text A”	”Text B”	”Text A”	”Text A, Text B”

Every column is assigned to a special type. Depending on the type the fields are treated differently while being processed. Taken from Asperger, Cieslik *et al.* [1]

2.4 Patients and Cells

2.4.1 ENDOpanel

The ENDOpanel was tested with spiked-in T47D cells, which were initially purchased from ATCC (USA). As described previously by Asperger *et al.* [56]: ”Cells were maintained in RPMI 1640 medium supplemented with 10% (v/v) fetal bovine serum, 100 units/mL penicillin/streptomycin and 0.025 mol/L HEPES in a humidified incubator at 37℃ with 5% CO_2. Cells (passage number ≤ 25) were authenticated regularly by Microsynth AG (Balgach, Switzerland) using STRS analysis. The last authentication was performed on May 22, 2018.” Further, the subsequent processing is identical to the CTC detection and isolation pipeline of the ESR1 sequencing project.

2.4.2 Estrogen Receptor Alpha Sequencing

Twenty-five patients with MBC, who had a PT of luminal subtype, were analyzed. The cohort (HER2/neu-negative MBC; CTC-positive) was selected from the DETECT III (NCT 01619111) and DETECT IV (NCT 02035813) trials [57]. ”All patients gave their informed consent for the use of their blood samples for CTC analysis and for translational research projects. Patients’ characteristics were anonymized by using sample identifiers (ethical approval MC-531 and MC-LKP-668).” [40]

A summary of the clinical patient data is shown in Table 6. Only patients without ESR1-hotspot mutations were included. Prior sequencing of CTCs was performed in a partnering laboratory.

2.5 Literature and Database Review

All mutations found in the ESR1 sequencing project were systematically reviewed. Pubmed and Google Scholar were used to find literature while COSMIC, ClinVar and dbSNP were utilized for database review. Search terms included mutation with three-letter codes, mutation with one-letter code, position of the amino acid plus the terms ”ESR1”, ”ERα” or ”Estrogen Receptor Alpha”.

Table 6: Clinical Patient Data (n = 25)

Tumor Size (TNM)		Tumor staging	
T1	3	I	1
T2	11	II	3
T3	7	III	6
T4	2	IV	6
TX	2	NA	9
Nodal Status (TNM)		Age	
N0	8	Mean	61.93
N+	15	Minimum	42
NX	2	Maximum	89
Metastasis Status (TNM)			
M0	15		
M1	6		
MX	4		

The eighth edition of UICC TNM classification was used.

2.6 CTC Detection and Processing

2.6.1 Enrichment and Enumeration of CTCs

Patient samples were collected with CellSave Preservative Tubes and initially processed in one of the participating DETECT laboratories within 96 hours. The CellSearch System was used for enrichment and enumeration of CTCs. Utilizing the CellTracks AutoPrep System with the CellSearch Circulating Tumor Cell Kit, a ferrofluid-based capture was performed. This step is based on a ferrofluid containing antibodies against the epithelial cell adhesion molecule (EpCAM), which is a transmembrane protein typically found on epithelial cells. Next, automatic cell labeling was performed with three immunofluorescent stainings: intracellular cytokeratins (CKs), 4',6-diamidino-2-phenylindole (DAPI) as well as CD45. CK is used to identify CTCs, while DAPI allows to check the nuclear integrity and CD45 enables the exclusion of leukocytes [23]. Finally, the enriched samples were transferred onto glass slides.

2.6.2 Isolation of CTCs and Other Single Cells

For the isolation of single cells the CellCelector was utilized, which is a combination of an inverted fluorescence microscope and an automated microfluidic cell picking robot [58]. The robotic arm transfers single cells via single-use high-precision glass capillaries into separate tubes. Live imaging allows for immediate transfer control. To prevent contamination the ALS Incubator FlowBox was used, which provides a stable and clean environment around the CellCelector.

2.6.3 Whole Genome Amplification

Before downstream analysis can be performed, the genomic DNA needs to be amplified. To achieve this the Ampli1 WGA kit was used according to the manufacturer's protocol. The manufacturer rates the kit suitable for a "balanced and complete amplification of the total DNA content of a single cell" [59] and deems it suitable for SNP and CNA analysis. Quality control was performed with the Ampli1 QC Kit, which consists of a PCR and the subsequent analysis of bands through gel electrophoresis. The PCR is a multiplex PCR with four markers, one of them is a primer for the KRAS fragment and the other three are primers for three long MSE1 fragments [60]. This allows to calculate the proprietary genome integrity index (GII) [60]. The GII is calculated by essentially counting the visible bands on the gel. When no band is visible, the GII is denoted as zero. If only the KRAS fragment is visible, a GII of one is achieved. A GII of two requires that one of the three long MSE1 fragments is visible, while a GII of three requires that two of the three long MSE1 fragments are detected [60]. Finally, a GII of four indicates that all three long MSE1 fragments are visible [60]. Cells with a GII of less than two were discarded.

2.6.4 Spike-In Experiment

The single cell workflow was validated with T47D cells (a breast cancer cell line) spiked into donor blood. Afterwards, the blood was processed by the CellSearch for CTC enrichment and labeling. Next, the cells were isolated with the CellCelector and the Ampli1 WGA and QC Kit used for whole genome amplification. Further, an ESR1 multiplex PCR was performed to validate the existence of the spiked-in cells in the healthy blood. Finally, the workflow was repeated multiple times with new T47D cells to perform the ENDOpanel.

2.6.5 ESR1 Sequencing

ESR1 DNA fragments were amplified and barcoded utilizing a multiplex PCR and sequenced using the Illumina MiSeq System as described previously [40]. In essence the WGA product underwent a PCR with ESR1 specific primers. Afterwards, the PCR product was processed in a second PCR, which applied sample-specific barcodes through barcoding primers onto the DNA to allow the MiSeq to distinguish the samples. In a final step, the samples were pooled and loaded onto the MiSeq sequencing platform.

2.6.6 NGS Library Preparation

The Agilent SureSelectQXT kit was utilized according to the manufacture's protocol with a custom library containing all exons from 12 genes (in alphabetical order: AKT1, AKT2, EIF4EBP1, ERBB2, ESR1, INPP4B, MTOR, PDL1, PDL2, PIK3CA, PIK3CB, RPS6KB1) and the complete range of PTEN. During the workflow, the samples were purified multiple times utilizing AMPure XP beads. For the experiment described in section 3.2.6 the recommended AMPure XP beads to DNA ratio was increased from 1:1 to 4:1.

2.6.7 Fragment Detection

For quality control, the fragment distribution during library preparation was detected with the Agilent 2100 Bioanalyzer. The distribution was detected twice: Once after MSE1 fragmentation and tagmentation using the Bioanalyzer DNA 1000 analysis kit. Once more after hybridization, capture and indexing using the Agilent Bioanalyzer DNA High Sensitivity kit.

2.7 In Silico Processing

Most tools described in this section were utilized using their respective version from Bioconda or the galaxy platform [61, 62].

2.7.1 Mapping and Variant Calling

For the following analysis, the human reference genome 38 (hg38) was used for genetic mapping. Reads were aligned to hg38 with BWA-MEM (0.7.17.1) [53]. Mutations in the aligned sequences were called using VarScan 2 [63] and annotated using SNPeff (Galaxy Version 4.3) [62, 64]. The reference protein sequence used for ERα (unless stated otherwise) is NP_000116.2. Sequence data was obtained from NCBI [65]. Statistical analysis was performed using R [66].

2.7.2 Coverage Calculations With the Custom CoverageReporter

Coverage was calculated by generating a pileup file with SAMtools (1.13) [67] and subsequent calculations with custom R scripts (CoverageReporter). The CoverageReporter received genomic start and end positions of the target (e.g., exon or restriction enzyme fragment) together with the pileup file as input. By iterating through each target area, the CoverageReporter creates multiple output files. One file contains the mean coverage per target:

$$\frac{\sum_{i=start}^{end} coverage_i}{end - start + 1}$$

Another output file contains the ratio of positions above certain cutoffs per target:

$$\frac{\sum_{i=start}^{end} g(i)}{end - start + 1}$$

with $g(i)$ as:

$$g(i) = \begin{cases} 0 & coverage_i < cutoff \\ 1 & coverage_i \geq cutoff \end{cases}$$

2.7.3 Virtual Reconstruction of Restriction Enzyme Fragments

To determine the impact of DNA fragmentation during WGA on the coverage of the ENDOpanel, a virtual reconstruction of the restriction enzyme fragments was required. For this purpose I programmed a custom R tool, which receives genomic ranges of targets as input and outputs the genomic ranges of the fragments. In this case the restriction enzyme is MSE1 and the targets are the exons of the ENDOpanel. The input is formatted as a *.bed* file, which consists of three required columns (chromosome, start and end position) plus the name and strand orientation of the exon. After obtaining the reference genome (hg38) and loading the *.bed* file containing the exon positions, the tool will iterate through every exon and apply the *fragment_coordinates* function.

The *fragment_coordinates* function will extract the reference sequence for the given genomic range and search for the given recognition pattern of the restriction enzyme, creating a vector of points at which the DNA will be cut, this vector can be empty if no recognition pattern is found inside the target area. Since in most cases, the first and last fragment will begin or end outside of the target area, the function then searches the first occurrence of the recognition pattern outside of the target area. These two points are appended to the beginning and end of the points vector. Finally, the *fragment_coordinates* function ends by returning the points vector.

Afterwards, the points vector is converted by the *fragment_dataframe* function into a data frame. Each row in the new data frame contains information about one fragment with chromosome, start/end position and name of the fragment. The name of the fragment is generated automatically from gene and exon name followed by an ascending number starting at zero. After each iteration, the fragment data frames are merged together. When the loop is completed, the fragment lengths are calculated and a fragment ID (based on fragment start and end position) is generated.

Next are two optional phases: deduplication and overhang cutting. Deduplication is performed by removing all entries with the same fragment ID. A duplicate entry can occur due to one fragment spanning more than one exon. Overhang cutting will trim the fragment boundaries to align them with the genomic ranges of the exons. This step is necessary as the coverage of a fragment can only be accessed inside the region that is covered by the ENDOpanel. Finally, the resulting data frame is saved as a *.csv* file.

2.7.4 ESR1 Hotspots and Potential Damage Map

For visualization of the distribution of previously described ESR1 mutations various studies [68–75] were aggregated using cBioPortal [76, 77]. Next, a lollipop diagram was created, illustrating the number of mutations per amino acid together with an overlay of functional domains and other characteristics of ESR1 [76, 77]. To create the potential impact heatmap (as seen in Fig. 17), every position from the FASTA reference sequence was extracted. A file containing every possible amino acid exchange was then created using a custom R Script. The variant list was then processed in batch by PolyPhen2 [78].

Table 7: Tunnel Profile Overview

ID	max. bottleneck radius [Å]	avg. length [Å]	avg. curvature	avg. throughput
1	1.07	23.83	1.43	0.54
2	1.03	12.37	1.21	0.53
3	1.01	22.10	1.71	0.44
4	0.92	13.56	1.23	0.42
5	0.92	17.52	1.36	0.40

Tunnels analyzed in ERα. Tunnels with an average throughput of less than 0.4 were discarded.

The amino acid exchanges and their predicted effect were plotted with ggplot2 [79].

2.7.5 Structure Files

The reference 3D models were obtained from the Protein Data Bank (PDB) [80] under the accession number 1QKU [81], 2R6Y [82] and 3ERT [38]. Mutant structures were created with ChimeraX [83] using the Dunbrack rotamere library [84]. Structure files for estradiol (DB00783), tamoxifen (DB00675), 4-hydroxytamoxifen (DB04468) and raloxifene (DB00481) were obtained from drugbank.ca [85].

2.7.6 Ligand Tunnel Analysis

Tunnel analysis was performed via CAVER Analyst 2.0 [86] and CAVER Web 1.0 [47]. Analysis of the ligand transport was performed with CaverDock [87]. Bound ligands were removed in the 3D models to enable unobstructed path finding into the LBD. Ligand tunnels calculated for the wild type (1QKU) are shown in Table 7, subsequent analysis was performed on tunnel 1. Calculations on mutated structures of 1QKU were done on the same tunnel based on location, since the tunnel numeration is based on the estimated throughput the nomenclature could change when a mutation affects the selected tunnel. A longitudinal section through the tunnel was obtained and the profile of the tunnel in the wild type and R394S mutant structure overlaid. Transport analysis was performed with estradiol, 4-hydroxytamoxifen and raloxifene. Ligand binding forces were calculated from the outside to the inside (LBD) of the tunnel, a distance of $0\mathring{A}$ denotes the outside edge of the tunnel. CAVER Web 1.0 can estimate two energy profiles by calculating upper- and lower-bound trajectories [47]. The lower-bound trajectory is generated by converting the tunnel in a set of consequent discs and calculating the ligand binding forces at each of these discs [47, 87]. However, this approach can lead to a ”flip” of the ligand, as the discs may skip over bottlenecks [47, 87]. On the other hand, upper-bound trajectories are generated as a contiguous trajectory where each calculation depends on the results of the previous disc [47, 87]. Since initial calculations of upper- and lower-bound forces showed no major discrepancies in the WT, subsequent calculations only included lower-bound values, as computational cost increases exponentially with upper-bound trajectories while a ”flip” of the ligand is unlikely in this case.

2.7.7 Receptor-Ligand Forces, Bonds and Predicted Structural Changes

For calculation of the receptor-ligand forces, BIOVIA Discovery Studio was used. The structure files were obtained from PDB as described above. The change in polarity is based on the hydropathy index [88].

3 Results

3.1 ViBiBa

3.1.1 Declaration on Own Contribution

ViBiBa was developed in collaboration with Hannah Asperger. We share the first authorship in the peer-reviewed publication about ViBiBa [1]. The workload for software development, project planning and manuscript writing was equally distributed between us.
The source code can be accessed online: `https://github.com/asperciesl/vibiba`

3.1.2 Definition of Requirements

The work on ViBiBa began after a literature and web review as the DETECT trial group was searching for a viable solution for its decentralized sample management. A catalog of requirements was created for an ideal software. First of all, the software should be able to merge sample databases together. Additionally, an overview of all available material per patient or per sample (when it is shipped between laboratories) should be generated. The ideal software should be able to perform the data transformation gracefully, e.g., allowing different upload formats and adjustments for laboratory-specific variants in notation. One real-life example is the notation of the sample identifier, e.g., the DETECT III kit with the ID 5000 can be written as DIII-5000, DETECT3-5000 or any other combination with different characters between the trial arm and ID (e.g., a space instead of a dash). This would require some kind of plugin support that allows to customize the upload process for each input file. In terms of security, there should be a right management system to control the data editing permissions. From a data protection perspective, the application should be self-hosted and, if possible, published as open source to allow for code inspection and adjustments. [1]

3.1.3 Design of ViBiBa's Functionality

ViBiBa is split into multiple modules, each built to process different user requests (Fig. 5). The user is mainly interested in the overview of samples and the detail view of individual specimens, which are read from the summary tables. To get to this point ViBiBa must first generate the data. Through the data insertion module, the user can upload new data, which is later processed in the core module of ViBiBa. One of the main goals during development was the seamless experience of browsing through ViBiBa. The user should not notice that different tables are required to process a request and always receive a readable output. [1]

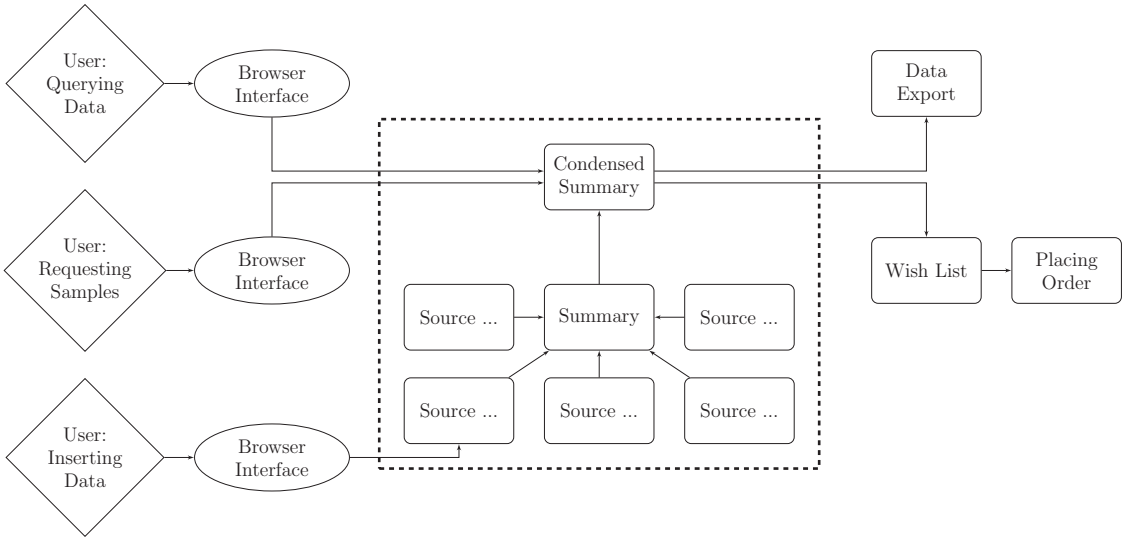


Figure 5: ViBiBa: Workflow of Selected Processes
The data flow when retrieving and manipulating information in ViBiBa.
Taken from Asperger, Cieslik *et al.* [1]

3.1.4 Exploring Data

After logging in, the user is greeted with the main overview of ViBiBa. Mainly, the ”condensed summary table” is fetched and output to the user. This data shows one entry per sample. Each sample can be expanded into a detailed view, which is then fetched from the ”summary table” and displays one entry per OU that has information on the selected sample. Allowing to identify where parts of the sample are stored and trace where downstream analysis has already been performed. The displayed data can be downloaded, e.g., as an excel file. However, ViBiBa does not store the detailed results of downstream analysis (e.g., sequencing data) to maintain the autonomy of the participating laboratories. [1]

3.1.5 Searching Samples

A more advanced query can be performed using the ViBiBa basic or advanced search. The search form allows to select the fields of interest and create a custom filter with mathematical operators ($<, >, =$). When more than one filter is created, they are automatically linked via an AND operation. After the search query is submitted, ViBiBa retrieves all matching results from the ”condensed summary table”. The results can be downloaded similarly to the process already mentioned. Further, the user can also put all or some samples from the search results onto the wish list, a function described below. [1]

3.1.6 Requesting Physical Samples

Multicenter trials with decentralized sample storage make translational experiments harder, as samples need to be exchanged between laboratories to study a cohort of interest.

