Integrated microfluidic tools for the characterization of protein/ DNA interactions *in vitro* and *in vivo*

Thèse N° 9233

Présentée le 25 janvier 2019

à la Faculté des sciences de la vie
Unité du Prof. Deplancke
Programme doctoral en biotechnologie et génie biologique

pour l'obtention du grade de Docteur ès Sciences

par

RICCARDO DAINESE

Acceptée sur proposition du jury

Prof. D. Trono, président du jury
Prof. B. Deplancke, directeur de thèse
Prof. S. Quake, rapporteur
Prof. R. Schlapbach, rapporteur
Prof. B. Fierz, rapporteur

2019

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

"Jeff, one day you'll understand that it's harder to be kind than clever."

Lawrence Preston Gise

"It’s over 9000!"

Vegeta

Acknowledgements

First of all, I would like to thank Bart for his mentorship, (lots of) patience and for believing in me.

I would then like to thank the committee members of my thesis, Prof. Ralph Schlapbach, Prof. Stephen Quake, Prof. Beat Fierz and Prof. Didier Trono, for reading and evaluating my work.

I would like to thank Alina, for her initial guidance and for laying the groundwork for my research.
I thank Marjan, Johannes and Wanze, for being nerds and for showing me how fun it can be to do technology development in a team.
I thank Julie, for her patience and for being an impressive help for everybody in the lab.
I thank Roel and Michael, for teaching me to like beer and for organizing most of the fun stuff that happened outside of work.
I thank Gerard, Daniel, Antonio and Vincent for their help on the FloChIP project.
I thank the rest of the lab members for just being fun people to work with.

I cannot thank my family enough.
I thank my dad for all the time spent together, all the important life lessons and for sharing with me the passion for the best sport in the world.
I thank my mom for being a loving person and giving me and my sister everything we could possibly need or ask for (food-wise and beyond).
I thank my sister for being the funniest “stronza” in the world. It was great to grow up together.
I thank my grandparents, there is nothing comparable to the love they are capable of.

I thank my friends from Santhià - Massimo, Francesco and Simone - for the best childhood and teenage years anybody can ever dream of. Brothers.

I thank my friends from Vercelli, Marco, Timmy, Deiv, Lapo, Ste, Frenk, Lore, Peila, Dindi, Caro, Roby, Iaia and Connie for being the most interesting, funny and stimulating group of people I ever met. Polli.

Finally, there is another person I cannot thank enough, Vera.
I thank Vera for changing my life, for making me a better version of myself and, simply, for making me happy.
(By the way, I thank Alessia as well for having played cupid!)

Abstract

The specific interaction between DNA and proteins constitutes one of the crucial elements in the regulation of gene expression. This thesis focuses on the development and optimization of two microfluidic-based technologies called SMiLE-seq and FloChIP that enable the sensitive and high-throughput analysis of two important aspects of protein/DNA interactions: respectively, 1) the transcription factor DNA-binding specificities and 2) the genome-wide distribution of chromatin-associated proteins.

In Chapter 1, we describe the core motivation, operating principles, optimization and results obtained with SMiLE-seq. As opposed to existing solutions, SMiLE-seq offers the possibility to screen a large library of randomized DNA for sequence-specific protein ligands in a miniaturized context. SMiLE-seq integrates in a single microfluidic chip the advantages of both HT-SELEX – i.e. large DNA library screens coupled to high-throughput sequencing – and MITOMI – i.e. microfluidic trapping of DNA/protein complexes. We first demonstrate as proof-of-principle that SMiLE-seq successfully recapitulates with robustness and reproducibility the binding specificities of known factors belonging to different species and TF families. Moreover, we show that SMiLE-seq reflects the energy binding landscapes of TFs with high accuracy. Subsequently, we target the numerous although largely uncharacterized family of TFs called KRAB-ZFPs. We initially set out to redesign the microfluidic and protocol architecture in order to attain maximal throughput and sensitivity. Next, we add the ability of probing the sensitivity of these factors to methylation by synthesizing randomized methylated DNA libraries. Finally, we proceed to test 101 KRAB-ZFPs with both methylated and non-methylated DNA. We obtained high confidence motifs for 43 factors, of which 22 we not sensitive to methylation, 10 yielded motifs only with methylated DNA and 10 only with non-methylated DNA. By integrating our SMiLE-seq data with published ChIP-exo data and *in silico* predictions, we develop a framework for systematically identifying which zinc fingers directly contribute to DNA binding for a given KRAB-ZFPs.

In Chapter 2, we describe the development and results obtained with FloChIP, a microfluidic implementation of the widespread ChIP-seq and sequential ChIP-seq protocols. FloChIP encompasses three main technological advances: 1) the multi-layered serface chemistry that allows any off-the-shelf antibody to be immobilized on-chip and and 2) a micropillar architecture that provides high surface-to-volume ratio and efficient removal of non-specific DNA. 3) The direct on-chip tagmentation of immunoprecipitated DNA. We validate our approach by first performing H3k27ac ChIP-seq on different number of sample inputs, i.e. from 500 to 10^6 cells. We show that although FloChIP provides good results with as low as 500 cells, the best trade-off between low-input and data quality is reached at 100'000 cells. Therefore, we proceed to prove the flexibility of our approach by performing ChIP-seq on different histone marks – namely H3k27ac, H3k27me3, H3k4me3, H3k4me1 and H3k9me3 – and comparing them to existing ENCODE data. Both region-specific profiles and enrichment scores show high similarity with publicly available data. Subsequently, we set out to demonstrate the feasibility of on-chip sequential-ChIP-seq by performing consecutive H3k4me3 and H3k27me3 IP on chromatin derived from mouse embryonic stem cells. Our data faithfully recapitulates previously studies and show enrichment of bivalent 4me3/27me3 in promoters associated to embryonic development. Finally, we illustrate the high-throughput feature of our technology by performing MEF2A-ChIP-seq on chromatin derived from 32 different lymphoblastoid cell lines.

Keywords: transcription factor, DNA binding specificity, ChIP-seq, sequential ChIP-seq, microfluidics.

Sommario

L'interazione tra DNA e proteine costituisce uno degli elementi principali nella regolazione dell'espressione genica. Questa tesi si concentra sullo sviluppo e ottimizzatione di due tecnologie a base di microfluidi, chiamate SMiLE-seq e FloChIP, che permettono in maniera efficiente e high-throughput di analizzare due aspetti importanti delle interazioni DNA/proteina: rispettivamente, 1) la specificità dei fattori di trascrizione e 2) la distribuzione genome-wide di protein associate alla cromatina.

Nel Capitolo 1 descriviamo la motivazione, i principi operativi, l'ottimizzazione e risultati ottenuti con SMiLE-seq. Contrariamente a soluzioni pre-esistenti, SMiLE-seq offre la possibilità di fare uno screening su una vasta libraria di DNA randomizzato in un contesto miniaturizzato. SMiLE-seq presenta in un solo chip i vantaggi di HT-SELEX - i.e. DNA screening su large scale in combinazione con sequenziamento high-throughout - e MITOMI - i.e. la cattura di complessi proteina/DNA a livello microfluidico. Inizialmente dimostriamo che SMiLE-seq riproduce fedelmente la specificità di legame di fattori conosciuti e appartenenti a diverse speci e famiglie di TFs. Inoltre, SMiLE-seq riflette l'energia di legame con grande accuratezza. Successivamente, prendiamo in considerazione la numerosa seppur poco studiata famiglia di TFs conosciuta come KRAB-ZFPs. Inizialmente, ci dedichiamo a riprogettare il sistema a microfluidic e il protocollo sperimentale al fine di ottenere massimo throughput e sensibilità. In secondo luogo, aggiungiamo la capacità di sondare la sensibilità di questi fattori alla metilazione tramite la sintesi di librerie di DNA metilate e randomizzate. Infine, procediamo a testare 101 KRAB-ZFPs sia con DNA metilato sia non-metilato. Otteniamo cosí binding motifs per 43 fattori: di questi 22 non presentano sensibilità a metilazione, 10 presentano motifs solo con DNA metilato e 11 solo con DNA non-metilato. Tramite l'integrazioni dei dati ottenuti con SMiLE-seq con dati ChIP-exo pubblici e simulazioni *in silico*, sviluppiamo un framework per identificare in maniere sistematica quali sono gli zinc fingers che contribuiscono direttamente al legame di DNA.

Nel Capitolo 2, descriviamo lo sviluppo e i risulati ottenuti con FloChIP, un'implementazione microfluidica dei diffusi protocolli di ChIP-seq e sequential ChIP-seq. FloChIP include tre

principali progressi tecnologici: 1) una surface chemistry multi-strato che permette a qualsisi anticorpo disponibile commercialmente di essere immobilizzato on-chip, 2) una architettura a microcolonne che offre un alto rapporto area/volume e una efficiente rimozione di DNA non specifico e 3) la tagmentazione on-chip del DNA immunoprecipitato. Validiamo il nostro approccio tramite H3k27ac ChIP-seq su un diverso numero di input cellular, da 500 a 10^6 cellule. Dimostriamo così che sebbene con FloChIP si ottengano buoni risulati con solo 500 cellule, il migliore trade-off tra low-input e qualità dei dati si ottiene con un input di 100'000 cellule. Quindi, procediamo ad illustrare la flessibilità della nostra tecnica eseguendo ChIP-seq su diversi tipi di histone marks - nel dettaglio H3k27ac, H3k27Sme3, H3k4me3, H3k4me1 and H3k9me3 - e paragonando i dati cosi ottenuti ai dati ENCODE. Sia i profili di specifiche regioni genomiche che le enrichment scores esibiscono grande similarità con i dati pubblici. Successivamente, ci proponiamo di dimostrare sequential-ChIP-seq on-chip eseguendo H3k4me3 e H3k27me3 IP in maniere consecutiva su cromatina ottenuta da cellule embrionali di topo. I nostri data riproducono fedelmente studi precedenti e presentano enrichment di 4me3/27me3 in promotori associati allo sviluppo embrionale. In fine, mostriamo le capacità high-throughput della nostra tecnologia eseguendo MEF2A-ChIP-seq su cromatina ottenuta da 32 linee cellulari di linfoblastoidi differenti.

Parole chiave: fattori di trascrizione, specificità di legame DNA, ChIP-seq, sequential ChIP-seq, microfluidica.

Table of Contents

List of figures

List of Supplementary figures

List of abbreviations

BSA – bovine serum albumin

bvScore – bivalency score

BYDV – wheat-germ-compatible barley yellow dwarf virus

ChIP – chromatin immunoprecipitation

ChIP-seq – chromatin immunoprecipitation followed by next generation sequencing

Ct – cycle threshold

DNA – deoxyribonucleic acid

ES cells – embryonic stem cells

FRiP – fraction of reads into peaks

GFP – green fluorescent protein

HM – histone marks

HPC – high-CpG

HT-SELEX – high-throughput systematic evolution of ligands by exponential enrichment

IP – Immunoprecipitation

KRAB-ZFPs – Krüppel-associated box domain zinc finger proteins

LCLs – lymphoblastoid cell lines

mESCs – mouse embryonic stem cells

MITOMI – Mechanically Induced Trapping of Molecular Interactions

NGS – next generation sequencing

PBM – protein binding microarrays

PBS – phosphate buffer saline

PCR – polymerase chain reaction

PFA - paraformaldehyde

RNA – ribonucleic acid

SMiLE-seq–selective microfluidics-based ligand enrichment followed by sequencing

TE – translation enhancers

TFs – transcription factors

TSS – transcription start sites

ZFPs –zinc-finger proteins

Chapter 1: Introduction

1.1 The relevance of Protein/DNA interactions

First introduced by Francis Crick (Crick, 1970), the central dogma of molecular biology states that "**DNA is transcribed into RNA and RNA is translated into protein**" (Fig. 1.1). Despite its simplicity, the central dogma of molecular biology provides a simple yet powerful framework to understand how information is encoded in biological systems:

- It introduces the three fundamental classes of biopolymers, i.e. DNA, RNA and proteins.
- It highlights the direction for the biological flow of information, i.e. DNA → RNA → Protein.
- It provides the terms used for describing the conversion of one species into the next, i.e. transcription (DNA → RNA) and translation (RNA → Protein).

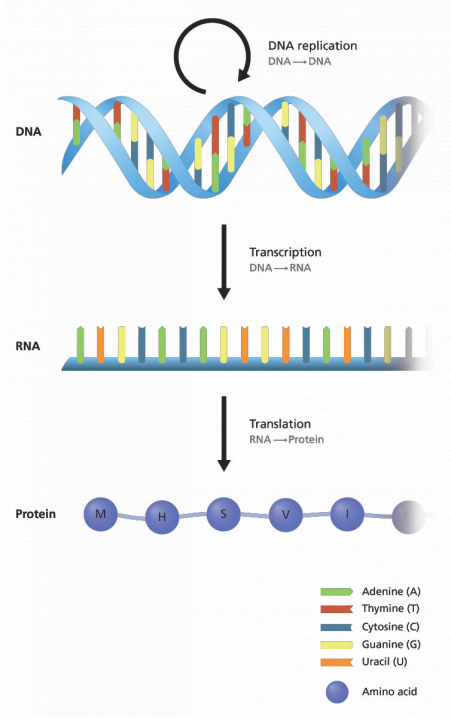


Figure 1.1: The flow of information between DNA, RNA and protein. Image credit: Genome Research Limited

Although true and experimentally validated, the central dogma somehow fails to deliver the sense of enormous complexity and diversity that biological systems have evolved throughout the years. In humans, for instance, a single zygote grows into ~37 trillion cells (Bianconi et al., 2013), all different (to some extent) from each other but all containing the exact same DNA makeup, i.e. the same **genome**. Therefore, it is clear that the genome not only encodes information on what proteins to express but also information on when and where to express those proteins. Surprisingly, only 1.5% of the human genome codes for proteins while more than 80% has been assigned a regulatory function (ENCODE Project Consortium, 2012). Thus, the majority of the genome is dedicated to integrating external information in order to efficiently regulate the expression of its smaller protein-coding portion. This external information, in turn, is relayed to the genome by some of the same protein species that are encoded in it by means of physical **protein/DNA interactions** (Fig. 1.2). Protein/DNA interactions constitute a major component of a much vaster collection of molecular interactions - the **gene regulatory network** - which collectively govern and orchestrate the expression of proteins in cells.

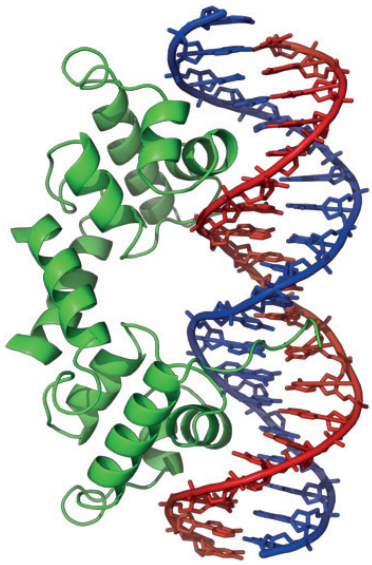


Figure 1.2: Molecular reconstruction of the lambda repressor helix-turn-helix transcription factor bound to its DNA target. Image credit: Wikipedia

From this observations, a more complex picture is drawn than the one outlined initially by the central dogma. In this picture, the link between proteins and DNA is not unidirectional but bidirectional in the sense that proteins themselves interact with DNA to regulate its transcription. Moreover, the human genome is made up of 3 billion "letters" (or base pairs) and there are between 1'200 and 1'300 estimated sequence-specific DNA-binding proteins, also called **transcription factors** (Vaquerizas et al., 2009). Given the absence of an accurate prediction model for protein/DNA interaction, the only way to study the complexity of this enormous network is to probe its interactions experimentally.

1.2 Techniques to study of Protein/DNA interactions

Historically, protein/DNA interactions have been experimentally probed mainly in two contexts: *in vitro* and *in vivo*.

1.2.1 *In vitro* techniques

In vitro techniques aim to probe the interaction between protein and DNA in isolation from their natural context, in order to exclude possible confounding factors due to unexpected interactions. The main goal of in vitro methods is to derive the sequence specificity and affinity of transcription factors (TFs).
Several *in vitro* solutions have been proposed over the years; of these, the most successful examples are: protein binding microarrays (PBM) (Berger et al., 2006) and high-throughput systematic evolution of ligands by exponential enrichment (HT-SELEX) (Jolma et al., 2010).

In **PBM** (Fig. 1.3), a double-stranded DNA library - either genomic (Mukherjee et al., 2004) or synthetic (Berger et al., 2006) - is immobilized on a microarray substrate and exposed to a TF of interest. After binding occurs, the unbound protein is washed away while the specifically bound TF remains attached to its corresponding DNA strand in a specific position of the microarray. A fluorescent antibody targeting the TF or its tag is later used to determine the position and hence, the sequence of the TF-bound DNA. PBMs have been extensively proven their efficiency and robustness in determining the DNA binding motifs of several mammalian TFs (Berger et al., 2008). The major drawback of PBMs lies the limited DNA sequence combination which they can test, which is ultimately constrained by the size of the microarray. As a matter of fact, the current upper limit is the space of all possible 12-mers of DNA.

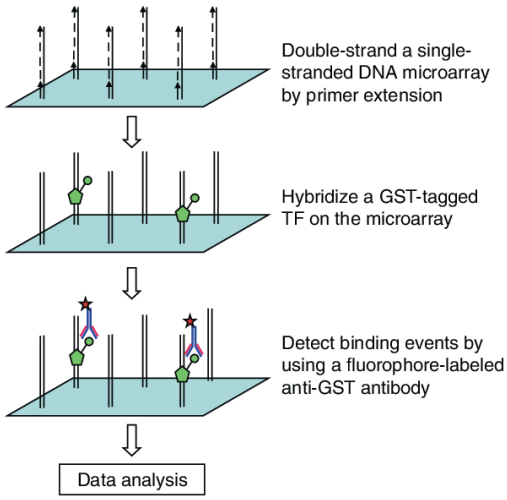


Figure 1.3: Workflow of protein binding microarrays. Copyright 2011 Springer.

On the other hand, in **HT-SELEX** (Fig. 1.4), it is the protein of interest which is immobilized inside the well of a 96-well plate (Jolma et al., 2010). As opposed to PBMs, HT-SELEX provides a much larger coverage of the DNA combinatorial space (x10^6 higher). The main differentiator for HT-SELEX is the fact that the DNA sequences are not immobilized on a solid support and can instead be freely floating in solution. This allows DNA randomers to be much longer as compared to PBMs, which translates into a much larger DNA search space. After exposing the TF with the random DNA library, the specifically bound DNA is collected, amplified and re-fed once more to the TF in order to achieve exponential enrichment of specific DNA binders. After usually three or four rounds of selection, the collected DNA is sequence by next generation sequencing (NGS) and the binding motifs are computationally derived. Another distinctive factor of HT-SELEX is the ability to multiplex different TFs in the same experiment. This is achieved thanks to molecular barcodes included in the synthetic DNA libraries. Given its advantages, HT-SELEX has become the most prolific *in vitro* technique for the derivation of TF binding specificities (Jolma et al., 2013), providing insights in the DNA-dependent formation of heterodimers (Jolma et al., 2015) as well as the impact of methylation on DNA binding (Yin et al., 2017).

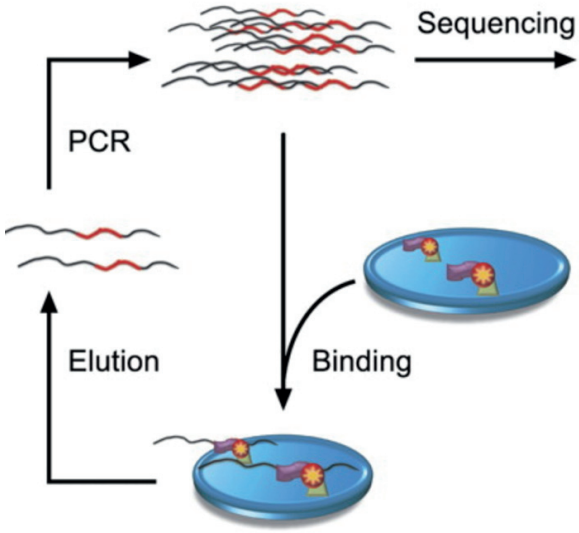


Figure 1.4: Overview of HT-SELEX selection process. Copyright 2010, Cold Spring Harbor Lab Press.

1.2.2 *In vivo* techniques

In vivo techniques probe the interaction between protein and DNA in their natural context, which comes with advantages and disadvantages. The main advantage is that not only can *in vivo* methods derive the binding preferences of the proteins under study, but they can also provide snapshots of their genome-wide distribution at a given time. The main disadvantage is the fact that, in their natural context, the DNA binding properties of a protein can be confounded by a myriad of other possible interactions.
Currently, the most widely adopted technique for in vivo protein/DNA interaction studies is chromatin immunoprecipitation followed by next generation sequencing (ChIP-seq).

ChIP-seq (Fig. 1.5) consists of several steps. Initially, cells are treated with a crosslinking agent, thereby effectively "freezing" all protein/DNA interactions. Subsequently, the genomic material is fragmented either by sonication or enzymatic digestion until all fragments are at a size compatible with subsequent DNA sequencing (i.e. around 150-200bp). Next, the fragmented chromatin is mixed with solutions containing antibody-bound microbeads: the antibody binds tightly to its target protein, whereas the microbeads are used in the separation procedure. Once antibody/protein/DNA complexes are formed, the microbeads are extracted from the solution

(usually by means of a dedicated magnet) washed and the DNA is finally eluted by high temperature and high salt buffers. Subsequently, the collected DNA is sequenced by NGS and the corresponding data is analyzed to study the genome-wide DNA-binding properties of the protein. Introduced in 2007 (Johnson et al. 2007), ChIP-seq has become an instrumental tool in identifying the functional elements of the human and other genomes (Landt et al., 2012). However, it remains a manually intensive, slow (>3 days) and low-throughput method and only a small number of protein targets can be probed at a time.

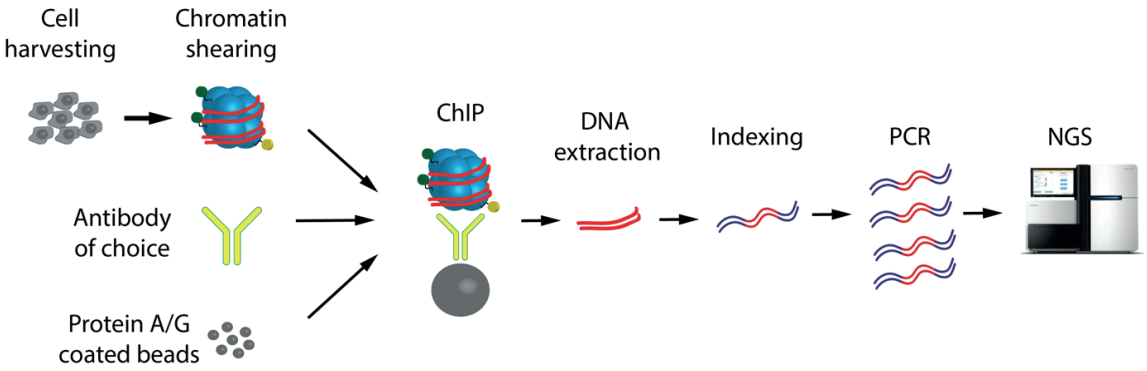


Figure 1.5: Overview of the steps and reagents required for ChIP-seq.

1.3 The potential for microfluidics applied to biology

All the techniques presented above are macroscale implementations, designed by humans for human's execution. However, the fundamental processes of biology happen at a very different scale, the microscale. Moreover, as also mentioned above, the complexity of the protein/DNA interaction network is such that the shear number of interactions to probe requires fast, automated and high-throughput methods. These important missing aspects, miniaturization, automation and throughput, are the key advantages offered by microfluidics.

1.3.1 What are microfluidic systems?

Microfluidics is both a science and a technology. The science addresses the study of fluids dynamics through micro-channels and micro-structures. The technology addresses with the fabrication of miniaturized devices inside of which fluids can be constrained. In general, microfluidics deals with very minute amount of fluids, from microliters to femtoliters. At these scales, fluids behave in different way and a plethora of useful properties originate because of this.

1.3.2 The advantages of microfluidics and their application to protein/DNA interactions.

Given their microscopic nature, microfluidic devices offer a series of advantages over traditional macroscale techniques. First of all, the very small size translates into reducing costs of reagents. Furthermore, the small size enables the development of highly integrated and sensitive tools capable of isolating and analyzing small quantities of precious samples. Examples of areas of molecular biology in which microfluidic chips have found successful include:

- single cell handling and analysis (Macosko et al., 2015, Fig. 1.6a)
- microfluidic flow cytometers and cell sorters (Nitta et al., 2018, Fig. 1.6b)
- single cell culturing chambers (Lecault et al., 2011, Fig. 1.6c)

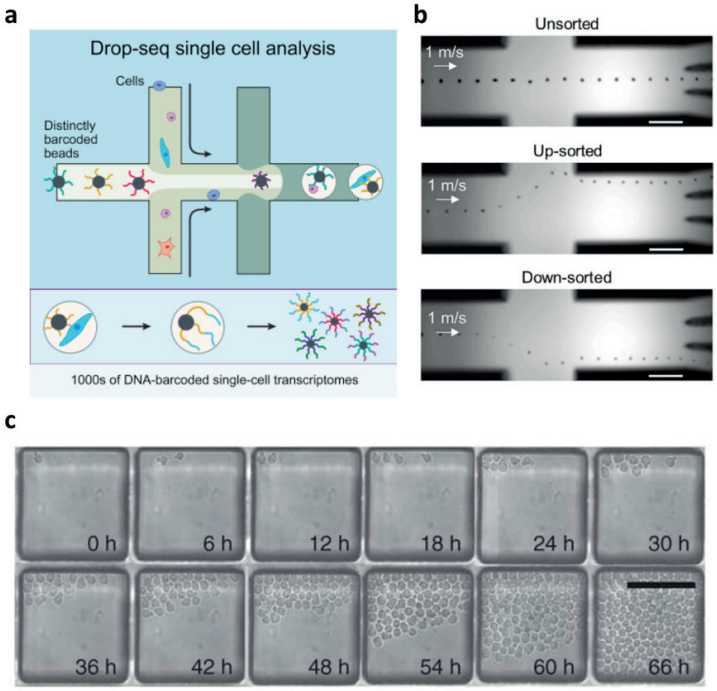


Figure 1.6: Examples of microfluidic applications to biology. a) Single cell handling and encapsulation. Copyright 2015, Cell Press. b) Single cell sorting. Copyright 2018, Cell Press. c) Single cell culturing. Copyright 2011, Nature Publishing group.

Nevertheless, the most relevant devices for the work in this thesis are the ones applied to the study of protein/DNA interactions. As per standard molecular biology techniques, also these devices can be subdivided in two categories, *in vitro* and *in vivo*, for which the two most notable examples are MITOMI (Maerkl and Quake, 2007) and MOWChIP-seq (Cao et al. 2015), respectively.

MITOMI (Fig. 1.7) stands for Mechanically Induced Trapping of Molecular Interactions. It is an approach that exploits high-throughput microfluidics in order to achieve sensitive measurement of the biophysical affinity between TFs and DNA. So far, MITOMI was used to study the binding energy profiles of a numerous TFs from a diverse set of organisms such as human, yeast, and E. coli both in monomeric and heterodimeric form (Maerkl and Quake, 2007; Isakova et al., 2016).

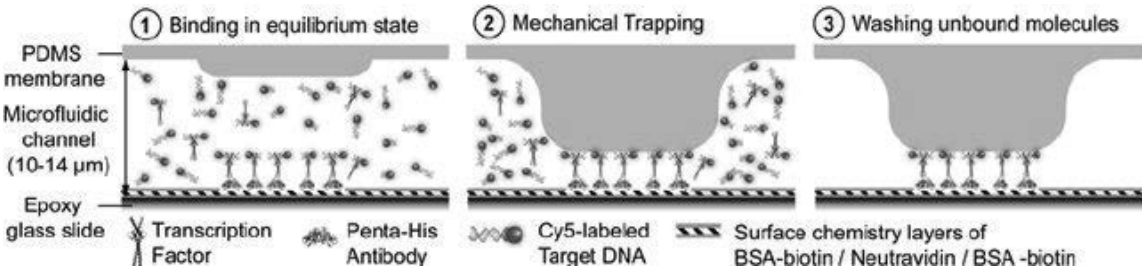


Figure 1.7: Schematic workflow of the MITOMI principle. Copyright 2012 Springer.

The key advantage of MITOMI is the speed at which the assay is conducted, thereby allowing to obtain a snapshot of the equilibrium state of the TF/DNA interaction and keeping the complexes intact. Other desirable advantages are the ability to purify proteins directly on-chip, low reagent consumption and small device footprint. However, one key disadvantage of this method is the requirement for know *a priori* the DNA-binding specificity of TF under study. This precludes the possibility of performing MITOMI on large number of TFs for which binding motifs have not been derived yet. In the previous section, it was mentioned that the efficient derivation of the DNA-binding specificity of TFs was a successful application of HT-SELEX. It is therefore intriguing to wonder if it would be possible borrow lessons learned from HT-SELEX and MITOMI and to devise a microfluidic solution that can be used to study the DNA-binding specificity of TFs in a sensitive and miniaturized manner. The design, implementation and optimization of such a device alongside with the results obtained with it are the main subject of Chapter 2.

MOWChIP-seq (Fig. 1.8) stands for microfluidic oscillatory washing–based ChIP-seq. It is an approach in which immunoprecipitation is carried out by a closely packed column of magnetic beads coated with an antibody of choice and localized within the walls of a microfluidic chamber. The chromatin sample is introduced into the microfluidic chip and flown through the small space in between the microbeads. Washing to remove unbound chromatin is then carried out in an oscillatory manner and with the help of a magnet in order to keep the microbeads on chip.

