

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Expanding the toolbox: Optimization of a synthetic regulatory locus for massively parallel reporter assays (MPRAs).

verfasst von | submitted by

Felix Haslhofer BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt |
Degree programme code as it appears on the
student record sheet:

UA 066 865

Studienrichtung lt. Studienblatt | Degree
programme as it appears on the student
record sheet:

Masterstudium Molecular Biology

Betreut von | Supervisor:

Assoz. Prof. Christa Bucker PhD

Zusammenfassung

Die zellspezifische Expression von Entwicklungsgenen wird meist von mehreren regulatorischen Sequenzen, sogenannten Enhancern, reguliert. Obwohl Enhancer Gene aus großer Distanz regulieren können, nimmt die Genexpression in minimaler regulatorischer Umgebung drastisch ab, wenn die Enhancer-Promoter Distanz erhöht wird. Die systematische Analyse von Genloci zeigt, dass Enhancer in komplexen Abhängigkeiten und Hierarchien mit anderen regulatorischen Elementen stehen. Mit aktuellen Plasmid-basierten Hochdurchsatzmethoden lassen sich diese Wechselwirkungen nur bedingt zeigen. Daher wurde von uns kürzlich ein synthetischer Genlocus konstruiert, der die komplexen Interaktionsmuster von endogenen Loci zeigen kann. Hier soll gezeigt werden, dass starke Enhancer oft mit schwächeren Enhancern synergistisch zusammenarbeiten: Ein starker Enhancer, der 75 kb entfernt von einem Promoter positioniert wird, zeigt Superadditivität mit 3 von 4 schwachen Enhancern, die sich bei 25 kb Distanz befinden. Je stärker ein Enhancer, desto besser kann er die genomische Distanz überwinden und desto weniger profitiert er von der Präsenz eines Enhancers in 75 kb Distanz. Durch die Optimierung mehrerer SOX2- und OCT4-Motive lässt sich zeigen, dass die Fähigkeit, genomische Distanz zu überwinden, eng mit der Stärke eines Enhancers zusammenhängt. Zuletzt wurde basierend auf einem Kassettenaustauschverfahren (RMCE) ein optimierter synthetischer Locus konstruiert, der eine Hochdurchsatz-Anwendung erlaubt. Damit lassen sich zukünftig tausende von regulatorischen Elementen testen, die Aufschluss darüber geben sollen, wie die Fähigkeit, Distanzen zu überwinden und mit anderen Elementen in Wechselwirkung zu stehen, in der DNA-Sequenz verankert ist.

Abstract

Expression of cell-type specific developmental genes is typically regulated by multiple enhancers, often located at large genomic distances from their target genes. Increasing enhancer-promoter distance at an ectopic location leads to a sharp drop of gene expression. However, the same elements often work from significantly larger distances in their endogenous context. Genetic dissections of individual loci showed that enhancers are not independent units but display intricate hierarchies and interdependencies. Current plasmid-based high-throughput analyses of enhancer-enhancer combinations do not display the diversity of interdependencies seen in genetic dissections. To address these drawbacks, we have previously built a synthetic locus that allows the bottom-up construction of regulatory landscapes. Here I show that synergy between strong and weak enhancers could be a wide-spread phenomenon: When fixing a strong enhancer at 75 kb from a minimal promoter, 3 of 4 weak enhancers at 25 kb show synergy. Additionally, there is a trend for enhancers with increased strength to act over larger distances and display less synergy if placed at 25 kb. Furthermore, by optimizing individual SOX2/OCT4 motifs I could show that enhancer strength and distance dependency are interconnected. Finally, I developed a more efficient synthetic locus (2.0), based on the large-serine recombinase Bxb1 and a cassette-exchange approach (RMCE). To optimize our enhancer integration system, I constructed an inducible Bxb1-expressing cell line containing an RMCE-based landing pad at 1.5 kb, suitable for high-throughput integrations. By integrating thousands of elements, we hope to understand the rules governing distance dependency and synergy.

Acknowledgements

I want to start by thanking both Christa and Songjie. Christa has been an amazing supervisor and I'm very lucky to have gotten the opportunity to work in her lab. I really enjoyed all our discussions and your passion for science (and enhancers!) is truly inspiring. I'm also extremely grateful for Songjie, the "cloning-god". Most of the work in Chapter I and II were directly supervised by him and he is one of the kindest and hardworking people I know. The enhancer-team, and the whole lab really, would not be the same without you. I'm very lucky to have you as a co-worker, and more importantly, as a friend.

I'm also extremely grateful for Maria, always creating such a happy and delightful atmosphere. I'm amazed at how much work you get done next to supervising students, thousands of hobbies and travelling. I wonder if you ever sleep sometimes. Also, Chapter II was done in collaboration with Maria.

I'm thankful for Isa and Tamara. Isa must be in competition with Maria in creating the most joyful atmosphere in the lab. She is a big part of why the lab feels more like a family, rather than a workplace and I'm proud that you got to write your first (and hopefully not last) book during the time I was here. I'm glad that through her project also Tamara could join our lab. She is clearly the most creative person in the lab, responsible for many interestingly shaped gastruloid pictures, alien group pictures and most importantly, cards that allow me to end any cat topic and switch to an enhancer topic during lunchtime.

A special thanks also to Marie, the main reason why the lab is running so smoothly. You have been nothing but supportive throughout the year and you are truly a great technician, manager and scientist.

I would also like to thank the *Leeboratory* including Martin, Anna, Kathi, Luis, Mattia, Michelle, Michi B., Michi O., Moritz and Tania.

Additionally, many thanks to the FACS facility (Thomas, Kitti and Endre) for conducting many of the sorting experiments.

Most importantly, I want to thank my parents Gabi and Manfred, which not only supported me throughout my whole studies but really throughout my whole life. While it was not always clear I would make it to university (or even Matura), you always supported every decision that I made, and nothing would have been possible without you. I don't take it for granted.

I also want to thank the second half of my family, Rene and Simone. Willingly or not, Rene has always been a role model in my life and a major reason why I decided to move to Vienna to pursue my studies. I'm glad that you brought Simone into our family.

Table of contents

Zusammenfassung	1
Abstract	2
Acknowledgements.....	3
Table of contents	4
Enhancers and Their Role in Evolution and Disease.....	7
Chapter I.....	9
Introduction	9
Enhancers: Overcoming Genomic Distance.....	9
A Primer on Enhancer Clusters: Lessons from Two Well-Studied Regulatory Networks.....	11
High-Throughput Assays for Enhancer-Enhancer and Enhancer-Promoter Combinations.....	13
A Versatile Synthetic Locus to Test Enhancer Interdependencies.	14
Tyrosine Recombinases and Their Use in Synthetic Biology.....	15
Engineering Directionality of loxP Sites.....	16
Orthogonal Recombination Systems by Mutating the Spacer.....	17
Aims.	17
Results.....	18
Selection of Weak Enhancers.	18
Intrinsic Strength and Ability to Overcome Genomic Distance Does Not Correlate for All Weak Enhancers.	20
Integration of Our Previous Dataset.....	22
Most Weak Enhancers Display Synergy with a Strong Enhancer.....	24
Discussion.....	27
Limitations.	30
Chapter II.....	31
Introduction.	31
Aims.	31
Results.....	32
Optimizing Motif Affinity Increases Enhancer Strength While Maintaining Sensitivity to Genomic Distance.....	32
Integration of an Old Map4k3 E1 Cell Line.....	36
Discussion.....	36
Chapter III.....	38

Introduction.	38
Limitations of the Enhancer Integration System.	38
The Principle of Recombination-Mediated Cassette Exchange.	38
Regulating Directionality of Recombination via Serine Recombinases.	39
Aims.	41
Results.	42
Generating a Doxycycline-Inducible Bxb1 Cell Line (iBxb1).	42
Generation of a Landing Pad Capable of Monitoring and Enriching Clones After RMCE.	43
Integration of a Landing Pad into iBxb1 for Testing RMCE Efficiencies.	45
Construction of a Targeting Vector for RMCE.	47
Cassette Exchange Is Efficient in Non-Selective Conditions Despite Unexpected Rearrangements.	48
Negative Selection Enriches Successful Recombinants.	49
Discussion.	53
Material and Methods.	55
ESC Maintenance.	55
Transfection.	55
PCR.	55
Restriction Digest.	55
Gibson Assembly.	56
Annealing Oligos.	56
Ligation.	56
Transformation of Bacteria.	56
Analysis of Flow Cytometry Data.	57
Cre/Flpo-Mediated Integration of Enhancers at Synthetic Locus.	57
Site Directed Mutagenesis.	57
Cloning of Flpo-IRFP.	58
Generation of Doxycycline-Inducible Bxb1 Cell Line (iBxb1).	58
Cloning of GFP Cassettes.	59
Cloning of Bxb1 Landing Pads.	60
Integration of an EGFP-T2A-TK LP into iBxb1-C4.	61
Cloning of Bxb1-attB Targeting Vector for RMCE.	61
Integrating eNanog via RMCE at 1.5 kb to Assess Exchange Efficiency.	61
Ganciclovir Selection of Recombinants.	62

Motif Analysis via JASPAR.	63
References	67
Supplementary Figures.....	77
Supplementary Notes	87
Supplementary Note 1.....	87
Single Enhancer Normalization.....	87
Comparing Enhancer Strength Across Datasets with Different Conditions.....	87
Dual Enhancer Normalization.	88
Extrapolating Values for Linear Regression.	88

Enhancers and Their Role in Evolution and Disease.

Most common traits and diseases have a large heritable component (Maher, 2008). Through genome-wide association studies (GWAS) the heritable component can be pinpointed to individual loci, albeit still with moderate success due to linkage disequilibrium (Uffelmann et al., 2021). The risk variants nominated by GWAS are predominantly overlapping with DNase-hypersensitivity sites and only a small number corresponds to promoters or exons (Maurano et al., 2012). Indeed, most risk variants overlap with cis-regulatory elements (CREs) or enhancers, respectively (Nasser et al., 2021). While an individual gene is expressed in many different cell states, enhancers are very specific (Nasser et al., 2021). In this way, through risk variants in enhancers, expression of a particular gene can be affected in very specific cell states, leading to a variety of diseases associated with the same gene. Additionally, many risk factors are also found in healthy individuals and much of the heritability of non-mendelian diseases and traits remain unexplained, which has been referred to as “hidden heritability” (Maher, 2008; Mohammed et al., 2024). This has been beautifully demonstrated by work on an enhancer regulating PKDCC expression, a protein involved in shaping the jaw (Mohammed et al., 2024). An “A to T” transversion in the enhancer is associated both with normal jaw diversity (a more protruding jaw) in humans but also with the pathological non-syndromic cleft lip/palate (nsCL/P). The SNP leads to disruption of a transcription factor (TF) motif and decreases PKDCC expression, leading to a protruding jaw. However, it is not sufficient for causing the clefting, but additional mutations or environmental stimuli during development are required to trigger the disease (Mohammed et al., 2024).

Sequence diversity in enhancers can not only help to understand phenotypical variation in humans but can shed light on evolution. While human and chimpanzee genomes are very similar, most of the differences map to cis-regulatory regions (Prescott et al., 2015; Suntsova and Buzdin, 2020). This has been explored by a decade of work on cranial neural crest cells (CNCCs), which are responsible for facial shape (S. Kim et al., 2024; Prescott et al., 2015). There is a small number of less-conserved, species-specific enhancer variants, whose activity correlates in both chimpanzee and human trans-regulatory environments, confirming that the cis-regulatory sequence is indeed responsible for the interspecies differences (Prescott et al., 2015). These enhancers are enriched with the coordinator motif and the strengthening or weakening of its TF-binding capacity correlates with gain or loss of typical enhancers marks as well as nearby gene expression of CNCC-specific genes (Prescott et al., 2015). Indeed, coordinator motif-containing enhancers (as well as enhancers that control expression of the TFs bound to it) do not only drive interspecies differences, but also variability between humans (S. Kim et al., 2024).

While SNPs and the associated changes in traits and diseases are probably the most common mutations, structural rearrangements such as deletions, duplications, inversion or translocation have been implicated in enhancer mistargeting (Friedman et al., 2024).

Because regulatory landscapes can span large genomic distances it is very likely that structural rearrangements can lead to aberrant enhancer-promoter interactions. Indeed, at the *Wnt6/Ihh/Epha4/Pax3* locus, various rearrangements lead to loss of TAD boundaries and enhancer mistargeting, leading to limb malformations (Lupiáñez et al., 2015). While this is mainly caused by a failure of insulation, an inversion or translocation on chromosome 3 places a *Gata2* enhancer in vicinity to a proto-oncogene, driving its expression from up to 200 kb away, and consequently leading to leukemia (Gröschel et al., 2014).

In conclusion, while changing motif affinities through SNPs is clearly important to regulate dosage and specificity of gene expression and consequently drive both evolution and disease, many genes are regulated by multiple regulatory elements, which additionally have the ability to communicate over large genomic distances. Understanding the complexity of these regulatory landscapes will be necessary to understanding the emergence of complex traits, disease and evolution.

Chapter I

Introduction

Enhancers: Overcoming Genomic Distance.

According to its very first definition an enhancer is a regulatory element with the ability to activate gene expression from a distance (Banerji et al., 1981). Despite substantial flexibility in linear enhancer-promoter distance, communication requires 3D proximity, functionally demonstrated by increased expression via forced looping (Deng et al., 2014; Weintraub et al., 2017) or decreased expression via insulating (Chakraborty et al., 2023; Tsang et al., 2023). While there are many elements working from very large distances, referred to as ultra-long-range enhancers (Kane et al., 2022; Long et al., 2020), most of these elements are found within a 100 kb radius from the gene they regulate (Gschwind et al., 2023). DNA is a polymer, therefore contact probability and distance between two loci scale according to a power-law (Pombo and Nicodemi, 2014). Moving further away from the promoter, the contact probability is reduced drastically in a non-linear fashion. Naively, one would assume that the influence of an enhancer on its promoter would scale accordingly. Surprisingly, also the contact probability and expression relate non-linearly (Jensen et al., 2024; Rinzema et al., 2022; Zuin et al., 2022). One can imagine that similar to TFs cooperating to achieve high levels of expression, subsequent promoter-enhancer contacts might cooperate in a timely fashion (Xiao et al., 2021; Zuin et al., 2022). Indeed, a Hill-function, famous for describing cooperative binding of oxygen to hemoglobin (Goutelle et al., 2008), fitted the data well for the *Sox2* promoter and the core SCR at an ectopic location (Zuin et al., 2022). In this way, through cooperativity, enhancers initially appear to be buffered and then dramatically sensitive towards changes in contact probability. Buffering has also been noted for the *Fgf5* gene, while the expression of *Prdm8* is reduced exponentially with contact probability (Jensen et al., 2024). While the exact mechanisms of enhancer-promoter communication at intermediate distances is still unclear, promoter dynamics do not seem to be exclusively responsible, as swapping the promoters of *Fgf5* and *Prdm8* does not change their different responses (Jensen et al., 2024). Additionally, a “kiss-and-run” model, leaving a transient “memory” at the promoter is not very likely (Karr et al., 2022). However, there is a case to be made for influences of the genomic “background”, usually describing unknown influences from the genomic environment at which the gene is located. For example, even though the *H19* enhancer and the SCR are similarly sized and at comparable distances to their target promoters, when switching the enhancers, *H19* fails to upregulate *Sox2*, while the SCR outperforms *H19* at its endogenous position, probably due to additional synergizing elements (Ordoñez et al., 2023). Additionally, when a classical enhancer is placed into a “neutral” genomic background it shapes its environment both in 3D organization but also by activating cryptic regulatory sequences (Georgiades et al., 2023).

Moving around different elements in different genomic environments precludes the disentanglement of the relative contribution of each factor involved. To tackle this

problem, we have previously constructed a synthetic regulatory locus in which the distance behavior of different enhancers is compared at the exact same locus, canceling out the effect of the genomic “background” and potential differences in scaling due to chromatin properties (Pombo and Nicodemi, 2014; Thomas et al., 2023). Because individual enhancers are thought to influence 3D chromatin structure only minorly, changes in linear distance should be proportional to changes in contact probability (Jensen et al., 2024; Zuin et al., 2022). Our first enhancer integrations at the synthetic locus have already indicated that enhancer strength influences both the absolute transcription levels at each distance, as well as the relative drop in expression (Thomas et al., 2023). In other words, enhancer-driven expression scales uniquely for each element, with a tendency of strong enhancers to barely lose any activity within 25 kb. This contrasts with a previous publication that predicted that the strength of an enhancer would only affect the absolute expression levels but the loss of expression over distance scales equally (Zuin et al., 2022).

While enhancers may or may not be buffered at close distances, at minimal regulatory loci even the strongest enhancers lose most of their activity beyond 50 kb and contact probability becomes negligible (Rinzema et al., 2022; Zuin et al., 2022). This raises the question on how these elements overcome distance at their endogenous loci, where they are sometimes located much further away. CTCF and cohesin are able to keep two distant loci in close proximity as well as increase the duration of contacts by reducing mobility (Mach et al., 2022). Indeed, long-range enhancers use cohesin to overcome large genomic distances (Hung et al., 2024; Kane et al., 2022; M. Kim et al., 2024). At the endogenous *Shh* locus a 100 kb enhancer does not require cohesin, while the enhancers beyond 400 kb do (Kane et al., 2022). Similarly, in a minimal regulatory environment cohesin contributes to enhancer-promoter proximity especially above 50 kb (Rinzema et al., 2022). However, while TADs are indeed dependent on cohesin and CTCF, 80% of enhancer-promoter loops are unaltered after acute depletion of either cohesin or CTCF, demonstrating that cohesin-mediated loop extrusion plays a minor role in regulating steady-state gene expression (Hsieh et al., 2022). Interestingly, while the β -globin μ LCR enhancer has been shown to affect 3D organization even after cohesin depletion and at close distances (Rinzema et al., 2022), two other minimal systems report negligible effects on contact probability between enhancer presence and absence (Jensen et al., 2024; Zuin et al., 2022). Also, at the very complex endogenous *Sox2* locus, where the SCR is responsible for almost 100% of transcriptional activity at the promoter, the heightened contact probability between the SCR and its promoter is largely unaffected upon deletion of the SCR, essentially uncoupling transcription from contact (Taylor et al., 2022). Confusingly, cohesin seems to be required for *Sox2*-SCR contacts, but expression levels are only reduced marginally upon cohesin depletion, once more questioning the necessity of contacts (Aljahani et al., 2022).

Clearly different factors, absent in the minimal synthetic loci, appear to help the SCR to communicate with its promoter. Interestingly, even though the SCR is fully necessary for endogenous *Sox2* expression, it does not work alone to achieve those levels (Brosh et al., 2023). The SCR synergizes with a plethora of regulatory elements distributed along the 100 kb stretch (Brosh et al., 2023). Similarly, the first enhancer integrations in our

synthetic locus showed that two enhancers can synergize to overcome the loss of activity caused by distance, potentially being the missing link to how enhancers maintain communication with their promoters at intermediate distances (Thomas et al., 2023).

A Primer on Enhancer Clusters: Lessons from Two Well-Studied Regulatory Networks.

Many genes that are essential for the identity of a specific cell state are regulated by multiple enhancers, often organized in clusters (Gschwind et al., 2023). The *Sox2* gene encodes a transcription factor necessary for pluripotency of ESCs (Zhang, 2014), while the α -globin genes necessary for oxygen transport are expressed in erythrocytes (Hardison, 2012). The α -globin expression is regulated by a super-enhancer consisting of 5 closely spaced elements (R1, R2, R3, Rm, R4) around 50-30 kb away from the two α -globin promoters, forming a ~65 kb sub-TAD that is specific to erythrocytes (Reviewed in Kassouf et al., 2023). *Sox2* expression in mouse ESCs is primarily regulated by a locus-control region (LCR or SCR) that is ~100 kb from the *Sox2* promoter (Li et al., 2014). *Sox2* and the SCR form a 100 kb TAD-like structure (Chakraborty et al., 2023). Between promoter and SCR are numerous additional regulatory elements that form an intricate network (Brosh et al., 2023). Both loci have been well described and most of what we know about super-enhancers originate from these two examples (Brosh et al., 2023; Kassouf et al., 2023). *Sox2* is especially useful, because it is expressed in ESCs, a cell-state easily maintained in tissue-culture. To study the α -globin locus, one needs to differentiate ESCs into erythrocytes, which could lead to differences between laboratories, based on the protocols they use. However, through differentiation the activation kinetics of the enhancers can be studied, which is not feasible for ESCs (Stolper et al., 2023). Additionally, while *Sox2* is necessary for maintaining the cell-state, its depletion leading to secondary effects, the depletion of the α -globin genes does not affect global gene expression (Blayney et al., 2023). Despite the largely different cell state, very similar rules have emerged. Both loci are largely indifferent to cohesin depletion and CTCF sites delimiting the borders of the (sub-) TADs are mainly required to block enhancers from affecting nearby genes (Aljahani et al., 2022; Chakraborty et al., 2023; Stolper et al., 2023; Tsang et al., 2023). More specifically, cohesin is required to upregulate α -globin expression during differentiation, but not for its maintenance (Stolper et al., 2023). At both loci, CTCF sites readily block enhancer-promoter communication when placed between promoter and enhancers, especially when the N-terminus is oriented towards the enhancers, consistent with cohesin primary extruding from enhancers (Tsang et al., 2023).

The α -globin locus serves as a cautionary tale. The advent of genome editing tools enabled knockouts of single regulatory elements to test their necessity. Determining sufficiency by testing regulatory elements in their endogenous locus without contribution of other regulatory elements remained out of scope. In this manner, R1 and R2 were found to be necessary, with minor contributions of R4 (Hay et al., 2016). Because the individual contributions summed up to 100%, their activity was speculated to combine additively, hence independent from one another (Hay et al., 2016). In recent years as large-scale bottom-up reconstructions of large loci became feasible, a test of sufficiency revealed a completely different picture (Blayney et al., 2023). While indeed R3, Rm and

R4 did not encode much activity on their own, R1 and R2 were not particularly sufficient either, driving only around 10% of alpha-globin expression each. Combining enhancers one by one, and thereby re-building the super-enhancer from bottom up, showed that the enhancers depend on one another and act synergistically. While R3, Rm and R4 were neither necessary nor sufficient, they were crucial for supporting¹ the classical enhancers R1 and R2. Additionally, R3, Rm and R4 are partially redundant, explaining why individual deletion did not render them particularly necessary. These non-classical enhancers have been coined “facilitator elements”, showing no intrinsic activity in luciferase-based assays or STARR-Seq, but being accessible and containing typical chromatin marks such as H3K4me1 and H3K27Ac (Blayney et al., 2023).

Similar relationships have been reported from the *Sox2* locus. DHS19 and DHS20 neither have intrinsic activity nor are they necessary (Brosh et al., 2023). However, when all the elements from the LCR are replaced by DHS19 or/and DHS20 together with a classical enhancer (DHS2526), the output is vastly stronger than the sum of their activities (Brosh et al., 2023). Synergistic interactions remain an intriguing problem, as one can never exclude the contribution of additional unknown elements even if the expression is completely abrogated, as their supporting activity might be attributed to the classical enhancer that was just removed (Fig. 1).

That the relative distance of these elements is of importance is exemplified by the α -globin SE. While R4 is the strongest “facilitator” and R3 is the weakest, replacing R4 by R3 rescues almost all the activity of R4 (Blayney et al., 2023). Similarly, at the *Myc* locus, enhancers interact more synergistically if they are closer together (Gschwind et al., 2023). This also includes H3K27Ac levels, which are synergistic at distances below 10 kb and mostly additive beyond that (Gschwind et al., 2023).

This shows, while individual TFs cooperate do modulate motif binding, groups of TFs (enhancers) regularly cooperate with other distant groups, via an unknown mechanism. It is very likely that just as motif affinity is defined by its context (Jindal and Farley, 2021), the influence of an enhancer on a distal promoter is defined by a specific grammar. Besides promoter expression, synergy can affect tissue-specificity and directionality. While individual enhancers might be broadly active, “AND gate logic” could confine expression to the intersecting cell-states (Bell et al., 2024; Crocker et al., 2015; Farley et al., 2015; Kim and Wysocka, 2023). Additionally, while individual enhancers work orientation-independent, super-enhancers show directionality as an emergent property (Kassouf et al., 2022). Synergism could also provide a larger playground in which enhancers can evolve by contributing to overcoming genomic distance. As an example, *Sox2* has proximal enhancers that are broadly active and encode around 17% and 22% of activity in mESCs, respectively, but are not sufficient to drive *Sox2* expression (Brosh et al., 2023; Ordoñez et al., 2023). This allows distal enhancers, such as the SCR at 100 kb or the neural enhancer at 400 kb, that could never reach significant levels of expression on their own, to influence *Sox2* expression significantly (Brosh et al., 2023;

¹ The directionality of this relationship, which element supports which element, is unclear.

Chakraborty et al., 2023). Without synergism, the space of elements to evolve would be confined to a tight space around the gene.

To dissect these hierarchies, a bottom-up approach, synthetically reconstructing regulatory networks, is extremely useful. Despite their usefulness, deciphering the determinants of enhancer interaction can be challenging, due to vastly different genomic contexts, distances between elements as well as number and types of elements involved. Ideally, a limited number of enhancers need to be placed into a uniform chromatin background² at precisely defined distances to the same promoter (Thomas et al., 2023).

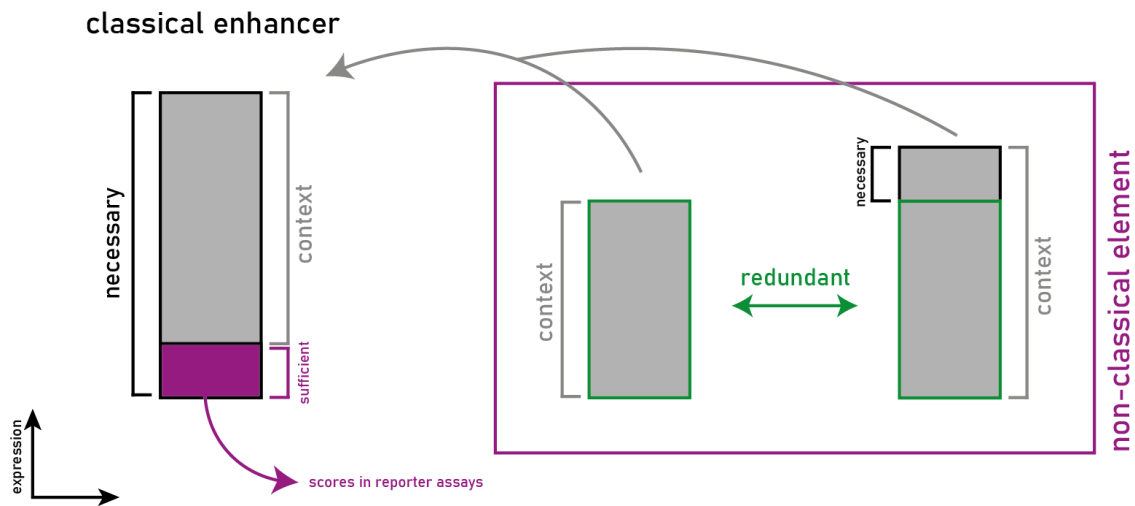


Figure 1: Classical vs. non-classical elements (based on Kassouf et al., 2023). Classical enhancers are sufficient to activate gene expression. A test of necessity might reveal a bigger contribution, due to supporting elements. These non-classical enhancers are not sufficient to activate gene expression on their own. They can also be partially redundant.

High-Throughput Assays for Enhancer-Enhancer and Enhancer-Promoter Combinations.

Massively parallel reporter assays (MPRAs) are a great resource to identify classical enhancers genome-wide based on their ability to activate a minimal promoter (Arnold et al., 2013; Peng et al., 2020), analyze sequence determinants of enhancers (Melnikov et al., 2012; Patwardhan et al., 2012; Sahu et al., 2022) and promoters (Patwardhan et al., 2009; Van Arensbergen et al., 2017) as well as promoter-enhancer compatibility (Bergman et al., 2022; Martinez-Ara et al., 2022; Zabidi et al., 2015). Although they exist in many different variants, they are remarkably similar, even when comparing semi-random integrations vs. episomal assays (Klein et al., 2020). They allow the analysis of hundreds of thousands of variants but are devoid of consistent chromatin context and distance between elements and do not allow the detection of non-classical enhancers, such as facilitator elements. Enhancer strength as measured by reporter assays also does not correlate well with known enhancer marks such as H3K27Ac and accessibility (Peng et al., 2020). For mouse and human, promoters were found to be broadly compatible with enhancers (Bergman et al., 2022; Martinez-Ara et al., 2022). While the

² Ideally in “neutral” chromatin.

intrinsic strength of both enhancer-promoter pairs and enhancer-enhancer pairs were found to scale multiplicatively, another MPRA found that most enhancers combine additively (Bergman et al., 2022; Loubiere et al., 2023; Martinez-Ara et al., 2023, 2022). This clearly differs from the complex interactions and interdependencies seen at endogenous loci (Blayney et al., 2023; Brosh et al., 2023).

A Versatile Synthetic Locus to Test Enhancer Interdependencies.

While MPRA is useful to compare enhancers on a large scale in a similar context, they do not display the complex intricacies seen when reconstructing endogenous regulatory environments, possibly due to the absence of distance effects (Loubiere et al., 2023; Martinez-Ara et al., 2023). Conversely, the individual contributions of elements in complex regulatory environments are near impossible to dissect. To combine the best of both worlds, our laboratory has recently constructed a synthetic locus, consisting of a minimal promoter driving mCherry and three orthogonal loxP sites at different distances (Fig. 2) (Thomas et al., 2023). Using Cre-mediated insertion of a targeting vector, one or multiple enhancers can be precisely targeted to any one of three positions, based on the spacer-variant (WT, 2272, m2) and the excision is prevented through the lox66/71-system. The backbone is then removed via orthogonal Flpo/FRT systems. The effect on the minimal promoter can be read out by flow cytometry. The locus was constructed in the v6.5 cell line, a hybrid mESC strain (C57BL/6 and 129/sv) (Rideout et al., 2000). In that way the alleles can be distinguished and reads from sequencing-based methods like ChIP-Seq, or 3C-derivatives can be specifically assigned to the allele carrying the synthetic locus. Similarly, ectopic enhancers are amplified from the CAST/EiJ genome, to distinguish ectopic enhancer variants from endogenous enhancer variants. The synthetic locus is positioned within the β -globin genes, which are not active in ESCs and do not carry activating or repressive chromatin marks (Lienert et al., 2011; Thomas et al., 2023). The locus is also devoid of specific chromatin interactions (Hsieh et al., 2020). In summary, this locus allows us to compare the effect of single or combinations of enhancers on a minimal promoter at controllable genomic distances and in a constant genomic background without confounding influences from neighboring elements, chromatin modifications or chromatin topology.

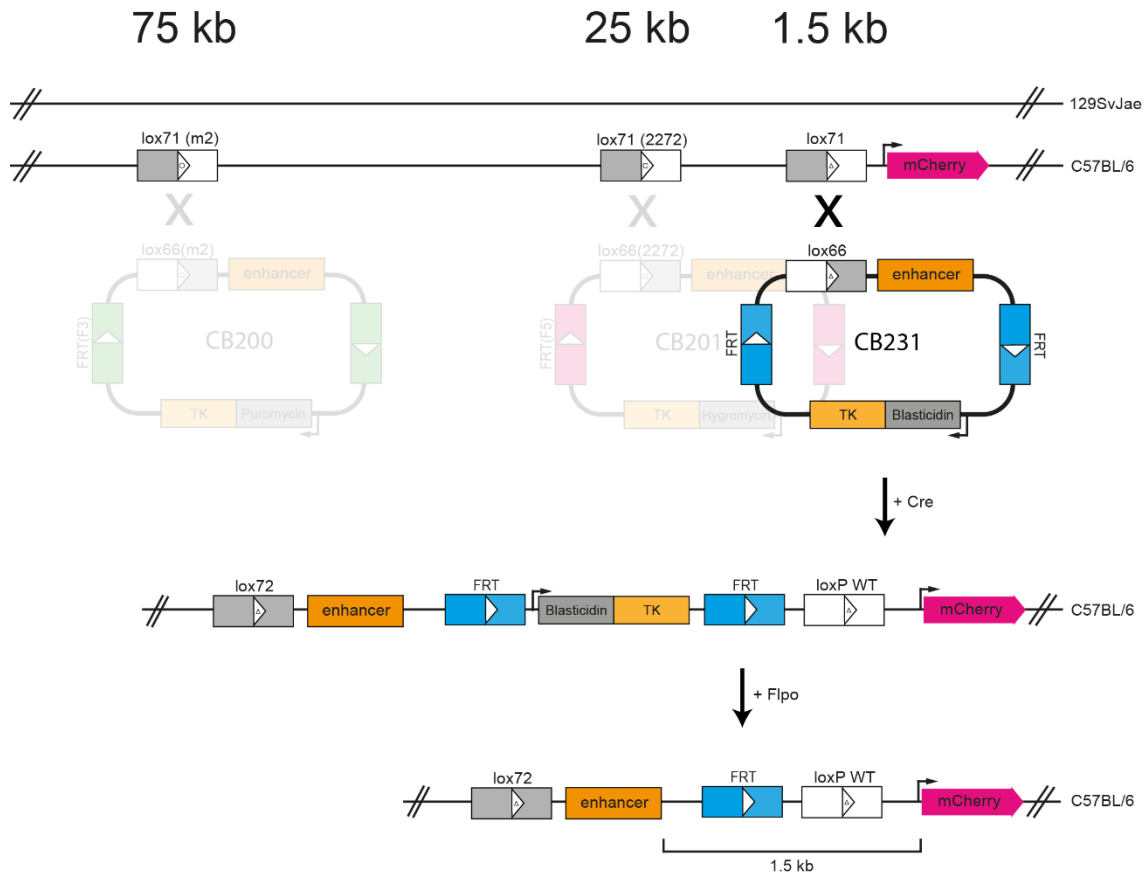


Figure 2: A synthetic locus for reconstruction of regulatory environments to measure enhancer distance dependency and enhancer cooperativity. The locus is located within chromosome 7 of the C57BL/6 allele, which can be distinguished from the 129SvJae allele via SNPs. Three loxP variants are placed at different distances from a minimal promoter driving mCherry expression. Three vectors target the enhancer to the three locations. The backbone of the integrated targeting vector is removed with the Fip/FRT system.

Tyrosine Recombinases and Their Use in Synthetic Biology.

Generating synthetic cell lines requires tools to precisely integrate the sequence of interest at the desired location in the genome. Our synthetic locus is largely based on the action of the tyrosine recombinases Cre and Fip. Chapter III will deal with the optimization of this process.

