Dissertation submitted to the
Combined Faculty of Natural Sciences and Mathematics
of the Ruberto Carola University Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

Presented by
M.Sc. Konrad Herbst
born in Potsdam

Oral examination: December 2nd, 2021

Scalable approaches for gene tagging and genome walking sequencing

Referees: Prof. Dr. Michael Knop
Prof. Dr. Lars M. Steinmetz

Contents

Abstract

A central question in the life sciences is how an observed phenotype is realized by a given biological system. One explanatory factor is certainly the genotype of such a system which can easily be characterized due to the development of high-throughput next-generation sequencing (NGS) technologies. This allows to generate many hypotheses which are best tested by manipulating genotypes and observing the resulting change in phenotype. However, respective high-throughput technologies for genotype manipulation are lagging behind.

This thesis presents technical advancements in the field of genome engineering and next-generation sequencing which allow to construct and characterize collections of mutants with tagged genes.

First, an improved version of a targeted NGS strategy is introduced, which is termed Tn5-Anchor-Seq. This protocol for genome walking sequencing allowed to characterize unknown sequences adjacent to known genomic sites so that for example tag integrations could be mapped. It builds upon the concept of tagmentation and was designed with scalability, efficiency and sensitivity in mind.

Second, CRISPR-Cas12a-assisted tag library engineering (CASTLING) is presented as a high-throughput pooled strategy for gene tagging in the yeast *Saccharomyces cerevisiae*. It was implemented and validated using a set of 215 genes which were simultaneously targeted. Furthermore, genome-wide targeting was explored revealing that $\sim$50 % of yeast genes can be covered within one single experiment. Factors important for further application of this technology were identified and are discussed.

Third, the insights gained during the development of the CASTLING strategy motivated the application of their concepts for single gene tagging in mammalian cells. Usually, mammalian gene targeting is relatively inefficient and laborious. Therefore, a convenient CRISPR-Cas12a-assisted PCR tagging strategy was developed. Several targeted NGS approaches including Tn5-Anchor-Seq supported the validation of this technology. Furthermore, these analyses allowed the characterization of an experimental artifact associated with mammalian PCR tagging which can most likely be explained by aberrant tag expression.

Finally, Tn5-Anchor-Seq was applied to the characterization of a diagnostic RT-LAMP assay for SARS-CoV-2 detection. The high scalability of the resulting LAMP-sequencing protocol allowed to sequence RT-LAMP reactions from 768 patient samples and by that

helped to validate the sensitivity and specificity of this assay. This was important for deploying additional testing capacity during the COVID-19 pandemic.

In conclusion, this thesis introduces and showcases scalable approaches for pooled and single gene taggings in yeast and mammalian cells. In this context also an improved genome walking procedure was implemented which furthermore supported the establishment of a diagnostic assay for SARS-CoV-2 detection.

Zusammenfassung

Eine zentrale Frage in den Biowissenschaften ist, wie ein beobachteter Phänotyp durch ein gegebenes biologisches System realisiert wird. Ein mögliche Erklärung ist sicherlich der Genotyp eines solchen Systems, der dank der Entwicklung von Technologien zur Hochdurchsatz-Sequenzierung (NGS) leicht charakterisiert werden kann. Dadurch lassen sich viele Hypothesen aufstellen, die am ehesten durch die Manipulation der Genotypen und der Beobachtung der daraus resultierenden Phänotypen überprüft werden können. Allerdings sind die entsprechenden Hochdurchsatztechnologien zur Manipulation von Genotypen nur bedingt vorhanden.

In der vorliegenden Arbeit werden technische Fortschritte auf den Gebieten der Genome-Editierung und der Hochdurchsatzsequenzierung vorgestellt, die es ermöglichen, Sammlungen von Mutanten mit gentechnisch markierten Genen zu erstellen und zu charakterisieren.

Zunächst wird eine verbesserte Version einer gezielten Strategie zur Hochdurchsatzsequenzierung eingeführt, die als Tn5-Anchor-Seq bezeichnet wird. Mit diesem Protokoll können unbekannte Sequenzen charakterisiert werden, die neben bekannten genomischen Bereichen liegen, auch Genom-Walking-Verfahren genannt. Damit können zum Beispiel Integrationen heterologer Sequenzen kartiert werden. Es baut auf dem Konzept der Tagmentierung auf und wurde mit Blick auf Skalierbarkeit, Effizienz und Sensitivität entworfen.

Zweitens wird das CRISPR-Cas12a-assisted tag library engineering (CASTLING) als Hochdurchsatz-Strategie für das gemischte Herstellen von Integrationsmutanten der Hefe *Saccharomyces cerevisiae* vorgestellt. Sie wurde anhand einer Auswahl von 215 gleichzeitig zu verändernden Genen implementiert und validiert. Darüber hinaus wurde untersucht, ob genomweit alle Gene gleichzeitig zugänglich sind, wobei sich zeigte, dass ∼50 % der Hefegene in einem einzigen Experiment verändert werden konnten. Faktoren, die für die weitere Anwendung dieser Technologie wichtig sind, wurden identifiziert und werden diskutiert.

Drittens motivierten die bei der Entwicklung der CASTLING-Strategie gewonnenen Erkenntnisse die Anwendung ihrer Konzepte für die gentechnische Markierung einzelner Gene in Säugetierzellen. Normalerweise ist das Gen-Targeting in Säugetierzellen relativ ineffizient und aufwändig. Daher wurde eine praktische CRISPR-Cas12a-gestützte Strategie zum PCR tagging (PCR-gestützen Genmarkierung) entwickelt. Mehrere gezielte

Ansätze der Hochdurchsatzsequenzierungs, darunter Tn5-Anchor-Seq, unterstützten die Validierung dieser Technologie. Darüber hinaus ermöglichten diese NGS Analysen die Charakterisierung eines experimentellen Artefakts, welches im Zusammenhang mit dem PCR tagging in Säugetierzellen auftrat und welches höchstwahrscheinlich durch eine unbeabsichtigte Expression des Markierungsgens erklärt werden kann.

Schließlich wurde Tn5-Anchor-Seq für die Charakterisierung eines diagnostischen Assays basierend auf reverse Transkriptase Schleifen-vermittelter isothermaler Amplifikation (RT-LAMP) für den Nachweis von SARS-CoV-2 eingesetzt. Die hohe Skalierbarkeit des resultierenden LAMP-sequencing genannten Protokolls ermöglichte die Sequenzierung von RT-LAMP-Reaktionen von 768 Patientenproben. Damit konnte die Sensitivität und Spezifität dieses Assays validiert werden. Dies war wichtig für die Bereitstellung zusätzlicher Testkapazitäten während der COVID-19-Pandemie.

Zusammenfassend befasst sich diese Arbeit mit skalierbaren Ansätzen für gemischte und einzelne gentechnische Markierungen von Genen in Hefe- und Säugetierzellen. In diesem Zusammenhang wurde auch ein verbessertes Genom-Walking-Verfahren implementiert, das darüber hinaus die Etablierung eines diagnostischen Assays zum Nachweis von SARS-CoV-2 unterstützt hat.

List of Figures

List of Tables

List of Abbreviations

DSB double strand break

HR homologous recombination

NHEJ non-homologous end joining

NGS next-generation sequencing

gDNA genomic DNA

ssDNA single-stranded DNA

dsDNA double-stranded DNA

ORF open reading frame

CDS coding sequence

UMI unique molecular identifier

RT-qPCR reverse transcription quantitative PCR

RT-LAMP reverse transcription loop-mediated isothermal amplification

CT threshold cycle

List of Publications

Related to thesis

- **Herbst, K.**, Meurer, M., Kirrmaier, D., Anders, S., Knop, M., Thi, V.L.D. (2021). Colorimetric RT-LAMP and LAMP-sequencing for Detecting SARS-CoV-2 RNA in Clinical Samples. *Bio-Protocol*, 11 (6), e3964-e3964, doi:10.21769/BioProtoc.3964

  – **first authorship**

- Thi, V.L.D.. **Herbst, K.***, Boerner, K.*, Meurer, M.*, Kremer, L. P. M., Kirrmaier, D., Freistaedter, A., Papagiannidis, D., Galmozzi, C., Stanifer, M. L., Boulant, S., Klein, S., Chlanda, P., Khalid, D., Miranda, I. B., Schnitzler, P., Kräusslich, H.-G., Knop, M., Anders, S. (2020). A colorimetric RT-LAMP assay and LAMP-sequencing for detecting SARS-CoV-2 RNA in clinical samples. *Science Translational Medicine*, 12(556), doi:10.1126/scitranslmed.abc7075

  – ***shared second authorship**

- Fueller, J*. **Herbst, K.***, Meurer, M.*, Gubicza, K., Kurtulmus, B., Knopf, J. D., Kirrmaier, D., Buchmuller, B. C., Pereira, G., Lemberg, M. K., Knop, M. (2020). CRISPR-Cas12a-assisted PCR tagging of mammalian genes. *Journal of Cell Biology*, 219(6), doi:10.1083/jcb.201910210

  – ***shared first authorship**

- Buchmuller, B. C.*, **Herbst, K.***, Meurer, M., Kirrmaier, D., Sass, E., Levy, E. D., Knop, M. (2019). Pooled clone collections by multiplexed CRISPR-Cas12a-assisted gene tagging in yeast. *Nature Communications*, 10, 2960, doi:10.1038/s41467-019-10816-7

  – ***shared first authorship**

- Meurer, M.*, Duan, Y.*, Sass, E.*, Kats, I., **Herbst,K.**, Buchmuller, B. C., Dederer, V., Huber, F., Kirrmaier, D., Štefl, M., Van Laer, K., Dick, T. P., Lemberg, M. K., Khmelinskii, A., Levy, E. D., Knop, M. (2018). Genome-wide C-SWAT library for high-throughput yeast genome tagging. *Nature Methods*, 15(18), 598-600, doi:10.1038/s41592-018-0045-8

  – **co-authorship**

Further contribution

- Štefl, M. **Herbst, K.**, Rübsam, M., Benda, A., Knop, M. (2020). Single-color Fluorescence Lifetime Cross-Correlation Spectroscopy *in vivo*. *Biophysical Journal*, 119(7), 1359-1370, doi:10.1016/j.bpj.2020.06.039

  – **second authorship**

Acknowledgments

First of all, I would like to thank Michael Knop who I not only see as my supervisor but also as my mentor. You gave me the opportunity to fall in love with yeast as one of the coolest model systems but moreover to learn what it means to be a scientist. In your lab I had the freedom to expose myself to a broad range of projects, approaches and ideas. In particular, your innovative and challenging thinking and our scientific and not-so-scientific discussions are highly stimulating to me. I am very honored for your continuous and encouraging support and guidance.

Furthermore, I would like to thank Lars Steinmetz for participating in my PhD development as a member of my thesis advisory committee and as referee for this thesis. You always and without hesitation found the time to help me and provide professional input which I consider as a great privilege. Not only as my third thesis advisory committee member I would also like to thank Simon Anders with whom I shared fascinating thoughts and time as a collaborator and as a friend.

Speaking of collaborators. Unfortunately there is not the space to thank all of them in length but I would especially like to mention Emmanuel Levy, Ehud Sass, Marius Lemberg, Andreas Decker and Viet Loan Dao Thi with whom working was an exceptional pleasure.

In addition to Michael, Matthias Meurer and Daniel Kirrmaier are the constants of the Knop lab. Both are great colleagues and exceptionally helpful; Matthias' insightful approach to a new project and experience is inspirational and Daniel's efficient and reliable way of working is a pleasure. Benjamin Buchmuller is a former group member and fellow graduate student with whom I enjoyed working a lot and whose eagerness and productivity was highly motivating. I had the honor to further publish work with two other former lab members, Julia Füller and Martin Štefl, and it was also their enthusiasm which made these studies possible. With Yuan Qiang Duan I pursued a project which unfortunately could not be fit into this thesis but it was a pleasure to work with him as well. Robin Burk is a former colleague with whom I shared work during the pandemic which was a particular stressful time but somehow he made it enjoyable and fun. With Krisztina Gubicza and Christian Reinbold I long shared the office and the lab which made working there especially comfortable. In addition, I also shared many running rounds with Christian and while this kept our bodies fit our engaging discussions about everything were as stimulating for the mind. I am very grateful for all the other Knop lab members who were such wonderful colleagues to me over the past years.

I would also like to thank the students I supervised during my PhD. Thanks to Carla Castignani, Jan Dohnálek, Nils Leibrock, Kaisa Pakari and Alvaro Mendoza for keeping up with me and I hope they learned at least as much as I did.

Ilia Kats is one close friend who I for once thank for encouraging me to join the Michael's group in the first place. Also he and another Knop lab alumni, Florian Huber, are part of my Heidelberg nerd connection which I hope continue long beyond my PhD. This also goes for Hanna and Nils Kurzawa who are two very important friends helping me to find the balance in life through various discussions, long hikes, game nights and boulder sessions. Especially for the latter also Sophie Winter and Dimitris Papagiannidis cannot go unmentioned and I am incredible happy to have met them during my PhD. Florian Schmidt, a fellow former iGEM warrior remains to be a close friend and it was through one of our enthusiastic scientific discussions that I learned about the tagmentation method which became so fundamental to the work presented in this thesis.

Without question I feel deeply encouraged and supported by my family to pursue my scientific career. In particular, there was my grandfather who ignited my appetite for the sciences. There is my mother who nurtured this appetite and my dad from whom I learned the love for nature and craftsmanship and of course also my sister for having my back.

And there is my own little family. Our son Thomas arrived while I was in the middle of writing this thesis. I am infinitely thankful for you being such a wonderful and balanced child and I am so much looking forward to (again and further) discover the world with you. Finally, there is my partner in life and science, Sophie Herbst. We met at the start of our studies and during that time you became my biggest supporter and most valuable critic. With you I can share my thoughts, ideas and doubts be it scientifically or not. You keep me grounded and it is with you that I want to continue to grow. It is impossible to express how grateful I am to share my life with you.

Contributions

The original Anchor-Seq protocol utilizing vectorette adapters was developed by our collaborators Ehud Sass and Emmanuel Levy (Weizman institute, Israel). Emmanuel Levy also implemented the original computational workflow to validate the C-SWAT library. We further discussed the potential adaptation of UMIs with a modified adapter design which I then implemented. I designed and implemented the Tn5-Anchor-Seq protocol. Some initial experiments not further discussed in this thesis but related to the establishment of tagmentation were conducted by the Master student Jan Dohnálek, who conducted a summer internship and whom I supervised. Daniel Kirrmaier performed the dilution series experiment which was planned by me. I performed all computational analysis related to Tn5-Anchor-Seq and supported the original analysis related to the C-SWAT library validation.

The CASTLING strategy was designed and implemented by Michael Knop, Benjamin C. Buchmuller and Matthias Meurer with some suggestions from my side. Daniel Kirrmaier provided technical assistance. I designed and performed the experiment with the small nuclear CASTLING pool and performed the experiment for genome-wide library construction together with Matthias Meurer with technical help from Krisztina Gubicza. Data analysis was jointly performed by Benjamin C. Buchmuller and me.

During the PCR tagging project all mammalian cell culture work was performed by Julia Füller as well as the tagged cells counting assays and genomic DNA extractions. Matthias Meurer designed and performed cloning of tagging plasmids. Cell counting visualizations were realized by Michael Knop. Experiments were planned by me, Julia Füller, Matthias Meurer, Marius Lemberg and Michael Knop. Daniel Kirrmaier performed Tn5-Anchor-Seq experiments related to the project and I analyzed and visualized the resulting sequencing data. I planned and conducted the amplicon sequencing experiment and analysis including visualizations. I planned the test for the end-to-end ligation which was conducted by Daniel Kirrmaier. Krisztina Gubicza implemented a web-application for designing PCR tagging primers with input from me, Michael Knop and Matthias Meurer.

The LAMP-sequencing strategy was designed by me. Sequencing of the RT-LAMP reactions by LAMP-sequencing was jointly planned by me and Daniel Kirrmaier who also performed the experiment. I analyzed the resulting NGS data and performed additional analysis and visualization for the remaining part of the project. The RT-LAMP assay was implemented by Matthias Meurer, Viet Loan Dao Thi, Kathleen Boerner, Daniel Kirrmaier and Michael Knop.

Throughout all of these involvements I contributed substantially to strategic discussions.

1 Introduction

1.1 Characterization of the genotype-to-phenotype relationship

Establishing the genotype to phenotype relationship is one of the central tasks in biology as such relationship can still not be easily predicted [1]. A central aid to find such relations is the manipulation of a genotype followed by the characterization of the resulting phenotypic change. Examples of genetic perturbations which might be of interest to assess biological questions are knocking genes in or out [2]. Furthermore, the regulatory context of a gene can be altered to explore expression changes. Finally, alterations of the coding sequence of genes can be performed to characterize and dissect the function of the gene product. This includes gene tagging with reporter genes. Frequently used examples for such reporter genes are fluorescent protein tags which can inform on abundance, subcellular function and stability [3, 4, 5]. In the past genetic perturbations were performed in a random, untargeted fashion and such approaches are still useful today [6]. Targeted approaches are an alternative which will be discussed in more detail below (refer to section 1.2). Technologies from genome-wide strain collections to systemic methodologies were historically quite often pursued using budding yeast *Saccharomyces cerevisiae* as model system mainly because it is especially amenable to genetic modifications and relatively easy to handle in the laboratory [7]. Within recent years, several advances have opened up new possibilities for functional characterizations also of other biological model systems such as mammalian cells. The most notable advance is based on CRISPR-Cas systems which allow to easily program genome editing endeavors [8, 9, 10, 11].

High-throughput studies often relied on arrayed cell line or strain collections (libraries) where genotypes have *a priori* known positions for example on an agar plate. A phenotype can be directly connected to its causing genotype by the array position given that arrayed assays can be used for phenotypic assessment of the library. However, one disadvantage of arrayed libraries is that their construction and use is laborious and that arrayed assays have limited throughput. This might arguably be a limitation for future functional genomics endeavors. A conceptually different approach with the potential to overcome these limitations can be realized with pooled libraries. They have the advantage of being much less resourceful in handling than an arrayed library and phenotype determination of all genotypes can be performed in parallel. Among others, pooled library assays potentially provide improved phenotypic sensitivity because all cell lines are exposed to exactly the same conditions. In contrast, arrayed libraries are for example affected by spatial effects. A disadvantage of pooled libraries is the challenge of genotype traceability. Nowadays,

methods for the characterization of pooled libraries are either based on next-generation sequencing (NGS) of phenotypically separated subsets of such pools or spatially resolved genotyping [1].

1.2 Gene targeting and tagging

Often it is desirable to alter the genomic sequence of a biological system under investigation to study the impact of the resulting genotype on phenotype outcome. Gene targeting is the process in which such alterations are performed in a specific manner. Such alterations can include gene disruptions, fusions with terminal or internal tag coding sequences for altered control or function, or site-directed mutagenesis at the endogenous gene locus [12, 13].

1.2.1 A brief history of gene targeting

One of the first reports of gene targeting in eukaryotes was presented in yeast by the lab of Gerald Fink in 1978. This study showed that a yeast strain auxotrophic for leucin as a consequence of a mutation in the leu2 gene could be transformed to prototrophy using a non-replicative vector which contained the intact LEU2 gene. Subsequent genetic analysis confirmed that the vector often had facilitated gene conversion of the mutant leu2 gene to its functional version LEU2. Homologous recombination (HR) had already then been proposed to play an important role in mediating such specific alterations [14]. Building on this knowledge it was later shown that recombinant DNA can be used to site-specifically delete the respective gene at its endogenous locus if this DNA carried the sequence of the gene of interest which has been disrupted by a marker gene [15]. Around the same time the lab of Mario Capecchi first reported HR of exogenous non-replicating plasmid DNA in mammalian somatic cells [16] by microinjecting the DNA into the nucleii of mouse embryonic stem cells. Later studies showed that it can be used for the purpose of gene targeting [17] and gene disruption [18]. Linearization of the recombinant DNA greatly improved transformation efficacy indicating that the co-occurrence of a double strand break (DSB) stimulated recombination in yeast [19] as well as in mammalian cells [20]. Gene targeting in mammalian cells turned out to be more difficult than in yeast cells as many genomic integrations of the recombinant DNA occurred at random instead of the targeted sites indicating non-homologous mechanisms for DSB repair [20]. Mario Capecchi and Oliver Smithies together with Martin Evans received the Nobel Prize in Physiology or Medicine in 2007 for their achievements related to gene targeting in mice.

1.2.2 PCR targeting

A major simplification of gene targeting in yeast was to assemble the respective recombinant construct by PCR. This was done by using primers bearing short terminal sequences of at least 30 nucleotides homologous to the target gene [21, 22]. This ”PCR targeting”

provided the immediate advantage of more rapid conduction of gene targeting experiments in yeast as only a single PCR amplification instead of several tedious molecular cloning steps was required to generate the respective material for transformation. Conceptually an additional advantage was the modularization of the gene targeting approach. On the one hand, the information directing the gene targeting construct to a specific genomic location is solely encoded on the 5' homologies of the primers. On the other hand, the template used for PCR encodes functional elements to fulfill the purpose of the gene targeting experiment. Usually, a selection marker is included in this functional module for the rapid identification of successfully transformed clones. In addition, other functional elements can be included such as regulatory sequences for example for promoter substitution experiments or terminal protein tag sequences for gene tagging experiments. The use of generic primer binding sites allows to freely combine primers and functional module templates to construct new targeting modules. The full modularity of this approach further facilitates rapid experimentation and still is one of the fundamental molecular techniques in yeast research [23, 24, 25].

1.2.3 How programmable endonucleases revolutionize gene targeting

The ease of gene targeting in *Saccharomyces cerevisiae* has been unmatched in other organisms for some time which was probably one of the reasons why this organism was seen as such an exquisite model system for functional studies [7]. In addition, although relatively efficient in yeast, approximately only one in 10^6 to 10^7 cells in a gene targeting experiment is transfomed to the desired genotype which prohibits more complex endeavors such as pooled gene targeting [24]. The most promising avenue to increase success rates of gene targeting efforts is to induce DSBs at the target site to stimulate homologous recombination. A range of molecular tools were used for site-specific DSB induction in different gene targeting applications across a range of organisms. These include site-specific meganucleases such as I-SceI and I-CreI, zinc-finger nucleases (ZFN), or TALENs [12, 13]. These endonucleases helped to perform a variety of successful gene targeting studies. It became clear that simultaneous DSB induction using target site-specific nucleases increased targeting efficiencies to a level which allowed for the use of HR templates with homology length as short as 50 nucleotides [26]. Nevertheless, the application of these endonucleases was generally hampered by the lack of accessibility, versatility and specificity. Another class of site-specific endonucleases are CRISPR-Cas endonucleases which were originally described to aid bacterial immunity but have been rapidly repurposed for gene targeting and genome engineering applications [8, 11]. These endonucleases have the advantage that they are guided to a particular genomic locus by a site-specific RNA. Therefore, designing a new site-directed endonuclease is as simple as exchanging the sequence of the guiding RNA. This versatility explains why CRISPR effectors are currently the most attractive solution for site-directed DSB induction.

1.3 DNA double strand break repair in eukaryotic cells

The repair of DSBs is a central prerequisite for efficient gene targeting in eukaryotic cells. I will therefore briefly summarize the current knowledge on DSB repair in eukaryotic cells in the following section.

As critical DNA lesions double strand breaks can impose an immediate threat on genome integrity eventually resulting in severe cellular dysfunction and disease. Moreover, DSBs can also be purposefully involved in several important cellular mechanisms such as meiotic chromosome segregation, resolution of stalled replication forks and antigen receptor differentiation in immune cells of vertebrates. Cellular systems across all kingdoms of life have therefore evolved various mechanisms to resolve DSBs. The following discussion will concentrate on DNA double strand repair in eukaryotic cells but many of the central factors are deeply conserved as exemplified by the conservation of the central homologous recombination protein Rad51 (recA in *E. coli*) indicating the high evolutionary pressure for DNA maintenance [27].

If sequences of reasonable homology are available they can serve as template for correct recovery of the original sequence. This process is known as Homologous Recombination (HR). If such a template is not available structural integrity of the DNA is ensured by non-templated ligation of the free DNA ends using a pathway termed Non Homologous End Joining (NHEJ). This pathway is imperfect in that it can be often accompanied by minimal sequence alterations (reviewed in ref [28]). Usually, one of these two pathways result in the resolution of a DSB. In case of their failure eukaryotic cells can revert to more promiscuous but less accurate repair mechanisms which are alternative End joining (a-EJ) and single strand annealing (SSA). The subsequent discussion will focus on HR and NHEJ because of their higher relevance for gene targeting and normal cell physiology.

Apart from the availability of a homologous template several other factors play a role in DNA repair pathway choice. An important one is the cell cycle. NHEJ is active throughout the cell cycle while HR is especially upregulated during S and to a smaller extend in G2 phase [29]. This permits that the sister chromatid more likely serves as homologous template for DSB repair by HR instead of the homologous chromosome [30, 29].

Decondensation and remodeling of dense chromatin surrounding a DSB is enhanced by the activation of Poly(ADP-ribose) polymerase (PARP) and signaling of the central DNA damage response kinase ataxia telangiectasia mutated protein (ATM) facilitating accessibility of the DSB. If DNA ends of the DSB are freely accessible and present only very short or no single stranded DNA protrusions, they are rapidly bound and protected from resection by the factor Ku, a heterodimeric complex of Ku70 and Ku80. Given the right circumstances such as cell cycle state and chromatin context binding of the Ku will be

challenged by factors which will direct DSB repair towards HR. Otherwise DSB repair will be committed to NHEJ by factors which bring the two Ku protected termini of the DSB into close proximity (i.e. formation of a synaptic complex) to facilitate their ligation. In case the DSB already provides longer ssDNA protrusions HR will engage directly at the DSB site without Ku binding [31]. The following sections will discuss both options in more detail.

1.3.1 Non-homologous End Joining (NHEJ)

NHEJ is the DSB repair pathway of choice for a majority of cases in mammalian cells [32] but is also physiological essential for V(D)J recombination during lymphocyte maturation [33]. The free DNA termini of a DSB are first bound by the highly abundant Ku complex (Ku70-Ku80; yeast: Yku70 and Yku80) which prevents extensive DNA end resection and serves as scaffold for further NHEJ factors [34]. The DNA damage response and chromatin-associated protein p53-binding protein 1 (53BP1; yeast: Rad9) promotes NHEJ through its effectors RAP1-interacting factor 1 (RIF1) [35] and the recently identified shieldin complex [36, 37]. The central ligase acting in NHEJ is a complex of the DNA ligase IV (LIG4; yeast: Dnl4) with X-ray repair cross-complementing 4 (XRCC4; yeast: Lif1) [38, 39]. The DNA termini of the DSB need to be spatially aligned through synapsis formation in order be competent for ligation. The ligation of the DNA ends is stabilized by the utilization of terminal microhomologies of up to four nucleotides [40]. Recent *in vitro* work has further defined the central role of Ku, XRCC4-LIG4 and XRCC4-like factor (XLF; yeast: Nej1) in promoting synapsis and facilitating end-joining. A flexible synapsis of Ku-bound DNA can be formed by XRCC4-LIG4 which on its own is sufficient to mediate ligation of compatible ends. Ligation is further enhanced by directly aligning the ends in a closed synapsis facilitated by XLF [41]. In case the DNA ends of the DSB are incompatible further ligation attempts are facilitated by end processing with nucleases and polymerases. A major nuclease activity in mammalian cells is provided by Artemis which is activated through phosphorylation by autophosphorylated DNA-dependent protein kinase catalytic subunit (DNA-PKcs) bound to Ku-loaded DNA [42]. Once activated, Artemis can cut DNA ends at single-strand-to-double-strand boundaries including 5' and 3' overhangs and various other obstacles to end-to-end ligation [43]. The two Pol X family polymerases Polμ and Polλ (yeast: Pol4) are important for template independent and dependent nucleotide addition respectively [44, 45].

Multiple rounds of processing and ligation attempts and an incremental use of a multitude of factors seem to allow for highly promiscuous and flexible DSB repair by NHEJ [28]. The complete process is more complex than described here and many open questions regarding the NHEJ pathway remain.

1.3.2 Homologous Recombination (HR)

The HR pathway in eukaryotic cells mediates genetic exchange and faithful chromosome segregation during meiosis. In somatic cells, HR ensures correct replication in addition to DSB repair. The extend of end resection at DSBs contributes to DSB repair pathway choice and longer resections of the DNA ends are required for commitment to HR in comparison to NHEJ. The DSB is first bound by a complex of MRE11-RAD50-NBS1 (yeast: Mre11-Rad50-Xrs2 (MRX)) [46, 47]. This also activates the central DSB response kinase ATM (yeast: Tel1) [48]. The nuclease activity of MRE11 leads to short-range (<300 nucelotides) nicking and resection so that 3' single-stranded ends are revealed. The nuclease activity of MRE11 requires interaction with CtBP-interacting protein (CtIP, yeast: Sae2) [49]. Further long-range resections are performed by two parallel pathways which involve either exonuclease 1 (yeast: Exo1) or the endonuclease DNA2 (yeast: Dna2) together with the Bloom syndrome helicase (BLM; yeast: Sgs1) [50]. This leads to displacement of the KU complex which eventually covers the DSB ends. The long 3' ssDNA ends are initially protected by binding to the three-member replication protein A (RPA) which also prevents secondary structure formation [51]. A dynamic structure of a nucleoprotein filament is formed by binding of the ssDNA to RAD51. In mammals loading of RAD51 is mediated by BRCA2 and Rad51 paralogs [52], while in yeast loading of Rad51 is predominantly mediated by Rad52 [53, 54]. The RAD51-ssDNA nucleoprotein filament mediates the homology search of HR and the sister chromatid is the preferred substrate over homologous chromosomes and ectopic homologous sequences [29]. Upon detection of adequate homology strand invasion occurs by the nucleoprotein filament and a DNA DNA heteroduplex is formed [55]. Displacement of the non-base-paired strand at the heteroduplex side forms a structure known as the displacement loop (D-loop) [56]. The invaded filament is used as template for nascant strand synthesis predominantly by DNA polymerase δ(Polδ) which works together with PCNA and RFC1-5 [57, 58]. After strand synthesis the D-loop must be resolved which is often mediated by BLM-TOPOIIIα-RMI1-RMI2 (yeast: Sgs1-Top3-Rmi1). During mitosis the major pathway is synthesis-dependent strand annealing (SDSA). It is the most conservative way of D-loop resolution and does not involve cross-overs which could potentially lead to loss-of-heterozygosity [59]. Another form of D-loop resolution involves double Holliday junctions (dHJs) which can be resolved by crossover and non-crossover events [60]. Finally, if HR fails to terminate the strand synthesis, D-loops are resolved by a strategy of break-induced replication which is highly mutagenic [61].

1.4 CRISPR Cas

CRISPR-Cas systems have been recognized as the adaptive immune system of prokaryotes. In general, they are characterized by the presence of so called Clustered regularly interspaced short palindromic repeats (CRISPR) arrays in the host genome and CRISPR-

associated (Cas) proteins which perform functions associated with CRISPR array maintenance and immunity. The immunity is mediated by three phases: adaptation, expression and interference. Adaptation to an unknown virus or plasmid is realized by the acquisition of short DNA sequences (spacer) from fragments of the genome of the invader (protospacer) into the CRISPR array within the genome of the host cell. Each spacer within the array is flanked by direct repeat sequences which contributed to the original discovery of this type of locus. The CRISPR array is expressed as one long transcript (pre-crRNA) which is further processed either by host factors or specialized Cas proteins into mature CRISPR RNAs (crRNAs) where each crRNA contains one spacer. The third phase of interference is characterized by the complexation of the crRNA with one or several Cas proteins to form a ribonucleoprotein effector. Watson-Crick base pairing of the crRNA and the target sequence triggers endonucleolytic cleavage of the target genome by the effector complex. Important for recognition of the target sequence is the presence of a characteristic protospacer-adjacent motif (PAM). This motif is absent in the CRISPR array which ensures that the CRISPR-Cas effector complex does not target the CRISPR array itself [62].

Historically, CRISPR arrays were first recognized bioinformatically as recurrently and prevalently occurring across bacteria and archea species [63]. The term "CRISPR" was coined and adjacent operons were recognized which code for the CRISPR-associated (Cas) proteins [64]. It was later observed that the spacers within the repeats are homologous to invasive DNA species such as the genomes of bacteriophages and plasmid sequences. This prompted the hypothesis that the CRISPR array might serve as memory for an interference system to protect these prokaryotes against invasion [65, 66, 67]. Shortly after by showing that the CRISPR-Cas system of *Streptococcus thermophilus* is indeed an adaptive immune system against bacteriophage infection this hypothesis was experimentally confirmed [68]. The *in vitro* reconstitution of essential parts of the CRISPR-Cas system of *Streptococcus pyogenes* showed that one CRISPR associated protein, Cas9, is an RNA-guided DNA-targeting endonuclease which can easily be reprogrammed for new targets [69, 70]. This work opened up the possibility for revolutionary new biotechnological developments starting with the application of Cas9 for genome editing via NHEJ- and HR-mediated repair in human [71, 72] and bacterial cells [73]. Another single RNA-guided DNA-targeting endonuclease interesting for biotechnological applications, Cas12a (formerly CRISPR-associated endonuclease in *Prevotella* and *Francisella* 1 [Cpf1]) was subsequently identified [74].

So far, natural CRISPR-Cas systems could be categorized into two classes, six types and 33 subtypes [75]. The two main classes of CRISPR-Cas systems are differentiated by the organisation of the effector Cas proteins, which consist of multiple proteins in the case of class 1 while class 2 members are single protein effectors. Cas9 and Cas12a as single

protein effectors therefore belong to the class 2 CRISPR-Cas systems [75]. Nevertheless, Cas9 and Cas12a differ in several aspects. The Cas9 ribonucleoprotein complex consists of one protein (Cas9) and two RNA components (crRNA and trans-activating RNA [tracRNA]) in nature. It has been shown that the two RNA components can be combined into one ~100 nucleotides long chimeric molecule [69]. In contrast, Cas12a requires no tans-activating RNA so that a 42 nucleotides long crRNA is sufficient for targeting [74]. In addition, Cas12a can on its own process its crRNA while Cas9 requires tracrRNA and host RNAse III as additional factors [76, 74]. Upon target site binding Cas9 performs a blunt end cut proximal to the recognition site. This is mediated by HNH and RuvC nuclease domains which are part of the Cas9 protein [69]. Cas12a contains a single RuvC nuclease domain and performs a staggered end cut with 5' overhangs distal from the recognition site [74]. The PAM of Cas9 is generally G-rich and placed at the 3'end of the target site. In contrast, Cas12a has a T-rich PAM and is placed at the 5' end of the target site [74]. As a consequence CRISPR-Cas systems containing the Cas9 effector are categorized into type II while Cas12a-containing CRISPR-Cas systems belong to type V of the class 2 systems [75].

1.5 Next-generation sequencing (NGS) and genome walking sequencing

1.5.1 History and application of NGS

In order to study the function of biopolymers such as proteins or nucleic acids it is often essential to determine their primary sequence. Especially sequencing of DNA has progressed in an astonishing pace within the last decades and adaptations of these technologies now penetrate most areas of biology [77]. In the late 1970s the chemical cleavage method by Maxam and Gilbert and the chain termination method by Sanger and Coulson allowed to efficiently sequence DNA for the first time [78, 79]. The principle of the chain termination method was adapted more widely and is still used today with some technological improvements. The experimentally observed sequence of bases in a nucleic acid molecule is what is called a read in sequencing. Initially, determining few or even a single read was a resourceful investment. Automation and massive parallelization increased the throughput while the cost decreased several orders of magnitude allowing to determine the genome sequence of an individual almost within one day for ~1,000 $. These technological advances are collectively termed next-generation sequencing (NGS) [77]. Sequencing technologies can be broadly categorized into short-read (35-700 bp) and long-read sequencing technologies (>700 bp). In general, sample DNA is first fragmented into an appropriate size range and common DNA sequences (adapters) are added which are needed for clonal amplification to construct the sequencing library. In addition, the constant sequences of the adapter serve as starting point to conduct the sequencing reaction. These molecules then serve for templated incorporation of nucleotides which is coupled to a distinguishable signal. The nowadays most common platforms are provided by the company Illumina.

For the Illumina platforms DNA of interest is attached to P5 and P7 adapter sequences. These two sequences are needed to immobilize and amplify the DNA molecules on the glass surface of a flow cell by bridge amplification resulting in clonal clusters. Sequencing is then performed by sequentially exchanging the reagents within the flow cell. A DNA polymerase is allowed to proceed on the template molecules withing the clusters but because the used nucleotides are terminally blocked, only one nucleotide is incorporated per cycle. Because it is fluorescently labeled in different colors this nucleotide can then be detected using a microscopic system coupled to the flow cell. The labeling and blocking of the incorporated nucleotides is reversible so that it can be removed after image acquisition and before the next cycle can be conducted. Using this approach one base of a read is determined per cycle [80]. The majority of sequence data analyzed in this thesis was generated using this technology.

The technology underlying the products from Illumina are therefore considered as a cyclic reversible termination type of sequencing by synthesis reactions. Another type would be single-nucleotide addition in which the four different nucleotides are sequentially added to the polymerase reaction. Nucleotide incorporation can then be converted into a detectable signal for example by an enzymatic cascade converting pyrophosphate into light (as in the first commercial NGS system by 454/Roche) or the pH change using semiconductor technology (as in the Ion Torrent systems). In contrast to sequencing by synthesis some commercial systems (e.g. Complete Genomics/MGI) rely on sequencing by ligation in which nucleotide incorporation is mediated by a DNA ligase instead of a DNA polymerase. [81]

A more recent development are technologies for real-time single-molecule sequencing which are sometimes also considered belonging to the third generation. The development of single-molecule optics with zero-mode waveguides allow to directly observe nucleotide incorporation of single polymerase-mediated synthesis reactions. This approach has been commercialized by Pacific Biosciences. An alternative strategy utilizes nanopores for sequencing. It is based on the realization that an ion current flow through a nanopore is modulated when a nucleic acid polymer also passes through the pore. By discretizeing the movement of the nucleic acid polymer with an engineered nucleic acid binding protein the current change becomes distinguishable based on the sequence of the nucleic acid polymer. The company Oxford Nanopore Technologies has commercialized this technology. Both aforementioned approaches allow for long read sequencing. [81, 77]

NGS technologies can be used for *de novo* assembly of a reference genome for a certain species and resequencing to identify genetic variants by aligning sequence data of a different sample to such a reference genome [77]. Especially for genome assembly of repetitive genomes long-range linkage between genomic loci is critical and NGS technologies have

profited from the idea to sequence both ends of the NGS library molecules, a principle known as paired-end sequencing [82]. In addition, NGS can be used as a molecule counting device which allows to quantify a wide range of cellular parameters. Probably NGS is most widely applied for measuring transcript abundance (i.e. transcriptomics) but other applications are possible [77]. In this thesis NGS will be used to characterize genetic variants of a gene editing experiment with respect to an expected outcome (i.e. resequencing) and counting of these variants within a mixture of cell lines.

1.5.2 NGS library preparation using Tn5 tagmentation

A methodological advance in NGS library preparation was the utilization of *in vitro* transposition reactions with a hyperactive variant of the Tn5 transposon also known as tagmentation [83]. Wild-type Tn5 is a composite transposon containing antibiotic resistance genes flanked by two highly homologous sequences of which one contains the gene for expressing the Tn5 transposase enzyme. The homologous sequences contain end sequences which are required by the transposase for recognition and transposition. This transposition follows a 'cut-and-paste' mechanism in which the transposon sequence is excised from a donor DNA sequence and randomly inserted into target DNA. Naturally, transposition occurs with low frequency but hyperactive mutants of the transposase and the transposon end sequences (denoted ME; mosaic ends) have been identified which allows for *in vitro* and *in vivo* biotechnological applications. (reviewed in ref [84])

DNA structures containing just one copy of the ME sequence can serve as potent transposition substrate and lead to random DNA fragmentation. This realization fueled the application for rapid NGS library preparations [83]. It has been commercialized in the Nextera product line from Epicentre Biotechnologies (now Illumina) but open-source protocols also have been published [85, 86].

1.5.3 Genome walking in the era of NGS

Often the purpose of a NGS experiment is not to generate sequence information across the whole genome but rather to characterize specific loci in detail. Therefore, enrichment for the informative subset of the sample is necessary before NGS is conducted, a process considered as targeted sequencing.

In principle, PCR can be conducted to amplify the loci of interest. Nevertheless, characterization of unknown sequences adjacent to a known genomic site is particular challenging because specific enrichment must be achieved but locus specific primers for PCR cannot be designed in a straight-forward manner. Solving this problem is for example a requirement for the identification of all insertion sites in certain mutants (e.g. in transposon mutagenesis or knock-in experiments) or for sequencing the proximity of genomic targets in an unbiased manner (e.g. in genome engineering or gene therapy). The approaches aiming

at addressing this challenge are denoted as genome, chromosome or primer walking [87, 88].

Already more than 30 years ago before the advent of NGS first genome walking protocols were developed to enable sequencing of the long DNA inserts in yeast artificial chromosomes (reviewed in ref [89]). These approaches can be broadly categorized based on how DNA is treated before conducting the PCR for enrichment [87, 88]. In fragmentation-based genome walking protocols the DNA is first fragmented either using restriction enzymes or randomly using methods such as sonication. The fragmented DNA is then ligated to generic adapters (or cassettes) which serve as second primer binding site during PCR [90]. The design of the adapter ensures that the amplification can only proceed from molecules which contain the site of interest (e.g. vectorette PCR; [91]). A notable variation are protocols based on inverse PCR for which the fragmented DNA is circularized by intra-molecular ligation. The two sequences flanking the site of interest are linked for this and the resulting molecules are amplified using a PCR with primers which specifically bind outwards at both ends of the side of interest [92, 93]. A second set of genome-walking protocols is primer-based which means that the PCR is conducted with specific primers for the site of interest and random or degenerated primers so that PCR can proceed semi-selectively [94]. In a third variation of genome walking protocols primers specific for the site of interest are first extended to generate ssDNA encompassing the site of interest and its adjacent sequence. The ssDNA molecules are then processed to be tailed with polynucleotide sequences or ligated to generic adapters so that they can serve as templates in a PCR [95, 96]. Specificity of primer- and primer extension-based protocols is further increased by performing nested PCRs with two different primers specific for the site of interest. The integration of genome walking protocols with next-generation sequencing allows to characterize many unknown sequence in a targeted but unbiased manner which allows to address the aforementioned challenges [97, 98] (reviewed in [99]).

