University of Natural Resources
and Life Sciences, Vienna

Studying drug resistance and activating mutations in EGFR

MASTER'S THESIS

for the attainment of the academic degree

Diplom-Ingenieurin

submitted by

Anja Wagner, BSc
Matriculation number 01240612

Department of Chemistry, Division of Biochemistry

Supervisor: Univ.Prof. Mag.rer.nat. Dr.rer.nat. Christian Obinger

Co-Supervisor: Dipl.-Ing. Dr. Michael Traxlmayr

Vienna, November 2018

Acknowledgements

First, I want to thank Christian Obinger for the opportunity to do my master's thesis in the Protein Biochemistry group.

Second, I want to thank Michael Traxlmayr who gave me the chance to work on this highly interesting project. His supervision during my thesis not only provided me with a lot of know-how in the lab but also sparked my personal enthusiasm for this topic and scientific research. Thank you for always having an open ear even for the smallest issues.

Additionally, I want to thank the whole BioB for the great work environment. I really enjoyed my time there.

Furthermore, I want to thank Peter Sykacek for the much appreciated help with the analysis of our sequencing data.

I want to express my gratitude to my parents and sister, who always support me in everything I do.

Also, I want to thank Bettina, who kindly offered to read my thesis.

Finally, I want to thank Julian for helping me with so many things regarding my thesis and being there for me no matter what.

Kurzfassung

Der epidermale Wachstumsfaktorrezeptor EGFR ist eine Rezeptor-Tyrosinkinase, die durch zelluläre Signalwege Proliferation induziert und Apoptose reguliert. In einigen Krebsarten ist EGFR konstitutiv aktiviert, was oft von aktivierenden Mutationen, die Ligandenunabhängigkeit des Rezeptors für die Initiierung von Signalwegen herbeiführen, verursacht wird. Die therapeutische Behandlung von Krebs mit dereguliertem EGFR erfolgt mit monoklonalen Antikörpern (z.B. Cetuximab) oder Tyrosinkinase-Inhibitoren (z.B. Erlotinib), deren Wirksamkeit allerdings in den meisten Fällen durch Resistenzmutationen aufgehoben wird. Eine weitere Erschwerung der Therapie stellt die oft fehlende Kenntnis des Effektes einer bestimmten Mutation dar. In dieser Masterarbeit wurde ein *in vitro* Test basierend auf gerichteter Evolution mittels „cell-surface display“ und Durchflusszytometrie verwendet, um EGFR-Mutationen zu charakterisieren. EGFR-Varianten mit einem bestimmen Phänotyp, wie ligandenunabhängige Aktivierung, Cetuximab-Resistenz, Erlotinib-Resistenz oder Verlust der EGFR-Funktion, wurden aus zufallsmutierten EGFR-Bibliotheken selektiert. Diese EGFR-Varianten wurden anschließend mit Illumina „next generation sequencing“ analysiert. Dabei wurde unter anderem festgestellt, dass die Mutation T790M, welche die am häufigsten vorkommende Erlotinib-Resistenzmutation in der Klinik ist, auch in den beiden unabhängigen Selektionen für Erlotinib-Resistenz am stärksten angereichert wurde. Das zeigt, dass es möglich ist, die Resistenz-Mechanismen aus der Klinik mit diesem „high-throughput“ *in vitro* Test zu simulieren. Zusammenfassend wurde in dieser Masterarbeit gezeigt, dass die Selektion von EGFR-Varianten mit aktivierendem beziehungsweise resistenzvermittelndem Phänotyp durchführbar ist, was diese neue Methode zu einem nützlichen Werkzeug macht, um den Einblick in das Mutationsspektrum von EGFR zu vertiefen.

Abstract

Constitutive signaling of the epidermal growth factor receptor (EGFR), a receptor tyrosine kinase regulating cellular proliferation and survival, is associated with various forms of cancer. This aberrant signaling is often caused by activating mutations, i.e. mutations abrogating EGFR's ligand dependency for signaling, and can be targeted with monoclonal antibodies (e.g. cetuximab) or tyrosine kinase inhibitors (e.g. erlotinib). However, the effectiveness of those drugs is often neutralized due to the acquisition of resistance mutations. Another obstacle for the treatment of cancers harboring EGFR mutations is the lacking knowledge about the exact phenotype of certain clinically observed mutations. In the present work an *in vitro* assay utilizing directed evolution by means of mammalian cell display and fluorescence activated cell sorting was used to study EGFR mutations. EGFR libraries created by error-prone PCR were selected for different phenotypes, such as ligand-independent EGFR activation, cetuximab resistance, erlotinib resistance and loss-of-function, i.e. mutations impairing the function of the protein. The obtained libraries were analyzed by Illumina next generation sequencing. This revealed for example that the mutation T790M, which is the most frequent erlotinib resistance mutation in cancer samples, was also the mutation with the highest enrichment in two independent erlotinib resistance selection experiments. This demonstrates that the resistance mechanisms occurring in the clinic can be simulated and studied in this high-throughput *in vitro* assay. In conclusion, this study shows that the selection of EGFR libraries containing potential activating and resistance mutations is feasible and that the assay constitutes a useful tool to further elucidate the mutational landscape of EGFR in the context of cancer.

Contents

CHAPTER 1

Introduction

In 2011 Hanahan and Weinberg published an updated version of their famous paper on the concept of the "hallmarks of cancer" [1]. Those postulated hallmarks comprise "sustaining proliferative signaling", which can be achieved by a variety of factors including alterations of the proliferation signal receptor proteins themselves [1]. One of the receptor proteins facilitating proliferative signaling in mammalian cells is the epidermal growth factor receptor (EGFR, also known as Her1 or ErbB1). EGFR is a transmembrane receptor tyrosine kinase which constitutes the starting point for different downstream signaling cascades finally leading to cellular proliferation and survival [2]. In a healthy, somatic cell those pathways are tightly controlled to maintain the overall homeostasis between tissue growth and depletion. In various types of malignant diseases, however, this homeostasis becomes deregulated in favor of constant cell proliferation, among others caused by aberrant signaling of EGFR and other ErbB family members [2, 3]. This positions EGFR at the foundation of one of the most essential traits of cancer cells and therefore naturally draws the attention of research towards regaining control over the signaling molecule. A quest which turned out to be most challenging.

1.1 EGFR - an overview on structure and function

Over time a remarkable body of research regarding EGFR's structure has accumulated, elaborately reviewed by Kovacs *et al.* [4]. In the following section an overview on those structural insights and the cell signaling pathways affected will be given.

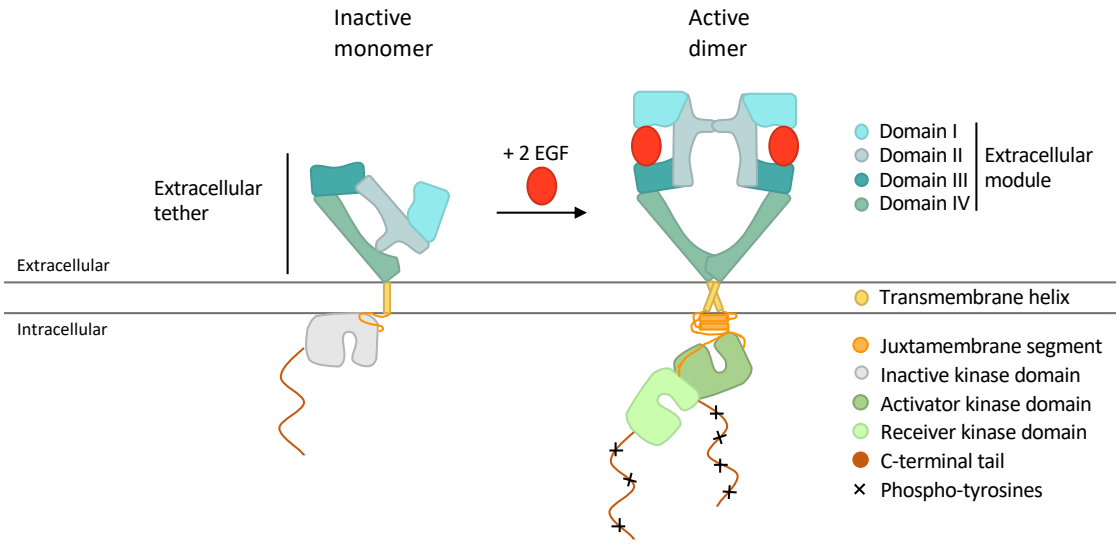

Figure 1.1: Schematic depiction of EGFR's canonical activation mechanism. EGFR's extracellular domain transitions from the inactive monomeric conformation (often referred to as "tether") to the active dimeric form upon binding of a ligand. This ultimately leads to the assembly of the asymmetric kinase domain dimer and phosphorylation of tyrosine residues on the C-terminal tails [4].

EGFR consists of a ligand-binding extracellular module, comprising four distinct domains, connected by a single transmembrane α-helix to its intracellular part, consisting of the juxtamembrane segment, the catalytic kinase domain and the C-terminal tail (Fig. 1.1) [4]. The inactive monomeric state of EGFR is maintained by a "tether" formed by domain II and domain IV of the extracellular module [4]. Domain II also harbours a loop-like "dimerization arm", which is inaccessible in the tethered conformation [4]. Upon binding of a ligand, for example epidermal growth factor (EGF), the extracellular module transitions from the tethered into an upright conformation, exposing the beforehand shielded dimerization arm. Subsequently, EGFR can either

homodimerize or heterodimerize with another receptor tyrosine kinase of the ErbB-family in a "back-to-back" manner [4]. Importantly, the extracellular domain ensures control of the intracellular kinase domain's enzymatic activity by establishing a dependency on extracellular ligand binding, which is necessary to release the conformational autoinhibition [4].

This extracellular association is transmitted through dimerization of the two transmembrane helices via N-terminal GxxxG-like motifs to the intracellular parts of the dimerization partners [5, 6]. The transmembrane helix is followed by the so-called "juxtamembrane segment", which itself consists of two functional entities termed segment A and B (JM-A, JM-B) [7]. In the inactive state, JM-A is likely to interact via its positively charged residues with the negatively charged plasma membrane [8]. In comparison, it is proposed that upon activation the segment forms an antiparallel helical dimer [7]. The subsequent JM-B forms a latch that links the two adjacent kinase domains to form an asymmetric dimer [7]. Thereby the N-lobe of the so-called "receiver" kinase domain interacts with the C-lobe of the "activator" kinase domain [9]. The asymmetric dimer formation is crucial for enzymatic activity, whereby the activator, as the names suggests, induces the active conformation of the receiver kinase domain [9].

In the active state of the kinase domain (Fig. 1.2 (a)) the asparagine of a conserved Asp-Phe-Gly (DFG) motif is oriented towards the catalytic center allowing in combination with an extended activation loop the coordination of Mg-ATP [4]. The so-called αC-helix located in the N-lobe is in close proximity to the ATP-binding site, stabilized by a salt bridge between a lysine and a glutamic acid residue [4]. In the inactive conformation (Fig. 1.2 (b)) this salt bridge is disrupted and the helix turned away from the catalytic cleft.

Upon activation, tyrosine residues on the C-terminal tail and in turn also in the activation loop of the kinase domain are phosporylated [4]. Those phospho-tyrosines serve as docking-sites for different downstream signaling molecules with SH2 and/or PTB domains [4]. Thereby a link between a variety of incoming signals and a cell's nucleus via different signaling pathways is provided. Those signaling pathways initiated by the activation of EGFR include the

RAS/RAF/MEK/ERK as well as the PI3K/AKT/mTOR pathway, leading to proliferation and survival of targeted cells [2]. A detailed view on the EGFR signaling network is for example given by Citri and Yarden [10]. Besides EGF many other ligands, for example transforming growth factor-alpha (TGF-α), are able to activate EGFR in distinct manners [11].

The generally accepted view includes inactive EGFR monomers transitioning into dimers upon activation by ligand binding. However, as summarized by Kovacs *et al.*, different studies suggest that the inactive form of the receptor consists of an equilibrium between the monomeric form and inactive dimers lacking bound ligands [4]. Moreover, a model for receptor multimerization upon activation has been proposed [12], further increasing the functional complexity of EGFR.

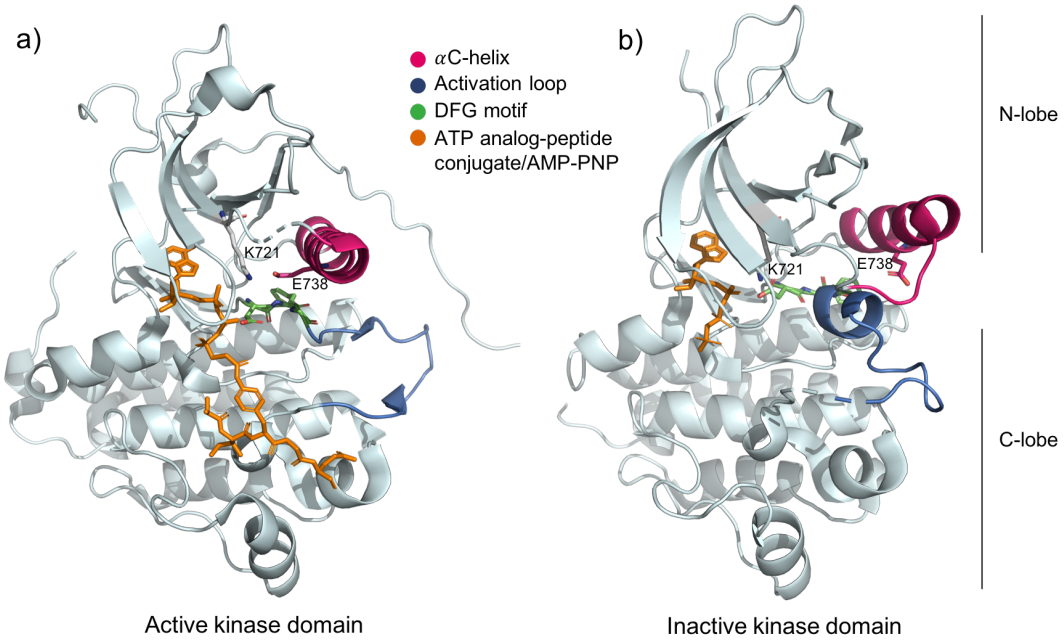

Figure 1.2: EGFR's kinase domain. **(a)** Crystal structure of the kinase domain in the active conformation with an ATP analog-peptide conjugate bound (Protein Data Bank (PDB) 2GS6 [9]). The so-called αC-helix is oriented towards the ATP-binding site which brings the indicated glutamic acid (E738) in position to form a salt bridge with the indicated lysine (K721). The activation loop is in its extended conformation. **b)** Crystal structure of the inactive kinase domain with AMP-PNP bound (PDB 2GS7 [9]). The αC-helix is rotated outward and the ionic contact between K721 and E738 disrupted. The activation loop adopts a helical conformation [4].

1.2 Aberrant EGFR signaling

As initially mentioned, aberrant EGFR signaling plays a pivotal role in many different forms of malignant diseases. In those cancers EGFR is frequently found to be either overexpressed as a result of gene amplification or altered by point mutations, insertions or deletions (indels) [2, 3]. Those point mutations and indels render the receptor signaling ligand-independent and are therefore termed "activating mutations". Consequently, EGFR has become the target of various attempts to inhibit this aberrant signaling and thereby inhibit cell proliferation and ultimately tumor growth.

In the last decades research led to the development and market approval of extracellularly targeting monoclonal antibodies, for example cetuximab and panitumumab, as well as intracellularly targeting small molecule tyrosine kinase inhibitors, like gefitinib and erlotinib [2]. The clinical application of those drugs revealed two somewhat game-changing aspects. First, patients with EGFR containing activating mutations show higher response rates to kinase inhibitors [13, 14] and second, in many cases initially responsive tumors develop resistance to the respective inhibitory drug as summarized by Chong and Jänne [15]. The first aspect emphasizes the emerging importance of personalized medicine based on mutational profiles in cancer treatment. Regarding drug resistance, knowledge about mutations with resistance potential present in naive tumor tissue is certainly valuable to choose a fitting therapy. In the case of secondary resistance mutations acquired by the bulk cell mass upon drug treatment, which is fostered by a high mutational instability inherent to many tumor cells and clonal selection, a race between resistance mechanisms and available treatment options starts. Until now a new drug does not go without a new resistance mutation, resulting to date in the fourth generation of tyrosine kinase inhibitors [16, 17].

Those insights combined with high throughput sequencing led to a large body of mutational data available to researchers and clinicians. These data can be viewed for example in the "Catalogue of Somatic Mutations in Cancer" (COSMIC) provided by the Sanger Institute, which is a curated database of mutations from various sources related to cancer [18]. However, upon a more detailed study of the provided data, it becomes clear that in many cases the exact phenotype of a

mutation, which is required to understand tumor behavior and to adapt drug treatment, is poorly known. In the following section a more detailed view on the findings regarding activating mutations as well as cetuximab resistance and erlotinib resistance mutations, all three being subjects of the present thesis, will be given. In the course of this, also the mechanism of the mentioned inhibitors will be discussed.

1.3 Activating mutations

EGFR requires in its wild type form (EGFR-WT) the binding of a ligand for activation. This interaction triggers a conformational change releasing autoinhibitory, intramolecular associations and promoting the formation of an EGFR dimer which in turn facilitates enzymatic activity [4]. As already mentioned, research unraveled that in different types of cancer certain forms of EGFR are constitutively signaling and largely independent from the availability of a ligand [2, 3]. Those aberrations, referred to as activating mutations, appear either in combination with overexpression or at physiological levels of the receptor [2, 3].

A very common example of a constitutively signaling EGFR mutant is the EGFR variant III (EGFRvIII), which lacks a large part of the extracellular domain encoded by exons 2-7 and is often associated with EGFR overexpression [19]. Another deletion related to aberrant signaling is located in exon 19 of the *EGFR* gene [14]. Besides deletions also point mutations with activating ability are described and are very often found in lung cancer patients [15]. Regarding activating mutations, this thesis focuses on point mutations facilitating signaling in the absence of a suitable ligand.

The most prominent activating point mutation is located in the kinase domain, precisely in the activation loop, and consists of a substitution of the leucine residue at position 858 (throughout this thesis the nomenclature including the N-terminal 24 amino acid membrane-targeting signal peptide is used) by an arginine (L858R) [15]. Interestingly, this point mutation was identified in non-small cell lung cancer (NSCLC) tumor material of patients responsive to the tyrosine kinase inhibitors gefitinib and erlotinib [13, 14]. This led to the assumption that EGFR containing

certain activating mutations is more susceptible to tyrosine kinase inhibitors than EGFR-WT. Later, it was established that the L858R activating mutant is indeed more sensitive to erlotinib inhibition compared to EGFR-WT and also that the L858R mutant's K_M value for its substrate ATP is slightly decreased [20, 21].

In a study published in 2012, the possible activating mechanism of the L858R mutation was studied by long-timescale molecular dynamics simulations [22]. The authors suggested that besides the active and inactive conformation of the kinase domain's N-lobe, a third state exists which is characterized by an intrinsic disorder located especially at the αC-helix [22]. This local disorder is considered to be suppressed in favor of the active conformation by asymmetric dimerization of two kinase domains. Conversely the authors figured that a mutation promoting the active conformation of the kinase domain on the expense of the disordered state leads to increased dimerization [22]. This assumption was supported by different experimental procedures for the L858R mutant which, as the authors stated, is consistent with the finding that *in vivo* the L858R mutant requires and promotes dimerization for activity [22].

Approximately 85% of all mutations in EGFR found in lung cancer patients comprise either the L858R kinase domain mutation or the exon 19 deletion [15]. Furthermore, most of the other mutations found in tumor tissue expressing aberrantly signaling EGFR are located in the receptors kinase domain [15]. However, considering the body of data created in the last years regarding structure and function of EGFR, it is evident that the kinase domain is not the only location where mutations can exhibit an activating effect. Even though those mutations might appear rarely compared to for example L858R, still they can be highly relevant in individual cases.

1.4 Cetuximab - targeting EGFR extracellularly

Monoclonal antibodies (mAbs) were the first attempt to inhibit EGFR signaling as soon as the link between cancer and EGFR was drawn [2]. The chimeric mAb cetuximab, commercially available under the name Erbitux®, gained market authorization by the FDA in 2004 for appli-

cation in colorectal cancer [2].

1.4.1 Mechanism of EGFR inhibition by cetuximab

As already mentioned, upon ligand binding to EGFR's domain III of the extracellular module a shift from the tethered into the upright conformation is induced [4]. This conformational change facilitates accessibility of the dimerization arm which is required for the association with another EGFR monomer forming the "back-to-back" dimer [4]. Cetuximab competes with the EGFR ligand EGF for interaction with domain III [23]. Binding of the antibody consequently prevents binding of EGF and further sterically hinders adoption of the upright conformation of the extracellular domain [23].

1.4.2 Cetuximab resistance

After initiation of treatment with cetuximab, primary response is followed by resistance to the drug in a time frame of months [24]. In a study from 2015 several mutations mostly located in the cetuximab epitope were found *in vivo* as well as *in vitro* after treatment with the antibody [24]. Post-treatment patient biopsies suffering from colorectal cancer revealed three mutations in EGFR, namely R451C, I491M and S492R. In a colorectal cancer cell *in vitro* model acquiring resistance after cetuximab treatment, the I491M point mutation was again enriched, thus supporting the clinical sequencing data. Additionally, two other mutations at position 464 from serine to leucine and position 465 from glycine to arginine were found. Except for R451C, all mutations are located in the cetuximab epitope and showed the potential to disrupt the interaction of the antibody with the receptor [24].

1.5 Tyrosine kinase inhibitors

Besides mAbs, also intracellularly acting small molecule tyrosine kinase inhibitors (TKIs) directed against EGFR are clinically approved for the treatment of cancers expressing aberrantly signaling EGFR [2]. Until now, driven by the recurrent acquisition of resistance against those drugs, the fourth generation of EGFR-directed tyrosine kinase inhibitors is under investigation

[16, 17]. In this thesis the focus lies on the first generation TKI erlotinib, commercially available under the name Tarceva®, as a model for TKI resistance.

1.5.1 Mechanism of EGFR inhibition by erlotinib

Erlotinib exhibits its inhibitory function by directly targeting the catalytic unit of EGFR, the kinase domain. Binding of erlotinib competes with binding of Mg-ATP required for the enzymatic phospho-group transfer to tyrosine residues on the C-terminal tail of EGFR [25]. Interestingly, it has been found that only a subset of patients treated with erlotinib shows a response to the inhibitor [13, 14]. As already mentioned, this effect has been associated with the presence of activating mutations in EGFR's kinase domain, for example L858R and exon 19 deletions [13, 14].

1.5.2 Erlotinib resistance

In the majority of cases erlotinib treatment results in the acquisition of resistance after some time, which on the level of EGFR alterations is mostly caused by the well-described resistance mutation T790M [26]. T790 is referred to as the "gatekeeper residue" due to its location at the entrance of the kinase domain's ATP-binding cleft [27]. Initially, it was assumed that the substitution to the larger methionine blocks the binding of TKIs such as erlotinib [28]. However, it was found that compared to the reversible binding of erlotinib, irreversible inhibitors with a comparable structure are still able to downregulate EGFR kinase activity in the presence of T790M [29]. Considering the increased affinity of the T790M mutant to ATP compared to the erlotinib-sensitive L858R mutant and the fact that erlotinib competes with ATP for binding, it was suggested that a shift of the equilibrium of ATP-bound vs. erlotinib-bound EGFR towards ATP-bound EGFR contributes to the resistance mechanism of the L858R/T790M double mutant [27]. Interestingly, the decreased ATP affinity of the L858R mutant is only restored to wild type level by T790M and no increased ATP affinity of the T790M single mutant could be observed in the EGFR-WT background [27]. Nevertheless, it was reported that the kinase domain activity of the T790M mutant is 5-fold increased compared to EGFR-WT in the absence of EGF [27].

1.6 Loss-of-function mutations

Clearly, there are also mutations which cause the abrogation of EGFR's function in the presence of a suitable ligand by for example disruption of relevant structures and intramolecular contacts. In this thesis those mutations are from now on called "loss-of-function" mutations. Compared to mutations conferring an activating or resistance phenotype, loss-of-function mutations are usually not found by screening of mutational profiles in cancer patients but rather by *in-vitro* approaches aiming to elucidate the molecule's structure to function relation. In the course of experiments regarding the activation mechanism of EGFR's kinase domain, which led to the discovery of the asymmetric dimer, mutations impairing the formation of this asymmetric dimer were discovered [9]. Another study concerned with the function of the intracellular juxtamembrane segment showed that mutations in the C-terminal part of this region, corresponding to JM-B, abrogates the activity of EGFR [30]. There are plenty of other studies revealing loss-of-function mutations, but those studies are mostly covering a distinct region of EGFR. Studying mutations which impair the function of EGFR throughout the whole molecule might on the one hand be valuable to deepen the understanding of its structure to function relation. On the other hand, those insights could be useful to improve inhibitors targeting EGFR by rational design.

1.7 Directed evolution - getting the big picture

Considering the diversity of mutations with unknown phenotypes found in cancer patients collected in the COSMIC, it is clear that a method to study many mutations simultaneously would be highly useful. In a previous work in our laboratory [31], an assay based on mammalian cell display and directed evolution was developed to meet the requirement of functional data for the full mutational profile in EGFR and provide predictive insights on which resistance mutations arise upon treatment with a given inhibitor. The latter possibly facilitates the choice of effective drug combinations. The principle of the assay will be shortly explained in the following and can be viewed in Fig. 1.3.