Integration of a sequence at a specific locus occurs usually in one of two ways, either utilizing DNA-repair systems and homologous recombination or through site-specific recombination (SSRs) (Zheng et al., 2024). While the CRISPR/Cas9 system has evolved to induce double-strand breaks, SSRs are specialists in integrating, excising or inverting sequences. Fip is naturally encoded on the 2 μ plasmid of *Saccharomyces cerevisiae*, which is a small (6.3 kb) selfish, multicopy plasmid (Chan et al., 2013). Its function is to invert the replication fork to convert bidirectional replication into rolling-circle replication when the plasmid copy-number is low and resolve the plasmid multimers into monomers (Chan et al., 2013). Cre circularizes the linear P1-phage genome when undergoing lysogeny and resolves plasmid-dimers that would interfere with the equipartition of the low-copy number phage-plasmid (Austin et al., 1981; Sternberg and Hamilton, 1981).

While Cre was just codon-optimized (Shimshek et al., 2002), Flp has been extensively evolved and optimized for mammalian cells and referred to as Flpo (Buchholz, 1996; Buchholz et al., 1998; Raymond and Soriano, 2007; Schaft et al., 2001). SSRs are highly efficient, but are limited by their recombination sites, which need to be placed beforehand.

Tyrosine recombinases differ from serine recombinases (Chapter III) mechanistically. While serine recombinases produce double-strand break-intermediates, tyrosine recombinases work via Holliday junctions. The mechanism of Flp and Cre has been reviewed extensively and should be briefly summarized here (Chan et al., 2013; Chen, 2000; Chen and Rice, 2003). Both Flp and Cre recognize a recombination site composed of two 13-bp inverted repeats and a central 8-bp spacer, called the FRT³ and loxP site⁴, respectively. Since the binding sites are inverted repeats, the spacer is the only sequence conferring directionality. If the spacer is symmetrical, directionality is lost (Senecoff and Cox, 1986). A recombination event starts with each recombination site being bound by a dimer (Chan et al., 2013; Chen, 2000; Chen and Rice, 2003). Once the two recombination sites come together, a so-called synapsis is formed. Due to the symmetry of the synaptic complex (a pseudo-tetramer) only half of the four subunits are active at a time. Two transesterification reactions with a transient 3' phosphotyrosyl linkage lead to a Holliday junction-intermediate. The Holliday junction can only be resolved if the spacers of the two recombination sites can base-pair, but its exact sequence is flexible. That is the basis of designing orthogonal (non-reactive) FRT or loxP pairs. Lastly, an isomerization step switches the half-site activity to the other pair of monomers and two transesterification reactions resolve the Holliday junction, leading to the recombination product (Chan et al., 2013; Chen, 2000; Chen and Rice, 2003).

At our synthetic locus two loxP sites (one in the genome, one in the plasmid) undergo interchromosomal recombination, leading to an integration. Two FRT sites flank the backbone of the plasmid in a head-to-tail orientation. The intrachromosomal recombination of the FRT sites leads to an excision.

Both Cre and Flpo have been widely used due to not requiring any accessory proteins, having no preference for DNA topology and recombining integration-, excision-, and inversion-reactions⁵ (Yarmolinsky and Hoess, 2015).

Engineering Directionality of loxP Sites.

Integration and excision are two sides of the same reaction. In comparison to serine recombinases (Chapter III) directionality of the reaction is not regulated for Cre and Flp and because intrachromosomal reactions are kinetically and thermodynamically favored, the product is usually an excision (Baer and Bode, 2001; Ringrose et al., 1998). Since they do not insert sequences in their biological context, they do not need to regulate this

³ Flp recognition target

⁴ Locus of crossover(x) - the "P" was assigned to the phage lox-site, as they also found pseudo-lox sites in the bacterial host (loxB), which are probably not relevant for the biological function of the Cre/lox system (Yarmolinsky and Hoess, 2015).

⁵ Lambda-Int only integrates or excises the lambda-phage (direct repeats). It also needs accessory proteins and requires negative supercoiling of the attP site, making it not particularly useful for genomic engineering, despite being well studied. Invertases do not recombine direct repeats and resolvases do not recombine inverted repeats (Yarmolinsky and Hoess, 2015).

process. To allow integration for synthetic biology, mutant loxP sites were developed (Albert et al., 1995; Araki et al., 1997). While lox66 has a mutation in the upstream (left) binding site, lox71 has a mutation in the downstream (right) binding site (Shaw et al., 2021). Both are similarly active as loxP, but once recombined an active loxP(WT) site and an inactive lox71/66 (lox72) mutant emerges, making the recombination essentially unidirectional (Fig. 2, (Albert et al., 1995; Araki et al., 1997; Shaw et al., 2021)).

Orthogonal Recombination Systems by Mutating the Spacer.

The ability to resolve the Holliday junction during a recombination event depends on base-pairing of the spacer sequence (Chen and Rice, 2003). Mutating the spacer without losing recombination efficiency and thereby generating orthogonal loxP or FRT pairs set the frame for the first recombination-mediated cassette exchange (RMCE) approaches (Chapter III) and integrating multiple constructs in the same cell (Sauer, 1996).

Already in 1986, the 8-bp spacer of FRT was found to be largely flexible (Senecoff and Cox, 1986; Umlauf and Cox, 1988), which was also initially assumed for loxP (Hoess et al., 1986), but later turned out to be more complicated (Lee and Saito, 1998; Missirlis et al., 2006). For Cre, most of the single and double mutants appeared to have significantly lower efficiency or did not recombine at all (Lee and Saito, 1998) although a more unbiased mutational screening approach showed that heavily mutated spacers could reach efficiencies beyond the WT loxP (Missirlis et al., 2006). Conveniently, for both FRT and loxP as little as a single point mutation can be enough to reduce the recombination efficiency drastically when used in combination with the WT FRT/loxP or a different mutant, which led to the principle of orthogonal/non-compatible pairs (Hoess et al., 1986; Langer, 2002; Lee and Saito, 1998; Missirlis et al., 2006; Schlake and Bode, 1994; Umlauf and Cox, 1988). F3 and F5 are 1-bp mutants used in our synthetic locus for removing the backbone of the targeting vectors at 75 kb and 25 kb, respectively (Fig. 2, (Schlake and Bode, 1994)). While the 1-bp mutant loxP511 still maintained slight cross-reactivity with loxP(WT) (Feng et al., 1999, p. 199; Kolb, 2001; Langer, 2002; Lee and Saito, 1998), the 2-bp mutant loxP2272⁶ showed no cross-reactivity but slightly lower recombination efficiency and is used for our 25 kb landing pad (Lee and Saito, 1998). Another mutational spacer screen developed the loxPm2 variant (Langer, 2002), which was confirmed to be not only incompatible with loxP(WT) but also highly efficient (Kim et al., 2007). We used loxPm2 for the 75 kb landing pad.

In Chapter I, I utilized a synthetic platform that uses the mutant spacers to form three orthogonal integration-excision systems (Thomas et al., 2023). In Chapter III the spacer of the serine recombinase Bxb1 is used for an RMCE approach.

Aims.

In a previous publication our laboratory has shown that strong enhancers like the proximal enhancer of *Nanog* (eNanog), the E1 enhancer of *Map4k3* and the E2 enhancer of *Rybp* at 75 kb can synergize with the weak E2 enhancer of *Map4k3* at 25 kb (Thomas et al., 2023). Whether this is a special property of *Map4k3* E2 or if weak elements generally synergize with stronger elements was not tested. Therefore, I selected four

⁶ loxP2272 is still highly relevant. The Big-IN technology developed for creating large synthetic regulatory networks is based on a RMCE approach with loxP and loxP2272 (Brosh et al., 2023).

previously untested weak enhancers and integrated them at 25 kb with or without *eNanog* at 75 kb to assess cooperativity. Additionally, the weak enhancers were integrated at 1.5 kb to assess their intrinsic strength and to measure their ability to overcome genomic distance. Assuming that enhancers do not differentially affect the 3D organization around them (Jensen et al., 2024; Zuin et al., 2022), we can revisit our previous observation that the ability of enhancers to overcome distance is a function of enhancer strength.

Results.

Selection of Weak Enhancers.

The *Sox2* locus provides a rich set of well characterized enhancers (Brosh et al., 2023). I selected 2 enhancers (E19, E20) that did not show any STARR-Seq activity and synergized strongly with classical elements. Additionally, a classical enhancer (E15) was chosen. E15 showed STARR-Seq activity and together with E10 was necessary for 22% activity (Brosh et al., 2023). A knockout of E15 was also shown to alleviate the enhancing effect of condensate-proximity on burst size and burst frequency at the *Sox2* gene (Du et al., 2024). Since our previous publication showed that *Rybp* E2 can synergize with *Map4k3* E2, but its endogenous partner *Rybp* E1 was not analyzed, I also selected this element (Thomas et al., 2023). In comparison to *Rybp* E2, E1 is bound by lower levels of p300 and shows less H3K27Ac (Thomas et al., 2023). The sizes of the enhancers were selected based on p300 binding. The enhancers were amplified via PCR from gDNA of *Mus musculus castaneus* (CAST/EiJ) and cloned into the targeting vectors for the 1.5 and 25 kb position (Fig. 2, CB231, CB201)

To account for clonal variation an extensive replication strategy was sought out (Fig. 3). Two separate clones (A and B) were cultured after integration of the targeting vector. The backbone was removed independently with Flpo and 2-3 clones were analyzed each. These 4-6 biological replicates were measured every second day to get 4-5 technical replicates, leading to ~20-24 data points per construct.

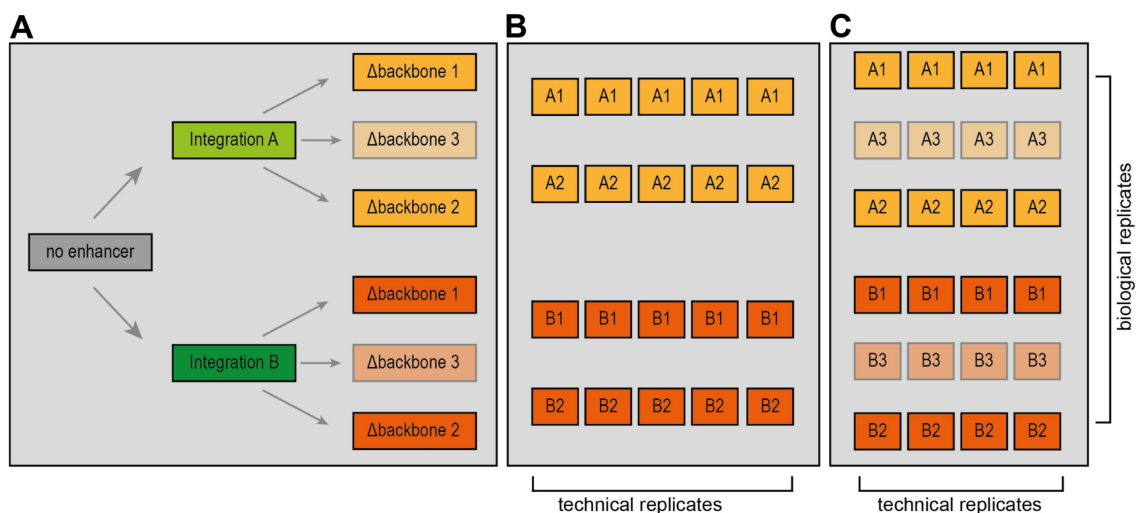


Figure 3: Replication strategy. A) Each enhancer is integrated twice separately. For each integration the backbone is removed with Flpo and 2 clones each are analyzed for single enhancer integrations and 3 clones each for dual enhancer integrations. B) Single enhancer constructs are measured every other day, five times in total. C) Dual enhancers constructs are measured four times in total.

After Cre-mediated integration of the targeting vector and antibiotic selection, the percentage of positive clones after genotyping varied extensively and was a function of genomic distance, declining from 1.5 kb to 25 kb. The lowest efficiency was found at 75 kb (Table 1). Conversely, Flpo-based backbone removal achieved ~60-80% efficiency, independent of genomic position (Table 2).

Table 1: Positive clones after Cre-mediated insertion and antibiotic selection.

	1.5 kb	25 kb	75 kb
Rybp E1	46%	8%	/
Sox2 E15	30%	20%	/
Sox2 E19	64%	20%	/
Sox2 E20	46%	16%	/
Rybp E2 3OPT	/	16%	16%
Rybp E2 allOPT	7%	9%	6%
Rybp E2 new	18%	6%	7%
Map4k3 E1 3OPT	18%	5%	3%
Map4k3 E1 allOPT	44%	19%	5%
Map4k3 WT new	41%	13%	2%
Average	35 +/- 18%	13 +/- 6%	7 +/- 5%

Table 2: Positive clones after Flpo-mediated backbone removal and antibiotic selection.

	1.5 kb	25 kb	75 kb
Rybp E1	16 %	79%	/
Sox2 E15	32%	79%	/
Sox2 E19	25%	42%	/
Sox2 E20	74%	38%	/
Rybp E2 3OPT	/	71%	/
Rybp E2 allOPT	93%	20%	82%
Rybp E2 new	95%	/	69%
Map4k3 E1 3OPT	85%	94%	90%
Map4k3 E1 allOPT	86%	88	80%
Map4k3 WT new	70%	58%	92%
Average	64 +/- 29%	63 +/- 24%	83 +/- 8%

Intrinsic Strength and Ability to Overcome Genomic Distance Does Not Correlate for All Weak Enhancers.

The expression of mCherry caused by the 4 weak enhancers at 1.5 kb and 25 kb was measured via flow cytometry (Fig. 4A and Fig. 4B). For each measurement the mean expression was calculated, and the no enhancer cell line (#1) was subtracted (Supplementary Note 1). Additionally, I included *Map4k3* E2 as a control to enable comparison to our previously published enhancers. Linear regression was performed between each pair of technical replicates (Fig. S1, Fig. S2). The residuals of the 5 technical replicates were remarkably low, with R^2 ranging from 0.90 to 0.98, indicating highly consistent results, even at lower expression levels. Interestingly, replicates measured closer together were clearly more similar to each other than replicates measured further apart and residuals between replicates clearly decreased over time (Fig. S3). While the slope of the regression was up to ~20% different when analyzing early replicates (1 and 2), slopes of later replicates are remarkably close to 1. The y-intercept, indicating systematic differences between measurements was a lot better after subtracting the background (#1), with replicate 4 being a bit of an outlier (Fig. S3). Additionally, clones from the same Cre-mediated integration were more similar than clones from independent integrations and the median expression difference between clones carrying the same enhancer was close to zero (Fig. 4C).

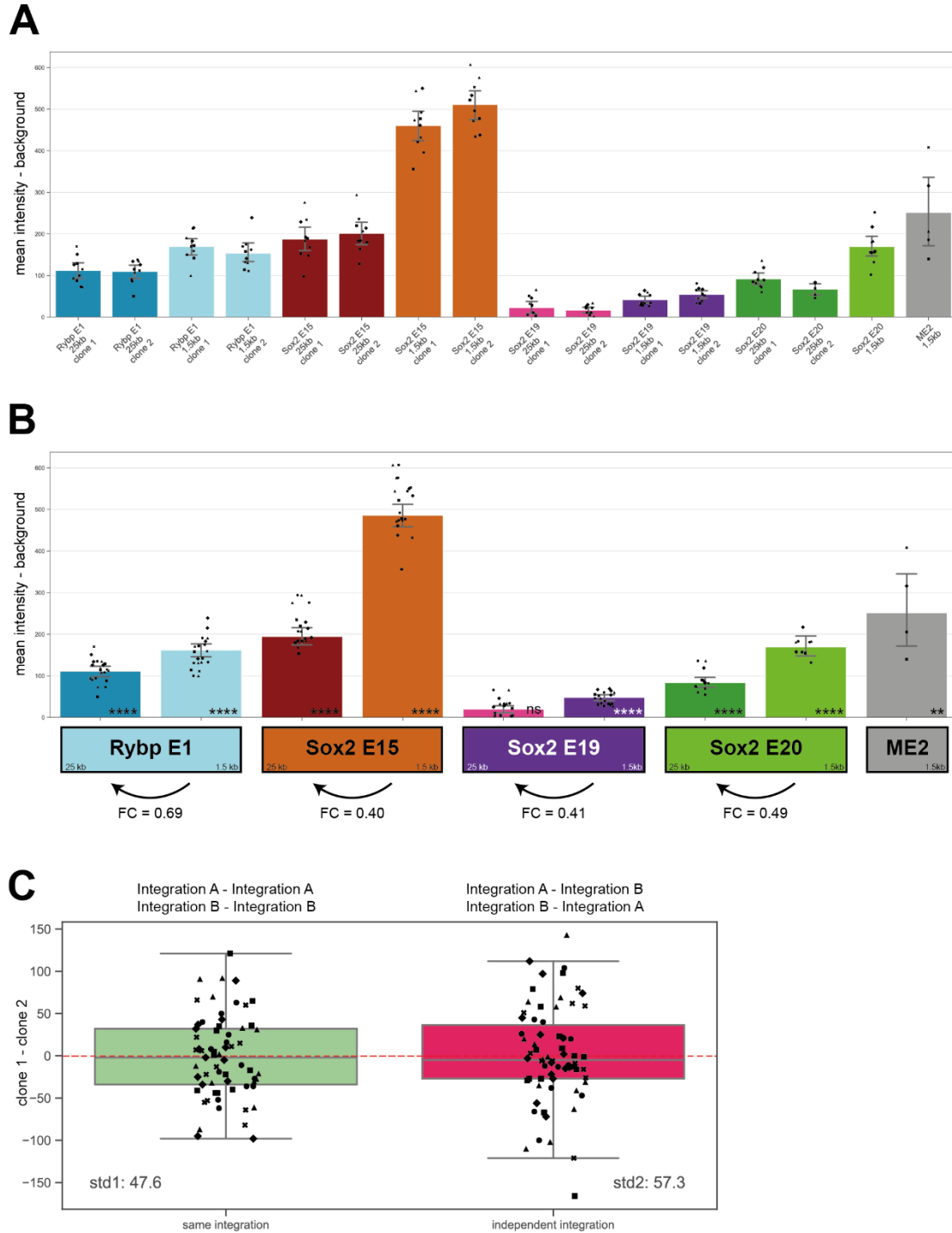


Figure 4: Weak enhancers and distance dependency. Each dot represents the mean mCherry intensity subtracted by the background mCherry intensity from the same technical replicate. A) Each integration is depicted as a barplot. Different symbols correspond to different technical replicates. B) Integrations of the same construct are combined. The significance of being different from background is indicated as asterisks on the bottom of the bars. The fold-change (FC) between 25 kb and 1.5 kb (drop in activity) is indicated. C) Two samples from the same construct are subtracted. Either the two samples originate from the same integration (green) or from separate/independent integrations. The difference should deviate from zero if the compared mean intensities are very different.

All 4 enhancers had low intrinsic activity, as defined by expression at 1.5 kb. Sox2 E15 had the strongest activity, in concordance with being a classical enhancer. The strength of Sox2 E15 was comparable to *Map4k3* E1 of our previous dataset (Thomas et al., 2023). To analyze the loss of activity over distance, the fold-change was computed (Supplementary Note 1, Fig. 4B). While Sox2 E15, Sox2 E19 and Sox2 E20 maintained very similar expression levels at 25 kb (~40-50%), *Rybp* E1 was a clear outlier, maintaining ~70% of its expression levels.

Integration of Our Previous Dataset.

To find trends in enhancer distance behavior I needed to integrate information from our previous enhancer datasets (Thomas et al., 2023). To enable a direct comparison of enhancer strength, I first subtracted the background mCherry expression and then divided this value by the background to account for scaling differences, e.g., the laser power (Fig. 5, Supplementary Note 1). Previously, we only had one weak enhancer (*Map4k3* E2) that we used for dual enhancer integrations (Thomas et al., 2023). The new enhancers are all weaker than *Map4k3* E2, with the exception of Sox2 E15.

Plotting the enhancer strength vs. distance dependency showed that there is a small but significant correlation for stronger elements to act over larger distances (Fig. 5C). *Rybp* E1 and *Map4k3* E2 clearly deviate into opposite directions from that correlation. The maintenance of activity of *Rybp* E1 seems to be comparable to its neighboring element *Rybp* E2, which is at least 3 times as strong. *Map4k3* E2 on the other hand loses the most activity over distance, despite 3 elements being up to 2-fold weaker.

To assess whether *Rybp* E1 and *Map4k3* E2 could be special elements I utilized the JASPAR database to analyze the presence of transcription factor motifs (Rauluseviciute et al., 2024). Comparing the 9 enhancers with each other, most elements contained a highly overlapping set of motifs. Interestingly, *Rybp* E1 and Sox2 E19 were highly unique, both containing 6 and 7 motifs for TFs, respectively, that were absent in any other enhancer (Fig. 5D). *Map4k3* E2 was unique in containing the least number of motif instances and lowest transcription factor diversity (Fig. 5E and Fig. 5F). When normalized to the enhancer length, these values are comparable to the other enhancers (not shown). *Map4k3* E2 is simply a very small element.

Overall, 3 out of 4 weak enhancers showed similar distance dependency, while *Rybp* E1 is remarkably similar to *Rybp* E2, an element ~3-fold stronger originating from the same locus.

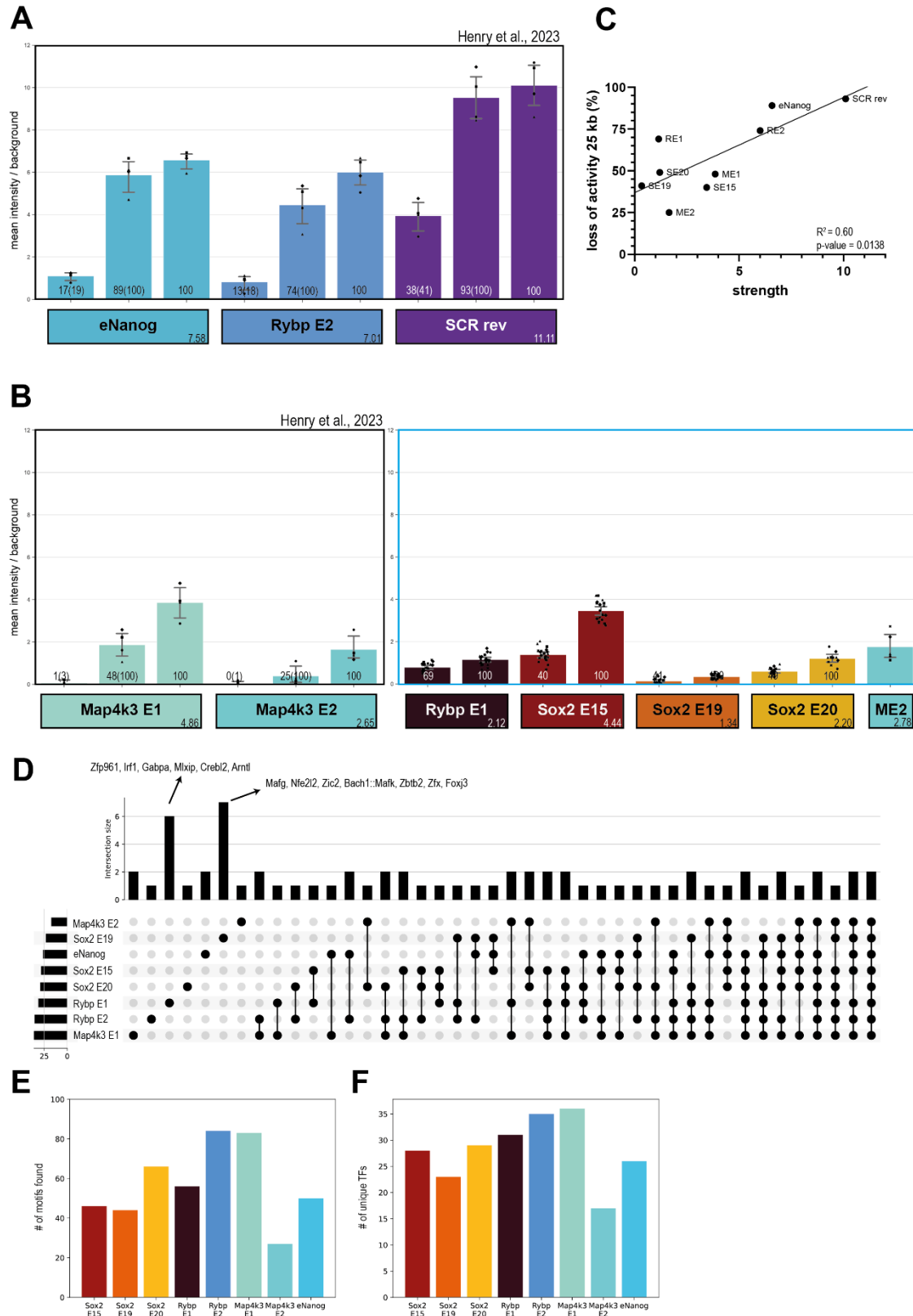


Figure 5: Comparison of distance dependency and strength of all previously measured enhancers (Thomas et al., 2023). A, B) Measured intensities are subtracted and divided by the background. They represent the fold-enrichment over background. Values at the bottom of the bar represent loss of activity either relative to 1.5 kb or 25 kb (in parenthesis) as measured by the fold change of background subtracted mean intensities. Values on the bottom right of the boxes with the enhancer names indicate the intrinsic

strength (1.5 kb position) of the enhancer. A) Strong enhancers at 75 kb, 25 kb and 1.5 kb. B) Intermediate to weak enhancers at all three positions for the Thomas et al. dataset and only 25 kb and 1.5 kb for my dataset. C) Linear regression. Strength as measured by 1.5 kb and loss of activity over 25 kb correlated moderately well. D, E, F) JASPAR motif analysis, filtered for TFs expressed in mESCs (Buecker et al., 2014; Rauluseviciute et al., 2024). D) UpSet plot. E) Number of motifs found. F) Number of non-repetitive motifs found (proxy for TF diversity).

Most Weak Enhancers Display Synergy with a Strong Enhancer.

The proximal enhancer of *Nanog*, located at 5 kb from its endogenous promoter, is the strongest single enhancer element that we tested (Fig. 5A, (Thomas et al., 2023)). It maintains ~17% of its activity at 75 kb. To test whether some of the weak enhancers can synergize with *eNanog*, dual enhancer cell lines were generated. Integration of two elements in our system works best if both targeting vectors are first integrated, before the backbones are removed in a single step with Flpo. This allows for selection of both integrands, as some of the integrated vectors can be excised through low but significant activity of lox72 during Cre-mediated integration of the second vector. In this way we generated 4 dual enhancer combinations. We started with one clone of *eNanog* at 75 kb, cultured 2 independent integrations of each weak enhancer at 25 kb and *analyzed* 3 clones each after Flpo removal (Fig. 3). These 6 biological replicates were analyzed 4 times to obtain 24 data points. Similar to single enhancer cell lines, replicates measured closer together correlated stronger and R^2 tended to increase over time (Fig. S4, Fig. S5, Fig. S6).

To compute an expected value, I assumed an additive model. The amount of synergy can then be described as the FC of observed vs. expected and can be compared directly between different measurements (Supplementary Note 1).

All weak enhancers, with the notable exception of Sox2 E20, showed strong synergy with *eNanog* (Fig. 6). In terms of distance dependency, absolute strength at 25 kb or 1.5 kb and loss of activity over distance, Sox2 E20 is average. Utilizing the JASPAR database, Sox2 E20 is also average in number of unique or overlapping TFs with *eNanog* (Rauluseviciute et al., 2024). It is very likely that Sox2 E20 is simply an outlier and might work synergistic with other elements.

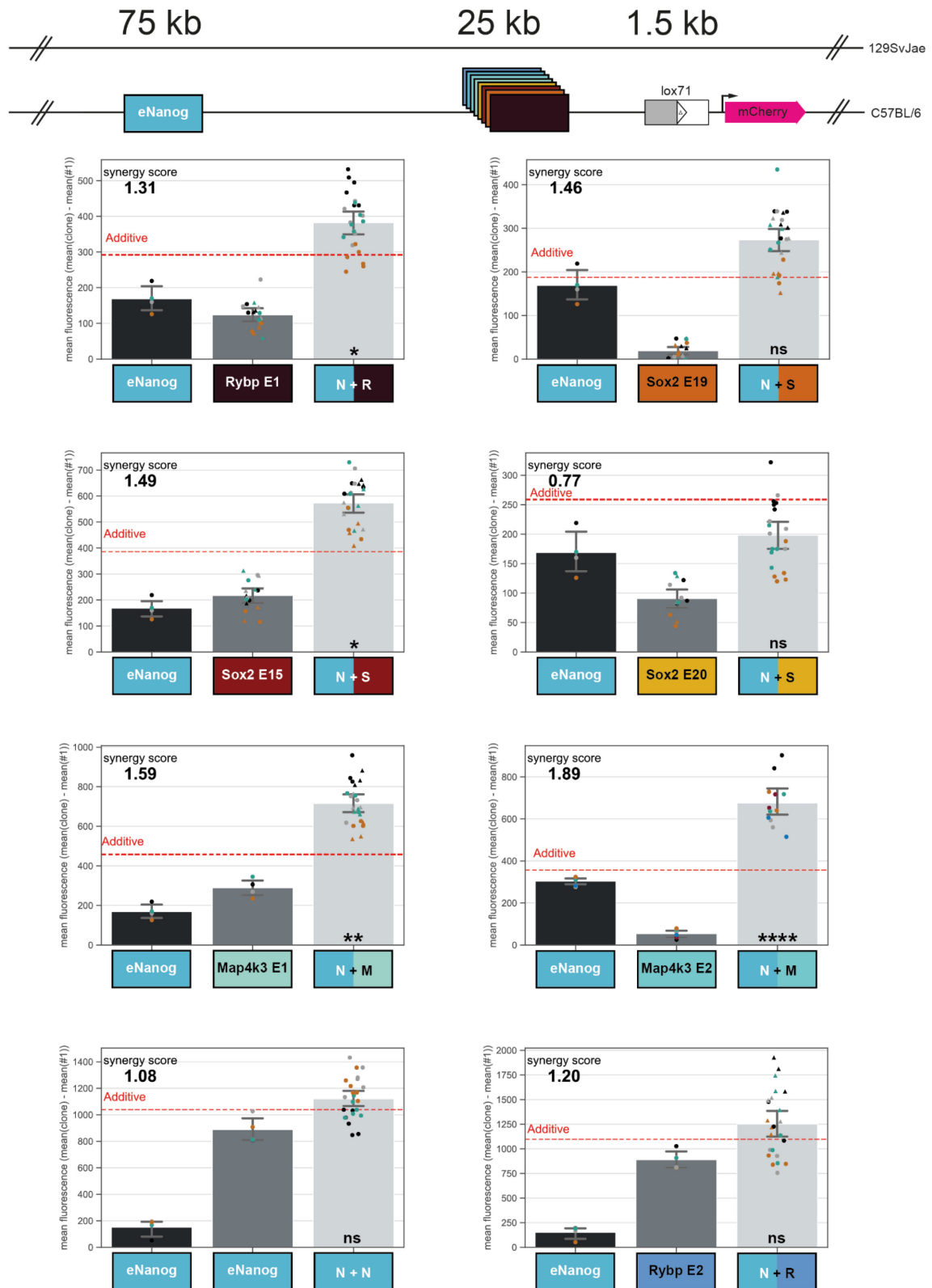


Figure 6: Dual enhancer integrations including previously measured combinations (Thomas et al., 2023). Colors of dots represent different technical replicates. Different symbols represent independent integrations (Fig. S7). Not all constructs have independent integrations. The order of the bars aligns with the upper figure (first 75 kb, then 25 kb, lastly 75 kb + 25 kb). Significance in being different from the additive model (two-tailed independent t-test) is always indicated at the bottom of the 3rd bar. Synergy score is equal to fold-change of observed vs. expected under the additive model.

The other 3 weak enhancers show synergy, albeit at lower levels than the previously published *Map4k3* E2. Integrating synergy data from *Map4k3* E2 and three previously established cell lines with *Map4k3* E1, *Rybp* E2 or another copy of *eNanog* at 25 kb shows that there is a highly significant trend for enhancers with more distance dependency to show more synergy with *eNanog* (Fig. 6A). This is very evident for *Rybp* E1 and E2, which have very different intrinsic activities but similar distance dependency and therefore similar synergy with *eNanog*. Also, *Sox2* E15 and *Sox2* E19 have very different activities but similar loss of activity over distance and therefore similar cooperativity with *eNanog*. This could indicate that *eNanog* helps the “inside” enhancer to compensate for the loss of activity. Indeed, the synergistic expression is close to the sum of *eNanog* at 75 kb with the other enhancers at 1.5 kb (Fig. 6B).

