

Inferring structural properties of protein-DNA binding using high-throughput sequencing. The paradigm of GATA1, KLF1 and their complexes GATA1/FOG1 and GATA1/KLF1. Insights into the transcriptional regulation of the erythroid cell lineage.



Spyros Oikonomopoulos

Wellcome Trust Centre For Human Genetics

&

Keble College, University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Hilary Term 2014

Inferring structural properties of protein-DNA binding using high-throughput sequencing. The paradigm of GATA1, KLF1 and their complexes GATA1/FOG1 and GATA1/KLF1. Insights into the transcriptional regulation of the erythroid cell lineage.

Abstract

GATA1 and KLF1 are transcription factors that regulate genes which are important for the development of erythroid cells. The GATA1 transcriptional co-factor FOG1 has been shown to be essential in a wide range of GATA1 dependent cellular functions. Here we tried to understand the diverse mechanisms by which GATA1 and KLF1 recognize their binding sites, how the GATA1 recognition mechanisms are affected by complexation with either FOG1 or KLF1 and how the GATA1 recognition mechanisms affect the transcriptional regulation of the erythroid differentiation. We profiled the DNA binding specificities/affinities of a GATA1 fragment (mGATA1_{NC}), that contains only the two DNA binding domains (N-terminal and C-terminal Zn finger), and the DNA binding specificities/affinities of a KLF1 fragment (mKLF1₂₅₇₋₃₅₈), that contains the three DNA binding domains, using a novel methodology that combines EMSA with high throughput sequencing (EMSA-seq (Wong et al., 2011a)). We also profiled the DNA binding specificities of the C-terminal Zn finger of GATA1 alone (mGATA1_C), the wt-mGATA1, the wt-mGATA1/wt-mFOG1 complex and the mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex. At first, we confirmed that the N-terminal Zn finger of GATA1 has a strong preference for the “GATC”

motif, whereas the C-terminal Zn finger of GATA1 has a strong preference for the “GATA” motif. Next, we found that in the mGATA1_{NC}, both DNA binding domains can bind simultaneously a wide range of different positional combinations of the co-occurring “GATA” and “GATC” motifs, on the same DNA sequence. The wt-mGATA1 did not show the ability to bind in the same co-occurring motifs implying an effect of the non-DNA binding domains of the protein in the regulation of its DNA binding specificities. On the contrary, complexation of wt-mGATA1 with the wt-mFOG1 partially restored its ability to bind in a now limited range of different positional combinations of the co-occurring “GATA” and “GATC” motifs, on the same DNA sequence. Similar observations were made for other pairs of GATA1 N-terminal and C-terminal Zn finger specific motifs. We then projected the GATA1 DNA binding specificities/affinities in vivo and we classified the GATA1 ChIP-seq peaks in low, medium or high affinity based on the number of the GATA1 motifs. We noticed that high affinity GATA1 ChIP-seq peaks tend to appear in regions with low nucleosome occupancy. We also noticed that GATA1 ChIP-seq peaks in the enhancer regions are usually high affinity whereas GATA1 ChIP-seq peaks in the proximal promoter regions are usually low affinity. Additionally, we observed that high affinity GATA1 ChIP-seq peaks are usually found in regions with increased levels of H3K4me2 and are associated with a higher decrease in the H3K4me3 levels on the TSS of the nearby genes. None of these GATA1 related in vivo observations were found for the KLF1 ChIP-seq positions. These findings significantly advance our understanding of the DNA binding properties of GATA1, KLF1 and their complexes and give an insight on the importance of the GATA1 DNA binding affinities in the regulation of the erythroid transcriptional program. Overall the work establishes an experimental and analytical framework to investigate how transcriptional co-factors can change the DNA binding specificities of specific transcription factors and how integration of the transcription factor DNA binding affinities with in vivo data can give novel insights into the transcriptional regulation.

Declaration

Unless otherwise indicated, all the work in this thesis was performed solely by the author at the Wellcome Trust Centre For Human Genetics, University of Oxford.

The work reported in this thesis has not been submitted for any other degree in this or any other university or institute of learning.

Acknowledgements

My thanks go to:

- My supervisor Dr Ioannis Ragoussis for his help and support throughout all these years.
- All the members of the Genomics Research lab (Dr. Daniel Wong, Dr Ana Teixeira, Dr Carme Camps, Dr Hani Choudhry) for their useful advice.
- The funding agency for providing the necessary means to accomplish part of this work under the framework of FP7 Marie Curie Initial Training Network “Integer” (Integrated Gene Regulation: From DNA sequence to Nuclear Structure).

Abbreviations

AKT	protein kinase B
ALAS2	aminolevulinate, delta-, synthase 2
AML1/RUNX1	acute myeloid leukemia 1 protein
AR	androgen receptor
AREA	nitrogen regulatory protein areA
BCL2	B-cell lymphoma 2
BFU-e	burst-forming unit-erythroid
BMP4	bone morphogenetic protein 4
BRG1	Brahma-related gene 1
P/CAF	P300/CBP-associated factor
CBF-beta	core-binding factor subunit beta
CBP	CREB-binding protein
CD117	mast/stem cell growth factor receptor Kit
CD34	hematopoietic progenitor cell antigen CD34
CD36	Thrombospondin receptor/Platelet glycoprotein 4
CD71	Transferrin receptor protein 1
CDX4	Caudal-type homeobox protein 4
CFP1	CXXC-type zinc finger protein 1/CpG-binding protein
CFU-e	erythroid Colony-Forming Unit
ChIP	Chromatin Immunoprecipitation
COMPASS	Complex Proteins Associated with Set1
CREB	cAMP response element-binding protein
Ct	PCR cycle number
CtBP	C-terminal-binding protein 1
CTD	C-terminal domain tail
DMSO	Dimethyl sulfoxide

EBF	Early B-Cell Factor
C/EBPa	CCAAT enhancer binding protein alpha
EDTA	Ethylenediaminetetraacetic acid
EGTA	Ethylene glycol-bis(2-aminoethylether)-N,N,N',N'-tetraacetic acid
EMSA	Electrophoretic Mobility Shift Assay
EPO	Erythropoietin
EpoR	Erythropoietin receptor
ER-alpha	Estrogen receptor alpha
FACS	Fluorescence-Activated Cell Sorting
FBS	Fetal Bovine Serum
FDR	False Discovery Rate
FLI1	Friend Leukemia Virus Integration 1
FRAP	Fluorescence Recovery After Photobleaching
GST	Glutathione S-transferase
HA	hemagglutinin
HBB	hemoglobin beta
HDAC1	Histone deacetylase 1
HDAC2	Histone deacetylase 2
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HIF1a	Hypoxia-inducible factor 1-alpha
His-tag	polyhistidine tag
HOXA9	Homeobox protein Hox-A9
HOXB7	Homeobox protein Hox-B7
HPCs	Haemopoietic Progenitor Cells
HPLC	High Performance Liquid Chromatography
HSCs	hematopoietic stem cells
IL3	Interleukin-3

IL6	Interleukin-6
IPTG	isopropyl beta-D-1-thiogalactopyranoside
IRF-1	Interferon regulatory factor 1
JARID1/KDM5	Lysine-specific demethylase 5
KDM2A	Lysine-specific demethylase 2A
KLF1	Krueppel-like zinc finger transcription factor 1
KLF3	Krueppel-like zinc finger transcription factor 3
KLF4	Krueppel-like zinc finger transcription factor 4
KLF8	Krueppel-like zinc finger transcription factor 8
LCR	Locus Control Region
LDB1	LIM domain-binding protein 1
LMO2	LIM domain only protein 2
LSD1	Lysine-specific histone demethylase 1A
MAPK	Mitogen-activated protein kinase 1
MED1/TRAP220	Thyroid hormone receptor-associated protein complex 220 kDa component
MEL	Murine erythroleukemia (MEL) cell
MLL1	Myeloid/lymphoid or mixed-lineage leukemia protein 1/Histone-lysine N-methyltransferase 2A
MLL2	Myeloid/lymphoid or mixed-lineage leukemia protein 2/Histone-lysine N-methyltransferase 2D
MLL3	Myeloid/lymphoid or mixed-lineage leukemia protein 3/Histone-lysine N-methyltransferase 2C
MLL4	Myeloid/lymphoid or mixed-lineage leukemia protein 3/Histone-lysine N-methyltransferase 2B
MOZ	Monocytic leukemia zinc finger protein/Histone acetyltransferase KAT6A

NF-E2p45	Nuclear factor, erythroid-derived 2 45 kDa subunit
NMR	Nuclear magnetic resonance
NURD	Nucleosome Remodeling Deacetylase
p300	Histone acetyltransferase p300
p53	Cellular tumor antigen p53
PBM	Protein Binding Microarray
PBMs	Protein Binding Microarrays
PF4	Platelet factor 4
PI3K	Phosphoinositide 3-kinase
PIC	preinitiation complex
PPAR-gamma	Peroxisome proliferator-activated receptor gamma
PVDF	Polyvinylidene difluoride
PWM	Position Weight Matrix
RBC	Red Blood Cell
RNAPII	RNA polymerase II
SCL/TAL1	Stem cell protein/T-cell acute lymphocytic leukemia protein 1 homolog
SDS-PAGE	Sodium Dodecyl Sulfate PolyAcrylamide Gel Electrophoresis
SELEX	Systematic Evolution of Ligands by EXponential Enrichment
SAGE	Serial Analysis of Gene Expression
SET1A	SET domain-containing protein 1A
SET1B	SET domain-containing protein 1B
SF9	Spodoptera frugiperda cells clone 9
STAT5	Signal transducer and activator of transcription 5
SWI/SNF	SWItch/Sucrose NonFermentable
TBE	Tris-borate-EDTA
TCF/LEF1	T-cell-specific transcription factor

TF	Transcription Factor
TFIID	Transcription factor II D
TFIIH	Transcription factor II Human
TFs	Transcription Factors
TIMP-1	Tissue Inhibitor Of Metalloproteinases 1
TPO	Thrombopoietin
TSS	Transcription Start Site
USH	Protein U-shaped
ZFPM1	Zinc Finger Protein, FOG Family Member 1

Contents

Abstract	i
Declaration.....	iv
Acknowledgements	v
Abbreviations.....	vi
CHAPTER 1 INTRODUCTION.....	1
1.1 Hematopoiesis	1
1.2 Erythroid differentiation.....	2
1.3 Transcription factors and transcriptional co-factors that control the differentiation of the erythroid lineage	7
1.3.1 GATA1.....	7
1.3.1.1 Functional domains of GATA1	8
1.3.2 TAL1	10
1.3.3 FOG1.....	11
1.3.3.1 Functional domains of FOG1	12
1.3.3.2 Effect of GATA1/FOG1 complexation	13
1.3.4 KLF1	15
1.3.4.1 Effect of KLF1 on gene regulation in the erythroid lineage.....	16
1.3.5 Additional GATA1 complexes.....	18
1.4 Epigenetic changes in chromatin during erythroid differentiation.....	19
1.5 Models of erythropoiesis.....	24
1.5.1 Isolation of fetal liver cells from E14.5 mice	24
1.5.2 Differentiation of a mouse erythroleukaemia cell line (MEL cells)	25
1.5.3 Differentiation of GATA1-/- mouse ES cells upon GATA1 restoration	25
1.5.4 Differentiation of human CD34+ cells from peripheral blood	26
1.5.5 Differentiation of human K562 cells.....	26
1.6 Experimental strategies to investigate transcription factor-DNA interactions.....	29
1.6.1 High-throughput methods to characterize protein-DNA interactions in vitro	29
1.6.1.1 Sequencing based methods to characterize protein-DNA interactions	29
1.6.1.2 Microarray based methods to characterize protein-DNA interactions	35
1.6.1.3 Additional platforms to characterize protein-DNA interactions.....	36
1.6.2 Methods to characterize protein-DNA interactions in vivo	37
1.7 Definition of transcription factor binding specificities.....	38

1.8 Importance of transcription factor DNA binding specificities and their relevance on the induction of gene expression	39
Aims of the project.....	40
CHAPTER 2 MATERIALS AND METHODS	41
2.1 Materials	41
2.2 Amino acid sequence of proteins that were used in this study.....	41
2.3 Expression, purification and characterization of the protein and the protein complexes used in this study	45
2.3.1 Protein expression and purification from insect cells	45
2.3.2 Protein expression and purification from bacteria cells	47
2.3.3 Electrophoretic Mobility Shift Assay (EMSA) for testing the functionality of proteins and the protein complexes	50
2.3.4 Western Blot for the characterization of the isolated proteins and protein complexes.....	51
2.3.5 EMSA-Western for the characterization of the proteins and protein complexes in the DNA band shifts	52
2.4 EMSA-seq experiments	52
2.4.1 Creation of the 20mer degenerate DNA libraries that were used in the EMSA-seq experiments.....	52
2.4.2 Creation of protein-DNA binding isotherms with fluorescence polarization assays	54
2.4.3 Electrophoretic Mobility Shift Assay (EMSA) for the EMSA-seq experiments	56
2.4.4 EMSA-seq library preparation	59
2.4.5 Bioinformatic processing of the sequenced files from the EMSA-seq experiments	60
2.4.6 EMSA-seq analysis	62
2.4.7 Calculation of the EMSA-seq coefficient (e).....	63
2.4.8 DNA binding motif creation	67
2.4.9 Statistical methods	67
2.4.10 Binding models of the purified proteins and their protein complexes.....	68
2.5 FOG1 ChIP-seq	69
2.5.1 ChIP-seq of FOG1 from murine erythroleukemia cells.....	69
2.5.2 Data processing of the FOG1 ChIP-seq	69
2.6 Analysis of publicly available ChIP-seq datasets.....	70
CHAPTER 3 RESULTS.....	72
3.1 Expression, purification and DNA band shifts of GATA1 and KLF1 proteins and of the GATA1/FOG1, GATA1/KLF1 complexes.....	74

3.1.1 Creation of fusion proteins with specific purification tags.....	74
3.1.1.1 Kinetics of protein production of wt-GATA1 and wt-GATA1/wt-FOG1 in the baculovirus expression system	77
3.1.1.2 Protein production in the bacterial expression system.....	81
3.1.1.3 Purification of proteins from the baculovirus or the bacterial expression system	81
3.1.2 DNA band shift of a 20mer degenerate DNA library from the GATA1 and KLF1 proteins and the complexes GATA1/FOG1 and GATA1/KLF1	90
3.2 EMSA-seq can be used to calculate both the DNA binding specificities and the DNA binding affinities (dissociation constant-Kd) of a transcription factor	99
3.2.1 Non-cooperative competition model in an EMSA-seq reaction.....	99
3.2.2 Cooperative competition model in an EMSA-seq reaction.....	108
3.3 Comparison of published GATA1 DNA binding specificities.....	116
3.4 Investigation of the DNA binding specificities and the DNA binding affinities of a transcription factor using EMSA-seq	119
3.4.1 DNA binding specificities of mGATA1 _{NC}	125
3.4.2 The N-terminal GATA1 Zn finger binds to “GATC” and “GATG” motifs (DNA binding specificities of mGATA1 _C).....	132
3.4.2.1 Absence of the N-terminal Zn finger changes the binding specificities of the C-terminal Zn finger.....	135
3.4.2.2 Absence of the N-terminal Zn finger eliminates the binding in palindromic sequences	138
3.4.3 Full length GATA1 binds on the DNA with only the C-terminal Zn finger, the N-terminal Zn finger shows no binding.....	141
3.4.4 Binding of two molecules of GATA1 independently or as a dimer on the DNA, changes the distance distribution between the specific GATA1 N-terminal and the C-terminal Zn finger DNA binding motifs	145
3.4.5 Integration of the GATA1 EMSA-seq results with other in vitro data.....	156
3.4.5.1 DNA motif distances of mGATA1 _{NC} , mGATA1 _C and their relationship with solved crystal structures	156
3.4.5.2 DNA motif distances of mGATA1 _{NC} , mGATA1 _C and their relationship with biochemical studies	162
3.4.6 DNA binding specificities of KLF1.....	169
3.4.7 Binding cooperativity as is reflected in the EMSA-seq coefficient (“e”) indicates two different modes of binding for GATA1 and KLF1	174
3.4.8 Effect of FOG1 and KLF1 complexation on the DNA binding specificities of GATA1.....	182
3.4.8.1 Complexation of GATA1 with FOG1 increases the preference for GATAA and GATC containing motifs.....	182

3.4.8.2 Complexation with FOG1 increases the preference for the GATC{N} ₁₋₂ GATA motifs	185
3.4.8.3 DNA binding specificities of mGATA1 _{NC} /mFOG1 _{finger1} , mGATA1 _{NC} /mFOG1 _{finger56} , mGATA1 _{NC} /mFOG1 _{finger9}	200
3.4.8.4 Similarities in the protein sequence of the FOG1 fingers may explain the similar DNA binding specificities	208
3.4.8.5 Functional redundancy of the FOG1 fingers in vivo may explain the similar DNA binding specificities	211
3.4.9 Effect of KLF1/GATA1 complexation on the binding specificities of GATA1	213
3.4.9.1 Complexation of KLF1 with GATA1 changes the distance distribution between the C-terminal and the N-terminal Zn finger of GATA1	215
3.4.10 Effect of KLF1/GATA1 complexation on the binding specificities of KLF1	228
3.4.11 The orientation of the GATA1-KLF1 complex in vivo indicates a mechanism of local DNA looping	231
3.5 Projection of the EMSA-seq DNA binding affinities/specificities on the GATA1 related erythropoietic gene regulation models	238
3.5.1 Calculation of the DNA binding affinities of GATA1/KLF1 ChIP-seq peaks from the “Kd” values of each mGATA1 _{NC} /mKLF1 ₂₅₇₋₃₅₈ 8mer	238
3.5.2 GATA1 ChIP-seq binding affinities, assigned from the EMSA-seq results, are representative of the GATA1 DNA binding turnover in vivo	244
3.5.2.1 Prediction of the kinetic recruitment of the protein upon GATA1 restoration in a GATA1 -/- cell line	244
3.5.3 EMSA-seq binding affinity distributions of ChIP-seq peaks can be used to characterize the similarity between ChIP-seq experiments of the same protein under different conditions	254
3.5.4 GATA1 low affinity ChIP-seq peaks are closer to the TSS. KLF1 does not show this pattern	256
3.5.5 GATA1 high affinity ChIP-seq peaks are present in enhancers whereas GATA1 low affinity ChIP-seq peaks are present in promoters. KLF1 high/low affinity ChIP-seq peaks are equally probable to appear in enhancers or promoters.	258
3.5.6 GATA1 high affinity ChIP-seq peaks are present in regions of low nucleosome occupancy and of reduced DNase I hypersensitivity whereas GATA1 low affinity ChIP-seq peaks are present in regions of high nucleosome occupancy and of increased DNase I hypersensitivity. KLF1 high/low affinity ChIP-seq peaks are equally probable to appear in regions of high/low nucleosome occupancy and of high/low DNase I hypersensitivity	261
3.5.7 Co-factors change the DNA binding specificities of GATA1 and can appear in low affinity GATA1 ChIP-seq peaks exclusively	280
3.5.7.1 DNA binding specificities of GATA1 change upon complexation with the TAL1 protein. Results from ChIP-seq data	281

3.5.7.2 Projection of the EMSA-seq results on ChIP-seq data from the erythropoiesis lineage shows that the DNA binding specificities of GATA1 change upon complexation with the p300 protein and that the GATA1/p300 overlapping sites appear in low affinity GATA1 ChIP-seq peaks.....	282
3.5.7.3 Projection of the EMSA-seq results on ChIP-seq data from the erythropoiesis lineage shows that the DNA binding specificities of GATA1 do not change in the presence of NF-E2 and that the GATA1/NF-E2 overlapping sites appear in low, medium and high affinity GATA1 ChIP-seq peaks.....	283
3.5.7.4 Projection of the EMSA-seq results on ChIP-seq data from multipotential hematopoietic progenitors shows that the DNA binding specificities of GATA1 do not change in the presence of PU.1 and that the GATA1/PU.1 overlapping sites appear in low, medium and high affinity GATA1 ChIP-seq peaks.....	284
3.5.7.5 Projection of the EMSA-seq results on ChIP-seq data from the megakaryopoiesis lineage shows that the DNA binding specificities of GATA1 change in the presence of FLI1 and that the GATA1/FLI1 overlapping sites appear in low affinity GATA1 ChIP-seq peaks.....	285
3.5.7.6 Projection of the EMSA-seq results on ChIP-seq data from the megakaryopoiesis lineage shows that the DNA binding specificities of GATA1 change in the presence of RUNX1 and that the GATA1/RUNX1 overlapping sites appear in low affinity GATA1 ChIP-seq peaks	286
3.5.7.7 GATA1/KLF1 overlapping sites but not GATA1/TAL1 overlapping sites appear in low affinity GATA1 ChIP-seq peaks in the TER119-/TER119+ mouse model of erythropoiesis.....	287
3.5.8 Binding affinities of ChIP-seq peaks give insights into the transcriptional regulatory mechanisms during erythroid and megakaryocytic differentiation.....	300
3.5.8.1 GATA1 high affinity ChIP-seq peaks are present in regions with increased levels of H3K4me2 and are associated with a higher decrease in the H3K4me3 levels of the closest TSS	300
3.5.8.2 The DNA binding affinity of the GATA1 ChIP-seq peaks cannot discriminate the upregulated and the downregulated genes during erythroid differentiation.....	303
3.5.8.3 The change in the TSS related H3K4me3 levels during erythroid differentiation, for both the upregulated and the downregulated genes, is correlated with the DNA binding affinity of the closest promoter associated GATA1 ChIP-seq peak	305
3.5.8.4 The same gene in different lineages, regulated by GATA1 from different genomic positions (10 kb +/- of the gene TSS) which have different binding affinities, shows a change in the H3K4me3 levels that corresponds to the binding affinity of the GATA1 ChIP-seq peak in the specific lineage.....	324
3.5.8.5 The GATA1/GATA2 switch occurs in sites where the GATA1 DNA binding affinity is higher than the GATA2 binding affinity.....	333
3.5.8.6 Low and high affinity GATA1 ChIP-seq peaks control different functional categories of genes.....	339

Results summary	343
CHAPTER 4 DISCUSSION	345
4.1 EMSA-seq is a versatile platform for studying protein-DNA interactions	345
4.2 The microscopic properties of the GATA1 DNA binding domains are regulated by both intrinsic and extrinsic factors.....	347
4.3 The macroscopic properties of the GATA1 DNA binding affinity distribution characterize the transcriptional regulation of erythropoiesis	349
4.3.1 The EMSA-seq derived DNA binding affinities of the GATA1 ChIP-seq peaks reflect the relative DNA residence time of GATA1	349
4.3.2 Complexes of GATA1 with some of its transcriptional co-factors, that can selectively increase its binding affinities, can exclusively appear in genomic positions where GATA1 has short residence time when it binds alone	351
4.3.3 The DNA residence time of GATA1 affects the nucleosome occupancy levels of the binding site. It additionally affects the levels of the H3K4me2 and the H3K4me1 histone marks on the binding site and the levels of the H3K4me3 on the nearest TSS	353
4.3.4 GATA1 can affect transcription by modulating the H3K4me3 levels of the closest TSS	358
4.3.5 GATA1 modulates the H3K4me3 levels through an unknown mechanism.....	360
4.3.6 The high occurrence of GATA1 binding motifs in enhancers probably increases the likelihood of interaction with the compatible promoter	364
Future experiments	369
References	372

CHAPTER 1 INTRODUCTION

1.1 Hematopoiesis

Hematopoiesis is a highly organized multi-step process that involves the proliferation, differentiation and maturation of a small population of pluripotent hematopoietic stem cells (HSCs) into various specialized and distinct blood cell types. During this process, cellular specification is initiated by a set of early transcription factors. Commitment and differentiation along a particular lineage is regulated by late transcription factors that cooperate or antagonize with the early transcription factors.

The differentiation program regulates either the lymphoid lineage or the myeloid lineage. Each of these lineages is regulated by a different set of transcription factors (Figure 1).

The formation and/or maintenance of the HSCs is attributed to the transcription factors RUNX1, TAL1, MLL, GATA2, LMO2, CDX4, SOX17 (Orkin and Zon, 2008).

The common lymphoid progenitor (CLP) gives rise to all the cells of the lymphoid lineage. The Ikaros transcription factor along with members of the Aiolos family commits the cells through the lymphoid lineage and blocks the commitment through the myeloid lineage (John and Ward, 2011). Subsequently, Pax5, PU.1, EBF, E2A are necessary for the B-cell lineage commitment (Nutt and Kee, 2007) whereas GATA3 and E2A for the T-cell lineage commitment (Laiosa et al., 2006).

The common myeloid progenitor (CMP) produces the following two lineages:

i) The myeloid lineage, that leads to the production of granulocytes and macrophages (GM lineage), is regulated mainly by the function of the ETS transcription factors PU.1 and C/EBPa (Wolff and Humeniuk, 2013) and secondarily by the transcription factors GATA1, GATA2, HOXB7, IRF-1 (Valledor et al., 1998).

ii) The erythroid/megakaryocyte lineage (MegE lineage) is regulated by the GATA1, KLF1 transcription factors (Hattangadi et al., 2011) for the erythroid lineage and by the GATA1, FLI1 for the megakaryocytic lineage (Starck et al., 2003).

The interplay between GATA1 and PU.1 is the main regulator of the myeloid lineage specification. High PU.1 and low GATA1 results in myeloid differentiation. Low PU.1 and high GATA1 leads to erythroid differentiation (Galloway et al., 2005, Rhodes et al., 2005).

In order the cells to differentiate along the erythroid lineage, the core erythroid set of genes needs to be upregulated with the simultaneous downregulation of the core myeloid set of genes. The regulation involves an interplay between transcription factors, chromatin modifiers and non-coding regulatory RNAs (miRNAs) (Hattangadi et al., 2011).

1.2 Erythroid differentiation

In mammals, hematopoietic stem cells (HSCs) that give rise to the early erythroid progenitors (primitive erythroblasts that express embryonic globin proteins), appear de novo from the hemangioblasts of the yolk sac (Figure 2A). In parallel, HSCs that give rise to definite erythroblasts (they express adult globin proteins) appear de novo from the hemogenic endothelium of the aorta-gonad mesonephros (AGM) region. Definite

erythroblasts also appear in the placenta, either through de novo creation of HSCs from the hemogenic endothelium or through colonisation from the AGM HSCs. Later in development, HSCs migrate in the fetal liver where they establish a population of rapidly proliferating pluripotent stem cells that can give rise to one or a few hematopoietic lineages. Just before birth, HSCs and progenitor cells begin to populate the bone marrow giving rise to a quiescence population of pluripotent stem cells (Cumano and Godin, 2007, Orkin and Zon, 2008, Laird et al., 2008). In the erythroid lineage, the first committed progenitors, that have characterised from ex-vivo studies, are the BFU-e cells (burst-forming unit-erythroid) that give rise to the CFU-e population (colony-forming unit-erythroid) (Elliott et al., 2008). The CFU-e cells are more rapidly proliferating than the BFU-e cells and they give rise to mature erythrocytes after undergoing a decrease in their cell size, haemoglobinization and enucleation (Figure 2B).

The control of the level of red blood cells (RBCs) is mediated through the kidney-derived cytokine erythropoietin (EPO). EPO stimulates the proliferation and the differentiation of the CFU-e. In case of stress erythropoiesis (chronic anemia), the CFU-e population is insufficient to support the maintenance of the RBCs levels and the BFU-e population increases (Dolznic et al., 2006). The BFU-e population responds to many hormones in addition to EPO including SCF (stem cell factor) (Perry et al., 2007), IGF-1 (insulin like growth factor 1) (Miyagawa et al., 2000), glucocorticoids (GCs) (Bauer et al., 1999), IL3 (Migliaccio et al., 2002), IL6.

Upon EPO stimulation, the CFU-e cells undergo two cell divisions in the first day and they up-regulate the transferrin receptor, the TER119 protein and other erythroid-important genes. On the second day, two or three rapid cell divisions occur with cells mainly in the S and M stage. The following days there is a complete repression of all gene transcription which is followed by chromatin condensation, nuclear condensation and enucleation (Hattangadi et al., 2011).

EPO regulates transcription by signalling mainly through STAT5 (Socolovsky et al., 2001) and secondary by the PI3K/AKT pathway (Ghaffari et al., 2006) and the RAS/MAPK pathway (Zhang and Lodish, 2007). STAT5 subsequently co-operates with the erythroid specific transcription factor GATA1 for the erythroid maturation (Wierenga et al., 2010). EPO also induces the phosphorylation of GATA1 at Ser₁₃₀ through the PI3K/AKT pathway (Kadri et al., 2005). This phosphorylation enhances the expression of the *TIMP-1* gene although the global role of the phosphorylation event is not well understood (Kadri et al., 2005).

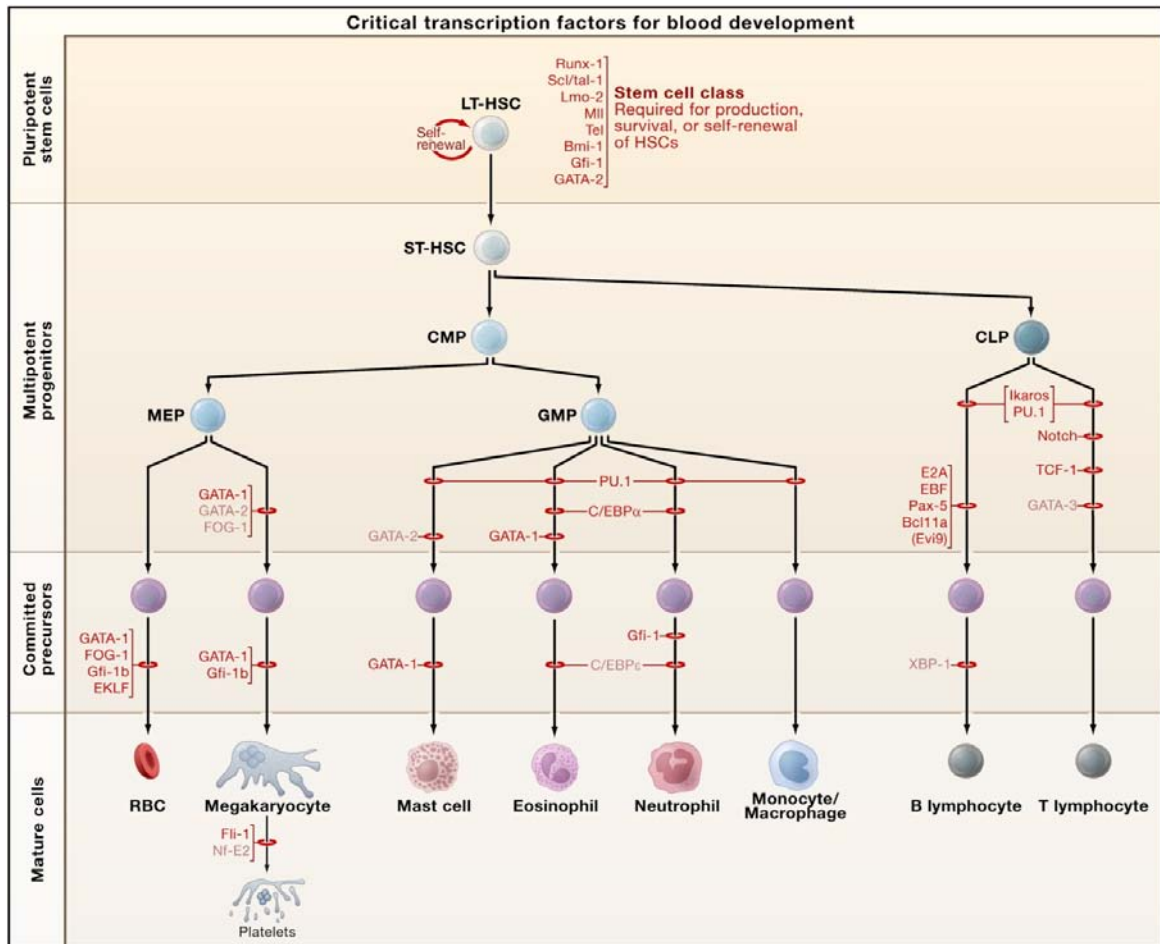


Figure 1. Hematopoietic developmental hierarchy. Progenitor, committed and differentiated cells are shown in the myeloid and the lymphoid lineage. The abbreviations are the following: LT-HSC (Long Term Hematopoietic Stem Cell), ST-HSC (Short Term Hematopoietic Stem Cell), MPP (Multipotent Progenitor), CMP (Common Myeloid Progenitor), MEP (Megakaryocyte/Erythroid Progenitor Cell), GMP (Granulocyte/Monocyte progenitor), CLP (Common Lymphoid Progenitor), RBC (Red Blood Cell). From (Orkin and Zon, 2008).

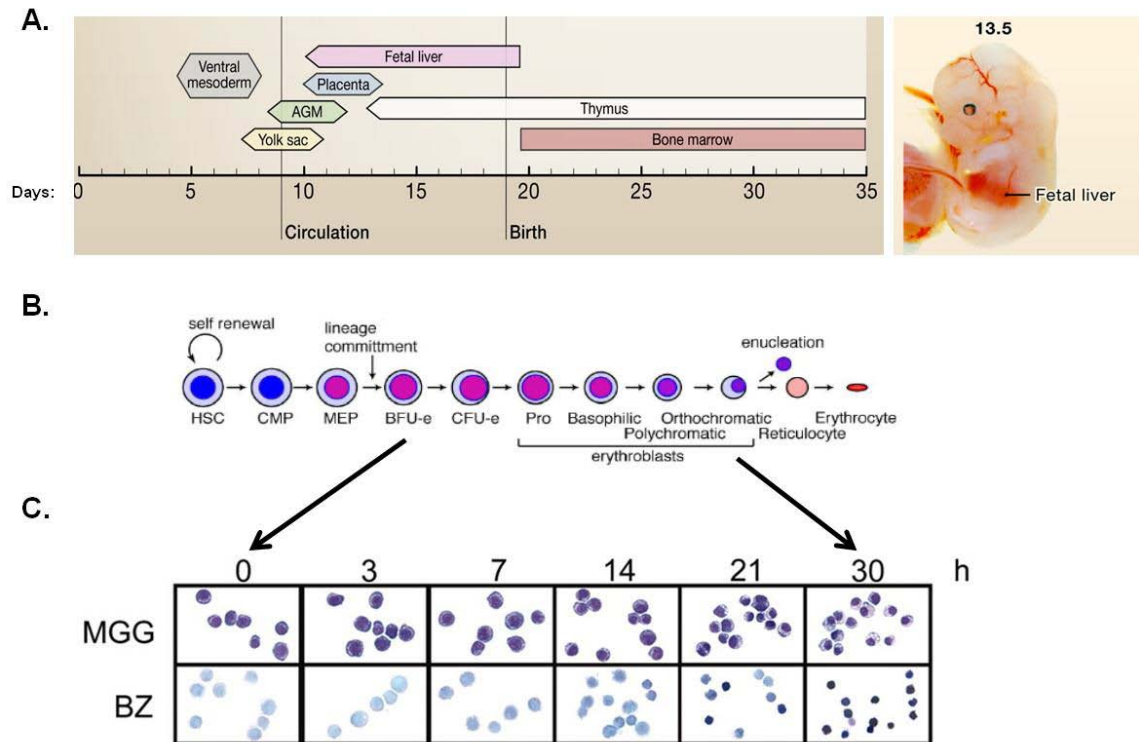


Figure 2. A) Hematopoiesis sites change during development. Adapted from (Orkin and Zon, 2008). Hematopoiesis progenitors first appear in the yolk sac blood islands. Then they appear sequentially at the aorta-gonad mesonephros region (AGM), placenta, and fetal liver. Eventually they appear in the thymus and the bone marrow. B) Different stages of erythroid maturation from the hematopoietic stem cell. The abbreviations are the following: HSC (Hematopoietic Stem Cell), CMP (Common Myeloid Progenitor), MEP (Megakaryocyte/Erythroid Progenitor), BFU-e (erythroid Burst-Forming Unit), CFU-e (erythroid Colony-Forming Unit), Pro (pro-erythroblasts). Adapted from (Wu et al., 2011). C) GATA1 mediated erythroid maturation in the G1E/G1E-ER4 model. The morphologic changes (MGG, May–Grunwald Giemsa stain) and accumulation of hemoglobin (BZ, benzidine stain for hemoglobin) are apparent after activation of GATA1. The time points indicate time after GATA1 induction with estradiol. From (Welch et al., 2004).

1.3 Transcription factors and transcriptional co-factors that control the differentiation of the erythroid lineage

1.3.1 GATA1

GATA1 is a member of the GATA family. The GATA family has 6 members (GATA1-6). All the proteins are developmental regulators. GATA1, GATA2 and GATA3 are used during the differentiation of the hematopoietic lineages whereas GATA4, GATA5 and GATA6 control the development of heart and lung (Brewer and Pizzey, 2006, Molkentin, 2000). GATA2 and GATA3 are also found in endothelial cells and in the central nervous system (Nardelli et al., 1999). GATA4, GATA5 and GATA6 might also have a functional role during muscle and gastrointestinal development (Jacobsen et al., 2002, Molkentin et al., 2000, Koutsourakis et al., 1999).

GATA1 is necessary for the differentiation of the erythroid and the megakaryocytic lineage. GATA1 is expressed in primitive and definitive erythroid cells (Fujiwara et al., 1996, Leonard et al., 1993), megakaryocytes (Romeo et al., 1990, Martin et al., 1990), eosinophils (Zon et al., 1993) and mast cells (Martin et al., 1990). GATA2 maintains the compartment of stem and progenitor cells (Tsai et al., 1994, Tsai and Orkin, 1997, Ling et al., 2004). GATA3 is required for the development and function of the T-helper-2 T-cells (Ho et al., 2009).

GATA1 regulates the proliferation, differentiation and apoptosis of the erythroid lineage cells. For example GATA1-deficient embryonic stem cells are able to contribute to all different tissues in chimeric mice except for the mature red blood cells because they die through apoptosis when they reach the pro-erythroblast stage (Pevny et al., 1995). GATA1 exerts its actions through cooperative regulation with other co-factors. These co-

factors include both DNA binding (KLF1, TAL1, PU.1) and non-DNA binding proteins (FOG1, RUNX1). Depending on the context and its associated co-factors, GATA1 can act as an activator or repressor of transcription.

In order to understand the GATA1 function, it is important to understand its structural organization as presented below.

1.3.1.1 Functional domains of GATA1

The DNA binding domains of the proteins of the GATA family have two C₂C₂ zinc fingers (the N-terminal Zn finger and the C-terminal Zn finger) (Figure 3A). The N and C termini of the proteins are divergent whereas the two zinc fingers are conserved across the six mammalian proteins (Bresnick et al., 2005). Additionally, immediately after the C-terminal Zn finger there is a highly conserved region that contains the nuclear localisation signal (NLS) and a region of basic amino acids where post translational modifications occur. This region of basic amino acids is also important for DNA binding and transcriptional activity (Crossley and Orkin, 1994, Hung et al., 1999). The C-terminal Zn finger of GATA1 mediates DNA binding in the “WGATAR” motif (Yang and Evans, 1992), whereas the N-terminal Zn finger stabilises the DNA interactions in some genomic locations (Trainor et al., 1996). The N-terminal Zn finger was shown to bind distinct motifs from the C-terminal Zn finger and specifically the “GATC” motif (Newton et al., 2001). Flanking residues of the zinc fingers also play role in the binding (Ghirlando and Trainor, 2003). The N-terminal and C-terminal Zn fingers of GATA1 share a homologous core region and therefore they show similar binding preference for the first three nucleotides of the “GATN” motif. The specificity in the recognition of the “GATC” motif from the N-terminal Zn finger versus the “GATA” motif is given by the C-terminal basic tail that follows the Zn core module (Ghirlando and Trainor, 2003). A point mutation in the N-terminal Zn finger (R216Q)

abolishes the DNA binding of the N-terminal Zn finger (Newton et al., 2001) and decreases the affinity of GATA1 for regions that have both “GATA” and “GATC” motifs (Yu et al., 2002) indicating that both the N-terminal Zn finger and the C-terminal Zn finger can bind simultaneously.

The zinc fingers themselves interact with other protein co-factors (Figure 3A). Proteins that have been reported to interact physically with GATA1 include the co-factor FOG1 (Tsang et al., 1997), histone acetyltransferase CBP/p300 (Blobel et al., 1998), the mediator complex subunit MED1 (Stumpf et al., 2006), the oncogene LMO2 (Osada et al., 1995), the transcription factors KLF1 (EKLF), Sp1 (Merika and Orkin, 1995) and PU.1 (Rekhtman et al., 1999). These interactions can involve only the N-finger, as is the case of FOG1 (Fox et al., 1998), or occur in combination with the C-finger, as is the case of Sp1, or involve only the C-terminal Zn finger, for example EKLF (Gregory et al., 1996, Merika and Orkin, 1995), CBP/p300 (Blobel et al., 1998) and GATA1 itself (Mackay et al., 1998). GATA1 has been shown to recruit the chromatin remodeler BRG1 which can mediate higher order looping (Kim et al., 2009c, Hu et al., 2011, Kim et al., 2009b).

Lack of the N-terminal 80 amino acids of the GATA1 protein (short isoform GATA1s) appears to inhibit the terminal erythroid differentiation (Shimizu et al., 2001). This truncated short form contributes to cases of acute megakaryoblastic leukaemia (AMKL) (Wechsler et al., 2002) and transient myeloproliferative disorders (TMD) (Xu et al., 2003, Greene et al., 2003) that are seen in patients with Down syndrome. Loss of interaction of GATA1 with RUNX1, which interacts in the N-terminal region of GATA1 (Elagib et al., 2003), might be responsible for this failure of megakaryocyte maturation although is not widely accepted (Xu et al., 2006).

The function of GATA1 is also regulated through post-translational modifications such as acetylation, phosphorylation and sumoylation (Bresnick et al., 2012). Acetylation from

CBP/p300 has been shown to increase the DNA-binding affinity of GATA1 either through increase in the intrinsic DNA binding affinity (Boyes et al., 1998a, Hung et al., 1999) or through inhibition of self association (Nishikawa et al., 2003) or through stabilization of the GATA1 complexes with other proteins (Lamonica et al., 2006). Phosphorylation seems to have a cell line specific role. After erythroid differentiation, phosphorylation of GATA1 increases its DNA binding activity in K562 cells. In MEL and BA/F3 cells, the erythroid differentiation does not have any effect in the GATA1 DNA binding activity because all the phosphorylation sites, except one, are already phosphorylated in the undifferentiated state (Partington and Patient, 1999, Yu et al., 2005). Sumoylation of GATA1, in the Lys 137 amino acid residue, is important for the control of genes from the GATA1/FOG1 complex through either a direct effect on the transcriptional activation of the gene or by provoking a different nuclear compartmentalization of the whole locus (Lee et al., 2009a) .

The interaction of GATA1 with a variety of proteins plays an important role in hematopoiesis, since it leads to the transcriptional activation or repression of GATA1 target genes. Work has been undertaken to sort out which of these interactions are functionally significant and determine the mechanisms by which they modulate the GATA1 activity.

Below we present some of the transcriptional co-factors of GATA1, the TAL1 protein, the FOG1 protein and the KLF1 protein (Figure 4).

1.3.2 TAL1

SCL/TAL1 is essential for the development of the HSCs (Porcher et al., 1996) and the differentiation of erythrocytes and megakaryocytes (Hall et al., 2003, Mikkola et al., 2003).

It binds the E-box (CANNTG) and is part of a pentameric complex that involves GATA1, the bHLH protein E2A, the LIM domain containing co-factors LMO2 and LDB1 (Cantor and Orkin, 2002). The GATA1 and the TAL1/E2A complex bind the DNA whereas LMO2 makes the bridge between the transcription factors, either as a single molecule or as a homodimer (Valge-Archer et al., 1994, Wilkinson-White et al., 2011). This pentameric complex can bind in composite sites that have E-box and GATA1 motifs with a 9-11 bp spacer between them (Wadman et al., 1997, Kassouf et al., 2010). However, these proteins can co-localise at certain chromatin sites lacking composite elements. The GATA1-LMO2-LDB1-E2A-TAL1 pentameric complex is mainly responsible for the activation of gene expression (Kassouf et al., 2010, Kassouf et al., 2008). SCL/TAL1 interacts also with co-repressor complexes like histone deacetylases (HDAC1, HDAC2), histone demethylases (LSD1), chromatin-remodelling factors (BRG1), the ETO-2, Mtgr1 and Gfi-1b proteins (Anantharaman et al., 2011). SCL/TAL1 interacts with the co-activators CBP/p300 and P/CAF (Anantharaman et al., 2011).

1.3.3 FOG1

The FOG (Friend of Gata) family of transcription co-factors consists of FOG1 and FOG2. FOG1 is expressed in erythroid and megakaryocytic cells, and is co-expressed with GATA1 during development (Tsang et al., 1997). The phenotype of FOG1 null mice closely resembles the GATA1 knockout phenotype. Mutant mice of both lineages die from anemia caused by a defect in primitive and definitive erythropoiesis. The FOG1 null mice have a complete elimination of the megakaryocytic lineage, revealing a GATA1-independent role of FOG1 in megakaryopoiesis (Tsang et al., 1998). GATA1 N-terminal Zn finger mutants (GATA1^{V205G}) that cannot interact with FOG1 are unable to rescue the maturation of a GATA1-deficient erythroid cell line. Thus GATA1 and FOG1 coordinate

erythroid cell maturation by activating erythroid-specific genes and repressing genes associated with the undifferentiated state (Crispino et al., 1999).

Other studies have shown that the GATA1/FOG1 interaction represses the GATA1 activation of genes which are important for the differentiation of mast cells and eosinophil cells, thus committing the cells towards a megakaryocyte/erythroid lineage. Exogenous expression of FOG1 in eosinophils leads to loss of eosinophil markers and the appearance of a multipotent phenotype (Querfurth et al., 2000). Exogenous expression of FOG1 in mast cell progenitors induces features of the erythrocytic and megakaryocytic lineage (Cantor et al., 2008). GATA1^{V205G} knock-in mouse embryos have an increased mast-cell potential, and aberrant expression of mast cell genes in erythroid and megakaryocytic cells (Cantor et al., 2008).

FOG1 also appears in T-cells where it associates with GATA3 in order to repress the T-helper-2 (Th2) cells development (Zhou et al., 2001). FOG2 is expressed in the heart, brain, liver and lungs. It can also interact with GATA containing proteins (Chlon and Crispino, 2012).

1.3.3.1 Functional domains of FOG1

FOG1 has nine zinc fingers. The zinc fingers comprise of two types: four C₂H₂ type zinc fingers (finger 2, 3, 4, 8) and five C₂HC type zinc fingers (finger 1, 5, 6, 7, 9) (Figure 3B). Fingers 2, 3, 4 have low-DNA binding affinity but the protein has not been shown to bind DNA without the presence of GATA1 (Pal et al., 2004a). Four fingers contact the GATA1 N-terminal Zn finger. FOG1 fingers 1, 6 and 9 have been suggested to interact strongly with GATA1 whereas finger 5 interacts weakly (Fox et al., 1999). FOG1 facilitates GATA1

chromatin occupancy (Letting et al., 2004, Pal et al., 2004a) and the GATA1-dependent chromatin looping (Vakoc et al., 2005). Also, FOG1 has an N-terminal domain that binds the NuRD chromatin-remodeling complex (Hong et al., 2005, Miccio et al., 2010) (Figure 3B). FOG1 also binds the co-repressor CtBP (Turner and Crossley, 1998, Katz et al., 2002) (Figure 3B), which associates with the histone demethylase LSD1 (Lysine Specific Demethylase 1). The FOG1 region between Zn fingers 6 and 7 mediates the contact. N-terminal truncated translational isoforms of FOG1 lack the NuRD domain, altering the complexes with which GATA1 interacts (Snow and Orkin, 2009). The ability of FOG1 to bind the NuRD complex is required for the GATA switch (DNA binding sites where GATA1 replaces GATA2 during differentiation) as mice expressing a knock-in FOG1, defective in NuRD complex binding, have impaired GATA switches (Gao et al., 2010).

FOG1 appears mainly in the hematopoietic cells where it interacts with GATA proteins but it has being also found in gonads, intestine and heart.

Mutations on the GATA1 N-terminal Zn finger (V205M, G208S but not R216Q, D218G) have been shown to inhibit the complexation with FOG1 (Liew et al., 2005)

1.3.3.2 Effect of GATA1/FOG1 complexation

GATA1 can activate (Crispino et al., 1999) and repress genes (Johnson et al., 2007b, Pal et al., 2004b) through FOG1 dependent and FOG1 independent mechanisms. GATA1, but not GATA1^{V205G}, can activate the globin and *Band3* (*Slc4a1*) genes. GATA1, but not GATA1^{V205G}, can repress the *GATA2* gene.

More specifically FOG1 has a direct effect on the GATA1 binding itself. It has been shown that in the absence of FOG1, GATA1 itself cannot out compete GATA2 binding in places where the GATA1/GATA2 switch is necessary (Pal et al., 2004a). This was shown for the *GATA2* gene where GATA1 needs to bind and repress the gene during erythroid differentiation. GATA1 substitutes GATA2 at the -2.8kb loci upstream of the gene (Pal et al., 2004a).

This implies that the complexation changes the binding specificities of GATA1. Additionally, complexation with FOG1 is necessary in order to establish a favourable binding environment through histone acetylation. Complexation with FOG1 can increase the acetylation of the H3 histone at the K9 position (H3K9ac) on the binding site of the complex. This has been shown to be necessary for the formation of the HS2 on the beta globin locus (Letting et al., 2004). Lack of acetylation causes lack of the HS2 site. Thus the unique chromatin environment might be used as a recognition signature for GATA1 binding or GATA1 can modulate the chromatin surrounding itself something that might help to recruit other factors on the locus. On the other hand, GATA1 binds on the hypersensitive site HS3 of the beta globin locus in a FOG1 independent manner (Letting et al., 2004). Nevertheless, in the *GATA2* locus, the binding of the GATA1/FOG1 complex decreases the acetylated H3 histone (Pal et al., 2004a, Letting et al., 2004).

FOG1 interacts with the chromatin remodelling complex, the NuRD complex, which can mediate both repression and activation of specific targets in a context dependent manner (Miccio et al., 2010, Hong et al., 2005, Gregory et al., 2010, Gao et al., 2010). NuRD is a multisubunit complex that includes the HDAC1 and HDAC2 deacetylases and the Mi2 α/β ATPase proteins which leads to alteration of both the histone location and acetylation (Ramirez and Hagman, 2009). FOG1 mutations that abolish the FOG1/NuRD interaction do not permit the FOG1/GATA1 complex to repress *GATA2* and mast cell specific genes (Hong et al., 2005). On the contrary, although the FOG1/GATA1 complexation is retained

there is an induction of gene expression (Gregory et al., 2010). Interaction of FOG1 with CtBP mediates FOG1-mediated *GATA1* repression in other loci (Fox et al., 1999, Hong et al., 2005).

GATA1 requires FOG1 to create higher order chromatin loops at the beta globin locus (Kim et al., 2007b, Vakoc et al., 2005, Deng et al., 2012) and the *C-Kit* promoter (Jing et al., 2008). Specifically wild type *GATA1*, but not *GATA1*^{V205G}, is able to activate the beta globin (*HBB*) gene by bringing in close proximity the LCR and the promoter of the *HBB* gene. Similarly the *GATA1*/FOG1 interaction is necessary to down regulate the *C-Kit* gene by bringing together the promoter and the upstream enhancer element. In the latter case, the complex must first remove *GATA2* and then induce the looping.

1.3.4 KLF1

KLF1 expression is restricted in erythroid cells and their precursors. The binding domain of KLF1 consists of 3 C-terminal C₂H₂ zinc fingers (Miller and Bieker, 1993) (Figure 3C). Although there is no available structure of a KLF1-DNA complex, similarities with the SP1 protein predicts that each zinc finger binds to three nucleotides on the G-rich strand of the motif (Oka et al., 2004). The N-terminus of KLF1 has been shown to interact with co-activators such as p300, CBP, P/CAF, TFIIF (Mas et al., 2011, Zhang and Bieker, 1998) or co-repressors such as the NuRD complex (Siatecka et al., 2007) (Figure 3C). Similar to *GATA1*, KLF1 acetylation enhances its transcriptional activity at least in the beta globin (*HBB*) promoter (Zhang and Bieker, 1998) by enhancing its interaction with the SWI/SNF complex but not by increasing its DNA binding affinity (Zhang et al., 2001). The Zn fingers can interact with co-repressors like Sin3A and HDAC1 (Chen and Bieker, 2004, Chen and Bieker, 2001). The co-repressor function is speculated to be necessary in the early

hematopoietic cells (Chen and Bieker, 2004). KLF1 and FLI1 cross-antagonize for binding to GATA1 and the subsequent differentiation towards an erythroid or a megakaryocytic lineage respectively (Starck et al., 2003).

KLF1 null mice are embryonic lethal due to a defect in definitive erythropoiesis (Perkins et al., 1995, Nuez et al., 1995). These embryos show a defect in hemoglobin expression and they die from anemia. GATA1 binds the zinc-finger domain of KLF1 mainly via its C-terminal Zn finger (Merika and Orkin, 1995). This interaction is hypothesized to occur upon DNA binding of both transcription factors as they have not been co-immunoprecipitated from cells in the absence of DNA (Tallack et al., 2010). Reporter assays demonstrated that GATA1 transcriptional activity is increased by these interactions (Merika and Orkin, 1995). It has also been described a significant overlap between GATA1 peaks and KLF1 peaks (Tallack et al., 2010). Interaction between GATA1 and KLF1 may play a crucial role in bringing regulatory elements such as enhancers and LCRs in close proximity to promoters by promoting the formation of DNA loops (Drissen et al., 2005, Tallack et al., 2010). It has also been suggested that a KLF1-GATA1-TAL1 transcription factor network efficiently recruits p300 in the overlapped positions (Tallack et al., 2012) and along with the greater GATA1 pentameric complex (GATA1/E2A/{LMO2}₂/LDB1/TAL1) orchestrate the erythroid response (Soler et al., 2010).

1.3.4.1 Effect of KLF1 on gene regulation in the erythroid lineage

KLF1 has been shown to act as both activator and repressor (Tallack et al., 2012, Feng et al., 1994, Chen and Bieker, 2001, Drissen et al., 2005, Hodge et al., 2006, Siatecka et al., 2007, Chen and Bieker, 2004). KLF1 null mice have been used to find genes regulated by KLF1. Because the livers of KLF1 null mouse embryos do not have erythroblasts, the

gene regulation was only shown for the erythroid progenitors. KLF1 has been shown to be directly or indirectly involved in the upregulation of 690 genes and in the downregulation of 118 genes in mouse liver erythroid progenitors (Tallack et al., 2012). Only one of these repressed genes had KLF1 near the promoter (less than 1000bp). The rest of the repressed genes might be regulated by other KLFs such as KLF3, KLF8 and KLF12 (Tallack et al., 2012). Thus on the erythroid progenitors, KLF1 functions mainly as activator.

Wild type mouse embryo liver erythroblasts show 2447 downregulated genes and 633 upregulated ones in comparison with the mouse embryo liver erythroid progenitors. The KLF1 genomic positions are not the same in the erythroid progenitors and in the erythroblasts. To find the potential effect of KLF1 on the nearby genes during the erythroid differentiation, peaks that appear only in erythroblasts have been used. KLF1 is bound in the promoter (less than 10kb from the TSS) of 692 of the downregulated genes in erythroblasts and 246 of the upregulated ones (Pilon et al., 2011). Thus KLF1 acts as a repressor during the differentiation from erythroid progenitors to erythroblasts. This indicates that EKLF can act either as an activator or repressor depending on its association with other proteins (Chen and Bieker, 2004).

It has long been recognized that KLF1 itself is involved in the three dimensional organization of a gene-locus (Drissen et al., 2004, Schoenfelder et al., 2010). KLF1 was shown to be important to recruit the beta globin genes to the active chromatin hub substructure (ACH), a term used to describe the spatial clustering of the regulatory elements of the beta-globin locus, namely the hypersensitive sites (HS1–HS6) of the locus control region (LCR), the 5' upstream of the locus control region hypersensitive sites HS-60, HS-62 and the 3' downstream of the locus control region hypersensitive site HS1 (Palstra et al., 2003). It has also been shown that KLF1 was able to recruit KLF1 regulated genes in a limited number of transcription factor foci (Schoenfelder et al., 2010). It is

important to point out that in contrast to the GATA1 effect on the transcriptional status of the genes, KLF1 regulated genes are not dependent on KLF1 for expression. On the contrary KLF1 causes an increase in their expression during the maturation of the erythroid cell lineage (Drissen et al., 2005, Hodge et al., 2006, Schoenfelder et al., 2010). Thus, KLF1 is likely to function as a co-activator.

1.3.5 Additional GATA1 complexes

Some other prominent GATA1 interactions with other complexes are the following: GATA1 and PU.1 antagonize each other via direct physical interaction (Nerlov et al., 2000). The DNA-binding domains of both proteins are involved (Rekhtman et al., 1999), but the mechanisms by which these transcription factors antagonize each other are different. GATA1 prevents the PU.1 binding to its cofactor c-Jun (Zhang et al., 1999), while PU.1 blocks GATA1 DNA binding (Zhang et al., 2000).

The GATA1 N-finger has also been found to bind directly the Mediator complex through the MED1/TRAP220 subunit (Stumpf et al., 2006). Mediator is a large multi subunit complex (consisting of 22-28 subunits) that binds the Polymerase II and regulates its transcriptional activity. In multiple systems, it has been shown to be involved at least partially in bridging the interaction between upstream transcription factors and general transcription initiation factors (Kuras et al., 2003, Kim and Lis, 2005, Blazek et al., 2005). Although this implies that GATA1 might have a direct effect on gene transcription by recruitment of RNA polymerase II, only a restricted set of genes has been shown to be affected by this recruitment (Pope and Bresnick, 2013, Pope and Bresnick, 2010).

It has also been noted that the transcriptional outcome in multiple loci can be regulated from multiple co-factors albeit with different degrees of redundancy (Pope and Bresnick, 2010).

1.4 Epigenetic changes in chromatin during erythroid differentiation

Although post translational modifications of histones are usually the result of gene regulation, they give insight into the mechanisms that control gene regulation. Some patterns have emerged from the studies that have been performed either genome wide or by interrogating specific loci.

As with embryonic stem cells, bivalent genes that have both activating (H3K4me3) and repressor marks (H3K27me3), in their promoters, have been found (Cui et al., 2009). In a model of differentiation from CD133+ progenitors to CD36+ pro-erythroblasts, 19% of the 2910 bivalent genes lost the H3K27me3 mark during differentiation. Simultaneously the genes had H3K4me1, H3K9me1 and poised RNA polymerase II which indicates that the genes were primed for activation upon loss of the repressor marker (Cui et al., 2009).

Another subset of genes in multipotential hematopoietic stem cells showed enrichment for the H3K4me2 but not the H3K4me3 markers. These markers usually co-exist. This set of genes is enriched with lineage-specific hematopoietic genes. Upon erythroid differentiation, the erythroid specific genes from this subset gained H3K4me3 and were induced whereas non-erythroid genes lost the H3K4me2 and were repressed. Thus the H3K4me2 mark, at least in this subset, acts as a bivalent switch (Orford et al., 2008).

Upon reconstitution of GATA1 in a GATA1 null erythroblast cell line no significant change in the H3K4me3 and H3K27me3 status of genes has been noticed. On the other hand, RNA polymerase II elongation marks, such as H3K16ac and H3K79me2 are better classifiers of the change in gene expression during erythroid differentiation (Wu et al., 2011).

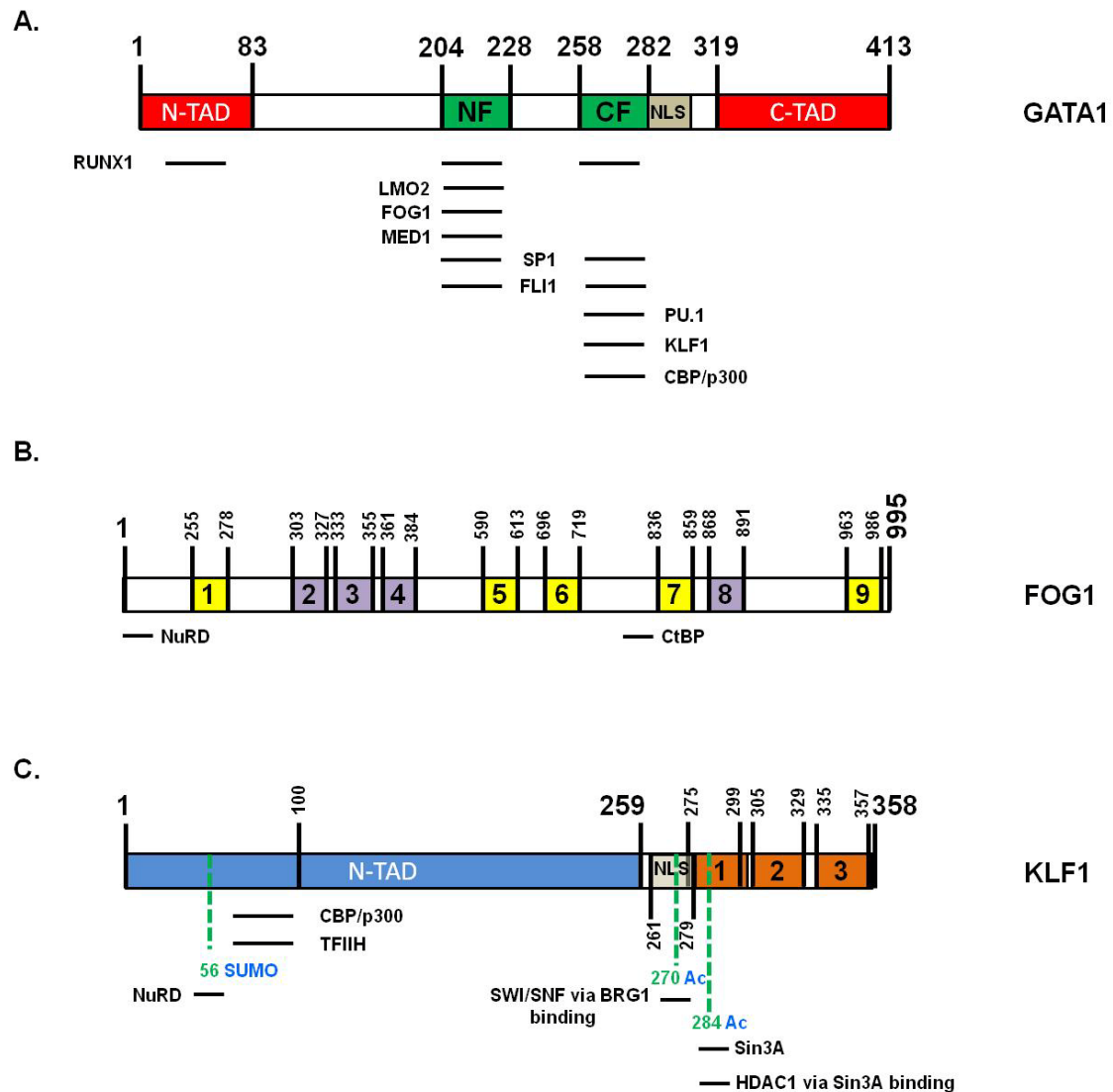
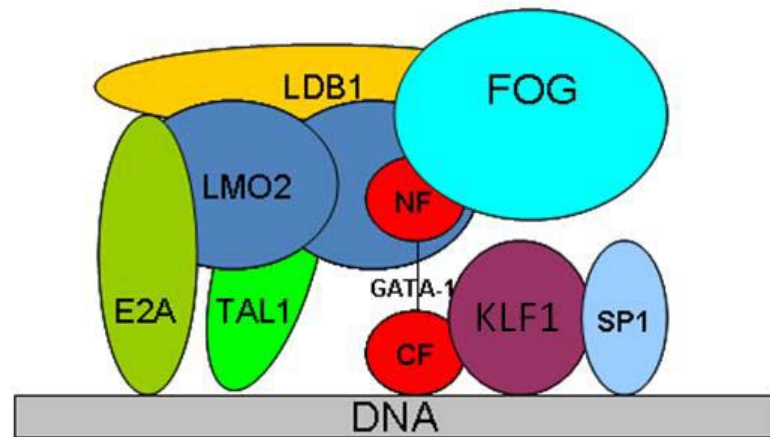


Figure 3. A) Structural organisation of GATA1. The two C_2C_2 Zn fingers are presented as green coloured boxes. The NF represents the N-terminal Zn finger, the CF represents the C-terminal Zn finger. The transactivation domains are presented as red coloured boxes. The N-TAD represents the N-terminal Transactivation Domain (acidic transactivation domain), the C-TAD represents the C-terminal Transactivation Domain (proline-rich transactivation domain) (Kaneko et al., 2012). The interaction regions of GATA1 with other proteins are shown with black horizontal bars along with the respective protein name (see section 1.3.1 and 3.5.7). The brown box represents a region with a high frequency of basic amino acids where the nuclear localization signal (NLS) lies. The picture is adapted from (Kim and Bresnick, 2007). B) Structural organisation of FOG1. The C_2HC Zn fingers

are present as purple coloured boxes. The C₂H₂ Zn fingers are present as yellow coloured boxes. The interaction regions of FOG1 with other proteins are shown with black horizontal bars along with the respective protein name (see section 1.3.3). The picture is adapted from (Chlon and Crispino, 2012). C) Structural organisation of KLF1. The three C₂H₂ Zn fingers are presented as orange coloured boxes. The N-terminal Transactivation Domain (N-TAD) (proline-rich transactivation domain) is presented as a blue coloured box. The light brown box represents a region with a high frequency of basic amino acids where the nuclear localization signal (NLS) lies. The vertical green dotted lines represent amino acids with post translational modifications (blue letters), namely sumoylation (SUMO) and acetylation (Ac). The interaction regions of KLF1 with other proteins are shown with black horizontal bars along with the respective protein name (see section 1.3.4). Regions where the black horizontal bars cross amino acids with post translational modifications, indicate that the co-factors need the post-translational modification in order to be able to bind. The picture is adapted from (Yien and Bieker, 2013). The numbers represent the position of the first and last amino-acid in each domain.

A.



B.

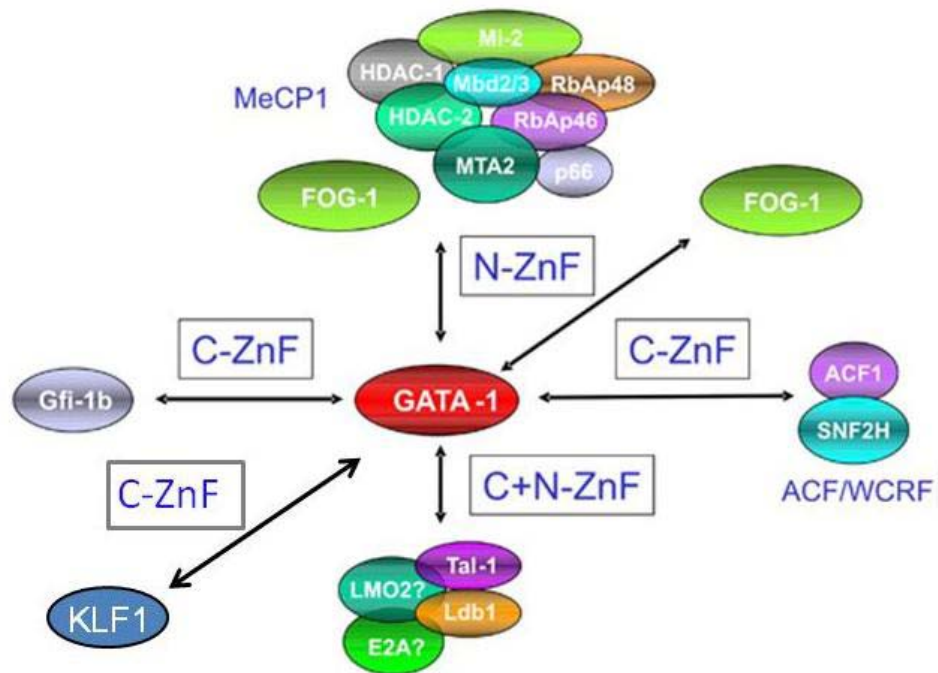


Figure 4. Multiprotein complexes of GATA1. Interactions of GATA1 with LMO2, LDB1, E2A, TAL1, FOG1, EKLF (KLF1), SP1 are presented (A). Not all proteins exist in the same complex at the same time (B). The N-ZnF indicates interaction with the N-terminal Zn finger of GATA1. The C-ZnF indicates interaction with the C-terminal Zn finger of GATA1. (A) adapted from (Wilkinson-White et al., 2011). (B) from <http://www.erasmusmc.nl/cellbiology/research/research-groups/grosveld/1410123/>.

1.5 Models of erythropoiesis

Below we present some of the models that have been used to understand the differentiation process from an erythroid progenitor to a late erythroblast:

1.5.1 Isolation of fetal liver cells from E14.5 mice

The majority of liver cells in this developmental stage are erythroid in different stages of maturation. These cells can be grouped in five populations based on the amount of two proteins: glycophorine A (TER119) and the transferrin receptor (CD71) (Figure 5). These subpopulations are the following (Zhang et al., 2003):

1. R1 (CD71^{med}TER119^{low}) (primitive progenitor cells and proerythroblasts, benzidine negative)
2. R2 (CD71^{high}TER119^{low}) (proerythroblasts and early basophilic erythroblasts, benzidine negative)
3. R3 (CD71^{high}TER119^{high}) (early and late basophilic erythroblasts, mixture of weak benzidine positive and benzidine negative)
4. R4 (CD71^{med}TER119^{high}) (chromatophilic and orthochromatophilic erythroblasts, all benzidine positive)
5. R5 (CD71^{low}TER119^{high}) (late orthochromatophilic erythroblasts and reticulocytes, all benzidine positive)

The TER119⁻ erythroid progenitors (CFU-e, BFU-e enriched populations) (R1, R2 populations) and TER 119⁺ erythroblasts (R3, R4, R5 populations) show low and high amount of GATA1 respectively.

1.5.2 Differentiation of a mouse erythroleukaemia cell line (MEL cells)

The spleen has a supportive role in erythropoiesis especially under conditions that reduce the oxygen to tissues. In normal mice, it contributes 10% in erythropoiesis. In phlebotomized mice this contribution increases to 40%. Upon inoculation of susceptible mice with the Friend virus complex, the mice develop polyclonal proliferation of splenic erythroblasts in two stages that will eventually cause erythroleukemia. These clones demonstrate p53 inactivation, activation of the ETS family transcription factor PU.1 and constitutive activation of the EpoR (Ney and D'Andrea, 2000). These cells are blocked to the erythroblast stage and they can undergo terminal erythroid differentiation upon DMSO induction (Friend et al., 1971). During the differentiation process the cells exit from the cell cycle and they show a 100-fold increase of the adult β -major globin gene transcription and haemoglobin synthesis (Scher et al., 1973, Sawado et al., 2001).

1.5.3 Differentiation of GATA1^{-/-} mouse ES cells upon GATA1 restoration

This model was created from GATA1 null mouse ES cells (G1E) that were cultured under conditions that favour the development of erythroid cell colonies (Weiss et al., 1997). Because GATA1^{-/-} erythroid precursors arrest at the proerythroblast stage and undergo apoptosis (Weiss and Orkin, 1995, Weiss et al., 1994, Pevny et al., 1995), a human *Bcl-2* cDNA was introduced to decouple developmental arrest from apoptosis but the cells eventually became BCL2 independent. The G1E cells resemble immature erythroblasts (Welch et al., 2004) (BFU-e, CFU-e). Upon GATA1 restoration and in the presence of erythropoietin, the cells undergo terminal erythroid maturation (late orthochromatic erythroblasts) (Figure 2C). GATA1 restoration is achieved either through lentiviral expression or through an estradiol inducible translocation of GATA1-ER to the nucleus in a GATA1 stable G1E cell line (G1E-ER4 cells). It should be noted that the G1E cells upon

GATA1 restoration and in the presence of TPO (thrombopoietin) and IL3, the cells differentiate towards megakaryocytes.

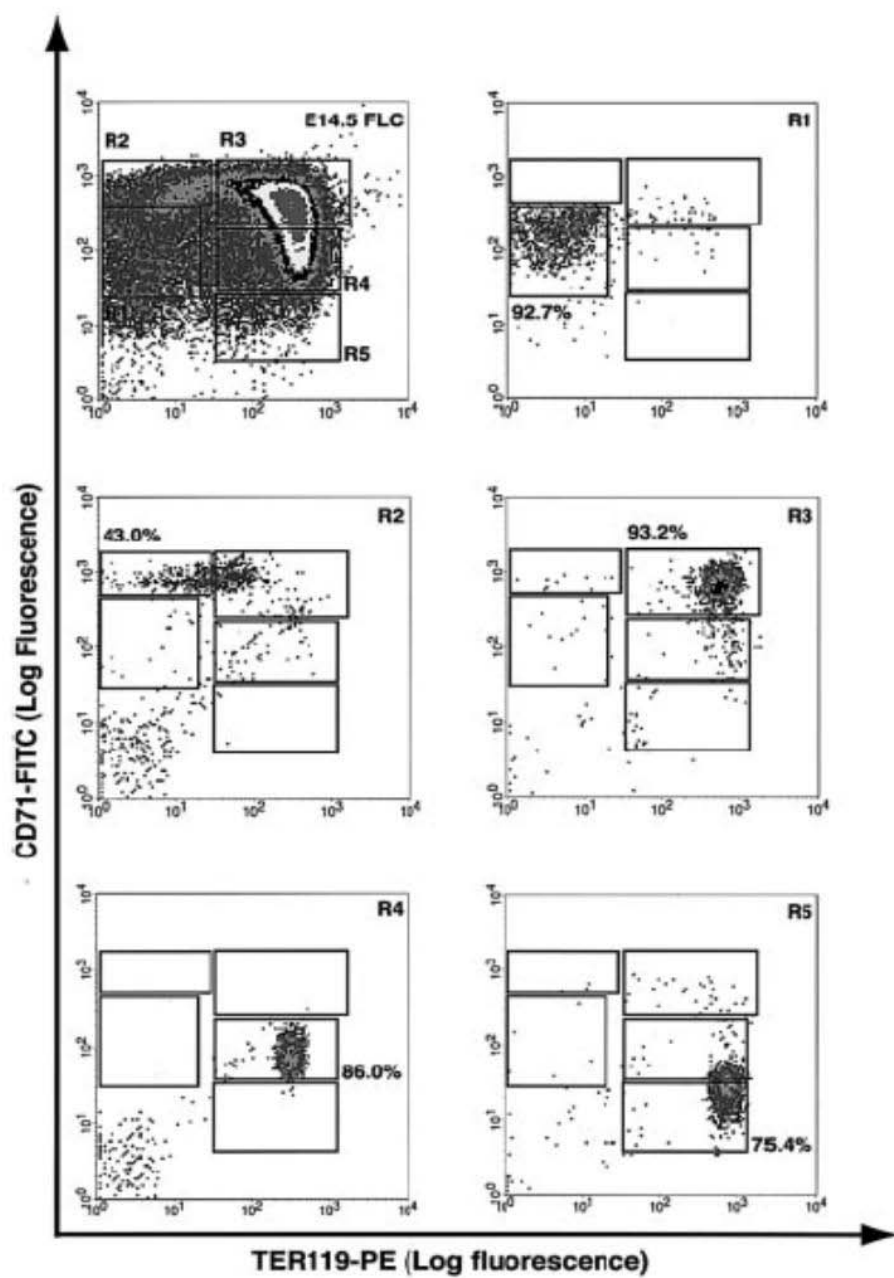
1.5.4 Differentiation of human CD34+ cells from peripheral blood

Human CD34+ cells from peripheral blood which represent hematopoietic progenitor cells can be differentiated into pro-erythroblasts (CD36+) using a cocktail of different cytokines including EPO (erythropoietin) ex vivo. The CD34+ cells in the presence of TPO (thrombopoietin) can be differentiated towards the megakaryocytic pathway.

1.5.5 Differentiation of human K562 cells

K562 cells is a multipotential hematopoietic cell line derived from a patient with chronic myelogenous leukaemia (CML). Erythroid differentiation of K562 cells is induced with hemin (Sutherland et al., 1986). K562 are also able to differentiate in megakaryocytes (Sutherland et al., 1986).

A.



B.

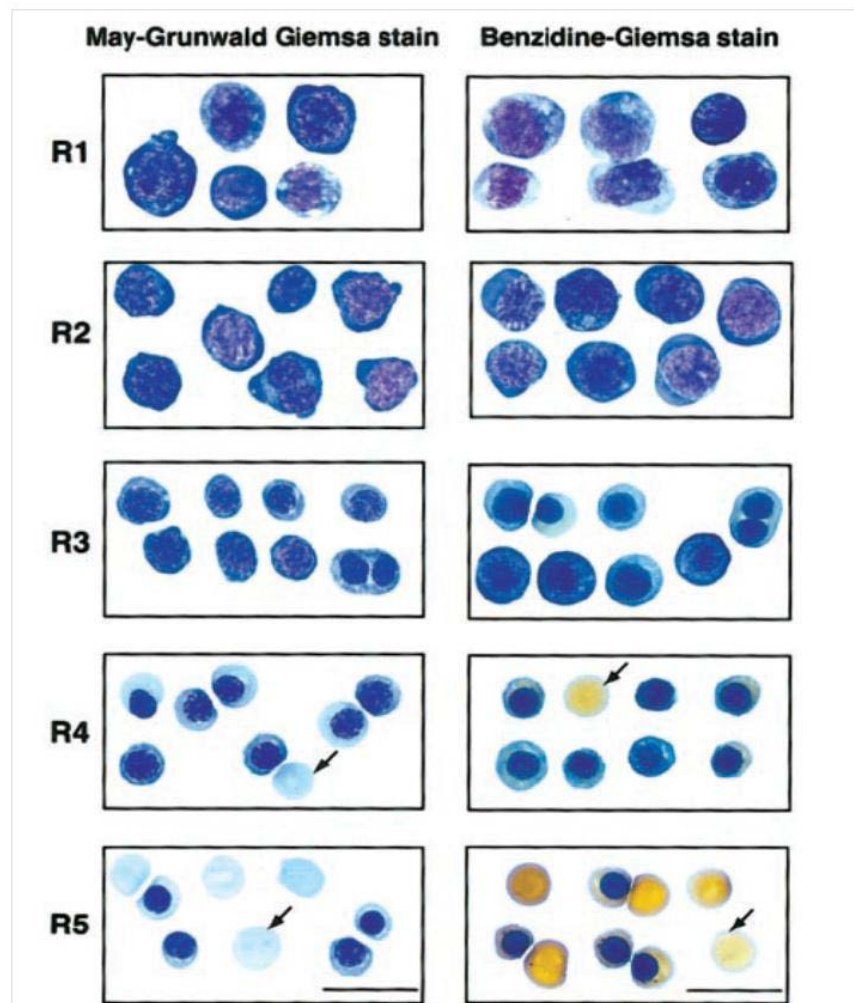


Figure 5 (previous and current page). A) Population of cells from the mouse fetal liver of E14.5. The cells were labeled with an antibody against CD71 (transferrin receptor) and TER119 (glycophorine A) (axes indicate relative logarithmic fluorescence units) and FACS sorted into 5 distinct populations (R1 to R5, see text for more details). The gating results are indicated. B) The morphologic changes (May-Grunwald Giemsa stain) and accumulation of hemoglobin (BZ, benzidine stain for hemoglobin) in the R1 to R5 populations is shown. From (Zhang et al., 2003).

1.6 Experimental strategies to investigate transcription factor-DNA interactions

Transcription factor (TF)-DNA interaction studies involve the characterisation of the DNA sequences that the protein prefers, the genomic location of the protein under specific conditions and its effect on gene expression. These studies involve in vitro assays like Electrophoretic Mobility Shift Assays (EMSA), in vivo assays like Chromatin Immunoprecipitation (ChIP) and reporter assays of specific promoter/enhancer sequences.

The classical in vitro approaches lacked the high-throughput requirements for a systematic and comprehensive characterization of TF-DNA interactions. The comprehensive characterization of TF binding profiles only became feasible with the advent of microarray and high-throughput sequencing based methods (Jolma and Taipale, 2011) (Figure 6).

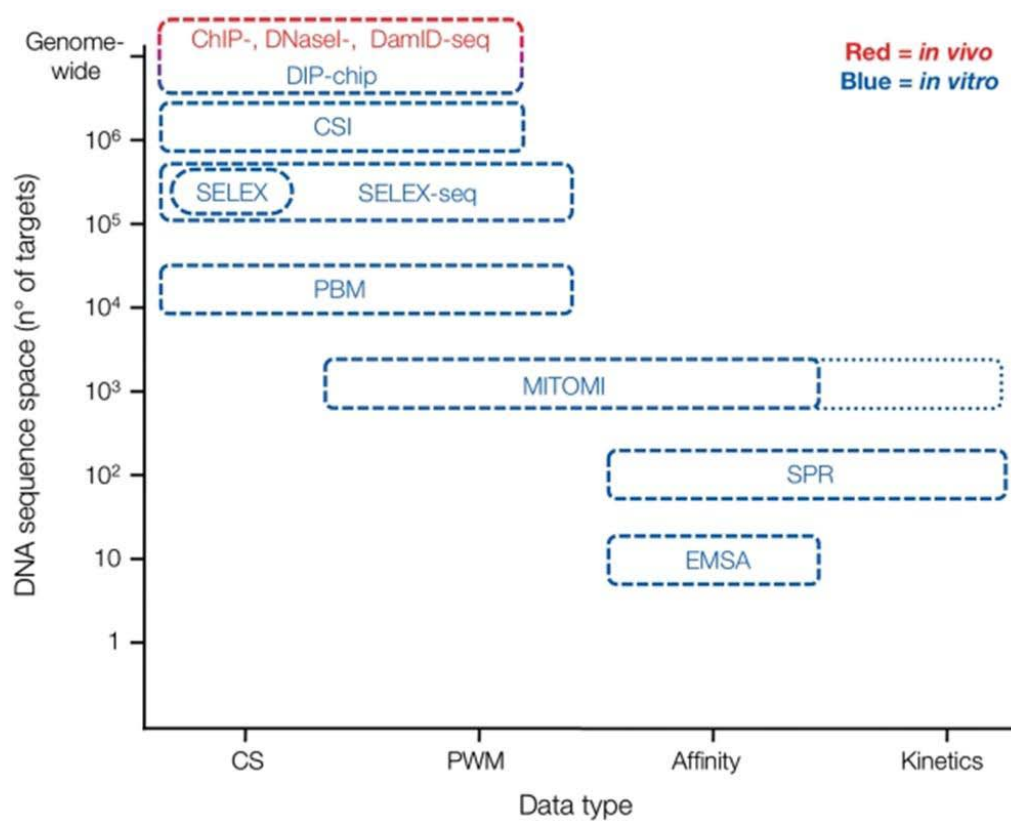
1.6.1 High-throughput methods to characterize protein-DNA interactions in vitro

1.6.1.1 Sequencing based methods to characterize protein-DNA interactions

Systematic Evolution of Ligands by Exponential Enrichment (SELEX) was the first experimental technique that could determine de novo TF binding sites. Purified TFs are added to the library of DNA sites and the bound and unbound sequences are separated by various biochemical procedures. Due to the limitations of conventional sequencing, the original approaches relied on a high number of selection cycles that resulted in the selection of only a **few high-affinity binding sites**, something that was not capturing

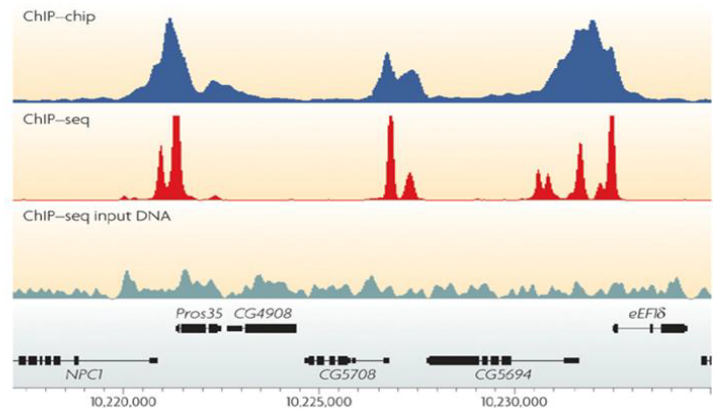
comprehensively the complex relationship between sequence composition and binding affinity for a TF (Gerland et al., 2002, Zhao et al., 2009). To overcome this limitation, in vitro selection was coupled with high throughput sequencing technologies. The first method SELEX-SAGE (Roulet et al., 2002) used the SAGE (Serial Analysis of Gene Expression) technique to sequence the selected DNA sites. The development of next-generation sequencing platforms saw the use of these technologies for sequencing all the bound DNA sites. Six methodologies based on similar principles were currently presented Bind-n-Seq (Zykovich et al., 2009), SELEX-seq (Jolma et al., 2010, Slattery et al., 2011), HT-seq (Zhao et al., 2009), EMSA-seq (Wong et al., 2011a), PB-seq (Guertin et al., 2012), HiTS-FLIP (Nutiu et al., 2011) (Figure 6). The advantage of these techniques is that there is no inherent limit to the length of the binding site that can be investigated. SELEX-seq involves multiple rounds of binding and amplification and in each round the bound DNA is sequenced. The binding affinities of the protein are calculated by taking into account all the rounds and giving major emphasis in the last rounds. On the contrary in EMSA-seq (Figure 7) only the first round of selection is used to extract relative binding affinities for a vast number of DNA sequences. These relative affinities are represented as fold enrichment of sequenced DNA fragments versus the non-selected pool. Although higher affinity sites have a higher probability of being bound by the TF, after a single selection most of the bound sequences are still medium to low affinity because they greatly exceed the number of high-affinity sequences. PB-seq conceptually is similar to EMSA-seq but the protein is incubated with naked, sheared genomic DNA.

A.

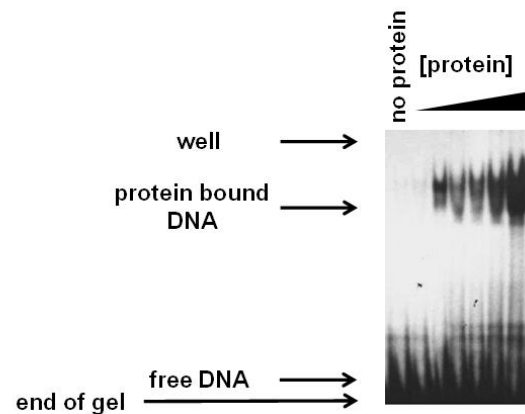


B.

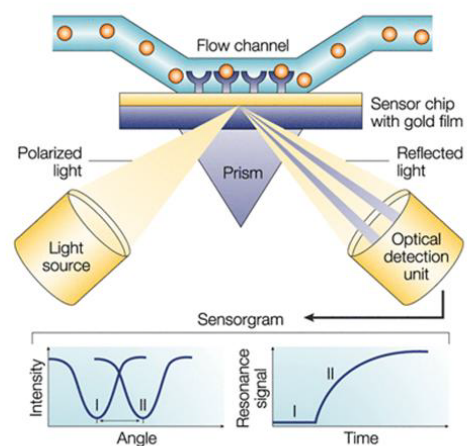
Chromatin
Immunoprecipitation combined
with microarrays (ChIP-chip) or
sequencing (ChIP-seq)



Electrophoretic Mobility
Shift Assay
(EMSA)



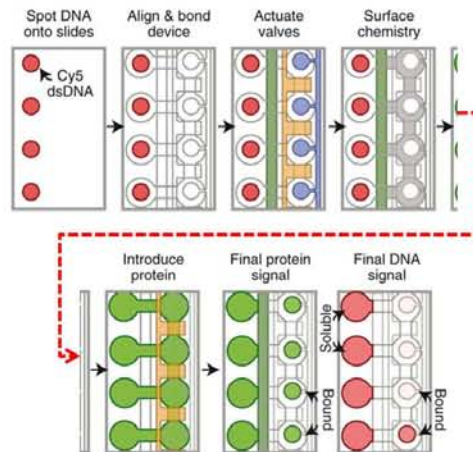
Surface Plasmon Resonance
(SPR)



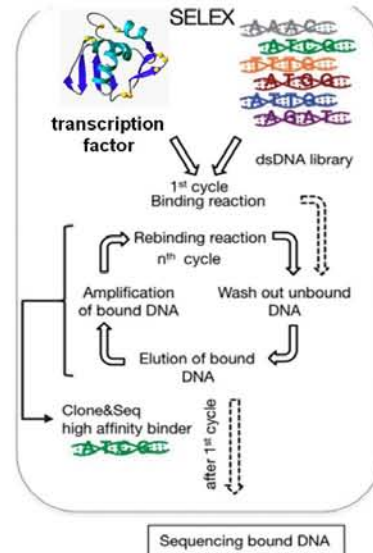
(continued on the next page)

(continued from previous page)

Mechanically Induced Trapping of Molecular Interactions (MITOMI)



Systematic Evolution of Ligands by Exponential Enrichment (SELEX/SELEX-seq)



Protein Binding Microarrays (PBM)

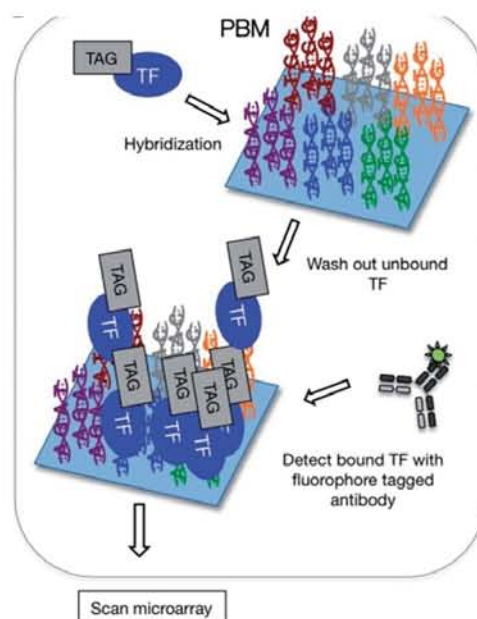


Figure 6 (previous three pages). Experimental techniques to study protein DNA interactions (A). The techniques illustrated in (A) are: Electrophoretic Mobility Shift Assay (EMSA) (in (B) picture adapted from (Kim et al., 2007a)), Surface Plasmon Resonance (SPR) (in (B) picture from (Cooper, 2002)), Mechanically Induced Trapping of Molecular Interactions (MITOMI) (in (B) picture from (Fordyce et al., 2010)), Protein Binding Microarrays (PBM) (in (B) picture from (Geertz and Maerkl, 2010)), Systematic Evolution of Ligands by Exponential Enrichment (SELEX/SELEX-seq) (in (B) picture from (Geertz and Maerkl, 2010)), Cognate Site Identification (CSI) (Warren et al., 2006), Chromatin Immunoprecipitation (ChIP) (in (B) picture from (Park, 2009)), DNA adenine methyltransferase identification (DamID), DNA Immunoprecipitation (DIP), DNase I assays. The abbreviations of x-axis in (A) are the following: CS (Consensus sequence), PWM (Position Weight Matrix). Adapted from (Geertz and Maerkl, 2010).

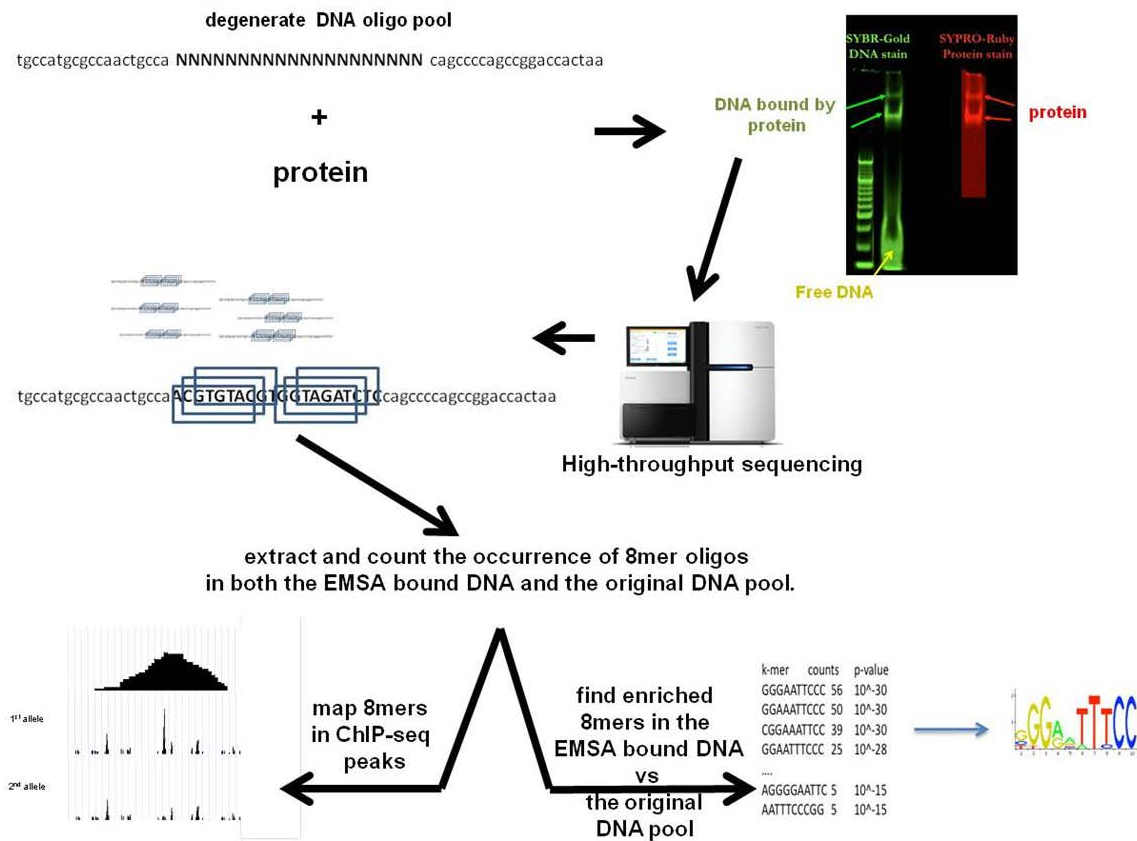


Figure 7. Outline of the EMSA-seq approach. DNA isolated from one round of EMSA experiments is sequenced with next generation sequencing platforms. K-mers (for example 8mers) are deconvoluted from the bound and the initial degenerate DNA oligo pool. Enriched k-mers are used to create DNA binding models for the selected transcription factor. Enriched k-mers are also projected in vivo in ChIP-seq binding sites. There is no restriction in the length of the degenerate region of the DNA library (usually 20 bases). Adapted from (Wong et al., 2011a).

1.6.1.2 Microarray based methods to characterize protein-DNA interactions

DNA microarray chips preceded the development of high throughput sequencing. Here the binding reactions were performed on immobilized double-stranded DNA oligonucleotide arrays (Protein Binding Microarrays, PBM) (Warren et al., 2006, Berger et al., 2006, Bulyk et al., 2001, Carlson et al., 2010). The protein of interest is allowed to bind to a PBM and the binding events are detected with fluorophore coupled antibodies against the protein of interest (Figure 6). Signal intensities are analyzed and interpreted as differential binding profiles. The early limitation of this technology was the number of different DNA molecules that could be spotted on the glass. Bioinformatic approaches were developed to permit the incorporation of the maximum number of overlapping DNA oligos of interest in the least amount of sequence space (de-Bruijn approach) (Mukherjee et al., 2004). Recent advances permitted the fabrication of high-density arrays, having all permutations of a 10-mer sequence. Using such PBMs a broad range of TFs was comprehensively examined (Badis et al., 2009).

1.6.1.3 Additional platforms to characterize protein-DNA interactions

Surface Plasmon Resonance (Paul et al., 2009) offers a low throughput investigation of the interaction of protein-DNA complexes. It can be made relatively high-throughput by using surfaces with arrays of different DNA molecules (Shumaker-Parry et al., 2004). The main advantage of this technique is that one can measure accurate TF-binding kinetics, specifically association and dissociation constants. Recently, a microfluidics based platform has been presented, which employs a detection method based on the trapping of TF-DNA interactions in chambers (MITOMI) (Fordyce et al., 2010). This approach is also able to give accurate TF-DNA dissociation constants in the nanomolar scale.

1.6.2 Methods to characterize protein-DNA interactions in vivo

In vitro approaches usually give the binding specificities of a protein against a specific DNA sequence. In contrast in vivo techniques, determine the locations within the genome where the TFs bind and provide candidate genes that they are likely to regulate. Different alternatives of the immunoprecipitation of cross-linked chromatin have been proposed like DNA immunoprecipitation-chip (DIP-chip) (Liu et al., 2005), DNA adenine methyltransferase immunoprecipitation (DamID) (van Steensel and Henikoff, 2000), ChIP-chip (Ren et al., 2000), ChIP-seq (Johnson et al., 2007a). Their main difference lies in the resolution of the genomic signal that they provide. The most widely adopted methods with the highest resolution were the ChIP-chip and its successor ChIP-seq. In these two approaches, sheared DNA fragments that were bound to the protein of interest are analysed by either hybridization on tiling microarrays (ChIP-chip) or by sequencing with high-throughput sequencing technologies (ChIP-seq). ChIP-seq has improved resolution, lower noise levels, and a higher dynamic range (Johnson et al., 2007a). However, the resolution of the binding locations is not sufficient to identify the binding sites precisely, only to indicate a region of 100 or more nucleotide bases in which the TF complex resides. Recently ChIP assays combined with Exonuclease digestion (ChIP-exo) have been used to narrow down candidate sequences in the immunoprecipitated regions (Rhee and Pugh, 2011). An alternative approach for extracting high resolution DNA binding sites from in vivo data is to overlap the genomic positions where the protein binds with a digital DNase I footprinting profile (Hesselberth et al., 2009). In this case dense mapping at single-nucleotide resolution of the individual DNase I cleavage sites enables the detailed identification of the DNA sequence where the transcription factor binds (Hesselberth et al., 2009). In all the above examples, applying motif discovery algorithms to the sequences of those binding regions can usually identify the main overrepresented features of the binding motif but the accuracy of these models is limited. This approach is

unsatisfactory for the distinction of direct and indirect protein–DNA interactions (Gordan et al., 2009). On the other hand, knowing the binding specificities of a TF for a big variety of DNA sequences, if the TF binds to regions of the genome that lack strong binding sites, it is an indication that the TF binds indirectly or that it binds to a weak binding site that requires cooperativity with other factors. In both cases, that information leads to the search for additional interacting factors.

1.7 Definition of transcription factor binding specificities

The term specificity means the ability of a TF to discriminate between different DNA sequences (Stormo and Zhao, 2010). This requires knowing the affinity to the entire set of potential binding sites in order to construct models that allow the complete specificity to be accurately represented. Other than just using the consensus sequence for the highest affinity sites, the simplest and most widely used DNA affinity model is a position weight matrix (PWM) (Stormo, 2000). In the simplest form of this model, a score is calculated for each possible base at each position in the binding site. The sum of the individual base scores gives a total score for that sequence. This allows the model to provide a score to all possible binding sites for the protein. The PWM model assumes that each position independently contributes to the binding. Although the DNA binding specificities of some TFs can adequately be explained with such a simple model for some of them two or more PWMs need to be employed to explain the modes of binding (Badis et al., 2009). For this reason, more complex models assuming dinucleotide interdependency (Sharon et al., 2008) or higher order interactions (Bulyk et al., 2002), that do not assume additivity of the individual bases, have been proposed. Still these models have not found a wide usage. Other approaches involve reporting for every possible bound DNA molecule of a certain length (k-mer), of an enrichment score versus the non-selected DNA molecules (Berger

et al., 2006, Newburger and Bulyk, 2009). These are the most complete approaches considering that they report both low and high affinity binding sites. The classification of different families of transcription factors and of different members of the same family is based on the whole spectrum of binding affinities (Badis et al., 2009, Berger et al., 2008, Jolma et al., 2013). It has also been shown that low affinity interactions are important in vivo (Schmidt et al., 2010, Tanay, 2006, Bradley et al., 2010). Thus a comprehensive representation of the transcription factor binding landscape is important to understand the biological role of the protein.

1.8 Importance of transcription factor DNA binding specificities and their relevance on the induction of gene expression

High-throughput investigation of the DNA binding specificities of a large number of transcription factor families was undertaken recently in two major studies for both human and mouse proteins (Jolma et al., 2013, Badis et al., 2009). Although these investigations exposed the complexity of DNA interactions of each individual protein family, still the way these binding specificities synergize with each other in order to affect gene expression was not very clear until recently. By exploiting high throughput promoter arrays combined with single cell analysis, the effect on gene expression has been slowly exposed (Sharon et al., 2012, Melnikov et al., 2012).

Aims of the project

1. Devise a novel methodology to investigate the DNA binding specificities of a protein using high-throughput sequencing platforms.
2. Create the experimental framework, of the novel methodology, in order to extract a comprehensive picture of the DNA binding preference of a specific protein.
3. Create the analytical framework, of the novel methodology, that will help in the interpretation of the acquired results.
4. Use the novel methodology to characterise the DNA binding specificities and affinities of two proteins that belong to different families, namely the erythropoiesis related transcription factors GATA1 and KLF1.
5. Assess the resolution and the sensitivity of the novel methodology by investigating how the DNA binding specificities of GATA1 change upon complexation with its transcriptional co-factors, FOG1 or KLF1.
6. Investigate how the DNA binding specificities of KLF1 change upon complexation with GATA1.
7. Assess whether the in vitro derived DNA binding specificities can be found in vivo, in GATA1 and KLF1 ChIP-seq data, taking into account the limited resolution in detailed motif characterisation that can be inferred from the ChIP-seq studies.
8. Decipher how the DNA binding affinities of GATA1 and KLF1 ChIP-seq peaks are associated with the regulation of gene expression during erythropoiesis. Specifically, examine the distribution of the GATA1 or KLF1 DNA binding affinities in enhancer, distal or proximal promoter regions and in the presence or absence of other GATA1 transcriptional co-factors. Additionally, examine how the DNA binding affinities of GATA1 and KLF1 ChIP-seq peaks are correlated with the change of histone post translational modifications locally and in the promoter of nearby genes.

CHAPTER 2 MATERIALS AND METHODS

2.1 Materials

The SF9 cells were cultured in insect cell culture medium (Sf-900[®] II SFM, Life Technologies, USA) supplemented with 10% FBS (Fetal Bovine Serum Gold (GE Healthcare)). The SF9 cells were grown in sterilized glass flasks or Erlenmeyer polycarbonate cell culture flasks (Corning). The BL21 (DE3) Escherichia coli (Merck, UK) were cultured in Luria Broth (Life Technologies, USA). The custom made DNA retardation gels were created using the following materials: empty mini 1.0 mm Novex[®] Gel Cassettes (Life Technologies, USA), Novex[®] TBE Running Buffer (5X) (Life Technologies, USA), 30% Acrylamide/BIS Solution 29:1 (Biorad), Ammonium persulfate (Sigma), Novex[®] Empty Gel Cassette Combs mini 1.0 mm 1 well or 2 wells or 5 wells (Life Technologies, USA). The following materials were purchased from the indicated companies: HEPES (Sigma), KCL (Sigma), Poly(dI-dC) (Sigma), EDTA (Sigma), EGTA (Sigma), ZnSO₄ (Sigma), MgCl₂ (Sigma), glycerol (Fisher Chemical), Trizma[®] base (Sigma), DTT (Sigma), Triton X-100[®] (Sigma).

2.2 Amino acid sequence of proteins that were used in this study

The protein sequences that were used along with their purification tags are the following:

- Mus musculus GATA1 amino acid sequence (mGATA1_{NC}): aa 200-318 (N-terminal: FLAG-tag, C-terminal: 6x His-tag) of wild type GATA1 (UniProt ID: P17679)

MADYKDDDDKEARECVNCGATATPLWRRDRTGHYLCNACGLYHKMNGQNRPLIRPKK
RMIVSKRAGTQCTNCQTTTTTLWRRNASGDPVCNACGLYFKLHQVNRPLTMRKDG IQT
RNRKASGKGKKKRGKHHHHHH

- Mus musculus GATA1 amino acid sequence (mGATA1_C): aa 249-318 (N-terminal: FLAG-tag, C-terminal: 6x His-tag) of wild type GATA1 (UniProt ID: P17679)

MADYKDDDDKIVSKRAGTQCTNCQTTTTTLWRRNASGDPVCNACGLYFKLHQVNRPLT
MRKDG IQTRNRKASGKGKKKRGKHHHHHH

- Mus musculus GATA1 amino acid sequence (wt-mGATA1): aa 1-413 (C-terminal: 6x His-tag) of wild type GATA1 (UniProt ID: P17679)

MDFPGLGALGTSEPLPQFVDSALVSSPSDSTGFFSSGPEGLDAASSSTSPNAATAAASA
LAYYREAEAYRHSPVFQVYPLLNSMEGIPGGSPYASWAYGKTALYPASTVCPSHEDAPS
QALDQEGKSNNTFLDTLKTERTLSPDLLTLGTALPASLPVTGSAYGGADFPSPFFSPTGS
PLSSAAYSSPKFHGSLPLAPCEARECVNCGATATPLWRRDRTGHYLCNACGLYHKMNG
QNRPLIRPKKRMIVSKRAGTQCTNCQTTTTTLWRRNASGDPVCNACGLYFKLHQVNRPL
TMRKDG IQTRNRKASGKGKKKRGSNLAGAGAAEGPAGGFMMVAGSSSSGNCGEVASG
LALGTAGTAHLYQGLGPVVLSPVSHLMPFPGPLLGSPTTSFPTGPAPTTSSTS VIAPLS
SKHHHHHH

- Mus musculus FOG1 amino acid sequence (wt-mFOG1): aa 1-995 (C-terminal: FLAG-tag) of wild type FOG1 (UniProt ID:O35615)

MSRRKQSNPRQIKRSLRDMEAGEEAKAMDSSPKEQEAPDPEAPAIEEPPSPREDVSP
PAVPAPPESPEDPEDMEGQEQLEMRPQDEEKEEKEEEAAMASPSWGPEEELALQDGQ
RCVRRARLSLTEGLSWGPFYGSIQTRALSPEREPEGPAVTLMVDESCWLRMLPQVLTEEA
ANSEIYRKDDALWCRVTKVVPSSGGLLYVRLVTEPHGAPRHPVQEPVEPGGLAPVHTDIQ
LLPQQAGMASILATAVINKDVFPCKDCGIWYRSENLQAHLLYYCASRQRAGSPVSATEE
KPKETYPNERVCPFPQCRKSCPSASSLEIHMRSHSGERPFCVCLICLSAFTTKANCERHLK
VHTDTLSGVCHNCGFISTTRDILYSHLVTNHMCQPGSKGEIYSPGAGHPAAKLPPDSLA
GFQQHSLMHSPVLPADKAPTPSSGLDSKAEVTNGETRVPPQNGGSSSPAARTIKVEA
AEEPEATRASGPGEPGPQAPSRTSPHSPNPVRVKTELSSPTPGSSPGPGELTMAGTL
FLPQYVFSPDAGTTTVPTAPQASEILAKMSELVHNRLQQGAGSSGAAGTPTGLFSGTKG
ATCFECEITFNNINNFYVHKRLYCSGRRAPEDPPTVRRPKAATGPAPAPAGAAAEPDPS
RSSPGPGPREEEASGTTTPEAEAAGRGSEGSQSPGSSVDDAEDDPSRTLCEACNIRFS
RHETYTVHKRYYCASRHDPPIRRPPAPTAPGPAAPALTAPPVTRRRRRKLYELPAAGA
PPPAAGPAPVPVPSPTAELPSSPRPGSASAGPAPALSPSPVPDGPIDLSKRPRRQSPD
APTALPALADYHECTACRVSFHSLEAYLAHKKYSCPAAPLRTTALCPYCPPNGRVRGDL
VEHLRQAHGLQVAKPAASPGAEPRTPAERAPRDSPDGRAPRSPSPAPENTPSDPADQG
ARTPSKGPPAPAPAPGGGGGHRYCRLCNIRFSSLSTFIAHKKYYCSSHAAEHVKDYKDD
DDK

- Mus musculus KLF1 amino acid sequence (mKLF1₂₅₇₋₃₅₈): aa 257-358 (N-terminal: 6xHis-tag, C-terminal: FLAG-tag) of wild type KLF1 (UniProt ID: P46099)

MAHHHHHHVGTGTAPPKRSRRTLAPKRQAAHTCGHEGCGKSYTKSSHLKAHLRTHTGE
KPYACSWDGCDFWRFARSDELTRHYRKHTGHRPFCCGLCPRAFSRSDHLALHMKRHLA
DYKDDDDK

- Mus musculus FOG1 finger 1 amino acid sequence (mFOG1_{finger1}): aa 241–295
(N-terminal: FLAG-tag, C-terminal: HA-tag) of wild type FOG1 (UniProt
ID:O35615)

MAGYPYDVDPDYAGMASILATAVINKDVFPCCKDCGIWYRSERNLQAHLLYYCASRQRAGS
P VSATEEKP**DYKDDDDK**

- Mus musculus FOG1 finger 5, 6 amino acid sequence (mFOG1_{finger56}): aa 563-760
(N-terminal: FLAG-tag, C-terminal: HA-tag) of wild type FOG1 (UniProt
ID:O35615)

MAGYPYDVDPDYAGVHNRLQQGAGSSGAAGTPTGLFSGTKGATCFECEITFNNINNFYV
HKRLYCSGRRAPEDPPTVRRPKAATGPARAPAGAAAEPDPSRSSPGPGPREEEASGTT
TPEAEAAGRGSEGSQSPGSSVDDAEDDPSRTLCEACNIRFSRHETYTVHKRYYCASH
DPPPRR PPAPTTAPGPAAPALTAPPVRTRRRRKLYEL**DYKDDDDK**

- Mus musculus FOG1 finger 9 amino acid sequence (mFOG1_{finger9}): aa 945–995
(N-terminal: FLAG-tag, C-terminal: HA-tag) of wild type FOG1 (UniProt
ID:O35615)

MAGYPYDVDPDYAGPSKGPPAPAPAGGGGGHRYCRLCNIRFSSLSTFIAHKKYCSSH
AA EHVK**DYKDDDDK**

The mGATA1, mFOG1 and mKLF1 cDNA were a gift from Dr. John Strouboulis
(Alexander Fleming BSRC, Athens, Greece).

2.3 Expression, purification and characterization of proteins and protein complexes used in this study

2.3.1 Protein expression and purification from insect cells

Baculoviruses were generated in SF9 cells with the use of the Bac-to-Bac Baculovirus Expression System (Life Technologies, USA). Briefly, the plasmids of pFASTBAC Dual (Life Technologies, USA) containing the relevant cDNAs were subjected to recombination with the baculoviral genome in DH10Bac bacterial cells (Life Technologies, USA) as per manufacturer's protocol. The resulting recombinant viral genomes were isolated with the BACMAX DNA purification kit (Epicentre Biotechnologies, USA). Two rounds of viral amplification were performed before protein production. During the first round, the viral particles were produced in a non-steering culture in a 6 well CELLSTAR plate (Greiner bio-one, USA), by transfecting 2×10^6 SF9 cells with 2 μ g of viral recombinant genome and 120 μ l of FuGENE-6 (Roche Diagnostics Ltd, UK) for 5 days. The second round of viral amplification was done by infecting a 1×10^6 cells/ml culture of SF9 cells with the first round of viral particles (dilution factor 1:40 times) for 5 days. Subsequently, the protein expression was done by infecting 1×10^6 cells/ml cultures of SF9 cells with the appropriate amount of viral particles (M.O.I) for the appropriate amount of time. For the wt-mGATA1 alone and the wt-mGATA1/wt-mFOG1 complex the protein expression was done for two and three days respectively as mentioned in the main text. The ratio of viral particles to cells and the protein production period were optimised so as:

- i) In case of undiluted virus, the survival of the SF9 cells, at the end of the protein production period, is at least 90% of the initial number of cells.
- ii) In case of diluted virus, the total number of the SF9 cells at the end of the protein production period is no more than 5×10^6 cells/ml.

- iii) In case of diluted virus, the optimal viral dilution was selected as the one that uses the minimum amount of viral particles that give the maximum protein production per specific number of SF9 cells.
- iv) Degradation fragments of the protein were tried to be minimised by diluting the virus.