EGFR gene is randomly mutated by error-prone PCR to create a library containing virtually all

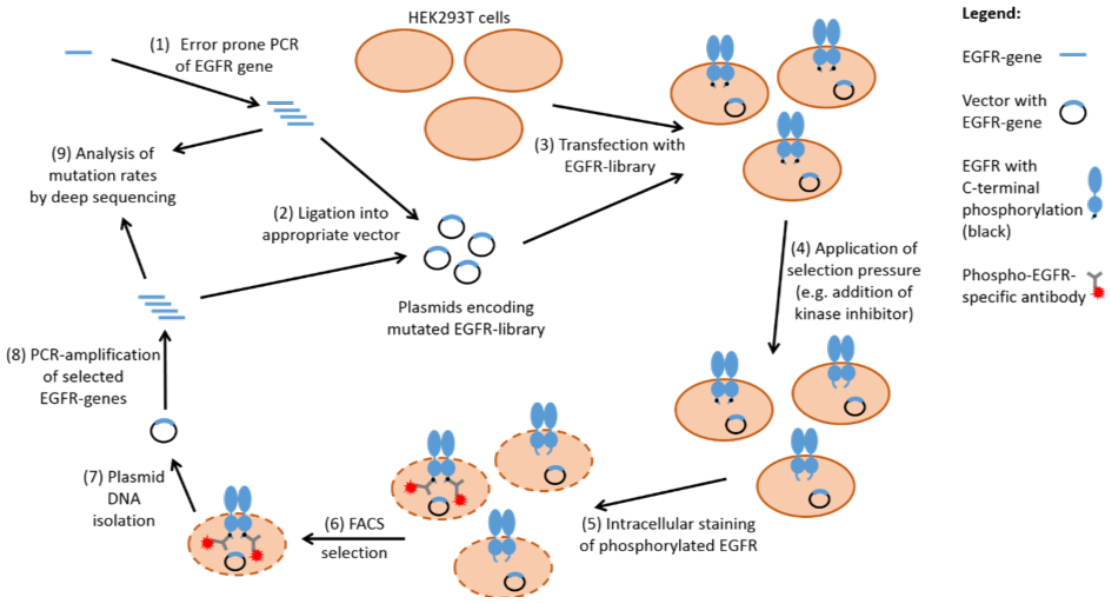

Figure 1.3: Schematic diagram of the previously developed assay to enrich *EGFR* libraries for certain phenotypes. © Michael Traxlmayr

possible single nucleotide *EGFR* variants. The library is then ligated into a vector containing the SV40 origin, allowing for episomal replication. Human embryonic kidney 293T (HEK 293T) cells, which exhibit very low endogenous EGFR expression and are capable of episomal replication, are subsequently transiently transfected with this *EGFR* library. Only very little amounts of plasmid DNA are used to ensure that most transfected cells contain only one plasmid. This is important to avoid that an *EGFR* variant is later co-selected despite the lack of the phenotype selected for. Subsequently, various selection pressures can be applied to the cells, such as incubation with a TKI. HEK 293T cells are afterwards harvested and prepared in a way that allows for intracellular staining with antibodies targeting certain phospho-tyrosines on EGFR as well as the c-myc-tag fused to *EGFR* gene. Thereby, phosphorylation of tyrosine residues on EGFR's C-terminal tail normalized to EGFR expression levels can be directly utilized for detection of EGFR activation. Importantly, this leaves the result of the assay independent of downstream signaling cascades as would be the case with cellular proliferation as readout. HEK 293T cells expressing an EGFR variant exhibiting a desired phenotype, for example erlotinib resistance, are then selected by fluorescence-activated cell sorting (FACS). Afterwards, the plasmid DNA

containing the selected *EGFR* variants is isolated, amplified by PCR, again ligated and propagated in *E. coli* for another selection round. Alternatively, next generation sequencing is used to analyze the mutational profile of the final, selected library.

CHAPTER 2

Aim of the thesis

Aberrant EGFR signaling is one of the driving forces for tumor growth in different types of cancer. In many patients, targeting EGFR with monoclonal antibodies and tyrosine kinase inhibitors results in growth inhibition and tumor regression. Eventually, however, resistance to those inhibitors develops. Efficient targeting of EGFR by a rational combination of drugs based on the mutational profile of the patient requires prior knowledge about the phenotype exhibited by the given mutations. In a previous work, an assay based on directed evolution and mammalian cell display, aiming to gain valuable insights on EGFR mutations, was developed. In the present work, certain aspects of this assay will be further studied and if possible improved. Then the assay will be utilized to study on the one hand activating mutations conferring ligand-independent EGFR signaling and on the other hand resistance mutations to the small molecule tyrosine kinase inhibitor erlotinib and the monoclonal antibody cetuximab. Additionally, it will be attempted to select a library containing loss-of-function mutations, i.e. mutations abrogating the function of EGFR in the presence of ligand. Two selection rounds based on the beforehand explained assay will be conducted for every phenotype mentioned. The resulting libraries of EGFR mutants will be further analyzed by flow cytometry and sequenced by Illumina next generation sequencing to obtain reliable results also for less frequent mutations. Enriched mutations will be examined in available crystal structures to gain a first insight on possible activating and resistance mechanisms.

CHAPTER 3

Materials and Methods

3.1 Cell culture techniques

3.1.1 HEK 293T cell culture

For cultivation of HEK 293T cells, Dulbecco's Modified Eagles Medium (DMEM) – high glucose (Sigma-Aldrich) containing 10% fetal bovine serum (FBS) and 100 U/mL penicillin-streptomycin (both from Gibco®, Thermo Fisher Scientific) was used as complete growth medium. A HEK 293T cell aliquot containing approximately 10^6 cells, frozen at passage 7 and stored in liquid nitrogen, was thawed at room temperature. The whole cell aliquot was pipetted into a 15 mL Falcon tube (Greiner Bio-One) containing 10 mL complete growth medium. The cells were suspended in the medium by inverting the tube for a few times and centrifuged at 1000 rpm and room temperature for 5 minutes in a VWR MegaStar 1.6 R centrifuge. Afterwards, the supernatant was carefully removed and the cell pellet resuspended in 5 mL complete growth medium tempered at 37°C. Subsequently, the cell suspension was transferred into a T25 cell culture flask (Greiner Bio-One) and incubated at 37°C and 5% CO_2 in a Heraeus HERAcell 150 incubator. One day after thawing of cells the medium was exchanged to remove cell debris. Initially, cells were subcultivated after 50-60% confluency was reached.

3.1.2 HEK 293T cell subcultivation

HEK 293T cells were passaged after three and four days of growth, respectively. At this time point cells had normally reached 70-80% confluency. Therefore, the present medium was carefully removed and the cell monolayer was washed with 5 mL Ca^{2+}/Mg^{2+}-free Dulbecco's phosphate buffered saline (PBS) from Gibco®. The cells were detached from the culture flask with 5 mL PBS and resuspended by carefully pipetting up and down several times to disintegrate cell clumps. Depending on the subcultivation ratio, usually 1:10 or 1:20, an aliquot of the cell suspension was transferred into a new T75 cell culture flask (Greiner Bio-One) containing 13.5 mL complete growth medium tempered at 37°C. The cells were suspended in the new medium by gentle shaking of the cell culture flask and incubated at 37°C and 5% CO_2 in a Heraeus HERAcell 150 incubator. Before and after passaging, morphology and confluency of cells were examined under the inverse microscope.

3.1.3 HEK 293T cell cryopreservation

The HEK 293T cell layer at 80% confluency of a T75 cell culture flask, corresponding to approximately 1.9 x 10^7 cells, was detached with PBS and the cell suspension was centrifuged at 1000 rpm and 4°C for 5 minutes. The supernatant was carefully removed and the cell pellet resuspended in 6.5 mL complete growth medium containing 5% dimethylsulfoxide (DMSO) from Sigma-Aldrich by carefully pipetting up and down several times. Immediately, 1 mL aliquots were pipetted into 2.0 mL TPP® cryotubes (Sigma-Aldrich) resulting in approximately 3 x 10^6 cells per tube. The tubes were transferred into an isopropanol filled freezing chamber (Nalgene®), which ensures a cooling-rate of 1°C per minute, and afterwards into a -80°C freezer. For long time storage, HEK293T cells were kept in a liquid nitrogen tank at -196°C.

3.2 Library preparation from sorted plasmids in *E. coli*

3.2.1 Amplification of *EGFR* gene from sorted plasmids by PCR

EGFR gene was amplified from sorted and isolated plasmids via PCR. Conditions for the PCR reactions were chosen based on previous optimization experiments in our lab [31]. Two con-

secutive amplifications, first with 32 cycles and second with 14 cycles (optimized in the present thesis, see Section 3.2.2 and Section 4.1.2), using the Phusion® High-Fidelity DNA polymerase from New England Biolabs® (NEB) were done. The reactions were prepared based on the amounts of reagents for one reaction displayed in Table 3.1. One PCR reaction had a total volume of 50 µL.

Table 3.1: PCR reaction set-up. Amount of primers (EGFR_epPCR_fwd and EGFR_epPCR_rev) and template depends on the number of cycles (*32 cycles for the first PCR/14 cycles for the second PCR).

Reagent	Stock concentration	Concentration in the reaction	Amount per reaction [µL]
ddH_2O			to reach 50 µL in total
Phusion® GC Buffer	5x	1x	10
dNTPs	10 mM each	200 µM each	1
EGFR_epPCR_fwd	10 µM	*0.5 µM/0.075 µM	*2.5/0.375
EGFR_epPCR_rev	10 µM	*0.5 µM/0.075 µM	*2.5/0.375
DMSO	100%	10%	5
Phusion® HF polymerase	2 U/µL	1 U/µL	0.5
Template			*5/0.5

Phusion® GC Buffer, DMSO and deoxynucleotide (dNTP) solution mix were also purchased from NEB. The used primers (EGFR_epPCR_fwd and EGFR_epPCR_rev) were designed in a previous work in our lab [31]. 5 µL of isolated *EGFR* plasmids were used as template for the first amplification, whereas 0.5 µL of the previous PCR product were used for the second amplification. As positive control *EGFR*-WT plasmid was used in different concentrations. For that purpose, a 1 ng/µL solution of *EGFR*-WT plasmid was diluted 1:500 which corresponds to 10 pg in 5 µL. Then, four serial dilutions were made, resulting in 1 pg, 0.1 pg, 0.01 pg and 0.001 pg of *EGFR*-WT plasmid DNA in 5 µL. ddH_2O served as negative control. At first the PCR master mix, containing all reagents except for template DNA, was assembled on ice. Before addition of Phusion® polymerase the master mix was inverted a few times and spun down to achieve an optimal buffer environment for the enzyme, especially considering the high concentration of DMSO in the reaction. 45 µL or 49.5 µL of master mix were pipetted into PCR tubes, followed by addition of 5 or 0.5 µL template DNA, respectively. In case of experiments with many PCR reactions, mixing of master mix and template DNA was also done on ice. Reactions were

mixed by inverting several times and spun down prior to transfer into the Arktik™ thermocycler with preheated lid. Cycling conditions were also based on previously done PCR optimization experiments in our lab [31] and are displayed in Table 3.2.

Table 3.2: PCR cycling conditions. The displayed durations and temperatures were used for all PCR amplifications. Either 32 or 14 cycles were done.

Temperature [°C]	Duration [min:sec]	Cycle number
98	0:30	
98	0:10	
72	2:30	32/14
72	10:00	
4	until removal of tubes	

3.2.2 Testing of different PCR conditions for the second *EGFR* gene amplification

For the second amplification various cycle numbers as well as primer concentrations were tested. PCR reactions were prepared according to Table 3.1 and Section 3.2.1, except for volumes of primers and template DNA. Cycle conditions were chosen as for the first 32 cycle amplification, except for the number of cycles. In Table 3.3 all different combinations of cycle numbers, template DNA volumes as well as primer concentrations are shown. Again, ddH_2O was used as negative control. For the positive control the same volume as for the *EGFR* libraries of a 1 ng/µL *EGFR*-WT plasmid dilution was used.

Table 3.3: PCR optimization. All possible combinations of the depicted parameters were tested.

	Cycle number	Primer concentration [µM]	Template [µL]
Experiment 1	16/20	0.2/0.5	0.5
Experiment 2	16/12	0.05/0.1/0.2	0.5/0.35

3.2.3 TAE-agarose gel electrophoresis

Preparative as well as analytic TAE-agarose gel electrophoresis was used throughout library preparations to remove unspecific PCR products as well as to ensure the presence and correct size of the product. The 1% agarose gels were prepared with Tris-Acetate-EDTA (TAE) buffer (pH 8.3) and LE agarose (Biozym). Furthermore, SYBR™ Safe DNA gel stain (Invitrogen) was added to reach a dilution of 1:10,000 in the gel. TAE buffer (pH 8.3) was also used as running buffer for the electrophoresis. DNA samples were mixed with loading dye (either gel loading dye, purple (6x) from NEB or DNA gel loading dye (6x) from Thermo Fisher Scientific). The GeneRuler 1 kb DNA ladder (Thermo Fisher Scientific) was used as standard. Ladder as well as samples were applied on the gel and the current was set to 120 V for 35 to 50 minutes. Afterwards, DNA bands were visualized with the Gel Doc™ XR+ (BioRad).

3.2.4 Purification of PCR products from TAE-agarose DNA gels

The *EGFR* gene bands were purified from TAE-agarose gels using the GE Healthcare illustra GFX PCR DNA and Gel Band purification kit. Bands were excised from the gel and transferred into weighed Eppendorf tubes. 10 µL capture buffer from the GFX kit was added for every 10 mg of DNA gel. The agarose gel was dissolved in capture buffer by heating to 60°C and shaking with 450 rpm in a thermoblock. The kit was further used according to the manual instructions, except for the wash step which was repeated once to improve removal of components which might interfere with following library preparation steps. Elution was done with cell culture grade water. The concentration of the purified PCR products was measured on a NanoDrop™ 1000 (Thermo Fisher Scientific). Until further processing, samples were stored at -20°C.

3.2.5 Restriction digest of purified *EGFR* gene

For double restriction digest the enzymes XbaI (20 U/µL) and KpnI-HF (20 U/µL) from NEB were used. According to the manufacturer 1 U of enzyme is able to cut 1 µg DNA at 37°C in a reaction volume of 50 µL. Restriction digest reactions were prepared to contain at least a 10-fold excess of enzyme to ensure thorough digestion of DNA samples. Furthermore, the volume of enzyme, which is stored in glycerol, was kept at a maximum of 10% of the total reaction volume

to prevent impaired activity of the enzymes as proposed by the manufacturer. In Table 3.4 an exemplary restriction digest reaction is shown.

Table 3.4: Restriction digest set-up

Reagent	Concentration	Volume [µL]
DNA		75
Cut Smart Buffer	10x	9
XbaI	20 U/µL	3
KpnI-HF	20 U/µL	3

Incubation was carried out for at least two hours at 37°C in a thermoblock without shaking. Instead of heat inactivation, enzymes were removed by preparative TAE-agarose gel electrophoresis as described in Section 3.2.3.

3.2.6 DNA purification with Monarch PCR & DNA Cleanup kit (5 µg)

In the case that the 260 nm to 230 nm ratio of the purified restriction digest samples was below 1.0, indicating the presence of components which might later interfere with ligation, another purification step using the Monarch PCR & DNA Cleanup (5 µg) kit from NEB was performed. According to the manual instructions dsDNA > 2kb (plasmids, gDNA) should be diluted with DNA cleanup binding buffer in a ratio of 2:1 (binding buffer:DNA), whereas dsDNA < 2kb (some amplicons, fragments) 5:1. Since the PCR product is not a plasmid but larger than 2 kb a ratio of 4:1 was chosen in a previous work in our lab [31]. The kit was further used according to the manual instructions. Elution was done with cell culture grade water.

3.2.7 Preparation of the pSF-CMV-SV40 Ori Sbfl vector for ligation

The amplified, digested and purified *EGFR* genes were ligated into a vector under the control of a cytomegalovirus (CMV) promoter. The vector further contained a kanamycin resistance gene and a SV40 origin, which allows for episomal replication in cells expressing the SV40 large T antigen. For the first selection experiment a previously prepared vector was used [31]. For the second selection experiment the pSF-CMV-SV40 Ori Sbfl vector was newly propagated in *E. coli* and restriction digested. For that purpose, an overnight liquid culture of *E. coli* contain-

ing the pSF-CMV-SV40 Ori Sbfl vector was prepared by inoculation of 20 mL of prewarmed LB medium supplemented with 50 µg/mL kanamycin with a previously prepared cryoculture. Overnight incubation was done at 37°C on a shaker with 180 rpm. For isolation of the vector from *E.coli* the Monarch Plasmid MiniPrep Kit (NEB) was used according to manual instructions. Elution was done with 40 µL cell culture grade water. The vector was double digested with XbaI (20 U/µL) and KpnI-HF (20 U/µL) from NEB as previously described. After restriction digest the vector was treated with a 10-fold excess of calf intestinal phosphatase (10 U/µL) from NEB at 37°C without shaking. After incubation for at least one hour the vector was purified by TAE-agarose gel electrophoresis followed by purification with the illustra GFX PCR DNA and Gel Band purification kit (GE Healthcare) as described previously. The prepared vector was stored at -20°C until further use.

3.2.8 Ligation

ElectroLigase® (NEB) was used for ligation of *EGFR* gene into the pSF-CMV-SV40 Ori Sbfl vector. Vector and insert were combined at a molar ratio of 1:1.125, which was tested in previously performed experiments in our lab [31]. All ligation reactions were prepared in a total volume of 11 µL according to the ratios displayed in Table 3.5. Vector without insert served as negative control for the ligation reaction. All components were combined and sufficiently mixed. Incubation was performed for 90 minutes at room temperature, followed by heat inactivation of the enzyme at 65°C for 15 minutes in a thermoblock. For purification of the ligation product, which results in a higher transformation efficiency, the Monarch PCR & DNA Cleanup kit (5 µg) was used as described before in Section 3.2.6. 7 µL cell culture grade water per sample were used to elute the DNA.

Table 3.5: Ligation reaction set-up

Reagent	Amount
ddH2O	to reach 11 µL in total
Insert	30-100 ng
Vector	30-100 ng
ElectroLigase® Reaction buffer	5 µL
ElectroLigase®	1 µL

3.2.9 Transformation of NEB 10-beta electrocompetent *E. coli*

Since a high transformation efficiency with a relatively large plasmid (around 8 kb for *EGFR* gene and vector backbone) is required to cover the *EGFR* library diversity, the 10-beta electrocompetent *E. coli* from NEB were chosen to propagate the library plasmids. *E. coli* vials were thawed for 10 minutes, gently mixed and divided into 45 µL aliquots on ice. 2 µL of previously prepared ligation product (*EGFR* libraries and ligation negative control) were mixed with the cells and transferred into a cooled Gene Pulser®/MicroPulser™ electroporation cuvette with a 0.1 cm gap (BioRad). Electroporation was conducted with the Gene Pulser Xcell™ electroporator with parameters set to 2.0 kV, 200 Ω and 25 µF. Immediately afterwards, 955 µL SOC outgrowth medium (NEB) at 37°C were directly pipetted into the electroporation cuvette and mixed with the cell suspension by carefully pipetting up and down once. The suspension was then transferred into an Eppendorf tube and immediately incubated at 37°C with gentle shaking in a thermoblock. After all transformations were completed, the cell suspension was transferred into 5 mL polystyrene round-bottom tubes (BD Falcon™) and incubated at 37°C and 210 rpm. The total incubation time at 37°C was one hour. After incubation, 1:10, 1:100 and 1:1000 dilutions of the transformed *E. coli* were prepared in LB Medium. For the ligation negative control a 1:10 dilution was prepared. The remaining cell suspension was pipetted into Erlenmeyer flasks with 50 mL LB Medium including 50 µg/mL kanamycin and incubated overnight at 37°C and 180 rpm. In addition, 100 µL of every dilution were distributed on kanamycin-selective (50 µg/mL kanamycin) LB agar plates. The plates were also incubated at 37°C overnight.

Furthermore, it was evaluated if transformation with directly isolated *EGFR*-library plasmids from sorted HEK 293T cells is also feasible, which would be desirable due to the lack of random mutation shuffling during PCR amplification. The same procedure of transformation as described above was used, however transformation was also conducted with 4 µL of *EGFR* library plasmids corresponding to double the volume which is used for transformation with amplified *EGFR* libraries. Additionally, not only dilutions of the transformed cell suspension but also 100 µL of undiluted cell suspension were plated. The remaining cell suspension was centrifuged at 2000 x g and room temperature for 3 minutes, the supernatant poured off and the cell

pellet resuspended in the remaining SOC outgrowth Medium. This concentrated suspension was also spread on selective LB agar plates with 1% kanamycin.

EGFR library plasmids were isolated from transformed *E. coli* overnight cultures with the Monarch Plasmid MiniPrep Kit (NEB). 1.4 mL cell suspension for each *EGFR* library were used for plasmid isolation. The kit was used according to manual instructions. Elution was done with 50 µL cell culture grade water. The plasmid DNA was stored at -20°C until further use.

3.2.10 Preparation of cryocultures

For every *EGFR* library a cryoculture was prepared. For that purpose, *E. coli* overnight culture was mixed with 100% sterile glycerol in 1.5 mL Nalgene™ general long-term storage cryogenic tubes to yield a suspension with 15% glycerol. The tubes were gently inverted a few times and transferred to -80°C for long-term storage.

3.3 FACS and flow cytometric analysis of EGFR libraries

3.3.1 HEK 293T cell seeding for transfection

Depending on the number of cells required for each experiment, HEK 293T cells were either seeded in T75 or T25 cell culture flasks and 6 or 24 well plates (all from Greiner Bio-One) at a confluency of 30%. Table 3.6 shows which cell culture vessels were used for each FACS sort round per library, each control sample and each sample for flow cytometric library comparison experiments.

Table 3.6: Seeding of HEK 293T cells. *For the second sort T25 flasks or multiple 6 well plates were used.

Experiment	Culture vessel [cm^2]	DMEM [mL]	Seeded cell number	Cell number at 100% confluency
1^{st} sort	75	13.5	7.35 x 10^6	2.45 x 10^7
2^{nd} sort	*25	6.5	2.45 x 10^6	8.17 x 10^6
Sort controls	9.6	2.5	9.41 x 10^5	3.14 x 10^6
Library comparison	1.9	0.5	1.86 x 10^5	6.20 x 10^5

At first, one T25 cell culture flask per library was used in case of the second sort round. However, detachment of cells from T25 cell culture flasks yielded more cell debris compared to cells seeded in 6 well plates. Consequently, from then on, cells were seeded in three or four wells of a 6 well plate per library instead of one T25 cell culture flask. For cell seeding the present complete growth medium was removed from subcultured HEK 293T cells and cells were washed with 5 mL PBS. Cells were detached from the culture flask with 5 mL PBS and transferred into a 50 mL Falcon tube (Greiner Bio-One). An aliquot of the cells was used for further subcultivation as described in Materials and Methods 3.1.2. The remaining cell suspension was pipetted into the above mentioned cell culture vessels containing DMEM (37°C) with 10% FBS and importantly without antibiotics, as advised by the manufacturer of the later used transfection reagent, to yield a cell confluency of 30%. Immediately after addition of cell suspension the culture vessel was gently shaken to distribute the cells evenly. The cells were incubated at 37°C and 5% CO_2.

3.3.2 Transfection of HEK 293T cells

24 hours after cells were seeded, transient transfection with different *EGFR*-containing plasmids was carried out. At that point the cell confluency was usually between 50-60%. As transfection reagent the TransIT-X2® from Mirus BIO LLS was used. Furthermore, inert carrier plasmid DNA (pCT-CON2 CD20, usually used for yeast display) was added. The formation of DNA-transfection reagent complexes was carried out in Opti-MEM® I reduced serum medium (Gibco®) as recommended by the manufacturer of the transfection reagent. For the first sort round *EGFR* plasmid DNA was added to reach a final dilution of 1:1500 compared to the suggested 1 µg/µL. For the second sort round a 1:3000 dilution and for all *EGFR* library comparison experiments a 1:1000 dilution was used. The high dilutions of plasmid DNA are necessary to achieve the uptake of only one individual *EGFR* plasmid in most transfected HEK 293T cells. The addition of the carrier plasmid facilitates efficient transfection despite the low *EGFR* plasmid concentrations. Depending on which cell culture vessel was used, transfection mixtures for one flask or well were prepared according to Table 3.7.

Opti-MEM® I reduced serum medium and TransIT-X2® transfection reagent were warmed at room temperature and the transfection reagent was further gently vortexed before use. After-

Table 3.7: Transfection reaction set-up

Reagent	Final DNA dilution	T75	T25	6 well plate	24 well plate
Opti-MEM® I		1.9 mL	633 µL	250 µL	50 µL
EGFR plasmid DNA (1 ng/µL)	1:1000	19 µL	6.33 µL	2.5 µL	0.5 µL
	1:1500	12.7 µL	4.22 µL	1.67 µL	0.33 µL
	1:3000	6.33 µL	2.11 µL	0.83 µL	0.16 µL
pCTCON2-CD20 (1 µg/µL)		19 µL	6.33 µL	2.5 µL	0.5 µL
*Trans*IT-X2®		57 µL	19 µL	7.5 µL	1.5 µL

wards, components were assembled in either 50 mL Falcon or Eppendorf tubes in the above displayed order. After addition of transfection reagent, all components were mixed by gently pipetting up and down a few times and the reactions incubated for 30 minutes to allow formation of DNA-transfection reagent complexes. Following incubation, the transfection mix was carefully pipetted onto the cell layer and the cell culture vessels were gently rocked to ensure optimal distribution. Subsequently, cells were incubated overnight at 37°C and 5% CO_2.

3.3.3 Serum starvation

16 to 18 hours before the transfected HEK 293T cells were prepared for FACS, serum starvation was induced to remove EGFR-activating growth factors like EGF present in the added FBS. This procedure in combination with addition of EGFR inhibitors applied the necessary selection pressure to the EGFR libraries. The present medium was removed from the cell layer and DMEM without any additives was carefully added according to the volumes shown in Table 3.6. In case of experiments requiring the addition of the EGFR inhibitors erlotinib and osimertinib (AZD9291), erlotinib*HCl or AZD9291*HCl were added to achieve a final concentration of 500 nM. Since the EGFR inhibitors are dissolved in DMSO, in case of EGFR library comparison experiments DMSO was also added to cells not receiving any inhibitor to ensure comparability. After medium exchange cells were incubated at 37°C and 5% CO_2.

3.3.4 Preparation of HEK 293T cells for FACS and flow cytometric analysis

After 16 to 18 hours of serum starvation, the present medium was carefully removed from the cell layer. Cells were detached from the used culture vessel and resuspended in PBS. For FACS sort experiments the cell suspension was divided into a certain number of 15 mL Falcon tubes to reach approximately 3.0 x 10^6 cells per tube. For library comparison experiments 6.2 x 10^5 cells were transferred to each 5 mL polystyrene round-bottom tube (BD Falcon™), from now on referred to as FACS tubes. In Table 3.8 PBS volumes for detachment of cells and number of tubes, to which cells from one vessel were divided, are shown.

Table 3.8: Preparation of cells for flow cytometry

Cell culture vessel	PBS	Number of tubes
T75 flask	8 mL	8 x 15 mL Falcon
T25 flask	4 mL	3 x 15 mL Falcon
6 well plate	2 mL/well	1 x 15 mL Falcon
24 well plate	90 µL/well	1 x 5 mL FACS tube

In case that cells from one culture vessel were stained in more than one Falcon tube, the resuspended cells were first transferred to one Falcon tube and centrifuged at 500 x g and 4°C for 5 minutes in a Heraeus Multifuge 1S-R. The supernatant was removed and the cell pellet resuspended by pipetting up and down in a certain volume of PBS to obtain 270 µL of cell suspension per Falcon tube. After resuspension of cells, 270 µL each were transferred into the designated number of Falcon tubes. For library comparison experiments, the cells from one 24 well plate well were stained in one FACS tube, therefore cells were detached and resuspended with 90 µl PBS and the whole suspension was directly transferred into one FACS tube.

3.3.5 EGFR inhibition with cetuximab

A 6 µM cetuximab solution was prepared in PBS supplemented with 1% bovine serum albumin (cold ethanol fraction from Sigma-Aldrich), from now on referred to as PBSA. For all FACS sorts selecting for cetuximab resistance, the cetuximab solution was added to each tube to reach a final concentration of 100 nM cetuximab. After this step, cells were shortly vortexed at intermediate speed and incubated for 20 minutes at 4°C.

3.3.6 Addition of EGF

Following the incubation of cells with cetuximab, a 1 µg/mL EGF dilution was added to samples requiring stimulation with EGF in 15 second intervals to reach a final concentration of 100 ng/mL in each sample. To all samples that did not receive EGF an equal volume of PBSA was added. After addition of EGF or PBSA cells were shortly vortexed at intermediate speed. EGF-treated samples were incubated for exactly 5 minutes at room temperature.