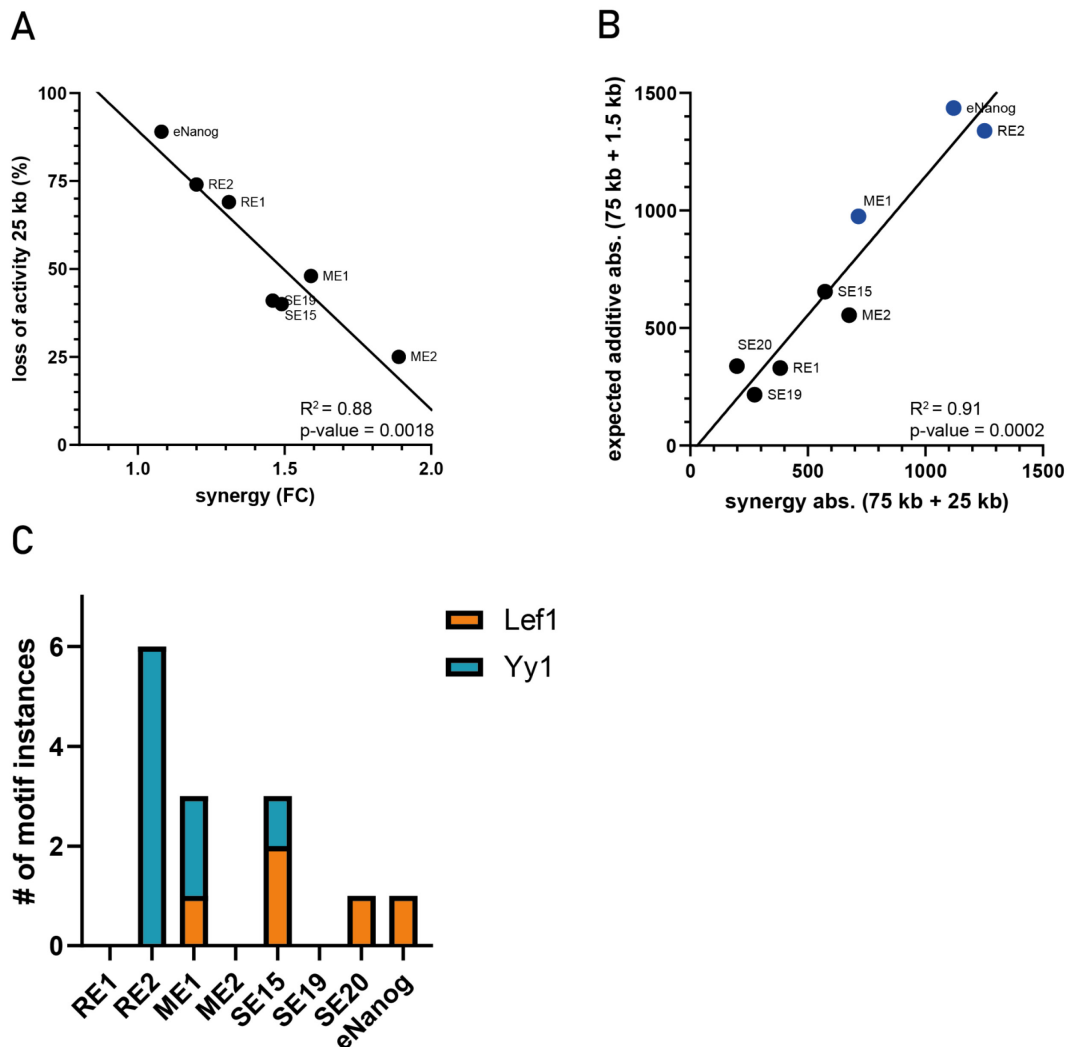


Figure 7: Trends in enhancer strength, distance dependency, synergy and motifs. Data from this figure is again a combination of my constructs and a previously published dataset (Thomas et al., 2023). A) Loss of activity as percentage (fold-change of 25 kb vs. 1.5 kb). Synergy as fold-change of observed vs. expected under an additive model. B) Theoretical values expected from a combination of 75 kb and 1.5 kb under an additive model. Synergy abs. (75kb and 25 kb) corresponds to observed absolute values. Expected additive abs. values had to be extrapolated for the blue dots (Supplementary Note 1). C) Number of high affinity LEF1 and YY1 motifs found according to the JASPAR database (Rauluseviciute et al., 2024).

With additional dual enhancer cell lines from our previous publication (Thomas et al., 2023) and from unpublished data, 6 of 7 enhancers show synergy with *eNanog* (excluding *eNanog* with itself). While the exact rules governing the extent of synergy might be more complex, the system is very sensitive in detecting differences between elements and a high-throughput approach will allow us to find these rules.

Discussion.

Using our synthetic locus, I analyzed the activity of 4 weak enhancers at two positions close to a minimal promoter and the combinatorial activity of *eNanog* at 75 kb with a weak enhancer at 25 kb. Integrating our previous dataset and other unpublished cell lines showed that in contrast to a previous publication (Zuin et al., 2022) there was a moderate trend for stronger elements to work over larger distances. Often these observations are interpreted as “saturation”. In other words, because the enhancer is so strong, it does not matter whether its contact probability and therefore its influence on the promoter is reduced, since it is still sufficient to keep the promoter engaged at maximum levels. However, to my knowledge this is purely speculative. The number of TFs bound is thought to scale linearly with expression (Doughty et al., 2024). Additionally, saturation of enhancer-promoter pairs occurs only if the promoter contains built-in enhancers (Bergman et al., 2022; Martinez-Ara et al., 2022). If both enhancer and promoter contain similar motifs, it is thought that they act on the same rate-limiting step, thereby minimizing the effect of the enhancer on the promoter (Bell et al., 2024; Bergman et al., 2022). Weak promoters, such as our minimal promoter, do not saturate with strong enhancers (Bergman et al., 2022). Also combining two enhancers with one promoter in a MPRA did not reveal any upper limit (Martinez-Ara et al., 2023), although there was slight saturation in a STARR-Seq design (Loubiere et al., 2023). Additionally, continuous bursting, the clear physical limit of a promoter, is rarely observed (Suter et al., 2011). Therefore, the “saturative behavior” might not originate from promoter dynamics but from distance effects. As suggested in Zuin et al. a more fitting term would be “buffering” (Zuin et al., 2022). Instead of promoter saturation overshadowing the fact that each enhancer has the same distance dependency, strong enhancers are probably just less sensitive to distance. This can be rationalized with the TAG-model, in which stronger enhancers can compensate for the loss of activity by a greater TF-gradient, spanning a larger space (Karr et al., 2022). Additionally, apart from the well-known CTCF, other TFs such as LEF1 (Bower et al., 2024) and YY1 (Gao et al., 2023) were shown to contribute to overcoming enhancer-promoter distance. Indeed, all our intermediate to strong enhancers have high affinity sites for either LEF1 or YY1 (Fig. 6C). *Rybp* E2 appears to be a very special element, having 6 YY1 motifs.

While in our previous dataset it seemed as if distance dependency and strength are tightly linked, there are some remarkable exceptions. Despite being around 3-fold weaker than *Rybp* E2, *Rybp* E1 had remarkably similar distance dependency (Fig. 5). Similarly, *Sox2* E15 and *Sox2* E19 shared the same distance dependency, despite their 3-fold difference in activity. Whether enhancer distance dependency contains a locus-specific component by evolving in a unique regulatory environment remains to be

investigated. A decoupling of strength and distance dependency would be also compatible with the TAG-model (Karr et al., 2022).

Integrating the new enhancers at 25 kb with *eNanog* at 75 kb showed significant synergy for 3 of 4 elements. Analyzing previously generated datasets revealed that there is a highly significant linear relationship between distance dependency and synergy. Elements that lose their activity rapidly benefit more from *eNanog* presence (Fig. 7A). It was previously suggested that in multi-enhancer loci, the outermost distal enhancer firmly connects to the promoter (e.g., through loop formation) while intermediate enhancers stabilize this interaction (Song et al., 2019). Recently also YY1-bound genome organizing elements (YGOs) were identified at the *Igλ* locus (Gao et al., 2023). YGOs show H3K27Ac and low TF-binding as well as strong YY1-binding, but no intrinsic enhancer strength as measured by STARR-Seq and therefore form a similar (possibly overlapping) class to the facilitator elements from the α -globin locus or the range extenders (REX) from the *Shh* locus (Blayney et al., 2023; Bower et al., 2024). Through dimerization, eRNA binding and interaction with disordered domains of TFs, YY1 is thought to create a hub-like structure between multiple enhancers (Weintraub et al., 2017). Additionally, cohesin can affect enhancer-promoter contacts (Rinzema et al., 2022) and clusters of enhancers recruit cohesin synergistically (Georgiades et al., 2023). This shows that synergy can clearly be caused by changing 3D genome organization. However, analysis of TF motif presence showed that none of our weak enhancers can be regarded as YGO elements or REX elements (Fig. 7C). Although *Rybp* E2 contains an unusually high number of YY1 motifs, it is a classical enhancer. It would be interesting to see whether exchanging *eNanog* with *Rybp* E2 would boost synergy. Indeed, in our previous publication *Rybp* E2 and *Map4k3* E2 show a ~2.7-fold increase to the additive model, despite *Rybp* E2 being weaker than *eNanog* (Thomas et al., 2023).

An open question is whether the strong enhancer supports the weak enhancers or vice versa. While it might be partially reciprocal, the loss of activity over distance promoting instead of diminishing synergy (Fig. 7A) argues against the possibility that the weak enhancer helps the outside enhancer to come into closer contact. Rather the outside enhancer might help to buffer the loss of activity of the weak enhancer. If that was the case, one would expect that the synergistic expression levels should be close to the sum of *eNanog* at 75 kb and the weak enhancer at 1.5 kb. Not only is that the case theoretically (Fig. 7B), but when keeping *eNanog* fixed at 75 kb and moving *Map4k3* E2 to 1.5 kb, the total expression levels are remarkably similar, but synergy decreases as *Map4k3* E2 is moved closer to the promoter (Thomas et al., 2023). In other words, with *eNanog* at 75 kb, moving the inside enhancer from 1.5 kb to 25 kb reduces its own contribution while maintaining total expression levels, leading to the appearance of synergism. Clearly, not only enhancer-promoter but also enhancer-enhancer distance is changed, which could be a confounding factor. This could be alleviated by studying synergism between two elements far away from a promoter, making the change in enhancer-promoter distance negligible.

We have previously found that synergism between two elements also depends on the order (Thomas et al., 2023). While I did not explore that further with the new elements,

we have previously shown that the inside enhancer at 25 kb does not have any specific strength requirements, while the outside (distal) position clearly needed to be a strong enhancer to show synergy. This is also consistent with the distal enhancer supporting the inside element. Conversely, if the inside element would indeed help to stabilize interaction between the distal enhancer and the promoter, as suggested for chain-enhancers (Song et al., 2019) and YGOs (Gao et al., 2023), there is no reason why a compaction of the locus by a strong inside enhancer would not increase the influence of the weaker outside enhancer to the promoter, resulting in synergy. However, if one assumes synergy originating from the outside position, the element would need sufficient strength to overcome the large distance and help the inside element.

That synergism is not exclusively a function of distance is exemplified by another recent observation that 2 copies of *eNanog* behave additively at 75 kb and 25 kb (Fig. 6). Despite synergism being more of a rule rather than the exception at the *Sox2* locus, the classical enhancer DHS24 also behaves additively, when inserted twice (Brosh et al., 2023). Requiring enhancer or motif diversity for synergy could be caused by different TFs working on different rate-limiting steps, leading to “kinetic synergy” (Bell et al., 2024). While this “kinetic synergy” happens at the promoter, recruitment of chromatin remodeling factors could also lead to non-linear effects of TF-binding. In this case enhancer-enhancer proximity would be more important than enhancer-promoter distance and would not appear in episomal reporter assays, since they are not chromatinized (Doughty et al., 2024). Comparing number of overlapping or unique TFs between two elements did not perform well in predicting the amount of synergy, either because distance effects mask the contribution of TFs, because motif presence is not a good predictor of TF binding and cofactor recruitment, or because synergy is indeed simply a distance factor. For example, *eNanog* maintains 89% of its activity at 25 kb and does not necessarily need the outside enhancer to support its activity. Indeed, from the distance dependency vs. synergy relationship alone we would not predict it to synergize (Fig. 7A). Also, while we did see additivity for *Map4k3* E1 with itself at 75 kb and 25 kb in a recently generated cell line (not shown), this was only barely significant and could be caused by the low activity of *Map4k3* E1 at 75 kb. It would be intriguing to see how different enhancers at 75 kb support a plethora of enhancers at 25 kb. Detecting synergistic levels higher than the sum of 75 kb and 1.5 kb could argue against my proposed model, that the strong distal enhancer does not contribute beyond supporting the inside enhancer. In this case, one would need to disentangle the individual contributions of increased enhancer-promoter distance and decreased enhancer-enhancer distance.

While the rules governing distance dependency of single enhancers might be more complex and not reducible to a single factor, synergy appears to be more unidimensional, with genomic distance being the major contributing factor (Blayney et al., 2023; Loubiere et al., 2023; Martinez-Ara et al., 2023; Thomas et al., 2023). Our synthetic locus allows us to dissect these rules, even with a minimal amount of regulatory elements. A high-throughput approach combined with machine learning will allow us to disentangle these aspects further (Chapter III).

Limitations.

We use mCherry protein expression as a proxy for enhancer activity. While RNA levels and protein levels do not necessarily correlate, this can be partially explained by protein synthesis and degradation kinetics (Semrau et al., 2017). In a similar synthetic system, it has been recently shown that GFP intensity correlates well with mRNA levels measured by RNA-FISH (Zuin et al., 2022). It is therefore unlikely that enhancers differentially regulate mCherry expression post-transcriptionally. Additionally, our synthetic locus contains two convergent CTCF sites between 75 kb and 25 kb. I discussed the connection between enhancers and CTCF extensively in the Introduction. It is possible that CTCF/cohesin-mediated loop extrusion underlies synergy and/or differences in distance dependency. The motif analysis via JASPAR might not be particularly informative because motif presence alone is not a good predictor of TF binding, especially if filtered for high-affinity motifs, as discussed in Chapter II.

Chapter II

Introduction.

The fundamental unit of an enhancer is the transcription factor motif. Motifs are not independent units but are part of a complex sequence grammar (Reviewed in Jindal and Farley, 2021). The building blocks (motifs with specific affinities) are arranged in specific order, spacing and orientation, each contributing to enhancer activity (Jindal and Farley, 2021). It follows that the activity of each motif is a function of its sequence context. In an elegantly designed saturation mutagenesis experiment, an 8-bp library was pasted into different positions of a developmental enhancer (Reiter et al., 2023). Strikingly, the majority of differences in activity between positions could not be explained by motif sequence (Reiter et al., 2023). However, owing to recent advances in machine learning, predicting (and designing) enhancers with specific strength (De Almeida et al., 2022) and tissue specificity (de Almeida et al., 2024; Taskiran et al., 2024) from sequence became feasible.

As discussed in “Enhancers and Their Role in Evolution and Disease.” SNPs within motifs are a frequent cause of phenotypic variability, disease and evolution. Predicting the effects of SNPs is still challenging, partly because most motifs are imperfect (Crocker et al., 2016, 2015; Farley et al., 2015). Such low affinity binding sites are crucial, since they are highly context dependent and allow expression only in the appropriate cell state (Crocker et al., 2016; Jindal and Farley, 2021; Lim et al., 2024). Lim et al. has previously shown that increasing the affinity of low affinity binding sites generally increases enhancer strength and can lead to ectopic expression (Lim et al., 2024). This was exemplified by the ETS-A motif in the ZRS enhancer, which is predicted to have 85% less affinity than the “perfect” motif. Interestingly, SNPs that increased the affinity by about 1.6-fold, as found in natural pathogenic variants, lead to ectopic expression and polydactyly, probably through increased TF residence time (Lim et al., 2024).

To understand how motif affinities affect gene expression mechanistically, it has been shown that lowering motif affinities and hence TF residence time, lowers transcriptional bursting duration (Donovan et al., 2019). Pomp et al. has recently shown that TF binding is required throughout a bursting event and that TFs can exchange to prolong bursts, bridging the gap between short TF binding kinetics and longer bursting duration. Changing motif affinities reduces the residence time of the TFs and lowers the average time a TF is bound to any of several motifs, lowering bursting duration (Pomp et al., 2024).

Aims.

While the effects of motif grammar on enhancer strength and tissue specificity have been well explored, the connection between enhancer strength and its ability to overcome genomic distance is still unknown. In Chapter I, I could show that enhancer strength and distance dependency are loosely connected, with stronger enhancers maintaining transcriptional levels at larger distances (Fig. 5C). In collaboration with the laboratory of

Emma Farley, we have decided on optimizing several SOX2 and OCT4 binding sites within *Map4k3* E1, an enhancer of intermediate strength. Similar to Chapter I, I integrated these variants at 3 positions to a minimal promoter driving mCherry (75 kb, 25 kb, 1.5 kb) to disentangle the effects on intrinsic enhancer strength to overcome genomic distance. Apart from biological implications, the results should primarily indicate how sensitive our approach is to detect small changes in enhancer strength, especially since the absolute expression levels will be drastically reduced over large distances.

Results.

Optimizing Motif Affinity Increases Enhancer Strength While Maintaining Sensitivity to Genomic Distance.

In ESCs the well-studied TFs SOX2 and OCT4 are both necessary for maintaining pluripotency (Yeo and Ng, 2013) and in combination with KLF4 and MYC are able to reprogram fibroblasts into ESCs (Takahashi and Yamanaka, 2006). SOX2 and OCT4 are both pioneer factors (Balsalobre and Drouin, 2022), show cooperative binding (Ambrosetti et al., 1997) and co-occupy many genes important for pluripotency (Chen et al., 2008; Marson et al., 2008). Similar to their previously published papers, Emma Farley's laboratory analyzed *Map4k3* E1 for low-affinity SOX2 and OCT4 binding sites and identified one SOX2 motif near one composite OCT4-SOX2 motif (Fig. 8, Lim et al., 2024, 2024). To increase the affinity, they have compared the sequence to protein binding microarray (PBM) data (Hume et al., 2015). The affinities are displayed normalized to the motif with the highest affinity, 1.00 indicating the highest possible value (Fig. 8). In this way they optimized the motifs in our *Map4k3* E1 enhancer sequence either through 3 mutations (3OPT) or 10 mutations (allOPT, Fig. 8). While 3OPT only partially optimizes the binding sites with an average increase of ~1.5-fold affinity, allOPT optimizes all 3 motifs to its maximum value, which corresponds to an average ~2.2-fold increase in affinity.

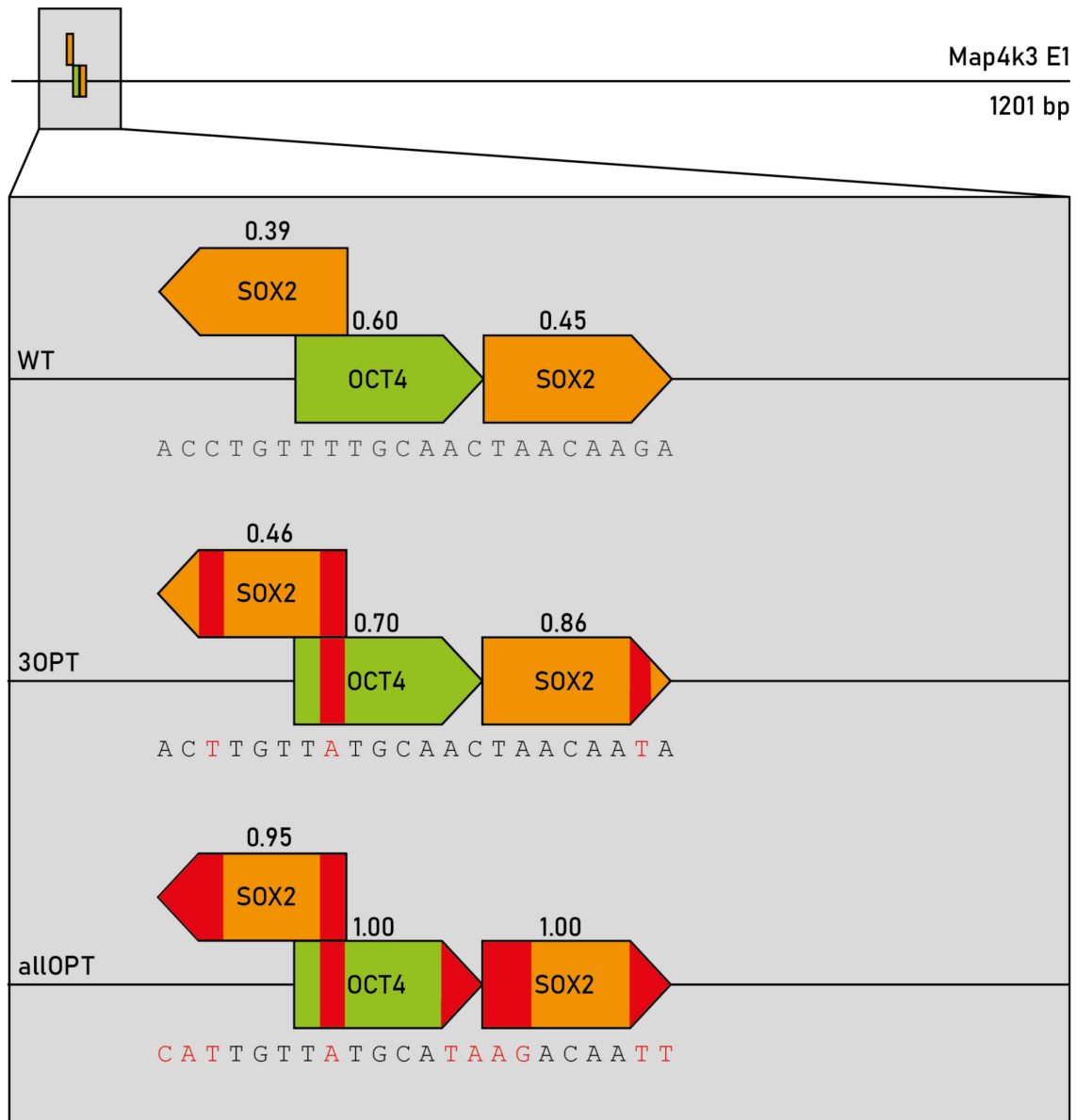


Figure 8: Optimization of SOX2 and OCT4 binding sites in *Map4k3* E1. The motif location is shown within the enhancer. Below the different variants are displayed. Mutations are shown in red. Affinities as measured by PBM are displayed above the boxes indicating motifs.

Using a site-directed mutagenesis approach, I introduced these optimizations into *Map4k3* E1 and cloned the enhancer into the three targeting vectors for 1.5 kb, 25 kb and 75 kb, respectively. As a control, I also integrated the wild-type *Map4k3* E1. This allowed not only a direct comparison to the optimized variants but also to the WT variants generated a year ago that already went through many passages. Analogous to the weak enhancer integrations in Chapter I, I attempted to remove the backbone of the integrated targeting vector individually from two separate integrations to account for clonal effects (Fig. 9A). While I could generate 9/9 constructs, I was only able to remove the backbone from two separate clones in 4/9 constructs (Fig. 9B). Similar to the weak enhancers, the individual integrations are highly similar and therefore the data points were combined (Fig. 9C, Fig. 9D). The 75 kb positions had very low expression levels overall and 2/5 75 kb constructs were not significantly different from background mCherry expression (#1).

In total, I analyzed a single replicate of 48 individual cell lines for wild-type *Map4k3* E1 and 15 cell lines for the 3OPT variant (Fig. 9B). Surprisingly, intrinsic strength as measured by the 1.5 kb position was not significantly different from the WT enhancer (Fig. 9D). Conversely, both the 25 kb and 75 kb positions were significantly different, with 3OPT at 25 kb showing a ~1.2 increase in expression and 3OPT at 75 kb showing a ~1.9-fold decrease in expression. Whether biologically informative or not, it shows that even small differences ($\sim\Delta 32$) in mean mCherry intensity can be detected reliably.

In a second round of analysis, I selected 29 of the 48 WT clones and repeated flow-cytometry, but now in comparison to 38 clones of the allOPT variant (Fig. 9B). Here, both the expression at 1.5 kb and 25 kb increased by ~1.2 fold, although the 25 kb position did not meet the traditional criteria for significance (Fig. 9D). While the 75 kb position also showed a small trend for increased expression, at least in one of the two individual integrations (Fig. 9C), the absolute levels were probably too low to be able to detect them. The allOPT variant seemed to maintain its distance dependency, despite increasing its activity significantly.

While I only measured allOPT and 3OPT once, in both cases the newly generated WT *Map4k3* E1 clones were measured. Therefore, these represent two technical replicates. Despite that, overall expression for the WT constructs in the 3OPT measurement looked significantly higher than in the allOPT measurement. This can be partially explained by overall increased expression levels, but also due to a lower background, as seen in the raw intensity plot (Fig. S8).

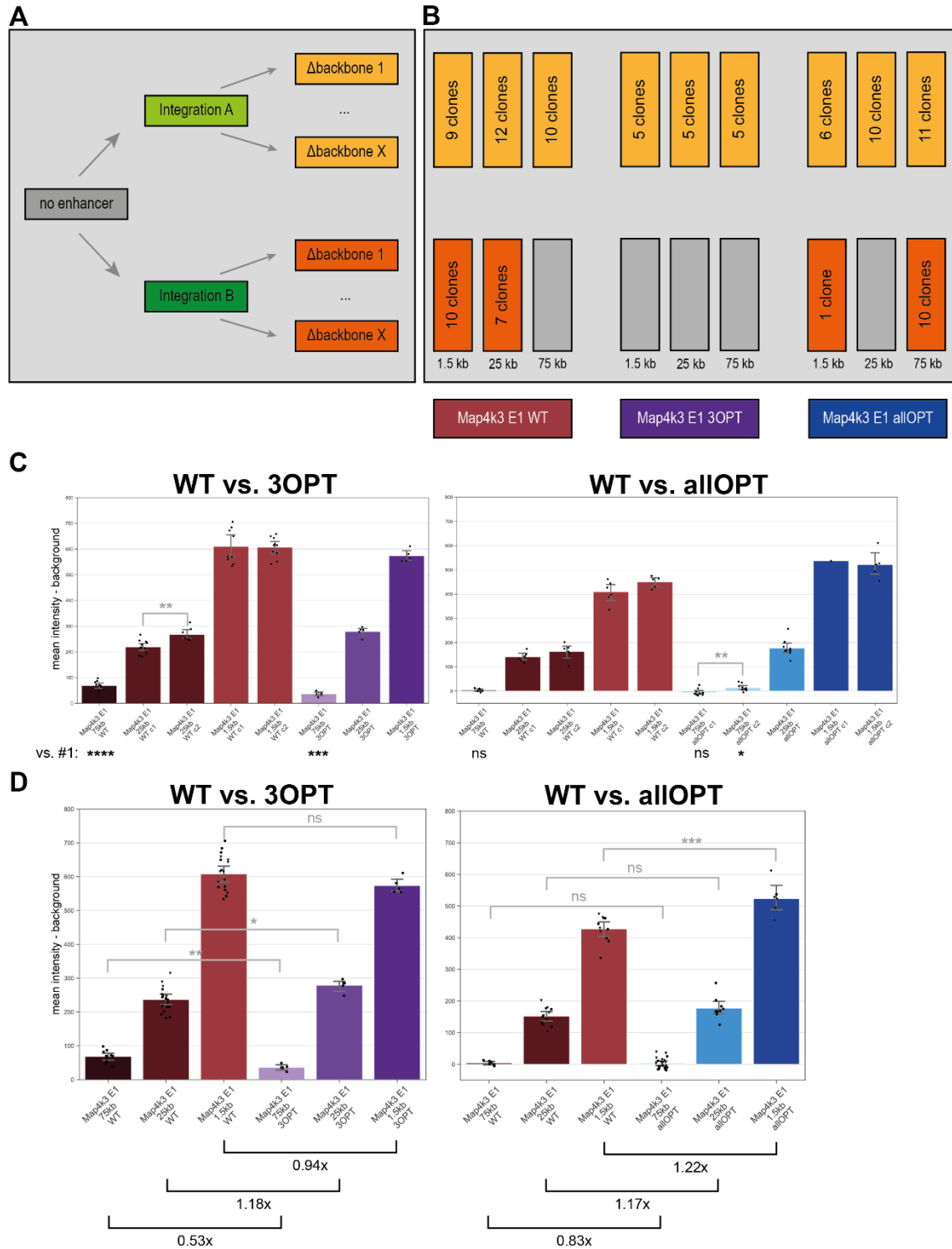


Figure 9: Enhancer strength and distance dependency of optimized *Map4k3* E1 variants. A) Replication strategy. B) Number of clones measured for each construct. There are no technical replicates. C) Mean mCherry intensity subtracted by background (#1). Individual integrations are displayed. Asterisks within the plot indicate whether two independent integrations of the same construct are significantly different. Asterisks (and “ns”) indicate whether the 75 kb positions are different from background (#1). D) Independent integrations are combined. Asterisks in the plot show whether two constructs are significantly different. Numbers under the plot indicate the fold-change of expression.

Integration of an Old *Map4k3* E1 Cell Line.

Independent of the measurements done for this thesis, a year ago WT *Map4k3* E1 cell lines were generated at all three positions, albeit only with one clone each (Thomas et al., 2023). To analyze whether gene expression levels stay comparable even after many passages and independent derivation of the constructs, I compared the background subtracted fold-enrichment over the background (Fig. 10, Supplementary Note 1). Despite small differences that can be explained by only having measured one biological replicate per position in the older dataset, the constructs were remarkably similar.

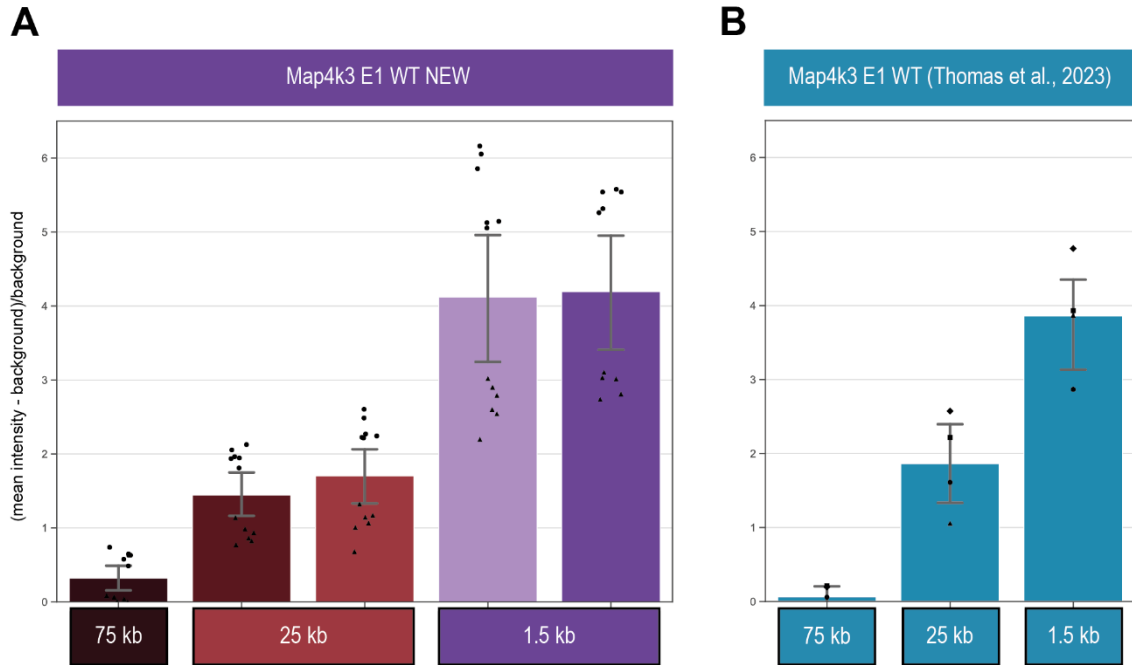


Figure 10: Comparison of newly generated *Map4k3* E1 WT with old constructs previously published (Thomas et al., 2023). A) Newly generated *Map4k3* E1. Both technical replicates are pooled and displayed as dots or triangles. B) Previously published *Map4k3* E1. The four different dots per bar represent four different technical replicates.

Discussion.

Here I integrated three versions of an enhancer with different SOX2 and OCT4 motif affinities at the three distances of our synthetic locus. Using a high number of different clones showed that clearly the effect of the enhancer on the minimal promoter and the readout of mCherry expression via flow cytometry is very specific and precise. The accuracy is limited by measuring no technical replicate. We have seen in Chapter I that different clones from the same replicate cluster together and are distinct from clones between replicates, with a tendency of earlier replicates to differ more (Fig. 4, Fig. S3, Fig. S6). This can be partially attributed to a fluctuating background (#1). Therefore, to maintain accuracy, it is necessary to measure multiple technical replicates, while the number of biological replicates can be limited due to high consistency between clones (Fig. 4C). Varying levels of background has a bigger influence on lower absolute intensity levels. While the 75 kb position of the WT *Map4k3* E1 is significantly different from background in the 3OPT measurement, it is similar to background in the alloPT

measurement (Fig. 9C). This has no biological meaning but can be attributed to fluctuating levels of baseline mCherry expression (Fig. S8).

While the absence of technical replicates precludes any biological interpretation, I am optimistic that the system is sensitive enough to detect small changes in expression, especially when combined with multiple clones from independent integrations and multiple measurements.

Chapter III

Introduction.

Limitations of the Enhancer Integration System.

In the first version of the enhancer-integration strategy, we first integrated the full targeting plasmid, before removing the backbone in a second step. Both the plasmid integration and the backbone removal were followed by extensive and time-consuming genotyping steps. Hence, the integration of a single enhancer took around a month on average but could last several months if either of the individual steps failed. Picking clones, genotyping, expanding and backfreezing twice in a row is a time-consuming and resource-intensive procedure, but necessary due to low integration efficiencies⁷. If the efficiency of integration would be increased, the screening for positive clones could be reduced to a single step.

While our previous results have demonstrated that synergism is widely spread among enhancers and follows specific rules, a large dataset with hundreds to thousands of elements will be necessary to move from anecdotal evidence to generalizable rules. Furthermore, such an extended dataset might form the basis to incorporate machine learning algorithms to scan for correlations between enhancer sequence and its ability to bridge genomic distance or interact with other elements. High integration efficiencies are needed, which will determine the number of elements we can test and the quality of our analysis. In this Chapter, I have designed a recombination-mediated cassette-exchange strategy (RMCE) based on a highly efficient serine recombinase (Bxb1). I tested its efficiency by exchanging a strong enhancer at the 1.5 kb position. Finally, I'm going to discuss further optimization steps and pitfalls in using this system for our synthetic locus.

The Principle of Recombination-Mediated Cassette Exchange.

Recombination-mediated cassette exchange is an alternative to the sequential integration-excision approach that we described in Chapter I and II by using a double-reciprocal crossover via a pair of incompatible recombination sites. In essence, if we wanted to convert our previous insertion-excision approach into a RMCE-strategy, we would only have to move one of the FRT sites from the targeting vector to the landing pad. This was done in the early 2000s and has been coined “dual-RMCE⁸” (Lauth, 2002). Conceptually the same happens, the plasmid is integrated and the backbone removed. The difference, while in the integration-excision approach only one site can integrate the plasmid, the RMCE approach allows any of the two recombination sites on the landing pad to integrate the plasmid and the remaining pair will delete the backbone. This should

⁷ Efficiencies multiply! Assuming 7% positive clones after selection for Cre and 83% after Flpo treatment, the final efficiency is ~6% (Table 1, Table 2). This means that we need to pick ~86 clones to obtain 5 positive clones. If we enrich in-between, almost every single clone surviving GCV treatment will have the backbone removed. In this way we can pick and analyze lots of clones.

⁸ The cassettes are “froxed”, based on “floxed” but with one loxP exchanged to FRT.

increase the efficiency, since integration is the rate-limiting step (Baer and Bode, 2001; Low et al., 2022).

