The dispersion of aromatic residues in TF IDRs controls a molecular trade-off between activity and specificity

Inaugural-Dissertation
to obtain the academic degree
Doctor rerum naturalium (Dr. rer. nat.)

submitted to the Department of Biology, Chemistry, Pharmacy
of Freie Universität Berlin

by

Julian Naderi

2024

This work was conducted between November 2019 and June 2024 under the supervision of Denes Hnisz, PhD at the Max Planck Institute for Molecular Genetics in Berlin, Germany.

1st reviewer:	Denes Hnisz, PhD
2nd reviewer:	Prof. Dr. Sigmar Stricker
Date of defense:	25 September 2024

Acknowledgements

What a ride! When I think back to my first days at the MPIMG, I see myself standing in front the whiteboard in the hallway, experiencing the bootcamp à la Denes. I remember it being terrifying presenting papers you selected in a one-on-one. In retrospect, I noticed that these bootcamps tell a lot about you and your relationship with your students. I got the impression that you don't do this to test our knowledge or to teach us how to think about the most important publications in the field. I think you invest the time because you care about how we think and who we are on a personal and scientific level. This feeling accompanied me throughout my entire time as a student in your lab.

Over the years, I have experienced many highs and, inevitably, a few lows. I am privileged to say the highs far outnumbered the lows, and even in those moments of doubt, I never felt left alone. Thank you, Denes, for teaching me to think with clarity and scientific rigor, for showing me the importance of taking responsibility for my actions, and for always making me feel supported. I admire you as a scientist. Even after five years of working together I get inspired discussing ideas and experiments, much like I did during those early bootcamp days. I wish you all the best for the future. You once told me that a mentee's goal is to surpass their mentor. I am convinced you will achieve this, and I will strive to reach this goal myself.

My journey was far from a solo endeavor. I am immensely proud to have been part of an extraordinary department, surrounded by colleagues who I am certain will shape the future of science. Many thanks go to Alex Meissner for creating such a unique environment that allowed me to conduct top-level science with unparalleled freedom. Special gratitude to Abhi and Adri, my late-night science buddies, for our discussions about scientific problems no one else cares about for a reason, yet were fascinating for us. I am pretty sure that some these discussions saved me from turning away from science in particularly dry sections of my time at the institute. I am already excited to see what you will achieve in the future.

Beyond colleagues, I found friends. I am deeply thankful to each one of you - Philine, Christina, Maxi, Ida, Danny and Fabian. You turned each day into a shared adventure rather than just another day of work. I wish you all the best. You are a remarkable group of people. Keep up your attitude!

To the Hnisz lab, we share a bond forged by struggling with the most challenging scientific question: What does Denes want? I am deeply thankful to each one of you for your critical feedback and support. Special thanks to Alex and Yaotian who both tremendously helped me

during my time at the institute. I am proud to be a co-author on our paper which would not be close to how impressive it is now without your contribution. Additionally, thanks to Thomas, Martin, Gregoire and Gözde whose insights significantly shaped the trajectory of my research.

I want to thank my family and friends who, always took initiative to keep in touch during times I was extremely busy working. “I have to feed my cells” has been my excuse too often to dodge meeting you. Thank you for your persistence.

Lastly, I want to thank Sigmar. Your guidance throughout my years in the IMPRS has been very valuable. Your feedback during meetings has always been a tremendous help. Therefore, I am especially excited to have you on board during my defense.

Thank you all for being part of this remarkable experience.

I hereby declare that I alone am responsible for the content of my doctoral dissertation and that I have only used the sources or references cited in this dissertation.

Julian Naderi
Berlin, 25 September 2024

Parts of this dissertation have been previously published in:

Christou-Kent M, Cuartero S, Garcia-Cabau C, et al. CEBPA phase separation links transcriptional activity and 3D chromatin hubs. *Cell Rep*. 2023;42(8):112897. doi:10.1016/j.celrep.2023.112897.

Naderi, J., Magalhaes, A.P., Kibar, G. et al. An activity-specificity trade-off encoded in human transcription factors. *Nat Cell Biol* 26, 1309–1321 (2024). https://doi.org/10.1038/s41556-024-01411-0.

Table of contents

List of tables

List of Figures

Abstract

Cell-type specification is guided by transcription factors (TFs) that control specificity and activity of transcription by binding to regulatory DNA elements, such as enhancers. TFs are modular proteins, consisting of structured DNA-binding domains, which allow binding to TF-specific DNA motifs, and intrinsically disordered regions (IDRs), which often harbor “minimal activation domains” that control the activity of the TF. Both, DNA-binding specificity and transcriptional activity of TFs have been extensively studied using ChIP-sequencing and activation domain screens. However, the mechanisms by which TFs establish specific gene expression programs remain poorly understood, as TF-binding or the presence of a minimal activation domain within a TF IDR do not necessarily correlate with target gene expression. Recent studies suggest that non-linear sequence features of TF IDRs facilitate the formation of transcriptional condensates, contributing to both the activity and specificity of TFs, thus indicating a relationship between these two features.

In the following, I provide evidence for an evolutionary trade-off between the activity and specificity in human transcription factors encoded as submaximal dispersion of aromatic residues in their IDRs. I identified multiple human TFs that display significant dispersion of aromatic residues in their IDRs, resembling imperfect prion-like sequences. Mutation of dispersed aromatic residues reduced transcriptional activity, while increasing aromatic dispersion in multiple human TFs enhanced transcriptional activity. Furthermore, sequence optimization by increasing aromatic dispersion enhanced *in vitro* reprogramming efficiency, promoted liquid-like features of condensates formed *in vitro*, and led to more promiscuous DNA-binding in cells. Together with recent work on enhancer elements, these results suggest an important evolutionary role of suboptimal features in transcriptional control. I propose that engineering of amino acid features that alter condensation may be a strategy to optimize TF-dependent processes, including cellular reprogramming.

Zusammenfassung

Die Differenzierung von Zellen wird von Transkriptionsfaktoren (TF) gesteuert. TF kontrollieren Spezifität, sowie Aktivität der Transkription, indem sie regulatorische DNA-Elemente wie Enhancer binden. TF sind modulare Proteine die zum einen, aus strukturierten DNA-Bindedomänen bestehen, welche eine spezifische Interaktion mit TF-Bindemotiven ermöglichen. Des Weiteren, bestehen TF aus intrinsisch ungeordneten Regionen (IDR), die häufig Aktivierungsdomänen enthalten, welche die Aktivität des TF regulieren. Sowohl die DNA-Bindungsspezifität als auch die transkriptionelle Aktivität von TF wurden umfassend mittels ChIP-Sequenzierung und Aktivierungsdomänen Screens untersucht. Die Mechanismen, durch die TF eine spezifische Aktivierung ihrer Zielgene hervorrufen, sind jedoch nicht vollständig verstanden, da das Binden eines TF an ein Zielgen oder die Präsenz einer Aktivierungsdomäne in einem TF nicht zwangsläufig mit der Expression eines Zielgens korreliert. Jüngste Studien deuten darauf hin, dass nicht-lineare Sequenzmerkmale von TF IDR die Bildung von transkriptionellen Kondensaten erleichtern und somit zu Spezifität als auch zur Aktivität von TF beitragen. Dies impliziert eine bisher nicht charakterisierte Anhängigkeit dieser beiden Merkmale zueinander.

Im Folgenden präsentiere ich Nachweise eines evolutionären Kompromisses zwischen der Aktivität und Spezifität von humanen TF, welcher sich in Form einer submaximalen Verteilung aromatischer Aminosäuren in TF IDR äußert. In meiner Arbeit identifizierte ich mehrere humane TF welche eine signifikante Verteilung von aromatischen Aminosäuren in ihren IDR aufweisen und somit Prion-ähnlichen Sequenzen ähneln. Die Mutagenese dieser aromatischen Aminosäuren verringerte die transkriptionelle Aktivität, während eine optimale Verteilung aromatischer Aminosäuren in mehreren humanen TF deren transkriptionelle Aktivität steigerte. Darüber hinaus erhöhte eine Sequenzoptimierung die Reprogrammierungseffizienz mehrerer TF *in vitro*, förderte biophysikalische Merkmale von gebildeten Kondensaten *in vitro* und führte zu einer weniger spezifischen DNA-Bindung in Zellen. Zusammen mit jüngsten Arbeiten zur Funktion von Enhancer-Elementen, deuten diese Ergebnisse auf eine wichtige evolutionäre Rolle suboptimaler Merkmale in der Transkriptionskontrolle hin. Meine Arbeit lässt darauf schließen, dass eine rationale Modifikation von Aminosäuremerkmalen, welche die Kondensation von Proteinen beeinflusst, eine Strategie zur Optimierung TF-abhängiger Prozesse, einschließlich der zellulären Reprogrammierung ist.

Introduction

Transcriptional dynamics and regulatory mechanisms in eukaryotic cells

Transcription is the process of converting information encoded in DNA to RNA molecules. In eukaryotic cells, RNA polymerase II binds to promoter regions of genes to synthesize a complementary RNA molecule to the template DNA strand [1,2]. The resulting RNA molecule, known as messenger RNA (mRNA), carries the genetic information from the DNA in the nucleus to the cytoplasm, where it serves as a template for protein synthesis during translation.

Every cell of an organism contains an identical set of genetic information, yet not all cells express all genes or the same set of genes at the same time. Transcriptional regulation, is the mechanism by which genes are selectively activated or repressed. Transcriptional regulation enables spatio-temporal selectivity of genes, particularly important during embryonic development as it directs differentiation of diverse cell types from a single fertilized oocyte and remains essential during maturation and maintenance of cells ensuring the formation of functional tissue [3,4].

Over time, models of transcriptional regulation have evolved, reflecting the contemporary understanding and technological advances of each era. Early insights on gene regulation were made in prokaryotes, where transcriptional regulation is mediated by transcription factors (TF) binding bacterial promoters, cis-regulatory elements (CREs) on DNA [5]. Bacterial promoters are typically composed of three core regions essential for the initiation of gene expression [5,6]: the “-35 region” and “-10 region” located approximately 35 and 10 base pairs upstream of the transcription start site (TSS), respectively. Bacterial promoters encode consensus sequences that facilitate efficient recognition and accessibility of the promoter by the sigma factor, a central subunit of the bacterial RNA polymerase [7]. In addition, upstream elements (UE), located further upstream of the TSS, have been reported to modulate promoter activity by interactions with transcription factors and the alpha subunit of RNA polymerase in *Escherichia coli* [8,9]. This stoichiometric model of transcriptional regulation rapidly gained popularity since many findings and mechanisms discovered in bacteria were found to be transferable to higher model organisms such as *Caenorhabditis elegans*, *Drosophila melanogaster*, and eventually mammals like *Mus musculus* and *Homo sapiens*. Even though transcriptional regulation in these organisms is much more complex and multilayered, the basic principles of CRE recognition by TFs and subsequent co-factor interaction still apply [10,11].

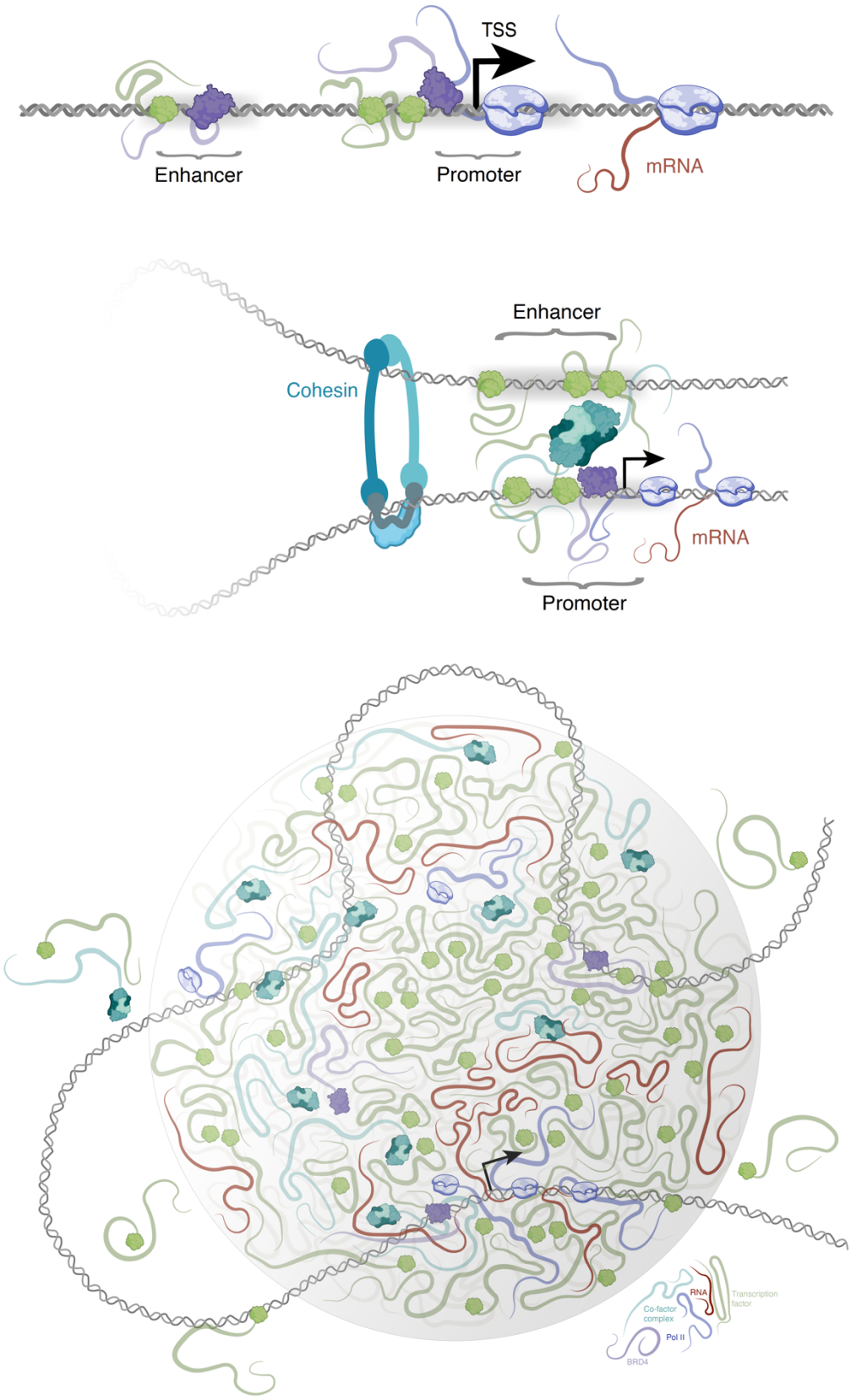


Figure 1: Schematic visualization of models of transcriptional regulation. *(top) Stoichiometric model of transcriptional regulation in eukaryotic cells. (middle) DNA-loop extrusion model. (bottom) Transcriptional condensate model. TSS, transcription start site.*

In eukaryotic cells, transcription is controlled by regulatory proteins such as TFs binding to promoter and enhancer regions to induce gene expression through recruitment of transcriptional machinery including RNA Polymerase II and co-factors (Figure 1, top) [12]. Eukaryotic core promoters are typically GC-rich DNA regions located ± 50 base pairs around the TSS and encode specific binding motifs for TFs. Furthermore, the structure of a promoter ensures proper assembly and loading of the RNA polymerase II preinitiation complex, composed of general TFs and RNA polymerase II [13–15]. Beyond the recruitment of activator or repressor TFs, the discovery of epigenetic marks on histone tails - disordered regions of nucleosome core components - has added complexity to eukaryotic gene control [16,17]. Chromatin remodelers, such as histone acetyltransferases (HATs), histone deacetylases (HDACs), histone methyltransferases (HMTs), and histone demethylases (HDMs), mediate the deposition and removal of these marks in a cell type dependent context. And specific histone modifications are associated with certain transcriptional states of a locus. For instance, methylation of histone H3 lysine 4 (H3K4me) generally indicates active transcription and is found at promoters of actively transcribed genes [18]. Histone acetylation, such as acetylation of histone H3 lysine 27 (H3K27ac), is often observed at active enhancers [19]. The addition of an acetyl group neutralizes the positive charge on histones, weakening their interaction with DNA, and thus, loosens the chromatin structure for enhanced accessibility by TFs and other regulatory proteins [16,17]. H3K27ac also aids in recruiting transcriptional co-activators like the bromodomain-containing protein 4 (BRD4), leading to increased gene expression [20].

BRD4 serves as a good example of a transcriptional co-activator. BRD4 interacts with acetylated lysines on histone tails through its bromodomain and participates in the regulation of its target genes by bridging the modified chromatin to transcriptional machinery [21]. BRD4 binds to acetylated histone H4 in enhancers and promoters of active genes [22]. It serves as a scaffold for the assembly of the positive transcription elongation factor (P-TEFb) that phosphorylates the C-terminal domain (CTD) of RNA polymerase II, thereby enhancing its processivity and promoting transcriptional elongation [23].

The eukaryotic adaptation of the stoichiometric model of transcriptional regulation prevailed for some time as it successfully explained mechanisms of how sequential DNA-protein and protein-protein interactions can induce transcription at a target gene. However, the model fell short in explaining how enhancers, specialized DNA sequences that, when bound by TFs and co-activators like BRD4, interact with promoters, given the potential for kilobases of intervening DNA eventually enhancing gene expression of their associated genes [24]. Pioneering work has revealed that the orientation or positioning of an enhancer relative to its promoter is often independent from its regulatory function [25]. The partial independence of

enhancers from their positioning along the DNA can be explained through the three-dimensional architecture of the eukaryotic nucleus. When considering 3D chromatin conformation, it becomes evident that an enhancer, though distant from its associate promoter considering DNA sequence, may in fact be in close proximity within the spatial organization of the nucleus [26–29].

The chromatin loop extrusion model serves as a framework for understanding the spatial organization of chromatin and its implication for gene regulation (Figure 1, middle). 3D chromatin structure is created by interactions of the ring-shaped cohesin protein complex with DNA-binding factors like CTCF and YY1 [30,31]. Cohesin, upon binding to DNA, is proposed to extrude the DNA strand through its ring-like structure in an ATP-dependent process [32,33]. This process is typically halted by boundary elements, such as CTCF or YY1, which are sequence-specific DNA-binding proteins that serve as architectural organizers of the genome [26]. The resulting loop can bring enhancers into close proximity with promoters facilitating the physical interaction necessary for the regulation of gene expression [28,34]. By organizing the three-dimensional folding of chromatin, loop extrusion establishes a higher-order genomic structure that defines the spatial and temporal context of transcriptional activity. This spatial organization is critical, as it determines which enhancers interact with which promoters, thereby influencing the precise patterns of gene expression that drive cellular identity and differentiation [35]. The interaction of enhancer and promoter sequences is facilitated by the Mediator complex contributing to transcriptional initiation by recruitment of RNA polymerase II [36]. Therefore, the Mediator complex helps to coordinate the transcriptional machinery required for controlled gene expression [37].

The models discussed here account for many foundational aspects of gene expression in both prokaryotic and eukaryotic cells, providing explanations for a broad range of observed phenomena. However, some observations have highlighted limitations in these models, including transcriptional bursting, molecular crowding, dynamics in transcriptional response and to some extend specificity in transcriptional regulation [38,45]. Consequently, new ideas have emerged to address these shortcomings.

The condensate model for transcriptional control is a relatively recent addition to the array of models for eukaryotic gene regulation, focusing on the role of phase-separated nuclear bodies, often referred to as transcriptional condensates (Figure 1, bottom) [38]. Transcriptional condensates are proteinaceous liquid-like bodies, or high-density assemblies of typically 50 - 100 nm of size that associate with *loci* of active transcription in cells [39–41]. They consist out of transcriptionally associated proteins like TFs, transcriptional co-activators like the Mediator

complex, RNA polymerase II and its associated factors but also RNA and DNA have shown to play important roles in the formation and maintenance of these membrane-less organelles [39,42–46]. The formation of transcriptional condensates creates environments with high concentrations of TFs, transcriptional co-activators and RNA polymerase II leading to probabilistically higher interaction efficiency [39,42]. Furthermore, it has been shown that transcriptional condensates dynamically assemble and disassemble, enabling spatio-temporal control over gene expression [45,47].

As our understanding of transcriptional regulation advances, it becomes evident that dynamic and spatial aspects of transcriptional regulation play crucial roles. This leads us to explore the field of biomolecular condensates, which provides valuable insights into the dynamics and specificity of biomolecular interactions.

Biomolecular condensates in transcriptional regulation

Liquid-liquid phase separation (LLPS) is a physicochemical process whereby a system transitions from a mixed to a de-mixed state, resulting in the segregation of distinct phases [48]. This phenomenon is observable in the context of solvent interactions, where intrinsic chemical properties dictate phase compatibility. A textbook example of LLPS is the behavior of water and oil. Water, being hydrophilic, favors interactions with polar and charged entities through formation of strong hydrogen bonds. Oil molecules, which are nonpolar and engage primarily in van der Waals and other hydrophobic interactions, cannot form such bonds and are excluded in the presence of the strong cohesive forces of water. In any given context, these properties do not change; water remains polar, oil remains nonpolar, and they do not mix. The system naturally favors a state of lower free energy, which includes minimizing the unfavorable interactions between water and oil [49].

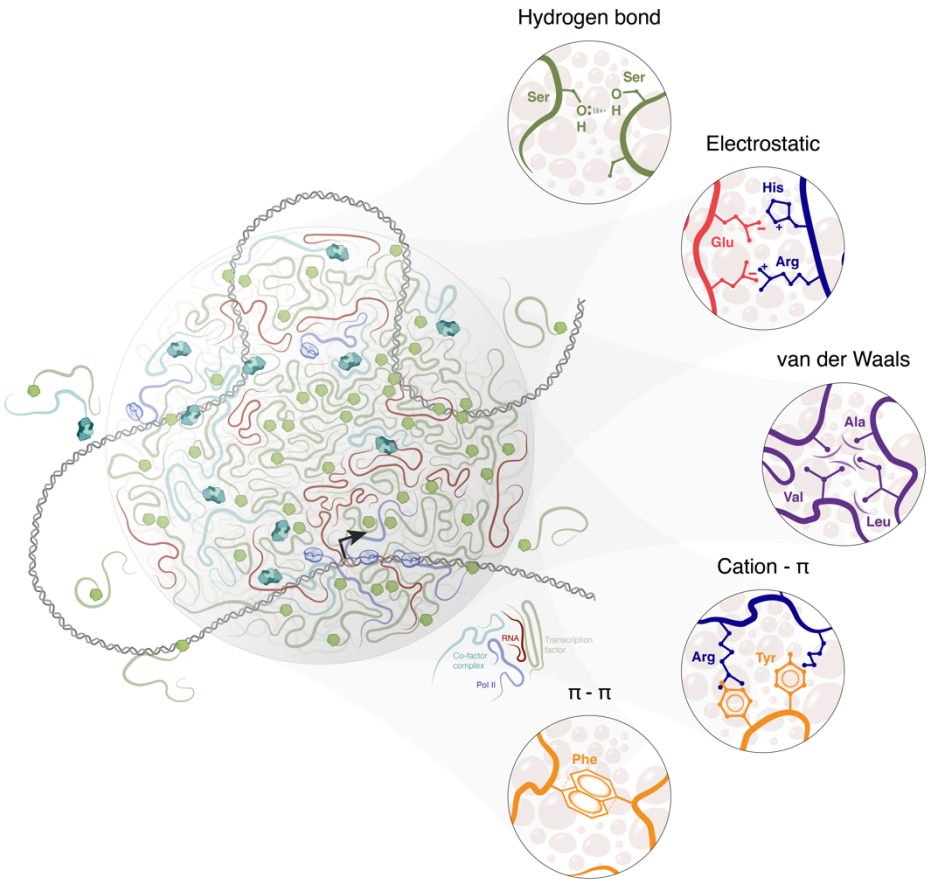


Figure 2: Weak multivalent interactions facilitate biomolecular condensate formation. *(left) Schematic of a transcriptional condensate at an actively transcribed gene. (right) Different types of weak multivalent interactions that can contribute to condensation.*

In cellular biology, the formation of biomolecular condensates is a much more complex process compared to our textbook example, as it is driven by the intrinsic properties of various biomolecules that interact together. Biomolecular condensates form in live cells through LLPS or coacervation, where macromolecules such as proteins and nucleic acids segregate into distinct phases within the crowded cellular environment [48,50]. Unlike simple solvents such as water and oil, whose phase behavior is governed by repulsive forces due to their polar and nonpolar natures, biomolecular condensates arise from attractive interactions between macromolecules (Figure 2). These attractive interactions exceed the critical energy needed to overcome the entropy of solvation. This interplay of forces includes hydrophobic interactions, electrostatic charges, hydrogen bonding and van der Waals forces among the molecules involved [51].

The process of LLPS in cells is influenced by various factors, including temperature, pressure, pH and concentration. Biomolecules remain miscible in aqueous solution until reaching their solubility limit, the critical saturation concentration (c_{sat}) – the threshold at which phase transitions occur due to macromolecular interactions becoming energetically more favorable than interactions with water. Consequently, two immiscible liquid phases form: a highly concentrated dense phase (c_D) and a dilute equilibrium phase (c_L) (Figure 3) [51].

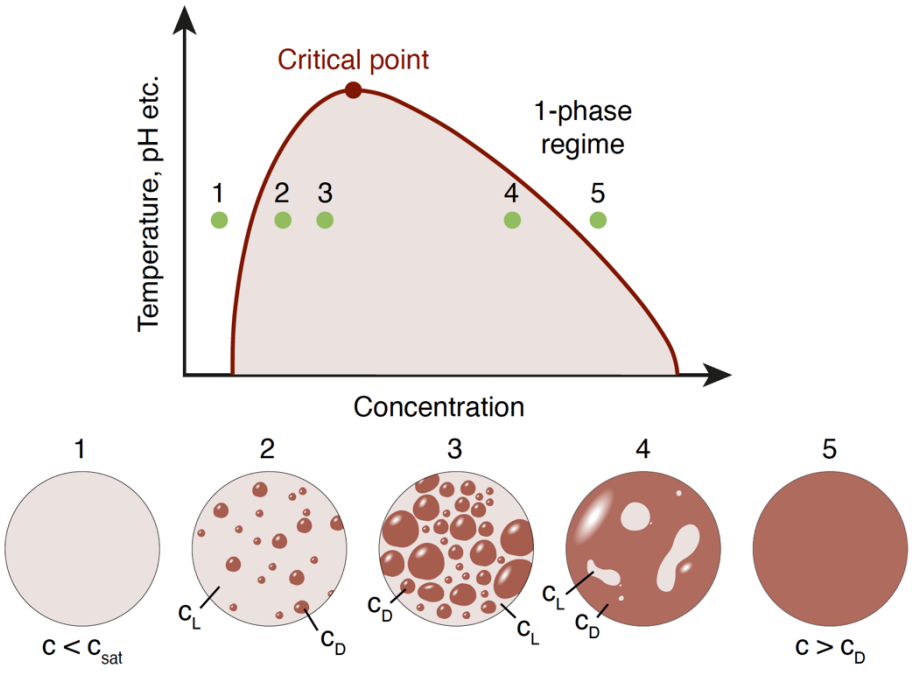


Figure 3: Schematic representation of a phase diagram. *In equilibrium, phase transitions are the consequence of changes in environmental conditions such as pH, temperature or protein concentration. c_D, concentrated dense phase, c_L, dilute equilibrium phase, c, concentration, c_{sat}, critical saturation concentration. Adapted from Alberti et al.*

Condensation is facilitated by multivalent interactions among macromolecules, which can be established through structured protein domains as well as through weak multivalent interactions among amino acid side chains within intrinsically disordered regions (IDRs) of proteins involved [52,53]. Under physiological conditions, IDRs lack stable secondary structures and often contain regions of low sequence complexity that tend to interact with low complexity regions of the same type [54,55]. The widely acknowledged “stickers and spacers” hypothesis attempts to explain how unique amino acid residues contribute to multivalent interactions between IDRs. In this model, aromatic amino acid residues act as “stickers”, mediating inter- and intramolecular interactions, while the surrounding non-hydrophobic amino acid residues serve as “spacers” to prevent hydrophobic collapse [56].

In a cell, biomolecular condensates create membrane-less compartments that concentrate essential factors for particular biomolecular processes [51]. Among the first described biomolecular condensates is the nucleolus [57,58]. Assembled around arrays of ribosomal DNA, this multi-phase system of different proteins and RNA components serves as an assembly line for ribosome subunit biogenesis [59]. The nucleolus features three distinct phase-separated layers, each maintained by specialized “scaffolding” proteins: nucleophosmin (NPM1) for the granular component, fibrillarin (FIB1) for the dense fibrillar component and treacle ribosome biogenesis factor 1 (TCOF1) for the fibrillar center [60,61]. These scaffolding proteins are essential for the integrity of their respective phase-separated compartments, providing a framework for their assembly. Conversely, most proteins in each nucleolar layer are not essential for their formation. These proteins are known as “clients”, which engage with scaffolding proteins via weak, multivalent or specific structural interactions [62].

Beyond the nucleolus, a large number of different biomolecular condensates has been identified within both, the cytoplasm and the nucleus. These condensates typically concentrate proteins that execute specialized molecular functions. For instance, heterochromatin condensates are associated with gene silencing [63], while splicing speckles are involved in mRNA processing [64] and the phase-separated nuclear pore complex is essential for molecule transport [65]. Additionally, condensates play critical roles in RNA storage [66], chromatin remodeling [67], and transcriptional regulation [39], each compartmentalizing distinct processes to enhance cellular efficiency.

Recent work highlighted the role of condensate formation in the regulation of gene expression. Observations of discrete nuclear clusters, approximately 50 - 100 nm in size, containing RNA polymerase II have led to the discovery of liquid-like assemblies at sites of active transcription [47]. The formation of transcriptional condensates is nucleated with the binding of TFs to DNA

motifs within accessible promoters and enhancers. This is followed by recruitment of significant quantities of co-activator molecules like BRD4, Mediator or β-catenin and RNA polymerase II through processes like LLPS or coacervation [39,46]. Together, these proteins are assumed to act as scaffolds for the formation and maintenance of a liquid-like condensate, prompting a rapid and strong gene expression response to various stimuli. Transcriptional condensates are highly dynamic with a half-life ranging from minutes to hours [68]. Transcription of enhancer RNAs, which are short non-coding transcripts from condensate-associated enhancers, is known to regulate condensate dynamics. In a feedback loop, low concentrations of highly charged RNA molecules promote condensation by increasing valency of the environment and thus serving a scaffolding role, while high concentrations promote condensate dissolution due to electrostatic repulsion [45]. Therefore, condensate dynamics are thought to underlie the phenomenon of transcriptional bursting observed in eukaryotic gene expression [68]. Rapid kinetics are essential for the functionality of transcriptional condensates. While some biomolecular condensates, like P-bodies, exhibit more gel-like properties manifesting as higher viscosity, the liquid-like features of transcriptional condensates seem to be crucial for their function [40]. The resulting rapid kinetics ensure efficient turnover of transcriptional machinery during the active process of transcription and are likely regulated by sequence features within IDRs of client proteins, such as the amount or the distribution of charged or hydrophobic amino acid residues, that interact with scaffold proteins [44,69,70].

In combination with stoichiometric models of biomolecular processes, the concept of biomolecular condensate formation provides a more complete picture of spatio-temporal and dynamic processes involved in the regulation of cellular function. It explains non-stochiometric enrichment of TFs observed at *loci* of active transcription and offers a conceptual foundation for further research into the outstanding question of how approximately 1,500 human TFs, with high sequence variability, interact specifically and in a controlled manner with a single holoenzyme of RNA polymerase II, despite lacking a unifying domain for direct interaction [71].

The importance of transcriptional condensates for transcription factor function

Transcription factors are the central elements of transcriptional regulation across cell types. TFs are structurally modular and typically composed of domains that control DNA-binding, transcriptional activation or repression, dimerization and ligand interaction [72]. Due to their modular architecture, they are extremely versatile molecules, required to selectively bind specific DNA-binding motifs within enhancer and promoter regions and subsequently activate or repress target gene expression [70].

TFs contain a DNA-binding domain that selectively and efficiently recognizes and binds specific DNA motifs, thus determining the target genes of the respective factor [73]. DBDs are structured domains that present a protruding surface in order to contact a DBD-specific motif of DNA base pairs with high affinity. They primarily target the major groove of the DNA double helix, establishing contacts through direct and water-mediated hydrogen bonds, as well as non-polar van der Waals interactions [74]. The domain structure of DBDs is highly conserved across species, and human TFs are classified into families based on the structure of their DBD (Figure 4) [74–76]. The major TF families in eukaryotic cells are C2H2-zinc finger (ZF) including Krüppel associated box (KRAB) domain containing proteins, Homeodomain, basic helix-loop-helix (bHLH), basic leucine zipper (bZIP) and nuclear hormone receptors (NHR) [73]. With to date 705 out of 1455 identified human TFs belonging to the C2H2-ZF family, it is the biggest group of transcription factors [40]. KRAB-domain containing C2H2-ZF typically exert a repressive role on the bound locus by recruitment of the TRIM28 complex, leading to chromatin remodeling that silences gene expression [77,78]. In contrast, other TF families are known to contain potent activators of gene expression, such as the Homeodomain or the bZIP family [78,79]. Some other families like the bHLH family have the unique capability to exert dual functions on gene expression by forming homo- and heterodimers before binding to DNA, with dimerization being necessary for stable DNA interaction. This dual capacity to activate and repress in the same cellular context can be attributed to different dimerization outcomes, where one configuration may be activating while a dimer containing a different partner can be repressive [80,81].

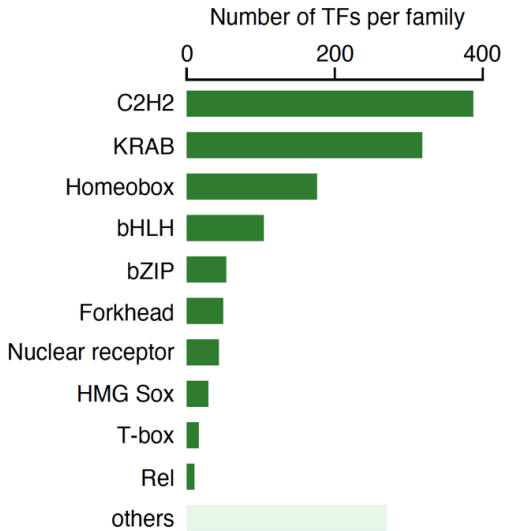


Figure 4: Overview of human transcription factor families and number of members.

Most TFs bind to accessible nucleosome-depleted DNA at active enhancers and promoters. However, a specialized subset known as pioneering TFs can

bind to nucleosome arrays *in vitro* and penetrate condensed heterochromatic sites in cells, inducing gene expression from previously repressed *loci* [82,83]. TFs with pioneering ability are crucial especially during early embryonic development as they regulate cell identity genes essential for cellular differentiation and cell type maintenance [84,85].

Despite major advances in understanding TF specificity by studying DBDs and TF-binding motifs, the question of how TF specificity is achieved still prevails. By today, several sequence-dependent and independent factors have been shown to contribute to TF-binding specificity, such as the affinity of a DBD to its cognate DNA-binding motif, the TF concentration in the cell, co-factor availability, and the chromatin context at the targeted *locus* [86]. But how do TFs selectively bind motifs within target CREs, while the overall number of accessible TF-binding motifs far exceeds the number of target *loci* [87]? And how do DBDs with highly similar high-affinity binding motifs bind very different low-affinity binding motifs [88]? Recently, research by Naama Barkai and her team has revealed that not only the DBDs of yeast TFs but also sequence features within their IDRs contribute to DNA-binding specificity and subsequent activation of target gene expression [89,90]. Therefore, we have to change our understanding of TF modularity and start considering the entire protein as a functional unit that influences binding specificity instead of focusing only at the DBD.

Early research by Steve McKnight and his team demonstrated that the capacity of a TF to activate target gene expression is typically not encoded within the structured DBD. Instead, this function resides in a distinct, non-structured region of the protein. When a TF binds to CREs via its DBD, gene expression is facilitated by its activation domain (AD) [79]. ADs are short, often disordered linear sequence motifs within TFs that adopt a specific secondary structure when in close proximity to a transcriptional co-activator. Such reversible interactions facilitate the transcriptional activation of the locus through subsequent recruitment of transcriptional machinery [91–93]. This function has made particularly strong activation domains like the herpes simplex virus VP16 AD to versatile tools in many functional studies [94].

The computational prediction of ADs in TF sequences is challenging since most ADs are located withing lager IDRs [95]. Thus, structure cannot serve as a reference, and instead, sequence composition along with non-linear sequence features must be considered as information carriers. Several studies have predicted ADs in human TFs [95–97]. Experimental evidence from yeast transcription factor screens has shown that especially sequences enriched in acidic and hydrophobic residues are critical for interactions with the Mediator complex, with acidic residues preventing hydrophobic motifs from collapsing, thereby exposing them most efficiently to the solvent [95,96]. Nonetheless, the predicted domains fail to

recapitulate the full spectrum of human TF function. A novel approach of genome-wide screens for ADs has been enabled by reduced costs of DNA fragment synthesis and next-generation sequencing. In such screens, TF sequences are tiled into 20 to 40 amino acid segments, fused to a DNA-binding domain, and analyzed for activity in high-throughput reporter assays. AD screens have led the labs of Mikko Taipale and Lacramioara Bintu to identify many formerly uncharacterized activation and repression domains [98–100]. Yet, the number of experimentally validated ADs is small, and the presence of an AD does not always correlate with the activity of the full-length IDR sequence.

The sole presence of DNA-binding motifs withing human CREs does not explain the specificity of TFs to their targets. Furthermore, genome wide screens for minimal ADs in human TFs fail to explain the entire scope of transcriptional activation. Besides harboring minimal ADs, IDRs play an important role in formation and function of transcriptional condensates. And the perturbation of transcriptional condensate formation has been shown to affect gene expression at the affected *locus* [39]. Therefore, I propose that the ability of a TF to transactivate specific target genes is inherently linked to its ability to condense and to partition into transcriptional condensates.

Transcription factors drive cell fate determination

Transcription factors are known for their tissue-specific expression. Out of the approximately 1500 mammalian TFs, around 300 are expressed in any given adult tissue at a time. About two-thirds of these are considered housekeeping TFs, which are ubiquitously expressed across most cell types. The rest are tissue-specific TFs, crucial for the maintenance of cell type characteristics [101]. Alongside housekeeping and tissue-specific TFs, developmental TFs are expressed exclusively during embryonic development in a tightly regulated spatio-temporal manner and play key roles in lineage specification [102] .

During embryonic development, a single-cell zygote, a fertilized oocyte, undergoes several rounds of cleavage to form a multicellular embryo. Its descendant cells will differentiate into variety of cell types, each with unique functions, ultimately forming a functional organism. Cell type-specific developmental TFs are already active in the inner cell mass of the blastocyst during the pre-implantation phase [103]. In these pluripotent embryonic stem cells (ESCs), the transcription factors Oct4, Sox2 and Nanog (OSN) are central to a transcriptional network that stabilizes the pluripotent state of ESCs. They promote self-renewal by activating genes crucial for stem cell maintenance and repressing genes inducing differentiation [104]. OSN show highly cooperative binding to similar targets and act synergistically to sustain their own expression, thereby forming a positive feedback loop [105,106]. The absence of any of these key TFs disrupts the core transcriptional circuitry, de-stabilizes the pluripotent state leading to differentiation [107].

From gastrulation onward, morphogenetic signals guide pluripotent stem cells to form the three primary germ layers: endoderm, ectoderm and mesoderm. These cell types are the precursors for all body structures and organs [108]. The differentiation of these germ layers is induced by morphogens and driven by the expression of lineage defining “master transcription factors” such as GATA6 and SOX17 for the endoderm, E2A for neural ectoderm, and T and CDX2 for mesoderm [109–111]. Master TFs define cell lineages and sit atop the transcriptional regulatory hierarchy [112]. From these germ layers, specialized tissues and cell types like neurons, bone, or liver are formed, driven by specific master transcription factors like HNF4α for the liver, NGN2 for neurons and RUNX2 for bone, which direct the transcriptional programs essential for cell-type specification and maturation [113–115].

Master TFs have such profound influence on cell fate that they can reprogram a differentiated somatic cell into a different cell type (Figure 5). The forced expression of a master TF can initiate the activation of target genes transforming the transcriptome and epigenome towards those characteristics of the TF-associated cell lineage [116]. This method, known as “direct

reprogramming" (also referred to as somatic lineage conversion or trans-differentiation), was discovered using master TFs such as MYOD1, C/EBPα, and NGN2 in various cell types [31,117–119]. However, the variety of cell types that can be created by direct reprogramming protocols is confined by the limited number of master TFs and epigenetic barriers that are not easily crossed by simple overexpression of these factors [120].

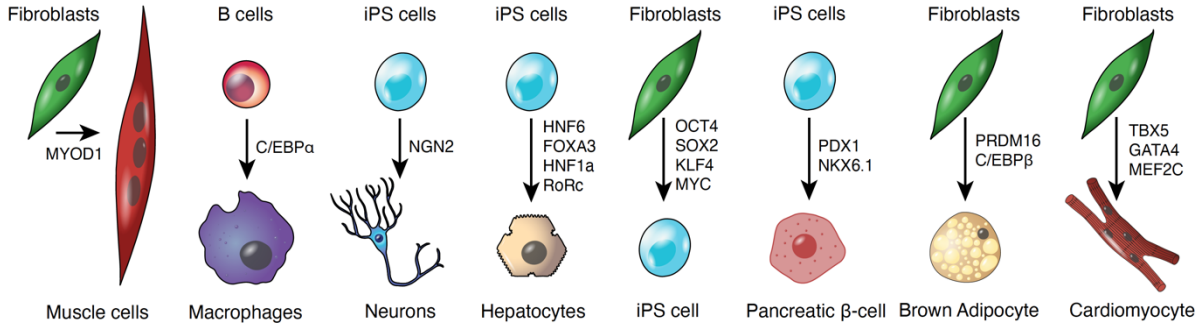


Figure 5: Transcription factors direct cell fate. *Schematic of master transcription factors used in direct reprogramming protocols. Adapted from Graf & Enver.*

In 2006, Takahashi and Yamanaka revolutionized the field of cellular reprogramming with their landmark study [121]. They identified a set of master TFs – Oct4, Sox2, Klf4 and c-myc – first in mice and then a year later in human [122]. These factors were shown to reprogram fibroblasts into pluripotent, stem cell-like cells when ectopically expressed. Such induced pluripotent stem cells (iPSCs) may now be used to potentially generate all cell types of the human body through either chemical stimulation or TF-mediated differentiation. To date, human iPSCs have been differentiated into more than 50 cell types *in vitro*, following protocols that include chemical stimulation, the addition of signaling factors, or TF overexpression [120]. Among these cell types are multipotent cells like hematopoietic or neural stem cells [123,124]; more terminally differentiated cells like osteoblasts [125], natural killer cells [126] and macrophages [127], as well as cells of significant medical importance, such as hepatocytes [128], various types of neurons [129,130], and insulin-producing pancreatic β-cells [131].

Despite these advances, most differentiation protocols currently in use fail to transition to an *in vivo* stage relevant for clinical applications such as cell replacement therapy, with some exceptions including cardiomyocyte, hepatocyte and pancreatic β-cell reprogramming [132–134]. A major factor contributing to this limitation is the generally low efficiency of reprogramming *in vitro*, as well as incomplete maturation states of the resulting cells. While cell type maturation *in vivo* often surpasses maturation *in vitro* due to the given signaling context in the living organism, the issue of low reprogramming efficiency remains [120]. Thus, there is a critical need for optimized protocols that can overcome this gap and transition more reprogramming protocols to clinically relevant stages.

Advances in the prediction of disordered protein regions

Recent work has highlighted the importance of intrinsically disordered regions in protein function, leading to the development of numerous tools for predicting protein disorder from amino acid sequences. Experimental validation of protein disorder has been challenging as it aims to detect the absence of a feature, in this case structure, making the conceptualization of IDRs elusive.
Different approaches are followed to experimentally determine whether a protein contains a disordered region, each with its own technical biases. Indirect evidence for protein disorder can be inferred from the absence of residues in results of X-ray crystallography experiments, though it should be noted that longer IDRs are often intentionally removed since they impede crystal formation [135]. In contrast, direct evidence of protein disorder can be obtained by assigning IDRs based on comparisons of residue-wise deviations of NMR structures with X-ray crystallography data [136]. Due to the low throughput and tedious nature of these experimental approaches, predictive tools have gained popularity for their ability to generate data rapidly and in high-throughput.

While all of these tools process the same amino acid input, they follow diverse methodologies to formulate their predictions. Understanding the features and parameters underlying each tool's predictions is essential to accurately interpret their outputs and the minor discrepancies that often arise in parallel comparisons. To date, more than 40 disorder predictors have been documented [135]. Some, such as DISOPRED [137], IUPred [138] and PONDR [139] can predict and depict disordered regions within a protein of interest in seconds. Each of these predictors uses a distinct approach, resulting in unique biases. While DISOPRED incorporates information on evolutionary conservation, IUPred relies on pairwise interaction energies for each residue in a protein sequence. PONDR uses a range of machine learning models, including neural networks and support vector machines.

Advancing the field, Alphafold2 released in 2021 took a significant leap in the development of prediction algorithms, addressing shortcomings in accuracy especially when little information is available on the evolutionary history of proteins, their homology to solved structures, or pairwise evolutionary correlations (Figure 6) [140]. Alphafold2 applies a new machine learning based approach by integrating physical and biological information together with multiple sequence alignments into its design of its deep learning models to predict the three-dimensional structure a protein will adapt solely based on its amino acid sequence [140]. For the first time, it has been possible to obtain high-confidence predictions that include disordered regions within a 3D protein model, providing context for small structured motifs embedded

within disordered sequence stretches. Additionally, the computation of the pLDDT (predicted Local Distance Difference Test) score has facilitated the mapping of disordered regions with high accuracy.

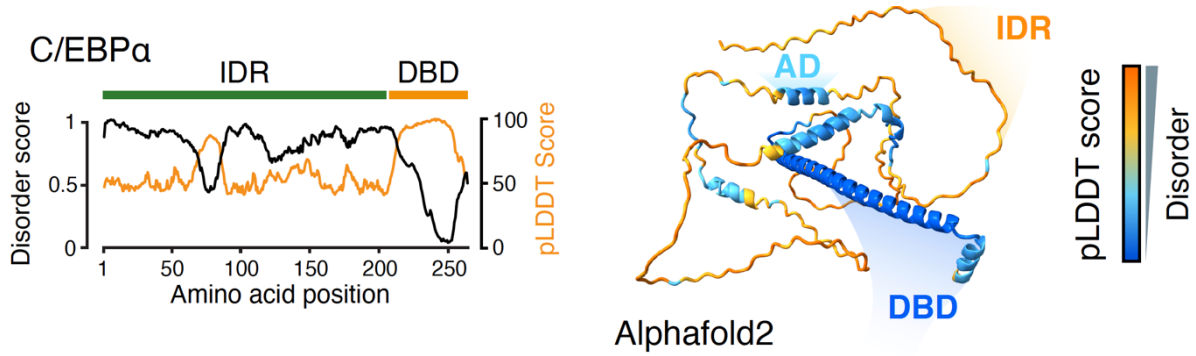


Figure 6: Prediction of disordered protein regions. *(left) Schematic representation of protein disorder along the amino acid sequence of C/EBPa using a disorder score (Metapredict v2, black) and the pLDDT score (Alphafold2, orange). IDR, intrinsically disordered region, DBD, DNA-binding domain. (right) Alphafold2 model of C/EBPa. Structure is colored according to pLDDT score. AD, activation domain.*

Given the diversity of methodologies used in the field, the preferred predictor in our study is Metapredict v2 [141], which adopts a meta-analytic strategy by integrating results from various tools like Alphafold2s pLDDT score and other common disorder predictors to create a conservative and comprehensive estimate.

Aims of this study

The sole presence of DNA-binding motifs within human CREs does not explain the specificity of TFs to their targets. Furthermore, predictive models and genome wide screens for minimal activation domains in human TFs fail to explain the entire scope of transcriptional activation. Employing the model of condensate formation for transcriptional regulation, my aim is to identify and characterize non-linear sequence features encoded within TF IDRs that mediate a potential relationship between specificity and activity of human TFs. Building on prior research, I focus on a sequence feature previously discovered in prion-like domains (PLDs) of RNA-binding proteins – the dispersed patterning of aromatic amino acids [69]. Periodically arranged aromatic amino acids in PLD-containing proteins have been shown to promote liquid-liquid phase separation *in vitro* and the dispersion of these aromatic amino acids in PLDs has been noted to affect the biophysical properties of the resultant condensates. Moreover, it has been demonstrated that PLDs, such as the N-terminal IDR of the human FUS protein, can confer transcriptional activity in reporter assays. Given that both transactivation and condensate formation are critical aspects of TF functionality, I set out to determine if dispersion of aromatic amino acids is a feature present in IDRs of human TFs and, if so, to discern how aromatic dispersion contributes to TF function. I also intend to test whether this sequence feature is the hitherto missing link that inherently connects transactivation with the condensation ability of human TFs.

First, I will apply a proteome-wide approach to detect and quantify the dispersion of aromatic residues. Additionally, I will deploy an IDR mutagenesis screen to assess the significance of this sequence feature on the transactivation strength of various human TF IDRs in luciferase reporter assays. In parallel, I will examine the effect of altered aromatic dispersion on the homotypic condensation propensity of these TF IDRs.

Second, informed by my findings, I plan to understand the influence of optimized aromatic dispersion in TF IDRs on gene regulation within live cells by creating endogenous knock-in lines of a candidate TF mutant followed by target gene expression analysis.

Finally, I want to examine the influence of optimized aromatic dispersion on TF function in dynamic reprogramming systems. Following the hypothesis that optimized aromatic dispersion increases transcriptional activity, I aim to test if sequence optimization of reprogramming TF IDRs can enhance reprogramming efficiencies in established protocols and therefore be used as an optimization strategy for *in vivo* applications such as cell replacement therapy.

Materials and Methods

This section has been adapted from [142].

Ethics statement
The research in this study complied with all relevant ethical regulations, and was approved by the Max Planck Institute for Molecular Genetics.

Cell culture
HAP1 cells were a kind gift from the Aktas Lab (MPIMG). HAP1 cells were cultured in IMDM (Gibco) supplemented with 10% fetal bovine serum (FBS) (Gibco) and 1% penicillin/streptomycin (Gibco). Cells were split at 80-90% confluence. Medium was changed every day.

HEK293T cells (ATCC: CRL-3216) were cultured in knockout DMEM (Gibco) containing 15% FBS, supplemented with 1X GlutaMAX supplement (Gibco), 1X non-essential amino acids (Gibco), 1% penicillin/streptomycin and 0.05mM ß-mercaptoethanol (Gibco). Cells were split at 80-90% confluence. Medium was changed every 2-3 days.