The proteins were purified based on their purification tags as per manufacturer's protocol. In more detail, for proteins that had the His-tag, and were purified first again the His-tag and secondly with DNA affinity purification (wt-mGATA1 and wt-mGATA1/wt-mFOG1 complex), the purification protocol was as follows:

Ni-NTA beads for purification of His-tagged proteins were purchased from Novagen (Merk, UK). One litter of infected SF9 cells was harvested in "**lysis**" buffer (50 mM NaH_2PO_4 pH 8.0, 300 mM NaCl, 10mM Imidazole, 2x complete EDTA-free Protease Inhibitors (Roche)) and sonicated for 2.5 minutes with 10 pulses On-Off. Then, 20% Triton X-100 for a final concentration of 0.5% was added to the lysate and the lysate was left rotating at 4°C for 30 minutes. Eventually, the debris were removed via centrifugation at 16000xg for 30 minutes. Then the cleared lysate was incubated with the Ni-NTA beads for 30 minutes, washed 3 times in "**wash**" buffer (50 mM NaH_2PO_4 pH 8.0, 300 mM NaCl, 20mM Imidazole) and eluted with "**elution**" buffer (50 mM NaH_2PO_4 pH 8.0, 300 mM NaCl, 250mM Imidazole). Afterwards, the buffer was exchanged with "**DNA-bind**" buffer (10mM Hepes pH 7.5, 0.3mM EDTA, 0.6mM EGTA, 50mM KCl, 10% glycerol, 0.05% Triton X-100, 1.5mM DTT) and the eluted proteins were incubated with a mixture of 3 different biotinylated DNA oligonucleotides conjugated with Streptavidin-Agarose beads (Sigma-Aldrich) for 2 hours, washed 3 times for 5 minutes with the "**DNA-wash**" buffer (10mM Hepes pH 7.5, 0.3mM EDTA, 0.6mM EGTA, 150mM KCl, 10% glycerol, 0.05% Triton X-100, 1.5mM DTT) and eluted in "**DNA-elution**" buffer (10mM Hepes pH 7.5, 0.3mM EDTA, 0.6mM EGTA, 500mM KCl, 10% glycerol, 0.05% Triton X-100, 1.5mM DTT). Finally, the DNA-elution buffer was exchanged with "**DNA-bind**" buffer with

ultracentrifugation columns (Amicon Ultra-0.5 mL Centrifugal Filters, Millipore, USA) (30kDa cut off for the wt-mGATA1 alone or 100kDa cut off for the wt-mGATA1/wt FOG1 complex) and the proteins were stored at -80°C. The purified recombinant proteins were detected by SDS-PAGE and Coomassie Blue staining.

The double-stranded purification oligonucleotides used were AACTGATAAGGAT (Perkins et al., 1989), GCGCTCAGAGATAAGGCCTTG (Yu et al., 2002), GATCTCCGGCAACTGATAAGGATTCCCTG (Eisbacher et al., 2003). All the single stranded purification oligonucleotides were obtained from METABION international AG (Martinsried, Germany). Prior to use the oligonucleotides were prepared as follows. The positive and the negative strand were annealed via incubation in NEB Buffer 3 at 94°C for 1 minute and then subsequently for an additional 69 cycles of 1 minute each coupled to a per-cycle, step-wise decrease of 1°C. A pre-annealed oligo mixture (712.5 µl) was conjugated with the streptavidin-agarose beads for 1 hour at room temperature and subsequently washed 3 times in “DNA-bind” buffer before incubation with the purified protein.

2.3.2 Protein expression and purification from bacteria cells

For the bacterial expression of one protein fragment per bacterial cell the pET-45(b+) DNA vector (Millipore, USA) was used. For the simultaneous expression of two protein fragments per bacterial cell the pETDuet-1 DNA vector (Millipore, USA) was used. The expression was performed in BL21 (DE3) Escherichia coli (Merck, UK). The mGATA1_{NC}, mGATA1_C, mKLF1₂₅₇₋₃₅₈ were expressed as fusion proteins with a HIS tag and a FLAG-tag. In case of the mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex, the mGATA1_{NC} was expressed as fusion protein with an HA tag and a HIS tag whereas the mKLF1₂₅₇₋₃₅₈ was expressed as fusion protein with an HA tag and a FLAG-tag. In case of the mGATA1_{NC}/mFOG1_{finger1}, the

mGATA1_{NC}/mFOG1_{finger56} or the mGATA1_{NC}/mFOG1_{finger9} complexes, the mGATA1_{NC} was expressed as fusion protein with an HA tag and a HIS tag and the mFOG1_{finger1}/mFOG1_{finger56}/mFOG1_{finger9} were expressed as fusion proteins with an HA tag and a FLAG-tag. All the proteins were expressed through induction of the bacterial cells with 0.2 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) (Sigma) at 37°C for 5 hours. The bacterial cells were harvested in “**FLAG binding**” buffer (50mM Tris pH 7.5, 5mM EDTA, 150mM NaCl, 0.1% NP-40, 2x complete EDTA-free Protease Inhibitors (Roche,UK)). The pellets were sonicated for 2.5 minutes with 10 pulses On-Off. Then, 20% Triton X-100 for a final concentration of 0.5% was added to the lysate and the lysate was left rotating at 4°C for 30 minutes. Eventually, the debris were removed via centrifugation at 16000xg for 30 minutes. A two-step purification procedure was then performed as follows.

In the first step the cleared lysate was incubated with beads that were coated with an anti-FLAG antibody (ANTI-FLAG[®] M2 Affinity Gel, Sigma, A2220) (approximately 40 µl per 2ml of cell lysate) at 4°C for 2 hours. The beads were then washed 2 times with 5ml of “**FLAG binding**” buffer. Next, the beads were washed 1 more time with 5ml of FLAG binding buffer, the supernatant was discarded, and 50 µl of FLAG binding buffer was left on top of the beads. The elution of the bound protein was done with a FLAG peptide (3x FLAG peptide, Sigma, F4799). Specifically, 200-450 µl of “**FLAG elution**” buffer solution was put to the resin. The “**FLAG elution**” buffer contained 180 µg/ml of 3x FLAG peptide in “**FLAG binding**” buffer. The eluted protein was then buffer exchanged in “**Ni-NTA bind**” buffer (50 mM NaH₂PO₄ pH 8.0, 300 mM NaCl, 10mM Imidazole) with ultracentrifugation columns (Amicon Ultra-0.5 ml Centrifugal Filters, Millipore, USA) (10kDa cut off for the mGATA1_{NC} or the mKLF1₂₅₇₋₃₅₈ or the mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex or the mGATA1_{NC}/mFOG1_{finger1} or mGATA1_{NC}/mFOG1_{finger56} or mGATA1_{NC}/mFOG1_{finger9} complexes and 3kDa for the mGATA1_C).

The second purification step was performed with Ni-coated beads against the HIS tag ('Ni-NTA His-Bind Resin' system, Merck, #70666). The eluted proteins in “**Ni-NTA bind**” buffer were incubated with the Ni-NTA beads for 30 minutes, washed 3 times in “**wash**” buffer (50 mM NaH₂PO₄ pH 8.0, 300 mM NaCl, 20mM Imidazole) and eluted with “**elution**” buffer (50 mM NaH₂PO₄ pH 8.0, 300 mM NaCl, 250mM Imidazole).

To further clean the GATA1_{NC} and mKLF1₂₅₇₋₃₅₈ proteins instead of column chromatography, two ultracentrifugation steps were also performed. At first the protein solution was passed from a 30kDa cut-off ultracentrifugation column (Amicon Ultra-0.5 ml Centrifugal Filters, Millipore, USA) and the flow through was passed from a 10kDa cut-off ultracentrifugation column (Amicon Ultra-0.5 ml Centrifugal Filters, Millipore, USA). Then, the retained fraction of the last column was kept. During the last two ultracentrifugation steps, the buffer was exchanged with “**storage**” buffer (10mM Hepes pH 7.5, 50mM KCl, 10% glycerol, 0.05% Triton X-100, 1.5mM DTT).

To clean the GATA1_{NC}/FOG1_{finger1}, GATA1_{NC}/FOG1_{finger56} and GATA1_{NC}/FOG1_{finger9} protein complexes two ultracentrifugation steps were again performed. At first the protein solution was passed from a 50kDa cut-off ultracentrifugation column (Amicon Ultra-0.5 ml Centrifugal Filters, Millipore, USA) and then the flow through was passed from a 10kDa cut-off ultracentrifugation column (Amicon Ultra-0.5 ml Centrifugal Filters, Millipore, USA). Then the retained fraction of the last column was kept. During the last two ultracentrifugation steps, the buffer was exchanged with “**storage**” buffer.

To clean the GATA1_C protein, two ultracentrifugation steps were also performed as described above. At first the protein solution was passed from a 10kDa cut-off ultracentrifugation column (Amicon Ultra-0.5 ml Centrifugal Filters, Millipore, USA) and the flow through was passed from a 3kDa cut-off ultracentrifugation column (Amicon Ultra-0.5 ml Centrifugal Filters, Millipore, USA). Then the retained fraction of the last column was

kept. During the last two ultracentrifugation steps, the buffer was again exchanged with “storage” buffer.

The proteins were aliquoted and stored at -80°C.

2.3.3 Electrophoretic Mobility Shift Assay (EMSA) for testing the functionality of proteins and the protein complexes

The purified wt-mGATA1 or the wt-mGATA1/wt-mFOG1 complex were assayed for DNA binding with one of the following DNA molecules:

- i) A biotin-labelled double-stranded oligonucleotide (see section 2.3.1) (GCGCTCAGAGATAAGGCCTTG).
- ii) A biotin-labelled DNA library with a 20-mer degenerate region (see section 2.4.1).

For the shift assays 20 fmoles of biotin-labelled oligonucleotides were incubated with 1.5 µg of protein for 1 hour at room temperature in binding buffer (10mM Hepes pH 7.5, 0.3mM EDTA, 0.6mM EGTA, 50mM KCl, 10% glycerol, 0.05% Triton X-100, 1.5mM DTT). Protein-DNA binding complexes were resolved by electrophoresis using commercial DNA retardation gels in 0.5X Novex® TBE Running Buffer (Life technologies, USA) at 150 Volt for 1 hour, transferred to Blotting-Nylon 66 type B positive membranes (Fluka), UV cross-linked, and visualized using the Chemiluminescent Nucleic Acid Detection System (Pierce, USA) as per manufacturer's protocol. 6% TBE NuPAGE® gels (Life technologies, USA) were used in order to separate the wt-mGATA1 and the wt-mGATA1/wt-mFOG1 protein-DNA complexes from the unbound DNA.

For super-shift assays, the GATA1 antibody (N1: sc-266, Santa Cruz Biotechnology, USA) and the FOG1 antibody (M-20: sc-9361, Santa Cruz Biotechnology, USA) were concentrated 10X and buffer exchanged in 10mM Tris-Cl pH 7.4, 30mM KCl. Before

addition of the biotin-labelled oligonucleotides, 7 µg of the 10X concentrated antibodies or non-immune IgG (Santa Cruz Biotechnology ,USA) were incubated with 1.5 µg of the purified protein for 1 hour at room temperature in the binding buffer (10mM Hepes pH 7.5, 0.3mM EDTA, 0.6mM EGTA, 50mM KCl, 10% glycerol, 0.05% Triton X-100, 1.5mM DTT).

We also suggest calculating the concentration of the functional GATA1 protein in the purified sample from the EMSA reaction by titrating an increased amount of a “GATA” containing oligonucleotide. When the binding saturates, it reveals the concentration of the GATA1 protein used (Mott et al., 2004).

2.3.4 Western Blot for the characterization of the isolated proteins and protein complexes

The protein samples were separated on a reducing 4–12% Bis-Tris NuPAGE® (Life technologies, USA) and electroblotted onto Invitrolon® PVDF membranes (Life technologies, USA). The membrane was blocked with a solution of 5% non-fat dried milk powder (AppliChem) in 1x phosphate-buffered saline (pH 7.4) (**blocking buffer**) (Sigma) for 60 min. The membrane was then incubated overnight with the primary antibody against GATA1 (N1: sc-266, Santa Cruz Biotechnology, USA), FOG1 (M-20: sc-9361, Santa Cruz Biotechnology, USA), FOG1 (A-20: sc-9362, Santa Cruz Biotechnology, USA), anti-FLAG (ab1162, Abcam, Cambridge), anti-6X His-tag (ab18184, Abcam, Cambridge) at a dilution of 1:1000, 1:3000, 1:3000 1:2000, 1:2000 respectively in blocking buffer, after which the membrane was washed four times. It was then incubated for 60 min with the secondary antibody (alkaline phosphatase-conjugated IgG) at a dilution of 1:30000 in blocking buffer. After another four washes, antigen-antibody complexes were detected using ECL plus (GE Healthcare) and the signal was exposed with Hyperfilm ECL (GE Healthcare).

2.3.5 EMSA-Western for the characterization of the proteins and protein complexes in the DNA band shifts

At first 600 ngs of a “GATA” containing double-stranded oligonucleotide (see section 2.3.1) (GCGCTCAGAGATAAGGCCTTG) were incubated with 1.5 µg of wt-mGATA1 or the wt-mGATA1/wt-mFOG1 complex for 1 hour at room temperature in binding buffer (10mM Hepes pH 7.5, 0.3mM EDTA, 0.6mM EGTA, 50mM KCl, 10% glycerol, 0.05% Triton X-100, 1.5mM DTT). Protein-DNA binding complexes were resolved by electrophoresis using 6% custom made 5 wells DNA retardation gels (see section 2.1) at 150 Volt for 1 hour. The gels were then immediately electro-blotted on an Invitrolon® PVDF membrane (Life technologies, USA) for 1 hour with a Trans-Blot® SD Semi-Dry Transfer Cell device (Biorad). Before the transfer, the PVDF membrane was equilibrated in 0.5X Novex® TBE Running Buffer (Life technologies, USA) for 10 minutes. Then the blocking and the incubation with the primary and the secondary antibodies were performed as described in the Western Blot section (see section 2.3.4).

2.4 EMSA-seq experiments

2.4.1 Creation of the 20mer degenerate DNA libraries that were used in the EMSA-seq experiments

HPLC purified 20mer degenerate DNA library was ordered from METABION international AG (Martinsried, Germany). The template of the degenerate 20mer DNA library is the following:

5-tgccatgcgccaactgccA NNNNNNNNNNNNNNNNNNNNNN Cagcccc agccggaccac tAA-3'

Each template was double stranded with the following barcoded primers which are complementary with the 3' of the DNA library:

Group A

- 1 . CCATGCATttagtggtccggctggggct**
- 2. AACATGCAtttagtggtccggctggggct**
- 3. TTGCATGCttagtggtccggctggggct**
- 4. GGTGCATGttagtggtccggctggggct**

Group B

- 5. CCACTGACttagtggtccggctggggct**
- 6. AAGACTGAtttagtggtccggctggggct**
- 7. GGTGACTGttagtggtccggctggggct**
- 8. TTCTGACTttagtggtccggctggggct**

Group C

- 9. AAGATCGAtttagtggtccggctggggct**
- 10. TTCGATCGttagtggtccggctggggct**
- 11. CCTCGATCttagtggtccggctggggct**
- 12. GGATCGATttagtggtccggctggggct**

This design permitted us to have 12 barcoded libraries separated in 3 groups. In replicate experiments only libraries that belong in the same group were used. This was done in order to avoid sequencing in the Illumina platform, nucleotide positions that do not have an equal representation of ACGT (low complexity positions). These sequencing positions can cause miscalculation of the quality scores of each base (PHRED score).

The double stranding of the 20mer DNA library was performed as follows. At first 9 µl from a 100µM solution of the 20mer DNA library were mixed with 9 µl from a 100µM solution of a barcoded primer in a 20 µl reaction containing 1X NEB3 buffer (New England Biolabs,

UK). Then the sequences in the reaction were annealed via incubation at 94°C for 1 minute and subsequently for an additional 69 cycles of 1 minute each coupled to a per-cycle, step-wise decrease of 1°C. Afterwards the library was double-stranded by incubating the 18 µl reaction with 10U (1U per 1 ug of DNA but no more than 0.67 U/µl) of Large Klenow Fragment (M0210S, New England Biolabs, UK), 1X NEB2 buffer (New England Biolabs, UK) and 167µM of dNTPs (New England Biolabs, UK) in a 30 µl reaction. The reaction was incubated at 20°C for 30 minutes and then at 37°C for 15-30 minutes. Phenol:chloroform extraction was followed with sodium acetate precipitation and at least two final washes in 70% ethanol. The library was reconstituted in water at a concentration of 0.5 µg/µl, aliquoted and stored at -20°C.

2.4.2 Creation of protein-DNA binding isotherms with fluorescence polarization assays

The fluorescence anisotropy technique is based on the following principle: The DNA is labelled with a fluorophore, in our case 6-FAM (6-Carboxyfluorescein) (6-FAM-DNA). Increasing amounts of the protein are incubated with the same amount of 6-FAM-DNA. Free 6-FAM-DNA and the complex-bound 6-FAM-DNA are illuminated with polarized light. Because the 6-FAM-DNA in the complex tumbles more slowly, a larger proportion of its emitted light remains polarised (anisotropic) (Figure 8). The difference between the anisotropy values of the free DNA and the complex bound DNA is used to create the binding isotherm of the protein-DNA interaction.

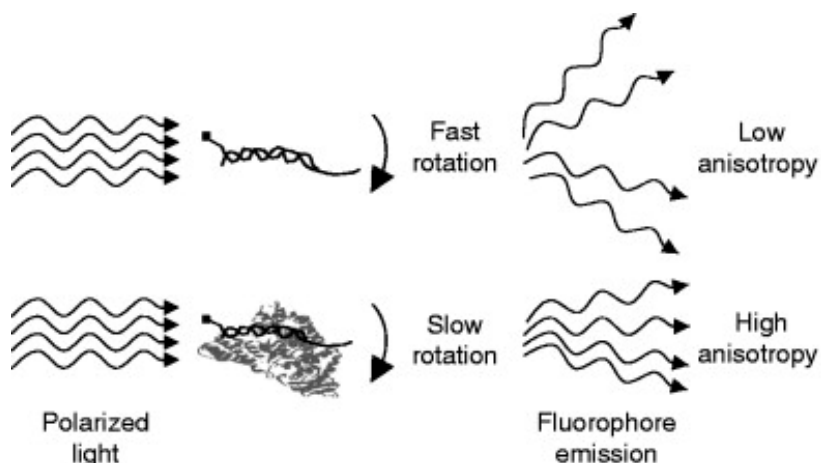


Figure 8. Biophysical principle of fluorescence anisotropy. From (LiCata and Wowor, 2008).

More specifically, 100 fmoles of 6-FAM (6-Carboxyfluorescein) labelled DNA oligonucleotides (the concentration of the 6-FAM-DNA was measured with NanoDrop 1000 Spectrophotometer (Thermo scientific) at 260 nm) were incubated with increasing amounts of protein in a 20 μ l buffer containing 10mM HEPES pH 7.5, 50mM KCl, 0.06% Triton X-100, 1.5mM DTT, 10% Glycerol. The reactions were incubated for 1 hour at room temperature before collecting the raw fluorescence values. Data were collected with a POLARstar Omega (LABTECH, Germany) instrument. Measurements for the fluorescein moiety of the DNA molecules were taken with an excitation filter at 494nm and an emission filter at 519nm. Three replicates were used per experiment. For each condition the mean and the standard deviation were calculated.

One of the following 6-FAM-DNA were used per experiment:

- i) A double-stranded oligonucleotide, 6-FAM labelled only in the forward strand (see section 2.3.1) (GCGCTCAGAGATAAGGCCTTG).
- ii) A DNA library with a 20-mer degenerate region. During the double stranding, the barcoded primer number 9 was 6-FAM labelled (see section 2.4.1).

The fluorescence anisotropy (r) of a sample was calculated from the following equation, as described previously (Weinberg et al., 2004), using the instruments own software:

$$r = \frac{I_{VV} - G * I_{VH}}{I_{VV} + 2 * G * I_{VH}}$$

where (I_{VV}) and (I_{VH}) are the fluorescence intensities measured with a vertically and horizontally oriented emission polariser respectively. The G-factor (G) represents the instrument sensitivity ratio towards vertically and horizontally polarized light. It is calculated from the instrument automatically.

2.4.3 Electrophoretic Mobility Shift Assay (EMSA) for the EMSA-seq experiments

The concentration of the degenerate DNA library and protein used in each EMSA-seq experiment is presented in Table I. The EMSA-seq reactions for the mGATA1_{NC}, the mGATA1_C, the mKLF1₂₅₇₋₃₅₈, the mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex or the mGATA1_{NC}/mFOG1_{finger1}, mGATA1_{NC}/mFOG1_{finger56}, mGATA1_{NC}/mFOG1_{finger9} complexes were performed in a 20 μ l reaction for 1 hour at room temperature. The buffer consisted from 10mM HEPES pH 7.5, 50mM KCl, 0.06% Triton X-100, 1.5mM DTT, 10% Glycerol. To avoid any interference in the kinetic analysis from a potential competitor, apart from the very simple reaction buffer, we used a reduced concentration of 0.01 ng/ μ l Poly(dI-dC).

The EMSA-seq reactions for the wt-mGATA1 and the wt-mGATA1/wt-mFOG1 complex were performed in a 10 μ l of reaction for 1 hour at room temperature. Because the proteins were isolated with DNA affinity chromatography, in order to reduce potential non specific binding, we used a buffer consisted from 10mM HEPES pH7.5, 118mM KCL, 0.06% Triton X-100, 1.5mM DTT, 10% Glycerol, 0.3 mM EDTA, 0.6mM EGTA, 10 μ M

ZnSO₄, 0.5 mM MgCl₂, 10 ng/μl Poly(dI-dC). Each individual salt (KCL, ZnSO₄, MgCl₂) was titrated independently in fluorescence anisotropy assays and the minimum concentration was selected that did not interfere with the overall binding and eliminated the subtle binding curves in the negative control. The negative control represents a lysate that lacks the specific protein and that was purified with the same methodology as the one that contains the specific protein.

The reactions were then loaded on the DNA retardation gels (commercial or custom made) and the protein-DNA binding complexes were resolved by electrophoresis in 0.5X Novex[®] TBE Running Buffer (Life technologies, USA) at 150 Volt for 1 hour. The shifted DNA bands were processed as in section 2.4.4.

6% TBE NuPAGE[®] gels (Life technologies, USA) were used in order to separate the wt-mGATA1 and the wt-mGATA1/wt-mFOG1 protein-DNA complexes from the unbound DNA. 8% custom-made 1 well polyacrylamide DNA retardation gels (see section 2.1) were used to separate the mGATA1_{NC} protein. 8% custom-made 2 well polyacrylamide DNA retardation gels were used to separate the mKLF1₂₅₇₋₃₅₈, the GATA1_{NC}/mKLF1₂₅₇₋₃₅₈, the GATA1_{NC}/FOG1_{finger1}, the GATA1_{NC}/FOG1_{finger56} or the GATA1_{NC}/FOG1_{finger9} bound DNA from the unbound one. 10% custom-made 2 well polyacrylamide DNA retardation gels were used to separate the GATA1_C bound from the unbound DNA.

Protein and DNA concentrations in the fluorescence anisotropy assay				Protein and DNA concentrations in the EMSA-seq reaction			
mGATA1 _{NC}	conditions	nM of protein	nM DNA		conditions	nM of protein	nM DNA
	1	254.3692315	5.15625		1	2657.649731	53.8725
	2	350.3576207	5.15625		2	3660.536422	53.8725
	3	446.34601	5.15625		3	4663.423112	53.8725
	4	542.3343992	5.15625		4	5666.309803	53.8725
	5	638.3227885	5.15625		5	6669.196494	53.8725
	6	734.3111777	5.15625		6	7672.083185	53.8725
	7	830.299567	5.15625		7	8674.969876	53.8725
	8	926.2879562	5.15625		8	9677.856566	53.8725
	9	1022.276345	5.15625		9	10680.74326	53.8725
mKLF1 ₂₅₇₋₃₅₈	conditions	nM of protein	nM DNA		conditions	nM of protein	nM DNA
	1	757.7952222	5.15625		1	7917.444481	53.8725
	2	969.7659137	5.15625		2	10132.11427	53.8725
	3	1128.743932	5.15625		3	11793.11661	53.8725
	4	1303.619753	5.15625		4	13620.21918	53.8725
	5	1588.013764	5.15625		5	16591.56781	53.8725
mGATA1 _{NC} /mFOG1 _{finger1}	conditions	nM of protein	nM DNA		conditions	nM of protein	nM DNA
	1	44847.35603	53.87640752		1	13866.39032	53.87640752
	1	13866.39032	53.87640752		1	2692.249142	53.87640752
mGATA1 _{NC} /mFOG1 _{finger9}	conditions	nM of protein	nM DNA		conditions	nM of protein	nM DNA
	1	5734.490672	53.87640752		2	5734.490672	53.87640752
mGATA1 _{NC} /mFOG1 _{finger56}	conditions	nM of protein	nM DNA		conditions	nM of protein	nM DNA
	1	10768.99657	53.87640752		3	10768.99657	53.87640752
wt-mGATA1	conditions	nM of protein	nM DNA		conditions	uM of protein	uM DNA
	1	140 (51.5)	14 (5.15)		1	22.52	1.79
wt-mGATA1/wt-mFOG1	conditions	nM of protein	nM DNA		conditions	uM of protein	uM DNA
	1	140 (51.5)	14 (5.15)		1	22.52	1.79

Table I. Concentration of protein and DNA used in the EMSA-seq reactions presented. Additionally the respective concentration in the fluorescence anisotropy assay is presented. In the wt-mGATA1, due to the higher concentration of DNA used during the fluorescence anisotropy assay, the number in the parenthesis represents proportionally scale down values so that the concentration of DNA is similar with the one used in the mGATA1_{NC} and the mKLF1₂₅₇₋₃₅₈ experiments.

2.4.4 EMSA-seq library preparation

In order to avoid damaging the DNA sequences of the library, the visualisation of the shifted DNA band in the DNA retardation gel was performed with a Dark Reader blue transilluminator (Clare Chemical Research). For every EMSA-seq experiment, the DNA band of the DNA retardation gel, that had the protein-DNA complex, was cut out of the gel with a scalpel. The DNA was then eluted from the acrylamide gel using the QIAEX II gel extraction kit (Qiagen). The DNA was eluted for 12 to 24 hours overnight and the purification of DNA from contaminants was performed as per manufacturer's protocol. The library preparation was performed with the NEB ChIP-seq Library Prep Master Mix Set for Illumina (E6240L, New England Biolabs, UK) as per manufacturer's protocol.

We started by pooling replicate experiments with different indexes. Library quantification was performed with both a spectrophotometric method that uses a fluorescence dye (QUBIT high sensitivity kit) (Life technologies, USA) and through qPCR. In the qPCR method (IQ SYBR Green Supermix) (Biorad), a standard curve was created based on a 10 fold serial dilution of the original 20mer library. The forward primer of the qPCR used was 5'-tgccatgcgccaactgcc-3' and the reverse primer used was one from the indexes, for example 5'-AAGATCGAtttagtggtccggctggggct-3'.

After creation of the Illumina libraries, the different conditions were pooled again in order to run in the same flow cell lane. The pooling of the libraries was done after calculating their concentration with both the QUBIT high sensitivity kit (Life Technologies, USA) and the KAPA Library Preparation Kit for Illumina (Kapa Biosystems Inc, USA).

2.4.5 Bioinformatic processing of the sequenced files from the EMSA-seq experiments

The bioinformatic processing of sequencing files (FASTQ files) was done with in house Perl scripts. In more detail, the sequencing produced two files for each sample. A file that contained the sequencing results of the forward strand of the reads and a file that contained the sequencing results of the reverse strand of the same reads. Because the sequenced read (20bp degenerate region + 20bp left (5') fixed flanking DNA sequence + 20bp right (3') fixed flanking DNA sequence) was less than the total sequencing length (50 bp from the start of the forward strand and 50 bp from the start of the reverse strand) the sequencing information for each read was combined into one file. In all the sequenced reads, bases with a base quality score (Phred score) less than 25 were masked with the letter "N" using the `fastq_masker` function from the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Subsequently, we used BOWTIE (Langmead et al., 2009) to map the 5' fixed flanking DNA sequence (tgccatgcgccaactgcca) on the sequenced reads using the default parameters with the exception of the following ones " --seedmms 1 --seedlen 30 --maqerr 41 -m 3 -k 3 ". We then removed the 5' fixed flanking DNA sequence whereas we used the 3' fixed flanking DNA sequence to separate the multiplexed libraries based on their barcode sequence. Afterwards, we removed the 3' fixed flanking DNA sequence and we were left with the 20bp degenerate region. In the 20bp degenerate region we deconvoluted k-mers with the "Meryl" software (http://sourceforge.net/apps/mediawiki/k-mer/index.php?title=Main_Page). The number of 20bp sequenced reads used for the k-mer deconvolution in the different experiments is presented in Table II. K-mers of length 4, 5, 6, 7 and 8 base pairs were extracted. For each k-mer the number of occurrence in each database was calculated. Replicates were used to calculate an average value for each k-mer.

experimental conditions	replicate 1 number of reads	library barcode	replicate 2 number of reads	library barcode	replicate 3 number of reads	library barcode	Illumina index
mGATA1nc EMSA-seq condition 1	1820726	2	1580063	3	2585783	4	1
mGATA1nc EMSA-seq condition 2	1860309	2	1592434	3	1624011	4	2
mGATA1nc EMSA-seq condition 3	2168637	2	1408212	3	1920504	4	3
mGATA1nc EMSA-seq condition 4	2040416	2	1725614	3	1391825	4	4
mGATA1nc EMSA-seq condition 5	2008742	2	1321850	3	1481341	4	5
mGATA1nc EMSA-seq condition 6	1778022	2	1372400	3	2269688	4	6
mGATA1nc EMSA-seq condition 7	2936368	2	1305405	3	1963967	4	7
mGATA1nc EMSA-seq condition 8	2479845	2	1608896	3	2023694	4	8
mGATA1nc EMSA-seq condition 9	1683694	2	1972896	3	2676696	4	9
mGATA1nc dimer EMSA-seq condition 1			1336587	3	2082849	4	5
mGATA1c EMSA-seq condition 1	2501882	2	1881155	3	1919320	4	10
wt-mGATA1	2618938	1	1738070	12			1
wt-mGATA1/wt-mFOG1	1919455	1	2387610	12			3
mGATA1nc/mFOG1 _{finger1} EMSA-seq	1770515	2	1352239	3	1373917	4	8
mGATA1nc/mFOG1 _{finger9} EMSA-seq	1403135	2	1275265	3	1465421	4	9
mGATA1nc/mFOG1 _{finger56} EMSA-seq condition 1	1666033	2	1434198	3	1816197	4	10
mGATA1nc/mFOG1 _{finger56} EMSA-seq condition 2	1610171	2	1134554	3	1624407	4	11
mGATA1nc/mFOG1 _{finger56} EMSA-seq condition 3	2214844	2	1670069	3	2001280	4	12
mKLF1257-358 EMSA-seq condition 1	1708018	2	1382296	3	1492503	4	13
mKLF1257-358 EMSA-seq condition 2	1983618	2	1502802	3	2022155	4	14
mKLF1257-358 EMSA-seq condition 3	2112372	2	1592642	3	1912248	4	15
mKLF1257-358 EMSA-seq condition 4	1936171	2	1351449	3	1703525	4	16
mKLF1257-358 EMSA-seq condition 5	2154256	2	1485121	3	1804531	4	18
mKLF1257-358 EMSA-seq condition 6	2566286	2	1671305	3	1912801	4	19
mGATA1nc/mKLF1257-358 main band EMSA-seq condition 1	2683659	2	1930170	3	1924496	4	20
mGATA1nc/mKLF1257-358 upper band EMSA-seq condition 1	1388379	2	1222903	3	1788332	4	23
original 20mer DNA library	2040571	2	1617538	3	1614719	4	11

Table II. Number of sequenced reads per experiment per replicate. The different conditions represent the ones presented in Table I. The different library barcodes represent the ones presented in section 2.4.1. The libraries of each experimental condition were mixed together in the first multiplexing step. Afterwards, experimental conditions that were going to run on the same lane of a sequencing flow cell in the Illumina sequencing platform were mixed together and barcoded with the indicated Illumina indexes (second multiplexing step).

2.4.6 EMSA-seq analysis

We noticed that in the replicates, in each EMSA-seq experiment, the fold change of a specific k-mer versus the unbound DNA 20mer library might show some variation. On the contrary, the rank ordering of the k-mer values in the replicates is much more robust. If we just consider the replicates, without correcting for this aspect, then we lose the sensitivity to detect differentially expressed k-mers because the inflated variance biases the test statistic. Although we can find differentially bound k-mers using rank statistics, these techniques are not as sensitive as parametric statistics. For this reason, in order to normalise the replicates we used quantile normalisation of the fold change values as follows:

For each replicate, we calculate the fold change of each k-mer against the unbound DNA 20mer library. We then order the values of the different k-mers, from the highest to the lowest for each one replicate independently. Starting from the highest trio of values of the three replicates we calculate the average value. We then assigned to the highest value k-mer in each individual replicate, the highest average value. Then to the second highest value k-mer in each individual replicate, the second highest average value and so on. We next transform the fold change values back to reads with the following equation:

$$\text{number of normalised reads in sample} = \text{number of reads in unbound 20mer library} * \text{fold change}$$

We then calculated the differentially bound k-mers by using the **exactTest** function in the R package **EdgeR**. Because the function gives p-values for two tailed test, and we are only interested to find bound k-mers that are enriched we divided each p-value with 2. We also corrected the p-values for multiple hypotheses testing using the **p.adjust** function in the R package with method **bonferroni**. We then used p-values with cut-off less than 1×10^{-6} to find enriched 8mers. This p-value cut-off was produced from the following: We

are testing 32896 8-mers and for each test we try to reject the null hypothesis, that there is no enrichment between the bound and the unbound DNA libraries, at a significance level of 0.05. So the adjusted significance level (p-value cut-off) for all the 32896 tests will be $0.05*(1/32896)=1.5*10^{-6}$.

We usually normalise the k-mer values from different libraries with the total number of sequenced reads in each library. As we discuss in the main text, for each protein we found the statistical enrichment of each k-mer for the EMSA-seq condition with the lowest amount of protein to DNA ratio.

The calculation of the dissociation constant values ("Kd") was performed as discussed in the main text.

2.4.7 Calculation of the EMSA-seq coefficient (e)

To calculate the Kd and the EMSA-seq coefficient ("e") coefficient as presented in the main text we used the "nls" function in R. All the parameters were default except for the parameter "algorithm" which was equal to "port". This was selected in order to put constraints on the calculated Kd and "e" parameters and to avoid the algorithm to fail (the value of the constraints used were lower=c(.001, .001) and upper=c(5000000, 50000000), two values for each one of the "Kd" and "e" respectively). In case that the algorithm could not converge (something that was very rare), we increased the number of iterations to 500 (maxiter = 500) and we also forced the algorithm to output the last calculated values of the parameters (warnOnly = TRUE). Overall the function (we will call it function **EQ**) was as follows:

```
nls(((d)/(1-d))) ~ (((c^(e))/((r^(1/1))*(1+(((1-(ratio))*3180)/Kmix))^(e))))^(1)), start = list(r = 1,e=1), control = list(maxiter = 500, warnOnly = TRUE), algorithm = "port", lower=c(.001,.001), upper=c(5000000,50000000))
```

where

d is the fraction of bound DNA for a specific 8mer (maximum 1),

e is the EMSA-seq coefficient,

The number **3180** represents the concentration (nM) of the total 8mers of the DNA library used in the EMSA-seq reaction. In more detail, we used 40 ngs of the 20mer DNA library (mean MW= 37123.7 Da) in each 20 µl reaction. This equals to DNA concentration 265 nM in the reaction. Because each 20mer has twelve 8mers we multiplied the 265 nM DNA concentration with 12 and we ended up having the concentration of 8mers (3180 nM).

ratio is the fraction of bound DNA library,

For both the “Kd” and “e” we computed also the confidence intervals. We used the “confint” function in the library MASS in R at 95% confidence interval. We call the maximum value of Kd ($K_{dk-mer1_upper95\%}$), the minimum value of the Kd ($K_{dk-mer1_lower95\%}$), the maximum value of the “e” ($e1_upper_95\%$) and the minimum value of the “e” ($e1_lower_95\%$). In order to calculate the normalised values of the $K_{dk-mer1_upper95\%}$ ($K_{dk-mer1_normalised_upper_95\%}$) and the $K_{dk-mer1_lower95\%}$ ($K_{dk-mer1_normalised_lower_95\%}$) from (eq 18) we had to assign an “e” value. Although we intuitively expect that for each k-mer the confidence interval values of the “Kd” and “e” parameters to be related, which means that the maximum value of the “e” ($e1_upper_95\%$) might be appropriate to normalise the maximum value of Kd ($K_{dk-mer1_upper95\%}$) and similarly for the lower values, we found that this is not always the case. For this reason, we normalised the $K_{dk-mer1_upper95\%}$ and the $K_{dk-mer1_lower95\%}$ by using the same “e” value that we used to normalise the $K_{dk-mer1}$ as follows:

$$K_{dk-mer1_upper95\%_normalised} = (K_{dk-mer1_upper95\%})^{\frac{1}{e1}}$$

$$K_{dk-mer1_lower95\%_normalised} = (K_{dk-mer1_lower95\%})^{\frac{1}{e1}}$$

For every k-mer we compare whether the co-operative model with the EMSA-seq coefficient fits the data better than the simpler equilibrium model without the EMSA coefficient as follows:

For each one of the model we found the SSR (sum-squared residuals):

$$\text{SSR} = \sum_{i=1}^N [(y_i)_{\text{observed}} - (y_i)_{\text{predicted}}]^2$$

This was calculated in R as follows:

sum(predict(EQ)-((d)/(1-(d)))^2)

where

d is the fraction of bound DNA for a specific 8mer (maximum 1),

EQ is the function presented earlier

In order to compare the two models we used the score of the F statistic:

$$F = \frac{\text{SSRs} - \text{SSRc}}{\text{SSRc}} * \frac{N - m_c}{m_c - m_s}$$

where

SSRs is the sum-squared residuals of the simple model (the model with fewer parameters)

SSRc is the sum-squared residuals of the complex model (the model with more parameters)

N is the number of data points. This number is 27 for the mGATA1_{NC} (3 replicates of 9 conditions) and 18 for mKLF1₂₅₇₋₃₅₈ (3 replicates of 6 conditions).

m_s is the number of parameters for the simple model. In our case is 1 (one parameter: “Kd”).

m_c is the number of parameters for the complex model. In our case is 2 (two parameters: “Kd” and “e”).

We then calculated the p-value of the F statistic (F score) using the “pf” function in R. Because the pf(X=x) is the cumulative distribution that describes the probability that a real value random variable X will be found at a value less than or equal to x [pf(X=x)=P(X≤x)] and we want the probability of a value more or equal to x [P(X≥x)], then the p-value of the F statistic is given as follows:

$$P(f(X) \geq f(X=x)) = 1 - pf(X=x)$$

And more specifically, the pf in the R function is taking the following values:

pf(q, df1, df2, ncp, lower.tail = TRUE, log.p = FALSE)

where

q equals the F score

$$\mathbf{df1} = m_c - m_s$$

$$\mathbf{df2} = N - m_c$$

We then corrected the p-value for multiple hypotheses testing using the Bonferroni correction. We used the “p.adjust” function in R with method “Bonferroni”.

The goodness of fit was assessed with the Pearson correlation (r^2) (function “cor” in R with method=“Pearson”).

2.4.8 DNA binding motif creation

The motifs of Figure 28 and Figure 29 were created as follows. We started with a **core** GATA, GATT, GATC, GATG motif and we counted the number of times that each nucleotide appears in positions -2, -1, +1, +2. For example for the -2 position we counted the number of occurrence of the following 6mers **ANGATA**, **CNGATA**, **GNGATA**, **TNGATA** in both the sample and the input library. We then found the fold change of occurrence of the sample versus the input library. The fold change represents a normalised occurrence value independently of the original number of k-mers in the input library. The fold change was represented as a number with maximum 2 decimals. We then multiplied the fold changes with an arbitrary number of 100 to transform the decimal values into real numbers. Afterwards, we created a file with sequences where each one of the examined k-mers was appearing the number of times of the fold change multiplied with 100 as mentioned previously. Finally, we used the Weblogo software (Crooks et al., 2004) to represent the motifs. We also represented with a heatmap the frequency of occurrence of the 8mers [ACGT][ACGT]**core**[ACGT][ACGT], where core is one of the already mentioned motifs. The frequency of occurrence was represented as fold change. The heatmaps were drawn with the R function `heatmap.2` in the package `gplots`.

2.4.9 Statistical methods

Except otherwise mentioned in the main text, the statistical methods used in some of the chapters of the manuscript are the following:

In order to find the statistical significance of the differences between the 3 groups in Figure 104 we focused on the median value and we performed a permutation analysis. At first we found the difference in the values of the GATA1 and GATA2 specific groups. We

combined both groups and we randomly assigned each value to either a GATA1 or a GATA2 specific group so that the number of values in each category is equal to the initial GATA1 and GATA2 specific groups. We did this 1000 times. We then examined in how many cases, in the permutation samples, the value of the difference between the median value of the GATA1 and the GATA2 specific peak group is higher than the difference between the initial GATA1 specific and GATA2 specific groups (median value=1.75). We found that none of the 1000 permutation samples (empirical p-value=0.001) had a higher difference. Similarly we performed the permutation analysis in the GATA2 specific and the GATA1/GATA2 overlapping groups. We again found none of the 1000 subsamples with a higher difference (median value=1.75). The median value of the GATA2 specific group is 1.45.

2.4.10 Binding models of the purified proteins and their protein complexes

In order to create the binding models of Figures 34-36, 41, 49-54, 56, 57, 64 and 65 we used the following approach. For example for Figure 34 we selected from the enriched mGATA1_{NC} 8mers, the 8mers that contained initially the “GATA” motif (“GATA” containing 8mers) and then the 8mers that contained the “GATC” motif (“GATC” containing 8mers). We then used the sequenced 20 bp reads of the bound DNA library (20mer reads). We found 20mer reads that contained at least one 8mer from the “GATA” containing 8mers (20mer “GATA” group). Next from the 20mer “GATA” group we found the reads that contained at least one 8mer from the “GATC” containing 8mers (20mer “GATA-GATC” group). Afterwards, in the 20mer “GATA-GATC” group we found the distance between a “GATA” motif and its closest “GATC” motif. Lastly, we created the bar plot of Figure 34. For the background distribution we performed the same approach in the original 20mer degenerate DNA library.

2.5 FOG1 ChIP-seq

2.5.1 ChIP-seq of FOG1 from murine erythroleukemia cells

Murine erythroleukemia (MEL) cells and MEL cells with biotinylated FOG1S_{Met82} (bioFOG1S_{Met82}) (translational isoform (S-short form) that starts from the methionine which appears in position 82 in the wt-mFOG1 (Snow and Orkin, 2009)) were cultured as previously described (Weiss et al., 1997, Woo et al., 2008). The biotinylation tag was introduced in the C-terminus of FOG1S_{Met82} and the biotin labelling was performed as described in (Yu et al., 2009). For erythroid differentiation, induction of MEL cells was done by adding 1.7% DMSO (Sigma) (final concentration) to the medium for 24 hr. The cells were cross-linked with EGS (Ethylene glycol-bis(succinic acid N-hydroxysuccinimide ester)) (Sigma). Streptavidin pull down assays were performed on chromatin prepared from DMSO-induced MEL cells as described (de Boer et al., 2003). Streptavidin immunoprecipitated material from untransfected DMSO-induced MEL cells was used as control (BirA sample). All the samples and their respective controls were kindly provided to us by Dr. John Strouboulis (Alexander Fleming BSRC, Athens, Greece). The material was processed for Illumina high-throughput sequencing.

2.5.2 Data processing of the FOG1 ChIP-seq

Sequences were mapped to the non repeat-masked reference mouse (build m37) genome using BOWTIE (Langmead et al., 2009). We accepted uniquely mapped reads with maximum 2 mismatches in the first 28 bases and total Phred quality score of

mismatches equal or less than 70. In vivo DNA bound regions (ChIP-seq peaks) were found by running MACS (Zhang et al., 2008) on each ChIP-seq sample against the matching control. DNA binding positions with a false discovery rate (FDR) of <1% were selected. For the $\text{bioFOG1S}_{\text{Met82}}$ ChIP-seq, peak finding was carried out on each ChIP-seq sample against the matching BirA sample. Because the BirA ChIP-seq has a high number of peaks that arise from biotinylated proteins other than the $\text{bioFOG1S}_{\text{Met82}}$, the empirical FDR that MACS uses will overestimate the FDR. For this reason we found GATA1 binding positions using cisGenome (Ji et al., 2008) which uses an FDR calculated from a binomial distribution fitted on genomic windows with small number of sequencing reads. Again, FDR of <1% was selected. The adequacy of the FDR cut-off was examined by visual inspection of the peaks on the UCSC browser. For the $\text{bioFOG1S}_{\text{Met82}}$ ChIP-seq 473 ChIP-seq peaks with FDR<1% were identified (In a similar FOG1 ChIP-seq dataset, with DMSO induced MEL cells around 800 peaks were identified, Frank Grosveld personal communication).

2.6 Analysis of publicly available ChIP-seq datasets

For the ChIP-seq datasets presented in Figures 94, 97, 99 and 103 the number of ChIP-seq reads that were crossing the 400bp region, with the TSS in the middle, was calculated as the maximum signal that appears in a wig file (genomic signal in the region/vertical concatenation of all the reads, that cross the region, presented as peak height) (<http://genome.ucsc.edu/goldenPath/help/wiggle.html>) of the presented datasets for the specific genomic regions. For Figures 79-83 the number of sequenced reads that were crossing the 400bp region, with the ChIP-seq peak centre in the middle, was calculated similarly. The sequenced tags of the public datasets were used with the given genomic coordinates otherwise in case of un-mapped data, the mapping was done similarly with section 2.5.2. In case of processed data the genomic coordinates were transformed, when

necessary, in coordinates of another genome assembly with the LiftOver tool (<http://genome-mirror.moma.ki.au.dk/cgi-bin/hgLiftOver>).

CHAPTER 3 RESULTS

In Figure 9 we present the process that we are following in the next chapters in order to produce the DNA binding specificities/affinities of the examined proteins and their extrapolation on in vivo derived datasets. In more detail, we initially expressed and purified the proteins and the protein complexes of interest. We then performed EMSA-seq with a 20mer degenerate DNA library in multiple concentrations of protein against a fixed amount of DNA, depending on the available quantity of the expressed protein. Afterwards, we extracted the DNA binding models (specificities and/or affinities) from the protein bound DNA sequences, depending on the experimental design and the examined protein. We then projected the DNA binding specificities and/or affinities in genomic positions, where the examined proteins have been found to bind in vivo, in order to investigate their relationship with the presence of different aspects of the transcriptional regulation for example histone post-translational modifications, nucleosome occupancy levels, DNase I hypersensitivity levels and differentially expressed genes during erythropoiesis.

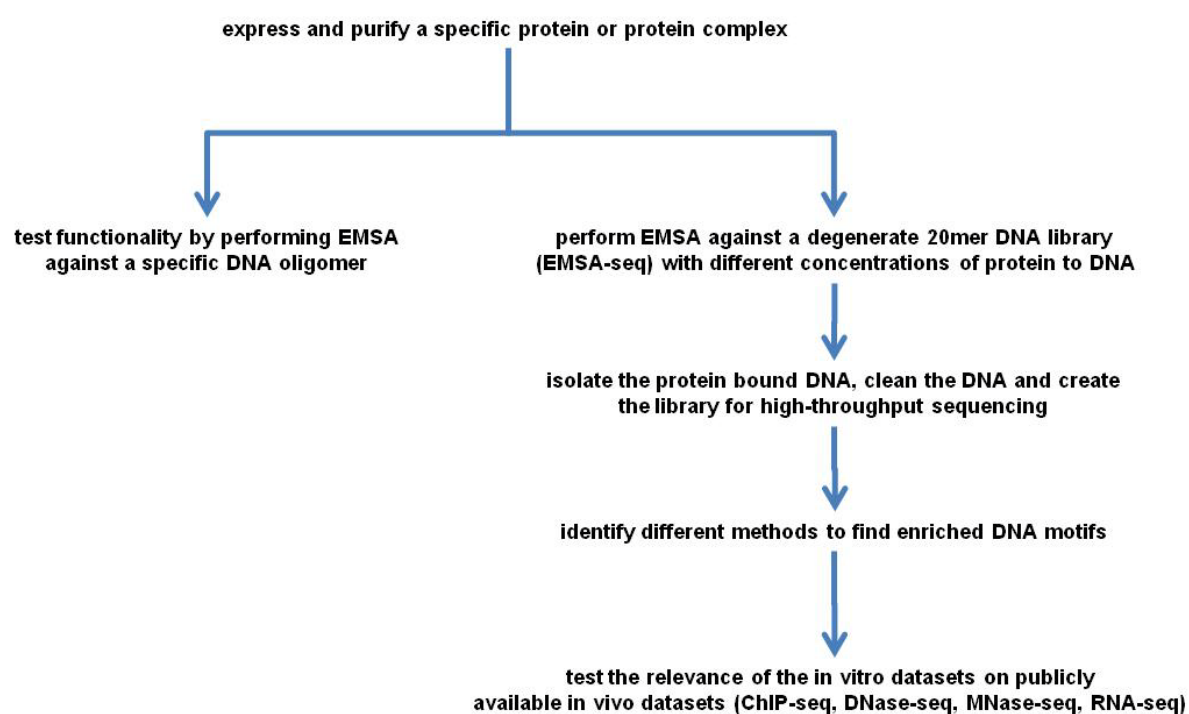


Figure 9. The scheme shows the process followed in this study in order to acquire in vitro derived DNA binding affinities/specificities of specific proteins or protein complexes and the subsequent integration with in vivo data.

3.1 Expression, purification and DNA band shifts of GATA1 and KLF1 proteins and of the GATA1/FOG1, GATA1/KLF1 complexes

At first we expressed the proteins and the protein complexes in vitro. Subsequently we performed EMSA-seq. We then analyzed the resulting sequences in order to infer binding specificities and other kinetic properties characteristic of the DNA binding.

3.1.1 Creation of fusion proteins with specific purification tags

We created fusion proteins and protein fragments with either the poly-histidine tag (His-tag) or both the His-tag and the FLAG peptide tag (FLAG-tag) for the following GATA1 related proteins, the full length mouse GATA1 protein (aa 1-413) (wt-mGATA1), mouse GATA1 with only the N-terminal Zn finger and the C-terminal Zn finger (aa 200-318) (mGATA1_{NC}), mouse GATA1 with only the C-terminal Zn finger (aa 258-282) (mGATA1_C) (Figure 10A). We purified the wt-mGATA1 from a baculovirus expression system whereas we purified the mGATA1_{NC} and the mGATA1_C from a bacterial expression system.

We additionally created fusion proteins and protein fragments with either the FLAG-tag or both the hemagglutinin tag (HA-tag) and the FLAG-tag for the following FOG1 related proteins, full length mouse FOG1 (aa 1-995) (wt-mFOG1), the mouse FOG1 Zn finger 1 (aa 241-295) (mFOG1_{finger1}), the mouse FOG1 Zn finger 5/6 (aa 563-730) (mFOG1_{finger56}), the mouse FOG1 Zn finger 9 (aa 945-995) (mFOG1_{finger9}) (Figure 10B). The wt-mFOG1 in complexation with wt-mGATA1 or alone was purified from a baculovirus expression system. The mFOG1_{finger1}, mFOG1_{finger56}, mFOG1_{finger9} in complexation with mGATA1_{NC} or alone were purified from a bacterial expression system.

We also created a mouse KLF1 protein (aa 257-358) (mKLF1₂₅₇₋₃₅₈) fused to a His-tag and a FLAG-tag (Figure 10C). This region spans the 3 Zn fingers of the protein. The mKLF1₂₅₇₋₃₅₈ was expressed in a bacterial expression system alone or in complexation with mGATA1_{NC}. We were not able to express full length KLF1 in the baculovirus expression system.

The specificity of the complexation between mGATA1_{NC} and the FOG1 Zn fingers has been shown previously. Briefly, the mFOG1 Zn finger 1, mFOG1 Zn finger 6 and mFOG1 Zn finger 9 have been shown to bind specifically with the mGATA1 N-terminal Zn finger (aa 200–254) using a Yeast Two-Hybrid screen and GST pull down assays (Tsang et al., 1997, Fox et al., 1998, Fox et al., 1999).

We also aimed to check the specificity of the wt-mGATA1/wt-mFOG1 complex by co-expressing the V205G mutant of mGATA1 (mGATA1^{V205G}) with the wt-mFOG1. The mGATA1^{V205G} cannot bind wt-mFOG1 (Crispino et al., 1999). The protein expression levels of mGATA1^{V205G} were very low which did not permit us to check the complexation under the same expression and purification conditions as the wt-mGATA1. We also tried to reconstitute the complex in vitro by co-incubating wt-mFOG1 and wt-mGATA1 or mGATA1_{NC}. In either case, we could not see a band shift given from the complex itself (data not shown). We interpret this result as an indication that the complex is specific.

The specificity of mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ fingers has been shown with GST pull down assays (Merika and Orkin, 1995). In more detail, the mKLF1 Zn fingers (aa 286-376) have been shown to bind specifically with the mGATA1 C-terminal Zn finger (aa 229–333).

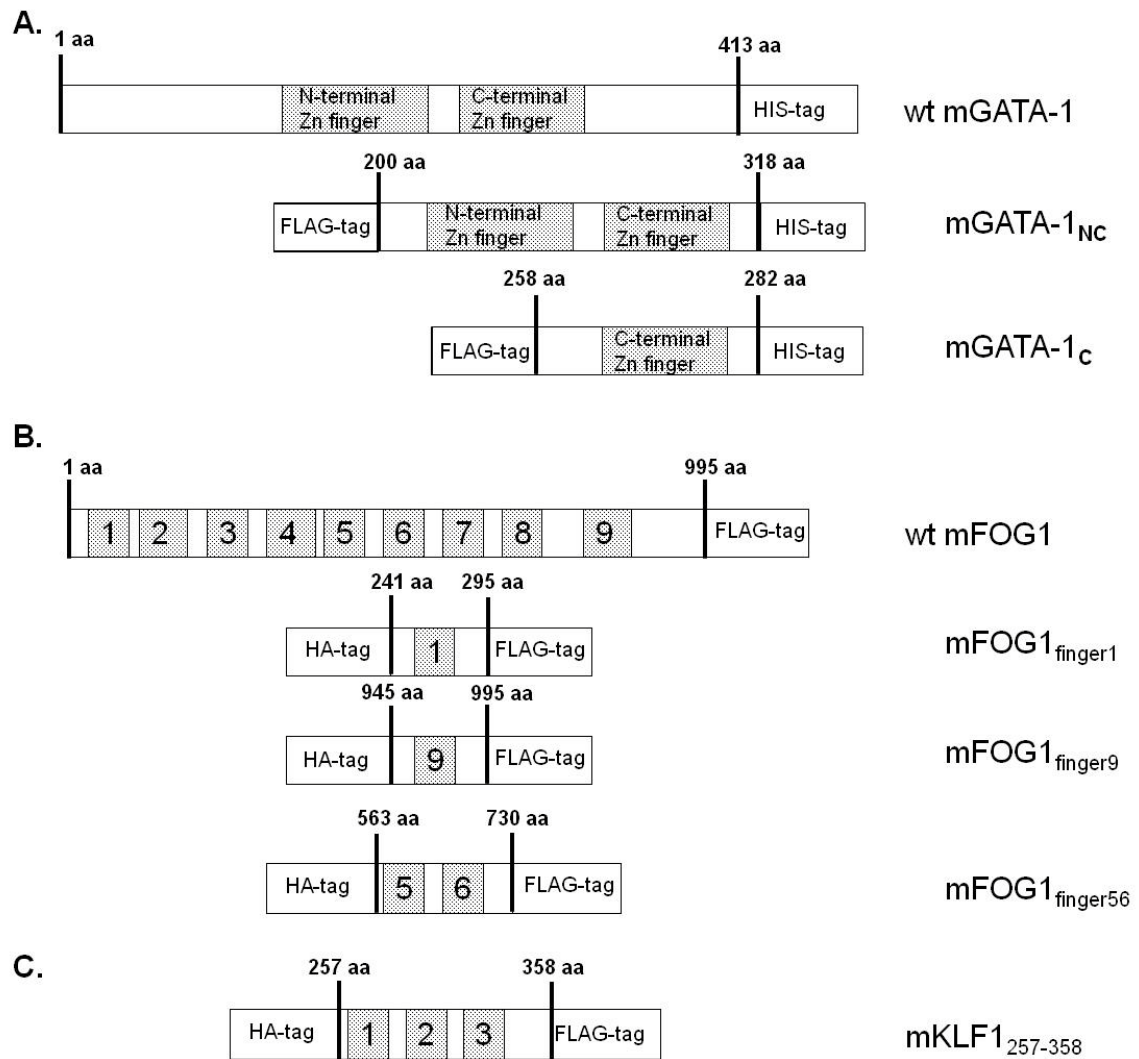


Figure 10. Full length proteins and protein fragments that were used in this study. The boxes with the black dots represent the zinc fingers. The tags that were used for purification are shown for each protein. The numbers in the beginning and end of each protein fragment represent the amino acids that correspond to the wild type protein. The different groups of protein and protein fragments, presented in the figure, consist of the mGATA1 group (A), the mFOG1 group (B) and the mKLF1 protein fragment (C).

3.1.1.1 Kinetics of protein production of wt-GATA1 and wt-GATA1/wt-FOG1 in the baculovirus expression system

The optimal levels of protein expression of wt-mGATA1 and wt-mGATA1/wt-mFOG1 were identified as described below. It is important to stress out that the wt-mGATA1 protein is produced at low levels whereas the wt-mFOG1 protein is produced in large amounts. For this reason, when we expressed the wt-mGATA1/wt-mFOG1 complex, we focused in optimising the protein expression of the wt-mGATA1 protein.

At first we calculated the growth curve of the SF9 cells along a course of 4 days. In Figure 11A we see a duplication time of 24 hours approximately. Starting at 1×10^6 cells/ml we reach after 3 days at approximately 6×10^6 cells/ml. Because, at this cell density, the cells are already overcrowded and stressed, we restricted our protein production at maximum 3 days.

We then diluted the viral titer 10 , 10^2 , 10^3 , 10^4 , 10^5 times in order to find the minimum amount of viral particles that produces the same or more protein as the undiluted titer after 3 days of protein production (Figure 11B). We concluded that dilution of the initial viral titer 10^2 times is the optimal one. The same results were taken for both the wt-mGATA1 and the wt-mGATA1/wt-mFOG1 complex.

We then checked the time-course of protein expression during the 3 days period. A representative image for the wt-mGATA1/wt-mFOG1 complex is shown in Figure 11C. We can see that, in the total protein lysate, the protein expression levels of the complex are highest after 3 days of protein production. On the contrary, the protein expression levels of wt-mGATA1 are highest after 2 days and then they decline (Figure 11D).

Based on the above we selected 3 days of protein expression for the wt-mGATA1/wt-mFOG1 complex and 2 days of protein expression for the wt-mGATA1 protein alone at

the pre-specified titer. We also notice, in the total lysate (Figure 11C), that after 3/2 days of protein production we have the appearance of degradation fragments but the purification procedure removes the majority of them (Figure 14B).

In case where we purified the mFOG1 protein alone from the SF9 cells, the optimisation of the expression was done similarly as the mGATA1 protein expression (data not shown).

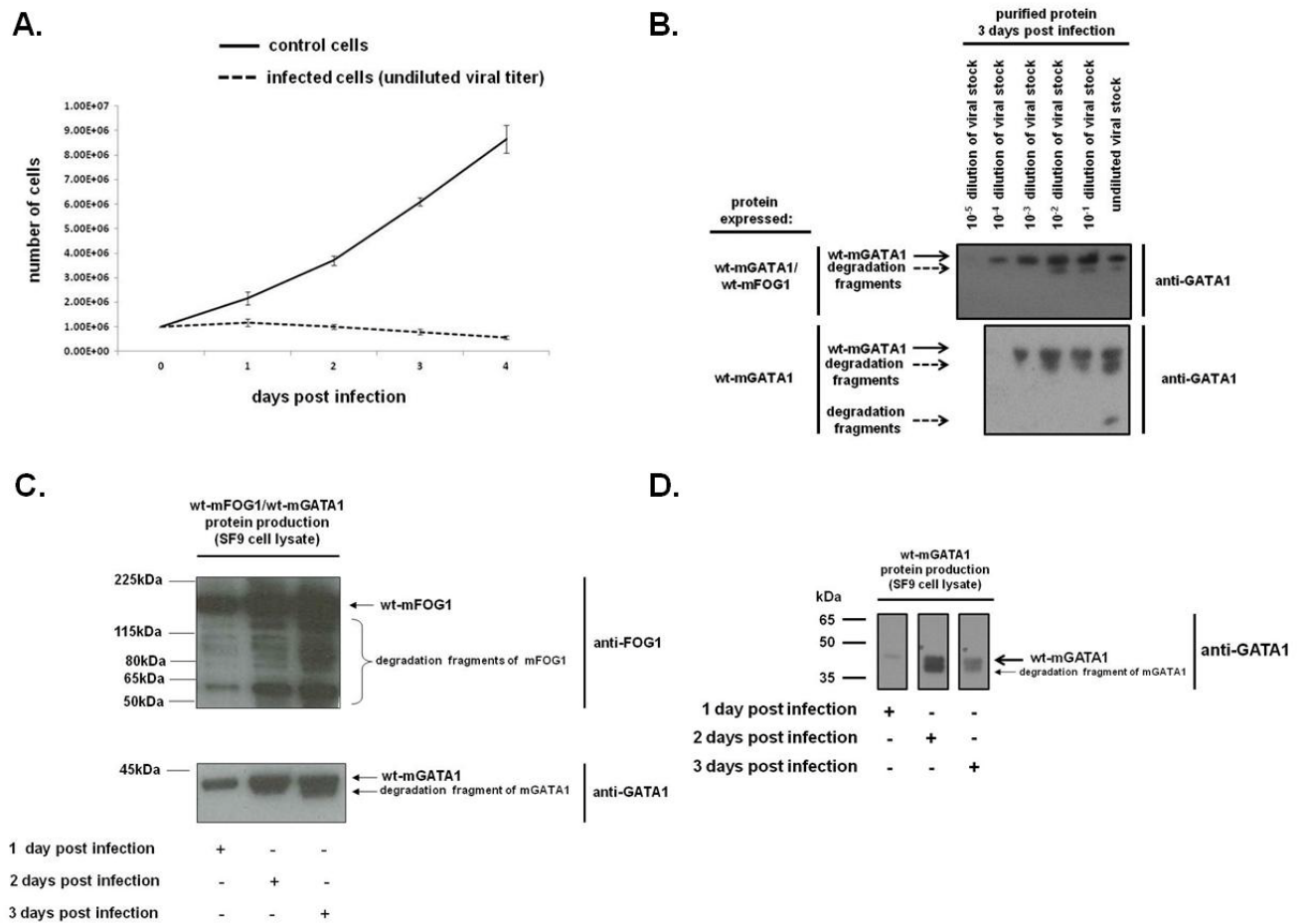


Figure 11. Protein expression in the baculovirus expression system. A) Growth curve of SF9 cells in the presence and absence of undiluted wt-mGATA1 viral titer (starting concentration 1×10^6 cells/ml). After 3 days of growth the SF9 cell concentration has increased to 6×10^6 cells/ml for the control cells and it has decreased to 0.7×10^6 cells/ml for the infected cells with the undiluted wt-mGATA1 viral titer. The growth curves of SF9 cells for the diluted viral titer lie between the two presented curves. B) Expression levels of wt-mGATA1 when it is expressed alone or in complexation with wt-mFOG1. The SF9 cells were expressing the wt-mGATA1/wt-mFOG1 complex or the wt-mGATA1 protein alone for 3 days. The SF9 cells were infected with different viral titers as indicated. In each condition the protein was purified with a two step purification protocol as described in the materials and methods section. C) Expression levels of wt-mFOG1 and wt-mGATA1 in the

total lysate of SF9 cells after 1, 2, 3 days post infection. The cells were expressing the wt-mGATA1/wt-mFOG1 complex. The viral titer used corresponds to the 10^{-2} dilution of the viral stock titer. D) Expression levels of wt-mGATA1 in the total lysate of SF9 cells after 1, 2, 3 days post infection. The cells were expressing the wt-mGATA1 protein alone. The viral titer used corresponds to the 10^{-2} dilution of the viral stock titer.

3.1.1.2 Protein production in the bacterial expression system

The mGATA1_{NC}, mGATA1_C, mGATA1_{NC} /mFOG1_{finger1}, mGATA1_{NC} /mFOG1_{finger9}, mGATA1_{NC} /mFOG1_{finger56}, mKLF1₂₅₇₋₃₅₈ proteins were all expressed in bacterial cells (as described in detail in the material and methods section). Briefly, the bacterial cells were incubated for 5 hours at 37°C after IPTG induction (Figure 12C). We also tested incubation at 25°C for 5 hours but the amount of produced protein was lower (data not shown).

3.1.1.3 Purification of proteins from the baculovirus or the bacterial expression system

We used the following purification schemes for the isolation of the proteins (Figure 12):

1. For the isolation of the wt-mGATA1 protein, in the baculovirus expression system, the first step of purification was against the 6x-HIS tag with Ni-NTA beads and the second step of purification was with DNA affinity chromatography (as described in the materials and methods section) (Figure 12A).
2. For the isolation of mGATA1_{NC}, mGATA1_{CF} and mKLF1₂₅₇₋₃₅₈ proteins in the bacterial expression system, the first step of purification was against the FLAG-tag with anti-FLAG beads and the second step of purification was against the 6x-HIS tag with Ni-NTA beads (as described in the materials and methods section) (Figure 12C).
3. For the isolation of mGATA1_{NC}/mFOG1_{finger1}, mGATA1_{NC}/mFOG1_{finger56} and mGATA1_{NC}/mFOG1_{finger9} complexes, in the bacterial expression system, the first step of purification was against the FLAG-tag of mFOG1 fragments with anti-FLAG

beads and the second step of purification was against the 6x-HIS tag of the mGATA1_{NC} with Ni-NTA beads (Figure 12C).

4. For the isolation of the mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex, in the bacterial expression system, the first step of purification was against the FLAG-tag of mKLF1₂₅₇₋₃₅₈ with anti-FLAG beads and the second step of purification was against the 6x-HIS tag of the mGATA1_{NC} with Ni-NTA beads (Figure 12C).
5. For the isolation of the wt-mGATA1/wt-mFOG1 complex three alternative schemes were tested (Figure 12B):
 - i) The first step of purification was against the 6xHIS tag of mGATA1 with Ni-NTA beads and the second step of purification was with DNA affinity chromatography (NiNTA/DNA oligo approach).
 - ii) The first step of purification was against the 6xHIS tag of mGATA1 with Ni-NTA beads and the second step of purification was against the FLAG-tag of mFOG1 with anti-FLAG beads (NiNTA/anti-FLAG approach).
 - iii) The first step of purification was against the FLAG-tag of mFOG1 with anti-FLAG beads and the second step of purification was with DNA affinity chromatography (anti-FLAG/DNA oligo approach).

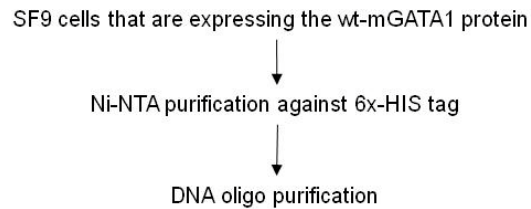
As seen in Figures 15,16 and 13B we successfully isolated the mGATA1_{NC}, mGATA1_{CF} and wt-mGATA1 proteins and the mGATA1_{NC}/mFOG1_{finger1}, mGATA1_{NC}/mFOG1_{finger56}, mGATA1_{NC}/mFOG1_{finger9}, mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complexes. All the proteins were functional because they were able to shift a “GATA” motif containing DNA sequence (Figure 13A for the wt-mGATA1 protein, for the rest of the proteins data are not shown).

In Figure 14B we see that for the wt-mGATA1/wt-mFOG1 complex, the anti-FLAG/DNA oligo approach did not isolate the complex. The NiNTA/anti-FLAG approach isolated the

complex but was not functional because no DNA band shift was detected during the EMSA experiments with either a “GATA” motif containing DNA sequence (Figure 14A) or the 20mer degenerate DNA libraries (data not shown). Thus we selected the NiNTA/DNA oligo approach that isolated the protein complex (Figure 14B) and that was giving a DNA band shift (Figure 14A).

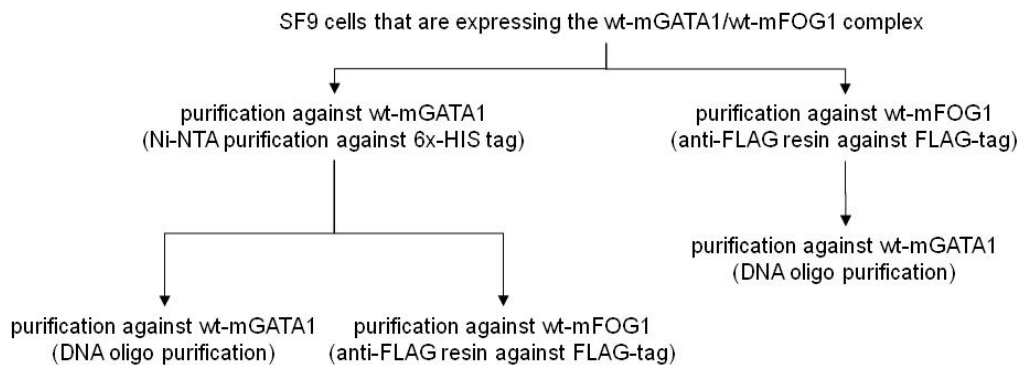
A.

wt-mGATA1 protein purification scheme



B.

wt-mGATA1/wt-mFOG1 complex purification scheme



C.

**purification scheme of the following proteins and protein complexes:
mGATA1_{NC}, mGATA1_C, mKLF1₂₅₇₋₃₅₈, mGATA1_{NC}/mFOG1 fingers, mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈**

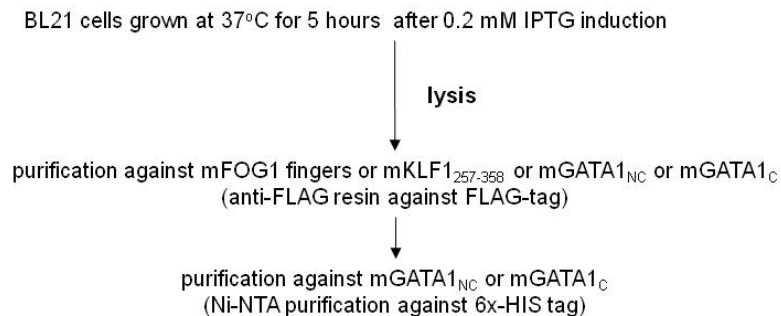


Figure 12. A) Purification scheme of the wt-mGATA1 protein from the baculovirus expression system. B) Purification scheme of the wt-mGATA1/wt-mFOG1 complex from the baculovirus expression system. Three different schemes were tested. C) Purification scheme of the presented protein fragments and protein complexes from the bacterial expression system.

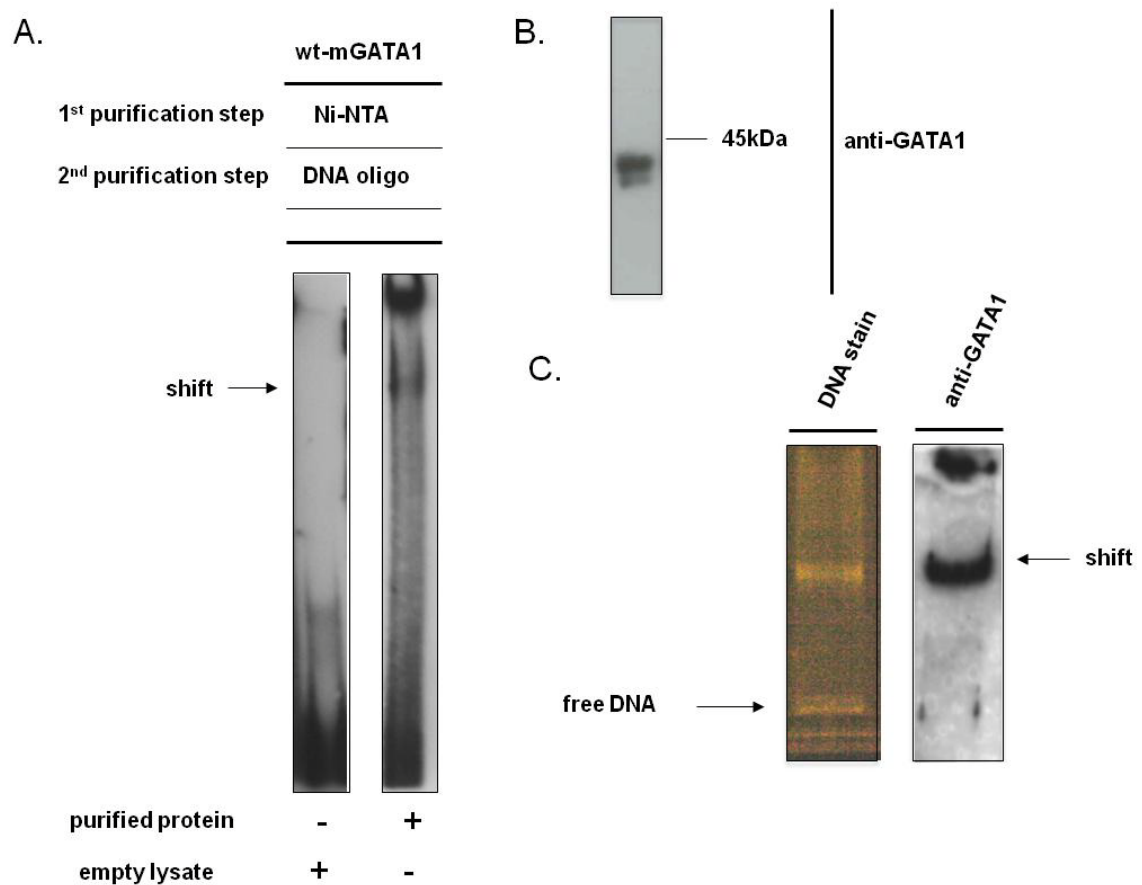
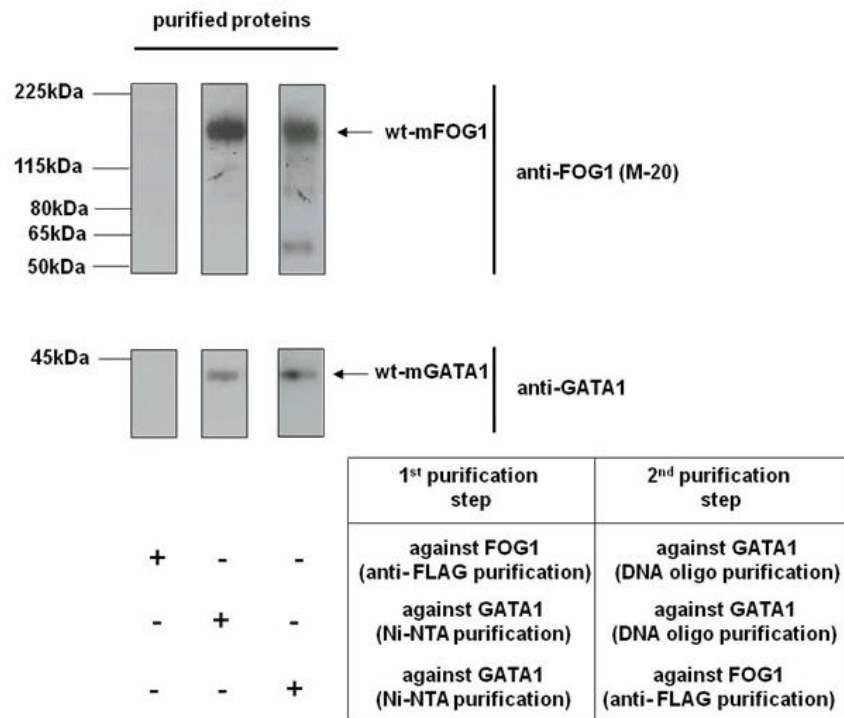
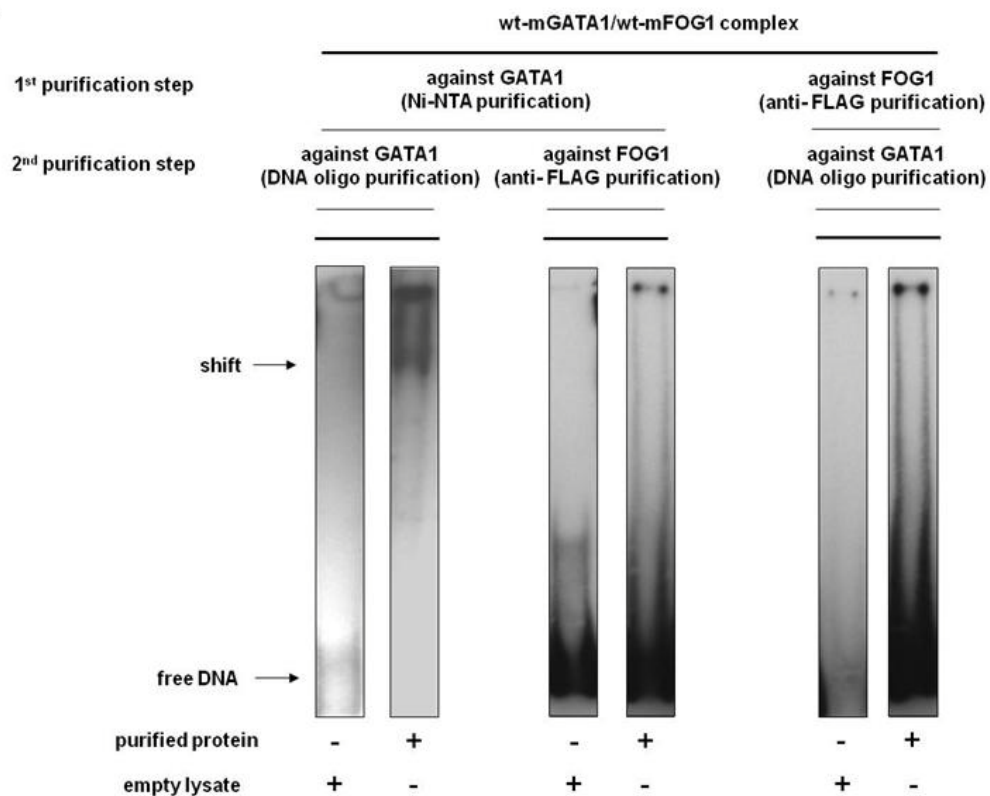


Figure 13. Purification of wt-mGATA1. The wt-mGATA1 protein was successfully purified (B). The functionality of wt-mGATA1 was shown with a band shift against a “GATA” sequence containing oligo (lane 2 on A). SF9 lysate that was not expressing wt-mGATA1 and that was used in the same purification procedure as the wt-mGATA1 containing lysate did not show any shift (lane 1 on A). Wt-mGATA1 was also able to shift the 20mer degenerate DNA library (C). The specificity of the band shift was shown with the EMSA-Western technique (see materials and methods section) that was able to detect GATA1 on the band shift (C).

A.



B.



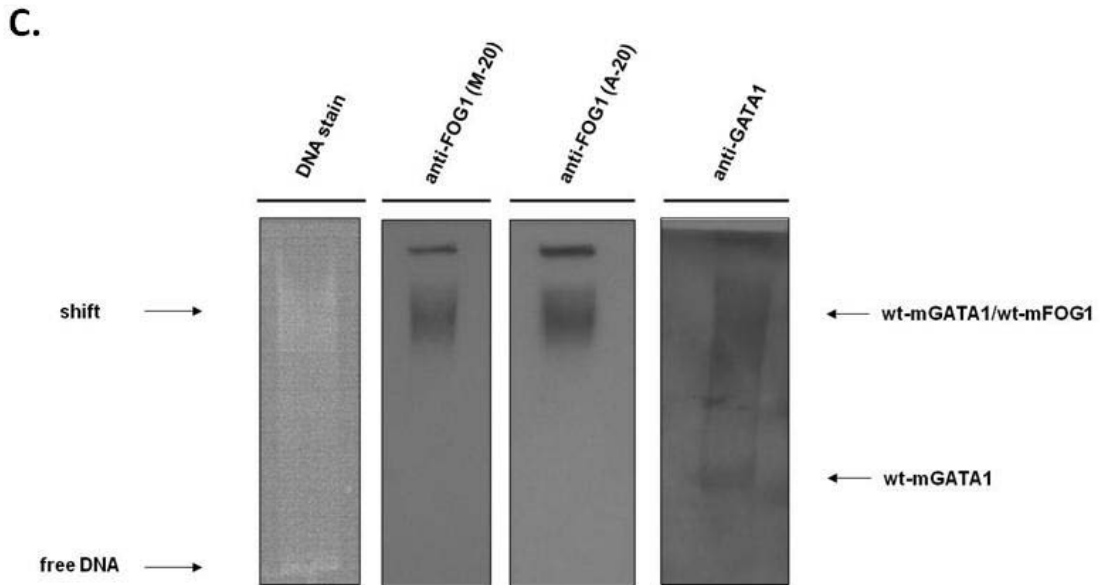


Figure 14 (previous and current page). The complex wt-mGATA1/wt-mFOG1 was successfully purified with the NiNTA/DNA oligo approach and NiNTA/anti-FLAG approach but not the anti-FLAG/DNA oligo approach (see text for details) (A). The functionality of the wt-mGATA1/wt-mFOG1 complex was shown with a band shift against a “GATA” sequence containing oligo (B). Only the NiNTA/DNA oligo approach gave a functional complex. SF9 lysate that was not expressing the complex and that was used in the same purification procedure as the complex containing lysate did not show any shift (B). The complex wt-mGATA1/wt-mFOG1 was also able to shift the 20mer degenerate DNA library (C). The specificity of the band shift was shown with the EMSA-Western technique (see materials and methods) that was able to detect mGATA1 and mFOG1 on the band shift (C). mFOG1 was detected with two antibodies. The “M-20” anti-FOG1 ab binds near the N-terminus of mFOG1 whereas the “A-20” anti-FOG1 ab binds near the C-terminus of mFOG1 indicating that the complex with full length mFOG1 binds the shifted DNA.

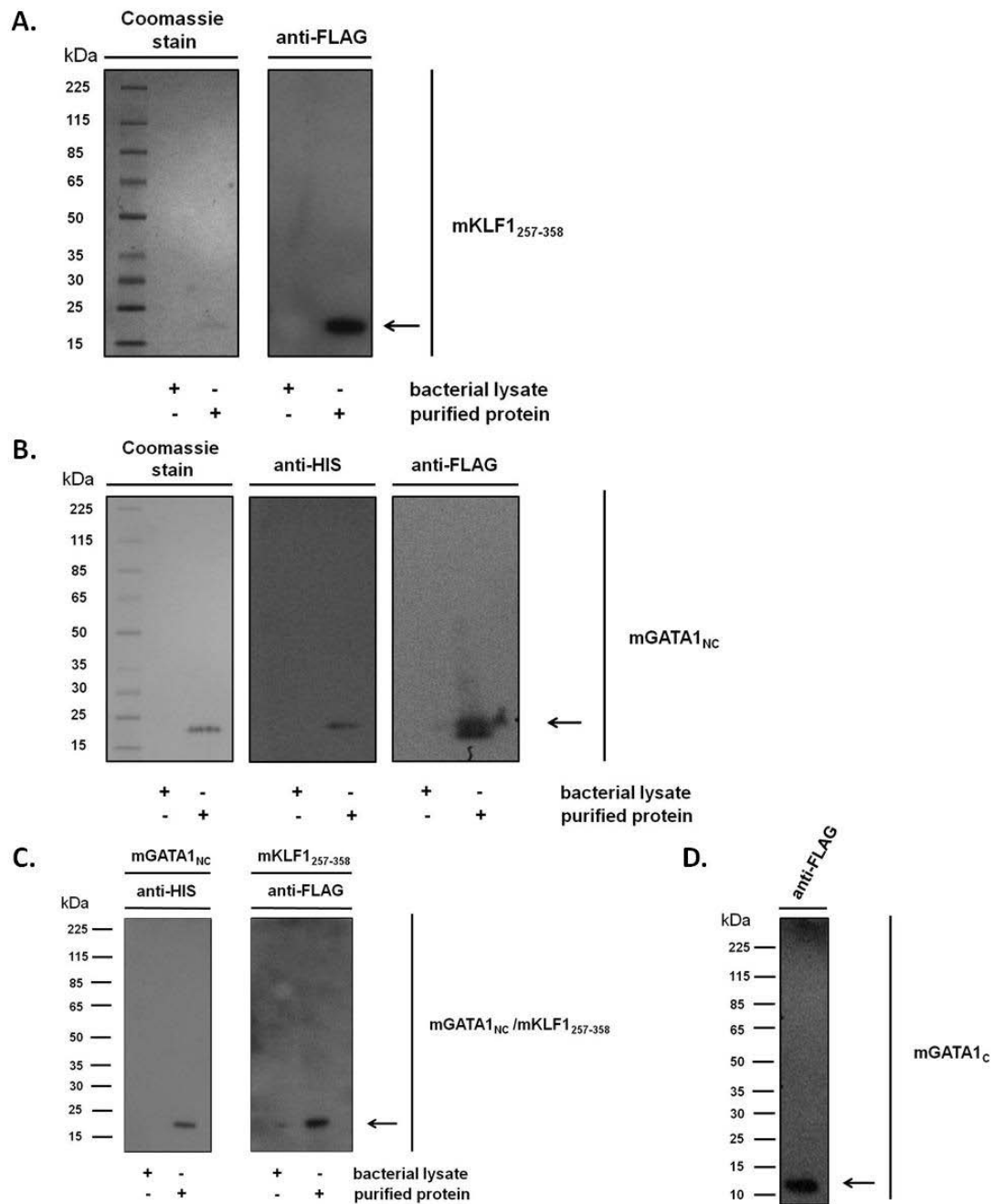


Figure 15. Purification of bacterially produced mGATA1_{NC}, mGATA1_C, mKLF1₂₅₇₋₃₅₈, mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈. The gel with the Coomassie stained protein and the western blot against the FLAG-tag is shown for mKLF1₂₅₇₋₃₅₈ (A). The gel with the Coomassie stained protein and the western blot against the FLAG-tag and the His-tag are shown for mGATA1_{NC} (B). The western blot against the FLAG-tag and the His-tag is shown for the mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex (C). The western blot against the FLAG-tag for the mGATA1_C protein is shown (D).

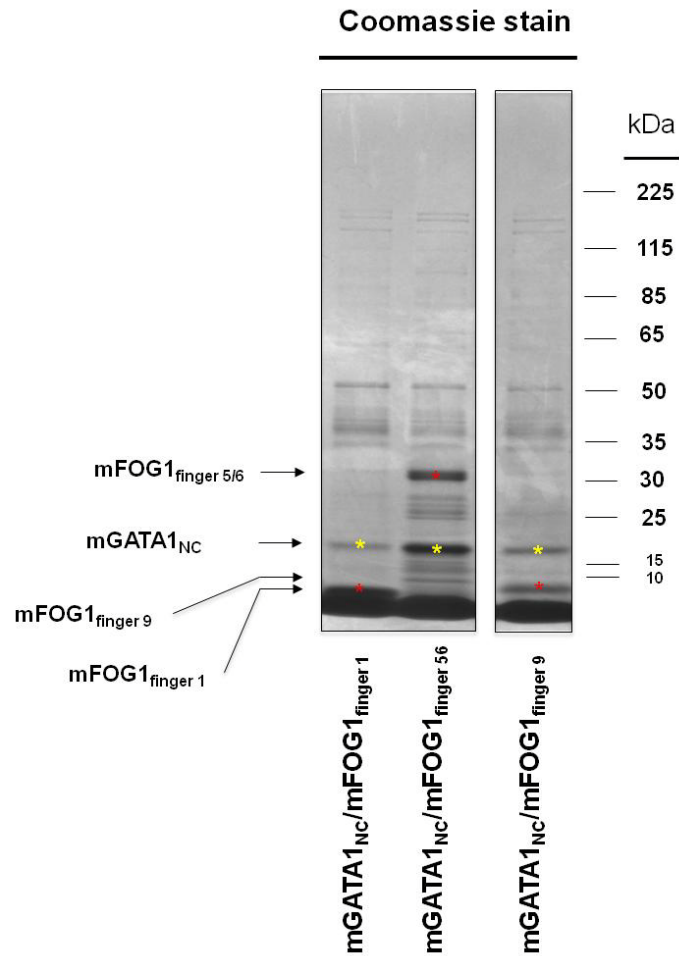


Figure 16. Purification of bacterially produced mGATA1_{NC}/mFOG1_{finger1}, mGATA1_{NC}/mFOG1_{finger56}, mGATA1_{NC}/mFOG1_{finger9} complexes. The Coomassie stained gel is shown. The complexes were purified with one step purification against the FLAG-tag of the FOG1 fragments. The same fraction of eluate was loaded on every lane. The position of mGATA1_{NC} is shown (yellow asterisk). The position of FOG1 fragments is also shown (red asterisk). We also notice that in the mGATA1_{NC}/mFOG1_{finger56} lane both proteins mGATA1_{NC} and mFOG1_{finger56} are presented in equal amounts which indicates that the complexation is favoured. The same for the mGATA1_{NC}/mFOG1_{finger9} complex. On the contrary in the mGATA1_{NC}/mFOG1_{finger1} the isolated mGATA1_{NC} is far less than the mFOG1_{finger1}. This indicates that the complexation is not favoured.

3.1.2 DNA band shift of a 20mer degenerate DNA library from the GATA1 and KLF1 proteins and the complexes GATA1/FOG1 and GATA1/KLF1

To identify the DNA binding landscape of the purified proteins we performed EMSA-seq experiments (Wong et al., 2011a) (Figure 7). Each protein was incubated with a specific amount of a 20mer degenerate DNA library.

In order to calculate dissociation constants (K_d), the EMSA-seq experiments for the mGATA1_{NC} and the mKLF1₂₅₇₋₃₅₈ proteins were performed on different protein/DNA ratios. Based on the availability of the protein we performed 9 different protein/DNA ratios for the mGATA1_{NC} and 6 different protein/DNA ratios for the mKLF1₂₅₇₋₃₅₈ (Figure 17).

For the complexes: wt-mGATA1/wt-mFOG1, mGATA1_{NC}/mFOG1_{finger1}, mGATA1_{NC}/mFOG1_{finger56}, mGATA1_{NC}/mFOG1_{finger9} and mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ we performed 1, 1, 3, 1 and 3 different protein/DNA ratios respectively but due to the limited amount of protein we were unable to calculate the binding fraction (Figure 17).

In Figures 18-20, 13C and 14C we can see the DNA band shifts of the 20mer degenerate DNA library for the wt-mGATA1, mGATA1_{NC}, mGATA1_C, mKLF1₂₅₇₋₃₅₈, wt-mGATA1/wt-mFOG1, mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈, mGATA1_{NC}/mFOG1_{finger1}, mGATA1_{NC}/mFOG1_{finger56} and mGATA1_{NC}/mFOG1_{finger9}.

In Figure 18 we notice that with increasing concentration of mGATA1_{NC} the band of the dimeric mGATA1_{NC} complex appears. Because not all the 20mer DNA library can be bound by two mGATA1_{NC} molecules, the creation of the trimeric protein-DNA complex depends on the motif content of the DNA sequence. The two mGATA1_{NC} molecules do not probably exist as a dimeric complex in solution (or they exist in very low concentration) but two independent mGATA1_{NC} molecules can bind simultaneously in permissive 20mer

sequences. Whether this close co-occurrence permits the creation of a true GATA1 dimeric complex is unknown. As expected at high protein concentration of mGATA1_{NC} the band of the free DNA is almost absent.

In Figure 20A-C we notice that the mGATA1_{NC}/mFOG1_{finger1}, mGATA1_{NC}/mFOG1_{finger56} and mGATA1_{NC}/mFOG1_{finger9} complexes give band shifts of the 20mer degenerate DNA library that migrate similarly with the band shifts from the mGATA1_{NC} alone. This indicates that either the DNA binding competes the protein-protein interactions of mGATA1 and mFOG1 fragments or that the migration inside the gel causes the complex to dissociate. As will be discussed later in the text, we believe that the DNA binding competes the mGATA1/mFOG1 interaction in vitro. On the other hand, we see that the wt-mFOG1/wt-mGATA1 gives a band shift as a complex (Figure 14C). This indicates that only the full length mFOG1 protein can stably interact with GATA1 when GATA1 binds DNA.

The mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ gives two band shifts the “upper” band (slow migrating band) that can be attributed to the intact mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex bound on the DNA and the “main” band (fast migrating band) that can be attributed to the dissociated mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex (mGATA1_{NC} bound on the DNA alone and mKLF1₂₅₇₋₃₅₈ bound on the DNA alone) (Figure 20D). In the EMSA-seq analysis that is presented in the next paragraphs we used both the slow and the fast migrating bands.

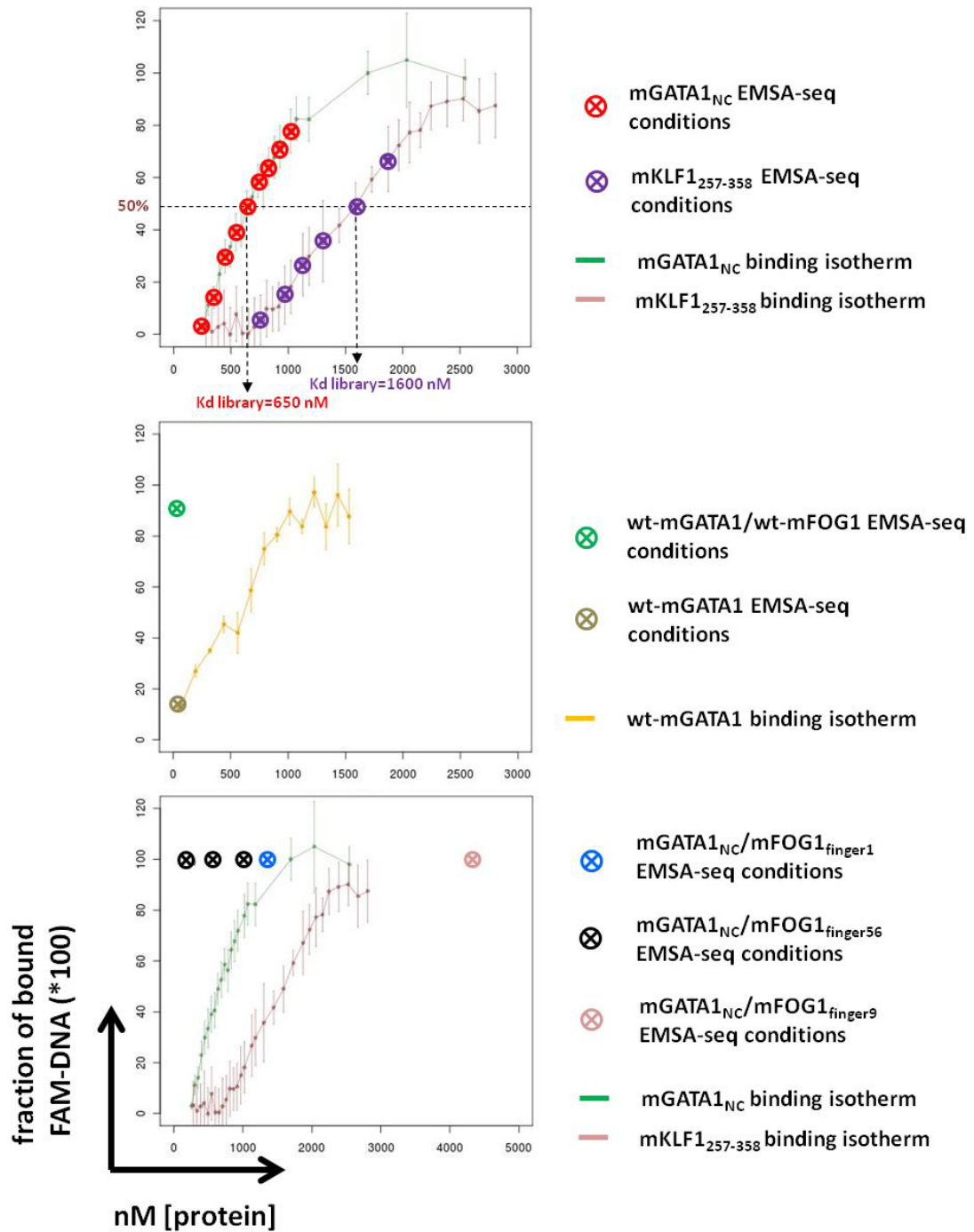
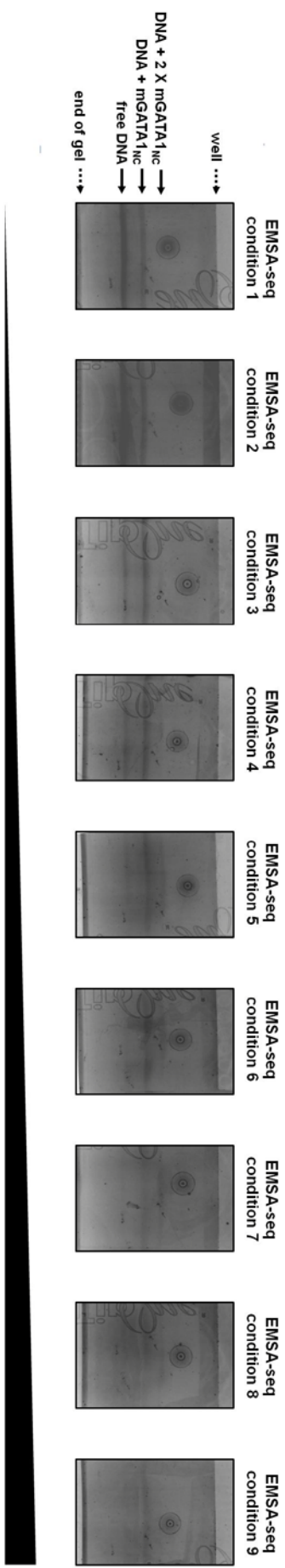


Figure 17. Experimental design and fraction (%) of the 20mer degenerate DNA library bound in each condition. The percentage of bound DNA library was assessed with fluorescence anisotropy (see materials and methods section). Binding isotherms were available only for the mGATA1_{NC}, mKLF1₂₅₇₋₃₅₈ and wt-mGATA1. The position of the other protein complexes (wt-mGATA1/wt-mFOG1, mGATA1_{NC}/mFOG1_{finger1},

mGATA1_{NC}/mFOG1_{finger56}, mGATA1_{NC}/mFOG1_{finger9}) indicates only the concentration of the protein used. Due to the competition between different DNA sequences for the same protein we do not expect the kinetic behaviour of the mGATA1_{NC} and mKLF1₂₅₇₋₃₅₈ to follow a one binding site model. The K_d of the bound libraries is the concentration of the protein where half of the DNA library is bound.



[mGATA1_{NC}]

Figure 18 (previous page). DNA band shifts of the 20mer degenerate DNA library for the mGATA1_{NC} protein. The nine EMSA-seq conditions are presented. The band shifts for protein-DNA complexes with one (DNA + mGATA1_{NC}) and two (DNA +2 X mGATA1_{NC}) mGATA1_{NC} molecules are presented. The pictures are representative of one replicate. The inverse of the original images is presented in order for the bands to be clearly visible. Although the exposure of the images and the image processing were done with the same way, and we used one lane gels during the experiments, we notice that at high protein concentration the intensity of the DNA band is becoming fainter. We attribute this to migration problems that have to do with the high protein concentration. The concentric cycles at the centre of the image and the letters at the edge have nothing to do with the experiment and are part of the imaging tray used during the image capture.

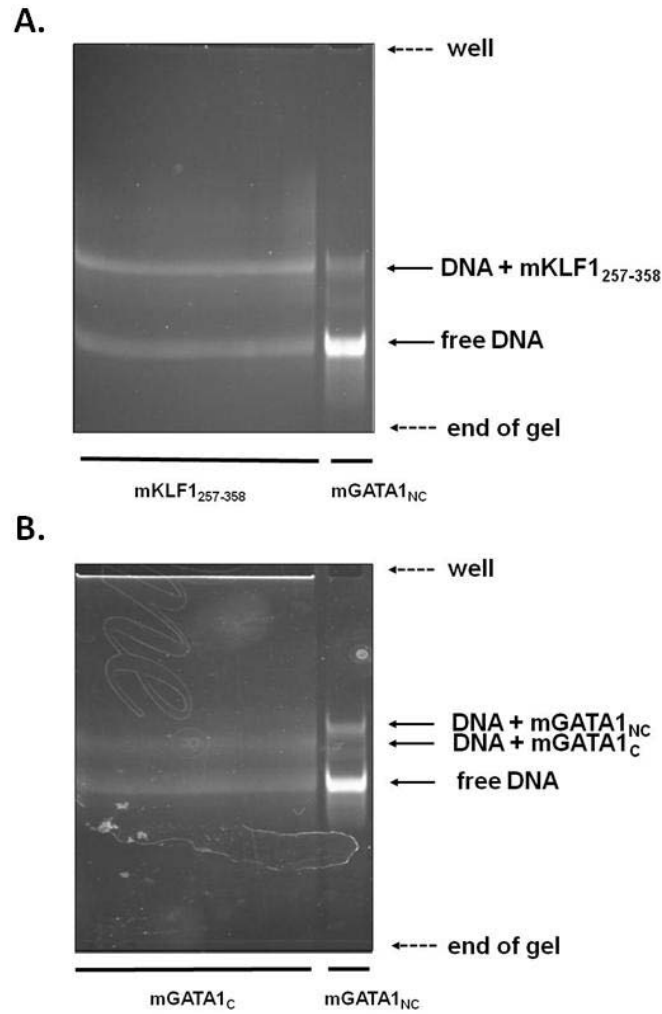


Figure 19. DNA Band shifts of the 20mer degenerate DNA library for the mKLF1₂₅₇₋₃₅₈ (A) and mGATA1_C (B) proteins. The DNA band shift from the mGATA1_{NC} protein is shown for comparison reasons. Due to the smaller MW of the protein, the DNA band shift from the mGATA1_C protein migrates faster than the mGATA1_{NC} one. The concentric cycles at the centre of image (B) and the letters at the edge have nothing to do with the experiment and are part of the imaging tray used during the image capture.

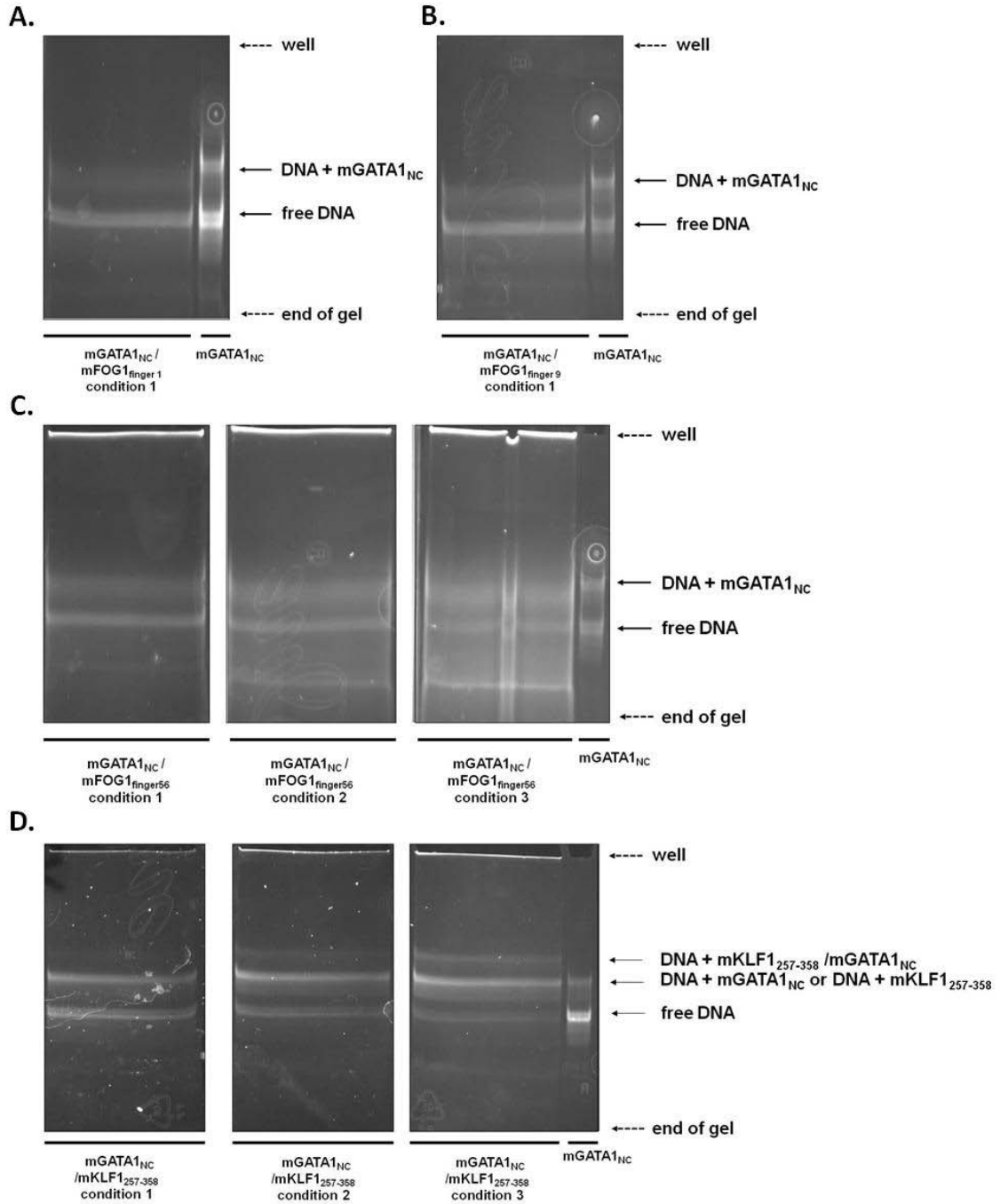


Figure 20. DNA Band shifts of the 20mer degenerate DNA library for the complexes mGATA1_{NC}/mFOG1_{finger1} (A), mGATA1_{NC}/mFOG1_{finger9} (B), mGATA1_{NC}/ mFOG1_{finger56} (C), and mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ (D). The DNA band shift from the mGATA1_{NC} is shown for comparison reasons. The three different conditions with increasing ratio of protein/DNA is shown for the mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ and mGATA1_{NC}/ mFOG1_{finger56} complexes. Except from the fast migrating band (“main” band), the slow migrating (“upper” band) (see text) of the mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex is visible in all the conditions. No other band except

the fast migrating band is visible for the mGATA1_{NC}/mFOG1_{finger1}, mGATA1_{NC}/mFOG1_{finger9}, mGATA1_{NC}/ mFOG1_{finger56} complexes in any of the conditions. We also notice that the migration of the shifted band is anomalous in the mGATA1_{NC}/mFOG1_{fingers} indicating probably a continuous competition between the FOG1 fingers and the DNA for the GATA1 binding. The concentric cycles at the centre of the image and the letters at the edge have nothing to do with the experiment and are part of the imaging tray used during the image capture.

At first we created a mechanistic model of the binding of a transcription factor on the DNA in an EMSA-seq reaction. This permitted us to calculate the dissociation constant “ K_d ” of a protein/DNA ligand interaction.

3.2 EMSA-seq can be used to calculate both the DNA binding specificities and the DNA binding affinities (dissociation constant- K_d) of a transcription factor

3.2.1 Non-cooperative competition model in an EMSA-seq reaction

We modeled the binding behavior of the protein against a degenerate 20mer DNA library as a combination of independent binding sites that compete with each other for the same amount of protein (multiple equilibria binding kinetics and competition kinetics) (Figure 21).

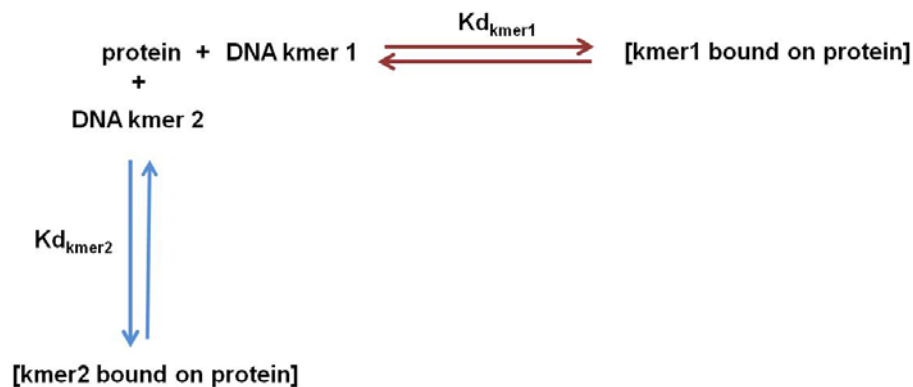


Figure 21. Modelling the binding of the DNA sequences in an EMSA-seq experiment. The model shows the competition between each k-mer for binding the specific amount of protein in the EMSA-seq experiment. The example of two k-mers is presented.

Experimentally, we titrated the protein against a fixed amount of the 20mer DNA library (Figures 17 and 18). We monitored the fraction of the bound DNA library with fluorescence anisotropy, as described in the material and methods section. Then, for each binding isotherm, we calculated the amount of total DNA molecules bound on the different protein/DNA ratio conditions. By taking into account:

- i) The fraction of each individual k-mer in the total number of k-mers from the sequenced DNA libraries of the different protein/DNA ratio conditions .
- ii) The total DNA molecules bound on the different protein/DNA ratio conditions.

we calculated the percentage of each individual DNA k-mer bound in each condition. We then used these numbers in the thermodynamic model and we calculated the K_d for each individual k-mer.

Below we model the binding of the protein with the 32896 different 8mers of the 20mer library (65536 different 8mers in total that are collapsed in 32640 reverse complement 8mers and 256 palindromic 8mers). The modeling is similar for k-mers of different length.

In more detail, we hypothesized that the binding behavior of the DNA library in each condition will be the sum of all the independent binding events. We express this as follows:

$$r_{\text{total}} = r_{\text{k-mer1}} + r_{\text{k-mer2}} + \dots + r_{\text{k-merN}} \quad (\text{eq 1})$$

where

r_{total} is the fraction of all the DNA library bound as calculated from the molecular anisotropy experiments, $r_{\text{k-merN}}$ for $N=\{1, 2, \dots, 32896\}$ is the fraction of each individual k-mer bound (in this case each k-mer is one of the 32896 available 8mers).

The above equation can be represented spectroscopically as shown in Figure 22.

In each different protein/DNA ratio condition each k-mer is competing with the other k-mers to bind the protein (Figure 21). For example, for each of the two k-mers of Figure 21, the binding affinities are expressed by the dissociation constants **Kd_{k-mer1}** and **Kd_{k-mer2}** as follows:

$$\mathbf{Kd_{k-mer1} = \frac{[\text{protein not bound on any k-mer}][\text{free k-mer1}]}{[\text{k-mer1 bound on protein}]}} \quad (\text{eq 2})$$

$$\mathbf{Kd_{k-mer2} = \frac{[\text{protein not bound on any k-mer}][\text{free k-mer2}]}{[\text{k-mer2 bound on protein}]}} \quad (\text{eq 3})$$

where, for example, **[free k-mer1]** is the DNA concentration of unbound k-mer1 and **[k-mer1 bound on protein]** is the DNA concentration of bound k-mer1.

The general equation of the dissociation constant for each 8mer is :

$$\mathbf{Kd_{k-merN} = \frac{[\text{protein not bound on any k-mer}][\text{free k-merN}]}{[\text{k-merN bound on protein}]}} \quad (\text{eq 4})$$

where **N={1, 2, ..., 32896}**

In each different protein/DNA ratio condition, the total concentration of the protein is given as follows :

$$[\text{total protein}] = [\text{protein not bound on any k-mer}] + \sum_{N=1}^{32896} [\text{k-merN bound on protein}]$$

(eq 5)

We introduce the dissociation constants ($Kd_{k\text{-mer}N}$) as follows:

If we expand eq 5 we have:

$$[\text{total protein}] = [\text{protein not bound on any k-mer}] + [\text{k-mer1 bound on protein}] +$$

$$[\text{k-mer2 bound on protein}] + \dots + [\text{k-merN bound on protein}]$$

(eq 6)

The $[\text{protein not bound on any k-mer}]$, $[\text{k-mer2 bound on protein}]$, $[\text{k-merN bound on protein}]$ in eq 6 are replaced by $Kd_{k\text{-mer}1}$, $Kd_{k\text{-mer}2}$ and $Kd_{k\text{-mer}N}$ respectively from eq 2, eq 3, eq 4:

$$[\text{total protein}] = \frac{Kd_{k\text{-mer}1} \cdot [\text{k-mer1 bound on protein}]}{[\text{free DNA k-mer1}]} * \left(1 + \sum_{N=2}^{32896} \frac{[\text{free DNA k-merN}]}{Kd_{k\text{-mer}N}}\right) + [\text{k-mer1 bound on protein}]$$

(eq 7)

By solving for $[\text{k-mer1 bound on protein}]$ we obtain:

$$[\text{k-mer1 bound on protein}] = \frac{[\text{total protein}] \cdot [\text{free DNA k-mer1}]}{[\text{free DNA k-mer1}] + Kd_{k\text{-mer}1} * \left(1 + \sum_{N=2}^{32896} \frac{[\text{free DNA k-merN}]}{Kd_{k\text{-mer}N}}\right)}$$

(eq 8)

The spectroscopic representation of eq 8 is the one shown in Figure 22B.