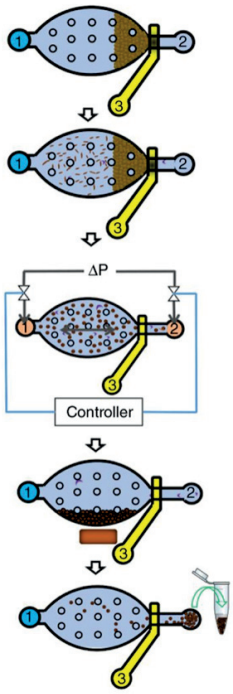


Figure 1.8: Schematic sequence of steps for MOWChIP-seq.

The main advantages of MOWCHIP-seq are the short experimental time-frame (45 minutes versus the 4 hours for manual ChIP-seq) and its sensitivity (chromatin landscapes can be obtained for as low as 100 cells versus the minimum 1 million cells of manual ChIP-seq). The main drawback of MOWCHIP-seq is that it can address only one protein species at a time, hence it suffers from being as low-throughput as manual ChIP-seq. In Chapter 3, I describe a novel microfluidic solution that not only is sensitive and fast, but it also performs in a high-throughput, parallel and automated way.

1.4 KRAB zinc finger proteins

One of the main focuses of this thesis is to apply microfluidics to the study of DNA-binding specificities of a large transcription factor family called the Krüppel-associated box domain zinc finger proteins (KRAB-ZFPs). In humans there are some 350 KRAB-ZFPs (Fig. 1.9) and their function has been mainly associated with repressing the activity of transposable retroelements during embryonic development (Ecco et al., 2017, Imbeault et al., 2017).

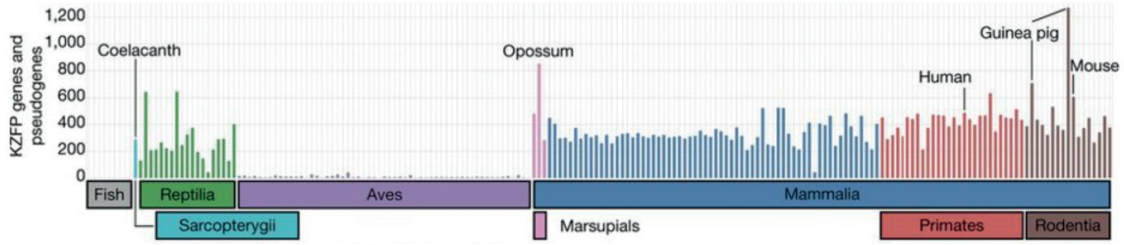


Figure 1.9: KRAB-ZFP count in different species genomes.
Copyright 2017 Nature Publishing Group.

KRAB-ZFPs are characterized by two domains with distinct function: an N-terminal KRAB domain, which mediates the factor repressive activity, and a C-terminal array of C2H2 zinc fingers that confer their DNA binding ability (Fig. 1.10).

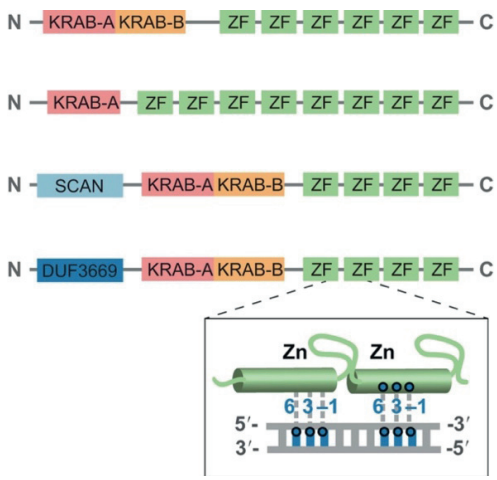


Figure 1.10: KRAB-ZFP structure and domains.
Copyright 2017 The Company of Biologists.

Human KRAB-ZFPs have between 2 and 40 fingers per factor and, even if the DNA-binding properties of individual zinc fingers is fairly well understood (Najafabadi et al., 2015, Persikov et al., 2014), the *a priori* prediction of the DNA specificity of entire zinc finger arrays remains a challenge. C2H2 zinc fingers are normally arranged in two β-barrels and one α-helix conformation coordinated by a zinc ion (Fig. 1.11)

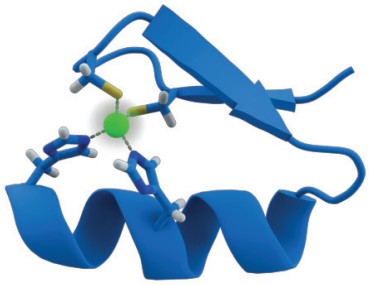


Figure 1.11: C2H2 zinc fingers structure.

The main determinant of the of the DNA-binding specificity of each zinc finger are the side chains of the amino-acids located at positions -1,2,3 and 6 with respect to the beginning of the α-helix motif. Conventionally, it has been thought that each zinc finger follows a one-finger-three-base rule, where each finger recognizes a specific set of three DNA base pairs. However, recent structural information (Patel et al., 2018) and comparison between predicted and ChIP-seq-based DNA binding motifs (Chapter 2) reveals widespread deviation of zinc finger arrays from this rule. In particular, Patel and colleagues demonstrate by structural analysis that the 11 zinc fingers of the mouse Zfp568 interact with DNA by contacting 2, 3 or 4 bases per finger. Moreover, they also demonstrate that not all zinc fingers of the array engage in DNA binding, which is consistent with other observations suggesting that zinc fingers not directly involved in DNA binding could be implicated in other types of interactions, e.g. with RNA or proteins (Ecco et al., 2017).

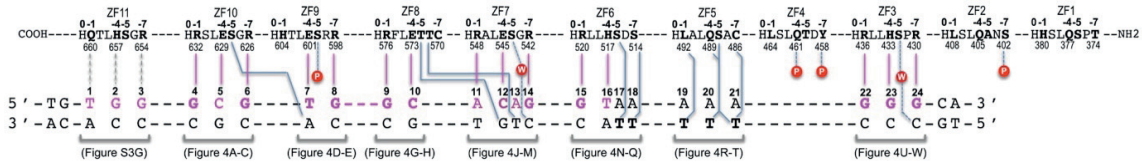


Figure 1.12: DNA interactions schematic of the zinc finger array of the mouse Zfp568. Copyright 2018 Cell Press.

To further complicate the picture, it has been reported that certain KRAB-ZFPs such as ZFP57 exclusively bind methylated DNA (Quenneville et al., 2011). Taken together, these observations seem to suggest that the DNA binding specificity of KRAB-ZFPs is influenced by several factors beyond the mere DNA sequence and individual zinc finger specificity. These factors may include DNA methylation, 3D-arrangements of the zinc finger array and protein-protein or RNA-protein interactions (Ecco et al., 2017). In order to begin to disentangle the complexity of the DNA-binding mechanisms KRAB-ZFPs, we apply our in-house microfluidic solutions to understand which factors are capable of *in vitro* DNA binding, i.e. in isolation from co-factors, and which of these factors are sensitive to the methylation state of their DNA substrate.

Chapter 2. SMiLE-seq

Smile-seq and Smile-seq v2.0: the next generation of tools to study the binding specificities of transcription factors

Alina Isakova[1], Riccardo Dainese[1], Evgeniia Pankevich[1], Bart Deplancke[1]

Institute of Bioengineering, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

2.1 Introduction

The DNA-binding specificities of transcription factors are an essential aspect in the regulatory dynamics of gene networks. Collective efforts aimed at extensively cataloguing these specifies have resulted in several online databases, e.g. TRANSFAC (Matys et al., 2006), HOCOMOCO (Kulakovskiy et al., 2016), or UniPROBE (Mathelier et al., 2014), which collectively account for 601 human transcription factor binding motifs. Nevertheless, as also mentioned in the introduction, it has been estimated that mammals express between 1'300 and 2'000 TFs (Vaquerizas et al., 2009), which suggests further experimental efforts have to be conducted in order to complete the specificity catalog. Of this missing data, a large portion belongs to transcription factors characterized by C2H2 zinc-finger-mediated DNA binding (Deplancke et al., 2016). There is a number of speculated reasons as to why these TFs have so far resisted experimental characterization, e.g. incomplete TF expression context, the need for co-factors and possibly the simple fact that existing technologies are not sensitive enough to detect certain types of TF/DNA interactions. Therefore, in order to obtain a deeper understanding of our genome architecture, it is important to devise alternative methods that can independently and quantitatively provide DNA-binding information.

In this chapter, I present two developmental stages of a microfluidic technology, SMiLE-seq (selective microfluidics-based ligand enrichment followed by sequencing), that allows for the sensitive and robust derivation of DNA-binding specificities of human transcription factors. SMiLE-seq combines the advantages of both MITOMI and HT-SELEX in order to screen a large library of random DNA with the added benefits of performing assay in a microfluidic context, i.e. low sample requirement, low reagent consumption and fast reactions. In the first developmental stage, I followed the guidance of a senior PhD and later postdoc, Alina Isakova, who has conceived the original experimental pipeline and obtained several SMiLE-seq results. We show that we successfully benchmark SMiLE-seq on several monomers belonging to distinct structural families and species. Moreover, we show that the distribution of k-mers in SMiLE-seq libraries correlates positively with MITOMI derived binding energy landscapes. This first stage is the subject of a first landmark paper, which encompasses all of Alina's major findings (Isakova et al, 2017). Subsequently, I show that in order to tackle the most numerous and largely unexplored TF family,

the Krüppel-associated box (KRAB)-containing C2H2 zinc-finger proteins (ZFPs), I set out to further develop our technology. As a result, SMiLE-seq v2.0 is a high-throughput, unbiased and highly multiplexable implementation which we use to obtain the largest catalogue of *in vitro*-derived KRAB ZFPs motifs to date and gain insights in their DNA binding properties.

2.2 SMiLE-seq

2.2.1 The SMiLE-seq device and its operation

As mentioned above SMiLE-seq combines features of both MITOMI and HT-SELEX. MITOMI provides a great framework for the expression and immune-based capture of tagged transcription factors. In particular, as opposed to other methods requiring protein expression, MITOMI does not require protein purification prior the actual experiment. The reason for this is that in the majority of cases the required protein is obtained through bacterial- or yeast-based expression, which which require the extra step of protein purification before the following assays. On the other hand, MITOMI adopts a different route and makes use of the so-called "cell-free" *in vitro* expression systems. As the name suggests, these are commercially available cellular lysates that can express the required protein by simple mixing and incubation with the appropriate plasmid. Despite their ease of use and very short turnaround time, these systems are very expensive and, for instance, bacterial-based expression is often a more cost-effective choice for obtaining large amounts of protein. However, being a microfluidic technology, MITOMI does require only minute amounts of protein, which, in turn can be expressed by very small fractions of the *in vitro* expression system, thereby amply justifying and encouraging their use.

Another notable feature of MITOMI is the way transcription factors are immobilized on-chip. Indeed, in MITOMI the clever integration between multilayer microfluidics and surface chemistry provides a way to controllably pull-down the TFs in very confined regions called MITOMI buttons. The MITOMI buttons are circular areas of the microfluidic chip with a diameter of ~250μm in which the ceiling of the microchannel can be made to collapse by applying pressure on channels above (hence called "control channels"), therefore closing like a button and trapping the molecular species beneath it. By selectively and consecutively flowing under the MITOMI button biotinyalted-BSA, neutravidin and a biotinylated-anti-GFP antibody, while closing the MITOMI button at appropriate times (details in the Methods section), it is possible to specifically immobilize the biotinyalted antibody under the area of MITOMI button (Fig. 2.1).

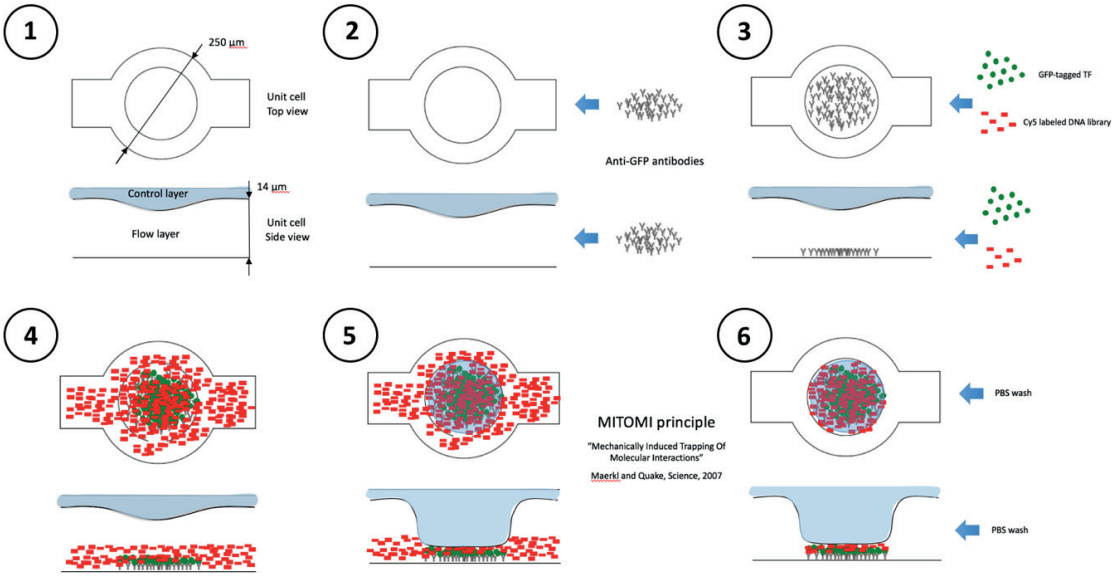


Figure 2.1: Detailed schematic of the MITOMI procedure.

In SMiLE-seq, we adopted these two great features of MITOMI, while at the same time adding the power of randomized DNA and molecular barcodes, in order to achieve very high-throughput in terms of screened DNA sequences. This second set of features was inspired by the seminal work of Jolma et al. and their HT-SELEX platform. In HT-SELEX proteins are obtained through standard bacterial expression, which, as we discussed, is a more laborious approach as compared to *in vitro* expression systems. However, the clever synthesis of their DNA library allowed them to achieve a very robust and multiplexed method, which resulted in great success in characterizing new DNA binding motifs (Jolma et al., 2013). Briefly, the original HT-SELEX library consisted of four main components: 1) Illumina-compatible primer binding sites for amplification, 2) Illumina-compatible primer binding sites for sequencing, 3) a TF-specific barcode in order to make experiments multiplexed and 4) a DNA random region of 14bp. Intuitively, this library construct not only allowed them to sample for each TF a large number of DNA sequences (namely 4^14= 268'435'456), but also to do so in a multiplexed format where each TF is assigned a specific molecular barcode (Fig. 2.2).

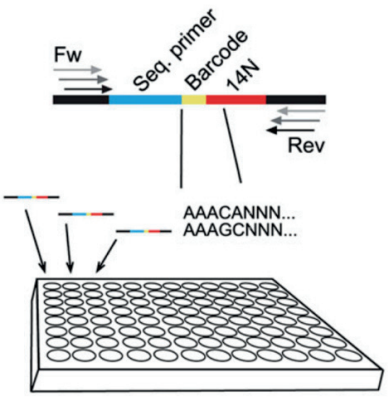


Figure 2.2: HT-SELEX random library composition.

The result of the cross-fertilization between the approaches reported in MITOMI and HT-SELEX is what we call selective microfluidics-based ligand enrichment followed by sequencing, i.e. SMiLE-seq. Our approach consists in a modified MITOMI device (Fig. 2.3) which harbours 8 consecutive MITOMI buttons. As in MITOMI, the inlets are separated by microvalves that control the sequential introduction reagents in order to immobilise antibodies in the center of the MITOMI buttons. The outlets are used both for ridding the device of surplus reagents and for washing away unbound DNA. The microcapillary pumps are introduced at the lower and upper ends of each MITOMI button in order to provide passive sample loading.

The second important aspect of SMiLE-seq is the synthesis of the barcoded and randomised library. As opposed to HT-SELEX, we opted for a larger random region, hence allowing us to sample an even larger sequence space as compared to HT-SELEX. Moreover, we introduced two small TF-specific barcodes at each end of the random region, in order to increase the reliability of post-experiment bioinformatic demultiplexing. The final, 121bp construct, already includes portions of sequencing adapters directly within the random DNA library design.

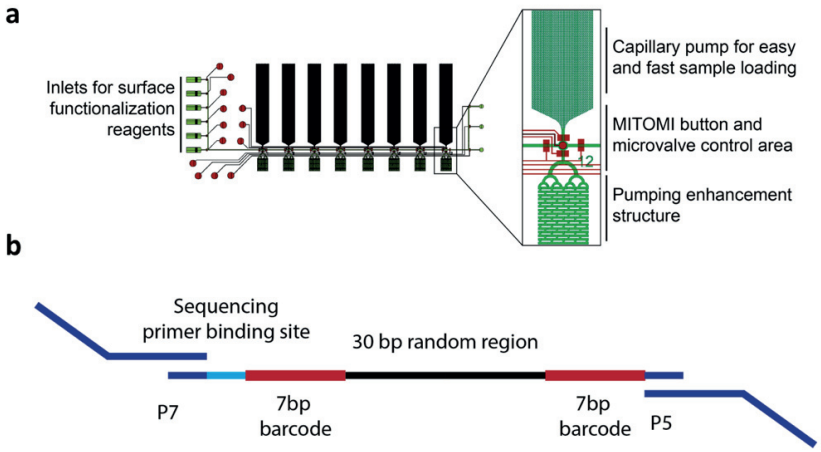


Figure 2.3: The two fundamental components of SMiLE-seq. a) The 8 unit MITOMI-based microfluidic chip. b) The Illumina-compatible random DNA library.

Despite adopting with minor modifications already established solutions, SMiLE-seq required substantial optimization in order to provide efficient enrichment of the specifically TF-bound DNA. Importantly, we reasoned that given the ability to wash the interior of microfluidic channels, it should be possible to simply remove unbound by extensive washing with a pure saline buffer like PBS. This notion would allow us to enrich for the wanted DNA in a very convenient manner while, at the same time, avoiding laborious rounds of exponential enrichment. The resulting key steps of the workflow are depicted in Fig. 2.4:

1) Express transcription factor of interested as a GFP-fusion in an *in vitro* expression mix.
2) Mix expressed TF with barcode DNA library and load onto antibody-functionalized device.
3) Perform MITOMI on TF/DNA complexes.
4) Wash unbound DNA while keeping MITOMI button closed.
5) Open MITOMI button and elute bound DNA.
6) Sequence bound DNA and compute motif logo.

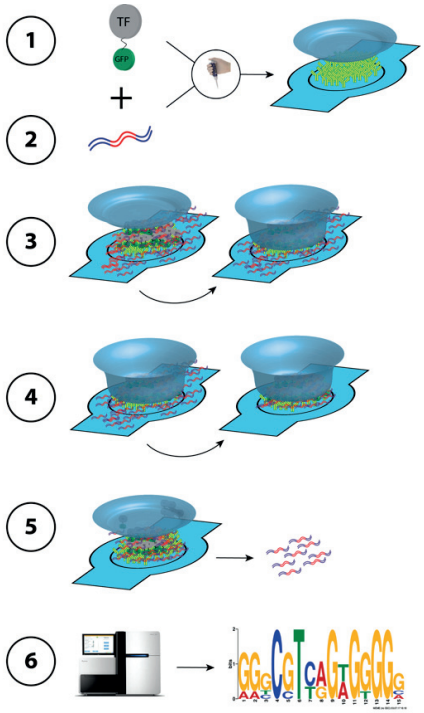


Figure 2.4: The SMiLE-seq workflow.

2.2.2 SMiLE-seq identifies motifs for a wide range of TFs

In order to benchmark our technology, we set out to obtain motifs from a number of previously characterized factors (Fig 2.6). Based on detectable GFP signal, we observed the correct expression of 58 (6 *Drosophila*, 12 mouse and 40 human) out of the randomly selected 60 transcription factors belonging to different families (Isakova et al, 2017). With SMiLE-seq we were able to obtain highly enriched motifs and for all of them. Moreover, transcription factors tested in replicates also showed high correlation (Supp. Fig. 2.1), therefore demonstrating the reproducibility of SMiLE-seq. Seeking an external validation of the retrieved motifs, we compared them to public databases through a TOMTOM-based search (Gupta et al., 2007). The top resulting TOMTOM matches, including motifs derived by independent methods like ChIP-seq, PBMs and SELEX, corresponded to the SMiLE-seq-derived motifs for all tested 58 motifs (Isakova et al, 2017). These results establish the validity of motifs obtained through SMiLE-seq and, in general,

advocate the reliability of our technology for its application to a large variety of transcription factors.

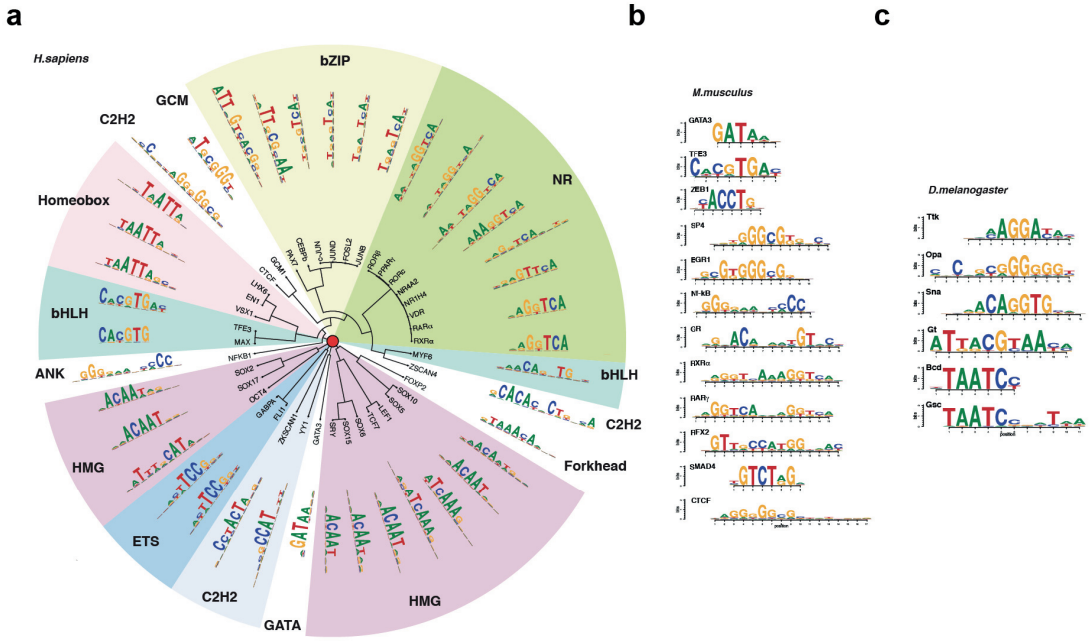


Figure 2.5: TF specificities obtained for the first 60 factors benchmark set. Copyright 2017, Nature Publishing Group (Isakova et al, 2017).

2.2.3 SMiLE-seq data reflects the energy binding landscapes of TFs

As mentioned previously, a distinctive aspect of SMiLE-seq is its ability to capture TF/DNA complexes at the steady state. We reasoned that this steady state could be reflected by the distribution and relative abundance of DNA sequences obtained by next generation sequencing, thereby providing direct information on the DNA binding energy landscapes for a given TF. In order to test this hypothesis, we considered two transcription factors, the human MAX and the mouse Erg1, for which the DNA affinity profiles had been previously obtained (Maerkl et al., 2007 and Geerz et al., 2012, respectively). Subsequently, we evaluated the k-mer distribution of the same TFs for three independent datasets obtained by HT-SELEX, PBMs and SMiLE-seq (Fig. 2.5

and Supp. Fig. 2.2). For both factors, we found that SMiLE-seq datasets show higher and reproducible correlation as compared to the other methods (Isakova et al, 2017).

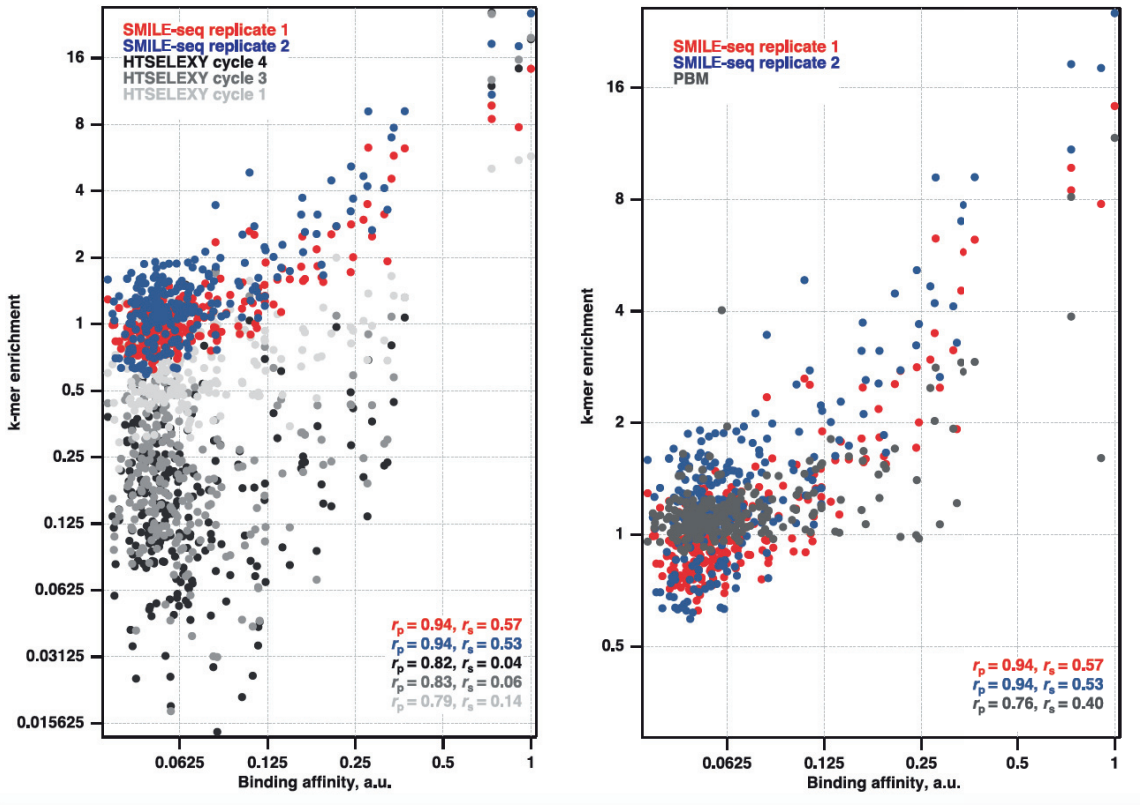


Figure 2.6: SMiLE-seq ability to reflect energy binding landscapes. Left panel: correlation analysis between k-mer enrichment and binding affinity for TF MAX derived both in SMiLE-seq and HT-SELEX. Right panel: correlation analysis between k-mer enrichment and binding affinity for TF MAX derived both in SMiLE-seq and PBM. Copyright 2017, Nature Publishing Group (Isakova et al, 2017).

2.2.4 SMiLE-seq demonstrates the feasibility of its application to KRAB-ZFPs

As mentioned in the introduction, the Krüppel-associated box domain zinc finger proteins (KRAB-ZFPs) constitute the largest family of transcriptional factors in higher vertebrates. They typically display an N-terminal KRAB domain, which is implicated in co-factor binding, and an array of DNA-binding C2H2 zinc fingers located C-terminally (Urrutia et al., 2003). Despite their large number, the DNA binding properties of most KRAB-ZFPs remains poorly characterized. The number of zinc fingers varies widely across the KRAB-ZFPs, from 4 to 24, which would suggest

an equal diversity in the size of DNA binding sites. Nevertheless, recent evidence demonstrates how the DNA binding of long tandem arrays can substantially diverge from the one-finger-three-bases rule of well-studied three-finger systems (Patel et al., 2018). Moreover, it appears that different DNA substrates can also induce distinct tandem conformations in which the individual zinc fingers re-arrange and diverge from their usual DNA-binding pattern (Patel et al., 2018). It therefore comes as no surprise that the several computational tools aimed at predicting the DNA binding motifs of C2H2 zinc finger proteins have thus far had limited success with KRAB-ZFPs (Persikov et al., 2014, Najafabadi et al., 2015, Patel et al., 2018). This multimodal and unpredictable behaviour suggests that KRAB-ZFPs may be characterized by a more complex DNA-recognition code than previously anticipated. Hence, it is necessary to devise innovative methods that can allow for flexible, sensitive and high-throughput testing of the several KRAB-ZFPs and in different DNA-binding contexts. In order to investigate whether SMiLE-seq would be applicable to the investigation of such a difficult family of TFs, we selected randomly 24 full-length KRAB-ZFPs of which 19 successfully expressed in our MITOMI-based expression system. Out of these initial set, we obtained motifs for 9 factors (Fig. 2.7, Isakova et al, 2017), i.e. with an approximate success rate of 37%, which is significantly higher than the one showed by other macroscale efforts like HT-SELEX (Yin et al., 2017 – tested 105 KRAB-ZFPs and obtained motifs for 16, i.e. with success rate of 15%). Subsequently, we evaluated to what extent the motifs that we derived were conforming to results obtained by online prediction tools. We then resorted to obtain the respective predicted motifs for all 9 TFs by using the Zinc Finger Recognition Code tool (http://zifrc.ccbr.utoronto.ca/, Najafabadi et al., 2015) and compare them to the SMiLE-seq motifs. Overall, the comparison shows a very limited ability of the only tool to predict either the main features any given motif.