V6.5 mouse embryonic stem cells (mESCs) were a kind gift from the Hochedlinger Lab (HSCI). Cells were cultured on irradiated primary Mouse Embryonic Fibroblasts (MEFs) in knockout DMEM containing 15% FBS, supplemented with 1X GlutaMAX supplement, 1X non-essential amino acids, 1% penicillin/streptomycin, 0.05mM ß-mercaptoethanol and 1000 U/ml leukemia inhibitory factor (LIF). Medium was changed every day.

ZIP13K2 human induced pluripotent stem cells (iPSCs), derived from fetal dermal fibroblasts, were a kind gift from the Müller Lab (MPIMG). Cells were cultured on Matrigel (Corning) pre-coated culture plates in mTeSR+ (STEMCELL Technologies) supplemented with 1% penicillin/streptomycin. Cells were split at 75-80% confluence. Medium was changed every day.

Kelly cells (DSMZ: ACC-355) were cultured in DMEM (Gibco) containing 10% FBS, supplemented with 1X GlutaMAX supplement and 1% penicillin/streptomycin. Cells were split at 80-90% confluence. Medium was changed every 2-3 days.

SH-SY5Y cells (DSMZ: ACC-209) were cultured in RPMI (Gibco) containing 10% FBS, supplemented with 1X GlutaMAX supplement and 1% penicillin/streptomycin. Cells were split at 80-90% confluence. Medium was changed every 2-3 days.

RCH-rtTA cells were derived from the RCH-ACV lymphoblastic leukemia cell line [143]. RCH-rtTA cells and derivates were cultured in RPMI (Gibco) containing 10% FBS, supplemented with 1% glutamine (Gibco),1% penicillin/streptomycin (Thermo) and 550 µM β-mercaptoethanol. Cells were maintained at a density of 0.1-6x10^6 cells/ml.

C2C12 mouse myoblasts were a kind gift from the Stricker Lab (FU Berlin). Cells were cultured in high glucose DMEM (Gibco) supplemented with 10% FBS and 1% penicillin/streptomycin. Cells were split at 70-75% confluence and medium was changed every 2 days.

U2OS cells were a kind gift from the Kinkley Lab (MPIMG). Cells were cultured in DMEM supplemented with 10% FBS and 1% penicillin/streptomycin. Cells were split at 80-90% confluence. Medium was changed every 2-3 days.

If not stated differently, all cells were cultured under standard conditions at 37°C and 5% CO_2. All reported cell lines were checked for mycoplasma contamination and tested negative.

Genomic DNA extraction

Genomic DNA of cultured cells was extracted by using the GeneJET Genomic DNA Purification Kit (Thermo Scientific) following the manufacturer's instructions. Concentrations of eluted DNA were measured using NanoDrop2000 (Thermo Scientific).

Generation of DNA constructs for protein purification

For the purification of mEGFP or mCherry labeled fusion proteins we amplified sequences from codon optimized gene fragments listed in section (Protein sequences and SLIMs) (Twist Bioscience) for HOXD4 wildtype, HOXD4 AroLITE A, HOXD4 AroLITE G, HOXD4 AroLITE S, HOXD4 AroPLUS, HOXD4 AroPLUS patched, HOXD4 AroPLUS LITE, HOXD4 AroPLUS patched LITE, HOXD4 AroPERFECT, HOXC4 wildtype, HOXC4 AroLITE S, HOXC4 AroPERFECT, HOXB1 wildtype, HOXB1 AroLITE A, C/EBPα wildtype, C/EBPα AroLITE A, C/EBPα AroPERFECT IS15, C/EBPα AroPERFECT IS10, NGN2 wildtype, NGN2 AroLITE A and NGN2 AroPERFECT C intrinsically disordered regions with primers. The amplified gene fragments were cloned into a pET45-mEGFP or pET45-mCherry backbone [41], linearized by restriction digest with AscI (NEB) and HindIII (NEB), via NEBuilder HiFi Assembly. All sequences of interest were cloned C-terminally to the fluorescence marker.

Protein purification

Overexpression of recombinant protein in BL21 (DE3) (NEB) was performed as described [40]. *E. coli* pellets were resuspended in 25 ml of ice-cold Buffer A (50 mM Tris pH 7.5, 500 mM

NaCl, 20 mM Imidazole) supplemented with cOmplete protease inhibitors (Sigma) and 0.1% Triton X-100 (Thermo, 851110) and sonicated for 10 cycles (15 s ON, 45 s OFF) on a Qsonica Q700 sonicator. Bacteria lysate was cleared by centrifugation at 15,500g for 30 minutes at 4°C. For protein purification the Äkta Avant 25 chromatography system was used. The entire cleared lysate was loaded onto a cOmplete His-Tag purification column (Merck) pre-equilibrated in Buffer A. The loaded column was washed with 15 column volumes (CV) of Buffer A. The fusion protein was eluted in 10 CV of Elution Buffer (50 mM Tris pH 7.5, 500 mM NaCl, 250 mM Imidazole) and diluted 1:1 in Storage Buffer (50 mM Tris pH 7.5, 125 mM NaCl, 1 mM DTT, 10% Glycerol). The fractions enriched for GFP were pooled after His-affinity purification and manually loaded through an injection valve connected to a 500 μl capillary tube onto an equilibrated Superdex 200 increase 10/300 GL column (Cytiva). The loaded column was equilibrated with 0.15 CV of ice-cold Buffer A supplemented with cOmplete protease inhibitors. Fusion proteins were eluted with 1.1 CV of ice-cold Buffer A supplemented with cOmplete protease inhibitors. Elution fractions were pooled. Eluates were further concentrated by centrifugation at 10,000g for 30 minutes at 4°C using 3000 MWCO Amicon Ultra centrifugal filters (Merck). The concentrated fraction was diluted 1:100 in Storage Buffer, re-concentrated, and stored at -80°C.

In vitro droplet assay

For *in vitro* droplet formation experiments, we measured the concentration of purified mEGFP-IDR fusion proteins with a NanoDrop2000 and subsequently diluted protein preparations to the required concentration with Storage Buffer (50 mM Tris pH 7.5, 125 mM NaCl, 1 mM DTT, 10% Glycerol). The *in vitro* droplet formation assay was performed as previously described [41]. Protein preparations were mixed 1:1 with 5 μl 20% PEG-8000 in de-ionized water (w/v) and equilibrated for 30 minutes at room temperature. The resulting 10 μl was pipetted on a chambered coverslip (Ibidi). Images were acquired after 3 minutes of equilibration on the slide, using an LSM880 confocal microscope equipped with a Plan-Apochromat-63x/1.40 oil DIC objective with a 2.5x zoom, resulting in a lateral pixel resolution of 0.04 μm. Quantification of condensate formation was based on at least 10 images acquired in at least two independent image series per condition.

Image analysis of *in vitro* droplet formation

Protein droplets were detected using ZEN blue 3.4 Image Analysis and Intellesis software packages. By use of a previously trained Intellesis model in spectral mode an image segmentation of individual pixels into objects (droplet area) or background (image background) was achieved. A minimum cutoff of 120 nm in diameter was applied on identified objects. Relative amounts of condensed protein were calculated by division of the sum of

mEGFP signal in objects defined as droplet area by the overall sum of mEGFP signal in the field of view. All values were calculated using R-Studio. Plots were generated using GraphPad PRISM9. To fit data to a sigmoidal curve we applied the in-build non-linear regression function (Sigmoidal; x is concentration).

Fluorescence recovery after photobleaching (FRAP)

In vitro droplets for FRAP experiments were formed as described above without 30 minutes of pre-assembly at room temperature at a protein concentration of 25 μM. Droplets were bleached immediately after pipetting the protein mixture onto the slide by using 488 nm light at 70% laser power in 10 iterations. Bleaching was performed on a central region of a settled single droplet. Fluorescence recovery was measured over a time course of 60 seconds in 2-second intervals. Quantification of FRAP data was based on at least ten images acquired in at least two independent image series per condition. The resulting signal recovery was normalized to background and fitted to a power law model in Microsoft Excel. All figures were generated using GraphPad PRISM9.

Generation of DNA constructs for transactivation assays

To study transactivation strength of transcription factor IDRs we amplified sequences from codon optimized gene fragments listed in section (Protein sequences and SLIMs) (Twist Bioscience) for HOXD4 wildtype, HOXD4 AroLITE A, HOXD4 AroLITE G, HOXD4 AroLITE S, HOXD4 AroPERFECT, HOXD4 AroPERFECT-1, HOXD4 AroPERFECT-2, HOXD4 wildtype YPWM(-), HOXD4 AroPERFECT YPWM(+), HOXD4 wildtype (N), HOXD4 WT(N)-FUSNxs, HOXC4 wildtype, HOXC4 AroLITE S, HOXB1 wildtype, HOXB1 AroLITE A, NANOG wildtype, NANOG AroLITE A, EGR1 wildtype, EGR1 AroLITE A, EGR1 AroSCRAMBLED, EGR1 AroPATCHY3, EGR1 AroPATCHY1, NFAT5 wildtype, NFAT5 AroLITE A, C/EBPα wildtype, C/EBPα AroLITE A, C/EBPα AroPERFECT IS15, C/EBPα AroPERFECT IS15+1, C/EBPα AroPERFECT IS15+2, C/EBPα AroPERFECT IS10, C/EBPα wildtype (N), C/EBPα wildtype (N)-IS15, C/EBPα wildtype (N)-FUSN, C/EBPα wildtype (N)-FUSNxs, FUSN, FUSNxs, NGN2 wildtype, NGN2 AroLITE A, NGN2 AroPERFECT, MYOD1 wildtype C, MYOD1 AroPERFECT C, MYOD1 AroLITE C, OCT4 wildtype N, OCT4 wildtype C, OCT4 AroLITE N, OCT4 AroLITE C, OCT4 AroPERFECT N, OCT4 AroPERFECT C, PDX1 wildtype, PDX1 AroLITE, PDX1 AroPERFECT, FOXA3 wildtype C, FOXA3 AroLITE C, FOXA3 AroPERFECT C, S6Y AroPATCHY1, S6Y AroPATCHY3, S6Y AroPERFECT, D6Y AroPERFECT intrinsically disordered regions with primers. Amplified gene fragments were cloned into a pGAL4 (Addgene #145245) backbone, linearized with AsiSI (NEB) and BsiWI (NEB) via NEBuilder HiFi Assembly.

Generation of DNA constructs for TF IDR tiling assays

To control for the potential creation of short linear motifs in TF IDR mutants, we tiled the HOXD4 wildtype, HOXD4 AroPERFECT, C/EBPα wildtype, C/EBPα AroPERFECT IS15, OCT4 wildtype C-, OCT4 AroPERFECT C-, MYOD1 wildtype C-, MYOD1 AroPERFECT C-, EGR1 wildtype and EGR1 AroSCRAMBLED IDRs into 40 amino acid segments with 20 amino acid overlaps. We amplified all 40 amino acid tiles in steps of 20 amino acids starting from the first amino acid of the sequence with primers. Amplified gene fragments were cloned into a pGAL4 (Addgene #145245) backbone, linearized with AsiSI (NEB) and BsiWI (NEB) via NEBuilder HiFi Assembly.

Transactivation assay

The transactivation activity of TF IDRs was assayed using the Dual-Glo Luciferase Assay system (Promega). V6.5 Mouse embryonic stem cells were seeded on gelatin pre-coated 24-well plates with a density of $1x10^5$ cells per cm^2. For feeder-free culture conditions, mESC medium was supplemented with 2x leukemia inhibitory factor (LIF). HEK-293T, SH-SY5Y, Kelly cells and C2C12 mouse myoblasts were seeded on 24-well plates with a density of $1x10^5$ cells per cm^2. After 24 hours, every well was transfected with 200 ng pGal4 empty vector control or the equimolar amount of the expression construct carrying an IDR of interest, 250 ng of the *Firefly* luciferase expression vector (Promega) and 15 ng of the *Renilla* luciferase expression vector (Promega) using FuGENE HD transfection reagent (Promega) following the manufacturer's instructions. After 24 hours, cells were washed once with PBS and lysed in 100 µl of 1x Passive Lysis Buffer (Promega) for 15 minutes on a shaker at room temperature. Subsequently, 10 µl of cell lysate was pipetted onto a white bottom 96-microwell plate in duplicates or triplicates followed by quantification of *Firefly* and *Renilla* using the Dual-Glo Luciferase Assay System Quick Protocol for 96-well plates (Promega). Triplicate data was normalized to *Renilla* luminescence of the respective well and finally normalized to the empty vector control. Data are shown as mean ± SD. All data shown were generated of three independent biological replicates. All data were plotted with GraphPad PRISM9. To assess statistical significance, two-sided unpaired t-tests were performed.

Generation of DNA constructs for locus re-construction assays

To confirm mutant-specific regulation of C/EBPα target promoters and enhancers, we amplified promoter and enhancer regions of GBP5, FAM98A and S100A with primers. Amplified fragments were cloned into a pGL3-Basic vector (Promega), linearized with BamHI (NEB) and SalI (NEB) in case of an enhancer region or with HindIII (NEB) and KpnI (NEB) in case of a promoter via NEBuilder HiFi Assembly. Full length C/EBPα wildtype and C/EBPα

AroPERFECT IS15 sequences for over-expression were cloned into a pGAL4 (Addgene #145245) backbone, linearized with EcoRI (NEB) and AsiSI (NEB) via NEBuilder HiFi Assembly.

Locus re-construction with pGL3 reporter assays

Transcription factor activity at genomic *loci* was assayed using the Dual-Glo Luciferase Assay system (Promega). V6.5 Mouse embryonic stem cells were seeded on gelatin pre-coated 24-well plates with a density of $1x10^5$ cells per cm^2. For feeder-free culture conditions, mESC medium was supplemented with 2x LIF. After 24 hours, every well was transfected with 200 ng of plasmid containing a C/EBPα wildtype or AroPERFECT IS15 overexpression cassette, 250 ng of pGL3-Basic control or an equimolar amount of the pGL3 construct carrying enhancer/promoter sequences of interest and 15 ng of the *Renilla* luciferase expression vector (Promega) using FuGENE HD transfection reagent (Promega) following the manufacturer's instructions. After 24 hours, cells were washed once with PBS and lysed in 100 µl of 1x Passive Lysis Buffer (Promega) for 15 minutes on an orbital shaker at room temperature. Subsequently, 10 µl of cell lysate was pipetted onto a white bottom 96-microwell plate in triplicates followed by quantification of *Firefly* and *Renilla* signal using the Dual-Glo Luciferase Assay System Quick Protocol for 96-well plates (Promega). Triplicate data was normalized to *Renilla* luminescence of the respective well, and normalized to the pGL3-Basic vector control. Data are shown as mean ± SD. All data shown were generated of three independent biological replicates. All data were plotted with GraphPad PRISM9. To assess statistical significance, two-sided unpaired t-tests were performed.

Western Blot

Cultured cells were washed twice in PBS and lysed in RIPA buffer (Thermo Scientific) for 30 min at 4ºC on an orbital shaker. Subsequently, the cell lysate was centrifuged for 20 min at 20,000g. The cleared lysate was transferred to a new tube and quantified by BCA assay (Thermo Scientific). 20 µg of extracted protein was run on a 4-12% NuPAGE SDS gel and transferred onto a PVDF membrane using an iBlot2 Dry Gel Transfer Device (Invitrogen) following manufacturer's instructions. To detect Gal4-fusion proteins, 50 µg of extracted protein was used. Membranes were blocked with 5% skim milk in TBST and incubated with primary antibodies over night at 4ºC. Primary antibodies used in this study include IFI16 (Santa Cruz Biotechnology, sc-8023, 1;200), GFP (Invitrogen, A11122, 1:2000), HSP90 (BD, 610419, 1:4000), ARHGAP4 (Santa Cruz Biotechnology, sc-376251, 1:200), ESX1 (Santa Cruz Biotechnology, sc-365740, 1:200), GAL4-DBD (Santa Cruz Biotechnology, sc-510, 1:200), GATA6 (RnD, AF1700, 1:1000) and FLAG (Merck, F1804, 1:2000). HRP-conjugated secondary antibodies Peroxidase-AffiniPure Donkey Anti-Goat IgG (JacksonImmuno,705-

035-147, 1:5000), Peroxidase IgG Fraction Monoclonal Mouse Anti-Rabbit IgG (JacksonImmuno ,211-032-171, 1:5000) and Peroxidase AffiniPure Goat Anti-Mouse IgG (JacksonImmuno, 115-035-174, 1:1000) were used against the host species and visualized with HRP substrate SuperSignal West Dura (Thermo Scientific).

LacO-LacI tethering assay

For LacO-LacI tethering experiments, we used a vector containing CFP-LacI followed by a multiple cloning site (MCS). MED1-IDR and POLR2-CTD plasmids were cloned via digestion with AsiSI (NEB) and BsiWI (NEB) via NEBuilder HiFi Assembly Master Mix. Tethering experiments were adapted from [40]. Imaging was performed on live cells 48 hours after transfection of 100 ng of CFP-LacI-HOXD4 wildtype, HOXD4 AroPERFECT, C/EBPα wildtype or C/EBPα AroPERFECT IS15 plasmid and 100 ng of MED1-IDR-YFP-NLS or POLR2-CTD-YFP-NLS into U2OS cells using FuGENE HD transfection reagent. Images were acquired using an LSM880 confocal microscope equipped with a Plan-Apochromat-63x/1.40 oil DIC objective with a 2x zoom. Laser intensities were adjusted prior to imaging to prevent possible channel bleed. Images were acquired across 2 biological replicates.

LacO-LacI tethering assay analysis

For LacO-LacI image analysis regions of interest corresponding to CFP-LacI-IDR fusion proteins were detected manually based on the cyan channel using ImageJ v 2.0.0. Mean intensities on these selected regions of interest were measured in both, YFP and CFP channels. Background intensity of the YFP channel was defined using a mean intensity measurement of a random nuclear region of same size and shape as the primary region of interest. Enrichment of YFP signal in regions of interest, predefined by the CFP signal, was calculated by dividing YFP mean signal intensity of the region of interest by the YFP mean signal intensity of the random nuclear region. Values were plotted as indicated using GraphPad PRISM9.

RNA isolation and quantitative Real-Time PCR (qRT-PCR)

RNA from cultured cells was extracted using the Direct-zol™ RNA MicroPrep Kit (Zymo Research) following the manufacturer's instructions. Subsequently, 1 µg of extracted RNA was used as input material for cDNA synthesis with the RevertAid First Strand cDNA Synthesis Kit (Thermo Scientific) using random hexamer primers following the manufacturer's instructions. Synthesized cDNA was diluted 1:10 with H_2O and stored at -20 °C. qPCR was performed using 2X PowerUP SYBR green master mix (Applied Biosystems) and primers.

Generation of HOXD4 GFP knock-in and knockout lines

For an endogenous knock-in of mEGFP-tagged HOXD4 variants, we cloned a synthesized, codon optimized sequence listed in section (Protein sequences and SLIMs) for HOXD4 wildtype, AroPERFECT or AroPLUS (Twist Bioscience) into a pUC19 backbone (Addgene #50005) linearized by restriction digest with BamHI (NEB) and HindIII (NEB). Besides the aforementioned HOXD4 coding sequences, the repair template contained N- and C-terminal homology regions for the HOXD4 genomic locus amplified from HAP1 genomic DNA, a synthesized GS-linker sequence (Sigma) and a mEGFP fluorescent protein sequence amplified from a pET45 plasmid. All plasmids were cloned by Gibson Assembly using the NEBuilder® HiFi DNA Assembly Kit (NEB).

The endogenous *HOXD4* locus was targeted by two guide RNAs (*Table 1*) cutting the N- or C-terminus of the HOXD4 coding sequence respectively.

Table 1: Guide RNA protospacer sequences

sgRNA	Gene	Protospacer sequence
hHOXD4_sgRNA_1	*HOXD4*	GCTGACGACCTTATAGAAGTG
hHOXD4_sgRNA_3	*HOXD4*	TGCAAATACTCCTCGCACGG

Both guide RNA sequences were cloned into the sgRNA-Cas9 vector px459 (Addgene #62988). Repair template and guide RNA vectors were co-transfected into HAP1 cells using Lipofectamine3000 (Thermo) following manufacturer instructions in a molar ratio of 5:1:1. To screen for functional integration, the transfected cells were sorted for mEGFP fluorescent protein expression by flow cytometry after 4 days, and a second time after 11 days. Positive cells were seeded as single cells on 96-multiwell plates. After expansion, clones were genotyped for the correct integration by PCR on extracted genomic DNA. Positive clones of every HOXD4-expressing line were selected on similar mEGFP expression levels. To generate a HOXD4 knockout cell line, HAP1 cells were transfected with both guide RNAs only. After 4 days, cells were seeded as single cells on 96-multiwell plates by flow cytometry and genotyped for HOXD4 deletion by PCR on extracted genomic DNA and quantitative real-time PCR on synthesized cDNA.

Imaging of HAP1 HOXD4 knock-in cells

For imaging of HOXD4 knock-in cells, $2x10^4$ cells were seeded onto chambered coverslips (Ibidi). After 24 hours, cells were washed with PBS and fixed with 4% paraformaldehyde for 15 minutes at room temperature. Cells were permeabilized with PBS supplemented with 0.1% Tween-20 (Sigma) for 5 minutes and PBS supplemented with 0.25% Tween-20 for 15 minutes.

Cells were then stained with primary antibody against GFP (Invitrogen, A11122, 1:500) and secondary goat anti-rabbit Alexa Fluor 594 antibody (Jackson Immuno, 2338059, 1:500). Nuclei were stained with 0.25µg/ml DAPI. Images were acquired with a Stellaris 8 confocal microscope and a Plan-Apochromat 100x/1.40 oil CS2 objective (Leica). For the analysis of sub-nuclear localization, a mosaic of at least 100 tile regions was imaged for each condition over two replicates. Object quantification was performed using ZEN 3.4 software (Zeiss). Briefly, DAPI counter stain was used to segment objects after Gaussian smoothing.

KAPA Stranded mRNA-seq of HAP1 HOXD4 knock-in cells

HAP1 cells were seeded with a density of $1x10^5$ cells per 6-well and cultured for 3 days until 80% confluency was reached. RNA was extracted using the Direct-zol™ RNA MicroPrep Kit (Zymo Research) following the manufacturer's instructions. 1 µg of RNA of each sample was used as input for library preparation using the KAPA Stranded mRNA-Seq Kit (Roche) following the manufacturer's instructions. Unique Dual-Indexed Set-B (UDI; KAPA biosystems) adapters were ligated and the library was amplified for 8 cycles. Libraries were sequenced on a Novaseq6000 as Paired-end 100 with 50 million fragments per library.

Generation of doxycycline-inducible HOXD4 overexpression systems in HAP1 cells

To generate a doxycycline-inducible overexpression system of HOXD4, we randomly integrated the coding sequences of HOXD4 wildtype, AroPERFECT and AroPLUS into HAP1 cells by using the PiggyBac transposon system.

N-terminally FLAG-tagged coding sequences of human HOXD4 wildtype, AroPERFECT or AroPLUS with a downstream 5xGS-linker were cloned into a backbone of the inducible Caspex expression vector (Addgene #97421) linearized by restriction digest with NcoI (NEB) and KpnI (NEB). Carrier plasmids and PiggyBac transposase expression vector (SBI, PB210PA-1) were co-transfected into HAP1 wildtype cells using Lipofectamine 3000 Transfection Reagent (Thermo Scientific) following the manufacturer's instructions in a molar ratio of 6:1. The transfected bulk population was screened for integration by addition of 2 µg / ml puromycin (Gibco) to the cell culture medium 24h after transfection for a total of 4 days. Bulk populations of every condition were induced by addition of 2 µg / ml doxycycline (Sigma) and screened for matching mEGFP expression levels across conditions by flow cytometry using a BD FACS Celesta. For the generation of clonal HOXD4 overexpression lines, bulk cells were single-cell sorted by using a BD FACS Aria II. HAP1 HOXD4 cells were directly sorted into wells of a 96-multiwell plate. Wells without any cells or with more than 2 cells were discarded. The other clones were expanded and eventually tested for HOXD4 expression level

upon DOX induction by flow cytometry. Cells with most similar expression levels were selected for further experiments.

Imaging of HAP1 HOXD4 PiggyBac overexpression cells

For sub-nuclear localization analysis of HOXD4 mutants, HAP1 cells with integrated HOXD4 overexpression cassettes were seeded onto chambered coverslips (Ibidi). After 24 hours, culture medium was substituted with medium containing 2µg/ml doxycycline to induce expression of *HOXD4* transgenes. On the next day, cells were washed with PBS and fixed with 4% paraformaldehyde for 15 minutes at room temperature. Afterwards, cells were stained with 0.25µg/ml DAPI (Invitrogen). Images were acquired with a Stellaris 8 confocal microscope and a Plan-Apochromat 100x/1.40 oil CS2 objective (Leica). For the analysis of sub-nuclear localization, a mosaic of at least 100 tile regions was imaged for each condition over two replicates. Object quantification was performed using ZEN 3.4 software (Zeiss). Briefly, DAPI counter stain was used to segment objects after Gaussian smoothing. Mean mEGFP intensities were then individually calculated for each segmented nucleus and the granularity was calculated by dividing standard deviation of mEGFP signal of each nucleus by the corresponding mean mEGFP signal using customer ImageJ/FIJI routines [144].

C/EBPα mediated B-cell to macrophage reprogramming

To induce C/EBPα-mediated B-cell to macrophage reprogramming, infected RCH-rtTA cells were seeded at $0.3x10^6$ cells/ml in RCH culture medium supplemented with 10 ng/ml of each, IL-2 (200-03, Preprotech) and CSF-1 (315-03B, Preprotech), as well as 2 µg/ml of doxycycline. The macrophage reprogramming was monitored by flow cytometry. Briefly, blocking was carried out for 10 min at room temperature using a 1:20 dilution of human FcR binding inhibitor (eBiosciences, 16-9161-73). Subsequently, cells were stained with antibodies against CD19 (APC-Cy7 Mouse anti-Human CD19, BD Pharmingen, 557791, 1:200) and Mac-1 (APC Mouse Anti-Human CD11b/Mac-1, BD Pharmingen, 550019, 1:200) at 4 ºC for 20 min in the dark. After washing, DAPI counterstaining was performed just before analysis. All analyses were performed using a LSR Fortessa (BD Biosciences). Data analysis was completed using FlowJo (v10) software.

FACS analysis of CD66a and FCGR2A

CD66 and FCGR2A expression levels were monitored by FACS analysis during C/EBPα-mediated B-cell to macrophage reprogramming. RCH-rtTA cells expressing either doxycycline-inducible CEBPα wildtype or AroPERFECT IS15 were seeded at $0.5x10^6$ cells/mL in RCH culture medium supplemented with 10 ng/ml of each, IL-2 (200-03, Preprotech) and CSF-1 (315-03B, Preprotech), as well as 2 µg/ml of doxycycline. Cells were collected at 24h

and 48h. Blocking was carried out for 10 min at room temperature using a 1:20 dilution of human FcR binding inhibitor (eBiosciences, 16-9161-73). Subsequently, cells were stained with antibodies against CD66a (Alexa Fluor 647 anti-human CD66a, BioLegend, 398905, 1:250) and FCGR2A (PE anti-human FCGR2A, BioLegend, 305503, 1:200) at 4 ºC for 20 min in the dark. After washing, DAPI counterstaining was performed just before analysis. All analyses were performed using LSR Fortessa (BD Biosciences). Data analysis was completed using FlowJo (v10) software.

Single cell RNA-seq (scRNA-Seq) data generation

One week after induction of C/EBPα-mediated B-cell to macrophage reprogramming, cells were collected and washed twice in PBS to remove dead cells and debris. Cells were resuspended in solution at a density of 700 cells/µl. We used the Chromium Next GEM Single Cell 3' technology for generating gene expression libraries from single cells. Briefly, gel beads-in-emulsion (GEMs) are generated by combination of barcoded Single Cell 3' v3.1 Gel Beads, a Master Mix containing cells, and Partitioning Oil on a Chromium Next GEM Chip G. To achieve single cell resolution, cells are delivered at a limiting dilution, such that the majority (~90-99%) of generated GEMs contain no cell, while the remainder largely contain a single cell. Immediately following GEM generation, gel beads were dissolved, primers were released, and any co-partitioned cell was lysed. Primer (containing an Illumina TruSeq Read 1, 16 nt 10x Barcode, 12 nucleotide unique molecular identifier and 30 nucleotide poly-dT sequence) were mixed with the cell lysate and a Master Mix containing reverse transcription (RT) reagents. Incubation of the GEMs produced barcoded, full-length cDNA from poly-adenylated mRNA. After incubation, GEMs were broken and pooled fractions were recovered. Silane magnetic beads were used to purify the first-strand cDNA from the post GEM-RT reaction mixture, which includes leftover biochemical reagents and primers. Barcoded, full-length cDNA was amplified via PCR to generate sufficient mass for library construction. cDNA was analyzed using Agilent Bioanalyzer assay (Agilent) to check size distribution profile and quantification. Only 25% of cDNA was used to 3' Gene Expression Library construction. Enzymatic fragmentation and size selection were used to optimize the cDNA amplicon size. TruSeq Read 1 (read 1 primer sequence) was added to the molecules during GEM incubation. P5, P7, a sample index, and TruSeq Read 2 (read 2 primer sequence) were added via End Repair, A-tailing, Adaptor Ligation, and PCR. The final libraries contained the P5 and P7 primers used in Illumina bridge amplification. Final libraries were analyzed using Agilent Bioanalyzer assay to estimate the quantity and check size distribution, and were then quantified by qPCR using the KAPA Library Quantification Kit (KapaBiosystems).

C/EBPα-GFP Chromatin immunoprecipitation-sequencing (ChIP-seq)

To study chromatin association of C/EBPα wildtype and AroPERFECT IS15, we performed ChIP-seq in RCH-rtTA C/EBPα wildtype and AroPERFECT IS15 cells 24 and 48 hours after induction of C/EBPα mediated macrophage reprogramming. The protocol was previously described [145]. Five million cells were collected and cross-linked for 10 min using 1% formaldehyde and quenched using a final concentration of 0.125 M glycine. After a wash in cold PBS and centrifugation, pellets were lysed in 500 µl pre-cooled SDS lysis buffer (1% SDS, 10mM EDTA, 50 mM Tris pH8 and 1x PIC) and incubated on ice for 15 min. Chromatin was sheared on a Bioruptor pico sonicator (Diagenode) at 4°C for 18 cycles of 30s ON and 30s OFF. After sonication, the solution was clarified by centrifugation at 1,000g for 5 min at 4°C and supernatant was transferred to a low-bind tube and mixed with 900µl of ChIP dilution buffer (0.01% SDS, 1.1% Triton X-100, 1.2 mM EDTA, 16.7 mM Tris-HCl pH 8.0, 167 mM NaCl, 1x PIC) containing antibody-coupled beads (10µL anti-GFP, clone 3E6 A-11120, ThermoFisher and 35 µL of protein G magnetic beads, 10003D, ThermoFisher). Five percent were saved as input and the samples were incubated overnight at 4°C under rotation. The beads were then collected and washed with 500µL of low salt buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl pH 8.0, 150 mM NaCl), high salt buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl pH 8.0, 500 mM NaCl), RIPA-LiCl buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA, 250 mM LiCl, 0.5% NP-40, 0.5% Na-DOC) and twice with TE buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA). Beads were then collected and eluted in 70µL of Elution buffer (10 mM Tris-HCl pH 8.0, 5 mM EDTA, 300 mM NaCl, 0.5% SDS) and incubated with proteinase K for 1h at 55°C and then overnight at 65°C to reverse the cross-linking. Beads were collected and transferred to a new tube and a second step of elution was performed with 30 µL of Elution buffer. DNA was finally purified with a Qiagen MinElute column and 3 ng of DNA were used to construct sequencing libraries using NEBNext® Ultra™ DNA Library Prep Kit for Illumina (E7370L). Libraries were sequenced on Illumina NextSeq2000 instruments using the 50 nucleotides single-end mode in order to obtain around 50 million reads per sample.

Generation of doxycycline-inducible NGN2 overexpression systems in human iPS cells

To generate a doxycycline-inducible overexpression system of NGN2, we randomly integrated the coding sequences of NGN2 wildtype, AroLITE and AroPERFECT into ZIP13K2 cells by using the PiggyBac transposon system.

N-terminally FLAG-tagged coding sequences of human NGN2 wildtype, AroLITE or AroPERFECT listed in section (Protein sequences and SLIMs) (Twist Bioscience) with a downstream T2A tag (Sigma) were cloned into a backbone of the inducible Caspex expression

vector (Addgene #97421) linearized by restriction digest with NcoI (NEB) and KpnI (NEB). Carrier plasmids and PiggyBac transposase expression vector (SBI, PB210PA-1) were co-transfected into ZIP13K2 wildtype cells using Lipofectamine Stem Transfection Reagent (Thermo Scientific) following the manufacturer's instructions in a molar ratio of 6:1. The transfected bulk population was screened for integration by addition of 2 µg / ml puromycin (Gibco) to the cell culture medium 24h after transfection for a total of 4 days. Surviving cells were seeded at low density under addition of 1x Y-27632 Rho-kinase inhibitor (biogems, 1293823) for the first 24 hours and expanded for several days until colonies derived from single cells were big enough to be picked and cultured separately. Clones of every condition were induced by addition of 2 µg/ml doxycycline (Sigma) and screened for matching mEGFP expression levels across conditions by flow cytometry using a BD FACS Celesta instrument.

NGN2-mediated neural differentiation of human iPSCs

We adapted our protocol for the differentiation of human iPSCs into neurons by overexpression of NGN2 from [146]. ZIP13K2 cells with integrated NGN2 overexpression cassette were cultured on Matrigel (Corning) pre-coated 10 cm culture plates. When cultures reached a confluency of approximately 80%, 2 µg/mL doxycycline (Sigma) was added to the culture medium to induce expression of the NGN2 transgene. After 24 hours, induced cultures were sorted for mEGFP expressing cells by flow cytometry unsing a BD FACS Aria II instrument. Positive cells were seeded on Matrigel pre-coated 96-well microclear plates (Greiner bio-one) in mTeSR+ and 1x Rho-kinase inhibitor at a density of $2x10^3$ cells / cm^2. On day 2, mTeSR+ medium was replaced by N2B27 neural cell culture medium (recipe) supplemented with 5 µg/ml human BDNF (biotechne). N2B27 medium was changed every day for a total of 4 days. Living cells were stained with 0.25 µg/ml Hoechst and Spy650-TUB (1:2000) (Spirochrome) and incubated in the microscope prior to image acquisition to equilibrate and thermalize all materials.

Live-cell imaging of human iPSC derived neurons

Living cells were imaged using the Celldiscoverer 7 Imaging Platform (Zeiss), in wide field mode running under ZEN Blue v 3.1 and full environmental control (5% v/v CO_2, 100% humidity, 37°C). Final experiments were performed using the Plan-Apochromat 20x / NA=0.7 objective, a 2x tube lens (Zeiss), and captured on an Axiocam 506 (Zeiss) with 3x3 binning resulting in a lateral pixel resolution of 0.347 µm/pixel. The fully automated imaging approach typically captured 20-40% of individual well surface. Focus stabilization was achieved by surface method in each third tile region. All images were acquired with one or two additional transmitted light or contrasting method (Brightfield, Oblique or Phase Gradient Contrast) channel. Each individual image position was acquired in consecutive sections of 3 slices

surrounding the focus position with a z-spacing of 0.63 μm to ensure the acquisition of each and every neurite. All parameters were kept identical during the experimental time course. The resulting large overview tile scan underwent a maximum-intensity projection and subsequent channel stitching using the nuclear counterstain (Hoechst) as reference. Cell numbers (Hoechst) and neurite density (SPY650) were quantified based on the respective channel.

Image analysis of nuclei and neurite densities in differentiated neurons

Wide field images were acquired using a 20x Air Objective (NA=0.7) a 2x optical post magnification on a Celldiscoverer 7 under ZEN 3.2 Blue (Zeiss). For each well and replicate a mosaic of 201 tile regions was imaged. A definite hardware focus was defined as the center for 3 slices of a consecutive z-stack with a slice distance of 0.34 µm. Image acquisition was performed using a Zeiss Axiocam 506, in a 3x3 binning mode, resulting in a lateral resolution of 0.34 µm/pixel. The resulting images were projected using Maximum Intensity Projection (MIP) in ZEN 3.4 (Zeiss) dedicated on analysis workstation. Object quantification was performed in the image analysis module in ZEN 3.4 (Zeiss). Briefly, within MIPs nuclei were identified by nuclear counter staining using Otsu intensity thresholds after faint smoothing (Gauss: 2,0), close by objects were segmented downstream by standard water shedding. Neurites were segmented by fixed intensity threshold on the respective staining without any water shedding.

KAPA Stranded mRNA-seq of ZIP13K2 NGN2 PiggyBac cells

At day 5 of NGN2-mediated neural differentiation, ZIP13K2 induced iNeurons were harvested following the Direct-zol™ RNA MicroPrep Kit (Zymo Research) standard protocol. 1 µg of RNA of each sample was used as input for library preparation using the KAPA Stranded mRNA-Seq Kit (Roche) following the manufacturer's instructions. Unique Dual-Indexed Set-B (UDI; KAPA biosystems) adapters were ligated and the library was amplified for 8 cycles. Libraries were sequenced on a Novaseq6000 as Paired-end 100 with 50 million fragments per library.

FLAG-NGN2 Chromatin immunoprecipitation-sequencing (ChIP-seq)

To study chromatin association of NGN2 wildtype, AroLITE and AroPERFECT, we performed ChIP-seq experiments in ZIP13K2 NGN2 wildtype, AroLITE and AroPERFECT-expressing cells 24 and 48 hours after induction of NGN2-mediated neural differentiation.

Cells were detached with Accutase solution (Sigma), washed twice in PBS and fixed in rotation with 1% formaldehyde for 10 minutes at room temperature. Subsequently, the reaction was quenched by addition of glycine leading to a final concentration of 125 mM. Per replicate, three

million cells were used as starting material. In brief, we followed the ChIPmentation protocol version 3 for histone marks and transcription factors [147]. Cells were lysed in lysis buffer 3 (10 mM Tris-HCl pH 8.0, 100 mM NaCl, 1 mM EDTA pH 8.0, 0.5 mM EGTA, 0.1% sodium deoxycholate and 0.5% N-laurosylsarcosine) supplemented with 1x cOmplete protease inhibitor cocktail. Afterwards, chromatin was sonicated for 10 minutes using a Covaris E220evolution focused-ultrasonicator with 2% duty cycles, 105 W peak incident power and 200 cycles per burst. Lysates were clarified by centrifugation for 10 minutes at 20,000g. 10% of the clarified lysate was put aside as input control. The remaining lysate was mixed with 50 µL of equilibrated anti-FLAG antibody (Merck, F1804, 1 µg total) coupled to Dynabeads™ Protein G magnetic beads (Invitrogen) and incubated on a 3D-shaker over night at 4°C. The next day, samples were washed twice in TF-wash buffer I (20 mM Tris-HCl pH 7.4, 150 mM NaCl, 0.1% sodium dodecyl sulfate, 1% Triton-X-100 and 2 mM EDTA pH 8.0) followed by two washes in TF-wash buffer III (10mM Tris-HCl pH 8.0, 250 mM LiCl, 1% Triton-X-100, 0.7% sodium deoxycholate and 1 mM EDTA pH 8.0) and a final wash with 10 mM Tris-HCl pH 8.0. All samples were tagmented for 5 minutes at 37°C using the Illumina Tagment DNA kit and immediately put on ice. Tagmented chromatin was washed in ice-cold wash buffer I and TET buffer (10 mM Tris-HCl pH 8.0, 5 mM EDTA pH 8.0, 0.2% Tween-20) twice each, and reverse-crosslinked for 1h at 55°C and 9h at 65°C in the presence of 300 mM NaCl and proteinase K (Ambion). Subsequently, DNA was purified using AMPureXP beads. Sequencing libraries were amplified using the Kapa HiFi HotStart Ready Mix (Roche) and Nextera custom primers (Illumina) [148] for a total of 12 cycles and paired-end sequenced on an NovaSeq6000 (Illumina) with a depth of ~50 million fragments per library.

Generation of doxycycline-inducible MYOD1 overexpression lines in C2C12 cells

To generate a doxycycline-inducible overexpression system of MYOD1, we randomly integrated the coding sequences of MYOD1 wildtype, AroLITE, AroPERFECT C and AroLITE C into C2C12 cells by using the PiggyBac transposon system.

N-terminally FLAG-tagged coding sequences of human MYOD1 wildtype, AroLITE, AroPERFECT C or AroLITE C listed in section (Protein sequences and SLIMs) (Twist Bioscience) with a downstream T2A sequence (Sigma) were cloned into a backbone of the inducible Caspex expression vector (Addgene #97421) linearized by restriction digest with NcoI (NEB) and KpnI (NEB). Carrier plasmids and PiggyBac transposase expression vector (SBI, PB210PA-1) were co-transfected into C2C12 wildtype cells using Lipofectamine3000 transfection reagent (Thermo Scientific) following the manufacturer's instructions in a molar ratio of 6:1. The transfected bulk population was screened for integration by addition of 2 µg /ml puromycin (Gibco) to the cell culture medium 24h after transfection for a total of 4 days.

Cells of every condition were induced by addition of 2 µg / ml doxycycline (Sigma) and screened for matching mEGFP expression levels across conditions by flow cytometry using a BD FACS Celesta instrument.

MYOD1-mediated myogenic differentiation of C2C12 myoblasts

C2C12 myoblasts with integrated MYOD1 overexpression cassette were seeded on chambered µ-Slide 8 Well ibiTreat coverslips (Ibidi). Upon reaching 85-90% confluence, 2 µg/ml doxycycline was added to the culture medium to induce MYOD1 transgene expression. Medium was changed every day over a total of 3 days. For imaging, cells were washed with PBS and fixed with 4% PFA for 15 minutes at room temperature. Cells were counterstained with DAPI.

Image analysis of differentiated C2C12 myotubes

Wide field Images were acquired using a 20x Air Objective (NA=0.7) a 2x optical post magnification on a Celldiscoverer 7 under ZEN 3.2 Blue (Zeiss). For each well and replicate a mosaic of 49 tile regions was covered. We defined the definite hardware focus as the center for 3 slices of a consecutive z-stack with a slice distance of 0.34 µm. Image acquisition was performed using a Zeiss Axiocam 506, in a 3x3 binning mode, resulting in a lateral resolution of 0.34 µm/pixel. The resulting images were projected using Maximum Intensity Projection (MIP) in ZEN 3.4 (Zeiss) on a dedicated Zeiss analysis workstation. Quantification of Fusion scores was conducted by implementation of a simple hierarchy order which was built within the image analysis module in ZEN 3.4 (Zeiss). We designed two segregating parent-classes by fixed intensity thresholds based on mEGFP signal resulting in fused myotubes (MT) and non-myotubes (NMT). Within these primary regions, nuclei were identified. Secondary objects were identified exclusively within primary objects (MT, NMT) by applying gaussian smoothing and fixed intensity thresholds on the nuclear counter staining, followed by standard water shedding the respective fluorescence image. All nuclei objects were filtered according to an area in between 30 and 300 μm^2.

KAPA Stranded mRNA-seq of C2C12 MYOD1 PiggyBac cells

At day 3 of MYOD1-mediated myogenic differentiation, MYOD1 induced myotubes were harvested following the Direct-zol™ RNA MicroPrep Kit (Zymo) standard protocol. 1 µg of RNA of each sample was used as input for library preparation using the KAPA Stranded mRNA-Seq Kit (Roche) following the manufacturer's instructions. Unique Dual-Indexed Set-B (UDI; KAPA biosystems) adapters were ligated and the library was amplified for 8 cycles. Libraries were sequenced on a Novaseq6000 as Paired-end 100 with 50 million fragments per library.

Identification of intrinsically disordered protein regions
Intrinsically disordered protein regions (IDRs) were predicted using the Metapredict v2 network at default settings [141].

Sequence disorder and pLDDT calculations
Disorder scores and pLDDT scores were calculated using Metapredict v2, and score plots were plotted using the built-in Metapredict graph plotting function.

Alphafold structure prediction
Alphafold models were downloaded from UniprotDB [149]. Models were rendered using UCSF ChimeraX, and were colored according to pLDDT scores.

Omega score (Ω_{Aro}) calculation
The Omega score was calculated using a modified localCIDER version [150]. Since the Omega score function is not length normalized, we adapted the python code to allow for variable interspace size referred in the package as the so-called blob size. This parameter is now calculated by dividing the sequence length over the fraction of aromatic residues. For this analysis only IDRs with a minimum of 3 aromatic residues were included. The mean of a random score was defined as the mean of 1000 kappa score calculations of randomly shuffled sequences from the original sequence. ggplot2 [151] was used for plotting violin plots and custom R to generate a distribution plot for the mean of random. One-way ANOVA with a post-Tukey test was used to compare IDR sets.

Bulk RNA-seq analysis
RNA-seq data from HAP1 and ZIP13K2 cells were mapped to a custom human genome hg38 assembly including the integrated mEGFP sequence using STAR aligner [152]. Count read tables were generated by the same software. C2C12 RNA-seq data was mapped to the mm10 mouse genome assembly using the above-mentioned software. Differential expression analysis was performed by the DEseq2 package [153] in R version 4.2. [154]. Differentially expressed genes were defined as having a fold-change bigger or equal to 1.5, *P*-value from Benjamini–Hochberg method smaller or equal to 0.01, and a minimum mean read count across the experiment samples of 50 reads. For the HAP1 data set, KO samples were compared to parental lines, and AroPERFECT and AroPLUS were compared to the HOXD4 wildtype line. For ZIP13K2 datasets, the NGN2 wildtype line was compared to the parental ZIP13K2 line. NGN2 AroLITE and AroPERFECT were compared to the NGN2 wildtype line. Genes were considered as NGN2 targets if they were differentially expressed in the parental ZIP13K2 vs NGN2 wildtype comparison and had a peak assigned in the NGN2 wildtype ChIP-Seq analysis. For C2C12 experiments, we compared the gene expression in the MYOD1

wildtype line to parental C2C12 cell gene expression and AroLITE, AroLITE-C, AroPERFECT and AroPERFECT-C variants to MYOD1 wildtype. Principal component analyses were carried out using the “PCAPlot” function from the DEseq2 package on the normalized read matrix that was transformed by using the variance stabilizing transformation (VST) function from the DEseq2 package and plotted by ggplot2. Volcano plots were plotted in ggplot2. Heatmaps were plotted with the help of the “ComplexHeatmap” package [155] in R, and cluster analysis was done by k-means clustering using the “cluster” [156] package in R.

Gene set enrichment analysis of the MYOD1 RNA-seq was conducted using “GSEAPreranked” v6.0.12 [157] with 1000 permutations on a ranked list of genes from the comparisons AroPERFECT-C vs. wildtype and wildtype vs. Parental sorted by Wald statistic (stat) [153] against the Wikipathways cell adhesion gene set in *Mus musculus* [158]. Empirical *P*-values were used for generating plots. Highest ranking genes in the AroPERFECT-C vs. wildtype comparison that were defined as MYOD1 targets were highlighted in the volcano plots.

Single-cell RNA-seq analysis

Data pre-processing
The single-cell RNA-seq datasets were processed with 10x Genomics' Cell Ranger pipeline version 3.1.0 [159] and mapped to a custom human hg38 genome assembly including mEGFP and codon-optimized C/EBPα wildtype, C/EBPα AroPERFECT IS15, and C/EBPα AroPERFECT IS10 sequences. The Cell Ranger hdf5 files were processed by Seurat package version 4.0.6 [160].

Filtering and normalization
Cells with more than 2000 expressed genes and genes with more than 5 reads across the samples were considered for analysis. Further filtering was done by removing cells with more than 20% mitochondrial read counts and less than 5% ribosomal read counts. The top 10 genes associated with PCA components were checked for mitochondrial and ribosomal genes. Next, cells were scored for cell cycle, and gene expression of S and G2M associated genes was regressed to eliminate any dependence on cell cycle to clustering. Doublets were identified and filtered out. mEGFP, C/EBPα wildtype, C/EBPα AroPERFECT IS15, and C/EBPα AroPERFECT IS10 reads were then used to identify mEGFP positive cells. Subsequently, expression of filtered cells was transposed to the metadata so it would not affect clustering. Finally, the Harmony package was used to batch correct the 3 libraries.

Cluster identification

Cluster identification was carried out using Seurat's built-in functions FindvariableGenes, RunPCA, RunUMAP, and FindClusters by first identifying the genes with the highest variation across all samples and cell types, building a shared nearest neighbor graph, and then running the Louvain algorithm on it. The number of clusters was determined by the optimum of the modularity function from the Louvain algorithm. The number of mEGFP-positive cells was then calculated for each cluster and was used to filter untransformed cell clusters; mainly cluster 0 and cluster 2.

Assignment of cell types to clusters

Cell type cluster assignment was based on a comparison of the market set from bulk RNA-seq experiment from [161] and augmented by both RNA velocity analysis and known markers for both, B-cell and macrophage cell types. In brief, RNA-Seq data and marker sets were retrieved from [161]. Raw fastq files were mapped against the human v38 genome assembly using STAR and aligned reads were counted. Raw count data was then processed in DESeq2 and normalized by VST transformation. Marker set VST data was retrieved and clustered according to the methods described in [161] and each gene was assigned to a gene cluster for Early, Early-Inter, Inter1, Inter2, Inter-Late, Late1, and Late2 as described in the publication. This assignment was referred to as “Choi et al. differentiation clusters”. To quantify the number of genes that are higly expressed in each single cell cluster, single cell gene expression was averaged within each single cell cluster and normalized to z-score. Normalized gene expression for the marker set above was clustered by k-means clustering with k= 8 in an efford to separate each single cell cluster by expression profiles. A heatmap was generated using complexHeatmap to visualize the expression profiles across clusters. This analysis helped to define B-cell and macrophage populations and assigned them to differentiation stages. Pseudotime and PAGA graph analysis also was used to visualize the trajectory of differentiating cells by giving temporal context to the single cell clusters.

Differential expression analysis

Inter-cluster differential expression analysis was performed by Wilcox test using the FindMarkers function with default settings and inter-sample cluster differential expression analysis between wildtype and IS15 cells in cluster 7 using the FindMarkers function with DESeq2 function. A q-value cutoff of 0.05 was used to define differentially expressed genes for the Wilcox test, and an adjusted *P*-value from Benjamini–Hochberg method of 0.05 was used for the inter-sample test. Volcano plots and bar plots were plotted in ggplot2, violin, UMAP and feature plots by Seurat's VlnPlot, FeaturePlot, and DimPlot function. Dot plots were plotted using a custom function to modify the output of the complexHeatmap package.

RNA velocity

We generated loop files necessary for RNA velocity using velocyto [162] and exported barcodes, expression matrix, metadata and UMAP coordinates from Seurat to csv files. scVelo [163] was used to build the manifold, calculate and visualize RNA velocity using a generalized dynamical model to solve the full transcriptional dynamics. A PAGA graph [164] was calculated using this model to visualize cell trajectories. Pseudotime was calculated by Markov diffusion process and plotted by a scVelo bult-in function.

