High-quality, high-throughput measurement of protein-DNA binding using HiTS-FLIP

Vincent Roman Wolowski

München 2016

Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig-Maximilians-Universität München

High-quality, high-throughput measurement of protein-DNA binding using HiTS-FLIP

Vincent Roman Wolowski

aus

Rosenheim, Deutschland

2016

Erklärung:

Diese Dissertation wurde im Sinne von §7 der Promotionsordnung vom 28. November 2011 von Herrn Dr. Johannes Söding betreut.

Eidesstattliche Versicherung:

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, am 2. März 2016

Vincent Wolowski

Dissertation eingereicht am: 04.04.2016

1. Gutachter: Dr. Johannes Söding
2. Gutachter: PD Dr. Dietmar Martin

Mündliche Prüfung am: 10.05.2016

Acknowledgments

My PhD thesis and the related research was a highly interdisciplinary and collaborative endeavor, and I would like to acknowledge the people who have influenced, assisted and motivated me during the past years.

Firstly, I would like to thank Dr. Johannes Söding for his great supervision, his inspiration, his interesting and creative ideas and the many fruitful discussions we had.

I am also grateful to Dr. Stefan Krebs for handling the Genome Analyzer II-x, conducting the experiments and providing crucial insights regarding NGS sequencing and HiTS-FLIP. In addition, my thanks goes to Dr. Christophe Jung for his participation in my PhD project and for validating the HiTS-FLIP Kds with his fluorescence anisotropy based assay. Furthermore, I would like to thank Gregor Witte for participating in the HiTS-FLIP research and providing his expertise on protein handling. Many thanks also to Dr. Armin Meier for his numerous contributions and the fascinating discussions I could share with him.

I am also grateful to Dr. Dietmar Martin, Prof. Klaus Förstemann, Prof. Mario Halic, Prof. Ulrike Gaul and Prof. Roland Beckmann for offering their time as members of my dissertation committee.

Furthermore, I would like to thank the transcription team Mark Heron, Anja Kiesel, Matthias Siebert and Dr. Phillipp Torkler for their stimulating feedback and help during my PhD time. My gratitude also goes to Susanne Vorberg, Stefan Seemayer, Jessica Andreani, Markus Maier, Maria Hauser, Holger Hartmann, Jörn Marialke and Eckhart Guthörlein for providing a nice atmosphere in our group.

I highly appreciate the financial support I received by BioSysNet, the Bavarian Research Network for Molecular Biosystems, and would like to thank Dr. Ulrike Kaltenhauser and Prof. Horst Domdey for their role as organizers.

In addition, I would like to thank all the scientists inside and outside the Gene Center with whom I had the pleasure to discuss various scientific ideas.

Lastly, I would like to deeply thank my mother and my sister for their continuous support and encouragement during my entire PhD period.

Summary

In order to understand in more depth and on a genome wide scale the behavior of transcription factors (TFs), novel quantitative experiments with high-throughput are needed. Recently, HiTS-FLIP (High-Throughput Sequencing-Fluorescent Ligand Interaction Profiling) was invented by the Burge lab at the MIT (Nutiu et al. (2011)). Based on an Illumina GA-IIx machine for next-generation sequencing, HiTS-FLIP allows to measure the affinity of fluorescent labeled proteins to millions of DNA clusters at equilibrium in an unbiased and untargeted way examining the entire sequence space by determination of dissociation constants (Kds) for all 12-mer DNA motifs. During my PhD I helped to improve the experimental design of this method to allow measuring the protein-DNA binding events at equilibrium omitting any washing step by utilizing the TIRF (Total Internal Reflection Fluorescence) based optics of the GA-IIx. In addition, I developed the first versions of XML based controlling software that automates the measurement procedure. Meeting the needs for processing the vast amount of data produced by each run, I developed a sophisticated, high performance software pipeline that locates DNA clusters, normalizes and extracts the fluorescent signals. Moreover, cluster contained k-mer motifs are ranked and their DNA binding affinities are quantified with high accuracy. My approach of applying phase-correlation to estimate the relative translative offset between the observed tile images and the template images omits resequencing and thus allows to reuse the flow cell for several HiTS-FLIP experiments, which greatly reduces cost and time. Instead of using information from the sequencing images like Nutiu et al. (2011) for normalizing the cluster intensities which introduces a nucleotide specific bias, I estimate the cluster related normalization factors directly from the protein images which captures the non-even illumination bias more accurately and leads to an improved correction for each tile image. My analysis of the ranking algorithm by Nutiu et al. (2011) has revealed that it is unable to rank all measured k-mers. Discarding all the clusters related to previously ranked k-mers has the side effect of eliminating any clusters on which k-mers could be ranked that share submotifs with previously ranked k-mers. This shortcoming affects even strong binding k-mers with only one mutation away from the top ranked k-mer. My findings show that omitting the cluster deletion step in the ranking process overcomes this limitation and allows to rank the full spectrum of all possible k-mers. In addition, the performance of the ranking algorithm is drastically reduced by my insight from a quadratic to a linear run time. The experimental improvements combined with the sophisticated processing of the data has led to a very high accuracy of the HiTS-FLIP dissociation constants (Kds) comparable to the Kds measured by the very sensitive HiP-FA assay (Jung et al. (2015)). However, experimentally HiTS-FLIP is a very challenging assay. In total, eight HiTS-FLIP experiments were performed but only one showed saturation, the others exhibited protein aggregation occurring at the amplified DNA clusters. This biochemical issue could not be remedied. As example TF for studying the details of HiTS-FLIP, GCN4 was chosen which is a dimeric, basic leucine zipper TF and which acts as the master regulator of the amino acid starvation response in Saccharomyces cerevisiae (Natarajan et al. (2001)). The fluorescent dye was mOrange. The HiTS-FLIP Kds for the TF GCN4 were validated by the HiP-FA assay and a Pearson correlation coefficient of $R = 0.99$ and a relative error of $\delta = 30.91\%$ was achieved. Thus, a unique and comprehensive data set of utmost quantitative precision was obtained that allowed to study the complex binding behavior of GCN4 in a new way. My downstream analyses reveal that the known 7-mer consensus motif of GCN4, which is TGACTCA, is modulated by its 2-mer neighboring flanking regions spanning an affinity range over two orders of magnitude from a Kd= 1.56 nM to Kd= 552.51 nM. These results suggest that the common 9-mer PWM (Position Weight Matrix) for GCN4 is insufficient to describe the binding behavior of GCN4. Rather, an additional left and right flanking nucleotide is required to extend the 9-mer to an 11-mer. My analyses regarding mutations and related $\Delta\Delta G$ values suggest long-range interdependencies between nucleotides of the two dimeric half-sites of GCN4. Consequently, models assuming positional independence, such as a PWM, are insufficient to explain these interdependencies. Instead, the full spectrum of affinity values for all k-mers of appropriate size should be measured and applied in further analyses as proposed by Nutiu et al. (2011). Another discovery were new binding motifs of GCN4, which can only be detected with a method like HiTS-FLIP that examines the entire sequence space and allows for unbiased, de-novo motif discovery. All these new motifs contain GTGT as a submotif and the data collected suggests that GCN4 binds as monomer to these new motifs. Therefore, it might be even possible to detect different binding modes with HiTS-FLIP. My results emphasize the binding complexity of GCN4 and demonstrate the advantage of HiTS-FLIP for investigating the complexity of regulative processes.

Contents

List of Figures

Abbreviations

AD	Activation domain
BCL	Illumina base calls per cycle
BFGS	Broyden-Fletcher-Goldfarb-Shannon
BSA	Bovine serum albumin
BunDLE-seq	Binding to Designed Library, Extracting and Sequencing
bZIP	Basic Leucine Zipper Domain
Cas9	CRISPR associated protein 9
ChIP-seq	Chromatin immunoprecipitation sequencing
CRE	Cyclic AMP response element
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DBD	DNA binding domain
DBP	DNA binding profile
DFT	Discrete Fourier Transform
DNA	Deoxyribonucleic acid
DSB	Double Strand Break
dsDNA	Double-stranded DNA
ER	Endoplasmic reticulum
ETS	E26 transformation-specific or E-twenty-six
Exd	Extradenticle

FA	Fluorescence anisotropy
FFT	Fast Fourier Transform
FP	Fluorescence polarization
FWHM	Full width half maximum
GA	Genome Analyzer
GAAC	General amino acid control
HDR	Homology directed repair
HiP-FA assay	High performance fluorescence anisotropy assay
HiTS-FLIP	High-Throughput Sequencing-Fluorescent Ligand Interaction Profiling
Hox	Homeobox
HT-SELEX	High throughput SELEX
iPSC	Induced pluripotent stem cell
L-BFGS-B	Limited Memory Boxed BFGS
LoG	Laplacian of Gaussian
LPS	Lipopolysaccharide
MITOMI	Mechanically induced trapping of molecular interactions
MLE	Maximum-likelihood estimation
NGS	Next Generation Sequencing
NHEJ	Non-homologous end joining
PBM	Protein Binding Microarray
PDB	Protein Data Bank
POU	Pituitary-specific Pit-1, Octamer transcription factor proteins Oct-1 and Oct-2, neural Unc-86 transcription factor
PSF	Point spread function

PWM	Position weight matrix
RGN	RNA-Guided Nucleases
RNA	Ribonucleic acid
SELEX	Systematic evolution of ligands by exponential enrichment
SELEX-seq	Systematic evolution of ligands by exponential enrichment sequencing
svb	Shavenbaby
TALEN	Transcription activator-like effector nuclease
TF	Transcription factor
TFBS	Transcription Factor Binding Sites
TIRF	Total Internal Reflection Fluorescence
UAS	Upstream activating sequences
Ubx	Ultrabithorax
UCE	Ultraconserved elements
UPR	Unfolded protein response
UPRE	Unfolded protein response element
WGS	Whole genome sequencing
ZFN	Zinc finger nuclease

1 Introduction

Gene regulation is a fundamental process in molecular biology and essential for all living organisms since it enables cell differentiation, maintenance, division, and adaptability to the environment. The regulation of genes, i.e. when and at what rate proteins are expressed in a cell, occurs at a variety of different stages. The first step is termed transcription in which a particular segment of DNA is converted into RNA by the enzyme RNA polymerase. In this phase, transcription factors (TFs) play a crucial role and have an important influence on cell fate through the interpretation of regulatory DNA within the genome of an organism.

A defining feature of transcription factors is that they contain one or more DNA-binding domains (DBDs), which attach to specific DNA sequences adjacent to the genes they regulate. Despite intensive research, a comprehensive understanding of the underlying mechanisms by which TFs select in vivo binding sites and alter gene expression remains still unclear (Slattery et al. (2014)).

One key question concerning the DNA-binding specificity is how TFs can very precisely identify their functional binding sites (typically ~5-15 bp long) in a cellular environment at the right location and time. A related question here is how the transcriptional behaviour of various genes can be understood from their DNA sequence and how the bindings of TFs to these sequences are determinants of prediction for gene expression. In order to systematically tackle this question a full exploration of the entire sequence space is needed which determines the binding affinity landscape of a TF, the range of affinities for every possible sequence combination up to a certain length (Segal and Widom (2009)). This TF specific binding affinity landscape leads to a distinct distribution of molecule binding configurations for a particular sequence, and consequently, to a distinct transcriptional behaviour for any given combination of DNA sequence and binding concentrations (Segal and Widom (2009)).

The full in vitro measurement of the binding affinity landscape of a TF forms its in vitro DNA binding profile (DBP) (Wang et al. (2011)). Stated in Wang et al. (2011) there are several important insights related to the in vitro DBP of a TF such as follows:

- the generation of accurate DNA-binding models (such as position weight matrices, PWMs)

- identification of all DNA-binding sites and target genes of TFs in the whole genome
- construction of transcription regulatory networks
- biomedical applications, such as transcription therapy, which uses TFs as targets for disease therapy (Li and Sethi (2010); Stellrecht and Chen (2011); Yeh et al. (2013)), or as another example, designing artificial TFs as means in human gene therapy to turn off malfunctioning, disease causing genes (Asuka et al. (2014))

Using in vivo instead of in vitro data has the limitations that the genomic regions identified are typically hundreds of base pairs long and the derived binding specificities might also reflect the specificities of other factors (Segal and Widom (2009)).

Therefore, studying in vitro DBPs of TFs is an essential research field with far reaching implications for understanding basic molecular mechanisms and finding cures for diseases. The research focus of this thesis deals with a new experimental method for the in vitro measurement of the binding affinity landscape of a TF, called HiTS-FLIP (High-Throughput Sequencing-Fluorescent Ligand Interaction Profiling) (Nutiu et al. (2011)), which allows to measure binding affinities of all possible k-mers (DNA motifs of sequenced reads) up to the length of 12 bp. So far high-throughput in vitro methods, such as protein-binding microarrays (Bulyk et al. (1999); Mukherjee et al. (2004)) and microfluidic platforms (Maerkl and Quake (2007)) only allowed to measure all possible ~8-10 bp sequences.

Utilizing next-generation sequencing (NGS) technology for measuring DNA and RNA binding proteins to explore the entire sequence space and to determine all relevant thermodynamic properties for each DNA motif is required to move the understanding of gene regulation forward and elucidate cellular mechanisms and regulatory networks on a system wide level in ultimate depth. Methods like HiTS-FLIP (Nutiu et al. (2011)), HiTS-RAP (Tome et al. (2014)) and RNA-MaP (Buenrostro et al. (2014)) have given already important examples how powerful the realization of such an approach is and what novel biological insights can be reached. These experiments were carried out on repurposed Illumina's NGS platforms, which can be envisioned to be the basis for such kind of studies. Besides the biochemical protocols, a crucial part in the analysis due to the different requirements and the large data volumes being produced is a well crafted software pipeline capable to handle all data processing needs for the arising scientific use cases. To be of wide applicability such a software pipeline should be designed as a general, modular set of user selectable algorithms and components as an open-source platform providing all processing steps from the image analysis up to the determination of equilibrium constants. In addition, the pipeline should enable different analysis techniques and automated tests for comparing different results leading to the most accurate biological insights for the data at hand.

In the following sections, the biological background, main scientific questions involved, details on the studied TF GCN4, related experimental methods, building blocks of HiTS-FLIP, the performed experiments, the pipeline and its components, and the downstream analyses and biological discoveries are described in depth.

2 TF-DNA recognition

2.1 Motivation

Gene regulation in vivo is a very complex and multi-layered process with numerous players involved. This includes the nucleotide sequence, 3D structure and flexibility of TFs and their binding sites, TF–DNA binding in the presence of cofactors, cooperative DNA-binding of TFs, chromatin accessibility and nucleosome occupancy, indirect cooperativity via competition with nucleosomes, pioneer TFs that bind to nucleosomal DNA, and DNA methylation (Slattery et al. (2014)). In addition, interactions exist among all of these factors, which might alter binding in a cell type-specific manner and in different modes at different time points during development (Slattery et al. (2014)). Up to now, the mechanisms by which TFs select in vivo binding sites and alter gene expression remain unclear (Slattery et al. (2014)). There is still much to discover and learn about TF-DNA interactions. Despite the artificial setting, in vitro experiments can greatly elucidate various aspects how TFs bind DNA in a bottom-up approach, providing building blocks for an improved and ever increasing understanding of the inner workings of transcriptional gene regulation. A better understanding of TF-DNA binding requires the ability to quantitatively model TF binding to accessible DNA as its first basic step, before additional in vivo components can be considered.

As an example from Zhao et al. (2012), improved specificity models that are based on in vitro binding data can be very useful for assessing how consistent in vivo location data are with the expected binding sites. When predicted genomic binding sites are not observed in ChIP-seq data, one can usually assume that those locations are not accessible. But when binding is observed in locations without predicted binding sites, or with only very low predicted affinity sites, that implies either indirect or cooperative binding mediated through some other factor(s) that binds directly to the DNA (Gordân et al. (2011)). Such indirect and cooperative binding events can lead to the discovery of interacting TFs that coordinately control gene expression. But to be confident about which ChIP-seq peaks are not due to direct binding one needs an accurate model for the specificity of the TF.

In the following, some of the most important research questions regarding TF-DNA interaction mechanisms are described for which in vitro methods are valuable research tools and thus amenable to HiTS-FLIP.

2.2 Base readout

Base readout (also called "direct readout") is the formation of hydrogen bonds or hydrophobic contacts with functional groups of the DNA bases, primarily in the major groove (Seeman et al. (1976)). The preference for a given nucleotide at a specific position is mainly determined by physical interactions between the amino acid side chains of the TF and the accessible edges of the base pairs that are contacted. These contacts include direct hydrogen bonds, water-mediated hydrogen bonds, and hydrophobic contacts. The underlying question here is to what effect does the DNA sequence dictate and control the TF binding behavior? Differently phrased, if we know a certain stretch of DNA sequence to what extent can we accurately predict for a given TF its binding affinity to this sequence? A prominent example for base readout is the formation of bidentate hydrogen bonds between arginine residues and guanine bases in the major groove of DNA (Honig and Shakked (2012)). In vitro methods such as PBM and SELEX-seq have been applied in many research projects to determine the sequence specificities and binding profiles of various TFs from different TF families.

In (Wei et al. (2010)) all human and mouse ETS (E26 transformation-specific or E-twenty-six) factors were analysed which are characterized by an evolutionary conserved ETS domain and play important roles in cell development, cell differentiation, cell proliferation, apoptosis, tissue remodeling as well as cancer progression (Oikawa and Yamada (2003)). ETS factor DNA-binding profiles were determined by microwell-based TF-DNA binding specificity assays as well as PBMs. Both approaches revealed that the ETS-binding profiles cluster into four distinct classes, and for a member of each class the specificities were confirmed in vivo using ChIP-seq showing that enrichment of ETS class PWMs matched well with ChIP-seq peak sequences.

Another study (Franco-Zorrilla et al. (2014)) characterized sequence specificity of 63 plant TFs representing 25 families using PBM. Analyses of co-regulated genes and transcriptomic data from TF mutants showed the functional significance of over 80% of all identified sequences and of at least one target sequence per TF. Strong overrepresentation of DNA motifs determined in vitro was obtained with sequences in the promoters of deregulated genes in mutant or overexpressing genotypes.

Finally, (Orenstein and Shamir (2014)) analysed 162 human and mouse TFs regarding their sequence specificity using in vitro methods HT-SELEX and PBM and good predictive power was shown for in vivo binding applying ChIP-seq data (eight most informative positions for DNA motifs, AUC of 0.732 and 0.719, p-value=0.18 Wilcoxon signed-rank test).

2.3 Shape readout

Shape readout (also called “indirect readout”) is the recognition of the 3D structure of the DNA double helix (Rohs et al. (2009a)). Since DNA shape is a function of the nucleotide sequence, an important question is if DNA shape is a direct determinant of protein-DNA recognition. It has long been recognized that every base pair has a unique hydrogen-bonding signature in the major groove, but that this is not the case in the minor groove (Rohs et al. (2009b)). Thus, the expectation has been that the recognition of specific DNA sequences would take place primarily in the major groove by the formation of a series of amino-acid- and base-specific hydrogen bonds (Garvie et al. (2001)).
It was shown in (Rohs et al. (2009b)) that the binding of arginine residues to narrow minor grooves is a widely used mode for protein–DNA recognition. This readout mechanism exploits the phenomenon that narrow minor grooves strongly enhance the negative electrostatic potential of the DNA. Thus, the marked enrichment of arginines in narrow regions of the DNA minor groove provides the basis for a new DNA recognition mechanism that is used by many families of DNA-binding proteins (Rohs et al. (2009b)). The minor-groove geometry was analysed with the software *Curves44* (reference in Rohs et al. (2009b)) using in vitro data, all 1031 crystal structures of protein–DNA complexes in the PDB that have any amino acid contacting base atoms. According to (Rohs et al. (2009b)) protein side chains contact the minor groove in 69% of those structures that have at least one helical turn of DNA.
Abe et al. (2015) teased base and shape readout apart in the context of Hox-DNA binding by mutating residues that, in a co-crystal structure, only recognize DNA shape. Hox genes (also known as homeotic genes) contain a DNA sequence known as the homeobox and are organized on the chromosome in the same order as their expression along the anterior-posterior axis of the developing animal (Pearson et al. (2005)), very different from many other genes which are scattered randomly in the genome. Hox proteins are transcription factors that control the body plan of an embryo along the anterior-posterior (head-tail) axis and bind to enhancers where they either activate or repress genes (Pearson et al. (2005)). Complexes made in (Abe et al. (2015)) with these mutants lost the preference to bind sequences with specific DNA shape features. However, introducing shape-recognizing

residues from one Hox protein to another swapped binding specificities in vitro, studied with SELEX-seq, and affected gene regulation in vivo analysing embryos. Therefore, Abe et al. (2015) concluded that shape readout is a direct and independent component of binding site selection by Hox proteins.

Zhou et al. (2015) integrated 3D DNA shape information derived from SELEX-seq into the modeling of TF binding specificities. Four distinct shape features were applied, namely minor groove width, propeller twist, roll, and helix twist, which had been shown to be important for protein-DNA recognition in specific cases (Zhou et al. (2015), references therein). Using support vector regression, quantitative models of TF binding specificity based on PBM data were trained for 68 mammalian TFs. Their results showed that shape-augmented PBM-trained models compared favorably to sequence-based models.

2.4 Interdependence of individual nucleotides

Currently, the most widely used mathematical representation of TF specificity is the position weight matrix (PWM) model (Stormo (2000)). This model assumes the positions within the binding site are independent, and the contribution at one position of the binding site to the overall affinity does not depend on the identity of nucleotides in other positions of the site.

According to (Siggers and Gordân (2014a)), disagreement with a PWM model may be due to:

(i) a protein having multiple binding modes, which will require multiple PWMs, or

(ii) poor or biased parameterization of the PWM model.

PWMs can capture low-affinity binding sites but must be explicitly parameterized using low-affinity binding data (Weirauch et al. (2013)).

Quantitative analysis of high-throughput binding data has shown that PWMs are a good quantitative model for most TFs (Zhao et al. (2012)). In (Zhao et al. (2012)), the results of a quantitative analysis were achieved using more than 400 TF specificity data obtained by the universal PBM technology (Berger et al. (2006)), which are available in the UniPROBE database (Robasky and Bulyk (2011)). Using the binding energy estimate by maximum likelihood for PBM program, (BEEML-PBM, Zhao et al. (2012)), to parameterize specificity models of varying complexity it was found that improvements from incorporating interactions between positions are usually small, although there were some significant exceptions. The interactions between neighboring bases are stronger than interactions between non-neighboring bases as found by (Zhao et al. (2012)). This pattern of nearest-neighbor interactions holds true for the zinc finger class, which has 89 members including C2H2, C4, C6, and GATA zinc finger domains. It also includes the nuclear transcription factor Hnf4a. The 25 TFs of the zipper class, including the basic leucine zipper (bZIP) and the basic helix-loop-helix (bHLH) domains, like Yap1, HLH-26, Myf6, Jundm2, Cbf1, GCN4, appear to have benefited the most from the inclusion of nearest-neighbor interactions, consistent with previous information (Berger et al. (2006); Maerkl and Quake (2007); Nutiu et al. (2011)). By contrast, none of the 24 high-mobility group (HMG) TFs benefited substantially from including adjacent dinucleotide energy contributions. While there are data showing nonindependence between positions for at least some HMG proteins, those appear to be relatively minor contributions overall, as found previously for several zinc finger proteins (Bulyk et al. (2002)). In summary, the finding of (Zhao et al. (2012)) demonstrate that some TF families are more likely to require interaction models than others and that GCN4 shows the pattern of nearest-neighbor interactions.

GCN4 binds DNA as a homodimer where each monomer binds optimally to the half-site sequence 5'-TGAC-3'(Ellenberger et al. (1992a); Sellers et al. (1990a)). Using HiTS-FLIP, Nutiu et al. (2011) discovered that substitutions at positions T_1, G_2 and A_3 in the GCN4 consensus 7-mer motif, $T_1G_2A_3C_4T_5C_6A_7$ resulted in larger increases in Kd, i.e. greater weakening of binding, than at the corresponding positions T_1', G_2'and A_3'of the right half-site. This result confirmed the asymmetry in binding to the 7-mer consensus caused by the preference for C at the 4th position, with stronger binding observed to the left than to the right half-site (Sellers et al. (1990b)). By examining pairwise substitutions relative to the consensus on the inferred change in Gibbs free energy, Nutiu et al. (2011) revealed extensive interdependence. The incremental effect on binding of a second mismatch in the same half-site was consistently lower than the effect of the corresponding mismatch in the opposite half-site, that is, two mismatches in the same half-site disrupt binding less than a single mismatch in each half-site (Nutiu et al. (2011)). Nutiu et al. (2011) suggest a model in which a substitution at one position in a half-site tends to weaken the interaction of the associated GCN4 monomer with other positions in the same half-site, perhaps through a subtle protein conformational change, making interactions between the other monomer and half-site more critical. Therefore, models that assume independence, such as the commonly used PWM model, cannot accurately capture the complex DNA binding affinity landscape of GCN4. Instead, Nutiu et al. (2011) advocate the use of the full spectrum of Kd values estimated by HiTS-FLIP for all k-mers of appropriate size, e.g. 8, 9, 10, 11 or 12 bp, depending on the specific protein and depth of data.

2.5 Effect of spacing and orientation

Combinatorial transcription factor binding is essential for cell-type-specific gene regulation (Ng et al. (2014)). One question here is to what extent constrained spacing and orientation of multiple interacting TFs are critical for regulatory element activity, another question how different spacing and orientation variations act as additional determinants of specificity and allow the modulation of binding behaviour of a single TF.

As described in (Siggers and Gordân (2014b)), GCN4 dimers can bind to biparte sites with half-sites (TGAC/G) separated by variable-length spacers. For example, the two half-sites can be bound by GCN4 overlapping or adjacent (Gordân et al. (2011); Zhu et al. (2009)).

Jolma et al. (2013) used HT-SELEX and observed formation of dimers for a large set of TFs, with strong orientation and spacing preferences. These preferences were applied by Jolma et al. (2013) to further classify TF subfamilies that had identical primary specificities. In addition, Jolma et al. (2013) showed that models incorporating adjacent dinucleotides, dimer spacing and orientation preferences improved modeling of TF binding to DNA and that the dimer model can be generalized to analyze large heteromeric TF-DNA complexes. Dimer orientation and spacing preferences could be used to further classify some factors that showed similar monomer binding specificities. For example, the ETS class I factors ERG, ETS1, and ELK1 preferred to bind to different homodimeric sites (Jolma et al. (2010)). Similarly, both T box factors and forkhead proteins displayed one type of monomer specificity but seven and three distinct dimeric spacing/orientation preferences, respectively. Next, Jolma et al. (2013) tested whether orientation and spacing preference matrix could be used to improve prediction of sequences enriched by TBX20, a factor that binds to a dimeric site where the same monomer is found in multiple different orientation and spacing configurations. For this purpose, Jolma et al. (2013) generated expected-observed plots for all possible combinations of two 4-mers with gaps of different length between them (gapped 8-mers). A model that incorporated spacing and orientation preferences described enriched gapped 8-mers much better (R2 = 0.67 compared to 0.44) than a simple PWM. Many TF families could be further subclassified by Jolma et al. (2013) based on more subtle differences in specificity within the families or on a combination of monomer specificity and spacing and orientation preferences. For example, nuclear receptors are known to bind to dimeric sites that vary in both specificity and spacing of the half-sites (Pardee et al. (2011)). Clear classification of nuclear receptors to different specificity groups has, however, not been accomplished (Jolma et al. (2013)). The systematic analysis described in (Jolma et al. (2013)) allowed classification of nuclear receptors to 12 classes based on a combination of half-site and dimer orientation and spacing preferences. Similarly, although all T box proteins bound to identical half-sites, seven different classes could be identified based on spacing and orientation preferences (Jolma et al. (2013)). ETS class I proteins also displayed three distinct dimer orientations and spacings. A more complex classification of factors was necessary for bZIP proteins, which are known to vary in both specificity and spacing of the half-sites (Badis et al. (2009); Kim and Struhl (1995)). Jolma et al. (2013) found that many bZIP proteins bind to two sites and that the specificities form a tiled pattern, where in many cases, two factors shared one site and also each bound to another separate site. Such a tiled organization of TF specificity allows a complex control of target genes based on the expression and activity of the particular bZIP factors present in a given cell. The binding of TFs to DNA is commonly modeled based on a PWM that assumes independence of binding of protein to individual bases. Several alternative models that do not make this independence assumption and instead use a larger set of parameters to describe TF-DNA binding have been developed (for example, Agius et al. (2010); Roulet et al. (2002)). Based on their observation that adjacent bases commonly affect each other, and that many TFs bind DNA as monomers or dimers, Jolma et al. (2013) developed two models for TF binding that incorporate these features. The first model was a simple replacement for a PWM that is based on a first-order Markov chain. This model takes into account the effect of adjacent bases and models binding of factors that bind to A or T stretches significantly better than a conventional PWM. The second model developed by Jolma et al. (2013) takes into account the spacing and orientation preferences of dimeric sites. This improved models for TFs that bind to DNA both as monomers and dimers or as multiple different dimers.

2.6 Multiple DBDs

The DNA-binding domains of eukaryotic transcriptional activators play a key role in selective promoter activation by tethering activation domains to the appropriate promoters and by coordinating the assembly of specific sets of transcription factors on these promoters (Herr and Cleary (1995)).

One example for proteins with multiple DBDs are POU (for Pit, Oct, UNC) proteins, which are eukaryotic transcription factors containing a bipartite DNA binding domain referred to as the POU domain (Herr and Cleary (1995)). The POU domain is the conserved DNA binding domain of a family of gene regulatory proteins. It consists of a POU-specific domain and a POU homeodomain, connected by a variable linker region.

Oct-1 is a ubiquitously expressed POU domain transcription factor and can bind to different DNA sites using different arrangements of its two DNA binding domains POU$_S$ and POU$_H$ (Klemm and Pabo (1996); Verrijzer et al. (1992)).

2.7 Influence of flanking nucleotides

Eukaryotic cells often express, at the same time, TFs with highly similar DNA binding motifs but distinct in vivo targets. Currently, it is not well understood how TFs with seemingly identical DNA motifs achieve unique specificities in vivo. What could be possible influences?

Siggers et al. (2012) examined how the DNA bases flanking 10-bp kB sites affect the binding to ten different dimers from mouse and human to a wide-ranging set of 3285 potential kB site sequences. In order to determine whether PBM-determined dimer-specific differences correlated with dimer specific binding differences in vivo, Siggers et al. (2012) examined an NF-kB ChIP dataset in which ChIP-chip was performed on LPS-stimulated human macrophages and a high correlation was found.

Gordân et al. (2013) used custom PBMs to analyze TF specificity for putative binding sites in their genomic sequence context. Examining yeast TFs Cbf1 and Tye7, Gordân et al. (2013) found that binding sites of these bHLH TFs (i.e., E-boxes) are bound differently in vitro and in vivo, depending on their genomic context. Cbf1 and Tye7 have highly similar DNA binding specificities according to consensus sequences PWMs from ChIP-chip data (Harbison et al. (2004)), or PWMs from universal PBM data (Zhu et al. (2009)). Computational analyses with regression-based models by Gordân et al. (2013) elucidated that sequence features not only in the proximal but also the distal flanks contribute to different DNA binding specificity. Namely, the DNA shape features in flanking regions are distinct for binding sites preferred by Cbf1 versus Tye7, and the genomic sequences flanking the E-Box motif contribute to explaining the differences in in vivo DNA binding between Cbf1 and Tye7 (Gordân et al. (2013)). This suggests that nucleotides outside E-box binding sites contribute to specificity by influencing the three-dimensional structure of DNA binding sites. Thus, the local shape of target sites might play a widespread role in achieving regulatory specificity within TF families (Gordân et al. (2013)).

In (Levo et al. (2015a)), a new method named BunDLE-seq (Binding to Designed Library, Extracting and Sequencing) was developed by the authors that provided quantitative measurements of TF binding to thousands of fully designed sequences of 200 bp in length within a single experiment. For the yeast TFs GCN4 and GAL4, Levo et al. (2015a) demonstrated that sequences outside the core TF binding sites profoundly affected TF binding, and that TF-specific models based on the sequence or DNA shape of the regions flanking the core binding site are highly predictive of the measured differential TF binding in vivo. These observations demonstrate the need for a more comprehensive understanding of the various factors influencing TF binding to regulatory sequences, going beyond the characterization of core binding sites (Levo et al. (2015a)). Notably, the selected TFs are structurally distinct and are representatives of the two most abundant yeast TF families (basic leucine zipper, bZIP, class and zinc cluster domain class, respectively, Hahn and Young (2011)). The conclusion Levo et al. (2015a) arrived at was that whereas sequences sharing the well characterized strong binding site for either GCN4 or GAL4 showed pronounced differences in binding, a simple TF-specific model accounting for 3-bp flanks successfully predicted these differences, and that DNA shape features provide a mechanistic explanation for the effect of flanking sequences.

2.8 Different binding modes

Another phenomenon by which TFs can differentiate their DNA binding behavior is by different binding modes.

Fordyce et al. (2012a) investigated Hac1, a S. cerevisiae bZIP TF involved in the highly conserved unfolded protein response (UPR). In S. cerevisiae, two main proteins are responsible for enacting the UPR: Ire1, a transmembrane kinase/endonuclease, and Hac1 (Fordyce et al. (2012a)). Unfolded proteins bind to the Ire1 domain facing the ER lumen, triggering its oligomerization and activation of its cytoplasmic endonuclease domain. Once activated, Ire1 cleaves Hac1 mRNA at two sites and tRNA ligase rejoins the severed exons via an unconventional spliceosome independent mechanism (Chapman et al. (1998)). This splicing removes an intron to produce a new transcript (denoted Hac1i mRNA; “i” for “induced”), thereby relieving translational inhibition exerted by the intron. Following translation of the spliced mRNA, Hac1i is translocated to the nucleus, where it regulates a large set of UPR-responsive genes (Rüegsegger et al. (2001)). Despite the central role played by Hac1i in activating the UPR, the rules by which Hac1i recognizes UPR target genes remain unclear. To obtain an unbiased assessment of Hac1i binding preferences, Fordyce et al. (2012a) used a microfluidic platform, MITOMI (mechanically induced trapping of molecular interactions, Fordyce et al. (2010)), to measure relative binding affinities ($\triangle\triangle G$) between Hac1i and 70 bp double-stranded oligonucleotides containing overlapping instances of all possible 8 bp combinations. In vivo studies of Hac1i are complicated by both the very short half-life of the Hac1i isoform derived from the spliced mRNA and the tendency of bZIP transcription factors to homo- and heterodimerize (Fordyce et al. (2012a)). Therefore by necessity, in vitro approaches provide a particularly valuable tool for accurately defining binding preferences (Fordyce et al. (2012a)). In addition, Fordyce et al. (2012a) analyzed expression of reporter genes driven by a variety of Hac1i mutants to identify the protein residues required for target site recognition. Fordyce et al. (2012a) discovered that Hac1i bind both long (11-13 bp), extended UPRE-1-like motif called extended core UPRE-1 or xcUPRE-1, and compact (6-7 bp) UPRE-2 DNA target sites. The 12-bp sequence of xcUPRE-1 is 5'-GGACAGCGTGTC-3'and the 6-bp sequence of UPRE-2 is 5'-TACGTG-3'. Fordyce et al. (2012a) suggest that changes in the conformation of Hac1, from the N-terminal region of extended homology, leads to recognition of one site or the other. To what purpose does Hac1i recognize multiple distinct sites? For the glucocorticoid receptor, DNA sequences can act as allosteric ligands, inducing conformational changes to preferentially recruit specific cellular co-factors with functional consequences for transcriptional activation (Meijsing et al. (2009)). A similar scenario may apply to Hac1i, and perhaps to other bZIP family members, although additional studies will be required to determine whether changes in protein conformation within the DNA binding domain can propagate elsewhere within the protein (Fordyce et al. (2012a)). Alternatively, dual site recognition could represent a snapshot in evolutionary time of a transcriptional network rewiring event in progress. According to this notion, it may have been advantageous to place an additional set of target genes under Hac1i control, perhaps as a handoff of some other transcriptional program (Fordyce et al. (2012a)). In this light, it is interesting to note that the Hac1i-driven transcription program in S. cerevisiae has been split into multiple transcriptional branches in metazoans, indicating evolutionary network plasticity (Fordyce et al. (2012a)).

2.9 Weak binding sites

TFs can specifically utilize low-affinity DNA-binding sites to regulate genes (Siggers and Gordân (2014a)). TF binding to low-affinity DNA sites can provide a mechanism for interpreting both spatial (Cotnoir-White et al. (2011); Struhl (1987)), and temporal (Rowan et al. (2010)) TF gradients that often arise during development to control where and when genes are expressed. Analysis of genome-wide binding data has also provided evidence that low-affinity sites are under wide-spread evolutionary selection (Jaeger et al. (2010); Tanay (2006)) and that their inclusion can greatly improve quantitative models of TF binding and gene regulation used for predicting segmentation patterns during early embryonic development in Drosophila (Segal et al. (2008)). Utilization of sites selected to be lower affinity than an optimal sequence opens the door for functionally relevant sites to deviate strongly from the consensus sequence and may not be well represented by a particular binding model (Siggers and Gordân (2014a)). For example, a comprehensive analysis of DNA binding by NF-κB dimers identified numerous lower affinity, non-traditional sites that differ significantly from the consensus sites and are not captured by the widely used PWMs (Siggers and Gordân (2014a); Wong et al. (2011)). According to Tanay (2006), transcription factors bind DNA stochastically and it is therefore expected that they would be interacting with promoters at different levels of specificity, depending on an affinity that is determined (at least partially) by the DNA sequence. Tanay (2006) developed an algorithm that predicts DNA-binding energies from sequences and ChIP data across a wide dynamic range of affinities and used them to reveal widespread functionality of low-affinity transcription factor binding in S.cerevisiae. Instead of focusing on a set of a few dozens of high-specificity hits for each TF, ChIP experiments are analyzed quantitatively in (Tanay (2006)), using (possibly noisy) estimates on TF-binding affinities for thousands of promoters. Applying PWMs for sequence-based predictions of TF affinities and comparing these predictions to ChIP binding ratios Tanay (2006) was able to test if low-specificity binding detected by ChIP provides quantitative indication to variability in in vivo binding strengths, or is by and large a noisy indication to biological cases of high-specificity targets. The results by (Tanay (2006)) showed that PWM predictions and ChIP binding ratios were highly correlated, thereby suggesting that binding of TFs to low-affinity promoters occurs abundantly in vivo, is determined by promoter sequences, and constitutes a substantial fraction of the interaction between TFs and DNA. One way to test whether these abundant weak TF–gene interactions carry functional relevance is to estimate their level of evolutionary conservation. Taking evolution into account, the predicted TF binding energies of orthologous promoters from different yeast species were shown to be more conserved than expected by neutrality (Tanay (2006)). Conservation analysis by (Tanay (2006)) suggested that selection due to a single TF may affect significant parts of the S.cerevisiae genome (10%-20%), much more than expected by purifying selection on strict binding sites. This finding was supported by analysis of gene expression. In conditions that activate a TF, one may associate the TF-binding affinity with a measurable change in gene expression for a large part of the genome (10% and more). According to these results, low-affinity TF–gene interactions are important features of genomic regulatory programs, with possible roles in fine-tuning the transcriptional phenotype and in providing abundant evolutionary raw material for its continuous modification. According to the results, conservation of energy is detectable in a large number of promoters, greatly exceeding the top few affinity percentiles predicted to have significant binding sites. For example, Gcn4 and Cbf1 are estimated to affect roughly 10% of the genome (Gcn4 may affect more weakly an additional 10%). The conservation of energies predicted for other TFs may be even broader. Mbp1 and Ume6 conservation peak at the top 5%, but remain significant on up to half of the affinity spectrum. For several of the TFs, conservation is observed on a significant fraction of the genome (10%–20%), reflecting widespread selection on the binding energy of promoters lacking high-affinity binding sites. The study by (Tanay (2006)) demonstrates that we can use ChIP experiments, so far considered to indicate only high-affinity TF targets, to quantify weak transcriptional interactions and combine them with promoter sequence analysis. One can therefore exploit comprehensive ChIP experiments to outline an "analog" model for transcriptional networks, and to explore the role of low-specificity, probabilistic TF–DNA interactions in genomic regulatory programs. According to the evolutionary and gene expression analysis reported in (Tanay (2006)), it is likely that many of the low-specificity transcriptional interactions in yeast are weakly functional. According to (Tanay (2006)) it is shown that for substantial parts of the genome, the total binding energy (and not just the existence of a binding site) is conserved and that on average, promoters with low predicted binding affinities can still generate gene expression. Evolutionarily, transcriptional programs in which a discrete logic is softened by a combination of low-affinity interactions may be more flexible. Such programs can allow changes to be gradually accumulated, therefore alleviating selective pressure on specific loci (e.g., classical binding sites) and increasing their ability to evolve. If binding of a TF to low-affinity promoters is functionally important, one would expect to observe selection operating not only on individual binding sites, but also on the total affinity of each promoter to that TF. A gene weakly regulated by a TF may be pushed to remain so in the course of evolution, but the pressure would not be focused on a specific locus but would be dispersed over the entire promoter, selecting for the integrated binding energy over many possible weak loci.

Raijman et al. (2008) developed a probabilistic model for the evolution of promoter regions in yeast, combining the effects of regulatory interactions of many different transcription factors. The model expressed explicitly the selection forces acting on transcription factor binding sites in the context of a dynamic evolutionary process. Raijman et al. (2008) examined the evolutionary dynamics in Saccharomyces species promoters and revealed relatively weak selection on most binding sites. Moreover, according to the estimates of (Raijman et al. (2008)), strong binding sites are constraining only a fraction of the yeast promoter sequence that is under selection. Using their new techniques, Raijman et al. (2008) was able to express a substantial part of the current functional knowledge on gene regulation in evolutionary terms and evaluate observed patterns of divergence and conservation based on this model. Specifically, Raijman et al. (2008) used their models to study the intensity of selection on TFBSs and to estimate the amount of promoter region under selection due to high specificity TFBSs. Given their results, it is evident that even on very short evolutionary time scales transcriptional regulation in yeast is highly dynamic. Taken together, it can be hypothesized that much of the functionality of transcriptional networks is encoded in ways other than strong TFBSs, and that due to high levels of redundancy, binding sites are under continuous remodeling (Raijman et al. (2008), references therein). Rather than being a deterministic and sparse network, transcriptional programs may be shaped as dense, noisy networks that are continuously changing during evolution.

Jaeger et al. (2010) used recently published universal PBM data on the in vitro DNA binding preferences of these proteins for all possible 8-base-pair sequences, and examined the evolutionary conservation and enrichment within putative regulatory regions of the binding sequences of a diverse library of 104 non-redundant mouse TFs spanning 22 different DNA-binding domain structural classes. These 8-mers occur preferentially in putative regulatory regions of the mouse genome, including CpG islands and non-exonic ultraconserved elements (UCEs). Jaeger et al. (2010) found that not only high affinity binding sites, but also numerous moderate and low affinity binding sites, are under negative selection in the mouse genome. The results of (Jaeger et al. (2010)) indicate that many of the sequences bound by these proteins in vitro, including lower affinity DNA sequences, are likely to be functionally important in vivo. Taken together, Jaeger et al. (2010) provide evidence supporting that lower affinity TF binding sites, as determined from PBMs, serve evolutionarily conserved, in vivo regulatory functions.

Segal et al. (2008) showed that in Drosophila embryonic development low affinity TF binding sites are important in gene regulation.

Crocker et al. (2015) demonstrated that the Hox protein Ultrabithorax (Ubx) in complex with its cofactor Extradenticle (Exd) bound specifically to clusters of very low affinity sites in enhancers of the shavenbaby (svb) gene of Drosophila. These low affinity sites conferred specificity for Ubx binding in vivo, but multiple clustered sites were required for robust expression when embryos developed in variable environments. Although most individual Ubx binding sites are not evolutionarily conserved, the overall enhancer architecture - clusters of low affinity binding sites - is maintained and required for enhancer function. Natural selection therefore works at the level of the enhancer, requiring a particular density of low affinity Ubx sites to confer both specific and robust expression. The results by (Crocker et al. (2015)) helped to explain previous difficulties with bioinformatic prediction of functional Hox binding sites, because low affinity sites are difficult to detect reliably. Indeed, the low affinity sites that implement Hox regulation within svb enhancers share little similarity with canonical Hox or Hox-Exd binding sites. Consequently, a very large number of seemingly disparate DNA sequences can confer low affinity binding for Hox proteins. If Hox-Exd sites are often clustered in the genome, then signals from genome-wide ChIP-seq will reflect binding to the entire cluster and the signals associated with individual low affinity sites may be difficult to discern from noise. Identification of important low affinity sites will require a change in computational approaches to analyzing genome-wide data. Currently, it is de rigueur to apply an arbitrary threshold to genome-wide data and then to analyze only signals above this threshold. This approach is likely to bias detection toward high affinity sites, whose functions may be distinct from those of clusters of low affinity sites (Crocker et al. (2015)).

Afek and Lukatsky (2013) showed with an equilibrium biophysical model for protein-DNA binding that non-consensus protein-DNA binding in yeast is statistically enhanced, on average, around functional Reb1 motifs that are bound as compared to nonfunctional Reb1 motifs that are unbound. The landscape of non-consensus protein-DNA binding around functional CTCF motifs in human demonstrated a more complex behavior (Afek and Lukatsky (2013)). In particular, human genomic regions characterized by the highest CTCF occupancy, showed statistically reduced level of nonconsensus protein-DNA binding. The findings by Afek and Lukatsky (2013) suggest that non-consensus protein-DNA binding is fine-tuned around functional binding sites using a variety of design strategies. Two quite different design strategies for non-consensus protein-DNA binding are pointed out by (Afek and Lukatsky (2013)) which might be operational in the genome:

1) The first design strategy (positive design) enhances the level of non-consensus protein-DNA binding in the vicinity of binding sites. Such an enhancement might guide sequence-specific TFs toward their specific binding sites, greatly speeding up their diffusion (Berg et al. (1981)). The existence of an optimal strength for nonspecific protein-DNA binding has been demonstrated theoretically in the past (Slutsky and Mirny (2004)), and once such an optimal strength is exceeded, the diffusion of TFs slows down (Slutsky and Mirny (2004)).

2) The second design strategy (negative design) is quite the opposite: it reduces the level of non-consensus protein-DNA binding in the vicinity of binding sites. Such strategy might statistically reduce the competition of CTCF with other, nonspecific TFs, near specific CTCF binding sites, thus facilitating specific binding.

Afek and Lukatsky (2013) suggested that such non-consensus binding landscape provides a background surrounding specific DNA motifs, and possibly regulating the kinetics of transcription regulators in their search for such specific motifs (Afek and Lukatsky (2013), references therein). Therefore, the predicted non-consensus protein-DNA binding mechanism could represent yet an additional layer of transcriptional regulation operating in vivo, which influences genome-wide protein-DNA binding preferences in an eukaryotic cell.

2.10 Off-target occurrences

Studying protein-DNA interactions in vitro not only spurs basic research investigating underlying mechanisms and principles of TF binding behavior but also crucial biomedical applications. Gene therapy is based on the principle of the genetic modification of living cells for use in treating various disorders. The final goal of gene therapy is to cure patients who suffer from genetic disorders, including cancer, congenital and infectious diseases (Liu and Fan (2014)). One approach is based on targeted genome editing using custom made nucleases, such as zinc finger nucleases (ZFNs) (Urnov et al. (2010)), transcription activator effector nucleases (TALENs) (Joung and Sander (2013)), and the clustered regulatory interspaced short palindromic repeat Cas9 (CRISPR-Cas9) RNA-guided nuclease system (RGNs) (Sander and Joung (2014)). These customized nucleases have enabled efficient and targeted genome editing in a wide variety of cell types and organisms, including human induced pluripotent stem cells (iPSCs) (Tsai and Joung (2014)). As stated in (Tsai and Joung (2014)), DNA double-stranded breaks (DSBs) induced by these customizable nucleases can be repaired by one of two competing pathways in the cell: error-prone nonhomologous end-joining (NHEJ), which leads to variable length insertion/deletion mutations (indels), or homology-directed repair (HDR), which can be used to introduce precise alterations directed by a homologous DNA template.

There was recently a successful clinical trial regarding HIV patients that were treated with ZFN-mediated CCR5-modified autologous CD4 T cells (Tebas et al. (2014)). For HIV to enter host cells, CD4 antigens and chemokine receptors, such as CCR5 or CXCR4, are required to invade macrophages and T-helper lymphocytes (Stone et al. (2013)). A 32 bp homozygous deletion between the transmembrane domains of CCR5 (the CCR5D32 mutation) results in a frameshift mutation in which affected individuals display high resistance to HIV-1 infection (Samson et al. (1996)). Functional knockout of CCR5 in autologous CD4 T cells of a small cohort of patients revealed that in one out of four enrolled subjects, the viral load remained undetectable at the time of treatment (Tebas et al. (2014)). Similarly, TALEN and CRISPR-Cas9 have been tested experimentally for efficient disruption of CCR5 and CXCR4 (Hu et al. (2014), references therein) and taking them into consideration for clinical trials is anticipated. Whether or not the strategies targeting HIV-1 entry can reach a sterile and permanent cure of AIDS remains to be seen.

A key question here is if the engineered nuclease act at any genomic locations besides its intended site, i.e. are there any off-targets? This is critically important because unintended, off-target modifications in cell populations can lead to unexpected functional consequences in both research and therapeutic contexts, where functional consequences of even low frequency mutations can be of significant concern (Tsai and Joung (2014)). Even though, there is active research and new studies conducted (Tsai and Joung (2014), references therein), the full genome-wide spectrum of off-target mutations induced by engineered nucleases remains as yet unclear. Whole-genome sequencing (WGS) with fold-coverage tries to address this issue but suffers from two main hurdles, i.e. systematic sequencing artifacts can make it difficult to discern nuclease-induced alterations, and WGS is currently impractical for identifying lower frequency off-target mutations (Tsai and Joung (2014)). A method like HiTS-FLIP which allows to examine the entire sequence space in an unbiased way is ideally suited to explore any off-target effects across the entire genome.

3 Background of HiTS-FLIP

3.1 Introduction

The Illumina Genome Analyzer builds millions of distinct clusters on a flow cell, each consisting of several hundred to around one thousand identical DNA molecules. Clusters are sequenced by synthesis in situ, with individual fluorescently tagged nucleotides visualized using a charge-coupled device camera to reconstruct the DNA sequence of each cluster (Bentley et al. (2008)).

Nutiu et al. (2011) reasoned that fluorescently tagged proteins could be added to the flow cell and their binding to each DNA cluster visualized in the same way as fluorophore-tagged nucleotides. Protein bound clusters could subsequently be matched to the corresponding DNA sequences based on their position in the flow cell, enabling direct observation of the DNA binding preferences of the fluorescently tagged protein.

3.2 Protocol

The HiTS-FLIP protocol shown in Figure 3.1 consists of the following steps:

1) Illumina-based NGS experiment, determining bases for ~100 million clusters of genomic or random synthetic DNA. Most imaging systems have not been designed to detect single fluorescent events, so amplified templates are required to increase the fluorescent signals, for which Illumina uses solid-phase amplification (Fedurco et al. (2006)).

2) Denaturation of the second DNA strand since it was build with modified, i.e. fluorescently labeled and 3' blocking group attached nucleotides during DNA sequencing. The modification itself or remaining inefficiently cleaved terminators during phasing can lead to side-effects for the protein binding to the DNA clusters.

3) Washing step to remove denaturated nucleotides from the flow cell.

4) Resynthesis of second DNA strand with unmodified nucleotides to obtain double-stranded DNA.

5) Adding fluorescent labeled proteins in different concentration steps to the flow cell without any washing steps.

6) Equilibration depending on the on-rate of the examined protein.
7) Laser excitation and imaging of the flow cell for each concentration step by the TIRF-based optics system of the GA-IIx.
8) Registration for each concentration step of the fluorescent signals from each tiff image onto tile-based DNA cluster reference positions in order to map intensities to the corresponding DNA sequences.
9) Intensity extraction from the registered fluorescent signals.
10) Normalization of the extracted intensities.
11) Ranking of k-mers according to their intensities.
11) Fitting a sigmoidal function to the normalized intensities for each k-mer to obtain Kds and thus a quantitative binding affinity landscape.

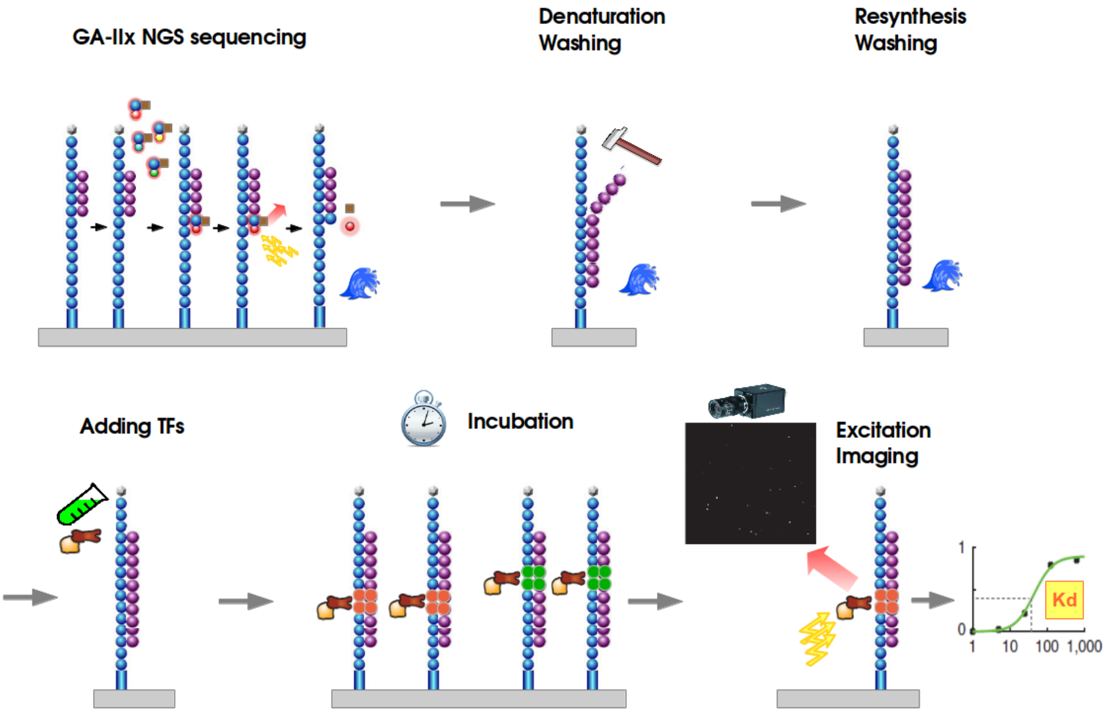


Figure 3.1: Overview of the HiTS-FLIP protocol and its different steps.

3.3 XML Encoding

The GA-IIx is operated by so called XML recipes that encode the biochemical steps of the sequencing protocol as XML dialect containing different commands to control the hardware of the GA-IIx. These commands can be used for encoding the entire HiTS-FLIP protocol. The following provides an example of a few XML encoded protocol steps.

XML HiTS-FLIP example

```
<TileSelection>
    <Incorporation>
        <Lane Index=”1”><RowRange Max=”60” Min=”1” /></Lane>
        <Lane Index=”2”><RowRange Max=”60” Min=”1” /></Lane>
        <Lane Index=”3”><RowRange Max=”60” Min=”1” /></Lane>
        </Incorporation>
    <ReadPrep>
        <Row Index=”5” />
        <Row Index=”26” />
        <Row Index=”45” />
    </ReadPrep>
</TileSelection>

<Chemistry Name=”Protein_conc_625nM_2h”>
    <PumpToFlowcell Solution=”13” AspirationRate=”50” DispenseRate=”2500” Volume=”205” />
    <Wait Duration=”600000” />
    <Temp Temperature=”20” />
    <PumpToFlowcell Solution=”13” AspirationRate=”50” DispenseRate=”2500” Volume=”10” />
    <Wait Duration=”600000” />
    <PumpToFlowcell Solution=”13” AspirationRate=”50” DispenseRate=”2500” Volume=”10” />
    <Wait Duration=”600000” />
    <PumpToFlowcell Solution=”13” AspirationRate=”50” DispenseRate=”2500” Volume=”10” />
    <Wait Duration=”5400000” />
    <Temp Temperature=”20” />
    <TempOff />
</Chemistry>
```

3.4 Optics of GA-IIx

The GA-IIx has two excitation lasers and two filters. shown in Figure 3.2, in order to distinguish between four fluorescent signals (Bentley et al. (2008)). The excitation wavelength of the red laser is 660 nm, and of the green laser is 532 nm (Bentley et al. (2008)). Emission wavelengths are not published by Illumina. The fluorescent dyes Illumina uses for DNA sequencing are probably related to Alexa dyes. Alexa Fluor 555 and Alexa Fluor 647 dyes provide higher confidence than Cy3 and Cy5 dyes in determining significant differences in gene expression on microarrays (Staal et al. (2005)). The fluorescent dye used by Nutiu et al. (2011) is mOrange, a fluorescent protein monomer with excitation wavelength of 548 nm and emission wavelength of 562 nm (Shaner et al. (2005)). Because dimerization and specific DNA binding involves residues situated at the

C terminus of the protein the N-terminal fusion of the GCN4-mOrange construct should have minimal effect on DNA binding characteristics (Hope and Struhl (1986)).

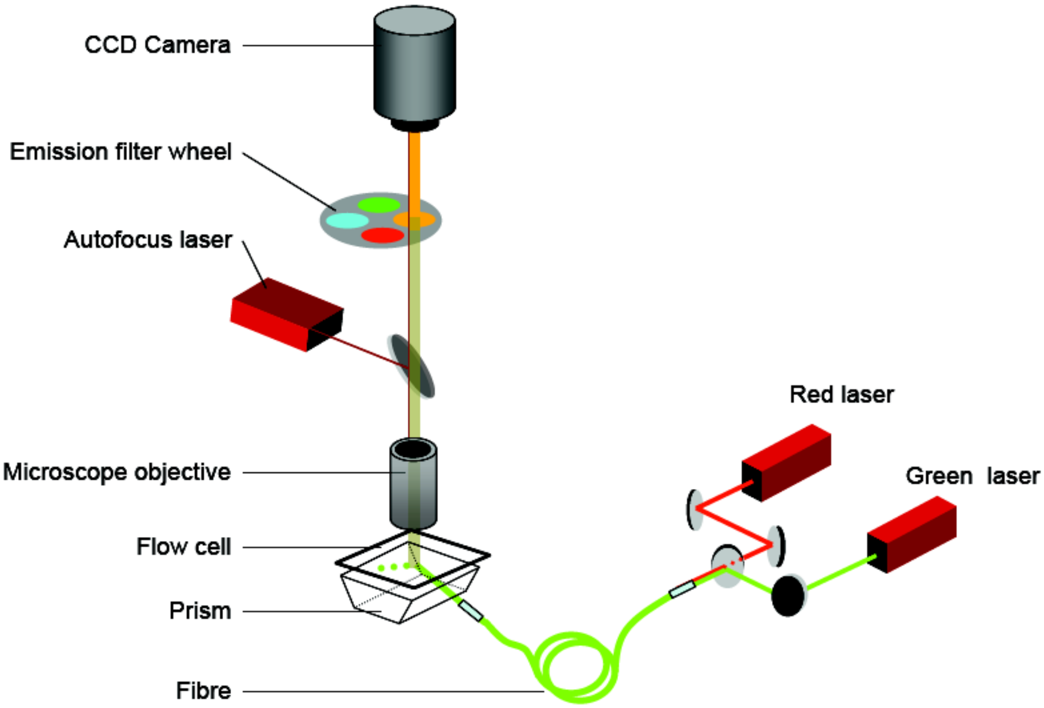


Figure 3.2: Figure adapted from (Bentley et al. (2008)). Overview of the GA-IIx optical components for imaging the flow cell. Red and green lasers provide excitation beams that are directed along an optical fibre and through a prism which is in contact with the flow cell. Excitation of fluorescent nucleotides incorporated into DNA clusters on the inner surface of the flow cell leads to a base-specific emission that passes through an objective and a filter wheel and the signal is collected by a CCD camera. Autofocus utilises a third laser (635 nm) that is projected through the objective onto the flow cell (Bentley et al. (2008)).

The GA-IIx applies a TIRF based optics which creates an evanescent wave reaching only ~100-200 nm into the flow cell (Bentley et al. (2008)). Therefore, only those proteins which are bound to the DNA clusters on the inner surface of the flow cell, illustrated in Figure 3.3, are excited making any washing during a HiTS-FLIP run unnecessary.

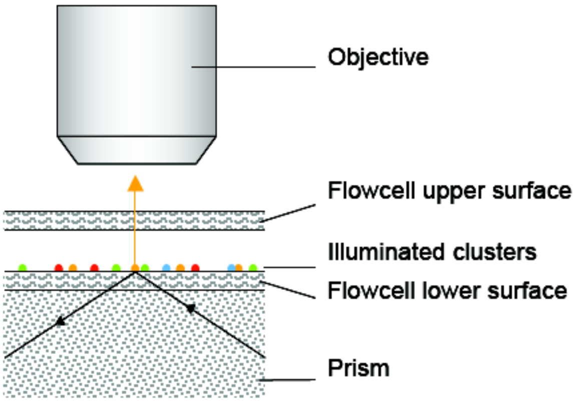


Figure 3.3: Figure adapted from (Bentley et al. (2008)). Total internal reflection of the incident excitation beam at the glass-buffer interface generates an evanescent wave that excites the clusters on the surface. The fluorescence emission is captured by a custom made microscope objective, passed through a filter and is then projected onto a CCD. The evanescent wave excitation technique maximises the sensitivity of signal detection while minimising background noise (Bentley et al. (2008)).

3.5 Imaging of the flow cell

In each cycle, the flow cell is imaged in a series of non-overlapping regions. The flow cell is physically divided into eight separate lanes, each lane is virtually divided into two columns, and each column is further virtually divided into 60 tiles (Bentley et al. (2008)). A tile is the area that gets imaged during a DNA sequencing or HiTS-FLIP run. An illustration is provided by Figure 3.4.

The CCD camera is stationary and the flow cell is moved under the camera in order to image each tile in each cycle. Four images are taken per tile, one for each base. Each GA-IIx image is a 1888 x 2048 pixel 16 bit gray-scale TIFF (though only 12 bits contain data). The tile size is 0.5274 mm^2, the tile is roughly square which gives an approximate width and height of 0.7262 mm (personal communication with the Illumina tech support, March 2014), each pixel covers ca. 0.14 μm^2 and on average ca. 3×3 pixels comprise one cluster object. The time for imaging a single tile is $\sim$ 2.7 sec (personal communication with the Illumina tech support, March 2014).

Because of the finite accuracy of the movements of the motion stage, images taken at different sequencing cycles have random translational offsets with respect to each other (Bentley et al. (2008)). Furthermore, images taken in different frequency channels have

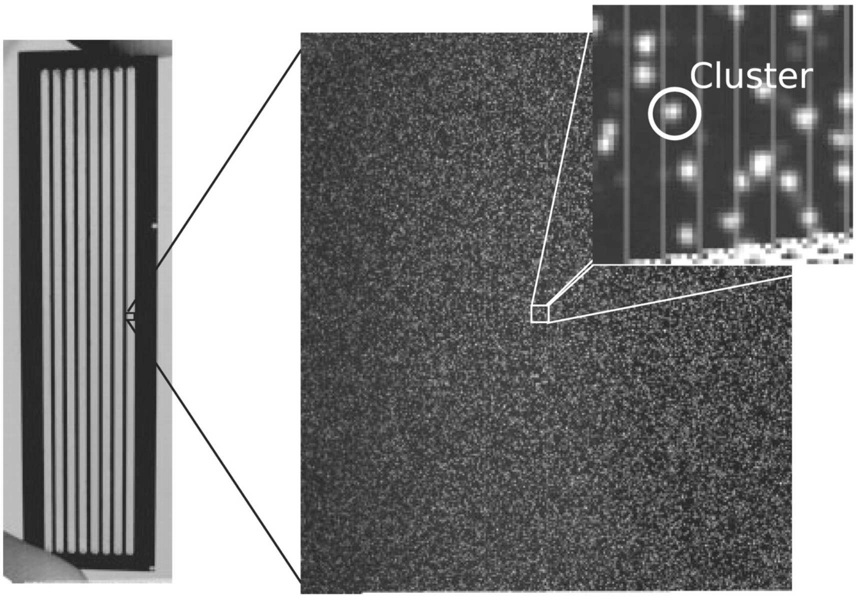


Figure 3.4: Figure adapted from (Whiteford et al. (2009)). It shows the Illumina GA-IIx flow cell with its eight lanes and a zoom-in on one tile and its DNA clusters.

different optical paths and wavelengths and experience further, albeit smaller, translations and scale transformations (Bentley et al. (2008)).

In order to correct for the image shifts and scalings, the cluster positions that were extracted from the four images taken in the first five cycles are super-imposed to construct a "reference image" for each tile containing all detected clusters. Transformations of the image coordinates to later cycles are then obtained from a cross-correlation of the taken images in later cycles to the reference images.

4 Biology of GCN4

4.1 Introduction

The eukaryotic transcriptional activator protein GCN4 is a transcription factor in S.cerevisiae and belongs to the bZIP family of DNA-binding proteins, which has more than 50 known members from yeast, mammalian and plant cells (Krylov (2001)). The name arose because leucines occur every seven amino acids in the dimerization domain and are critical for dimerization and DNA binding (Krylov (2001)). GCN4 binds specifically to HIS3 promoters of yeast amino acid biosynthetic genes, which code for enzymes required to synthesize all 20 major amino acids (Hope and Struhl (1987)). In general, transcription factors from the bZIP family recognize promoter and enhancer regions of transcribed genes and, together with other protein factors, contribute to the efficiency by which RNA polymerase binds and initiates transcription.

4.2 Composition

In total, GCN4 comprises 281 amino acids and is structured into two transcriptional activation domains (ADs), the highly charged basic motif, which constitutes the DNA binding domain, and the leucine-zipper as the dimerization domain. The leucine zipper is located towards the C-terminus and its helical extensions that make up the basic region towards the N-terminus (Krylov (2001)).

The two transcriptional activation domains (residues 1–100 and 101–134) are unrelated in sequence apart from their acidic character (Brzovic et al. (2011)). These tandem acidic ADs act in conjunction with the coactivators Mediator, SAGA, and SWI/SNF (Brzovic et al. (2011), references therein).

Figure 4.1 provides an overview of the composition of GCN4 and Figure 4.2 shows the parallel coiled-coil structure of GCN4 ZIP homodimer.

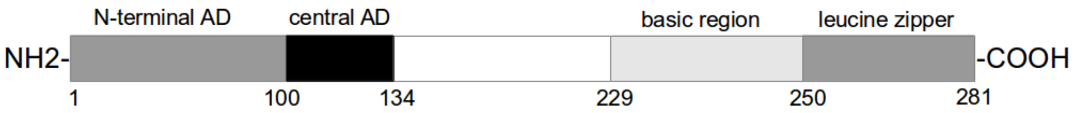


Figure 4.1: Overview of the composition of GCN4. The positions of the two ADs, the basic region and the leucine zipper are shown.