1.6 SARS-CoV-2 detection by nucleic acid amplification reactions

At the end of 2019 a new severe acute respiratory syndrome (SARS) coronavirus named SARS-CoV-2 was first described in Wuhan, China [100]. Infection with this virus causes severe respiratory syndrome in patients, known as coronavirus disease 2019 (COVID-19) [101]. It quickly spread globally causing an ongoing pandemic as of writing of this thesis. Detection of SARS-CoV-2 has since been instrumental in determining infected people to facilitate early identifications and tracing of outbreak situations. Therefore, SARS-CoV-2 detection is critical for informing stakeholders to decide on countermeasures [102]. The earliest, most widely applied and one of the most sensitive diagnostic test for SARS-CoV-2 detection is reverse transcription quantitative PCR (RT-qPCR) of viral RNA isolated from patient samples [103]. The RT-qPCR reaction involves virus

specific oligonucleotide primers which aid reverse transcription of viral RNA by a reverse transcriptase included in the reaction. The resulting DNA is subsequently amplified by the means of a PCR mediated by the primers and a thermostable *Thermus aquaticus* (Taq) DNA polymerase [104, 105]. Reactions with successful amplifications are either identified by specific fluorescently labeled oligonucleotide probes or unspecific fluorescent DNA dyes. The execution of a RT-qPCR assay therefore requires specialized equipment [106, 107]. Especially during the beginning of the pandemic the required resources to perform these kinds of assays, i.e. the reagents to isolate RNA from patient samples, the RT-qPCR reagents and thermocyclers with fluorometric readout needed for qPCRs were quickly becoming a bottleneck. Consequently, the demand was high for alternative diagnostic assays for SARS-CoV-2 detection. Several nucleic acid amplification reactions were proposed as alternatives to the PCR assay to aid SARS-CoV-2 genome detection, one of which is reverse transcription loop-mediated isothermal amplification (RT-LAMP) [108, 109, 110]. The execution of this particular assay seemed exceptionally appealing because it had proven exquisite specificity and sensitivity in other settings [111, 112]. Similar as in the RT-qPCR assay, viral RNA is first reverse transcribed with the aid of a reverse transcriptase and oligonucleotide primers included in the reaction [108, 113]. The reaction further involves a *Bacillus stearothermophilus* (Bst) DNA polymerase which in comparison to the Taq polymerase used in PCR has high strand-displacing activity allowing for isothermal reaction conditions. Exponential amplification is facilitated by the design of the oligonucleotide primers which mediate a rather complicated reaction process (Figure 4.1) [108, 109, 110]. Although a fluorometric readout can inform about the success of an amplification in a RT-LAMP reaction, alternative readouts have been established including ones as simple as change in color (colorimetric RT-LAMP). For this the polymerase reaction is only weakly buffered so that its pH will increase as DNA is amplified. Phenol red included in the reaction is used as pH indicator resulting in a color change from red/pink to yellow upon successful amplification [114]. This possibility together with the isothermal reaction conditions greatly simplifies the technical requirements for performing the assay. In addition, LAMP reactions are relatively robust with respect to reaction contamination raising the question whether RNA isolation from patient samples could be omitted and instead be used directly with the RT-LAMP assay [115].

1.7 Objectives of this thesis

We need more high-throughput approaches for the manipulation of genotypes to be able to test the hypotheses generated through large biological data sets. One considerable challenge is the tagging of genes (i.e. knock-ins) in cells. In addition, the genomic characterization of such cells is difficult to achieve in an unbiased manner. The following aims address these tasks.

1.7.1 Aim 1: Development a targeted strategy for genome walking NGS

Our previously published Anchor-Seq strategy is a targeted NGS protocol for the characterization of tagged cells. Unfortunately, its broader application is hampered by several shortcomings. The first aim is to utilize tagmentation for rapid NGS library preparation. Together with newly designed adapters the resulting Tn5-Anchor-Seq protocol will allow for more scalable applications. Anchor-Seq and Tn5-Anchor-Seq will be important for the subsequent aims.

1.7.2 Aim 2: A protocol for pooled gene tagging in yeast cells

Methods for single gene tagging in yeast based on HR are well established. Nevertheless, targeting several hundreds or thousands genes in parallel in a targeted fashion is not feasible with these classical approaches. The second aim is to develop a method CASTLING which utilizes the programmable endonuclease CRISPR-Cas12a for increasing HR gene tagging efficiency. The method, named CASTLING, will be implemented by first targeting a small set of 215 genes. Then the potential to tag genes genome-wide is explored. The findings from this study will be important to inform future directions for method development.

1.7.3 Aim 3: A protocol for single gene tagging in mammalian cells

In contrast to yeast, single gene tagging in mammalian cells is laborious and inefficient. The concepts of CASTLING (Aim 2) of utilizing CRSIPR-Cas12a and HR could also be applied to mammalian cells with some modifications. The third aim is therefore to generate respective tagging constructs by PCR. Extensive characterization of this mammalian PCR tagging strategy will be performed to identify any unwanted off-target effects.

1.7.4 Aim 4: Characterization of an diagnostic assay for SARS-CoV-2

Finally, during the beginning of the COVID-19 pandemic the need to increase testing capacity was evident. The fourth aim is to test a colorimetric RT-LAMP assay for SARS-CoV-2 detection and determine its diagnostic potential. One property of the new Tn5-Anchor-Seq protocol is to allow for the characterization of several hundred samples in parallel. It is therefore applied for this study as LAMP-sequencing to validate these RT-LAMP reactions.

Altogether, this work aims at developing and improving strategies for genome walking sequencing and gene tagging in eukaryotic cells to make them highly scalable and generally applicable.

2 Methods

2.1 Tagmentation mediated Anchor-Seq

Most methods were as published in ref.[116] and [117].

2.1.1 Anchor-Seq enrichment

"Genomic DNA (gDNA) was isolated from a saturated overnight culture (approximately 2×10^8 cells) using YeaStar Genomic DNA Kit (Zymo Research). Genomic DNA (125 µL at 15 ng/µL in ultrapure water) was fragmented by sonication to 800–1000 bp in a microTUBE Snap-Cap AFA Fiber on a Covaris M220 focused ultrasonicator (Covaris Ltd.). In our hands, 51 s shearing time per tube, a peak incident power of 50 W, a duty factor of 7%, and 200 cycles per burst robustly yielded the required size range. Adapters were prepared by combining 50 µM of the respective Watson and Crick oligonucleotides [...]. Each mixture was heated up to 95 °C for 5 min, followed by cooling to 23 °C in a large water bath over the course of at least 30 min. Annealed adapters were stored at –20 °C until use. We prepared an equimolar mixture of annealed adapters that contained either none, one, or two additional bases inserted after the UMI (halfY-Rd2-Watson and halfY-Rd2-NN-Crick) to increase heterogeneity of the sequencing library. The fragmented genomic DNA (55.5 µL) were end repaired and dA tailed (NEBNext Ultra End Repair/dA-Tailing Module, New England Biolabs) and ligated to 1.5 µL of the 25 µM annealed adapter mix (NEBNext Ultra Ligation Module, New England Biolabs). Products larger than 400 bp were purified by gel excision (using NuSieve, described above) and eluted in 50 µL 5 mM Tris-HCl (pH = 8.5). SIC integration sites were enriched by PCR (NEBNext Ultra Q5 Master Mix, New England Biolabs) using 12 µL of the eluate with suitable pairs of adapter- and SIC-specific primers. Initial denaturation was 98 °C (30 s), followed by 15 cycles of 98 °C (10 s), and 68 °C (75 s). Final extension was carried out at 65 °C (5 min). Reactions were purified using Agencourt AMPure XP beads (0.9 vol, Beckman Coulter). The fragments were further enriched in a second PCR using the custom-designed primers Ill-ONP-P7-bi7NN and Ill-ONP-P5-bi5NN [and a size selection was performed by gel extraction (250–600 bp)] [...]" [116]

2.1.2 Tn5-Anchor-Seq enrichment

"In detail, 100 ng/µl Tn5(E54K,L372P) transposase (purified according to [86]) was loaded with 1.25 µM annealed adapters (P5-UMI-gri501…506-ME.fw, Tn5hY-Rd2-Wat-SC3) in 50 mM Tris-HCl (pH 7.5) by incubating the reaction for 1 h at 23 °C. Tagmentation reactions were prepared by mixing loaded transposase with 1 µg gDNA and tagmentation

buffer (10 mM Tris-HCl, pH 7.5, 10 mM $MgCl_2$, and 25 % [vol/vol] dimethylformamide) and incubating for 10 min at 55 °C. For our batch of Tn5 transposase, we achieved reasonable tagmentation using an enzyme/gDNA mass ratio of 0.75. Tagmentation reactions were purified by bead purification (AMPure XP, Beckman Coulter) according to the manufacturer's instructions. Total eluates were used as input for a first PCR reaction with cassette- and Tn5 adapter–specific primers (5Btn-hmNeong.rv, P5.fw) with NEB-Next Q5 HotStart polymerase (New England BioLabs) with 15 cycles of 68 °C and 1 min elongation. Biotinylated amplicons were first purified by column purification (Macherey-Nagel) and then enriched using Dynabeads MyOne Streptavidin C1 beads (Invitrogen) according to the manufacturer's protocol. These beads were then used as input of a second PCR with cassette- and Tn5 adapter–specific primers with NEB-Next Q5 HotStart polymerase (New England BioLabs) with 25 cycles of 68 °C and 1 min elongation. PCR products were size selected for 400–550 bp using a 2 % agarose/TAE gel and column purification (Macherey-Nagel)." [117]

2.1.3 quantitative PCR (qPCR)

Quantitative PCR reactions were set up using the LightCycler 480 SYBR Green I Master Mix and run on a LightCycler 480 System (both Roche). Quantitative cycles (C_T or threshold cycle) were averaged over three technical replicates. For each condition two reference reactions were performed with primer pairs targeting the N-termini of the genes *ALG9* (*YNL219C*) and *SPS100* (*YHR139C*). The C_T value of these two measurements were averaged to derive the reference C_T value. Enrichment was calculated following the procedure of Livak and Schmittgen [118]:

$$\begin{aligned}\Delta\Delta C_T &= \Delta C_{T,enriched} - \Delta C_{p,input} \\ &= (C_{T,enriched,gene_x} - C_{T,enriched,gene_{reference}}) - (C_{T,input,gene_x} - C_{T,input,gene_{reference}})\end{aligned}$$

The enrichment is then calculated as:

$$enrichment = 2^{-\Delta\Delta C_T}$$

2.1.4 Tn5-Anchor-Seq sequencing and computational analysis

Libraries were first quantified and then sequenced on a NextSeq 550 sequencing system (Illumina) using 300 cycles paired-end reagent kit (reading 150 nucleotides per read). A spike-in of 20 % with phiX gDNA (Illumina) was performed to counteract homogeneity at the beginning of the second read.

Samples were demultiplexed with bcl2fastq (Illumina) using barcodes in the Illumina-specific adapters (introduced during tagmentation and second PCR). Reads were filtered and trimmed using a custom script (Julia v0.6.0 and BioSequences v0.8.0). The resulting

Table 2.1: Oligonucleotides used for qPCR this study. Sources for primer sequences of all other methods can be cound in the given references.

Name	Sequence (5'-to-3')	target / purpose
ENT4-q-rev	CGAAGAAGACAGTGAAGATACG	YLL038C / sample
YFH1-q-rev	CCTTCTCAATGGGGAGTG	YDL120W / sample
CDC1-q-rev	CCTTTTTGTTTCTGTGATGC	YDR182W / sample
GAS5-q-rev	CCTCTTCATCTTCATCTTCC	YOL030W / sample
mNeongreen_41nt.rv	CCGAAGATGTGCAATTCGTGAGTAGC	mNeonGreen cassette / sample
YER166W-q	ACAAAAGCTGGGCAACACG	YER166W /sample
YGR111W-q	CCAGACTTTGACATTTTCACATGG	YGR111W /sample
YJL154C-q	TTTATTCACGGGGATGAGTCC	YJL154C /sample
YCR012W-q	GCCAGTCGACTTCATCATTGC	YCR012W / sample
YDR155C-q	GAAAAGGGATTCGGCTACGC	YDR155C /sample
ALG9-qPCR-FP	TGTTTAATCCGGGCTGGTTC	ALG9 / standard
ALG9-qPCR-RP	AGTGGACAGATAGCGTAGAGAG	ALG9 / standard
sps_qpcr_f2	GCTTCTGCAACACCACTTTAC	SPS100 / standard
sps_qpcr_r2	CGTTTGACTGGCTACCAGATAC	SPS100 / standard

set of trimmed reads was aligned to the genomic sequence of *S. cerevisiae* strain S288C (R64-1-1) using bowtie2 (v2.4.2, [119]). Aligned reads were counted using a custom script (Python v3.9.2 with HTSeq v0.13.5, [120]) and the S288C gene annotation (R64-1-1). Counts were analysed in R (v4.0.4) and tidyverse.

2.2 CASTLING

The methods for this part of the thesis were originally written for ref.[116] which were jointly written by Benjamin Buchmuller and me. Custom material such as strains, plasmids and oligonucleotide sequences can be found in ref.[116].

2.2.1 SICs for individual genes

"Individual SICs were generated by PCR using a corresponding plasmid template [...] and using primers [...] that introduced the required 5 and 3 homology arms along with a locus-specific crRNA spacer. Cycling conditions for VELOCITY DNA polymerase-based amplification (Bioline) were 97 °C for 3 min, followed by 30 cycles of 97 °C (30 s), 63 °C (30 s), 72 °C (2 min 30 s), and a final 72 °C (5 min) extension hold. The reactions were column purified and adjusted to equal SIC concentration before yeast cell transformation." [116]

2.2.2 SICs for pooled libraries

"The oligonucleotide pools used in this study were synthesized by either CustomArray Inc. (pools A[...]), Twist Bioscience (pools B1 and B2), or Agilent Technologies ([genome-wide pool]), and reconstituted in TE in case they arrived lyophilized. Pool dilution and annealing temperature were optimized in each case to yield a uniform product of the expected length [...] To keep library member representation as uniform as possible, using more input material and higher annealing temperatures is desirable, as this will usually require fewer PCR cycles for amplification of the full-length synthesis product. All [...] pools were designed to allow for amplification in 15 cycles using Herculase II

DNA polymerase (Agilent Technologies) with forward primer pool-FP2 (or pool-FP3, as indicated) and reverse primers pool-RP2 (or pool-RP3). Cycling conditions were: 95 °C for 2 min, followed by six cycles of 95 °C (20 s), touch down from 67 °C (20 s, $T = -1$ °C per cycle), 75 °C (30 s), then nine cycles of 95 °C (20 s), 67 °C (20 s), 72 °C (30 s), and a final 72 °C (5 min) extension hold. Primers and truncated oligonucleotides (<75 bp) were removed using NucleoSpin Gel and PCR clean-up columns (Machery-Nagel GmbH & Co. KG). Feature cassettes were amplified by PCR using cognate cassette-FP and cassette-RP and any compatible plasmid template (50 ng[…]) under the following conditions: 97 °C for 3 min, followed by 30 cycles of 97 °C (30 s), 63 °C (30 s), 72 °C (2 min 30 s), and a final 72 °C (5 min) extension hold. The reaction was treated with DpnI (New England Biolabs) in situ and cleaned-up using NucleoSpin Gel and PCR clean-up columns. For PCR, VELOCITY high-fidelity DNA polymerase (Bioline) was used with the manufacturer's reaction mix supplemented with 500 µM betaine (Sigma-Aldrich). For analysis, 2 µL of the reaction were used for DNA gel electrophoresis (0.8% or 2.0% agarose in TAE (Tris-acetate-EDTA)[…])." [116]

"Circularization of the amplified oligonucleotide pool (0.8 pmol) with the amplified feature cassette (0.2 pmol) was performed using NEBuilder HiFi DNA Assembly Master Mix (New England Biolabs) in a total reaction volume of 20 µL at 50 °C for 30 min. For analysis by DNA gel electrophoreses, 10 µL of the reaction were used (0.8 % agarose in TAE[…])." [116]

"To amplify selectively the circular product from [the previous step], rolling circle amplification (RCA) using phi29 was used. First, the annealing mixture was set up (total volume: 5 µL in a PCR tube) using 1 µL of the crude or gel-purified circularization reaction, 2 µL exonuclease-resistant random heptamers (500 µM, Thermo Fisher Scientific), 1 µL of annealing buffer (stock: 400 mM Tris-HCl, 50 mM $MgCl_2$, pH = 8.0), and 1 µL of water. For annealing, the mixture was heated to 94 °C for 3 min and cooled down in thermocycler at 0.5 °C/s to 4 °C. Then, 15 µL amplification mixture were added (consisting of 2.0 µL 10× phi29 reaction buffer, 2.0 µL 100 mM dNTP mix, 0.2 µL 100× bovine serum albumin, 10 mg/mL, and 0.6 µL phi29 DNA polymerase; all from New England Biolabs). Amplification was allowed to proceed for 12–18 h at 30 °C, followed by heat inactivation of the enzymes at 80 °C for 10 min. For analysis by DNA gel electrophoresis (0.8 % agarose in TAE), 0.5 µL of this reaction was used." [116]

"To release the SICs, 20 U of the restriction enzyme BstXI (New England Biolabs) were added directly to the amplification reaction and the mixture was incubated for 3 h at 37 °C. Typically, such a reaction yielded 10–20 µg of SICs. For DNA gel electrophoresis, 1 µL was used[…]." [116]

2.2.3 SIC fidelity estimation by NGS

"The oligonucleotide pools were analyzed by NGS […] after PCR amplification and after recombineering, including UMIs for de-duplication […]. For the PCR amplicons, fragments with UMIs were generated using 200 ng starting material (purified by ethanol precipitation) in two cycles of PCR with Herculase II Fusion DNA Polymerase (Agilent Technologies) using an equimolar mixture of P023poolseqNN-primers (1 mM final concentration) in a 25 µL reaction. Cycling conditions were based on the manufacturer's recommendations (62 °C annealing, 30 s elongation). The reactions were purified with NucleoSpin Gel and PCR clean-up columns using diluted NTI buffer (1:5 in water) to facilitate primer depletion, and the fragments eluted in 20 µL 5 mM Tris-HCl (pH = 8.5) each. To remove residual primers, 7 µL of eluate were treated with 0.5 µL exonuclease I (Escherichia coli, New England Biolabs) in 1× Herculase II reaction buffer (1 h, 37 °C) and heat inactivated (20 min, 80 °C). The reaction was used without further purification as input for a second PCR (Herculase II Fusion DNA Polymerase, 30 cycles, 72 °C annealing, 30 s elongation) to introduce indexed Illumina-TruSeq-like adapters (primer Ill-ONP-P7-bi7NN and Ill-ONP-P5-bi5NN). The products were size selected on a 3 % NuSieve 3:1 Agarose gel (Lonza), purified using NucleoSpin Gel and PCR clean-up columns, and quantified on a Qubit Fluorometer (dsDNA HS Assay Kit, Thermo Fisher Scientific) and by quantitative PCR (qPCR) (NEBNext Library Quant, New England Biolabs, LightCycler 480, Roche). SIC pools were processed likewise using tRNA-seqNN and mNeon-seqNN as primers to introduce UMIs. All samples were pooled according to the designed complexity and sequenced on a NextSeq 550 system (Illumina) with 300 cycle paired-end chemistry." [116]

"We sequenced the oligonucleotide pool after PCR amplification, and the SIC pool obtained from the recombineering procedure[…]. In the latter instance, fragments compatible with Illumina NGS were generated digesting the products of RCA with Bts I (55 °C, 90 min, New England Biolabs) and SalI-HF (37 °C, 90 min, New England Biolabs). The fragments were column purified, diluted to 100 ng/µL, and blunted using 1 U/µg mung bean nuclease under the appropriate buffer conditions (New England Biolabs). The DNA fragments of 150–200 bp length were gel extracted on 3 % NuSieve 3:1 Agarose (Lonza). Both samples were sequenced by GATC Biotech AG (Konstanz, Germany) using Illumina MiSeq 150 paired-end NGS technology." [116]

2.2.4 SIC transformation

"For transformation of individual SICs or SIC pools, Cas12a-family proteins were transiently expressed by making frozen competent cells using either yeasts strains with GAL1-controlled Cas12a proteins grown in YP (1 % yeast extract + 2 % peptone) or SC (synthetic complete) medium containing 2 % (w/v) raffinose and 2 % (w/v) galactose as carbon source. For transformation [24], the heat shock was extended to 40 min and no dimethyl

sulfoxide was added. Recovery of cells that required selection for dominant antibiotic resistance markers (G-418, hygromycin B and clonNAT [25]) was allowed for 5–6 h at room temperature in YP-Raf/Gal (yeast extract peptone dextrose medium containing raffinose and galactose) or YPD (yeast extract peptone dextrose) to proceed prior to plating them on corresponding selection plates." [116]

"SIC pools were transformed at a total of 1 µg per 100 µL of frozen competent yeast cells (approximately 2×10^8 cells). Per library approximately 5 of such transformation reactions were combined corresponding to a yeast culture volume of 50 to 100 mL (OD600 = 1.0) to generate the competent cells. The number of transformants per library was calculated from serial dilutions. Replica plating on selective plates was used to exclude transiently transformed clones. After outgrowth, libraries were harvested in 15 % glycerol and stored at –80 °C. For subsequent experiments, including genotyping, approximately 10,000 cells per clone were inoculated in YPD, diluted to OD600 = 1.0 (approximately 50 mL of culture), and grown overnight. If necessary, a second dilution was performed to obtain cells in exponential growth phase." [116]

"Each transformation mixture was split into two parts containing 1/20 (libA) or 19/20 (libB) of the volume. The largest sample was plated onto four $25 \times 25\,\mathrm{cm}^2$ square plates with YPD + G-418. No replica plating was performed before the libraries were cryopreserved in 2.5, 10, and 50 mL of 15 % glycerol, respectively." [116]

"For libraries libC, and the small nuclear library[…], the transformation mixture was plated onto two $25 \times 25\,\mathrm{cm}^2$ plates with YPD + hygromcycin B." [116]

2.2.5 Fluorescence microscopy

"Cells were inoculated at an OD600 = 0.5 per condition in 5 mL low-fluorescent SC medium (SC-LoFlo [121]) from cryopreservation stocks and grown overnight, followed by dilution to OD600 = 0.1 in 20 mL SC-LoFlo the next morning and imaging during mid-exponential growth in the afternoon. Cells were attached to glass-bottom 96-well microscopy plates (MGB096-1-2-LG-L, Matrical) using concanavalin A coating [122]. High-resolution fluorescence micrographs were taken on a Nikon Ti-E epifluorescence microscope equipped with a 60x ApoTIRF oil-immersed objective (1.49 NA, Nikon), a 2048 × 2048 pixel (6.5 µm), an sCMOS camera (Flash4, Hamamatsu), and an autofocus system (Perfect Focus System, Nikon) with either bright field, 469/35 excitation and 525/50 emission filters, or 542/27 excitation and 600/52 emission filters (all from Semrock except 525/50, which was from Chroma). For each condition, a z-stack of 10 planes at 0.5 µm distance was acquired each with a bright field, a short (75 % excitation intensity, 10 ms) and a long fluorescence exposure (100 % excitation intensity, 100 ms) regimen. For display, the fluorescent image stacks were z-projected for maximum intensity, and cell boundaries taken from out-of-focus bright field images. For imaging cells of the small nuclear [libraries],

cells were inoculated from cryopreservation stocks and grown overnight in selective synthetic media (SC with monosodium glutamate and hygromycin B). The next morning, the cells were diluted in the same medium and grown to mid-exponential phase. Z-stacks were acquired using 17 planes and 0.3 µm spacing between planes." [116]

2.2.6 Fluorescence-activated cell sorting

"A homogenous population of small cells (mostly in the G1 phase of the cell cycle) were selected using forward and side scatter. Single cells were sorted according to fluorescence intensity using fluorescence-activated cell sorting performed on a FACSAria III (BD Diagnostics) equipped for the detection of green fluorescent proteins (excitation: 488 nm,; long pass: 502LP,; bandpass: 530/30). We first isolated cells (three million in total), which represented roughly the 30 % most fluorescent cells in library #1.1 […] as judged by comparison to cells from strain ESM356-1, which was used as a negative control. The population of fluorescent cells was then grown to exponential phase and sorted into eight fractions (bins) of 125,000 cells each (except for 62,500 cells sorted into bin 8) using bin sizes of roughly 5 % (bin 1), 20 %, 20 %, 20 %, 25 %, 5 %, 5 %, 1 % (bin 8) according to the $\log_{10}$-transformed intensity of fluorescence emission of small (G1) cells. Sorted pools were grown overnight and the cells were harvested for genomic DNA extraction and target enrichment NGS by Anchor-Seq." [116]

2.2.7 Library characterization by Anchor-Seq

Anchor-Seq enrichment was performed as described in section 2.1.1 and the samples were sequenced using Illumina technology as follows.

"After size selection by gel extraction (250–600 bp), NGS library concentrations were measured by Qubit Fluorometer (dsDNA HS Assay Kit, Thermo Fisher Scientific) and by qPCR (NEBNext Library Quant, New England Biolabs, LightCycler 480, Roche). Furthermore, their size distribution was verified either on a Fragment Analyzer (Advanced Analytical Technologies Inc) or by gel electrophoresis of the qPCR product. Quantified libraries were sequenced […] or on a NextSeq 550 sequencing system (both Illumina, 300 cycle paired end). If necessary, 10–15 % phiX gDNA was spiked in to increase sequence complexity." [116]

"For MinION nanopore sequencing, the first PCR was carried out as described [in section 2.1.1] for library #1.1 (using 20 cycles) to introduce barcodes for multiplexing FACS bins on the same sequencing run, column purified, and the NGS library was prepared for 1D sequencing by ligation (SQK-LSK108) according to the manufacturer's protocols (Oxford Nanopore Technologies). Sequencing was performed on a MinION device using R9.4 chemistry (Oxford Nanopore Technologies). Samples were multiplexed considering the number of different clones present in a pool, bin size, gDNA yield after extraction,

and yield of the first PCR." [116]

2.2.8 Insertion junction sequencing of non-fluorescent cells

"Cells from library 1a were grown in selective synthetic media (SC with monosodium glutamate and hygromycin B) for approximately eight generations, and non-fluorescent cells were sorted into glass-bottom 384-microscopy plates using a FACS Aria III as described [in section 2.2.6]. The absence of fluorescence was confirmed by fluorescence microscopy and 60 non-fluorescent clones were pooled and grown overnight to full density. Anchor-Seq amplicons were prepared as described [in section 2.2.7] using primers NegCells-NNN[…]. The amplicons were size selected (~600 bp) and cloned using the NEB PCR Cloning Kit (New England Biolabs). The resulting amplicons were Sanger sequenced at Eurofins Genomics (Cologne, Germany)." [116]

2.2.9 Illumina NGS data analysis and read counting

"Raw reads (150 bp paired-end) were trimmed and de-multiplexed using a custom script written in Julia v0.6.0 with BioSequences v0.8.0. Read pairs were retained upon detection of basic Anchor-Seq adapter features. Next, these reads were aligned to a reference with all targeted loci using bowtie2 [119] v2.3.3.1. Such references comprised the constant sequence starting from the feature cassette amplified by PCR and 600 bp of the respective proximal genomic sequence of *S. cerevisiae* strain S288C (R64-2-1). For off-target analysis, the constant Anchor-Seq adapter features were trimmed off the reads. The remaining variable sequence of the reads was then aligned with bowtie2 to the complete and unmodified genome sequence of *S. cerevisiae* strain S288C (R64-2-1). A read pair that aligned to the reference was counted if both reads of the pair were aligned, such that the forward read started at the constant region of the Anchor-Seq adapter-specific primers. In addition, we set the requirement that the inferred insert size was longer than the sequence provided for homologous recombination during the tagging reaction. Counting was implemented using a custom script (Python v3.6.3 with HTSeq 0.9.1 [120] and pysam 0.13). In case UMIs were included in the Anchor-Seq adapter design, they were normalized for sequencing errors using UMI-tools (version 0.5.3)[123] ." [116]

"For analysis of data obtained from amplicon sequencing (i.e., from PCR and SIC amplification reactions), the reads were either denoised from sequencing errors using dada2 (version 1.5.2)[124] to evaluate fidelity and abundance or directly aligned with bowtie2 to a reference build from the designed oligonucleotides. Denoised reads were assigned to loci based on the minimal hamming distance to designed oligonucleotides." [116]

2.2.10 Analysis of nanopore sequencing data and read counting

"Nanopore sequencing yields very long reads. Therefore, the reference was assembled as aforementioned but using 2000 bp of the locus-specific sequences plus the constant se-

quence of the cassette enriched by the Anhor-Seq reaction. MinION data were basecalled using the Albacore Sequencing Pipeline Software v2.0.2 (Oxford Nanopore Technologies). For data analysis, a custom script was used to extract and de-multiplex informative sequence segments from all reads based on approximate matching of amplicon features (e.g., the constant region of the vectorette or feature cassette; Julia v0.6.0 with BioSequences v0.8.0, see above). Matching with a Levenshtein distance of 1 was sufficient to discriminate between the barcodes used in this study. Then, the extracted sequence segments were aligned to the reference using minimap2 (v2.2-r409)[125], using the default parameters (command line option: "-ax map-ont") for mapping of long noisy genomic reads. Only reads that mapped to the beginning of the reference were counted using a custom shell script." [116]

Calculation of fluorescence intensity estimates are outlined in detail in ref.[116].

2.3 Mammalian PCR tagging

The methods for this part of the thesis were originally written for ref.[117] for which I wrote the section on targeted amplicon NGS and Anchor-Seq and Tn5-Anchor-Seq applications. Custom material such as oligonucleotide sequences can be found in ref.[117].

2.3.1 Tissue culture and transfection

"HEK293T […] cells were grown in DMEM high glucose (Life Technologies) supplemented with 10 % (vol/vol) FBS (Gibco) […] [and] were grown at 37 °C with 5 % CO_2 and regularly screened for mycoplasma contamination." [117]

"Transfection of HEK293T […] cells was performed using Lipofectamine 2000 (Invitrogen) according to protocol of the manufacturer and using a 24-well format. If not stated otherwise, 500 ng Cas12a plasmid and 500n g of the PCR cassette were used for transfection of 1 well in a 24-well plate." [117]

2.3.2 Cell counting and fluorescence microscopy

"[F]or live-cell imaging, cells were split 24 h after transfection into eight-well µ-slides (Ibidi). Analyses of transfected cells were performed 3 d after transfection or as described in the figure legends. Cells were stained with Hoechst 33342 (4 µg/ml in PBS, Thermo Fisher Scientific) for 5 min, and then the medium was changed to FluoroBrite (Thermo Fisher Scientific) supplemented with 10 % FBS (Gibco) and 20 mM Hepes-KOH, pH 7.4 (Thermo Fisher Scientific)." [117]

"For counting and imaging, different microscopes were used: a Nikon Ti-E widefield epifluorescence microscope or a DeltaVision, each with 60× oil immersion objectives (1.49 NA, Nikon; 1.40 NA, DeltaVision). Z stacks of 11 planes with 0.5 µm spacing were recorded

with 100 ms exposure time. Single-plane images and maximum intensity Z projections are shown. Subcellular localizations were identified and scored visually." [117]

"For cell counting, random fields of view were inspected in the HOECHST/DAPI channel, and all nuclei present in the entire field of view were counted. Cells containing transfected fluorescent protein–expressing cassettes were then counted subsequently in the same fields of view using the appropriate illumination wavelengths. In some experiments, counting was done using images recorded in the same manner." [117]

2.3.3 Preparation of genomic DNA (gDNA)

"gDNA for experiments shown in all figures except [for fidelity and on-/off-target analysis] was isolated from HEK293T cells […]. After washing with PBS, confluently grown cells from a well on a 6-well plate were lysed in 600 µl SNET buffer (20 mM Tris, pH 8.0, 400 mM NaCl, 5 mM EDTA, pH 8.0, and 1 % SDS), and 2 µl of RNase A (10 mg/ml RNase A, 10 mM Tris-HCl, pH 8.0, and 10 mM $MgCl_2$) was added for 30 min at room temperature. Afterwards, proteinase K (20 mg/ml proteinase K, 50 mM Tris-HCl, pH 8.0, 1.5 mM $CaCl_2$, and 50 % glycerol) was added for another 30 min at room temperature. Proteins were precipitated using 200 µl 3 M K-acetate solution, followed by precipitation of the DNA with isopropanol and washing with 70 % ethanol. DNA was dried and dissolved in TE (10 mM Tris and 1 mM EDTA) buffer. gDNA for [fidelity and on-/off-target analysis] experiments were purified according to the instructions of the manufacturer using the High Pure PCR Template Preparation Kit (Roche) followed by RNase A digest and a final purification with the High Pure PCR Product Purification Kit (Roche)." [117]

2.3.4 Targeted next-generation sequencing of tagged and wild-type alleles

"Tag- and wild type–specific amplicons from cells […] were generated from 200 ng gDNA using junction-specific primers by a two-step nested PCR with Velocity polymerase (Bioline). The first PCR reaction was performed for 15 cycles with 60 °C annealing and 30 s elongation and then purified with AMPure XP PCR beads (Beckman Coulter). The second PCR was performed for 15 cycles for wild type–specific and for 21 cycles for tag-specific amplicons, respectively, using 60 °C annealing and 30 s elongation. PCR products were size selected by gel electrophoresis on 2 % agarose/TAE and gel extracted by column purification (Macherey-Nagel). Amplicons were paired-end sequenced with 500 cycles on a MiSeq system (Illumina) using the Amplicon-EZ (150–500 bp) service by Genewiz to acquire at minimum 13,123 reads per sample. Paired reads were merged and aligned to the respective expected amplicon references using CRISPResso (v2.0.29; [[126]]) with parameters “cleavage_offset,” 1; and “window_around_sgrna,” 0. Mutations were subsequently quantified using a custom R script excluding primer binding sites in the analysis." [117]

2.3.5 NGS of gDNA by Anchor-Seq and Tn5-Anchor-Seq

Sequencing libraries for data presented in Figure 3.20 were prepared as described in section 2.2.7 with UMI-containing adapters [117].

"Quantified libraries were sequenced paired-end with 300 cycles on a NextSeq 550 sequencing system (Illumina) with a spike-in of 20 % phiX gDNA library (Illumina). Raw reads were trimmed from technical sequences (adapter and cassette sequences) using custom scripts (Julia v0.6.0 and BioSequences v0.8.0). The trimmed reads were aligned to the human reference genome (Genome Reference Consortium Human Build 38 for alignment pipelines) using bowtie2 (v2.3.3.1; [119]). Template cassette sequences were included in the reference genome as decoy. Aligned reads were grouped with UMI tools [123] based on UMIs included in the Anchor-Seq adapters. Enriched integration sites were further evaluated and counted using IGV (v2.4.10; [[127]])." [117]

Sequencing libraries for fidelity and on-/off-target analysis were prepared by Tn5-Anchor-Seq. For this tagmentation was performed as in section 2.1.2 using primers published in [117]. The resulting Tn5-Anchor-Seq sequencing libraries were sequenced as mentioned above.

"Raw reads were trimmed as already mentioned, but aligned to the human reference genome supplemented with PCR cassette sequences with bwa mem (v0.7.17-r1188; [128]). Mapped insertion sites were summarized by a custom R script and further evaluated and counted using IGV (v2.4.10; [127])." [117]

2.4 RT-LAMP assay and LAMP-sequencing for SARS-CoV-2 detection

The methods for this part of the thesis were originally written for ref.[115] for which I wrote the section on the LAMP-sequencing method. A protocol version of these methods was published as ref.[129] which was written by me. Custom material such as oligonucleotide sequences can be found in ref.[115] and ref.[129].

2.4.1 Clinical sample handling

"Specimens were collected as nasopharyngeal and oropharyngeal flocked swabs in Amies medium (eSwab, Copan Italia). The sample collection happened as part of the routine operation of Heidelberg University Hospital and at public testing stations set up by the City of Heidelberg [...]. Collected samples were transported in sterile containers, delivered to the diagnostic laboratory within a few hours, and then examined directly or stored at 4 °C until further processing. Samples were processed in a biosafety level 2 cabinet until inactivation by heat or mixing with a lysis buffer." [115]

2.4.2 RNA isolation and RT-qPCR

"RNA was isolated from nasopharyngeal and oropharyngeal swab specimens using QIAGEN kits (QIAGEN, Hilden, Germany); either automated on the QIASymphony (DSP Virus/Pathogen Mini Kits) or QIAcube (QIAamp Viral RNA Mini Kits) devices or manually (QIAamp Viral RNA Mini Kits). […] RT-qPCR for the quantification of the SARS-CoV-2 viral genome was performed using kits and reagents from TIB MOLBIO Syntheselabor, Berlin, Germany. The kits were used according to the manufacturer's instruction and contained the primer/probe sets developed based on the published Sarbeco primer set [103]. Per 20-µl reaction, the master mix contained 5.4 µl of RNAse free water, 4.0 µl of LightCycler Multiplex RNA Virus Master (Roche, Basel, Switzerland), 0.5 µl of LightMix Modular SARS and Wuhan CoV E gene (cat. no. 53-0776-96; TIB MOLBIOL Syntheselabor GmbH, Berlin, Germany) or LightMix Modular SARS and Wuhan CoV N gene (cat. no. 53-0775-96; TIB MOLBIOL), 0.5 µl of LightMix Modular EAV RNA Extraction Control (cat. no. 66-0909-96; TIB MOLBIOL), and 0.1 µl of reverse transcriptase enzyme (LightCycler Multiplex RNA Virus Master, Roche, Basel, Switzerland). The master mix (10 µl) was distributed per reaction into 96-well plates, and 10 µl of purified RNA was added per well. The performance of the RT-qPCR was validated using a positive control for the E gene. A total of 1000 molecules of E gene RNA per RT-qPCR reaction correspond to a CT $\approx$ 30." [115]

2.4.3 RT-LAMP primer design and positive control

"The RT-LAMP primer sets used in this study have been designed by Zhang et al. [130] against ORF1a and N gene and were synthesized by Sigma-Aldrich (synthesis scale, 0.025 µmol; purification, desalt; solution, water). An RNA-positive control for the N gene was amplified from a short fragment from 2019-nCoV_N_Positive control plasmid [Integrated DNA Technologies (IDT), 10006625] with oligonucleotides T7-GeneN-Fragment.for and GeneN-Fragment.rev including the T7 promoter and a subsequent IVT with the MEGAscript T7 Kit (Invitrogen) purified using the RNeasy MinElute Cleanup Kit (QIAGEN)." [115]

2.4.4 RT-LAMP assay

"Assays were assembled in total reaction volumes of either 12.5 µl (for LAMP assays using isolated RNA) […]. Master mixes were prepared at room temperature for each reaction immediately before use with either 6.25 or 10 µl, respectively, of the WarmStart Colorimetric RT-LAMP 2X Master Mix (M1800, New England Biolabs) and 1.25 or 2 µl, respectively, of the 10× primer mix, filled up to 11.5 or 19 µl with nuclease-free water (AM9937, Ambion). Values given are for one reaction: For a 96-well plate, 100 times larger volumes were used, and the LAMP mix was distributed to the wells of a 96-well plate (4ti-0960/C, Brooks Life Sciences or 0030128672, Eppendorf) before pipetting 1 µl

of sample into each well of the plate [...]. Plates were prepared immediately before use to limit exposure of the LAMP reagents to atmospheric CO_2 (to prevent acidification of the reaction) and kept on an ice-cold metal block. Plates were sealed using a transparent adhesive foil (GK480-OS, Kisker Biotech), and the reactions were incubated in a PCR cycler at 65 °C for 15 to 60 min with the lid heated to 75 °C. To perform measurements at the indicated time points, the reactions were taken out of the PCR cycler and placed into an ice cold metal block for 30 s. [...] Absorbance measurements were performed with a Spark Cyto or Infinite M200 (Tecan) at 434 and 560 nm with 25 flashes. These two peaks from phenol red are strongly changing during the acidification of the reaction (434 nm absorbance is increased, 560 nm absorbance is decreased). To obtain a good readout of the color change, absorbance at 560 nm was substracted from the one at 434 nm. This difference was denoted ΔOD." [115]

"Sensitivity and specificity values were obtained from count tables as follows: Specificity of the RT-LAMP assay was calculated as the fraction of RT-qPCR–negative samples that were also negative in the RT-LAMP assay. Sensitivity for a given CT interval was calculated as the fraction of all samples with an RT-qPCR CT value in that interval that was positive in the RT-LAMP assay. In both cases, 95 % confidence intervals were calculated by interpreting the fractions of counts as binomial rates and then using Wilson's method for binomial confidence intervals as implemented in the R package binom. The R code used to perform analyses and produce figures can be found on GitHub, together with all data tables: `https://github.com/anders-biostat/LAMP-Paper-Figures`." [115]

2.4.5 LAMP-sequencing

Tn5-Anchor-Seq was adapted for sequencing RT-LAMP reactions as detailed below.

"[T]ransposon adapters containing well-defining barcodes and unique molecular identifiers (UMIs) were annealed by mixing 25 µM oligos (P5-UMI-xi5001…5096-ME.fw, Tn5hY-Rd2-Wat-SC3) in 5 µM tris-HCl (pH 8), incubating at 99 °C for 5 min, and slowly cooling down to 20 °C within 15 min in a thermocycler. Transposons were assembled by mixing Tn5(E54K, L372P) transposase (100 ng/µl) [purified according to [86]] with 1.25 µM annealed adapters in 50 mM Tris-HCl (pH 7.5) and incubating the reaction for 1 hour at 23 °C. Tagmentation was carried out by mixing 1.2 µl of the RT-LAMP product (~200 ng DNA) with 1.5 µl of loaded transposase in freshly prepared tagmentation buffer [10 mM [tris(hydroxymethyl)methylamino]propanesulfonic acid) (TAPS)] (pH 8.5), 5 mM $MgCl_2$, and 10 % (v/v) dimethylformamide] using a Liquidator 96 Manual Pipetting System (Mettler Toledo). The reactions were incubated at 55 °C for 10 min. Reactions were stopped by adding SDS to a final concentration of 0.033 %. Tagmented DNA of each plate was pooled and size-selected using a two-step AMPureXP bead (Beckman Coulter) purification to target for fragments between 300 and 600 bp. First, 50 µl of pooled reaction was

mixed with 50 µl of water and bound to 55 µl of beads to remove large fragments. To further remove small fragments, the supernatant of this reaction was added to 25 µl of fresh beads and further purified using two washes with 80 % ethanol before the samples were finally eluted in 10 µl of 5 mM tris-HCl (pH 8). One PCR per plate with 1 µl of the eluate and RT-LAMP–specific and Tn5-adapter–specific primers (P7nxt-GeneN-A-LBrc and P7-xi7001…7016, P5.fw) was performed using NEBNext Q5 HotStart polymerase (New England Biolabs) with two cycles at 62 °C for annealing and 90 s elongation, followed by two cycles at 65 °C for annealing and 90 s elongation, and 13 cycles at 72 °C annealing and 90 s elongation. All PCR reactions were combined and 19 % of this pool was size-selected for 400 to 550 bp using a 2 % agarose/tris-acetate-EDTA gel and column purification (Macherey-Nagel). The final sequencing library was quantified by qPCR (New England Biolabs) and sequenced with a paired-end sequencing run on a NextSeq 550 machine (Illumina) with 20 % phiX spike-in and 136 cycles for the first read, 11 cycles to read the 11-nt-long plate index (i7) and 20 cycles to read the 11-nt-long well index (i5) and the 9-nt-long UMI."[115]

"For trimming of the reads (i.e., removal of P7 Illumina adapter sequences), cutadapt (version 2.8) [131] was used. For validation of the origin of the sequence of the LAMP product (Figure 3.28b), 10^7 reads were randomly selected and used for the analysis. Reads were mapped to the SARS-CoV-2 reference genome (NC_045512.2) [132], using bwa-mem with default settings (version 0.7.17-r1188) [128] . Virus genome coverage was determined with the samtools depth command (version 1.10) [133]. Using bwa-mem, 80.6 % of reads could be mapped to the virus genome (Figure 3.28c, 3.30a). To analyze the remaining sequences, a k-mer analysis using a custom script was performed. Using 9-mers, this matched 93.5 % of the nonmapped reads with a maximal Levenshtein distance of two to one of the LAMP primers or their reverse complement sequences (Figure 3.28b). This is explained by the fact that LAMP products can consist of complex sequence rearrangements."[115]

"For classification of samples by LAMP-sequencing, reads were assigned to wells and counted using custom scripts [`https://github.com/anders-biostat/LAMP-Paper-Figures/tree/master/LAMP-sequencing_raw_read_processing`]. A read was considered as a match to SARS-CoV-2 N gene if at least one of three short sequences (~13 nt, marked orange in fig. S4A) not covered by RT-LAMP primers was found in the read, otherwise it was counted as unmatched. Sequencing reads were grouped by UMI and by position of the matched sequence with the aim of removing PCR duplicates. A sample was considered if more than 200 total UMIs were observed and called positive if more than 10,000 virus-matching UMIs were observed."[115]

3 Results

3.1 Development of Anchor-Seq for genome walking sequencing

From time to time it is desirable to characterize only a certain region of the genome. In particular, following a gene targeting experiment, successful editing needs to be validated, ideally, by amplification and sequencing. In case of sequence knock-ins and specifically for gene taggings the correct insertion site must be confirmed.