ChIP-seq analysis

ChIP-seq data from C/EBPα and NGN2 experiments were mapped to a custom human genome hg38 assembly using BWA v0.7.17 [165]. Samtools was used for SAM to BAM conversion, sorting and indexing [166]. The Genome Analysis Toolkit v4 [167] was used to remove duplicated reads. Peak calling was performed using MACS3 v3.0.0 b1 [168] using the input of the respective sample as reference. Analysis and differential peak calling were done using DiffBind v3.6.5. Normalization was done using the native method and background input. We used the DEseq2 method for differential peak calling with an FDR threshold set to 0.01. Peaks were visualized using the DiffBind “plotprofile” function with default settings for general profiles. PCA analysis was run on normalized count samples and plotted with DiffBind.

Results

Sequence composition of human IDRs is not sufficient to explain function.

Approximately 25% of the human proteome is predicted to be disordered and more than half of all human proteins contain an intrinsically disordered region. Historically, these regions have been understudied, and to this day, their influence on cellular function remains not fully understood. To systematically assess differences and similarities in biochemical characteristics of human intrinsically disordered regions, I utilized the curated Uniprot reference proteome of *Homo sapiens* to create a comprehensive overview (Figure 7). Following additional filtering of reference protein sequences, disordered regions were predicted proteome-wide using Metapredict v2 [141].

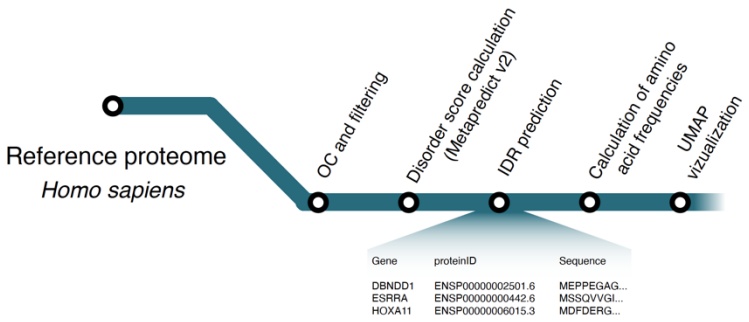


Figure 7: Schematic of the workflow for proteome wide IDR predictions. *IDRs were predicted using Metapredict v2. For visualization, we constructed UMAP plots based on an amino acid frequency matrix.*

I calculated amino acid frequencies proteome-wide, as well as for predicted IDRs.

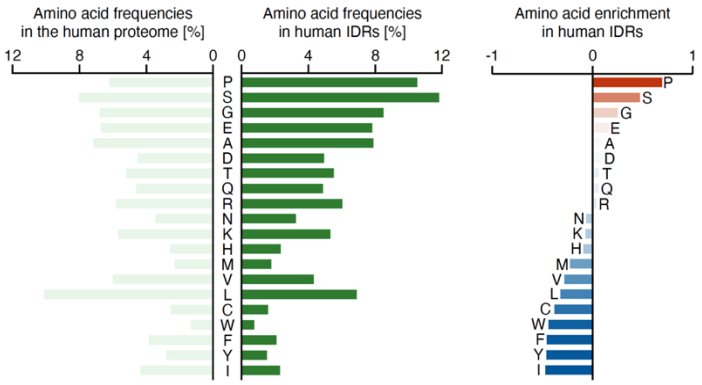


Figure 8: Amino acid frequencies of the human proteome and in predicted IDRs. *(left) Calculated amino acids frequencies for every amino acid in all open reading frames of the human reference proteome and the predicted IDR set. (right) Enrichment of amino acids within regions of predicted disorder compared to the reference proteome. The color of the bars corresponds to the degree of enrichment of the respective amino acid.*

The strongest enrichment in disordered regions was calculated for proline (P), serine (S), glycine (G) and glutamine (E), while disordered regions were depleted for the aromatic amino acids phenylalanine (F), tyrosine (Y) and tryptophan (W) as well as for amino acids with strong

hydrophobic side-chains like isoleucine (I) and leucine (L) (Figure 8). Surprisingly, despite its disorder-reducing nature, alanine (A) showed slight enrichment in human IDRs [169]. To more efficiently resolve IDR sequence characteristics, amino acid frequencies for every of the 15,347 predicted IDRs in the human proteome were calculated. This 20-dimensional matrix was then visualized as a UMAP (Uniform Manifold Approximation and Projection) plot, where each dot represents one IDR, highlighted in the color corresponding to its most abundant amino acid (Figure 9). This two-dimensional phenotype space encompasses the full spectrum of human IDR compositions, with major discriminants being hydrophobicity and charge, and reveals no relationship between length and sequence composition (Figure 9).

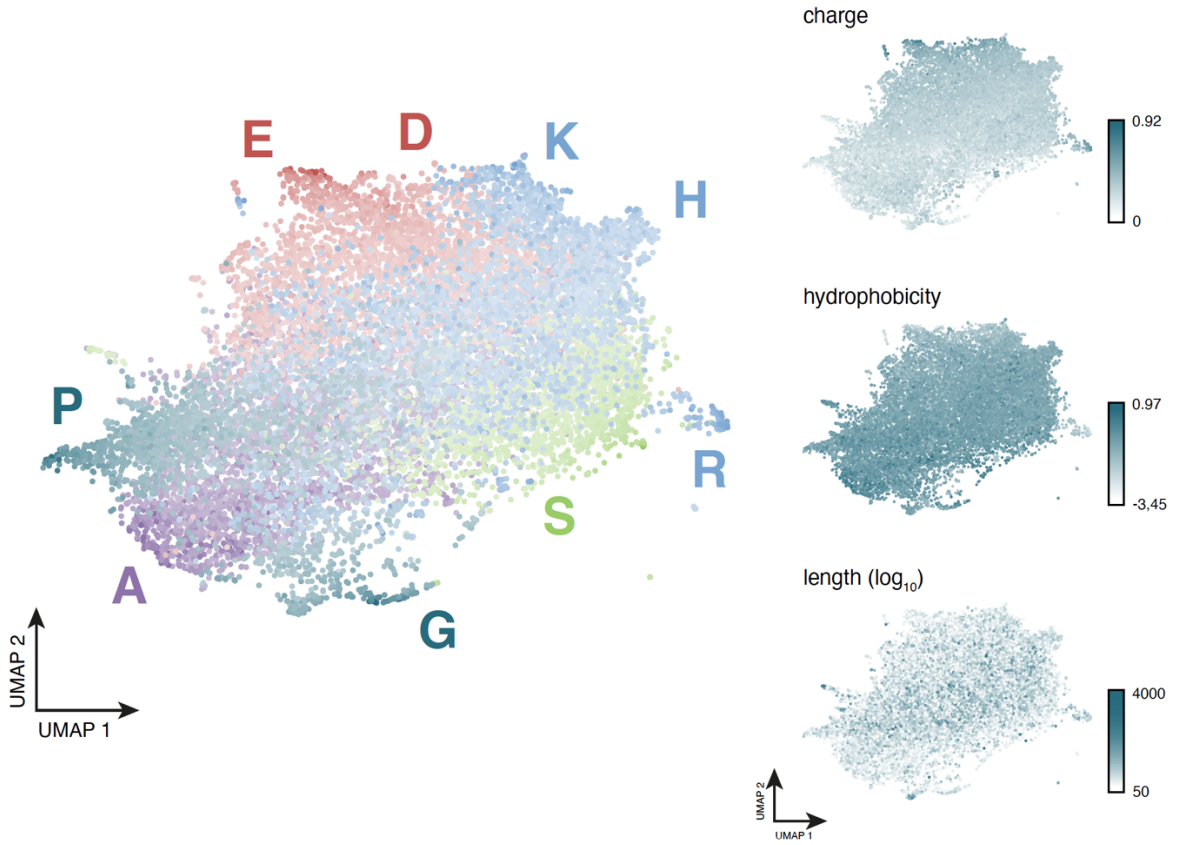


Figure 9: The compositional phenotype space of human IDRs. *(left) UMAP visualization of predicted human IDR sequences. Each dot represents one IDR sequence highlighted in a color corresponding the most represented amino acid in the respective sequence. S, serine (green), A, alanine (purple), K, lysine (blue), H, histidine (blue), R, arginine (blue), E, aspartic acid (red), D, glutamic acid (red), P, proline (teal), G, glycine (teal). (right) UMAP plots of the human IDR phenotype space highlight sequence charge, hydrophobicity, and length (log_{10}).*

The amino acid composition alone, did not separate IDRs into functionally relevant groups. Moreover, amino acid composition did not drive separation of IDRs with similar subcellular localization of the respective protein or association with either membrane-bound or membrane-less compartments, when using subcellular localization annotations from the Human Protein Atlas (Figure 10) [170]. Therefore, sequence composition of human IDRs alone

seems to be insufficient to explain sequence specific partitioning of molecules into membrane-less nuclear organelles like the nucleolus or nuclear speckles.

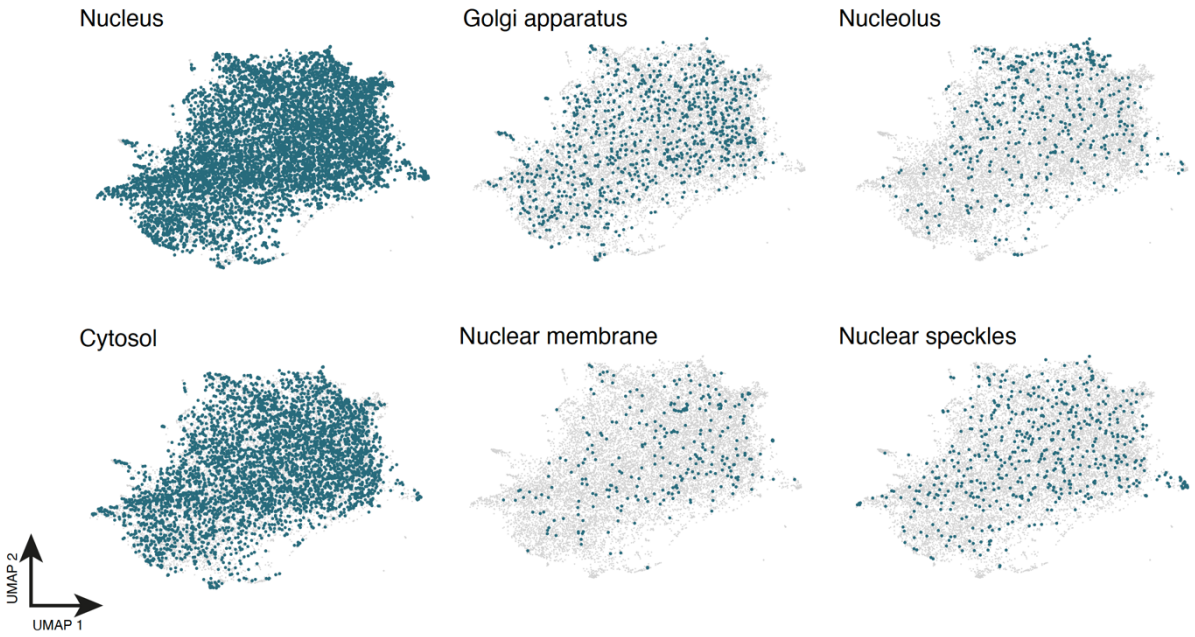


Figure 10: The amino acid composition of human IDRs does not explain sub-cellular localization of the respective protein. *Shown are UMAP visualizations of the human IDR phenotype space. Highlighted IDRs are part of proteins associated with annotated membrane-bound and membrane-less compartments following Human Protein Atlas annotation.*

I focused on approximately 1,500 human transcription factors specifically, using a previously published catalog of human TFs [40]. This revealed high compositional variability of human IDR sequences (Figure 11). When segregating TF IDRs by their respective TF family annotation, only minor compositional relatedness was observed. A dense cluster of histidine-rich IDRs mostly containing IDRs of KRAB-ZF transcription factors, could be attributed to the highly conserved and partly disordered KRAB domain [171]. Additionally, I observed many TF IDRs with high alanine content; however, overall, the sequence composition of IDRs alone failed to elucidate TF function.

Figure 11: Compositional variability of human transcription factor IDRs. *UMAP visualization of the human IDR phenotype space. (left) Highlighted IDRs are encoded by transcription factors using published human TF annotation. (right) Transcription factor IDRs colored according to TF family classification. ZF, zinc finger; KRAB, Krüppel associated box; bHLH, basic helix-loop-helix; bZIP, basic leucine-zipper.*

Condensation and transcriptional activity are inherently linked features of TF function

Intrinsically disordered regions of human transcription factors are essential for the formation and function of transcriptional condensates. However, the sequence features mediating condensation and subsequent transcriptional activation are not fully understood. Since neither the amino acid composition nor predicted minimal activation domains within human TF IDRs explain the full spectrum of TF function, I took inspiration from non-linear sequence features that have been associated with condensate formation. One such feature is the dispersion of aromatic amino acids in PLD-containing proteins [69].

To gain initial insights into the importance of aromatic amino acids in TF IDRs, I selected the three human transcription factors HOXB1, HOXD4 and HOXC4 for functional testing. I predicted the IDRs of the three factors and designed “AroLITE” mutant sequences in which all aromatic amino acids in the wild-type IDR were mutated to alanine or serine, as indicated (Figure 12a).

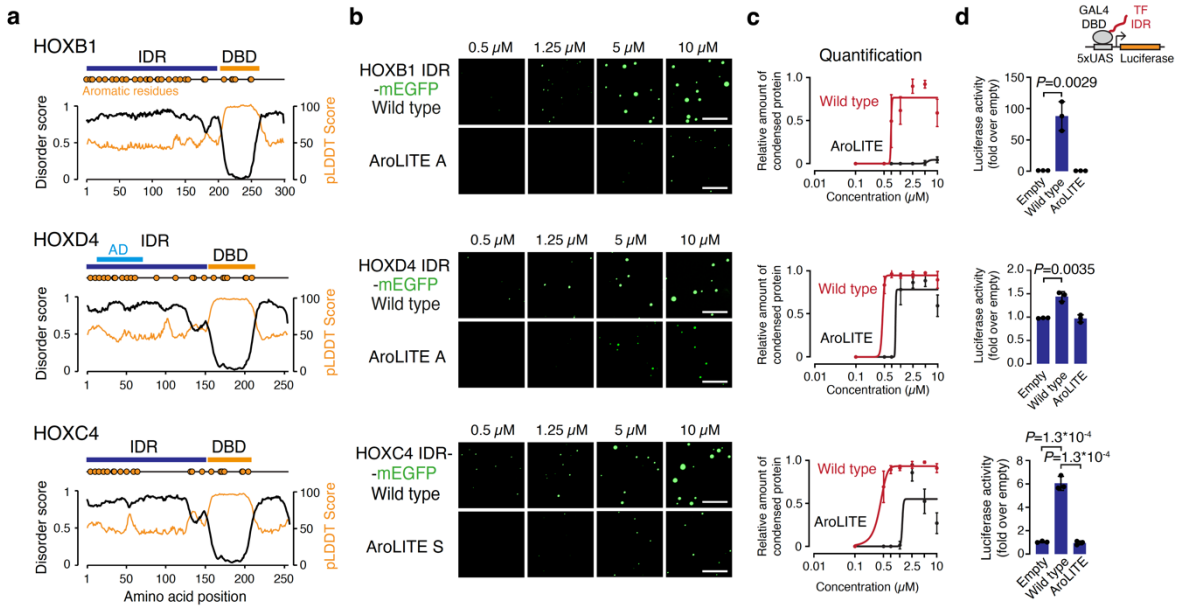


Figure 12: Aromatic amino acids in TF IDRs contribute to transactivation and condensate formation in vitro. *(a) Disorder plots for HOXB1, HOXD4 and HOXC4 predicted by Metapredict v2 (black) and Alphafold pLDDT (yellow). Predicted minimal activation domains (AD) highlighted in light blue. (b) Representative images of droplet formation of purified, recombinant HOXB1, HOXD4 and HOXC4 IDR-mEGFP proteins. Scale bar: 5 µm. Data generated by Yaotian Zhang (c) Quantification of the droplet assays. Data displayed as mean ± SD. N = 10 images from 2 replicates. The curve was generated as a non-linear regression to a sigmoidal curve function. Data generated by Yaotian Zhang (d) Schematic and results of luciferase reporter assays. Luciferase values were normalized against an internal Renilla control, and the values are displayed as percentages of the activity measured using an empty vector. Data displayed as mean ± SD, from three biological replicates. P-values are from two-sided unpaired t-tests.*

I expressed recombinant fusion proteins of the wild type and AroLITE IDRs tagged with mEGFP in *E. coli*. The propensity of these fusion proteins to undergo homotypic condensation was then tested in a concentration dependent manner in *in vitro* droplet formation assays by incubation of the purified fusion protein in the presence of the crowding agent polyethylene

glycol-8000. All sequences tested formed droplets in a concentration dependent manner (Figure 12b). The droplets formed exhibited hallmarks of LLPS such as fusion with one another or by wetting the microscopy slide. The phase behavior of the IDRs was quantified by calculating the relative amount of condensed protein at a given concentration (Figure 12c). Surprisingly, AroLITE mutants of all three proteins tested showed reduced droplet formation and an increased critical saturation concentration (c_{sat}) compared to the wild type. Consequently, when tested in luciferase reporter assays for transcriptional activity, all three AroLITE mutants - fused to a GAL4 DNA binding domain and co-transfected with a 5xUAS-driven luciferase reporter into mouse embryonic stem cells - demonstrated reduced activity (Figure 12d). While the wild type sequences of HOXB1, HOXD4 and HOXC4 displayed strong to moderate transcriptional activity, the activity of all AroLITE mutants was negligible, comparable to the empty control.

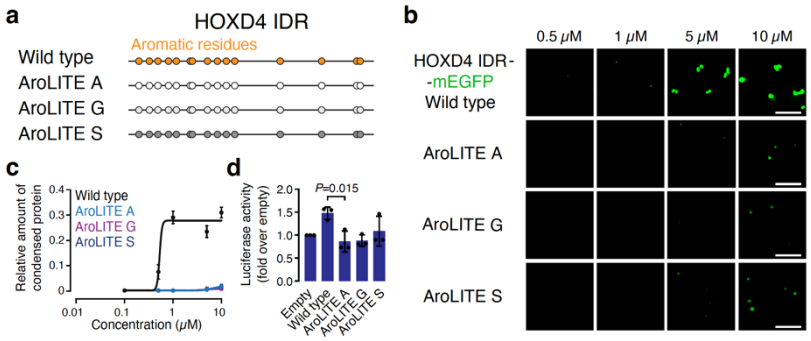


Figure 13: Depleting aromatic residues in the HOXD4 IDR abolishes transcriptional activity and condensate formation *in vitro*. *(a) Schematic of mutated aromatic residues in the HOXD4 IDR. (b) Representative images of droplet formation of purified, recombinant HOXD4 wild type and AroLITE IDR-mEGFP proteins. Scale bar: 5 µm. (c) Quantification of the droplet assays. Data displayed as mean ± SD. N = 10 images from 2 replicates. The curve was generated as a non-linear regression to a sigmoidal curve function. (d) Values are displayed as percentages of the activity measured using an empty vector. Data displayed as mean ± SD, from three biological replicates. P-values are from two-sided unpaired t-tests.*

To rule out an alanine-specific phenotype, I mutated all aromatic residues in the HOXD4 IDR not only to alanine but also to serine and glycine, two other small, disorder-promoting amino acids (Figure 13a). Mutating aromatic residues in the HOXD4 IDR into any of the three amino acids decreased droplet formation *in vitro* and abolished transcriptional activity (Figure 13b-d). This result was supported when testing three additional candidates EGR1, NANOG and NFAT5 (Figure 14a). Furthermore, the reduction of transcriptional activity was not specific to mouse embryonic stem cells (mESCs) as reporter assays performed in immortalized embryonic kidney cells (HEK293T) and neuroblastoma cells (Kelly, SH-SY5Y) using the HOXB1 IDR showed similar results (Figure 14b).

Aromatic amino acids appear to be key for transcriptional activity of the TF IDRs tested. Furthermore, the presence of aromatic residues is indispensable for the ability of the TF IDRs,

to undergo efficient homotypic condensation *in vitro.* This suggests that aromatic amino acids contribute to a sequence feature that inherently links two key aspects of TF function: condensation and transactivation.

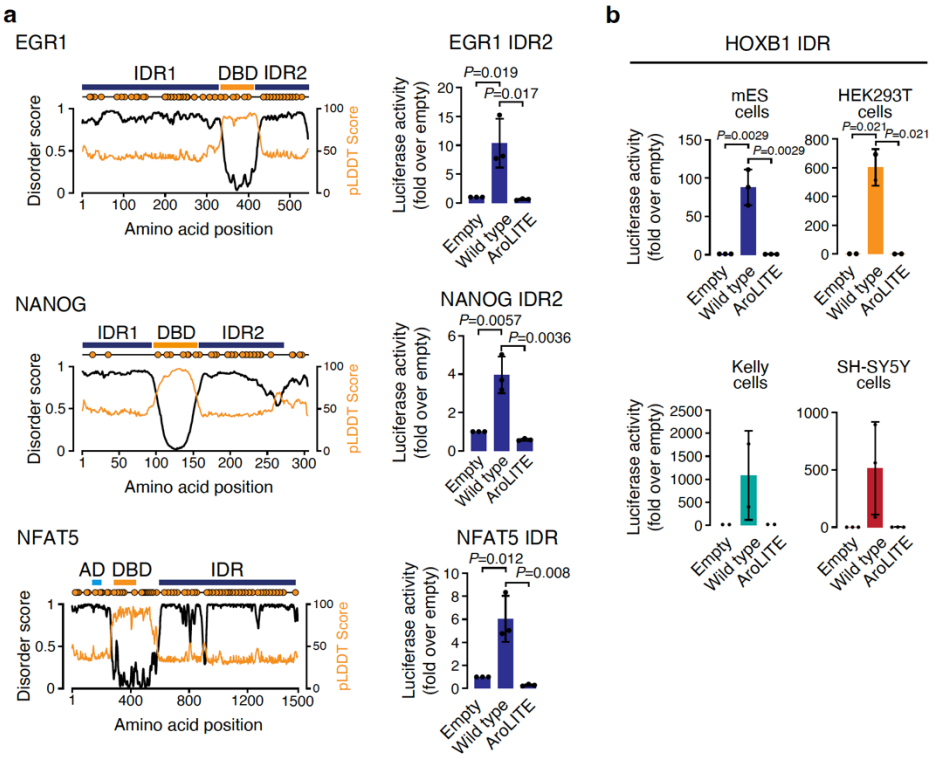


Figure 14: Reduction of transcriptional activity upon mutagenesis of aromatic residues is TF and cell line independent. (a) (left) *Disorder plots for EGR1, NANOG and NFAT5 predicted by Metapredict v2 (black) and Alphafold pLDDT (yellow). Predicted minimal activation domains (AD) highlighted in light blue. (right) Luciferase values displayed as percentages of the activity measured using an empty vector. Data displayed as mean ± SD, from three biological replicates. P-values are from two-sided unpaired t-tests. (b) Luciferase values displayed as percentages of the activity measured using an empty vector. Data displayed as mean ± SD, from three biological replicates for mESCs and two biological replicates for the other cell types. P-values are from two-sided unpaired t-tests.*

Suboptimal dispersion of aromatic amino acids in transcription factor IDRs

To evaluate the dispersion of aromatic amino acids in TF IDRs, Ω_{Aro} score calculations were performed. This patterning parameter, previously described by the lab of Tanja Mittag, quantifies the patterning of aromatic amino acids within a query sequence [69]. It assesses dispersion by calculating the Ω_{Aro} score, which is the normalized standard deviation of the distances between aromatic amino acids. A comparison of the result with 1000 randomly shuffled sequences of the same composition enables the calculation of an empirical p-value.

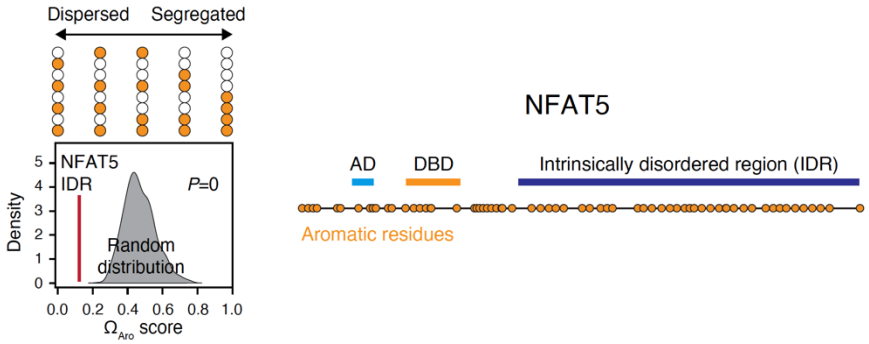


Figure 15: The Ω_{Aro} score as a patterning parameter for the dispersion of aromatic residues. *(left) Omega plot of the NFAT5 IDR. Empirical P-value is reported. (right) Positioning of aromatic residues in NFAT5. AD, activation domain; DBD, DNA-binding domain. Analysis was performed and data was plotted by Sebastian Mackowiak.*

The Ω_{Aro} score calculation for the NFAT5 IDR yielded a low Ω_{Aro} score of 0.124 with an empirical P-value = 0, indicating that the 30 aromatic amino acids in the NFAT5 disordered region are more uniformly dispersed than in all 1,000 of the randomly generated sequences with the same amino acid composition, implying a dispersion that is more pronounced than expected by random chance (Figure 15).

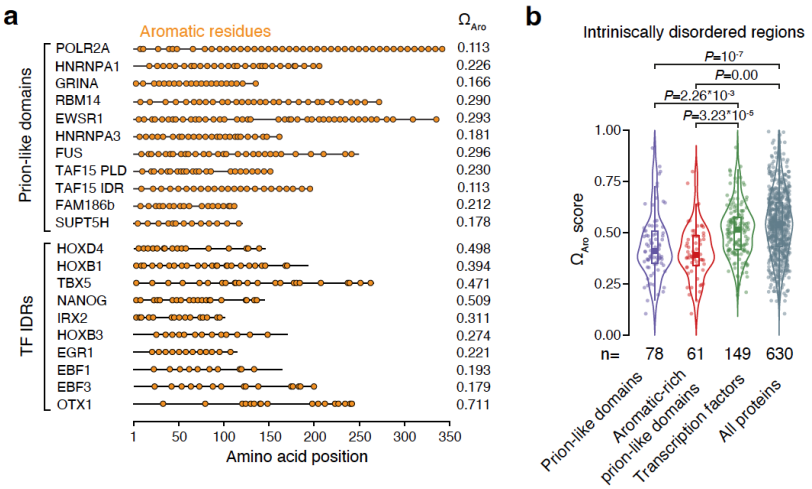


Figure 16: Dispersion of aromatic residues in human TF IDRs is submaximal. *(a) Schematic models of the patterning of aromatic residues in prion-like domains and TF IDRs including Ω_{Aro} scores. (b) Omega scores of IDRs in various protein classes. P-values are from one-way ANOVA with Tukey's multiple comparisons post-test. Analysis was performed and data was plotted by Alexandre Magalhães.*

While aromatic dispersion was notable in some transcription factor IDRs, such as NFAT5 or the C-terminal IDR of EGR1, the overall dispersion in human TFs appeared to be submaximal when compared to that in PLD-containing proteins, like HNRNPA1 or FUS (Figure 16a). To evaluate the prominence of this sequence feature compared to other protein classes, Ω_{Aro} scores were calculated proteome wide. We identified PLDs using the PLAAC prediction tool as previously described [172]. A direct comparison of Ω_{Aro} scores between human TF IDRs and PLDs indicated that, on average, human TF IDRs have higher Ω_{Aro} scores compared to PLDs, including aromatic-rich PLDs (Figure 16b). This suggests that human TF IDRs encode submaximal aromatic dispersion.

Optimized dispersion of aromatic residues enhances transcriptional activity

To determine whether the dispersion of aromatic residues in TF IDRs is a non-linear sequence feature that controls transcriptional activity, I selected EGR1, a candidate TF with a C-terminal IDR encoding 13 highly dispersed aromatic residues (Ω_{Aro} = 0.242, p < 0.01). Assessing the transactivation strength of the EGR1 IDR in reporter assays revealed moderate activity. I designed mutants of the EGR1 IDR with varying dispersion of aromatic residues while keeping the overall sequence composition constant. First, I created an AroSCRAMBLED mutant with a random distribution of the 13 aromatic residues along the IDR sequence (Ω_{Aro} = 0.564), and two additional mutants in which I clustered the aromatic residues into randomly distributed groups of 4 to 5 residues (Ω_{Aro} = 0.999) or into a single contiguous stretch of 13 residues (Ω_{Aro} = 0.999). Testing the three mutant sequences along with the wild type sequence revealed a decrease in transcriptional activity in the mutants that correlated with the degree of aromatic dispersion in the respective sequence (Figure 17).

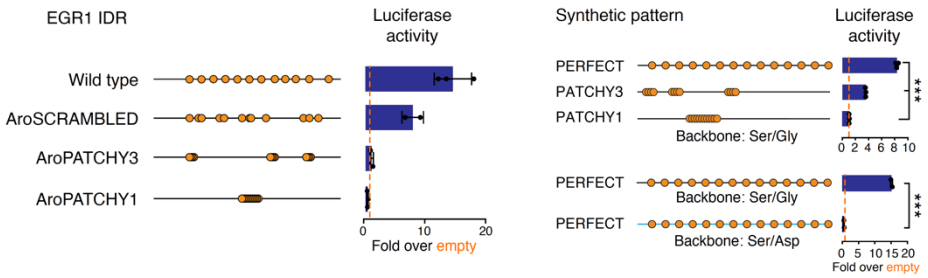


Figure 17: Dispersion of aromatic residues correlates with transcriptional activity. *(left) Reporter assays with the EGR1 IDR. (right) Reporter assays with synthetic sequences. Luciferase values displayed as percentages of the activity measured using an empty vector. Data displayed as mean ± SD, from three biological replicates P-values are from two-sided unpaired t-tests. ***: P<0.001*

This observation held true using synthetic sequences of 100 amino acids in length. The composition of these synthetic sequences was inspired by sequence features of prion-like domains. I introduced tyrosine residues with varying degrees of aromatic dispersion into an amino acid backbone composed of randomly arranged serine and glycine residues (1:1). In reporter assays, I measured transcriptional activity in the synthetic sequence featuring the PERFECT pattern, which encoded optimally dispersed tyrosine residues with fixed distances to neighboring aromatic residues. Analogous to the experiment with the EGR1 IDR, the activity of PATCHY3 and PATCHY1 mutants decreased compared to the PERFECT, proportional to the degree of aromatic dispersion. To further investigate this sequence feature in a system mimicking acidic activation domains of eukaryotic transcription factors, I generated an additional PERFECT mutant, replacing all glycine residues in the backbone with aspartic acid [173]. In an acidic environment, the perfectly dispersed aromatic residues failed to mediate transcriptional activity (Figure 17). Therefore, the transcriptional activity of the EGR1 IDR appears to depend not only on the presence of aromatic amino acids but also on their

arrangement, with optimally dispersed residues promoting transactivation most effectively, all within the constraints of the amino acid backbone in which the pattern is embedded.

Based on the findings in synthetic PLD-like IDRs, I set out to investigate if optimizing aromatic dispersion can enhance the transcriptional activity of human TF IDRs. To this end, I choose HOXD4, a homeobox transcription factor containing an IDR of 140 amino acids in its N-terminus. I selected the HOXD4 IDR because it contains a significant number of aromatic residues and encoded submaximal aromatic dispersion with a low Ω_{Aro} score in the first 71 amino acids (Ω_{Aro} = 0.210), but a low density of aromatic residues towards the C-terminal end of the IDR, resulting in an overall Ω_{Aro} score of 0.510 (Figure 18a).

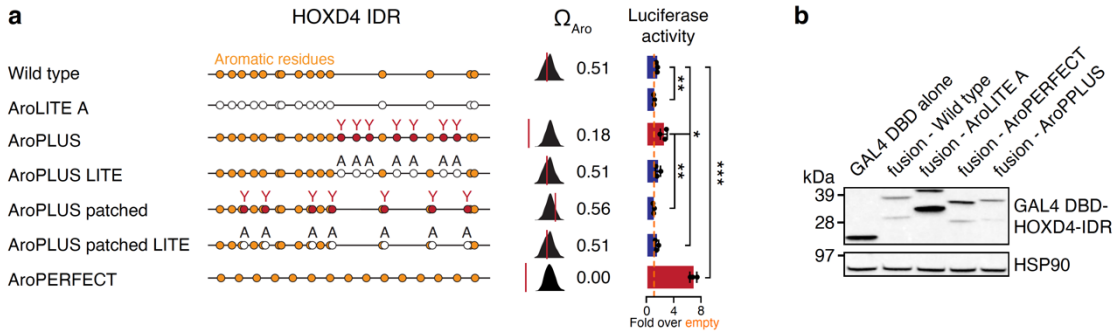


Figure 18: Optimized aromatic dispersion increases transcriptional activity of the HOXD4 IDR. *(a) (left) Schematic models of the HOXD4 IDR variants tested with Ω_{Aro} scores (right) Reporter assays of HOXD4 IDRs. Luciferase values displayed as percentages of the activity measured using an empty vector. Data displayed as mean ± SD, from three biological replicates P-values are from two-sided unpaired t-tests. *: P<0.05, **: P<0.01. (b) Western blot of overexpressed GAL4-fusion proteins. HSP90 was used as a housekeeping control.*

I followed two conceptually different approaches to optimize aromatic dispersion. Firstly, I sought to extend the N-terminal pattern of the HOXD4 IDR, which already demonstrates a high degree of aromatic dispersion by mutagenesis of seven amino acids of the C-terminal part to tyrosine. Thereby, I expanded the dispersed region to the end of the IDR, culminating in an overall Ω_{Aro} score of 0.180. In luciferase reporter assays, this "AroPLUS" mutant exhibited significantly higher transcriptional activity compared to the wild type IDR (P<0.05, t-test). Creating an "AroPLUS LITE" mutant, by mutating the same seven amino acids to alanine, did not lead to increased activity. Moreover, mutating seven amino acids directly adjacent to existing aromatic residues did not enhance activity, therefore excluding the possibility that the enhanced activity is caused by an increased number of aromatic amino acids. Secondly, I engineered an "AroPERFECT" mutant by uniformly dispersing all aromatic residues from the wild type sequence without altering the backbone composition. This mutant demonstrated a ~5-fold increase in activity compared to the wild type (P<0.001, t-test) (Figure 18a). Differences in luciferase activity were not attributable to variations in protein expression levels, as confirmed by Western Blot using a Gal4-DBD-specific antibody (Figure 18b).

I further dissected this gain-of-function effect with additional mutant sequences of the HOXD4 IDR. To determine if the enhanced activity was due to optimizing the dispersion of aromatic residues, I hypothesized that shifting the optimally dispersed pattern by one or two amino acid positions to the left should maintain the effect. However, testing these sequences in the luciferase reporter assay showed that shifting the optimally dispersed pattern by one position negated the enhanced effect of the AroPERFECT mutant. Moreover, shifting the pattern by two positions only marginally increased activity compared to the wild type sequence (Figure 19). These results corresponded with the count of small inert residues adjacent to aromatic residues in each sequence (Figure 20).

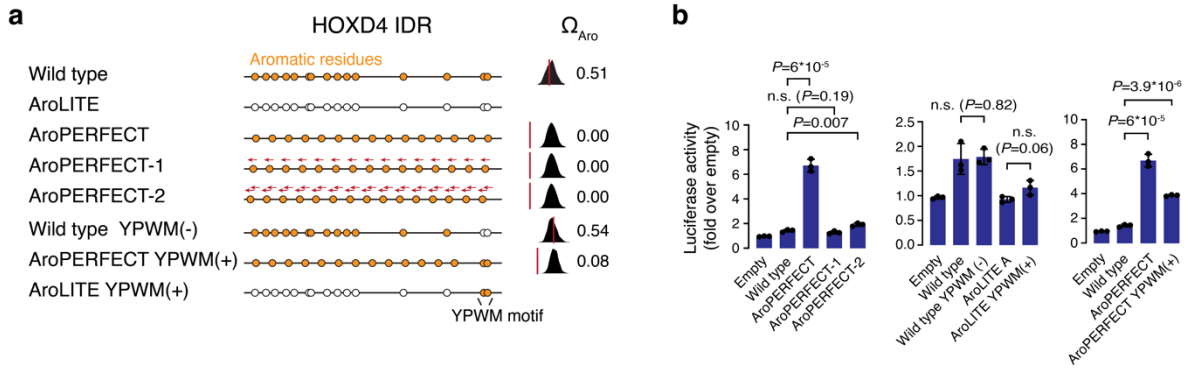


Figure 19: Optimized aromatic dispersion enhances transcriptional activity within the constraints of the backbone sequence. *(a) Schematic models of the HOXD4 IDR variants tested with Ω_{Aro} scores. (b) Reporter assays of HOXD4 IDRs. Luciferase values displayed as percentages of the activity measured using an empty vector. Data displayed as mean ± SD, from three biological replicates P-values are from two-sided unpaired t-tests.*

I considered the possibility that redistribution of aromatic residues can influence the integrity and function of short linear motifs (SLiMs), such as interaction interfaces with co-regulators. The HOXD4 IDR contains a documented SLiM near its C-terminus. This “YPWM motif” is known to be essential for proper heterodimerization of HOXD4 with PBX factors [174]. To ascertain whether the loss of the interaction with PBX factors was responsible for enhanced transcriptional activity, I engineered mutant IDRs where I mutated the aromatic residues within the SLiM to alanine (Wild type YPWM(-)), an AroPERFECT YPWM(+) mutant with an intact SLiM and nearly optimal dispersion of aromatic residues, and an AroLITE YPWM(+) mutant with an intact SLiM but all other aromatic residues mutated to alanine. Reporter assays showed no significant difference in transcriptional activity between wild type and wild type YPWM(-), AroLITE and AroLITE YPWM(+), or AroPERFECT

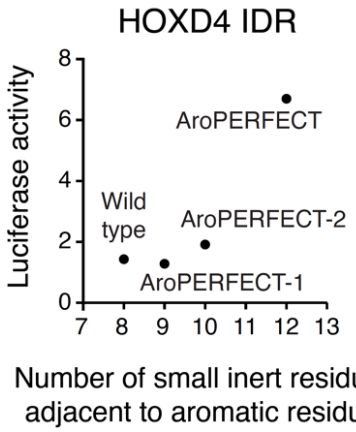


Figure 20: Increase in transcriptional activity of HOXD4 IDR variants correlates with the number of small inert residues adjacent to aromatic residues.

and AroPERFECT YPWM(+), indicating that, under the conditions tested, the YPWM motif in the HOXD4 IDR does not contribute to its activity.

The HOXD4 IDR contains a predicted minimal activation domain within its N-terminal region. Controlling for the potential creation of another minimal activation domain that could lead to enhanced transcriptional activity, I tiled the HOXD4 wild type and AroPERFECT IDRs into tiles of 40 amino acids in length, with an overlap of 20 amino acids. This approach allowed me to map the predicted activation domain in the wild type sequence to a region spanning amino acids 20 to 60 (Figure 21a). Comparing the activities of the individual tiles with the activity of the respective full-length sequence, I observed that the wild type tile containing the activation domain exhibited stronger activity than the complete IDR sequence. Additionally, there was a noticeable decrease in the activity of the corresponding AroPERFECT tile. No tile of the AroPERFECT sequence matched or exceeded the activity of the full-length IDR. Consequently, the creation of a minimal activation domain in the AroPERFECT sequence can be ruled out for being the reason for enhanced transcriptional activity.

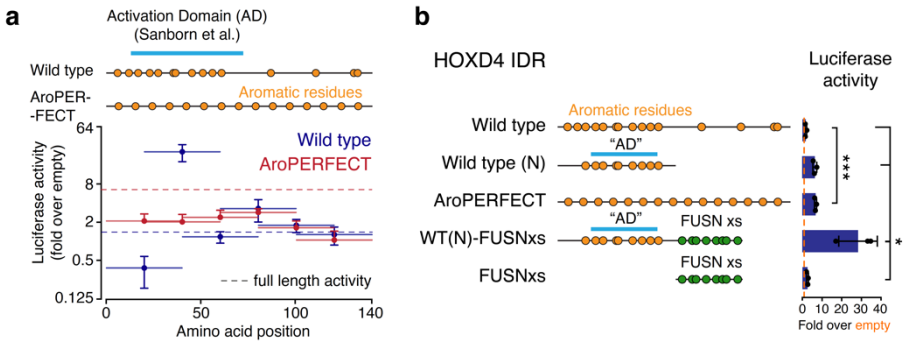


Figure 21: A minimal activation domain in the HOXD4 IDR synergizes with the PLD-specific sequence feature of aromatic dispersion. *(a) Reporter assays of 40 amino acid HOXD4 IDR tiles. Luciferase values displayed as percentages of the activity measured using an empty vector. Data displayed as mean ± SD, from three biological replicates. (b) Reporter assays of HOXD4 IDR complementation experiments with the PLD-containing protein FUS. Data displayed as mean ± SD, from two biological replicates.*

Notably, the activation domain of HOXD4 is located precisely within the N-terminal region, which displays a high degree of aromatic dispersion. In an IDR complementation experiment, I substituted the C-terminal part of the HOXD4 IDR, starting from amino acid position 60, with the N-terminal portion of the PLD of human FUS (FUSNxs) (Figure 21b). The N-terminal part of the HOXD4 IDR alone (Wild type (N)) demonstrated marginally higher transcriptional activity compared to the full-length sequence, and FUSNxs alone displayed minimal transcriptional activity. Remarkably, the WT(N)-FUSNxs fusion strongly activated the luciferase reporter, resulting in a ~15-fold increase in signal intensity.

In summary, the dispersion of aromatic residues seems to be a critical non-linear sequence feature that regulates transcriptional activity of TF IDRs within the constraints imposed by

other sequence features present in IDRs, such as SLiMs, minimal activation domains, and amino acid composition of the spacer sequences.

Optimized dispersion of aromatic residues enhances liquid-like features of HOXD4 condensates *in vitro*

As the dispersion of aromatic residues has been discovered in the context of a PLD-specific sequence feature that regulates biophysical properties of condensates formed by PLD-containing proteins such as the RNA-binding protein HNRNPA1, I was interested in examining the impact of optimized aromatic dispersion on condensate formation of TF IDRs [69]. Consequently, I focused on HOXD4 as my extensively studied example. I purified recombinant mEGFP-tagged HOXD4 wild type, AroLITE, AroPLUS, AroPLUS LITE, AroPLUS patched, AroPLUS patched LITE and AroPERFECT IDRs and conducted *in vitro* droplet formation assays to evaluate the propensity of these fusion proteins to form homotypic condensates in the presence of a crowding agent (PEG-8000) (Figure 22a).

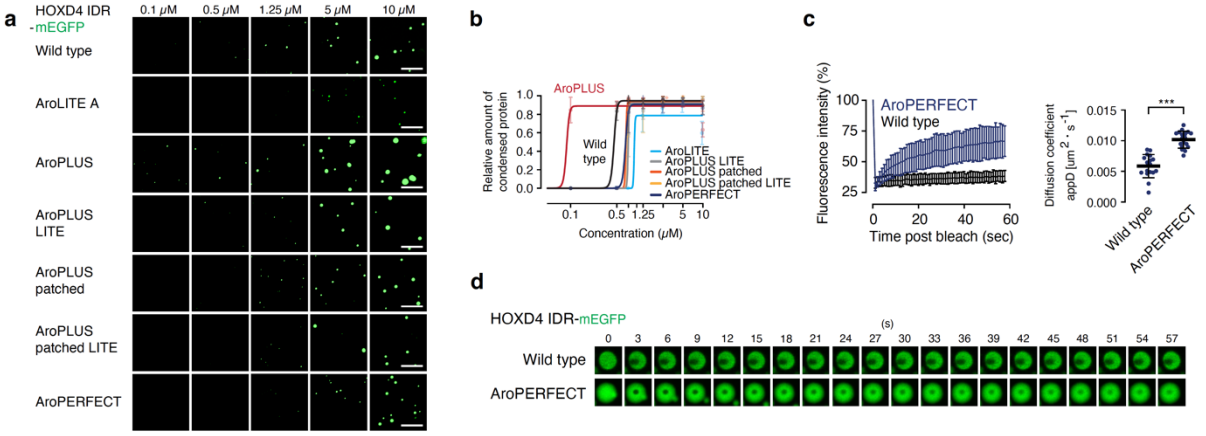


Figure 22: Optimized aromatic dispersion enhances liquid-like features in HOXD4 condensates *in vitro*. *(a) Representative images of droplet formation of purified, recombinant HOXD4 wild type and mutant IDR-mEGFP proteins. Scale bar: 5 µm. Data generated by Yaotian Zhang (b) Quantification of the droplet assays. Data displayed as mean ± SD. N = 15 images from 3 replicates. The curve was generated as a non-linear regression to a sigmoidal curve function. Data generated by Yaotian Zhang (c) (left) Fluorescence intensity of HOXD4 wild type and HOXD4 AroPERFECT in vitro droplets before, during and after photobleaching. Data displayed as mean ± SD. N = 20 images from two replicates. (right) Calculation of the apparent diffusion coefficient. P-values are from two-sided unpaired t-tests. ***: P<0.001. (d) Representative images of HOXD4 in vitro droplets before, during and after photobleaching.*

While the AroLITE mutant displayed reduced droplet formation compared to the wild type, the AroPLUS mutant, containing seven additional aromatic residues, formed droplets at lower concentrations and thus reached c_{sat} earlier than the wild type droplets or droplets of any of the control sequences AroPLUS LITE, AroPLUS patched and AroPLUS patched LITE which exhibited droplet formation comparable to the wild type (Figure 22b). Notably, despite the AroPERFECT mutant sequence having the highest transcriptional activity, it showed no marked difference in droplet formation compared to wild type or AroPLUS control sequences. Prior research on aromatic dispersion in PLD-containing proteins suggests that patterning of aromatic amino acids alters the biophysical properties of condensates formed. To investigate this, I performed fluorescence recovery after photobleaching (FRAP) experiments on

homotypic *in vitro* condensates of HOXD4 wild type and AroPERFECT (Figure 22c-d). HOXD4 wild type droplets did not recover fluorescence signal after bleaching a circular area within a droplet, indicating gel-like properties. In contrast, fluorescence signal in AroPERFECT IDR droplets exhibited faster recovery over time, suggesting more liquid-like properties. Recovery rates were quantified by calculating an apparent diffusion coefficient as previously reported [39]. Recovery rates revealed a significantly higher diffusion rate in droplets formed by HOXD4 AroPERFECT as compared to the wild type ($P < 0.001$, t-test). Replication of these results with the HOXC4 IDR, which has a similar aromatic residue patterning to the HOXD4 IDR but a chemically distinct spacer composition, confirmed the findings (Figure 23). The HOXC4 AroPERFECT IDR demonstrated a significant increase in transcriptional activity compared to the wild type in reporter assays (Figure 23a). Both IDRs formed homotypic droplets *in* vitro in a concentration dependent manner with similar c_{sat} (Figure 23b). FRAP experiments revealed a more liquid-like state of droplets formed by HOXC4 AroPERFECT IDR compared to the wild type, with a significant increase in the apparent diffusion coefficient ($P < 0.001$, t-test) (Figure 23d-e).

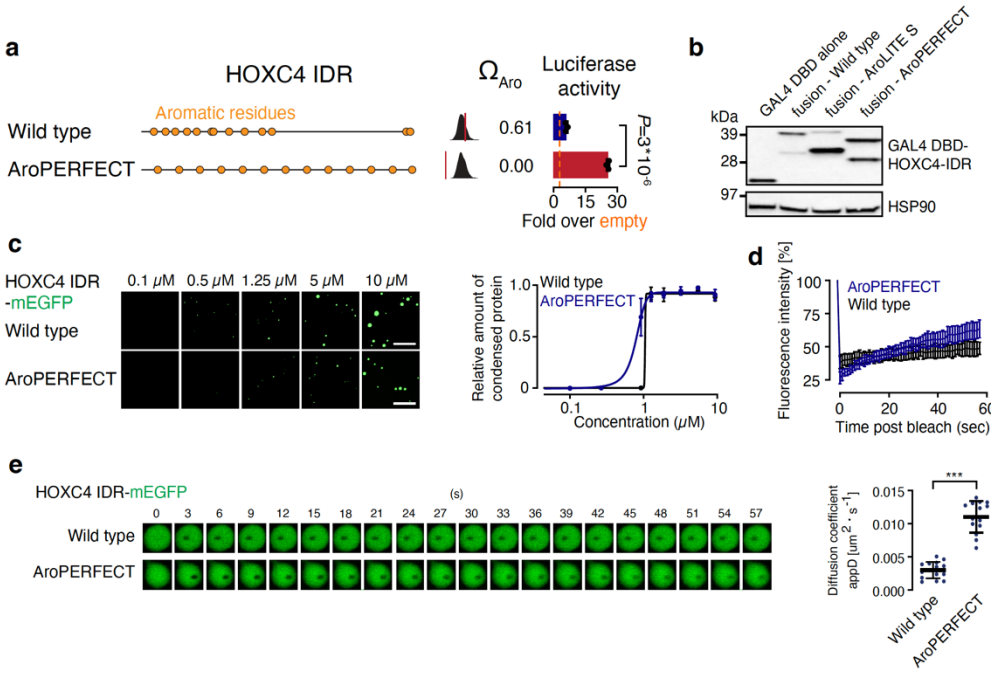


Figure 23: Optimized aromatic dispersion enhances transcriptional activity and liquid-like features of the HOXC4 IDR. *(a) (left) Schematic models of the HOXC4 IDR variants tested with Ω_{Aro} scores. (right) Reporter assays of HOXC4 IDR versions. Luciferase values displayed as percentages of the activity measured using an empty vector. Data displayed as mean ± SD, from three biological replicates. P-values from two-sided unpaired t-test. (b) Western blot of overexpressed GAL4-fusion proteins. HSP90 was used as a housekeeping control. (c) (left) Representative images of droplet formation of purified, recombinant HOXC4 wild type and mutant IDR-mEGFP proteins. Scale bar: 5 µm.(right) Quantification of the droplet assays. Data displayed as mean ± SD. N = 15 images from 3 replicates. The curve was generated as a non-linear regression to a sigmoidal curve function. Data generated by Yaotian Zhang (d) (top) Fluorescence intensity of HOXC4 wild type and HOXC4 AroPERFECT in vitro droplets before, during and after photobleaching. Data displayed as mean ± SD. N = 20 images from two replicates. (bottom) Calculation of the apparent diffusion coefficient. P-values are from two-sided unpaired t-tests. ***: P<0.001. (e) Representative images of HOXC4 in vitro droplets before, during and after photobleaching.*

In summary, this suggest that the dispersion of aromatic residues within TF IDRs is a non-linear sequence feature, bridging the two critical aspects of transcription factor function: condensation and transactivation. This link is supported by the correlation between the enhanced liquid-like properties of TF IDR condensates *in vitro* and increased transcriptional activity in a cell-based reporter system.