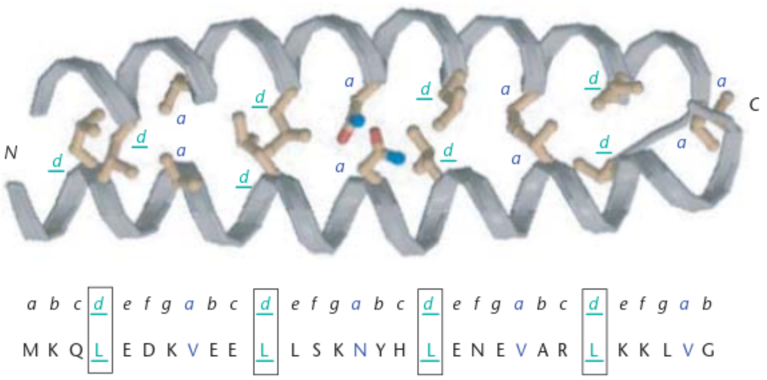

N
C
d
a
a b c d e f g a b c d e f g a b c d e f g a b c d e f g a b
M K Q L E D K V E E L L S K N Y H L E N E V A R L K K L V G

Figure 4.2: Figure adapted from (Hakoshima (2005)). Parallel coiled-coil structure of GCN4 ZIP homodimer (PDB accession code 1gd2). The main chains of the two peptide chains are represented as ribbons in gray. The side chains participating in the dimer association are represented as stick models with carbon atoms in brown, nitrogen atoms in blue and oxygen atoms in red. The positions of the heptad repeat are labeled a–g. The d-positioned leucines are boxed and highlighted in green with underline. The a-positioned residues are highlighted in blue.

4.3 DNA binding

GCN4 forms a homodimeric complex with each monomer recognizing half of a symmetric or nearly symmetric DNA site (Hollenbeck et al. (2002)). GCN4 binds to two optimal targets, i.e. asymmetric pseudo-palindrome AP-1 9-mer site 5'-ATGACTCAT-3' and the symmetric palindrome ATF/CREB 10-mer site 5'-ATGACGTCAT-3', which has one base pair inserted in the middle of the recognition site (Hill et al. (1986)). The recognition site, ATGA(C/G)TCAT, is inherently asymmetric because it contains an odd number of base pairs and because mutation of the central C-G base pair strongly reduces specific DNA binding (Sellers et al. (1990b)).

From this asymmetry, (Sellers et al. (1990b)) suggested that GCN4 interacts with nonequivalent and possibly overlapping half-sites, ATGAC and ATGAG, that have

different affinities. In vitro, GCN4 bound efficiently to the sequence ATGACGTCAT, whereas it failed to bind to ATGAGCTCAT or ATGATCAT (Sellers et al. (1990b)). The authors of (Sellers et al. (1990b)) concluded that:
1) GCN4 specifically recognizes the central base pair,
2) The optimal half-site for GCN4 binding is ATGAC, not ATGAG, and
3) GCN4 is a surprisingly flexible protein that can accommodate the insertion of a single base pair in the center of its compact binding site.

The DNA binding domain of GCN4 is flexible and partially disordered in the absence of DNA targets (Wobbe et al. (1990)), however, the entire bZIP domain becomes fully helical when bound to DNA (Ellenberger et al. (1992b); König and Richmond (1993)). Each monomer of the GCN4 fragment forms a smoothly curved, continous alpha helix (Brzovic et al. (2011)). The leucine zipper region of the monomers pack into a coiled coil, essentially identical to the isolated leucine zipper (Brzovic et al. (2011)). The two alpha helices diverge from the dimer axis in a segment comprising the junction between the leucine zipper and the basic regions (Brzovic et al. (2011)). This fork creates a smooth bend in each alpha helix which displaces the basic regions away from the dimer interface so that they can pass through the major groove of DNA, with one alpha helix on each side of the DNA.

The flexibility of the bZIP motif is central to its binding to DNA (Harbury et al. (1993)). The GCN4 bZIP domain, like that of other bZIP proteins, is unfolded in the absence of DNA and becomes structured only on binding to its target (Harbury et al. (1993), references therein). The crystal structure of the bZIP-DNA complex provides clues about the functions of this flexibility. Flexibility is required to dock and to dissociate the protein and DNA. The protein encloses the binding site, forming a mutually complementary interface (Ellenberger et al. (1992b)).

The ability of dimers to discriminate between related DNA sequences is independent of the zipper region and is specified by amino acids both in the basic region and in the linker region immediately N-terminal to the beginning of the leucine zipper (Agre et al. (1989); Metallo and Schepartz (1994)).

In the absence of DNA, the DNA-binding region is not structured, but upon DNA binding, it becomes alpha helical, lying in the major groove of the DNA (Krylov (2001)). Each helical extension of the leucine zipper can bind up to 5 base pairs in a sequence-specific manner and thus, the dimer can bind up to 10 base pairs without crossing the DNA backbone (Krylov (2001)). For the bZIP dimer to bind DNA, the leucine zipper has to interact in parallel and in heptad register to place both basic regions in the major groove Krylov (2001). One structural feature of the leucine zipper that accomplishes this heptad register is a nearly invariant asparagine in the position of the leucine zipper of bZIP proteins (Krylov (2001)). Contacts with the DNA are mediated by residues between positions 234 and 249, and contacts to the bases are made by only five residues: Asn235, Ala238, Ala239, Ser242 and Arg243 (Harbury et al. (1993)). The core of the DNA-binding interface contains Asn235, which forms hydrogen bonds to bases C2 and T3 in each DNA half-site (Ellenberger et al. (1992b)). This key role for Asn235 is consistent with its absolute conservation in bZIP proteins that recognize the AP-1 sequence. Ala239, Ala240 and Ser242 make van der Waals contact with bases T1 and T3, and solvent is excluded from the binding site by the side chains of residues 240-243. Lys231 makes a water-mediated contact with base A4 in one half-site, perhaps accounting for the preference for purines at this position in the binding site (Ellenberger et al. (1992b)). Arg243 plays the special role of adapting the symmetric protein to the asymmetric binding site. One Arg243 side chain 'reads out' the G base in the central base pair, and the other contacts phosphates of C0 and A1 on the opposite DNA strand.

McHarris and Barr (2014) performed all-atom molecular dynamics simulations of the full-length GCN4 protein as well as three truncated variants and observed consistent sequence-specific protein-DNA contacts across all of their simulations, confirming the critical role of Asn235, Ala239, and Arg243 as identified by mutation experiments (Suckow et al. (1993)). Overall, the GCN4 bZIP-DNA crystal structures show that only four highly conserved amino acids in each basic region of the monomer make direct contacts to bases in the DNA major groove: Asn235, Ala238, Ala239, and Arg243, which highlighted by the Figure 4.3.

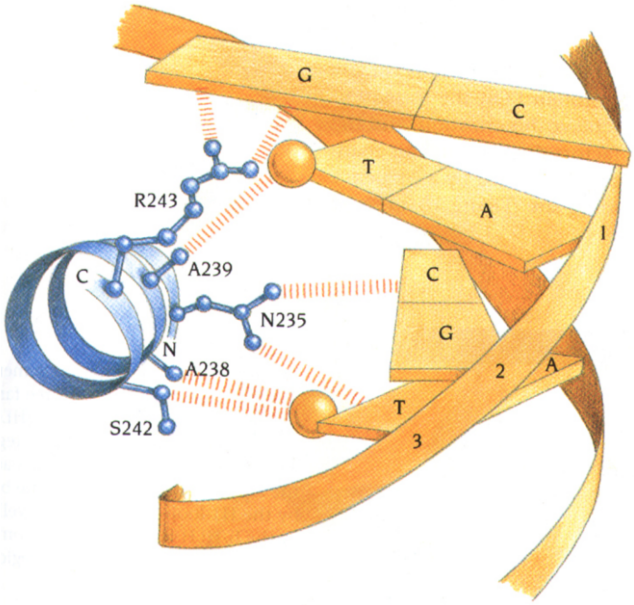


Figure 4.3: Figure adapted from (Alberts et al. (2007)). GCN4 binds to DNA with both specific and nonspecific contacts. 4 amino acid side chains form sequence-specific contacts. Asn235 is at the center of the interaction area and strictly conserved in all bZIP family members.

Binding of bZIP proteins to DNA results in dynamic effects on both DNA and protein structure (Lee (1992)). The helical transition that occurs in the basic region upon DNA binding might result in changes in overall protein conformation, which could influence interaction with other transcriptional components (Lee (1992)).

Several lines of evidence suggest that protein-DNA recognition involves non-identical contacts between GCN4 monomers and half-sites in the target DNA (Hope and Struhl (1987)).

1) First, neither the native HIS3 site nor any of the presumptive regulatory sequences in 14 other promoters activated by GCN4 are perfectly symmetric (Hill et al. (1986)).

2) Second, some symmetrical changes of the HIS3 regulatory site do not have equivalent effects on DNA binding affinity or transcriptional activation (Hill et al. (1986)).

3) Third, GCN4 binding is reduced significantly when the central C of the HIS3 site is changed to any other base including G, its symmetric counterpart (Hill et al. (1986)). This suggests that the central base pair is part of a half-site recognized by a GCN4 monomer, and given the odd number of base pairs in the palindrome, it follows that the protein-DNA interactions at the half-sites cannot possibly be identical, even for the optimal sequence. These considerations also suggest that the half-sites overlap at the

central base pair, and the overlap might conceivably be more extensive.

Hollenbeck and Oakley (2000) have found that the bZIP protein GCN4 can also bind with high affinity to DNA sites containing only a single GCN4 consensus half-site. Quantitative DNA binding and affinity cleaving studies support a model in which GCN4 binds as a dimer, with one monomer making specific contacts to the consensus half-site and the other monomer forming nonspecific contacts that are nonetheless important for binding affinity (Hollenbeck and Oakley (2000)). Given that one of the two half-sites in the consensus AP-1 site appears to be more important for GCN4 binding, multiple substitutions in the second half-site may have only a modest effect on complex stability (Hollenbeck and Oakley (2000)).

Half-site recognition by bZIP proteins may be biologically significant. Several GCN4- and AP-1-responsive promoters have binding sites that contain only one-half of the consensus core sequence (Hollenbeck and Oakley (2000), references therein). These results suggest that half-site binding may play a role in the regulation of gene activation in vivo.

Presumably, one monomer of the GCN4 dimer contacts the left-half site and the central base pair, whereas the monomer interacting with the right half-site does not contact the central position (Sellers et al. (1990b)). This view of the GCN4-DNA interaction accounts for why alterations in the right half-site are tolerated better than symmetrically equivalent alterations in the left half-site (Oliphant et al. (1989)).

The crystal structure of GCN4 complexed with its target AP-1 site (RCSB Protein Data Bank, PDB code: 1YSA), which was solved by (Ellenberger et al. (1992b)), reveals that while Arg243 of one GCN4 monomer specifically contacts the central guanine nucleotide, Arg243 from the other monomer forms non-specific hydrogen bonds with the DNA backbone (Selvaraj et al. (2002)). This observation, along with mutational and DNA binding studies, indicates GCN4–DNA binding to be inherently asymmetric, and suggests that the specific recognition of a single half-site by one GCN4 monomer may be more important than recognition by the other.

The importance of the central C-G base pair and the asymmetry of the GCN4 recognition sequence strongly support the model that GCN4 dimers bind to nonequivalent half-sites (Oliphant et al. (1989)).

It seems likely that asymmetrical contacts made with the central C-G base pair cause the GCN4 dimer to be shifted from the center of the site. In the 7-bp core, GCN4 probably interacts more avidly with the left half-site (positions -1, -2, and -3) than with the right half-site (positions +1, +2, and +3), because deviations generally occur to the right of the central base (Oliphant et al. (1989)).

In contrast to the relative importance of the left side of the core, flanking positions in the right half-site (positions +4, +5, and +6) contribute more to GCN4 binding than equivalent positions in the left half-site (positions -4, -5, and -6) do, perhaps to compensate for the relative weakness of the right side of the core (Oliphant et al. (1989)). According to Chan et al. (2007), there are three different binding modes of GCN4. Dimeric binding of basic regions on DNA full site, dimeric binding of basic regions on DNA half site and monomeric binding of basic regions on DNA half site, as shown in Figure 4.4.

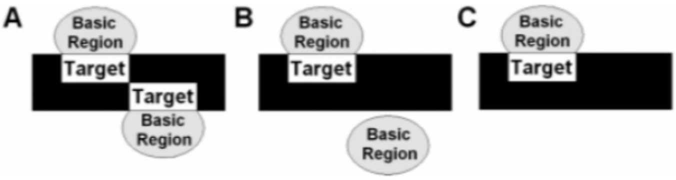


Figure 4.4: Figure adapted from (Chan et al. (2007)). (A) Dimeric binding of basic regions on DNA full site. Both basic regions of the dimer bind to target DNA half sites selectively. (B) Dimeric binding of basic regions on DNA half site. Only one basic region of the dimer binds selectively to the target DNA half site; the other basic region interacts nonspecifically with DNA. (C) Monomeric binding of basic regions on DNA half site. No protein dimerization occurs.

4.4 3D structure

As shown by Figure 4.5 GCN4 forms a “chopstick-like” homodimer of alpha helices at the DNA-binding interface. In the crystal structure of bZIP-DNA complexes, the dimeric protein binds to a DNA site with dyad symmetry, each monomer of the bZIP factor recognizing one half-site.

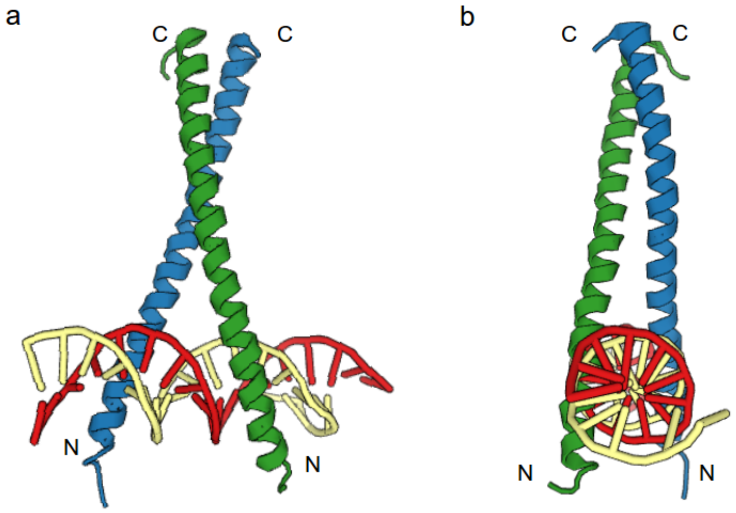


Figure 4.5: DNA binding of GCN4 in dimeric oligomerization state as described in Ellenberger et al. (1992a). (a) The bZIP dimer binds in the major groove of the DNA. Each bZIP protomer is a smoothly curved, continuous α-helix. The carboxy-terminal residues of the monomers pack together as a coiled coil, which gradually diverges to allow the basic region residues to follow the major groove of either DNA half site. This divergence of the bZIP monomers corresponds to an unwinding of the coiled-coil super helix, with a slight righthanded rotation of basic region residues about the α-helical axis of each chain and a lateral displacement of each monomer along the helical axis of the DNA. The DNA in the complex is straight, and its conformation is in the B form across the region contacted by the protein. (b) View down the DNA axis. The basic region residues amino-terminal to the point of DNA contact are in a straight, α-helical conformation. The amino-terminal residues of the basic region do not wrap around the back side of the binding site.

4.5 Dimer and monomer pathway

Dimerization of bZIP transcription factor GCN4 is linked to the folding of its C-terminal leucine zipper domain. However, monomeric GCN4, lacking a folded leucine zipper, also recognizes the DNA site with dimerization taking place on the DNA (Cranz et al. (2004)). In Cranz et al. (2004) the kinetics of DNA recognition by unfolded monomeric and folded dimeric derivatives of GCN4 were reported using a 19 bp dsDNA containing a palindromic CRE site (5'-ATGACGTCAT-3'). The rate of DNA binding of both monomeric and dimeric GCN4 has a bimolecular rate constant of 3-5 × 10^8 M^{-1} s^{-1}, which is near the diffusion limit (10^9 M^{-1} s^{-1} according to Alberty and Hammes (1958); Eigen and Hammes (2006)). Because the rate of dimerization of GCN4 is slower (1.7 × 10^7 M^{-1} s^{-1}) than the rate of DNA association, the formation of the dimeric GCN4-DNA complex through consecutive binding of two monomers (monomer pathway) is faster when starting from free monomers. Figure 4.6 provides an illustration of the dimeric and monomeric pathway with the related rate constants. The results presented by (Cranz et al. (2004)) support facilitated and rapid target recognition by the monomeric transcription factor. However, DNA binding of preformed folded dimeric GCN4 is as rapid as complex formation through the monomer pathway. Therefore, the monomer and dimer pathways are kinetically equivalent if monomeric and dimeric GCN4 are at equilibrium. Hence, the dimer pathway may also have a role under in vivo conditions. However, the observed rapid rates of DNA binding could not be accounted for if formation of a dimeric bZIP peptide had to precede DNA binding (Cranz et al. (2004), references therein). Thus, it has been proposed that monomeric transcription factors can recognize DNA and that these monomers dimerize while bound to DNA (Kim and Little (1992)). This has been confirmed by experiment for several dimeric transcription factors (Cranz et al. (2004), references therein). A monomer binding pathway may increase specificity and prevent the transcription factor from becoming trapped at nonspecific DNA sites (Cranz et al. (2004), references therein).

The results of Cranz et al. (2004) demonstrate that in the isolated system they studied, which is composed of a 19-mer dsDNA target and the 62-residue C-terminal DNA-binding domain of GCN4, both the monomeric and the dimeric transcription factor recognize the palindromic CRE target site at the same rapid rate. The association rate of the monomer is virtually the same as that of the dimer, 5 × 10^8 M^{-1} s^{-1} (Cranz et al. (2004)), however the monomer pathway is more rapid than the dimer pathway when starting from two monomeric GCN4 proteins and no dimer, but not when monomeric and dimeric GCN4 are at equilibrium. In a cellular environment, an equilibrium mixture of monomeric and dimeric transcription factors may be competing for DNA sites (Cranz et al. (2004)). Both monomeric and dimeric GCN4 can bind to DNA at a very rapid rate and, therefore, the monomer-dimer equilibrium of the free bZIP factor does not affect the overall rate of DNA recognition (Berger et al. (1998)). The monomer and dimer pathways are thermodynamically equivalent and preference for the monomer pathway is kinetic (Berger et al. (1998)). When the bZIP factor slides along the DNA, non-specific binding should be weak. Because binding strength correlates with the number of possible interactions between peptide and DNA (von Hippel and Berg (1989)), the monomeric basic region may slide along the DNA more easily than the dimer (Berger et al. (1998)). Unspecific DNA binding of the dimer could also be stronger because of more nonspecific electrostatic interactions (Cranz et al. (2004)). Less steric hindrance may also contribute to a faster diffusion rate of the monomer (Berger et al. (1998)).

Finally, accessory proteins influence the strength of the transcription factor-DNA complex. The rates of target finding and DNA binding through a monomer or dimer pathway could differ, depending on whether such accessory proteins bind to the monomeric or dimeric transcription factor, or both (Cranz et al. (2004)).

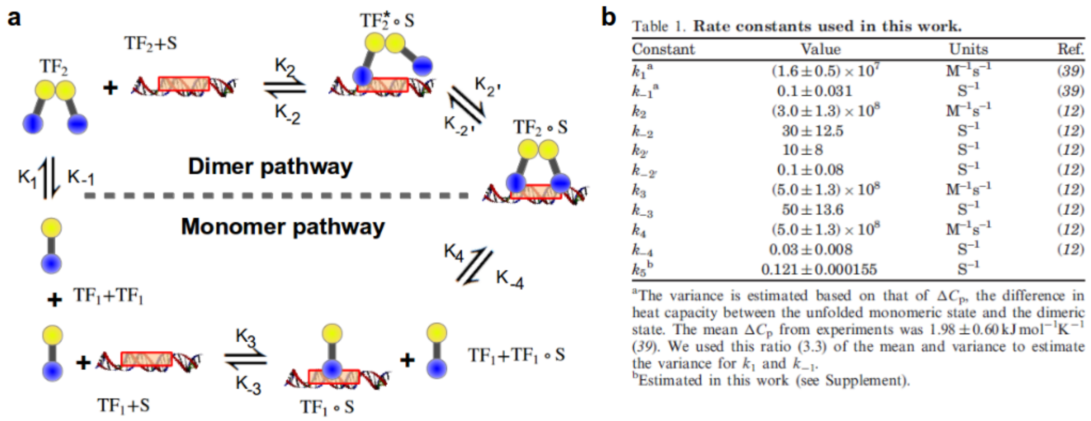


Table 1. **Rate constants used in this work.**

Constant	Value	Units	Ref.
k_1[a]	$(1.6 \pm 0.5) \times 10^7$	$M^{-1}s^{-1}$	*(39)*
k_{-1}[a]	0.1 ± 0.031	S^{-1}	*(39)*
k_2	$(3.0 \pm 1.3) \times 10^8$	$M^{-1}s^{-1}$	*(12)*
k_{-2}	30 ± 12.5	S^{-1}	*(12)*
$k_{2'}$	10 ± 8	S^{-1}	*(12)*
$k_{-2'}$	0.1 ± 0.08	S^{-1}	*(12)*
k_3	$(5.0 \pm 1.3) \times 10^8$	$M^{-1}s^{-1}$	*(12)*
k_{-3}	50 ± 13.6	S^{-1}	*(12)*
k_4	$(5.0 \pm 1.3) \times 10^8$	$M^{-1}s^{-1}$	*(12)*
k_{-4}	0.03 ± 0.008	S^{-1}	*(12)*
k_5[b]	0.121 ± 0.000155	S^{-1}	

[a] The variance is estimated based on that of ΔC_p, the difference in heat capacity between the unfolded monomeric state and the dimeric state. The mean ΔC_p from experiments was 1.98 ± 0.60 kJ mol^{-1}K^{-1} (*39*). We used this ratio (3.3) of the mean and variance to estimate the variance for k_1 and k_{-1}.
[b] Estimated in this work (see Supplement).

Figure 4.6: Overview of the dimer and monomer pathway of GCN4 and its rate constants. Figure based on (Cranz et al. (2004); Yang et al. (2007)), table adapted from (Yang et al. (2007)). (a) Dimer and monomer pathway. (b) Rate constants involved in the dimer and monomer pathway.

4.6 Gene regulation

It has been known for many years that GCN4 stimulates the transcription of more than 30 amino acid biosynthetic genes, representing 12 different pathways, in response to starvation for any of several amino acids (Hinnebusch and Natarajan (2002)). This regulatory response is known as general amino acid control (GAAC) (Hinnebusch and Natarajan (2002), references therein). Figure 4.7 shows the schematic representation of functional categories of GCN4 target genes. In two publications (Jia et al. (2000); Natarajan et al. (2001)) in which cDNA microarrays were used to conduct a genome-wide transcriptional profiling analysis of gene expression it was shown that GCN4 induces (directly or indirectly) a much larger set of genes, encompassing 10% or more of the yeast genome. Hence, GAAC is much broader with regard to the range of stimuli that elicit the response and the ensemble of genes that are transcriptionally induced (Hinnebusch and Natarajan (2002)). The broad transcriptional response controlled by GCN4 suggests that GCN4 acts as a master regulator of gene expression.

Mascarenhas et al. (2008) showed that GCN4 is required for the response to peroxide stress in S.cerevisiae. Hydrogen peroxide stress damages many intracellular targets and affects diverse cellular processes. The response to oxidative stress requires extensive reprogramming of transcription and translation. Translational control of GCN4 expression and transcriptional control of GCN4 target genes are key components of this adaptive response (Mascarenhas et al. (2008)).

One important cofactor is GAL11 which has three conserved GCN4-binding domains that bind GCN4 with micromolar affinity (Brzovic et al. (2011), references therein). These multiple, weak GCN4-GAL11 interactions additively contribute to overall transcription activation and illustrate an important principal of GAL11 recruitment by GCN4: GCN4 binds GAL11 not by a single high-affinity and high-specificity interaction but rather by multiple low-affinity interactions (Brzovic et al. (2011)).

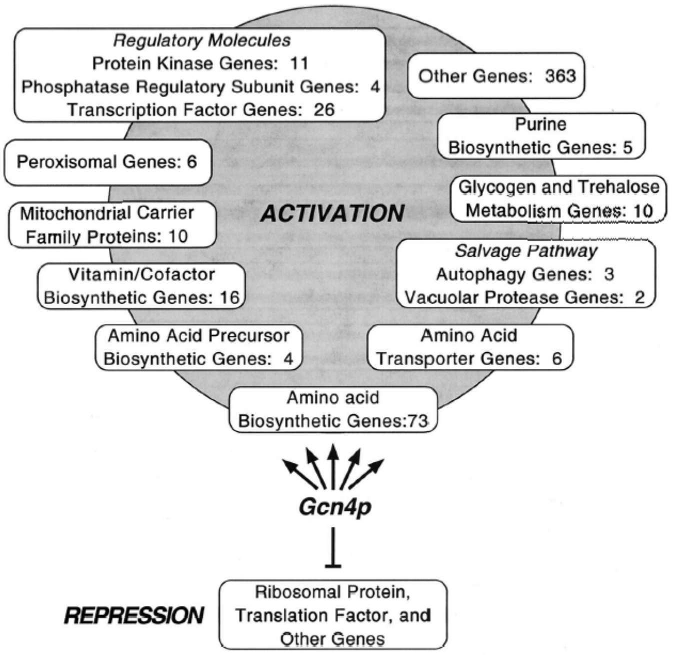

Regulatory Molecules
Protein Kinase Genes: 11
Phosphatase Regulatory Subunit Genes: 4
Transcription Factor Genes: 26
Other Genes: 363
Purine Biosynthetic Genes: 5
Peroxisomal Genes: 6
Glycogen and Trehalose Metabolism Genes: 10
Mitochondrial Carrier Family Proteins: 10
ACTIVATION
Salvage Pathway
Autophagy Genes: 3
Vacuolar Protease Genes: 2
Vitamin/Cofactor Biosynthetic Genes: 16
Amino Acid Precursor Biosynthetic Genes: 4
Amino Acid Transporter Genes: 6
Amino acid Biosynthetic Genes:73
Gcn4p
REPRESSION
Ribosomal Protein, Translation Factor, and Other Genes

Figure 4.7: Figure adapted from (Hinnebusch and Natarajan (2002)). Schematic representation of functional categories of GCN4 target genes. When GCN4 is induced under conditions of histidine starvation, it elicits the transcriptional activation of at least 539 genes, designated GCN4 targets (shown above GCN4 in the activation group).

4.7 Ribonuclease activity

Nikolaev et al. (2010) showed in vitro that c-Jun and GCN4 possessed weak but distinct ribonuclease activity and could likely catalyze degradation of RNA in vivo. In a follow-up study (Nikolaev (2011)) delineated structural details of RNA binding by the GCN4 leucine zipper motif by solution NMR experiments and elucidated that only the dimeric (coiled coil) leucine zipper conformation is capable of binding RNA. The authors hypothesized that catalytic activity of bZIP proteins in vivo will primarily be associated with the DNA-bound form of the dimeric TFs. While in other cellular contexts bZIP motifs may have little or no activity due to the prevalence of the monomer form.

5 Pipeline

5.1 Overview of the HiTS-FLIP pipeline

Figure 5.1 shows the main processing steps of the HiTS-FLIP pipeline. The Appendix section 9.10 provides a summary of the parameters, input and output.

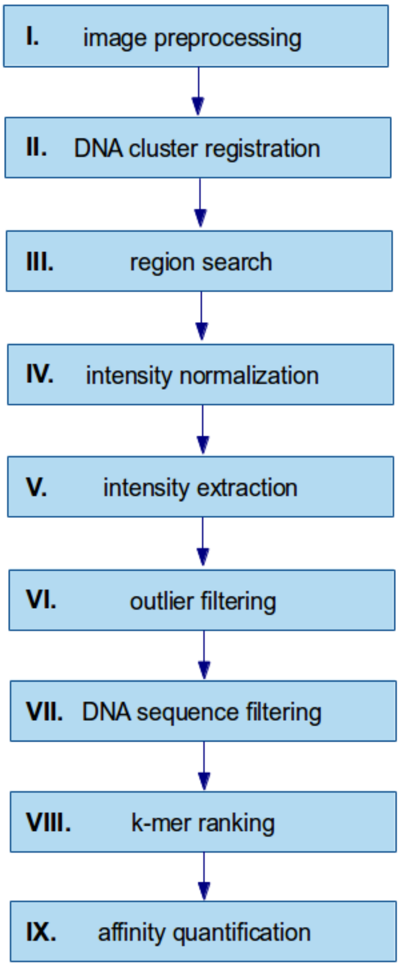

I. image preprocessing
II. DNA cluster registration
III. region search
IV. intensity normalization
V. intensity extraction
VI. outlier filtering
VII. DNA sequence filtering
VIII. k-mer ranking
IX. affinity quantification

Figure 5.1: Overview of the HiTS-FLIP pipeline and its components.

5.2 Image preprocessing

Since the raw data produced by a HiTS-FLIP experiment are 16 bit tif images, the first step in the image processing part of the pipeline is to enhance the bright spots in the images which denote the DNA clusters bound with fluorescently tagged proteins for registration with the template coordinates. This is achieved by convoluting the image with a Laplacian of Gaussian (LoG) filter which combines a Gaussian low-pass filter reducing noise and a Laplacian operator for emphasizing edges and thus better separation of the DNA clusters (Parker (2010)). The details regarding the underlying theory and the implementation are explained in Appendix section 9.1.

5.2.1 Results

In the HiTS-FLIP pipeline a LoG filter with the parameter $\sigma = 0.7644$ (using FWHM$= 1.8$ pixels) and 5×5 pixel kernel (shown in Appendix section 9.2) was applied. The processing of the protein images by the LoG filter was only applied for the cluster registration step, subsequent operations in the pipeline are carried out on the unfiltered protein images. The following two figures exemplify the filtering result with respect to the tif image of tile 6 of lane 2 at concentration 125 nM (cycle 96) from experiment 18.08.2014.

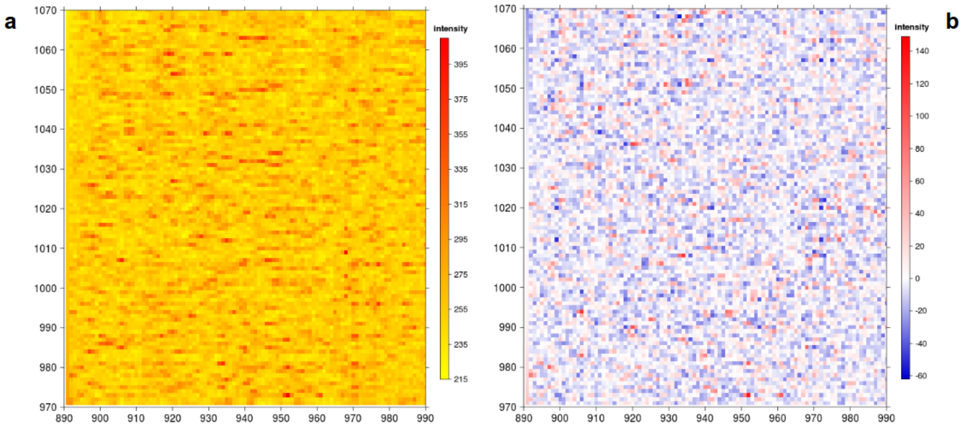


Figure 5.2: Unfiltered image and LoG filtered image. (a) shows the unfiltered image and (b) the image after filtering with the LoG filter ($\sigma = 0.7644$, kernel: 5×5 pixel, shown in Appendix section 9.2).

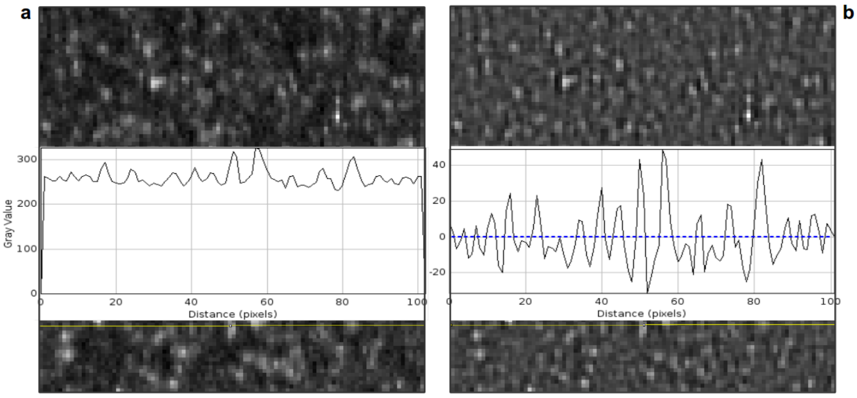


Figure 5.3: Intensity profile of unfiltered image and LoG filtered image. The intensity profile of the centered 100×100 pixel subimage at $y = 1052$ pixel is shown for (a) the unfiltered and (b) the LoG filtered image ($\sigma = 0.7644$, kernel: 5×5 pixel, shown in Appendix section 9.2). The better separation of the intensity peaks (denoting the DNA clusters) is clearly visible.

5.3 DNA cluster registration

The flow cell is mounted on a sledge which is mechanically moved during the imaging process. The CCD camera of the GA-IIx is stationary (personal communication with the Illumina tech support, January 2014). During the process of moving the flow cell, there is an x, y offset for each tile. Therefore, clusters are shifted across imaging cycles and have to be aligned so that the observed intensities can be related to the correct DNA sequences.

5.3.1 Template images

As a reference onto which all shifted clusters are aligned x, y coordinates are used that represent cluster positions without any distortion by translation or other transformations per tile. Images containing these reference coordinates are called templates and the process of aligning shifted images to these templates is called registration. The cluster positions of the template images are created at the beginning of the NGS sequencing by a spot finding procedure in the Illumina RTA pipeline as described in (Inc. (2011c)) which results in x, y coordinates for each single tile stored in the *pos* text files and *.locs* files or in compressed form as *.clocs* files.

In order to register an observed image to its related template image, the template cluster positions are used to create an artificial image which is then correlated with the observed image. In the following sections, the theoretical framework underlying the implementation in Illumina's OLB (Off-line Basecaller) pipeline version 1.9.4 (Inc. (2011a)), which has been adapted here for the HiTS-FLIP pipeline, is explained.

5.3.2 PSF of DNA cluster

Each reference cluster position in the *pos* text file is convolved with a point spread function (PSF). The PSF describes the response of an imaging system to a point source (Shaw and Rawlins (1991)), i.e. the fluorescent signal of a DNA cluster in this case, and is approximated by the $2d$ Gaussian function shown in equation 5.1. An isotropic Gaussian is a reasonable model of a circularly symmetric blob as demonstrated by (Zhang et al. (2007)). A 5×5 mask of discrete pixel values (shown in 9.4) was used to represent equation 5.1.

$$PSF_{cluster} = A \times \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \tag{5.1}$$

A: amplitude of Gaussian (set to 1.0 here).
σ: bandwidth of the filter kernel.

The FWHM (full width at half maximum) of a DNA cluster denotes the spread of the PSF of a DNA cluster which is estimated to be 1.8 pixel. Figure 5.4 illustrates the connection between FWHM, PSF and σ.

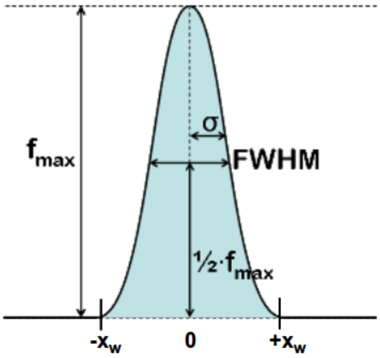


Figure 5.4: The relationship between FWHM and σ is shown schematically. Adapted from URL: `https://wiki.uio.no/mn/safe/nukwik/index.php/KJM-FYS_5920_Lab_Exercise_2_-_Student_Report`

If the filter is centered at the origin, the mean is 0 and the FWHM is the distance between the $-x_w$ and the $+x_w$ that produces the half of the peak. For the normal distribution, the mean is the same as the mode (i.e. peak) and x_w needs to be found that will result in:

$$f(x_w) = \frac{1}{2}f(x_{max}) = \frac{1}{2}f(\mu) \tag{5.2}$$

$$\exp\left(-\frac{x_w^2}{2\sigma^2}\right) = \frac{1}{2}\exp\left(-\frac{(\mu-\mu)^2}{2\sigma^2}\right) \tag{5.3}$$

$$-\frac{x_w^2}{2\sigma^2} = -\ln(2) \tag{5.4}$$

$$x_w^2 = 2\sigma^2\ln(2) \tag{5.5}$$

$$x_w = \pm\sqrt{2\ln(2)}\sigma \tag{5.6}$$

$$FWHM == +x_w - (-x_w) = 2x_w = 2\sqrt{2\ln(2)}\sigma \approx 2.3548\sigma \tag{5.7}$$

Therefore, the value of σ used for generating the template clusters is

$$\sigma = \frac{1.8}{2.3548} \approx 0.7644$$

Figure 5.5 gives an example of an observed image and the related template image from one selected tile.

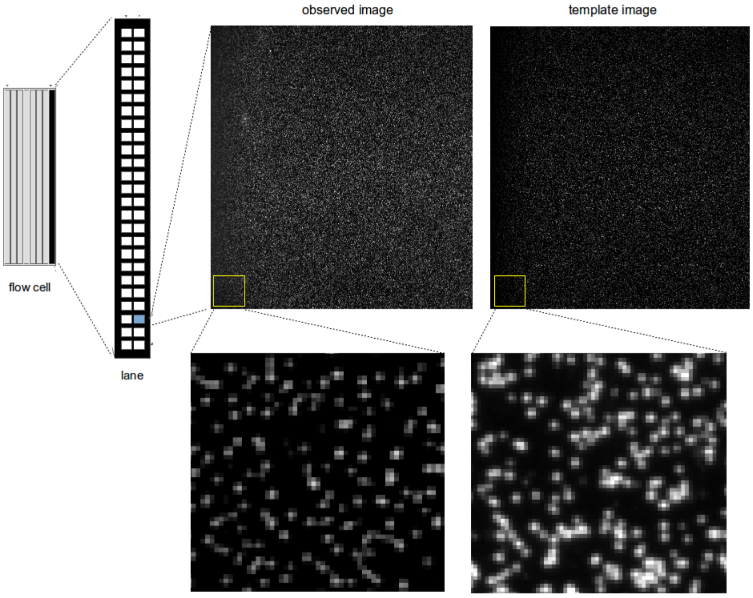


Figure 5.5: Overview of observed and template image from one selected tile. On the left hand side a flow cell is displayed, where one lane is magnified and one tile is zoomed-in. Magnified subareas are shown for the observed and related template images which need to be correlated. Template images are created as described in section 5.3.2.

5.3.3 Phase correlation

How can the template and observed images be aligned without any landmarks? Due to the phase correlation method (Kuglin and Hines (1975)) the translational x, y offsets can be estimated. Kuglin and Hines (1975) observed that information about the displacement of one image with respect to another is included in the phase component of the cross-power spectrum, i.e. the Fourier transform of the cross-correlation function of the images that measures the similarity as a function of the lag of one image relative to the other.
According to the Fourier shift property the following equations hold (Goshtasby (2012)):

$$\mathscr{F}\{g(t-a)\} = \int\limits_{-\infty}^{\infty} g(t-a)\exp\left(-2\pi ift\right)dt, a \in \mathbb{R} \tag{5.8}$$

$$= \int\limits_{-\infty}^{\infty} g(u)\exp\left(-2\pi if(u+a)\right)du, u = t-a \tag{5.9}$$

$$= \exp\left(-2\pi ifa\right)\int\limits_{-\infty}^{\infty} g(u)\exp\left(-2\pi ifu\right)du \tag{5.10}$$

$$= \exp\left(-2\pi ifa\right)G(f) \tag{5.11}$$

The original function $g(t)$ is shifted in time (or in space) by a constant amount, therefore it should have the same magnitude since the frequency content of $G(f)$ remains unchanged. A delay in time (or shift in space) only alters the phase of $G(f)$ but not the magnitude. Let the image g_a be a shifted version of the image g_b by (x_0, y_0) (Goshtasby (2012)):

$$g_a(x,y) = g_b(x-x_0, y-y_0) \tag{5.12}$$

After taking the discrete Fourier transform (DFT) of both images,

$$\mathscr{F}\{g_a\} = G_a(u,v), \mathscr{F}\{g_b\} = G_b(u,v) \tag{5.13}$$

the following relationship is obtained due to the shift property of the Fourier transform:

$$R(u,v) = \frac{G_a G_a^*}{|G_a G_a^*|} \tag{5.14}$$

where * denotes the complex conjugate.

$$= \frac{G_a G_a^* \exp\left(-2\pi i(ux_0 + vy_0)\right)}{|G_a G_a^* \exp\left(-2\pi i(ux_0 + vy_0)\right)|} \tag{5.15}$$

$$= \frac{G_a G_a^* \exp\left(-2\pi i(ux_0 + vy_0)\right)}{|G_a G_a^*|} \tag{5.16}$$

$$= \exp\left(-2\pi i(ux_0 + vy_0)\right) \tag{5.17}$$

Equation 5.17 is obtained since the phase of the denominator is zero and its magnitude of the imaginery exponential is one. The phase correlation function, which is the normalized cross-correlation function, is obtained by applying the inverse Fourier transform to $R(u, v)$:

$$r = \mathscr{F}^{-1}\{R\} \tag{5.18}$$

The translational shift can be determined as the location of the peak in r:

$$(\Delta x, \Delta y) = \underset{(x,y)}{\operatorname{argmax}}\{r\} \tag{5.19}$$

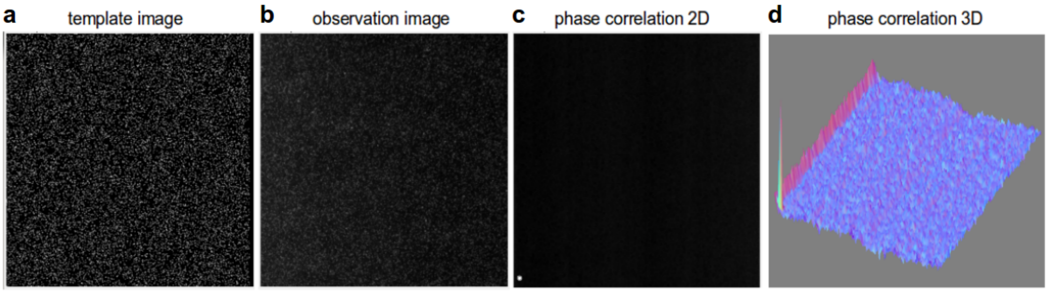


Figure 5.6: Phase correlation of template and observed image. Images created with the FFT filter and 3D Surface Plot of ImageJ (Abràmoff et al. (2004); Schneider et al. (2012)). From left to right: (a) template image containing the reference positions, (b) image taken during protein cycle, (c) 2d phase correlation image with peak in the lower left corner, and (d) as 3d image to highlight the peak.

The pixel based shifts are refined by fitting a Gaussian to the 3 × 3 pixel area around the detected shifts to subpixel resolution. The Levenberg-Marquardt (Levenberg (1944);

Marquardt (1963)) fit is used to interpolate a peak in order to determine its maximum to sub-pixel accuracy.
The fast Fourier transform (FFT) (Cooley and Tukey (1965a)) allows to compute the DFT in $O(nm \log(nm))$ for an image with size $n \times m$, similarly for the inverse Fourier transform. The multiplication of transforms in the frequency domain has a negligible cost of $O(nm)$. The phase correlation method is insensitive to occlusions and brightness change and it is remarkably robust against noise (Kuglin and Hines (1975)).

5.3.4 Implementation

The implementation of the cluster registration using phase correlation as described above has been based on modified code from Illumina's OLB pipeline version 1.9.4 (Inc. (2011a)). The input are the *pos* files produced by the Illumina RTA (Real Time Analysis) pipeline (Inc. (2011c)) which contain the template cluster positions per tile, and the tif images of the T channel for each protein cycle. The output are the x,y coordinates of the cluster positions of the observed images for each protein cycle.

5.3.5 Estimation of scaling

The following approach is based on modified code from Illumina's OLB pipeline version 1.9.4 (Inc. (2011a)). In order to estimate scaling and its role in the cluster registration, each image that contains the measured fluorescent signals and its related template image is divided into four quadrants. For each quadrant a subregion is taken for which the phase correlation is calculated. Figure 5.7 provides an example.

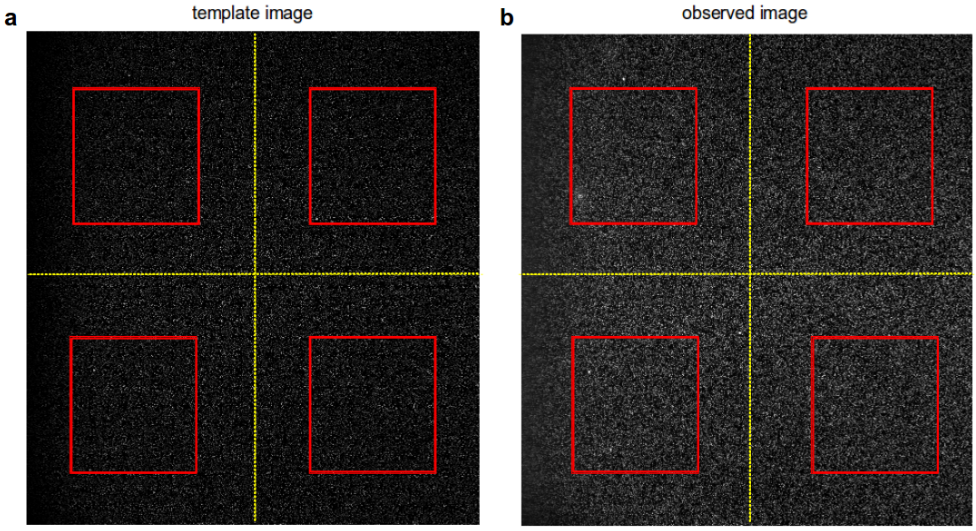


Figure 5.7: Image regions used for calculating the x,y scaling factors. Division of template (a) and observed image (b) into four quadrants (yellow) with subregions (red) used for calculating the scaling factors.

For each pair of correlated subregions between template and observation, shifts in x and y direction are obtained and then used for a linear regression by which x, y offset (intercept) and x, y scaling parameters (slope) are calculated that encapsulate the affine transformation.

Determining x offset and x scaling parameter by linear regression:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} \tag{5.20}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{5.21}$$

where n = 4 for the four subregions, x_i, y_i are the coordinates of the Δx shift, $\hat{\alpha}$ is the x

offset, $\hat{\beta}$ is the x scaling, y offset and y scaling parameter are determined accordingly.

5.3.6 Assessment of transformation parameters

It was tested if further transformation parameters besides scaling like rotation and shearing would be relevant. The Fourier-Mellin transformation (Derrode and Ghorbel (2001)) allows to estimate rotation by utilizing the Mellin transformation and the ECC algorithm (Evangelidis and Psarakis (2008)) uses a nonlinear similarity measure for the image alignment problem. Both methods showed that rotation is not occurring, whereas scaling and shearing are negligible. As shown in Figure 5.8 the translational offset in x and y direction is the dominating factor.

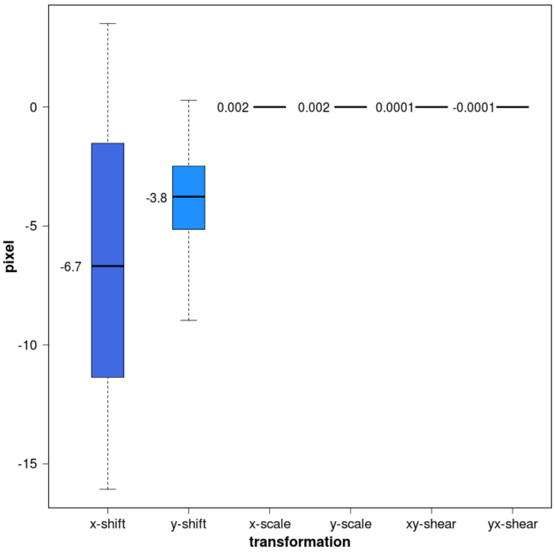


Figure 5.8: Overview of the transformation parameters for the experiment 12.02.2015, cycle 41 (concentration 135 nM) for tiles 1 to 120. Translational offsets were determined by phase correlation, scaling parameters by linear regression, and shearing parameters were estimated by the ECC algorithm (Evangelidis and Psarakis (2008)).

5.3.7 Investigation of overlap with local maxima

In order to quantify the accuracy of the translation parameters for capturing the geometric transformation, I crafted two tests. The first test investigates a necessary condition for the correct mapping of the clusters into the images which is the overlap (or spatial proximity) to local maxima in the image. The second test examines if the right local maxima are matched by the cluster mapping.

Firstly, the local maxima in the observed image are determined for a particular tile, i.e. tile 21, cycle 95 (25 nM), lane 2 of experiment 18.08.2014. Two methods (Neubeck and Van Gool (2006); Schmid (2006)) have been compared which detected largely the same local maxima for the observed image (74.4% of intersecting local maxima).

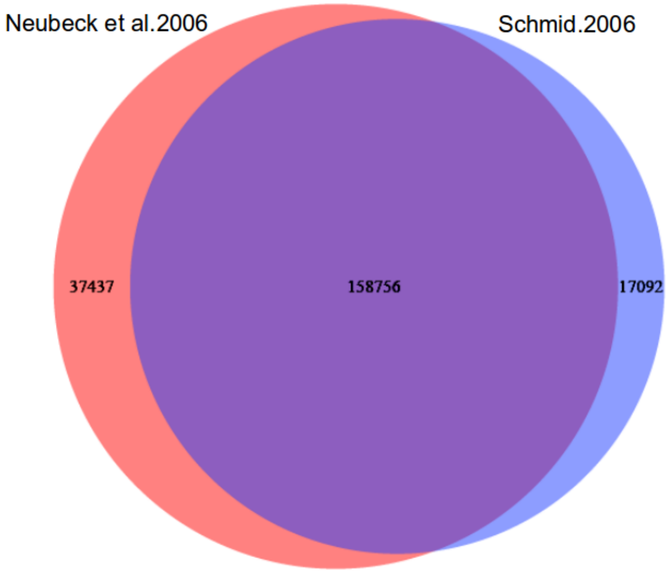


Figure 5.9: Number of detected local maxima and overlap of the two compared methods by (Neubeck and Van Gool (2006); Schmid (2006)).

The parameter settings were as follows:
Algorithm by Schmid (2006):

- Height tolerance: 2.0 pixel, maxima are accepted only if protruding more than this value from the ridge to a higher maximum.
- Threshold: 10.0 pixel, minimum height of a maximum.

Algorithm by Neubeck and Van Gool (2006):

- Minimum distance to other local maximum: 1.0 pixel.
- Threshold: 10.0 pixel, value below which a maximum will be rejected.

Procedure for measuring the overlap of mapped clusters with local maxima:
1) Retrieve all local maxima from the observed image by the methods of (Neubeck and Van Gool (2006); Schmid (2006)).
2) Map all clusters into the observed image by the translational transformation parameters

determined by phase correlation.

3) Divide image into 32 × 32 grid cells, resulting in $\frac{2048}{32} \times \frac{1888}{32} = 64 \times 59 = 3776$ cells.

4) Within each grid cell sort clusters and local maxima by brightness in descending order.

5) Measure Euclidean distance between mapped cluster positions and local maxima.

6) If distance $<$ threshold, the mapping is regarded as correct, else incorrect.

As threshold the median cluster distance is taken per tile which varies between 2.2 and 3.1 pixels among different tiles. Figure 5.10 shows the error map for tile 21, cycle 95 (protein concentration 25 nM), lane 2 from experiment 18.08.2014 using as average cluster distance 2.2 pixels. If the borders are included the percentage mapping error is 2.25%, if the borders are excluded it is only 0.96%.

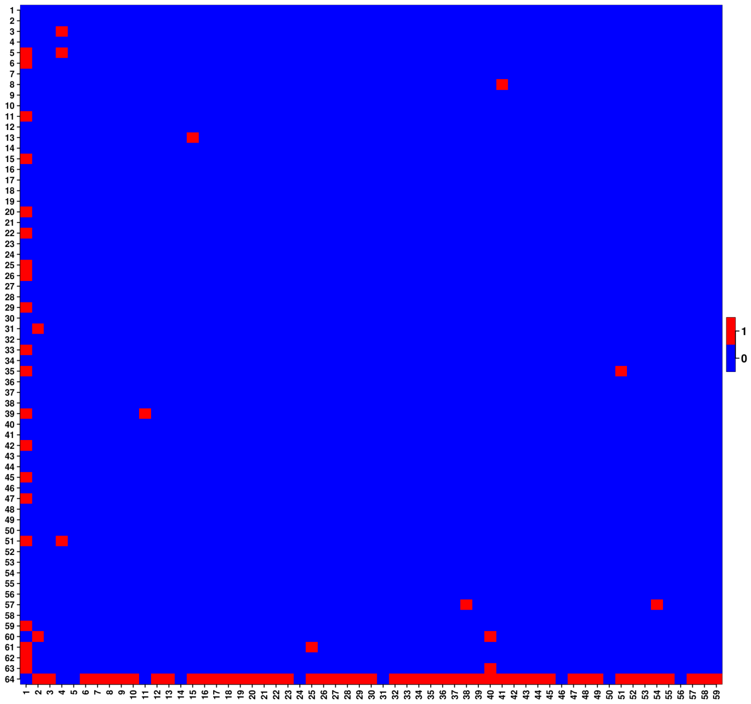


Figure 5.10: Quantification of the cluster registration precision. Mapping errors are displayed for the different 32 × 32 cells of tile 21, cycle 95, lane 2 from experiment 18.08.2014. Red color denotes cells with mapping errors, blue color cells with correct mapping.

5.3.8 Investigation of motif occurrences

Another validation for the translational transformation is if the occurrences of binding motifs adhere to the ranking that would be expected. Figure 5.11 shows the median intensities for different GCN4 motifs on tile 21, cycle 95 (protein concentration 25 nM), lane 2 from experiment 18.08.2014. The median intensity and therefore the binding affinity of GCN4 to DNA decreases the larger the Hamming distance becomes with respect to the consensus motif. If the mapping of the clusters due to the estimated transformation parameters would be biased and adulterated, such a decrease could not be achieved.

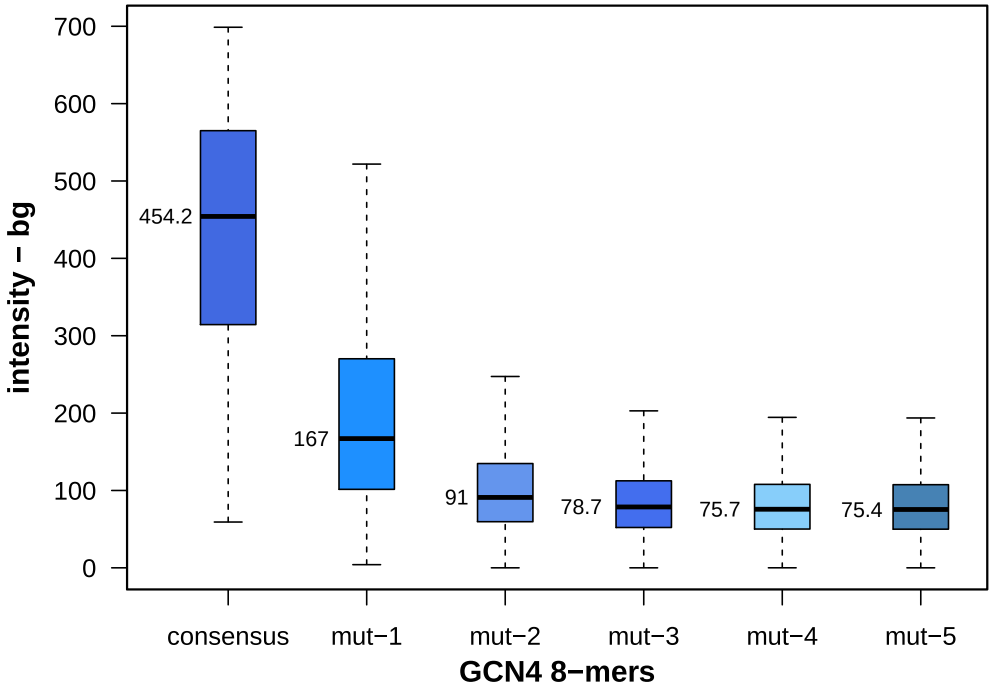


Figure 5.11: Median intensities for different GCN4 motifs after cluster registration. The larger the Hamming distance from the consensus, the smaller the intensities become providing evidence for a correct mapping.

5.3.9 Results

It can be concluded that a translational offset in x and y direction is the dominant transformation that affects observed images due to moving the sledge onto which the flow cell is mounted during imaging. The values of the x and y offsets can be effectively determined by phase correlation allowing to map reference cluster positions onto observed image cluster positions and thus aligning consistently protein intensities at different concentrations with the related DNA sequences.

Since protein images are quite different from sequencing images, the mapping accuracy regarding protein images needs to be assessed. Estimating the mapping accuracy by the spatial proximity of mapped cluster positions to local maxima in the observed protein image, which is a necessary condition for the correctness of the cluster position transformation, yields a mapping error of 2.25%, if borders are included, and of only 0.96%, if borders are excluded as displayed in Figure 5.10. Approaching the mapping accuracy for a protein image by the brightness of GCN4 motifs and their expected binding affinity enables a "semantic“ verification checking if clusters are mapped consistently onto related local maxima in the images. As Figure 5.11 demonstrates the median intensities decrease with increasing Hamming distance as expected given that the binding affinity is lowered by an increased number of mutations. The transformation therefore must provide the right mapping otherwise the decrease of the intensities with increase in Hamming distance could not be observed. In summary then protein images can be registered with high accuracy even though they are different with respect to the fluorescent signals compared to sequencing images. A significant advantage connected with this finding is that resequencing is unnecessary and can be omitted thus allowing the reuse of the flow cell for several HiTS-FLIP experiments, which greatly reduces cost and time.

5.4 Local region search

Some of the cluster positions after being transformed do not overlap with the local maxima positions (representing the DNA cluster positions) identified in the observed images. This can be due to rounding to discrete pixel coordinates that shift cluster positions (most often within an one pixel neighborhood) away from the related local maxima positions. The search space consisting of all the image pixels can be divided into separate, disjunct regions defining the local, non-overlapping neighborhood of the clusters within which clusters can be shifted onto local maxima thereby increasing the accuracy of the intensity extraction as described in section 5.6. The technique that can be utilized to achieve this is called “region labeling” ((Burger and Burge, 2009b, pp. 5–17)), described in the following subsection.

5.4.1 Region labeling

During the “region labeling” process connected components are uniquely labeled based on a given heuristic. Here, the connected components are the cluster positions and their local regions, and the “labeling” technique provides a way to demarcate the local cluster regions from each other. The labeling procedure occurs in the following way:
1) Initialize a 2d matrix representing the imaged tile with 0 as initial values for each cell.
2) Iterate through all cluster positions and for each cluster add 1 to the cluster position itself and its 8-connected neighborhood pixels.

After all clusters have been processed by the labeling procedure the outcome is that the local, disjoint search region of each cluster are defined by the label “1”, and overlaps of search regions are marked by the label “2”. Therefore, pixels with the label “1” can be used to shift clusters onto local maxima.
Figure 5.12 gives an example. The left side (a) shows a subarea of the image of tile 10, cycle 96 (concentration 125 nM), lane 2 of experiment 18.08.2014. Pixels that represent mapped cluster coordinates are framed in green and hatched, separate local maxima are framed in red. There are five clusters of which two are displaced next to the related local maxima. Figure 5.12 (b) displays the local search regions A to E composed of the 8-connected neighborhoods for each of the five clusters where the label ”1“ denotes pixels that are included in the search region and pixels labeled ”2” are excluded. Since the local maxima are included (labeled ”1“) in the related search area, the positions of the two clusters can be shifted.

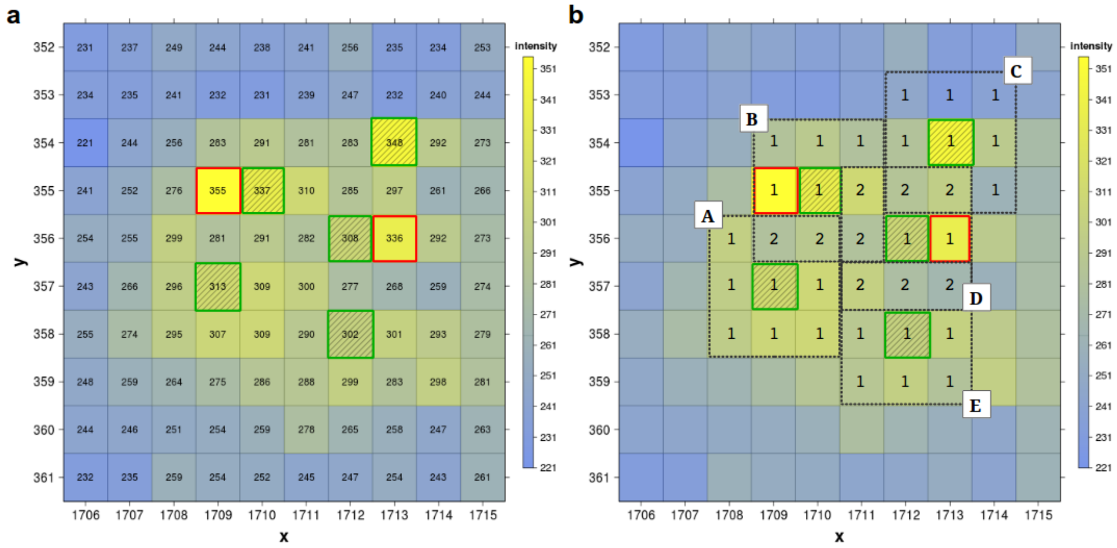


Figure 5.12: Search space of five mapped clusters. (a) Subarea of an imaged tile that shows five clusters (outlined in green and hatched), and local maxima (outlined in red). (b) Clusters with their local search regions A to E and labels "1" denoting included pixel and labels "2" denoting excluded pixels.

5.4.2 Shifting clusters

The process of defining local cluster regions by region labeling and searching these regions for local maxima onto which clusters can be shifted can be executed iteratively. An overview of all iteratively one pixel shifted cluster positions from tile 10, cycle 96 (concentration 125 nM), lane 2 of experiment 18.08.2014 is provided in the Figure 5.13. In total, there are 205337 identified DNA clusters on the tile, in round-1 14.5% of these clusters are shifted by one pixel, in round-2 0.7% of these clusters are shifted by one pixel, in round-3 0.02% of these clusters are shifted by one pixel, and in round-4 0.001% of these clusters are shifted by one pixel, after which all clusters overlap with the local maxima.

5.4.3 Implementation

I developed the region search and cluster shifting using Java (Gosling (2000)) and ImageJ (Abràmoff et al. (2004); Schneider et al. (2012)). The input are the protein images and the x,y coordinates of the mapped clusters. The output are the updated x,y coordinates of the clusters.

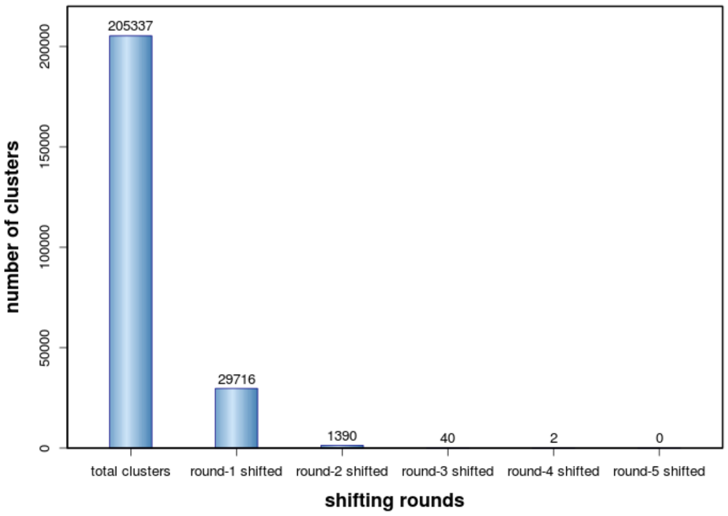


Figure 5.13: Shifting process with different iterations during which cluster positions are overlaid onto local maxima.

5.4.4 Results

A certain portion (10% - 20%) of the mapped cluster positions do not overlap with the local maxima positions (representing the DNA cluster positions) identified in the observed protein images. “Region labeling” is an elegant and effective method for dividing the search space (all image pixels) into separate, disjunct regions defining the local, non-overlapping neighborhood of the clusters marking the area within clusters can be shifted onto local maxima. This overlay process of mapped cluster positions onto local maxima positions can be executed iteratively and within a small number of iterations all cluster positions are adjustable.

5.5 Image normalization

As the main bias which obfuscates intensities and impedes the quantification of the fluorescent signals I identified uneven illumination occurring in the images taken during the protein cycles depending on the spatial positions of the DNA clusters in the flow cell.

5.5.1 Possible causes for non-even illumination

There are various scientific publications that discuss this uneven illumination bias and provide explanations for possible causes.
According to (Waters (2009)) fluorescence emission is generally proportional to the intensity of the illuminating light (except when fluorophore ground state depletion occurs). Therefore, if an uniform fluorescent sample is unevenly illuminated, the resulting fluorescence will usually be uneven as well (Waters (2009)). Uneven illumination can be extremely detrimental to quantitative measurements because it may cause the intensity of an object in one area of the field of view to measure differently than the intensity of an object of equal fluorophore concentration in another area of the field of view (Waters (2009)).
Because of the inherent imperfections of the image formation process, microscopical images are often corrupted by intensity variations manifesting themselves as large area intensity gradients not present in the original scene (Inoué (2013)). This phenomenon is usually referred to as shading, or intensity non-uniformity, or intensity inhomogeneity (Likar and Pernuš (2000)).
This phenomenon can also be named as vignetting, i.e. a brightness attenuation away from the image center often resulting in the outer image edges being significantly darker than the center (Marty et al. (2007); Zheng et al. (2009)).
Uneven illumination may originate from inaccurate object preparation and mounting or from imperfections in the image acquisition process. In the latter case, shading may arise from nonuniform background illumination, departing from Köhler illumination, imperfect, dirty, or dusty optics, uneven spatial sensitivity of the video camera, dark-level camera response, or camera non-linearity (Likar and Pernuš (2000), references therein).
According to (Zheng et al. (2009)) several mechanisms may be responsible for vignetting effects. Some arise from the optical properties of camera lenses, the most prominent of which is off-axis illumination falloff or the cos^4 law (Reiss (1945)). This contribution to vignetting results from foreshortening of the lens when viewed from increasing angles from the optical axis (Klein and Furtak (2013)). Other sources of vignetting are geometric in

nature. For example, light arriving at oblique angles to the optical axis may be partially obstructed by the field stop or lens rim (Zheng et al. (2009)). Leong et al. (2003) states that vignetting may be attributed to multiple factors from the illumination filament, the design of the light path between the camera and the microscope, or the behavior of the imaging device.

5.5.2 Illustration of non-even illumination

Figure 5.14 shows a thumbnail image (produced by the Illumina RTA pipeline (Inc. (2011c))) of tile 1, cycle 46 (concentration 125 nM), T channel, lane 2 from experiment 18.08.2014 with nine selected areas and their magnified view. The uneven illumination is clearly visible.