ViBiBa has a sample ordering module to facilitate easier sample exchange. Users can put a specimen on a ”wish list” at the click of a button, which is similar to the shopping cart on modern internet stores. After selecting all desired samples, the user can go to the ”check out” process where a priority and comment may be added. Lastly, the request can be submitted. This request can be approved automatically or a board of people can be notified to decide on the request. The option to regulate the flow of samples can be important to protect valuable samples of specific subgroups of patients. After approval of the request, the participating laboratories that store parts of the requested specimen are notified by email and they can see a list of requested samples together with the comment of the requesting user and a shipping address. [1]

3.1.7 Automatic Flagging

As mentioned previously, ViBiBa tries to improve the creation of viable cohorts for translational research. ViBiBa supports custom filters that automatically assign a sample to a specific cohort. The filtered samples will be marked by a flag icon and a description of the cohort. As during most single cell analyses the cell is destroyed in the process, this will help to preserve valuable specimens. [1]

3.1.8 Harmonization

ViBiBa allows for a greater comparability between laboratories by harmonization of data and enforcing standardized notation. This is more than just a side effect of the database, as the active processing of various data sources paves the way to a common standardized dataset in the trial group. Most traditional values like clinical blood parameters or date/time values are already highly standardized. In contrast, the notation and storage of WGA QC data is not standardized and every laboratory utilizes a different data storage approach. As mentioned before, ViBiBa deploys custom plugin scripts for its data sources to create a standardized dataset.

3.1.9 Administration

Right management is one of the key features of ViBiBa. Write permissions are assigned on a per-OU basis to restrict data modification to the dataset contributed by the OU. To increase transparency, all data sources can be viewed by every user in ”read only” mode. While a dedicated administrative interface is currently under development, the default way of managing ViBiBa is via a direct interaction with the MySQL database. The MySQL tables are commented and changes on configuration tables like the ”field list” (specifying all available columns and their properties) are automatically propagated to the auto-generated tables like the summary tables. ViBiBa allows the deployment of multiple separated databases, which makes it possible to share server resources and only requires one set of login credentials per user for an unlimited number of databases/trials.

The default docker configuration comes with a bundled phpMyAdmin instance to allow the administrators an easy initial overview over the internal MySQL structure. For production deployment the phpMyAdmin instance is disabled for security reasons, to prevent access to the database with the default credentials through a browser.

3.1.10 Data Protection

At the moment ViBiBa is designed to only store information about biological specimens with a patient and sample ID without attached clinical data about the patient. This design choice is intentional as data security for potential patient identifying data is much stricter. To comply with general data protection regulations, ViBiBa eliminates all patient data except for the pseudonymized ID. In future updates inclusion of patient data is planned, but an assessment of the required data security has to be performed first. As mentioned in the methods section, ViBiBa's default deployment method is via docker containers, which allows a separation of the application services and the host environment. As ViBiBa is an open source software, the deployment methods are not restricted and together with the SSO capabilities they enable a modern security approach. The default installation uses only the default HTTP protocol without SSL security, as it is expected that the end user deploys own certificates with a reverse proxy like NGINX or other solutions which then encrypt the web traffic.

3.2 ENDOpanel

3.2.1 Overall Coverage

To validate if the ENDOpanel covers the expected genomic ranges, the coverage per position had to be calculated. The required depth for the subsequent analysis was set to ≥ 10 reads. Other authors describe a reliable SNP detection above a coverage of four to six reads [89, 90], but as this work focuses on single cell analysis, a higher depth was chosen to be less susceptible to artifacts. As seen in Fig. 6, the ENDOpanel covers 74% of positions with a depth of ≥ 10 reads. Further analysis was required to validate if all known cancer hotspots are covered and to generate a hypothesis for the missing coverage.

3.2.2 Coverage per Exon

To get a better understanding of the regions with a worse coverage, I calculated the mean coverage per exon. 166 of 220 (75%) observed exons displayed a coverage of $\geq 50\%$ (Fig. 7 a), while 12 exons were not covered at all. There is a direct positive relationship between exon length and coverage (Spearman's rank correlation: $\rho = 0.33$, $p < 0.001$). This indicates that longer exons have a higher probability to get good coverage. There is a drop of coverage for exons that are shorter than 60 bp, dropping from 73% above 60 bp to 48% (Fig. 7 b).

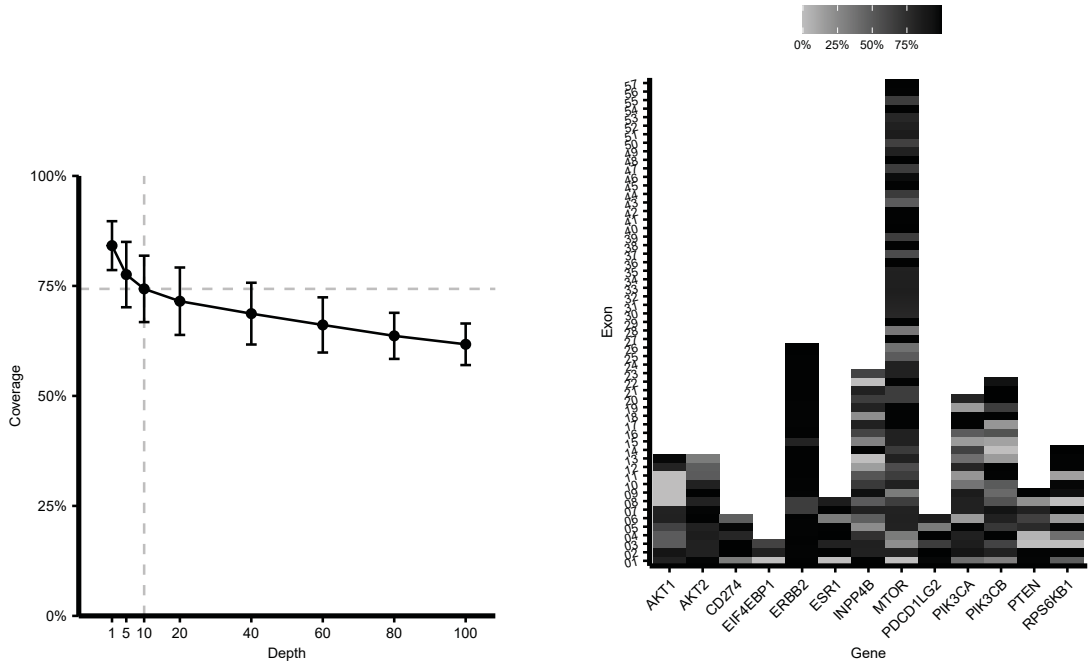


Figure 6: ENDOpanel Overview
a) Coverage report **b)** Gene and exon coverage heatmap

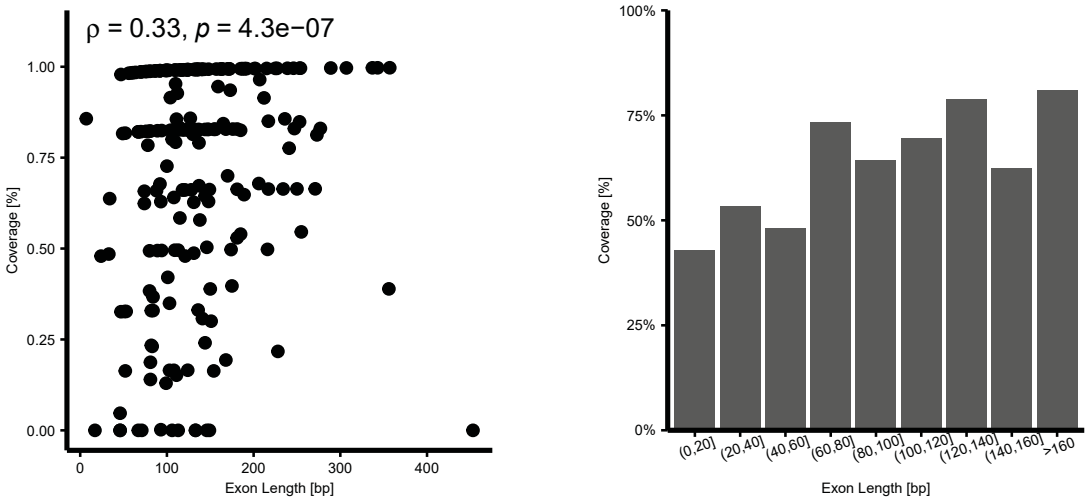


Figure 7: Exon Length Analysis
a) Correlation of exon length and coverage (depth ≥ 10 reads)
b) Coverage (depth ≥ 10 reads) of exon length bins

3.2.3 Coverage of Known Mutational Hotspots

To validate the ENDOpanel for real world usage, I calculated the coverage of already known SNP hotspots [91]. The method counts a hotspot as covered when a depth of at least 10 reads is reached. In total, 256 genomic positions were observed (one SNP hotspot can include up to three genomic positions) (Fig. 8). A total of 88% of known hotspots are covered at least once in the six tested samples, while only 30% of hotspots were covered in all tested samples. The genomic hotspot positions not covered at all were mainly on the first exon of PIK3CA and the first exon of ESR1.

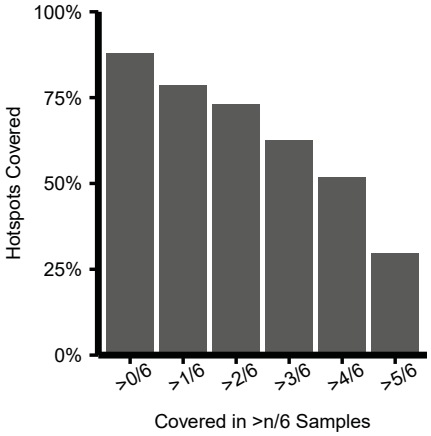


Figure 8: Panel Coverage of Known SNP Hotspots (Depth ≥ 10)

3.2.4 Coverage per MSE1 Fragment

As the DNA of the CTCs was amplified using the Ampli1 WGA Kit, it is enzymatically fragmented by the MSE1 restriction enzyme. MSE1 has the DNA recognition pattern 5'-TTAA-3'. Since the recognition pattern is not distributed equally, fragments of different sizes are created. One hypothesis is that the coverage of the ENDOpanel is associated with the size of the MSE1 fragments. Subsequently, the mean coverage per MSE1 fragment was calculated. 185 of 287 (64%) observed MSE1 fragments displayed a coverage of $\geq \ 50\%$ (Fig. 9 a). 19 MSE1 fragments were not covered at all. Spearman's rank

correlation shows a positive relationship between MSE1 fragment length and coverage ($\rho = 0.37, p < 0.001$). Longer fragments thus have a significantly higher probability to get good coverage. Fragments below 150 bp show a steep drop in coverage from 71% to 45% (Fig. 9 b).

3.2.5 Exon and MSE1 Fragment Length

Both exon length and the length of MSE1 fragments correlate with coverage of the ENDOpanel. Next, a possible confounding of the two variables needs to be checked. The distribution of the exon lengths resembles a Gaussian bell curve (Fig. 10 a). In contrast, the MSE1 fragment lengths accumulate at the lower end with a long trail, resembling an inverse or reciprocal function (Fig. 10 b). Subsequently, a Spearman correlation containing every MSE1 fragment with its length and the length of the associated exon length was calculated. Spearman correlation does not show a significant relationship (Fig. 11 $\rho = 0.024, p = 0.68$).

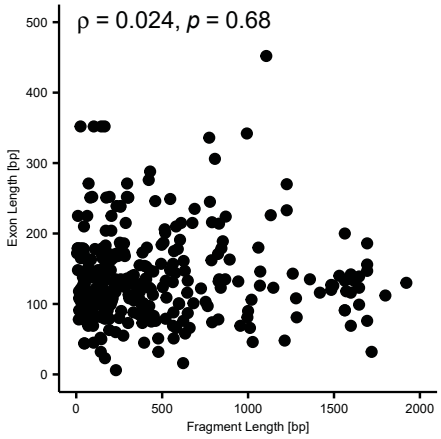
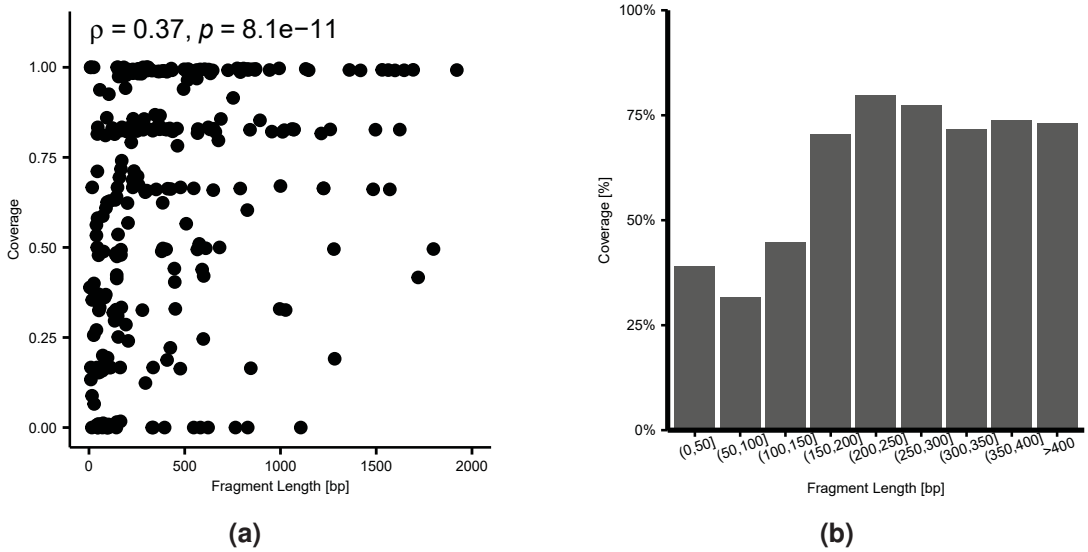


Figure 11: Exon Length Against MSE1 Fragment Length



(a) **(b)**

Figure 9: MSE1 Fragment Analysis
a) Correlation of MSE1 fragment length and the coverage (depth ≥ 10 reads)
b) Coverage (depth ≥ 10 reads) of MSE1 fragment length bins

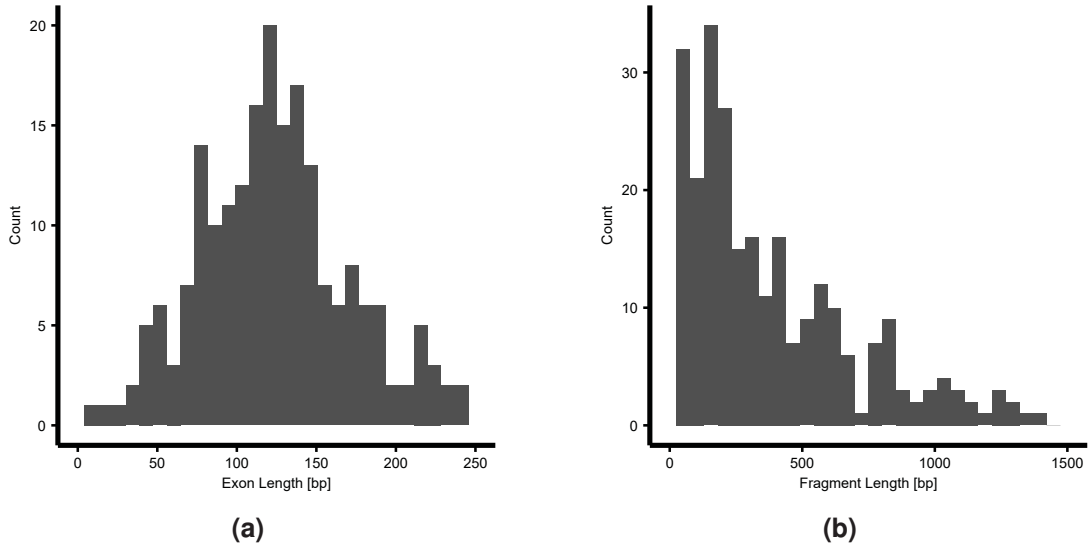


Figure 10: Exon and MSE1 Fragment Histograms
a) Histogram of exon length **b)** Histogram of MSE1 fragment length