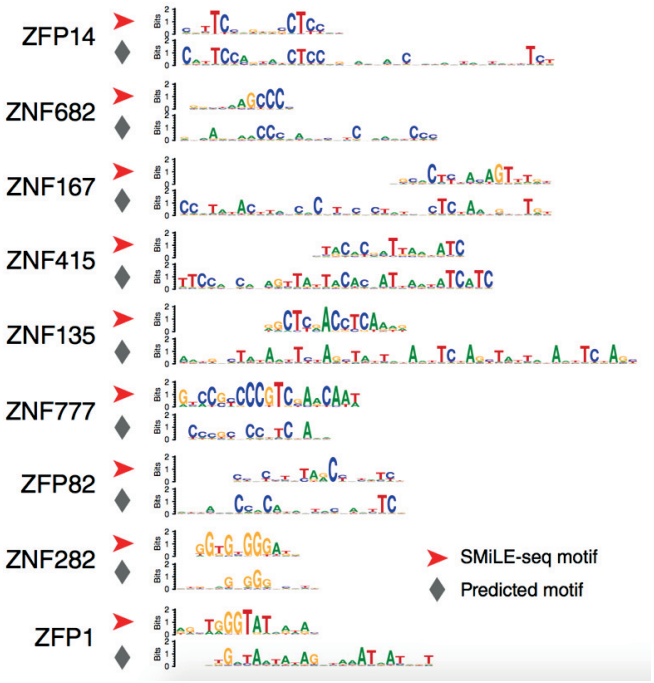


Figure 2.7: First set of SMiLE-seq derived KRAB-ZFPs motifs (Isakova et al, 2017).

These preliminary results highlight the potential of SMiLE-seq as a tool to investigate in a systematic way the complex DNA-binding patterns of KRAB-ZFPs. However, before systematically tackling the entire TF family, we set out to re-develop our tool further in order to increase the throughput, sensitivity and scope of our future studies. In particular, we aimed to address the following limitations:

- Throughput.

  The current SMiLE-seq device can process eight TFs in parallel, which is in the low end of the spectrum of throughput that can be achieved with microfluidic tools. We reason that we careful re-design, it would be possible to increase the procedural output of the device.

- Inter-experiment multiplexing.

  The original SMiLE-seq library does not contain allow for inserting molecular barcodes during the amplification step. As a consequence, two different SMiLE-seq libraries cannot be sequenced in the same Illumina sequencing run. Remarkably, the frequency with which SMiLE-seq experiments can be performed is far higher than for Illumina sequencing runs.

This creates a situation in which several SMiLE-seq libraries sit idle for weeks before they can be processed, therefore delaying the whole discovery process.

- Analysis.

  The current SMiLE-seq analysis pipeline is based on the MEME suite (Bailey et al., 1994), which is an algorithm that does not allow, for computational reasons, to conveniently process more than a few thousands reads for a given TF. Since tens or hundreds of thousands of sequences are readily available, this analytical constraint constitute a limit to the overall sensitivity of the *de novo* motif discovery process.

- Library bias.

  Randomized synthetic DNA libraries are supposed to present a uniform distribution of nucleotides and no enrichment should be detected when analysing input libraries, i.e. libraries amplified and sequenced without being enriched through SMiLE-seq. Nevertheless, we observed significant input library bias with over-enriched sequences that show statistical levels of enrichment similar to the ones obtained for low-affinity TFs. This constitutes another obstacle towards the efficient derivation of high-quality motifs from TFs such as the KRAB-ZFPs that exhibit a wide range DNA-binding modes and affinities.

The new approach that resulted from addressing all the above limitations, SMiLE-seq v2.0, provides an unbiased, high-throughput and multiplexable new version of its original implementation.

2.3 SMiLE-seq v2.0

2.3.1 The SMiLE-seq v2.0 chip

As mentioned in the previous section, the new SMiLE-seq implementation that we aimed to developed addressed four specific limitations: throughput, experiment multiplexing, analysis pipeline and library bias. In order to address the first of these limitations, i.e. throughput, we re-designed the microfluidic chip in its entirety. Our specific goals in this regard were to increase the number of MITOMI buttons that could be utilized in a single experimental run. In the first SMiLE-seq implementation, each MITOMI button was connected to bulky capillary pumps, which helped the loading of reagents while, at the same time, increasing the footprint of the microfluidic device. In SMiLE-seq v2.0, we circumvented the need for capillary pumps by connecting each sample inlet to western-blot tips, which could then, in turn, be easily connected to external macroscale pressure sources (Fig. 2.8). In practice, we reduced the area footprint of each MITOMI button by substituting capillary pumps with readily available external pressure sources and a simple but clever macro-to-micro interface. This allowed us to confine in the limited of a glass slide several 32 MITOMI buttons, hence a 4-fold increase compared to the previous SMiLE-seq version (Fig. 2.8c). The new microfluidic design consists of 4 rows of 8 MITOMI buttons each. Moreover, each button is connected to the same reagents inlets but has different reagents outlets. This enables us to perform the same MITOMI-based surface functionalization in parallel for all rows, thereby effectively performing 4 SMiLE-seq experiments at once, while at the same separately collecting the bound DNA library individually from each row. In the next section, I explain how this row-based microfluidic architecture, in combination with a new DNA library design enables cost-effective experimental multiplexing.

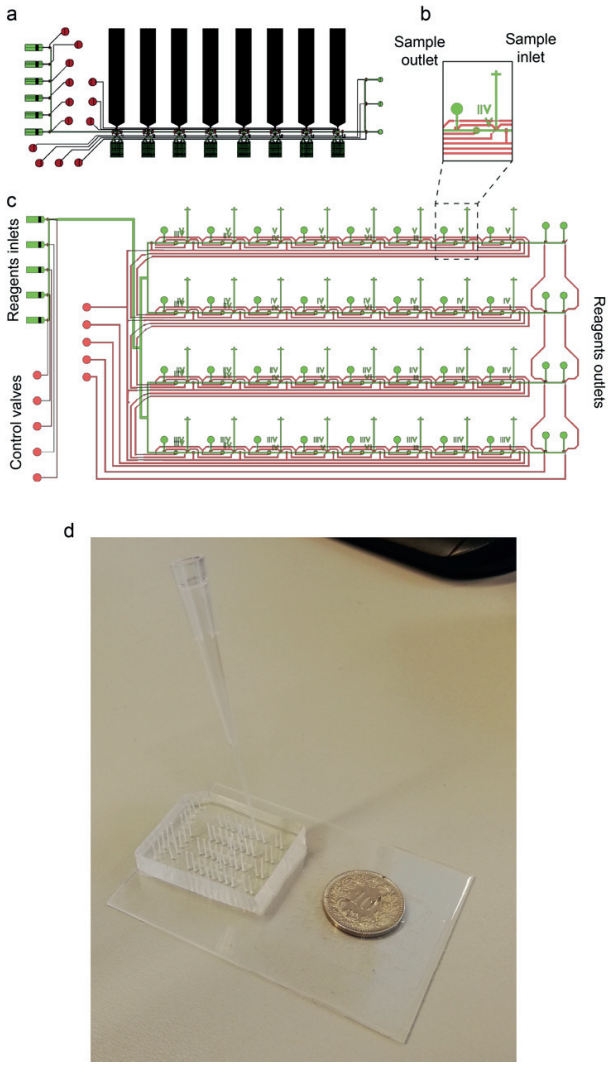


Figure 2.8: The SMiLE-seq v2.0 chip. a) Schematic layout of the original SMiLE-seq chip for comparison purposes. b) Inset of the MITOMI button unit of SMiLE-seq v2.0. c) Schematic layout of the SMiLE-seq v2.0 chip. d) Photograph of a SMiLE-seq v2.0 chip connected to a tip for sample dispensing.

2.3.2 The SMiLE-seq v2.0 library

The major drawback related to the original SMiLE-seq library design was the inability to sequence SMiLE-seq libraries at the same frequency as the one at which they could be generated. As a result, several libraries would be stored for weeks without being sequenced, which delay the whole troubleshooting and discovery process. In order to address this issue, we modified the SMiLE-seq library in order to make it compatible with the highly-multiplexed Nextera library preparation protocol (Fig. 2.9). The important modifications to the library design are:

- The introduction of Nextera primer binding sites (Nextera adapters) at each end of the DNA library sequence. This allowed for standard and readily available Nextera primers to be used for library amplification, instead of the custom designed original SMiLE-seq primers. The important aspect of Nextera primers is that they contain molecular barcodes which enables libraries amplified with different primers to be sequenced in the same sequencing run and de-multiplexed bioinformatically. As a result, the introduction of Nextera adapters in the library design allows for scalable experimental multiplexing, i.e. any number of SMiLE-seq v2.0 libraries can be sequenced in the same sequencing run, as long as they are amplified with a different set of Nextera amplification primers.
- The replacement of two 7bp barcodes with a 10bp barcode as several other independent barcoding strategies have shown to be sufficient for effective multiplexing. This barcode should not be confused with the barcodes introduced at the amplifications stage by the Nextera primers. Indeed, the Nextera primers are used to multiplex libraries, whereas the SMiLE-seq v2.0 barcodes are used to multiplex samples. In this manner, we effectively achieve a double-multiplexing process in which N_{tf} transcription factors can be sequenced in the same sequencing run, where $N_{tf}=N_{Nb}*N_{Sb}$ (N_{Nb} is the available number of nextera barcodes and N_{Sb} is the available number of SMiLE-seq barcodes). This double-multiplexing constitutes a significant improvement because it not only allows to save time by eliminating the waiting times between each SMiLE-seq library sequencing, but it also saves significant capital in terms of DNA synthesis cost.
- The expansion of the DNA random region from 30bp to 40bp, in order to accommodate TFs with potential long motifs like the KRAB ZFPs.

Overall, the new library design and its combination with the row-based microfluidic architecture of SMiLE-seq v2.0 allows to test the DNA binding specificities of TFs at a much greater throughput than and for the same cost as the previous SMiLE-seq version.

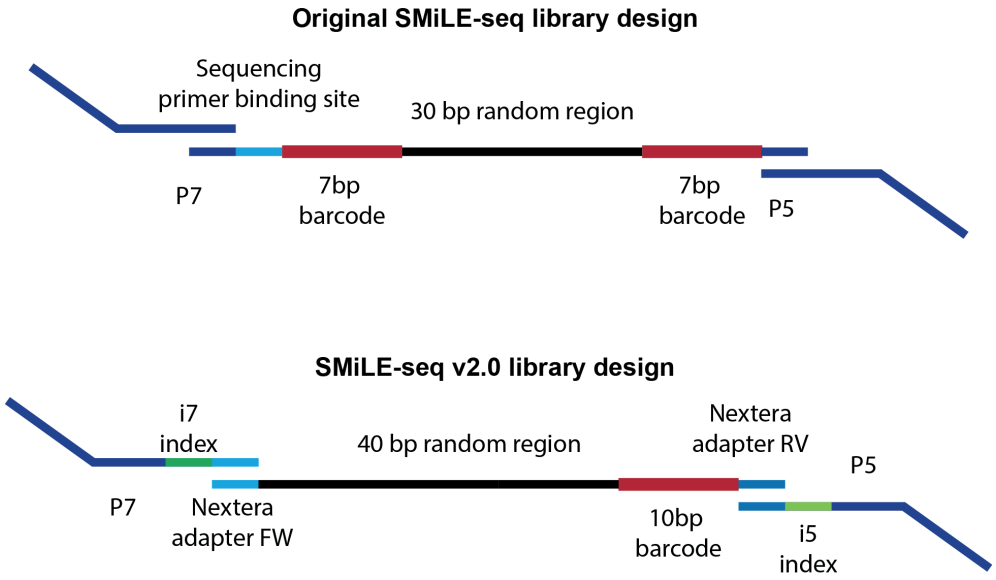


Figure 2.9: The SMiLE-seq v2.0 library design

2.3.3 The SMiLE-seq v2.0 analysis pipeline

As mentioned in the previous section, the limitations regarding the original SMiLE-seq procedure that we aimed to tackle are two-fold:

1. The original pipeline for *de novo* motif discovery was based on the command line and/or web-based tool MEME (Bailey et al., 1994). The main constraint that this tool posed was the upper limit of sequences (~1000 reads) that could be processed before the computation became prohibitively slow. This drawback has to do with the inner algorithms of the tool, for which the number of operations to perform increases quadratically with the number of input sequences and the motif size. Nevertheless, even though SMiLE-seq libraries are normally sequenced as simple spike-ins, they yield in the order of tens of thousands of sequences. As a result, by using MEME we are forced to limit the analysis to less than 10% of our sequencing data, thereby ignoring all the information contained in the remaining

90%. This drawback can prove significantly problematic for low-affinity TFs, for which enrichment of specifically bound sequences may be very low. Therefore, it is essential to establish an analysis framework that can take as input a much larger set of DNA sequences.

2. The previous SMiLE-seq pipeline also ignores the significant bias of synthetic random DNA. As a matter of fact, even if synthetic libraries are supposed to have a uniform distribution of bases and no over-represented sequences, we found that the opposite is true. In order to quantify this, we sequenced a subset of input DNA libraries without first enriching them through SMiLE-seq. Ideally, when subjected to motif discovery, these libraries should show no statistically significant enrichment of any motifs. As reported in Fig. 2.10, each input library shows different statistically significant biases, with P-values ranging from 1e-65 to 1e-22.

   Unfortunately, another drawback of the MEME suite is the inability to complement the motif discovery with background correction, a solution that would compensate for this experimental bias and decrease the likelihood of false positives.

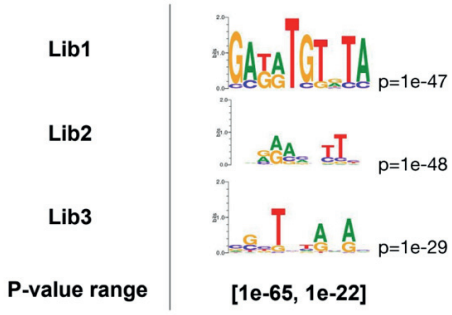


Figure 2.10: Examples of bias in random synthetic DNA libraries.

The solution that we adopted in order to address these issues consisted in shifting to a new analysis software, called HOMER (Heinz et al., 2010). HOMER allows for *de novo* motif discovery with large numbers of input DNA sequences, variable length and, importantly, also accepts as input files containing the background sequences to be corrected for. With HOMER, we can take advantage of the entire information content of the SMiLE-seq v2.0 libraries, while, at the same time, accurately discovering *de novo* motifs in an unbiased manner.

2.3.4 Generation of a synthetic methylated DNA library

It has been shown in previous studies that CpG methylation may be an important modulator of the DNA binding of transcription factors (Yin et al., 2017). In order to add this extra dimension to study of the DNA binding specificity of KRAB-ZFPs, we set out to artificially introduce CpG methylation in the SMiLE-seq v2.0 DNA libraries. To achieve this, we treated half of each of our DNA libraries with a CpG Methyltransferase (*M.SssI*) which methylates all cytosine residues of the double-stranded DNA within CpG dinucleotides (Fig. 2.11a). Before using the methylated libraries in SMiLE-seq experiments, we confirmed successful CpG methylation by methylation-specific restriction analysis. In order to do this, we run both methylated and non-methylated libraries on agarose gel before and after restriction with the enzyme *BstBI*, a frequent cutter whose restriction ability is impaired by CpG methylation. As can be seen in Fig. 2.11b, after digestion only the non-methylated library shows significant degradation, whereas the methylated library remains at the nominal size of 121bp.

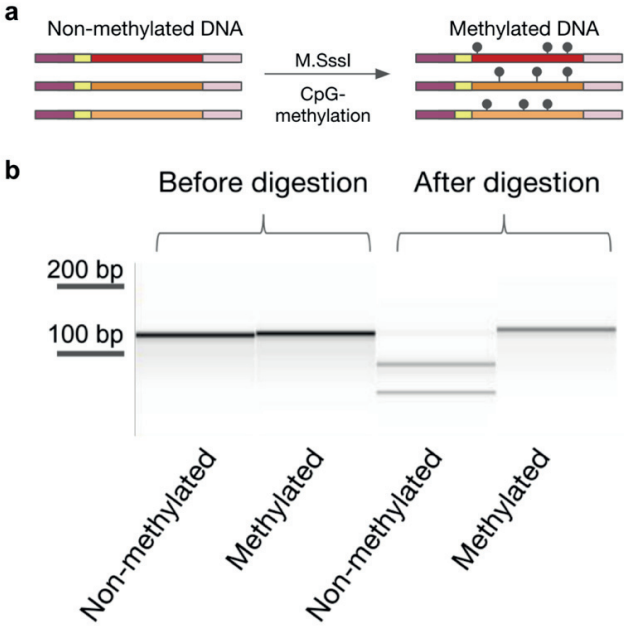


Figure 2.11: The SMiLE-seq v2.0 methylated DNA library. a) Schematic depiction of the methyltransfare-based library conversion. b) Restriction analysis of the methylated and non-methylated DNA library.

The generation of methylated DNA libraries is an important one as it increases the scope of the SMiLE-seq motif discovery process, by allowing to not only infer the sequence preferences of transcription factors but also their sensitivity to the methylation state of the DNA substrate.

2.3.5 The DNA binding specificities of KRAB ZFPs

After establishing a more powerful experimental framework with a new chip, new library design and new analysis pipeline we set out to process an initial set of 101 KRAB ZFPs. These factors were chosen based on motif size considerations made in light of a preliminary analysis of recently published ChIP-exo data on KRAB-ZFPs (Imbeault et al., 2017). We derived motifs for all KRAB-ZFPs using their respective peak files and subsequently sorted each motif based on size (Supp. Fig. 2.3). We reasoned that factors with very long motifs – i.e. from 30 to 50 basepairs – would have low probability of yielding motifs in our experimental setting due to the difficulty of covering the extremely high number of sequence combinations. Therefore, for SMiLE-seq, we prioritized factors with shorter *in vivo* - i.e. derived from ChIP-exo - motifs.
Of the initial set of 101 KRAB-ZFPs probed through SMiLE-seq, we obtained enriched motifs for 43 factors (Supp. Table 1); of these, 22 yielded motifs for both methylated and non-methylated DNA (referred to as met-Independent factors), 10 only for methylated DNA (methylation sensitive type A or metA factors), 11 only for non-methylated DNA (methylation sensitive type A metB factors).

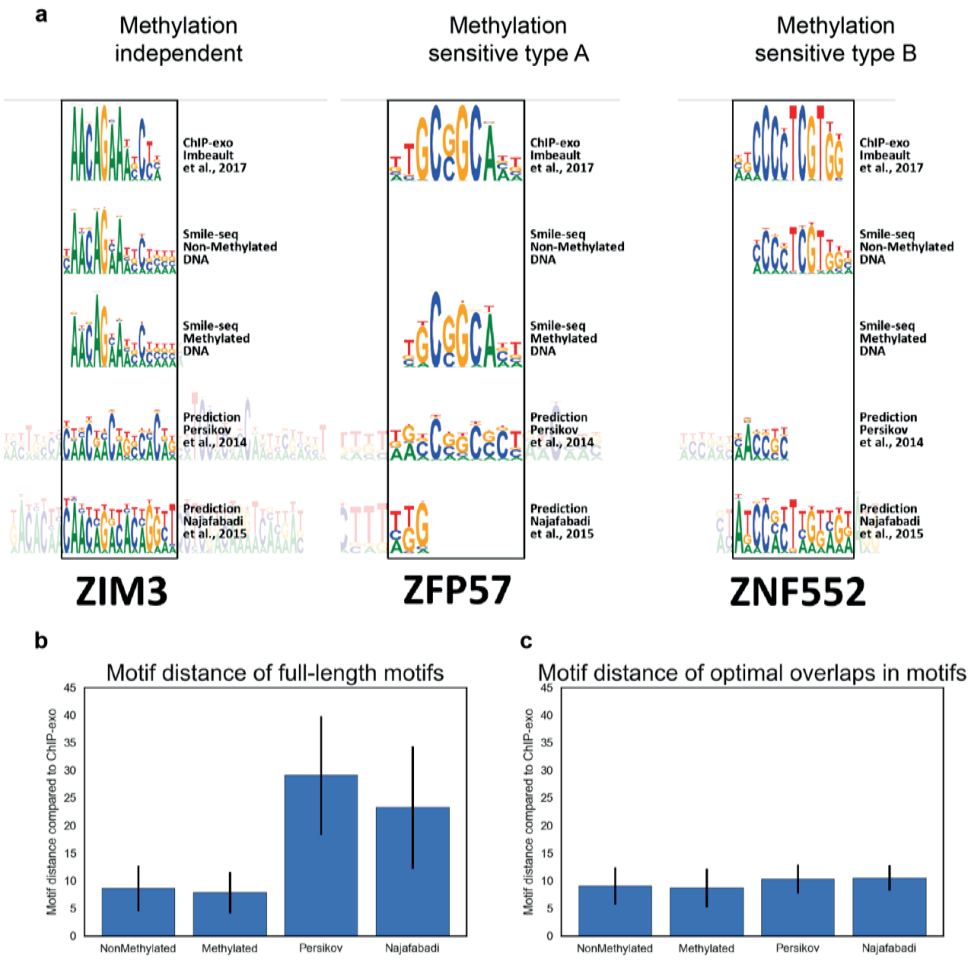


Figure 2.12: Example of the three types of methylation sensitivity observed in our SMiLE-seq results.

In order to validate the motifs obtained, we compared them to the respective ChIP-exo motifs as well as the motifs predicted using two dedicated web-based tools (Persikov et al., 2014, Najafabadi et al., 2015) (Fig 2.12a., Supp. Fig. 2.4). We observed that, while SMiLE-seq faithfully recapitulates ChIP-exo motifs, the predicted motifs fell short of predicting the correct motif in most cases (Fig. 2.12b). However, we reasoned that the online tools tend to predict motifs based on the assumption that all zinc fingers bind DNA – which is reflected by the fact that the predicted motifs are much longer than the actual SMiLE-seq and ChIP-exo motifs. Moreover, by manual inspection we noticed that certain portions of the predicted motifs were similar to the *in vitro* and *in vivo* results. We therefore wondered whether the low performance the prediction tools were due

mostly to overestimating the size of the motifs and not the DNA-binding specificities of individual zinc fingers. To test this hypothesis, we aligned the predicted motifs to the ChIP-exo ones and calculated the motif difference considering only the overlapping portion of the aligned motif. Indeed, we observed that the overall similarity of the aligned motifs was much higher and became comparable to the similarity between ChIP-exo and SMiLE-seq motifs (Fig. 2.12b). This suggests that the major limitation of available prediction tools is not the ability to predict the specificity of individual zinc fingers or zinc finger arrays. Rather the main challenge is to understand which zinc fingers actually contribute to DNA binding and which ones don't. In this direction, we set out to systematically identify for each KRAB-ZFP the subset of zinc fingers that likely contributed to binding. In order to achieve this, we took advantage of the ability of the web tool provided by Persikov et al. of providing the predicted binding specificity of any submitted zinc finger array and, for each TF, we queried the tool with different subsets of consecutive zinc fingers. After downloading the predicted motif for each zinc finger subset, we selected the one that was most similar to the ChIP-exo and SMiLE-seq motifs, thus effectively shortlisting which zinc fingers are predicted to directly bind DNA. For instance, the KRAB-ZFP ZIM3 contains 11 zinc fingers and the predicted motif is 33 base pairs long. Nevertheless, the *in vivo* and *in vitro* motifs are much shorter (15bp) which suggests that a minimum of 5 out of the 11 zinc fingers contributes to binding. By applying the systematic approach described above we predicted that zinc fingers 5-9 are likely to be the ones responsible for the observed DNA specificity of the factor (Fig. 2.13a). Similar observations were made for all the other KRAB-ZFPs tested with SMiLE-seq (Supp. Fig. 2.5), including metA and metB factors such as ZFP57 and ZNF133, respectively (Fig. 2.13b,c).

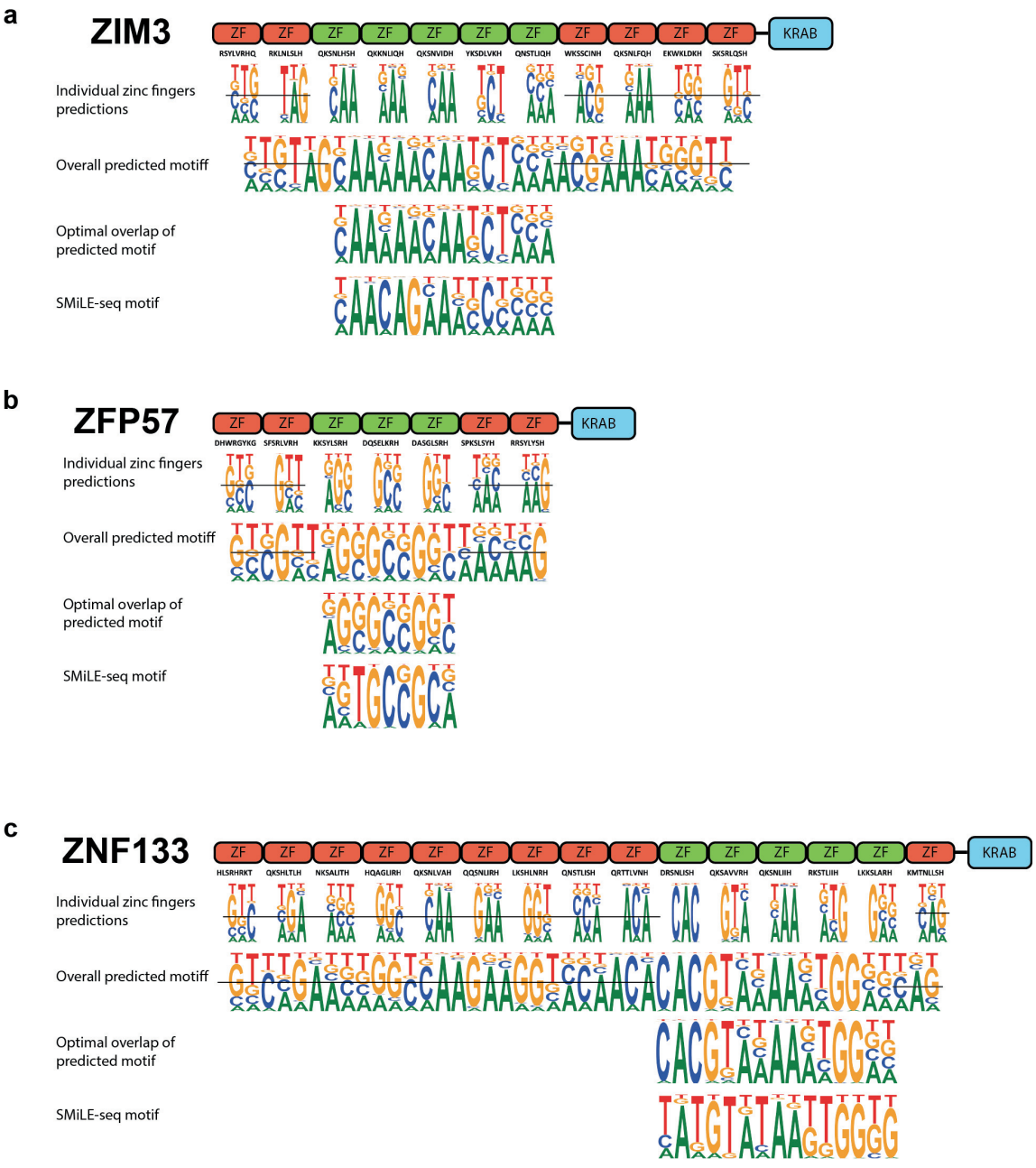


Figure 2.13: Example of zinc finger binding prediction for three KRAB-ZFPs, i.e. ZIM3 (a), ZFP57 (b) and ZNF133 (c).

In order to explore the origins of the observed sensitivity of some factors to methylation, we took into consideration 5 type of phylogenetic trees based on different types of similarities between the examined factors: 1) similarity based on full protein sequence, 2) similarity based on the sequence of all zinc fingers, 3) similarity between the alpha helices of all zinc fingers, 4) similarity between the alpha helices of the zinc fingers that are predicted to bind DNA and 5) similarity between experimentally derived motifs. Surprisingly, none of these approaches successfully clustered the methylation-sensitive factors in specific subgroups (Fig. 2.14, Supp. Fig. 2.6).

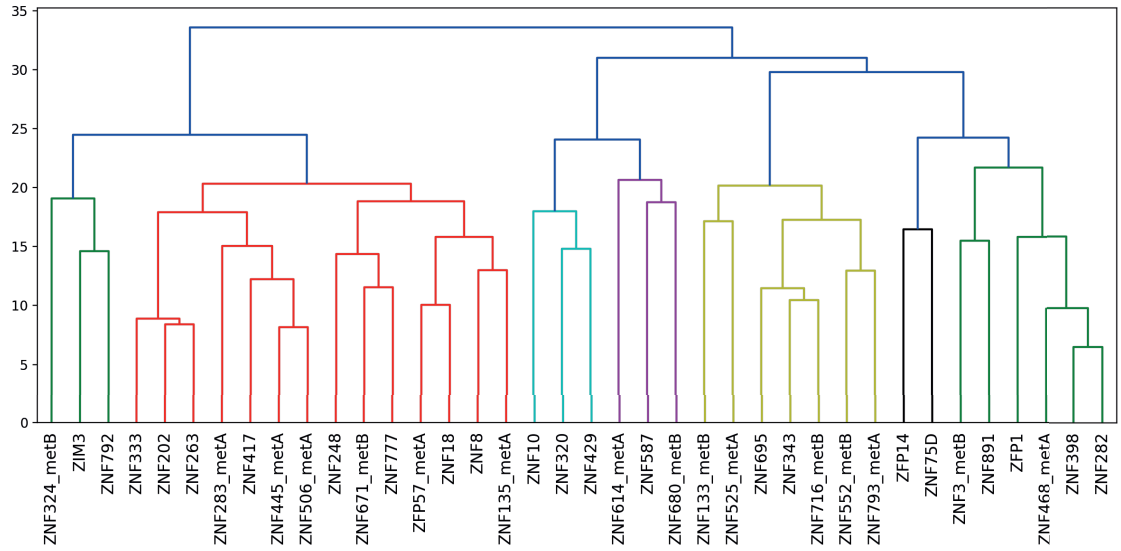


Figure 2.14: Dendrogram obtained considering the similarity between experimentally derived motifs.