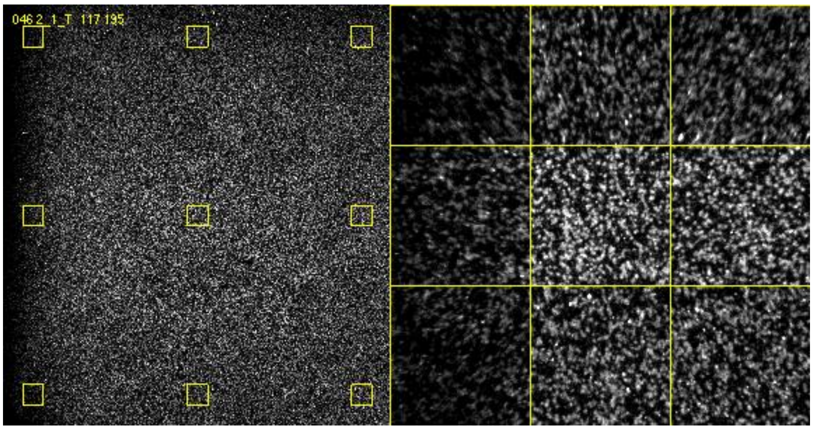


Figure 5.14: Thumbnail image of a tile showing uneven illumination. Image of tile 1, cycle 46 (concentration 125 nM), T channel, lane 2 of experiment 18.08.2014 with magnified subareas.

Figure 5.15 shows the image of tile 10, cycle 46 (concentration 25 nM), T channel, lane 4 from experiment 13.06.2013. Figure 5.15 (a) has been processed using ImageJ (Abràmoff et al. (2004); Schneider et al. (2012)) applying its rolling ball algorithm based on (Sternberg (1983)) with a radius of 40 pixels. This algorithm uses a ball as a structuring element and performs the morphological operation top-hat transform (Dougherty et al. (2003)). The result is an estimate of the local background in different regions of the image. Here it is apparent that there are different patches of varying brightness. Figure 5.15 (b) shows the image of the same tile processed in the following way. The image was divided into 32 × 32 pixel regions, and for each region the mean of the 20 dimmest pixels was calculated representing the background of the region. These background values comprise the displayed intensities.

Figure 5.16 shows on the left the gray level tif image of tile 10, cycle 46 (concentration 25 nM), lane 4 from experiment 13.06.2013, with three vertical selection lines (in yellow), and on the right the related intensity profiles for these selections regarding the region background. The regions and related backgrounds are calculated such that the image is divided into 32 × 32 pixel regions and for each region the background is determined as the mean of the 20 dimmest pixels. Difference in brightness as well as intensity drop off at the borders is eminent.

Figure 5.17 correlates the intensity of spike-in clusters, i.e. DNA clusters with the exact same insert sequence TGCAGGAATGACTCATTGAAGGTTAGATCGGAAGAG, with the related local background, calculated as mean of the dimmest 10 pixel of a 17 × 17 pixel window around the spike-in cluster, for the different concentrations of experiment 13.06.2013 on lane 4. During all protein cycles a strong correlation is observable.

Figure 5.18 correlates the local background, calculated as mean of the dimmest 10 pixel of a 17 × 17 pixel window around the spike-in cluster, of the spike-in clusters across the different protein cycles of experiment 13.06.2013 on lane 4. During all protein cycles a very strong correlation is displayed showing that the non-even illumination effect is stationary across imaging cycles.

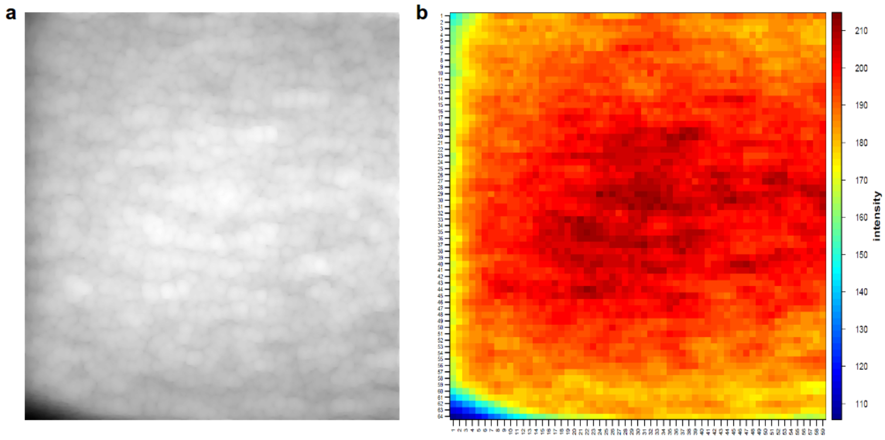


Figure 5.15: Uneven illumination. (a) Image processed with the rolling ball algorithm (Sternberg (1983)) with a radius of 40 pixels. (b) Same imaged tile as in (a) as heat map depicting the background intensities of 32×32 pixel regions. For each region the mean of the 20 dimmest pixels was calculated representing the background of the region.

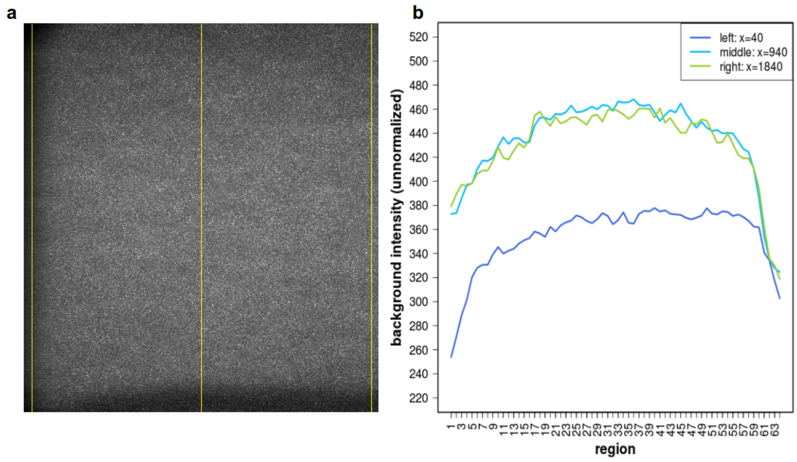


Figure 5.16: Intensity profile of region background for a representative image displaying uneven illumination. (a) Gray level tif image of tile 10, cycle 46 (concentration 25 nM), lane 4 from experiment 13.06.2013 with three vertical selection lines marked in yellow. The x coordinates are 40, 940 and 1840. (b) Intensity profiles for the three selection lines regarding the region background. The y axis shows the unnormalized intensity of the region background, the x axis shows the different region (32×32 pixels) along the y direction of the selection lines. For each region the mean of the 20 dimmest pixels was calculated representing the background of the region.

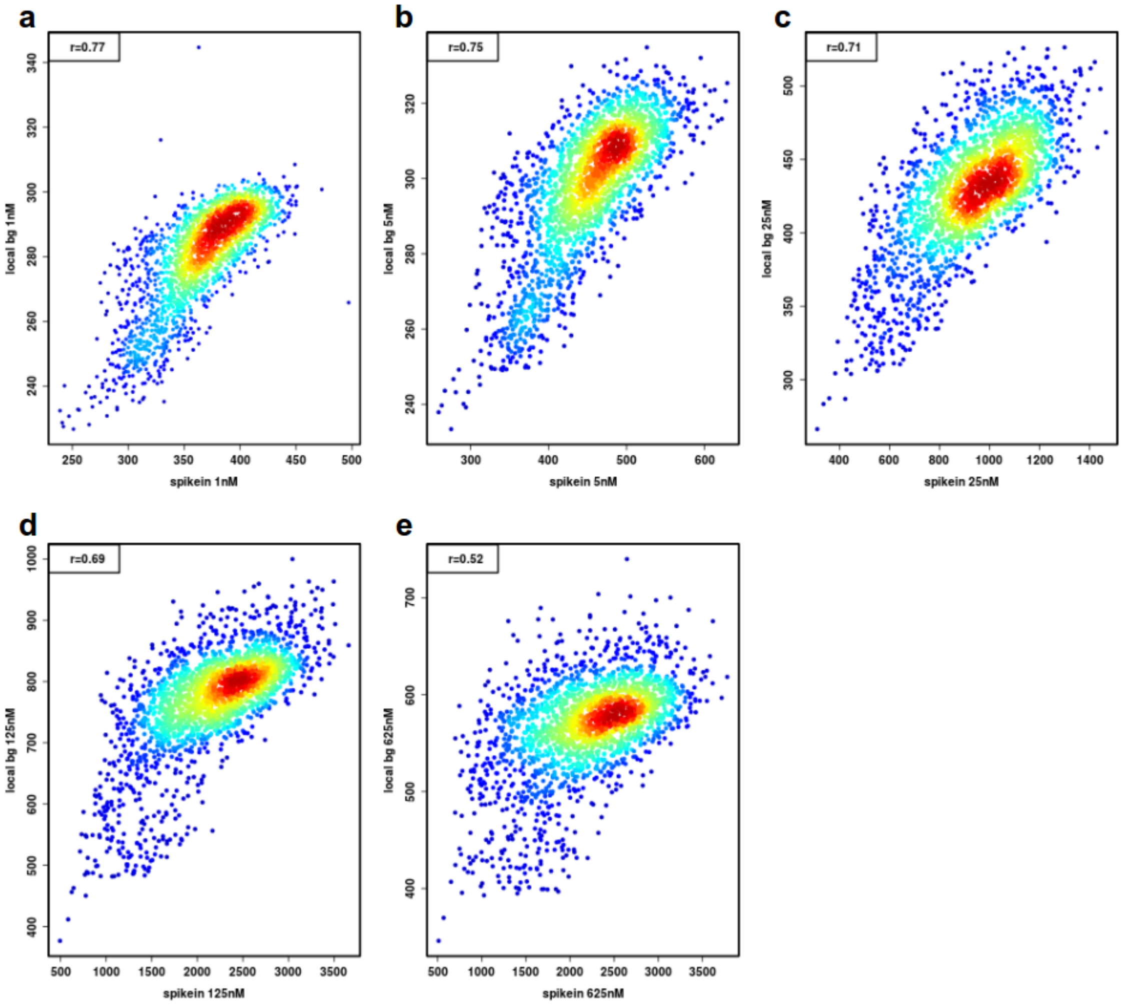


Figure 5.17: Correlation of spike-in cluster intensity with local background intensity of experiment 13.06.2013 on lane 4. (a) to (e) denote the different concentrations. The local background is calculated as the mean of the dimmest 10 pixel of a 17 × 17 pixel window around the spike-in cluster.

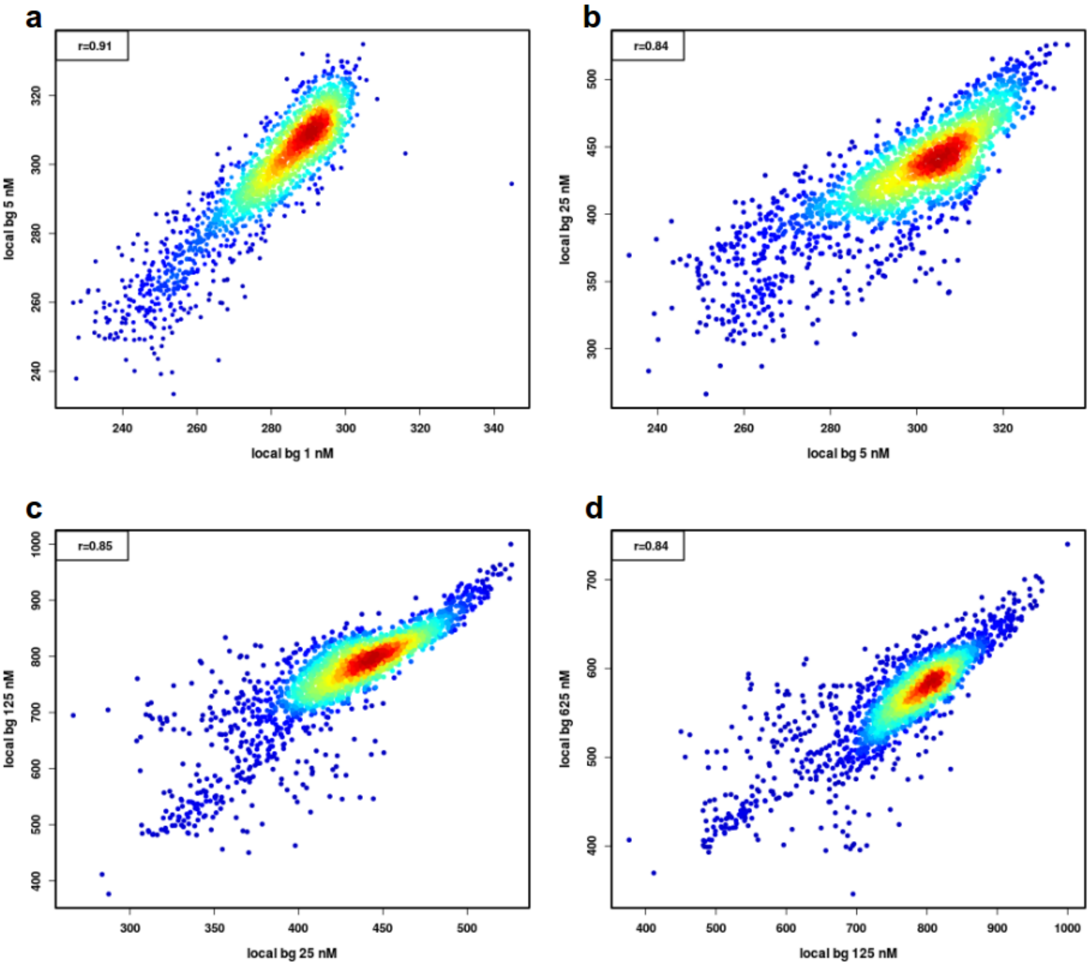


Figure 5.18: Correlation of the local background intensity of the spike-in clusters across the different protein cycles of experiment 13.06.2013 on lane 4. (a) 1 nM compared to 5 nM. (b) 5 nM compared to 25 nM. (c) 25 nM compared to 125 nM. (d) 125 nM compared to 625 nM. The local background is calculated as the mean of the dimmest 10 pixel of a 17 × 17 pixel window around the spike-in cluster.

5.5.3 Methods for non-even illumination correction

There are several different approaches to correct for non-even illumination. Correction methods can be prospective when a calibration protocol and extra images are acquired, or retrospective when the only data available is the image itself (Reyes-Aldasoro (2009)). Since a flat-field image which captures the background without any foreground objects cannot be taken during a HiTS-FLIP run, retrospective correction is required.

There are different retrospective methods. Most existing bias correction methods assume that the bias field is multiplicative, slowly varying, and tissue independent (Kubecka et al. (2010)).

The first class of correction algorithms apply filtering with low pass, homomorphic or morphological operators as it is a simple and intuitive way of removing low frequency shading components (Reyes-Aldasoro (2009)).

A second class of algorithms use surface fitting methods (Hou et al. (2006); Russ (2011)) requiring the selection of a number of points on the background, either manually or automatically, and the background is obtained by the fitting of a parametric surface (Kubecka et al. (2010)). The polynomial fit method is based on the assumption that the variation of the intensity of the background image can be obtained by the fitting of a polynomial function to the intensity values of a number of points selected in the background of the image (Tomazevic et al. (2002)). It approximates an image by a polynomial and uses the orthogonality relation of the Legendre polynomials to expand an image as a double sum of those functions. The sum is then evaluated to produce an image that approximates a projection onto the space of polynomial images (Babaloukas et al. (2011)).

A third class of algorithms perform entropy minimisation (Likar et al. (2000); Vovk et al. (2006)) as it is assumed that the shading introduces extra information to the image, which manifests itself as a higher entropy. For example, in (Likar et al. (2000)) a parametric polynomial surface that minimises the entropy is assumed to be the shading component.

5.5.4 Linear model of the image formation

A widespread linear model of the image formation (Beckers et al. (1994); Leahy et al. (2012); Likar et al. (2000)) which describes the relation between the true image $U(x, y)$ and the acquired image $N(x, y)$ is the following:

$$N(x, y) = U(x, y)S_m(x, y) + S_a(x, y) \tag{5.22}$$

$N(x, y)$: acquired, intensity non-uniform image.
$U(x, y)$: true image.
$S_m(x, y)$: multiplicative shading component.
$S_a(x, y)$: additive shading component.

Shading correction is concerned with finding the corrected image $\hat{U}(x, y)$ which optimally estimates the true image $U(x, y)$ from the acquired image $N(x, y)$ (Likar et al. (2000)):

$$N(x, y) \xrightarrow{shading\,correction} \hat{U}(x, y) \approx U(x, y) \tag{5.23}$$

The shading corrected image $\hat{U}(x, y)$ can easily be calculated by inverting the image formation model:

$$\hat{U}(x, y) = \frac{N(x, y) - \hat{S}_A(x, y)}{\hat{S}_M(x, y)} \tag{5.24}$$

where $\hat{S}_A(x, y)$ and $\hat{S}_M(x, y)$ are estimates of the additive and multiplicative shading component. The problem of shading correction can thus be viewed as the problem of estimating the additive and multiplicative shading components (Likar et al. (2000)).

5.5.5 Estimation of the additive shading component

In order to estimate retrospectively the appropriate additive shading component $S_a(x, y)$ a pixel window around the cluster was taken and the local background was determined as the mean of the dimmest 5% pixels of this pixel window. In the following analyses, the pixel window size is scrutinized. The left side of Figure 5.19 shows the local background intensity for different pixel window sizes, calculated as the mean of the dimmest 5% pixels within the pixel window, taken over all clusters from tile 10, cycles 93 to 97, lane 2 from experiment 18.08.2014. The distribution of the clusters is represented as density. The right side of the Figure 5.19 shows the local background intensity distribution as box plots for the different window sizes.

For the following analysis a 17×17 pixel window was chosen. A 17×17 pixel window consists of 289 pixels, and given the experimental setup of ca. 200000 DNA clusters per tile the average number of clusters in a 17×17 pixel region is 15. A DNA cluster can be up to 9 pixels (sometimes even 10-12 pixels) in size and thus there are then 135 foreground and 154 background pixels in a 17×17 pixel region. The number of the dimmest pixels within a 17×17 pixel window used for calculating the local background

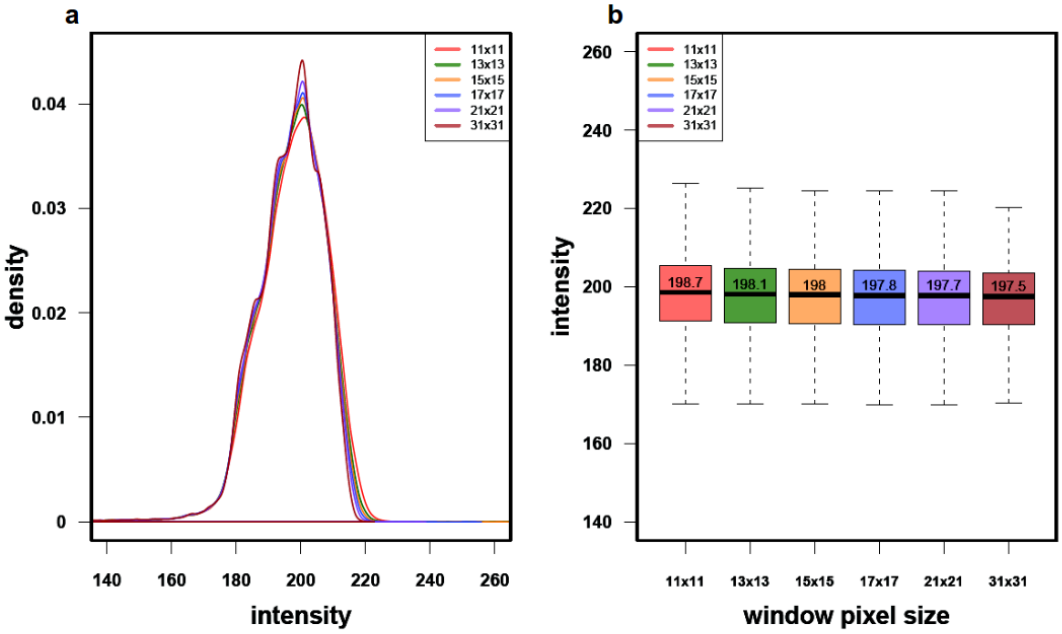


Figure 5.19: Different window sizes and related background intensities. (a) Local background intensity for different pixel window sizes, calculated as the mean of the dimmest 5% pixels within the pixel window. (b) Local background intensity distribution as box plots for the different window sizes.

was analyzed and the result is shown in the Figure 5.20. The local background was processed over all clusters from tile 10, cycles 93 to 97, lane 2 from experiment 18.08.2014. Taking 5, 10 or 20 dimmest pixels within the 17 × 17 pixel window does not lead to significantly different values.

There are hundreds of occurrences of a particular 7-mer on a tile during a HiTS-FLIP experiment. Figure 5.21 shows on the left the occurrences (yellow dots) of the 7-mer TGACTCA (reverse complement TGAGTCA) on the tile 10, cycle 95 (concentration 25 nM), T channel, lane 2 of experiment 18.08.2014. On the right the occurrences of all the first 20 ranked (see section 5.9 for details regarding the ranking method) 7-mers is displayed.

Since the occurrences of k-mer motifs can be employed for analyzing the effects of the normalization by subtraction of local background (mean of 5% dimmest pixels from pixel window around cluster), the first 20 ranked 7-mers and their related unnormalized and normalized intensities were used for measuring the variance and Kruskal Wallis statistics. The Kruskal–Wallis one-way analysis of variance by ranks (Kruskal and Wallis

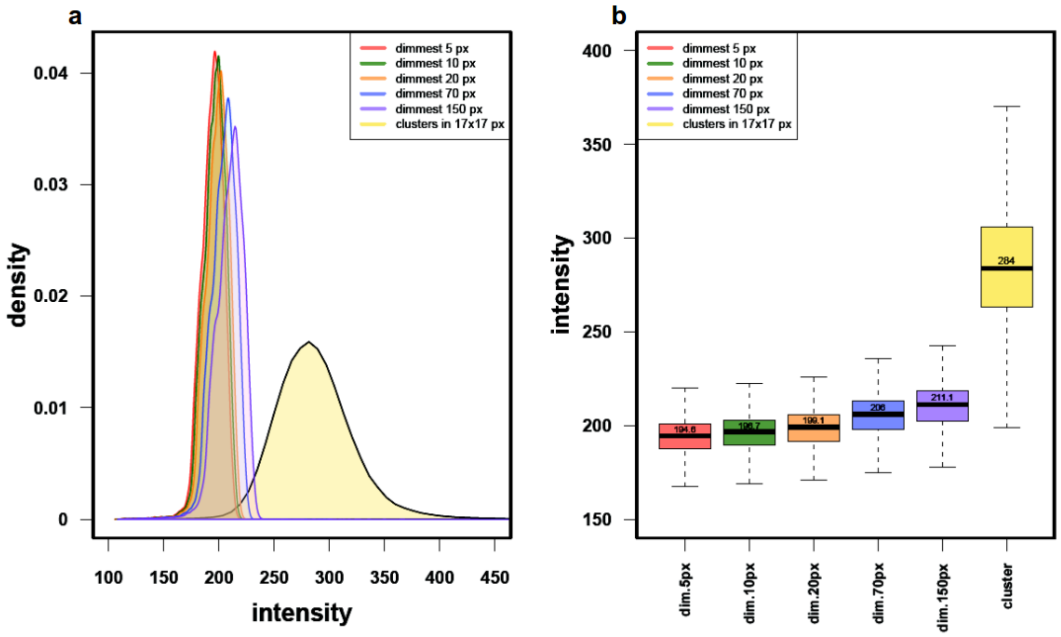


Figure 5.20: Different number of dimmest pixels and related background intensities, with the addition of the cluster intensity. (a) Local background intensities for different numbers of dimmest pixels of a 17 × 17 pixel window, with the addition of the cluster intensity. (b) Same as (b) but visualized as a box plot.

(1952)) was used as test statistics since the distribution of the intensities of a k-mer is not normally distributed, and the variances and sample sizes among the k-mers are different. Figure 5.22 shows for different pixel window sizes the variance, the Kruskal Wallis test statistics K and the related p-values for the first 20 ranked 7-mers on the tile 10, cycle 95 (concentration 25 nM), T channel, lane 2 of experiment 18.08.2014. Since the p-values turn out to be nearly zero, at a 0.05 significance level the null hypothesis that the different 7-mers with their intensities are identical can be rejected. Thus, the 7-mers are different and the larger the Kruskal Wallis test statistics K is the more different the 7-mers are from one another. The largest value (513.86) occurs for the window size of 15 pixels.

Figure 5.23 shows the outcome when calculating the Kruskal Wallis test statistics K for the first 20 ranked 7-mers using tiles 1 to 120, T channel, lane 2 of experiment 18.08.2014 for the concentration 5 nM, 25 nM, 125 nM, and 625 nM and selecting the size of the window for which K is maximal. Since 15 pixel as window size is the distinguished value,

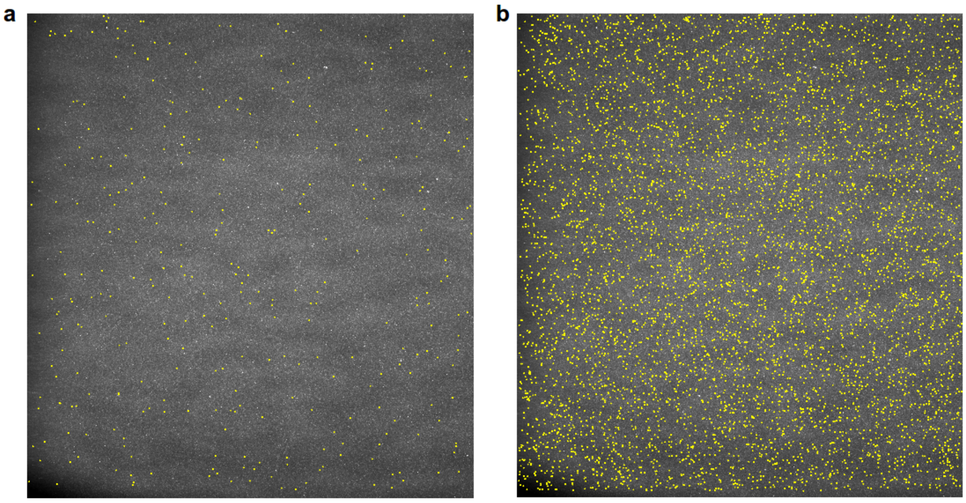


Figure 5.21: Occurrence and distribution of 7-mers on the tile 10, cycle 95 (concentration 25 nM), T channel, lane 2 of experiment 18.08.2014, shown as yellow dots. (a) Occurrences of 7-mer TGACTCA and its reverse complement TGAGTCA. (b) Occurrences of all the first 20 ranked 7-mers and the related reverse complements.

the window size for calculating the local background around a cluster was chosen to be 15 × 15 pixels and the local cluster background was calculated as the mean of the 5% dimmest pixels from these 15 × 15 pixels.

5.5.5.1 Implementation

I developed the calculation of the local cluster background using Java (Gosling (2000)) and ImageJ (Abràmoff et al. (2004); Schneider et al. (2012)). The input are the protein images and the x,y coordinates of the clusters. The output are the local background intensity values for each cluster, calculated as the 15 × 15 pixel window around the cluster and the mean of the 5% dimmest pixels from this 15 × 15 pixel window.

5.5.5.2 Results

Since the largest Kruskal Wallis test statistics K value occurs for the window size of 15 pixels, the size of the pixel window around a cluster was chosen as 15 × 15 pixel window and the local background intensity value as the mean of the 5% dimmest pixels from this 15 × 15 pixel window.

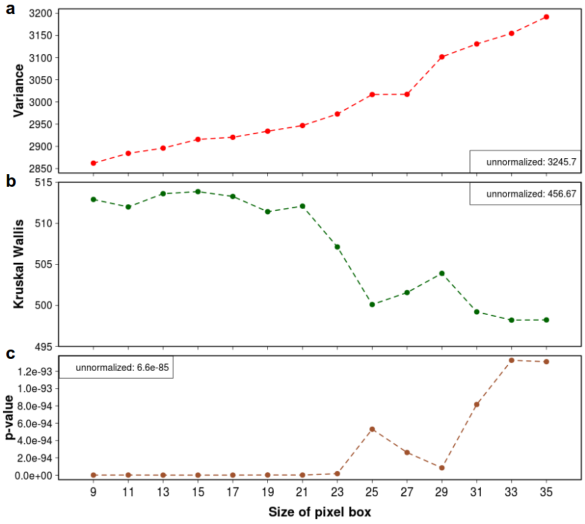


Figure 5.22: Different measures for assessing the normalization by local background subtraction using the first 20 ranked 7-mers on the tile 10, cycle 95 (concentration 25 nM), T channel, lane 2 of experiment 18.08.2014. (a) Variance of the different 7-mers and their intensities. (b) Kruskal Wallis test statistics K values. (c) p-values for Kruskal–Wallis one-way analysis of variance by ranks (Kruskal and Wallis (1952)).

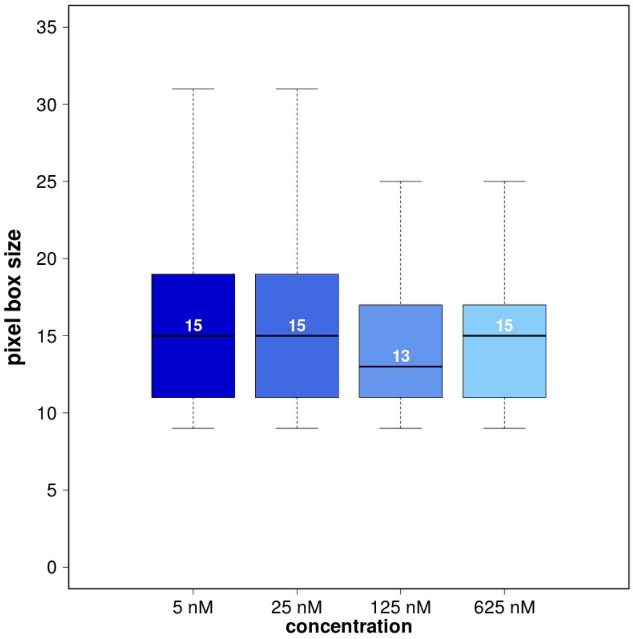


Figure 5.23: Pixel size of windows at different concentrations for which the maximal Kruskal Wallis test statistics K was used for all 120 tiles of lane 2 and the first 20 ranked 7-mers.

5.5.6 Estimation of the multiplicative shading component

For estimating retrospectively the multiplicative shading component $S_m(x, y)$ various different techniques can be applied such as Gaussian filtering (Babaloukas et al. (2011); Leong et al. (2003)), homomorphic filtering (Delac et al. (2006); Etemadnia and Asharif (2004); Wen-Cheng and Xiao-Jun (2013)), morphological operators (Babaloukas et al. (2011); Michálek et al. (2010); Wang et al. (2014)), anisotropic diffusion (Black et al. (1998); Hama and Al-Ani (2013); Liu (2013); Tschumperle and Deriche (2005)), surface fitting by higher-order polynomial for approximating the background (Zhang et al. (2014)), and entropy (as a measure of global intensity uniformity) minimization based methods (Likar et al. (2000)).

I used Gaussian filtering here as a linear, low pass filter based upon the assumption that the uneven illumination is a low frequency signal. Therefore, low pass filtering can be used to extract it from an image. The objects of interest, i.e. DNA clusters, are smaller than the variation of the background and the background has a different intensity than the clusters. Blurring the image with a Gaussian filter including background as well as cluster pixels is based upon the known fact that all intensity measurements are a mixture of signal and background (Waters and Swedlow (2007)). The resulting smoothed image is considered an estimate of the background of the image (Babaloukas et al. (2011); Leong et al. (2003)).

This filtering can be achieved by convolving the image $I(x, y)$ with a Gaussian kernel. The Gaussian function $G(x, y)$ is defined by:

$$G_\sigma(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \tag{5.25}$$

where σ defines the effective spread of the function. The effect of this function is to delimit the spatial frequencies in an image, resulting in loss of edge definition and averaging of intensity values. The larger the value of the parameter σ, the greater the smoothing effect. The aim is to smooth the image until it is devoid of cluster features but retains the weighted average intensity across the image corresponding to the underlying illumination pattern.

5.5.6.1 Implementation

I developed the Gaussian filtering using Java (Gosling (2000)) and ImageJ (Abràmoff et al. (2004); Schneider et al. (2012)), using the GaussianBlur component of the ImageJ framework. The input are the unnormalized protein images and the output are the Gaussian based normalized protein images.

5.5.6.2 Assessment of different σ values

Figure 5.24 shows for different σ values the variance, the entropy of the background, the Kruskal Wallis test statistics K and the related p-values for the first 20 ranked ((see section 5.9 for details regarding the ranking method)) 7-mers on the tile 10, cycle 95 (concentration 25 nM), T channel, lane 2 of experiment 18.08.2014. The background pixels of the image were determined by dividing the image into 32 × 32 pixel regions and taking the mean of the dimmest 20 pixels. The entropy H of the image is based on Shannon's entropy (Shannon (2001)) and was calculated for these background pixels in the following way:

$$H = -\sum_{k=0}^{M-1} p_k \log_2(p_k) \tag{5.26}$$

where
M: is the number of gray levels in the image.
$p_k = \dfrac{n_k}{M \times N}$: is the probability associated with the gray level k with n_k being the number of pixels with grayscale k and $M \times N$ the size of the image.

The various pixels in an image may be considered to be symbols produced by a discrete information source with the gray level as its states and the entropy is a measure of their information content. High entropy images have a great deal of contrast from one pixel to the next whereas a uniform distribution of gray levels results in a low entropy. The lower the entropy in Figure 5.24 the more uniform the intensity of the background pixels. The Kruskal Wallis test statistics K value is maximal (566.21) for a σ of 33 pixel which is still linked to a low entropy value (3.021). The lowest entropy value is except for $\sigma = 1$ occurring for $\sigma = 14$ (2.953).

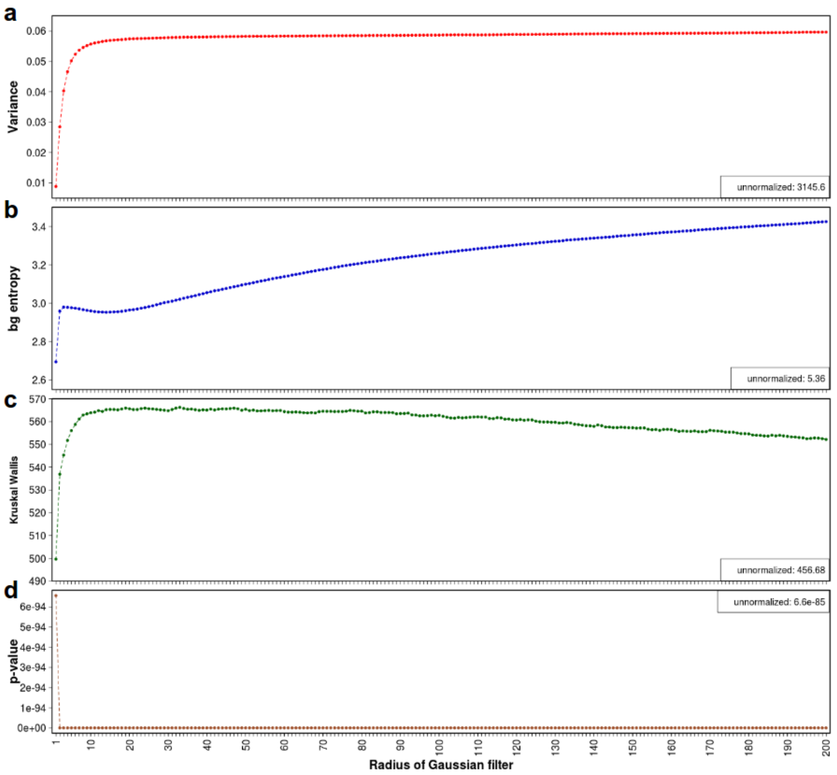


Figure 5.24: Different measures for assessing the normalization by division of Gaussian filtered image with different σ (radius) using the first 20 ranked 7-mers on the tile 10, cycle 95 (concentration 25 nM), T channel, lane 2 of experiment 18.08.2014. (a) Variance of the different 7-mers and their intensities. (b) Entropy of background pixels. (c) Kruskal Wallis test statistics K values. (d) p-values Kruskal–Wallis one-way analysis of variance by ranks (Kruskal and Wallis (1952)).

Figure 5.25 shows the outcome when calculating the Kruskal Wallis test statistics K for the first 20 ranked 7-mers using tiles 1 to 120, T channel, lane 2 of experiment 18.08.2014 for the concentration 5 nM, 25 nM, 125 nM, and 625 nM and selecting the radius of the Gaussian filter kernel for which K is maximal. For the subsequent processing in the pipeline a σ value of 30 pixel was chosen since this is in the range of the values determined by the maximal Kruskal Wallis test statistics K value, and it is around 10 times bigger than the size of a DNA cluster which is on average 3 pixels in width.

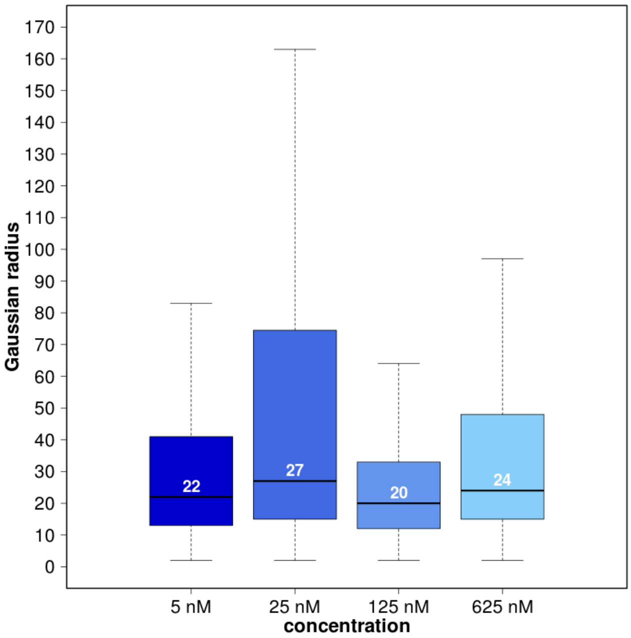


Figure 5.25: Different Gaussian filter radius values (σ) at different concentrations for which the maximal Kruskal Wallis test statistics K was used for all 120 tiles of lane 2 and the first 20 ranked 7-mers.

Figure 5.26 illustrates the outcome when both additive and multiplicative shading correction is applied. The additive shading correction is the subtraction of the local background intensity from the related cluster intensity calculated as a 15 × 15 pixel window around the cluster and taking the mean of the dimmest 5% pixels. The multiplicative shading correction is the division of a Gaussian filtered image with the different radius values as shown in the Figure 5.26.

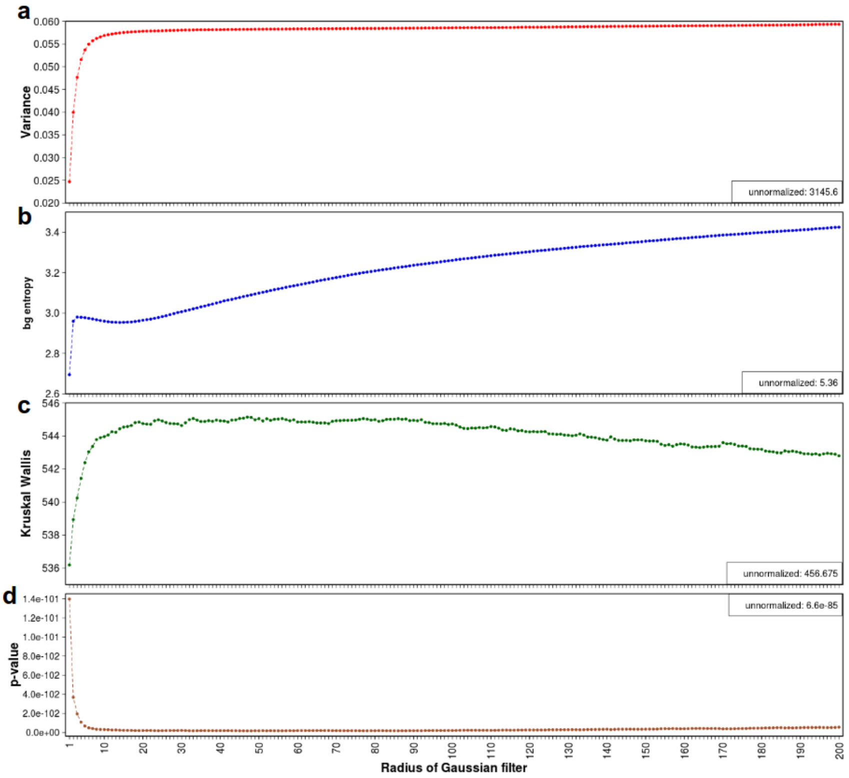


Figure 5.26: Different measures for assessing the normalization by subtraction of local background and division of Gaussian filtered image with different σ (radius) using the first 20 ranked 7-mers on the tile 10, cycle 95 (concentration 25 nM), T channel, lane 2 of experiment 18.08.2014. (a) Variance of the different 7-mers and their intensities. (b) Entropy of background pixels, only for the multiplicative shading correction. (c) Kruskal Wallis test statistics K values. (d) p-values Kruskal–Wallis one-way analysis of variance by ranks (Kruskal and Wallis (1952)).

5.5.6.3 Weighting factors

In addition to the Gaussian filtering, a weighting factor has been applied such that for increasing concentration levels the related cluster intensities increase as well. Using only the Gaussian filtering for the normalization leads to intensities on the same intensity level across the different concentrations. As an estimate for the increasing amount of protein linked with each consecutive concentration level, the unbound proteins in the flow cell have been used. The amount of unbound proteins can be determined by the global background intensity across an entire protein image, calculated in the following way:
1) Divide each protein image into 32×32 grid cells, resulting in $\frac{2048}{32} \times \frac{1888}{32} = 64 \times 59 = 3776$ cells.
2) For each cell, take the mean of the dimmest 20 pixels as the local background intensity of the cell.
3) Take the median of all these cell backgrounds as the global background intensity of the related protein image.
The global background intensity of a protein image at a higher concentration has been put in relation to the global background intensity of this protein image at the lowest concentration since the increase in protein amount relative to the starting concentration is required. More formally, the normalization with weighting can be expressed in the following way:

$$I_{C_i}^{norm} = \frac{I_{C_i}^{unnorm}}{\hat{I}_{C_i}^{mult}} \quad \text{for } i = 1 \tag{5.27}$$

$$I_{C_i}^{norm} = \frac{I_{C_i}^{unnorm}}{\hat{I}_{C_i}^{mult}} \times \frac{B_i^{global}}{B_1^{global}} \quad \text{for } i = 2..n \tag{5.28}$$

where
$I_{C_i}^{norm}$: is the normalized cluster intensity at concentration i for tile t.
$I_{C_i}^{unnorm}$: is the unnormalized cluster intensity at concentration i for tile t.
B_i^{global}: is the global background intensity at concentration i for tile t.
B_1^{global}: is the global background intensity at the lowest concentration ($i = 1$) for tile t.

5.5.6.4 Results

$\sigma = 30$ pixels was chosen as value for smoothing the Gaussian filtered image. The weighting factors were determined as described above.

5.5.7 Comparison of additive and multiplicative shading correction

Applying the Kruskal Wallis test statistics K values as the underlying measure the quality of the different normalization approaches can be compared. Table 5.1 summarizes the findings. The first 20 ranked 7-mers on the tile 10, cycle 95 (concentration 25 nM), T channel, lane 2 of experiment 18.08.2014 were used.

method	*max K value*
subtraction of local background (15 × 15, 5% dimmest)	513.86
division by Gaussian filtered image ($\sigma = 30$)	566.21

Table 5.1: Overview of the maximal Kruskal Wallis test statistics K values for the different normalization methods.

Since the highest Kruskal Wallis test statistics K value was achieved by the division by the Gaussian filtered image, this method was applied as normalization for the HiTS-FLIP pipeline.
An explanation why this is suitable here is given by (Russ (2011)). If the image acquisition device is logarithmic (such as film), then subtraction of the background image point by point from each acquired image is correct. If the camera or sensor is linear (CCD have a linear photometric response, Mullikin et al. (1994)), then the correct procedure is to divide the acquired image by the background. The difference is easily understood because in the linear case the corrected result is *Image / Background*, and when the signals are logarithmic, the division is accomplished by subtraction: *Log(Image) - Log(Background)* (Russ (2011)).

Figure 5.27 shows the result of applying the division of the Gaussian smoothed image regarding the tif image of tile 10, cycle 46 (concentration 25 nM), lane 4 from experiment 13.06.2013. The left side shows the intensity profile for three representative vertical lines at $x = 40$ pixel, $x = 940$ pixel, and $x = 1840$ pixel for the unnormalized intensities, and the right side shows the intensity profile after normalization by the Gaussian filter. The difference in intensities coming from non-even illumination is drastically reduced.

The additive and the multiplicative shading correction can also be compared regarding the Kd based correlation with the HiP-FA Kds (see section 5.9.1.7) and the associated relative errors. The additive shading correction consists of the subtraction of the local cluster background, which is calculated as the mean of the 5% dimmest pixels from a

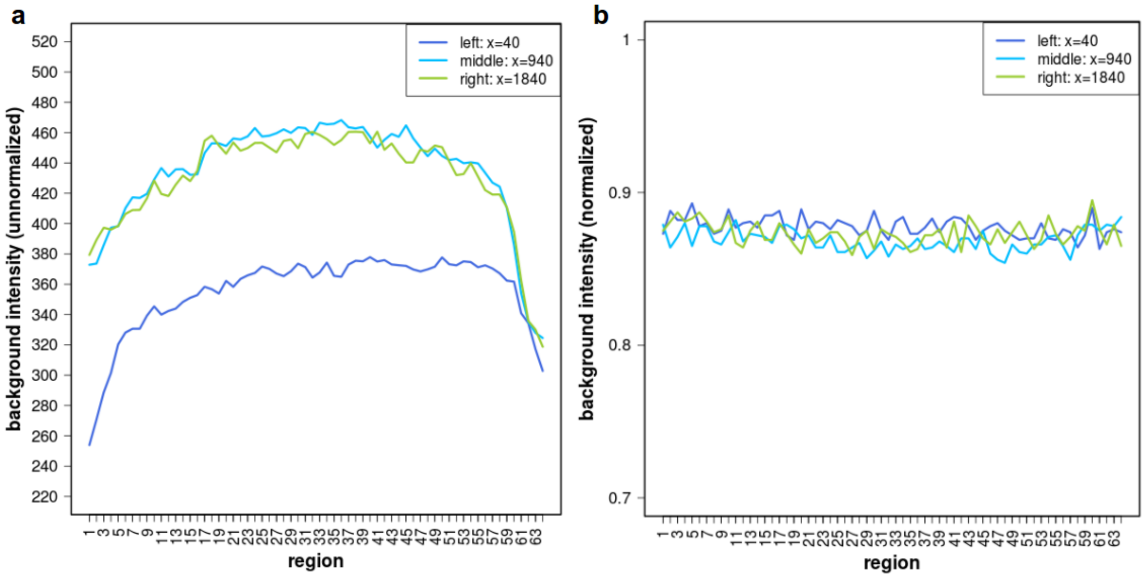


Figure 5.27: Intensity profile for three representative vertical lines at $x = 40$ pixel, $x = 940$ pixel, and $x = 1840$ pixel for the unnormalized and normalized intensities. (a) Unnormalized intensities. (b) Normalized intensities.

15 × 15 pixel box around each DNA cluster. The multiplicative shading correction is made up by the division of the original image by itself, smoothed by a Gaussian filter with $\sigma = 30$ pixels and the weighting factors (5.5.6.3). The intensity extraction method was applied as described in section 5.6. As ranking method the heuristic based ranking was used as described in subsection 5.9.1 but without discarding clusters that contained ranked motifs. The fitting was done as described in section 5.10. As shown in Table 5.2, the correction of the additive and the multiplicative shading component yield similar results.

method	R	δ
subtraction of local background (15 × 15, 5% dimmest)	0.99	50.48%
division by Gaussian filtered image ($\sigma = 30$)	0.99	30.91%

Table 5.2: Comparison of the additive and the multiplicative shading correction with respect to the correlation of Kds and relative errors as measured by HiP-FA as described in section 5.9.1.7. Normalization is carried out by the division of the original image by itself, smoothed by a Gaussian filter with $\sigma = 30$ pixels.

5.5.8 Comparison with sequencing image based normalization

The normalization of the DNA cluster intensities of the protein images by Nutiu et al. (2011) utilizes the fluorescent signals from the sequencing images for correcting the effect of cluster size and of cluster tile position on the variability of the cluster intensities. The normalization factor *nf* of a certain cluster has the following form:

$$nf = \frac{1}{n}\sum_{i=1}^{n} \max(A, C, G, T) \tag{5.29}$$

where
n: is the number of sequence cycles.
A, C, G, T: are the fluorescent signals from the four nucleotides that are incorporated during each sequencing cycle.

For each sequencing cycle the brightest fluorescent signal from the four channels is used. The normalization factor *nf* is then taken as divisor of the related cluster intensity for all the measured protein cycles.

Normalizing for cluster size is unnecessary since a cluster would only be brighter if the density of the DNA template strands during amplification would vary. However, solid phase amplification is a very uniform process leading to clusters with very similar density of DNA template strands in the center of a DNA colony (Mercier and Slater (2005); Mercier et al. (2003)). Starting with the primers they form a very dense and uniform carpet on the surfaces of the flow cell. Solid phase DNA amplification leads in three distinct steps (annealing, extension, and denaturation) to the growth of a colony of molecules attached to the surface and located in the same region. In (Mercier et al. (2003)) a Monte Carlo lattice model was used to study solid phase amplification. In a follow-up examination, (Mercier and Slater (2005)) applied Brownian dynamics and came to similar conclusions. According to (Mercier and Slater (2005); Mercier et al. (2003)) the density at the center of the colony can be expected to be somewhat higher than at the fringe. When a molecule is completely surrounded by others, its free end tends to move away from the surface (like in a dense polymer brush). Therefore, after a few cycles, a molecule at the center of the colony (which is thus surrounded by others) will have a smaller duplication probability (its free end is less likely to find a matching primer on the surface). Because of this phenomenon, a DNA colony can be characterized by a roughly constant density and grows outwards, i.e., from its perimeter. Since the intensity extraction methods described in section 5.6 make use of the central cluster pixel, uniform DNA template density can be expected for the central cluster pixel across different motif-containing clusters.

A particular issue concerning the use of the fluorescent signals from the four nucleotides is that their brightness varies quite drastically as shown in Figure 5.28. For example, the signal in the G channel is almost four times as bright as the signal in the C channel, and the signal in the T channel is nearly twice as large as the signal in the C channel. Therefore, the normalization using the four different channels is biased depending on the nucleotide specific composition of the DNA cluster sequence.

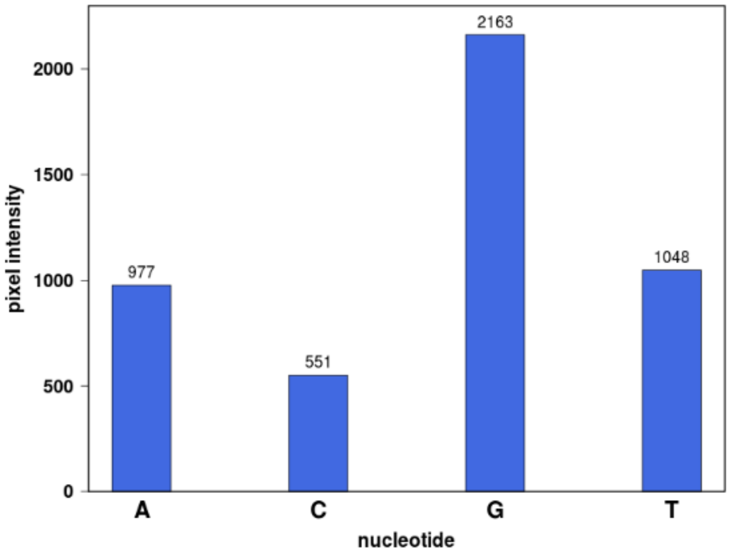


Figure 5.28: Difference of fluorescent intensity from the four nucleotides. The spike-in clusters from experiment 13.06.2013 were used and the median intensity for each channel during the sequence cycles is shown here.

In order to evaluate how efficient the sequence image based normalization by Nutiu et al. (2011) is for correcting the non-even illumination bias depending on the cluster tile position Kds have been calculated and compared with the Kds from an alternative assay. For details on the HiP-FA assay see section 5.9.1.7. The underlying data set for the analyses in this section are the data from lane 2 of experiment 18.08.2014. Preprocessing of the tif images was done as described in section 5.2, the cluster position transformation was carried out as described in section 5.3, and the shifting of mapped clusters was performed as explained in section 5.4. The normalization of the cluster intensities was executed as described above using the averaged CIF intensities produced from the Illumina pipeline during sequencing as done by Nutiu et al. (2011). Intensity extraction was performed as stated in section 5.6, image outlier detection was carried out as described in section 5.7, and DNA cluster sequence filtering was applied as stated in section 5.8. The k-mer ranking was performed as described in section 5.9.1.5. The Hill based fitting for determining the Kds was done as described in section 5.10.

There is an inferior agreement (R=0.86, δ=13.31%) between HiP-FA Kds and HiTS-FLIP Kds using the sequence images for normalization as shown by Figure 5.29, compared to using the protein images directly (R=0.99, δ=30.91%) as shown by Figure 5.52. In conclusion, using a single protein image thus allows to estimate directly the non-even illumination in a precise fashion without having to process all sequence images and introducing a nucleotide specific bias.

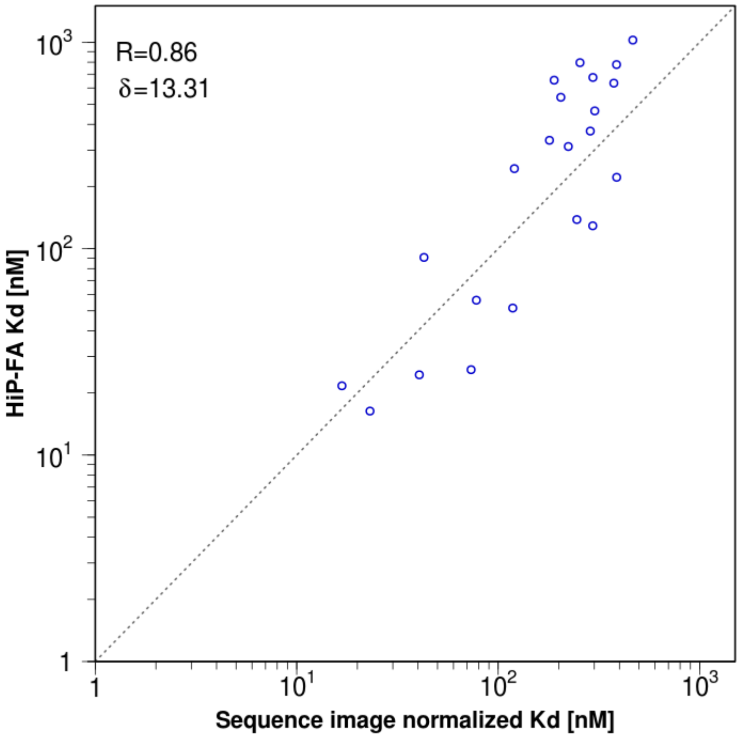


Figure 5.29: Validation of sequence image based normalized Kds leads to an inferior agreement between HiP-FA Kds and HiTS-FLIP Kds compared to using the protein images directly (see Figure 5.29).

5.6 Intensity extraction

After estimation of the transformation parameters that allows to localize the DNA clusters in the tile images, and image normalization to correct for non-even illumination, the intensities associated with the DNA clusters need to be extracted from the tif images. The following sections investigate different methods for the intensity extraction.

5.6.1 Implementation

I implemented the following methods the nearest neighbor intensity extraction, the Gaussian based intensity extraction, the intensity extraction based on the average of 2×2 pixel area, the intensity extraction based on the brightest 2×2 pixel area, the intensity extraction based on the average of 3×3 pixel area, the bilinear intensity extraction, and the bicubic intensity extraction using Java (Gosling (2000)) and ImageJ (Abràmoff et al. (2004); Schneider et al. (2012)). The intensity extraction based on the weighted area coverage is based on modified code from Illumina's OLB pipeline version 1.9.4 (Inc. (2011a)). The input are the normalized protein images and the x,y coordinates of the cluster positions. The output are the extracted cluster intensities.

5.6.2 Nearest neighbor intensity extraction

Given the coordinates (x_0, y_0) of a point, where x_0 and y_0 are floating-point numbers, and assuming u is the integer part of x_0 and v is the integer part of y_0, the rectangular neighborhood defined by pixels (u, v), $(u, v+1)$, $(u+1, v)$, and $(u+1, v+1)$ contain the point (x_0, y_0), as shown in Figure 5.30.
More formally:
The pixel closest to a given continuous point (x_0, y_0) is found by rounding the x_0 and y_0 coordinates independently to integral values:

$$I_{cluster} = \hat{I}(x_0, y_0) = I(u_0, v_0) \tag{5.30}$$

with
$u_0 = round(x_0) = \lfloor x_0 + 0.5 \rfloor$
$v_0 = round(y_0) = \lfloor y_0 + 0.5 \rfloor$
The computational complexity of nearest-neighbor resampling is on the order of n comparisons if the image contains n pixels, thus $\mathcal{O}(n)$ Goshtasby (2012).

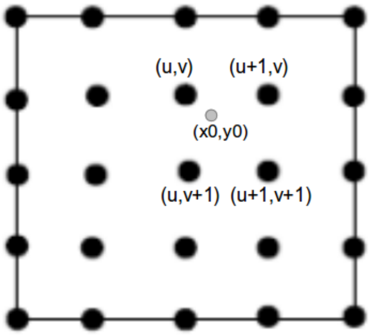


Figure 5.30: The nearest neighbor intensity for (x_0, y_0) is (u, v). Here, $u_0 = u$ and $v_0 = v$.

5.6.3 Gaussian based intensity extraction

One natural way to describe the intensity distribution of an amplified DNA cluster is by a 2d Gaussian function which simulates the blurring effect and the variance of Gaussian (σ) changes linearly with the axial axis. It has been shown that a 2d Gaussian function is suitable to represent the PSF of point sources in fluorescent microscopic images (Stallinga and Rieger (2010); Zhang et al. (2007)). Thus the intensity at a subpixel location l can be estimated from the intensities of a small number of pixels at discrete locations surrounding l. The following formula has been used:

$$f(x,y) = \frac{A}{2\pi\sigma_x^2\sigma_y^2} \exp\left(-\left(\frac{(x-x_0)^2}{2\sigma_x^2} + \frac{(y-y_0)^2}{2\sigma_y^2}\right)\right) + O \tag{5.31}$$

where (x_0, y_0) is the position of the peak (the center), σ_x and σ_y is the Gaussian width in x and y direction, A is the amplitude and O is the offset.
For each cluster a pixel box of size $w = 5$ around its position has been used as a subset of intensities for the fit. The initialization of the six parameters is carried out such as:
x_0: x coordinate of peak amplitude.
y_0: y coordinate of peak amplitude.
σ_x: $w/10$ (heuristic for $\sigma = FWHM/\sqrt{8ln2}$).
σ_y: $w/10$ (heuristic for $\sigma = FWHM/\sqrt{8ln2}$).
A: intensity of (x_0, y_0) minus offset.
O: minimal intensity within w.
The cluster intensity is then:

$$I_{cluster} = \hat{I}(x_0, y_0) = \hat{I} + \hat{O} \tag{5.32}$$

Figure 5.31 depicts the intensities of a DNA cluster from tile 21, cycle 95 (25 nM), lane 2 of experiment 18.08.2014, which is suitable to be fitted by the Gaussian function.
However, there are clusters for which a Gaussian cannot be fitted. The following intensities of Figure 5.32 of a DNA cluster from tile 21, cycle 95 (25 nM), lane 2 of experiment 18.08.2014 emphasize this situation. It can be observed that there is only one distinguished pixel that determines the cluster intensity.

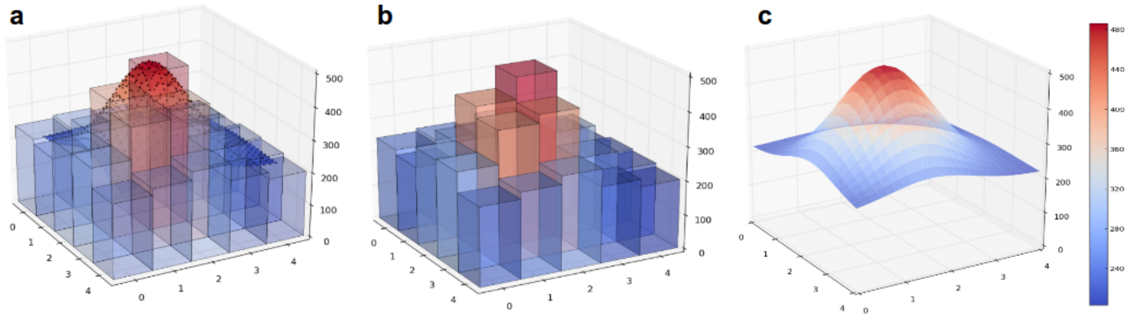


Figure 5.31: Cluster intensities with good Gaussian fit. (**a**) shows as 3d bar plot the cluster intensities and the overlaid Gaussian fit. (**b**) shows the cluster intensities, (**c**) shows the Gaussian fit.

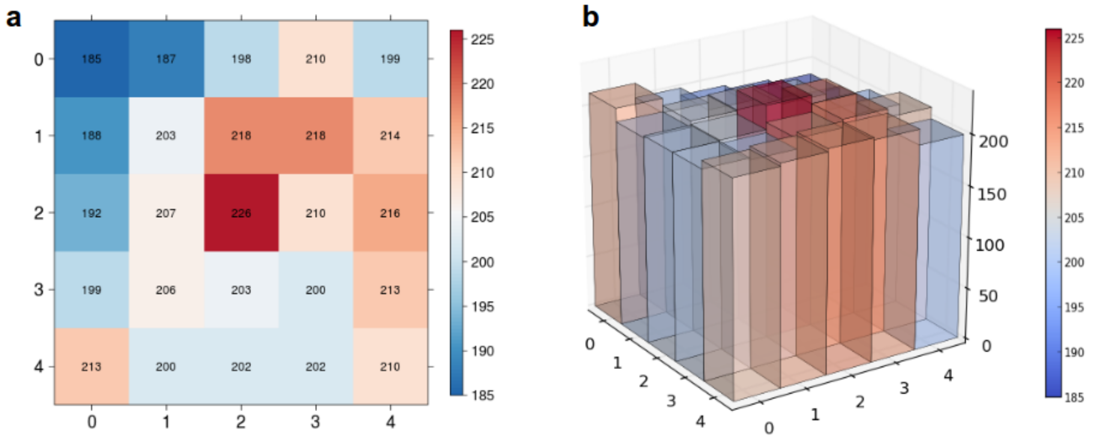


Figure 5.32: Cluster intensities for which no Gaussian fit is possible. (**a**) shows the intensities as a 2d heat map, (**b**) shows the intensities as a 3d bar plot.

5.6.4 Intensity extraction based on average of 2×2 pixel area

The cluster intensity for (x_0, y_0) can be defined as the averaged intensity over the closest four pixel neighbors ($n = 4$), illustrated by Figure 5.33.

$$I_{cluster} = \hat{I}(x_0, y_0) = \frac{1}{n} \sum_{j=0}^{n-3} \sum_{i=0}^{n-3} I(u + i * a, v + j * b) \tag{5.33}$$

where

$$a = \begin{cases} 1, & \text{if } x_0 - u_0 > 0 \\ -1, & \text{otherwise} \end{cases}$$

$$b = \begin{cases} 1, & \text{if } y_0 - v_0 > 0 \\ -1, & \text{otherwise} \end{cases}$$

$u_0 = round(x_0) = \lfloor x_0 + 0.5 \rfloor$

$v_0 = round(y_0) = \lfloor y_0 + 0.5 \rfloor$

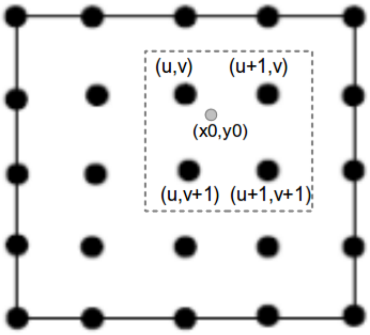


Figure 5.33: The intensity for (x_0, y_0) is the average intensity from the four neighboring pixels (u, v), $(u + 1, v)$, $(u, v + 1)$, and $(u + 1, v + 1)$.

5.6.5 Intensity extraction based on brightest 2×2 pixel area

A variation of the previous intensity extraction method is to use the brightest 2×2 pixel window which includes the cluster pixel, illustrated by Figure 5.34.

$$I_{cluster} = \hat{I}(x_0, y_0) = max(I_{A_1}, I_{A_2}, I_{A_3}, I_{A_4}) \tag{5.34}$$

where

$$I_{A_1} = \frac{1}{n}(I(u - 1, v) + I(u, v) + I(u, v + 1) + I(u - 1, v + 1)) \tag{5.35}$$

$$I_{A_2} = \frac{1}{n}(I(u,v) + I(u+1,v) + I(u+1,v+1) + I(u,v+1)) \tag{5.36}$$

$$I_{A_3} = \frac{1}{n}(I(u,v-1) + I(u+1,v-1) + I(u+1,v) + I(u,v)) \tag{5.37}$$

$$I_{A_4} = \frac{1}{n}(I(u-1,v-1) + I(u,v-1) + I(u,v) + I(u-1,v)) \tag{5.38}$$

with $n = 4$.

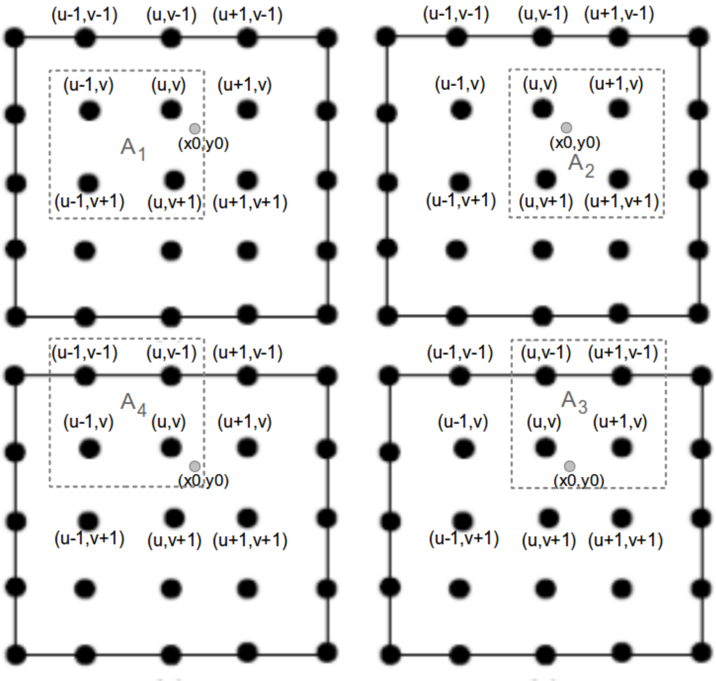


Figure 5.34: Example of the four 2×2 pixel windows around (x_0, y_0), the average intensity of the brightest 2×2 pixel window is chosen to be the intensity of (x_0, y_0).