3.3.7 Methanol fixation

After incubation with EGF for 5 minutes, 3 mL for Falcon tubes and 1 mL for FACS tubes, respectively, of ice-cold methanol (100%) were added to all samples in 15 second intervals. Importantly, before and after addition of methanol, cells were vortexed to avoid cell clumps. Cells were subsequently incubated at 4°C for 30 minutes.

3.3.8 Antibody staining

After methanol fixation and subsequent incubation, cells were washed twice with PBSA. At first, PSBA was added to the cells in methanol up to a total volume of 12 mL for Falcon tubes and 4 mL for FACS tubes. Cells were centrifuged at 500 x g for 5 minutes at 4°C. The supernatant was decanted and 12 mL (Falcon tubes) and 4 mL (FACS tubes) PBSA, respectively, were added to the cell pellets. Again, cells were centrifuged at 500 x g for 5 minutes at 4°C. After decanting the supernatant, PBSA was added in a way that after resuspension of the cell pellet, 800 µL or 80 µL of cell suspension could be transferred to new Falcon or FACS tubes, respectively. This transfer step is necessary to ensure that all cells are stained with the same concentration of antibody. A staining master mix containing either a rabbit monoclonal anti-phospho-EGFR antibody targeting the phosphorylated tyrosine at position 998 (clone C24A5) or a rabbit monoclonal anti-phospho-EGFR antibody targeting the phosphorylated tyrosine at position 1092 (clone D7A5) of EGFR (both Cell Signaling Technology®) was prepared in PBSA. Additionally, an anti-c-myc mAB (clone 9E10) conjugated to Alexa Fluor 488 (Thermo Fisher Scientific) was added to the master mix, to detect the c-myc-tag fused to *EGFR*. The master mix was added to each sample to reach a final dilution of 1:800 for all three antibodies. After addition of the

antibody master mix, samples were shortly vortexed and incubated at room temperature for 30 minutes in the dark. This procedure was followed by two wash steps comprising addition of 12 mL or 4 mL, depending on tube volume, of PBSA and subsequent centrifugation at 500 x g and 4°C for 5 minutes. In contrast to the anti-c-myc-tag antibody, which is directly conjugated to the fluorophore Alexa Fluor 488, the phospho-specific antibodies require the addition of a secondary antibody for detection. Therefore, a secondary staining master mix containing a polyclonal anti-rabbit IgG (H+L), F(ab´)2 fragment conjugated to Alexa Fluor 647 (Cell Signaling Technology®) in PBSA was prepared. Before addition of the secondary antibody, cell samples were again transferred into new Falcon/FACS tubes as described before. The secondary antibody master mix was added to reach a final dilution of 1:500. Again, cell samples were shortly vortexed and incubated at room temperature for 30 minutes in the dark. Finally, cells were again washed twice with 12 mL or 4 mL PBSA followed by centrifugation at 500 x g and 4°C for 5 minutes. The supernatant was decanted and the liquid remaining in the tubes was later used to resuspend the cells for FACS or flow cytometric analysis.

3.3.9 FACS and flow cytometric analysis

All FACS experiments were conducted at the St. Anna Children's Cancer Research Institute (CCRI) FACS core unit on a BD FACSAria™ Fusion cell sorter. Flow cytometric analysis was conducted either at the before mentioned FACS core unit on a BD LSR Fortessa™ cell analyzer or at the University of Natural Resources and Life Sciences using a BD FACSCanto™.

3.3.10 Isolation of *EGFR* library plasmids from sorted HEK 293T cells

After FACS sorting of EGFR phosphorylation-positive HEK 293T cells, *EGFR* library plasmids were isolated from cells using the QIAprep Spin Miniprep kit (QIAGEN). The original protocol was modified by addition of Proteinase K to suit the rather unusual application of plasmid isolation from mammalian cells which was established in previous experiments in our lab [31]. At first, 2 x 10^5 non-transfected HEK293T cells, which were treated equally to transfected cells until methanol fixation and the following wash step, were added as carrier cells to the sorted HEK 293T cells. This step is required to obtain a visible cells pellet during the plasmid isola-

tion, which would otherwise be impossible due to the low number of sorted cells. Furthermore, a sample containing only non-transfected cells served as negative control for the isolation and following amplification reactions. After addition of carrier cells, all samples were centrifuged at 1500 x g and 4°C for 5 minutes. The supernatant was carefully removed and cell pellets were re-suspended in 250 µL buffer "P1" provided by the Miniprep kit. The following cell lysis step was extended by the use of Proteinase K and further by addition of pCTCON2-CD20 plasmid serving as carrier DNA. The latter improves the adsorption of *EGFR* plasmid DNA to the Miniprep kit column and also facilitates DNA measurement since the *EGFR* plasmid DNA alone is below the detection level of the used NanoDrop™ 1000 photometer. Therefore, 250 µL buffer "P2" were added to the cell suspension followed by careful inversion of the Eppendorf tube. Next, 0.8 µg of the pCTCON2-CD20 plasmid were added followed by immediate addition of 10 µL Proteinase K (QIAGEN). Eppendorf tubes were inverted six times and immediately transferred into a pre-heated thermoblock at 56°C. Samples were incubated for 10 minutes at 56°C without shaking. Afterwards, samples were cooled at room temperature for 4 minutes followed by neutralization of the lysis buffer using 350 µL of the buffer "N3" and careful inversion of the Eppendorf tubes. Samples were incubated on ice for 5 minutes and finally centrifuged at 18 000 x g and 4°C for 10 minutes. From this point on, the Miniprep kit was used according to the manual instructions. The supernatant was carefully pipetted on the Miniprep columns and samples were centrifuged at 16 000 x g for one minute. Columns were washed twice, at first with 500 µL buffer "PB" and then with 700 µL buffer "PE". After each step, columns were centrifuged at 16 000 x g for one minute and the flow-through was decanted. Elution was performed at 16 000 x g for one minute with cell culture grade water after incubation for 5 minutes at 37°C, which improves elution of DNA from the columns.

3.4 Sanger sequencing

After the second FACS sort round of both experiments various single clones were analyzed to check whether full length *EGFR* was isolated from sorted HEK 293T cells and propagated in *E. coli*. 4 mL LB Medium with 50 µg/mL kanamycin each were inoculated with colonies from the kanamycin-selective LB agar plates of each *EGFR* library and incubated overnight at

37°C and 210 rpm. The *EGFR* plasmids were isolated with the Monarch Plasmid MiniPrep kit (NEB), concentrations measured on the NanoDrop™ 1000 and diluted to yield 12 µL with a DNA concentration of 80 ng/µL. Those samples were sent to the company Microsynth for Sanger sequencing. To gain a first insight on selected mutations as well as knowledge in which structural regions of EGFR mutation hotspots can be found, also 96 well plate Sanger sequencing was performed at Microsynth. Furthermore, Sanger sequencing of *EGFR* libraries was used to check for enrichment during the selection process.

3.5 Next generation sequencing

The twice sorted *EGFR* libraries of both experiments were analyzed by Illumina next generation sequencing at the next generation sequencing core facility of the Vienna Biocenter (VBCF NGS unit, www.vbcf.ac.at). As sequencing platform Illumina's HiSeq2500 system was chosen to obtain approximately 190 million reads choosing the settings 125 bp and paired-end reads. This relatively large number of reads was attempted to ensure also identification of variants which are present at only low frequencies in the selected libraries.

3.5.1 Preparation of *EGFR* libraries for Illumina sequencing

Amplification of *EGFR* libraries by PCR was carried out at the University of Natural Resources and Life Sciences, fluorescence NanoDrop™ measurements and shearing of *EGFR* libraries at the VBCF NGS unit. Except for amplification of *EGFR* libraries by PCR, fluorescence NanoDrop™ measurements and sample fragmentation, all preparations to obtain sequencing ready libraries were performed by staff of the VBCF NGS unit.

3.5.1.1 Amplification by PCR

Sequencing of *EGFR* libraries incorporated into a vector backbone would lead to unnecessary consumption of sequencing capacity since not only *EGFR* but also the vector backbone would be sequenced. Therefore, *EGFR* libraries were initially amplified by PCR using the *EGFR* library-containing plasmids as template. To maximize the *EGFR* product yield as well as minimize

unspecific PCR products, different PCR cycle numbers as well as template DNA amounts were evaluated (data not shown). Finally, all PCRs were conducted according to Tables 3.1 and 3.2 with 14 cycles, a final primer concentration of 0.075 µM each and 5 ng DNA template. To remove as much unspecific PCR product as possible, PCR reactions were separated by a preparative TAE-agarose gel electrophoresis. Subsequently, the *EGFR* bands were purified with the GE Healthcare illustra GFX PCR DNA and Gel Band purification kit as described before.

3.5.1.2 Fluorescent NanoDrop™ measurement

In order to obtain the actual DNA concentration of the *EGFR* library samples to be sequenced, concentrations were not only measured on a regular NanoDrop™ 1000 but also on a NanoDrop™ 3300 fluorospectrometer at the VBCF NGS unit. A picogreen aliquot, obtained from the VBCF NGS unit was diluted 1:200 in Tris-EDTA buffer. 1 µL of undiluted *EGFR* library DNA was mixed with 4 µL Tris-EDTA buffer and 5 µL diluted picogreen solution. The NanoDrop™ 3300 fluorospectrometer was blanked with Tris-EDTA buffer and all samples were measured in triplicate.

3.5.1.3 *EGFR* gene fragmentation

Illumina sequencing requires samples to have a certain length of fragments. Therefore, fragmentation was performed on the S220 focused-ultrasonicator from Covaris. At first, different ultrasonication durations ranging from 15 seconds to 3.5 minutes were tested to achieve the optimal fragment size distribution for the *EGFR* gene. Fragmentation was carried out in microTUBE AFA Fiber Pre-Slit Snap-Cap 6x16 mm tubes obtained from Covaris. One Covaris tube was filled with 110 µL ddH_2O and 2 µL of *EGFR* library DNA were added. Immediately, 2 µL were withdrawn for analysis and pipetted into an Eppendorf tube, marking time point one. Fragmentation duration was set on the Covaris ultrasonicator and the Covaris tube with sample was placed in the sonicator. After every time point, 2 µL of sample were drawn from the Covaris tube for fragment analysis. Based on the fragment size distribution analysis, which was performed by staff of the VBCF NGS Unit, 15 seconds were determined to be the optimal ultrasonication duration for *EGFR*. 4 µL of each *EGFR* library sample, with concentrations between

36.7 ng/µl and 74.7 ng/µL, as determined by Fluorescence NanoDrop™ 3300 measurements, were fragmented.

3.5.2 Analysis of Illumina next generation sequencing data

The analysis of the obtained Illumina next generation sequencing data was performed in collaboration with Peter Sykacek from the University of Natural Resources and Life Sciences. Samples were analyzed based on the Genome Analysis Tool Kit 4 (GATK4) best practice. Briefly, the samples, which were sequenced on one single lane and therefore multiplexed prior to sequencing, were received already demultiplexed from the VBCF NGS unit. Unmapped reads were aligned to the reference using the Burrows-Wheeler Aligner with the BWA-MEM algorithm. Duplicate reads were marked using the MarkDuplicates tool (Picard) and the base quality scores recalibrated. For variant calling Mutect2 in the tumor only mode was used. The frequency of a given mutation was calculated by dividing the allele depth of that mutation by the total allele depth of all four nucleotides at that position.

CHAPTER 4

Results

4.1 Assay optimization

Throughout the selections and library preparations it became evident that some steps in the assay require further optimization to provide satisfying results. Especially the PCR amplification of full-length *EGFR*, consisting of 3,630 basepairs (bps), proved to be a challenge. Additionally, lysis of sorted HEK 293T cells, an essential step to maintain library diversity, was further optimized. Besides that, cloning efficiency of isolated and amplified *EGFR* libraries was evaluated.

4.1.1 Lysis of HEK 293T cells with Proteinase K

The QIAprep Spin Miniprep kit (QIAGEN) combined with Proteinase K was used to isolate *EGFR* plasmid DNA from sorted HEK 293T cells. However, since this method led to varyingly successful isolations (data not shown), it was subject to further optimization. As suggested by the manufacturer of the QIAprep Spin Miniprep kit, Proteinase K was added to the provided lysis buffer. Importantly, the lysis buffer has a pH of approximately 13, which led to the assumption that the activity of the enzyme in the buffer is rather short-lived. Furthermore, the enzyme also requires an elevated temperature of approximately 56°C for optimal activity. Therefore, it was examined whether incubation of the reaction at room temperature before transfer onto a thermoblock, which can occur if a lot of samples are handled simultaneously, impacts the result

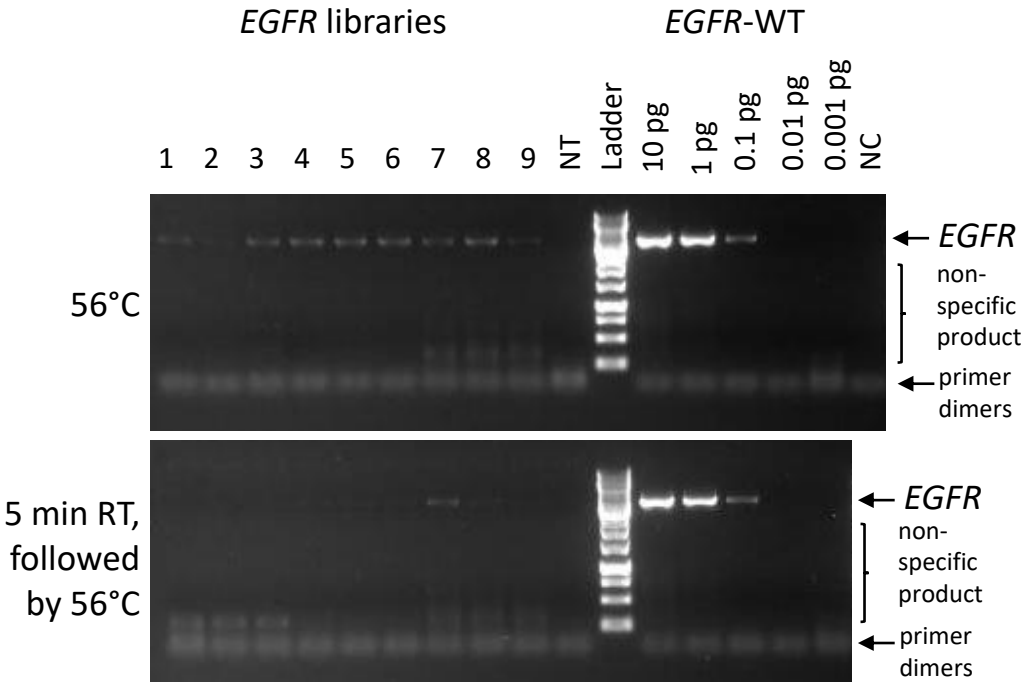

Figure 4.1: DNA isolation and amplification. Upper panel: Immediate transfer to 56°C after addition of Proteinase K. Lower panel: Incubation at room temperature after addition of Proteinase K followed by transfer to 56°C. The isolated *EGFR* libraries were amplified by PCR (*EGFR* libraries, 1-9). Non-transfected HEK 293T cells, that were subject to the same lysis and amplification procedure as the *EGFR* library samples, served as assay negative control (NT). Various amounts of wild type *EGFR*-containing vector used as template served as PCR positive control (*EGFR*-WT), ddH_2O instead of template served as PCR negative control (NC).

of the isolation. Each HEK 293T cell sample was split into two tubes after FACS selection. After the samples were treated with lysis buffer and Proteinase K, they were either incubated for 5 minutes at room temperature followed by incubation at 56°C or immediately transferred to 56°C. After isolation, PCR amplification of *EGFR* was performed and visualized by TAE-agarose gel electrophoresis. Thereby, the result of the isolation was indirectly analyzed. As can be seen in Figure 4.1, DNA bands at the position of *EGFR* gene were only visible for the immediately transferred samples with the exception of lane 7. Compared to that, the positive controls (Fig. 4.1, *EGFR*-WT) had almost the same intensity. This suggests that the difference in *EGFR* amplification product results from the different isolation procedures of the otherwise equally treated samples. Furthermore, in the case of one library also the amount of either not

full-length *EGFR* or non-specific PCR product was higher when incubated at room temperature before transfer to 56°C (Fig. 4.1, lane 1-3, bands slightly above the "primer dimers" band). Consequently, cell samples were immediately transferred to 56°C after addition of Proteinase K in subsequent experiments.

4.1.2 Testing of different PCR conditions for *EGFR* amplification

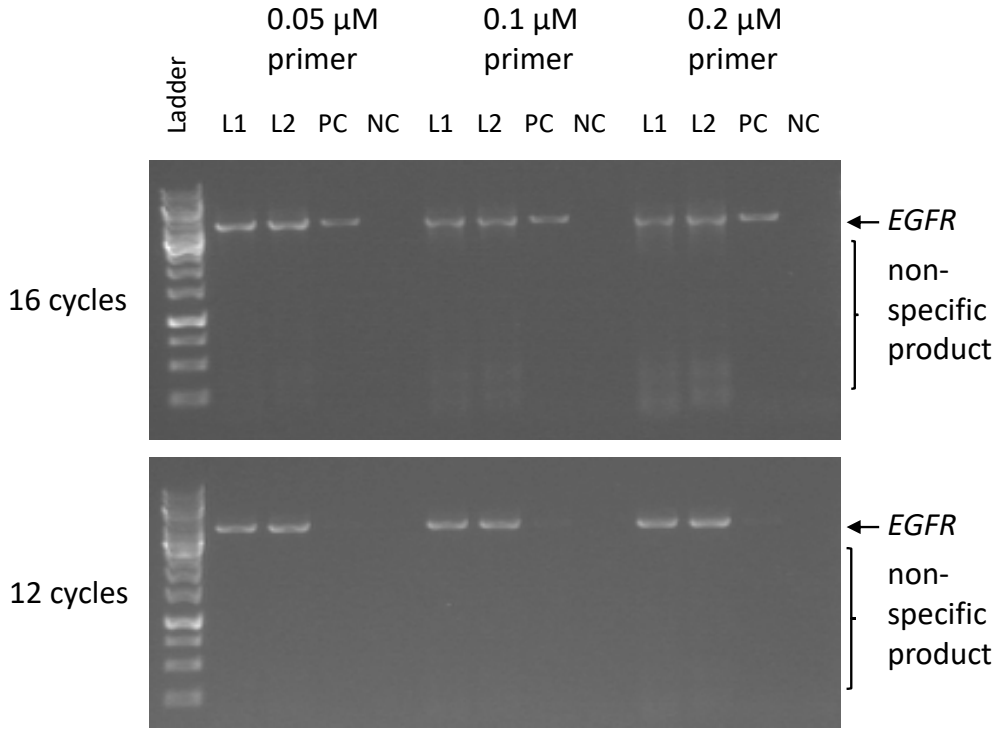

Figure 4.2: PCR optimization. Different PCR cycle numbers and primer concentrations were tested. Two different *EGFR* libraries were used as template (L1, L2). *EGFR*-WT template (0.5 ng) was used as PCR positive control (PC) and ddH_2O instead of template as PCR negative control (NC).

Since plasmid DNA isolation from HEK 293T cells yielded too little amount of DNA for direct transformation of *E. coli*, the selected *EGFR* libraries required initial amplification by PCR. Therefore, a first PCR step with 32 cycles was followed by a second PCR amplification. However, the second PCR of the already once amplified *EGFR* gene generated substantial amounts of non-specific PCR product and was therefore subject to optimization. Varying cycle repeats as

well as PCR primer amounts, as exemplarily depicted in Figure 4.2, were tested. Clearly, it can be seen that with increasing PCR cycles and primer concentrations also the amount of unspecific product increases. Consequently, 14 PCR cycles with a primer concentration of 0.075 µM were chosen for all further second PCR amplifications, which followed the initial 32 cycle PCR of the isolated *EGFR* libraries.

4.1.3 Evaluation of cloning efficiency

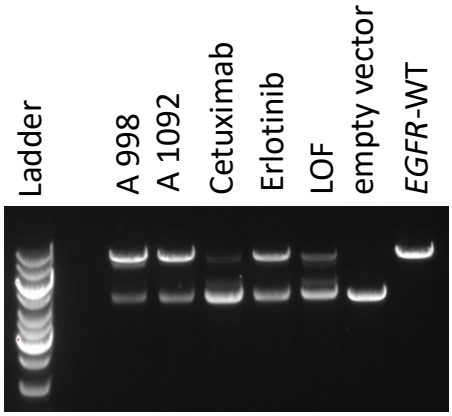

Figure 4.3: Restriction digest of *EGFR* libraries. Plasmids encoding the libraries selected for activating phenotype with either an antibody targeting the phospho-tyrosine at position 998 (A 998) or 1092 (A 1092), cetuximab or erlotinib resistance (Cetuximab, Erlotinib) and loss-of-function phenotype (LOF) were digested with KpnI-HF and subsequently analyzed by agarose gel electrophoresis.

The amplification of *EGFR* from isolated plasmids required re-ligation into an appropriate vector before transformation of *E. coli*. After transfection of HEK 293T cells with the prepared library plasmids, it was observed by flow cytometry that for some libraries the EGFR expression had been relatively low (data not shown). Additionally, upon single clone analysis via Sanger sequencing, many clones contained only the empty vector. Therefore, a single restriction digest with KpnI-HF was performed to assess the proportions of insert-containing vector compared to empty vector in the *EGFR* libraries isolated from transformed *E. coli*. A previously prepared plasmid containing *EGFR*-WT as well as the empty vector were used as controls. Figure 4.3 shows that for all libraries a varying proportion of vector did not contain an insert. In the case

of the library sorted for cetuximab resistance as well as the library sorted for loss-of-function phenotype, the majority of the vectors were empty. Therefore, those libraries were again ligated with a newly prepared vector. Furthermore, it can be seen that especially in the above mentioned libraries also apparently truncated *EGFR* had been incorporated in the vector. In this context, it has to be taken into account that *E. coli* cells containing only the empty vector have a clear growth advantage compared to cells transformed with the insert-containing vector, which is almost twice as long. Certainly, this circumstance shifts the proportions in favor of empty vector and highlights the importance of a high ligation efficiency preceding *E. coli* transformation, which can be examined by this easy and quick method.

4.2 FACS selection of HEK 293T cells expressing EGFR libraries

Fluorescence-activated cell sorting (FACS) was used to select HEK 293T cells expressing EGFR variants exhibiting a certain desired phenotype. In total, two independent selection experiments were conducted to examine the reproducibility of the mutational profile obtained in the selected EGFR libraries for a given phenotype. Each selection experiment comprised two selection rounds, displayed in Figure 1.3.

In the first experiment, an EGFR-WT library was selected for activating mutations and cetuximab resistance. In both cases phosphorylation of the tyrosine residue at amino acid position 998 (pY998) was used as marker for activation of EGFR. Additionally, a library containing the EGFR activating mutation L8585R (EGFR-L858R) was selected for erlotinib resistance. Erlotinib inhibits EGFR containing activating mutations, like L8585R, more potently compared to EGFR-WT [20], and is therefore primarily effective in patients with those mutations [13]. This is the reason for the choice of an EGFR-L858R library instead of the EGFR-WT library for this particular selection. Again, a pY998-specific antibody was used for detection of EGFR activation.

In the second experiment, activating mutations were selected with another antibody targeting the phospho-tyrosine at amino acid position 1092 (pY1092) in addition to the pY998-specific

antibody, which was used in a parallel selection. Thereby, the suitability of both antibodies for the selection of activating mutations and potential differences in the enriched mutations were studied. Furthermore, the erlotinib resistance selection from the first experiment was repeated. In order to potentially deepen the understanding of EGFR's structure and function, an EGFR-WT library was also selected for loss-of-function mutations, that is mutations abrogating the function of EGFR. The erlotinib resistance and loss-of-function selections from the second experiment were conducted with the pY998-specific antibody.

Since the outcome of a selection highly depends on the definition of phosphorylation-positive and phosphorylation-negative cell populations, applied sort gates as well as a numerical overview of the conducted selection experiments will be given below.

4.2.1 FACS gating for different phenotypes

In Figure 4.4 FACS plots of the second selection round for different phenotypes are depicted exemplarily. The gates for the selection of HEK 293T cells expressing pre-selected EGFR libraries as well as positive and negative controls for each individual sort are shown. The x-axis corresponds to the expression of EGFR, detected by an antibody targeting the c-myc-tag fused to the C-terminus of the protein. On the y-axis the phosphorylation of EGFR, which was used as marker for activation, detected with the pY998-specific antibody is displayed. In case of the pY1092-specific antibody, which was also used in the selections for activating mutations, the cell populations appeared to be similarly positioned and therefore only one antibody is shown here. In the absence of any selection pressure, phosphorylation positively correlates with expression of EGFR resulting in a diagonal cell population, as can be seen for the positive controls displayed in the first column of Figure 4.4. Additionally, it is evident that the bulk population of HEK 293T cells was not expressing EGFR, which is due to the fact that only very low concentrations of plasmid DNA were used for transfection, as has been described earlier (see Section 1.7). The sort gates were adjusted using the negative control for every library (Fig. 4.4, second column), which will be described in more detail below.

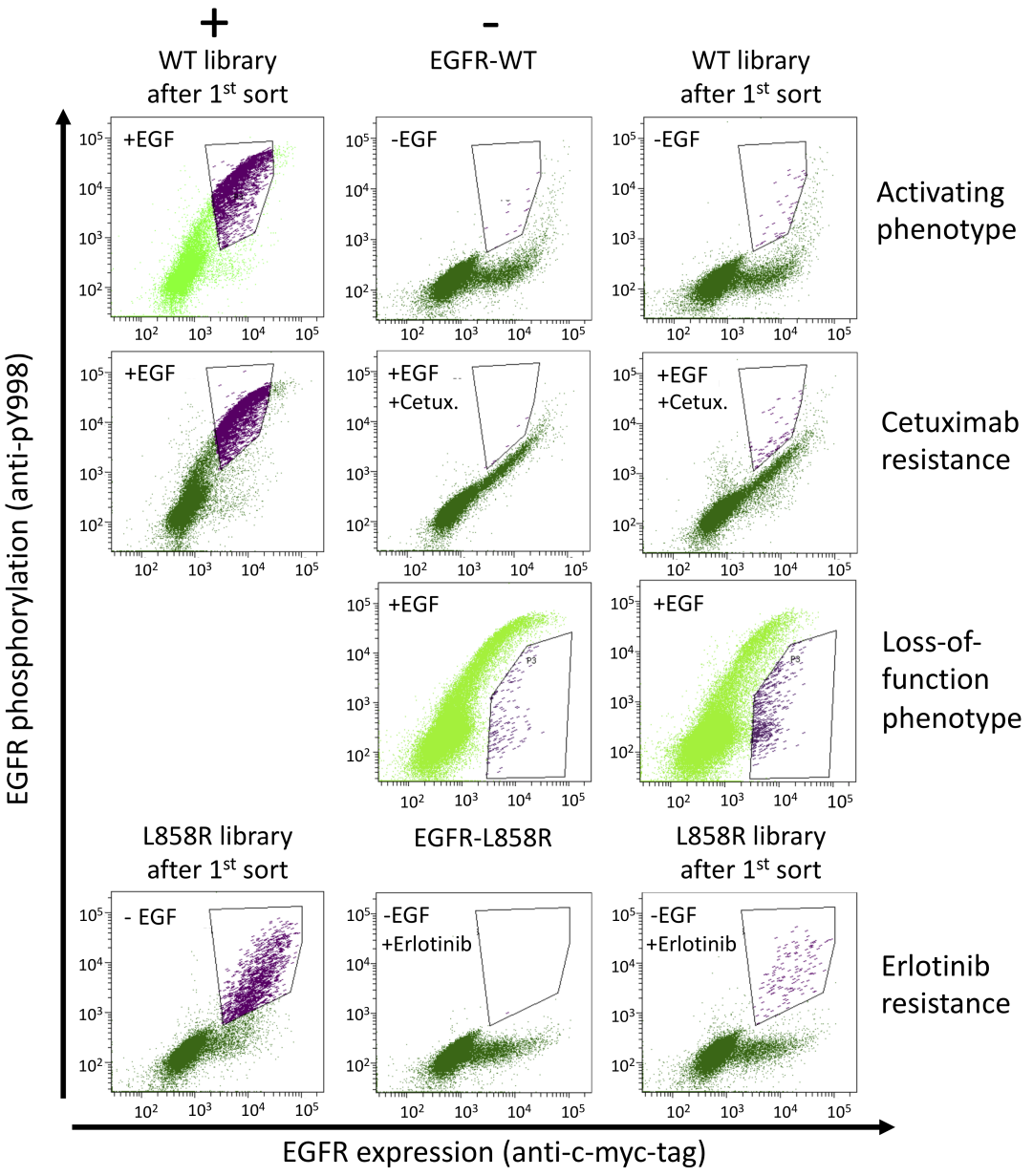

Figure 4.4: Gating strategy in the second selection round for different phenotypes. Transfection of HEK 293T cells with pre-selected EGFR-WT libraries or EGFR-L858R libraries for phenotypes shown on the right was followed by application of selection pressures shown in every dot plot. Libraries without selection pressure served as positive control (+), EGFR-WT or EGFR-L858R with selection pressure as negative control (-). Phosphorylation was detected with a pY998-specific antibody (y-axis), expression with a c-myc-tag-specific antibody (x-axis). 500 nM erlotinib, 100 nM cetuximab and 100 ng/mL EGF, respectively, were applied as indicated. Phosphorylation-positive cells are indicated as dashes, residual cells as dots.