The concept of RMCE was first implemented with Flp and Cre in the early 90s (Bouhassira et al., 1997; Schlake and Bode, 1994). Its advantage was that one could first place a landing pad with positive and negative selection cassettes randomly into the genome and subsequently, after exchange to the sequence of choice, select for the absence of the landing pad. Therefore, no selection cassette is left behind that could affect the expression of the transgene (Chen et al., 2004). Efficiencies without antibiotic selection were seldom reported, possibly due to its limited importance initially (Bouhassira et al., 1997; Karreman et al., 1996; Lee and Saito, 1998; Schlake and Bode, 1994). Rather they were reported as the number of drug-resistant colonies containing the correct integration/exchange product, which is usually a convolute of transfection efficiency, recombination efficiency and selection efficiency. Still, for Cre-based RMCE approaches most of the clones recovered after selection did not contain the correct exchange and (if measured or estimated) the recombination frequencies did not reach or were close to 1% without selection (Feng et al., 1999; Kameyama et al., 2010; Kolb, 2001). The cross-reactivity of loxP variants being an additional problem (Feng et al., 1999; Kolb, 2001; Langer, 2002; Lee and Saito, 1998), hope for more efficient approaches came from another class of recombinases, the serine recombinases.

Regulating Directionality of Recombination via Serine Recombinases.

Next to the multiple classifications we encountered in Chapter I, recombinases can also be divided into regulated and unregulated recombinases. While Cre and Flp have no control over whether they catalyze insertion or excision, requiring sophisticated mutational systems (lox66/71), the founding member of the tyrosine recombinases, λ -Int, uses the attP/B system to exert control over insertion and excision of its host genome (Reviewed in Yarmolinsky and Hoess, 2015). Unfortunately, λ -Int is not useful for genome engineering, due to its complexity, including large attachment sites, requirement for accessory factors and supercoiling constraints. The introduction of serine recombinases such as ϕ C31 and Bxb1 combined the best of both worlds (Stark, 2014):

1. Functionally, they use the same directionality system as λ -Int. They recombine 2 different attachment sites attP and attB to form attL(attBP') and attR(attB'P). The reaction is unidirectional because attL and attR can only recombine in the presence of specialized recombination directionality factors (RDF) (Stark, 2014).
2. They are as simple as Cre and Flp in their requirements, having no supercoiling constraints, catalyzing different types of reactions (insertion, excision, inversion) and requiring no additional factors and therefore also having very short attachment sites (Stark, 2014).

Mechanistically (Reviewed in Fogg et al., 2014), serine recombinases differ from tyrosine recombinases by attacking all 4 strands simultaneously with the catalytic serine leading to a double-strand break. Then the two dimers rotate 180° relative to one another and base-pairing of a central dinucleotide is checked. If correct, a second round of transesterification takes place to ligate the ends. Otherwise Bxb1 will rotate again and re-ligate the original strands (Fogg et al., 2014). Interestingly, the 2-bp dinucleotide is

functionally analogous to the 8-bp spacer of Flp and Cre, being the only requirement for directionality (Ghosh et al., 2006, 2003). That means, even though, unlike attP or FRT sites, attP and attB sites differ substantially in their sequence, they are functionally symmetrical (Ghosh et al., 2003). Additionally, because Bxb1 forms a synapsis in both the parallel and antiparallel direction (Cre and Flp do so only in antiparallel direction (Ennifar, 2003; Grainge et al., 2000)), a non-palindromic spacer is required for maintaining directionality (Ghosh et al., 2003). The central dinucleotide also allows the synthesis of orthogonal systems. For Bxb1, next to the wildtype “GT” dinucleotide, also “CT”, “GA” and “GC” have been shown to readily recombine in vitro, but “GC” lacks the directionality due to being palindromic (Ghosh et al., 2003). “GA” has been tested in vivo, showing even higher recombination efficiencies and high orthogonality with the wildtype “GT” (Jusiak et al., 2019; Low et al., 2022).

φC31 was initially used frequently, due to the presence of pseudosites in the mammalian genome, which allowed integration of transgenes without needing to place a landing pad beforehand (Chalberg et al., 2006; Keravala et al., 2006). However, this is a major disadvantage for site-specific integration. Bxb1 was initially reported to have no pseudosites in human or mouse cells (Low et al., 2022; Russell et al., 2006), consistent with its off-target integration being generally very low (Durrant et al., 2023; Low et al., 2022; Pandey et al., 2024). However, more recently 3 pseudosites targeted by attP have been found (Pandey et al., 2024). Because the target of attP is the shorter attB-site, pseudosites for attP are less likely and installation of attP landing pads is preferable. Bxb1 also outperforms φC31 in integration efficiency in diverse mammalian cells (Durrant et al., 2023; Jusiak et al., 2019; Xu et al., 2013). Additionally, apart from Bxb1, many serine recombinases are toxic for eukaryotic cells, possibly due to the ds-break intermediate that is created (Xu et al., 2013).

Thus far, only a handful publications reported the absolute integration efficiency of Bxb1. Both Duportet et al. and Jusiak et al. report around 10% transfection normalized integration efficiency when attP(GT) is placed into HEK293, HELA or CHO cells (Duportet et al., 2014; Jusiak et al., 2019). Jusiak et al. additionally find that the mutant variant attP(GA) is twice as efficient, leading to 20% integration efficiency (Jusiak et al., 2019).

Pandey et al. and Durrant et al. report efficiencies that are not normalized to transfection efficiency but are also in non-selective conditions (Durrant et al., 2023; Pandey et al., 2024). Pandey et al. report around 10-20% recombination efficiency in HEK293 cells. They also find that whether pre-installation of attB or attP is more efficient depends on the genomic locus (Pandey et al., 2024). Durrant et al. only report a maximum of 3% integration efficiency in HEK293 cells. Interestingly, they used the attB-site for the landing pad, not attP (Durrant et al., 2023).

Low et al. use an entirely different system, injecting the targeting vector into zygotes and comparing integration with cassette exchange. They find that RMCE is twice as efficient as integration. They also speculate that the increased efficiency for RMCE might stem from having two attachment sites that can be targeted for the initial integration step (Low et al., 2022).

Davis et al. and Matreyek et al. use Bxb1 for library integration and achieve around 5% integration efficiency when a Bxb1-expression vector is co-transfected. When Bxb1 is stably integrated on the landing pad, they achieve up to 20% integration efficiency (Davis et al., 2020; Matreyek et al., 2019).

Overall, using Bxb1 via an RMCE approach in our synthetic locus should allow us to reach unprecedented efficiencies in enhancer integrations, more than suitable for high-throughput applications.

Aims.

Our previous efforts to integrate enhancers at precise positions from a minimal promoter required Cre-mediated integration of the targeting vector and subsequent removal of the backbone via Flpo. Here I will design and test an updated strategy, using the highly efficient Bxb1 in combination with RMCE, converting the two-step enhancer integration-excision strategy into a single-step exchange. By exchanging a newly designed GFP-based landing pad with a strong enhancer at 1.5 kb, I will measure the exchange efficiency by gain of mCherry and loss of GFP (Fig. 11). Using an enhancer instead of another fluorophore will specifically increase mCherry expression only upon integration at the right position. Off-target or random integrations will not cause mCherry increase. Finally, I will discuss several problems and optimization steps that need to be tackled to move from single enhancer integrations to high-throughput approaches, surveilling thousands of enhancer-enhancer interactions.

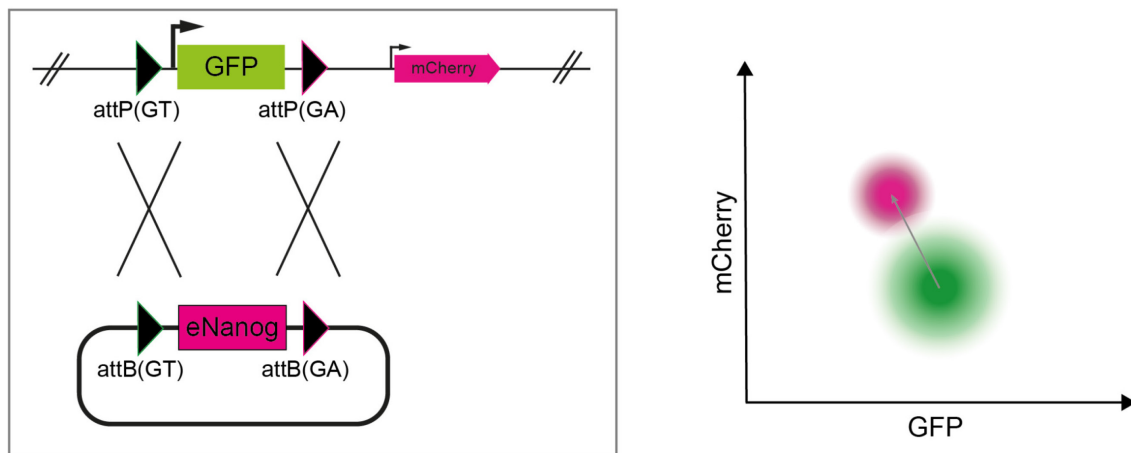


Figure 11: Overview of RMCE strategy. By exchanging a GFP-based landing pad with a strong enhancer GFP expression should be lost, and enhancer-mediated activation of the minimal promoter should lead to high mCherry expression.

Results.

Generating a Doxycycline-Inducible Bxb1 Cell Line (iBxb1).

A factor limiting recombination efficiencies can be the absence of sufficient recombinase or targeting vector in a given cell. To ensure that each cell expresses Bxb1 reliably, I generated a stable cell line containing doxycycline (dox) inducible Bxb1. The Bxb1-P2A-IRFP670 construct was a kind gift of Moritz Gaidt. A vector for PiggyBac-mediated insertion was generated, containing two terminal repeats (TR) flanking a Bxb1-P2A-IRFP670 expression cassette but no selectable marker (CB020-Bxb1- Δ puro Δ loxP, Fig. 2A). In this way, Bxb1-expressing cells can be enriched by gating for IRFP⁺ cells using FACS, while not occupying resistance genes unnecessarily.

To establish the inducible cell line (iBxb1), the cell line containing the synthetic construct was co-transfected with a vector expressing PiggyBac transposase and CB020-Bxb1- Δ puro Δ loxP. I induced Bxb1-IRFP expression for two days with doxycycline and sorted the population for IRFP-expressing cells (Fig. 2B). After a second round of induction, I received a final population containing >50% IRFP⁺ cells, which was single cell sorted in 96-well plates and genotyped (Fig. 2B, Fig. S9). 10 clones were picked for further analysis (Fig. 2C). The relevant parameters for finding a suitable clone are the mean expression level, the consistent inducibility and a low leakiness. The mean expression level should be moderately high, with low expression limiting the efficiency of recombination and high expression promoting toxicity. Based on mean expression level, 2 clones were chosen for further analysis (Fig. 2D). After withdrawal of doxycycline the IRFP levels dropped and were comparable to the parental cell line not transfected with Bxb1 (negative control). A second round of induction restored IRFP levels to its initial levels. In conclusion, a doxycycline inducible cell line was established (iBxb1-C4), which showed consistent inducibility and low leakiness.

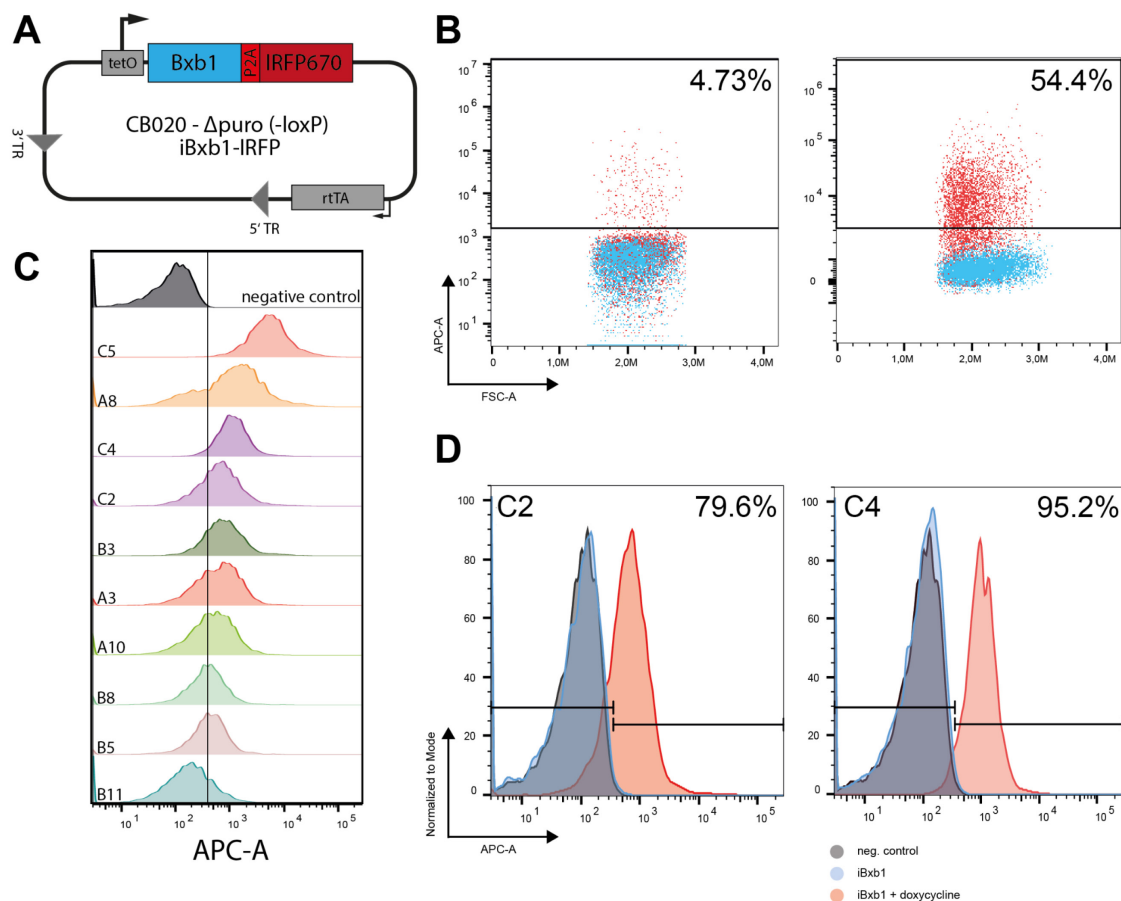


Figure 12: Establishment of iBxb1. A) Plasmid for integrating the inducible Bxb1 via PiggyBac. B) FACS-based enrichment of IRFP expressing clones. Left: First sorting. Right: Second sorting. C) IRFP expression of 10 expanded iBxb1 clones. D) Two selected clones. IRFP expression after doxycycline induction (red). Negative control corresponds to the parental cell line (#1, no Bxb1 integrated).

Generation of a Landing Pad Capable of Monitoring and Enriching Clones After RMCE.

Applying RMCE for high-throughput purposes requires enrichment of successfully exchanged clones. To perform negative selection, TK and/or GFP can be incorporated into the landing pad (LP). A decline in GFP expression after exchange can be used as indication for the exchange efficiency, given that most LP-cells express GFP higher than a negative control. I also considered using a destabilized GFP (GFPDD), since this would ideally show the exchange even earlier than the regular GFP.

To construct a landing pad that I could easily insert at the 1.5 kb position for testing, I used the pre-existing loxP66-containing targeting vector (CB231). However, instead of “targeting” another enhancer (as Chapter I and II) to the 1.5 kb position, I chose to integrate several different Bxb1 landing pads for testing, instead of relying on tedious CRISPR/Cas9 knock-in. An important decision was whether to place the attP-site or the attB-site into the genome. Recent publications achieving high integration efficiencies placed the attP in the genome and the attB in the targeting vector (Jusiak et al., 2019; Low et al., 2022). Comparing attP-LPs and attB-donors with attB-LPs and attP-donors showed that the relative performance is highly locus dependent (Pandey et al., 2024). However, due to its shorter length, there are several attB pseudosites in HEK293 cells,

of which at least one is regularly targeted by attP (Pandey et al., 2024). Taking these considerations into account, I placed 52-bp attP sites into CB231. Even though the minimal length for attP is 48-bp, two recent publications (Durrant et al., 2023; Ghosh et al., 2003; Jusiak et al., 2019) used a 52-bp variant, therefore I included 4 additional base pairs (Fig. 13A). Between the attP sites, two short random sequences were placed, both for technical reasons (allowing Gibson assembly) but also for designing specific primers later. Different GFP expression cassettes with or without TK were cloned and tested via transient transfection (Fig. 13B). I decided to use EGFP-T2A-TK, as it is both highly expressed and allows for antibiotic selection against “non-exchanged” clones. From here on the attP(GT)-EGFP-T2A-TK-attP(GA) construct is simply referred to as landing pad.

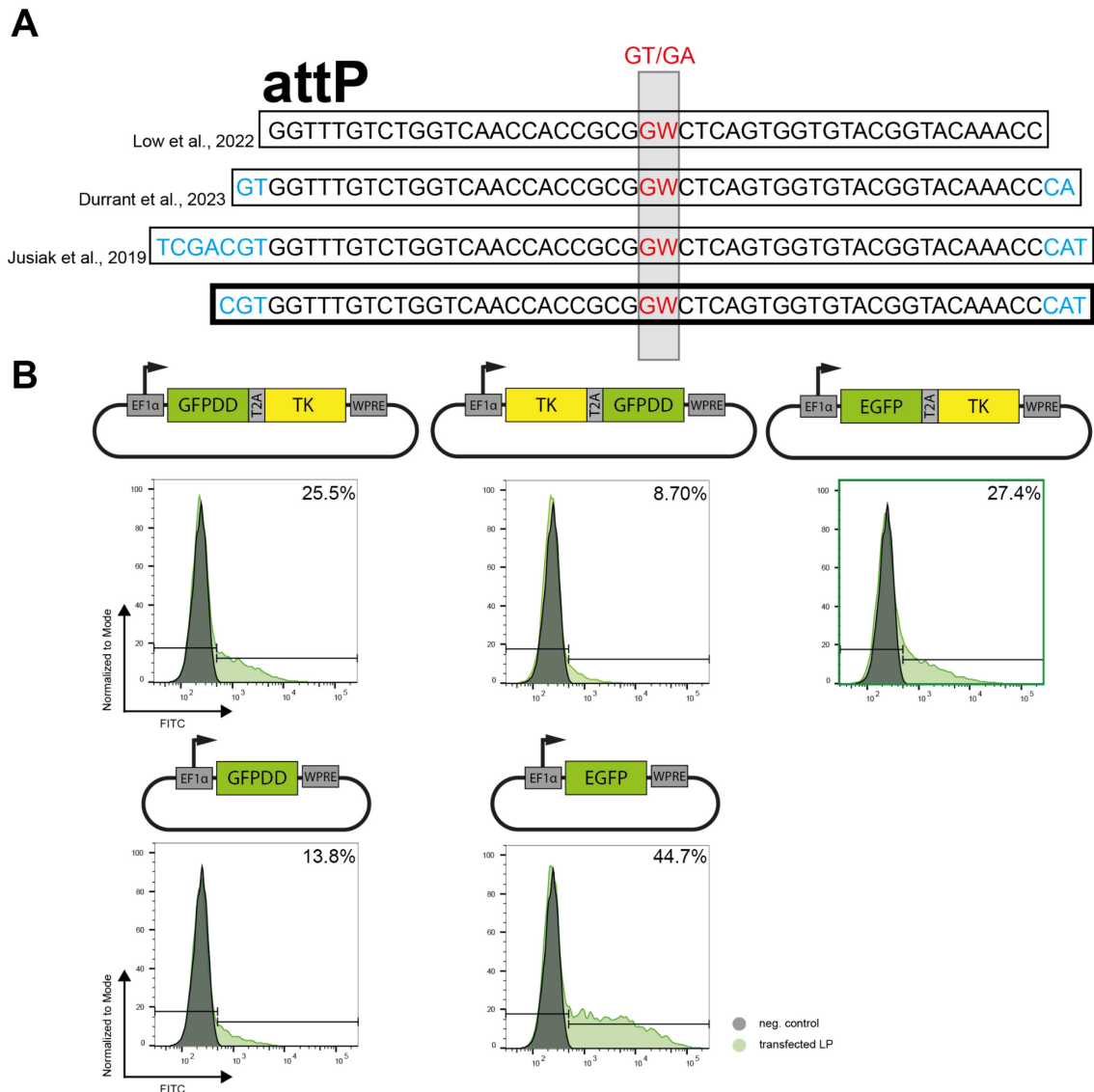


Figure 13: Cloning and testing of GFP-based landing pads. A) selection of attP sites for landing pads. Spacer in the middle is either GT or the mutant GA. The fourth sequence is used for constructing the landing pads. B) Transient transfection of the five landing pads generated.

Integration of a Landing Pad into iBxb1 for Testing RMCE Efficiencies.

Next, I used the pre-existing Cre-loxP system to precisely integrate the landing pad at 1.5 kb (Fig. 2). After integration of the LP via Cre and subsequent genotyping (Fig. S10A), 7 clones were analyzed twice for GFP expression, once in the +dox and -dox condition, respectively (Fig. 14A). Here the relevant parameter was not mean expression but the percentage of GFP positive cells relative to a negative control (iBxb1-C4). The higher the number of cells that can be clearly distinguished from GFP-expressing cells, the better we can enrich for the absence of the LP when we perform RMCE. The highest expressing clone was H2, being ~60% GFP positive (Fig. 14A). iBxb1-C4H2 was then further used for removing the backbone.

Backbone removal is required for testing, since we observed that enhancers sometimes fail to activate mCherry if the backbone is located in the genome, probably due to the highly expressed resistance genes. Because both the LP and the backbone express TK, negative selection cannot be used for backbone removal. I added an IRFP expression cassette to the Flpo-vector and sorted transfected cells based on IRFP into 96-well plates. 6 genotyped clones were expanded and tested for GFP expression (Fig. 14B, Fig. S10A). The highest expressing clone (D12) was selected for testing.

In conclusion, I established iBxb1-C4H2D12, which is capable of expressing Bxb1 at high levels upon dox-induction, expresses high levels of GFP to indicate a successful exchange through a sharp drop of GFP expression and allows selection for the absence of the landing pad, hence presence for the sequence of interest.

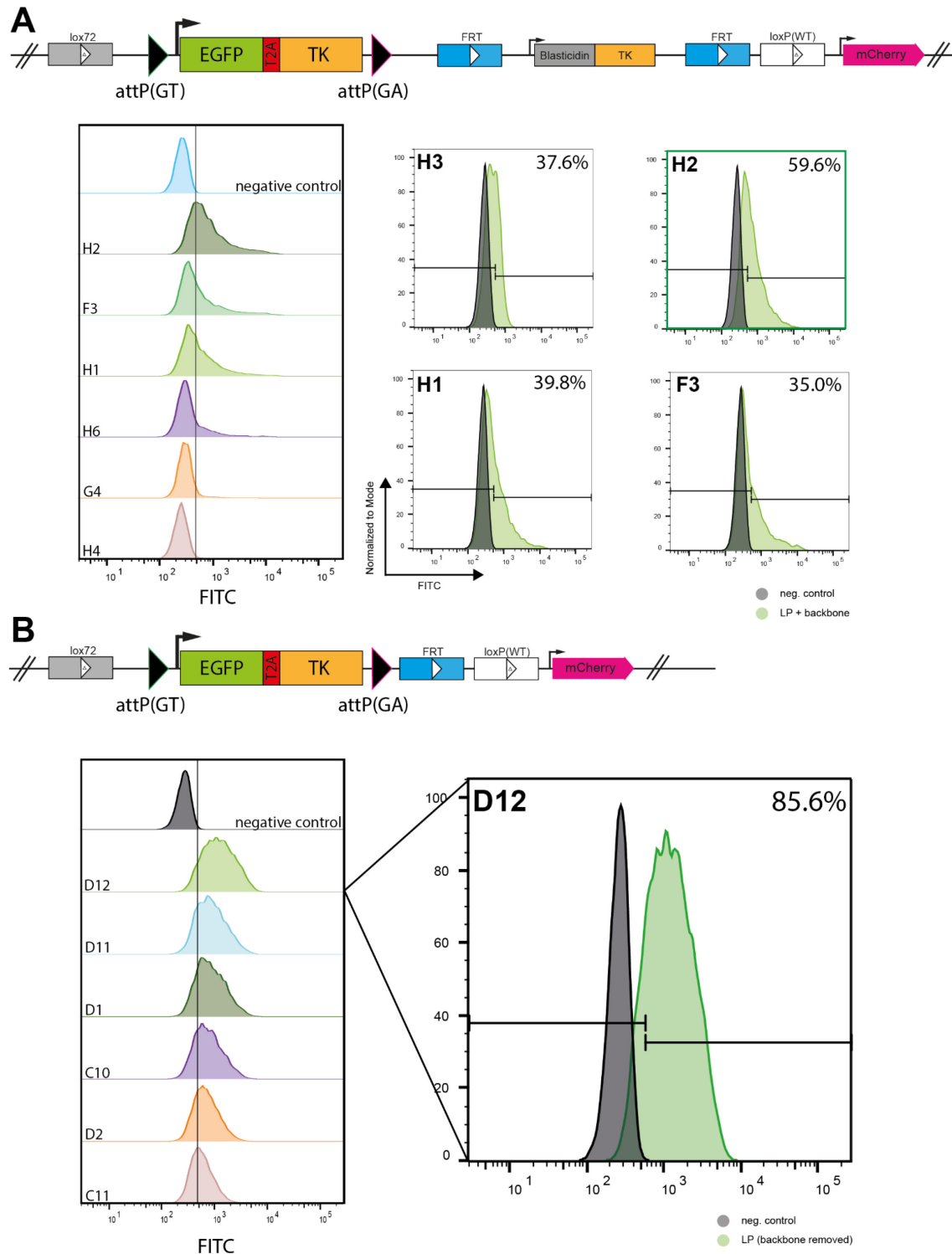


Figure 14: Integration of LP into synthetic locus of iBxb1 at 1.5 kb. A) Analysis of GFP expression after integration of plasmid carrying the LP. Of six clones, four were selected for a second round of analysis. The negative control is iBxb1-C4 (without LP). B) Analysis of GFP expression of the final iBxb1-LP cell line (after backbone removal of integrated plasmid). Final clone is shown (D12). Negative control is iBxb1-C4 (without LP).

Construction of a Targeting Vector for RMCE.

The main requirements for the optimal targeting vector were a small size and two attB(GT/GA) sites flanking a restriction site for integration of a sequence of interest. Instead of the minimal 38-bp attB sequence, I used a longer 47-bp version to make sure it is comparable to high efficiency systems previously published (Jusiak et al., 2019; Low et al., 2022). The vector backbone was based on pGEM-T. For purposes of testing the exchange efficiency, I integrated the proximal enhancer of *Nanog* (*eNanog*) between the attB sites (Fig. 15B). *eNanog* is a strong enhancer and around 90% of cells are clearly distinguishable based on mCherry (Chapter I, Fig. 15C). Additionally, only if *eNanog* is integrated close to the minimal promoter, mCherry expression will increase. This distinguishes it from using another fluorophore (e.g., BFP), which would increase its expression even in off-target positions. A promoter trap method is not suitable for our purposes, since a promoter close to the integrated enhancer could influence mCherry expression, interfering with measuring enhancer strength. Together with the loss of GFP upon exchange, this should enable us to visualize this subpopulation with flow cytometry.

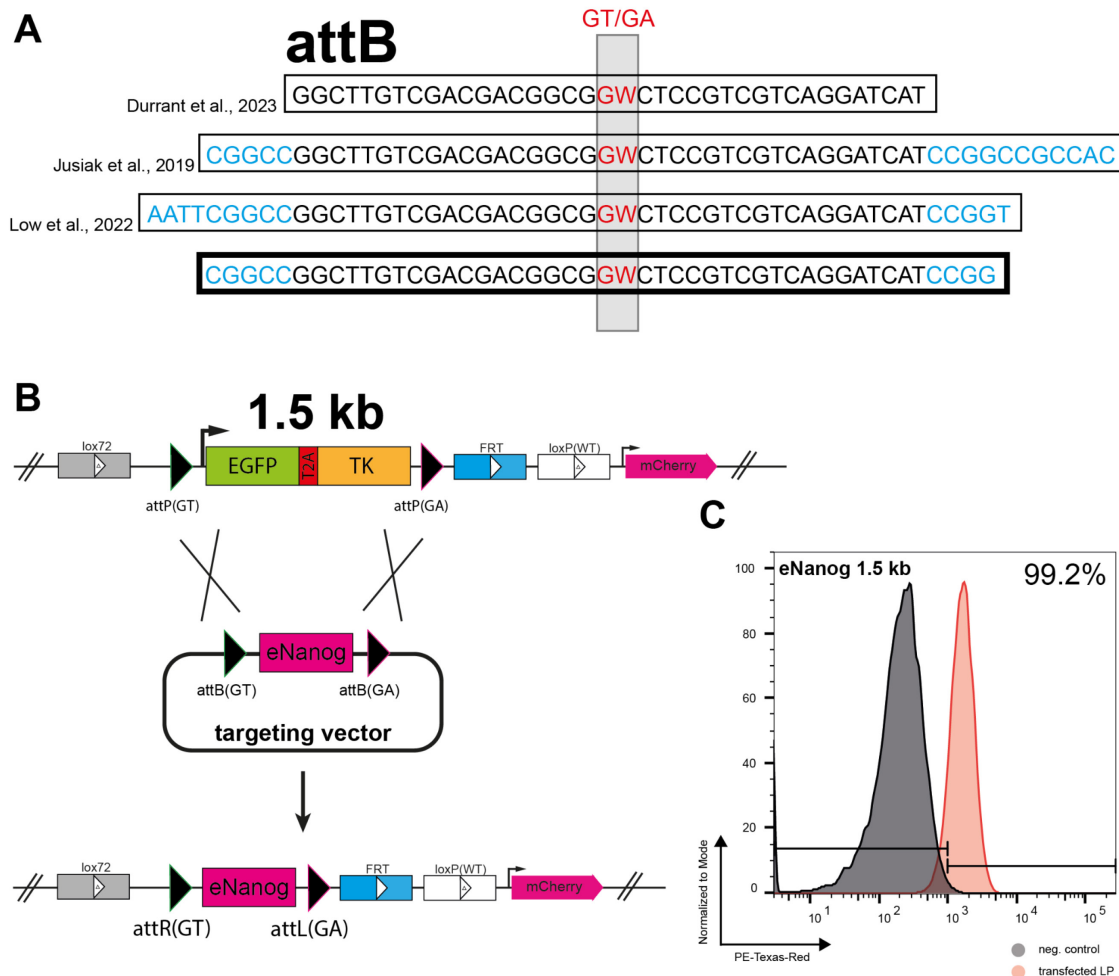


Figure 15: Construction of targeting vector for RMCE. A) Selection of attB sites for targeting vector. The fourth sequence is used for constructing the targeting vector. B) Scheme of exchange reaction for measuring the efficiency. C) mCherry expression in a cell line with *eNanog* at 1.5 kb. Almost 100% of cells should be mCherry⁺ and indicate exchange reliably.

Cassette Exchange Is Efficient in Non-Selective Conditions Despite Unexpected Rearrangements.

I hypothesized that exchanging the EGFP-T2A-TK cassette with *eNanog* allows for tracking the exchange via decreased GFP and increased mCherry expression (**Fig. 11**). Bxb1 expression was induced two days prior to transfection of the targeting vector and terminated two days after transfection by withdrawal of doxycycline to limit potential toxic effects of long-term expression. 2 days after transfection, a subpopulation containing around 0.7% of the cells is clearly mCherry positive and GFP negative (Fig. 16A). This subpopulation further increased to 1.2-1.4% at day 4.

One pitfall was that GFP expression was reduced drastically at day 2 even in non-transfected cells. I postulated that part of the reduction in GFP was caused directly by Bxb1 expression, not by Bxb1-mediated exchange. Expectedly, supplying doxycycline to non-transfected cells did not affect mCherry expression but reduced GFP levels transiently (Fig. S11A). GFP expression recovered upon withdrawal of doxycycline. This could be caused by Bxb1 binding to the attP sites, even in the absence of the targeting vector and possibly blocking the transcription machinery.

During the first exchange experiment the measured IRFP intensity was almost 2-fold higher than previous measurements without transfection (Fig. S11B). To investigate whether this is caused by transfection, I compared the IRFP expression throughout the three FACS measurements in a second exchange experiment. Indeed, IRFP expression was again boosted 1.5-fold at day 2 but recovered to expected levels at day 4. I concluded that the stressful conditions caused by transfection lead to temporarily elevated fluorophore intensities.

The prokaryotic backbones of most plasmids are known to trigger an immune response in eukaryotic cells and the size can limit the transfection efficiency (Vandermeulen et al., 2011). To test whether the absence of the backbone might boost the exchange efficiency, I transfected linear dsDNA (gel-purified PCR product of attB(GT)-*eNanog*-attB(GA)). This led to a severely reduced number of mCherry+GFP- cells and was not very efficient, possibly due to insufficient purity after gel purification.

To confirm that the correct exchange product has formed, I used the transfected population of day 6 and sorted the mCherry+GFP- fraction and mCherry-GFP+ fraction into separate tubes. Genotyping showed that only the mCherry+GFP- fraction contained the correct exchange product while only the mCherry-GFP+ fraction contained the intact landing pad (Fig. 16B). Both junctions of *eNanog* were sequenced and showed the successful exchange including the recombined attR and attL sites (Fig. 16C). Surprisingly, the strongest band in both the mCherry+ and mCherry- fraction corresponded to an unexpected band (Fig. 16C). Re-analysis of primer binding sites and subsequent sequencing of the PCR product identified this band as the backbone of the targeting vector (Fig. 17, see "*Negative selection enriches successful recombinants.*").

The transfection efficiency being ~60%, the overall exchange efficiency is around 2.0-2.3%. 70% can be excluded based on GFP levels, enabling efficient FACS-based enrichment.

1. If the exchange efficiency is around 1.4%, why did we obtain only 48 clones (~0.012%⁹)? Are the selection conditions too harsh?
2. If 50% of the clones did not exchange the LP for *eNanog*, why did they survive GCV treatment, especially considering that GCV treatment usually kills most of the GCV-sensitive cells and apparently even many of the resistant ones?