3.2.6 Increasing Magnetic Beads Volume During DNA Purification

Another hypothesis is that the coverage of short fragments depends on the amount of AMPure XP beads used during the purification steps. At first, I purified a DNA ladder with different ratios of AMPure XP beads. Afterwards, the purified sample was analyzed with the Agilent Bioanalyzer using the DNA 1000 analysis kit as described in the methods section. As the beads to DNA ratio increased, shorter fragments were captured more often (Fig. 12). To further test the hypothesis, the manufacturer's protocol was altered to compare the original 1:1 beads to DNA ratio with a 4:1 ratio. The following compar-

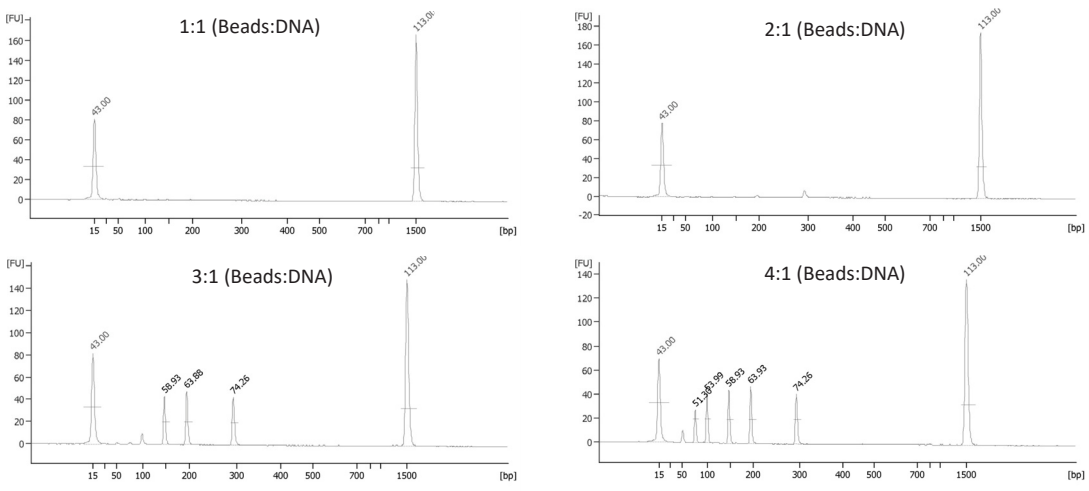
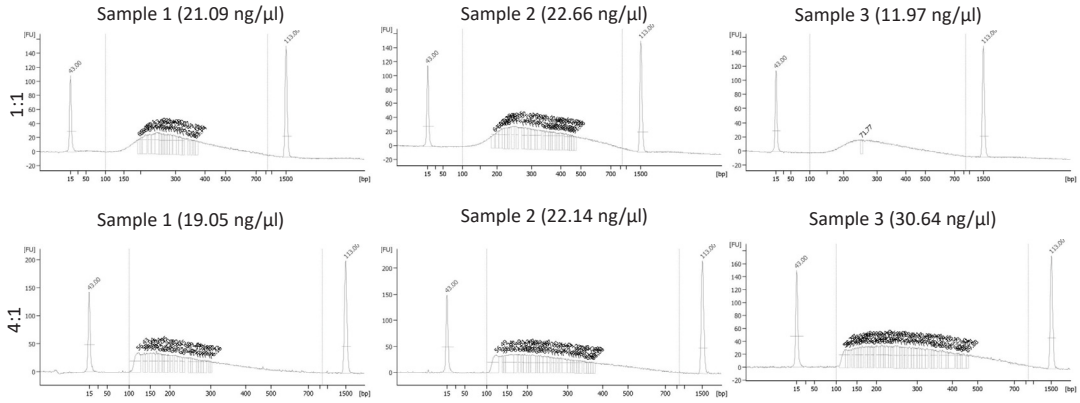
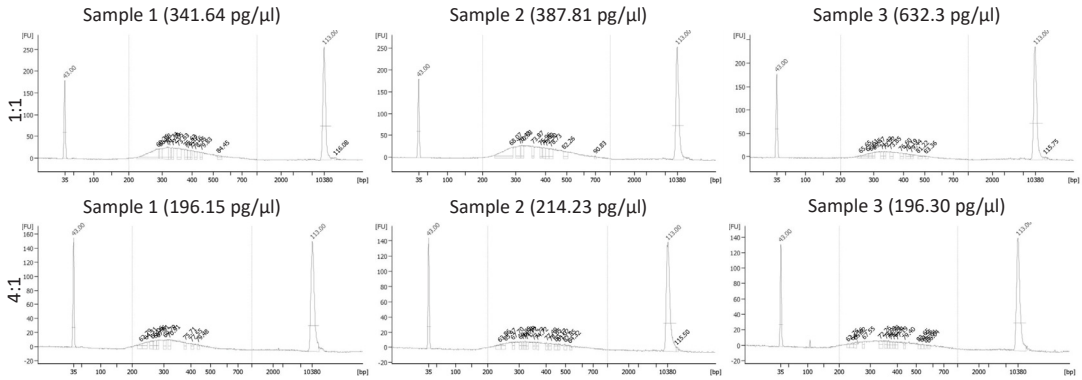


Figure 12: Fragment Recovery with Different AMPure XP Beads to DNA Ratios

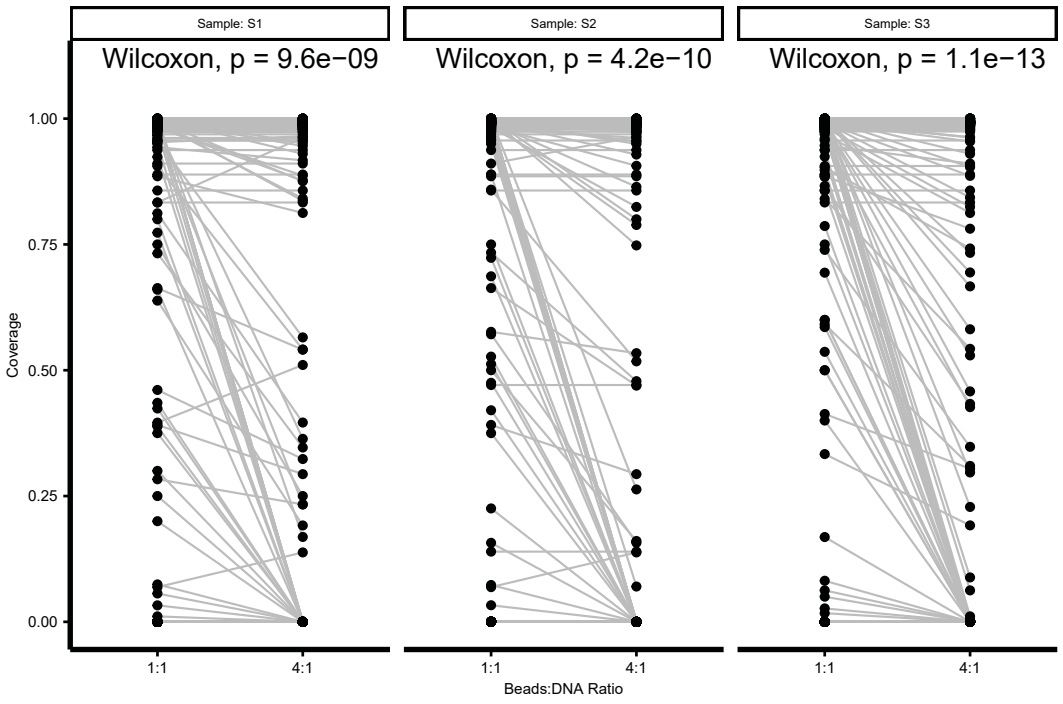
ison was performed using three samples. During the two quality control steps (on the Bioanalyzer platform) shorter fragments were detected more often in the batch with the 4:1 ratio (Fig. 13 a,b). In contrast, the total DNA amount was not significantly altered after tagmentation (Fig. 13 a). Further, the DNA amount was significantly decreased after the hybridization, capture and indexing steps (Fig. 13 b; $p < 0.05$). After sequencing, the coverage per MSE1 fragment was significantly lower than in the matched samples (Fig. 13 c). While the first synthetic comparisons demonstrated a theoretical benefit of a higher beads to DNA ratio (Fig. 12), this could not be reproduced in the real workflow (Fig. 13 c).



(a)



(b)



(c)

Figure 13: ENDOpanel Performance with Different AMPure XP Beads to DNA Ratios

a) Fragment size distribution after tagmentation **b)** Fragment size distribution after capture **c)** Correlation of MSE1 fragment length and coverage (depth ≥ 20 reads)

3.3 Estrogen Receptor Alpha Sequencing

3.3.1 Patient Characteristics

In this study the ESR1 mutation status of CTCs (enriched via CellSearch and isolated with the CellCelector) obtained from the blood of 25 MBC patients was analyzed. PTs of all patients enrolled in this study were ER positive. All of them received endocrine therapy: 76% of the patients were treated with an AI or SERM, 24% of the patients received GnRH analogs. CellSearch analysis of blood samples revealed CTC counts between 2 and approximately 517 CTCs per 7.5 ml of blood. The median CTC count per 7.5 ml of blood was 43. The characteristics of the patient cohort can be found as described in the methods sections (Table 6).

3.3.2 CTC Count Correlates With ESR1 Mutational Burden

As it is already known that the CTC count correlates with the clinical outcome, it is speculated that an increased CTC count would also increase the ratio of ESR1-mutant CTCs, as this could potentially lead to endocrine resistance. Since not all CTCs collected from the cohort were sequenced (as some patients exhibited more than 500 CTCs), the ratio of ESR1-mutant CTCs that were sequenced per patient was determined (Fig. 14). Subsequently, I correlated the CTC count from the CellSearch with the ESR1 mutational burden on the analyzed CTCs. A positive correlation (Spearman's rank correlation: $\rho = 0.52, p < 0.01$) between measured CTC count and ratio of ESR1-mutant CTCs could be shown (Fig. 14).

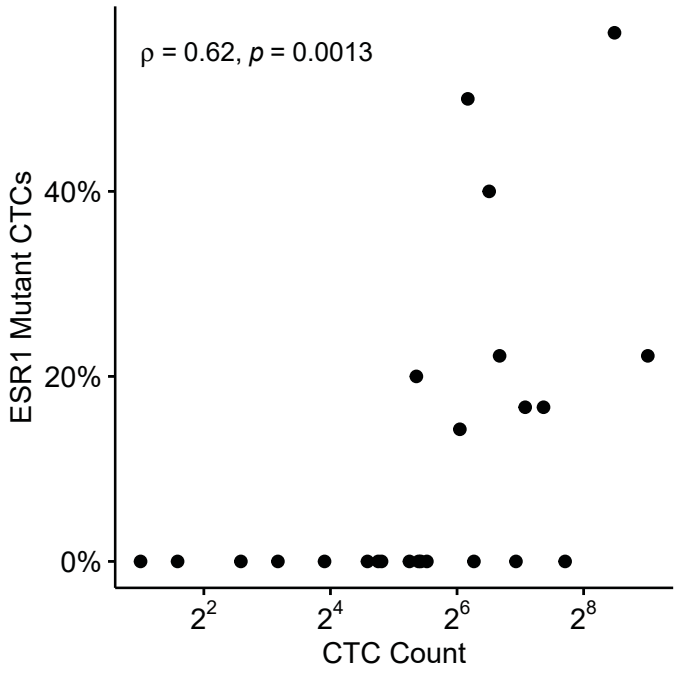


Figure 14: CTC Count against CTC ESR1 Mutational Burden

Table 8: ESR1 Mutations, Literature Review & Amino Acid Characteristics

		Wild Type		Mutant	
		Polarity	Hydropathy I.	Polarity	Hydropathy I.
M176I		Nonpolar	1.9	Nonpolar	4.5
Y195H		Polar	-1.3	Basic polar	-3.2
C205R	[92]	Nonpolar	2.5	Basic polar	-4.5
P222Q	[93]	Nonpolar	-1.6	Polar	-3.5
L242I	[94]	Nonpolar	3.8	Nonpolar	4.5
M250I	[95]	Nonpolar	1.9	Nonpolar	4.5
S294R	[96]	Polar	-0.8	Basic polar	-4.5
A307D	[97]	Nonpolar	1.8	Acidic polar	-3.5
H356Y	[98]	Basic polar	-3.2	Polar	-1.3
Q375K		Polar	-3.5	Basic polar	-3.9
W383R	[95]	Nonpolar	-0.9	Basic polar	-4.5
R394S	[99–103]	Basic polar	-4.5	Polar	-0.8
K492R		Basic polar	-3.9	Basic polar	-4.5
M528T	[95]	Nonpolar	1.9	Polar	-0.7
Q565R	[95]	Polar	-3.5	Basic polar	-4.5
L568F		Nonpolar	3.8	Nonpolar	2.8
A593S	[104]	Nonpolar	1.8	Polar	-0.8

Detected ESR1 mutations present on at least one CTC, but not in germline.
Hydropathy Index: A positive value indicates a hydrophobic amino acid [88]

3.3.3 (Novel) ESR1 Mutations on CTCs

The isolated CTCs' genomic DNA was amplified via Ampli1 WGA and sequenced with a previously validated approach by Franken *et al.* [40]. I successfully identified 17 different somatic ESR1 SNPs (not present in germline) in 10 patients (Fig. 15 & Table 8). Twenty-one CTCs harbored ESR1 mutations; One SNP (H356Y) could be detected in two patients in a total of five CTCs. Two mutations (R394S and W383R) affected "critical amino acids" involved in ligand binding [105]. Additionally, R394 is involved in a hydrogen bond between ERα and its ligand estradiol. To determine the functional relevance of the point mutations I performed a literature and database review (e.g., through COSMIC [106]). I included direct matches (identical amino acid exchange) and indirect matches (same position in the amino acid sequence but different or no amino acid exchange). Additionally, I included matches from non-breast cancer patients in the review. In four cases I found a direct match (M250I, S294R, W383R, R394S); in eight cases only indirect matches were found (C205R, P222Q, L242I, A307D, H356Y, M528T, Q565R, A593S). In a total of five mutations, I could not find any previous mention of the point mutation (M176I, Y195H, Q375K, K492R, L568F; last review July 2021). As seen in Fig. 16, ESR1 does not seem to have a broader clustered region of higher mutational burden. Instead, a few already known oncogenic mutations create singular peaks while the rest of the SNPs are evenly scattered throughout the gene. To better visualize potential areas of interest, I created an ERα mutation impact map and highlighted mutations detected in the study (Fig. 17).

The mutations detected in the cohort seem to follow the clusters of ”probably damaging” regions in ERα. Partly, those regions coincide with the zinc finger domain and the hormone-binding region.

3.3.4 R394S May Alter ERα LBD Cavity

After initial evaluation of the found mutations, I selected some for further in silico analysis to demonstrate a possible impact on endocrine resistance. In this step, I focused on the amino acid exchange R394S (arginine to serine). This SNP was found in a CTC from a patient who was previously treated with tamoxifen and subsequently with an AI. To better illustrate the possible effects of the amino acid exchange, I created a hydrogen bond analysis of ERα in complex with estradiol. The residue R394 creates a hydrogen bond with the estrogen ligand (as previously described) (Fig. 18 a). This hydrogen bond does not exist in the R394S mutant (not shown). Further, I theorized that the residue may be involved in the ligand transport into the buried LBD of ERα. To examine this hypothesis, I calculated possible ligand tunnels (Table 7). Subsequent tunnel analysis was performed on the tunnel with the highest throughput. Next, I created a plot of the diameter of the tunnel to detect possible bottlenecks that may change with residue mutation (Fig. 18 b). In comparison to the wild type, the R394S mutant displays a wider tunnel with a bottleneck radius of $2\ \mathring{A}$ instead of $1\ \mathring{A}$ (Fig. 18 b). To better visualize the difference of the SNP on the profile of the tunnel, I created a virtual cross section through the tunnel at the position of residue R394 (Fig. 18 c). In this figure one can observe the impact of the amino acid exchange on position 394 as the tunnel widens (Fig. 18 c). Finally, I computed a ligand transport analysis for estradiol, 4-hydroxytamoxifen (the active metabolite of tamoxifen) and raloxifene with the wild type and R394S-mutant ERα structure (Fig. 18 d). The serine residue reduces the calculated energy that is required for estradiol to traverse the tunnel from 44.2 kcal/mol to 3.4 kcal/mol. Similarly, the peak energy for the transport of tamoxifen is reduced from 48.2 kcal/mol to 15.1 kcal/mol. This means that in comparison, the required energy for estradiol is reduced by a factor of 13 while the energy barrier for tamoxifen is only reduced by a factor of 3.2. Additionally, SERMs are known to traverse through a different tunnel, which allows them to exert their inhibitory effect on helix 12 of ERα [38].