In the case of the selection for activating mutations, where EGFR-WT in the absence of EGF stimulation was used as negative control (Fig. 4.4, first row, middle), it can be seen that only in a certain EGFR expression range EGF is required for phosphorylation. Above this expression level, EGFR phosphorylation was detected even in the absence of EGF. This effect is visible as tail-like population in the plots of the negative control for activating mutations and the activating library itself (Fig. 4.4, first row, middle and right). This EGFR-overexpressing population was deliberately excluded from the sort gate since activation of EGFR might be due to its overexpression in the absence of a mutation. Accordingly, mutations which lead to overexpression of EGFR and exhibit no activating phenotype at normal expression levels are not selected for in this assay. Furthermore, it is evident that the appearance of the phosphorylation-negative population was rather diffuse in the case of this selection compared to other negative controls. This made the adjustment of the selection gate more difficult as will be discussed in Chapter 5.

The plots for the selection of cetuximab-resistant EGFR in the presence of EGF and cetuximab show that the diagonal cell population shifted towards the bottom if the cells had been pre-incubated with cetuximab before addition of EGF. Compared to the selection for activating mutations, the phosphorylation-negative population was more defined, which allowed for the sort gate to be positioned directly above the diagonal (Fig. 4.4, second row).

Importantly, compared to all other selections the erlotinib resistance selection was based on the EGFR-L858R library. As expected, this library exhibited an activating phenotype in the absence of EGF (Fig. 4.4, bottom row, left) and was highly susceptible to inhibition by erlotinib (Fig. 4.4, bottom row, middle). In this case, even high expression levels did not markedly impact the level of inhibition as was observed for other selections.

For the selection of mutations conferring a loss-of-function phenotype, i.e. the absence of phosphorylation in the presence of EGF stimulation, EGFR-WT treated with EGF was used as control to adjust the gate (Fig. 4.4, third row, middle). Accordingly, the gate was positioned below the diagonal of the main cell population, i.e. in the area of cells which were expressing EGFR that

is not activated despite the presence of its ligand EGF. In this case no positive control, which could be a mutation abrogating the function of EGFR, was available.

4.2.2 FACS selections in terms of numbers

In Table 4.1 numerical values for all selection experiments are shown. The library diversities of the EGFR-WT library as well as the EGFR-L858R library, which resemble the initial, unselected libraries created by error-prone PCR, were obtained from a previous work in our lab [31]. Importantly, all further library diversities correspond to the number of sorted HEK 293T cells for each selection round. As described before, the c-myc-tag fused to EGFR allows to consider only cells which are actually expressing EGFR for detection of phosphorylation and selection. The number of the analyzed c-myc-tag positive cells compared to the library diversity was used to calculate the coverage of the library diversity. In the case of the first sort of the initial libraries, full coverage of the library diversity was not given. However, if it is considered that *EGFR* gene consists of 3,630 nucleotides, each of which can be mutated to 3 other nucleotides, theoretically only 10,890 variants would be required to cover the maximum diversity, if only considering single nucleotide variants. In practice, single coverage is not sufficient and at least a 10-fold coverage of the maximum theoretical diversity should be reached, which was achieved for every starting library. For example, in the case of the first experiment and selection for activating mutations with 5.6 x 10^5 analyzed EGFR-expressing cells, the theoretical single nucleotide diversity was covered 51-fold. Since the total number of cells was divided into a variable number of FACS tubes for analysis, the value for "sorted cells" represents an average of all tubes.

4.3 Analysis of final libraries

After both selection experiments, flow cytometric analysis, Sanger sequencing as well as Illumina next generation sequencing were performed to evaluate the final libraries that had been sorted twice.

Table 4.1: Numerical overview of FACS selections. * The term "library diversity" in the case of the initial, unsorted EGFR-WT library and EGFR-L858R library refers to the number of *E. coli* clones obtained after error-prone PCR and subsequent ligation. In the case of the sorted libraries (after 1st sort and after 2nd sort) the term refers to the number of sorted HEK 293T cells and not the number of individual *E. coli* clones obtained after ligation.

	Selection	Library	Library diversity*	Analyzed c-myc-positive cells	Sorted cells (% of c-myc-positive)	Library diversity coverage (-fold)	Theoretical single nucleotide diversity (10 890 variants) coverage (-fold)
1st experiment	**Activating mutations (pY998)**	EGFR-WT lib.	7.6×10^5	5.6×10^5	0.4	0.7	51
		after 1st sort	2.0×10^3	9.9×10^4	1.5	49	
		after 2nd sort	1.5×10^3				
	Cetuximab resistance	EGFR-WT lib.	7.6×10^5	4.9×10^5	1.4	0.6	45
		after 1st sort	6.6×10^3	9.4×10^4	2.2	14	
		after 2nd sort	2.1×10^3				
	Erlotinib resistance	EGFR-L858R lib.	1.6×10^6	3.3×10^5	0.5	0.2	30
		after 1st sort	1.6×10^3	8.1×10^4	5.6	51	
		after 2nd sort	4.6×10^3				
2nd experiment	**Activating mutations (pY998)**	EGFR-WT lib.	7.6×10^5	5.2×10^5	0.5	0.7	48
		after 1st sort	2.5×10^3	8.8×10^4	2.4	35	
		after 2nd sort	2.1×10^3				
	Activating mutations (pY1092)	EGFR-WT lib.	7.6×10^5	3.0×10^5	0.9	0.4	28
		after 1st sort	2.7×10^3	1.4×10^5	3.5	51	
		after 2nd sort	4.7×10^3				
	Erlotinib resistance	EGFR-L858R lib.	1.6×10^6	3.0×10^5	0.2	0.2	27
		after 1st sort	6.2×10^2	8.1×10^4	10.3	130	
		after 2nd sort	8.3×10^3				
	Loss-of-function mutations	EGFR-WT lib.	7.6×10^5	2.9×10^5	6.1	0.4	27
		after 1st sort	1.8×10^4	1.0×10^4	14.6	6	
		after 2nd sort	1.5×10^4				

4.3.1 Flow cytometric analysis

After each selection experiment the obtained libraries were analyzed by flow cytometry in a single experiment. Libraries sorted once or twice were compared with the unsorted library on which the respective selection was based on as well as EGFR-WT or the EGFR-L858R mutant, respectively (Fig. 4.5 and 4.6). Furthermore, cross-reactivity of twice sorted libraries was tested by application of selection pressures other than the respective library was selected for (Fig. 4.7 and Fig. 4.8).

In general, compared to the flow cytometric analysis of the second selection experiment, approximately 10-times less cells were analyzed in the first experiment, with roughly 2,000 single cells per sample. This circumstance renders the results from the first experiment less reliable and highlights the importance of a sufficient number of cells for those flow cytometric comparisons. This applies especially in the case of libraries with comparably few phosphorylation-positive cells, like the ones selected for activating mutations and cetuximab resistance mutations. As explained before, the bulk population of analyzed cells did not express EGFR. Furthermore, EGFR activated by means of overexpression was excluded from the gate for phosphorylation-positive cells. The diagrams in Figure 4.5 (b), 4.6 (b) and 4.7 (b) show the percentage of cells with activated EGFR in relation to the percentage of c-myc-positive, therefore EGFR-expressing, single cells per library. In Figure 4.8 those percentages are directly displayed in the dot plots. It should be noted that the percentages of phospho-EGFR-positive cells is highly dependent on the position and shape of the gates. Therefore, caution is required when comparing those percentages between the different experiments (Fig. 4.5 and 4.6) or between the different rows (representing different phenotypes) within a figure. The necessary analysis of all libraries obtained in both selection experiments in one single flow cytometry experiment will be subject to future research.

4.3.1.1 Enrichment of activating and resistance phenotypes

In Figure 4.5 (a) first row and 4.6 (a) first and second row, the flow cytometry plots of all libraries selected for activating mutations, i.e. phosphorylation in the absence of the ligand EGF, are shown. Phosphorylation was detected either by the pY998-specific or the pY1092-specific

antibody. As already mentioned, the use of two different antibodies for detection of phosphotyrosines enabled the comparison of their suitability for selection experiments as well as differences in terms of enriched mutations. The flow cytometry gate corresponding to cells expressing activated EGFR was adjusted using the respective negative control, namely EGFR-WT without EGF present. In the first experiment, the already once selected library resulted in slightly less activated EGFR-expressing cells than the unselected library (Fig. 4.5 (b) left). This is most likely due to the small number of analyzed cells and thereby clearly shows the limitations of this particular experiment. Detection of activated EGFR by means of the pY1092-specific antibody was not done in the first experiment. In the second selection experiment, the events in the gate for cells expressing activated EGFR increased from 0.1% for EGFR-WT to 3.0% for the EGFR-WT library after two selection rounds with the pY1092-specific antibody (Fig. 4.6 (b) second diagram from left). In comparison, the library selected twice for activating phenotype using the pY998-specific antibody resulted in 0.9% of cells expressing activated EGFR, compared to 0.1% for EGFR-WT (Fig. 4.6 (b) first diagram from left). This led to the assumption that the pY1092-specific antibody is more suitable for the selection of EGFR activating mutations, which will be further addressed in Section 5.2.

Selection of a cetuximab resistance library was done in only one experiment and is shown in Figure 4.5 (a) second row. The final library sorted twice for cetuximab resistance led to 5.9% of EGFR-expressing and cetuximab-resistant cells upon transfection. In comparison, EGFR-WT resulted in 0.7% of cells present in the cetuximab resistance gate (Fig. 4.5 (b) middle).

Among the final libraries for activating mutations or inhibitor resistance, the highest number of phosphorylation-positive cells was achieved by selection for erlotinib resistance (Fig. 4.5 (a) third row and 4.6 (a) fourth row). In both experiments a substantial fraction of the EGFR-expressing cell population shifted towards the phosphorylation-positive gate. In the first selection experiment 30.6% of the EGFR-expressing cells were present in the phosphorylation-positive gate and in the second selection experiment 21.1% were reached (Fig. 4.5 (b) right and Fig. 4.6 (b) second from right).

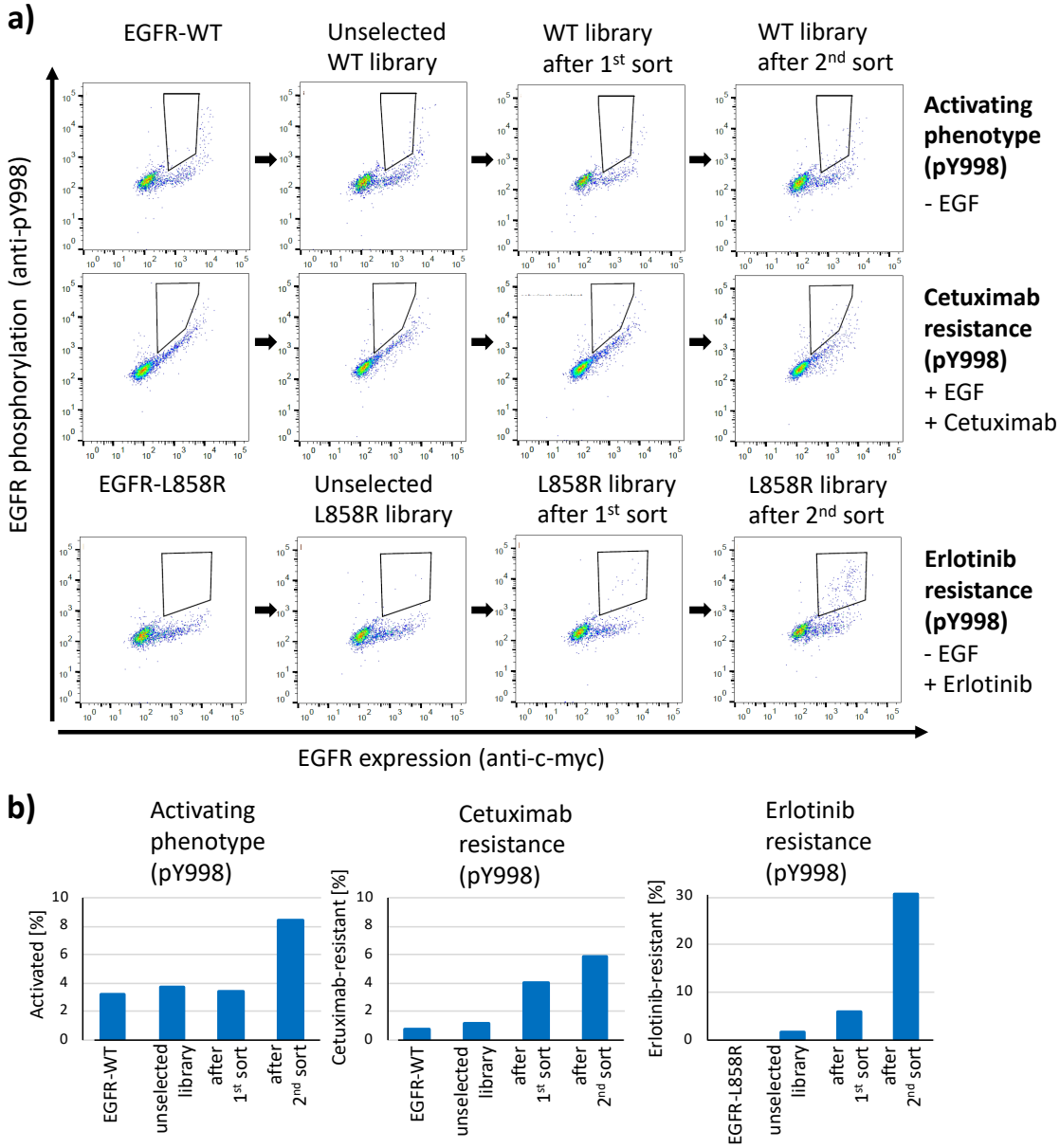

Figure 4.5: Libraries selected in the first experiment. **(a)** Dot plots of HEK 293T cells transfected with EGFR-WT/L858R, unselected libraries, once and twice sorted libraries for different phenotypes followed by application of selection pressures as indicated on the right. Phosphorylation was detected with a pY998-specific antibody (y-axis), expression with a c-myc-tag-directed antibody (x-axis). 500 nM erlotinib, 100 nM cetuximab and 100 ng/mL EGF, respectively, were applied as indicated. **(b)** Percentages of phosphorylation-positive HEK 293T cells normalized by percentages of EGFR-expressing HEK 293T cells conferred by activating, cetuximab resistance and erlotinib resistance libraries.

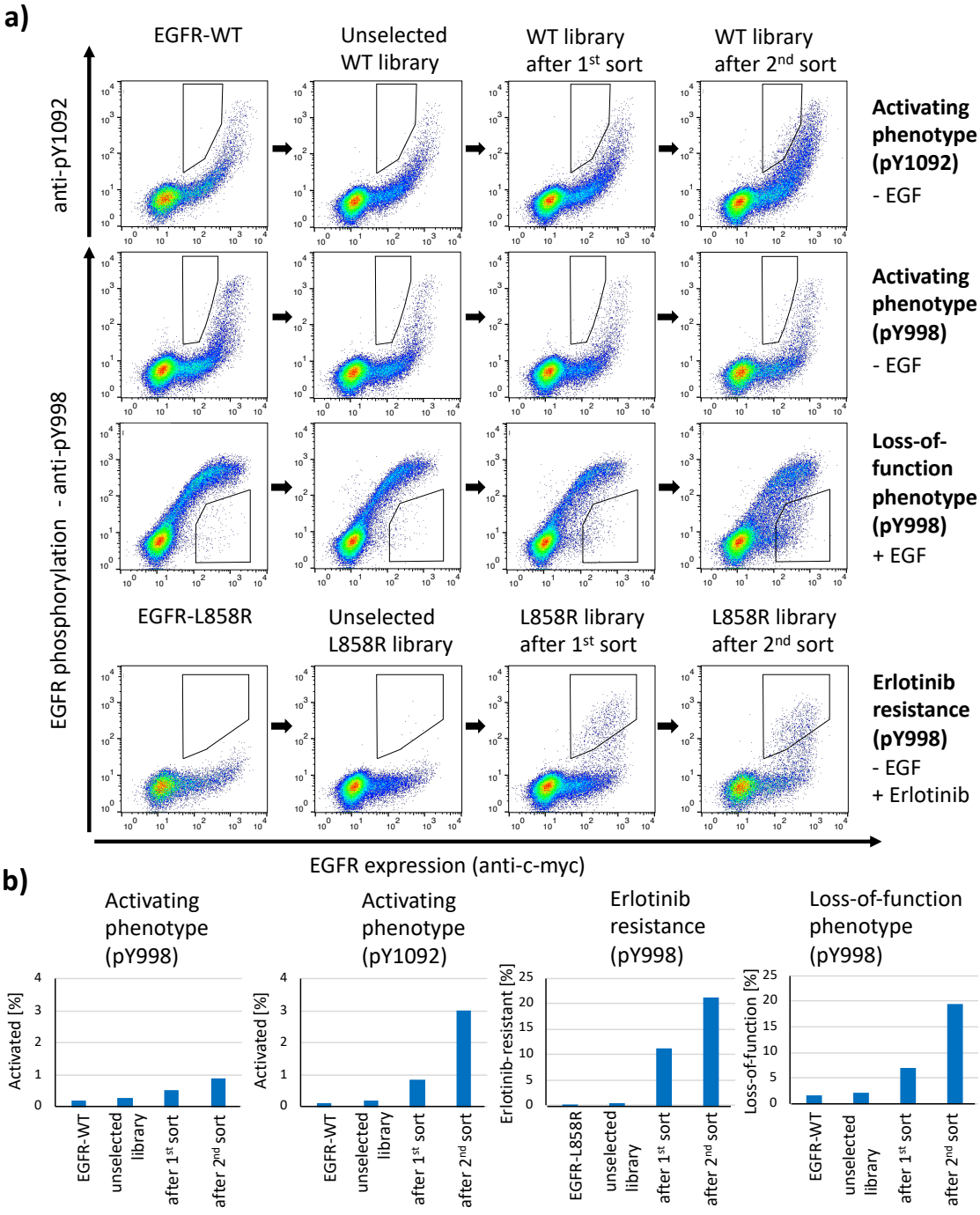
a)
EGFR-WT
Unselected WT library
WT library after 1st sort
WT library after 2nd sort
anti-pY1092
Activating phenotype (pY1092)
- EGF
Activating phenotype (pY998)
- EGF
Loss-of-function phenotype (pY998)
+ EGF
EGFR-L858R
Unselected L858R library
L858R library after 1st sort
L858R library after 2nd sort
Erlotinib resistance (pY998)
- EGF
+ Erlotinib
EGFR phosphorylation - anti-pY998
EGFR expression (anti-c-myc)
b)
Activating phenotype (pY998)
Activating phenotype (pY1092)
Erlotinib resistance (pY998)
Loss-of-function phenotype (pY998)
Activated [%]
Erlotinib-resistant [%]
Loss-of-function [%]
EGFR-WT
unselected library
after 1st sort
after 2nd sort
EGFR-L858R

Figure 4.6: Libraries selected in the second experiment. (a) Dot plots of HEK 293T cells transfected with EGFR-WT/L858R, unselected libraries, once and twice sorted libraries for different phenotypes followed by application of selection pressures as indicated on the right. Phosphorylation was detected with a pY998/pY1092-specific antibody (y-axis), expression with a c-myc-tag-directed antibody (x-axis). 500 nM erlotinib or 100 ng/mL EGF were applied as indicated. **(b)** Percentages of phosphorylation-positive/negative (LOF) HEK 293T cells normalized by percentages of EGFR-expressing HEK 293T cells conferred by activating pY998, activating pY1092, erlotinib resistance and loss-of-function libraries.

In the case of selection for loss-of-function mutations, a substantial fraction of the diagonal cell population clearly shifted towards the lower right quadrant of the dot plot (Fig. 4.6 (a) third row). The twice sorted library gave rise to approximately 19.4% of cells expressing EGFR that is inactive in the presence of EGF (Fig. 4.6 (b) right).

4.3.1.2 Cross-reactivity

Potential cross-reactivity of selected libraries was examined by application of different selection pressures (Fig. 4.7 and Fig. 4.8). Additionally, differences between the use of either pY998- or pY1092-directed antibodies for selection of activating mutations were further elucidated, as will be shown in the following. In this case, all flow cytometric data originate from the same experiment with the exception of the cetuximab resistance library tested for activating phenotype and vice versa, which has to be considered separately (Fig. 4.8).

The first and second row of Figure 4.7 (a) show the comparison of pY998- and pY1092-directed antibodies in the context of the activating phenotype. In the case of the antibody targeting pY1092 for detection of activation (Fig. 4.7 (a) first row and (b)), the library which was in the first place selected for activating mutations using the same antibody (pY1092) yielded more cells expressing activated EGFR in the absence of EGF than the library selected with a different antibody targeting pY998. However, less expected, in the case of detection with the pY998-specific antibody (Fig. 4.7 (a) second row and (b)), again the library selected by pY1092 detection gave rise to a higher percentage of phosphorylation-positive cells. In fact, irrespective of the detection antibody, both libraries resulted in the same percentage of cells expressing activated EGFR,

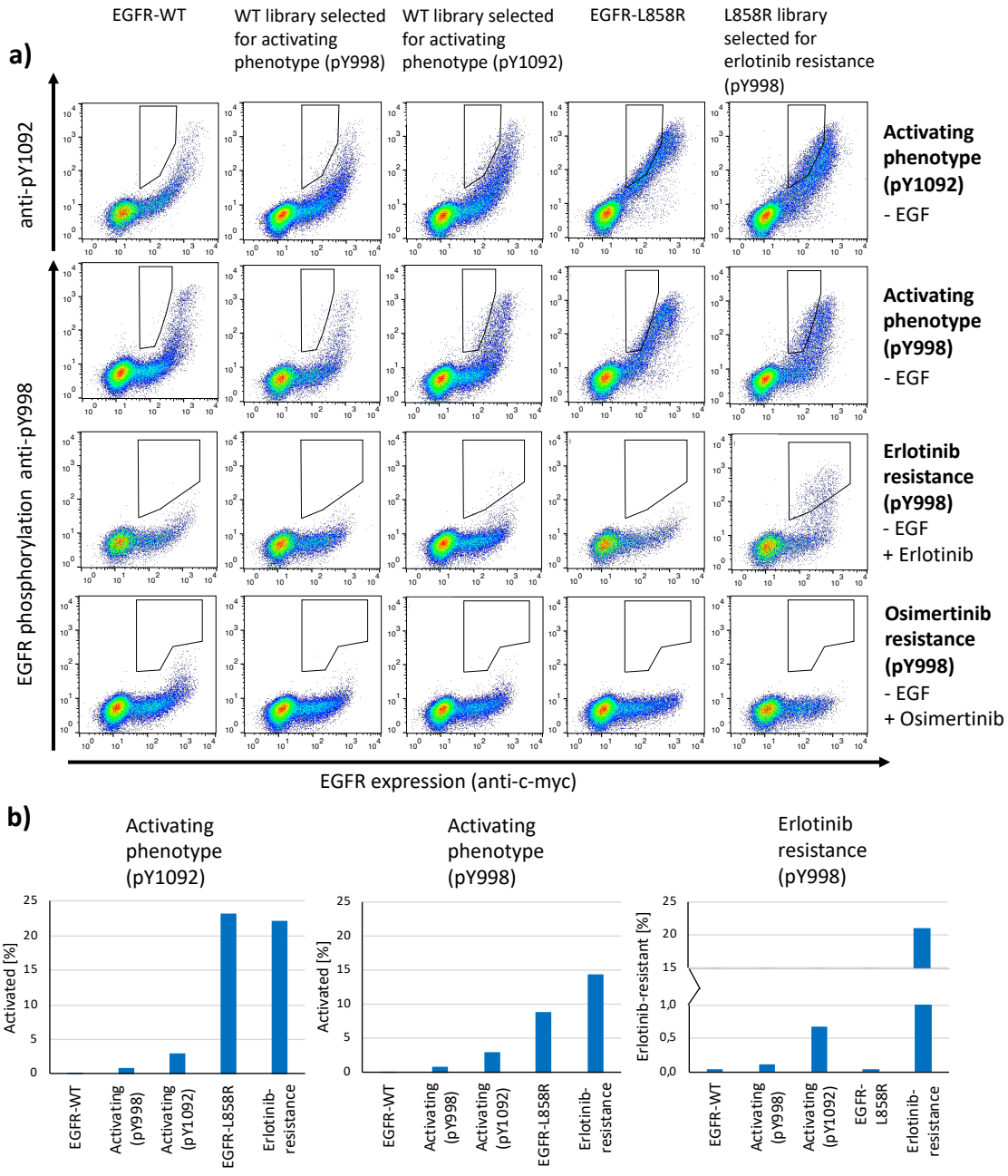
a)
EGFR-WT
WT library selected for activating phenotype (pY998)
WT library selected for activating phenotype (pY1092)
EGFR-L858R
L858R library selected for erlotinib resistance (pY998)
anti-pY1092
EGFR phosphorylation anti-pY998
Activating phenotype (pY1092) - EGF
Activating phenotype (pY998) - EGF
Erlotinib resistance (pY998) - EGF + Erlotinib
Osimertinib resistance (pY998) - EGF + Osimertinib
EGFR expression (anti-c-myc)
b)
Activating phenotype (pY1092)
Activating phenotype (pY998)
Erlotinib resistance (pY998)
Activated [%]
Erlotinib-resistant [%]
EGFR-WT
Activating (pY998)
Activating (pY1092)
EGFR-L858R
Erlotinib-resistance

Figure 4.7: Cross-reactivity of libraries (second experiment). (a) Dot plots of HEK 293T cells transfected with EGFR-WT/L858R and twice sorted libraries as indicated at the top followed by application of selection pressures as indicated on the right. Phosphorylation was detected with a pY998/pY1092-specific antibody (y-axis), expression with a c-myc-tag-directed antibody (x-axis). 500 nM erlotinib or 500 nM osimertinib were applied as indicated. **(b)** Percentages of phosphorylation-positive HEK 293T cells normalized by percentages of EGFR-expressing HEK 293T cells conferred by different libraries tested for activating pY1092, activating pY998 and erlotinib resistance phenotype.