Optimized aromatic dispersion enhances TF function in cells

To investigate the impact of optimized dispersion in the HOXD4 IDR in live cells at endogenous expression levels, I generated HOXD4 knock-in cell lines in human HAP1 cells using the Cas9 system. The human HAP1 cell line, derived from a myeloid leukemia line (KBM7), was selected for its near-haploid genotype, rapid doubling time, and the low-level expression of HOXD4. I targeted the endogenous locus of HOXD4 using two single-guide RNAs that showed sequence complementarity to the start and stop codons of the canonical HOXD4 coding sequence. Successful targeting involved the excision of the entire open reading frame of HOXD4, and subsequent replacement with a synthetic sequence encoding a fusion protein comprising the full-length HOXD4 wild type, AroPERFECT or AroPLUS protein fused to a mEGFP-tag, separated by a GS-linker (Figure 24a). In parallel, I created a HOXD4 knock-out cell line using the same single-guide RNAs as those in the knock-in experiments.

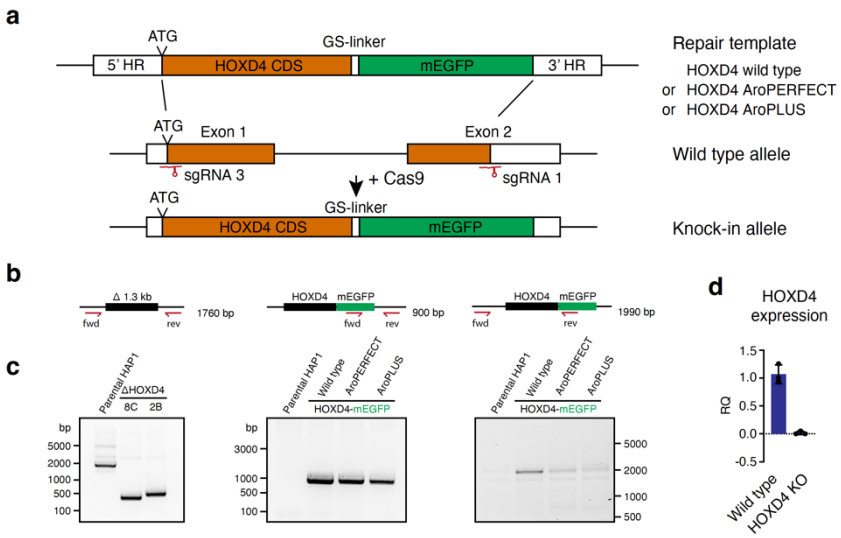


Figure 24: Integration strategy for endogenous HOXD4 knock-in cell lines. *(a) Scheme of mEGFP knock-in strategy at the HOXD4 locus. (b) Scheme of the PCR genotyping strategy of the HAP1 cell lines. (c) PCR genotyping of HAP1 cell lines. (d) HOXD4 gene expression levels quantified as RQ value in HAP1 wild type and HAP1 HOXD4 knock-out cells by quantitative real-time PCR. Data represented as mean ± SD from three technical replicates.*

I confirmed successful targeting and integration of the knock-in sequences by genotyping using PCR primers flanking the HOXD4 locus, in combination with primers targeting the mEGFP-tag within the introduced sequence (Figure 24b). A shift in PCR product size indicated efficient targeting and excision of the HOXD4 coding sequence in two independent clonal HOXD4 knockout lines (Figure 24c). Furthermore, qPCR analysis of HOXD4 expression levels, using isolated RNA from the HOXD4 knockout 2B clone converted to cDNA, revealed a complete loss of HOXD4 expression in the respective cell line compared to the expression level in parental HAP1 cells (Figure 24d). PCR amplification of DNA including the knock-in

sequences demonstrated successful integration of the transgene alleles into the genome of HAP1 cells (Figure 24c).

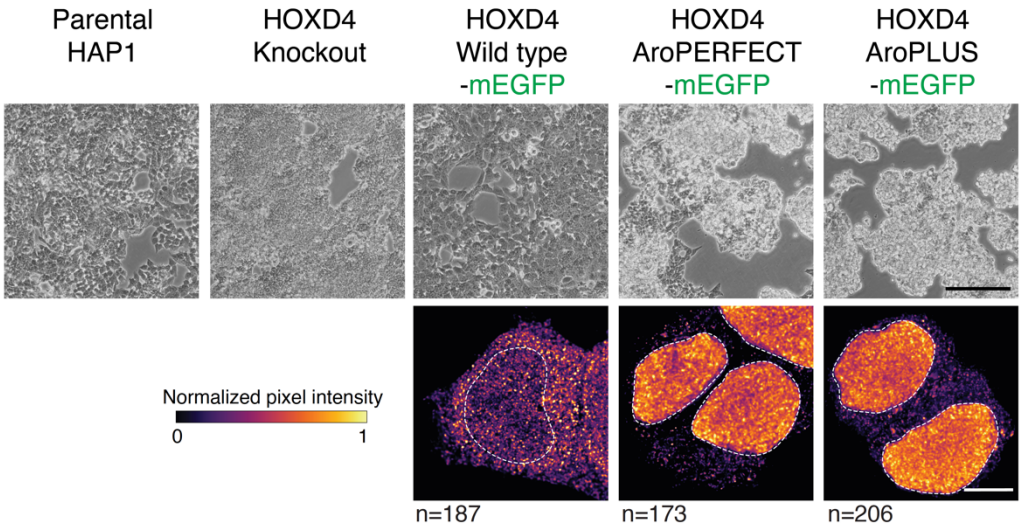


Figure 25: Maximized aromatic dispersion in the HOXD4 IDR changes the morphology of HAP1 cells. *(top) Differential interference contrast (DIC) microscopy of the indicated cell lines. Scale bar is 0.4 mm. (bottom) Representative fluorescence microscopy images of cell nuclei. The fusion proteins were visualized using anti-GFP immunofluorescence in fixed cells. The normalized signal intensity was calculated by dividing standard deviation of mEGFP signal of each nucleus by the corresponding mean mEGFP signal. Scale bar is 10 µm. Images acquired by Hannah Wieler.*

To my surprise, the introduction of HOXD4 AroPERFECT-mEGFP and AroPLUS-mEGFP mutants into the genome of HAP1 cells altered their morphology (Figure 25). While the parental HAP1 cell line and HAP1 HOXD4 wild type-mEGFP cells grew as a uniform monolayer with distinct cell-cell boundaries, the HAP1 HOXD4 AroPERFECT-mEGFP and AroPLUS-mEGFP cells exhibited a colony-forming, three-dimensional growth pattern. These cells generally appeared smaller with increased granularity. Furthermore, AroPERFECT-mEGFP and AroPLUS-mEGFP cells detached more readily from the cell culture dish compared to parental HAP1 or HAP1 HOXD4 wild type-mEGFP cells. Although HOXD4 knockout cells maintained a properly attached monolayer, their size shifted in a manner similar to that observed in the AroPERFECT-mEGFP and AroPLUS-mEGFP cells. Confocal microscopy following immunofluorescence staining of HOXD4 wild type-mEGFP, AroPERFECT-mEGFP and AroPLUS-mEGFP cells with a GFP antibody revealed modest enrichment of HOXD4 wild type-mEGFP in the nuclei, whereas the expression levels of HOXD4 AroPERFECT-mEGFP and AroPLUS-mEGFP were significantly higher than in the wild type, resulting in pronounced nuclear enrichment and the formation of intense nuclear clusters (Figure 25).

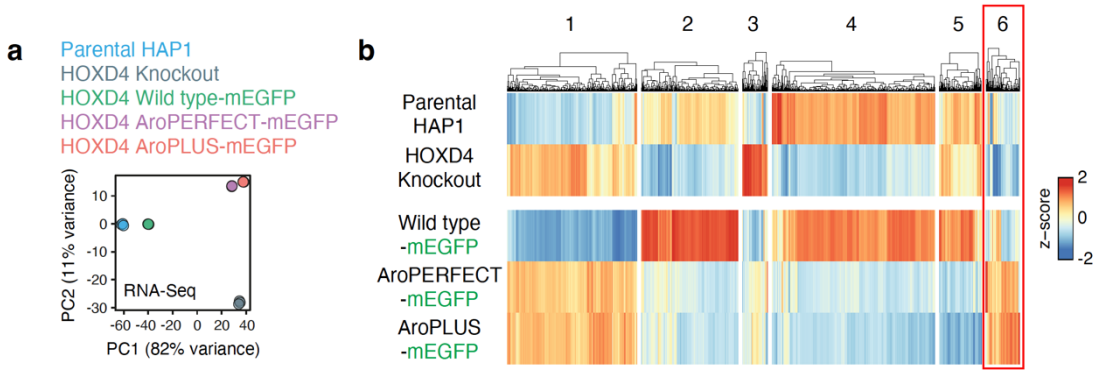


Figure 26: Maximized aromatic dispersion in the HOXD4 IDR is associated with altered gene specificity. *(a) Principal component (PC) analysis of the RNA-Seq expression profiles of HAP1 wild type, HOXD4 knockout and indicated knock-in cell lines. (b) Heatmap analysis of RNA-Seq data in the five cell lines. Expression values are represented by scaling and centering VST transformed read count normalized values (z-score). K-means clustering was used to define the clusters. Data was analyzed and plotted by Alexandre Magalhães.*

Bulk RNA-sequencing revealed transcriptional differences between parental HAP1 cells and HOXD4 knockout and knock-in cell lines. A principal component analysis (PCA) of the approximately 16,000 quantified transcripts showed global similarities between the parental HAP1 cell line and HOXD4 wild type-mEGFP cells. The profiles of HOXD4 AroPERFECT-mEGFP and AroPLUS-mEGFP cells, however, were distinct from those of the parental and HOXD4 wild type-mEGFP cells, as well as the HOXD4 knockout cells (Figure 26a). When examining differentially expressed genes between cell lines, 1,133 HOXD4 target genes were identified based on differential expression between the conditions of HOXD4 wild type-mEGFP and HOXD4 knockout. Most of these target genes were downregulated in the HOXD4 knockout, but some also exhibited increased expression levels (cluster 1 & 3) (Figure 26b). A majority (76%) of the differentially expressed genes in the AroPERFECT-mEGFP and AroPLUS-mEGFP cells were deregulated similarly to the HOXD4 knockout condition. Additionally, 396 genes were uniquely upregulated in AroPERFECT-mEGFP and AroPLUS-mEGFP cells but downregulated in the knockout. These genes (cluster 6) were defined as HOXD4 gain-of-function targets and their expression levels were validated using Western Blot (Figure 27b). One of the cluster 6 genes was HOXD4 itself, which reportedly autoregulates its own expression through a positive feedback loop [175–177]. Other differentially expressed genes of cluster 6 were ARHGAP4, IFI16, ESX1 and GATA6, with ESX1 and GATA6 also showing increased protein levels in the HOXD4 knockout (Figure 27b). To investigate the impact of optimized aromatic dispersion in the HOXD4 IDR on target gene expression at equal expression levels, I randomly integrated doxycycline inducible versions of HOXD4 wild type, AroPERFECT and AroPLUS into the genome of human HAP1 cells using the PiggyBac transposon system.

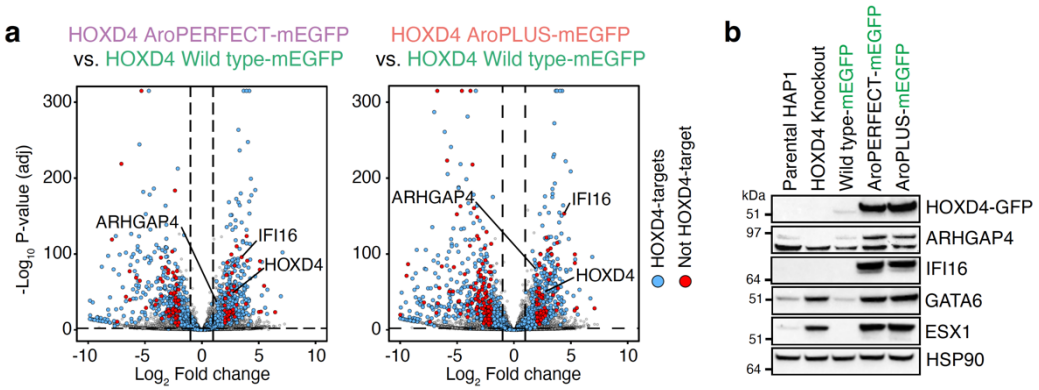


Figure 27: Differential expression between HAP1 HOXD4 knock-in lines. *(a) Differential expression analysis of HAP1 HOXD4 AroPERFECT-mEGFP and HOXD4 AroPLUS-mEGFP cells versus HOXD4 wild type-mEGFP cells. HOXD4 target genes are highlighted in blue, non-HOXD4 target genes are highlighted in red. P-values from Benjamini-Hochberg method. Data was analyzed and plotted by Alexandre Magalhães. (b) Western blot analysis of HOXD4-mEGFP, ARHGAP4, IFI16, GATA6 and ESX1 in the indicated cell lines. HOXD4-mEGFP proteins were probed with an anti-GFP anybody. HSP90 is shown as loading control.*

In this system, the expression of HOXD4 can be induced by addition of doxycycline to the cell culture medium, thus rendering it independent of any autoregulation of HOXD4 itself. Employing this approach, I successfully reproduced the morphological changes observed in the HOXD4 AroPERFECT-mEGFP and AroPLUS-mEGFP knock-in cell lines. Following 14 days of continuous doxycycline induction, the AroPERFECT and AroPLUS cells began to display a more colony-forming, three-dimensional growth pattern, while cells expressing HOXD4 wild type and the uninduced HAP1 cells, maintained growth as an adherent monolayer (Figure 28).

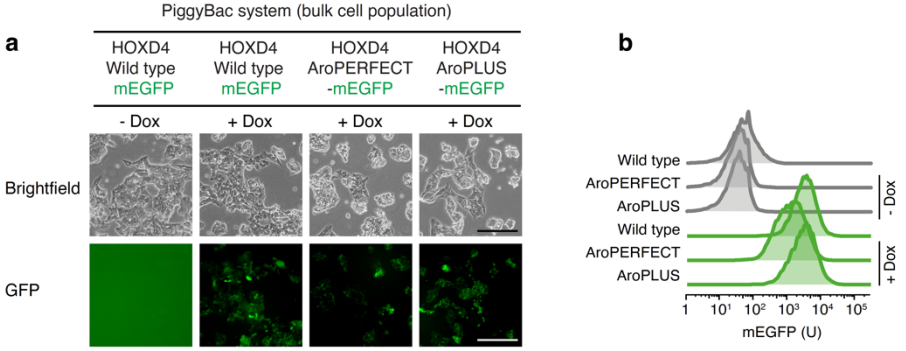


Figure 28: Overexpression of HOXD4 wild type, AroPERFECT and AroPLUS at comparable levels confirms morphological knock-in phenotypes. *(a) (top) Differential interference contrast microscopy of the indicated cell lines. Scale bar is 0.4 mm. (bottom) Fluorescence microscopy images. Cells were imaged 14 days after constant doxycycline induction. (b) Flow cytometry analysis of mEGFP expression in HAP1 HOXD4-mEGFP PiggyBac cell lines after 14 days of Dox induction. A representative quantification is shown. Data normalized to mode.*

I confirmed comparable expression levels of HOXD4 in the PiggyBac cell lines using flow cytometry, 48 hours after induction of transgene expression. Although I achieved comparable expression levels of HOXD4 wild type-mEGFP and HOXD4 AroPLUS-mEGFP, I could not generate a cell line with equivalent expression levels of HOXD4 AroPERFECT-mEGFP. Therefore, I utilized a cell line expressing approximately 50% less HOXD4 protein for subsequent analyses. Confocal microscopy of the HOXD4 expressing cells, after

immunofluorescence staining with an mEGFP specific antibody, revealed nuclear localization of all three HOXD4 variants. At these high expression levels, all three variants formed nuclear clusters (Figure 29a). However, image analysis of cells expressing equal levels of HOXD4 wild type, AroPERFECT or AroPLUS indicated a significant increase in signal granularity for HOXD4 AroPERFECT and AroPLUS compared to the wild type (Figure 29b).

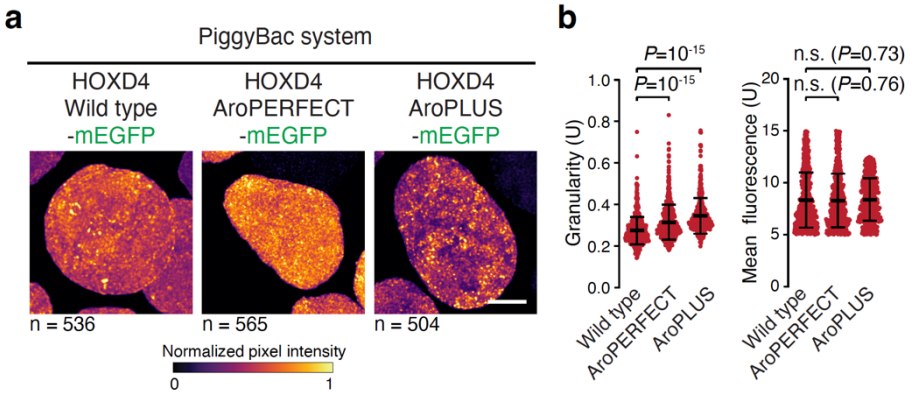


Figure 29: HAP1 cells overexpressing HOXD4 versions with maximized aromatic dispersion show increased signal granularity. *(a) Representative images of HAP1 HOXD4 wild type-mEGFP, HOXD4 AroPERFECT-mEGFP and HOXD4 AroPLUS-mEGFP nuclei after 24h of HOXD4 expression. The fusion proteins were visualized using mEGFP fluorescence in fixed cells. The normalized signal intensity was calculated by dividing standard deviation of mEGFP signal of each nucleus by the corresponding mean mEGFP signal. Number of individual nuclei per condition is displayed. Scale bar is 5 µm. (b) Granularity scores of nuclei, with corresponding mean nuclear mEGFP intensities. Data are displayed as mean ± SD from two biological replicates. P-values are from two-sided unpaired t-tests. Images were acquired and data was analyzed by Hannah Wieler.*

To validate the findings from our RNA sequencing experiment using the HAP1 HOXD4 knock-in lines, I conducted qPCR on cDNA from HOXD4 PiggyBac lines to check for differential expression of cluster 6 genes: ARHGAP4, IFI16, ESX1 and GATA6. As I was unsuccessful in generating a HOXD4 AroPERFECT-mEGFP cell line that expressed the transgene at levels comparable to the HOXD4 wild type-mEGFP and AroPLUS-mEGFP, I excluded this clone from further analysis. Following 14 days of continuous doxycycline induction, all four candidate genes exhibited increased expression levels in the HOXD4 AroPLUS-mEGFP condition than in the HOXD4 wild type-mEGFP, thus replicating the differential expression phenotype observed in the knock-in cell lines (Figure 30).

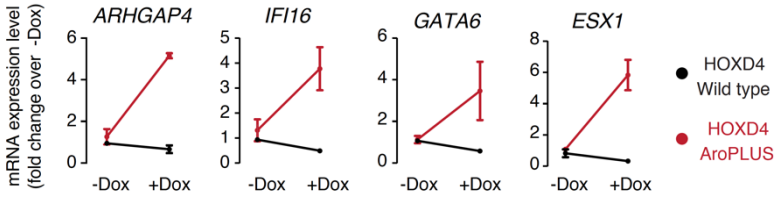


Figure 30: Differential expression of HOXD4 target and non-target genes upon HOXD4 overexpression. *Gene expression levels quantified as fold change in HAP1 PiggyBac clones, measured by quantitative real-time PCR after 14 days of constant doxycycline induction. Data represented as mean ± SD from two biological replicates.*

Optimized aromatic dispersion facilitates RNAPII interaction

The C-terminal domain (CTD) of RNAPII is an integral component of the transcription pre-initiation complex and is necessary for transcriptional activation. To further dissect the link between optimized aromatic dispersion, transcriptional activity and condensation even further, I tested HOXD4 wild type and AroPERFECT IDRs for their ability to recruit the CTD of RNAPII into cellular condensates. I utilized a previously described cell based tethering system that makes use of an integrated array of LacO binding motifs in the genome of human U2OS osteosarcoma cells (Figure 31a) [178]. By fusing the HOXD4 IDRs of interest to a CFP-tagged LacI-DBD and co-transfecting it with the YFP-tagged RNAPII-CTD I was able to measure the efficiency of the LacI fusion protein tethered to the LacO array, thereby creating an artificial condensate, to recruit RNAPII-CTD into the condensed body (Figure 31b).

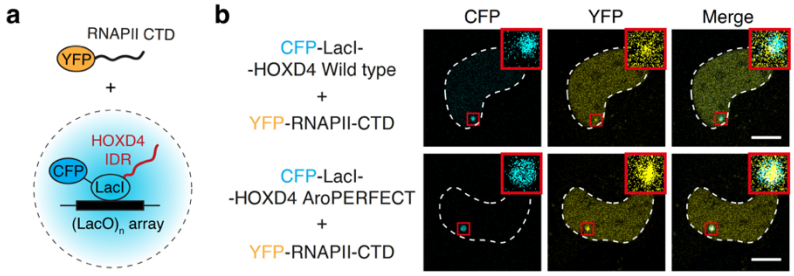


Figure 31: HOXD4 wild type and AroPERFECT recruit RNAPII-CTD into cellular condensates in U2OS cells. *(a) Schematic model of the condensate tethering system. (b) Fluorescence images of ectopically expressed YFP-RNAPII CTD in live U2OS cells co-transfected with the indicated CFP-LacI-HOXD4 IDR fusion constructs. Dashed line is the nuclear contour. Scale bar is 10 µm.*

Both, HOXD4 wild type and AroPERFECT IDR were capable of recruiting RNAPII-CTD into the condensate more efficient than the YFP-only control. However, the IDR of HOXD4 AroPERFECT recruited RNAPII-CTD significantly more efficient than the wild type (Figure 32a). Protein levels of the HOXD4 IDRs quantified by CFP signal intensities within the condensed region did not explain the altered enrichment (Figure 32b). This data suggests that the enhanced activity of HOXD4 AroPERFECT is associated with facilitated RNAPII interaction.

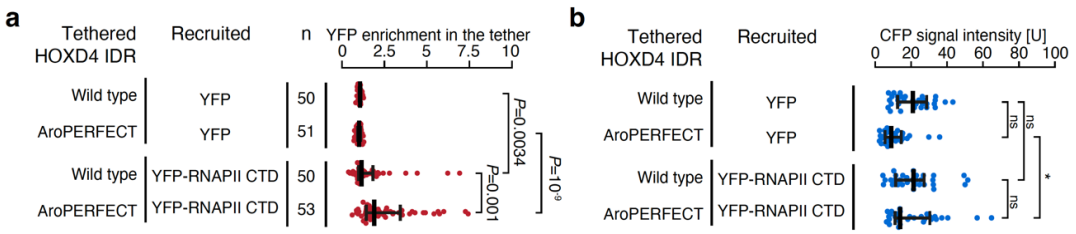


Figure 32: Optimized aromatic dispersion in the HOXD4 IDR facilitates RNAPII-CTD recruitment to cellular condensates. *(a) Quantification of the relative YFP signal intensity in the tether foci. Data displayed as mean ± SD from two biological replicates, P-values are from two-sided unpaired t-tests. (b) Control quantification of CFP fluorescence intensity in the tethered foci. Data represented as mean ± SD, N = number of cells shown, from two biological replicates. P-values are from 2-way ANOVA multiple comparisons tests. *:P<0.05*

Optimized aromatic dispersion as a generalizable approach to enhance TF-mediated direct reprogramming

The overexpression of transcription factors can reprogram cell identity. Direct reprogramming of one cell type to another has been described for various cell types using different transcription factors [179]. However, low reprogramming efficiencies are currently an obstacle to transitioning many such protocols from the *in vitro* stage to *in vivo* applications, such as cell replacement therapy. Consequently, there is a need for enhanced reprogramming efficiencies.

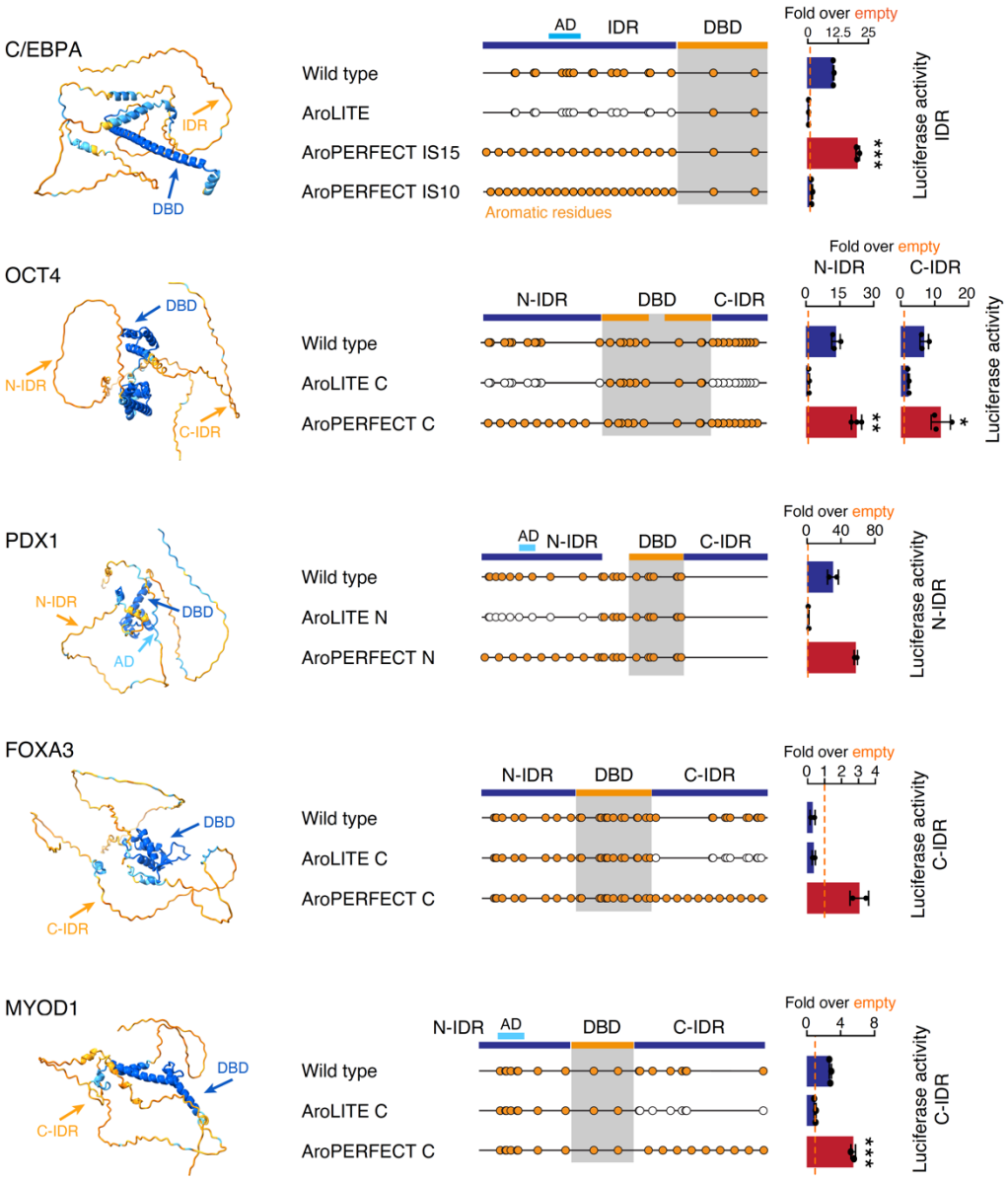


Figure 33: Optimizing aromatic dispersion enhances activity of multiple reprogramming TFs. *(left) Alphafold2 models of C/EBPA, OCT4, PDX1, FOXA3 and MYOD1. (center) Schematic models of C/EBPA, OCT4, PDX1, FOXA3 and MYOD1 wild type and mutant sequences. (right) Results of luciferase reporter assays. Luciferase values displayed as percentages of the activity measured using an empty vector. Data displayed as mean ± SD, from 2-3 biological replicates. P-values from two-sided unpaired t-test. *: P < 0.05, **: P < 0.01, ***: P < 0.001. Note that shown AroPERFECT IDRs have stronger transactivation capacity than their respective wild type sequences.*

I tested the effect of optimized aromatic dispersion on well-known reprogramming transcription factors. For several TFs, I designed AroPERFECT mutants as previously described and tested their transcriptional activity in reporter assays (Figure 33). Out of 14 transcription factors tested (excluding HOXD4 and HOXC4), I succeeded in generating mutants with enhanced transcriptional activity compared to their respective wild type sequences for five.

C/EBPα, a master transcription factor of the myeloid lineage, contains a N-terminal disordered region and a C-terminal bZIP-DNA binding domain. I generated a C/EBPα AroLITE mutant by substituting all aromatic residues in the IDR with alanine. Additionally, I designed two AroPERFECT mutants. The first, “AroPERFECT IS15”, with an interspacer length of 15 amino acids, for which I reintroduced all 16 endogenous aromatic residues of the C/EBPα wild type IDR in an optimally dispersed pattern. The second, “AroPERFECT IS10”, to which I added eight additional tyrosine residues to the optimally dispersed pattern to achieve a pattern with an interspacer length similar to that found in HOXD4 and HOXC4. In luciferase reporter assays, the wild type IDR exhibited moderate activity. While the AroLITE mutant, as expected, showed no activity, the AroPERFECT IS15 mutant demonstrated significantly stronger activity compared to the wild type. Notably, the AroPERFECT IS10 mutant failed to transactivate the luciferase reporter, exhibiting activity comparable to the AroLITE mutant. Additionally, I optimized the N- and C-terminal IDR of the pluripotency-associated transcription factor OCT4, as well as the N-terminal IDR of the pancreatic master TF PDX1, and the C-terminal IDRs of the liver-specific TF FOXA3 and the muscle-specific master TF MYOD1. In all cases, an AroLITE mutant of the respective sequence led to a loss of transcriptional activity, while the introduction of optimal aromatic dispersion enhanced activity (Figure 33). Again, differences in luciferase activity were not attributable to variations in protein expression levels, as confirmed by Western Blot (Figure *34*).

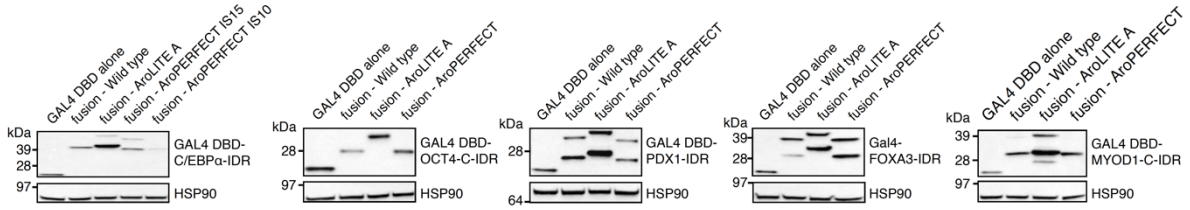


Figure 34: Expression levels of Gal4-fusion proteins do not explain transcriptional differences in reporter assays. *Western blot of GAL4-DBD and (left to right) GAL4-DBD-C/EBPA-IDR-, GAL4-DBD-OCT4-IDR-, GAL4-DBD-PDX1-IDR-, GAL4-DBD-FOXA3-IDR-, and GAL4-DBD-MYOD1-IDR- fusion proteins in HEK293T cells 24 hours after transfection using a GAL4-DBD specific antibody. HSP90 serves as a loading control. Wild type and AroPERFECT mutants are expressed at comparable levels.*

For the remaining factors tested, I could not measure enhanced transcriptional activity upon introduction of the optimized sequence feature (Figure 35).

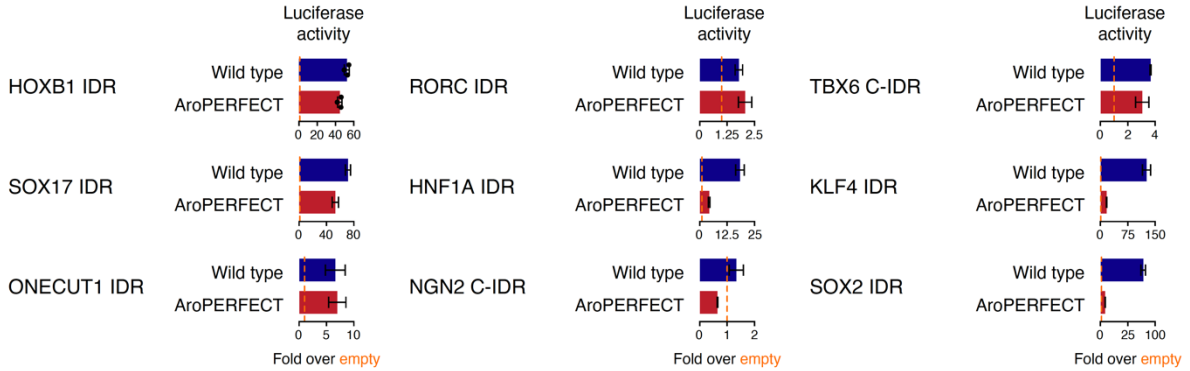


Figure 35: Non-successful sequence optimization in human transcription factors. *Results of luciferase reporter assays. Luciferase values displayed as percentages of the activity measured using an empty vector. Data displayed as mean ± SD, from 2-3 biological replicates. Note that shown AroPERFECT IDRs do not have stronger transactivation capacity than their respective wild type sequences.*

To investigate the effects of optimized aromatic dispersion on cellular reprogramming, I focused on C/EBPα for further functional studies. The C/EBPα AroPERFECT IS15 IDR displayed enhanced transcriptional activity, prompting additional functional assays to characterize the mutant. The C/EBPα wild type IDR contains a predicted and experimentally validated minimal activation domain [79]. To ascertain whether I had inadvertently created an additional minimal activation domain in the AroPERFECT IS15 IDR during the rearrangement of aromatic residues, I tiled the entire IDR sequence into overlapping fragments of 40 amino acids in length and assessed the activity of each fragment in reporter assays. I localized the minimal activation domain to amino acids 80 to 120 (Figure 36a).

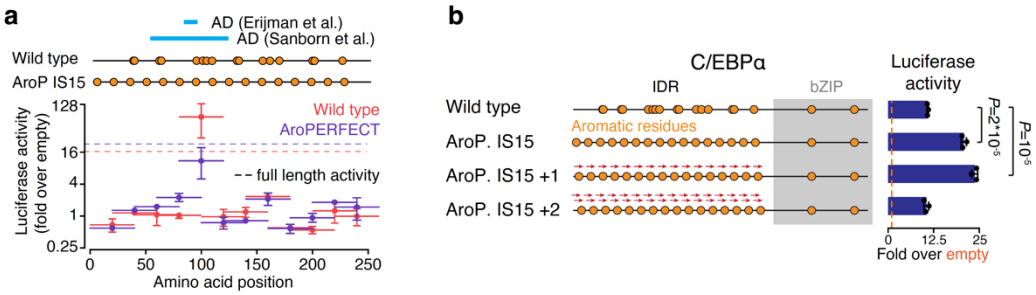


Figure 36: Optimized aromatic dispersion enhances C/EBPα transcriptional activity within the constraints of the backbone sequence. *(a) Results of a C/EBPα IDR tiling experiment using luciferase reporter assays. Data are displayed as mean ± SD from three biological replicates with two technical replicates each. The activities of the full-length IDRs are indicated with dashed horizontal lines. AD, activation domain. (b) (left) Schematic models of wild type and mutant C/EBPα proteins. The position of the bZIP DNA binding domain is highlighted with a grey box. (right) Results of C/EBPα luciferase reporter assays. Luciferase values were normalized against an internal Renilla control, and the values are displayed as percentages normalized to the activity measured using an empty vector. Data are displayed as mean ± SD from three biological replicates with three technical replicates each. P-values are from two-sided unpaired t-test.*

In both the wild type and AroPERFECT IS15 sequences, this tile demonstrated high activity; however, activity was markedly reduced in the AroPERFECT IS15 fragment compared to its respective wild type fragment. Additionally, no other fragment accounted for the enhanced

transcriptional activity of the full-length AroPERFECT IS15 IDR sequence, negating the hypothesis that a second minimal activation domain was created by chance. Moreover, by shifting the optimized aromatic pattern by one amino acid position toward the C-terminus, I recaptured the enhanced transactivation effect (Figure 36b). When I shifted the pattern by two amino acid positions toward the C-terminus, the activity observed was similar to that of the wild type IDR.

To assess the impact of this PLD-specific sequence feature on C/EBPα function, I deconstructed the function of the minimal activation domain and conducted IDR complementation assays. Initially, I evaluated the activity of the first 120 amino acids of the C/EBPα IDR, which included the experimentally validated minimal activation domain (referred to as Wild type (N)) (Figure 37). This segment exhibited a marginally higher activity than the wild type and the AroPERFECT IS15. Subsequently, I substituted the sequence C-terminal to the minimal activation domain with that of the AroPERFECT IS15, creating a fusion protein (WT(N)-IS15) while preserving the overall length and sequence composition. This chimera amplified transcriptional activity of the IDR, indicating an additive or synergistic influence of the minimal activation domain and the PLD-specific sequence feature. This observation prompted us to create a fusion of the wild type (N) with the N-terminal PLD of the human FUS protein (WT(N)-FUSN), which demonstrated comparable activity to the WT(N)-IS15 and increased activity relative to both, the wild type (N) and FUSN alone. Finally, I designed a fusion with a segment of the FUS PLD that matched the aromatic amino acid count of the C-terminal portion of the C/EBPα IDR (WT(N)-FUSNxs). This construct exhibited the highest activity among all the sequences tested, underscoring that transcriptional activity can be augmented by dispersion of aromatic residues within the structural framework of the C/EBPα IDR such as its minimal activation domains.

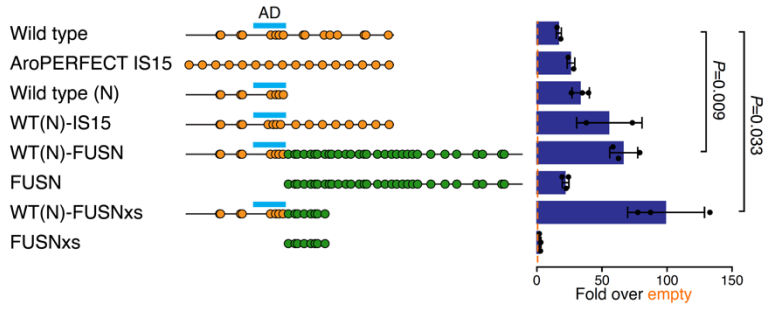


Figure 37: A minimal activation domain in the HOXD4 IDR synergizes with the optimized non-linear sequence feature. *Results of luciferase reporter assays using C/EBPα IDR constructs. Luciferase values were normalized against an internal Renilla control, and the values are displayed as percentages normalized to the activity measured using an empty vector. Data displayed as mean ± SD with N = 3 biological replicates. P-values are from a two-sided unpaired t-tests.*

Recombinantly expressed and purified C/EBPα IDRs formed *in vitro* droplets in a concentration dependent manner when exposed to crowding agent. The AroLITE IDR exhibited a reduced propensity to form droplets, indicated by a higher c_{sat} compared to the wild type or the AroPERFECT IS15 IDR; both of the latter showing similar behaviors. Conversely, the AroPERFECT IS10 mutant, containing 8 additional aromatic residues, underwent phase transition at lower concentrations, suggesting a lower c_{sat} than the wild type. This was quantified by calculating the relative amount of condensed protein under each condition at increasing concentrations (Figure 38a).

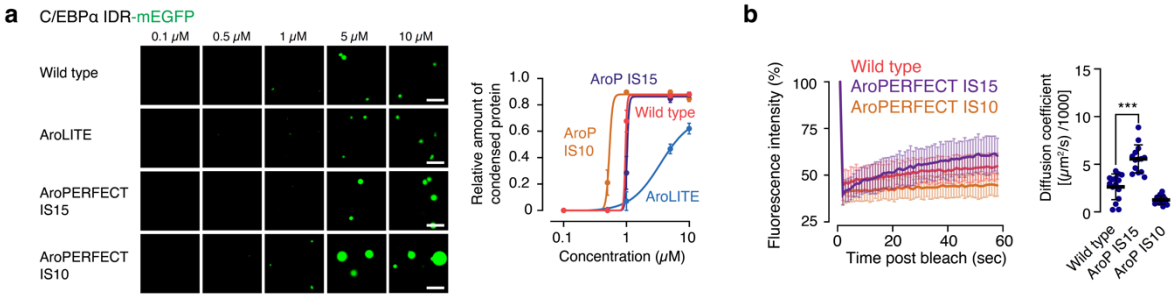


Figure 38: Optimized aromatic dispersion in the C/EBPα IDR enhances liquid-like features of condensates formed *in vitro. (a) (left) Representative images of droplet formation of purified C/EBPα IDR-mEGFP fusion proteins at the indicated concentrations in droplet formation buffer. Scale bar is 5 μm. (right) The relative amount of condensed protein per concentration quantified in the droplet formation assays. Data are displayed as mean ± SD. N = 10 images from 2 replicates. The curve was generated as a non-linear regression to a sigmoidal curve function. (b) (left) Fluorescence intensity of C/EBPα wild type, AroLITE and AroPERFECT IS15 IDR in in vitro droplets before, during and after photobleaching. Data are displayed as mean ± SD. N = 14 droplets from two replicates. (right) Calculation of the apparent diffusion coefficient. P-values are from two-sided unpaired t-tests. ***: P<0.001.*

FRAP experiments showed more efficient fluorescence signal recovery in droplets formed by the AroPERFECT IS15 compared to those formed by the wild type, resulting in a higher apparent diffusion coefficient (Figure 38b). Surprisingly, even with the AroPERFECT IS10 mutant's increased propensity to form droplets, the droplets did not regain fluorescence signal after photobleaching, thus exhibiting a more gel- or solid-like state, which corresponded to a low apparent diffusion coefficient.

Finally, I conducted LacO-LacI tethering experiments using the C/EBPα wild type and AroPERFECT IS15 IDRs in conjunction with the RNAPII-CTD. These experiments demonstrated that both IDRs could effectively recruit RNAPII-CTD to the LacO array. Quantitative analysis of YFP signal intensities revealed a significantly stronger recruitment of RNAPII-CTD into condensed areas formed by the AroPERFECT IS15 IDR compared to the wild type IDR (Figure 39). This increased recruitment was not attributable to differences in expression levels or to the efficiency with which the tether was recruited to the LacO array.

In summary, optimizing aromatic dispersion within transcription factors critical for cellular reprogramming appears to be a promising strategy for enhancing reprogramming efficiencies, at least *in vitro*.

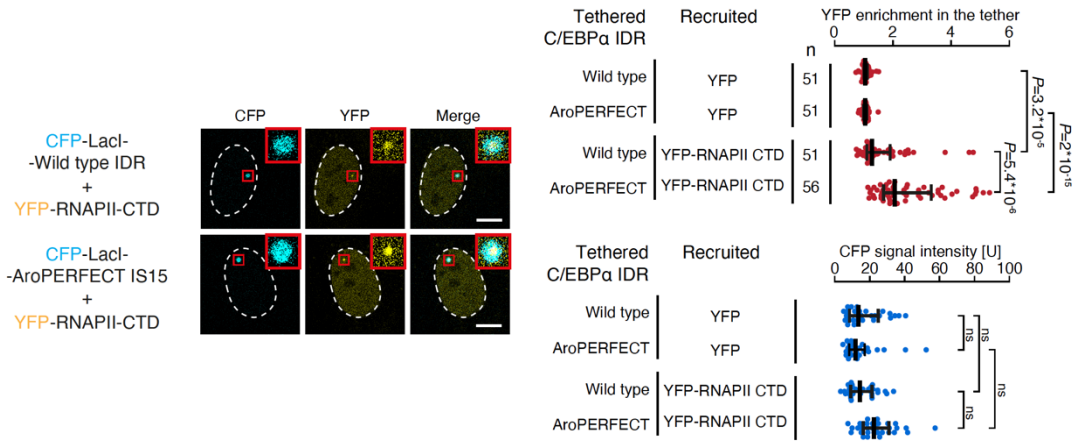


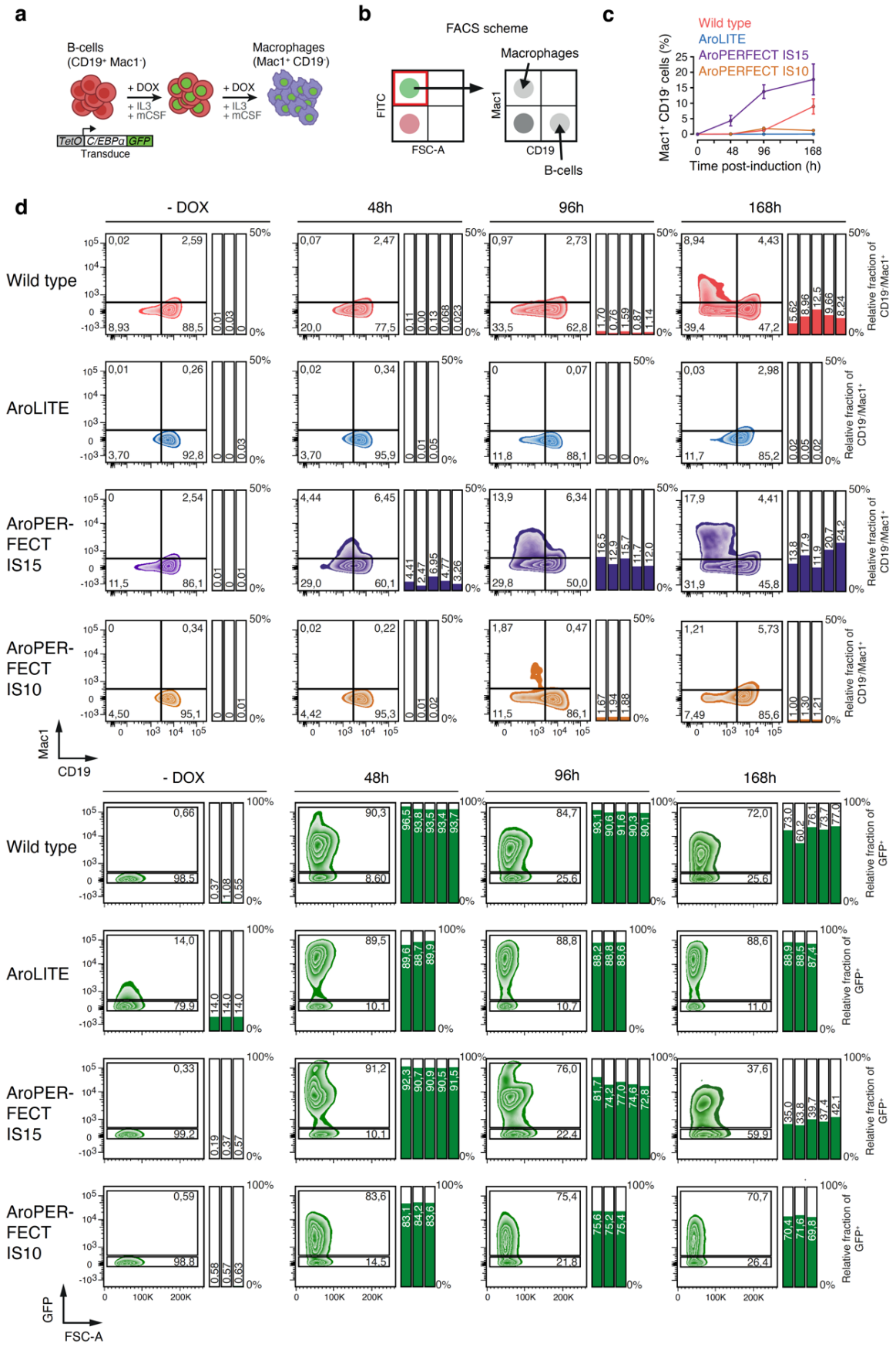
Figure 39: Optimized aromatic dispersion in the C/EBPα IDR facilitates RNAPII-CTD recruitment to cellular condensates. *(left) Fluorescence images of ectopically expressed YFP-RNAPII CTD in live U2OS cells co-transfected with the indicated CFP-LacI- C/EBPα IDR fusion constructs. Dashed line is the nuclear contour. Scale bar is 10 µm. (top right) Quantification of the relative YFP signal intensity in the tether foci. Data displayed as mean ± SD from two biological replicates, P-values are from two-sided unpaired t-tests. (bottom right) Control quantification of CFP fluorescence intensity in the tethered foci. Data represented as mean ± SD, N = number of cells shown, from two biological replicates. P-values are from 2-way ANOVA multiple comparisons tests.*

Optimized aromatic dispersion enhances C/EBPα-mediated macrophage reprogramming

To determine if the enhanced transcriptional activity of the C/EBPα AroPERFECT IS15 results in improved overall function, a previously characterized B-cell to macrophage reprogramming protocol was applied [161]. RCH-rtTA human leukemic B-cells were virally transduced to stably integrate GFP-tagged versions of full-length C/EBPα wild type, AroLITE, AroPERFECT IS15 and AroPERFECT IS10, all under the control of a doxycycline response element.

Following the induction of C/EBPα expression, the B-cells underwent reprogramming into terminally differentiated post-mitotic macrophages over the span of one week (Figure 40a). Cellular reprogramming efficiency was assessed using flow cytometry analysis to monitor the transduced cell population, measuring the expression of the B-cell-specific marker CD19 and the macrophage marker Mac1 (Figure 40b). A minority of cells in the AroLITE condition showed induction of the transgene in the absence of doxycycline, which was noted for further analysis. After 48 hours of continuous doxycycline induction, a majority of cells in all four conditions expressed the GFP-tagged C/EBPα transgene. This expression remained stable for up to one week of culture in the presence of doxycycline, with a slight decrease in efficiency due to the technical nature of the experimental setup (Figure 40d). Examining marker gene expression in B-cells expressing C/EBPα wild type, a sequential increase in $CD19^-/Mac1^-$ cells was observed over time. After one week, cells expressed Mac1, leading to a calculated $CD19^-/Mac1^+$ macrophage population of 8.94% (Figure 40c). Expression of the C/EBPα AroLITE mutant, failed to reprogram the B-cells, only resulting in a mild reduction of $CD19^+$ cells by 11.7% after one week.