For example, a feasible strategy was needed to validate clones during the construction of a large collection of *Saccharomcyes cerevisiae* strains in the lab. In each strain of this collection a different gene is C-terminally tagged with a heterologous DNA cassette [134]. The purpose of this library required determination of the exact sequences of the region encompassing the junction of the cassette and the genomic integration site. For single strains this would have been easily achievable by PCR amplification and sequencing with gene specific primers. Nevertheless, such an effort would have been an unrealistic effort for the 6014 strains originally targeted during the construction of the collection.

A protocol needed to be established which would allow for specific enrichment of the region adjacent to the termini of the tagging cassette followed by NGS to determine the sequences of these enriched regions. For this vectorette PCR which allows for such targeted amplifications [91] was adapted to an Illumina NGS workflow. In this protocol the sequence of the integrated tag serves as anchor for specific amplification of molecules containing the junction between the tag and adjacent sequences. Consequently, this protocol was termed Anchor-Seq (Figure 3.1) [134].

The original Anchor-Seq procedure can be summarized as follows: After genomic DNA (gDNA) extraction, the DNA was sheared using focused sonication which creates randomly fragmented molecules with a variety of end conformation (e.g. 5' and 3' ssDNA overhangs). These ends were homogenized using an enzymatic end repair and A tailing. Vectorette adapters were ligated to these molecules and excess adapters removed. Cassette-adjacent sequences were amplified by PCR using tagging cassette-specific and a vectorette-adapter-specific primers (Figure 3.1a). The design of the vectorette adapter with partially unmatching sequences allows for exponential amplification only after DNA polymerase extension had initiated from the binding of the cassette-specific primer. Hence, only molecules which contain this tagging-cassette-specific binding site can serve as template in the PCR reaction allowing for selective enrichment of cassette adjacent sequences (Figure 3.1b). The cassette-specific primer was selected to allow to sequence upstream

of the C-terminal cassette, i.e. the 3'-terminal coding sequence of the tagged gene was sequenced. Because this enrichment approach prior to NGS is invariant to the sequence adjacent to the anchor site all strains of the library could be pooled into one sample and sequenced simultaneously using barcodes introduced during the PCR steps [134, 135].

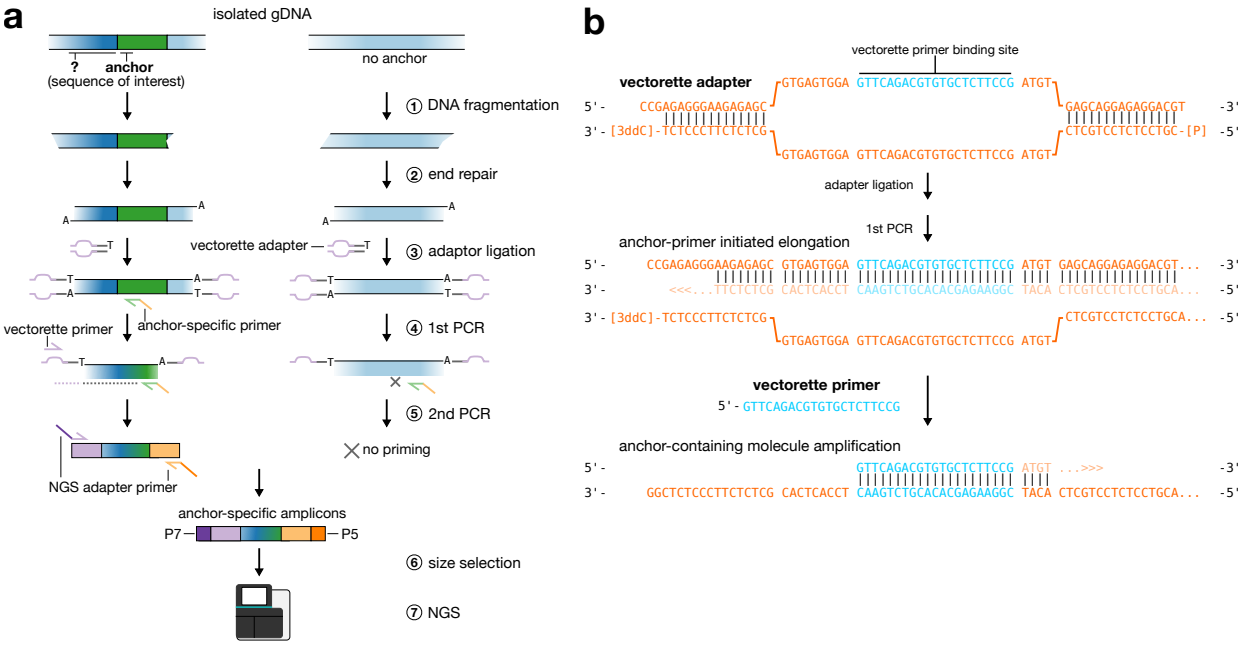


Figure 3.1: Anchor-Seq workflow using vectorette PCR. **(a)** Isolated genomic DNA is fragmented by sonication. Enzymatic end repair prepares DNA fragments for AT-ligation with pre-annealed vectorette adapters. In a first round of PCR only molecules containing the anchor sequence can serve as template for amplification because the vectorette primer has the same sequence as the vectorette adapter. This means it can only bind after DNA extension has initiated from the anchor-specific primer. A second PCR is used for further amplification and attachment of sequencing adapters. Before NGS a size selection is performed to remove short amplicons which do not contain enough sequence information. **(b)** Sequences of the vectorette adapter and primers conferring specificity to the first PCR step in (a). A 3-dideoxycytosine group (3ddC) is used to inhibit adapter concatenation.

3.1.1 Tagmentation mediated Anchor-Seq (Tn5-Anchor-Seq)

The original Anchor-Seq protocol involves random fragmentation of the gDNA by mechanical shearing with sonication. The ends of the resulting fragments then need to be repaired by enzymatic treatment so that they can be used for adapter ligation. A major limitation of this multi-step approach is that it is time- and cost-intensive and therefore difficult to upscale leading to an inherent low overall experimental throughput. Therefore, multiplexing of a larger number of samples (more than 30 samples) will be increasingly tedious. In addition, the specificity of the enrichment could be improved as I readily observed that 40-70 % of reads did not contain the expected sequence elements of the anchor or adapter and consequently needed to be discarded as off-target amplifications.

I therefore set out to develop an improved Anchor-Seq protocol which aimed at both increased scalability and specificity. In order to achieve this goal I integrated into the existing workflow two experimental approaches which are already widely applied in NGS

protocols. First, tagmentation can serve as an alternative for the introduction of technical adapter sequences which are required for performing the sequencing in NGS library preparations. It simplifies the experimental procedure by combining the steps of gDNA fragmentation and adapter ligation into one single enzymatic reaction [83]. Secondly, higher specificity in PCR-based NGS library enrichment was accomplished by using a nested PCR design.

I combined these ideas into a new Anchor-Seq protocol (Tn5-Anchor-Seq) in which tagmentation substitutes the mechanical shearing, end repair and adapter ligation steps of the original Anchor-Seq protocol unifying the adapter introduction into a single protocol step. The tagmentation adapter contains an index which means that samples can be pooled after tagmentation and before the subsequent PCR steps. The second sequencing adapter contains an additional index and is introduced during the second PCR. By utilizing different anchor-specific binding sites for each PCR reaction a nested PCR design is achieved to increase specificity. Furthermore, when the anchor-specific primer of the first PCR is biotinylated it might be possible to further increase specificity by performing a biotin-streptavidin-capture step before the second PCR (Figure 3.2a). The technical sequences of the tagmentation adapters were redesigned to represent the halfY structure of an annealed long and short oligonucleotide more commonly used for tagmentation experiments and to alleviate some of the hurdles we observed earlier for the vectorette design (Figure 3.2b and c). The sequences are the same as the library molecules of Illumina Nextera protocol which allows for direct adaptation of the commercial sequencing kit without the use of custom sequencing primers. These sequences also encompass the mosaic sequence (ME) required by the Tn5 transposase complex for binding to the DNA adapters (Figure 3.2b). The P5 and the P7 sites determine which molecule side is sequenced first during paired-end sequencing. In the original Anchor-Seq protocol the first read always started directly before the anchor sequence. This meant that the initial bases of the first read are the same for all sequencing reads and this reduced diversity often results in poor Illumina run quality and loss of overall read output [136]. In contrast, placing the P5 sequence at the randomly inserted tagmentation adapter naturally leads to increased variability of the initial bases of each read and therefore should help to improve overall run output and quality (Figure 3.2c). The tagmentation adapter not only contains a barcode (i5) which can be used for sample multiplexing but also a unique molecular identifier (UMI) which can be used to computationally correct for PCR bias [137]. For most of the experiments which are presented here an Illumina NextSeq 550 machine was used which reads the i5 index in the tagmentation adapter in 3' to 5' direction. This means that one can control reading of the UMI by adding the respective number of sequencing cycles to index read 2. The anchor-specific primer which contains the P7 site can also be barcoded so that higher levels of multiplexing can be achieved. Including a heterogeneity spacer

after the sequencing primer binding site (Figure 3.2c) improves run quality further for the second read.

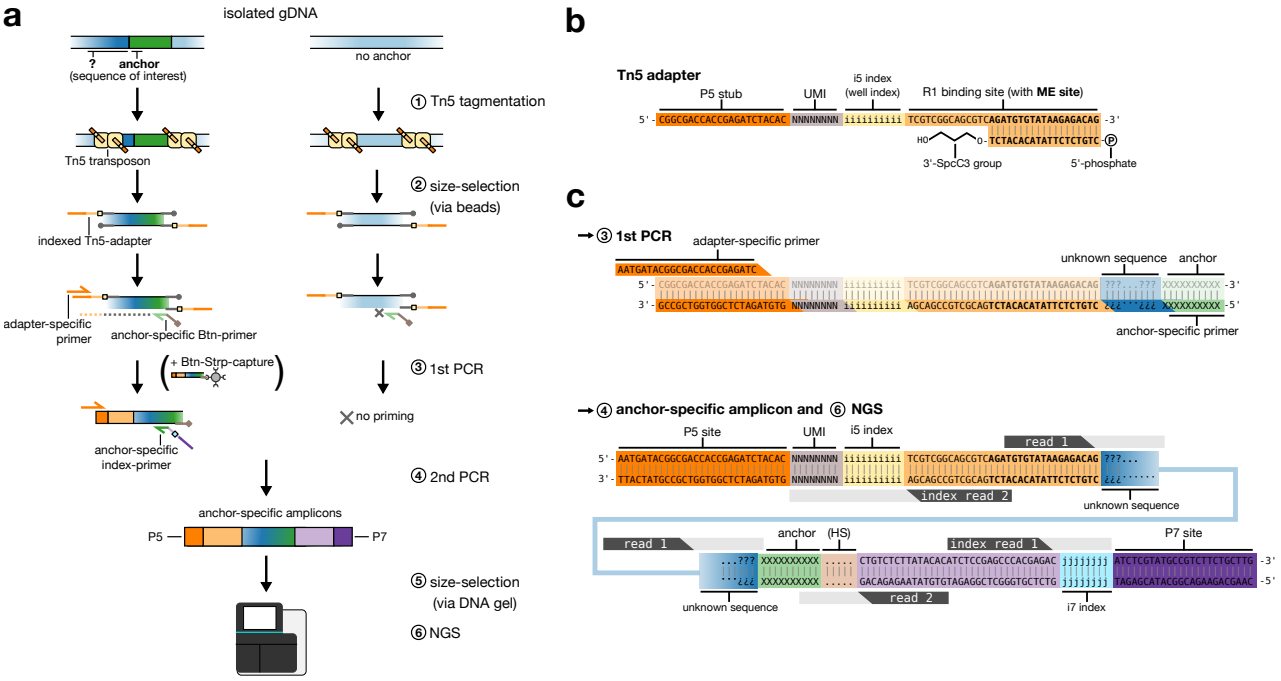


Figure 3.2: Anchor-Seq workflow using tagmentation (Tn5-Anchor-Seq). **(a)** Adapters for anchor-specific molecule amplification are introduced to isolated genomic DNA via tagmentation. This step also randomly fragments DNA to a certain size range. Bead-based size-selection removes adapters which have not been incorporated. As in the original Anchor-Seq protocol only molecules containing an anchor sequence can proceed for amplification although the design of the adapter is different (see also panel b and c). This first PCR can be performed with a biotinylated primer which allows for optional biotin-streptavidin-capture before the second PCR. The second NGS adapter is introduced during the second PCR using a binding site upstream of the first anchor sequence. This nested PCR design helps to improve specificity. Amplicons are size-selected via gel electrophoresis and sequenced using a modified sequencing routine (see panel c). **(b)** Sequences of the tagmentation adapter which follows a halfY design. A 3'-C3-Spacer group is used to inhibit primer extension by the DNA polymerase in the first PCR step which would preclude specific amplification of the anchor-containing molecules. The adapter also contains a unique molecular identifier (UMI) for the correction of potential PCR bias. **(c)** Sequences of anchor-containing molecules after first (top) and second PCR (bottom). The P7 site can contain a heterogeneity spacer (HS) which improves diversity of the start of read 2 and therefore sequencing quality. The single NGS reads for sequencing insert and indices are also shown. Note, that the index read 2 is used to sequence i5 index as well as the UMI.

First, it needed to be validated if the halfY adapter design has acceptable performance in comparison to the earlier used vectorette adapter design. An experiment was performed in which genomic DNA was framgented by sonication and either vectorette or halfY adapter were ligated to the end repaired fragments. The genomic DNA was prepared from pooled yeast libraries which were constructed using the CASTLING strategy (discussed in more detail in the following section 3.2) and contained a mixture of strains with C-terminally tagged genes at various strain concentrations. Four different genes tagged in those libraries were quantified by qPCR and the enrichment with respect to the quantification in the input genomic DNA was determined (Figure 3.3). Compared to the vectorette adapter the halfY adapter design resulted in a 30-fold reduced enrichment. Nevertheless, after

enrichment the abundance of these genes was still almost 30 million fold over the genomic DNA abundance and was therefore considered as being sufficient for successful genome walking sequencing.

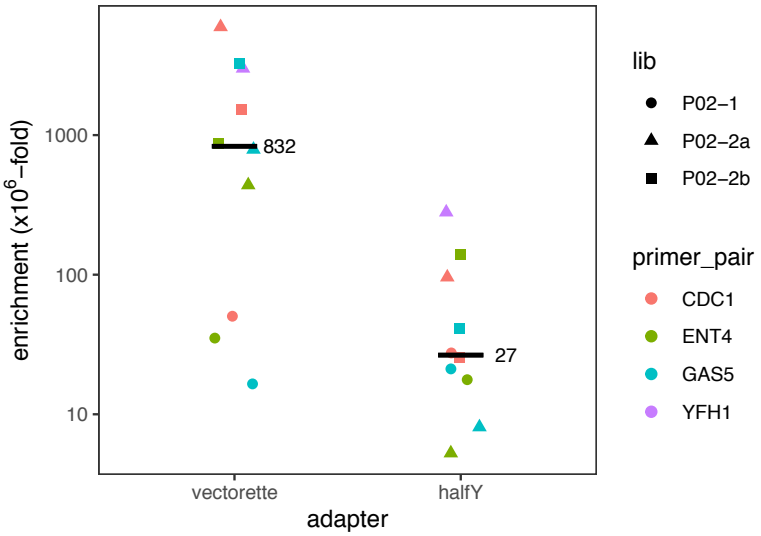


Figure 3.3: Enrichment factors of two different Anchor-Seq adapter designs. Genomic DNA from three different pooled C-terminal tag libraries were used as samples. The libraries (P02-1/-2a/-2b) were prepared by the CASTLING strategy (see following section 3.2). The respective DNA was used as input for an Anchor-Seq enrichment but using either the standard vectorette or halfY adapter design. Four different genes tagged in those libraries were quantified by qPCR (primer_pairs) and enrichment was calculated with respect to the input genomic DNA. For each gene the mean of three technical replicates is shown. Horizontal bars indicate median of all quantification and the bars are further labeled with the respective enrichment value (x10^6-fold).

After validating that enrichment can be performed with a halfY adapter design the next step was to establish conditions for tagmentation which perform well in the Anchor-Seq context. A critical aspect for Anchor-Seq performance is the size distribution of the NGS library molecules. The longer the molecules are the more they inform about the sequence adjacent of the anchor sequence. Molecule size distributions after tagmentation are partially determined by the concentration ratio of Tn5 enzyme to genomic DNA. This mainly depends on the activity of the Tn5 preparation and should be determined for every batch individually [86]. I performed a tagmentation experiment in which I added various amounts of loaded Tn5 complex to a defined amount of genomic DNA (ratio 0.6 to 10) and used this reactions as input for the first PCR of the proposed Tn5-Anchor-Seq protocol. Material from the first PCR step was then used for the second PCR step and both PCR reactions were resolved by DNA gel electrophoresis to inspect approximate fragment size distributions (Figure 3.4). I could observe that amplicon sizes decreased with increasing amounts of loaded Tn5 complex in the reaction. A broad size distribution between approximately 0.3 and 3 kbp was observed for reactions in which 0.6- to 2.5-times loaded Tn5 complex was used with respect to the input DNA concentration. Since this size distribution is favorable for the Tn5-Anchor-Seq application I decided

to use a concentration ratio (Tn5:gDNA) of 0.75 for all subsequent experiments. This value is flexible enough to tolerate some variability in the input DNA concentrations (approximately 3-fold differences). An additional reaction in which SDS was included to reduce enzyme activity indicates that amplicon complexity is reduced when the available Tn5 complex becomes limited in the tagmentation reaction (see single band and banding pattern for 0.4 % +S samples in Figure 3.4b and c respectively). To achieve high resolution results it is therefore advantageous to not reduce Tn5 concentration too much.

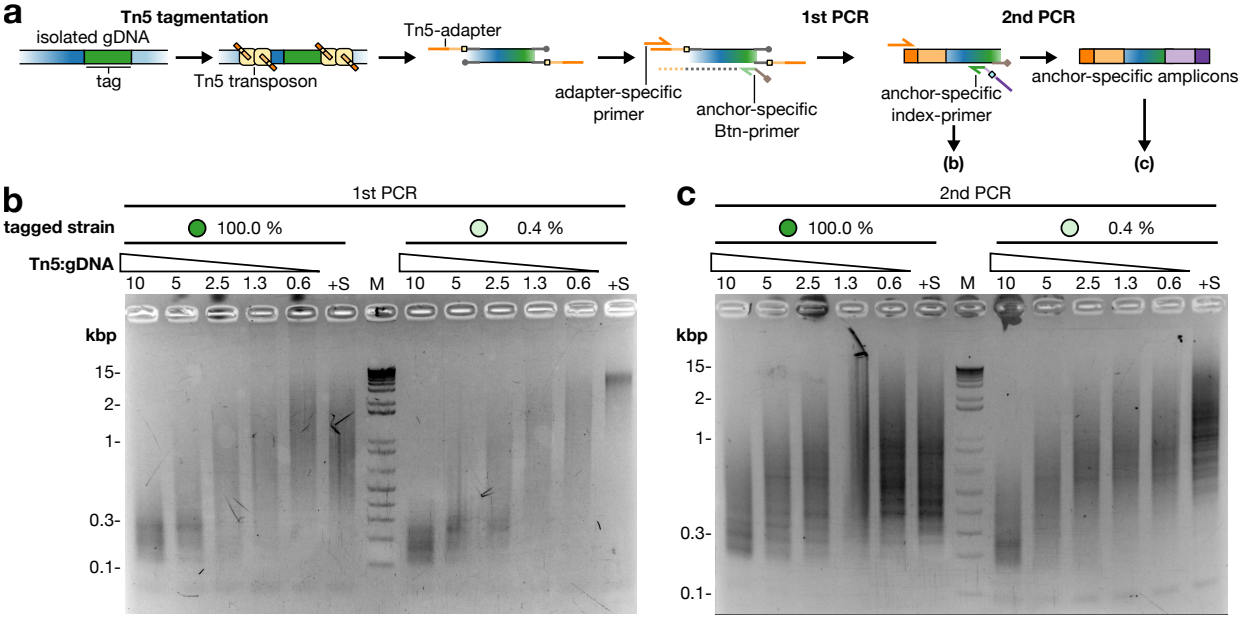


Figure 3.4: Ratio of Tn5 enzyme to DNA determines fragment size distribution. **(a)** Tn5 tagmentation reactions were set up with different amounts of Tn5 complex added to the same amount of DNA. Tagmented DNA was used as input for two subsequent PCRs to generate Tn5-Anchor-Seq amplicons. The PCR products were analyzed by DNA gel electrophoresis to evaluate amplicon size distributions. **(b)** A yeast strain in which the gene *SGF29* was C-terminally fused to mNeonGreen was used as tagged strain. Genomic DNA was either isolated from this strain alone (100 %) or from a mixture of this strain with an untagged strain (BY4741) to simulate reduced abundance of this genotype (0.4 %). 2.5 ng/µL of gDNA was used for tagmentation with the indicated Tn5 concentrations (gDNA concentration multiplied by Tn5:gDNA factor). First PCR reactions were set up using the whole tagmentation reaction as template and run for 30 cycles. Reactions were resolved by DNA gel electrophoresis. An additional reaction in which a 2.5 Tn5:gDNA reaction was performed in the presence of 0.2 % SDS (+S) indicates reduced tagmentation. As marker (M) 1 Kb Plus DNA Ladder (ThermoFisher Scientific) was used. **(c)** Second PCR reactions analyzed as in panel b. PCR reactions were set up with 1 µL of the first PCR reactions and run for 20 cycles.

After establishing the general conditions of the tagmentation reactions for Tn5-Anchor-Seq I needed to evaluate the enrichment achieved with this protocol. Genomic DNA from three pooled tag libraries (prepared by genome-wide CASTLING, see following section 3.2) were used as starting material for Anchor-Seq preparations using the halfY adapter or Tn5-Anchor-Seq preparations including or excluding the biotin-streptavidin-capture step (Figure 3.5). The size distribution of the respective Anchor-Seq amplicons after the second PCR showed various outcomes for the different preparation approaches (Figure 3.5a). Interestingly, Biotin-streptavidin-capture lead to a broader and more uniform size

distribution over the range of 0.15 to 3 kbp when compared to the Tn5-Anchor-Seq preparation without biotin-streptavidin-capture step. Amplicons of size 0.4 to 0.8 kbp were extracted from the agarose gels and used as input for quantification of selected strains in the CASTLING pool by qPCR. This analysis showed an enrichment of approximately sevenfold for the Tn5-Anchor-Seq without capture step in comparison to the Anchor-Seq protocol with halfY adapters (Figure 3.5b). Performing Tn5-Anchor-Seq with the additional biotin-streptavidin capture step seemed to moderately improve enrichment further by approximately threefold (Figure 3.5b). This let me to conclude that the newly developed Tn5-Anchor-Seq protocol can be used to selectively enrich for anchor-containing molecules of interest. The inclusion of the optional biotin-streptavidin-capture step leaves further room for experimental adjustments depending on the experimental demands.

Another important question which needed to be addressed was the performance of the Tn5-Anchor-Seq protocol in an actual NGS experiment. To test if the protocol achieves reproducible quantification, Tn5-Anchor-Seq reactions of a controlled sample were prepared. The 15 plates of an arrayed yeast library in which each strain represents a different C-terminally tagged gene were pooled in various amounts representing a dilution series spanning approximately six orders of magnitudes. Genomic DNA of this library was then used to perform Tn5-Anchor-Seq reactions either in UP sequencing direction or for UP and DOWN sequencing simultaneously (Figure 3.6a). This confirmed high reproducibility and linearity over five orders of magnitude showcasing the potential of Tn5-Anchor-Seq for genotype quantification. At very low genotype fractions (below one million percent) some deviations from the linear trend can be observed which could hint at small number effects and dilution errors (Figure 3.6b). For the sample in which UP and DOWN amplicons were amplified within the same Tn5-Anchor-Seq reaction I noticed that five times more reads were observed for the UP than for the DOWN amplicon direction. This indicates that the primers required for the individual Tn5-Anchor-Seq amplicon directions might need to be carefully fine-tuned in case both directions should be enriched within the same reaction.

In summary, I could implement an adaptation of Anchor-Seq to tagmentation resulting in the Tn5-Anchor-Seq protocol which allows a more scalable introduction of the Anchor-Seq adapters.

3.1.2 Redesign of the computational workflow for genome walking NGS

The computational pipeline which was initially used to validate the C-SWAT library was specifically designed for this particular purpose. I will briefly summarize this original design. A reference database of sequences was constructed by concatenating the part of the constant sequence of the tagging cassette spanned by the sequencing read and 140 nt of the 3'-terminal sequence of the coding sequences of each ORF excluding the stop

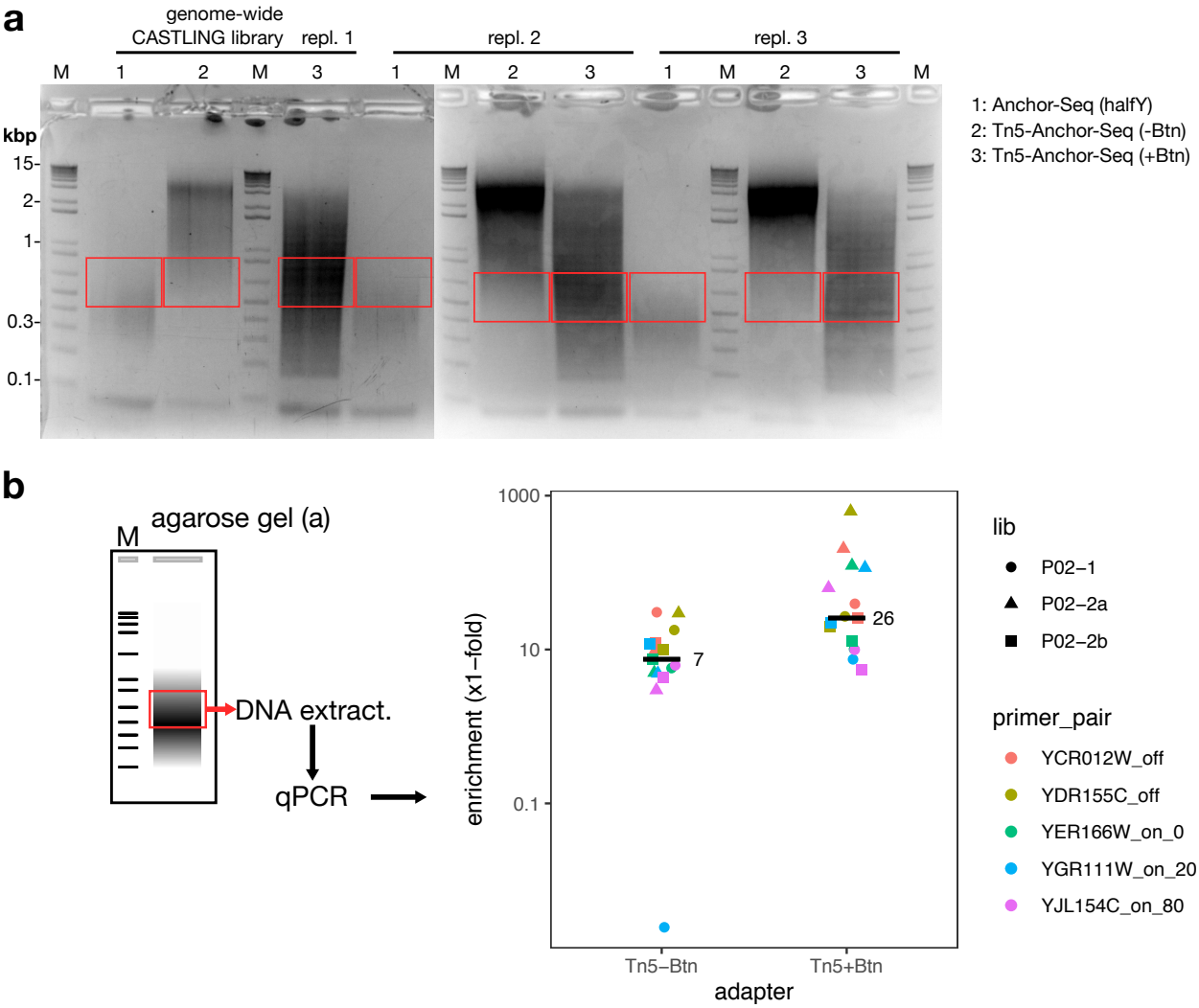


Figure 3.5: Enrichment by Tn5-Anchor-Seq . Pooled genome-wide C-terminal yeast libraries (CASTLING library; see following section 3.2) were processed either with the Anchor-Seq protocol using halfY adapters or with the proposed Tn5-Anchor-Seq protocol excluding or including biotin-streptavidin-capture. **(a)** The different reactions exhibit different size distributions in DNA gel electrophoresis. The red rectangles indicate excised gel sections used analyzed by qPCR (see panel b). As marker (M) the 1 Kb Plus DNA Ladder (ThermoFisher Scientific) was used. **(b)** Relative quantification of five selected tagged genes (primer_pair) in each library by qPCR. DNA enriched by the different Anchor-Seq reactions was size-selected by gel electrophoresis (panel a). Enrichment was calculated for Tn5-Anchor-Seq without (Tn5-Btn) or with biotin-streptavidin capture (Tn5+Btn) with respect to the Anchor-Seq reaction performed with halfY adapter. For each gene the mean of three technical replicates is shown. Horizontal bars indicate median of all quantification and the bars are further labeled with the respective enrichment value.

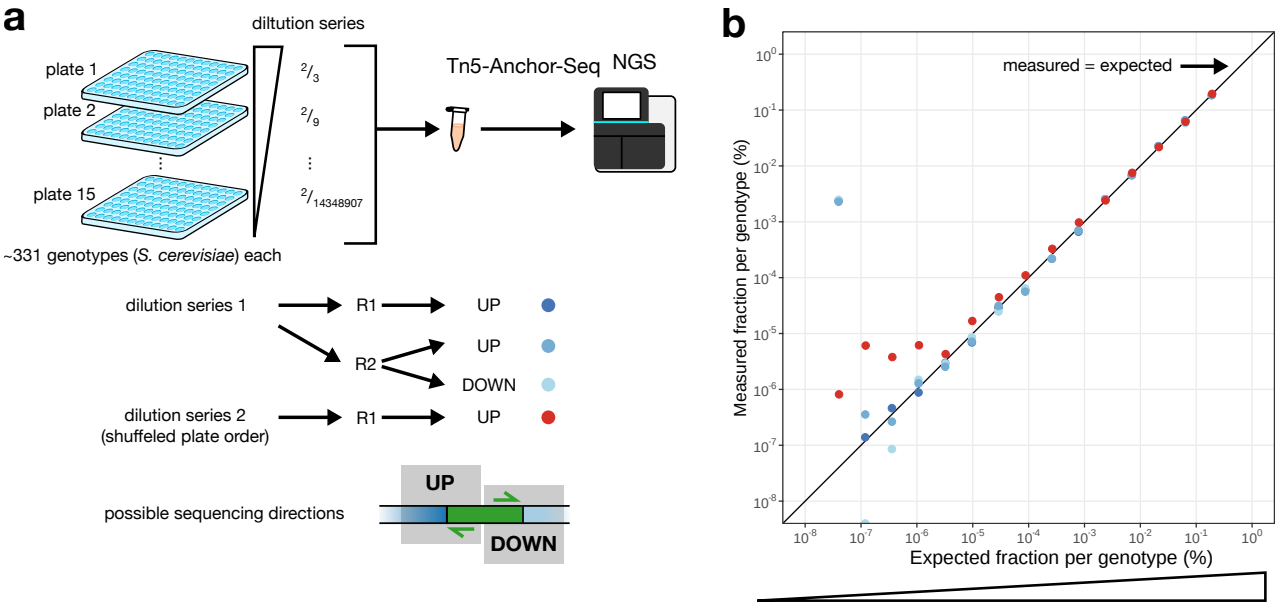


Figure 3.6: Robustness and linearity of Tn5-Anchor-Seq . **(a)** An arrayed genome-wide yeast library in which each strain represents a different gene tagged with a C-terminal tag was pooled in various amounts to create a controlled dilution series which was then analyzed using the Tn5-Anchor-Seq protocol. **(b)** Plotting genotype quantification measured by Tn5-Anchor-Seq against the expected fraction of these genotypes based on their dilution factor revealed high correspondence and linearity (diagonal line with slope 1) over a wide dynamic range. Each point represents the mean of genotype abundancies within one library plate. Some divergence from the expected trend was observed for the genotypes of very low abundance (below 1 millionth percent) with one likely outlier for dilution series 1.

codon. Strict matching of each sequencing read was used to assign the most likely ORF the read originated from. Because errors in the tagging junction might prohibit strict matching, reads were split into two halfs which were matched individually. Additionally, for matching, frameshifts by minus three to plus three nucleotides were used to account for deletion or insertion mutations. This simple assignment procedure then allowed to identify tag junction sequences which were consistent with the expected sequence after successful tagging.

This computational workflow however had several limitations which I needed to address for potential later applications of Anchor-Seq . First, this procedure could only be used to evaluate insertion sites which were known *a priori* so that a respective reference sequence could be added to the database. Second, the analysis was confined to the first read only although paired-end sequencing was performed. The second read contains sequence information adjacent to the vectorette primer more distal from the anchor sequence. This sequence can help to further improve identification of the insertion site especially in case of redundant genomic regions. Third, strict matching although applied in several variations might not account for all possible mutations and more sophisticated alignment procedures can be used. Fourth, in the beginning demultiplexing of samples could be performed with Illumina's software as the sample-specific barcodes were present as barcodes within the sequencing adapters. Later versions of the Anchor-Seq protocol made

also use of more complex barcode variations (internal and in sequencing adapter) and unique molecular identifiers which needed to be accounted for during the downstream computational analysis.

In order to address all these issues, I designed a computational workflow which works in the reverse direction (Figure 3.8). Constant sequences are first trimmed off the sequencing reads. This step also serves as quality control to remove read pairs which are not informative because the molecules from which they originated were too short or they result from unspecific amplification and do not contain the expected constant sequences. The trimmed reads are then aligned to a reference genome of the species in which the insertion experiment was performed. This leads to a pile-up of reads at the insertion site and this alignment is characteristic in that it has a random start position at the site of the Anchor-Seq adapter and a specific position at the transition site to the integrated sequence (Figure 3.8). The alignments are then counted by grouping reads by these specific positions and identifying overlapping genomic features (e.g. ORF boundaries).

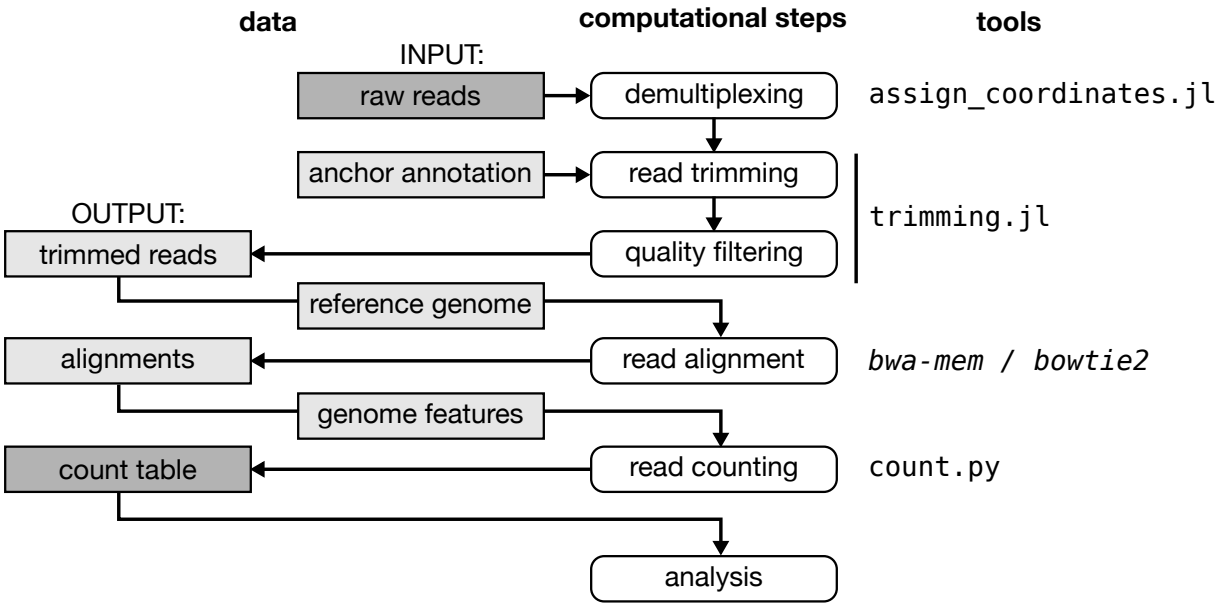


Figure 3.7: Design of the computational Tn5-Anchor-Seq workflow. A custom workflow was designed to process Tn5-Anchor-Seq sequencing data starting with raw Tn5-Anchor-Seq reads and resulting in a count table which can then further analyzed depending on the Tn5-Anchor-Seq application. The workflow is modular in that most individual computational steps are implemented using single tools. Italic tool names are referring to external software while custom scripts written for this workflow are not italic.

In this section I showed how the Anchor-Seq protocol was updated to a more versatile and scalable workflow using tagmentation. Furthermore, I presented how the redesign of the computational Anchor-Seq workflow enables more comprehensive analysis. The following sections will be concerned with different applications of the Anchor-Seq and Tn5-Anchor-Seq protocols and how they can help to implement and validate various experimental methodologies.

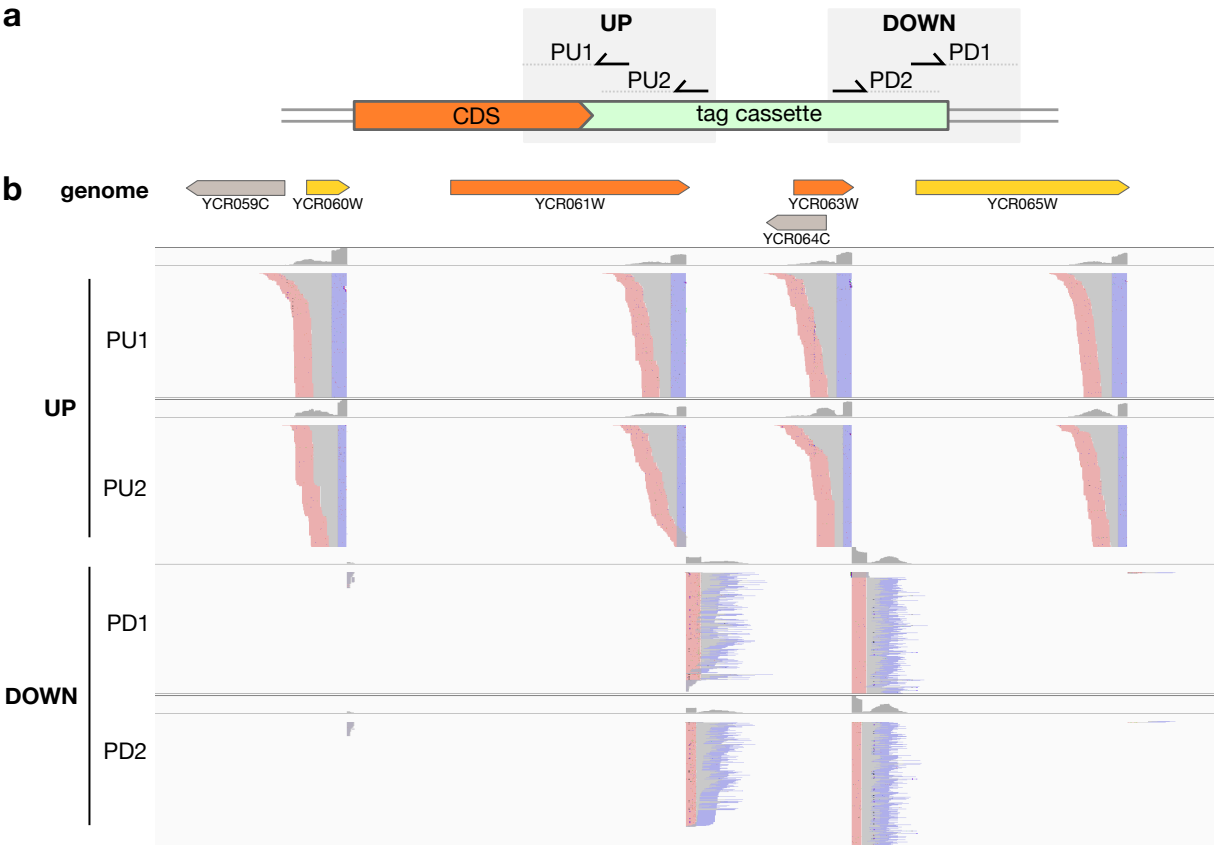


Figure 3.8: Exemplary view of an Tn5-Anchor-Seq alignment. Several C-SWAT strains in which *Saccharomyces cerevisiae* genes had been tagged with a C-terminal tag cassette were pooled and analysed by Tn5-Anchor-Seq . **(a)** Four different Tn5-Anchor-Seq reactions were prepared and sequenced. Two UP and two DOWN sequencing reactions each with two different anchor primers (PU1 and PU2 or PD1 and PD2 respectively) were used. **(b)** Each row represents one of the four Tn5-Anchor-Seq reactions. Reads were trimmed and aligned to the *Saccharomyces cerevisiae* reference genome as described in the main text. Read alignments piled up at the 3' end of the ORFs (before the stop codon) as this is where the C-terminal tag cassette is integrated. Tn5-Anchor-Seq alignments exhibited a strict end at one side where the cassette genome junction is situated and a random end to the other side where the Tn5-Anchor-Seq adapter is ligated to the randomly framgented DNA. Two ORFs showed such an alignment for all four sequencing reactions (orange) while only the UP alignments are present for two ORFs (yellow) indicating that the cassette was joined at its 3' end to a different genomic position than expected. Two ORFs exhibited no alignment indicating that they were not tagged (gray).

3.2 Pooled tag library construction in the the yeast *Saccharomyces cerevisiae* with CASTLING

Arrayed *Saccharomcyes cerevisiae* libraries such as the C-SWAT collection have and will continue to be useful resources for biological studies in this organism. However, their initial creation is time and resource intensive and their application is limited to the genetic background they were constructed in. Therefore, arrayed library construction and the analysis thereof will become a bottleneck in future functional genomics studies. Pooled, 'shotgun'-like approaches might provide a promising alternative [1]. So far, several methodologies have been developed which allow for pooled yeast library construction using microarray-synthesized oligonucleotide pools and CRISPR-Cas9-mediated gene targeting. These methodologies are limited in that they allow only for the introduction of single-nucleotide polymorphisms and small indels [138, 139, 140, 141, 142, 143]. For more complex genotypes such as the tagging of endogenous genes large genomic insertions are required. Therefore, we set out to conduct a proof-of-principle study for the pooled construction of tagging genotypes in *Saccharomyces cerevisiae* to further extend the available toolset for functional genetics in this organism [116].

The strategy which we termed CRISPR-Cas12a-assisted tag library engineering (CASTLING) builds upon the following ideas. First, the efficiency of gene tagging must be increased to allow for pooled library construction. This is achieved by inducing DSBs at the target sites and simultaneously providing a repair template to stimulate HR. For DSB induction the crRNA-guided DNA-endonuclease CRISPR-Cas12a is used because the coverage of potential target sites at gene termini is higher compared to CRISPR-Cas9 which is important for gene tagging applications [116]. Second, all the information required for this targeting should encoded in the same molecule. This is achieved by including also the gene for crRNA expression in the tagging cassette. Consequently, these molecules were termed self-integrating cassettes (SICs). The crRNA structure of CRISPR-Cas12a allows for its incorporation in the SIC design which would have been difficult with CRISPR-Cas9 [116]. Finally, a molecular recombineering strategy needed to be implemented which allows for the pooled construction of such SICs. These SIC pools are then transformed into a yeast strain which expresses CRISPR-Cas12a leading to the recovery of a pooled yeast library.