---

namely 0.9% for the library selected with the pY998-specific antibody and 3.0% for the library selected with the pY1092-specific antibody (Fig. 4.7 (b)). This suggests that either the detection with the pY1092-specific antibody itself is more efficient or that phosphorylation of the tyrosine at position 1092 is more suited as marker for activation. In case of the detection of the EGFR-L858R mutant, which is constitutively activated in the absence of ligand, with the pY1092-specific antibody, the diagonal cell population was more defined and shifted stronger into the phosphorylation-positive gate (Fig. 4.7 (a) first row, second from right). The same applied for the final erlotinib resistance library (Fig. 4.7 (a) first row, right), which was based on the EGFR-L858R library and was therefore anticipated to retain its activating phenotype.

Regarding erlotinib resistance, it is especially interesting how libraries potentially containing activating mutations behave, since it has been observed that erlotinib more potently inhibits EGFR harboring the L858R mutation than wild type EGFR [20]. Figure 4.7 (third row) clearly shows that a small fraction of the library selected for activating mutations by using the pY1092-specific antibody for detection of activation shifted into the erlotinib resistance gate. In comparison, the library selected with the pY998-specific antibody is not distinguishable from the behavior of EGFR-WT. This indicates that the activating library, selected with the pY1092-specific antibody, contained, among others, mutations that confer erlotinib resistance.

Besides the first generation TKI erlotinib, the third generation TKI osimertinib (AZD9291), which inhibits EGFR harboring the erlotinib resistance mutation T790M [32], was applied to the selected libraries. As with erlotinib, also osimertinib showed stronger inhibition of the

EGFR-L858R mutant compared to EGFR-WT (Fig. 4.7 (a) last row). Neither the library selected for erlotinib resistance, nor both activating libraries displayed any significant osimertinib resistance. In fact, the bulk population of EGFR-expressing cells transfected with the activating libraries seems to be even slightly more inhibited by osimertinib treatment in relation to EGFR-WT.

Furthermore, it was studied whether the cetuximab resistance library exhibits any activating phenotype and vice versa. Since cetuximab inhibits EGFR signaling indirectly by trapping the

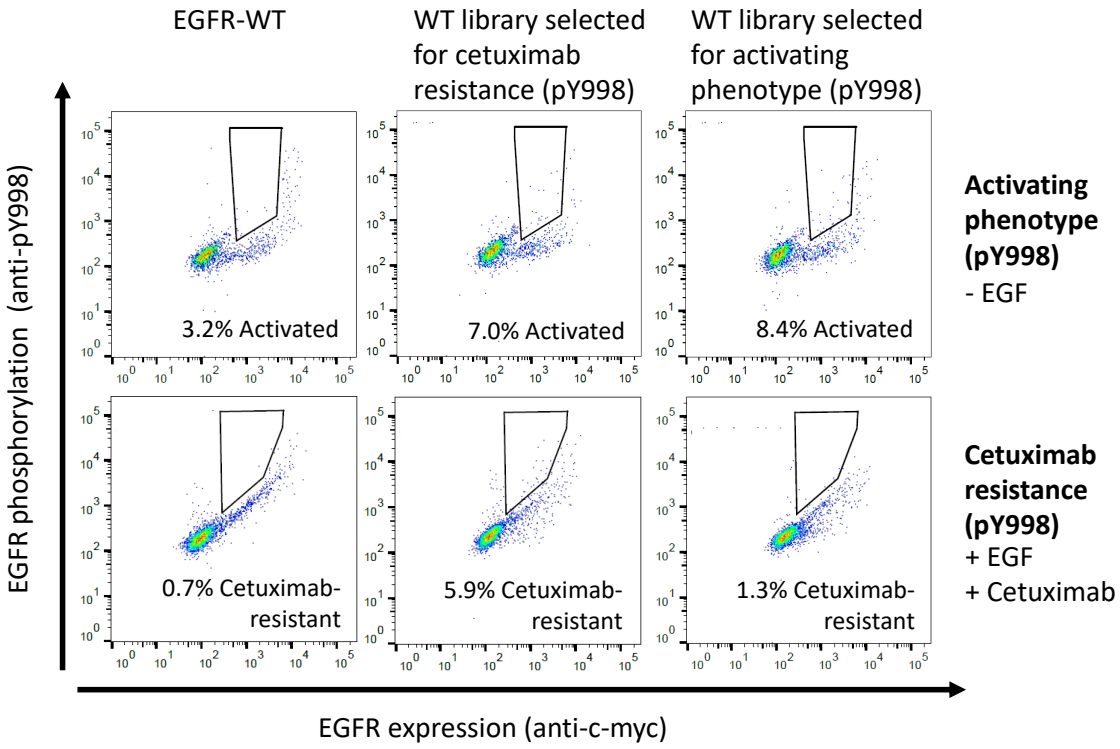

Figure 4.8: Cross-reactivity of cetuximab resistance and activating libraries (first experiment). (a) Dot plots of HEK 293T cells transfected with EGFR-WT and twice sorted libraries for cetuximab resistance and activating phenotype followed by application of selection pressures as indicated on the right. Phosphorylation was detected with a pY998-specific antibody (y-axis), expression with a c-myc-tag-specific antibody (x-axis). 100 nM cetuximab and 100 ng/mL EGF, respectively, were applied as indicated. Percentages of phosphorylation-positive HEK 293T cells normalized by percentages of EGFR-expressing HEK 293T cells are shown in every dot plot.

molecule extracellularly in the inactive conformation [23], it might be possible that a mutation stabilizing the active conformation confers cetuximab resistance. Interestingly, both libraries resulted in approximately the same percentage of cells expressing activated EGFR in the absence of ligand, namely 7.0% and 8.4%, respectively (Fig. 4.8, first row). However, the library selected for activating mutations appears to be less cetuximab-resistant (Fig. 4.8, second row, right). This indicates that the selected cetuximab resistance mutations exhibit, at least to some extent, an activating phenotype. Importantly, it needs to be considered that the flow cytometry experiments for cetuximab resistance were conducted with a comparably little number of cells, requiring caution regarding interpretation of the results.

4.3.2 Sanger sequencing

To obtain a first insight into possibly enriched mutations, Sanger sequencing chromatograms of all initial as well as already selected libraries were analyzed at certain positions of interest. Those positions included for example the codon for threonine at position 790 in the protein, which was assumed to be partly mutated in the library for erlotinib resistance. Furthermore, the selected libraries were based on either *EGFR*-WT or *EGFR*-L858R libraries, what rendered them to a certain degree distinguishable from each other. Since libraries selected for different phenotypes were handled together throughout all assay steps, the possibility of cross-contamination, especially in the light of apparent cross-reactivity of some libraries (see Section 4.3.1.2), had to be precluded. Of course, this is only possible if the frequency of the mutation is above a certain threshold. Still, Sanger sequencing resembles a useful method to check for enrichment and to exclude contaminations.

Exemplarily, the sequence chromatograms for the selection of erlotinib resistance from the initial, unselected library to the final, twice sorted library are shown in Figure 4.9 for both experiments. First, the codon for the L858R mutation, on which the selected library was based and therefore served as cross-contamination control, is displayed. As can be seen, the unselected *EGFR*-L858R library clearly contained the codon for arginine, CGG, instead of leucine, CTG. This is also the case for the once and twice sorted erlotinib resistance library. Second, in the final,

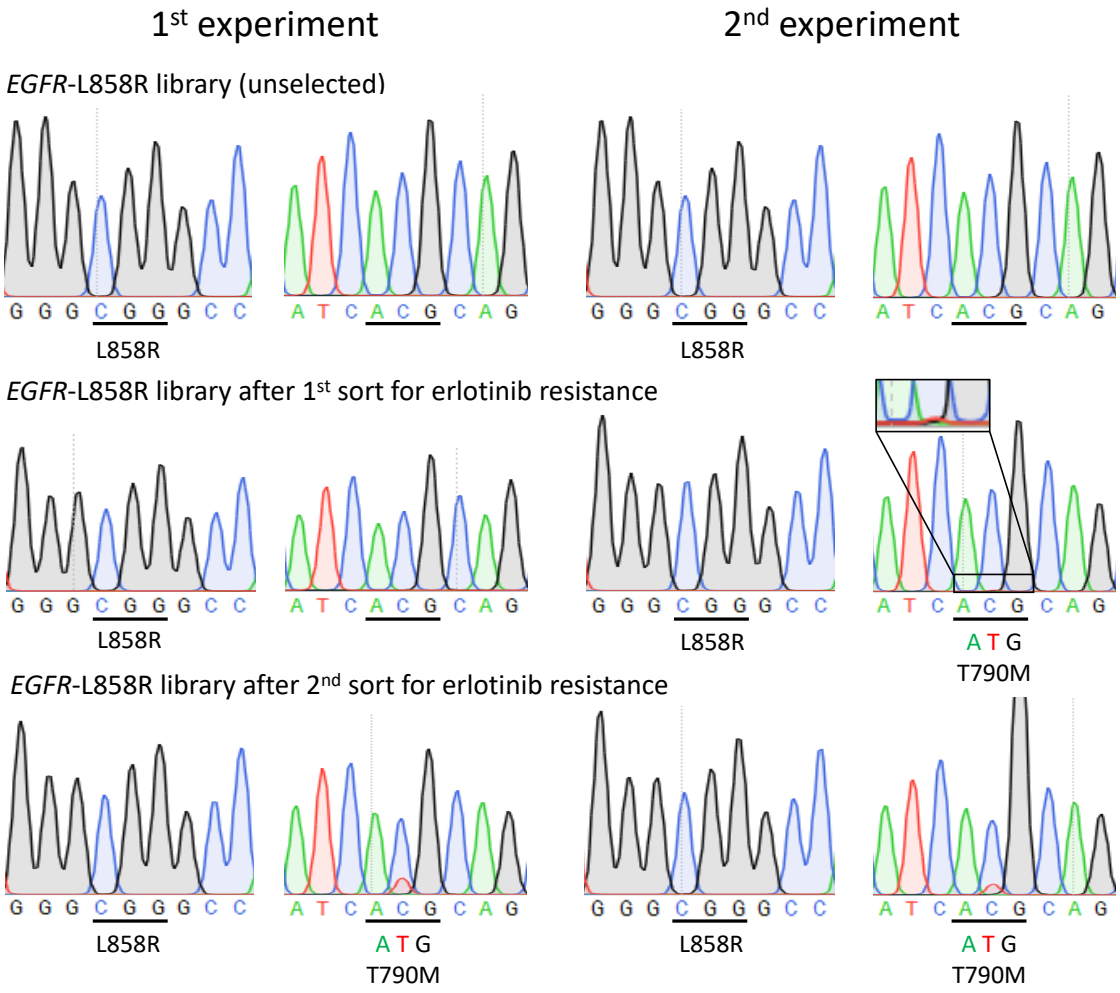

Figure 4.9: Sanger sequencing chromatograms. Exemplarily, chromatograms from the unselected *EGFR*-L858R library (top), the once (middle) and twice (bottom) sorted library from the first (left) and second (right) experiment are shown. The codons for L858R (CGG) and T790 (ACG) are underlined, the mutated codon for T790M (ATG) is indicated below the wild type codon.

twice sorted library the second base of the codon for threonine at position 790, ACG, showed a clear mixed-peak consisting of the wild type cytosine and the mutant thymidine. This mutation results in the exchange of threonine by methionine in the protein, resembling the erlotinib resistance mutation T790M, that is known from the clinic. Importantly, in the second experiment, the mixed-peak for T790M was already visible in the chromatogram after one selection round for erlotinib resistance, as magnified in Figure 4.9 (second row). Naturally, the base call quality in

this area of the chromatogram was evaluated and the phred quality scores for all displayed bases were found to be above 50, supporting the validity of the presence of the respective mutation.

4.3.3 Illumina next generation sequencing

To identify all mutations that were enriched above a certain threshold, Illumina next generation sequencing was performed. Thereby, approximately 20 million reads were obtained per library. Figure 4.10, 4.11 and 4.12 show all called mutations for *EGFR*-WT, the initial, unselected libraries and all final, twice selected libraries.

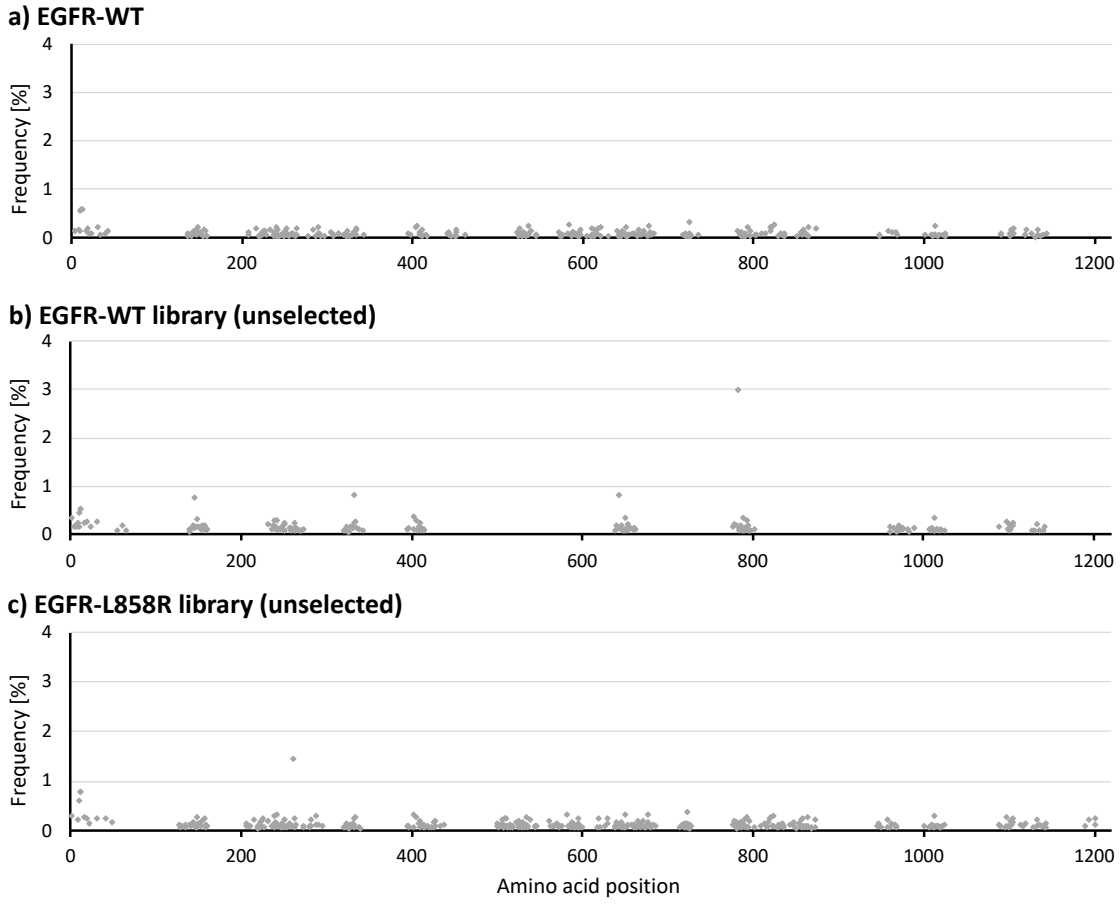

Figure 4.10: Unfiltered Illumina next generation sequencing data. **(a)** *EGFR*-WT **(b)** unselected *EGFR*-WT library and **(c)** unselected *EGFR*-L858R library. The frequency of nucleotide mutations is plotted against the amino acid position.

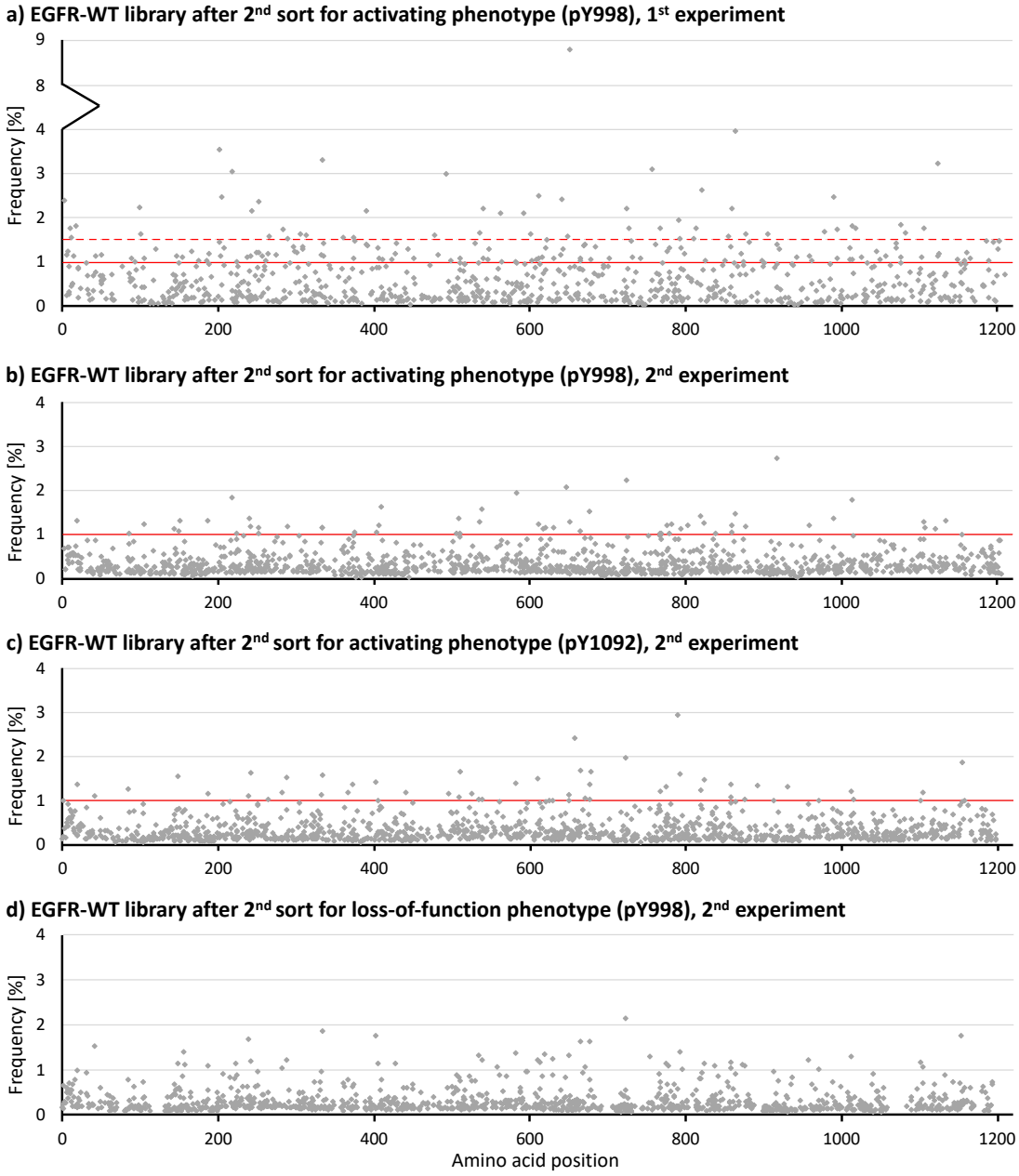

Figure 4.11: Unfiltered Illumina next generation sequencing data. *EGFR*-WT library sorted twice for **(a)** activating phenotype (pY998) from the first experiment **(b)** activating phenotype (pY998) from the second experiment **(c)** activating phenotype (pY1092) and **(d)** loss-of-function phenotype (pY998). The 1% frequency threshold is indicated by a red line, the 1.5% frequency threshold in the case of (a) by a dashed, red line. The frequency of nucleotide mutations is plotted against the amino acid position.

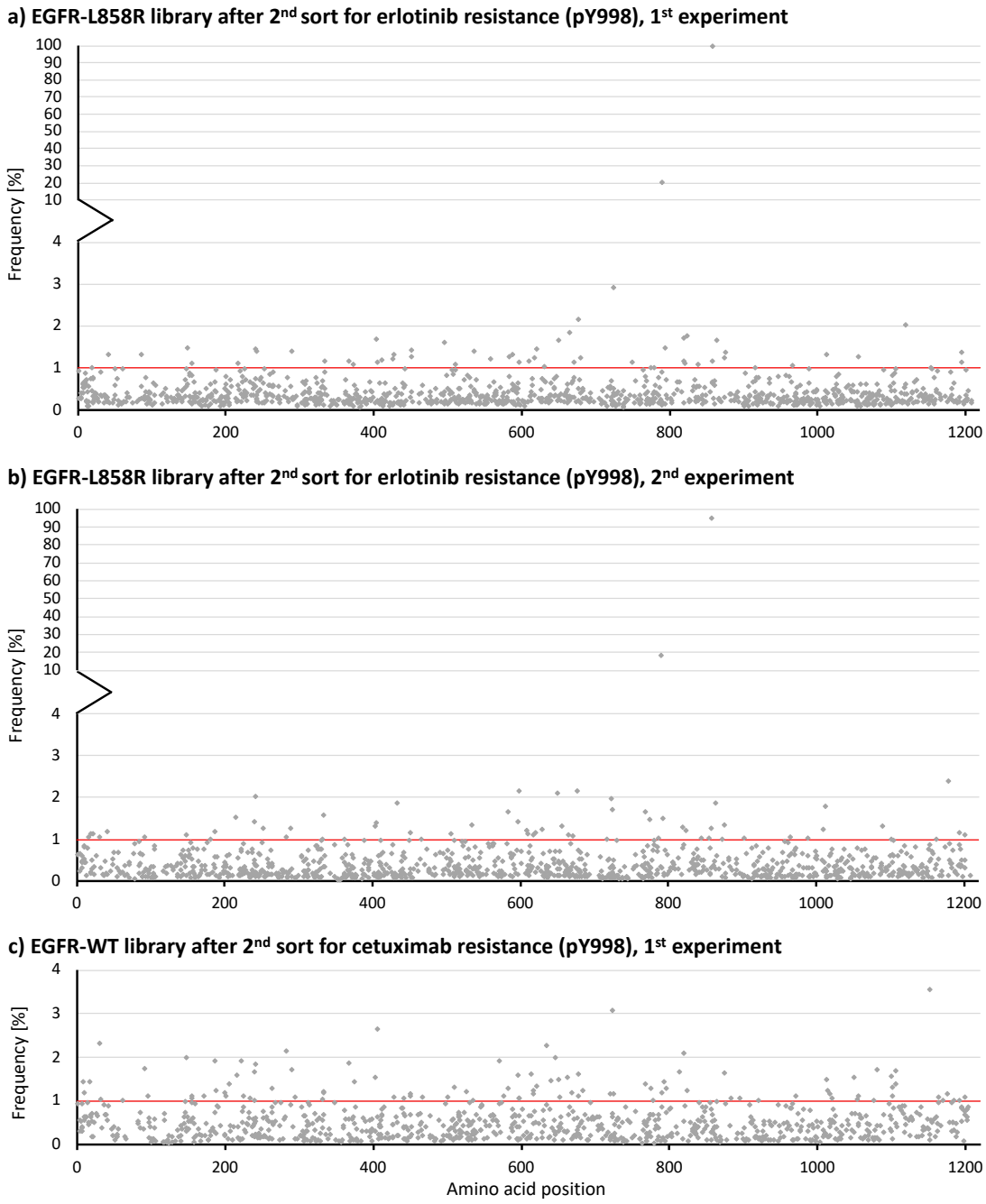

Figure 4.12: Unfiltered Illumina next generation sequencing data. *EGFR*-L858R library selected twice for **(a)** erlotinib resistance from the first experiment **(b)** erlotinib resistance from the second experiment and **(c)** EGFR-WT library sorted twice for cetuximab resistance. The 1% frequency threshold is indicated by a red line. The frequency of nucleotide mutations is plotted against the amino acid position.

For better comparability with already known mutations, the frequency of the nucleotide mutations was plotted against the amino acid position. The sequencing diagrams for *EGFR*-WT, *EGFR*-WT library and *EGFR*-L858R library (Fig. 4.10) elucidate that called mutations were clustered and the clusters were located in approximately the same areas of the protein. The mutational frequency in those unselected libraries was, as expected, low, except for one mutation in the *EGFR*-WT library with a frequency of approximately 3.0% and another in the *EGFR*-L858R library with a frequency of 1.5%. Interestingly, both mutations were either not called or substantially depleted in the selected libraries.

Initially, it became evident by Sanger sequencing of individual library clones that, regardless of the libraries' phenotype observed in the flow cytometric analysis, a rather large number of mutations was present throughout the whole *EGFR* gene (data not shown). This finding was confirmed by deep sequencing as can be seen in Figure 4.11 and 4.12. It is evident that most of the present mutations did not exceed a frequency of approximately 1% (threshold indicated as red line), except in the library containing activating mutations from the first experiment in which the mutation frequency was generally higher (Fig. 4.11 (a), 1.5% threshold indicated as dashed, red line). Therefore, an arbitrary threshold of 1% was set in order to focus on the mutations that were most enriched. All mutations below the 1% threshold were excluded from further analysis.