5.6.6 Intensity extraction based on average of 3×3 pixel area

The neighboring pixels can be increased to a 3×3 pixel window, illustrated by Figure 5.35.

$$I_{cluster} = \hat{I}(x_0, y_0) = \frac{1}{n} \sum_{j=v_0-1}^{v_0+1} \sum_{i=u_0-1}^{u_0+1} I(i,j), n = 9 \tag{5.39}$$

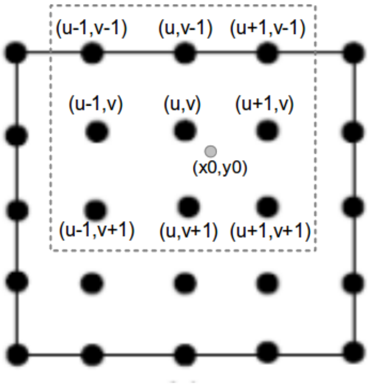


Figure 5.35: The intensity for (x_0, y_0) is the average intensity from the nine neighboring pixels.

5.6.7 Bilinear intensity extraction

Bilinear interpolation is used when values at random position on a regular 2d grid (discrete pixel values) need to be determined. Existing values are interpolated at fixed grid location to compute values anywhere else on the grid.

Interpolation takes place both in x- and y-direction, hence the name bilinear (Demant et al. (2013)). The result of bilinear interpolation is independent of which axis is interpolated first and which second. Bilinear interpolation uses the distance-weighted average of the four nearest pixel values to estimate a new pixel value (Goshtasby (2012)). The weight on each of the four pixel values is based on the computed pixel's distance (in 2d space) from each of the known points.

First, the four closest (surrounding) pixels are determined. Then, two horizontal linear interpolations are done, obtaining $I(\Delta u, v)$ and $I(\Delta u, v+1)$. Finally, a third vertical linear interpolation is carried out to obtain $I(x_0, y_0)$. An illustration of this calculation is provided by Figure 5.36.

According to (Goshtasby (2012)) bilinear interpolation can be defined as:

$$I_{cluster} = \hat{I}(x_0, y_0) = w_1{*}I(u{+}1, v{+}1){+}w_2{*}I(u, v{+}1){+}w_3{*}I(u{+}1, v){+}w_4{*}I(u, v) \quad (5.40)$$

where

$$w_1 = \Delta u \Delta v = (x_0 - u)(y_0 - v) \quad (5.41)$$

$$w_2 = (1 - \Delta u)\Delta v = (u + 1 - x_0)(y_0 - v) \quad (5.42)$$

$$w_3 = \Delta u(1 - \Delta v) = (x_0 - u)(v + 1 - y_0) \quad (5.43)$$

$$w_4 = (1 - \Delta u)(1 - \Delta v) = (u + 1 - x_0)(v + 1 - y_0) \quad (5.44)$$

Computationally, resampling by bilinear interpolation requires on the order of n multiplications if the reference image contains n pixels. Therefore, nearest-neighbor and bilinear interpolation have the same computational complexity, although nearest-neighbor is several times faster than bilinear interpolation (Goshtasby (2012)).

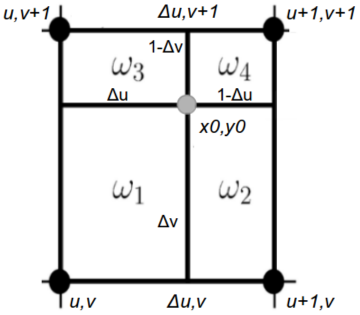


Figure 5.36: Bilinear interpolation. For a given position (x_0, y_0), the interpolated value is computed from the intensity values of the four closest pixels $(u, v + 1), (u + 1, v + 1), (u + 1, v), (u, v)$ in two steps. First the intermediate values $(\Delta u, v)$ and $(\Delta u, v + 1)$ are computed by linear interpolation in the horizontal direction between (u, v) and $(u + 1, v)$, and $(u, v + 1)$ and $(u + 1, v + 1)$, respectively, where Δu is the distance to the nearest pixel to the left of x_0. Subsequently, the intermediate values $(\Delta u, v)$ and $(\Delta u, v{+}1)$ are interpolated in the vertical direction, where Δv is the distance to the nearest pixel below y_0.

5.6.8 Bicubic intensity extraction

Bicubic interpolation is an extension of cubic interpolation for interpolating data points on a 2d regular grid. The interpolated surface is smoother than corresponding surfaces obtained by bilinear interpolation or nearest-neighbor interpolation. Bicubic interpolation can be accomplished using either Lagrange polynomials, cubic splines, or cubic convolution algorithm (Burger et al. (2009)). In contrast to bilinear interpolation, which only takes 4 pixels (2 × 2) into account, bicubic interpolation considers 16 pixels (4 × 4). Images resampled with bicubic interpolation are smoother and have fewer interpolation artifacts. Bicubic interpolation occurs in two steps. According to (Burger et al. (2009)), at first, a one-dimensional cubic interpolation is performed in the horizontal direction with $w_{cub}(x)$ over the four pixel intensities $I(u_i, v_j)$ in four lines. Then, the result $\hat{I}(x_0, y_0)$ is computed by a one-dimensional cubic interpolation in the vertical direction over the intermediate results $p_0 ... p_3$. An illustration of this calculation is provided by Figure 5.37.

$$I_{cluster} = \hat{I}(x_0, y_0) = \sum_{j=0}^{3} \Big[w_{cub}(y_0 - v_j) \sum_{i=0}^{3} \big[I(u_i, v_i) w_{cub}(x_0 - u_i) \big] \Big] \tag{5.45}$$

with $u_i = \lfloor x_0 \rfloor - 1 + i$ and $v_j = \lfloor y_0 \rfloor - 1 + i$, and where

$$w_{cub}(x) = \begin{cases} (a+2)|x|^3 - (a+3)|x|^2 + 1 & \text{for } |x| \leq 1 \\ a|x|^3 - 5a|x|^2 + 8a|x| - 4a & \text{for } 1 < |x| < 2 \\ 0 & \text{otherwise} \end{cases}$$

with $a = -0.5$ (another common value is -0.75).

The value $p_j = \sum_{i=0}^{3} \big[I(u_i, v_i) w_{cub}(x_0 - u_i) \big]$ denotes the intermediate result of the cubic interpolation in the x direction in line j. The interpolation is based on a 4×4 neighborhood of pixels and requires a total of $16 + 4 = 20$ additions and multiplications. This computation of bicubic interpolation is on the order of n^2 multiplications, thus $\mathcal{O}(n^2)$ (Goshtasby (2012)).

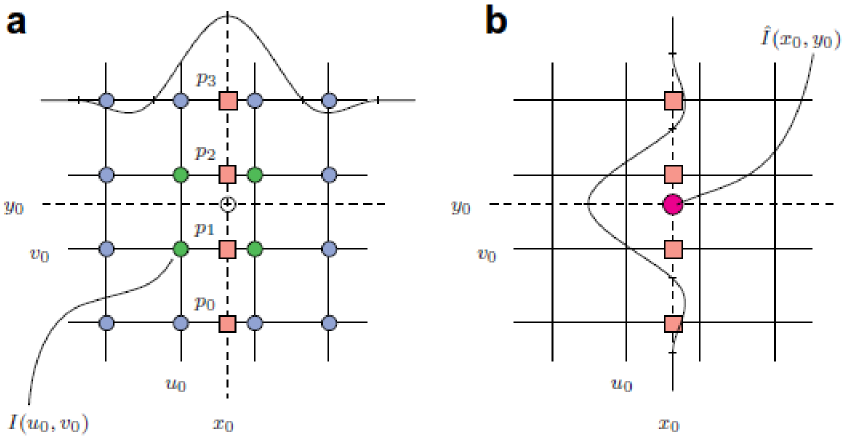


Figure 5.37: Bicubic interpolation in two steps. The discrete image I is to be interpolated at some continuous position (x_0, y_0). (a) In step 1, a one-dimensional interpolation is performed in the horizontal direction with $w_{cub}(x)$ over four pixels $I(u_i, v_j)$ in four lines. One intermediate result p_j is computed for each line j. (b) In step 2, the result $\hat{I}(x_0, y_0)$ is computed by a single cubic interpolation in the vertical direction over the intermediate results $p_0 ... p_3$. Adapted from Burger et al. (2009).

5.6.9 Intensity extraction based on weighted area coverage

Another variation of an intensity extraction method can be derived by adjusting weights due to the area coverage around the central cluster pixel in the following way. This idea is based on modified code from Illumina's OLB pipeline version 1.9.4 (Inc. (2011a)).

$$I_{cluster} = \hat{I}(x_0, y_0) = \frac{1}{A}\sum_{i=1}^{n} w_i I(x_i, y_i) = \frac{1}{A}\sum_{i=1}^{n} w_i \big[w_c I(x_c, y_c) + w_N \sum_{j=1}^{4} I(x_j^N, y_j^N)\big] \tag{5.46}$$

where
A = area of coverage
n = number of pixels p^A overlapped by A
w_i = weight based on overlap between A and the area of pixel p_i^A

$$I(x_i, y_i) = w_c I(x_c, y_c) + w_N \sum_{j=1}^{4} I(x_j^N, y_j^N) \tag{5.47}$$

where
$I(x_c, y_c)$ = intensity of cluster pixel covered by A
$I(x_j^N, y_j^N)$ = intensity of 4-connected neighbor pixels of p_i^A
w_c = weight of central pixel (x_c, y_c)
w_N = weight of neighboring pixel (x_j^N, y_j^N)

Figure 5.38 depicts an example. The point p for which the intensity is extracted is $(2.1, 2.4)$ which is formalized above by (x_c, y_c). The area A around p is denoted by the red square. Here $n = 6$, since 6 pixels are affected by the overlap of A. The six pixels p_i^A of p affected by the overlap of A are marked by the blue frames. The weights w_i in Figure 5.38, determined by the overlap between A and the area of pixel p_i^A, are $w_1 = 0.15 \times 0.85 = 0.1275$, $w_2 = 0.15 \times 0.65 = 0.0975$, $w_3 = 1.0 \times 0.65 = 0.65$, $w_4 = 0.35 \times 0.65 = 0.2275$, $w_5 = 0.35 \times 0.85 = 0.2975$, and $w_6 = 1.0 \times 0.85 = 0.85$.

The parameter settings for the HiTS-FLIP pipeline are based on Illumina's OLB pipeline version 1.9.4 (Inc. (2011a)):
$A = 1.5^2$ pixels
$w_c = 5.0$
$w_N = 0.9$

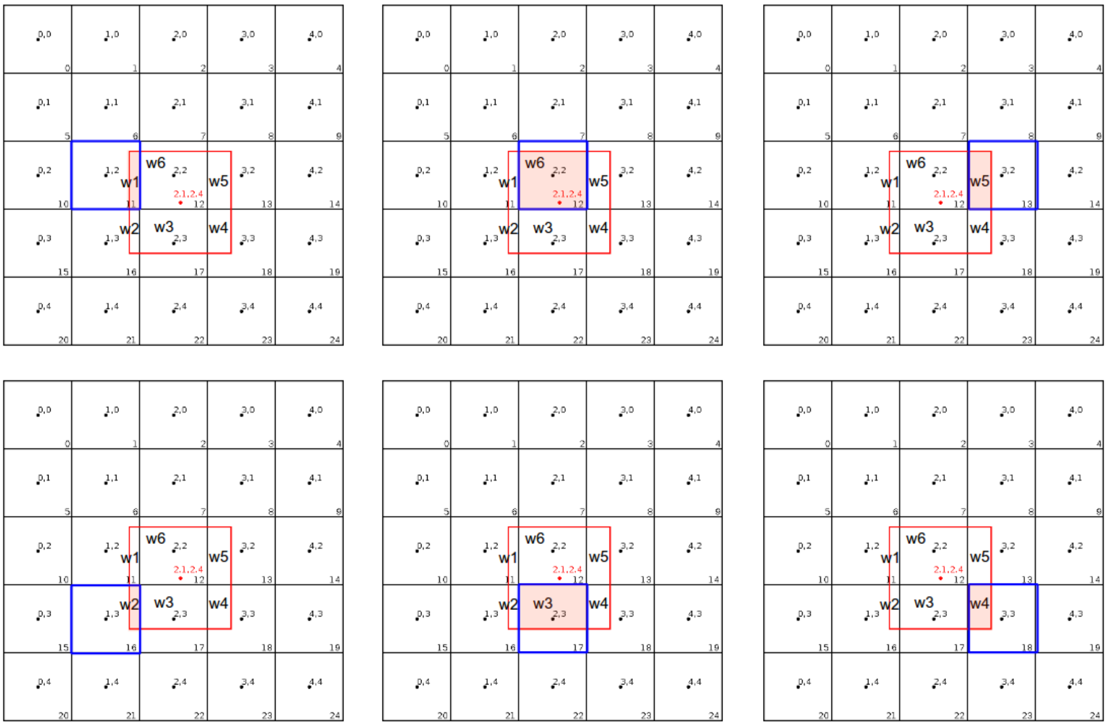


Figure 5.38: Intensity extraction based on weighted area coverage. The point p for which the intensity is extracted is $(2.1, 2.4)$. Area A is denoted by the red frame. A has size 1.5^2 pixels. Affected neighboring pixels by the overlap with A are denoted by the blue frames. Here $n = 6$, since 6 pixels are affected by the overlap. The weights w_i here are the weights w_1 to w_6 (red colored areas). For example, $w_1 = 0.15 \times 0.85 = 0.1275$.

5.6.10 Comparison of different intensity extraction methods

The different intensity extraction methods were compared by the correlation to the Kds measured by HiP-FA as described in section 5.9.1.7.

The underlying data set for this comparison are the data from lane 2 of experiment 18.08.2014. Preprocessing of the tif images was done as described in section 5.2, the cluster position transformation was performed as described in section 5.3, and normalization of the cluster intensities was executed as detailed in section 5.5, Shifting of mapped clusters was done as explained in section 5.4, image outlier detection was carried out as described in section 5.7, and DNA cluster sequence filtering was applied as stated in section 5.8. The k-mer ranking was carried out as stated in subsection 5.9.1 but without any cluster deletion. The Hill based fitting for determining the Kds was done as described in section 5.10.

The results of the comparison are shown in the Table 5.3. The Pearson's product-moment correlation coefficients show a high correlation for all methods except for the Gaussian based intensity extraction method. However, the relative error δ is the smallest for the intensity extraction method based on weighted area coverage. Therefore, this method has been chosen for the HiTS-FLIP pipeline with the settings stated in subsection 5.6.9.

method	R	δ
Nearest neighbor intensity extraction	0.98	38.83%
Gaussian based intensity extraction	0.86	664.27%
Intensity extraction based on average of 2 × 2 pixel area	0.98	44.36%
Intensity extraction based on brightest of 2 × 2 pixel area	0.98	45.28%
Intensity extraction based on average of 3 × 3 pixel area	0.98	54.44%
Bilinear intensity extraction	0.98	50.47%
Bicubic intensity extraction	0.98	35.6%
Intensity extraction based on weighted area coverage	0.99	30.91%

Table 5.3: Comparison of the different intensity extraction methods with respect to the correlation of Kds and relative errors as measured by HiP-FA as described in section 5.9.1.7.

5.7 Image outlier detection

Besides non-even illumination dust particles and air bubbles can also obfuscate cluster intensities. The contamination by dust particles and air bubbles can vary quite drastically from experiment to experiment. Dust particles are a contamination appearing as very bright spots on the tile images as shown by Figure 5.39. Air bubbles reach the flow cell by the syringe pump system and can sometimes cover a large area of a tile as illustrated by Figure 5.40. Filtering out clusters affected by dust particles and air bubbles reduces false positives since their bright appearance is only artificial and not due to a high amount of bound protein.

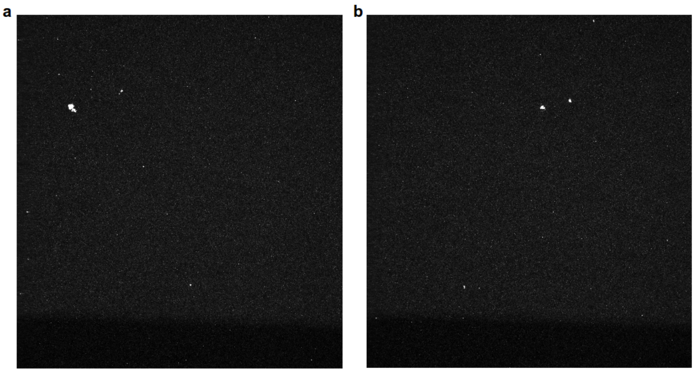


Figure 5.39: Dust particles on imaged tiles. (a) and (b) show two images with dust particles (bright spots) contaminating the imaged tile area.

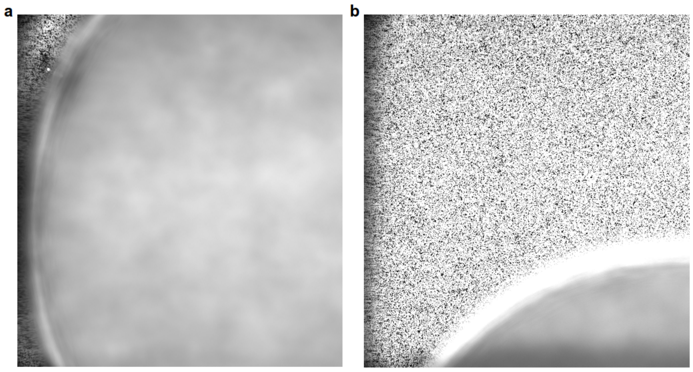


Figure 5.40: Example of air bubbles covering the imaged tile area. (a) Image where more than 80% of all the DNA clusters on the tile are affected by the bubble. (b) Image where a portion of the tile is covered by an air bubble, and ca. 21% of all the DNA clusters on the tile are affected.

5.7.1 Detection approach for air bubbles

Since dust particles and air bubbles distinguish themselves drastically by their appearance and size from DNA clusters and background, they can be easily detected and affected DNA clusters can be filtered out. The following steps comprise the detection of an air bubble as illustrated by Figure 5.41.

In *step 1*, a sharpening filter is applied to the image using a 3 × 3 convolution kernel increasing contrast and accentuating details. The implementation of the sharpening filter is based on the method sharpen() from ImageJ's component ImageProcessor, URL: `http://rsbweb.nih.gov/ij/docs/guide/146-29.html`. Then the image is binarized using Otsu's method (Otsu (1975)) using the ImageJ plugin implementation by C. Mei et al., URL: `http://rsb.info.nih.gov/ij/plugins/otsu-thresholding.html`. Otsu's method is a threshold based binarization algorithm that aims to maximize the inter-class variance and does no require any user defined parameters (Otsu (1975)).

In *step 2*, a blob detection for identifying particle objects is carried out by connected component labeling (Chang et al. (2004)), which is implemented in the ImageJ library IJBlob (Wagner and Lipinski (2013)). The biggest blob is considered as an air bubble.

In *step 3*, the concave contour line, reaching into the image, is detected.

In *step 4*, a circle is fitted to the detected bubble that allows to differentiate between all the pixels that belong to the air bubble and the pixels that lie outside.

Finally in *step 5*, all pixels belonging to the air bubble (> 30000 pixels) are marked (here in red) so that they can be distinguished and filtered out.

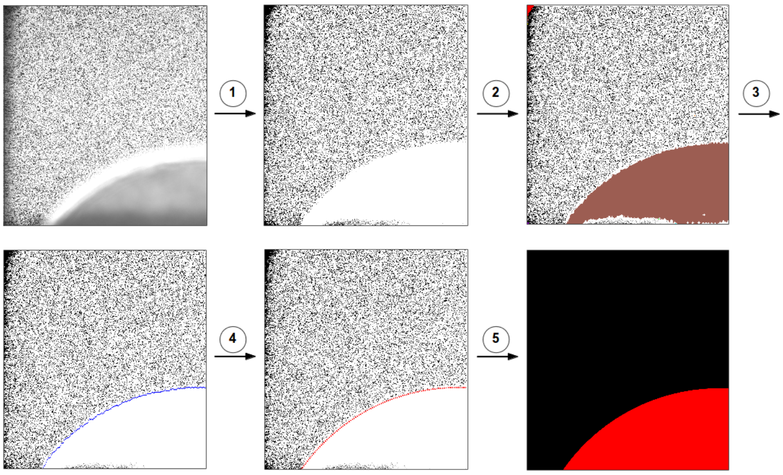


Figure 5.41: Overview of the different steps in the air bubble detection process.

5.7.2 Detection approach for dust particles

The following steps make up the detection of a dust particle as illustrated by Figure 5.42. In *step 1*, a sharpening filter is applied to the image using a 3 × 3 convolution kernel increasing contrast and accentuating details. The implementation of the sharpening filter is based on the method sharpen() from ImageJ's component ImageProcessor, URL: `http://rsbweb.nih.gov/ij/docs/guide/146-29.html`. Then the image is binarized using Otsu's method (Otsu (1975)) using the ImageJ plugin implementation by C. Mei et al., URL: `http://rsb.info.nih.gov/ij/plugins/otsu-thresholding.html`. Otsu's method is a threshold based binarization algorithm that aims to maximize the inter-class variance and does no require any user defined parameters (Otsu (1975)).
In *step 2*, a blob detection for identifying particle objects is carried out by connected component labeling (Chang et al. (2004)), which is implemented in the ImageJ library IJ Blob (Wagner and Lipinski (2013)).
In *step 3*, blobs representing particles are identified by the size of their outer contour line (between 30 and 30000 pixels) and marked (here in red) as regions to be filtered out.

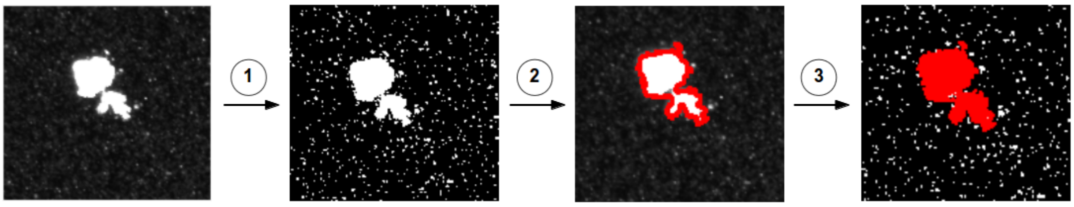


Figure 5.42: Overview of the different steps in the dust particle detection process.

5.8 DNA sequence filtering

It is crucial to include only correctly identified bases in the analysis. Since base calling is a probabilistic process, a certain threshold needs to be applied in order to confidently determine a reliable base accuracy.

5.8.1 Per base sequence quality plot

Figure 5.43 shows an overview of the range of quality values across all bases at each position from the FASTQ file of lane 2 of experiment 18.08.2014. For each position a BoxWhisker type plot is drawn. According to (Andrews (2010–2015b)) the elements of the plot are as follows:

- The central red line is the median value
- The yellow box represents the inter-quartile range (25-75%)
- The upper and lower whiskers represent the 10% and 90% points
- The blue line represents the mean quality
- The y-axis on the graph shows the quality scores. The higher the score the better the base call.

The background of the graph divides the y-axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). The quality of calls on most platforms will degrade as the run progresses, so it is common to see base calls falling into the orange and red area towards the end of a read.

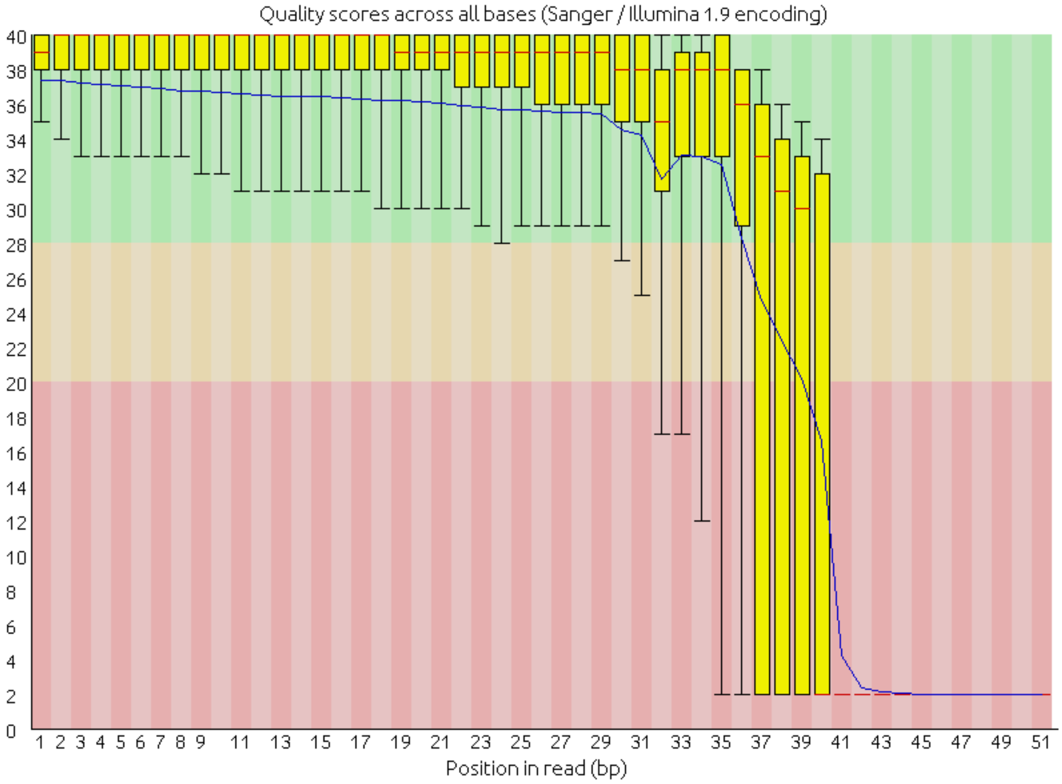


Figure 5.43: Per base sequence quality plot for lane 2 of experiment 18.08.2014, produced with the FastQC tool (Andrews (2010–2015a)).

5.8.2 Per sequence quality scores plot

The per sequence quality score report allows one to see if a subset of the sequences have universally low quality values (Andrews (2010–2015b)). It is often the case that a subset of sequences will have universally poor quality, often because they are poorly imaged (on the edge of the field of view), however such low quality subsets should represent only a small percentage of the total sequences.

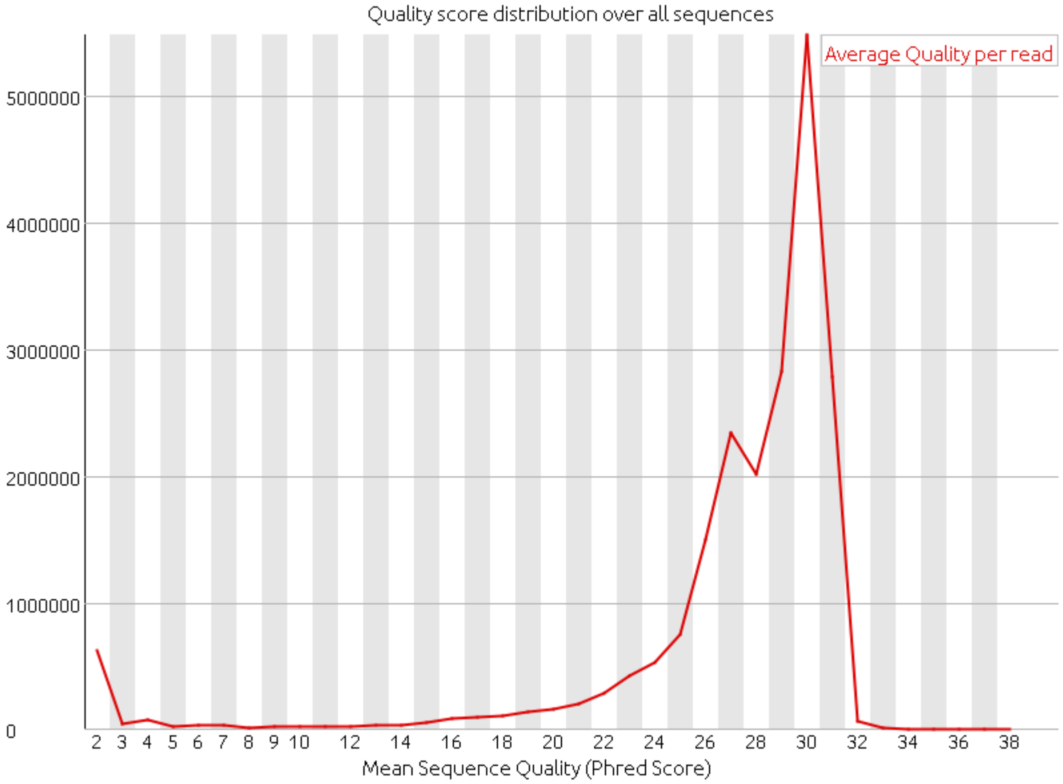


Figure 5.44: Per sequence quality scores plot for lane 2 of experiment 18.08.2014, produced with the FastQC tool (Andrews (2010–2015a)).

5.8.3 Phred quality scores

A Phred quality score (Q-score) is a prediction of the probability of an error in base calling and the most common metric used to assess the accuracy of a sequencing platform (Inc. (2011b, 2014)). During a sequencing run with the GA-IIx, a quality score is assigned to each base call for every cluster, on every tile, for every sequencing cycle. The GA-IIx generates per-cycle BCL basecall files which are then converted to per-read FASTQ files where both the sequence letter and quality score are each encoded with a single ASCII character. A high quality score implies that a base call is more reliable and less likely to be incorrect. Q-scores are defined as a property that is logarithmically related to the base calling error probabilities P (Ewing and Green (1998)):

$$Q = -10 \log_{10} P \tag{5.48}$$

For base calls with a quality score of $Q30$, one base call in 1000 is predicted to be incorrect (error probability 0.001). This quality measure for the base call accuracy has been applied in the HiTS-FLIP pipeline.

5.8.4 Implementation

I implemented the parser of the quality score in Java (Gosling (2000)). The input is a FASTQ file and the output is the related quality score as integer.

5.9 K-mer ranking

A crucial part besides the cluster position transformation, the normalization against the uneven illumination bias and the intensity extraction is the ranking of the binding motifs according to their affinity to the underlying DNA cluster sequence.

5.9.1 Heuristic ranking algorithm

To estimate Kd values for short k-mers contained in longer sequences, one must be able to assign the binding intensity to the correct k-mer (Nutiu et al. (2011)).
Because the oligonucleotides clustered on the flow cell are relatively short (25 variable nucleotides insert, but 150 nucleotides in total), it is extremely rare for a specific k-mer sequence to occur in a cluster more than once, for example roughly 0.1% of 7-mers occur more than once in any of the clusters (Nutiu et al. (2011)). Because of this fact, Nutiu et al. employed an iterative algorithm to assign binding intensities to k-mers that works in the following way:
For each sequence of size k, the median intensity at a GCN4 concentration of 125 nM was calculated over all clusters containing a certain k-mer or its reverse complement. The k-mer with the highest median intensity at 125 nM was selected, then its dissociation constant was determined based on the median intensities, and finally all clusters containing this k-mer or its reverse complement were removed from further calculations.
Nutiu et al. repeated this procedure iteratively until all remaining k-mers had $Kd > 1\mu M$, selecting sequences with affinity that could not be explained by occasional overlap with stronger binding sites. Due to computational limitations, Nutiu et al. only performed the iterative removal of clusters for 9-mers containing 8-mers that had $Kd < 1\mu M$. Likewise, Nutiu et al. only performed iterative removal of clusters for 10-mers containing 9-mers that had $Kd < 1\mu M$ and so on for longer k-mers.
Thus, Nutiu et al. made the assumption that all of the intensity of the clusters containing the top k-mer comes from binding to that specific k-mer, and not to others. In the case of 8-mers, for example, 95% of the time there will not be two 8-mers having $Kd < 1\mu M$ in the same cluster. After removing clusters containing this k-mer, Nutiu et al. make the same assumption for the k-mer with the next highest median binding intensity in the remaining clusters.

The Nutiu et al. ranking algorithm has the following form.
L: number of k-mers k to be ranked.
N: number of DNA clusters c on one lane.
C: number of different concentrations.
C': number of concentrations without concentration level used for ranking.
i_c: intensity of DNA cluster c.
i_k*-list*: DNA cluster intensity list of k-mer k at selected ranking concentration.
m_L*-list*: median intensity list of all k-mers at selected ranking concentration.
top$_k$: top ranked k-mer k after each iteration.
top$_L$*-list*: list of all ranked k-mers.

```
while $L > 0$ do
    for all $k$ in $L$ do
        for all $c$ in $N$ do
            if $k \in c$ or $revcomp(k) \in c$ then
                $i_k\text{-}list[index_c] = i_c$
        $m_L\text{-}list[index_k] = median(i_k\text{-}list)$
    $sort(m_L\text{-}list)$
    $top_k = first(m_L\text{-}list)$
    $top_L\text{-}list[index_k] = top_k$
    for $j$ in $C'$ do
        $median(top_k)$
    $L = L - 1$
    $N = N - \{top_k \in c$ or $revcomp(top_k) \in c\}$
```

The run time complexity for one iteration is given by:

$$\begin{aligned}N * \mathcal{O}(l_c + l_k) + L * \mathcal{O}(n_k \log(n_k)) + \mathcal{O}(L \log(L)) + C' * \mathcal{O}(n_{top\text{-}k} \log(n_{top\text{-}k})) + \mathcal{O}(n_c)\\ = \mathcal{O}(l_c + l_k) + \mathcal{O}(n_k \log(n_k)) + \mathcal{O}(L \log(L)) + \mathcal{O}(n_{top\text{-}k} \log(n_{top\text{-}k})) + \mathcal{O}(n_c)\end{aligned} \tag{5.49}$$

since for N remaining clusters $\mathcal{O}(l_c + l_k)$ operations have to be done with l_c the length of the cluster c and $2 * l_k$ the length of the k-mer k and its reverse complement to determine if k or its reverse complement is contained in the cluster sequence, for L remaining k-mers $\mathcal{O}(n_k \log(n_k))$ median values have to be calculated where each k-mer k has on average n_k intensity values, L many k-mers have to be sorted in $\mathcal{O}(L \log(L))$ to get the current top k-mer and n_c many clusters containing the current top k-mer have to be removed in $\mathcal{O}(n_c)$. The time complexity of each loop through the remaining k-mers is mainly dominated by $\mathcal{O}(l_c + l_k) + \mathcal{O}(n_k \log(n_k))$ which are the most expensive operations. Given a k-mer of length 11, on average a particular 11-mer or its reverse complement occurs with the following probability at least once at a certain DNA cluster:

$$p_{occ} = 1 - (\frac{|A|^{l_k} - 2}{|A|^{l_k}})^{(n-l_k+1)} \tag{5.50}$$

where
A: alphabet of 4 letters.
l_k: length of k-mer.
n: number of variable nucleotides per DNA cluster.

For a particular 11-mer and its reverse complement it results in $p_{occ} = 0.000007152$ or around 179 occurrences of that particular 11-mer per lane.

This shows that the number N of clusters increases fairly slowly since the longer a k-mer k the fewer clusters can be removed per iteration. The total number L of all possible k-mers is given by $|A|^{l_k}$. For 11-mers, this is 4194304.

In order to rank all k-mers L, the execution has to iterate at first through n many k-mers, then through $n - 1$ many k-mers, and so on which amounts to a quadratic run time as shown by equation 5.51.

$$\mathcal{O}(n + (n-1) + (n-2) + ... + 2 + 1) = \mathcal{O}(\frac{n \times (n-1)}{2}) = \mathcal{O}(n^2) \tag{5.51}$$

5.9.1.1 Optimization of the heuristic ranking algorithm

Figure 5.45 gives an overview of the execution of the optimized ranking algorithm.

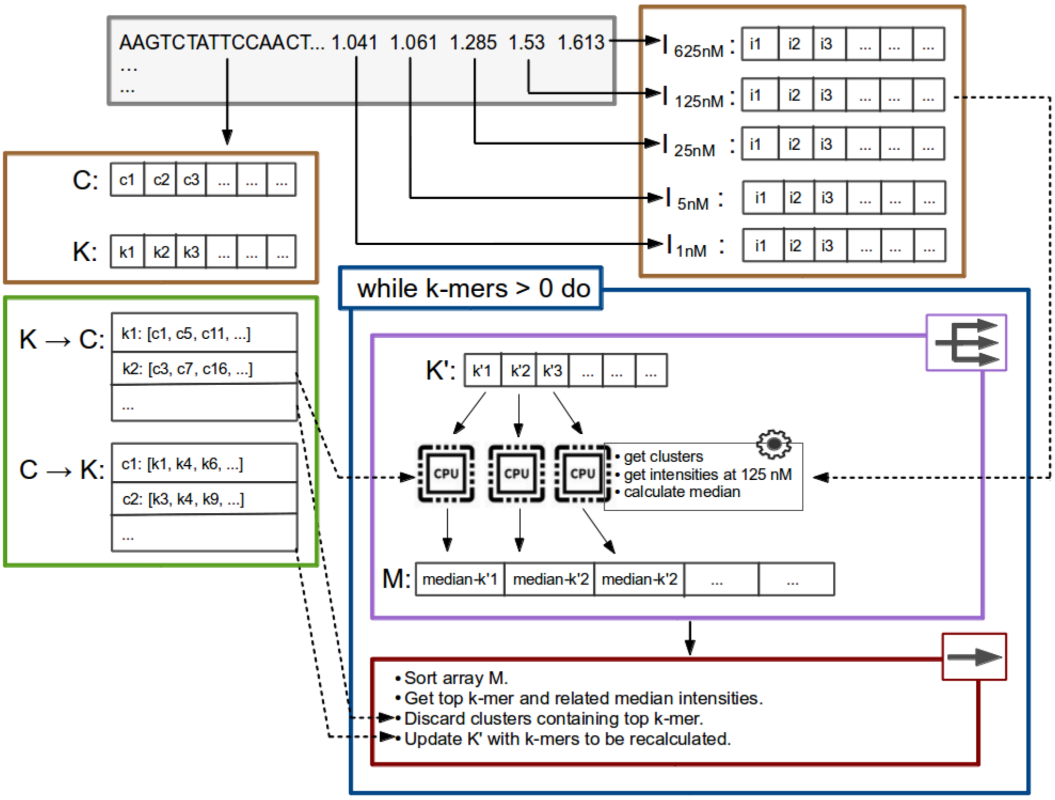

AAGTCTATTCCAACT... 1.041 1.061 1.285 1.53 1.613
...
...
I_{625nM}: i1 i2 i3 ... ...
I_{125nM}: i1 i2 i3 ... ...
I_{25nM}: i1 i2 i3 ... ...
I_{5nM}: i1 i2 i3 ... ...
I_{1nM}: i1 i2 i3 ... ...
C: c1 c2 c3 ... ...
K: k1 k2 k3 ... ...
K → C: k1: [c1, c5, c11, ...]
k2: [c3, c7, c16, ...]
...
C → K: c1: [k1, k4, k6, ...]
c2: [k3, k4, k9, ...]
...
while k-mers > 0 do
K': k'1 k'2 k'3 ... ...
CPU
• get clusters
• get intensities at 125 nM
• calculate median
M: median-k'1 median-k'2 median-k'2 ... ...
• Sort array M.
• Get top k-mer and related median intensities.
• Discard clusters containing top k-mer.
• Update K' with k-mers to be recalculated.

Figure 5.45: Overview of the ranking procedure. Gray box: input file consisting of cluster sequences and related intensities at the different concentration levels. Brown boxes: Data structures for storing the information from the input file. Array C stores the cluster sequences, arrays I store the cluster intensities, one array for each concentration level. Array K contains all the embedded k-mers from all the cluster sequences from the input file. Green box: Lookup maps to speed up the processing. The $K \rightarrow C$ map provides a lookup for a certain k-mer k for all the clusters c which contain k or its reverse complement. The $C \rightarrow K$ map provides for a certain cluster c a lookup for all the k-mers k that are embedded in the sequence of c. Blue box: Main execution loop for ranking the k-mers. Purple box: Median intensity calculation of the k-mers in array K' in a parallel fashion. Red box: Sequential part of the ranking procedure, which includes sorting array M, getting top k-mer and related median intensities, discarding clusters containing top ranked k-mer, and updating K' with k-mers for which the median intensities need to be recalculated.

At the beginning the DNA cluster sequences and the related cluster intensities at the different concentration levels are read in from a file and stored in the related arrays C (for the cluster sequences) and I (for the cluster intensities, one array for each concentration level). In addition array K is created (for all the embedded k-mers in the cluster sequences). There are several improvements that can be made to speed-up the execution of the ranking algorithm.

1) Lookup maps

In an initial phase before the ranking starts, two important lookup maps are created (green colored box in Figure 5.45).

The $K \rightarrow C$ map provides a lookup for a certain k-mer k for all the clusters c which contain k or its reverse complement. This lookup helps to calculate the median intensity values for the k-mers to be ranked much quicker than by looping through k-mers and clusters. In addition, it gives immediate access to all the clusters that need to be discarded due to containing the current top ranked k-mer after each ranking iteration.

The $C \rightarrow K$ map provides for a certain cluster c a lookup for all the k-mers k that are embedded in the sequence of c. This lookup enables to determine quickly the set of k-mers for which the median intensities need to be recalculated due to the discarding of clusters which contained a previously top ranked k-mer.

2) Efficient calculation of median intensity values

The execution time can be reduced to a great extent if only those median intensity values are recalculated for which the related k-mers have been affected by discarded clusters. Only in such a situation do the intensity values change. The array K' contains the k-mers for which the intensities need to be recalculated.

3) Parallelization of k-mer intensity calculation

As illustrated by the purple box, the median intensity calculation of the k-mers in array K' can be parallelized for each k-mer k'. The array M which contains all the median intensity values for the different k-mers k' can be accessed by index without any blocking of the various threads.

5.9.1.2 Execution time of the optimized heuristic ranking algorithm

Figure 5.46 shows the execution time for one ranking iteration, i.e. determining the current top k-mer, its median intensity values and discarding clusters, for k-mers of length 11 nt using around 12 million DNA cluster sequences of length 25 nucleotides from experiment 18.08.2014. The purple dots at 1 and 4 CPU cores display execution times measured on an Intel Core i5-200K quad-core processor with 16 GB RAM. As expected applying parallelization allows to gain roughly a factor of 4 in speed-up. The blue triangles represent inferred execution times.

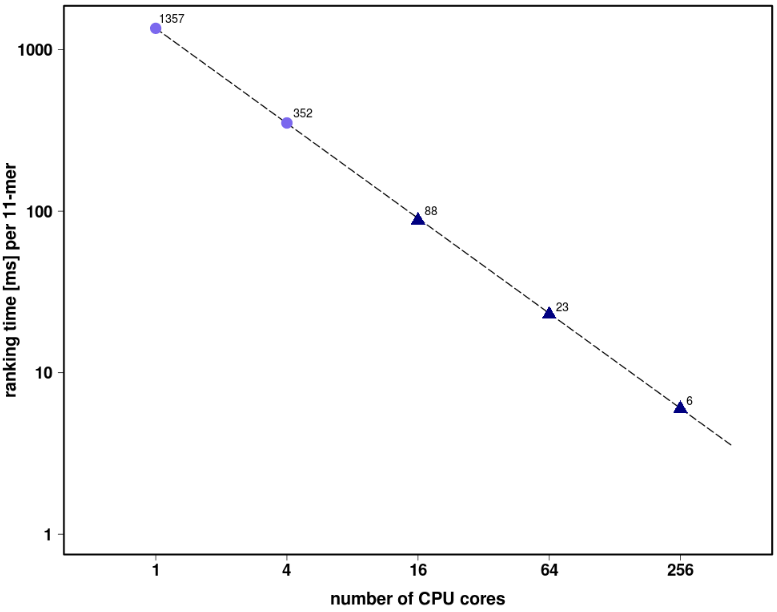


Figure 5.46: Speedup of the ranking algorithm from single processor to multiple processors. The purple dots are measured execution times, the blue triangles represent inferred execution times.

There are $4^{11} = 4194304$ possible 11-mers. Ranking all possible 11-mers using a computing cluster with 64 CPU cores would take around 27 hours, and with 256 CPU cores around 7 hours.

5.9.1.3 Implementation

I implemented the optimized heuristic ranking algorithm in C++ (Stroustrup (1986)) using OpenMP (Dagum and Enon (1998)) for parallelization.
The input are

- the length of the k-mer to be ranked
- number of ranked k-mers
- the concentration at which the ranking should be achieved
- the cluster sequences
- the different cluster intensities for the increasing concentrations

The output are the ranked k-mer motifs with their different intensities.

5.9.1.4 Issues with discarding DNA clusters

The total number of all possible 7-mers is $4^7 = 16384$. During the ranking, a k-mer and its reverse complement is treated equivalent, therefore there would be $\frac{16384}{2} = 8192$ ranking iterations. However, only 4728 ranking iterations can be executed and determine k-mers since 3464 k-mers cannot be ranked due to a lack of clusters in which they can occur ($4728 \cdot 2 + 3464 \cdot 2 = 16384 = 4^7$). Thus, in total 42% of all 7-mers cannot be ranked.
The Hamming distance between two strings of equal length is the number of positions at which the corresponding symbols are different (Hamming (1950)). In other words it measures the minimum number of substitutions required to change one string into the other, or the minimum number of errors that could have transformed one string into the other. In the example here, the top ranked 7-mer TGACTCA was used as reference.
The number of possible mutations *mut* of a DNA motif m with length l is calculated as follows:
z: number of nucleotide mutations.
l: length of DNA motif m.
n: sites in m to be mutated.

$$mut = z^n \binom{l}{n} \tag{5.52}$$

Figure 5.47 displays the accumulated number of discarded clusters during ranking (a), the loss of 7-mers during ranking which cannot be ranked (b), the total number of 7-mers that cannot be ranked due to cluster deletion (c), and in percent how many 7-mers with a certain Hamming distance from the top ranked 7-mer TGACTCA could be processed and ranked (d). On average, between 15000 and 17000 clusters are discarded while ranking one 7-mer (total number of clusters is 10620667). The analysis is based on the data from experiment 18.08.2014, lane 2. The cluster position transformation was performed as described in section 5.3, normalization of the cluster intensities was executed as detailed in section 5.5, and DNA cluster sequence filtering was applied as stated in section 5.8.

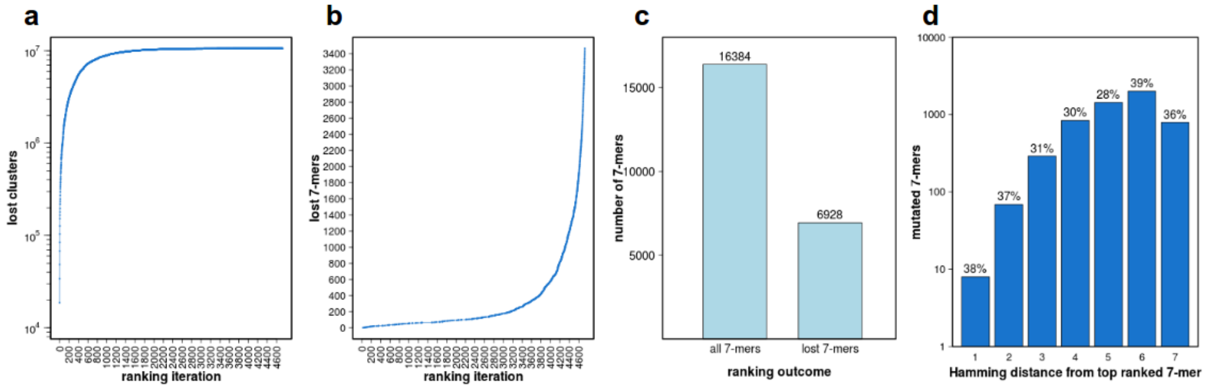
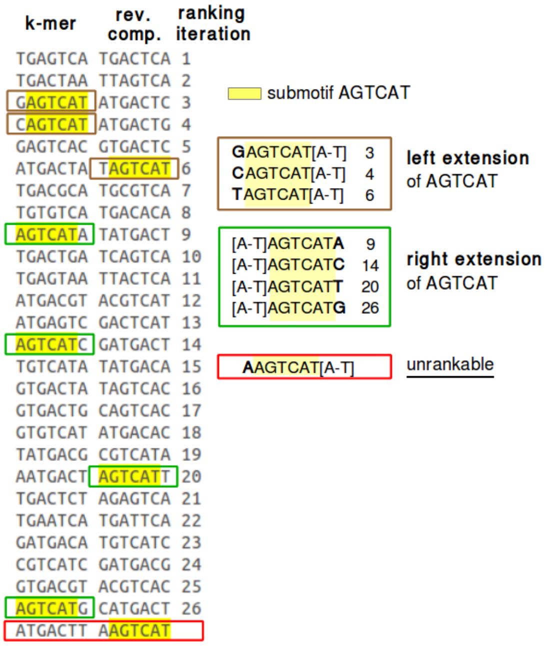


Figure 5.47: Accumulated loss of clusters and 7-mers during ranking, total number of unrankable 7-mers and percentage of 7-mers that can be ranked grouped by their Hamming distance from reference 7-mer TGACTCA. (a) Accumulated number of discarded clusters during ranking. (b) Accumulated loss of 7-mers during ranking which cannot be ranked anymore due to cluster deletion. (c) All possible 7-mers and total number of unrankable 7-mers. (d) Percentage of 7-mers with a certain Hamming distance from the top ranked 7-mer TGACTCA that could be processed and ranked. Only less than half of all the 7-mers for each Hamming distance category can be ranked.

Figure 5.48 exemplifies very clearly the underlying issue in discarding DNA clusters. The three columns represent the k-mer sequence, the reverse complement and the ranking iteration. In the list of ranked k-mers shown in Figure 5.48 there occurs the 6-mer submotif AGTCAT, highlighted in yellow, which can be flanked to the left and right side by other nucleotides. The brown and green boxes display the extensions of the submotif to the left- and to the right-hand side respectively. The 7-mer motif AAGTCAT (ATGACTT), emphasized in red, cannot be ranked anymore even though it is a high affinity motif because there is no DNA cluster available anymore to provide an intensity for it. The reason is that previously in the ranking process all possible right-hand extensions by A, C, T and G of the submotif (large green box) were already ranked (left-hand side is “context-averaged” over A-T) and thus any DNA cluster is eliminated which could contain the submotif with an one nucleotide right-hand extension. It is the 7-mer motif AAGTCAT, with an extension of A to the left of the submotif AGTCAT,

that cannot be ranked anymore since all other left-hand extensions of the submotif by G, C and T (large brown box) occurred in the ranking (right-hand side is "context-averaged" over A-T) before all the possible right-hand extensions of the submotif were ranked and consumed all clusters containing AAGTCAT. As a result, the 7-mers:
GAGTCAT
CAGTCAT
TAGTCAT
occur in the ranking, but not
AAGTCAT
which makes it impossible to investigate systematically the effect of mutating the leftmost nucleotide position. All these 7-mers are important since they share the 5-mer submotif AGTCA with the top ranked 7-mer TGAGTCA.



k-mer	rev. comp.	ranking iteration
TGAGTCA	TGACTCA	1
TGACTAA	TTAGTCA	2
GAGTCAT	ATGACTC	3
CAGTCAT	ATGACTG	4
GAGTCAC	GTGACTC	5
ATGACTA	TAGTCAT	6
TGACGCA	TGCGTCA	7
TGTGTCA	TGACACA	8
AGTCATA	TATGACT	9
TGACTGA	TCAGTCA	10
TGAGTAA	TTACTCA	11
ATGACGT	ACGTCAT	12
ATGAGTC	GACTCAT	13
AGTCATC	GATGACT	14
TGTCATA	TATGACA	15
GTGACTA	TAGTCAC	16
GTGACTG	CAGTCAC	17
GTGTCAT	ATGACAC	18
TATGACG	CGTCATA	19
AATGACT	AGTCATT	20
TGACTCT	AGAGTCA	21
TGAATCA	TGATTCA	22
GATGACA	TGTCATC	23
CGTCATC	GATGACG	24
GTGACGT	ACGTCAC	25
AGTCATG	CATGACT	26
ATGACTT	AAGTCAT	

Figure 5.48: Discarding DNA clusters during the ranking process leads to unrankable k-mers. The three columns represent the k-mer sequence, the reverse complement and the ranking iteration. The submotif AGTCAT is marked in yellow. Possible right-hand extensions by A, C, T and G of the submotif are shown in the green box on the right. Left-hand extensions by G, C and T of the submotif are shown in the brown box on the right. The unrankable k-mer AAGTCAT is enframed in red. AAGTCAT cannot be ranked anymore because all clusters that contain this 7-mer are discarded due to previously ranked 7-mers that occurred in these clusters as well.

5.9.1.5 Heuristic ranking algorithm without cluster discharge

Since discarding clusters during the k-mer ranking process has the side effect of losing important k-mers due to the lack of clusters in which they can occur, one approach is to omit the deletion of clusters entirely. For their ranking algorithm, Nutiu et al. (2011) made the assumption that the entire intensity of a cluster containing the top k-mer comes from binding to that specific k-mer, and not to other k-mers embedded in the cluster. After removing clusters containing this k-mer, Nutiu et al. (2011) make the same assumption for the k-mer with the next highest median binding intensity in the remaining clusters. Not deleting any clusters means that they can be reused for the ranking of the following k-mers which could lead to falsely assigning intensities coming from higher ranked k-mers. Figure 5.49 illustrates this issue. Here, the two 11-mers
ATGAGTCATTG
and
TATATGAGTCA
share the submotif ATGAGTCA and occur in the same cluster sequence displayed in Figure 5.49.

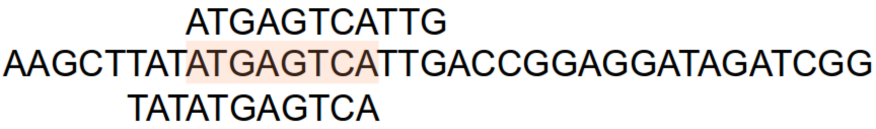


Figure 5.49: Overlap of 11-mers occurring in the same cluster sequence.

However, this problem becomes only apparent if higher ranked k-mers exclusively co-occur with lower ranked k-mers in the same cluster sequences. If a high ranked k-mer only co-occurs in a small portion of all the cluster sequences in which a lower ranked k-mer is found, then the influence of the high ranked k-mer is negligible. Using the median for robust averaging over the collected cluster intensities aids in diminishing the influence from high ranked k-mers. In addition, the longer the length of the ranked k-mers the less like it becomes to find clusters in which they co-occur. The longest length of k-mers which can be selected with a sufficient number of counts given our experiments is 11 nt (average count number is 30). Since it is well known that flanking nucleotides have a significant influence on the binding affinity of the TF, larger k-mers can more accurately capture the binding behavior.

The analysis in Figure 5.50 investigates for all ranked 11-mers with Kd $< 1\mu M$ (in total around 20000 11-mers), which share the 8 nt long submotif ATGAGTCA or its reverse complement TGACTCAT (resulting in 256 11-mers), how often they occur in the clusters and how large the fraction of any other 11-mer is that co-occurs with them. In Figure 5.50 the total number of clusters in which the 11-mers sharing the submotif ATGAGTCA or its reverse complement TGACTCAT occur are shown in light blue, and for each of these 11-mers the highest fraction of the co-occurring 11-mer is displayed in dark blue. For the 8 nt long overlapping submotif here the average number of clusters affected by the highest fraction of co-occurring 11-mers is 33%. Even for a 10 nt long overlapping submotif, the average number of clusters is only 35%. Therefore, there are always many more clusters in which the 11-mers do not occur together, thus allowing the 11-mers to differentiate their intensities.

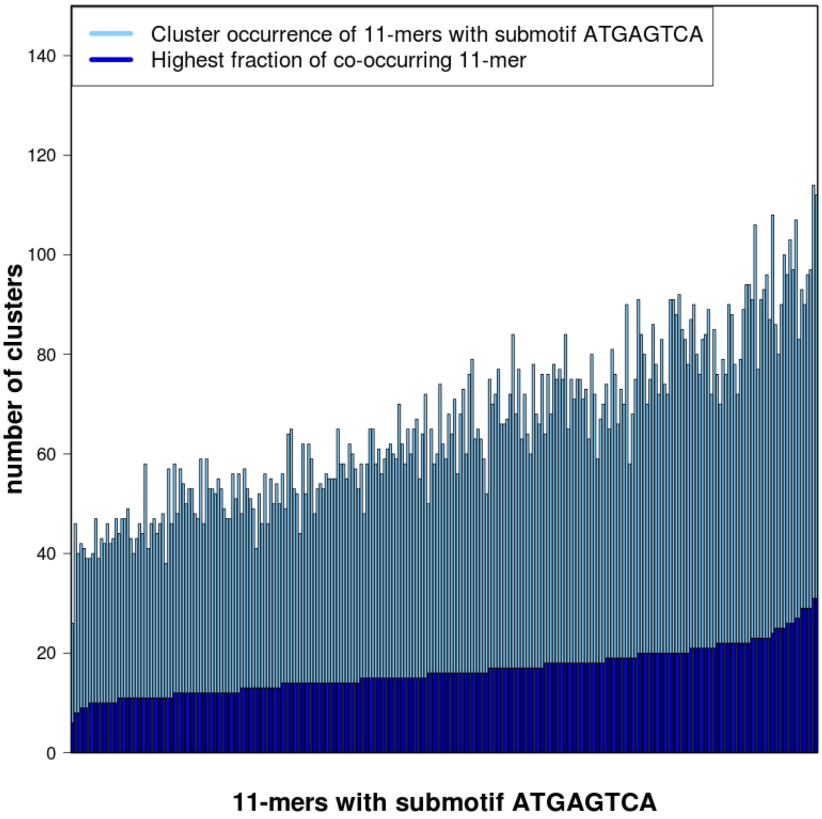


Figure 5.50: Co-occurrences of 11-mers in clusters.

5.9.1.6 Run time reduction by omitting cluster deletion

As shown by equation 5.51 the ranking procedure with cluster deletion is quadratic. Since cluster discharge is unnecessary as demonstrated by the analysis in subsection 5.9.1.5, the run time of the ranking procedure can be significantly decreased from a quadratic to a linear run time. With the $K \to C$ map as data structure, explained in section 5.9.1.1, a single iteration is sufficient to rank all k-mers.

5.9.1.7 Validation of ranking by Kds

The underlying data set for the analyses in this section are the data from lane 2 of experiment 18.08.2014. Preprocessing of the tif images was done as described in section 5.2, the cluster position transformation was carried out as described in section 5.3, the shifting of mapped clusters was performed as explained in section 5.4, normalization of the cluster intensities was executed as detailed in section 5.5, intensity extraction was performed as stated in section 5.6 , image outlier detection was carried out as described in section 5.7, and DNA cluster sequence filtering was applied as stated in section 5.8. The Hill based fitting for determining the Kds was done as described in section 5.10.

One way to validate the correctness of the ranking algorithm is by comparing the related Kds with the Kds measured by an alternative assay for a selection of k-mers covering the sequence space. The Kds obtained by the heuristic ranking algorithm without cluster discharge and subsequent affinity quantification were validated with a very sensitive, medium throughput fluorescence anisotropy (HiP-FA) assay, developed by C.Jung in the Gaul lab at the Gene Center (Jung et al. (2015)).

Anisotropy can be measured when a fluorescent molecule is excited with polarized light. The ratio of emission intensity in each polarization plane, parallel and perpendicular relative to the excitation polarization plane, gives a measure of anisotropy, often referred to as “fluorescence polarization” (FP) (Chen (2009)). This anisotropy is proportional to the Brownian rotational motion of the fluorophore and changes in anisotropy occur when the fluorescent small molecule binds to a much larger molecule affecting its rotational velocity (Chen (2009)).

The HiP-FA assay utilizes 396 well plates and provides a measure of the rotational speed of a fluorescently labeled species, which are DNA oligomers here. GCN4 fused with mOrange was used just like for a HiTS-FLIP experiment and its binding to DNA increases the molecular weight and thereby decreases the rotational speed of the fluorescently labeled DNA oligomer, resulting in increased FA (Jung et al. (2015)). The HiP-FA assay is competitive, in which TF and Cy5-labeled reference DNA are mixed at fixed concentrations and embedded together into an agarose gel. In the titration wells of the plate, the TF concentration is in molar excess over the labeled reference DNA, thereby ensuring its complete binding to protein (Jung et al. (2015)). The reference DNA is labeled with Cy5, a dye that proves well suited for FA measurements. Unlabeled competitor DNA is added on top of the agarose and establishes a concentration gradient throughout the gel, whose shape changes over time (Jung et al. (2015)). As the competitor DNA diffuses through the matrix it competes with the Cy5-reference DNA for binding to the TF, resulting in a dynamically changing FA signal of the Cy5-reference DNA. This process allows to measure, over time, a continuous titration series within a single well and results in hundreds of measurement points for fitting binding curves and determining Kds. Figure 5.51 provides an example for the measurement points and fitted binding curve regarding ATGACTCA embedded in the oligomer GGTATGACTCATGGCC. The detection of binding constants of the HiP-FA assay lies within the range of 10^{-10} to 10^{-3} molar.

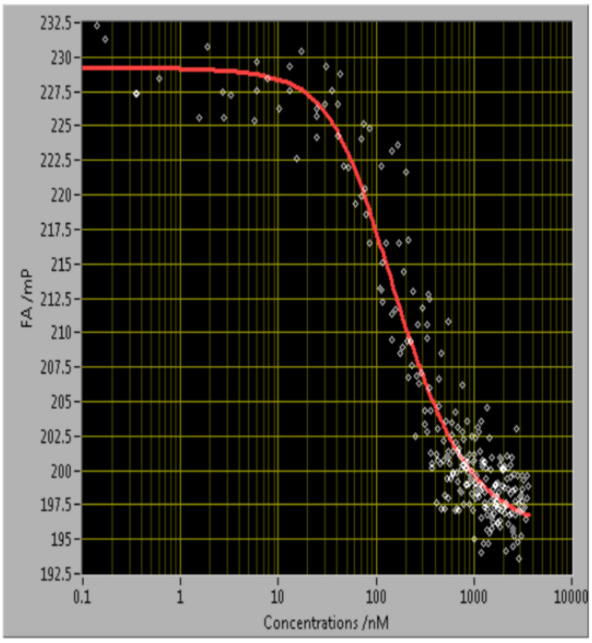


Figure 5.51: Measurement of ATGACTCA embedded in the oligo GGTATGACTCATG-GCC by the HiP-FA assay (Kd=21.66 nM). The white colored dots denote measurement points and the red line is the fitted binding curve.

There is an excellent agreement between HiP-FA Kds and HiTS-FLIP Kds for the heuristic ranking algorithm without cluster discharge with a Pearson product-moment correlation coefficient R=0.99 and a relative error δ=30.91% which is shown by the correlation plot of Figure 5.52 for 25 11-mers with a Kd range from 3.49 nM to 875.36 nM. The Appendix lists the details on the HiP-FA and HiTS-FLIP Kds (9.4), and the fits and parameters of the HiTS-FLIP Kds (9.8).

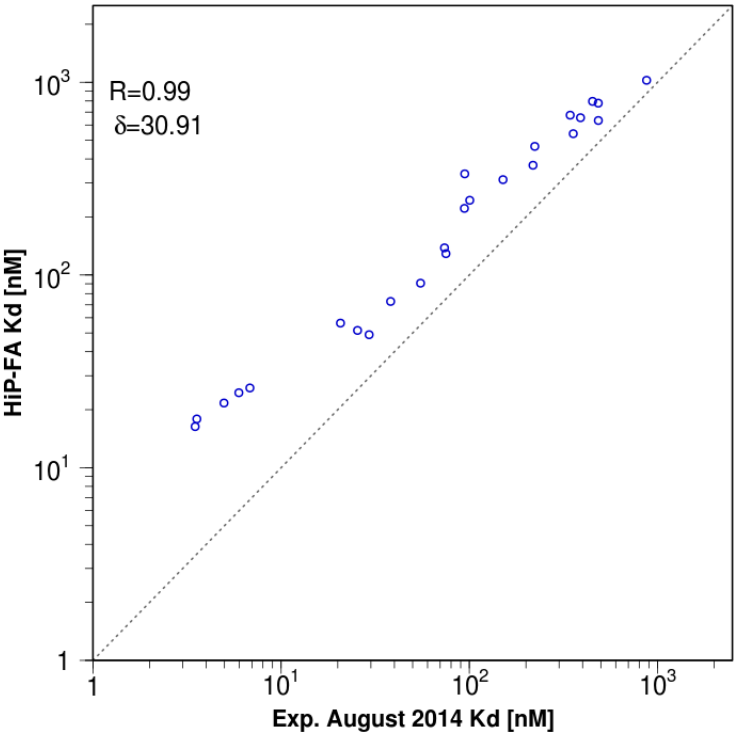


Figure 5.52: Validation of HiTS-FLIP Kds with HiP-FA Kds.

5.9.2 Maximum likelihood based ranking

5.9.2.1 Notation

The following notation is introduced for describing the probabilistic model in this chapter.
$n \in \{1, ..., N\}$: Cluster index over all N cluster. N can be here between 80.000 and 350.000 clusters per tile. That is between $120 * 80.000 = 9.600.000$ and $120 * 350.000 = 42.000.000$ per lane, and between $7 * 9.600.000 = 67.200.000$ and $7 * 42.000.000 = 294.000.000$ per flow cell, i.e. $N \approx 10^7$ or 10^8.
K: Number of *k-mers* to rank. 4^l where l is the length of the *k-mer*, 4 denoting the four nucleotides.
$w \in \{1, ..., 4^k\}$: *k-mer* word index for the different *k-mers*.
I_n^{exp}: Experimentally measured intensity of cluster n.
I_n^{pred}: Predicted intensity of cluster n by probabilistic model.
S_n: Sequence S of cluster n.
ϕ_w: Contribution of the word w to the intensity of a particular cluster containing w by specific binding (proportional to the occupancy of the TF on this word). Each ϕ_w is a parameter whose value has to be learned from the measured data.
$\vec{\phi}$: Vector of all parameters ϕ_w, i.e. all embedded words w of a cluster n.
ϕ_{ub}: Single parameter capturing the contribution to cluster intensity by unspecific binding of the TF.
$W \subseteq \{1, ..., 4^k\}$: Set of words that are potentially contributing to the TF binding.
$W_n \subseteq W$: Set of words from W that occur as substring in the sequence S_n of cluster n: $W_n = \{w \in W \mid w \subseteq S_n\}$
$M = \sum_{n=1}^{N} |W_n|$: Number of words to consider when running through all N clusters.

5.9.2.2 Bayesian approach to ranking k-mers

In order to avoid deletion of clusters and thereby losing important k-mers in the ranking, a probabilistic model based on Bayes' theorem can be applied. A Bayesian based machine learning scheme infers each contribution of the embedded k-mer to the related cluster intensity and yields the highest, estimated likelihood for the k-mer intensity. Bayes' theorem has the following general form (Weisstein (2009)):

$$P(\text{model}|\text{data}) = \frac{P(\text{data}|\text{model}) \cdot P(\text{model})}{P(\text{data})} \tag{5.53}$$

Regarding HiTS-FLIP experiments, the data consist of all the measured cluster intensities

I_n^{exp} and the model is composed of all the ϕ_w. Thus, the expression becomes:

$$p(\phi_w|I_n^{exp}) = \frac{p(I_n^{exp}|\vec{\phi}) \cdot p(\phi_w)}{p(I_n^{exp})} \tag{5.54}$$

$p(\phi_w|I_n^{exp})$, i.e. the probability for the intensities of the embedded *k-mers* (words w) in the cluster sequence given the measured intensity of cluster n, is calculated by Bayes' rule:
$p(I_n^{exp}|\vec{\phi})$: the probability of measuring the intensity I of a cluster n given its sequence embedded set of *k-mers* (words w).
$p(\phi_w)$: the probability of the *k-mer* (word w) to contribute to a certain cluster intensity.
$p(I_n^{exp})$: the probability of measuring the intensity I for a certain cluster n.