Figure 40: (next page) Sequence optimization of the C/EBPα IDR enhances macrophage reprogramming. *(a) Schematic model of C/EBPα-mediated B-cell to macrophage reprogramming. (b) Scheme of FACS analysis strategy for quantification of macrophage reprogramming efficiency. (c) FACS quantification of GFP+ RCH-rtTA cells encoding C/EBP*α *overexpression cassettes. Cells were quantified for the level of the macrophage marker Mac1 and B-cell marker CD19, 48h, 96h and 168h after transgene induction. Data displayed as mean ± SD with N = 5 (Wild type, AroPERFECT IS15) or 3 (AroLITE, AroPERFECT IS10) biological replicate experiments. (d) Flow cytometry analysis of Mac1 and CD19 expression in differentiating RCH-rtTA cells after induction of C/EBPα constructs with doxycycline. The lines separating the quadrants of the plot indicate the gating strategy to categorize the population into Mac1/CD19 positive or negative. The barplots show the percentage of Mac1+ CD19- cells among the $mEGFP^+$ cell population in every replicate that corresponds to each condition. Concatenated data is shown (top sub-panel). Flow cytometry analysis of mEGFP expression in differentiating RCH-rtTA cells. Gates indicate cell populations considered as $mEGFP^+$ or $mEGFP^-$. The barplots on the right depict the percentage of the $mEGFP^+$ cell population in every replicate that correspond to each condition. Concatenated data is shown. Data was generated by Gregoire Stik.*


a
B-cells
(CD19+ Mac1-)
+ DOX
+ IL3
+ mCSF
+ DOX
+ IL3
+ mCSF
Macrophages
(Mac1+ CD19-)
TetO C/EBPa GFP
Transduce
b
FACS scheme
Macrophages
FITC
Mac1
FSC-A
CD19
B-cells
c
Wild type
AroLITE
AroPERFECT IS15
AroPERFECT IS10
Mac1+ CD19- cells (%)
Time post-induction (h)
d
- DOX
48h
96h
168h
Wild type
AroLITE
AroPER-FECT IS15
AroPER-FECT IS10
Relative fraction of CD19-/Mac1+
Mac1
CD19
Relative fraction of GFP+
GFP
FSC-A

In contrast, the cell population expressing the C/EBPα AroPERFECT IS15 mutant showed an increase in $CD19^-/Mac1^-$ cells after 48 hours of transgene expression. Moreover, an increase in the $CD19^-/Mac1^+$ cell population was observed, which initially comprised 4.44% of the total after 48 hours and became more substantial over time, eventually reaching 17.9% of the total cell population after one week of transgene expression. This effectively doubled the reprogramming efficiency compared to the C/EBPα wild type protein. Similar to the AroLITE mutant, the AroPERFECT IS10 failed to induce macrophage marker expression, only leading to a small $CD19^-/Mac1^-$ population of 7.49% of the total.

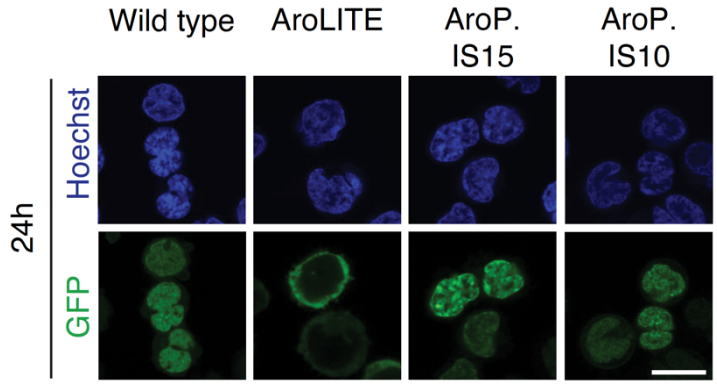


Figure 41: Sub-cellular localization of C/EBPα wild type and mutants in RCH-rtTA cells. *Fluorescence microscopy images of differentiating RCH-rtTA cells expressing GFP-tagged C/EBPα proteins are displayed 24h after transgene induction. Scale bar is 10 µm. Images were acquired by Gregoire Stik.*

Fluorescence microscopy confirmed the expression of the C/EBPα transgenes (Figure 41). Within 24 hours of induction, GFP signal in differentiating B-cells was detected in all four conditions. Notably, while the C/EBPα wild type, AroPERFECT IS15 and AroPERFECT IS10 localized to the cell nucleus, the AroLITE mutant was found predominantly in the cytoplasm and appeared to be enriched at the cellular membrane.

To gain insights on the consequences of optimized aromatic dispersion on the transcriptional programs driven by C/EBPα in differentiating B-cells, single-cell RNA-sequencing was performed on cells expressing C/EBPα wild type, AroPERFECT IS15 and AroPERFECT IS10 seven days after induction of transgene expression. I excluded C/EBPα AroLITE from this experiment due to its cytoplasmic localization and included the transcriptionally inert C/EBPα AroPERFECT IS10 as a negative control. Across the three samples, eight clusters containing cells with similar transcriptional profiles were identified (Figure 42a). The numbers of successfully transduced cells in each cluster were calculated based on GFP reads, omitting clusters 0 and 2 from further analysis as they contained almost no GFP-positive cells (Figure 42b).

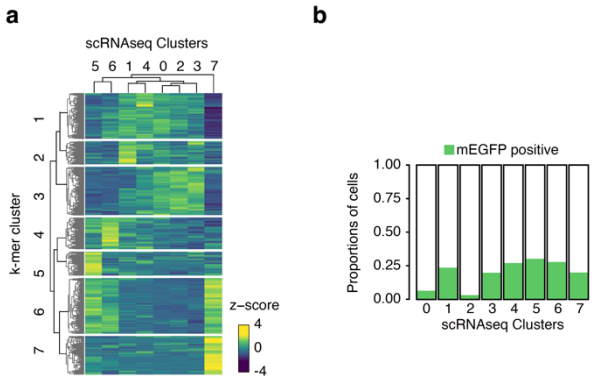


Figure 42: Characterization of single-cell RNA-Seq clusters *(a) Average expression for each cluster was normalized by vst and centered (z-score). K-means clustering was used to define the heatmap clusters. (b) Quantification of mEGFP-positive cells in the initial clusters. Cluster 0 and 2 contain virtually no mEGFP-positive cells, and were therefore removed from downstream analyses. Data was generated by Gregoire Stik and analyzed by Alexandre Magalhães.*

Cell types were assigned to clusters by referencing published marker gene sets from bulk RNA-seq data of cells subjected to the same protocol [161]. This set included known markers for B-cell and more defined marker genes for intermediate differentiation states like Early, Early-Intermediate, Differentiating-, and Late Macrophages.

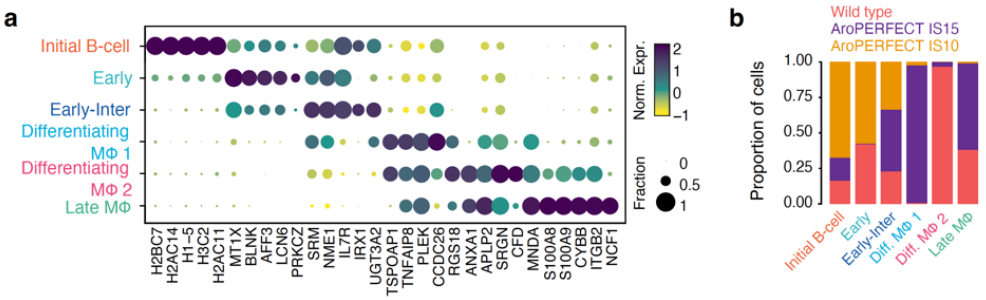


Figure 43: Cell-state annotations for single-cell RNA-Seq clusters using marker gene expression. *(a) Top 5 differentially expressed genes per cluster. These gene show specific expression signatures associated with each cluster and are used as differentiation stage markers. (b) Sample proportions for each cluster. Differentiating macrophage 1 is wild type-specific and Differentiating macrophage 2 is AroPERFECT IS15-specific. AroPERFECT IS10 cells are absent from the macrophage clusters. Data was generated by Gregoire Stik and analyzed by Alexandre Magalhães.*

Comparing the marker gene expression from these differentiation stages with expression profiles in our single-cell clusters allowed the identification of B-cells and macrophages and helped to assign the remaining clusters to intermediate differentiation stages as per the original annotation from the reference publication (Figure 43a). The contribution of each sample to each cluster was calculated and plotted the combined single-cell RNA cell state map as a UMAP (Figure 43b, Figure 44a). Pseudotime analysis of the differentiating cells indicated that initial B-cells progressed through early and intermediate differentiation states, ultimately expressing macrophage markers and culminating as terminally differentiated macrophages. This trajectory was validated by tracking normalized CD19 expression

alongside early and late macrophage markers such as CD14, PTPRC and ITGAM, the gene encoding Mac1 (Figure 44b).

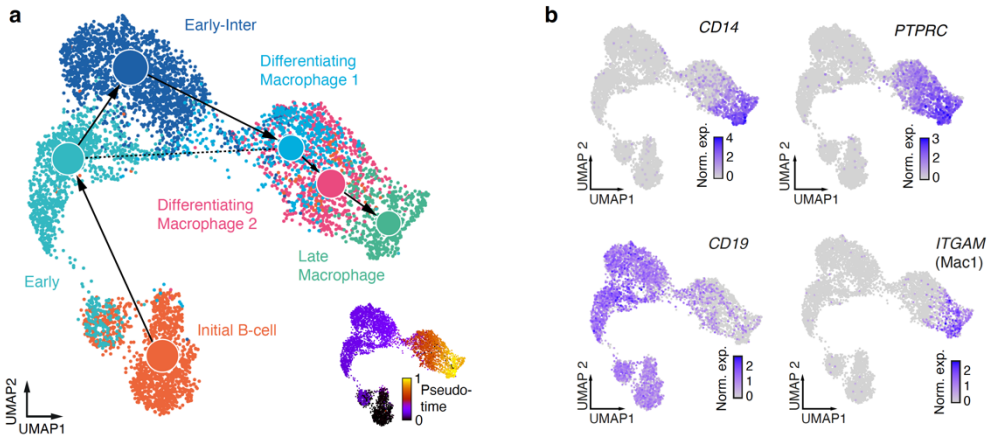


Figure 44: Graph-based clustering (UMAP) of the scRNA-Seq data of C/EBPα-mediated reprogramming. *(a) Clusters were annotated based on marker genes and previous work. Overlayed is the Partition-based graph abstraction (PAGA) showing cell trajectory based on dynamic modeling of RNA velocity. The inset is a pseudotime plot. (b) Combined UMAP colored CD14 and PTPRC, CD19 and ITGAM (MAC1) gene expression. These markers are associated with macrophage differentiation. Data was generated by Gregoire Stik and analyzed by Alexandre Magalhães.*

By considering GFP-positive cells in Differentiating Macrophage 1, Differentiating Macrophage 2 and Late Macrophage clusters as reprogrammed, the enhanced reprogramming efficiency observed in flow cytometry was corroborated at the RNA level; with an increase from about 50% in the wild type to about 80% in the AroPERFECT IS15 culture (Figure 45). As observed in previous experiments, the C/EBPα AroPERFECT did not reprogram B-cells into macrophages, with most cells remaining in Early, Early-Intermediate cell states or retaining B-cell identity (Figure 43b). Interestingly, cells in the Differentiating Macrophage 1 and 2 clusters originated predominantly from different cultures, indicating transcriptional differences between the populations, that eventually converged into a single Late Macrophage cluster.

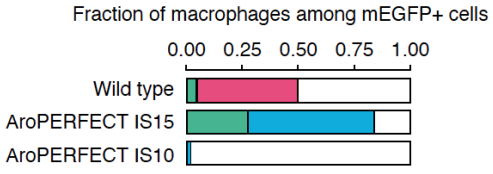


Figure 45: Sequence optimization of the C/EBPα IDR enhances macrophage reprogramming based on scRNA-Seq data. *Quantification of mEGFP-positive cells in macrophage clusters. Data was generated by Gregoire Stik and analyzed by Alexandre Magalhães.*

A differential expression analysis between cells expressing C/EBPα wild type and C/EBPα AroPERFECT IS15 in the Late Macrophage cluster revealed highly similar transcriptional profiles, with some macrophage associated marker genes differentially expressed (Figure 46a). The differential expression of two genes – CD66 (encoded by CEACAM8) and FCGR2A – was confirmed by flow cytometry analysis 48 hours after transgene induction (Figure 46b).

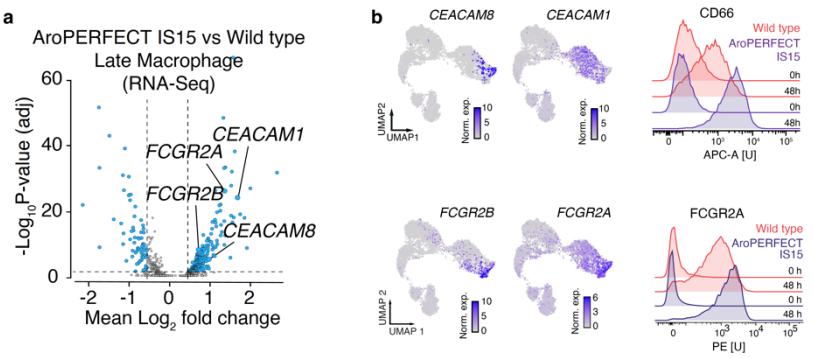


Figure 46: C/EBPα wild type and AroPERFECT IS15 exhibit differential marker gene expression. *(a) Volcano plot of differentially expressed genes in the Late Macrophage cluster for wild type vs. AroPERFECT IS15 samples. Differentially expressed target genes (Benjamini–Hochberg method, P<0.05) are highlighted in blue. Data was generated by Gregoire Stik and analyzed by Alexandre Magalhães. (b) (left) Combined UMAP colored on CEACAM8, CEACAM1, FCGR2B and FCGR2A expression. (right) Flow cytometry analysis of CD66 and FCGR2A expression in differentiating GFP⁺ RCH-rtTA cells 0h and 48h after induction of C/EBPα overexpression. Data normalized to mode. Data was generated by Gregoire Stik.*

Additionally, a set of genes uniquely expressed in cells of AroPERFECT IS15-expressing cells was identified (Figure 47), suggesting slightly altered gene specificity.

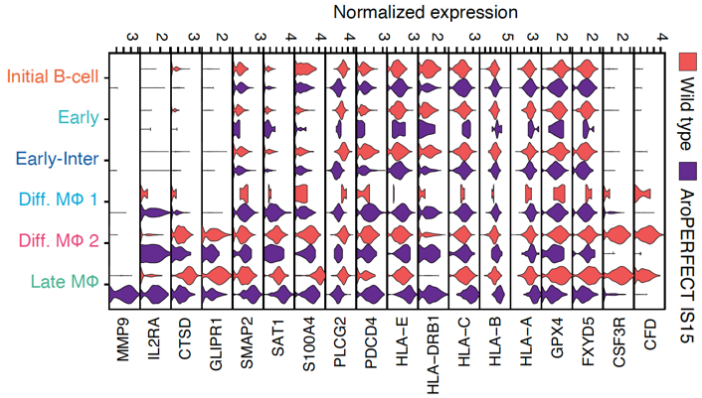


Figure 47: C/EBPα wild type and AroPERFECT IS15 show altered gene specificity. *Stacked violin plots for selected DEGs in the Late macrophage cluster between AroPERFECT IS15 and wild type. Most genes seem to be expressed in other clusters with the exceptions of MMP9. CSF3R and CFD seem to be wild type-specific while IL2RA is AroPERFECT IS15-specifically expressed. Data was generated by Gregoire Stik and analyzed by Alexandre Magalhães.*

In conclusion, I present evidence for the enhanced reprogramming efficiency of a C/EBPα mutant with optimized aromatic dispersion within its IDR. I quantified overall reprogramming efficiency by transcriptional changes in differentiating cells and macrophage marker gene expression at the protein level. Macrophages expressing C/EBPα wild type and AroPERFECT IS15 showed transcriptional similarities but exhibited signs of altered gene specificity of the overexpressed factors.

Optimized aromatic dispersion in C/EBPα enhances genomic binding and alters DNA-binding specificity

Optimized aromatic dispersion within the C/EBPα IDR led to improved cellular reprogramming efficiency *in vitro* and subtly modified gene specificity. To characterize the molecular basis underlying these observations, ChIP-Seq was performed on B-cells expressing C/EBPα wild type and C/EBPα AroPERFECT IS15, at 24 and 48 hours following the induction of transgene expression. Clonal lines expressing C/EBPα wild type (clone B12) and C/EBPα AroPERFECT IS15 (clone B8) were generated to control for equivalent expression levels, with GFP signal intensities quantified by flow cytometry to confirm comparability (Figure 48a). To ensure consistent binding affinities across both protein versions, I made use of the GFP-tag on both constructs and utilized an GFP antibody for immunoprecipitation experiments.

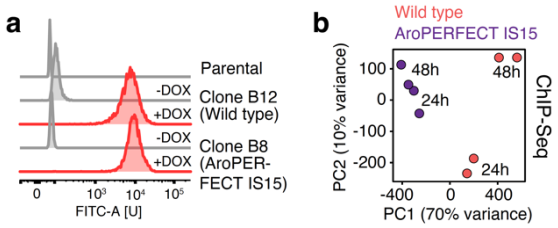


Figure 48: Global differences in genomic binding upon sequence optimization of the C/EBPα IDR. *(a) Flow cytometry analysis of GFP expression in RCH-rtTA clonal cell lines expressing GFP-tagged versions of C/EBPα. Data normalized to mode. (b) Principal component analysis of the ChIP-Seq peak profiles for wild type and AroPERFECT IS15 C/EBPα expressing cells 24h and 48h after induction of C/EBPα expression (PC1 vs. PC2). Data was generated by Gregoire Stik and analyzed by Alexandre Magalhães.*

Principal component analysis of the ChIP-Seq results demonstrated a correlation between replicates within each condition (Figure 48b). Additionally, clustering of AroPERFECT IS15-expressing samples was observed alongside a temporal trend delineated by principal component 2. Peak calling across samples identified an overlap in binding sites between C/EBPα wild type and C/EBPα AroPERFECT IS15. Generally, both factors bound to the same genomic locations; however, C/EBPα AroPERFECT IS15 exhibited, on average, higher read densities at these binding sites compared to C/EBPα wild type (Figure 49a). When assessing differential binding, the number of peaks with higher read densities in the AroPERFECT IS15 sample was two magnitudes larger than the number of peaks with higher read densities in the wild type. Several differentially bound peaks in the AroPERFECT IS15 sample were located within enhancer and promoter regions regulating macrophage marker gene expression. Notably, some of these genes were upregulated in the Late Macrophage cluster of the scRNA-Seq experiment in the AroPERFECT IS15 sample (Figure 49b).

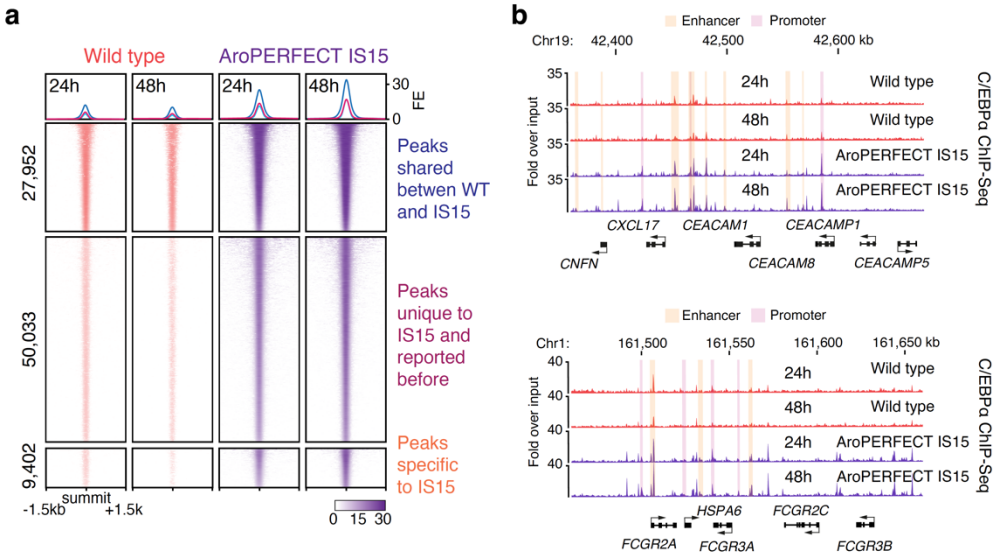


Figure 49: Sequence optimization in the C/EBP*α* IDR enhances genomic binding. *(a) Heatmap representation of ChIP-Seq read densities of C/EBPα wild type and AroPERFECT IS15 within a 1.5kb window around all shared C/EBPα peaks, and differentially enriched peaks in C/EBPα AroPERFECT IS15. "Peaks unique to IS15 and reported before" denote binding sites differentially enriched in IS15-binding that overlap C/EBPα peaks reported in previous literature. FE: enrichment. (b) C/EBPα AroPERFECT IS15 shows enhanced binding at the CEACAM gene cluster and at the FCGR2A locus. Displayed are genome browser tracks of ChIP-Seq data of C/EBPα wild type and AroPERFECT IS15 in RCH-rtTA cells, 24 and 48 hours after C/EBPα expression. Co-ordinates are hg38 genome assembly co-ordinates. Data was generated by Gregoire Stik and analyzed by Alexandre Magalhães.*

Additional efforts were made to characterize differences in genomic binding between C/EBPα wild type and AroPERFECT IS15. In the analysis of called peaks across samples, 27,952 peaks were discerned common to both the wild type and AroPERFECT IS15 samples. Moreover, 59,435 peaks were identified as unique to C/EBPα AroPERFECT IS15 in at least one condition tested. Out of these approximately 60,000 peaks, 50,033 were previously documented, as verified by cross-referencing with other C/EBPα ChIP-Seq datasets, which included a variety of cell types, not limited to B-cells or macrophages, but also encompassing e.g., liver. The remaining 9,402 peaks were specific to the C/EBPα AroPERFECT IS15 mutant (Figure 49a).

C/EBPα AroPERFECT IS15 exhibited altered gene specificity when compared to the wild type protein. This result was unanticipated as introduction of the optimally dispersed pattern of aromatic residues did not affect the C-terminal bZIP DNA-binding domain or its immediately adjacent amino acids. Association of differential genomic binding to motif composition at the bound regions revealed enrichment of canonical bZIP transcription factor motifs beneath the previously identified peak sets: "Peaks shared between WT and IS15", "Peaks unique to IS15 and reported before", and "Peaks specific to IS15". This analysis revealed a strong enrichment of the canonical CEBPA binding motif in the first two peak sets (Figure 50a). However, for the "Peaks specific to IS15", a decrease in CEBPA motif enrichment was noted, coupled with a

marginal increase in enrichment for the canonical binding motifs of bZIP TFs CEBPB and NFIL3 (Figure 50a), suggesting a specific loss of DNA binding specificity in C/EBPα AroPERFECT IS15.

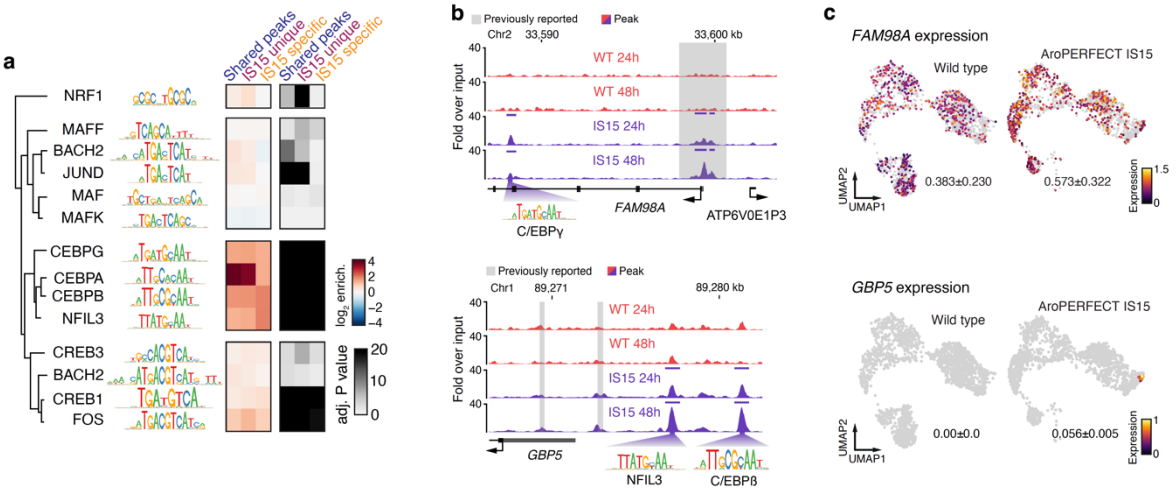


Figure 50: Sequence optimization of the C/EBP*α* IDR alters DNA-binding specificity. *(a) Enrichment scores of bZIP TF motifs, and adjusted P-values of enrichment at the three indicated peak sets. P-values from Benjamini-Hochberg method. (b) C/EBPα AroPERFECT IS15 shows enhanced binding at the FAM98A and GBP5 loci. Displayed are genome browser tracks of ChIP-Seq data of C/EBPα 24 and 48 hours after C/EBPα induction. Co-ordinates are hg38 genome assembly co-ordinates. (c) UMAPs colored on FAM98A and GBP5 expression. The numbers denote the mean expression +/-SD in the whole samples. Data was generated by Gregoire Stik and analyzed by Alexandre Magalhães.*

This finding was validated by examining *loci* specifically bound by C/EBPα AroPERFECT IS15 that were in close proximity to genes differentially expressed in our scRNA-Seq dataset. In the showcased examples, the specific association of C/EBPα AroPERFECT IS15 with non-canonical DNA binding motifs in proximity of FAM98A or GBP5 corresponded with AroPERFECT IS15-specific differential expression of these genes (Figure 50b-c).

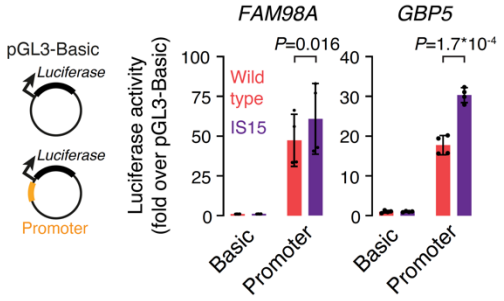


Figure 51: Locus reconstitution assays show increased transcriptional activity of C/EBP*α* AroPERFECT IS15. *Luciferase assays using the indicated reporter plasmids co-transfected with expression vectors encoding either wild type (red bars) or AroPERFECT IS15 (purple bars) C/EBPα. Luciferase values were normalized against an internal Renilla control, and the values are displayed as percentages of the activity measured using the 'basic' vector. Data are displayed as mean ± SD from four biological replicates. P-values are from two-sided unpaired t-tests.*

To affirm a regulatory relationship between genomic regions bound by C/EBPα AroPERFECT IS15 and the expression levels of target genes, luciferase reporter assays were conducted. Detected peak regions bound by C/EBPα AroPERFECT IS15 were cloned upstream of a

luciferase reporter gene to assess potential regulatory functions. Co-transfection of this reporter plasmid with plasmids encoding either C/EBPα wild type or C/EBPα AroPERFECT IS15 was followed by analysis of luciferase activity after 24 hours (Figure 51). This analysis consistently revealed significantly higher luciferase activity in cells expressing C/EBPα AroPERFECT IS15 compared to the wild type, emphasizing an enhanced interaction of the mutant protein with the cloned regulatory region.

In summary, the enhanced reprogramming efficiency of the C/EBPα AroPERFECT IS15 mutant can be attributed to its facilitated recruitment of the transcriptional regulator RNAPII-CTD, as well as its stronger association with canonical C/EBPα targets via the canonical CEBPA binding motif. Conversely, the C/EBPα AroPERFECT IS15 mutant displayed altered gene specificity by associating with non-canonical bZIP binding motifs, resulting in nonspecific gene activation.

Optimized aromatic dispersion in NGN2 enhances neuronal differentiation

Optimized aromatic dispersion within the C/EBPα IDR increased transcriptional activity and enhanced macrophage reprogramming efficiency. To test if this enhancement was solely dependent on increased transcriptional activity, I selected a reprogramming TF that did not exhibit increased transcriptional activity upon optimization of aromatic dispersion. Due to its thoroughly characterized function in neuronal differentiation, I chose the neuronal master transcription factor Neurogenin-2 (NGN2). NGN2 is a member of the basic helix-loop-helix bHLH family of transcription factors and is composed of a N-terminal disordered region, a central bHLH DBD, and a C-terminal IDR. The C-terminal IDR contains five aromatic amino acids with no significant aromatic dispersion. Since the N-terminal IDR only contained one aromatic residue, our mutagenesis approach was confined to the C-terminal region. I engineered both an AroLITE and an AroPERFECT mutant of the C-terminal IDR of NGN2 and assessed their transcriptional activity using luciferase reporter assays (Figure 52).

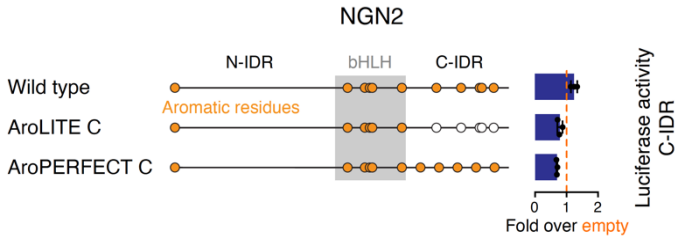


Figure 52: Sequence optimization of the NGN2 C-terminal IDR does not increase transcriptional activity. *(left) Schematic models of wild type and mutant NGN2 proteins. The position of the bHLH DNA binding domain is highlighted with a grey box. (right) Results of NGN2 luciferase reporter assays. Luciferase values were normalized against an internal Renilla control, and the values are displayed as percentages normalized to the activity measured using an empty vector (dashed orange line). Data are displayed as mean ± SD from three biological replicates.*

None of the IDRs demonstrated meaningful activity relative to a control vector, indicating that sequence optimization following our design principles did not affect transcriptional activity of the NGN2 C-IDR. I then investigated whether optimized aromatic dispersion in the C-IDR enhances liquid-like condensate properties *in vitro*. Recombinantly expressed NGN2 wild type C, AroLITE C and AroPERFECT C were purified and examined for condensate formation (Figure 53a).

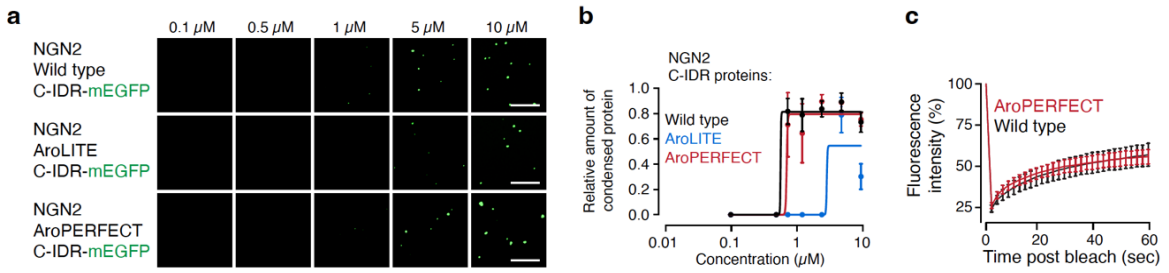


Figure 53: Optimized aromatic dispersion in the NGN2 C-IDR does not significantly alter liquid-like features of *in vitro* condensates. *(a) Representative images of droplet formation of purified NGN2 C-terminal IDR-mEGFP proteins. Scale bar: 5 µm. Data was generated by Yaotian Zhang (b) The relative amount of condensed protein per concentration quantified in the droplet formation assays. Data are displayed as mean ± SD. N = 10 images from 2 replicates. The curve was generated as a non-linear regression to a sigmoidal curve function. Data was analyzed by Yaotian Zhang (c) Fluorescence intensity of NGN2 wild type and AroPERFECT IDR in in vitro droplets before, during and after photobleaching. Data are displayed as mean ± SD. N = 20 droplets from two biological replicates.*

All three IDRs formed condensates in a concentration-dependent manner in the presence of crowding agent. While NGN2 wild type C and AroPERFECT C showed similar propensities to form condensates, as measured by the relative amount of condensed protein, the AroLITE mutant demonstrated reduced condensation, resulting in a higher c_{sat} (Figure 53b). FRAP analyses on *in vitro* condensates formed by NGN2 wild type C and AroPERFECT C showed comparable recovery rates, with the AroPERFECT C variant displaying a slight but not significant increase in recovery. Thus, augmenting aromatic dispersion in the C-terminal IDR of NGN2 neither increased transcriptional activity nor enhanced the liquid-like properties of *in vitro* condensates.

Functional tests on the NGN2 AroPERFECT IDR did not indicate any changes in the ability of NGN2 to direct cell fate. NGN2 can direct human induced pluripotent stem cells (hiPSCs) to differentiate into induced neurons (iNeurons) within 14 days when ectopically expressed [146]. To determine if sequence optimization in the NGN2 AroPERFECT mutant enhances neuronal differentiation, I stably integrated sequences encoding FLAG-tagged versions of NGN2 wild type, AroLITE and AroPERFECT into the genome of ZIP13K2 hiPSCs, using the PiggyBac system (Figure 54). These were monocistronically linked by a T2A self-cleavage sequence to mEGFP. Transgene expression was induced by addition of doxycycline to the culture medium. Upon induction of NGN2 expression, iPSCs differentiated into iNeurons over 14 days. To ensure equivalent expression levels of the three NGN2 variants, I created clonal lines for each variant and selected clones with similar expression levels, confirmed by anti-FLAG immunofluorescence signal intensity comparison (Figure 54b). In all conditions, FLAG-NGN2 signal was localized to the nucleus. To note, the expression levels of AroLITE and AroPERFECT lines were higher than the signal measured in NGN2 wild type expressing cells.

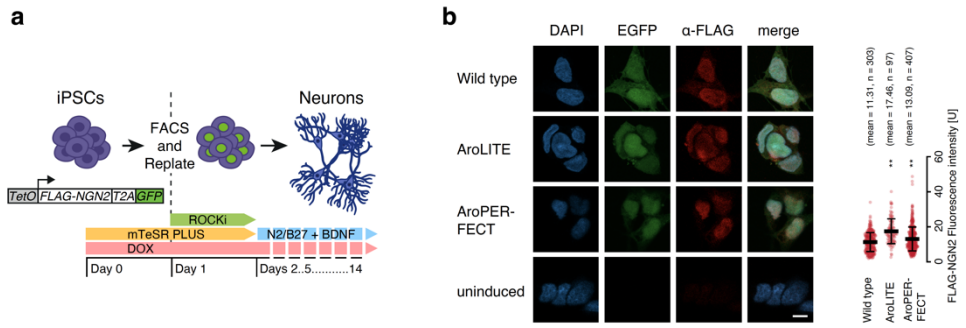


Figure 54: NGN2-mediated differentiation of hiPSCs into iNeurons at comparable expression levels. *(a) Schematic model of the NGN2-mediated hIPSC to neuron differentiation experiment. ROCKi: Rho-kinase inhibitor. (b) (left) Fluorescence microscopy images of differentiating ZIP13K2 cells expressing FLAG-tagged versions of NGN2 at 48h. NGN2-FLAG was visualized with an anti-FLAG antibody. GFP signal is the endogenous mEGFP fluorescence signal of mEGFP. Scale bar: 5 µm. (right) Quantification of FLAG-NGN2 signal. Data displayed as mean ± SD. N = number of cells from one biological replicate. P-values are from two-sided unpaired t-test. **: P < 0.01.*

To exclude uninduced cells from subsequent analysis, I performed FACS sorting 24 hours after induction. The cells were then re-plated at a defined cell density on Matrigel coated imaging slides.

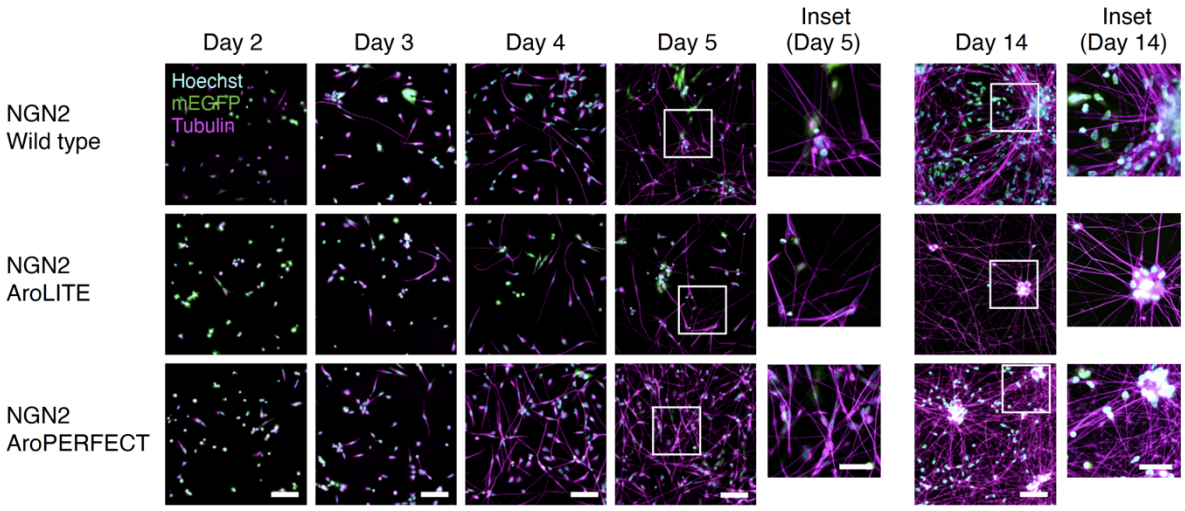


Figure 55: Live cell imaging of differentiating hiPSCs. *Representative fluorescence microscopy images of differentiating human iPSCs expressing the indicated NGN2 proteins. Tubulin staining is in magenta, nuclear counterstain (Hoechst) in blue, NGN2-T2A-mEGFP is green. Scale bar is 0.1 mm. Scale bar of insets is 0.05 mm.*

From days two to five following NGN2 induction, I assessed reprogramming efficiency through high-throughput microscopy, utilizing a tubulin-specific live cell dye to visualize morphological changes of the differentiating neurons and Hoechst as a nuclear counterstain (Figure 55). On day 14 after induction, I acquired additional images to assess the morphological state of the differentiating cells. By day three, neural projections appeared in all conditions, with neural morphology evident after five days and neural hub formation after 14 days of differentiation. Using Hoechst staining for segmentation and quantification of nuclei and a tubulin dye as a reference for the area covered by neurites, I noted a reduced differentiation efficiency in cells

expressing NGN2 AroLITE compared to wild type and a significant increase in differentiation efficiency in cells expressing the AroPERFECT mutant after 5 days (Figure 56). Due to the complexity of 3D neuronal hub formation, I excluded day 14 data from the quantification, as a reliable nuclei segmentation was not feasible.

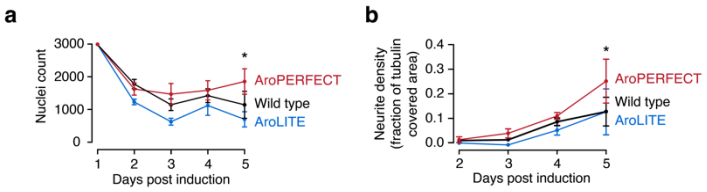


Figure 56: Sequence optimization of the NGN2 C-terminal IDR enhances neuronal differentiation. *(a) Quantification of the number of cells based on Hoechst nuclear staining in the NGN2-mediated differentiation experiments. Data are displayed as mean ± SD. N = 6 images from 2 independent experiments. P-value from a two-sided unpaired t-test. *: P < 0.05. (b) Quantification of neurite density based on tubulin staining in the NGN2-mediated differentiation experiments. Data are displayed as mean ± SD. N = 6 images from 2 independent experiments. P-value from a two-sided unpaired t-test. *: P < 0.05.*

To dissect the molecular mechanisms behind the enhanced reprogramming efficiency of NGN2 AroPERFECT, bulk RNA-Seq was performed on parental human iPSCs, NGN2 wild type, AroLITE and AroPERFECT iNeurons five days after induction of transgene expression.

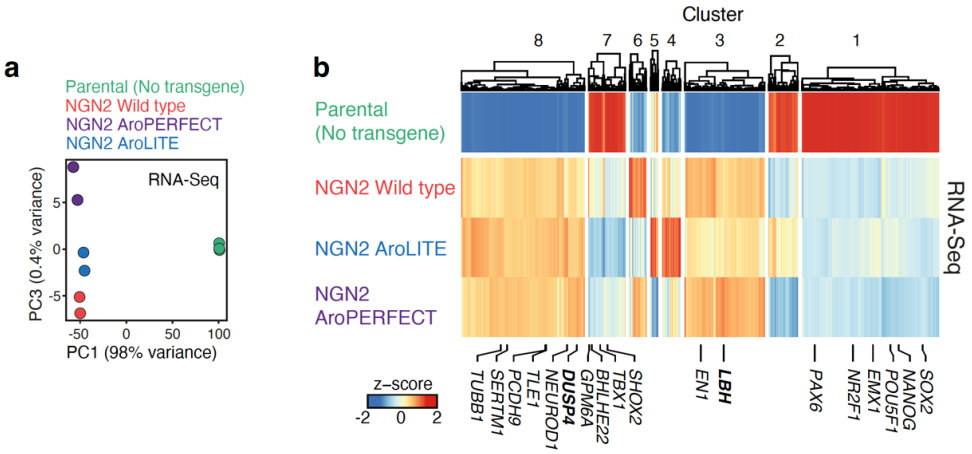


Figure 57: Global transcriptional changes in iNeurons generated by NGN2 wild type, AroLITE or AroPERFECT overexpression. *(a) Principal component analysis of the RNA-Seq expression profiles of parental ZIP13K2 hIPSCs, and hIPSCs expressing the indicated NGN2 transgenes. (b) Heatmap analysis of RNA-Seq data in the four cell lines. Genes were clustered using k-means clustering on expression values. Expression values are represented by scaling and centering VST transformed read count normalized values (z-score). Data was analyzed and plotted by Alexandre Magalhães.*

Principal component analysis of the RNA-Seq data revealed that the transcriptomes of the iNeurons were largely similar to each other but markedly distinct from the transcriptional profile of parental iPSCs (Figure 57a). A differential expression analysis followed by hierarchical clustering of the genes revealed eight principal clusters (Figure 57b). Cluster 1 encompassed genes downregulated in all iNeuron samples as opposed to the parental iPSCs, including pluripotency associated genes POU5F1, SOX2 and NANOG. Cluster 2 consisted of genes

downregulated in all iNeuron samples, albeit to a reduced degree in AroLITE and to a greater extend in AroPERFECT when compared to the wild type. Clusters 3 and 8 contained genes that were upregulated in all iNeuron samples relative to the parental cells, which included genes associated with neuronal differentiation such as SERTM1, NEUROD1 and DUSP4 suggesting a consistent shift towards a neuronal phenotype [180]. NGN2 target genes were determined based on differential expression between parental iPSCs and NGN2 wild type, as well as genomic binding identified in the ChIP-Seq analysis (Figure 60).

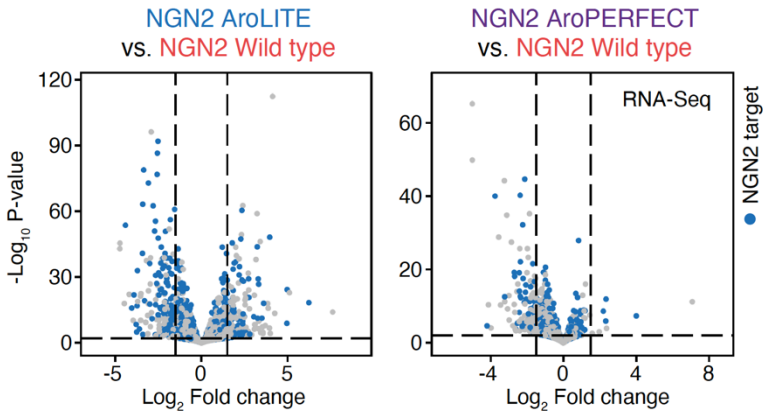


Figure 58: Differential gene expression in differentiating iNeurons. *Differential expression analysis of hIPSCs expressing the indicated transgenes. NGN2 target genes are highlighted in blue. P-values from Benjamini-Hochberg method. Data was analyzed and plotted by Alexandre Magalhães.*

The analysis showed highly similar transcriptional profiles for NGN2 AroLITE compared to NGN2 wild type and NGN2 AroPERFECT compared to NGN2 wild type. There was a predominant downregulation of the majority of differentially expressed genes in the AroLITE and AroPERFECT iNeurons, five days after transgene induction, encompassing direct NGN2 targets and genes differentially expressed in an indirect way (Figure 58).

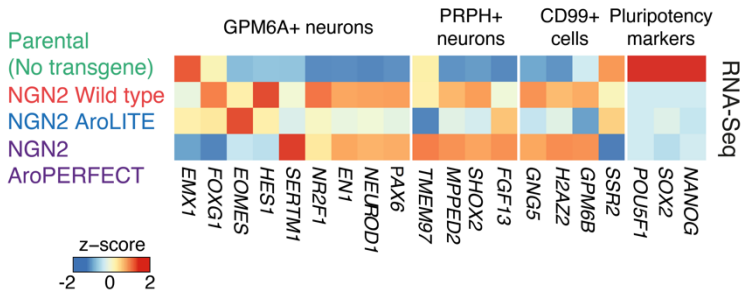


Figure 59: Neuronal marker gene expression in differentiating iNeurons. *Marker gene analysis from selected genes from single cell cluster markers in NGN2-induced neural differentiation experiments. Data was analyzed and plotted by Alexandre Magalhães.*

However, when assessing neuronal marker gene expression in iNeurons, differences were noted: HES1 was specifically upregulated in NGN2 wild type cells, EOMES was upregulated in NGN2 AroLITE cells, and SERTM1 predominantly expressed in NGN2 AroPERFECT cells (Figure 59).

To ascertain differences in genomic binding, ChIP-Seq experiments were conducted on differentiating iPSCs 24 and 48 hours after induction of NGN2 wild type, AroLITE or AroPERFECT.

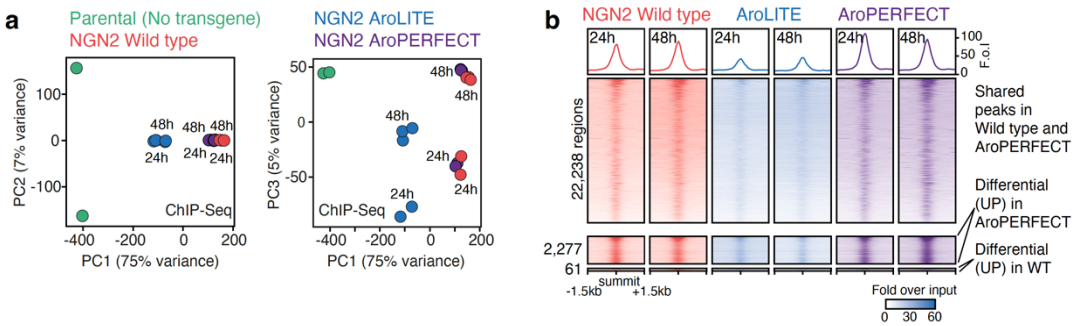


Figure 60: Sequence optimization of the C-terminal IDR of NGN2 enhances genomic binding *(a) Principal component analysis of the NGN2 ChIP-Seq peak profiles. (b) Heatmap representation of ChIP-Seq read densities of NGN2 wild type, AroLITE and AroPERFECT-expressing cells within a 1.5kb window around all shared NGN2 peaks (top), differentially enriched peaks in NGN2 AroPERFECT (center) and differentially enriched peaks in NGN2 wild type (bottom). F.o.I: fold over input. Data was analyzed and plotted by Alexandre Magalhães.*

Principal component analysis demonstrated clustering of NGN2 wild type and AroPERFECT samples for each time point, while AroLITE-expressing samples diverged from other conditions (Figure 60a). Peak calling across the six conditions revealed variability of peak densities at bound regions (Figure 60b). NGN2 AroPERFECT exhibited marginally higher read densities at these regions than the wild type after 24 hours of transgene expression. Read densities after 48 hours were comparable. Regardless of the time point, the AroLITE mutant displayed a marked reduction in read densities at *loci* bound by NGN2 wild type or AroPERFECT. 22,238 peaks were identified as equally bound by NGN2 wild type and AroPERFECT, 2,277 peaks were differentially bound by NGN2 AroPERFECT in at least one condition, and a mere 61 peaks were differentially bound by the wild type protein in at least one condition. When examining genomic *loci* of genes associated with neuronal differentiation, increased signal was detected at promoters in the AroPERFECT sample at 24 hours, compared to the wild type, with a notable absence of NGN2 AroLITE at these regions (Figure 61).

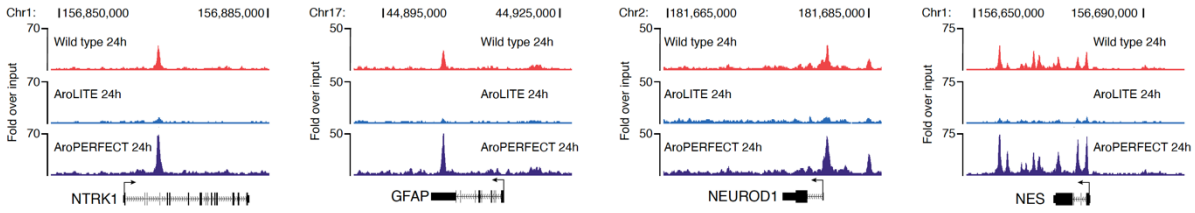


Figure 61: Increased read densities of NGN2 AroPERFECT at neuronal marker gene-associated *loci*. *(left to right) NGN2 differential binding at the NTRK1, GFAP, NEUROD1 and NES locus. Displayed are genome browser tracks of ChIP-Seq data after 24 hours of NGN2 expression. Co-ordinates are hg38 genome assembly co-ordinates.*

An analysis of bHLH TF binding motif enrichment beneath the identified peaks revealed enrichment of the canonical NGN2 motif in the shared peak set, alongside motifs for OLIG1, NEUROD1 and FERD3L (Figure 62). Co-operativity between the factors mentioned is expected as bHLH TFs require homo- or heterodimerization for effective DNA binding [181]. A decrease in enrichment for the canonical NGN2 motif was observed in both NGN2 wild type and AroPERFECT-specific peak sets.