3.2.1 SICs allow for pooled tag library construction with CASTLING

Gene tagging is facilitated by the usage of SICs because of several reasons. Once such a molecule enters a cell the crRNA is expressed from a gene encoded on the SIC itself. The CRISPR-Cas12a enzyme also expressed in the cell forms a complex with this crRNA and can then introduce a DSB near the target site. The created lesion is repaired using the tagging cassette as template for HR (Figure 3.9). It was first tested if this SIC design allows for increased tagging efficiencies for individual genes. SIC constructs for

a C-terminal fluorescent tag targeting several highly expressed genes were transformed into yeast cells expressing the CRISPR-Cas12a enzyme from *Francisella novicida U112* (FnCas12a) which was confirmed to be functional in this organism [116, 144]. Tagging efficiency, quantified by the number of colonies after transformation, revealed a ~50-fold to several thousand-fold increase compared to transformations in cells which did not express the CRISPR-Cas12a enzyme (Figure 3.10a). Furthermore, quantification of the fraction of fluorescent colonies showed that tagging fidelity increased from 67-91 % to 96-99 % (Figure 3.10a).

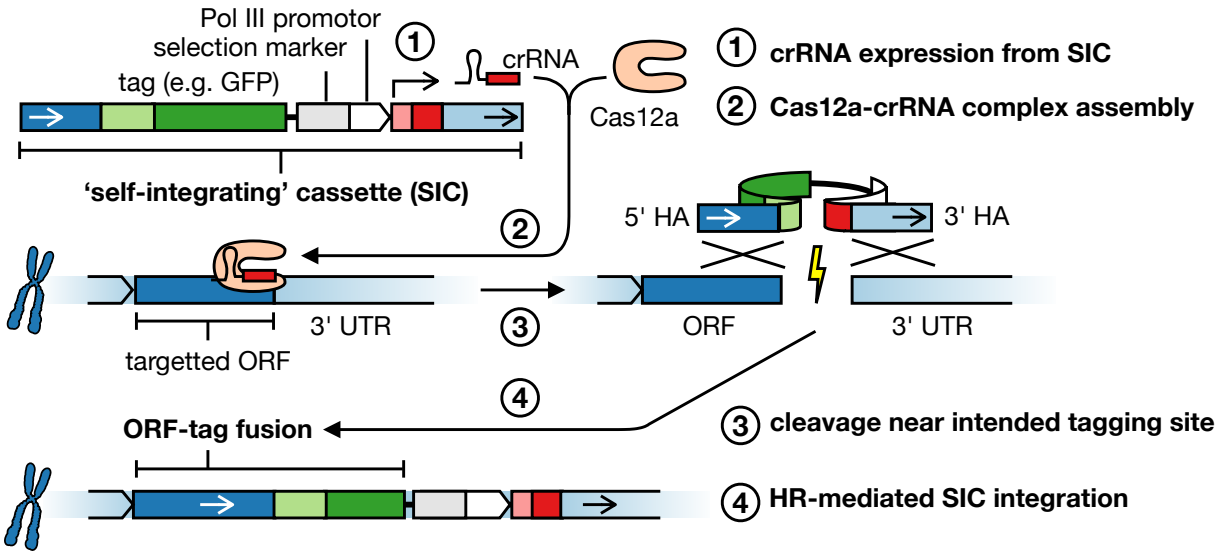


Figure 3.9: Self integrating cassette (SICs) principle for gene tagging. SICs contain all the genetic information required for gene targeting with large insertions (i.e. gene tagging). Here the SIC concept is illustrated for C-terminal gene tagging. The crRNA for guiding the CRISPR-Cas12a endonuclease is expressed from a dedicated gene placed at the 3'end of the SIC (1). The Cas12a-crRNA complex is formed (2) and introduces a DSB close to the target site at the C-terminus of the target gene (3). Directed by the terminal homology arms (5'HA and 3'HA) the SIC is further used as template for repair of the genomic lesion by HR (4). *The figure was reproduced in modified form [116] with permission and was jointly created by Benjamin Buchmuller, Michael Knop and me.*

All the genetic information required for gene tagging are combined in SICs. This becomes especially useful if one wants to utilize oligonucleotide pool synthesis to design highly multiplexed genome engineering efforts like it is necessary for pooled library construction. Oligonucleotide pools allow to include thousands of individually designed sequences very cost-efficiently in the same experimental design so that system-wide studies become feasible. The target information (i.e. homology arms and crRNA sequences) can be entirely encoded in the relatively short oligonucleotides while the tagging cassette encodes for the tag functionality. The SIC concept therefore separates the target information by coding it on individual sequences in the oligonucleotide pool from the functionality information which is programmed on a generic tagging cassette to be integrated. Co-integration from two simultaneously transformed SIC species was rarely observed which indicated the feasibility of specific gene targeting in pools of many different SICs [116]. The challenge is to combine both information so that the cassette carries homologies on either site of

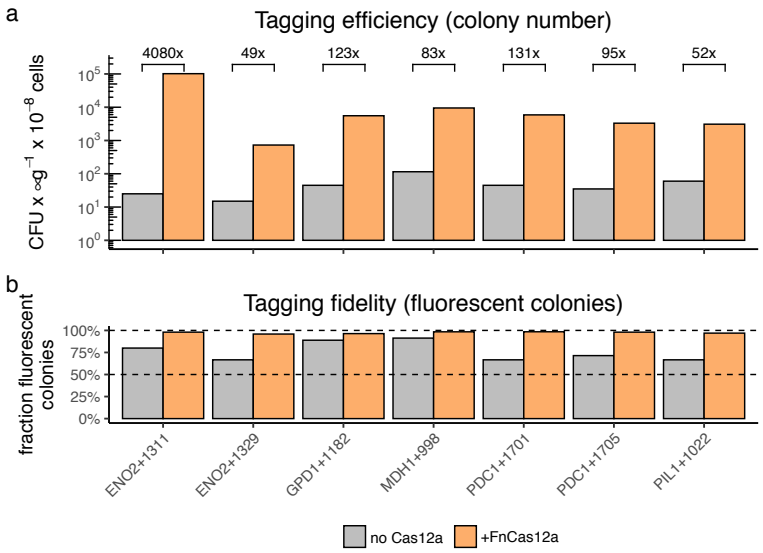


Figure 3.10: SICs enhance tagging of single genes. SICs (targeted genomic positions labeled at the bottom) were individually transformed into wild-type yeast cells or cells expressing the CRISPR enzyme FnCas12a. **(a)** Tagging efficiency was quantified as normalized colony number after transformation. The approximate fold change of colony numbers from transformations with and without FnCas12a expression is indicated above each bar pair. **(b)** The fraction of all colonies exhibiting fluorescence was quantified. Since the tagged genes are all highly expressed this fraction is indicative for the tagging fidelity. *The figure was updated from [116] and the underlaying data was accquired by Benjamin Buchmuller and Matthias Meurer.*

the tagging cassette without any formation of chimeras between molecules in the pool. For this a molecular cloning strategy was devised and implemented as a procedure to generate such molecules fully *in vitro*. The oligonucleotide pool is first amplified by PCR using primers with homologies to the tagging cassette. These two molecule species are combined by *in vitro* homology-directed assembly (using the NEBuilder HiFi DNA Assembly kit from NEB). This leads to circularization of the molecules which are then used as templates for rolling circle amplification (RCA) with phi29 DNA polymerase. A restriction site included in the oligonucleotide design between the two homology arms is then used to release monomeric tagging cassettes from the concatemeric RCA product. The resulting material is then used for transformation into *Saccharomyces cerevisiae* for pooled tag library construction. Such a pooled library can then be analyzed by a cellular assay which can fractionate the library into phentoypic bins (e.g. fluorescence-activated cell sorting). These bins can then be individually genotyped using a sequencing strategy for tag insertion site determination (e.g. (Tn5-)Anchor-Seq). These steps comprise the full CASTLING strategy (Figure 3.11) [116].

3.2.2 CASTLING characterization using a library with defined phenotype

In order to evaluate the CASTLING approach an experiment was conducted in which a defined subset of the budding yeast genes was targeted. I selected 215 genes for which the gene products under standard growth condition are localized to the nuclear

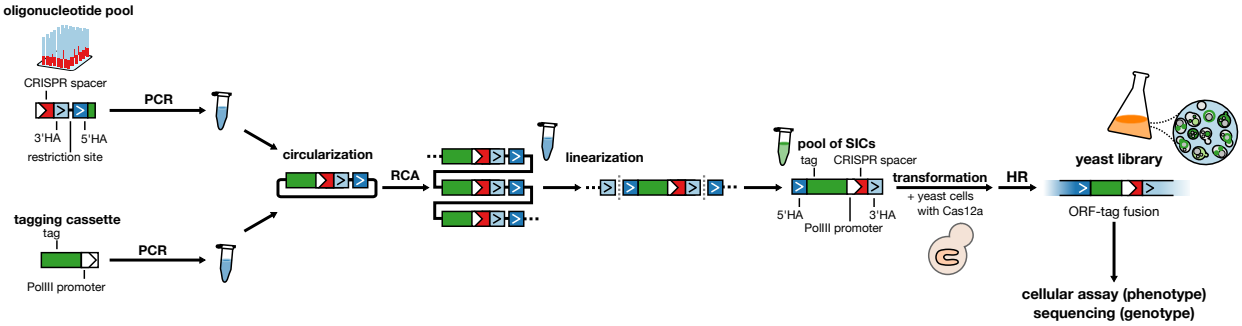


Figure 3.11: CRISPR-Cas12a-assisted tag library engineering (CASTLING). The CASTLING strategy encompasses the steps required for SIC pool generation and transformation to retrieve a pooled yeast clone library. Individual sequences in the oligonucleotide pool contain all the gene-specific elements of SICs while the tagging cassette encodes tag functionality and selection marker. After oligonucleotide and tagging cassette amplification by PCR both molecule species are combined using homology directed *in vitro* recombineering leading to circular molecules. These then serve as templates in a rolling circle amplification (RCA) reaction using phi29 DNA polymerase. By enzymatic digestion of the concatemeric RCA products individual SICs are released. The resulting SIC pool is then tranformed into a yeast strain which also expresses the CRISPR-Cas12a endonuclease and a pooled library is retrieved in which each clone carries a different tagging genotype. This library can then be phenotypically characterized by cellular assays, fractionated into individual phenotype bins and genotyped with (Tn5-)Anchor-Seq .

compartment [4] and have appropriate expression strength to be easily detected by microscopy [134, 145]. In total, 1,577 oligonucleotides (on average seven oligonucleotides per gene) were designed which covered all suitable PAM sites within 30 bps of the stop codon of the selected genes. This pool design was purchased independently three times from two different suppliers to also investigate the impact of production-specific differences on tagging efficiency with CASTLING. Input amounts of the oligonucleotide pools were adjusted so that 20 cycles of PCR yielded enough material for the subsequent CASTLING steps. Noteworthy, we needed to use about 270-times more material for pool A from supplier A than for pools B1 and B2 from supplier B. Pool B2 was also amplified using approximately seven-times more starting material to test how genotype complexity of the resulting CASTLING library is dependent on this initial step. For each condition SIC pools were generated in duplicates and transformed individually into a yeast strain expressing FnCas12a yielding 30,000 to 95,000 clones per condition, i.e. more than 100 clones per targeted gene and condition (Figure 3.12a).

I performed fluorescence microscopy of the resulting libraries and quantified the observed fluorescence localization phenotypes (Figure 3.12b). This allowed me to determine that tagging fidelity ranged from 90 to 95 % based on the fraction of cells exhibiting the expected nuclear localization signal (Figure 3.12c). The tagging fidelity of a pooled library constructed by CASTLING is therefore almost as high compared to the one achieved for individual genes when DSBs are induced in parallel with CRISPR-Cas12a. The remaining cells exhibited either almost exclusively cytoplasmic fluorescence (0-4 %) or no fluorescence at all (2-8 %) (Figure 3.12c). By sequencing the junction between tag and genome of individual clones I observed predominantly single base deletions (Supplementary Figure

4.2). A likely explanation for unsuccessful tagging events are errors in the oligonucleotides used for CASTLING as small deletions are typically introduced during their chemical synthesis [146].

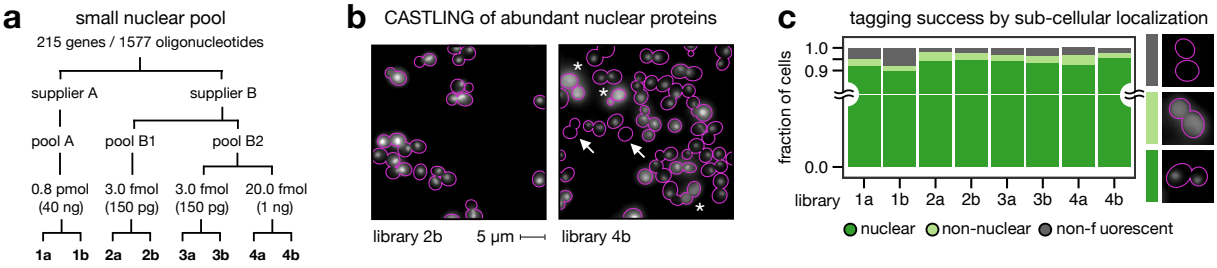


Figure 3.12: Small CASTLING library with nuclear localization phenotype. **(a)** A pool was designed targeting genes which show localization to the nuclear compartment when C-terminally tagged with a fluorescent protein tag. To explore library quality determinants the pool was purchased several times from different suppliers and used to prepare eight CASTLING libraries with a C-terminal mNeonGreen tag. The amount of oligonucleotide pool used for SIC pool production is indicated. **(b)** Representative mNeonGreen fluorescence micrographs from two CASTLING libraries. Most cells show fluorescence signal localized to the nuclear compartment. In addition, some cells exhibit no fluorescence (indicated with an arrow) or diffuse cytoplasmic localization (indicated with an asterisk). **(c)** Quantification of the three phenotypes described in panel b across all eight CASTLING libraries. At least 1,000 cells per library were used for quantification. *The figure was reproduced in modified form from [116] with permission and was jointly created by Benjamin Buchmuller and me.*

Next, the genotype complexity of the different steps of the CASTLING library preparation protocol needed to be quantitatively characterized. For this, NGS libraries were prepared for the reactions after PCR of the oligonucleotide pools and after RCA. Furthermore, the final yeast libraries were analyzed by Anchor-Seq (Figure 3.13). Sequencing was performed in DOWN direction of the CASTLING cassettes covering the crRNA sequence so that the oligonucleotides could be identified. UMIs were included in these preparations to correct for potential PCR bias introduced using the NGS library preparations. Overall, it could be observed that each step of the CASTLING preparations reduces the genotype complexity due to sequence losses (Figure 3.13a and b). Pools from supplier B performed better than from supplier A. Best performance was achieved for the replicates using the pool B2 and higher amounts of starting material (i.e. reaction 4a and 4b). For this pool approximately 50 % of the designed sequences were present in these SIC preparations but because several SICs have been designed for each gene this accounted for the recovery of approximately 90 % of the targeted genes indicating that at least for this pool design of medium complexity high coverage is feasible (Figure 3.13a and b). The likelihood of loosing certain oligonucleotide sequences during RCA corresponded to their abundance after PCR (Figure 3.13c).

Furthermore, the correlation between replicates after PCR was preserved while it was markedly reduced after RCA (Figure 3.14a). A more detailed investigation of one replicate indicated that 44 % of the oligonucleotide sequences did not change more than twofold and 91 % of the sequences did not change more than tenfold after RCA. Sequences were more

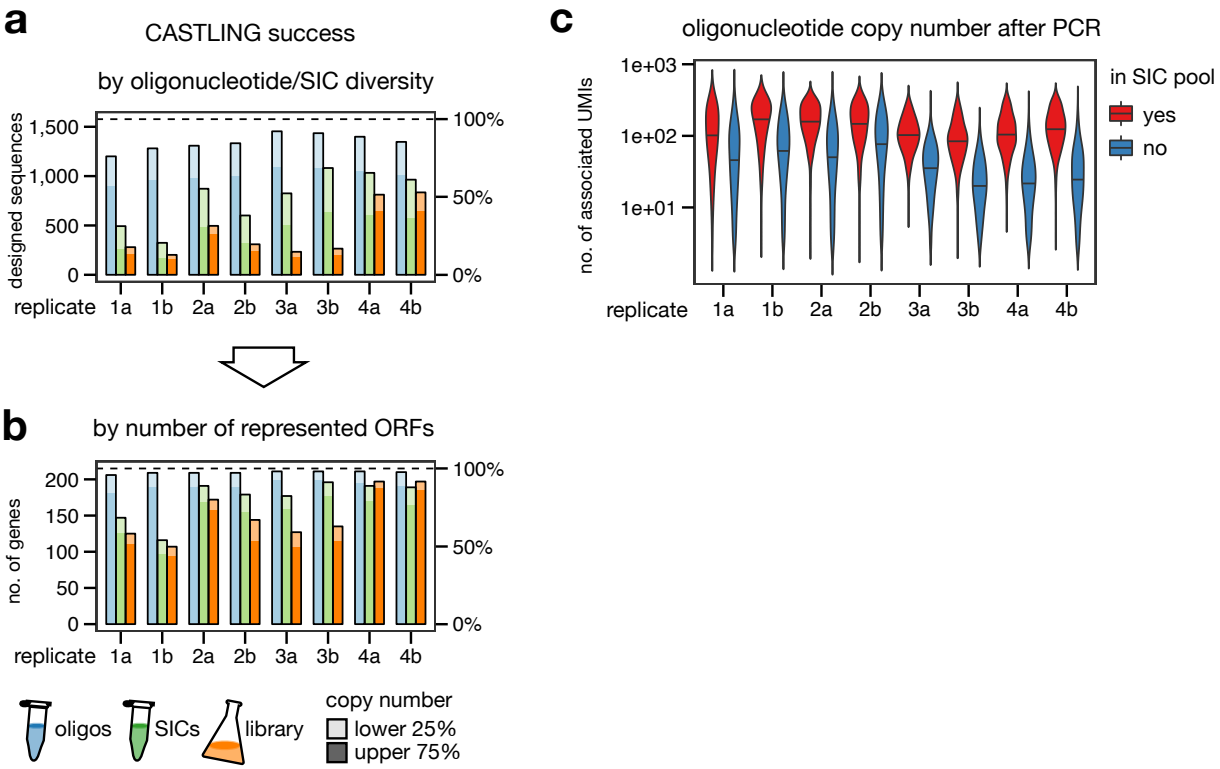


Figure 3.13: Genotype coverage of the eight small CASTLING libraries. Eight small CASTLING libraries (replicate 1a to 4b) were quantifely assessed by NGS at individual steps of the CASTLING process. **(a)** Sequence diversity of oligonucleotide pools after PCR amplifications (blue) and of SIC pools after RCA reactions (green) and yeast libraries (orange) was determined by NGS using UMIs. The full pool design encompassed 1,577 distinct sequences. All sequence instances with the lowest 25 % copy number are light shaded because of their lower confidence. **(b)** Diversity as quantified in panel a but focusing on gene coverage. The full pool design encompassed 215 distinct genes. **(c)** For each CASTLING replicate the copy number based on UMI counts was quantified for sequences which were retained (red) or lost (blue) during SIC construction. *The figure was reproduced in modified form from [116] with permission and was jointly created by Benjamin Buchmuller and me.*

likely to be depleted than to be enriched during this CASTLING step (Figure 3.14b).

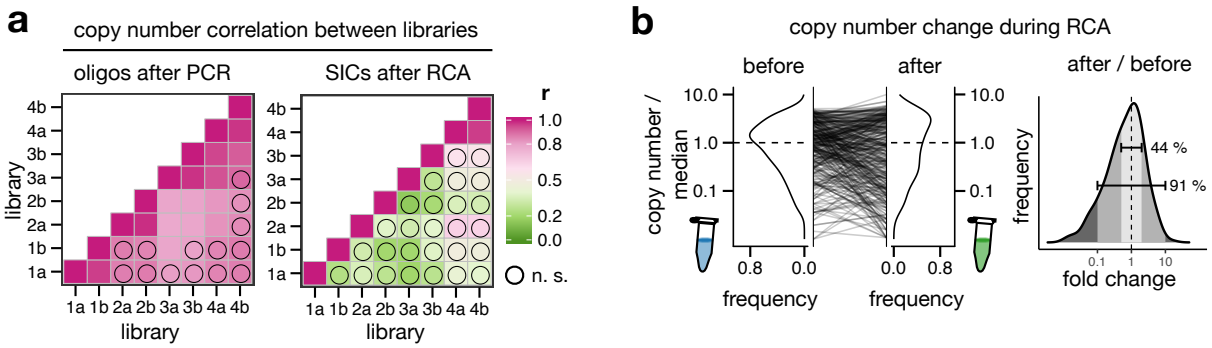


Figure 3.14: Reproducibility during CASTLING library preparation. **(a)** Pearson correlation coefficients (r) of sequence abundances between individual replicates after oligonucleotide pool amplification with PCR (left) or after RCA (right). Nonsignificant correlations (n.s.; $p > 0.05$) are indicated with circles. **(b)** Copy number changes of oligonucleotide sequences from PCR (before) to RCA step (after) in replicate 1a. On the left, kernel density estimates for the copy numbers normalized to the median copy number in that sample are shown. The distribution of fold changes of individual sequences is indicated on the right. 44 % of the sequences changed no more than twofold while 91 % of sequences changed no more than tenfold. *The figure was reproduced in modified form from [116] with permission and was jointly created by Benjamin Buchmuller and me.*

In order to inspect how balanced genotype representation is in the individual steps of the CASTLING process the sequence abundance distributions were examined (Figure 3.15). If each genotype in a pool had the same abundance, the pool would be perfectly balanced. Such an abundance distribution would be characterized by a straight line with a slope of one while increasingly skewed abundance distributions would increasingly bend upwards. Already after the PCR step the abundance distribution was unbalanced and up to 20 % of the designed sequences were absent (Figure 3.15a). As one would expect, each subsequent step of the CASTLING process increased the skew in abundances further. The replicates were in better concordance after the PCR step than after the SIC preparation or the library transformation. Furthermore, when the SIC preparation was less skewed in abundance the genotype representation of the final yeast library was higher which confirms that the abundance bias is propagated from step to step. The replicates with higher coverage generally exhibited more balanced abundance distributions than the replicates with lower coverage. Overall this analysis indicates that optimizations focusing on balanced genotype representation in the SIC pools has the potential to improve genotype coverage of the final library.

3.2.3 CASTLING for genome-wide library construction

Having validated that the CASTLING strategy can be used to construct pooled yeast libraries of intermediate complexity we turned to the question how well the strategy scales for genome-wide tagging endeavors. A SIC pool was designed based on optimized selection rules covering 5,940 genes (89 % of all yeast genes) with 27,000 oligonucleotide sequences [116]. Three yeast libraries were constructed. For the first two libraries 30

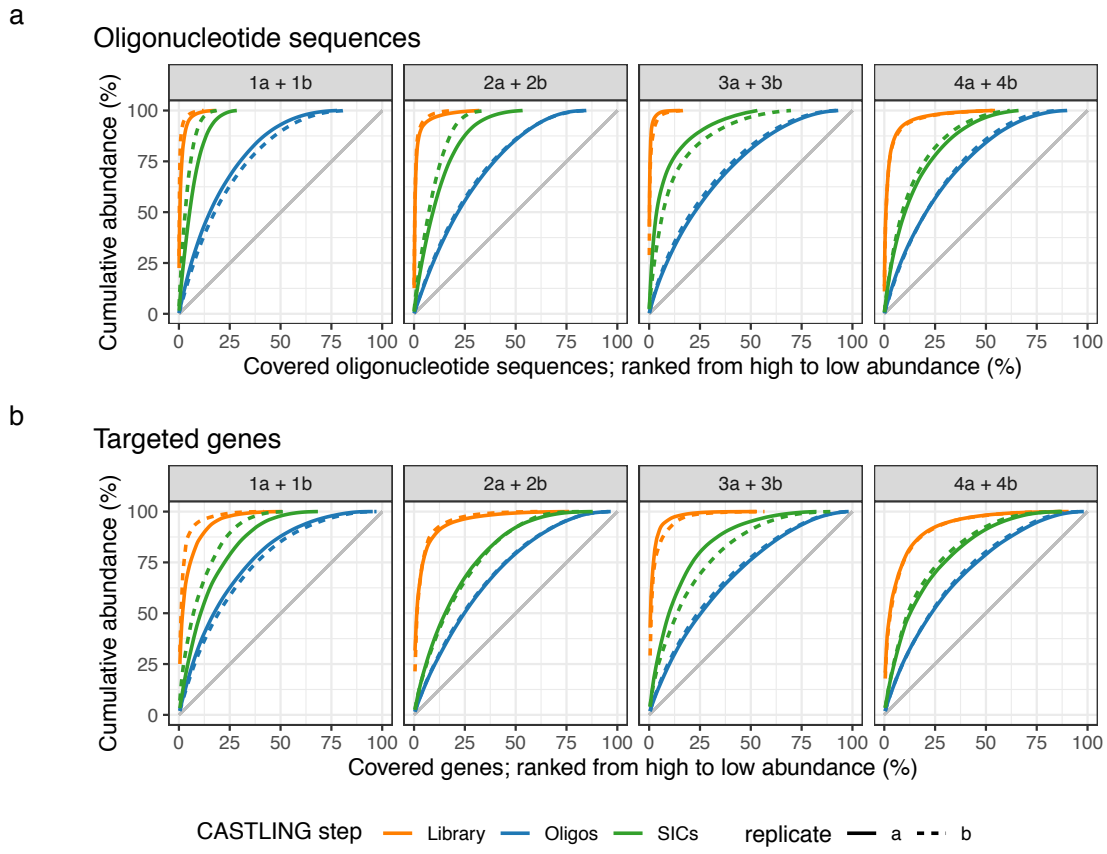


Figure 3.15: Genotype abundance distributions across the CASTLING process. Genotype coverage at different steps of the CASTLING process inspected by cumulative sequence abundance distributions. **(a)** The contribution of a designed oligonucleotide sequence to all sequences (i.e. the cumulative abundance in percent) observed in a particular CASTLING step (colors), pool amplification and replicate (facet and linetype) are indicated with respect to the rank of this oligonucleotide sequence (ordered high to low, from left to right), given in percent of the pool design. The gray line with slope one represents the cumulative distribution for an ideal sequence pool, where each sequence has the same abundance. **(b)** Same as panel a but instead for each targeted gene by summing up the UMI counts of the respective designed oligonucleotide sequences.

RCA reactions were pooled and SICs were generated. To explore how the number of recovered clones influences retrieved genotype complexity a large library encompassing 704,000 clones (LibA) and a small library covering 44,000 clones (LibB) was recovered after SIC transformation. To explore the impact of the RCA reaction on genotype complexity a third yeast library with an intermediate size of 116,000 clones was recovered (LibC) from a transformation with a SIC pool generated from pooling two additional RCA reactions (Figure 3.16). The genotype complexity of the yeast libraries was characterized with Anchor-Seq . The three libraries altogether covered 76 % of all targeted genes. One third of all targeted genes were covered reproducibly in each of the three libraries indicating that a considerable genotype coverage can be robustly achieved with CASTLING. Notably, almost 50 % coverage of a genome-wide library could be observed when analyzing LibB for which 7-fold more clones than targeted genes were recovered. Analyzing LibA instead which had 119-fold more clones than targeted genes resulted only in a moderate increase in total genotype coverage of 64 %. Almost all genes covered in the small library LibB were also covered in the larger LibA library as it would be expected when the same SIC pool is used for transformation. On the other hand 20 % of genes observed in library LibC which was constructed using the independent RCA pool preparation were unique to this library while only 14 % of genes observed in the large library LibA were unique. This indicates that the recovered genotype complexity depends more on independent SIC pool preparations than on the number of recovered clones.

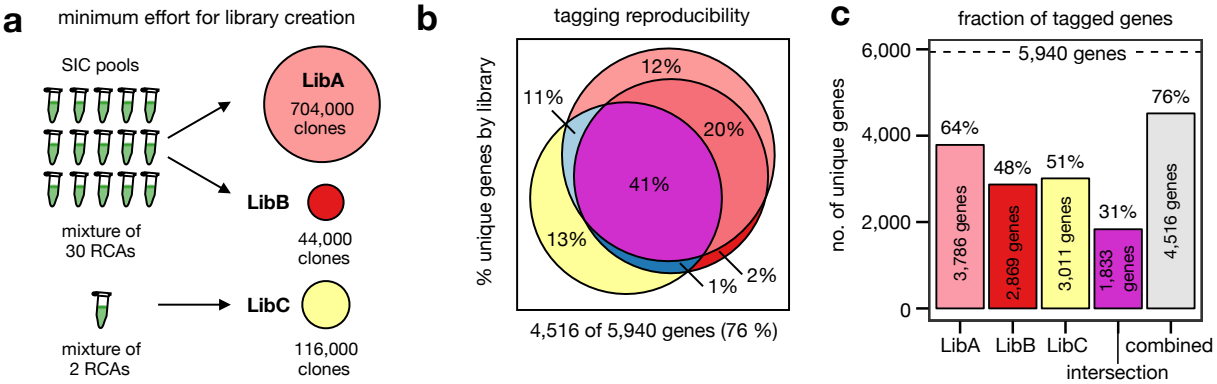


Figure 3.16: Genome-wide library construction with CASTLING. A CASTLING pool design targeting 5,940 yeast genes was used to prepare several libraries to explore the retrieved tag genotype coverage. **(a)** Experimental design to explore different determinants of genotype complexity. Three CASTLING libraries were constructed. To probe the influence of the retrieved number of clones after transformation the same SIC pool prepared from many RCA reactions was used to recover a large library and a small library, encompassing 119-fold and 7-fold more clones than targeted genes respectively. To compare how the SIC preparation influences coverage a third library of intermediate size (20-fold more clones than targeted genotypes) was prepared from an independent SIC pool which was prepared from fewer RCA reactions. **(b)** Number of identified genes in each of the three libraries and overlap thereof. Percentages are given with respect to the total number of genes observed across all libraries (4,516 genes, 76 % of targeted genes). **(c)** Comparison of covered genes achieved per library. *The figure was reproduced in modified form from [116] with permission and was jointly created by Benjamin Buchmuller and me.*

3.2.4 Application of CASTLING for proteome profiling

We also performed a proof-of-concept experiment to showcase the use of a CASTLING library for screening applications. A CASTLING library with 109,000 clones in which genes had been tagged with a C-terminal mNeonGreen tag was used for protein abundance profiling using fluorecence intensity as a proxy. Fluorescence-activated cell sorting (FACS) was performed on this library to enrich for fluorescent cells and the genotype complexity of the pre- and post-FACS libraries was determined by Anchor-Seq . Altogether both libraries covered 2,335 genes of which 848 were enriched and detected in both libraries, 283 were only detected in the post-FACS library (i.e. they were enriched), 732 were depleted and detected in both libraries and 472 were only detected in the pre-FACS library (i.e. they were also depleted). Taken together, this yields 48 % enriched and 52 % depleted genes. Across three previous flow cytometry studies between 34 % and 54 % [147, 148, 149, 150] of the 4,159 genes tagged in the Yeast GFP Clone Collection [4] could be detected which shows that the fraction of FACS-enriched genes we observe is as expected. The post-FACS library was further fractionated into eight fluorescence bins using a second round of FACS. For each fluorescence bin Anchor-Seq libraries were prepared to determine which genes are contained in those bins. Because this experiment was primarily used to showcase a potential application workflow of a CASTLING workflow we turned to performing the NGS on a MinION device from Oxford Nanopore Technologies which gave us access to the sequencing data more readily than the Illumina-based NGS which we usually performed. Based on the number of cells and the reads observed for each gene in each bin I was able to recover a fluorescence profile estimate for each genotype. However, the read depth of that experiment was relatively low and we acquired 18,638 informative reads which allowed me to generate an abundance estimate based on the fluorescence intensity for 435 genes. I compared these abundance estimates to the results from previous studies [150]. A relatively high Spearman correlation coefficient of 0.63 was observed with an earlier flow cytometry data set (Figure 3.17, Supplementary Figure 4.3). Further comparisons to data sets generated by mass-spectrometry and immunoblotting revealed several outliers which appeared to be unique to the flow-cytometry dataset. It was therefore questionable if the sparse resolution of this small data set might impede conclusive results. Therefore, I compared correlation of protein abundance estimates between studies either using their full coverage of proteins or only using the subset detected in our proof-of-concept experiment (Figure 3.17c). Because correlation coefficients were relatively similar for both cases (Pearson correlation coefficient of 0.74) I concluded that this sparse data set can still be used for comparison to other studies. In general, I observed that all protein abundance studies using the original GFP Clone Collection clustered together while our small data set was more similar to measurements performed with mass-spectrometry. This was also observed for the earlier measurements by our laboratory using a C-SWAT collection (Supplementary Figure 4.3). This indicates that

previous flow-cytometry data sets have a certain systematic bias originating from the used GFP yeast collection. The *de novo* construction of this CASTLING library allowed to reveal such biases. In addition, expression of several tagged genes were detected which were absent in previous studies and we confirmed those by individual gene tagging [116].

In summary, CASTLING outlines a technology to generate pooled libraries of complex tag genotypes and this section highlighted how Anchor-Seq can be used to validate and quantitatively assess such libraries.

3.3 Mammalian PCR tagging

In contrast to the PCR targeting strategy in *Saccharomyces cerevisiae*, endogenous gene tagging in mammalian cells has been a laborious and inefficient endeavor so far [12]. This quite often prohibits rapid and direct testing of a hypothesis as the required genotypes usually are not easily accessible and require multiple molecular cloning steps. As the application of the concept of self-integrating cassettes (SIC) for pooled gene tagging in yeast turned out to be highly successful it prompted the question if it could also be useful for simplified gene targeting in mammalian cells. As mentioned in the introduction, PCR targeting is used for cloning-free generation of constructs for gene targeting which simplifies experimentation. Conceptually, PCR targeting allows for modularization by separating the information which kind of modification should be introduced and to where it should be directed to. The SIC concept embraces the same ideas which makes it perfectly suitable for PCR targeting in mammalian cells.

Similar to PCR tagging in yeast [25], a streamlined mammalian PCR tagging protocol would include a toolbox with a wide range of tagging constructs with tags of different functionalities for various applications. The constructs would be designed following the same set of rules which allows for the generic combination of gene-specific primer pairs and function-specific tag modules maximizing the use of already available resources. Furthermore, this can also simplify cloning in case somebody wishes to extend the toolbox which fosters the adaptation across the community.

The following chapter is concerned with the establishment of a streamlined PCR targeting strategy for mammalian cells. The strategy was tailored for C-terminal tagging of mammalian genes and was termed ”PCR tagging”. The focus will mainly be on the characterization of possible repair outcomes and artifacts related to this protocol as this was my primary contribution to this project.

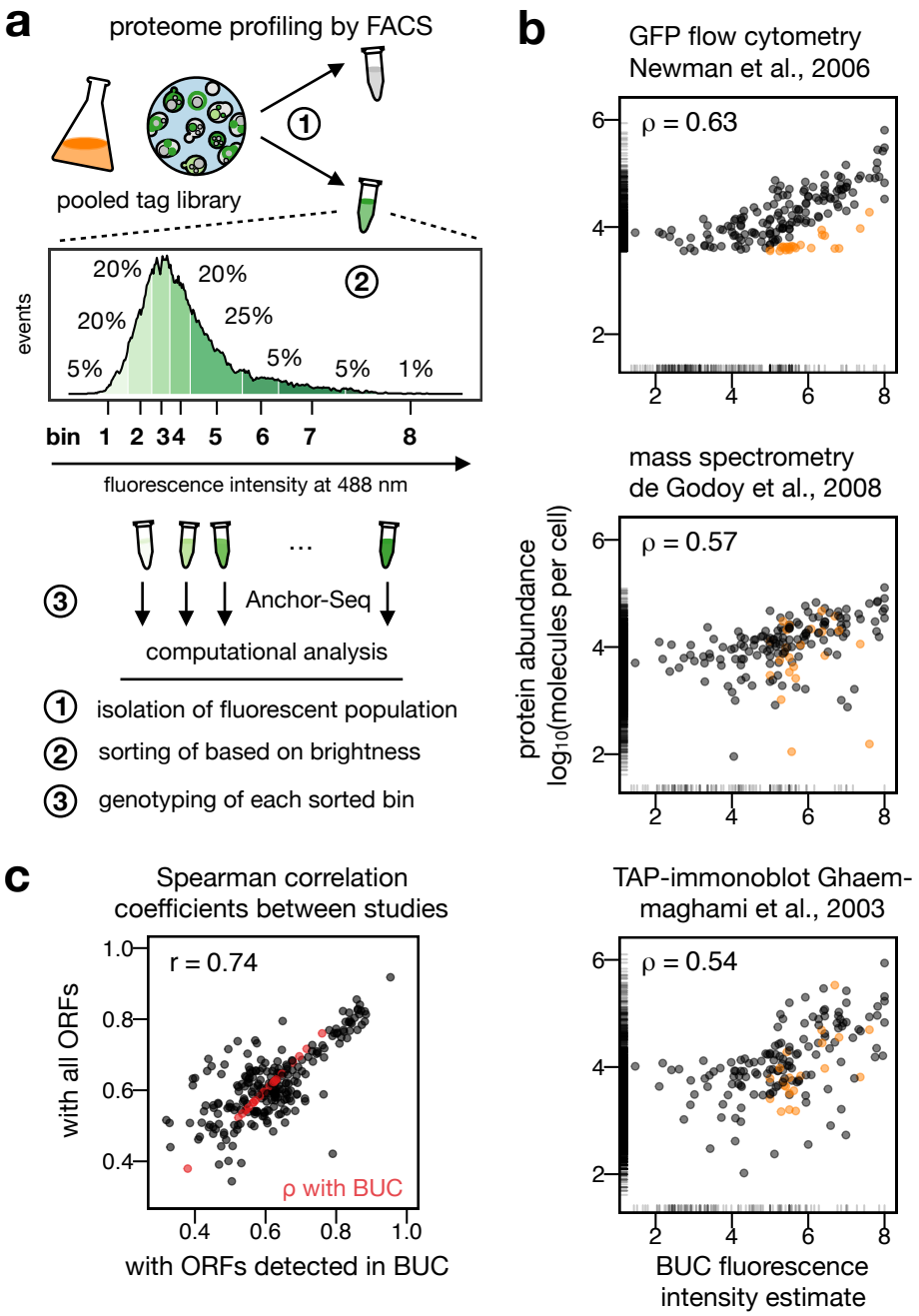


Figure 3.17: Protein abundance estimation using a CASTLING library. A CASTLING library in which genes were C-terminally tagged with mNeonGreen were fractionated by FACS and each fraction was analyzed by Anchor-Seq to infer fluorescence intensity profiles for individual genes. **(a)** The mNeonGreen CASTLING library was first enriched for fluorescent cells using a first round of FACS. The enriched library was then fractionated into bins based on fluorecence intensity using a second round of FACS. Gene abundances in each bin were quantified by Anchor-Seq and the underlying fluorescence intensity profiles were determined. From those tagged gene abundance estimates were retrieved (see Methods). **(b)** Comparison of abundance estimate from this experiment to earlier studies in which various technologies have been used to estimate protein abundances. Some genes were marked as outliers (orange) in comparison to the study by Newman *et al.* (top plot) to show that their placement is not recapitulated with other studies (lower two plots) and therefore seem to be unique to this study. Marginal lines at either axis site indicate abundance estimate for genes detected in only one of the compared data sets. **(c)** Spearman correlation coefficients (ρ) between studies calculated either using all the genes detected in those study (y-axis) or using only the 435 genes detected by Anchor-Seq after the second FACS (x-axis). *The figure was reproduced in modified form from [116] with permission and was jointly created by Benjamin Buchmuller and me.*

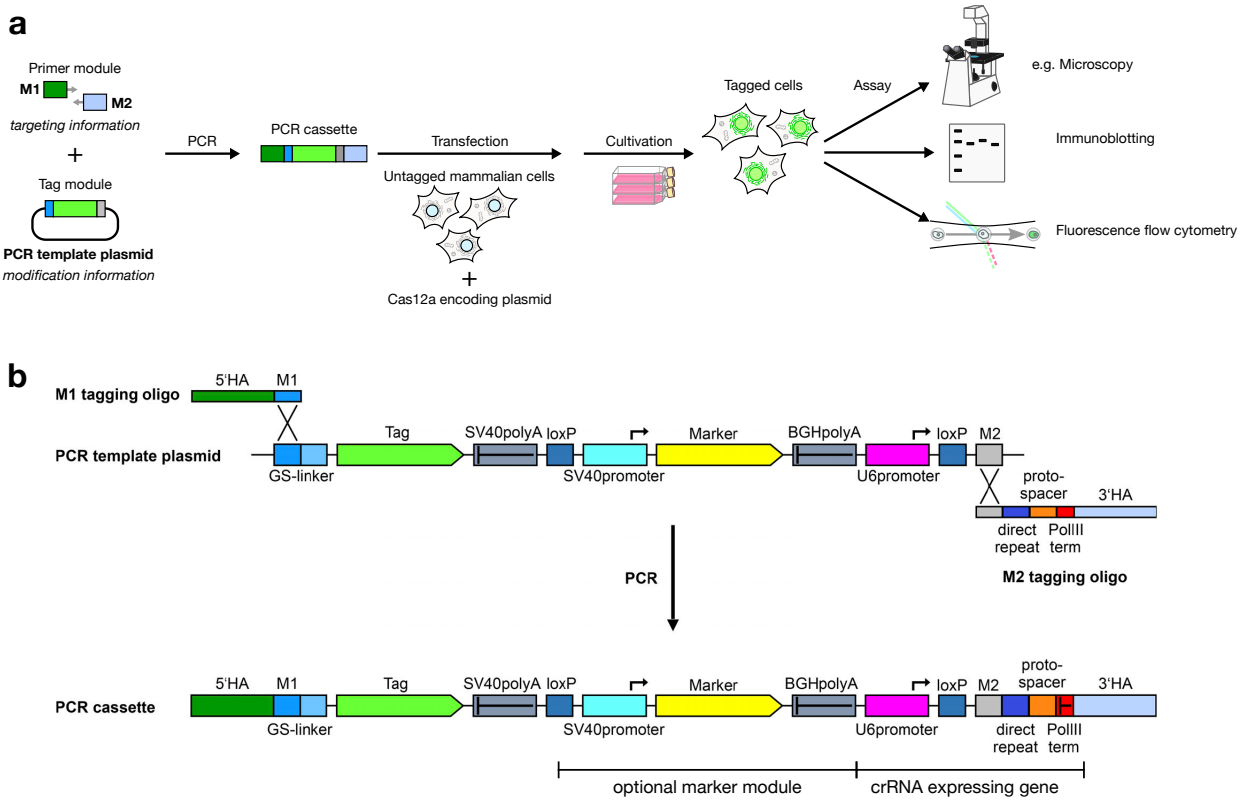


Figure 3.18: General workflow for PCR tagging in mammalian cells using SICs. **(a)** PCR tagging in mammalian cells allows for streamlined experimentation. **(b)** PCR with gene-specific M1 and M2 primers using PCR tagging plasmid as template yields linear DNA framgents (PCR cassettes) which follow the SIC design. This includes a gene for PolIII-driven crRNA expression downstream of the C-terminal tag. A marker for the selection of transformed cells is optionally included in the cassette. Homologous sequence (HA, homology arms) for gene targeting are introduced with the M1 and M2 primers. *Panel b was reproduced in modified form from [117] with permission and was originally created by Matthias Meurer.*

3.3.1 Self-integrating cassettes for efficient tagging of endogenous genes in mammalian cells

The experiments presented in this section were conducted and analyzed by Julia Füller, Matthias Meurer and Michael Knop. They are included here for completeness.