Surprisingly, the clones not corresponding to the fully exchanged *eNanog* were found to contain the backbone of the targeting vector (Fig. 17B). The presence of these “backbone clones” could have several reasons:

1. **Only one site recombined, integrating the whole targeting vector.** The possible rearrangements are shown in Fig. 17A. This was the most likely option, considering that the full *eNanog* exchange and the “backbone clones” are (almost) mutually exclusive (Fig. 17B). Additionally, the “backbone clones” also have consistently low but above background GFP expression, while the full exchange has no GFP expression (Fig. 17D). This would be consistent with a single integration, since the origin of replication of the backbone would be placed right in front of the GFP-T2A-TK promoter, which could lead to interference. If indeed the integration occurred only partially, then repeated induction of Bxb1 could possibly lead to additional recombination of the remaining attachment sites. However, when we induced Bxb1 again, mCherry expression was not rescued in any of the “backbone clones” (Fig. 17E). Furthermore, I could not amplify any of the expected constructs that would be expected from recombining only attB(GT). Overall, the mutual exclusivity of full *eNanog* exchange or “backbone clone” could also be a simple correlation, originating from the fact that each cell only receives a limited amount of targeting vector. If this vector has recombined, there might not be another chance for a second integration. The correlation between “backbone clone” and low GFP expression could simply be a selection bias, since every “backbone clone” that expresses high GFP also expresses high TK and will die in GCV presence. However, this would mean that a subset of cells containing the LP are irreversibly silenced, since GFP expression also did not recover in the “backbone clones” several days after the selection medium was removed. Irreversible silencing is unlikely, since it would lead to a decreasing pool of GFP-expressing cells, which we did not observe.
2. **Integration of targeting vector through attB pseudosite.** However, at this point no attB pseudosites have been published. This is precisely why I chose to integrate attP as the acceptor in the landing pad.
3. **Random integration of the targeting vector.** Independent random integrations would lead to different rearrangements of the targeting vector. That the same $\frac{1}{5}$ ¹⁰ of the targeting vector can be amplified in 24 clones hints towards a specific integration. Random integrations are also expected to occur at a lower frequency.

⁹ Of the 200,000 cells transfected, $\frac{1}{2}$ were plated on a 10 cm dish 24h after transfection (2 cell-cycles). This equals ~400,000 cells. At an exchange efficiency of 1.5% we should get 6000 GCV resistant clones. 48 clones correspond to 0.012% exchange efficiency. Essentially, we lose 99.2% of the clones.

¹⁰ The targeting vector is 4225 bp. The amplified fragment is 838 bp long.

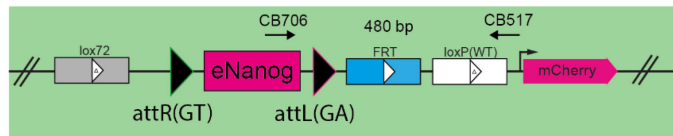
4. **The presence of a second landing pad.** A second LP integration could happen randomly or via a loxP pseudosite. Subsequent treatment with Flpo could lead to a plethora of potential rearrangements. Since we selected a high GFP expressing cell, a bias towards multiple integrations was possible. Additionally, I got an unexpected PCR product that proved to be part of the backbone of the landing pad (not shown). Since I know that the backbone of the landing pad at 1.5 kb is clearly removed, based on multiple PCRs and sequencing, this must correspond to an additional integration, even though surprisingly, every single clone is blasticidin-sensitive. Additionally, I could amplify a band corresponding to parts of GFP in multiple clones (including the fully exchanged *eNanog* clones that do not express GFP). A clear pattern was not observed.

All of the above, more probable than not, argue for the presence of a second landing pad or some unknown rearrangement at the 1.5 kb landing pad. Future Cre-based landing pad integrations need to be carefully checked for the presence of off-target effects.

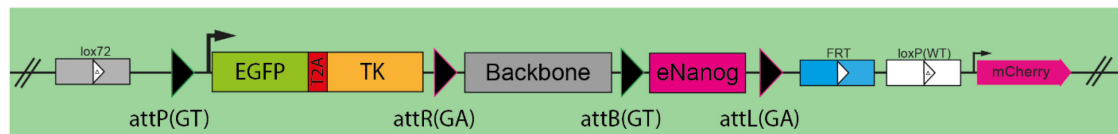
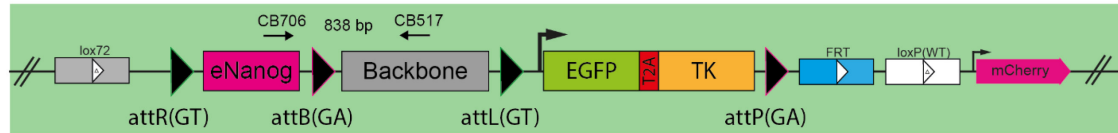
The death of most clones after GCV selection was not surprising, as we have already encountered similar issues after Flpo-based backbone removal. A potential effect of GCV selection on cells that should be resistant is likely (See Discussion), however, a systematic analysis was not feasible, since I was unable to determine whether multiple TKs were potentially present in the different cell lines produced. This should be reconsidered after a cell line with a single LP was generated.

A

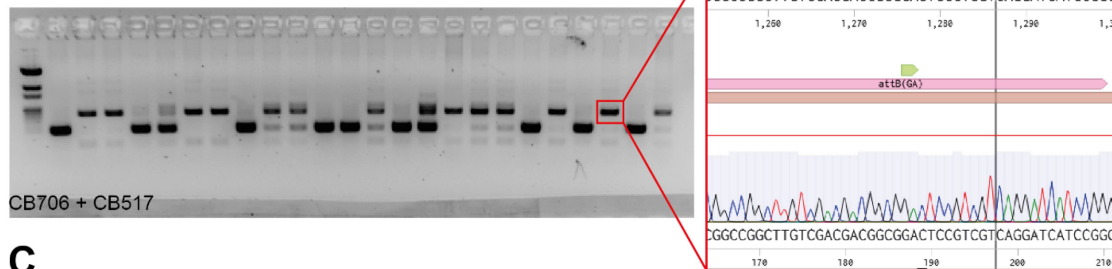
Full exchange:



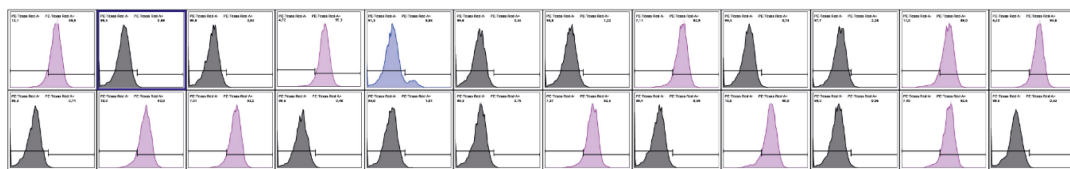
Single integrations:



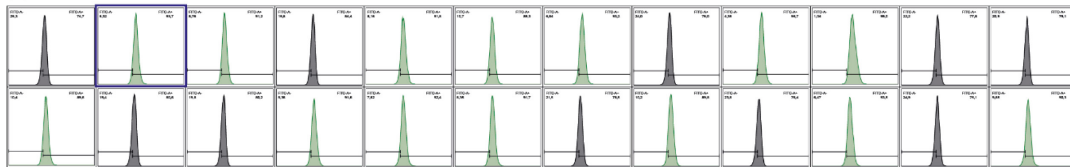
B



C



D



E

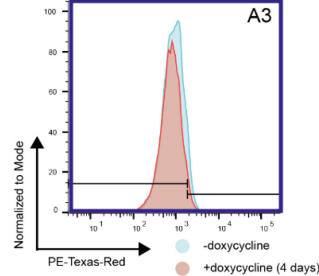


Figure 17: Single clones after RMCE and ganciclovir selection. A) Expected constructs for full exchange or single exchange. B) Genotyping of single clones after RMCE and ganciclovir selection. Lower band corresponds to full exchange. Upper band corresponds to backbone of the targeting vector. C) mCherry expression of clones seen in B). The peaks colored in pink show high mCherry expression as expected for *eNanog* integration at 1.5 kb. D) GFP expression of clones seen in B). GFP expression is downregulated in

all of the clones, but the ones colored in green have slightly elevated levels. E) Second round of doxycycline treatment for the clones A2 and A3 (only A3 shown), which contain the backbone of the targeting vector. mCherry is not upregulated.

Discussion.

Bxb1 is a frequently used recombinase in mammalian cells, outperforming most other recombinases in integration efficiency, limited off-target effects and low toxicity (Durrant et al., 2023; Jusiak et al., 2019; Low et al., 2022; Pandey et al., 2024; Xu et al., 2013). Here I employed Bxb1 together with a cassette exchange strategy, which enables us to integrate enhancers in a single step, reducing the time for enhancer integrations about 50% and allowing us to perform high-throughput experiments. While our previous sequential Cre/Flpo-based integration-excision strategy could in principle be adapted to a similar timescale, Bxb1 should ideally increase the efficiency of the rate-limiting plasmid integration step (Baer and Bode, 2001), while the RMCE strategy increases the probability of integration by placing two recombination sites into the genome. Despite high integration/exchange efficiencies published, our RMCE efficiency still falls short by about an order of magnitude, reaching 1.4% in non-selective conditions. Gating on GFP expression allows enrichment to ~4% (not normalized to transfection efficiency). With an average death rate of 2/3 after single-cell sorting into 96-well plates, one would get around 1-2 positive clones per 96-well plate. For high-throughput approaches, assuming 100 enhancers with 100x coverage, you would need 10,000 viable clones. At a 1.4% efficiency, this would require sorting of ~700,000 cells and you would end up with 10,000 exchanged clones in 230,000 GFP negative cells. For larger libraries, e.g., 30,000 enhancers at 20x coverage, you would need to sort 40 million cells, and you will run into limitations due to sorting speed.

While switching from FACS-based enrichment to magnetic beads-based enrichment is an option (Durrant et al., 2023), negative selection would already greatly improve our system. After TK selection I obtained 48 clones, of which 50% seemed to have the full *eNanog* exchanged, while the other 50% had some rearrangement or off-target integration, possibly due to the integration of a second landing pad (Fig. 17). 24 positive clones are sufficient for single enhancer integrations, and I could additionally integrate a negative selection gene into the backbone of the targeting vector to remove all targeting vector integrations that did not fully exchange. For high-throughput integrations, 48 surviving clones, which means losing 99% of the 1.4% properly exchanged cells, is clearly not sufficient. Additionally, each clone that has the backbone of the targeting vector integrated, is a lost opportunity for integration at the right position. The consistent PCR band and the high percentage of “backbone clones” clearly argues against random integration and I found evidence for the presence of a second landing pad, which possibly sequesters many of the targeting vectors, limiting our efficiency measured based on mCherry and GFP expression. The supply of the targeting vector being one of the most critical steps explains why our efficiency is still very low. It will be crucial to screen for off-target integrations of the landing pad. ddPCR could be used to determine the GFP copy numbers (Inniss et al., 2017). Alternatively, a form of one-sided PCR called UDiTaS could be used, in which the genome is tagmented by Tn5 (Giannoukos et al., 2018). A LP-primer in combination with a primer annealing to the tagmentation site can reveal the

position of the LP (Durrant et al., 2023; Pandey et al., 2024). Eliminating off-target LPs could boost the efficiency at least 2-fold and possibly beyond, reaching 5% transfection normalized recombination efficiency. At the same time, high GFP-expressing integrations will always have a bias towards additional off-target integrations. This bias can be reduced by elevating GFP expression levels of the landing pad, e.g., by exchanging the EF-1 α promoter to the CAG promoter (Dou et al., 2021; Wang et al., 2017). Additionally, the WPRE should be exchanged to a poly(A), to avoid read-through into the minimal promoter driving mCherry.

At the same time, we will need to evaluate whether GCV-treatment might be too harsh. Already for the Cre/Flpo system we repeatedly observed only a handful or even no clones after Flpo treatment, despite backbone excision being a very efficient process and high transfection efficiency. It might be recommendable to postpone GCV selection from day 2 after transfection to day 3 or 4, since residual TK could still be around. On the other hand, it might be reasonable to evaluate whether the presence of ~99% TK expressing cells could influence the survival of the 1% non-expressing cells, e.g., through taking up modified GCV from dead cells through the medium. Exchanging the medium more frequently, plating the cells in lower density or completely switching from TK to iCaspase9 (Zhou et al., 2015) or DTV (Araki et al., 2006) could resolve this issue.

PCR-amplified libraries were reported to work very well with Bxb1 (Durrant et al., 2023). Despite that, dsDNA fragments of the targeting vector (without backbone) decreased exchange efficiency drastically. Switching from gel-extraction to a PCR purification kit could increase the efficiency. To improve the overall recombination efficiency, we should further consider switching from wild-type Bxb1 to the previously published evoBxb1, which would allow us to boost integration efficiency up to 2.4-fold (Pandey et al., 2024). Additionally, supplying higher amounts of targeting vector is crucial. Durrant et al. showed that when comparing the electroporation of 1 μ g with electroporation of 3 μ g or 3 μ g twice in succession, the integration efficiency increased almost 4-fold (Durrant et al., 2023).

While all these improvements might not combine additively, they should allow us to hit the 10% exchange efficiency mark. In this way we should be able to screen tens of thousands of enhancers in their ability to bridge genomic distance and cooperate with other elements.

Material and Methods.

ESC Maintenance.

Murine ESCs were cultured in base medium consisting of EmbryoMax® DMEM/F12 (with L-Glutamine, without HEPES), 4 mg/mL AlbuMAX™ II Lipid-Rich Bovine Serum Albumin (GIBCO™), 1x MACS® NeuroBrew®-21 (with Vitamin A), 1x MEM NEAA (GIBCO™), 100 U/mL Penicillin-Streptomycin (GIBCO™), 1x Sodium Pyruvate (GIBCO™) and 55 mM 2-Mercaptoethanol (GIBCO™), which was supplemented with 3.3 µM CHIR-99021 (MedChem Express), 0.8 µM PD0325901 (Selleckchem) and 10 ng/mL hLIF (provided by the VBCF Protein Technologies Facility, <https://www.viennabiocenter.org/facilities/>), referred to as 2i/LIF medium. mESCs were maintained on tissue culture dishes which were first coated with 6 µg/mL of Poly-L-ornithine hydrobromide (Sigma-Aldrich) in 1x PBS for 1 hour at 37°C and subsequently with 1.2 µg/mL of Laminin from Engelbreth-Holm-Swarm murine sarcoma basement membrane (Sigma-Aldrich) in 1x PBS for 1 hour at 37°C. 96-well plates were usually coated with 10 µg/mL Human Fibronectin Protein (CF, Bio-Techne GmbH) in 1x PBS for 1 hour at room temperature. Cells were passaged every other day by incubating the cells in Trypsin-EDTA solution (Sigma-Aldrich) for 5 minutes, adding 1:1 volume of 2i/LIF containing 10% Fetal Bovine Serum (FBS, Sigma-Aldrich), centrifuging the cells for 5 minutes at 100 g (rcf) and resuspending them in 2i/LIF.

Transfection.

200,000 cells were seeded in a 6-well plate. 3-4 hours later, for one well, 250 µL of Opti-MEM™ (Thermo Fisher Scientific) was incubated with 3 µL of Lipofectamine® 2000 Transfection Reagent (Thermo Fisher Scientific) for 5 minutes. Simultaneously, 250 µL Opti-MEM™ was added to a maximum of 1.5 µg of DNA. After the 5 minutes, the two volumes were combined and incubated for 15 minutes. The 500 µL were added dropwise to the cells. Medium was exchanged 5-6 hours after the transfection.

PCR.

Every PCR was performed with Phusion® High-Fidelity DNA Polymerase (NEB). For extractions 4 x 40 µL reactions were used. A 40 µL reaction consisted of 8 µL Phusion Green HF Buffer (Thermo Fisher Scientific), 1 µL dNPTs, 1 µL 10 µM forward primer, 1 µL 10 µM reverse primer, 1 µL of DNA template, 2 µL of polymerase, 2 µL of DMSO and 24 µL of H₂O. For genotyping or colony PCR the reaction was downscaled to 20 µL, while maintaining 1 µL of DNA.

Restriction Digest.

For preparative plasmid digestion around 2000-4000 ng was used in a 100 µL reaction, consisting of 2 µL of enzyme 1, 2 µL of enzyme 2 (optional), 10 µL of 10x rCutSmart and H₂O. Restriction digests were generally performed overnight, otherwise uncut vector might still have been present, which could have led to false positive clones after e.g., ligation and transformation. A negative control consisted of a 20 µL reaction with 2 µL of 10x rCutSmart, 200 ng of plasmid and H₂O. The next day, 10 µL of each reaction was

mixed with 6x loading dye and analyzed on a gel. If the digest was successful, the rest of the plasmid was purified with the PCR-purification kit, since this resulted in higher cloning success, possibly due to increased purity.

Gibson Assembly.

The insert was amplified via PCR and primers contained ~20 bp homology arms for the destination vector. If the PCR was performed on a different plasmid, the PCR product was purified via gel extraction to prevent the presence of any parental vector. If amplification happened from gDNA, a PCR purification kit was used. The destination vector was cut via restriction digest and purified with the PCR purification kit or gel extraction. The 5x isothermal buffer was self-made and consisted of 25% PEG-8000, 500 mM Tris-HCl (pH = 7.5), 50 mM MgCl₂, 50 mM DTT, 1 mM of each dNTP, 5mM NAD and H₂O filled to a total volume of 3 mL. The final Gibson-Mix consisted of 2 µL of 5x isothermal buffer, 0.2 µL T5 exonuclease (0.008 units), 0.125 µL Phusion DNA polymerase (0.025 units) and 0.5 µL Taq DNA ligase (2 units), which was ~ 3 µL in total volume. 50 ng of backbone and 1:1 insert (molar ratio) were added and filled up to 10 µL with H₂O. The reaction was incubated at 50°C for 30 minutes and half of the volume (5 µL) was transformed.

Annealing Oligos.

Oligos were annealed by using 0.5 µL of T4 PNK, 1 µL of T4 Ligase Buffer and 1 µL of 100 µM oligos each. The reaction was filled up to a total volume of 10 µL with H₂O. In the thermocycler the mixture was first incubated at 37°C for 30 minutes, then fully denatured at 95°C for 5 minutes and then slowly annealed by decreasing the temperature continuously for 15 minutes to reach 25°C.

Ligation.

For Ligation 1 µL of T4 Ligase, 2 µL of T4 Ligase Buffer and 100 ng of backbone were mixed. If the insert was two annealed oligos, 1 µL of the annealing reaction was added. If two inserts (each two annealed oligos) were ligated with the backbone, 0.5 µL of the annealing reaction was added each. If the insert was a purified PCR product, a 3:1 molar ratio (insert:backbone) was used. The reaction was filled up to 10 µL with H₂O and incubated either for 2 hours at 16°C or 4°C overnight. Half the volume (5 µL) was transformed.

Transformation of Bacteria.

100 µL of competent DH5α cells were thawed for 5 minutes on ice. Afterwards the DNA was added (maximum 5 µL). After 30 minutes of incubation a heat shock was performed at 42°C for 1 minute. The bacteria were recovered for 2 minutes on ice and 800 µL of LB was added. Then the bacteria were shaken at 37°C for one hour, spun down and resuspended in a smaller volume to plate it on ampicillin plates, which are incubated overnight at 37°C.

Analysis of Flow Cytometry Data.

Analysis via flow cytometry was performed with BD LSRFortessa™ Cell Analyzer (BD Life Sciences - Biosciences). For sorting Cytex Aurora CS™ Cell Sorter (Cytex® Biosciences) or BD FACSMelody™ Cell Sorter (BD Life Sciences - Biosciences) were used. FCS files were analyzed with FlowJo™ software (BD Life Sciences – Biosciences, version 10.5.3). The mean intensities were extracted. The #1 cell line (synthetic locus with basal mCherry expression) was used for normalization, usually by subtracting (for more details see Supplementary Note 1). Barplots with integrated scatterplots were made with Python 3.10.10. The error bars show the 95% confidence interval. Fig. 4B shows significance for being different from background (#1) via an independent two-sample *t*-test. For the expected additive value for dual enhancer integrations, the mean of the 75 kb position and the mean of the 25 kb position were added for each replicate and the mean was calculated. Significance in being different from the additive model in Fig. 6 was calculated with an independent two-sample *t*-test, staying open for changes in either direction (subadditivity or superadditivity). For Fig. 9C there was only one background value (one replicate). Therefore, a one-sample *t*-test was performed. For Fig. 9D an independent two-sample *t*-test was used. Fig. 5C, Fig. 7, Fig. S3 and Fig. S6 were made with GraphPad Prism 10.1.0 for Windows (GraphPad Software, Boston, Massachusetts USA, www.graphpad.com).

Cre/Flpo-Mediated Integration of Enhancers at Synthetic Locus.

Enhancers are amplified from the CAST/EiJ mouse genome (Chapter I) and integrated into the targeting vectors via Gibson assembly (CB200, CB201 and CB231 for 75 kb, 25 kb and 1.5 kb). To integrate the enhancers into the different genomic loci, 200,000 cells (#1 with PCP) were seeded in 6-well plates and co-transfected with 800 ng of targeting vector and 200 ng of Cre-tdTomato plasmid. After the successful expression of the fluorophore was checked the next day, cells were expanded to 10 cm dishes to facilitate picking individual clones after selection. 2 days after transfection 2i/LIF was exchanged with 2i/LIF containing puromycin (2 µg/mL), hygromycin (400 µg/mL) or blasticidin (10 µg/mL), depending on the targeting vector. Selection medium was exchanged every other day. Once most of the cells died and individual colonies were visible (~1 week of selection), colonies were picked and genotyped. Clones containing the integrated targeting vectors were expanded. For dual enhancer integrations, the second enhancer is integrated at this point and selection is performed for both positions, e.g., puromycin and hygromycin for 75 kb and 25 kb. To remove the plasmid backbone, 200,000 cells of a positive clone were seeded and transfected with 2 µg of a Flpo-expressing plasmid (CB233). Cells were expanded 1 day after transfection and selection with 500 ng/mL ganciclovir (Invivogen) was started 2 days after transfection. The selection medium was exchanged every other day and colonies were picked after 5-9 days of selection. Genotyping for presence of enhancer and absence of backbone was performed and positive clones were expanded.

Site Directed Mutagenesis.

Mutagenesis primers for amplifying the whole targeting vector containing the *Map4k3* E1 (CB254) were designed. The 3 and 10 mismatches necessary for 3OPT and allOPT,

respectively, were distributed among the two primers (CB936, CB937, CB938, CB939). The PCR was performed using the Q5®-Polymerase as a 2x 40 µL reaction. Afterwards parental plasmid was removed by adding 1 µL of DpnI directly to the PCR reaction and incubated overnight at 37°C. The product was gel purified. To phosphorylate the ends, I used 15 µL (600 ng) of the product and added 1 µL T4 PNK, 5 µL 10x T4 Ligase Buffer and 29 µL of H₂O and incubated it for 30 minutes at 37°C. A 20 µL blunt-end ligation was performed for 2 hours at 16°C with 8 µL (100 ng) of phosphorylated vector. 5 µL (25 ng) of the ligation product was transformed. Initially, the alloPT construct had a 1 bp deletion at the ligation junction. Designing new primers (CB975, CB976) and repeating the site directed mutagenesis solved this issue. After obtaining the 3OPT and alloPT version for the 1.5 kb targeting vector, I simply transferred the mutated enhancer to the 25 kb and 75 kb targeting vector with Gibson assembly, leading to 6 targeting vectors that were transformed (CB306, CB307, CB308, CB309, CB310, CB311). For enhancer integrations we used the #1 (PCP+, no Flpo) cell line, also used in Chapter I for weak enhancer integrations.

Cloning of Flpo-IRFP.

Our present Flpo-containing plasmid (CB233) did not contain a functional fluorophore. I used CB022 as a backbone to replace EGFP by Flpo and puromycin by IRFP. EGFP was removed by cutting with AgeI and NotI. Flpo was amplified from CB233 and inserted instead of EGFP into CB022. CB022-Flpo was cut with BamHI and ApaI to replace the puromycin ORF. A vector containing a IRFP670-T2A-Neo cassette was a kind gift of Michelle Huth and was inserted instead of puromycin into CB022-Flpo (CB022-Flpo-IRFP). To test IRFP expression 1.5 µg was transfected into mESCs. This resulted in around 20% IRFP+ cells. The mean intensity (~470) was clearly different from background intensity (~90).

Generation of Doxycycline-Inducible Bxb1 Cell Line (iBxb1).

I utilized a pre-existing plasmid (CB020) that contained a doxycycline inducible *Fgf5* gene, a rtTA3-IRES-puromycin cassette and terminal repeats (TR) for PiggyBac-mediated insertion. I removed the puromycin cassette since it is used for 75 kb integrations in the synthetic locus, and I wanted to use FACS-based enrichment to select for the integrands later. The cassette was removed using the restriction enzymes EcoRI and XhoI and adding back two annealed oligos that re-establish EcoRI and XhoI via ligation. In this way any ORF could be still integrated after the IRES, if needed (CB020Δpuro).

A plasmid carrying Bxb1-P2A-IRFP670 was a gift from Moritz Gaidt. I removed *Fgf5* with flanking restriction sites from CB020Δpuro (EcoRI and XhoI). Bxb1-P2A-IRFP670 was PCR-amplified with >20 bp overhangs adding homology arms for Gibson assembly and a Kozak sequence for transcription initiation. The cassette was integrated via Gibson assembly (CB020-Bxb1-Δpuro).

CB020 still contained loxP sites flanking the rtTA3-IRES cassette. I found another vector (CB237) that originated from CB020 but had the loxP sites already removed. I amplified the UbqCprom-rtTA3 sequence from CB237 including homology arms and cut CB020

with Sall and NotI to remove the loxP flanked UbqCprom-rtTA3 sequence. The new UbqCprom-rtTA3 sequence not flanked by loxP was integrated via Gibson assembly, establishing the final plasmid CB020-Bxb1-ΔpuroΔloxP.

For transfection the *3x loxP no PCP (no Flpo)* cell line was used, herein referred to as #1. The #1 in Chapter III corresponds to this cell line, while the #1 that is the basis for enhancer integrations in Chapter I and II contains an additional PCP-GFP-Neo integration. 200,000 cells were transfected with 600 ng CB020-Bxb1-ΔpuroΔloxP and 1 μg of PiggyBac transposase-expressing plasmid (Day 1). On day 2 the medium was exchanged to 1 μg/mL of doxycycline to induce Bxb1-IRFP expression. On day 3 I expanded $\frac{3}{4}$ of the transfected cells in the 6-well plate to a 10 cm dish and on the next day (48 hours of doxycycline) the cells were 30-40% confluent. They were collected with 2 mL of trypsin for 8 minutes and 8 mL of 10% FBS. After centrifugation they were resuspended in 1 mL of 2i/LIF and were ready for sorting with Cytex Aurora CS™ Cell Sorter (Cytex® Biosciences). #1 served as a negative control. 33,000 cells were clearly IRFP positive (Fig. 12B) and sorted into a falcon tube and seeded in a 10 cm dish. On day 8 the cells grew dense enough, so I induced Bxb1-IRFP expression again with doxycycline for another round of enrichment. 48 hours later (day 10) cells were again sorted via FACS and around 50% were clearly IRFP positive (Fig. 12B). The highest expressing 18% were now directly sorted into two fibronectin-coated 96-well plates, to receive 192 clones. A week later (day 17) around 50% of the wells had a visible clone, meaning that I had around 96 clones left. Cells were lysed and genotyped with CB946 and CB002 (Fig. S9). Close to 100% of the cells had a band for the CB020-Bxb1-ΔpuroΔloxP construct. 10 of these iBxb1 clones were expanded into 12-well plates and subsequently seeded in doxycycline and no doxycycline conditions. Together with #1, they were tested for basal IRFP expression and induction potential with BD LSRFortessa™ Cell Analyzer (BD Life Sciences – Biosciences, Fig. 12C). Mean intensity levels differed over one order of magnitude between clones. The lowest with a mean of 347, to the highest with a mean of 7748, which corresponds to 16% and 99% of IRFP-positive cells, compared to #1, respectively. The no doxycycline fraction was indistinguishable from #1, hence there was no leakiness. I kept four clones. Of these four clones I took the doxycycline-induced fraction and put it into 2i/LIF for two days, then again induced it for two days with doxycycline and measured the induction potential with BD LSRFortessa™ Cell Analyzer (BD Life Sciences - Biosciences). Indeed, IRFP expression could be repeatedly induced without silencing effects. I selected clone 4 due to its high Bxb1-IRFP expression levels, which was subsequently called iBxb1-C4.

Cloning of GFP Cassettes.

The following expression cassettes had to be cloned: GFPDD, GFPDD-T2A-TK, TK-T2A-GFPDD, EGFP and EGFP-T2A-TK. EGFP was already existing in a fully functional expression cassette in CB022 (EF-1α promoter, Kozak-sequence, stop-codon, WPRE). EGFP was then exchanged by other ORFs. EGFP was excised with AgeI and NotI. This removed the Kozak-sequence and EGFP, including the stop-codon. I started by inserting GFPDD* and T2A-GFPDD* into CB022 (asterisk indicates stop-codon). The GFPDD* and T2A-GFPDD* ORFs were amplified from a pre-existing KLF5-T2A-GFPDD vector, kindly gifted by Michael Oberhuemer. The primers contained homology arms for Gibson

assembly. Additionally, the forward primers re-established the *AgeI* site, and the reverse primer re-established the *NotI* site. The forward primer for GFPDD* also added a Kozak-sequence. CB022-GFPDD was successfully cloned.

CB022-T2A-GFPDD* was then cut with *AgeI* to insert TK before T2A. TK was amplified from a pre-existing EGFP-T2A-TK vector (but not in CB022 backbone). Through the Gibson deletion method, the stop codon of TK was removed. A Kozak-sequence was added at the beginning. This PCR product was inserted to establish the CB022-TK-T2A-GFPDD vector. To clone the GFPDD-T2A-TK* cassette, T2A-TK* was amplified from the EGFP-T2A-TK vector. The GFPDD vector was cut with *NotI*, which ends ~12 bp after the stop-codon. The homology arm on the forward primer that amplified T2A-TK* was designed in a way to skip the 12 bp and the 3 bp stop-codon at the end of the GFPDD ORF. In this way, with the Gibson deletion method, T2A-TK* was added perfectly to GFPDD and CB022-GFPDD-T2A-TK was cloned.

All expression cassettes, with the exception of EGFP-T2A-TK, were now in a CB022-backbone. The GFP constructs were tested for proper function via transient transfection into mESCs (Fig. 13B).

Cloning of Bxb1 Landing Pads.

I utilized the pre-existing integration-excision system used in Chapter I and II to construct a landing pad (LP) that I could integrate at the 1.5 kb position. CB231 is the targeting vector for the 1.5 kb position. I constructed a sequence consisting of the two 52-bp attP sites (GT and GA spacer), flanked by restriction sites with different overhangs (*NotI*, *BamHI*). Between the two attP sites two random 20-bp sequences flanking a *EcoRI* site in the middle were inserted. This large sequence was split into two parts (four oligos) that were 78-79 bp long. *EcoRI* and the two random sequences in the middle allowed integration of a LP, since the two attP sites are so similar that they cannot be used as homology arms for directional integration via Gibson assembly. *NotI* and *BamHI* on the ends allowed a directional integration. The decision to use *NotI* and *BamHI* overhangs for inserting the entire construct into CB231 was made because the size of the oligos was close to a size limit (80 bp), therefore another pair of 20-bp homology arms at the end was not feasible. CB231 only had a *NotI* restriction site. To add the *BamHI* restriction site, two short oligos were constructed and annealed. This “restriction platform” was essentially inserted just like an enhancer via Gibson assembly (CB231 + restriction platform). This vector was then cut with *NotI* and *BamHI* and gel-purified to prevent presence of the parental vector. The two pairs of oligos for the attP sites were annealed and ligated in a single step for 2 hours at 16°C. I received around 100 clones of which 24 were analyzed via colony PCR (CB005 + CB374) and 13 turned out to contain the inserts. This construct was now called CB231 + 2x2 attP.

The GFP-containing cassettes that were constructed in CB022 were now amplified with overhangs corresponding to the random sequences flanking *EcoRI* of CB231 + 2x2 attP. CB231 + 2x2x attP was cut with *EcoRI* and Gibson assembly was performed to construct the final plasmids CB231-EGFP-attP-Bxb1, CB231-GFPDD-attP-Bxb1, CB231-TK-T2A-GFPDD-attP-Bxb1, CB231-GFPDD-T2A-TK-attP-Bxb1 and CB231-EGFP-T2A-TK-attP-Bxb1.

Integration of an EGFP-T2A-TK LP into iBxb1-C4.

The CB231-EGFP-T2A-TK-attP-Bxb1 construct allows Cre/Flpo-mediated integration at 1.5 kb. 200,000 cells were transfected with 800 ng of Cre and 200 ng of CB231-EGFP-T2A-TK-attP-Bxb1 and selected with blasticidin. According to genotyping, 42% contained the integration (Fig. S10). 9 clones were analyzed twice for GFP expression, once in the +dox and -dox condition, respectively (6 clones shown in Fig. 14A). I selected the highest expressing GFP cell line, which was H2 with 60% GFP+ (iBxb1-C4H2, Fig. 14A). iBxb1-C4H2 was transfected with the newly generated Flpo-IRFP plasmid to remove the backbone, since I could not select for absence of the backbone (TK) with ganciclovir, due to the LP encoding TK as well. 200,000 cells were transfected with 1.5 µg of Flpo-IRFP and sorted for IRFP-positive cells with Cytex Aurora CS™ Cell Sorter (Cytex® Biosciences) after two days into two fibronectin-coated 96-well plates. Around 30% of the 192 IRFP+ clones survived in the 96-well plate. From the ~60 clones 16 had the backbone removed according to genotyping (Fig. S10). 12 clones were expanded and GFP expression was measured again to ensure that the presence of the backbone had not silenced the GFP-cassette, and the landing pad was fully functional (6 clones shown in Fig. 14B). GFP expression ranged from 47.5% to 85.6% GFP+. The highest expressing clone (D12) was selected for testing (iBxb1-C4H2D12, Fig. 14B). D1 and D11 were kept as a backup.

Cloning of Bxb1-attB Targeting Vector for RMCE.

The backbone of the targeting vector is based on the pGem®-T vector (Promega). I recently inserted an A-tailed PCR-product containing the 75 kb lox71(m2). I excised the insert with SphI and Sall. Similar to the CB231 + 2x2 attP construct, I designed a sequence that contained the 47-bp attB(GT) and attB(GA) flanking two 20 bp random sequences and 2 restriction sites (NotI and NcoI) at the center. In this way, these restriction sites with the random sequences can be used for Gibson assembly. The whole sequence was flanked by SphI and Sall overhangs for insertion into pGem-T. The sequence was divided into a left and right part (with an overlapping sequence at the NotI site), and each was ordered as an upper and lower oligo. The matching pairs of oligos were annealed to get dsDNA fragments. Then these two inserts were simultaneously ligated into pGEM-T (pGEM-T-attB). To measure the exchange not only by loss of GFP but also gain of mCherry, a strong enhancer needed to be integrated. The *Mus musculus castaneus* (CAST/EiJ)-variant of the proximal enhancer of *eNanog* was amplified from CB223 and inserted into the NotI-cut pGEM-T-attB via Gibson assembly (pGEM-T-attB-*eNanog*). The NotI restriction site was maintained through part of the overhangs.