Therefore, based on our preliminary results it appears that methylation sensitivity did not evolve from a common ancestor and that methylation can have different effects even on members of the same family of paralogs (Supp. Fig. 1.6). Another puzzling observation we made was that several of the factors that we identified as methylation sensitive, did not display the canonical CG dinucleotide. By analyzing the dinucleotide distribution in all the ChIP-exo derived motifs we also observed a significant bias against CG as compared to other dinucleotides (Fig. 2.15).

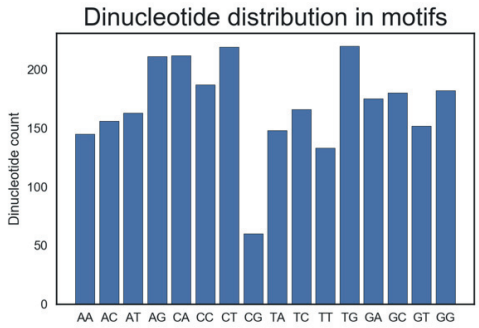


Figure 2.15: Dinucleotide distribution in ChIP-exo motifs.

Despite being consistent with the phenomenon of CG suppression and underrepresentation in the human genome (Lander et al., 2001), this observation is somehow at odds with the importance of KRAB-ZFPs in controlling imprinting and mediating DNA methylation (Ecco et al., 2017). It is also remarkable to notice that prediction tools tend to predict quite accurately the binding specificity of both non- and methylation sensitive factors (including the known ZFP57) without integrating information about methylation specificity at all. This suggests that the methylation sensitivity of certain factors might originate independently from the binding specificity conferred by their respective ZF contacts.

2.4 Discussion

In this chapter, I presented the development and optimization of a microfluidic-based tool, SMiLE-seq, for the study of DNA binding specificities of transcription factors *in vitro*. In SMiLE-seq we adopt and merge features of both HT-SELEX and MITOMI. As a result, our solution allows to screen a very large amount of DNA binders in a sensitive and miniaturized fashion. With the first version of this technology, we set out to demonstrate its feasibility by analysing several factors belonging to different species and TF families. Of the randomly selected 60 factors, 58 (6 Drosophila, 12 mouse and 40 humans) yielded the expected motif with high reproducibility. Moreover, by comparing SMiLE-seq data and the experimentally derived binding affinity of the factors MAX and Erg1, we found the SMiLE-seq recapitulates better than existing methods the energy landscapes of TF/DNA binding. After successfully passing the proof-of-principle stage, we proceeded to investigate the applicability SMiLE-seq to a family of transcription factors, the KRAB-ZFPs, for which the vast majority of DNA binding motifs are still missing. In order to do this, we processed with SMiLE-seq 24 KRAB-ZFPs, of which 9 yielded highly enriched motifs that closely reflected *in vivo* binding. Given the low success rate with this initial set of KRAB-ZFPs, we set out to further develop our technology by exploring ways to increase its sensitivity and throughput. As a result of these efforts, we arrived at a new version of SMiLE-seq, called SMiLE-seq v2.0 that presents the following improvements: 1) a 4-fold increase in the number of factors processed for each microfluidic chip; 2) Nextera-compatible libraries which allow for any number of SMiLE-seq experiments to be processed in one sequencing run; 3) a longer random region in the DNA library in order to expand the combinatorial space covered; 4) methylated DNA libraries in order to investigate the sensitivity of TFs to cytosine methylation; 5) a new analysis pipeline that allows for background-corrected *de novo* motif discovery with a much larger number of input sequences. With SMiLE-seq v2.0, we processed 101 KRAB-ZFPs for both methylated and non-methylated DNA configurations, for a total of 202 experiments. We retrieved highly enriched motifs for 43 factors – 22 yielded motifs for both non- and methylated DNA, 10 only for methylated DNA and 11 only for non-methylated DNA – therefore providing the largest individual dataset on the *in vitro* binding properties of KRAB-ZFPs. All of the motifs thus obtained faithfully resembled motifs originated from the analysis of a large KRAB-ZFP-based ChIP-exo dataset recently published (Imbeault et al., 2017). Subsequently, we compared ChIP-exo and SMiLE-seq motifs to the ones obtained using two dedicated prediction tools (Najafabadi et al., 2015, Persikov

et al., 2014). We noticed that both tools systematically overestimated the size of the experimentally derived motifs while, at the same time, closely resembling these motifs only in specific portions of their predictions. By isolating these "highly similar" portions for each factor, we found that overall the main limitation of existing prediction tools is not the ability to predict the DNA specificity of zinc fingers, but rather the ability to discern which fingers contribute to DNA binding and which ones do not. By mapping the optimal portions of predicted motifs to the responsible zinc fingers, we were able to identify for each factor the specific subset of zinc fingers that were likely to be responsible for binding. It has previously been noted that KRAB-ZFPs may present motifs that are much shorter than as expected by the number of zinc fingers in each factor (Ecco et al., 2017). Our preliminary results confirm this view: considering the 43 SMiLE-seq-positive factors, the average number of zinc fingers per factor is 10±3 whereas the predicted number of zinc fingers directly contributing to binding is only 4.5±0.7. Phylogenetic analysis based on different similarity criteria of the methylation specific factors did not provide, for the moment, insight in the evolutionary origin of methylation sensitivity. Moreover, we also noticed that several of the identified methylation specific factors do not contain the canonical CG dinucleotide in their motif, which is consistent with the underrepresentation of the same dinucleotide considering all experimental motifs and the human genome dinucleotide distribution as well. These observations, together with the fact that prediction tools seem to perform well without integrating any methylation-specific information, raises more questions on the mechanisms of methylation detection of these factors. While further experimental validation will be required to validate these results, through SMiLE-seq applied to truncated TF forms or MITOMI to confirm methylation specific factors, we believe the presented framework will prove a powerful addition towards deciphering the DNA recognition code of KRAB-ZFPs.

2.5 Methods

2.5.1 Protein expression

Every SMiLE-seq experiment begins with the *in vitro* expression of a GFP-tagged transcription factor. After testing different systems, we identified the TnT® Coupled Wheat Germ Extract System from Promega as the most reproducible and robust solution for human transcription factors expression. In order to achieve high quantities of GFP-tagged TFs, we subcloned by LR reaction (Gateway cloning system) the TFs of interest into a custom-made vector called pF3a-eGFP, which contains translation enhancers (TE) from the wheat-germ-compatible barley yellow dwarf virus (BYDV) and a C-terminal eGFP tag.

2.5.2 Target DNA library preparation

The random DNA library was prepared through primer extension of single stranded Ultramers® ordered directly from IDT. For both SMiLE-seq and SMiLE-seq v2.0 the Ultramers® were extended by a Cy5-labelled primer (/5Cy5/CAA GCA GAA GAC GGC ATA CG, also from IDT) and a Klenow-based extension reaction (NEB Cat No M0212).
The library synthesis occurs with the following reaction:

- 5µl Buffer 2 (NEB)
- 5µl dNTPs
- 0.5µl Cy5 labeling primer (500 µM)
- 1.5µl Library oligos (200 µM)
- 37µl dH2O

And the following thermal cycle:

- 94°C for 5 min
- 50°C for 60 sec
- place tubes on ice
- add 1µl of Klenow 3' – 5' exo- (NEB Cat No M0212)

- 37°C for 60 min
- keep at 0°C

Use MinElute (Qiagen) to purify the double-stranded libraries, elute in 12 µl of EB. Dilute the libraries 1:10 in ddH2O or elution buffer (Qiagen).

2.5.3 Chip fabrication

Microfluidic devices were fabricated using standard multilayer soft-lithography, similarly to previous examples (Maerkl and Quake, 2007, Unger et al. 2000). To these end, two layouts were designed using a dedicated software (L-Edit Tanner Tools), one for the so-called control layer and one for the so-called flow layer. Both layouts were transferred onto chromium masks using photolithography. For the flow layer, the layout was transferred onto a silicon wafer coated with AZ9260 positive resist 10 µm of thickness. After development, the positive resist microstructures were re-flown for 1 minute at 130°C in order to obtain round-shaped channels, which are required for the correct functioning of the microfluidic microvalves. For the control layer, the layout was transferred onto a silicon wafer coated with SU8 negative resist 10µm of thickness. The microfluidic flow layer, PDMS at 5:1 (prepolymer/curing agent w/w ration) was cast and cured on the flow layer wafer at a thickness of 4mm, whereas for the control layer the PDMS 20:1 was spun onto the control wafer in order to achieve a thickness of ~20µm. Both layers were partially cured at 80°C for 30 minutes, then the flow layer was peeled off its wafer and aligned on the control layer. Curing was finalized for a minimum of 90 minutes at 80°C. After curing, the assembled device was peeled off the control layer and holes were punched with a reusable biopsy punch. After punching the device was cleaned with isopropanol and bonded to a glass slide by plasma treatment.

2.5.4 Control lines preparation

All pressure sources are connected to the same number of plastic tubes. Each tube is color-coded as in 2.17d. In order to connect the tubes of the control lines to the chip, fill the tubes with water using a syringe (the water should fill ~2-5 cm of the tube). Connect all the control lines one-by-one by plugging the tubes into the chip. Actuate one-by-one each valve. After a 10-30 seconds, the water should start reaching the control channels which should be visible under the microscope.

It is important to make sure that the valves close according to the predefined experimental scheme and that all valves are completely filled with water.

2.5.5 Surface preparation

For the SMiLE-seq surface preparation the following steps are required:

1. Collect and place on ice the following reagents:
   - Biotin-BSA at 2mg/ml
   - Neutravidin at 1mg/ml
   - PBS
   - Anti-GFP antibody aliquots (1:50 dilution from 100ug/µl stock)

2. Take 20 µl of BSA and 15 µl of PBS, NA and anti-GFP using a western-blot pipette tip and place them in the reagent inlets of the chip. Connect each tip the with blue pressure tubes coming from the flow line pressure manifold.
3. When all tips and tubes have been properly connected (all valves are still closed) bring the pressure of flow line to 8PSI.
4. Open one-by-one the inlet pressure channels and check visually that the order of the scheme has been respected. When an inlet channel gets pressurized, the respective valve is slightly “pushed up” by its inlet channel.
5. When all inlets are pressurized, release the biotin-BSA channel by opening the respective valve and let the biotin-BSA go everywhere on chip and remove the air in the main channels.
6. Wait for 10-15 minutes until the inlet areas of the other reagents gets filled with the respective reagents and no air is left in the chip.
7. Set timer to 10 minutes and start it. Bring inlet pressure to 2.5psi, mark on the pipette tips with a marker the position of the reagents (in order to keep track of how much liquid is introduced in the chip. It is very important not to let any air get inside the chip.
8. The SMiLE-seq surface functionalization sequence is as follows:
   - 10 min of biotin-BSA.

- 1min of PBS wash.
- 10 min of Neutravidin
- PBS wash for 1 min and then close MITOMI button.
- Wash another minute with PBS, then 10min of biotin-BSA.
- 1min PBS wash.
- Antibody 1 minute.
- After 1 min of Antibody, open the MITOMI button and wait 10min.
- After 10 min of Antibody, 1min PBS wash.
- Extract all western blot tips from the chip (apart from PBS tip).

2.5.6 Sample loading

Add 4 µl of DNA library to the pre-incubated *in vitro* expression mix, vortex, centrifuge to let all possible debris go to the bottom (14'000 rcf, 4 min). For each inlet, take 6-8 µl of TF using a western pipette tip and load each sample in a different 2mm inlet hole. The samples are ejected in the hole and the tip is discarded. Make sure you are touching the bottom of the chip while ejecting the liquid. Extract slowly the pipette tip while you are ejecting the samples in order to avoid that the liquids tops off the inlet hole.

Once all samples have been transferred to the 2mm inlet holes, close the MITOMI button. By using the little "air gun" connected to the flow line pressure, start pushing the samples in the chip. This is done by opening the channel connected to it and placing the air gun over one of the 2mm inlet holes. The pressure created by the air coming out of the gun (around 3-4 PSI) in this case is going to push the liquid from the hole to its respective outlet. 2-10 seconds after placing the gun on top of a 2mm inlet hole, you should see some liquid coming of its respective outlet; when this happens move on to the next sample. After priming all sample inlets, open the MITOMI button and push in the rest samples using the air gun (pressure 2-6 PSI). Be careful not to let any air be introduced in the chip. It is not essential to push exactly all of the sample liquid inside the chip, i.e. after 70-80% of the sample has been pushed into the chip, enough TF should have been trapped under the MITOMI button. Check regularly how much liquid is left in the 2mm holes by tilting the chip and looking from the side at the 2mm inlet holes or looking at the level of liquid in the

western blot tips if you used them. After all samples have been introduced into the chip, leave the chip at RT for around 2 hours in order for the TF/Ab complexes to reach steady state.

2.5.7 Chip washing

When the 2 hours have passed turn on the microarray scanner and login into the scanner computer. Close the MITOMI button B3 and wash all the MITOMI button areas with PBS. Before scanning, open the MITOMI button for 30 seconds and close again (this is done to prevent elution of non-bound DNA trapped under the button). Wash extensively with PBS in order to remove all unbound DNA.

2.5.8 Chip scanning

In order to confirm the correct expression and pull-down of the factor the microfluidic SMiLE-seq chip can be scanned with a microarray scanner. To do this, pull off the pipette tips and wash briefly the external surface of the chip with ethanol and a tissue. Insert the chip into the scanner and wait for the lamp to reach the appropriate temperature. Select the filter A488 and initialise scan. When the MITOMI buttons in use are localised, a bright white circle in the middle of the button should be visible. It is recommended to also scan for the Cy5 signal in each button in order to identify specific sources of contamination, e.g. dust particles, onto which non-specific DNA is accumulated.

2.5.9 DNA elution and library preparation

Take 30 µl of SMiLE-seq Elution Buffer with a western blot tip, insert into A5 inlet. Open B1, Assemble the system for elution: a metal pin is inserted into the plastic tube on one end and into the library outlet of the chip on the other end. Purify with a PCR purification kit and elute with 20 µl of EB.

For sample amplification, two qPCR runs are needed – one with volume 15µl to determine the number of cycles for amplification, and a second one of 50µl for final amplification.

Make the following mix for both reaction (total volume 65ul):

NEBNext Ultra II Master Mix	32.5 µl
Purified DNA	20 µl
Primers (i5 and i7 Illumina adapter sequences)	0.5 µl per primer
SYBR (1X)	0.5 µl
EB (elution buffer)	to 65 µl

The first qPCR consists of 5 cycles for the whole volume of the reaction (65µl) with the following settings:

Temperature (°C)	Time	N of cycles
72	5 min	1
98	30 sec	1
98	10 sec	5
63	30 sec	
72	1 min	
72	1 min	1

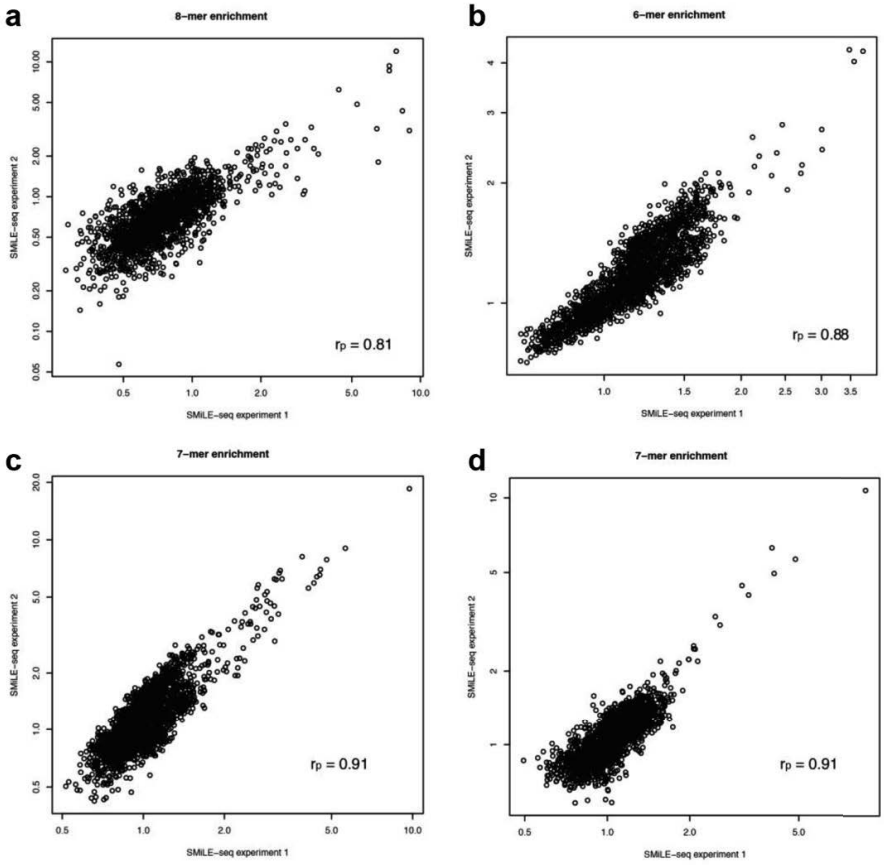
In the second qPCR step 15 µl out the initial 65 µl are cycled for 20 cycles in order to determine the correct number of amplification cycles in order to avoid overamplification. Finally, the remaining 50 µl are amplified for the determined number of cycles.

Before sequencing, the amplified DNA is purified, its size distribution and concentration measured through Fragment Analyzer (High Sensitivity NGS Fragment Analysis Kit), and Qubit (dsDNA HS Assay Kit), respectively.

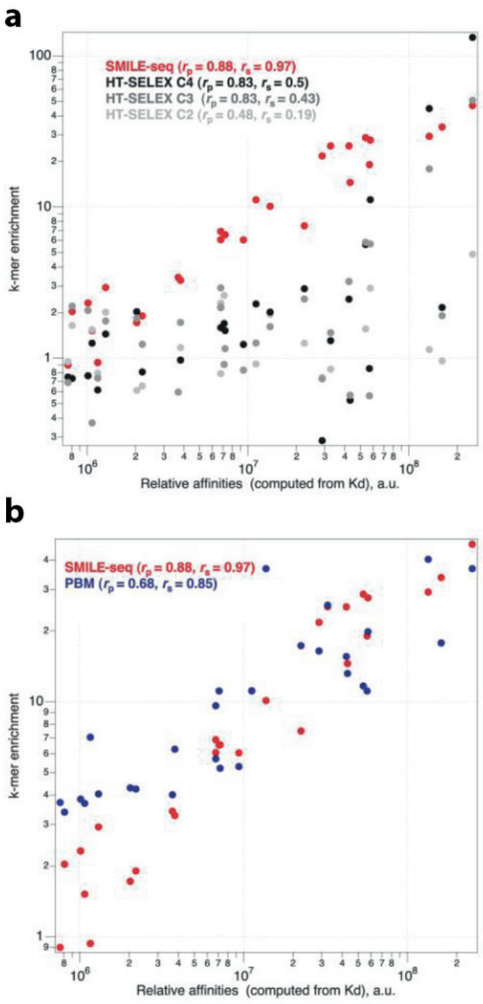
2.5.10 Analysis

DNA reads are demultiplexed and duplicates are removed using FastxTools scripts. PWM are calculated using the HOMER suite and especially the command *findMotifs.pl*. As input to the command, we used SMiLE-seq libraries of up to 100'000 reads each. For background compensation, the same number of reads taken from an "input library", i.e. libraries not enriched by SMiLE-seq, was used for each TF analysed. Recovered motifs are compared to existing data as well as predictions by means of custom Python scripts. In particular, we programmed scripts that systematically queried the online tools with the KRAB-ZPs protein sequences and downloaded the response in PWM form.

2.6 Supplementary Figures



Supplementary Figure 2.1: SMiLE-seq reproducibility. Pearson correlation of the top 2000 k-mers, from two independent SMiLE-seq experiments for the TFs PAX7 (a), SRY (b), MAX (c) and FLI1 (d).



Supplementary Figure 2.2: PBM, HT-SELEX and SMiLE-seq data correlation. a) correlation analysis between k-mer enrichment and binding affinity for TF Egr1 derived both in SMiLE-seq and HT-SELEX. Right panel: correlation analysis between k-mer enrichment and binding affinity for TF Egr1 derived both in SMiLE-seq and PBM. Copyright 2017, Nature Publishing Group.