The parameters of the right-hand side of equation 5.54, i.e. ϕ_w, can be approximated by maximum-likelihood estimation (MLE). The principle of maximum likelihood yields a choice for the values of all the ϕ_w that makes the observed data, the measured cluster intensities I_n^{exp}, most probable. The MLE can be obtained by maximizing the objective function or by minimizing the negative objective function. For numerical stability the logarithm is usually taken of the objective function. The objective function can be simplified to *posterior probability* $\propto$ *likelihood* $\cdot$ *prior probability*:

$$p(\vec{\phi}|I_n^{exp}) \propto p(I_n^{exp}|\vec{\phi}) \cdot p(\phi_w) \tag{5.55}$$

The objective function can be iteratively optimized by L-BFGS-B (Limited Memory Boxed BFGS) (Byrd et al. (1995)), an limited-memory extension of the BFGS (Broyden–Fletcher–Goldfarb–Shanno) algorithm (Broyden (1970); Fletcher (1970); Goldfarb (1970); Shanno (1970)) with simple bound constraints of the form $l_i \leq x_i \leq u_i$ where l_i and u_i are per-variable constant lower and upper bounds. The bounds were utilized here for the ranking of the k-mers in order to enforce that the intensities cannot be negative.

5.9.2.3 Probabilistic model

The following equations make up the model.

$$I_n^{pred} = \sum_{w \in W_n} \phi_w + \phi_{ub} \tag{5.56}$$

The underlying assumption is that the predicted intensity of a cluster n is the sum of all embedded *k-mer* intensities plus the contribution of some unspecific binding.

$$p(\vec{\phi}|I_n^{exp}) \propto p(I_n^{exp}|\vec{\phi}) \cdot p(\phi_w) \tag{5.57}$$

$$\prod_{n=1}^{N} p(\vec{\phi}|I_n^{exp}) \propto \prod_{n=1}^{N} p(I_n^{exp}|\vec{\phi}) \cdot \prod_{w \in W} p(\phi_w) \tag{5.58}$$

$$NLP = -\log\big(\prod_{n=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(I_n^{exp}-I_n^{pred})^2}{\sigma^2}} \cdot \prod_{w \in W} (e^{-C\phi_w} \cdot H(\phi_w))\big) \tag{5.59}$$

$$= \sum_{n=1}^{N} \big(-\log(\frac{1}{\sigma\sqrt{2\pi}}) + \frac{1}{2\sigma^2} \cdot (I_n^{exp} - I_n^{pred})^2\big) + C \sum_{w \in W} \big(\phi_w + \infty \cdot I(\phi_w < 0)\big) \tag{5.60}$$

$$= -N\log(\frac{1}{\sigma\sqrt{2\pi}}) + \frac{1}{2\sigma^2} \sum_{n=1}^{N} (I_n^{exp} - I_n^{pred})^2 + C \sum_{w \in W} \big(\phi_w + \infty \cdot I(\phi_w < 0)\big) \tag{5.61}$$

$$= -N\log(\frac{1}{\sigma\sqrt{2\pi}}) + \frac{1}{2\sigma^2} \sum_{n=1}^{N} (I_n^{exp} - \sum_{w \in W_n} \phi_w - \phi_{ub})^2 + C \sum_{w \in W} \big(\phi_w + \infty \cdot I(\phi_w < 0)\big) \tag{5.62}$$

The likelihood, how the cluster intensity is generated by the embedded *k-mers*, is modeled by a Gaussian distribution (I_n^{pred} represents μ as the expected value). The prior constrains the *k-mer* intensities to be positive by the unit step function H and models the expected distribution of the *k-mer* intensities in such a way that small values are predominantly likely.

$$\frac{\partial NLP}{\partial \phi_w} = \frac{1}{\sigma^2} \sum_{n=1}^{N} \big((I_n^{exp} - \sum_{w \in W_n} \phi_w - \phi_{ub}) \cdot I(w \in W_n)\big) + C \cdot 1 \tag{5.63}$$

$$\frac{\partial NLP}{\partial \phi_{ub}} = \frac{1}{2\sigma^2} \sum_{n=1}^{N} \big(2 \cdot (I_n^{exp} - \sum_{w \in W_n} \phi_w - \phi_{ub}) \cdot (-1)\big) \tag{5.64}$$

$$= -\frac{1}{\sigma^2} \sum_{n=1}^{N} \big(I_n^{exp} - \sum_{w \in W_n} \phi_w - \phi_{ub}\big) \tag{5.65}$$

$$\frac{\partial NLP}{\partial \sigma} = \frac{N\sqrt{2\pi}}{\sigma} - \frac{1}{\sigma^3}\sum_{n=1}^{N}(I_n^{exp} - \sum_{w \in W_n} \phi_w - \phi_{ub})^2 \tag{5.66}$$

$$\frac{\partial NLP}{\partial \phi_\sigma} = 0 \tag{5.67}$$

$$\frac{N\sqrt{2\pi}}{\sigma} - \frac{1}{\sigma^3}\sum_{n=1}^{N}\left(I_n^{exp} - \sum_{w \in W_n} \phi_w - \phi_{ub}\right)^2 = 0 \tag{5.68}$$

$$\sigma^2\sqrt{2\pi} = \frac{1}{N}\sum_{n=1}^{N}\left(I_n^{exp} - \sum_{w \in W_n} \phi_w - \phi_{ub}\right)^2 \tag{5.69}$$

$$\sigma = \sqrt{\frac{1}{N\sqrt{2\pi}}\sum_{n=1}^{N}\left(I_n^{exp} - \sum_{w \in W_n} \phi_w - \phi_{ub}\right)^2} \tag{5.70}$$

5.9.2.4 Implementation

Together with Armin Meier the ML based ranking procedure was implemented in C++ (Stroustrup (1986)). As optimizer L-BFGS-B was applied using the Fortran library by Zhu et al. (1997) together with the included C wrapper. For parallelization OpenMP (Dagum and Enon (1998)) was used.
The input are

- the length of the k-mer to be ranked
- the concentration at which the ranking should be achieved
- the cluster sequences
- the different cluster intensities for the increasing concentrations

The output are the ranked k-mer motifs with their different intensities.

5.9.2.5 Run time complexity

The run time complexity for one iteration is given by:

$$\begin{aligned} &\mathcal{O}(K) + \mathcal{O}(N) + \mathcal{O}(K) + \mathcal{O}(K \cdot M) \\ &\qquad = \mathcal{O}(K + N + KM) \end{aligned} \tag{5.71}$$

since all *k-mers* K have to be updated after each iteration, σ is evaluated for all clusters N, the gradient for phi_w and for phi_{ub} is recalculated for all *k-mers* K, and finally the run time complexity of L-BFGS-B for computing the direction p_k and x_{k+1} is $\mathcal{O}(K \cdot M)$, where M are the last input and gradient differences (usual values are 10 or 20). Since $M \ll K$ the run time complexity becomes:

$$\mathcal{O}(K + N) \tag{5.72}$$

Figure 5.53 shows the execution time for ranking one 11-mer, using around 12 million DNA cluster sequences of length 25 nucleotides. The purple dots at 1 and 4 CPU cores display execution times measured on an Intel Core i5-200K quad-core processor with 16 GB RAM. The gained parallel speed up here is 1.79. The blue triangles represent inferred execution times.

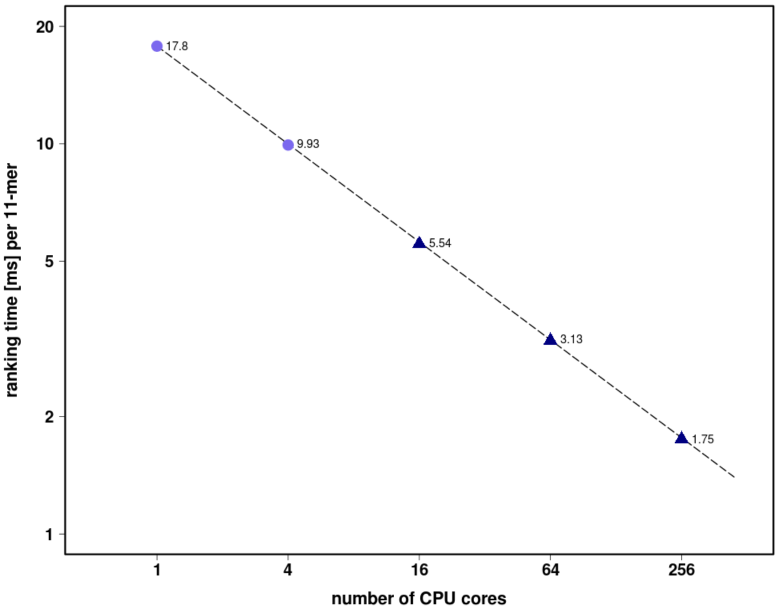


Figure 5.53: Speedup of the probabilistic ranking algorithm from single processor to multiple processors. The purple dots are measured execution times, the blue triangles represent inferred execution times.

There are $4^{11} = 4194304$ possible 11-mers. Ranking all possible 11-mers using a computing cluster with 64 CPU cores would last around 3 hours and 37 minutes, and with 256 CPU cores around 2 hours.

5.9.3 Discussion

The two different ranking methods, the heuristic ranking algorithm without cluster discharge and the maximum likelihood (ML) based ranking method, mainly differ in their underlying assumptions. The heuristic ranking algorithm assumes only one binding site per DNA cluster sequence to which a GCN4 molecule can bind and thus generate the related cluster intensity. The ML based ranking method includes in its modeling that the cluster intensity is produced by multiple k-mers embedded in the cluster sequence to which GCN4 molecules can bind, thereby partitioning the cluster intensity among its constituent k-mers in proportion to their binding affinity. In order to compare these different ranking approaches, the related Kds were validated by the HiP-FA assay (Jung et al. (2015)).

The underlying data set for the analyses in this section are the data from lane 2 of experiment 18.08.2014. Preprocessing of the tif images was done as described in section 5.2, the cluster position transformation was carried out as described in section 5.3, the shifting of mapped clusters was performed as explained in section 5.4, normalization of the cluster intensities was executed as detailed in section 5.5, intensity extraction was performed as stated in section 5.6 , image outlier detection was carried out as described in section 5.7, and DNA cluster sequence filtering was applied as stated in section 5.8. Ranking was performed with respect to the heuristic ranking algorithm without cluster discharge as described in subsection 5.9.1.5. The ML based ranking was carried out as described in section 5.9.2. The Hill based fitting for determining the Kds was done as described in section 5.10.

Kds were correlated in Figure 5.54. The ML ranking was performed for all quadruple mutations in relation to the top ranked 11-mer TATGACTCATA (TATGAGTCATA). Matching this data set with the 25 sequences and HiP-FA validated Kds resulted in 25 11-mers with a Kd range from 3.0 nM to 1171.25 nM. Using these 25 11-mers the Kds for the heuristic ranking without cluster deletion as well as for the ML based ranking have an excellent agreement with the HiP-FA measured Kds. The Appendix section 9.7 lists the different Kds. The correlation is slightly higher regarding the heuristic ranking without cluster deletion than the ML based ranking ($R = 0.99$ versus $R = 0.97$). Since these two different ranking approaches lead to very similar rankings as demonstrated by the Kd based correlation with an alternative, highly sensitive assay, it can be concluded that the assumption of one binding site per cluster sequence of length 25 bp as given by the experimental design is sufficient and the heuristic ranking without cluster deletion is a good working solution for the k-mer ranking. This finding is in agreement with a recent publication by Levo et al. (2015b), which examined the length of the binding site of GCN4 and showed that the known 9 bp binding site (ATGACTCAT, (Hill et al. (1986))) together with 3 bp flanks best captured the differential binding of GCN4 observed in their measurements. Given the length of the GCN4 binding site (up to 15 bp as shown by (Levo et al. (2015b))) and additional steric hindrance, it seems very likely that only one GCN4 molecule can bind to a DNA cluster sequence of 25 bp length.

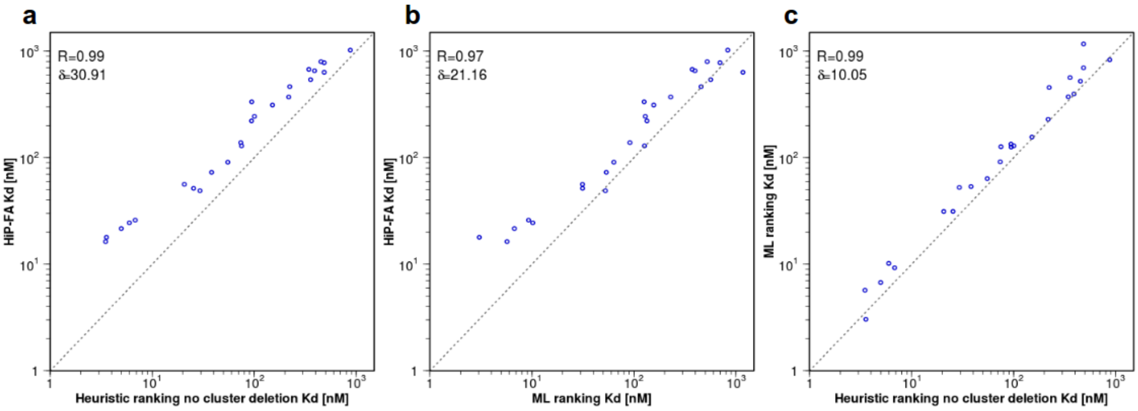


Figure 5.54: Comparison of Kd correlation for heuristic ranking without cluster deletion and ML based ranking. The ML ranking was performed for all quadruple mutations in relation to the top ranked 11-mer TATGACTCATA (TATGAGTCATA). Matching this data set with the 25 sequences and HiP-FA validated Kds resulted in 25 11-mers with a Kd range from 3.0 nM to 1171.25 nM. (a) Correlation of Kds for heuristic ranking without cluster deletion with HiP-FA Kds ($R = 0.99$,$\delta = 30.91\%$). (b) Correlation of Kds for ML based ranking method with HiP-FA Kds ($R = 0.97$,$\delta = 21.16\%$). (c) Correlation of Kds for heuristic ranking without cluster deletion with ML based ranking ($R = 0.99$,$\delta = 10.05\%$).

5.10 Affinity quantification

For determining the Kd of a certain k-mer the Hill equation (Hill (1910)) shown in equation 5.73 was used and fitted to the measured and median averaged fluorescent intensities of the k-mer.

$$F_{obs} = s \times \frac{[TF]^h}{[TF]^h + Kd^h} + o \tag{5.73}$$

where
F_{obs}: are the observed fluorescent intensities for a particular k-mer, averaged over all clusters containing this k-mer, for the different TF concentrations used in the experiment.
s: is a scaling factor obtained from the top binding k-mer and applied for all weaker binding k-mers.
o: is a global offset used for all k-mers.
$[TF]$: TF concentration used in the experiment.
Kd: is the dissociation constant of the TF to the DNA sequence.
h: is the Hill coefficient of binding.

The fitting procedure is as follows:
1) Intensities are transformed so that they are in the range $[0 - 1]$, done by dividing all intensities by the greatest intensity at the greatest concentration. For the experiment performed at 18.08.2014 this is the greatest intensity at 625 nM.
2) All k-mers are sorted by their intensity at 125 nM in descending order.
3) A global offset o is subtracted from the intensities of each k-mer. This offset is an estimate for the unspecific binding by using the median of the dimmest 0.1% of all the ranked k-mers at the smallest concentration. Regarding the experiment 18.08.2014 the value was $o = 0.26793$ using the ranked 11-mers for the concentration at 5 nM.
4) The scaling factor s is determined for the first ranked k-mer at 125 nM. This scaling factor is then fixed and applied for all other k-mers in the same way as done by Nutiu et al. (2011).
5) The Hill equation 5.73 with the fixed scaling factor s is applied for all k-mers for getting the Kds.

5.10.1 Implementation

I used the nls function (Bates and Chambers (1992); Bates and Watts (1988)) from the R stats package for estimating the parameters of the Hill equation by nonlinear least-squares (Fox and Weisberg (2010)). As nls option the "port" algorithm was selected which allows bounds constraints and uses a quasi-Newton method. The Kd was constrained to be ≥ 1, the Hill coefficient was constrained to be between $[2 \geq h \geq 1]$ and the scaling factor s was constrained to be ≥ 0.1 for the top k-mer and fixed for all lower ranked k-mers.

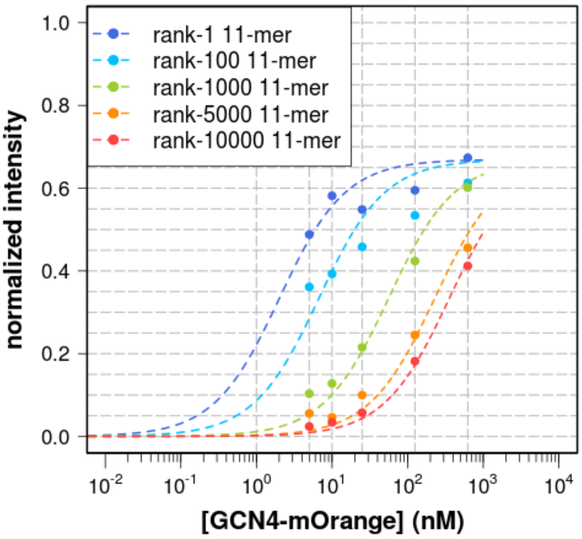
5.10.2 Binding curves

Figure 5.55 shows binding curves for selected 11-mers, the Appendix lists the first 50 11-mers and their Kds (see 9.5), and the first 50 11-mers ranked by intensity at 125 nM (see 9.6). The underlying data set for Figure 5.55 are the data from lane 2 of experiment 18.08.2014. Preprocessing of the tif images was done as described in section 5.2, the cluster position transformation was carried out as described in section 5.3, the shifting of mapped clusters was performed as explained in section 5.4, normalization of the cluster intensities was executed as detailed in section 5.5, intensity extraction was performed as stated in section 5.6 , image outlier detection was carried out as described in section 5.7, and DNA cluster sequence filtering was applied as stated in section 5.8. Ranking was performed using the heuristic ranking algorithm without cluster discharge as described in subsection 5.9.1.5. The Hill based fitting for determining the Kds was done as described above with the offset $o = 0.26793$ as an estimate for the unspecific binding.

It can be noticed that the fluorescent intensities do not follow perfect saturation curves. This aspect is a lesser issue since the intensities for the different 11-mers differentiate themselves appropriately relative to one another leading to increasing Kds in proportion to decreasing binding affinity as validated in subsection 5.9.3 by the excellent correlation with the HiP-FA assay (Jung et al. (2015)).

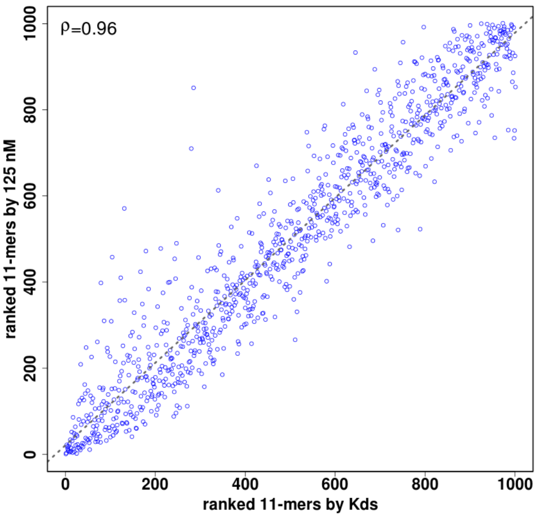
An additional validation is provided by Figure 5.56 which shows a very high correlation of motifs ranked by intensity and by Kd from experiment 18.08.2014.

This result demonstrates that the decrease in binding affinity, which is measured by the decreasing fluorescent intensities from the DNA clusters, is well captured by the related Kds. This even suggests that relative affinities based on intensities alone could describe the binding behavior sufficiently.



.

Figure 5.55: Binding curves for 11-mers from experiment 18.08.2014. The processing of the data is described in section 5.10



.

Figure 5.56: Correlation of motifs ranked by intensity and by Kd from experiment 18.08.2014 which results in a Spearman's rank correlation coefficient of $\rho = 0.96$. The processing of the data is described in section 5.10

5.10.3 Discussion

Using a global offset is based on the observation that at a low concentration level, all clusters are predominantly bound unspecifically. At higher concentration levels, clusters with high affinity binding sites are bound by a higher amount of protein at the high affinity binding sites than clusters with low affinity binding sites where binding still occurs largely unspecifically. Subtracting the median taken over all cluster intensities per concentration level results in a decrease of intensities for the top ranked 11-mers at the highest concentration 125 nM and 625 nM as shown in Appendix section 9.9.

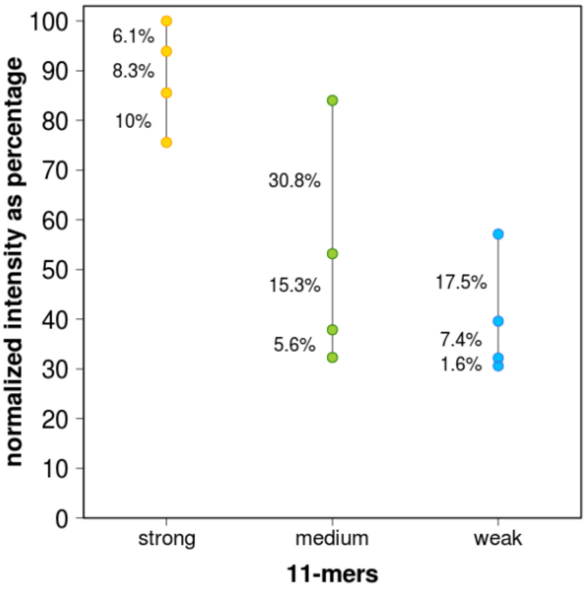


Figure 5.57: Intensity as percentage for three selected 11-mers with Kd=2.8 nM, Kd=172.9 nM and Kd=765.5 nM. The annotation shows the relative increase in intensity from 5 nM to 25 nM, 25 nM to 125 nM, and 125 nm to 625 nM.

As shown in Figure 5.57 the relative intensity increase for a strong binder (CGATGACTCAC, Kd=2.8 nM) is much lower then for a medium (ATTTGTCATAA, Kd=172.9 nM) or weak binder (CGTCACCCCAT, Kd=765.5 nM). This outcome highlights that clusters with high affinity binding sites cannot be normalized for unspecific binding in the same way as clusters with medium or low affinity binding sites at higher concentrations.

6 Experiments

6.1 Protein expression and purification

The GCN4+mOrange plasmid from the Burge Lab at the MIT, described in (Nutiu et al. (2011)), was received and used. According to (Nutiu et al. (2011)), a carrier vector was used to build the GCN4 fused to mOrange sequence. The GFP-GCN4 sequence was PCR amplified from the vector pME2126 provided by the G. H. Braus laboratory and inserted in the carrier vector between the Bgl2 and Not1 restriction sites, with a Spe1 site between the GFP and GCN4 coding sequences. Consequently mOrange, missing the stop codon (Clontech), was introduced in place of GFP. The whole mOrange-GCN4 sequence was then cloned into the pET151/D-TOPO vector (Invitrogen) according to the manufacturer's instructions. The final construct generated a 6xHis-mOrange-GCN4 fusion gene that was verified by sequencing and transformed into BL21Star bacteria (Invitrogen). Protein production was induced by 1 mM IPTG for 4 h at 37°C. The protein was purified using a Ni-NTA Fast Start kit (Qiagen) following the manufacturer's protocol. The purity of the protein was verified on a NuPAGE 10% Bis-Tris Urea gel (Invitrogen).

6.2 Library design

The library design was performed as described in (Nutiu et al. (2011)). pChip_bot_R and pChip_top_R were annealed to form adaptor R, and pChip_bot_L and pChip_top_R were annealed to form adaptor L. The samples were heated up at 95°C in a heat block for 5 min, and then the heat block was left to cool down to 25°C. The library pChip_N25 was ligated (25°C; 20 min) to the adaptors R and L. The library was PCR amplified (12 cycles). The PCR product was purified on a 6% TBE PAGE gel. The ~135 bp band was eluted out from the gel, ethanol precipitated and quantified by Bioanalyzer.
Figure 6.1 shows the complete sequence of a DNA cluster in the flow cell which is 150 nucleotides long. The insert is 36 nucleotides long, with two adapter sequences at each end (3 nucleotides: AAG, and 8 nucleotides: TAGATCGG) which leave 25 nucleotides

for the variable region.

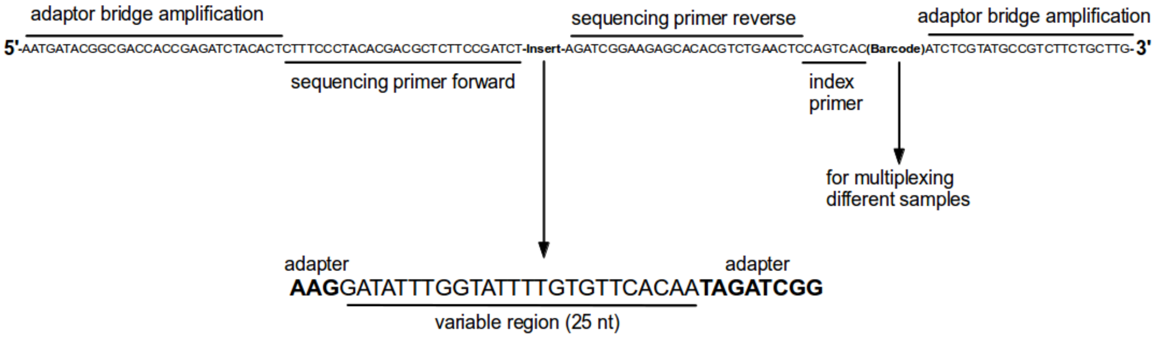

adaptor bridge amplification
sequencing primer reverse
adaptor bridge amplification
5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-Insert-AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC(Barcode)ATCTCGTATGCCGTCTTCTGCTTG-3'
sequencing primer forward
index primer
for multiplexing different samples
adapter
adapter
AAGGATATTTGGTATTTTGTGTTCACAATAGATCGG
variable region (25 nt)

Figure 6.1: Complete DNA cluster sequence.

6.3 Cluster generation, linearization, blocking and primer hybridization

The DNA cluster generation, linearization, blocking and primer hybridization was accomplished following (Nutiu et al. (2011)). The DNA clusters were grown using the Illumina standard protocol, starting from ~3–4 pM template to give a density of ~150000-200000 DNA clusters per tile. The DNA clusters were linearized and blocked using standard protocol. The sequencing primer was hybridized using standard protocol.

6.4 DNA Sequencing

The DNA sequencing was performed in the following way. 36 cycles of sequencing were performed using standard protocol. At the end of sequencing a final cleavage step was added.

6.5 GA-IIx modifications

The GA-IIx was modified for allowing lane-by-lane control as described in (Gravina et al. (2013)).

6.6 XML recipe modifications

According to (Nutiu et al. (2011)) to avoid the delivery of scan mix before protein imaging, the ImageCyclePump.xml config file (C:\Illumina\SCS2.6\DataCollection\bin\Config) was modified as follows:
“<ImageCyclePump On=”true” AutoDispense = ”false”>”
was changed to
“<ImageCyclePump On=”false” AutoDispense = ”false”>”.
The different HiTS-FLIP experiments were automatically carried out by special XML recipes for delivering the varying protein amounts at different concentrations into the flow cell, applying equilibration time and performing imaging.

6.7 Data processing

For the analyses of the following experiments the data was processed in the subsequent manner. Preprocessing of the tif images was done as described in section 5.2, the cluster position transformation was performed as described in section 5.3, the shifting of mapped clusters was carried out as explained in section 5.4, normalization of cluster intensities by local background subtraction was executed as detailed in section 5.5, intensity extraction was performed as stated in section 5.6, image outlier detection was carried out as described in section 5.7, and DNA cluster sequence filtering was applied as stated in section 5.8. The ranking procedure was applied as explained in subsection 5.9.1.5. CIF intensities were produced by the Illumina pipeline (SCS version 2.10 and RTA vrsion 1.13).

6.8 Listing of GCN4 HiTS-FLIP experiments

The following listing shows all the performed HiTS-FLIP experiments using GCN4 and mOrange.

- Experiment by Nutiu et al.: Details in Appendix section 9.12.
- Experiment 03.04.2013: Details in Appendix section 9.13.
- Experiment 13.06.2013: Details in Appendix section 9.14.
- Experiment 28.03.2014: Details in Appendix section 9.15.
- Experiment 11.08.2014: Details in Appendix section 9.16.
- Experiment 18.08.2014: Details in Appendix section 9.17.
- Experiment 12.02.2015: Details in Appendix section 9.18.
- Experiment 06.03.2015: Details in Appendix section 9.19.
- Experiment 14.04.2015: Details in Appendix section 9.20.

6.9 Discussion

Reviewing all the experiments and the data of lane 2 from Nutiu et al. (2011) makes apparent that the experiment from 18.08.2014 has been overall the most successful result. In the intensity course for this experiment I can observe a saturation occurring for the best 8-mer binding motif ATGACTCA (TGAGTCAT) and a steep decline in intensities over the first 20 ranks that flattens out into unspecific binding. The binding curve for the top 8-mer demonstrates saturation as I would expect it since the binding sites should all be eventually occupied fully - if no artificial aggregation happens. Weaker binding sites that appear at lower rank order do not display any saturation. The half-sites TGA(C/G) which are bound by dimeric GCN4 are strongly enriched and a first decrease is observable towards the 200th rank. Similarly, the Hamming distance from the best 8-mer binder, ATGACTCA (TGAGTCAT), increases in general towards the 200th rank since the more dissimilar the motifs are compared to the best binder, the weaker the binding affinity and the further down in the ranking they occur. The striking antagonistic peaks, correlating with the motifs (GA)GTGT, are discussed in section 7.5 in more detail.

The HiTS-FLIP experiment by Nutiu et al. (2011) can be regarded as the second best result since it shows the attributes connected to a valid outcome as discussed for the experiment from 18.08.2014 however with less specificity and with the application of a washing step after each pump-in of protein concentration. At higher concentrations 125 nM and 625 nM the decline in intensity is much less steep, and the overall intensity is a factor around 5-6 higher than in the experiment from 18.08.2014. The other diagnostic plots reveal a similar trend to the experiment from 18.08.2014.

What is the reason for the successful experiment from 18.08.2014? Why did all the other experiments fail? There are the following differences between the experiments from 03.04.2013, 13.06.2013, 28.03.2014 and the experiments from 11.08.2014 and 18.08.2014.

1) Use of different primer

A different primer, i.e. the Illumina read 1 sequencing primer, has been applied for a more efficient resynthesis of the second DNA strand. An ineffective resynthesis leading to a decreased amount of dsDNA after the sequencing step in the HiTS-FLIP protocol is the likely cause for the massive unspecific binding, i.e. all DNA clusters and related 8-mer motifs are bound equally resulting in an uniform intensity as exemplified with the intensity course of the experiment from 06.03.2013. If there is less dsDNA than the specific binding of GCN4 must be drastically reduced. Therefore, much lower intensity levels should be displayed in the intensity course plot. In contrast, much higher intensities can be observed, for the concentrations at 125 nM and 625 nM by more than a factor of 10.

Thus, there is not only a reduction in specific binding but a massive increase in unspecific binding that affects various different DNA clusters. This could arise from the electrostatic interaction between the negatively charged phosphate groups of the DNA backbone and the positively charged GCN4 molecules. Another possibility is that single-stranded DNA or at least only partially double stranded DNA exhibits a higher flexibility than fully double stranded DNA, and thus the close proximity in the region of amplified DNA clusters could lead to inter-chain base pairing forming a mesh which provokes unspecific aggregation of GCN4 molecules. Moreover, intra-chain base pairing and base stacking interactions of single stranded DNA could also form individual structures which decrease specific binding.

2) Change of flow cell buffer

In order to avoid nonspecific interaction of proteins with the surface of the flow cell reaction chamber, BSA (bovine serum albumin) is commonly used as blocking reagent derived from the serum of cows. In addition, Tween-20 (Polysorbate-20) is a non-ionic detergent and can help to prevent nonspecific binding. DNA-protein interactions is mostly driven by interactions with phosphates and in order to avoid breaking these interactions by too much salt 150 mM NaCl was used here. In the experiments from 11.08.2014 and 18.08.2014 additionally MgCl2 and KCl was added to the running buffer. Magnesium ions prevent nonspecific electrostatic interactions between protein molecules and DNA in solution and thus enhances sequence-specific DNA binding. An additional effect is achieved by potassium chloride (KCl) ions. In Moll et al. (2002), it was shown that magnesium and potassium chloride ions prevent nonspecific electrostatic interactions between CREB and DNA in solution.

3) Quality control for dsDNA resynthesis

In the experiments from 11.08.2014 and 18.08.2014, a new quality control was integrated to test the resynthesis of the second DNA strand. A primer (0.01 μM) with an Alexa-like dye (detectable in the C channel) was hybridised to the flow cell primer oligos before the resynthesis. This primer should be displaced by Klenow polymerase if the resynthesis occurs at the related DNA cluster.

4) Different fluidics setup

In (Gravina et al. (2013)), the authors describe how the GA-IIx can be modified to enable lane-by-lane sequencing. For the experiment from 11.08.2014 (and the following experiments) this modification was carried out.

The differences between the experiments from 12.02.2015, 06.03.2015 and 14.04.2015 compared to the experiment from 18.08.2014 were the following:

1) Ten concentrations

For the experiments 12.02.2015, 06.03.2015 and 14.04.2015 ten concentrations instead of five were used. In the experiments from 06.03.2015 and 14.04.2015 after the first five concentrations the second DNA strand of the clusters was denatured and resynthesized.

2) Higher concentration of the fluorescent primer for dsDNA quality check

In the experiments from 11.08.2014 and 18.08.2014, only 0.01 μM of the primer with the Alexa-like dye was used. In the experiments from 12.02.2015, 06.03.2015 and 14.04.2015 0.1 μM was used.

6.10 Comparison with experiment from 18.08.2014

In the following, the experiments from Nutiu et al. (2011), 13.06.2013, 12.02.2015, 06.03.2015 and 14.04.2015 are compared with the experiment from 18.08.2014 regarding the behavior of unbound GCN4, the behavior of unspecific GCN4 binding, cluster density per tile and the template density per cluster.

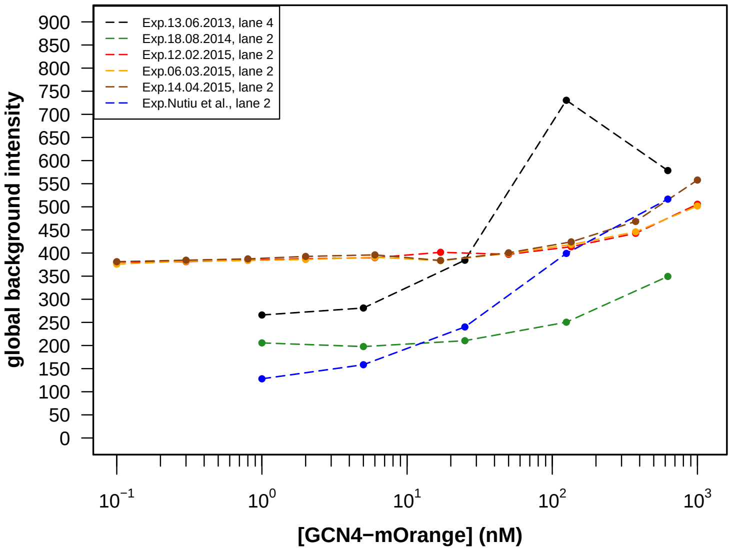


Figure 6.2: Behavior of unbound GCN4 compared for different experiments.

The global background in Figure 6.2 is calculated by dividing each tile image into 32 x 32 pixel regions, taking the mean of the dimmest 20 pixels and then taking the median over all these region means. These tile medians are averaged over the entire lane and produce the global background for each concentration step. The strongest increase in unbound GCN4 molecules occurs for the experiment from 13.06.2013. I have no explanation for the drop in intensity at the highest concentration 625 nM. The experiment from 18.08.2014 as well as the experiments from 2015 show only a slight increase in unbound GCN4 molecules related to the starting amount of unbound GCN4 molecules. One possible explanation for seeing the highest amount of unbound GCN4 molecule in the experiment from 13.06.2013 could be that the unspecific binding is also the highest in the experiment from 13.06.2013 and due to dissociation from these unspecifically bound DNA clusters the amount of unbound GCN4 molecule near the clusters is increased.

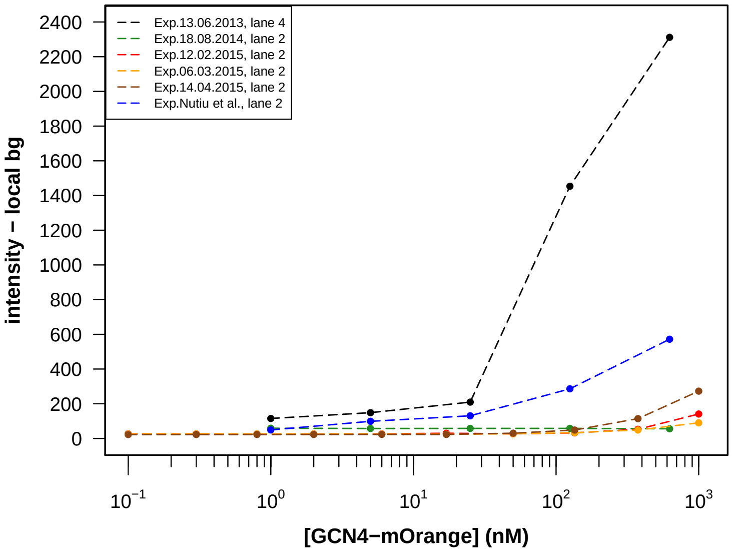


Figure 6.3: Behavior of unspecific GCN4 binding compared for different experiments.

In Figure 6.3 the T-channel median intensity calculated over all DNA clusters of the entire lane is used here as an approximation of the unspecific binding of GCN4. The largest increase can be seen for the experiment by Nutiu et al. (2011) and the experiment from 13.06.2013. The experiment from 18.08.2014 shows no increase.

The cluster density per tile varies slightly among the experiments as shown by Figure 6.4 but to an insignificant degree. Regarding the experiments from 2013, 2014 and 2015, the sequencing protocol was the same and the slight increase could be due to differences in sequencing kits.

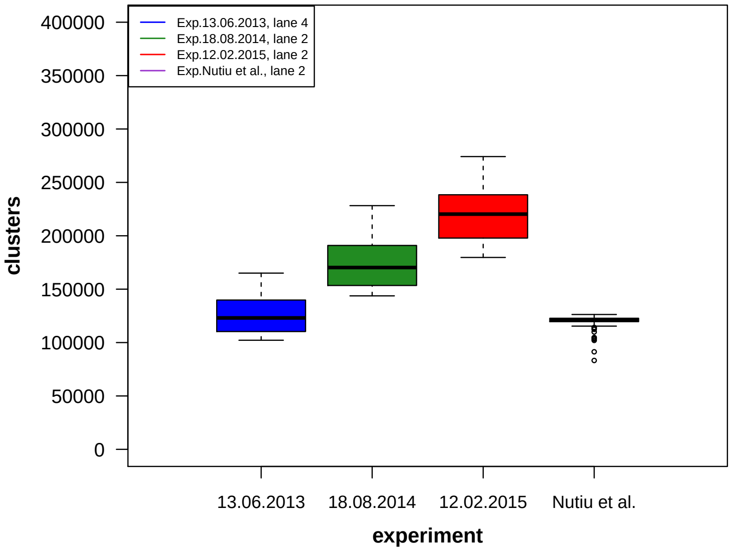


Figure 6.4: Cluster density per tile compared for different experiments.

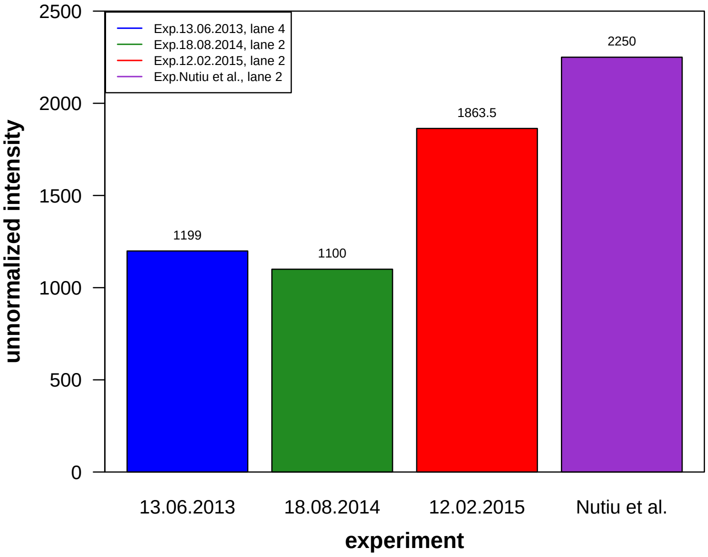


Figure 6.5: Template density per cluster compared for different experiments.

The assumption for Figure 6.5 is that the intensity of a DNA cluster of a sequencing image would be all the brighter the higher the density of the templates in this cluster is. As shown in Figure 6.5 the A-channel median intensities of all the clusters from the second sequencing cycle are compared across the different experiments. This seems to be the greatest difference regarding the experiment from 12.02.2015.

7 Biological results

7.1 Data processing

For the insights and findings described in this chapter, the data from lane 2 of experiment 18.08.2014 is processed in the following manner. Preprocessing of the tif images was done as described in section 5.2, the cluster position transformation was performed as described in section 5.3, the shifting of mapped clusters was carried out as explained in section 5.4, normalization of the cluster intensities was executed as detailed in section 5.5, intensity extraction was performed as stated in section 5.6 , image outlier detection was carried out as described in section 5.7, and DNA cluster sequence filtering was applied as stated in section 5.8. The ranking procedure was applied as explained in subsection 5.9.1.5. The Hill based fitting for determining the Kds was done as described in section 5.10. See the Appendix section 9.10 for a summary of the parameters, input and output.

7.2 Consistency of results from experiment 18.08.2014

Besides the excellent correlation of Kds shown in section 5.10, ranking k-mers with increasing length demonstrates the consistency of the experimental results with known and well established scientific findings regarding the binding behavior of GCN4 as highlighted by the Figure 7.1.

The motif TGAC is the top ranked 4-mer. Biochemical and crystallographic analysis of a complex containing GCN4 bound to the AP-1 site by (Ellenberger et al. (1992a); Sellers et al. (1990a)) has indicated that the optimal half-site is TGAC.

The motif ATGAC is the top ranked 5-mer. (Sellers et al. (1990b)) showed for the first time that the optimal half-site is ATGAC, not ATGAG. In (Stanojevic and Verdine (1995)), it was shown experimentally by a DNase I protection assay that GCN4 binds more strongly to 5'-ATGAC (the consensus half site) than to 5'-ATGAG (the non-consensus half site), δT_m ca. 43°C, T_m half-melting transition.

The known 6-mer consensus motif of GCN4, 5'-TGACTC-3' as discovered by (Arndt and Fink (1986); Gartenberg et al. (1990)) is the top ranked 6-mer.

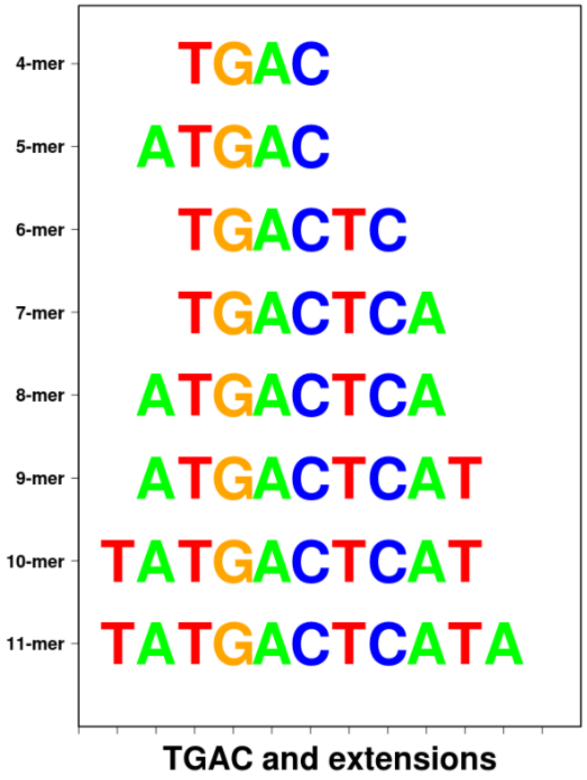


Figure 7.1: Ranking k-mers with increasing length yields TGAC as strongest binding 4-mer, the preferred half-site by GCN4, optimal and subsequently extensions of it.

The pseudosymmetric sequence 5'-TGA(C/G)TCA-3' has been identified from a comparison of enhancer sites in GCN4-dependent promoters (Hill et al. (1986)) and from in vitro selection experiments (MAVROTHALASSITIS et al. (1990); Oliphant et al. (1989)). The heptanucleotide consensus motif TGACTCA is the top ranked 7-mer here.
GCN4 recognizes the pseudo-symmetric 9 bp AP-1 (ATGACTCAT) site in vivo (Hill et al. (1986)) which is the top ranked 9-mer.
All these results are further evidence that the ranking yields biological accurate top binders.

Figure 7.2 shows the enrichment of TGAC, TGAG and TGAT among the first 100 8-mers ranked at 125 nM. Since TGAC is the preferred half-site by GCN4, it is strongly overrepresented.

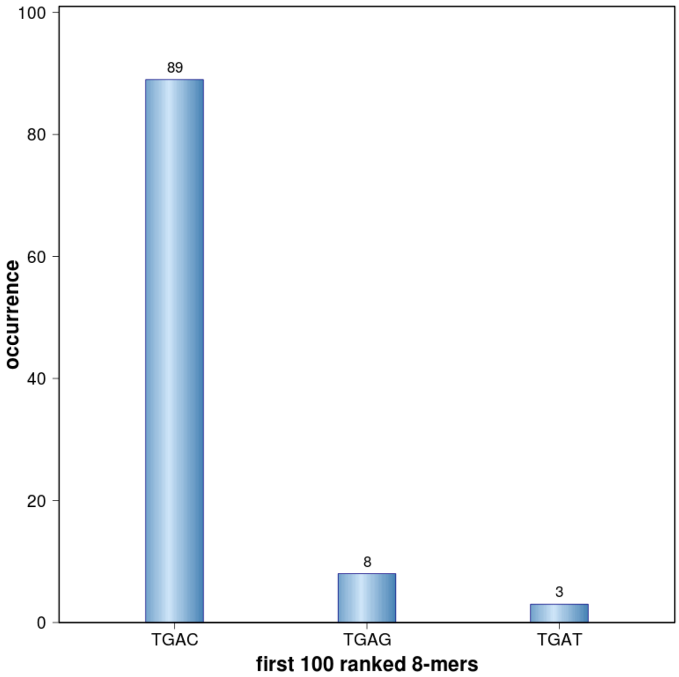


Figure 7.2: Enrichment of GCN4 dimeric half-sites among the first 100 8-mers ranked at 25 nM.

In Figure 7.3 the sequence logo for the top 100 9-mers ranked at 125 nM is compared with the PWM of GCN4 from the ScerTF database (Spivak and Stormo (2012)), a comprehensive database of 1226 motifs for Saccharomyces cerevisiae TFs from 11 different sources. The creators of the ScerTF database identified a single matrix for each TF that best predicts in vivo data by benchmarking matrices against chromatin immunoprecipitation and TF deletion experiments. In addition, in vivo data were also used to optimize thresholds for identifying regulatory sites with each matrix.

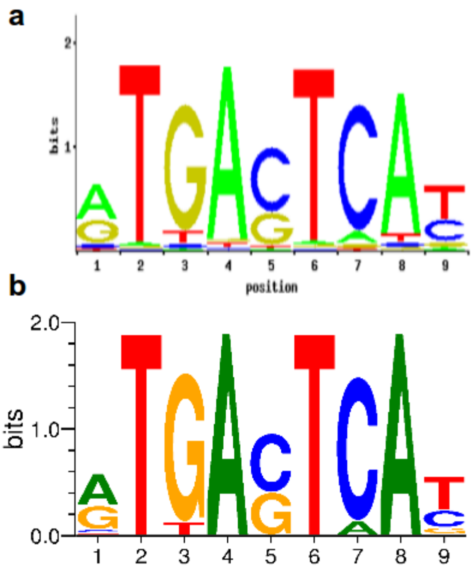


Figure 7.3: Sequence logos for GCN4 9-mers. (a) Sequence logo for GCN4 from the ScerTF database (Spivak and Stormo (2012)). (b) Sequence logo created from the top 100 9-mers ranked at 125 nM.

7.3 Single and double mutation analysis

Figure 7.4 displays all single mutations as $\Delta\Delta G$ values of the first ranked 11-mer TATGACTCATA (reverse complement TATGAGTCATA). The $\Delta\Delta G$ values to all other three mutated nucleotides were averaged for each position. The $\Delta\Delta G$ value, the change in Gibbs free energy (or the binding affinity) is defined by:

$$\Delta G = RT \ln(Kd) \tag{7.1}$$

$$\Delta\Delta G = \Delta G(k) - \Delta G(r) = RT \ln(\frac{Kd_k}{Kd_r}) \tag{7.2}$$

where
R: ideal gas constant: 8.3144598 $J/(molK)$
T: temperature (here 293.15 K for 20°C)
$\Delta G(k)$: Gibbs free energy of a ranked k-mer k.
$\Delta G(r)$: Gibbs free energy of the top ranked k-mer r, used as reference.
Kd_k: Kd of ranked k-mer k.
Kd_r: Kd of top ranked k-mer r, used as reference.

It is clearly visible that the left half-site is more important for binding since mutating nucleotides of this site disrupts the binding affinity more strongly than mutations occurring in the right half-site. In Sellers et al. (1990b); Stanojevic and Verdine (1995) it was demonstrated experimentally that the optimal half-site is 5'-ATGAC and not 5'-ATGAG. The quantitative analysis of Ellenberger et al. (1992b) elucidated that Arg243 in the monomer bound to the cognate half site contributes more to the specific interaction than the other monomer. According to Ellenberger et al. (1992b) the Arg243 side chain makes markedly different contacts at the central base pair: in the specifically bound monomer, Arg243 makes bidentate hydrogen bonds to N7 and O6 of the central guanine, whereas Arg243 in the other monomer donates hydrogen bonds to the DNA phosphodiester backbone.

Examining each single mutation individually shows large differences on the binding affinity of GCN4. Figure 7.5 lists each mutation of the top ranked 11-mer (a), and a comparison is provided to the sequence logo (Schneider and Stephens (1990)) of the PWM of GCN4 from the ScerTF database (Spivak and Stormo (2012)) regarding the nucleotide

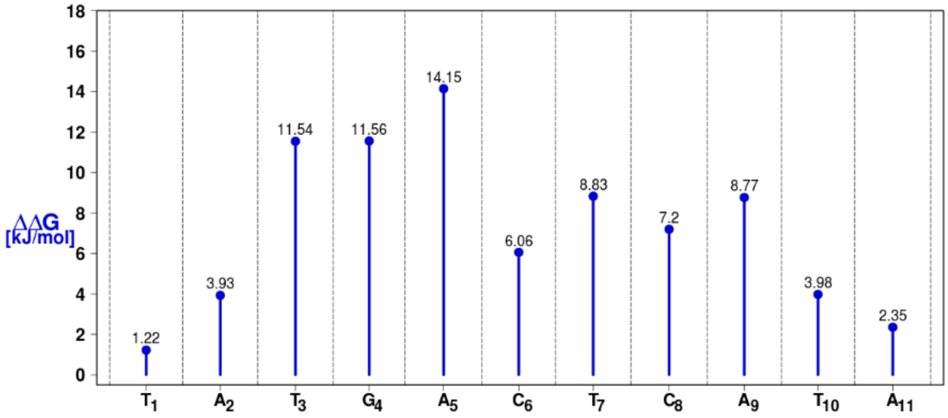


Figure 7.4: All single mutations of first ranked 11-mer TATGACTCATA (TATGAGTCATA) as $\Delta\Delta G$ values averaged over all three mutated nucleotides.

probability for the different positions (b). A sequence logo displays the frequencies of bases or amino acids at each position, as the relative heights of letters, along with the degree of sequence conservation as the total height of a stack of letters, measured in bits of information (Schneider and Stephens (1990)). The bigger the letter appears at a certain position in the sequence logo of Figure 7.3 and the higher the probability of the nucleotide in Figure 7.5 (b), the smaller the related $\Delta\Delta G$ value in Figure 7.5 (a). The probabilities are listed in Appendix 9.11. Using the positions of the 11-mer:
Position 2: mutation A → G: lowest $\Delta\Delta G$ value (1.53).
Position 3: mutation T → A,C,G: only high $\Delta\Delta G$ values (9.83, 11.52, 13.29).
Position 4: mutation G → T: lowest $\Delta\Delta G$ value (7.0).
Position 5: mutation A → C,G,T: only high $\Delta\Delta G$ values (15.43, 12.64, 14.37).
Position 6: mutation C → G: lowest $\Delta\Delta G$ value (0.0).
Position 7: mutation T → A,C,G: only high $\Delta\Delta G$ values (7.22, 11.32, 7.96).
Position 8: mutation C → A: lowest $\Delta\Delta G$ value (3.43).
Position 9: mutation A → T: lowest $\Delta\Delta G$ value (6.44).
Position 10: mutation T → C: lowest $\Delta\Delta G$ value (1.82).
This agreement is further evidence that the HiTS-FLIP Kds are valid and very accurate.

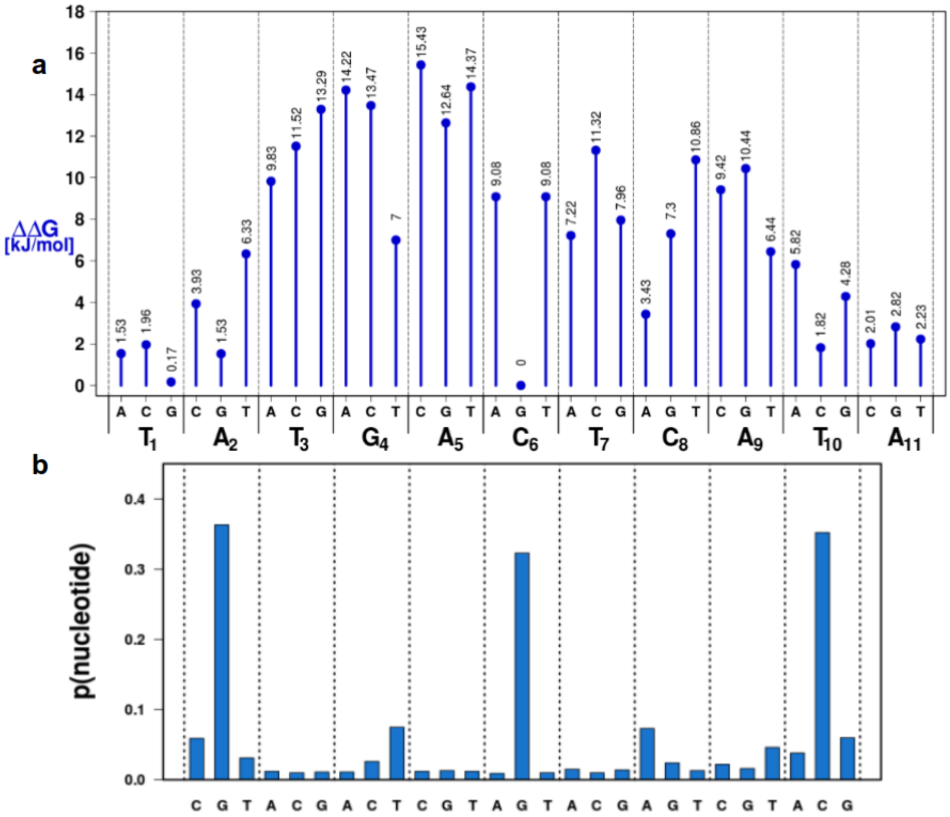


Figure 7.5: All single mutations of first ranked 11-mer TATGACTCATA (TATGAGTCATA) as $\Delta\Delta G$ values and comparison to the PWM of GCN4 from the ScerTF database (Spivak and Stormo (2012)). (a) All single mutations of first ranked 11-mer TATGACTCATA (TATGAGTCATA) as $\Delta\Delta G$ values. (b) Probabilities from the GCN4 PWM from the ScerTF database (Spivak and Stormo (2012)).

In the Figure 7.6 all single and double mutations of the first ranked 11-mer TATGACT-CATA (TATGAGTCATA) with their related $\Delta\Delta G$ values are displayed. The single mutations occur on the antidiagonal, on which the double mutations in the lower and upper triangular matrix are symmetrically mirrored. The heat map shows again the left half-site is more important for binding than the right half-side. Double mutations have the most detrimental effect on the binding if they mutate both half-sites. The finding was pointed out by Nutiu et al. (2011) which suggested a model in which a substitution at one position in a half-site tends to weaken the interaction of the associated GCN4 monomer with other positions in the same half-site, perhaps through a subtle protein conformational change, making interactions between the other monomer and half-site more critical.

To pick one example, an interesting aspect regarding compensation and long range influence is visible for

mutating $\mathrm{A}_2 \rightarrow \mathrm{G}$ and mutating $\mathrm{C}_8 \rightarrow \mathrm{A}$ ($\Delta\Delta G = 5.42$ kJ/mol).

The other mutations at position 2

$\mathrm{A}_2 \rightarrow \mathrm{C}$ ($\Delta\Delta G = 7.51$ kJ/mol), and

$\mathrm{A}_2 \rightarrow \mathrm{T}$ ($\Delta\Delta G = 10.44$ kJ/mol)

have a more weakening effect on the binding energy.

Figure 7.7 shows a heat map where the values are calculated as $\Delta\Delta G(k_{i,j}) - (\Delta\Delta G(k_i) + \Delta\Delta G(k_j))$, where i and j are the positions in the first ranked 11-mer TATGACTCATA which are mutated.

Having a higher $\Delta\Delta G$ value here means that the double mutation has a stronger detrimental effect on binding than the two single mutations would provoke. This effect only occurs for the positions 6 and 7, or 6 and 8.

$\mathrm{C}_6 \rightarrow \mathrm{G}$ and $\mathrm{T}_7 \rightarrow \mathrm{A}$ or $\mathrm{T}_7 \rightarrow \mathrm{G}$.

$\mathrm{C}_6 \rightarrow \mathrm{G}$ and $\mathrm{C}_8 \rightarrow \mathrm{G}$.

Having a lower $\Delta\Delta G$ value here means that the double mutation has a less detrimental effect on binding than the two single mutations would provoke, and thus hints at a synergistic effect linking two nucleotide positions.

Neighboring effects between dinucleotides are clearly apparent, more strongly in the left half-site than the right half-site. The effect concerns most strongly the positions 3, 4 and 5 in the left half-site. There are long-range effects visible between the two GCN4 half-sites related to positions 3, 4 and 5 in the left half-site and positions 7, 8 and 9 in the right half-site.

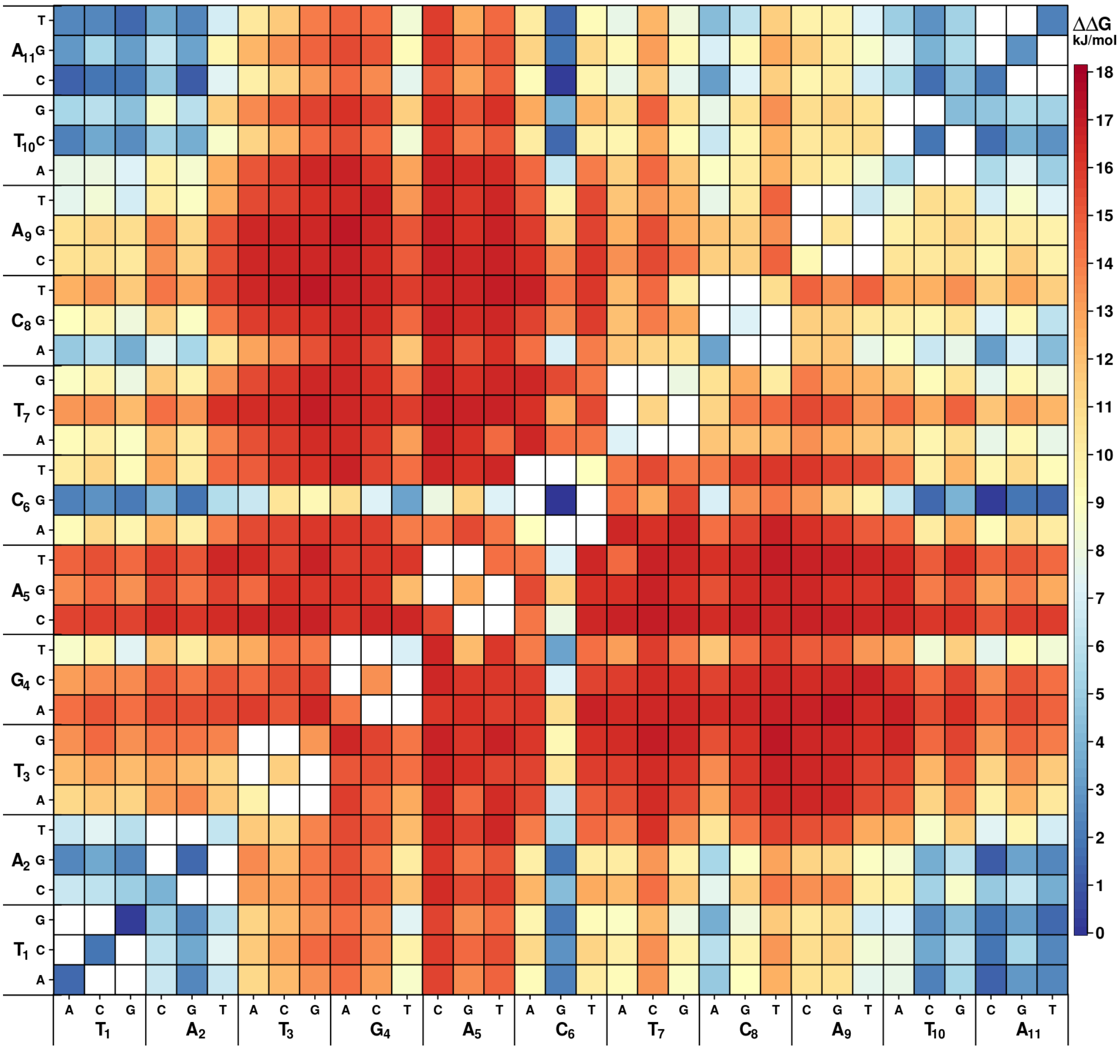


Figure 7.6: All single and double mutations of first ranked 11-mer TATGACTCATA (TATGAGTCATA).

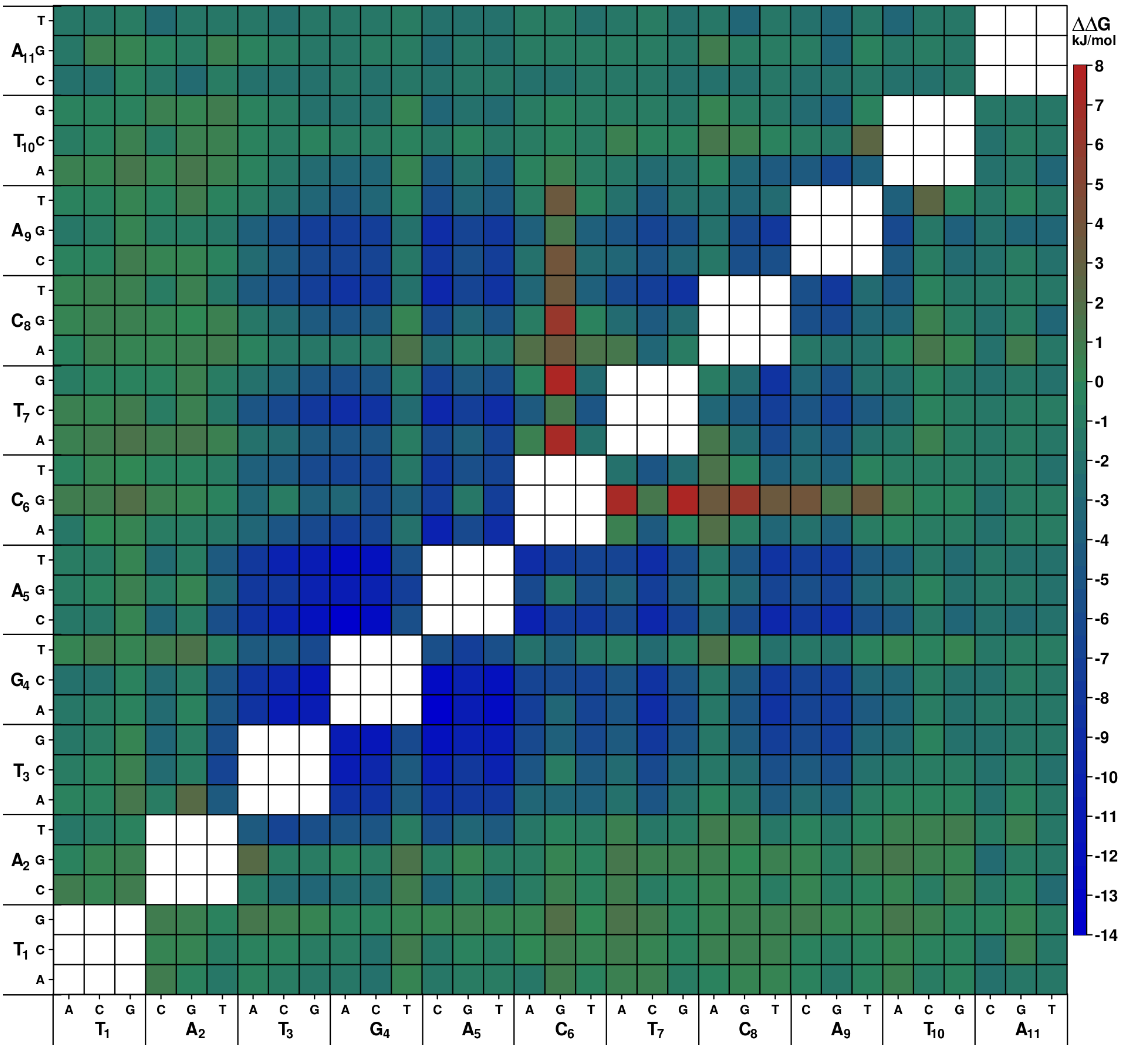


Figure 7.7: Effect of double mutations in comparison to individual single mutations. The values in the heat map are calculated as $\Delta\Delta G(k_{i,j}) - (\Delta\Delta G(k_i) + \Delta\Delta G(k_j))$, where i and j are the positions in the first ranked 11-mer TATGACTCATA which are mutated.

7.4 Influence of flanking nucleotides on binding affinity

If the inner positions 3, 4 and 8, 9 of the top 11-mer TATGACTCATA is mutated, the binding affinity is drastically changed (Figure 7.8 (a)). If the mutations occur in the flanking region (positions 1, 2 and 10, 11), the binding affinity is gradually modulated from 1.56 Kd to 552.51 Kd (Figure 7.8 (b)), more than a 100-fold in total. Therefore, the known 9-mer PWM (Spivak and Stormo (2012)) is insufficient for describing the binding affinity of GCN4, and as the HiTS-FLIP data demonstrate the 11-mer TATGACTCATA together with the mutations in the flanking regions provide the spectrum of binding affinities of which GCN4 is capable of. This finding is related to (Levo et al. (2015a)) where it was demonstrated that flanking sequences of core binding sites affect the binding of GCN4 (flanking 3-mers besides the 9-bp core motif) using a novel experimental assay termed BunDLE-seq.