We then rearrange eq 8 in order to bring it in a form that will be convenient to fit it in the data for each individual k-mer:

$$\frac{[\text{k-mer1 bound on protein}]}{[\text{free DNA k-mer1}]} = \frac{[\text{total protein}]}{[\text{free DNA k-mer1}] + Kd_{\text{k-mer1}} * (1 + \sum_{N=2}^{32896} \frac{[\text{free DNA k-merN}]}{Kd_{\text{k-merN}}})} \quad (\text{eq 9})$$

If we divide both $[\text{k-mer1 bound on protein}]$ and $[\text{free DNA k-mer1}]$ with $[\text{total DNA k-mer1}]$ we will have the fraction of bound and free k-mer1 respectively :

$$\begin{aligned} \frac{[\text{k-mer1 bound on protein}]}{[\text{free DNA k-mer1}]} &= \frac{\frac{[\text{k-mer1 bound on protein}]}{[\text{total DNA k-mer1}]}}{\frac{[\text{free DNA k-mer1}]}{[\text{total DNA k-mer1}]}} \\ &= \frac{\frac{[\text{k-mer1 bound on protein}]}{[\text{total DNA k-mer1}]}}{\frac{[\text{total DNA k-mer1}] - [\text{k-mer1 bound on protein}]}{[\text{total DNA k-mer1}]}} \\ &= \frac{r_{\text{k-mer1}}}{1 - r_{\text{k-mer1}}} \end{aligned}$$

So eq 9 can be rewritten as:

$$\frac{r_{\text{k-mer1}}}{1 - r_{\text{k-mer1}}} = \frac{[\text{total protein}]}{[\text{free DNA k-mer1}] + Kd_{\text{k-mer1}} * (1 + \sum_{N=2}^{32896} \frac{[\text{free DNA k-merN}]}{Kd_{\text{k-merN}}})} \quad (\text{eq 10})$$

Because the dissociation constants (**Kd_{k-merN}**) of eq 10 are unknown we have to substitute them with the dissociation constant of the 20mer degenerate DNA library.

The dissociation constant for the total amount of the 20mer degenerate DNA library (Kd of the library, Figure 17) is the following:

$$\mathbf{Kd_{kmix}} = \frac{[\text{protein not bound on any k-mer}] * [\text{free library}]}{[\text{library bound on protein}]} \quad (\text{eq 11})$$

where **Kd_{kmix}** is the amount of protein where 50% of the 20mer degenerate DNA library is bound.

The concentration of the amount of DNA bound on the protein is given:

$$[\text{library bound on protein}] = \sum_{N=1}^{32896} [\text{k-merN bound on protein}]$$

If we expand the above equation we have:

$$\begin{aligned} [\text{library bound on protein}] &= [\text{k-mer1 bound on protein}] \\ &+ [\text{k-mer2 bound on protein}] + \dots + [\text{k-merN bound on protein}] \end{aligned} \quad (\text{eq 12})$$

If we substitute eq 2, eq 3, eq 4 and eq 11 in eq 12 we have:

$$\frac{[\text{free DNA k-mers}]}{\mathbf{Kd_{kmix}}} = \sum_{N=1}^{32896} \frac{[\text{free DNA k-merN}]}{\mathbf{Kd_{k-merN}}}$$

By rearranging we have:

$$\frac{[\text{free DNA k-mers}]}{Kd_{k\text{-mer1}}} - \frac{[\text{free DNA k-mer1}]}{Kd_{k\text{-mer1}}} = \sum_{N=2}^{32896} \frac{[\text{free DNA k-merN}]}{Kd_{k\text{-merN}}} \quad (\text{eq 13})$$

By substituting eq 13 to eq 10 we have:

$$\frac{r_{k\text{-mer1}}}{1-r_{k\text{-mer1}}} = \frac{[\text{total protein}]}{[\text{free DNA k-mer1}] + Kd_{k\text{-mer1}} * \left(1 + \frac{[\text{free DNA k-mers}]}{Kd_{k\text{-mer1}}} - \frac{[\text{free DNA k-mer1}]}{Kd_{k\text{-mer1}}}\right)}$$

which equals to:

$$\frac{r_{k\text{-mer1}}}{1-r_{k\text{-mer1}}} = \frac{[\text{total protein}]}{Kd_{k\text{-mer1}} * \left(1 + \frac{[\text{free DNA k-mers}]}{Kd_{k\text{-mer1}}}\right)}$$

The above competition model implies that we have each 8mer alone on the tube. This is not representative of the 20mer degenerate DNA library where each 20mer molecule has 12 overlapping 8mers. If we correct for this feature then we have the following:

(total number of 8mer molecules in the DNA library pool)=12*(number of 20mer molecules)

So the concentration of non bound 8mers is:

$$[\text{free DNA k-mers}] = 12 * [\text{free 20mer DNA sequence of the library}]$$

Consequently:

$$\frac{r_{k\text{-mer}1}}{1-r_{k\text{-mer}1}} = \frac{[\text{total protein}]}{Kd_{k\text{-mer}1} * \left(1 + \frac{12 * [\text{free 20mer DNA sequence of the library}]}{Kd_{\text{mix}}}\right)} \quad (\text{eq 14})$$

The **Kd_{mix}** is substituted with values 650nM of protein for the mGATA1_{NC} experiment and 1600nM of protein for the mKLF1₂₅₇₋₃₅₈ experiment (Figure 17).

The unknown **Kd_{k-mer1}** can be calculated by minimising the least square fit of a non linear curve of the $\frac{r_{k\text{-mer}1}}{1-r_{k\text{-mer}1}}$ against the **[total protein]**. We note here that the above equations imply that the binding isotherm starts from the beginning of x axis at x=0. As can be seen in Figure 17, the binding isotherm starts from the value approximately 250nM of protein. This is attributed to the sensitivity of the instrument where although there is protein-DNA binding in the solution, it is below the detection levels. To calculate the Kd for every k-mer we thus subtracted from the **Kd_{mix}** the 250nM value in each experiment. We then add 250nM in the final **Kd_{k-mer1}** value.

The final value of **Kd_{k-mer1}** is also corrected for the total number of different k-mers that we have which indicates the total number of protein molecules allocated in each different DNA k-mer. For example the corrected **Kd_{k-mer1}** for an 8mer is:

$$Kd_{k\text{-mer}1 \text{ corrected}} = Kd_{k\text{-mer}1} * \frac{1}{32896}$$

where the number 32896 represents the fact that we have 32896 different 8mers

For the 8mers the values of the **Kd_{k-mer1 corrected}** is in the picoMolar scale which is in good agreement with the DNA concentration of the 8mers in the EMSA-seq reaction used.

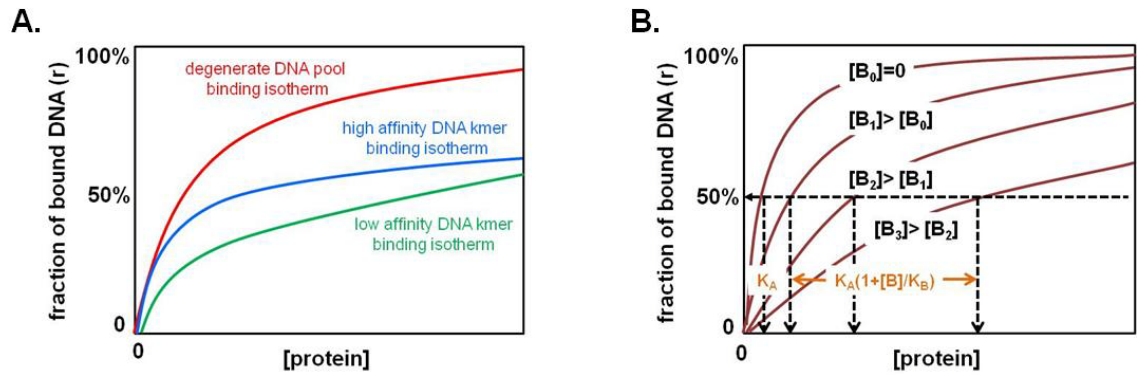


Figure 22 . A) Binding of a transcription factor to two DNA molecules of different affinity. The individual binding isotherms for one high and one low affinity DNA molecule are shown along with the binding isotherm of a solution that contains a mixture of low and high affinity DNA molecules (“DNA pool”) (it is preferred the DNA pool to have equimolar amounts of different DNA molecules but usually there are deviations due to manufacturing reasons). The **[protein]** represents the concentration of the transcription factor. The fraction of the DNA molecules bound by the transcription factor is presented with “r”. The “DNA pool” curve is the binding isotherm observed during an EMSA-seq experiment. The different affinity DNA molecules represent different k-mers. B) Binding isotherms of a specific DNA k-mer in the presence of increasing concentrations of another higher affinity DNA k-mer. The binding isotherms represent the competition between the two DNA k-mers. The **[protein]** represents the concentration of the transcription factor. The **[B₀]**, **[B₁]**, **[B₂]**, **[B₃]** represent different concentrations of the competitor (**[B₀]**<**[B₁]**<**[B₂]**<**[B₃]**). In the presence of a higher concentration of the competitor, the amount of protein needed to bind 50% of the examined k-mer is increased by an amount that depends in both the concentration of the competitor and the binding affinity of the competitor for the examined k-mer (Adapted from (Bisswanger, 2008)).

3.2.2 Cooperative competition model in an EMSA-seq reaction

Because we are testing the binding behaviour of 32896 k-mers simultaneously we expect deviations from the binding equilibrium as presented in eq 2. These deviations might be attributed to an unknown biophysical behaviour or it might be the result of a non specific carry-over of 8mers due to the overlapping nature of 8mers on the 20mer sequence. The latter is assumed to be part of the experimental error. For this reason we are introducing in the equation an extra parameter, the EMSA-seq coefficient (e). The way we are introducing it is similar to the Hill coefficient that is used to describe the binding cooperativity which occurs from allosteric changes. Although the previous model might not be of a direct relevance with our data, our rational for using this equation is the following:

- i) Without overfitting, the extra parameter will be able to describe changes in the expected shape of the binding isotherm curve as the Hill coefficient does,
- ii) The value of the parameter might reflect a biological phenomenon as the Hill coefficient also does.

The EMSA-seq coefficient (e) is related to the binding equilibrium as follows:



The **[k-mer1 bound on protein]** is calculated experimentally. The **[free k-mer1]** can be calculated analytically (**[free k-mer1]=[total k-mer1]-[k-mer1 bound on protein]**). The **[protein not bound on any k-mer]** can also be calculated analytically (**[protein not bound on any k-mer]=[total protein]-[k-mer1 bound on protein]**). The previous calculation assumes that the DNA/protein binding is 1:1. The binding equilibrium equation with the EMSA-seq coefficient (e) tests this “1:1 binding” assumption which means, for example, whether the protein binds more or less DNA molecules of k-mer1 than expected

by the classical protein-DNA equilibrium. If the binding follows the classical protein DNA equilibrium then $e=1$. If there are deviations from this equilibrium $e \neq 1$.

Because the experiment is designed so that we have a fixed amount of DNA against an increasing amount of protein then the EMSA-seq coefficient (e) value will have the following meaning:

- If $e > 1$ we define that the binding shows negative cooperativity. This means that in order to have a specific amount of bound DNA, more protein is needed than expected. So, for example, the protein binds less DNA molecules of a specific k-mer than expected by the classical protein-DNA equilibrium.
- If $e < 1$ we define that the binding shows positive cooperativity. This means that in order to have a specific amount of bound DNA, less protein is needed than expected. So, for example, the protein binds more DNA molecules of a specific k-mer than expected by the classical protein-DNA equilibrium
- If $e = 1$ we have no-cooperativity.

In more detail, the introduction of the EMSA-seq coefficient in eq 14 is as follows:

We again assume the antagonistic model of Figure 21. For each of the two k-mers, the binding affinities are again expressed by the dissociation constants Kd_{k-mer1} and Kd_{k-mer2} respectively now with the EMSA-seq coefficient introduced:

$$Kd_{k-mer1} = \frac{[\text{protein not bound on any k-mer}]^{e1} \cdot [\text{free k-mer1}]}{[\text{k-mer1 bound on protein}]} \quad (\text{eq 15})$$

$$Kd_{k-mer2} = \frac{[\text{protein not bound on any k-mer}]^{e2} \cdot [\text{free k-mer2}]}{[\text{k-mer2 bound on protein}]} \quad (\text{eq 16})$$

where “**e1**” and “**e2**” are the EMSA-seq coefficients of k-mer1 and k-mer2.

In each protein/DNA ratio in the EMSA-seq experiment the total concentration of the protein is the one presented in eq 6. The dissociation constant (**Kd_{mix}**) for the total amount of DNA library bound has also been presented in eq 11. The concentration of the amount of DNA bound on the protein is given in eq12. If we substitute eq12 in eq 11 and the derived equation in eq 6 we have the following:

$$[\text{total protein}] = [\text{protein not bound on any k-mer}] + \frac{[\text{protein not bound on any k-mer}] * [\text{free library}]}{\mathbf{Kd_{mix}}}$$

This can be rewritten as:

$$[\text{total protein}] = [\text{protein not bound on any k-mer}] * \left(1 + \frac{[\text{free library}]}{\mathbf{Kd_{mix}}}\right)$$

The [protein not bound on any k-mer] is replaced with the **Kd_{k-mer1}** from eq 15 as follows:

$$[\text{total protein}] = \left(\frac{\mathbf{Kd_{k-mer1}} * [\text{k-mer1 bound on protein}]}{[\text{free DNA k-mer1}]} \right)^{1/e1} * \left(1 + \frac{[\text{free library}]}{\mathbf{Kd_{mix}}}\right)$$

By solving for [k-mer1 bound on protein] we obtain:

$$[\text{k-mer1 bound on protein}]^{1/e1} = \frac{[\text{total protein}] * [\text{free DNA k-mer1}]^{1/e1}}{\mathbf{Kd_{k-mer1}}^{1/e1} * \left(1 + \frac{[\text{free library}]}{\mathbf{Kd_{mix}}}\right)}$$

This equals to:

$$\frac{[\text{k-mer1 bound on protein}]^{1/e1}}{[\text{free DNA k-mer1}]^{1/e1}} = \frac{[\text{total protein}]}{\mathbf{Kd_{k-mer1}}^{1/e1} * \left(1 + \frac{[\text{free library}]}{\mathbf{Kd_{mix}}}\right)}$$

This can be written as:

$$\frac{[\text{k-mer1 bound on protein}]}{[\text{free DNA k-mer1}]} = \left(\frac{[\text{total protein}]}{K_{d\text{k-mer1}}^{1/e_1} * (1 + \frac{[\text{free library}]}{K_{d\text{k-mix}}})} \right)^{e_1}$$

And eventually:

$$\frac{r_{\text{k-mer1}}}{1-r_{\text{k-mer1}}} = \left(\frac{[\text{total protein}]}{K_{d\text{k-mer1}} * (1 + \frac{12 * [\text{free 20mer DNA sequence of the library}]}{K_{d\text{k-mix}}})} \right)^{e_1} \quad (\text{eq 17})$$

In the equilibrium model the $K_{d\text{k-mer1}}$ takes values from the **[total protein]** because there is a linear relationship between the $\frac{[\text{k-mer1 bound on protein}]}{[\text{free DNA k-mer1}]}$ and the **[total protein]**.

In the above model the $K_{d\text{k-mer1}}$ takes values from the **[total protein]^{e1}** because the $\frac{[\text{k-mer1 bound on protein}]}{[\text{free DNA k-mer1}]}$ relates to the “**e1**” power of **[total protein]**.

Thus in order to be able to compare the calculated $K_{d\text{k-mer1}}$ of individual k-mers we have to normalise the values as follows:

$$K_{d\text{k-mer1_normalised}} = (K_{d\text{k-mer1}})^{\frac{1}{e_1}} \quad (\text{eq 18})$$

We also need to correct the $K_{d\text{k-mer1}}$ value as has been presented earlier. So for the 8mers we will have:

$$K_{d\text{k-mer1_normalised_corrected}} = K_{d\text{k-mer1_normalised}} * \frac{1}{32896}$$

We also note that the equilibrium model (eq 14) is a nested model of the model that has the EMSA-seq coefficient for $e=1$ (eq 17). Because we introduced a second parameter (e , EMSA-seq coefficient), we expect to fit the data better. In order to check whether the better fit was not due to the extra degrees of freedom but it was a genuine better fit, we used an F-test to compare the non-cooperative ($e=1$) with the cooperative ($e \neq 1$) model.

In Figure 23 we see that for the “AATG**G**ATC” 8mer the cooperative model fits the EMSA-seq data better than the non-cooperative model. On the contrary, for the “GCG**A**TAAA” 8mer, the cooperative model fits the data equally well as the non-cooperative model.

In Figure 24 we compare the calculated K_d from the EMSA-seq (units in concentration of the titrated protein, $K_{d_{\text{protein}}}$), with the K_d that was presented in the literature (units in concentration of the titrated DNA probe, $K_{d_{\text{DNA_probe}}}$) (Merika and Orkin, 1993) for the following 6mers “AG**A**TAA”, “T**G**ATTG”, “G**G**ATAG”, “T**G**ATTA”. We inspect a good concordance which indicates that the $K_{d_{\text{protein}}}$ derived values, in a high throughput way, do indeed represent a good approximation of the binding affinities.

We note here that (Guertin et al., 2012), in their PB-seq platform, presented also a way to calculate absolute dissociation constants (“ K_d ”). Their approach differs in the following:

- i) The authors did only one condition with a specific protein/DNA ratio in contrast to the multiple conditions for the mGATA1_{NC} or mKLF1₂₅₇₋₃₅₈ EMSA-seq in our experiments.
- ii) In (Guertin et al., 2012) the read depth of each k-mer is transformed into “ K_d ” values by experimentally calculating the “ K_d ” for two specific k-mers and calculating the “ K_d ” values for the rest of the k-mers relative to the two experimentally calculated ones. This implies that k-mers with the same read depth will eventually have the same “ K_d ”. By exploring our EMSA-seq data we realised that the relative values of the read depth of the same k-mers in two different

protein/DNA ratios is not fixed and for example, in one condition one k-mer might have more reads than the other one whereas in another condition the reverse might happen. We assign this phenomenon to the competitive nature of the EMSA-seq reaction. For this reason, we feel that our approach of fitting a competition model in the different experimental conditions is much more comprehensive.

Our method is similar to the (Nutiu et al., 2011). The authors titrated the protein against a specific amount of DNA that was immobilised on a microarray slide. They visualised the binding of the fluorescently tagged protein on the DNA sequences with an “Illumina” sequencing machine. Although their method is laborious and available to a limited number of labs, their data produced binding affinity “Kd” values for the examined protein-DNA interactions. The authors used the Hill equation in their analysis that incorporated the Hill co-efficient as we have already discussed. In contrast to us (see later sections), they failed to appreciate the importance of this coefficient.

Lastly, (Slattery et al., 2011) presented a model for inferring the relative affinities (Kd) of k-mers from multiple rounds of SELEX-seq. Although this model reflects the behaviour of k-mers in multiple rounds of enrichment, it still does not incorporate any competition between the k-mers.

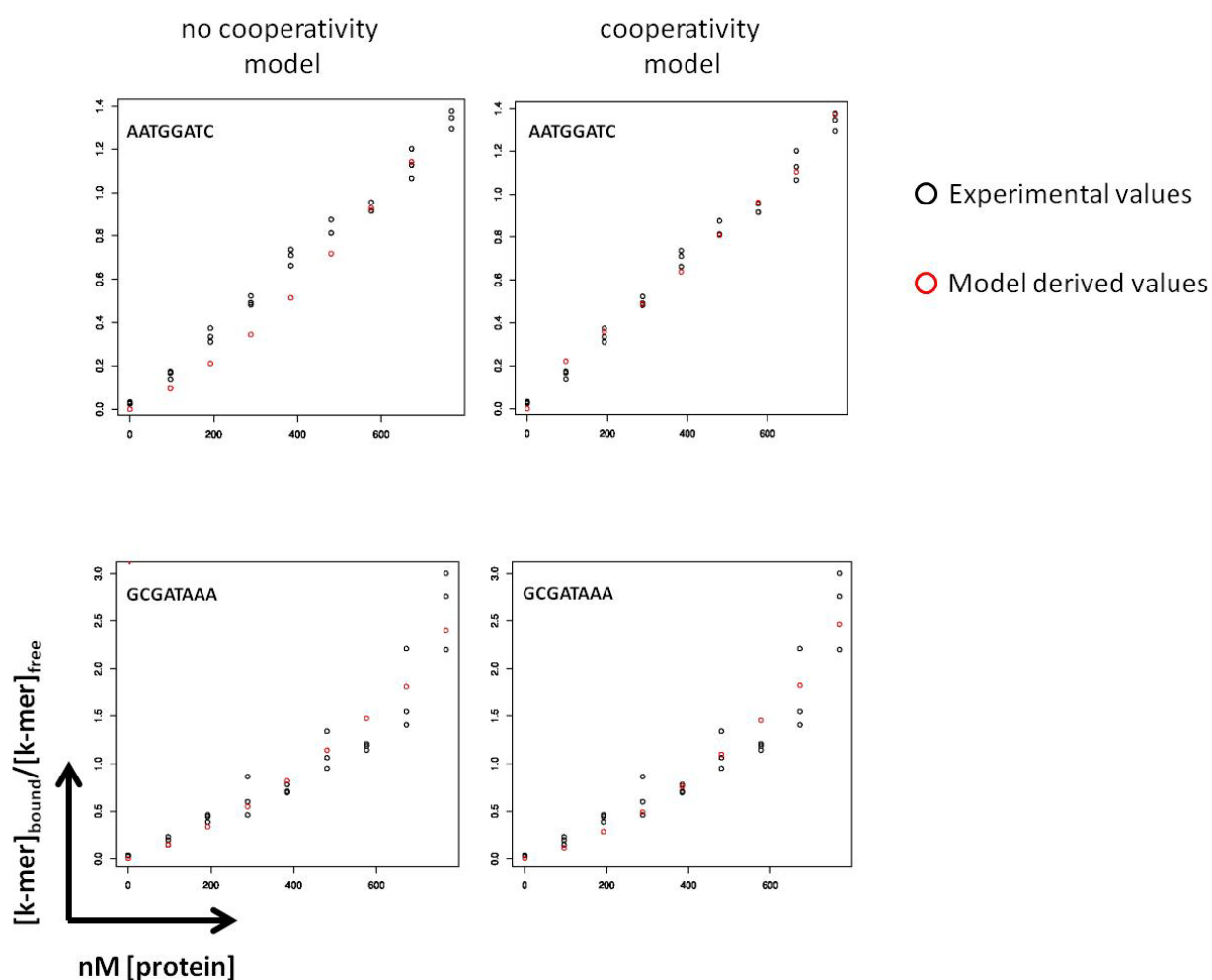


Figure 23. Comparison between the non-cooperative and the cooperative model for the “AATGGATC” and “GCGATAAA” 8mers. The black circles represent the derived values for the three replicates from the nine conditions of the mGATA1_{NC} EMSA-seq experiments. The y axis represents the fraction of bound 8mer. The x axis represents the concentration of the mGATA1_{NC} in the nine different EMSA-seq conditions (concentration of mGATA1_{NC} in Figure 17 minus 250 nM which represents the region with the lack of signal in Figure 17). The red circles represent the expected values from either the non-cooperative or the cooperative model.

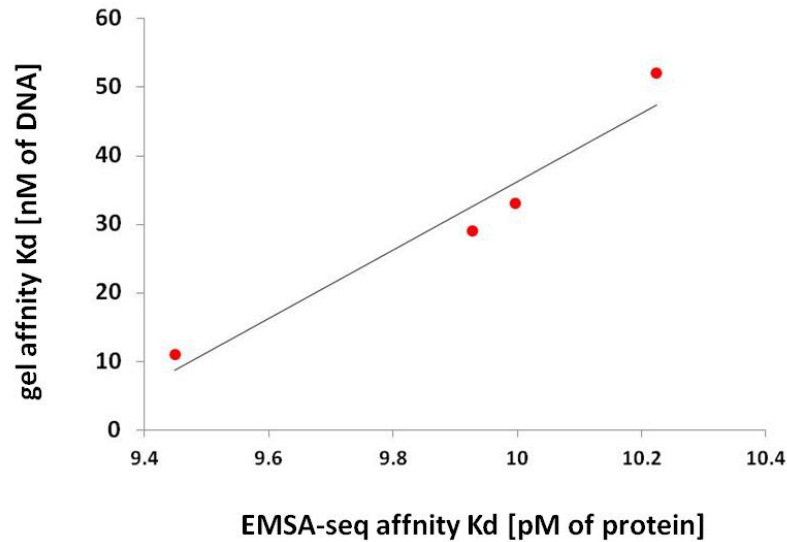


Figure 24. Comparison of reported Kd affinities (Merika and Orkin, 1993) and the respective binding affinities calculated from the EMSA-seq of mGATA1_{NC}. We note that (Merika and Orkin, 1993) have titrated the DNA probes against a specific amount of protein. Thus, the Kd units are in nM of DNA ($K_{d_{DNA_probe}}$). On the other hand, we titrated the protein against a specific amount of DNA. The Kd units of the EMSA-seq are in pM of protein ($K_{d_{protein}}$). The 6mers used are “AGATAA” ($K_{d_{DNA_probe}}=11\text{nM}$, EMSA-seq $K_{d_{protein}}=9.44\text{ pM}$), “TGATTG” ($K_{d_{DNA_probe}}=29\text{nM}$, EMSA-seq $K_{d_{protein}}=9.92\text{ pM}$), “GGATAG” ($K_{d_{DNA_probe}}=52\text{nM}$, EMSA-seq $K_{d_{protein}}=10.22\text{ pM}$), “TGATTA” ($K_{d_{DNA_probe}}=33\text{nM}$, EMSA-seq $K_{d_{protein}}=10\text{ pM}$). The EMSA-seq affinity is the median value of the 8mers that contain the respected 6mer.

3.3 Comparison of published GATA1 DNA binding specificities

Since the discovery of GATA1 as one of the major transcriptional regulators of erythropoiesis, numerous studies have been performed to find the DNA binding specificities of the protein. The early studies employed Sanger sequencing combined with multiple rounds of SELEX (Systematic Evolution of Ligands by Exponential Enrichment) (Merika and Orkin, 1993, Ko and Engel, 1993). Years later the SELEX approaches were combined with high-throughput sequencing (SELEX-seq) (Jolma et al., 2010) in order to get a detailed image of the DNA binding landscape of GATA1. In parallel, ChIP-chip and ChIP-seq studies extracted binding motifs from GATA1 binding sites in vivo (Figure 25).

All these studies concluded that the GATA1 DNA binding motif is the WGATAR. GATA1 can also recognize the “GATT” and “GATC” motifs but with lower affinity. (Ko and Engel, 1993, Whyatt et al., 1993).

The N-terminal and the C-terminal Zn fingers of GATA1 have been shown to be able to bind DNA independently of each other. The N-terminal GATA1 Zn finger binds DNA with lower affinity than the C-terminal GATA1 Zn finger and it shows different DNA binding specificities than the C-terminal Zn finger (Whyatt et al., 1993, Ko and Engel, 1993). The GATA1 N-terminal Zn finger has a preference for “GATC” core motifs (Newton et al., 2001, Martin and Orkin, 1990, Pedone et al., 1997). The C-terminal Zn finger has been shown to bind the core “GAT(A/T)” motif (Martin and Orkin, 1990; Pedone et al., 1997; Newton et al., 2001).

It has also been shown that the GATA1 motifs can appear as palindromes. In this case it has been hypothesized that the N-terminal Zn finger is used to stabilize the interaction while the C-terminal Zn finger binds in a suboptimal motif site. For example, it has been shown that the GATA1 C-terminal Zn finger can bind in a suboptimal motif site (“GATG”),

if the N-terminal Zn finger can bind a nearby motif (“GTAAC”) (Whyatt et al., 1993). Another example involves the non-canonical motif GNN[TC]AAG where the first G is interacting with the C terminal finger whereas the last four bases are interacting with the N-terminal finger (Whyatt et al., 1993). These observations have not been discussed further in later studies mainly because it involved non canonical motifs.

In case that the GATA1 C-terminal Zn finger binds its optimal “GATA” motif, the simultaneous binding of the GATA1 N-terminal Zn finger increases the binding affinity of the sequence for the protein. This has been shown for the palindromic motif $\overleftarrow{\text{ATC}}\text{WGATA}\overrightarrow{\text{A}}$ (Trainor et al., 1996).

Recently, ChIP-seq studies have validated all the above in vitro experiments. From the in vivo binding sites, it has been shown that GATA1 can bind to either a single motif [W]GATAA or to a palindromic motif $\overleftarrow{\text{CATC}}\text{tGATAAG}\overrightarrow{\text{A}}$ (Yu et al., 2009, Wei et al., 2011, Horiuchi et al., 2011, Fujiwara et al., 2009) .

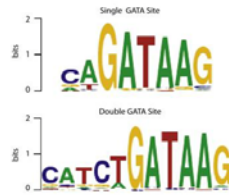
Below, based on the EMSA-seq results, we will explore each one of the mentioned behaviors of GATA1 namely:

- i) Its preference for “GATA” containing motifs.
- ii) The DNA binding specificities of its Zn fingers alone or together.
- iii) Its preference for palindromic sequences.

A.

GATA1 motifs from published studies:

De novo PWM from GATA1 ChIP-seq peaks (Yu et al., 2009)



Enriched 6mers in GATA1 ChIP-seq peaks (Cheng et al., 2009):

AGATAA
TGATAA
GATAAG
GATAAC

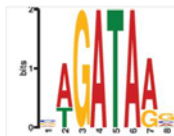
Enriched 6mers in GATA1 ChIP-chip peaks (Zhang et al., 2009):

AGATAA
AGATAG
TGATAA
TGATAG
WGATAA
WGATAR

B.

motif of the GATA1/FOG1 complex from published studies:

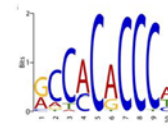
De novo PWM from ChIP-seq peaks containing only the GATA-1/FOG1 complex (Chlon et al., 2012)



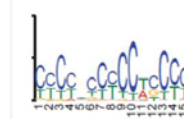
C.

KLF1 motifs from published studies:

De novo PWM from KLF1 ChIP-seq peaks (Tallack et al., 2010)



De novo PWM from KLF1 ChIP-seq peaks (Pilon et al., 2011)



D.

motif of the GATA1/KLF1 complex from published studies:

suggested PWM of the GATA1/KLF1 complex from the KLF1 ChIP-seq peaks that overlap with GATA1 ChIP-seq peaks (Tallack et al., 2010)

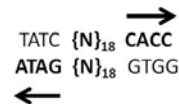


Figure 25. A) DNA binding specificities of GATA1 from published work. B) DNA binding specificities of the GATA1/FOG1 complex from published work. C) DNA binding specificities of KLF1 from published work. D) DNA binding specificities of the GATA1/KLF1 complex from published work. The motifs have been extracted from the indicated ChIP-seq and ChIP-chip studies.

3.4 Investigation of the DNA binding specificities and the DNA binding affinities of a transcription factor using EMSA-seq

3.4.1 General considerations of the EMSA-seq analysis

To extract the DNA binding specificities from the EMSA-seq experiment we focused on 8mers. It should be noted here that in order to find high affinity 8mers we followed 2 approaches:

- i) In the experiments where we have DNA binding isotherms we assigned to all the 8mers a “Kd” value, as has been described earlier. This gave an unbiased, “cut-off free” approach where each 8mer contributes a high, medium or low binding affinity.
- ii) In the experiments where we do not have a DNA binding isotherm we compared the condition with the lowest protein/DNA ratio (the reason for this selection is presented in Figure 26) versus the original DNA pool. We used statistical inference to find 8mers that are more frequent in the bound DNA fraction compared to the original DNA pool. This gave us only a fraction of enriched 8mers (compared to the total number of 8mers) (Table III) and their respective fold change (Figure 27). We also found statistically enriched 8mers for the EMSA-seq experiments that we had binding isotherms in order to be able to compare them with the rest of the EMSA-seq experiments.

EMSA-seq experiments	Enriched 8mers (p-value <10 ⁻⁶)
GATA1 _{NC}	2122
GATA1 _C	4763
GATA1 _{NC} /KLF1 ₂₅₇₋₃₅₈ main band	2230
GATA1 _{NC} /KLF1 ₂₅₇₋₃₅₈ upper band	4218
KLF1 ₂₅₇₋₃₅₈	193

Table III. Number of statistically enriched 8mers in each of the different EMSA-seq experiments. The 8mers were selected as described in material and methods (binomial test, p-value $\leq 10^{-6}$). As it is argued in the text, for every different protein only the EMSA-seq condition with the lowest ratio of protein to DNA was used to infer statistically enriched 8mers.

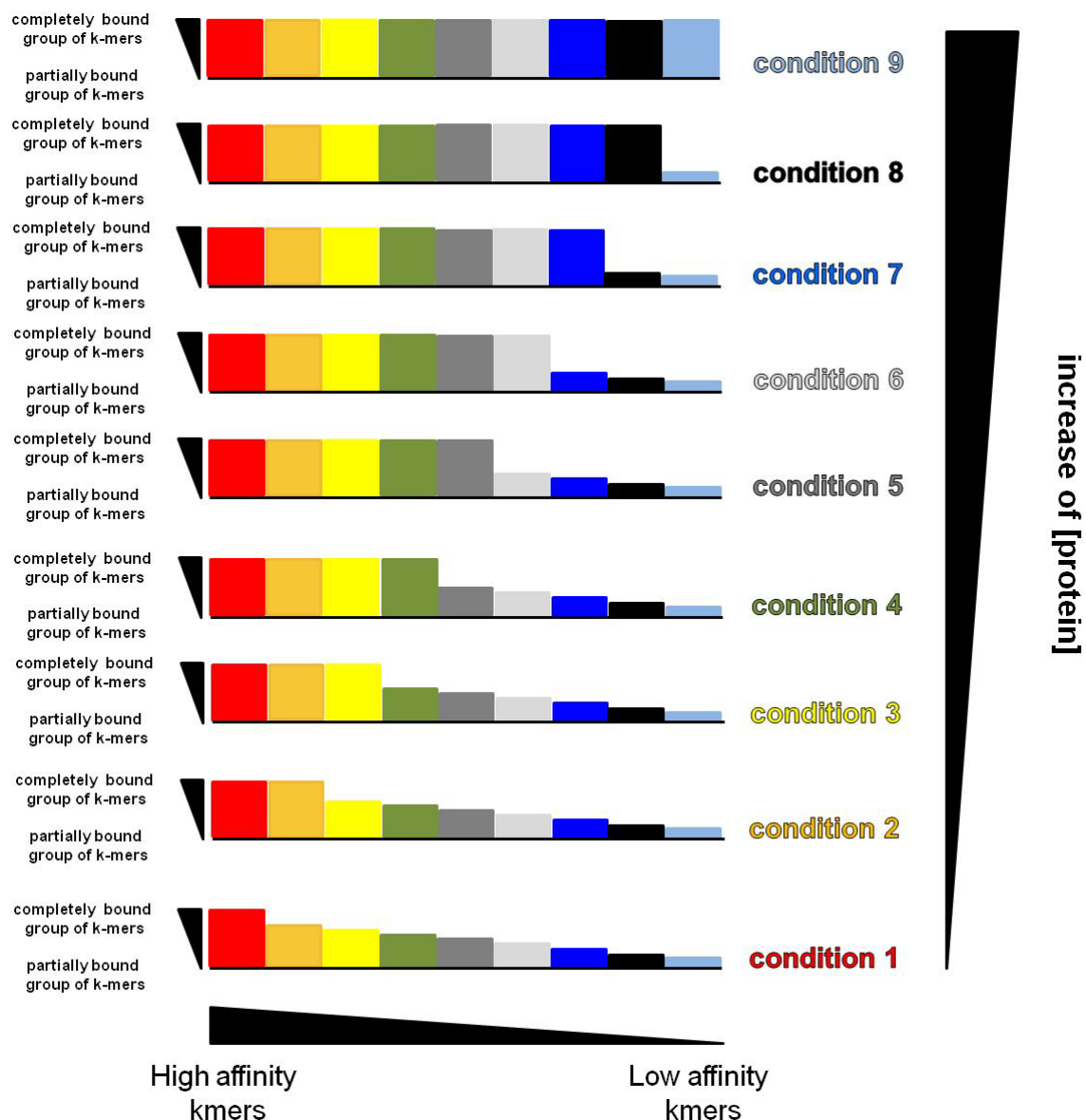
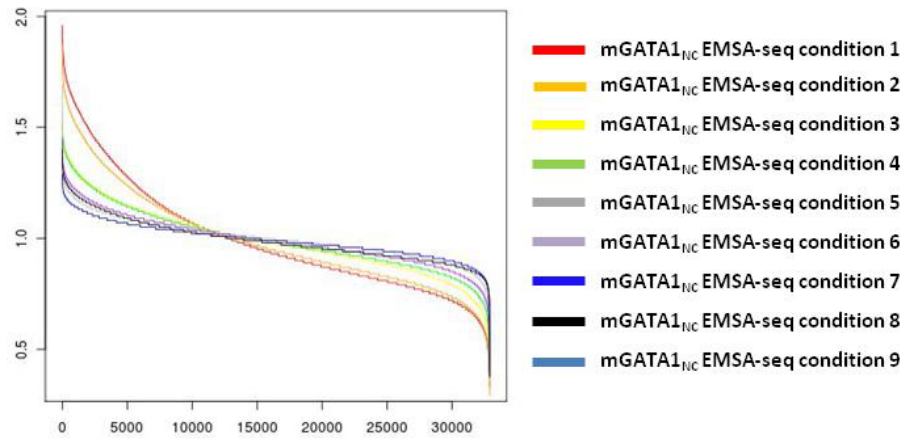


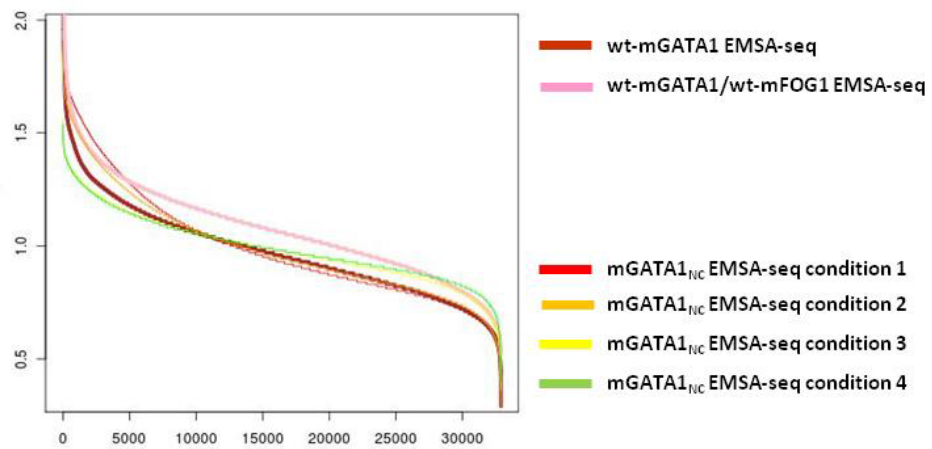
Figure 26. Kinetic behaviour of the EMSA-seq reaction in the different conditions of the mGATA1_{NC} experiment. Condition 1 represents a low protein/DNA ratio. Condition 9 represents a high protein/DNA ratio. The k-mers are separated in 9 categories (boxes of different colours) based on their affinity values. Each colour represents the different group of k-mers that is bound mainly in the condition of the same colour (coloured letters on the right). In each graph, the y axis represents the proportion of the bound molecules of each group. The protein binds first the high affinity k-mers until all of them are selected by the protein. Then the protein moves to bind the medium affinity k-mers. And the protein finally

binds the low affinity k-mers. When for example the protein selects all the high affinity k-mers, then the relative frequencies with each other (but not with the medium or low affinity k-mers) will look similar like the original unbound DNA library. This behaviour will not create a problem in the calling of enriched k-mers (fold change versus the original degenerate 20mer DNA library) in the low protein/DNA ratio conditions where only a fraction of high/medium affinity k-mers is bound (Figure 27A, conditions 1 and 2 have a significant number of k-mers with a high fold change). In the conditions where we increase the protein/DNA ratio, the normalisation will fail as more and more of the k-mers will be sampled until complete binding and the bound library will look similar to the original unbound DNA library (Figure 27A, in conditions 7-9 the majority of k-mers show low fold change). The suggested way to circumvent this problem is to assume additivity of the conditions where for example the fold change of k-mers in condition 4 will not be: $(\text{normalised number of k-mers in condition 4})/(\text{normalised number of k-mers in original DNA library})$ but $[(\text{normalised number of k-mers in condition 1})+(\text{normalised number of k-mers in condition 2})+(\text{normalised number of k-mers in condition 3})+(\text{normalised number of k-mers in condition 4})]/(\text{normalised number of k-mers in original DNA library})$. The different colours represent the different groups of k-mers bound in each EMSA-seq experimental condition from the high affinity ones (red colour) to the low affinity ones (cyan colour).

A.



B.



C.

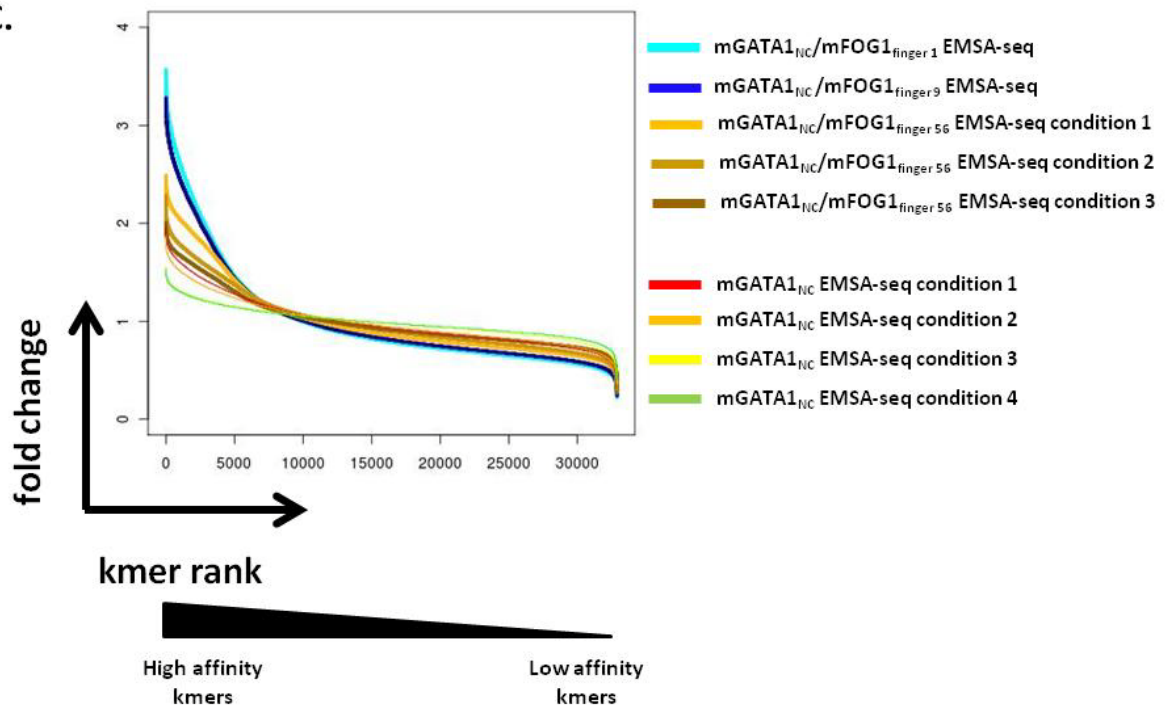


Figure 27 (previous page). A) Fold change distribution for the different conditions of the mGATA1_{NC} EMSA-seq experiments. The mGATA1_{NC} EMSA-seq condition 1 represents the low protein/DNA ratio condition. The mGATA1_{NC} EMSA-seq condition 9 represents the high protein/DNA ratio condition. The other conditions 2-8 are intermediate between 1 and 9 and represent conditions with increasing protein/DNA ratio (see text). B) Fold change distribution for the wt-mGATA1 EMSA-seq and the wt-mGATA1/wt-mFOG1 EMSA-seq experiments. For comparison the fold change distributions of the mGATA1_{NC} EMSA-seq conditions 1-4 are presented. C) Fold change distribution for the mGATA1_{NC}/mFOG1_{finger1}, mGATA1_{NC}/mFOG1_{finger9}, mGATA1_{NC}/mFOG1_{finger56} EMSA-seq experiments. The mGATA1_{NC}/mFOG1_{finger56} EMSA-seq condition 1 represents the low protein/DNA ratio condition. The mGATA1_{NC}/mFOG1_{finger56} EMSA-seq condition 2 represents the medium protein/DNA ratio condition. The mGATA1_{NC}/mFOG1_{finger56} EMSA-seq condition 3 represents the high protein/DNA ratio condition. For comparison, the fold change distributions of the mGATA1_{NC} EMSA-seq conditions 1-4 are presented. The k-mers were ranked from the one that shows the highest fold change (position 1) to the one that shows the lowest fold change (position 32896).

3.4.1 DNA binding specificities of mGATA1_{NC}

Based on the depth of sequencing of the EMSA-seq experiments, we have the statistical power to detect enriched k-mers up to 9 nucleotides in total when we compare with the original degenerate DNA library. Nevertheless, we selected an 8mer in order to see whether we can expand the 6mer motif presented in the literature. As seen in Table III, from the 32896 8mers that were counted from the EMSA-seq experiment, we found 2122 enriched 8-mers for the mGATA1_{NC} protein (binomial test p-value <0.00006) (15.5% of total). At first we searched which 4mers are enriched in these top 8mers. We found that “GATA” (occurrence in 36% of the enriched sequences) and “GATT” (26%) are the most frequent 4mers followed by “GATC” (10%) and “GATG” (7%) and finally by “GAGA” (7%) (Figure 28A). Thus the core motif that GATA1 prefers is presented in Figure 28B.

We then expanded the 4mer motif up to an 8mer motif by checking the base preference of the -1/-2 and +1/+2 positions (Figure 29). We notice that the “GATA” and the “GATT” are the less restricted motifs implying that the core 4 nucleotides are adequate for binding (Figure 29). As can be seen from the heat maps of Figure 29, the “GATC”, “GATG” and “GAGA” have extra flanking nucleotides restrictions implying that these nucleotides are necessary to stabilize the binding. More specifically whereas the +1 position in “GATA” and “GATT” seem to be degenerate the -1 position shows a specific restriction towards “ATC” for the “GATA” core motif and mainly towards “AT” for the “GATT” core. In the “GATA” motif we see that the G nucleotide in position -1 of the “GATA” motif is preferred only 58% of time (37 occurrences, orange colorful boxes in the heat map, versus 64 overall possible occurrences) whereas the “G” and “C” nucleotides in the “GATT” core motif are preferred 26% (17 occurrences, orange colorful boxes in the heat map, versus 64 overall possible occurrences) and 60% (39 occurrences, orange colorful boxes in the heat map, versus 64 overall possible occurrences) of the time respectively. For both the

“GATA” and “GATT” the -2/+2 position are degenerate (N[ATC_G]GATANN, N[ATC_G]GATTNN).

As far as it concerns “GATC”, the “G”, “C”, “T”, “A” in the -1 position are preferred 19%, 41%, 70% and 80% of time respectively (12, 26, 45, 51 occurrences respectively, orange colorful boxes in the heat map, versus 64 overall possible occurrences) . The “G”, “C”, “T”, “A” in the +1 position are preferred 41%, 19%, 80% and 70% of time respectively (26, 12, 51, 45 occurrences respectively, orange colorful boxes in the heat map, versus 64 overall possible occurrences). The -2/+2 position are degenerate (N[ATC_G]GATC[TA_G]N). For the “GATG” motif, there is a preference for the “A” and “T” nucleotides in both the (-1) and (+1) position. The -2 position is degenerate whereas in the +2 position there is a preference for the “A” and “T” nucleotides (N[AT]GATG[AT][AT]). For “GAGA”, the -2, -1, +2 positions are degenerate whereas the +1 position is restricted to contain only Thymidine (NNGAGATN).

We also checked whether the “GATA”/“GATT”/“GATC”/“GAGA” 4mers are present in the top or bottom of the first 2122 enriched 8mers. In Figure 28C,D we see that the “GATT” and “GAGA” are equally present in the top and bottom 8mers. On the contrary, “GATA” 4mers are present mainly on the top 8mers whereas “GATC” and “GATG” are present mainly on the bottom 8mers. **This confirms for one more time that “GATA” and “GATT” are the core motifs with the highest binding affinity whereas “GATC” and “GATG” are the core motifs with the lowest binding affinity.**

It is useful to mention here the structural properties on the DNA binding of the GATA1 C-terminal Zn finger. When bound to DNA, the chicken GATA1 C-terminal Zn finger folds into a structure that consists from four beta strands (β 1, β 2, β 3, β 4) (residues 1-26), an alpha helix (residues 27-37) and a long loop (carboxy-terminal tail) (residues 37-51) (Figure 30A). The alpha helix and the loop that connects the β 2 and the β 3 strands orient

in the major groove of DNA whereas the carboxy-terminal tail wraps around the DNA and lies in the minor groove (Omichinski et al., 1993). NMR studies have shown that the following amino-acids mainly make contacts with the DNA bases, Thr14, Leu15, Arg17, Asn27, Leu31, Leu35 (Omichinski et al., 1993, Tjandra et al., 1997) (Figure 30A). Arg17 and Asn27 make hydrogen bonds and/or salt bridges with the nucleotide bases or the phosphate backbone. Thr14, Leu15, Leu31 and Leu35 make hydrophobic contacts with the nucleotide bases (Omichinski et al., 1993). It has been shown that the binding energy of the GATA1 C-terminal Zn finger with the DNA is attributed mainly to the amino acids that create hydrogen bonds and/or salt bridges with the DNA whereas the hydrophobic interactions with the DNA do not have a strong role (Mott et al., 2004, Vonderfecht et al., 2008). For the “GATA” motif, the protein creates hydrogen bonds with the Gua of position 1, the Ade of position 2, the Ade of position 4 and the complementary nucleotide Ade of position 3. This might explain the strong preference for the “GAT” in the first 3 nucleotides of the motif and the “AT” preference for the 4th nucleotide (Mott et al., 2004).

percentage of 8mers that show each one of the 16 different 4mers was calculated. Then the frequency numbers were pooled and a z-score value was assigned to them. B) Logo representation of the enriched 4mers. C) Heat map of the presence of the 4mers (blue signal) on the most enriched to the lowest enriched 8mers. D) Cumulative distribution of the presence of the 4mers on the most enriched to the lowest enriched 8mers.

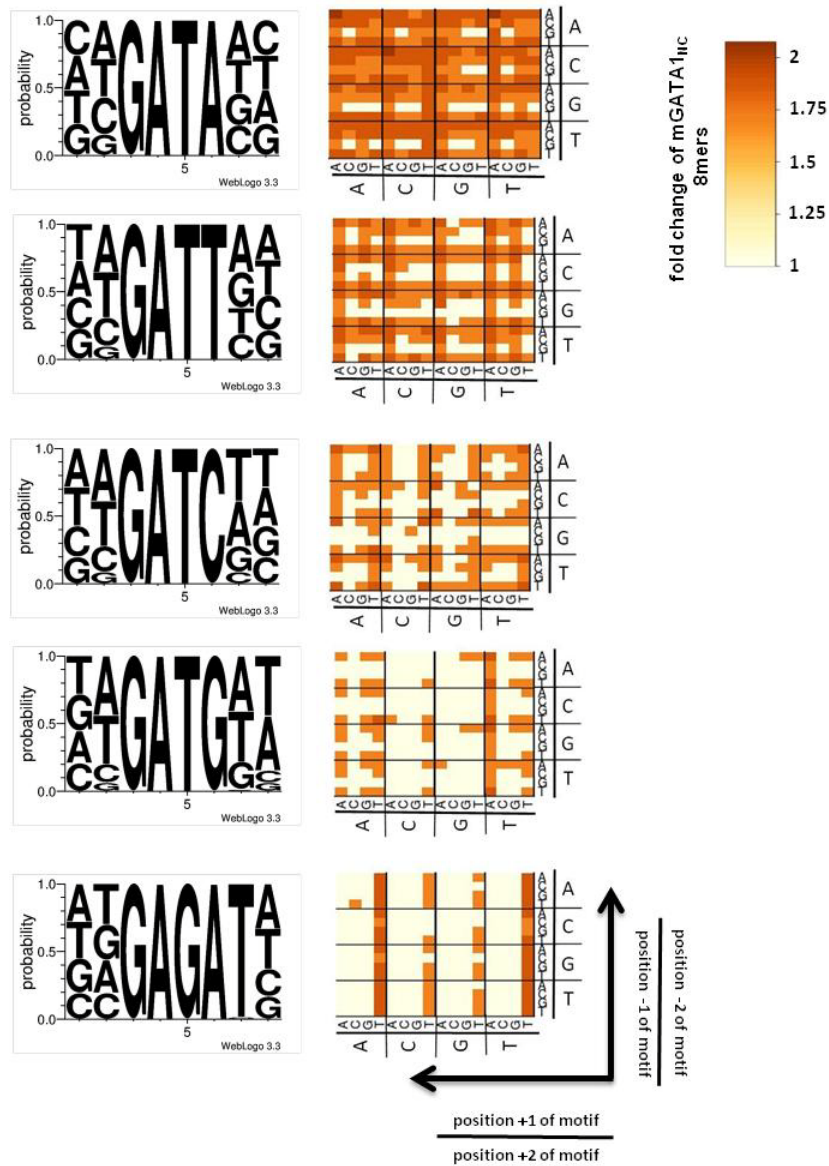
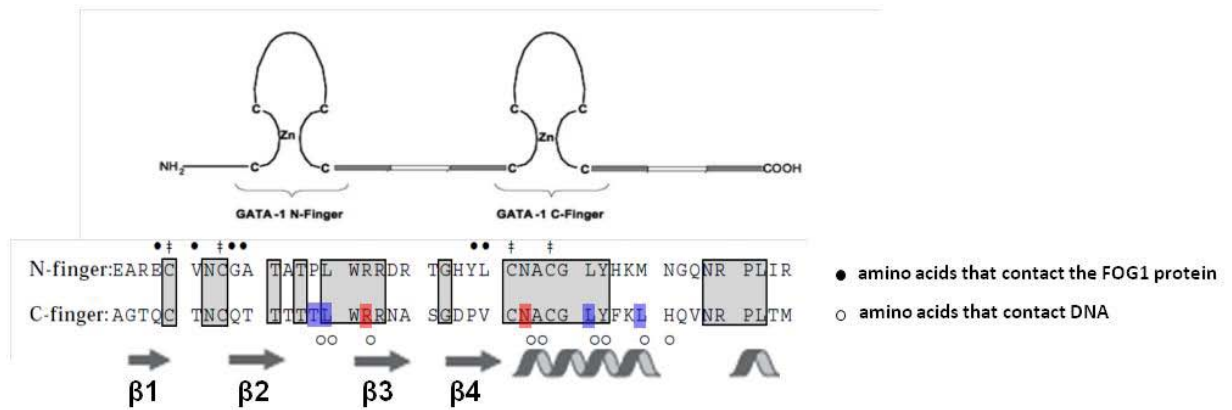


Figure 29. Nucleotide flank restriction of the core GATA1 4mer motifs ("GATA", "GATT", "GATC", "GATG", "GAGA") from the mGATA1_{NC} EMSA-seq experiment. The heat map represents the fold change for the mGATA1_{NC} 8mers with different nucleotides on positions -1, -2, +1 and +2. The logo on the left represents in a more compact way the enrichment seen on the heat map but without taking into account particular dinucleotide correlations.

A.



B.

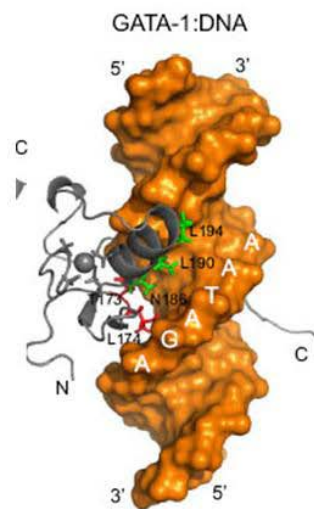


Figure 30. A) Folding of the GATA1 N-terminal and C-terminal Zn fingers. The grey arrow represents a beta strand whereas the grey spiral represents an alpha helix. The grey boxes represent regions of conservation between the two Zn fingers. The filled circles indicate amino acids that make contacts with the FOG1 protein. The open circles indicate amino acids that make contacts with the DNA sequence. The blue boxes indicate amino acids that make contacts with the DNA sequence. The blue boxes indicate amino acids that form hydrogen bonds with the DNA sequence. The red boxes indicate amino acids that are making hydrophobic contacts with the DNA sequence. Adapted from (Newton et al., 2001) and (Kowalski et al., 1999). B) Model of the GATA1 C-terminal Zn finger binding on the major groove of the DNA sequence. The amino acid L174 represents the Leu15 of the text, the rest of the amino acid numbers are changed accordingly. From (<http://www.faculty.ucr.edu/~mmaduro/meds.htm>).

3.4.2 The N-terminal GATA1 Zn finger binds to “GATC” and “GATG” motifs (DNA binding specificities of mGATA1_C)

We then examined the DNA binding specificities of the C-terminal Zn finger of GATA1 alone (mGATA1_C). In Figure 31A, we see that absence of the N-terminal Zn finger results in a reduced frequency of “GATC” (from 10% to 5%) and “GATG” (from 7% down to 3%) motifs in the top 2122 enriched 8mers. **This indicates that the GATA1 N-terminal Zn finger DNA binding motifs are the “GATC” and the “GATG”.** The fact that we do not see a complete depletion of “GATC” and “GATG” implies that the C-terminal Zn finger can bind to them albeit with a lower affinity. We also notice an increase in frequency for the 8mers containing the “GATT” motif (from 26% to 35%). In Figure 31C,D we see that the mGATA1_C “GATC” containing 8mers have a lower enrichment than the mGATA1_{NC} “GATC” containing 8mers. **This implies that the N-terminal Zn finger can only bind high affinity “GATC” containing 8mers. The C-terminal is able to stabilise the binding in low affinity “GATC” motifs.** In Figure 31B, we see that absence of the N-terminal Zn finger causes a re-distribution of the “GATA” motif along the 8mer sequence. In the mGATA1_{NC} 8mers the “GATA” motif is shifted towards the 4th and 5th position. In the mGATA1_C 8mers, the “GATA” motif is equally found in all the positions. The “GATT” motif shows a smaller shift towards position 1 and 2. The “GATG” or “GAGA” motifs do not show any difference in the mGATA1_{NC} versus the mGATA1_C 8mers. **This indicates that the N-terminal Zn finger is involved in simultaneous DNA binding mainly with the “GATA” motif and at a lower extend with the “GATT” motif. The “GATC” or “GATG” motifs are bound by either the C-terminal Zn finger or by the N-terminal Zn finger alone.**

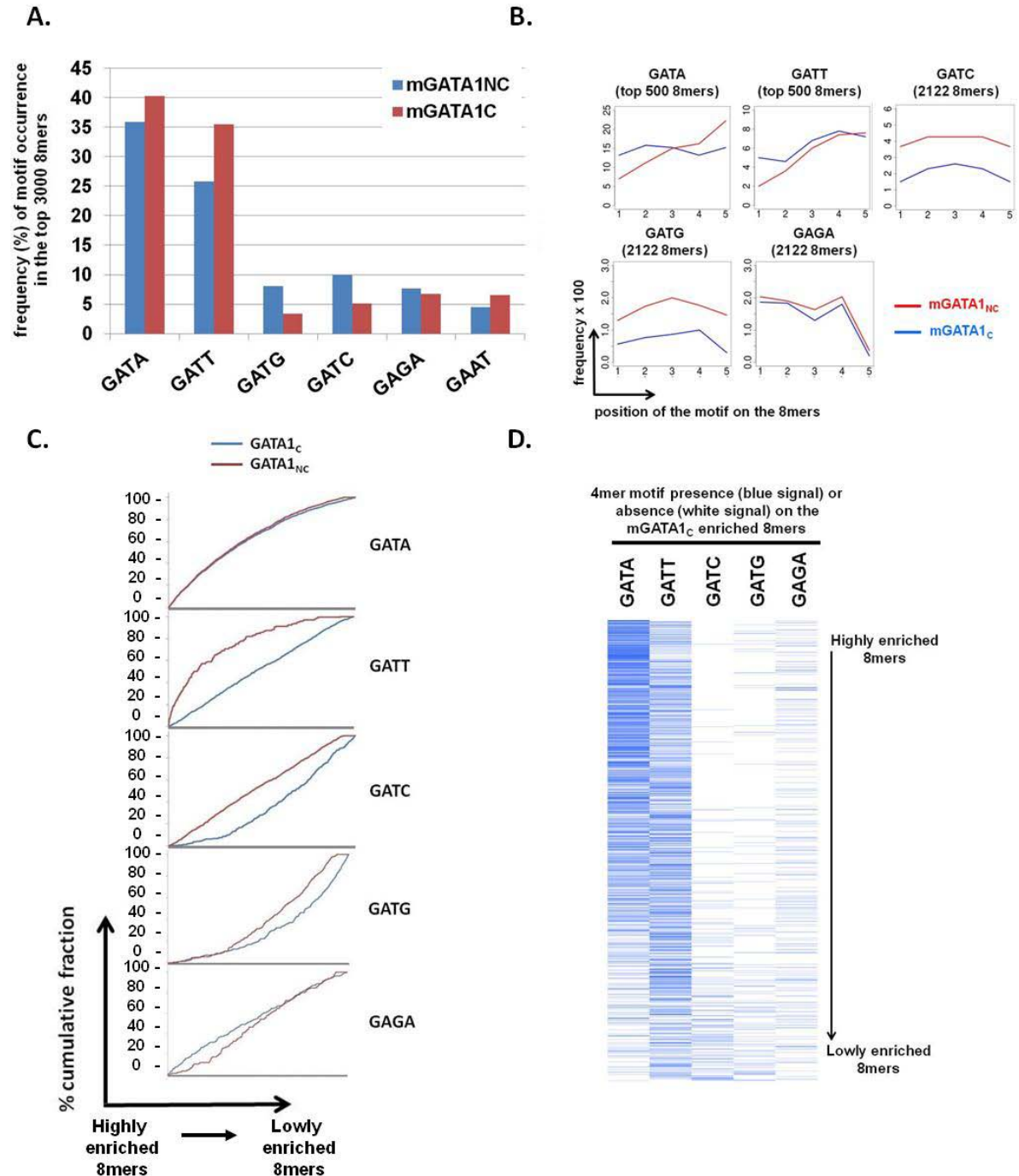


Figure 31. Binding specificities of mGATA1_C A) Frequency of occurrence (%) of the “GATA”, “GATT”, “GATC”, “GATG”, “GAGA” motifs in the enriched 8mers of either the mGATA1_C (top 2122 8mers) or the mGATA1_{NC} (2122 8mers). B) Distribution of the “GATA”, “GATT”, “GATC”, “GATG”, “GAGA” motif along the enriched 8mers for either the enriched mGATA1_{NC} (top 2122 8mers or top 500 8mers) or the enriched mGATA1_C 8mers (2122 8mers or top 500 8mers). C) Cumulative fraction of the “GATA”, “GATT”, “GATC”,

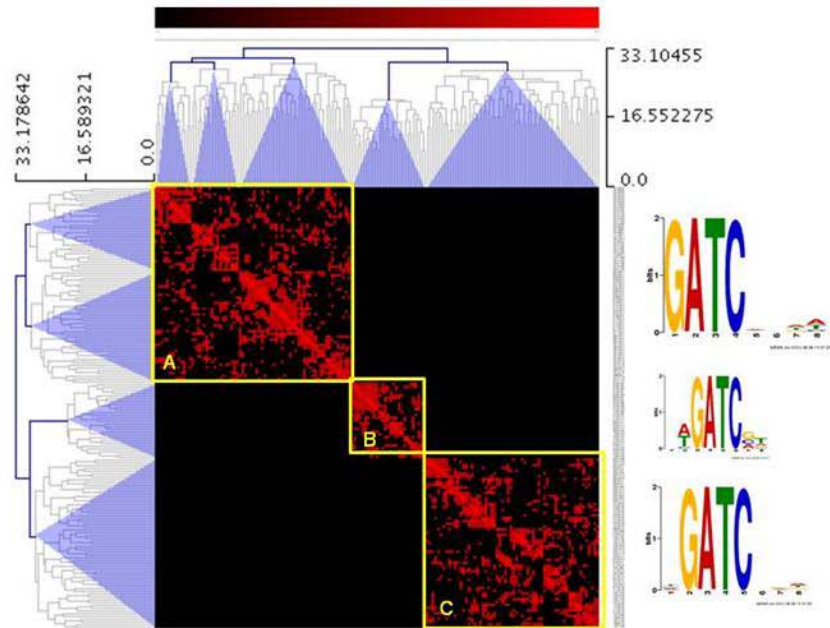
"GATG", "GAGA" motif appearance in the top 2122 enriched mGATA1_C 8mers. The 8mers are sorted by fold change from the highest (on the left of the x-axis) to the lowest (on the right of the x-axis). D) "GATA", "GATT", "GATC", "GATG", "GAGA" motif appearance in the top 2122 enriched mGATA1_C 8mers. The 8mers are sorted by fold change from the highest to the lowest.

3.4.2.1 Absence of the N-terminal Zn finger changes the binding specificities of the C-terminal Zn finger

We also grouped the GATC containing 8mers in 8mers that are present in both mGATA1_{NC} and mGATA1_C (GATA1 C-terminal Zn finger specific 8mers) and in 8mers that are present only in mGATA1_{NC} (GATA1 N-terminal Zn finger specific 8mers) (Figure 32). The GATA1 N-terminal Zn finger specific 8mers represent 8mers that only the N-terminal Zn finger binds. The GATA1 C-terminal Zn finger specific 8mers represent 8mers that only the C-terminal Zn finger binds. We clustered the 8mers based on their similarity. We used as a similarity measure the “Hamming distance”. The “Hamming distance” corresponds to the number of positions that are the same between two 8mers. The clustering gave us three groups of 8mers (groups A, B, C). Their main difference lies in the position of the “GATC” motif occurrence. In group A the “GATC” motif appears in position 1. In group B the “GATC” motif appears in position 3 and in group C the “GATC” motif appears in position 2. In the 8mers that the C-terminal Zn finger binds, we see an increase in the occurrence of 8mers that are part of group B (“GATC” in position 3) and a reduction of group A (“GATC” in position 1) related 8mers. The group C size is also larger. **This indicates that the C-terminal Zn finger needs extra interactions with the nucleotide bases or the phosphate backbone upstream of the motif to stabilise the binding.** We also notice some flank restrictions. The C-terminal Zn finger shows an increased preference for [AT]GATC (group B) and GATC[AT] (group C and A). **This again indicates that extra base interactions are needed for the GATA1 C-terminal Zn finger to stabilise the binding. On the contrary, the GATA1 N-terminal Zn finger can bind the “GATC” without the need for extra base interactions.**

A.

GATA1 N-terminal Zn finger specific "GATC" motif containing 8mers



B.

GATA1 C-terminal Zn finger specific "GATC" motif containing 8mers

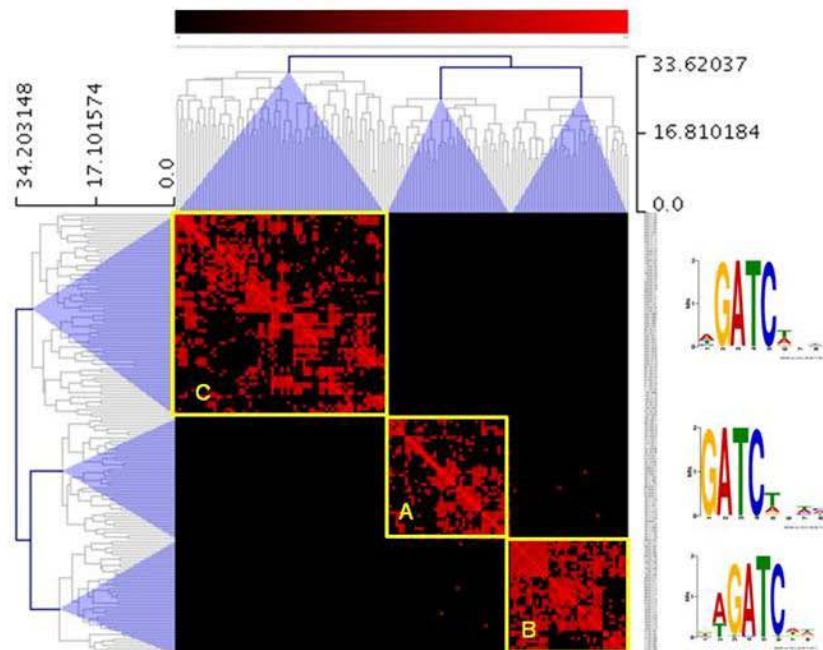


Figure 32. Binding specificities of the C-terminal Zn finger and the N-terminal Zn finger of GATA1 for the enriched "GATC" containing 8mers. The binding specificities were derived from 8mers that were bound from both the GATA1 N-terminal and the C-terminal Zn fingers (mGATA1_{NC} enriched 8mers) or the GATA1 C-terminal Zn finger alone (mGATA1_C

enriched 8mers). A) GATA1 N-terminal Zn finger specific 8mers that represent “GATC” containing 8mers that appear only in the mGATA1_{NC} bound fraction. B) GATA1 C-terminal Zn finger specific 8mers that represent “GATC” containing 8mers that appear in both the mGATA1_{NC} and the mGATA1_C bound fraction. Both (A) and (B) pictures were derived through hierarchical clustering of the DNA sequence similarity of the enriched 8mers. The similarity measure used was the “Hamming distance”. Each 8mer was assigned a similarity score against the rest of the 8mers. The similarity score can take values from 5 to 8 (red boxes on the clustergram, the higher the intensity of red the higher the similarity). A similarity score of 5, for example, indicates that in the same 5 nucleotide positions of the 8mer DNA sequence, the two 8mers have the same nucleotide base. In order to cluster only the more similar 8mers, a pair of 8mers with similarity score less than 5 was assigned a score of 0 (black boxes on the clustergram). The hierarchical clustering was performed with the default parameters of the MeV software (<http://www.tm4.org/mev.html>). The clusters were assigned in groups “A”, “B” or “C” by manual curation. The 8mers of each group were then used for de novo motif discovery (presented motifs) using the “MEME” software (Bailey et al., 2009).

3.4.2.2 Absence of the N-terminal Zn finger eliminates the binding in palindromic sequences

In the previous paragraphs we provided proofs, similarly with the already existing literature, that the GATA1 C-terminal Zn finger binds mainly the “GATA” motif and secondary the “GATT” motif. The N-terminal GATA1 Zn finger motif binds the “GATC” motif. **Considering that the EMSA-seq platform can interrogate a vast majority of the DNA binding space simultaneously, we assumed that it can give indications of the N-terminal and the C-terminal Zn finger distance distribution in an unbiased way.** For this reason, we examined in the bound 20mer library the distance distribution between the “GATC” motif (as representative of the GATA1 N-terminal Zn finger binding) and the “GATA” or “GATT” motif (as representative of the GATA1 C-terminal Zn finger binding). In Figure 33 we present the interpretation of the graphs.

In Figure 34 we see that the 20mer sequences bound by the mGATA1_C protein (the GATA1 C-terminal Zn finger) show a tandem occurrence of the “GATC” and “GATA”/“GATT” motifs (**GATCGATA, GATCGATT, GATAGATC, GATTGATC**). **We hypothesize that the occurrence of two closely found motifs (in tandem, no spacer between them) will increase the local binding affinity.** We exclude binding of two C-terminal Zn fingers in tandem because the steric hindrance cannot permit their coexistence in such a short distance (also the migration distance from the free DNA front indicates that only one protein molecule is bound on the DNA) (Figure 19).

On the contrary, the 20mer sequences bound by the mGATA1_{NC} protein show a **broad distance distribution** between the “GATC” and “GATA”/“GATT” motifs. More specifically the preferred distance distribution for the GATA-GATC conformation is mainly **GATC{N}₀₋₁₁GATA** and **GATA{N}₀₋₃GATC** and secondary **GATA{N}₄₋₁₁GATC**. Similarly for the GATA-GATT conformation the preferred distance distribution is again mainly the **GATC{N}₀₋**

$_{11}\text{GATT}$ and $\text{GATT}\{\text{N}\}_{0-3}\text{GATC}$ and secondary the $\text{GATT}\{\text{N}\}_{4-11}\text{GATC}$. We note that these distance distributions is asymmetric with the GATC-GATA/GATT conformation to have a broad distance dependence whereas the GATA/GATT-GATC conformation has a narrow distance dependence. This might be an indication that the flexible spacer between the two domains need to wrap around one of them in order for both of them to be able to bind in the GATA/GATT-GATC conformation. This wrapping around one of the domains might reduce the flexibility of conformational changes for the whole protein.

We also checked the distance distribution between the “GATG” and the “GATA”/“GATT” motif. We have already seen that the GATA1 N-terminal Zn finger shows a “GATG” preference. In Figure 35 we see that the 20mer sequences bound by the mGATA1_C protein show a tandem occurrence of the “GATG” and “GATA”/“GATT” motifs (**GATGATA**, **GATGATT**, **GATAGATG**, **GATTGATG**). The conformation where the motifs are in different strands showed preference for **TATCGATG**, **GATGTATC** and **AATCGATG**, **GATGAATC**. The 20mer sequences bound by the mGATA1_{NC} protein show mainly the same pattern as the mGATA1_C. In addition there is a preference for the **GATG**{N}_{0,1}GATA and **GATG**{N}₁GATT conformations. The conformation where the motifs are in different strands showed the extra preferences **GATG**{N}₁TATC and **GATG**{N}₁AATC.

From this we can conclude either one of the following:

- i) The N-terminal Zn finger can bind DNA independently from the C-terminal Zn finger. This means that for example the **GATGATA**, **GATGATT**, **GATAGATC** and **GATTGATC** motifs are bound by either the N-terminal or the C-terminal Zn finger.
- ii) The same preference of the mGATA1_C and mGATA1_{NC} for the “GATG”-“GATA”/“GATT” motif distance indicates that only the C-terminal Zn finger of GATA1 is involved in the DNA binding in both cases. The reduction of the

presence of the “GATG” motifs in the mGATA1_C enriched 8mers (Figure 31) indicates that the effect may not be from the reduced binding of the GATA1 N-terminal Zn finger but from the allosteric effects of the GATA1 N-terminal Zn finger on the GATA1 C-terminal Zn finger. This practically means that absence of the N-terminal Zn finger does not permit the C-terminal Zn finger to bind a subset of the “GATG” containing 8mers.

As we have seen earlier, the above are in accordance with the lack of re-distribution of the “GATG” along the sequence of the mGATA1_C enriched 8mers versus the mGATA1_{NC} enriched 8mers (Figure 31).

Although the GATA1 N-terminal Zn finger binds with high affinity to the “GATC” and “GATG” motifs it has also been shown that it binds with low affinity to the “GATA” or “GATT” motifs. For this reason, we examined whether we can see a distance distribution between the “GATT” and the “GATA” motif. In Figure 36 we see that the 20mer sequences bound by the mGATA1_C protein show the following preference **GATAGATT**, **GATTGATA** and **GATTATC**. As expected the 20mer sequences bound by the mGATA1_{NC} protein show enrichment of the same pattern as the mGATA1_C with strong preference for the **GATTATC** motif. They also show the extra binding motifs **GATT**{N}₀₋₁₁**GATA**, **GATA**{N}₀₋₁₁**GATT** when the motifs are on the same strand and the motifs **TATC**{N}₀₋₁₁**GATT**, **GATT**{N}₀₋₁₁**TATC** when the motifs are on different strands.

For the GATA-GATA distance the mGATA1_C shows preference only for the **GATAGATA** and the **GATATC** motifs. The mGATA1_{NC} shows similar preference. The mGATA1_{NC} also likes the **GATA**{N}₁₋₁₀**GATA** motifs.

3.4.3 Full length GATA1 binds on the DNA with only the C-terminal Zn finger, the N-terminal Zn finger shows no binding

The previous results were based on examining the binding of either the N-terminal Zn finger and the C-terminal Zn finger together or the C-terminal Zn finger alone. To investigate whether the N-terminal and C-terminal GATA1 tails have an effect on the DNA binding of the protein we performed EMSA-seq with the full length GATA1. We then checked the distance distribution between the “GATC” or “GATG” and the “GAT(A/T)” motif .

In Figure 49 we do not see any distance distribution between the “GATA” and the “GATC” motif. **Thus the N-terminal and C-terminal tails inhibit the binding of the N-terminal Zn finger on the DNA. We also did not see the “GATC” enrichment that was based only on the binding of the C-terminal Zn finger alone (GATAGATC or GATCGATA). This implies that the tails affect both the N-terminal and the C-terminal Zn finger. On the other hand the binding of the C-terminal Zn finger on the GATGATA was not affected significantly (Figure 52).**

For the GATA-GATT distance we found only the footprint of the C finger binding by using either the mGATA1_{NC} specific 8mers or the wt-mGATA1 specific 8mers. In particular we only see enrichment of **GATTATC**, **GATAGATT** and **GATTGATA**. This is similar with the GATG-GATA motif distance distribution.

On the contrary the GATA-GATA distance pattern is more similar to the mGATA1_{NC} pattern with preference for the **GATA{N}₀₋₈GATA** and the **GATATATC** motifs.

The above indicate that the GATA1 N-terminal and C-terminal tails restrict the DNA binding in a “GATC” motif specific manner which probably involves the N-terminal

Zn finger binding. We can hypothesize that the N-terminal tails and/or C-terminal tails are inhibiting the N-terminal Zn finger DNA binding which inhibits, in his turn, the C-terminal Zn finger DNA binding. So, the GATA1 N-terminal Zn finger cannot see the “GATC” motif, no structural change can occur and the GATA1 C-terminal Zn finger DNA binding is inhibited. On the contrary the “GATGATA” and “GATTTATC” motifs, that represent the C-terminal Zn finger binding, are adequate to impose structural changes on the C-terminal Zn finger that will prevent the N-terminal Zn finger and/or N-terminal/C-terminal tails inhibition. Interestingly, the GATA{N}₀₋₈GATA motifs seem that are able to overcome the N-terminal/C-terminal tails inhibition on both the N-terminal and C-terminal Zn fingers by permitting the DNA binding of both Zn fingers simultaneously.

We do not believe that this behaviour of full length GATA1 is artificial. Newton et al (Newton et al., 2001) have shown that the transactivation activity of DNA sequences containing only the “GATC” motif is similar for either the case when the GATA1 N-terminal Zn finger binds alone or the case when the GATA1 C-terminal Zn finger binds alone or the case when wt-GATA1 binds. They showed this by using mutant versions of the wt-GATA1 in which they were disrupting the folding of either the GATA1 N-terminal or the GATA1 C-terminal Zn finger. On the contrary, DNA sequences that contain only the “GATA” motif can be bound mainly by the GATA1 C-terminal Zn finger or the wt-GATA1 protein. The transactivation activity is twice as much, compared to the transactivation of the “GATC” containing DNA sequences from the GATA1 C-terminal Zn finger or the wt-GATA1 protein, implying that the “GATA” motif is of a higher binding affinity than the “GATC” motif, for the GATA1 C-terminal Zn finger as we have already seen. The above show that the “GATC” motif is of equal affinity for both the C-terminal and the N-terminal Zn finger of GATA1. Nevertheless, the “GATC” motif in all the binding spectrum of the GATA1 C-terminal Zn finger is at the bottom binding affinity level whereas the “GATC” motif in all the binding spectrum of the GATA1 N-terminal Zn finger is at the top binding affinity level.

This indicates that the binding affinity scale of the GATA1 N-terminal Zn finger overall is lower than the binding affinity scale of the GATA1 C-terminal Zn finger.

On the contrary, the “GATA” motif appears at the top binding affinity level of the GATA1 C-terminal Zn finger and at the bottom binding affinity level of the GATA1 N-terminal Zn finger. This is the reason that the “GATA” containing DNA sequences of (Newton et al., 2001) show an increased transactivation difference when they are bound by the GATA1 C-terminal Zn finger compared to the case when they are bound by the GATA1 N-terminal Zn finger.

Although this targeted examination of the specific motifs for each GATA1 Zn finger did indeed permit the calibration of the relative binding affinities between them, the nature of the experiment is rather binary, where the presence or absence of the appropriate GATA1 Zn finger either keeps the transactivation at the same levels as the wild type protein or it eliminates it completely. On the contrary, Newton et al (Newton et al., 2001) showed that in the DNA sequence from the promoters of the *ALAS2* and *PF4* genes, which both have a mixture of GATA1 N-terminal and C-terminal Zn fingers specific DNA binding motifs, the transactivation of the wt-GATA1 protein is lower than the transactivation of either the GATA1 C-terminal or the N-terminal Zn finger alone. Because both promoters contain a mixture of “GATC”, “GATG”, “GATT” and “GATA” motifs (Figure 37) we can hypothesize the following scenario. If the GATA1 N-terminal Zn finger or the GATA1 C-terminal Zn finger are alone they might be free to bind in multiple low and high binding affinity positions, whereas if both the GATA1 N-terminal and the C-terminal Zn fingers need to bind simultaneously then the binding might be restricted to some of the available positions. Nevertheless, knowing that the co-operative binding of the two GATA1 Zn fingers increases the affinity of all the protein for the binding site and that the high binding affinity sites of the GATA1 N-terminal Zn finger are in reality at the bottom part of the DNA binding affinities of the C-terminal Zn finger, then we would expect from the mutagenesis experiments that the transactivation activity of the wt-GATA1 to be higher than the

transactivation activity from each one of its Zn fingers, which is the reverse from what Newton et al (Newton et al., 2001) saw. On the contrary, if the N-terminal and C-terminal Zn fingers in the wild type protein restrict each other's binding then we would expect that by eliminating, through the mutagenesis, the structure of either one of them the intact GATA1 Zn finger will be able to bind unrestricted on the DNA sequence. In this case the expectation is that the transactivation activity of each one of the GATA1 Zn fingers alone should be higher than the transactivation activity of the wild-type GATA1 protein. Indeed, this is something that Newton et al (Newton et al., 2001) observed for the *ALAS2* and *PF4* promoters. Thus, the results of Newton et al (Newton et al., 2001) are in accordance with our model where the interaction of the GATA1 N-terminal and C-terminal Zn fingers with the GATA1 N-terminal and the C-terminal tails and with each other, restrict the DNA specificities of the wild type GATA1 protein.

3.4.4 Binding of two molecules of GATA1 independently or as a dimer on the DNA, changes the distance distribution between the specific GATA1 N-terminal and the C-terminal Zn finger DNA binding motifs

During the mGATA1_{NC} band shifts in some conditions we observed a slow migration band (Figure 18). We attribute this band to DNA molecules that are bound by two protein molecules simultaneously. Binding of GATA containing DNA molecules by two GATA1 proteins has also been observed in crystal structures (Bates et al., 2008).

For the GATA-GATC distance the GATA1 dimer showed mainly a preference for GATA{N}₃₋₁₁**GATC** and GATT{N}₃₋₁₁**GATC** and a lower preference for **GATC**{N}₃₋₇GATA, **GATC**{N}₃₋₆GATA (Figure 34). The GATAGATC, GATTGATC and at a lower extend the **GATC**GATA motifs are also present. The **GATC**GATT motif is absent along with the GATA{N}_{1,2}**GATC**, GATT{N}_{1,2}**GATC**, **GATC**{N}_{1,2}GATA and **GATC**{N}_{1,2}GATT motifs.

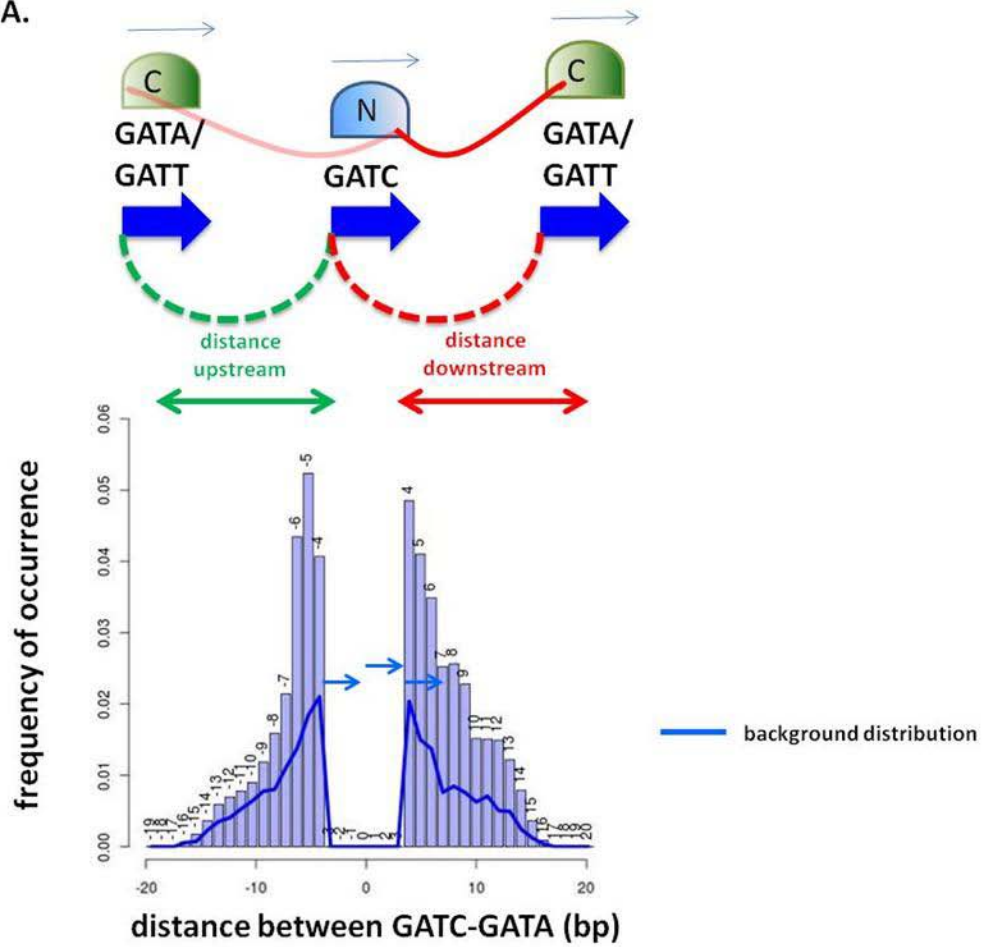
For the GATA-GATG distance the GATA1 dimer showed similar pattern with the monomeric GATA1. We also noticed an increase of GATA{N}₁₋₈**GATG**, GATT{N}₁₋₈**GATG**, **GATG**{N}₁₋₈TATC (Figure 35).

For the GATA-GATT distance the GATA1 dimer showed a similar pattern with the mGATA1_{NC} for motifs that were on the same strand (Figure 36). We notice here that there is a reduced frequency for the **GATT**{N}₁₋₁₁GATA distance. For motifs that were on opposite strand the **GATT**ATC is present. On the other hand, the TATC{N}₁₋₁₁**GATT** and **GATT**{N}₁₋₃TATC motifs are significantly reduced. There is also a significant increase of the **GATT**{N}₄₋₁₄TATC motif (Figure 36).

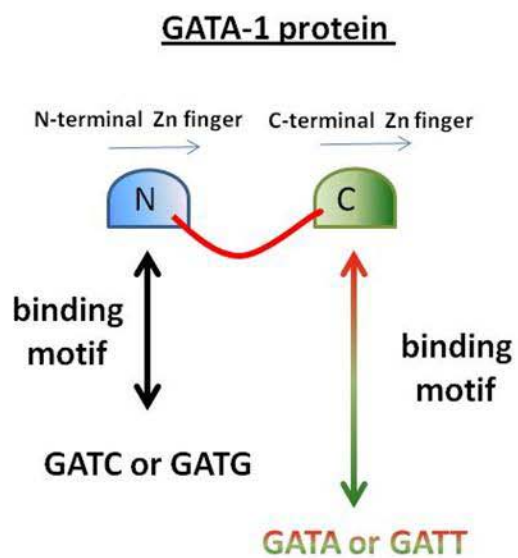
Similarly for the GATA-GATA distance, the pattern is similar with the mGATA1_{NC} one with the exception that there is a significant increase in the **GATA**{N}₅₋₁₃TATC motif.

The binding model of the mGATA1_{NC} on the DNA sequence, based on all the above observations, is presented in Figure 38.

A.



B.



C.

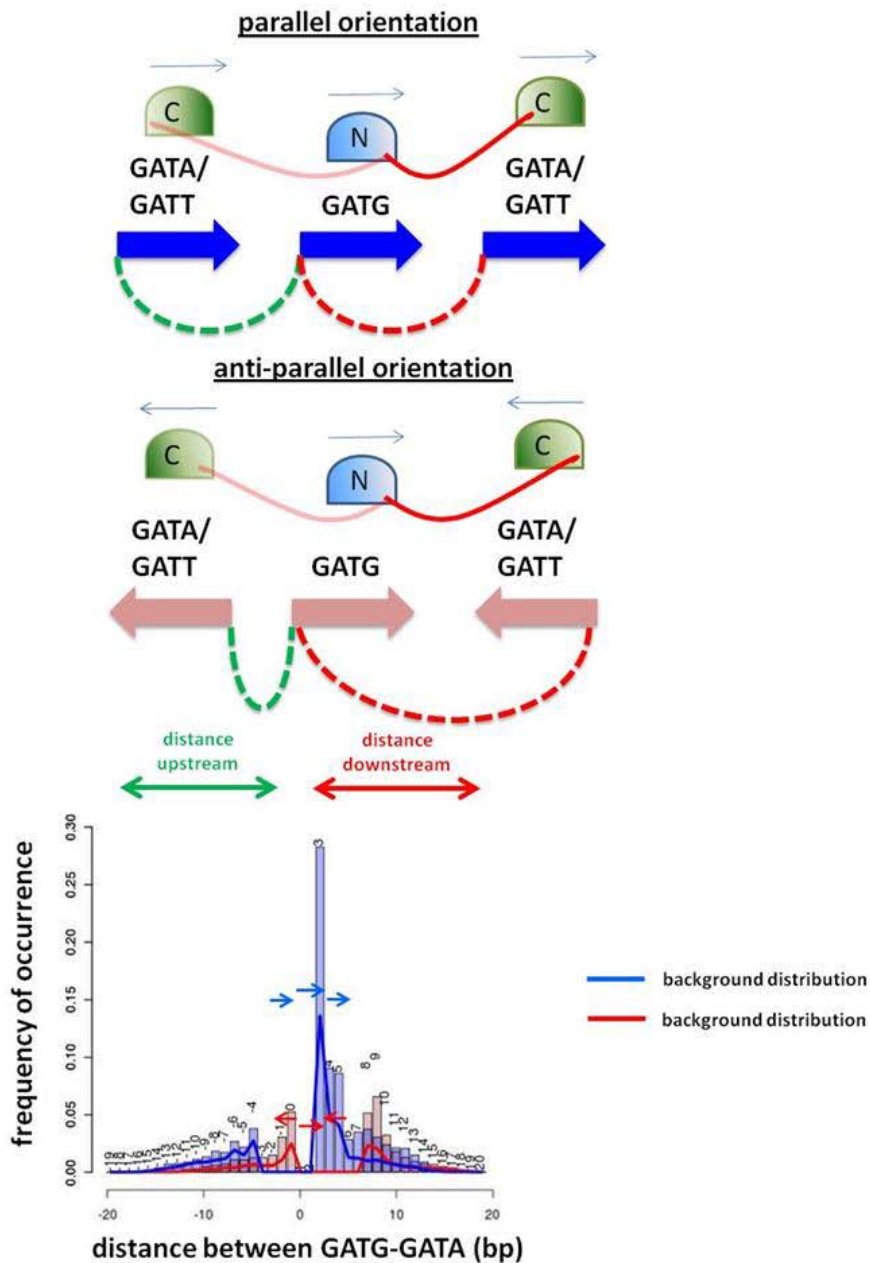


Figure 33 (previous two pages). Interpretation of the distance distribution graphs between the “GATA”/“GATT” motif and the “GATC” motif (A) or the “GATA”/“GATT” motif and the “GATG” motif (C). The working hypothesis is presented in (B). In more detail, the GATA1 N-terminal Zn finger is assumed to bind the “GATC” or “GATG” motif whereas the C-terminal Zn finger is assumed to bind the “GATA” or “GATT” motif. The x-axis in the bar plots of (A) represents the distance in bp between the beginning of the “GATC” motif and the beginning of the “GATA”/“GATT” motif as indicated in the schematic graphs. The value

of the distance in bp is indicated above the bars in the plots (the absolute value of the indicated number represents the distance). Negative values represent the “GATA”/”GATT” motif being upstream of the “GATC” motif. Positive values represent the “GATA”/”GATT” motif being downstream of the “GATC” motif. The appearance of the “GATA”/”GATT” motif and the “GATG” motif on the same sequence can have two orientations the parallel one and the anti-parallel one (C) depending on which strand the motifs are. The interpretation of the parallel orientation is similar to the GATA/GATT-GATC plots (blue bars in the bar plot). In the anti-parallel orientation, the “GATA”/”GATT” motif upstream of the “GATG” motif faces a different direction to the “GATG” motif, whereas the “GATA”/”GATT” motif downstream of the “GATG” is in the same direction as the “GATG” motif . Similar plots like the GATA/GATT-GATG distance plots can be created for the GATA-GATT or the GATA-GATA distance plots.

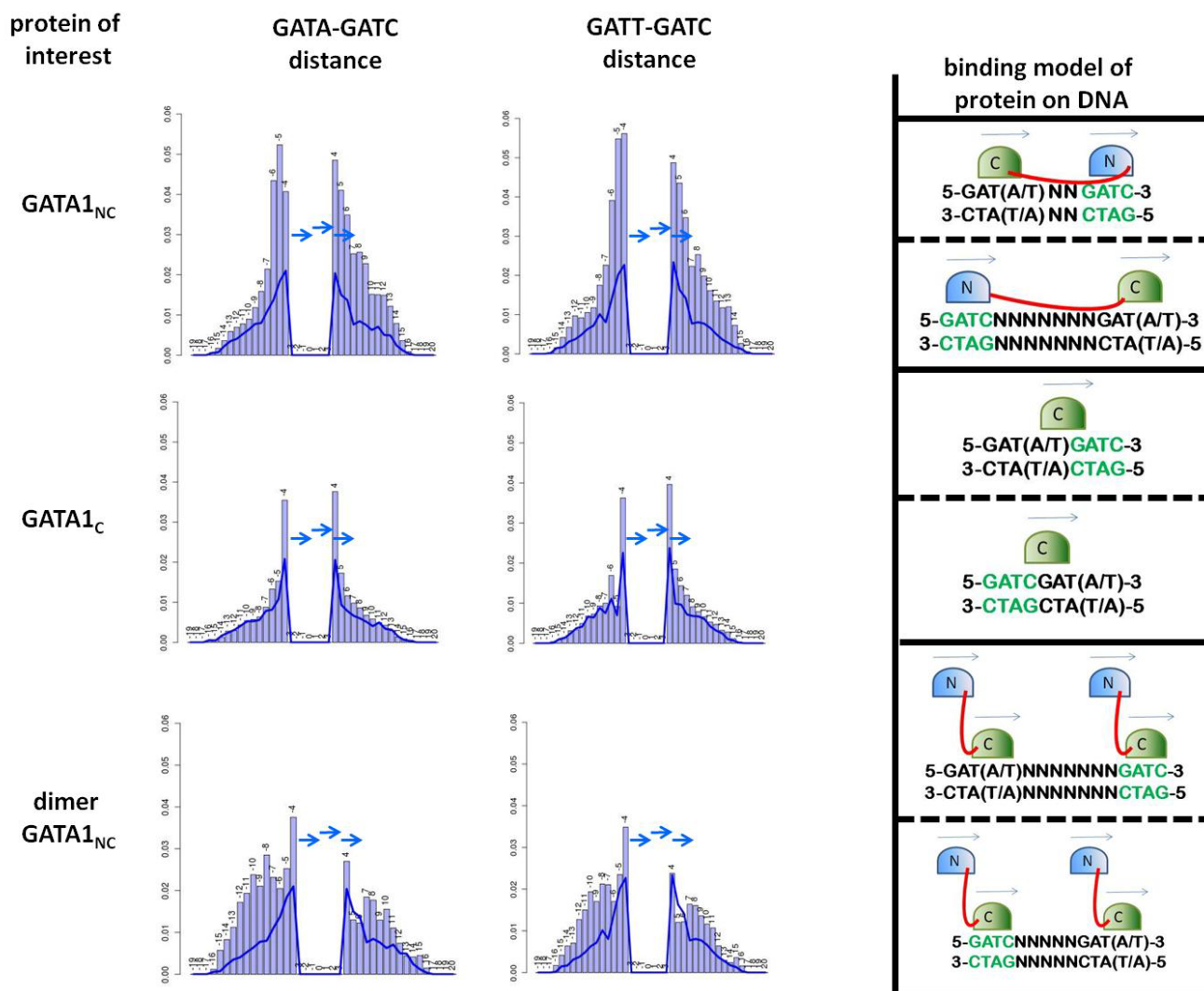


Figure 34. Distance graphs between the “GATA” and the “GATC” motifs or between the “GATT” and the “GATC” motifs for the monomeric mGATA1_{NC}, mGATA1_C and the dimeric mGATA1_{NC}. The distance graphs for the monomeric and dimeric mGATA1_{NC} were created with the 2122 enriched 8mers of the monomeric mGATA1_{NC} (see materials and methods section for further details). The distance graphs for the mGATA1_C were created with the top 2122 enriched 8mers of the mGATA1_C. The binding models of the protein on DNA are also shown. For further explanation about the graph see Figure 33. The blue arrows represent the orientation of the motifs as appear in Figure 33A.

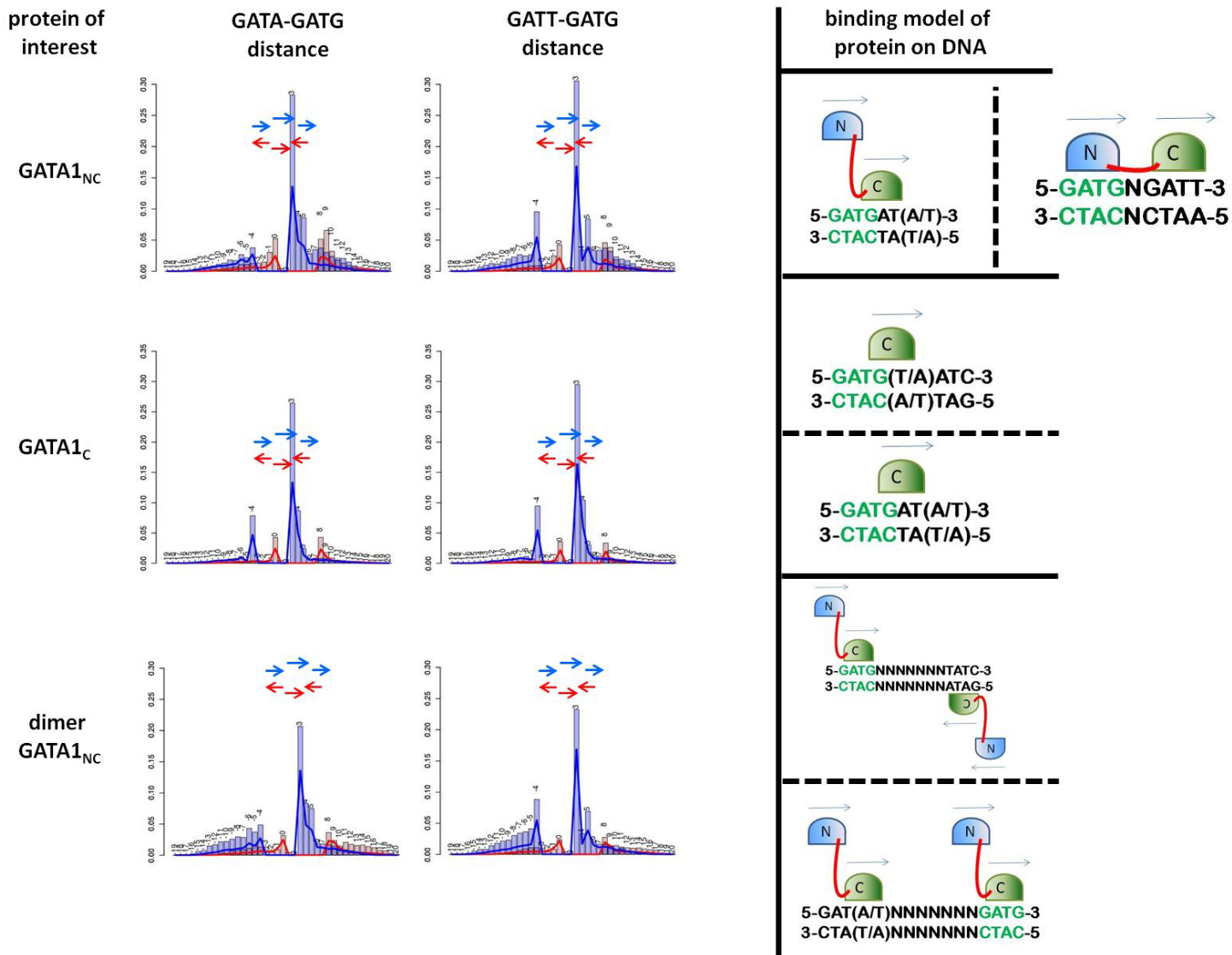


Figure 35. Distance graphs between the “GATA” and the “GATG” motifs or between the “GATT” and the “GATG” motifs for the monomeric mGATA1_{NC}, mGATA1_C and the dimeric mGATA1_{NC}. The distance graphs for the monomeric and dimeric mGATA1_{NC} were created with the 2122 enriched 8mers of the monomeric mGATA1_{NC} (see materials and methods section for further details). The distance graphs for the mGATA1_C were created with the top 2122 enriched 8mers of the mGATA1_C. The binding models of the protein on DNA are also shown. For further explanation about the graph see Figure 33. The blue and red arrows represent the orientation of the motifs as appear in Figure 33C.

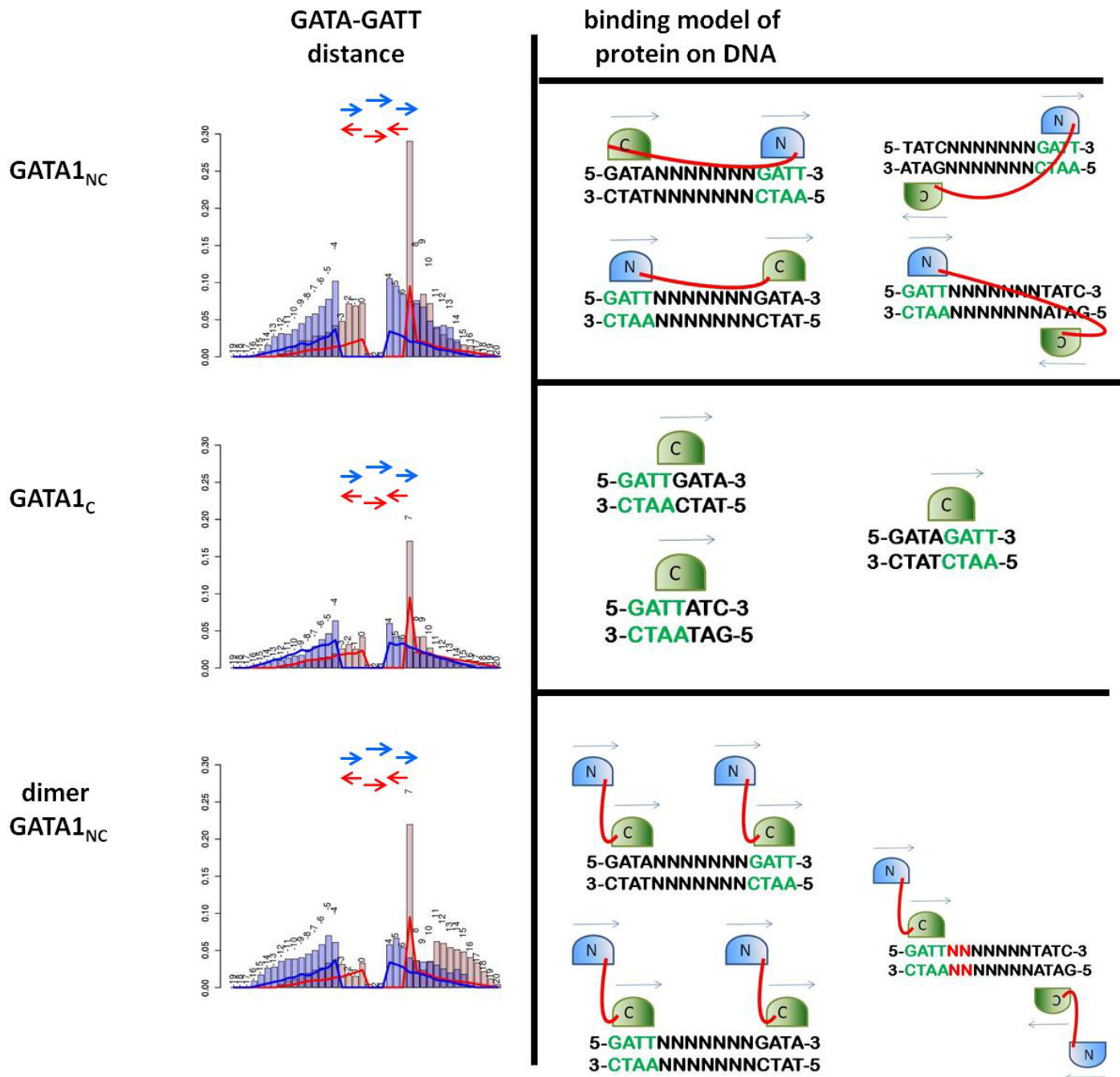


Figure 36. Distance graphs between the “GATA” and the “GATT” motifs for the monomeric mGATA1_{NC}, mGATA1_C and the dimeric mGATA1_{NC}. The distance graphs for the monomeric and dimeric mGATA1_{NC} were created with the 2122 enriched 8mers of the monomeric mGATA1_{NC} (see materials and methods section for further details). The distance graphs for the mGATA1_C were created with the top 2122 enriched 8mers of the mGATA1_C. The binding models of protein on DNA are also shown. The red nucleotides in the binding models of the third row indicate that short distances between the motifs are

not favoured. For further explanation about the graph see Figure 33. The blue and red arrows represent the orientation of the motifs as appear in Figure 33C.

(mouse *ALAS2* promoter , from -300 to +37 nucleotide position):

GTCCTCTCCACGTGGACTTGCCTTTAGTGGCCTCCTTCATGGCTGAGGGA
ACTGAAGGTCAGAGGGATGGTTCTACAGAAGGTGCCCTGGATTATAAGG
ATGGTCTGATCTCAAATCGAATGAGTTAGTTTCACTACCCCAGAGGGAG
TAGAGGAAGAAAACAGAAAGAGAAGGCACAGAACAAAGGGGAGTGAAAG
GA GACTACTTTGCTTTTATCTTTAGCAGAGGCCTGAGATACCTTAAGGCCG
AGGCGGAAGGGGGTGGGGTGGGGTGGGCGGGGCTGAGCAGGAGGCTG
GGA TGAAAGCTGCTTCCCAGCCCAGCCCACAGCACTCACCG

(mouse *PF4* promoter , from -355 to +19 nucleotide position):

CACCACAATCGGCAGGAAGAGTGACAGTGGCTTTGGAGTTACGTGACTGG
AAGCCAAGAAGGCAACACATTGATACTGAAGGTGTGTGTTCTGCAAACCC
GATGGCGATCTCAGAAAGTTCTCTGCCCCATACAGCATACCTTTTGCTAA
ATTCCAAGTCCTACTGTTTCTCACTCTGTAATTGTTGATATCAATTACACA
TTGGAGATGTGGATCCTGCTGCCAGCTACTAGTCAAGACAGCTGGCCTCA
GCTTTTTTTTTTTTTTTTTTCTGCTCTGTTTTATTTAATCTTGGCTGGC
CAGATCCCAATTACTGTTCCACAAGTGTCAATTGCTTCTGCAGAGCACTTC
CTCATCCTGTATCCTGGGTTTCCGG

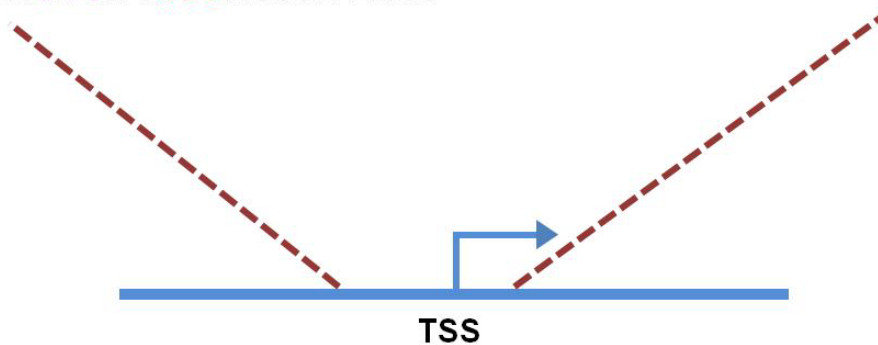


Figure 37. Nucleotide sequence of the promoter region for the *ALAS2* and the *PF4* gene between the indicated genomic positions. The GATA1 motifs “GATA”, “GATT”, “GATG”, “GATC” are indicated with different colours.

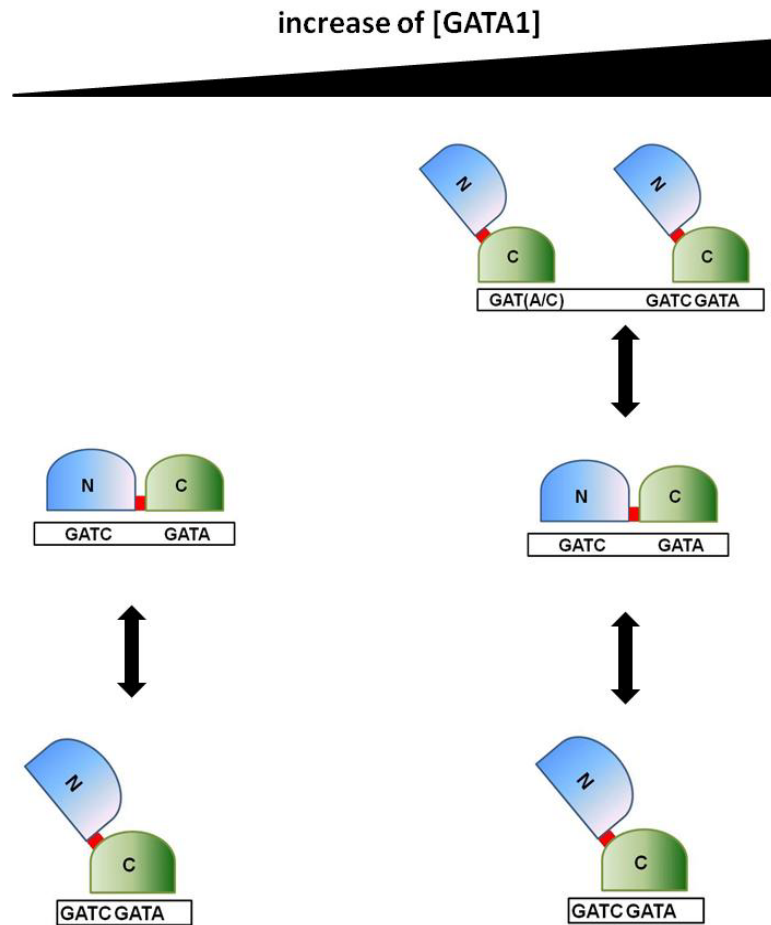


Figure 38. General binding model of the GATA1_{NC} protein derived from the EMSA-seq experiments. In low concentration of protein, GATA1 can bind with either the C-terminal Zn finger alone or with both the N-terminal Zn finger and the C-terminal Zn finger. In high concentration of protein, GATA1 can also bind as a dimer. Probably only the C-terminal Zn fingers of the dimeric GATA1 molecules are used for binding when the sequence has two “GATA” motifs. In case of the dimeric GATA1, when the sequence has only one “GATC” motif and one “GATA” motif, the dimer can still probably bind with the C-terminal Zn fingers. In this situation, the interaction of the two GATA1 proteins will probably stabilise the binding.