To first test if SICs lead to efficient and robust C-terminal gene tagging of single genes, subcellular localization of mNeonGreen-tagged proteins was used as an easily quantifiable and distinguishable phenotype. The human embryonic kidney cell line 293T (HEK293T) was chosen for the following experiments because it is a commonly used model cell line and its parental cell line 293 exhibits high transfection efficiencies [151, 152]. Sixteen genes were selected based on their high abundance and distinct subcellular localization for targeting with a C-terminal mNeonGreen tag. Respective gene-specific primer pairs were designed: The M1 primer consists of an annealing site for the 5'end of the tagging module and the 5' homology arm (HA). This homologous sequence encompasses 90 nt of the end of the targeted gene including the last coding exon until the stop codon. The M2 primer contains the annealing site of the 3'end of the tagging construct and the 3'HA which encompasses 55 nt of the genomic sequence following the stop codon. In addition, this primer contains a gene-specific sequence encoding for a crRNA targeting the genomic locus around the stop codon of the targeted gene and a short T_6 element for termination of polymerase III (PolIII) transcription. The mNeonGreen tagging module includes the C-terminal mNeonGreen tag with a linker peptide, a heterologous 3'end to terminate expression of the mNeonGreen gene fusion and the U6 PolIII promoter to drive crRNA expression. (Figure 3.18b)

The cassettes which are constructed by PCR reconstitute the SIC design. PCR cassettes for each gene were individually transfected into HEK293T cells together with a helper plasmid for Cas12a expression. The cells were cultured for three days and examined by fluorescence microscopy. The expected localization of mNeonGreen signal was observed for 0.2 % to 13 % of cells across all the inspected genes (Figures 3.19a,b). In addition, control transfections were performed in which co-transfection with the Cas12a plasmid were omitted or the PCR cassettes lacked either homology arms or crRNA sequence (Figure 3.19c). No cells with the expected localization signal were observed in those controls which indicates that Cas12a expression and the individual functionalities of the SIC are required for successful gene tagging. When the homology arms and the crRNA of the PCR cassette targeted different genes very few cells were observed which showed mNeonGreen localization of the gene targeted by the crRNA. Therefore the tagging cassette was integrated at the site of the DSB although the respective homologous sequences for HR were absent indicating that repair was mediated by alternative pathways such as NHEJ. Tagging cassette and crRNA expression construct could also be provided as separate molecules during transfection resulting in similar numbers of tagged cells than when

the crRNA was expressed directly from the tagging cassette in case of the SIC design. This indicated, that a PCR tagging approach using SICs leads to successful and specific endogenous gene modifications in mammalian cells.

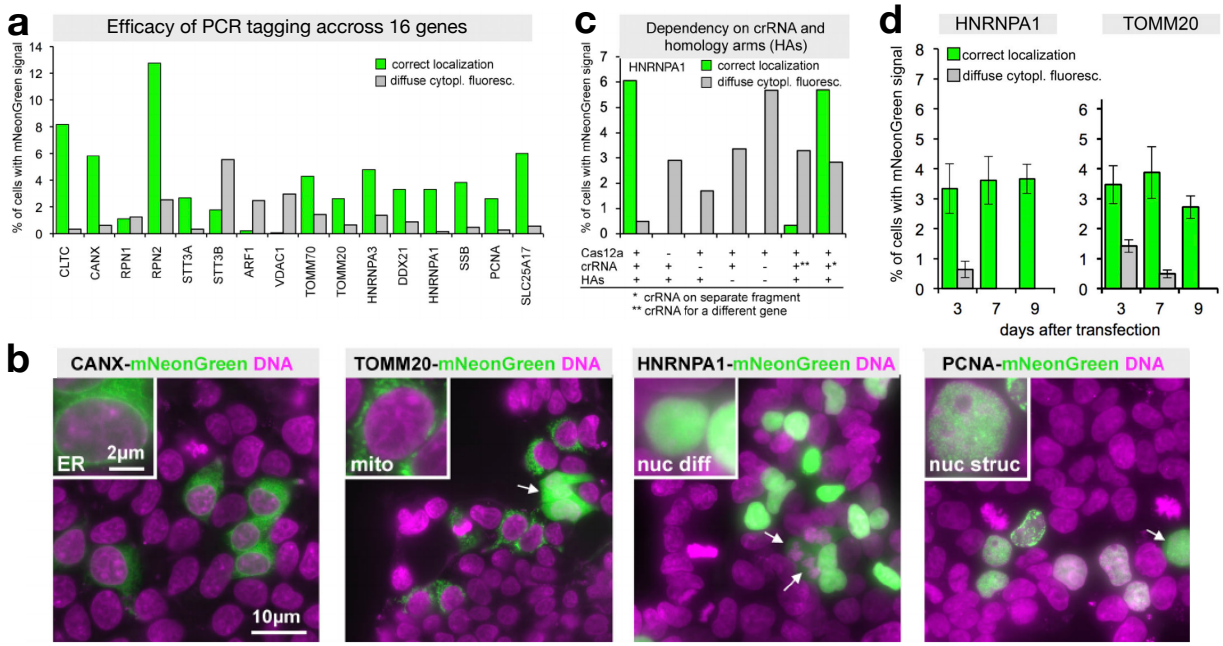


Figure 3.19: Using PCR tagging with SICs for single gene targeting in mammalian cells. PCR cassettes (SICs) for C-terminal gene tagging with mNeonGreen were transfected into HEK293T cells. Live cells were imaged three days after transfection by fluorecence microscopy. Successful C-terminal tagging was scored by mNeonGreen signal while total cells were counted using DNA stained nuclei by HOECHST dye. **(a)** The efficacy of PCR tagging was determined as the fraction of cells exhibiting the expected subcellular phenotype for 16 selected genes. The fraction of cells exhibiting a diffuse cytoplasmic phenotype was also determined. **(b)** Representative micrographs for the four genes *CANX* (ER localization), *TOMM20* (mitochondrial localization), *HNRNPA1* (diffuse nuclear localization) and *PCNA* (structured nuclear localization) are shown. Cells exhibiting diffuse cytoplasmic signal are indicated with arrows. DNA was stained with HOECHST dye. **(c)** Control PCR taggings of HNRNPA1 in which different components of the SIC design were omitted indicating that the method depends on all components to work. * The crRNA was expressed from a separate molecule than the PCR cassette (SIC). ** The crRNA sequence in the SIC targets *CANX* while the homology arms direct to *HNRNPA1* leading to off-target cleavage and integration most likely by NHEJ for a very small fraction of cells (<0.02 %, five cells in entire well). **(d)** PCR tagging of two selected genes in HEK293T cells and quantification of cells with localized and diffuse fluorescent signal over several days of cultivation. Data from three replicates. Error bars represent SD. *The figure was reproduced in modified form from [117] with permission and the original version was created by Julia Füller, Matthias Meurer and Michael Knop.*

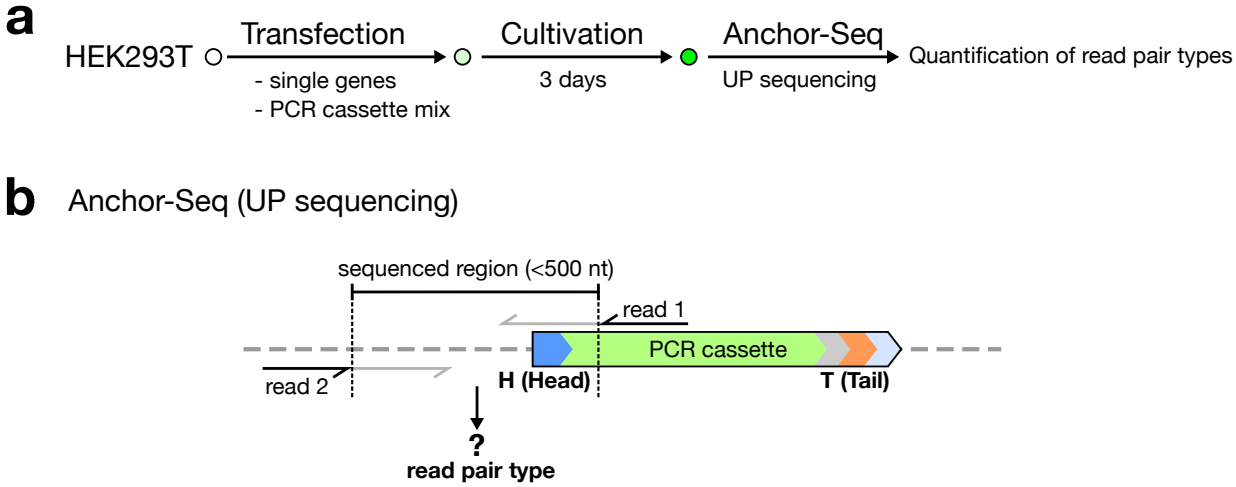
3.3.2 Aberrant tag expression results in cytoplasmic artifact for C-terminal mNeonGreen taggings

During inspection of cells transfected with PCR cassettes for C-terminal tagging with mNeonGreen another phenotype than the correct fluorescence localization was frequently observed across all inspected genes. A considerable number of cells exhibited a diffuse cytoplasmic fluorescent signal with high variability in cell-to-cell intensity (Figure 3.19a, b). This phenotype was also present in the control transfection experiments in which individual PCR tagging components were omitted and which did not yield successful gene

taggings. This indicated that the diffuse cytoplasmic signal might be an artifact from the transfected PCR cassettes themselves (Figure 3.19c) and consequently required further investigations. Counting and classifying cells over the course of several days after transfection revealed differing stability of the correct localization as well as the diffuse cytoplasmic phenotype. While the fraction of cells with correctly localized fluorescence signal remained relatively constant, the fraction of cells exhibiting cytoplasmic fluorescence gradually decreased over the course of nine days (Figure 3.19d). This implied, that the molecular species responsible for this artifact might be genetically unstable and therefore potentially of extrachromosomal origin.

The question arose which genetic structure of the PCR cassette might cause expression of mNeonGreen protein. In order to characterize the DNA sequences upstream of the mNeonGreen tag without any biases Anchor-Seq was performed with genomic DNA extracted from cells three days after they had been transfected with C-terminal mNeonGreen constructs targeting different genes (Figure 3.20a). The resulting dataset allowed to categorize and quantify the nature of molecules which contain the PCR cassette sequence (Figure 3.20b). Sequences indicative of correct integration events were observed although in very low numbers (Figure 3.20c). The remaining majority of read pairs did not cover a sequence beyond the cassette sequence (i.e. beyond the homology arms). These types of molecules arose either from frequent tagmentation events close to the anchor sequence or from many PCR cassette molecules remaining in the culture three days after transfection. In case of the latter prolonged cultivation would dilute PCR cassette molecules further and by this could improve sensitivity for correct integration events. In addition, a considerable number of read pairs indicated end-to-end fusions of PCR cassette sequences. Head-to-tail fusions (i.e. ligation events between 5' and 3'end of the PCR cassettes) were on average six-times more prevalent than head-to-head fusions. In contrast, because the Anchor-Seq reaction was performed in UP direction only, tail-to-tail fusions could not be observed. These findings revealed that the transfected PCR cassettes frequently fused end-to-end which was most likely mediated by NHEJ. In addition, cells were also transfected with a mixture of PCR cassettes targeting different genes. Performing the same type of analysis as before revealed that this time head-to-head and head-to-tail fusion occurred between ends of the same gene (homo fusions) as well as between ends of different genes (hetero fusions). This further supported end-to-end ligation of molecules by DNA damage repair pathways such as NHEJ and indicated that at least a fraction of ligation events is intermolecular. The occurrences of head-to-head homo fusions and head-to-head and head-to-tail hetero fusions were approximately equal while head-to-tail homo fusions were approximately six-times more prevalent than the other three fusion events. The finding from the transfection experiments with mixed PCR cassettes concurred with the results from the transfection experiments targeting just a single gene. The observed bias

for head-to-tail homo fusions could indicate that most ligation events are intramolecular. In summary, these results show that the PCR cassette are frequently ligated together preferentially in a head-to-tail conformation.



c

observed read pair types (%)	*RPN2* +	*STT3A* +	*TOMM70* +	*TOMM20* +	PCR casette mix +	PCR casette mix -	Cas12a
correct integration	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	
PCR cassette	85.0	81.3	83.7	89.3	90.2	89.4	
template plasmid	0.5	0.3	0.1	<0.1	<0.1	<0.1	
HH homo fusion	2.4	3.0	2.1	1.2	0.8	1.3	
HH hetero fusion	-	-	-	-	1.0	1.4	
HT (homo)	12.1	15.4	14.1	9.5	6.8	6.5	
HT (hetero)	-	-	-	-	1.2	1.4	

Figure 3.20: Characterization of diffuse cytoplasmic signal by stability and sequencing. (a) Schematic outline of the Anchor-Seq experiment to quantify PCR tagging outcomes. **(b)** Anchor-Seq was performed so that the sequence upstream of the PCR cassette could be determined (i.e. UP sequencing). This allowed to classify and quantify read pair types. **(c)** Tabulation of the read pair type quantification (in percent total read pairs) for transfection of four single PCR cassettes and a mixture of PCR cassettes. Read pairs could either indicate correct integration, PCR cassette, template plasmid or head-to-head or head-to-tail fusions. For transfections of PCR cassette mixtures these two types were further subdivided into homo and hetero fusions. *The figure was reproduced in modified form from [117] with permission.*

The head-to-tail conformation brings the PolIII promoter used for crRNA expression into close proximity with the 5'-end of the mNeonGreen coding sequence. I proposed the hypothesis that this provides a potential explanation for the diffuse cytoplasmic phenotype. It would require PolII-driven transcription from the PolIII promoter and use of some early ATG start codon in the tag sequence for translation. The result would be aberrant tag expression in which the tag would not be fused to an upstream coding sequence of a gene and instead be expressed on its own (Figure 3.21a). In the case of the PCR targeting with a C-terminal mNeonGreen cassette this would mean the expression of only mNeonGreen protein and not a fusion protein which would then result in the observed diffuse cytoplasmic fluorescence signal. Indeed, the transfection of a modified PCR cassette in

which the ATG start codon was removed resulted in reduction in cells with cytoplasmic fluorescence phenotype (Figure 3.21b). In an additional experiment the U6 PolIII promoter together with a Kozak sequence was cloned in front of the mNeonGreen tag which resulted in constitutive diffuse cytoplasmic GFP signal. Furthermore, the U6 promoter was substituted with either the 7SK or the H1 PolIII promoters which exhibit lower or higher PolII activity respectively than the U6 PolII promoter [153, 154]. The resulting 7SK promoter construct exhibited a decrease in cellular fluorescence intensity while the H1 promoter construct exhibited an increase in cellular fluorescence intensity with respect to the U6 promoter construct (Matthias Meurer, personal communication).

NHEJ requires moderate end resection for efficient ligation [40]. Earlier studies in which PCR products where used for transfection have already shown that modified primers can attenuate NHEJ acting on these molecules [155]. As the end-to-end ligation of PCR cassettes is most likely mediated by NHEJ it would be expected that use of primer modifications should also reduce the occurrence of cells with a cytoplasmic phenotype. Indeed, it was observed that PCR cassettes generated with primers modified with phosphorothioate and biotin for several different genes resulted in fewer cells with cytoplasmic phenotype than PCR cassettes created with unmodified primers. Although the effect was not very strong the trend was always the same (Figure 3.21c). These results further supported the hypothesis, that PCR cassettes are frequently ligated together head-to-tail by NHEJ.

Further direct evidence for ligation of PCR cassette ends was acquired by performing a timecourse experiment and subsequent PCR amplifications to specifically probe end-to-end fusions (Figure 3.22). Genomic DNA was extracted from cells which were transfected with a PCR cassette for C-terminal tagging of the gene *HNRNPA1* with mNeonGreen. The cells were harvested 6, 18 and 30 days after transfection. The DNA was then subjected to PCR reactions with a primer pair which covered the potential head-to-tail junction of the PCR cassette. Two additional PCR reactions were performed in comparison in which only one of each primer was used to probe the presence of molecules with head-to-head and tail-to-tail ligations. No specific head-to-head and head-to-tail amplicons could be observed despite the presence of some unspecific amplification products. On the contrary the reaction which included both primers clearly showed the presence of head-to-tail specific amplicons which were still present 30 days after transfection. This indicates the continued persistence of molecules with head-to-tail fusions in the cells. Additional unspecific amplicons which are mostly present in the later points in time hint to competitive amplifications which usually arises when the primer specific template is very diluted. An amplicon with a size comparable to the head-to-tail amplicon in the transfection samples was observed in a control PCR reaction which was performed with the material used for transfection. Nevertheless, it was considered an unspecific by-product because it appeared at much lower amount than an amplification product of residual PCR

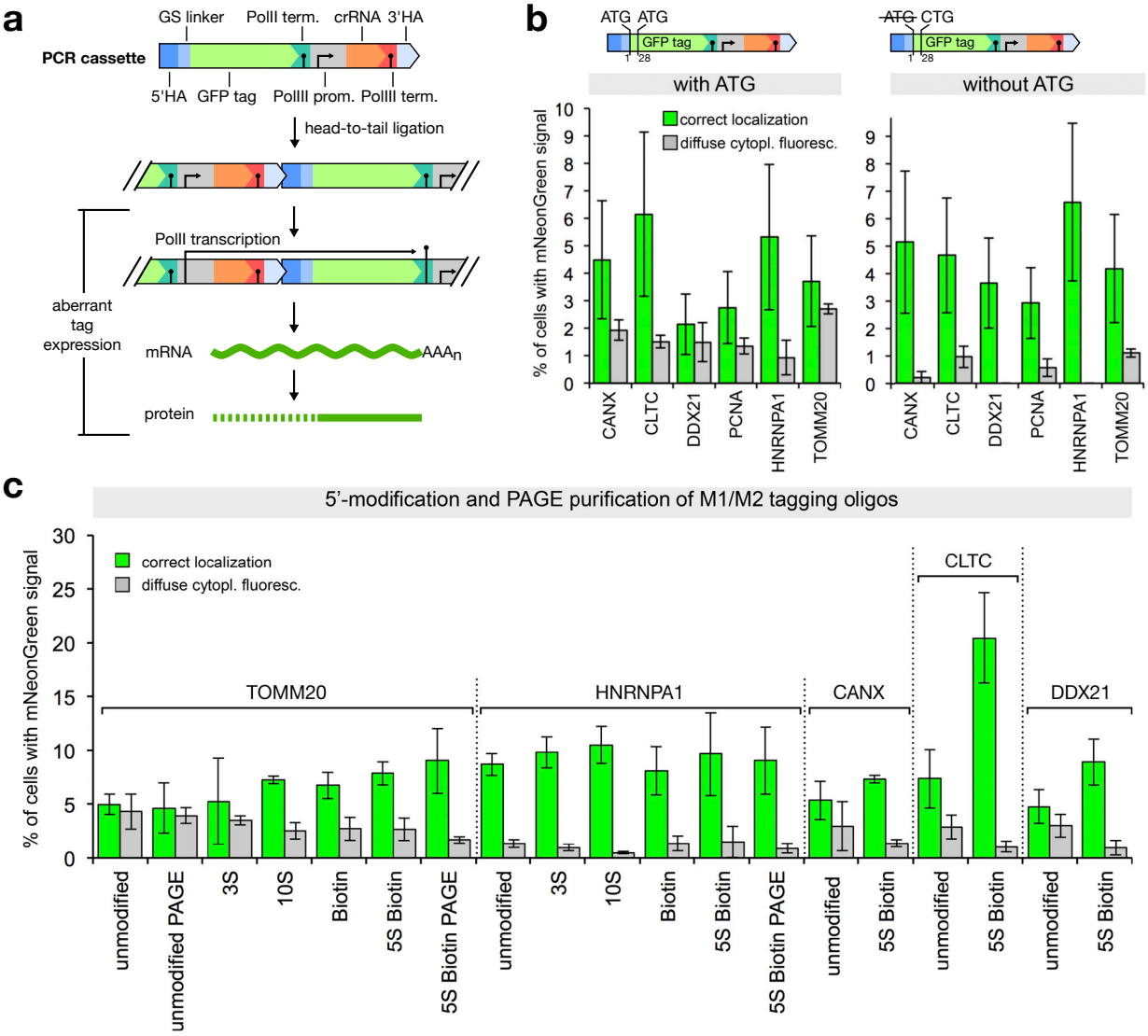


Figure 3.21: Diffuse cytoplasmic signal could be explained by aberrant tag expression. **(a)** Model how head-to-tail ligation of PCR cassettes could lead to aberrant tag expression. **(b)** Removal of initial start codon from mNeonGreen tag reduces the number of cells with diffuse cytosolic signal. For some genes this artifact is not completely eliminated suggesting that abberant tag expression might be driven by cryptic start codons e.g. in the crRNA or homology arm sequences. Tagging efficacies were estimated three days after transfections of HEK293T cells. Data from three replicates; error bars represent SD. **(c)** 5'-end modification of M1/M2 primers used for PCR cassette generation aim to reduce NHEJ and therefore head-to-tail ligation of PCR cassettes. In consequence reduced numbers of cells with diffuse cytoplasmic signal are observed across several genes and types of modifications. Primers were either PAGE- or cartridge-purified if not noted otherwise. Modifications included three (3S), five (5S) or ten (10S) phosphorothioate bonds or biotin. The experiment was analysed as in panel b. *The figure was reproduced in modified form from [117] with permission and the original version was created by Julia Füller, Matthias Meurer and Michael Knop.*

cassette plasmid which due to its larger size is actually unfavored in the PCR reaction. In addition, this by-product appears at a very different size when a tagging PCR for a different gene was used as template (data not shown). In summary, these results indicate that head-to-tail ligation molecules are the dominating ligation species in the cells and that they are diluted over time. Nevertheless, they remain present in the cells for up to 30 days which could indicate that a fraction of ligated molecules is stably integrated into the genome.

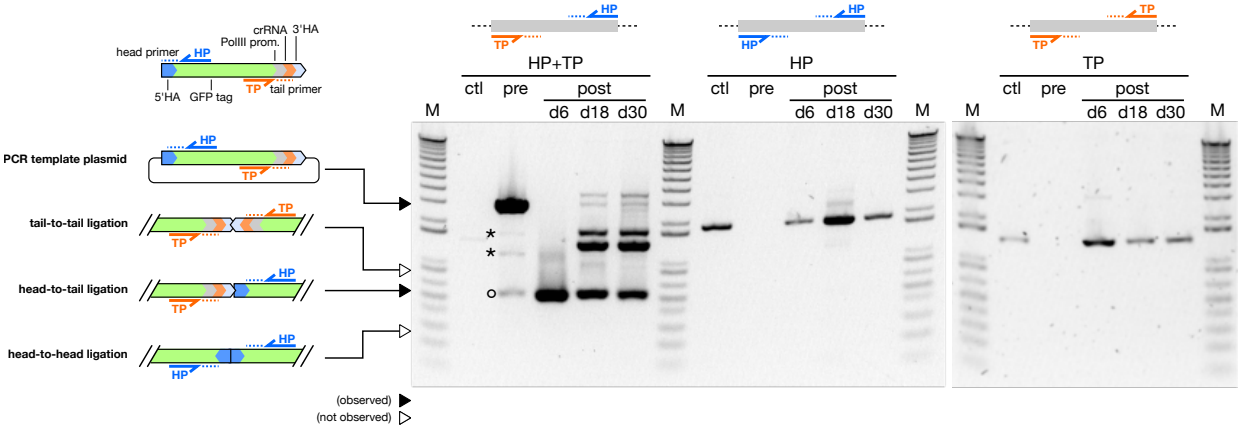


Figure 3.22: Specific test for detecting end-to-end ligations of PCR cassettes. HEK293T cells were transfected with mNeonGreen PCR cassettes for the *HNRNPA1* gene. Genomic DNA was extracted 6 (d6), 18 (d18) and 30 (d30) days after transfection (post) and PCR reactions were either performed with tail specific primer (TP) or head specifc primer (HP) or both (HP+TP) to specifically detect the ligation products indicated on the left. As control, genomic DNA from HEK293T cells (ctl) and the PCR cassette material before transfection (pre) was also used as PCR template. The PCR reaction with both primers (HP+TP) of the PCR cassette (pre) exhibits several additional bands which are likely artifacts of this particular PCR amplification and of different origin as the amplificates in the genomic DNA as they are not present in the single primer PCR reactions (*) or have a very different length (°) in a transfection for the gene *CANX* (data not shown). Both primers lead to unspecific amplification when genomic DNA was present in the reaction. As marker (M) the 1 Kb Plus DNA Ladder (ThermoFisher Scientific) was used. *The experiment was jointly performed by Daniel Kirrmaier and me.*

3.3.3 Sequence fidelity at the target locus

Critical aspects of gene editing experiments are how well the intended modification was achieved and if any unintended modifications occurred either at the target site or at any off-target sites at other sites of the genome. At the on-target site several repair outcomes are possible. Either the PCR cassette is used as template for HR and the desired repair outcome is achieved. Alternatively, the DSB at the endogenous locus cassette might be re-ligated by the means of NHEJ or repaired by HR using for example the sister chromatid. Finally, the PCR cassette could also be ligated into the DSB between the free DNA ends in either orientation leading to a duplication of the homologous sequences as repair outcome (Figure 3.23a).

The primers used to construct PCR cassettes are relatively long so that the likelihood that they contain errors is very high. Consequently, it was desirable to quantify the sequence variation within the junction between tag and coding sequence to evaluate if

HR selects against such errors. In addition, sequence errors are also likely to occur if the targeted locus is not repaired using the PCR cassette as template and instead is rejoined by NHEJ or other more erroneous DSB repair pathways. In order to determine the fidelity of the tag junction I used genomic DNA harvested from HEK293T transfected with a mNeonGreen PCR cassette targeting either of the genes *HNRNPA1*, *CANX* or *CLTC*. To reduce the interference of concatenated PCR cassettes, modified primers were used for the preparation of the PCR cassettes and the cells were harvested after a prolonged cultivation of 18 days after transfection. The target sequence for these PCR taggings resulted in Cas12a cutting either before (*CANX*), at (*CLTC*) or after (*HNRNPA1*) the stop codon (Figure 3.23b). I performed site-specific PCR to amplify a short sequence of approximately 230 nt spanning the junction between coding and tag sequence of the tagged allele. DNA electrophoresis of the tag PCR indicated the presence of a second amplicon species (Figure 3.23c). The approximate length of the shorter species coincided well with the estimated length of the desired repair outcome resulting from HR, whereas the length of the longer species indicated duplication of the homology arm as it would be the case if NHEJ was used by the cell instead for DSB repair. Although a PCR analysis like this is not strictly quantitative it nevertheless suggests that a considerable fraction of repair outcomes were facilitated by HR.

To focus the analysis on the HR-mediated PCR cassette integration product the shorter amplicon of the tag PCR was purified from the gel and used for NGS. Between 84 % (*CANX*) and 93 % (*CLTC*) of reads of the tag amplicons matched the expected sequence which meant that a very high fraction of editing events resulted in the desired tag junction sequence (Figure 3.23d). When the positions of the different sequence variants (i.e. deletions, insertions and substitutions) were inspected it was apparent that variants mainly occurred at positions close to the tag coding sequence within the genome-cassette-junction. This indicates that the errors were most likely already present in the PCR cassette due to primer errors as HR is a highly accurate DSB repair process. HR seems to miss more errors in the homologous sequence of the template the closer the error is to the DSB (i.e. the shorter the homologous sequence is) as it has been noticed before in the CASTLING study. A notable enrichment of deletions within the target sequence was observed for the PCR tagging of *CANX*. For this gene the PAM sequence of the crRNA target site is reconstituted after HR with the PCR cassette so that the tag allele is still a target for the CRISPR-Cas12a complex. The potential introduction of deletions by this targeting will likely be detrimental for the expression of the tagged gene. Consequently, this finding helped to improve the rules for M1/M2 primer design by penalizing target sequences with similar position and orientation with respect to the stop codon as the one used for *CANX* (Figure 3.23e).

To this end the fidelity of the tag allele was inspected but it was also interesting to

investigate the occurrence of mutations in the wild-type allele. Such mutations could be the result of erroneous DSB repair without using the repair template leading to faulty reconstitution of the wild-type allele. I therefore specifically amplified and sequenced the wild-type locus to inspect possible mutations introduced by PCR tagging. Of all reads 7 to 12 % exhibited alterations of the wild-type sequence (Figure 3.23e). These alterations encompassed primarily deletions which were similar to the case of the *CANX* tag allele enriched within the region of the cut side. Such mutations are likely a result of mutagenic NHEJ events. Since these mutations occurred around the stop codon they probably influence the expression of the targeted gene which will be discussed in more detail below.

3.3.4 Application of Tn5-Anchor-Seq for unbiased detection of on- and off-targets

Apart from the fidelity of the tag junction it needed to be determined if the PCR cassette had integrated elsewhere in the genome. Furthermore, it was interesting to investigate if such off-target integrations occurred systematically at certain genomic hot-spots and in dependence of Cas12a expression. Therefore repeated transfections of PCR cassettes for the three genes *HNRNPA1*, *CANX* and *CLTC* were analysed by Tn5-Anchor-Seq to determine potential off-targets without prior knowledge of the integration sites (Figure 3.24). On-target integrations were only observed in those replicates in which the Cas12a helper plasmid had been co-transfected. Off-target integration events were observed across the whole genome irrespective whether the Cas12a helper plasmid had been co-transfected or not. None of these individual integration events were observed in two or more replicates. Both observations lead to the conclusion that off-target integrations occur randomly and independently of Cas12a expression and therefore also independently of the PCR tagging strategy. Overall, off-target integrations seem to be a frequent by-product of transfections of linear DNA at least in the HEK293T cell line used in these experiments.

In summary, SICs allow for a streamlined and efficient PCR tagging strategy applicable in mammalian cells. During transfection the linear PCR tagging cassettes preferentially undergo intramolecular ligation leading to a diffuse cytoplasmic signal in case of the C-terminal mNeonGreen tag. The various tagging outcomes were characterized revealing specific integration of the PCR cassettes while off-target integrations can occur independent of the presence of the CRISPR-Cas12a endonuclease.

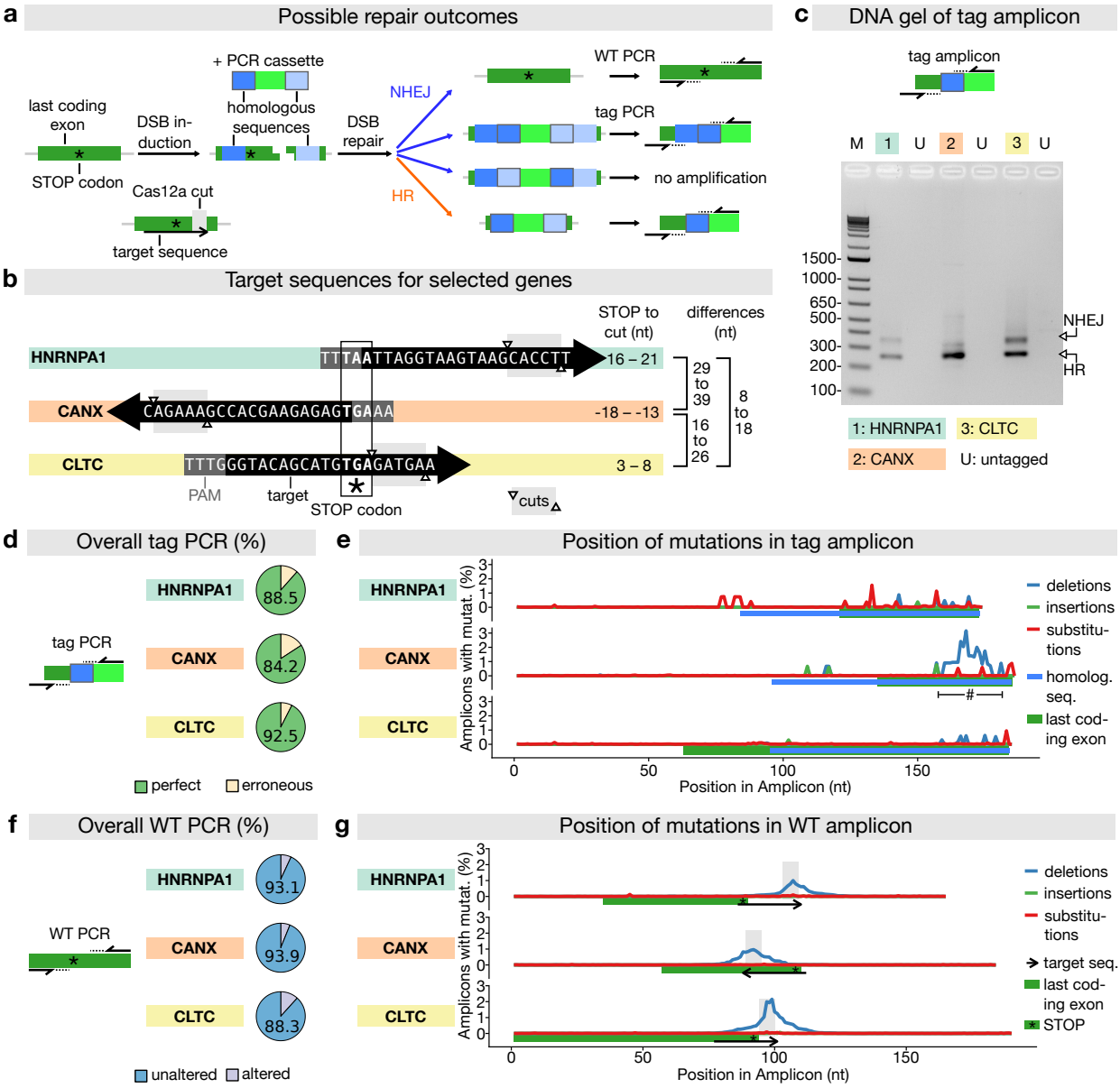


Figure 3.23: Amplicon sequencing to determine on-target fidelity. HEK293T cells were transfected with mNeonGreen tagging cassettes targeting either *HNRNPA1*, *CANX* or *CLTC*. Genomic DNA was prepared 18 days after transfection and on-target amplicons either specific for the wild-type (WT) or the tag allele were amplified by PCR and analyzed by NGS (approx. 10,000 reads per amplicon and gene). **(a)** Schematic of the expected repair outcomes and how they are amplified by PCR. DSB are either repaired by HR or NHEJ. In NHEJ the PCR cassette might be ligated into the targeted locus leading to duplication of the homologous sequences. **(b)** Three genes were targeted which exhibited different target postions. Distances between stop codon and DSB sites are indicated. **(c)** DNA electrophoresis of the tag amplicons for the three targeted genes. For each, two different amplicons could be observed originating either from HR- or NHEJ-mediated repair of the DSB (see also panel a). HR bands were cut, gel purified and used for NGS. Differences in length of NHEJ and HR amplicon correspond to the distance differences of the three genes as indicated in panel b. As marker (M) 1 Kb Plus DNA Ladder (ThermoFisher Scientific) was used. PCR reactions with genomic DNA from untransfected HEK293T cells served as untagged control (U). **(d)** After NGS the fraction of reads perfectly matching the expected tag amplicon sequence was determined. **(e)** Mutations in erroneous reads were further analyzed by type and occurrence within the amplicon. The range of the last coding exon and the homologous sequence of the PCR cassette are also indicated. Note that for *CANX* a high rate of deletions was observed around the cut site of the Cas12a complex (#). **(f)** Same analysis as in panel d for the wild-type amplicon. **(g)** Same analysis as in panel e for the wild-type amplicon. The range of the last coding exon and the position of the of the target sequence is indicated. The position of the Cas12a cuts is indicated by a gray box. *The figure was reproduced in modified form from [117] with permission.*

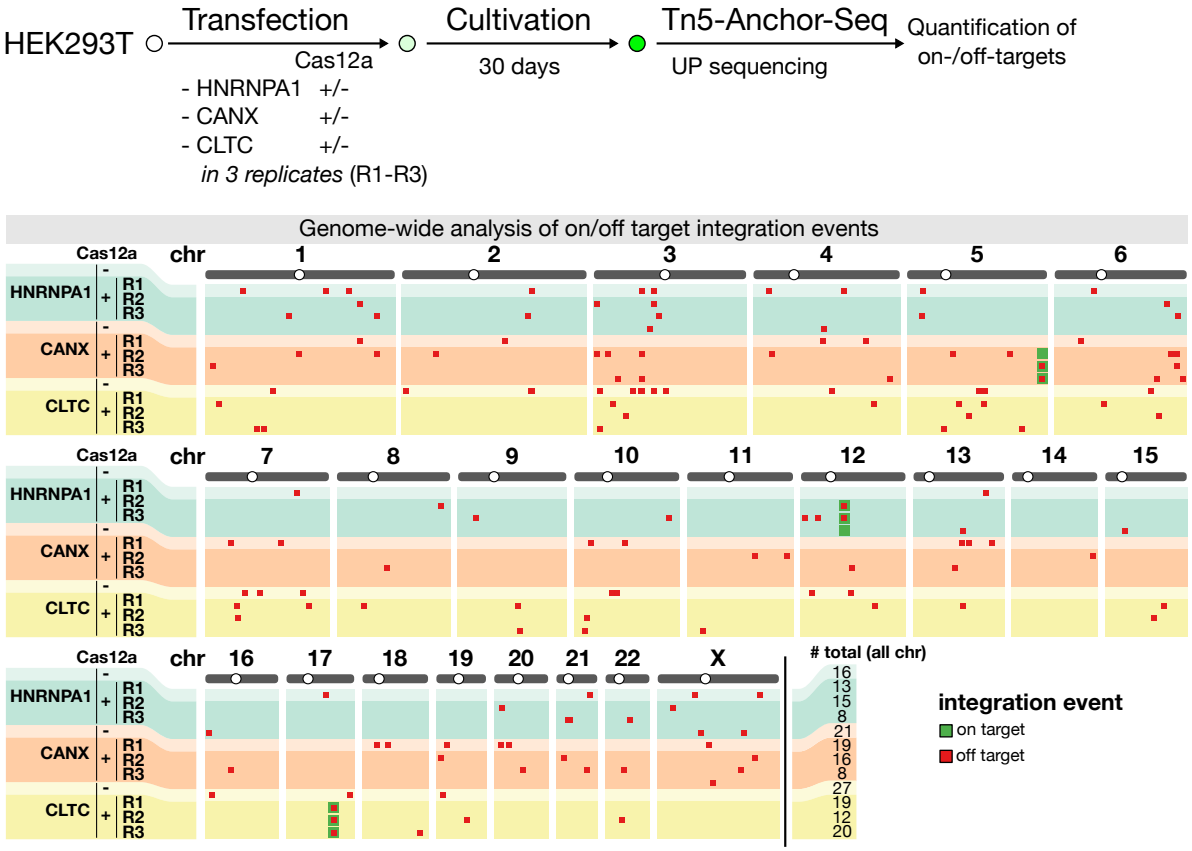


Figure 3.24: On- and off-target quantification by Tn5-Anchor-Seq. HEK293T cells were transfected with PCR cassettes targeting one of three different genes in triplicates with or without Cas12a helper plasmid. Genomic DNA was extracted 30 days after transfection and used for unbiased genome walking sequencing by Tn-Anchor-Seq. The genomic position of the resulting integration events for each of the replicates is indicated. On-target integrations (green) were detected in all replicates when Cas12a was also expressed while off-target integrations (red) were detected across the genome. Some integration events were observed very close to the desired target site but did not yield successful taggings and were therefore scored as off-target integrations. The total number of events detected in a replicate is indicated at the right of chromosome X. *The figure was reproduced in modified form from [117] with permission.*

3.4 RT-LAMP for SARS-CoV-2 detection

3.4.1 Characterization of a colorimetric RT-LAMP assay for diagnostic SARS-CoV-2 detection

As a promising alternative to the RT-qPCR assay, RT-LAMP was quickly adopted for SARS-CoV-2 detection by several groups which proposed LAMP respective primer sets. Even though the performance and validity of most of these primers proved to be excellent, validation had been performed on artificial samples with *in vitro* transcriped (IVT) RNA as synthetic SARS-CoV-2 genomes [156, 157, 130]. A systematic comparison in a diagnostic setting with real-world patient samples was therefore lacking. In an effort to fill this gap several RT-LAMP primer sets were evaluated in March 2020 within the Knop group using IVT RNA. After the N gene primer set published by Zhang *et al.* [130] performed best (Figure 3.25a) the diagnostic performance of this primer set was further assessed with surplus samples from the diagnostic lab of the virology department in Heidelberg. RT-LAMP samples were quantified by the difference in absorbance at wavelength 434 and 560 nm, i.e. the absorbance maxima of phenol red at higher and lower pH respectively (Figure 3.25b). Finally, isolated RNA from 768 patient samples was quantified for SARS-CoV-2 by RT-qPCR and further assessed using the RT-LAMP assay (Figure 3.26 and Figure 3.27a). For samples with a RT-qPCR result of CT<30 very high specificity (i.e. negative RT-qPCR samples which also scored negative in the RT-LAMP assay) of 99.7 % (Wilson's 95 % confidence interval: 98.9 to 99.9 %) was observed in the RT-LAMP assay. The sensitivity (i.e. positive RT-qPCR samples also scoring positive in the RT-LAMP assay) was with 97.5 % (Wilson's 95 % confidence interval: 91.4 to 99.3 %) similarly high (Figure 3.27b). The set of samples with a RT-qPCR result of CT<30 comprised high to intermediate viral titers as the RT-qPCR assay used was calibrated for approximately 1000 viral template molecules for CT=30 according to the manufacturers quality certification. For samples with CT values higher than 30 (i.e. lower viral titers) RT-LAMP reactions exhibited false negative results most of the time. The RT-LAMP reaction were performed with ten times less template compared to RT-qPCR reactions. This indicates that the limit of detection of the RT-LAMP assay is approximately 100 molecules. This observation was also in concordance with earlier results when the RT-LAMP assay was established with IVT RNA as artificial template [156, 157, 130]. Nevertheless, several samples turned yellow upon prolonged incubation of more than 35 minutes at 65 °C. Since this occurred across positive and negative RT-qPCR samples it suggested spurious amplification. An alternative cause could have been that RT-qPCR positive samples with low viral titers resulted in delayed amplification of minute amounts of viral RNA. It was therefore important to investigate the cause of these delayed positive RT-LAMP reactions.