Selecting a few mutated 11-mers highlights the influence of the nucleotides from the flanking region as well as the sensitivity with which HiTS-FLIP is able to measure Kds. In Figure 7.9 (a) the nucleotide at position 1 (upper four 11-mers) is changed (T to G, A, and C), and reveals subtle differences in Kds. The nucleotide at position 2 (lower four 11-mers) is changed (A to G, C, and T), provoking larger differences in Kds. Changing T to C at position 1 of TTTGACTCATA (bottom 11-mer) has similar strength in Kd change than the mutation occurring at position 2, emphasizing the importance of the outer flanking position. In Figure 7.9 (b) same scenario as (a) but with mutations taking place in the right flank. Changing the nucleotide at position 11 (upper four 11-mers) from A to C,G and T invokes a 2-fold change in Kds (Kd=1.56 nM to Kd=4.97 nM). Mutating the nucleotide at position 10 (lower four 11-mers) from T to C, G and A provokes larger differences in Kds. However, changing A to G at position 11 resulting in TATGACTCAAG (bottom 11-mer) increases the Kd=17.04 nM to Kd=32.38 nM by almost 2-fold, emphasizing again the importance of the outer flanking position for the binding affinity.

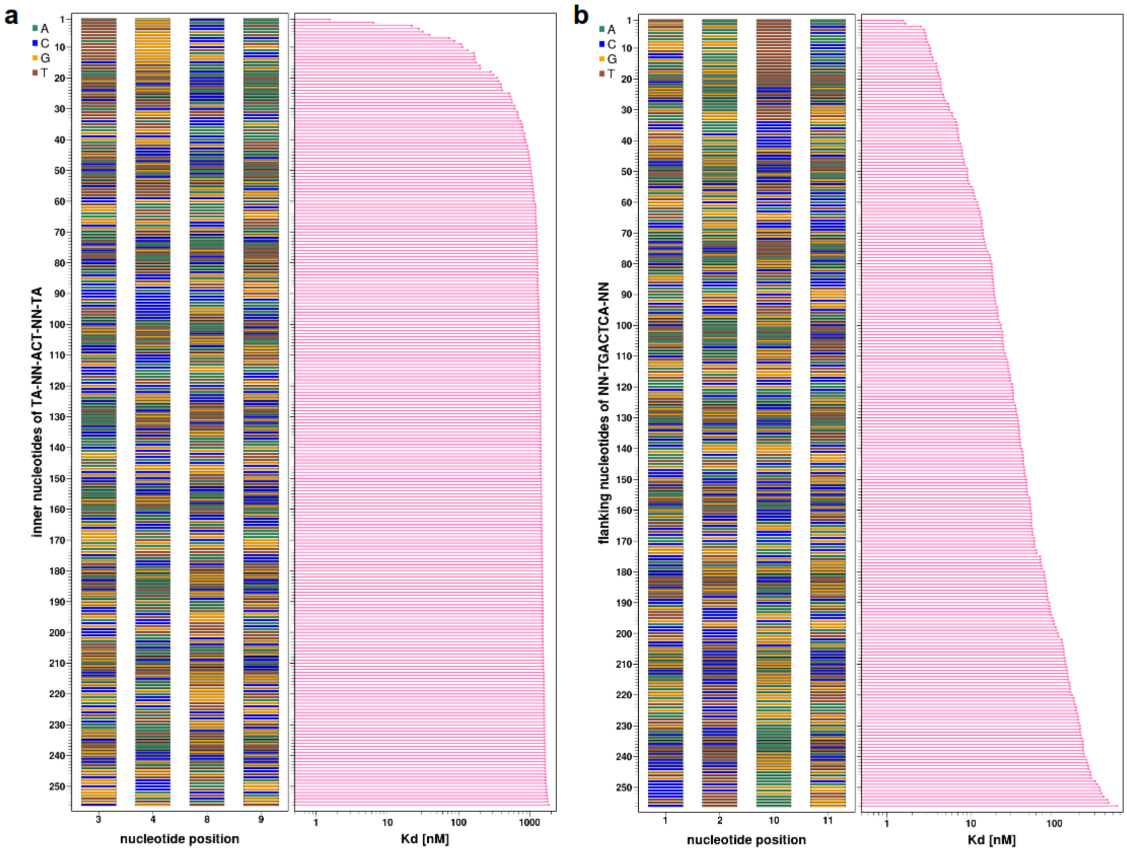


Figure 7.8: Effect of quadruple mutations. (a) Mutations occurring at positions 3, 4 and 8, 9 (TA-NN-ACT-NN-TA), and (b) in flanking positions 1, 2 and 10, 11 (NN-TGACTCA-NN) of top 11-mer.

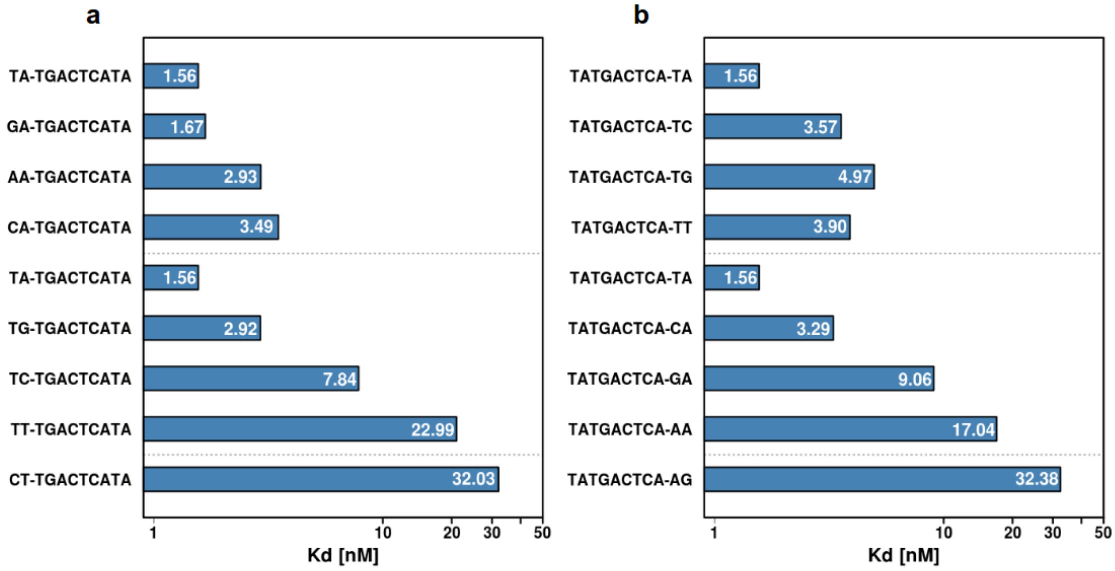


Figure 7.9: Details regarding the effect of flanking nucleotides on binding affinity shown by selected 11-mers. (a) Mutations occurring in the left flank, (b) Mutations occurring in the right flank.

7.5 Discovery of new GCN4 binding motifs

In the ranking conducted on the data from experiment 18.08.2014, very unique antagonistic peaks could be observed as shown by Figure 7.10.

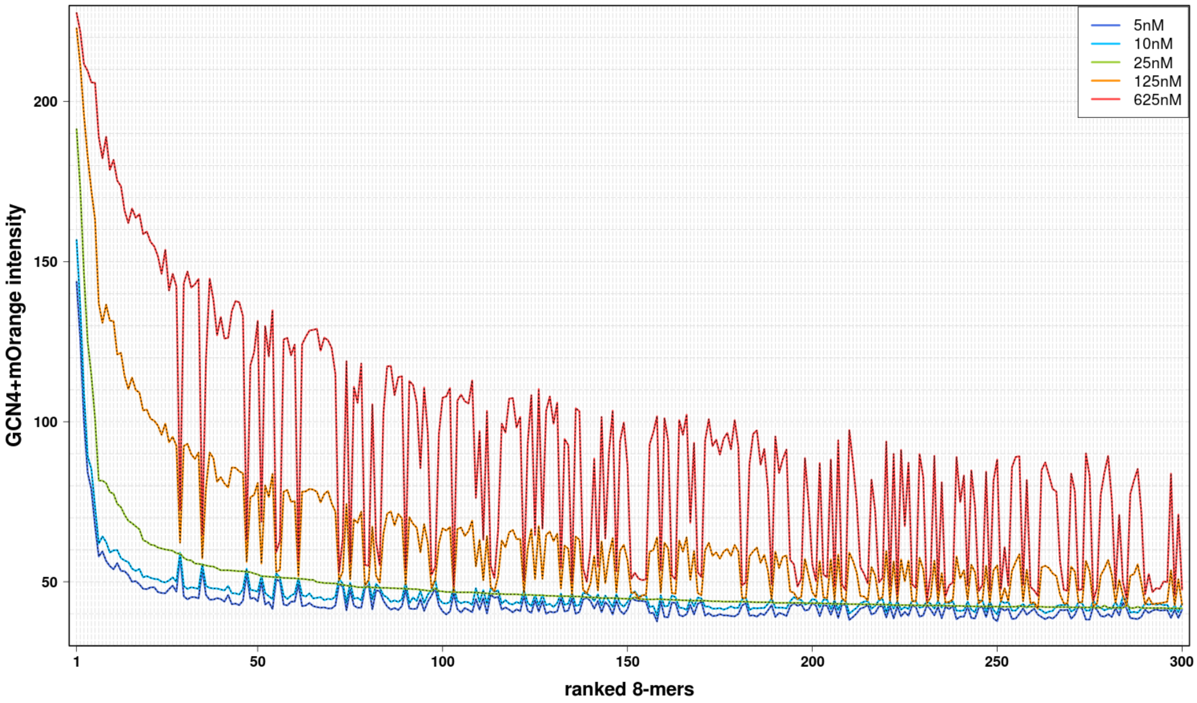


Figure 7.10: 8-mers ranked at 25 nM of experiment 18.08.2014 with antagonistic peaks.

In Figure 7.11, the peaks are annotated and the corresponding 8-mer motifs are listed. The submotif GTGT occurs in all of these peaks. The hypothesis is that at low concentrations GCN4 occurs mainly as monomer and binds in this oligomerization state specifically to the motif GTGT. At higher concentrations GCN4 has largely dimerized and occurs predominantly as dimer binding specifically to motifs containing the half-site TGAC (or a close derivative).

Correlating the occurrence of intensity down peaks at 625 nM with the occurrence of the submotif GTGT yields a perfect agreement (Spearman rank correlation coefficient $\rho = 1$), Figure 7.12. Wherever there is the submotif GTGT occurring, there is an intensity down peak at 625 nM. Down peaks have been determined by fitting a cubic smoothing spline to the ranked 8-mer intensities and classifying the intensity as a down peak if it lies below the fitted spline line. There are five exceptions:

rank 36: TGACGTCA (TGACGTCA)

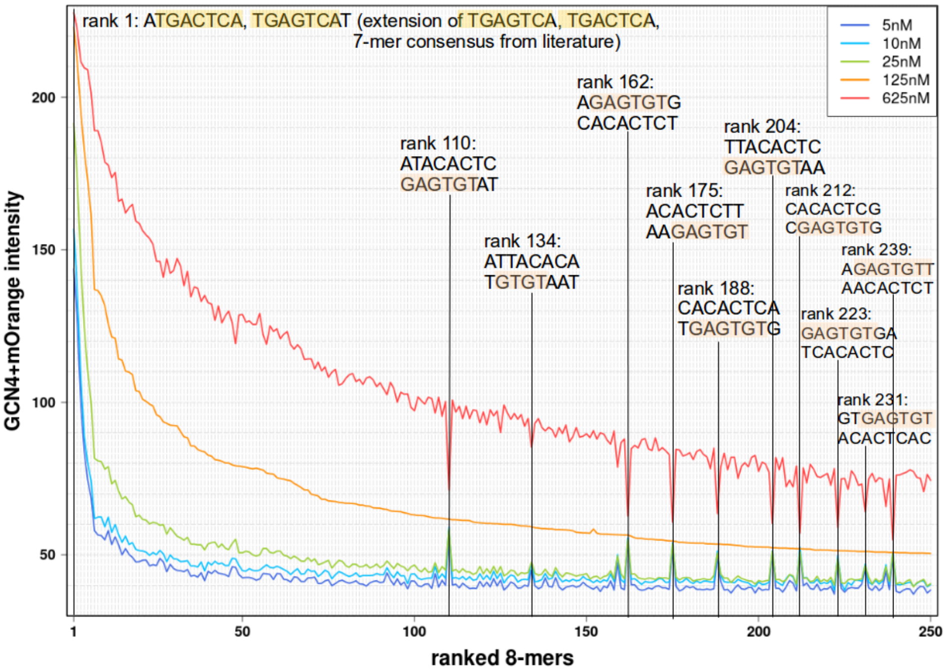


Figure 7.11: 8-mers ranked at 125 nM of experiment 18.08.2014 with annotations of the peaks.

The 8-mer contains the optimal half-site TGAC and can be classified rather not as a down peak (Figure 7.12, first green dot).

rank 48: TGACGTGT (ACACGTCA)

Here the 8-mer contains GTGT but also the optimal half-site TGAC, and cannot really be regarded as a down peak (Figure 7.12, second green dot).

rank 103: AAAAAAAA (TTTTTTTT)

Here, clearly a down peak is realized (Figure 7.12, third green dot) but GTGT does not occur as submotif. The most likely explanation is that GCN4 binds to this stretch of adenine (or thymine) nucleotides as a monomer as well.

rank 276: AAAAAAAT (ATTTTTTT)

Here, clearly a down peak is realized (Figure 7.12, fourth green dot) but GTGT does not occur as submotif. A possible explanation is that GCN4 binds to this stretch of adenine (or thymine) nucleotides as a monomer as well.

rank 282: AGATTGTA (TACAATCT)

Here, the submotif GATTGT is a degenerated version of GAGTGT and seems to be bound by GCN4 likewise.

Similarly, correlating the occurrence of intensity up peaks at 625 nM with the occurrence

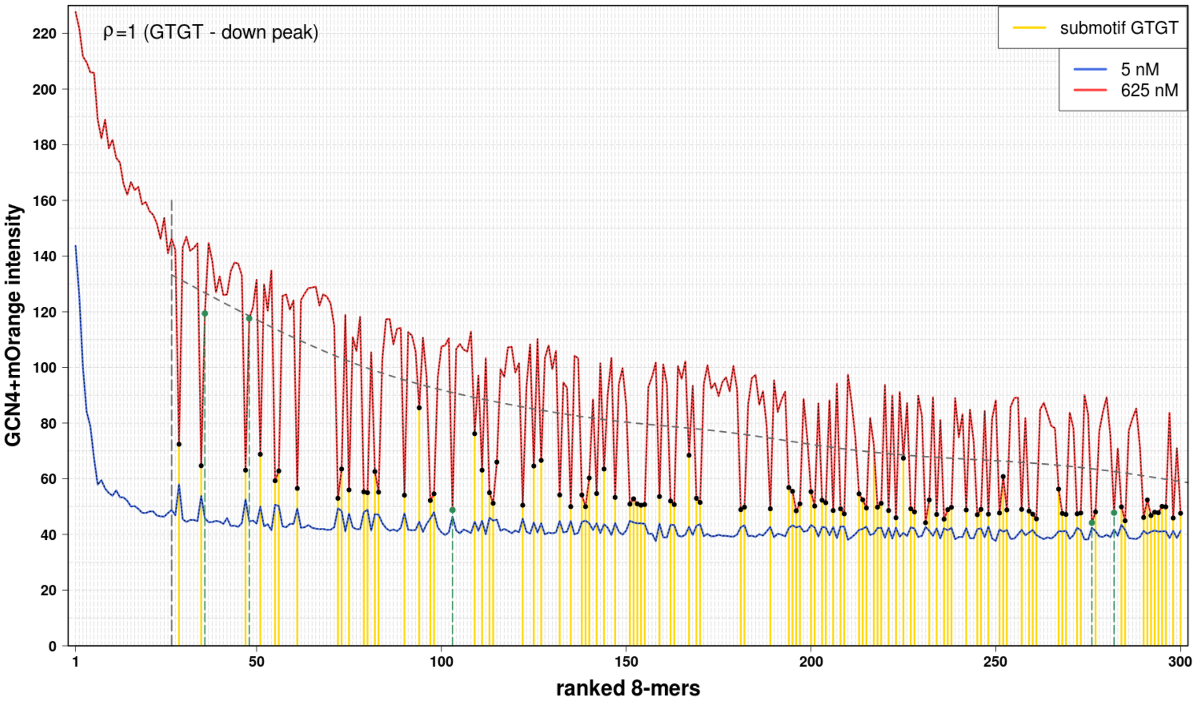


Figure 7.12: Rank correlation between intensity down peaks occurring at 625 nM and the occurrence of the submotif GTGT. The yellow lines denote intensity down peaks at 625 nM, green lines are the exceptions as discussed above.

of the submotif TGAC yields a perfect agreement (Spearman rank correlation coefficient $\rho = 1$), Figure 7.13. Wherever there is the submotif TGAC occurring, there is an intensity up peak at 625 nM.

Ranking k-mers of different length (4, 5 nt and so on) at concentration 5 nM and selecting the top ranked k-mer for which an intensity down peak at 625 nM occurred leads to the following result (Figure 7.14). For k-mers of length greater 8 nt, no intensity down peaks at 625 nM or intensity up peaks at lower concentration could be observed anymore.

Selecting the first 100 occurrences of 8-mers containing the submotif GTGT and aligning them on this submotif results in the sequence logo shown in Figure 7.15. Preferably, the left flank is made up of GA or GT, and the right flank A or G and T or A.

The lower the concentration of GCN4 is, the more the ratio of monomers and dimers should be shifted towards monomers. The more monomers occur in the flow cell, the higher the amount of bound DNA clusters containing the motif GTGT in their sequence. Therefore, a motif enrichment of GTGT should be visible in the ranking the lower the concentration is at which the ranking takes place, which is displayed in Figure 7.16.

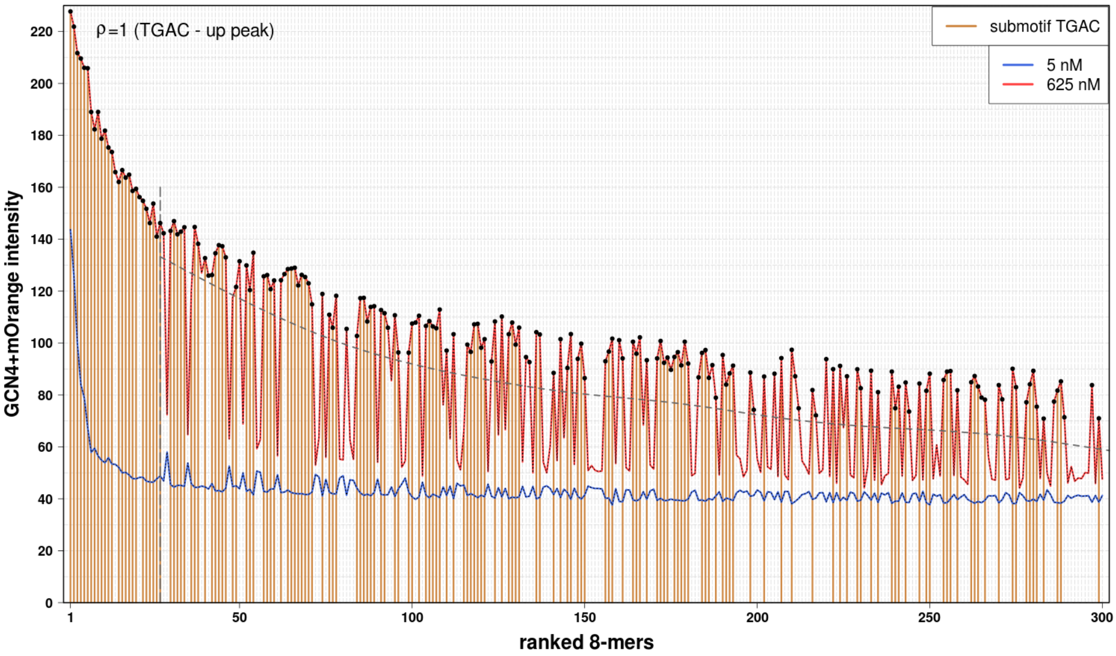


Figure 7.13: Rank correlation between intensity up peaks occurring at 625 nM and the occurrence of the submotif TGAC. The brown lines denote intensity up peaks at 625 nM.

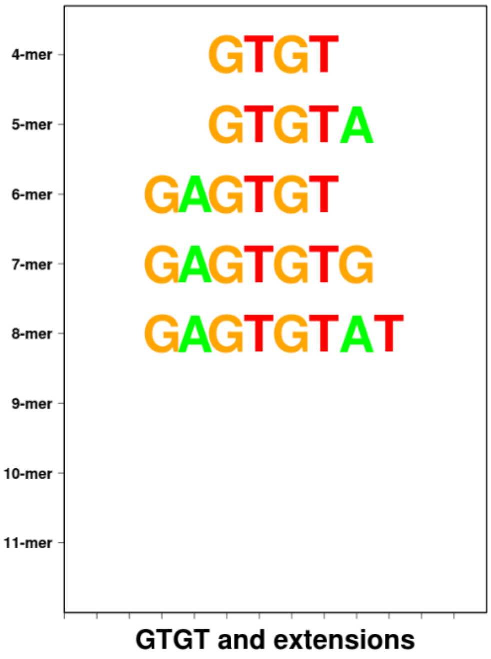


Figure 7.14: Extension of submotif GTGT.

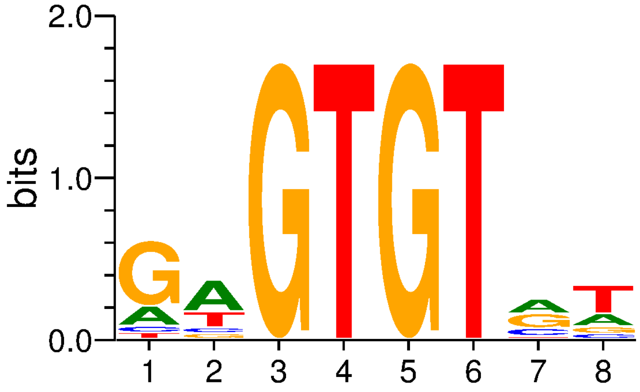


Figure 7.15: Sequence logo for the submotif GTGT.

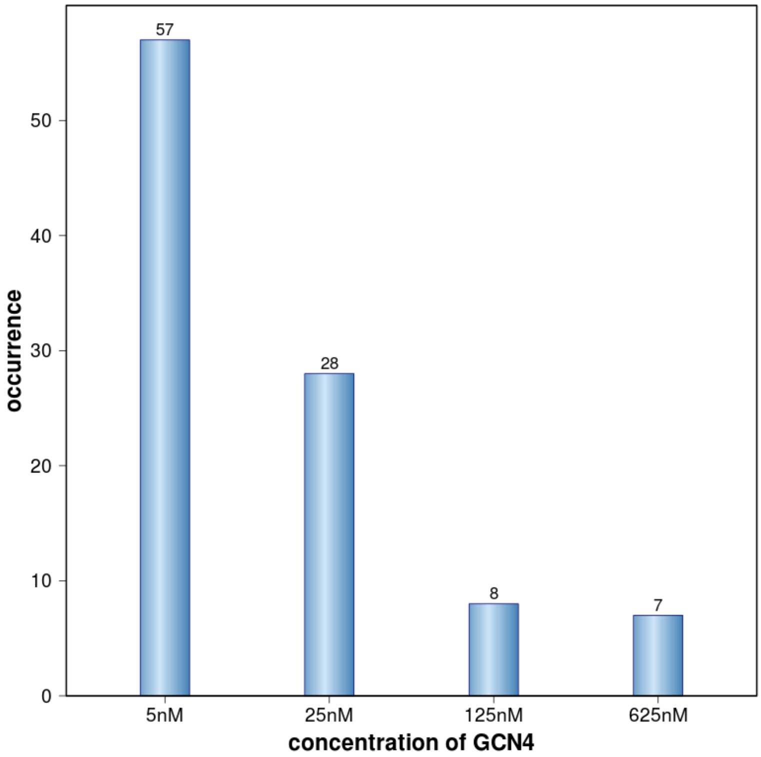


Figure 7.16: Enrichment of GTGT among the first 100 8-mers ranked at different concentrations.

7.6 Literature based evidence for GTGT affinity and monomer binding

When unfolded proteins accumulate in the endoplasmic reticulum (ER), a signal is sent across the ER membrane into the nuclear and cytoplasmic compartments. There, effector proteins respond by upregulating the transcription of a characteristic set of target genes and slowing general translation, and the cell is enabled to tolerate and survive conditions which compromise protein folding in the ER. This reaction to ER stress is known as the unfolded protein response (UPR), a signal transduction pathway that communicates between the ER and the nucleus (Patil and Walter (2001)). In (Patil et al. (2004)), the authors analyzed the promoters of UPR target genes computationally, identifying as candidate upstream activating sequences (UASs) short sequences that are statistically overrepresented. They tested the most promising of these UASs for biological activity, and identified two novel unfolded protein response elements (UPREs), which are necessary and sufficient for UPR activation of promoters. (Patil et al. (2004)) demonstrated that Gcn4p is required for normal induction of UPR transcription, both in the context of artificial promoters containing any of the known UPREs and in the context of the native promoters of most target genes. Both Hac1p and Gcn4p bind target gene promoters to stimulate transcriptional induction, and UPRE-2 can be activated by Gcn4p alone, and it is bound by Gcn4p either as a homodimer or a monomer (Patil et al. (2004)). Both UPRE-1 and UPRE-2 contain GTGT as submotif as shown in Figure 7.17.

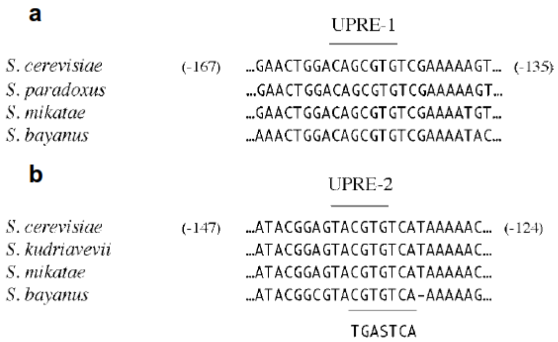

a
UPRE-1
S. cerevisiae (-167) ...GAACTGGACAGCGTGTCGAAAAAGT... (-135)
S. paradoxus ...GAACTGGACAGCGTGTCGAAAAAGT...
S. mikatae ...GAACTGGACAGCGTGTCGAAAATGT...
S. bayanus ...AAACTGGACAGCGTGTCGAAAATAC...
b
UPRE-2
S. cerevisiae (-147) ...ATACGGAGTACGTGTCATAAAAAC... (-124)
S. kudriavevii ...ATACGGAGTACGTGTCATAAAAAC...
S. mikatae ...ATACGGAGTACGTGTCATAAAAAC...
S. bayanus ...ATACGGCGTACGTGTCA-AAAAAG...
TGASTCA

Figure 7.17: Multiple alignment of UPRE-1 and UPRE-2 from three budding yeasts. (a) A segment of the KAR2/YJL034W promoter and homologs. The core sequence of UPRE-1 is indicated. (b) A segment of the ERO1/YML130C promoter and homologs. The core sequence of UPRE-2 is indicated. Figure adapted from Patil et al. (2004).

In (Fordyce et al. (2012b)) it was shown by using MITOMI (mechanically induced trapping of molecular interactions) that Hac1, a bZIP TF like GCN4, possesses two distinct binding modes: (1) to short (6-7 bp) UPRE-2-like motifs (containing GTGT) and (2) to significantly longer (11-13 bp) extended UPRE-1-like motifs.

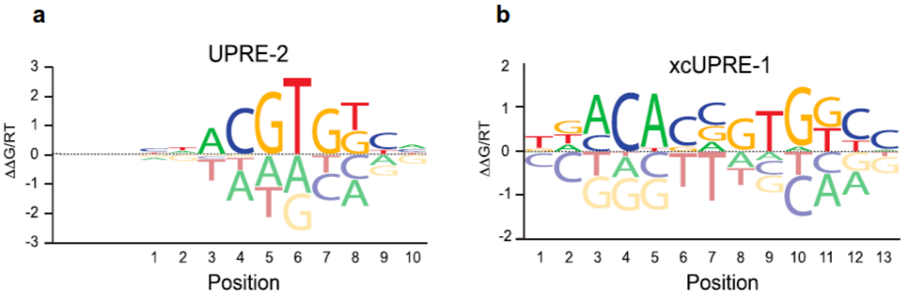


Figure 7.18: Nucleotide binding preferences of Hac1 as affinity logos (Foat et al. (2006)) derived from relative affinities. (a) Affinity logo for UPRE-2. (b) Affinity logo for xcUPRE-1. Figure adapted from Fordyce et al. (2012b).

Cranz et al. (2004) demonstrated experimentally that GCN4 can bind to DNA as unfolded monomeric and folded dimeric derivatives of GCN4. The association rate of the monomer is virtually the same as that of the dimer, 5×10^8 M^{-1} s^{-1} (Cranz et al. (2004)). Because the rate of dimerization of GCN4 is slower (1.7×10^7 M^{-1} s^{-1}) than the rate of DNA association, the formation of the dimeric GCN4-DNA complex through consecutive binding of two monomers (monomer pathway) is faster when starting from free monomers. Thus, if GCN4 occurs largely as monomers, the monomer DNA binding pathway is preferred. The following Figure, adapted from (Cranz et al. (2004)), shows that a GCN4 monomer mutant can bind to DNA in a stable manner. CRE^{F}_{19} is a double-stranded 19-mer oligonucleotide containing the CRE site with the fluorescence marker NBD attached to a phosphorothioate bond preceding the recognition site. $\mathrm{C62GCN4}^{one-leg}_{SS}$ is a monomer derivative of the wild type GCN4 which makes contact with only one half-site.

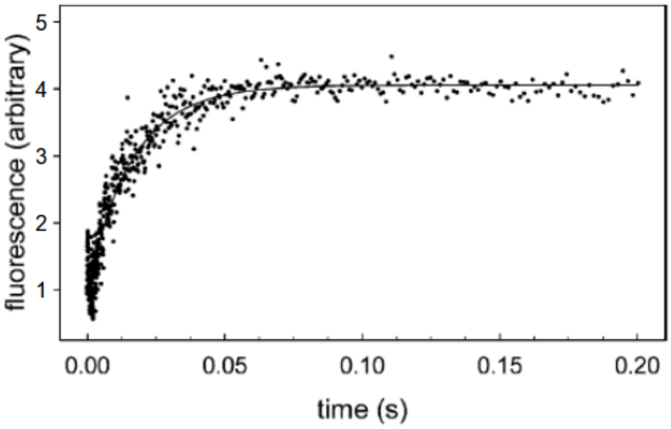


Figure 7.19: GCN4 monomer binding to DNA. Reaction of 35 nM one-legged derivative C62GCN4$_{SS}^{one-leg}$ with 35 nM CRE$_{19}^{F}$. Figure adapted from (Cranz et al. (2004)).

8 Conclusion and Outlook

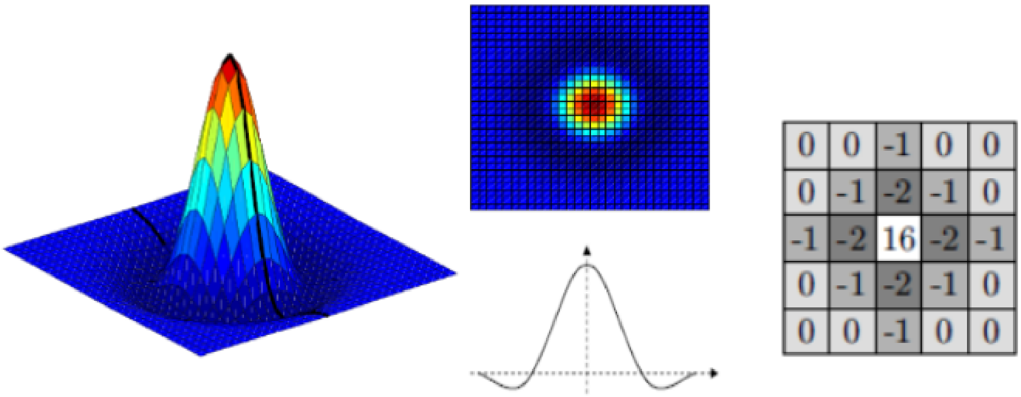
Repurposing an Illumina GA-IIx NGS sequencing machine, it is possible to measure in parallel binding events to hundreds of millions of DNA clusters at equilibrium. This enables the measurement of accurate dissociation constants for the entire sequence space of all possible mutations up to a k-mer length of 12 nucleotides as shown by Nutiu et al. (2011). My approach of applying phase-correlation to estimate the relative translative offset between the observed tile images and the template images omits resequencing and thus allows to reuse the flow cell for several HiTS-FLIP experiments, which greatly reduces cost and time. Instead of using information from the sequencing images like Nutiu et al. (2011) for the normalization of cluster intensities which introduces a nucleotide specific bias, I estimate the cluster related normalization factors directly from the protein images which captures the non-even illumination bias more accurately and leads to an improved correction for each tile image. My analysis of the ranking algorithm by Nutiu et al. (2011) has revealed that it is unable to rank all measured k-mers. Discarding all the clusters related to previously ranked k-mers has the side effect of eliminating any clusters on which k-mers could be ranked that share submotifs with previously ranked k-mers. This shortcoming affects even strong binding k-mers with only one mutation away from the top ranked k-mers. My analysis shows that omitting the cluster deletion step in the ranking process overcomes this limitation and can rank the full spectrum of all possible k-mers. In addition, the performance of the ranking algorithm is drastically reduced from a quadratic to a linear run time. The TIRF optics of the GA-IIx allows to avoid any washing step, done by Nutiu et al. (2011), and to measure the binding events at equilibrium. The experimental improvements combined with the sophisticated processing of the data led to a very high accuracy of the HiTS-FLIP Kds comparable to the Kds measured by the very sensitive HiP-FA assay (Jung et al. (2015)). However, as evident from all the performed experiments, HiTS-FLIP is so far not a robust assay for achieving saturated binding curves, and how to setup optimal experimental conditions and to handle best protein aggregation occurring at the amplified DNA clusters needs further investigation. Nevertheless, we achieved a successful experiment (18.08.2014) resulting in a unique, quantitative data set and utilizing the related Kds for investigating the binding

behavior of GCN4 has shed more light on the complexity of its DNA association.
Given the obtained insights from the down stream analyses I could demonstrate that the common 9-mer PWM for GCN4 is insufficient to describe the binding behavior of GCN4. Rather, an additional left and right flanking nucleotide is required to extend the 9-mer to an 11-mer whereby the influence of the flanking nucleotides is taken into account which modulates the binding affinity a 100-fold. My analyses regarding mutations and related $\Delta\Delta G$ values suggest long-range interdependencies between nucleotides of the two dimeric half-sites of GCN4 and thus models assuming positional independence, like the PWM, are not able to embody such effects. Instead, the full spectrum of affinity values for all k-mers of appropriate size should be measured and applied as originally proposed by Nutiu et al. (2011). Another important discovery were completely new binding motifs of GCN4, which can only be detected with a method like HiTS-FLIP that examines the entire sequence space and allows for de-novo motif discovery in an unbiased way. All these new motifs contain the submotif GTGT and the evidence collected suggests that GCN4 binds as monomer to these new motifs. Therefore, it might be even possible to detect different binding modes with HiTS-FLIP.
Future steps are further experimental improvements to turn HiTS-FLIP into a robust assay. This might even require to adapt the hardware of the GA-IIx to be better suited for a DNA-protein binding experiment. Another possibility might be to use relative affinities based on intensities alone which could describe the binding behavior sufficiently. As a research topic, a very promising scenario would be to investigate simultaneously multiple proteins in the flow cell to study their cooperativity or perhaps even their antagonistic binding effects. If the DNA clusters in the flow cell could be methylated, the effect of methylation marks on DNA binding of certain TFs could be studied in depth, by measuring the binding behavior first on unmethylated and then on methylated DNA. As it was demonstrated by (Buenrostro et al. (2014); Tome et al. (2014)) the HiTS-FLIP assay can be used for measuring the affinity of RNA binding proteins. A possible research direction along this line would be the study of transcriptomics where the entire transcriptome of interest could be examined in the flow cell. Another important application for future research is the design of custom agents like TALENs for genome editing in the field of personalized medicine where genome-wide off-target effects need to be studied for which HiTS-FLIP is a very suitable tool. Finally, on the bioinformatics level further investigation on a mathematical formalism is required to fully capture the complexity of the binding behavior of a TF like GCN4.

9 Appendix

9.1 Details regarding the LoG filter

LoG filter computes the weighted difference between the center pixel and the surrounding pixels and thus reacts most strongly to local intensity peaks. Other names of the LoG filter are Marr-Hildreth-Operator or Mexican hat filter, since it has the shape of a positive peak in a negative dish and is thus an "inverted Mexican hat" (Wu et al. (2010)). The parameter σ controls the width of the peak, which is related to the amount of smoothing. The edge positions can be determined by the zero-crossings in the LoG-filtered image, Figure 9.1.



0	0	-1	0	0
0	-1	-2	-1	0
-1	-2	16	-2	-1
0	-1	-2	-1	0
0	0	-1	0	0

Figure 9.1: LoG filter as continuous function and pixel kernel. Here the continuous function as well as a 5×5 pixel kernel are shown for the LoG filter, adapted from (Burger et al. (2009)).

The definition of the Laplacian operator (Vinogradov and Hazewinkel (2001)):

$$\triangle f = \nabla^2 = \nabla \cdot \nabla f = \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2} \tag{9.1}$$

where f is a twice-differentiable real-valued function. As denoted by the definition, the Laplacian of f is the sum of all the second partial derivatives in the Cartesian coordinates x_i. The definition of the Gaussian filter in two dimensions is the following (Vinogradov and Hazewinkel (2001)):

$$G_\sigma(x,y) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \tag{9.2}$$

The Gaussian filter is separable in the x and y directions and can thus be written as the product of two 1d Gauss functions:

$$G_\sigma(x,y) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right)\right)\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{y^2}{2\sigma^2}\right)\right) \tag{9.3}$$

Definition of the LoG filter, where $f(x,y)$ represents the image as a function $\mathbb{R}^2 \rightarrow \mathbb{R}$:

$$\triangle[G_\sigma(x,y) \star f(x,y)] = [\triangle G_\sigma(x,y)] \star f(x,y) = LoG \star f(x,y) \tag{9.4}$$

Since the convolution of two functions f and g, denoted by $\star$, is defined as the integral of the product of the two functions after one is reversed and shifted:

$$(f \star g)(t) \overset{def}{=} \int\limits_{-\infty}^{\infty} f(\tau)g(t-\tau)\,d\tau \tag{9.5}$$

The convolution operation is commutative: $f \star g = g \star f$
because, upon the substitution $\sigma = t - \tau$:

$$(f \star g)(t) = \int\limits_{-\infty}^{\infty} g(\sigma)f(t-\sigma)\,d\sigma = (g \star f)(t) \tag{9.6}$$

Therefore

$$\int\limits_{-\infty}^{\infty} f(\tau)g(t-\tau)\,d\tau = \int\limits_{-\infty}^{\infty} f(t-\tau)g(\tau)\,d\tau \tag{9.7}$$

which proofs

$$\triangle[G_\sigma(x,y) \star f(x,y)] = [\triangle G_\sigma(x,y)] \star f(x,y) \tag{9.8}$$

It is therefore equal to firstly convolute the Gaussian filter with the image and then apply the Laplacian operator on this modified image, or to apply the Laplacian operator on the Gaussian filter and then use this modified filter to convolve it with the image. The key aspect here is that the latter is computational more efficient since the modified filter can be prepared in advance as a result of its image independence.
The derivation of the LoG filter is:

$$\frac{\partial^2}{\partial^2 x} G_\sigma(x,y) = \frac{1}{2\pi\sigma^2} \frac{x^2-\sigma^2}{\sigma^4} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \tag{9.9}$$

$$\frac{\partial^2}{\partial^2 y} G_\sigma(x,y) = \frac{1}{2\pi\sigma^2} \frac{y^2-\sigma^2}{\sigma^4} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \tag{9.10}$$

$$LoG = \triangle G_\sigma(x,y) = \frac{\partial^2}{\partial^2 x} G_\sigma(x,y) + \frac{\partial^2}{\partial^2 y} G_\sigma(x,y) \tag{9.11}$$

$$= \frac{1}{2\pi\sigma^2} \frac{x^2-\sigma^2}{\sigma^4} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) + \frac{1}{2\pi\sigma^2} \frac{y^2-\sigma^2}{\sigma^4} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \tag{9.12}$$

$$= \frac{1}{\pi\sigma^4} \left(\frac{x^2+y^2}{2\sigma^2} - 1\right) \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \tag{9.13}$$

In order to be applicable to images and their discrete pixel values, the LoG filter has to be discretized by sampling the function in the above equation into a $(2k+1) \times (2k+1)$ filter kernel (also called mask) for an appropriate value of k. It is proposed (Klette (2014)) to use a window size of $\lceil 6\sqrt{2}\sigma \rceil \times \lceil 6\sqrt{2}\sigma \rceil$, i.e. the smallest integer equal to or larger than the $6\sqrt{2}\sigma$. The value of σ, the amount of smoothing, needs to be estimated for the given image data.
Similarly, the convolution regarding discrete values changes in the following manner. The discrete convolution can be defined as a “shift and multiply” operation, where the kernel is shifted over the image and its value is multiplied with the corresponding pixel values of the image. For a square kernel with size $M \times M$, the output image is calculated with

the formula (Burger and Burge (2009a)):

$$g(i,j) = \sum_{m=-\frac{M}{2}}^{\frac{M}{2}} \sum_{n=-\frac{M}{2}}^{\frac{M}{2}} LoG(m,n) f(i-m, j-n) \tag{9.14}$$

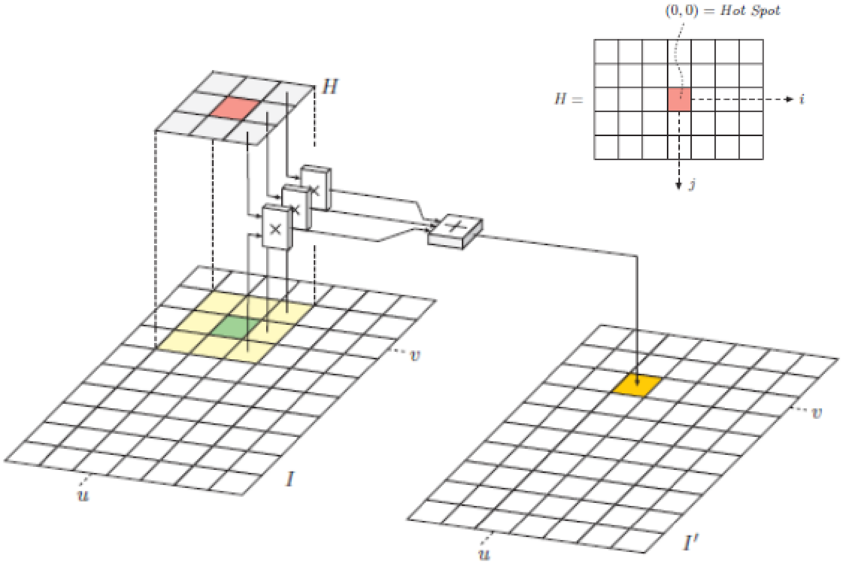


Figure 9.2: Filtering process by mask. The filter matrix H is placed with its origin at position (u, v) on the image I. Each filter coefficient $H(i, j)$ is multiplied with the corresponding image pixel $I(u + i, v + j)$, the results are added, and the final sum is inserted as the new pixel value $I'(u, v)$. Figure adapted from (Burger and Burge (2009a)).

Applying the LoG filter in the spatial domain to an image is a simple process as illustrated in Figure 9.2. According to (Burger and Burge (2009a)), the following steps are performed at each image position (u, v):

1) The filter matrix H is moved over the original image I such that its origin $H(0, 0)$ coincides with the current image position (u, v).

2) All filter coefficients $H(i, j)$ are multiplied with the corresponding image element $I(u + i, v + j)$, and the results are added.

3) Finally, the resulting sum is stored at the current position in the new image $I'(u, v)$

An additional speedup can be achieved by applying the Fourier transform and turning the LoG filter into a frequency filter. According to the convolution theorem (Lim (1990)) point-wise multiplication of the Fourier transformed kernel and Fourier transformed image in the frequency domain is equivalent to convolution in the spatial domain. The Fourier transform of the convolution of two functions is the product of their Fourier transforms:

$$F[h \star f] = F[h]F[f] \tag{9.15}$$

The inverse Fourier transform of the convolution of two functions is the product of their Fourier transforms:

$$F^{-1}[fh] = F^{-1}[f] \star F^{-1}[h] \tag{9.16}$$

Therefore:

$$g(x, y) = F^{-1}(F(h)F(f)) \tag{9.17}$$

The discrete Fourier transformation (DFT) and its inverse are used and performed by fast Fourier transformation (FFT) (Cooley and Tukey (1965b)) which reduces the computation time from $O(n^2m^2)$ to the almost linear complexity of $O(nm\log(nm))$ for an image with size $n \times m$.

9.1.1 Padding

An important technical detail is padding (Rao et al. (2011)). When doing a DFT the resulting frequency domain representation of the function is periodic, leading to circular convolution. This means that without padding the image properly, results from one side of the image will wrap around to the other side of the image. Padding allows space for this wrap-around to occur without contaminating actual output pixels. According to (Burger and Burge (2009a); Rao et al. (2011)) there are several different methods how padding can be done:
1) Zero-padding:
Zero-padding completes the borders of the image with zero valued pixels.
2) Boundary reflection:
Padded pixels are computed by reflecting the input image pixels about the border.
3) Pixel replication:
Pixel replication is done by copying the nearest border pixel.
4) Linear extrapolation:

Linear extrapolation seeks to extend the image as if it were continuing along a linear ramp off the edge of the image. The ramp is made up of the border pixels, the computed padded pixels and the input pixels that are a reflection of the padded pixel about the border. For each line in x (or column for y) the padding elements added to the line (or column) are a linear combination of the first and last element of that input line (or column).

5) Weighted mean:

A weighted mean, $\frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$, can be used in the following way:

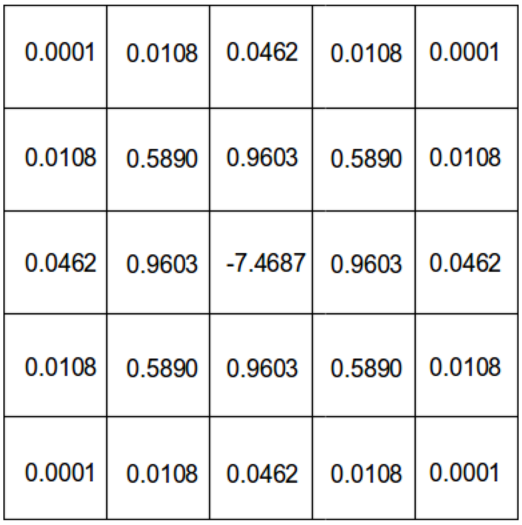
```
for $x = O_x \; to \; W - O_x$
  for $y = 0 \; to \; O_y - 1$
    $I[x * H + y] = \frac{(y + O_y + 1) * I[x * H + O_y] + (O_y - y) * I[x * H + H - O_y - 1]}{2 * O_y + 1}$
    $I[x * H - y - 1] = \frac{(O_y - y) * I[x * H + O_y] + (y + O_y + 1) * I[x * H + H - O_y - 1]}{2 * O_y + 1}$
for $y = 0 \; to \; H$
  for $x = 0 \; to \; O_x$
    $I[x * H + y] = \frac{(x + O_x + 1) * I[O_x * H + y] + (O_x - x) * I[(W - O_x - 1) * H + y]}{2 * O_x + 1}$
    $I[(W-x-1)*H+y] = \frac{(O_x - x) * I[O_x * H + y] + (x + O_x + 1) * I[(W - O_x - 1) * H + y]}{2 * O_x + 1}$
```

x, y: control variables for running through the padding area.
I: image pixel values of 1d image array, accessed in column-major order.
W: image width.
H: image height.
O_x: x offset of padding area.
O_y: y offset of padding area.

9.1.2 Implementation

The implementation of the LoG filter is based on the ImageJ (Abràmoff et al. (2004); Schneider et al. (2012)) plugin by Dimiter Prodanov, URL: `http://rsb.info.nih.gov/ij/plugins/mexican-hat/index.html`. The input for the LoG filter are the tif images taken during the protein cycles of a HiTS-FLIP experiment. The output of the LoG filter are tif images which are LoG filtered and stored separately besides the unfiltered images. These filtered images are only used during the cluster registration process.

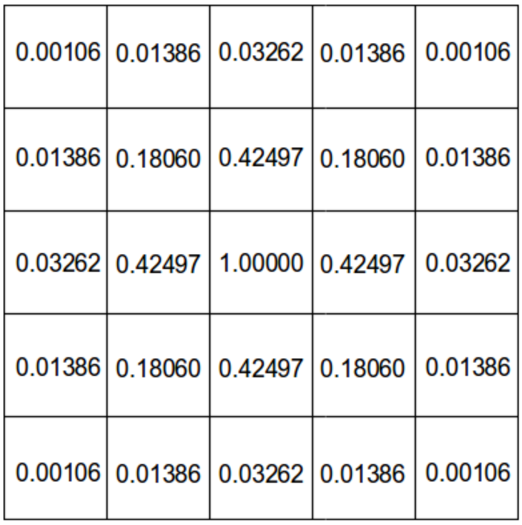
9.2 Pixel kernel of LoG filter



0.0001	0.0108	0.0462	0.0108	0.0001
0.0108	0.5890	0.9603	0.5890	0.0108
0.0462	0.9603	-7.4687	0.9603	0.0462
0.0108	0.5890	0.9603	0.5890	0.0108
0.0001	0.0108	0.0462	0.0108	0.0001

Figure 9.3: Pixel kernel of LoG filter used for filtering the protein images before cluster registration as described in section 5.2.

9.3 Pixel mask for template cluster generation



0.00106	0.01386	0.03262	0.01386	0.00106
0.01386	0.18060	0.42497	0.18060	0.01386
0.03262	0.42497	1.00000	0.42497	0.03262
0.01386	0.18060	0.42497	0.18060	0.01386
0.00106	0.01386	0.03262	0.01386	0.00106

Figure 9.4: Pixel mask for template cluster generation used for cluster registration as described in section 5.3.

9.4 HiP-FA Kds and HiTS-FLIP Kds

Oligo Sequence	*HiP-FA Kd*	*11-mer*	*rev. complement*	*HiTS-FLIP Kd*
GGTATGAGTCATGGCC	16.34	CATGACTCATA	TATGAGTCATG	3.49
GGGTATGACTCATCCC	17.9	GATGAGTCATA	TATGACTCATC	3.57
GGTATGACTCATGGCC	21.66	CATGAGTCATA	TATGACTCATG	4.97
GGTGTGACTCATGGCC	24.49	TGTGACTCATG	CATGAGTCACA	5.96
GGTGTGAGTCATGGCC	25.93	CATGACTCACA	TGTGAGTCATG	6.82
GGGTATGACTGATCCC	49.0	TATGACTGATC	GATCAGTCATA	29.36
GGTGTGACTAATGGCC	51.57	CATTAGTCACA	TGTGACTAATG	25.47
GGTCTGACTCATGGCC	56.28	TCTGACTCATG	CATGAGTCAGA	20.67
GGGTATGACACATCCC	72.9	GATGTGTCATA	TATGACACATC	38.19
GGTATGACTCTTGGCC	90.69	CAAGAGTCATA	TATGACTCTTG	55.08
GGTATGACACATGGCC	129.05	CATGTGTCATA	TATGACACATG	75.13
GGTATGACGCATGGCC	138.35	TATGACGCATG	CATGCGTCATA	73.82
GGTTGACTAATTGGCC	221.65	AATTAGTCAAC	GTTGACTAATT	94.2
GGTATGACTCGTGGCC	244.16	TATGACTCGTG	CACGAGTCATA	100.53
GGTTCAGTCATTGGCC	312.45	AATGACTGAAC	GTTCAGTCATT	151.02
GGTTTAGTCATTGGCC	334.9	GTTTAGTCATT	AATGACTAAAC	94.57
GGTATGACTAGTGGCC	371.13	TATGACTAGTG	CACTAGTCATA	218.02
GGTATGACGTATGGCC	464.89	TATGACGTATG	CATACGTCATA	222.8
GGTCTGACGCATGGCC	541.01	TCTGACGCATG	CATGCGTCAGA	356.78
GGTGTGTGACATGGCC	633.51	CATGTCACACA	TGTGTGACATG	485.0
GGTTGAGTAATTGGCC	654.88	GTTGAGTAATT	AATTACTCAAC	389.87
GGTATGACGCGTGGCC	675.05	TATGACGCGTG	CACGCGTCATA	343.61
GGTTTACGTCATGGCC	779.26	TTTACGTCATG	CATGACGTAAA	484.6
GGTATCCGTCATGGCC	796.11	TATCCGTCATG	CATGACGGATA	451.76
GGTTCACTCATTGGCC	1024.26	GTTCACTCATT	AATGAGTGAAC	875.36

Table 9.1: Listing of HiP-FA Kds and HiTS-FLIP Kds as used in Figure 5.52.

9.5 First 50 11-mers and Kds of experiment 18.08.2014

k-mer	*reverse complement*	*5 nM*	*10 nM*	*25 nM*	*125 nM*	*625 nM*	*s*	*h*	*o*	*Kd*
TATGAGTCATA	TATGACTCATA	0.5094	0.58539	0.54504	0.59656	0.65996	0.65345	1.0	0.26793	1.562
GTATGAGTCAT	ATGACTCATAC	0.48893	0.59662	0.55888	0.59435	0.7001	0.65345	1.0	0.26793	1.639
GATGACTCATA	TATGAGTCATC	0.52161	0.54095	0.56534	0.56021	0.70031	0.65345	1.0	0.26793	1.675
ATGACTCATCT	AGATGAGTCAT	0.49438	0.53619	0.55156	0.58751	0.65243	0.65345	1.0	0.26793	2.002
ATGAGTCATTT	AAATGACTCAT	0.48092	0.50015	0.57188	0.60437	0.67185	0.65345	1.0	0.26793	2.308
ATGACTCATAT	ATATGAGTCAT	0.47886	0.5008	0.55385	0.59616	0.6987	0.65345	1.0	0.26793	2.397
TGTGACTCATC	GATGAGTCACA	0.45742	0.52625	0.53085	0.60089	0.63088	0.65345	1.0	0.26793	2.533
CTATGAGTCAT	ATGACTCATAG	0.46423	0.51583	0.53634	0.54436	0.6169	0.65345	1.0	0.26793	2.576
ATGAGTCATTG	CAATGACTCAT	0.45821	0.50526	0.54544	0.59213	0.6612	0.65345	1.0	0.26793	2.634
ATGAGTCATCC	GGATGACTCAT	0.4704	0.48661	0.55131	0.57627	0.64366	0.65345	1.0	0.26793	2.645
ATATGACTCAC	GTGAGTCATAT	0.44357	0.52211	0.54127	0.61358	0.66862	0.65345	1.0	0.26793	2.655
TTATGACTCAT	ATGAGTCATAA	0.45723	0.50757	0.5344	0.58649	0.64432	0.65345	1.0	0.26793	2.683
GATGAGTCATT	AATGACTCATC	0.47392	0.47208	0.54077	0.56632	0.65326	0.65345	1.0	0.26793	2.784
GTGAGTCATCG	CGATGACTCAC	0.45643	0.47417	0.55203	0.63203	0.69053	0.65345	1.0	0.26793	2.839
AATGAGTCATC	GATGACTCATT	0.462	0.49839	0.51383	0.5503	0.60172	0.65345	1.0	0.26793	2.85
TGATGACTCAT	ATGAGTCATCA	0.46976	0.4677	0.54736	0.5525	0.64873	0.65345	1.0	0.26793	2.858
TTATGAGTCAT	ATGACTCATAA	0.45605	0.47486	0.55826	0.56793	0.67091	0.65345	1.0	0.26793	2.883
TAATGAGTCAT	ATGACTCATTA	0.46094	0.47192	0.54076	0.58924	0.64198	0.65345	1.0	0.26793	2.917
TGTGACTCATA	TATGAGTCACA	0.46651	0.47027	0.53736	0.55506	0.64264	0.65345	1.0	0.26793	2.921
AATGACTCATA	TATGAGTCATT	0.45821	0.47303	0.54988	0.56019	0.63575	0.65345	1.0	0.26793	2.931
GGTGAGTCATA	TATGACTCACC	0.44559	0.49564	0.52003	0.54077	0.62309	0.65345	1.0	0.26793	3.05
ATGACTCATCG	CGATGAGTCAT	0.4766	0.45116	0.52399	0.53008	0.5873	0.65345	1.0	0.26793	3.089
ATGACTCACCT	AGGTGAGTCAT	0.4249	0.49133	0.54957	0.61397	0.66863	0.65345	1.0	0.26793	3.093
CTATGACTCAT	ATGAGTCATAG	0.47359	0.47573	0.48307	0.4886	0.59275	0.65345	1.0	0.26793	3.158
GGTGAGTCATC	GATGACTCACC	0.45643	0.45047	0.54348	0.58211	0.64687	0.65345	1.0	0.26793	3.166
TGATGAGTCAT	ATGACTCATCA	0.46278	0.46766	0.5051	0.54184	0.64795	0.65345	1.0	0.26793	3.174
ATATGACTCAT	ATGAGTCATAT	0.45382	0.4757	0.50297	0.55118	0.61736	0.65345	1.0	0.26793	3.217
GATGACTCATC	GATGAGTCATC	0.45853	0.44545	0.53345	0.57123	0.61337	0.65345	1.0	0.26793	3.266
TGTGAGTCATA	TATGACTCACA	0.43687	0.46754	0.53154	0.59219	0.65074	0.65345	1.0	0.26793	3.295
ATGAGTCATAC	GTATGACTCAT	0.43986	0.47635	0.51545	0.54151	0.56698	0.65345	1.0	0.26793	3.349
CATGACTCATC	GATGAGTCATG	0.45836	0.45761	0.50302	0.53251	0.66041	0.65345	1.0	0.26793	3.355
GTGTGAGTCAT	ATGACTCACAC	0.43544	0.46773	0.51057	0.58776	0.6457	0.65345	1.0	0.26793	3.435
ATGAGTCATCG	CGATGACTCAT	0.44752	0.46422	0.50615	0.52511	0.59696	0.65345	1.0	0.26793	3.442
AGATGACTCAT	ATGAGTCATCT	0.43984	0.4439	0.53489	0.57949	0.65917	0.65345	1.0	0.26793	3.491
CATGACTCATA	TATGAGTCATG	0.43941	0.44418	0.53596	0.58288	0.62563	0.65345	1.0	0.26793	3.493
GATGAGTCATA	TATGACTCATC	0.43202	0.47573	0.50357	0.51423	0.57922	0.65345	1.0	0.26793	3.565
ATGACTCATTG	CAATGAGTCAT	0.41149	0.48292	0.52718	0.53538	0.63319	0.65345	1.0	0.26793	3.573
GTGAGTCATAC	GTATGACTCAC	0.42419	0.46348	0.50789	0.52318	0.60191	0.65345	1.0	0.26793	3.759
AAATGAGTCAT	ATGACTCATTT	0.42864	0.46674	0.47179	0.56492	0.57778	0.65345	1.0	0.26793	3.836
AGTGAGTCATC	GATGACTCACT	0.41998	0.44252	0.51264	0.60542	0.68125	0.65345	1.0	0.26793	3.863
GTGTGACTCAT	ATGAGTCACAC	0.43258	0.4532	0.48668	0.54466	0.59946	0.65345	1.0	0.26793	3.864
CTATGAGTCAC	GTGACTCATAG	0.41806	0.44676	0.51245	0.58565	0.67234	0.65345	1.0	0.26793	3.878
AATGAGTCATA	TATGACTCATT	0.42893	0.47635	0.46883	0.47887	0.54602	0.65345	1.0	0.26793	3.898
ATGTGACTCAT	ATGAGTCACAT	0.41178	0.46572	0.49654	0.57378	0.60188	0.65345	1.0	0.26793	3.898
TGTGAGTCATT	AATGACTCACA	0.40209	0.44162	0.5406	0.60928	0.6582	0.65345	1.0	0.26793	3.937
GGATGAGTCAT	ATGACTCATCC	0.42637	0.45758	0.48481	0.52938	0.60988	0.65345	1.0	0.26793	3.937
GGTGACTCATA	TATGAGTCACC	0.42592	0.44821	0.47313	0.55936	0.61677	0.65345	1.0	0.26793	4.084
TTATGACTAAT	ATTAGTCATAA	0.36986	0.46365	0.53787	0.61997	0.71686	0.65345	1.0	0.26793	4.104
GTGAGTCATAG	CTATGACTCAC	0.4078	0.45395	0.49708	0.54414	0.6328	0.65345	1.0	0.26793	4.129
CCTATGACTCA	TGAGTCATAGG	0.40384	0.46154	0.48123	0.57557	0.64035	0.65345	1.0	0.26793	4.148

Table 9.2: First 50 11-mers and HiTS-FLIP Kds of experiment 18.08.2014 as used in Figure 5.55.

9.6 First 50 11-mers ranked at 125 nM of exp. 18.08.2014

k-mer	*reverse complement*	*5 nM*	*10 nM*	*25 nM*	*125 nM*	*625 nM*
GTGAGTCATCG	CGATGACTCAC	2.9897	3.0629	3.38428	3.71443	3.9559
TTATGACTAAT	ATTAGTCATAA	2.63238	3.01947	3.32583	3.66469	4.06457
ATGACTCACCT	AGGTGAGTCAT	2.85954	3.13375	3.37411	3.63992	3.86551
ATATGACTCAC	GTGAGTCATAT	2.93663	3.26076	3.33985	3.63832	3.86547
TGAGTCATCAA	TTGATGACTCA	2.72001	2.86207	3.22004	3.62958	3.88587
TGTGAGTCATT	AATGACTCACA	2.76539	2.92856	3.3371	3.62056	3.82245
GATATGACTCA	TGAGTCATATC	2.68541	2.99135	3.05187	3.62009	3.84844
AGTGAGTCATC	GATGACTCACT	2.83923	2.93227	3.2217	3.60462	3.9176
ATGAGTCATTT	AAATGACTCAT	3.09078	3.17014	3.46621	3.60029	3.8788
ATGACTCACTG	CAGTGAGTCAT	2.59429	2.72297	3.12548	3.5867	3.80671
TGTGACTCATC	GATGAGTCACA	2.99377	3.27787	3.29683	3.58591	3.70971
GATTAGTCATA	TATGACTAATC	2.52781	2.76174	3.0561	3.58234	3.86143
TTCATGACTCA	TGAGTCATGAA	2.136	2.23907	2.7688	3.57634	4.10505
AAAATGAGTCA	TGACTCATTTT	2.39901	2.66312	2.93668	3.56961	3.63804
TATGAGTCATA	TATGACTCATA	3.20831	3.52195	3.35543	3.56806	3.82972
ATGACTCATAT	ATATGAGTCAT	3.08227	3.1728	3.39177	3.56639	3.98963
ATTATGACTCA	TGAGTCATAAT	2.56942	2.72326	3.17365	3.56601	3.82771
GTATGAGTCAT	ATGACTCATAC	3.12382	3.56829	3.41254	3.55893	3.99539
TCTGACTCATT	AATGAGTCAGA	2.49694	2.5861	2.98953	3.55822	3.87298
TGTGAGTCATA	TATGACTCACA	2.90896	3.03554	3.29971	3.55	3.79166
ATGAGTCATTG	CAATGACTCAT	2.99703	3.19123	3.35706	3.54978	3.83484
AGTGAGTCATG	CATGACTCACT	2.50189	2.66709	3.04382	3.53849	4.06535
TAATGAGTCAT	ATGACTCATTA	3.00832	3.05363	3.33776	3.53783	3.75553
TGACTCATATC	GATATGAGTCA	2.4091	2.61786	3.08495	3.53375	3.98912
GTGTGAGTCAT	ATGACTCACAC	2.90306	3.03631	3.21316	3.53172	3.77089
ATGACTCATCT	AGATGAGTCAT	3.14633	3.31889	3.38232	3.53068	3.79863
TTATGACTCAT	ATGAGTCATAA	2.99299	3.20074	3.31149	3.52651	3.76518
CTATGAGTCAC	GTGACTCATAG	2.83133	2.94977	3.22092	3.52304	3.88081
ATGAGTCACAG	CTGTGACTCAT	2.74099	2.92125	3.12247	3.51667	3.73115
ATGACTCACTT	AAGTGAGTCAT	2.62523	2.83754	3.09863	3.51389	3.8083
CATATGACTCA	TGAGTCATATG	2.25206	2.32113	2.82388	3.51208	3.89855
ATGACTCATGA	TCATGAGTCAT	2.62321	2.78808	3.17572	3.51175	3.68855
CATGACTCATA	TATGAGTCATG	2.91945	2.93911	3.31792	3.51159	3.68804
ACATGAGTCAC	GTGACTCATGT	2.44077	2.6947	3.01403	3.51065	3.76905
GGTGAGTCATC	GATGACTCACC	2.9897	2.96507	3.34899	3.50843	3.77569
TCGATGACTCA	TGAGTCATCGA	2.49706	2.63466	2.98263	3.50798	3.74253
GTATTAGTCAT	ATGACTAATAC	2.54593	2.66158	3.02763	3.50788	3.99939
ATGACTCACAG	CTGTGAGTCAT	2.67608	2.84112	3.22176	3.50674	3.67334
TCAATGACTCA	TGAGTCATTGA	2.45188	2.4456	3.12649	3.5061	3.94763
CATGACTCATT	AATGAGTCATG	2.75076	2.8929	3.20336	3.50602	3.83574
ATGTGAGTCAT	ATGACTCACAT	2.60487	2.77032	3.17036	3.50439	3.75787
TTGTGAGTCAT	ATGACTCACAA	2.72186	2.90006	3.11572	3.50072	3.79934
TGAGTCATCAC	GTGATGACTCA	2.34131	2.53457	2.93627	3.49943	3.86426
AGATGACTCAT	ATGAGTCATCT	2.92122	2.93799	3.3135	3.49758	3.82648
GTGAGTCATTT	AAATGACTCAC	2.6112	2.80436	3.13299	3.49503	3.78482
ACTATGACTCA	TGAGTCATAGT	2.44169	2.66124	2.97468	3.49396	3.90605
TGTGACTCATG	CATGAGTCACA	2.59835	2.69527	3.06662	3.49033	3.81104
ATGAGTCACGC	GCGTGACTCAT	2.56056	2.63998	2.9431	3.48928	3.82363
GTGACTCATAT	ATATGAGTCAC	2.67399	2.98078	3.25098	3.489	3.79223
ATGAGTCATCC	GGATGACTCAT	3.04733	3.11424	3.38131	3.48431	3.76245

Table 9.3: First 50 11-mers ranked at 125 nM of experiment 18.08.2014.