3.4.5 Integration of the GATA1 EMSA-seq results with other in vitro data

3.4.5.1 DNA motif distances of mGATA1_{NC}, mGATA1_C and their relationship with solved crystal structures

Although the distance distributions represent potential protein conformations upon DNA binding the best way to check whether they can reflect the DNA-protein interaction conformations is to compare them with solved crystal structures. DNA Binding by a single GATA zinc finger has been analysed by NMR on the chicken GATA1 C-terminal Zn finger (Omichinski et al., 1993, Tjandra et al., 1997, Clore et al., 1994) and on its fungal homologue AREA (Starich et al., 1998a, Starich et al., 1998b). Crystal structures of the GATA zinc fingers bound to DNA have been analyzed on the mouse GATA3 C-terminal Zn finger alone (Bates et al., 2008) and on the human GATA3 C-terminal Zn finger and N-terminal Zn finger when they are together on the protein sequence (Chen et al., 2012). Below we present some mechanistic details from these binding models. We eventually compare them with our models. We should always have in mind that the crystal structure represents a protein DNA conformation that was able to crystallize under the specific crystallization conditions. The crystallization conditions might not always favor the most thermodynamically stable conformation for the protein-DNA complex. Crystal stacking effects might also distort the protein-DNA binding. **From our frequency data of the motif distance distributions, we can assign the most frequent conformation as the most thermodynamically favorable one (independently from the generic DNA sequence context).** The unbiased nature of our experimental setup permits us to be able to approximate the thermodynamically stable condition and infer the dynamics of the protein-DNA complexes. For this reason the frequency data from the EMSA-seq can be assumed that they lie along the lines of simulations from molecular

dynamics experiments and from NMR dynamics experiments (Kleckner and Foster, 2011). Below we present the GATA3/DNA binding models from the crystal structures of the above mentioned studies and the projection of our data on them. We use the GATA3/DNA binding models because at the time of writing this manuscript the only available protein/DNA crystallographic models from the proteins of the GATA family are the ones from GATA3.

Chen et al (Chen et al., 2012) have crystallized the N/C-terminal Zn fingers of human GATA3 (hGATA3_{NC}) along with the following palindromic molecules a) AATGTCCATCTGATAAGAC (**GATApal**) and b) CCTAAATCAGAGATAACC. They noticed that the hGATA3_{NC} can bind in the **GATApal** with the C-terminal Zn finger bound on the “GATA” motif and the N-terminal Zn finger bound on the “CATC” (“GATG”) motif. In the second sequence, they noticed that the hGATA3_{NC} can bind simultaneously on two different DNA molecules that are aligned in parallel in the crystal structure. In this case the GATA1 C-terminal Zn finger can bind on the “GATA” motif whereas the N-terminal Zn finger can bind on the GATT motif of the other molecule. **They interpret this as an indication that GATA1 can bridge two DNA binding sites that are distant apart.** They also showed that this can happen in solution. **These observations are similar to what we have seen where the GATA1 N-terminal Zn finger and the GATA1 C-terminal Zn finger can bind motifs that are further distant apart as long as a flexible DNA sequence exists between them (Figure 39).**

Bates et al (Bates et al., 2008) also presented the crystal structure of proximal GATA sites with the C-terminal Zn finger of mouse GATA3 (mGATA3_C). This complement the NMR analyses of the chicken GATA1 C-terminal Zn finger (cGATA1_C) (Omichinski et al., 1993). Bates et al (Bates et al., 2008) used the following sequences:

- i) A sequence that contains **two “GATA”** motifs on the same strand and **one “GATT”** motif in tandem with one “GATA” motif but on the reverse strand. We will call the



sequence “**OPP complex**” and it is the following, TTGATAAATCAGAGATAACC (The AATC represents the GATT sequence). In the crystal structure, two molecules of mGATA3_C bind the second “GATA” motif (colored green) and the “GATT” motif.

- ii) A sequence that contains **two** “GATA” motifs, that are four nucleotides apart, on different strands. Bates et al call the sequence “**ADJ**” and it is the following, TTCT[→]GATAA[←]GA[←]CTTATCTGA. In the crystal structure two molecules of mGATA3_C bind the “GATA” motifs. EMSA also showed that two wt-mGATA3 bind on this sequence
- iii) A sequence that contains **two** “GATA” motifs, that are three nucleotides apart, on different strands. Bates et al call it sequence “**OPP**” and it is the following, TAT[←]TATCAGA[→]GATAAAC. EMSA showed that either one or two wt-mGATA3 can bind on this sequence depending on the protein concentration.
- iv) A sequence that contains **one** “GATA” motif and **one** “GATC” motif on different strands. This sequence was exactly the same as the sequence “**OPP**” with the exception that one “GATA” motif was replaced with the “GATC” motif. Bates et al, call it sequence “**OPP A-C**” and it is the following, TAG[→]GATCAGA[→]GATAAAC. EMSA showed that one mGATA3 can bind this sequence.

From the above sequences, Bates et al presented 2 crystal structures (for sequences (i),(ii)) and proposed 2 hypothetical models (for sequences (iii),(iv)).

On our side, we assumed that the mGATA1 binding affinities will be representative of the mGATA3 binding preference. We hypothesized that the binding affinity values (Kd) that were extracted from EMSA-seq will be able to reproduce to some extent the binding conformation of the protein-DNA complex. In Figure 40 we see the projection of the Kd values of the mGATA1_{NC} 8mers on the sequences (i),(ii),(iii) and (iv). **We notice that the**

EMSA-seq binding affinities (K_d) represent a very good approximation of the thermodynamically favorable protein-DNA conformation.

For the “**ADJ**” sequence we can see two positions with high DNA binding affinity (K_d approximately 10 pM) at the edges of the sequence. These two positions are separated by a region of low DNA binding affinity (K_d approximately 13 pM). Two molecules of the wt-mGATA3 can bind simultaneously at this sequence at the two high binding affinity regions.

For the “**OPP A-C**” sequence, we can see one position with high DNA binding affinity (K_d approximately 10 pM) at the right edge of the sequence. At the left edge, the mean binding affinity is around 12 pM (mGATA1_{NC} 8mers with K_d equal to 11pM, 12pM and 13pM). The C-terminal Zn finger of wt-mGATA3 can probably bind on the high DNA binding affinity region on the right whereas the N-terminal Zn finger can bind on the low DNA binding affinity region on the left.

For the “**OPP**” sequence we can see that the binding affinity throughout the sequence is homogeneous, with K_d around 10pM. Nevertheless, we can see a trend where the right part of the sequence has higher binding affinity (K_d approximately 9.5 pM) and the left part of the sequence has lower binding affinity (K_d approximately 10.5 pM). In low protein concentration one molecule of wt-mGATA3 binds on the DNA whereas in high protein concentration two molecules of wt-mGATA3 bind on the DNA. We can assume a metastable situation where at low protein concentration the C-terminal Zn finger of wt-mGATA3 binds the 9.5 pM region and the N-terminal Zn finger binds the 10.5 pM region, whereas at high protein concentration the C-terminal Zn finger of a second wt-mGATA3 molecule can compete the N-terminal Zn finger, of the first molecule, for binding at the 10.5 pM region.

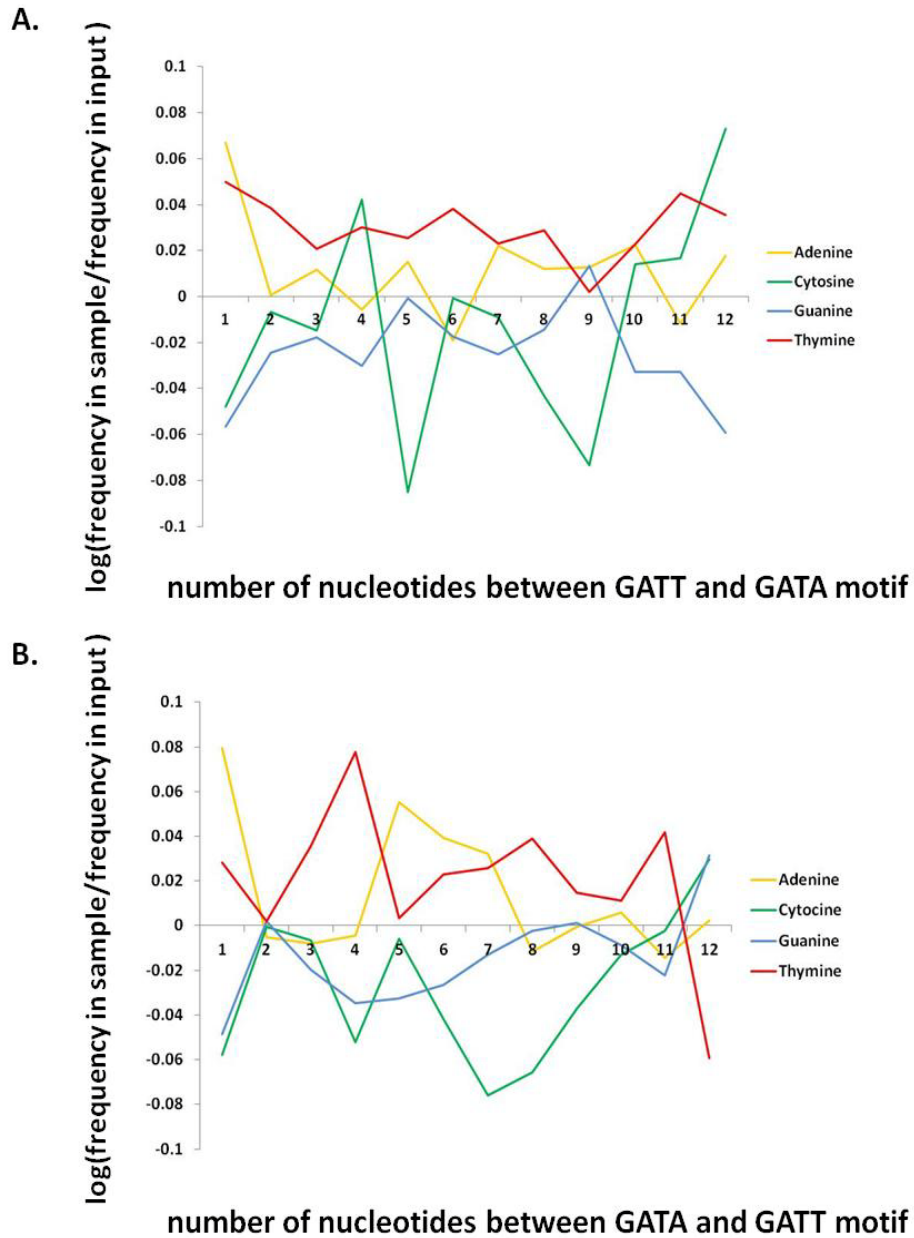


Figure 39. Fold change of the nucleotide frequency occurrence (Adenine, Cytosine, Guanine, Thymine) between the “GATA” and the “GATT” motifs for the mGATA1_{NC} EMSA-seq condition 1 experiment (sample) versus the original 20mer degenerate DNA library (input). The distribution of the fold change for both the GATT-GATA (A) and GATA-GATT (B) orientations is shown. In case where there are two or more nucleotides between the “GATA” and the “GATT” motifs, the occurrence of each one of the four nucleotide bases is summed across all nucleotides for both the sample and the input independently and the fold change is calculated.

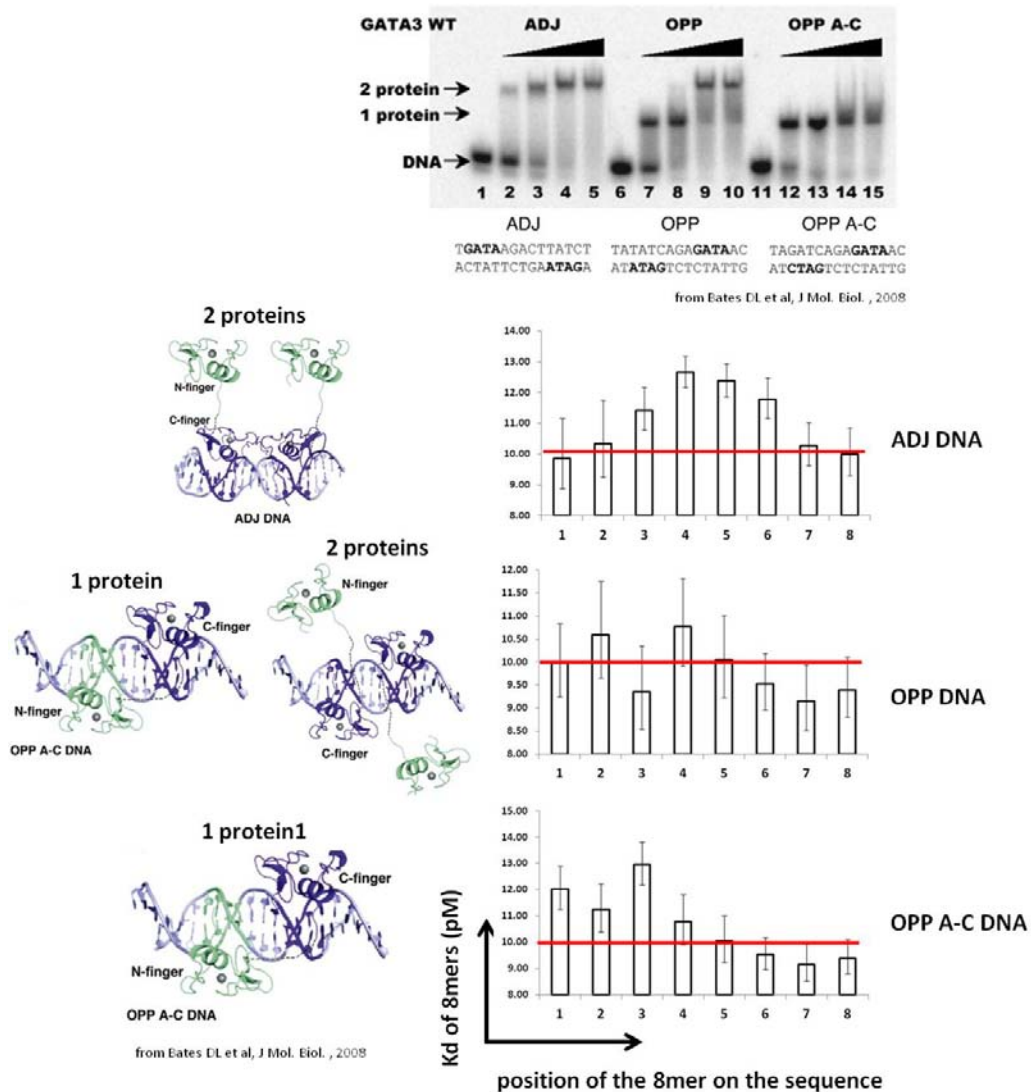


Figure 40. Comparison between the proposed models of GATA3 binding on 3 different DNA sequences (**ADJ**, **OPP**, **OPP A-C**) and the projected DNA binding affinities (Kd) of the mGATA1_{NC} 8mers, on the overlapping 8mers of the DNA sequence. No available EMSA-seq Kd values exist for GATA3 so the mGATA1_{NC} Kd were used. The GATA3 binding models were from (Bates et al., 2008). We also present the EMSA shifts of the same DNA sequences bound by GATA3 where it is visible the migration of the protein-DNA complexes that have one and two GATA3 proteins (Bates et al., 2008).

3.4.5.2 DNA motif distances of mGATA1_{NC}, mGATA1_C and their relationship with biochemical studies

In the sequences that are presented in this section it is advisable for the reader to focus on the existence or not of the “**GATA**”, “**GATC**”, “**GATC**”, “**GATT**” motifs and how big is the spacing between them in case that they co-occur in the same sequence.

Trainor et al (Trainor et al., 2000) have shown that the hGATA1_{NC} binds a sequence that contains **one** “**GATC**” and **one** “**GATA**” motif (GGATCAGATAAG) with higher affinity than a sequence that contains **two** “**GATA**” motifs (GGATAAGATAAG, one nucleotide spacer between the “GATA” motifs) ($K_d=8\text{nM}$ for the GATC/GATA sequence versus $K_d=22\text{nM}$ for the GATA/GATA sequence). This was independent of the sequence context because it was noticed for another set of sequences (AGCTTCGGGATAAAGATCTTAAATTC versus TCGCTcagAGATAAGGCCTTCAGAGTGC). Also, competition experiments showed that the hGATA1_{NC} bound to a sequence mentioned in the literature as **GATApal** (AGTCCATCTGATAAGACTTCAGTGCTGCCC / α -globin promoter) can be effectively competed from a sequence that contains both “GATA and GATC” motifs (AGCTTCGGGATAAAGATCTTAAATTC) but not only “GATA” motifs (AGCTTCGGTTGCAGATAAACATTGAATTCA). The cGATA1_C showed the reverse behavior.

DNase footprinting experiments showed that the cGATA1_C protects only the GATA motif whereas the hGATA1_{NC} protects both the “GATA and GATC” motifs or the “GATA and GATG” motifs (Trainor et al., 2000).

As we have already seen from the EMSA-seq results, the above indicate that both the N-terminal and the C-terminal Zn finger of hGATA1_{NC} bind simultaneously on

the DNA sequence. Apart from the GATA-GATC distance plots, as we will discuss later, the EMSA-seq platform through the EMSA-seq coefficient (“e”) did indeed indicate to us the distinct role of the “GATC” motif (as an indication of the simultaneous binding of both the GATA1 N-terminal and the C-terminal Zn finger) when it is compared with the other GATA1 motifs.

Trainor et al (Trainor et al., 2000) also showed that the transactivation induced from the “GATC” containing sequence $\xrightarrow{\quad} \text{(GATCCTGCAGATAAACATG)} \xrightarrow{\quad}$ was lower than the sequence that has the two “GATA” motifs $\xrightarrow{\quad} \text{(GGATAAGATAAG)} \xrightarrow{\quad}$. This implies that either different GATA1 complexes bind the two sequences or that the protein has different conformations (which will eventually lead to complexation with different factors) upon binding on the two different sequences. As we have seen from the EMSA-seq, the GATA1 C-terminal Zn finger cannot interact with two “GATA” motifs that are one nucleotide spacer (Figure 41). On the contrary, the GATA1 N-terminal and GATA1 C-terminal Zn fingers can simultaneously bind two “GATA” motifs, independently of one another (Figure 41). Thus, the differential transactivation behavior should be assigned to the N-terminal Zn finger of GATA1. We have already hypothesized on the context of the wt-mGATA1 the “GATC” motif specific N-terminal Zn finger specific binding behavior. All the above confirm the validity of the EMSA-seq results.

Interestingly, the data of Trainor et al (Trainor et al., 2000) also show that the $\xrightarrow{\quad} \text{AGCTTCGGGATAAAGATCTTAAATTC} \xrightarrow{\quad}$ sequence that contained both GATA and GATC motifs did not show any transactivation at all independently of the number of repeats in the reporter plasmid. In the $\xrightarrow{\quad} \text{AGCTTCGGGATAAAGATCTTAAATTC} \xrightarrow{\quad}$ we notice the “GATANNGATC” orientation whereas in the $\xrightarrow{\quad} \text{GATCCTGCAGATAAACATG} \xrightarrow{\quad}$ we notice the “GATCNNNNNGATA” orientation. Based on the EMSA-seq results both orientations, with the given number of spacers, are permitted. Still, it is interesting to point

out that the EMSA-seq results show a clear preference for the “**GATC-GATA**” orientation rather than the “**GATA-GATC**” orientation (the asymmetric distance pattern discussed earlier) and this might be a reflection of the transactivation behavior. Maybe complexation with other co-factors, necessary for the transactivation activity, will put extra conformational strains on the already restricted “**GATA-GATC**” orientation whereas the flexibility of the “**GATC-GATA**” orientation might make the protein more resistant to conformational strains.

Lastly, we present some other sequences that were tested for binding from the hGATA1_{NC} or the hGATA1_C protein. Trainor et al (Trainor et al., 1996) showed that if you take the chicken GATA1 promoter (TCGCT**TATCAGATA**AGGCCT**TATC**AGTGC) and you mutate the second “**GATA**” motif (TCGCT**TATCAGATA**AGGCCTTcagAGTGC) (we will call it “sequence 1”) the hGATA1_{NC} or full length hGATA1 proteins bind with higher affinity than the hGATA1_C protein. If you also mutate the “**TATC**” motif (TCGCTcag**AGATA**AGGCCTTcagAGTGC) (we will call it “sequence 2”) only the hGATA1_C protein can bind with high affinity. In “sequence 1” the two “GATA” motifs have the TATC{N}₁GATA orientation. In Figure 41 this is an orientation that the mGATA1_{NC} shows a preference. On the contrary, the mGATA1_C does not have a preference.

Similarly the AGTCCATCT**GATA**AGACT**TATC**TGCTGCCC (we will call it “sequence 3”) but not the AGTCCATCT**GATA**AGACTTcagTGCTGCCC sequence (we will call it “sequence 4”) was bound with higher affinity from the hGATA1_{NC} rather than the hGATA1_C protein whereas the AGCTTCGGTTGCAG**GATA**AACATTGAATTCA (β/e sequence) was bound equally well from both of them. In “sequence 3” the two “GATA” motifs have the GATA{N}₅TATC orientation. As we have seen in Figure 41 this is an orientation that the mGATA1_{NC} does not show any preference. On the contrary, only the dimeric mGATA1_{NC} shows preference. Although, Trainor et al do not provide any evidence whether hGATA1_{NC} binds as a monomer or as a dimer, it is interesting to

speculate that the observed higher affinity of “sequence 3” versus “sequence 4” can be attributed to the binding of two hGATA1_{NC} molecules. **Although it is not unreasonable to speculate that, because two hGATA1_{NC} proteins can bind nearby “GATA” motifs, we should expect two hGATA1_C to be able to bind the same orientation of “GATA” motifs, the EMSA-seq results of Figure 41 indicates to us that these expectation is false. Thus, the dimeric conformation of GATA1 is a special state that might: i) stabilize the binding and/or ii) reduce the steric hindrance of nearby binding sites by forcing the proteins into a specific conformation.**

Ghirlando et al (Ghirlando and Trainor, 2000) have shown that the GATA1 protein, upon binding on a DNA sequence, bends it at a specific angle irrespectively of the DNA sequence. The authors prove that this bending is mediated only from the GATA1 C-terminal Zn finger binding on the DNA sequence although both Zn fingers of GATA1 bind on the DNA sequence. This bending probably results from the insertion of the C-terminal basic residues, that are present after the GATA1 C-terminal Zn finger (Figure 3), on the minor groove of the DNA sequence (Ghirlando and Trainor, 2000). The presence of the GATA1 N-terminal Zn finger does not have any effect on the overall bending of the DNA sequence although the authors used a variety of DNA sequences with different distances between the GATA1 C-terminal and the N-terminal Zn finger binding sites. The authors claim that this indicates that GATA1 changes structure upon binding rather than the DNA itself is bending in order to accommodate the binding of the two GATA1 Zn fingers. In the GATA-GATC distance plots we saw that there exist positional combinations of motifs where the C-terminal Zn finger and the N-terminal Zn finger of GATA1 bind with no distance restriction between them (for example mGATA1_{NC} binds mainly the **GATC{N}_kGATA** motif, for $k=0..8$). Additionally, in the same distance plots we also found positional combinations of motifs where the distance between the GATA1 Zn fingers is restricted to nearby motifs (**GATA{N}_kGATC** distance, for $k \in (0,3)$). This behavior points towards the existence of a very flexible spacer that connects the GATA1 DNA binding

domains. Overall, the module “Zn finger-spacer-Zn finger” has shown a big variety of conformations. This indeed confirms the hypothesis of Ghirlando et al (Ghirlando and Trainor, 2000) that GATA1 presents a variety of conformational changes upon DNA binding.

On the other hand, we examined whether the DNA sequence between the “GATT” motif and the “GATA” motif helps the binding. We hypothesized that if only the GATA1 protein changes conformation in order to be able to accommodate the binding then the “A”, “C”, “G”, “T” nucleotide distribution for each nucleotide position between the “GATT” and the “GATA” motifs should be similar. In Figure 39 we see that the “A”, “T” nucleotides are more frequent than the “G”, “C” nucleotides. **Considering that the “A”, “T” nucleotides make the DNA sequence more flexible in bending than the “G”, “C” nucleotides this indicates that the DNA sequence itself facilitates the GATA1 binding. Thus, in contrary to Ghirlando et al (Ghirlando and Trainor, 2000), we argue that both the GATA1 conformational changes and the DNA content between the motifs help the binding. We can expect different DNA bending conformations based on the DNA content of the sequences.**

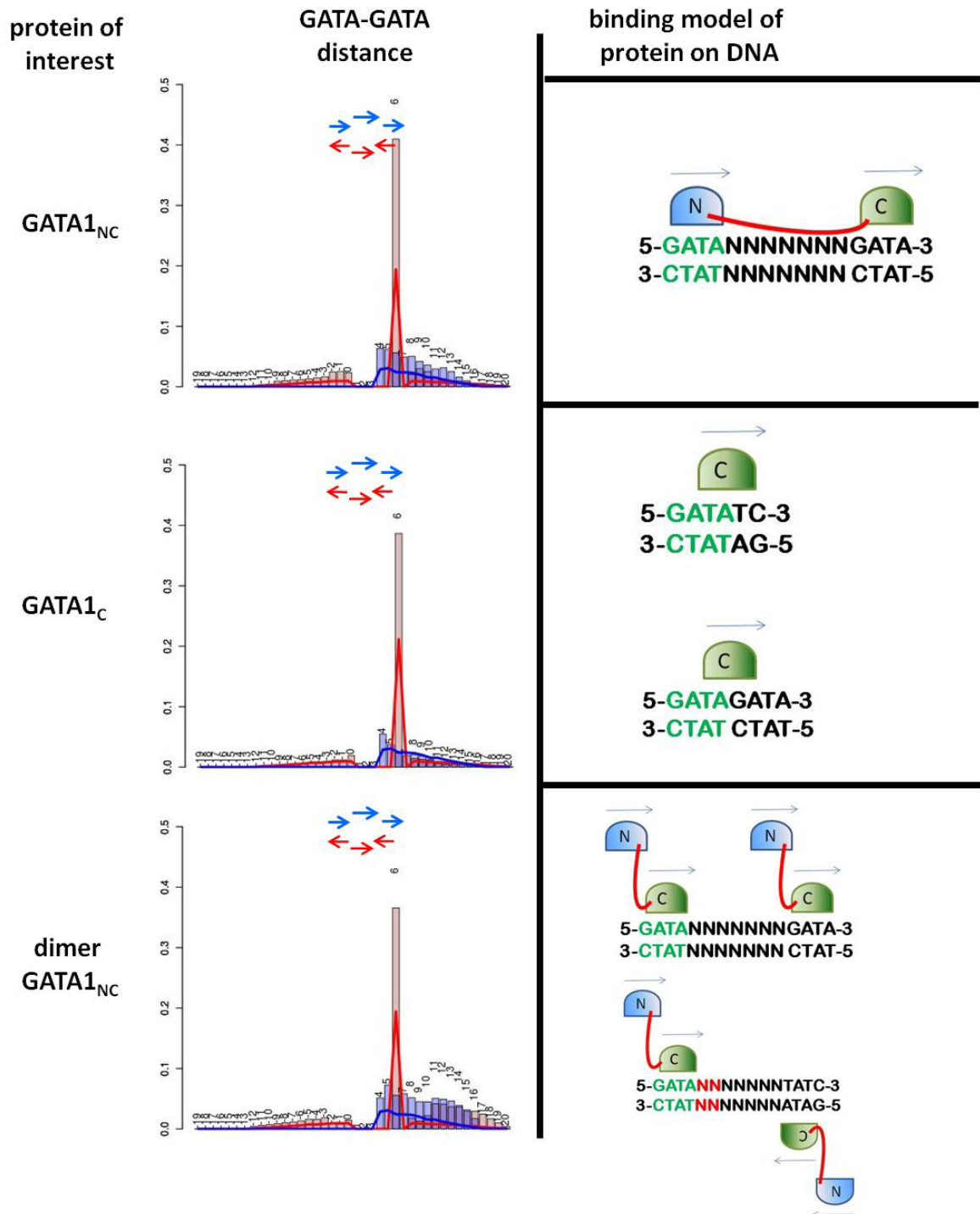


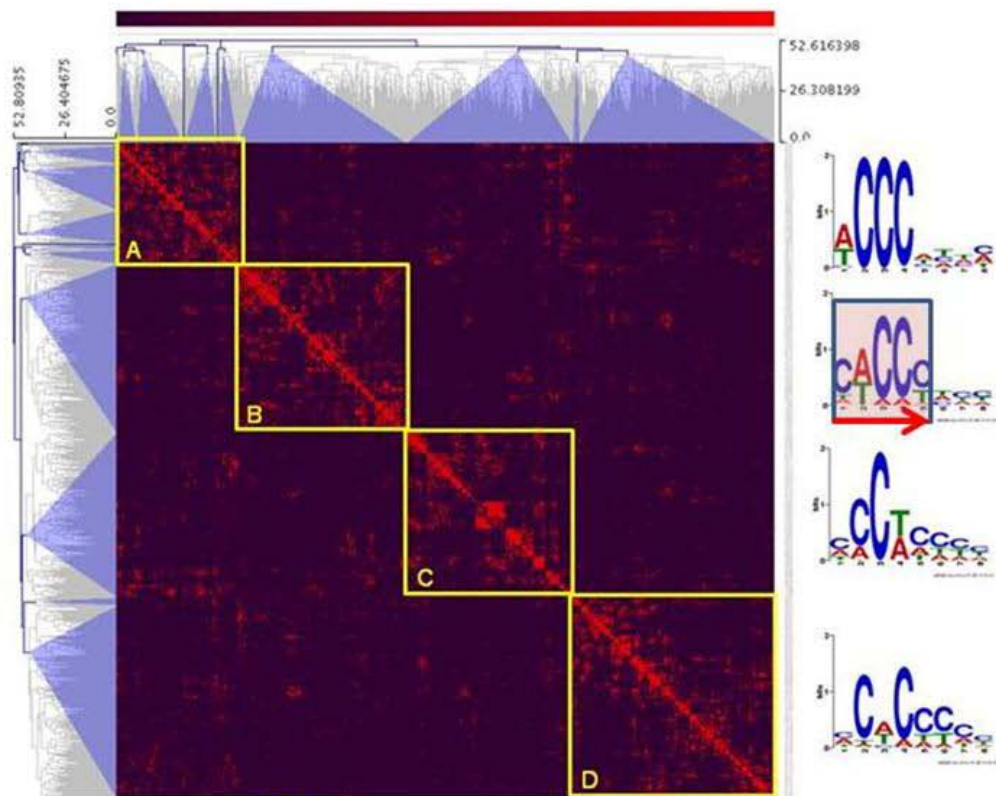
Figure 41. Distance graphs between the “GATA” and the “GATA” motifs for the monomeric mGATA1_{NC}, mGATA1_C and the dimeric mGATA1_{NC}. The distance graphs for the monomeric and dimeric mGATA1_{NC} were created with the 2122 enriched 8mers of the monomeric mGATA1_{NC} (see materials and methods section for further details). The distance graphs for the mGATA1_C were created with the top 2122 enriched 8mers of the mGATA1_C. The binding models of protein on DNA are also shown. The red nucleotides in

the binding models of the third row indicate that short distances between the motifs are not favoured. For further explanation about the graph see Figure 33. The blue and red arrows represent the orientation of the motifs as appear in Figure 33C.

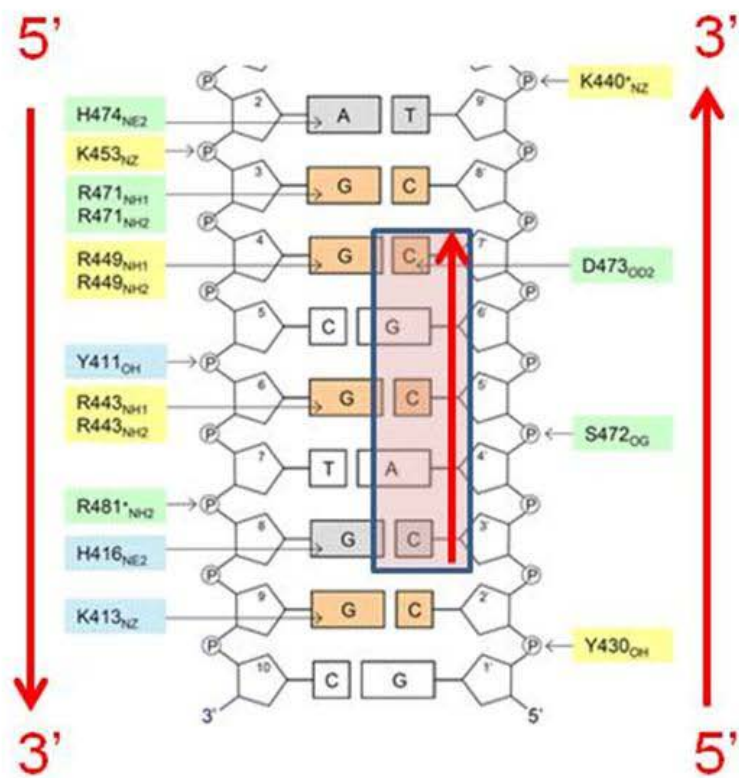
3.4.6 DNA binding specificities of KLF1

KLF1, as other members of the same family, has been shown to bind the “GT-box” or “CACC” sites (Bieker, 2001, Kaczynski et al., 2003). KLF1, in more detail, has been shown to have preference for the CAACC element (Miller and Bieker, 1993) and mutations that change the CCACACCCCT sequence to CCACATTCCT or CCACACCGCT or CCACACCCGT, in the promoter of the beta globin gene, can cause beta thalassemia (Feng et al., 1994). From the 32896 8mers that were counted from the EMSA-seq experiment, we found enriched 193 8-mers (binomial test p-value $<10^{-6}$) (0.58% of total). In Figure 42A (groups A, B, D), we see that the protein prefers the C[AT]CCC motif. We also used the crystal structure of KLF4 in contact with the **GCCACGCCT** DNA sequence in order to speculate about potential contacts between the protein and the DNA sequence (Figure 42B). In Figure 42C we see that the amino acids responsible for the DNA contacts and their relative distance between them are conserved in all the three KLF1 Zn fingers. Thus we hypothesized that the C[AT]CC motif is the structural footprint of the KLF1 Zn finger 1 and 2. To have a more focused picture for the preferred motif we used the CACC as a core motif in order to find flank restrictions on the -1, +1 positions (Figure 43). We see that the expanded motif is [AC]CACC[CT].

A.



B.

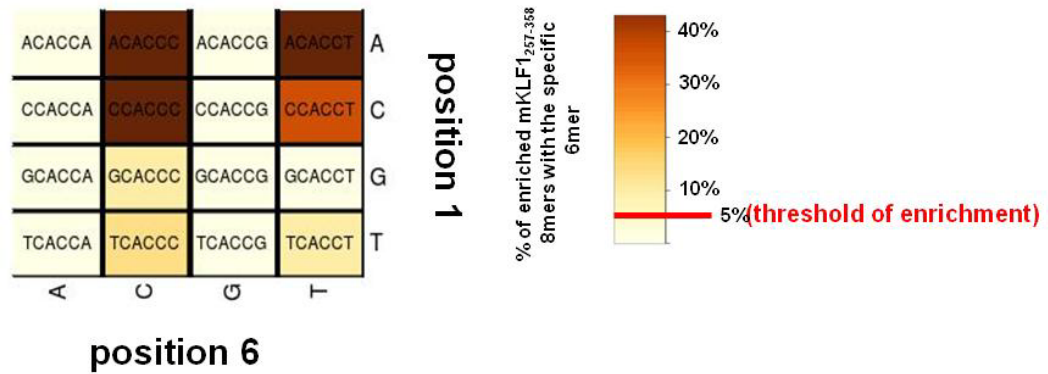


sp Q60793 KLF4_MOUSE	MRQPPGESDMVSDALLPSFSTTFASGPAGREKTLRPAGAPTNRWREELSHMKRLPPLPGR	60
sp P46099 KLF1_MOUSE	-----MASAETVLPSISTLTT---LGQFLDTQEDFLKWWRSEETQDLGPGPPNPT	47
	:::*:**::: : * . : **,* : :	*
sp Q60793 KLF4_MOUSE	PYDLAATVATDLESGGAGAACSNNPALLARRETETEFDLLDLDLFIENSLTHQESVAAT	120
sp P46099 KLF1_MOUSE	GPSLVHVSLSKEDPSGED-----DERDVTCAWDPLFLTN-FPGSESPGTS	91
	. * . : : : ** : . . . * * : : * : . . ** . : :	
sp Q60793 KLF4_MOUSE	VTTASASSSSSPASSGPASAPSTCSFSYPPIRAGDPGVAAANTGGGLLYSRESAPPPTA	180
sp P46099 KLF1_MOUSE	RTCALAPSVGPVAQFEPPESLG-----AYAGGPGLVTGPLGSEEHTSWAHPTPRPP	142
	* : : . * . . . * * : , * , * : . . * . * . . * . .	
sp Q60793 KLF4_MOUSE	PFNLADINDVSPSGGFVAELLRPELDVPYIPQPQQPPGGGLMGKFVLKASLTTPGSEYS	240
sp P46099 KLF1_MOUSE	-----APEPFVAPALAPGLAPKAQPSYSDSRAGS--VGGFPPAGLAVP----	184
	. . *** * * * * * . . . * . : * . : * , * : *	
sp Q60793 KLF4_MOUSE	SPSVISVSKGSPDGSHPVVVVAPYSGGPPRMCPKIKQEAVPSCTVSRSLAEHLASAGPQLSN	300
sp P46099 KLF1_MOUSE	-----AAPGAPYG-----LLSGYPALYP-----APQYQG	208
	. : * : * * . * * : * . ** . .	
sp Q60793 KLF4_MOUSE	GHRPNTHDFPLGRQLPTRTPTPLSPEELLNSRDCHPGLPLPPGFHPHPGNYPFPFLPDQM	360
sp P46099 KLF1_MOUSE	-----HFQLFRGLAAPSAGGTAPPSFLN-----CLGPG-----	236
	. * * * * . : : : * . : **	
sp Q60793 KLF4_MOUSE	QSQVPSLVHYQELMPGPSCLPEEKPCKRGRRSWPRKRKTATHCTDYACCGKTYTKSSH LKAH	420
sp P46099 KLF1_MOUSE	-TVATELGATAIAGDAGLSPGTAPPKRSRRTLAPKRQAHAHTCGHEGCGKSYSKSSH LKAH	295
	: . . . * : . . * . * . * . * . * . * . * . * . * . * . * . * . * . * . * .	
sp Q60793 KLF4_MOUSE	LRTHTGEPKYHCDWDGCGMWFARSDELTRHYRKHTGHRPFQCQKCDRAFSRSDHLALHMK	480
sp P46099 KLF1_MOUSE	LRTHTGEPKYACSWDCDWRFARSDELTRHYRKHTGHRPFCCGLCPRAFSSRSDHLALHMK	355
	***** * . * . * . * . * . * . * . * . * . * . * . * . * . * . * . * . * .	
sp Q60793 KLF4_MOUSE	RHF 483	
sp P46099 KLF1_MOUSE	RHL 358	
	**:	

171

contact with the DNA bases. The rest coloured amino acids are making contacts with the phosphate backbone.

A.



B.

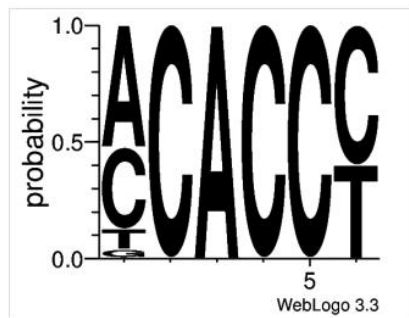


Figure 43. A) Frequency of occurrence of the 16 different 6mers from the DNA motif “[ACGT]CACC[ACGT]” in the 193 statistically enriched mKLF1₂₅₇₋₃₅₈ 8mers. The heat map represents the enrichment for different nucleotides on positions 1 and 6 of the DNA motif “[ACGT]CACC[ACGT]”. The occurrence of each 6mer is measured as percentage of 8mers that show the specific 6mer. Similarly with Figure 28, the 5% threshold of enrichment represents, for the numbers of the frequency of occurrence, a z-score value of 3. This z-score value has occurred through permutation analysis. In more detail, 193 8mers were randomly subsampled, without substitution, 1000 times from the 32896 different 8mers. In each sample, the percentage of 8mers that show each one of the 16 different 6mers was calculated. Then the frequency numbers were pooled and a z-score value was assigned to them. B) The logo represents in a more compact way the enrichment seen on the heat map but without taking into account particular dinucleotide correlations of positions 1 and 6.

3.4.7 Binding cooperativity as reflected in the EMSA-seq coefficient (“e”) indicates two different modes of binding for GATA1 and KLF1

As was discussed earlier our mechanistic model of the EMSA-seq reaction can give us an indication for the binding cooperativity of the k-mers through the EMSA-seq coefficient (“e”). In Figure 44 we can see the EMSA-seq coefficient (“e”) distributions for the mGATA1_{NC} and the mKLF1₂₅₇₋₃₅₈ 8-mers that show a better fit with the binding cooperativity model rather than the non-cooperativity one. We found 1904 8mers for GATA1 and 9804 8mers for KLF1, out of 32896 8mers, that show cooperativity. From the GATA1 8mers all of them have EMSA-seq coefficient (“e”) value less than 1. On the contrary, from the KLF1 8mers all of them have EMSA-seq coefficient (“e”) values more than 1. **Considering that we have assumed that the “e” value represents a biological phenomenon this difference in the distribution might indicate different functional properties of the two proteins.** Indeed, in Figure 45 we see that the GATA1 8mers that showed positive cooperativity are enriched in the high binding affinity 8mers. If this behaviour did not have a biological role then we would expect the GATA1 8mers that showed positive cooperativity to be equally distributed between the high and the low binding affinity 8mers.

Here, we have to stress that the Zn fingers of GATA1 and KLF1 belong to different fold groups (Krishna et al., 2003). The GATA1 zinc fingers belong to the “treble clef” fold group whereas the KLF1 zinc fingers belong to the “C₂H₂ like” fold group (Krishna et al., 2003). This difference of the DNA binding domains gives to the whole protein different functional properties. For example the C₂H₂ zinc fingers (Lam et al., 2011), as the ones that appear in KLF1, are usually found in tandem arrays (ZFAs) and each one of these fingers makes contact with three (or more) sequential bases. This modular assembly of ZFAs permits the host protein to recognize a DNA sequence that is the concatenation of each finger’s binding specificity.

To examine in more detail the different functional properties, we grouped the 8mers for each protein in the ones that show and in the ones that do not show cooperativity. In the KLF1 groups, we focused on 4mers that were over-represented in the group that **does not show** cooperativity (Figure 46). In the GATA1 groups, we focused on 4mers that were over-represented in the group that **shows** cooperativity (Figure 47).

In Figure 47, we see that the **“GATC”** motif is more frequent in 8mers that **show** positive cooperativity. **This might reflect the fact that GATA1 binding implicates first the binding of the C-terminal Zn finger on the “GATA” motif and consequently the binding of the N-terminal Zn finger on the “GATC” motif. So the “GATC” motif appears more frequently than expected (negative “e” values) because it is bound every time that the C-terminal Zn finger binds the “GATA” motif.**

In Figure 46, we see that the **“CCC”** motif is more frequent in 8mers that **do not show** negative cooperativity. KLF1 consists of three Zn fingers. Similarly with GATA1 we hypothesized that KLF1 might have a similar binding behavior where the binding of one of the three fingers might be necessary and adequately to stabilize the binding and the other Zn fingers confer extra specificity and increased affinity. The alternative is the simultaneous binding of the 3 Zn fingers where each one contributes equally and independently of the others on the final binding. **The last model does not exclude that one of the Zn fingers drives the specificity of binding but that the binding cannot be stabilized on the absence of the others. Our data favors the last model.** Specifically the fact that 29.8% of the total number of 8mers shows negative cooperativity **indicates that the KLF1 fingers in contrast to GATA1 need to bind simultaneously in order to stabilize the binding.** We conclude this due to the overlapping nature of the 8mers on the 20mer sequence. We hypothesize that only very few 8mers will reflect the simultaneous binding of all the 3 KLF1 Zn fingers. These 8mers will drive the binding

behavior of the whole library. The rest might represent binding of each of the fingers independently and their contrast with the binding behavior of the whole library permits us to notice the less than expected appearance. In Figure 42A we saw that the “**CCC**” is part of the core motif. Because it appears after the [A/T] nucleotides on the motifs, if we compare it with the crystal structure of KLF4 we can conclude that it represents the region where the **second Zn finger** of KLF1 has extensive contacts (5-GCCCA**CGC**CT-3). **Because the “CCC” is overrepresented in the 8mers that do not show negative cooperativity, we propose that the second Zn finger of KLF1 is responsible for the initial DNA binding specificity and its motifs are used as “seed” recognition motifs. We define the “seed” recognition motifs as “motifs” that are necessary to initiate the binding but they are not adequate to maintain the binding. For this reason, the first and the third finger are used to stabilize the binding.**

It is interesting to point out the following. KLF1 has a role in switching the expression from fetal to adult hemoglobin (Nuez et al., 1995, Perkins et al., 1995). Mutations of human KLF1 that affect the DNA binding domains of the protein have been associated with the presence of high levels of fetal hemoglobin (HbF) in adults (Borg et al., 2011). Seven KLF1 missense mutations in people with high HbF appear in the second Zn finger whereas only one missense mutation has been recorded in the first Zn finger and none missense mutations have been recorded in the third Zn finger (<http://www.ithanet.eu/db/ithagenes?action=list&gn=17> and (Borg et al., 2011)).

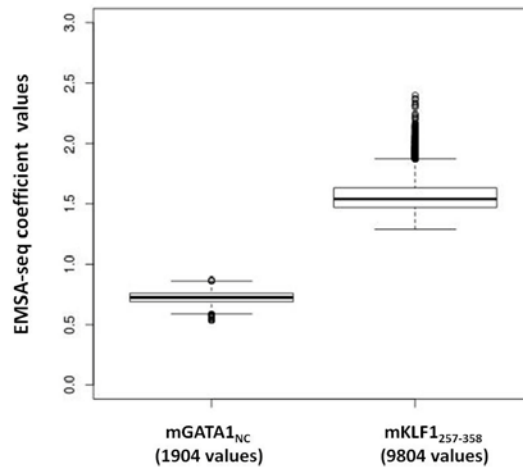


Figure 44. Distribution of the EMSA-seq coefficient values (“e”) for the mGATA1_{NC} and the mKLF1₂₅₇₋₃₅₈ proteins. For the mGATA1_{NC} and the mKLF1₂₅₇₋₃₅₈ 1094 and 9804 values were used respectively.

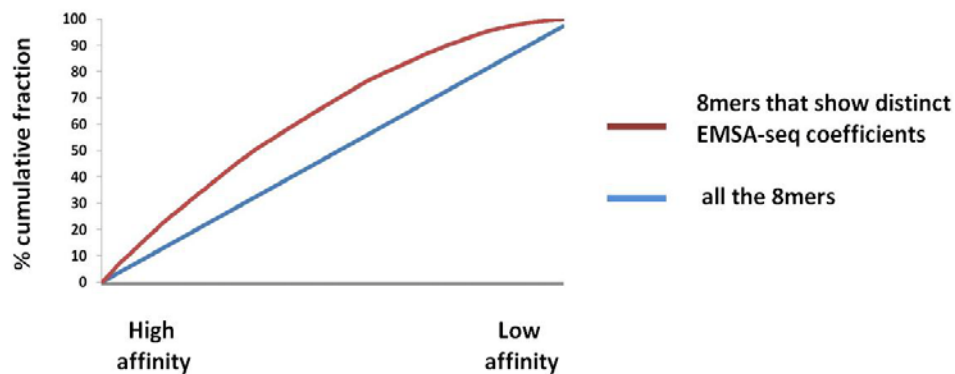
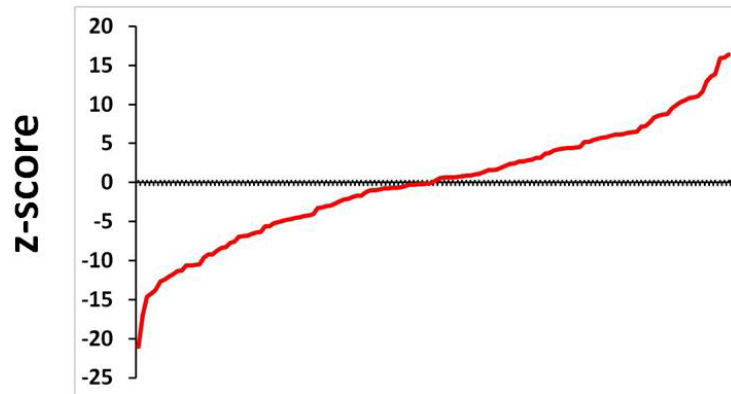


Figure 45. Cumulative distribution of the presence of mGATA1_{NC} 8mers that show distinct EMSA-seq coefficients (red curve). The 8mers are sorted from the most enriched to the lowest enriched. The mGATA1_{NC} 8mers that show distinct EMSA-seq coefficients are enriched with high affinity 8mers. The blue line represents the hypothetical situation where an 8mer group is equally populated with high and low affinity 8mers.

depletion of a 4mer on the group of mKLF1₂₅₇₋₃₅₈ 8mers that show distinct EMSA-seq coefficients
(negative z-score represents lowest enrichment)



15 4mers with
the lowest z-score

4mer	z-score
G <u>CCC</u>	-21.07
GATA	-17.00
ATAA	-14.65
TACA	-14.23
ATTA	-13.79
<u>CCC</u> A	-12.68
AGTA	-12.43
AATA	-12.02
GTTA	-11.69
ATAC	-11.35
AACA	-11.22
ATAT	-10.64
ATCA	-10.60
TATA	-10.57
<u>CCCC</u>	-10.50

4mers that are
part of the KLF1 motif

4mer	z-score
<u>CACC</u>	-0.74
<u>CTCC</u>	1.33

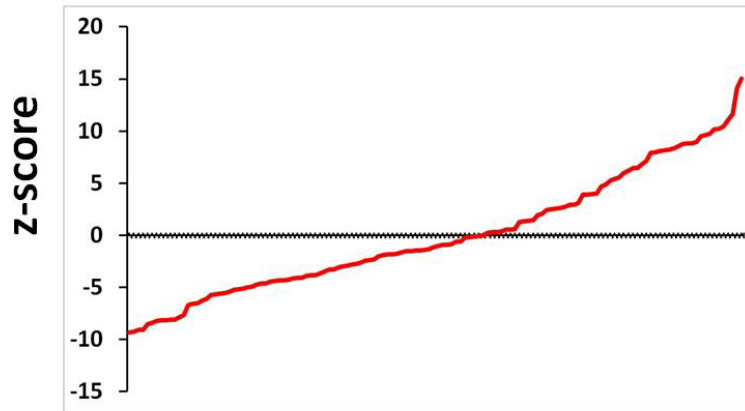
16 4mers with
the highest z-score

4mer	z-score
CTCA	8.73
CAAC	8.78
AGAG	9.51
AAAA	9.86
AGCA	10.27
CAAA	10.49
ACCA	10.84
AAAG	10.93
GCAA	11.09
CTTC	11.59
ATGG	12.96
GAAA	13.56
AAGC	13.87
CCAA	15.97
GGAA	15.99
TCCA	16.39

Figure 46. Z-score distribution of the differential presence of 4mers in mKLF1₂₅₇₋₃₅₈ 8mers that have distinct EMSA-seq coefficients and in mKLF1₂₅₇₋₃₅₈ 8mers that do not have. The first 15 4mers with the highest and lowest z-score values are presented. Positive z-score means that the 4mer is enriched in mKLF1₂₅₇₋₃₅₈ 8mers that have distinct EMSA-seq coefficients. Negative z-score means that the 4mer is enriched in mKLF1₂₅₇₋₃₅₈ 8mers that do not have distinct EMSA-seq coefficients. The z-score value has been calculated through permutation analysis. In more detail, 9804 8mers were randomly subsampled, without substitution, 1000 times from the 32896 different 8mers. In each sample, the

number of times each one of the 1024 different 4mers appears in the 8mers, was calculated. For each 4mer, the mean value and the standard deviation of the values of the number of times the 4mer appears in each one of the 1000 samples, was used to assign a z-score in the 8mer groups that either have or have not distinct EMSA-seq coefficients.

enrichment of a 4mer on the group of mGATA1_{NC} 8mers that show distinct EMSA-seq coefficients
(positive z-score represents highest enrichment)



<u>15 4mers with the lowest z-score</u>		<u>4mers that are part of the GATA1 motif</u>		<u>15 4mers with the highest z-score</u>	
4mer	z-score	4mer	z-score	4mer	z-score
AGCC	-9.33	<u>CATC(GATG)</u>	-0.08	AAAG	8.32
<u>GATA</u>	-9.28			AATG	8.55
<u>AATC(GATT)</u>	-9.08			CGTC	8.73
GGCA	-9.06			AGTG	8.83
AGGC	-8.56			CTCA	8.85
ATAG	-8.40			CACA	8.97
GTTA	-8.24			GAAC	9.51
CGCC	-8.21			AAGA	9.60
AATA	-8.16			GGAA	9.76
GACA	-8.13			TGAA	10.15
GCCA	-8.09			ACTC	10.19
CGGC	-7.85			CGAG	10.44
GCCC	-7.68			ACGA	11.05
GGTA	-6.74			GAGA	11.63
ATAA	-6.57			AGAA	14.13
				<u>GATC</u>	15.02

Figure 47. Z-score distribution of the differential presence of 4mers in mGATA1_{NC} 8mers that have distinct EMSA-seq coefficients and in mGATA1_{NC} 8mers that do not have. The first 15 4mers with the highest and lowest z-score values are presented. Positive z-score means that the 4mer is enriched in mGATA1_{NC} 8mers that have distinct EMSA-seq coefficients. Negative z-score means that the 4mer is enriched in mGATA1_{NC} 8mers that do not have distinct EMSA-seq coefficients. The z-score value has been calculated through permutation analysis. In more detail, 1094 8mers were randomly subsampled, without substitution, 1000 times from the 32896 different 8mers. In each sample, the number of times each one of the 1024 different 4mers appears in the 8mers, was

calculated. For each 4mer, the mean value and the standard deviation of the values of the number of times the 4mer appears in each one of the 1000 samples, was used to assign a z-score in the 8mer groups that either have or have not distinct EMSA-seq coefficients.

3.4.8 Effect of FOG1 and KLF1 complexation on the DNA binding specificities of GATA1

In order to find how FOG1 and KLF1 change the binding specificities of GATA1 we performed EMSA-seq of the wt-mGATA1/wt-mFOG1 complex, the mGATA1_{NC}/mFOG1_{finger1} complex, the mGATA1_{NC}/mFOG1_{finger56} complex, the mGATA1_{NC}/mFOG1_{finger9} complex and of the mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex.

3.4.8.1 Complexation of GATA1 with FOG1 increases the preference for GATAA and GATC containing motifs

As we have seen the wt-mGATA1 shows reduced binding specificities in vitro compared to the mGATA1_{NC}. For this reason we will compare the binding specificities of wt-mGATA1/wt-mFOG1 against the mGATA1_{NC} ones. In Figure 48A we see that FOG1 increases the preference for GATAA in the top 200 8mers of the wt-mGATA1/wt-mFOG1 (from 63 times in mGATA1_{NC} alone (31.5% of total 8mers) to 117 in wt-mGATA1/wt-mFOG1 complex (58.5% of total 8mers)). We also see an increase of GATC containing 8mers from 2 in mGATA1_{NC} alone (1%) to 19 8mers in the complex (5%). The "GATA.G" motif also shows an increase from 4 8mers in mGATA1_{NC} alone (2%) to 30 8mers in the complex (15%). On the contrary the AGATA and the GATT show reduced occurrence. The AGATA is reduced from 81 8mers in mGATA1_{NC} alone (40.5%) to 57 8mers in the complex (28.5%). The GATT is reduced from 43 8mers in mGATA1_{NC} alone (21.5%) to 22 8mers in the complex (11%). We also checked the distribution of motifs throughout all the enriched 8mers with cumulative plots (Figure 48C). We see that the GATAA, GATC and GATA.G are indeed appearing more frequently in the highest enriched 8mers of the complex than the mGATA1_{NC} alone (Figure 48B,C).

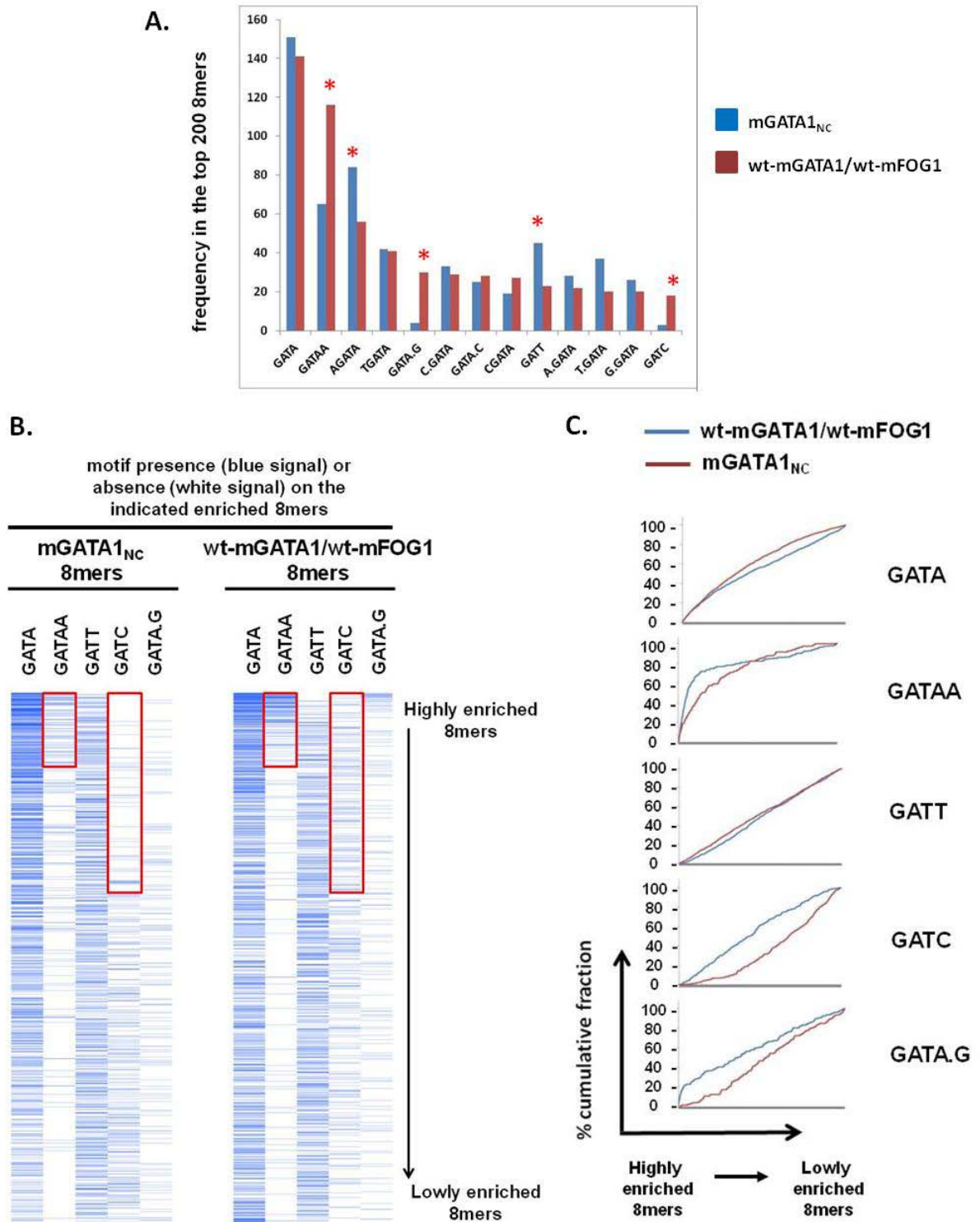


Figure 48. Binding specificities of wt-mGATA1/wt-mFOG1. A) Frequency of occurrence of the indicated motifs in the top 200 8mers for both the mGATA1_{NC} protein and the wt-mGATA1/wt-mFOG1 complex. B) "GATA", "GATAA", "GATT", "GATC", "GATA.G" motif appearance in the top 2122 enriched 8mers for both the mGATA1_{NC} and wt-mGATA1/wt-mFOG1. The 8mers are sorted by fold change from the highest to the lowest. C)

Cumulative fraction of "GATA", "GATAA", "GATT", "GATC", "GATA.G" motif appearance in the top 2122 enriched 8mers. The 8mers are sorted by fold change from the highest to the lowest.

3.4.8.2 Complexation with FOG1 increases the preference for the $\text{GATC}\{N\}_{1-2}\text{GATA}$ motifs

At first, we checked the distance distribution between the “**GATA**”/”**GATT**” and “**GATC**” motifs. In Figure 49 we see that the distance distribution is affected in a “direction” and a “motif” dependent manner. Specifically, if we compare the distance distribution of the “**GATA**”-”**GATC**” co-occurring motifs between the wt-mGATA1 alone and the wt-mGATA1/wt-mFOG1 complex we see an increased preference for the $\text{GATC}\{N\}_{1-2}\text{GATA}$ motifs. We can also see that for the “**GATT**”-”**GATC**” co-occurring motifs, the $\text{GATC}\{N\}_{1-2}\text{GATT}$ motifs preference is not very prominent. This might explain the reduction in the occurrence frequency of the “**GATT**” motifs at least in the top 200 8mers of the complex. We note here that the distance distribution of the “**GAT(A/T)**”-”**GATC**” co-occurring motifs of the mGATA1_{NC} (Figure 34) covers all the potential distance distributions that the protein can have for these specific motifs. Under this perspective, assuming that FOG1 forces the wt-mGATA1 to sample a part of its binding landscape, we can argue that there is no preference for the $\text{GATA}\{N\}_{1-2}\text{GATC}$ motifs (Figure 55). This behavior gives directionality in the binding with an unknown physiological role.

In Figure 49, we also notice, for the wt-mGATA1/wt-mFOG1 complex, an increase occurrence of the $\text{GATC}\{N\}_{5-10}\text{GATA}$ motifs (either by using the enriched mGATA1_{NC} 8mers or the enriched wt-mGATA1/wt-mFOG1 8mers).

In Figure 50 and Figure 51 we present the GATA-GATC and GATT-GATC distance distributions in the GATA1 ChIP-seq peaks called in the experiments of (Chlon et al., 2012). Chlon et al (Chlon et al., 2012) grouped the GATA1ChIP-seq peaks in three categories based on whether FOG1 was present or absent in the GATA1 binding sites. The three categories are:

- i) GATA1 specific peaks (these represent positions where the complex wt-mGATA1/wt-mFOG1 binds alone),
- ii) GATA1_{V205G} specific peaks (these represent positions where wt-mGATA1 binds alone)
- iii) The rest of the GATA1 ChIP-seq peaks (these represent positions where wt-mGATA1 binds alone or a mixture of wt-mGATA1 alone and of wt-mGATA1/wt-mFOG1 complex bind).

We used enriched 8mers from both the mGATA1_{NC} alone (Figure 50) and the wt-mGATA1/wt-mFOG1 complex (Figure 51).

We see that the **GATA1 ChIP-seq peaks overall** show a preference mainly for **GATC{N}₉GATA**, **GATC{N}₉GATT** and secondary for **GATC{N}₁GATA**, **GATC{N}₁GATT** and **GATT{N}₂GATC**. The **wt-mGATA1/wt-mFOG1 specific ChIP-seq peaks** show a preference for **GATC{N}₁GATA** (the fluctuation of the frequency of occurrence for the rest of the positions is of lower height). The **wt-mGATA1 specific ChIP-seq peaks** do not show this preference (the fluctuation of the frequency of occurrence for the rest of the positions is of equal height). Both the **wt-mGATA1/wt-mFOG1 specific ChIP-seq peaks** and the **wt-mGATA1 specific ChIP-seq peaks** do not show preference for a specific “GATT-GATC” distance. This is in accordance with our in vitro data for both the **wt-mGATA1 protein alone (Figure 49)** and the **wt-mGATA1/wt-mFOG1 complex (Figure 49)**. The only difference is the length of the spacer between the “GATC” and the “GATA” motif for the **wt-mGATA1/wt-mFOG1** complex. Where in vitro it is one or two nucleotides, in vivo seems to be one nucleotide. This might be attributed to other protein-protein interactions that might restrict the protein to accommodate one nucleotide between the motifs. Also, the fact that the **GATA1 ChIP-seq peaks overall** show preference for extra “GATA-GATC” and “GATT-GATC” distances indicates that other GATA1 co-factors change the GATA1 DNA binding specificities.

In Figure 52 we can see the GATA-GATG and GATT-GATG distance distributions for both the wt-mGATA1 alone and the wt-mGATA1/wt-mFOG1 complex. As far as it concerns the GATA-GATG distance distribution, the wt-mGATA1/wt-mFOG1 complex shows a similar preference like the wt-mGATA1 alone for the **GATGATA** motif (this preference could be seen only with the enriched 8mers from the wt-mGATA1/wt-mFOG1 complex but not with the enriched 8mers from the mGATA1_{NC}). Additionally we see a small preference for the TATC{N}₁**GATG** motif and a strong preference for the **GATG**{N}₁TATC motif. The **GATG**{N}₁TATC motif preference could be seen with both the enriched 8mers from the wt-mGATA1/wt-mFOG1 complex and the enriched 8mers from the mGATA1_{NC}. The wt-mGATA1/wt-mFOG1 complex showed no preference for the GATT-GATG distance.

In Figure 53 we can see the GATA-GATT distance distributions for the wt-mGATA1/wt-mFOG1 complex. Similarly with the wt-mGATA1 alone, the wt-mGATA1/wt-mFOG1 complex shows a preference for **GATTATC**. The wt-mGATA1/wt-mFOG1 complex, also, shows the extra preference for TATC{N}_{1,2}**GATT** and for **GATT**{N}_{1,2}TATC.

In Figure 54 we can see the GATA-GATA distance distributions for the wt-mGATA1/wt-mFOG1 complex. Compared with the wt-mGATA1 alone, we see an increased preference for the **GATATC** and TATC{N}_{1,2}**GATA** motifs.

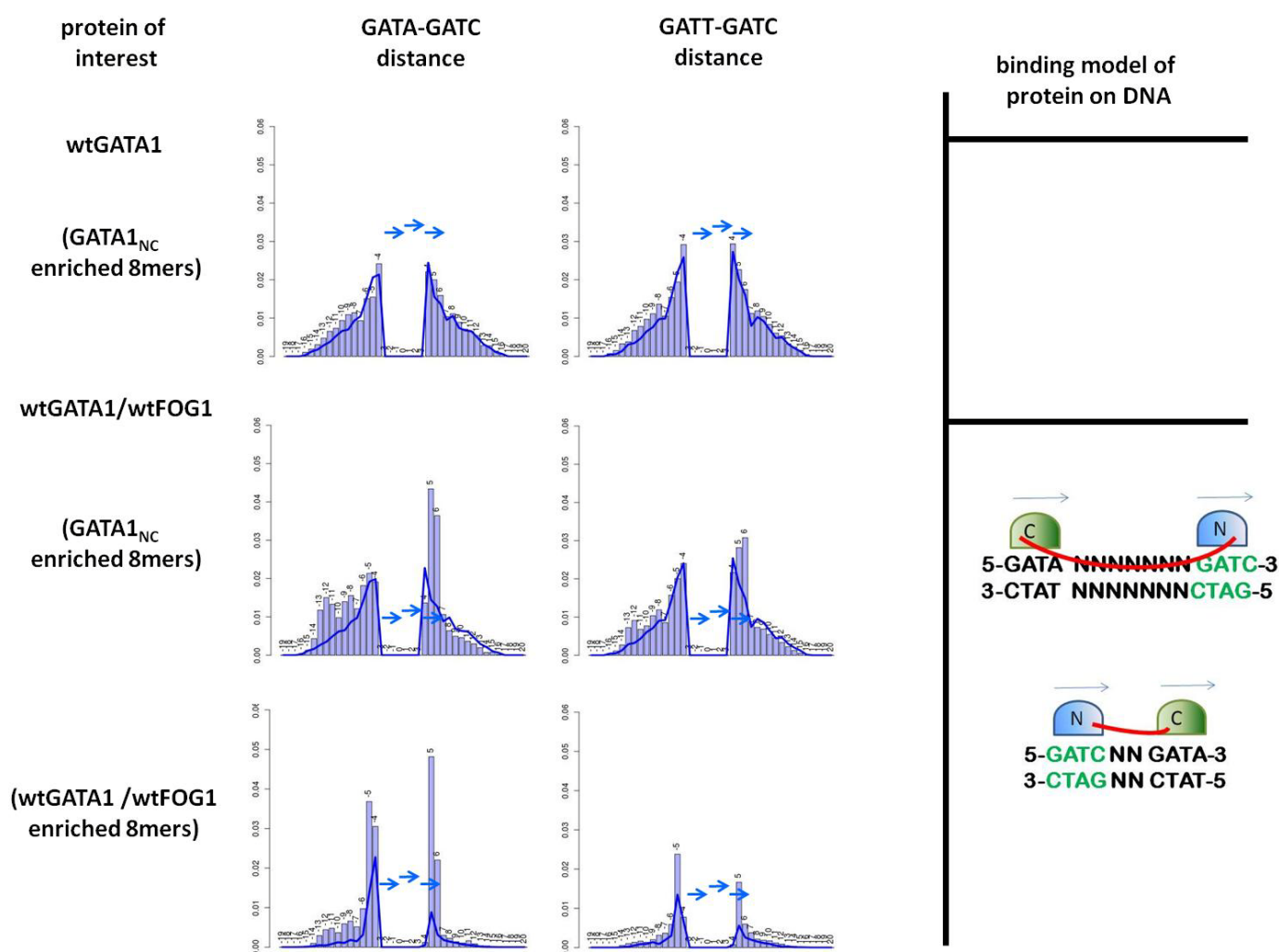


Figure 49. Distance graphs between the “GATA” and the “GATC” motifs or between the “GATT” and the “GATC” motifs for the wt-mGATA1 protein and the wt-mGATA1/wt-mFOG1 complex. The distance graphs for the wt-mGATA1 were created with the 2122 enriched 8mers of the monomeric mGATA1_{NC} (first row). The enriched 8mers from the wt-mGATA1 were not used due to the lack of an adequate number of enriched 8mers containing a “GATC” motif. The distance graphs for the wt-mGATA1/wt-mFOG1 were created with the 2122 enriched 8mers of the monomeric mGATA1_{NC} (second row) or the top 2122 enriched 8mers from the wt-mGATA1/wt-mFOG1 complex (third row). The binding models of the protein on DNA are also shown. For further explanation about the

graph see Figure 33. The blue arrows represent the orientation of the motifs as appear in Figure 33A.

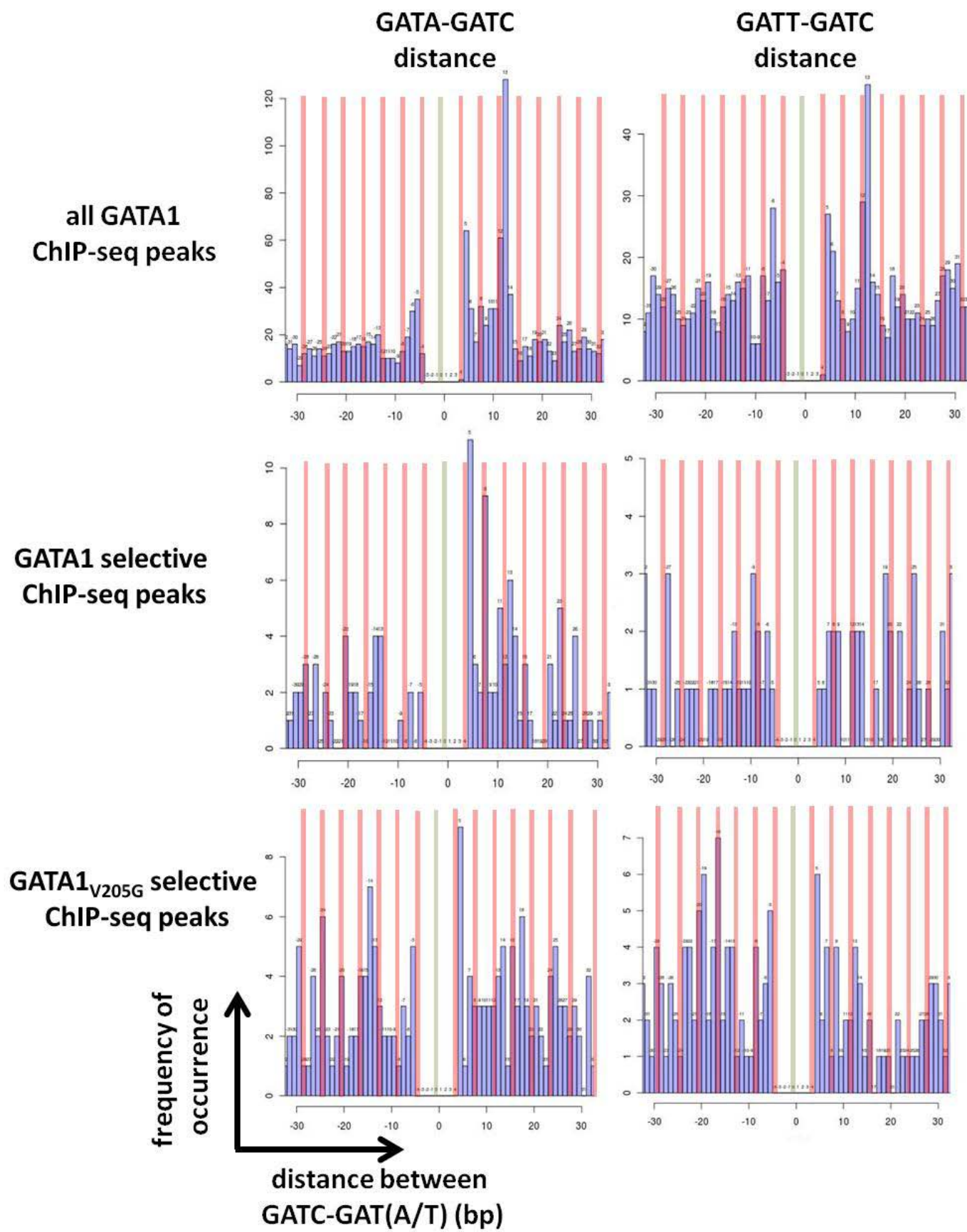


Figure 50 (previous page). Distance graphs (blue bars) between the “GATA” and the “GATC” motifs or between the “GATT” and the “GATC” motifs for the all the GATA1 ChIP-seq peaks, the GATA1 selective ChIP-seq peaks (fraction of peaks where only wt-mGATA1/wt-mFOG1 is bound), the GATA1_{V205G} selective ChIP-seq peaks (fraction of peaks where only wt-mGATA1 is bound). The distance graphs were created with the 2122 enriched 8mers from the monomeric mGATA1_{NC}. Distances were calculated similarly with Figure 34. The green bar indicates position 0. The red bars indicate positions on the x-axis every four nucleotides (for example the first red bars right and left of the green bar in the centre indicate positions +4 and -4 respectively).

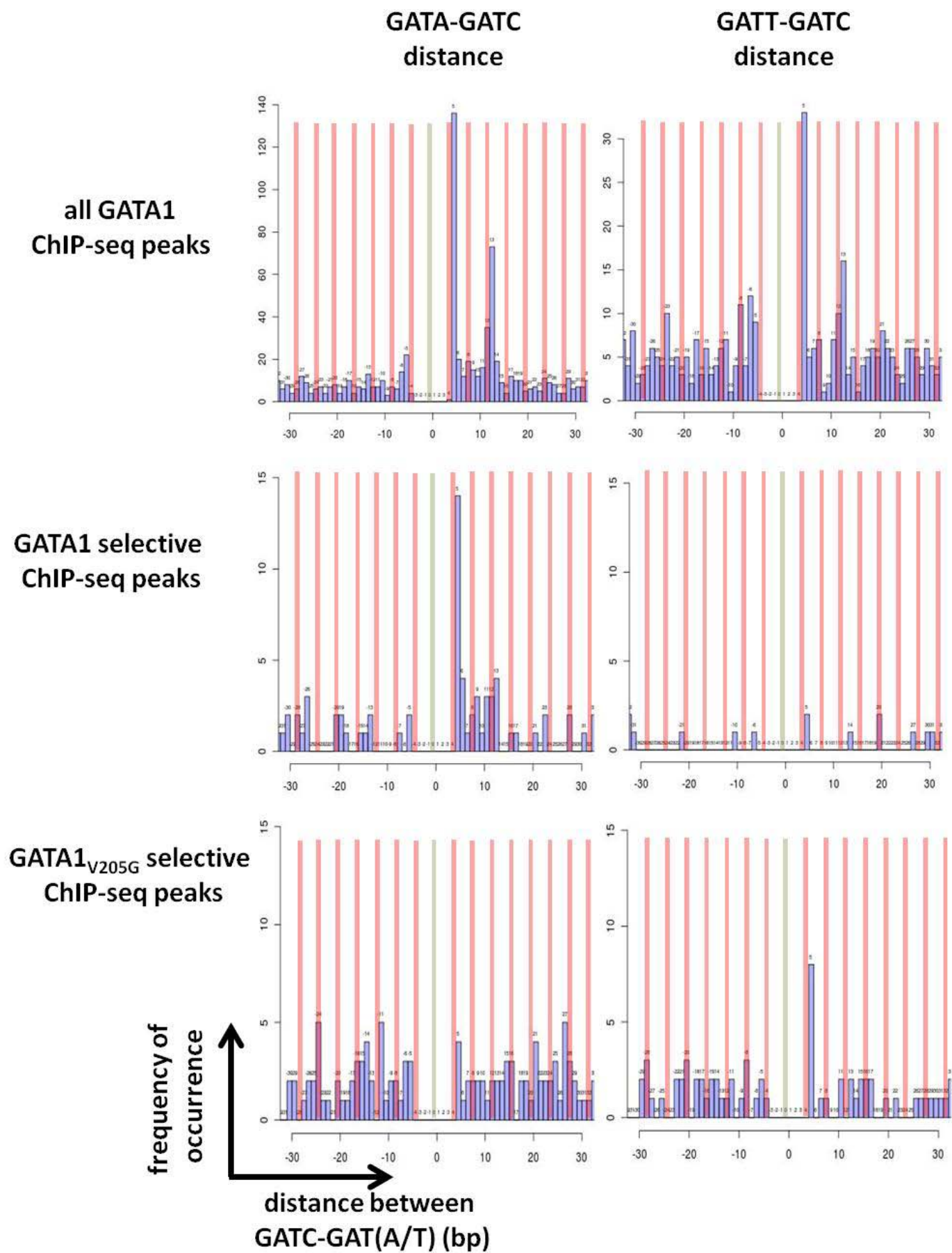


Figure 51 (previous page). Distance graphs between the “GATA” and the “GATC” motifs or between the “GATT” and the “GATC” motifs for the all the GATA1 ChIP-seq peaks, the GATA1 selective ChIP-seq peaks (fraction of peaks where only wt-mGATA1/wt-mFOG1 is bound), the GATA1_{V205G} selective ChIP-seq peaks (fraction of peaks where only wt-mGATA1 is bound). The distance graphs were created with the top 2122 enriched 8mers from the wt-mGATA1/wt-mFOG1 complex. Distances were calculated similarly with Figure 34. The green bar indicates position 0. The red bars indicate positions on the x-axis every four nucleotides (for example the first red bars right and left of the green bar in the centre indicate positions +4 and -4 respectively).

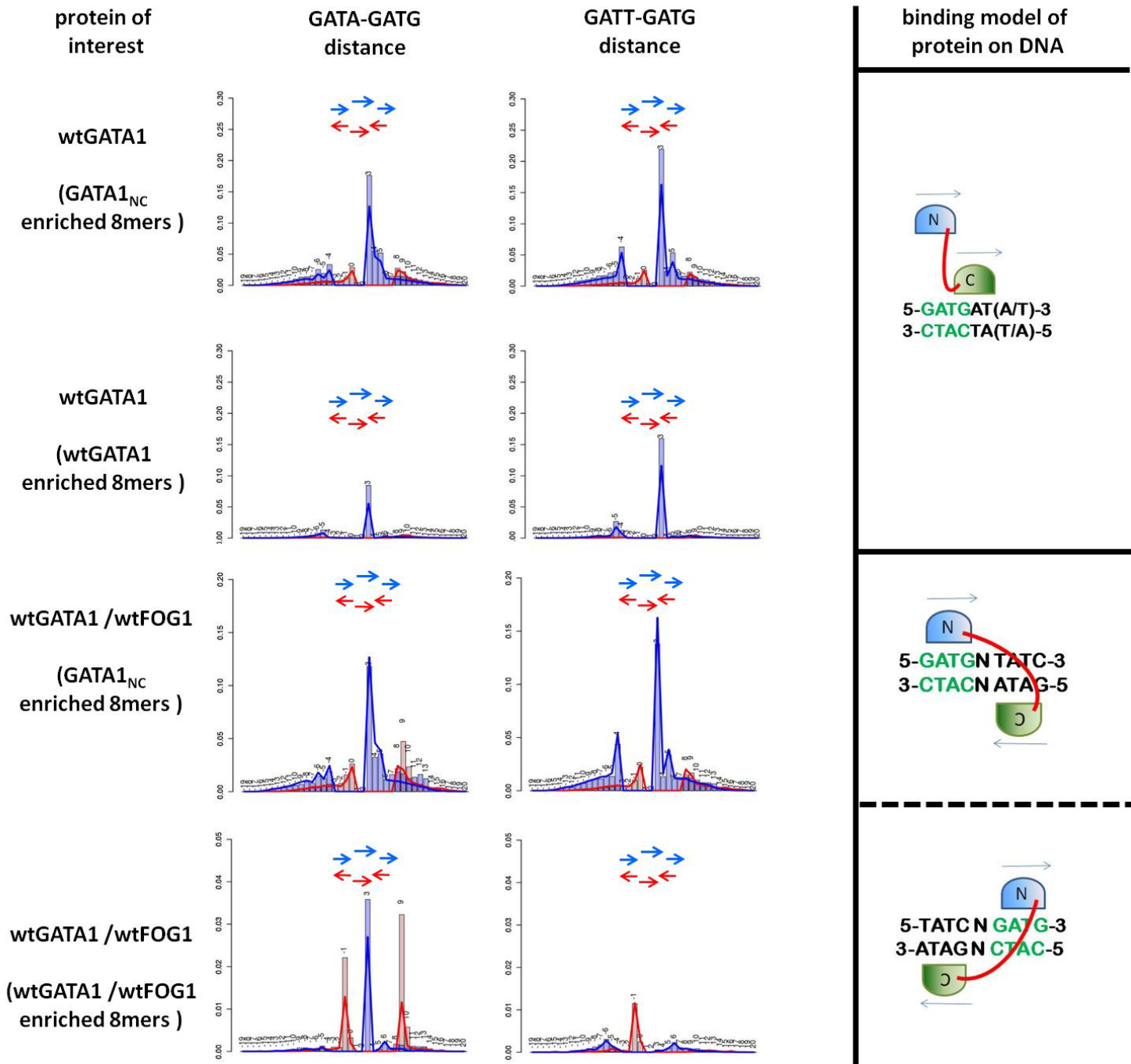


Figure 52. Distance graphs between the “GATA” and the “GATG” motifs or between the “GATT” and the “GATG” motifs for the wt-mGATA1 protein and the wt-mGATA1/wt-mFOG1 complex. The distance graphs for the wt-mGATA1 were created with the 2122 enriched 8mers of the monomeric mGATA1_{NC} (first row) or the top 2122 enriched 8mers of the wt-mGATA1 protein (second row). The distance graphs for the wt-mGATA1/wt-mFOG1 were created with the 2122 enriched 8mers of the monomeric mGATA1_{NC} (third row) or the top 2122 enriched 8mers of the wt-mGATA1/wt-mFOG1 complex (fourth row).

The binding models of the protein on DNA are also shown. For further explanation about the graph see Figure 33. The blue and red arrows represent the orientation of the motifs as appear in Figure 33C.

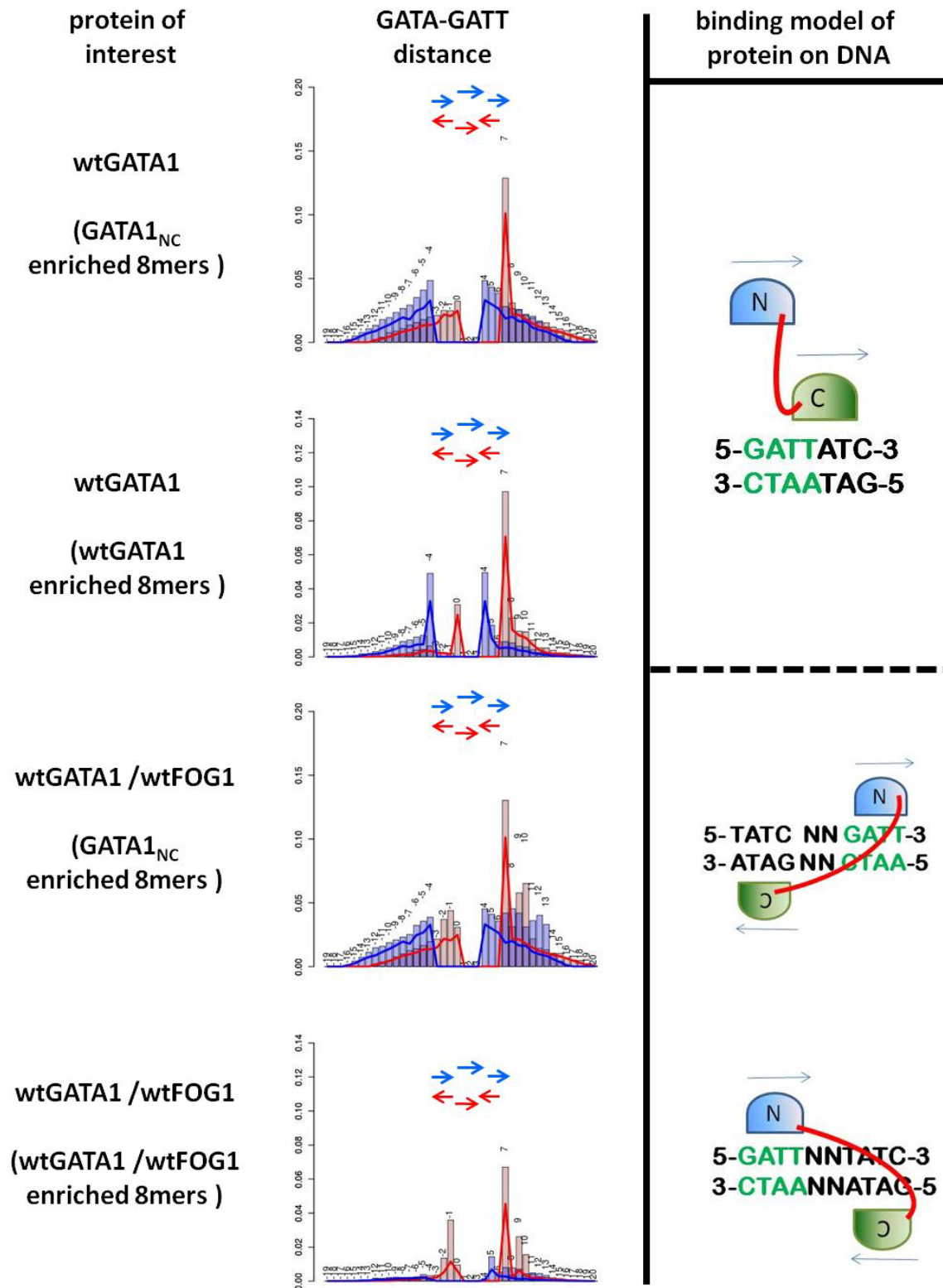


Figure 53. Distance graphs between the “GATA” and the “GATT” motifs for the wt-mGATA1 protein and the wt-mGATA1/wt-mFOG1 complex. The distance graphs for the wt-mGATA1 were created with the 2122 enriched 8mers of the monomeric mGATA1_{NC}

(first row) or the top 2122 enriched 8mers of the wt-mGATA1 protein (second row). The distance graphs for the wt-mGATA1/wt-mFOG1 were created with the 2122 enriched 8mers of the monomeric mGATA1_{NC} (third row) or the top 2122 enriched 8mers of the wt-mGATA1/wt-mFOG1 complex (fourth row). The binding models of the protein on DNA are also shown. **For further explanation about the graph see Figure 33. The blue and red arrows represent the orientation of the motifs as appear in Figure 33C.**

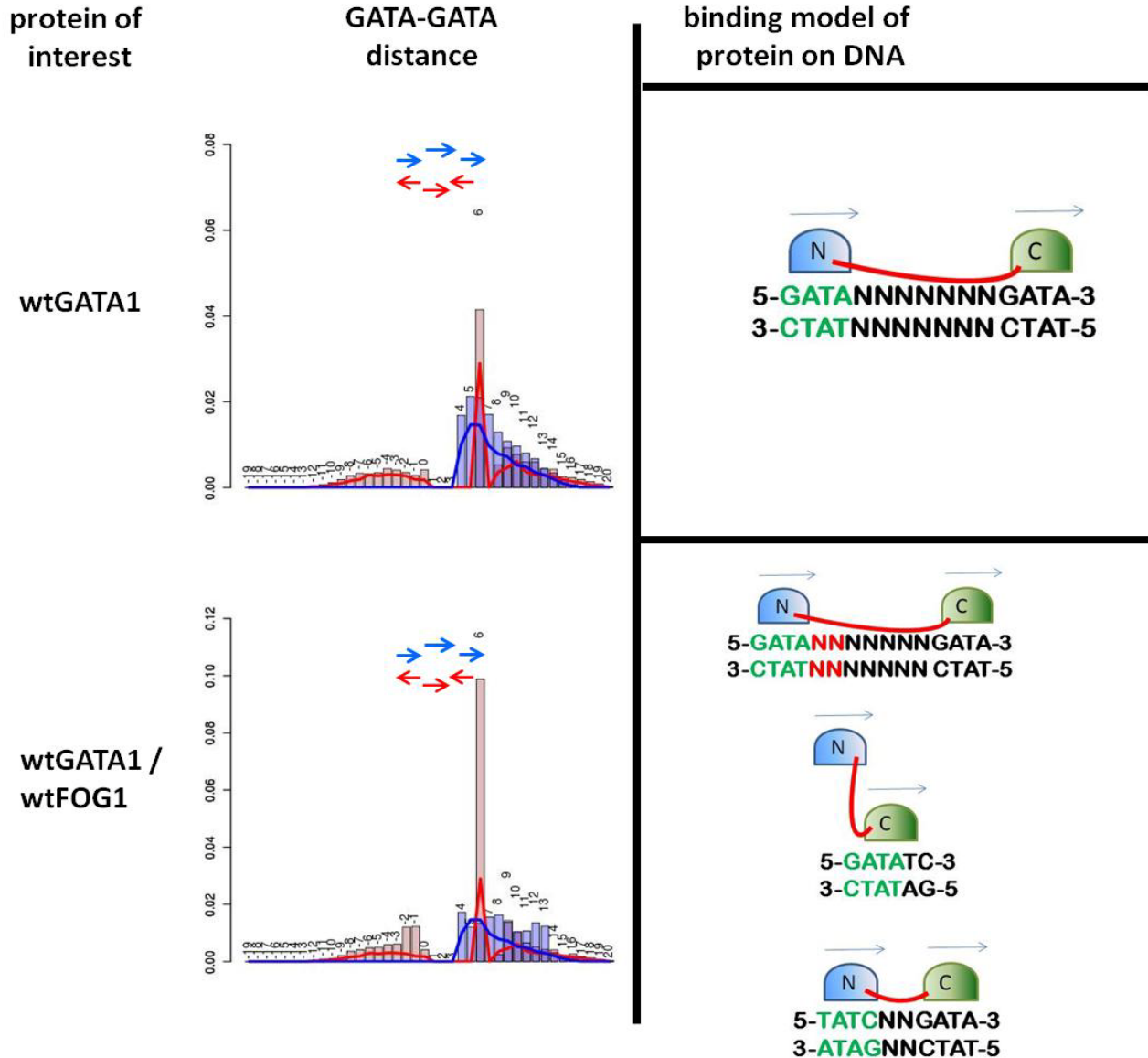


Figure 54. Distance graphs between the “GATA” and the “GATA” motifs for the wt-mGATA1 protein and the wt-mGATA1/wt-mFOG1 complex. The distance graphs for the wt-mGATA1 were created with the top 2122 enriched 8mers of the wt-mGATA1 protein. The distance graphs for the wt-mGATA1/wt-mFOG1 were created with the top 2122 enriched 8mers of the wt-mGATA1/wt-mFOG1 complex. The binding models of the protein on DNA are also shown. The red nucleotides in the binding models of the second row indicate that short distances between the motifs are not favoured. For further explanation about the graph see Figure 33. The blue and red arrows represent the orientation of the motifs as appear in Figure 33C.

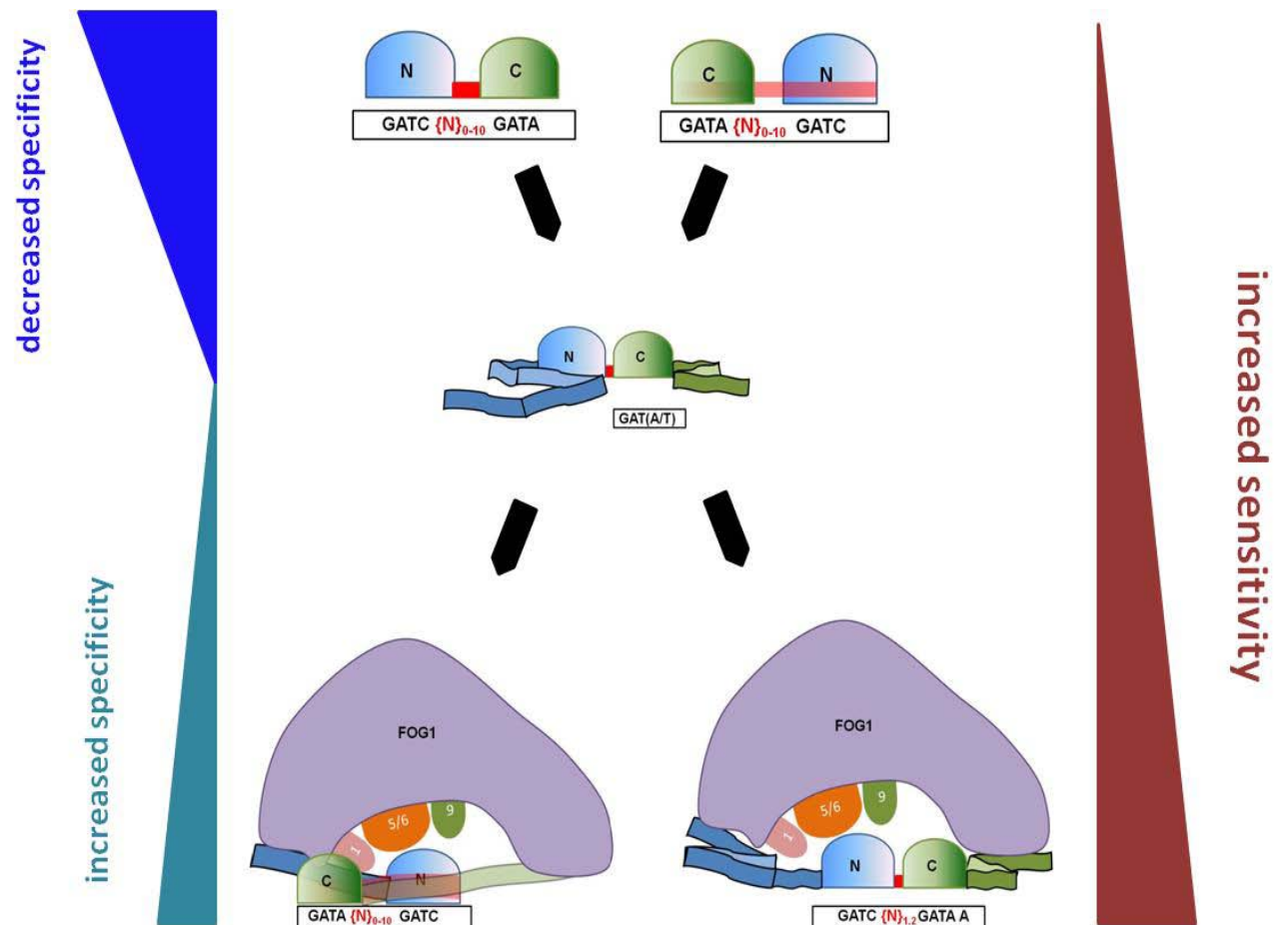


Figure 55. Model of the effect of FOG1 on the DNA binding specificities of GATA1. The two DNA binding domains of GATA1 give to the protein a broad DNA binding specificity. The N-terminal and C-terminal tails of GATA1 restrict the DNA binding specificities of the protein to a very narrow spectrum of the DNA motifs. Interaction with FOG1 restores only a specific fraction of the unrestricted DNA binding landscape. Although overall GATA1 shows reduced binding specificities, this behaviour increases its ability to discriminate the fraction of the GATA1 binding sites where the presence of FOG1 is necessary.

3.4.8.3 DNA binding specificities of mGATA1_{NC}/mFOG1_{finger1}, mGATA1_{NC}/mFOG1_{finger56}, mGATA1_{NC}/mFOG1_{finger9}

To further validate the wt-mGATA1/wt-mFOG1 results we investigated the binding specificities of each FOG1 fingers independently. As discussed earlier the FOG1 fingers 1, 6, 9 are responsible for the interaction with GATA1 (Fox et al., 1999). For this reason we profiled the binding specificities of mGATA1_{NC}/mFOG1_{finger1}, mGATA1_{NC}/mFOG1_{finger56}, mGATA1_{NC}/mFOG1_{finger9}. As has been already discussed these complexes dissociate during the DNA binding reaction (Figure 20). If the binding of the DNA with the mGATA1_{NC} antagonizes the binding of the FOG1 fingers with the mGATA1_{NC} then we might expect the DNA recognition footprint of the complex to exist on the DNA bound library and in its respective k-mers.

In Figure 56 we can see the GATA-GATC distance distribution of the mGATA1_{NC}/mFOG1_{finger1}, mGATA1_{NC}/mFOG1_{finger56} and mGATA1_{NC}/mFOG1_{finger9} complexes. The pattern is similar with the mGATA1_{NC} pattern with the exception of an increased frequency for the GATA{N}₃₋₁₀**GATC** motifs. This preference appears in all of the three complexes. As we have seen, this pattern has also been found in the wt-mGATA1/wt-mFOG1 when either the enriched mGATA1_{NC} 8mers or the wt-mGATA1/wt-mFOG1 8mers were projected on the 20mer bound library (Figure 49). When we projected on the wt-mGATA1/wt-mFOG1 20mer bound library the enriched mGATA1_{NC}/mFOG1_{finger1}, mGATA1_{NC}/mFOG1_{finger56} and mGATA1_{NC}/mFOG1_{finger9} 8mers the GATA{N}₁₋₉**GATC** motif preference was much more prominent (Figure 57).

In Figure 56 we also point out the lack of a strong preference for the **GATC**{N}₁₋₂GATA motifs. This can be attributed to one of the following factors:

- i) The co-existence of all the three FOG1 fingers on the wild type protein forces GATA1 to show preference for the **GATC**{N}₁₋₂GATA motifs.

- ii) As was discussed in the previous chapter, the reduced specificities of wt-mGATA1, which equals to lack of preference for the GATC{N}₁₋₂GATA motifs, permits the protein upon binding with FOG1 to have specific conformational changes that will show this GATC{N}₁₋₂GATA motif specific preference. This means, that the FOG1 fingers cannot force mGATA1_{NC} to show the GATC{N}₁₋₂GATA motif specific preference because mGATA1_{NC} has already the conformation to bind on this motif whereas wt-mGATA1 does not have it.

From our perspective we favor the second interpretation (Figure 55, Figure 60).

We also notice that all the fingers show the same binding pattern and the same DNA binding specificities which implies a redundant role in the binding (Figure 58). Seventy five percent of the top 500 enriched 8mers for the mGATA1_{NC}/mFOG1_{finger1}, the mGATA1_{NC}/mFOG1_{finger9} and the mGATA1_{NC}/mFOG1_{finger56} complex are common between all the complexes (Figure 58A). Nevertheless we caution here, that because approximately sixty eight percent of the top 500 enriched 8mers for the mGATA1_{NC} alone and the mGATA1_{NC}/mFOG1 finger complexes is common between the two conditions (Figure 58B) but the rest, thirty two percent of the top 500 enriched mGATA1_{NC}/mFOG1 finger 8mers, shows a wide range of increased binding affinity (lower binding affinity in the mGATA1_{NC}, the blue bars in lane 2, 3, 4 of Figure 58C are lower than lane 1), further exploration needs to be undertaken to examine whether the non-common group has particular patterns that can discriminate the FOG1 fingers.

In Figure 27 we see that the fold change distribution of the mGATA1_{NC}/mFOG1_{finger1} 8mers and the mGATA1_{NC}/mFOG1_{finger9} 8mers, are higher than the mGATA1_{NC}/mFOG1_{finger56} (lowest protein/DNA ratio condition) and the mGATA1_{NC} ones. In Figure 17 we also notice that the protein concentration of the mGATA1_{NC}/mFOG1_{finger1}

and the mGATA1_{NC}/mFOG1_{finger9} complexes is more than the mGATA1_{NC} alone and the mGATA1_{NC}/mFOG1_{finger56} (lowest protein/DNA ratio condition). Considering that the binding specificities of the mGATA1_{NC}/mFOG1 fingers is overall similar with the mGATA1_{NC} ones (Figure 58B), the previous observation **indicates significant competition between binding of either the DNA or the FOG1 fingers on the mGATA1_{NC} protein.** The fact that the mGATA1_{NC}/mFOG1_{finger56} fold change distribution of 8mers is less than the mGATA1_{NC}/mFOG1_{finger1} and mGATA1_{NC}/mFOG1_{finger9} ones imply that **the binding of the mFOG1_{finger56} on the mGATA1_{NC} has less strength to withstand the DNA competition, than the binding of the mFOG1_{finger1} and mFOG1_{finger9} on the mGATA1_{NC}.** Here we have to remember Figure 16 where we saw that upon purification against the tagged FOG1 fingers from the mGATA1_{NC}/mFOG1_{finger56} and mGATA1_{NC}/mFOG1_{finger9} complexes, the mFOG1_{finger56} and mFOG1_{finger9} appears in equimolar concentration with the mGATA1_{NC}. This implies that the complexation of mGATA1_{NC} with mFOG1_{finger56} or mFOG1_{finger9} is favored. On the contrary in the mGATA1_{NC}/mFOG1_{finger1} lane of Figure 16, mGATA1_{NC} is far less than mFOG1_{finger1} meaning that the complexation is not favored. Based on these in gel protein-protein interaction affinities we were expecting, in EMSA-seq, the mGATA1_{NC}/mFOG1_{finger9} and mGATA1_{NC}/mFOG1_{finger56} complexes to behave similarly and different from the mGATA1_{NC}/mFOG1_{finger1} complex. **Thus, the fact that, in EMSA-seq, only the mGATA1_{NC}/mFOG1_{finger9} and mGATA1_{NC}/mFOG1_{finger1} complexes behave similarly indicates that the competition of the DNA with the FOG1 fingers for the mGATA1_{NC} binding is another feature of the GATA1/FOG1 protein complex.** Although, the FOG1 fingers might force similar binding specificities on the mGATA1_{NC} as far as it concerns the DNA motif, their differences and the resulting fine/critical regulation might derive from the strength of the DNA binding that each one causes (Figure 60).

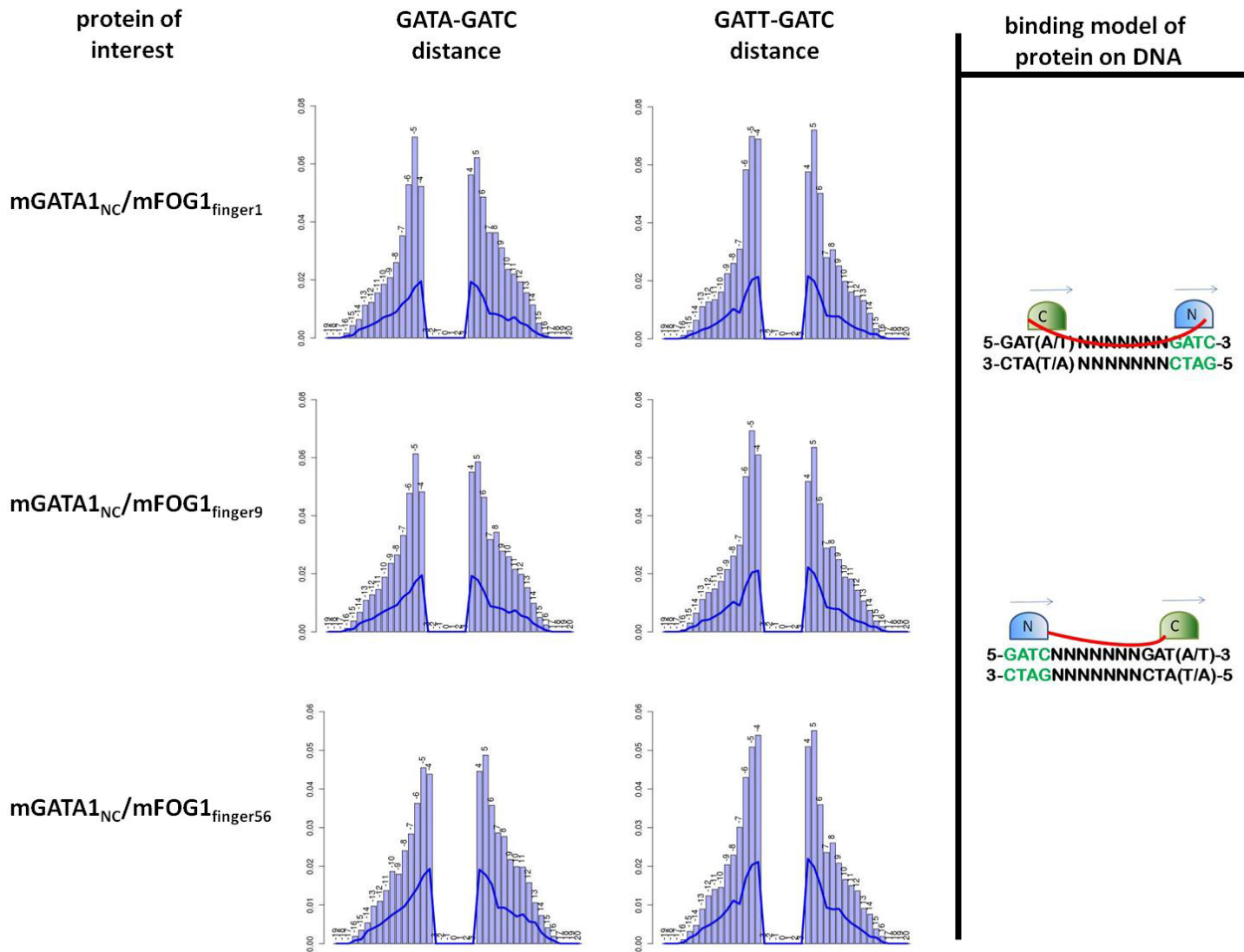


Figure 56. Distance graphs between the “GATA” and the “GATC” motifs or between the “GATT” and the “GATC” motifs for the mGATA1_{NC}/mFOG1_{finger1}, mGATA1_{NC}/mFOG1_{finger9}, mGATA1_{NC}/mFOG1_{finger56} bound sequences. The distance graphs were created with the top 2122 enriched 8mers from the mGATA1_{NC}/mFOG1_{finger1}, mGATA1_{NC}/mFOG1_{finger9}, mGATA1_{NC}/mFOG1_{finger56} complexes respectively. The binding models of protein on DNA are also shown. For further explanation about the graph see Figure 33A.

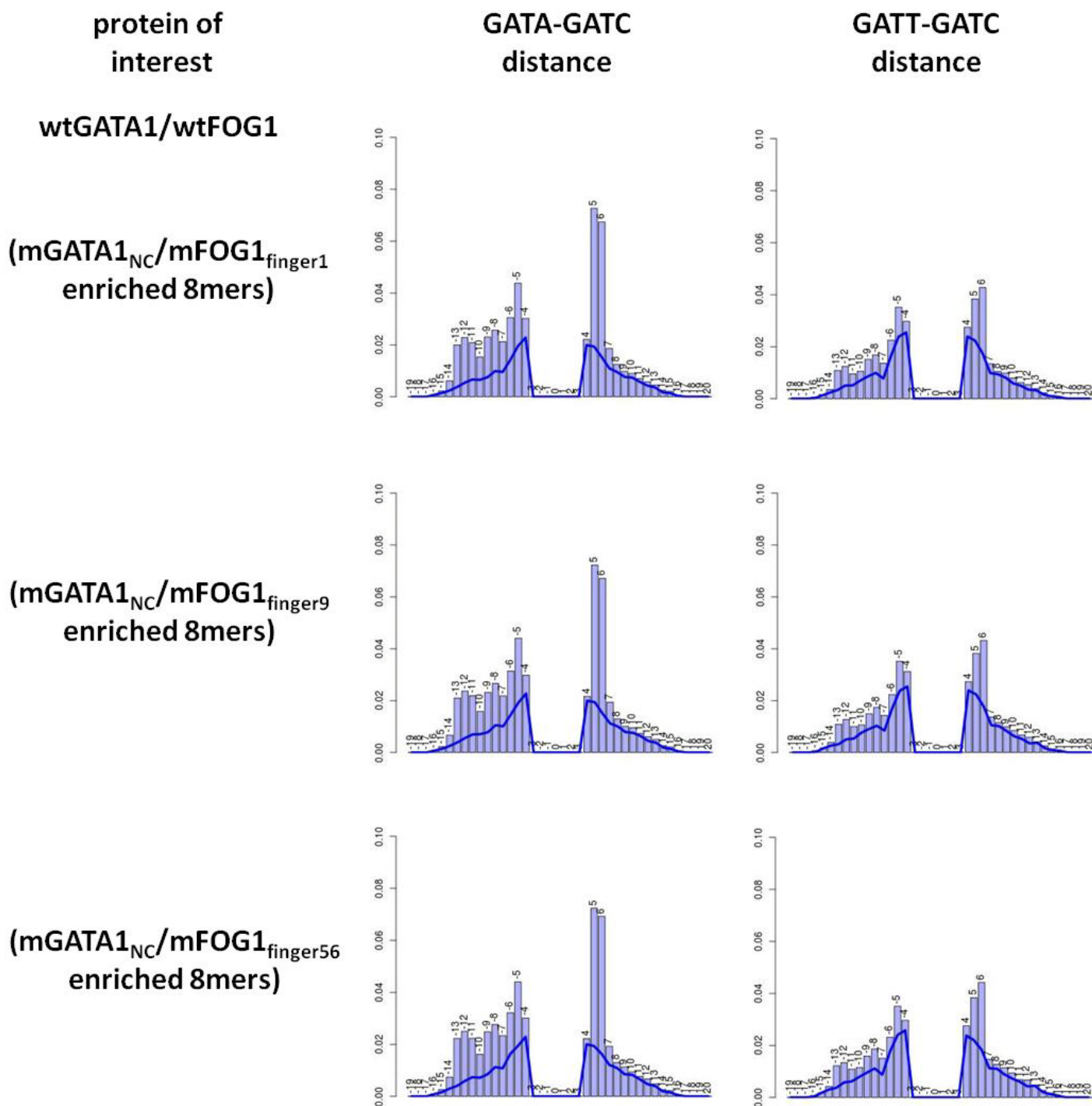


Figure 57. Distance graphs between the “GATA” and the “GATC” motifs or between the “GATT” and the “GATC” motifs for the wt-mGATA1/wt-mFOG1 bound sequences. The distance graphs were created with the enriched 8mers from the mGATA1_{NC}/mFOG1_{finger1}, mGATA1_{NC}/mFOG1_{finger9}, mGATA1_{NC}/mFOG1_{finger56} complexes as indicated. The binding

models of protein on DNA are also shown. For further explanation about the graph see Figure 33A.