As already mentioned, in all libraries mutations were distributed throughout the whole *EGFR* gene. A comparison between them revealed that a substantial amount of mutations was also present in every library. To a certain extent overlaps were suspected, however, those overlaps occurred not only in libraries selected for an activating or certain resistance phenotype, but also in the library selected for loss-of-function mutations. It was assumed that enriched mutations in the loss-of-function library should be rather complementary to mutations in all other libraries and certainly not overlapping. Therefore, the sequencing data were further examined, which revealed a certain pattern inherent to those overlapping mutations, rendering them artifacts of the library preparation process. Those artifacts were most probably caused by the polymerase used for amplification of selected *EGFR* mutants, which will be discussed in full detail in Sec-

tion 5.5. Consequently, all mutations that were present in libraries selected for an activating or resistance phenotype and concurrently in the loss-of-function library and were not relatively enriched to a certain extent in one of those libraries, were excluded from further analysis. A threshold of 8-fold enrichment of a mutation compared to the loss-of-function library was arbitrarily chosen to consider the mutation a ”true“ mutation (i.e. not an artifact caused by the

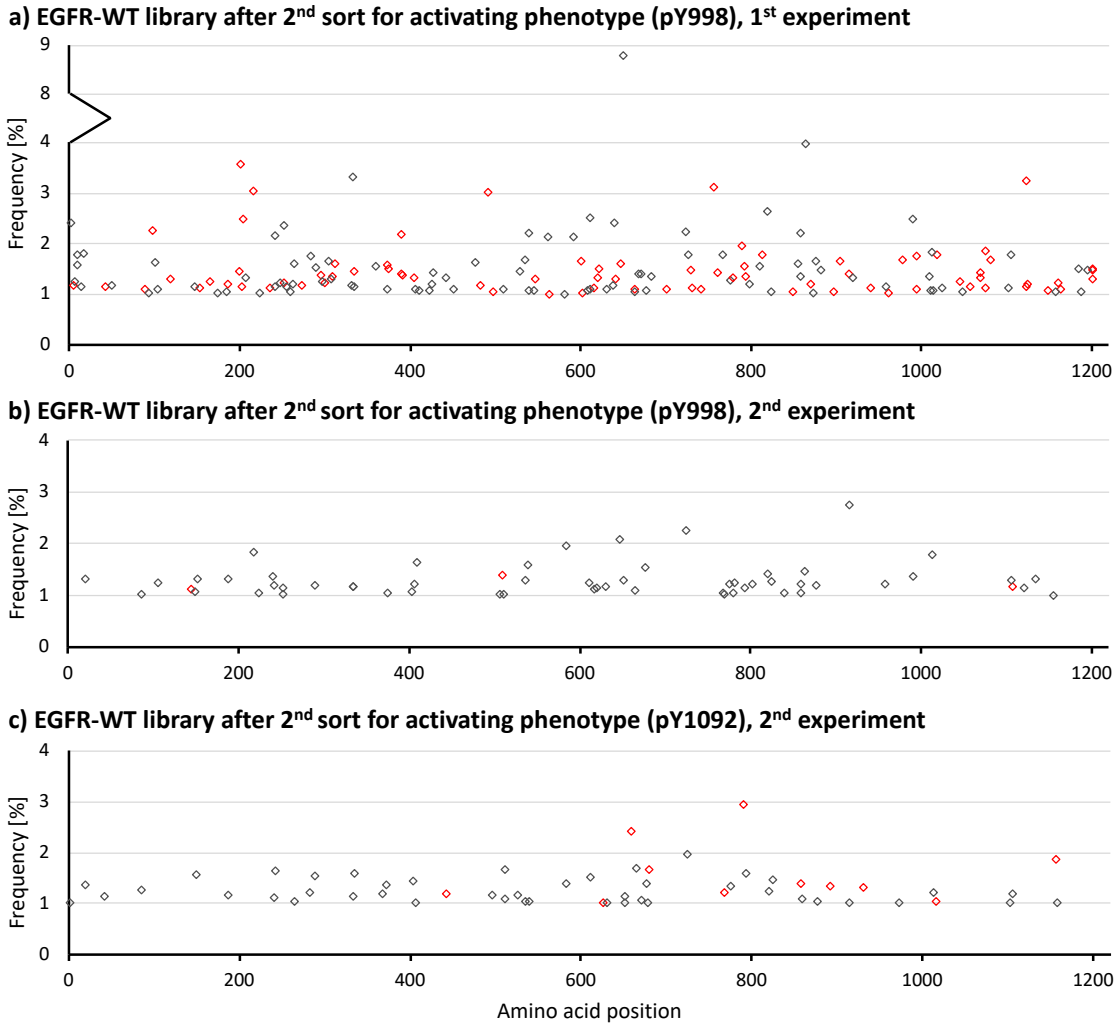

Figure 4.13: Mutations above the 1% threshold. *EGFR*-WT library selected twice for **(a)** activating phenotype (pY998) from the first experiment **(b)** activating phenotype (pY998) from the second experiment and **(c)** activating phenotype (pY1092). Mutations not present in the loss-of-function library or at least 8-fold enriched are shown in red. Mutations below 8-fold enrichment are shown in grey.

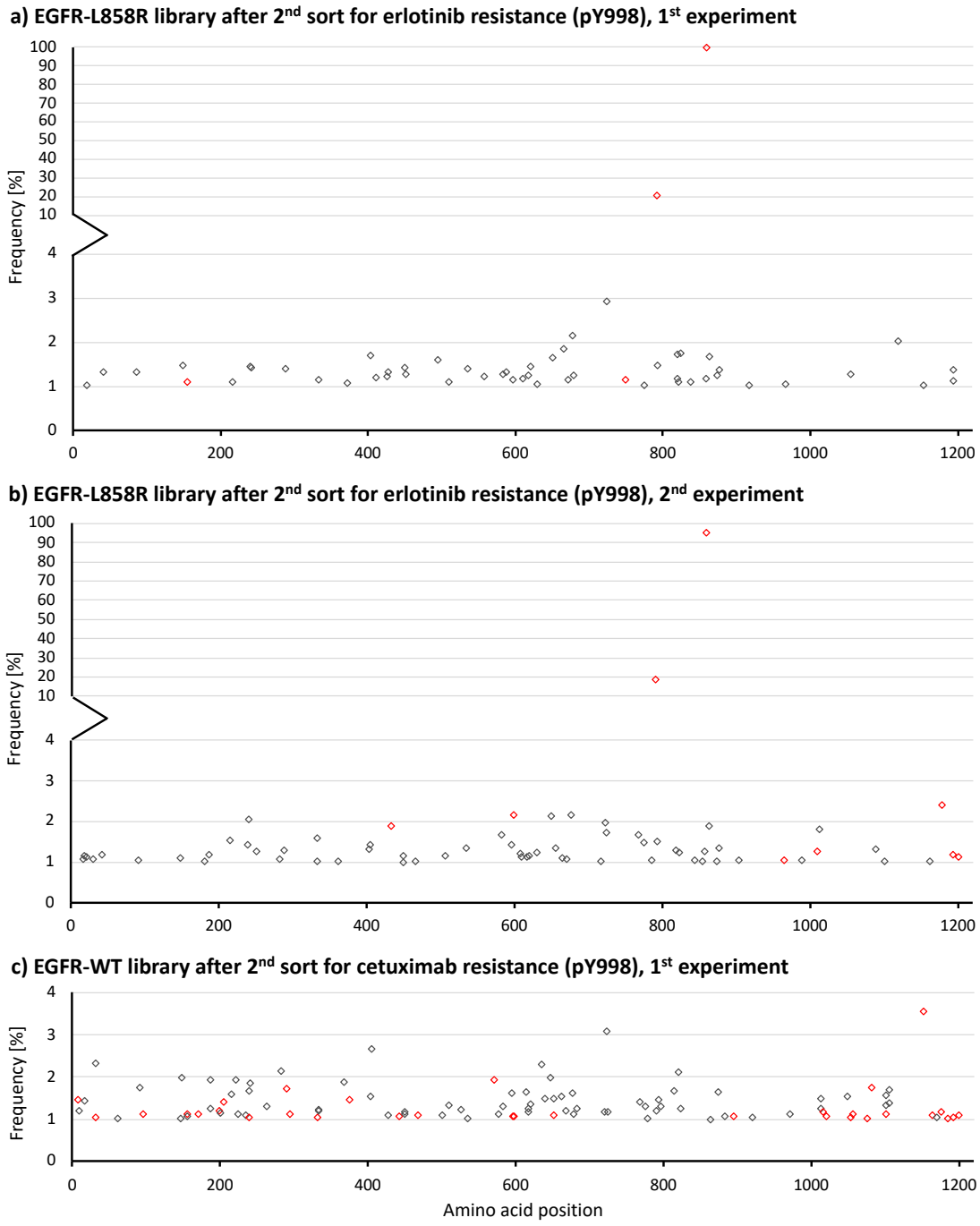

Figure 4.14: Mutations above the 1% threshold. *EGFR*-L858R library selected twice for **(a)** erlotinib resistance (1st experiment) **(b)** erlotinib resistance (2nd experiment) and **(c)** EGFR-WT library sorted twice for cetuximab resistance. Mutations not present in the loss-of-function library or at least 8-fold enriched are shown in red, mutations below 8-fold enrichment in grey.

experimental procedure). In Figure 4.13 and 4.14 the distribution of the mutations above the 1% threshold is displayed. Furthermore, mutations below the 8-fold enrichment compared to the loss-of-function library are shown in grey, whereas mutations above are shown in red.

Those steps reduced the number of relevant mutations greatly for all libraries. For reasons of clarity, the mutations passing both thresholds are again shown in Figure 4.15 and summarized in Table 4.2, 4.3 and 4.4. In the case of the library selected for activating mutations from the first experiment, which had in general a higher mutational rate, only mutations above a frequency of 1.5% are listed in Table 4.4. Apart from mutations resulting in a change of amino acid at a certain position, also mutations leading to a codon translating to the same amino acid, so-called silent mutations, were present. Furthermore, most unexpectedly, stop codons as well as frameshifts passed the applied thresholds. After removal of the library preparation artifacts by only considering mutations that were enriched at least 8-fold compared to the loss-of-function library, also the strong overlap between individual selections was lost. This appears reasonable in the case of libraries selected for different phenotypes, however, also selections for the same phenotype from different experiments lacked shared mutations. Notably, T790M was an exception to this circumstance. Finally, all listed mutations were compared with the data present in the COSMIC.

As expected, the mutation T790M was substantially enriched in the libraries selected for erlotinib resistance (Fig. 4.15 (c) and Table 4.2). This enrichment was observed in both independent experiments at almost the same frequency (18.6% and 20.5%). Besides T790M, another mutation in the kinase domain, A750T, was present at a frequency of 1.1%. Interestingly, most of the other enriched mutations were located in the C-terminal tail of EGFR. Additionally, the presence of the L858R mutation, on which the erlotinib resistance libraries were based, could be detected with frequencies of 99.7% and 95.3%, respectively. This confirms that L858R is required for efficient EGFR signaling in this experimental setup, i.e. the absence of EGF and presence of erlotinib, and was therefore largely maintained during the selection process.

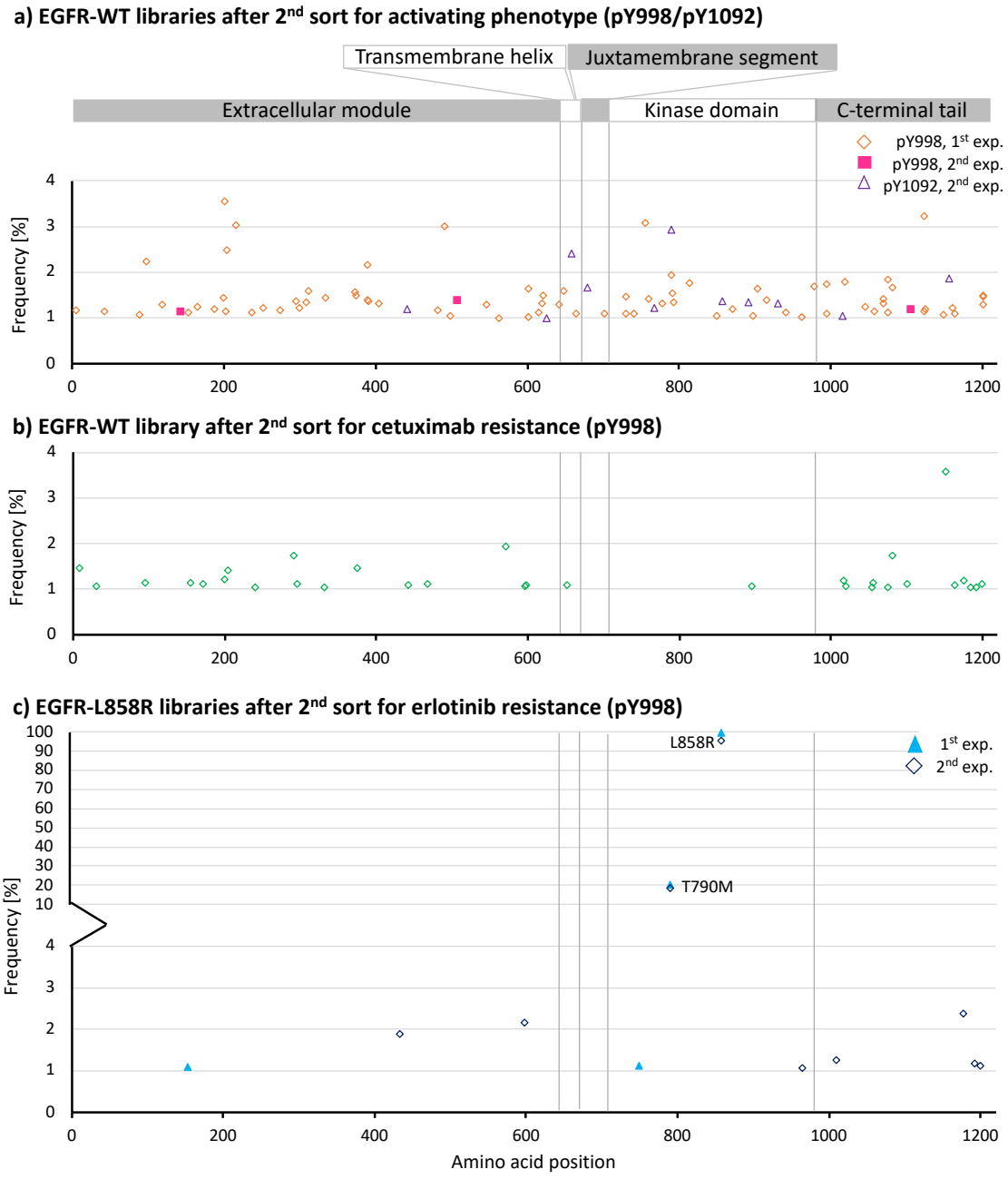

Figure 4.15: Mutations passing the 1% frequency and 8-fold enrichment threshold for each phenotype. **(a)** Activating mutations selected with the pY998-specific antibody from the first (orange) and second (pink) experiment or selected with the pY1092-specific antibody (violet), **b)** cetuximab resistance mutations (green) and **(c)** erlotinib resistance mutations from the first (light blue) and second (dark blue) experiment. EGFR's structural domains are shown at the top.

Table 4.2: Erlotinib resistance mutations. All mutations present in the *EGFR*-L858R libraries selected for erlotinib resistance passing the 8% threshold for enrichment compared to the loss-of-function library and reaching a frequency of 1% are listed. For all presented mutations the number of COSMIC entries are listed (as of 18/10/2018).

Selection	Amino acid positon	Mutation	Frequency [%]	COSMIC
1st experiment **Erlotinib resistance**	155	Ala → Thr	1.1	-
	750	Ala → Thr	1.1	1
	790	Thr → Met	20.5	1348
2nd experiment **Erlotinib resistance**	434	Gly → Asp	1.9	-
	599	Val → Ala	2.2	-
	790	Thr → Met	18.6	1348
	965	Ile → Ile	1.1	-
	1010	Val → Ala	1.3	-
	1179	Lys → Glu	2.4	-
	1194	Asn → Thr	1.2	-
	1201	Ala → Ala	1.1	3

Regarding the library selected for cetuximab resistance, mutations were not limited to the cetuximab epitope. Interestingly, the kinase domain, spanning residues 707 to 979, contained only one mutation. All other mutations were mainly located in the extracellular module or the C-terminal tail (Fig. 4.15 (b) and Table 4.3).

As can be seen in Table 4.4, the activating mutation L858R, which is the most prominent point mutation in lung cancer [15], was present in the library sorted for activating phenotype using the antibody targeting pY1092. In contrast, the mutation was not enriched above the 1% threshold in the other activating libraries where the pY998-specific antibody was used. This indicates a different suitability of the two antibodies for selection of at least this particular mutation, which will be further discussed in Section 5.2. Additionally, mutations spaced throughout the whole protein were present as was expected to some extent for the selection of an activating phenotype. In Section 5.7 certain selected mutations of all libraries will be discussed in a structural and functional context.

The loss-of-function library itself contained only a single mutation not present in other libraries.

Apart from this mutation, the loss-of-function library showed the same pattern inherent to the mutations present in all libraries, and was therefore only used as mutational negative control.

Table 4.3: Cetuximab resistance mutations. All mutations present in the *EGFR*-WT library selected for cetuximab resistance passing the 8% threshold for enrichment compared to the loss-of-function library and reaching a frequency of 1% are listed. For all presented mutations the number of COSMIC entries are listed (as of 18/10/2018).

Selection	Amino acid position	Mutation	Frequency [%]	COSMIC
1st experiment **Cetuximab resistance**	8	Gly → Gly	1.5	-
	32	Gln → STOP	1.1	-
	96	Val → Met	1.1	-
	156	Leu → Gln	1.1	-
	171	Asp → frameshift	1.1	-
	200	Trp → frameshift	1.2	-
	205	Glu → Glu	1.4	-
	240	Cys → Ser	1.0	-
	291	Cys → Phe	1.7	-
	295	Cys → frameshift	1.1	-
	332	Pro → Leu	1.0	-
	375	Ala → Val	1.5	-
	442	Ser → Arg	1.1	-
	467	Lys → Asn	1.1	-
	571	Cys → Cys	1.9	-
	597	Ala → Ser	1.1	-
	598	Gly → STOP	1.1	-
	652	Gly → Trp	1.1	1
	896	Asp → Asn	1.1	-
	1016	Tyr → STOP	1.2	-
	1020	Gln → Gln	1.1	-
	1054	Gly → Arg	1.0	-
	1056	Gln → STOP	1.1	-
	1076	Ala → Thr	1.0	-
	1081	Ser → Thr	1.7	-
	1101	Pro → Pro	1.1	-
	1151	Phe → Leu	3.6	-
	1164	Gln → Arg	1.1	1
	1176	Phe → Ser	1.2	-
	1184	Asn → Lys	1.0	-
	1192	Ala → Ala	1.0	-
	1199	Arg → Met	1.1	-

Table 4.4: Activating mutations. All mutations present in the *EGFR*-WT libraries selected for activating phenotype passing the 8% threshold for enrichment compared to the loss-of-function library and reaching either a frequency of 1.5% (1st experiment, activating mutations (pY998)) or 1% (2nd experiment, activating mutations (pY998 and pY1092) are listed. For all presented mutations the number of COSMIC entries are listed (as of 18/10/2018).

Selection	Amino acid position	Mutation	Frequency [%]	COSMIC
1st experiment **Activating mutations (pY998)**	98	Arg → Gln	2.3	5
	202	Ala → Thr	3.6	-
	205	Glu → STOP	2.5	-
	217	Gln → STOP	3.1	-
	312	Gly → Glu	1.6	-
	373	Pro → Leu	1.6	1
	390	Gln → STOP	2.2	-
	492	Ser → Arg	3.0	6
	601	Gly → Ala	1.6	-
	648	Thr → Ile	1.6	-
	756	Asn → Asn	3.1	-
	790	Thr → Met	2.0	1348
	792	Leu → His	1.5	1
	813	Tyr → Tyr	1.8	-
	904	Val → Val	1.7	-
	978	Tyr → Cys	1.7	-
	995	Ser → frameshift	1.8	-
	1018	Ile → Ile	1.8	-
	1076	Ala → Val	1.9	-
	1081	Ser → Asn	1.7	-
	1123	Pro → Gln	3.2	-
2nd experiment				
Activating mutations (pY998)	144	Leu → Met	1.1	-
	508	Ala → Ala	1.4	-
	1106	Gln → Gln	1.2	-
Activating mutations (pY1092)	442	Ser → Ile	1.2	1
	626	Tyr → Cys	1.0	-
	658	Leu → Gln	2.4	1
	679	Leu → Leu	1.7	-
	768	Ser → Ile	1.2	264
	790	Thr → Met	2.9	1348
	858	Leu → Arg	1.4	10229
	892	Thr → Thr	1.4	-
	931	Glu → Gly	1.3	-
	1016	Tyr → Asn	1.0	-
	1156	His → His	1.9	-

CHAPTER 5

Discussion

5.1 Assay optimization

Establishment of a new assay naturally requires a substantial amount of evaluation and optimization. In this case, especially the isolation and amplification of *EGFR* variants between selection rounds proved to be challenging and demanded certain adaptations.

First of all, lysis of sorted HEK 293T cells, as described in Section 3.3.10, can become a bottleneck of the assay. During the second selection round, on average approximately 6,000 cells were sorted (compare Table 4.1). Consequently, only very small and therefore by conventional spectrophotometry unquantifiable amounts of plasmid DNA were recovered. On the one hand, the large number of different mutations found by Illumina next generation sequencing suggests that also a considerable number of individual *EGFR* variants was isolated. On the other hand, libraries selected for the same phenotype in independent experiments shared no common mutations, except for T790M. This could be at least partly attributed to insufficient lysis of cells and the concomitant lack of library diversity coverage. The correct use of the QIAprep Spin Miniprep kit (QIAGEN) combined with Proteinase K, which was part of the conducted assay optimization, is certainly important for *EGFR* variant recovery. It was shown that Proteinase K requires immediate transfer to its optimal activity temperature, 56°C, to achieve efficient plas-

mid DNA isolation. However, a method to directly quantify recovered plasmid DNA per sorted HEK 293T cell would be of great help to better understand the potential limitations of this assay step.

As mentioned before, only very small amounts of *EGFR* library plasmid were isolated from sorted HEK 293T cells. This necessitated amplification of plasmid DNA by PCR before transformation of *E. coli* was feasible. Even though those PCR amplifications were optimized during the present and a previous master thesis [31], problems regarding especially the presence of truncated *EGFR* remained to various extent. Ultimately, it is clear that direct transformation of *E. coli* with isolated plasmids containing selected *EGFR* variants would on the one hand save time during each selection round and on the other hand potentially improve the quality of produced *EGFR* libraries.

5.2 The choice of an antibody for phosphorylation detection

For the selection of activating mutations two different antibodies, targeting either the phosphorylated tyrosine residue at position 998 or 1092 of EGFR, were used for detection of activation. This was done on the one hand to assess the suitability of both antibodies for detection and on the other hand to see whether selection for phosphorylation of different tyrosines would result in a similar or different set of activating mutations. As can be seen in Table 4.4, it is evident that the latter is the case for our experiments. However, since even the same phospho-tyrosine-specific antibody and the same selection pressure led to the enrichment of different mutations, this circumstance certainly cannot be attributed to the antibody choice. Regarding the suitability for detection of activated EGFR and ultimately selection of activating mutations, different experiments led to the assumption that the pY1092-specific antibody is the better choice. Importantly, only in the activating library selected with the pY1092-specific antibody, the activating mutation L858R, which is the most frequently observed point mutation in lung cancer [15], was detected above the 1% threshold for the frequency of appearance. Agilent Technologies states that the error-prone PCR kit, which was used for the construction of the randomly mutated libraries, introduces the transversion from thymidine to guanine required for the L858R mutation at a

frequency of 4.7% (https://www.agilent.com/cs/library/usermanuals/Public/200550.pdf). Considering the possible transversions and transitions, this particular nucleotide switch has the lowest frequency besides the transversion from guanine to cytosine (4.1%). Compared to that, the transition from cytosine to thymidine, which is required for the introduction of T790M, appears at a frequency of 25.5%. In the light of this, it is possible that the L858R mutation was underrepresented in the initial EGFR-WT library and consequently, by chance, "missed" in the later selections using the pY998-targeting antibody.

In Figure 5.1 the phosphorylation levels of EGFR-WT and the L858R mutant in the absence of EGF stained with either the pY998-specific or the pY1092-targeting antibody is shown. Clearly, it can be seen that in the case of the antibody targeting the phospho-tyrosine at position 1092, the diagonal cell population expressing the L858R mutant is less scattered and better separated

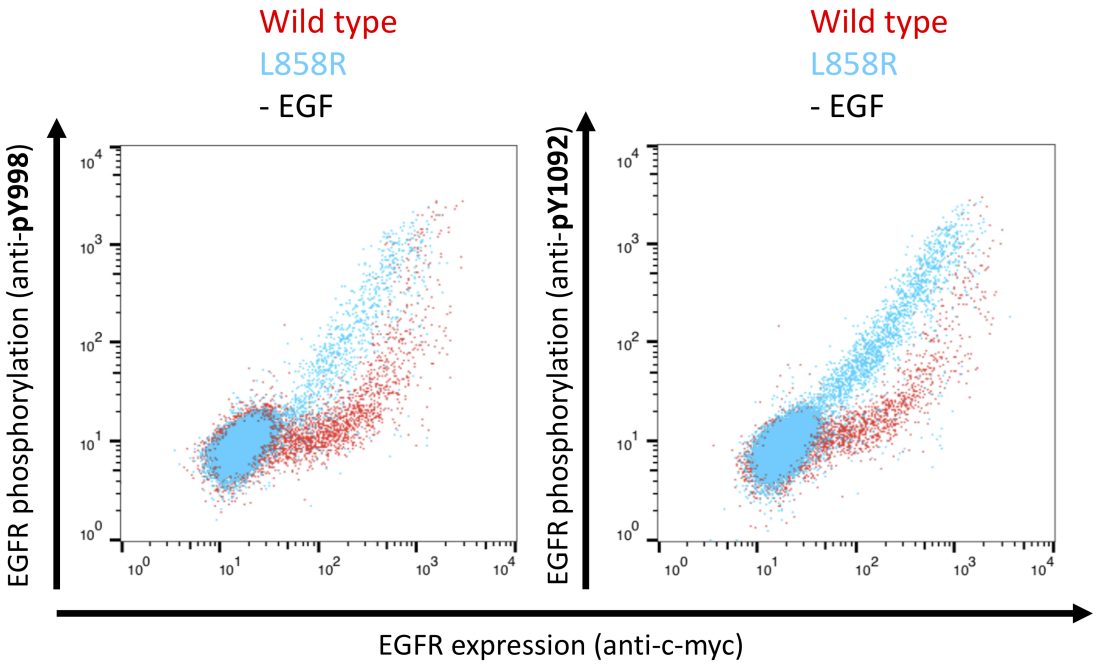

Figure 5.1: Antibody comparison for detection of EGFR phosphorylation. HEK 293T cells were transfected with plasmids encoding *EGFR*-WT (red) or the L858R mutant (blue) and studied by flow cytometry. Phosphorylation was either detected with an antibody targeting pY998 (left) or pY1092 (right) in the absence of EGF. EGFR expression was detected with a c-myc-tag-specific antibody.

from the cells with wild type EGFR. This suggests that the enrichment of the L858R mutant in the library containing activating mutations selected by pY1092 detection may be the result of better separation from EGFR-WT, on which the library was based. This effect is also visible with the erlotinib resistance library, which was based on the L858R mutant, tested for activating phenotype (Fig. 4.7, first and second row, right). The experiment for cross-reactivity also shows that regardless of the antibody used for detection, the library selected for activating mutations using the pY1092-specific antibody resulted in a higher percentage of activated EGFR (Fig. 4.7 (b)). Taken together, it appears that the pY1092-specific antibody is the better choice for the selection of activating mutations. Still, it remains to be tested if this holds also true for the selection of resistance mutations.

5.3 FACS gating for different phenotypes

Besides the choice of the right detection antibody, also the individual adjustment of sort gates highly influences the outcome of selection experiments. In the case of the selection for erlotinib-resistant EGFR variants, the L858R mutant, which was the negative control used to adjust the sort gate, was highly inhibited by erlotinib (Fig. 4.4, bottom row). This made a distinct separation of cells expressing phosphorylation-positive and therefore erlotinib-resistant EGFR from the phosphorylation-negative bulk population feasible.