9.7 HiP-FA Kds and HiTS-FLIP Kds by heuristic and ML ranking

11-mer	*rev.comp.*	*HiP-FA Kd*	*Heuristic ranking Kd*	*ML ranking Kd*
CATGACTCATA	TATGAGTCATG	16.34	3.49	5.71
GATGAGTCATA	TATGACTCATC	17.9	3.57	3.05
CATGAGTCATA	TATGACTCATG	21.66	4.97	6.74
TGTGACTCATG	CATGAGTCACA	24.49	5.96	10.2
CATGACTCACA	TGTGAGTCATG	25.93	6.82	9.28
TATGACTGATC	GATCAGTCATA	49.0	29.36	52.62
CATTAGTCACA	TGTGACTAATG	51.57	25.47	31.34
TCTGACTCATG	CATGAGTCAGA	56.28	20.67	31.3
GATGTGTCATA	TATGACACATC	72.9	38.19	53.68
CAAGAGTCATA	TATGACTCTTG	90.69	55.08	63.65
CATGTGTCATA	TATGACACATG	129.05	75.13	126.77
TATGACGCATG	CATGCGTCATA	138.35	73.82	91.39
AATTAGTCAAC	GTTGACTAATT	221.65	94.2	134.35
TATGACTCGTG	CACGAGTCATA	244.16	100.53	129.1
AATGACTGAAC	GTTCAGTCATT	312.45	151.02	156.27
GTTTAGTCATT	AATGACTAAAC	334.9	94.57	125.91
TATGACTAGTG	CACTAGTCATA	371.13	218.02	229.43
TATGACGTATG	CATACGTCATA	464.89	222.8	455.95
TCTGACGCATG	CATGCGTCAGA	541.01	356.78	564.66
CATGTCACACA	TGTGTGACATG	633.51	485.0	1171.25
GTTGAGTAATT	AATTACTCAAC	654.88	389.87	396.29
TATGACGCGTG	CACGCGTCATA	675.05	343.61	373.0
TTTACGTCATG	CATGACGTAAA	779.26	484.6	697.52
TATCCGTCATG	CATGACGGATA	796.11	451.76	522.22
GTTCACTCATT	AATGAGTGAAC	1024.26	875.36	830.72

Table 9.4: Listing of HiP-FA Kds and HiTS-FLIP Kds by heuristic and ML ranking as used in Figure 5.54.

9.8 Fits of HiTS-FLIP Kds with subtraction of global offset

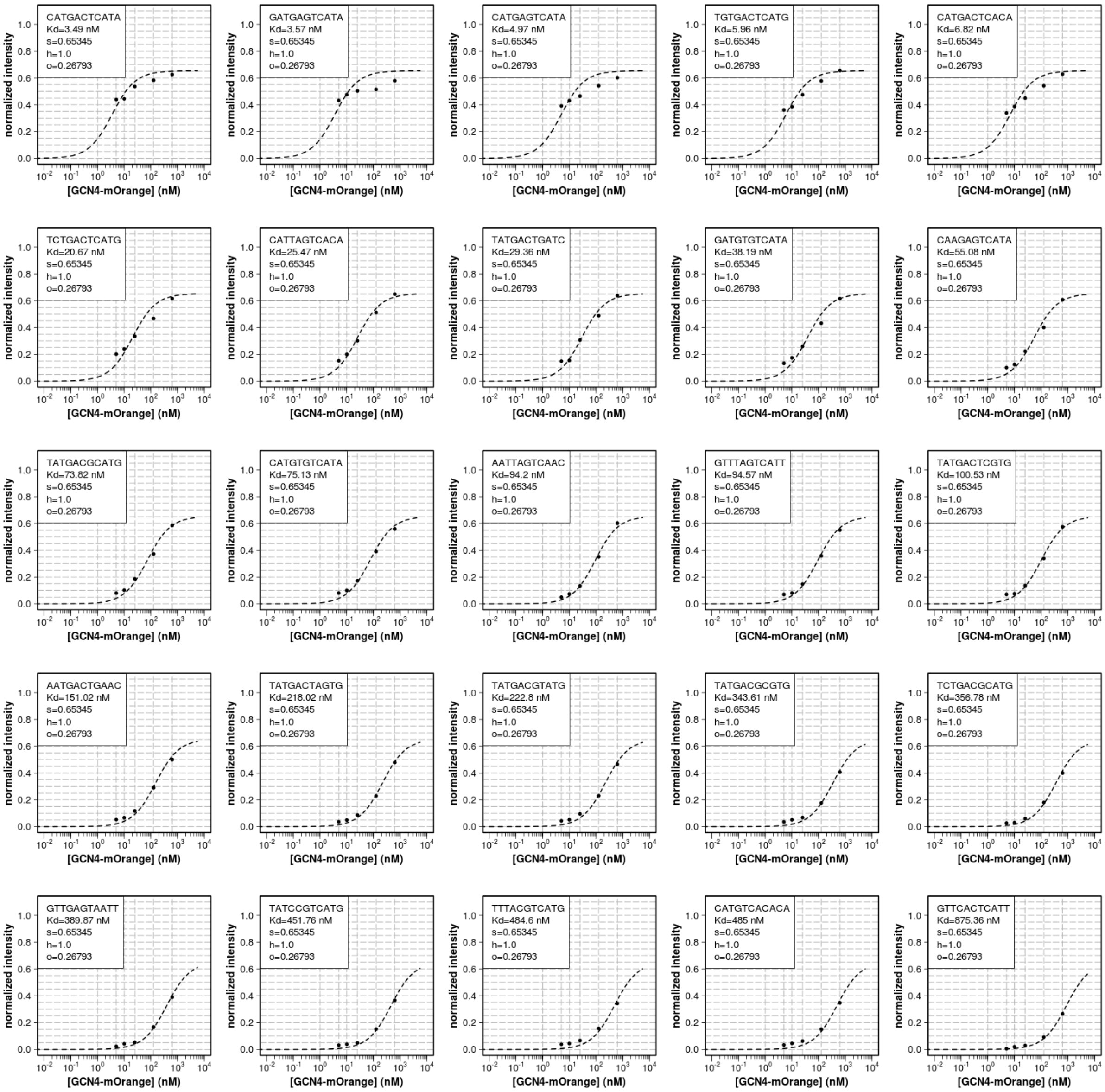


Figure 9.5: Fits for HiTS-FLIP Kds with parameters as used in Figure 5.52.

9.9 Fits of HiTS-FLIP Kds with subtraction of median of cluster intensities per concentration

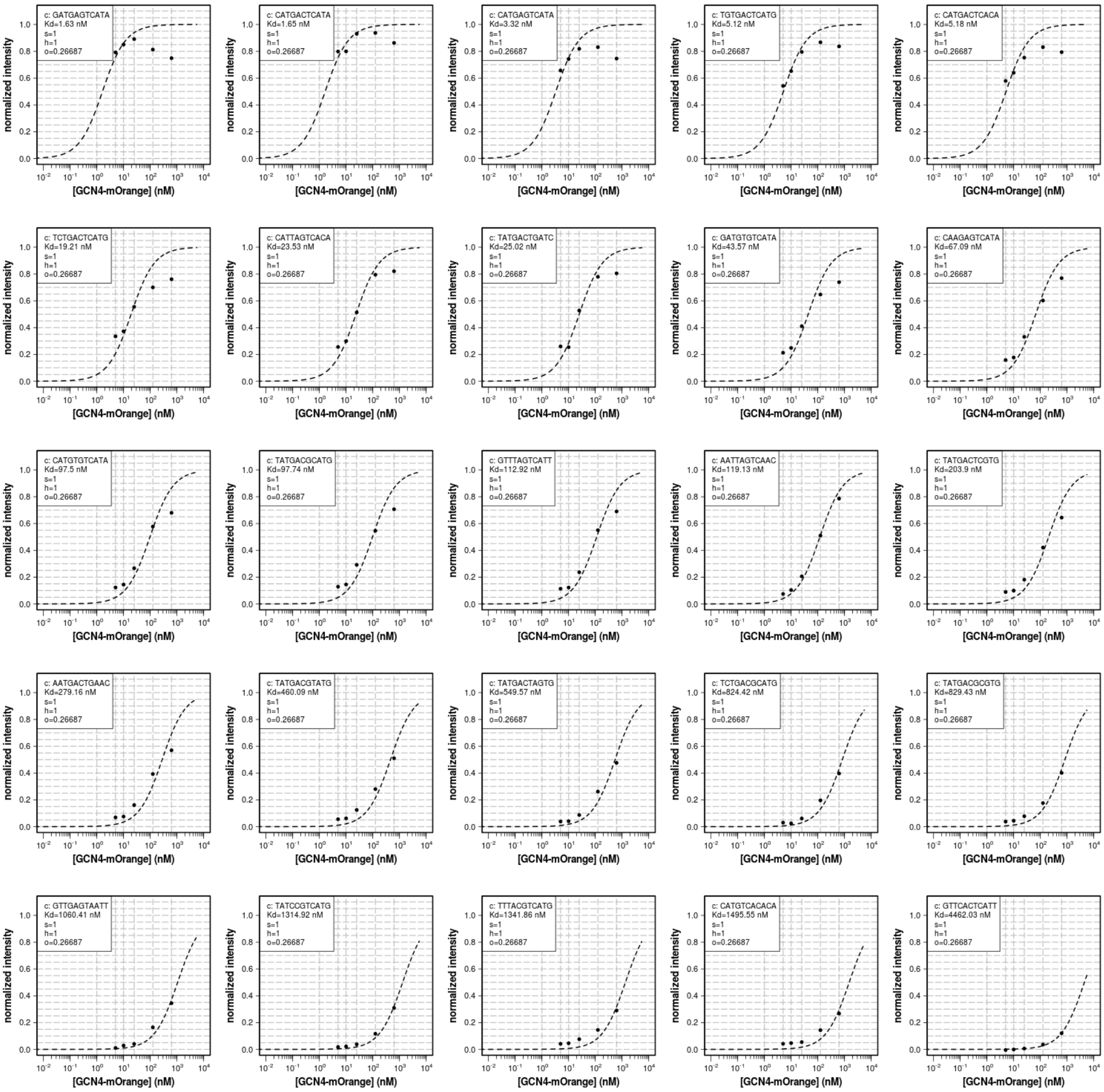


Figure 9.6: Fits for HiTS-FLIP Kds with subtraction of median of cluster intensities per concentration.

9.10 Summary of HiTS-FLIP methods, parameters, input and output

LoG filter (Section 5.2)
Goal:
Reducing noise and emphasizing edges to improve separation of DNA clusters for protein images.
Parameters:

- $\sigma = 0.764$.
- 5×5 pixel kernel, see 9.3.

Input:
tif image from protein cycle.
Output:
LoG filtered image.

Cluster registration (Section 5.3)
Goal:
Aligning protein images to connect fluorescent cluster intensities with nucleotide sequences.
Parameters:

- template cluster: amplitude $A = 1.0$.
- template cluster: $\sigma = 0.7644$.
- 5×5 pixel mask, see 9.4.

Input:

- tif image from protein cycle.
- pos file containing the template cluster positions.

Output:
Translational offset $\Delta x, \Delta y$ for each protein image.

Local region search (Section 5.4)
Goal:
Overlapping of mapped cluster positions with local maxima in a protein image.
Parameters:

- local neighborhood: distance 1 pixel from cluster position.

Input:
x,y coordinates of cluster positions.
Output:

Shifted clusters (on average 10% - 20% of all clusters per tile) which overlap again with local maxima.

Image normalization (Section 5.5)

Goal:

Correction of uneven illumination in the protein images by Gaussian based filtering.

Parameters:

- Gaussian filter: $\sigma = 30$ pixels.
- Weighting factor: $\frac{B_i^{global}}{B_1^{global}}$ for $i = 2..n$. Details in section 5.5.6.3.

Input:

Tif image of protein cycle.

Output:

Normalized protein image with reduced uneven illumination.

Intensity extraction based on weighted area coverage (Section 5.6.9)

Goal:

Extraction of cluster intensities from protein images with the weighted area coverage method.

Parameters:

- $A = 1.5^2$ pixels.
- $w_c = 5.0$.
- $w_n = 0.9$.

Input:

- protein image.
- x,y coordinates of cluster positions.

Output:

Cluster intensities.

Dust particle detection (Section 5.7.2)

Goal:

Reducing false positives by removing dust particles and affected clusters from protein images.

Parameters:

- threshold: $30 - 30000$ pixels for classifying dust particles.

Input:

Protein image.

Outcome:

Set of pixels identifying dust particles.

Air bubble detection (Section 5.7.1)
Goal:
Reducing false positives by removing air bubbles and affected clusters from protein images.
Parameters:

- threshold: > 30000 pixels for classifying air bubbles.

Input:
Protein image.
Outcome:
Set of pixels identifying air bubbles.

DNA sequence filtering (Section 5.8)
Goal:
Removing clusters with erroneous bases.
Parameters:

- FASTA quality score: Q30.

Input:
FASTQ files.
Outcome:
Filtered cluster sequences with high quality base calls.

K-mer ranking (Section 5.9.1.5)
Goal:
Ranking k-mers of different length for DNA motif finding by heuristic ranking without cluster deletion.
Parameters:

- heuristic ranking without cluster deletion.

Input:

- length of the k-mer to be ranked.
- number of ranked k-mers.
- concentration at which the ranking should be achieved.
- cluster sequences.
- different cluster intensities for the increasing concentrations.

Outcome:
Ranked k-mer motifs with their different intensities.

Affinity quantification (Section 5.10)
Goal:
Determining dissociation constants for each ranked k-mer.
Parameters:

- s: scaling factor obtained from the top binding k-mer and applied for all weaker binding k-mers.
- $[TF]$: transcription factor concentration used in the experiment.
- Kd: the dissociation constant of the TF to the DNA sequence.
- h: the Hill coefficient of binding.
- o: the global offset, an estimate for the unspecific binding by using the median of the dimmest 0.1% of all the ranked k-mers at the smallest concentration. Regarding the experiment 18.08.2014 the value was $o = 0.26793$ of the ranked 11-mers for the concentration at 5nM.

Input:
Intensities at each concentration for each k-mer.
Outcome:
Dissociation constants for each ranked k-mer.

9.11 Position Frequency Matrix for Aligned Matrix GCN4

Nucleotide	*Prob.*								
A	0.547	0.012	0.011	0.964	0.009	0.015	0.073	0.915	0.038
C	0.059	0.010	0.026	0.012	0.657	0.010	0.891	0.022	0.352
G	0.363	0.011	0.889	0.013	0.323	0.014	0.024	0.016	0.060
T	0.031	0.968	0.075	0.012	0.010	0.962	0.013	0.046	0.550

Table 9.5: Position Frequency Matrix for Aligned Matrix GCN4 based on the ScerTF database (Spivak and Stormo (2012)). Probabilities are as used in Figure 7.5.

9.12 Experiment by Nutiu et al.

Figure 9.7 shows the analysis plots for the HiTS-FLIP experiment lane 2 as performed by (Nutiu et al. (2011)).

The 8-mer ranking yields ATGACTCA (TGAGTCAT) and ATGAGTCA (TGACTCAT) as the first two top placed motifs which are extension of the known pseudosymmetric 7-mer sequence 5'-TGA(C/G)-TCA-3' Oliphant et al. (1989). Even though a washing step of 2 min before each imaging cycle was applied to reduce unspecific binding, the intensity course does not show saturation for the best binder.

There seems to be saturation occurring for the motif ATGACTCA (TGAGTCAT), however not already at 625 nM.

The half-sites TGAC or TGAG of the dimer consensus motif 5'-TGA(C/G)-TCA-3' Oliphant et al. (1989) are enriched and predominantly bound among the first 200 ranked 8-mers. As pointed out by Hollenbeck and Oakley (2000) GCN4 can bind with high-affinity and in a specific manner to DNA sites containing only the single consensus half-site 5'-TGAC-3'.

The Hamming distance of the ranked 8-mer motifs from the 8-mer consensus ATGACTCA (TGAGTCAT) overall increases as the ranking proceeds deeper into the sequence space resulting in weaker GCN4 binding motifs.

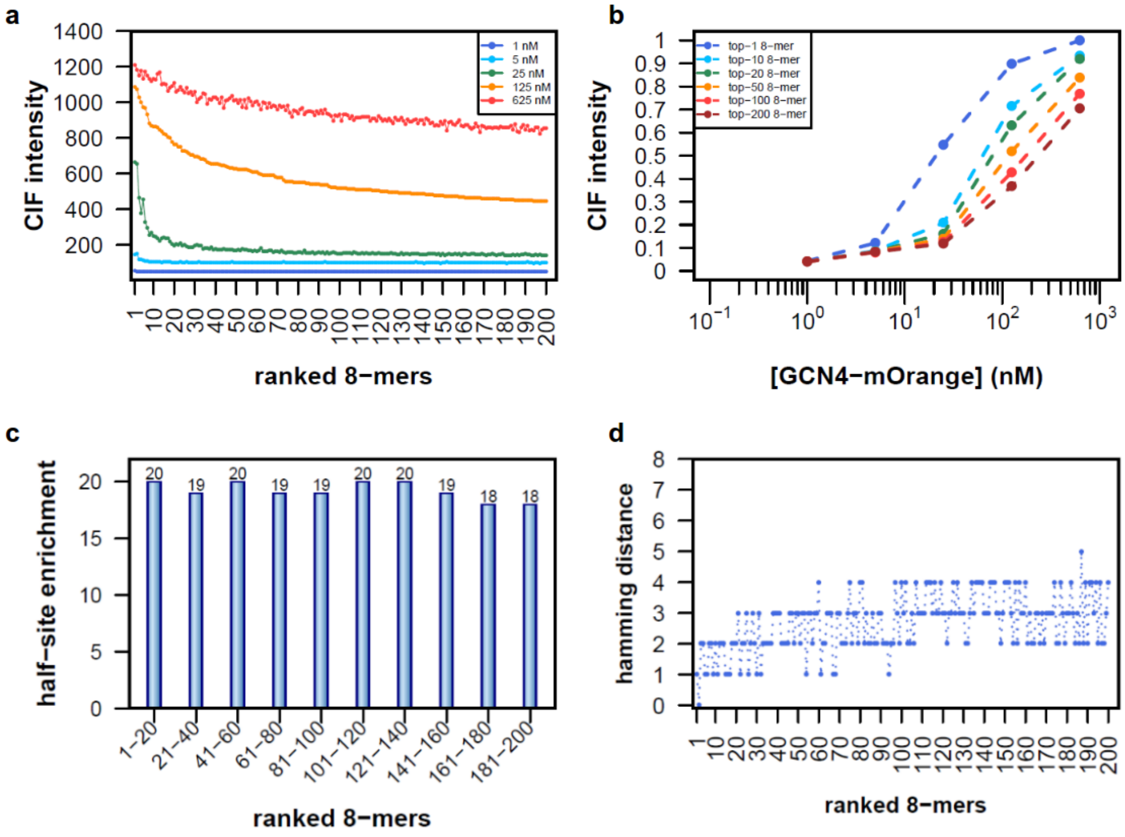


Figure 9.7: Analyses of the experiment by (Nutiu et al. (2011)). (a) Intensity course for the first 200 ranked 8-mers ranked at 125 nM on lane 2. (b) Binding curves for selected 8-mers ranked at 125 nM on lane 2. (c) Enrichment of half-sites (TGAC or TGAG) for the first 200 8-mers ranked at 125 nM on lane 2. (d) Hamming distance from consensus ATGACTCA for the first 200 8-mers ranked at 125 nM on lane 2.

9.13 Experiment 03.04.2013

In the following section the details regarding the experiment 03.04.2013 are described.

9.13.1 Sample

As sample GCN4 fused with mOrange was used as described previously.

9.13.2 Flow cell buffer

The flow cell buffer was composed of PBS + 0.3 mg/ml BSA + 0.1% Tween-20.

9.13.3 Protocol

The random DNA library N25 was used as described previously. Data from lane 4 were employed. Five concentrations were applied, 1 nM, 5 nM, 25 nM, 125 nM, 625 nM. Delivery rate of the protein solution was 50 µl/min. Equilibration time was 1 h at 20 °C. No washing was applied, the next concentration level was continuously titrated into the flow cell. The protein cycles were: cycle 44: 1 nM, cycle 45: 5 nM, cycle 46: 25 nM, cycle 47: 125 nM, cycle 48: 625 nM.

9.13.4 Data analysis

Figure 9.8 shows the main analysis plots for the experiment 03.04.2013 lane 4.

The 8-mer ranking yields TTAGATAA (TTATCTAA) and TAGATAAG (CTTATCTA) as the first two top placed motifs, which do not contain the half-sites TGAC or TGAG. The 8-mer consensus ATGACTCA (TGAGTCAT) does not occur among the first 200 ranks. The intensity courses do not decline gradually and most DNA clusters seem to be bound non-specifically.

No saturation is occurring here, GCN4 (or at least mOrange) molecules keep aggregating at the DNA clusters. The first ranked 8-mer motif is similarly bound as the 200th ranked 8-mer motif.

The half-sites TGAC or TGAG of the dimer consensus motif 5'-TGA(C/G)-TCA-3' are almost completely depleted at the first 20 ranked 8-mer motifs.

There is no increase in Hamming distance from the 8-mer consensus ATGACTCA (TGAGTCAT) observable.

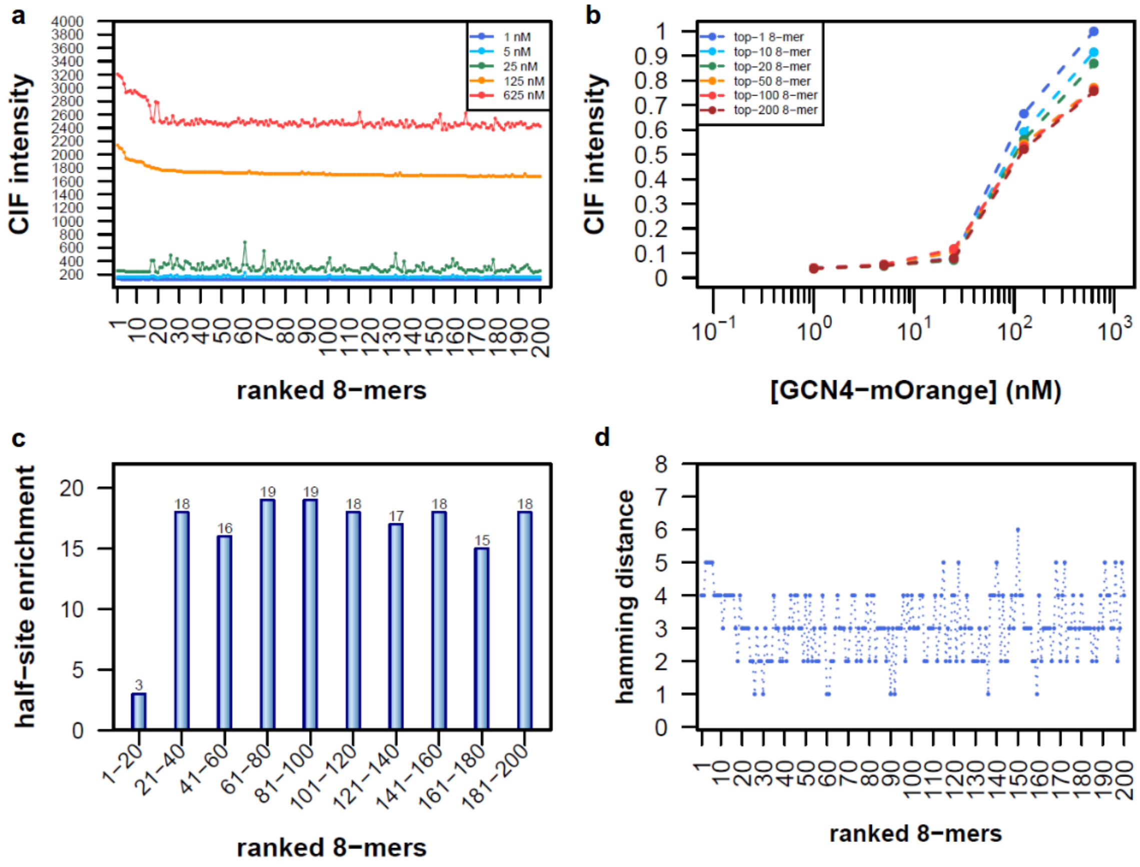


Figure 9.8: Analyses of the experiment 03.04.2013 lane 4. (a) Intensity course for the first 200 ranked 8-mers ranked at 125 nM on lane 4. (b) Binding curves for selected 8-mers ranked at 125 nM on lane 4. (c) Enrichment of half-sites (TGAC or TGAG) for the first 200 8-mers ranked at 125 nM on lane 4. (d) Hamming distance from consensus ATGACTCA for the first 200 8-mers ranked at 125 nM on lane 4.

9.14 Experiment 13.06.2013

In the following section the details regarding the experiment 13.06.2013 are described.

9.14.1 Sample

As sample GCN4 fused with mOrange was used as described previously.

9.14.2 Flow cell buffer

The flow cell buffer was composed of PBS + 0.3 mg/ml BSA + 0.1% Tween-20.

9.14.3 Protocol

The random DNA library N25 was used as described previously. Data from lane 4 were employed. Five concentrations were applied, 1 nM, 5 nM, 25 nM, 125 nM, 625 nM. Delivery rate of the protein solution was 50 µl/min. Equilibration time was 1 h at 20 °C. No washing was applied during the protein cycles, the next concentration level was titrated into the flow cell. Washing in between with PBS/TWEEN/BSA and PBS/Tween. The protein cycles were: cycle 44: 1 nM, cycle 45: 5 nM, cycle 46: 25 nM, cycle 47: 125 nM, cycle 48: 625 nM.

9.14.4 Data analysis

Figure 9.9 shows the main analysis plots for the experiment 03.04.2013 lane 4.
The 8-mer ranking yields AAGAGTCA (TGACTCTT) and AGTCATGT (ACATGACT) as the first two top placed motifs. The 8-mer consensus ATGACTCA (TGAGTCAT) occurs at rank 4. The intensity courses do not decline and all DNA clusters seem to be bound non-specifically.
No saturation is occurring here, GCN4 (or at least mOrange) molecules keep aggregating at the DNA clusters in an equal fashion.
The half-sites TGAC or TGAG of the dimer consensus motif 5'-TGA(C/G)-TCA-3' are enriched.
There is only a small increase in the Hamming distance observable from the 8-mer consensus ATGACTCA (TGAGTCAT).

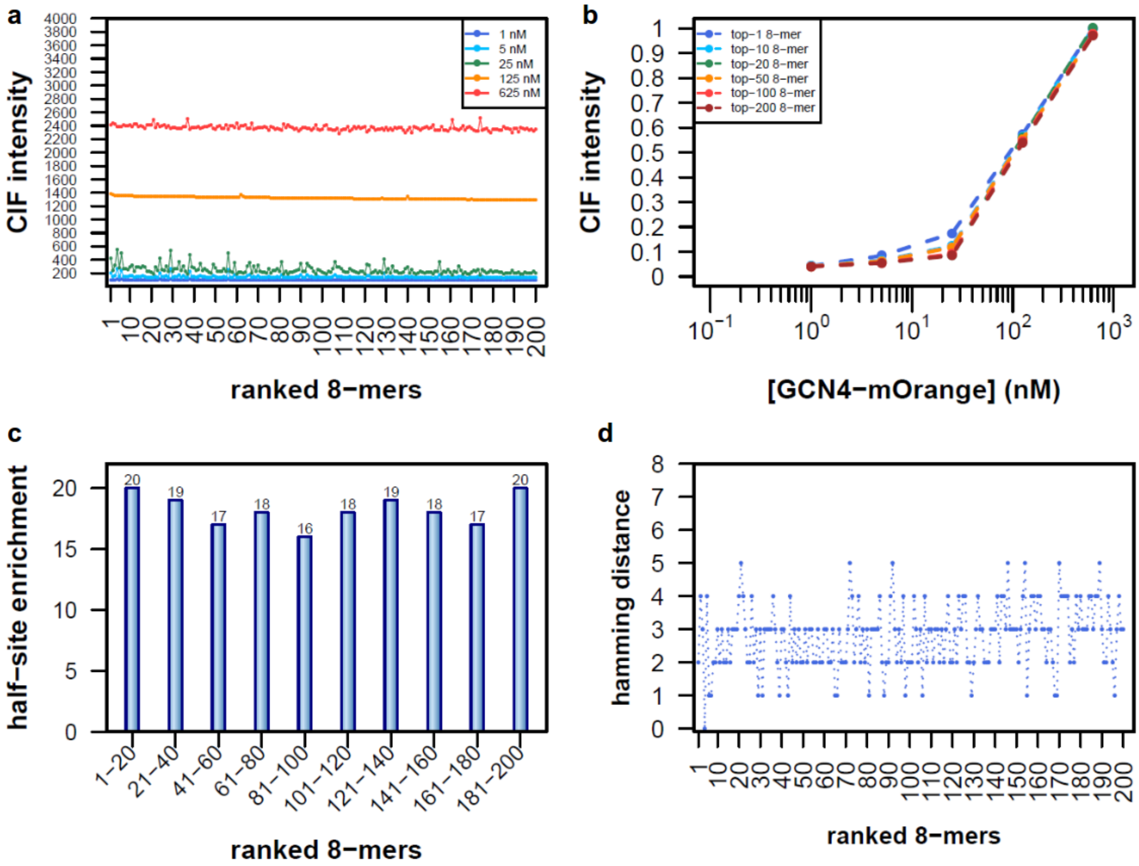


Figure 9.9: Analyses of the experiment 13.06.2013 lane 4. (a) Intensity course for the first 200 ranked 8-mers ranked at 125 nM on lane 4. (b) Binding curves for selected 8-mers ranked at 125 nM on lane 4. (c) Enrichment of half-sites (TGAC or TGAG) for the first 200 8-mers ranked at 125 nM on lane 4. (d) Hamming distance from consensus ATGACTCA for the first 200 8-mers ranked at 125 nM on lane 4.

9.15 Experiment 28.03.2014

In the following section the details regarding the experiment 28.03.2014 are described.

9.15.1 Sample

As sample GCN4 fused with mOrange was used as described previously.

9.15.2 Flow cell buffer

The following flow cell buffer compositions were applied.
Lane 1: PBS + 0.3 mg/ml BSA + 0.1% Tween-20.
Lane 2: PBS + 0.3 mg/ml BSA + 0.1% Tween-20 + 200 ng/ml poly(dI-dC).
Lane 3: PBS + 0.3 mg/ml BSA + 0.1% Tween-20.

9.15.3 Protocol

The random DNA library N25 was used. Five concentrations were applied, i.e. 1 nM, 5 nM, 25 nM, 125 nM, 625 nM.
Lane 1:
Equilibration time was 2 h at 20 °C. No washing was applied during the protein cycles, the next concentration level was titrated into the flow cell.
Lane 2:
Equilibration time was 1 h at 20 °C. No washing was applied during the protein cycles, the next concentration level was titrated into the flow cell.
During protein cycles 37-41, the following delivery was used:
150 µl protein solution
10 min wait
10 µl protein solution
10 min wait
10 µl protein solution
10 min wait
10 µl protein solution
1:30 h wait
Lane 3:
Same procedure as in (Nutiu et al. (2011)), i.e. 30 min equilibration time and 2 min wash step.
The protein cycles on lane 1 were: cycle 37: 1 nM, cycle 38: 5 nM, cycle 39: 25 nM, cycle 40: 125 nM, cycle 41: 625 nM.

The protein cycles on lane 2 were: cycle 42: 1 nM, cycle 43: 5 nM, cycle 44: 25 nM, cycle 45: 125 nM, cycle 46: 625 nM.
The protein cycles on lane 3 were: cycle 47: 1 nM, cycle 48: 5 nM, cycle 49: 25 nM, cycle 50: 125 nM, cycle 51: 625 nM.
There was a problem with lane 2, cycles 42 and 43 were performed correctly. Cycles 44 to 46 occurred without protein solution pumped into the flow cell. XML protocol was changed and restarted by cycle 44. Between cycle 42 and 43 there was a break of 3 h, and after cycle 43 washing occurred.

9.15.4 Data analysis

The following sections describe the main analysis plots for the experiment 28.03.2014.

9.15.4.1 Lane 1

Analysis results for lane 1 illustrated by Figure 9.10.
The 8-mer ranking yields ATGAGTCA (TGACTCAT) and ATGACTCA (TGAGTCAT) as the first two top placed motifs. The intensity courses do not decline and basically all DNA clusters seem to be bound non-specifically.
No saturation is occurring here, GCN4 (or at least mOrange) molecules keep aggregating at the DNA clusters in a similar way.
The half-sites TGAC or TGAG of the dimer consensus motif 5'-TGA(C/G)-TCA-3' are less enriched here.
Overall there is an increase in Hamming distances from the 8-mer consensus ATGACTCA (TGAGTCAT) for increasing ranking depth observable.

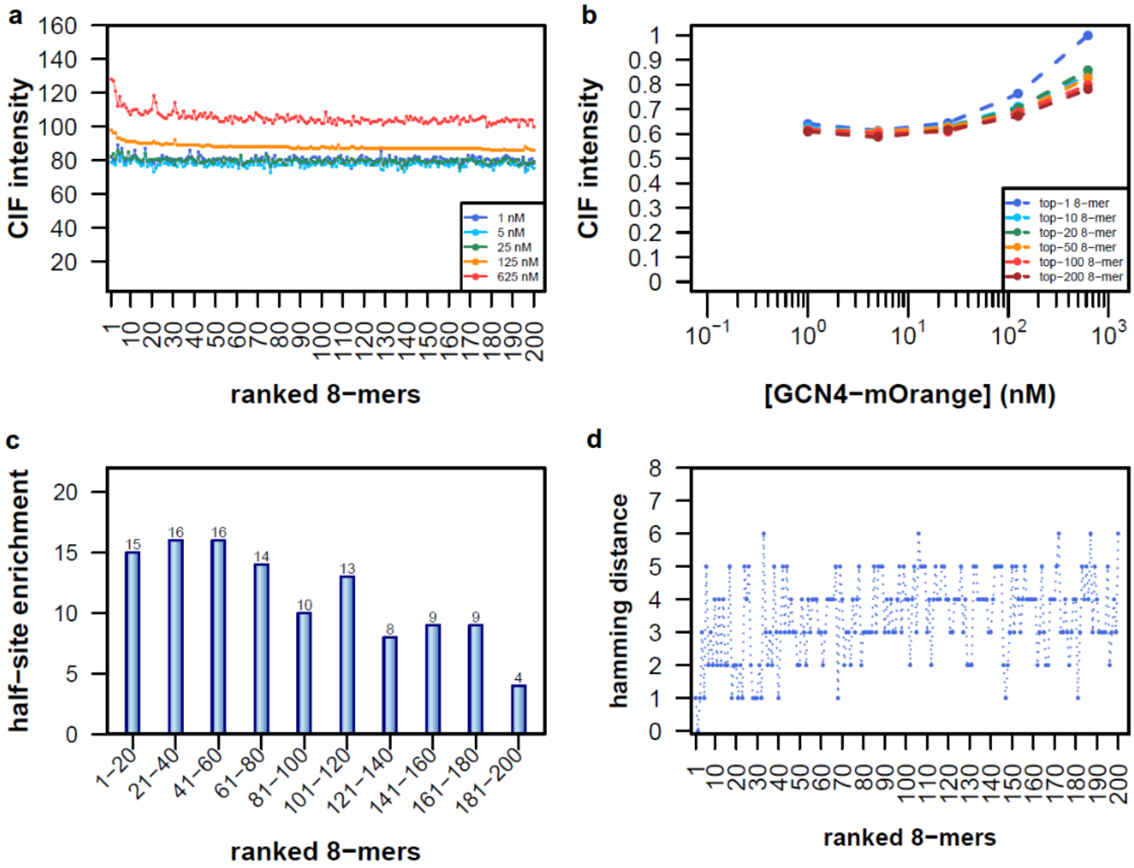


Figure 9.10: Analyses of the experiment 28.03.2014 lane 1. (a) Intensity course for the first 200 ranked 8-mers ranked at 125 nM on lane 1. (b) Binding curves for selected 8-mers ranked at 125 nM on lane 1. (c) Enrichment of half-sites (TGAC or TGAG) for the first 200 8-mers ranked at 125 nM on lane 1. (d) Hamming distance from consensus ATGACTCA for the first 200 8-mers ranked at 125 nM on lane 1.

9.15.4.2 Lane 2

Analysis results for lane 2 illustrated by Figure 9.11.

The 8-mer ranking yields ATGAGTCA (TGACTCAT) and ATGACTCA (TGAGTCAT) as the first two top placed motifs. The intensity courses do not decline after the 10th rank and basically all DNA clusters seem to be bound non-specifically.

No saturation is occurring here, GCN4 (or at least mOrange) molecules keep aggregating at the DNA clusters in a similar way. In addition, the intensities at 1 nM are higher than at 5 nM, and at 25 nM higher than at 125 nM which might point to a problem with the syringe pumps of the GA-IIx.

The half-sites TGAC or TGAG of the dimer consensus motif 5'-TGA(C/G)-TCA-3' are enriched among the first ranked 8-mers and the occurrences decline among weaker binding motifs.

Overall there is an increase in Hamming distances from the 8-mer consensus ATGACTCA (TGAGTCAT) with increasing ranking depth observable.

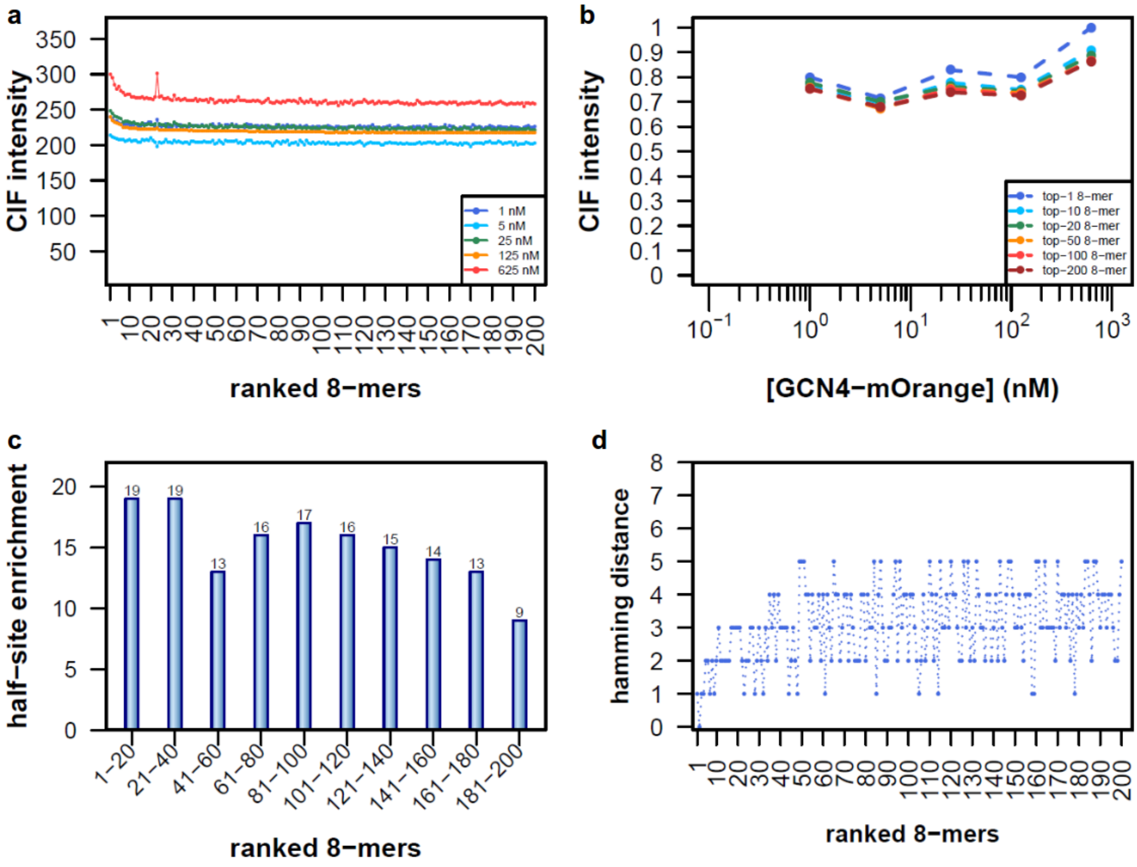


Figure 9.11: Analyses of the experiment 28.03.2014 lane 2. (a) Intensity course for the first 200 ranked 8-mers ranked at 125 nM on lane 2. (b) Binding curves for selected 8-mers ranked at 125 nM on lane 2. (c) Enrichment of half-sites (TGAC or TGAG) for the first 200 8-mers ranked at 125 nM on lane 2. (d) Hamming distance from consensus ATGACTCA for the first 200 8-mers ranked at 125 nM on lane 2.

9.15.4.3 Lane 3

Analysis results for lane 3 illustrated by Figure 9.12.
The 8-mer ranking yields TTATATAA (TTATATAA) and TAGATAAG (CTTATCTA) as the first two top placed motifs. The 8-mer consensus ATGAGTCA (TGACTCAT) does not occur among the first 200 ranked 8-mers. The intensity courses do not decline after the 30th rank and from rank 30 onwards all DNA clusters seem to be bound non-specifically.
No saturation is occurring here, GCN4 (or at least mOrange) molecules keep aggregating at the DNA clusters in an equal fashion.
The half-sites TGAC or TGAG of the dimer consensus motif 5'-TGA(C/G)-TCA-3' are not enriched among the first 40 ranked 8-mer motifs.
There is no increase in Hamming distances from the 8-mer consensus ATGACTCA (TGAGTCAT) with increasing ranking depth observable.

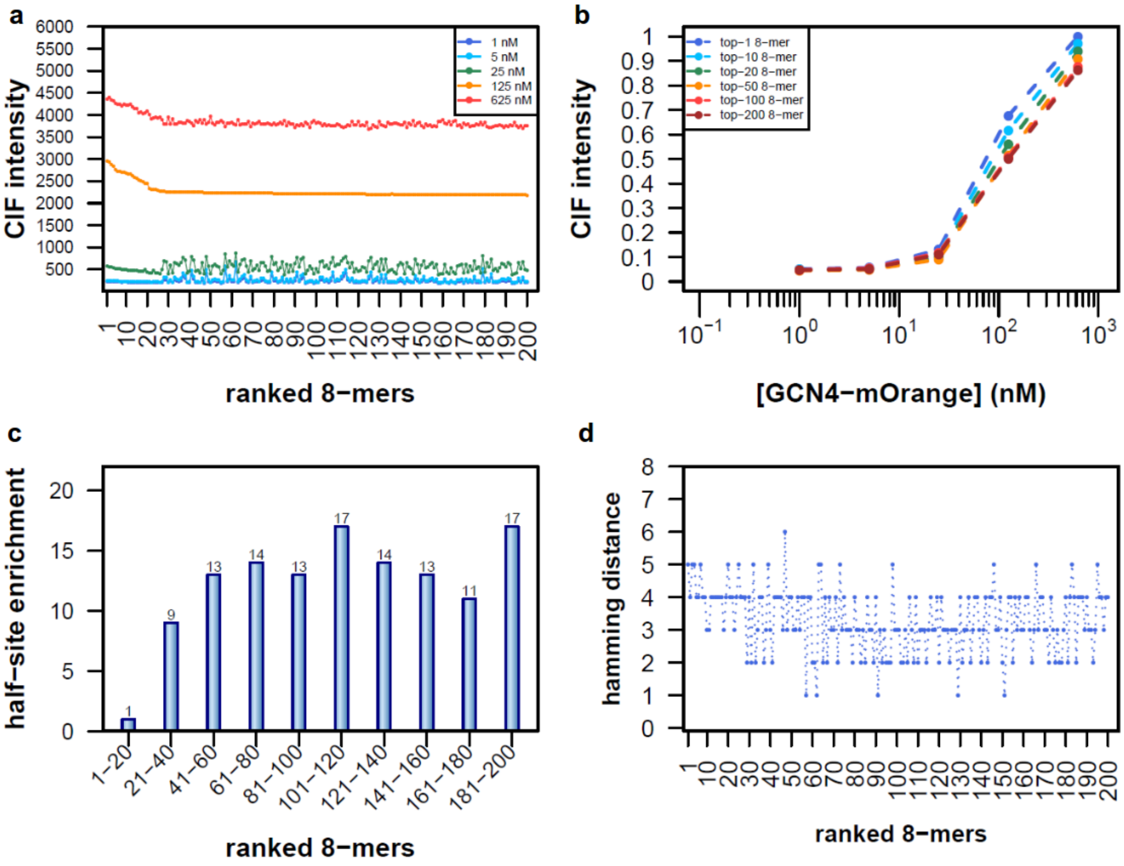


Figure 9.12: Analyses of the experiment 28.03.2014 lane 3. (a) Intensity course for the first 200 ranked 8-mers ranked at 125 nM on lane 3. (b) Binding curves for selected 8-mers ranked at 125 nM on lane 3. (c) Enrichment of half-sites (TGAC or TGAG) for the first 200 8-mers ranked at 125 nM on lane 3. (d) Hamming distance from consensus ATGACTCA for the first 200 8-mers ranked at 125 nM on lane 3.

9.16 Experiment 11.08.2014

In the following section the details regarding the experiment 11.08.2014 are described.

9.16.1 Sample

As sample GCN4 fused with mOrange as described previously.

9.16.2 Flow cell buffer

The flow cell buffer was composed of PBS + 5 mM $MgCl_2$ + 60 mM KCl + 0.3 mg/ml BSA + 0.1% Tween-20.

9.16.3 Protocol

The random DNA library N25 was used as described previously. Five concentrations were applied, i.e. 1 nM, 5 nM, 25 nM, 125 nM, 625 nM. Imaging was also performed at 0 nM. A different primer, i.e. the Illumina read 1 sequencing primer, from the previous experiments was used that enabled a more efficient resynthesis of the second DNA strand. As a control check, a primer (0.01 µM) with an Alexa-like dye (detectable in the C channel) was hybridised to the flow cell primer oligos before the resynthesis. This primer should be displaced by Klenow polymerase if the resynthesis occurs at the related DNA cluster. At cycle 52: Fluorescently labeled primer hybridisation, there should be an even signal in the C channel. At cycle 53: After Klenow reaction, there should be a weaker signal in the C channel where the DNA clusters are positioned (if dsDNA synthese has happened). Every 10 min Klenow mix was pumped into the flow cell here.
Lane 1:
Equilibration time was 1 h at 20 °C. No washing was applied during the protein cycles, the next concentration level was titrated into the flow cell. Every ten minutes protein solution was pumped into the flow cell during protein cycles 55 to 59. Protein cycles for lane 1 were: cycle 54: 0 nM, cycle 55: 1 nM, cycle 56: 5 nM, cycle 57: 25 nM, cycle 58: 125 nM, cycle 59: 625 nM.

9.16.4 Data analysis

Figure 9.13 shows the main analysis plots for the experiment 11.08.2014 lane 1.
The 8-mer ranking yields ATACACTC (GAGTGTAT) and ACACTCTT (AAGAGTGT) as the first two top placed motifs. The 8-mer consensus ATGACTCA (TGAGTCAT) occurs at rank 30. The intensities at the different concentrations are not increasing properly. The intensity courses only decline very marginally. The peak at rank 30 occurs for the 8-mer consensus ATGACTCA (TGAGTCAT), averaging over 1106 DNA

clusters, and the peak at rank 81 occurs for the 8-mer motif GTGACTCA (TGAGTCAC), averaging over 990 DNA clusters. The half-site TGAC occurs at ranks 27, 30, 81 and 132, and the GCN4 7-mer consensus motif TGACTCA occurs at ranks 30 and 81. Nearly all other ranked 8-mers contain the submotif GTGT (ACAC).
No saturation is occurring here and the concentration levels are not increasing properly. The half-sites TGAC or TGAG of the dimer consensus motif 5'-TGA(C/G)-TCA-3' are not enriched. However, the submotif GTGT (ACAC) is enriched among the entire 200 ranked 8-mer as displayed in Figure 9.14.
There is no increase in Hamming distance observable from the 8-mer consensus ATGACTCA (TGAGTCAT).

There is an decrease in cycle 53 regarding the fluorescent signal in the C-channel coming from the primer with the Alexa-like dye as shown by Figure 9.15.

An additional quality control concerning the dsDNA synthese could be applied here by using a threshold, e.g. 10% quantile, in order to exclude clusters for which the resynthesis was less efficient (Figure 9.16).

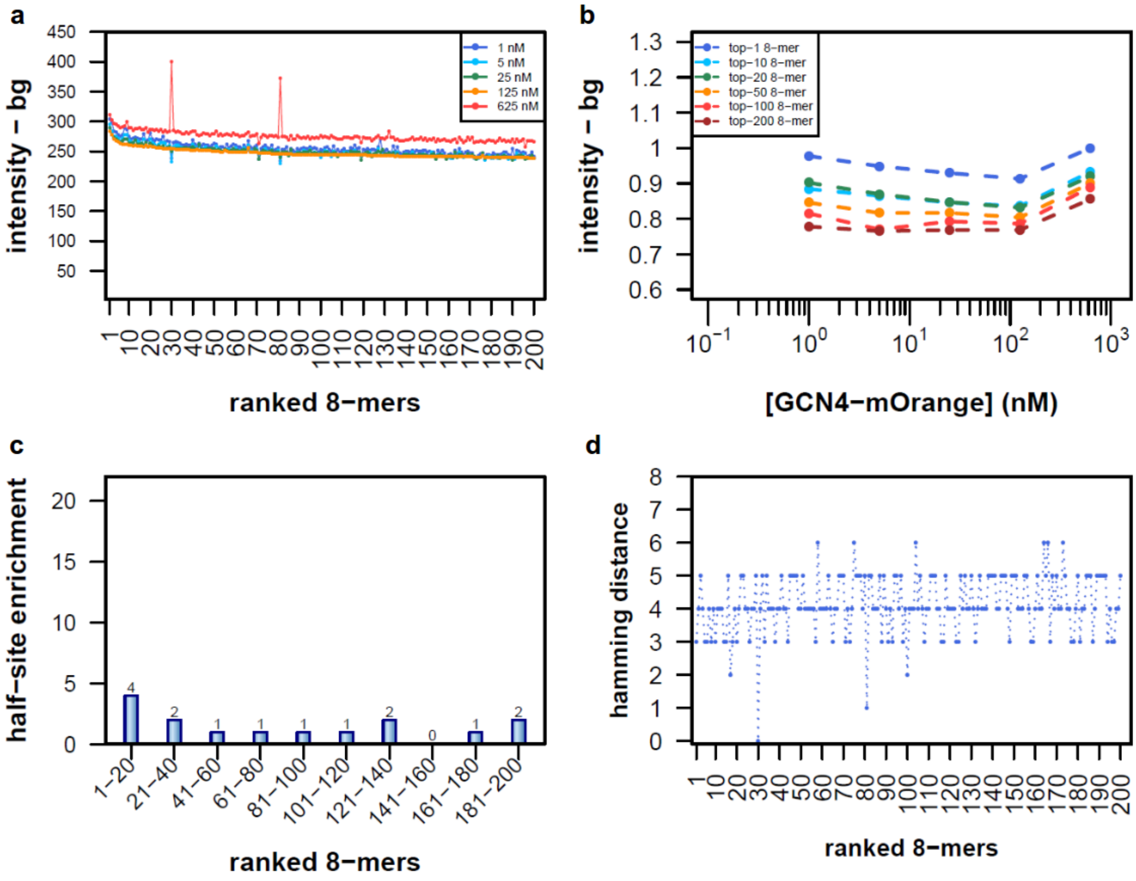


Figure 9.13: Analyses of the experiment 11.08.2014 lane 1. (a) Intensity course for the first 200 ranked 8-mers ranked at 125 nM on lane 1. (b) Binding curves for selected 8-mers ranked at 125 nM on lane 1. (c) Enrichment of half-sites (TGAC or TGAG) for the first 200 8-mers ranked at 125 nM on lane 1. (d) Hamming distance from consensus ATGACTCA for the first 200 8-mers ranked at 125 nM on lane 1.

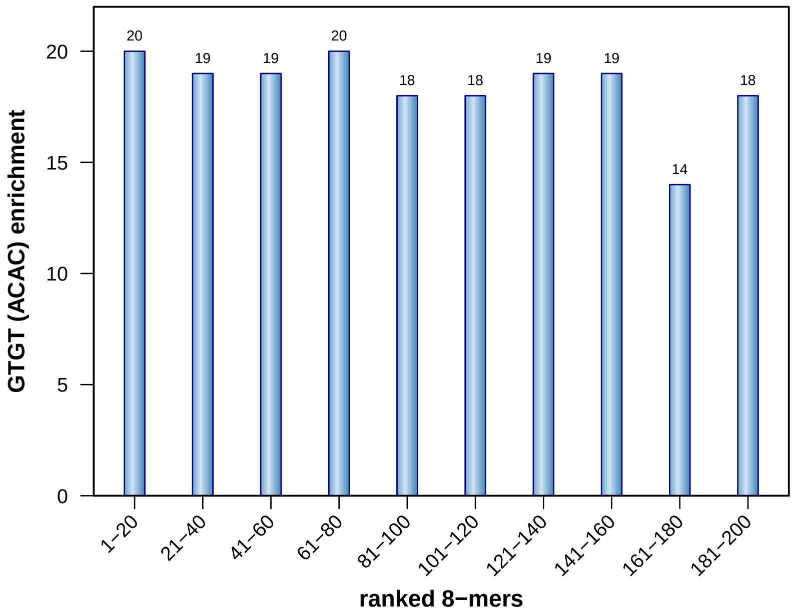


Figure 9.14: Enrichment of GTGT (ACAC) for the first 200 8-mers ranked at 125 nM.

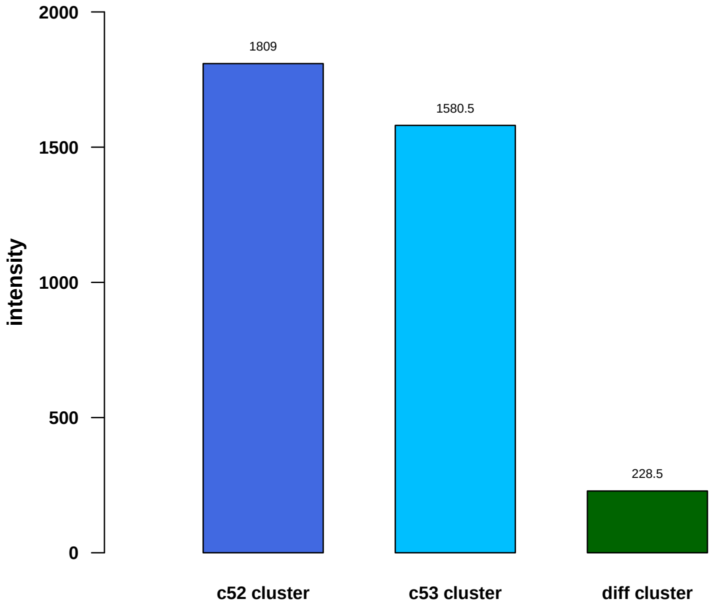


Figure 9.15: C-channel signals from cycle 52 and 53 before and after dsDNA synthesis.

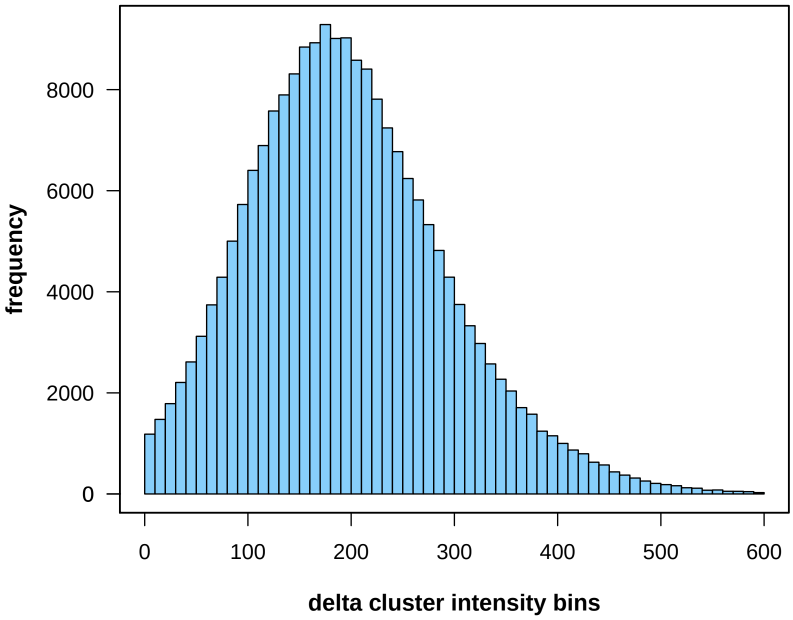


Figure 9.16: Histogram of cluster intensity differences (cycle 52 minus cycle 53).

9.17 Experiment 18.08.2014

In the following section the details regarding the experiment 18.08.2014 are described.

9.17.1 Sample

As sample GCN4 fused with mOrange was used as described previously.

9.17.2 Flow cell buffer

The flow cell buffer for the different lanes was composed as follows:
Lane 1: HiTS-FLIP buffer as described in Nutiu et al. (2011).
Lane 2: PBS + 5 mM $MgCl_2$ + 60 mM KCl + 0.3 mg/ml BSA + 0.1% Tween-20.
Lane 3: PBS + 5 mM $MgCl_2$ + 60 mM KCl + 0.3 mg/ml BSA + 0.1% Tween-20.

9.17.3 Protocol

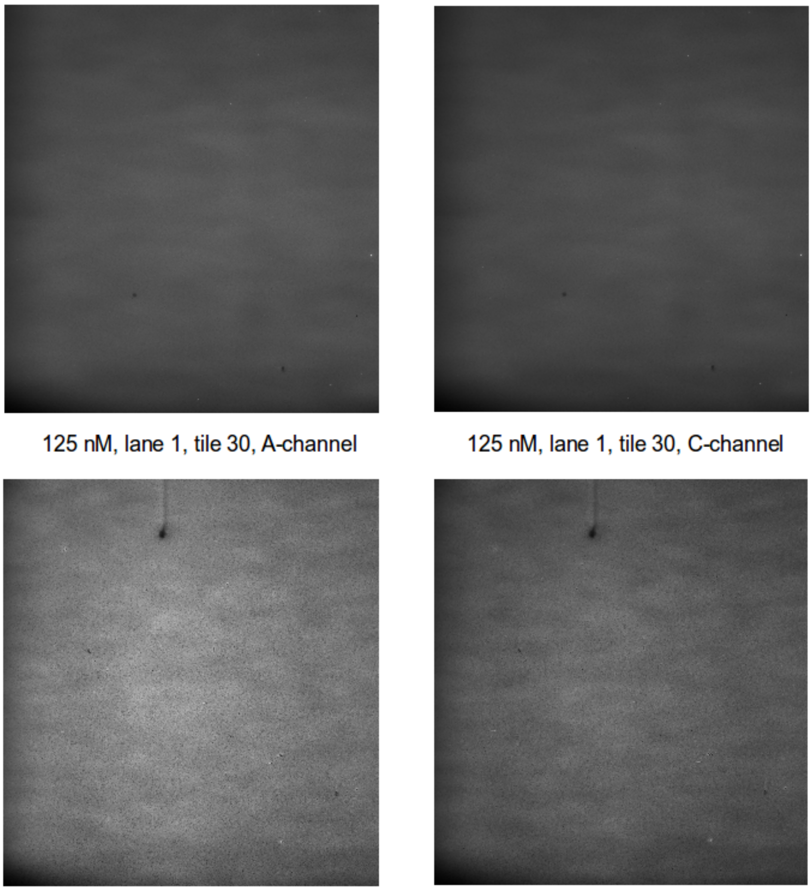
The flow cell from March 2014 (experiment 28.03.2014) was reused and no (re)sequencing was done here. Five concentrations were applied, i.e. 1 nM, 5 nM, 25 nM, 125 nM, 625 nM. Imaging was also performed at 0 nM.
Lane 2:
Equilibration time was 1 h at 20 °C. No washing was applied during the protein cycles, the next concentration level was titrated into the flow cell. Cycle 90: Fluorescently labeled primer hybridisation (0.01 µM). Cycle 91: dsDNA synthesis. The Protein cycles were: cycle 92: 0 nM, cycle 93: 1 nM, cycle 94: 5 nM, cycle 95: 25 nM, cycle 96: 125 nM, cycle 97: 625 nM.
Lane 3:
Equilibration time was 30 min at 20 °C. No washing was applied during the protein cycles, the next concentration level was titrated into the flow cell. Cycle 96: Fluorescently labeled primer hybridisation (0.01 µM). Cycle 97: dsDNA synthesis. The protein cycles were: cycle 98: 0 nM, cycle 99: 1 nM, cycle 100: 5 nM, cycle 101: 25 nM, cycle 102: 125 nM, cycle 103: 625 nM.

9.17.4 Data analysis

The following sections show the main analysis plots for the experiment 18.08.2014.

9.17.4.1 Lane 1

Regarding lane 1, there was a problem at the concentrations of 125 nM and 625 nM as highlighted by Figure 9.17 and Figure 9.18. At concentration 125 nM, brightness is uniform and clusters seem to be bound unspecifically. At concentration 625 nM, the images at the G and T channel show a high brightness across the whole tile, and no individual clusters can be identified. The brightness for the images at the A and C channel is less pronounced but still very uniform so that individual clusters are hard to identify. What could have caused this issue? It seems to be a problem with the clusters on lane 1, perhaps poor primer annealing occurred. Therefore, clusters are only bound by GCN4 unspecifically and the amount of unbound GCN4 molecules in the background is increased, especially at the highest concentration.



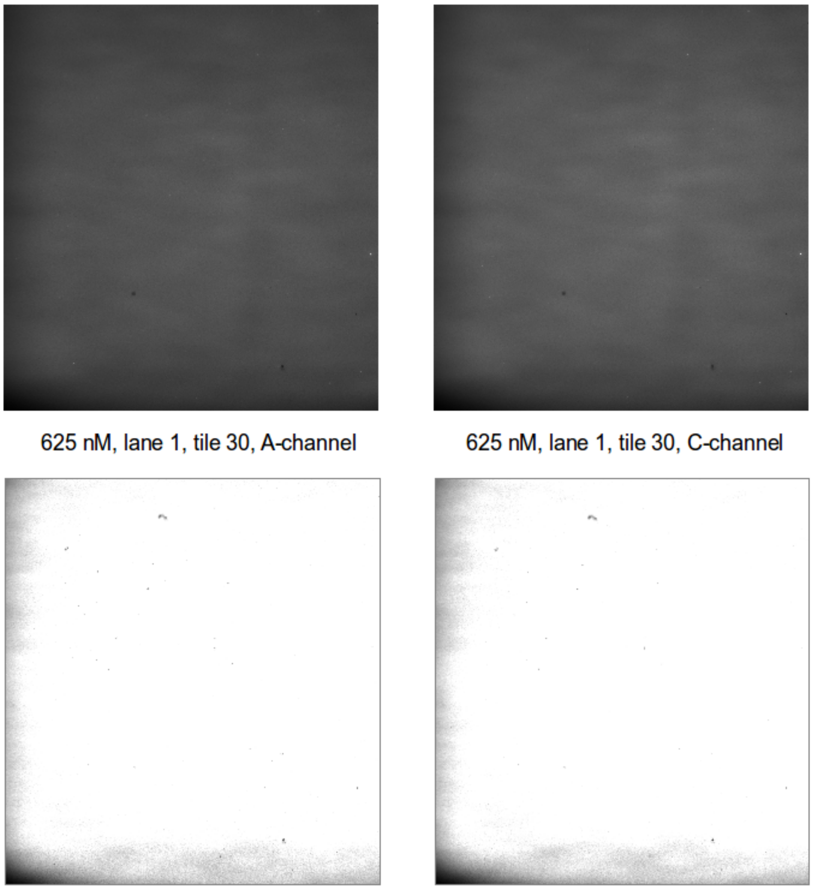
125 nM, lane 1, tile 30, A-channel

125 nM, lane 1, tile 30, C-channel

125 nM, lane 1, tile 30, G-channel

125 nM, lane 1, tile 30, T-channel

Figure 9.17: Four images from lane 1, tile 30 from the different channels at 125 nM.



625 nM, lane 1, tile 30, A-channel

625 nM, lane 1, tile 30, C-channel

625 nM, lane 1, tile 30, G-channel

625 nM, lane 1, tile 30, T-channel

Figure 9.18: Four images from lane 1, tile 30 from the four different channels at 625 nM.

9.17.4.2 Lane 2

Figure 9.19 shows the main analysis plots for the experiment 18.08.2014 lane 2.
The 8-mer ranking yields ATGACTCA (TGAGTCAT) and ATGAGTCA (TGACTCAT) as the first two top placed motifs. The preeminent, antagonistic peaks occur at the following ranks:
rank 110: ATACACTC (GAGTGTAT)
rank 162: AGAGTGTG (CACACTCT)
rank 175: ACACTCTT (AAGAGTGT)
rank 188: CACACTCA (TGAGTGTG)
Those peaks occur at DNA clusters containing the motif (GA)GTGT, however only where a half-site of the dimeric consensus motif, TGA(C/G), is not involved. Concentration at 1 nM is rather 10 nM (problem with the fluidics).
Saturation is occurring here. Concentration at 1 nM is rather 10 nM (problem with the fluidics).
The half-sites TGAC or TGAG of the dimer consensus motif 5'-TGA(C/G)-TCA-3' are enriched.
There is an increase in Hamming distance from the 8-mer consensus ATGACTCA (TGAGTCAT) observable with increasing ranking depth.

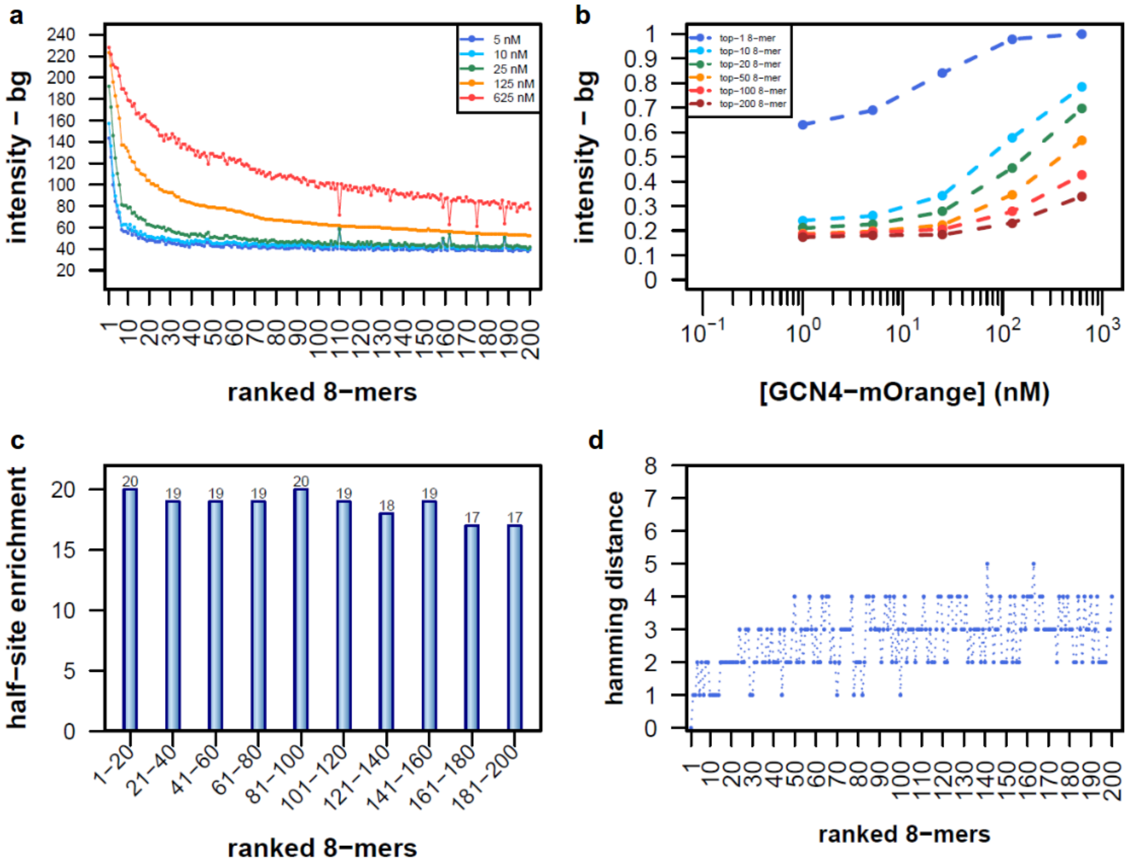


Figure 9.19: Analyses of the experiment 18.08.2014 lane 2. (a) Intensity course for the first 200 ranked 8-mers ranked at 125 nM on lane 2. (b) Binding curves for selected 8-mers ranked at 125 nM on lane 2. (c) Enrichment of half-sites (TGAC or TGAG) for the first 200 8-mers ranked at 125 nM on lane 2. (d) Hamming distance from consensus ATGACTCA for the first 200 8-mers ranked at 125 nM on lane 2.