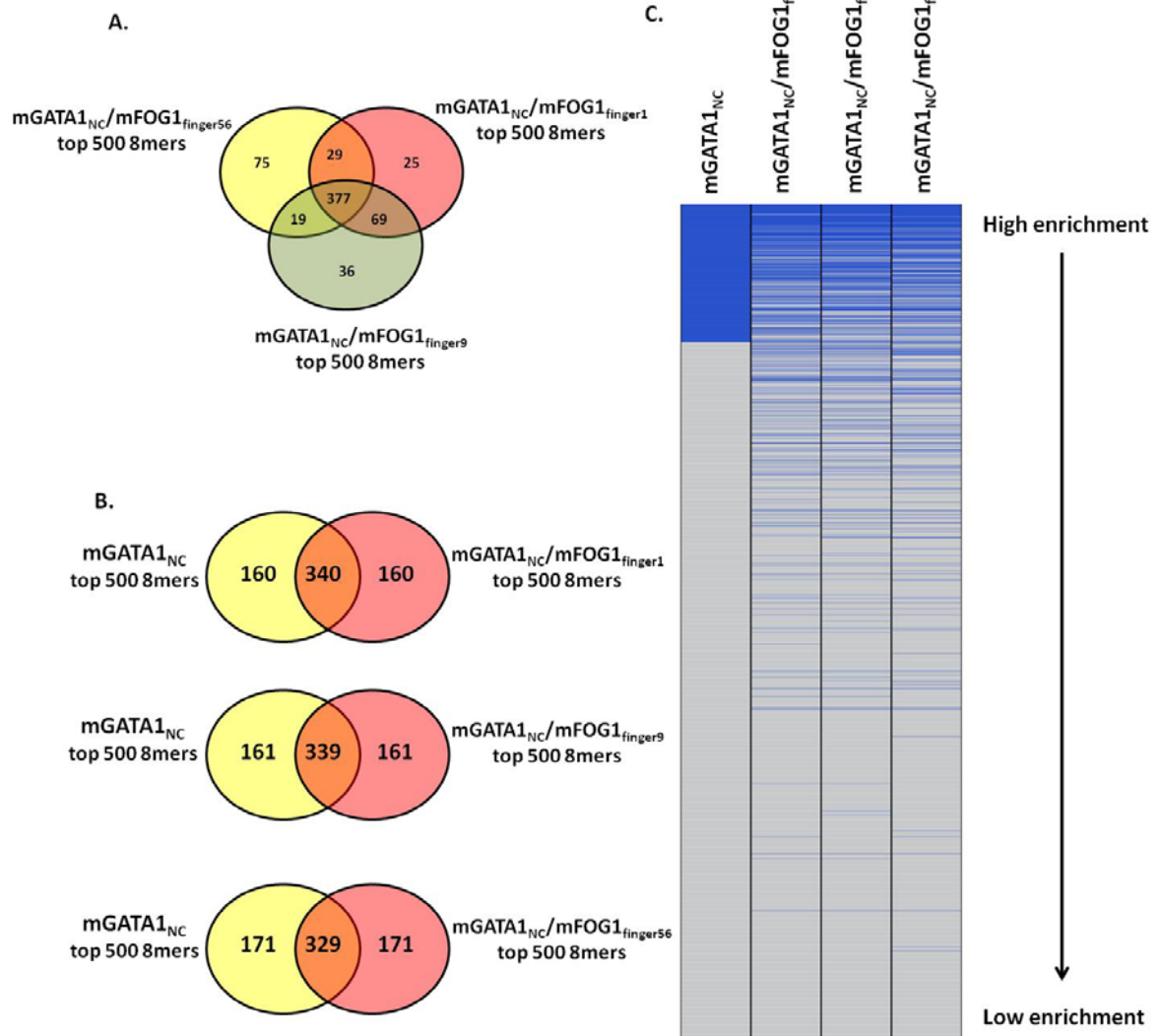


Figure 58. A) The top 500 enriched 8mers from each mGATA1_{NC}/mFOG1 fingers combination were selected and there overlap is presented. B) The top 500 enriched 8mers from each mGATA1_{NC}/mFOG1 fingers combination were overlapped with the top 500 enriched 8mers from mGATA1_{NC}. C) The 2122 enriched 8mers from mGATA1_{NC} were sorted from the highest enriched to the lowest enriched. The top 500 enriched 8mers of mGATA1_{NC} are marked as blue (lane 1). The top 500 enriched 8mers from the mGATA1_{NC}/mFOG1_{finger1} (lane 2), mGATA1_{NC}/mFOG1_{finger9} (lane 3), mGATA1_{NC}/mFOG1_{finger56} (lane 4) complexes were selected and there position in the sorted 2122 enriched 8mers of mGATA1_{NC} was marked with blue.

3.4.8.4 Similarities in the protein sequence of the FOG1 fingers may explain the similar DNA binding specificities

To give insights into the potential functional redundancy of the FOG1 fingers we focused on available crystal structures of FOG1 fingers. For this reason, we used the crystal structure of the *Drosophila* U-shaped (USH) protein that is a FOG1 homologue. The USH protein has been shown that it can interact with both the *Drosophila* homologue of GATA1 (Pannier) (Haenlin et al., 1997) and the mouse GATA1 (Fox et al., 1999).

Three out of the five C₃H₁ fingers of the USH protein have been shown to be able to interact with the N-terminal Zn finger of mouse GATA1 (mGATA1) (Fox et al., 1999). Crystal structure of the finger 1 from the USH protein bound on the N-terminal mGATA1 Zn finger showed that the following residues are important for the binding S218_{dFOG-F1}, S219_{dFOG-F1}, T222_{dFOG-F1}, E224_{dFOG-F1}, A225_{dFOG-F1}, Y229_{dFOG-F1}, Y230_{dFOG-F1} (Liew et al., 2005). As can be seen in Figure 59 the Y229_{dFOG-F1} and the Y230_{dFOG-F1} are conserved in all the mouse FOG1 Zn fingers, the A225_{dFOG-F1} and the T222_{dFOG-F1} are conserved in two out of the three mouse FOG1 Zn fingers. The residue that is substituting the Alanine₂₂₅ is Valine in mFOG1_{finger56} and the Threonine₂₂₂ is Asparagine in mFOG1_{finger1}. Both Alanine and Valine have hydrophobic side-chains and the authors claim that this substitution does not affect the binding (Liew et al., 2005). Both Threonine and Asparagine have polar uncharged side chains. So the substitutions are with amino acids with similar properties. The S218_{dFOG-F1} and S219_{dFOG-F1} are also conserved in three out of the four proteins. But Serine is substituted with Arginine in both cases (mFOG1_{finger1} and mFOG1_{finger56} respectively). Although both amino acids have polar side chains of different length, the Arginine side chain has a positive charge. So although they have different properties the charge might be neutralized from a nearby negatively charged amino acid. The E224_{dFOG-F1} does not show any conservation. **This indicates that in the FOG1 fingers the**

majority of residues that are required for the GATA1 binding are conserved, which might explain their similar effect on the DNA binding specificities of GATA1.

We also note here that the dissociation of the mGATA1_{NC}/mFOG1_{finger1} upon DNA binding might not be unexpected. Wilkinson-White et al. (Wilkinson-White et al., 2011) have presented a weak association constant (K_a approx 1×10^{-5} M) for the mGATA1_{NF} and dFOG1_{finger1} complexation. This is not a FOG1 specific behavior as another GATA1 co-factor LMO2 interacts with the mGATA1 N-terminal Zn finger with similar affinities (K_a approx 10^4 - 10^5 M⁻¹) (Wilkinson-White et al., 2011). Although the DNA binding surface of GATA1 N-terminal Zn finger is in opposite site of the mFOG1 binding surface (Liew et al., 2005, Wilkinson-White et al., 2011) the weak association contact might not be adequate to stabilize the simultaneous binding of DNA and the FOG1 fingers. Nevertheless, it has been shown that a GATC containing DNA sequence from the *ALAS2* promoter, can bind purified mGATA1 NF in complex with purified mFOG1_{finger6} (Liew et al., 2005). This might also indicate that due to the degenerate nature of the library there might exist sequences that are more effective in causing dissociation of the complex upon binding.



Figure 59. Sequence alignment of the mFOG1 fingers 1, 9, 56 and the finger 1 of the *Drosophila* homolog of FOG1 (USH). ClustalW2 was used for multiple alignment. The ClustalW2 colour code is as follows: red represents small hydrophobic/aromatic amino acid (except Y), blue represents acidic amino acid, magenta represents basic amino acid (except H), green represents hydroxyl/sulfhydryl containing amino acids (including G). The orange vertical boxes represent amino acids that make contacts with the GATA1 N-terminal Zn finger. The yellow vertical boxes represent conserved amino acids across all proteins. The arrows indicate the Cysteines (C) and Histidines (H) that are necessary to coordinate the zinc ion. The finger 1 of the USH protein corresponds to amino acids 200-235 of the full length protein.

3.4.8.5 Functional redundancy of the FOG1 fingers in vivo may explain the similar DNA binding specificities

Fox et al (Fox et al., 1999) have shown that the FOG1 fingers 1, 6, 9 interact equally well with the GATA1 N-terminal Zn finger and that mutations, in the GATA1 N-finger, that disrupt the interaction of GATA1 N-finger with the FOG1 fingers have the same effect for all the fingers. Fox et al (Fox et al., 1999) claim that this implies that the FOG1 fingers and GATA1 interact with the same contact surface which hints that the FOG1 fingers have a functional redundancy. Indeed Fox et al (Fox et al., 1999) showed that mutations in some but not in all of its FOG1 fingers, that disrupt the binding with GATA1, **reduce but do not eliminate** the ability of the FOG1/GATA1 complex to repress a *KLF1* promoter containing reporter gene. We can hypothesize that this is due to the reduced interaction of GATA1 with FOG1.

Also, the fact that FOG1 has multiple redundant fingers it might imply that FOG1 can interact with multiple GATA1 proteins simultaneously. Cantor et al (Cantor et al., 2002) argues against it as it was shown that FOG1 mutants that interact with GATA1 with only one Zn finger can save terminal maturation in a FOG1^{-/-} erythroid progenitor cell line. He also noticed that the GATA1 interaction with only the FOG1 finger 1 or FOG1 finger 6 leads to a more substantial rescue of the erythroid maturation than the GATA1 interaction with only the FOG1 finger 9. This implies that the FOG1 fingers 1, 6 and 9 are not functionally equivalent in vivo and in the context of the whole FOG1 protein. These dissimilarities might be attributed to other reasons except the DNA binding specificities. Nevertheless, the fact that either of the FOG1 finger 1, 6 or 9 can save the erythroid maturation indicates that all these fingers can substitute each other in a very general perspective.

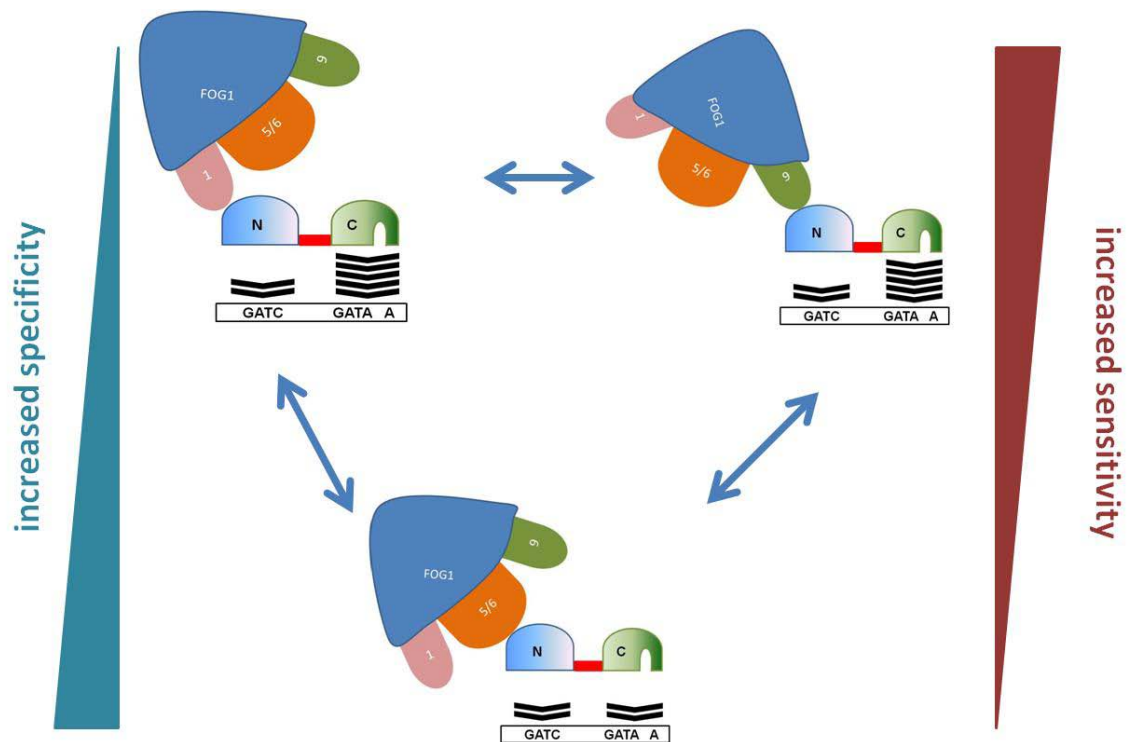


Figure 60. Model of the effect of the FOG1 fingers 1, 9 and 5/6 on the DNA binding affinities of GATA1. The black arrows indicate levels of binding affinity. The more black arrows appear on top of a DNA motif the higher the affinity of the DNA binding domain of the protein for the specific motif. The FOG1 fingers 1 and 9 of the GATA1 protein complexes compete with the DNA for the GATA1 binding. Only high affinity DNA motifs can effectively compete the interaction of the FOG1 fingers 1 or 9 with GATA1. This gives increased sensitivity on the specific DNA motif recognition from the GATA1/FOG1_{finger1} and GATA1/FOG1_{finger 9} complexes. The FOG1 fingers 5/6 do not show this behaviour.

3.4.9 Effect of KLF1/GATA1 complexation on the binding specificities of GATA1

In Figure 61A we present the frequency of flanking nucleotides (four upstream and four downstream) for the “**GATA**” motif on the bound 20mer DNA library for both the mGATA1_{NC} alone and the mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex. We also present the fold change of the frequency of occurrence of the nucleotides in the mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex in comparison with the mGATA1_{NC} alone. We notice that there is an increased frequency of mainly the Gua nucleotide and secondary of the Ade or Thy nucleotides at positions +1, +2, +3, +4 downstream of the “**GATA**” motif. The Cyt nucleotide did not show similar increase. At position -1 there is an increased frequency of the Ade and Thy nucleotides. Generally upstream of the “**GATA**” motif, there is mainly an increase in occurrence of the Ade and Thy nucleotides and secondary of the Cyt nucleotide. The Gua nucleotide did not show a significant increase in frequency. In Figure 61B we see that the distance distribution between the GATA1 motifs and the KLF1 motifs indicates that the GATA1-KLF1 binding orientation, on the 20mer bound sequences, is preferred rather than the KLF1-GATA1 binding orientation. Both GATA1 motifs and KLF1 motifs in Figure 61B are assumed to be orientation independent, meaning that we measure the distance between them independently on which strand they occur (either in the same or in different strands).

In Figure 61A there is the expectation that both the Gua and the Cyt nucleotides should be equally over-represented downstream of GATA1. In Figure 42 from the crystal structure of KLF4 we see extensive contacts of the KLF4 protein with the Gua nucleotides and not with the Cyt nucleotides. The increase of the Gua nucleotides downstream of the “**GATA**” motif might be an indication that the KLF1 protein tries to bind in a parallel or an anti-parallel orientation with the GATA1 motif (GATA{N}GG[AT]G / GATA{N}G[AT]GG) (Figure 62G-K). Thus, the behavior in Figure 61A indicates to us that by having a specific reference point (in this case the “**GATA**” motif) permit us to see a strand specific

preference in binding which reflects the true binding contacts. The binding model of the GATA1-KLF1 complex on a DNA sequence is presented in Figure 61C.

In Figure 17 we have seen that the KLF1 binding is weaker than the GATA1 binding. To further confirm that this is the case and is not the result of the different protein isolation batches we used the sequencing results from the mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ main band (dissociated mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex) where both proteins are expected to be present in equal amounts in the EMSA-seq reaction. If GATA1 has indeed higher affinity than KLF1 then we expect to have more GATA1 related 8mers with higher fold change than KLF1 ones. For this reason we took the top 150 mGATA1_{NC} 8mers and the top 150 mKLF1₂₅₇₋₃₅₈ 8mers (no 8mer is common between them). Then we ranked the 8mers of the mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ main band from the highest to the lowest and we tested in which position the mGATA1_{NC} specific 8mers and the mKLF1₂₅₇₋₃₅₈ specific 8mers appear. We noticed that the mGATA1_{NC} 8mers appear higher than the mKLF1₂₅₇₋₃₅₈ 8mers (Willcoxon test, $p\text{-value} < 10^{-6}$). **Thus indeed GATA1 has a higher affinity than KLF1. In vivo, the implications are that GATA1 probably interacts with KLF1 in order to stabilize its binding (increase the time that KLF1 occupies a specific genomic position).**

In Figure 63 we also notice that the position distribution of the GATA1 8mers on the sequenced 20mer bound library of the upper band (intact mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex) is shifted towards the edges of the 20mer. On the contrary the KLF1 8mers show no bias in the position distribution on the 20mer sequences and they are equally distributed throughout. The GATA1 shift towards the edge of the 20mer implies that the 20mer sequence might try to accommodate the binding from both proteins. Both proteins need at least 10 nucleotides each (1 full turn of DNA) in order to have efficient contacts with the major and minor groove of the DNA. Considering that GATA1 has a higher affinity than KLF1, this is something unexpected. We would have hypothesized that GATA1

drives the binding of the complex, so it can bind with no restriction throughout the 20mer sequence (no bias in the position distribution of the GATA1 8mers) and KLF1 follows with accessory binding, so KLF1 can now bind wherever it is available (expected bias in the position distribution of the KLF1 8mers). **The fact that this is not happening indicates that the extensive contacts of the three KLF1 Zn fingers drive the binding of the complex and can effectively outcompete the binding of the one and sometimes two GATA1 Zn fingers. Maybe, KLF1 acts as a rigid structure during the binding that forces the simultaneous binding of the three Zn fingers whereas GATA1 might be more flexible in the binding and can accommodate the binding of only one of its Zn fingers.**

3.4.9.1 Complexation of KLF1 with GATA1 changes the distance distribution between the C-terminal and the N-terminal Zn finger of GATA1

As discussed earlier we checked the distance distribution between the GATA-GATC, GATT-GATC, GATA-GATT, GATA-GATA for the KLF1/GATA1 complex. We avoided using the enriched 8mers of the mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex because they represent a mixture of the mGATA1_{NC} and mKLF1₂₅₇₋₃₅₈ 8mers. On the contrary, we projected the mGATA1_{NC} 8mers on the sequenced 20mer bound library of the mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex and we created the motif distance plots of Figure 64 . We see that the overall distributions of the mGATA1_{NC} alone and of the mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex are similar. The exception is that we see some increase in the frequency of the long bi-motif distances (orange boxes in Figure 64). Specifically we see an increased preference for GATA{N}₃₋₁₀**GATC**, GATA{N}₀₋₈**GATG**, **GATT**{N}₃₋₁₀TATC, **GATA**{N}₃₋₁₀TATC. **This indicates that the interaction of the KLF1 with the C-terminal Zn finger of GATA1 pushes the N-terminal Zn finger further apart and thus**

the mGATA1_{NC} of the mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex is able to bind bi-motifs with long separation between them.

We have to point out here that the increase in the preference for the GATA{N}₃₋₁₀**GATC** was also noticed in the mGATA1_{NC}/mFOG1 fingers complexes. Additionally, the increase in the preference for GATA{N}₃₋₁₀**GATC**, GATA{N}₀₋₈**GATG**, **GATT**{N}₃₋₁₀TATC and **GATA**{N}₃₋₁₀TATC was also noticed for the mGATA1_{NC} dimer. **This indicates that although mGATA1_{NC} interacts with different co-factors KLF1, FOG1 fingers and probably itself (dimer state) the functional outcome is funneled through a limited variety of conformational changes, which is reflected in the change of the mGATA1_{NC} DNA binding specificities.** As we have seen in the wt-mGATA1, the N-terminal and C-terminal tails change significantly the DNA binding specificities of GATA1. **Therefore, we can hypothesize that the addition in the system of the effect of the GATA1 tails might increase the variety of the functional outcomes and/or fine-tune the final response, as we have seen for the wt-mGATA1/wt-mFOG1 complex (increased preference for GATC{N}_{1,2}GATA).**

We also created the motif distance plots of the dissociated mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex (main band). We again projected the mGATA1_{NC} 8mers on the sequenced 20mer bound library (Figure 65). We see that the overall distribution for the **GATA-GATT** and **GATA-GATA** distances of the mGATA1_{NC} alone (Figure 36, Figure 41) and of the dissociated mGATA1_{NC} from the complex is similar. The frequency of occurrence is lower but this probably is the cause of the existence of KLF1 bound 20mers on the bound library that forces us to overestimate the total number of GATA1 related 20mers on the bound library. On the contrary, the dissociated complex shows a very restricted pattern of only the **GATGATA**, **GATGATT** motifs for the **GATG-GATA** distance and the **GATG-GATT** distance respectively. For the **GATC-GATA** distance and the **GATC-GATT** distance, the dissociated complex shows similar restriction with preference for mainly the **GATAGATC**,

GATCGATA, GATTGATC, GATCGATT motifs and to a lesser extend for the **GATC{N}₁₋₈GATA, GATC{N}₁₋₈GATT** motifs. As we have seen, for the mGATA1_{NC}/mFOG1 fingers complexes, the differences between them and the mGATA1_{NC} were attributed to the FOG1 fingers effect on the DNA binding specificities of mGATA1_{NC}. Similarly here, the differences between the dissociated mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex and the mGATA1_{NC} probably reflect a biological phenomenon. **The fact that the intact mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex and the dissociated mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex have different GATA1 related DNA binding specificities, which are in both cases different than the mGATA1_{NC} ones, indicates that the intact and the dissociated complex represent two different populations of the mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex with different properties.**

Because the **GATC-GAT(A/T)** and the **GATG-GAT(A/T)** distance distribution patterns of the mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ dissociated complex look similar with the mGATA1_C ones, we can hypothesize the following model. In the mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex, we can hypothesize that KLF1, before it dissociates, forces the C-terminal Zn finger of GATA1 to bind on the 20mer sequence irrespectively of the N-terminal Zn finger sequence preference. Thus, KLF1 either inhibits the GATA1 N-terminal Zn finger DNA binding or it increases the affinity of (stabilizes) the GATA1 C-terminal Zn finger DNA binding. As we have already discussed, DNA sequences that contain both the **“GATC”** and **“GATA”** motifs have a higher binding affinity for the **GATA1_{NC} alone** than sequences that have only **“GATA”** motifs (Trainor et al., 2000). **Thus we favor a model, where KLF1 increases the binding affinity of the C-terminal Zn finger and this is adequate to increase the GATA1 binding in “GATA” containing sequences without the necessity for KLF1 to bind nearby. We argue here, that KLF1 probably affects the rate of association (“Ka”) of GATA1 with the DNA sequence and not the rate of dissociation of GATA1 from the DNA sequence.**

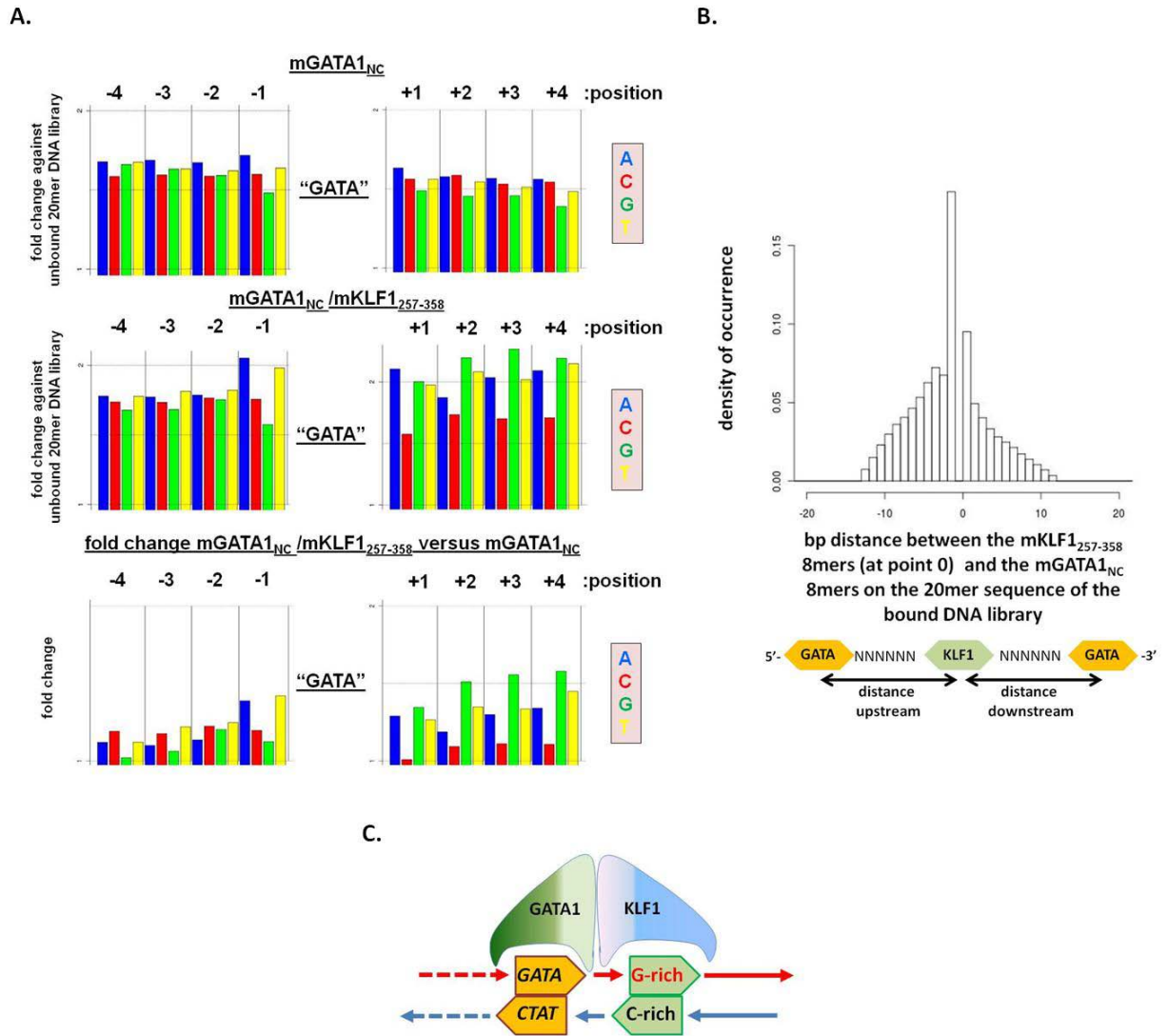
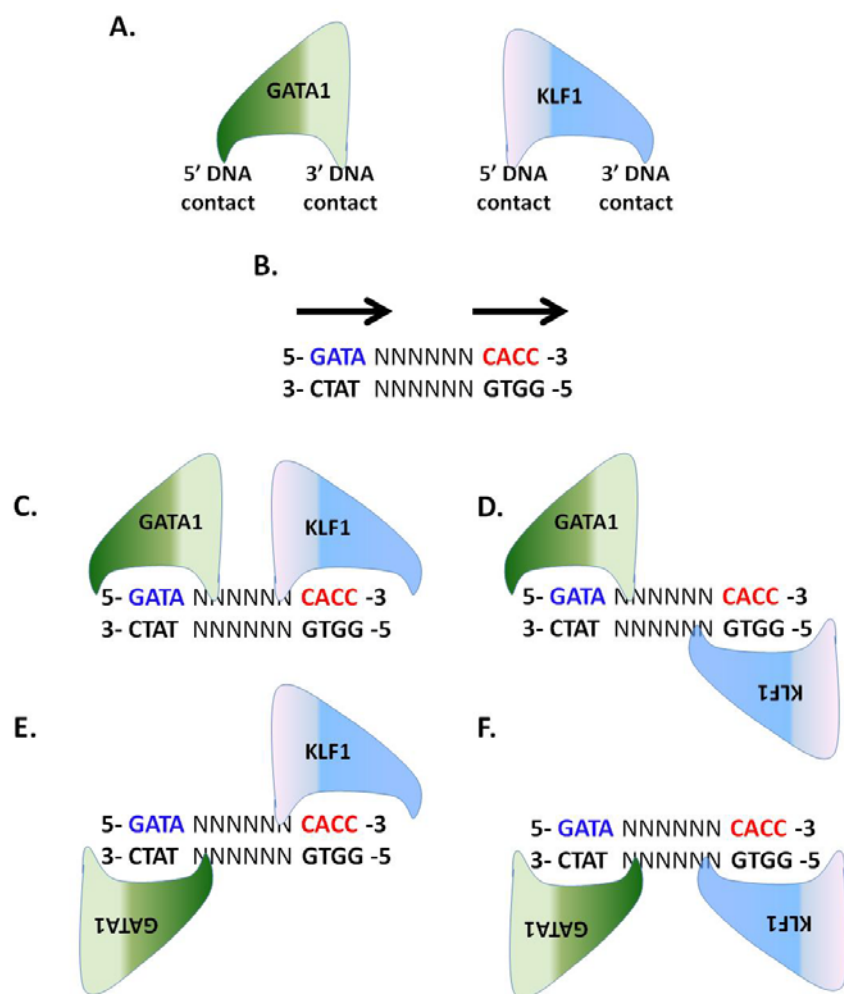


Figure 61. A) Frequency of nucleotides upstream and downstream of the “GATA” motif for both the $mGATA1_{NC}$ alone and the intact $mGATA1_{NC}/mKLF1_{257-358}$ complex ($mGATA1_{NC}/mKLF1_{257-358}$ upper band). The positive (+1 to +4) and the negative (-4 to -1) numbers represent the nucleotide positions downstream and upstream of the indicated motif. The nucleotides are indicated with different colours: Adenine (blue), Cytosine (red), Guanine (green), Thymine (yellow). B) Distance distribution between the 2122 enriched $mGATA1_{NC}$ 8mers and the 193 $mKLF1_{257-358}$ 8mers on the 20mer sequences of the degenerate DNA library bound from the intact $mGATA1_{NC}/mKLF1_{257-358}$ complex ($mGATA1_{NC}/mKLF1_{257-358}$ upper band). Both the GATA1 and the KLF1 motifs are assumed to be orientation independent (see main text). C) Model for the interaction of the GATA1

and KLF1 proteins. KLF1 is assumed to be present with a specific orientation relative to the G-rich side of the “C[AT]CCC” motif. GATA1 is also assumed to be present with a specific orientation relative to the “GATA” nucleotides side of the “GATA” motif. The two proteins interact so that the “GATA” nucleotides side, of the “GATA” motif, and the “G-rich” side, of the “C[AT]CCC” motif, look at the same orientation.



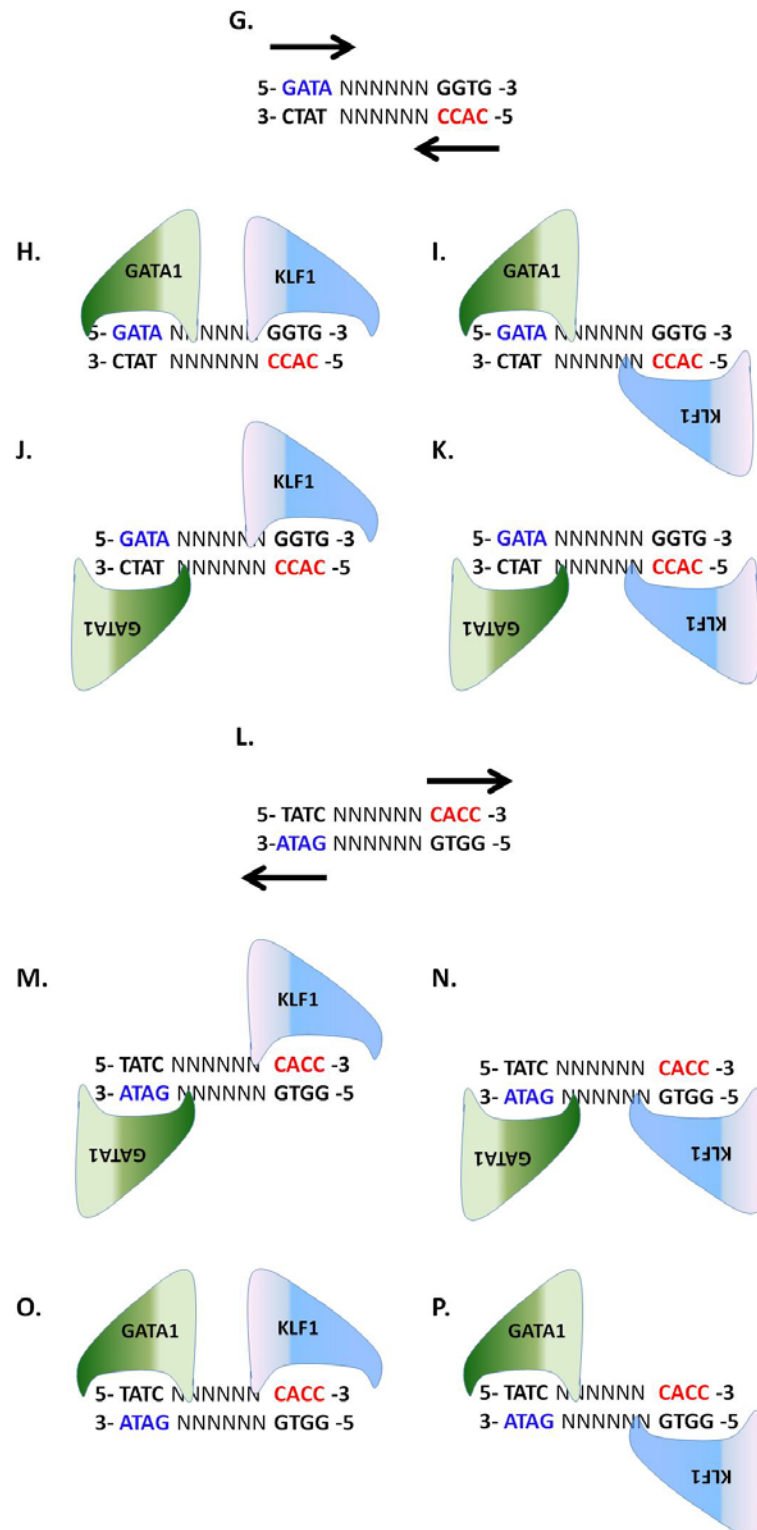


Figure 62 (previous two pages). Working hypothesis for the orientation of the GATA1 and KLF1 proteins on DNA sequences where both GATA1 binding sites and KLF1 binding sites exist. A) The GATA1 protein is presented with the green scheme and the KLF1 protein is presented with the blue scheme. Both schemes represent the proteins as asymmetric units with the light colour end of the scheme making contact with the 3' or 5' part of the DNA, and the dark colour end of the scheme making contact with the 5' or 3' part of the DNA, for the GATA1 or KLF1 respectively. B-P) Potential orientation of the GATA1 and the KLF1 proteins on the DNA sequence for the case where the "GATA" nucleotides side of the GATA1 motif ("GATA" motif) and the C-rich side of the KLF1 motif ("CACC" motif) are facing towards the same direction (B-F) or the case where they are facing different directions (G-P).

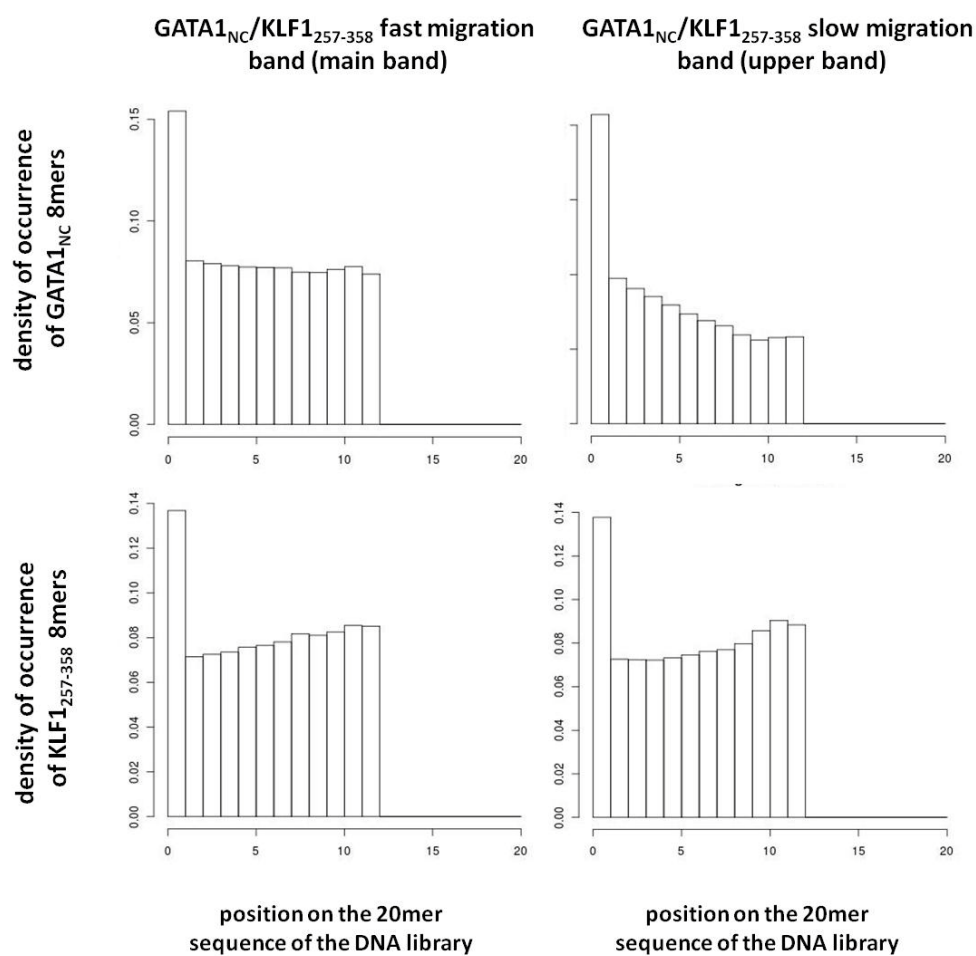
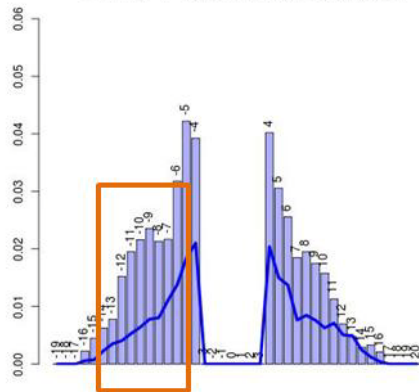
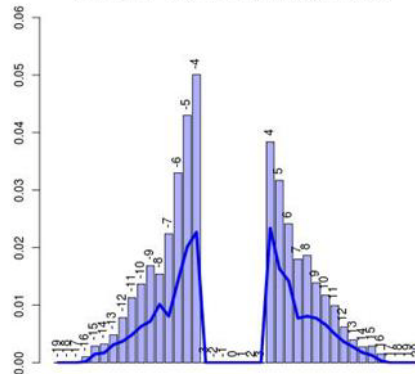


Figure 63. Distribution of the mGATA1_{NC} and mKLF1₂₅₇₋₃₅₈ 8mers positions on the 20 bp sequence of the bound degenerate DNA library for the mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex. The position distribution of both the slow (upper) and fast (main) migrating bands are shown. The slow migrating band represents the DNA band shift of the intact mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex. The fast migrating band represents the DNA band shift of the dissociated mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex.

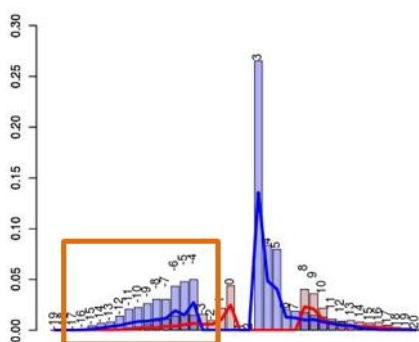
GATA-GATC distance



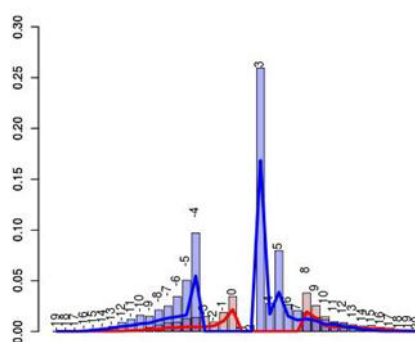
GATT-GATC distance



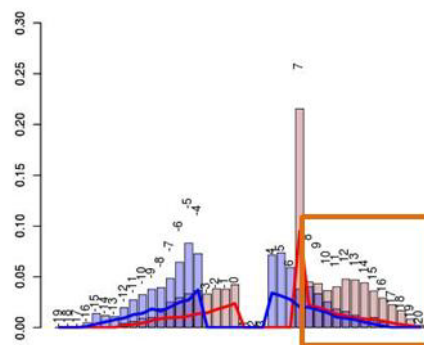
GATA-GATG distance



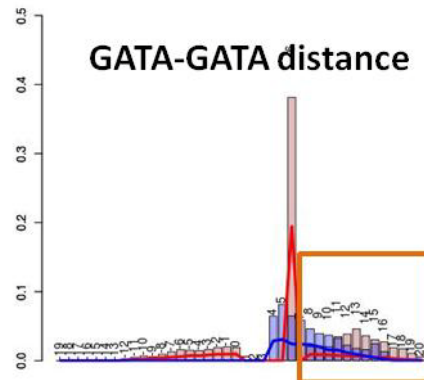
GATT-GATG distance



GATA-GATT distance



GATA-GATA distance



binding model of
protein on DNA

KLF1 protein

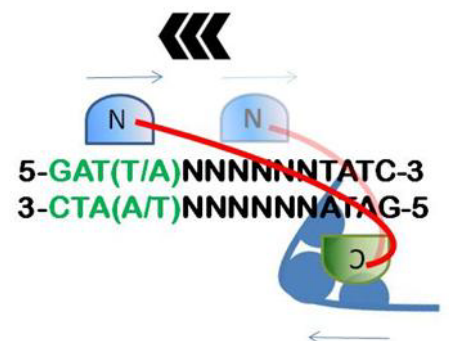
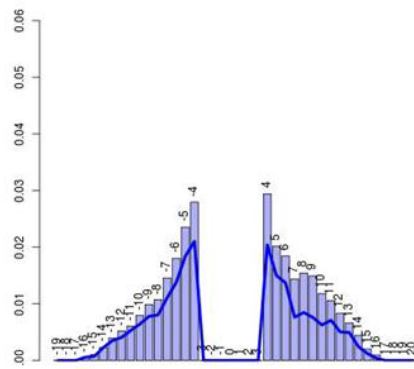
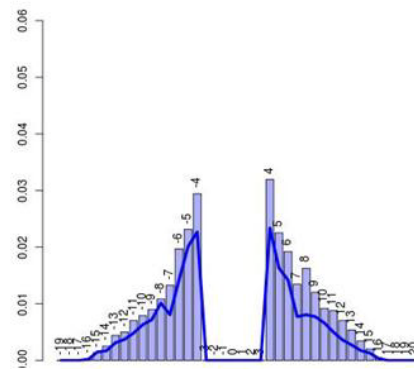


Figure 64 (previous page). Distance graphs between the “GATA” and the “GATC” motifs, the “GATT” and the “GATC” motifs, the “GATA” and the “GATG” motifs, the “GATT” and the “GATG” motifs, the “GATA” and the “GATT” motifs or between two co-occurring “GATA” motifs. The graphs represent the distance pattern for the bound 20mer sequences of the intact mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex (mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ upper band). The distance graphs were created with the 2122 enriched 8mers from the mGATA1_{NC} experiment. The binding models of the protein on DNA are also shown. For further explanation about the graph see Figure 33. The blue arrows represent the orientation of the motifs as appear in Figure 33A.

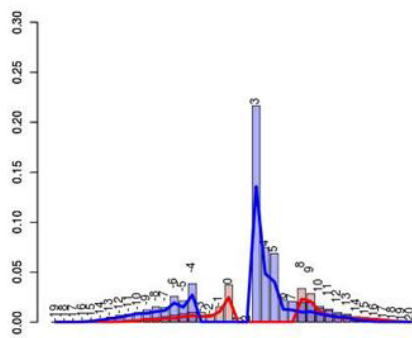
GATA-GATC distance



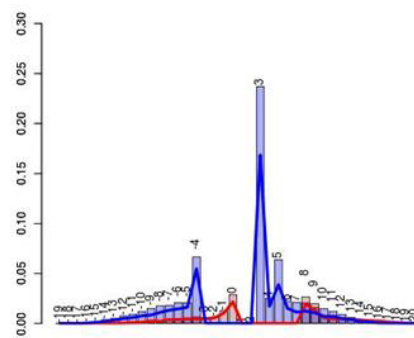
GATT-GATC distance



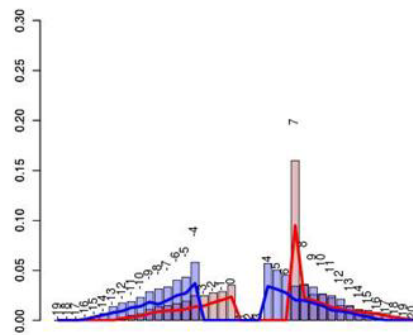
GATA-GATG distance



GATT-GATG distance



GATA-GATT distance



GATA-GATA distance

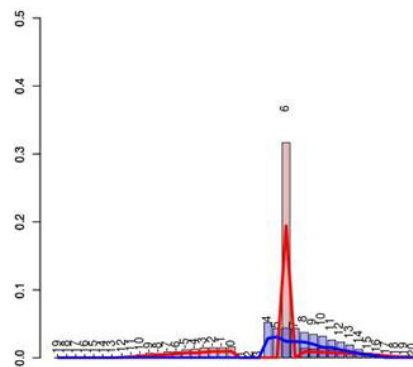
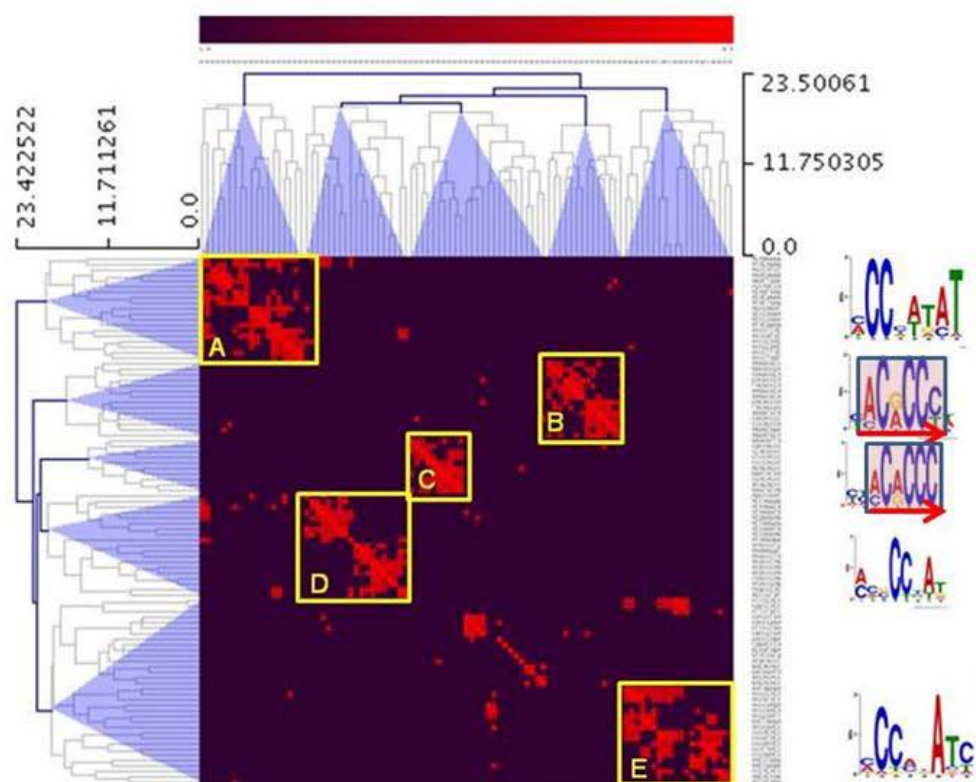


Figure 65 (previous page). Distance graphs between the “GATA” and the “GATC” motifs, the “GATT” and the “GATC” motifs, the “GATA” and the “GATG” motifs, the “GATT” and the “GATG” motifs, the “GATA” and the “GATT” motifs or between two co-occurring “GATA” motifs. The graphs represent the distance pattern for the bound 20mer sequences of the dissociated mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex (mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ main band). The distance graphs were created with the 2122 enriched 8mers from the mGATA1_{NC} experiment. The binding models of the protein on DNA are also shown. For further explanation about the graph see Figure 33C.

3.4.10 Effect of KLF1/GATA1 complexation on the binding specificities of KLF1

In Figure 66 we see KLF1 8mers that are enriched upon complexation with GATA1. We see an enrichment of the AC[GA]CCC motif (group B,C). Upon projection of the motif on the KLF4/DNA crystal structure, we can hypothesize that this enriched motif represents mainly the binding of the KLF1 Zn fingers 2 and 3. Thus complexation with GATA1 increases the binding of the KLF1 Zn finger 2 and 3 and reduces the binding of the KLF1 Zn finger 1.

A.



B.

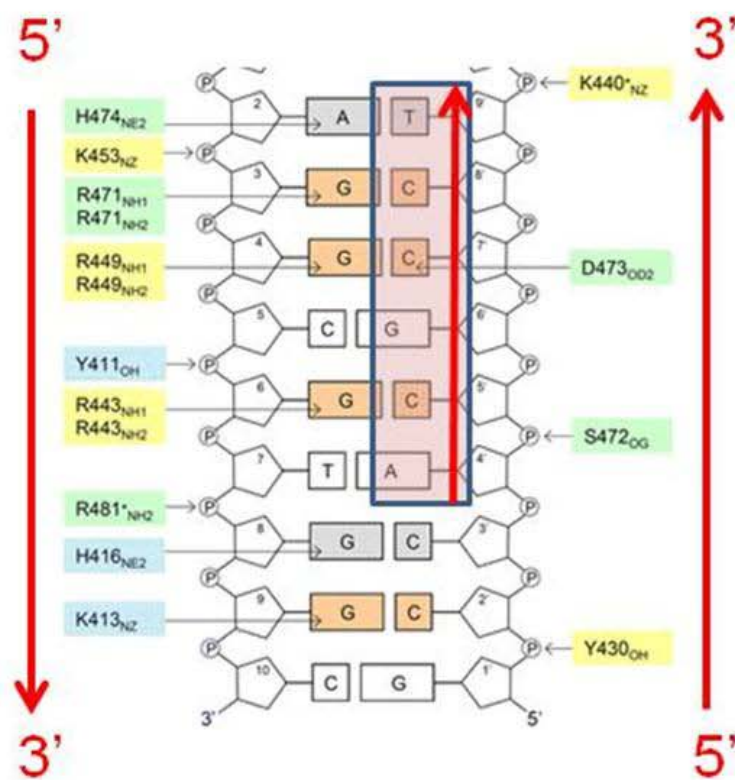


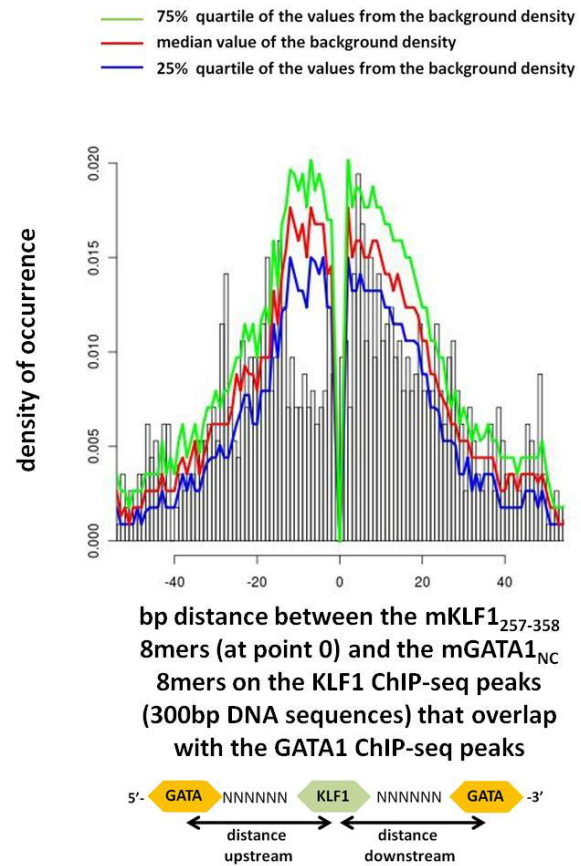
Figure 66 (previous page). A) Enriched 8mers from the EMSA-seq experiment of the intact mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ complex (mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ upper band). The 8mers were clustered based on their Hamming distance (maximum 2 mismatches). B) Hypothetical position of the group B and the group C motifs on the KLF4 crystal structure (The KLF4 - DNA interactions graph is from (Schuetz et al., 2011)).

3.4.11 The orientation of the GATA1-KLF1 complex in vivo indicates a mechanism of local DNA looping

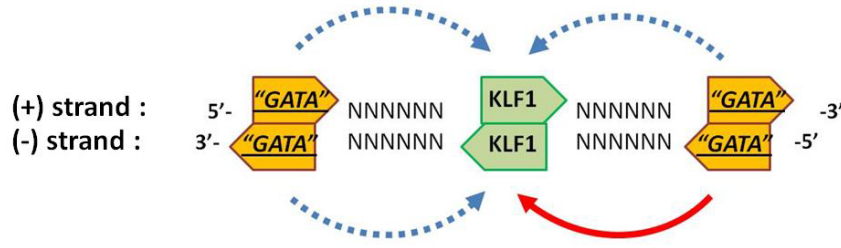
We checked whether in vivo, the in vitro behaviour of the KLF1-GATA1 complex is retained. We used positions where both the GATA1 ChIP-seq peaks and the KLF1 ChIP-seq peaks overlap in the TER119-/TER119+ model of erythropoiesis. We also used positions where the GATA1 ChIP-seq peaks and the KLF1 ChIP-seq peaks do not overlap in the same model. Similarly with Figure 61B, we examined the distance distribution of the GATA1, KLF1 motifs in the overlapping GATA1, KLF1 ChIP-seq positions (Figure 67A). We also examined this distance distribution in a set of 1000 different samples, of the same size as the overlapping set, taken from the GATA1 ChIP-seq peaks that do not overlap with KLF1. Although we were expecting similar image like the Figure 61B, with an enrichment above the background distribution, in Figure 67A we see that the GATA1-KLF1 orientation is preferred less than the background. On the other hand the KLF1-GATA1 orientation is preferred equal with the background. This differential presence of the two orientations of the GATA1/KLF1 complex, although it did not show any enrichment, points towards a functional role for the KLF1-GATA1 orientation. This is in contrast with the in vitro results where the GATA1-KLF1 orientation was preferred. Because the distances were calculated in orientation independent motifs and considering that we saw that in vitro the “GATA” nucleotides side, of the “GATA” motif, is frequently found, on the same strand, with the “G-rich” side of the “C[AT]CCC” motif (Figure 61C), we decided to examine this in vitro preference in the GATA1/KLF1 overlapping ChIP-seq peaks. At first we checked whether the association (see legend of Figure 67B) of the “GATA” nucleotides side, of the “GATA” motif with either the “G-rich” side or the “C-rich” side, of the “C[AT]CCC” motif is statistically significant, meaning that it is not found in the 1000 sets of the background distribution. Indeed, we found that it is statistically significant for the KLF1-GATA1 orientation (Figure 67B). We then examined whether the “GATA”

nucleotides side, of the “GATA” motif, is frequently found, on the same strand, with the “G-rich” side or the “C-rich” side, of the “C[AT]CCC” motif. We noticed that the “GATA” nucleotides side, of the “GATA” motif is frequently found, on the same strand, with the “C-rich” side of the “C[AT]CCC” motif (Figure 67C). We then assumed that the in vitro orientation of the GATA1-KLF1 complex is the optimal for the complex (preference for the “GATA” nucleotides side of the “GATA” motif, to be frequently found in association with the “G-rich” side of the “C[AT]CCC” motif). Thus, we conclude that the DNA loops in order to be able to accommodate the proteins in the correct orientation (Figure 68).

A.



B.



C.

% of the Gua and Cyt nucleotides occurrence on the (-) strand of the KLF1 8mers for the statistically significant association ("GATA" motif in minus strand- KLF1 8mers) (red arrow in figure (B))

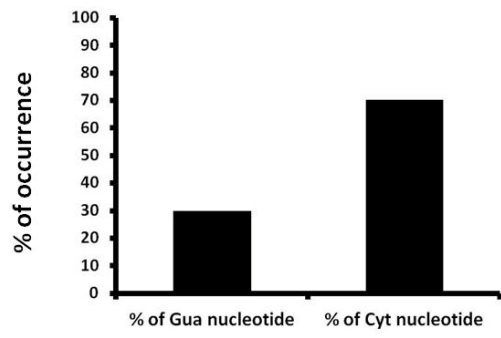


Figure 67 (previous page). A) Distance distribution between the non directional GATA1 8mers and the non directional KLF1 8mers in the GATA1/KLF1 overlapping ChIP-seq peaks in the TER119+ erythroblasts (vertical bars). The continuous lines (green, red and blue) represent the distance distribution of the GATA1 8mers and the KLF1 8mers in 100 different groups of the GATA1 ChIP-seq peaks that do not overlap with the KLF1 ChIP-seq peaks (background density). Each one of the 100 background samples was of equal size as the GATA1/KLF1 overlapping ChIP-seq group. B) Association between the “GATA” nucleotides side of the GATA1 motif and the GC content of the orientation independent KLF1 8mers in the GATA1/KLF1 overlapping ChIP-seq peaks of figure (A). The red arrow indicates statistically significant deviations (over-representation of the Cyt nucleotides, z-score=6) from the expected background distribution of 1000 subsamples from the GATA1 ChIP-seq peaks that do not overlap with the KLF1 ChIP-seq peaks (same number of peaks as the GATA1/KLF1 overlapping ChIP-seq peaks). For the “GATA” nucleotides side of the GATA1 motif in either the plus or the minus strand, in either the GATA-KLF1 or the KLF1-GATA1 orientation (four combinations in total, four arrows), we calculated the total number of either the Cyt or the Gua nucleotides, that appear in the same strand as the “GATA” nucleotides strand, of the orientation independent KLF1 motifs. For each peak the closest KLF1 8mer to the “GATA” motif was considered. In case of multiple “GATA” motifs of the same orientation, in a specific ChIP-seq peak, the mean occurrence of either the Cyt or the Gua nucleotides was calculated for the specific ChIP-seq peak. In case of multiple “GATA” motifs of different orientations, in a specific ChIP-seq peak, the occurrence of either the Cyt or the Gua nucleotides was assigned independently for the specific ChIP-seq peak in each one of the different background distributions. The occurrence of either the Cyt or the Gua nucleotides was summed across all the examined ChIP-seq peaks. A z-score was calculated by performing similar analysis in the background subsamples and creating eight background distributions of the total number of the Cyt or the Gua nucleotides for each one of the four associations (four for the Cyt nucleotides and four for the Gua nucleotides). Statistically

significant associations in the GATA1/KLF1 overlapping ChIP-seq group were the ones where the total number of either the Cyt or the Gua nucleotides were at least four standard deviations higher than the mean ($z\text{-score} \geq 4$) of the relevant background distribution. The blue dotted arrows represent non-statistically significant associations. C) Percentage of the Cyt nucleotide versus the Gua nucleotide in the KLF1 motif of the statistically significant association of figure (B). The GATA1 ChIP-seq peaks were from (Xu et al., 2012) (GEO accession number: GSM970258). The KLF1 ChIP-seq peaks were from (Su et al., 2013) (GEO accession number: GSM1067275).

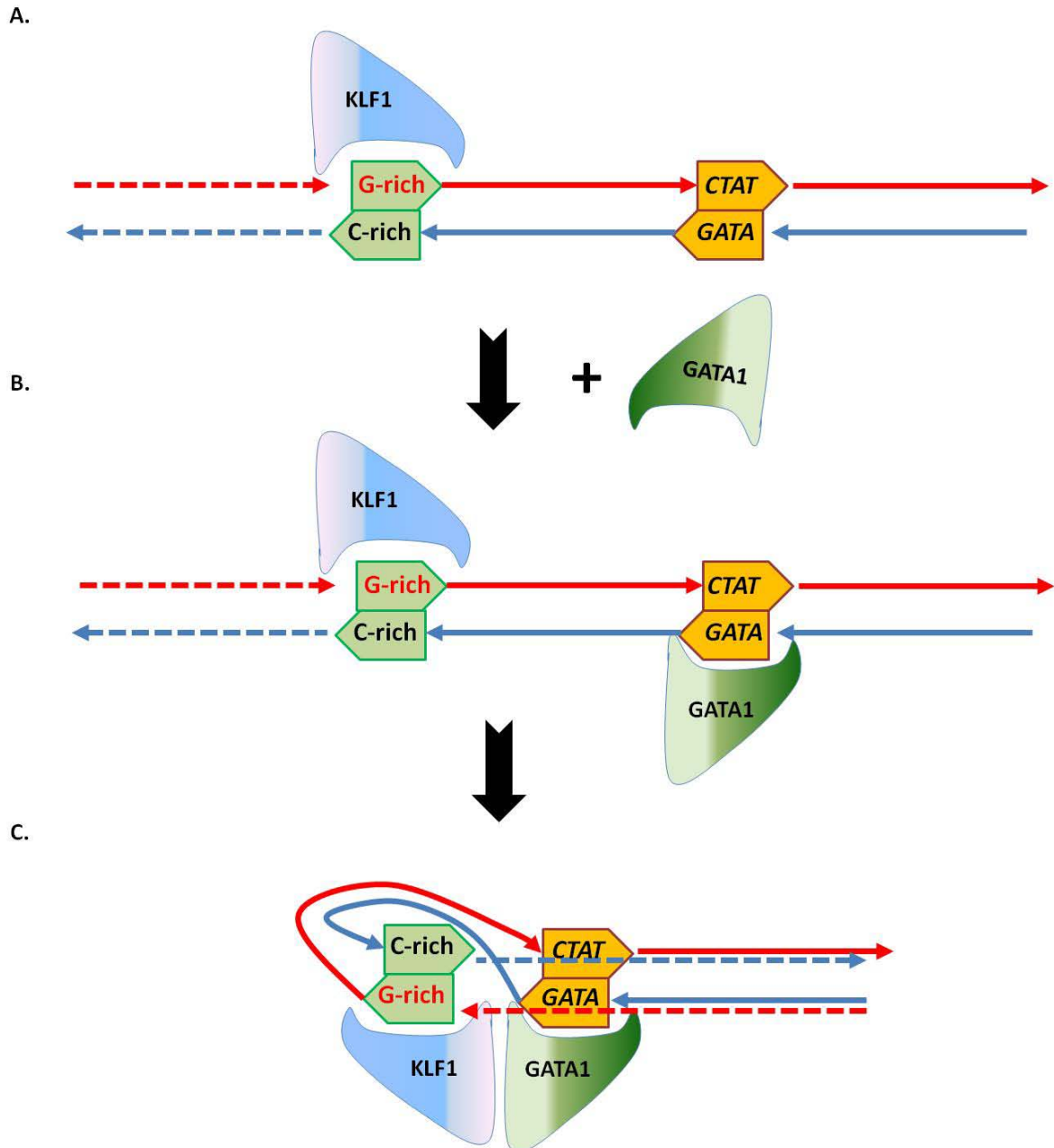


Figure 68. Model for the interaction of KLF1 and GATA1 in vivo. A) KLF1 is assumed to be present with a specific orientation relative to the G-rich side of the “C[AT]CCC” motif. B) GATA1 is also assumed to be present with a specific orientation relative to the “GATA” nucleotides side of the “GATA” motif. C) We assume that KLF1 is present on the DNA before GATA1 arrives (Vernimmen et al., 2007). Upon GATA1 arrival the 2 proteins interact in a specific way, that needs looping of the DNA sequence which lies between them, so that the “GATA” nucleotides side, of the “GATA” motif, and the “G-rich” side, of

the “C[AT]CCC” motif, look at the same orientation. The functional significance of this orientation might be the fact that the DNA sequence upstream of the KLF1 motif and the DNA sequence downstream of the GATA1 motif are brought closer together which either facilitates the recruitment of specific protein complexes and/or inhibits the recruitment of nucleosomes.

3.5 Projection of the EMSA-seq DNA binding affinities/specificities on the GATA1 related erythropoietic gene regulation models

We projected the EMSA-seq DNA binding affinities/specificities of GATA1 and its co-factors on GATA1 related ChIP-seq data. We used data from both the mouse G1E/G1ME model of erythropoiesis/megakaryopoiesis, the mouse embryonic liver erythroblasts model [TER119- progenitor erythroblasts (R1, R2 populations)/TER119+ mature erythroblasts (R3, R4, R5 populations)] and finally the human CD34+ progenitors/CD36+ mature erythroblasts model (see introduction section 1.5). In Figure 69 we see that the human and mouse GATA1 proteins in the DNA binding domain (N-terminal Zn finger and C-terminal Zn finger) are highly conserved. This permits us to generalise the GATA1 behaviour independently of the model system used.

At first we created a thermodynamic model to score the GATA1 and KLF1 ChIP-seq peaks with the “Kd” values of the mGATA1_{NC} and mKLF1₂₅₇₋₃₅₈ 8mers respectively.

3.5.1 Calculation of the DNA binding affinities of GATA1/KLF1 ChIP-seq peaks from the “Kd” values of each mGATA1_{NC}/mKLF1₂₅₇₋₃₅₈ 8mer

In order to calculate the binding affinities of a ChIP-seq peak we used the approach that is presented below. At first, we make the following assumptions:

- i) Its ChIP-seq peak sequence comprises of a group of k-mers (k-mer1, k-mer2, ..., k-merN) with $N = (\text{length of the ChIP-seq peak sequence}) - k$. In our case $k=8$, the length of a ChIP-seq peak sequence is 300 bp and eventually $N=292$ bp.
- ii) The ChIP-seq peak binding affinity is independent of the k-mer order on the sequence.

- iii) The protein is going to bind to each k-mer independently. The total protein bound on the sequence is the sum of the binding on the independent k-mers. This also implies that the overlapping nature of k-mers is not considered at all and each k-mer is treated independently of the others.

So the total amount of protein bound is:

$$[\text{Total protein bound on sequence}] = \sum_{N=1}^{32896} [\text{k-merN bound on protein}] \quad (\text{eq 19})$$

From eq 2, eq 3 we have:

$$[\text{k-mer1 bound on protein}] = \frac{[\text{protein not bound on any k-mer}] * [\text{free k-mer1}]}{Kd_{\text{k-mer1}}} \quad (\text{eq 20})$$

$$[\text{k-mer2 bound on protein}] = \frac{[\text{protein not bound on any k-mer}] * [\text{free k-mer2}]}{Kd_{\text{k-mer2}}} \quad (\text{eq 21})$$

$$[\text{total protein bound on sequence}] = \frac{[\text{protein not bound on any k-mer}] * [\text{free sequence}]}{Kd_{\text{of sequence}}} \quad (\text{eq 22})$$

So if we substitute in eq 19 the eq 20, eq 21, eq 22 we have:

$$\frac{[\text{protein not bound on any k-mer}][\text{free sequence}]}{Kd_{\text{of sequence}}} = \sum_{N=1}^{32896} \frac{[\text{protein not bound on any k-mer}][\text{free k-merN}]}{Kd_{\text{k-merN}}}$$

(eq 23)

We also make the assumption that at any given point only one of the k-mers can bind.

Then because k-mer1 and k-mer2 appear on the sequence in 1:1 ratio we have:

$$[\text{free k-mer1}] = [\text{free k-mer2}] = [\text{free sequence}]$$

We note here that even similar k-mers are treated independently. So eq 23 can be simplified as:

$$\frac{1}{Kd_{\text{of sequence}}} = \sum_{N=1}^{32896} \frac{1}{Kd_{\text{k-merN}}}$$

By rearranging we have:

$$Kd_{\text{of sequence}} = \frac{\prod_{N=1}^{292} Kd_{\text{k-merN}}}{\sum_{N=1}^{292} \frac{\prod_{N=1}^{292} Kd_{\text{k-merN}}}{Kd_{\text{k-merN}}}}$$

(eq 24)

For proteins/complexes that we did not have EMSA-seq “Kd” affinities of 8mers we scored their ChIP-seq peaks with the number of statistically enriched 8mers. For proteins that we have “Kd” affinities of 8mers, we also scored their ChIP-seq peaks with the number of statistically enriched 8mers when we compare them with proteins which we did not have “Kd” affinities.

All the ChIP-seq peaks binding affinity “Kd” scores that are calculated using “Kd” values derived from the EMSA-seq experiments we will call them **EMSA-seq binding affinity Kd scores**. The lower the EMSA-seq binding affinity Kd scores the higher the binding affinity.

In case that the enrichment value of the k-mers (the fold change of the frequency of the k-mer in the bound fraction versus the frequency of the k-mer in the original 20mer DNA degenerate pool) has to be used then the **EMSA-seq binding affinity Kd scores** is intuitively similar with the average fold change of all the k-mers in each ChIP-seq peak. We caution here that when the average fold change of all the k-mers from two different proteins/protein complexes for the same ChIP-seq peak are compared, a normalised fold change value for each protein needs to be used. As we have seen the absolute fold change depends on the protein/DNA ratio used, on the affinity (“Kd”) of the protein for a DNA sequence and on the effect that co-factors can have on the DNA binding affinities (competition). The normalised fold change score is something that we did not explore.

For the above mentioned reasons, we used as an alternative score the number of statistically enriched EMSA-seq k-mers that are present in each one of the ChIP-seq peaks. As expected this score gives equal weight in two ChIP-seq peaks that have the same number of k-mers, something that causes loss of sensitivity. **On the other hand, we argue that the presence or absence of an enriched k-mer in two ChIP-seq peaks with different number of k-mers is probably more important than their absolute affinity values due to the cooperative nature of k-mers which we discussed in an earlier chapter.**

The number of statistically enriched EMSA-seq k-mers that are present in each one of the ChIP-seq peaks we will call them **EMSA-seq k-mer number score**. The higher the EMSA-seq k-mer number score the higher the binding affinity.

In the following chapters, we are using the **EMSA-seq binding affinity Kd score** when we compare the GATA1 or KLF1 ChIP-seq peaks binding affinities against a biological phenomenon (nucleosome occupancy, DNase I accessibility, H3K4me3/2 levels, distance of a ChIP-seq peak from the TSS, presence of the ChIP-seq peak on enhancers or promoters).

We used the **EMSA-seq k-mer number score** when we compare the affinity distributions of different proteins that at least one of them does not have Kd affinity values for each one of the k-mers. We also used this score when we compare the binding affinities of a fraction of the GATA1 ChIP-seq peaks dataset against the binding affinities of the rest of the dataset (GATA1 ChIP-seq peaks that overlap with a co-factor versus GATA1 ChIP-seq peaks that do not overlap). This was selected because the distributions that are compared are discrete and have much more data points (so there is a range of values that can give more confidence for the biological phenomenon that we observe) and the scoring function is intuitively straightforward.

In some cases, we also confirmed that both scoring schemes (EMSA-seq binding affinity Kd score, EMSA-seq k-mer number score) give similar results.

3.5.2 GATA1 ChIP-seq binding affinities, assigned from the EMSA-seq results, are representative of the GATA1 DNA binding turnover in vivo

3.5.2.1 Prediction of the kinetic recruitment of the protein upon GATA1 restoration in a GATA1 ^{-/-} cell line

We hypothesized that our in vitro DNA binding affinities will be representative of the protein turnover at the occupied ChIP-seq positions in vivo. To check whether this assumption is true, we used data from the GATA1 restoration in the G1E model of erythropoiesis. These are genome wide data of the GATA1 kinetic recruitment at its binding sites from a 24 hours time course ChIP-seq study. There have also been two published reports of the kinetic recruitment of GATA1 in specific loci in the same model assessed through ChIP-qPCR. The genome wide data include these loci.

We will start with the low-throughput studies. These studies examined the α -globin locus, the β -globin locus and the GATA2 locus. The positions that have been tested with ChIP-qPCR are noted with red asterisks on Figures 70-72. These positions are:

- i) For the α -globin locus the α -globin promoter and the α -globin enhancer HS-26 (Pope and Bresnick, 2010).
- ii) For the β -globin locus the β -major promoter (gene *Hbb-b1*) and the β -globin enhancer HS2 (Pope and Bresnick, 2010).
- iii) For the GATA2 locus the Hypersensitive site (HS) positions -77, -3.9, -2.8, +9.5 (Grass et al., 2006).

At first we checked in our ^{bio}FOG1 ChIP-seq dataset whether these positions show a FOG1 ChIP-seq signal. Although the FOG1 ChIP-seq was performed in MEL cells we assumed that it is representative of the presence of FOG1 in the G1E model. Indeed we

found FOG1 signal in all the previous positions with the exception of the GATA2 promoter (GATA1 ChIP-seq peaks -3.9, -2.8) and the GATA1 ChIP-seq peak inside the GATA2 gene (+9.5). All the mentioned positions, including the GATA2 promoter peaks and the GATA2 intragenic peak, have been shown in the literature to have FOG1 (Pope and Bresnick, 2010, Grass et al., 2006). Lack of FOG1 might indicate poor ChIP efficiency or cell-type specific preferences. In the pre-mentioned positions we checked whether the kinetics of recruitment reflect the affinity of binding. This means that ChIP-seq positions, that attract the protein quicker than other regions, should have higher binding affinity (**high values of EMSA-seq k-mer number scores**).

We first focus on the alpha and beta globin regions (promoter of the *Hba-a1* and *Hbb-b1* gene and the HS-24 and HS-2 sites respectively). Because these positions have FOG1 protein we focused on the binding affinities of the wt-mGATA1/wt-mFOG1 complex. The Hypersensitive Sites recruit GATA1 quicker than the promoters. Indeed the **EMSA-seq k-mer number scores** are higher in the HS sites than in the promoters. We then focused on the GATA2 loci. The -77, -3.9, -2.8 and +9.5 all have been shown to have FOG1. It has been also shown that the -77 and the -2.8 position GATA1 recruitment is higher than the -3.9 and +9.5 position. The **EMSA-seq k-mer number scores** at the -2.8 and the +9.5 are higher than the -77 and -2.8 positions. The **EMSA-seq k-mer number scores** of the -2.8 and +9.5 are similar. This discrepancy might be caused because the GATA1 protein might be in a complexation with other factors that increase or decrease its binding affinity.

We then used the genome wide ChIP-seq data of the kinetic recruitment of GATA1 in its binding sites. The ChIP-seq of GATA1 was done 3 h, 7 h, 14 h and 24 h after GATA1 restoration in the G1E model of erythropoiesis. We expressed the ChIP-seq reads for each binding position in each time point as a fraction of the number of reads of the 24 hours GATA1 ChIP-seq (Figure 73). We hypothesized that high affinity binding sites (**high EMSA-seq k-mer number score**) should have a higher fraction of the 24 hours ChIP-seq reads (high saturation) than lower affinity sites (**low EMSA-seq k-mer number score**).

(low saturation). Indeed we notice that for the 3 hours and the 7 hours GATA1 ChIP-seq the high binding affinity sites show high saturation whereas the low affinity sites show low saturation (Figure 73). The 14 hour GATA1 ChIP-seq dataset showed equal saturation across the high and low affinity binding sites implying that the binding of the protein on the genome has reached steady state levels (Figure 73).

All the above indicated that the EMSA-seq binding affinities represent the GATA1 DNA binding turnover in vivo and potentially can be used to infer mechanistic principles of the system (Figure 74).

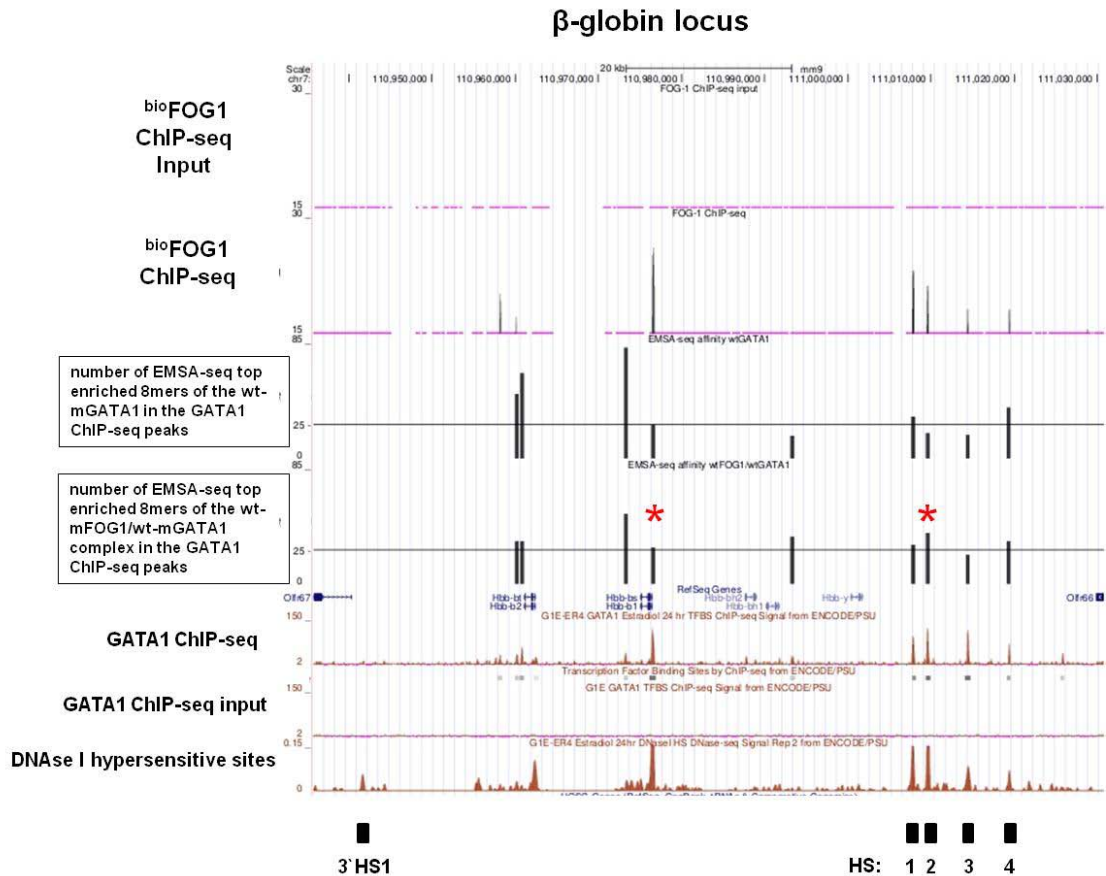


Figure 70. Profile of the beta-globin locus (chr7:110.930.000-111.030.000). The EMSA-seq derived DNA binding affinities of the GATA1 ChIP-seq peaks (tracks with black columns) are shown along with the GATA1 ChIP-seq peaks themselves (orange tracks). The ^{bio}FOG1 ChIP-seq peaks from the differentiated MEL cells and their respective input are also been shown. The red stars represent regions where the kinetic studies of the GATA1 recruitment after GATA1 restoration in the G1E model has been performed (see text). The row below the gene track represents published DNase I hypersensitive sites (HS). The GATA1 ChIP-seq peaks and the DNase I hypersensitive sites are derived from differentiated G1E cells for 24 hours after GATA1 restoration and in the presence of erythropoietin. (*The mouse beta-globin locus consists from the following genes from left to right: the Hbb-y gene encodes the embryonic beta-globin-like y protein (Hansen et al., 1982), the Hbb-bh1 gene encodes the minor early embryonic beta-globin-like z protein (Farace et al., 1984), the Hbb-bh2 and Hbb-bh3 are pseudogenes (Phillips et al., 1984, Hutchison et al., 1984), the Hbb-b1 encode the adult beta-globin major chain, the Hbb-b2*

encode the adult beta-globin minor chain (Konkel et al., 1979). Similar to the LCR in the human beta-globin locus, the mouse beta-globin locus consists of 6 hypersensitive regions termed: HS1 (Shehee et al., 1989), HS2 (Moon and Ley, 1990), HS3, HS4 (Jimenez et al., 1992), HS5 , HS6)

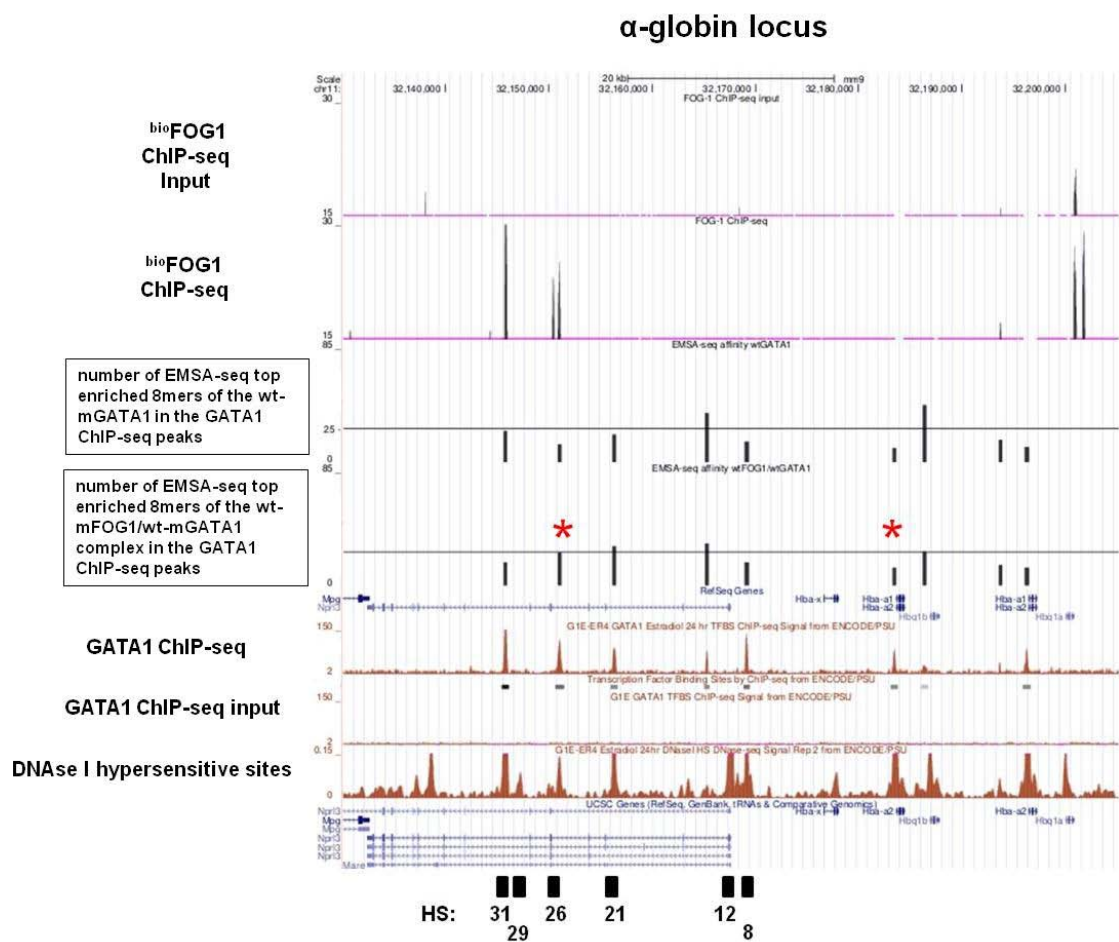


Figure 71. Profile of the alpha-globin locus (chr11:32.130.000-32.210.000). The rows in the picture are the ones presented in Figure 70.

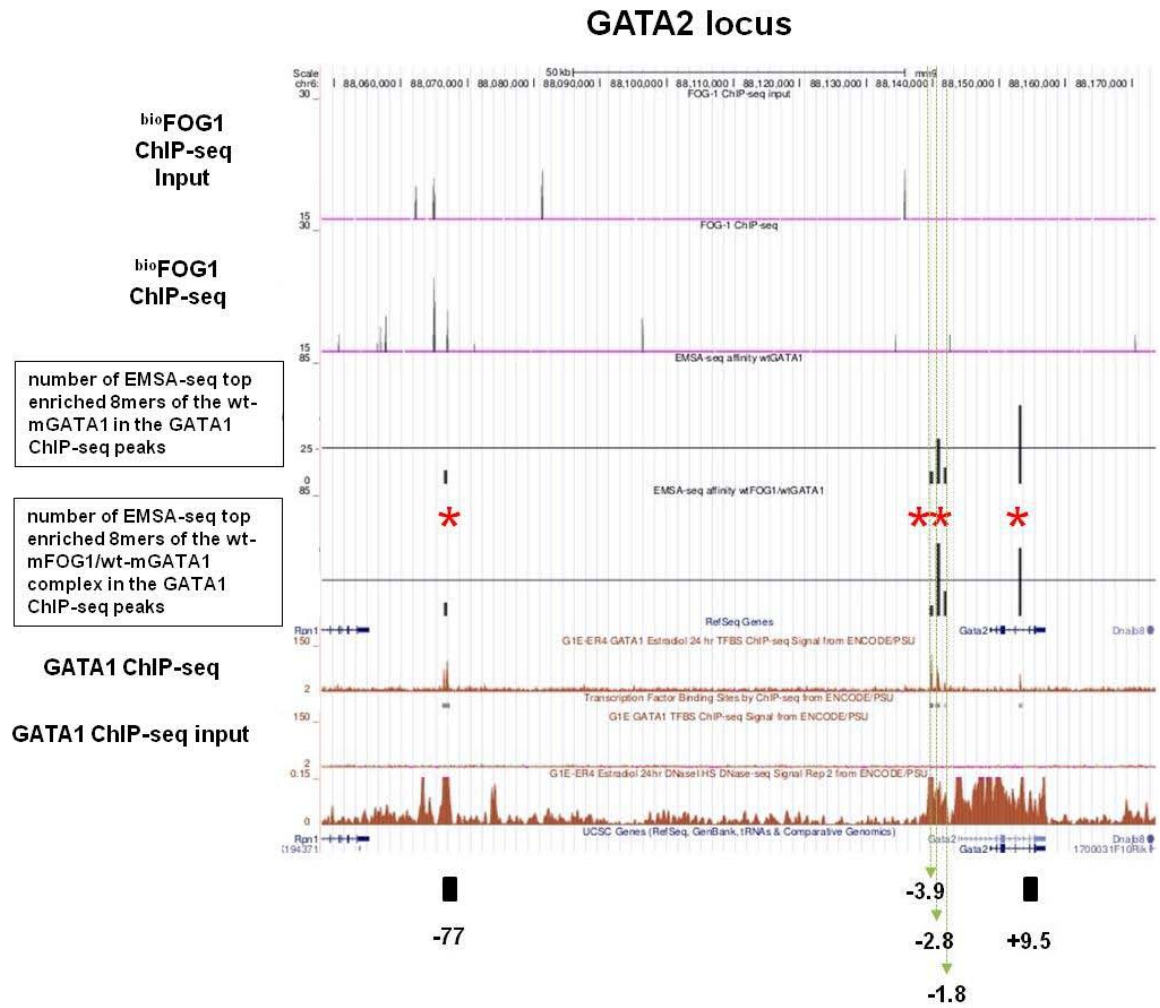


Figure 72. Profile of the GATA2 locus (chr6:88.050.000-88.170.000). The rows in the picture are the ones presented in Figure 70.

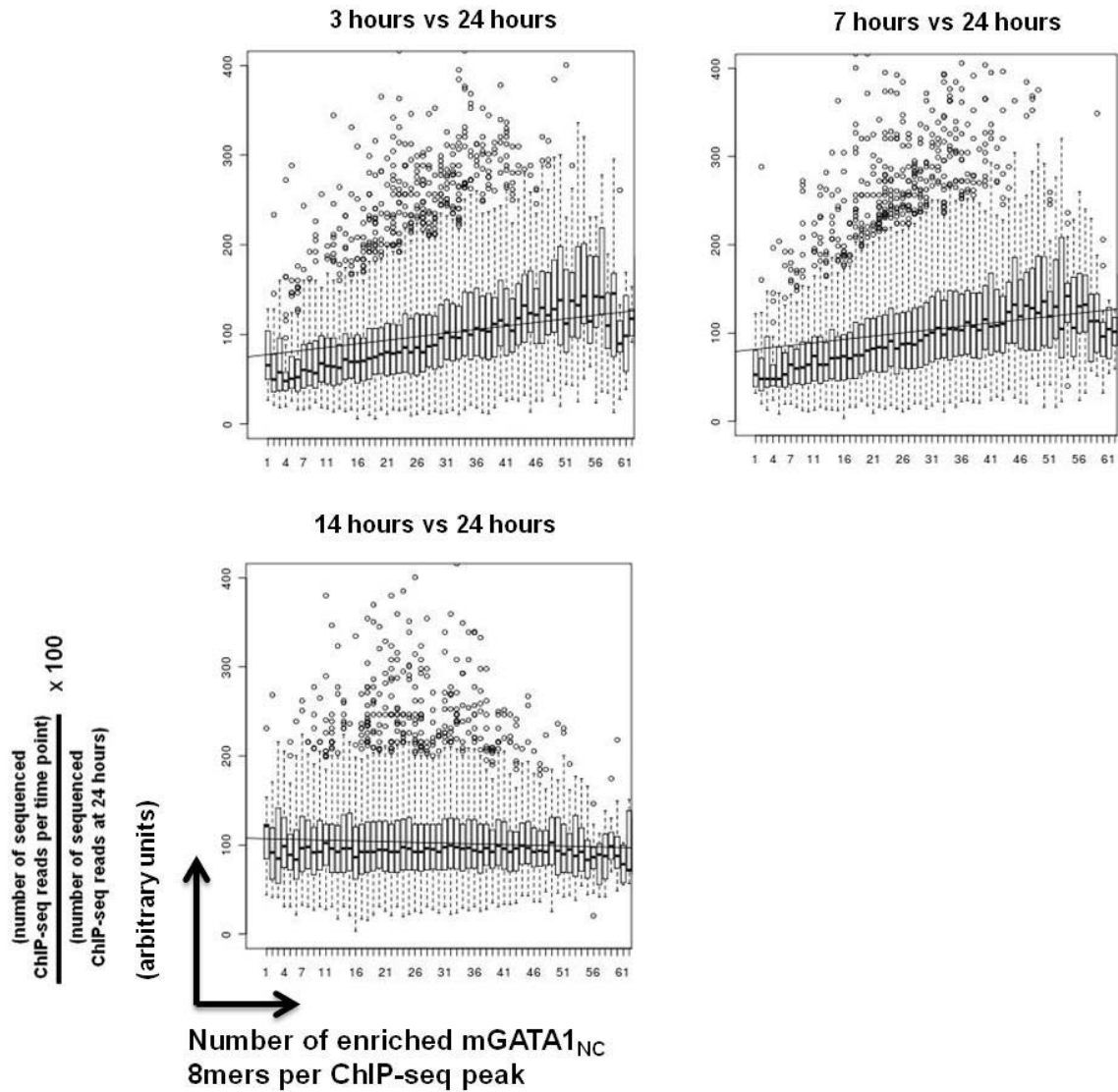


Figure 73. Comparison of the GATA1 ChIP-seq signal upon GATA1 restoration in a GATA1 null mouse cell line (G1E). The pictures represent four time points (3h, 7h, 14h, 24h after GATA1 restoration) of a 24h time-course. In each graph the ChIP-seq signal of the indicated time-points was compared with the ChIP-seq signal at 24 hours. The y axis represents per ChIP-seq peak the ratio of the number of sequenced reads per time point versus the number of sequenced reads at 24 hours. Because we expect the 24 hour ChIP-seq signal to have per position more reads than the 3h, 7h or 14 hours one the y axis represents arbitrary units. Each ChIP-seq peak was scored with the 2122 enriched 8mers of mGATA1_{NC}.

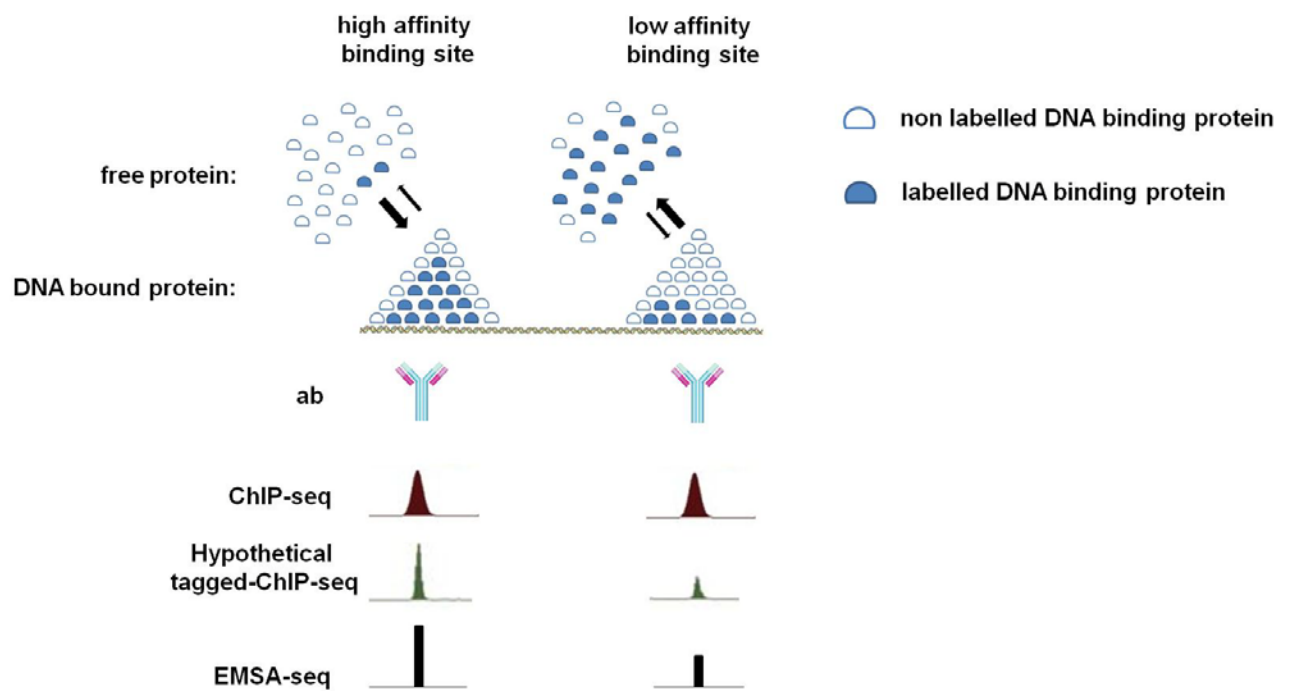


Figure 74. Working model hypothesis of the relationship between ChIP-seq and EMSA-seq. We assume that we have two ChIP-seq positions where the chromatin can be immunoprecipitated equally well. These two positions have different DNA binding affinities (we will call them a low DNA binding affinity position and a high DNA binding affinity position). In a single cell, we will find in the nucleus two copies of each specific position. We also assume that in the nucleus we have a significant amount of our transcription factor. This implies that the binding of the transcription factor in each specific position is near saturation. So independently of the strength of binding of each position, the positions will always be bound from the transcription factor. The number of molecules from the same transcription factor that can be bound in each position will depend only on the steric hindrance of nearby binding positions which has to do with the organisation of the binding sites on the sequence (motif grammar (Spitz and Furlong, 2012)). For example, if 10 molecules of the transcription factor can bind in both positions then the ChIP-seq signal from these two positions will be of the same strength. On the contrary, the difference in the DNA binding affinity of these two positions will affect the DNA binding turnover of the protein in these two positions. Low affinity binding sites will have a high DNA binding

turnover. High affinity binding sites will have a low DNA binding turnover. This practically means the following: If at $t_0=0$ we have 10 protein molecules in the low DNA binding affinity site and 10 protein molecules in the high DNA binding affinity site then at $t_1>t_0$ the high DNA binding affinity position will have retained the same 10 protein molecules whereas the low DNA binding affinity position will have 10 different molecules. We can argue that if each binding event involves considerable structural changes and protein-protein interactions we can hypothesize that in the low DNA binding affinity sites the rate of occurrence of a potential enzymatic event, from a GATA1 co-factor, will be slower than at the high DNA binding affinity sites where the reaction is uninterrupted. On the other hand, we can also argue that the high binding turnover in the low affinity sites, versus the high affinity sites, replenishes the enzyme needed on the specific locus. Assuming that the more time the same enzyme stays on the locus, the lower is its processivity, the quick replenishment keeps the rate of the enzymatic reaction high. Some studies have shown that putting a tag on the protein and permit it to compete the endogenous protein can give indication of the binding turnover (Competition ChIP (Lickwar et al., 2012)) (the hypothetical tagged-ChIP-seq row on the image). This technique although laborious is still limited from the resolution of the ChIP-seq technique which does not have the linearity of the EMSA-seq. An alternative approach is to have a cell line that lacks the protein of interest something that is not always feasible. Upon reconstitution of the examined protein someone can follow the recruitment in specific genomic positions with a ChIP-seq time course study as was used here for GATA1 (see main text, section 3.5.2.1).

3.5.3 EMSA-seq binding affinity distributions of ChIP-seq peaks can be used to characterize the similarity between ChIP-seq experiments of the same protein under different conditions

At first we checked how similar are the GATA1 ChIP-seq peak affinity distributions of the available datasets for the different models (Figure 75). We assigned the binding affinity of each ChIP-seq peak as the number of top 2122 enriched mGATA1_{NC} 8mers mapped on each one of the GATA1 peaks. We only found a human CD36+ dataset from (Su et al., 2013) that showed a skewed distribution in comparison with similar datasets from human CD36+ pro-erythroblasts cells. On the other hand independently of the differentiation state (TER119- versus TER119+), the model or the species (human versus mouse) the GATA1 affinity distributions are similar. Thus we suggest that the EMSA-seq binding affinity distributions of ChIP-seq peaks can be used as a quality control of ChIP-seq datasets. We also excluded the human CD36+ dataset that showed the skewed distribution from further analysis.

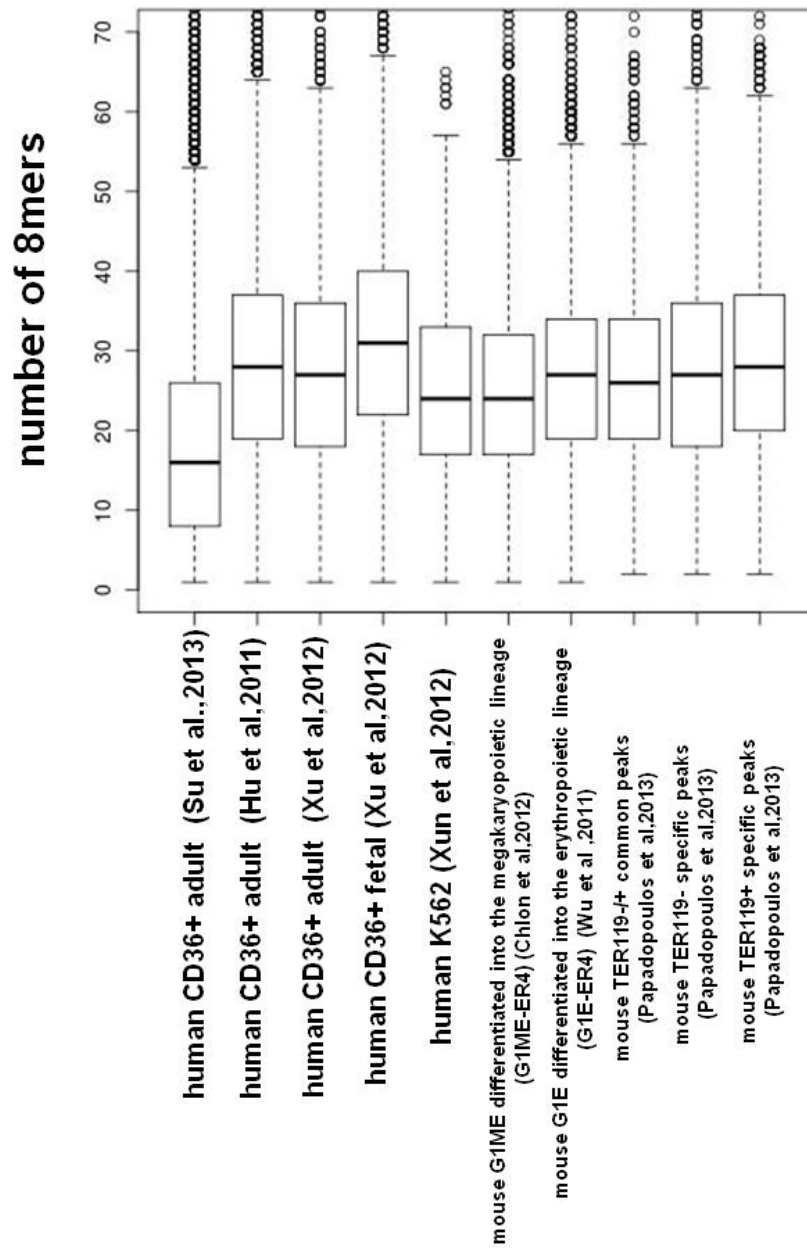


Figure 75. DNA binding affinity distributions of the GATA1 ChIP-seq peaks from different published studies. The affinity is expressed as the number of 8mers per ChIP-seq peak from the 2122 enriched mGATA1_{NC} 8mers. The experiments cover different models of erythropoiesis and megakaryopoiesis.

3.5.4 GATA1 low affinity ChIP-seq peaks are closer to the TSS. KLF1 does not show this pattern

We assigned the GATA1 ChIP-seq peaks from human peripheral blood-derived CD36+ pro-erythroblasts to the nearest TSS (10kb upstream and downstream from the TSS).

From 29952 GATA1 peaks in CD36+ differentiated cells we assigned 9929 (33%) peaks in the nearest TSS. Of these 4760 (47% of 9929) peaks lie upstream of the TSS and 5169 (53% of 9929) lie downstream of the TSS. Overall the peaks upstream of the TSS are associated with 3918 genes and the peaks downstream of the TSS are associated with 4203 genes.

Similarly for the KLF1 peaks in the CD36+ cell line we assigned 20394 (out of 56027) peaks in the nearest TSS. Of these 8238 (40% of 20394) peaks lie upstream of the TSS and 12156 (60% of 20394) lie downstream of the TSS. Overall the peaks upstream of the TSS are associated with 6585 genes and the peaks downstream of the TSS are associated with 9551 genes.

We checked the distribution of ChIP-seq peak binding affinities (**EMSA-seq binding affinity Kd scores**) near the TSS (10kb upstream or downstream of the TSS) for both GATA1 and KLF1. In Figure 76 we see that the low binding affinity GATA1 ChIP-seq peaks are more frequent near the TSS. By performing the same analysis on the KLF1 ChIP-seq peaks we do not see any spatial restriction of the ChIP-seq peak binding affinities. In this case high and low affinity ChIP-seq peaks are equally distributed near and distant from TSS.

We then checked whether there might exist a functional role for this differential presence of the GATA1 ChIP-seq peak binding affinities by assigning GATA1 ChIP-seq peaks in enhancer regions and examining their binding affinity distributions.

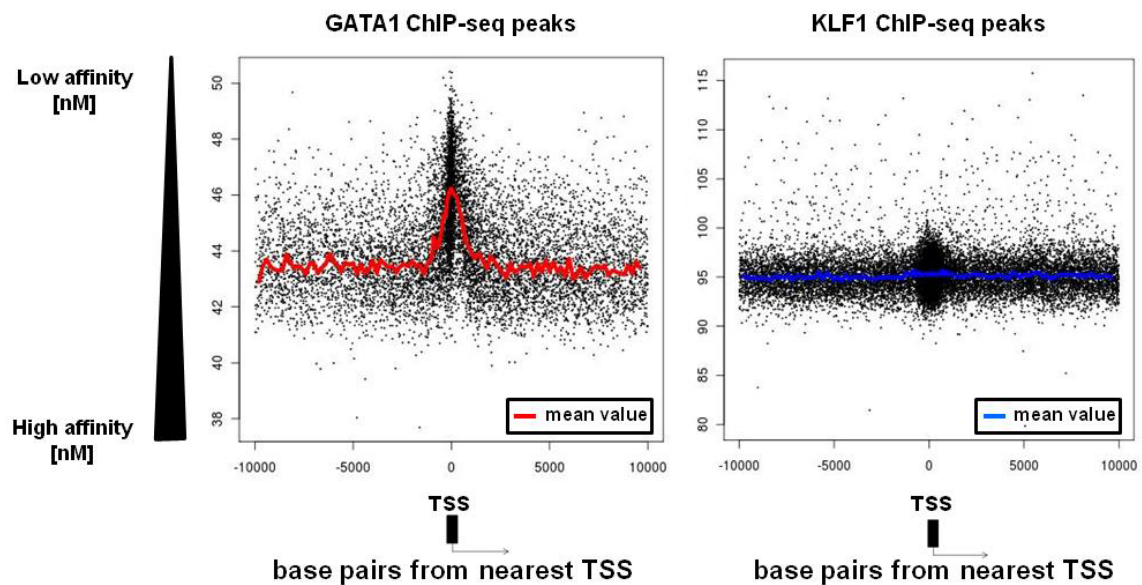


Figure 76. A) Scatter plot of the relationship of the GATA1 ChIP-seq peak DNA binding affinities with the GATA1 ChIP-seq peak distance from the nearest TSS in the CD36+ cells. The region +/- 10 kb from the TSS is shown. The red line represents the mean value of the GATA1 ChIP-seq peak DNA binding affinities in a 100 bp window. B) Scatter plot of the relationship of the KLF1 ChIP-seq peak DNA binding affinities with the KLF1 ChIP-seq peak distance from the nearest TSS in the CD36+ cells. The region +/- 10 kb from the TSS is shown. The blue line represents the mean value of the KLF1 ChIP-seq peak DNA binding affinities in a 100 bp window. The GATA1 ChIP-seq peaks were from (Xu et al., 2012) (GEO accession number: GSM970258) (cell type: proerythroblasts (ProEs) from adult human bone marrow CD34+ cells). The KLF1 ChIP-seq peaks were from (Su et al., 2013) (GEO accession number: GSM1067275) (cell type: R3/R4 erythroid cells derived from human bone marrow CD34+ cells)

3.5.5 GATA1 high affinity ChIP-seq peaks are present in enhancers whereas GATA1 low affinity ChIP-seq peaks are present in promoters. KLF1 high/low affinity ChIP-seq peaks are equally probable to appear in enhancers or promoters.

We used the ChromHMM (Ernst and Kellis, 2012) to segment the genome of the CD36+ erythropoiesis model in 6 states based on a number of epigenetic marks (Wu et al., 2011). Each of these chromatin states represents a combination of four chromatin modifications (H3K4me1, H3K4me3, H3K27me3, H3K9me3). Each GATA1 ChIP-seq peak was assigned in one of these 6 chromatin states. In more detail, state 2 is indicative of enhancers (enriched H3K4me1), state 1 is indicative of promoters (enriched H3K4me1 and H3K4me3), state 3 is indicative of bivalent marks (enriched H3K4me3 and H3K27me3) and state 4 (enriched H3K27me3) and state 5 (enriched H3K9me3) are indicative of repressed marks (Zhou et al., 2011). We then focused in states 1 and 2. In Figure 77 we notice that high affinity GATA1 ChIP-seq peaks appear in enhancers (state 2) whereas low affinity GATA1 ChIP-seq peaks appear in promoters (state1).

We also assigned the KLF1 ChIP-seq peaks in the same chromatin states but we did not notice any discriminatory presence of high or low binding affinity KLF1 ChIP-seq peaks in enhancers and promoters (Figure 77).

The same results were taken when we used the segmentation of the G1E-ER4 erythropoiesis model genome in 6 chromatin states based on a combination of the same four chromatin modifications (Wu et al., 2011) (data not shown).

In Figure 78 we present a model for the discriminatory preference of the high/medium/low GATA1 or KLF1 DNA binding affinities to appear in enhancers/distal promoter elements and proximal promoter elements.

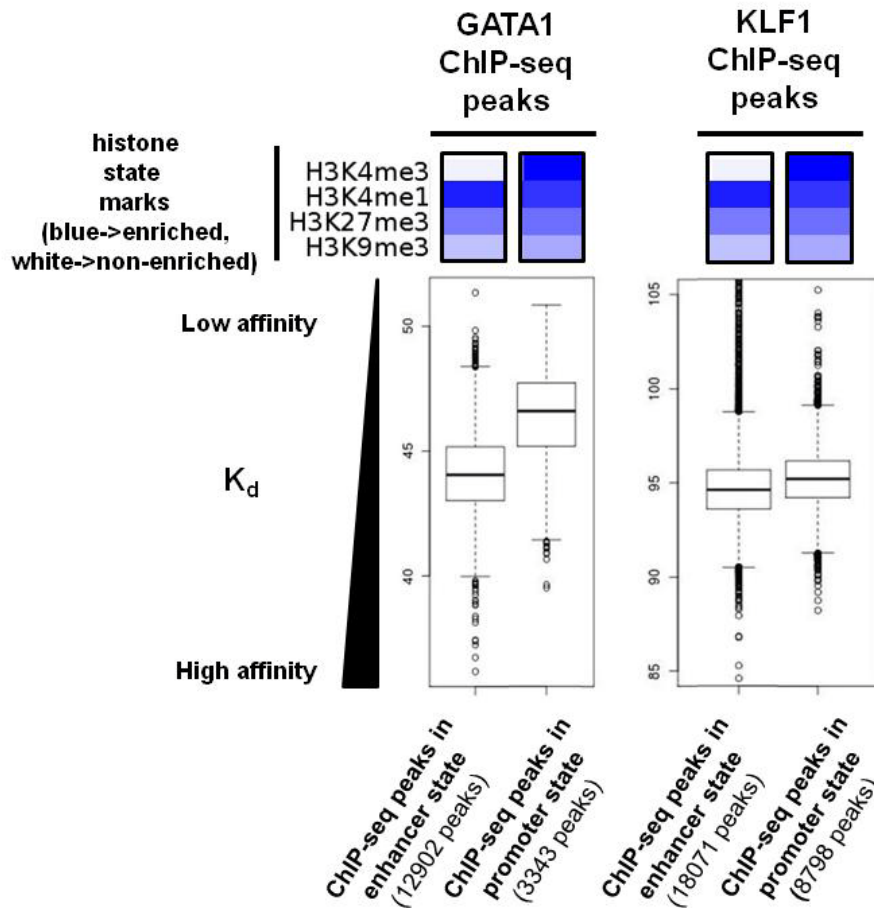


Figure 77. Box plots with the distribution of the GATA1 and KLF1 ChIP-seq peak binding affinities in enhancers and promoters for the human CD36+ pro-erythroblasts. Four histone modifications (H3K4me3, H3K4me1, H3K27me3, H3K9me3) were used to segment the genome of the CD36+ pro-erythroblasts in enhancer-like histone state (H3K4me1 high, H3K4me3 low) and promoter-like histone state (H3K4me1 high, H3K4me3 high). The GATA1 and KLF1 ChIP-seq peaks were the same as the ones in Figure 76. The H3K4me3, H3K4me1, H3K27me3, H3K9me3 were from (Cui et al., 2009) (GEO accession number: GSM317597, GSM317596, GSM317594, GSM317599 respectively).