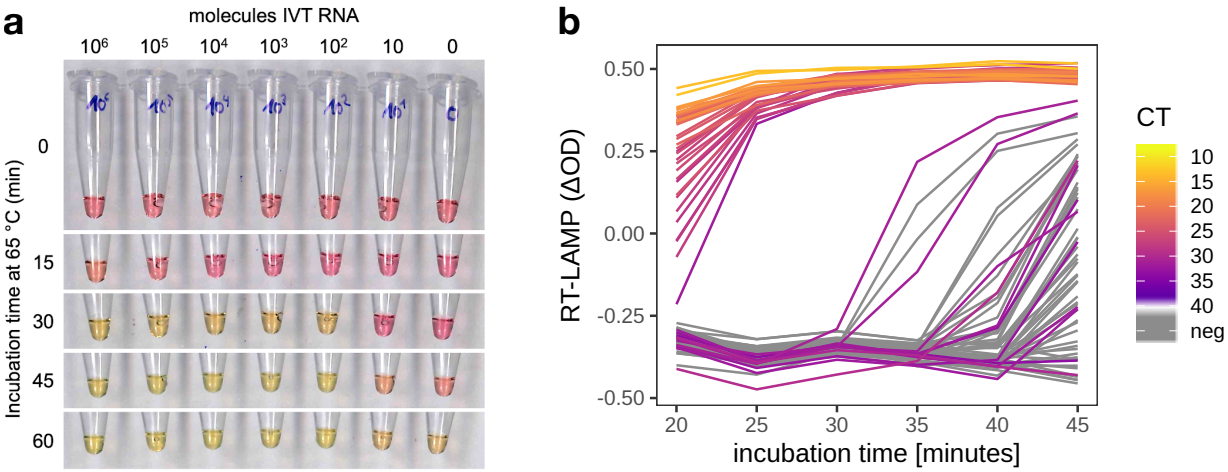


Figure 3.25: A colorimetric RT-LAMP assay for SARS-CoV-2 detection. **(a)** The indicated number of SARS-CoV-2 *N* gene molecules generated by IVT were added to individual colorimetric RT-LAMP reactions. The reactions were incubated at 65 °C, moved briefly to ice at different points in time and imaged using the color scanner function of an office copy machine. A color change from red to yellow indicates production of DNA in the RT-LAMP reaction which until 30 minutes is specific for *N* gene positive samples with at least ~100 molecules IVT RNA. **(b)** Spectrophotometric absorbance measurements allow quantification of the red-to-yellow color change of the colorimetric RT-LAMP assay. The extracted RNA from 95 clinical pharyngeal swab specimens was quantified using the RT-LAMP and the RT-qPCR assay. Color change of the RT-LAMP assay is quantified as difference in absorbance (ΔOD values) at the two maxima of phenol red which is used as pH indicator in this assay: $\Delta OD = OD_{434\,nm} - OD_{560\,nm}$. Positive samples yield a ΔOD value of between 0.3 and 0.4. CT values from RT-qPCR were acquired using the E-Sarbeco primer set [103]. *The figure was reproduced in modified form from [115] with permission. The underlaying data was accquired by Matthias Meurer.*

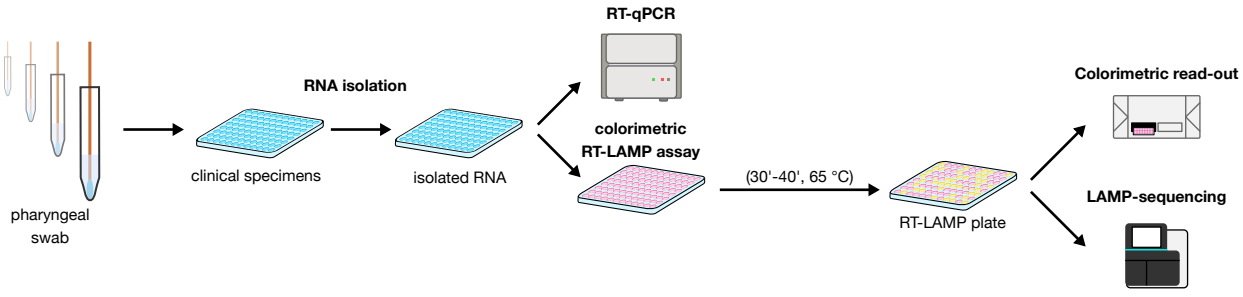


Figure 3.26: Overall workflow of the LAMP-sequencing study. RNA was isolated from pharyngeal swab specimens and used as input for RT-qPCR and colorimetric RT-LAMP assays respectively. The RT-LAMP reactions were analyzed either using a spectrometric plate reader for colorimetric read-out or were used as input for LAMP-sequencing.

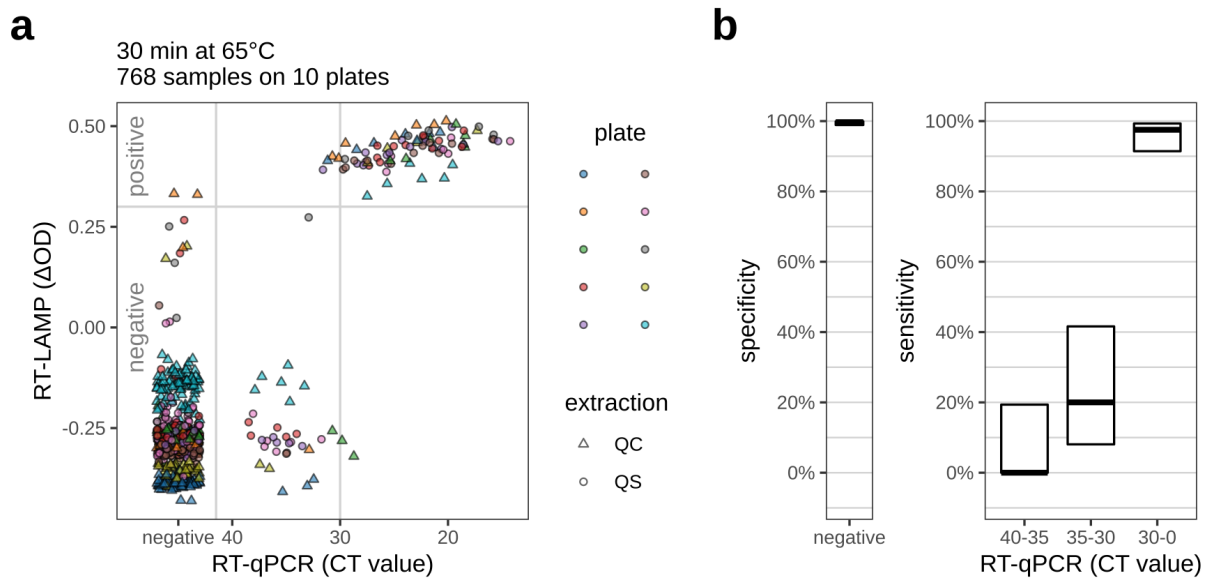


Figure 3.27: Sensitvity and specificity of the RT-LAMP assay. **(a)** Comparison of RT-LAMP to RT-qPCR assay results for 768 clinical samples distributed across 10 96-well plates (different point colors). Over the course of the study RNA was extracted using two different extraction methods (different point shapes; QC: column-based QiaCube method; QS: bead-based QiaSymphony method). RT-LAMP results were recorded after 30 minutes of incubation at 65 °C. **(b)** Specificity (left) and sensitivity (right) of the RT-LAMP assay based on the data presented in panel a. The sensitivity is stratified by RT-qPCR derived CT value. Horizontal lines represent either specificity or sensitivity estimate respectively while the Wilson's binominal confidence interval is given by the boxes. *The figure was reproduced in from [115] with permission. The data was jointly analyzed by Simon Anders and me.*

3.4.2 Development of the LAMP-sequencing protocol

In order to more thoroughly investigate the nature of the samples which were positive in the RT-LAMP assay after prolonged incubation I devised a sequencing strategy based on Tn5-Anchor-Seq which provides several advantages for this purpose (Figure 3.28a). First, library preparation is relatively unbiased in that sample amplification is invariant to the sequences apart from the anchor sequence. This allows to "focus" sequencing on the LAMP reaction product but at the same time allows for amplification irrespective of the remaining molecule sequence which is important because the nature of the LAMP reaction produces complex concatemeric amplification products [108]. Second, adapter sequences for library preparation are introduced to the LAMP molecules by the means of tagmentation which reduces the number of steps and time required per sample. Finally, multiplexing of samples is performed right from the start during tagmentation which furthermore results in fewer processing steps per sample. A plate with 96 individual barcoded tagmentation adapters (P5 site) was used to assemble 96 different Tn5 transposon complexes. These were used to tagment the 96 samples of one RT-LAMP assay plate individually. Consequently, the samples of one assay plate could be combined into one sample reducing complexity of the following steps. Excess transposon adapters were removed by bead-based size selection and a PCR was performed with LAMP reaction- and

transposon adapter-specific oligonucleotide primers respectively. The Illumina adapter on the side of the LAMP reaction-specific primer (P7 site) also contains a barcode which then codes for the plate identity of the sample. Because sample complexity was expected to be reduced compared to the normal application of Anchor-Seq for genomic DNA, the biotin-streptavidin-capture step was not considered necessary which further simplified sample processing. The second PCR was performed with an anchor-specific bridging primer, plate-specifc index primers and the Tn5 adapter-specific primers for target enrichment. After the second PCR another gel-based size selection was included to ensure that mostly longer molecules are sequenced. This ensures that sequencing reads are long enough to span the full region targeted by the RT-LAMP reaction. The LAMP-specific primer was designed to overlap with the binding site of one of the primers with which the RT-LAMP reaction was performed. The region of the virus which is amplified during the RT-LAMP assay contains short sequence stretches which are not covered by any LAMP primer (Figure 3.28b). These virus-specific sequences could only occur in a read sequence if they already have been present in the template, i.e. they were indicative of viral genome in the sample. After sequencing and sample assignment sequencing reads were therefore classified into virus-matching and -unmatching reads. The tagmentation adapter also contained UMIs which were used to collapse PCR duplicates (Figure 3.28c).

I observed that for most positive samples the UMI counts of virus-containing sequences were at least ten-fold higher than for negative samples which easily allows to classify virus-positive and -negative samples (Figure 3.29a and b). For a few samples only few reads were observed although some of these samples were clearly positive RT-LAMP reactions as judged by the banding pattern after agarose gel electrophoresis (Figure 3.29 and Supplementary Figure 4.4). As the majority of negative samples exhibited considerable total UMI counts, low UMI counts likely originated from sample processing problems during LAMP-sequencing library preparation, e.g. not enough material was transferred from the LAMP reactions into the tagmentation reactions. Overall, 14 out of 768 samples (~1.8 %) exhibited this problem and they were not considered for further analysis. This allowed to make a call for 754 of 768 (98.2 %) samples (Figure 3.29b), which confirmed most of the RT-LAMP results for CT<30 samples. Two RT-LAMP positive samples which scored negative in the RT-qPCR were also clearly negative in the LAMP-sequencing assay. The sequencing therefore confirmed the RT-qPCR result for these samples (Figure 3.29b and c). All of the samples which turned yellow upon prolonged incubation of ~40 minutes scored negative in the LAMP-sequencing experiment which confirmed that these were in fact all spurious amplifications independent of the presence of virus titers (Figure 3.29d).

In order to further confirm specificity of the reaction, LAMP-sequencing reads were aligned to the SARS-CoV-2 reference sequence (Figure 3.28c). Of all reads 80.6 % unambiguously mapped to the short segment of the N gene which is targeted by the used RT-LAMP primer

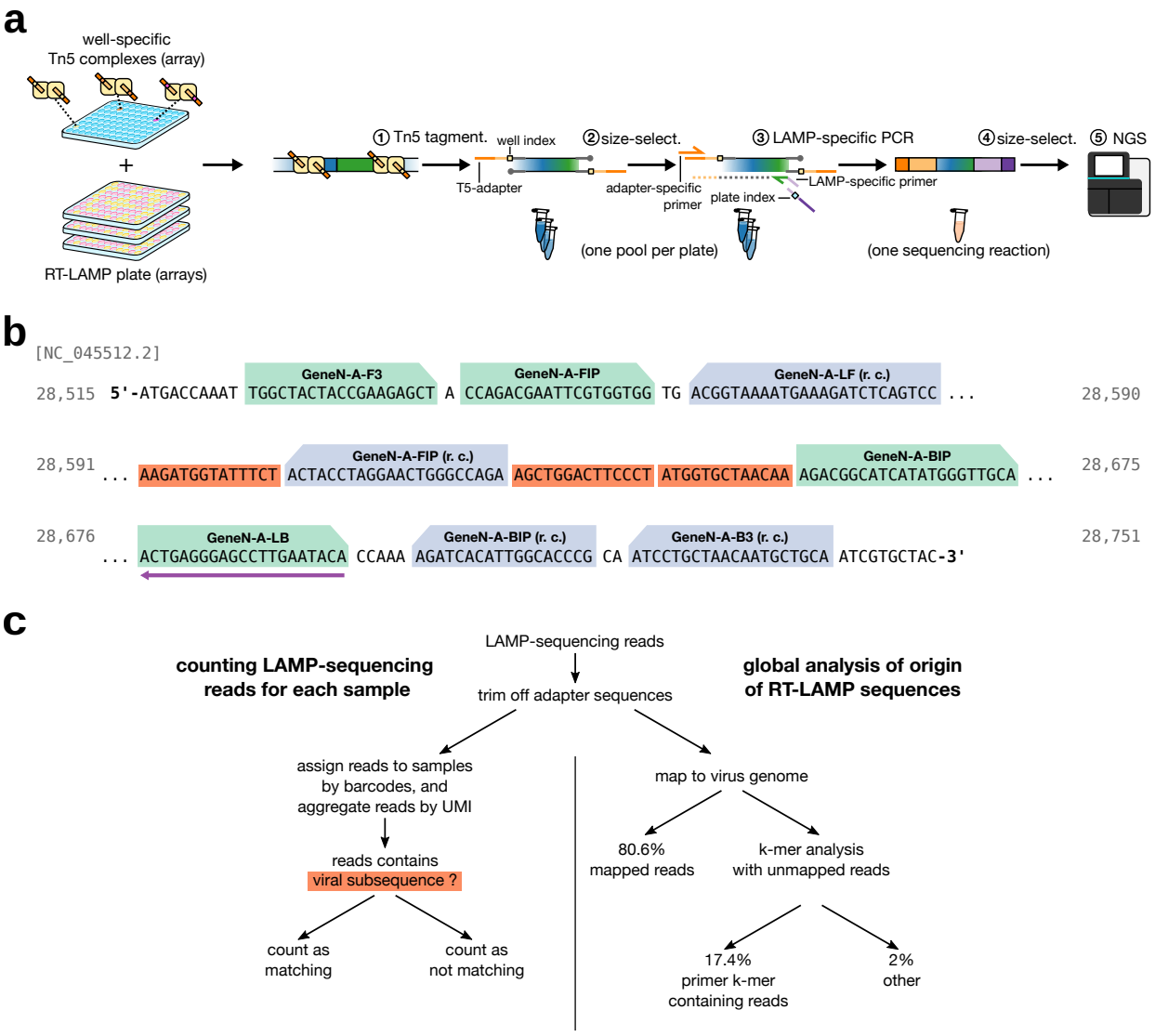


Figure 3.28: LAMP-sequencing workflow. **(a)** Colorimetric RT-LAMP assay plates are each tagmented with an array of well-specific indexed Tn5 complexes and then pooled to one sample per assay plate. These samples are enriched by a nested PCR with plate-specific indices resembling Tn5-Anchor-Seq without biotin-streptavidin-capture. After PCR the samples are pooled to one sequencing reaction, size-selected and analyzed by NGS. **(b)** Region of the N gene in the SARS-CoV-2 genome (NC_045512.2) covered by the different parts of the RT-LAMP primers (green and blue boxes). Some viral subsequences are not covered by primers (orange boxes) and were used for classifying "virus-matching" and "non-matching" reads. The purple arrow spans the anchor sequence used for LAMP-sequencing NGS library preparation. **(c)** Schematic of the computational analysis of the LAMP-sequencing data. *The figure was reproduced in modified form from [115] with permission.*

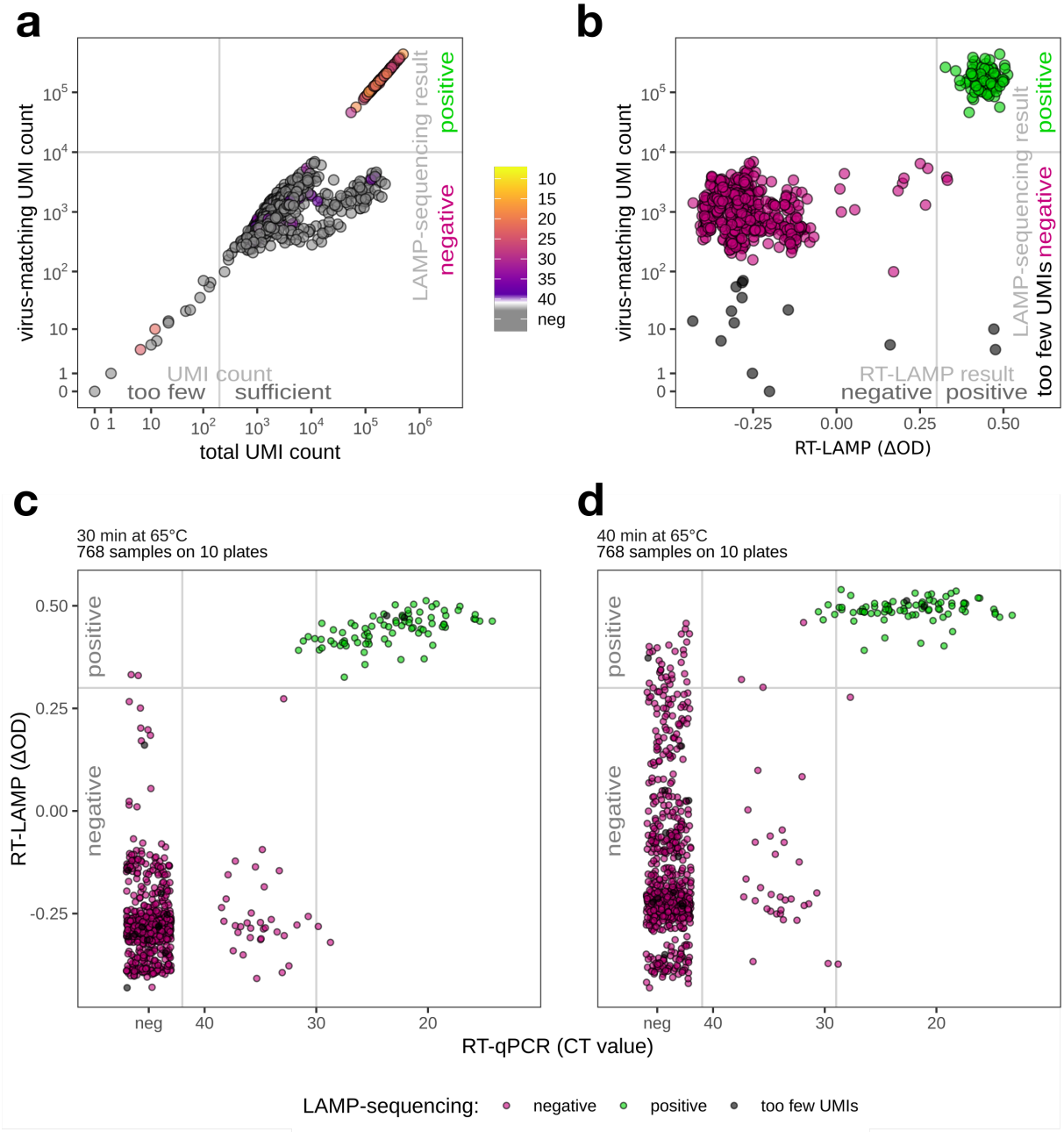


Figure 3.29: LAMP-sequencing read counting. **(a)** Applied thresholds for sample classification based on UMI counts. The total UMI count threshold of 200 removed 14 outlier samples for which too few reads were observed to make a call (see also Supplementary Figure 4.4). A threshold of 10^4 virus-matching UMIs separated the LAMP-sequencing positive and negative samples. **(b)** Comparison of LAMP-sequencing and RT-LAMP results for the 768 samples. **(c)** LAMP-sequencing results mapped onto RT-LAMP (at 30 minutes incubation) and RT-qPCR results. **(d)** Same as panel c but with RT-LAMP results after 40 minutes incubation (45 minutes for one plate). At this point in time the RT-LAMP samples were recovered for LAMP-sequencing. *The figure was reproduced in modified form from [115] with permission. The underlaying data was accquired by Daniel Kirrmaier under my supervision.*

set, confirming a highly specific amplification of the SARS-CoV-2 genome (Figure 3.30a). To explore the possibility of unspecific amplification at lower frequency the remaining unmapped reads were subjected to kmer analysis to explore their sequence composition. The resulting k-mers constituting 89.8 % of unmatched sequences (17.4 % of all reads) carried sequences of the LAMP primers used for the assay implicating that these primers explain most of the non-viral sequence content of the LAMP reactions (Figure 3.30b). Approximately 2 % of sequences could neither be explained by mapping to the viral genome sequence nor by unbiased k-mer analysis (Figure 3.30b). As this rate was relatively low their sequence content was not further investigated. It could be concluded that the RT-LAMP reaction is highly specific for the SARS-CoV-2 genome.

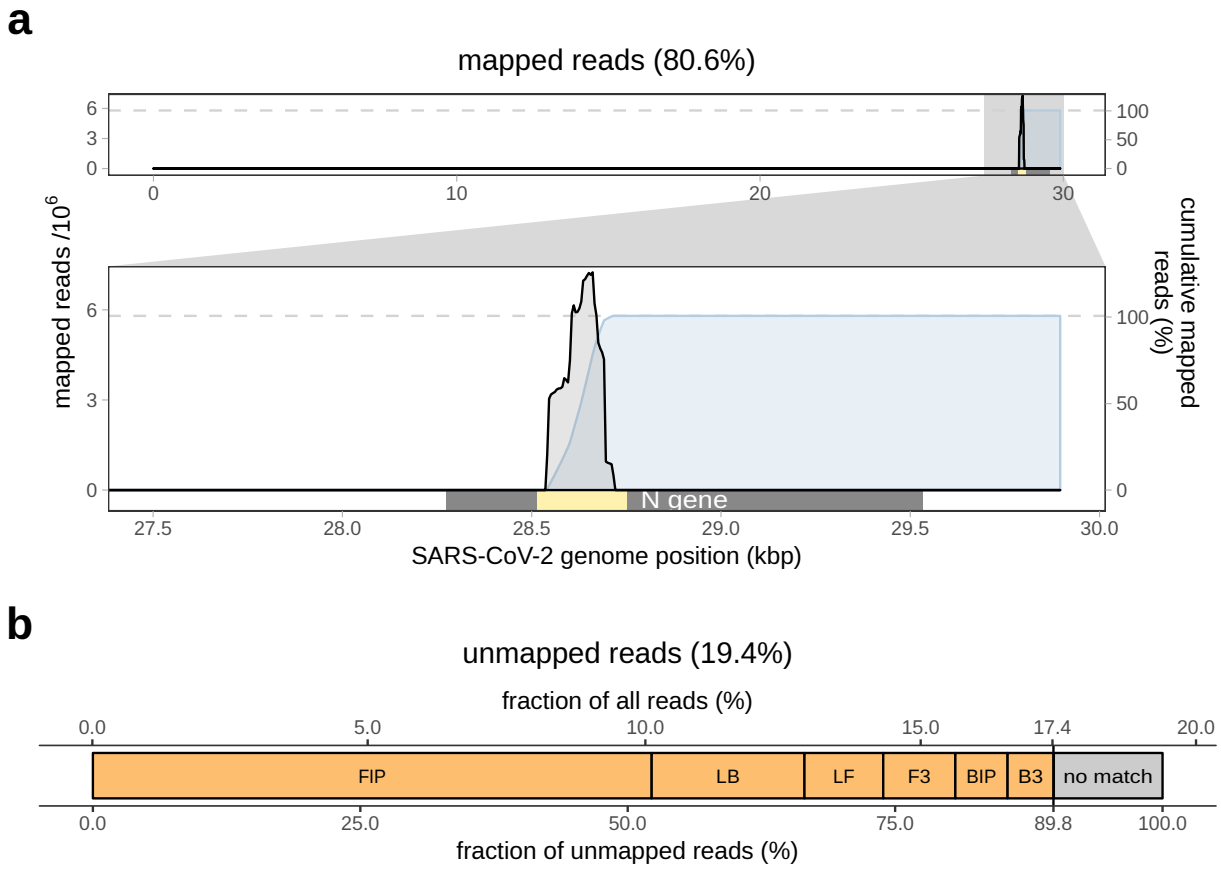


Figure 3.30: Origin of RT-LAMP sequences. **(a)** After adapter trimming 80.6 % of reads could be mapped onto a single locus of the SARS-CoV-2 genome (NC_045512.2) which is coinciding with the region of the N gene amplified by the RT-LAMP primer set. **(b)** A k-mer analysis (9-mers, maximal Levenshtein-distance of 2) was performed with the remaining 19.4 % unmatched reads indicating that 17.4 % of all reads contained sequences from one of the RT-LAMP primers. The origin remained undetermined for 2 % of all reads. *The figure was reproduced in modified form from [115] with permission.*

4 Discussion

4.1 Tagmentation-mediated Anchor-Seq allows to streamline genome walking sequencing

The targeted amplification and characterization of an unknown sequence adjacent to a genomic site of interest is known as genome walking and finds application for various biologically relevant questions. This includes the mapping or confirmation of insertion sites in random and targeted knock-in mutants which are for example created by transposon mutagenesis or gene targeting respectively. Another application is the characterization of genomic targets and their proximal sequences in an unbiased manner which is relevant for example for genome engineering and gene therapy. Anchor-Seq is a targeted next-generation strategy which allows to perform such genome walking sequencing [134].

One established genome walking approach is termed vectorette PCR [91]. The initial Anchor-Seq protocol was implemented by integrating vectorette PCR with a standard NGS library preparation including sonication, A tailing and ligation [134]. The protocol has been used by our laboratory to characterize and validate the C-SWAT collection, which is an arrayed library of *Saccharomyces cerevisiae* knock-in mutants. A custom computational pipeline was used to confirm the knock-in integration sites based on the expected outcome. Both parts, preparation and NGS of the Anchor-Seq samples and their computational analysis were well-designed for the particular purpose of validating the C-SWAT collection. However, further applications of this protocol were hindered by several unfavorable design aspects which motivated further protocol improvements. Similar as alternative protocols, Anchor-Seq requires several experimental steps before individual samples are barcoded for pooling which made it difficult to scale the protocol up for more than 30 samples [87, 88, 134]. In addition, it was favorable to re-design the computational analysis which had the advantage that less *a priori* knowledge is required and samples can be analyzed in a more unbiased manner.

Therefore, the Tn5-Anchor-Seq protocol was developed to succeed the Anchor-Seq protocol. Its central design element is the use of tagmentation for introducing generic adapters into the DNA of the sample [83]. This enzymatic reaction substitutes physical fragmentation of the DNA by sonication, end repair and adapter ligation. The introduction of tagmentation is a considerable simplification of the Anchor-Seq protocol as this reaction takes about 10 minutes rather than hours to complete [85, 86]. A prerequisite for tagmentation is that the generic adapter design which was initially based on a vectorette structure is replaced with a halfY structure (i.e. two annealed oligos which differ in

length). I confirmed that the halfY adapters provides an enrichment of approximately 30 million fold which I considered sufficient for targeted sequencing applications. One commonly used Tn5 transposase variant established for tagmentation carries the mutations E54K and L372P [85]. Recently, it was reported that an additional mutation in the DNA binding domain of Tn5 transposase (R27S) resulted in a stronger dependence of the tagmented DNA size distribution on the enzyme-DNA concentration ratio [86]. For the studies presented in this thesis it was decided to use the original Tn5 transposase containing only two mutations which is not as dependent on enzyme-DNA concentration ratio as the three residue mutant Tn5 transposase. The rational was that the Tn5-Anchor-Seq protocol should be as invariant as possible with respect to variation in the input DNA amount. It was also critical to determine the optimal Tn5 transposase to DNA concentration ratio. On the one hand it was observed that too high concentrations led to highly fragmented DNA. This results in sequencing reads which are too short to be informative about the anchor-adjacent sequences. On the other hand too low Tn5 transposase to DNA concentrations resulted in too few tagmented molecules which finally led to low resolution Anchor-Seq results. I showed that an optional biotin-enrichment-capture step can be performed between the PCR steps which allows for moderate improvements in enrichment. For most parts of this thesis NGS with Illumina short-read technology was used. I observed that Tn5-Anchor-Seq without biotin-streptavidin-capture led to longer molecules which makes it attractive for experiments relying on long-read sequencing technologies. The reproducibility and linearity of the Tn5-Anchor-Seq results was confirmed over five orders of magnitude. This experiment also revealed that care must be taken, if one wants to simultaneously use several anchor sequences in the same PCR reaction during Tn5-Anchor-Seq preparation. In this particular experiment, UP- and DOWN-stream genomic sequences of an integrated cassette were amplified. After sequencing, five-times more reads were observed for the UP than for the DOWN amplification indicating that the primers used for the UP direction performed better than the primers for the DOWN direction. This means that primer designs must be carefully balanced if such pooled amplifications are performed.

Tn5-Anchor-Seq is not the only protocol describing genome walking sequencing using tagmentation for introduction of the generic adapters before PCR. To my knowledge the first integration of tagmentation and genome walking sequencing was communicated as the Tagmentation-Based Mapping (TagMap) protocol by David Stern. It was used to characterize the insertion sites of mobile genetic elements in *Drosophila* lines [158]. A similar purpose serves the piggyBac insertion site sequencing (PBISeq) protocol used for mapping transposon insertion sites in the yeast species *Candida albicans* [159]. Insertion site mapping also plays an important role in gene therapy applications. The tagmentation-assisted PCR (tag-PCR) protocol was published for characterizing piggyBac-, retrovirus-

and lentivirus-based integration systems in mammalian cells [160]. A quantitative protocol using UMIs was developed as the iPool-Seq protocol to identify new virulance factors in a screen with pooled mutants of the fungus *Ustilago maydis* which is a maize pathogen [161]. UMIs also have been implemented in the Uni-Directional Targeted Sequencing (UDiTaS) protocol which has been developed for the unbiased detection and characterization of complex genome editing events such as translocations, inversions and insertions in mammalian cells [162]. Finally, Tagmentation-based tag integration site sequencing (TTISS) was also developed for genome editing experiments and specifically applied to determine specificity of different Cas9 variants [163]. The protocols differ in several aspects. PBISeq and tag-PCR use the commercial Nextera kit reagents from Illumina [160, 159]. The Nextera technology is actually intended for untargeted amplifications. Two generic adapters are introduced during tagmentation with these kits which can result in high background by off-target amplification lowering sensitivity for anchor site detection. The remaining protocols use a custom transposase preparation with a single generic adapter which circumvents such problems. UMI incorporation in the iPool-Seq protocol requires exchanging sequencing primers while the UDiTaS adapter design only requires slight adaptations of the Illumina NGS run parameters [161, 162]. Specificity is increased in the TTISS protocol using a nested PCR design while iPool-Seq uses biotin-streptavidin-capture [161, 163]. Multiplexing in all protocols is implemented at the PCR stage with the exception of the UDiTaS protocol which performs multiplexing at the tagmentation stage [162]. Performing the barcoding for multiplexing already at the tagmentation stage has the advantage that samples can be pooled as early as possible during the protocol which can potentially reduce the number of reactions needed to be handled.

Tn5-Anchor-Seq nevertheless is unique in that it tries to unify most of the advantages of the other protocols while avoiding their disadvantages. Custom transposon adapters lead to tagmented molecules with the same generic adapter so that all molecules can serve as templates in the PCR amplification step. This should improve resolution of the experiment [164]. The adapter design of Tn5-Anchor-Seq optimizes run performance and no custom sequencing primers are required as the adapters are directly compatible with sequencing primers of the standard Illumina kit. Specificity is improved by using two subsequent PCR steps with a nested primer design. It can be improved even further by incorporating a biotin-streptavidin-capture step between the two PCRs. UMIs are a central design element of the Tn5-Anchor-Seq adapter and easily recorded by slightly changing the Illumina NGS run parameters. Finally, the Tn5-Anchor-Seq adapters are barcoded. This allows to uniquely label samples during the tagmentation step of the protocol so that they can be pooled before conducting the subsequent PCR amplification steps. This reduces experimental burden and cost. Individual PCR reactions can again be

barcoded during the second PCR step which further increases the number of processable samples. This feature of the protocol was critical for implementing LAMP-sequencing, a specific application of Tn5-Anchor-Seq to characterize several hundred RT-LAMP reactions in parallel. More details on the application for SARS-CoV-2 detection are described in dedicated sections of this thesis. Although not implemented here, even higher levels of multiplexing can be achieved through digital encoding of sample identity when insertion sites are specific for samples. In this strategy samples are pooled repetitively such that each sample occurs in a unique order of sample pools. Sample identity is therefore encoded by the insertion site absence or presence in all pools. Results for individual samples can be inferred *in silico* after sequencing [165, 166]. In addition to the experimental improvements a more generic computational workflow was designed. The previous workflow used for validation of the C-SWAT library was focused on *a priori* known insertion sites whereas I designed the current workflow so that no such prior information is required. The computational workflow is build in a modular fashion which will allow for simplified adaptation for future experiments or changes in the Tn5-Anchor-Seq design.

In comparison to Anchor-Seq and Tn5-Anchor-Seq, inverse PCR would have the advantage to simultaneously provide information on upstream and downstream sequences adjacent to the inserted sequence of interest. This allows for insertion site determination with high confidence. Inverse PCR protocols have been successfully applied for genome walking NGS for the characterization of insertion sites in complex samples (e.g. saturated transposition experiments in *Saccharomyces cerevisiae* [167]). Nevertheless, inverse PCR requires optimal ligation condition to favor circularization of single molecules by intramolecular ligation. These circular molecules are the substrates required for specific amplification of target sites of interest. Such ligation reactions are difficult to scale as they need to be conducted with diluted concentrations of input material. This complicates high-throughput protocol development based on inverse PCR. As Anchor-Seq typically yields a high coverage at insertion sites correlative observation for more than one sequencing direction can confirm an insertion site. One alternative approach might be motivated by the observation that Tn5 transposase remains physically linked to the DNA after tagmentation. A two-staged indexing approach based on dilution and pooling could then be used to highly multiplex these linked molecules so that they can be assigned together *in silico* after sequencing. An NGS protocol based on this principle has been termed contiguity preserving transposase sequencing (CPT-seq) and has for example allowed for haplotype resolved sequencing [168, 169]. The Tn5-Anchor-Seq protocol is compatible with such a barcoding approach when UP and DOWN amplifications are performed within the same reaction. Since this strategy helps to associate reads originating from the same molecule linked UP and DOWN sites could be inferred after sequencing.

The use of tagmentation provides the potential to further streamline the Tn5-Anchor-Seq

protocol. The tagmentation reaction has not only been shown to work with purified DNA but also with minimally processed mammalian and yeast cells [170, 171]. Omission of the DNA extraction step in the Tn5-Anchor-Seq protocol provides the opportunity to further reduce sample processing time which could allow for more rapid and cost-efficient insights. Later protocol applications could therefore profit from the investigation if tagmentation of minimally processed cells would yield reliable results.

4.2 CASTLING: An avenue for pooled library construction of complex genotypes

Resolving the relationship between an observed phenotype and its causal genotype remains a fundamental challenge in biology. For this, screens in which genetic perturbations are conducted for all genes of an organism in a systematic manner are particularly instrumental [172]. Such screens and the construction of respective genotype collections (i.e. libraries) are conducted either in an arrayed or in a pooled format. Arrayed libraries have the advantage that the genotype of a certain cell line is directly mapped to the physical storing position on the array grid of the collection. However, because every single cell line construction must be performed on its own the effort scales proportionally with the intended coverage of the library and therefore the effort is relatively high for genome-wide collections. Furthermore, the library is restricted to the genetic background it was constructed in and because of the initially high investment, the arrayed library approach lacks versatility. As an alternative, pooled library constructions have the potential to reduce this initial investment [1]. Two technological advancements of the recent years fueled the development of pooled library construction approaches. On the one hand, genetic alterations are more accessible since the introduction of CRISPR-Cas endonucleases as they allow for precise and targeted genetic alterations in an easily programmable fashion [173, 174]. On the other hand, microarray-based synthesis of long oligonucleotides provide the necessary material to program these genetic alterations for several thousand targets at reasonable cost [146, 175]. Pooled library construction with CRISPR-Cas9 targeting has been performed in bacteria [138], yeast [138, 139, 140, 141, 142, 143] and mammalian cells [176] but these developments were limited to small genetic alterations comprising few nucleotides. In contrast, complex alterations such as large insertions are required for example for experiments involving genetic tags.

We introduced CASTLING which allows for pooled library construction with gene tagging genotypes in *Saccharomcyes cerevisiae* [116]. The strategy applies the concept of self-integrating cassettes (SICs), genetic constructs which combine all the information required for CRISPR-Cas12a-assisted gene targeting on the same molecules. Using this concept gene targeting is enhanced by directing CRISPR-Cas12a to the target gene of interest for DSB induction. We could show that with the CASTLING strategy tagging

efficiency and fidelity are considerably improved which allows for pooled library constructions. CASTLING encompasses an *in vitro* cloning strategy to construct SIC pools from microarray-synthesized oligonucleotide pools, the use of these SIC pools to construct yeast libraries and finally the genotyping of such libraries using Anchor-Seq. The integrated SICs provide the anchoring sequences for Anchor-Seq and therefore provide access to the genotypes in the pool directly while most other approaches only rely on indirect inference of genotype counts from sequencing crRNAs or other barcodes [138, 139, 140, 141, 142, 143].

Starting with a selected subset of 215 yeast genes it was shown how CASTLING can indeed yield robust and comprehensive pooled libraries with tag genotypes. The genes were selected based on an easily quantifiable phenotype when C-terminally tagged with mNeonGreen. This allowed to determine a tagging fidelity of 90 % for the pooled CASTLING process which recapitulates the results of individual gene taggings with single SICs. When I inspected the junction between the coding sequence of the endogenous gene and the tag the most likely explanation for failed taggings were deletions introduced by the oligonucleotides used to convey the targeting information [116, 146]. The likelihood for an oligonucleotide to contain errors increases with its length and since relatively long sequences are needed for the CASTLING strategy, a substantial fraction of the oligonucleotide pool molecules will contain errors [116]. Given that the tagging fidelity of a CASTLING experiment is similar to that of single gene taggings it can be concluded that the *in vitro* and *in vivo* CASTLING processes select against those erroneous oligonucleotide sequences likely because several steps are facilitated by considerable sequence homologies. When the individual steps of the SIC pool construction during CASTLING were sequenced it was observed that oligonucleotide sequences are more likely to be depleted than enriched during the RCA step and that this amplification seems to be stochastic as only moderate reproducibility was observed for replicative reactions [116].

Genotyping CASTLING libraries with the small nuclear pool design revealed that a coverage of up to 90 % of the targeted genes was reproducibly possible. When attempting a genome-wide pool design approximately 50 % coverage was achieved. Interestingly, the number of recovered clones after transformation only had a limited impact on gene coverage while using distinct SIC pool preparations increased genotype complexity. This is explained by the observation that the genotype distribution is skewed because of the overrepresentation of some genotypes in the pool. This imbalance in sequence representation was already observed in the starting material of the CASTLING process which is the amplified oligonucleotide pool. The overrepresentation of sequences is problematic in that it reduces the coverage of targeted genes which can be achieved with a certain number of clones in a library [177]. As it was observed for the CASTLING libraries with genome-wide pool design increasing recovered clone numbers alone did not alleviate

library coverage in a scalable manner.

It would therefore be interesting to further investigate how the CASTLING protocol could be improved. First experiments were conducted in the lab in which the yield of the *in vitro* homology-directed assembly reaction was improved so that the RCA can be conducted with higher template amounts. As this should reduce the skewed genotype representation of the SIC preparation step of CASTLING this potentially could improve reproducibility. This assumption remains to be tested by sequencing of the individual reaction steps.

Additionally, several strategies could be explored to reduce the skew in the sequence abundance distribution by removing highly abundant sequences. These strategies can be subdivided into untargeted and targeted approaches. One untargeted approach would be to perform *in vitro* compartmentalization of minute enzymatic reaction volumes with water-in-oil emulsions which has been shown to alleviate problems of bulk PCR and other reactions [178]. In contrast, one could also perform untargeted depletion of overrepresented sequences by using thermostable duplex-specific crab nuclease which has been used for cDNA [179] and RNA-Seq [180] library normalization. As a targeted alternative, *in vitro* depletion with Cas9 ribonucleoproteins specific for undesirable (i.e. overrepresented) sequences has been already used for NGS library normalization [181] and CRISPR pool customization [182]. Finally, several CASTLING libraries could be constructed in series using oligonucleotide pool designs which are incrementally updated by removing the overrepresented oligonucleotide sequences. These suggestions provide opportunities for improving the efficiency of the CASTLING protocol in the future.

In general, a heterogeneous pooled library is fractionated for analysis into bins of differing phenotype which are then individually genotyped by NGS [1]. Perspective applications for CASTLING libraries depend on the type of tag used for genotype construction and the assay which can be used for phenotypic characterization. Of particular interest are fluorescent protein tags because they can inform about protein abundance, localization and stability [4, 5]. Previously, pooled libraries constructed with fluorescent protein tags could be analyzed most easily with fluorescence-activated cell sorting (FACS) and one application of this approach for protein abundance profiling is showcased in this thesis. Unfortunately, the accessible phenotypic space is limited when FACS is applied. More recently, the application space of pooled tag libraries constructed with CASTLING is broadened by the development of image-activated cell sorting (IACS) which integrates high-content image acquisition with machine learning for decision making and cell sorting (reviewed in ref. [183]). An alternative approach for pooled library sorting based on more complex cellular features than with FACS is based on cells which express a photo-convertible marker. This is achieved by identifying interesting cells by microscopic imaging. Laser targeting is used to label such cells by photoconversion of the marker. The

labeled cells are subsequently sorted by FACS [184]. High-content microscopy also has been shown to enable phenotypic characterization and subsequent *in situ* genotyping of pooled libraries [185, 186]. This approach was integrated with CRISPR-mediated library construction in bacteria [187] and mammalian cells [188, 189]. Finally, pooled CRISPR screens have been integrated with single-cell RNA-seq enabling the interrogation of a massive number of phenotypes [190, 191, 192, 193]. CASTLING extends the available genetic modifications for this type of high-throughput genetic analysis.

In summary, CASTLING provides a strategy to create pooled libraries of complex tag genotype collections for functional genetics experimentation and shot-gun cell biology using targeted pooled screens.

4.3 Single-gene and modular PCR tagging in mammalian cells

From early on gene targeting of large heterologous constructs in mammalian cells required long homologous sequences of several hundred nucleotides which necessitates tedious cloning for HR donor template generation for every single target [16]. Later, it became clear that simultaneous DSB induction using target site-specific nucleases increased targeting efficiencies to a level which allowed for the use of HR templates with homology length as short as 50 nucleotides [26]. Cloning-free generation of the HR template by PCR was therefore feasible. The CRISPR-Cas9 and -Cas12a endonucleases are exceptionally appealing as site-specific endonucleases because of how easily they can be repurposed to a new target. However, earlier published strategies based on these endonucleases still provided separate components for crRNA expression for endonuclease targeting and tagging construct for repair by HR. This contradicted the aim of the simplified PCR tagging idea [194, 195, 196, 197, 198]. However, the SIC concept by design unifies both components for CRISPR-mediated tagging in the same molecule and therefore allows for a protocol which just requires a single PCR reaction to produce all required reagents for targeting a gene of interest in mammalian cells. Together with my colleagues I presented the consequent implementation and characterization of this procedure [117].

We showed that the SIC design, originally developed for CRISPR-enhanced gene tagging in yeast cells (see section 3.2; ref. [116]) can also facilitate a PCR tagging strategy in mammalian cells [117]. For this a SIC is generated by performing one PCR-reaction using target-specific primers and a template containing the tag sequence. For C-terminal tagging the reverse primer provides homologies to the 3'UTR of the targeted gene as well as the crRNA sequence guiding DSB induction around the stop codon of the targeted gene. The crRNA expression is driven by a PolIII promoter included in the tag template. Homologies to the coding sequence of the target gene are introduced using the forward primer. The resulting PCR products, the PCR cassettes, are transfected into mammalian cells together with a helper plasmid for CRISPR-Cas12a expression. When all these

components are provided they allow for efficient and specific gene tagging. Nevertheless, an unexpected phenotype of diffuse cytoplasmic fluorescence was observed during transfection of PCR cassettes for C-terminal gene tagging with mNeonGreen. This artifact occurred independently from CRISPR-Cas12a expression and was unstable. [117].