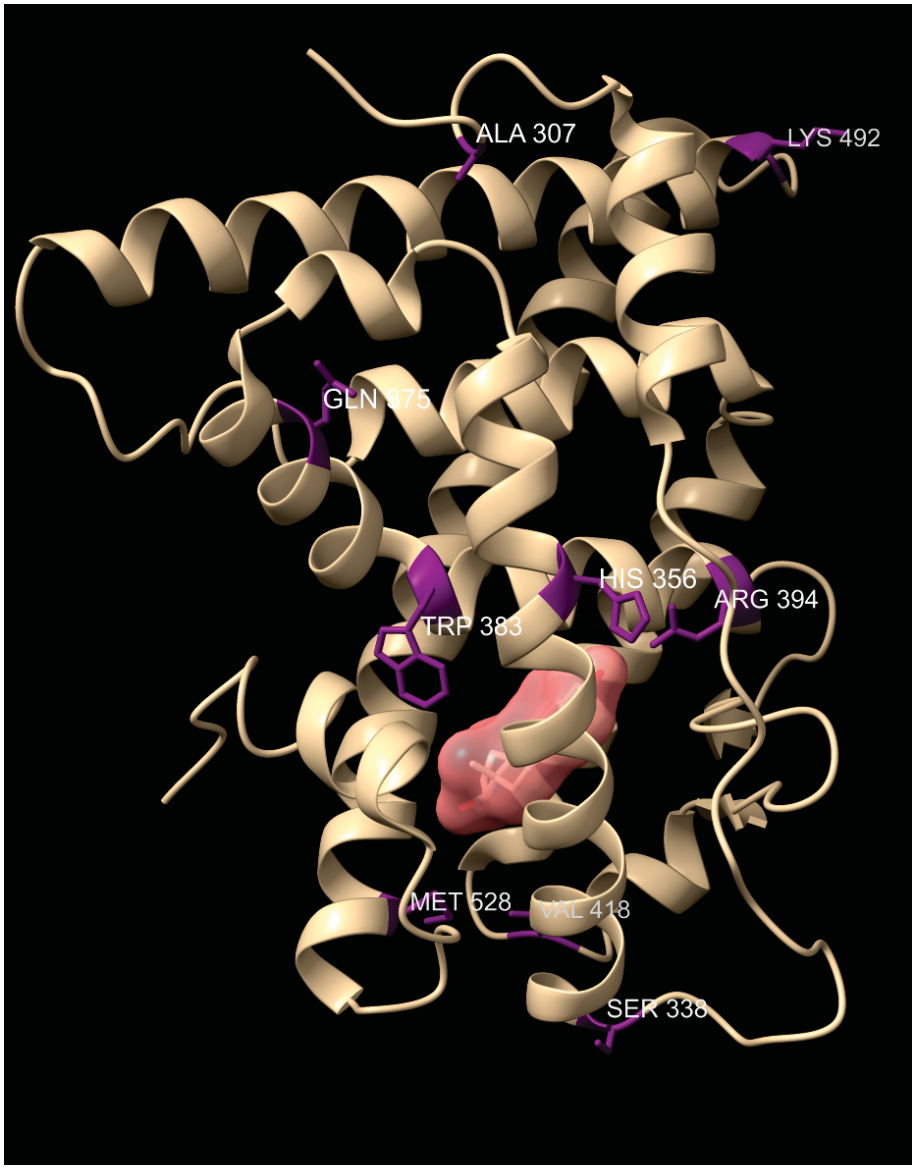


Figure 15: 3D Structure of ERα with Mutational Annotations
Mutated residues found in the cohort are colored in purple.
The ligand (estradiol) is colored in red.

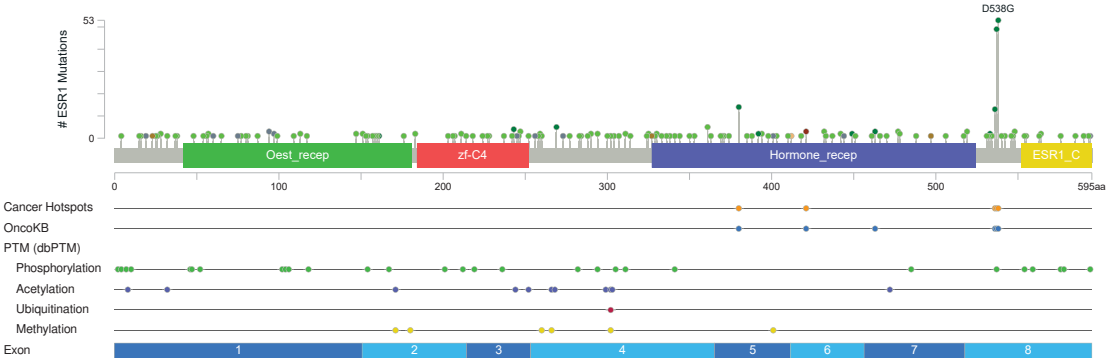


Figure 16: Known ESR1 Mutations Described in cBioPortal Using Multiple Pan-Cancer Studies
[68–76]

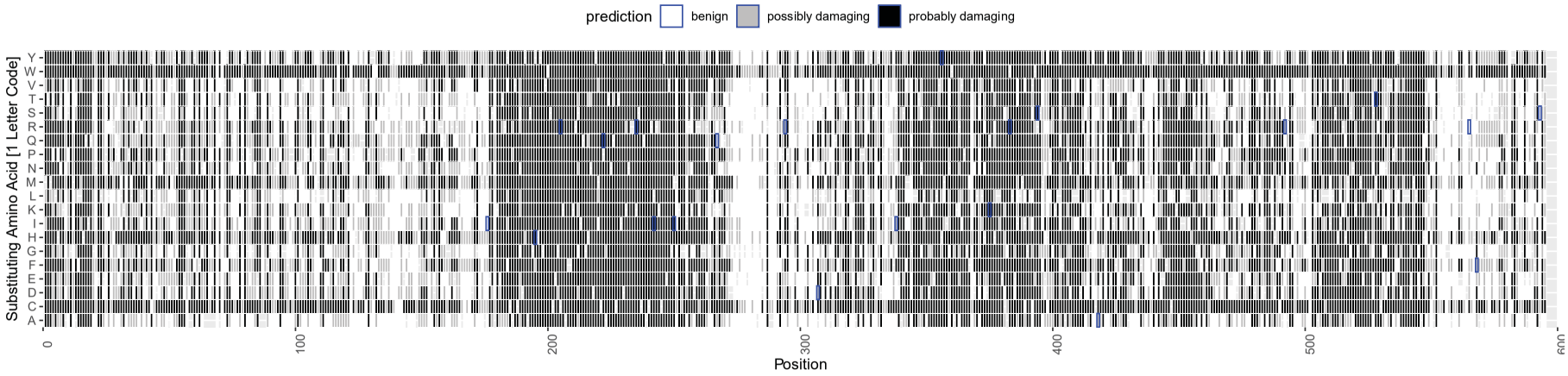
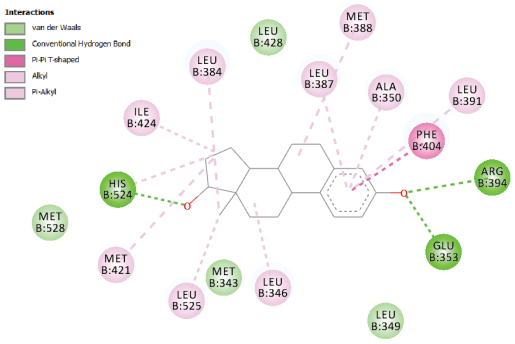


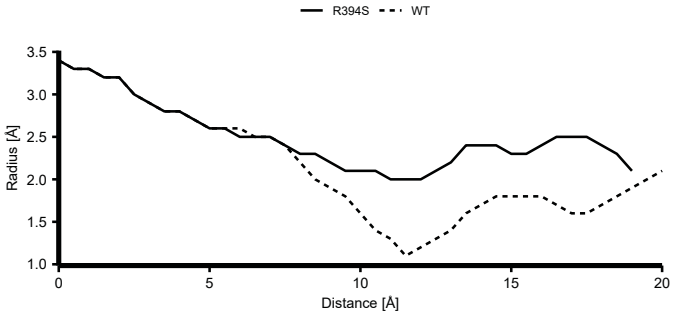
Figure 17: Mutation Impact Map of ESR1

Predicted effect of any possible amino acid exchange in ESR1 using PolyPhen2.

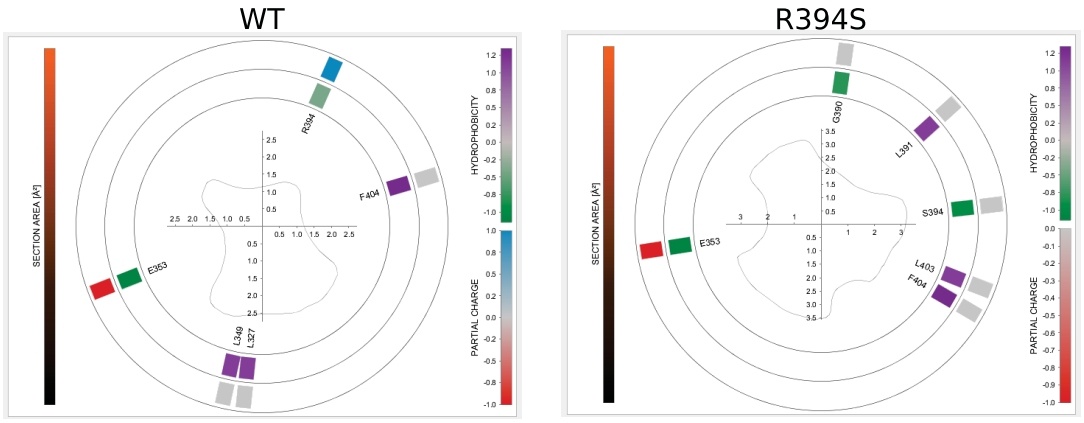
Encircled in blue are the SNPs found in the cohort.



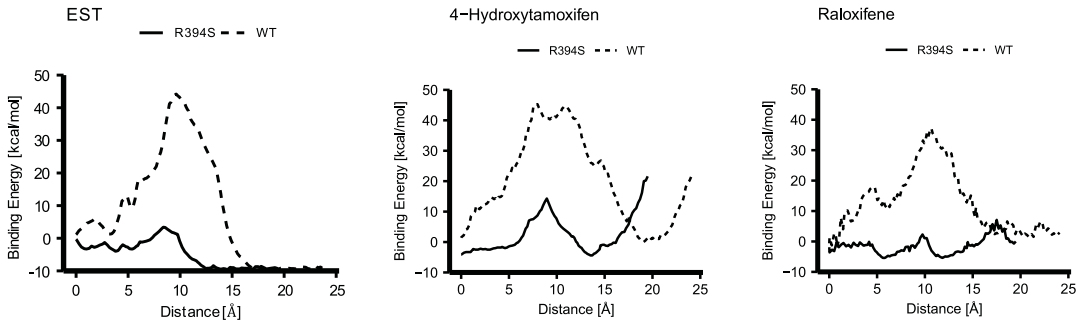
(a)



(b)



(c)



(d)

Figure 18: ESR1 R394S Computational Analysis

a) Interaction diagram of ERα in complex with estradiol. **b)** Longitudinal section profile of a ligand tunnel in WT and mutant ERα **c)** Cross section profile of a ligand tunnel in WT and mutant ERα **d)** Ligand transport analysis for estradiol, 4-hydroxytamoxifen and raloxifene in WT and mutant ERα

3.3.5 W383R, M528T May Alter Tamoxifen Interaction With ERα

While evaluating the 3D positioning of W383 and M528, I found that both mutations are positioned around the ”tail” of bound 4-hydroxytamoxifen in crystallization studies (see Fig. 19 a,b). Since this part of 4-hydroxytamoxifen is of special interest, as it confers the selective inhibiting effect of tamoxifen on ERα, I further investigated these two SNPs. In both cases, the amino acid exchange introduced a polar amino acid into the protein, implying an additional positive charge in physiologic pH. Tamoxifen has a formal neutral charge but the active metabolite 4-hydroxytamoxifen is predicted to have a positive charge of one at physiological pH levels (Fig. 19 c). It can be speculated that the subsequent change in forces alters the interaction of tamoxifen with ERα, as the effect of tamoxifen is described to heavily rely on the physical obstruction of helix 12 [38]. Furthermore, I calculated receptor-ligand forces using BIOVIA Discovery Studio, as described in the methods section above. I found that in the wild type protein neither W383 nor M528 exert a force on tamoxifen, which is unsurprising as both amino acids are nonpolar.

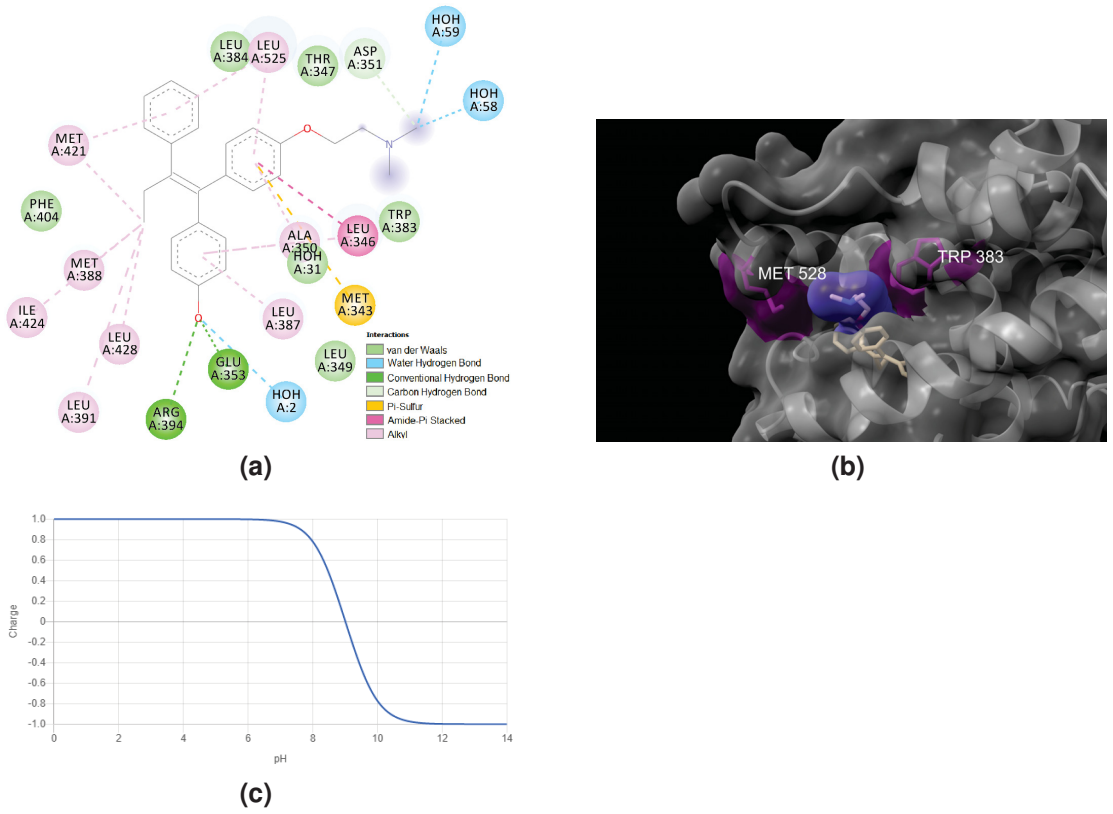


Figure 19: ESR1 W383 Computational Analysis

a) ERα in complex with 4-hydroxytamoxifen: Interaction Diagram

b) 3D structure of ERα in complex with 4-hydroxytamoxifen - colored in purple are two mutant residues in the cohort

c) Calculated charge of 4-hydroxytamoxifen. Indicating a positive charge under physiologic conditions [107].

4 Discussion

4.1 The Importance of Data Science in Cancer Research

With the advancement of high-throughput technologies like NGS, data science has become an integral part of everyday cancer research. Especially the analysis of single cells produces immense amounts of data, which requires special tools for thorough analysis. Sequencing of single cells comes with unique problems, as the loss of data is more prominent than in bulk sequencing approaches. Not only the type of amplification (through WGA or lack thereof) influences results of genomic sequencing, but also does every other downstream step have a great impact on the end result compared with bulk sequencing with proportionally higher DNA amounts. These biases need to be addressed with biological and computational means. Although variant calling of SNP in single cells can be challenging, emerging tools like Monovar take advantage of multiple sequenced single cells and utilize the additional data by comparing the coverage and minor allele frequencies between cells thus reducing false positive and negative rates [108]. Another field made possible by single cell genomic sequencing is phylogenomics which describes the evolutional tree of cancer cells and brings different samples into relation. Furthermore, studies like the recent multi-parametric analysis from my working group (coauthored by me) by Franken *et al.*, evaluate CTC and biopsy specimens and places the cells into relation. In this study, we utilized the tool Cloe which deduces information about sub-clones from the frequency of mutations [109, 110]. Studies like ours cast a positive outlook on the upcoming years, as we could theoretically track the evolution of cancer in real time and adjust the therapy regimen accordingly. Lähnemann *et al.* describe eleven challenges in singe cell data science, a main point is focused on the current difficulties in phylogenomics [111]. In essence, they point out that big data volumes overwhelm current algorithms and that different sources (SNP, CNV, ...) of data cannot be reliably combined into a single model [111].