Compared to that, activating mutations were selected in the mere absence of any EGFR ligand. As can be seen in Figure 4.4 (upper row, middle), ligand-independent signaling still took place at high EGFR expression levels, corresponding to the tail like cell population visible in the dot plot. This circumstance made the adjustment of the sort gate comparably difficult. If the gate is positioned rather close to that tail like population, it is possible that also EGFR variants without an activating phenotype are sorted. Figure 4.4 (upper row, middle) shows, that compared to the selections for resistance to inhibitors, an elevated number of cells from the negative control (EGFR-WT in the absence of EGF) was already present in the sort gate. This entails that the enrichment of activated EGFR is less efficient compared to selections in which the separation of active and non-active EGFR is more distinct. If, on the contrary, the gate is positioned more

stringent, diversity of activating mutations might be lost.

Concluding it can be said, that in the case of the selection for activating mutations, it can be of relevance to test different selection gates to optimize enrichment and better understand the appearance of certain mutations in the dot plots. Another possibility to refine this particular selection might be to constrict the level of EGFR expression to a range in which ligand-independent signaling does not take place.

5.4 Linking flow cytometric observations to mutational profiles

After each selection experiment, the enrichment of cells expressing phosphorylated EGFR in the presence of the respective selection pressure as well as cross-reactivity were analyzed by flow cytometry. First of all, it became clear that especially for the analysis of selections with comparably low enrichment, a large number of cells is required to obtain reliable results. Therefore, the flow cytometry data for the comparison of the libraries from the first experiment (Fig. 4.5) has to be interpreted with caution. Still, similar conclusions regarding the efficiency of individual selections compared to those from the analysis of the second experiment can be drawn.

On the one hand, the selection for erlotinib resistance led to a considerable enrichment of EGFR mutants which were active in the presence of erlotinib (Fig. 4.5, (b) right, and Fig. 4.6, (b) second diagram from right). This can be mainly attributed to the presence of the resistance mutation T790M (Table 4.2) and therefore nicely resembles the situation in the clinic, where T790M is by far the most frequently observed resistance mutation in erlotinib-treated lung cancer [26]. On the other hand, enrichment of EGFR variants active in the absence of ligand turned out to be more difficult, with comparably low percentages of positive cells (Fig. 4.5, (b) left, and Fig. 4.6, (b) first and second diagram from left). Those differences are assumed to arise at least partly from the rather difficult separation of the EGFR-overexpressing and therefore phosphorylation-positive cell population from the cell population expressing an EGFR mutant active at normal expression levels, as has been discussed in the previous Section 5.3.

Interestingly, the library selected for activating mutations using the pY1092-specific antibody showed low level erlotinib resistance, as can be seen in Figure 4.7, third row. Illumina sequencing revealed that the T790M mutation was not only enriched in the erlotinib resistance libraries but also in certain activating libraries (Table 4.4). Moreover, the mutation was detectable in every library including the unselected EGFR-WT library and EGFR-L858R library, albeit at low frequencies (0.3% and 0.2% for the EGFR-WT library and the EGFR-L858R library, respectively). As already stated, the used error-prone PCR kit introduces the transition from cytosine to thymidine required for T790M at a frequency of 25.5% which corresponds to the second highest nucleotide exchange. Therefore, it is likely that the T790M mutation was already slightly overrepresented in the unselected libraries. Furthermore, as already mentioned, literature reports that the T790M mutation leads to a 5-fold increased kinase activity compared to EGFR-WT in the absence of EGF, which indicates a weak, *de novo* activating phenotype of this mutation [27]. Besides the erlotinib resistance libraries, the highest frequency of T790M (2.9%) was detected in the selection for ligand-independent activation using the pY1092-specific antibody (Table 4.4). This is consistent with the highest percentage of erlotinib resistance of this library among activating libraries observed by flow cytometry (Fig. 4.7 (b), right). In the selections for activating phenotype using the pY998-specific antibody, T790M was present at 2.0% in the library from the first experiment (Table 4.4), which implies also in this case low level erlotinib resistance. However, as already mentioned, a larger number of cells would have been required to reliably analyze this effect by flow cytometry. Compared to that, in the pY998 activating library from the second experiment the T790M mutation did not reach the 1% threshold and was therefore excluded by this filter.

Another surprising observation provided the analysis of the cetuximab resistance library in the absence of EGF. Figure 4.8 (upper row) shows that almost the same percentage of cells expressing the cetuximab resistance library were phosphorylation-positive compared to the activating library. As already mentioned, this observation requires further experiments for verification. Surprisingly, Illumina sequencing showed that the T790M mutation was also present at a frequency of 1.2% in the cetuximab resistance library, however, only enriched 5-fold compared to

the loss-of-function library and therefore not listed in Table 4.3. If the activating phenotype of the library selected for cetuximab resistance would be indeed reproducible, a possible explanation for the enrichment of activating mutations could be found in the set-up of this particular selection, as will be explained below.

Before FACS selection, cells were serum starved for 16 to 18 hours to reduce background phosphorylation. Shortly before fixation of cells with methanol, incubation of cells with cetuximab at 4°C was followed by stimulation with EGF. It was assumed that cetuximab-resistant EGFR variants would still be able to bind EGF and phosphorylate their tyrosines despite the presence of the antibody, which would consequently lead to the selection of a cetuximab-resistant library. However, during the period of serum starvation, EGFR variants which are active in the absence of EGF were also able to phosphorylate their tyrosine residues. Since the cetuximab incubation was done at 4°C, it is reasonable to assume, that this incubation did not significantly reduce the phosphorylation level that had been present before the addition of cetuximab. Moreover, EGFR internalization from the cellular membrane was most likely downregulated, leaving the receptors active and independent of inhibition with cetuximab. Those EGFR variants were wrongly considered cetuximab-resistant in the following flow cytometry experiment, while they were actually exhibiting an activating phenotype and cells expressing those variants were selected. Consequently, the selected EGFR library consisted of a mixture of cetuximab-resistant variants and activating mutants. This would explain why the apparent cetuximab resistance library was also active in the absence of ligand (Fig. 4.8, first row).

In comparison, the library selected for activating mutations in the same experiment showed low-level cetuximab resistance (Fig. 4.8, second row). This could be attributed to the rather unexpected presence of the S492R mutation, which is associated with cetuximab resistance [24], in this activating library. Besides that, another residue at position 442 was mutated in both the pY1092 activating library and cetuximab resistance library, albeit to different amino acids (Table 4.3 and 4.4). Nevertheless, both mutations are located in the cetuximab epitope. In this context, it would be interesting to study whether those mutations influence the binding of EGF, whose

binding site partially overlaps with that of cetuximab [23]. Theoretically, bovine EGF should not be present in the culture medium before selection of activating mutations due to serum starvation. However, since no wash step was conducted before medium exchange, it is possible that bovine EGF had been present at very low concentrations. Consequently, a mutation which enhances the affinity of bovine EGF to EGFR might exhibit on the one hand activating properties and on the other hand confer low level cetuximab resistance in this particular set-up. Clearly, this experiment has to to be repeated with a sufficient number of cells to draw any evidence-supported conclusions.

5.5 Overlapping mutations - a possible explanation

During the conducted selection experiments, it became evident by Sanger sequencing that some mutations were not only present in one distinct library but could also be found in other libraries selected for different phenotypes. Since Sanger sequencing of a certain number of clones is certainly not enough to determine the frequency of a mutation, Illumina next generation sequencing was used to elucidate the mutational profile of the obtained libraries. On the contrary to what was expected, next generation sequencing confirmed the beforehand observed mutational overlaps not only between libraries selected for an activating or resistance phenotype but also the loss-of-function phenotype (Fig. 4.13 and 4.14, all mutations that are not substantially enriched compared to the loss-of-function library are shown in grey). As has been described with the cancer related EGFRvIII [19], it is possible that a mutation leads to decreased activity in the presence of ligand (potentially explaining enrichment in the loss-of-function library), however, is still constitutively active at a low level in the absence of ligand (potentially explaining enrichment in the libraries selected for activating phenotype). Still, it is highly unlikely that many different mutations are enriched in all libraries at approximately the same frequency. This finding prompted the search for possible explanations, which will be discussed in the following.

At first, cross-contamination between libraries was considered to be the source of those overlaps. However, many of the overlapping mutations were present at almost the same frequency in all selected libraries and all libraries contained either entirely the wild type codon at position

858 or the L858R mutation. Therefore, it can be assumed that if a certain overlapping mutation is present due to contaminations, also a substantial amount of contamination with either the L858R mutant or *EGFR*-WT has to be visible. Since this is not the case, the possibility of cross-contamination between all samples was excluded. Another explanation would be the introduction of sequencing errors by Illumina sequencing itself at certain positions of the gene. This would explain the similar frequencies of overlapping mutations throughout the libraries. As can be seen in Figure 4.10, not only the selected libraries but also the unselected *EGFR*-WT library and *EGFR*-L858R library as well as non-mutated *EGFR*-WT were sequenced. Those samples also contained the overlapping mutations, however, at a substantially lower frequency (compare the overlapping (grey) mutations in Fig. 4.13 and 4.14 with mutations in Fig. 4.10, where almost all mutations are below 0.5%). Thus, overlapping mutations were only found in libraries that underwent the selection process, irrespective of the type of selection pressure that had been applied. Since it is highly unlikely that the frequency of sequencing errors introduced by Illumina sequencing differs only in samples which were not subject to selections, the theory of sequencing errors was also discarded.

The second above mentioned theory, even though not itself the solution to the mystery, holds the clue to the most likely explanation. It is evident that Illumina sequencing is not the only step in the conducted selection assay where correct amplification of DNA sequences is required. As already explained, *EGFR* libraries were subject to numerous PCR cycles. More precisely, in total well over 100 PCR cycles were done in one selection experiment with two selection rounds and a final amplification for Illumina sequencing. The Phusion® High-Fidelity DNA polymerase, which was used for all amplifications, has an error rate of 9.5 x 10^{-7} with the used Phusion® GC buffer (User Guide: Phusion High-Fidelity DNA Polymerase from Thermo Scientific™, as referred to by NEB). However, this error rate is most certainly an average value, which means that certain sequence patterns will most likely be amplified with an increased error rate. Therefore, it is assumed that such a large number of PCR cycles, despite the overall low error rate, introduces certain mutations. Also the different frequency of the mutations in the selected libraries compared to the unselected libraries and the wild type is explainable by this theory, since the

latter were only subject to a comparably small number of PCR cycles.

To test the hypothesis, that the Phusion® polymerase introduced certain mutation patterns, the mutations in the loss-of-function library were further analyzed. It was found that 44 mutations were present above a frequency of 1% in the loss-of-function library, whereby only one of those is unique to the loss-of-function library. All other mutations were enriched at least in two other libraries selected for activating or resistance phenotypes. Of the 44 mutations above 1%, 23 share a guanine to adenine substitution, while 18 comprise a cytosine to thymidine exchange and 3 lead to a frameshift. Considering that a cytosine to thymidine exchange on the one strand entails a guanine to adenine substitution on the complementary strand and vice versa, it became clear that in fact all point mutations detected above 1% in the loss-of-function library are exchanges of guanine to adenine. Interestingly, 27 of those mutations are located in the sequence GCA. In Table 5.1, a selection of those mutations, namely all reaching at least a frequency of 1.5%, is given. As can be seen, 5 out of 7 point mutations are located in the GCA motif. The cytosine to thymidine mutation at position 3461 (grey) is unique to the loss-of-function library and also not located in the GCA motif.

Table 5.1: Mutations with a frequency of at least 1.5% present in the loss-of-function library. The listed mutations, which were present in at least two other libraries except for one mutation, which is unique to the loss-of-function library (highlighted in grey), were further analyzed to deduce potential sequence patterns. The sequence motif, GCA, possibly causing amplification errors of the used Phusion® High-Fidelity DNA polymerase is highlighted in red. NT..nucleotide, AA..amino acid.

NT position	AA position	NT change	Initial codon	New codon	AA change	Frequency [%]	Sequence motif
125	42	G → A	GGC	GAC	G → D	1.5	GGGCACT
719	240	G → A	TGC	TAC	C → Y	1.7	CTGCACA
1001	334	G → A	CGC	CAC	R → H	1.9	CCGCAAA
1209	403	GT → G		frame shift		1.8	
1994	665	G → A	GGC	GAC	G → D	1.6	CGGCCTC
2030	677	G → A	CGC	CAC	R → H	1.7	GCGCACG
2171	724	G → A	GGC	GAC	G → D	2.2	CGGCACG
3461	1154	C → T	CCT	CTT	P → L	1.8	

Ultimately, those findings suggest that the polymerase shows an increased error rate at the sequence GCA and is indeed the source of the mutational overlaps. In the current setting of the assay, PCR amplification of *EGFR* libraries is indispensable. This entails the need for a control that enables the distinction between those artificially introduced mutations and mutations exhibiting a certain phenotype. If the assumption that mutations present in the library selected for loss-of-function should not be found in libraries requiring a functional molecule holds true, the loss-of-function selection can serve as such a control. In this context, it needs to be considered that the variant calling algorithms employed by Mutect2 only call variants if a genomic region is considered to be "active", i.e. only if variability above a certain threshold occurs. Thereby, also the occurrence of mutation clusters in samples with an overall low mutation rate, as is the case for EGFR-WT and the unselected libraries (Fig. 4.10), can be explained. The threshold for a mutation to be considered relevant was set at a frequency of 1% and at least 8-fold enrichment compared to the loss-of-function library. This implies that if a mutation has a frequency of 1% in a certain library, its frequency requires to be below 0.125% in the loss-of-function library, which in turn is below the required threshold for a region to be considered active by Mutect2. In such cases the mutation would be missing in the loss-of-function library and wrongly regarded as "true" mutation in the library selected for a resistance or activating phenotype. It is possible to check for the presence of those low-frequency mutations in the respective Bam-file, which is a file format that contains the sequence alignment data, of the loss-of-function library. Nevertheless, as further discussed below, a substantially improved enrichment of mutations exhibiting a phenotype would be desirable.

5.6 A problem of numbers

In all libraries, except the one selected for erlotinib resistance, unique mutations are indistinguishable from artifacts in terms of frequency of appearance (compare Fig. 4.13 and Fig. 4.14). Furthermore, there is also no difference in frequency between mutations which could actually have a certain phenotype and silent mutations as well as mutations introducing stop codons or frameshifts (compare Tables 4.4, 4.3 and 4.2). Transfection of HEK 293T cells with the selected activating libraries and cetuximab resistance libraries resulted in relatively low percentages of

phosphorylation positive cells (Fig. 4.5 and Fig. 4.6). For this reason, it is in turn not possible that a considerable number of mutations is substantially enriched, as was the case for the T790M mutation in the erlotinib resistance library. Furthermore, it is assumed that a larger number of mutations throughout EGFR is able to activate signaling in the absence of ligand, than are able to abrogate erlotinib binding to the comparably small ATP-binding cleft. This renders the identification of relevant mutations by their frequency of appearance difficult. Of course, every mutation enriched in a library has to be tested independently to confirm any observed activating or resistance effects. Still, it would improve the assay if the frequency of "true" mutations could be raised beyond the levels of artifacts and mutations abrogating the function of EGFR.

Furthermore, the observation was made that the mutation A289V, which is known for its activating properties [33], had a frequency of over 1% in every activating library but was still not 8-fold enriched compared to the loss-of-function library. Naturally, the question arises if this mutation did in fact contribute to the activating phenotype of the libraries and was just not enriched compared to the loss-of-function library for some reason, or if the mutation indeed exhibited no activating phenotype in this setting. A higher enrichment of a phosphorylation-positive subpopulation of cells would help to answer this question, assuming that relevant mutations would then be present at a higher frequency.

Usually, to obtain a higher enrichment, more selection rounds would be conducted. However, this would also mean to increase the artifacts introduced by PCR. Therefore, the more stringent adjustment of the sort gate to achieve a higher enrichment from the beginning should be considered.

5.7 Functional and structural consideration of enriched mutations

It is evident that, especially in the light of the above discussed aspects, every mutation enriched in the selected libraries requires to be cloned and tested for its apparent function. Nevertheless, certain mutations, which are either known from the literature or are located at structurally interesting positions, will be discussed in the following.

5.7.1 Activating mutations

The mutations that were present in the libraries selected for activating mutations are spaced throughout the whole protein, as can be seen in Table 4.4. Besides the ubiquitous T790M substitution and the mutations located in the cetuximab/EGF epitope, a few other interesting positions showed amino acid exchanges.

One of those is a glycine to alanine exchange at position 601, detected at a frequency of 1.6% in the activating library selected with the pY998-specific antibody in the first experiment. According to the crystal structure of the inactive (i.e. tethered) extracellular module of EGFR, this mutation is located at the interface between domains II and IV forming the tether (Fig. 5.2 (b)). Even though the glycine is only substituted by a small, non-polar amino acid, the mutation might have an effect on the conformation of the tether, which in turn can influence the equilibrium between the tethered and upright form of the extracellular module.

Another mutation in the extracellular module was found at a frequency of 1.0% in the activating library selected with the pY1092-specific antibody and consists of a substitution introducing a cysteine instead of a tyrosine at position 626. This position is located in domain IV and is adjacent to the transmembrane helix of EGFR. As can be seen in Figure 5.2 (c), according to the crystal structure of the upright dimer with bound EGF, this residue is placed at the dimerization interface of domain IV. Therefore, it can be assumed that introduction of a cysteine would lead to the formation of a disulfide bridge connecting the two EGFR molecules, which in turn potentially stabilizes the active dimeric conformation of the extracellular module.

The serine at position 442 was found to be mutated to arginine at a frequency of 1.1% in the cetuximab resistance library, but to isoleucine at a frequency of 1.2% in the selection for activating mutations using the pY1092-specific antibody. Therefore, the mutated position is also shown in this context, as can be seen in Figure 5.2 (a). The residue is not only located in the cetuximab epitope but also in the proximity of the EGF epitope. Besides the possibility that this mutation was co-selected and does not exhibit an activating phenotype, it might be possible that certain

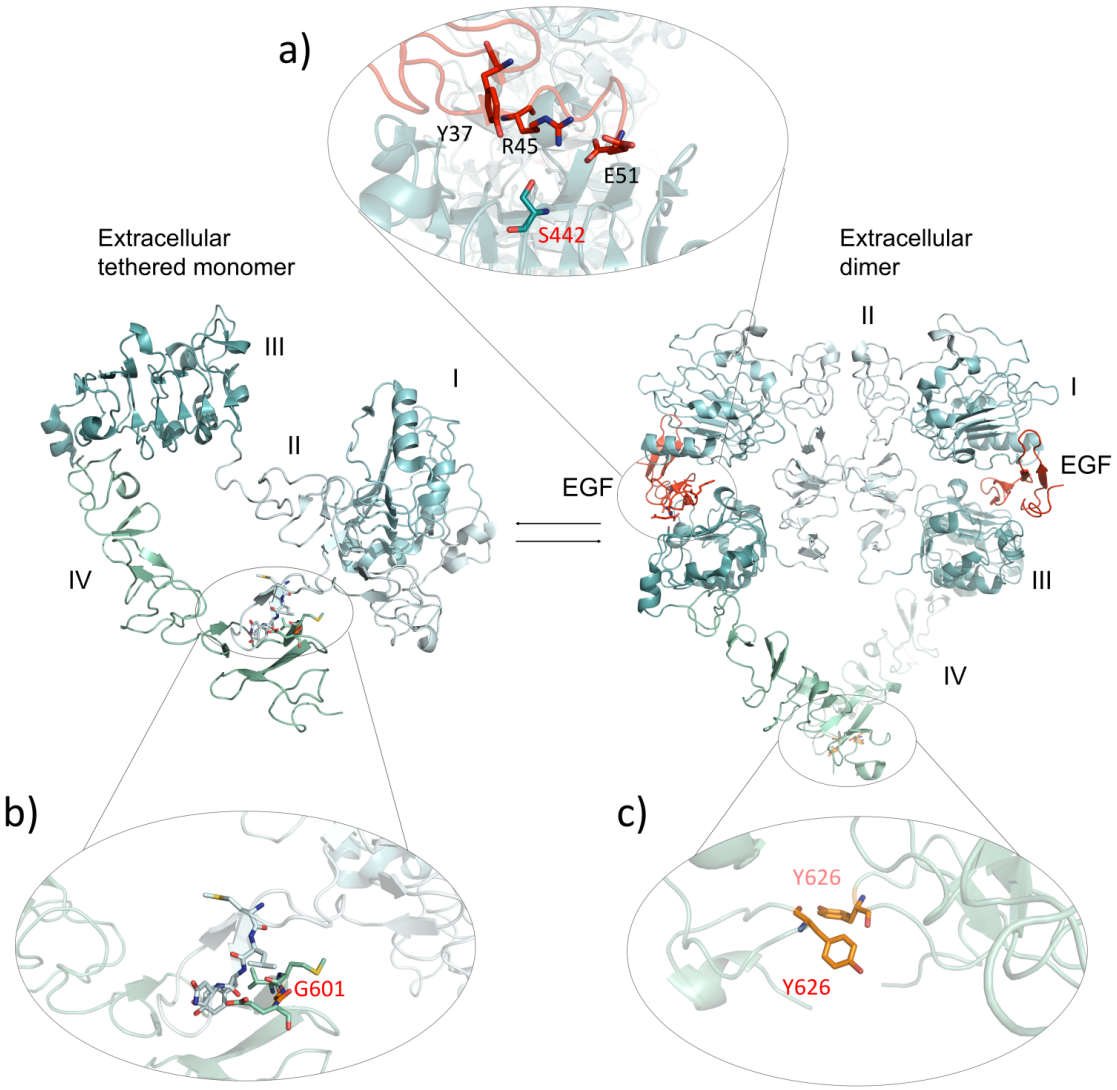

Figure 5.2: Potential activating mutations located in EGFR's extracellular module. The monomeric, tethered extracellular domain of EGFR (left, PDB 1NQL [34]) adopts an upright position upon binding of ligand EGF (red) and dimerizes via the dimerization arm of domain II (right, PDB 3NJP [35]). Potential activating mutations affecting the equilibrium of monomeric and dimeric EGFR are highlighted; **(a)** S442, which was found to be mutated to isoleucine in a library selected for activating phenotype, and adjacent residues from EGF **(b)** G601, which was found to be mutated to alanine, and surrounding residues **(c)** Y626, highlighted on both monomers dimerizing at domain IV, was found to be mutated to cysteine.

interactions with EGF increase the affinity for the ligand. As already mentioned, serum starvation did not exclude the presence of very low-level bovine EGF concentrations, which could be

a possible explanation for the occurrence of mutations affecting EGF-binding. Interestingly, the S442I mutation is listed once in the COSMIC, originating from whole exome sequencing of a sample from a lung cancer patient who received no prior treatment [36].

Downstream of the previous mutations in the *EGFR* gene, a mutation leading to the substitution of leucine by glutamine at position 658 was detected at a frequency of 2.4% in the activating library selected with the pY1092-specific antibody. This mutation is located in EGFR's transmembrane helix spanning residues 643 to 669. The transmembrane helix contains an N- and a C-terminal GxxxG-motif which serve as dimerization interface [5, 6]. It has been shown that the N-terminal dimerization is more stable and likely resembles the association of the active EGFR dimer, whereas C-terminal dimerization is associated with an inactive EGFR dimer [5, 6]. Supportive of those results is the finding that mutations stabilizing the C-terminal association and disrupting the N-terminal dimerization inhibit EGFR activity [5]. Interestingly, the found mutation is placed between the N-terminal (residues 648-653) and C-terminal (residues 661-665) dimerization interface. Clearly, a large, polar amino acid like glutamine tends to interact with other polar structures and avoids contact with the hydrophobic interior of the cell membrane. Consequently, it can be assumed that the L658Q mutation promotes dimerization of two transmembrane helices. However, due to the location of the mutation, it is unclear if it promotes the N-terminal or C-terminal dimerization or rather a completely different type of association.

As expected for the selection of activating mutations, some mutations were also enriched in the kinase domain. For example, in the library selected with the pY1092-specific antibody, the mutation S768I was detected at a frequency of 1.2%. Importantly, the S768I point mutation located in the kinase domain has been found more often in tumor samples, with 264 listings in the COSMIC (as of 18/10/2018). As can be seen in Figure 5.3 (a), the S768 residue is located at the dimerization interface of the activator and receiver kinase domain forming the asymmetric dimer. In a recent study from 2018, it was shown that S768 is a potential site for posttranslational phosphorylation and this modification helps to control the activity of EGFR [37]. An S768I mutant showed substantially higher phosphorylation levels of two different tyrosine residues compared

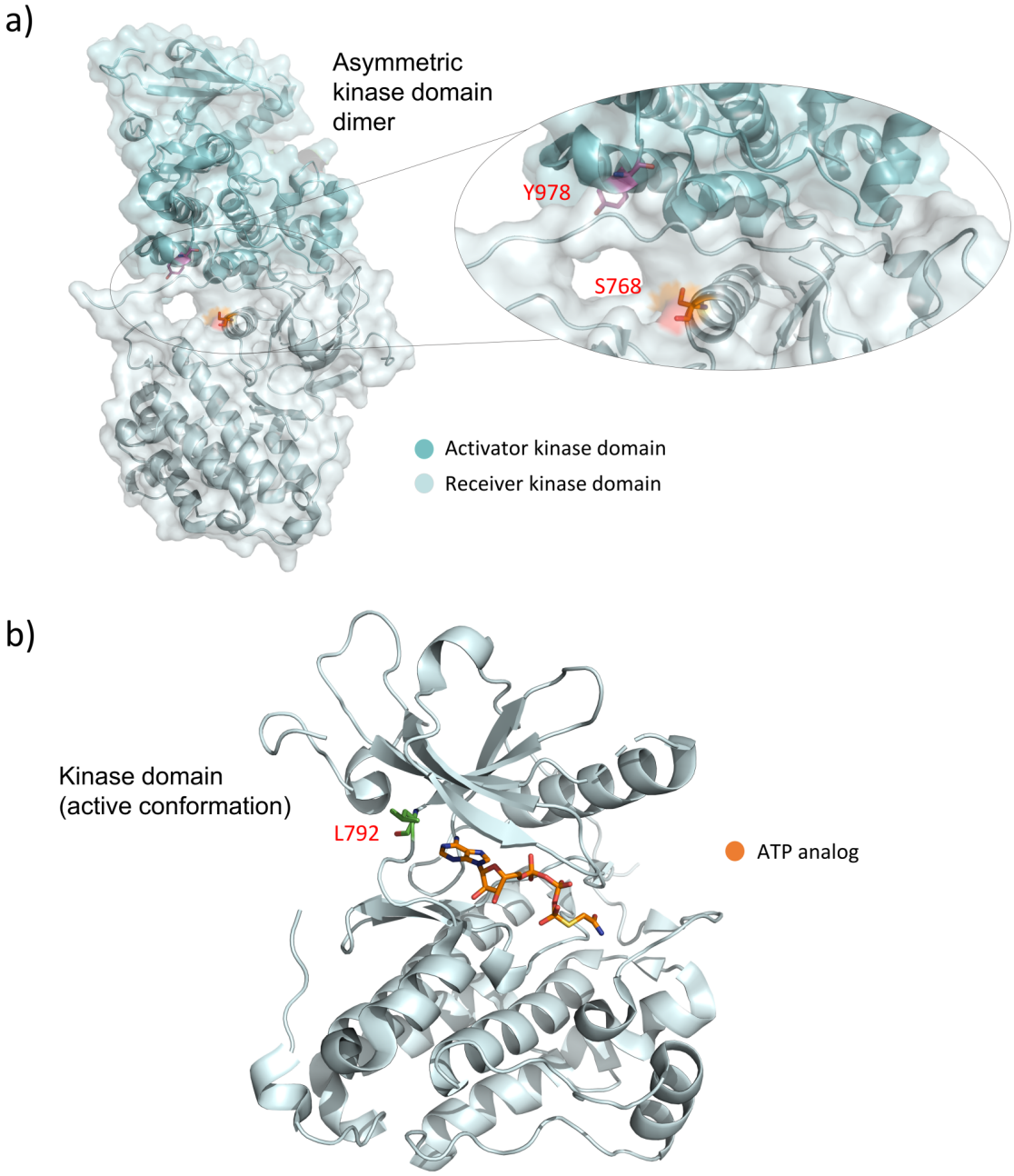

Figure 5.3: Potential activating mutations located in EGFR's kinase domain. **(a)** Residues found mutated in activating phenotype libraries at the interface of the asymmetric kinase domain dimer (PDB 2GS6 [9]) are highlighted in orange (S768 found mutated to isoleucin) and pink (Y978 found mutated to cysteine). **(b)** L792 (green), which was mutated to histidine in an activating phenotype library, is located in the ATP-binding cleft of the active kinase domain, which has an ATP analog-peptide conjugate bound (peptide moiety not shown here; PDB 2GS6 [9]).

to EGFR-WT as well as two other mutants containing negatively charged residues instead of the serine [37]. The latter were used to mimick the constitutive phosphorylation of S768. Regarding the inhibitory effect of this posttranslational modification, the authors suggested that phosphorylation of S768, which is located at the dimerization interface between receiver and activator kinase domain, obstructs this association [37]. It was assumed that the hydrophobic isoleucine not only impairs phosphorylation at position 768 but also enhances dimer formation by minimizing unfavourable polar contacts [37].