Earlier studies have established that transfected DNA molecules are readily concatemerized into tandem head-to-tail structures via HR. In case the DNA is linear, intramolecular ligation by NHEJ leads to circularization of the exogenous material. These circular molecules then preferentially recombine intermolecularly into concatemers in a tandem head-to-tail conformation mediated by HR [16, 17]. Head-to-tail concatemers were furthermore very recently reported in the context of CRISPR-Cas9-mediated genome editing experiments in mice [199]. There are several lines of evidence presented in this thesis which support the model of *in vivo* concatemerization of the transfected PCR cassettes via this mechanism. Anchor-Seq of transfections with PCR cassettes for single genes as well as cassette mixtures indicated that the linear PCR cassette molecules circularize preferentially by intramolecular ligation in head-to-tail conformation likely mediated by NHEJ. This was further validated in junction-specific PCR reactions. Further indirect evidence for head-to-tail ligation of the PCR cassette comes from the occurrence of the cytoplasmic tag artifact itself. Head-to-tail ligation brings the 5'end of the coding sequence of the tag into close proximity with the PolIII promoter which usually drives expression of the crRNA gene. It has been shown that PolIII promoters can also allow for PolII expression, including the U6 promoter used here [153, 154]. The observed diffuse cytoplasmic GFP signal therefore likely originates from expression of mNeonGreen which is not fused to any upstream coding sequence. The mNeonGreen expression is driven by the U6 PolIII promoter in head-to-tail ligated PCR cassettes. This would further suggest that inhibiting NHEJ and therefore preventing end-to-end ligation would reduce this artifact as it was observed when modified primers were used for PCR cassette preparation. In addition, changing the expression context for example by removing potential start codons from the mNeonGreen coding sequence reduced the occurrence of cells with cytoplasmic fluorescence. It can be expected that independent tag expression upon head-to-tail ligation depends on various circumstances. For example after NHEJ the resulting DNA sequence needs to allow for efficient translation. This likely explains why some PCR cassettes exhibit more cytoplasmic artifacts than others. The decrease of the number of cells with diffuse cytoplasmic signal over time is an indication that most molecules which are ligated in head-to-tail conformation are genetically unstable and therefore likely extrachromosomal. Nevertheless, even after 30 days of cultivation these head-to-tail ligation products could be detected by PCR and had therefore likely chromosomally integrated. This is evidence that at least a small fraction of head-to-tail ligations involve more than one copy of the PCR cassette in tandem which supports the model that the linear PCR cassettes

are not only intramolecularly ligated by NHEJ but also intermolecularly concatemerized by HR as suggested by Folger *et al.* [16, 17].

Integration events involving concatemerized PCR cassettes are unproblematic for the expression of the tagged endogenous gene as the PCR cassette includes a heterologous sequence for proper termination. However, independent tag expression from downstream head-to-tail junctions leading to fusions between the U6 PolIII promoter and the tag coding sequence might be considered an obstacle of the PCR tagging approach. It is quite likely that only a few molecules provide the necessary genetic sequence for efficient and independent tag expression. One indication for this is that cells with diffuse cytoplasmic signal were rarely observed after a prolonged time of cultivation after transfection although head-to-tail ligation products were still well detectable by PCR as mentioned above. The integration of concatemerized PCR cassettes could be avoided using the following strategy. Folger *et al.* observed that also the circular DNA of intact plasmids is a competent substrate for HR similar to linerarized DNA which circularized in the cells. However, plasmid DNA and circularized linear DNA are substrates for HR only for a limited time after transfection [200]. For CRISPR-Cas9 mediated gene targeting a plasmid excision strategy has been proposed. In this strategy the HR donor is cloned into a plasmid which is then transfected into cells. In addition, these cells also express sgRNA-Cas9 enzyme complexes targeting the genomic locus of interest as well as the plasmid containing the HR donor. Consequently, DSBs are introduced on the plasmid so that the HR template is excised and available for HR-mediated repair inside the cells [201]. Based on the results by Folger *et al.* it seems likely that the donor plasmid undergoes HR and form head-to-tail concatemers upon transfection. Nevertheless, after CRISPR-Cas9 targeting and excision the HR template molecules would again be single copies. Because transfected DNA seems to be able to concatemerize only for a limited time it can be assumed that the linearized HR template molecules will not concatemerize again if there is some temporal delay between concatemerization and plasmid excision. Therefore, no integration of concatemerized HR template molecules should be observed for a plasmid excision strategy. It would be interesting to perform a respective experiment for the PCR tagging strategy. On the contrary, cloning SICs into plasmids would abrogate one of the advantages of the otherwise cloning-free PCR tagging approach.

The mammalian C-terminal tagging experiments of highly expressed and localized genes with the mNeonGreen tag presented in this thesis allow for an easily distinguishable phenotype. It was generally desirable to determine the fidelity of the junction between the coding sequences of the endogenous gene and the tag. In order to do so I used NGS of the PCR-amplified junctions to quantify the occurrence of mutations. Tagging fidelity was high as more than 80 % of the reads exhibited the expected junction sequence. I could observe that the likelihood of observing an error was generally higher towards the

beginning of the tag coding sequence indicating that they might originate from incorrect primer sequences which are then only partially corrected by mismatch repair during HR. This is in agreement with the observation that HR becomes less efficient in tolerating errors the further they are away from the DSB [202]. Analysis of the tagging experiment of the gene *CANX* also revealed an opportunity for design rule optimizations for the PCR tagging primers. In this case the targeting sequence was positioned so that DSB was induced before the stop codon. Furthermore, the PCR cassette integration reconstituted a sufficient target site so that the resulting sequence was again subject to further cutting by the CRISPR-Cas12a complex. This lead to an increased rate of deletions most likely introduced by NHEJ of the resulting DSB. Such constellations can be generally avoided when target sites are placed so that the DSB is induced only after the stop codon.

A similar phenomenon was observed when I analyzed the wild-type allele in an analogues manner. Around 10 % of the reads carried a mutation in an approximately 20 nucleotide wide window around the target site where the DSB should occur. These numbers are an underestimation as the genomic DNA was extracted from unselected cells and therefore combine the events from altered, unaltered and untransfected cells. Nevertheless, assuming that transfection efficiencies in HEK293T cells are usually around 50-90 % [151, 152], these numbers should be off by no more than two-fold. In case the free ends of the Cas12-mediated DSB are directly repaired by NHEJ the wild-type allele is reconstituted and can again be a target for the Cas12a complex. This cycle can repeat until the sequence of the target site is either disrupted when the PCR cassette is integrated (either by HR or NHEJ) or by mutations introduced during DSB repair. One can expect that CRISPR-Cas12a-induced DSBs which have a four nucleotide overhang are very efficient substrates for NHEJ and might often be repaired without errors. It would be interesting to repeat the experiment with cells which were selected for a stably integrated PCR cassette using a selection marker or to characterize individual clonal lines to be able to better estimate the prevalence of modifications of the wild-type locus in successfully targeted cells. As a potential consequence of stimulating mutations at the terminal side of the wild-type locus, expression might be hampered, because these mutations can trigger surveillance pathways such as nonsense-mediated decay [203].

It might be possible to further optimize the mammalian PCR tagging protocol by including procedures to modulate DSB repair pathway selection in the target cells. Several strategies have been shown to enhance editing efficiencies in CRISPR-Cas experiments by stimulating HR activity. These strategies include small molecules or heterologous protein expression to restrict DSB repair to a certain cell cycles phase, or to either block NHEJ or activate HR. In addition, tethering of the HR repair template to the DSB has been reported to improve editing efficiencies (reviewed in [204]). Some of these approaches are currently under investigation for integration with the mammalian PCR tagging strategy.

Frequently, it has been noticed that CRISPR-Cas9 targeting can be unspecific and that targeting at unintended sites occurs [205, 206, 207]. CRISPR-Cas12a had been reported to be more specific but the concern persisted that the PCR tagging strategy itself could lead to increased off-target integrations [208, 209, 210, 211]. To characterize to which extend off-target integration plays a role in the PCR tagging strategy HEK293T cells were transfected with PCR cassettes in the presence or absence of co-transfected CRIPSR-Cas12a helper plasmid. Tn5-Anchor-Seq was used to determine genomic integrations of PCR cassettes without the prior knowledge of genomic position. This data on the one hand confirmed that PCR targeting at on-target sites required CRISPR-Cas12a expression as such sites were not observed in transfection in which the CRISPR-Cas12a helper plasmid was omitted. On the other hand, this data showed that several off-target integrations of the PCR cassette occurred for each transfection but that these integration events appeared to be randomly scattered throughout the genome and that integration was independent on CRISPR-Cas12a expression. It could therefore be concluded that random integrations of the PCR cassettes are a frequent by-product of transfections as it had been previously observed [16, 212]. As CRISPR-Cas12a-mediated cleavage is not a requirement for these off-target integrations of the PCR cassette the most likely explanation for their presence is the occurrence of spontaneous DSBs and HR-independent repair thereof. It would be interesting to further investigate how these off-target integrations generalize to other types of donor DNA (e.g. circular double-stranded or linear single-stranded DNA) or other cell lines than HEK293T which was used for the experiments presented here.

Another possibility to improve PCR tagging could be to include a step in which cells with random integrations of the PCR cassette are selectively removed. A respective approach termed positive negative selection (PNS) was previously introduced for gene targeting in mice [213]. In this strategy the gene targeting construct contains two markers. One marker confers drug resistance (i.e. positive selection) and the other marker confers sensitivity to another drug (i.e. negative selection). The HR donor template is structured in a way so that only the positive selection marker is genomically integrated if HR occurs with the donor template. In case the construct is randomly integrated by NHEJ both markers are preferentially integrated because the negative selection marker extends the homology arm to one side. In consequence, the integrating cell becomes sensitive to the drug used for negative selection. After transfection cells are selected with both drugs to selectively kill cells which either have integrated the donor template by NHEJ or which have not integrated the template at all. This strategy allowed for an approximately 2000-fold increase in cells with HR events over NHEJ events [213]. If selection would be an option in a PCR tagging experiment one could add a suitable counter-selectable marker to the PCR cassette for example by fusion PCR using one homology arm as overlap. Nevertheless, this strategy first needs to be tested because few nucleotides of non homologous sequences

at the end of the PCR cassettes impaired targeting efficiency [117] and the homologous sequences used in the original PNS study had a length of several thousand nucleotides while PCR tagging involves very short homologous sequences of less than 100 nucleotides.

In summary, mammalian PCR tagging provides an CRISPR-Cas12a-mediated route for specific tag genotype construction in mammalian cells. As with every gene targeting strategy, it cannot prevent the need to confirm the correct integration [214] but because this strategy omits the need for elaborate cloning of genetic constructs it lower the bar for performing gene targeting experiments [117].

4.4 LAMP-sequencing validates a diagnostic SARS-CoV-2 assay

As of writing of this thesis the COVID-19 pandemic is still an ongoing threat to the world community. While vaccination programs have allowed to return back to some sort of normality in privileged countries the spread of the COVID-19 causing virus SARS-CoV-2 remains a problematic reality [215]. Therefore, countermeasures such as extensive diagnostic testing to reveal and control outbreak situations remain critically important [216].

One of the most sensitive assays for SARS-CoV-2 detection is based on RT-qPCR and is a widely applied testing solution [103, 102]. Especially during the beginning of the pandemic reagents for this assay were limited which exposed the need to evaluate alternative assay modalities. Here the diagnostic potential of a colorimetric RT-LAMP assay was evaluated under real-world conditions with patient samples. In addition, an adaptation of Tn5-Anchor-Seq termed LAMP-sequencing supported the validation of this RT-LAMP assay [115].

A published colorimetric RT-LAMP assay for SARS-CoV-2 detection was established with a spectrometric readout [130, 115]. In total, extracted RNA from 768 pharyngeal swab samples was evaluated with this assay. In addition, the same material had been quantified for SARS-CoV-2 genomes using a gold standard RT-qPCR assay. The samples covered a wide range of virus titers which allowed to test the diagnostic potential of the SARS-CoV-2 RT-LAMP assay. Across all samples a high specificity was observed. For samples with a high to intermediate viral load, i.e. CT values below 30, the sensitivity was also high. However, the sensitivity dropped considerably for samples with lower viral titers (CT>30). According to the manufacturer of the RT-qPCR assay used here, a CT value of 30 corresponds to ∼1000 viral genomes in the RT-qPCR reaction. It is difficult to say whether an individual with such a viral load would still be contagious. In any case, the RT-LAMP assay proved valuable in detecting cases with high to intermediate viral loads.

One difficulty while implementing the RT-LAMP assay was the occurrence of positive RT-LAMP reactions upon prolonged incubation. It was initially unclear if this was the result of inconsistent specific amplification due to low viral titers or due to spurious

amplification [217]. It was therefore necessary to further characterize the LAMP products of these samples. I designed an adaptation of Tn5-Anchor-Seq for highly multiplexed sequencing of RT-LAMP reaction products which was termed LAMP-sequencing. In addition to the adapter-specific primer the second PCR was performed with plate-specific index primers and a LAMP-amplicon-specific bridging primer. Such a bridging primer allows to adapt a new anchor sequence while the plate-specific index primers can remain the same which reduces the investment for Tn5-Anchor-Seq protocol adaptations.

Almost all RT-LAMP assay results were confirmed by LAMP-sequencing. It needs to be stressed that LAMP-sequencing can only inform about the nature of the amplified LAMP product and not about the absence of template in case the LAMP amplification failed. However, RT-LAMP reactions which turned positive after a prolonged incubation time contained no evidence for amplification of viral sequences. Instead, only sequences covering the LAMP assay primers sequences were detected. This finding supports the hypothesis that without template LAMP primer sets might still initiate priming at a lower rate leading to unspecific amplifications. Such artifacts could by avoided by including sequence specific probes although this could only be implemented at the expense of the simple assay readout by reaction color [218].

One disadvantage of the colorimetric RT-LAMP assay is that no internal control can be included in contrast to the RT-qPCR assay because only the presence or absence of DNA amplification is scored. Multiplexed LAMP assays using differently labeled fluorecence probes have been reported to be problematic and this readout alternative would again require specialized equipment [219]. LAMP-sequencing provides an alternative readout modality for RT-LAMP assays. This might also hold true for multiplexed RT-LAMP reactions but would require further investigations.

As alternative to LAMP-sequencing the LAMP primers of the assay could also be barcoded directly so that samples can be pooled right after the RT-LAMP reaction. However, the barcode sequences within the LAMP primers can have an impact on the amplification efficiency and hence must be carefully selected for every new LAMP primer set [220].

Taken together, LAMP-sequencing confirmed that a colorimetric RT-LAMP can be used as SARS-CoV-2 diagnostic. The assay performs very well for the detection of high to intermediate viral loads. Together with its low-cost and scalability this assay is therefore exceptionally well suited for sentinel testing. Our lab explored and implemented such a strategy which also relied on self-sampled saliva specimens [221]. In principle, LAMP-sequencing would allow to characterize a large number of such diagnostic samples because of its high multiplexing capabilities.

4.5 Conclusion

In this thesis I presented my work on developing scalable technologies for gene tagging in yeast as well as mammalian cells and their characterization by NGS.

Cataloging natural diversity by next-generation sequencing and multi-parametric omic studies creates a multitude of testable hypotheses. Yet, experimental approaches to currently generate respective genetic manipulations to test such hypothesis are lagging behind. One obstacle is the widespread construction of arrayed mutant collections as their scalability in terms of resources and effort is limited. Here I was involved in the implementation and characterization of the CASTLING methodology. This approach applies the programmable endonuclease CRISPR-Cas12a and oligonucleotide pools to generate pooled collections of *Saccharomcyes cerevisiae* strains with tagged genes. The method was extensively validated and potential shortcomings were identified. The ultimate goal of single constructions of pooled collections with genome-wide coverage was explored. This work paves the way to generate complex genetic alterations in a pooled format for functional genetics applications. I discussed how CASTLING will develop its full potential as soon as appropriate assay technologies become available.

In a collaborative effort I have continued with applying the ideas from the CASTLING strategy to enhance gene tagging in mammalian cells. Previous approaches are experimentally tedious and relatively inefficient. The application of CASTLING-like DNA cassettes in mammalian cells allowed for a simplified PCR-based tagging strategy. Parameters influencing efficiency and quality of the tagging were identified including mapping of potential off-target integrations of the PCR tagging cassette. Along the way an experimental artifact was observed which was explained by arbitrary tag expression. I suggested a model based on concatemerized PCR tagging cassettes which would explain this artifact. The model was supported by direct and circumstantial evidence. Taken together, this study provides a strategy for streamlined gene tagging in mammalian cells which reduces the effort needed to gain biological insights.

Central for these studies was the development of an improved NGS strategy for genome walking sequencing which allows for the targeted enrichment of unknown sequences adjacent to genomic sites of interest. The substitution of several experimental steps of our original genome walking protocol Anchor-Seq with one *in vitro* tagmentation step allowed for more rapid and efficient experimentation. In addition, this Tn5-Anchor-Seq protocol was influenced by ideas from several complementary approaches and unified their advantages. I implemented and used throughout my work versatile computational workflows to analyze the Tn5-Anchor-Seq data. In the end I provided ideas on how the Tn5-Anchor-Seq protocol could be further applied and improved.

An important aspect which motivated the development of Tn5-Anchor-Seq was scalabil-

ity. During the beginning of the SARS-CoV-2 pandemic in 2020 our lab was involved in implementing a diagnostic RT-LAMP assay for viral genome detection. I helped to validate this assay by designing and performing an application of Tn5-Anchor-Seq termed LAMP-sequencing which allowed to sequence hundreds of RT-LAMP reactions. This analysis was critical in that it validated the quality and feasibility of the RT-LAMP assay as an alternative modality for large scale SARS-CoV-2 testing. In addition, this work confirmed that the Tn5-Anchor-Seq approach is indeed highly scalable and can be easily and rapidly adopted to new applications.

In conclusion, this thesis provides and applies several technological advancements in the field of NGS, genome engineering and diagnostic testing which open up new opportunities to elucidate fundamental biological questions.

Bibliography

[1] Michael Lawson and Johan Elf. "Imaging-Based Screens of Pool-Synthesized Cell Libraries". In: *Nature Methods* 18.4 (2021), pp. 358–365. DOI: 10.1038/s41592-020-01053-8.

[2] Guri Giaever et al. "Functional Profiling of the Saccharomyces cerevisiae Genome". In: *Nature* 418.6896 (2002), pp. 387–391. DOI: 10.1038/nature00935.

[3] Sina Ghaemmaghami et al. "Global Analysis of Protein Expression in Yeast". In: *Nature* 425.6959 (2003), pp. 737–741. DOI: 10.1038/nature02046.

[4] Won-Ki Huh et al. "Global Analysis of Protein Localization in Budding Yeast". In: *Nature* 425.6959 (2003), pp. 686–691. DOI: 10.1038/nature02026.

[5] Anton Khmelinskii et al. "Tandem Fluorescent Protein Timers for in Vivo Analysis of Protein Dynamics". In: *Nature Biotechnology* 30.7 (2012), pp. 708–714. DOI: 10.1038/nbt.2281.

[6] Tim van Opijnen and Henry L. Levin. "Transposon Insertion Sequencing, a Global Measure of Gene Function". In: *Annual Review of Genetics* 54.1 (2020), pp. 337–365. DOI: 10.1146/annurev-genet-112618-043838.

[7] Guri Giaever and Corey Nislow. "The Yeast Deletion Collection: a Decade of Functional Genomics". In: *Genetics* 197.2 (2014), pp. 451–465. DOI: 10.1534/genetics.114.161620.

[8] Rodolphe Barrangou and Philippe Horvath. "A Decade of Discovery: CRISPR Functions and Applications". In: *Nature Microbiology* 2.7 (2017), p. 17092. DOI: 10.1038/nmicrobiol.2017.92.

[9] Gavin J. Knott and Jennifer A. Doudna. "CRISPR-Cas Guides the Future of Genetic Engineering". In: *Science* 361.6405 (2018), pp. 866–869. DOI: 10.1126/science.aat5011.

[10] F. Zhang. "Development of CRISPR-Cas Systems for Genome Editing and Beyond". In: *Quarterly Reviews of Biophysics* 52.nil (2019), e6. DOI: 10.1017/s0033583519000052.

[11] Adrian Pickar-Oliver and Charles A. Gersbach. "The Next Generation of CRISPR-Cas Technologies and Applications". In: *Nature Reviews Molecular Cell Biology* 20.8 (2019), pp. 490–507. DOI: 10.1038/s41580-019-0131-5.

[12] Yutaka Yamamoto and Susan A. Gerbi. "Making Ends Meet: Targeted Integration of DNA Fragments By Genome Editing". In: *Chromosoma* 127.4 (2018), pp. 405–420. DOI: 10.1007/s00412-018-0677-6.

[13] Svetlana A. Smirnikhina, Arina A. Anuchina, and Alexander V. Lavrov. "Ways of Improving Precise Knock-In By Genome-Editing Technologies". In: *Human Genetics* 138.1 (2018), pp. 1–19. DOI: 10.1007/s00439-018-1953-5.

[14] Albert Hinnen, James B. Hicks, and Gerald R. Fink. "Transformation of Yeast." In: *Proceedings of the National Academy of Sciences* 75.4 (1978), pp. 1929–1933. DOI: 10.1073/pnas.75.4.1929.

[15] David Shortle, James E. Haber, and David Botstein. "Lethal Disruption of the Yeast Actin Gene By Integrative DNA Transformation". In: *Science* 217.4557 (1982), pp. 371–373. DOI: 10.1126/science.7046050.

[16] Kim R. Folger et al. "Patterns of Integration of DNA Microinjected Into Cultured Mammalian Cells: Evidence for Homologous Recombination Between Injected Plasmid DNA Molecules." In: *Molecular and Cellular Biology* 2.11 (1982), pp. 1372–1387. DOI: 10.1128/mcb.2.11.1372.

[17] Kim R. Folger, Kirk R. Thomas, and Mario R. Capecchi. "Analysis of Homologous Recombination in Cultured Mammalian Cells". In: *Cold Spring Harbor Symposia on Quantitative Biology* 49.0 (1984), pp. 123–138. DOI: 10.1101/sqb.1984.049.01.016.

[18] Oliver Smithies et al. "Insertion of DNA Sequences Into the Human Chromosomal β-globin Locus By Homologous Recombination". In: *Nature* 317.6034 (1985), pp. 230–234. DOI: 10.1038/317230a0.

[19] Terry L. Orr-Weaver, Jack W. Szostak, and Rodney J. Rothstein. "Yeast Transformation: a Model System for the Study of Recombination." In: *Proceedings of the National Academy of Sciences* 78.10 (1981), pp. 6354–6358. DOI: 10.1073/pnas.78.10.6354.

[20] Kirk R. Thomas, Kim R. Folger, and Mario R. Capecchi. "High Frequency Targeting of Genes To Specific Sites in the Mammalian Genome". In: *Cell* 44.3 (1986), pp. 419–428. DOI: 10.1016/0092-8674(86)90463-0.

[21] A Baudin et al. "A Simple and Efficient Method for Direct Gene Deletion In *Saccharomyces cerevisiae*". In: *Nucleic Acids Research* 21.14 (1993), pp. 3329–3330. DOI: 10.1093/nar/21.14.3329.

[22] Palaniyandi Manivasakam et al. "Micro-Homology Mediated PCR Targeting Insaccharomyces Cerevisiae". In: *Nucleic Acids Research* 23.14 (1995), pp. 2799–2800. DOI: 10.1093/nar/23.14.2799.

[23] Achim Wach et al. "Heterologous HIS3 Marker and GFP Reporter Modules for PCR-Targeting in Saccharomyces cerevisiae". In: *Yeast* 13.11 (1997), pp. 1065–1075. DOI: 10.1002/(sici)1097-0061(19970915)13:11<1065::aid-yea159>3.0.co;2-k.

[24] Michael Knop et al. "Epitope Tagging of Yeast Genes Using a PCR-Based Strategy: More Tags and Improved Practical Routines". In: *Yeast* 15.10B (1999), pp. 963–972.

DOI: 10.1002/(sici)1097-0061(199907)15:10b<963::aid-yea399>3.0.co;2-w.

[25] Carsten Janke et al. "A Versatile Toolbox for PCR-Based Tagging of Yeast Genes: New Fluorescent Proteins, More Markers and Promoter Substitution Cassettes". In: *Yeast* 21.11 (2004), pp. 947–962. DOI: 10.1002/yea.1142.

[26] Salvatore J. Orlando et al. "Zinc-Finger Nuclease-Driven Targeted Integration Into Mammalian Genomes Using Donors With Limited Chromosomal Homology". In: *Nucleic Acids Research* 38.15 (2010), e152–e152. DOI: 10.1093/nar/gkq512.

[27] Zhenguo Lin et al. "Origins and Evolution of the recA/RAD51 Gene Family: Evidence for Ancient Gene Duplication and Endosymbiotic Gene Transfer". In: *Proceedings of the National Academy of Sciences* 103.27 (2006), pp. 10328–10333. DOI: 10.1073/pnas.0604232103.

[28] Bailin Zhao et al. "The Molecular Basis and Disease Relevance of Non-Homologous DNA End Joining". In: *Nature Reviews Molecular Cell Biology* 21.12 (2020), pp. 765–781. DOI: 10.1038/s41580-020-00297-8.

[29] Minoru Takata et al. "Homologous Recombination and Non-Homologous End-Joining Pathways of DNA Double-Strand Break Repair Have Overlapping Roles in the Maintenance of Chromosomal Integrity in Vertebrate Cells". In: *The EMBO Journal* 17.18 (1998), pp. 5497–5508. DOI: 10.1093/emboj/17.18.5497.

[30] L. C. Kadyk and L. H. Hartwell. "Sister Chromatids Are Preferred Over Homologs As Substrates For Recombinational Repair In Saccharomyces cerevisiae." In: *Genetics* 132.2 (1992), pp. 387–402. DOI: 1322387[pii].

[31] Raphael Ceccaldi, Beatrice Rondinelli, and Alan D. D'Andrea. "Repair Pathway Choices and Consequences At the Double-Strand Break". In: *Trends in Cell Biology* 26.1 (2016), pp. 52–64. DOI: 10.1016/j.tcb.2015.07.009.

[32] Andrea Beucher et al. "ATM and Artemis Promote Homologous Recombination of Radiation-Induced DNA Double-Strand Breaks in G2". In: *The EMBO Journal* 28.21 (2009), pp. 3413–3427. DOI: 10.1038/emboj.2009.276.

[33] Guillermo E. Taccioli et al. "Impairment of V(D)J Recombination in Double-Strand Break Repair Mutants". In: *Science* 260.5105 (1993), pp. 207–210. DOI: 10.1126/science.8469973.

[34] Eleni P. Mimitou and Lorraine S. Symington. "Ku Prevents Exo1 and Sgs1-dependent Resection of DNA Ends in the Absence of a Functional MRX Complex Or Sae2". In: *The EMBO Journal* 29.19 (2010), pp. 3358–3369. DOI: 10.1038/emboj.2010.193.

[35] Cristina Escribano-Díaz et al. "A Cell Cycle-Dependent Regulatory Circuit Composed of 53BP1-RIF1 and BRCA1-CtIP Controls DNA Repair Pathway Choice". In: *Molecular Cell* 49.5 (2013), pp. 872–883. DOI: 10.1016/j.molcel.2013.01.001.

[36] Sylvie M. Noordermeer et al. "The Shieldin Complex Mediates 53BP1-dependent DNA Repair". In: *Nature* 560.7716 (2018), pp. 117–121. DOI: 10.1038/s41586-018-0340-7.

[37] Rajat Gupta et al. "DNA Repair Network Analysis Reveals Shieldin As a Key Regulator of NHEJ and PARP Inhibitor Sensitivity". In: *Cell* 173.4 (2018), 972–988.e23. DOI: 10.1016/j.cell.2018.03.050.

[38] Ulf Grawunder et al. "Activity of DNA Ligase IV Stimulated By Complex Formation With XRCC4 Protein in Mammalian Cells". In: *Nature* 388.6641 (1997), pp. 492–495. DOI: 10.1038/41358.

[39] Thomas E. Wilson, Ulf Grawunder, and Michael R. Lieber. "Yeast DNA Ligase IV Mediates Non-Homologous DNA End Joining". In: *Nature* 388.6641 (1997), pp. 495–498. DOI: 10.1038/41365.

[40] Howard H. Y. Chang et al. "Different DNA End Configurations Dictate Which NHEJ Components Are Most Important for Joining Efficiency". In: *Journal of Biological Chemistry* 291.47 (2016), pp. 24377–24389. DOI: 10.1074/jbc.m116.752329.

[41] Bailin Zhao et al. "The Essential Elements for the Noncovalent Association of Two DNA Ends During Nhej Synapsis". In: *Nature Communications* 10.1 (2019), p. 3588. DOI: 10.1038/s41467-019-11507-z.

[42] Aaron A. Goodarzi et al. "DNA-PK Autophosphorylation Facilitates Artemis Endonuclease Activity". In: *The EMBO Journal* 25.16 (2006), pp. 3880–3889. DOI: 10.1038/sj.emboj.7601255.

[43] Howard H.Y. Chang, Go Watanabe, and Michael R. Lieber. "Unifying the DNA End-Processing Roles of the Artemis Nuclease". In: *Journal of Biological Chemistry* 290.40 (2015), pp. 24036–24050. DOI: 10.1074/jbc.m115.680900.

[44] Stephanie A. Nick McElhinny et al. "A Gradient of Template Dependence Defines Distinct Biological Roles for Family X Polymerases in Nonhomologous End Joining". In: *Molecular Cell* 19.3 (2005), pp. 357–366. DOI: 10.1016/j.molcel.2005.06.012.

[45] John M. Pryor et al. "Essential Role for Polymerase Specialization in Cellular Nonhomologous End Joining". In: *Proceedings of the National Academy of Sciences* 112.33 (2015), E4537–E4545. DOI: 10.1073/pnas.1505805112.

[46] Benjamin E. Nelms et al. "In Situ Visualization of DNA Double-Strand Break Repair in Human Fibroblasts". In: *Science* 280.5363 (1998), pp. 590–592. DOI: 10.1126/science.280.5363.590.

[47] Michael Lisby et al. "Choreography of the DNA Damage Response". In: *Cell* 118.6 (2004), pp. 699–713. DOI: 10.1016/j.cell.2004.08.015.

[48] Ji-Hoon Lee and Tanya T. Paull. "ATM Activation By DNA Double-Strand Breaks Through the Mre11-Rad50-Nbs1 Complex". In: *Science* 308.5721 (2005), pp. 551–554. DOI: 10.1126/science.1108297.

[49] Alessandro A. Sartori et al. "Human CtIP Promotes DNA End Resection". In: *Nature* 450.7169 (2007), pp. 509–514. DOI: 10.1038/nature06337.

[50] Eleni P. Mimitou and Lorraine S. Symington. "Sae2, Exo1 and Sgs1 Collaborate in DNA Double-Strand Break Processing". In: *Nature* 455.7214 (2008), pp. 770–774. DOI: 10.1038/nature07312.

[51] Tomohiko Sugiyama, Elena M. Zaitseva, and Stephen C. Kowalczykowski. "A Single-Stranded DNA-Binding Protein Is Needed for Efficient Presynaptic Complex Formation By the Saccharomyces cerevisiae Rad51 Protein". In: *Journal of Biological Chemistry* 272.12 (1997), pp. 7940–7945. DOI: 10.1074/jbc.272.12.7940.

[52] Ryan B. Jensen, Aura Carreira, and Stephen C. Kowalczykowski. "Purified Human BRCA2 Stimulates Rad51-mediated Recombination". In: *Nature* 467.7316 (2010), pp. 678–683. DOI: 10.1038/nature09399.

[53] Akira Shinohara and Tomoko Ogawa. "Stimulation By Rad52 of Yeast Rad51-Mediated Recombination". In: *Nature* 391.6665 (1998), pp. 404–407. DOI: 10.1038/34943.

[54] James H. New et al. "Rad52 Protein Stimulates DNA Strand Exchange By Rad51 and Replication Protein A". In: *Nature* 391.6665 (1998), pp. 407–410. DOI: 10.1038/34950.

[55] Zhucheng Chen, Haijuan Yang, and Nikola P. Pavletich. "Mechanism of Homologous Recombination From the RecA-ssDNA/dsDNA Structures". In: *Nature* 453.7194 (2008), pp. 489–494. DOI: 10.1038/nature06971.

[56] Thijn van der Heijden et al. "Homologous Recombination in Real Time: DNA Strand Exchange By RecA". In: *Molecular Cell* 30.4 (2008), pp. 530–538. DOI: 10.1016/j.molcel.2008.03.010.

[57] Xuan Li et al. "PCNA Is Required for Initiation of Recombination-Associated DNA Synthesis By DNA Polymerase δ". In: *Molecular Cell* 36.4 (2009), pp. 704–713. DOI: 10.1016/j.molcel.2009.09.036.

[58] Wade M. Hicks, Minlee Kim, and James E. Haber. "Increased Mutagenesis and Unique Mutation Signature Associated With Mitotic Gene Conversion". In: *Science* 329.5987 (2010), pp. 82–85. DOI: 10.1126/science.1191125.

[59] Grzegorz Zapotoczny and Jeff Sekelsky. "Human Cell Assays for Synthesis-Dependent Strand Annealing and Crossing Over During Double-Strand Break Repair". In: *G3 (Bethesda)* 7.4 (2017), pp. 1191–1199. DOI: 10.1534/g3.116.037390.

[60] Anna H. Bizard and Ian D. Hickson. "The Dissolution of Double Holliday Junctions". In: *Cold Spring Harbor Perspectives in Biology* 6.7 (2014), a016477–a016477. DOI: 10.1101/cshperspect.a016477.

[61] Ralph Scully et al. “DNA Double-Strand Break Repair-Pathway Choice in Somatic Mammalian Cells”. In: *Nature Reviews Molecular Cell Biology* 20.11 (2019), pp. 698–714. DOI: 10.1038/s41580-019-0152-0.

[62] Philip M. Nussenzweig and Luciano A. Marraffini. “Molecular Mechanisms of CRISPR-Cas Immunity in Bacteria”. In: *Annual Review of Genetics* 54.1 (2020), pp. 93–120. DOI: 10.1146/annurev-genet-022120-112523.

[63] Francisco J. M. Mojica et al. “Biological Significance of a Family of Regularly Spaced Repeats in the Genomes of Archaea, Bacteria and Mitochondria”. In: *Molecular Microbiology* 36.1 (2000), pp. 244–246. DOI: 10.1046/j.1365-2958.2000.01838.x.

[64] Ruud Jansen et al. “Identification of Genes That Are Associated With DNA Repeats in Prokaryotes”. In: *Molecular Microbiology* 43.6 (2002), pp. 1565–1575. DOI: 10.1046/j.1365-2958.2002.02839.x.

[65] Francisco J M Mojica et al. “Intervening Sequences of Regularly Spaced Prokaryotic Repeats derive From Foreign Genetic Elements.” In: *Journal of Molecular Evolution* 60.2 (2005), pp. 174–82. DOI: 10.1007/s00239-004-0046-3.

[66] C. Pourcel, G. Salvignol, and G. Vergnaud. “CRISPR Elements in Yersinia Pestis Acquire New Repeats By Preferential Uptake of Bacteriophage DNA, and Provide Additional Tools for Evolutionary Studies”. In: *Microbiology* 151.3 (2005), pp. 653–663. DOI: 10.1099/mic.0.27437-0.

[67] Alexander Bolotin et al. “Clustered Regularly Interspaced Short Palindrome Repeats (CRISPRs) Have Spacers of Extrachromosomal Origin”. In: *Microbiology* 151.8 (2005), pp. 2551–2561. DOI: 10.1099/mic.0.28048-0.

[68] Rudolphe Barrangou et al. “CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes”. In: *Science* 315.5819 (2007), pp. 1709–1712. DOI: 10.1126/science.1138140.

[69] Martin Jinek et al. “A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity”. In: *Science* 337.6096 (2012), pp. 816–821. DOI: 10.1126/science.1225829.

[70] Giedrius Gasiunas et al. “Cas9-crRNA Ribonucleoprotein Complex Mediates Specific DNA Cleavage for Adaptive Immunity in Bacteria”. In: *Proceedings of the National Academy of Sciences* 109.39 (2012), E2579–E2586. DOI: 10.1073/pnas.1208507109.

[71] Le Cong et al. “Multiplex Genome Engineering Using CRISPR/Cas Systems”. In: *Science* 339.6121 (2013), pp. 819–823. DOI: 10.1126/science.1231143.

[72] Preshant Mali et al. “RNA-Guided Human Genome Engineering Via Cas9”. In: *Science* 339.6121 (2013), pp. 823–826. DOI: 10.1126/science.1232033.

[73] Wenyan Jiang et al. “RNA-Guided Editing of Bacterial Genomes Using CRISPR-Cas Systems”. In: *Nature Biotechnology* 31.3 (2013), pp. 233–239. DOI: 10.1038/nbt.2508.

[74] Bernd Zetsche et al. “Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System”. In: *Cell* 163.3 (2015), pp. 759–771. DOI: 10.1016/j.cell.2015.09.038.

[75] Kira S. Makarova et al. “Evolutionary Classification of CRISPR-Cas Systems: a Burst of Class 2 and Derived Variants”. In: *Nature Reviews Microbiology* 18.2 (2019), pp. 67–83. DOI: 10.1038/s41579-019-0299-x.

[76] Elitza Deltcheva et al. “CRISPR RNA Maturation By Trans-Encoded Small RNA and Host Factor RNAse III”. In: *Nature* 471.7340 (2011), pp. 602–607. DOI: 10.1038/nature09886.

[77] Jay Shendure et al. “DNA Sequencing At 40: Past, Present and Future”. In: *Nature* 550.7676 (2017), pp. 345–353. DOI: 10.1038/nature24286.

[78] A. M. Maxam and W. Gilbert. “A New Method for Sequencing DNA.” In: *Proceedings of the National Academy of Sciences* 74.2 (1977), pp. 560–564. DOI: 10.1073/pnas.74.2.560.

[79] F. Sanger, S. Nicklen, and A. R. Coulson. “DNA Sequencing With Chain-Terminating Inhibitors”. In: *Proceedings of the National Academy of Sciences* 74.12 (1977), pp. 5463–5467. DOI: 10.1073/pnas.74.12.5463.

[80] David R. Bentley et al. “Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry”. In: *Nature* 456.7218 (2008), pp. 53–59. DOI: 10.1038/nature07517.

[81] Sara Goodwin, John D. McPherson, and W. Richard McCombie. “Coming of Age: Ten Years of Next-Generation Sequencing Technologies”. In: *Nature Reviews Genetics* 17.6 (2016), pp. 333–351. DOI: 10.1038/nrg.2016.49.

[82] Al Edwards et al. “Automated DNA Sequencing of the Human HPRT Locus”. In: *Genomics* 6.4 (1990), pp. 593–608. DOI: 10.1016/0888-7543(90)90493-e.

[83] Andrew Adey et al. “Rapid, Low-Input, Low-Bias Construction of Shotgun Fragment Libraries By High-Density in Vitro Transposition”. In: *Genome Biology* 11.12 (2010), R119. DOI: 10.1186/gb-2010-11-12-r119.

[84] William S. Reznikoff. “Transposon Tn5”. In: *Annual Review of Genetics* 42.1 (2008), pp. 269–286. DOI: 10.1146/annurev.genet.42.110807.091656.

[85] Simone Picelli et al. “Tn5 Transposase and Tagmentation Procedures for Massively Scaled Sequencing Projects”. In: *Genome Research* 24.12 (2014), pp. 2033–2040. DOI: 10.1101/gr.177881.114.

[86] Bianca P. Hennig et al. “Large-Scale Low-Cost NGS Library Preparation Using a Robust Tn5 Purification and Tagmentation Protocol”. In: *G3* 8.1 (2018), pp. 79–89. DOI: 10.1534/g3.117.300257.

[87] Claudia Leoni et al. “Genome Walking in Eukaryotes”. In: *FEBS Journal* 278.21 (2011), pp. 3953–3977. DOI: 10.1111/j.1742-4658.2011.08307.x.

[88] Alfred J. Arulandhu et al. “DNA Enrichment Approaches To Identify Unauthorized Genetically Modified Organisms (GMOs)”. In: *Analytical and Bioanalytical Chemistry* 408.17 (2016), pp. 4575–4593. DOI: 10.1007/s00216-016-9513-0.

[89] André Rosenthal. “PCR Amplification Techniques for Chromosome Walking”. In: *Trends in Biotechnology* 10.nil (1992), pp. 44–48. DOI: 10.1016/0167-7799(92)90167-t.

[90] Venkatakrishna Shyamala and Giovanna Ferro-Luzzi Ames. “Genome Walking By Single-Specific-Primer Polymerase Chain Reaction: SSP-PCR”. In: *Gene* 84.1 (1989), pp. 1–8. DOI: 10.1016/0378-1119(89)90132-7.

[91] J. Riley et al. “A Novel, Rapid Method for the Isolation of Terminal Sequences From Yeast Artificial Chromosome (YAC) Clones”. In: *Nucleic Acids Research* 18.10 (1990), pp. 2887–2890. DOI: 10.1093/nar/18.10.2887.

[92] Tony Triglia, M. Gregory Peterson, and David J. Kemp. “A Procedure For *In Vitro* Amplification of DNA Segments That Lie Outside the Boundaries of Known Sequences”. In: *Nucleic Acids Research* 16.16 (1988), pp. 8186–8186. DOI: 10.1093/nar/16.16.8186.

[93] H. Ochman, A. S. Gerber, and D. L. Hartl. “Genetic Applications of An Inverse Polymerase Chain Reaction”. In: *Genetics* 120.3 (1988), pp. 621–3.

[94] Jay D. Parker, Peter S. Rabinovitch, and Glenna C. Burmer. “Targeted Gene Walking Polymerase Chain Reaction”. In: *Nucleic Acids Research* 19.11 (1991), pp. 3055–3060. DOI: 10.1093/nar/19.11.3055.

[95] Paul R. Mueller and Barbara Wold. “In Vivo Footprinting of a Muscle Specific Enhancer By Ligation Mediated PCR”. In: *Science* 246.4931 (1989), pp. 780–786. DOI: 10.1126/science.2814500.

[96] Anna Paruzynski et al. “Genome-Wide High-Throughput Integrome Analyses By nrLAM-PCR and Next-Generation Sequencing”. In: *Nature Protocols* 5.8 (2010), pp. 1379–1395. DOI: 10.1038/nprot.2010.87.

[97] G. P. Wang et al. “HIV Integration Site Selection: Analysis By Massively Parallel Pyrosequencing Reveals Association With Epigenetic Modifications”. In: *Genome Research* 17.8 (2007), pp. 1186–1194. DOI: 10.1101/gr.6286907.

[98] Jeffrey D. Gawronski et al. “Tracking Insertion Mutants Within Libraries By Deep Sequencing and a Genome-Wide Screen for Haemophilus Genes Required in the Lung”. In: *Proceedings of the National Academy of Sciences* 106.38 (2009), pp. 16422–16427. DOI: 10.1073/pnas.0906627106.

[99] Mariateresa Volpicella et al. “Genome Walking By Next Generation Sequencing Approaches”. In: *Biology* 1.3 (2012), pp. 495–507. DOI: 10.3390/biology1030495.

[100] Na Zhu et al. “A Novel Coronavirus From Patients With Pneumonia in China, 2019”. In: *New England Journal of Medicine* 382.8 (2020), pp. 727–733. DOI: 10.1056/nejmoa2001017.

[101] Jason Phua et al. “Intensive Care Management of Coronavirus Disease 2019 (COVID-19): Challenges and Recommendations”. In: *The Lancet Respiratory Medicine* 8.5 (2020), pp. 506–517. DOI: 10.1016/s2213-2600(20)30161-2.

[102] Ralph Weissleder et al. “COVID-19 Diagnostics in Context”. In: *Science Translational Medicine* 12.546 (2020), eabc1931. DOI: 10.1126/scitranslmed.abc1931.

[103] Victor M. Corman et al. “Detection of 2019 Novel Coronavirus (2019-nCoV) By Real-Time RT-PCR”. In: *Eurosurveillance* 25.3 (2020), nil. DOI: 10.2807/1560-7917.es.2020.25.3.2000045.

[104] Russell Higuchi et al. “Simultaneous Amplification and Detection of Specific DNA Sequences”. In: *Nature Biotechnology* 10.4 (1992), pp. 413–417. DOI: 10.1038/nbt0492-413.

[105] Russell Higuchi et al. “Kinetic PCR Analysis: Real-Time Monitoring of DNA Amplification Reactions”. In: *Nature Biotechnology* 11.9 (1993), pp. 1026–1030. DOI: 10.1038/nbt0993-1026.