Another emerging field of cancer research is the investigation and integration of clinical data with high throughput and image data. Platforms like cBioPortal publish significant amounts of data and make them accessible to every researcher in the world [76]. Furthermore, as electronic patient surveys and modern sensors become more advanced, the data we can analyze is more diverse than ever before. This leads to data sets of high dimensionality, which may contain insights to an equally complex disease. The advent of open science repositories promises to accelerate research in key areas, but it is too early to know if significant clinical benefit is generated by them. Recently open science has received a boost in attention, as open research and exchange of big data sets became essential during the Covid-19 pandemic and funding for open platforms grew rapidly [112].

4.2 ViBiBa

While LB analysis is not new, it is still a challenge to obtain relevant numbers of CTCs and ctDNA for translational research. Even though CTCs are getting more attention as possible independent predictors of therapy response, we suffer from limited knowledge due to low availability. In contrast to classic clinical cancer trials where blood or tissue specimens are analyzed, CTCs create an increased logistical overhead. The analysis of CTCs is time-sensitive and requires advanced equipment, which is not readily available at every clinic. This creates the need to ship patient samples to capable laboratories. A further hurdle is the low positivity rate in entities like breast cancer, as many patients do not display CTCs at all or only a low count of CTCs. Translational research on CTCs requires access to a big pool of CTCs, ideally with multiple samples from the same patient. Modern CTC trials like DETECT [32] are designed to be decentralized multicenter studies. In this setting, every specimen is sent to a random laboratory, often separating samples from the same patient. [1]

While clinically established parameters like CTC count and immunohistochemistry results are reported directly into clinical trial management systems (CTMS), newer CTC downstream analysis (that may not be standardized yet) is not centrally reported. This leads to a loss of data, as none of the participating laboratories knows to which extent samples are available and in which form these specimens have already been processed. Previous studies like the European Human Frozen Tumor Tissue Bank (TuBaFrost) [113] have shown that the adoption of a common database can be hindered by a complex upload process. Thus ViBiBa tries to minimize the effort a participating laboratory must undertake to an absolute minimum. The upload procedure of ViBiBa allows for a "*drag and drop*" import of excel or csv files without changing the file structure at all. ViBiBa facilitates sufficient cohort sizes, which enables translational research like the ENDOpanel or the ESR1 sequencing project. The functionality of ViBiBa is not limited to multicenter applications, as recent findings suggest that keeping track of samples is challenging even in single center studies [114].

ViBiBa is not the first virtual biobanking solution. One predecessor is TuBaFrost from 2006 [115]. While TuBaFrost shares the general idea of decentralized storage, it relies heavily on standardized data input, which ultimately led to the termination of the project as participation was low [113]. The European bone tumors network (EuroBoNeT) [113] is the direct successor to TuBaFrost and broadened the range of stored sample types, but ultimately did not reach widespread adoption. [1]

ViBiBa fulfills all previously defined requirements for the DETECT trial group. It is capable of building a standardized database from multiple inhomogeneous data sources, while making it easy for the user to upload and explore data. Previous applications failed in the long run due to low adoption rates. In contrast, the experience with ViBiBa in the DETECT trial group gives a positive outlook, while the long-term adoption still remains to be seen. Our application tries to avoid known pitfalls by simplifying the process as much as possible. [1]

4.3 LB: Perspectives and Recent Advancements

4.3.1 Precision Medicine (LB) Tumor Boards

Molecular tumor boards, also referred to as precision medicine tumor boards, are slowly getting traction in clinical care. With the approval of targeted therapies, companion diagnostic tests in breast cancer and an increase in general complexity of treatment algorithms, a dedicated group of experts is required to decide upon the best care for a patient. One example is the recently FDA-approved drug alpelisib, a PI3K inhibitor which can be utilized in combination with fulvestrant in HR-positive HER2-negative patients with a PIK3CA mutation [116]. This mutation can be detected with the therascreen companion diagnostic kit on genomic DNA from solid tumor tissue or ctDNA [116]. Further therapies that molecular tumor boards have to consider are PARP inhibitors in patients with BRCA1/2 mutations, resistance against AI in ESR1 mutant patients and PD-1 inhibitor pembrolizumab in patients with microsatellite instability [117]. Such molecular tumor boards can extend regular breast cancer tumor boards which mainly consist of gynecologists, oncologists, pathologists and radiologists. Additional positions in precision medicine tumor boards include bioinformaticians, genetic counselors and molecular pathologists [117]. A molecular tumor board described by Sultova *et al.* ordered NGS tests on all 95 patients included in the tumor board and performed NGS analysis more than once on four patients [117]. In this population Sultova *et al.* found 41 patients with actionable mutations and recommended 15 diagnostic tests and 49 treatment plans [117]. Some patients received more than one possible treatment plan, as the NGS testing revealed more than one actionable target. Further, only 9 treatment plans were pursued, resulting in an actual impact on treatment in less than one tenth of patients [117]. The concept of precision medicine tumor boards was also tested in the randomized, open-label, multicenter SHIVA trial [118]. In this French study by Le Tourneau *et al.* patients with a metastatic cancer not responding to primary treatment were included. Subsequently, 741 patients were screened and 195 randomized into experimental (molecular tumor board) and control (physician's choice) groups. Unfortunately, no significant difference in PFS could be demonstrated ($p = 0.41$) [118]. While results from the SHIVA trial may look discouraging at first glance, they rather point towards the remaining research that has to be performed prior to routine NGS screenings of cancer patients. Furthermore, the study was also criticized for its design, which allowed for the usage of therapies with unknown performance for a given target, thus undermining the aspect of targeted therapies [119]. Medications inside their respective indication for targeted therapy (e.g., alpelisib) lead to a better outcome in patients [116]. Consequently, off-label use of targeted therapies based on NGS or other molecular profiling should be reserved for clinical trials and not incorporated into tumor boards. A recent review by Larson *et al.* analyzed 14 published studies of molecular tumor boards and over 3,000 patients [119]. In their review, they found heterogenous data and subsequently were not able to perform a meta-analysis. One of the reviewed studies focused on breast cancer and found that none of

the 43 patients shared identical abnormality profiles [120]. Parker *et al.* demonstrated a mean latency of 23 days between availability of results and the recommendation of the molecular tumor board [120]. Further, the authors found that 40% of patients received the recommended treatment. Moreover, when comparing patients who received a matching targeted therapy with patients who did not, the authors found a beneficial PFS increase of 5.1 vs. 2.4 months respectively ($p = 0.029$) [120].

4.3.2 LB Diagnostic Tests

The first FDA approved LB diagnostic test is the CellSearch system which received approval in 2004 for clinical usage of CTC enumeration in MBC patients [121]. In the following decade, the CellSearch system expanded its approval towards the monitoring of colorectal and prostate cancer, without significant alteration of the test itself [121]. In 2016 the FDA granted the cobas EGFR Mutation Test v2 clearance for EGFR mutational testing in cfDNA to guide therapy decisions in non-small cell lung cancer [122]. A leading example for CTC diagnostics is the detection of the AR-V7 splice variant in metastatic castration-resistant prostate cancer. The splice variant confers ligand independent activation of the androgen receptor and is linked to a worse outcome [24]. As the detection of this splice site variant currently requires mRNA or protein measurement, it cannot be performed with cfDNA [24]. Potential new tests include new pan-cancer LB approaches, which cover a broad spectrum of genes to streamline the analysis of LB specimens independently of tumor entities. Recently the first pan-cancer LB genetic test got FDA approval. Guardant360 CDx is a cfDNA NGS test targeting 55 genes, copy number alterations in two genes and four fusion genes [123]. Further, the test is approved as a companion diagnostic test in non-small cell lung cancer as the detection of EGFR mutations, insertions and deletions allows to identify patients who may benefit from the novel drugs osimertinib and amivantamab [123].

4.3.3 Detection of CTCs and Associated Biases

Further, a possible hurdle is the reliable detection of CTCs. As the gold standard for CTC enrichment is based on the EpCAM marker, the subsequent analysis relies on the expression of the epithelial cell adhesion molecule. It is speculated that EpCAM positive CTCs themselves are not the relevant cells behind metastasis and only function as a surrogate marker for other CTCs with metastatic potential [124]. This hypothesis is built upon the EMT paradigm (as described in section 1.6). As carcinomas (e.g., breast cancer) are derived from epithelial cells, they have to undergo EMT and acquire mesenchymal characteristics to break through the basement membrane to invade neighboring tissues [19, 125]. Similarly, CTCs that suppress epithelial markers like EpCAM are thought to be more aggressive while remaining undetected by the CellSearch technology [124]. Research into these $\text{EpCAM}^{\text{low}}$ cells yields mixed results, as some trials find no correlation between number of $\text{EpCAM}^{\text{low}}$ CTCs with survival [126]. In non-small cell lung cancer,

a combined approach for CTC enrichment via EpCAM, EGFR and HER3 antibodies has been developed [127]. Utilizing this antibody cocktail, the authors were able to enrich more CTCs but stopped short of demonstrating benefits from the detection of additional CTCs [127].

4.3.4 CTC Count and Image Analysis

One of the contributing factors of uncertainty in CTC detection is the subjectivity of image-based detection, as human operators have to judge if an image of a cell displays a CTC or not. In a trial by Zeune *et al.* 15 trained reviewers were tasked to evaluate 100 objects. Only once all 15 reviewers agreed that a particle constitutes a CTC. Moreover, the agreement of the reviewer (measured as Fleiss' κ with 0 = no agreement and 1 = perfect agreement) was measured as $\kappa = 0.38$, demonstrating a moderate agreement between the reviewers [128]. Image analysis driven by artificial intelligence and predefined selection criteria as tested in the ACCEPT software is the first step to a more objective classification of CTCs and could subsequently improve the utility and reliability of CTC NGS approaches [129].

4.3.5 Prospects of Genomic Analysis in LB

Genomic analysis of LB is one of the stepping stones towards precision medicine. Even mutations in rare sub-clones of the solid tumor are detectable in CTCs. Heitzer *et al.* define "private CTC mutations" as mutations that are only present on CTCs without detection in tumor material. In their study Heitzer *et al.* find that after ultra deep sequencing 85% of "private CTC mutations" are found in tumor material in frequencies as low as 2% [130]. For sequencing of CTCs the single cells need to be isolated, requiring specialized machinery, time and trained operators. Additionally, the isolation of single cells comes with the risk of losing cells in the process. As most patients harbor a very small number of CTCs, the loss of a rare cell is detrimental for downstream analysis [131]. Moreover, CTCs can be analyzed for copy number alterations with Carter *et al.* successfully predicting chemosensitivity based on CTC CNV analysis of small cell lung cancer in 83% of patients and demonstrating a significant difference in PFS (n=31) [132].

4.4 ENDOpanel

The central aim of the ENDOpanel is getting a broader picture of the mutational (endocrine resistance) landscape in single cells. In this work, the ENDOpanel could be established as a reliable method for single cell analysis. While the ENDOpanel does not cover 100% of all targeted positions, I could identify potential hurdles in the workflow that lead to worse results when using WGA products.

4.4.1 Coverage

The coverage analysis demonstrated that 74% of all positions could be sufficiently covered, whereas 13 exons were not detected by the ENDOpanel at all. To test if the ENDOpanel could be used in real-world scenarios, I calculated the coverage of known cancer hotspots. Since 88% of hotspots were covered at least in one sample, a broad range of potential regions can be detected. The data shows good initial coverage which can certainly be improved. A comparison to other LB NGS panels can be found below (section 4.4.4).

4.4.2 DNA Fragments

Utilizing significant evidence, I propose that shorter DNA fragments experience less coverage in the final sequencing result. The DNA fragments are created by the MSE1 fragmentation during the WGA of CTCs, making this pitfall especially interesting in single cell analysis. Since the recognition pattern of MSE1 is not scattered evenly throughout the human genome, the resulting fragments are of varying length. This could be due to insufficient DNA capture, e.g., during purification steps. During the single cell workflow, the DNA is cut multiple times: once during the Ampli1 WGA utilizing the MSE1 restriction enzyme and once more during the SureSelectQXT library preparation utilizing a transposase-mediated DNA fragmentation. While I could enhance the capture rate of short fragments in synthetic tests, this effect could not be recreated in the real workflow. One possibility is that more off-target DNA and other byproducts were captured by the increased magnetic beads volume. An additional indicator for worse coverage is the significantly lower DNA amount when utilizing more magnetic beads. Ultimately, the ENDOpanel could perform better when using non-WGA (pooled DNA) products or when utilizing a different library preparation like the SureSelectXT, which relies on physical shearing (using the Covaris system), as this could potentially lessen the ratio of short fragments. While enzymatic fragmentation relies on the distribution of DNA recognition patterns (which are mostly random), physical shearing leads to fragments of roughly equal length.

4.4.3 CTC Versus ctDNA Sequencing

To bring the results of the ENDOpanel into context, they need to be compared against other LB sequencing approaches. When comparing gene panels, the quantity and quality of the initial DNA need to be taken into account as well. Sequencing of single cells comes with the inherent need for WGA, which subsequently involves the risk of allelic dropouts. On the other hand, ctDNA cannot be analyzed on its own, as it forms only a fraction of the cell-free DNA (cfDNA) inside blood samples. Additionally, the fraction of tumor DNA inside cfDNA varies in the range of 0.1% to 90%, creating a high background noise when analyzing cfDNA with low tumor DNA load [133]. Consequently, methods with

high sensitivity for singular mutations, without broader coverage are utilized in cfDNA (e.g., BEAMing), while NGS approaches are currently only possible and interpretable in patients with higher ctDNA loads [133]. In comparison, single cell analysis has no natural background noise from other cells. These drawbacks have limited the clinical utility of CTCs and ctDNA to detection of residual disease, while precision medicine capabilities are slowly emerging [133]. Furthermore, ctDNA is naturally fragmented while the degree of fragmentation can vary by entity and tumor load [134].

4.4.4 Other LB Gene Panels

One example in the field of ctDNA is the 180-gene PredicinePLUS panel as demonstrated by Davis *et al.* [135]. PredicinePLUS covers 565 kb of genomic sequences with a target sequence depth over 20,000 [135]. Unfortunately, the study does not report a mean sequencing depth. Instead, it states that 40 of 43 samples reach a depth of over 3,000 reads on 90% of the target genomic range. The study finds a similar correlation between ESR1 mutational burden in LB and CTC count ($p = 0.0017$) as shown in Fig. 14. A more targeted approach with a smaller genomic range of 5995 bases was demonstrated by Forshew *et al.* producing an average sequencing depth of 3000 reads in solid tumor biopsy specimens and a mean depth of 650 in cfDNA [136]. Another example is the TruSight Oncology 500 kit from Illumina, a hybrid based approach targeting 523 cancer genes [137, 138]. In an exhaustive study Liu *et al.* compared CTC NGS analysis in fixed and fresh cells with different pre-processing steps [139]. The authors utilized the Qiagen GeneRead DNAseq Colorectal Cancer Panel and compared two WGA kits: REPLI-g and WGA4 [139]. Liu *et al.* defined a target sequencing depth of ≥ 10 reads. The target was reached by 97.7% of genomic positions from the fresh cells batch with REPLI-g WGA [139]. The coverage dropped to 89.7% for fixed cells with REPLI-g WGA and only reached 48.2% on fixed cells with the WGA4 kit [139]. In comparison, the ENDOpanel achieved an overall coverage of 74% with the same depth of ≥ 10 reads. The lower coverage could be explained by the different WGA (Ampli1) used by the ENDOpanel as the resulting fragmentation of the DNA could lead to a worse coverage, as already discussed. When compared with other panels, the ENDOpanel offers a viable approach towards single cell NGS analysis. Possible improvements of the ENDOpanel coverage are discussed below.

4.4.5 Choice of WGA Kit

As already described in section 1.12.1, numerous WGA methods are currently on the market. In a recent review, Biezuner *et al.* compare seven kits for single cell WGA which are commercially available [140]. One of the tested kits (Ampli1) was also utilized for the ENDOpanel and the ESR1 sequencing. Conveniently, the reviewers also isolated single cells with the CellCelector, similarly to the presented workflow of the ENDOpanel [140]. To measure the performance of the different kits, the authors only considered amplicons

on the X chromosome to avoid confounding through allelic dropouts and had to eliminate amplicons containing the MSE1 recognition pattern. The observed metrics were genome coverage, reproducibility and error rate [140]. Ampli1 showed superior coverage with a median of 1095.5 (out of 1585) amplicons covered with the closest competitor being REPLI-g with a median of 918 amplicons [140]. Although Ampli1 has the strongest overall coverage, it also demonstrates variability with coverage in some cells dropping well below 200 amplicons [140]. Further, the authors analyze reproducibility by calculating the number of intersecting amplicons in all cell pairs [140]. In this analysis Ampli1 once again showed the best performance, but the authors also found that every WGA kit induced a systematic bias in amplicons that are covered [140]. At last, Biezuner *et al.* determined the error rate which resembles the frequency of mutations during the in vitro amplification [140]. As anticipated by the authors, the kits based on isothermal multiple displacement amplification (e.g., REPLI-g) produce a lower error rate in contrast to the Ampli1 kit which utilizes a linker adapter PCR [140]. In my working group, we decided to utilize Ampli1 for single cell genomic DNA amplification, as it results in the highest coverage [140]. While REPLI-g generates fewer errors, it also sometimes fails to amplify the genome at all and results in a lower, less reliable coverage [140].