ZKSCAN3

ZKSCAN5

ZNF2

ZNF3

ZNF7

ZNF8

ZNF10

ZNF17

ZNF18

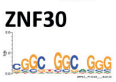
ZNF19

ZNF25

ZNF26

ZNF28

ZNF30



ZNF33A

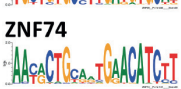
ZNF33B

ZNF44

ZNF45

ZNF69

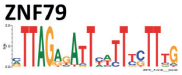
ZNF74



ZNF75D

ZNF77

ZNF79



ZNF81

ZNF84

ZNF85

ZNF90

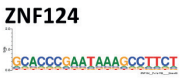
ZNF93

ZNF100

ZNF101

ZNF114

ZNF124



ZNF132

ZNF133

ZNF135

ZNF136

ZNF141

ZNF154

ZNF169

ZIM2%5F1

ZNF157

ZNF182

ZNF184

ZNF189

ZNF197

ZNF202

ZNF205

ZNF211

ZNF212

ZNF214

ZNF223

ZNF12

ZFP1

ZFP14

ZFP57

ZFP69

ZFP69B

ZFP90

ZIK1

ZIM3

ZKSCAN2

ZNF331

ZNF333

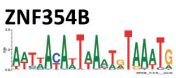
ZNF334

ZNF337

ZNF343

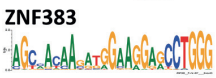
ZNF354A

ZNF354B



ZNF382

ZNF383



ZNF398

ZNF417

ZNF419

ZNF425

ZNF429

ZNF430

ZNF431

ZNF432

ZNF433

ZNF273

ZNF418

ZNF248

ZNF254

ZNF257

ZNF263

ZNF264

ZNF266

ZNF267

ZNF274

ZNF282

ZNF283

ZNF284

ZNF285

ZNF287

ZNF300

ZNF302

ZNF304

ZNF311

ZNF317

ZNF320

ZNF324

ZNF324B

ZNF486

ZNF487P

ZNF492

ZNF496

ZNF506

ZNF519

ZNF525

ZNF527

ZNF528

ZNF529

ZNF530

ZNF534

ZNF540

ZNF484

ZNF222

ZNF180

ZNF181

ZNF224

ZNF225

ZNF235

ZNF543

ZNF439

ZNF440

ZNF441

ZNF442

ZNF443

ZNF445

ZNF454

ZNF460

ZNF468

ZNF479

ZNF480

ZNF483

ZNF485

ZNF610

ZNF764

ZNF776

ZNF846

ZNF587

ZNF605

ZNF611

ZNF613

ZNF614

ZNF615

ZNF616

ZNF619

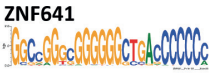
ZNF620

ZNF621

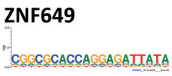
ZNF626

ZNF627

ZNF641



ZNF649



ZNF662

ZNF667

ZNF669

ZNF671

ZNF674

ZNF675

ZNF680

ZNF547

ZNF548

ZNF549

ZNF550

ZNF552

ZNF555

ZNF557

ZNF558

ZNF561

ZNF562

ZNF564

ZNF565

ZNF566

ZNF567

ZNF570

ZNF571

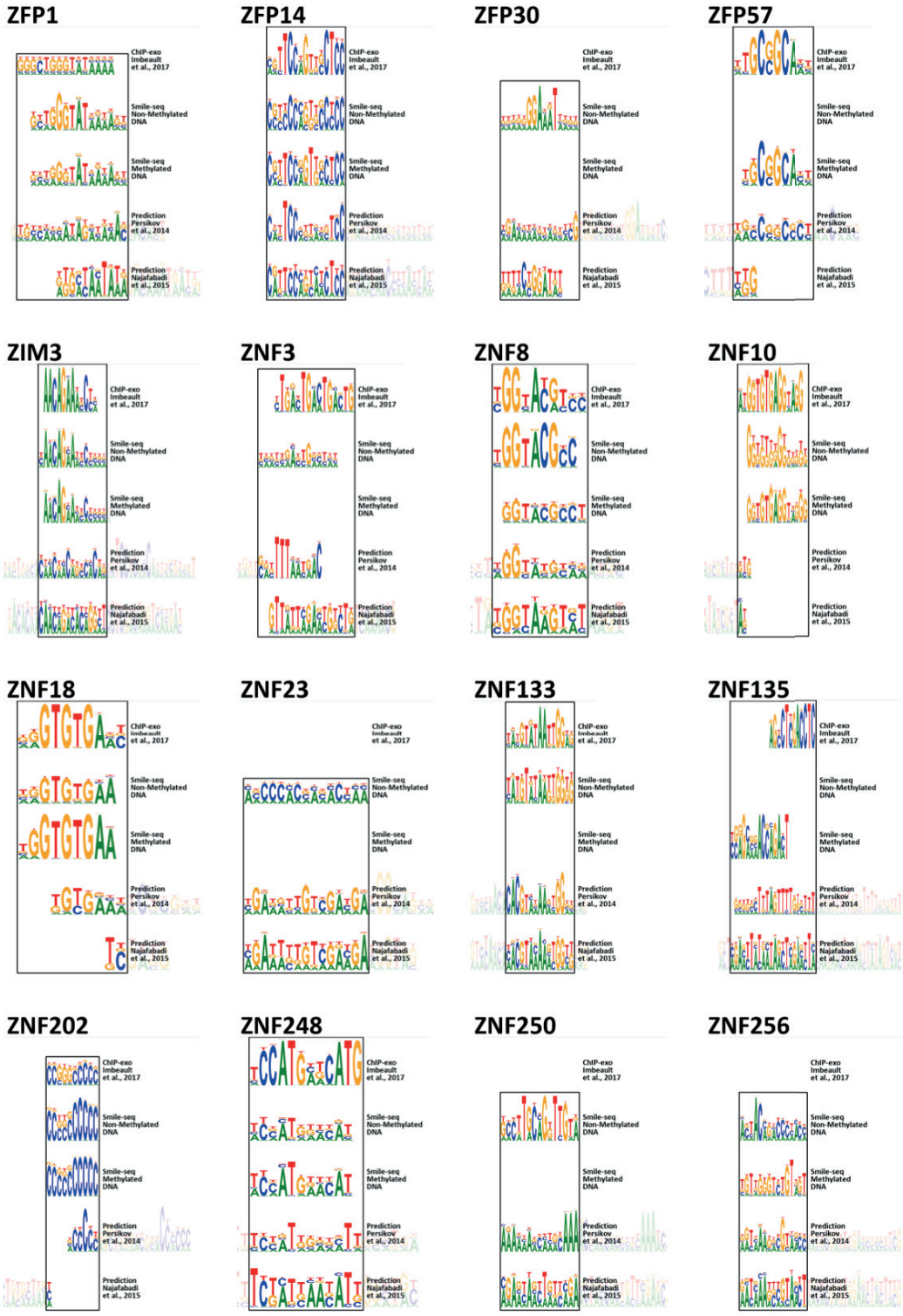
ZNF573

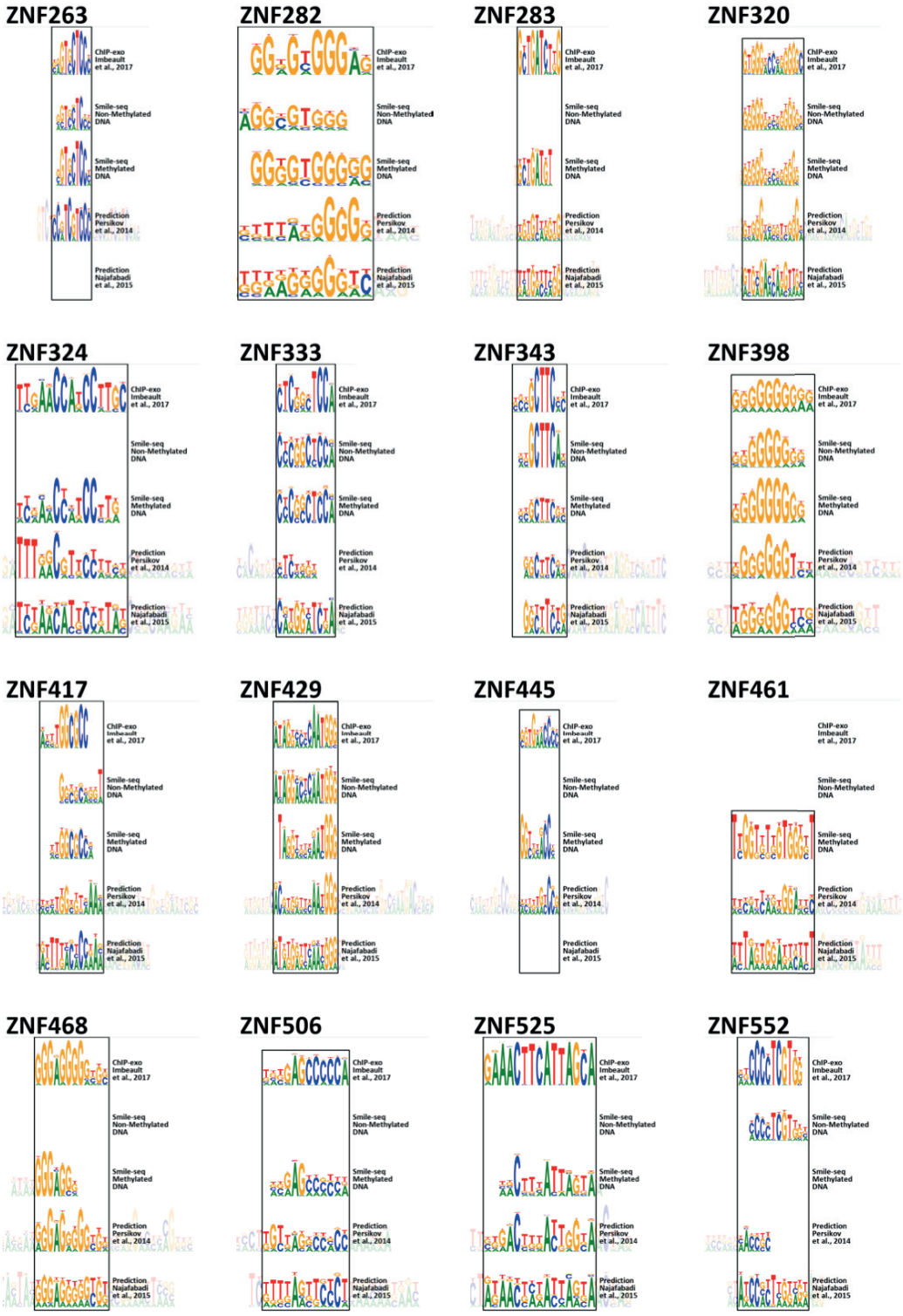
ZNF582

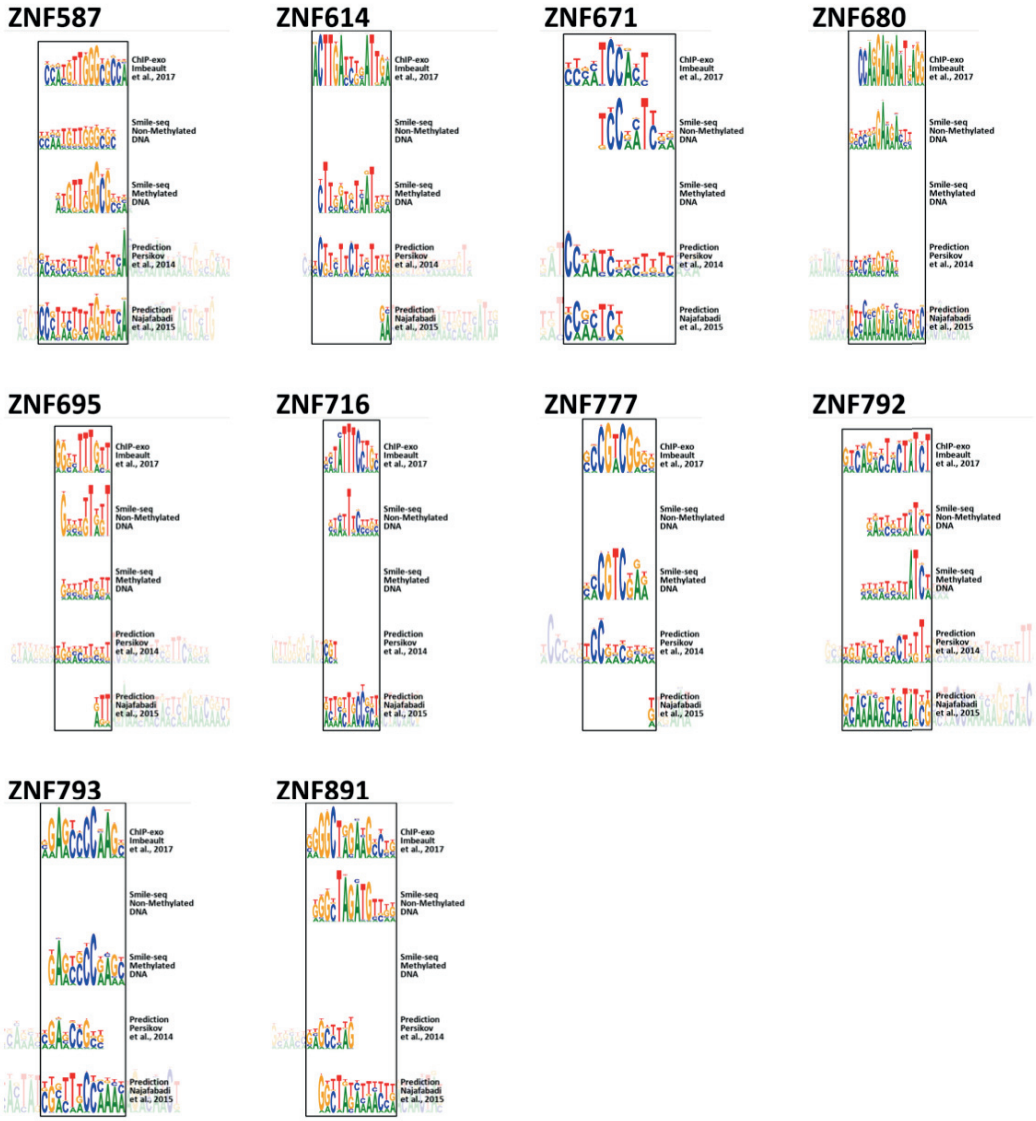
ZNF584

ZNF585A

Supplementary Figure 2.3: KRAB-ZFPs DNA-binding motifs derived from ChIP-exo data.


ZFP1
ZFP14
ZFP30
ZFP57
ZIM3
ZNF3
ZNF8
ZNF10
ZNF18
ZNF23
ZNF133
ZNF135
ZNF202
ZNF248
ZNF250
ZNF256
ChIP-exo Imbeault et al., 2017
Smile-seq Non-Methylated DNA
Smile-seq Methylated DNA
Prediction Persikov et al., 2014
Prediction Najafabadi et al., 2015


ZNF263
ZNF282
ZNF283
ZNF320
ZNF324
ZNF333
ZNF343
ZNF398
ZNF417
ZNF429
ZNF445
ZNF461
ZNF468
ZNF506
ZNF525
ZNF552
ChIP-exo Imbeault et al., 2017
Smile-seq Non-Methylated DNA
Smile-seq Methylated DNA
Prediction Persikov et al., 2014
Prediction Najafabadi et al., 2015



Supplementary Figure 2.4: Motifs obtained with SMiLE-seq and respective ChIP-exo and predicted motifs.

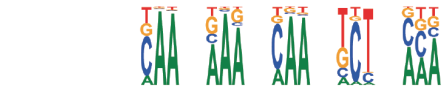
ZNF202
CNSHLVRH RSSHLARH WTSDLVRH QKSVLTTH EHRRYLAH HSAAFAKH RRDHLVRH RGYHLIRH

ZFP1
RKSRLSVH QRSTLRLH QNSQLIIH KKSNLIIH QKFELTTH HQSNLIVH HKANLIKH FKSLLISH

ZFP14
RRSTLSQH QRAHLIRH VLQELTQH VHQQLARH QCTHLTRH CGPDLRVH ICQQLTVH LRQQLVRH SYSQLISH LLSQLTQH LYSFLTQH QHSHLTQH

ZFP57
DHWRGYKG SFSRLVRH KKSYLSRH DQSELKRH DASGLSRH SPKSLSYH RRSYLYSH

ZIM3
RSYLVRHQ RKLNLSLH QKSNLHSH QKKNLIQH QKSNVIDH YKSDLVKH QNSTLIQH WKSSCINH QKSNLFQH EKWKLDKH SKSRLQSH

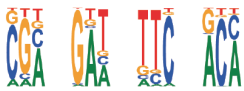


ZNF3
YSSALVRH YSSGLIQH QSSHLYQH RSSTLIHH WSSTLTHH CSSALILH QSSHLIQH RTSDLIQH

ZNF8
QSSHLIRH HSSSLAQH RTTCLFLH QNSSLGRH HSTHLTVH QNSSLVQH HNAHLTVH

ZNF10
NTSNLIGY QNSPFIVH RKNDLIKH QSSALIVH RSSHLYSH QRSHLVVH QSTHLILH HSSRLIRH WFSHLVTH RSSHLIGH WRSNLTRH

ZNF18
RNSQLIFH RSSDFVKH DFSGLRHH QRSNFNRH WSSSLDKH



ZNF75D
RRSSLLRH QNSHLIKH HNTNLHTH WSSDLNKH VSSDLIKH

ZNF133
HLSRHRKT QKSHLTLH NKSALITH HQAGLIRH QKSNLVAH QQSNLIRH LKSHLNRH QNSTLISH QRTTLVNH DRSNLISH QKSAVVRH QKSNLIIH RKSTLIIH LKKSLARH KMTNLLSH

ZNF135
HSSALIEH NSSALTKH QIAPLIQH FRSSFSQH QSIHLTQH HSSSLTKH QITPLIQH QSTLLTEH HSSSLSQH QSTHLTQH QLAPLIQH QSSLLIEH

ZNF248
EKSNLTQH QKPHLTNH VKSNLTEH HRSALTVH VKSNLIVH EKSALTKH QRSVLTKH

ZNF468
QMSSLVYH QTSSLIIH RDSHLAQH RKSNLERH RKSHLERH RLSTLARH YNSYLAKH HNSSLFIH QKRYLACH CSSLLKKH

ZNF263
NNSNLIRH QNTHLTRH CNSNLHRH HSSNLLRH RSSHLVIH QGMHLTRH HRSNLIRH HSSNRIRH RSSRLMSH

ZNF282
YKESLKDH RKQNLLKH RKEHLQNH CHSGLIRH VRKSLIIH

ZNF283
SGYQLSVH WGSSLVKH RGYHLTQH CGSSLVQH SGYQLTRH RGYHLSQH CGSSLIQH WGYQLTRH RGYQLTQH WGSSLAKH SAYQLNVH HGSKLVQH

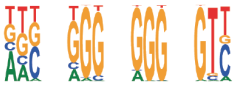
ZNF320
AEHQKIPF SPSHLIRH RKSHLERH TKAYLACH SDSRLAEH QRATLAGH RNSVLVIH QTSHLVYH QKATLACH HDSHLAKH CKSHLEIH

ZNF324
KSSDLLKH QTSHLTQH HSSSLVRH HGSNLSQH RNSHLIQH QGSSLFKH HSSNLTQH KGAVLLSH ERPALFHH

ZNF333
YSSDLIRH YSSNLRRH RNFNLILH QPSSLRSH EHSSLKTH TSTHLNVH RLSTLKSH EPSSLRKH QSSHLIVH HSSSLTVH

ZNF343
DRSTLIRH QKSNLSRH SKSILNRH EKSSFIRH DKSTLRKH QNSDLIKH DKSTLIIH RKSLLLVH QKSNLIRH DKSTLIVH HKSNLIRH

ZNF398
KQTLKDHL RKHHLMKH RKEHLLNH GHSALIRH LKISLLLH DQARLTSH HPSRLTYH SQDMLLTH



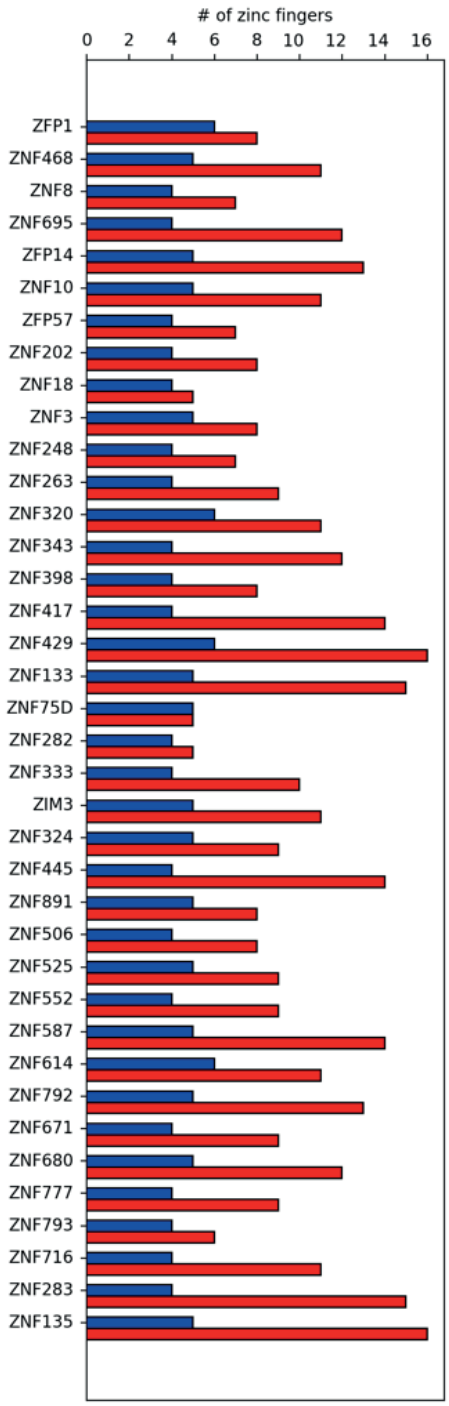
ZNF417
DVFHFADH AYLHQHQK RYVSFSNH RKSSLIQH QKGSLISH QKGNLIQH QKFCFINH QKGNLVQH YRSHLTEH RKYHLLVH NKNCVTIH SSSALHVH ECSSLIKH RSSTLLHH

ZNF429
RSSRLTQH HSSNLSSH LSSRLTQH QSSNLNSH RSSHLTSH YSSTLTRH QSSRLTRH WSSTLTKH RSSTLTSH ISSTFTKH RYSTLTTH HYSTLTNH TNHKRIYV MLSQLTQH

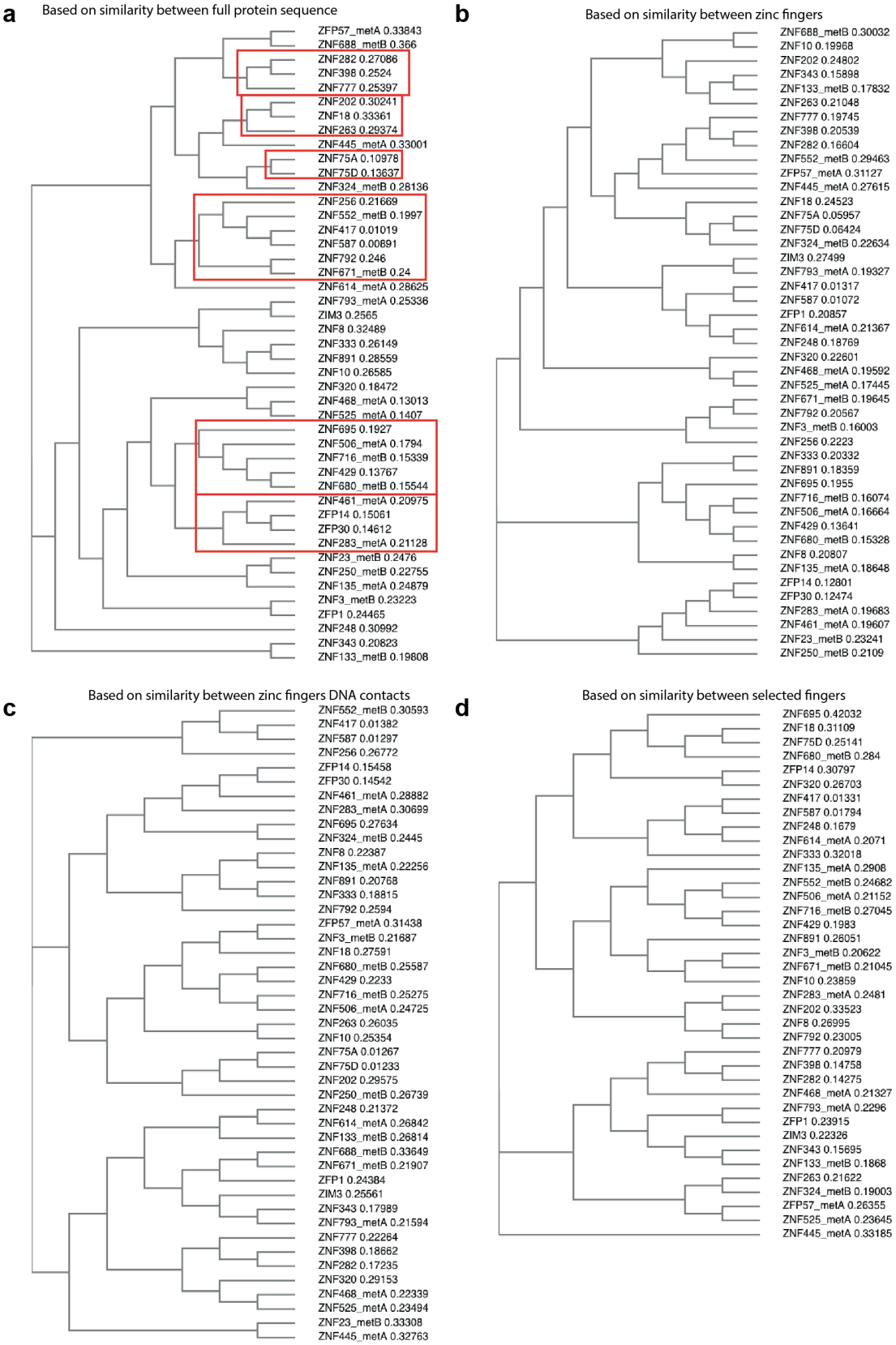
ZNF445
HSSHLAYH WSSNCARH RLSAYRLH CKSYVLEH WRSNFTRH RKKTLVDH YRSAFIVH NHSFLLIH WSSNLYRH RKRTLLDH RNYRLVNH GRHTLSSH KSSQLISH WSSNLARH

ZNF506
WSSALNKH AFSTLTEH WYSSLSKH RSSNLTKH SSSTLTKH HPATLFSH QSCNLTTH QSSHLTTH

Supplementary Figure 2.5: Predicted DNA binding zinc fingers for the factors that yielded motifs in SMiLE-seq



Supplementary Figure 2.6: Summary of difference between number of zinc fingers in each factor and number of zinc fingers that are predicted to bind.


a
Based on similarity between full protein sequence
ZFP57_metA 0.33843
ZNF688_metB 0.366
ZNF282 0.27086
ZNF398 0.2524
ZNF777 0.25397
ZNF202 0.30241
ZNF18 0.33361
ZNF263 0.29374
ZNF445_metA 0.33001
ZNF75A 0.10978
ZNF75D 0.13637
ZNF324_metB 0.28136
ZNF256 0.21669
ZNF552_metB 0.1997
ZNF417 0.01019
ZNF587 0.00691
ZNF792 0.246
ZNF671_metB 0.24
ZNF614_metA 0.28625
ZNF793_metA 0.25336
ZIM3 0.2565
ZNF8 0.32489
ZNF333 0.26149
ZNF891 0.28559
ZNF10 0.26585
ZNF320 0.18472
ZNF468_metA 0.13013
ZNF525_metA 0.1407
ZNF695 0.1927
ZNF506_metA 0.1794
ZNF716_metB 0.15339
ZNF429 0.13767
ZNF680_metB 0.15544
ZNF461_metA 0.20975
ZFP14 0.15061
ZFP30 0.14612
ZNF283_metA 0.21128
ZNF23_metB 0.2476
ZNF250_metB 0.22755
ZNF135_metA 0.24879
ZNF3_metB 0.23223
ZFP1 0.24465
ZNF248 0.30992
ZNF343 0.20823
ZNF133_metB 0.19808
b
Based on similarity between zinc fingers
ZNF688_metB 0.30032
ZNF10 0.19968
ZNF202 0.24802
ZNF343 0.15898
ZNF133_metB 0.17832
ZNF263 0.21048
ZNF777 0.19745
ZNF398 0.20539
ZNF282 0.16604
ZNF552_metB 0.29463
ZFP57_metA 0.31127
ZNF445_metA 0.27615
ZNF18 0.24523
ZNF75A 0.05957
ZNF75D 0.06424
ZNF324_metB 0.22634
ZIM3 0.27499
ZNF793_metA 0.19327
ZNF417 0.01317
ZNF587 0.01072
ZFP1 0.20857
ZNF614_metA 0.21367
ZNF248 0.18769
ZNF320 0.22601
ZNF468_metA 0.19592
ZNF525_metA 0.17445
ZNF671_metB 0.19645
ZNF792 0.20567
ZNF3_metB 0.16003
ZNF256 0.2223
ZNF333 0.20332
ZNF891 0.18359
ZNF695 0.1955
ZNF716_metB 0.16074
ZNF506_metA 0.16664
ZNF429 0.13641
ZNF680_metB 0.15328
ZNF8 0.20807
ZNF135_metA 0.18648
ZFP14 0.12801
ZFP30 0.12474
ZNF283_metA 0.19683
ZNF461_metA 0.19607
ZNF23_metB 0.23241
ZNF250_metB 0.2109
c
Based on similarity between zinc fingers DNA contacts
ZNF552_metB 0.30593
ZNF417 0.01382
ZNF587 0.01297
ZNF256 0.26772
ZFP14 0.15458
ZFP30 0.14542
ZNF461_metA 0.28882
ZNF283_metA 0.30699
ZNF695 0.27634
ZNF324_metB 0.2445
ZNF8 0.22387
ZNF135_metA 0.22256
ZNF891 0.20768
ZNF333 0.18815
ZNF792 0.2594
ZFP57_metA 0.31438
ZNF3_metB 0.21687
ZNF18 0.27591
ZNF680_metB 0.25587
ZNF429 0.2233
ZNF716_metB 0.25275
ZNF506_metA 0.24725
ZNF263 0.26035
ZNF10 0.25354
ZNF75A 0.01267
ZNF75D 0.01233
ZNF202 0.29575
ZNF250_metB 0.26739
ZNF248 0.21372
ZNF614_metA 0.26842
ZNF133_metB 0.26814
ZNF688_metB 0.33649
ZNF671_metB 0.21907
ZFP1 0.24384
ZIM3 0.25561
ZNF343 0.17989
ZNF793_metA 0.21594
ZNF777 0.22264
ZNF398 0.18662
ZNF282 0.17235
ZNF320 0.29153
ZNF468_metA 0.22339
ZNF525_metA 0.23494
ZNF23_metB 0.33308
ZNF445_metA 0.32763
d
Based on similarity between selected fingers
ZNF695 0.42032
ZNF18 0.31109
ZNF75D 0.25141
ZNF680_metB 0.284
ZFP14 0.30797
ZNF320 0.26703
ZNF417 0.01331
ZNF587 0.01794
ZNF248 0.1679
ZNF614_metA 0.2071
ZNF333 0.32018
ZNF135_metA 0.2908
ZNF552_metB 0.24682
ZNF506_metA 0.21152
ZNF716_metB 0.27045
ZNF429 0.1983
ZNF891 0.26051
ZNF3_metB 0.20622
ZNF671_metB 0.21045
ZNF10 0.23859
ZNF283_metA 0.2481
ZNF202 0.33523
ZNF8 0.26995
ZNF792 0.23005
ZNF777 0.20979
ZNF398 0.14758
ZNF282 0.14275
ZNF468_metA 0.21327
ZNF793_metA 0.2296
ZFP1 0.23915
ZIM3 0.22326
ZNF343 0.15695
ZNF133_metB 0.1868
ZNF263 0.21622
ZNF324_metB 0.19003
ZFP57_metA 0.26355
ZNF525_metA 0.23645
ZNF445_metA 0.33185

Supplementary Figure 2.7: Phylogenetic trees obtained using different similarity criteria. a) similarity based on full protein sequence – paralog groups are highlighted in red, b) similarity based on the sequence of all zinc fingers, c) similarity between the alpha helices of all zinc fingers, d) similarity between the alpha helices of the zinc fingers that are predicted to bind DNA.

Supplementary Table 1: Summary of KRAB-ZFPs tested with SMiLE-seq

Factors tested			Motifs obtained		Met-independent		MetA motifs		MetB
ZFP1	ZNF431		ZFP1		ZFP1		ZFP57		ZNF3
ZFP14	ZNF436		ZFP14		ZFP14		ZNF135		ZNF133
ZFP30	ZNF440		ZFP30		ZIM3		ZNF283		ZNF552
ZFP57	ZNF443		ZFP57		ZNF8		ZNF445		ZNF671
ZFP82	ZNF445		ZIM3		ZNF10		ZNF468		ZNF680
ZIM3	ZNF461		ZNF3		ZNF18		ZNF506		ZNF716
ZKSCAN3	ZNF468		ZNF8		ZNF75D		ZNF525		ZNF891
ZKSCAN5	ZNF485		ZNF10		ZNF202		ZNF614		
ZNF2	ZNF506		ZNF18		ZNF248		ZNF793		(no ChIP-exo motif)
ZNF3	ZNF519		ZNF23		ZNF263		ZNF324		ZFP30
ZNF8	ZNF525		ZNF75A		ZNF282				ZNF23
ZNF10	ZNF528		ZNF75D		ZNF320				ZNF250
ZNF18	ZNF540		ZNF133		ZNF333				ZNF688
ZNF23	ZNF547		ZNF135		ZNF343				
ZNF30	ZNF548		ZNF202		ZNF398				
ZNF75A	ZNF549		ZNF248		ZNF417				
ZNF75D	ZNF552		ZNF250		ZNF429				
ZNF100	ZNF557		ZNF263		ZNF587				
ZNF101	ZNF558		ZNF282		ZNF695				
ZNF114	ZNF561		ZNF283		ZNF777				
ZNF132	ZNF562		ZNF320		ZNF792				
ZNF133	ZNF564		ZNF324						
ZNF135	ZNF570		ZNF333		(no ChIP-exo motif)				
ZNF140	ZNF571		ZNF343		ZNF75A				
ZNF141	ZNF582		ZNF398						
ZNF154	ZNF585A		ZNF417						
ZNF184	ZNF587		ZNF429						
ZNF197	ZNF611		ZNF445						
ZNF202	ZNF613		ZNF468						
ZNF205	ZNF614		ZNF506						
ZNF211	ZNF616		ZNF525						
ZNF222	ZNF641		ZNF552						
ZNF235	ZNF649		ZNF587						
ZNF248	ZNF671		ZNF614						
ZNF250	ZNF680		ZNF671						
ZNF256	ZNF682		ZNF680						
ZNF263	ZNF688		ZNF688						
ZNF274	ZNF695		ZNF695						
ZNF282	ZNF713		ZNF716						
ZNF283	ZNF714		ZNF777						
ZNF320	ZNF716		ZNF792						
ZNF324	ZNF737		ZNF793						
ZNF324B	ZNF777		ZNF891						
ZNF333	ZNF780A								
ZNF337	ZNF783								
ZNF343	ZNF792								
ZNF345	ZNF793								
ZNF398	ZNF812								
ZNF417	ZNF891								
ZNF429									

Chapter 3. FloChIP

A microfluidic multiprocessor for large-scale, individual or sequential ChIP of histone marks and transcription factors

Riccardo Dainese[1,2], Vincent Gardeaux[1,2], Gerard Llimos[1,2], Antonio Carlos Alves Meireles-Filho[1,2], Daniel Alpern[1,2], Bart Deplancke[1,2,]

[1]Institute of Bioengineering, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland

[2]Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland

3.1 Introduction

The genome-wide distribution and dynamics of protein-DNA interactions constitute a fundamental aspect of gene regulation. Chromatin immunoprecipitation followed by next generation sequencing (ChIP-seq, Johnson et al. 2007) has become the most widespread technique for mapping DNA-protein interactions genome-wide. ChIP-seq has been successfully applied to dozens of TFs, histone modifications, chromatin modifying complexes, and other chromatin-associated proteins in humans and other model organisms. The ENCODE and modENCODE consortia have performed more than 8,000 ChIP-seq experiments, which have enhanced our collective understanding of how gene regulatory processes are orchestrated in humans as well as several model organisms (Landt et al., 2012). In addition, ChIP-seq proved to be essential to acquire new insights into genomic organization (Kasowski et al., 2013, Waszak et al., 2015) and into the mechanisms underlying genomic variation-driven phenotypic diversity and disease susceptibility (Deplancke et al., 2016, Lehner et al., 2013, Albert et al., 2015). More specifically, this assay proved crucial in determining the DNA binding properties of hundreds of TFs (Lambert et al., 2018). Nevertheless, in comparison to other widespread NGS-based methods – e.g. RNA-seq (Kolodziejczyk et al., 2015), ATAC-seq (Buenrostro et al., 2015), and Hi-C (Ramani et al., 2017) – ChIP-seq lags behind in some key metrics, i.e. throughput, sensitivity, modularity, and automation, which hinder its wider adoption and reproducibility. For example, while RNA-seq can now be regularly performed on hundreds or thousands of single cells using readily available workflows, ChIP-seq has largely remained labor intensive and limited to few samples per run, each composed of millions of cells. Moreover, while a typical pre-amplification RNA-seq workflow consists of only three steps – i.e. cell lysis, RNA capturing and reverse transcription – ChIP-seq typically involves several pre-amplification steps (crosslinking, lysis, fragmentation, immunoprecipitation, end-repair and adapter ligation). Finally, any given RNA transcript is present in each cell in numerous copies, which increases the likelihood of its capture and detection, whereas, on the other hand, each locus-specific protein-DNA contact occurs a maximum of two times in a diploid cell. The combination of these idiosyncratic differences, together with the lack of enabling solutions, has thus far prevented the ChIP-seq technology, as opposed to other NGS-based methods, to reach its full potential in terms of adoption and overall utility.