Integrating eNanog via RMCE at 1.5 kb to Assess Exchange Efficiency.

iBxb1-C4H2D12 contains both the inducible Bxb1-IRFP as well as the LP expressing EGFP-T2A-TK. Bxb1 expression was induced 48h before transfection with doxycycline. I also transfected non-induced cells, to assess whether changes in intensity could be attributed to transfection alone. Additional controls included non-induced and induced cells but without transfection as well as iBxb1-C4, that contained Bxb1 but no LP. Non-induced transfected cells, induced non-transfected cells and induced transfected cells are displayed in Fig. 16A. 48 hours after induction 1.5 µg of pGEM-T-attB-eNanog was

transfected. Additionally, a PCR-product of attB-eNanog-attB was generated with CB606 and CB006, adding 84-bp and 122-bp overhangs to the ends of attB and having a total size of 1508 bp. The PCR-product was gel-purified and both the A260/A280 and A260/A230 indicated high purity. 1.5 µg and 535 ng of the PCR-product were transfected. The PCR-product is approximately ~2.8-fold smaller than the full targeting vector, which means that 535 ng is the same number of molecules as 1.5 µg of targeting vector. The cells were analyzed on day 2, 4 and 6 after transfection for mCherry, GFP and IRFP expression with BD LSRFortessa™ Cell Analyzer (BD Life Sciences – Biosciences, Fig. 16A). For each measurement 1/2 of the cells were taken for analysis and 1/10 was passaged. At day 2 the entire population slightly shifted to higher intensity levels of PE-Texas Red-A (mCherry) and APC-A (IRFP) in the transfected cell lines (Fig. S11B). A few mCherry+/GFP- cells were already visible. GFP expression was decreased in the entire population as long as doxycycline was supplied (Fig. S11B). By day 4 PE-Texas Red-A and APC-A had recovered and the highest percentage of mCherry+/GFP- was measured. At day 8 of the first RMCE test the population was sorted for mCherry+GFP- and mCherry-GFP- with BD FACSMelody™ Cell Sorter (BD Life Sciences - Biosciences). The mCherry+GFP- subpopulation was still around 1%. The fractions were passaged, and genomic DNA was extracted. CB706 and CB517 were used to amplify the right junction of the exchanged *eNanog* (Fig. 16B). CB013 and CB517 were used to amplify the right junction of the landing pad. The PCR products were gel-purified and Sanger-sequenced (Fig. 16C).

In the first round of RMCE testing, doxycycline was withdrawn at day 4. At day 6 a slight drop in the fraction of mCherry+/GFP- cells was visible. To reduce toxicity, during the second RMCE doxycycline was withdrawn at day 2. The mCherry+GFP- fraction was now maintained at least until day 6.

Ganciclovir Selection of Recombinants.

24 hours after transfection of the first RMCE experiment, 1/2 of the doxycycline-induced and pGEM-T-attB-eNanog transfected cells were expanded separately to a 10 cm dish and 500 ng/mL ganciclovir was added the day after. The cells were selected until day 6 post transfection and the selection medium was exchanged every other day. Most cells seemed to have died at day 6 and the few remaining colonies were in a bad condition (irregular shapes and many dead cells). To relieve selection pressure and recover the remaining colonies, I withdrew ganciclovir at day 6 and grew them for another four days. At day 10, 48 colonies remained, which were picked into a fibronectin-coated 96-well plate. The cells were genotyped with CB706 and CB517 as well as CB013 and CB517, as described in “Integrating *eNanog* via RMCE at 1.5 kb to assess exchange efficiency (Fig. 17B). 26 of the 48 clones had *eNanog* integrated (Fig. 7B). All 48 clones in the 96-well plates were also analysed for mCherry and GFP expression with the plate-reader of BD LSRFortessa™ Cell Analyzer (BD Life Sciences – Biosciences, Fig. 17C and Fig. 17D). Two correct clones (A1 and C3) and two “backbone clones” (A2 and A3) were expanded and gDNA was extracted and sequenced (Fig. 17B). Additionally, the clones A2 and A3 were cultured in doxycycline for four days and analysed with BD LSRFortessa™ Cell Analyzer (BD Life Sciences - Biosciences), to check whether Bxb1 expression could remove the backbone of the integrated targeting vector (Fig. 17E).

Motif Analysis via JASPAR.

For the motif analysis in Fig. 5 motifs were fetched from the 2024 version of the JASPAR database via the python extension (Rauluseviciute et al., 2024). I used the core collection for mouse (accession number = 10090, all_versions = true). A total of 384 motifs were extracted, corresponding to 223 unique TFs. The count matrix for each motif was converted into a position frequency matrix (divide each nucleotide count by the total number of counts at that position) and a pseudocount was added. The matrix was converted into a position weight matrix (position-specific scoring matrix, pssm) by assuming a background probability of 0.25 for each nucleotide and computing the log2-fold change (observed vs. expected) of each nucleotide at each position. Then the minimal score and the maximal score of the motif was calculated and a threshold was set at 90% between the minimum and the maximum. E.g., if the minimum score was -40 and the maximum was 20, the threshold would be a score ≥ 14 . Additionally, I used the RNA-Seq dataset from Buecker et al. to determine which TFs are likely to be expressed in mESCs (Buecker et al., 2014). I applied an FKPM threshold of >1 . That meant that 11786 genes (51%) were “expressed”.

Table 3: Motif analysis via JASPAR.

Enhancer	Motifs fetched	Distinct TFs	# of motif instances	TFs expressed	# of motif instances filtered for expressed TFs
Map4k3 E1	143	96	254	36	83
Map4k3 E2	68	44	94	17	27
eNanog	113	81	170	26	50
Rybp E1	99	74	178	31	56
Rybp E2	132	93	241	35	84
Sox2 E15	71	56	111	28	46
Sox2 E19	82	62	122	23	44
Sox2 E20	95	74	167	29	66
SCR	251	158	1161	58	353

Table 4: Primers and Oligos.

Name	Database Number	Sequence (5'-3')
CSII EF seq reverse (Genotyping)	CB022	CCAGTCAATCTTTCACAAAT
M13 forward	CB005	GTTTTCCAGTCACGAC
WPRE 3' out	CB013	CTGTGTTGCCACCTGGATTCTG
MS2 fgf5 sequencing primer R	CB374	CACCTACTCAGACAATGCGATGC
STAgR colony reverse	CB517	TTACGGTTCCTGGCCTTTTG
Nanog_prox add seq	CB706	AAATTACGTCGCCCTTGGGA
Map4k3-E1-3OPT-fwd	CB936	TAACAATACTGAGTCAGAGAGAAAC
Map4k3-E1-3OPT-rev	CB937	GTTGCATAACAAGTATAATAACAA TAG

Map4k3-E1-allOPT-fwd	CB938	TAAGACAATTCTGAGTCAGAGAG AAAC
Map4k3-E1-allOPT-rev	CB939	TGCATAACAATGATAATAACAATA GCTGC
Seq-Bxb1-fwd-1	CB944	CGACCTGAATAGACGCGGCG
Seq-Bxb1-fwd-2	CB945	CGCAGCCAAGAACACATGGC
Seq-Bxb1-fwd-3	CB946	GGACGACGGCTCCGGGATGG
Seq-Pa01-fwd-1	CB947	GGAATGATATACCTGACGTC
Seq-Pa01-fwd-2	CB948	ACAATGGTCTCTTCCGATTC
Seq-Pa01-fwd-3	CB949	GCGAGGCCAAGCGCAGCGAC
GA_Bxb1/Pa01-P2A-IRFP670- WPRE_fwd	CB950	GCACCGGTGGTACCGACTCGGT GACGTCTAGAGCCACC
GA_Bxb1/Pa01-P2A-IRFP670- WPRE_rev	CB951	GCGGAGGCCACGCGTAGATCAG GTCGACGGTATCGATG
NcoI_HpaI (puro removal CB020)_ fwd	CB952	CATGGTTTGGTATCTGCGCTCGT T
NcoI_HpaI (puro removal CB020)_ rev-NEW	CB953	AACGAGCGCAGATACCAAAC
GA_Flpo_to_CB022_fwd	CB956	CTTAAAAGATCCAGCCACCAATG AGCCAGTTCGACATCCT
GA_Flpo_to_CB022_rev	CB957	CAGAGGTTGATTGTGACGCTCA GATCCGCCTGTTGATGT
Restriction_platform_CB231_librar y_oligo 1	CB958	TACATTATACGAACGGTAGCGCG GCCGCTCTATGGATCCGCGAAG TTCCTATTCTCTAG
Restriction_platform_CB231_librar y_oligo 2	CB959	ATGTAATATGCTTGCCATCGCGC CGGCGAGATACCTAGGCGCTTC AAGGATAAGAGATC
NotI-attP(GT)-EcoRI_oligo_1	CB960	GGCCGCGTGGTTTGTCTGGTCA ACCACCGCGGTCTCAGTGGTGT ACGGTACAAACCCATAGTCAGCA ACCAGGTGTGGG
NotI-attP(GT)-EcoRI_oligo_2-NEW	CB961	AATTCCCACACCTGGTTGCTGAC TATGGGTTTGTACCGTACACCAC TGAGACCGCGGTGGTTGACCAG ACAAACCACGC
EcoRI-attP(GA)-BamHI_oligo_1	CB962	AATTCCCAGTCACGTAGCGATAG CGGTGGTTTGTCTGGTCAACCAC CGCGGACTCAGTGGTGTACGGT ACAAACCCAG
EcoRI-attP(GA)-BamHI_oligo_2- NEW	CB963	GATCCTGGGTTTGTACCGTACAC CACTGAGTCCGCGGTGGTTGAC CAGACAAACCACCGCTATCGCTA CGTGACTIONGG
HA_AgeI_Kozak_GFPDD/BFPDD_ fwd	CB964	CTTAAAAGATCCAGCCACCACCG GTGCCACCATGGTGAGCAAGGG CGAG
HA_GFPDD/BFPDD_rev	CB965	CAGAGGTTGATTGTGACGCGG CCGCCGCGCAGTGCCTACACAT TG
HA_AgeI-T2A-GFPDD_fwd	CB966	CTTAAAAGATCCAGCCACCACCG GTGAGGGCAGAGGAAGTCTCC

T2A-TK_to_GFPDD-fwd	CB967	CTTCTGCTAGGATCAATGTGGAG GGCAGAGGCAGTCTGCTG
T2A-TK_to_GFPDD-rev	CB968	CAGAGGTTGATTGTGACGCTCA GTTAGCCTCCCCCATCT
TK_into_T2AGFPDD_fwd	CB969	GCTTAAAAGATCCAGCCACCATG TCCATGCCCACGCTACTGCGG
TK_into_T2AGFPDD(nostop)_rev	CB970	AGGAGACTTCCTCTGCCCTCGTT AGCCTCCCCCATCTCCCG
Nanog-proximal-to-attB_fwd		CGACTGCTGTATTTTCATCGCGGC CGCGTAAAGCAGCACAAAGCCTT
Nanog-proximal-to-attB_rev		GGCATCGTTGCCATGGTCGCAT GCTTACTAATCCCAGTTA
CB022-cassette_to_CB231_fwd		AGTCAGCAACCAGGTGTGGGAT CCGGTGCCCGTCAGTGGG
CB022-cassette_to_CB231_rev		GCTATCGCTACGTGACTGGGGC GGGGAGGCGGCCCAAAGG
IRFP-Neo-casette_GA-fwd		CTTTGGGCGCCTCCCCGCGGG GTAGGGGAGGCGCTTTTC
IRFP-Neo-casette_GA-rev		GGGAAAAAGCCATACCAATGCCA TAGAGCCCACCGCATCC
SphI-attB(GT)-NotI-up	CB971	CCGGCCGGCTTGTCGACGACGG CGGTCTCCGTCGTCAGGATCATC CGGTCGACTGCTGTATTTTCATCG C
SphI-attB(GT)-NotI-down	CB972	GGCCGCGATGAAATACAGCAGT CGACCGGATGATCCTGACGACG GAGACCGCCGTCGTCGACAAGC CGCCCGGCATG
NcoI-attB(GA)-EcoRI-up	CB973	GGCCGCGACCATGGCAACGATG CCTACTCTCCCGGCCGGCTTGTC GACGACGGCGGACTCCGTCGTC AGGATCATCCGGG
NcoI-attB(GA)-EcoRI-down	CB974	TCGACCCGGATGATCCTGACGA CGGAGTCCGCCGTCGTCGACAA GCCGGCCGGGAGAGTAGGCATC GTTGCCATGGTCGC
Map4k3-E1-allOPT-fwd_NEW	CB975	CATAAGACAATTCTGAGTCAGAG AG
Map4k3-E1-allOPT-rev_NEW	CB976	CATAACAATGATAATAACAATAGC TGC
GA_rtTA_fwd	CB1023	GCCTCCCCGCATCGATACCGGA TCTACGCGTGGCCTCCGC
GA_rtTA_rev	CB1024	AGGGAGAGGGGCGGAATTTGC

Table 5: Plasmids.

Name	Database Number
pB tetO Fgf5 Puro	CB020
pB Ef1 GFP Ubi Puro	CB022
lox66(m2) FRT(F3) PuroTK	CB200
lox66(2272) FRT(F5) HygroTK	CB201
lox66(m2) FRT(F3) PuroTK + eNanog proximal	CB223

lox66(WT) FRT(WT) BlastiT _K	CB231
Flpo + non-functional dsRed	CB233
pB tetO Podxl ubiquitin Puro w/o loxP	CB237
lox66(WT) FRT(WT) BlastiT _K + Map4k3 E1	CB254
lox66(m2) FRT(F3) PuroTK + Map4k3 E1 3OPT	CB306
lox66(2272) FRT(F5) HygroTK + Map4k3 E1 3OPT	CB307
lox66(WT) FRT(WT) BlastiT _K + Map4k3 E1 3OPT	CB308
lox66(m2) FRT(F3) PuroTK + Map4k3 E1 allOPT	CB309
lox66(2272) FRT(F5) HygroTK + Map4k3 E1 allOPT	CB310
lox66(2272) FRT(F5) HygroTK + Map4k3 E1 allOPT	CB311
Cre-tdTomato	/
IRFP670-T2A-Neo (<i>Michelle Huth</i>)	/
CB022-Flpo (Puro)	/
CB022-Flpo-IRFP	/
p146 pLenti Bxb1 IRFP670 (<i>Moritz Gaidt</i>)	/
CB020 Δ puro	/
CB020-Bxb1- Δ puro	/
CB020-Bxb1- Δ puro Δ loxP	/
PiggyBac transposase	/
KLF5-T2A-GFPDD (<i>Michael Oberhuemer</i>)	/
pGEM-T-GFP-cassette (step4+loxP66) EGFP-T2A-TK (<i>Songjie Feng</i>)	/
CB022-GFPDD	/
CB022-T2A-GFPDD	/
CB022-TK-T2A-GFPDD	/
CB022-GFPDD-T2A-TK	/
CB231 + restriction platform	/
CB231 + 2x2 attP	/
CB231-EGFP-attP-Bxb1	/
CB231-GFPDD-attP-Bxb1	/
CB231-TK-T2A-GFPDD-attP-Bxb1	/
CB231-GFPDD-T2A-TK-attP-Bxb1	/
CB231-EGFP-T2A-TK-attP-Bxb1	/
pGEM-T-CB815-CB635 75 kb lox site sequencing	/
pGEM-T-attB-Bxb1	/
pGEM-T-attB-Bxb1 + eNanog	/

References

- Albert, H., Dale, E.C., Lee, E., Ow, D.W., 1995. Site-specific integration of DNA into wild-type and mutant *lox* sites placed in the plant genome. *The Plant Journal* 7, 649–659. <https://doi.org/10.1046/j.1365-313X.1995.7040649.x>
- Aljahani, A., Hua, P., Karpinska, M.A., Quililan, K., Davies, J.O.J., Oudelaar, A.M., 2022. Analysis of sub-kilobase chromatin topology reveals nano-scale regulatory interactions with variable dependence on cohesin and CTCF. *Nat Commun* 13, 2139. <https://doi.org/10.1038/s41467-022-29696-5>
- Ambrosetti, D.C., Basilico, C., Dailey, L., 1997. Synergistic activation of the fibroblast growth factor 4 enhancer by Sox2 and Oct-3 depends on protein-protein interactions facilitated by a specific spatial arrangement of factor binding sites. *Mol Cell Biol* 17, 6321–6329.
- Araki, K., Araki, M., Yamamura, K., 2006. Negative Selection with the Diphtheria toxin A fragment Gene Improves Frequency of Cre-Mediated Cassette Exchange in ES Cells. *The Journal of Biochemistry* 140, 793–798. <https://doi.org/10.1093/jb/mvj208>
- Araki, K., Araki, M., Yamamura, K. -i., 1997. Targeted integration of DNA using mutant *lox* sites in embryonic stem cells. *Nucleic Acids Research* 25, 868–872. <https://doi.org/10.1093/nar/25.4.868>
- Arnold, C.D., Gerlach, D., Stelzer, C., Boryń, Ł.M., Rath, M., Stark, A., 2013. Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science* 339, 1074–1077. <https://doi.org/10.1126/science.1232542>
- Austin, S., Ziese, M., Sternberg, N., 1981. A novel role for site-specific recombination in maintenance of bacterial replicons. *Cell* 25, 729–736. [https://doi.org/10.1016/0092-8674\(81\)90180-X](https://doi.org/10.1016/0092-8674(81)90180-X)
- Baer, A., Bode, J., 2001. Coping with kinetic and thermodynamic barriers: RMCE, an efficient strategy for the targeted integration of transgenes. *Current Opinion in Biotechnology* 12, 473–480. [https://doi.org/10.1016/S0958-1669\(00\)00248-2](https://doi.org/10.1016/S0958-1669(00)00248-2)
- Balsalobre, A., Drouin, J., 2022. Pioneer factors as master regulators of the epigenome and cell fate. *Nat Rev Mol Cell Biol* 23, 449–464. <https://doi.org/10.1038/s41580-022-00464-z>
- Banerji, J., Rusconi, S., Schaffner, W., 1981. Expression of a β -globin gene is enhanced by remote SV40 DNA sequences. *Cell* 27, 299–308. [https://doi.org/10.1016/0092-8674\(81\)90413-X](https://doi.org/10.1016/0092-8674(81)90413-X)
- Bell, C.C., Balic, J.J., Talarmain, L., Gillespie, A., Scolamiero, L., Lam, E.Y.N., Ang, C.-S., Faulkner, G.J., Gilan, O., Dawson, M.A., 2024. Comparative cofactor screens show the influence of transactivation domains and core promoters on the mechanisms of transcription. *Nat Genet*. <https://doi.org/10.1038/s41588-024-01749-z>
- Bergman, D.T., Jones, T.R., Liu, V., Ray, J., Jagoda, E., Siraj, L., Kang, H.Y., Nasser, J., Kane, M., Rios, A., Nguyen, T.H., Grossman, S.R., Fulco, C.P., Lander, E.S., Engreitz, J.M., 2022. Compatibility rules of human enhancer and promoter sequences. *Nature* 607, 176–184. <https://doi.org/10.1038/s41586-022-04877-w>
- Blayney, J.W., Francis, H., Rampasekova, A., Camellato, B., Mitchell, L., Stolper, R., Cornell, L., Babbs, C., Boeke, J.D., Higgs, D.R., Kassouf, M., 2023. Super-enhancers include classical enhancers and facilitators to fully activate gene expression. *Cell* 186, 5826–5839.e18. <https://doi.org/10.1016/j.cell.2023.11.030>
- Bouhassira, E.E., Westerman, K., Leboulch, P., 1997. Transcriptional Behavior of LCR Enhancer Elements Integrated at the Same Chromosomal Locus by Recombinase-Mediated Cassette Exchange. *Blood* 90, 3332–3344. <https://doi.org/10.1182/blood.V90.9.3332>
- Bower, G., Hollingsworth, E.W., Jacinto, S., Clock, B., Cao, K., Liu, M., Dziulko, A., Alcaina-Caro, A., Xu, Q., Skowronska-Krawczyk, D., Lopez-Rios, J., Dickel, D.E., Bardet, A.F., Pennacchio, L.A., Visel, A., Kvon, E.Z., 2024. Conserved *Cis* -Acting Range Extender Element Mediates Extreme Long-Range Enhancer Activity in Mammals. <https://doi.org/10.1101/2024.05.26.595809>

- Brosh, R., Coelho, C., Ribeiro-dos-Santos, A.M., Ellis, G., Hogan, M.S., Ashe, H.J., Somogyi, N., Ordoñez, R., Luther, R.D., Huang, E., Boeke, J.D., Maurano, M.T., 2023. Synthetic regulatory genomics uncovers enhancer context dependence at the Sox2 locus. *Molecular Cell* 83, 1140–1152.e7. <https://doi.org/10.1016/j.molcel.2023.02.027>
- Buchholz, F., 1996. Different thermostabilities of FLP and Cre recombinases: implications for applied site-specific recombination. *Nucleic Acids Research* 24, 4256–4262. <https://doi.org/10.1093/nar/24.21.4256>
- Buchholz, F., Angrand, P.-O., Stewart, A.F., 1998. Improved properties of FLP recombinase evolved by cycling mutagenesis. *Nat Biotechnol* 16, 657–662. <https://doi.org/10.1038/nbt0798-657>
- Buecker, C., Srinivasan, R., Wu, Z., Calo, E., Acampora, D., Faial, T., Simeone, A., Tan, M., Swigut, T., Wysocka, J., 2014. Reorganization of Enhancer Patterns in Transition from Naive to Primed Pluripotency. *Cell Stem Cell* 14, 838–853. <https://doi.org/10.1016/j.stem.2014.04.003>
- Chakraborty, S., Kopitchinski, N., Zuo, Z., Eraso, A., Awasthi, P., Chari, R., Mitra, A., Tobias, I.C., Moorthy, S.D., Dale, R.K., Mitchell, J.A., Petros, T.J., Rocha, P.P., 2023. Enhancer–promoter interactions can bypass CTCF-mediated boundaries and contribute to phenotypic robustness. *Nat Genet* 55, 280–290. <https://doi.org/10.1038/s41588-022-01295-6>
- Chalberg, T.W., Portlock, J.L., Olivares, E.C., Thyagarajan, B., Kirby, P.J., Hillman, R.T., Hoelters, J., Calos, M.P., 2006. Integration Specificity of Phage ϕ C31 Integrase in the Human Genome. *Journal of Molecular Biology* 357, 28–48. <https://doi.org/10.1016/j.jmb.2005.11.098>
- Chan, K.-M., Liu, Y.-T., Ma, C.-H., Jayaram, M., Sau, S., 2013. The 2 micron plasmid of *Saccharomyces cerevisiae*: A miniaturized selfish genome with optimized functional competence. *Plasmid* 70, 2–17. <https://doi.org/10.1016/j.plasmid.2013.03.001>
- Chen, X., Vega, V.B., Ng, H.-H., 2008. Transcriptional Regulatory Networks in Embryonic Stem Cells. *Cold Spring Harb Symp Quant Biol* 73, 203–209. <https://doi.org/10.1101/sqb.2008.73.026>
- Chen, Y., 2000. Crystal Structure of a Flp Recombinase–Holliday Junction Complex Assembly of an Active Oligomer by Helix Swapping. *Molecular Cell* 6, 885–897. [https://doi.org/10.1016/S1097-2765\(00\)00086-1](https://doi.org/10.1016/S1097-2765(00)00086-1)
- Chen, Y., Rice, P.A., 2003. New Insight into Site-Specific Recombination from Flp Recombinase-DNA Structures. *Annu. Rev. Biophys. Biomol. Struct.* 32, 135–159. <https://doi.org/10.1146/annurev.biophys.32.110601.141732>
- Chen, Z.Y., He, C.Y., Meuse, L., Kay, M.A., 2004. Silencing of episomal transgene expression by plasmid bacterial DNA elements in vivo. *Gene Ther* 11, 856–864. <https://doi.org/10.1038/sj.gt.3302231>
- Crocker, J., Abe, N., Rinaldi, L., McGregor, A.P., Frankel, N., Wang, S., Alsawadi, A., Valenti, P., Plaza, S., Payre, F., Mann, R.S., Stern, D.L., 2015. Low Affinity Binding Site Clusters Confer Hox Specificity and Regulatory Robustness. *Cell* 160, 191–203. <https://doi.org/10.1016/j.cell.2014.11.041>
- Crocker, J., Preger-Ben Noon, E., Stern, D.L., 2016. Chapter Twenty-Seven - The Soft Touch: Low-Affinity Transcription Factor Binding Sites in Development and Evolution, in: Wassarman, P.M. (Ed.), *Current Topics in Developmental Biology, Essays on Developmental Biology, Part B*. Academic Press, pp. 455–469. <https://doi.org/10.1016/bs.ctdb.2015.11.018>
- Davis, J.E., Insigne, K.D., Jones, E.M., Hastings, Q.A., Boldridge, W.C., Kosuri, S., 2020. Dissection of c-AMP Response Element Architecture by Using Genomic and Episomal Massively Parallel Reporter Assays. *Cell Systems* 11, 75–85.e7. <https://doi.org/10.1016/j.cels.2020.05.011>
- De Almeida, B.P., Reiter, F., Pagani, M., Stark, A., 2022. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nat Genet* 54, 613–624. <https://doi.org/10.1038/s41588-022-01048-5>
- de Almeida, B.P., Schaub, C., Pagani, M., Secchia, S., Furlong, E.E.M., Stark, A., 2024. Targeted design of synthetic enhancers for selected tissues in the *Drosophila* embryo. *Nature* 626, 207–211. <https://doi.org/10.1038/s41586-023-06905-9>

Deng, W., Rupon, J.W., Krivega, I., Breda, L., Motta, I., Jahn, K.S., Reik, A., Gregory, P.D., Rivella, S., Dean, A., Blobel, G.A., 2014. Reactivation of Developmentally Silenced Globin Genes by Forced Chromatin Looping. *Cell* 158, 849–860. <https://doi.org/10.1016/j.cell.2014.05.050>

Donovan, B.T., Huynh, A., Ball, D.A., Patel, H.P., Poirier, M.G., Larson, D.R., Ferguson, M.L., Lenstra, T.L., 2019. Live-cell imaging reveals the interplay between transcription factors, nucleosomes, and bursting. *EMBO J* 38, e100809. <https://doi.org/10.15252/embj.2018100809>

Dou, Y., Lin, Y., Wang, T., Wang, X.-Y., Jia, Y., Zhao, C., 2021. The CAG promoter maintains high-level transgene expression in HEK293 cells. *FEBS Open Bio* 11, 95–104. <https://doi.org/10.1002/2211-5463.13029>

Doughty, B.R., Hinks, M.M., Schaepe, J.M., Marinov, G.K., Thurm, A.R., Rios-Martinez, C., Parks, B.E., Tan, Y., Marklund, E., Dubocanin, D., Bintu, L., Greenleaf, W.J., 2024. Single-molecule chromatin configurations link transcription factor binding to expression in human cells. <https://doi.org/10.1101/2024.02.02.578660>

Du, M., Stitzinger, S.H., Spille, J.-H., Cho, W.-K., Lee, C., Hijaz, M., Quintana, A., Cissé, I.I., 2024. Direct observation of a condensate effect on super-enhancer controlled gene bursting. *Cell* S0092867423013375. <https://doi.org/10.1016/j.cell.2023.12.005>

Duportet, X., Wroblewska, L., Guye, P., Li, Y., Eyquem, J., Rieders, J., Rimchala, T., Batt, G., Weiss, R., 2014. A platform for rapid prototyping of synthetic gene networks in mammalian cells. *Nucleic Acids Research* 42, 13440–13451. <https://doi.org/10.1093/nar/gku1082>

Durrant, M.G., Fanton, A., Tycko, J., Hinks, M., Chandrasekaran, S.S., Perry, N.T., Schaepe, J., Du, P.P., Lotfy, P., Bassik, M.C., Bintu, L., Bhatt, A.S., Hsu, P.D., 2023. Systematic discovery of recombinases for efficient integration of large DNA sequences into the human genome. *Nat Biotechnol* 41, 488–499. <https://doi.org/10.1038/s41587-022-01494-w>

Ennifar, E., 2003. Crystal structure of a wild-type Cre recombinase-loxP synapse reveals a novel spacer conformation suggesting an alternative mechanism for DNA cleavage activation. *Nucleic Acids Research* 31, 5449–5460. <https://doi.org/10.1093/nar/gkg732>

Farley, E.K., Olson, K.M., Zhang, W., Brandt, A.J., Rokhsar, D.S., Levine, M.S., 2015. Suboptimization of developmental enhancers. *Science* 350, 325–328. <https://doi.org/10.1126/science.aac6948>

Feng, Y.-Q., Seibler, J., Alami, R., Eisen, A., Westerman, K.A., Leboulch, P., Fiering, S., Bouhassira, E.E., 1999. Site-specific chromosomal integration in mammalian cells: highly efficient CRE recombinase-mediated cassette exchange. *Journal of Molecular Biology* 292, 779–785. <https://doi.org/10.1006/jmbi.1999.3113>

Fogg, P.C.M., Colloms, S., Rosser, S., Stark, M., Smith, M.C.M., 2014. New Applications for Phage Integrases. *Journal of Molecular Biology* 426, 2703–2716. <https://doi.org/10.1016/j.jmb.2014.05.014>

Friedman, M.J., Wagner, T., Lee, H., Rosenfeld, M.G., Oh, S., 2024. Enhancer–promoter specificity in gene transcription: molecular mechanisms and disease associations. *Exp Mol Med* 56, 772–787. <https://doi.org/10.1038/s12276-024-01233-y>

Gao, Z., Wang, M., Smith, A., Boyes, J., 2023. YY1 Binding to Regulatory Elements That Lack Enhancer Activity Promotes Locus Folding and Gene Activation. *Journal of Molecular Biology* 435, 168315. <https://doi.org/10.1016/j.jmb.2023.168315>

Georgiades, E., Harrold, C.L., Roberts, N., Kassouf, M., Riva, S.G., Sanders, E., Francis, H.S., Blayney, J., Oudelaar, A.M., Milne, T.A., Higgs, D.R., Hughes, J., 2023. Active regulatory elements recruit cohesin to establish cell-specific chromatin domains (preprint). *Genomics*. <https://doi.org/10.1101/2023.10.13.562171>

Ghosh, P., Kim, A.I., Hatfull, G.F., 2003. The Orientation of Mycobacteriophage Bxb1 Integration Is Solely Dependent on the Central Dinucleotide of attP and attB. *Molecular Cell* 12, 1101–1111. [https://doi.org/10.1016/S1097-2765\(03\)00444-1](https://doi.org/10.1016/S1097-2765(03)00444-1)

- Ghosh, P., Wasil, L.R., Hatfull, G.F., 2006. Control of Phage Bxb1 Excision by a Novel Recombination Directionality Factor. *PLoS Biol* 4, e186. <https://doi.org/10.1371/journal.pbio.0040186>
- Giannoukos, G., Ciulla, D.M., Marco, E., Abdulkarim, H.S., Barrera, L.A., Bothmer, A., Dhanapal, V., Gloskowski, S.W., Jayaram, H., Maeder, M.L., Skor, M.N., Wang, T., Myer, V.E., Wilson, C.J., 2018. UDiTaS™, a genome editing detection method for indels and genome rearrangements. *BMC Genomics* 19, 212. <https://doi.org/10.1186/s12864-018-4561-9>
- Goutelle, S., Maurin, M., Rougier, F., Barbaut, X., Bourguignon, L., Ducher, M., Maire, P., 2008. The Hill equation: a review of its capabilities in pharmacological modelling. *Fundamental Clinical Pharmacology* 22, 633–648. <https://doi.org/10.1111/j.1472-8206.2008.00633.x>
- Grainge, I., Buck, D., Jayaram, M., 2000. Geometry of site alignment during Int family recombination: antiparallel synapsis by the Flp recombinase. *Journal of Molecular Biology* 298, 749–764. <https://doi.org/10.1006/jmbi.2000.3679>
- Gröschel, S., Sanders, M.A., Hoogenboezem, R., de Wit, E., Bouwman, B.A.M., Erpelinck, C., van der Velden, V.H.J., Havermans, M., Avellino, R., van Lom, K., Rombouts, E.J., van Duin, M., Döhner, K., Beverloo, H.B., Bradner, J.E., Döhner, H., Löwenberg, B., Valk, P.J.M., Bindels, E.M.J., de Laat, W., Delwel, R., 2014. A Single Oncogenic Enhancer Rearrangement Causes Concomitant EVI1 and GATA2 Dereglulation in Leukemia. *Cell* 157, 369–381. <https://doi.org/10.1016/j.cell.2014.02.019>
- Gschwind, A.R., Mualim, K.S., Karbalayghareh, A., Sheth, M.U., Dey, K.K., Jagoda, E., Nurtdinov, R.N., Xi, W., Tan, A.S., Jones, H., Ma, X.R., Yao, D., Nasser, J., Avsec, Ž., James, B.T., Shamim, M.S., Durand, N.C., Rao, S.S.P., Mahajan, R., Doughty, B.R., Andreeva, K., Ulirsch, J.C., Fan, K., Perez, E.M., Nguyen, T.C., Kelley, D.R., Finucane, H.K., Moore, J.E., Weng, Z., Kellis, M., Bassik, M.C., Price, A.L., Beer, M.A., Guigó, R., Stamatoyannopoulos, J.A., Aiden, E.L., Greenleaf, W.J., Leslie, C.S., Steinmetz, L.M., Kundaje, A., Engreitz, J.M., 2023. An encyclopedia of enhancer-gene regulatory interactions in the human genome (preprint). *Genomics*. <https://doi.org/10.1101/2023.11.09.563812>
- Hardison, R.C., 2012. Evolution of Hemoglobin and Its Genes. *Cold Spring Harbor Perspectives in Medicine* 2, a011627–a011627. <https://doi.org/10.1101/cshperspect.a011627>
- Hay, D., Hughes, J.R., Babbs, C., Davies, J.O.J., Graham, B.J., Hanssen, L.L.P., Kassouf, M.T., Oudelaar, A.M., Sharpe, J.A., Suciu, M.C., Telenius, J., Williams, R., Rode, C., Li, P.-S., Pennacchio, L.A., Sloane-Stanley, J.A., Ayyub, H., Butler, S., Sauka-Spengler, T., Gibbons, R.J., Smith, A.J.H., Wood, W.G., Higgs, D.R., 2016. Genetic dissection of the α -globin super-enhancer in vivo. *Nat Genet* 48, 895–903. <https://doi.org/10.1038/ng.3605>
- Hoess, R.H., Wierzbicki, A., Abremski, K., 1986. The role of the *loxP* spacer region in Φ 1 site-specific recombination. *Nucl Acids Res* 14, 2287–2300. <https://doi.org/10.1093/nar/14.5.2287>
- Hsieh, T.-H.S., Cattoglio, C., Slobodyanyuk, E., Hansen, A.S., Darzacq, X., Tjian, R., 2022. Enhancer–promoter interactions and transcription are largely maintained upon acute loss of CTCF, cohesin, WAPL or YY1. *Nat Genet* 54, 1919–1932. <https://doi.org/10.1038/s41588-022-01223-8>
- Hsieh, T.-H.S., Cattoglio, C., Slobodyanyuk, E., Hansen, A.S., Rando, O.J., Tjian, R., Darzacq, X., 2020. Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. *Molecular Cell* 78, 539–553.e8. <https://doi.org/10.1016/j.molcel.2020.03.002>
- Hume, M.A., Barrera, L.A., Gisselbrecht, S.S., Bulyk, M.L., 2015. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein–DNA interactions. *Nucleic Acids Research* 43, D117–D122. <https://doi.org/10.1093/nar/gku1045>
- Hung, T.-C., Kingsley, D.M., Boettiger, A.N., 2024. Boundary stacking interactions enable cross-TAD enhancer–promoter communication during limb development. *Nat Genet* 56, 306–314. <https://doi.org/10.1038/s41588-023-01641-2>
- Inniss, M.C., Bandara, K., Jusiak, B., Lu, T.K., Weiss, R., Wroblewska, L., Zhang, L., 2017. A novel Bxb1 integrase RMCE system for high fidelity site-specific integration of mAb expression cassette in CHO Cells. *Biotech & Bioengineering* 114, 1837–1846. <https://doi.org/10.1002/bit.26268>