4.4.6 Possible Enhancements of the ENDOpanel

Independently of the utilized NGS approach, the user can reduce PCR biases and errors with molecular barcodes, which label individual DNA fragments with unique adapter sequences. Multiple studies could show a better detection limit when utilizing molecular barcodes in LB. Masunaga *et al.* demonstrated an improvement of the detection limit from 1% minor allele frequency to 0.1% in ESR1 sequencing in cfDNA [141]. Moreover, De Luca *et al.* demonstrated the use of molecular barcodes in a WGA free approach on CTC pools of 2 to 5 cells with a nearly perfect recall ratio of 35 of 37 CTC pools and no false positive mutations [142]. Such a pre-processing step could be integrated into the ENDOpanel and similar NGS panels as it promises to improve the clinical utility of such a panel. Molecular barcodes in a WGA free approach could eliminate the inherent downsides of WGA in CTC sequencing. On the other hand, this method requires the pooling of CTCs while decreasing the spatial resolution of the sequencing results. Another way to improve the coverage without utilizing molecular barcodes, would be through bulk sequencing of CTCs. To capture a high enough number of CTCs for sequencing without WGA, a novel approach such as diagnostic leukapheresis (DLA) is required. Leukapheresis typically targets mononuclear cells while the blood of the patient flows continuously through an extracorporeal centrifuge [143]. Moreover, this method can be used to enrich CTCs as they have a similar density as mononuclear cells [143]. The patient's veins are punctured on both arms, blood is withdrawn from the one arm and returned through the other puncture site [143]. Crucially, this method allows to screen a higher blood volume of the patients when compared to CellSave tubes, which are normally in use for CTC analysis. A recent single center study demonstrated a 200 fold estimated increase in CTC

capture rate. The authors estimated a median screened blood volume of 2,770 ml and reported no severe adverse events [143]. Further, the ENDOpanel could benefit from a switch from the SureSelectQXT platform to the SureSelectXT platform as this could create an equal distribution of fragments without the risk of loss of short fragments.

4.5 Estrogen Receptor Alpha Sequencing

4.5.1 ESR1 as a Predictive Marker

ERα is a well-described target for endocrine breast cancer therapy. Likewise, mutations in ERα are a heavily researched cause for resistance to endocrine therapy. This was not always the case, they were initially not associated with clinical significance, as ESR1 mutations are rare in the primary tumor [144]. Moreover, LB specimen is known to be rare in early stages of breast cancer, further reducing the chance of ESR1 mutation detection [144]. A study by Takeshita *et al.* found an increase in ESR1 mutations in cfDNA in later treatment lines without a significant difference in time to treatment failure in ESR1 mutant patients [145]. Another study found that both overall and progression-free survival are reduced in patients harboring ESR1 mutations [144]. In a recent meta-analysis, both ctDNA levels and ESR1 mutation burden were found to be associated with survival, prompting the question how the two predictors influence each other [144]. Furthermore, clinical trials begin to perform LB tests with as little as 15 days of latency after treatment onset, raising the question if a rapid switch in treatment regimes should be performed as a response to an arising resistant sub-clone [144, 146]. As described by Carausu *et al.* the evidence for the predictive value of ESR1 in LB is sparse due to the limitations of LB trials, e.g., small positivity rate [144]. This hurdle is a core aspect of ViBiBa, as already described.

Currently, ESR1 mutational analysis is focused on hotspot regions (e.g., Y537 and D538) as most of the mutations inside the hotspot regions display promising preclinical data for sensitivity to fulvestrant (an SERD) [41]. Moreover, some studies find significant evidence for ESR1 mutations, conferring resistance against AI therapy through ligand-independent activation [144]. In contrast, SERM usage in ESR1 mutant cancers yields mixed results while SERDs like fulvestrant appear to still be able to degrade mutant ERα (e.g., D538G) [144]. While the correlation between the pure presence of ESR1 mutations and CTC count is already known [147], I could show that the frequency of non-hotspot ESR1 mutations is also significantly correlated with the CTC count. This links our understanding between two independent prognostic markers, CTC count and the ESR1 mutational burden (including non-hotspot regions) [148].

Many studies focus on methods like digital droplet PCR (ddPCR) which cover only a narrow genomic range often targeting already known hotspot mutations. This may lead to an under-representation of non-hotspot mutations in ESR1. Since treatment naive PTs are known to harbor less ESR1 mutations than cancer tissue under endocrine therapy (e.g., metastases) [18] and big public repositories like TCGA focus on the PTs, further

under-reporting of ESR1 non-hotspot mutations can be expected. To utilize ESR1 as a predictive marker outside of hotspot regions, it is paramount to establish a workflow to evaluate their potential effects on endocrine resistance development. This will become eminently important if only one or a few CTCs with functionally unclear ESR1 mutations are detected.

4.5.2 Other ESR1 LB Studies

Other studies focused mainly on cfDNA and subsequent in vitro analysis via mutant cell culture experiments [149, 150]. A recent work on cfDNA by Jeannot *et al.* utilized ddPCR for ESR1 mutation detection in LB [150]. Their study detects mutations only in the codons 380, 536, 537, and 538 accounting for 12 DNA bases of the 6327 coding base pairs [150]. As the study focused on cfDNA, a low limit of detection is required to find rare sub-clones in patients with a low ctDNA fraction inside the cfDNA sample. The limit of detection calculated by Jeannot *et al.* is as low as 0.07 % minor allele frequency. In contrast, this level of precision is not required when analyzing single cells, as the minor allele frequency does not reach such low levels when only two alleles are inspected.

4.5.3 Novel Resistance Mechanisms

In this study I analyzed three selected mutations in more depth with in silico methods. The keyhole-lock-key model describes the ligand tunnel as a filter for the ligand in front of the LBD [43]. This opens up new possibilities for research of (endocrine) treatment resistance. Mutations in the ligand tunnels of ERα are not widely studied and may be under-reported in light of infrequent sequencing outside of ESR1-hotspot regions. Further, in other proteins, it was already demonstrated that a mutation of a tunnel forming residue can lead to a different protein-ligand kinetic [151–153]. Modern tools allow us to calculate tunnels through the protein and to discover residues at tunnel bottlenecks, which may be critical determinants in ligand transport [47, 86, 87]. Contrary to other protein-ligand complexes, there is no clear path in ERα that a ligand would take to reach the deep-buried LBD [154]. I identified R394 as a potential bottleneck, whose mutation to serine alters the radius of the ligand tunnel. Knowing that the patient received tamoxifen allows us to perform simulations of the interactions with the (mutated) ERα. The data points towards a change in the keyhole-lock-key behavior of ERα when interacting with tamoxifen and estradiol. While the affinity towards estradiol increased tenfold, only a mild increase in affinity was noted for tamoxifen. This change may result in an observed lower efficacy of tamoxifen in R394S mutant cells and could therefore represent a novel mechanism of endocrine resistance. Further in the W383R and M528T mutants, I hypothesize that the interaction between tamoxifen and ERα could be altered. Activation of ERα is induced by conformation change of helix 12, which can be physically inhibited, e.g., by the tail of bound tamoxifen [38]. According to the physicochemical properties of the mutated residues (positively charged receptor residue with a positively charged tamoxifen ligand)

this repelling force could hamper ligand interaction and thus possibly confer tamoxifen resistance [155, 156]. Additionally, while I did not attempt any in silico calculations for the H356Y SNP, this novel mutation occurred in two patients in a total of five cells.
The mutation impact heatmap is a first approach on mapping all possible ESR1 mutations. With the discovery of new in silico predictions, the heatmap can certainly be fine-tuned more accurately by iterating through simulations targeted on specific mutated protein structures and specific drugs. In future studies, the aim should be to expand our knowledge on possible treatment resistance mechanisms in ERα that can be modeled in silico and used to create a simple mutation and drug-specific chart on ESR1 mutation significance.

4.5.4 Limitations of In Silico Approaches

Currently, in silico analysis is limited by our understanding of proteins and their behavior/folding on a molecular level and limited data availability for model systems (e.g., protein spectrometry data). The protein model chosen for the demonstrated calculations is deducted from the ligand-bound form. An apo-protein spectrometry version would be more fitting, but high-resolution data is still lacking in this field.

4.5.5 Precision Medicine: Timely Evaluation of Novel Mutations

The presented work focuses on CTCs and in silico methods to determine a possible impact of SNPs on endocrine treatment response. Although in vitro approaches offer a broad spectrum of analysis, they are also associated with a high investment of time. In the clinical setting, a novel mutation has to be assessed in a timely manner to evaluate treatment options for the patient. One advantage of the presented method is that the current treatment regime of the patient can be directly fed into the in silico workflow. This enables us to perform targeted bioinformatic analysis with the mutant ERα amino acid and the (potentially) affected drugs.

4.6 Conclusion

ViBiBa was successfully launched inside the DETECT trial program and the source code published under an open source license. It is in daily usage at the time of writing and maintained with regular updates. User engagement is adequate and a few sample transfers have already been completed through ViBiBa.

The ENDOpanel could be utilized as a viable approach for CTC analysis. While it displays some shortcomings, I produced multiple hypotheses for the loss of coverage. These hypotheses were analyzed in detail and some of them demonstrated evidence for possible improvement. A major hurdle will be future cost reductions of competing technologies like whole exome sequencing, which will allow to sequence all genes without a specific focus *a priori*.

In the ESR1 sequencing project, I selected a patient cohort with an already established lack of ESR1-hotspot mutations. CTCs from this cohort were analyzed via NGS of the coding ESR1 genomic range to access for possible (novel) ESR1 mutations. I demonstrated the feasibility of this targeted sequencing of ESR1 and illustrated how some of the described mutations may impact endocrine resistance in breast cancer. This work generated new hypotheses for endocrine resistance mechanisms utilizing in silico approaches and showed how these approaches can be personalized for a specific patient with a specific treatment, making it an option for individual analyses while remaining cost efficient.

List of Figures

List of Tables

References

1. Asperger, H. *et al.* ViBiBa: Virtual BioBanking for the DETECT Multicenter Trial Program - Decentralized Storage and Processing. *Translational Oncology* **14,** 101132. ISSN: 1936-5233. pmid: 34051621 (Aug. 2021).
2. Robert Koch-Institut & Gesellschaft Der Epidemiologischen Krebsregister In Deutschland E.V. Krebs in Deutschland 2015/2016. https://edoc.rki.de/handle/176904/6012.3 (2019).
3. Antoniou, A. *et al.* Average Risks of Breast and Ovarian Cancer Associated with BRCA1 or BRCA2 Mutations Detected in Case Series Unselected for Family History: A Combined Analysis of 22 Studies. *American Journal of Human Genetics* **72,** 1117–1130. ISSN: 0002-9297. pmid: 12677558 (May 2003).
4. Nicolini, A., Ferrari, P. & Duffy, M. J. Prognostic and Predictive Biomarkers in Breast Cancer: Past, Present and Future. *Seminars in Cancer Biology* **52,** 56–73. ISSN: 1096-3650. pmid: 28882552 (Pt 1 Oct. 2018).
5. Perou, C. M. *et al.* Molecular Portraits of Human Breast Tumours. *Nature* **406,** 747–752. ISSN: 0028-0836. pmid: 10963602 (Aug. 17, 2000).
6. Kreienberg, R. *et al.* Interdisziplinäre S3-Leitlinie für die Diagnostik, Therapie und Nachsorge des Mammakarzinoms. *Senologie - Zeitschrift für Mammadiagnostik und -therapie* **10,** 164–192. ISSN: 1611-6453, 1611-647X (Sept. 16, 2013).
7. Beatson, G. T. On the Treatment of Inoperable Cases of Carcinoma of the Mamma: Suggestions for a New Method of Treatment, with Illustrative Cases. *Transactions. Medico-Chirurgical Society of Edinburgh* **15,** 153–179. ISSN: 0267-2790. pmid: 29584099 (1896).
8. Pike, M. C., Spicer, D. V., Dahmoush, L. & Press, M. F. Estrogens, Progestogens, Normal Breast Cell Proliferation, and Breast Cancer Risk. *Epidemiologic Reviews* **15,** 17–35. ISSN: 0193-936X. pmid: 8405201 (1993).
9. Allison, K. H. *et al.* Estrogen and Progesterone Receptor Testing in Breast Cancer: ASCO/CAP Guideline Update. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* **38,** 1346–1366. ISSN: 1527-7755. pmid: 31928404 (Apr. 20, 2020).
10. Huang, B., Warner, M. & Gustafsson, J.-Å. Estrogen Receptors in Breast Carcinogenesis and Endocrine Therapy. *Molecular and Cellular Endocrinology* **418 Pt 3,** 240–244. ISSN: 1872-8057. pmid: 25433206 (Dec. 15, 2015).

11. Burstein, H. J. *et al.* Adjuvant Endocrine Therapy for Women With Hormone Receptor-Positive Breast Cancer: ASCO Clinical Practice Guideline Focused Update. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* **37,** 423–438. ISSN: 1527-7755. pmid: 30452337 (Feb. 10, 2019).

12. Musgrove, E. A. & Sutherland, R. L. Biological Determinants of Endocrine Resistance in Breast Cancer. *Nature Reviews. Cancer* **9,** 631–643. ISSN: 1474-1768. pmid: 19701242 (Sept. 2009).

13. Clarke, R., Tyson, J. J. & Dixon, J. M. Endocrine Resistance in Breast Cancer–An Overview and Update. *Molecular and Cellular Endocrinology* **418 Pt 3,** 220–234. ISSN: 1872-8057. pmid: 26455641 (Dec. 15, 2015).

14. Ingle, J. N. *et al.* Fulvestrant in Women with Advanced Breast Cancer after Progression on Prior Aromatase Inhibitor Therapy: North Central Cancer Treatment Group Trial N0032. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* **24,** 1052–1056. ISSN: 1527-7755. pmid: 16505423 (Mar. 1, 2006).

15. Robinson, D. R. *et al.* Activating ESR1 Mutations in Hormone-Resistant Metastatic Breast Cancer. *Nature Genetics* **45,** 1446–1451. ISSN: 1546-1718. pmid: 24185510 (Dec. 2013).

16. Chandarlapaty, S. *et al.* Prevalence of ESR1 Mutations in Cell-Free DNA and Outcomes in Metastatic Breast Cancer: A Secondary Analysis of the BOLERO-2 Clinical Trial. *JAMA oncology* **2,** 1310–1315. ISSN: 2374-2445. pmid: 27532364 (Oct. 1, 2016).

17. Jeselsohn, R., Buchwalter, G., De Angelis, C., Brown, M. & Schiff, R. ESR1 Mutations—a Mechanism for Acquired Endocrine Resistance in Breast Cancer. *Nature Reviews. Clinical Oncology* **12,** 573–583. ISSN: 1759-4782. pmid: 26122181 (Oct. 2015).

18. Toy, W. *et al.* ESR1 Ligand-Binding Domain Mutations in Hormone-Resistant Breast Cancer. *Nature Genetics* **45,** 1439–1445. ISSN: 1546-1718. pmid: 24185512 (Dec. 2013).

19. Weinberg, R. A. *The Biology of Cancer* Second edition. 876 pp. ISBN: 978-0-8153-4219-9 978-0-8153-4220-5 (Garland Science, Taylor & Francis Group, New York, 2014).

20. Pantel, K. & Alix-Panabières, C. Circulating Tumour Cells in Cancer Patients: Challenges and Perspectives. *Trends in Molecular Medicine* **16,** 398–406. ISSN: 1471-499X. pmid: 20667783 (Sept. 2010).

21. Barradas, A. M. C. & Terstappen, L. W. M. M. Towards the Biological Understanding of CTC: Capture Technologies, Definitions and Potential to Create Metastasis. *Cancers* **5,** 1619–1642. ISSN: 2072-6694. pmid: 24305653 (Dec. 4, 2013).

22. Yu, M., Stott, S., Toner, M., Maheswaran, S. & Haber, D. A. Circulating Tumor Cells: Approaches to Isolation and Characterization. *The Journal of Cell Biology* **192,** 373–382. ISSN: 1540-8140. pmid: 21300848 (Feb. 7, 2011).

23. Allard, W. J. *et al.* Tumor Cells Circulate in the Peripheral Blood of All Major Carcinomas but Not in Healthy Subjects or Patients with Nonmalignant Diseases. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* **10,** 6897–6904. ISSN: 1078-0432. pmid: 15501967 (Oct. 15, 2004).

24. Ignatiadis, M., Sledge, G. W. & Jeffrey, S. S. Liquid Biopsy Enters the Clinic - Implementation Issues and Future Challenges. *Nature Reviews. Clinical Oncology* **18,** 297–312. ISSN: 1759-4782. pmid: 33473219 (May 2021).