In addition to the standard ChIP protocol, a modification of its workflow involving sequential chromatin immunoprecipitation (sequential-ChIP) has also been adopted to infer genomic co-occurrence of two distinct protein targets. In principle, sequential-ChIP consists in performing ChIP twice on the same input chromatin, which leads to a multiplication of the inefficiencies mentioned above. Therefore, not only does sequential-ChIP show the same limitations as regular ChIP-seq, but these also come in an augmented form due to its sequential nature. As a result, despite the multi-dimensional information provided, sequential-ChIP has also resisted wider adoption.

In recent years, several attempts have been made to alleviate some of the limitations of the ChIP-seq and sequential-ChIP approaches. Gasper et al. (2014) and Aldridge et al. (2013) addressed the issue of automation by implementing the manual steps of a conventional ChIP-seq workflow on robotic liquid handling systems. However, in these examples, automation came at the expense of sensitivity, which remains in the range of tens of millions of cells per experiment. van Galen et al. (2015) and Chabbert et al. (2015) addressed the issue of throughput by barcoding and pooling chromatin samples before immunoprecipitation (IP). Although van Galen and colleagues also proved that their approach leads to higher sensitivity (500 cells per ChIP), both methods are not automated and were shown to work only for histone marks. Ma et al. (2018) and Rotem et al. (2015) addressed the limits of sensitivity with two different microfluidic-based strategies. Ma et al. focused on improving the efficiency of the IP step by confining it within microfluidic channels. Although these researchers showed good IP efficiency down to as few as 30 cells, their approach requires impractical antibody-oligo conjugates, is not automated and was not shown to work for TFs. On the other hand, Rotem et al. achieved the remarkable feat of performing ChIP-seq in a single cell by integrating the concept of chromatin barcoding and pooling into a single droplet-based microfluidic chip. However, even though the barcoding step has indeed single cell resolution, the most critical step – i.e. the IP step – is performed manually on 100 cells. As a result, their approach, which was also shown to work only for histone marks so far, yielded sparse single cell data and thousands of assays are needed to identify specific cell subpopulation signatures. In a notable effort to simplify the sequential-ChIP workflow, Weiner et al. (2016) complement the immunoprecipitation steps with sequential chromatin barcoding, thus achieving a high degree of multiplexing. However, their approach increases the number of experimental steps which makes it significantly more labor-intensive given that the workflow is not automated. In

summary, previous valuable attempts at improving the technology selectively address specific limitations but typically at the expense of or ignoring others.

In this work, we tackle all the three major limitations of current ChIP-seq and sequential-ChIP solutions (throughput, sensitivity and automation), by developing a microfluidic multiprocessor that we named FloChIP. We show that high quality one-day and parallelized ChIP-seq for histone marks (down to 500 cells) and TFs (100'000 cells) is achieved through a combination of microvalves, microstructures, flexible surface chemistry and on-chip chromatin tagmentation. Moreover, by designing an interconnected and modular device, FloChIP enables straightforward re-immunoprecipitation of eluted chromatin, effectively enabling sequential-ChIP which allows us to probe bivalent chromatin with unprecedented ease.

3.2 FloChIP

3.2.1 FloChIP is engineered for automated, bead-less and miniaturized ChIP-seq

The two core elements of FloChIP's technology are assembly of a multilayered "totem" of molecular species (Fig. 3.1a, Supp. Fig. 3.1a) and an engineered pattern of high surface-to-volume micropillars (Fig. 3.1b). The totem is based on strong although non-covalent molecular interactions and culminates with the immobilization of an antibody of choice prior to immunoprecipitation. The first layer is obtained by flowing on-chip a concentrated solution of biotinylated-BSA, which passively but thoroughly adsorbs to the hydrophobic walls of the microfluidic device. This layer has both an insulating role, i.e. it prevents non-specific adsorption of chromatin to the walls, and a docking role for the next layer, which is obtained by flowing on-chip a solution containing neutravidin that strongly binds to the biotin groups of the first layer. The third layer is formed by flowing a solution of biotinylated-protein A/G, which gets firmly immobilized by the unsaturated binding sites of the neutravidin layer. Protein A/G is a recombinant protein used in a variety of immunoassays due to its ability to strongly bind to a large number of different antibodies. This ability is retained by FloChIP's totem which thus constitutes a general substrate for antibody pull-down (Fig. 3.1a).

For the successful initiation of the antibody-capturing totem, the only substrate requirement is the hydrophobic surface of the polymer. Therefore, we set out to optimize the topology of the microfluidic channels having three main goals in mind: obtaining as much surface area as possible, miniaturizing the overall device footprint and ensuring flawless distribution of chemical species – i.e. without dead volumes where undesired chromatin could accumulate. This optimization strategy led to a design encompassing an array of micropillars of rhomboidal cross-section with the major axis aligned to the direction of the flow (Fig. 3.1b). To achieve a total estimated surface area that yields a sufficiently complex post-IP DNA library, the micropillar pattern is then repeated multiple times across each IP-lane (Fig. 3.1c).

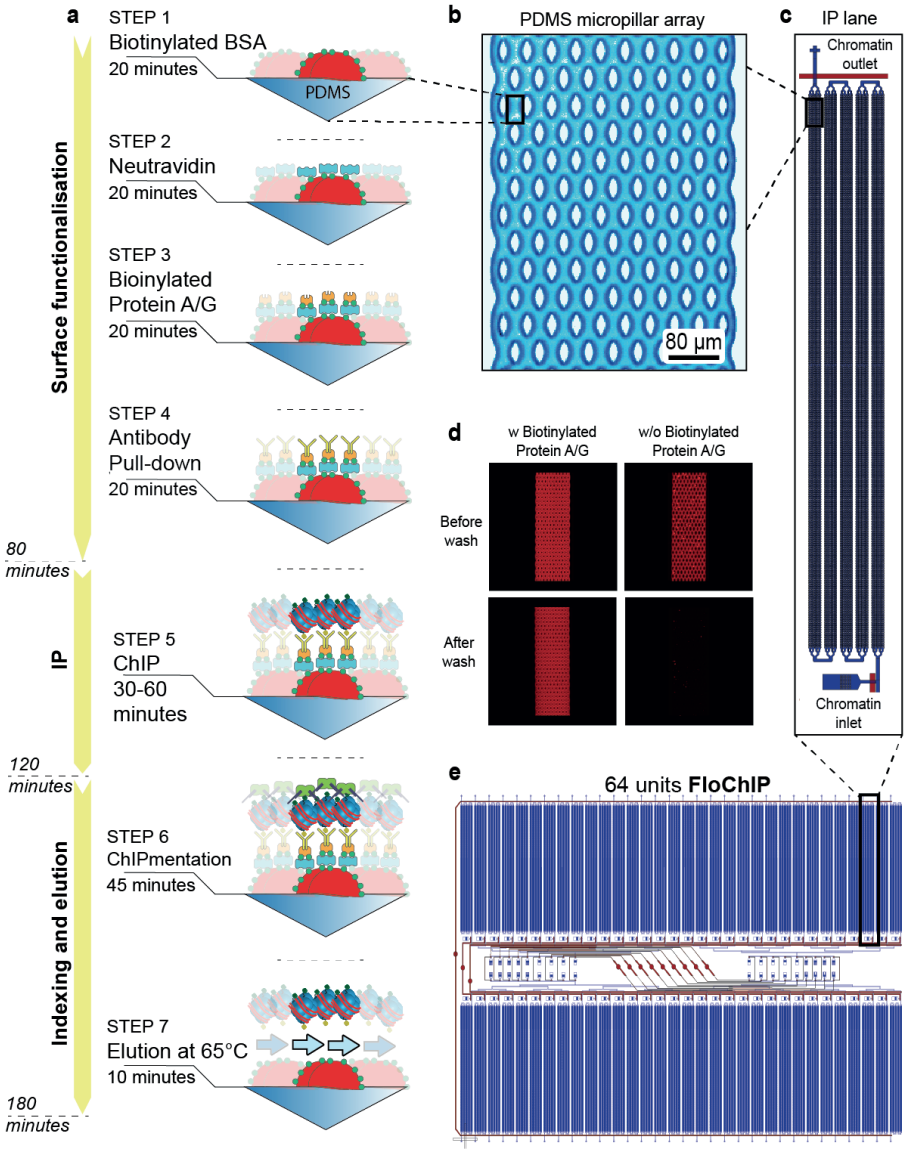


Figure 3.1. FloChIP technical innovation. a) FloChIP's processing phases in descending chronological order. b) Top-view microscopy picture of a portion containing numerous micro-pillars. c) Top-view schematic of one IP lane. d) Fluorescence micrographs showing the requirement for biotin-BSA in the correct formation of FloChIP's totem. e) Top-view schematic of the high-throughput 64-units FloChIP device, flow channels are in blue and control channels in red.

With the goal of visually confirming the outcome of the combination between totem assembly and the micropillar array and that every element is essential to this end, we first sought to IP chromatin derived from a HeLa H2B-mCherry cell line using an anti-mCherry antibody. The resulting fluorescence micrographs confirmed that each layer of the totem is necessary for successful IP of cellular chromatin (Fig. 3.1c, Supp. Fig. 3.1b).

The IP-lane is the fundamental element of the FloChIP architecture and it can itself be repeated *n* times, where *n* is the desired throughput of the device. For our low-throughput initial tests we used an 8-lanes FloChIP device (Supp. Fig. 3.1c), whereas for high-throughput experiments we utilized a 64-unit device (Fig. 3.1e). To gain accurate flow control, automation and multiplexing, a network of Quake-style microfluidic valves was added to the design. Moreover, by actuating distinct sets of valves, different multiplexing modes can be achieved with the same microfluidic architecture. For instance, we named "FloChIP mode 1" the option of multiplexing one sample into different IP units, hence equally distributing the same sample across multiple lanes, enabling multiple parallel IPs involving distinct antibodies (Fig. 3.2a). Alternatively, "FloChIP mode 2" provides the option of coating the whole device with one antibody, thus achieving sample multiplexing (Fig. 3.2b). We note that both multiplexing modes are compatible with the direct ChIP approach. However, FloChIP is also fully compatible with the indirect ChIP strategy, in which the chromatin is pre-incubated with an antibody before flowing the sample-antibody mixture on-chip. Interestingly, we noticed that for low-affinity antibodies (e.g. certain TF antibodies), the indirect ChIP was preferable to the direct one (data not shown). On the other hand, all tested histone marks antibodies showed high affinity for their epitopes and both direct and indirect ChIP yielded an equal data quality (data not shown). Regardless of the chosen approach, it is important to emphasize that, due to its microfluidic nature, FloChIP's IP step can be considerably shorter (30 to 60 minutes) than for the other macroscale alternatives.

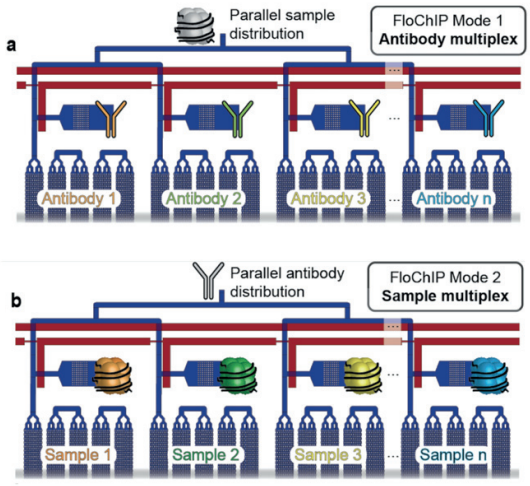


Figure 3.2. Schematics of FloChIP operational modes**.** a) Schematic depiction of FloChIP's *mode 1*: antibody multiplex. Each IP lane is functionalized separately by introducing different antibodies through the individual inlets. During IP, one sample is introduced through the common inlet and distributed equally across all IP lanes. b) Schematic depiction of FloChIP's *mode 2*: sample multiplex. One antibody solution is introduced through the common inlet and distributed equally across all IP lanes. During IP, each IP lane is loaded separately by introducing different samples through the individual inlets.

3.2.2 FloChIP reliably reproduces ENCODE data across a wide range of input cells

To benchmark the reliability of our technology and its multiplexing features as well as its overall sensitivity, we first set out to empirically estimate the overall binding capacity of each IP lane. To this end, we performed FloChIP in multiplexing mode 2, i.e. "antibody multiplex", by functionalizing the whole chip with an anti-H3K27ac antibody and immunoprecipitating different chromatin dilutions, from 1 million down to 500 cells. Since the amount of DNA that was typically recovered from the chip tended to be too small to be measured directly, we used the number of amplification cycles needed to reach a given cycle threshold (Ct) as an indirect estimator of DNA amount. Following this metric, we observed that for lower input amounts, a progressively greater number of amplification cycles is required to reach the same Ct value (Supp. Fig. 3.2a), indicating that below 100'000 cells, FloChIP functions in below-saturation conditions. We therefore estimated that FloChIP's inner surface saturates at approximately 100'000 cells. Nevertheless, we also obtained positive and stable fold enrichment results (Supp. Fig. 3.2b) and good genomic coverage (Fig. 3.3a) across the whole series of dilutions tested, suggesting that FloChIP can be carried out efficiently below and above its saturation point. To obtain a genome-wide perspective on its dynamic range, we sequenced FloChIP's derived libraries for four dilutions, i.e. 100'000, 50'000, 5'000 and 500 cells. Although the rate of uniquely mapped reads remained high for all samples (Supp. Fig. 2c), the fraction of reads falling into peaks (FRiP score) slightly decreased with lowering input amounts – from over 60% for 100'000 cells, to just above 10% for 500 cells (Supp. Fig. 2d). Nevertheless, genome-wide analysis of the obtained libraries revealed expected accumulation of reads into regions in proximity of transcription start sites (TSS, Fig. 3.3b). Moreover, genome-wide correlations demonstrate the high accuracy of our approach by showing high correlation between all library pairs (between $R^2 = 0.78$ and $R^2 = 0.98$), including Encode-FloChIP pairs among which the highest correlation was obtained for the 100'000 cells samples, i.e. $R^2 = 0.91$.

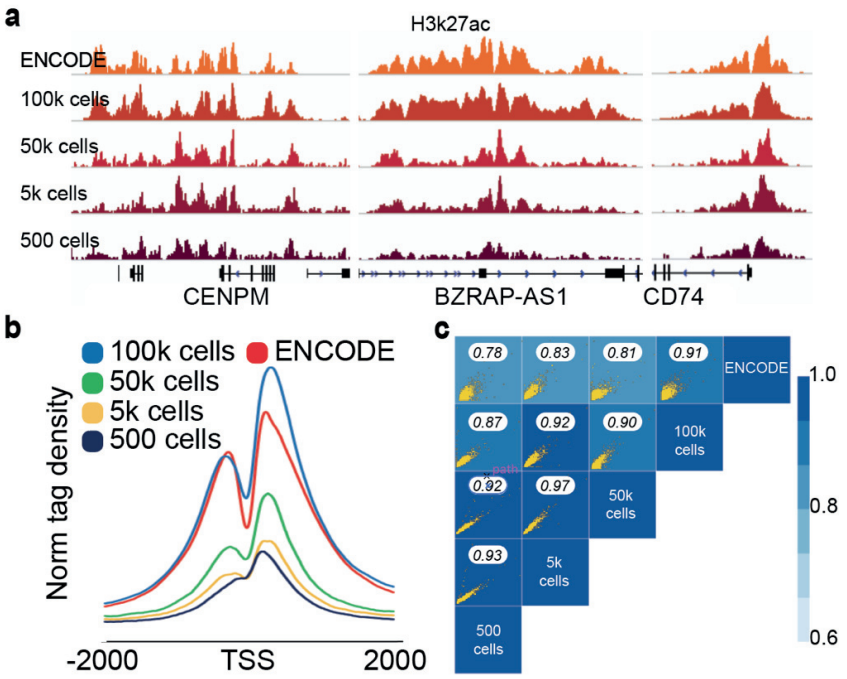


Figure 3.3. FloChIP data on cell number dilutions**.** a) H3k27ac profiles obtained by FloChIP with decreasing cell numbers. For comparison, ENCODE data generated by conventional ChIP-seq are also shown. b) Normalized read density profiles around transcription start sites for samples of decreasing cell numbers and ENCODE. c) Genome-wide correlation between pairs of samples with decreasing cell numbers and ENCODE.

After establishing 100'000 cells as the optimal trade-off point between IP-lane saturation and data quality, we set out to evaluate the reproducibility of our approach with other genomic targets. To this end, by using FloChIP's mode 1: "sample multiplex", we ChIPed in parallel 5 histone marks (H3K27ac, H3K4me3, H3K27me3, H3K4me1 and H3K9me3), going from chromatin to sequencing-ready libraries, in just one day. By visual inspection of locus specific genomic regions, we found that the obtained signal tracks closely resemble those of Encode (Fig. 3.4a). In addition, to evaluate FloChIP's performance more precisely, we determined the extent of genome-wide correlation between FloChIP and Encode datasets. Comparison of signal intensities between the respective datasets confirmed an overall high genome-wide correlation (H3K4me3: $R^2 = 0.82$, H3K27ac: $R^2 = 0.88$, H3K4me1: $R^2 = 0.91$, H3K27me3: $R^2 = 0.56$, H3K9me3: $R^2 = 0.86$; Fig. 3.4b). Moreover, comparison in terms of the FRiP score showed that, despite the ChIP input for Encode being two orders of magnitude greater than that of FloChIP, our technology consistently yields

highly enriched libraries, with FRiP scores between 1.07x and 4.12x higher for FloChIP compared to Encode (expect for H3k27me3, Fig. 3.4b). These data show that FloChIP can be used to robustly generate chromatin landscapes for histone marks with a wide input dynamic range.

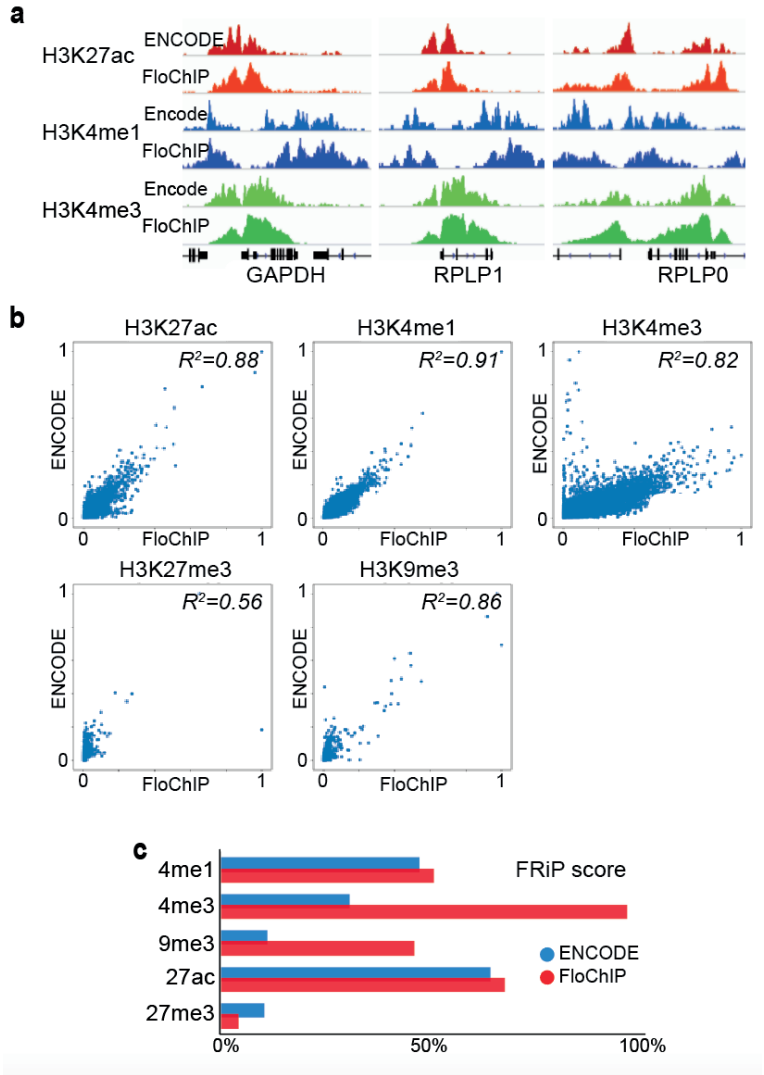


Figure 3.4. Comparison of FloChIP and ENCODE data**.** a) Signal tracks for H3k27ac, H3k4me1 and H3k4me3 profiles obtained by FloChIP are shown. For comparison, ENCODE data generated by conventional ChIP-seq are also shown. b) Genome-wide correlation plots between FloChIP (x axis) and ENCODE (y axis) data for all targets tested, i.e. H3k27ac, H3k4me1, H3k9me3, H3k27me3 and H3k4me3. c) Comparison in terms of fraction of reads in peaks (FRiP) between FloChIP and ENCODE for histone mark samples.

3.2.3 FloChIP "sequential-IP" mode provides genome-wide information on bivalent promoters

Conventional ChIP-seq provides information on the genome-wide localization of one specific protein or histone modification at a time. However, DNA regulatory elements generally harbor the interaction of several transcription factors and histone modifications in order to regulate gene expression (Deplancke et al., 2016, Spitz et al., 2012). For instance, it has been shown that promoters showing both repressive (H3k27me3) and activating (H3k4me3) marks are a characteristic feature in embryonic stem (ES) cells (Mikkelsen et al., 2007, Bernstein et al., 2006). This class of promoters have been originally named "bivalent" and are strongly associated to key developmental genes. In order to obtain direct information on the genomic location of bivalent promoters, a variant of the standard ChIP protocol called sequential-ChIP was developed. Sequential-ChIP relies on the consecutive IP of two different antigens and, as opposed to simply intersecting two ChIP-seq datasets, provides unbiased information on bivalent regions. Despite the advantage of sequential ChIP over standard ChIP in discerning true bivalency, its manual involvement and impracticality have thus far prevented widespread usage. Moreover, due to the inefficiency of the method, few studies have so far performed sequential ChIP followed by next generation sequencing (sequential-ChIP-seq), since most of them relied on qPCR to validate putative bivalent regions (sequential ChIP-qPCR). To address the technical limitations of the current sequential-ChIP workflow, we exploited FloChIP's intrinsic modularity, highly efficient IP and multiplexing features to derive the example of an automated and miniaturized sequential-ChIP solution (Fig. 3.5).

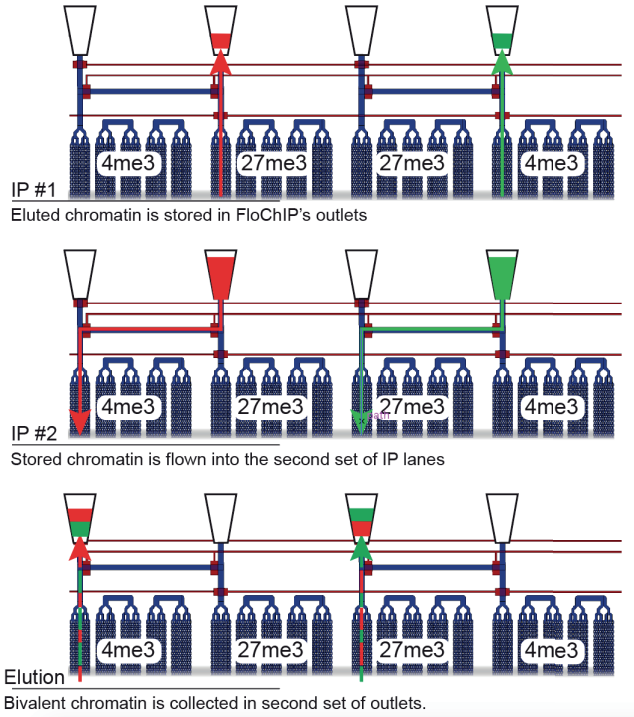


Figure 3.5. Operational schematics of FloChIP in sequential IP mode. a) FloChIP's sequential IP steps in descending chronological order for the case of H3k4me3-H3k27me3. Chromatin coming from the first IP is collected into off-chip reservoirs connected to device. Following collection, the control channels are actuated in a way to isolate the first IP lane from the chromatin, while opening the path to the second IP lane. At this point, the chromatin flown into the second pre-functionalised IP lane. Finally, the bivalent chromatin is eluted again in off-chip reservoirs.

We validated our approach by focusing on bivalent chromatin given its well-studied role in embryonic development. Specifically, we acquired genome-wide direct co-occupancy profiles for H3K27me3 and H3K4me3 in mouse embryonic stem cells (mESCs) in both IP directions – i.e. H3K27me3 first followed by H3K4me3 (H3K27me3/H3K4me3) and vice versa. As mentioned above, H3K4me3 and H3K27me3 bivalency has been originally attributed to promoters of developmental genes, leading to the hypothesis that a bivalent state maintains genes in a poised state (Mikkelsen et al 2007). In a previous study, it has been suggested that, based on the promoter read coverage comparison of ChIP-seq and sequential-ChIP-seq data, promoters show three distinct patterns of bivalency, i.e. pseudo bivalency, partial bivalency and full bivalency[21]. However, we consider these classes an artificial construct that does reflect the more fine-grained distribution of bivalency levels. Therefore, instead of assigning promoters to specific classes, we compute for each TSS a "bivalency score" (bvScore, Fig. 3.6b, details in the Methods section). To

evaluate the performance of sequential FloChIP, we focused on three distinct regions which have been previously used as proof-of-concept models by Bernstein and colleagues using ChIP-qPCR and sequential ChIP-qPCR to illustrate the methylation status difference among 4me3-only (*Tcf4* TSS), 27me3-only (upstream of *Hoxa3*) and bivalent (*Irx2* TSS) regions. FloChIP-based genomic profiles (Fig. 3.6a, Supp. Fig. 3.3a) and bvScore distributions (Fig. 3.6b) validated these previous findings as we observed that the *Tcf4* promoter shows high H3K4me3 but low H3K27me3 enrichment, and thus very low bivalency (*bvScore=0.83)*. In contrast, *Hoxa3* was mainly marked by H3K27me3, with low H3K4me3 and bivalency signals (*bvScore=0.44)*. Finally, the TSS of *Irx2* showed true bivalency (*bvScore=3.34*), with all four genomic tracks showing high coverage. In addition to considering specific loci, we also validated our data on a genome-wide scale by achieving high correlation with the results obtained by Weiner et. al using their Co-ChIP system (Supp. Fig. 3.3b). Our results are also consistent with a study by Mikkelsen and colleagues[21], who analyzed the genome-wide co-occurrence of H3K4me3 and H3K27me3 in mESCs by conventional ChIP-seq. Their findings suggested that, at the embryonic stage, most "high-CpG" (HPC) promoters, are associated with intervals of H3K4me3 enrichment, while the remaining ~22% appear to be bivalent. While our analysis confirmed that the majority (80%) of HPC promoters is marked by H3K4me3, as reflected by the green color-coded region in Fig. 3.6b, we found that the remaining 20% of promoters is bivalent (~15%, blue color-coded region) but also marked by mainly H3K27me3 (~5%, red color-coded region) especially for very low bivalency scores.

Finally, as an independent validation of our analysis, we performed gene ontology enrichment on the first one thousand promoters with the highest bivalency score. As expected, we found that these promoters are highly enriched in genes involved in a number of developmental processes, from anatomical structure development to neurogenesis (Fig. 3.6c).

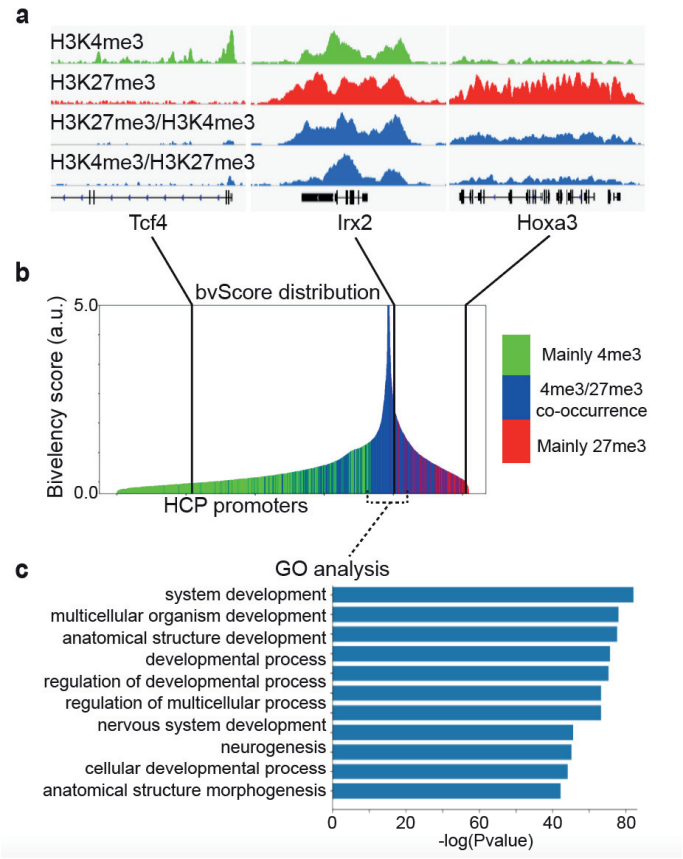


Figure 3.6. FloChIP sequential ChIP-seq results. a) Locus-specific signal tracks for the two individual IP libraries (H3k4me3 and H3k27me3) as well as the corresponding sequential IP samples (H3k27me3/H3k4me3 and H3k4me3/ H3k27me3). b) Bivalency score distribution for HCP promoters. The color-codes reflect the relative abundance of the two individual marks for each promoter. c) Gene Ontology enrichment analysis for the first one thousand promoters with the highest bivalency score.