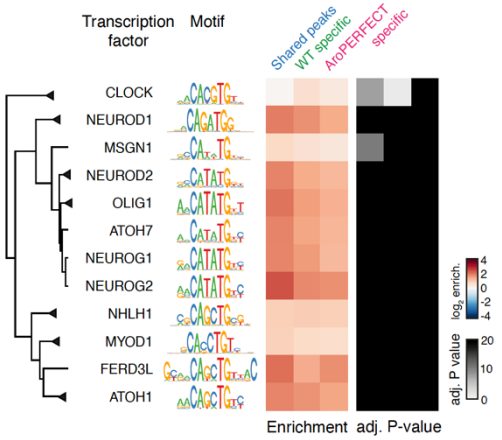


Figure 62: Sequence optimization of the NGN2 C-terminal IDR alters genomic DNA-binding specificity. *Enrichment scores of bHLH TF motifs, and adjusted P-values. P-values from Benjamini-Hochberg method. Data was analyzed and plotted by Alexandre Magalhães.*

Despite no detectable difference in transcriptional activity for NGN2 AroPERFECT C, sequence optimization led to more efficient neuronal reprogramming. The bulk RNA-seq data indicated highly similar transcriptional profiles among differentiated iNeurons with a pronounced neuronal signature and exhibited differences in early neuronal marker gene expression, which may indicate altered gene specificity of the overexpressed factors. ChIP-Seq data confirmed closely aligned binding profiles of NGN2 wild type and AroPERFECT with enriched read densities for AroPERFECT-expressing cells, but genomic depletion of NGN2 AroLITE.

Optimized aromatic dispersion enhances myotube differentiation

As a third example of the enhanced effect of reprogramming TFs with an optimized sequence feature within their IDRs, I introduced optimal aromatic dispersion into the disordered regions of the myogenic master TF MYOD1. MYOD1 is a member of the bHLH family of transcription factors and, like NGN2, consists of an N-terminal IDR, a central bHLH DNA-binding domain and a C-terminal IDR. The N-terminal IDR of MYOD1 contains a predicted minimal activation domain (Figure 63).

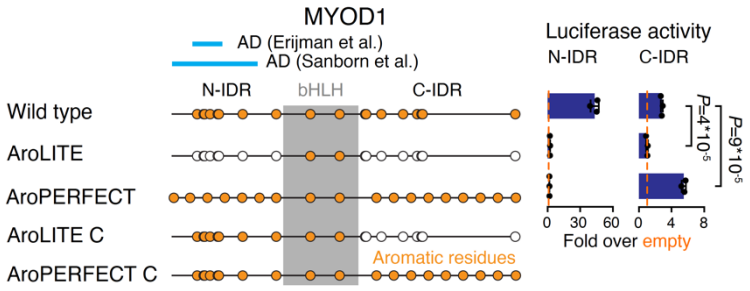


Figure 63: Sequence optimization of the C-terminal MYOD1 IDR enhances transcriptional activity. *(left) Schematic models of wild type and mutant MYOD1 proteins. The position of the bHLH DNA binding domain is highlighted with a grey box. AD, activation domain. (right) Results of luciferase reporter assays in C2C12 mouse myoblasts. Luciferase values were normalized against an internal Renilla control, and the values are displayed as percentages normalized to the activity measured using an empty vector. Data are displayed as mean ± SD from three biological replicates. P-values are from two-sided unpaired t-tests.*

To evaluate the transcriptional activity of both disordered regions, I tested the wild type sequences alongside the AroLITE and AroPERFECT mutants in luciferase reporter assays. The N-terminal wild type IDR, which harbors the minimal activation domain, displayed strong luciferase activity, as expected (Figure 63). When comparing the transcriptional activity of the wild type sequence to the AroLITE and AroPERFECT mutant sequences, both mutants showed a complete loss of activity, indicating the structural integrity of the minimal activation domain - comprising four aromatic residues - is critical for transcriptional activity of the N-terminal MYOD1 IDR.

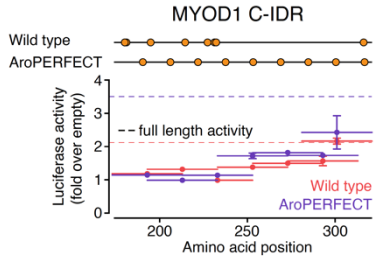


Figure 64: Enhanced transcriptional activity of MYOD1 AroPERFECT C was not driven by the creation of a minimal activation domain. *Results of MYOD1 C-IDR tiling experiment using luciferase reporter assays. Data are displayed as mean ± SD from three biological replicates with two technical replicates each. The activities of the full-length IDRs are indicated with dashed horizontal lines.*

I then tested the C-terminal disordered region and observed moderate transcriptional activity in the wild type IDR, which was lost in the AroLITE mutant. However, the AroPERFECT mutant, encoding optimized aromatic dispersion, demonstrated a significant increase in transcriptional activity. To verify that no additional minimal activation domains were inadvertently created in the AroPERFECT mutant, I tiled the C-terminal IDR of MYOD1 into fragments of 40 amino acids in length with 20 amino acids overlaps and tested their individual transcriptional activity in reporter assays (Figure 64). None of the tiles within the MYOD1 AroPERFECT C-IDR demonstrated transcriptional activity that could explain the increased activity of the complete IDR sequence.

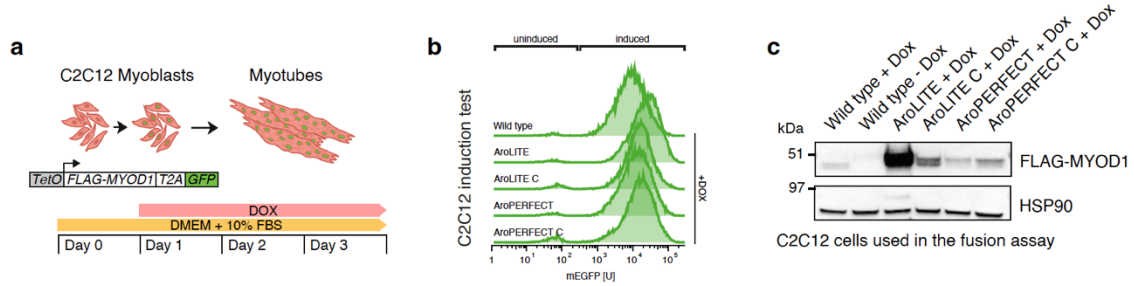


Figure 65: MYOD1-mediated differentiation of mouse myoblasts into myotubes at comparable expression levels. *(a) Schematic model of the MYOD1-mediated myotube differentiation experiment. (b) Flow cytometry analysis of mEGFP expression in mouse C2C12 PiggyBac cell lines 24 hours after doxycycline induction. (c) Western blot of FLAG-MYOD1 fusion proteins in differentiating C2C12 cells 24 hours after transgene induction. Wild type and AroPERFECT mutants are expressed at comparable levels. HSP90: loading control*

An increase in transcriptional activity led us to test whether a MYOD1 mutant with optimized dispersion in its C-terminal IDR could facilitate myotube differentiation. Thus, I integrated T2A-mEGFP-tagged versions of MYOD1 wild type, MYOD1 AroLITE, MYOD1 AroPERFECT, MYOD1 AroLITE C and MYOD1 AroPERFECT C into the genome of mouse C2C12 myoblast cells using the PiggyBac system (Figure 65a). I confirmed comparable expression of the integrated transgenes across conditions using flow cytometry analysis and Western blot, utilizing the FLAG-tag present in all MYOD1 variants (Figure 65b-c). Notably, MYOD1 AroLITE demonstrated higher expression levels in both the flow cytometry analysis and Western blot than all other conditions. Apart from that, expression levels of MYOD1 transgenes were comparable.

The forced expression of MYOD1 induces the differentiation of C2C12 myoblast cells into myotubes. To evaluate the reprogramming efficiency of MYOD1 AroPERFECT C, I induced the expression of MYOD1 wild type, AroLITE, AroLITE C, AroPERFECT and AroPERFECT C in C2C12 myoblasts by adding doxycycline to the growth medium. Over three days, I quantified myotube formation in differentiating C2C12 cells through high-throughput live-cell microscopy and subsequent image analysis. I used the monocistronically expressed mEGFP

in differentiating cells as a cytoplasmic marker and Hoechst as a nuclear counterstain to assess myotube formation. After three days, MYOD1 wild type, AroPERFECT C and AroLITE C formed myotube-like syncytia by fusing cell membranes with adjacent cells (Figure 66). MYOD1 AroLITE and AroPERFECT (not shown) did not form myotubes. This suggests that the sequence integrity of the N-terminal minimal activation domain of MYOD1 is vital for efficient myogenic differentiation. I quantified the proportion fused cells using the nuclear Hoechst stain as a marker to identify nuclei. Subsequently, I used the cytoplasmic mEGFP to delineate cell-boundaries and classified cells with more than two nuclei resulting from cell fusion as differentiated.

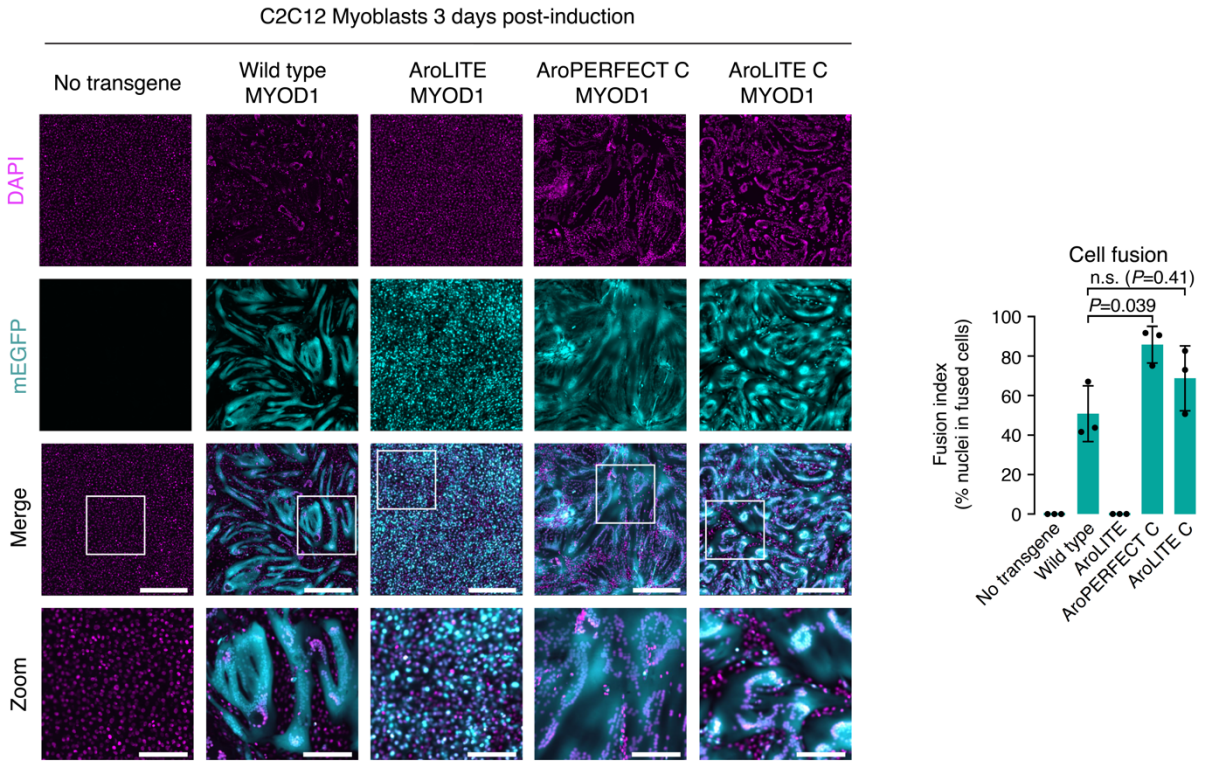


Figure 66: Sequence optimization of the MYOD1 C-terminal IDR enhances myotube formation. *(left) Representative fluorescence microscopy images of differentiating myoblasts expressing the indicated MYOD1 proteins at day 3 after Dox induction. The mEGFP signal of the MYOD1-T2A-mEGFP construct was used as a cytoplasmic marker and is shown in cyan. Nuclear counterstain (Hoechst) is shown in magenta. Scale bar is 0.5 mm. (right) Quantification of MYOD1 driven myotube differentiation efficiency. Fusion coefficient was calculated as the percentage of nuclei in cells containing at least 3 nuclei. Data are displayed as mean ± SD. N = 15 images from three biological replicates. P-values are from two-sided unpaired t-tests.*

The resulting fusion index represents the proportion of myoblasts that efficiently differentiated into multinucleated myotubes. This analysis indicated a significant increase in myotube formation in the cells expressing MYOD1 AroPERFECT C compared to the wild type MYOD1 and a marginal, but not significant, increase in myotube formation for the MYOD1 AroLITE C condition, while MYOD1 AroLITE-expressing cells failed to differentiate into myotubes (Figure 66).

To further dissect the molecular basis of the enhanced reprogramming efficiency of the MYOD1 AroPERFECT C mutant, bulk RNA-seq of differentiating myoblasts was carried out three days following transgene induction. Principal component analysis revealed transcriptional differences among the conditions, reflecting the reprograming outcome for each MYOD1 variant (Figure 67a). Notably, MYOD1 AroLITE did not exhibit meaningful similarity to either the uninduced C2C12 cells or the differentiating and myotube-forming MYOD1 wild type, AroPERFECT C and AroLITE C conditions. Subsequent differential expression analysis of the RNA-Seq data allowed for the definition of MYOD1 target genes, which were genes differentially expressed between the uninduced C2C12 cells and those expressing MYOD1 wild type.

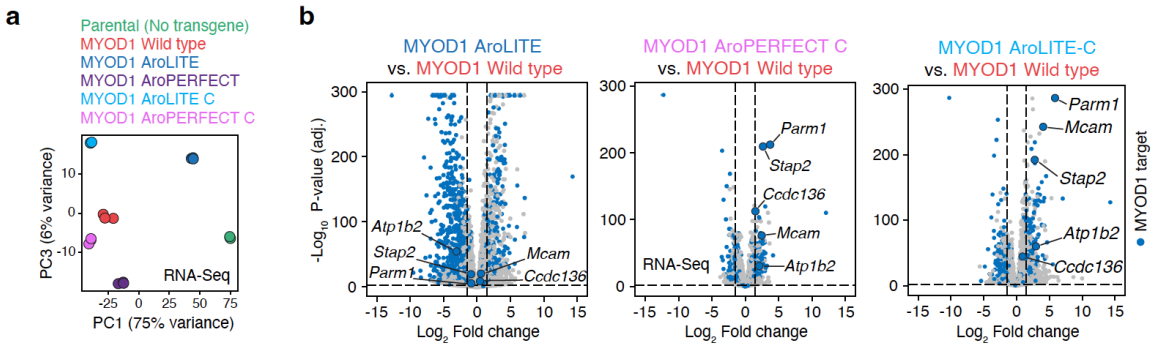


Figure 67: Differential expression in differentiating myotubes. *(a) Principal component analysis of RNA-Seq expression profiles of parental C2C12 cells, and cells expressing the indicated MYOD1 transgenes. (b) Differential expression analysis of C2C12 MYOD1 AroLITE, C2C12 MYOD1 AroLITE-C and C2C12 MYOD1 AroPERFECT-C -expressing cells versus C2C12 cells expressing wild type MYOD1. MYOD1 target genes are highlighted in blue. P-values from Benjamini-Hochberg method. Data was analyzed and plotted by Alexandre Magalhães.*

This revealed a pronounced differential expression of both MYOD1 target genes and non-target genes between the MYOD1 wild type and AroLITE conditions. Additionally, there were more similar transcriptional profiles observed between MYOD1 wild type and MYOD1 AroLITE C, as well as between MYOD1 wild type and AroPERFECT C, with the latter showing 290 differentially expressed genes, 197 of which were direct MYOD1 target genes. Gene set enrichment analysis (GSEA) revealed that these 197 differentially expressed MYOD1 target genes were particularly enriched for genes involved in cell adhesion. Therefore, providing a functional relationship between altered gene expression patterns and the enhanced reprogramming efficiency (Figure 68).

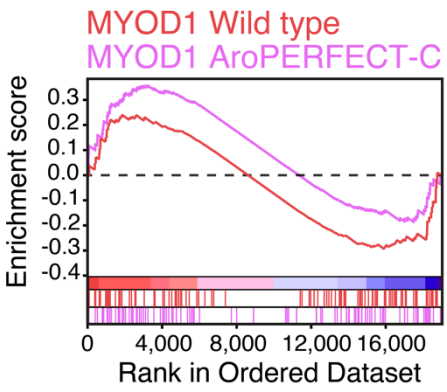


Figure 68: Differentially expressed genes are involved in cell adhesion. *Gene set enrichment analysis (GSEA) of differentially expressed genes in the MYOD1 AroPERFECT C RNA-Seq sample. Data was analyzed and plotted by Alexandre Magalhães.*

In light of the substantial transcriptional differences between cells expressing MYOD1 AroLITE and other MYOD1 variants, I conducted a Gene Ontology (GO) term analysis on genes that were upregulated in MYOD1 AroLITE-expressing cells compared to the wild type condition. This analysis revealed significant enrichment for GO terms such as “ossification”, “skeletal tissue development”, and “osteoblast differentiation” (Figure 69).

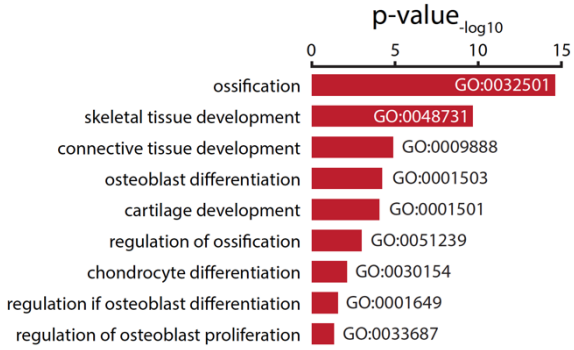


Figure 69: Differentially expressed genes in MYOD1 AroLITE vs. MYOD1 wild type are associated with osteoblast differentiation. *GO term analysis of genes upregulated in MYOD1 AroLITE compared to MYOD1 wild type. Empirical P-values are plotted.*

Given the established role of MYOD1 in the regulation of osteoblast differentiation in C2C12 cells, I examined the expression of myogenic and osteogenic marker genes in MYOD1 wild type and AroLITE cells (Figure 70) [182]. I observed a downregulation of myogenic marker genes MYH2, MYMX and MB, and an upregulation of osteogenic marker genes BMP4, GREM2 and SP7 in cells expressing MYOD1 AroLITE compared to those expressing wild type MYOD1. Additionally, I noted the expression of BMP and TGFβ signaling pathway components.

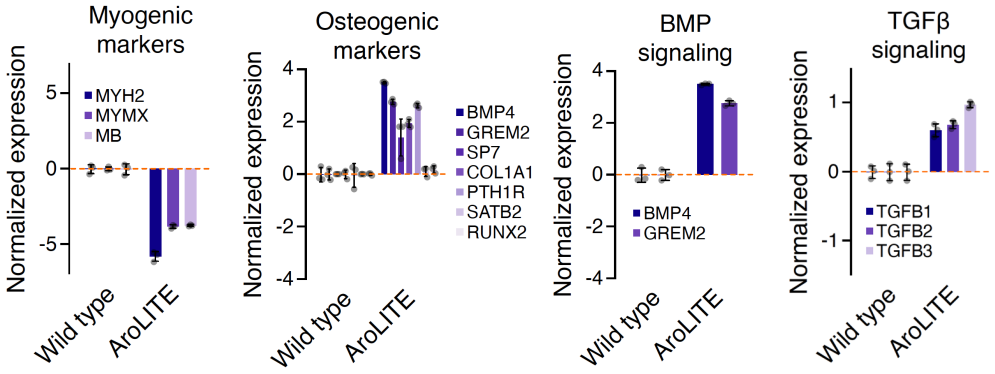


Figure 70: Differential expression of myogenic and osteogenic marker genes in MYOD1 AroLITE- compared to MYOD1 wild type-expressing cells. *Normalized RNA-expression values of myogenic and osteogenic marker genes, and BMP and TGFβ signaling markers. Expression is normalized to MYOD1 wild type.*

Taken together, introducing optimized aromatic dispersion into the MYOD1 C-IDR increases the transcriptional activity of the factor and enhances myotube formation in C2C12 myoblast cells, a process that is likely associated with transcriptional changes affecting genes implicated in cell adhesion. Moreover, I identified a specific function of MYOD1 AroLITE in promoting osteoblast differentiation, suggesting a dominant-negative effect of the transcriptionally inert mutant.

Discussion

Intrinsically disordered regions of transcription factors play a crucial role in transcriptional regulation by contributing to the formation and maintenance of transcriptional condensates [39]. Liquid-liquid phase separation, one of the physicochemical processes driving biomolecular condensate formation, is facilitated by weak-multivalent interactions among amino acid side chains in disordered regions of factors involved. Consequently, the first part of my study aimed at examining the amino acid composition of IDRs across the human proteome, seeking non-linear sequence features in functionally associated proteins indicative of condensate-specific localization or function.

The amino acid composition of human IDRs alone fails to explain the functional properties of the respective protein

When compared to the global amino acid makeup of the human proteome, the amino acid composition of predicted IDRs showed enrichment of amino acids that promote disorder – namely proline, serine and glycine – and a depletion of hydrophobic and aromatic amino acids (Figure 8). This finding aligns with expectations and reinforces the validity of the disorder prediction methodology. The distinctive cyclic structure of proline restricts protein backbone flexibility, disrupting secondary structures such as alpha-helices [183]. Glycine, being the smallest amino acid, imparts flexibility to the polypeptide chain. Serine's polar nature allows it to engage in hydrogen bonding with surrounding water molecules, potentially de-stabilizing fixed structures [184]. Conversely, the depletion of hydrophobic amino acids from IDRs was anticipated as hydrophobic side chains promote folding and structural stability by contributing to the formation of a hydrophobic core characteristic of ordered protein domains.

Major discriminants in IDR sequence composition are the proportions of charged residues, particularly driven by IDRs rich in lysine and glutamic acid, and overall hydrophobicity influenced by alanine-rich IDR sequences (Figure 9). Although I did not identify sequence compositions specific to any sub-cellular compartment, I observed an enrichment of the negatively charged amino acids glutamic acid (E) and aspartic acid (D), as well as the positively charged lysine (K), in IDRs associated with nucleolar localization (Figure 10). The presence of D/E-rich tracts and K-blocks, which have been implicated in determining localization preferences of nucleolar proteins among different phase-separated layers of the nucleolus [185], was described as a non-linear sequence feature affecting specific partitioning into condensates.

Examining the amino acid composition of transcription factor IDRs, I observed enrichment for IDRs with a high content of alanine (A). This is particularly noteworthy as alanine, with its weak hydrophobic nature is classified as a structure-promoting amino acid through facilitation of alpha-helix formation [186]. The presence of homopolymeric repeats of alanine have been implicated in transcription factor function and linked to disease when mutated [40].

Taken together, these findings suggest that amino acid composition alone is insufficient to categorize IDRs into functional groups. This insight was expected given that the variability in amino acid composition is constrained to the 20 amino acids encoded by eukaryotic cells. Such a limitation would not offer the diversity necessary to ensure accurate protein partitioning across all described condensates in a context dependent manner. Consequently, in the following I focused on effects of more specific protein-protein interactions mediated by short linear motifs (SLiMs) and non-linear sequence features.

Non-linear sequence features encoded in IDRs contribute to protein function

Protein-protein interactions can be mediated by SLiMs: sequences embedded within larger IDRs that can adopt transient secondary structures upon binding to the correct interaction partners. One example is the conserved “YPWMK” motif located in the N-terminal IDR of HOXD4, which facilitates an interaction with PBX1 for target gene regulation during embryonic development [187]. These structured interactions tend to be robust, providing significant stability and specificity. However, SLiM-mediated interactions alone, especially in the light of transcriptional condensate formation, do not fully account for the spectrum of TF functions and fail to explain phenomena such as molecular crowding and non-stoichiometric enrichment of effector molecules [38,188]. Recent work on IDRs has identified numerous non-linear sequence features that contribute to the overall functionality of the molecule by influencing the biophysical properties of cellular condensates. They do so by dictating the partitioning behavior of the factor into functionally related condensates, or determining the protein composition of the formed condensate [40,69,70,185,189]. Interestingly, the precise positioning of residues that contribute to a sequence feature within the entire sequence has only marginal influence on their function. Thus, it is assumed that these non-linear sequence features, composed of amino acids with distinctive side-chain properties, likely contribute to multivalent weak interactions among proteins within the same environment, rather than to specific structure mediated interactions with a single interaction partner.

One of these features, the dispersion of aromatic residues in prion-like domains of RNA binding proteins influences the biophysical properties of *in vitro* condensates [69]. Moreover, the

PLD of the human protein FUS has demonstrated transcriptional activity in reporter assays [190]. Consequently, I hypothesized that the dispersed distribution of aromatic residues might be a non-linear sequence feature encoded in TF IDRs, that plays a central role in regulating and linking essential functional aspects of TF biology, such as transcriptional activity and condensation.

Aromatic residues in TF IDRs facilitate TF condensation and transactivation

I selected TF candidates that encode traces of aromatic dispersion, hypothesizing that this sequence feature influences TF function. In line with the model of aromatic "stickers", candidate IDRs showed enrichment for serine and glycine, reminiscent of the "spacer" sequences in PLDs (Figure 12). Mutating all aromatic residues in the IDRs of these factors to alanine increased the c_{sat} of the respective proteins in *in vitro* droplet formation assays. Aromatic residues within disordered protein sequences can participate in non-covalent interactions, such as π-π and cation- π interactions, with other aromatic side chains or cations in spatial proximity [191]. Given the stickers-and-spacers model, mutating aromatic residues to the hydrophobic alanine decreases the overall valence of the TF, consequently reducing the cumulative binding affinity of the IDRs, resulting in a higher c_{sat}. Homotypic condensation observed *in vitro* provides valuable insights into the inter- and intramolecular interactions among IDRs with identical sequences. Caution is warranted to avoid overinterpreting these results, since the correlations to cellular function within a condensate comprising a variety of proteins at much lower endogenous concentrations may not be direct. However, mutagenesis across multiple candidates suggests that aromatic residues play a central role in condensate formation by participating in weak multivalent interactions, irrespective of the backbone composition.

Reduced binding affinities of IDRs may account for the observed reduction in transcriptional activity of AroLITE mutants. Transcriptional activation depends on the recruitment of co-activator molecules, such as the Mediator complex. Consequently, reduced binding affinities of TF IDRs may lead to decreased recruitment efficiency of co-factors by compromising condensate formation at endogenous expression levels, thereby hindering co-factor recruitment through LLPS-mediated condensate partitioning. Evaluating the activity of various TF IDRs and specifically testing the HOXB1 IDR for its transcriptional activity across different cell lines suggests that aromatic residues within the IDR sequences tested are not part of an extensive linear sequence feature that interacts exclusively with a single co-factor, given their cell type specific expression (Figure 14) [192].

An alternative explanation for altered TF activity could involve changes in DNA-binding efficiency. In the reporter system used in this study, a GAL4 DNA-binding domain was N-terminally fused to all IDRs investigated, and IDR mutagenesis did not alter the GAL4-DBD sequence. However, recent research indicates that the IDRs of transcription factors in yeast contribute to promoter recognition, thereby influencing DNA-binding efficiency and specificity [70]. This study identified dispersed aliphatic amino acids as key determinants of TF genomic localization in yeast, a role that aromatic residues have not yet been reported to play.

Transcription factors encode suboptimal aromatic dispersion

To quantify the extend of dispersion of aromatic amino acids in TF IDRs, I employed the patterning parameter Ω_{Aro}, as previously described by Martin *et al* [69]. Proteome-wide quantification of aromatic dispersion revealed that some TFs, such as the stress response factor NFAT5, contain aromatic residues in their IDRs with a dispersion far more pronounced than would be expected by random chance, thus resembling a PLD-specific sequence feature (Figure 15). However, in overall, the dispersion of aromatic residues was less pronounced in TF IDRs compared to PLDs. Prion-like domains are present in a subset of RNA-binding proteins that localize to the nucleus and regulate different aspects of RNA metabolism including splicing. They are characterized by a high content of small polar amino acids, such as serine and glycine, and periodic patterning of aromatic residues within this framework. PLDs enable proteins to engage in reversible aggregation, which is considered crucial for the formation of condensed dynamic ribonucleoprotein (RNP) granules, such as splicing speckles [193]. The term "prion-like" is derived from the similarity of these domains to prions in their ability to transition between different conformational states [194]. However, unlike pathogenic prions, the aggregation of proteins containing prion-like domains in the context of RNP granule formation is typically regulated and functional. Nonetheless, several human proteins with PLDs are associated with neurodegenerative diseases such as TDP-43 and FUS, which are implicated in amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD), tau in Alzheimer's disease, and α-synuclein in Parkinson's disease [195–197]. The presence of a PLD in these proteins is linked to their propensity to aggregate, a common pathological feature in these diseases. However, PLDs themselves are not inherently pathogenic but may confer a strong tendency to facilitate biomolecular condensation, which can become pathological under certain conditions, such as mutations or cellular stress. It has been hypothesized that the sequence feature characterizing PLDs serves as an evolutionary safeguard, driving LLPS to avert irreversible protein aggregation [198].

For TFs, the ability to partition into transcriptional condensates is considered central for their function. Consequently, aromatic dispersion could facilitate this process. It seems counterintuitive to observe submaximal aromatic dispersion in TF IDRs relative to PLD-containing proteins, given the significant role of condensation in TF functionality. The sequence feature regulating condensation and transcriptional activity of TF IDRs seems to be suboptimal.

Altering aromatic dispersion influences TF activity in both, enhancing and inhibitory manners

I decreased the aromatic dispersion in the C-terminal IDR of EGR1, which displayed significant dispersion in the wild type sequence (Figure 17). Conversely, I increased aromatic dispersion in the HOXD4 and C/EBPα IDRs (AroPERFECT and AroPERFECT IS15, respectively), where this sequence feature was less pronounced. The transcriptional activity of these mutant sequences was in line with the degree of aromatic dispersion, indicating its role as a key regulatory feature within TF IDRs (Figure 18, Figure 33). In line with observations made on AroLITE mutants, the addition of aromatic residues to the HOXD4 IDR (AroPLUS) resulted in increased transcriptional activity and condensation propensity, likely due to augmented molecular valency. Subsequent mutagenesis revealed that the “YPWM” SLiM did not influence activity in our reporter assay, suggesting either the absence of PBX1, which interacts with this motif, in mESCs, or limitations of the luciferase reporter’s 5xUAS promoter in binding the GAL4-HOXD4-IDR-PBX1 complex (Figure 19). Hence, the luciferase activity primarily reflected function of non-linear sequence features within the IDR.

Furthermore, I observed a correlation between the transcriptional activity of HOXD4 and C/EBPα IDR sequences and the presence of small inert residues (G, S, A, P) adjacent to aromatic residues (Figure 20). This suggests that the interaction potential of aromatic 'stickers' is modulated by neighboring small inert amino acids exposing the aromatic side chain, as bulky or charged side chains in close proximity could affect exposure by spatial hindrance or electrostatic forces. Additionally, non-linear sequence features within TF IDRs can synergize with embedded minimal activation domains. For instance, mutants of the C/EBPα IDR combining the N-terminal minimal activation domain with the C-terminal segment of the IDR exhibiting maximized aromatic dispersion or both, HOXD4 and C/EBPα IDR combining the N-terminal minimal activation domain with the N-terminal PLD of FUS displayed greater transcriptional activity than the activation domains or the FUS PLD alone (Figure 21, Figure 37).

Optimal aromatic dispersion enhances liquid-like features of TF condensates *in vitro*

FRAP experiments with HOXD4 AroPERFECT and C/EBPα AroPERFECT IS15 mutants revealed changes in the biophysical properties of TF IDR condensates compared to wild type sequences, with increased apparent diffusion coefficients indicating more dynamic molecular behavior within a condensate (Figure 22, Figure 38). Aromatic amino acids, when adjacent, are predisposed to forming stable aromatic clusters though π-π interactions, creating expanded delocalized electron systems. Aromatic clusters show high stability and the probability of their formation is determined by the distance between the centers of each ring [199]. In the AroPERFECT mutants, the strategic arrangement of aromatic residues with maximized distances between aromatic side chains potentially reduces the likelihood of cluster formation within the constraints of the backbone sequence, possibly enhancing condensate liquidity, and consequently, transcriptional activity as liquid-like properties of the formed condensates *in vitro* correlate with facilitated recruitment of RNAPII-CTD into condensates in live cells (Figure 32, Figure 39).

These findings provide valuable insights into principles underlying transcriptional activation where the dispersion of aromatic amino acids within TF IDRs acts as a molecular grammar that modulates transcriptional activity in conjunction with stoichiometric TF-co-factor interactions mediated by minimal activation domains. The data imply that the number and dispersion of aromatic residues within IDRs govern both the valence, affecting cumulative binding affinity, and the cluster formation, influencing liquid dynamics in condensates that are critical for co-factor recruitment and, therefore, the transcriptional output of the locus.

Optimized aromatic dispersion enhances condensation and transcriptional activity in cells

To investigate the HOXD4 AroPERFECT and AroPLUS mutants within a more endogenous cellular framework, I integrated the full-length mutants into the endogenous genomic site of HAP1 cells (Figure 25). These cells expressing HOXD4 AroPERFECT and AroPLUS exhibited morphological changes distinct from the HOXD4 KO phenotype, indicative of a gain-of-function effect. RNA sequencing experiments revealed an upregulation of HOXD4 and its target genes, consistent with its auto-regulatory role (Figure 26). Concurrently, I observed a downregulation of HOXD4 targets enriched with PBX1 targets, defined based on existing PBX1 ChIP-Seq data in the same cell line. Hence, the data imply that the “YPWM” motif within the HOXD4 IDR contributes to its gene regulatory function in HAP1 cells, and that mutagenesis of the motif (AroPERFECT) or proximity placement of aromatic amino acids (AroPLUS) interfere with PBX1 interaction, leading to a significant loss of target gene

specificity that resembles the KO phenotype. Nonetheless, both HOXD4 AroPERFECT and AroPLUS demonstrated an upregulation of a robust set of PBX1 independent HOXD4 targets and non-targets, alongside increased signal granularity in immunofluorescence microscopy in cell lines overexpressing HOXD4 wild type and mutant protein at comparable levels, corroborating that optimized aromatic dispersion in the HOXD4 IDR influences condensation and transcriptional activity in an endogenous context (Figure 27, Figure 29).

The transcriptional activity of reprogramming TFs can be optimized

To gain broader insights into TF function, I set out to apply *in vitro* direct reprogramming protocols using well-characterized master TFs. I engineered AroPERFECT mutants for 16 developmental transcription factors, with most of them commonly utilized in *in vitro* and some in *in vivo* reprogramming protocols (Figure 33, Figure 35). Success was partial, with only 7 out of 16 mutants showing an increased transcriptional activity to their wild type counterparts. For certain factors, including KLF4, SOX2, HOXB1 and SOX17, I noticed strong transcriptional activity in the wild type sequence that was reduced upon introducing optimized aromatic dispersion, possibly due to the disruption of minimal activation domains or SLiMs that mediate stoichiometric co-factor interactions essential for TF activity [200,201]. For other factors, such as RORC, TBX6 and ONECUT1, transcriptional activity remained unchanged post-optimization, suggesting irrelevance of aromatic dispersion to their function or the possibility that the altered IDRs act rather as structural linkers than regulatory domains. Lastly, in the case of NGN2, no discernible transcriptional activity was detected in either wild type or AroPERFECT mutant, indicating no assigned role in transcriptional activity for this sequence.

Optimal aromatic dispersion in TF IDRs enhances reprogramming efficiency

Increased transcriptional activity upon sequence optimization of reprogramming TFs should manifest in enhanced reprogramming efficiencies *in vitro*. To test this, I utilized C/EBPα to direct cell fate in a B-cell to macrophage direct reprogramming assay (Figure 40). C/EBPα expression reprograms pre-leukemic B-cells into macrophages within a seven-day period. Macrophages reprogramed by overexpression of C/EBPα wild type or AroPERFECT IS15 demonstrated expression of the macrophage-specific differentiation marker Mac1, indicative of successful reprogramming. Conversely, B-cells expressing C/EBPα AroLITE or AroPERFECT IS10 variants did not undergo successful reprogramming. Fluorescence microscopy revealed a cytoplasmic localization of the C/EBPα AroLITE mutant, correlating with a loss-of-function phenotype in the reprogramming assay presumably due to the factor's absence from chromatin (Figure 41). Meanwhile, the unsuccessful reprogramming by C/EBPα

AroPERFECT IS10 cannot be attributed to its cellular localization, as the factor showed nuclear localization, akin to the wild type and AroPERFECT IS15 conditions. *In vitro* droplet formation assays with purified C/EBPα-IDR-mEGFP fusions implied a reduced c_{sat} for AroPERFECT IS10 relative to the wild type and AroPERFECT IS15, likely due to increased valence leading to higher cumulative binding affinity (Figure 38). Such enhancement in valence could be responsible for the observed depletion of transcriptional activity in reporter assays. Previous research has linked reduced c_{sat} and increased cumulative binding affinities to impaired cofactor interaction, suggesting that, after mutagenesis, the factors may preferentially interact with like molecules rather than with key transcriptional cofactors, thus resulting in transcriptionally inert homotypic condensates [40].

Seven days post-induction, cells expressing C/EBPα AroPERFECT IS15 displayed an enhanced reprogramming efficiency, based on Mac1 marker expression, when compared to the cells expressing C/EBPα wild type. Single-cell RNA sequencing at the same timepoint revealed an analogous effect at the RNA level, with an augmented fraction of cells corresponding to clusters expressing both early and late macrophage markers (Figure 44, Figure 45). Notably, transcriptional differences were observable between cells expressing C/EBPα wild type and AroPERFECT IS15, particularly during earlier macrophage differentiation stages, with cells of single conditions populating clusters, partially overlapping in UMAP space, almost exclusively. These transcriptional differences were less pronounced in late-differentiated macrophages, and insufficient to preclude convergence of both conditions into a shared cluster (Figure 46). ChIP-Seq analyses of differentiating macrophages, conducted 24h and 48h after C/EBPα induction, showed a global increase in genomic binding for the AroPERFECT IS15 mutant relative to the wild type at largely overlapping *loci* (Figure 49).

There are several plausible mechanisms to account for the enhanced reprogramming efficiency observed with the C/EBPα AroPERFECT IS15 mutant. Higher read densities at shared peaks in ChIP-Seq experiments can be caused by an augmentation in TF-DNA binding affinity, not solely reliant on the bZIP DNA-binding domain of C/EBPα but also potentially enhanced by the IDR, which may contribute to promoter recognition and, thus, influence DNA-binding efficiency. Alternatively, the increased occupancy of C/EBPα AroPERFECT IS15 on DNA could be indicative of longer retention times. Considering the propensity of C/EBPα to participate in transcriptional condensate formation, one might hypothesize that enhanced liquid-like features of C/EBPα droplets *in vitro* could mirror faster condensate dynamics in live cells, resulting in a more frequent assembly of condensates and more efficient condensate maintenance, driven by accelerated diffusion of effector proteins into and out of these assemblies. Consequently, this could manifest in sustained active phases (“on-times”) at

enhancer and promoter regions, culminating in a higher density of functional molecules at these regulatory sites.

Aromatic dispersion in TF IDRs controls a molecular trade-off between activity and specificity

Enhanced genomic binding of the C/EBPα AroPERFECT IS15 mutant resulted in more promiscuous interactions with bZIP TF-binding motifs, leading to a diminished binding specificity (Figure 50). Considering the overexpression system used, it is likely that canonical C/EBPα binding sites within the genome of RCH-rtTa cells are saturated by C/EBPα under both experimental conditions. Given the overabundance of C/EBPα relative to its binding sites, the AroPERFECT mutant's lack of binding specificity may display as facilitated interaction with similar binding motifs of related bZIP transcription factors, accounting for the AroPERFECT IS15-specific peak set. Such indiscriminate binding at atypical target sites can drive gene expression at *loci* not conventionally regulated by C/EBPα (Figure 51). No striking enrichment for an alternative binding motif was evident in the AroPERFECT IS15-specific peak set, suggesting that the loss of specificity does not result from a preferential interaction with a distinct motif, but rather from a broader affinity for motifs resembling the canonical one.

I show data supporting the idea of a molecular trade-off, encoded within TF IDRs in form of the dispersion of aromatic amino acids, that regulates the balance between transcriptional activity, facilitated by effective formation and maintenance of transcriptional condensates, and DNA-binding specificity, jointly governed by the IDR and the DNA-binding domain of a transcription factor. This model implies that high activity compromises the factor's DNA-binding specificity. Consequently, for interactions requiring both high specificity and high activity, additional elements such as short linear motifs and minimal activation domains might be necessary to retain function. This theory elucidates the design challenges of AroPERFECT mutants for transcription factors like KLF4, SOX2, HOXB1 and SOX17, where the native sequence exhibits high activity mediated by minimal activation domains containing aromatic residues. It suggests a hierarchical model in which the functions of activation domains or SLiM-mediated interactions take precedence over those mediated by non-linear sequence features.

Linear and non-linear sequence features act together to regulate TF target gene expression

MYOD1 is a suitable candidate for studying the effects of non-linear sequence features and minimal activation domains on transcriptional activity within the same molecule. As a bHLH transcription factor, MYOD1 comprises an N-terminal IDR harboring a potent minimal

activation domain that drives myogenic differentiation, and a C-terminal IDR devoid of predicted minimal activation domains or SLiMs (Figure 63). Optimized aromatic dispersion in the activating N-terminal IDR resulted in diminished transcriptional activity, presumably due to disruption of the minimal activation domain characterized by aromatic residues. Conversely, sequence optimization in the C-terminal IDR led to increased transcriptional activity. These findings from reporter assays highlight the proposed hierarchy between non-linear and structural sequence motifs.

After verification, that the enhanced transcriptional activity in the C-terminal IDR of MYOD1 was not a consequence of unintended creation of structural elements (Figure 64), differentiation assays were conducted to assess the capability of both wild type and mutant MYOD1 in driving mouse myoblasts to differentiate into myotubes (Figure 66). Although all MYOD1 versions with an intact N-terminal activation domain mediated myotube formation with different efficiencies, versions with mutations within this domain failed to direct cell fate along this trajectory. A mutant featuring a sequence-optimized C-terminal IDR, while preserving structural integrity of the N-terminal activation domain, demonstrated enhanced differentiation efficiency compared to the wild type. This mutant also exhibited differential expression of MYOD1 target genes, involved in cell adhesion, a critical aspect of myotube formation (Figure 68). Notably, the AroLITE mutant, with all aromatic residues in both, N- and C-terminal IDRs substituted with alanine, failed to induce myogenic differentiation, instead promoting trans-differentiation of cells resembling osteoblasts by expression of osteogenic markers and BMP and TGFβ pathway activation (Figure 69, Figure 70). While myoblasts and osteoblasts share a common mesenchymal precursor, MYOD1 as a myogenic transcription factor has not been associated with osteoblast differentiation [202]. Therefore, the transcriptionally inactive MYOD1 AroLITE variant appears to exert a specific dominant negative role in cell fate specification of C2C12 cells controlling myogenic and osteogenic transcriptional programs.

Sequence suboptimization as a consequence of a molecular trade-off between TF functions

Sequence suboptimality has been observed and described in transcriptional regulation within TF DNA-binding motifs in cis-regulatory elements such as enhancers [203]. Previous work on the sea urchin *Ciona intestinalis* by Farley *et al.* has experimentally validated that flanking nucleotides of core binding motifs of GATA and ETS transcription factors within enhancer regions can be mutated to enhance transcriptional activity of the controlled locus, resulting from a stronger TF-DNA interaction. Consequently, this enhanced transcriptional activity was accompanied by a loss of the tissue-specific expression pattern in *C. intestinalis* embryos.

Similar results were observed when an affinity-optimized ZRS enhancer was created to regulate expression of Shh in the developing mouse limb, leading to polydactyly phenotypes [204]. In both cases, while the overall enhancer sequence was optimal for its function, in the developing embryo, the functional feature encoded - the DNA-binding motif sequence - was suboptimized to ensure appropriate gene activation at the respective locus. Thus, sequence suboptimization of DNA-binding motifs within developmental enhancers appears to regulate an important evolutionary trade-off between transcriptional activity and specificity.

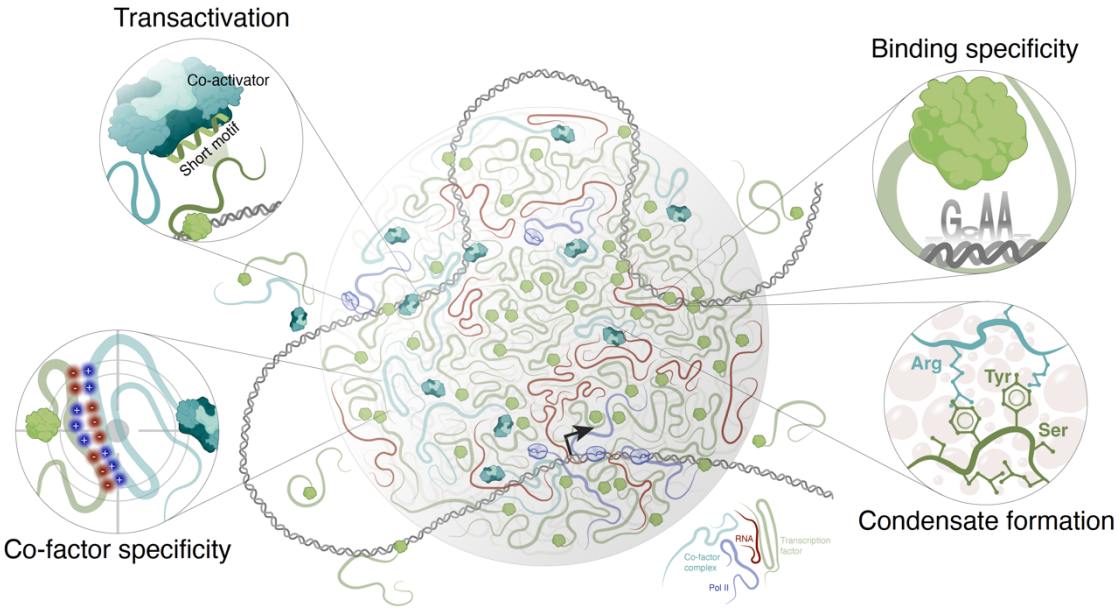


Figure 71: Schematic overview of different functional aspects of transcription factor biology within the framework of the transcriptional condensate model for gene regulation.

The principle of suboptimization has been extensively used in fields such as mechanical engineering, where it describes the concept of intentionally underperforming a certain system feature, for example a motorcycle engine, for the benefit of the systems overall performance. Designing an engine to produce a high level of horsepower and torque to maximize performance, while the vehicles transmission or body panels are not sufficient to handle this power, can lead to increased stress on the construction and potentially reduce the longevity of the entire system. This principle is equally applicable to biological systems, where a single protein may have multiple functional features contributing to its overall function [205]. For instance, a human transcription factor functions by regulating target gene expression, but this function is governed by a complex interplay of factors such as stability, SLiMs, transcriptional activity, DNA-binding specificity, size, and condensation propensity. These features, all encoded within the same amino acid sequence, interact in a delicate balance, exemplifying a molecular trade-off.

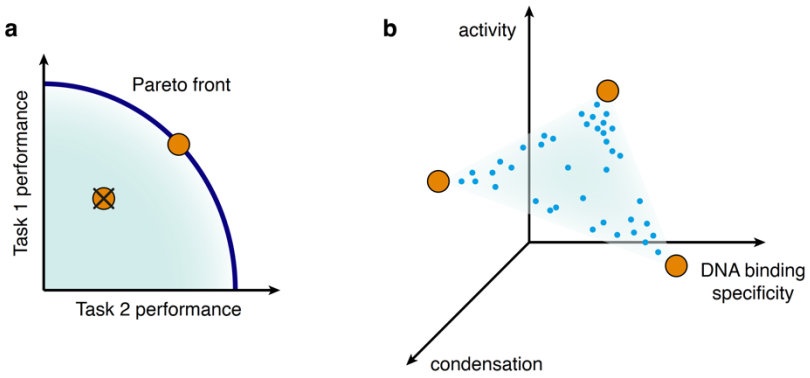


Figure 72: Pareto optimality principle adapted to the phenotype space of human transcription factors. *(a) Two-dimensional phenotype space with the theoretical Pareto front highlighted in dark blue. (b) Schematic representation of a three-dimensional phenotype space using functional features of transcription factors. Note that a two-dimensional triangular Pareto front is generated. Figure adapted from Shoval et al.*

To understand the trade-offs in human transcription factors, one can apply the Pareto front concept, as outlined by Shoval *et al.* This concept suggests that evolutionarily shaped molecules represent sets of designs with optimal trade-offs among all their functional features, and that deviations from these designs are likely to be selected against [206]. Within this framework, all optimal designs together constitute a Pareto front in a multidimensional phenotype space where no single functional feature of a design can be enhanced without compromising another. Following that rational would infer that the functional features of TFs cannot be improved without diminishing the overall performance of the factor in a physiological context. The precise mechanisms of how these trade-offs are manifested in TF biology and how they are encoded in the TF sequence are elusive.

In accordance with the Pareto front concept applied to human TFs, two predictions can be made:

First, assuming that the dispersion of aromatic residues is a sequence feature regulating a molecular trade-off in TF biology, altering aromatic dispersion should simultaneously impact TF functional features both positively and negatively.

Ambivalence in transcriptional activity upon changes in aromatic dispersion has been observed in experiments with EGR1 and HOXD4 IDR mutants (Figure 17, Figure 18). Specifically, a decrease in aromatic dispersion within the EGR1 IDR resulted in reduced transcriptional activity, whereas an increase in aromatic dispersion within the HOXD4 IDR enhanced transcriptional activity.

Second, enhancing one or several functional features inevitably required a compromise in the overall performance of the TF.

Enhanced transcriptional activity of the C/EBPα AroPERFECT IS15 mutant resulted in more promiscuous interactions with bZIP TF-binding motifs. Consequently, I observed a diminished binding specificity of C/EBPα AroPERFECT IS15 illustrating a molecular trade-off between TF activity and specificity regulated by the dispersion of aromatic amino acids within the C/EBPα IDR.

In the two systems examined (C/EBPα and MYOD1), the anticipated compromise in the overall performance of the AroPERFECT mutants tested was not observed, suggesting that the systems studied do not fully capture the breadth of transcription factor functions, such as those observed in developing embryos, where TFs regulate gene expression at endogenous expression levels in a spatial and temporal context. Although not investigated in this study, I anticipate that introducing AroPERFECT mutants of C/EBPα or MYOD1 into developing embryos will result in a critical loss-of-function phenotype, disrupting the delicate equilibrium of tissue-temporal expression crucial for proper development. In the *in vitro* reprogramming assays performed, transcriptional mis-regulation can be overridden by high concentrations of master TFs dictating differentiation, regardless of off-target expression. Therefore, AroPERFECT mutants of master TFs could potentially be used to optimize *in vitro* reprogramming protocols, which often face limitations when transitioning to *in vivo* application such as cell replacement therapy due to low conversion efficiencies [120].