- Jensen, C.L., Chen, L.-F., Swigut, T., Crocker, O.J., Yao, D., Bassik, M.C., Ferrell, J.E., Boettiger, A.N., Wysocka, J., 2024. Long range regulation of transcription scales with genomic distance in a gene specific manner. <https://doi.org/10.1101/2024.07.19.604327>
- Jindal, G.A., Farley, E.K., 2021. Enhancer grammar in development, evolution, and disease: dependencies and interplay. *Developmental Cell* 56, 575–587. <https://doi.org/10.1016/j.devcel.2021.02.016>
- Jusiak, B., Jagtap, K., Gaidukov, L., Duportet, X., Bandara, K., Chu, J., Zhang, L., Weiss, R., Lu, T.K., 2019. Comparison of Integrases Identifies Bxb1-GA Mutant as the Most Efficient Site-Specific Integrase System in Mammalian Cells. *ACS Synth. Biol.* 8, 16–24. <https://doi.org/10.1021/acssynbio.8b00089>
- Kameyama, Y., Kawabe, Y., Ito, A., Kamihira, M., 2010. An accumulative site-specific gene integration system using cre recombinase-mediated cassette exchange. *Biotech & Bioengineering* 105, 1106–1114. <https://doi.org/10.1002/bit.22619>
- Kane, L., Williamson, I., Flyamer, I.M., Kumar, Y., Hill, R.E., Lettice, L.A., Bickmore, W.A., 2022. Cohesin is required for long-range enhancer action at the Shh locus. *Nat Struct Mol Biol* 29, 891–897. <https://doi.org/10.1038/s41594-022-00821-8>
- Karr, J.P., Ferrie, J.J., Tjian, R., Darzacq, X., 2022. The transcription factor activity gradient (TAG) model: contemplating a contact-independent mechanism for enhancer–promoter communication. *Genes Dev.* 36, 7–16. <https://doi.org/10.1101/gad.349160.121>
- Karreman, S., Hauser, H., Karreman, C., 1996. On the Use of Double FLP Recognition Targets (FRTs) in the LTR of Retroviruses for the Construction of High Producer Cell Lines. *Nucleic Acids Research* 24, 1616–1624. <https://doi.org/10.1093/nar/24.9.1616>
- Kassouf, M., Ford, S., Blayney, J., Higgs, D., 2023. Understanding fundamental principles of enhancer biology at a model locus: Analysing the structure and function of an enhancer cluster at the α -globin locus. *BioEssays* 45, 2300047. <https://doi.org/10.1002/bies.202300047>
- Kassouf, M.T., Francis, H.S., Gosden, M., Suci, M.C., Downes, D.J., Harrold, C., Larke, M., Oudelaar, M., Cornell, L., Blayney, J., Telenius, J., Xella, B., Shen, Y., Sousos, N., Sharpe, J.A., Sloane-Stanley, J., Smith, A., Babbs, C., Hughes, J.R., Higgs, D.R., 2022. Multipartite super-enhancers function in an orientation-dependent manner (preprint). *Molecular Biology*. <https://doi.org/10.1101/2022.07.14.499999>
- Keravala, A., Portlock, J.L., Nash, J.A., Vitrant, D.G., Robbins, P.D., Calos, M.P., 2006. PhiC31 integrase mediates integration in cultured synovial cells and enhances gene expression in rabbit joints. *The Journal of Gene Medicine* 8, 1008–1017. <https://doi.org/10.1002/jgm.928>
- Kim, M., Wang, P., Clow, P., Chien, I., Wang, X., Peng, J., Chai, H., Liu, X., Lee, B., Ngan, C.Y., Yue, F., Milenkovic, O., Chuang, J.H., Wei, C.-L., Casellas, R., Cheng, A., Ruan, Y., 2024. Multifaceted roles of cohesin in regulating transcriptional loops. <https://doi.org/10.1101/2024.03.25.586715>
- Kim, M.S., Kim, W.H., Lee, G.M., 2007. A simple analysis system for the estimation of recombination efficiency using fluorescence-activated cell sorting. *Journal of Biotechnology* 127, 373–384. <https://doi.org/10.1016/j.jbiotec.2006.07.027>
- Kim, S., Morgunova, E., Naqvi, S., Goovaerts, S., Bader, M., Koska, M., Popov, A., Luong, C., Pogson, A., Swigut, T., Claes, P., Taipale, J., Wysocka, J., 2024. DNA-guided transcription factor cooperativity shapes face and limb mesenchyme. *Cell* 187, 692–711.e26. <https://doi.org/10.1016/j.cell.2023.12.032>
- Kim, S., Wysocka, J., 2023. Deciphering the multi-scale, quantitative cis-regulatory code. *Molecular Cell* 83, 373–392. <https://doi.org/10.1016/j.molcel.2022.12.032>
- Klein, J.C., Agarwal, V., Inoue, F., Keith, A., Martin, B., Kircher, M., Ahituv, N., Shendure, J., 2020. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat Methods* 17, 1083–1091. <https://doi.org/10.1038/s41592-020-0965-y>
- Kolb, A.F., 2001. Selection-Marker-Free Modification of the Murine β -Casein Gene Using a lox2722 Site. *Analytical Biochemistry* 290, 260–271. <https://doi.org/10.1006/abio.2000.4984>
- Langer, S.J., 2002. A genetic screen identifies novel non-compatible loxP sites. *Nucleic Acids Research* 30, 3067–3077. <https://doi.org/10.1093/nar/gkf421>

- Lauth, M., 2002. Stable and efficient cassette exchange under non-selectable conditions by combined use of two site-specific recombinases. *Nucleic Acids Research* 30, 115e–1115. <https://doi.org/10.1093/nar/gnf114>
- Lee, G., Saito, I., 1998. Role of nucleotide sequences of loxP spacer region in Cre-mediated recombination. *Gene* 216, 55–65. [https://doi.org/10.1016/S0378-1119\(98\)00325-4](https://doi.org/10.1016/S0378-1119(98)00325-4)
- Li, Y., Rivera, C.M., Ishii, H., Jin, F., Selvaraj, S., Lee, A.Y., Dixon, J.R., Ren, B., 2014. CRISPR Reveals a Distal Super-Enhancer Required for Sox2 Expression in Mouse Embryonic Stem Cells. *PLoS ONE* 9, e114485. <https://doi.org/10.1371/journal.pone.0114485>
- Lienert, F., Wirbelauer, C., Som, I., Dean, A., Mohn, F., Schübeler, D., 2011. Identification of genetic elements that autonomously determine DNA methylation states. *Nat Genet* 43, 1091–1097. <https://doi.org/10.1038/ng.946>
- Lim, F., Solvason, J.J., Ryan, G.E., Le, S.H., Jindal, G.A., Steffen, P., Jandu, S.K., Farley, E.K., 2024. Affinity-optimizing enhancer variants disrupt development. *Nature* 626, 151–159. <https://doi.org/10.1038/s41586-023-06922-8>
- Long, H.K., Osterwalder, M., Welsh, I.C., Hansen, K., Davies, J.O.J., Liu, Y.E., Koska, M., Adams, A.T., Aho, R., Arora, N., Ikeda, K., Williams, R.M., Sauka-Spengler, T., Porteus, M.H., Mohun, T., Dickel, D.E., Swigut, T., Hughes, J.R., Higgs, D.R., Visel, A., Selleri, L., Wysocka, J., 2020. Loss of Extreme Long-Range Enhancers in Human Neural Crest Drives a Craniofacial Disorder. *Cell Stem Cell* 27, 765–783.e14. <https://doi.org/10.1016/j.stem.2020.09.001>
- Loubiere, V., De Almeida, B.P., Pagani, M., Stark, A., 2023. Developmental and housekeeping transcriptional programs display distinct modes of enhancer-enhancer cooperativity in *Drosophila* (preprint). *Genomics*. <https://doi.org/10.1101/2023.10.10.561770>
- Low, B.E., Hosur, V., Lesbirel, S., Wiles, M.V., 2022. Efficient targeted transgenesis of large donor DNA into multiple mouse genetic backgrounds using bacteriophage Bxb1 integrase. *Sci Rep* 12, 5424. <https://doi.org/10.1038/s41598-022-09445-w>
- Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S.A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., Spielmann, M., Visel, A., Mundlos, S., 2015. Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell* 161, 1012–1025. <https://doi.org/10.1016/j.cell.2015.04.004>
- Mach, P., Kos, P.I., Zhan, Y., Cramard, J., Gaudin, S., Tünnermann, J., Marchi, E., Eglinger, J., Zuin, J., Kryzhanovska, M., Smallwood, S., Gelman, L., Roth, G., Nora, E.P., Tiana, G., Giorgetti, L., 2022. Cohesin and CTCF control the dynamics of chromosome folding. *Nat Genet* 54, 1907–1918. <https://doi.org/10.1038/s41588-022-01232-7>
- Maher, B., 2008. Personal genomes: The case of the missing heritability. *Nature* 456, 18–21. <https://doi.org/10.1038/456018a>
- Marson, A., Foreman, R., Chevalier, B., Bilodeau, S., Kahn, M., Young, R.A., Jaenisch, R., 2008. Wnt signaling promotes reprogramming of somatic cells to pluripotency. *Cell Stem Cell* 3, 132–135. <https://doi.org/10.1016/j.stem.2008.06.019>
- Martinez-Ara, M., Comoglio, F., Van Arensbergen, J., Van Steensel, B., 2022. Systematic analysis of intrinsic enhancer-promoter compatibility in the mouse genome. *Molecular Cell* 82, 2519–2531.e6. <https://doi.org/10.1016/j.molcel.2022.04.009>
- Martinez-Ara, M., Comoglio, F., van Steensel, B., 2023. Large-scale analysis of the integration of enhancer-enhancer signals by promoters. <https://doi.org/10.1101/2023.08.11.552995>
- Matreyek, K.A., Stephany, J.J., Chiasson, M.A., Hasle, N., Fowler, D.M., 2019. An improved platform for functional assessment of large protein libraries in mammalian cells. *Nucleic Acids Research* gkz910. <https://doi.org/10.1093/nar/gkz910>

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutayavin, T., Stehling-Sun, S., Johnson, A.K., Canfield, T.K., Giste, E., Diegel, M., Bates, D., Hansen, R.S., Neph, S., Sabo, P.J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S.R., Kaul, R., Stamatoyannopoulos, J.A., 2012. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* 337, 1190–1195. <https://doi.org/10.1126/science.1222794>

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C.G., Kinney, J.B., Kellis, M., Lander, E.S., Mikkelsen, T.S., 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* 30, 271–277. <https://doi.org/10.1038/nbt.2137>

Missirlis, P.I., Smailus, D.E., Holt, R.A., 2006. A high-throughput screen identifying sequence and promiscuity characteristics of the loxP spacer region in Cre-mediated recombination. *BMC Genomics* 7, 73. <https://doi.org/10.1186/1471-2164-7-73>

Mohammed, J., Arora, N., Matthews, H.S., Hansen, K., Bader, M., Walsh, S., Shaffer, J.R., Weinberg, S.M., Swigut, T., Claes, P., Selleri, L., Wysocka, J., 2024. A common cis-regulatory variant impacts normal-range and disease-associated human facial shape through regulation of PKDCC during chondrogenesis. *eLife* 13, e82564. <https://doi.org/10.7554/eLife.82564>

Nasser, J., Bergman, D.T., Fulco, C.P., Guckelberger, P., Doughty, B.R., Patwardhan, T.A., Jones, T.R., Nguyen, T.H., Ulirsch, J.C., Lekschas, F., Mualim, K., Natri, H.M., Weeks, E.M., Munson, G., Kane, M., Kang, H.Y., Cui, A., Ray, J.P., Eisenhaure, T.M., Collins, R.L., Dey, K., Pfister, H., Price, A.L., Epstein, C.B., Kundaje, A., Xavier, R.J., Daly, M.J., Huang, H., Finucane, H.K., Hacohen, N., Lander, E.S., Engreitz, J.M., 2021. Genome-wide enhancer maps link risk variants to disease genes. *Nature* 593, 238–243. <https://doi.org/10.1038/s41586-021-03446-x>

Ordofiez, R., Zhang, W., Ellis, G., Zhu, Y., Ashe, H.J., Ribeiro-dos-Santos, A.M., Brosh, R., Huang, E., Hogan, M.S., Boeke, J.D., Maurano, M.T., 2023. Genomic context sensitizes regulatory elements to genetic disruption. <https://doi.org/10.1101/2023.07.02.547201>

Pandey, S., Gao, X.D., Krasnow, N.A., McElroy, A., Tao, Y.A., Duby, J.E., Steinbeck, B.J., McCreary, J., Pierce, S.E., Tolar, J., Meissner, T.B., Chaikof, E.L., Osborn, M.J., Liu, D.R., 2024. Efficient site-specific integration of large genes in mammalian cells via continuously evolved recombinases and prime editing. *Nat. Biomed. Eng.* <https://doi.org/10.1038/s41551-024-01227-1>

Patwardhan, R.P., Hiatt, J.B., Witten, D.M., Kim, M.J., Smith, R.P., May, D., Lee, C., Andrie, J.M., Lee, S.-I., Cooper, G.M., Ahituv, N., Pennacchio, L.A., Shendure, J., 2012. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* 30, 265–270. <https://doi.org/10.1038/nbt.2136>

Patwardhan, R.P., Lee, C., Litvin, O., Young, D.L., Pe'er, D., Shendure, J., 2009. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol* 27, 1173–1175. <https://doi.org/10.1038/nbt.1589>

Peng, T., Zhai, Y., Atlasi, Y., Ter Huurne, M., Marks, H., Stunnenberg, H.G., Megchelenbrink, W., 2020. STARR-seq identifies active, chromatin-masked, and dormant enhancers in pluripotent mouse embryonic stem cells. *Genome Biol* 21, 243. <https://doi.org/10.1186/s13059-020-02156-3>

Pombo, A., Nicodemi, M., 2014. Physical mechanisms behind the large scale features of chromatin organization. *Transcription* 5, e28447. <https://doi.org/10.4161/trns.28447>

Pomp, W., Meeussen, J.V.W., Lenstra, T.L., 2024. Transcription factor exchange enables prolonged transcriptional bursts. *Molecular Cell* 84, 1036-1048.e9. <https://doi.org/10.1016/j.molcel.2024.01.020>

Prescott, S.L., Srinivasan, R., Marchetto, M.C., Grishina, I., Narvaiza, I., Selleri, L., Gage, F.H., Swigut, T., Wysocka, J., 2015. Enhancer Divergence and cis-Regulatory Evolution in the Human and Chimp Neural Crest. *Cell* 163, 68–83. <https://doi.org/10.1016/j.cell.2015.08.036>

Rauluseviciute, I., Riudavets-Puig, R., Blanc-Mathieu, R., Castro-Mondragon, J.A., Ferenc, K., Kumar, V., Lemma, R.B., Lucas, J., Chèneby, J., Baranasic, D., Khan, A., Fornes, O., Gundersen, S., Johansen, M., Hovig, E., Lenhard, B., Sandelin, A., Wasserman, W.W., Parcy, F., Mathelier, A., 2024. JASPAR 2024:

20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* 52, D174–D182. <https://doi.org/10.1093/nar/gkad1059>

Raymond, C.S., Soriano, P., 2007. High-Efficiency FLP and Φ C31 Site-Specific Recombination in Mammalian Cells. *PLoS ONE* 2, e162. <https://doi.org/10.1371/journal.pone.0000162>

Reiter, F., De Almeida, B.P., Stark, A., 2023. Enhancers display constrained sequence flexibility and context-specific modulation of motif function. *Genome Res.* 33, 346–358. <https://doi.org/10.1101/gr.277246.122>

Rideout, W.M., Wakayama, T., Wutz, A., Eggan, K., Jackson-Grusby, L., Dausman, J., Yanagimachi, R., Jaenisch, R., 2000. Generation of mice from wild-type and targeted ES cells by nuclear cloning. *Nat Genet* 24, 109–110. <https://doi.org/10.1038/72753>

Ringrose, L., Lounnas, V., Ehrlich, L., Buchholz, F., Wade, R., Stewart, A.F., 1998. Comparative kinetic analysis of FLP and cre recombinases: mathematical models for DNA binding and recombination. *Journal of Molecular Biology* 284, 363–384. <https://doi.org/10.1006/jmbi.1998.2149>

Rinzema, N.J., Sofiadis, K., Tjalsma, S.J.D., Verstegen, M.J.A.M., Oz, Y., Valdes-Quezada, C., Felder, A.-K., Filipovska, T., Van Der Elst, S., De Andrade Dos Ramos, Z., Han, R., Krijger, P.H.L., De Laat, W., 2022. Building regulatory landscapes reveals that an enhancer can recruit cohesin to create contact domains, engage CTCF sites and activate distant genes. *Nat Struct Mol Biol* 29, 563–574. <https://doi.org/10.1038/s41594-022-00787-7>

Russell, J.P., Chang, D.W., Tretiakova, A., Padidam, M., 2006. Phage Bxb1 Integrase Mediates Highly Efficient Site-Specific Recombination in Mammalian Cells. *BioTechniques* 40, 460–464. <https://doi.org/10.2144/000112150>

Sahu, B., Hartonen, T., Pihlajamaa, P., Wei, B., Dave, K., Zhu, F., Kaasinen, E., Lidschreiber, K., Lidschreiber, M., Daub, C.O., Cramer, P., Kivioja, T., Taipale, J., 2022. Sequence determinants of human gene regulatory elements. *Nat Genet* 54, 283–294. <https://doi.org/10.1038/s41588-021-01009-4>

Sauer, B., 1996. Multiplex Cre/lox recombination permits selective site-specific DNA targeting to both a natural and an engineered site in the yeast genome. *Nucleic Acids Res* 24, 4608–4613.

Schaft, J., Ashery-Padan, R., Van Der Hoeven, F., Gruss, P., Stewart, A.F., 2001. Efficient FLP recombination in mouse ES cells and oocytes. *Genesis* 31, 6–10. <https://doi.org/10.1002/gene.1076>

Schlake, T., Bode, J., 1994. Use of Mutated FLP Recognition Target (FRT) Sites for the Exchange of Expression Cassettes at Defined Chromosomal Loci. *Biochemistry* 33, 12746–12751. <https://doi.org/10.1021/bi00209a003>

Semrau, S., Goldmann, J.E., Soumillon, M., Mikkelsen, T.S., Jaenisch, R., van Oudenaarden, A., 2017. Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells. *Nat Commun* 8, 1096. <https://doi.org/10.1038/s41467-017-01076-4>

Senecoff, J.F., Cox, M.M., 1986. Directionality in FLP protein-promoted site-specific recombination is mediated by DNA-DNA pairing. *J Biol Chem* 261, 7380–7386.

Shaw, D., Serrano, L., Lluch-Senar, M., 2021. Lox'd in translation: contradictions in the nomenclature surrounding common lox-site mutants and their implications in experiments. *Microbiology* 167. <https://doi.org/10.1099/mic.0.000997>

Shimshek, D.R., Kim, J., Hübner, M.R., Spergel, D.J., Buchholz, F., Casanova, E., Stewart, A.F., Seeburg, P.H., Sprengel, R., 2002. Codon-improved Cre recombinase (iCre) expression in the mouse. *Genesis* 32, 19–26. <https://doi.org/10.1002/gene.10023>

Song, W., Sharan, R., Ovcharenko, I., 2019. The first enhancer in an enhancer chain safeguards subsequent enhancer-promoter contacts from a distance. *Genome Biol* 20, 197. <https://doi.org/10.1186/s13059-019-1808-y>

Stark, W.M., 2014. The Serine Recombinases. *Microbiol Spectr* 2, 2.6.05. <https://doi.org/10.1128/microbiolspec.MDNA3-0046-2014>

- Sternberg, N., Hamilton, D., 1981. Bacteriophage P1 site-specific recombination. *Journal of Molecular Biology* 150, 467–486. [https://doi.org/10.1016/0022-2836\(81\)90375-2](https://doi.org/10.1016/0022-2836(81)90375-2)
- Stolper, R.J., Tsang, F.H., Georgiades, E., Hansen, L.L.P., Downes, D.J., Harrold, C.L., Hughes, J.R., Beagrie, R.A., Davies, B., Kassouf, M.T., Higgs, D.R., 2023. Loop extrusion by cohesin plays a key role in enhancer-activated gene expression during differentiation (preprint). *Molecular Biology*. <https://doi.org/10.1101/2023.09.07.556660>
- Suntsova, M.V., Buzdin, A.A., 2020. Differences between human and chimpanzee genomes and their implications in gene expression, protein functions and biochemical properties of the two species. *BMC Genomics* 21, 535. <https://doi.org/10.1186/s12864-020-06962-8>
- Suter, D.M., Molina, N., Naef, F., Schibler, U., 2011. Origins and consequences of transcriptional discontinuity. *Current Opinion in Cell Biology*.
- Takahashi, K., Yamanaka, S., 2006. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* 126, 663–676. <https://doi.org/10.1016/j.cell.2006.07.024>
- Taskiran, I.I., Spanier, K.I., Dickmanken, H., Kempynck, N., Pančíková, A., Ekşi, E.C., Hulselmans, G., Ismail, J.N., Theunis, K., Vandepoel, R., Christiaens, V., Mauduit, D., Aerts, S., 2024. Cell-type-directed design of synthetic enhancers. *Nature* 626, 212–220. <https://doi.org/10.1038/s41586-023-06936-2>
- Taylor, T., Sikorska, N., Shchuka, V.M., Chahar, S., Ji, C., Macpherson, N.N., Moorthy, S.D., De Kort, M.A.C., Mullany, S., Khader, N., Gillespie, Z.E., Langroudi, L., Tobias, I.C., Lenstra, T.L., Mitchell, J.A., Sexton, T., 2022. Transcriptional regulation and chromatin architecture maintenance are decoupled functions at the *Sox2* locus. *Genes Dev.* 36, 699–717. <https://doi.org/10.1101/gad.349489.122>
- Thomas, H.F., Feng, S., Huber, M., Loubiere, V., Vanina, D., Pitasi, M., Stark, A., Buecker, C., 2023. Enhancer cooperativity can compensate for loss of activity over large genomic distances (preprint). *Molecular Biology*. <https://doi.org/10.1101/2023.12.06.570399>
- Tsang, F.H., Stolper, R.J., Hanifi, M., Cornell, L.J., Francis, H.S., Davies, B., Higgs, D.R., Kassouf, M.T., 2023. The characteristics of CTCF binding sequences contribute to enhancer blocking activity. <https://doi.org/10.1101/2023.09.06.556325>
- Uffelmann, E., Huang, Q.Q., Munung, N.S., De Vries, J., Okada, Y., Martin, A.R., Martin, H.C., Lappalainen, T., Posthuma, D., 2021. Genome-wide association studies. *Nat Rev Methods Primers* 1, 59. <https://doi.org/10.1038/s43586-021-00056-9>
- Umlauf, S.W., Cox, M.M., 1988. The functional significance of DNA sequence structure in a site-specific genetic recombination reaction. *The EMBO Journal* 7, 1845–1852. <https://doi.org/10.1002/j.1460-2075.1988.tb03017.x>
- Van Arensbergen, J., FitzPatrick, V.D., De Haas, M., Pagie, L., Sluimer, J., Bussemaker, H.J., Van Steensel, B., 2017. Genome-wide mapping of autonomous promoter activity in human cells. *Nat Biotechnol* 35, 145–153. <https://doi.org/10.1038/nbt.3754>
- Vandermeulen, G., Marie, C., Scherman, D., Pr  at, V., 2011. New Generation of Plasmid Backbones Devoid of Antibiotic Resistance Marker for Gene Therapy Trials. *Molecular Therapy* 19, 1942–1949. <https://doi.org/10.1038/mt.2011.182>
- Wang, X., Xu, Z., Tian, Z., Zhang, X., Xu, D., Li, Q., Zhang, J., Wang, T., 2017. The EF-1 α promoter maintains high-level transgene expression from episomal vectors in transfected CHO-K1 cells. *Journal of Cellular and Molecular Medicine* 21, 3044–3054. <https://doi.org/10.1111/jcmm.13216>
- Weintraub, A.S., Li, C.H., Zamudio, A.V., Sigova, A.A., Hannett, N.M., Day, D.S., Abraham, B.J., Cohen, M.A., Nabeta, B., Buckley, D.L., Guo, Y.E., Hnisz, D., Jaenisch, R., Bradner, J.E., Gray, N.S., Young, R.A., 2017. YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* 171, 1573–1588.e28. <https://doi.org/10.1016/j.cell.2017.11.008>
- Xiao, J.Y., Hafner, A., Boettiger, A.N., 2021. How subtle changes in 3D structure can create large changes in transcription. *eLife* 10, e64320. <https://doi.org/10.7554/eLife.64320>

- Xu, Z., Thomas, L., Davies, B., Chalmers, R., Smith, M., Brown, W., 2013. Accuracy and efficiency define Bxb1 integrase as the best of fifteen candidate serine recombinases for the integration of DNA into the human genome. *BMC Biotechnol* 13, 87. <https://doi.org/10.1186/1472-6750-13-87>
- Yarmolinsky, M., Hoess, R., 2015. The Legacy of Nat Sternberg: The Genesis of Cre- *lox* Technology. *Annu. Rev. Virol.* 2, 25–40. <https://doi.org/10.1146/annurev-virology-100114-054930>
- Yeo, J.-C., Ng, H.-H., 2013. The transcriptional regulation of pluripotency. *Cell Res* 23, 20–32. <https://doi.org/10.1038/cr.2012.172>
- Zabidi, M.A., Arnold, C.D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., Stark, A., 2015. Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* 518, 556–559. <https://doi.org/10.1038/nature13994>
- Zhang, S., 2014. Sox2, a key factor in the regulation of pluripotency and neural differentiation. *WJSC* 6, 305. <https://doi.org/10.4252/wjsc.v6.i3.305>
- Zheng, Y., Li, Y., Zhou, K., Li, T., VanDusen, N.J., Hua, Y., 2024. Precise genome-editing in human diseases: mechanisms, strategies and applications. *Sig Transduct Target Ther* 9, 47. <https://doi.org/10.1038/s41392-024-01750-2>
- Zhou, X., Di Stasi, A., Brenner, M.K., 2015. iCaspase 9 Suicide Gene System, in: Walther, W., Stein, U. (Eds.), *Gene Therapy of Solid Cancers: Methods and Protocols*. Springer, New York, NY, pp. 87–105. https://doi.org/10.1007/978-1-4939-2727-2_6
- Zuin, J., Roth, G., Zhan, Y., Cramard, J., Redolfi, J., Piskadlo, E., Mach, P., Kryzhanovska, M., Tihanyi, G., Kohler, H., Eder, M., Leemans, C., Van Steensel, B., Meister, P., Smallwood, S., Giorgetti, L., 2022. Nonlinear control of transcription through enhancer–promoter interactions. *Nature* 604, 571–577. <https://doi.org/10.1038/s41586-022-04570-y>

Supplementary Figures

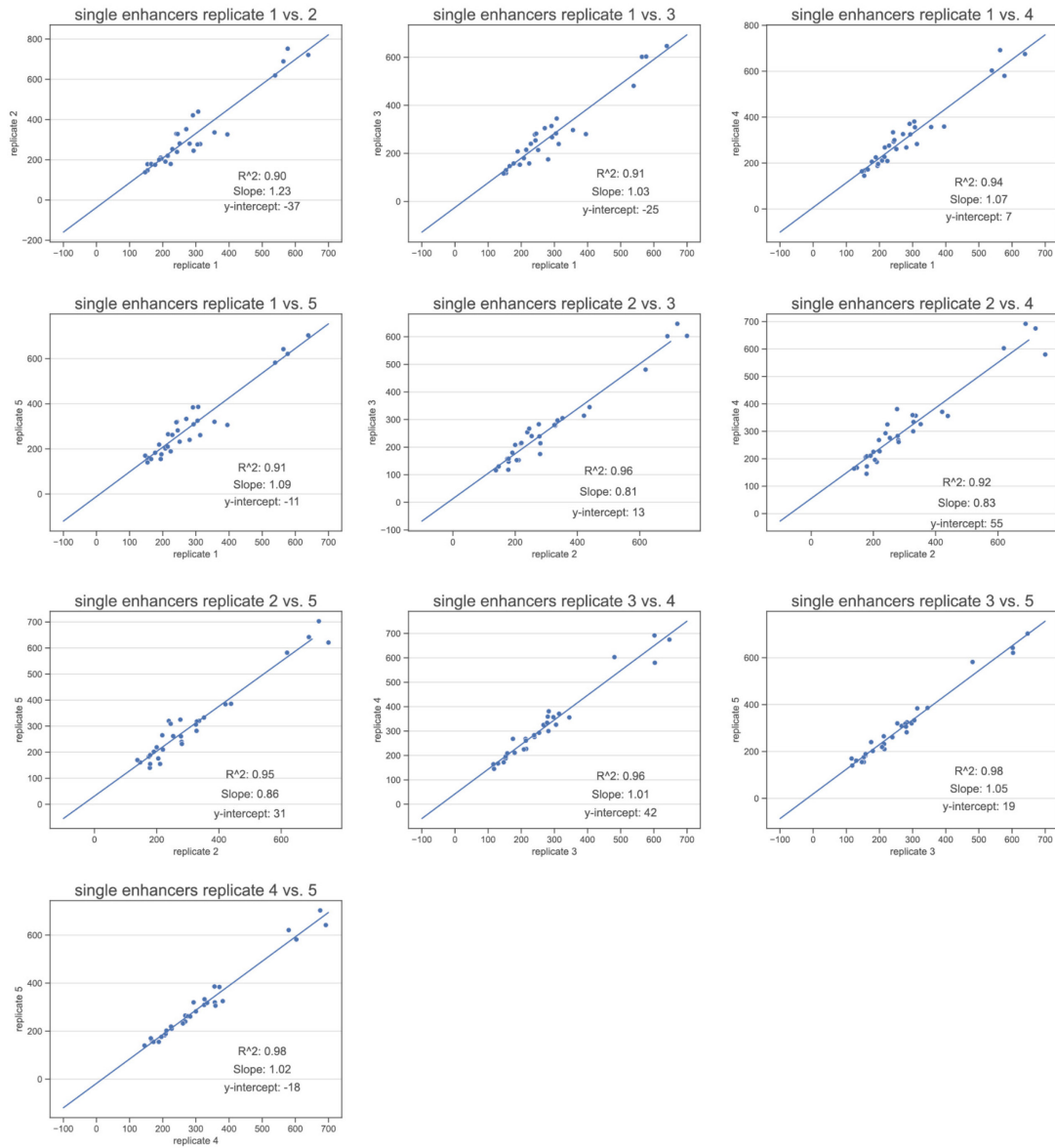


Figure S1: Linear regression of single enhancer replicates. Calculated with raw values.