Taken together, data indicates that FloChIP's "sequential IP" mode provides the first example of an automated, low-input (100'000 cells) and rapid (between 5-6 hours) sequential-ChIP-seq workflow for the study of bivalent promoters.

3.2.4 FloChIP is capable of ChIPing TFs in "high-throughput" mode

As mentioned in the introduction, previous attempts at improving the sensitivity and multiplexing ability of ChIP-seq experiments were shown to perform well only in the context of histone modifications. The reason for this is that performing TF ChIP-seq poses additional challenges as compared to histone marks (HM) including the fact that i) TF/DNA interactions are less abundant and less robust than HM/DNA interactions and ii) antibodies for TFs normally show lower affinity for their epitopes as compared to HM antibodies. These challenges, whose severity varies on a case by case basis, translate into the need for greater sample inputs and longer incubation times. Indeed, in our experience with FloChIP, we also experienced these challenges and for most of the TF antibodies that we tested, FloChIP's indirect method – i.e. with 2-4 hours antibody/chromatin pre-incubation in tubes – appeared to be the only way to obtain high quality results (data not shown). Nevertheless, by slowly and intermittently flowing the pre-incubated antibody/chromatin mixture on-chip, we succeeded in performing TF immunoprecipitation on only 100'000 cells, proving for the first time the feasibility of miniaturized and automated TF ChIP-seq (Fig. 3.7).

After establishing a working protocol for TF ChIP-seq, we set out to concurrently demonstrate the high-throughput capabilities of our device. To this end, by using half of the 64 IP lanes of the FloChIP device, we performed MEF2-A ChIP-seq on chromatin derived from 32 different lymphoblastoid cell lines (LCLs) (Fig. 3.7a) Before sequencing, we verified the immunoprecitation quality of each library by qPCR (Fig. 3.7b). Amplification results indicate consistently good fold enrichment across the 32 IP lanes (*log2(Fold enrichment), mean=5.7, stdev=1.5*). Another positive aspect of FloChIP is its apparent internal normalization effect on the immunoprecipitated DNA. To note, even without normalizing the input chromatin across the 32 cell lines, the amount of recovered DNA after FloChIP was extremely similar. In fact, all post-ChIP DNA libraries were amplified the same number of cycles (17 PCR cycles), adding to the convenience of our solution. We attribute this particular feature to the fact the FloChIP provides very efficient IP and reaches saturation levels even with lowly abundant samples. When saturated, the microfluidic device cannot IP any more chromatin and, given that the geometrical structure of each IP lane is the same, this leads to uniform amounts of DNA recovered even from very different samples. Subsequently, we sequenced at low coverage all 32 samples and observed variation in the percentage of uniquely mapped reads (Fig. 3.7c, *mean=42%, stdev=22%*), which in turn translated into variable genomic coverage of the libraries and variable number of peaks called (Fig.

3.7d and Fig. 3.7e). Despite this variable coverage, we were able to assess the quality of the generated libraries through different criteria. By visual inspection, both genomic profiles and genome wide TSS annotation suggested that the obtained reads accumulated as expected near transcription start sites and that the respective regions and were clearly visible in the genome browser (Fig. 3.7d and Supp. Fig. 3.4a). Moreover, in order to analyze the genome-wide agreement of the 32 datasets in an unbiased way, we considered a set of peaks obtained by merging all the 32 alignment files together and counted the number of reads mapped within each peak of each library. This allowed us to observed high pairwise correlation of the 32 libraries (Supp. Fig. 3.4b) as well as FRiP scores (Fig. 3.4f, *mean=6.9%, stdev=2.7%*) similar to the ones obtained for Encode with its own data (i.e 4.7%). Finally, we examined the enrichment of the MEF2-A motif for each set of peaks, generate individually from each sample. Despite the low number of peaks of some libraries, the expected motif is found in all libraries with a p-value lower 0.001 (Fig. 3.4f, *-log(Pvalue), mean=9.1, stdev=6.9*).

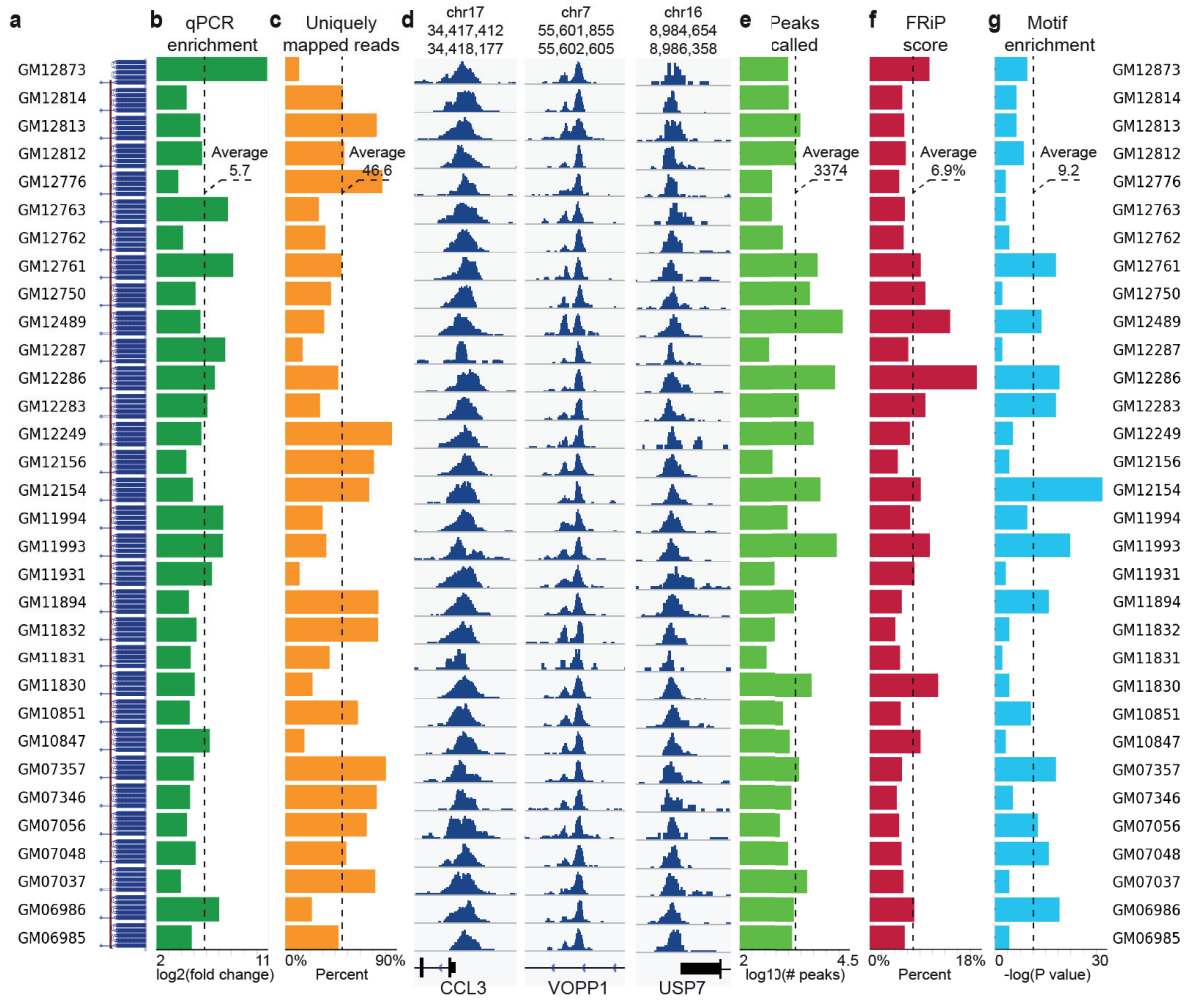


Figure 3.7. FloChIP TFs data. a) List of the 32 cell lines used in this study. b) qPCR enrichment for each library. The average across all libraries *log2(fold change)* is 5.7. c)Percent of mapped reads for each library. The average mapping rate across all libraries is 46.6%. d) Signal tracks are reported for each library for three different genomic regions. e) Number of peaks called for each library (3374 peaks on average). f) FRiP score for each library (6.9% on average). g) MEF-2A motif enrichment for each library (a *–log(Pvalue)* of 9.2 on average)

3.3 Discussion

The interaction between DNA and proteins constitutes a fundamental aspect of gene regulation. ChIP-seq allows to probe DNA-protein interactions on a genome-wide scale, thus achieving high-throughput in terms of DNA sequence space coverage. On the other hand, for what concerns throughput in terms of proteins and biological samples, ChIP-seq remains at the lowest level possible, with only one protein species and one sample tested per experiment. Aggravating this aspect, the long and manually intensive protocol prevents straightforward development towards higher throughput. Community-led efforts like ENCODE have therefore been put in place in order to perform ChIP-seq for a large number of proteins and cell types. However, despite the valuable data generated, ENCODE still sampled only a small portion of a much larger combination space. In addition to limited throughput and manual involvement, standard ChIP-seq is also restricted by the input requirements for biological material. The requirement for at least one million cells, has precluded ChIP-seq from performing reliably on smaller but possibly biologically relevant samples. Understanding the impact of these limitations, several groups attempted to improve the original protocol. However, these attempts have addressed specifically certain issues while overlooking others. In this study, we address all major ChIP-seq limitations by introducing a new technology, FloChIP, that allows for rapid, high-throughput, automated and sensitive chromatin immunoprecipitation. The two core technological aspects of FloChIP are its surface chemistry and its microfluidic architecture. The former confers FloChIP the ability to perform solid-state bead-less IP with most off-the-shelf antibodies, while the latter provides the structural substrate for miniaturized IP, rapid washing, multiplexing and straightforward automation.

Following IP, another distinctive feature of FloChIP is the direct on-chip tagmentation of captured chromatin. As also shown by Schmidl et al. (Schmidl et al., 2015), on bead-bound chromatin, direct solid-state tagmentation reduces time, cost and input requirements of ChIP experiments. To our knowledge, this is the first time solid-state tagmentation is shown to function efficiently and reproducibly on the walls on a microfluidic chip and, in general, on a substrate other than microbeads.

In order to demonstrate its reliability and applicability, we performed FloChIP for a variety of targets and samples. Initially, we aimed to empirically gain insights into FloChIP's dynamic input range. By obtaining high H3k27ac qPCR fold enrichment and high correlation with the respective ENCODE data for inputs ranging between 10^6 and 500 cells, we show that FloChIP can be used

across a wide range of inputs. Next, as a more comprehensive benchmark, we performed FloChIP for four more histone marks, namely H3k27me3, H3k4me3, H3k4me1 and H3k9me3. Despite the much lower input used for FloChIP, our results showed high correlation with ENCODE data and superior FRiP scores, thus advocating the robustness and efficiency of our approach. Next, by designing pair-wise interconnect IP lanes, we show that the chromatin eluted after the first IP step can be easily re-directed into a second IP lane, therefore achieving straightforward sequential immunoprecipitation. We validated FloChIP's sequential IP by recapitulating previously published qPCR and sequencing data on bivalent promoters in mouse embryonic stem cells. To the best of our knowledge, this is the fastest (1/2 day) and most sensitive (100'000 cells) example of sequential ChIP-seq. Moreover, this is the first automated, microfluidic and bead-less example of sequential ChIP-seq. Finally, we sought to simultaneously demonstrate FloChIP's applicability on TFs and throughput by ChIPping MEF-2A from 32 different lymphoblastoid cell lines. Overall, our data demonstrate that FloChIP is a robust, sensitive and high-throughput all-in-one ChIP-seq solution. Given its advantages and wide applicability, we believe FloChIP has great promise for establishing itself as a widely adopted tool for the study of genome-wide protein-DNA interactions.

3.4 Methods

3.4.1 Chromatin preparation

Cell fixation

GM12878 cells (5-10 millions) were harvested, washed once with PBS and resuspended in 1ml crosslinking buffer (1% PFA in PBS) for 10 minutes shaking. Crosslinking was stopped by adding 50µl of 2.5M glycine and shaking for other 5 minutes. Fixed cells were then washed twice with ice-cold PBS, pelleted, deprived of the supernatant, snap frozen and stored and -80°C.

Lysis and sonication

The frozen cell pellet was resuspended in ice-cold PBS at 4°C agitating for 30 minutes, spun at 1000g for 5 minutes, resuspended in lysis buffer (50 mM Hepes pH 7.8, 140 mM NaCl, 1mM EDTA, 0.5% NP40, 10% glycerol, 0.25% Triton and freshly added protease inhibitor), incubated with mild agitation for 10 minutes, spun for 5 minutes at 1000g, resuspended in nuclei wash buffer (20 mM Tris-HCl pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA and freshly added protease inhibitor), incubated with mild agitation for 10 minutes, spun for 5 minutes at 1000g and resuspended in sonication buffer (20 mM Tris-HCl pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.5% Na-Deoxycholate, 0.5% N-laurosylsarcosine and freshly added protease inhibitor). Nuclei were sonicated on a covaris E220 machine with the following settings: 140W intensity, 5% duty factor and 200 bursts/cycle. Chromatin was then aliquoted (~100'000 cells/aliquot) in PCR tubes and snap frozen until ChIP.

3.4.2 FloChIP

Device fabrication

Microfluidic designs were generated using Tanner L-Edit and fabricated using multilayer standard soft lithography (Thorsten et al., 2002) at the EPFL center for microtechnology. Briefly, designs were first transferred to chrome masks using a VPG200 pattern generator (Heidelberg

Instruments). Subsequently, microfluidic molds were assembled on silicon wafers with SU8 photoresist for the control layer and AZ9260 positive resist for the flow layer using a SUSS ACS200 Gen3 system (SUSS MicroTec). Microfluidics chips were fabricated by first separately casting PDMS onto the SU8 and the AZ9260 wafers with two different PDMS/curing agent ratios (20:1 and 5:1, respectively), partially curing for 30 minutes at 80°C, peeling off the PDMS from the AZ9260 wafer and aligning it to the SU8 wafer in order the reconstitute the wanted pattern. The chips were finally fully cured at 80°C for one hour and half, peeled off, holed and plasma-bonded to clean glass slides or to PDMS-coated petri dishes.

Experimental setup

Automated control of the FloChIP experimental workflow is obtained by a system of components including: 1) MATLAB software, 2) a standard laptop, 3) a WAGO fieldbus controller (ModBus 750-881), 4) FESTO 3/2 way 24V miniature solenoid valves, 5) compressed air building supply (Supp. Fig 3.1) and 6) a PCR machine. Tygon tubing and western blot tips are used to interface the microfluidic chip and the solenoid valves.
FloChIP is, in essence, a method consisting of the sequential introduction of different reagents into a custom-designed microfluidic chip. This sequence of reagents can be programmed with simple scripting commands which are, in turn, translated into sequences of solenoid valve actuations and releases. The concerted action of the solenoid valves, belonging to both the control layer and the flow layer of the chip, realizes in an automated fashion the required surface chemistry, immunoprecipitation and tagmentation reactions. On-chip temperature control is achieved by placing the microfluidic device on top of a PCR machine with flat heat-block and starting a pre-programmed temperature sequence in sync with the MATLAB script.

FloChIP operation

A FloChIP experiment starts pre-loading the control lines with distilled water and activating all valves (at a pressure of 25-30 PSI for the control lines and 2.5-5 PSI for the flow valves). Subsequently, all the reagents required for the surface chemistry (i.e. biotin-BSA, neutravidin, PBS and biotin-protein A/G, antibodies), IP (chromatin), washes (low-salt, high-salt and LiCL

buffers), tagmentation (Tn5 buffer) and elution (SDS buffer), are loaded into pipette tips and inserted into the inlets of the microfluidic device. At this stage, all valves are closed and there is no possible cross-talk between any of the reagents above. Immediately after completing the insertion of the tips, the automated protocol is launched by running the respective script. The protocol entails, in sequential order, the following steps: 20 minutes of BSA-biotin (2mg/ml), 30 seconds of PBS wash, 20 minutes of Neutravidin (1mg/ml), 30 seconds of PBS wash, 20 minutes of biotin-protein A/G (2mg/ml) and 30 seconds of PBS wash. Depending on whether direct or indirect ChIP is performed, immunoprecipitation is carried out in two different ways. (direct ChIP) or loading of the pre-incubated antibody/chromatin mix (indirect ChIP).

For direct ChIP, following biotin-protein A/G, the antibody or antibodies of choice are loaded on chip for 20 minutes. Moreover, within direct ChIP, it is possible to operate the chip in two distinct multiplexing modes, either antibody multiplex, in which microvalves are actuated in such a way that every IP unit is functionalized with a different antibody, or sample multiplex, in which all IP units are functionalized with the same antibody. The antibodies used in this study were (Abcam antibodies: anti-H3k27ac ab4729, anti-H3k4me3 ab8580, anti-H3k4me1 ab8895, anti-H3k9me3 ab8898, anti-H3k27me3 ab6147; Santa-cruz antibodies: anti-PU.1 sc-390405 and anti-MEF2a anti-MEF2A sc-17785). Following antibody loading and quick PBS wash, chromatin samples are loaded on chip by opening and closing the respective microvalves. These ON/OFF cycles, usually of 2 or 5 minutes, are performed in order to ensure that the chromatin spends enough time inside the micropillar array for the epitopes to be efficiently recognized by the corresponding antibody.

For indirect ChIP, the antibody and chromatin are incubated for 2 or 4 hours in a PCR tube prior the loading on-chip. During the IP step, the antibody/chromatin mixes are loaded into the chip in separate IP units by utilizing the same ON/OFF cycles as mentioned above.

Both for direct and indirect ChIP, the overall immunoprecipitation is performed at room temperature time spans between 30 and 60 minutes, depending on the amount of chromatin mix to be processed.

Following immunoprecipitation, rapid salt washes are performed to eliminate non-specific binding: 5 minutes of low-salt buffer (20 mM Tris pH 8.0, 150 mM NaCl, 2mM EDTA, 1% TritonX-100, 0.1% SDS), 5 minutes of high-salt buffer (20 mM Tris pH 8.0, 500 mM NaCl, 2mM EDTA, 1% TritonX-100, 0.1% SDS) and 5 minutes of LiCl buffer (20 mM Tris pH 8.0, 250 mM LiCl, 2mM EDTA, 1% NP40, 1% Na-Deoxycholate).

Following washes, Tn5 buffer (10 mM Tris pH 8.0, 5 mM $MgCl_2$) is slowly flown on-chip at 37°C for 45 minutes. This step ensures the complete tagmentation of the immunoprecipitated chromatin. Following Tn5 buffer and a 5-minutes low-salt wash to remove excess adapters, SDS buffer (10 mM Tris pH 8.0, 200 mM NaCl, 1mM EDTA, 1% SDS) is loaded on-chip at 65°C for 10 minutes in order to elute the antibody-bound chromatin from the device. The eluate is independently collected from each IP lane into PCR tubes and decrosslinked at 65°C for 4 hours. Following decrosslinking, DNA is purified in Qiagen EB buffer using Qiagen MinElute purification kits.

FloChIP operation for sequential ChIP

For sequential ChIP, instead of eluting the chromatin in SDS buffer, elution is performed by saturating the antibody with a given elution peptide (ab1342 for H3k4me3, ab1782 for H3k27me3 and ab24404 for H3k27ac, Abcam – Peptide elution buffer: 20µl of IP buffer, 2µg of an antibody-specific peptide). This way, the eluted chromatin from a given IP lane can be directly re-immunoprecipitated in the subsequent IP lane. Following elution, the chromatin is collection into a western-blot tip inserted in the specific chip outlet. Subsequently, by closing the microvalves connecting the first IP lane and the outlets while opening the ones connecting the outlet and the second IP lane, the chromatin is re-flown on-chip for the second immunoprecipitation. This second immunoprecipitation is also performed using ON/OFF cycles of 2 minutes each. The total time for the second ChIP is also between 30 and 60 minutes. Finally, after all the chromatin has been re-flown on-chip, the salt washes are repeated and elution is achieved using the standard SDS buffer.

3.4.3 ChIP-qPCR

Following FloChIP, qPCR was used to evaluate IP efficiency prior to next generation sequencing. qPCR was performed on a StepOnePlus™ (primer sequences: H3k27ac_FW CCACCCTGCACTTACGATG, H3k27ac_RV TGAGCTCCCTGTCTCTCCTC, H3k4me3_FW CGGGGGCTGCCCAAAGTTTCA, H3k4me3_RV ATTGGGGAAATTGCAGAGCGAGC, H3k27me3_FW, H3k4me1_FW, H39me3_FW GTCCGGGTCTGACTGTCTTG, H3k27me3_RV, H3k4me1_RV, H39me3_RV ACTGCACTGGGTTCACGAAG). Each qPCR reaction was composed of 10µl Applied Biosystems™ PowerUp™ SYBR™ Green Master Mix,

0.8µl of a 10µM forward primer solution, 0.8µl of a 10µM reverse primer solution, 2µl of DNA and water to a final volume of 20µl. The cycling program was the following: 2 minutes at 50°C, 2 minutes at 95°C and [5 seconds at 95°C, 20 seconds at 60°C]x60 cycles. Fold enrichment values were obtained as ratios between the percent of input of the expected positive and negative regions genomic regions.

3.4.4 NGS Library preparation

NGS library were prepared by mixing 20µl of purified DNA with 2.5µl of forward Nextera adapter, 2.5µl of reverse Nextera adapter, 32.5µl of NebNext master mix, 0.5µl of 1x SYBR green and water to 65µl. First, 5 pre-amplification cycles are run as follows: 5 minutes at 72°C, 30 seconds at 98°C and [10 seconds at 98°C, 30 seconds at 63°C, 60 seconds at 72°C]x5 cycles. Subsequently, 15µl out of the original 65µl are separated and amplified for 20 more cycles in order to estimate the optimal number of amplification cycles: 30 seconds at 98°C and [10 seconds at 98°C, 30 seconds at 63°C, 60 seconds at 72°C]x20 cycles. Finally, the remaining 50µl were amplified for N cycles, where N is the rounded up Ct value determined in the previous reaction.
DNA was size selected using AMPure XP beads in order to obtain a size distribution between 150bp and 500bp. Concentrations were measure with Qubit (ThermoFisher), size distribution was profiled with Fragment analyzer (AATI) and libraries were sequenced on an Illumina NextSeq 500.

3.4.5 FloChIP reads mapping and processing

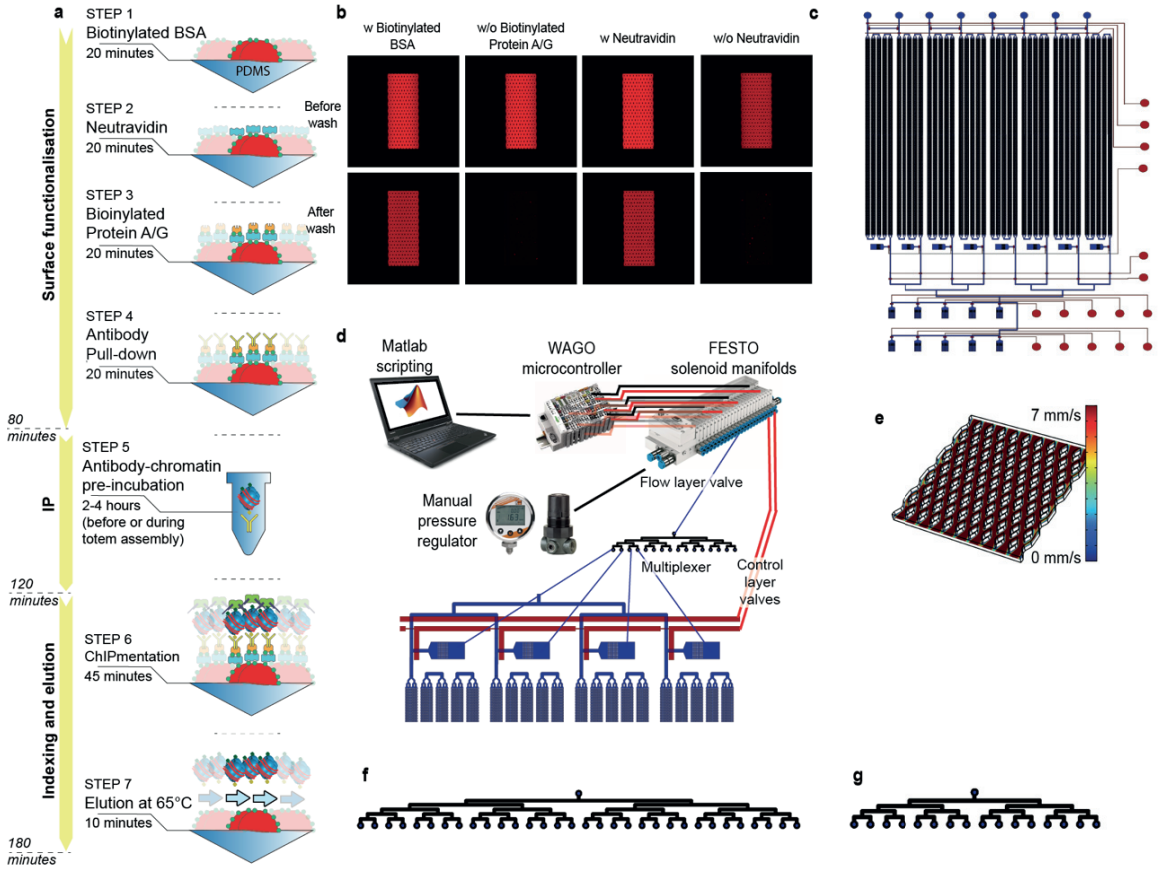
Sequencing reads were mapped to the human (hg38 and hg19) and mouse (mm10) genomes using STAR (Dobin et al., 2013) with default parameters. Uniquely mapped reads were used to call peaks using HOMER command *findPeaks.pl* with the appropriate flag, i.e. *–histone* for histone marks and *–factor* for transcription factors. FRiP scores we calculated using HOMER's command *annotatePeaks.pl* and divided the total number of reads that fall within peaks by the total number of mapped reads. Correlation plots were generated using *annotatePeaks.pl* on a common peak file, either Encode's peak files or, alternatively, the overlapping set of peaks between Encode and FloChIP datasets.

3.4.6 Bivalency score calculation

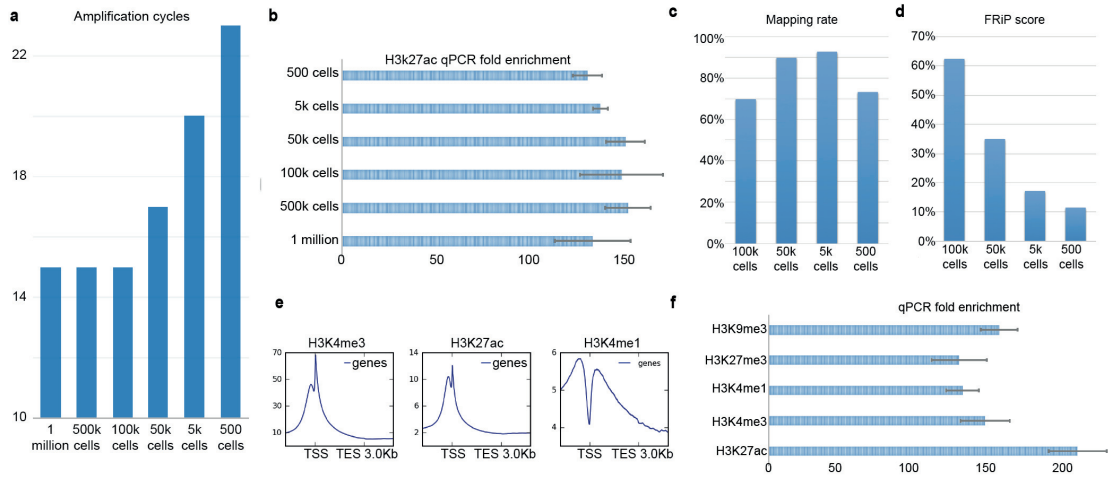
The bvScore is assigned to each promoter and is intended to take into account both the intersection between two ChIP-seq datasets as well as the agreement between the respective sequential-ChIP-seq datasets. Accordingly, the bvScore can be expressed as the product of the co-occurrence score (cScore), which measures the relative coverage of the two ChIP-seq tracks, and the agreement score (aScore), which measures the relative coverage of the two sequential-ChIP-seq tracks. We define the cScore as $(nmr_{4i}+nmr_{27i})/(nmr_{4i}-nmr_{27i})$, where nmr_{4i} and nmr_{27i} are the normalized number of mapped reads in promoter *i* for H3K4me3 and H3K27me3, respectively. The higher the value of the cScore for a promoter, the more similar is the occupancy of the two marks on that promoter. A positive cScore indicates prevalence of H3K4me3 while a negative cScore indicates prevalence of H3K27me3. We define the aScore as the absolute value of $(nmr_{4/27i}+nmr_{27/4i})/(nmr_{4/27i}-nmr_{27/4i})$, where $nmr_{4/27i}$ and $nmr_{27/4i}$ are the number of mapped reads in promoter *i* for the two sequential-ChIP-seq experiments. The higher the aScore of a promoter, the more similar is the coverage of the two sequential-ChIP-seq datasets on that promoter. Finally, the bivalency score is thus defined as *bvScore=abs(cScore*aScore)* or equivalently *bvScore=log(abs(cScore*aScore))*.

Gene ontology analysis was performed using the online tool http://geneontology.org/page/go-enrichment-analysis.