25. Janni, W. J. *et al.* Pooled Analysis of the Prognostic Relevance of Circulating Tumor Cells in Primary Breast Cancer. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* **22,** 2583–2593. ISSN: 1557-3265. pmid: 26733614 (May 15, 2016).

26. Bidard, F.-C. *et al.* Clinical Validity of Circulating Tumour Cells in Patients with Metastatic Breast Cancer: A Pooled Analysis of Individual Patient Data. *The Lancet. Oncology* **15,** 406–414. ISSN: 1474-5488. pmid: 24636208 (Apr. 2014).

27. Cristofanilli, M. *et al.* Circulating Tumor Cells: A Novel Prognostic Factor for Newly Diagnosed Metastatic Breast Cancer. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* **23,** 1420–1430. ISSN: 0732-183X. pmid: 15735118 (Mar. 1, 2005).

28. Bidard, F.-C. *et al.* Circulating Tumor Cells in Breast Cancer Patients Treated by Neoadjuvant Chemotherapy: A Meta-analysis. *Journal of the National Cancer Institute* **110,** 560–567. ISSN: 1460-2105. pmid: 29659933 (June 1, 2018).

29. Banys-Paluchowski, M., Krawczyk, N. & Fehm, T. Potential Role of Circulating Tumor Cell Detection and Monitoring in Breast Cancer: A Review of Current Evidence. *Frontiers in Oncology* **6,** 255. ISSN: 2234-943X. pmid: 27990412 (2016).

30. Rose Brannon, A. *et al.* Enhanced Specificity of Clinical High-Sensitivity Tumor Mutation Profiling in Cell-Free DNA via Paired Normal Sequencing Using MSK-ACCESS. *Nature Communications* **12,** 3770. ISSN: 2041-1723. pmid: 34145282 (June 18, 2021).

31. Cheng, D. T. *et al.* Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. *The Journal of molecular diagnostics: JMD* **17,** 251–264. ISSN: 1943-7811. pmid: 25801821 (May 2015).

32. Schramm, A. *et al.* Therapeutic Intervention Based on Circulating Tumor Cell Phenotype in Metastatic Breast Cancer: Concept of the DETECT Study Program. *Archives of Gynecology and Obstetrics* **293,** 271–281. ISSN: 1432-0711. pmid: 26354331 (Feb. 2016).

33. Fruman, D. A. *et al.* The PI3K Pathway in Human Disease. *Cell* **170,** 605–635. ISSN: 1097-4172. pmid: 28802037 (Aug. 10, 2017).

34. Agoulnik, I. U., Hodgson, M. C., Bowden, W. A. & Ittmann, M. M. INPP4B: The New Kid on the PI3K Block. *Oncotarget* **2,** 321–328. ISSN: 1949-2553. pmid: 21487159 (Apr. 2011).

35. Campbell, R. A. *et al.* Phosphatidylinositol 3-Kinase/AKT-mediated Activation of Estrogen Receptor Alpha: A New Model for Anti-Estrogen Resistance. *The Journal of Biological Chemistry* **276,** 9817–9824. ISSN: 0021-9258. pmid: 11139588 (Mar. 30, 2001).

36. Sridharan, S. & Basu, A. Distinct Roles of mTOR Targets S6K1 and S6K2 in Breast Cancer. *International Journal of Molecular Sciences* **21,** E1199. ISSN: 1422-0067. pmid: 32054043 (Feb. 11, 2020).

37. Jia, M., Dahlman-Wright, K. & Gustafsson, J.-Å. Estrogen Receptor Alpha and Beta in Health and Disease. *Best Practice & Research. Clinical Endocrinology & Metabolism* **29,** 557–568. ISSN: 1878-1594. pmid: 26303083 (Aug. 2015).

38. Shiau, A. K. *et al.* The Structural Basis of Estrogen Receptor/Coactivator Recognition and the Antagonism of This Interaction by Tamoxifen. *Cell* **95,** 927–937. ISSN: 0092-8674. pmid: 9875847 (Dec. 23, 1998).

39. Skafar, D. F. Formation of a Powerful Capping Motif Corresponding to Start of ”Helix 12” in Agonist-Bound Estrogen Receptor-Alpha Contributes to Increased Constitutive Activity of the Protein. *Cell Biochemistry and Biophysics* **33,** 53–62. ISSN: 1085-9195. pmid: 11322512 (2000).

40. Franken, A. *et al.* Detection of ESR1 Mutations in Single Circulating Tumor Cells on Estrogen Deprivation Therapy but Not in Primary Tumors from Metastatic Luminal Breast Cancer Patients. *The Journal of molecular diagnostics: JMD* **22,** 111–121. ISSN: 1943-7811. pmid: 31669227 (Jan. 2020).

41. Toy, W. *et al.* Activating ESR1 Mutations Differentially Affect the Efficacy of ER Antagonists. *Cancer Discovery* **7,** 277–287. ISSN: 2159-8290. pmid: 27986707 (Mar. 2017).

42. Reinert, T., Saad, E. D., Barrios, C. H. & Bines, J. Clinical Implications of ESR1 Mutations in Hormone Receptor-Positive Advanced Breast Cancer. *Frontiers in Oncology* **7,** 26. ISSN: 2234-943X. pmid: 28361033 (2017).

43. Kingsley, L. J. & Lill, M. A. Substrate Tunnels in Enzymes: Structure-Function Relationships and Computational Methodology. *Proteins* **83,** 599–611. ISSN: 1097-0134. pmid: 25663659 (Apr. 2015).

44. Eschenmoser, A. One Hundred Years Lock-and-Key Principle. *Angewandte Chemie International Edition in English* **33,** 2363–2363. ISSN: 0570-0833, 1521-3773 (Jan. 3, 1995).

45. Tripathi, A. & Bankaitis, V. A. Molecular Docking: From Lock and Key to Combination Lock. *Journal of Molecular Medicine and Clinical Applications* **2.** ISSN: 2575-0305. pmid: 29333532 (2017).

46. *Protein Engineering Handbook* (eds Lutz, S. & Bornscheuer, U. T.) 3 pp. ISBN: 978-3-527-31850-6 978-3-527-33123-9 (Wiley-VCH, Weinheim, 2009).

47. Stourac, J. *et al.* Caver Web 1.0: Identification of Tunnels and Channels in Proteins and Analysis of Ligand Transport. *Nucleic Acids Research* **47,** W414–W422. ISSN: 1362-4962. pmid: 31114897 (July 2, 2019).

48. *Whole Genome Amplification: Methods and Protocols* (ed Kroneis, T.) *Springer Protocols* **v. 1347.** 284 pp. ISBN: 978-1-4939-2989-4 (Humana Press, New York, 2015).

49. Klein, C. A. *et al.* Comparative Genomic Hybridization, Loss of Heterozygosity, and DNA Sequence Analysis of Single Cells. *Proceedings of the National Academy of Sciences of the United States of America* **96,** 4494–4499. ISSN: 0027-8424. pmid: 10200290 (Apr. 13, 1999).

50. *Next Generation Sequencing and Data Analysis* 1st ed. 2021 (ed Kappelmann-Fenzl, M.) 218 pp. ISBN: 978-3-030-62490-3 978-3-030-62489-7 (Springer, Cham, 2021).

51. *Covaris AFA: DNA/RNA Shearing for NGS* Online; accessed 4-Jan-2022. https://www.covaris.com/dna-rna-shearing-for-ngs/.

52. *Illumina: Patterned Flow Cells* Online; accessed 4-Jan-2022. https://emea.illumina.com/science/technology/next-generation-sequencing/sequencing-technology/patterned-flow-cells.html.

53. Li, H. & Durbin, R. Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics (Oxford, England)* **25,** 1754–1760. ISSN: 1367-4811. pmid: 19451168 (July 15, 2009).

54. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing Depth and Coverage: Key Considerations in Genomic Analyses. *Nature Reviews. Genetics* **15,** 121–132. ISSN: 1471-0064. pmid: 24434847 (Feb. 2014).

55. Combe, T., Martin, A. & Di Pietro, R. To Docker or Not to Docker: A Security Perspective. *IEEE Cloud Computing* **3,** 54–62. ISSN: 2325-6095 (Sept. 2016).

56. Asperger, H. *et al.* Progesterone Receptor Membrane Component 1 Regulates Lipid Homeostasis and Drives Oncogenic Signaling Resulting in Breast Cancer Progression. *Breast cancer research: BCR* **22,** 75. ISSN: 1465-542X. pmid: 32660617 (July 13, 2020).

57. *United States National Library of Medicine (NLM), clinical trials registry* Online; accessed 21-Dec-2021. https://clinicaltrials.gov/ct2/home.

58. *Automated single cell and colony picking system* Online; accessed 21-Dec-2021. www.als-jena.com/cellcelector-cell-and-colony-picking-system.html.

59. *Ampli1 WGA Kit* Online; accessed 21-Dec-2021. http://www.siliconbiosystems.com/ampli1-wga-kit.

60. Polzer, B. *et al.* Molecular Profiling of Single Circulating Tumor Cells with Diagnostic Intention. *EMBO molecular medicine* **6,** 1371–1386. ISSN: 1757-4684. pmid: 25358515 (Nov. 2014).

61. Grüning, B. *et al.* Bioconda: Sustainable and Comprehensive Software Distribution for the Life Sciences. *Nature Methods* **15,** 475–476. ISSN: 1548-7105. pmid: 29967506 (July 2018).

62. Afgan, E. *et al.* The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2018 Update. *Nucleic Acids Research* **46,** W537–W544. ISSN: 1362-4962. pmid: 29790989 (July 2, 2018).

63. Koboldt, D. C. *et al.* VarScan 2: Somatic Mutation and Copy Number Alteration Discovery in Cancer by Exome Sequencing. *Genome Research* **22,** 568–576. ISSN: 1549-5469. pmid: 22300766 (Mar. 2012).

64. Cingolani, P. *et al.* A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila Melanogaster Strain W1118; Iso-2; Iso-3. *Fly* **6,** 80–92. ISSN: 1933-6942. pmid: 22728672 (2012 Apr-Jun).

65. National Center for Biotechnology Information (NCBI). https://www.ncbi.nlm.nih.gov/protein/NP_000116.2.

66. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2020). https://www.R-project.org/.

67. Danecek, P. *et al.* Twelve Years of SAMtools and BCFtools. *GigaScience* **10,** giab008. ISSN: 2047-217X. pmid: 33590861 (Feb. 16, 2021).

68. Zehir, A. *et al.* Mutational Landscape of Metastatic Cancer Revealed from Prospective Clinical Sequencing of 10,000 Patients. *Nature Medicine* **23,** 703–713. ISSN: 1546-170X. pmid: 28481359 (June 2017).

69. Robinson, D. R. *et al.* Integrative Clinical Genomics of Metastatic Cancer. *Nature* **548,** 297–303. ISSN: 1476-4687. pmid: 28783718 (Aug. 17, 2017).

70. Miao, D. *et al.* Genomic Correlates of Response to Immune Checkpoint Blockade in Microsatellite-Stable Solid Tumors. *Nature Genetics* **50,** 1271–1281. ISSN: 1546-1718. pmid: 30150660 (Sept. 2018).

71. Hyman, D. M. *et al.* HER Kinase Inhibition in Patients with HER2- and HER3-mutant Cancers. *Nature* **554,** 189–194. ISSN: 1476-4687. pmid: 29420467 (Feb. 8, 2018).

72. Samstein, R. M. *et al.* Tumor Mutational Load Predicts Survival after Immunotherapy across Multiple Cancer Types. *Nature Genetics* **51,** 202–206. ISSN: 1546-1718. pmid: 30643254 (Feb. 2019).

73. Rosen, E. Y. *et al.* TRK Fusions Are Enriched in Cancers with Uncommon Histologies and the Absence of Canonical Driver Mutations. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* **26,** 1624–1632. ISSN: 1557-3265. pmid: 31871300 (Apr. 1, 2020).

74. Bolton, K. L. *et al.* Cancer Therapy Shapes the Fitness Landscape of Clonal Hematopoiesis. *Nature Genetics* **52,** 1219–1226. ISSN: 1546-1718. pmid: 33106634 (Nov. 2020).

75. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-Cancer Analysis of Whole Genomes. *Nature* **578,** 82–93. ISSN: 1476-4687. pmid: 32025007 (Feb. 2020).

76. Cerami, E. *et al.* The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery* **2,** 401–404. ISSN: 2159-8290. pmid: 22588877 (May 2012).

77. Gao, J. *et al.* Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Science Signaling* **6,** pl1. ISSN: 1937-9145. pmid: 23550210 (Apr. 2, 2013).

78. Adzhubei, I. A. *et al.* A Method and Server for Predicting Damaging Missense Mutations. *Nature Methods* **7,** 248–249. ISSN: 1548-7105. pmid: 20354512 (Apr. 2010).

79. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis* 2nd ed. 2016. 1 p. ISBN: 978-3-319-24277-4 (Springer International Publishing : Imprint: Springer, Cham, 2016).

80. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28,** 235–242. ISSN: 0305-1048. pmid: 10592235 (Jan. 1, 2000).

81. Gangloff, M. *et al.* Crystal Structure of a Mutant hERalpha Ligand-Binding Domain Reveals Key Structural Features for the Mechanism of Partial Agonism. *The Journal of Biological Chemistry* **276,** 15059–15065. ISSN: 0021-9258. pmid: 11278577 (May 4, 2001).

82. Dai, S. Y. *et al.* Prediction of the Tissue-Specificity of Selective Estrogen Receptor Modulators by Using a Single Biochemical Method. *Proceedings of the National Academy of Sciences of the United States of America* **105,** 7171–7176. ISSN: 1091-6490. pmid: 18474858 (May 20, 2008).

83. Goddard, T. D. *et al.* UCSF ChimeraX: Meeting Modern Challenges in Visualization and Analysis. *Protein Science: A Publication of the Protein Society* **27,** 14–25. ISSN: 1469-896X. pmid: 28710774 (Jan. 2018).

84. Shapovalov, M. V. & Dunbrack, R. L. A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. *Structure (London, England: 1993)* **19,** 844–858. ISSN: 1878-4186. pmid: 21645855 (June 8, 2011).

85. Wishart, D. S. *et al.* DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Research* **46,** D1074–D1082. ISSN: 1362-4962. pmid: 29126136 (Jan. 4, 2018).

86. Jurcik, A. *et al.* CAVER Analyst 2.0: Analysis and Visualization of Channels and Tunnels in Protein Structures and Molecular Dynamics Trajectories. *Bioinformatics (Oxford, England)* **34,** 3586–3588. ISSN: 1367-4811. pmid: 29741570 (Oct. 15, 2018).

87. Vavra, O. *et al.* CaverDock: A Molecular Docking-Based Tool to Analyse Ligand Transport through Protein Tunnels and Channels. *Bioinformatics (Oxford, England)* **35,** 4986–4993. ISSN: 1367-4811. pmid: 31077297 (Dec. 1, 2019).

88. Kyte, J. & Doolittle, R. F. A Simple Method for Displaying the Hydropathic Character of a Protein. *Journal of Molecular Biology* **157,** 105–132. ISSN: 0022-2836. pmid: 7108955 (May 5, 1982).

89. Li, H., Ruan, J. & Durbin, R. Mapping Short DNA Sequencing Reads and Calling Variants Using Mapping Quality Scores. *Genome Research* **18,** 1851–1858. ISSN: 1088-9051. pmid: 18714091 (Nov. 2008).

90. Li, R. *et al.* SNP Detection for Massively Parallel Whole-Genome Resequencing. *Genome Research* **19,** 1124–1132. ISSN: 1088-9051. pmid: 19420381 (June 2009).

91. Chang, M. T. *et al.* Accelerating Discovery of Functional Mutant Alleles in Cancer. *Cancer Discovery* **8,** 174–183. ISSN: 2159-8290. pmid: 29247016 (Feb. 2018).

92. *Obtained from COSMIC under the accession number COSM6382421; Accessed on 2021-05-05*

93. *Obtained from COSMIC under the accession number COSM450699; Accessed on 2021-05-05*

94. *Obtained from COSMIC under the accession number COSM4506677; Accessed on 2021-05-05*

95. Herynk, M. H. & Fuqua, S. A. W. Estrogen Receptor Mutations in Human Disease. *Endocrine Reviews* **25,** 869–898. ISSN: 0163-769X. pmid: 15583021 (Dec. 2004).

96. *Obtained from COSMIC under the accession number COSM6926145; Accessed on 2021-05-05*

97. *Obtained from COSMIC under the accession number COSM6936639; Accessed on 2021-05-05*

98. Bálint, M., Jeszenői, N., Horváth, I., Ábrahám, I. M. & Hetényi, C. Dynamic Changes in Binding Interaction Networks of Sex Steroids Establish Their Non-Classical Effects. *Scientific Reports* **7,** 14847. ISSN: 2045-2322. pmid: 29093525 (Nov. 1, 2017).

99. Vitale, S. R. *et al.* TP53 Mutations in Serum Circulating Cell-Free Tumor DNA As Longitudinal Biomarker for High-Grade Serous Ovarian Cancer. *Biomolecules* **10,** E415. ISSN: 2218-273X. pmid: 32156073 (Mar. 7, 2020).