Outlook

TF-mediated *in vivo* reprogramming is gaining increasing importance for clinical applications such as cell replacement therapy, with each new cell type derived from human iPSCs offering new opportunities for *in vivo* applications. Unfortunately, many *in vitro* systems fail to transition to the stage of *in vivo* application due to incomplete maturation of the desired cell type or low conversion efficiencies [120]. Sequence-optimized TF mutants enhance reprogramming efficiency *in vitro*, with minimal off-target gene expression. Therefore, optimizing aromatic dispersion in reprogramming TFs offers a promising method to refine protocols for *in vivo* interventions. Previous strategies aimed at enhancing transcriptional activity of TFs have involved either mutating the DNA-binding domain or incorporating strong viral activation domains such as the commonly used VP16 domain to the TF [207]. These alterations often resulted in strong non-specific binding and the activation of unrelated genes. The sequence optimization approach described in this study aims to boost TF function with minimal interference to the TF sequence, thereby improving reprogramming efficiency *in vitro* while ensuring minimal differential gene expression.

A promising approach to mitigate tissue damage following stroke involves *in vivo* reprogramming of stroke-affected brain tissue into functional neurons through viral transduction with TF overexpression vectors [208,209]. However, this *in vivo* conversion is hampered by suboptimal reprogramming efficiency. To address this, I engineered an AroPERFECT mutant of the human neuronal master TF NGN2, optimizing aromatic dispersion in its C-terminal IDR (Figure 52). NGN2 AroPERFECT demonstrated enhanced *in vitro* reprogramming efficiency of human iPSCs into induced Neurons, and genomic binding with minimal off-target interactions, consequently showing negligible differential gene expression five days post NGN2 induction, compared to the wild type TF (Figure 56, Figure 58, Figure 60). To evaluate the potential of NGN2 AroPERFECT *in vivo*, I engineered an AroPERFECT version of the conserved mouse NGN2 orthologue. In a pilot experiment, the reprogramming efficiency of mouse NGN2 AroPERFECT on primary mouse astrocytes, extracted from five to seven days old pups, was assessed following transduction with viral vectors carrying NGN2 wild type and AroPERFECT overexpression cassettes (Figure 73). Seven days later, reprogramming efficiency was evaluated using beta-3-tubulin immunofluorescence staining, which indicated an increase in efficiency from approximately 30% with wild type NGN2 to nearly 75% using the sequence-optimized TF. Current research is directed towards delineating the transcriptomic and genomic binding profiles of the overexpressed factors, with the aim of laying the groundwork for the first TF-mediated *in vivo* reprograming experiments utilizing TFs sequence-optimized for maximal aromatic dispersion.

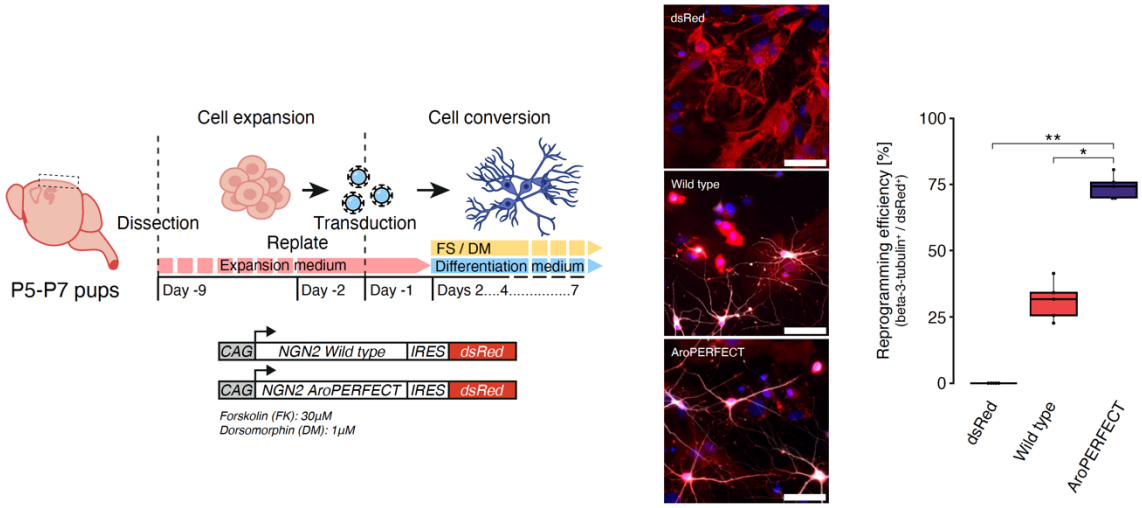


Figure 73: TF-mediated reprogramming of primary mouse astrocytes into iNeurons using a sequence-optimized NGN2 mutant. *(left) Schematic of the experimental workflow. (center) Immunofluorescence imaging of differentiated iNeurons using the negative control dsRED, NGN2 wild type or NGN2 AroPERFECT. DsRed (red), Hoechst (blue), beta-3-tubulin (white). (right) Quantification reprograming efficiency by calculation of the fraction of dsRed+/beta-3-tubulin+ cells. P-values from unpaired two-sided t-test. *: P< 0.05, **: P< 0.01. Data generated and analyzed by Giacomo Masserdotti & Sofia Pushkareva.*

Appendix

List of abbreviations

3'HR	Three-prime homology region
5'HR	Five-prime homology region
A	Alanine
AD	Activation domain
ALS	Amyotrophic Lateral Sclerosis
ANOVA	Analysis of Variance
appD	Apparent Diffusion coefficient
ARHGAP4	Rho GTPase-activating protein 4
AroP	AroPERFECT
ATCC	American Type Culture Collection
ATP	Adenosine triphosphate
BACH2	BTB domain and CNC homolog 2
BDNF	Brain derived neural factor
bHLH	Basic helix-loop-helix
BHLHE22	Basic Helix-Loop-Helix Family Member E22
BMP	Bone Morphogenetic Protein
BMP4	Bone Morphogenetic Protein 4
Bp	Base pair
BRD4	Bromodomain-containing protein 4
bZIP	Basic leucine zipper
C	Cysteine
C-IDR	C-terminal IDR
c-myc	MYC proto-oncogene
C/EBPA	CCAAT/enhancer-binding protein alpha
CD14	Cluster of Differentiation 14
CD19	Cluster of Differentiation 19
CD66	Cluster of Differentiation 66
cDNA	C
CDX2	Caudal type homeobox 2
CEACAM1	Carcinoembryonic antigen-related cell adhesion molecule 1
CEACAM8	Carcinoembryonic antigen-related cell adhesion molecule 8
CEBPB	CCAAT/enhancer binding protein beta
CEBPG	CCAAT/enhancer binding protein gamma
CFD	Complement factor D
CFP	Cyan fluorescent protein
ChIP-Seq	Chromatin immunoprecipitation-Sequencing
Chr	Chromosome
CNFN	Cornifelin

CO_2	Carbon dioxide
CRE	Cis-regulatory element
CREB1	CAMP responsive element binding protein 1
CREB3	CAMP responsive element binding protein 3
CRISPR/Cas9	Clustered Regularly Interspaced Short Palindromic Repeats and CRISPR-associated protein 9
Csat	Critical saturation concentration
CSF-1	Colony-Stimulating Factor 1
CSF3R	Colony stimulating factor 3 receptor
CTCF	CCCTC-binding factor
CTD	C-terminal domain
CTSD	Cathepsin D
CV	Column volumes
CXCL17	C-X-C motif chemokine ligand 17
D	Aspartic acid
D/E-rich	Aspartic acid/ glutamic acid-rich
DBD	DNA-binding domain
DIC	Differential Interference Contrast
DNA	Deoxyribunucleic acid
Dox	Doxycycline
DTT	Dithiothreitol
DUSP4	Dual Specificity Phosphatase 4
E	Glutamic acid
E.coli	Escherichia coli
E2A	E2A immunoglobulin enhancer-binding factors E12/E47
EBF1	Early B-cell Factor 1
EBF3	Early B-cell Factor 3
EDTA	Ethylenediaminetetraacetic acid
EGR1	Early Growth Response 1
EGTA	Ethylene glycol-bis(β-aminoethylether)-N,N,N',N'-tetraacetic acid
EMX1	Empty Spiracles Homeobox 1
EN1	Engrailed Homeobox 1 gene.
EOMES	Eomesodermin
ESX1	ESX homeobox 1
EWSR1	EWS RNA-Binding Protein 1
F	Phenylalanine
F.o.I	Fold over input
FACS	Fluorescence-Activated Cell Sorting
FAM98A	Family with sequence similarity 98 member A
FBS	Fetal Bovine Serum
FCGR2A	Fc fragment of IgG receptor IIa
FCGR2B	Fc fragment of IgG receptor IIb

FCGR2C	Fc fragment of IgG receptor IIc
FCGR3A	Fc fragment of IgG receptor IIIa
FCGR3B	Fc fragment of IgG receptor IIIb
FE	Fold enrichment
FERD3L	Fer3-Like BHLH Transcription Factor
FGF13	Fibroblast Growth Factor 13
FIB1	Fibrillarin 1
FITC	Fluorescein isothiocyanate
FOS	Fos proto-oncogene
FOXA3	Forkhead box A3
FOXG1	Forkhead Box G1
FRAP	Fluorescence recovery after photobleaching
FTD	Frontotemporal Dementia
FUS	Fused in sarcoma
FUSN	N-terminal region of Fused in sarcoma
FUSNxs	Shortened N-terminal region of Fused in sarcome
Fwd	Forward
FXYD5	FXYD domain-containing ion transport regulator 5
G	Glycine
GATA6	GATA Binding Protein 6
GBP5	Guanylate binding protein 5
GEM	Gel-beads in-emulsion
GFAP	Glial Fibrillary Acidic Protein
GFP	Green fluorescent protein
GLIPR1	GLI pathogenesis-related 1
Gly	Glycine
GNG5	G Protein Subunit Gamma 5
GO	Gene ontology
GPM6A	Glycoprotein M6A
GPM6B	Glycoprotein M6B
GPX4	Glutathione peroxidase 4
GREM2	Gremlin 2
GRINA	Glutamate Receptor, Ionotropic, N-Methyl D-Aspartate-Associated Protein 1
GS-linker	Glycine/Serine-linker
GSEA	Gene set enrichment analysis
H	Histidine
H23K27ac	Histone-3-Lysine-24 acetylation
H2AZ2	H2A Histone Family Member Z2
H3K4me	Histone-3-Lysine-4 methylation
HAT	Histone acetyltransferase
HDAC	Histone deacetylase
HDM	Histone demethylase

HEK293T	Human embryonic kidney 293 cells with a SV40 T-antigen
HES1	Hes Family BHLH Transcription Factor 1
hiPSC	Human induced pluripotent stem cell
HLA-A	Human leukocyte antigen A
HLA-B	Human leukocyte antigen B
HLA-C	Human leukocyte antigen C
HLA-DRB1	Human leukocyte antigen DR beta 1
HLA-E	Human leukocyte antigen E
HMG	High mobility group
HMT	Histone methyltransferase
HNF1A	Hepatocyte nuclear factor 1-alpha
HNF4A	Hepatocyte nuclear factor 4 alpha
HNRNPA1	Heterogeneous Nuclear Ribonucleoprotein A1
HNRNPA3	Heterogeneous Nuclear Ribonucleoprotein A3
HOXB1	Homeobox B1
HOXB3	Homeobox B3
HOXC4	Homeobox C4
HOXD4	Homeobox D4
HSP90	Heat Shock Protein 90
HSPA6	Heat shock protein family A member 6
I	Isoleucine
IDR	Intrinsically disordered region
IFI16	Interferon gamma-inducible protein 16
IgG	Immunoglobulin G
IL-2	Interleukin 2
IL2RA	Interleukin 2 receptor alpha
iNeurons	Induced neurons
IRX2	Iroquois Homeobox 2
IS10	Interspacer 10
IS15	Interspacer 15
ITGAM	Integrin subunit alpha M
JUND	Jun D proto-oncogene
K	Lysine
K-blocks	Lysine blocks
Kb	Kilo base
kDa	Kilo Dalton
KLF4	Krüppel-like factor 4
KO	Knockout
KRAB	Krüppel-associated box
KRAB-ZF	KRAB-zinc finger
L	Leucine
LacI	Lac repressor
LacO	Lac operator

LBH	Limb Bud and Heart Development
LiCl	Lithium chloride
LIF	Leukemia inhibitory factor
LLPS	Liquid-liquid phase separation
Log	Logarythm
M	Methionine
Mac1	Integrin subunit alpha M
MAF	V-maf avian musculoaponeurotic fibrosarcoma oncogene homolog
MAFF	V-maf avian musculoaponeurotic fibrosarcoma oncogene homolog F
MAFK	V-maf avian musculoaponeurotic fibrosarcoma oncogene homolog K
MB	Myoglobin
MCS	Multiple cloning site
MEF	Mouse embryonic fibroblasts
mEGFP	Monomeric enhanced green fluorescent protein
mESC	Mouse embryonic stem cells
MIP	Maximum intensity projection
mm	Millimeter
MMP9	Matrix metallopeptidase 9
MPIMG	Max-Planck Institute for molecular genetics
MPPED2	Metallophosphoesterase Domain Containing 2
mRNA	Messenger RNA
MWCO	Molecular weight cut-off
MYH2	Myosin Heavy Chain 2
MYMX	Myomixer
N	Asparagine
N-IDR	N-terminal IDR
n.s.	Not significant
NaCl	Sodium chloride
NES	Nestin
NEUROD1	Neurogenic Differentiation 1
NFAT5	Nuclear factor of activated T-cells 5
NFIL3	Nuclear factor, interleukin 3 regulated
NGN2	Neurogenin 2
NHR	Nuclear hormone receptor
NMR	Nuclear Magnetic Resonance
Norm. exp.	Normalized expression
NPM1	Nucleophosmin 1
NR2F1	Nuclear receptor subfamily 2 group F member 1
NRF1	Nuclear respiratory factor 1
NTRK1	Neurotrophic Receptor Tyrosine Kinase 1
OCT4	Octamer-binding transcription factor 4
OLIG1	Oligodendrocyte Transcription Factor 1

ONECUT1	One cut homeobox 1
OSN	OCT4, SOX2, NANOG
OTX1	Orthodenticle Homeobox 1
P	Proline
P-TEFb	Positive Transcription Elongation Factor b
P-value	Probability value
PAX6	Paired Box 6
PBX	Pre-B-cell leukemia transcription factor
PBX1	Pre-B-cell leukemia transcription factor 1
PCA	Principal component analysis
PCDH9	Protocadherin 9
PCR	Polymerase chain reaction
PDCD4	Programmed cell death protein 4
PDX1	Pancreatic and duodenal homeobox 1
PEG-8000	Polyethylene Glycol 8000
PFA	Paraformaldehyde
PIC	Protease inhibitor cocktail
PLCG2	Phospholipase C gamma 2
PLD	Prion-like domain
pLDDT	Predicted Local Distance Difference Test
POLR2A	RNA Polymerase II subunit A
PTPRC	Protein tyrosine phosphatase, receptor type C
Q	Glutamine
R	Arginine
RBM14	RNA Binding Motif Protein 14
Rev	Reverse
RNA	Ribonucleic adic
RNA-Seq	RNA-Sequencing
RNAPII	RNA Polymerase II subunit A
RNP	Ribonucleoprotein
ROCKi	Rho-kinase inhibitor
RORC	RAR-related orphan receptor C
RQ	Relative quantification
RT-qPCR	Reverse-transcriptase quantitative polymerase chain reaction
rtTA	Reverse tetracycline-controlled transactivator
RUNX2	Runt-related transcription factor 2
S	Serine
S100A4	S100 calcium-binding protein A4
SAT1	Spermidine/spermine N1-acetyltransferase 1
scRNA-Seq	Single cell RNA-Sequencing
SD	Standard deviation
SDS	Sodium dodecyl sulfate
Ser	Serine

SERTM1	Serine Rich And Transmembrane Domain Containing 1
sgRNA	Single guide RNA
SHOX2	Short Stature Homeobox 2
SLiM	Short linear motif
SMAP2	Small ArfGAP2
SOX17	SRY-box transcription factor 17
SOX2	SRY-box transcription factor 2
SP7	Sp7 Transcription Factor
SSR2	Signal Sequence Receptor Subunit 2
SUPT5H	SPT5 Homolog
T	Threonine
TAF15	TATA-Box Binding Protein Associated Factor 15
TBST	Tris-buffered saline with Tween
TBX1	T-Box 1
TBX5	T-Box 5
TBX6	T-box 6
TCOF1	Treacle ribosome biogenesis factor 1
TDP-43	TAR DNA-binding protein 43
TetO	Tetracycline operator
TF	Transcription factor
TGFβ	Transforming Growth Factor beta
TLE1	Transducin Like Enhancer Of Split 1
TMEM97	Transmembrane Protein 97
TRIM28	Tripartite motif-containing 28
TSS	Transcription start site
TUBB1	Tubulin Beta 1 Class VI
UAS	Upstream activator sequence
UE	Upstream element
uM	Micro molar
UMAP	Uniform Manifold Approximation and Projection
V	Valine
v2	Version 2
VP16	Viral protein 16
VST	Variance stabilizing transformation
W	Tryptophan
WT	Wild type
Y	Tyrosine
YFP	Yellow fluorescent protein
YY1	Yin Yang 1
ZF	Zinc finger
ZRS	Zone of polarizing activity regulatory sequence

Protein sequences and SLIMs

Aromatic residues are in bold face, Short linear motifs (L/F/Y/W XX L/F/Y/W) are underlined. The sequences are the translated protein sequences used in the *in vitro* droplet formation and transactivation assays. The SLiM counts are listed at the bottom.

HOXD4 IDR wild type

MVMSS**Y**MVNSK**Y**VDPK**F**PPCEE**Y**LQGG**Y**LGEQGAD**YY**GGGAQGAD**F**QPPGL**Y**PRPD**F**G
EQP**F**GGSGPGPGSALPARGHGQEPGGPGGH**Y**AAPGEPCPAPPAPPPAPLPGARA**Y**SQSD
PKQPPSGTALKQPAVV**Y**P**W**MKKV

HOXD4 IDR AroLITE A

MVMSSAMVNSKAVDPKAPPCEEALQGGALGEQGADAAGGGAQGADAQPPGLAPRPDAG
EQPAGGSGPGPGSALPARGHGQEPGGPGGHAAAPGEPCPAPPAPPPAPLPGARAASQSD
PKQPPSGTALKQPAVVAPAMKKV

HOXD4 IDR AroLITE S

MVMSSSMVNSKSVDPKSPPCEESLQGGSLGEQGADSSGGGAQGADSQPPGLSPRPDSG
EQPSGGSGPGPGSALPARGHGQEPGGPGGHSAAPGEPCPAPPAPPPAPLPGARASSQSD
PKQPPSGTALKQPAVVSPSMKKVMSRGP**Y**SIVSPKC

HOXD4 IDR AroLITE G

MVMSSGMVNSKGVDPKGPPCEEGLQGGGLGEQGADGGGGGAQGADGQPPGLGPRPD
GGEQPGGGSGPGPGSALPARGHGQEPGGPGGHGAAPGEPCPAPPAPPPAPLPGARAGS
QSDPKQPPSGTALKQPAVVGPGMKKVMSRGP**Y**SIVSPKC

HOXD4 IDR AroPLUS

MVMSS**Y**MVNSK**Y**VDPK**F**PPCEE**Y**LQGG**Y**LGEQGAD**YY**GGGAQGAD**F**QPPGL**Y**PRPD**F**G
EQP**F**GGSGPG**Y**GSALPAR**Y**HGQEP**Y**GPGGH**Y**AAPGEPCP**Y**PPAPPP**Y**PLPGARA**Y**SQSD
PK**Y**PPSGTA**Y**KQPAVV**Y**P**W**MKKV

HOXD4 IDR AroPLUS LITE

MVMSS**Y**MVNSK**Y**VDPK**F**PPCEE**Y**LQGG**Y**LGEQGAD**YY**GGGAQGAD**F**QPPGL**Y**PRPD**F**G
EQP**F**GGSGPGAGSALPARAHGQEPAGPGGH**Y**AAPGEPCPAPPAPPPAPLPGARA**Y**SQSD
PKAPPSGTAAKQPAVV**Y**P**W**MKKV

HOXD4 IDR AroPLUS patched

MVMSS**Y**MVNSK**Y**VDPK**FY**PPCEE**Y**LQGG**YY**LGEQGAD**YY**GGGAQGAD**F**QPPGL**YY**PRP
D**F**GEQP**FY**GGSGPGGSALPARHGQEPGPGGH**YY**AAPGEPCPPPAPPPPLPGARA**YY**SQS
DPKPPSGTAKQPAVV**YY**P**W**MKKV

HOXD4 IDR AroPLUS patched LITE
MVMSS**Y**MVNSK**Y**VDPK**F**APPCEE**Y**LQGG**Y**ALGEQGAD**YY**GGGAQGAD**F**QPPGL**Y**APRP
D**F**GEQP**F**AGGSGPGGSALPARHGQEPGPGGHA**Y**AAPGEPCPPPAPPPPLPGARA**Y**ASQS
DPKPPSGTAKQPAVVA**Y**P**W**MKKV

HOXD4 IDR AroPERFECT
MVMSSMVN**Y**SKVDPKPP**F**CEELQGGL**Y**GEQGADGG**Y**GAQGADQP**F**PGLPRPDG**F**EQPG
GSGP**F**GPGSALPA**Y**RGHGQEPG**Y**GPGGHAAP**Y**GEPCPAPP**Y**APPPAPLP**Y**GARASQSD**Y**
PKQPPSGT**Y**ALKQPAVV**W**PMKKV

HOXD4 IDR AroPERFECT -1
MVMSSMV**Y**NSKVDPKP**F**PCEELQGG**Y**LGEQGADG**Y**GGAQGADQ**F**PPGLPRPD**F**GEQPG
GSG**F**PGPGSALP**Y**ARGHGQEP**Y**GGPGGHAA**Y**PGEPCPAP**Y**PAPPPAPL**Y**PGARASQS**Y**D
PKQPPSG**Y**TALKQPAV**W**VPMKKV

HOXD4 IDR AroPERFECT -2
MVMSSM**Y**VNSKVDPK**F**PPCEELQG**Y**GLGEQGAD**Y**GGGAQGAD**F**QPPGLPRP**F**DGEQPG
GS**F**GPGPGSAL**Y**PARGHGQE**Y**PGGPGGHA**Y**APGEPCPA**Y**PPAPPPAP**Y**LPGARASQ**Y**SD
PKQPPS**Y**GTALKQPA**W**VVPMKKV

HOXD4 IDR wild type YPWM(-)
MVMSS**Y**MVNSK**Y**VDPK**F**PPCEE**Y**LQGG**Y**LGEQGAD**YY**GGGAQGAD**F**QPPGL**Y**PRPD**F**G
EQP**F**GGSGPGPGSALPARGHGQEPGGPGGH**Y**AAPGEPCPAPPAPPPAPLPGARA**Y**SQSD
PKQPPSGTALKQPAVVAPAMKKV

HOXD4 IDR AroLITE YPWM(+)
MVMSSAMVNSKAVDPKAPPCEEALQGGALGEQGADAAGGGAQGADAQPPGLAPRPDAG
EQPAGGSGPGPGSALPARGHGQEPGGPGGHAAAPGEPCPAPPAPPPAPLPGARAASQSD
PKQPPSGTALKQPAVV**Y**P**W**MKKV

HOXD4 IDR AroPERFECT YPWM (+)

MVMSSMVN**Y**SKVDPKPP**F**CEELQGGL**Y**GEQGADGG**Y**GAQGADQP**F**PGLPRPDG**F**EQPG
GSGP**F**GPGSALPA**Y**RGHGQEPG**Y**GPGGHAAP**Y**GEPCPAPP**Y**APPPAPLP**Y**GARASQSD**Y**
PKQPPSGTALKQPAVV**Y**P**W**MKKV

HOXD4 WT (N)
MVMSSMVNSKVDPK**F**PPCEE**Y**LQGG**Y**LGEQGAD**YY**GGGAQGAD**F**QPPGL**Y**PRPD**F**GEQ
P**F**GGSGPGPGSAL

HOXD4 WT(N)-FUSNxs
MVMSSMVNSKVDPK**F**PPCEE**Y**LQGG**Y**LGEQGAD**YY**GGGAQGAD**F**QPPGL**Y**PRPD**F**GEQ
P**F**GGSGPGPGSALASND**Y**TQQATQS**Y**GA**Y**PTQPGQG**Y**SQQSSQP**Y**GQQS**Y**SG**Y**SQSTD
TSG**Y**GQSS

FUSN
ASND**Y**TQQATQS**Y**GA**Y**PTQPGQG**Y**SQQSSQP**Y**GQQS**Y**SG**Y**SQSTDTSG**Y**GQSS**Y**SS**Y**G
QSQNTG**Y**GTQSTPQG**Y**GSTGG**Y**GSSQSSQSS**Y**GQQSS**Y**PG**Y**GQQPAPSSTSGS**Y**GSSS
QSSS**Y**GQPQSGS**Y**SQQPS**Y**GGQQQS**Y**GQQQS**Y**NPPQG**Y**GQQNQ**Y**NSSSGGGGGGGG
GGN**Y**GQDQSSMSSGGGSGGG**Y**GNQDQSGGGGSGG**Y**GQQDRGGRGRGGSGGGGGG
GGGG**Y**NRSSGG**Y**EPRGRGGGRGGRGGMGGSDRGG**F**NK**F**GGPRDQGSRHDSEQDNSD
NNTI

FUSNxs
ASND**Y**TQQATQS**Y**GA**Y**PTQPGQG**Y**SQQSSQP**Y**GQQS**Y**SG**Y**SQSTDTSG**Y**GQSS

HOXC4 IDR wild type
MIMSS**Y**LMDSN**Y**IDPK**F**PPCEE**Y**SQNS**Y**IPEHSPE**YY**GRTRESG**F**QHHHQEL**Y**PPPPPRPS
YPERQ**Y**SCTSLQGPGNSRGHGPAQAGHHHPEKSQSLCEPAPLSGASASPSPAPPACSQP
APDHPSSAASKQPIV**Y**P**W**MKMSRGP**Y**SIVSPKC

HOXC4 IDR AroLITE S
MIMSSALMDSNAIDPKAPPCEEASQNSAIPEHSPEAAGRTRESGAQHHHQELAPPPPPRPS
APERQASCTSLQGPGNSRGHGPAQAGHHHPEKSQSLCEPAPLSGASASPSPAPPACSQP
APDHPSSAASKQPIVAPAMKMSRGP**Y**SIVSPKC

HOXC4 IDR AroPERFECT

MIMSSLM**Y**DSNIDPKPPC**F**EESQNSIPEH**Y**SPEGRTRESG**F**QHHHQELPPP**Y**PPRPSPERQ
S**Y**CTSLQGPGNS**Y**RGHGPAQAGH**Y**HHPEKSQSLC**Y**EPAPLSGASA**Y**SPSPAPPACS**Y**QPA
PDHPSSA**Y**ASKQPIVPM**W**KV

HOXB1 IDR wild type
MD**Y**NRMNS**F**LE**Y**PLCNRGPSA**Y**SAHSAPTS**F**PPSSAQAVDS**Y**ASEGR**Y**GGGLSSPA**F**QQ
NSG**Y**PAQQPPSTLGVP**F**PSSAPSG**Y**APAACSPS**Y**GPSQ**YY**PLGQSEGDGG**YF**HPSS**Y**GA
QLGGLSDG**Y**GAGGAGPGP**Y**PPQHPP**Y**GNEQTAS**F**APA**Y**ADLLSEDKETPCPSEPNTPTAR
T**F**D**W**MKVKRNPPKTAKVSEPGL

HOXB1 IDR AroLITE A
MDANRMNSALEAPLCNRGPSAASAHSAPTSAPPSSAQAVDSAASEGRAGGGLSSPAAQQ
NSGAPAQQPPSTLGVPAPSSAPSGAAPAACSPSAGPSQAAPLGQSEGDGGAAHPSSAGA
QLGGLSDGAGAGGAGPGPAPPQHPPAGNEQTASAAPAAADLLSEDKETPCPSEPNTPTAR
TADAMKVKRNPPKTAKVSEPGL

HOXB1 IDR AroPERFECT
MD**Y**NRMNSLE**Y**PLCNRGP**Y**SASAHSA**F**PTSPPSS**F**AQAVDSA**Y**SEGRGGG**Y**LSSPAQQ**F**
NSGPAQQ**Y**PPSTLGV**F**PPSSAPS**Y**GAPAACS**Y**PSGPSQP**Y**LGQSEGD**Y**GGHPSSG**F**AQL
GGLS**Y**DGGAGGA**Y**GPGPPPQ**Y**HPPGNEQ**Y**TASAPAA**F**DLLSEDK**Y**ETPCPSE**Y**PNTPTAR
FTDMKVKR**W**NPPKTAKVSEPGL

NANOG IDR wild type
KQVKT**WF**QNQRMKSKR**W**QKNN**W**PKNSNGVTQKASAPT**Y**PSL**Y**SS**Y**HQGCLVNPTGNLP
M**W**SNQT**W**NNST**W**SNQTQNIQS**W**SNHS**W**NTQT**W**CTQS**W**NNQA**W**NSP**FY**NCGEESLQS
CMQ**F**QPNSPASDLEAAL

NANOG IDR AroLITE A
KQVKTAAQNQRMKSKRAQKNNAPKNSNGVTQKASAPTAPSLASSAHQGCLVNPTGNLPM
ASNQTANNSTASNQTQNIQSASNHSANTQTACTQSANNQAANSPAANCGEESLQSCMQA
QPNSPASDLEAAL

EGR1 IDR wild type
LRQKDKKADKSVVASSATSSLSS**Y**PSPVATS**Y**PSPVTTS**Y**PSPATTS**Y**PSPVPTS**F**SSPGSS
T**Y**PSPVHSG**F**PSPSVATT**Y**SSVPPA**F**PAQVSS**F**PSSAVTNS**F**SASTGLSDMTAT**F**SPRTIEIC

EGR1 IDR AroLITE A

LRQKDKKADKSVVASSATSSLSSAPSPVATSAPSPVTTSAPSPATTSAPSPVPTSASSPGSS
TAPSPVHSGAPSPSVATTASSVPPAAPAQVSSAPSSAVTNSASASTGLSDMTATASPRTIEIC

EGR1 IDR AroSCRAMBLED
LRQKDKKADKSVVASSATSSLSS**Y**PSPVA**F**T**Y**SPSPVTTSPSP**Y**AT**Y**TSPSPVPTSSS**F**PGS
S**YF**TPSPVHSGP**Y**SPSVATTSSVPPAPAQVSSPSSAV**F**TNS**F**SASTG**F**LSDMTATSPRTIEIC

EGR1 IDR AroPATCHY3
LRQKDKKADKSVVASSATSSLSS**YYYY**PSPVATSPSPVTTSPSPATTSPSPVPTSSSPGSST
PSPVHSGPSPSVATT**YFFY**SSVPPAPAQVSSPSSAVTNS**FFFF**SASTGLSDMTATSPRTIEIC

EGR1 IDR AroPATCHY1
LRQKDKKADKSVVASSATSSLSSPSPVATSPSPVTTSPSPATTSPSPVPTSSSPGSST**YYYY**
FYYFFFFFPSPVHSGPSPSVATTSSVPPAPAQVSSPSSAVTNSSASTGLSDMTATSPRTIEIC

NFAT5 IDR wild type
TMVKKEISSPARPCS**F**EEAMKAMKTTGCNLDKVNIIPNALMTPLIPSSMIKSEDVTPMEVTAE
KRSSTI**F**KTTKSVGSTQQTLENISNIAGNGS**F**SSPSSSHLPSENEKQQQIQPKA**Y**NPETLTTI
QTQDISQPGT**F**PAVSASSQLPNSDALLQQATQ**F**QTRETQSREILQSDGTVVNLSQLTEASQ
QQQQSPLQEQAQTLQQQISSNI**F**PSPNSVSQLQNTIQQLQAGS**F**TGSTASGSSGSVDLVQ
QVLEAQQQLSSVL**F**SAPDGNENVQEQLSADI**F**QQVSQIQSGVSPGM**F**SSTEPTVHTRPDN
LLPGRAESVHPQSENTLSNQQQQQQQQQQVMESSAAMVMEMQQSICQAAAQIQSEL**F**PS
TASANGNLQQSPV**Y**QQTSHMMSALSTNEDMQMQCEL**F**SSPPAVSGNETSTTTTQQVATPG
TTM**F**QTSSSGDGEETGTQAKQIQNSV**F**QTMVQMQHSGDNQPQVNL**F**SSTKSMMSVQNS
GTQQQGNGL**F**QQGNEMMSLQSGN**F**LQQSSHSQAQL**F**HPQNPIADAQNLSQETQGSL**F**HS
PNPIVHSQTSTTSSEQMQPPM**F**HSQSTIAVLQGSSVPQDQQSTNI**F**LSQSPMNNLQTNTVA
QEA**FF**AAPNSISPLQSTSNSEQQAA**F**QQQAPISHIQTPMLSQEQAQPPQQGL**F**QPQVALGS
LPPNPMPQSQQGTM**F**QSQHSIVAMQSNSPSQEQQPPPPRRPLPPLPLQQSIL**F**SNQNTMA
TMASPKQPPPNMI**F**NPNQNPMANQEQQNQSI**F**HQQSNMAPMNQEQQPMQ**F**QSQSTVSS
LQNPGPTQSESSQTPL**F**HSSPQIQLVQGSPSSQEQQVTL**F**LSPASMSALQTSINQQDMQQ
SPL**Y**SPQNNMPGIQGATSSPQPQATL**F**HNTAGGTMNQLQNSPGSSQQTSGM**F**L**F**GIQNNC
SQLLTSGPATLPDQLMAISQPGQPQNEGQPPVTTLLSQQMPENSPLASSINTNQNIEKIDLLV
SLQNQGNNLTGS**F**

NFAT5 IDR AroLITE A
TMVKKEISSPARPCSAEEAMKAMKTTGCNLDKVNIIPNALMTPLIPSSMIKSEDVTPMEVTAE
KRSSTIAKTTKSVGSTQQTLENISNIAGNGSASSPSSSHLPSENEKQQQIQPKAANPETLTTI

QTQDISQPGTAPAVSASSQLPNSDALLQQATQAQTRETQSREILQSDGTVVNLSQLTEASQ
QQQQSPLQEQAQTLQQQISSNIAPSPNSVSQLQNTIQQLQAGSATGSTASGSSGSVDLVQ
QVLEAQQQLSSVLASAPDGNENVQEQLSADIAQQVSQIQSGVSPGMASSTEPTVHTRPDN
LLPGRAESVHPQSENTLSNQQQQQQQQQQVMESSAAMVMEMQQSICQAAAQIQSELAPS
TASANGNLQQSPVAQQTSHMMSALSTNEDMQMQCELASSPPAVSGNETSTTTTQQVATPG
TTMAQTSSSGDGEETGTQAKQIQNSVAQTMVQMQHSGDNQPQVNLASSTKSMMSVQNS
GTQQQGNGLAQQGNEMMSLQSGNALQQSSHSQAQLAHPQNPIADAQNLSQETQGSLAH
SPNPIVHSQTSTTSSEQMQPPMAHSQSTIAVLQGSSVPQDQQSTNIALSQSPMNNLQTNTV
AQEAAAAAPNSISPLQSTSNSEQQAAAQQQAPISHIQTPMLSQEQAQPPQQGLAQPQVALL
GSLPPNPMPQSQQGTMAQSQHSIVAMQSNSPSQEQQQQQQQQQQQSILASNQNTMATM
ASPKQPPPNMIANPNQNPMANQEQQNQSIAHQQSNMAPMNQEQQPMQAQSQSTVSSLQ
NPGPTQSESSQTPLAHSSPQIQLVQGSPSSQEQQVTLALSPASMSALQTSINQQDMQQSP
LASPQNNMPGIQGATSSPQPQATLAHNTAGGTMNQLQNSPGSSQQTSGMALAGIQNNCS
QLLTSGPATLPDQLMAISQPGQPQNEGQPPVTTLLSQQMPENSPLASSINTNQNIEKIDLLVS
LQNQGNNLTGSA

C/EBPα IDR wild type
MRGRGRAGSPGGRRRRPAQAGGRRGSPCRENSNSPMESAD**F**Y**E**AEPRPPMSSHLQSPP
HAPSSAA**F**G**F**PRGAGPAQPPAPPAAPEPLGGICEHETSIDISA**Y**IDPAA**F**NDE**F**LADL**F**QHSR
QQEKAKAAVGPTGGGGGGD**F**D**Y**PGAPAGPGGAVMPGGAHGPPPG**Y**GCAAAG**Y**LDGRLE
PL**Y**ERVGAPALRPLVIKQEPREEDEAKQLALAGL**F**P**Y**QPPPPPPPSHPHPHPPPAHLAAPHL
Q**F**QIAHCGQ

C/EBPα IDR AroLITE A
MRGRGRAGSPGGRRRRPAQAGGRRGSPCRENSNSPMESADAAEAEPRPPMSSHLQSPP
HAPSSAAAGAPRGAGPAQPPAPPAAPEPLGGICEHETSIDISAAIDPAAANDEALADLAQHS
RQQEKAKAAVGPTGGGGGGDADAPGAPAGPGGAVMPGGAHGPPPGAGCAAAGALDGRL
EPLAERVGAPALRPLVIKQEPREEDEAKQLALAGLAPAQPPPPPPPSHPHPHPPPAHLAAPH
LQAQIAHCGQ

C/EBPα IDR AroPERFECT IS15
MRG**F**RGRAGSPGGRRRRPA**Y**QAGGRRGSPCRENSN**F**SPMESADEAEPRPPM**F**SSHLQS
PPHAPSSAA**F**GPRGAGPAQPPAPPA**Y**APEPLGGICEHETSI**F**DISAIDPAANDELAD**F**LQHSR
QQEKAKAAVG**F**PTGGGGGGDDPGAPA**Y**GPGGAVMPGGAHGPP**Y**PGGCAAAGLDGRLEP
YLERVGAPALRPLVIK**Y**QEPREEDEAKQLALA**F**GLPQPPPPPPPSHPH**Y**PHPPPAHLAAPHL
QQ**F**IAHCGQ

C/EBPα IDR AroPERFECT IS15 +1

MRGR**F**GRAGSPGGRRRRPAQ**Y**AGGRRGSPCRENSNS**F**PMESADEAEPRPPMS**F**SHLQS
PPHAPSSAAG**F**PRGAGPAQPPAPPAA**Y**PEPLGGICEHETSID**F**ISAIDPAANDELADL**F**QHSR
QQEKAKAAVGP**F**TGGGGGGDDPGAPAG**Y**PGGAVMPGGAHGPPP**Y**GGCAAAGLDGRLEP
L**Y**ERVGAPALRPLVIKQ**Y**EPREEDEAKQLALAGL**F**LPQPPPPPPPSHPHP**Y**HPPPAHLAAPHL
QQI**F**AHCGQ

C/EBPα IDR AroPERFECT IS15 +2

MRGRG**F**RAGSPGGRRRRPAQA**Y**GGRRGSPCRENSNSP**F**MESADEAEPRPPMSS**F**HLQS
PPHAPSSAAGP**F**RGAGPAQPPAPPAAP**Y**EPLGGICEHETSIDI**F**SAIDPAANDELADLQ**F**HSR
QQEKAKAAVGPT**F**GGGGGGDDPGAPAGP**Y**GGAVMPGGAHGPPPG**Y**GCAAAGLDGRLEP
LE**Y**RVGAPALRPLVIKQE**Y**PREEDEAKQLALAGL**F**PQPPPPPPPSHPHPH**Y**PPPAHLAAPHL
QQIA**F**HCGQ

C/EBPα IDR AroPERFECT IS10

MRGRGRAGS**F**PGGRRRRPAQ**F**AGGRRGSPCR**Y**ENSNSPMESA**Y**DEAEPRPPMS**F**SHLQ
SPPHAP**F**SSAAGPRGAG**F**PAQPPAPPAA**F**PEPLGGICEH**Y**ETSIDISAID**Y**PAANDELADL**F**Q
HSRQQEKAK**F**AAVGPTGGGG**F**GGDDPGAPAG**F**PGGAVMPGGA**F**HGPPPGGCAA**F**AGLD
GRLEPL**F**ERVGAPALRP**F**LVIKQEPREE**Y**DEAKQLALAG**Y**LPQPPPPPPP**Y**SHPHPHPPPA**Y**
HLAAPHLQQI**Y**AHCGQ

C/EBPα IDR WT (N)

MRGRGRAGSPGGRRRRPAQAGGRRGSPCRENSNSPMESAD**FY**EAEPRPPMSSHLQSPP
HAPSSAA**F**G**F**PRGAGPAQPPAPPAAPEPLGGICEHETSIDISA**Y**IDPAA**F**NDE**F**LADL**F**QHS

C/EBPα IDR WT(N)-IS15

MRGRGRAGSPGGRRRRPAQAGGRRGSPCRENSNSPMESAD**FY**EAEPRPPMSSHLQSPP
HAPSSAA**F**G**F**PRGAGPAQPPAPPAAPEPLGGICEHETSIDISA**Y**IDPAA**F**NDE**F**LADL**F**QHS
RQQEKAKAAVG**F**PTGGGGGGDDPGAPA**Y**GPGGAVMPGGAHGPP**Y**PGGCAAAGLDGRL
EP**Y**LERVGAPALRPLVIK**Y**QEPREEDEAKQLALA**F**GLPQPPPPPPPSHPH**Y**PHPPPAHLAA
PHLQQ**F**IAHCGQ

C/EBPα IDR WT(N)-FUSN

MRGRGRAGSPGGRRRRPAQAGGRRGSPCRENSNSPMESAD**FY**EAEPRPPMSSHLQSPP
HAPSSAA**F**G**F**PRGAGPAQPPAPPAAPEPLGGICEHETSIDISA**Y**IDPAA**F**NDE**F**LADL**F**QHS
ASND**Y**TQQATQS**YGAY**PTQPGQG**Y**SQQSSQP**Y**GQQS**YSGY**SQSTDTSG**Y**GQSS**YSSY**G
QSQNTG**Y**GTQSTPQG**Y**GSTGG**Y**GSSQSSQSS**Y**GQQSS**YPGY**GQQPAPSSTSGS**Y**GSSS

QSSS**Y**GQPQSGS**Y**SQQPS**Y**GGQQQS**Y**GQQQS**Y**NPPQG**Y**GQQNQ**Y**NSSSGGGGGGGG
GGN**Y**GQDQSSMSSGGGSGGG**Y**GNQDQSGGGGSGG**Y**GQQDRGGRGRGGSGGGGGG
GGGG**Y**NRSSGG**Y**EPRGRGGGRGGRGGMGGSDRGG**F**NK**F**GGPRDQGSRHDSEQDNSD
NNTI

C/EBPα IDR WT(N)-FUSNxs
MRGRGRAGSPGGRRRRPAQAGGRRGSPCRENSNSPMESAD**FY**EAEPRPPMSSHLQSPP
HAPSSAA**F**G**F**PRGAGPAQPPAPPAAPEPLGGICEHETSIDISA**Y**IDPAA**F**NDE**F**LADL**F**QHS
ASND**Y**TQQATQS**YGAY**PTQPGQG**Y**SQQSSQP**Y**GQQS**YSGY**SQSTDTSG**Y**GQSS

NGN2 wild type
M**F**VKSETLELKEEEDVLVLLGSASPALAALTPLSSSADEEEEEEPGASGGARRQRGAEAGQ
GARGGVAAGAEGCRPARLLGLVHDCKRRPSRARAVSRGAKTAETVQRIKKTRRLKANNRE
RNRMHNLNAALDALREVLPT**F**PEDAKLTKIETLR**F**AHN**Y**I**W**ALTETLRLADHCGGGGGLPG
AL**F**SEAVLLSPGGASAALSSSGDSPSPAST**W**SCTNSPAPSSSVSSNSTSP**Y**SCTLSPASPA
GSDMD**YW**QPPPPDKHR**Y**APHLPIARDCI

NGN2 AroLITE A
MAVKSETLELKEEEDVLVLLGSASPALAALTPLSSSADEEEEEEPGASGGARRQRGAEAGQ
GARGGVAAGAEGCRPARLLGLVHDCKRRPSRARAVSRGAKTAETVQRIKKTRRLKANNRE
RNRMHNLNAALDALREVLPT**F**PEDAKLTKIETLR**F**AHN**Y**I**W**ALTETLRLADHCGGGGGLPG
AL**F**SEAVLLSPGGASAALSSSGDSPSPASTASCTNSPAPSSSVSSNSTSPASCTLSPASPAG
SDMDAAQPPPPDKHRAAPHLPIARDCI

NGN2 AroPERFECT
M**F**VKSETLELKEEEDVLVLLGSASPALAALTPLSSSADEEEEEEPGASGGARRQRGAEAGQ
GARGGVAAGAEGCRPARLLGLVHDCKRRPSRARAVSRGAKTAETVQRIKKTRRLKANNRE
RNRMHNLNAALDALREVLPT**F**PEDAKLTKIETLR**F**AHN**Y**I**W**ALTETLRLADHCGGGGGLPG
AL**F**SEAVLLSPGGASAA**W**LSSSGDSPSPASTS**Y**CTNSPAPSSSVSSN**Y**STSPSCTLSPASP
A**W**GSDMDQPPPPDKHR**Y**APHLPIARDCI

NGN2 N-IDR wild type
M**F**VKSETLELKEEEDVLVLLGSASPALAALTPLSSSADEEEEEEPGASGGARRQRGAEAGQ
GARGGVAAGAEGCRPARLLGLVHDCKRRPSRARAVSRGAKTAETVQRIKKTRRLKANNRE
RNRMHNLNAA

NGN2 N-IDR AroPERFECT

M**F**VKSETLELKEEEDV**W**LVLLGSASPALAALT**Y**TPLSSSADEEEEEE**Y**PGASGGARRQRGAE
WAGQGARGGVAAGAE**Y**GCRPARLLGLVHDC**W**KRRPSRARAVSRGA**Y**KTAETVQRIKKTR
R**Y**LKANNRERNRMHNL**W**NAA

NGN2 C-IDR wild type

AVLLSPGGASAALSSSGDSPSPAST**W**SCTNSPAPSSSVSSNSTSP**Y**SCTLSPASPAGSDMD
YWQPPPPDKHR**Y**APHLPIARDCI

NGN2 C-IDR AroLITE A

AVLLSPGGASAALSSSGDSPSPASTASCTNSPAPSSSVSSNSTSPASCTLSPASPAGSDMD
AAQPPPPDKHRAAPHLPIARDCI

NGN2 C-IDR AroPERFECT

AVLLSPGGASAA**W**LSSSGDSPSPASTS**Y**CTNSPAPSSSVSSN**Y**STSPSCTLSPASPA**W**GSD
MDQPPPPDKHR**Y**APHLPIARDCI

MYOD1 wild type

MELLSPPLRDVDLTAPDGSLCS**F**ATTDD**FY**DDPC**F**DSPDLR**FF**EDLDPRLMHVGALLKPEE
HSH**F**PAAVHPAPGAREDEHVRAPSGHHQAGRCLL**W**ACKACKRKTTNADRRKAATMRERR
RLSKVNEA**F**ETLKRCTSSNPNQRLPKVEILRNAIR**Y**IEGLQALLRDQDAAPPGAAAA**FY**APG
PLPPGRGGEH**Y**SGDSDASSPRSNCSDGMMD**Y**SGPPSGARRRNC**Y**EGA**YY**NEAPSEPRP
GKSAAVSSLDCLSSIVERISTESPAAPALLLADVPSESPPRRQEAAAPSEGESSGDPTQSPD
AAPQCPAGANPNPI**Y**QVL

MYOD1 AroLITE A

MELLSPPLRDVDLTAPDGSLCSAATTDDAADDPCADSPDLRAAEDLDPRLMHVGALLKPEE
HSHAPAAVHPAPGAREDEHVRAPSGHHQAGRCLLAACKACKRKTTNADRRKAATMRERR
RLSKVNEAAETLKRCTSSNPNQRLPKVEILRNAIRAIEGLQALLRDQDAAPPGAAAAAAAPG
PLPPGRGGEHASGDSDASSPRSNCSDGMMDASGPPSGARRRNCAEGAAANEAPSEPRP
GKSAAVSSLDCLSSIVERISTESPAAPALLLADVPSESPPRRQEAAAPSEGESSGDPTQSPD
AAPQCPAGANPNPIAQVL

MYOD1 AroPERFECT

M**F**ELLSPPLRDVDLTA**F**PDGSLCSATTDDDDP**Y**CDSPDLREDLDPRLM**F**HVGALLKPEEHS
HPA**F**AVHPAPGAREDEHVR**F**APSGHHQAGRCLLAC**F**KACKRKTTNADRRKAATMRERRRL

SKVNEA**F**ETLKRCTSSNPNQRLPKVEILRNAIR**Y**IEGLQALLRDQDAAPPGAAAAAPGPLPP
GRGGE**W**HSGDSDASSPRSNCS**F**DGMMDSGPPSGARRR**Y**NCEGANEAPSEPRPG**Y**KSA
AVSSLDCLSSIV**Y**ERISTESPAAPALLL**Y**ADVPSESPPRRQEAA**Y**APSEGESSGDPTQSP**Y**D
AAPQCPAGANPNPI**Y**QVL

MYOD1 IDR wild type N
MELLSPPLRDVDLTAPDGSLCS**F**ATTDD**FY**DDPC**F**DSPDLR**FF**EDLDPRLMHVGALLKPEE
HSH**F**PAAVHPAPGAREDEHVRAPSGHHQAGRCLL**W**ACKAC

MYOD1 IDR AroLITE A N
MELLSPPLRDVDLTAPDGSLCSAATTDDAADDPCADSPDLRAAEDLDPRLMHVGALLKPEE
HSHAPAAVHPAPGAREDEHVRAPSGHHQAGRCLLAACKAC

MYOD1 IDR AroPERFECT N
M**F**ELLSPPLRDVDLTA**F**PDGSLCSATTDDDDP**Y**CDSPDLREDLDPRLM**F**HVGALLKPEEHS
HPA**F**AVHPAPGAREDEHVR**F**APSGHHQAGRCLLAC**F**KAC

MYOD1 IDR wild type C
AAAA**FY**APGPLPPGRGGEH**Y**SGDSDASSPRSNCSDGMMD**Y**SGPPSGARRRNC**Y**EGA**YY**
NEAPSEPRPGKSAAVSSLDCLSSIVERISTESPAAPALLLADVPSESPPRRQEAAAPSEGES
SGDPTQSPDAAPQCPAGANPNPI**Y**QVL

MYOD1 IDR AroLITE A C
AAAAAAAPGPLPPGRGGEHASGDSDASSPRSNCSDGMMDASGPPSGARRRNCAEGAAA
NEAPSEPRPGKSAAVSSLDCLSSIVERISTESPAAPALLLADVPSESPPRRQEAAAPSEGES
SGDPTQSPDAAPQCPAGANPNPIAQVL