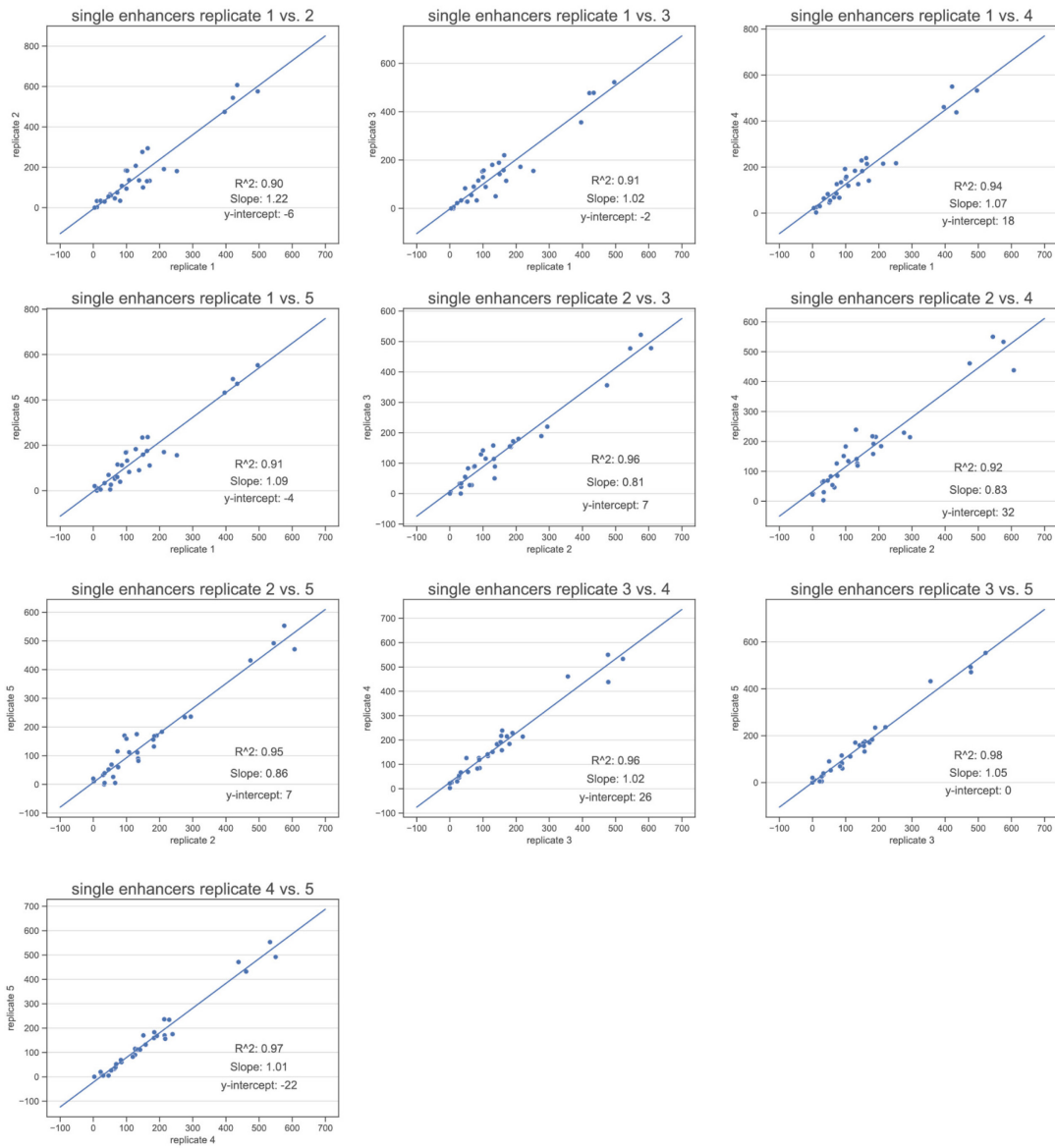


Figure S2: Linear regression of single enhancer replicates. Calculated with background (#1) subtracted values.

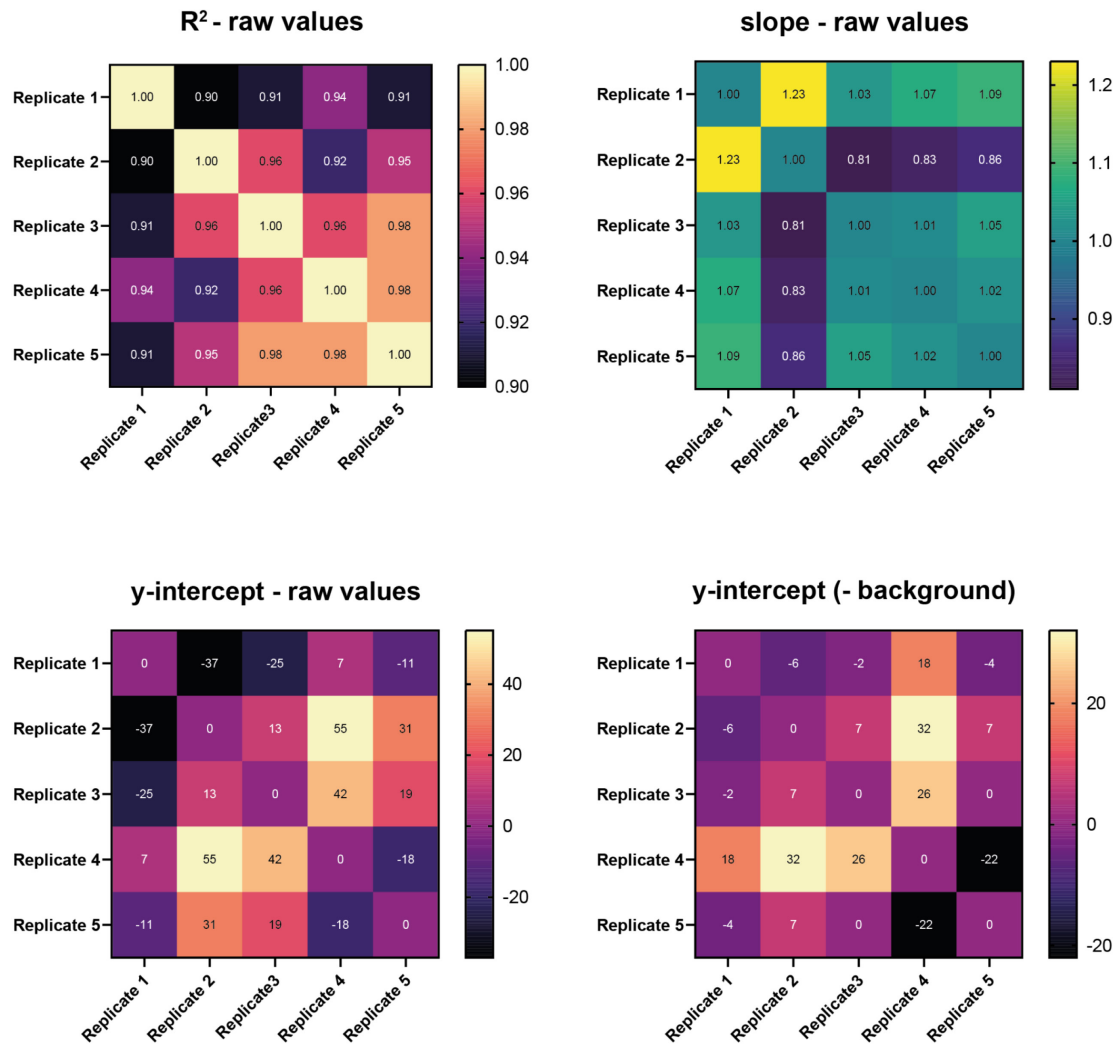


Figure S3: Linear regression of single enhancers - Overview.

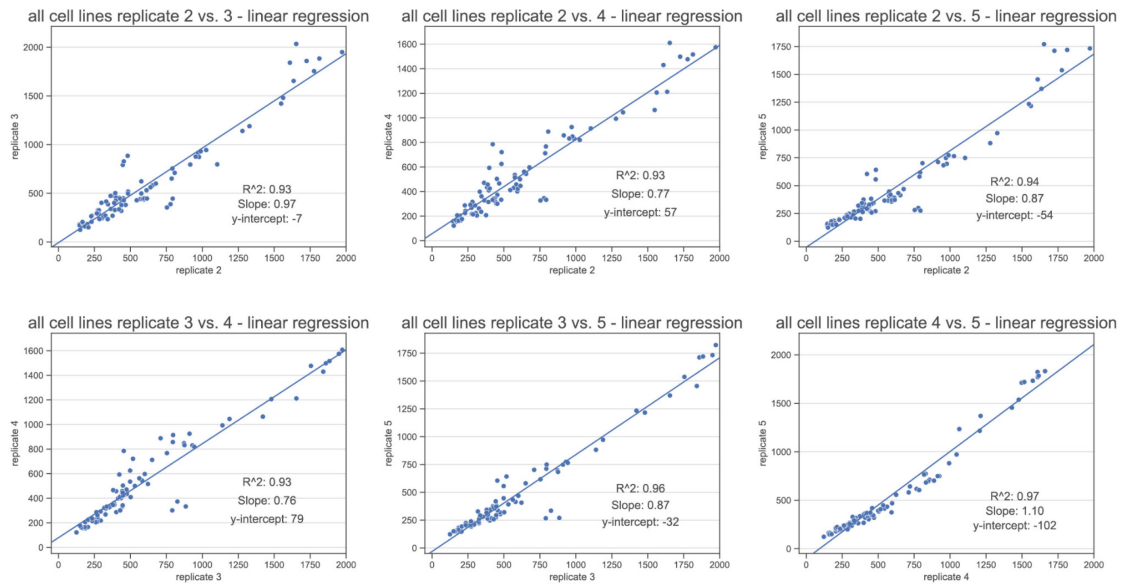


Figure S4: Linear regression of dual enhancer replicates. Calculated with raw values.

Linear Regression dual enhancers raw values.

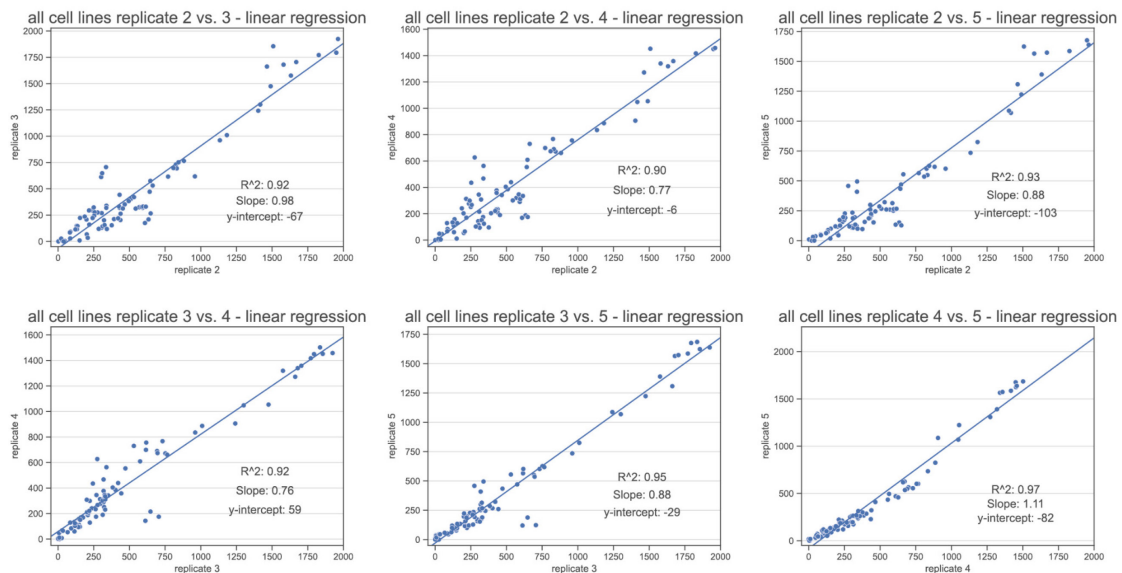


Figure S5: Linear regression of dual enhancer replicates. Calculated with background (#1) subtracted values.

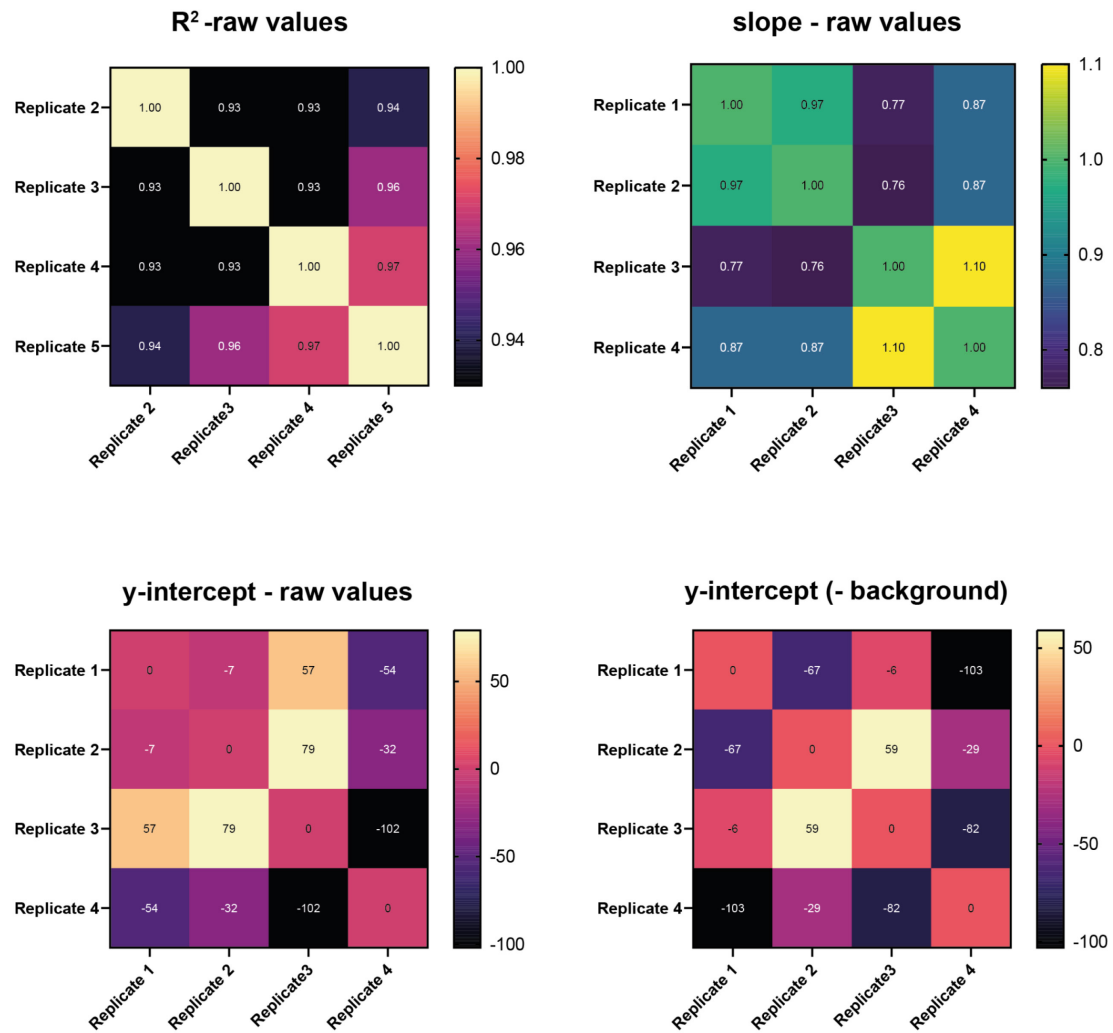


Figure S6: Linear regression of dual enhancers - Overview.

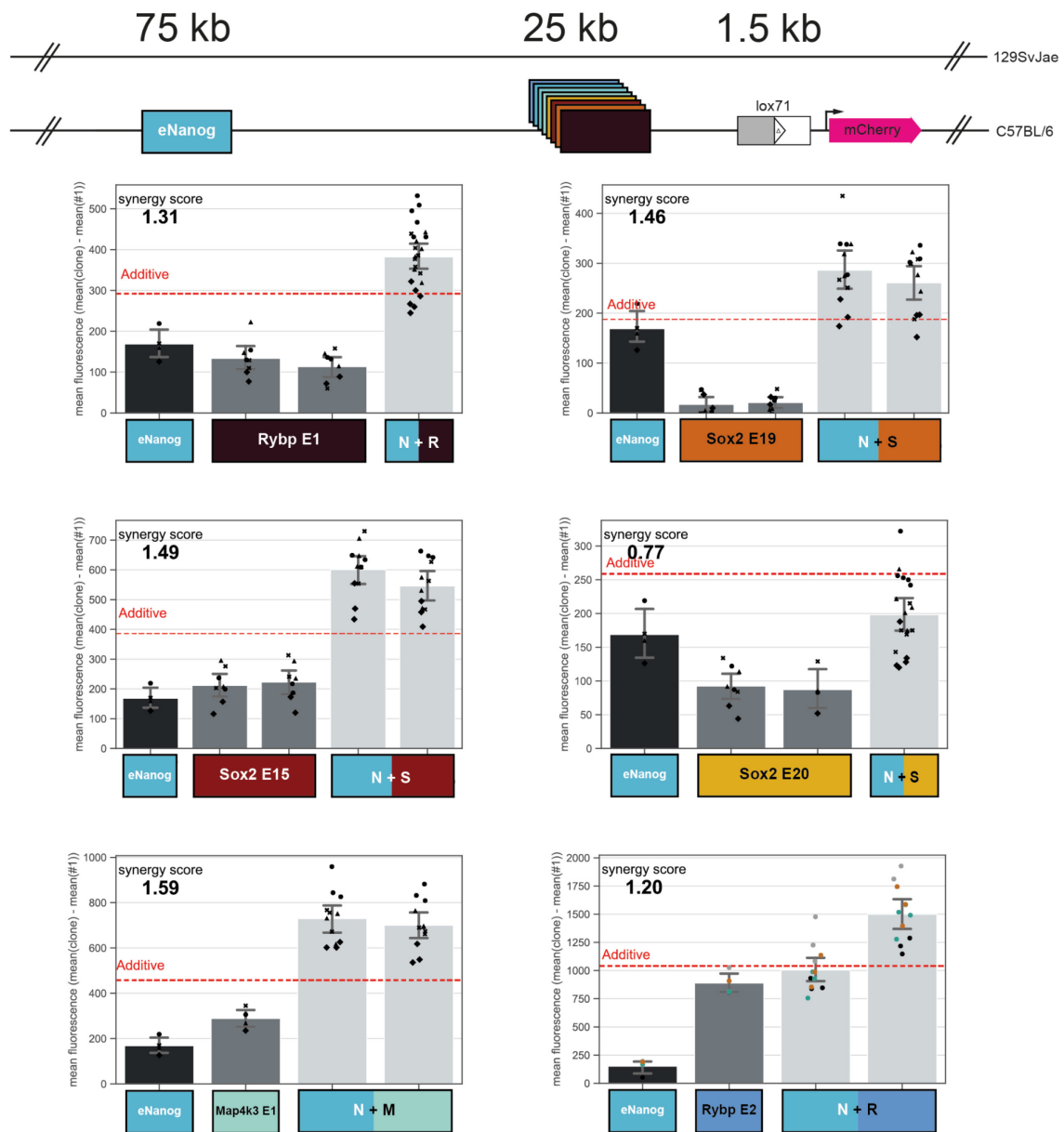


Figure S7: Dual enhancer integrations. Independent integrations are shown as separate bars.

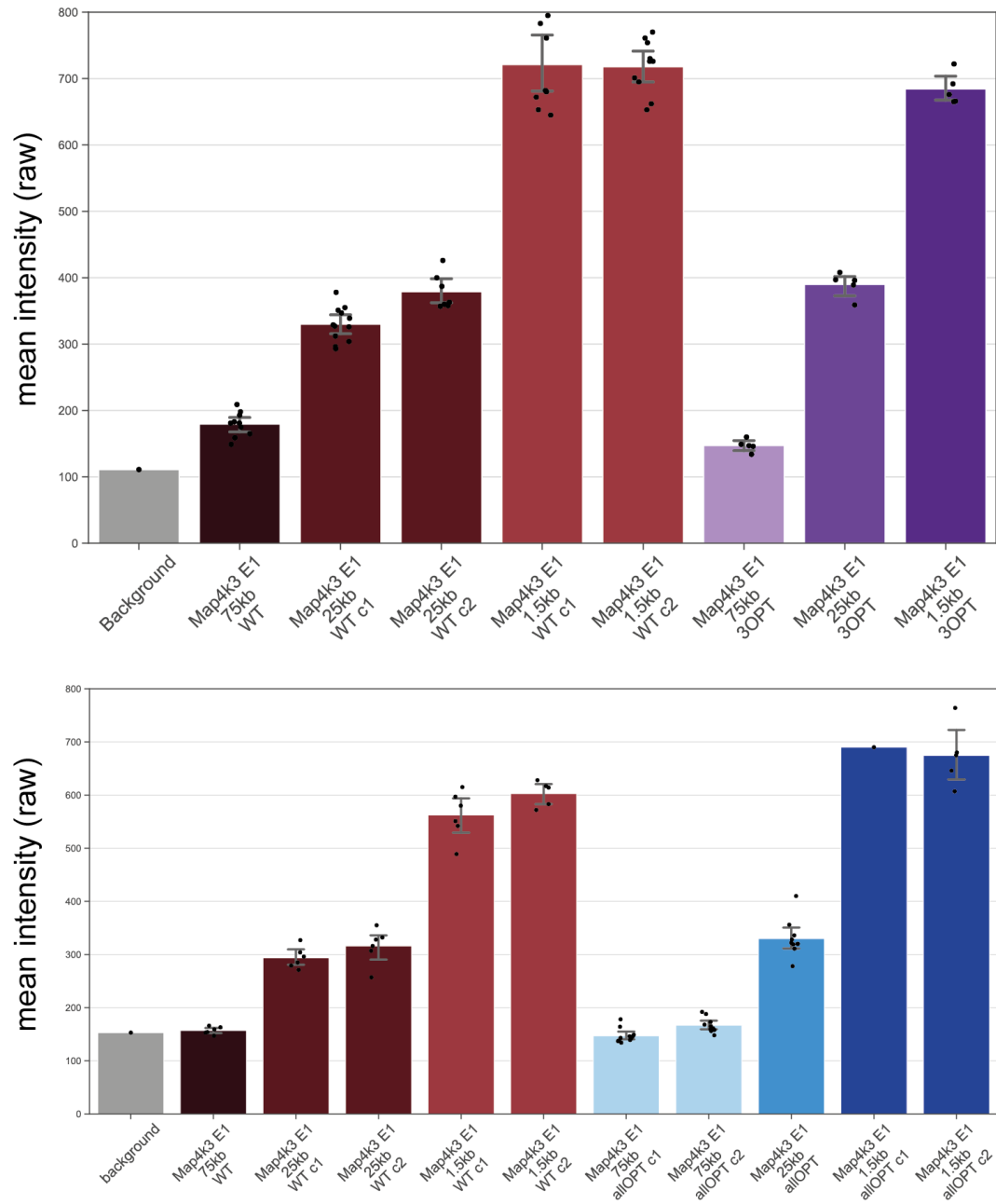


Figure S8: Raw values of *Map4k3* WT, 3OPT and allOPT.

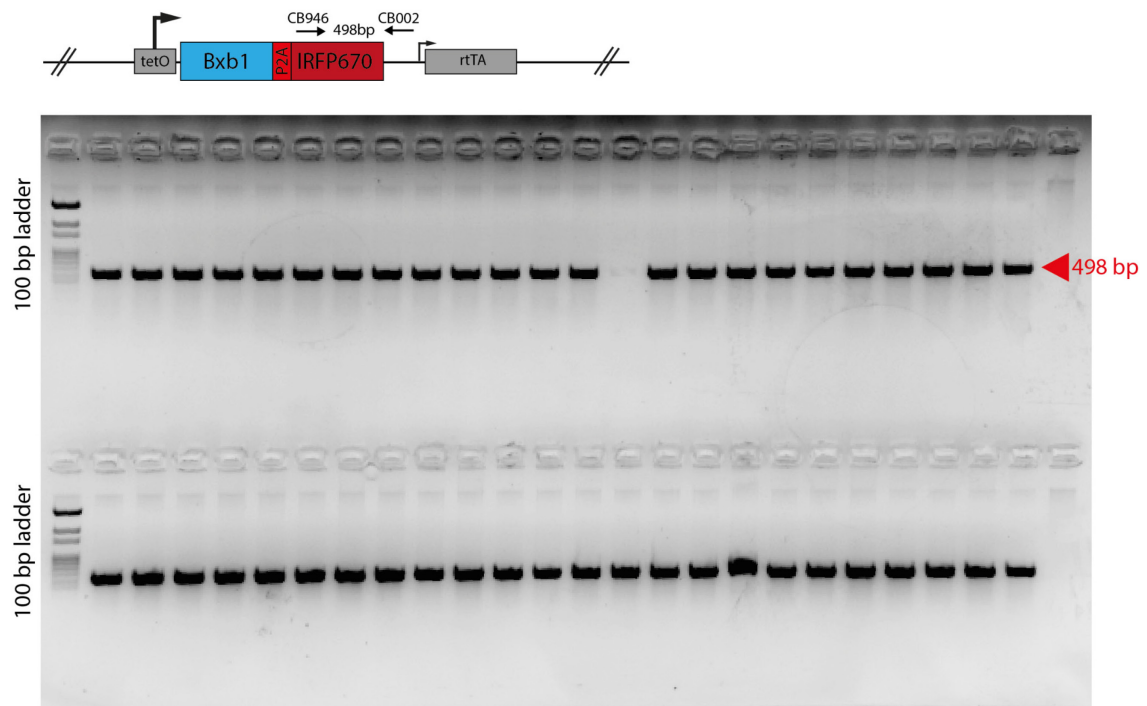


Figure S9: Genotyping of iBxb1.

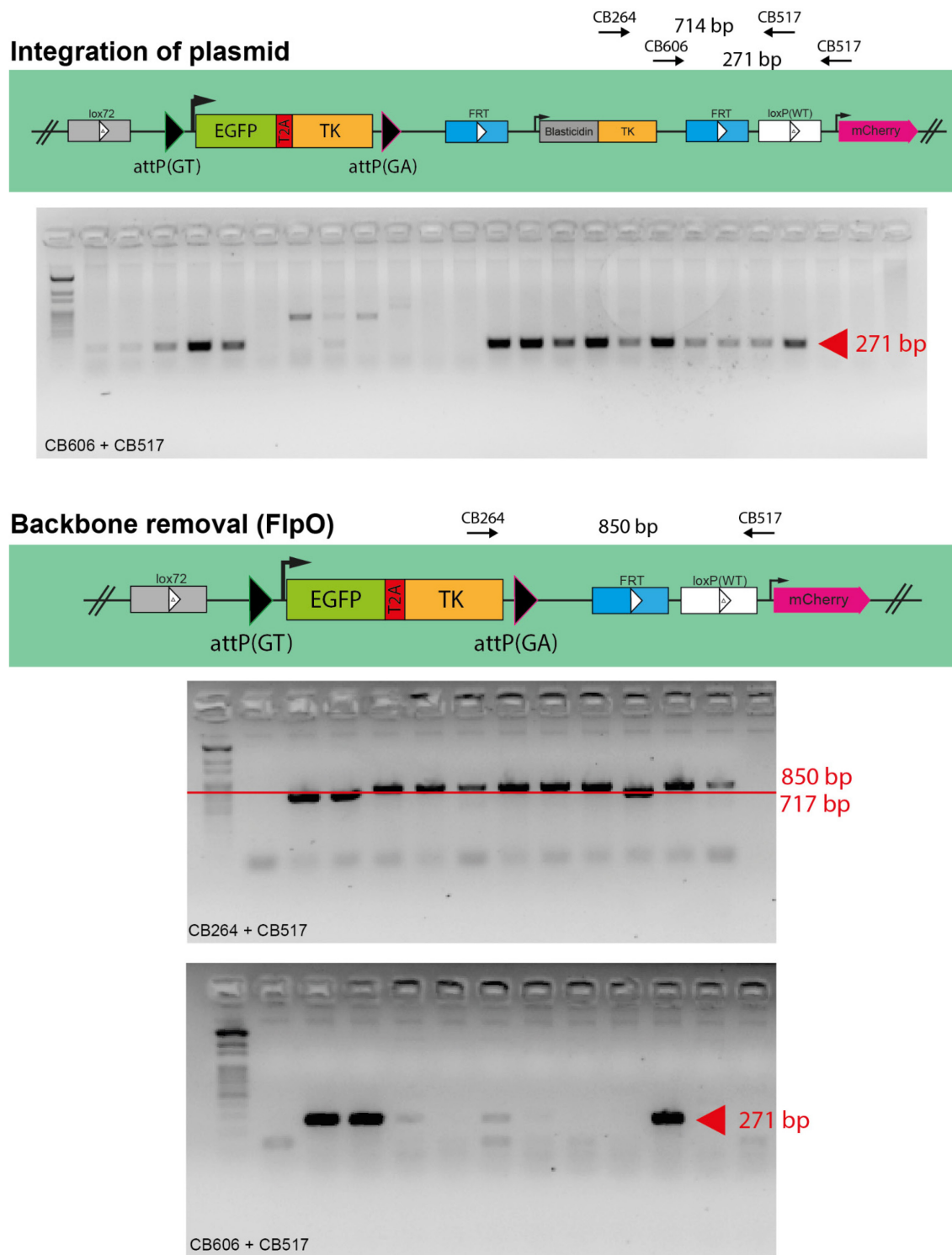


Figure S10: Genotyping of integration of LP into iBxb1-C4. Upper gel shows genotyping for Cre-mediated insertion of targeting vector carrying the LP. Lower gels show a PCR after FlpO treatment testing the absence of the backbone (upper) as well as presence of the backbone (lower).

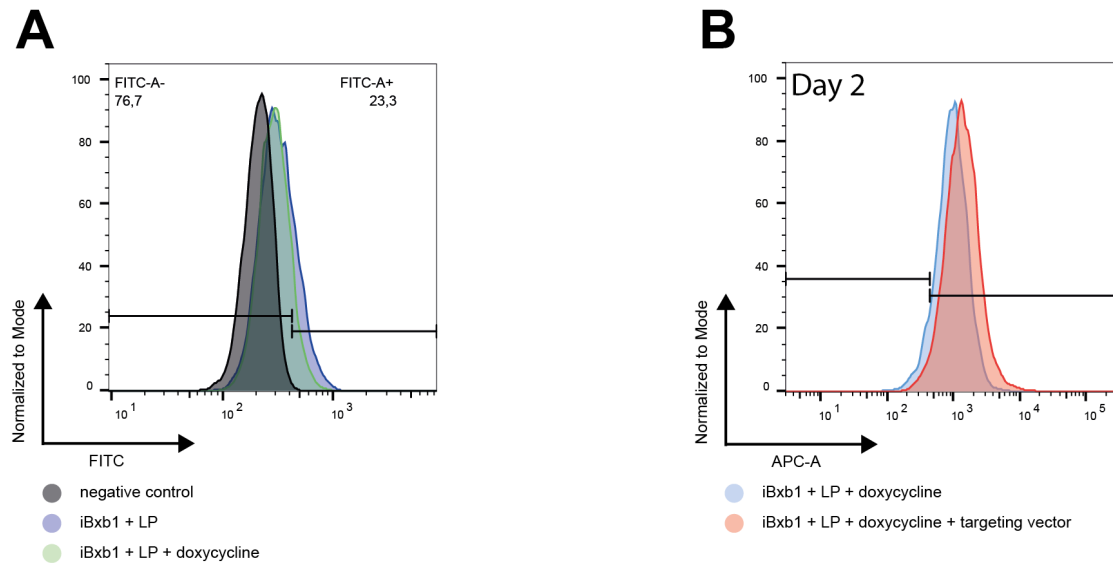


Figure S11: GFP problem, IRFP problem. A) Doxycycline treatment and the following Bxb1 expression downregulates GFP slightly. B) Transfection boosts measured APC-A (IRFP) intensity at day 2.

Supplementary Notes

Supplementary Note 1

Single Enhancer Normalization.

The most direct way to normalize against background mCherry expression within a dataset (same laser power, same device etc.) is to subtract the background (here #1).

Here is an enhancer at position 1 and position 2. The intensity measured is proportional to the concentration and the laser power.

$$\begin{aligned}I_{pos1} &\propto c_1 * P + c_{\#1} * P \\I_{pos2} &\propto c_2 * P + c_{\#1} * P \\I_{\#1} &\propto c_{\#1} * P\end{aligned}$$

Subtract the background.

$$\begin{aligned}I_{pos1-\#1} &\propto c_1 * P \\I_{pos2-\#1} &\propto c_2 * P\end{aligned}$$

When measured in the same conditions, P is the same. The difference or the ratio (FC) are both valid ways to compare them.

$$\begin{aligned}I_{pos1-\#1} - I_{pos2-\#1} &\propto (c_1 - c_2) * P \\ \frac{I_{pos1-\#1}}{I_{pos2-\#1}} &\propto \frac{c_1}{c_2}\end{aligned}$$

When comparing enhancer strengths within or between datasets with similar conditions, I usually used the difference of the background subtracted values (Fig. 4, Fig. 6, Fig. 9). When comparing the loss of activity over distance, I computed the fold change of the background subtracted values. That is because we are not interested in the absolute loss, but the relative loss of activity over distance.

Comparing Enhancer Strength Across Datasets with Different Conditions.

In Chapter I and Chapter II, I compared enhancers from measurements that happened more than a year apart (Fig. 5, Fig. 10). Clearly, the laser powers were changed. Because the laser power is proportional to the intensity, I can remove the dependency on power by dividing the background subtracted intensity by the background.

$$\begin{aligned}I_{dataset1-\#1/\#1} &\propto \frac{c_1 * P}{c_{\#1} * P} \propto \frac{c_1}{c_{\#1}} \\ I_{dataset2-\#1/\#1} &\propto \frac{c_2 * P}{c_{\#1} * P} \propto \frac{c_2}{c_{\#1}}\end{aligned}$$

Now the values are comparable between datasets but represent a fold-enrichment over the background.

Dual Enhancer Normalization.

The idea is to compare observed dual enhancer activities to the sum of the single enhancer activities, assuming an additive model (Fig. 6).

To compute an expected value, subtract the background from the single enhancers and add the background subtracted intensities.

$$\begin{aligned} I_1 &\propto c_1 * P + c_{\#1} * P - c_{\#1} * P \\ I_2 &\propto c_2 * P + c_{\#1} * P - c_{\#1} * P \end{aligned}$$

$$\begin{aligned} I_{1-\#1} &\propto c_1 * P \\ I_{2-\#1} &\propto c_2 * P \end{aligned}$$

$$I_{pos1-\#1} - I_{pos2-\#1} \propto (c_1 + c_2) * P$$

To compute the observed value, subtract the background from the dual enhancer cell line.

$$\begin{aligned} I_{1,2} &\propto c_{1,2} * P + c_{\#1} * P - c_{\#1} * P \\ I_{1,2-\#1} &\propto c_{1,2} * P \end{aligned}$$

Calculate the fold-change (synergy score).

$$FC = \frac{Observed}{Expected}$$

$$\frac{I_{1,2-\#1}}{I_{pos1-\#1} - I_{pos2-\#1}} = \frac{c_{1,2}}{(c_1 + c_2)}$$

Extrapolating Values for Linear Regression.

In Fig. 5B we have plotted the absolute values of the observed synergy vs. the absolute values that would be expected if the 75 kb position would combine additively with the 1.5 kb position. For the dots in blue, the 1.5 kb position was not measured recently, and had to be taken from a previous dataset (Thomas et al., 2023). To stay comparable, I used the fold-enrichment of background subtracted intensities over background.

$$I_{1.5kb-\#1+75kb-\#1_old/\#1_old} \propto \frac{c_{1.5kb+75kb}}{c_{\#1_old}}$$

This value was simply multiplied with the background of the synergy measurements ($\#1_new$).

$$\frac{C_{1.5kb+75kb}}{C_{\#1_old}} * C_{\#1_new} \propto I_{abs-old} \rightarrow \frac{C_{1.5kb+75kb}}{C_{\#1_new}} \propto I_{abs-new}$$