[106] Christian A. Heid et al. “Real Time Quantitative PCR.” In: *Genome Research* 6.10 (1996), pp. 986–994. DOI: 10.1101/gr.6.10.986.

[107] Carl T. Wittwer et al. “Continuous Fluorescence Monitoring of Rapid Cycle DNA Amplification”. In: *BioTechniques* 22.1 (1997), pp. 130–138. DOI: 10.2144/97221bi01.

[108] Tsugunori Notomi et al. “Loop-Mediated Isothermal Amplification of DNA”. In: *Nucleic Acids Research* 28.12 (2000), 63e–63. DOI: 10.1093/nar/28.12.e63.

[109] Kentaro Nagamine, Tetsu Hase, and Tsugunori Notomi. “Accelerated Reaction By Loop-Mediated Isothermal Amplification Using Loop Primers”. In: *Molecular and Cellular Probes* 16.3 (2002), pp. 223–229. DOI: 10.1006/mcpr.2002.0415.

[110] Norihiro Tomita et al. “Loop-Mediated Isothermal Amplification (LAMP) of Gene Sequences and Simple Visual Detection of Products”. In: *Nature Protocols* 3.5 (2008), pp. 877–882. DOI: 10.1038/nprot.2008.57.

[111] Tsugunori Notomi et al. “Loop-Mediated Isothermal Amplification (LAMP): Principle, Features, and Future Prospects”. In: *Journal of Microbiology* 53.1 (2015), pp. 1–5. DOI: 10.1007/s12275-015-4656-9.

[112] Marianna Soroka, Barbara Wasowicz, and Anna Rymaszewska. “Loop-Mediated Isothermal Amplification (LAMP): the Better Sibling of PCR?” In: *Cells* 10.8 (2021), p. 1931. DOI: 10.3390/cells10081931.

[113] S. Fukuta et al. “Detection of Japanese Yam Mosaic Virus By RT-LAMP”. In: *Archives of Virology* 148.9 (2003), pp. 1713–1720. DOI: 10.1007/s00705-003-0134-5.

[114] Nathan A. Tanner, Yinhua Zhang, and Thomas C. Evans. "Visual Detection of Isothermal Nucleic Acid Amplification Using pH-Sensitive Dyes". In: *BioTechniques* 58.2 (2015), pp. 59–68. DOI: 10.2144/000114253.

[115] Viet Loan Dao Thi et al. "A Colorimetric RT-LAMP Assay and LAMP-sequencing for Detecting SARS-CoV-2 RNA in Clinical Samples". In: *Science Translational Medicine* 12.556 (2020), eabc7075. DOI: 10.1126/scitranslmed.abc7075.

[116] Benjamin C. Buchmuller et al. "Pooled Clone Collections By Multiplexed CRISPR-Cas12a-Assisted Gene Tagging in Yeast". In: *Nature Communications* 10.1 (2019), p. 2960. DOI: 10.1038/s41467-019-10816-7.

[117] Julia Fueller et al. "CRISPR-Cas12a-Assisted PCR Tagging of Mammalian Genes". In: *Journal of Cell Biology* 219.6 (2020), nil. DOI: 10.1083/jcb.201910210.

[118] Kenneth J. Livak and Thomas D. Schmittgen. "Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2-ΔΔCT Method". In: *Methods* 25.4 (2001), pp. 402–408. DOI: 10.1006/meth.2001.1262.

[119] Ben Langmead and Steven L. Salzberg. "Fast Gapped-Read Alignment With Bowtie 2". In: *Nature Methods* 9.4 (2012), pp. 357–359. DOI: 10.1038/nmeth.1923.

[120] S. Anders, P. T. Pyl, and W. Huber. "HTSeq - a Python Framework To Work With High-Throughput Sequencing Data". In: *Bioinformatics* 31.2 (2014), pp. 166–169. DOI: 10.1093/bioinformatics/btu638.

[121] Mark A. Sheff and Kurt S. Thorn. "Optimized Cassettes for Fluorescent Protein Tagging In Saccharomyces cerevisiae". In: *Yeast* 21.8 (2004), pp. 661–670. DOI: 10.1002/yea.1130.

[122] Anton Khmelinskii and Michael Knop. "Analysis of Protein Dynamics with Tandem Fluorescent Protein Timers". In: *Methods in Molecular Biology*. Methods in Molecular Biology. Springer New York, 2014, pp. 195–210. DOI: 10.1007/978-1-4939-0944-5_13.

[123] Tom Smith, Andreas Heger, and Ian Sudbery. "UMI-Tools: Modeling Sequencing Errors in Unique Molecular Identifiers To Improve Quantification Accuracy". In: *Genome Research* 27.3 (2017), pp. 491–499. DOI: 10.1101/gr.209601.116.

[124] Benjamin J. Callahan et al. "DADA2: High-Resolution Sample Inference From Illumina Amplicon Data". In: *Nature Methods* 13.7 (2016), pp. 581–583. DOI: 10.1038/nmeth.3869.

[125] Heng Li. "Minimap2: Pairwise Alignment for Nucleotide Sequences". In: *Bioinformatics* 34.18 (2018), pp. 3094–3100. DOI: 10.1093/bioinformatics/bty191.

[126] Kendell Clement et al. "CRISPResso2 Provides Accurate and Rapid Genome Editing Sequence Analysis". In: *Nature Biotechnology* 37.3 (2019), pp. 224–226. DOI: 10.1038/s41587-019-0032-3.

[127] Helga Thorvaldsdottir, James T. Robinson, and Jill P. Mesirov. "Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Explo-

ration". In: *Briefings in Bioinformatics* 14.2 (2012), pp. 178–192. DOI: 10.1093/bib/bbs017.

[128] Heng Li. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs With BWA-MEM". In: *CoRR* (2013). arXiv: 1303.3997 [q-bio.GN].

[129] Konrad Herbst et al. "Colorimetric RT-LAMP and LAMP-sequencing for Detecting SARS-CoV-2 RNA in Clinical Samples". In: *Bio-Protocol* 11.6 (2021), nil. DOI: 10.21769/bioprotoc.3964.

[130] Yinhua Zhang et al. *Rapid Molecular Detection of SARS-CoV-2 (COVID-19) Virus RNA Using Colorimetric LAMP*. 2020. DOI: 10.1101/2020.02.26.20028373.

[131] Marcel Martin. "Cutadapt Removes Adapter Sequences From High-Throughput Sequencing Reads". In: *EMBnet.journal* 17.1 (2011), p. 10. DOI: 10.14806/ej.17.1.200.

[132] Fan Wu et al. *Complete Genome Characterisation of a Novel Coronavirus Associated With Severe Human Respiratory Disease in Wuhan, China*. 2020. DOI: 10.1101/2020.01.24.919183.

[133] Heng Li et al. "The Sequence Alignment/Map Format and SAMtools". In: *Bioinformatics* 25.16 (2009), pp. 2078–2079. DOI: 10.1093/bioinformatics/btp352.

[134] Matthias Meurer et al. "Genome-Wide C-SWAT Library for High-Throughput Yeast Genome Tagging". In: *Nature Methods* 15.8 (2018), pp. 598–600. DOI: 10.1038/s41592-018-0045-8.

[135] Uri Weill et al. "Genome-Wide SWAp-Tag Yeast Libraries for Proteome Exploration". In: *Nature Methods* 15.8 (2018), pp. 617–622. DOI: 10.1038/s41592-018-0044-9.

[136] Douglas W. Fadrosh et al. "An Improved Dual-Indexing Approach for Multiplexed 16s rRNA Gene Sequencing on the Illumina MiSeq Platform". In: *Microbiome* 2.1 (2014), p. 6. DOI: 10.1186/2049-2618-2-6.

[137] Teemu Kivioja et al. "Counting Absolute Numbers of Molecules Using Unique Molecular Identifiers". In: *Nature Methods* 9.1 (2011), pp. 72–74. DOI: 10.1038/nmeth.1778.

[138] Andrew D. Garst et al. "Genome-Wide Mapping of Mutations At Single-Nucleotide Resolution for Protein, Metabolic and Genome Engineering". In: *Nature Biotechnology* 35.1 (2016), pp. 48–55. DOI: 10.1038/nbt.3718.

[139] Meru J. Sadhu et al. "Highly Parallel Genome Variant Engineering With CRISPR-Cas9". In: *Nature Genetics* 50.4 (2018), pp. 510–514. DOI: 10.1038/s41588-018-0087-y.

[140] Kevin R. Roy et al. "Multiplexed Precision Genome Editing With Trackable Genomic Barcodes in Yeast". In: *Nature Biotechnology* 36.6 (2018), pp. 512–520. DOI: 10.1038/nbt.4137.

[141] Xiaoge Guo et al. “High-Throughput Creation and Functional Profiling of DNA Sequence Variant Libraries Using CRISPR-Cas9 in Yeast”. In: *Nature Biotechnology* 36.6 (2018), pp. 540–546. DOI: 10.1038/nbt.4147.

[142] Eilon Sharon et al. “Functional Genetic Variants Revealed By Massively Parallel Precise Genome Editing”. In: *Cell* 175.2 (2018), 544–557.e16. DOI: 10.1016/j.cell.2018.08.057.

[143] Zehua Bao et al. “Genome-Scale Engineering of *Saccharomyces cerevisiae* With Single-Nucleotide Precision”. In: *Nature Biotechnology* 36.6 (2018), pp. 505–508. DOI: 10.1038/nbt.4132.

[144] René Verwaal et al. “CRISPR/Cpf1 Enables Fast and Simple Genome Editing Of *Saccharomyces cerevisiae*”. In: *Yeast* 35.2 (2017), pp. 201–211. DOI: 10.1002/yea.3278.

[145] Benjamin Dubreuil et al. “YeastRGB: Comparing the Abundance and Localization of Yeast Proteins Across Cells and Libraries”. In: *Nucleic Acids Research* 47.D1 (2018), pp. D1245–D1249. DOI: 10.1093/nar/gky941.

[146] Sriram Kosuri and George M. Church. “Large-Scale De Novo DNA Synthesis: Technologies and Applications”. In: *Nature Methods* 11.5 (2014), pp. 499–507. DOI: 10.1038/nmeth.2918.

[147] John R. S. Newman et al. “Single-Cell Proteomic Analysis of *S. cerevisiae* Reveals the Architecture of Biological Noise”. In: *Nature* 441.7095 (2006), pp. 840–846. DOI: 10.1038/nature04785.

[148] Min-Woo Lee et al. “Global Protein Expression Profiling of Budding Yeast in Response To DNA Damage”. In: *Yeast* 24.3 (2007), pp. 145–154. DOI: 10.1002/yea.1446.

[149] George S. Davidson et al. “The Proteomics of Quiescent and Nonquiescent Cell Differentiation in Yeast Stationary-Phase Cultures”. In: *Molecular Biology of the Cell* 22.7 (2011), pp. 988–998. DOI: 10.1091/mbc.e10-06-0499.

[150] Brandon Ho, Anastasia Baryshnikova, and Grant W. Brown. “Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces Cerevisiae Proteome”. In: *Cell Systems* 6.2 (2018), 192–205.e3. DOI: 10.1016/j.cels.2017.12.004.

[151] Rosalie Maurisse et al. “Comparative Transfection of DNA Into Primary and Transformed Mammalian Cells From Different Lineages”. In: *BMC Biotechnology* 10.1 (2010), p. 9. DOI: 10.1186/1472-6750-10-9.

[152] Yao-Cheng Lin et al. “Genome Dynamics of the Human Embryonic Kidney 293 Lineage in Response To Cell Biology Manipulations”. In: *Nature Communications* 5.1 (2014), p. 4767. DOI: 10.1038/ncomms5767.

[153] Mohammad Rumi et al. “RNA Polymerase II Mediated Transcription From the Polymerase III Promoters in Short Hairpin RNA Expression Vector”. In: *Biochem-*

ical and Biophysical Research Communications 339.2 (2006), pp. 540–547. DOI: 10.1016/j.bbrc.2005.11.037.

[154] Zongliang Gao, Elena Herrera-Carrillo, and Ben Berkhout. "RNA Polymerase II Activity of Type 3 Pol III Promoters". In: *Molecular Therapy - Nucleic Acids* 12.nil (2018), pp. 135–145. DOI: 10.1016/j.omtn.2018.05.001.

[155] Jose A. Gutierrez-Triana et al. "Efficient Single-Copy HDR By 5' Modified Long dsDNA Donors". In: *eLife* 7.nil (2018), nil. DOI: 10.7554/elife.39468.

[156] Lin Yu et al. "Rapid Detection of Covid-19 Coronavirus Using a Reverse Transcriptional Loop-Mediated Isothermal Amplification (RT-LAMP) Diagnostic Platform". In: *Clinical Chemistry* 66.7 (2020), pp. 975–977. DOI: 10.1093/clinchem/hvaa102.

[157] Mohamed El-Tholotha, Haim H. Bau, and Jinzhao Song. *A Single and Two-Stage, Closed-Tube, Molecular Test for the 2019 Novel Coronavirus (COVID-19) at Home, Clinic, and Points of Entry*. 2020. DOI: 10.26434/chemrxiv.11860137.v1.

[158] David L. Stern. *Tagmentation-Based Mapping (TagMap) of Mobile DNA Genomic Insertion Sites*. 2016. DOI: 10.1101/037762.

[159] Zeyao Li et al. "Genome-Wide piggyBac Transposon-Based Mutagenesis and Quantitative Insertion-Site Analysis in Haploid Candida Species". In: *Nature Protocols* 15.8 (2020), pp. 2705–2727. DOI: 10.1038/s41596-020-0351-3.

[160] Motoharu Hamada et al. "Integration Mapping of piggyBac-Mediated CD19 Chimeric Antigen Receptor T Cells Analyzed By Novel Tagmentation-Assisted PCR". In: *EBioMedicine* 34.nil (2018), pp. 18–26. DOI: 10.1016/j.ebiom.2018.07.008.

[161] Simon Uhse et al. "In Vivo Insertion Pool Sequencing Identifies Virulence Factors in a Complex Fungal-Host Interaction". In: *PLOS Biology* 16.4 (2018), e2005129. DOI: 10.1371/journal.pbio.2005129.

[162] Georgia Giannoukos et al. "Uditas, a Genome Editing Detection Method for Indels and Genome Rearrangements". In: *BMC Genomics* 19.1 (2018), p. 212. DOI: 10.1186/s12864-018-4561-9.

[163] Jonathan L. Schmid-Burgk et al. "Highly Parallel Profiling of Cas9 Variant Specificity". In: *Molecular Cell* 78.4 (2020), 794–800.e8. DOI: 10.1016/j.molcel.2020.02.023.

[164] Ryan M. Mulqueen et al. "High-Content Single-Cell Combinatorial Indexing". In: *Nature Biotechnology* nil.nil (2021), nil. DOI: 10.1038/s41587-021-00962-z.

[165] Andrew L. Goodman et al. "Identifying Genetic Determinants Needed To Establish a Human Gut Symbiont in Its Habitat". In: *Cell Host & Microbe* 6.3 (2009), pp. 279–289. DOI: 10.1016/j.chom.2009.08.003.

[166] Daryl M. Gohl et al. "Large-Scale Mapping of Transposable Element Insertion Sites Using Digital Encoding of Sample Identity". In: *Genetics* 196.3 (2014), pp. 615–623. DOI: 10.1534/genetics.113.159483.

[167] Agnès H. Michel et al. “Functional Mapping of Yeast Genomes By Saturated Transposition”. In: *eLife* 6.nil (2017), nil. DOI: 10.7554/elife.23570.

[168] Sasan Amini et al. “Haplotype-Resolved Whole-Genome Sequencing By Contiguity-Preserving Transposition and Combinatorial Indexing”. In: *Nature Genetics* 46.12 (2014), pp. 1343–1349. DOI: 10.1038/ng.3119.

[169] Andrew Adey et al. “In Vitro, Long-Range Sequence Information for De Novo Genome Assembly Via Transposase Contiguity”. In: *Genome Research* 24.12 (2014), pp. 2041–2049. DOI: 10.1101/gr.178319.114.

[170] Jason D. Buenrostro et al. “Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position”. In: *Nature Methods* 10.12 (2013), pp. 1213–1218. DOI: 10.1038/nmeth.2688.

[171] Sibylle C. Vonesch et al. “Fast and Inexpensive Whole-Genome Sequencing Library Preparation From Intact Yeast Cells”. In: *G3 (Bethesda)* 11.1 (2020), nil. DOI: 10.1093/g3journal/jkaa009.

[172] Stephen G. Oliver. “From DNA Sequence To Biological Function”. In: *Nature* 379.6566 (1996), pp. 597–600. DOI: 10.1038/379597a0.

[173] Ophir Shalem, Neville E. Sanjana, and Feng Zhang. “High-Throughput Functional Genomics Using CRISPR-Cas9”. In: *Nature Reviews Genetics* 16.5 (2015), pp. 299–311. DOI: 10.1038/nrg3899.

[174] Ruth E. Hanna and John G. Doench. “Design and Analysis of CRISPR-Cas Experiments”. In: *Nature Biotechnology* 38.7 (2020), pp. 813–823. DOI: 10.1038/s41587-020-0490-7.

[175] Emily M. LeProust et al. “Synthesis of High-Quality Libraries of Long (150mer) Oligonucleotides By a Novel Depurination Controlled Process”. In: *Nucleic Acids Research* 38.8 (2010), pp. 2522–2540. DOI: 10.1093/nar/gkq163.

[176] Pierre Billon et al. “CRISPR-Mediated Base Editing Enables Efficient Disruption of Eukaryotic Genes Through Induction of Stop Codons”. In: *Molecular Cell* 67.6 (2017), 1068–1079.e4. DOI: 10.1016/j.molcel.2017.08.008.

[177] Louise Clarke and John Carbon. “A Colony Bank Containing Synthetic Col EI Hybrid Plasmids Representative of the Entire E. coli Genome”. In: *Cell* 9.1 (1976), pp. 91–99. DOI: 10.1016/0092-8674(76)90055-6.

[178] Richard Williams et al. “Amplification of Complex Gene Libraries By Emulsion PCR”. In: *Nature Methods* 3.7 (2006), pp. 545–550. DOI: 10.1038/nmeth896.

[179] Pavel A. Zhulidov et al. “Simple cDNA Normalization Using Kamchatka Crab Duplex-Specific Nuclease”. In: *Nucleic Acids Research* 32.3 (2004), 37e–37. DOI: 10.1093/nar/gnh031.

[180] Danos C. Christodoulou et al. “Construction of Normalized RNA-seq Libraries for Next-Generation Sequencing Using the Crab Duplex-Specific Nuclease”. In:

Current Protocols in Molecular Biology. Current Protocols in Molecular Biology. John Wiley & Sons, Inc., 2011, nil. DOI: 10.1002/0471142727.mb0412s94.

[181] W. Gu et al. "Depletion of Abundant Sequences By Hybridization (DASH): Using Cas9 To Remove Unwanted High-Abundance Species in Sequencing Libraries and Molecular Counting Applications". In: *Genome Biology* 17.1 (2016), p. 41. DOI: 10.1186/s13059-016-0904-5.

[182] Jiyeon Kweon et al. "CRISPR/Cas-Based Customization of Pooled CRISPR Libraries". In: *PLOS ONE* 13.6 (2018), e0199473. DOI: 10.1371/journal.pone.0199473.

[183] Cody A. LaBelle et al. "Image-Based Live Cell Sorting". In: *Trends in Biotechnology* 39.6 (2021), pp. 613–623. DOI: 10.1016/j.tibtech.2020.10.006.

[184] Nicholas Hasle et al. "High-throughput, Microscope-based Sorting To Dissect Cellular Heterogeneity". In: *Molecular Systems Biology* 16.6 (2020), nil. DOI: 10.15252/msb.20209442.

[185] George Emanuel, Jeffrey R. Moffitt, and Xiaowei Zhuang. "High-Throughput, Image-Based Screening of Pooled Genetic-Variant Libraries". In: *Nature Methods* 14.12 (2017), pp. 1159–1162. DOI: 10.1038/nmeth.4495.

[186] Michael J. Lawson et al. "In Situ Genotyping of a Pooled Strain Library After Characterizing Complex Phenotypes". In: *Molecular Systems Biology* 13.10 (2017), p. 947. DOI: 10.15252/msb.20177951.

[187] Daniel Camsund et al. "Time-Resolved Imaging-Based CRISPRi Screening". In: *Nature Methods* 17.1 (2019), pp. 86–92. DOI: 10.1038/s41592-019-0629-y.

[188] David Feldman et al. "Optical Pooled Screens in Human Cells". In: *Cell* 179.3 (2019), 787–799.e17. DOI: 10.1016/j.cell.2019.09.016.

[189] Andreas Reicher, Anna Koren, and Stefan Kubicek. "Pooled Protein Tagging, Cellular Imaging, and In Situ Sequencing for Monitoring Drug Action in Real Time". In: *Genome Research* 30.12 (2020), pp. 1846–1855. DOI: 10.1101/gr.261503.120.

[190] Atray Dixit et al. "Perturb-Seq: Dissecting Molecular Circuits With Scalable Single-Cell RNA Profiling of Pooled Genetic Screens". In: *Cell* 167.7 (2016), 1853–1866.e17. DOI: 10.1016/j.cell.2016.11.038.

[191] Britt Adamson et al. "A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response". In: *Cell* 167.7 (2016), 1867–1882.e21. DOI: 10.1016/j.cell.2016.11.048.

[192] Diego Adhemar Jaitin et al. "Dissecting Immune Circuits By Linking CRISPR-Pooled Screens With Single-Cell RNA-Seq". In: *Cell* 167.7 (2016), 1883–1896.e15. DOI: 10.1016/j.cell.2016.11.039.

[193] Paul Datlinger et al. "Pooled CRISPR Screening With Single-Cell Transcriptome Readout". In: *Nature Methods* 14.3 (2017), pp. 297–301. DOI: 10.1038/nmeth.4177.

[194] Qiupeng Zheng et al. “Precise Gene Deletion and Replacement Using the CRISPR/Cas9 System in Human Cells”. In: *BioTechniques* 57.3 (2014), nil. DOI: 10.2144/000114196.

[195] Mandana Arbab et al. “Cloning-Free CRISPR”. In: *Stem Cell Reports* 5.5 (2015), pp. 908–917. DOI: 10.1016/j.stemcr.2015.09.022.

[196] Ryan M. Sheridan and David L. Bentley. “Selectable One-Step PCR-Mediated Integration of a Degron for Rapid Depletion of Endogenous Human Proteins”. In: *BioTechniques* 60.2 (2016), nil. DOI: 10.2144/000114378.

[197] Brian R. Shy et al. “Co-Incident Insertion Enables High Efficiency Genome Engineering in Mouse Embryonic Stem Cells”. In: *Nucleic Acids Research* 44.16 (2016), pp. 7997–8010. DOI: 10.1093/nar/gkw685.

[198] Alexandre Paix et al. “Precision Genome Editing Using Synthesis-Dependent Repair of Cas9-induced DNA Breaks”. In: *Proceedings of the National Academy of Sciences* 114.50 (2017), E10745–E10754. DOI: 10.1073/pnas.1711979114.

[199] Boris V. Skryabin et al. “Pervasive Head-To-Tail Insertions of DNA Templates Mask Desired CRISPR-Cas9-mediated Genome Editing Events”. In: *Science Advances* 6.7 (2020), eaax2941. DOI: 10.1126/sciadv.aax2941.

[200] Kim R. Folger, Kirk Thomas, and Capecchi Mario R. “Nonreciprocal Exchanges of Information Between DNA Duplexes Coinjected Into Mammalian Cell Nuclei.” In: *Molecular and Cellular Biology* 5.1 (1985), pp. 59–69. DOI: 10.1128/mcb.5.1.59.

[201] Jian-Ping Zhang et al. “Efficient Precise Knockin With a Double Cut HDR Donor After CRISPR/Cas9-mediated Double-Stranded DNA Cleavage”. In: *Genome Biology* 18.1 (2017), p. 35. DOI: 10.1186/s13059-017-1164-8.

[202] Ranjith Anand et al. “Rad51-mediated Double-Strand Break Repair and Mismatch Correction of Divergent Substrates”. In: *Nature* 544.7650 (2017), pp. 377–380. DOI: 10.1038/nature22046.

[203] Jens Lykke-Andersen and Eric J. Bennett. “Protecting the Proteome: Eukaryotic Cotranslational Quality Control Pathways”. In: *Journal of Cell Biology* 204.4 (2014), pp. 467–476. DOI: 10.1083/jcb.201311103.

[204] Charles D. Yeh, Christopher D. Richardson, and Jacob E. Corn. “Advances in Genome Editing Through Control of DNA Repair Pathways”. In: *Nature Cell Biology* 21.12 (2019), pp. 1468–1478. DOI: 10.1038/s41556-019-0425-z.

[205] Xiaoling Wang et al. “Unbiased Detection of Off-Target Cleavage By CRISPR-Cas9 and TALENs Using Integrase-Defective Lentiviral Vectors”. In: *Nature Biotechnology* 33.2 (2015), pp. 175–178. DOI: 10.1038/nbt.3127.

[206] Richard L. Frock et al. “Genome-Wide Detection of DNA Double-Stranded Breaks Induced By Engineered Nucleases”. In: *Nature Biotechnology* 33.2 (2014), pp. 179–186. DOI: 10.1038/nbt.3101.

[207] Shengdar Q. Tsai et al. “Guide-Seq Enables Genome-wide Profiling of Off-Target Cleavage By CRISPR-Cas Nucleases”. In: *Nature Biotechnology* 33.2 (2014), pp. 187–197. DOI: 10.1038/nbt.3117.

[208] Hui K. Kim et al. “In Vivo High-Throughput Profiling of CRISPR-Cpf1 Activity”. In: *Nature Methods* 14.2 (2016), pp. 153–159. DOI: 10.1038/nmeth.4104.

[209] Daesik Kim et al. “Genome-wide Analysis Reveals Specificities of Cpf1 Endonucleases in Human Cells”. In: *Nature Biotechnology* 34.8 (2016), pp. 863–868. DOI: 10.1038/nbt.3609.

[210] Benjamin P. Kleinstiver et al. “Genome-wide Specificities of CRISPR-Cas Cpf1 Nucleases in Human Cells”. In: *Nature Biotechnology* 34.8 (2016), pp. 869–874. DOI: 10.1038/nbt.3620.

[211] Winston X. Yan et al. “BLISS Is a Versatile and Quantitative Method for Genome-Wide Profiling of DNA Double-Strand Breaks”. In: *Nature Communications* 8.1 (2017), p. 15058. DOI: 10.1038/ncomms15058.

[212] Shinta Saito, Ryo Maeda, and Noritaka Adachi. “Dual Loss of Human POLQ and LIG4 Abolishes Random Integration”. In: *Nature Communications* 8.1 (2017), p. 16112. DOI: 10.1038/ncomms16112.

[213] Suzanne L. Mansour, Kirk R. Thomas, and Mario R. Capecchi. “Disruption of the Proto-Oncogene Int-2 in Mouse Embryo-Derived Stem Cells: a General Strategy for Targeting Mutations To Non-Selectable Genes”. In: *Nature* 336.6197 (1988), pp. 348–352. DOI: 10.1038/336348a0.

[214] Birgit Koch et al. “Generation and Validation of Homozygous Fluorescent Knock-In Cells Using CRISPR-Cas9 Genome Editing”. In: *Nature Protocols* 13.6 (2018), pp. 1465–1487. DOI: 10.1038/nprot.2018.042.

[215] T. V. Padma. “Covid Vaccines To Reach Poorest Countries in 2023 - Despite Recent Pledges”. In: *Nature* 595.7867 (2021), pp. 342–343. DOI: 10.1038/d41586-021-01762-w.

[216] Catharina Boehme, Emma Hannay, and Madhukar Pai. “Promoting Diagnostics As a Global Good”. In: *Nature Medicine* 27.3 (2021), pp. 367–368. DOI: 10.1038/s41591-020-01215-3.

[217] Vijay J. Gadkar et al. “Real-Time Detection and Monitoring of Loop Mediated Amplification (LAMP) Reaction Using Self-Quenching and De-Quenching Fluorogenic Probes”. In: *Scientific Reports* 8.1 (2018), p. 5548. DOI: 10.1038/s41598-018-23930-1.

[218] Sanchita Bhadra et al. “High-Surety Isothermal Amplification and Detection of SARS-CoV-2”. In: *mSphere* 6.3 (2021), nil. DOI: 10.1128/msphere.00911-20.

[219] Woong Sik Jang et al. “Development of a Multiplex Loop-Mediated Isothermal Amplification (LAMP) Assay for On-Site Diagnosis of SARS CoV-2”. In: *PLOS ONE* 16.3 (2021), e0248042. DOI: 10.1371/journal.pone.0248042.

[220] Kerstin U. Ludwig et al. “LAMP-Seq Enables Sensitive, Multiplexed Covid-19 Diagnostics Using Molecular Barcoding”. In: *Nature Biotechnology* nil.nil (2021), nil. DOI: 10.1038/s41587-021-00966-9.

[221] Andreas Deckert et al. “Effectiveness and Cost-Effectiveness of Four Different Strategies for SARS-CoV-2 Surveillance in the General Population (CoV-Surv Study): a Structured Summary of a Study Protocol for a Cluster-Randomised, Two-Factorial Controlled Trial”. In: *Trials* 22.1 (2021), p. 39. DOI: 10.1186/s13063-020-04982-z.

Supplement

Supplementary Figures

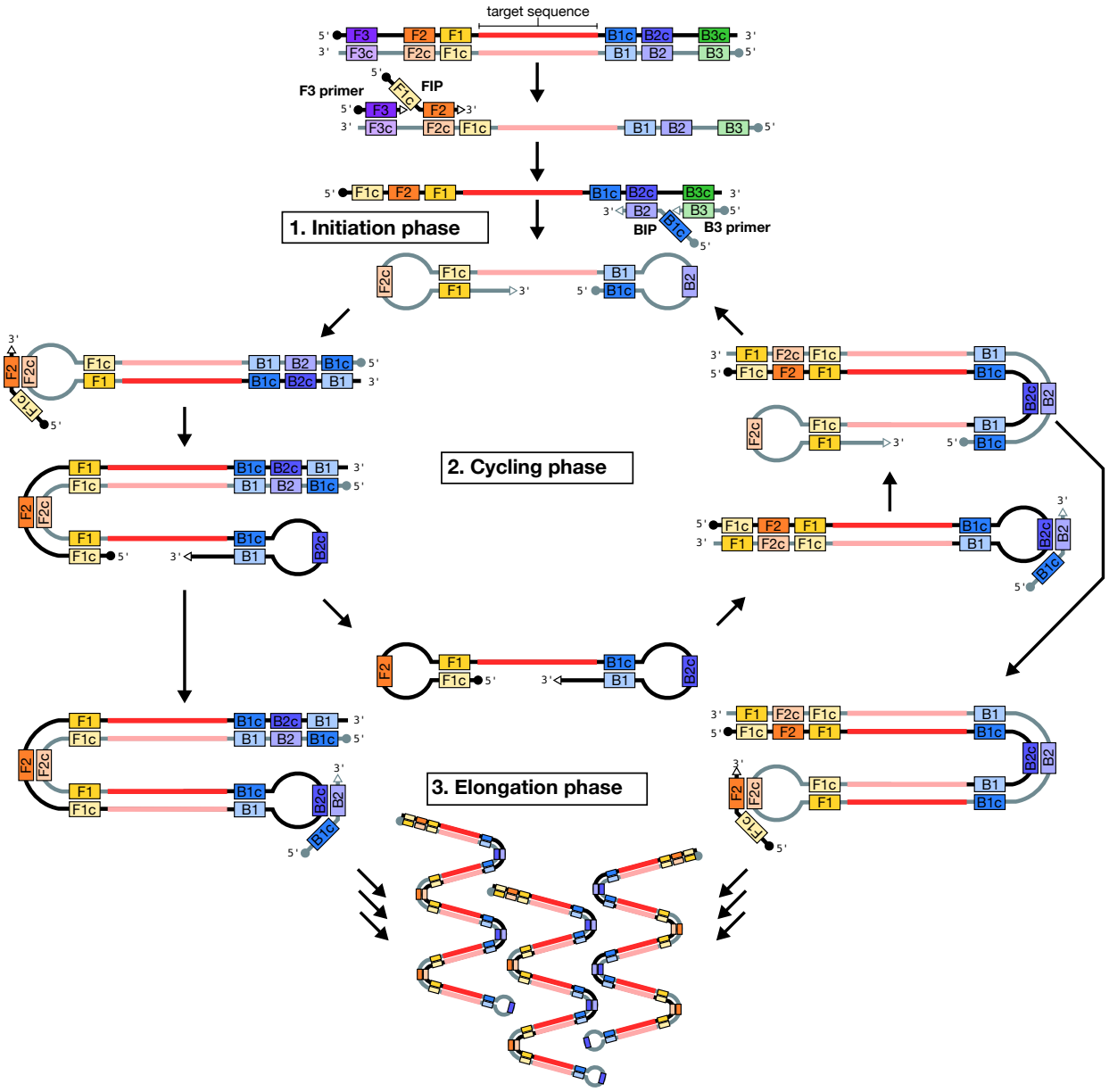


Figure 4.1: Simplified schematic overview of the LAMP reaction. Both DNA strands of the target region are indicated but only one route of the amplification reaction is shown which uses the reverse complement strand (light gray) as template. The target sequence is flanked by primer binding sites for forward primers (F1, F2, F3) and reverse primers (B1, B2, B3). LAMP primers (F3, FIP, B3 and BIP) encompass one or two primer binding sites. Complement primer binding sites are suffixed with a "c". The amplification reaction has three stages. In the initiation phase the template DNA strand is primed by primers FIP, BIP, F3 and B3. The resulting structure serves as template for the cycling phase which also uses the FIP and BIP primers. The reaction kinetics can be improved by including loop-mediated priming by primers LF and LB (not shown). During the elongation phase DNA molecules are further extended leading to long concatemers of the target region. In practice the reaction products are likely more complex and accompanied by side products with different structures.

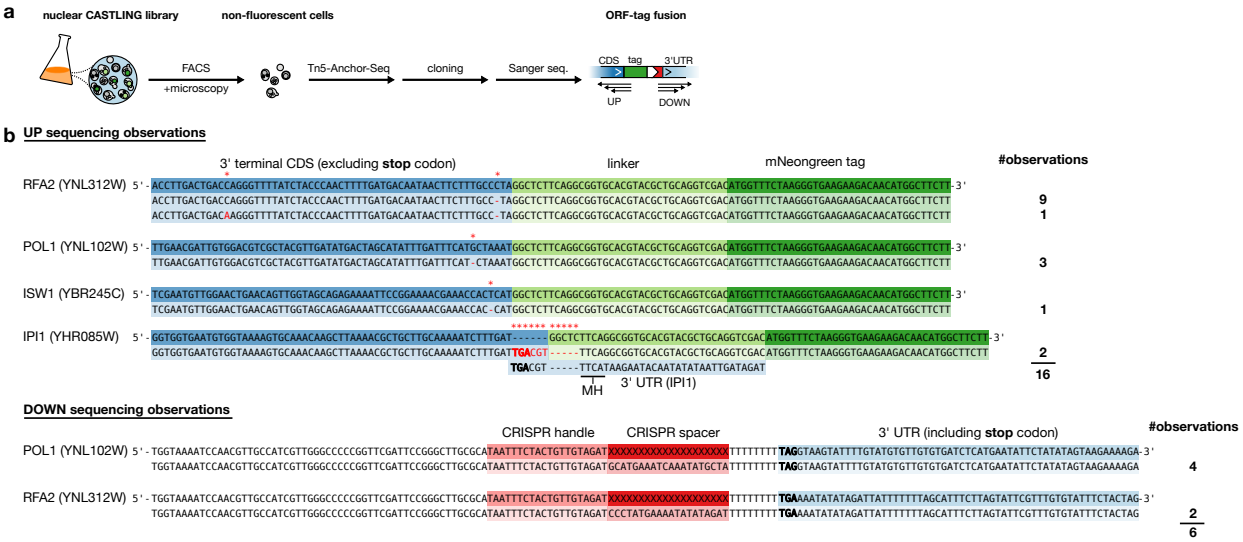


Figure 4.2: Genotyping of non-fluorescent cells of a small nuclear CASTLING library. (a) Non-fluorescent cells were retrieved from small nuclear yeast library 1a using fluorescence activated cell sorting (FACS) and confirmed by fluorescence microscopy. The junction upstream (UP) and downstream (DOWN) of the cassette insertion site was enriched using Tn5-Anchor-Seq resulting amplicons were cloned for Sanger sequencing. **(b)** Observed Sanger sequencing reads (transparent color) is compared to the expected sequence (filled color) of a correctly inserted cassette. Number of sequenced amplicon clones is indicated on the right (#observations). UP amplicons reveal mostly indel mutations (red asterisks) leading to frame-shift mutations in the polypeptide linker between tagged CDS and tag. For one ORF (IPI1) a more complex indel was observed which was most likely mediated by a microhomology (MH) present in the linker and 3'UTR of this gene. DOWN amplicons show that expected and observed sequence match without errors confirming correct integration of the cassette.

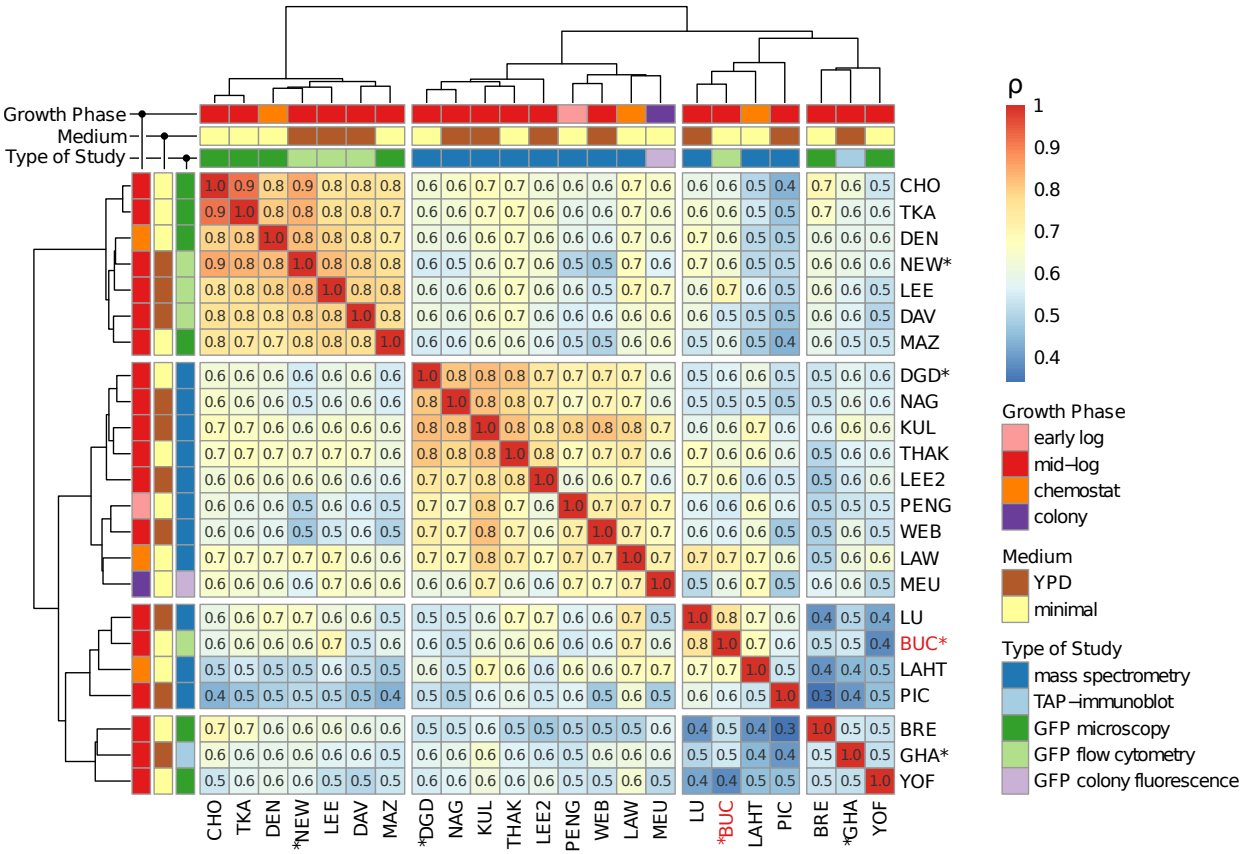


Figure 4.3: Cross-study correlations of protein abundance estimates. Heatmap of spearman correlation coefficients between protein abundance estimation using a CASTLING library (BUC, colored red) and earlier studies compiled in Ho *et al.* [150]. Except for BUC and MEU [134] the respective studies are abbreviated as by Ho *et al.*. Studies used for comparisons depicted in Figure 3.17 are indicated with an asterisk. Furthermore, different study characteristics (growth phase, medium and study type) as compiled in [150] are indicated. *The figure was reproduced from [116] with permission.*

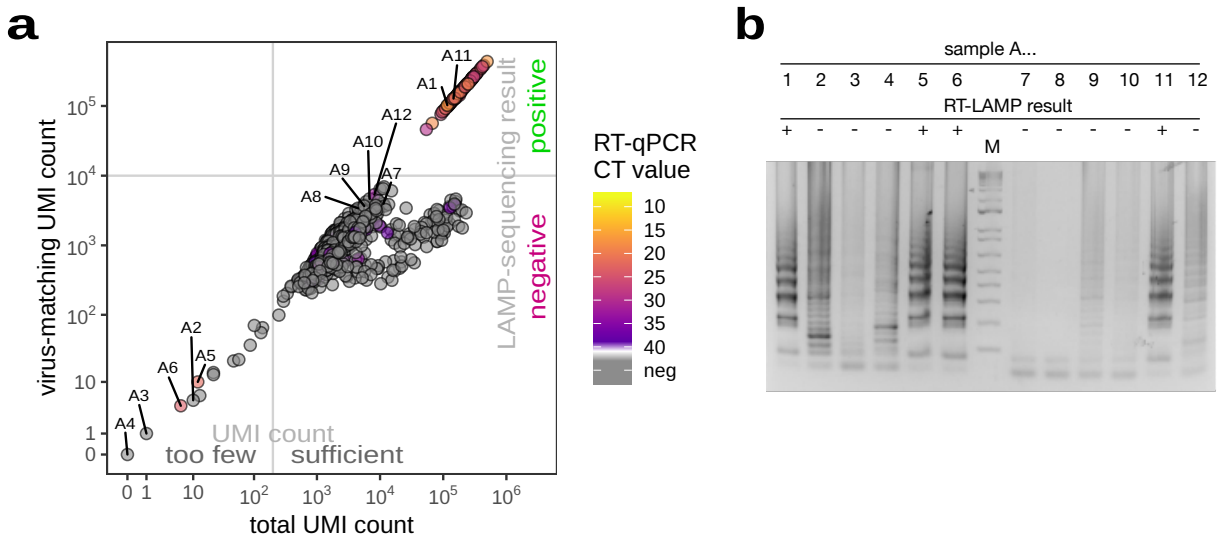


Figure 4.4: Failed reactions in the LAMP-sequencing experiment. **(a)** Same as Figure 3.29 with annotation of one particular row of an array plate from which five samples of the 14 samples with too few reads originate. **(b)** The five samples resulting in too few LAMP-sequencing reads in panel a clearly show DNA content in DNA gel electrophoresis two of which (A5 and A6) also show the expected LAMP banding pattern for positive samples. Contrary, two other samples (A7 and A8) do not seem to contain DNA although they resulted in considerable reads in LAMP-sequencing. It was therefore concluded that the failure to obtain reads for the 14 RT-LAMP samples with too few reads resulted from LAMP-sequencing multiplexing issues (e.g. pipetting errors) instead of lack of material after the RT-LAMP assay. *The figure was reproduced in modified form from [115] with permission.*