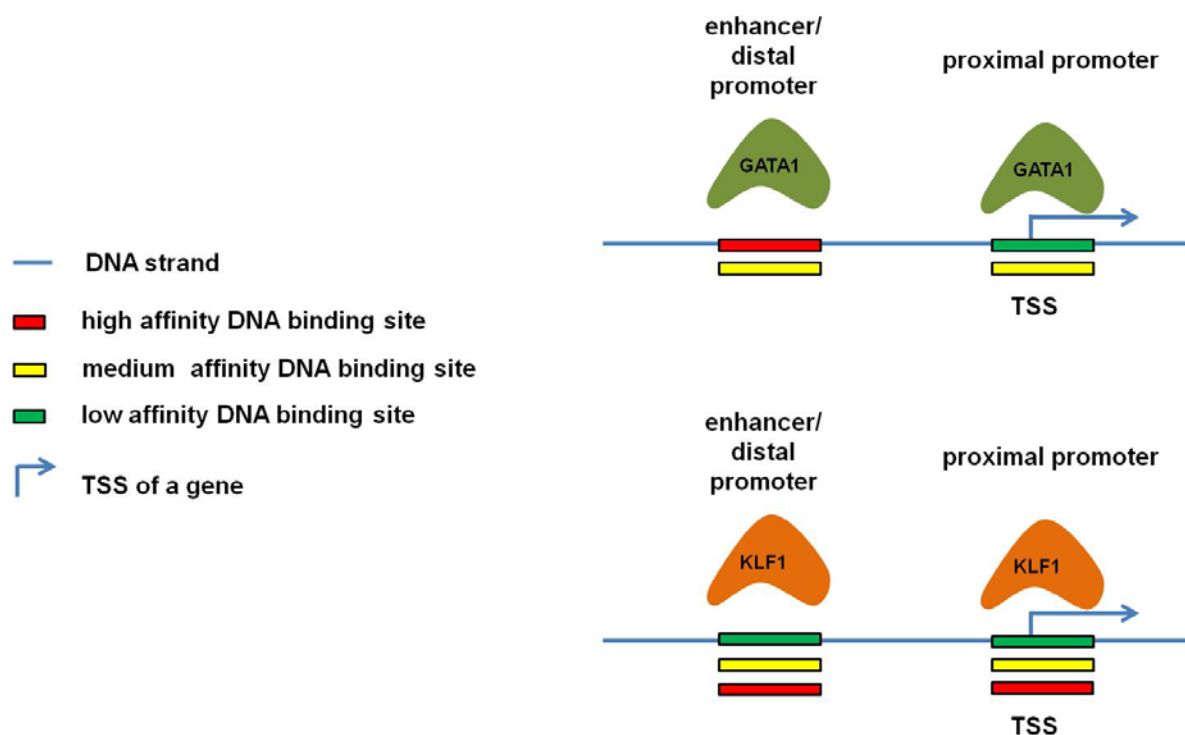


Figure 78. Binding models of GATA1 and KLF1 on enhancers and proximal/distal promoter elements. High affinity GATA1 DNA binding sites are more frequently found away from the TSS (enhancers/distal promoter elements) whereas low affinity GATA1 DNA binding sites are more frequently found near the TSS (proximal promoter elements). The KLF1 DNA binding sites show no specific preference. We define as distal promoter elements binding sites less than 10 kb and more than 1 kb upstream or downstream from the closest TSS. We also define as proximal promoter elements binding sites less than 1 kb upstream or downstream from the closest TSS.

3.5.6 GATA1 high affinity ChIP-seq peaks are present in regions of low nucleosome occupancy and of reduced DNase I hypersensitivity whereas GATA1 low affinity ChIP-seq peaks are present in regions of high nucleosome occupancy and of increased DNase I hypersensitivity. KLF1 high/low affinity ChIP-seq peaks are equally probable to appear in regions of high/low nucleosome occupancy and of high/low DNase I hypersensitivity

We used nucleosome occupancy data from the CD34⁺ progenitor/CD36⁺ mature erythroblasts model to assess whether the high affinity ChIP-seq peaks might compete the nucleosome occupancy. To avoid the high concentration of nucleosomes near the TSS we used peaks that are more than 10kb upstream or downstream from the closest TSS (Figure 79). Because we have GATA1 and KLF1 ChIP-seq peaks only from the CD36⁺ cells, we also examined the nucleosome pattern in the CD34⁺ cells in order to check whether the observed pattern pre-exists in the progenitors or it is acquired only in the differentiated cells.

In Figure 79 we notice an association between high affinity peaks and low nucleosome occupancy levels. More specifically, we observed that the higher it is the affinity of the ChIP-seq peaks, the lower it is the nucleosome occupancy on the ChIP-seq peaks. Assuming that we are interrogating binding events in the whole cell population, the previous observation indicates that high affinity GATA1 DNA binding sites are usually depleted of the local nucleosomes in each cell of the whole cell population whereas in the low affinity GATA1 DNA binding sites, the local nucleosomes are usually present in the majority of cells in the whole cell population (Figure 86). We assume that the nucleosome eviction is caused from the SWI/SNF chromatin remodelling complex that interacts with GATA1 (Kadam and Emerson, 2003). Similar graph for the KLF1 did show a similar trend but not as strong as the GATA1. The majority of the KLF1 ChIP-seq peak binding affinities are very similar.

Because the association was true for both the CD34⁺ progenitors and the CD36⁺ differentiated cells (Figure 79), we can also hypothesize that the protein prefers high affinity binding sites to be in lower nucleosome occupancy regions. This implies that the nucleosomes must first be remodelled from a GATA1 independent mechanism and then GATA1 is recruited. We caution here that we do not have GATA1 or KLF1 ChIP-seq peaks for the CD34⁺ progenitors and it is possible that a lot of the CD36⁺ GATA1 or KLF1 ChIP-seq positions do pre-exist in the CD34⁺ cells. It is also possible that GATA2 might exert the same effect on the nucleosome occupancy and it might bind in the same positions as GATA1 in the CD34⁺ progenitors. Although the previous imply that the local nucleosomes in the CD36⁺ GATA1 binding sites are already remodelled in the CD34⁺ progenitors cells from a pre-existing GATA1 or GATA2 complex or from another protein complex, we do not favour a model where GATA1 binds high affinity sites in regions with pre-existing low nucleosome occupancy but rather a model where the GATA1 presence and only is adequate to remodel the local nucleosomes. Regions with already remodelled nucleosomes might permit GATA1 to bind more effectively but it is not a prerequisite.

We also used DNase I hypersensitivity data from the TER119⁻/TER119⁺ erythropoiesis model. We used data available from the ENCODE project. Because there were no replicates in each dataset, we used datasets from two different populations of the TER119⁻ cells and two different populations of the TER119⁺ cells. These datasets represent cells in different stages of the maturation process as judged from the presence of the CD117 marker (c-Kit receptor) and the CD71 marker (transferring receptor). The maturation process starts with the CD117⁺, CD71⁻, TER119⁻ population that matures into the CD117⁺, CD71⁺, TER119⁻ population (Figure 80, Figure 82). The last population

matures into CD117+, CD71+, TER119+ cells and eventually into CD117-, CD71+, TER119+ cells (Figure 81, Figure 83).

We again used peaks that are more than 10 kb upstream or downstream from the closest TSS. We notice that the high affinity binding sites are associated with regions of low DNase-seq reads in both the TER199- and TER199+ cells (Figure 80, Figure 81). **This implies that these regions, although they have low nucleosome content, as we have seen earlier, are less sensitive to DNase I because GATA1 and its associated co-factors stay longer in these regions.** Conceptually this is similar with the digital DNase I footprinting (Hesselberth et al., 2009). We can also imagine that the lower presence of nucleosomes, that compete with the TFs for the DNA binding, might facilitate the stochastic formation of large GATA1 multiprotein complexes (Figure 86). KLF1 did not show this pattern (Figure 82, Figure 83).

We then used all the available DNase I hypersensitive regions of the cell, independently whether they are bound from GATA1 or KLF1, and we assigned a DNA binding affinity score (number of enriched mGATA1_{NC} 8mers or enriched mKLF1₂₅₇₋₃₅₈ 8mers) for both the GATA1 and the KLF1 protein (Figure 84). We examined whether GATA1 and KLF1 binds to low, medium and high affinity DNase I hypersensitive regions equally well or whether they prefer to bind in low or high affinity DNase I hypersensitive regions exclusively. In Figure 84C, D we see that GATA1 prefers to bind more frequently in low and medium affinity DNase I hypersensitive regions whereas KLF1 prefers to bind more frequently in low affinity DNase I hypersensitive regions. The fact that in both cases the high affinity DNase I hypersensitive regions are not preferred indicates one of the following:

- i) High affinity binding sites are avoided because the system can be regulated better when the proteins have a high DNA binding turnover.

- ii) Complexation of the protein with other co-factors forces it to bind in regions that can accommodate motifs for multiple proteins. This will lead in a lower occurrence of GATA1 or KLF1 motifs.

In Figure 85 we also present the DNA binding affinity distributions of the positions where the complex wt-mGATA1/wt-mFOG1 binds in comparison with the positions where wt-mGATA1 binds alone. We see that when wt-mGATA1 cannot interact with FOG1 (the mGATA1^{V205G} mutant protein), it goes and binds in positions with high binding affinity (high number of mGATA1_{NC} 8mers). On the contrary, when the wt-mGATA1 interacts with FOG1, it goes and binds exclusively in regions with low DNA binding affinity. Because GATA1 in the cell is probably always bound to a co-factor, **the previous observation might indicate that the system tries to avoid GATA1 to bind in high affinity sites.**

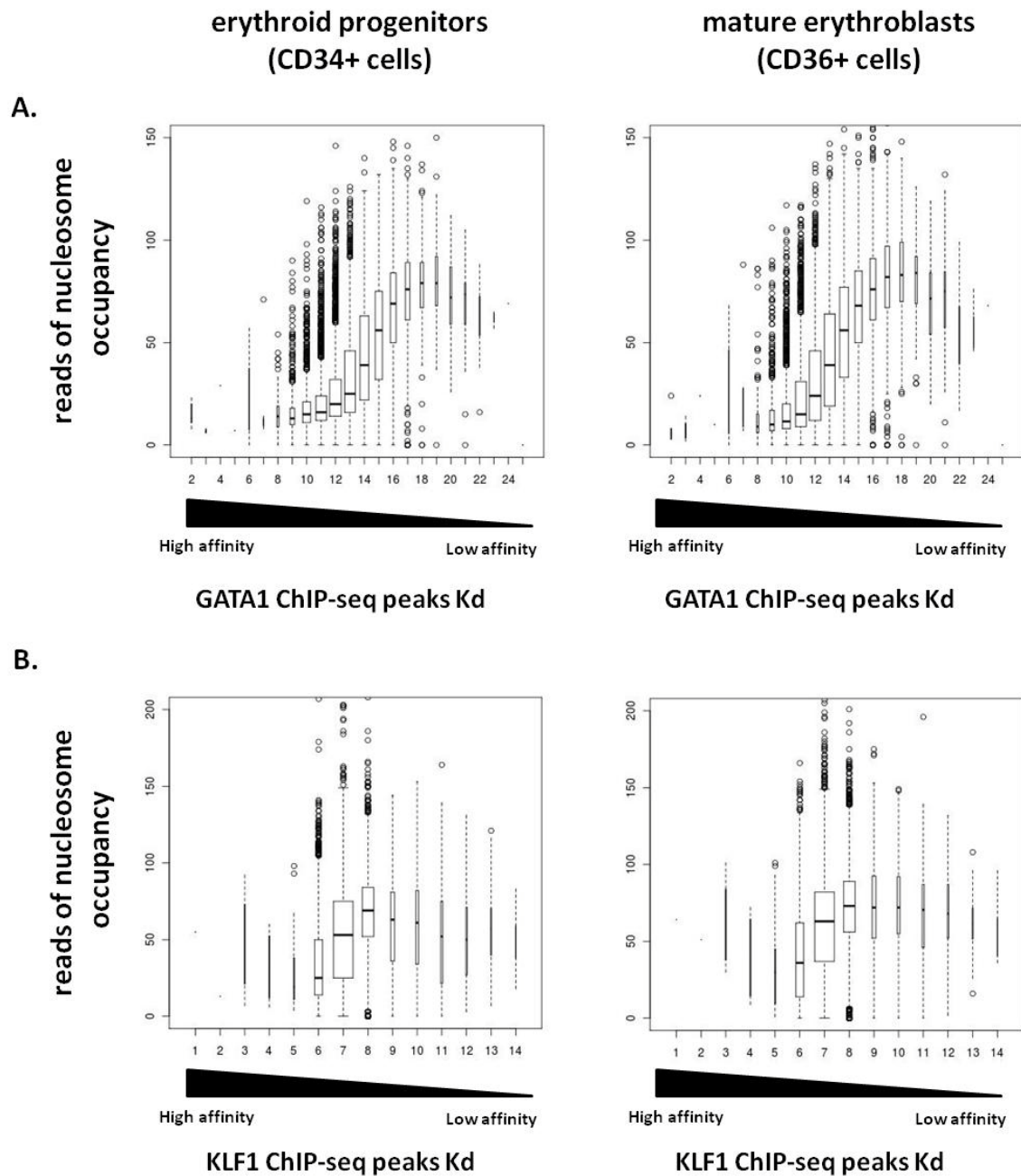


Figure 79. Nucleosome occupancy of the GATA1 ChIP-seq peaks (A) and of the KLF1 ChIP-seq peaks (B) in the CD34+ undifferentiated and CD36+ differentiated cells. Only GATA1 and KLF1 peaks that are more than 10 kb away, upstream or downstream, from the nearest TSS are presented. We used the EMSA-seq binding affinity Kd score to characterise the binding affinity of the GATA1 and KLF1 ChIP-seq peaks (x-axis). The ChIP-seq peaks were binned in 25 categories for GATA1 and 14 categories for KLF1 based on their EMSA-seq binding affinity Kd scores. The width of the box plot in each category is proportional to the number of ChIP-seq peaks in each category. The y axis

represents the number of nucleosome occupancy reads per ChIP-seq peak. The GATA1 ChIP-seq data, CD34+ nucleosome data and CD36+ nucleosome data were from (Hu et al., 2011) (GEO accession number: GSM651547, GSM651559, GSM651558 respectively) (cell type: human CD36+ cells derived from differentiated CD34+/CD113+ HSC cells). The KLF1 ChIP-seq peaks were from (Su et al., 2013) (GEO accession number: GSM1067275) (cell type: R3/R4 erythroid cells derived from human bone marrow CD34+ cells).

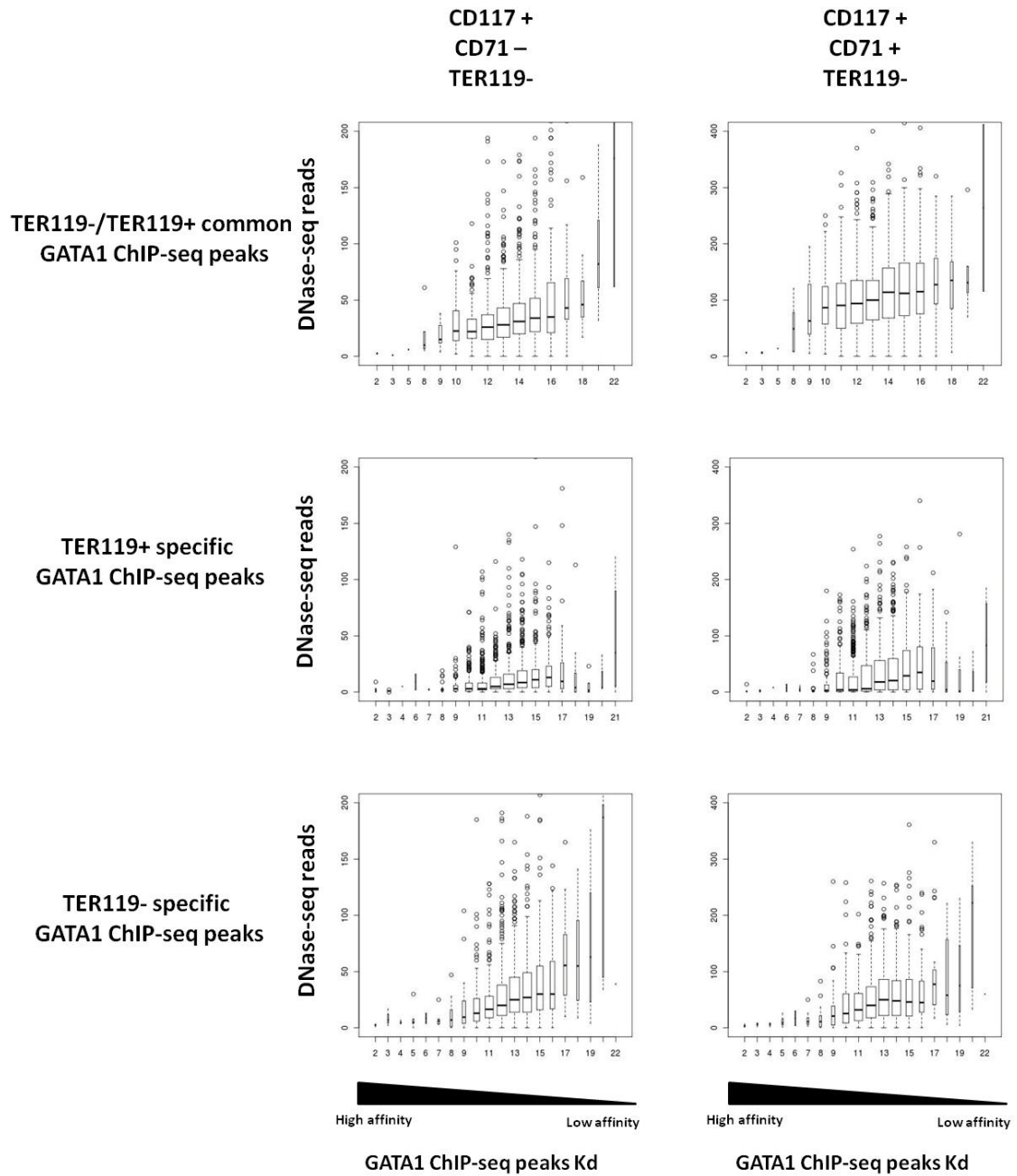


Figure 80. DNase I hypersensitivity of the GATA1 ChIP-seq peaks from two different populations of the TER119- cells (c-Kit receptor (CD117) present and transferring receptor (CD71) absent or present). Only GATA1 ChIP-seq peaks that are more than 10 kb away, upstream or downstream, from the nearest TSS are presented. We used the EMSA-seq binding affinity Kd score to characterise the binding affinity of the GATA1 ChIP-seq peaks (x- axis). The ChIP-seq peaks were binned in 21/22 categories based on their EMSA-seq binding affinity Kd scores. The width of the box plot in each category is proportional to the number of ChIP-seq peaks in each category. The y axis represents the number of the

DNase I hypersensitivity reads per ChIP-seq peak. The CD117+/CD71-/TER119- and CD117+/CD71+/TER119- DNase I hypersensitivity data were from the ENCODE/University of Washington project (GEO accession number: GSM1014155, GSM1014158 respectively). The GATA1 ChIP-seq peaks of the TER119- cells were from (Papadopoulos et al., 2013).

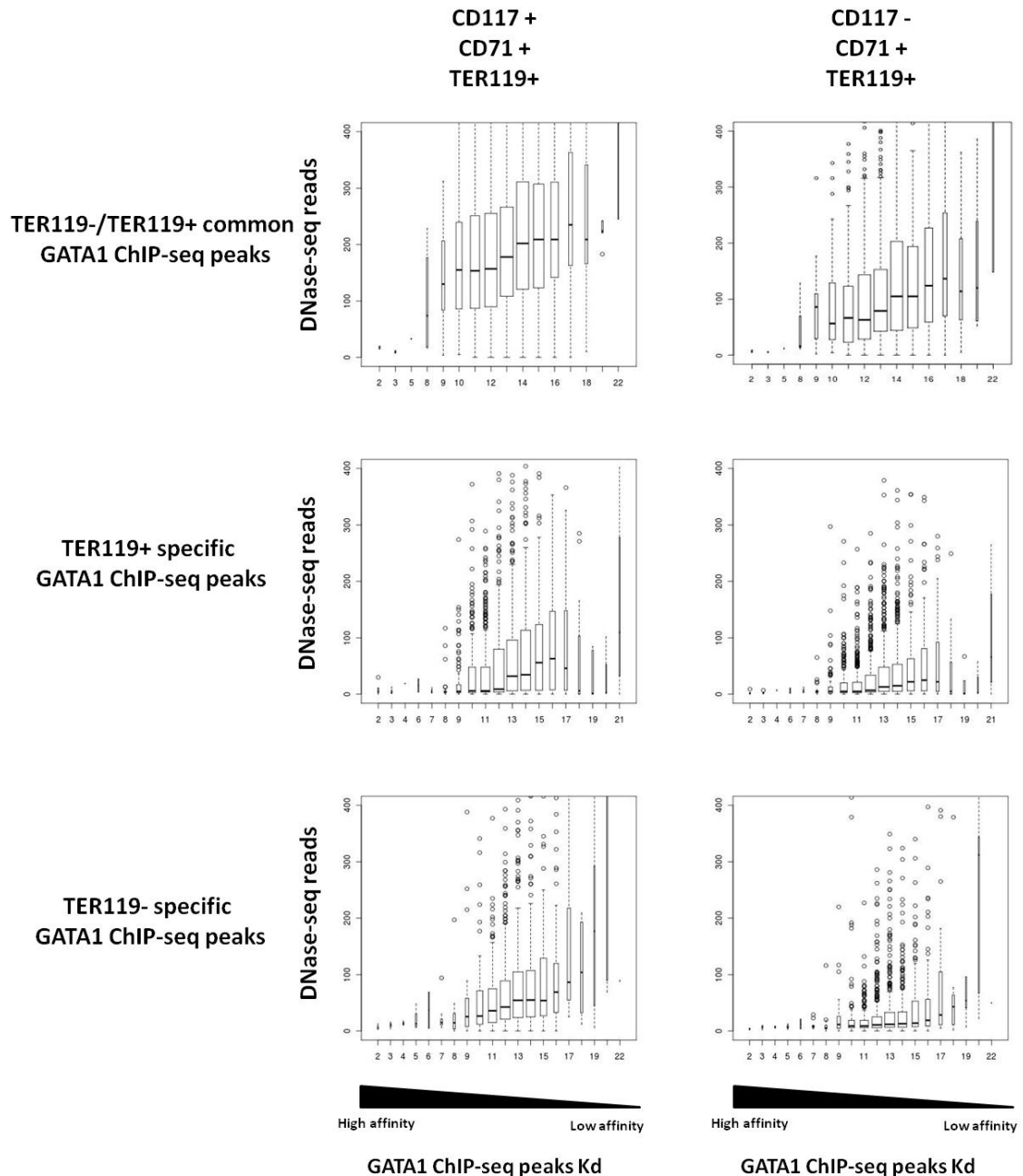


Figure 81. DNase I hypersensitivity of the GATA1 ChIP-seq peaks from two different populations of the TER119+ cells (c-Kit receptor (CD117) present or absent and transferring receptor (CD71) present). Only GATA1 ChIP-seq peaks that are more than 10 kb away, upstream or downstream, from the nearest TSS are presented. We used the EMSA-seq binding affinity Kd score to characterise the binding affinity of the GATA1 ChIP-seq peaks (x- axis). The ChIP-seq peaks were binned in 21/22 categories based on their EMSA-seq binding affinity Kd scores. The width of the box plot in each category is proportional to the number of ChIP-seq peaks in each category. The y axis represents the

number of the DNase I hypersensitivity reads per ChIP-seq peak. The CD117+/CD71+/TER119+ and CD117-/CD71+/TER119+ DNase I hypersensitivity data were from the ENCODE/University of Washington project (GEO accession number: GSM1014157, GSM1014156 respectively). The GATA1 ChIP-seq peaks of the TER119+ cells were from (Papadopoulos et al., 2013).

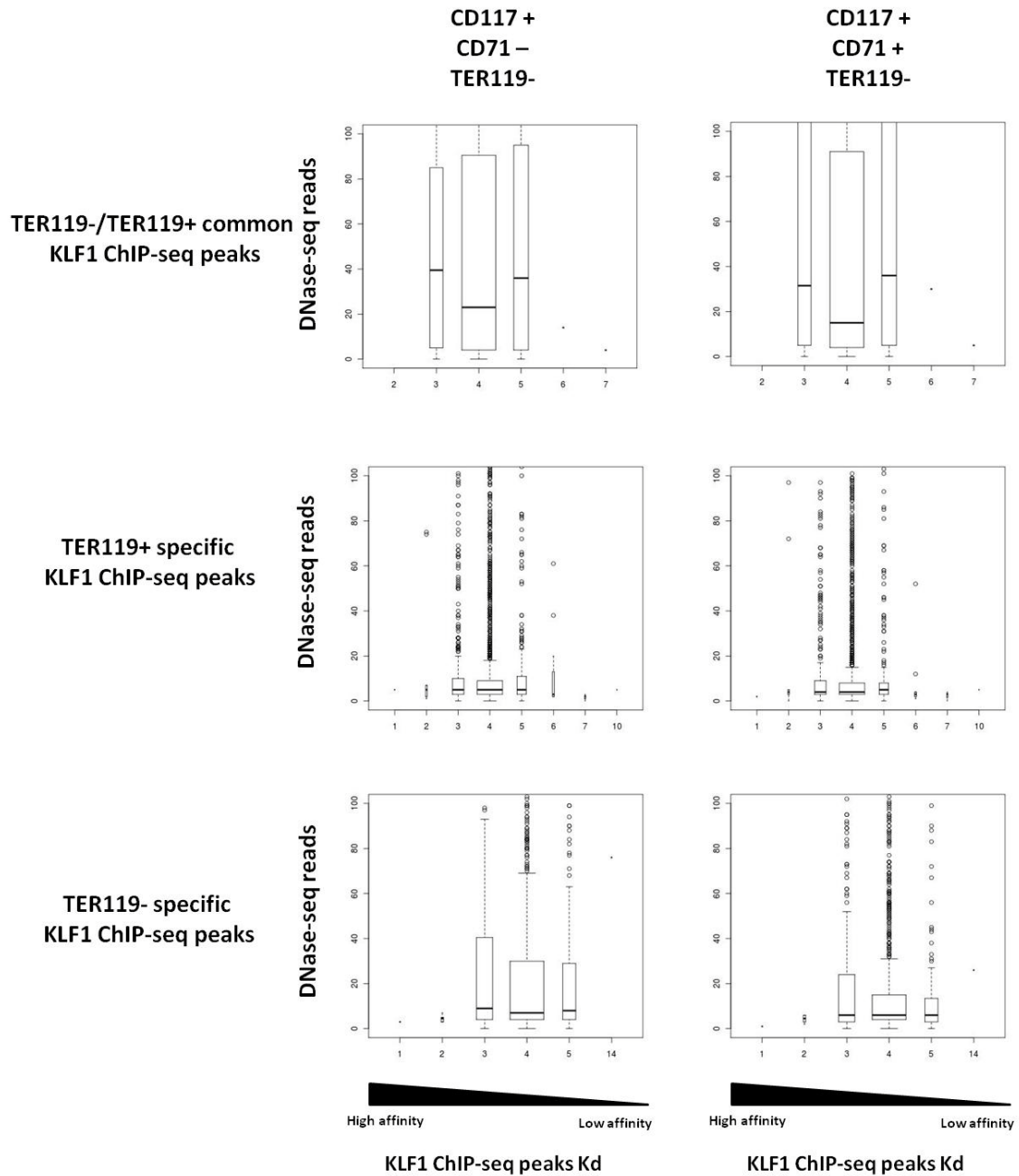


Figure 82. DNase I hypersensitivity of the KLF1 ChIP-seq peaks from two different populations of the TER119- cells (c-Kit receptor (CD117) present and transferring receptor (CD71) absent or present). Only KLF1 ChIP-seq peaks that are more than 10 kb away, upstream or downstream, from the nearest TSS are presented. We used the EMSA-seq binding affinity Kd score to characterise the binding affinity of the KLF1 ChIP-seq peaks (x- axis). The ChIP-seq peaks were binned in 7/10/14 categories based on their EMSA-seq binding affinity Kd scores. The width of the box plot in each category is proportional to the number of ChIP-seq peaks in each category. The y axis represents the number of

the DNase I hypersensitivity reads per ChIP-seq peak. The DNase I hypersensitivity data were similar with Figure 80. The KLF1 ChIP-seq peaks of the TER119- cells were from (Pilon et al., 2011) (GEO accession number: GSM1165492).

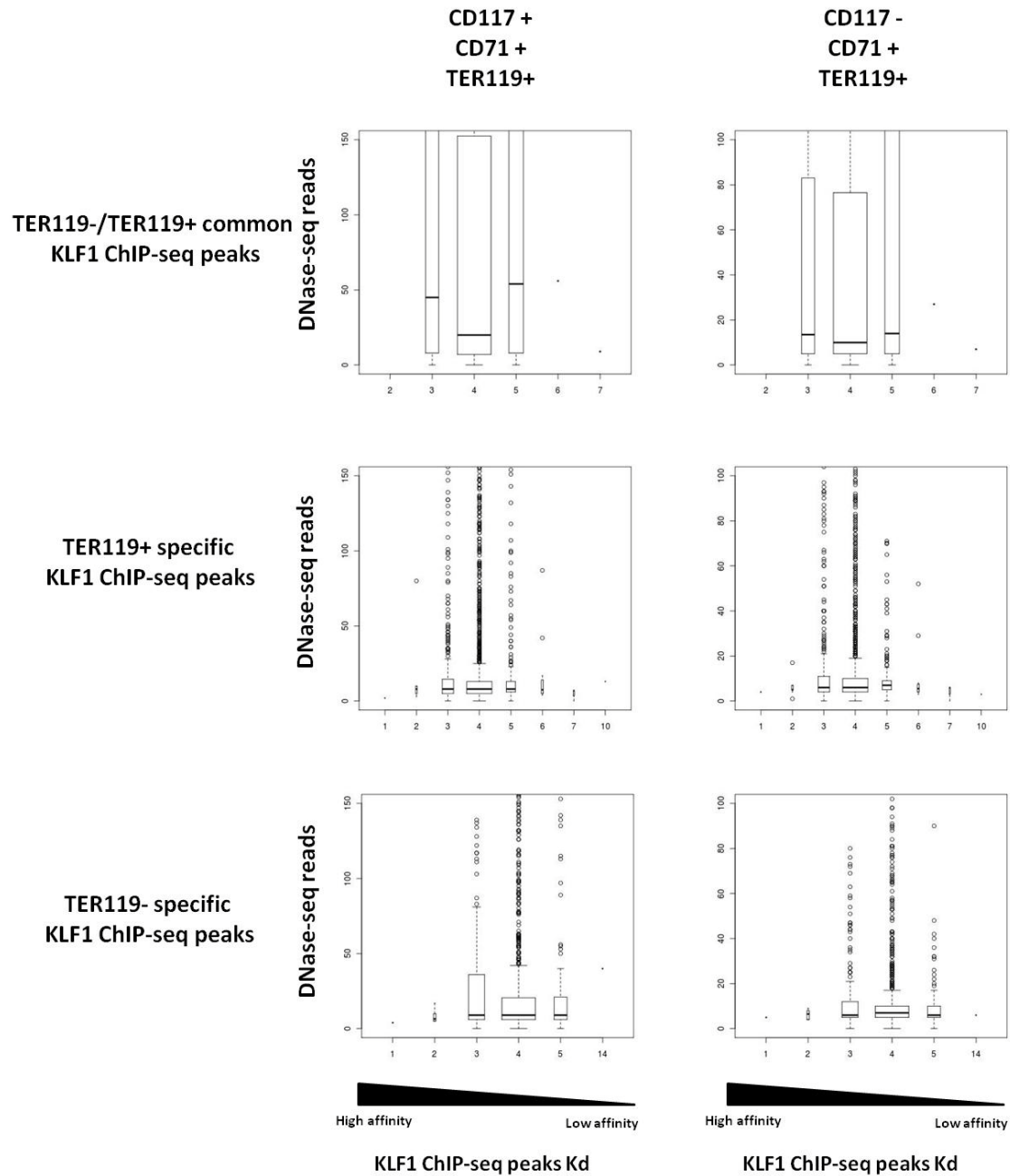


Figure 83. DNase I hypersensitivity of the KLF1 ChIP-seq peaks from two different populations of the TER119+ cells (c-Kit receptor (CD117) present or absent and transferring receptor (CD71) present). Only KLF1 ChIP-seq peaks that are more than 10 kb away, upstream or downstream, from the nearest TSS are presented. We used the EMSA-seq binding affinity Kd score to characterise the binding affinity of the KLF1 ChIP-seq peaks (x- axis). The ChIP-seq peaks were binned in 7/10/14 categories based on their EMSA-seq binding affinity Kd scores. The width of the box plot in each category is proportional to the number of ChIP-seq peaks in each category. The y axis represents the

number of the DNase I hypersensitivity reads per ChIP-seq peak. The DNase I hypersensitivity data were similar with Figure 81. The KLF1 ChIP-seq peaks of the TER119+ cells were from (Pilon et al., 2011) (GEO accession number: GSM1165493).

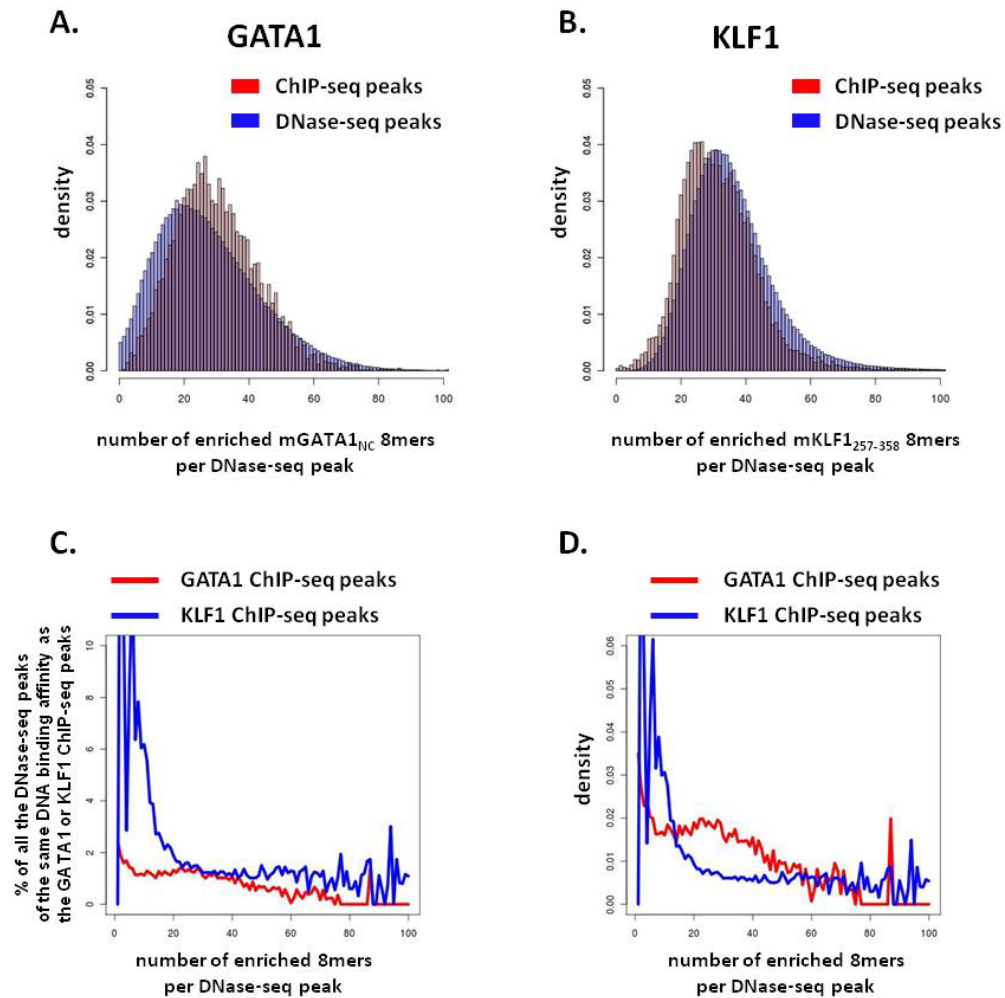


Figure 84. Relationship between the GATA1 or KLF1 DNA binding affinities of the DNase I hypersensitive sites in vivo (DNase-seq peaks) and the fraction of them that is bound by either GATA1 or KLF1 (ChIP-seq peaks). A) DNA binding affinity distributions for the GATA1 DNase-seq peaks and the GATA1 ChIP-seq peaks. The binding affinity of the peaks is expressed as the number of 8mers per peak from the 2122 enriched mGATA1_{NC} 8mers. B) Similar for KLF1. The binding affinity of the peaks is expressed as the number of 8mers per peak from the 193 enriched mKLF1₂₅₇₋₃₅₈ 8mers. C) Fraction of the DNase I hypersensitivity peaks, with the same DNA binding affinity, which are bound by GATA1 or KLF1. D) Same as (C) but the distributions are presented as density. More low rather than medium and high affinity DNase I hypersensitivity peaks, scored with the KLF1 8mers, are bound by KLF1 in vivo. On the contrary both low and medium rather than high affinity

DNase I hypersensitivity peaks, scored with the GATA1 8mers, are bound by GATA1 in vivo.

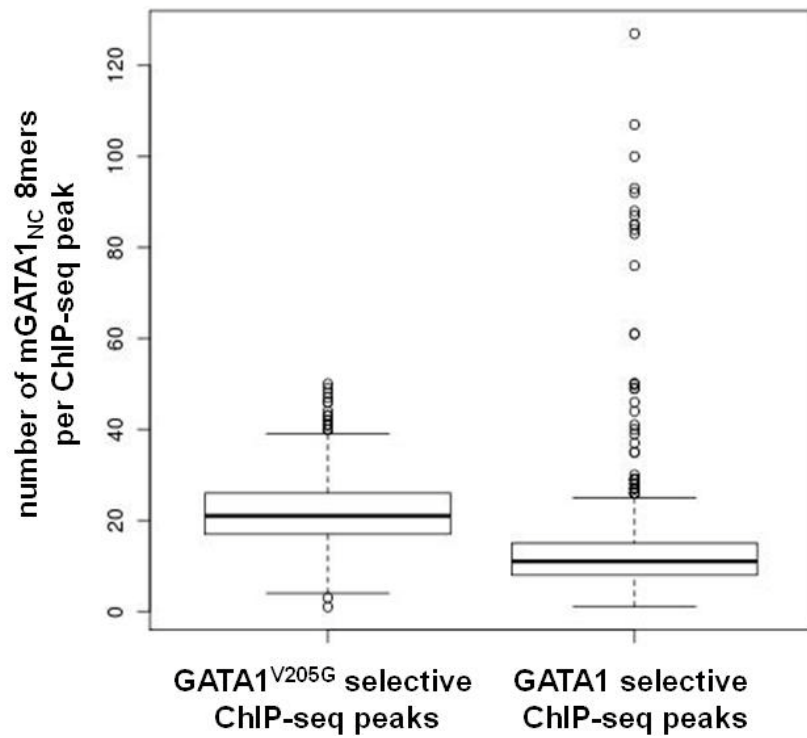


Figure 85. Affinity distributions of the GATA1 selective ChIP-seq peaks (fraction of peaks where only the wt-mGATA1/wt-mFOG1 is bound) and the GATA1^{V205G} selective ChIP-seq peaks (fraction of peaks where only the wt-mGATA1 is bound). The affinity is expressed as the number of 8mers per ChIP-seq peak from the 2122 enriched mGATA1_{NC} 8mers. The ChIP-seq groups are the same as the ones appearing in Figures 50 and 51.

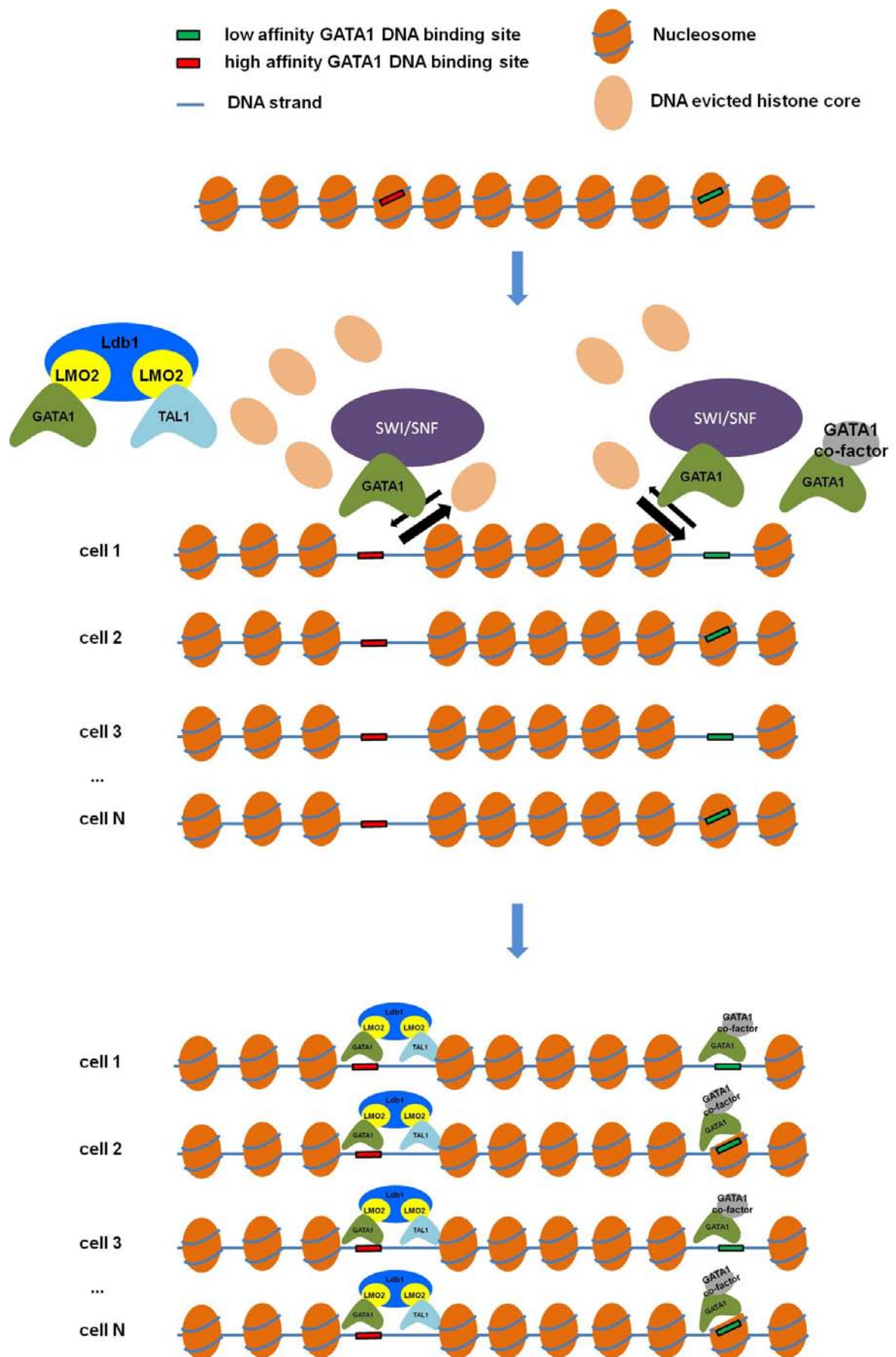


Figure 86 (previous page). Binding models of the GATA1 complexes on chromatin. High affinity GATA1 DNA binding sites are more probable to have their local nucleosome evicted simultaneously in the majority of cells from the whole cell population. Low affinity GATA1 DNA binding sites are less probable to have their local nucleosome evicted simultaneously in the majority of cells from the whole cell population. This behaviour in the whole cell population can reflect the frequency of the nucleosome eviction in a specific genomic position in a unique cell (black arrows). The higher the frequency of eviction (high affinity GATA1 DNA binding site), the more probable the stochastic formation of large GATA1 multiprotein complexes. The lower the frequency of eviction (low affinity GATA1 DNA binding site), the less probable the stochastic formation of large GATA1 multiprotein complexes. On the contrary, small multiprotein complexes of GATA1 can accommodate binding in regions with a lower frequency of nucleosome eviction. One of the functions of the GATA1 transcriptional co-factors, which increase the GATA1 DNA binding affinity (for example FLI1, p300), may be to quickly grab the complex on these positions. Because members of the GATA family can act as pioneer factors (see section 4.3.3) we cannot exclude the fact that small multiprotein complexes of GATA1 can accommodate binding even in the presence of a nucleosome. We assume that the nucleosome eviction is mediated from the SWI/SNF chromatin remodelling complex that interacts with GATA1 (Kadam and Emerson, 2003).

3.5.7 Co-factors change the DNA binding specificities of GATA1 and can appear in low affinity GATA1 ChIP-seq peaks exclusively

As we have seen from the EMSA-seq experiments of GATA1 and its protein complexes the presence of a co-factor can enhance the binding of a transcription factor in a subset of its DNA binding landscape. **We make the assumption that free GATA1 must bind in vivo with the same binding specificities as in vitro. This implies that the frequency of occurrence of specific motifs in vitro must match the frequency of occurrence of the same motifs in vivo. Deviations from this behaviour show that GATA1 cofactors can enhance the binding specificities for a subset of the GATA1 motifs.** To examine the change in the DNA binding specificities for GATA1 co-factors that we did not tested in vitro, we used the ChIP-seq binding positions where GATA1 overlaps with p300 and TAL1 in a CD36+ differentiation model. Both p300 and TAL1 have been shown to interact with GATA1. We also examined the potential change of the DNA binding specificities for the GATA1/PU.1, GATA1/FLI1 and GATA1/RUNX1 overlapping positions in the megakaryocyte differentiation model. PU.1, FLI1 and RUNX1 have all been shown to interact with GATA1. Finally, we used the GATA1 and NF-E2 overlapping sites as a negative control for our approach. Because GATA1 and NF-E2 do not interact with each other we do not expect to see a change in the GATA1 DNA binding specificities.

Apart from the change in the GATA1 DNA binding specificities we also examined the affinity distribution of the GATA1 overlapping peaks with each one of the above mentioned co-factors. As will be shown below TAL1 and PU.1 overlapping peaks do not show any preference for high or low affinity GATA1 ChIP-seq peaks. On the contrary p300, FLI1, RUNX1, KLF1 they all appear in GATA1 low affinity ChIP-seq peaks. This led us to speculate **that the role of p300, FLI1, RUNX1 and KLF1 is to reduce the GATA1 DNA binding turnover by increasing its affinity for the specific positions.**

3.5.7.1 DNA binding specificities of GATA1 change upon complexation with the TAL1 protein. Results from ChIP-seq data

The TAL1 protein always binds on the DNA as a heterodimer. It interacts with the bHLH E-proteins (E2A) in order to be able to bind the E-box DNA motif ("CANNTG"). TAL1 and GATA1 is part of a pentameric complex (TAL1/E2A/LMO2/LDB1/GATA1) (Wadman et al., 1997). The LIM-only domain protein (LMO2) and the LIM domain-binding protein (LDB1) bridge the interaction between GATA1 and TAL1 (El Omari et al., 2011, Anderson et al., 1998). E-box and GATA motifs separated by 9 to 12 nucleotides [GATA{N}₉₋₁₂CANNTG] have been reported in genomic positions where GATA1 and TAL1 co-exist (Vyas et al., 1999, Lahlil et al., 2004, Anderson et al., 1998, Kassouf et al., 2010). We checked in the ChIP-seq data whether the GATA1/TAL1 positions appear in low or high affinity GATA1 ChIP-seq peaks and whether the GATA1 motifs show deviations from the expected in vitro frequencies (Figure 88B). The histogram of Figure 88B shows that the TAL1 appears in GATA1 ChIP-seq peaks with no preference for either high or low affinity binding sites. The barplot of Figure 88B indicates that in comparison with the bar plot of Figure 87A there is a higher deviation (lower occurrence) of the "GATC" motif group from its binding equilibrium. This might be a reflection of the TAL1 binding in the GATA1 N-terminal Zn finger that potentially restricts its DNA binding capability. There is also a higher deviation (higher occurrence) from the binding equilibrium of the AGATA, TGATA, GATAA from the "GATA" motif group. This motif preference also has been noticed in a previous study where the authors present the motif [C/G][A/T]GATAAG as the preferred GATA1 sequence associated with TAL1 peaks (Kassouf et al., 2010). Members from the GATT motif group ([A/T/C]GATT, GATTA) and from the GATG group (GGATG, GATGG) showed a lower deviation from the binding equilibrium. This probably means that the protein acquired a partial structural flexibility to bind these motifs similar with the in vitro conditions where it does not interact with other co-factors.

3.5.7.2 Projection of the EMSA-seq results on ChIP-seq data from the erythropoiesis lineage shows that the DNA binding specificities of GATA1 change upon complexation with the p300 protein and that the GATA1/p300 overlapping sites appear in low affinity GATA1 ChIP-seq peaks

CREB-binding protein (CBP) and its paralog p300 are central regulators of gene expression. CBP/p300 can acetylate histones on H3K27 and H3K18 positions (Jin et al., 2011, Tie et al., 2009), it can acetylate other transcription factors and itself, it can function as a scaffold for recruiting other proteins or it can mediate the interaction of transcription factors with basal factors and the RNA polymerase II (Holmqvist and Mannervik, 2013). A few genomic regions are occupied by either p300 or CBP but the majority of them is occupied by both proteins (Ramos et al., 2010). CBP/p300 directly interacts with the GATA1 C-terminal Zn finger (Blobel et al., 1998). CBP/p300 act as a co-activator of GATA1 (Blobel et al., 1998). CBP/p300 acetylates GATA1 which stimulates its transcriptional activity (Boyes et al., 1998a). Two lysine-rich motifs (aa 243-246 and 312-315) at the C-terminus of the GATA1 zinc fingers are acetylated (Hung et al., 1999). GATA1 recruits CBP/p300 to chromatin sites, including the β -globin LCR and the *Hbb-b1* (b major) promoter and this is probably responsible for the H3 and H4 acetylation at these sites (Letting et al., 2003). We checked in the ChIP-seq data whether the GATA1/p300 positions appear in low or high affinity GATA1 ChIP-seq peaks and whether the GATA1 motifs show deviations from the expected in vitro frequencies (Figure 88A). p300 appears in GATA1 ChIP-seq peaks with low binding affinity. GATA1 acetylation increases the binding affinity of the protein (Boyes et al., 1998a). This might explain the preference of the complex for low affinity sites. From the bar plots of Figure 88A we see that both the “GATA” motif group and the “GATG” motif group show a higher deviation from the binding equilibrium (lower and higher occurrence respectively). As we have discussed in vitro the

“GATG” motif can be bound by both the N-terminal and the C-terminal Zn finger. The increased preference for “GATG” containing motifs might be caused by an increased binding from the GATA1 C-terminal Zn finger or from the GATA1 N-terminal Zn finger. An increased binding from the N-terminal Zn finger would cause GATA1 to bind with a higher affinity due to the cooperativity of the DNA binding domains, as we have discussed earlier. Because in vitro the “GATG” did not show any cooperativity, in contrast to what was noticed for the “GATC” motif, we can speculate that the increased preference for the “GATG” containing motifs would be caused from an increase in the binding of the C-terminal Zn finger. Thus acetylation and/or p300 interaction, probably affects the GATA1 C-terminal Zn finger in a way that decreases the preference for one motif and increases the preference for another motif. This might be one of the reasons that acetylation of GATA1 is required for chromatin occupancy (Lamonica et al., 2006).

3.5.7.3 Projection of the EMSA-seq results on ChIP-seq data from the erythropoiesis lineage shows that the DNA binding specificities of GATA1 do not change in the presence of NF-E2 and that the GATA1/NF-E2 overlapping sites appear in low, medium and high affinity GATA1 ChIP-seq peaks

The p45 nuclear factor-erythroid-derived 2 (NF-E2) is an heterodimer of a small Maf protein (NF-E2p18) and a hematopoietic specific Cap ‘n’ collar (CNC) family of basic leucine transcription factors (NF-E2p45) (Andrews et al., 1993, Gasiorrek et al., 2012, Mignotte et al., 1989). NF-E2 has been shown to have a major role in erythroid differentiation. The NF-E2p45 null cell line fails to express alpha and beta globins (Lu et al., 1994) and NF-E2p45 knockout mice show an increase in the population of early erythroblasts implying a partial block in the maturation progress (Gasiorrek et al., 2012). GATA1 and NF-E2 can have a synergistic effect upon binding on the same genomic

positions but they do not show any direct interaction (Woon Kim et al., 2011). The affinity distribution of the GATA1 and NF-E2 overlapping positions was similar with the affinity distribution from all the GATA1 ChIP-seq peaks, like the TAL1 graph of Figure 88B (Figure 88C). This is not something unexpected for a protein that does not interact with GATA1. Nevertheless in the bar plot of Figure 88C we see an increase preference for the “GATG” motif group. Although this is unexpected assuming that the protein is free to bind, the increase preference for GATG might permit the protein to co-occur in overlapping positions with NF-E2.

3.5.7.4 Projection of the EMSA-seq results on ChIP-seq data from multipotential hematopoietic progenitors shows that the DNA binding specificities of GATA1 do not change in the presence of PU.1 and that the GATA1/PU.1 overlapping sites appear in low, medium and high affinity GATA1 ChIP-seq peaks

PU.1 is a member of the ETS family of transcription factors. PU.1 is a myeloid and B-cell specific transcription factor and is a negative regulator of terminal erythroid differentiation. GATA1 and PU.1 suppress each other's activity by direct protein interaction (Rekhtman et al., 1999, Zhang et al., 1999, Nerlov and Graf, 1998). The GATA1-PU.1 interaction has been shown to reduce the GATA1 DNA binding (Zhang et al., 2000). PU.1 interacts with the C-terminal Zn finger of GATA1 (Zhang et al., 1999). We used GATA1 ChIP-seq data from erythroblasts (Wontakal et al., 2012) and PU.1 ChIP-seq data from multipotential hematopoietic progenitors (ES-EP cells) (Wilson et al., 2010). We found positions where both of them co-exist. We made the assumption that in these positions GATA1 and PU.1 will compete each other during the initial phase of the erythroid differentiation. The histogram of Figure 89C shows that the affinity distribution of the GATA1 and PU.1 overlapping positions was similar with the affinity distribution from all the GATA1 ChIP-seq

peaks. This shows that the inhibition of GATA1 from PU.1 needs to happen in all the binding spectrum of GATA1. In Figure 89C we also see that the motifs pattern of the GATA1/PU.1 overlapping ChIP-seq peaks is similar with the motifs pattern from all the GATA1 ChIP-seq peaks (Figure 87C) indicating that GATA1 does not show any specific DNA binding profile in the GATA1/PU.1 overlapping positions.

3.5.7.5 Projection of the EMSA-seq results on ChIP-seq data from the megakaryopoiesis lineage shows that the DNA binding specificities of GATA1 change in the presence of FLI1 and that the GATA1/FLI1 overlapping sites appear in low affinity GATA1 ChIP-seq peaks

Friend leukemia integration 1 (FLI1), is a member of the ETS family of transcription factors which bind to the DNA consensus sequence “GGA(A/T)”. FLI1 is an important regulator during megakaryocytic differentiation. FLI1 binds to both early and late megakaryocyte genes (Pang et al., 2006). FLI1 knockout mice are embryonic lethal and show severe dysmegakaryopoiesis along with other pathologies (Hart et al., 2000, Kawada et al., 2001). Both the N-terminal Zn finger and the C-terminal Zn finger of GATA1 was able to interact with FLI1 (Eisbacher et al., 2003). We used ChIP-seq data from primary human megakaryocytes that were derived from differentiated CD34+ HPCs in the presence of thrombopoietin and Interleukin-1 β . We checked in the ChIP-seq data whether the GATA1/FLI1 positions appear in low or high affinity GATA1 ChIP-seq peaks and whether the GATA1 motifs show deviations from the expected in vitro frequencies (Figure 89B). The histogram of Figure 89B shows that the FLI1 appears in GATA1 ChIP-seq peaks with low binding affinity. The bar plot of Figure 89B indicates that in comparison with the bar plot of Figure 87B there is a higher deviation (higher occurrence) of the “GATG” motif group, from its binding equilibrium. Half of the motifs of the “GATT” motif group show

higher deviation, higher ([G]GATT[C/T]) and lower ([A/T/C]GATT) occurrence, whereas the “GATA” motif group shows lower deviation ([A/C]GATA[A]).

3.5.7.6 Projection of the EMSA-seq results on ChIP-seq data from the megakaryopoiesis lineage shows that the DNA binding specificities of GATA1 change in the presence of RUNX1 and that the GATA1/RUNX1 overlapping sites appear in low affinity GATA1 ChIP-seq peaks

RUNX1 (AML1) is a member of the RUNT family of transcription factors. RUNX1 binds directly the sequence “TGTGGT” by forming a dimer with the other subunit, CBF β . RUNX1 is important for definitive hematopoiesis in the foetus (Okuda et al., 1996, Wang et al., 1996). In the adult, RUNX1 is necessary for megakaryocyte maturation (Ichikawa et al., 2004, Growney et al., 2005) and normal development of the T-lymphocytes and B-lymphocytes (Taniuchi et al., 2002, Kuo et al., 2008). RUNX1 is associated with the activation of gene targets by recruiting histone acetylases (p300, CBP) (Kitabayashi et al., 1998) or the transcriptional co-activators YAP1 (Yagi et al., 1999) and MOZ proteins (Kitabayashi et al., 2001). The amino (aa 1-85) and carboxy termini (aa 319-413) of the hGATA1 have been shown to have a direct interaction with the hRUNX1 (Elagib et al., 2003). Also the N-terminal and the C-terminal Zn finger of GATA1 have been shown to interact with RUNX1 (Waltzer et al., 2002, Xu et al., 2006). We used ChIP-seq data from primary human megakaryocytes that were derived from differentiated CD34+ HPCs in the presence of thrombopoietin and interleukin-1 β . We checked in the ChIP-seq data whether the GATA1/RUNX1 positions appear in low or high affinity GATA1 ChIP-seq peaks and whether the GATA1 motifs show deviations from the expected in vitro frequencies (Figure 89A). The histogram of Figure 89A shows that the RUNX1 appears in GATA1 ChIP-seq peaks with low binding affinity. The bar plot of Figure 89B indicates that similarly with FLI1

there is a higher deviation (higher occurrence) of the “GATG” motif group from its binding equilibrium ([TG]GATG[ATCG]). The “GATA” motif group shows again lower deviation ([A/C/G]GATA[A]), a pattern similar with the FLI1, whereas the “GATT” motif group shows higher (high occurrence: [G]GATT[C], low occurrence: [A]GATT) and lower ([C]GATT[A/G]) deviations. In general the FLI1 pattern and the RUNX1 pattern seemed similar and distinct from the p300, TAL1 and NF-E2 pattern. This cannot be attributed to the different cell lines used because the overall pattern of the GATA1 ChIP-seq peaks in Figure 87A and Figure 87B looks similar.

3.5.7.7 GATA1/KLF1 overlapping sites but not GATA1/TAL1 overlapping sites appear in low affinity GATA1 ChIP-seq peaks in the TER119-/TER119+ mouse model of erythropoiesis

We checked the TAL1/GATA1 overlapping positions behaviour in the TER119+/TER119- differentiation model. We separated the GATA1 ChIP-seq peaks in three categories. Peaks that appear only in the TER119+ cells. Peaks that appear only in the TER119- cells. And peaks that appear in both TER119+/TER119- cells. We see that in all these categories TAL1 has no preference for either high or low affinity GATA1 ChIP-seq peaks (Figure 90).

When we checked the KLF1/GATA1 ChIP-seq overlapping positions in the TER119+/TER119- differentiation model, we notice that KLF1 prefers to bind in low affinity GATA1 ChIP-seq peaks in all three categories (Figure 90).

We also examined the KLF1 binding affinity distribution of the KLF1/GATA1 ChIP-seq overlapping positions. We see no preference for either high or low affinity KLF1 ChIP-seq peaks in all three categories (Figure 91).

In Figure 92 we present the binding models of the complexes of GATA1 with the examined co-factors.

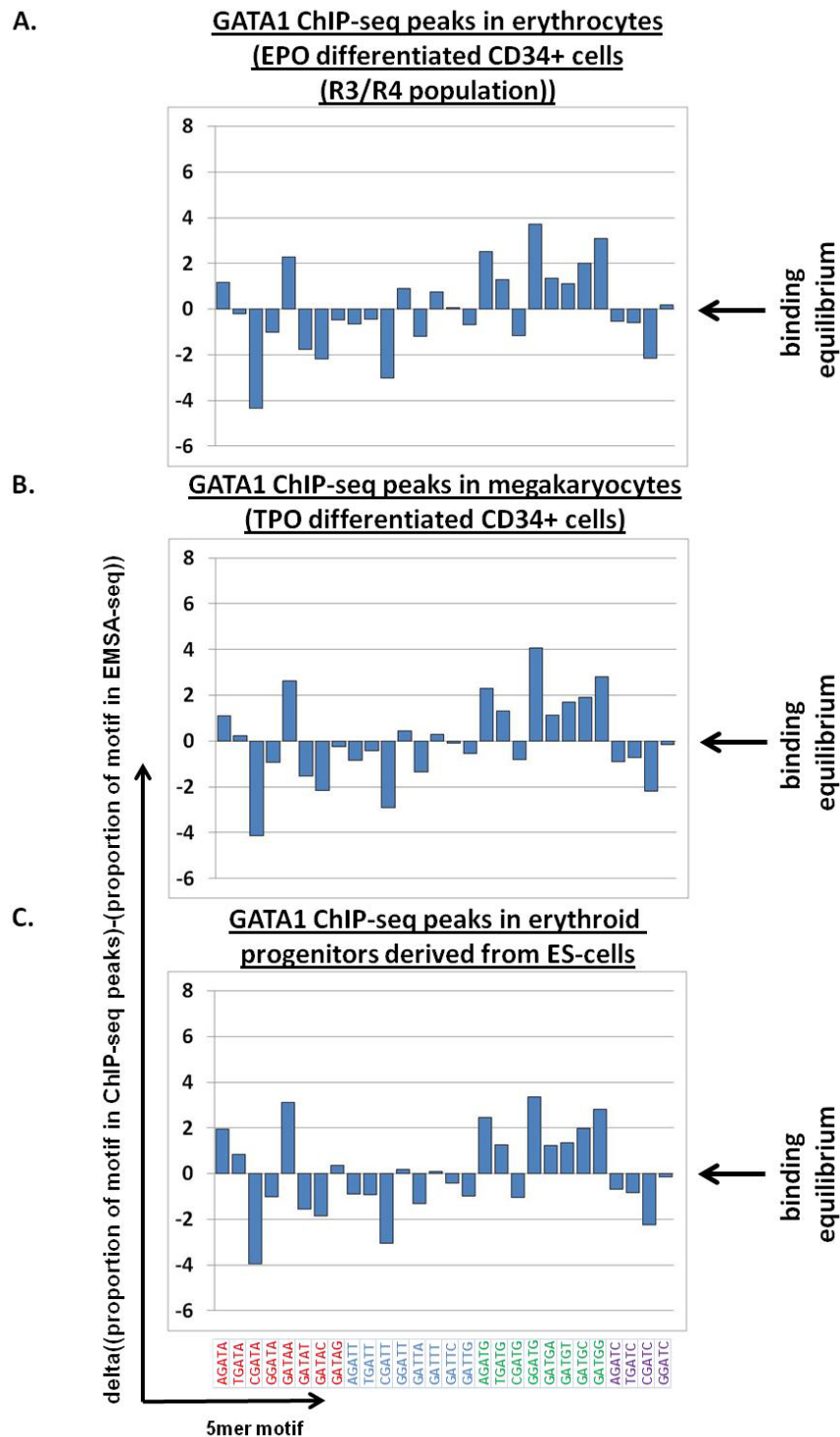


Figure 87. Difference between the proportion of the presented GATA1 5mer motifs in the indicated GATA1 ChIP-seq peaks minus the proportion of the same motifs in the mGATA1_{NC} EMSA-seq experiment of condition 1. The 5mer motifs are separated in the “GATA” motif group (red colour), in the “GATT” motif group (blue colour), in the “GATG”

motif group (green colour) and in the “GATC” motif group (purple colour). The zero value represents no difference from the expected EMSA-seq distribution of proportions which we assume that it represents the in vitro binding equilibrium. The GATA1 ChIP-seq peaks of figure (A) were from (Xu et al., 2012) (GEO accession number: GSM970258). The GATA1 ChIP-seq peaks of figure (B) were from (Tijssen et al., 2011). The GATA1 ChIP-seq peaks of figure (C) were from (Wontakal et al., 2012) (GEO accession number: GSM867156 and GSM867157 respectively).

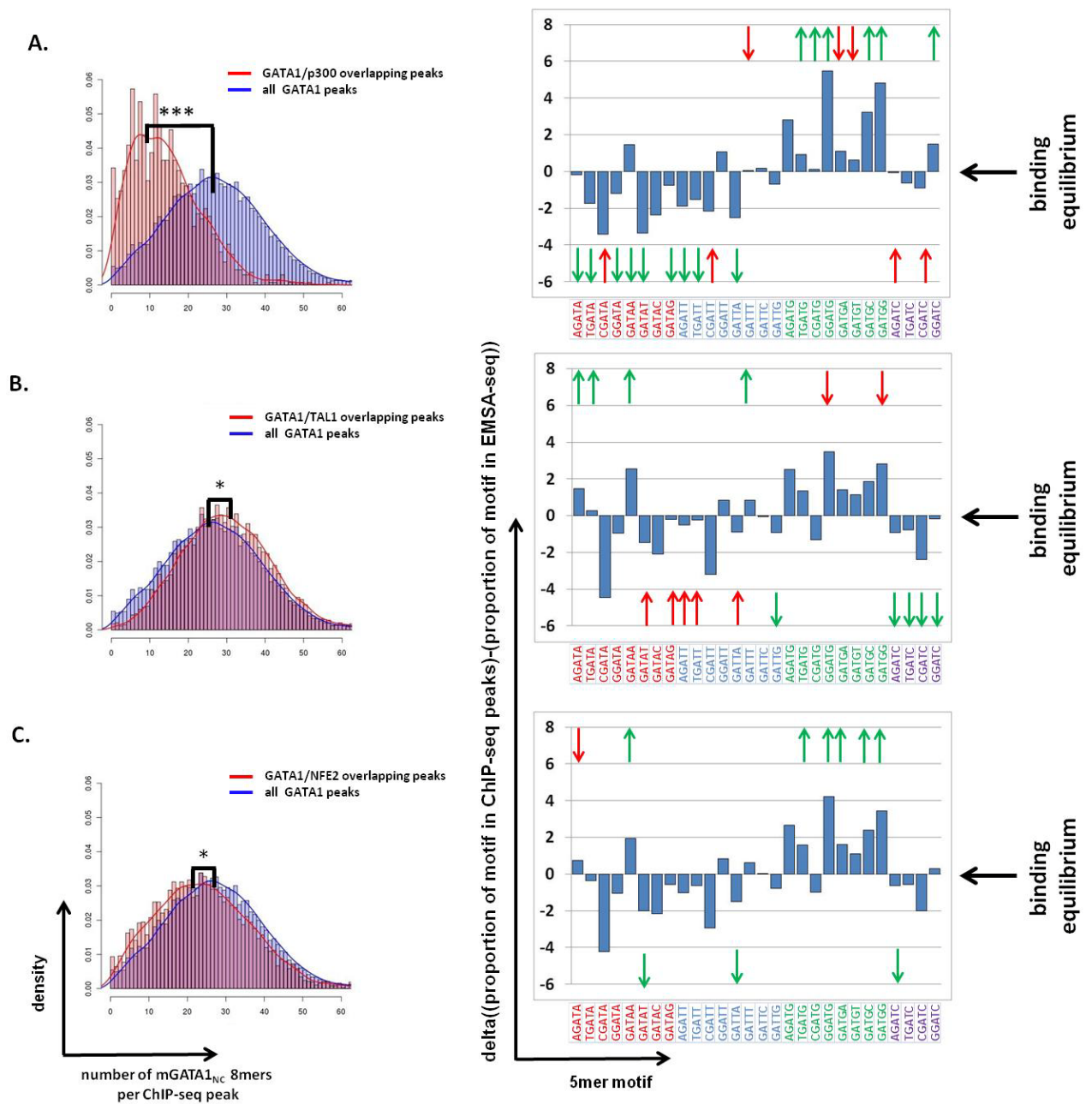


Figure 88. DNA binding affinity distributions of the GATA1 ChIP-seq peaks in sites where they overlap with the ChIP-seq peaks from the p300 (A) or the TAL1 (B) or the NF-E2 protein (C). Both the affinity distributions for all the GATA1 ChIP-seq peaks (blue line and blue histogram) and their overlapping peaks is shown (red line and red histogram). The bar plots in the second column represents the difference between the proportion of the indicated GATA1 5mer motifs in the overlapping ChIP-seq peaks minus the proportion of

the same motifs in the mGATA1_{NC} EMSA-seq experiment of condition 1. The 5mer motifs are separated in the “GATA” motif group (red colour), in the “GATT” motif group (blue colour), in the “GATG” motif group (green colour) and in the “GATC” motif group (purple colour). The values of the bar plots for figures (A), (B), (C) were compared with the ones from Figure 87A for the same 5mer and the difference is indicated with arrows. The green arrow indicates an increase in the absolute value of the bar plot whereas the red arrow indicates a decrease in the absolute value of the bar plot. The arrows on the bottom row of each bar plot graph represent the change of the value for the bars that are below the zero value. The arrows on the top row represent the change of the value for the bars that are above the zero value. The zero value represents no difference from the expected EMSA-seq distribution of proportions which we assume that it represents the in vitro binding equilibrium. The GATA1 ChIP-seq peaks of figures (A), (B), (C) were from (Xu et al., 2012) (GEO accession number: GSM970258) (cell type: proerythroblasts (ProEs) from adult human bone marrow CD34+ cells). The p300, TAL1, NF-E2 ChIP-seq peaks of figures (A), (B), (C) were from (Su et al., 2013) (GEO accession number: GSM1067278, GSM1067277, GSM1067276 respectively) (cell type: R3/R4 erythroid cells derived from human bone marrow CD34+ cells) (in both studies of (Xu et al., 2012) and (Su et al., 2013) the erythroid differentiation of human bone marrow CD34+ cells was performed with the same growth protocol).

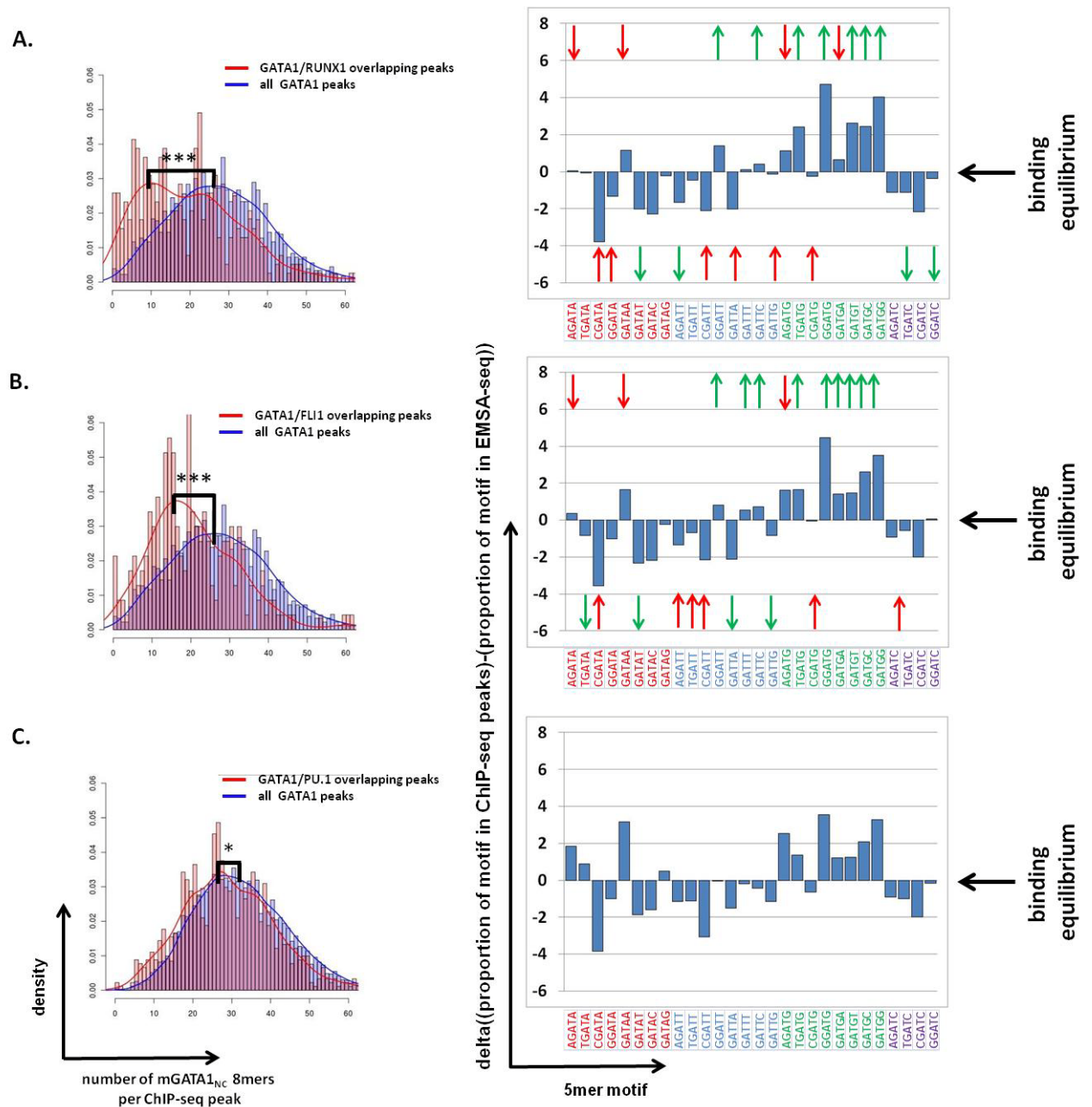


Figure 89. DNA binding affinity distributions of the GATA1 ChIP-seq peaks in sites where they overlap with the ChIP-seq peaks from the RUNX1 (A) or the FLI1 (B) or the PU.1 protein (C). The figure is similar with Figure 88. The values of the bar plots for figures (A), (B) were compared with the ones from Figure 87B whereas the values of the bar plots for figure (C) were compared with the ones from Figure 87C for the same 5mer. The GATA1, FLI1 and RUNX1 ChIP-seq peaks of figures (A), (B) were from (Tijssen et al., 2011) (cell type: megakaryocytes cultured from cord blood human CD34⁺ cells). The GATA1 ChIP-

seq peaks of figure (C) were from (Wontakal et al., 2012) (GEO accession number: GSM867156 and GSM867157) (cell type: ES-cell derived mouse erythroid progenitors). The PU.1 ChIP-seq peaks of figure (C) were from (Wilson et al., 2010) (cell type: HPC-7 Cell Line, early haematopoietic mouse cell line derived from ES cells).

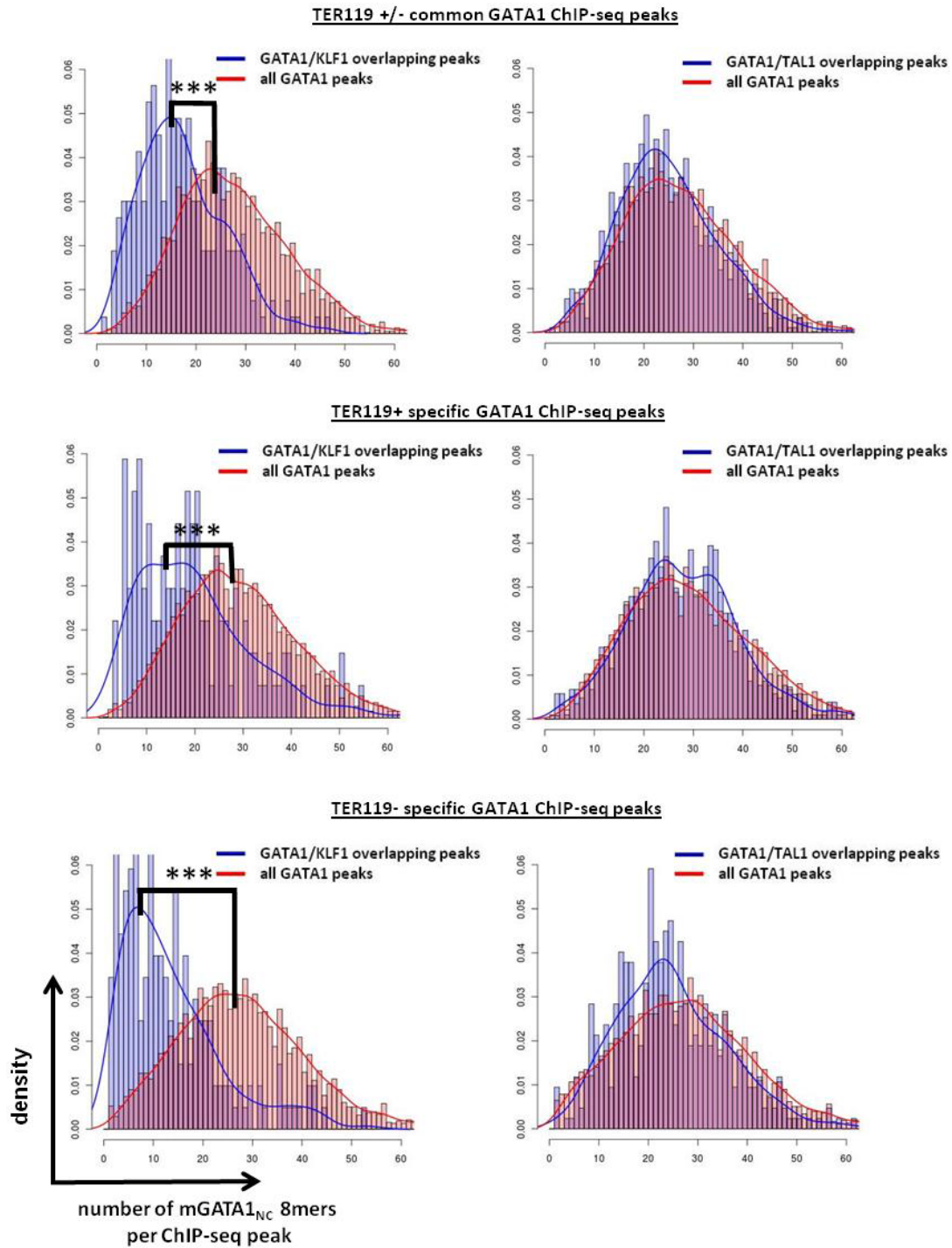


Figure 90. DNA binding affinity distributions of the GATA1 ChIP-seq peaks that overlap with the KLF1 or TAL1 ChIP-seq peaks (blue line and blue histogram) in comparison with the DNA binding affinity distributions of all the GATA1 ChIP-seq peaks (red line and red histogram). The GATA1 ChIP-seq peaks were from the TER119+/- model and the peaks were classified in 3 categories (common GATA1 ChIP-seq peaks in the TER119+ and TER119- cells, GATA1 ChIP-seq peaks that appear only in the TER119+ cells, GATA1

ChIP-seq peaks that appear only in the TER119- cells). The ChIP-seq peak affinity represents the number of 8mers per ChIP-seq peak from the 2122 enriched mGATA1_{NC} 8mers. The GATA1 ChIP-seq peaks were from (Papadopoulos et al., 2013). The KLF1 ChIP-seq peaks were from (Pilon et al., 2011) (GEO accession number: GSM1165492 for the TER119- and GSM1165493 for the TER119+).

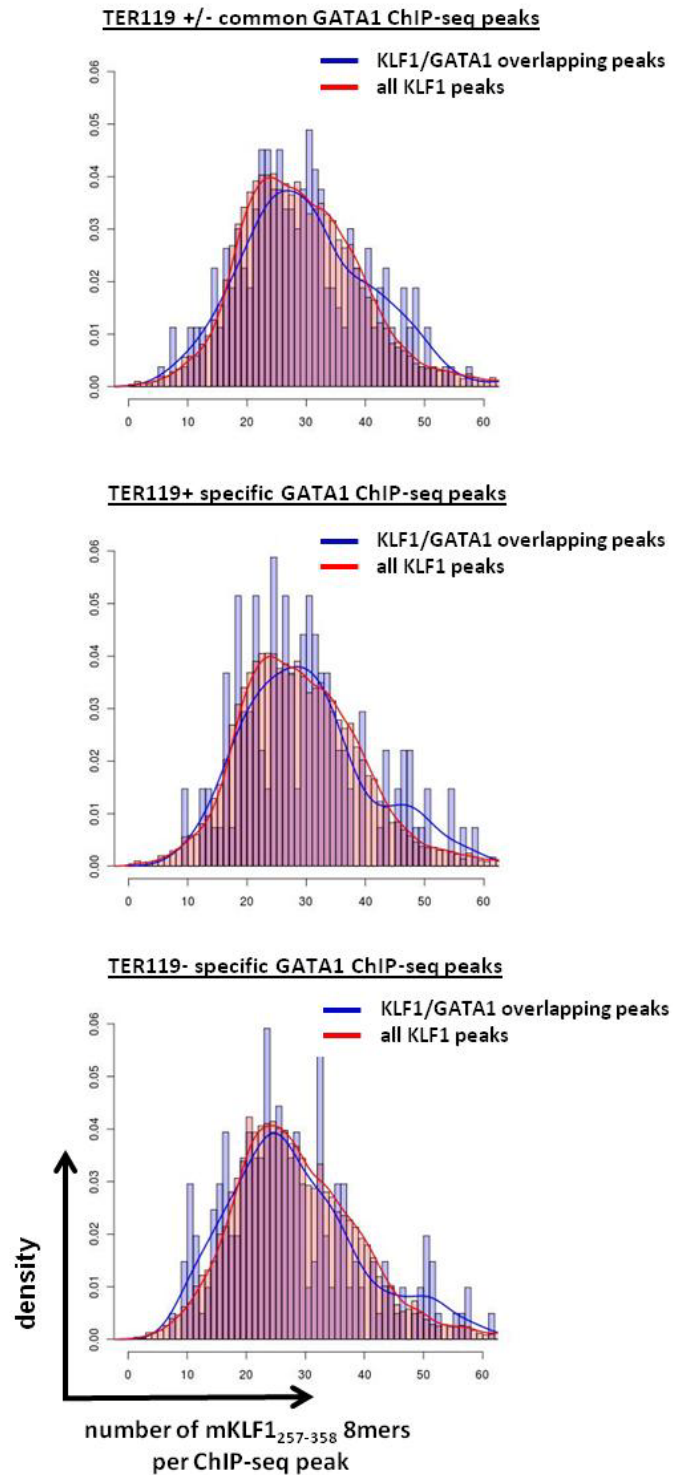


Figure 91. DNA binding affinity distributions of the KLF1 ChIP-seq peaks that overlap with the GATA1 ChIP-seq peaks (blue line and blue histogram) in comparison with the DNA binding affinity distributions of all the KLF1 ChIP-seq peaks (red line and red histogram). The KLF1 ChIP-seq peaks were from the TER119+/- model and the peaks were classified

in 3 categories (common KLF1 ChIP-seq peaks in the TER119+ and TER119- cells, KLF1 ChIP-seq peaks that appear only in the TER119+ cells, KLF1 ChIP-seq peaks that appear only in the TER119- cells). The ChIP-seq peak affinity represents the number of 8mers per ChIP-seq peak from the 193 enriched mKLF1₂₅₇₋₃₅₈ 8mers. The GATA1 ChIP-seq peaks were from (Papadopoulos et al., 2013). The KLF1 ChIP-seq peaks were from (Pilon et al., 2011) (GEO accession number: GSM1165492 for the TER119- and GSM1165493 for the TER119+).

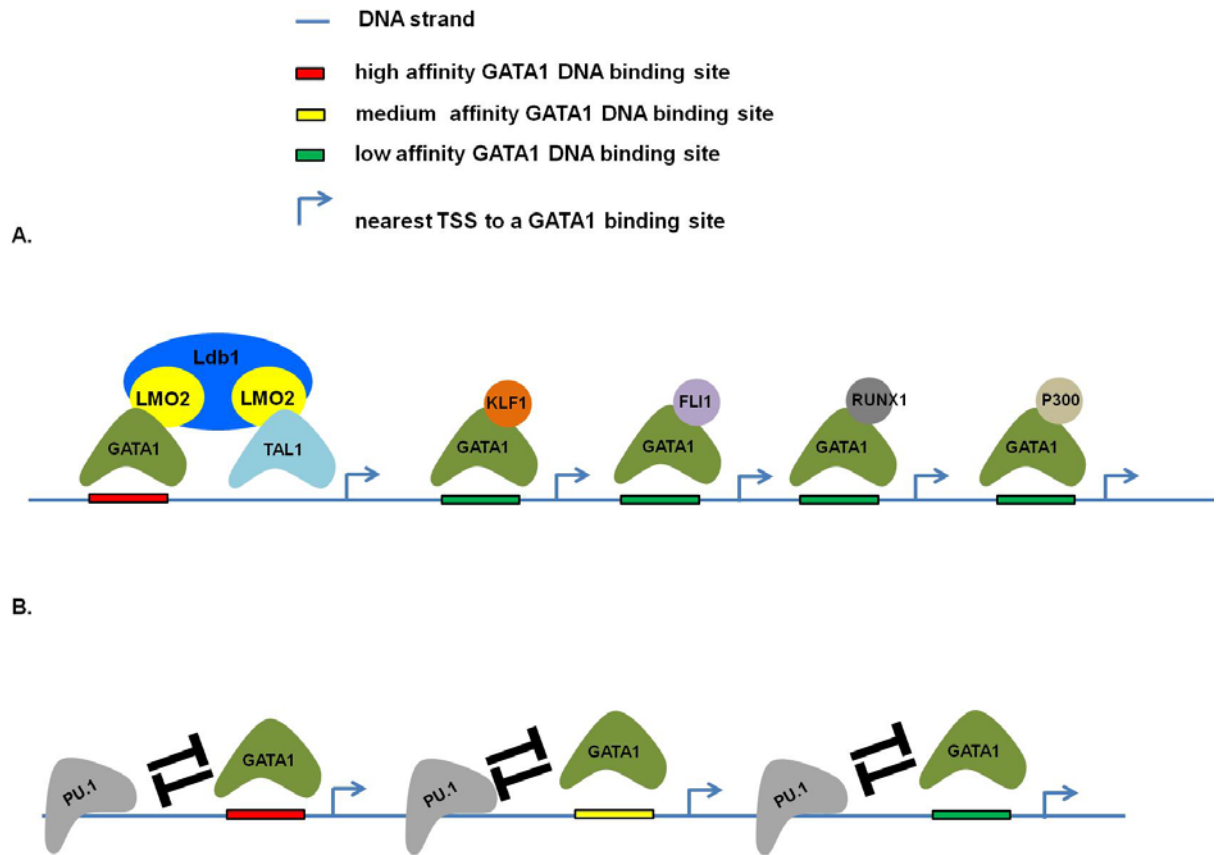


Figure 92. Model for the selective preference of genomic positions with different DNA binding affinities from the GATA1 complexes. A) The GATA/KLF1, GATA1/FLI1, GATA1/RUNX1 and GATA1/p300 complexes are more frequently found in genomic positions with low GATA1 DNA binding affinity. The GATA1/TAL1 complex (pentameric complex of GATA1, TAL1, LMO2, LDB1) is frequently found in genomic positions with medium to high GATA1 DNA binding affinity. B) The GATA1/PU.1 self-antagonistic complex is formed in genomic positions with low, medium and high GATA1 DNA binding affinity. This indicates the necessity to inhibit the GATA1 function independently of the GATA1 DNA binding affinity of the genomic position.

3.5.8 Binding affinities of ChIP-seq peaks give insights into the transcriptional regulatory mechanisms during erythroid and megakaryocytic differentiation

3.5.8.1 GATA1 high affinity ChIP-seq peaks are present in regions with increased levels of H3K4me2 and are associated with a higher decrease in the H3K4me3 levels of the closest TSS

We first checked the correlation between the presence of H3K4me3 and H3K4me2 on the TSS in relationship with the affinity of the nearby GATA1 ChIP-seq peaks in the CD36+ model of differentiation. We normalised the number of H3K4me3 and H3K4me2 reads on the TSS (+200bp) with the number of reads in the input. We also calculated the ratio of the H3K4me3 versus the H3K4me2 reads. Because the H3K4me3 is the result of the methylation of the H3K4me2, the H3K4me3/H3K4me2 ratio indicates whether in a given TSS the balance has been shifted towards the H3K4me3 or towards the H3K4me2. The normalisation with the input indicates that in the total amount of nucleosomes, that were isolated for the specific region, what fraction of them has the histone mark whereas the ratio of the histone marks indicates whether the frequency of occurrence of the specific histone marks have changed relative to each other independently of the total amount of nucleosomes (Figure 93).

In Figure 94 we notice that in general high DNA binding affinity GATA1 ChIP-seq peaks are associated with a lower ratio of H3K4me3/H3K4me2. In Figure 94A we can notice both a reduction of the H3K4me3 levels and an increase of the H3K4me2 levels. We notice here that the reduction of the H3K4me3 levels in the high affinity ChIP-seq peaks is not as prominent as the increase of the H3K4me2 levels. Low DNA binding affinity ChIP-seq peaks are associated with TSS regions with a higher ratio of H3K4me3/H3K4me2. The lower ratio of the H3K4me3/H3K4me2 fraction in the high DNA binding affinity

GATA1 ChIP-seq peaks can be caused from either a reduction in the H3K4me3 levels or from an increase in the H3K4me2 levels. Although intuitively we are expecting the aetiology to be assigned to the increased occurrence of the H3K4me2 levels in the total nucleosome levels for the high affinity GATA1 ChIP-seq peaks (Figure 94A middle graph), we caution here that the fraction represents the levels of the H3K4me2 and H3K4me3 per se independently from the total nucleosome levels (Figure 93). The total nucleosome levels are represented from the ratio of the reads of the histone mark versus the reads in the input. For this reason to decipher whether it is the H3K4me2 or the H3K4me3 levels that change significantly we created a scatter plot of the raw H3K4me3 and the H3K4me2 values for:

- i) The TSS of all the genes.
- ii) The TSS of genes that have a low or medium affinity GATA1 ChIP-seq peak near the TSS.
- iii) The TSS of genes that have a high affinity GATA1 ChIP-seq peak near the TSS.

In Figure 95 we see that only the median values of the H3K4me3 levels for all these categories differ, whereas the median values of the H3K4me2 levels are similar (median values of the H3K4me2 levels between 17-17.5). This indicates to us that GATA1 regulates directly the H3K4me3 levels whereas it exploits the local H3K4me2 levels, as will be discussed in the next paragraphs. We also notice a bimodal role of GATA1 in the increase of the H3K4me3 levels. Whereas the median value of the H3K4me3 levels for all the TSS are 113, the TSS that are associated with low and medium affinity GATA1 ChIP-seq peaks show an increased median value of 140. Contrary to the expectation for even higher H3K4me3 levels, the TSS that are associated with high affinity GATA1 ChIP-seq peaks show an increased median value of 133 compared to all the TSS but this value is lower than the one from the TSS of low and medium affinity GATA1 ChIP-seq peaks. We will discuss again this bimodal regulation in section 3.5.8.4. **We caution that we compare**

GATA1 affinities with steady state kinetics as this is represented from the histone modifications which might not give a clear direct association.

We also tested the above trend in the TER119-/TER119+ model. No raw data for the input were available so we could only examine the ratio H3K4me3/H3K4me2. We notice a similar trend for the GATA1 ChIP-seq peaks that are common between the TER119- and TER119+ cells (Figure 94B). The trend was not very clear for the GATA1 ChIP-seq peaks that were present only in the TER119- or the TER119+ population (TER119- and TER119+ specific ChIP-seq peaks respectively) (Figure 94C,D).

We then separated the promoters in GC rich (more than 50% of GC in the region -100, -1 upstream of the TSS) (2846 in total) and AT rich (less than 50% of GC in the region -100, -1 upstream of the TSS) (359 in total) promoters (Bajic et al., 2006). In Figure 96 we see that the reverse correlation between the H3K4me3/H3K4me2 ratio and the binding affinity of the nearby GATA1 peak is much more prominent in the AT rich (GC poor) promoters than the GC rich promoters. This can be attributed to some extent to the fact that there are much more high affinity GATA1 ChIP-seq peaks associated with AT rich promoters, as expected, than GC rich promoters. Nevertheless, because the range of values (upper and lower quantile) in each DNA binding affinity category is higher in the AT rich (GC poor) promoters than in the GC rich promoters (compare the range of values for the DNA binding affinity categories 7-12 versus categories 2-6 in Figure 96A,B) we might be able to conclude that counterbalancing mechanisms (SET1 methyltransferase complex) are present in the GC rich promoters (see section 4.3.5 in discussion) that keep the H3K4me3 levels under tight control. Consequently, high affinity GATA1 ChIP-seq peaks might not be able to effectively reduce the H3K4me3 levels in the GC rich promoters and their presence might be more productive, and preferably necessary, in the AT rich (GC poor) promoters.

We then checked whether GATA1 can affect the H3K4me2/H3K4me1 ratio. Because the H3K4me1 mark is enriched in positions away from the TSS we examined how the ratio of the H3K4me2 and H3K4me1 on the GATA1 ChIP-seq peak site changes with the ChIP-seq DNA binding affinity in the CD36+ model of erythroid differentiation. In Figure 97C we see that no change was observed which probably indicates that either GATA1 does not affect directly and independently at least one of the H3K4me2 and the H3K4me1 marks or that GATA1 affects directly and in the same way the deposition of the H3K4me2 and the H3K4me1 marks. In Figure 97A and Figure 97B we notice that the high affinity GATA1 binding sites are present in regions with high levels of H3K4me1 and H3K4me2 histone marks when we examined the occurrence of the histone marks in the total nucleosome levels. Considering that we have seen that the high affinity GATA1 binding sites are present in regions with low nucleosome occupancy and that it is known that the histone post translational modifications increase the turnover of the histones on chromatin (Zee et al., 2010, Dion et al., 2007), we can conclude that the post translational modifications must exist or pre-exist in order to create a permissive chromatin environment for GATA1 to bind. We favour a model where GATA1 can guide directly or indirectly the deposition of the H3K4me2 or H3K4me1 mark in order to facilitate the remodelling of the nucleosomes which eventually causes their eviction and the observed low nucleosome occupancy (Figure 93, Figure 107).

3.5.8.2 The DNA binding affinity of the GATA1 ChIP-seq peaks cannot discriminate the upregulated and the downregulated genes during erythroid differentiation

In the CD36+ differentiation model, it has been reported 1137 genes that are upregulated and 1186 genes that are downregulated (Xu et al., 2012). We checked whether the upregulated and the downregulated genes showed a different distribution of GATA1 ChIP-

seq binding affinities. Because we have seen that the region ± 1 kb from the TSS has a greater range of DNA binding affinities for the GATA1 ChIP-seq peaks (Figure 76), something that might imply a potential functional role, we checked the GATA1 ChIP-seq peak affinity distribution for this region.

In Figure 98 we see that the GATA1 ChIP-seq binding affinities distributions (**EMSA-seq k-mer number score**) are similar in both the upregulated (367 GATA1 ChIP-seq peaks) and the downregulated genes (251 GATA1 ChIP-seq peaks) for the region 1kb upstream and downstream of TSS. The KLF1 ChIP-seq binding affinities distributions (**EMSA-seq k-mer number score**) did not show also any distinct pattern (689 KLF1 ChIP-seq peaks correspond to upregulated genes and 710 KLF1 ChIP-seq peaks correspond to downregulated genes).

We, also, used available microarray data from a very detailed time course of the erythroid lineage developmental programme (May et al., 2013). The authors have clustered the genes in 30 groups based on their co-expression pattern during the differentiation program. Rather than examining the GATA1 binding affinities distribution in all the upregulated/downregulated genes we thought that if we restrict our attention in groups of genes that show similar expression pattern during the differentiation process, thus they should be regulated by the same mechanisms, then we might be able to deconvolve the potential effect of the GATA1 DNA binding affinities on gene expression in a clear way. This probably stands true because GATA1 appears mainly at the late stages of the differentiation program. Still even after the above stratification of genes the GATA1 DNA binding affinities could not be associated directly with the up-regulation or the down-regulation of the erythroid lineage genes (data not shown).

3.5.8.3 The change in the TSS related H3K4me3 levels during erythroid differentiation, for both the upregulated and the downregulated genes, is correlated with the DNA binding affinity of the closest promoter associated GATA1 ChIP-seq peak

We have already seen that the GATA1 ChIP-seq binding affinities are associated with a shift in the balance between the H3K4me2 and H3K4me3 in the CD36+ model. For this reason we hypothesized that GATA1 might regulate the balance between the H3K4me2 and H3K4me3 levels during the differentiation process. Because we did not have any available H3K4me3, H3K4me2 datasets for the human CD34+ progenitors we moved to the G1E model of erythropoiesis. Upon GATA1 restoration and in the presence of erythropoietin the cells can be differentiated into erythroblasts. In the G1E model we do not have available H3K4me2 data for the undifferentiated and the differentiated state. Although in Figure 94 we saw that the change that the H3K4me2 levels shows in relation to the GATA1 ChIP-seq binding affinities is more prominent than the H3K4me3 levels, still we hypothesized that a shift in the balance of the H3K4me3/H3K4me2 ratio will be reflected in the ratio (normalized H3K4me3 reads of TER119+)/(normalized H3K4me3 reads of TER119-). Thus, we asked what is the effect of GATA1 on the H3K4me3 levels upon GATA1 restoration. We also note here that the advantage of the G1E model is that there is no GATA1 bound on the genome in the undifferentiated state. On the contrary in the TER119-/TER119+ model GATA1 ChIP-seq peaks can be found in both the undifferentiated (TER119-) and the differentiated (TER119+) state.

The G1E cells express basal levels of GATA2. GATA1 and GATA2 can occupy the same binding site on the genome and it is very well documented the GATA1/GATA2 switch on these binding sites (Dore et al., 2012). To avoid any functional redundancy of GATA1 and GATA2 on the level of the H3K4me3 of nearby genes we focused only in the GATA1 binding sites that do not overlap with the GATA2 binding sites. We used 11492 GATA1

binding sites (Wu et al., 2011) and 18149 GATA2 bindings sites (GEO accession number: GSM777091 (Dore et al., 2012)) . After overlapping them we were left with 8530 GATA1 binding sites that do not show any GATA2 signal. We assigned the nearby genes and we found 125 upregulated genes and 148 downregulated genes that are in a distance of 10 kb upstream and downstream of a TSS. We eventually used TSS that have in either condition G1E basal and G1E-ER4 more than 20 H3K4me3 reads (50 upregulated genes and 75 downregulated genes).

We then correlated the binding affinity (**EMSA-seq k-mer number score**) of GATA1 ChIP-seq peaks with the change on the H3K4me3 status of the nearest TSS in the TER119+/TER119- cells. We performed this analysis in 4 groups of peaks that differ based on their distance from the TSS:

- i) GATA1 ChIP-seq peaks that are between 10 kb and 1 kb upstream from the TSS
- ii) GATA1 ChIP-seq peaks that are between 1 kb and 0 kb upstream from the TSS
- iii) GATA1 ChIP-seq peaks that are between 0 kb and 1 kb downstream from the TSS
- iv) GATA1 ChIP-seq peaks that are between 1 kb and 10 kb downstream from the TSS

We performed the same analysis for the upregulated and the downregulated genes independently (Figure 99).

In Figure 99 we see that in the region between 10 kb and 1 kb upstream from the TSS, the higher the binding affinity the higher the change on the H3K4me3 levels (the lower the ratio (normalized H3K4me3 reads of TER119+)/ (normalized H3K4me3 reads of TER119-)) in both the upregulated and the downregulated genes. For the region between 1 kb and 0 kb upstream from the TSS this relationship was prominent only for the downregulated genes and not for the upregulated genes. The same was observed for the region between

0 kb and 1 kb downstream from the TSS, where the downregulated genes showed a more prominent relationship than the upregulated genes. In the region between 1 kb and 10 kb downstream from the TSS, the relationship is prominent only for the upregulated genes whereas the downregulated genes do not show any correlation. Because in the region between 1 kb and 0 kb upstream from the TSS the upregulated and the downregulated genes did not show the same trend we hypothesize that the transcriptional outcome might be the result of a combination of factors and not of the GATA1 action exclusively. As we have already seen, the region 1 kb upstream from the TSS has a big range of DNA binding affinities for the GATA1 ChIP-seq peaks (Figure 76) and this might be an extra indication for its potential functional role in the fine regulation of the system.

Thus, as we have seen the GATA1 DNA binding affinities are associated with the H3K4me3, H3K4me2 levels (Figure 100) but not with the final outcome (upregulation/downregulation) of the gene expression. This implies that GATA1 acts independently on the H3K4me3, H3K4me2 levels along with other transcription factors or other transcriptional co-factors and the final integration of all these inputs results in the upregulation/downregulation of the respective genes.

	○ Nucleosome ● Nucleosome with the H3K4me3 mark ● Nucleosome with the H3K4me2 mark ● Nucleosome with the H3K4me1 mark			
	nucleosomes present on the specific locus in the total cell population	number of H3K4me3 nucleosomes in the total population of nucleosomes	number of H3K4me2 nucleosomes in the total population of nucleosomes	ratio of H3K4me3 versus H3K4me2 nucleosomes
TSS near low affinity GATA1 ChIP-seq peaks		4/16 (25%)	4/16 (25%)	4/4 (100%)
TSS near high affinity GATA1 ChIP-seq peaks		2/8 (25%)	4/8 (50%)	2/4 (50%)
	nucleosomes present on the specific locus in the total cell population	number of H3K4me2 nucleosomes in the total population of nucleosomes	number of H3K4me1 nucleosomes in the total population of nucleosomes	ratio of H3K4me2 versus H3K4me1 nucleosomes
binding position of low affinity GATA1 ChIP-seq peaks		4/16 (25%)	4/16 (25%)	4/4 (100%)
binding position of high affinity GATA1 ChIP-seq peaks		4/8 (50%)	4/8 (50%)	4/4 (100%)

Figure 93. Working model hypothesis of the frequency of occurrence of nucleosomes with specific histone marks in the total nucleosome population and relative to each other. The graph shows nucleosomes that appear in a TSS region that is near a high or low affinity GATA1 ChIP-seq peak. It also shows nucleosomes that appear on a high or low affinity GATA1 ChIP-seq peak. As we have seen high affinity GATA1 binding sites show lower levels of nucleosome occupancy (Figure 86). Additionally, we make the assumption that contacts of the GATA1 binding sites with the TSS region, due to local DNA looping, can affect similarly the nucleosome occupancy levels of the TSS (Figure 107). Nucleosomes with specific histone post-translational modifications (H3K4me3, H3K4me2, H3K4me1) are shown with colour. Their frequency of occurrence in the total nucleosome population are shown as percentage in the first two columns of the mathematical calculations. The third column represents the ratio of the nucleosomes with the specific histone post-

translational modifications relative to each other and independently of the total nucleosome population. In the TSS regions we can see that a reduction in the whole nucleosome population is reflected as an increase in the H3K4me2 levels (frequency of occurrence) although the number of H3K4me2 containing nucleosomes is the same. Additionally, although there is a reduction in the number of the H3K4me3 containing nucleosomes, their frequency of occurrence is the same in the lower nucleosome population compared to the bigger one. On the contrary, the ratio of the number of the H3K4me3 reads versus the number of the H3K4me2 reads in the two different nucleosome populations reflects the reduction in the number of the H3K4me3 containing nucleosomes. For the GATA1 binding sites, either low or high affinity ones, the reduction in the whole nucleosome population is reflected as an increase in the H3K4me2 and H3K4me1 levels (frequency of occurrence) although the number of the H3K4me2 and H3K4me1 containing nucleosomes is the same. The ratio of the number of the H3K4me2 reads versus the number of the H3K4me1 reads in the two different nucleosome populations reflects the lack of change in the number of the H3K4me2 and H3K4me1 containing nucleosomes.

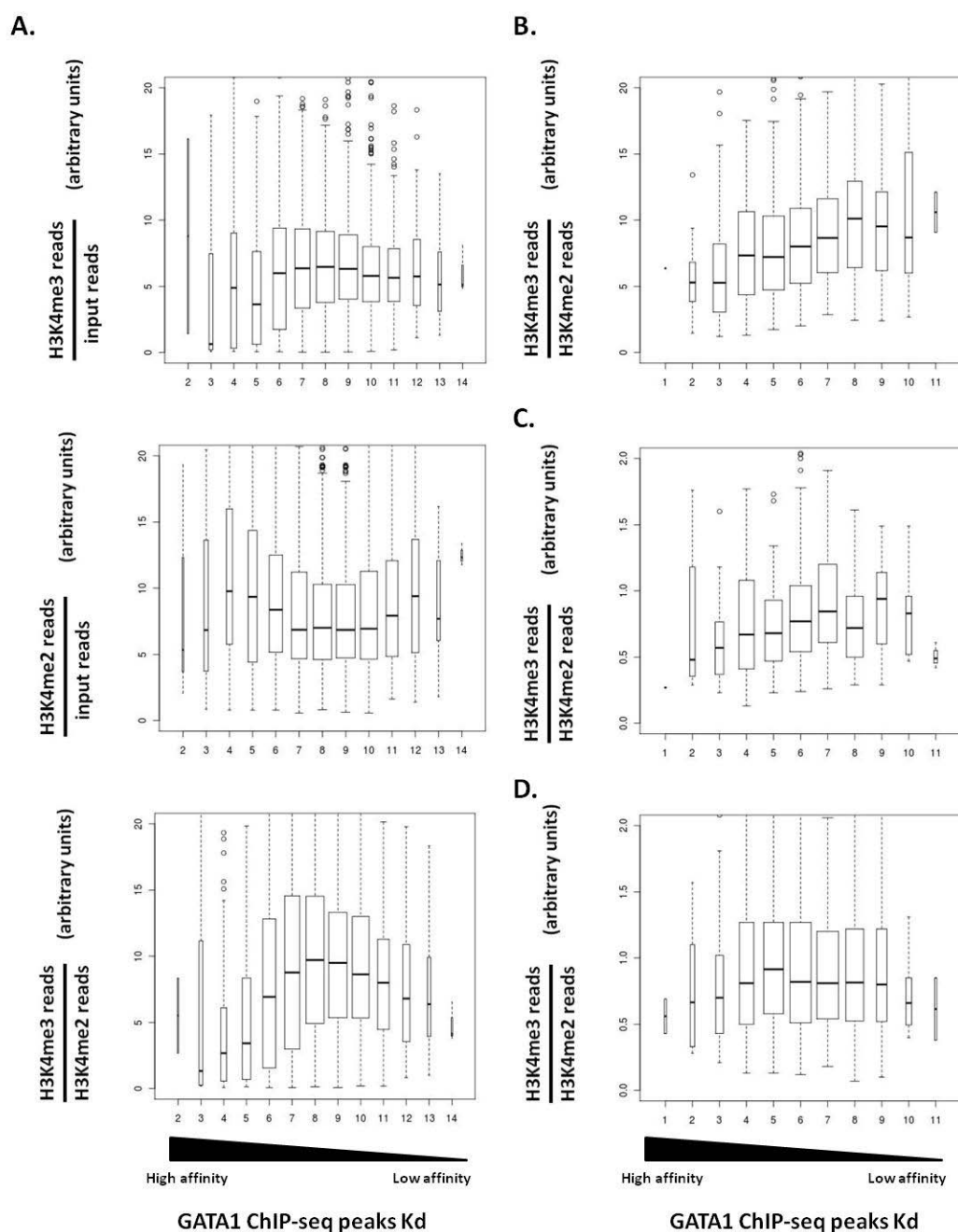


Figure 94. A) Relationship between the GATA1 ChIP-seq peak binding affinity (EMSA-seq binding affinity Kd scores) and the number of H3K4me3 reads on the nearest TSS in the CD36+ model (top graph). Only GATA1 ChIP-seq peaks that are less than 1 kb, upstream and downstream, of the TSS are considered (3404 peaks). The fraction of the H3K4me3 reads in the sample versus the reads in the input, that are 200 bp upstream or

downstream of the TSS, is presented. Similarly for the H3K4me2 reads (middle graph). Also the fraction of the H3K4me3 reads versus the H3K4me2 reads is presented (bottom graph). B) Relationship between the GATA1 ChIP-seq peak binding affinity ("Kd") and the fraction of the H3K4me3 reads versus the H3K4me2 reads on the nearest TSS (number of reads in the region 200 bp upstream or downstream of the TSS) for the common peaks (10 kb upstream and downstream of the TSS) in the TER119+/TER119- model. The same graph is presented for the TER119+ specific GATA1 ChIP-seq peaks (C) and the TER119- specific GATA1 ChIP-seq peaks (D). We used the EMSA-seq binding affinity Kd score to characterise the binding affinity of the GATA1 ChIP-seq peaks (x- axis). The ChIP-seq peaks were binned in 11/14 categories based on their EMSA-seq binding affinity Kd scores. The width of the box plot in each category is proportional to the number of ChIP-seq peaks in each category. The GATA1 ChIP-seq peaks of figure (A) were from (Hu et al., 2011) (GEO accession number: GSM651547) (cell type: human CD36+ cells derived from differentiated CD34+/CD113+ HSC cells). The H3K4me3 data, H3K4me2 data, H3K4me3 or H3K4me2 input data of figure (A) were from (Xu et al., 2012) (GEO accession number: GSM908039, GSM908037, GSM970263 respectively) (cell type: proerythroblasts (ProEs) from adult human bone marrow CD34+ cells). The GATA1 ChIP-seq peaks of figures (B), (C), (D) were from (Papadopoulos et al., 2013). The H3K4me3 data for TER119-, H3K4me2 data for TER119-, H3K4me3 data for TER119+ and H3K4me2 data for TER119+ of figures (B), (C), (D) were from (Wong et al., 2011b) (GEO accession number: GSM688818, GSM688817, GSM688809, GSM688808 respectively).

- All the TSS
- TSS that are near medium and low affinity GATA1 ChIP-seq peaks
- TSS that are near high affinity GATA1 ChIP-seq peaks
- Median value of the H3K4me3 (horizontal) and H3K4me2 (vertical) reads of all the TSS
- Median value of the H3K4me3 (horizontal) and H3K4me2 (vertical) reads of the TSS that are near medium and low affinity GATA1 ChIP-seq peaks
- Median value of the H3K4me3 (horizontal) and H3K4me2 (vertical) reads of the TSS that are near high affinity GATA1 ChIP-seq peaks

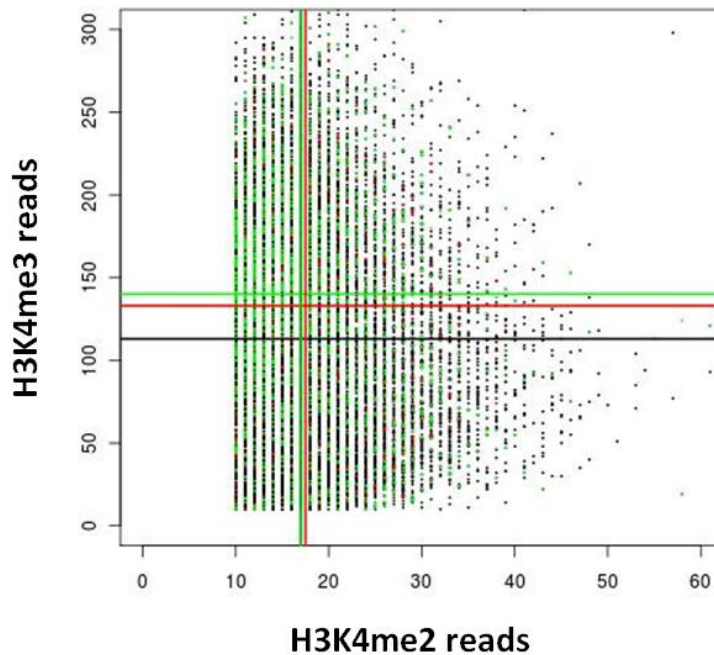


Figure 95. Scatter plot of the H3K4me3 and H3K4me2 reads that appear on the TSS of genes that belong to either one of the following three categories: i) All the TSS of genes, either coding or non coding, were downloaded from the RefSeq database (for the human genome assembly hg18, downloaded from <http://genome.ucsc.edu/>) (in the graph they correspond to black dots) (12674 values). Multiple isoforms of the same gene that have different TSS coordinates were kept whereas isoforms of the same gene that have the same TSS coordinates were considered only once. ii) TSS that are near medium and low affinity GATA1 ChIP-seq peaks (in the graph they correspond to green dots). The affinity of the GATA1 ChIP-seq peaks was assigned as the number of 8mers per ChIP-seq peak from the 2122 enriched mGATA1_{NC} 8mers. The low and medium affinity GATA1 ChIP-seq peaks are the ones with lower than 25 mGATA1_{NC} 8mers (2607 peaks). iii) TSS that are

near high affinity GATA1 ChIP-seq peaks (in the graph they correspond to red dots). The high affinity GATA1 ChIP-seq peaks are the ones with equal or higher than 25 mGATA1_{NC} 8mers (500 peaks). The TSS that are assigned near a low, medium or high affinity GATA1 ChIP-seq ($\pm$ 1 kb from a TSS) peak along with their respective H3K4me3 and H3K4me2 reads are the ones presented in Figure 94A. The horizontal and the vertical lines represent the median value of the H3K4me3 and H3K4me2 reads respectively for the three mentioned categories. TSS with at least 10 H3K4me3 and H3K4me2 reads were used in each category.