As shown in Figure 5.3 (a), another residue located at the asymmetric dimer interface of the two kinase domains was found to be mutated in a library selected for activating mutations. The tyrosine exchange to cysteine at position 978 was detected at a frequency of 1.7% in the library selected for activating mutations using the pY998-specific antibody from the first experiment. Due to its location it is possible that the mutation might activate EGFR by promoting asymmetric dimerization of the kinase domains.

Another interesting finding is the mutation L792H, detected at a frequency of 1.5% in the activating library selected in the first experiment with the pY998-specific antibody. This mutation is located at the ATP-binding cleft in the kinase domain and two residues from the well-described T790M mutation (Fig. 5.3 (b)). Surprisingly, mutations at this position have been associated with acquired *in vivo* as well as *in vitro* osimertinib (AZD9291) resistance in combination with the activating mutation L858R and the secondary resistance mutation T790M [38, 39]. *In vitro* experiments showed that the mutation also confers low-level osimertinib resistance in combination with just the L858R mutation [38]. However, an L792 mutation to proline was also identified in two non-small cell lung cancer patients without other mutations in EGFR's exons 18 to 21 [40]. Certainly, the effect of the L792 mutation alone, which resembles the situation in our selection, remains to be elucidated.

5.7.2 Cetuximab resistance mutations

Cetuximab resistance mutations are expected to be located in the cetuximab epitope on domain III of EGFR's extracellular module, thereby interfering with the binding of this antibody. However, as can be seen in Table 4.3, also other mutations, mainly in the extracellular module and C-terminal tail, met the 1% frequency and 8-fold enrichment threshold. This was possibly, at least partly, caused by the procedure of selection, as described in Section 5.4.

Nevertheless, the mutations S442R and K467N, both found at a frequency of 1.1% in the cetuximab resistance library, are located in the cetuximab epitope (Fig. 5.4 (a)). The lysine residue

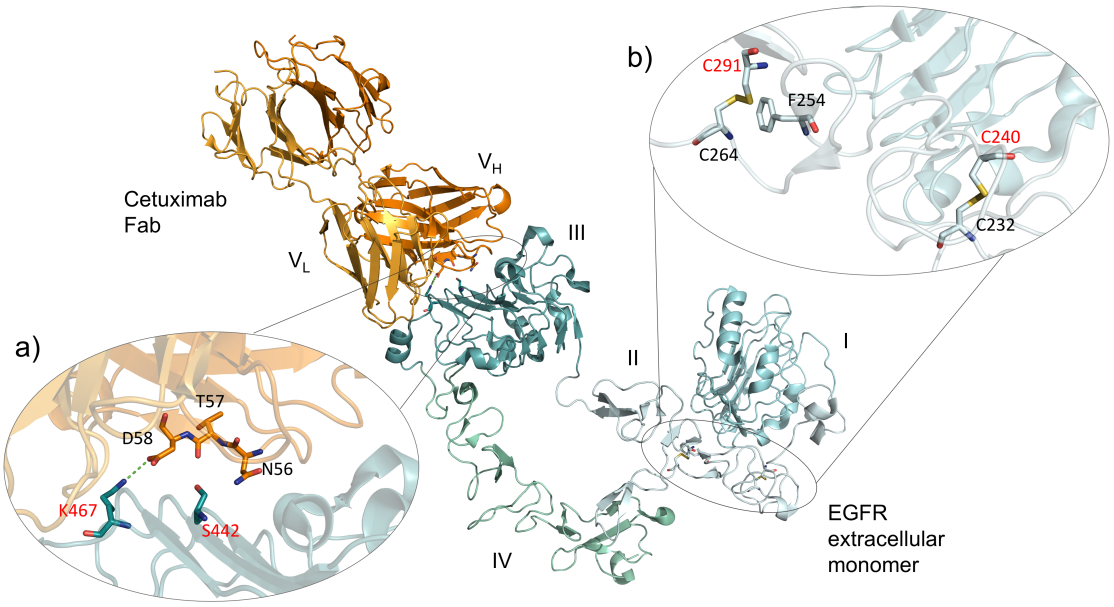

Figure 5.4: Potential cetuximab resistance mutations located in EGFR's extracellular module. The monomeric, tethered extracellular module of EGFR is shown with the cetuximab Fab bound to domain III (PDB 1YY9 [23]). **(a)** K467 and S442, found mutated to asparagine and arginine, respectively, in the cetuximab resistance library, are shown with adjacent residues from the cetuximab Fab. The ionic contact between K467 and D58, potentially interrupted by the K467N mutation, is indicated with a green dashed line. **(b)** The two disulfide bridges, which would be removed by substitution of the cysteine by serine at position 240 and phenylalanine at position 291, as found in the cetuximab resistance library, are shown alongside the corresponding cysteines forming the disulfide bridges and another phenylalanine in close vicinity of the introduced phenylalanine.

at position 467 was found to be mutated to threonine in a colorectal cancer patient with acquired resistance to cetuximab [24]. The abrogated binding of cetuximab to the K467T mutant was also confirmed *in vitro* in that study. Furthermore, the authors suggested that the removal of the positive charge which interacts with a negatively charged residue from cetuximab, as can be seen in Fig. 5.4 (a), is likely the mechanism of resistance [24]. In turn, this indicates that the introduction of another charge-neutral amino acid like asparagine exhibits a similar effect.

Figure 5.4 (a) also shows the position of the serine residue (S442), which was found to be mutated to arginine at a frequency of 1.1% in the cetuximab resistance selection, and adjacent residues at the interface of EGFR's domain III and cetuximab. The introduction of a large, positively charged amino acid like arginine might sterically interfere with the surrounding residues of the antibody and thereby disrupt other favorable contacts. As already stated, the position 442 was also mutated to isoleucine in the library selected for activating mutations using the pY1092-specific antibody at a frequency of 1.2%.

Another interesting observation, highlighted in Figure 5.4 (b), is the removal of two cysteine bridges in the tether-forming domain II by the mutations C240S and C291F. Those mutations were found at frequencies of 1.0% for C240S and 1.7% for C291F in the cetuximab resistance library. A mutation from cysteine to tyrosine at position 240 can be found in the COSMIC in a sample of a glioblastoma patient, which was analyzed in the course of a genomic characterization project conducted by "The Cancer Genome Atlas" (TCGA) [41].

5.7.3 Erlotinib resistance mutations

The selection for erlotinib-resistant EGFR variants clearly resembles the erlotinib resistance occurring in the clinic, where in a substantial fraction of the cases the T790M mutation appears after initiation of treatment with the inhibitor [26]. Besides T790M, which was present at a frequency of 20.5% and 18.6%, respectively, in the two erlotinib resistance libraries, another mutation in the kinase domain was detected.

This mutation from alanine to threonine at position 750 was present at a frequency of 1.1% in the erlotinib resistance library from the first experiment (Table 4.2). However, the same mutation was not enriched above the defined thresholds in the erlotinib resistance library from the second experiment. Nevertheless, the mutation is listed once in the COSMIC. The COSMIC entry originates from a study in which the outcome of routine EGFR screening in non-small cell lung cancer patients was retrospectively analyzed [42]. The A750T mutation was found in only one patient prior to TKI treatment and is listed as not previously described mutation. Furthermore, the patient harboring this mutation showed partial response to TKI treatment, but it is not stated if gefitinib or erlotinib had been administered. Moreover, the preceding treatment schedule for the individual patient is not declared [42]. Figure 5.5 shows the location of the mutated amino acid in the crystal structure of the kinase domain in the active conformation in complex with erlotinib. The mutation is located in the loop structure connecting a ß-strand with the so-called αC helix, which is oriented towards the ATP-binding cleft in the active conformation, in the N-lobe

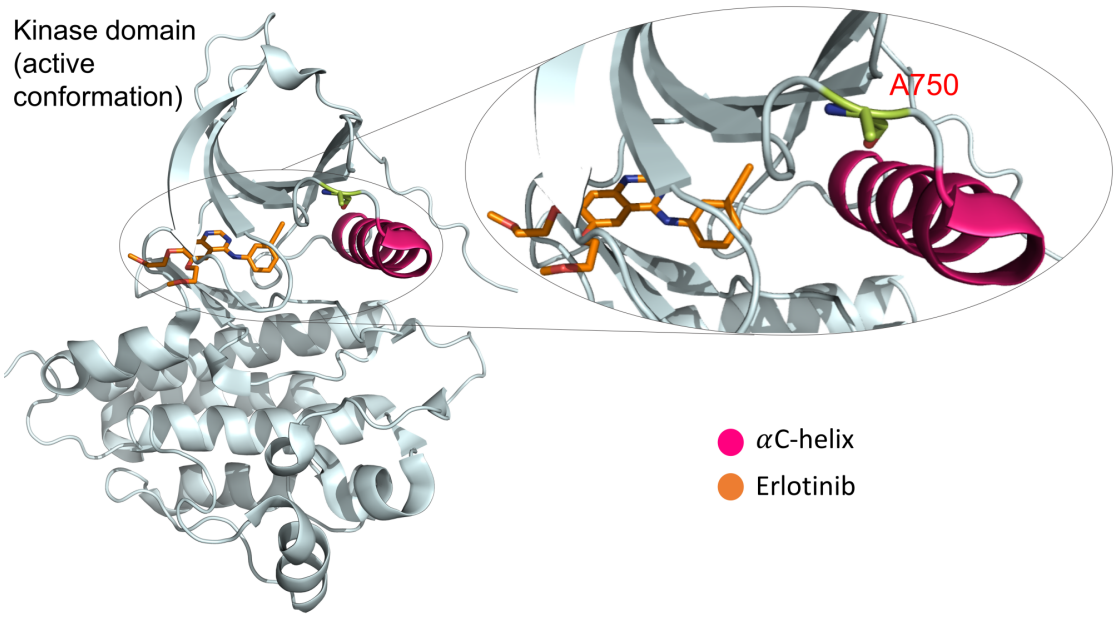

Figure 5.5: Potential erlotinib resistance mutation located in EGFR's kinase domain. A750, found mutated to threonine in an erlotinib resistance library, is shown in EGFR's kinase domain in the active conformation with bound erlotinib (PDB 1M17 [25]). The residue is located in the loop connecting a ß-strand with the so-called αC-helix.

of the kinase domain. Certainly, it is not directly evident if the substitution of the hydrophobic alanine to the polar threonine in this loop structure has any effect on the active conformation of the kinase domain or the ability of erlotinib to bind to the molecule.

In addition, several mutations were enriched in the C-terminal tail of EGFR in the erlotinib resistance library from the second experiment (I965I, V1010A, K1179E, N1194T and A1201A) as can be seen in Table 4.2. Notably, the silent mutation at position 1201 (A1201A) is listed three times in the COSMIC. However, those mutations were only present in one of the two libraries selected for erlotinib resistance. Furthermore, due to the structural flexibility of the tail, no complete crystal structure is available up to date.

CHAPTER 6

Concluding Remarks and Outlook

Despite the ever increasing insight on mutations causing aberrantly signaling EGFR, for many rare mutations, often only found in individual cancer patients, functional data is lacking. Additionally, the almost always occurring resistance to various EGFR inhibitors remains a substantial obstacle for successful cancer treatment [15]. In the present thesis randomly mutated *EGFR* libraries were selected for mutations conferring ligand-independent signaling (activating mutations), cetuximab resistance as well as erlotinib resistance using mammalian cell display and fluorescence activated cell sorting. The library selected for loss-of-function mutations was eventually only used as control to exclude mutational artifacts most likely introduced by PCR amplification of *EGFR* between selections. The produced libraries were sequenced by Illumina next generation sequencing and a subset of enriched mutations (above the set threshold of 1% frequency and 8-fold enrichment compared to the loss-of-function library) was examined in the available crystal structures. Certainly, every mutation still requires individual testing of its apparent phenotype. In this context, the relative enrichment of mutations in terms of frequency remains to be optimized, potentially by the application of more stringent sort gates. Also the selection for cetuximab resistance mutations with an improved protocol for the selection pressure application, excluding the possibility to simultaneously enrich activating mutations, is subject for future research. Furthermore, the suitability of phosphorylation detection with the pY1092-specific antibody, which produced superior results compared to the pY998-specific antibody in

the case of activating mutations, for the selection of other phenotypes remains to be tested. Another valuable addition to the assay would be the establishment of a method for reliable quantification of *EGFR* library plasmids isolated from sorted HEK 293T cells. Moreover, all produced libraries have to be compared in a single flow cytometry experiment for comparability. Furthermore, if it would be possible to omit the PCR amplification steps between library selections by improvement of *E. coli* transformation efficiency, also the introduction of PCR artifacts would be reduced. Concluding, it can be said that the first results from the evaluation of the developed assay gave valuable insights on its applicability and also revealed certain aspects which require further research. Still, it is fair to say that it was possible to enrich EGFR variants for different phenotypes and obtain a set of mutations, which potentially contributes to the elucidation of the mutational landscape of EGFR in cancer.

Acronyms

E. coli *Escherichia coli*.

bps Base pairs.

COSMIC Catalogue of somatic mutations in cancer.

DMEM Dulbecco's modified eagles medium.

DMSO Dimethylsulfoxide.

EGF Epidermal growth factor.

EGFR Epidermal growth factor receptor.

EGFR-L858R EGFR containing the activating mutation L858R.

EGFR-WT EGFR wild type.

EGFRvIII EGFR variant III.

FACS Fluorescence-activated cell sorting.

FBS Fetal bovine serum.

HEK 293T Human embryonic kidney 293T cells.

JM-A Juxtamembrane segment A.

JM-B Juxtamembrane segment B.

mAb Monoclonal antibody.

NSCLC Non-small cell lung cancer.

PBS Dulbecco's phosphate buffered saline.

PDB Protein data bank.

pY1092 Phospho-tyrosine at amino acid position 1092 in EGFR.

pY998 Phospho-tyrosine at amino acid position 998 in EGFR.

TGF-α Transforming growth factor-alpha.

TKI Tyrosine kinase inhibitor.

List of Figures

List of Tables

Bibliography

[1] D. Hanahan and R. A. Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674, 2011.

[2] Y. Yarden and G. Pines. The ERBB network: at last, cancer therapy meets systems biology. *Nature Reviews Cancer*, 12(8):553–563, 2012.

[3] C. L. Arteaga and J. A. Engelman. ERBB receptors: from oncogene discovery to basic science to mechanism-based cancer therapeutics. *Cancer cell*, 25(3):282–303, 2014.

[4] E. Kovacs, J. A. Zorn, Y. Huang, T. Barros, and J. Kuriyan. A structural perspective on the regulation of the epidermal growth factor receptor. *Annual review of biochemistry*, 84:739–764, 2015.

[5] N. F. Endres, R. Das, A. W. Smith, A. Arkhipov, E. Kovacs, Y. Huang, J. G. Pelton, Y. Shan, D. E. Shaw, D. E. Wemmer, et al. Conformational coupling across the plasma membrane in activation of the EGF receptor. *Cell*, 152(3):543–556, 2013.

[6] A. Arkhipov, Y. Shan, R. Das, N. F. Endres, M. P. Eastwood, D. E. Wemmer, J. Kuriyan, and D. E. Shaw. Architecture and membrane interactions of the EGF receptor. *Cell*, 152(3):557–569, 2013.

[7] N. Jura, N. F. Endres, K. Engel, S. Deindl, R. Das, M. H. Lamers, D. E. Wemmer, X. Zhang, and J. Kuriyan. Mechanism for activation of the EGF receptor catalytic domain by the juxtamembrane segment. *Cell*, 137(7):1293–1307, 2009.

[8] S. McLaughlin, S. O. Smith, M. J. Hayman, and D. Murray. An electrostatic engine model for autoinhibition and activation of the epidermal growth factor receptor (EGFR/ErbB) family. *The Journal of general physiology*, 126(1):41–53, 2005.

[9] X. Zhang, J. Gureasko, K. Shen, P. A. Cole, and J. Kuriyan. An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor. *Cell*, 125(6):1137–1149, 2006.

[10] A. Citri and Y. Yarden. EGF-ERBB signalling: towards the systems level. *Nature Reviews Molecular cell biology*, 7(7):505, 2006.

[11] D. M. Freed, N. J. Bessman, A. Kiyatkin, E. Salazar-Cavazos, P. O. Byrne, J. O. Moore, C. C. Valley, K. M. Ferguson, D. J. Leahy, D. S. Lidke, et al. EGFR ligands differentially stabilize receptor dimers to specify signaling kinetics. *Cell*, 171(3):683–695, 2017.

[12] Y. Huang, S. Bharill, D. Karandur, S. M. Peterson, M. Marita, X. Shi, M. J. Kaliszewski, A. W. Smith, E. Y. Isacoff, and J. Kuriyan. Molecular basis for multimerization in the activation of the epidermal growth factor receptor. *Elife*, 5:e14107, 2016.

[13] W. Pao, V. Miller, M. Zakowski, J. Doherty, K. Politi, I. Sarkaria, B. Singh, R. Heelan, V. Rusch, L. Fulton, et al. EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proceedings of the National Academy of Sciences*, 101(36):13306–13311, 2004.

[14] T. J. Lynch, D. W. Bell, R. Sordella, S. Gurubhagavatula, R. A. Okimoto, B. W. Brannigan, P. L. Harris, S. M. Haserlat, J. G. Supko, F. G. Haluska, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *New England Journal of Medicine*, 350(21):2129–2139, 2004.

[15] C. R. Chong and P. A. Jänne. The quest to overcome resistance to EGFR-targeted therapies in cancer. *Nature medicine*, 19(11):1389–1400, 2013.

[16] S. Wang, Y. Song, and D. Liu. EAI045: the fourth-generation EGFR inhibitor overcoming T790M and C797S resistance. *Cancer letters*, 385:51–54, 2017.

[17] H. Patel, R. Pawara, A. Ansari, and S. Surana. Recent updates on third generation EGFR inhibitors and emergence of fourth generation EGFR inhibitors to combat C797S resistance. *European journal of medicinal chemistry*, 142:32–47, 2017.

[18] S. A. Forbes, D. Beare, H. Boutselakis, S. Bamford, N. Bindal, J. Tate, C. G. Cole, S. Ward, E. Dawson, L. Ponting, R. Stefancsik, B. Harsha, C. Y. Kok, M. Jia, H. Jubb, Z. Sondka, S. Thompson, T. De, and P. J. Campbell. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Research*, 45(D1):D777–D783, 2017.

[19] H. K. Gan, A. N. Cvrljevic, and T. G. Johns. The epidermal growth factor receptor variant III (EGFRvIII): where wild things are altered. *The FEBS journal*, 280(21):5350–5370, 2013.

[20] K. D. Carey, A. J. Garton, M. S. Romero, J. Kahler, S. Thomson, S. Ross, F. Park, J. D. Haley, N. Gibson, and M. X. Sliwkowski. Kinetic analysis of epidermal growth factor receptor somatic mutant proteins shows increased sensitivity to the epidermal growth factor receptor tyrosine kinase inhibitor, erlotinib. *Cancer research*, 66(16):8163–8171, 2006.

[21] C.-H. Yun, T. J. Boggon, Y. Li, M. S. Woo, H. Greulich, M. Meyerson, and M. J. Eck. Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity. *Cancer cell*, 11(3):217–227, 2007.

[22] Y. Shan, M. P. Eastwood, X. Zhang, E. T. Kim, A. Arkhipov, R. O. Dror, J. Jumper, J. Kuriyan, and D. E. Shaw. Oncogenic mutations counteract intrinsic disorder in the EGFR kinase and promote receptor dimerization. *Cell*, 149(4):860–870, 2012.

[23] S. Li, K. R. Schmitz, P. D. Jeffrey, J. J. Wiltzius, P. Kussie, and K. M. Ferguson. Structural basis for inhibition of the epidermal growth factor receptor by cetuximab. *Cancer cell*, 7(4):301–311, 2005.

[24] S. Arena, B. Bellosillo, G. Siravegna, A. Martínez, I. Cañadas, L. Lazzari, N. Ferruz, M. Russo, S. Misale, I. González, et al. Emergence of multiple EGFR extracellular

mutations during cetuximab treatment in colorectal cancer. *Clinical cancer research*, 21(9):2157–2166, 2015.

[25] J. Stamos, M. X. Sliwkowski, and C. Eigenbrot. Structure of the epidermal growth factor receptor kinase domain alone and in complex with a 4-anilinoquinazoline inhibitor. *Journal of Biological Chemistry*, 277(48):46265–46272, 2002.

[26] H. Yu, M. E. Arcila, N. Rekhtman, C. S. Sima, M. F. Zakowski, W. Pao, M. G. Kris, V. A. Miller, M. Ladanyi, and G. J. Riely. Analysis of mechanisms of acquired resistance to EGFR TKI therapy in 155 patients with EGFR-mutant lung cancers. *Clinical cancer research*, 19(8):2240–2247, 2013.

[27] C.-H. Yun, K. E. Mengwasser, A. V. Toms, M. S. Woo, H. Greulich, K.-K. Wong, M. Meyerson, and M. J. Eck. The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. *Proceedings of the National Academy of Sciences*, 105(6):2070–2075, 2008.

[28] W. Pao, V. A. Miller, K. A. Politi, G. J. Riely, R. Somwar, M. F. Zakowski, M. G. Kris, and H. Varmus. Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain. *PLoS medicine*, 2(3):e73, 2005.

[29] E. L. Kwak, R. Sordella, D. W. Bell, N. Godin-Heymann, R. A. Okimoto, B. W. Brannigan, P. L. Harris, D. R. Driscoll, P. Fidias, T. J. Lynch, et al. Irreversible inhibitors of the EGF receptor may circumvent acquired resistance to gefitinib. *Proceedings of the National Academy of Sciences*, 102(21):7665–7670, 2005.

[30] M. R. Brewer, S. H. Choi, D. Alvarado, K. Moravcevic, A. Pozzi, M. A. Lemmon, and G. Carpenter. The juxtamembrane region of the EGF receptor functions as an activation domain. *Molecular cell*, 34(6):641–651, 2009.

[31] M. Teufl. Establishing an assay for studying the molecular mechanisms of drug resistance in cancer. *University of Natural Resources and Life Sciences, Vienna, Department of Chemistry, Division of Biochemistry*, 2017.

[32] D. A. Cross, S. E. Ashton, S. Ghiorghiu, C. Eberlein, C. A. Nebhan, P. J. Spitzler, J. P. Orme, M. R. V. Finlay, R. A. Ward, M. J. Mellor, et al. AZD9291, an irreversible EGFR TKI, overcomes T790M-mediated resistance to EGFR inhibitors in lung cancer. *Cancer discovery*, 4(9):1046–1061, 2014.

[33] J. C. Lee, I. Vivanco, R. Beroukhim, J. H. Huang, W. L. Feng, R. M. DeBiasi, K. Yoshimoto, J. C. King, P. Nghiemphu, Y. Yuza, et al. Epidermal growth factor receptor activation in glioblastoma through novel missense mutations in the extracellular domain. *PLoS medicine*, 3(12):e485, 2006.

[34] K. M. Ferguson, M. B. Berger, J. M. Mendrola, H.-S. Cho, D. J. Leahy, and M. A. Lemmon. EGF activates its receptor by removing interactions that autoinhibit ectodomain dimerization. *Molecular cell*, 11(2):507–517, 2003.

[35] C. Lu, L.-Z. Mi, M. J. Grey, J. Zhu, E. Graef, S. Yokoyama, and T. A. Springer. Structural evidence for loose linkage between ligand binding and kinase activation in the epidermal growth factor receptor. *Molecular and cellular biology*, 30(22):5432–5443, 2010.

[36] C. Vinayanuwattikun, F. Le Calvez-Kelm, B. Abedi-Ardekani, D. Zaridze, A. Mukeria, C. Voegele, M. Vallée, D. Purnomosari, N. Forey, G. Durand, et al. Elucidating genomic characteristics of lung cancer progression from in situ to invasive adenocarcinoma. *Scientific reports*, 6:31628, 2016.

[37] L.-C. Huang, K. E. Ross, T. R. Baffi, H. Drabkin, K. J. Kochut, Z. Ruan, P. D'Eustachio, D. McSkimming, C. Arighi, C. Chen, et al. Integrative annotation and knowledge discovery of kinase post-translational modifications and cancer-associated mutations through federated protein ontologies and resources. *Scientific reports*, 8:6518, 2018.

[38] Z. Yang, N. Yang, Q. Ou, Y. Xiang, T. Jiang, X. Wu, H. Bao, X. Tong, X. Wang, Y. W. Shao, et al. Investigating novel resistance mechanisms to third-generation EGFR tyrosine kinase inhibitor osimertinib in non-small cell lung cancer patients. *Clinical Cancer Research*, 24(13):3097–3107, 2018.

[39] K. Chen, F. Zhou, W. Shen, T. Jiang, X. Wu, X. Tong, Y. W. Shao, S. Qin, and C. Zhou. Novel mutations on EGFR Leu792 potentially correlate to acquired resistance to osimertinib in advanced NSCLC. *Journal of Thoracic Oncology*, 12(6):e65–e68, 2017.

[40] M.-S. Tsao, A. Sakurada, J.-C. Cutz, C.-Q. Zhu, S. Kamel-Reid, J. Squire, I. Lorimer, T. Zhang, N. Liu, M. Daneshmand, et al. Erlotinib in lung cancer-molecular and clinical predictors of outcome. *New England Journal of Medicine*, 353(2):133–144, 2005.

[41] Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061, 2008.

[42] M. Locatelli-Sanchez, S. Couraud, D. Arpin, R. Riou, P.-P. Bringuier, and P.-J. Souquet. Routine EGFR molecular analysis in non-small-cell lung cancer patients is feasible: exons 18-21 sequencing results of 753 patients and subsequent clinical outcomes. *Lung*, 191(5):491–499, 2013.