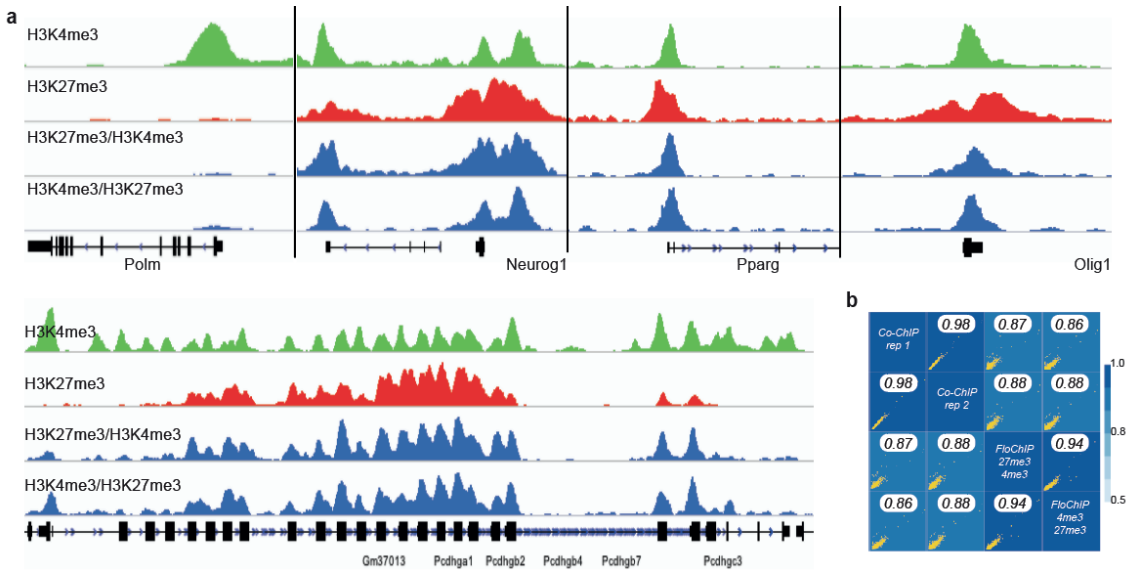
3.5 Supplementary Figures



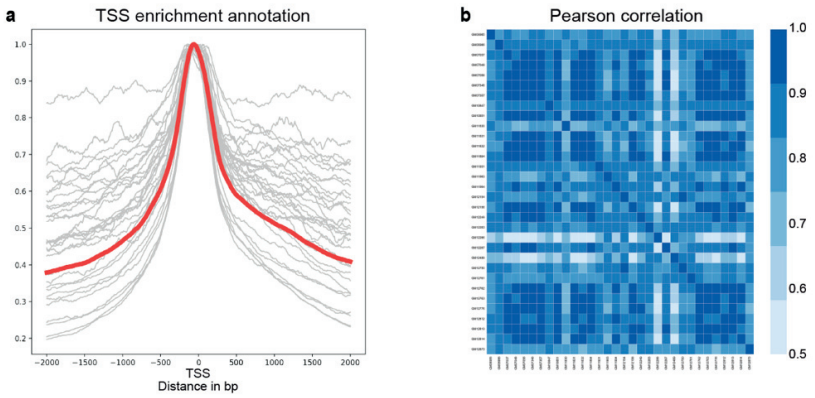
Supplementary Figure 3.1. FloChIP's setup schematics. a) FloChIP's processing phases in descending chronological order in the case of chromatin/antibody pre-incubation. b) Fluorescence micrographs showing the requirement for neutravidin and protein A/G in the correct formation of FloChIP's totem. c) Top-view schematic of the medium-throughput 8-units FloChIP device, flow channels are in blue and control channels in red. d) Schematic of FloChIP's electronic control system. e) Example of a COMSOL simulation used to optimise the device architecture. f) Microfluidic 64-outlets multiplexer for straightforward pressure distribution into FloChIP's device. g) Microfluidic 16-outlets multiplexer.



Supplementary Figure 3.2. Amplification and enrichment of FloChIP on histone marks. a) Amplification cycle statistics for samples of decreasing cell number, from 1 million to 500 cells. b). Fold enrichment statistics for samples of decreasing cell number, from 1 million to 500 cells. c) Mapping rate for samples of decreasing cell number, from 100'000 to 500 cells. d) FRiP score for samples of decreasing cell number, from 1 million to 500 cells. e) Normalized read density profiles around transcription start sites for H3k4me3, H3k27ac and H3k4me1. f) Fold enrichment statistics for histone mark samples, namely H3k4me3, H3k27ac, H3k27me3, H3k9me3 and H3k4me1.



Supplementary Figure 3.3 FloChIP's sequential IP results recapitulate previously published data. a) Signal tracks for individual and sequential IP libraries reported for the same loci originally shown in the seminal work of Mikkelsen and colleagues. b) Correlation results between FloChIP and the previously publishd Co-Chip data.



Supplementary Figure 3.4. Genome wide characterization FloChIP's of TF data. a) Normalized read density profiles around transcription start sites for all sequenced libraries (in red the average profile). b) Correlation results between all pairs of sequenced libraries

.

4 Conclusions and outlook

4.1 Research rational

Living organisms replicate, develop and function according to rules encoded in their own genome. These rules self-express in the form of gene regulatory networks, which consist of a vast amount of physical interactions between protein DNA and RNA. The sheer amount of these interactions, their diversity, their context-specificity, their microscale and transient nature constitute a major obstacle towards deciphering the underlying operational principles of the genome. The research community tackled this daunting challenge by continuously and rapidly developing new technologies to reverse-engineer the complexity of gene regulatory networks.

The first wave of technological innovation addressed the task of efficiently and cost-effectively identifying the "nodes" of gene regulatory networks, i.e. listing the functional elements of the network such as proteins, DNA regulatory regions and RNA species. This approach culminated in a plethora of new tools, among which on of the most notable example is perhaps next generation sequencing (NGS). Introduced commercially in the mid-2000s, NGS allowed researchers to sequence DNA faster, cheaper and at a much larger scale than ever before. As an historical perspective, it is noteworthy to compare the duration and cost of the Human Genome Project (1990–2003, $3 billion) with today's NGS-enabled standards (one day, <$5'000) to sequence a single human genome.

With the establishment of next generation sequencing as the technological gold standard for DNA sequencing, the focus of technical development shifted towards finding efficient ways to characterize the interaction between DNA and the other constituents of gene regulatory networks, i.e. RNA and proteins. Several of the resulting technologies leveraged on the throughput offered by NGS and are for this reason referred to as "NGS-based". These technologies focused on the isolation of molecular complexes containing genomic (or random) DNA, while relying on NGS for downstream sequencing, in order to obtain so-called genome-wide profiles.

As opposed to the first wave of technical developments, whose main achievements are now consolidated, privatized and commercially successful, the second stage of innovation is still in its infancy. For instance, the most adopted "first-wave" NGS solution is embodied by the Illumina sequencing instruments. They come in the form of workstations onto which sequencing cartridges are loaded and run. They are standardized, microfluidic-based, user-friendly and extremely high-throughput. One the other hand, the most adopted solutions belonging to the new wave of advances come for the most part in the form by publicly-shared laboratory protocols. For instance, even if ChIP-seq has been instrumental for the community – from classifying functional of DNA elements (ENCODE Project Consortium, 2012) to elucidating the impact of genetic variation on gene regulation (Deplancke et al., 2016) – it remains a poorly standardized, manually intensive and low-throughput approach.

In this thesis work, I report development and optimization of two microfluidic tools, SMiLE-seq and FloChIP, which complements standard macroscale techniques for the characterization of protein/DNA interactions in vitro and in vivo, respectively. In both instances, research efforts were tailored towards offering superior alternatives to existing methods according distinct although equally important metrics: user-friendliness, automation, throughput and sensitivity. In other words, I aimed to develop technologies that follow a similar "innovation trajectory" as the one that led to next generation sequencing.

I demonstrate the functionalities of the SMiLE-seq and FloChIP by first benchmarking them against publicly available data and secondly by providing examples of how their engineered microfluidic features render them superior alternatives to existing solutions.

4.2 SMiLE-seq summary

The initial motivation was that a large portion of the DNA-binding specificity of individual transcription factors was yet uncharacterized. Although established methods like protein binding microarrays (Berger et al., 2006) and HT-SELEX (Jolma et al., 2010) greatly expanded the list of known *in vitro* human TF-DNA specificity (Jolma et al., 2013), their difficulty in completing the

list left room for alternative or complementary solutions. We reasoned that the resistance of some factors towards *in vitro* characterization had to do with their inherent perhaps weak and/or transient DNA-binding modes. We therefore considered devising an alternative based on the MITOMI principle (Maerkl and Quake, 2007) which has been shown to be able to mechanically trap TF/DNA interactions with great sensitivity.

By capitalizing on our knowledge of microfluidics and molecular biology, we combined features of both MITOMI and HT-SELEX in order to design a novel NGS-based experimental framework capable of characterizing the binding specificities of TFs over a wide affinity range and with minimum reagent/sample consumption. We first proved the feasibility of our approach be recapitulating the binding specificities of 58 previously characterized factors coming from different species (6 *Drosophila*, 12 mouse and 40 human) and TF families. Next, we showed that SMiLE-seq data holds accurate information on the DNA binding energy landscapes indicating that microfluidic-based washing of unspecifically-bound DNA leads to superior ligands enrichment. Finally, we proceeded to address the challenge of obtaining DNA-binding motifs for those factors that so far resisted characterization by both HT-SELEX and PBMs.

The majority of these uncharacterized factors belong to the family of krüppel-associated box zinc-finger proteins (KRAB-KZFPs), a large family of ~350 transcription factors. Although, the exact function of several KRAB-ZFPs is poorly understood, a great portion of this TF family is implicated in repressing transposable retroelements during embryonic development. Moreover, it is known that the DNA-binding of certain KRAB-zinc fingers is methylation-specific, e.g. ZFP57 is known to selectively bind the methylated hexanucleotide TGCCGC. Therefore, in order to simultaneously increase the likelihood of deriving motifs and add another dimension to our study, we considered introducing in our assays methylated DNA libraries as well.

Given the theoretically large number of assays planned (700, i.e. 350 with normal and methylated DNA), we first decided to test the efficiency of SMiLE-seq on a small sample test and later opted for the re-design of the original workflow towards a version with higher throughput and sensitivity. Briefly, we achieved higher through modifying the architecture of the previous SMiLE-seq device and the sequence of the input DNA libraries. As opposed to 8 factors per experiment and one

library per sequencing run, the new system could host 32 parallel assays per chip while allowing a theoretical multiplexing ability of hundreds of libraries per sequencing run. To put this in perspective, considering that the average waiting time for sequencing SMiLE-seq libraries is ~1 week, the original SMiLE-seq would have processed the 700 KRAB-ZFPs assays in more than 1.5 years. On the other hand, the new SMiLE-seq version (SMiLE-seq v2.0) eliminates the sequencing waiting time as an experimental bottleneck. Instead, the assay delay becomes solely dependent on the throughput of microfluidic device. Considering an experiment per day, the new 32-factor SMiLE-seq device can potentially complete the 700 assays in 22 days.

With SMiLE-seq 2.0 we performed 202 assay – i.e. 101 factors for both methylated and non-methylated DNA – and retrieved motifs for 43 KRAB-ZFPs. These motifs closely match the motifs obtained by analysing recently published ChIP-exo data of the same factors. By integrating ChIP-exo and SMiLE-seq with *in silico* predictions we demonstrate that the major limitation of existing tools is the ability to discern which zinc fingers contribute to DNA binding in a given zinc finger array. Moreover, for all tested KRAB-ZFPs, we propose the subset of zinc fingers that actually determines the DNA binding specificity of the factor. Further experimental work is needed to validate our results.

4.3 FloChIP summary

For FloChIP, the scope of the project was to explore ways to address the major ChIP-seq limitations with one comprehensive solution. After years of development, FloChIP now provides rapid, high-throughput, automated and sensitive chromatin immunoprecipitation.
This was achieved by developing in parallel two aspects that together confer FloChIP its functionalities: its surface chemistry and its microfluidic architecture. The former enables solid-state bead-less IP with any off-the-shelf antibodies, while the latter provides the structural substrate for miniaturized IP, rapid washing, multiplexing and straightforward automation.

Beyond IP, another important feature of FloChIP is the direct on-chip tagmentation of captured chromatin which reduces time, cost and input requirements of ChIP experiments.

In order to demonstrate its reliability and applicability, we performed FloChIP for a number of targets and samples. Initially, we obtained high H3k27ac qPCR fold enrichment and high correlation with the respective ENCODE data for inputs ranging between 10^6 and 500 cells, thus demonstrating that FloChIP can be used across a wide range of inputs.

Next, we performed FloChIP for four more histone marks – i.e. H3k27me3, H3k4me3, H3k4me1 and H3k9me3. Despite the much lower input used for FloChIP, we demonstrate high correlation with ENCODE data and high FRiP scores, which proves the robustness and efficiency of the technology.

Subsequently, by connecting the IP lanes in a pair-wise fashion, we show the feasibility of sequential ChIP-seq on-chip. We validated FloChIP's sequential IP by recapitulating previously published qPCR and sequencing data on bivalent promoters in mouse embryonic stem cells. Our positive results demonstrate the feasibility of a very fast (6 hours) and sensitive (100'000 cells) sequential ChIP-seq. Finally, we sought to simultaneously demonstrate FloChIP's applicability on TFs and throughput by ChIPping MEF-2A from 32 different lymphoblastoid cell lines.

Overall, we expect that both SMiLE-seq and FloChIP will contribute to the characterization and understanding of protein/DNA interactions in the context of gene regulatory networks.

4.4 Outlook

During a long and intense explorative journey such as four years of hands-on and heads-down research, it is easy to get lost in the details of a project and lose track of the so-called "big picture". When it's finally the time to stop experiments, it becomes an interesting mental exercise to ask: what's next? What follows is my personal view on different aspects of the near and not so-near future of biological research.

4.4.1 SMiLE-seq and the specificity of transcription factors

In the immediate future, it is straightforward to imagine that the SMiLE-seq, together with HT-SELEX and PBM will be employed to obtain the DNA-binding specificities of the remaining ~500

human transcription factors. Depending on the research interest, dedicated funding and adoption of the respective technologies, it is likely that the TF/DNA specificity cataloguing will be extended to other species as well, e.g. mouse, *Drosophila*, *C. elegans*, the zebrafish and possibly other exotic species.

Once the DNA-binding specificities of most individual TFs are known, the climb uphill will only have started. I believe the main following challenges will be:

- understand how the formation of TF/protein complexes, i.e. co-factor binding, dimerization, trimerization etc. – will induces changes in the binding specific of individual TFs. A number of groups already reported indications which support the 'variable specificity' of TF-complexes (Slattery et al., 2011, Jolma et al., 2015) but the systematic examination of all possible dimers has not been conducted yet. Moreover, higher order oligomers haven't been tested yet because the shear number of possible combinations surpasses by far the capabilities of existing techniques.
- characterize the affinity landscapes for all TFs. Specificity is only one side of the coin and the quantification of the affinity of each TF to its target DNA sequences (and variants therein) will be equally important to characterize. This thorough biophysical characterization will be extremely important towards the accurate mathematical modeling of regulatory networks, their rewiring and their *de novo* engineering. In order to achieve this, one could envision large-scale systematic MITOMI-like assays. Unfortunately, the current state of MITOMI and other techniques does not allow yet to perform such large-scale studies in reasonable times/costs.
- design structural and computational tools the accurately model TF/DNA interactions. As data keeps being accumulated, a foreseeable consequence is the development of ever more accurate models that generalize well the specificity and affinity of TFs to DNA. These models will again be instrumental in the process of digitizing biology.

4.4.2 FloChIP and experimental biology

As mentioned before, FloChIP constitutes an attempt to integrate on a single microfluidic chip several steps of the long and tedious ChIP-seq technique. This miniaturization came with intrinsic advantages such as ease of automation, sensitivity and speed. I believe these qualities will translate

into wide adoption of FloChIP and, with it, wider adoption of ChIP-seq in general. In the post-genomic era, this will be crucial towards understanding the mechanisms of how genetic variants and environmental factors affect chromatin state and gene expression. Nevertheless, FloChIP's development is only just beginning. The ultimate goal is to minimize manual operation and maximize reproducibility. For example, FloChIP still requires chromatin that has been prepared manually, this introducing variability inter-experiment and inter-operator variability. FloChIP therefore still needs to be developed further in order to become a complete solution. Nevertheless, my experience with FloChIP so far proved me that it is possible to take an established multi-step experimental protocol, miniaturize it and thus obtaining a series of advantages over the current solution. In light of this I wonder: how many existing laboratory protocols exist that suffer from the same drawback/limitations as ChIP-seq that have not been innovated yet? From relatively new methods like RNA-seq to old-school protocols such DNA purification and PCR. Surprisingly, several examples can be found in the literature of microfluidic implementations of the above mentioned techniques. However, very few of them has been widely adopted. As a result, most common lab practices remain manual, slow and low-throughput. The biggest challenge for the future is then to understand the underlying reasons of the resistance of experimental biology to widespread automation and miniaturization, address them and move small but steady steps towards complete miniaturization and automation of experiments. Before digitizing biology, we need to digitizing experimental biology.

5. References

1. Albert, F.W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. Nat. Rev. Genet. 16, 197–212 (2015).
2. Aldridge, S. et al. AHT-ChIP-seq: a completely automated robotic protocol for high-throughput chromatin immunoprecipitation. Genome Biol., 14, p. R124 (2013)
3. Bailey, T.L. & Elkan, C. In Proc. Int. Conf. Intell. Syst. Mol. Biol. (Eds. Altman, R. et al.) 28–36 (AAAI Press, 1994).
4. Berger, M. F. et al. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nature Biotechnol. 24, 1429–1435 (2006).
5. Berger, M. F. et al. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell 133, 1266–1276 (2008).
6. Bernstein, B. E. et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125, 315–326 (2006).
7. Bianconi, E. et al. An estimation of the number of cells in the human body. Ann. Hum. Biol. 40, 463–471 (2013).
8. Buenrostro, J.D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 523, 486–490 (2015).
9. Cao, Z., Chen, C., He, B., Tan, K. & Lu, C. A microfluidic device for epigenomic profiling using 100 cells. Nat. Methods http://dx.doi.org/10.1038/nmeth.3488 (2015).
10. Chabbert, C. D. et al. A high-throughput ChIP-Seq for large-scale chromatin studies. Mol. Syst. Biol. 11, 777 (2015).
11. Crick, F. Central dogma of molecular biology. Nature 227, 561–563 (1970).
12. Deplancke, B., Alpern, D. & Gardeux, V. The genetics of transcription factor DNA binding variation. Cell 166, 538–554 (2016).
13. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013).
14. G. Ecco, M. Imbeault, D. Trono KRAB zinc finger proteins Development, 144 (2017), pp. 2719-2729.

15. Gasper, W. C. et al. Fully automated high-throughput chromatin immunoprecipitation for ChIP-seq: Identifying ChIP-quality p300 monoclonal antibodies. Sci. Rep. 4, 5152 (2014).
16. Geertz, M., Shore, D. & Maerkl, S.J. Massively parallel measurements of molecular interaction kinetics on a microfluidic platform. Proc. Natl. Acad. Sci. USA 109, 16540–16545 (2012).
17. Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L. & Noble, W.S. Quantifying similarity between motifs. Genome Biol. 8, R24 (2007).
18. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 38, 576–589 (2010).
19. Isakova, A., Berset, Y., Hatzimanikatis, V. & Deplancke, B. Quantification of cooperativity in heterodimer-DNA binding improves the accuracy of binding specificity models. J. Biol. Chem. 291, 10293–10306 (2016).
20. Isakova, A. et al. SMiLE-seq identifies binding motifs of single and dimeric transcription factors. Nat. Methods 14, 316–322 (2017).
21. Johnson, D., Mortazavi, A., Myers, R. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. Science 316, 1497–1502 (2007).
22. Jolma, A. et al. DNA-binding specificities of human transcription factors. Cell 152, 327–339 (2013).
23. Jolma, A. et al. DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature 527, 384–388 (2015).
24. Jolma, A. et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res. 20, 861–873 (2010).
25. Kasowski, M. et al. Extensive variation in chromatin states across humans. Science 342, 750–752 (2013).
26. Kinkley J. et al. reChIP-seq reveals widespread bivalency of H3K4me3 and H3K27me3 in $CD4^+$ memory T cells. Nat. Commun., 7, 12514 (2016)
27. Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C. & Teichmann, S.A. The technology and biology of single-cell RNA sequencing. Mol. Cell 58, 610–620 (2015).

28. Kulakovskiy, I.V. et al. HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. Nucleic Acids Res. 44 D1, D116–D125 (2016).
29. Lambert, S.A. et al. The human transcription factors. Cell 172, 650–665 (2018).
30. Lander, E. S. et al. Initial sequencing and analysis of the human genome. Nature 409, 860–921 (2001).
31. Landt, S. G. et al. ChIP-seq guidelines and practices used by the ENCODE and modENCODE consortia. Genome Res. (2012).
32. Lecault, V. et al. High-throughput analysis of single hematopoietic stem cell proliferation in microfluidic cell culture arrays. Nat. Methods 8, 581–586 (2011).
33. Lehner, B. Genotype to phenotype: lessons from model organisms for human genetics. Nature Rev. Genet. 14, 168–178 (2013).
34. Ma, S. et al. Low-input and multiplexed microfluidic assay reveals epigenomic variation across cerebellum and prefrontal cortex. Science Advances. 4, 4 (2018).
35. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161, 1202–1214 (2015).
36. Maerkl, S. J. & Quake, S. R. A systems approach to measuring the binding energy landscapes of transcription factors. Science 315, 233–237 (2007).
37. Mathelier, A. et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. 42, D142–D147 (2014).
38. Matys, V. et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 34, D108–D110 (2006).
39. Mikkelsen, T. S. et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448, 553–560 (2007).
40. Mukherjee, S. et al. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nature Genet. 36, 1331–1339 (2004).
41. Najafabadi, H.S. et al. C2H2 zinc finger proteins greatly expand the human regulatory lexicon. Nat. Biotechnol. 33, 555–562 (2015).
42. Nitta, N. et al. Intelligent Image-Activated Cell Sorting. Cell 175, 266-276 (2018)
43. Patel et al. DNA Conformation Induces Adaptable Binding by Tandem Zinc Finger Proteins. Cell. 173(1):221-233 (2018).

44. Patel, A., et al. DNA Conformation Induces Adaptable Binding by Tandem Zinc Finger Proteins. Cell 2018, 173, 221– 233.
45. Persikov, A.V. & Singh, M. De novo prediction of DNA-binding specificities for Cys2His2 zinc finger proteins. Nucleic Acids Res. 42, 97–108 (2014).
46. Quenneville, S. et al. (2011). In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. Mol. Cell 44, 361-372
47. Ramani, V. et al. Massively multiplex single-cell Hi-C. Nat. Methods 14, 263–266 (2017)
48. Rotem, A. et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. Nat. Biotechnol. 33, 1165–1172 (2015).
49. Schmidl, C., Rendeiro, A. F., Sheffield, N. C. & Bock, C. ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. Nat. Methods http://dx.doi.org/10.1038/nmeth.3542 (2015).
50. Slattery, M. et al. Cofactor binding evokes latent differences in DNA binding specificity between hox proteins. Cell 147, 1270–1282 (2011).
51. Spitz, F. & Furlong, E. E. Transcription factors: from enhancer binding to developmental control. Nature Rev. Genet. 13, 613–626 (2012).
52. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012)
53. Thorsen, T., Maerkl, S. J. & Quake, S. R. Microfluidic large-scale integration. Science 298, 580–584 (2002).
54. Unger, M. A., Chou, H.-P., Thorsen, T., Scherer, A. & Quake, S. R. Monolithic microfabricated valves and pumps by multilayer soft lithography. Science 288, 113–116 (2000).
55. Urrutia, R. KRAB-containing zinc-finger repressor proteins. Genome Biol. 4, 231 (2003).
56. van Galen, P. et al. A multiplexed system for quantitative comparisons of chromatin landscapes. Mol. Cell 61, 170–180 (2015).
57. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. Nature Rev. Genet. 10, 252–263 (2009)

58. Warkiani, M. E., Wu, L., Tay, A. K. P. & Han, J. Large-Volume Microfluidic Cell Sorting for Biomedical Applications. Annu. Rev. Biomed. Eng. 17, 1–34 (2015).
59. Waszak, S. M. et al. Population variation and genetic control of modular chromatin architecture in humans. Cell 162, 1039–1050 (2015).
60. Weiner, A. et al. Co-ChIP enables genome-wide mapping of histone mark co-occurrence at single-molecule resolution. Nat. Biotechnol. 34, 953–961 (2016).
61. Yin, Y. et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science 356, eaaj2239 (2017).

Riccardo Dainese

EPFL PhD Candidate in Bioengineering

I am passionate about automation and miniaturisation for biology and medicine.

Email: riccardo.dainese@epfl.ch

EXPERIENCE

Ecole polytechnique fédérale de Lausanne

Oct 2014 to Present

PHD Student in Bioengineering

iGEM assistant for EPFL teams 2015 and 2016 Courses taken and passed: - Theoretical Microfluidics - Innosuisse Business Concept - Machine learning (Coursera) - Deep learning (Udacity) - Electrochemical nano-bio-sensing and bio/CMOS interfaces

iGEM Competition

Jun 2014 to Sep 2014

Captain of team Gothenburg 2014

Best Foundational Advance Project In this exciting project, my team and I set out to build a synthetic gene circuit that we called: “Yeast age counter”. The ultimate goal was to make yeast cells express a different fluorescent protein depending on how many generations the cell has already undergone. This age counter would allow quick and easy fluorescence based separation of yeast cells coming from an age-wise heterogeneous population.

Biophotonics research group at Chalmers University

Jun 2014 to Aug 2014

Project Intern

In this project I worked with the Biophotonics research group at Chalmers University at the development of two gradient generators, one for a linear gradient and the second for a 2-fold logarithmic gradient. The goal of this project was to devise an efficient tool for studies of mammalian cell response and chemotaxis along gradients of chemoattractants or repellants.

EDUCATION

Udacity

2016 to 2017

Nanodegree in Machine Learning

In this Nanodegree, I delved deeper into the techniques and algorithms of machine learning with particular focus on Deep Learning. In my capstone project, I programmed an agent based on neural networks and policy gradients in order to solve the Cartpole's Open AI gym environment by using only raw pixels as input.

Chalmers University of Technology

2012 to 2014

Master in Nanotechnology, 5/5

At Chalmers I pursued an education focused on bionanotechnology, balancing my focus between theoretical and practical work. The courses I followed with respective grades are: Fundamentals of nanoscience, 5/5 Fundamentals of micro- and nanotechnology, 5/5 Nanomaterials chemistry, 5/5 Modeling and fabrication of micro/nanodevices, 5/5 Biophysical chemistry, 5/5 Applied optical spectroscopy, 5/5 Bionanotechnology, 5/5 Nanobioscience for information processing, 5/5. My master thesis represents a simple attempt to use mathematical modelling in order to study aging on asymmetrically dividing

cell populations. I use systems of differential equations, parallelised computational simulations and HTML representations in order to examine the population advantages of different single cell damage segregation strategies upon division.

Politecnico di Torino
2009 to 2011

Bachelor's degree in Biomedical engineering, 110/110

SKILLS

Microfabrication, Biotechnology, Microfluidics, Machine Learning, Mathematical Modeling, Molecular Biology, Python, Matlab, Mathematica, Arduino, HTML, PHP, JavaScript, Biophysics, Chromatin Immunoprecipitation, Multilayer microfluidics, Droplet microfluidics, Deep Learning, TensorFlow

PUBLICATIONS

SMiLE-seq identifies binding motifs of single and dimeric transcription factors
Jan 16, 2017

Alina Isakova, Romain Groux, Michael Imbeault, Pernille Rainer, Daniel Alpern, Riccardo Dainese, Giovanna Ambrosini, Didier Trono, Philipp Bucher & Bart Deplancke

Nature methods

Resolving the DNA-binding specificities of transcription factors (TFs) is of critical value for understanding gene regulation. Here, we present a novel, semiautomated protein–DNA interaction characterization technology, selective microfluidics-based ligand enrichment followed by sequencing (SMiLE-seq). SMiLE-seq is neither limited by DNA bait length nor biased toward strong affinity binders; it probes the DNA-binding properties of TFs over a wide affinity range in a fast and cost-effective fashion. We validated SMiLE-seq by analyzing 58 full-length human, mouse, and Drosophila TFs from distinct structural classes. All tested TFs yielded DNA-binding models with predictive power comparable to or greater than that of other in vitro assays. De novo motif discovery on all JUN–FOS heterodimers and several nuclear receptor-TF complexes provided novel insights into partner-specific heterodimer DNA-binding preferences. We also successfully analyzed the DNA-binding properties of uncharacterized human C2H2 zinc-finger proteins and validated several using ChIP-exo.

Systems Biology of Aging
Mar 18, 2017

Wiley *Johannes Borgqvist, Riccardo Dainese, Marija Cvijovic*

Aging is a highly complex, irreversible process, characterized by the accumulation of damage coupled with progressive functional decline, inevitably culminating in death. Aging can be viewed as a set of complex systems with emergent properties, which cannot be understood by simply analyzing the individual components. Therefore, a multidisciplinary approach where experimental work is integrated with mathematical modeling constitutes a powerful method in order to elucidate the multiple biological mechanisms encountered in aging. In this chapter, such a systems biology approach is exemplified with a mathematical description of the evolutionarily conserved damage accumulation theory.

LANGUAGES

English (Full professional proficiency), Italian (Native or bilingual proficiency), French (Limited working proficiency), Spanish (Limited working proficiency), Portuguese (Limited working proficiency)

**Ce document a été imprimé au Centre d'impression EPFL,
imprimerie climatiquement neutre, certifiée myClimate.**