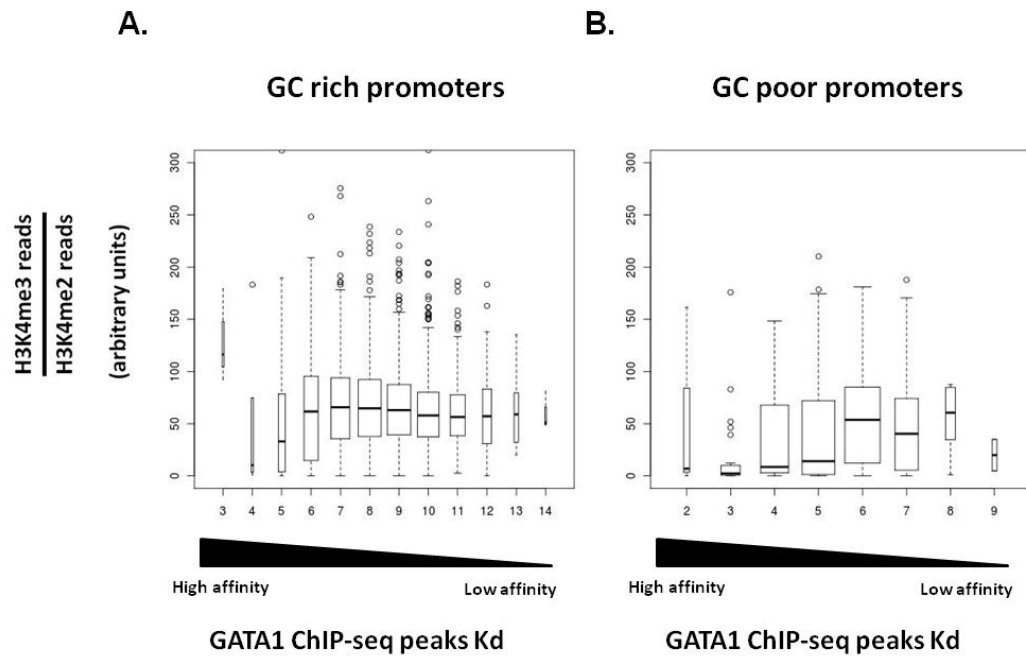


Figure 96. Relationship between the GATA1 ChIP-seq peak binding affinity (EMSA-seq binding affinity Kd scores) and the number of H3K4me3 and H3K4me2 reads on the nearest TSS in the CD36+ model. The figure is similar with the bottom graph of Figure 94A with the only difference that the TSS of Figure 94A have been separated in GC rich promoters (A) and GC poor promoters (B) (see main text).

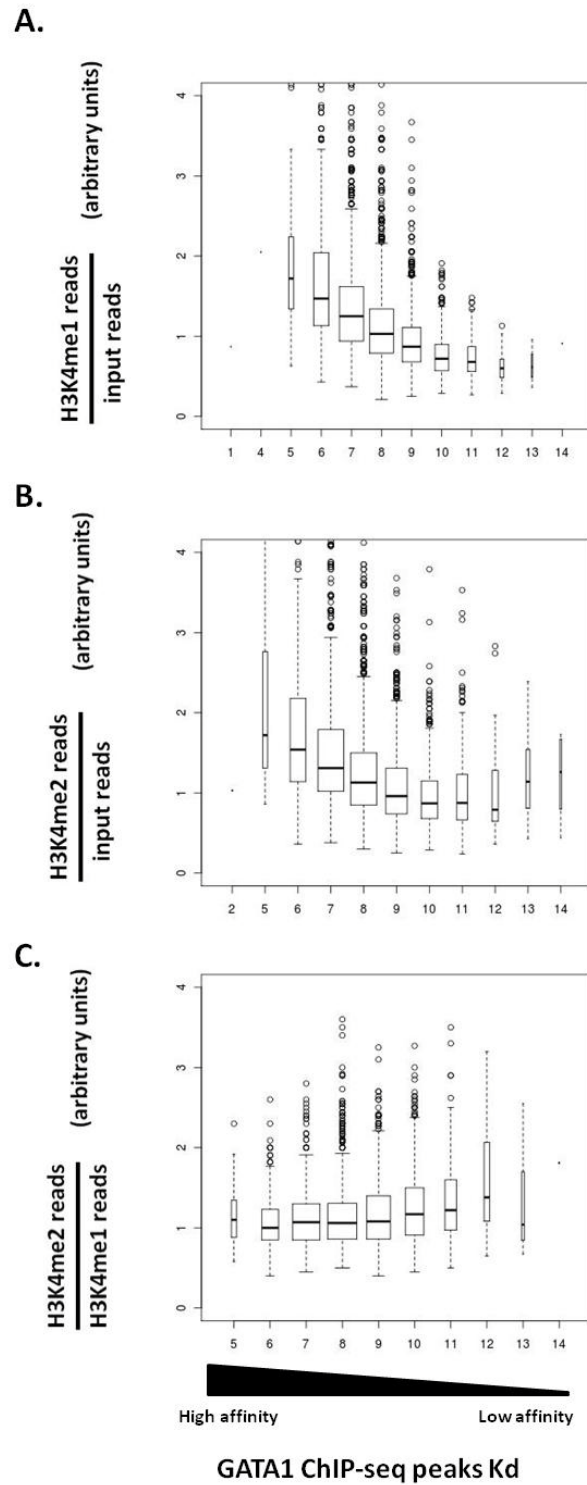


Figure 97. A) Relationship between the GATA1 ChIP-seq peak binding affinity (EMSA-seq binding affinity Kd scores) and the number of H3K4me1 reads on the GATA1 ChIP-seq peak binding sites in the CD36+ model. The fraction of the H3K4me1 reads in the sample versus the reads in the input, that are 200 bp upstream or downstream from the GATA1 ChIP-seq peak centre, is presented. Only GATA1 ChIP-seq peaks that are more than 10

kb, upstream and downstream, of a TSS are considered. Additionally, only GATA1 ChIP-seq peaks with more than 10 H3K4me1 reads are considered (7417 peaks). B) Similarly for the H3K4me2 reads. Again here only GATA1 ChIP-seq peaks with more than 10 H3K4me2 reads are considered (5940 peaks). C) The fraction of the H3K4me2 reads versus the H3K4me1 reads is presented. The figure is similar with (A). Only GATA1 ChIP-seq peaks with more than 10 H3K4me1reads and more than 10 H3K4me2 reads are considered (3973 peaks). We used the EMSA-seq binding affinity Kd score to characterise the binding affinity of the GATA1 ChIP-seq peaks (x- axis). The ChIP-seq peaks were binned in 14 categories based on their EMSA-seq binding affinity Kd scores. The width of the box plot in each category is proportional to the number of ChIP-seq peaks in each category. The GATA1 ChIP-seq peaks were from (Hu et al., 2011) (GEO accession number: GSM651547) (cell type: human CD36+ cells derived from differentiated CD34+/CD113+ HSC cells). The H3K4me1 data, H3K4me2 data, H3K4me1 or H3K4me2 input data of figure (A) were from (Xu et al., 2012) (GEO accession number: GSM908035, GSM908037, GSM970263 respectively) (cell type: proerythroblasts (ProEs) from adult human bone marrow CD34+ cells).

GATA1 or KLF1 ChIP-seq peaks +/- 1 kb from a TSS of a differentially expressed gene during erythroid differentiation

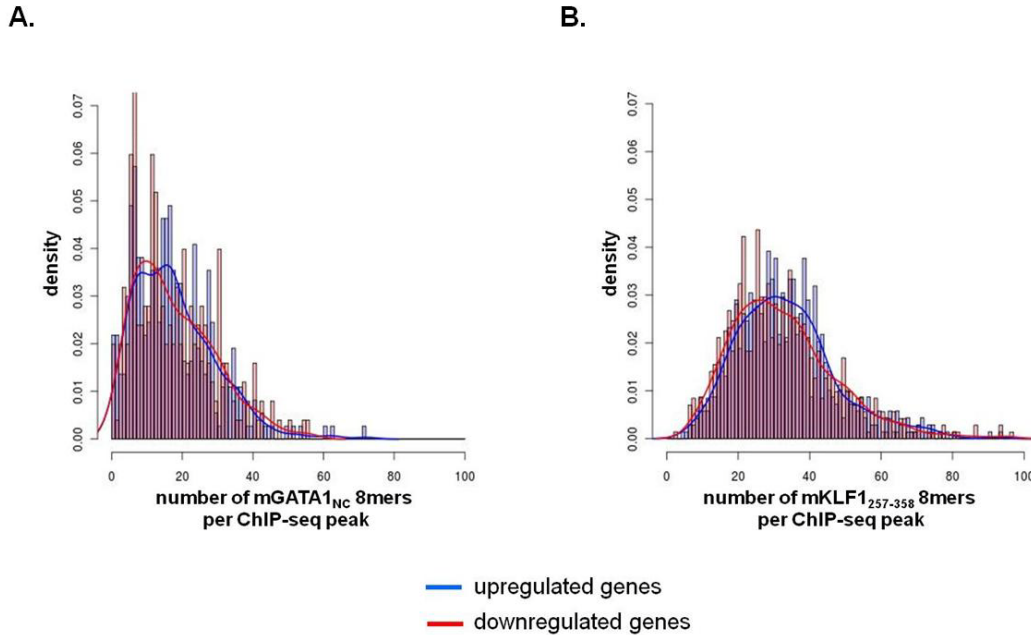


Figure 98. GATA1 ChIP-seq peak affinity distributions for the GATA1 ChIP-seq peaks that are 1 kb upstream or downstream from the TSS of the genes that show differential expression during erythroid differentiation (CD36+/CD34+ model). GATA1 ChIP-seq peaks for both the upregulated (blue histogram) and the downregulated (red histogram) genes are shown. The ChIP-seq peak affinity represents the number of 8mers per ChIP-seq peak from the 2122 enriched mGATA1_{NC} 8mers. B) Similarly for the KLF1 ChIP-seq peaks. The ChIP-seq peak affinity represents the number of 8mers per ChIP-seq peak from the 193 enriched mKLF1₂₅₇₋₃₅₈ 8mers. The GATA1 ChIP-seq peaks and the KLF1 ChIP-seq peaks of the CD36+ cells were used. The GATA1 ChIP-seq peaks were from (Xu et al., 2012) (GEO accession number: GSM970258) (cell type: proerythroblasts (ProEs) from adult human bone marrow CD34+ cells). The KLF1 ChIP-seq peaks were from (Su et al., 2013) (GEO accession number: GSM1067275) (cell type: R3/R4 erythroid cells derived from human bone marrow CD34+ cells). The differentially expressed genes

were from (Xu et al., 2012). Xu et al defined the A5 population of cells (adult human bone marrow CD34+ cells 5 days after erythroid differentiation, CD36+ cells) and the A0 population of cells (adult human bone marrow CD34+ cells 0 days after erythroid differentiation, CD34+ cells). The upregulated genes of the CD36+ cells were the ones that had high expression in the A5 cells and low expression in the A0 cells (A5_vs_A0 A5 high genes from (Xu et al., 2012)). The downregulated genes of the CD36+ cells were the ones that had high expression in the A0 cells and low expression in the A5 cells (A5_vs_A0 A0 high genes from (Xu et al., 2012)).

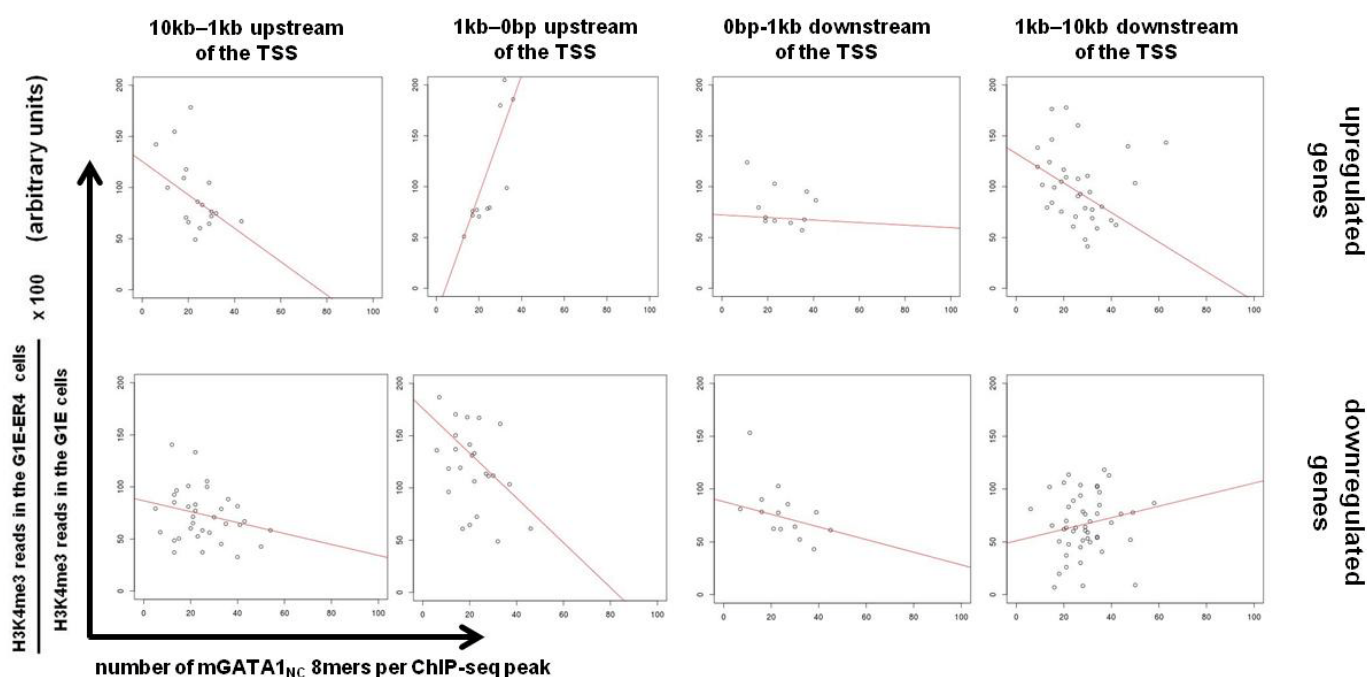


Figure 99. Relationship between the GATA1 ChIP-seq peak binding affinity and the change in the H3K4me3 levels for the genes that show differential expression during erythroid differentiation (G1E/G1E-ER4 model). GATA1 ChIP-seq peaks that are 10 kb upstream and downstream of the TSS are used. Both the upregulated and the downregulated genes are presented separately. The ChIP-seq peak affinity represents the number of 8mers per ChIP-seq peak from the 2122 enriched mGATA1_{NC} 8mers. The change in the levels of the H3K4me3 is presented as the ratio of the H3K4me3 reads in the G1E-ER4 cells versus the H3K4me3 reads in the G1E cells for the region 200 bp upstream and downstream of the TSS. The GATA1 ChIP-seq peaks are grouped, based on their distance from the closest TSS, in 4 categories (10 kb to 1 kb upstream of the TSS, 1 kb to 0 kb upstream of the TSS, 0 kb to 1 kb downstream of the TSS, 1 kb to 10 kb downstream of the TSS). We also note that we intended to express the H3K4me3 reads as the fraction of (number of H3K4me3 reads)/(number of input reads) in each of the G1E and G1E-ER4 conditions. Because in both the G1E and G1E-ER4 experiments, the

number of reads from the input that are assigned in each peak are less than 10 we assumed that the input library bias for each ChIP-seq position is negligible in either experiment. Thus, we only calculated the presented H3K4me3 ratio. The red line represents the median value trend. The GATA1 ChIP-seq peaks for the G1E-ER4 cells were from (Wu et al., 2011) (file: GATA1_G1EER4_replicatesCombined_peaks from <https://usegalaxy.org/library> (Data library name: Erythroid Epigenetic Landscape)). The H3K4me3 data for the G1E-ER4 cells were from (Wu et al., 2011) (file: H3K4me3_G1EER4_replicate1_mapped from <https://usegalaxy.org/library> (Data library name: Erythroid Epigenetic Landscape)). The H3K4me3 data for the G1E cells were from (Wu et al., 2011) (file: H3K4me3_G1E_replicate1a_mapped from <https://usegalaxy.org/library> (Data library name: Erythroid Epigenetic Landscape)). The RNA-seq derived differentially expressed genes were from (Wu et al., 2011) (file: g1e_er4_gene_exp from <https://usegalaxy.org/library> (Data library name: Erythroid Epigenetic Landscape))

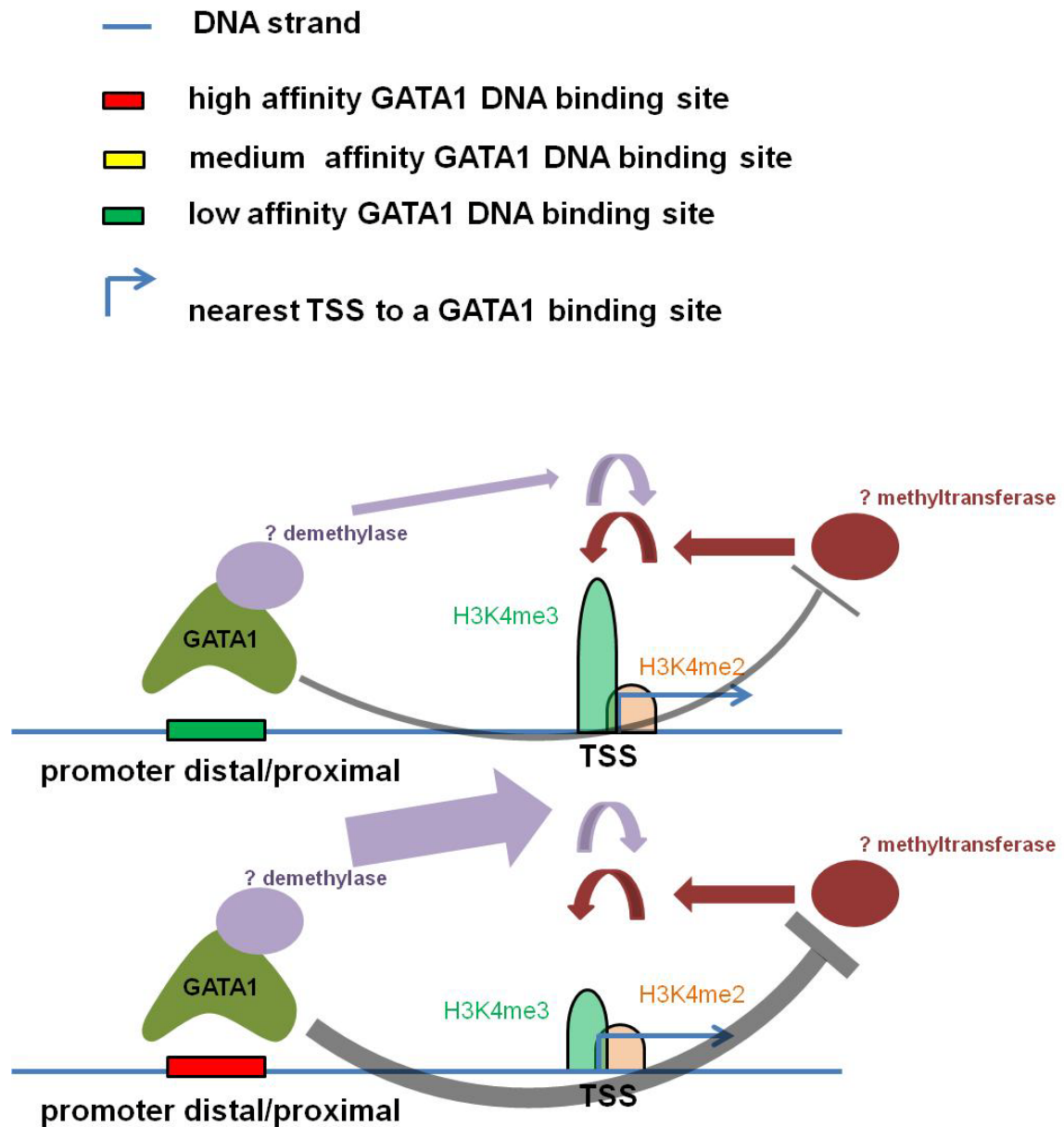


Figure 100. Model for the effect of GATA1 on the H3K4me3 levels of a nearby TSS. We hypothesize that GATA1 can affect directly the H3K4me3 levels on the nearest TSS. We can argue that GATA1 through an unknown mechanism (for example, a histone demethylase) shifts the H3K4me3 back to H3K4me2. The higher the DNA binding affinity of the GATA1 ChIP-seq peak, the more prominent this shift. The alternative model is that GATA1 indirectly controls the H3K4me3 deposition by inhibiting a prospective H3K4me2

methyltransferase (black inhibitory arrows). The two mechanisms can also occur simultaneously.

3.5.8.4 The same gene in different lineages, regulated by GATA1 from different genomic positions (10 kb +/- of the gene TSS) which have different binding affinities, shows a change in the H3K4me3 levels that corresponds to the binding affinity of the GATA1 ChIP-seq peak in the specific lineage

To further validate the effect of the GATA1 DNA binding affinities on the H3K4me3 levels, we used the G1ME megakaryopoiesis model and the G1E erythropoiesis model where GATA1 is involved in the differentiation of both lineages under different treatment protocols.

In the G1ME and the G1E model, GATA1 restoration is performed in the presence of thrombopoietin (Chlon et al., 2012) or erythropoietin (Wu et al., 2011) in order the cells to differentiate towards either the megakaryocyte lineage or the erythrocyte lineage respectively. In these two models, we checked whether there is a different distribution of ChIP-seq peak affinities in the GATA1 megakaryocyte specific peaks and the GATA1 erythrocyte specific peaks. We separated the ChIP-seq peaks in 3 categories: common erythrocyte/megakaryocyte peaks, erythrocyte specific peaks and megakaryocyte specific peaks.

We have already seen that the mean GATA1 ChIP-seq DNA binding affinity increases the further away from the TSS (Figure 76). For this reason, at first we explored whether the distribution of the GATA1 ChIP-seq peaks near the TSS is different between the erythrocyte and the megakaryocyte lineage. In Figure 101A we see that the GATA1 ChIP-seq peaks that appear in both lineages are more frequent near the TSS (less than 1000 bp upstream of TSS) (10.6% of the overlapped peaks versus 6.9% and 5.1% of the megakaryocyte and erythrocyte specific peaks respectively). The megakaryocyte specific peaks are more frequent in the distal intergenic regions (28.2% of the megakaryocyte

specific peaks versus 17.9% and 20.8% of the overlapped peaks and erythrocyte specific peaks respectively).

In Figure 102, we see the binding affinity distributions of the erythrocyte and megakaryocyte specific GATA1 ChIP-seq peaks. The binding affinity distribution of the erythrocyte specific GATA1 ChIP-seq peaks has overall higher values (more 8mers) than the binding affinity distribution of the megakaryocyte specific GATA1 ChIP-seq peaks (two-sided Kolmogorov-Smirnov (KS) test, $p\text{-value} < 2.2\text{e-}16$). In the previous paragraph, we saw that the megakaryocyte specific peaks are more frequent, the further away from the TSS, than the erythrocyte specific peaks. So, we would expect that they would show a tendency to have higher binding affinity values (**EMSA-seq k-mer number score**) than the erythrocyte specific ones. This is something that we do not observe. On the contrary, we saw that the erythrocyte specific peaks have overall higher values than the megakaryocyte specific peaks. This might indicate a general role of the ChIP-seq DNA binding affinity distribution in lineage specification.

In order to see whether this binding affinity distribution difference might have any physiological role, we focused on the common genes between the erythrocyte and the megakaryocyte GATA1 specific ChIP-seq peaks. As can be seen in Figure 101B we found 272 genes that are common in both categories. We then asked whether the peak k-mer distribution shift is retained in the common genes. As can be seen in figure 47B the peak k-mer distribution shift is much more prominent (two-sided Kolmogorov-Smirnov (KS) test, $p\text{-value} < 10^{-6}$). We have already associated the GATA1 high affinity peaks with a higher reduction in the H3K4me3 content on the TSS. We then focused on genes that showed a different GATA1 ChIP-seq peak binding site in the two different lineages 10 kb upstream of the promoter. We found 88 genes. We then tested the normalised H3K4me3 levels of these 88 common genes in cells that were differentiated towards an erythrocyte lineage versus cells that were differentiated towards a megakaryocyte lineage.

In Figure 103 we are presenting the scatter plots of the distribution of the normalised H3K4me3 signal value in both the megakaryocyte and the erythrocyte lineage of the G1ME and G1E models for the genes that are associated with megakaryocyte GATA1 specific ChIP-seq peaks, the genes that are associated with erythrocyte GATA1 specific ChIP-seq peaks and the selected 88 genes that have GATA1 ChIP-seq peaks in both megakaryocytes and erythrocytes. We also show the contour plots (which point toward the centre of the 2d density) of each of the scatter plots for the 3 categories. We see that the centre of the contour plots (the centre of the concentric circles) for the genes that are associated with megakaryocyte GATA1 specific ChIP-seq peaks (blue circles) and the genes that are associated with erythrocyte specific GATA1 ChIP-seq peaks (red circles) are distinct.

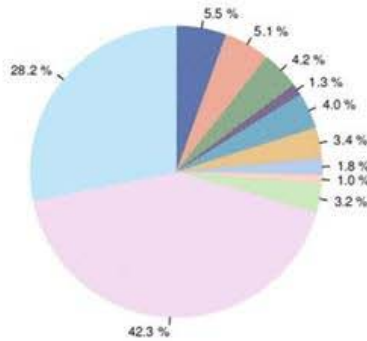
We also assume that the presence of GATA1 affects the total levels of H3K4me3 on a given TSS in a unimodal and unidirectional fashion (the higher the GATA1 ChIP-seq binding affinity, the lower the H3K4me3 levels). The alternative is to assume a bimodal role for GATA1 where the presence of GATA1 at low levels increases the H3K4me3 levels initially but eventually high levels of GATA1 leads to low H3K4me3 levels. If GATA1 was operating in a unimodal role then we would expect the blue and red contour centres to lie anti-parallel with the main diagonal. On the other hand if GATA1 was operating in a bimodal role and considering that the overall affinity distributions of the megakaryocyte GATA1 specific ChIP-seq peaks are lower than the ones from the erythrocyte specific GATA1 ChIP-seq peaks, then the blue and red contour centres are expected to lie in parallel with the diagonal, which is the case. Because the GATA1 ChIP-seq peaks of the common genes in the erythroid lineage have binding affinity peaks that are of higher values than the GATA1 ChIP-seq peaks of the common genes in the megakaryocyte lineage, then we expect the common genes to be much more similar with the scatter plot distributions of the erythroid specific GATA1 ChIP-seq peaks rather than the megakaryocyte specific GATA1 ChIP-seq peaks. Indeed in Figure 103 we can see that

the centre of the green contour lies near the centre of the red contour lines and far away from the centre of the blue contour lines. **Thus the dependence of the H3K4me3 levels on the binding affinity of GATA1 has a physiological role.** We also note here that a bimodal role for GATA1 in the regulation of the *Tac-2* gene expression, in an erythroid differentiation model, has been described in the literature (Johnson et al., 2006). Specifically the initial activation and the subsequent repression, during the erythroid differentiation process, of the *Tac-2* gene expression have been shown to depend on the GATA1 activity. Although FOG1 is present in the GATA1 binding site from the beginning of the differentiation process, only at the last stages of the differentiation the FOG1 repressing activity and the increased eviction of the GATA1/FOG1 complex from the DNA manifest in the repression of the *Tac-2* gene expression

A.

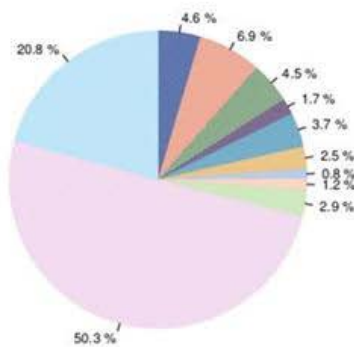
GATA1 ChIP-seq peaks genomic distribution

megakaryocyte specific GATA1 ChIP-seq peaks



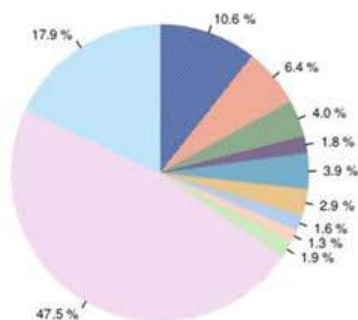
- Promoter (≤ 1000 bp): 5.5 %
- Promoter (1000–5000 bp): 5.1 %
- Promoter (5000–10000 bp): 4.2 %
- Downstream (≤ 1000 bp): 1.3 %
- Downstream (1000–5000 bp): 4.0 %
- Downstream (5000–10000 bp): 3.4 %
- 5'UTR: 1.8 %
- 3'UTR: 1.0 %
- Coding exon: 3.2 %
- Intron: 42.3 %
- Distal intergenic: 28.2 %

erythrocyte specific GATA1 ChIP-seq peaks



- Promoter (≤ 1000 bp): 4.6 %
- Promoter (1000–5000 bp): 6.9 %
- Promoter (5000–10000 bp): 4.5 %
- Downstream (≤ 1000 bp): 1.7 %
- Downstream (1000–5000 bp): 3.7 %
- Downstream (5000–10000 bp): 2.5 %
- 5'UTR: 0.8 %
- 3'UTR: 1.2 %
- Coding exon: 2.9 %
- Intron: 50.3 %
- Distal intergenic: 20.8 %

common megakaryocyte and erythrocyte GATA1 ChIP-seq peaks



- Promoter (≤ 1000 bp): 10.6 %
- Promoter (1000–5000 bp): 6.4 %
- Promoter (5000–10000 bp): 4.0 %
- Downstream (≤ 1000 bp): 1.8 %
- Downstream (1000–5000 bp): 3.9 %
- Downstream (5000–10000 bp): 2.9 %
- 5'UTR: 1.6 %
- 3'UTR: 1.3 %
- Coding exon: 1.9 %
- Intron: 47.5 %
- Distal intergenic: 17.9 %

B.

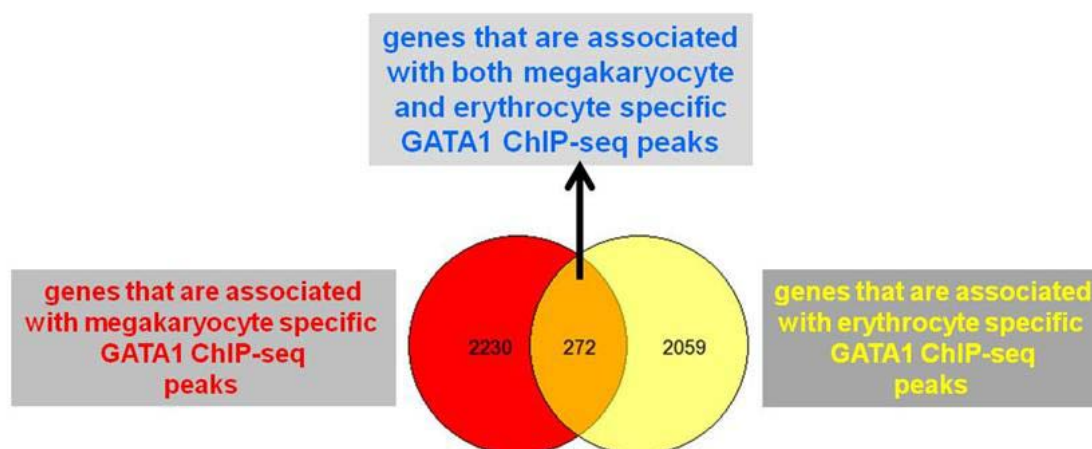
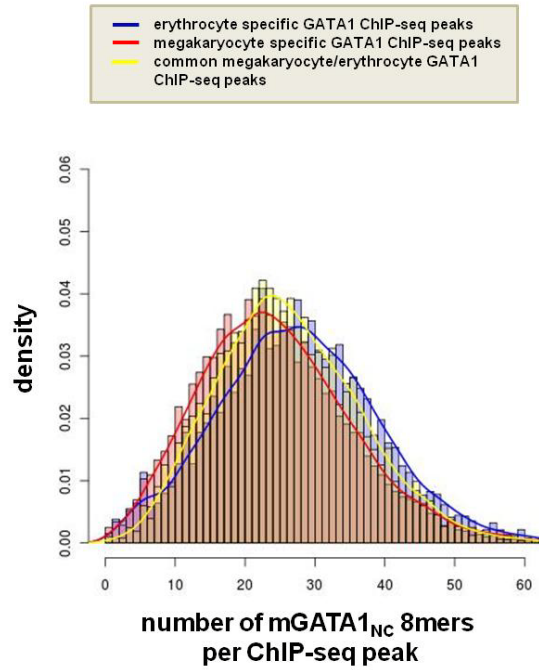


Figure 101 (previous and current page). A) Genomic distribution of the GATA1 ChIP-seq peaks in the G1E/G1ME model. GATA1 ChIP-seq peaks from the erythroid differentiation model (G1E/G1E-ER4) were overlapped with GATA1 ChIP-seq peaks from the megakaryocytic differentiation model (G1ME/GATA1 restored G1ME). The GATA1 ChIP-seq peaks were separated in 3 categories. The GATA1 ChIP-seq peaks that appear only in the erythroid differentiation model (erythrocyte specific), the GATA1 ChIP-seq peaks that appear only in the megakaryocytic differentiation model (megakaryocyte specific), the common GATA1 ChIP-seq peaks in both models. The genomic distribution of the ChIP-seq peaks in these three categories is shown. B) Overlap between the genes that are associated with the erythrocyte or megakaryocyte specific GATA1 ChIP-seq peaks (less than 10 kb upstream or downstream from the TSS). The GATA1 ChIP-seq peaks for the G1E-ER4 cells were from (Wu et al., 2011) (file: GATA1_G1EER4_replicatesCombined_peaks from <https://usegalaxy.org/library> (Data library name: Erythroid Epigenetic Landscape)). The GATA1 ChIP-seq peaks for the GATA1 restored G1ME cells were from (Dore et al., 2012) (GEO accession number: GSM777092).

A.



B.

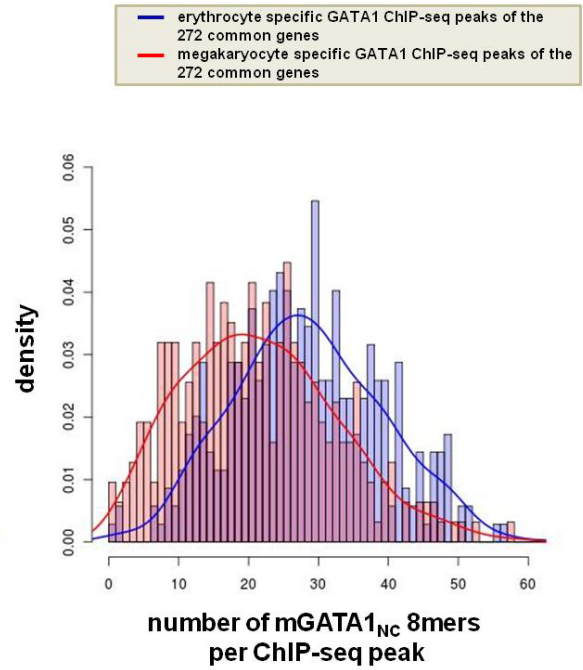


Figure 102. GATA1 ChIP-seq peak affinity distributions in the G1E/G1ME model under different conditions of differentiation. The GATA ChIP-seq peak categories are the ones used in Figure 101A. B) GATA1 ChIP-seq peak affinity distributions for the peaks that are associated with the common 272 genes from Figure 101B. Although the genes are the same the GATA1 ChIP-seq peaks are not the same. In both images, the ChIP-seq peak affinity represents the number of 8mers per ChIP-seq peak from the 2122 enriched mGATA1_{NC} 8mers.

- genes that are associated with erythrocyte specific GATA1 ChIP-seq peaks (± 10 kb)
- genes that are associated with megakaryocyte specific GATA1 ChIP-seq peaks (± 10 kb)
- genes that are associated with a GATA1 ChIP-seq peak in both lineages and the GATA1 ChIP-seq peak binding site, for the region 10kb upstream of the TSS, is different in the two lineages

- contour plot for the red group of genes
- contour plot for the blue group of genes
- contour plot for the green group of genes
- * centre of density for the red group of genes
- * centre of density for the blue group of genes
- * centre of density for the green group of genes

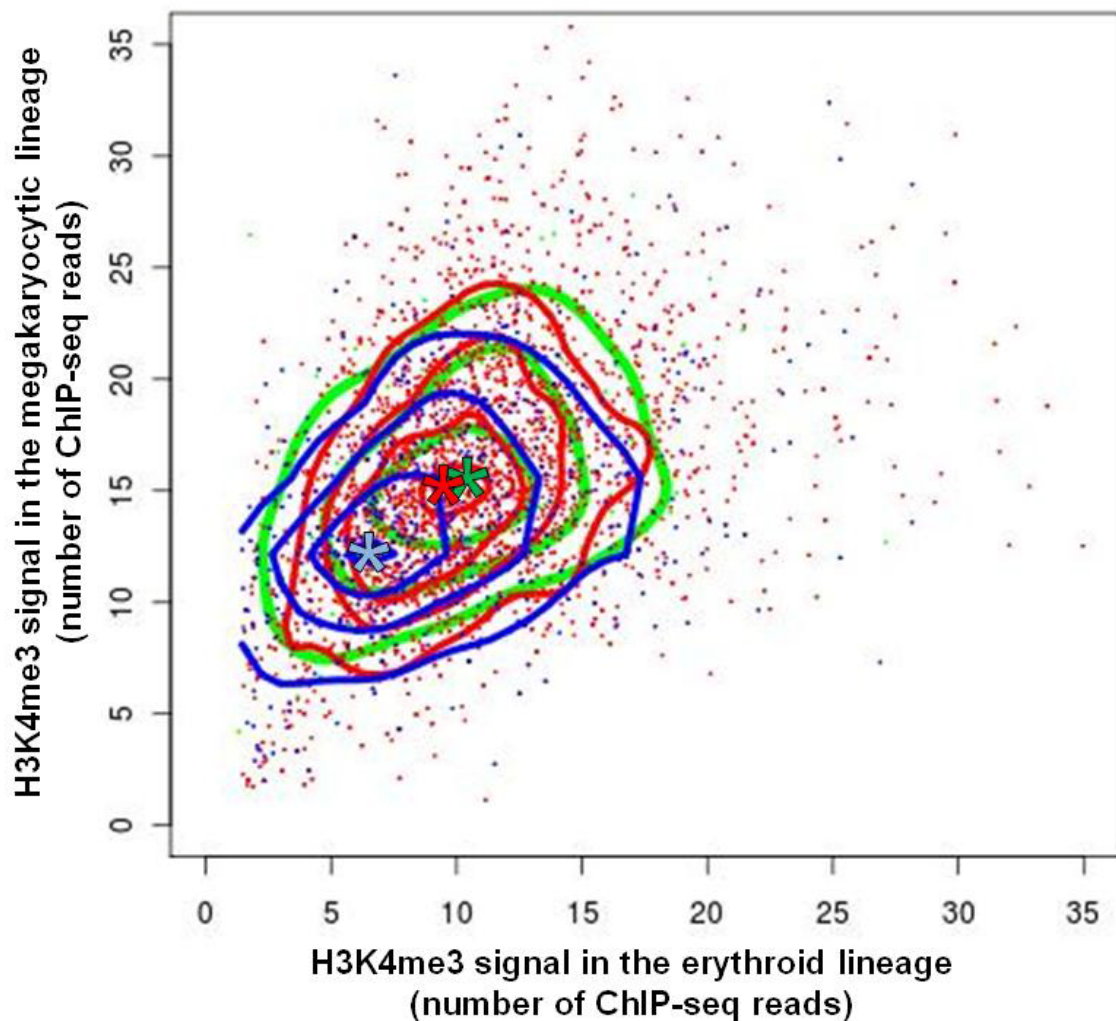


Figure 103. Scatter plot of the H3K4me3 signal, in the erythroid and the megakaryocytic lineage, for genes that are associated with a GATA1 ChIP-seq peak in either of the two lineages for the G1E/G1ME model. The genes are grouped in the ones that are

associated with erythrocyte specific GATA1 ChIP-seq peaks (red group) and the ones that are associated with megakaryocyte specific GATA1 ChIP-seq peaks (blue group) as have been presented in Figure 101A. A third group of genes is also presented. This group concerns genes (88 genes) that showed a different GATA1 ChIP-seq peak binding site in the two different lineages 10 kb upstream of the promoter (green group). The scatter plot shows the H3K4me3 signal on the TSS (+/- 200bp). Contour lines that indicate the position of the centre of density (asterisks) are also presented. The centre of density for the genes that are associated with the erythrocyte specific GATA1 ChIP-seq peaks and the megakaryocyte specific GATA1 ChIP-seq peaks is different (red versus blue asterisk). The distance difference between the red and the blue centre is statistically significant (1000 permutation samples, p-value=0.03*). On the other hand, the distance difference between the centre of density for the 88 common genes (green asterisk) and the centre of density for the genes that are associated with the erythrocyte specific GATA1 ChIP-seq peaks (red asterisk) is not statistically significant (1000 permutation samples, p-value=0.85). The GATA1 ChIP-seq peaks were the ones used in Figure 101. The H3K4me3 data for the G1E-ER4 cells were from the ENCODE/PSU project (GEO accession number: GSM777094 file: GSM946519_mm9_wgEncodePsuHistoneG1eer4e2H3k04me3ME0S129InputPk.broadPeak). The H3K4me3 data for the GATA1 restored G1ME cells were from (Dore et al., 2012) (GEO accession number: GSM777094 file: GSM777094_H3K4me3_A_peaks).

3.5.8.5 The GATA1/GATA2 switch occurs in sites where the GATA1 DNA binding affinity is higher than the GATA2 binding affinity

GATA2 is mainly expressed in hematopoietic stem cells, multipotent hematopoietic progenitors and erythroid precursors (Tsai et al., 1994, Tsai and Orkin, 1997). It can also be found in megakaryocytes, eosinophils and mast cells (Tsai and Orkin, 1997, Minegishi et al., 1999). During the erythroid maturation GATA2 levels slowly decline whereas GATA1 levels increase. GATA1 represses the expression of the GATA2 gene. GATA1 displaces GATA2 (GATA switch) from the promoter of the GATA2 gene by binding in the same position where GATA2 binds and this eventually shuts down the autoregulatory loop of GATA2 expression (Grass et al., 2003). Apart from the lack of the autoregulation loop, GATA1 mediates directly the repression of the GATA2 gene (Hong et al., 2005). This GATA switch happens in multiple loci (Dore et al., 2012).

We tried to explain the GATA1/GATA2 switch behaviour. We used a dataset derived from (Dore et al., 2012). The authors used ChIP-seq data from GATA2 in the undifferentiated G1ME cells (multipotent precursors) and from GATA1 in the differentiated into megakaryocytes G1ME cells. The authors then separated the GATA1 and GATA2 ChIP-seq peaks into 3 groups:

- i) Peaks where GATA1 and GATA2 overlaps. The authors claim that these are peaks that show a GATA1/GATA2 switch during the differentiation process.
- ii) GATA1 ChIP-seq peaks that do not show GATA2 in the undifferentiated state (GATA1 specific ChIP-seq peaks).
- iii) GATA2 ChIP-seq peaks that do not show GATA1 in the differentiated state (GATA2 specific ChIP-seq peaks).

Although the GATA2 ChIP-seq peaks were derived from an undifferentiated state and the GATA1 ChIP-seq peaks from the differentiated state this does not exclude the existence

of GATA2 peaks in the differentiated state as has been shown during the differentiation of the erythropoietic lineage (May et al., 2013).

Because we do not have GATA2 binding specificities we used the binding specificities of the protein with the closest similarity to GATA2 from the GATA family of proteins. In Figure 104A, B we see that the GATA3 protein sequence is more similar to GATA2 rather than GATA1. In Figure 104C we present per ChIP-seq peak the ratio of the binding affinity of the GATA1 protein versus the binding affinity of the GATA2 protein. The binding affinity of the GATA1 protein is expressed as the number of 8mers from the 2122 enriched mGATA1_{NC} 8mers. The binding affinity of the GATA2 protein is expressed as the number of 8mers from the top 2122 enriched mGATA3 8mers (Jolma et al., 2013). The higher the ratio the more probable the ChIP-seq peak to be bound from GATA1 (this assumes that both GATA1 and GATA2 are in equimolar concentrations, which is not exactly true during the differentiation of the megakaryocytic lineage). The higher the ratio the more probable the ChIP-seq peak to be bound from GATA1 (GATA1 has a higher binding affinity for this position than GATA2). The lower the ratio the more probable the ChIP-seq peak to be bound from GATA2 (GATA2 has a higher binding affinity for this position than GATA1). In Figure 104C we see that the GATA2 specific positions have a higher GATA2 binding affinity (lower median ratio) than the GATA1 specific positions (higher median ratio) (permutation test, empirical p-value=0). The GATA1 specific positions have a higher GATA1 binding affinity than the GATA2 specific positions. On the contrary, the overlapping GATA1/GATA2 positions have a higher GATA1 binding affinity than GATA2 binding affinity (lower median ratio) (permutation test, empirical p-value=0). This probably explains the ability of GATA1 to compete and displace GATA2 from the GATA1/GATA2 switch regions. In the GATA2 specific regions, GATA2 can resist the competition because it has higher binding affinity. This is in accordance with experiments where proteasome inhibitors, that stabilise GATA2 and permit the co-existence of high levels of GATA2 and GATA1 inside the cell, impair GATA switches (Lurie et al., 2008). This role of the DNA

sequence in the GATA1 switch, as has been presented above, is probably complementary with the indispensable role of the GATA1 co-factor, FOG1, in the GATA switch. FOG1 either facilitates the chromatin occupancy of GATA1 (Pal et al., 2004a) or mediates the GATA switch through the nucleosome remodelling activity of its NuRD co-factor (Gao et al., 2010).

A.

```

sp|P23772|GATA3_MOUSE MEVTADQPRWVSHHHPAVLNGQHPDTHHPGLGHSYME-AQYPLTEEVDVL 49
sp|009100|GATA2_MOUSE MEVAPEQPRWMAH--PAVLNAQHPDSHHPGLAHNYMEPAQLLPDEVDVF 48
***:.;****:.* *****.****:*****.*,*** ** .:****:

sp|P23772|GATA3_MOUSE FN-IDGQGNHVPSSYGNVSRATVQR-YPPTH--GSQVCRPPLLHG-SL 93
sp|009100|GATA2_MOUSE FNHLDSQGN---PYANPAHARARVSYSPAHARLTGGQMCRLHLLHSPGL 95
** :*,*** .*,*.:* .: *,*: * ,*:*** ***, .*

sp|P23772|GATA3_MOUSE PWLDGGKALSS---HHTASPNLSPFSKTSIH---HGSPG-PLSVYPPAS 136
sp|009100|GATA2_MOUSE PWLDGGKALSAHAHHSPMTVSPFSKTPHPSAAGAPGGPLSVYPPGAA 145
***** * * ***,*****:,* *: ** ***** *:

sp|P23772|GATA3_MOUSE SSS-----LAAGHSSPHLFTFPPTPKDVSPPDPSLS---TPGS 171
sp|009100|GATA2_MOUSE GSGGGGSGSSVASLTPTAAHSGSHLFGFPPTPKDVSPPDPTTGAASPAS 195
.,* :*,**.,*** *****:***** : ,*,*

sp|P23772|GATA3_MOUSE AGS---ARQDEKECLKYQVQLPDSMKLET-SHRSMTTLGGASSSAHH 216
sp|009100|GATA2_MOUSE SSAGGSVARGEDKDGKYQVSLSESMKMEGGSPLRPLATMG-TQPATH 244
.: ** :*: :****,*,****: * * :*: * :.:**

sp|P23772|GATA3_MOUSE PITTYPPYVP---EYSSGLFPPSSLLGGSPTFGCKSRPKARSSTEGRE 262
sp|009100|GATA2_MOUSE PIPTYPYVPAAHHDYGSLLFHPGGFLGGPASFTPKQSKARSCSEGE 294
** ,***,*** :*,*,** :.:****.*** * ,*,****.:****

sp|P23772|GATA3_MOUSE CVNCGATSTPLWRDGTGHYLCNACGLYHKMNGQNRPLIKPKRRLSAARR 312
sp|009100|GATA2_MOUSE CVNCGATATPLWRDGTGHYLCNACGLYHKMNGQNRPLIKPKRRLSAARR 344
*****:*****:*****:*****:*****

sp|P23772|GATA3_MOUSE AGTSCANCQTTTTLWRRNANGDPVCNACGLYKLNINRPLTMKKEGIG 362
sp|009100|GATA2_MOUSE AGTCCANCQTTTTLWRRNANGDPVCNACGLYKLNINRPLTMKKEGIG 394
***,*****:*****:*****:*****:*****

sp|P23772|GATA3_MOUSE TRNRKMSKSKKCKKVHDALEDFP---KSSSFNPAALSRRHSSLSHIS 407
sp|009100|GATA2_MOUSE TRNRKMSKSKKCKKGAECFEELSKCMQEKSPFSAALAGHMAPVGHLP 444
*****:*** :.:***. **.,*.:***: **.:*.:*

sp|P23772|GATA3_MOUSE PFSHSSHMLTTPPMHPPSGLSFGPHHPSSMVTAMG 443
sp|009100|GATA2_MOUSE PFSHSHGILPTPTPIHPSSSLSGFHPHPSSMVTAMG 480
*****:*,*.:****:*,*.:*** *****

```

B.

```

sp|P23772|GATA3_MOUSE MEVTADQPRWVSHHHPAVLNGQHPDTHHPGLGHSYMEAQYPLTEEVDVLF 50
sp|P17679|GATA1_MOUSE MDFFGLGALGTSEPLQFVDSALVSSPSDSTGFFSSGPEGLDAASSTSP 50
*:... .*. * :.: .: .*. .: .:..

sp|P23772|GATA3_MOUSE NIDGQGNHVPSSYGNVSRATVQRYPPTHGSGVCRPPLLHGSPLWLDGGK 100
sp|P17679|GATA1_MOUSE NAATAAASALAY-----REAEAYRHSVPVQVYPLNMEGIPGG- 90
* . . :.* * .: : * : * : * : * :

sp|P23772|GATA3_MOUSE ALSSHHTASPNLSPFSKTSIHGSPGPLSVYPPASSSSSLAAGHSSPHLF 150
sp|P17679|GATA1_MOUSE -----SPYASWAYGKTALYPAS----- 107
** : .:***:.*

sp|P23772|GATA3_MOUSE TFPPTPKDVSPPDPSLSTPGSAGSARQDEKECLKYQVQLPDSMKLETSHS 200
sp|P17679|GATA1_MOUSE --TVCPSHEDAPSQALEDQ--EGKSMNTFLDTLTKTERLSPDLLTGLTALP 153
. *: :*: :*. *: : : ** : ** :* *: .

sp|P23772|GATA3_MOUSE RGSMTTLGGASSSAHHPITTYPPYVFEYSSGLFPPSSLLGGSPTFGCKS 250
sp|P17679|GATA1_MOUSE -ASLPVTGSAYGGADFPSPFFSPTGSLSSAAYSSPKFHGSLP----- 195
.:** ,* .*. * .: * . **.,*.:**

sp|P23772|GATA3_MOUSE RPKARSSTEGRECVNCGATSTPLWRDGTGHYLCNACGLYHKMNGQNRPL 300
sp|P17679|GATA1_MOUSE ----LAPCEARECVNCGATATPLWRDRDTGHYLCNACGLYHKMNGQNRPL 241
.: *,*****:***** *****:*****

sp|P23772|GATA3_MOUSE IKPKRRLSAARRAGTSCANCQTTTTLWRRNANGDPVCNACGLYKLNINR 350
sp|P17679|GATA1_MOUSE IRPKRMIVSKRAGTQCTNCQTTTTLWRRNASGDPVCNACGLYKLNINR 291
*:**:* :.:****,*:*****:*****:*****:*****

sp|P23772|GATA3_MOUSE NRPLTMKKEGIQTRNRKMSKSKKCKKVHDALEDFPKSSS----- 390
sp|P17679|GATA1_MOUSE NRPLTMKDKGIQTRNRKASGKGGKRGSLAGAGAAEGPAGGFMVAGSS 341
*****:***** *,*.: * : * .:..

sp|P23772|GATA3_MOUSE -----FNPAALSRRHSSLSHISPFSSHMLTTPPMH--PP 425
sp|P17679|GATA1_MOUSE SSGNCGEASGLALGTAGTAHYQLGLPVVLSGPVSHLMPFPGPLLGSP 391
.:*. :. :*. : . **.: * *: *

sp|P23772|GATA3_MOUSE SGLSFGPHHPSSMVTAMG---- 443
sp|P17679|GATA1_MOUSE TSFPTGPAITTSSTSVIAPLSS 413
.:* ** .:* :.:

```

C.

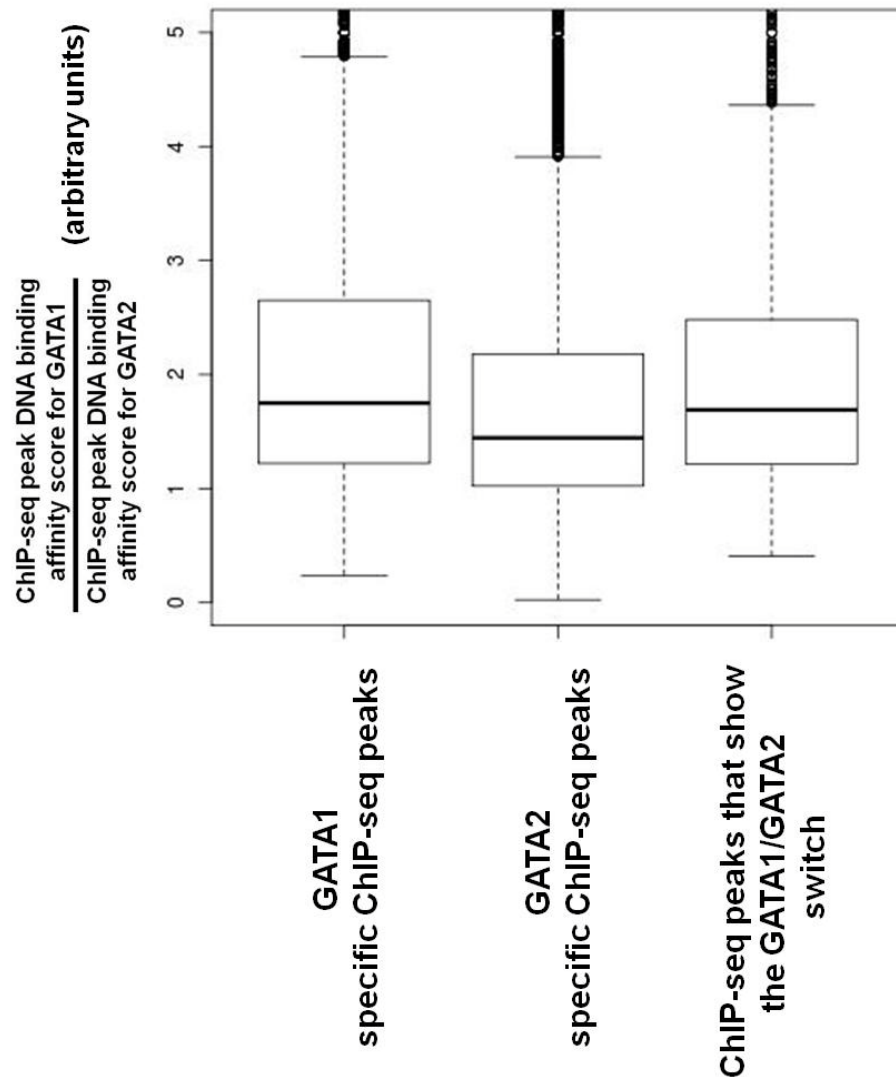


Figure 104 (previous and current page). A) Alignment of the protein sequences of mouse GATA2 and GATA3. The pink line represents the GATA3 peptide fragment used in the experiments of (Jolma et al., 2013). The green line represents the N-terminal and the C-terminal Zn finger of GATA2. B) Alignment of the protein sequences of mouse GATA1 and GATA3. The blue line represents the GATA1 peptide fragment of mGATA1_{NC}. The red line represents the N-terminal and the C-terminal Zn finger of GATA1. GATA1 has lower similarity with GATA3 than GATA2 in the region of the DNA binding domains (N-terminal Zn finger, C-terminal Zn finger, linker region between the two Zn fingers, C-terminal region after the C-terminal Zn finger). Due to the high similarity between the GATA2 and the

GATA3 DNA binding domains, the GATA3 enriched 8mers from (Jolma et al., 2013) were used as the hypothetical 8mers that GATA2 prefers to bind. C) Distributions of the ratio of the DNA binding affinity score for a given ChIP-seq peak for GATA1 versus the DNA binding affinity score for the same ChIP-seq peak for GATA2 (for details see text). The GATA1 and the GATA2 ChIP-seq peaks were separated in three groups: i) ChIP-seq peaks that are bound by both GATA1 in the differentiated state and GATA2 in the undifferentiated state. These are the peaks that GATA1 substitutes GATA2 during the differentiation (GATA1/GATA2 switch) (4799 peaks), ii) GATA1 ChIP-seq peaks that do not show GATA2 in the undifferentiated state (GATA1 specific ChIP-seq peaks) (7894 peaks), iii) GATA2 ChIP-seq peaks that do not show GATA1 in the differentiated state (GATA2 specific ChIP-seq peaks) (13261 peaks).

3.5.8.6 Low and high affinity GATA1 ChIP-seq peaks control different functional categories of genes

We also examined whether the GATA1 ChIP-seq binding affinities can classify the associated genes into different functional categories. We separated the GATA1 ChIP-seq peaks into low, medium and high affinity peaks. We took the first 800, middle 800 and bottom 800 peaks respectively. In Figure 105 we can see that the low affinity GATA1 ChIP-seq peaks are associated mainly with signal transduction pathways (erythropoietin/JAK-STAT signalling), the medium affinity GATA1 ChIP-seq peaks are associated with pathways of erythroid metabolism (heme biosynthetic pathway) whereas the high-affinity GATA1 ChIP-seq peaks are associated with housekeeping processes (glutathione maintenance) either erythroid specific or not. In the heme biosynthetic pathway, although a lot of the enzymes are regulated from medium affinity GATA1 ChIP-seq peaks, we can find either the same or different enzymes to be additionally regulated from low and high affinity GATA1 ChIP-seq peaks (Figure 106). This combinatorial regulation can probably create different rate limiting steps for the metabolic pathways.

biological processes enriched

pathways enriched

genes associated with low DNA binding affinity GATA1 ChIP-seq peaks

Term	P-Value
regulation of transcription	1.40E-006
regulation of RNA metabolic process	2.50E-006
regulation of transcription, DNA-dependent	4.00E-006
positive regulation of macromolecule biosynthetic process	4.50E-006
positive regulation of macromolecule metabolic process	4.90E-006
positive regulation of nitrogen compound metabolic process	5.90E-006
positive regulation of biosynthetic process	6.30E-006
positive regulation of cellular biosynthetic process	8.40E-006
transcription	1.10E-005
positive regulation of transcription	1.40E-005

Term	P-Value
Wnt signaling pathway	5.70E-003
Jak-STAT signaling pathway	1.70E-002
mTOR signaling pathway	3.70E-002
Cell cycle	3.80E-002
Steroid biosynthesis	4.60E-002
Renal cell carcinoma	5.00E-002
Chronic myeloid leukemia	6.80E-002
N-Glycan biosynthesis	6.90E-002
Porphyrin and chlorophyll metabolism	7.20E-002
Neurotrophin signaling pathway	7.50E-002
T cell receptor signaling pathway	7.50E-002
Oocyte meiosis	8.20E-002

genes associated with medium DNA binding affinity GATA1 ChIP-seq peaks

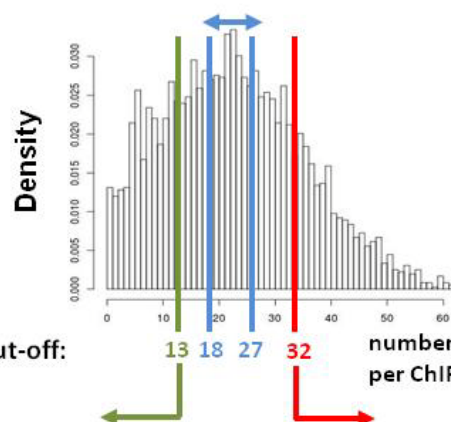
Term	P-Value
porphyrin metabolic process	9.90E-004
tetrapyrrole metabolic process	9.90E-004
heme metabolic process	1.70E-003
heme biosynthetic process	4.40E-003
cytoskeleton organization	7.40E-003
porphyrin biosynthetic process	8.10E-003
tetrapyrrole biosynthetic process	8.10E-003
hemopoiesis	8.80E-003
immune system development	1.00E-002
response to peptide hormone stimulus	1.10E-002

Term	P-Value
Porphyrin and chlorophyll metabolism	3.20E-003
Insulin signaling pathway	7.90E-003
Adipocytokine signaling pathway	1.00E-002
Glycosphingolipid biosynthesis	2.80E-002
Non-small cell lung cancer	3.40E-002
Renal cell carcinoma	3.80E-002
Acute myeloid leukemia	4.60E-002

genes associated with high DNA binding affinity GATA1 ChIP-seq peaks

Term	P-Value
positive regulation of translation	2.60E-004
regulation of I-kappaB kinase/NF-kappaB cascade	4.40E-004
positive regulation of cellular protein metabolic process	1.50E-003
nucleocytoplasmic transport	1.90E-003
gene silencing	1.90E-003
nuclear transport	2.20E-003
positive regulation of protein metabolic process	2.50E-003
nuclear import	3.20E-003
positive regulation of macromolecule metabolic process	4.80E-003
protein localization in nucleus	5.20E-003

Term	P-Value
Glutathione metabolism	5.10E-003
Nicotinate and nicotinamide metabolism	1.80E-002
NOD-like receptor signaling pathway	4.90E-002
Fc epsilon RI signaling pathway	5.00E-002
Ether lipid metabolism	6.20E-002
Purine metabolism	6.50E-002
Cytosolic DNA-sensing pathway	8.60E-002



ChIP-seq peaks DNA binding affinity cut-off:

number of mGATA1_{IC} 8mers
per ChIP-seq peak

Figure 105 (previous page). GATA1 ChIP-seq peaks with different DNA binding affinities regulate distinct groups of genes. We took 700 GATA1 ChIP-seq peaks with the lowest DNA binding affinities (less or equal to 13 mGATA1_{NC} 8mers per ChIP-seq peak), 700 GATA1 ChIP-seq peaks with the highest DNA binding affinities (more or equal to 32 mGATA1_{NC} 8mers per ChIP-seq peak) and 700 GATA1 ChIP-seq peaks that showed DNA binding affinities between the previous two groups (more or equal to 18 and less or equal to 27 mGATA1_{NC} 8mers per ChIP-seq peak). The genes that were associated with the GATA1 ChIP-seq peaks (less than 10kb upstream and downstream of the TSS) in each one of the previous three groups were used to find enriched biological processes and pathways. The binding affinity of the GATA1 ChIP-seq peaks is expressed as the number of 8mers per ChIP-seq peak from the 2122 enriched mGATA1_{NC} 8mers. The gene set enrichment analysis was performed with DAVID (Huang da et al., 2009). The GATA1 ChIP-seq peaks of figure (A) were from (Xu et al., 2012) (GEO accession number: GSM970258) (cell type: proerythroblasts (ProEs) from adult human bone marrow CD34+ cells).

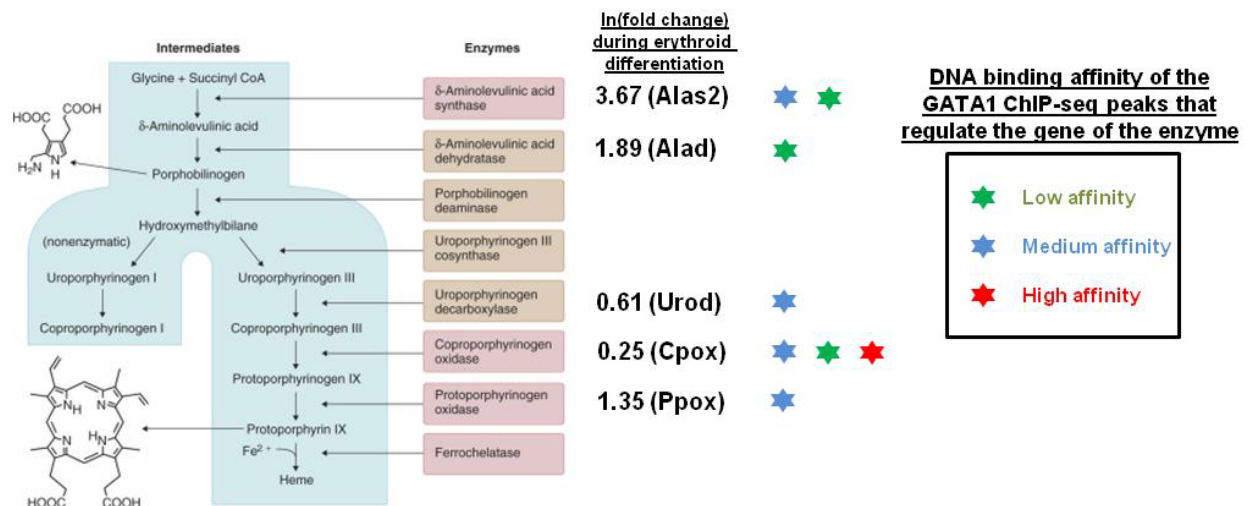


Figure 106. Heme biosynthetic pathway. The genes of the presented enzymes, that are controlled from different GATA1 ChIP-seq peaks, are shown inside the parenthesis, along with their gene expression fold change during the erythroid differentiation (G1E/G1E-ER4 model, (Wu et al., 2011)). Each asterisk represents a different GATA1 ChIP-seq peak. The colour of the asterisk represents the binding affinity category of the GATA1 ChIP-seq peak as presented in Figure 105. Adapted from (Cecil Medicine, 23rd edition).

Results summary

The results can be summarized in the following points:

- High-throughput profiling of a transcription factor DNA binding can be used to infer structural properties of binding. We saw the relative positioning of the GATA1 N-terminal and the C-terminal Zn fingers by examining the nucleotide distance distribution between their binding sites on the DNA sequence. We also saw that non DNA binding domains of GATA1 can significantly suppress the DNA binding specificities of the Zn fingers.
- Complexation of a co-activator/co-repressor can cause a shift on the DNA binding properties of a transcription factor (DNA binding motif and orientation). Transcriptional co-factors of GATA1 can relax the suppressive effect of the non DNA binding domains of GATA1 (for example complexation with FOG1) while they can force the two GATA1 Zn fingers to bind on the DNA sequence further apart between them (for example complexation with either FOG1 or KLF1).
- High-throughput estimation of dissociation constants in vitro are representative of the DNA binding turnover in vivo. GATA1 transcriptional co-factors can affect the GATA1 DNA binding turnover by increasing or decreasing the residence time of GATA1 on the DNA sequence. The increase in the residence time of GATA1 is mediated mainly in genomic positions where GATA1 can bind with low affinity (for example GATA1/p300 overlapping sites). The decrease in the residence time of GATA1 happens, equally probable, in genomic positions where GATA1 can bind with low, medium or high affinity (for example GATA1/PU.1 overlapping sites).
- DNA binding turnover can discriminate functional modules (enhancers, promoters). Enhancers usually have more binding sites for GATA1 than promoters indicating that high DNA residence time for GATA1 increases their chance of contacting the relevant promoter.

- The DNA binding turnover of GATA1 dictates the nucleosome occupancy levels. Regions with high DNA residence time for GATA1 show lower nucleosome occupancy levels.
- Coupled enzymatic activity (H3K4 tri/di-methylation) and DNA binding turnover can be used to infer direct/indirect associations in vivo. The DNA residence time of GATA1 near a TSS was found to be inversely correlated with the trimethylation levels of H3K4 on the TSS. The DNA residence time of GATA1 on its binding site was positively correlated with the dimethylation and monomethylation levels of H3K4 on its binding site.
- All the above are transcription-factor specific properties. In contrast with GATA1, KLF1 did not show any of the examined biological properties to depend on its DNA binding affinities.

CHAPTER 4 DISCUSSION

The current work has presented the EMSA-seq technique as a versatile platform for studying protein-DNA interactions. By using GATA1 as an example, the work has shown the following:

- i) GATA1 co-factors affect the **microscopic** properties of the GATA1 protein. Specifically, we saw changes of the GATA1 DNA binding specificities, which we assigned in the different movement of the two DNA binding Zn fingers relative to each other.
- ii) In vivo, the GATA1 DNA binding affinities of the genome affect the **macroscopic** properties of the system. Specifically, we saw how the DNA binding affinities of the genomic positions where GATA1 binds, affect the local nucleosome occupancy levels, the local H3K4me2 levels and the H3K4me3 levels on the TSS of a nearby gene.

4.1 EMSA-seq is a versatile platform for studying protein-DNA interactions

Initially, we and others used the EMSA-seq/SELEX-seq platforms to characterize the DNA binding specificities of different transcription factors from the same or different families (Jolma et al., 2013, Badis et al., 2009). Our original work with the EMSA-seq platform, presented an initial exploration of the EMSA-seq data by comparing the change in the DNA binding specificities of the different members of the NF-kappaB family (Wong et al., 2011a). Because the proteins were all part of the same family, the work was mainly focused on how different is the DNA binding motif logo of each member, from the general DNA binding motif logo that represents the whole NF-kappaB family.

In this work we expanded the potential of the EMSA-seq platform (Figure 7), both as far as it concerns the experimental design and the data analysis. For this reason, we introduced concepts from classical biochemistry. Specifically, we titrated the protein against a specific amount of DNA and we sequenced the bound DNA from each different protein/DNA ratio conditions (Figure 17). We used simple biochemical models to extract dissociation constant (“Kd”) values (Figure 22). We also introduced the EMSA-seq coefficient (e) as a value characteristic of the kinetic behavior of the DNA binding (Figure 46, Figure 47).

Because our protein of interest, GATA1, has two DNA binding domains (the N-terminal and the C-terminal Zn finger), we considered interrogating each one of them independently. We eventually managed to interrogate only one of them (C-terminal Zn finger). Nevertheless, we were able to assign independent and dissimilar motifs “**GATA**” and “**GATC/GATG**” to the C-terminal and the N-terminal DNA binding domains of GATA1 respectively (Figure 28, Figure 31). We then used these motifs as “**fiducial markers**” that permitted us to explore the movement of the two DNA binding domains relative to each other (**Figure 33**). This experimental design did indeed permit us to see, **for the first time and in a quantitative way**, how the two DNA binding domains of GATA1 behave when GATA1 is alone or in complex with other proteins. **This concept has little to do with the motif logos that are very often found in the literature.**

The above approach gave us insight into the system at a very detailed level. For this reason, we went to suggest that this technique complements molecular biophysics techniques like x-ray crystallography, NMR and molecular dynamics. The X-ray crystallography gives data with atomic resolution but its disadvantage is that it does not have a dynamic resolution, meaning that the structure is one of the potential conformations of the complex. On the contrary, the dynamic resolution of a specific protein-DNA complex can be derived from techniques like NMR and molecular dynamics.

Similarly, the EMSA-seq platform has shown a dynamic resolution at base pair level. As we have seen, the platform offers unbiased, experimentally derived frequencies of occurrence of specific conformations. This makes our approach even more powerful, as these frequencies represent the protein-DNA dynamics that NMR and molecular dynamics experiments are trying to calculate.

We also saw that the different protein-DNA ratios of EMSA-seq also gave us insights into the binding cooperativity of specific DNA motifs, another aspect of the structural conformation of the transcription factors. **We thus suggest that the EMSA-seq, but not SELEX-seq is more appropriate to capture structural properties of the transcription factors and give novel insights.** Although massive efforts have been undertaken to characterize the DNA binding specificities of a great variety of transcription factors, through automation based SELEX-seq (Jolma et al., 2013), we suggest that the derived binding motif logo is just an aspect of the binding of the transcription factor per se. **Equally and probably most important are the other aspects of binding as has been presented in this study.**

4.2 The microscopic properties of the GATA1 DNA binding domains are regulated by both intrinsic and extrinsic factors

As far as it concerns GATA1, we show that the N-terminal and the C-terminal tails are equally important as the DNA binding Zn fingers. We saw that the tails regulate the binding by reducing the DNA binding specificities of the whole protein (Figure 34, Figure 49). **This we hypothesize that reduces the stochastic recognition of the binding motifs on the genome and permits the protein to acquire increased sensitivity upon interaction with factors that can restore part of the original DNA binding preference (Figure 55).**

We also saw that complexation of GATA1 with other co-factors favors specific binding conformations. For the wt-GATA1/wt-FOG1 complex we saw that FOG1 forces GATA1 to show increased preference for the **GATC**{N}_{1,2}GATA conformations (Figure 49). It also increases in a lesser extent the binding preference for distant conformations GATA{N}₈₋₅**GATC** (Figure 49).

We then saw that the FOG1 Zn fingers 1, 9 and the 5/6 have the same effect on the conformation of the GATA1 DNA binding domains which reflects the change in the GATA1 DNA binding specificities (Figure 58, Figure 56). We noticed that the GATA1/FOG1 fingers complexes dissociate upon DNA binding something that we suggest reflects a true behavior (Figure 20). Although all three FOG1 fingers are redundant in their functional effect on the GATA1 conformation, the dissociation behavior along with the amount of competition from specific DNA sequences (the fold enrichment of the bound GATA1 8mers that we have seen) (Figure 27) have lead us to speculate that the different FOG1 fingers regulate the rate of transition between specific GATA1 conformations (Figure 60).

We also saw that the GATA1/KLF1 complexation shows preference for the GATA{N}₈₋₅**GATC** motif (Figure 64). This is similar with the effect of the FOG1 Zn fingers on GATA1 implying that there probably exists a limited panel of conformational changes with which GATA1 can respond to interaction with other transcriptional co-factors. The GATA1/KLF1 complex showed additional preference for distant GATA/GATG and GATA/GATA conformations (Figure 64).

4.3 The macroscopic properties of the GATA1 DNA binding affinity distribution characterize the transcriptional regulation of erythropoiesis

4.3.1 The EMSA-seq derived DNA binding affinities of the GATA1 ChIP-seq peaks reflect the relative DNA residence time of GATA1

Gene regulation has components that work on different time scales usually in parallel. From individual proteins to multiprotein complexes the system needs to organise the succession of events in a way that independently of the properties of its individual components, the whole system moves towards one direction in order to achieve the final outcome (Coulon et al., 2013).

Different studies have found that the stochasticity in the system can be reduced if the components that work on lower time scales initiate multiple events whereas the components that work on larger time scales create a permissive environment for the lower time scale events to occur (Fuda et al., 2009). For example, Fluorescence Recovery After Photobleaching (FRAP) studies have found that transcription factors (TF) are bound on the DNA on average for a few seconds (van Royen et al., 2011, McNally et al., 2000). Competition ChIP has shown that nucleosome remodelling takes minutes to hours (Dion et al., 2007). Additionally, transcription kinetics studies have shown that initiation of transcription takes several minutes (Larson et al., 2011). This implies that multiple cycles of binding and dissociation of the TF are happening until a productive event (a sequence of different histone post translational modifications or histone eviction or formation of a stable pre-initiation complex) passes a threshold and permits the system to move into the next step (Metivier et al., 2003).

All the above indicate that having a way to measure the residence time of the protein on the DNA can reveal mechanistic aspects of the system. From a technical point of view

chromatin immunoprecipitation (ChIP) techniques and all of its derivatives can be assumed that they can partially address this problem. The reason for this is that different methodologies can interrogate the same event on different time scales. For example, after estrogen induction, FRAP analysis of ER has shown retention times on the DNA at the scale of few seconds (Stenoien et al., 2001). On the contrary kinetic ChIP analysis showed retention time at the scale of minutes (Metivier et al., 2003). Although the experimental design of FRAP and ChIP are considerably different, the difference in the final results is attributed to the fact that ChIP determines the fraction of the DNA sequences, in all the cell population, that are bound from the transcription factor. This fraction is defined as the sum of the binding events in all cells, whereas in each cell the factor might be or might not be bound at the time of the experiment (Coulon et al., 2013).

In our case, we hypothesized that the EMSA-seq technique, although in vitro by nature and without directly interrogating any in vivo mechanisms, can under some assumptions reflect the binding behavior of the protein on specific genomic loci in an idealized cell (Figure 74).

For this reason, we projected the in vitro calculated GATA1 DNA binding affinities in vivo and we indeed saw that they are representative of the kinetic recruitment of GATA1 on the ChIP-seq peaks (Figure 73). We used data from a kinetic GATA1 ChIP-seq study after GATA1 induction. Although our GATA1 DNA binding turnover, which is expressed as GATA1 DNA binding affinities, has no time scale units, we believe that it represents the strength of binding in each genomic position. If the genomic positions are compared relative to each other, the GATA1 DNA binding turnover can represent the arbitrary fraction of time that GATA1 stays on the genomic position defined from the ChIP-seq peak.

We also do recognise that by using the EMSA-seq DNA binding affinities as a measure of the DNA binding turnover in the GATA1 binding sites, we are assuming that no-other factor can affect the DNA residence time of GATA1. This is a disadvantage comparing to the ChIP-seq studies that can hint for potential deviations from the expected behaviour albeit with a lower resolution. Nevertheless, as we have already seen, the genome wide distribution of the GATA1 DNA binding affinities revealed interesting properties of the transcriptional regulation. For this reason we believe that our approach is of a great interest in the general scientific community.

4.3.2 Complexes of GATA1 with some of its transcriptional co-factors, that can selectively increase its binding affinities, can exclusively appear in genomic positions where GATA1 has short residence time when it binds alone

Apart from the affinity of the DNA sequence itself, the residence time of a TF on the DNA is regulated by other mechanisms. For example, the eviction time of a TF from the DNA can be modulated (increased) by complexes that interact with the TF like chromatin remodelers (as in the case of the glucocorticoid receptor (GR) and the yeast Ace1 (Fletcher et al., 2002, Karpova et al., 2004) respectively), chaperones (as in the case of the GR and the yeast HSF1 (Zelin et al., 2012, Freeman and Yamamoto, 2002) respectively) or the proteasome (as in the case of GR and the yeast GAL4 (Ferdous et al., 2007, Stavreva et al., 2004) respectively). For other TFs, DNA binding can be stabilised. For example the heat shock factor (HSF) in *Drosophila* binds as a monomer with a low residence time in non-heat shock conditions whereas upon heat-shock it trimerizes and therefore it shows high residence time on the DNA (Yao et al., 2006). Similarly, competition ChIP showed increased stability on specific genomic loci for the human HIF1a (hypoxia inducible factor 1) under hypoxia induction (Yu and Kodadek, 2007) and for the

yeast GAL4 under the presence of galactose (Nalley et al., 2006). All these indicate that the ability to regulate the dynamic formation and disruption of the nucleoprotein complexes, which represent the intermediates of the eukaryotic gene expression, is an essential part of the system. This mechanism may make the system more or less sensitive to changes of signalling cascades by decreasing or increasing respectively the lifetime of the nucleoprotein complexes.

Although the effect of the GATA1 interacting proteins on the GATA1 DNA binding affinities has been established in vitro, we tried to examine whether we can infer this role genome wide from in vivo data, by projecting our EMSA-seq derived GATA1 binding affinities. This is very important because it will reveal the robust role of the DNA binding affinities in regulating the GATA1 DNA binding behaviour. Indeed we found that all of the examined DNA binding and non-DNA binding co-factors, with the exception of TAL1, tend to associate with low/medium GATA1 DNA binding affinity genomic positions (Figure 92). For example, as we saw the GATA1 co-factors FLI1 and p300 appear in low GATA1 DNA binding affinity genomic positions. Both have been shown to increase the DNA binding affinity of the GATA1 protein. **Thus, by using other GATA1 co-factors, the system can probably recruit and retain GATA1, for extended time, in genomic positions where it cannot afford to have multiple GATA1 binding sites.**

On the other hand, the GATA1 competitor PU.1 associates with the entire DNA binding affinity spectrum of the GATA1 genomic positions (Figure 92). This probably happens because the GATA1 exclusion from the lymphoid lineage needs to happen in as many as possible from the GATA1 binding sites.

As we discussed, one of the major complexes of GATA1 is a pentameric complex where the two main DNA binding proteins of the complex are GATA1 and TAL1 (Cantor and Orkin, 2002). We saw that TAL1 is found mainly in GATA1 ChIP-seq peaks of medium to high DNA binding affinity (Figure 90). **Considering that TAL1 is associated with**

transcriptional activation (Kassouf et al., 2010, Kassouf et al., 2008), the potential regulation of the complex formation from the GATA1 DNA binding affinities of the genomic position, indicates a direct role of the GATA1 DNA binding affinities in the transcriptional activation.

Another aspect of the regulation of the system comes from experiments which used a mutated form of GATA1 (GATA1^{V205G}) that could not bind FOG1. There, GATA1 showed additional preference for binding in high affinity genomic positions compared to the average DNA binding affinity from the ChIP-seq binding sites of the wild-type protein (Figure 85). **This implies that the GATA1 co-factors can reduce binding to high affinity sites in order to probably give fine regulation in the whole system.** This is in accordance to studies that have shown that reduced clearance of a TF from the DNA, through proteasome inhibition, can lead to reduced expression of its target genes (Lipford et al., 2005, Muratani et al., 2005, Reid et al., 2003).

4.3.3 The DNA residence time of GATA1 affects the nucleosome occupancy levels of the binding site. It additionally affects the levels of the H3K4me2 and the H3K4me1 histone marks on the binding site and the levels of the H3K4me3 on the nearest TSS

Transcription factor (TF) occupancy is influenced by nucleosome positioning. TFs and histones compete for DNA binding. TFs usually bind in nucleosome depleted regions (Liu et al., 2006, Li et al., 2011, John et al., 2011, Degner et al., 2012). TFs can also displace and/or reposition the nucleosomes by interacting with chromatin remodelers (SWI/SNF complex and CHD, ISWI, INO80 complexes respectively) (John et al., 2008, Vicent et al., 2009, Morris et al., 2014). All these indicate that nucleosome instability regulates gene

expression by facilitating the access of transcription factors to promoters and other cis-regulatory elements (Mito et al., 2007).

Regions with mobilised nucleosomes are sensitive to DNase I digestion. Genomic regions with active genes were recognised to be more susceptible to DNase I digestion than regions with non-transcribed genes (Weintraub and Groudine, 1976, Garel and Axel, 1976). Promoters and enhancers of active genes where the chromatin fibre is more relaxed are highly sensitive to DNase I digestion than the local genomic region (DNase I hypersensitive regions) (Wu, 1980).

Pioneer factors are TFs that can access their DNA binding sites in compact chromatin. Their binding and the consequent chromatin remodelling, needs to occur before the recruitment of additional TFs that are otherwise unable to bind (cooperative binding) (Spitz and Furlong, 2012). Pioneer factors can prime the binding sites by changing the post-translational modifications of histones, by recruiting histone variants or by organising the local nucleosomes (Zaret and Carroll, 2011). TFs that have been found to act as pioneer factors are the Forkhead box A (FOXA) (Krum et al., 2008, Hurtado et al., 2011), GATA2 acts as a pioneer factor for subsequent androgen receptor (AR) binding (Wang et al., 2007), GATA3 (Eeckhoute et al., 2007) and GATA4 act as pioneer factors for subsequent estrogen receptor (ER) binding (Miranda-Carboni et al., 2011), PU.1 (Ghisletti et al., 2010, Heinz et al., 2010).

GATA1 has been shown to collaborate with the SWI/SNF chromatin remodelling complex through interaction with BRG1 (Kadam and Emerson, 2003) in order to activate the transcription of the beta-globin genes. In the GATA1 enhancers, GATA1 acts as a pioneer factor that brings BRG1 with it (Boyes et al., 1998b, Escamilla-Del-Arenal and Recillas-Targa, 2008). Subsequently, BRG1 shifts the nucleosomes away from the binding sites of GATA1, something that facilitates the binding of TAL1 and the transcriptional activation of the nearby genes (Hu et al., 2011). KLF1 has been shown also to interact with chromatin

modifying complexes like CBP/p300 (Zhang et al., 2001) and the SWI/SNF chromatin remodelling complex through BRG1 (Brown et al., 2002, Kim et al., 2009b, Zhang et al., 2001, Kadam et al., 2000, Armstrong et al., 1998). Nevertheless, BRG1 is not a general transcriptional regulator because it is not necessary for the GATA1 dependent expression of the FOG1, GATA1 and KLF1 genes (Kim et al., 2007b, Kim et al., 2009c), whereas the expression of the alpha (Kim et al., 2009b) and beta globin genes (Kim et al., 2009c, Bultman et al., 2005) depends on his function. The H3 acetylation and H3K4me2 levels of the DNase I hypersensitive sites, the DNase I sensitivity and chromatin looping between promoter and enhancers, for both the alpha and the beta globin genes, have been shown to depend on the GATA1 and KLF1 recruitment of BRG1 (Kim et al., 2009c, Kim et al., 2009b).

Our working hypothesis was that if either the nucleosome occupancy or the histone post translational modifications are controlled directly by GATA1 or KLF1 they should show a positive or negative correlation with the DNA binding affinity of the GATA1 or KLF1 binding site. We do recognise that the same will happen if the binding affinity of the GATA1 or the KLF1 ChIP-seq peak depends on the levels of either the nucleosome occupancy or the histone post translational modifications, something that reflects the problem of causality. Nevertheless, here we mainly hypothesized that the levels of the nucleosome occupancy or the histone post translational modifications, are affected from the residence time of the GATA1 complexes on the DNA sequence. The DNA residence time is positively correlated with the binding affinity of the GATA1 ChIP-seq peaks.

We saw that the binding affinity of the GATA1 ChIP-seq peaks is positively correlated with the levels of the H3K4me1 and the H3K4me2 marks on the region of the ChIP-seq peaks (Figure 97). It is also positively correlated with the levels of the H3K4me2 mark on the region of the TSS (Figure 94). On the contrary, it is negatively correlated with the levels of the H3K4me3 mark on the region of the TSS (Figure 95). Additionally the binding affinity of the GATA1 ChIP-seq peaks is negatively correlated with the levels of nucleosome

occupancy (Figure 79) and positively correlated with the levels of DNase I hypersensitivity (Figure 81) on the region of the ChIP-seq peaks. Furthermore, the binding affinities of the KLF1 ChIP-seq peaks showed no associations with the histone post translational modifications similar with the GATA1 ones. The binding affinities of the KLF1 ChIP-seq peaks were associated with constant levels of the histone post translational modifications and of the DNase I hypersensitivity (Figure 83) and a very mild negative correlation with the nucleosome occupancy (Figure 79).

Because high affinity DNA binding sites for GATA1 in DNase I hypersensitive regions are not preferred equally well as low and medium affinity binding sites (Figure 84) we cannot exclude a mechanism of regulation where GATA1 binding in high affinity binding sites is regulated by permissive (high) levels of H3K4me2, H3K4me1 and low nucleosome occupancy levels. In this case another factor must act as a pioneer factor to enable GATA1 to bind by priming the GATA1 binding sites.

We can also envision a direct effect of GATA1 on transcription by facilitating the rate of assembly of the pre-initiation complex (PIC). It is generally assumed that regulation of transcription is mediated through regulation of the assembly of the pre-initiation complex (PIC) (Juven-Gershon et al., 2008, Fuda et al., 2009) although equally important in the mammalian cells is regulation of the promoter proximal pausing of RNA polymerase II (Adelman and Lis, 2012). The components of the PIC assemble stochastically. The residence time of its components can be modulated in order to increase the likelihood of its full assembly (Hager et al., 2009).

The PIC is constantly assemble and disassemble in the promoters due to competition with the nucleosomes. Reduced nucleosome occupancy in the promoters has been shown to lead to increased transcription after heat shock in yeast (Zanton and Pugh, 2006). A similar thing has been observed when the H3 histone is depleted in vivo (Gossett and Lieb, 2012). Nevertheless, nucleosome depletion does not control the time of transcription

(Grimaldi et al., 2014, Huebert et al., 2012) but controls proportionally the rate of transcription (Lee et al., 2004).

Further dissection of the interplay between the rate of nucleosome disassembly and the rate of assembly of the PIC has led to a biphasic transcriptional output. Kinetic analysis of transcription has revealed that gene expression can either be continuous or occur in bursts (Lionnet and Singer, 2012). Both in the continuous and the burst expression, the firing rate of transcription (**initiation rate** in the continuous expression or the total number of transcripts (size) per burst) can be controlled by the same mechanism whereas the mechanism that controls the period of inactivity after the burst, which defines the frequency of a burst, is different than the continuous expression (Larson et al., 2009). That is to say that the oscillation pattern of the chromatin (type and/or overall occupancy) in the promoter between a permissive and a non-permissive state can indicate the type of expression. Stochastic opening leads to continuous expression whereas the coordinated opening leads to burst expression (Kaern et al., 2005, Zenklusen et al., 2008). Indeed, nucleosome occupancy has been shown to affect the induction level of a gene (**frequency of a burst**) but not its expression level (**initiation rate/burst size**) (Lam et al., 2008, Bai et al., 2010). For example, exposed low or high affinity binding sites can recruit a factor that will remodel the nearby chromosomes in order to expose the rest of its binding sites, that are hidden by nearby nucleosomes. The low or high frequency of the gene induction (frequency of a burst) will depend from whether the **initial** exposed binding sites are of a low or high binding affinity respectively. The magnitude of gene induction (initiation rate) will depend on the **total** amount of exposed high or low affinity binding sites after the chromatin remodelling (Lam et al., 2008). Similarly, it has been shown that the burst size can be modulated from cis-regulatory motifs in the promoter (Suter et al., 2011).

One interpretation of our analysis results is that GATA1 modulates the local nucleosome occupancy levels. We tested GATA1 ChIP-seq peaks that were 10 kb away from the TSS to avoid any interference from the nucleosome structure in the promoters that might reflect the effect of additional gene related mechanisms (Figure 79). Nevertheless, we made the assumption that the positive correlation between the GATA1 ChIP-seq affinity and the nucleosome eviction of its binding site, must reflect a generic GATA1 mechanism. We also saw that high levels of H3K4me2 were present in regions of high affinity GATA1 binding sites indicating a direct or indirect role of GATA1 in the regulation of the H3K4me2 deposition (Figure 97). Because TSS that were near high affinity GATA1 binding sites also showed higher levels of H3K4me2 when they were compared to TSS that were near low affinity GATA1 binding sites (Figure 94), we can conclude that GATA1 probably keeps an open chromatin environment in each closest TSS through transient interactions between its binding site and the compatible TSS region (Figure 107). Additionally, the low DNase I hypersensitivity in the nucleosome evicted high affinity GATA1 ChIP-seq peaks indicates that large multiprotein complexes of GATA1 appear in GATA1 ChIP-seq peaks of high DNA binding affinity (Figure 81). These large complexes might also compete and prevent the re-assembly of the nucleosomes. **We can also hypothesize that the retention of GATA1 in a specific position for a long time is a requirement for the formation of large complexes.** Sequentially, the GATA1 complexes can affect the local chromatin environment.

4.3.4 GATA1 can affect transcription by modulating the H3K4me3 levels of the closest TSS

We saw that the higher the binding affinity of the GATA1 ChIP-seq peak the higher the decrease in the H3K4me3 levels on the nearest TSS upon erythroid differentiation after GATA1 restoration (Figure 99). Although high levels of H3K4me3 are associated with

actively transcribed genes, the absolute change in its levels does not correlate linearly with a similar change in the gene expression levels. In contrary to what it was expected, reports have suggested that reduction of the H3K4me3 levels due to high H3K4me3 demethylase activity can positively regulate transcription (Gildea et al., 2000, Lee et al., 2007c, Lloret-Llinares et al., 2008) . One of the mechanisms on which this was attributed, is that removal of the H3K4me3 mark from the promoter, left behind from the previous RNA polymerase II molecule, facilitates the recruitment of the next RNA polymerase II molecule on the promoter, something that permits fast **initiation rates** (Lloret-Llinares et al., 2012). While present H3K4me3 can recruit locally components of the core transcription factor TFIID (Vermeulen et al., 2007), chromatin-remodelers and histone modifying enzymes (Kim and Buratowski, 2009, Li et al., 2006, Santos-Rosa et al., 2003, Wysocka et al., 2006) thus maintaining a transcriptionally permissive environment for a short time even after its absence. Methylation of H3K4 can also regulate not only the initiation rate but also the frequency of the expression bursts. For example, the pulse frequency of the expression bursts, but not the burst size, has been shown to be inherited through mitosis in *Dictyostelium discoideum* via a mechanism that depends on H3K4 methylation (Muramoto et al., 2010).

We also proved that regulation of the H3K4me3 levels has a physiological role in vivo by bringing as an example the common genes that are regulated by GATA1 in both the erythrocyte and the megakaryocyte lineage (Figure 101). Because, the erythrocyte specific GATA1 ChIP-seq peaks of the common genes are of a higher binding affinity than the megakaryocyte specific GATA1 ChIP-seq peaks of the common genes (Figure 102), the levels of the H3K4me3 mark are lower in the erythrocyte lineage than in the megakaryocyte one (Figure 103).

Unexpectedly, all the above were noticed only for GATA1 whereas KLF1 did not show any similar strong pattern. Intriguingly, although both GATA1 and KLF1 are important for

erythropoiesis their genomic DNA binding affinities seem to regulate transcription in different ways.

We can also suspect that the DNA residence time of GATA1 can affect the expression pattern of the nearby gene. We have argued that high affinity GATA1 DNA binding sites are correlated hypothetically with higher initiation rates. If we combine these high initiation rates with the low nucleosome proximal promoter environment, for which the high affinity GATA1 DNA binding sites are responsible, we can hypothesize that the specific gene will show a continuous expression rather than a burst one (Figure 107). In yeast, **housekeeping** genes show a continuous expression whereas **stress response** genes show bursting expression (Zenklusen et al., 2008). We have already seen that GATA1 can regulate different classes of genes and that high affinity GATA1 DNA binding sites are regulating mainly housekeeping activities (Figure 105). So, both the yeast models and the way we have interpreted the results in our system are in good concordance.

4.3.5 GATA1 modulates the H3K4me3 levels through an unknown mechanism

The H3K4me3 mark is present at the promoters of both active and inactive genes even without the presence of transcriptional activators or RNA polymerase II (Vastenhouw et al., 2010). The increase of the H3K4me3 can be either the cause or the consequence of transcription. In mammals, the proteins responsible for the methylation of the H3K4me2 to H3K4me3 are the SET1A, SET1B, MLL1, MLL2, MLL3, MLL4 (Herz et al., 2013). SET1A and SET1B are the main proteins responsible for maintaining the H3K4me3 levels genome wide (Ardehali et al., 2011, Hallson et al., 2012, Wu et al., 2008). The MLL1 and MLL2 maintenance of the H3K4me3 levels is gene specific, for example in the HOX cluster (Wang et al., 2009), and developmental stage specific (oocytes) (Andreu-Vieyra et al., 2010). The MLL3 and MLL4 role in maintenance of the H3K4me3 levels is limited

(Vicent et al., 2011, Kim et al., 2009a, Lee et al., 2009c, Guo et al., 2012) and shows gene specificity, for example the p53 target genes (Lee et al., 2009b). SET1A and SET1B in order to methylate the H3K4me2 require the existence of monoubiquitinated H2B (Lee et al., 2007b, Vitaliano-Prunier et al., 2008, Zheng et al., 2010, Soares and Buratowski, 2012). The yeast SET1 and the mammal MLL1 have been shown to interact with the Serine 5 of the C-terminal domain tail (CTD) of the initiating form of RNA polymerase II a region that recruits also the complex responsible for the monoubiquitination of H2B (Ng et al., 2003, Wood et al., 2003a, Wood et al., 2003b, Milne et al., 2010, Muntean et al., 2010). It has also been shown that asymmetric H3 Arginine 2 (H3R2) di-methylation prevents the recruitment of SET1 in yeast and MLL1 in mammals (Kirmizis et al., 2007, Guccione et al., 2007, Hyllus et al., 2007). Unmethylated CpG islands can also recruit multiprotein complexes (COMPASS) that contain the SET1A and SET1B proteins (Thomson et al., 2010). The MLL1 and MLL2 can also act as a co-factor and interact with the activated estrogen receptor alpha (ER- α), PPAR- γ receptors, E2F, LEF1/TCF (Dreijerink et al., 2006, Dreijerink et al., 2009, Imachi et al., 2010, Mo et al., 2006, Sierra et al., 2006, Tyagi et al., 2007) whereas MLL3 and MLL4 can interact with p53 (Lee et al., 2009b). The removal of the methyl group from the H3K4me3 back to H3K4me2 is mediated from the Jarid1/KDM5 family of histone lysine demethylases (Secombe and Eisenman, 2007). Alternatively, because the mono-methylated, di-methylated and tri-methylated molecules of H3K4 have increasingly slower formation rates on bulk histones these modifications can be lost with passive dilution for example through DNA replication (Zee et al., 2010, Dion et al., 2007).

Depletion of either the CFP1 subunit (Clouaire et al., 2012) or the DPY-30 subunit (Jiang et al., 2011) of the SET1 H3K4 methyltransferases, that leads to a global reduction of the H3K4me3 levels, has a minimal effect on gene expression of most genes. Similar effect was noticed in yeast (Guillemette et al., 2011). H3K4me3 might regulate the processing of

transcripts by affecting the kinetics of RNA polymerase II transcription in early elongation (Terzi et al., 2011).

Based on the above the function of GATA1 might be attributed to the following:

- i) GATA1 might have a direct effect on the H3K4me3 levels by interaction with a histone lysine demethylase from the Jarid1 family. No interaction with a Jarid1 demethylase has been shown. GATA1 has only been shown to interact indirectly with an H3K4me1 or H3K4me2 demethylase LSD1 (Turner and Crossley, 1998, Katz et al., 2002). Thus, GATA1 might interact with an unknown H3K4me3 demethylase.
- ii) GATA1 might affect the recruitment of the SET1A/B from the RNA polymerase II by inducing the dephosphorylation of Ser 5 of the CTD domain. Similarly, GATA1 may reduce the levels of the monoubiquitination of H2B. No interaction of GATA1 with a RNA polymerase II Ser 5 phosphatase or an H2B deubiquitinase has been reported. We also checked whether high DNA binding affinities of GATA1 ChIP-seq peaks are associated with reduced levels of RNA polymerase II (lower levels of RNA polymerase II might cause lower levels of H3K4me3 deposition) but we did not find any association (data not shown).
- iii) GATA1 might increase the asymmetric H3 Arginine 2 (H3R2) di-methylation. The PRMT6 (protein arginine methyltransferase 6) catalyses the recruitment of asymmetric dimethyl H3R2 (Bedford and Clarke, 2009). No direct interaction between GATA1 and PRMT6 has been found although the expression of PRMT6 increases during erythroid differentiation of CD34+ hematopoietic progenitors (Herglotz et al., 2013). Only interaction between RUNX1 and PRMT6 in CD34+ hematopoietic progenitors has been reported (Herglotz et al., 2013). This interaction represses some of the RUNX1 target genes.
- iv) GATA1 might increase the methylation of CpG islands inhibiting the recruitment of the COMPASS complex. No direct interaction of GATA1 with the de novo DNA

methyltransferase Dnmt3a/b has been reported, only interaction with the maintenance methyltransferase Dnmt1 (Dr John Strouboulis, personal communication).

- v) GATA1 might inhibit the DNA binding of proteins that can recruit the histone methyltransferases. Also no interaction between GATA1 and H3K4me2 histone methyltransferases has been reported.

We have also noticed that the GATA1 dependent reduction of the H3K4me3 levels was not as prominent in GC rich promoters as in GC poor promoters (Figure 96). Although high affinity GATA1 ChIP-seq peaks are expected to be found in AT rich/GC poor promoters more frequently, the difference between the two groups permit us to hypothesize about a potential mechanism (Figure 96). Proteins that have the CxxC domain can bind specifically to unmethylated CpG (Voo et al., 2000). CFP1 is a protein containing a CxxC domain and it is part of both the SET1A and SET1B H3K4 methyltransferase complex (Lee and Skalnik, 2005, Lee et al., 2007a). CFP1 is found in CpG islands in mouse brains and its depletion leads to reduced H3K4me3 levels at specific CpG islands (Thomson et al., 2010). Furthermore, the CpG density in CpG islands correlates positively with the H3K4me3 levels (Illingworth et al., 2010). Similarly, the H3K36me2 demethylase KDM2A, which is a CxxC protein, is also being recruited in CpG islands (Blackledge et al., 2010). H3K36me2 inhibits transcriptional initiation in yeast through recruitment of histone deacetylase (HDAC) (Li et al., 2009). Depletion of H3K36me2 may contribute to increased transcription (Deaton and Bird, 2011). We can speculate that GATA1 acts antagonistically with the function of the recruited SET1 complex. In GC rich promoters SET1 can compete the reduction of the H3K4me3 levels caused by GATA1. This indicates that GATA1 probably affects directly the H3K4me3 levels and it is frequently found, and probably prefers, AT rich/GC poor promoters where it can regulate the H3K4me3 levels without significant competition.

4.3.6 The high occurrence of GATA1 binding motifs in enhancers probably increases the likelihood of interaction with the compatible promoter

Gene expression is regulated from the combinatorial action of core promoters, proximal promoter elements and enhancers as well as from the function of silencers, insulators and tethering elements (Lenhard et al., 2012, Bulger and Groudine, 2011, Petrykowska et al., 2008, Gaszner and Felsenfeld, 2006, Calhoun et al., 2002). Fine regulation of transcription is accomplished mainly by enhancers. Enhancers can activate transcription independently of their location, orientation or distance from the promoter region of the gene. Enhancers activate transcription by facilitating the recruitment of the transcription pre-initiation complex (Maston et al., 2006) or the release of paused RNA polymerase II (Ong and Corces, 2011).

The chromatin state (usually combination of H3K4me1, H3K4me2, H3K27me3, H3K27ac marks) at enhancers, in contrast to promoters, is cell type specific and correlates with the gene expression (Heintzman et al., 2009). Because the chromatin state is similar across many enhancers the specificity of each enhancer occurs from the presence of binding sites for multiple factors. This combinatorial binding of factors that can read the epigenetic state of an enhancer and impose a new one, can take place in a temporal manner during the cell differentiation and development (Rada-Iglesias et al., 2011).

The combinatorial binding of multiple TFs on an enhancer depends on their position, orientation and relative distance between them (Spitz and Furlong, 2012). Binding cooperativity of two or more transcription factors can induce gene expression, through their binding synergism, even when their levels are suboptimal for function (Szymanski and Levine, 1995). For example the developmental patterning in *Drosophila* embryos are established from TF gradients that utilise in distant regions, from their place of production in the embryo, either high affinity enhancers or cooperative DNA binding (Lebrecht et al.,

2005). The cooperativity can be either strict or flexible. Strict cooperativity means the enhancers cannot function optimally unless all TFs are present and all bind cooperatively with a specific position and orientation. For example, in the Interferon beta promoter (IFN β) multiple activators need to present their acidic activation domains together and simultaneously contact the co-activator p300/CBP (a structure called enhanceosome) (Merika et al., 1998, Merika and Thanos, 2001). Flexible cooperativity means the enhancers can function optimally when a fraction of all the possible factors, that can bind, have been recruited ("billboard" activity) (Arnosti and Kulkarni, 2005). A second flexible cooperativity model is when all the possible TFs, that can bind, need to be recruited but their position and orientation in the nucleoprotein complex can vary ("transcription factor collective" activity) (Junion et al., 2012).

In our case, we saw that low affinity GATA1 binding sites are usually near the TSS (+/- 1kb) whereas high affinity GATA1 binding sites are further away (Figure 76). As expected, the GATA1 binding sites in enhancer regions (defined as regions that show low H3K4me3 and high H3K4me1) are of a higher binding affinity than the TSS related GATA1 binding sites (high H3K4me3, low H3K4me1) (Figure 77). The KLF1 ChIP-seq peaks did not show so distinctively this discriminatory behaviour (Figure 76, Figure 77). We hypothesize that the occurrence of multiple DNA binding motifs for GATA1 would increase the frequency of contacts between the enhancer and the compatible promoter. Assuming that GATA1 complexes of the same functionality bind on a specific promoter, we envision a "billboard" activity for the GATA1 enhancers where the presence of multiple GATA1 motifs is redundant but important to increase the likelihood of occurrence of downstream events. This would facilitate the long range interactions that are mediated by the GATA1/FOG1 complex (Kim et al., 2007b, Vakoc et al., 2005, Deng et al., 2012). It has also been shown that an ectopic beta globin LCR can preferentially contact genes that are controlled by shared transcription factors (Noordermeer et al., 2011) and probably an increased frequency of events in the enhancers will facilitate this behaviour.

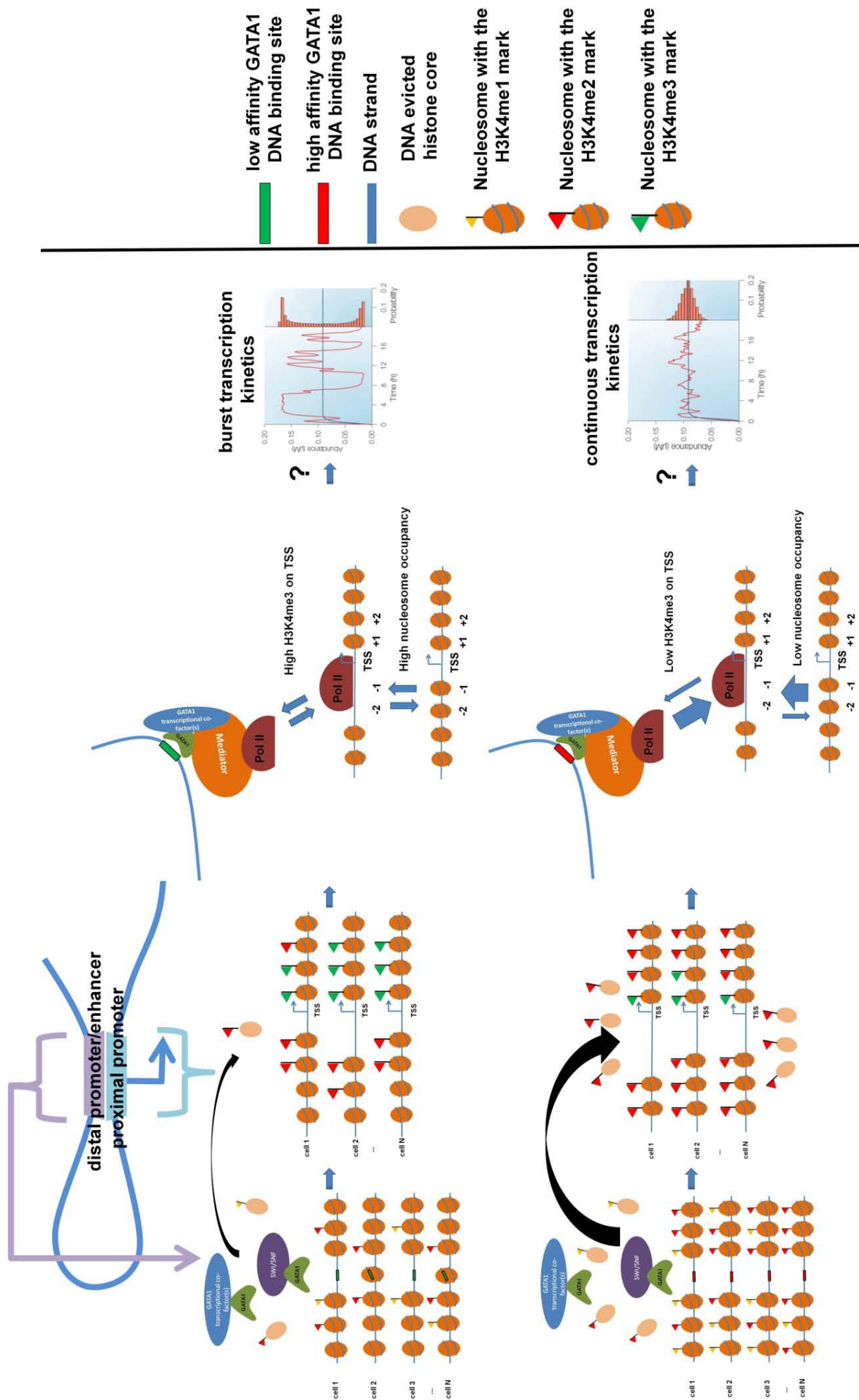


Figure 107 (previous page). Hypothetical model for the effect of the GATA1 DNA binding affinities on the local chromatin environment and on the chromatin environment of the TSS region and their relationship with RNA polymerase II loading and transcriptional kinetics. Enhancers/Distal promoter elements with high or low affinity GATA1 DNA binding sites show higher or lower eviction of the local nucleosomes respectively in the total cell population, due the chromatin remodelling function of the GATA1 co-factor the SWI/SNF complex. Eviction of the local nucleosomes might be facilitated by priming the nucleosomes with the histone marks H3K4me1 and H3K4me2. High or low occurrence of the H3K4me1/H3K4me2 marks in the total cell population is associated with a higher or lower eviction of the local nucleosomes. The priming can be regulated by GATA1 directly or indirectly. Close proximity between the enhancers/distal promoters and the proximal promoters affects the H3K4me1 levels, the H3K4me2 levels and the nucleosome occupancy levels of the TSS region in a similar way as the local nucleosomes of the GATA1 binding site in the enhancer/distal promoter region (black arrows, the thicker the arrow the stronger the effect). Additionally, close proximity between the enhancers/distal promoters and the proximal promoters affects the H3K4me3 levels on the TSS. High GATA1 DNA binding affinity enhancers/distal promoter elements, in contrast to low affinity enhancers/distal promoter elements, can reduce the H3K4me3 levels on the TSS. Reduction of the H3K4me3 levels can increase the loading of sequential RNA polymerase II molecules. The Mediator complex can bridge the interaction between GATA1 and RNA polymerase II (Stumpf et al., 2006). Simultaneously, due to the close proximity of the enhancers/distal promoter elements and the proximal promoters, increased chromatin remodelling activity from the high affinity GATA1 binding sites can increase the nucleosome eviction near the TSS. Increased RNA polymerase II loading and nucleosome eviction can lead to a continuous pattern of expression. On the contrary reduced RNA polymerase II loading and increased nucleosome formation can lead to a burst expression pattern. Periods of transcriptional activity can be followed from periods of nucleosome presence on the TSS and absence of RNA polymerase II with no

transcriptional activity. The images of the continuous and burst transcription kinetics are from (Kaern et al., 2005). Abundance represents the number of mRNA molecules produced per fraction of time. The question mark represents a hypothetical association. The different complexes of GATA1 for example either with the SWI/SNF complex or with another transcriptional co-factor can appear interchangeably on the same binding site. We can also hypothesize that low affinity binding sites where GATA1 complexates with transcriptional co-factors, which extend the residence time of GATA1 on the DNA, can have similar effects as the high affinity binding sites.

Future experiments

The study generated the following observations:

- We saw that wt-mGATA1 cannot bind in palindromic positions as efficiently as the wt-mGATA1/wt-mFOG1 complex. To test this we could use a minimal promoter reporter assay in HELA cells which lack endogenous GATA1 (Rodriguez et al., 2005). We will have to show that the induction of the reporter gene from wt-mGATA1 alone is similar for sequences containing $\text{GATA}\{\text{N}\}_{1-6}\text{GATC}$ and $\text{GATC}\{\text{N}\}_{1-6}\text{GATA}$. On the other hand, the expectation is that the wt-mGATA1/wt-mFOG1 complex will show an increased induction of the reported gene for sequences containing **GATC**NNGATA whereas the rest of the sequences will have to show similar activation as the wt-mGATA1 alone.