9.17.4.3 Lane 3

Figure 9.20 shows the main analysis plots for the experiment 18.08.2014 lane 3.
The 8-mer ranking yields ATGACTCA (TGAGTCAT) and ATGAGTCA (TGACTCAT) as the first two top placed motifs. The first four antagonistic peaks occur at the following ranks:
rank 28: AGAGTGTT (AACACTCT)
rank 33: ATACACTC (GAGTGTAT)
rank 56: ACACTCTT (AAGAGTGT)
rank 57: CGAGTGTT (AACACTCG)
Again, as for lane 2, those peaks occur at DNA clusters containing the motif (GA)GTGT, however only where a half-site of the dimeric consensus motif, TGA(C/G), is not involved. Concentration at 1 nM is rather 10 nM (problem with the fluidics).
No saturation is occurring here, GCN4 (or at least mOrange) molecules keep aggregating at the DNA clusters in an equal fashion.
The half-sites TGAC or TGAG of the dimer consensus motif 5'-TGA(C/G)-TCA-3' are enriched among the first 40 ranked 8-mer motifs but not as strongly as on lane 2, and (GA)GTGT containing 8-mer motifs are more dominant.
There is an increase in Hamming distances from the 8-mer consensus ATGACTCA (TGAGTCAT) observable for increasing ranking depth.

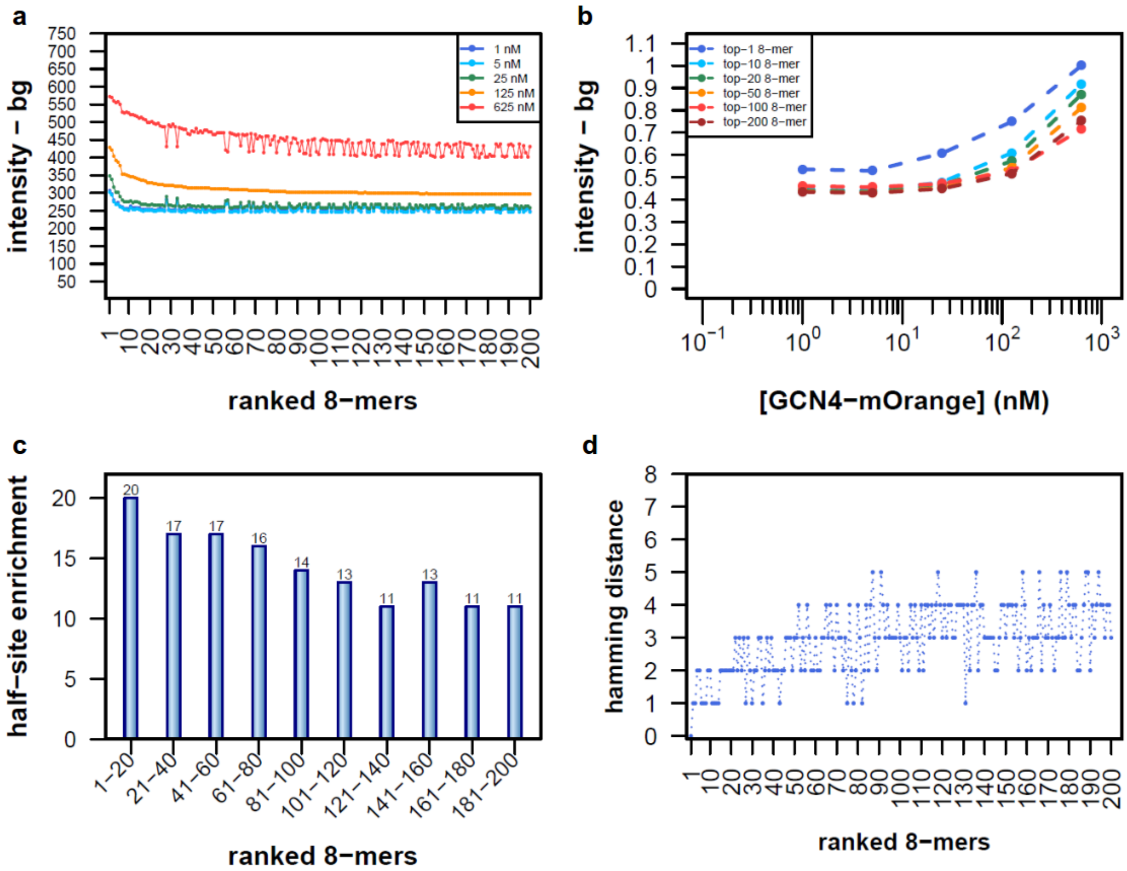


Figure 9.20: Analyses of the experiment 18.08.2014 lane 3. (a) Intensity course for the first 200 ranked 8-mers ranked at 125 nM on lane 3. (b) Binding curves for selected 8-mers ranked at 125 nM on lane 3. (c) Enrichment of half-sites (TGAC or TGAG) for the first 200 8-mers ranked at 125 nM on lane 3. (d) Hamming distance from consensus ATGACTCA for the first 200 8-mers ranked at 125 nM on lane 3.

9.18 Experiment 12.02.2015

In the following section the details regarding the experiment 12.02.2015 are described.

9.18.1 Sample

As sample GCN4 fused with mOrange was used as described previously.

9.18.2 Flow cell buffer

The flow cell buffer was composed of PBS + 5 mM $MgCl_2$ + 60 mM KCl + 0.3 mg/ml BSA + 0.1% Tween-20.

9.18.3 Protocol

The random DNA library N25 was used as described previously. Ten concentrations were applied, i.e. 0.1 nM, 0.3 nM, 0.8 nM, 2 nM, 6 nM, 17 nM, 50 nM, 135 nM, 375 nM, 1000 nM. Imaging was also performed at 0 nM. Lane 2 and 3 (replicate of lane 2) were used. Equilibration time was 30 min at 20 °C. No washing was applied during the protein cycles, the next concentration level was titrated continuously into the flow cell.
Lane 2:
Cycle 31: Fluorescently labeled primer hybridisation (0.1 µM).
Cycle 32: dsDNA synthesis.
The protein cycles were: cycle 33: 0 nM, cycle 34: 0.1 nM, cycle 35: 0.3 nM, cycle 36: 0.9 nM, cycle 37: 2 nM, cycle 38: 6 nM, cycle 39: 17 nM, cycle 40: 50 nM, cycle 41: 135 nM, cycle 42: 375 nM, cycle 43: 1000 nM.

9.18.4 Data analysis

Figure 9.21 shows the main analysis plots for the experiment 12.02.2015 lane 2.
The 8-mer ranking yields ATGACTCA (TGAGTCAT) and ATGAGTCA (TGACTCAT) as the first two top placed motifs. There is a steep increase visible in fluorescent intensity for the top 8-mer motif ATGACTCA (TGAGTCAT) at the concentrations 375 nM and 1000 nM. The motifs (GA)GTGT are only bound unspecifically beyond the 200th rank. There is no saturation occurring here. During the run there has been a problem with the fluidics so that the accurate amount of GCN4 was not pumped into the flow cell (135 nM is lower than 50 nM).
The half-sites TGAC or TGAG of the dimer consensus motif 5'-TGA(C/G)-TCA-3' are enriched.
There is an increase in Hamming distance observable from the 8-mer consensus ATGACTCA (TGAGTCAT).

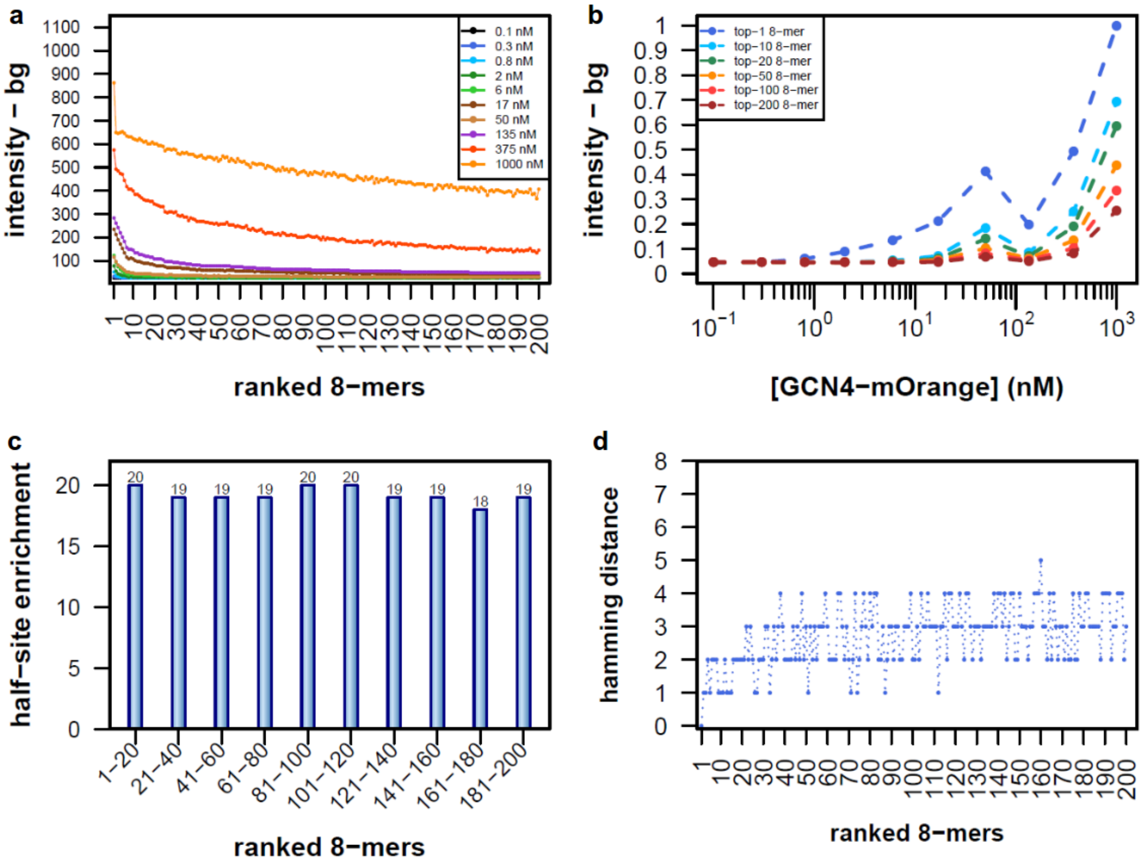


Figure 9.21: Analyses of the experiment 12.02.2015 lane 2. (a) Intensity course for the first 200 ranked 8-mers ranked at 125 nM on lane 2. (b) Binding curves for selected 8-mers ranked at 125 nM on lane 2. (c) Enrichment of half-sites (TGAC or TGAG) for the first 200 8-mers ranked at 125 nM on lane 2. (d) Hamming distance from consensus ATGACTCA for the first 200 8-mers ranked at 125 nM on lane 2.

There is an decrease in cycle 32 regarding the fluorescent signal in the C-channel coming from the primer with the Alexa-like dye as shown in Figure 9.22.
An additional quality control concerning the dsDNA synthese could be applied here by using a threshold, e.g. 10% quantile, in order to exclude clusters for which the resynthesis was less efficient (Figure 9.23).

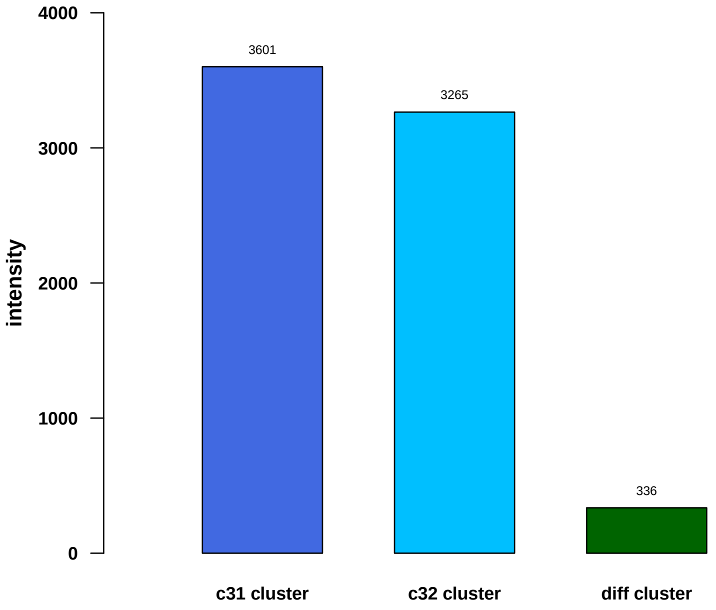


Figure 9.22: C-channel signals from cycle 31 and 32 before and after dsDNA synthesis shown by bar plot of experiment 12.02.2015.

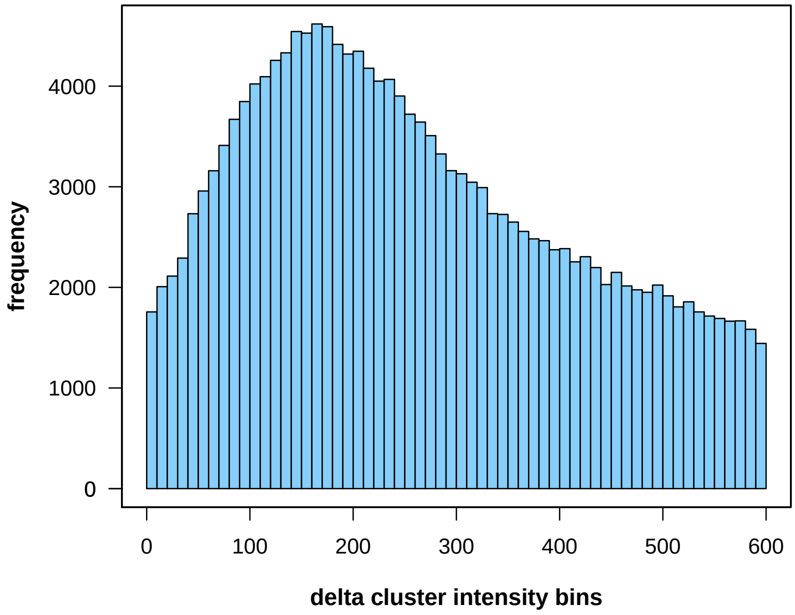


Figure 9.23: Histogram of cluster intensity differences (cycle 31 minus cycle 32) of experiment 12.02.2015.

9.19 Experiment 06.03.2015

In the following section the details regarding the experiment 06.03.2015 are described.

9.19.1 Sample

As sample GCN4 fused with mOrange was used as described previously.

9.19.2 Flow cell buffer

The flow cell buffer was composed of PBS + 5 mM $MgCl_2$ + 60 mM KCl + 0.3 mg/ml BSA + 0.1% Tween-20.

9.19.3 Protocol

The flow cell from February 2015 (experiment 12.02.2015) was reused and no (re)sequencing was done here. Ten concentrations were applied, i.e. 0.1 nM, 0.3 nM, 0.8 nM, 2 nM, 6 nM, 17 nM, 50 nM, 135 nM, 375 nM, 1000 nM. Imaging was also performed at 0 nM. After the first five concentration steps, denaturation and resynthesis of the second DNA strand was performed. Equilibration time was 30 min at 20 °C. No washing was applied during the protein cycles, the next concentration level was titrated into the flow cell.
Lane 2:
Cycle 44: Fluorescently labeled primer hybridisation (0.1 µM). Cycle 45: dsDNA synthesis. The protein cycles were: cycle 46: 0 nM, cycle 47: 0.1 nM, cycle 48: 0.3 nM, cycle 49: 0.9 nM, cycle 50: 2 nM, cycle 51: 6 nM, cycle 52: denaturation of second DNA strand, cycle 53: resynthesis of second DNA strand, cycle 54: 17 nM, cycle 55: 50 nM, cycle 56: 135 nM, cycle 57: 375 nM, cycle 58: 1000 nM.

9.19.4 Data analysis

Figure 9.24 shows the main analysis plots for the experiment 06.03.2015 lane 2.
The 8-mer ranking yields ATGACTCA (TGAGTCAT) and ATGAGTCA (TGACTCAT) as the first two top placed motifs. There is a steep increase visible in fluorescent intensity for the top 8-mer motif ATGACTCA (TGAGTCAT) at the concentrations 375 nM and 1000 nM. The motifs (GA)GTGT are only bound unspecifically beyond the 200th rank. There is no saturation occurring here.
The half-sites TGAC or TGAG of the dimer consensus motif 5'-TGA(C/G)-TCA-3' are enriched.
There is an increase in Hamming distance observable from the 8-mer consensus ATGACTCA (TGAGTCAT).

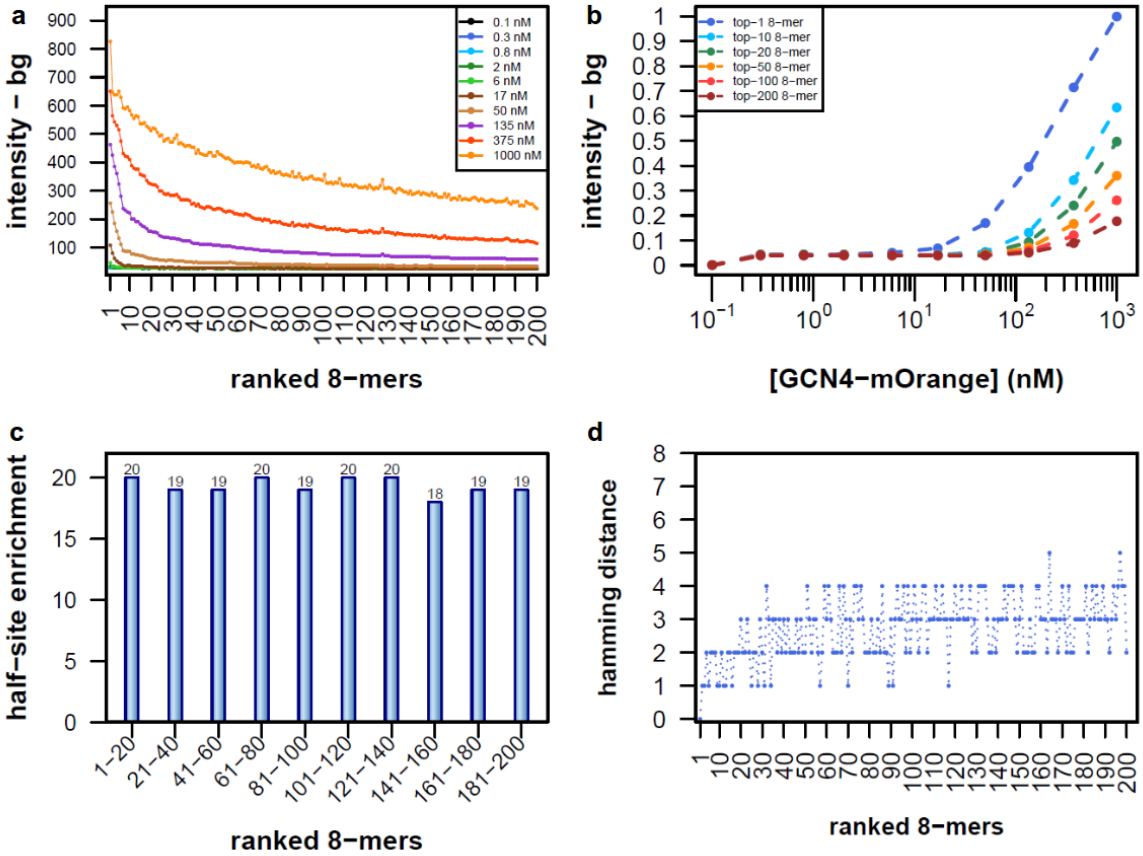


Figure 9.24: Analyses of the experiment 06.03.2015 lane 2. (a) Intensity course for the first 200 ranked 8-mers ranked at 125 nM on lane 2. (b) Binding curves for selected 8-mers ranked at 125 nM on lane 2. (c) Enrichment of half-sites (TGAC or TGAG) for the first 200 8-mers ranked at 125 nM on lane 2. (d) Hamming distance from consensus ATGACTCA for the first 200 8-mers ranked at 125 nM on lane 2.

9.20 Experiment 14.04.2015

In the following section the details regarding the experiment 14.04.2015 are described.

9.20.1 Sample

As sample GCN4 fused with mOrange was used as described previously.

9.20.2 Flow cell buffer

The flow cell buffer was composed of PBS + 5 mM $MgCl_2$ + 60 mM KCl + 0.3 mg/ml BSA + 0.1% Tween-20.

9.20.3 Protocol

The flow cell from February 2015 (experiment 12.02.2015) was reused and no (re)sequencing was done here. Ten concentrations were applied, i.e. 0.1 nM, 0.3 nM, 0.8 nM, 2 nM, 6 nM, 17 nM, 50 nM, 135 nM, 375 nM, 1000 nM. Imaging was also performed at 0 nM. After the first five concentration steps, denaturation and resynthesis of the second DNA strand was performed. Lane 2 and 3 (replicate of lane 2) was used. Equilibration time was 30 min at 20 °C. No washing was applied during the protein cycles, the next concentration level was titrated into the flow cell.
Lane 2:
Cycle 59: Fluorescently labeled primer hybridisation (0.1 µM). Cycle 60: dsDNA synthesis. The protein cycles were: cycle 61: 0 nM, cycle 62: 0.1 nM, cycle 63: 0.3 nM, cycle 64: 0.9 nM, cycle 65: 2 nM, cycle 66: 6 nM, cycle 67: denaturation of second DNA strand, Cycle 68: resynthesis of second DNA strand, cycle 69: 17 nM, cycle 70: 50 nM, cycle 71: 135 nM, cycle 72: 375 nM, cycle 73: 1000 nM

9.20.4 Data analysis

Figure 9.25 shows the main analysis plots for the experiment 14.04.2015 lane 2.
The 8-mer ranking yields ATGACTCA (TGAGTCAT) and ATGAGTCA (TGACTCAT) as the first two top placed motifs. There is again a steep increase visible in fluorescent intensity for the top 8-mer motif ATGACTCA (TGAGTCAT) at the concentrations 375 nM and 1000 nM. In addition, there are spikes occurring at the highest concentrations 375 nM and 1000 nM:
rank 7: TATGAA
rank 9: TATGAA
rank 11: TGTGAA

rank 13: TATGAA

rank 14: TGTGAA

rank 16: TGTGAA

The motifs (GA)GTGT are only bound unspecifically.
There is no saturation occurring here.
The half-sites TGAC or TGAG of the dimer consensus motif 5'-TGA(C/G)-TCA-3' are enriched but to a lesser extent since there is the new motif T(A/G)TGAA ranked here.
There is an increase in Hamming distance observable from the 8-mer consensus ATGACTCA (TGAGTCAT).

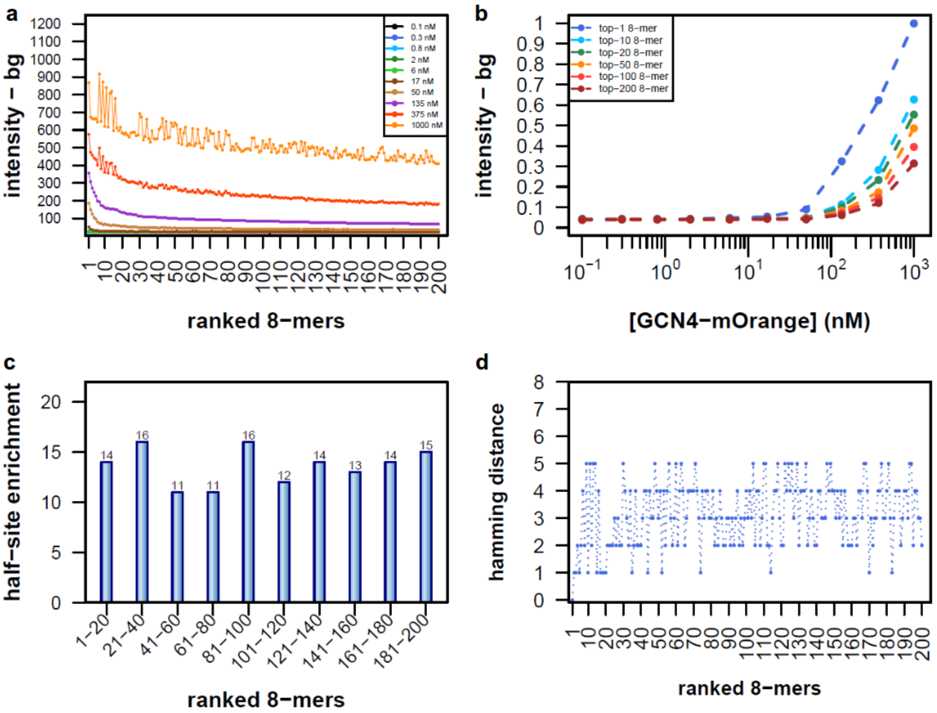


Figure 9.25: Analyses of the experiment 14.04.2015 lane 2. (a) Intensity course for the first 200 ranked 8-mers ranked at 125 nM on lane 2. (b) Binding curves for selected 8-mers ranked at 125 nM on lane 2. (c) Enrichment of half-sites (TGAC or TGAG) for the first 200 8-mers ranked at 125 nM on lane 2. (d) Hamming distance from consensus ATGACTCA for the first 200 8-mers ranked at 125 nM on lane 2.

Bibliography

N. Abe, I. Dror, R. Rohs, R. S. Mann, N. Abe, I. Dror, L. Yang, M. Slattery, T. Zhou, H. J. Bussemaker, and R. Rohs. Deconvolving the Recognition of DNA Shape from Sequence. *Cell* **2015**;*161*(2):1–12.

M. D. Abràmoff, P. J. Magalhães, and S. J. Ram. Image processing with ImageJ. *Biophotonics international* **2004**;*11*(7):36–42.

A. Afek and D. B. Lukatsky. Positive and negative design for nonconsensus protein-DNA binding affinity in the vicinity of functional binding sites. *Biophysical Journal* **2013**;*105*(7):1653–1660.

P. Agius, A. Arvey, W. Chang, W. S. Noble, and C. Leslie. High Resolution Models of Transcription Factor-DNA Affinities Improve In Vitro and In Vivo Binding Predictions. *PLoS Computational Biology* **2010**;*6*(9):e1000916.

P. Agre, P. F. Johnson, and S. L. McKnight. Cognate DNA binding specificity retained after leucine zipper exchange between GCN4 and C/EBP. *Science* **1989**;*246*(4932):922–926.

B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. Molecular Biology of the Cell (5th edit.) Garland Science. *New York* **2007**;.

R. A. Alberty and G. G. Hammes. Application of the theory of diffusion-controlled reactions to enzyme kinetics. *The Journal of Physical Chemistry* **1958**;*62*(2):154–159.

S. Andrews. FASTQC. `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`, **2010–2015**a.

S. Andrews. FASTQC Manual. `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`, **2010–2015**b.

K. Arndt and G. R. Fink. GCN4 protein, a positive transcription factor in yeast, binds general control promoters at all 5'TGACTC 3'sequences. *Proceedings of the National Academy of Sciences* **1986**;*83*(22):8516–8520.

E. Asuka, O. L. Garrett, W. Fang, S. E. Graham, and Z. A. Aseem. Controlling gene networks and cell fate with precision-targeted DNA-binding proteins and small-molecule-based genome readers. *Biochemical Journal* **2014**;*462*(3):397–413.

G. Babaloukas, N. Tentolouris, S. Liatis, A. Sklavounou, and D. Perrea. Evaluation of three methods for retrospective correction of vignetting on medical microscopy images utilizing two open source software tools. *Journal of microscopy* **2011**;*244*(3):320–4.

G. Badis, M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, et al. Diversity and complexity in DNA recognition by transcription factors. *Science* **2009**;*324*(5935):1720–1723.

D. Bates and J. Chambers. Statistical models in S, chapter 10 (Nonlinear models). **1992**.

D. M. Bates and D. G. Watts. Nonlinear regression analysis and its applications. 1988. *John Wiles & Sons, Inc* **1988**;.

A. Beckers, W. Bruijn, E. Gelsema, M. CLETOM-SOETEMAN, and H. Eijk. Quantitative electron spectroscopic imaging in bio-medicine: Methods for image acquisition, correction and analysis. *Journal of Microscopy* **1994**;*174*(3):171–182.

D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. a. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. J. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. D. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. a. Baybayan, V. a. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. a. Bridgham, R. C. Brown, A. a. Brown, D. H. Buermann, A. a. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, M. Chiara E Catenazzi, S. Chang, R. Neil Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. Fuentes Fajardo, W. Scott Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. a. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. a. Huw Jones, G.-D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. a. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. a. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. Ling Ng, S. M. Novo, M. J. O'Neill, M. a. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. Chris Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. Chiva Rodriguez, P. M. Roe, J. Rogers, M. C. Rogert Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. a. Smith, J. Ernest Sohna Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. Vandevondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, and A. J. Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **2008**;*456*(7218):53–59.

O. G. Berg, R. B. Winter, and P. H. von Hippel. Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry* **1981**;*20*:6929–6948.

C. Berger, L. Piubelli, U. Haditsch, and H. Rudolf Bosshard. Diffusion-controlled DNA recognition by an unfolded, monomeric bZIP transcription factor. *FEBS Letters* **1998**;*425*:14–18.

M. F. Berger, A. a. Philippakis, A. M. Qureshi, F. S. He, P. W. Estep, and M. L. Bulyk. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature biotechnology* **2006**;*24*(11):1429–1435.

M. J. Black, G. Sapiro, D. H. Marimont, and D. Heeger. Robust anisotropic diffusion. *Image Processing, IEEE Transactions on* **1998**;*7*(3):421–432.

C. G. Broyden. The convergence of a class of double-rank minimization algorithms 2. The new algorithm. *IMA Journal of Applied Mathematics* **1970**;*6*(3):222–231.

P. S. Brzovic, C. C. Heikaus, L. Kisselev, R. Vernon, E. Herbig, D. Pacheco, L. Warfield, P. Littlefield, D. Baker, R. E. Klevit, and S. Hahn. The Acidic Transcription Activator Gcn4 Binds the Mediator Subunit Gal11/Med15 Using a Simple Protein Interface Forming a Fuzzy Complex. *Molecular Cell* **2011**;*44*(6):942–953.

J. D. Buenrostro, C. L. Araya, L. M. Chircus, C. J. Layton, H. Y. Chang, M. P. Snyder, and W. J. Greenleaf. Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. *Nature biotechnology* **2014**;*32*(6):562–568.

M. L. Bulyk, E. Gentalen, D. J. Lockhart, and G. M. Church. Quantifying DNA-protein interactions by double-stranded DNA arrays. *Nature biotechnology* **1999**;*17*(6):573–577.

M. L. Bulyk, P. L. F. Johnson, and G. M. Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic acids research* **2002**;*30*(5):1255–1261.

W. Burger and M. J. Burge. Digital image processing: an algorithmic introduction using Java. Springer Science & Business Media, **2009**a.

W. Burger and M. J. Burge. Principles of Digital Image Processing: Core Algorithms. Springer, **2009**b.

W. Burger, M. J. Burge, M. J. Burge, and M. J. Burge. Principles of Digital Image Processing. Springer, **2009**.

R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* **1995**;*16*(5):1190–1208.

I. S. Chan, A. V. Fedorova, and J. a. Shin. The GCN4 bZIP targets noncognate gene regulatory sequences: Quantitative investigation of binding at full and half sites. *Biochemistry* **2007**; *46*:1663–1671.

F. Chang, C.-J. Chen, and C.-J. Lu. A linear-time component-labeling algorithm using contour tracing technique. *computer vision and image understanding* **2004**;*93*(2):206–220.

R. Chapman, C. Sidrauski, and P. Walter. Intracellular signaling from the endoplasmic reticulum to the nucleus. *Annual review of cell and developmental biology* **1998**;*14*:459–485.

T. Chen. A practical guide to assay development and high-throughput screening in drug discovery. CRC Press, **2009**.

J. Cooley and J. Tukey. An Algorithm for the Machine Calculation of Complex Fourier Series. **1965**a.

J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Mathematics of computation* **1965**b;*19*(90):297–301.

D. Cotnoir-White, D. Laperrière, and S. Mader. Evolution of the repertoire of nuclear receptor binding sites in genomes. *Molecular and Cellular Endocrinology* **2011**;*334*(1-2):76–82.

S. Cranz, C. Berger, A. Baici, I. Jelesarov, and H. R. Bosshard. Monomeric and Dimeric bZIP Transcription Factor GCN4 Bind at the Same Rate to Their Target DNA Site. *Biochemistry* **2004**;*43*:718–727.

J. Crocker, N. Abe, L. Rinaldi, A. P. McGregor, N. Frankel, S. Wang, A. Alsawadi, P. Valenti, S. Plaza, F. Payre, R. S. Mann, and D. L. Stern. Low Affinity Binding Site Clusters Confer Hox Specificity and Regulatory Robustness. *Cell* **2015**;*160*(1-2):191–203.

L. Dagum and R. Enon. OpenMP: an industry standard API for shared-memory programming. *Computational Science & Engineering, IEEE* **1998**;*5*(1):46–55.

K. Delac, M. Grgic, and T. Kos. Sub-image homomorphic filtering technique for improving facial identification under difficult illumination conditions. In International Conference on Systems, Signals and Image Processing, volume 1. **2006**; pages 21–23.

C. Demant, C. Demant, and B. Streicher-Abel. Industrial image processing: Visual Quality Control in Manufacturing. Springer, **2013**.

S. Derrode and F. Ghorbel. Robust and Efficient Fourier–Mellin Transform Approximations for Gray-Level Image Reconstruction and Complete Invariant Description. *Computer Vision and Image Understanding* **2001**;*83*(1):57–78.

E. R. Dougherty, R. A. Lotufo, and T. I. S. for Optical Engineering SPIE. Hands-on morphological image processing, volume 71. SPIE press Bellingham, **2003**.

M. Eigen and G. G. Hammes. Elementary steps in enzyme reactions (as studied by relaxation spectrometry). *Advances in Enzymology and Related Areas of Molecular Biology, Volume 25* **2006**;pages 1–38.

T. E. Ellenberger, C. J. Brandl, K. Struhl, and S. C. Harrison. The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted α helices: crystal structure of the protein-DNA complex. *Cell* **1992**a;*71*(7):1223–1237.

T. E. Ellenberger, C. J. Brandl, K. Struhl, and S. C. Harrison. The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted ?? helices: Crystal structure of the protein-DNA complex. *Cell* **1992**b;*71*:1223–1237.

H. Etemadnia and M. R. Asharif. Homomorphic filtering approach using HSV color space in automatic image shadow segmentation. *WSEAS Trans Systems* **2004**;*3*(3):1150–1154.

G. D. Evangelidis and E. Z. Psarakis. Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2008**;*30*(10):1858–1865.

B. Ewing and P. Green. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome research* **1998**;*8*(3):186–194.

M. Fedurco, A. Romieu, S. Williams, I. Lawrence, and G. Turcatti. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Research* **2006**;*34*(3).

R. Fletcher. A new approach to variable metric algorithms. *The computer journal* **1970**; *13*(3):317–322.

B. C. Foat, A. V. Morozov, and H. J. Bussemaker. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* **2006**;*22*(14):e141–e149.

P. M. Fordyce, D. Gerber, D. Tran, J. Zheng, H. Li, J. L. DeRisi, and S. R. Quake. De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. *Nature biotechnology* **2010**;*28*(9):970–975.

P. M. Fordyce, D. Pincus, P. Kimmig, C. S. Nelson, H. El-Samad, P. Walter, and J. L. DeRisi. PNAS Plus: Basic leucine zipper transcription factor Hac1 binds DNA in two distinct modes as revealed by microfluidic analyses. *Proceedings of the National Academy of Sciences* **2012**a; *109*(45).

P. M. Fordyce, D. Pincus, P. Kimmig, C. S. Nelson, H. El-Samad, P. Walter, and J. L. DeRisi. PNAS Plus: Basic leucine zipper transcription factor Hac1 binds DNA in two distinct modes as revealed by microfluidic analyses. *Proceedings of the National Academy of Sciences* **2012**b; *109*(45).

J. Fox and S. Weisberg. An R companion to applied regression. Sage, **2010**.

J. M. Franco-Zorrilla, I. Lopez-Vidriero, J. L. Carrasco, M. Godoy, P. Vera, and R. Solano. DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proceedings of the National Academy of Sciences* **2014**;*111*(6):2367–2372.

M. R. Gartenberg, C. Ampe, T. A. Steitz, and D. M. Crothers. Molecular characterization of the GCN4-DNA complex. *Proceedings of the National Academy of Sciences* **1990**;*87*(16):6034–6038.

C. W. Garvie, J. Hagman, and C. Wolberger. Structural studies of Ets-1/Pax5 complex formation on DNA. *Molecular Cell* **2001**;*8*:1267–1276.

D. Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of computation* **1970**;*24*(109):23–26.

R. Gordân, K. F. Murphy, R. P. McCord, C. Zhu, A. Vedenko, and M. L. Bulyk. Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights. *Genome Biology* **2011**;*12*(12):R125.

R. Gordân, N. Shen, I. Dror, T. Zhou, J. Horton, R. Rohs, and M. L. Bulyk. Genomic Regions Flanking E-Box Binding Sites Influence DNA Binding Specificity of bHLH Transcription Factors through DNA Shape. *Cell Reports* **2013**;*3*(4):1093–1104.

A. A. Goshtasby. Image registration: Principles, tools and methods. Springer Science & Business Media, **2012**.

J. Gosling. The Java language specification. Addison-Wesley Professional, **2000**.

M. T. Gravina, J. H. Lin, and S. S. Levine. Lane-by-lane sequencing using Illumina's Genome Analyzer II. *BioTechniques* **2013**;*54*(5):265–269.

S. Hahn and E. T. Young. Transcriptional regulation in saccharomyces cerevisiae: Transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. *Genetics* **2011**;*189*(3):705–736.

T. Hakoshima. Leucine Zippers. *eLS* **2005**;.

S. M. Hama and M. S. Al-Ani. Medical Image Enhancement Based on an Efficient Approach for Adaptive Anisotropic Diffusion. *International Journal of Advances in Engineering & Technology* **2013**;*6*(3):1424–1430.

R. W. Hamming. Error detecting and error correcting codes. *Bell System technical journal* **1950**; *29*(2):147–160.

C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J.-B. Tagne, D. B. Reynolds, J. Yoo, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature* **2004**;*431*(7004):99–104.

P. B. Harbury, T. Zhang, P. S. Kim, and T. Alber. A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science* **1993**;*262*(5138):1401–1407.

W. Herr and M. A. Cleary. The POU domain : versatility in transcriptional regulation by a flexible two-in-one DNA-binding domain. *Genes & development* **1995**;*9*:1679–1693.

A. V. Hill. The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *J Physiol (Lond)* **1910**;*40*:4–7.

D. E. Hill, I. A. Hope, J. P. Macke, and K. Struhl. Saturation mutagenesis of the yeast his3 regulatory site: requirements for transcriptional induction and for binding by GCN4 activator protein. *Science* **1986**;*234*(4775):451–457.

A. G. Hinnebusch and K. Natarajan. Gcn4p , a Master Regulator of Gene Expression , Is Controlled at Multiple Levels by Diverse Signals of Starvation and Stress Gcn4p , a Master Regulator of Gene Expression , Is Controlled at Multiple Levels by Diverse Signals of Starvation and Stress **2002**;*1*(1):22–32.

J. J. Hollenbeck, D. L. McClain, and M. G. Oakley. The role of helix stabilizing residues in GCN4 basic region folding and DNA binding. *Protein science : a publication of the Protein Society* **2002**;*11*:2740–2747.

J. J. Hollenbeck and M. G. Oakley. GCN4 binds with high affinity to DNA sequences containing a single consensus half-site. *Biochemistry* **2000**;*39*:6380–6389.

B. Honig and Z. Shakked. structures with Hoogsteen base pairs **2012**;*17*(4):423–429.

I. A. Hope and K. Struhl. Functional dissection of a eukaryotic transcriptional activator protein, GCN4 of Yeast. *Cell* **1986**;*46*(6):885–894.

I. a. Hope and K. Struhl. GCN4, a eukaryotic transcriptional activator protein, binds as a dimer to target DNA. *The EMBO journal* **1987**;*6*(9):2781–2784.

Z. Hou, S. Huang, Q. Hu, and W. L. Nowinski. A fast and automatic method to correct intensity inhomogeneity in MR brain images. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006, pages 324–331. Springer, **2006**;.

J. Hu, Y. Lei, W.-K. Wong, S. Liu, K.-C. Lee, X. He, W. You, R. Zhou, J.-T. Guo, X. Chen, X. Peng, H. Sun, H. Huang, H. Zhao, and B. Feng. Direct activation of human and mouse Oct4 genes using engineered TALE and Cas9 transcription factors. *Nucleic acids research* **2014**; *42*(7):4375–4390.

I. Inc. Off-Line Basecaller v1.9.4 User Guide. Illumina Inc., **2011**a.

I. Inc. Quality scores for next generation sequencing **2011**b;.

I. Inc. RTA 1.13, HCS 1.5, and SCS 2.10, Theory of Operation. Illumina Inc., **2011**c.

I. Inc. Understanding Illumina quality scores **2014**;.

S. Inoué. Video microscopy. Springer Science & Business Media, **2013**.

S. a. Jaeger, E. T. Chan, M. F. Berger, R. Stottmann, T. R. Hughes, and M. L. Bulyk. Conservation and regulatory associations of a wide affinity range of mouse transcription factor binding sites. *Genomics* **2010**;*95*(4):185–195.

M. H. Jia, R. A. LaRossa, J.-M. Lee, A. Rafalski, E. DeRose, G. Gonye, and Z. Xue. Global expression profiling of yeast treated with an inhibitor of amino acid biosynthesis, sulfometuron methyl. *Physiological Genomics* **2000**;*3*(2):83–92.

A. Jolma, T. Kivioja, J. Toivonen, L. Cheng, G. Wei, M. Enge, M. Taipale, J. M. Vaquerizas, J. Yan, M. J. Sillanpää, M. Bonke, K. Palin, S. Talukder, T. R. Hughes, N. M. Luscombe, E. Ukkonen, and J. Taipale. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Research* **2010**;*20*(6):861–873.

A. Jolma, J. Yan, T. Whitington, J. Toivonen, K. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. Wei, K. Palin, J. Vaquerizas, R. Vincentelli, N. Luscombe, T. Hughes, P. Lemaire, E. Ukkonen, T. Kivioja, and J. Taipale. DNA-Binding Specificities of Human Transcription Factors. *Cell* **2013**;*152*(1-2):327–339.

J. K. Joung and J. D. Sander. TALENs: a widely applicable technology for targeted genome editing. *Nature reviews Molecular cell biology* **2013**;*14*(1):49–55.

C. Jung, P. Bandilla, M. von Reutern, S. Lange, U. Unnerstall, and U. Gaul. High sensitivity measurement of transcription factor-DNA binding energies by automated fluorescence microscopy. *in press* **2015**;.

B. Kim and J. W. Little. Dimerization of a specific DNA-binding protein on the DNA. *Science (New York, NY)* **1992**;*255*(5041):203–6.

J. Kim and K. Struhl. Determinants of half-site spacing preferences that distinguish AP-1 and ATF/CREB bZIP domains. *Nucleic acids research* **1995**;*23*(13):2531–2537.

M. Klein and T. E. Furtak. Optik. Springer-Verlag, **2013**.

J. D. Klemm and C. O. Pabo. Oct-1 POU domain-DNA interactions: cooperative binding of isolated subdomains and effects of covalent linkage. *Genes & development* **1996**;*10*(1):27–36.

R. Klette. Concise computer vision. Springer, **2014**.

P. König and T. J. Richmond. The X-ray structure of the GCN4-bZIP bound to ATF/CREB site DNA shows the complex depends on DNA flexibility. *Journal of molecular biology* **1993**; *233*(1):139–154.

W. H. Kruskal and W. A. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association* **1952**;*47*(260):583–621.

D. Krylov. Leucine Zipper. *eLS* **2001**;pages 1–7.

L. Kubecka, J. Jan, and R. Kolar. Retrospective illumination correction of retinal images. *International journal of biomedical imaging* **2010**;*2010*:780262.

C. D. Kuglin and D. C. Hines. The phase correlation image alignment method. **1975**.

C. Leahy, A. O'Brien, and C. Dainty. Illumination correction of retinal images using Laplace interpolation. *Applied optics* **2012**;*51*(35):8383–8389.

K. a. Lee. Dimeric transcription factor families: it takes two to tango but who decides on partners and the venue? *Journal of cell science* **1992**;*103 (Pt 1*:9–14.

F. J. W.-M. Leong, M. Brady, and J. O. McGee. Correction of uneven illumination (vignetting) in digital microscopy images. *Journal of clinical pathology* **2003**;*56*(8):619–21.

K. Levenberg. A method for the solution of certain non–linear problems in least squares **1944**;.

M. Levo, E. Zalckvar, E. Sharon, A. C. Dantas Machado, Y. Kalma, M. Lotam-Pompan, A. Weinberger, Z. Yakhini, R. Rohs, and E. Segal. Unraveling determinants of transcription factor binding outside the core binding site. *Genome Research* **2015**a;page gr.185033.114.

M. Levo, E. Zalckvar, E. Sharon, A. C. Dantas Machado, Y. Kalma, M. Lotam-Pompan, A. Weinberger, Z. Yakhini, R. Rohs, and E. Segal. Unraveling determinants of transcription factor binding outside the core binding site. *Genome Research* **2015**b;page gr.185033.114.

F. Li and G. Sethi. Targeting transcription factor NF-κB to overcome chemoresistance and radioresistance in cancer therapy. *Biochimica et Biophysica Acta - Reviews on Cancer* **2010**; *1805*(2):167–180.

Likar, Maintz, Viergever, and Pernus. Retrospective shading correction based on entropy minimization. *Journal of Microscopy* **2000**;*197*(3):285–295.

B. Likar and F. Pernuš. Retrospective shading correction of microscopical images. In Proc. Czech Pattern Recognition Workshop, Eds. Svoboda, PT. **2000**; pages 15–20.

J. S. Lim. Two-dimensional signal and image processing. *Englewood Cliffs, NJ, Prentice Hall, 1990, 710 p* **1990**;*1*.

H. Liu. Adaptive Gradient-Based and Anisotropic Diffusion Equation Filtering Algorithm for Microscopic Image Preprocessing. *Journal of Signal and Information Processing* **2013**;*4*(01):82.

L. Liu and X. D. Fan. CRISPR-Cas system: A powerful tool for genome engineering. *Plant Molecular Biology* **2014**;*85*(3):209–218.

S. J. Maerkl and S. R. Quake. A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors. *Science* **2007**;*315*(5809):233–237.

D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics* **1963**;*11*(2):431–441.

G. D. Marty et al. Blank-field correction for achieving a uniform white background in brightfield digital photomicrographs. *BioTechniques* **2007**;*42*(6):716.

C. Mascarenhas, L. C. Edwards-Ingram, L. Zeef, D. Shenton, M. P. Ashe, and C. M. Grant. Gcn4 is required for the response to peroxide stress in the yeast Saccharomyces cerevisiae. *Molecular biology of the cell* **2008**;*19*(7):2995–3007.

G. MAVROTHALASSITIS, G. BEAL, and T. S. PAPAS. Defining target sequences of DNA-binding proteins by random selection and PCR: determination of the GCN4 binding sequence repertoire. *DNA and cell biology* **1990**;*9*(10):783–788.

D. M. McHarris and D. A. Barr. Truncated variants of the GCN4 transcription activator protein bind DNA with dramatically different dynamical motifs. *Journal of chemical information and modeling* **2014**;*54*(10):2869–2875.

S. Meijsing, M. Pufall, A. So, and D. Bates. DNA binding site sequence directs glucocorticoid receptor structure and activity. *Science* **2009**;*324*(5925):407–410.

J.-F. Mercier and G. W. Slater. Solid phase DNA amplification: a Brownian dynamics study of crowding effects. *Biophysical journal* **2005**;*89*(1):32–42.

J.-F. Mercier, G. W. Slater, and P. Mayer. Solid phase DNA amplification: a simple Monte Carlo Lattice model. *Biophysical journal* **2003**;*85*(4):2075–2086.

S. J. Metallo and a. Schepartz. Distribution of labor among bZIP segments in the control of DNA affinity and specificity. *Chemistry {&} biology* **1994**;*1*(3):143–151.

J. Michálek, M. Čapek, X. Mao, and L. Kubínová. Application Of Morphology Filters To Compensation Of Lateral Illumination Inhomogeneities In Confocal Microscopy Images **2010**;.

J. R. Moll, A. Acharya, J. Gal, A. a. Mir, and C. Vinson. Magnesium is required for specific DNA binding of the CREB B-ZIP domain. *Nucleic acids research* **2002**;*30*(5):1240–1246.

S. Mukherjee, M. F. Berger, G. Jona, X. S. Wang, D. Muzzey, M. Snyder, R. a. Young, and M. L. Bulyk. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nature genetics* **2004**;*36*(12):1331–1339.

J. C. Mullikin, L. J. van Vliet, H. Netten, F. R. Boddeke, G. Van der Feltz, and I. T. Young. Methods for CCD camera characterization. In IS&T/SPIE 1994 International Symposium on Electronic Imaging: Science and Technology. International Society for Optics and Photonics, **1994**; pages 73–84.

K. Natarajan, M. R. Meyer, M. Belinda, D. Slade, C. Roberts, G. Alan, M. J. Marton, B. M. Jackson, and A. G. Hinnebusch. Transcriptional Profiling Shows that Gcn4p Is a Master Regulator of Gene Expression during Amino Acid Starvation in Yeast Transcriptional Profiling Shows that Gcn4p Is a Master Regulator of Gene Expression during Amino Acid Starvation in Yeast. *Molecular and cellular biology* **2001**;*21*(13):4347–4368.

A. Neubeck and L. Van Gool. Efficient non-maximum suppression. In Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, volume 3. IEEE, **2006**; pages 850–855.

F. S. Ng, J. Schutte, D. Ruau, E. Diamanti, R. Hannah, S. J. Kinston, and B. Gottgens. Constrained transcription factor spacing is prevalent and important for transcriptional control of mouse blood cells. *Nucleic Acids Research* **2014**;*42*(22):13513–13524.

Y. Nikolaev. Rethinking Leucine Zipper : ribonuclease activity and structural dynamics of a ubiquitous oligomerization motif **2011**;.

Y. Nikolaev, C. Deillon, S. R. K. Hoffmann, L. Bigler, S. Friess, R. Zenobi, K. Pervushin, P. Hunziker, and B. Gutte. The leucine zipper domains of the transcription factors GCN4 and c-Jun have Ribonuclease Activity. *PLoS ONE* **2010**;*5*(5).

R. Nutiu, R. C. Friedman, S. Luo, I. Khrebtukova, D. Silva, R. Li, L. Zhang, G. P. Schroth, and C. B. Burge. Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nature biotechnology* **2011**;*29*(7):659–664.

T. Oikawa and T. Yamada. Molecular biology of the Ets family of transcription factors. *Gene* **2003**;*303*(1-2):11–34.

a. R. Oliphant, C. J. Brandl, and K. Struhl. Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Molecular and cellular biology* **1989**;*9*:2944–2949.

Y. Orenstein and R. Shamir. A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. *Nucleic Acids Research* **2014**;*42*(8):e63—-e63.

N. Otsu. A threshold selection method from gray-level histograms. *Automatica* **1975**;*11*(285-296):23–27.

K. Pardee, A. S. Necakov, and H. Krause. Nuclear receptors: small molecule sensors that coordinate growth, metabolism and reproduction. In A Handbook of Transcription Factors, pages 123–153. Springer, **2011**;.

J. R. Parker. Algorithms for image processing and computer vision. John Wiley & Sons, **2010**.

C. Patil and P. Walter. Intracellular signaling from the endoplasmic reticulum to the nucleus: the unfolded protein response in yeast and mammals. *Current opinion in cell biology* **2001**; *13*(3):349–355.

C. K. Patil, H. Li, and P. Walter. Gcn4p and novel upstream activating sequences regulate targets of the unfolded protein response. *PLoS Biology* **2004**;*2*(8).

J. C. Pearson, D. Lemons, and W. McGinnis. Modulating Hox gene functions during animal body patterning. *Nature Reviews Genetics* **2005**;*6*(12):893–904.

D. Raijman, R. Shamir, and A. Tanay. Evolution and selection in yeast promoters: Analyzing the combined effect of diverse transcription factor binding sites. *PLoS Computational Biology* **2008**;*4*(1):0077–0087.

K. R. Rao, D. N. Kim, and J. J. Hwang. Fast Fourier Transform-Algorithms and Applications. Springer Science & Business Media, **2011**.

M. Reiss. The cos4 law of illumination. *J Opt Soc Am* **1945**;*35*:283–288.

C. Reyes-Aldasoro. Retrospective shading correction algorithm based on signal envelope estimation. *Electron Lett* **2009**;*45*(9):454.

K. Robasky and M. L. Bulyk. UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein–DNA interactions. *Nucleic acids research* **2011**;*39*(suppl 1):D124–D128.

R. Rohs, S. M. West, P. Liu, and B. Honig. Nuance in the double-helix and its role in protein-DNA recognition. *Current Opinion in Structural Biology* **2009**a;*19*(2):171–177.

R. Rohs, S. M. West, A. Sosinsky, P. Liu, R. S. Mann, and B. Honig. The role of DNA shape in protein-DNA recognition. *Nature* **2009**b;*461*(7268):1248–1253.

E. Roulet, S. Busso, A. a. Camargo, A. J. G. Simpson, N. Mermod, and P. Bucher. High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nature biotechnology* **2002**;*20*(8):831–835.

S. Rowan, T. Siggers, S. a. Lachke, Y. Yue, M. L. Bulyk, and R. L. Maas. Precise temporal control of the eye regulatory gene Pax6 via enhancer-binding site affinity. *Genes and Development* **2010**;*24*:980–985.

U. Rüegsegger, J. H. Leber, and P. Walter. Block of HAC1 mRNA translation by long-range base pairing is released by cytoplasmic splicing upon induction of the unfolded protein response. *Cell* **2001**;*107*:103–114.

J. C. Russ. The image processing handbook. CRC press, **2011**.

M. Samson, F. Libert, B. J. Doranz, J. Rucker, C. Liesnard, C.-M. Farber, S. Saragosti, C. Lapouméroulie, J. Cognaux, C. Forceille, et al. Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature* **1996**; *382*(6593):722–725.

J. D. Sander and J. K. Joung. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nature biotechnology* **2014**;*32*(4):347–355.

M. Schmid. MaximumFinder. `http://rsb.info.nih.gov/ij/developer/api/ij/plugin/filter/MaximumFinder.html`, **2006**.

C. A. Schneider, W. S. Rasband, K. W. Eliceiri, et al. NIH Image to ImageJ: 25 years of image analysis. *Nat methods* **2012**;*9*(7):671–675.

T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic acids research* **1990**;*18*(20):6097–6100.

N. C. Seeman, J. M. Rosenberg, and a. Rich. Sequence-specific recognition of double helical nucleic acids by proteins. *Proceedings of the National Academy of Sciences of the United States of America* **1976**;*73*(3):804–808.

E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, and U. Gaul. Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature* **2008**;*451*(7178):535–540.

E. Segal and J. Widom. From DNA sequence to transcriptional behaviour: a quantitative approach. *Nature reviews Genetics* **2009**;*10*(7):443–456.

J. W. Sellers, A. Vincent, and K. Struhl. Mutations that define the optimal half-site for binding yeast GCN4 activator protein and identify an ATF/CREB-like repressor that recognizes similar DNA sites. *Molecular and cellular biology* **1990**a;*10*(10):5077–5086.

J. W. Sellers, a. C. Vincent, and K. Struhl. Mutations that define the optimal half-site for binding yeast GCN4 activator protein and identify an ATF/CREB-like repressor that recognizes similar DNA sites. *Molecular and cellular biology* **1990**b;*10*(10):5077–5086.

S. Selvaraj, H. Kono, and A. Sarai. Specificity of protein-DNA recognition revealed by structure-based potentials: Symmetric/asymmetric and cognate/non-cognate binding. *Journal of Molecular Biology* **2002**;*322*(5):907–915.

N. C. Shaner, P. A. Steinbach, and R. Y. Tsien. A guide to choosing fluorescent proteins. *Nature methods* **2005**;*2*(12):905–909.

D. F. Shanno. Conditioning of quasi-Newton methods for function minimization. *Mathematics of computation* **1970**;*24*(111):647–656.

C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review* **2001**;*5*(1):3–55.

P. J. Shaw and D. J. Rawlins. The point-spread function of a confocal microscope: its measurement and use in deconvolution of 3-D data. *Journal of Microscopy* **1991**;*163*(2):151–165.

T. Siggers, A. B. Chang, A. Teixeira, D. Wong, K. J. Williams, B. Ahmed, J. Ragoussis, I. A. Udalova, S. T. Smale, and L. Martha. Europe PMC Funders Group Principles of dimer-specific gene regulation revealed by a comprehensive characterization of NF- κ B family DNA binding **2012**;*13*(1):95–102.

T. Siggers and R. Gordân. Protein-DNA binding: Complexities and multi-protein codes. *Nucleic Acids Research* **2014**a;*42*(4):2099–2111.

T. Siggers and R. Gordân. Protein-DNA binding: Complexities and multi-protein codes. *Nucleic Acids Research* **2014**b;*42*(4):2099–2111.

M. Slattery, T. Zhou, L. Yang, A. C. Dantas Machado, R. Gordân, and R. Rohs. Absence of a simple code: how transcription factors read the genome. *Trends in Biochemical Sciences* **2014**; *39*(9):381–399.

M. Slutsky and L. a. Mirny. Kinetics of protein-DNA interaction: facilitated target location in sequence-dependent potential. *Biophysical journal* **2004**;*87*(6):4021–4035.

A. T. Spivak and G. D. Stormo. ScerTF: a comprehensive database of benchmarked position weight matrices for Saccharomyces species. *Nucleic acids research* **2012**;*40*(D1):D162–D168.

Y. C. Staal, M. H. van Herwijnen, F. J. van Schooten, and J. H. van Delft. Application of four dyes in gene expression analyses by microarrays. *BMC genomics* **2005**;*6*(1):101.

S. Stallinga and B. Rieger. Accuracy of the Gaussian Point Spread Function model in 2D localization microscopy **2010**;*18*(24):24461–24476.

D. Stanojevic and G. L. Verdine. Deconstruction of GCN4/GCRE into a monomeric peptide-DNA complex. *Nature Structural & Molecular Biology* **1995**;*2*(6):450–457.

C. M. Stellrecht and L. S. Chen. Transcription inhibition as a therapeutic target for cancer. *Cancers* **2011**;*3*(4):4170–4190.

S. R. Sternberg. Biomedical image processing. *Computer* **1983**;*1*(16):22–34.

D. Stone, H.-P. Kiem, and K. R. Jerome. Targeted gene disruption to cure HIV. *Curr Opin HIV AIDS* **2013**;*18*(9):1199–1216.

G. D. Stormo. DNA binding sites: representation and discovery. *Bioinformatics (Oxford, England)* **2000**;*16*(1):16–23.

B. Stroustrup. The C++ programming language. Pearson Education India, **1986**.

K. Struhl. The DNA-binding domains of the jun oncoprotein and the yeast GCN4 transcriptional activator protein are functionally homologous. *Cell* **1987**;*50*:841–846.

M. Suckow, B. von Wilcken-Bergmann, and B. Müller-Hill. Identification of three residues in the basic regions of the bZIP proteins GCN4, C/EBP and TAF-1 that are involved in specific DNA binding. *The EMBO journal* **1993**;*12*(3):1193–1200.

A. Tanay. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Research* **2006**;*16*(8):962–972.

P. Tebas, D. Stein, W. W. Tang, I. Frank, S. Q. Wang, G. Lee, S. K. Spratt, R. T. Surosky, M. A. Giedlin, G. Nichol, et al. Gene editing of CCR5 in autologous CD4 T cells of persons infected with HIV. *New England Journal of Medicine* **2014**;*370*(10):901–910.

D. Tomazevic, B. Likar, and F. Pernus. Comparative evaluation of retrospective shading correction methods. *Journal of Microscopy* **2002**;*208*(3):212–223.

J. M. Tome, A. Ozer, J. M. Pagano, D. Gheba, G. P. Schroth, and J. T. Lis. Comprehensive analysis of RNA-protein interactions by high-throughput sequencing-RNA affinity profiling. *Nature methods* **2014**;*11*(6):683–8.

S. Q. Tsai and J. K. Joung. What's changed with genome editing? *Cell Stem Cell* **2014**; *15*(1):3–4.

D. Tschumperle and R. Deriche. Vector-valued image regularization with PDEs: A common framework for different applications. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **2005**;*27*(4):506–517.

F. D. Urnov, E. J. Rebar, M. C. Holmes, H. S. Zhang, and P. D. Gregory. Genome editing with engineered zinc finger nucleases. *Nature reviews Genetics* **2010**;*11*(9):636–646.

C. P. Verrijzer, J. A. van Oosterhout, and P. C. van der Vliet. The Oct-1 POU domain mediates interactions between Oct-1 and other POU proteins. *Mol Cell Biol* **1992**;*12*(2):542–551.

I. M. Vinogradov and M. Hazewinkel. Encyclopaedia of mathematics. Kluwer Academic Publ, **2001**.

P. H. von Hippel and O. Berg. Facilitated target location in biological systems. *Journal of Biological Chemistry* **1989**;*264*(2):675–678.

U. Vovk, F. Pernuš, and B. Likar. Intensity inhomogeneity correction of multispectral MR images. *Neuroimage* **2006**;*32*(1):54–61.

T. Wagner and H.-G. Lipinski. IJBlob: An ImageJ Library for Connected Component Analysis and Shape Analysis. *Journal of Open Research Software* **2013**;*1*(1).

G. Wang, Y. Wang, H. Li, X. Chen, H. Lu, Y. Ma, C. Peng, Y. Wang, and L. Tang. Morphological Background Detection and Illumination Normalization of Text Image with Poor Lighting **2014**; .

J. Wang, J. Lu, G. Gu, and Y. Liu. In vitro DNA-binding profile of transcription factors: Methods and new insights. *Journal of Endocrinology* **2011**;*210*(1):15–27.

J. C. Waters. Accuracy and precision in quantitative fluorescence microscopy. *The Journal of cell biology* **2009**;*185*(7):1135–1148.

J. C. Waters and J. R. Swedlow. Techniques Interpreting Fluorescence Microscopy Images and Measurements. *Evaluating Techniques in Biochemical Research* **2007**;pages 36–42.

G. H. Wei, G. Badis, M. F. Berger, T. Kivioja, K. Palin, M. Enge, M. Bonke, A. Jolma, M. Varjosalo, A. R. Gehrke, J. Yan, S. Talukder, M. Turunen, M. Taipale, H. G. Stunnenberg, E. Ukkonen, T. R. Hughes, M. L. Bulyk, and J. Taipale. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *Embo J* **2010**;*29*(13):2147–2160.

M. T. Weirauch, A. Cote, R. Norel, M. Annala, Y. Zhao, T. R. Riley, J. Saez-Rodriguez, T. Cokelaer, A. Vedenko, S. Talukder, H. J. Bussemaker, Q. D. Morris, M. L. Bulyk, G. Stolovitzky, and T. R. Hughes. Evaluation of methods for modeling transcription factor sequence specificity. *Nature biotechnology* **2013**;*31*(2):126–34.

E. W. Weisstein. CRC encyclopedia of mathematics. CRC Press, **2009**.

W. Wen-Cheng and C. Xiao-Jun. A Segmentation Method for Uneven Illumination Particle Images. *Research Journal of Applied Sciences, Engineering and Technology* **2013**;*5*(4):1284–1289.

N. Whiteford, T. Skelly, C. Curtis, M. E. Ritchie, A. Löhr, A. W. Zaranek, I. Abnizova, and C. Brown. Swift: Primary data analysis for the Illumina Solexa sequencing platform. *Bioinformatics* **2009**;*25*(17):2194–2199.

C. R. Wobbe, J. P. Lee, S. C. I-larrisoni, and K. Struhl. Folding transition in the DNA-binding domain of GCN4 on specific binding to DNA. *Nature* **1990**;*347*:11.

D. Wong, A. Teixeira, S. Oikonomopoulos, P. Humburg, I. Lone, D. Saliba, T. Siggers, M. Bulyk, D. Angelov, S. Dimitrov, I. a. Udalova, and J. Ragoussis. Extensive characterization of NF-κB binding uncovers non-canonical motifs and advances the interpretation of genetic functional traits. *Genome Biology* **2011**;*12*(7):R70.

Q. Wu, F. Merchant, and K. Castleman. Microscope image processing. Academic press, **2010**.

H.-T. Yang, C.-P. Hsu, and M.-J. Hwang. An Analytical Rate Expression for the Kinetics of Gene Transcription Mediated by Dimeric Transcription Factors. *Journal of Biochemistry* **2007**; *142*(2):135–144.

J. E. Yeh, P. A. Toniolo, and D. A. Frank. Targeting transcription factors. *Current Opinion in Oncology* **2013**;*25*(6):652–658.

B. Zhang, J. Zerubia, and J.-C. Olivo-Marin. Gaussian approximations of fluorescence microscope point-spread function models. *Applied Optics* **2007**;*46*(10):1819–1829.

X. Zhang, Y. Xin, N. Xie, W. Jiang, and Y. Zhao. Shading Surface Estimation using Piecewise Polynomials for Binarizing Unevenly Illuminated Document Images **2014**;.

Y. Zhao, S. Ruan, M. Pandey, and G. D. Stormo. Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics* **2012**;*191*(3):781–790.

Y. Zheng, S. Lin, C. Kambhamettu, J. Yu, and S. B. Kang. Single-image vignetting correction. *IEEE transactions on pattern analysis and machine intelligence* **2009**;*31*(12):2243–2256.

T. Zhou, N. Shen, L. Yang, N. Abe, J. Horton, R. S. Mann, H. J. Bussemaker, R. Gordân, and R. Rohs. Quantitative modeling of transcription factor binding specificities using DNA shape. *Proceedings of the National Academy of Sciences* **2015**;page 201422023.

C. Zhu, K. Byers, R. McCord, Z. Shi, M. Berger, D. Newburger, K. Saulrieta, Z. Smith, M. Shah, M. Radhakrishnan, a. Philippakis, Y. Hu, F. De Masi, M. Pacek, a. Rolfs, T. Murthy, J. Labaer, and M. L. Bulyk. High-resolution DNA binding specificity analysis of yeast transcription factors. *Genome Res* **2009**;pages 556–566.

C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)* **1997**;*23*(4):550–560.