MYOD1 IDR AroPERFECT C
AAAAAPGPLPPGRGGE**W**HSGDSDASSPRSNCS**F**DGMMDSGPPSGARRR**Y**NCEGANEAP
SEPRPG**Y**KSAAVSSLDCLSSIV**Y**ERISTESPAAPALLL**Y**ADVPSESPPRRQEAA**Y**APSEGES
SGDPTQSP**Y**DAAPQCPAGANPNPI**Y**QVL

OCT4 IDR wild type N
MAGHLASD**F**A**F**SPPPGGGDGSAGLEPG**W**VDPRT**W**LS**F**QGPPGGPGIGPGSEVLGISPCP
PA**Y**E**F**CGGMA**Y**CGPQVGLGLVPQVGVETLQPEGQAGARVESNSEGTSSEPCADRPNAVK
LEKVEPTPEESQDMKALQKELEQ

OCT4 IDR wild type C

KGKRSSIE**W**SQREE**Y**EATGTP**F**PGGAVS**F**PLPPGPH**F**GTPG**Y**GSPH**F**TTL**Y**SVP**F**PEGEA **F**PSVPVTALGSPMHSN

OCT4 IDR AroLITE N

MAGHLASDAAASPPPGGGDGSAGLEPGAVDPRTALSAQGPPGGPGIGPGSEVLGISPCPP AAEACGGMAACGPQVGLGLVPQVGVETLQPEGQAGARVESNSEGTSSEPCADRPNAVKL EKVEPTPEESQDMKALQKELEQ

OCT4 IDR AroLITE C

KGKRSSIEASQREEAEATGTPAPGGAVSAPLPPGPHAGTPGAGSPHATTLASVPAPEGEA APSVPVTALGSPMHSN

OCT4 IDR AroPERFECT N

MAGHLASD**F**ASPPPGGGDGSAGL**F**EPGVDPRTLSQGPP**W**GGPGIGPGSEVLGI**W**SPCPP AECGGMACG**F**PQVGLGLVPQVGVE**Y**TLQPEGQAGARVES**F**NSEGTSSEPCADRP**Y**NAVK LEKVEPTPEE**F**SQDMKALQKELEQ

OCT4 IDR AroPERFECT C

KGKRSSIE**Y**SQREEE**Y**ATGTPP**F**GGAVSP**F**LPPGPH**F**GTPGGS**Y**PHTTLS**F**VPPEGE**Y**APS VPV**F**TALGSP**F**MHSN

PDX1 IDR wild type

MNGEEQ**YY**AATQL**Y**KDPCA**F**QRGPAPE**F**SASPPACL**Y**MGRQPPPPPPHP**F**PGALGALEQ GSPPDISP**Y**EVPPLADDPAVAHLHHHLPAQLALPHPPAGP**F**PEGAEPGVLEEPNRVQLP

PDX1 IDR AroLITE

MNGEEQAAAATQLAKDPCAAQRGPAPEASASPPACLAMGRQPPPPPPHPAPGALGALEQ GSPPDISPAEVPPLADDPAVAHLHHHLPAQLALPHPPAGPAPEGAEPGVLEEPNRVQLP

PDX1 IDR AroPERFECT

MN**Y**GEEQAATQLKDPC**Y**AQRGPAPESASPP**Y**ACLMGRQPPPPPP**F**HPPGALGALEQGS**F** PPDISPEVPPLAD**Y**DPAVAHLHHHLPA**F**QLALPHPPAGPPE**Y**GAEPGVLEEPNRV**F**QLP**F**

FOXA3 IDR wild type C

RRQKR**F**KLEEKVKKGGSGAATTTRNGTGSAASTTTPAATVTSPPQPPPPAPEPEAQGGED
VGALDCGSPASSTP**Y****F**TGLELPGELKLDAP**Y**N**F**NHP**F**SINNLMSEQTPAPPKLDVG**F**GG**Y**G
AEGGEPGV**YY**QGL**Y**SRSLLNAS

FOXA3 IDR AroLITE C
RRQKRAKLEEKVKKGGSGAATTTRNGTGSAASTTTPAATVTSPPQPPPPAPEPEAQGGED
VGALDCGSPASSTPAATGLELPGELKLDAPANANHPASINNLMSEQTPAPPKLDVGAGGAG
AEGGEPGVAAQGLASRSLLNAS

FOXA3 IDR AroPERFECT C
RRQKR**F**KLEEKVKKGGSG**Y**AATTTRNGTGSA**F**ASTTTPAATVTS**Y**PPQPPPPAPEPE**F**AQ
GGEDVGALDC**F**GSPASSTPTGLE**F**LPGELKLDAPNN**Y**HPSINNLMSEQT**Y**PAPPKLDVGG
GG**Y**AEGGEPGVQGLS**Y**RSLLNAS

S6Y AroPATCHY1
MSGSSSGSSGGSSSSGSSGSGGSSS**YYYYYYYYYYYYYY**SSGSGSSGSSSGGSSSGSS
GSSSGSSGGSSSSGSSGSGGSSSSSGSGSSGSSSGGSSSGS

S6Y AroPATCHY3
MSGSS**YYYY**SGSSGGSS**YYYY**SSGSSGSGGSSSSSGSGSSGSSSG**YYYY**GSSSGSSGS
SSGSSGGSSSSGSSGSGGSSSSSGSGSSGSSSGGSSSGSSS

S6Y AroPERFECT
MSGSSSG**Y**SSGGSS**Y**SSGSSG**Y**SGGSSS**Y**SSGSGS**Y**SGSSSG**Y**GSSSGS**Y**SGSSSG**Y**S
SGGSS**Y**SSGSSG**Y**SGGSSS**Y**SSGSGS**Y**SGSSSG**Y**GSSSGS**Y**

D6Y AroPERFECT
MSDDDSG**Y**SSDGDS**Y**SDDSDG**Y**SDDSDS**Y**DSDSDS**Y**DGSDDG**Y**GDDSDS**Y**DGDDSG**Y**S
DGDDS**Y**SSDDDG**Y**SDGDSD**Y**SDDDGS**Y**DGDSDG**Y**GSDDGD**Y**

For Figure 13d, the following sequences were used:

HOXD4 IDR Wild type
MVMSS**Y**MVNSK**Y**VDPK**F**PPCEE**Y**LQGG**Y**LGEQGAD**YY**GGGAQGAD**F**QPPGL**Y**PRPD**F**G
EQP**F**GGSGPGPGSALPARGHGQEPGGPGGH**Y**AAPGEPCPAPPAPPPAPLPGARA**Y**SQSD
PKQPPSGTALKQPAVV**Y**P**W**MKKVVSRGP**Y**SIVSPKC

HOXD4 IDR AroLITE A

MVMSSAMVNSKAVDPKAPPCEEALQGGALGEQGADAAGGGAQGADAQPPGLAPRPDAG
EQPAGGSGPGPGSALPARGHGQEPGGPGGHAAAPGEPCPAPPAPPPAPLPGARAASQSD
PKQPPSGTALKQPAVVAPAMKKVVSRGP**Y**SIVSPKC

HOXD4 IDR AroLITE S

MVMSSSMVNSKSVDPKSPPCEESLQGGSLGEQGADSSGGGAQGADSQPPGLSPRPDSG
EQPSGGSGPGPGSALPARGHGQEPGGPGGHSAAPGEPCPAPPAPPPAPLPGARASSQSD
PKQPPSGTALKQPAVVSPSMKKVVSRGP**Y**SIVSPKC

HOXD4 IDR AroLITE G

MVMSSGMVNSKGVDPKGPPCEEGLQGGGLGEQGADGGGGGAQGADGQPPGLGPRPD
GGEQPGGGSGPGPGSALPARGHGQEPGGPGGHGAAPGEPCPAPPAPPPAPLPGARAGS
QSDPKQPPSGTALKQPAVVGPGMKKVVSRGP**Y**SIVSPKC

References

1. Roeder RG, Rutter WJ. Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. *Nature*. 1969;224(5216):234-237. doi:10.1038/224234A0
2. Roeder RG, Rutter WJ. Specific nucleolar and nucleoplasmic RNA polymerases. *Proc Natl Acad Sci U S A*. 1970;65(3):675-682. doi:10.1073/PNAS.65.3.675
3. Pennetier S, Uzbekova S, Perreau C, Papillier P, Mermillod P, Dalbiès-Tran R. Spatio-temporal expression of the germ cell marker genes MATER, ZAR1, GDF9, BMP15,andVASA in adult bovine tissues, oocytes, and preimplantation embryos. *Biol Reprod*. 2004;71(4):1359-1366. doi:10.1095/BIOLREPROD.104.030288
4. Pangas SA, Rajkovic A. Transcriptional regulation of early oogenesis: in search of masters. *Hum Reprod Update*. 2006;12(1):65-76. doi:10.1093/HUMUPD/DMI033
5. Hawley DK, Mcclure WR. Compilation and analysis of Escherichia coli promoter DNA sequences. *Nucleic Acids Res*. 1983;11(8):2237-2255. doi:10.1093/NAR/11.8.2237
6. Estrem ST, Ross W, Gaal T, et al. Bacterial promoter architecture: subsite structure of UP elements and interactions with the carboxy-terminal domain of the RNA polymerase alpha subunit. *Genes Dev*. 1999;13(16):2134-2147. doi:10.1101/GAD.13.16.2134
7. Dombroski AJ, Walter WA, Record MT, Slegele DA, Gross CA. Polypeptides containing highly conserved regions of transcription initiation factor sigma 70 exhibit specificity of binding to promoter DNA. *Cell*. 1992;70(3):501-512. doi:10.1016/0092-8674(92)90174-B
8. Ross W, Gosink KK, Salomon J, et al. A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase. *Science*. 1993;262(5138):1407-1413. doi:10.1126/SCIENCE.8248780
9. Rao L, Ross W, Appleman JA, et al. Factor independent activation of rrnB P1. An “extended” promoter with an upstream element that dramatically increases promoter strength. *J Mol Biol*. 1994;235(5):1421-1435. doi:10.1006/JMBI.1994.1098
10. 100 years of biochemistry and molecular biology. The decade-long pursuit of a reconstituted yeast transcription system: the work of Roger D. Kornberg - PubMed. https://pubmed.ncbi.nlm.nih.gov/19847957/. Accessed April 9, 2024.
11. Cooper GM. Regulation of Transcription in Eukaryotes. 2000.
12. Tong Ihn Lee, Young RA. Transcription of eukaryotic protein-coding genes. *Annu Rev Genet*. 2000;34:77-137. doi:10.1146/ANNUREV.GENET.34.1.77
13. Resolution of factors required for the initiation of transcription by yeast RNA polymerase II - PubMed. https://pubmed.ncbi.nlm.nih.gov/2193032/. Accessed April 9, 2024.

14. Hampsey M. Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiol Mol Biol Rev*. 1998;62(2):465-503. doi:10.1128/MMBR.62.2.465-503.1998
15. Haberle V, Stark A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol*. 2018;19(10):621. doi:10.1038/S41580-018-0028-8
16. ALLFREY VG, LITTAU VC, MIRSKY AE. On the role of of histones in regulation ribonucleic acid synthesis in the cell nucleus. *Proc Natl Acad Sci U S A*. 1963;49(3):414-421. doi:10.1073/PNAS.49.3.414
17. ALLFREY VG, FAULKNER R, MIRSKY AE. ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS. *Proc Natl Acad Sci U S A*. 1964;51(5):786-794. doi:10.1073/PNAS.51.5.786
18. Dong X, Weng Z. The correlation between histone modifications and gene expression. *Epigenomics*. 2013;5(2):113-116. doi:10.2217/EPI.13.13
19. Creyghton MP, Cheng AW, Welstead GG, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A*. 2010;107(50):21931-21936. doi:10.1073/PNAS.1016071107/SUPPL_FILE/ST01.XLSX
20. Devaiah BN, Case-Borden C, Gegonne A, et al. BRD4 is a histone acetyltransferase that evicts nucleosomes from chromatin. *Nat Struct Mol Biol*. 2016;23(6):540-548. doi:10.1038/NSMB.3228
21. Dey A, Chitsaz F, Abbasi A, Misteli T, Ozato K. The double bromodomain protein Brd4 binds to acetylated chromatin during interphase and mitosis. *Proc Natl Acad Sci U S A*. 2003;100(15):8758. doi:10.1073/PNAS.1433065100
22. Gonzales-Cope M, Sidoli S, Bhanu N V., Won KJ, Garcia BA. Histone H4 acetylation and the epigenetic reader Brd4 are critical regulators of pluripotency in embryonic stem cells. *BMC Genomics*. 2016;17(1). doi:10.1186/S12864-016-2414-Y
23. Itzen F, Greifenberg AK, Bösken CA, Geyer M. Brd4 activates P-TEFb for RNA polymerase II CTD phosphorylation. *Nucleic Acids Res*. 2014;42(12):7577-7590. doi:10.1093/NAR/GKU449
24. Zabidi MA, Stark A. Regulatory enhancer–core-promoter communication via transcription factors and cofactors. *Trends Genet*. 2016;32(12):801. doi:10.1016/J.TIG.2016.10.003
25. Banerji J, Rusconi S, Schaffner W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*. 1981;27(2 Pt 1):299-308. doi:10.1016/0092-8674(81)90413-X

26. Ling JQ, Li T, Hu JF, et al. CTCF mediates interchromosomal colocalization between Igf2/H19 and Wsb1/Nf1. *Science*. 2006;312(5771):269-272. doi:10.1126/SCIENCE.1123191
27. Simonis M, Klous P, Splinter E, et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet*. 2006;38(11):1348-1354. doi:10.1038/NG1896
28. Fullwood MJ, Liu MH, Pan YF, et al. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*. 2009;462(7269):58-64. doi:10.1038/NATURE08497
29. Lieberman-Aiden E, Van Berkum NL, Williams L, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. 2009;326(5950):289-293. doi:10.1126/SCIENCE.1181369
30. Rubio ED, Reiss DJ, Welcsh PL, et al. CTCF physically links cohesin to chromatin. *Proc Natl Acad Sci U S A*. 2008;105(24):8309-8314. doi:10.1073/PNAS.0801273105
31. Weintraub H, Tapscott SJ, Davis RL, et al. Activation of muscle-specific genes in pigment, nerve, fat, liver, and fibroblast cell lines by forced expression of MyoD. *Proc Natl Acad Sci U S A*. 1989;86(14):5434-5438. doi:10.1073/PNAS.86.14.5434
32. Sanborn AL, Rao SSP, Huang SC, et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci U S A*. 2015;112(47):E6456-E6465. doi:10.1073/PNAS.1518552112
33. Davidson IF, Bauer B, Goetz D, Tang W, Wutz G, Peters JM. DNA loop extrusion by human cohesin. *Science*. 2019;366(6471):1338-1345. doi:10.1126/SCIENCE.AAZ3418
34. Zuin J, Roth G, Zhan Y, et al. Nonlinear control of transcription through enhancer–promoter interactions. *Nat 2022 6047906*. 2022;604(7906):571-577. doi:10.1038/s41586-022-04570-y
35. Robson MI, Ringel AR, Mundlos S. Regulatory Landscaping: How Enhancer-Promoter Communication Is Sculpted in 3D. *Mol Cell*. 2019;74(6):1110-1122. doi:10.1016/J.MOLCEL.2019.05.032
36. Ramasamy S, Aljahani A, Karpinska MA, et al. The Mediator complex regulates enhancer-promoter interactions. *Nat Struct Mol Biol*. 2023;30(7):991-1000. doi:10.1038/S41594-023-01027-2
37. Soutourina J. Transcription regulation by the Mediator complex. *Nat Rev Mol Cell Biol*. 2018;19(4):262-274. doi:10.1038/NRM.2017.115
38. Hnisz D, Shrinivas K, Young RA, Chakraborty AK, Sharp PA. A Phase Separation Model for Transcriptional Control. *Cell*. 2017;169(1):13-23. doi:10.1016/J.CELL.2017.02.007
39. Sabari BR, Dall'Agnese A, Boija A, et al. Coactivator condensation at super-

enhancers links phase separation and gene control. *Science*. 2018;361(6400). doi:10.1126/SCIENCE.AAR3958

40. Basu S, Mackowiak SD, Niskanen H, et al. Unblending of Transcriptional Condensates in Human Repeat Expansion Disease. *Cell*. 2020;181(5):1062-1079.e30. doi:10.1016/J.CELL.2020.04.018
41. Asimi V, Sampath Kumar A, Niskanen H, et al. Hijacking of transcriptional condensates by endogenous retroviruses. *Nat Genet 2022 548*. 2022;54(8):1238-1247. doi:10.1038/s41588-022-01132-w
42. Boija A, Klein IA, Sabari BR, et al. Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell*. 2018;175(7):1842-1855.e16. doi:10.1016/J.CELL.2018.10.042
43. Shrinivas K, Sabari BR, Coffey EL, et al. Enhancer Features that Drive Formation of Transcriptional Condensates. *Mol Cell*. 2019;75(3):549-561.e7. doi:10.1016/J.MOLCEL.2019.07.009
44. Guo YE, Manteiga JC, Henninger JE, et al. Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. *Nature*. 2019;572(7770):543-548. doi:10.1038/S41586-019-1464-0
45. Henninger JE, Oksuz O, Shrinivas K, et al. RNA-Mediated Feedback Control of Transcriptional Condensates. *Cell*. 2021;184(1):207-225.e24. doi:10.1016/J.CELL.2020.11.030
46. Zamudio A V., Dall'Agnese A, Henninger JE, et al. Mediator Condensates Localize Signaling Factors to Key Cell Identity Genes. *Mol Cell*. 2019;76(5):753-766.e6. doi:10.1016/J.MOLCEL.2019.08.016
47. Cho WK, Spille JH, Hecht M, et al. Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science (80-)*. 2018;361(6400):412-415. doi:10.1126/SCIENCE.AAR4199/SUPPL_FILE/AAR4199_S3.MOV
48. Shin Y, Brangwynne CP. Liquid phase condensation in cell physiology and disease. *Science*. 2017;357(6357). doi:10.1126/SCIENCE.AAF4382
49. Garde S. Physical chemistry: Hydrophobic interactions in context. *Nature*. 2015;517(7534):277-279. doi:10.1038/517277A
50. Alberti S, Gladfelter A, Mittag T. Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates. *Cell*. 2019;176(3):419. doi:10.1016/J.CELL.2018.12.035
51. Banani SF, Lee HO, Hyman AA, Rosen MK. Biomolecular condensates: organizers of cellular biochemistry. *Nat Rev Mol Cell Biol*. 2017;18(5):285-298. doi:10.1038/NRM.2017.7

52. Li P, Banjade S, Cheng HC, et al. Phase transitions in the assembly of multivalent signalling proteins. *Nat 2012 4837389*. 2012;483(7389):336-340. doi:10.1038/nature10879
53. Patel A, Lee HO, Jawerth L, et al. A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. *Cell*. 2015;162(5):1066-1077. doi:10.1016/j.cell.2015.07.047
54. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol*. 2005;6(3):197-208. doi:10.1038/NRM1589
55. Gonçalves-Kulik M, Mier P, Kastano K, et al. Low Complexity Induces Structure in Protein Regions Predicted as Intrinsically Disordered. *Biomolecules*. 2022;12(8). doi:10.3390/BIOM12081098
56. Choi JM, Dar F, Pappu R V. LASSI: A lattice model for simulating phase transitions of multivalent proteins. *PLoS Comput Biol*. 2019;15(10). doi:10.1371/JOURNAL.PCBI.1007028
57. Pederson T. The nucleolus. *Cold Spring Harb Perspect Biol*. 2011;3(3):1-15. doi:10.1101/CSHPERSPECT.A000638
58. Brangwynne CP, Mitchison TJ, Hyman AA. Active liquid-like behavior of nucleoli determines their size and shape in Xenopus laevis oocytes. *Proc Natl Acad Sci U S A*. 2011;108(11):4334-4339. doi:10.1073/PNAS.1017150108
59. Dubois ML, Boisvert FM. The Nucleolus: Structure and Function. *Funct Nucl*. January 2016:29. doi:10.1007/978-3-319-38882-3_2
60. Ferrolino MC, Mitrea DM, Michael JR, Kriwacki RW. Compositional adaptability in NPM1-SURF6 scaffolding networks enabled by dynamic switching of phase separation mechanisms. *Nat Commun*. 2018;9(1). doi:10.1038/S41467-018-07530-1
61. Maiser A, Dillinger S, Längst G, Schermelleh L, Leonhardt H, Németh A. Super-resolution in situ analysis of active ribosomal DNA chromatin organization in the nucleolus. *Sci Rep*. 2020;10(1). doi:10.1038/S41598-020-64589-X
62. Banani SF, Rice AM, Peeples WB, et al. Compositional Control of Phase-Separated Cellular Bodies. *Cell*. 2016;166(3):651-663. doi:10.1016/J.CELL.2016.06.010
63. Keenen MM, Brown D, Brennan LD, et al. HP1 proteins compact DNA into mechanically and positionally stable phase separated domains. *Elife*. 2021;10. doi:10.7554/ELIFE.64563
64. Xu S, Lai SK, Sim DY, Ang WSL, Li HY, Roca X. SRRM2 organizes splicing condensates to regulate alternative splicing. *Nucleic Acids Res*. 2022;50(15):8599. doi:10.1093/NAR/GKAC669
65. Hampoelz B, Schwarz A, Ronchi P, et al. Nuclear Pores Assemble from Nucleoporin Condensates During Oogenesis. *Cell*. 2019;179(3):671-686.e17.

doi:10.1016/J.CELL.2019.09.022
66. Feric M, Brangwynne CP. A nuclear F-actin scaffold stabilizes ribonucleoprotein droplets against gravity in large cells. *Nat Cell Biol*. 2013;15(10):1253-1259. doi:10.1038/NCB2830
67. Patil A, Strom AR, Paulo JA, et al. A disordered region controls cBAF activity via condensation and partner recruitment. *Cell*. 2023;186(22):4936-4955.e26. doi:10.1016/J.CELL.2023.08.032
68. Du M, Stitzinger SH, Spille JH, et al. Direct observation of a condensate effect on super-enhancer controlled gene bursting. *Cell*. 2024;187(2):331-344.e17. doi:10.1016/J.CELL.2023.12.005
69. Martin EW, Holehouse AS, Peran I, et al. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science*. 2020;367(6478):694-699. doi:10.1126/SCIENCE.AAW8653
70. Jonas F, Carmi M, Krupkin B, et al. The molecular grammar of protein disorder guiding genome-binding locations. *Nucleic Acids Res*. 2023;51(10):4831-4844. doi:10.1093/NAR/GKAD184
71. Shrinivas K, Sabari BR, Coffey EL, et al. Enhancer Features that Drive Formation of Transcriptional Condensates. *Mol Cell*. 2019;75(3):549-561.e7. doi:10.1016/J.MOLCEL.2019.07.009
72. Zhu L, Huq E. Mapping functional domains of transcription factors. *Methods Mol Biol*. 2011;754:167-184. doi:10.1007/978-1-61779-154-3_9
73. Lambert SA, Jolma A, Campitelli LF, et al. The Human Transcription Factors. *Cell*. 2018;172(4):650-665. doi:10.1016/J.CELL.2018.01.029
74. Harrison SC. A structural taxonomy of DNA-binding domains. *Nature*. 1991;353(6346):715-719. doi:10.1038/353715A0
75. Struhl K. The DNA-binding domains of the jun oncoprotein and the yeast GCN4 transcriptional activator protein are functionally homologous. *Cell*. 1987;50(6):841-846. doi:10.1016/0092-8674(87)90511-3
76. Nitta KR, Jolma A, Yin Y, et al. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *Elife*. 2015;4(4). doi:10.7554/ELIFE.04837
77. Rowe HM, Jakobsson J, Mesnard D, et al. KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature*. 2010;463(7278):237-240. doi:10.1038/NATURE08674
78. Yang P, Wang Y, Macfarlan TS. The role of KRAB-ZFPs in transposable element repression and mammalian evolution. *Trends Genet*. 2017;33(11):871. doi:10.1016/J.TIG.2017.08.006
79. Friedman AD, McKnight SL. Identification of two polypeptide segments of

CCAAT/enhancer-binding protein required for transcriptional activation of the serum albumin gene. *Genes Dev*. 1990;4(8):1416-1426. doi:10.1101/GAD.4.8.1416

80. Amati B, Brooks MW, Levy N, Littlewood TD, Evan GI, Land H. Oncogenic activity of the c-Myc protein requires dimerization with Max. *Cell*. 1993;72(2):233-245. doi:10.1016/0092-8674(93)90663-B
81. Massari ME, Murre C. Helix-Loop-Helix Proteins: Regulators of Transcription in Eucaryotic Organisms. *Mol Cell Biol*. 2000;20(2):429. doi:10.1128/MCB.20.2.429-440.2000
82. Cirillo LA, McPherson CE, Bossard P, et al. Binding of the winged-helix transcription factor HNF3 to a linker histone site on the nucleosome. *EMBO J*. 1998;17(1):244-254. doi:10.1093/EMBOJ/17.1.244
83. Barral A, Zaret KS. Pioneer factors: roles and their regulation in development. *Trends Genet*. 2024;40(2):134-148. doi:10.1016/J.TIG.2023.10.007
84. Lee CS, Friedman JR, Fulmer JT, Kaestner KH. The initiation of liver development is dependent on Foxa transcription factors. *Nature*. 2005;435(7044):944-947. doi:10.1038/NATURE03649
85. Soufi A, Donahue G, Zaret KS. Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. *Cell*. 2012;151(5):994-1004. doi:10.1016/J.CELL.2012.09.045
86. Pan Y, Tsai CJ, Ma B, Nussinov R. Mechanisms of transcription factor selectivity. *Trends Genet*. 2010;26(2):75. doi:10.1016/J.TIG.2009.12.003
87. Georges AB, Benayoun BA, Caburet S, Veitia RA. Generic binding sites, generic DNA-binding domains: where does specific promoter recognition come from? *FASEB J*. 2010;24(2):346-356. doi:10.1096/FJ.09-142117
88. Badis G, Berger MF, Philippakis AA, et al. Diversity and complexity in DNA recognition by transcription factors. *Science*. 2009;324(5935):1720-1723. doi:10.1126/SCIENCE.1162327
89. Jonas F, Carmi M, Krupkin B, et al. The molecular grammar of protein disorder guiding genome-binding locations. *Nucleic Acids Res*. 2023;51(10):4831-4844. doi:10.1093/NAR/GKAD184
90. Kumar DK, Jonas F, Jana T, Brodsky S, Carmi M, Barkai N. Complementary strategies for directing in vivo transcription factor binding through DNA binding domains and intrinsically disordered regions. *Mol Cell*. 2023;83(9):1462-1473.e5. doi:10.1016/J.MOLCEL.2023.04.002
91. Ma J, Ptashne M. Deletion analysis of GAL4 defines two transcriptional activating segments. *Cell*. 1987;48(5):847-853. doi:10.1016/0092-8674(87)90081-X
92. Sadowski I, Ma J, Triezenberg S, Ptashne M. GAL4-VP16 is an unusually potent

transcriptional activator. *Nature*. 1988;335(6190):563-564. doi:10.1038/335563A0

93. Piskacek M, Havelka M, Rezacova M, Knight A. The 9aaTAD Transactivation Domains: From Gal4 to p53. *PLoS One*. 2016;11(9). doi:10.1371/JOURNAL.PONE.0162842
94. Gilbert LA, Horlbeck MA, Adamson B, et al. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell*. 2014;159(3):647-661. doi:10.1016/J.CELL.2014.09.029
95. Staller M V., Holehouse AS, Swain-Lenz D, Das RK, Pappu R V., Cohen BA. A High-Throughput Mutational Scan of an Intrinsically Disordered Acidic Transcriptional Activation Domain. *Cell Syst*. 2018;6(4):444-455.e6. doi:10.1016/J.CELS.2018.01.015
96. Sanborn AL, Yeh BT, Feigerle JT, et al. Simple biochemical features underlie transcriptional activation domain diversity and dynamic, fuzzy binding to Mediator. *Elife*. 2021;10. doi:10.7554/ELIFE.68068
97. Erijman A, Kozlowski L, Sohrabi-Jahromi S, et al. A High-Throughput Screen for Transcription Activation Domains Reveals Their Sequence Features and Permits Prediction by Deep Learning. *Mol Cell*. 2020;79(6):1066. doi:10.1016/J.MOLCEL.2020.08.013
98. Tycko J, DelRosso N, Hess GT, et al. High-Throughput Discovery and Characterization of Human Transcriptional Effectors. *Cell*. 2020;183(7):2020-2035.e16. doi:10.1016/J.CELL.2020.11.024
99. Alerasool N, Leng H, Lin ZY, Gingras AC, Taipale M. Identification and functional characterization of transcriptional activators in human cells. *Mol Cell*. 2022;82(3):677-695.e7. doi:10.1016/J.MOLCEL.2021.12.008
100. DelRosso N, Tycko J, Suzuki P, et al. Large-scale mapping and mutagenesis of human transcriptional effector domains. *Nature*. 2023;616(7956):365-372. doi:10.1038/S41586-023-05906-Y
101. Zhou Q, Liu M, Xia X, et al. A mouse tissue transcription factor atlas. *Nat Commun*. 2017;8. doi:10.1038/NCOMMS15089
102. Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet*. 2012;13(9):613-626. doi:10.1038/NRG3207
103. Rodda DJ, Chew JL, Lim LH, et al. Transcriptional regulation of nanog by OCT4 and SOX2. *J Biol Chem*. 2005;280(26):24731-24737. doi:10.1074/JBC.M502573200
104. Loh YH, Wu Q, Chew JL, et al. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet*. 2006;38(4):431-440. doi:10.1038/NG1760
105. Boyer LA, Tong IL, Cole MF, et al. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*. 2005;122(6):947-956. doi:10.1016/J.CELL.2005.08.020

106. Chew J-L, Loh Y-H, Zhang W, et al. Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells. *Mol Cell Biol*. 2005;25(14):6031-6046. doi:10.1128/MCB.25.14.6031-6046.2005
107. Hough SR, Clements I, Welch PJ, Wiederholt KA. Differentiation of mouse embryonic stem cells after RNA interference-mediated silencing of OCT4 and Nanog. *Stem Cells*. 2006;24(6):1467-1475. doi:10.1634/STEMCELLS.2005-0475
108. Kiecker C, Bates T, Bell E. Molecular specification of germ layers in vertebrate embryos. *Cell Mol Life Sci*. 2016;73(5):923-947. doi:10.1007/S00018-015-2092-Y
109. Tsankov AM, Gu H, Akopian V, et al. Transcription factor binding dynamics during human ES cell differentiation. *Nature*. 2015;518(7539):344-349. doi:10.1038/NATURE14233
110. Yi S, Huang X, Zhou S, et al. E2A regulates neural ectoderm fate specification in human embryonic stem cells. *Development*. 2020;147(23). doi:10.1242/DEV.190298
111. Pham PD, Lu H, Han H, et al. Transcriptional network governing extraembryonic endoderm cell fate choice. *Dev Biol*. 2023;502:20-37. doi:10.1016/J.YDBIO.2023.07.002
112. Lee KW, Yeo SY, Gong JR, et al. PRRX1 is a master transcription factor of stromal fibroblasts for myofibroblastic lineage progression. *Nat Commun*. 2022;13(1). doi:10.1038/S41467-022-30484-4
113. Mammalian hepatocyte differentiation requires the transcription factor HNF-4alpha - PubMed. https://pubmed.ncbi.nlm.nih.gov/10691738/. Accessed April 9, 2024.
114. Hulme AJ, Maksour S, St-Clair Glover M, Miellet S, Dottori M. Making neurons, made easy: The use of Neurogenin-2 in neuronal differentiation. *Stem cell reports*. 2022;17(1):14-34. doi:10.1016/J.STEMCR.2021.11.015
115. Schroeder TM, Jensen ED, Westendorf JJ. Runx2: a master organizer of gene transcription in developing and maturing osteoblasts. *Birth Defects Res C Embryo Today*. 2005;75(3):213-225. doi:10.1002/BDRC.20043
116. Basu A, Tiwari VK. Epigenetic reprogramming of cell identity: lessons from development for regenerative medicine. *Clin Epigenetics*. 2021;13(1). doi:10.1186/S13148-021-01131-4
117. Battistelli C, Garbo S, Maione R. MyoD-Induced Trans-Differentiation: A Paradigm for Dissecting the Molecular Mechanisms of Cell Commitment, Differentiation and Reprogramming. *Cells*. 2022;11(21). doi:10.3390/CELLS11213435
118. Xie H, Ye M, Feng R, Graf T. Stepwise reprogramming of B cells into macrophages. *Cell*. 2004;117(5):663-676. doi:10.1016/S0092-8674(04)00419-2
119. Son EY, Ichida JK, Wainger BJ, et al. Conversion of mouse and human fibroblasts into functional spinal motor neurons. *Cell Stem Cell*. 2011;9(3):205-218.

doi:10.1016/J.STEM.2011.07.014

120. Wang H, Yang Y, Liu J, Qian L. Direct cell reprogramming: approaches, mechanisms and progress. *Nat Rev Mol Cell Biol*. 2021;22(6):410. doi:10.1038/S41580-021-00335-Z
121. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*. 2006;126(4):663-676. doi:10.1016/J.CELL.2006.07.024
122. Takahashi K, Tanabe K, Ohnuki M, et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*. 2007;131(5):861-872. doi:10.1016/J.CELL.2007.11.019
123. Szabo E, Rampalli S, Risueño RM, et al. Direct conversion of human fibroblasts to multilineage blood progenitors. *Nature*. 2010;468(7323):521-526. doi:10.1038/NATURE09591
124. Sareen D, Gowing G, Sahabian A, et al. Human induced pluripotent stem cells are a novel source of neural progenitor cells (iNPCs) that migrate and integrate in the rodent spinal cord. *J Comp Neurol*. 2014;522(12):2707-2728. doi:10.1002/CNE.23578
125. De Peppo GM, Marcos-Campos I, Kahler DJ, et al. Engineering bone tissue substitutes from human induced pluripotent stem cells. *Proc Natl Acad Sci U S A*. 2013;110(21):8680-8685. doi:10.1073/PNAS.1301190110/-/DCSUPPLEMENTAL/PNAS.201301190SI.PDF
126. Li Y, Hermanson DL, Moriarity BS, Kaufman DS. Human iPSC-Derived Natural Killer Cells Engineered with Chimeric Antigen Receptors Enhance Anti-tumor Activity. *Cell Stem Cell*. 2018;23(2):181-192.e5. doi:10.1016/J.STEM.2018.06.002
127. Yanagimachi MD, Niwa A, Tanaka T, et al. Robust and highly-efficient differentiation of functional monocytic cells from human pluripotent stem cells under serum- and feeder cell-free conditions. *PLoS One*. 2013;8(4). doi:10.1371/JOURNAL.PONE.0059243
128. Hay DC, Zhao D, Fletcher J, et al. Efficient differentiation of hepatocytes from human embryonic stem cells exhibiting markers recapitulating liver development in vivo. *Stem Cells*. 2008;26(4):894-902. doi:10.1634/STEMCELLS.2007-0718
129. Maroof AM, Keros S, Tyson JA, et al. Directed differentiation and functional maturation of cortical interneurons from human embryonic stem cells. *Cell Stem Cell*. 2013;12(5):559-572. doi:10.1016/J.STEM.2013.04.008
130. Theka I, Caiazzo M, Dvoretskova E, et al. Rapid generation of functional dopaminergic neurons from human induced pluripotent stem cells through a single-step procedure using cell lineage transcription factors. *Stem Cells Transl Med*. 2013;2(6):473-479. doi:10.5966/SCTM.2012-0133

131. Pagliuca FW, Millman JR, Gürtler M, et al. Generation of functional human pancreatic β cells in vitro. *Cell*. 2014;159(2):428-439. doi:10.1016/J.CELL.2014.09.040
132. Qian L, Huang Y, Spencer CI, et al. In vivo reprogramming of murine cardiac fibroblasts into induced cardiomyocytes. *Nature*. 2012;485(7400):593-598. doi:10.1038/NATURE11044
133. Wang Y, Zheng Q, Sun Z, et al. Reversal of liver failure using a bioartificial liver device implanted with clinical-grade human-induced hepatocytes. *Cell Stem Cell*. 2023;30(5):617-631.e8. doi:10.1016/J.STEM.2023.03.013
134. Hogrebe NJ, Ishahak M, Millman JR. Developments in stem cell-derived islet replacement therapy for treating type 1 diabetes. *Cell Stem Cell*. 2023;30(5):530-548. doi:10.1016/J.STEM.2023.04.002
135. Necci M, Piovesan D, Hoque MT, et al. Critical assessment of protein intrinsic disorder prediction. *Nat Methods 2021 185*. 2021;18(5):472-481. doi:10.1038/s41592-021-01117-3
136. Ota M, Koike R, Amemiya T, et al. An assignment of intrinsically disordered regions of proteins based on NMR structures. *J Struct Biol*. 2013;181(1):29. doi:10.1016/J.JSB.2012.10.017
137. Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics*. 2015;31(6):857-863. doi:10.1093/BIOINFORMATICS/BTU744
138. Erdos G, Pajkos M, Dosztányi Z. IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Res*. 2021;49(W1):W297-W303. doi:10.1093/NAR/GKAB408
139. Xue B, Dunbrack RL, Williams RW, Dunker AK, Uversky VN. PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim Biophys Acta*. 2010;1804(4):996-1010. doi:10.1016/J.BBAPAP.2010.01.011
140. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583-589. doi:10.1038/S41586-021-03819-2
141. Emenecker RJ, Griffith D, Holehouse AS. Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure. *Biophys J*. 2021;120(20):4312-4319. doi:10.1016/J.BPJ.2021.08.039
142. Naderi J, Magalhaes AP, Kibar G, et al. An activity-specificity trade-off encoded in human transcription factors. *Nat Cell Biol 2024 268*. 2024;26(8):1309-1321. doi:10.1038/s41556-024-01411-0
143. Jack I, Seshadri R, Garson M, et al. RCH-ACV: A lymphoblastic leukemia cell line with chromosome translocation 1;19 and trisomy 8. *Cancer Genet Cytogenet*.

1986;19(3-4):261-269. doi:10.1016/0165-4608(86)90055-5

144. Schindelin J, Arganda-Carreras I, Frise E, et al. Fiji: an open-source platform for biological-image analysis. *Nat Methods 2012 97*. 2012;9(7):676-682. doi:10.1038/nmeth.2019
145. Stik G, Vidal E, Barrero M, et al. CTCF is dispensable for immune cell transdifferentiation but facilitates an acute inflammatory response. *Nat Genet 2020 527*. 2020;52(7):655-661. doi:10.1038/s41588-020-0643-0
146. Zhang Y, Pak CH, Han Y, et al. Rapid single-step induction of functional neurons from human pluripotent stem cells. *Neuron*. 2013;78(5):785-798. doi:10.1016/J.NEURON.2013.05.029
147. Schmidl C, Rendeiro AF, Sheffield NC, Bock C. ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nat Methods*. 2015;12(10):963-965. doi:10.1038/NMETH.3542
148. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods 2013 1012*. 2013;10(12):1213-1218. doi:10.1038/nmeth.2688
149. Bateman A, Martin MJ, Orchard S, et al. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res*. 2023;51(D1):D523-D531. doi:10.1093/NAR/GKAC1052
150. Holehouse AS, Das RK, Ahad JN, Richardson MOG, Pappu R V. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophys J*. 2017;112(1):16-21. doi:10.1016/J.BPJ.2016.11.3200
151. Wickham H. ggplot2. 2016. doi:10.1007/978-3-319-24277-4
152. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21. doi:10.1093/BIOINFORMATICS/BTS635
153. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12). doi:10.1186/S13059-014-0550-8
154. R: The R Project for Statistical Computing. https://www.r-project.org/. Accessed February 24, 2024.
155. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*. 2016;32(18):2847-2849. doi:10.1093/BIOINFORMATICS/BTW313
156. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*. 2008;24(5):719-720. doi:10.1093/BIOINFORMATICS/BTM563

157. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545-15550. doi:10.1073/PNAS.0506580102/SUPPL_FILE/06580FIG7.JPG
158. Sepulveda JL, Gkretsi V, Wu C. Assembly and signaling of adhesion complexes. *Curr Top Dev Biol*. 2005;68:183-225. doi:10.1016/S0070-2153(05)68007-6
159. Zheng GXY, Terry JM, Belgrader P, et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun*. 2017;8. doi:10.1038/NCOMMS14049
160. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. *Cell*. 2021;184(13):3573-3587.e29. doi:10.1016/J.CELL.2021.04.048
161. Choi J, Lysakovskaia K, Stik G, et al. Evidence for additive and synergistic action of mammalian enhancers during cell fate determination. *Elife*. 2021;10. doi:10.7554/ELIFE.65381
162. La Manno G, Soldatov R, Zeisel A, et al. RNA velocity of single cells. *Nat 2018 5607719*. 2018;560(7719):494-498. doi:10.1038/s41586-018-0414-6
163. Bergen V, Lange M, Peidli S, Wolf FA, Theis FJ. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol*. 2020;38(12):1408-1414. doi:10.1038/S41587-020-0591-3
164. Wolf FA, Hamey FK, Plass M, et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol*. 2019;20(1):1-9. doi:10.1186/S13059-019-1663-X
165. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760. doi:10.1093/BIOINFORMATICS/BTP324
166. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *Gigascience*. 2021;10(2). doi:10.1093/GIGASCIENCE/GIAB008
167. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-1303. doi:10.1101/GR.107524.110
168. Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 2008;9(9). doi:10.1186/GB-2008-9-9-R137
169. Perez RB, Tischer A, Auton M, Whitten ST. Alanine and proline content modulate global sensitivity to discrete perturbations in disordered proteins. *Proteins*. 2014;82(12):3373-3384. doi:10.1002/PROT.24692
170. Uhlén M, Fagerberg L, Hallström BM, et al. Proteomics. Tissue-based map of the human proteome. *Science*. 2015;347(6220). doi:10.1126/SCIENCE.1260419
171. Urrutia R. KRAB-containing zinc-finger repressor proteins. *Genome Biol*. 2003;4(10).

doi:10.1186/GB-2003-4-10-231

172. Lancaster AK, Nutter-Upham A, Lindquist S, King OD. PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition. *Bioinformatics*. 2014;30(17):2501. doi:10.1093/BIOINFORMATICS/BTU310
173. Udupa A, Kotha SR, Staller M V. Commonly asked questions about transcriptional activation domains. *Curr Opin Struct Biol*. 2024;84. doi:10.1016/J.SBI.2023.102732
174. Morgan R, In Der Rieden P, Hooiveld MHW, Durston AJ. Identifying HOX paralog groups by the PBX-binding region. *Trends Genet*. 2000;16(2):66-67. doi:10.1016/S0168-9525(99)01881-8
175. Kmita M, Van Der Hoeven F, Zákány J, Krumlauf R, Duboule D. Mechanisms of Hox gene colinearity: transposition of the anterior Hoxb1 gene into the posterior HoxD complex. *Genes Dev*. 2000;14(2):198. doi:10.1101/gad.14.2.198
176. Pöpperl H, Bienz M, Studer M, et al. Segmental expression of Hoxb-1 is controlled by a highly conserved autoregulatory loop dependent upon exd/pbx. *Cell*. 1995;81(7):1031-1042. doi:10.1016/S0092-8674(05)80008-X
177. Popperl H, Featherstone1 MS. An autoregulatory element of the murine Hox-4.2 gene. *EMBO J*. 1992;11(10):3673. doi:10.1002/J.1460-2075.1992.TB05452.X
178. Chong S, Dugast-Darzacq C, Liu Z, et al. Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science*. 2018;361(6400). doi:10.1126/SCIENCE.AAR2555
179. Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. *Cell*. 2013;152(6):1237-1251. doi:10.1016/J.CELL.2013.02.014
180. Kim SY, Han YM, Oh M, et al. DUSP4 regulates neuronal differentiation and calcium homeostasis by modulating ERK1/2 phosphorylation. *Stem Cells Dev*. 2015;24(6):686-700. doi:10.1089/SCD.2014.0434
181. de Martin X, Sodaei R, Santpere G. Mechanisms of Binding Specificity among bHLH Transcription Factors. *Int J Mol Sci*. 2021;22(17). doi:10.3390/IJMS22179150
182. Hewitt J, Lu X, Gilbert L, Nanes MS. The muscle transcription factor MyoD promotes osteoblast differentiation by stimulation of the Osterix promoter. *Endocrinology*. 2008;149(7):3698-3707. doi:10.1210/EN.2007-1556
183. Morgan AA, Rubenstein E. Proline: The Distribution, Frequency, Positioning, and Common Functional Roles of Proline and Polyproline Sequences in the Human Proteome. *PLoS One*. 2013;8(1):53785. doi:10.1371/JOURNAL.PONE.0053785
184. Zhou HX, Pang X. Electrostatic Interactions in Protein Structure, Folding, Binding, and Condensation. *Chem Rev*. 2018;118(4):1691. doi:10.1021/ACS.CHEMREV.7B00305
185. King MR, Ruff KM, Lin AZ, et al. Macromolecular condensation organizes nucleolar sub-phases to set up a pH gradient. *Cell*. March 2024.

doi:10.1016/J.CELL.2024.02.029

186. McKay MJ, Afrose F, Koeppe RE, Greathouse D V. Helix formation and stability in membranes. *Biochim Biophys Acta - Biomembr*. 2018;1860(10):2108-2117. doi:10.1016/J.BBAMEM.2018.02.010
187. Phelan ML, Rambaldi I, Featherstone MS. Cooperative interactions between HOX and PBX proteins mediated by a conserved peptide motif. *Mol Cell Biol*. 1995;15(8):3989-3997. doi:10.1128/MCB.15.8.3989
188. Alerasool N, Leng H, Lin ZY, Gingras AC, Taipale M. Identification and functional characterization of transcriptional activators in human cells. *Mol Cell*. 2022;82(3):677-695.e7. doi:10.1016/J.MOLCEL.2021.12.008
189. Lyons H, Veettil RT, Pradhan P, et al. Functional partitioning of transcriptional regulators by patterned charge blocks. *Cell*. 2023;186(2):327-345.e28. doi:10.1016/J.CELL.2022.12.013
190. Haile S, Lal A, Myung JK, Sadar MD. FUS/TLS Is a Co-Activator of Androgen Receptor in Prostate Cancer Cells. *PLoS One*. 2011;6(9):e24197. doi:10.1371/JOURNAL.PONE.0024197
191. Infield DT, Rasouli A, Galles GD, Chipot C, Tajkhorshid E, Ahern CA. Cation-π interactions and their functional roles in membrane proteins. *J Mol Biol*. 2021;433(17):167035. doi:10.1016/J.JMB.2021.167035
192. Talukdar PD, Chatterji U. Transcriptional co-activators: emerging roles in signaling pathways and potential therapeutic targets for diseases. *Signal Transduct Target Ther 2023 81*. 2023;8(1):1-41. doi:10.1038/s41392-023-01651-w
193. Tauber D, Tauber G, Parker R. Mechanisms and Regulation of RNA Condensation in RNP Granule Formation. *Trends Biochem Sci*. 2020;45(9):764. doi:10.1016/J.TIBS.2020.05.002
194. Li L, McGinnis JP, Si K. Translational control by prion-like proteins. *Trends Cell Biol*. 2018;28(6):494. doi:10.1016/J.TCB.2018.02.002
195. Lagier-Tourenne C, Polymenidou M, Cleveland DW. TDP-43 and FUS/TLS: emerging roles in RNA processing and neurodegeneration. *Hum Mol Genet*. 2010;19(R1):R46-R64. doi:10.1093/HMG/DDQ137
196. Mandelkow EM, Mandelkow E. Tau in Alzheimer's disease. *Trends Cell Biol*. 1998;8(11):425-427. doi:10.1016/S0962-8924(98)01368-3
197. Stefanis L. α-Synuclein in Parkinson's Disease. *Cold Spring Harb Perspect Med*. 2012;2(2). doi:10.1101/CSHPERSPECT.A009399
198. Holehouse AS, Ginell GM, Griffith D, Böke E. Clustering of Aromatic Residues in Prion-like Domains Can Tune the Formation, State, and Organization of Biomolecular Condensates. *Biochemistry*. 2021;60(47):3566-3581.

doi:10.1021/ACS.BIOCHEM.1C00465/SUPPL_FILE/BI1C00465_SI_001.PDF
199. Chelli R, Gervasio FL, Procacci P, Schettino V. Stacking and T-shape competition in aromatic-aromatic amino acid interactions. *J Am Chem Soc*. 2002;124(21):6133-6143. doi:10.1021/JA0121639/ASSET/IMAGES/LARGE/JA0121639F00009.JPEG
200. Piskacek M, Havelka M, Rezacova M, Knight A. The 9aaTAD Transactivation Domains: From Gal4 to p53. *PLoS One*. 2016;11(9). doi:10.1371/JOURNAL.PONE.0162842
201. Slupsky CM, Sykes DB, Gay GL, Sykes BD. The HoxB1 hexapeptide is a prefolded domain: implications for the Pbx1/Hox interaction. *Protein Sci*. 2001;10(6):1244-1253. doi:10.1110/PS.50901
202. Huang W, Yang S, Shao J, Li YP. Signaling and transcriptional regulation in osteoblast commitment and differentiation. *Front Biosci*. 2007;12(8):3068-3092. doi:10.2741/2296
203. Farley EK, Olson KM, Zhang W, Brandt AJ, Rokhsar DS, Levine MS. Suboptimization of developmental enhancers. *Science*. 2015;350(6258):325-328. doi:10.1126/SCIENCE.AAC6948
204. Lim F, Solvason JJ, Ryan GE, et al. Affinity-optimizing enhancer variants disrupt development. *Nature*. 2024;626(7997):151-159. doi:10.1038/S41586-023-06922-8
205. Zaretsky JZ, Wreschner DH. Protein Multifunctionality: Principles and Mechanisms. *Transl Oncogenomics*. 2008;3(3):99. doi:10.4137/tog.s657
206. Shoval O, Sheftel H, Shinar G, et al. Evolutionary trade-offs, pareto optimality, and the geometry of phenotype space. *Science (80-)*. 2012;336(6085):1157-1160. doi:10.1126/SCIENCE.1217405/SUPPL_FILE/SHOVAL.SM.V2.PDF
207. Yaghmai R, Cutting GR. Optimized regulation of gene expression using artificial transcription factors. *Mol Ther*. 2002;5(6):685-694. doi:10.1006/MTHE.2002.0610
208. Li H, Chen G. In Vivo Reprogramming for CNS Repair: Regenerating Neurons from Endogenous Glial Cells. *Neuron*. 2016;91(4):728. doi:10.1016/J.NEURON.2016.08.004
209. Bocchi R, Masserdotti G, Götz M. Direct neuronal reprogramming: Fast forward from new concepts toward therapeutic approaches. *Neuron*. 2022;110(3):366-393. doi:10.1016/J.NEURON.2021.11.023