100. Mu, Y., Peng, S., Zhang, A. & Wang, L. Role of Pocket Flexibility in the Modulation of Estrogen Receptor Alpha by Key Residue Arginine 394. *Environmental Toxicology and Chemistry* **30,** 330–336. ISSN: 1552-8618. pmid: 21038436 (Feb. 2011).

101. El-Mowafy, A. M., Abou-Zeid, L. A. & Edafiogho, I. Recognition of Resveratrol by the Human Estrogen Receptor-Alpha: A Molecular Modeling Approach to Understand Its Biological Actions. *Medical Principles and Practice: International Journal of the Kuwait University, Health Science Centre* **11,** 86–92. ISSN: 1011-7571. pmid: 12123109 (2002 Apr-Jun).

102. Sharma, A. *et al.* Antagonists for Constitutively Active Mutant Estrogen Receptors: Insights into the Roles of Antiestrogen-Core and Side-Chain. *ACS chemical biology* **13,** 3374–3384. ISSN: 1554-8937. pmid: 30404440 (Dec. 21, 2018).

103. *Obtained from COSMIC under the accession number COSM7817121; Accessed on 2021-05-05*

104. *Obtained from COSMIC under the accession number COSM3829324; Accessed on 2021-05-05*

105. Mojica, W. D., Mojica, P., Sykes, D., *et al.* Critical Ligand Binding Sequences of the Esr1 Gene: What Role in the Treatment of Er (+) Breast Cancers? *North American Journal of Medicine and Science* **9** (2016).

106. Tate, J. G. *et al.* COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research* **47,** D941–D947. ISSN: 1362-4962. pmid: 30371878 (Jan. 8, 2019).

107. *Chemicalize was used for prediction of chemical properties, chemicalize.com Accessed: 2021-05-05*

108. Zafar, H., Wang, Y., Nakhleh, L., Navin, N. & Chen, K. Monovar: Single-Nucleotide Variant Detection in Single Cells. *Nature Methods* **13,** 505–507. ISSN: 1548-7105. pmid: 27088313 (June 2016).

109. Franken, A. *et al.* Multiparametric Circulating Tumor Cell Analysis to Select Targeted Therapies for Breast Cancer Patients. *Cancers* **13,** 6004. ISSN: 2072-6694. pmid: 34885114 (Nov. 29, 2021).

110. Marass, F., Mouliere, F., Yuan, K., Rosenfeld, N. & Markowetz, F. A Phylogenetic Latent Feature Model for Clonal Deconvolution. *The Annals of Applied Statistics* **10.** ISSN: 1932-6157. https://projecteuclid.org/journals/annals-of-applied-statistics/volume-10/issue-4/A-phyloge

netic-latent-feature-model-for-clonal-deconvolution/10.1214/16-AOAS986.full (2022) (Dec. 1, 2016).

111. Lähnemann, D. *et al.* Eleven Grand Challenges in Single-Cell Data Science. *Genome Biology* **21,** 31. ISSN: 1474-760X. pmid: 32033589 (Feb. 7, 2020).

112. Kadakia, K. T., Beckman, A. L., Ross, J. S. & Krumholz, H. M. Leveraging Open Science to Accelerate Research. *The New England Journal of Medicine* **384,** e61. ISSN: 1533-4406. pmid: 33761227 (Apr. 29, 2021).

113. Riegman, P. H. J., de Jong, B. W. D. & Llombart-Bosch, A. The Organization of European Cancer Institute Pathobiology Working Group and Its Support of European Biobanking Infrastructures for Translational Cancer Research. *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* **19,** 923–926. ISSN: 1538-7755. pmid: 20332270 (Apr. 2010).

114. Van Draanen, J. *et al.* Assessing Researcher Needs for a Virtual Biobank. *Biopreservation and Biobanking* **15,** 203–210. ISSN: 1947-5543. pmid: 27929677 (June 2017).

115. Riegman, P. H. J. *et al.* TuBaFrost 1: Uniting Local Frozen Tumour Banks into a European Network: An Overview. *European Journal of Cancer (Oxford, England: 1990)* **42,** 2678–2683. ISSN: 0959-8049. pmid: 17027254 (Nov. 2006).

116. Narayan, P. *et al.* FDA Approval Summary: Alpelisib Plus Fulvestrant for Patients with HR-positive, HER2-negative, PIK3CA-mutated, Advanced or Metastatic Breast Cancer. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* **27,** 1842–1849. ISSN: 1557-3265. pmid: 33168657 (Apr. 1, 2021).

117. Sultova, E. *et al.* NGS-guided Precision Oncology in Metastatic Breast and Gynecological Cancer: First Experiences at the CCC Munich LMU. *Archives of Gynecology and Obstetrics* **303,** 1331–1345. ISSN: 1432-0711. pmid: 33277683 (May 2021).

118. Le Tourneau, C. *et al.* Molecularly Targeted Therapy Based on Tumour Molecular Profiling versus Conventional Therapy for Advanced Cancer (SHIVA): A Multicentre, Open-Label, Proof-of-Concept, Randomised, Controlled Phase 2 Trial. *The Lancet. Oncology* **16,** 1324–1334. ISSN: 1474-5488. pmid: 26342236 (Oct. 2015).

119. Larson, K. L. *et al.* Clinical Outcomes of Molecular Tumor Boards: A Systematic Review. *JCO precision oncology* **5,** PO.20.00495. ISSN: 2473-4284. pmid: 34632252 (July 2021).

120. Parker, B. A. *et al.* Breast Cancer Experience of the Molecular Tumor Board at the University of California, San Diego Moores Cancer Center. *Journal of Oncology Practice* **11,** 442–449. ISSN: 1935-469X. pmid: 26243651 (Nov. 2015).

121. Millner, L. M., Linder, M. W. & Valdes, R. Circulating Tumor Cells: A Review of Present Methods and the Need to Identify Heterogeneous Phenotypes. *Annals of Clinical and Laboratory Science* **43,** 295–304. ISSN: 1550-8080. pmid: 23884225 (2013).

122. Kwapisz, D. The First Liquid Biopsy Test Approved. Is It a New Era of Mutation Testing for Non-Small Cell Lung Cancer? *Annals of Translational Medicine* **5,** 46. ISSN: 2305-5839. pmid: 28251125 (Feb. 2017).

123. *FDA Approval: Guardant360 CDx* Online; accessed 22-Feb-2022. https://www.accessdata.fda.gov/cdrh_docs/pdf20/P200010S001A.pdf.

124. Joosse, S. A. & Pantel, K. Biologic Challenges in the Detection of Circulating Tumor Cells. *Cancer Research* **73,** 8–11. ISSN: 1538-7445. pmid: 23271724 (Jan. 1, 2013).

125. Taube, J. H. *et al.* Core Epithelial-to-Mesenchymal Transition Interactome Gene-Expression Signature Is Associated with Claudin-Low and Metaplastic Breast Cancer Subtypes. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 15449–15454. ISSN: 1091-6490. pmid: 20713713 (Aug. 31, 2010).

126. De Wit, S. *et al.* EpCAMhigh and EpCAMlow Circulating Tumor Cells in Metastatic Prostate and Breast Cancer Patients. *Oncotarget* **9,** 35705–35716. ISSN: 1949-2553. pmid: 30479699 (Nov. 2, 2018).

127. Scharpenseel, H. *et al.* EGFR and HER3 Expression in Circulating Tumor Cells and Tumor Tissue from Non-Small Cell Lung Cancer Patients. *Scientific Reports* **9,** 7406. ISSN: 2045-2322. pmid: 31092882 (May 15, 2019).

128. Zeune, L. L. *et al.* How to Agree on a CTC: Evaluating the Consensus in Circulating Tumor Cell Scoring. *Cytometry. Part A: The Journal of the International Society for Analytical Cytology* **93,** 1202–1206. ISSN: 1552-4930. pmid: 30246927 (Dec. 2018).

129. Zeune, L. *et al.* Quantifying HER-2 Expression on Circulating Tumor Cells by ACCEPT. *PloS One* **12,** e0186562. ISSN: 1932-6203. pmid: 29084234 (2017).

130. Heitzer, E. *et al.* Complex Tumor Genomes Inferred from Single Circulating Tumor Cells by Array-CGH and next-Generation Sequencing. *Cancer Research* **73,** 2965–2975. ISSN: 1538-7445. pmid: 23471846 (May 15, 2013).

131. Kilgour, E., Rothwell, D. G., Brady, G. & Dive, C. Liquid Biopsy-Based Biomarkers of Treatment Response and Resistance. *Cancer Cell* **37,** 485–495. ISSN: 1878-3686. pmid: 32289272 (Apr. 13, 2020).

132. Carter, L. *et al.* Molecular Analysis of Circulating Tumor Cells Identifies Distinct Copy-Number Profiles in Patients with Chemosensitive and Chemorefractory Small-Cell Lung Cancer. *Nature Medicine* **23,** 114–119. ISSN: 1546-170X. pmid: 27869802 (Jan. 2017).

133. Corcoran, R. B. & Chabner, B. A. Application of Cell-free DNA Analysis to Cancer Treatment. *The New England Journal of Medicine* **379,** 1754–1765. ISSN: 1533-4406. pmid: 30380390 (Nov. 1, 2018).

134. Peneder, P. *et al.* Multimodal Analysis of Cell-Free DNA Whole-Genome Sequencing for Pediatric Cancers with Low Mutational Burden. *Nature Communications* **12,** 3230. ISSN: 2041-1723. pmid: 34050156 (May 28, 2021).

135. Davis, A. A. *et al.* Association of a Novel Circulating Tumor DNA Next-Generating Sequencing Platform with Circulating Tumor Cells (CTCs) and CTC Clusters in Metastatic Breast Cancer. *Breast cancer research: BCR* **21,** 137. ISSN: 1465-542X. pmid: 31801599 (Dec. 4, 2019).

136. Forshew, T. *et al.* Noninvasive Identification and Monitoring of Cancer Mutations by Targeted Deep Sequencing of Plasma DNA. *Science Translational Medicine* **4,** 136ra68. ISSN: 1946-6242. pmid: 22649089 (May 30, 2012).

137. Zhao, C. *et al. TruSight Oncology 500: Enabling Comprehensive Genomic Profiling and Biomarker Reporting with Targeted Sequencing* preprint (Bioinformatics, Oct. 22, 2020). http://biorxiv.org/lookup/doi/10.1101/2020.10.21.349100 (2022).

138. Verhein, K. C., Hariani, G., Hastings, S. B. & Hurban, P. *Abstract 3114: Analytical Validation of Illumina's TruSight Oncology 500 ctDNA Assay* in *Clinical Trials* Proceedings: AACR Annual Meeting 2020; April 27-28, 2020 and June 22-24, 2020; Philadelphia, PA (American Association for Cancer Research, Aug. 15, 2020), 3114–3114. http://cancerres.aacrjournals.org/lookup/doi/10.1158/1538-7445.AM2020-3114 (2022).

139. Liu, H. E. *et al.* Workflow Optimization of Whole Genome Amplification and Targeted Panel Sequencing for CTC Mutation Detection. *NPJ genomic medicine* **2,** 34. ISSN: 2056-7944. pmid: 29263843 (2017).

140. Biezuner, T. *et al.* Comparison of Seven Single Cell Whole Genome Amplification Commercial Kits Using Targeted Sequencing. *Scientific Reports* **11,** 17171. ISSN: 2045-2322. pmid: 34433869 (Aug. 25, 2021).

141. Masunaga, N. *et al.* Highly Sensitive Detection of ESR1 Mutations in Cell-Free DNA from Patients with Metastatic Breast Cancer Using Molecular Barcode Sequencing. *Breast Cancer Research and Treatment* **167,** 49–58. ISSN: 1573-7217. pmid: 28905136 (Jan. 2018).

142. De Luca, G. *et al.* Optimization of a WGA-Free Molecular Tagging-Based NGS Protocol for CTCs Mutational Profiling. *International Journal of Molecular Sciences* **21,** E4364. ISSN: 1422-0067. pmid: 32575430 (June 19, 2020).

143. Fehm, T. N. *et al.* Diagnostic Leukapheresis for CTC Analysis in Breast Cancer Patients: CTC Frequency, Clinical Experiences and Recommendations for Standardized Reporting. *Cytometry. Part A: The Journal of the International Society for Analytical Cytology* **93,** 1213–1219. ISSN: 1552-4930. pmid: 30551262 (Dec. 2018).

144. Carausu, M. *et al.* ESR1 Mutations: A New Biomarker in Breast Cancer. *Expert Review of Molecular Diagnostics* **19,** 599–611. ISSN: 1744-8352. pmid: 31188645 (July 2019).

145. Takeshita, T. *et al.* Clinical Significance of Plasma Cell-Free DNA Mutations in PIK3CA, AKT1, and ESR1 Gene According to Treatment Lines in ER-positive Breast Cancer. *Molecular Cancer* **17,** 67. ISSN: 1476-4598. pmid: 29482551 (Feb. 26, 2018).

146. O'Leary, B. *et al.* Early Circulating Tumor DNA Dynamics and Clonal Selection with Palbociclib and Fulvestrant for Breast Cancer. *Nature Communications* **9,** 896. ISSN: 2041-1723. pmid: 29497091 (Mar. 1, 2018).

147. Guttery, D. S. *et al.* Noninvasive Detection of Activating Estrogen Receptor 1 (ESR1) Mutations in Estrogen Receptor-Positive Metastatic Breast Cancer. *Clinical Chemistry* **61,** 974–982. ISSN: 1530-8561. pmid: 25979954 (July 2015).

148. Clatot, F. *et al.* Kinetics, Prognostic and Predictive Values of ESR1 Circulating Mutations in Metastatic Breast Cancer Patients Progressing on Aromatase Inhibitor. *Oncotarget* **7,** 74448–74459. ISSN: 1949-2553. pmid: 27801670 (Nov. 15, 2016).

149. Yanagawa, T. *et al.* Detection of ESR1 Mutations in Plasma and Tumors from Metastatic Breast Cancer Patients Using Next-Generation Sequencing. *Breast Cancer Research and Treatment* **163,** 231–240. ISSN: 1573-7217. pmid: 28283903 (June 2017).

150. Jeannot, E. *et al.* A Single Droplet Digital PCR for ESR1 Activating Mutations Detection in Plasma. *Oncogene* **39,** 2987–2995. ISSN: 1476-5594. pmid: 32042112 (Apr. 2020).

151. Kaushik, S. *et al.* Impact of the Access Tunnel Engineering on Catalysis Is Strictly Ligand-Specific. *The FEBS journal* **285,** 1456–1476. ISSN: 1742-4658. pmid: 29478278 (Apr. 2018).

152. Biedermannová, L. *et al.* A Single Mutation in a Tunnel to the Active Site Changes the Mechanism and Kinetics of Product Release in Haloalkane Dehalogenase LinB. *The Journal of Biological Chemistry* **287,** 29062–29074. ISSN: 1083-351X. pmid: 22745119 (Aug. 17, 2012).

153. Chaloupková, R. *et al.* Modification of Activity and Specificity of Haloalkane Dehalogenase from Sphingomonas Paucimobilis UT26 by Engineering of Its Entrance Tunnel. *The Journal of Biological Chemistry* **278,** 52622–52628. ISSN: 0021-9258. pmid: 14525993 (Dec. 26, 2003).

154. Oostenbrink, B. C., Pitera, J. W., van Lipzig MM, n., Meerman, J. H. & van Gunsteren WF, n. Simulations of the Estrogen Receptor Ligand-Binding Domain: Affinity of Natural Ligands and Xenoestrogens. *Journal of Medicinal Chemistry* **43,** 4594–4605. ISSN: 0022-2623. pmid: 11101351 (Nov. 30, 2000).

155. Bohacek, R. S. & McMartin, C. Multiple Highly Diverse Structures Complementary to Enzyme Binding Sites: Results of Extensive Application of a de Novo Design Method Incorporating Combinatorial Growth. *Journal of the American Chemical Society* **116,** 5560–5571. ISSN: 0002-7863, 1520-5126. https://pubs.acs.org/doi/abs/10.1021/ja00092a006 (2021) (June 1994).

156. Kangas, E. & Tidor, B. Electrostatic Complementarity at Ligand Binding Sites: Application to Chorismate Mutase. *The Journal of Physical Chemistry B* **105,** 880–888. ISSN: 1520-6106, 1520-5207. https://pubs.acs.org/doi/10.1021/jp003449n (2021) (Feb. 1, 2001).

5 Acknowledgments

I would like to thank everyone who supported me throughout the years of research that went into this work:
Professor Tanja Fehm and Professor Hans Neubauer for their valuable and reliable mentorship.
Professor Günter Niegisch for his regular review and feedback on my work.
Hannah Asperger for her tireless support and company in the laboratory.
André Franken for his day-to-day supervision.
Dorothee Köhler, Nora Hinssen, Dagmar Hohmann and Ellen Honisch for their help in the laboratory.
Jan Kaźmierczak, Torben Zader, Natalie Schneider and Ashton Hunt for proofreading.
Natalie Schneider for her love and her valuable critique of my work.
My family for their unconditional encouragement and support.