We can also use the mutant protein $\text{GATA1}^{\text{R216Q}}$. The N-terminal Zn finger of $\text{GATA1}^{\text{R216Q}}$ cannot bind DNA (Liew et al., 2005, Yu et al., 2002, Campbell et al., 2013) but it can bind FOG1. The expectation here is that only the C-terminal Zn finger will guide the binding something that will result in lack of preference for the **GATC**NNGATA configuration from the wt-mGATA1^{R216Q}/wt-mFOG1 complex.

- Enhancers show increased binding affinity compared to promoters. The hypothesis that is generated is whether there is a functional significance. Are the multiple binding sites indispensable for the function of the enhancer or they operate additively. We can use the same minimal reporter assays as previously. We can synthesize mutant versions of enhancers and test them in parallel either in low throughput or high throughput assays (Melnikov et al., 2012, Patwardhan et al., 2012). If the multiple binding sites in the enhancer operate additively then we expect the signal of the reporter gene to be proportional to the number of binding

sites. On the other hand if all the binding sites are necessary for the function of the enhancer the signal of the reporter gene will be binary, either the enhancer will be able to activate the reporter gene or not.

- We observed that high affinity GATA1 ChIP-seq peaks are associated with lower levels of H3K4me3. We can test the role of the H3K4me3 demethylase (JARID1) in mediating the reduced H3K4me3 levels by using chemical inhibitors or siRNA mediated inhibition. There is also the possibility that GATA1 is regulating indirectly the reduced H3K4me3 levels by inhibiting the methylation of H3K4me2 to H3K4me3. In this case we can take a gene regulated by GATA1 from a low affinity (or a high affinity) promoter associated binding position, mutagenise the sequence to a high affinity binding site (or to a low affinity binding site respectively) using a genome editing technology like CRISPR-Cas (Mali et al., 2013, Sander and Joung, 2014) and assess the effect on the H3K4me3 and the H3K4me2 levels of the nearby TSS.
- The genome editing of the GATA1 binding site can be used to assess the effect of GATA1 on the local histone-posttranslational modifications levels of the GATA1 binding sites, for example the levels of H3K4me1 and H3K4me2.
- We also observed that high affinity GATA1 ChIP-seq peaks are associated with lower nucleosome occupancy levels. Again, we can use the genome editing of the GATA1 binding site to assess the effect of GATA1 on the local nucleosome occupancy levels. Alternatively, we can use a genome wide approach by randomly inserting thousands of reporters in the genome (integrated reporters) (Akhtar et al., 2013). The integrated reporters will consist from a minimal promoter and a high binding affinity site for GATA1. We can then correlate the gene induction in the

different genomic loci with the nucleosome occupancy levels of the respective loci. The expectation is that if the nucleosome occupancy levels dictate the GATA1 binding on the high affinity site then the expression levels of the reporter gene will be negatively correlated with the nucleosome occupancy levels of the genomic loci. On the contrary, if the GATA1 binding on the high affinity site is independent of the local nucleosome levels (due to the remodelling activity of the GATA1 co-factors) then the expression levels of the reporter gene will show no correlation with the nucleosome occupancy levels of the genomic loci.

References

- ADELMAN, K. & LIS, J. T. 2012. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat Rev Genet*, 13, 720-31.
- AKHTAR, W., DE JONG, J., PINDYURIN, A. V., PAGIE, L., MEULEMAN, W., DE RIDDER, J., BERNIS, A., WESSELS, L. F., VAN LOHUIZEN, M. & VAN STEENSEL, B. 2013. Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell*, 154, 914-27.
- ANANTHARAMAN, A., LIN, I. J., BARROW, J., LIANG, S. Y., MASANNAT, J., STROUBOULIS, J., HUANG, S. & BUNGERT, J. 2011. Role of helix-loop-helix proteins during differentiation of erythroid cells. *Mol Cell Biol*, 31, 1332-43.
- ANDERSON, K. P., CRABLE, S. C. & LINGREL, J. B. 1998. Multiple proteins binding to a GATA-E box-GATA motif regulate the erythroid Kruppel-like factor (EKLF) gene. *J Biol Chem*, 273, 14347-54.
- ANDREU-VIEYRA, C. V., CHEN, R., AGNO, J. E., GLASER, S., ANASTASSIADIS, K., STEWART, A. F. & MATZUK, M. M. 2010. MLL2 is required in oocytes for bulk histone 3 lysine 4 trimethylation and transcriptional silencing. *PLoS Biol*, 8.
- ANDREWS, N. C., ERDJUMENT-BROMAGE, H., DAVIDSON, M. B., TEMPST, P. & ORKIN, S. H. 1993. Erythroid transcription factor NF-E2 is a haematopoietic-specific basic-leucine zipper protein. *Nature*, 362, 722-8.
- ARDEHALI, M. B., MEI, A., ZOBECK, K. L., CARON, M., LIS, J. T. & KUSCH, T. 2011. Drosophila Set1 is the major histone H3 lysine 4 trimethyltransferase with role in transcription. *EMBO J*, 30, 2817-28.
- ARMSTRONG, J. A., BIEKER, J. J. & EMERSON, B. M. 1998. A SWI/SNF-related chromatin remodeling complex, E-RC1, is required for tissue-specific transcriptional regulation by EKLF in vitro. *Cell*, 95, 93-104.
- ARNOSTI, D. N. & KULKARNI, M. M. 2005. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J Cell Biochem*, 94, 890-8.
- BADIS, G., BERGER, M. F., PHILIPPAKIS, A. A., TALUKDER, S., GEHRKE, A. R., JAEGER, S. A., CHAN, E. T., METZLER, G., VEDENKO, A., CHEN, X., KUZNETSOV, H., WANG, C. F., COBURN, D., NEWBURGER, D. E., MORRIS, Q., HUGHES, T. R. & BULYK, M. L. 2009. Diversity and complexity in DNA recognition by transcription factors. *Science*, 324, 1720-3.
- BAI, L., CHARVIN, G., SIGGIA, E. D. & CROSS, F. R. 2010. Nucleosome-depleted regions in cell-cycle-regulated promoters ensure reliable gene expression in every cell cycle. *Dev Cell*, 18, 544-55.
- BAILEY, T. L., BODEN, M., BUSKE, F. A., FRITH, M., GRANT, C. E., CLEMENTI, L., REN, J., LI, W. W. & NOBLE, W. S. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*, 37, W202-8.
- BAJIC, V. B., TAN, S. L., CHRISTOFFELS, A., SCHONBACH, C., LIPOVICH, L., YANG, L., HOFMANN, O., KRUGER, A., HIDE, W., KAI, C., KAWAI, J., HUME, D. A., CARNINCI, P. & HAYASHIZAKI, Y. 2006. Mice and men: their promoter properties. *PLoS Genet*, 2, e54.
- BATES, D. L., CHEN, Y., KIM, G., GUO, L. & CHEN, L. 2008. Crystal structures of multiple GATA zinc fingers bound to DNA reveal new insights into DNA recognition and self-association by GATA. *J Mol Biol*, 381, 1292-306.
- BAUER, A., TRONCHE, F., WESSELY, O., KELLENDONK, C., REICHARDT, H. M., STEINLEIN, P., SCHUTZ, G. & BEUG, H. 1999. The glucocorticoid receptor is required for stress erythropoiesis. *Genes Dev*, 13, 2996-3002.
- BEDFORD, M. T. & CLARKE, S. G. 2009. Protein arginine methylation in mammals: who, what, and why. *Mol Cell*, 33, 1-13.

- BERGER, M. F., BADIS, G., GEHRKE, A. R., TALUKDER, S., PHILIPPAKIS, A. A., PENA-CASTILLO, L., ALLEYNE, T. M., MNAIMNEH, S., BOTVINNIK, O. B., CHAN, E. T., KHALID, F., ZHANG, W., NEWBURGER, D., JAEGER, S. A., MORRIS, Q. D., BULYK, M. L. & HUGHES, T. R. 2008. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*, 133, 1266-76.
- BERGER, M. F., PHILIPPAKIS, A. A., QURESHI, A. M., HE, F. S., ESTEP, P. W., 3RD & BULYK, M. L. 2006. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol*, 24, 1429-35.
- BIEKER, J. J. 2001. Kruppel-like factors: three fingers in many pies. *J Biol Chem*, 276, 34355-8.
- BISSWANGER, H. 2008. *Enzyme kinetics : principles and methods*, Weinheim, Wiley-VCH.
- BLACKLEDGE, N. P., ZHOU, J. C., TOLSTORUKOV, M. Y., FARCAS, A. M., PARK, P. J. & KLOSE, R. J. 2010. CpG islands recruit a histone H3 lysine 36 demethylase. *Mol Cell*, 38, 179-90.
- BLAZEK, E., MITTLER, G. & MEISTERERNST, M. 2005. The mediator of RNA polymerase II. *Chromosoma*, 113, 399-408.
- BLOBEL, G. A., NAKAJIMA, T., ECKNER, R., MONTMINY, M. & ORKIN, S. H. 1998. CREB-binding protein cooperates with transcription factor GATA-1 and is required for erythroid differentiation. *Proc Natl Acad Sci U S A*, 95, 2061-6.
- BORG, J., PATRINOS, G. P., FELICE, A. E. & PHILIPSEN, S. 2011. Erythroid phenotypes associated with KLF1 mutations. *Haematologica*, 96, 635-8.
- BOYES, J., BYFIELD, P., NAKATANI, Y. & OGRYZKO, V. 1998a. Regulation of activity of the transcription factor GATA-1 by acetylation. *Nature*, 396, 594-8.
- BOYES, J., OMICHINSKI, J., CLARK, D., PIKAART, M. & FELSENFELD, G. 1998b. Perturbation of nucleosome structure by the erythroid transcription factor GATA-1. *J Mol Biol*, 279, 529-44.
- BRADLEY, R. K., LI, X. Y., TRAPNELL, C., DAVIDSON, S., PACHTER, L., CHU, H. C., TONKIN, L. A., BIGGIN, M. D. & EISEN, M. B. 2010. Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related *Drosophila* species. *PLoS Biol*, 8, e1000343.
- BRESNICK, E. H., KATSUMURA, K. R., LEE, H. Y., JOHNSON, K. D. & PERKINS, A. S. 2012. Master regulatory GATA transcription factors: mechanistic principles and emerging links to hematologic malignancies. *Nucleic Acids Res*, 40, 5819-31.
- BRESNICK, E. H., MARTOWICZ, M. L., PAL, S. & JOHNSON, K. D. 2005. Developmental control via GATA factor interplay at chromatin domains. *J Cell Physiol*, 205, 1-9.
- BREWER, A. & PIZZEY, J. 2006. GATA factors in vertebrate heart development and disease. *Expert Rev Mol Med*, 8, 1-20.
- BROWN, R. C., PATTISON, S., VAN REE, J., COGHILL, E., PERKINS, A., JANE, S. M. & CUNNINGHAM, J. M. 2002. Distinct domains of erythroid Kruppel-like factor modulate chromatin remodeling and transactivation at the endogenous beta-globin gene promoter. *Mol Cell Biol*, 22, 161-70.
- BULGER, M. & GROUDINE, M. 2011. Functional and mechanistic diversity of distal transcription enhancers. *Cell*, 144, 327-39.
- BULTMAN, S. J., GEBUHR, T. C. & MAGNUSON, T. 2005. A Brg1 mutation that uncouples ATPase activity from chromatin remodeling reveals an essential role for SWI/SNF-related complexes in beta-globin expression and erythroid development. *Genes Dev*, 19, 2849-61.
- BULYK, M. L., HUANG, X., CHOO, Y. & CHURCH, G. M. 2001. Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc Natl Acad Sci U S A*, 98, 7158-63.

- BULYK, M. L., JOHNSON, P. L. & CHURCH, G. M. 2002. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res*, 30, 1255-61.
- CALHOUN, V. C., STATHOPOULOS, A. & LEVINE, M. 2002. Promoter-proximal tethering elements regulate enhancer-promoter specificity in the *Drosophila* Antennapedia complex. *Proc Natl Acad Sci U S A*, 99, 9243-7.
- CAMPBELL, A. E., WILKINSON-WHITE, L., MACKAY, J. P., MATTHEWS, J. M. & BLOBEL, G. A. 2013. Analysis of disease-causing GATA1 mutations in murine gene complementation systems. *Blood*, 121, 5218-27.
- CANTOR, A. B., IWASAKI, H., ARINOBU, Y., MORAN, T. B., SHIGEMATSU, H., SULLIVAN, M. R., AKASHI, K. & ORKIN, S. H. 2008. Antagonism of FOG-1 and GATA factors in fate choice for the mast cell lineage. *J Exp Med*, 205, 611-24.
- CANTOR, A. B., KATZ, S. G. & ORKIN, S. H. 2002. Distinct domains of the GATA-1 cofactor FOG-1 differentially influence erythroid versus megakaryocytic maturation. *Mol Cell Biol*, 22, 4268-79.
- CANTOR, A. B. & ORKIN, S. H. 2002. Transcriptional regulation of erythropoiesis: an affair involving multiple partners. *Oncogene*, 21, 3368-76.
- CARLSON, C. D., WARREN, C. L., HAUSCHILD, K. E., OZERS, M. S., QADIR, N., BHIMSARIA, D., LEE, Y., CERRINA, F. & ANSARI, A. Z. 2010. Specificity landscapes of DNA binding molecules elucidate biological function. *Proc Natl Acad Sci U S A*, 107, 4544-9.
- CHEN, X. & BIEKER, J. J. 2001. Unanticipated repression function linked to erythroid Kruppel-like factor. *Mol Cell Biol*, 21, 3118-25.
- CHEN, X. & BIEKER, J. J. 2004. Stage-specific repression by the EKLF transcriptional activator. *Mol Cell Biol*, 24, 10416-24.
- CHEN, Y., BATES, D. L., DEY, R., CHEN, P. H., MACHADO, A. C., LAIRD-OFFRINGA, I. A., ROHS, R. & CHEN, L. 2012. DNA binding by GATA transcription factor suggests mechanisms of DNA looping and long-range gene regulation. *Cell Rep*, 2, 1197-206.
- CHLON, T. M. & CRISPINO, J. D. 2012. Combinatorial regulation of tissue specification by GATA and FOG factors. *Development*, 139, 3905-16.
- CHLON, T. M., DORE, L. C. & CRISPINO, J. D. 2012. Cofactor-mediated restriction of GATA-1 chromatin occupancy coordinates lineage-specific gene expression. *Mol Cell*, 47, 608-21.
- CLORE, G. M., BAX, A., OMICHINSKI, J. G. & GRONENBORN, A. M. 1994. Localization of bound water in the solution structure of a complex of the erythroid transcription factor GATA-1 with DNA. *Structure*, 2, 89-94.
- CLOUAIRE, T., WEBB, S., SKENE, P., ILLINGWORTH, R., KERR, A., ANDREWS, R., LEE, J. H., SKALNIK, D. & BIRD, A. 2012. Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. *Genes Dev*, 26, 1714-28.
- COOPER, M. A. 2002. Optical biosensors in drug discovery. *Nat Rev Drug Discov*, 1, 515-28.
- COULON, A., CHOW, C. C., SINGER, R. H. & LARSON, D. R. 2013. Eukaryotic transcriptional dynamics: from single molecules to cell populations. *Nat Rev Genet*, 14, 572-84.
- CRISPINO, J. D., LODISH, M. B., MACKAY, J. P. & ORKIN, S. H. 1999. Use of altered specificity mutants to probe a specific protein-protein interaction in differentiation: the GATA-1:FOG complex. *Mol Cell*, 3, 219-28.
- CROOKS, G. E., HON, G., CHANDONIA, J. M. & BRENNER, S. E. 2004. WebLogo: a sequence logo generator. *Genome Res*, 14, 1188-90.
- CROSSLEY, M. & ORKIN, S. H. 1994. Phosphorylation of the erythroid transcription factor GATA-1. *J Biol Chem*, 269, 16589-96.

- CUI, K., ZANG, C., ROH, T. Y., SCHONES, D. E., CHILDS, R. W., PENG, W. & ZHAO, K. 2009. Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell*, 4, 80-93.
- CUMANO, A. & GODIN, I. 2007. Ontogeny of the hematopoietic system. *Annu Rev Immunol*, 25, 745-85.
- DE BOER, E., RODRIGUEZ, P., BONTE, E., KRIJGSVELD, J., KATSANTONI, E., HECK, A., GROSVELD, F. & STROUBOULIS, J. 2003. Efficient biotinylation and single-step purification of tagged transcription factors in mammalian cells and transgenic mice. *Proc Natl Acad Sci U S A*, 100, 7480-5.
- DEATON, A. M. & BIRD, A. 2011. CpG islands and the regulation of transcription. *Genes Dev*, 25, 1010-22.
- DEGNER, J. F., PAI, A. A., PIQUE-REGI, R., VEYRIERAS, J. B., GAFFNEY, D. J., PICKRELL, J. K., DE LEON, S., MICHELINI, K., LEWELLEN, N., CRAWFORD, G. E., STEPHENS, M., GILAD, Y. & PRITCHARD, J. K. 2012. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, 482, 390-4.
- DENG, W., LEE, J., WANG, H., MILLER, J., REIK, A., GREGORY, P. D., DEAN, A. & BLOBEL, G. A. 2012. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell*, 149, 1233-44.
- DION, M. F., KAPLAN, T., KIM, M., BURATOWSKI, S., FRIEDMAN, N. & RANDO, O. J. 2007. Dynamics of replication-independent histone turnover in budding yeast. *Science*, 315, 1405-8.
- DOLZNIG, H., GREBIEN, F., DEINER, E. M., STANGL, K., KOLBUS, A., HABERMANN, B., KERENYI, M. A., KIESLINGER, M., MORIGGL, R., BEUG, H. & MULLNER, E. W. 2006. Erythroid progenitor renewal versus differentiation: genetic evidence for cell autonomous, essential functions of EpoR, Stat5 and the GR. *Oncogene*, 25, 2890-900.
- DORE, L. C., CHLON, T. M., BROWN, C. D., WHITE, K. P. & CRISPINO, J. D. 2012. Chromatin occupancy analysis reveals genome-wide GATA factor switching during hematopoiesis. *Blood*, 119, 3724-33.
- DREIJERINK, K. M., MULDER, K. W., WINKLER, G. S., HOPPENER, J. W., LIPS, C. J. & TIMMERS, H. T. 2006. Menin links estrogen receptor activation to histone H3K4 trimethylation. *Cancer Res*, 66, 4929-35.
- DREIJERINK, K. M., VARIER, R. A., VAN BEEKUM, O., JENINGA, E. H., HOPPENER, J. W., LIPS, C. J., KUMMER, J. A., KALKHOVEN, E. & TIMMERS, H. T. 2009. The multiple endocrine neoplasia type 1 (MEN1) tumor suppressor regulates peroxisome proliferator-activated receptor gamma-dependent adipocyte differentiation. *Mol Cell Biol*, 29, 5060-9.
- DRISSEN, R., PALSTRA, R. J., GILLEMANS, N., SPLINTER, E., GROSVELD, F., PHILIPSEN, S. & DE LAAT, W. 2004. The active spatial organization of the beta-globin locus requires the transcription factor EKLF. *Genes Dev*, 18, 2485-90.
- DRISSEN, R., VON LINDERN, M., KOLBUS, A., DRIEGEN, S., STEINLEIN, P., BEUG, H., GROSVELD, F. & PHILIPSEN, S. 2005. The erythroid phenotype of EKLF-null mice: defects in hemoglobin metabolism and membrane stability. *Mol Cell Biol*, 25, 5205-14.
- EECKHOUTE, J., KEETON, E. K., LUPIEN, M., KRUM, S. A., CARROLL, J. S. & BROWN, M. 2007. Positive cross-regulatory loop ties GATA-3 to estrogen receptor alpha expression in breast cancer. *Cancer Res*, 67, 6477-83.
- EISBACHER, M., HOLMES, M. L., NEWTON, A., HOGG, P. J., KHACHIGIAN, L. M., CROSSLEY, M. & CHONG, B. H. 2003. Protein-protein interaction between Fli-1 and GATA-1 mediates synergistic expression of megakaryocyte-specific genes through cooperative DNA binding. *Mol Cell Biol*, 23, 3427-41.
- EL OMARI, K., HOOSDALLY, S. J., TULADHAR, K., KARIA, D., VYAS, P., PATIENT, R., PORCHER, C. & MANCINI, E. J. 2011. Structure of the leukemia oncogene LMO2: implications for the assembly of a hematopoietic transcription factor complex. *Blood*, 117, 2146-56.

- ELAGIB, K. E., RACKE, F. K., MOGASS, M., KHETAWAT, R., DELEHANTY, L. L. & GOLDFARB, A. N. 2003. RUNX1 and GATA-1 coexpression and cooperation in megakaryocytic differentiation. *Blood*, 101, 4333-41.
- ELLIOTT, S., PHAM, E. & MACDOUGALL, I. C. 2008. Erythropoietins: a common mechanism of action. *Exp Hematol*, 36, 1573-84.
- ERNST, J. & KELLIS, M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods*, 9, 215-6.
- ESCAMILLA-DEL-ARENAL, M. & RECILLAS-TARGA, F. 2008. GATA-1 modulates the chromatin structure and activity of the chicken alpha-globin 3' enhancer. *Mol Cell Biol*, 28, 575-86.
- FARACE, M. G., BROWN, B. A., RASCHELLA, G., ALEXANDER, J., GAMBARI, R., FANTONI, A., HARDIES, S. C., HUTCHISON, C. A., 3RD & EDGELL, M. H. 1984. The mouse beta h1 gene codes for the z chain of embryonic hemoglobin. *J Biol Chem*, 259, 7123-8.
- FENG, W. C., SOUTHWOOD, C. M. & BIEKER, J. J. 1994. Analyses of beta-thalassemia mutant DNA interactions with erythroid Kruppel-like factor (EKLF), an erythroid cell-specific transcription factor. *J Biol Chem*, 269, 1493-500.
- FERDOUS, A., SIKDER, D., GILLETTE, T., NALLEY, K., KODADEK, T. & JOHNSTON, S. A. 2007. The role of the proteasomal ATPases and activator monoubiquitylation in regulating Gal4 binding to promoters. *Genes Dev*, 21, 112-23.
- FLETCHER, T. M., XIAO, N., MAUTINO, G., BAUMANN, C. T., WOLFORD, R., WARREN, B. S. & HAGER, G. L. 2002. ATP-dependent mobilization of the glucocorticoid receptor during chromatin remodeling. *Mol Cell Biol*, 22, 3255-63.
- FORDYCE, P. M., GERBER, D., TRAN, D., ZHENG, J., LI, H., DERISI, J. L. & QUAKE, S. R. 2010. De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. *Nat Biotechnol*, 28, 970-5.
- FOX, A. H., KOWALSKI, K., KING, G. F., MACKAY, J. P. & CROSSLEY, M. 1998. Key residues characteristic of GATA N-fingers are recognized by FOG. *J Biol Chem*, 273, 33595-603.
- FOX, A. H., LIEW, C., HOLMES, M., KOWALSKI, K., MACKAY, J. & CROSSLEY, M. 1999. Transcriptional cofactors of the FOG family interact with GATA proteins by means of multiple zinc fingers. *EMBO J*, 18, 2812-22.
- FREEMAN, B. C. & YAMAMOTO, K. R. 2002. Disassembly of transcriptional regulatory complexes by molecular chaperones. *Science*, 296, 2232-5.
- FRIEND, C., SCHER, W., HOLLAND, J. G. & SATO, T. 1971. Hemoglobin synthesis in murine virus-induced leukemic cells in vitro: stimulation of erythroid differentiation by dimethyl sulfoxide. *Proc Natl Acad Sci U S A*, 68, 378-82.
- FUDA, N. J., ARDEHALI, M. B. & LIS, J. T. 2009. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature*, 461, 186-92.
- FUJIWARA, T., O'GEEN, H., KELES, S., BLAHNIK, K., LINNEMANN, A. K., KANG, Y. A., CHOI, K., FARNHAM, P. J. & BRESNICK, E. H. 2009. Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. *Mol Cell*, 36, 667-81.
- FUJIWARA, Y., BROWNE, C. P., CUNNIFF, K., GOFF, S. C. & ORKIN, S. H. 1996. Arrested development of embryonic red cell precursors in mouse embryos lacking transcription factor GATA-1. *Proc Natl Acad Sci U S A*, 93, 12355-8.
- GALLOWAY, J. L., WINGERT, R. A., THISSE, C., THISSE, B. & ZON, L. I. 2005. Loss of gata1 but not gata2 converts erythropoiesis to myelopoiesis in zebrafish embryos. *Dev Cell*, 8, 109-16.
- GAO, Z., HUANG, Z., OLIVEY, H. E., GURBUXANI, S., CRISPINO, J. D. & SVENSSON, E. C. 2010. FOG-1-mediated recruitment of NuRD is required for cell lineage re-enforcement during haematopoiesis. *EMBO J*, 29, 457-68.
- GAREL, A. & AXEL, R. 1976. Selective digestion of transcriptionally active ovalbumin genes from oviduct nuclei. *Proc Natl Acad Sci U S A*, 73, 3966-70.

- GASIOREK, J. J., NOUHI, Z. & BLANK, V. 2012. Abnormal differentiation of erythroid precursors in p45 NF-E2(-/-) mice. *Exp Hematol*, 40, 393-400.
- GASZNER, M. & FELSENFELD, G. 2006. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet*, 7, 703-13.
- GEERTZ, M. & MAERKL, S. J. 2010. Experimental strategies for studying transcription factor-DNA binding specificities. *Brief Funct Genomics*, 9, 362-73.
- GERLAND, U., MOROZ, J. D. & HWA, T. 2002. Physical constraints and functional characteristics of transcription factor-DNA interaction. *Proc Natl Acad Sci U S A*, 99, 12015-20.
- GHAFFARI, S., KITIDIS, C., ZHAO, W., MARINKOVIC, D., FLEMING, M. D., LUO, B., MARSZALEK, J. & LODISH, H. F. 2006. AKT induces erythroid-cell maturation of JAK2-deficient fetal liver progenitor cells and is required for Epo regulation of erythroid-cell differentiation. *Blood*, 107, 1888-91.
- GHIRLANDO, R. & TRAINOR, C. D. 2000. GATA-1 bends DNA in a site-independent fashion. *J Biol Chem*, 275, 28152-6.
- GHIRLANDO, R. & TRAINOR, C. D. 2003. Determinants of GATA-1 binding to DNA: the role of non-finger residues. *J Biol Chem*, 278, 45620-8.
- GHISLETTI, S., BAROZZI, I., MIETTON, F., POLLETTI, S., DE SANTA, F., VENTURINI, E., GREGORY, L., LONIE, L., CHEW, A., WEI, C. L., RAGOISSIS, J. & NATOLI, G. 2010. Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. *Immunity*, 32, 317-28.
- GILDEA, J. J., LOPEZ, R. & SHEARN, A. 2000. A screen for new trithorax group genes identified little imaginal discs, the Drosophila melanogaster homologue of human retinoblastoma binding protein 2. *Genetics*, 156, 645-63.
- GORDAN, R., HARTEMINK, A. J. & BULYK, M. L. 2009. Distinguishing direct versus indirect transcription factor-DNA interactions. *Genome Res*, 19, 2090-100.
- GOSSETT, A. J. & LIEB, J. D. 2012. In vivo effects of histone H3 depletion on nucleosome occupancy and position in Saccharomyces cerevisiae. *PLoS Genet*, 8, e1002771.
- GRASS, J. A., BOYER, M. E., PAL, S., WU, J., WEISS, M. J. & BRESNICK, E. H. 2003. GATA-1-dependent transcriptional repression of GATA-2 via disruption of positive autoregulation and domain-wide chromatin remodeling. *Proc Natl Acad Sci U S A*, 100, 8811-6.
- GRASS, J. A., JING, H., KIM, S. I., MARTOWICZ, M. L., PAL, S., BLOBEL, G. A. & BRESNICK, E. H. 2006. Distinct functions of dispersed GATA factor complexes at an endogenous gene locus. *Mol Cell Biol*, 26, 7056-67.
- GREENE, M. E., MUNDSCHAU, G., WECHSLER, J., MCDEVITT, M., GAMIS, A., KARP, J., GURBUXANI, S., ARCECI, R. & CRISPINO, J. D. 2003. Mutations in GATA1 in both transient myeloproliferative disorder and acute megakaryoblastic leukemia of Down syndrome. *Blood Cells Mol Dis*, 31, 351-6.
- GREGORY, G. D., MICCIO, A., BERSENEV, A., WANG, Y., HONG, W., ZHANG, Z., PONCZ, M., TONG, W. & BLOBEL, G. A. 2010. FOG1 requires NuRD to promote hematopoiesis and maintain lineage fidelity within the megakaryocytic-erythroid compartment. *Blood*, 115, 2156-66.
- GREGORY, R. C., TAXMAN, D. J., SESHASAYEE, D., KENSINGER, M. H., BIEKER, J. J. & WOJCHOWSKI, D. M. 1996. Functional interaction of GATA1 with erythroid Kruppel-like factor and Sp1 at defined erythroid promoters. *Blood*, 87, 1793-801.
- GRIMALDI, Y., FERRARI, P. & STRUBIN, M. 2014. Independent RNA polymerase II preinitiation complex dynamics and nucleosome turnover at promoter sites in vivo. *Genome Res*, 24, 117-24.
- GROWNEY, J. D., SHIGEMATSU, H., LI, Z., LEE, B. H., ADELSPERGER, J., ROWAN, R., CURLEY, D. P., KUTOK, J. L., AKASHI, K., WILLIAMS, I. R., SPECK, N. A. & GILLILAND, D. G. 2005. Loss of Runx1 perturbs adult hematopoiesis and is associated with a myeloproliferative phenotype. *Blood*, 106, 494-504.

- GUCCIONE, E., BASSI, C., CASADIO, F., MARTINATO, F., CESARONI, M., SCHUCHLAUTZ, H., LUSCHER, B. & AMATI, B. 2007. Methylation of histone H3R2 by PRMT6 and H3K4 by an MLL complex are mutually exclusive. *Nature*, 449, 933-7.
- GUERTIN, M. J., MARTINS, A. L., SIEPEL, A. & LIS, J. T. 2012. Accurate prediction of inducible transcription factor binding intensities in vivo. *PLoS Genet*, 8, e1002610.
- GUILLEMETTE, B., DROGARIS, P., LIN, H. H., ARMSTRONG, H., HIRAGAMI-HAMADA, K., IMHOF, A., BONNEIL, E., THIBAUT, P., VERREAULT, A. & FESTENSTEIN, R. J. 2011. H3 lysine 4 is acetylated at active gene promoters and is regulated by H3 lysine 4 methylation. *PLoS Genet*, 7, e1001354.
- GUO, C., CHANG, C. C., WORTHAM, M., CHEN, L. H., KERNAGIS, D. N., QIN, X., CHO, Y. W., CHI, J. T., GRANT, G. A., MCLENDON, R. E., YAN, H., GE, K., PAPADOPOULOS, N., BIGNER, D. D. & HE, Y. 2012. Global identification of MLL2-targeted loci reveals MLL2's role in diverse signaling pathways. *Proc Natl Acad Sci U S A*, 109, 17603-8.
- HAENLIN, M., CUBADDA, Y., BLONDEAU, F., HEITZLER, P., LUTZ, Y., SIMPSON, P. & RAMAIN, P. 1997. Transcriptional activity of pannier is regulated negatively by heterodimerization of the GATA DNA-binding domain with a cofactor encoded by the u-shaped gene of Drosophila. *Genes Dev*, 11, 3096-108.
- HAGER, G. L., MCNALLY, J. G. & MISTELI, T. 2009. Transcription dynamics. *Mol Cell*, 35, 741-53.
- HALL, M. A., CURTIS, D. J., METCALF, D., ELEFANTY, A. G., SOURRIS, K., ROBB, L., GOTHERT, J. R., JANE, S. M. & BEGLEY, C. G. 2003. The critical regulator of embryonic hematopoiesis, SCL, is vital in the adult for megakaryopoiesis, erythropoiesis, and lineage choice in CFU-S12. *Proc Natl Acad Sci U S A*, 100, 992-7.
- HALLSON, G., HOLLEBAKKEN, R. E., LI, T., SYRZYCKA, M., KIM, I., COTSWORTH, S., FITZPATRICK, K. A., SINCLAIR, D. A. & HONDA, B. M. 2012. dSet1 is the main H3K4 di- and tri-methyltransferase throughout Drosophila development. *Genetics*, 190, 91-100.
- HANSEN, J. N., KONKEL, D. A. & LEDER, P. 1982. The sequence of a mouse embryonic beta-globin gene. Evolution of the gene and its signal region. *J Biol Chem*, 257, 1048-52.
- HART, A., MELET, F., GROSSFELD, P., CHIEN, K., JONES, C., TUNNACLIFFE, A., FAVIER, R. & BERNSTEIN, A. 2000. Fli-1 is required for murine vascular and megakaryocytic development and is hemizygously deleted in patients with thrombocytopenia. *Immunity*, 13, 167-77.
- HATTANGADI, S. M., WONG, P., ZHANG, L., FLYGARE, J. & LODISH, H. F. 2011. From stem cell to red cell: regulation of erythropoiesis at multiple levels by multiple proteins, RNAs, and chromatin modifications. *Blood*, 118, 6258-68.
- HEINTZMAN, N. D., HON, G. C., HAWKINS, R. D., KHERADPOUR, P., STARK, A., HARP, L. F., YE, Z., LEE, L. K., STUART, R. K., CHING, C. W., CHING, K. A., ANTOSIEWICZ-BOURGET, J. E., LIU, H., ZHANG, X., GREEN, R. D., LOBANENKOV, V. V., STEWART, R., THOMSON, J. A., CRAWFORD, G. E., KELLIS, M. & REN, B. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459, 108-12.
- HEINZ, S., BENNER, C., SPANN, N., BERTOLINO, E., LIN, Y. C., LASLO, P., CHENG, J. X., MURRE, C., SINGH, H. & GLASS, C. K. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*, 38, 576-89.
- HERGLOTZ, J., KUVARDINA, O. N., KOLODZIEJ, S., KUMAR, A., HUSSONG, H., GREZ, M. & LAUSEN, J. 2013. Histone arginine methylation keeps RUNX1 target genes in an intermediate state. *Oncogene*, 32, 2565-75.

- HERZ, H. M., GARRUSS, A. & SHILATIFARD, A. 2013. SET for life: biochemical activities and biological functions of SET domain-containing proteins. *Trends Biochem Sci*, 38, 621-39.
- HESSELBERTH, J. R., CHEN, X., ZHANG, Z., SABO, P. J., SANDSTROM, R., REYNOLDS, A. P., THURMAN, R. E., NEPH, S., KUEHN, M. S., NOBLE, W. S., FIELDS, S. & STAMATOYANNOPOULOS, J. A. 2009. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods*, 6, 283-9.
- HO, I. C., TAI, T. S. & PAI, S. Y. 2009. GATA3 and the T-cell lineage: essential functions before and after T-helper-2-cell differentiation. *Nat Rev Immunol*, 9, 125-35.
- HODGE, D., COGHILL, E., KEYS, J., MAGUIRE, T., HARTMANN, B., MCDOWALL, A., WEISS, M., GRIMMOND, S. & PERKINS, A. 2006. A global role for EKLF in definitive and primitive erythropoiesis. *Blood*, 107, 3359-70.
- HOLMQVIST, P. H. & MANNERVIK, M. 2013. Genomic occupancy of the transcriptional co-activators p300 and CBP. *Transcription*, 4, 18-23.
- HONG, W., NAKAZAWA, M., CHEN, Y. Y., KORI, R., VAKOC, C. R., RAKOWSKI, C. & BLOBEL, G. A. 2005. FOG-1 recruits the NuRD repressor complex to mediate transcriptional repression by GATA-1. *EMBO J*, 24, 2367-78.
- HORIUCHI, S., ONODERA, A., HOSOKAWA, H., WATANABE, Y., TANAKA, T., SUGANO, S., SUZUKI, Y. & NAKAYAMA, T. 2011. Genome-wide analysis reveals unique regulation of transcription of Th2-specific genes by GATA3. *J Immunol*, 186, 6378-89.
- HU, G., SCHONES, D. E., CUI, K., YBARRA, R., NORTHRUP, D., TANG, Q., GATTINONI, L., RESTIFO, N. P., HUANG, S. & ZHAO, K. 2011. Regulation of nucleosome landscape and transcription factor targeting at tissue-specific enhancers by BRG1. *Genome Res*, 21, 1650-8.
- HUANG DA, W., SHERMAN, B. T. & LEMPICKI, R. A. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4, 44-57.
- HUEBERT, D. J., KUAN, P. F., KELES, S. & GASCH, A. P. 2012. Dynamic changes in nucleosome occupancy are not predictive of gene expression dynamics but are linked to transcription and chromatin regulators. *Mol Cell Biol*, 32, 1645-53.
- HUNG, H. L., LAU, J., KIM, A. Y., WEISS, M. J. & BLOBEL, G. A. 1999. CREB-Binding protein acetylates hematopoietic transcription factor GATA-1 at functionally important sites. *Mol Cell Biol*, 19, 3496-505.
- HURTADO, A., HOLMES, K. A., ROSS-INNES, C. S., SCHMIDT, D. & CARROLL, J. S. 2011. FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nat Genet*, 43, 27-33.
- HUTCHISON, C. A., 3RD, HARDIES, S. C., PADGETT, R. W., WEAVER, S. & EDGELL, M. H. 1984. The mouse globin pseudogene beta h3 is descended from a premammalian delta-globin gene. *J Biol Chem*, 259, 12881-9.
- HYLLUS, D., STEIN, C., SCHNABEL, K., SCHILTZ, E., IMHOF, A., DOU, Y., HSIEH, J. & BAUER, U. M. 2007. PRMT6-mediated methylation of R2 in histone H3 antagonizes H3 K4 trimethylation. *Genes Dev*, 21, 3369-80.
- ICHIKAWA, M., ASAI, T., SAITO, T., SEO, S., YAMAZAKI, I., YAMAGATA, T., MITANI, K., CHIBA, S., OGAWA, S., KUROKAWA, M. & HIRAI, H. 2004. AML-1 is required for megakaryocytic maturation and lymphocytic differentiation, but not for maintenance of hematopoietic stem cells in adult hematopoiesis. *Nat Med*, 10, 299-304.
- ILLINGWORTH, R. S., GRUENEWALD-SCHNEIDER, U., WEBB, S., KERR, A. R., JAMES, K. D., TURNER, D. J., SMITH, C., HARRISON, D. J., ANDREWS, R. & BIRD, A. P. 2010. Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet*, 6, e1001134.
- IMACHI, H., MURAO, K., DOBASHI, H., BHUYAN, M. M., CAO, X., KONTANI, K., NIKI, S., MURAZAWA, C., NAKAJIMA, H., KOHNO, N., YAMASHITA, H., IWASE, H., HAYASHI, S., ISHIDA, T. & YAMAUCHI, A. 2010. Menin, a product of the MEN1

- gene, binds to estrogen receptor to enhance its activity in breast cancer cells: possibility of a novel predictive factor for tamoxifen resistance. *Breast Cancer Res Treat*, 122, 395-407.
- JACOBSEN, C. M., NARITA, N., BIELINSKA, M., SYDER, A. J., GORDON, J. I. & WILSON, D. B. 2002. Genetic mosaic analysis reveals that GATA-4 is required for proper differentiation of mouse gastric epithelium. *Dev Biol*, 241, 34-46.
- JI, H., JIANG, H., MA, W., JOHNSON, D. S., MYERS, R. M. & WONG, W. H. 2008. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol*, 26, 1293-300.
- JIANG, H., SHUKLA, A., WANG, X., CHEN, W. Y., BERNSTEIN, B. E. & ROEDER, R. G. 2011. Role for Dpy-30 in ES cell-fate specification by regulation of H3K4 methylation within bivalent domains. *Cell*, 144, 513-25.
- JIMENEZ, G., GALE, K. B. & ENVER, T. 1992. The mouse beta-globin locus control region: hypersensitive sites 3 and 4. *Nucleic Acids Res*, 20, 5797-803.
- JIN, Q., YU, L. R., WANG, L., ZHANG, Z., KASPER, L. H., LEE, J. E., WANG, C., BRINDLE, P. K., DENT, S. Y. & GE, K. 2011. Distinct roles of GCN5/PCAF-mediated H3K9ac and CBP/p300-mediated H3K18/27ac in nuclear receptor transactivation. *EMBO J*, 30, 249-62.
- JING, H., VAKOC, C. R., YING, L., MANDAT, S., WANG, H., ZHENG, X. & BLOBEL, G. A. 2008. Exchange of GATA factors mediates transitions in looped chromatin organization at a developmentally regulated gene locus. *Mol Cell*, 29, 232-42.
- JOHN, L. B. & WARD, A. C. 2011. The Ikaros gene family: transcriptional regulators of hematopoiesis and immunity. *Mol Immunol*, 48, 1272-8.
- JOHN, S., SABO, P. J., JOHNSON, T. A., SUNG, M. H., BIDDIE, S. C., LIGHTMAN, S. L., VOSS, T. C., DAVIS, S. R., MELTZER, P. S., STAMATOYANNOPOULOS, J. A. & HAGER, G. L. 2008. Interaction of the glucocorticoid receptor with the chromatin landscape. *Mol Cell*, 29, 611-24.
- JOHN, S., SABO, P. J., THURMAN, R. E., SUNG, M. H., BIDDIE, S. C., JOHNSON, T. A., HAGER, G. L. & STAMATOYANNOPOULOS, J. A. 2011. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat Genet*, 43, 264-8.
- JOHNSON, D. S., MORTAZAVI, A., MYERS, R. M. & WOLD, B. 2007a. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316, 1497-502.
- JOHNSON, K. D., BOYER, M. E., KANG, J. A., WICKREMA, A., CANTOR, A. B. & BRESNICK, E. H. 2007b. Friend of GATA-1-independent transcriptional repression: a novel mode of GATA-1 function. *Blood*, 109, 5230-3.
- JOHNSON, K. D., KIM, S. I. & BRESNICK, E. H. 2006. Differential sensitivities of transcription factor target genes underlie cell type-specific gene expression profiles. *Proc Natl Acad Sci U S A*, 103, 15939-44.
- JOLMA, A., KIVIOJA, T., TOIVONEN, J., CHENG, L., WEI, G., ENGE, M., TAIPALE, M., VAQUERIZAS, J. M., YAN, J., SILLANPAA, M. J., BONKE, M., PALIN, K., TALUKDER, S., HUGHES, T. R., LUSCOMBE, N. M., UKKONEN, E. & TAIPALE, J. 2010. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res*, 20, 861-73.
- JOLMA, A. & TAIPALE, J. 2011. Methods for Analysis of Transcription Factor DNA-Binding Specificity In Vitro. *Subcell Biochem*, 52, 155-73.
- JOLMA, A., YAN, J., WHITTINGTON, T., TOIVONEN, J., NITTA, K. R., RASTAS, P., MORGUNOVA, E., ENGE, M., TAIPALE, M., WEI, G., PALIN, K., VAQUERIZAS, J. M., VINCENTELLI, R., LUSCOMBE, N. M., HUGHES, T. R., LEMAIRE, P., UKKONEN, E., KIVIOJA, T. & TAIPALE, J. 2013. DNA-binding specificities of human transcription factors. *Cell*, 152, 327-39.
- JUNION, G., SPIVAKOV, M., GIRARDOT, C., BRAUN, M., GUSTAFSON, E. H., BIRNEY, E. & FURLONG, E. E. 2012. A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell*, 148, 473-86.

- JUVEN-GERSHON, T., HSU, J. Y., THEISEN, J. W. & KADONAGA, J. T. 2008. The RNA polymerase II core promoter - the gateway to transcription. *Curr Opin Cell Biol*, 20, 253-9.
- KACZYNSKI, J., COOK, T. & URRUTIA, R. 2003. Sp1- and Kruppel-like transcription factors. *Genome Biol*, 4, 206.
- KADAM, S. & EMERSON, B. M. 2003. Transcriptional specificity of human SWI/SNF BRG1 and BRM chromatin remodeling complexes. *Mol Cell*, 11, 377-89.
- KADAM, S., MCALPINE, G. S., PHELAN, M. L., KINGSTON, R. E., JONES, K. A. & EMERSON, B. M. 2000. Functional selectivity of recombinant mammalian SWI/SNF subunits. *Genes Dev*, 14, 2441-51.
- KADRI, Z., MAOUCHE-CHRETIEN, L., ROOKE, H. M., ORKIN, S. H., ROMEO, P. H., MAYEUX, P., LEBOULCH, P. & CHRETIEN, S. 2005. Phosphatidylinositol 3-kinase/Akt induced by erythropoietin renders the erythroid differentiation factor GATA-1 competent for TIMP-1 gene transactivation. *Mol Cell Biol*, 25, 7412-22.
- KAERN, M., ELSTON, T. C., BLAKE, W. J. & COLLINS, J. J. 2005. Stochasticity in gene expression: from theories to phenotypes. *Nat Rev Genet*, 6, 451-64.
- KANEKO, H., KOBAYASHI, E., YAMAMOTO, M. & SHIMIZU, R. 2012. N- and C-terminal transactivation domains of GATA1 protein coordinate hematopoietic program. *J Biol Chem*, 287, 21439-49.
- KARPOVA, T. S., CHEN, T. Y., SPRAGUE, B. L. & MCNALLY, J. G. 2004. Dynamic interactions of a transcription factor with DNA are accelerated by a chromatin remodeller. *EMBO Rep*, 5, 1064-70.
- KASSOUF, M. T., CHAGRAOUI, H., VYAS, P. & PORCHER, C. 2008. Differential use of SCL/TAL-1 DNA-binding domain in developmental hematopoiesis. *Blood*, 112, 1056-67.
- KASSOUF, M. T., HUGHES, J. R., TAYLOR, S., MCGOWAN, S. J., SONEJI, S., GREEN, A. L., VYAS, P. & PORCHER, C. 2010. Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. *Genome Res*, 20, 1064-83.
- KATZ, S. G., CANTOR, A. B. & ORKIN, S. H. 2002. Interaction between FOG-1 and the corepressor C-terminal binding protein is dispensable for normal erythropoiesis in vivo. *Mol Cell Biol*, 22, 3121-8.
- KAWADA, H., ITO, T., PHARR, P. N., SPYROPOULOS, D. D., WATSON, D. K. & OGAWA, M. 2001. Defective megakaryopoiesis and abnormal erythroid development in Fli-1 gene-targeted mice. *Int J Hematol*, 73, 463-8.
- KIM, D. H., LEE, J., LEE, B. & LEE, J. W. 2009a. ASCOM controls farnesoid X receptor transactivation through its associated histone H3 lysine 4 methyltransferase activity. *Mol Endocrinol*, 23, 1556-62.
- KIM, J. W., TANG, Q. Q., LI, X. & LANE, M. D. 2007a. Effect of phosphorylation and S-S bond-induced dimerization on DNA binding and transcriptional activation by C/EBPbeta. *Proc Natl Acad Sci U S A*, 104, 1800-4.
- KIM, S. I. & BRESNICK, E. H. 2007. Transcriptional control of erythropoiesis: emerging mechanisms and principles. *Oncogene*, 26, 6777-94.
- KIM, S. I., BRESNICK, E. H. & BULTMAN, S. J. 2009b. BRG1 directly regulates nucleosome structure and chromatin looping of the alpha globin locus to activate transcription. *Nucleic Acids Res*, 37, 6019-27.
- KIM, S. I., BULTMAN, S. J., JING, H., BLOBEL, G. A. & BRESNICK, E. H. 2007b. Dissecting molecular steps in chromatin domain activation during hematopoietic differentiation. *Mol Cell Biol*, 27, 4551-65.
- KIM, S. I., BULTMAN, S. J., KIEFER, C. M., DEAN, A. & BRESNICK, E. H. 2009c. BRG1 requirement for long-range interaction of a locus control region with a downstream promoter. *Proc Natl Acad Sci U S A*, 106, 2259-64.
- KIM, T. & BURATOWSKI, S. 2009. Dimethylation of H3K4 by Set1 recruits the Set3 histone deacetylase complex to 5' transcribed regions. *Cell*, 137, 259-72.

- KIM, Y. J. & LIS, J. T. 2005. Interactions between subunits of Drosophila Mediator and activator proteins. *Trends Biochem Sci*, 30, 245-9.
- KIRMIZIS, A., SANTOS-ROSA, H., PENKETT, C. J., SINGER, M. A., VERMEULEN, M., MANN, M., BAHLER, J., GREEN, R. D. & KOUZARIDES, T. 2007. Arginine methylation at histone H3R2 controls deposition of H3K4 trimethylation. *Nature*, 449, 928-32.
- KITABAYASHI, I., AIKAWA, Y., NGUYEN, L. A., YOKOYAMA, A. & OHKI, M. 2001. Activation of AML1-mediated transcription by MOZ and inhibition by the MOZ-CBP fusion protein. *EMBO J*, 20, 7184-96.
- KITABAYASHI, I., YOKOYAMA, A., SHIMIZU, K. & OHKI, M. 1998. Interaction and functional cooperation of the leukemia-associated factors AML1 and p300 in myeloid cell differentiation. *EMBO J*, 17, 2994-3004.
- KLECKNER, I. R. & FOSTER, M. P. 2011. An introduction to NMR-based approaches for measuring protein dynamics. *Biochim Biophys Acta*, 1814, 942-68.
- KO, L. J. & ENGEL, J. D. 1993. DNA-binding specificities of the GATA transcription factor family. *Mol Cell Biol*, 13, 4011-22.
- KONKEL, D. A., MAIZEL, J. V., JR. & LEDER, P. 1979. The evolution and sequence comparison of two recently diverged mouse chromosomal beta-globin genes. *Cell*, 18, 865-73.
- KOUTSOURAKIS, M., LANGEVELD, A., PATIENT, R., BEDDINGTON, R. & GROSVELD, F. 1999. The transcription factor GATA6 is essential for early extraembryonic development. *Development*, 126, 723-32.
- KOWALSKI, K., CZOLIJ, R., KING, G. F., CROSSLEY, M. & MACKAY, J. P. 1999. The solution structure of the N-terminal zinc finger of GATA-1 reveals a specific binding face for the transcriptional co-factor FOG. *J Biomol NMR*, 13, 249-62.
- KRISHNA, S. S., MAJUMDAR, I. & GRISHIN, N. V. 2003. Structural classification of zinc fingers: survey and summary. *Nucleic Acids Res*, 31, 532-50.
- KRUM, S. A., MIRANDA-CARBONI, G. A., LUPIEN, M., EECKHOUTE, J., CARROLL, J. S. & BROWN, M. 2008. Unique ERalpha cistromes control cell type-specific gene regulation. *Mol Endocrinol*, 22, 2393-406.
- KUO, Y. H., GERSTEIN, R. M. & CASTILLA, L. H. 2008. Cbfbeta-SMMHC impairs differentiation of common lymphoid progenitors and reveals an essential role for RUNX in early B-cell development. *Blood*, 111, 1543-51.
- KURAS, L., BORGGREFE, T. & KORNBERG, R. D. 2003. Association of the Mediator complex with enhancers of active genes. *Proc Natl Acad Sci U S A*, 100, 13887-91.
- LAHLIL, R., LECUYER, E., HERBLOT, S. & HOANG, T. 2004. SCL assembles a multifactorial complex that determines glycophorin A expression. *Mol Cell Biol*, 24, 1439-52.
- LAIOSA, C. V., STADTFELD, M. & GRAF, T. 2006. Determinants of lymphoid-myeloid lineage diversification. *Annu Rev Immunol*, 24, 705-38.
- LAIRD, D. J., VON ANDRIAN, U. H. & WAGERS, A. J. 2008. Stem cell trafficking in tissue development, growth, and disease. *Cell*, 132, 612-30.
- LAM, F. H., STEGER, D. J. & O'SHEA, E. K. 2008. Chromatin decouples promoter threshold from dynamic range. *Nature*, 453, 246-50.
- LAM, K. N., VAN BAKEL, H., COTE, A. G., VAN DER VEN, A. & HUGHES, T. R. 2011. Sequence specificity is obtained from the majority of modular C2H2 zinc-finger arrays. *Nucleic Acids Res*, 39, 4680-90.
- LAMONICA, J. M., VAKOC, C. R. & BLOBEL, G. A. 2006. Acetylation of GATA-1 is required for chromatin occupancy. *Blood*, 108, 3736-8.
- LANGMEAD, B., TRAPNELL, C., POP, M. & SALZBERG, S. L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10, R25.
- LARSON, D. R., SINGER, R. H. & ZENKLUSEN, D. 2009. A single molecule view of gene expression. *Trends Cell Biol*, 19, 630-7.

- LARSON, D. R., ZENKLUSEN, D., WU, B., CHAO, J. A. & SINGER, R. H. 2011. Real-time observation of transcription initiation and elongation on an endogenous yeast gene. *Science*, 332, 475-8.
- LEBRECHT, D., FOEHR, M., SMITH, E., LOPES, F. J., VANARIO-ALONSO, C. E., REINITZ, J., BURZ, D. S. & HANES, S. D. 2005. Bicoid cooperative DNA binding is critical for embryonic patterning in *Drosophila*. *Proc Natl Acad Sci U S A*, 102, 13176-81.
- LEE, C. K., SHIBATA, Y., RAO, B., STRAHL, B. D. & LIEB, J. D. 2004. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat Genet*, 36, 900-5.
- LEE, H. Y., JOHNSON, K. D., FUJIWARA, T., BOYER, M. E., KIM, S. I. & BRESNICK, E. H. 2009a. Controlling hematopoiesis through sumoylation-dependent regulation of a GATA factor. *Mol Cell*, 36, 984-95.
- LEE, J., KIM, D. H., LEE, S., YANG, Q. H., LEE, D. K., LEE, S. K., ROEDER, R. G. & LEE, J. W. 2009b. A tumor suppressive coactivator complex of p53 containing ASC-2 and histone H3-lysine-4 methyltransferase MLL3 or its paralogue MLL4. *Proc Natl Acad Sci U S A*, 106, 8513-8.
- LEE, J. H. & SKALNIK, D. G. 2005. CpG-binding protein (CXXC finger protein 1) is a component of the mammalian Set1 histone H3-Lys4 methyltransferase complex, the analogue of the yeast Set1/COMPASS complex. *J Biol Chem*, 280, 41725-31.
- LEE, J. H., TATE, C. M., YOU, J. S. & SKALNIK, D. G. 2007a. Identification and characterization of the human Set1B histone H3-Lys4 methyltransferase complex. *J Biol Chem*, 282, 13419-28.
- LEE, J. S., SHUKLA, A., SCHNEIDER, J., SWANSON, S. K., WASHBURN, M. P., FLORENS, L., BHAUMIK, S. R. & SHILATIFARD, A. 2007b. Histone crosstalk between H2B monoubiquitination and H3 methylation mediated by COMPASS. *Cell*, 131, 1084-96.
- LEE, N., ZHANG, J., KLOSE, R. J., ERDJUMENT-BROMAGE, H., TEMPST, P., JONES, R. S. & ZHANG, Y. 2007c. The trithorax-group protein Lid is a histone H3 trimethyl-Lys4 demethylase. *Nat Struct Mol Biol*, 14, 341-3.
- LEE, S., KIM, D. H., GOO, Y. H., LEE, Y. C., LEE, S. K. & LEE, J. W. 2009c. Crucial roles for interactions between MLL3/4 and INI1 in nuclear receptor transactivation. *Mol Endocrinol*, 23, 610-9.
- LENHARD, B., SANDELIN, A. & CARNINCI, P. 2012. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet*, 13, 233-45.
- LEONARD, M., BRICE, M., ENGEL, J. D. & PAPAYANNOPOULOU, T. 1993. Dynamics of GATA transcription factor expression during erythroid differentiation. *Blood*, 82, 1071-9.
- LETTING, D. L., CHEN, Y. Y., RAKOWSKI, C., REEDY, S. & BLOBEL, G. A. 2004. Context-dependent regulation of GATA-1 by friend of GATA-1. *Proc Natl Acad Sci U S A*, 101, 476-81.
- LETTING, D. L., RAKOWSKI, C., WEISS, M. J. & BLOBEL, G. A. 2003. Formation of a tissue-specific histone acetylation pattern by the hematopoietic transcription factor GATA-1. *Mol Cell Biol*, 23, 1334-40.
- LI, B., JACKSON, J., SIMON, M. D., FLEHARTY, B., GOGOL, M., SEIDEL, C., WORKMAN, J. L. & SHILATIFARD, A. 2009. Histone H3 lysine 36 dimethylation (H3K36me2) is sufficient to recruit the Rpd3s histone deacetylase complex and to repress spurious transcription. *J Biol Chem*, 284, 7970-6.
- LI, H., ILIN, S., WANG, W., DUNCAN, E. M., WYSOCKA, J., ALLIS, C. D. & PATEL, D. J. 2006. Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. *Nature*, 442, 91-5.
- LI, X. Y., THOMAS, S., SABO, P. J., EISEN, M. B., STAMATOYANNOPOULOS, J. A. & BIGGIN, M. D. 2011. The role of chromatin accessibility in directing the

- widespread, overlapping patterns of *Drosophila* transcription factor binding. *Genome Biol*, 12, R34.
- LICATA, V. J. & WOWOR, A. J. 2008. Applications of fluorescence anisotropy to the study of protein-DNA interactions. *Methods Cell Biol*, 84, 243-62.
- LICKWAR, C. R., MUELLER, F., HANLON, S. E., MCNALLY, J. G. & LIEB, J. D. 2012. Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. *Nature*, 484, 251-5.
- LIEW, C. K., SIMPSON, R. J., KWAN, A. H., CROFTS, L. A., LOUGHLIN, F. E., MATTHEWS, J. M., CROSSLEY, M. & MACKAY, J. P. 2005. Zinc fingers as protein recognition motifs: structural basis for the GATA-1/friend of GATA interaction. *Proc Natl Acad Sci U S A*, 102, 583-8.
- LING, K. W., OTTERSBAACH, K., VAN HAMBURG, J. P., OZIEMLA, A., TSAI, F. Y., ORKIN, S. H., PLOEMACHER, R., HENDRIKS, R. W. & DZIERZAK, E. 2004. GATA-2 plays two functionally distinct roles during the ontogeny of hematopoietic stem cells. *J Exp Med*, 200, 871-82.
- LIONNET, T. & SINGER, R. H. 2012. Transcription goes digital. *EMBO Rep*, 13, 313-21.
- LIPFORD, J. R., SMITH, G. T., CHI, Y. & DESHAIES, R. J. 2005. A putative stimulatory role for activator turnover in gene expression. *Nature*, 438, 113-6.
- LIU, X., LEE, C. K., GRANER, J. A., CLARKE, N. D. & LIEB, J. D. 2006. Whole-genome comparison of Leu3 binding in vitro and in vivo reveals the importance of nucleosome occupancy in target site selection. *Genome Res*, 16, 1517-28.
- LIU, X., NOLL, D. M., LIEB, J. D. & CLARKE, N. D. 2005. DIP-chip: rapid and accurate determination of DNA-binding specificity. *Genome Res*, 15, 421-7.
- LLORET-LLINARES, M., CARRE, C., VAQUERO, A., DE OLANO, N. & AZORIN, F. 2008. Characterization of *Drosophila melanogaster* JmjC+N histone demethylases. *Nucleic Acids Res*, 36, 2852-63.
- LLORET-LLINARES, M., PEREZ-LLUCH, S., ROSSELL, D., MORAN, T., PONSACOBAS, J., AUER, H., COROMINAS, M. & AZORIN, F. 2012. dKDM5/LID regulates H3K4me3 dynamics at the transcription-start site (TSS) of actively transcribed developmental genes. *Nucleic Acids Res*, 40, 9493-505.
- LU, S. J., ROWAN, S., BANI, M. R. & BEN-DAVID, Y. 1994. Retroviral integration within the Fli-2 locus results in inactivation of the erythroid transcription factor NF-E2 in Friend erythroleukemias: evidence that NF-E2 is essential for globin expression. *Proc Natl Acad Sci U S A*, 91, 8398-402.
- LURIE, L. J., BOYER, M. E., GRASS, J. A. & BRESNICK, E. H. 2008. Differential GATA factor stabilities: implications for chromatin occupancy by structurally similar transcription factors. *Biochemistry*, 47, 859-69.
- MACKAY, J. P., KOWALSKI, K., FOX, A. H., CZOLIJ, R., KING, G. F. & CROSSLEY, M. 1998. Involvement of the N-finger in the self-association of GATA-1. *J Biol Chem*, 273, 30560-7.
- MALI, P., YANG, L., ESVELT, K. M., AACH, J., GUELL, M., DICARLO, J. E., NORVILLE, J. E. & CHURCH, G. M. 2013. RNA-guided human genome engineering via Cas9. *Science*, 339, 823-6.
- MARTIN, D. I. & ORKIN, S. H. 1990. Transcriptional activation and DNA binding by the erythroid factor GF-1/NF-E1/Eryf 1. *Genes Dev*, 4, 1886-98.
- MARTIN, D. I., ZON, L. I., MUTTER, G. & ORKIN, S. H. 1990. Expression of an erythroid transcription factor in megakaryocytic and mast cell lineages. *Nature*, 344, 444-7.
- MAS, C., LUSSIER-PRICE, M., SONI, S., MORSE, T., ARSENEAULT, G., DI LELLO, P., LAFRANCE-VANASSE, J., BIEKER, J. J. & OMICHINSKI, J. G. 2011. Structural and functional characterization of an atypical activation domain in erythroid Kruppel-like factor (EKLF). *Proc Natl Acad Sci U S A*, 108, 10484-9.
- MASTON, G. A., EVANS, S. K. & GREEN, M. R. 2006. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet*, 7, 29-59.
- MAY, G., SONEJI, S., TIPPING, A. J., TELES, J., MCGOWAN, S. J., WU, M., GUO, Y., FUGAZZA, C., BROWN, J., KARLSSON, G., PINA, C., OLARIU, V., TAYLOR, S.,

- TENEN, D. G., PETERSON, C. & ENVER, T. 2013. Dynamic analysis of gene expression and genome-wide transcription factor binding during lineage specification of multipotent progenitors. *Cell Stem Cell*, 13, 754-68.
- MCNALLY, J. G., MULLER, W. G., WALKER, D., WOLFORD, R. & HAGER, G. L. 2000. The glucocorticoid receptor: rapid exchange with regulatory sites in living cells. *Science*, 287, 1262-5.
- MELNIKOV, A., MURUGAN, A., ZHANG, X., TESILEANU, T., WANG, L., ROGOV, P., FEIZI, S., GNIRKE, A., CALLAN, C. G., JR., KINNEY, J. B., KELLIS, M., LANDER, E. S. & MIKKELSEN, T. S. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol*, 30, 271-7.
- MERIKA, M. & ORKIN, S. H. 1993. DNA-binding specificity of GATA family transcription factors. *Mol Cell Biol*, 13, 3999-4010.
- MERIKA, M. & ORKIN, S. H. 1995. Functional synergy and physical interactions of the erythroid transcription factor GATA-1 with the Kruppel family proteins Sp1 and EKLF. *Mol Cell Biol*, 15, 2437-47.
- MERIKA, M. & THANOS, D. 2001. Enhanceosomes. *Curr Opin Genet Dev*, 11, 205-8.
- MERIKA, M., WILLIAMS, A. J., CHEN, G., COLLINS, T. & THANOS, D. 1998. Recruitment of CBP/p300 by the IFN beta enhanceosome is required for synergistic activation of transcription. *Mol Cell*, 1, 277-87.
- METIVIER, R., PENOT, G., HUBNER, M. R., REID, G., BRAND, H., KOS, M. & GANNON, F. 2003. Estrogen receptor-alpha directs ordered, cyclical, and combinatorial recruitment of cofactors on a natural target promoter. *Cell*, 115, 751-63.
- MICCIO, A., WANG, Y., HONG, W., GREGORY, G. D., WANG, H., YU, X., CHOI, J. K., SHELAT, S., TONG, W., PONCZ, M. & BLOBEL, G. A. 2010. NuRD mediates activating and repressive functions of GATA-1 and FOG-1 during blood development. *EMBO J*, 29, 442-56.
- MIGLIACCIO, G., DI PIETRO, R., DI GIACOMO, V., DI BALDASSARRE, A., MIGLIACCIO, A. R., MACCIONI, L., GALANELLO, R. & PAPAYANNOPOULOU, T. 2002. In vitro mass production of human erythroid cells from the blood of normal donors and of thalassemic patients. *Blood Cells Mol Dis*, 28, 169-80.
- MIGNOTTE, V., WALL, L., DEBOER, E., GROSVELD, F. & ROMEO, P. H. 1989. Two tissue-specific factors bind the erythroid promoter of the human porphobilinogen deaminase gene. *Nucleic Acids Res*, 17, 37-54.
- MIKKOLA, H. K., KLINTMAN, J., YANG, H., HOCK, H., SCHLAEGER, T. M., FUJIWARA, Y. & ORKIN, S. H. 2003. Haematopoietic stem cells retain long-term repopulating activity and multipotency in the absence of stem-cell leukaemia SCL/tal-1 gene. *Nature*, 421, 547-51.
- MILLER, I. J. & BIEKER, J. J. 1993. A novel, erythroid cell-specific murine transcription factor that binds to the CACCC element and is related to the Kruppel family of nuclear proteins. *Mol Cell Biol*, 13, 2776-86.
- MILNE, T. A., KIM, J., WANG, G. G., STADLER, S. C., BASRUR, V., WHITCOMB, S. J., WANG, Z., RUTHENBURG, A. J., ELENITOBA-JOHNSON, K. S., ROEDER, R. G. & ALLIS, C. D. 2010. Multiple interactions recruit MLL1 and MLL1 fusion proteins to the HOXA9 locus in leukemogenesis. *Mol Cell*, 38, 853-63.
- MINEGISHI, N., OHTA, J., YAMAGIWA, H., SUZUKI, N., KAWAUCHI, S., ZHOU, Y., TAKAHASHI, S., HAYASHI, N., ENGEL, J. D. & YAMAMOTO, M. 1999. The mouse GATA-2 gene is expressed in the para-aortic splanchnopleura and aorta-gonads and mesonephros region. *Blood*, 93, 4196-207.
- MIRANDA-CARBONI, G. A., GUEMES, M., BAILEY, S., ANAYA, E., CORSELLI, M., PEAULT, B. & KRUM, S. A. 2011. GATA4 regulates estrogen receptor-alpha-mediated osteoblast transcription. *Mol Endocrinol*, 25, 1126-36.
- MITO, Y., HENIKOFF, J. G. & HENIKOFF, S. 2007. Histone replacement marks the boundaries of cis-regulatory domains. *Science*, 315, 1408-11.

- MIYAGAWA, S., KOBAYASHI, M., KONISHI, N., SATO, T. & UEDA, K. 2000. Insulin and insulin-like growth factor I support the proliferation of erythroid progenitor cells in bone marrow through the sharing of receptors. *Br J Haematol*, 109, 555-62.
- MO, R., RAO, S. M. & ZHU, Y. J. 2006. Identification of the MLL2 complex as a coactivator for estrogen receptor alpha. *J Biol Chem*, 281, 15714-20.
- MOLKENTIN, J. D. 2000. The zinc finger-containing transcription factors GATA-4, -5, and -6. Ubiquitously expressed regulators of tissue-specific gene expression. *J Biol Chem*, 275, 38949-52.
- MOLKENTIN, J. D., TYMITZ, K. M., RICHARDSON, J. A. & OLSON, E. N. 2000. Abnormalities of the genitourinary tract in female mice lacking GATA5. *Mol Cell Biol*, 20, 5256-60.
- MOON, A. M. & LEY, T. J. 1990. Conservation of the primary structure, organization, and function of the human and mouse beta-globin locus-activating regions. *Proc Natl Acad Sci U S A*, 87, 7693-7.
- MORRIS, S. A., BAEK, S., SUNG, M. H., JOHN, S., WIENCH, M., JOHNSON, T. A., SCHILTZ, R. L. & HAGER, G. L. 2014. Overlapping chromatin-remodeling systems collaborate genome wide at dynamic chromatin transitions. *Nat Struct Mol Biol*, 21, 73-81.
- MOTT, B. H., BASSMAN, J. & PIKAART, M. J. 2004. A molecular dissection of the interaction between the transcription factor Gata-1 zinc finger and DNA. *Biochem Biophys Res Commun*, 316, 910-7.
- MUKHERJEE, S., BERGER, M. F., JONA, G., WANG, X. S., MUZZEY, D., SNYDER, M., YOUNG, R. A. & BULYK, M. L. 2004. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet*, 36, 1331-9.
- MUNTEAN, A. G., TAN, J., SITWALA, K., HUANG, Y., BRONSTEIN, J., CONNELLY, J. A., BASRUR, V., ELENITOBA-JOHNSON, K. S. & HESS, J. L. 2010. The PAF complex synergizes with MLL fusion proteins at HOX loci to promote leukemogenesis. *Cancer Cell*, 17, 609-21.
- MURAMOTO, T., MULLER, I., THOMAS, G., MELVIN, A. & CHUBB, J. R. 2010. Methylation of H3K4 is required for inheritance of active transcriptional states. *Curr Biol*, 20, 397-406.
- MURATANI, M., KUNG, C., SHOKAT, K. M. & TANSEY, W. P. 2005. The F box protein Dsg1/Mdm30 is a transcriptional coactivator that stimulates Gal4 turnover and cotranscriptional mRNA processing. *Cell*, 120, 887-99.
- NALLEY, K., JOHNSTON, S. A. & KODADEK, T. 2006. Proteolytic turnover of the Gal4 transcription factor is not required for function in vivo. *Nature*, 442, 1054-7.
- NARDELLI, J., THIESSON, D., FUJIWARA, Y., TSAI, F. Y. & ORKIN, S. H. 1999. Expression and genetic interaction of transcription factors GATA-2 and GATA-3 during development of the mouse central nervous system. *Dev Biol*, 210, 305-21.
- NERLOV, C. & GRAF, T. 1998. PU.1 induces myeloid lineage commitment in multipotent hematopoietic progenitors. *Genes Dev*, 12, 2403-12.
- NERLOV, C., QUERFURTH, E., KULESSA, H. & GRAF, T. 2000. GATA-1 interacts with the myeloid PU.1 transcription factor and represses PU.1-dependent transcription. *Blood*, 95, 2543-51.
- NEWBURGER, D. E. & BULYK, M. L. 2009. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res*, 37, D77-82.
- NEWTON, A., MACKAY, J. & CROSSLEY, M. 2001. The N-terminal zinc finger of the erythroid transcription factor GATA-1 binds GATC motifs in DNA. *J Biol Chem*, 276, 35794-801.
- NEY, P. A. & D'ANDREA, A. D. 2000. Friend erythroleukemia revisited. *Blood*, 96, 3675-80.
- NG, H. H., ROBERT, F., YOUNG, R. A. & STRUHL, K. 2003. Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. *Mol Cell*, 11, 709-19.

- NISHIKAWA, K., KOBAYASHI, M., MASUMI, A., LYONS, S. E., WEINSTEIN, B. M., LIU, P. P. & YAMAMOTO, M. 2003. Self-association of Gata1 enhances transcriptional activity in vivo in zebra fish embryos. *Mol Cell Biol*, 23, 8295-305.
- NOORDERMEER, D., DE WIT, E., KLOUS, P., VAN DE WERKEN, H., SIMONIS, M., LOPEZ-JONES, M., EUSSEN, B., DE KLEIN, A., SINGER, R. H. & DE LAAT, W. 2011. Variegated gene expression caused by cell-specific long-range DNA interactions. *Nat Cell Biol*, 13, 944-51.
- NUEZ, B., MICHALOVICH, D., BYGRAVE, A., PLOEMACHER, R. & GROSVELD, F. 1995. Defective haematopoiesis in fetal liver resulting from inactivation of the EKLF gene. *Nature*, 375, 316-8.
- NUTIU, R., FRIEDMAN, R. C., LUO, S., KHREBTUKOVA, I., SILVA, D., LI, R., ZHANG, L., SCHROTH, G. P. & BURGE, C. B. 2011. Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nat Biotechnol*, 29, 659-64.
- NUTT, S. L. & KEE, B. L. 2007. The transcriptional regulation of B cell lineage commitment. *Immunity*, 26, 715-25.
- OKA, S., SHIRAISHI, Y., YOSHIDA, T., OHKUBO, T., SUGIURA, Y. & KOBAYASHI, Y. 2004. NMR structure of transcription factor Sp1 DNA binding domain. *Biochemistry*, 43, 16027-35.
- OKUDA, T., VAN DEURSEN, J., HIEBERT, S. W., GROSVELD, G. & DOWNING, J. R. 1996. AML1, the target of multiple chromosomal translocations in human leukemia, is essential for normal fetal liver hematopoiesis. *Cell*, 84, 321-30.
- OMICHTINSKI, J. G., CLORE, G. M., SCHAAD, O., FELSENFELD, G., TRAINOR, C., APPELLA, E., STAHL, S. J. & GRONENBORN, A. M. 1993. NMR structure of a specific DNA complex of Zn-containing DNA binding domain of GATA-1. *Science*, 261, 438-46.
- ONG, C. T. & CORCES, V. G. 2011. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat Rev Genet*, 12, 283-93.
- ORFORD, K., KHARCHENKO, P., LAI, W., DAO, M. C., WORHUNSKY, D. J., FERRO, A., JANZEN, V., PARK, P. J. & SCADDEN, D. T. 2008. Differential H3K4 methylation identifies developmentally poised hematopoietic genes. *Dev Cell*, 14, 798-809.
- ORKIN, S. H. & ZON, L. I. 2008. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell*, 132, 631-44.
- OSADA, H., GRUTZ, G., AXELSON, H., FORSTER, A. & RABBITS, T. H. 1995. Association of erythroid transcription factors: complexes involving the LIM protein RBTN2 and the zinc-finger protein GATA1. *Proc Natl Acad Sci U S A*, 92, 9585-9.
- PAL, S., CANTOR, A. B., JOHNSON, K. D., MORAN, T. B., BOYER, M. E., ORKIN, S. H. & BRESNICK, E. H. 2004a. Coregulator-dependent facilitation of chromatin occupancy by GATA-1. *Proc Natl Acad Sci U S A*, 101, 980-5.
- PAL, S., NEMETH, M. J., BODINE, D., MILLER, J. L., SVAREN, J., THEIN, S. L., LOWRY, P. J. & BRESNICK, E. H. 2004b. Neurokinin-B transcription in erythroid cells: direct activation by the hematopoietic transcription factor GATA-1. *J Biol Chem*, 279, 31348-56.
- PALSTRA, R. J., TOLHUIS, B., SPLINTER, E., NIJMEIJER, R., GROSVELD, F. & DE LAAT, W. 2003. The beta-globin nuclear compartment in development and erythroid differentiation. *Nat Genet*, 35, 190-4.
- PANG, L., XUE, H. H., SZALAI, G., WANG, X., WANG, Y., WATSON, D. K., LEONARD, W. J., BLOBEL, G. A. & PONCZ, M. 2006. Maturation stage-specific regulation of megakaryopoiesis by pointed-domain Ets proteins. *Blood*, 108, 2198-206.
- PAPADOPOULOS, G. L., KARKOULIA, E., TSAMARDINOS, I., PORCHER, C., RAGOISSIS, J., BUNGERT, J. & STROUBOULIS, J. 2013. GATA-1 genome-wide occupancy associates with distinct epigenetic profiles in mouse fetal liver erythropoiesis. *Nucleic Acids Res*, 41, 4938-48.

- PARK, P. J. 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, 10, 669-80.
- PARTINGTON, G. A. & PATIENT, R. K. 1999. Phosphorylation of GATA-1 increases its DNA-binding affinity and is correlated with induction of human K562 erythroleukaemia cells. *Nucleic Acids Res*, 27, 1168-75.
- PATWARDHAN, R. P., HIATT, J. B., WITTEN, D. M., KIM, M. J., SMITH, R. P., MAY, D., LEE, C., ANDRIE, J. M., LEE, S. I., COOPER, G. M., AHITUV, N., PENNACCHIO, L. A. & SHENDURE, J. 2012. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol*, 30, 265-70.
- PAUL, S., VADGAMA, P. & RAY, A. K. 2009. Surface plasmon resonance imaging for biosensing. *IET Nanobiotechnol*, 3, 71-80.
- PEDONE, P. V., OMICHINSKI, J. G., NONY, P., TRAINOR, C., GRONENBORN, A. M., CLORE, G. M. & FELSENFELD, G. 1997. The N-terminal fingers of chicken GATA-2 and GATA-3 are independent sequence-specific DNA binding domains. *EMBO J*, 16, 2874-82.
- PERKINS, A. C., SHARPE, A. H. & ORKIN, S. H. 1995. Lethal beta-thalassaemia in mice lacking the erythroid CACCC-transcription factor EKLF. *Nature*, 375, 318-22.
- PERKINS, N. D., NICOLAS, R. H., PLUMB, M. A. & GOODWIN, G. H. 1989. The purification of an erythroid protein which binds to enhancer and promoter elements of haemoglobin genes. *Nucleic Acids Res*, 17, 1299-314.
- PERRY, J. M., HARANDI, O. F. & PAULSON, R. F. 2007. BMP4, SCF, and hypoxia cooperatively regulate the expansion of murine stress erythroid progenitors. *Blood*, 109, 4494-502.
- PETRYKOWSKA, H. M., VOCKLEY, C. M. & ELNITSKI, L. 2008. Detection and characterization of silencers and enhancer-blockers in the greater CFTR locus. *Genome Res*, 18, 1238-46.
- PEVNY, L., LIN, C. S., D'AGATI, V., SIMON, M. C., ORKIN, S. H. & COSTANTINI, F. 1995. Development of hematopoietic cells lacking transcription factor GATA-1. *Development*, 121, 163-72.
- PHILLIPS, S. J., HARDIES, S. C., JAHN, C. L., EDGELL, M. H. & HUTCHISON, C. A., 3RD 1984. The complete nucleotide sequence of a beta-globin-like structure, beta h2, from the [Hbb]d mouse BALB/c. *J Biol Chem*, 259, 7947-54.
- PILON, A. M., AJAY, S. S., KUMAR, S. A., STEINER, L. A., CHERUKURI, P. F., WINCOVITCH, S., ANDERSON, S. M., CENTER, N. C. S., MULLIKIN, J. C., GALLAGHER, P. G., HARDISON, R. C., MARGULIES, E. H. & BODINE, D. M. 2011. Genome-wide ChIP-Seq reveals a dramatic shift in the binding of the transcription factor erythroid Kruppel-like factor during erythrocyte differentiation. *Blood*, 118, e139-48.
- POPE, N. J. & BRESNICK, E. H. 2010. Differential coregulator requirements for function of the hematopoietic transcription factor GATA-1 at endogenous loci. *Nucleic Acids Res*, 38, 2190-200.
- POPE, N. J. & BRESNICK, E. H. 2013. Establishment of a cell-type-specific genetic network by the mediator complex component Med1. *Mol Cell Biol*, 33, 1938-55.
- PORCHER, C., SWAT, W., ROCKWELL, K., FUJIWARA, Y., ALT, F. W. & ORKIN, S. H. 1996. The T cell leukemia oncoprotein SCL/tal-1 is essential for development of all hematopoietic lineages. *Cell*, 86, 47-57.
- QUERFURTH, E., SCHUSTER, M., KULESSA, H., CRISPINO, J. D., DODERLEIN, G., ORKIN, S. H., GRAF, T. & NERLOV, C. 2000. Antagonism between C/EBPbeta and FOG in eosinophil lineage commitment of multipotent hematopoietic progenitors. *Genes Dev*, 14, 2515-25.
- RADA-IGLESIAS, A., BAJPAI, R., SWIGUT, T., BRUGMANN, S. A., FLYNN, R. A. & WYSOCKA, J. 2011. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, 470, 279-83.

- RAMIREZ, J. & HAGMAN, J. 2009. The Mi-2/NuRD complex: a critical epigenetic regulator of hematopoietic development, differentiation and cancer. *Epigenetics*, 4, 532-6.
- RAMOS, Y. F., HESTAND, M. S., VERLAAN, M., KRABBENDAM, E., ARIYUREK, Y., VAN GALEN, M., VAN DAM, H., VAN OMMEN, G. J., DEN DUNNEN, J. T., ZANTEMA, A. & T HOEN, P. A. 2010. Genome-wide assessment of differential roles for p300 and CBP in transcription regulation. *Nucleic Acids Res*, 38, 5396-408.
- REID, G., HUBNER, M. R., METIVIER, R., BRAND, H., DENGGER, S., MANU, D., BEAUDOUIN, J., ELLENBERG, J. & GANNON, F. 2003. Cyclic, proteasome-mediated turnover of unliganded and liganded ERalpha on responsive promoters is an integral feature of estrogen signaling. *Mol Cell*, 11, 695-707.
- REKHTMAN, N., RADPARVAR, F., EVANS, T. & SKOULTCHI, A. I. 1999. Direct interaction of hematopoietic transcription factors PU.1 and GATA-1: functional antagonism in erythroid cells. *Genes Dev*, 13, 1398-411.
- REN, B., ROBERT, F., WYRICK, J. J., APARICIO, O., JENNINGS, E. G., SIMON, I., ZEITLINGER, J., SCHREIBER, J., HANNETT, N., KANIN, E., VOLKERT, T. L., WILSON, C. J., BELL, S. P. & YOUNG, R. A. 2000. Genome-wide location and function of DNA binding proteins. *Science*, 290, 2306-9.
- RHEE, H. S. & PUGH, B. F. 2011. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, 147, 1408-19.
- RHODES, J., HAGEN, A., HSU, K., DENG, M., LIU, T. X., LOOK, A. T. & KANKI, J. P. 2005. Interplay of pu.1 and gata1 determines myelo-erythroid progenitor cell fate in zebrafish. *Dev Cell*, 8, 97-108.
- RODRIGUEZ, P., BONTE, E., KRIJGSVELD, J., KOLODZIEJ, K. E., GUYOT, B., HECK, A. J., VYAS, P., DE BOER, E., GROSVELD, F. & STROUBOULIS, J. 2005. GATA-1 forms distinct activating and repressive complexes in erythroid cells. *EMBO J*, 24, 2354-66.
- ROMEO, P. H., PRANDINI, M. H., JOULIN, V., MIGNOTTE, V., PRENANT, M., VAINCHENKER, W., MARGUERIE, G. & UZAN, G. 1990. Megakaryocytic and erythrocytic lineages share specific transcription factors. *Nature*, 344, 447-9.
- ROULET, E., BUSSO, S., CAMARGO, A. A., SIMPSON, A. J., MERMOD, N. & BUCHER, P. 2002. High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat Biotechnol*, 20, 831-5.
- SANDER, J. D. & JOUNG, J. K. 2014. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat Biotechnol*, 32, 347-55.
- SANTOS-ROSA, H., SCHNEIDER, R., BERNSTEIN, B. E., KARABETSOU, N., MORILLON, A., WEISE, C., SCHREIBER, S. L., MELLOR, J. & KOUZARIDES, T. 2003. Methylation of histone H3 K4 mediates association of the Isw1p ATPase with chromatin. *Mol Cell*, 12, 1325-32.
- SAWADO, T., IGARASHI, K. & GROUDINE, M. 2001. Activation of beta-major globin gene transcription is associated with recruitment of NF-E2 to the beta-globin LCR and gene promoter. *Proc Natl Acad Sci U S A*, 98, 10226-31.
- SCHER, W., PREISLER, H. D. & FRIEND, C. 1973. Hemoglobin synthesis in murine virus-induced leukemic cells in vitro. 3. Effects of 5-bromo-2'-deoxyuridine, dimethylformamide and dimethylsulfoxide. *J Cell Physiol*, 81, 63-70.
- SCHMIDT, D., WILSON, M. D., BALLESTER, B., SCHWALIE, P. C., BROWN, G. D., MARSHALL, A., KUTTER, C., WATT, S., MARTINEZ-JIMENEZ, C. P., MACKAY, S., TALIANIDIS, I., FLICEK, P. & ODOM, D. T. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, 328, 1036-40.
- SCHOENFELDER, S., SEXTON, T., CHAKALOVA, L., COPE, N. F., HORTON, A., ANDREWS, S., KURUKUTI, S., MITCHELL, J. A., UMLAUF, D., DIMITROVA, D. S., ESKIWI, C. H., LUO, Y., WEI, C. L., RUAN, Y., BIEKER, J. J. & FRASER, P.

2010. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat Genet*, 42, 53-61.
- SCHUETZ, A., NANA, D., ROSE, C., ZOCHER, G., MILANOVIC, M., KOENIGSMANN, J., BLASIG, R., HEINEMANN, U. & CARSTANJEN, D. 2011. The structure of the Klf4 DNA-binding domain links to self-renewal and macrophage differentiation. *Cell Mol Life Sci*, 68, 3121-31.
- SECOMBE, J. & EISENMAN, R. N. 2007. The function and regulation of the JARID1 family of histone H3 lysine 4 demethylases: the Myc connection. *Cell Cycle*, 6, 1324-8.
- SHARON, E., KALMA, Y., SHARP, A., RAVEH-SADKA, T., LEVO, M., ZEEVI, D., KEREN, L., YAKHINI, Z., WEINBERGER, A. & SEGAL, E. 2012. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol*, 30, 521-30.
- SHARON, E., LUBLINER, S. & SEGAL, E. 2008. A feature-based approach to modeling protein-DNA interactions. *PLoS Comput Biol*, 4, e1000154.
- SHEHEE, W. R., LOEB, D. D., ADEY, N. B., BURTON, F. H., CASAVANT, N. C., COLE, P., DAVIES, C. J., MCGRAW, R. A., SCHICHMAN, S. A., SEVERYNSE, D. M. & ET AL. 1989. Nucleotide sequence of the BALB/c mouse beta-globin complex. *J Mol Biol*, 205, 41-62.
- SHIMIZU, R., TAKAHASHI, S., OHNEDA, K., ENGEL, J. D. & YAMAMOTO, M. 2001. In vivo requirements for GATA-1 functional domains during primitive and definitive erythropoiesis. *EMBO J*, 20, 5250-60.
- SHUMAKER-PARRY, J. S., AEBERSOLD, R. & CAMPBELL, C. T. 2004. Parallel, quantitative measurement of protein binding to a 120-element double-stranded DNA array in real time using surface plasmon resonance microscopy. *Anal Chem*, 76, 2071-82.
- SIATECKA, M., XUE, L. & BIEKER, J. J. 2007. Sumoylation of EKLF promotes transcriptional repression and is involved in inhibition of megakaryopoiesis. *Mol Cell Biol*, 27, 8547-60.
- SIERRA, J., YOSHIDA, T., JOAZEIRO, C. A. & JONES, K. A. 2006. The APC tumor suppressor counteracts beta-catenin activation and H3K4 methylation at Wnt target genes. *Genes Dev*, 20, 586-600.
- SLATTERY, M., RILEY, T., LIU, P., ABE, N., GOMEZ-ALCALA, P., DROR, I., ZHOU, T., ROHS, R., HONIG, B., BUSSEMAKER, H. J. & MANN, R. S. 2011. Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell*, 147, 1270-82.
- SNOW, J. W. & ORKIN, S. H. 2009. Translational isoforms of FOG1 regulate GATA1-interacting complexes. *J Biol Chem*, 284, 29310-9.
- SOARES, L. M. & BURATOWSKI, S. 2012. Yeast Swd2 is essential because of antagonism between Set1 histone methyltransferase complex and APT (associated with Pta1) termination factor. *J Biol Chem*, 287, 15219-31.
- SOCOLOVSKY, M., NAM, H., FLEMING, M. D., HAASE, V. H., BRUGNARA, C. & LODISH, H. F. 2001. Ineffective erythropoiesis in Stat5a(-/-)5b(-/-) mice due to decreased survival of early erythroblasts. *Blood*, 98, 3261-73.
- SOLER, E., ANDRIEU-SOLER, C., DE BOER, E., BRYNE, J. C., THONGJUEA, S., STADHOUDERS, R., PALSTRA, R. J., STEVENS, M., KOCKX, C., VAN IJCKEN, W., HOU, J., STEINHOFF, C., RIJKERS, E., LENHARD, B. & GROSVELD, F. 2010. The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. *Genes Dev*, 24, 277-89.
- SPITZ, F. & FURLONG, E. E. 2012. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet*, 13, 613-26.
- STARCK, J., COHET, N., GONNET, C., SARRAZIN, S., DOUBEIKOVSKAIA, Z., DOUBEIKOVSKI, A., VERGER, A., DUTERQUE-COQUILLAUD, M. & MORLE, F. 2003. Functional cross-antagonism between transcription factors FLI-1 and EKLF. *Mol Cell Biol*, 23, 1390-402.

- STARICH, M. R., WIKSTROM, M., ARST, H. N., JR., CLORE, G. M. & GRONENBORN, A. M. 1998a. The solution structure of a fungal AREA protein-DNA complex: an alternative binding mode for the basic carboxyl tail of GATA factors. *J Mol Biol*, 277, 605-20.
- STARICH, M. R., WIKSTROM, M., SCHUMACHER, S., ARST, H. N., JR., GRONENBORN, A. M. & CLORE, G. M. 1998b. The solution structure of the Leu22-->Val mutant AREA DNA binding domain complexed with a TGATAG core element defines a role for hydrophobic packing in the determination of specificity. *J Mol Biol*, 277, 621-34.
- STAVREVA, D. A., MULLER, W. G., HAGER, G. L., SMITH, C. L. & MCNALLY, J. G. 2004. Rapid glucocorticoid receptor exchange at a promoter is coupled to transcription and regulated by chaperones and proteasomes. *Mol Cell Biol*, 24, 2682-97.
- STENOIEN, D. L., PATEL, K., MANCINI, M. G., DUTERTRE, M., SMITH, C. L., O'MALLEY, B. W. & MANCINI, M. A. 2001. FRAP reveals that mobility of oestrogen receptor-alpha is ligand- and proteasome-dependent. *Nat Cell Biol*, 3, 15-23.
- STORMO, G. D. 2000. DNA binding sites: representation and discovery. *Bioinformatics*, 16, 16-23.
- STORMO, G. D. & ZHAO, Y. 2010. Determining the specificity of protein-DNA interactions. *Nat Rev Genet*, 11, 751-60.
- STUMPF, M., WASKOW, C., KROTSCHER, M., VAN ESSEN, D., RODRIGUEZ, P., ZHANG, X., GUYOT, B., ROEDER, R. G. & BORGGREFE, T. 2006. The mediator complex functions as a coactivator for GATA-1 in erythropoiesis via subunit Med1/TRAP220. *Proc Natl Acad Sci U S A*, 103, 18504-9.
- SU, M. Y., STEINER, L. A., BOGARDUS, H., MISHRA, T., SCHULZ, V. P., HARDISON, R. C. & GALLAGHER, P. G. 2013. Identification of biologically relevant enhancers in human erythroid cells. *J Biol Chem*, 288, 8433-44.
- SUTER, D. M., MOLINA, N., GATFIELD, D., SCHNEIDER, K., SCHIBLER, U. & NAEF, F. 2011. Mammalian genes are transcribed with widely different bursting kinetics. *Science*, 332, 472-4.
- SUTHERLAND, J. A., TURNER, A. R., MANNONI, P., MCGANN, L. E. & TURC, J. M. 1986. Differentiation of K562 leukemia cells along erythroid, macrophage, and megakaryocyte lineages. *J Biol Response Mod*, 5, 250-62.
- SZYMANSKI, P. & LEVINE, M. 1995. Multiple modes of dorsal-bHLH transcriptional synergy in the Drosophila embryo. *EMBO J*, 14, 2229-38.
- TALLACK, M. R., MAGOR, G. W., DARTIGUES, B., SUN, L., HUANG, S., FITTOCK, J. M., FRY, S. V., GLAZOV, E. A., BAILEY, T. L. & PERKINS, A. C. 2012. Novel roles for KLF1 in erythropoiesis revealed by mRNA-seq. *Genome Res*, 22, 2385-98.
- TALLACK, M. R., WHITTINGTON, T., YUEN, W. S., WAINWRIGHT, E. N., KEYS, J. R., GARDINER, B. B., NOURBAKHSH, E., CLOONAN, N., GRIMMOND, S. M., BAILEY, T. L. & PERKINS, A. C. 2010. A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. *Genome Res*, 20, 1052-63.
- TANAY, A. 2006. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res*, 16, 962-72.
- TANIUCHI, I., OSATO, M., EGAWA, T., SUNSHINE, M. J., BAE, S. C., KOMORI, T., ITO, Y. & LITTMAN, D. R. 2002. Differential requirements for Runx proteins in CD4 repression and epigenetic silencing during T lymphocyte development. *Cell*, 111, 621-33.
- TERZI, N., CHURCHMAN, L. S., VASILJEVA, L., WEISSMAN, J. & BURATOWSKI, S. 2011. H3K4 trimethylation by Set1 promotes efficient termination by the Nrd1-Nab3-Sen1 pathway. *Mol Cell Biol*, 31, 3569-83.
- THOMSON, J. P., SKENE, P. J., SELFRIDGE, J., CLOUAIRE, T., GUY, J., WEBB, S., KERR, A. R., DEATON, A., ANDREWS, R., JAMES, K. D., TURNER, D. J.,

- ILLINGWORTH, R. & BIRD, A. 2010. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature*, 464, 1082-6.
- TIE, F., BANERJEE, R., STRATTON, C. A., PRASAD-SINHA, J., STEPANIK, V., ZLOBIN, A., DIAZ, M. O., SCACHERI, P. C. & HARTE, P. J. 2009. CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing. *Development*, 136, 3131-41.
- TIJSSEN, M. R., CVEJIC, A., JOSHI, A., HANNAH, R. L., FERREIRA, R., FORRAI, A., BELLISSIMO, D. C., ORAM, S. H., SMETHURST, P. A., WILSON, N. K., WANG, X., OTTERSBAACH, K., STEMPLE, D. L., GREEN, A. R., OUWEHAND, W. H. & GOTTGENS, B. 2011. Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. *Dev Cell*, 20, 597-609.
- TJANDRA, N., OMICHINSKI, J. G., GRONENBORN, A. M., CLORE, G. M. & BAX, A. 1997. Use of dipolar $1H-15N$ and $1H-13C$ couplings in the structure determination of magnetically oriented macromolecules in solution. *Nat Struct Biol*, 4, 732-8.
- TRAINOR, C. D., GHIRLANDO, R. & SIMPSON, M. A. 2000. GATA zinc finger interactions modulate DNA binding and transactivation. *J Biol Chem*, 275, 28157-66.
- TRAINOR, C. D., OMICHINSKI, J. G., VANDERGON, T. L., GRONENBORN, A. M., CLORE, G. M. & FELSENFELD, G. 1996. A palindromic regulatory site within vertebrate GATA-1 promoters requires both zinc fingers of the GATA-1 DNA-binding domain for high-affinity interaction. *Mol Cell Biol*, 16, 2238-47.
- TSAI, F. Y., KELLER, G., KUO, F. C., WEISS, M., CHEN, J., ROSENBLATT, M., ALT, F. W. & ORKIN, S. H. 1994. An early haematopoietic defect in mice lacking the transcription factor GATA-2. *Nature*, 371, 221-6.
- TSAI, F. Y. & ORKIN, S. H. 1997. Transcription factor GATA-2 is required for proliferation/survival of early hematopoietic cells and mast cell formation, but not for erythroid and myeloid terminal differentiation. *Blood*, 89, 3636-43.
- TSANG, A. P., FUJIWARA, Y., HOM, D. B. & ORKIN, S. H. 1998. Failure of megakaryopoiesis and arrested erythropoiesis in mice lacking the GATA-1 transcriptional cofactor FOG. *Genes Dev*, 12, 1176-88.
- TSANG, A. P., VISVADER, J. E., TURNER, C. A., FUJIWARA, Y., YU, C., WEISS, M. J., CROSSLEY, M. & ORKIN, S. H. 1997. FOG, a multitype zinc finger protein, acts as a cofactor for transcription factor GATA-1 in erythroid and megakaryocytic differentiation. *Cell*, 90, 109-19.
- TURNER, J. & CROSSLEY, M. 1998. Cloning and characterization of mCtBP2, a co-repressor that associates with basic Kruppel-like factor and other mammalian transcriptional regulators. *EMBO J*, 17, 5129-40.
- TYAGI, S., CHABES, A. L., WYSOCKA, J. & HERR, W. 2007. E2F activation of S phase promoters via association with HCF-1 and the MLL family of histone H3K4 methyltransferases. *Mol Cell*, 27, 107-19.
- VAKOC, C. R., LETTING, D. L., GHELDOLF, N., SAWADO, T., BENDER, M. A., GROUDINE, M., WEISS, M. J., DEKKER, J. & BLOBEL, G. A. 2005. Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. *Mol Cell*, 17, 453-62.
- VALGE-ARCHER, V. E., OSADA, H., WARREN, A. J., FORSTER, A., LI, J., BAER, R. & RABBITS, T. H. 1994. The LIM protein RBTN2 and the basic helix-loop-helix protein TAL1 are present in a complex in erythroid cells. *Proc Natl Acad Sci U S A*, 91, 8617-21.
- VALLEDOR, A. F., BORRAS, F. E., CULLELL-YOUNG, M. & CELADA, A. 1998. Transcription factors that regulate monocyte/macrophage differentiation. *J Leukoc Biol*, 63, 405-17.
- VAN ROYEN, M. E., ZOTTER, A., IBRAHIM, S. M., GEVERTS, B. & HOUTSMULLER, A. B. 2011. Nuclear proteins: finding and binding target sites in chromatin. *Chromosome Res*, 19, 83-98.

- VAN STEENSEL, B. & HENIKOFF, S. 2000. Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat Biotechnol*, 18, 424-8.
- VASTENHOEW, N. L., ZHANG, Y., WOODS, I. G., IMAM, F., REGEV, A., LIU, X. S., RINN, J. & SCHIER, A. F. 2010. Chromatin signature of embryonic pluripotency is established during genome activation. *Nature*, 464, 922-6.
- VERMEULEN, M., MULDER, K. W., DENISSOV, S., PIJNAPPEL, W. W., VAN SCHAIK, F. M., VARIER, R. A., BALTISSEN, M. P., STUNNENBERG, H. G., MANN, M. & TIMMERS, H. T. 2007. Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell*, 131, 58-69.
- VERNIMMEN, D., DE GOBBI, M., SLOANE-STANLEY, J. A., WOOD, W. G. & HIGGS, D. R. 2007. Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. *EMBO J*, 26, 2041-51.
- VICENT, G. P., NACHT, A. S., FONT-MATEU, J., CASTELLANO, G., GAVEGLIA, L., BALLARE, C. & BEATO, M. 2011. Four enzymes cooperate to displace histone H1 during the first minute of hormonal gene activation. *Genes Dev*, 25, 845-62.
- VICENT, G. P., ZAURIN, R., NACHT, A. S., LI, A., FONT-MATEU, J., LE DILY, F., VERMEULEN, M., MANN, M. & BEATO, M. 2009. Two chromatin remodeling activities cooperate during activation of hormone responsive promoters. *PLoS Genet*, 5, e1000567.
- VITALIANO-PRUNIER, A., MENANT, A., HOBEIKA, M., GELI, V., GWIZDEK, C. & DARGEMONT, C. 2008. Ubiquitylation of the COMPASS component Swd2 links H2B ubiquitylation to H3K4 trimethylation. *Nat Cell Biol*, 10, 1365-71.
- VONDERFECHT, T. R., SCHROYER, D. C., SCHENCK, B. L., MCDONOUGH, V. M. & PIKAART, M. J. 2008. Substitution of DNA-contacting amino acids with functional variants in the Gata-1 zinc finger: a structurally and phylogenetically guided mutagenesis. *Biochem Biophys Res Commun*, 369, 1052-6.
- VOO, K. S., CARLONE, D. L., JACOBSEN, B. M., FLODIN, A. & SKALNIK, D. G. 2000. Cloning of a mammalian transcriptional activator that binds unmethylated CpG motifs and shares a CXXC domain with DNA methyltransferase, human trithorax, and methyl-CpG binding domain protein 1. *Mol Cell Biol*, 20, 2108-21.
- VYAS, P., MCDEVITT, M. A., CANTOR, A. B., KATZ, S. G., FUJIWARA, Y. & ORKIN, S. H. 1999. Different sequence requirements for expression in erythroid and megakaryocytic cells within a regulatory element upstream of the GATA-1 gene. *Development*, 126, 2799-811.
- WADMAN, I. A., OSADA, H., GRUTZ, G. G., AGULNICK, A. D., WESTPHAL, H., FORSTER, A. & RABBITTS, T. H. 1997. The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins. *EMBO J*, 16, 3145-57.
- WALTZER, L., BATAILLE, L., PEYREFITTE, S. & HAENLIN, M. 2002. Two isoforms of Serpentine containing either one or two GATA zinc fingers have different roles in *Drosophila* haematopoiesis. *EMBO J*, 21, 5477-86.
- WANG, P., LIN, C., SMITH, E. R., GUO, H., SANDERSON, B. W., WU, M., GOGOL, M., ALEXANDER, T., SEIDEL, C., WIEDEMANN, L. M., GE, K., KRUMLAUF, R. & SHILATIFARD, A. 2009. Global analysis of H3K4 methylation defines MLL family member targets and points to a role for MLL1-mediated H3K4 methylation in the regulation of transcriptional initiation by RNA polymerase II. *Mol Cell Biol*, 29, 6074-85.
- WANG, Q., LI, W., LIU, X. S., CARROLL, J. S., JANNE, O. A., KEETON, E. K., CHINNAIYAN, A. M., PIENTA, K. J. & BROWN, M. 2007. A hierarchical network of transcription factors governs androgen receptor-dependent prostate cancer growth. *Mol Cell*, 27, 380-92.
- WANG, Q., STACY, T., BINDER, M., MARIN-PADILLA, M., SHARPE, A. H. & SPECK, N. A. 1996. Disruption of the Cbfa2 gene causes necrosis and hemorrhaging in the

- central nervous system and blocks definitive hematopoiesis. *Proc Natl Acad Sci U S A*, 93, 3444-9.
- WARREN, C. L., KRATOCHVIL, N. C., HAUSCHILD, K. E., FOISTER, S., BREZINSKI, M. L., DERVAN, P. B., PHILLIPS, G. N., JR. & ANSARI, A. Z. 2006. Defining the sequence-recognition profile of DNA-binding molecules. *Proc Natl Acad Sci U S A*, 103, 867-72.
- WECHSLER, J., GREENE, M., MCDEVITT, M. A., ANASTASI, J., KARP, J. E., LE BEAU, M. M. & CRISPINO, J. D. 2002. Acquired mutations in GATA1 in the megakaryoblastic leukemia of Down syndrome. *Nat Genet*, 32, 148-52.
- WEI, G., ABRAHAM, B. J., YAGI, R., JOTHI, R., CUI, K., SHARMA, S., NARLIKAR, L., NORTHRUP, D. L., TANG, Q., PAUL, W. E., ZHU, J. & ZHAO, K. 2011. Genome-wide analyses of transcription factor GATA3-mediated gene regulation in distinct T cell types. *Immunity*, 35, 299-311.
- WEINBERG, R. L., VEPRINTSEV, D. B. & FERSHT, A. R. 2004. Cooperative binding of tetrameric p53 to DNA. *J Mol Biol*, 341, 1145-59.
- WEINTRAUB, H. & GROUDINE, M. 1976. Chromosomal subunits in active genes have an altered conformation. *Science*, 193, 848-56.
- WEISS, M. J., KELLER, G. & ORKIN, S. H. 1994. Novel insights into erythroid development revealed through in vitro differentiation of GATA-1 embryonic stem cells. *Genes Dev*, 8, 1184-97.
- WEISS, M. J. & ORKIN, S. H. 1995. Transcription factor GATA-1 permits survival and maturation of erythroid precursors by preventing apoptosis. *Proc Natl Acad Sci U S A*, 92, 9623-7.
- WEISS, M. J., YU, C. & ORKIN, S. H. 1997. Erythroid-cell-specific properties of transcription factor GATA-1 revealed by phenotypic rescue of a gene-targeted cell line. *Mol Cell Biol*, 17, 1642-51.
- WELCH, J. J., WATTS, J. A., VAKOC, C. R., YAO, Y., WANG, H., HARDISON, R. C., BLOBEL, G. A., CHODOSH, L. A. & WEISS, M. J. 2004. Global regulation of erythroid gene expression by transcription factor GATA-1. *Blood*, 104, 3136-47.
- WHYATT, D. J., DEBOER, E. & GROSVELD, F. 1993. The two zinc finger-like domains of GATA-1 have different DNA binding specificities. *EMBO J*, 12, 4993-5005.
- WIERENGA, A. T., VELLENGA, E. & SCHURINGA, J. J. 2010. Down-regulation of GATA1 uncouples STAT5-induced erythroid differentiation from stem/progenitor cell proliferation. *Blood*, 115, 4367-76.
- WILKINSON-WHITE, L., GAMSJAEGER, R., DASTMALCHI, S., WIENERT, B., STOKES, P. H., CROSSLEY, M., MACKAY, J. P. & MATTHEWS, J. M. 2011. Structural basis of simultaneous recruitment of the transcriptional regulators LMO2 and FOG1/ZFPM1 by the transcription factor GATA1. *Proc Natl Acad Sci U S A*, 108, 14443-8.
- WILSON, N. K., FOSTER, S. D., WANG, X., KNEZEVIC, K., SCHUTTE, J., KAIMAKIS, P., CHILARSKA, P. M., KINSTON, S., OUWEHAND, W. H., DZIERZAK, E., PIMANDA, J. E., DE BRUIJN, M. F. & GOTTGENS, B. 2010. Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell*, 7, 532-44.
- WOLFF, L. & HUMENIUK, R. 2013. Concise review: erythroid versus myeloid lineage commitment: regulating the master regulators. *Stem Cells*, 31, 1237-44.
- WONG, D., TEIXEIRA, A., OIKONOMOPOULOS, S., HUMBURG, P., LONE, I. N., SALIBA, D., SIGGERS, T., BULYK, M., ANGELOV, D., DIMITROV, S., UDALOVA, I. A. & RAGOISSIS, J. 2011a. Extensive characterization of NF-kappaB binding uncovers non-canonical motifs and advances the interpretation of genetic functional traits. *Genome Biol*, 12, R70.
- WONG, P., HATTANGADI, S. M., CHENG, A. W., FRAMPTON, G. M., YOUNG, R. A. & LODISH, H. F. 2011b. Gene induction and repression during terminal erythropoiesis are mediated by distinct epigenetic changes. *Blood*, 118, e128-38.

- WONTAKAL, S. N., GUO, X., SMITH, C., MACCARTHY, T., BRESNICK, E. H., BERGMAN, A., SNYDER, M. P., WEISSMAN, S. M., ZHENG, D. & SKOULTCHI, A. I. 2012. A core erythroid transcriptional network is repressed by a master regulator of myelo-lymphoid differentiation. *Proc Natl Acad Sci U S A*, 109, 3832-7.
- WOO, A. J., MORAN, T. B., SCHINDLER, Y. L., CHOE, S. K., LANGER, N. B., SULLIVAN, M. R., FUJIWARA, Y., PAW, B. H. & CANTOR, A. B. 2008. Identification of ZBP-89 as a novel GATA-1-associated transcription factor involved in megakaryocytic and erythroid development. *Mol Cell Biol*, 28, 2675-89.
- WOOD, A., KROGAN, N. J., DOVER, J., SCHNEIDER, J., HEIDT, J., BOATENG, M. A., DEAN, K., GOLSHANI, A., ZHANG, Y., GREENBLATT, J. F., JOHNSTON, M. & SHILATIFARD, A. 2003a. Bre1, an E3 ubiquitin ligase required for recruitment and substrate selection of Rad6 at a promoter. *Mol Cell*, 11, 267-74.
- WOOD, A., SCHNEIDER, J., DOVER, J., JOHNSTON, M. & SHILATIFARD, A. 2003b. The Paf1 complex is essential for histone monoubiquitination by the Rad6-Bre1 complex, which signals for histone methylation by COMPASS and Dot1p. *J Biol Chem*, 278, 34739-42.
- WOON KIM, Y., KIM, S., GEUN KIM, C. & KIM, A. 2011. The distinctive roles of erythroid specific activator GATA-1 and NF-E2 in transcription of the human fetal gamma-globin genes. *Nucleic Acids Res*, 39, 6944-55.
- WU, C. 1980. The 5' ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. *Nature*, 286, 854-60.
- WU, M., WANG, P. F., LEE, J. S., MARTIN-BROWN, S., FLORENS, L., WASHBURN, M. & SHILATIFARD, A. 2008. Molecular regulation of H3K4 trimethylation by Wdr82, a component of human Set1/COMPASS. *Mol Cell Biol*, 28, 7337-44.
- WU, W., CHENG, Y., KELLER, C. A., ERNST, J., KUMAR, S. A., MISHRA, T., MORRISSEY, C., DORMAN, C. M., CHEN, K. B., DRAUTZ, D., GIARDINE, B., SHIBATA, Y., SONG, L., PIMKIN, M., CRAWFORD, G. E., FUREY, T. S., KELLIS, M., MILLER, W., TAYLOR, J., SCHUSTER, S. C., ZHANG, Y., CHIAROMONTE, F., BLOBEL, G. A., WEISS, M. J. & HARDISON, R. C. 2011. Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration. *Genome Res*, 21, 1659-71.
- WYSOCKA, J., SWIGUT, T., XIAO, H., MILNE, T. A., KWON, S. Y., LANDRY, J., KAUER, M., TACKETT, A. J., CHAIT, B. T., BADENHORST, P., WU, C. & ALLIS, C. D. 2006. A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature*, 442, 86-90.
- XU, G., KANEZAKI, R., TOKI, T., WATANABE, S., TAKAHASHI, Y., TERUI, K., KITABAYASHI, I. & ITO, E. 2006. Physical association of the patient-specific GATA1 mutants with RUNX1 in acute megakaryoblastic leukemia accompanying Down syndrome. *Leukemia*, 20, 1002-8.
- XU, G., NAGANO, M., KANEZAKI, R., TOKI, T., HAYASHI, Y., TAKETANI, T., TAKI, T., MITUI, T., KOIKE, K., KATO, K., IMAIZUMI, M., SEKINE, I., IKEDA, Y., HANADA, R., SAKO, M., KUDO, K., KOJIMA, S., OHNEDA, O., YAMAMOTO, M. & ITO, E. 2003. Frequent mutations in the GATA-1 gene in the transient myeloproliferative disorder of Down syndrome. *Blood*, 102, 2960-8.
- XU, J., SHAO, Z., GLASS, K., BAUER, D. E., PINELLO, L., VAN HANDEL, B., HOU, S., STAMATOYANNOPOULOS, J. A., MIKKOLA, H. K., YUAN, G. C. & ORKIN, S. H. 2012. Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. *Dev Cell*, 23, 796-811.
- YAGI, R., CHEN, L. F., SHIGESADA, K., MURAKAMI, Y. & ITO, Y. 1999. A WW domain-containing yes-associated protein (YAP) is a novel transcriptional co-activator. *EMBO J*, 18, 2551-62.
- YANG, H. Y. & EVANS, T. 1992. Distinct roles for the two cGATA-1 finger domains. *Mol Cell Biol*, 12, 4562-70.
- YAO, J., MUNSON, K. M., WEBB, W. W. & LIS, J. T. 2006. Dynamics of heat shock factor association with native gene loci in living cells. *Nature*, 442, 1050-3.

- YIEN, Y. Y. & BIEKER, J. J. 2013. EKLF/KLF1, a tissue-restricted integrator of transcriptional control, chromatin remodeling, and lineage determination. *Mol Cell Biol*, 33, 4-13.
- YU, C., NIAKAN, K. K., MATSUSHITA, M., STAMATOYANNOPOULOS, G., ORKIN, S. H. & RASKIND, W. H. 2002. X-linked thrombocytopenia with thalassemia from a mutation in the amino finger of GATA-1 affecting DNA binding rather than FOG-1 interaction. *Blood*, 100, 2040-5.
- YU, M., RIVA, L., XIE, H., SCHINDLER, Y., MORAN, T. B., CHENG, Y., YU, D., HARDISON, R., WEISS, M. J., ORKIN, S. H., BERNSTEIN, B. E., FRAENKEL, E. & CANTOR, A. B. 2009. Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. *Mol Cell*, 36, 682-95.
- YU, P. & KODADEK, T. 2007. Dynamics of the hypoxia-inducible factor-1-vascular endothelial growth factor promoter complex. *J Biol Chem*, 282, 35035-45.
- YU, Y. L., CHIANG, Y. J., CHEN, Y. C., PAPETTI, M., JUO, C. G., SKOULTCHI, A. I. & YEN, J. J. 2005. MAPK-mediated phosphorylation of GATA-1 promotes Bcl-XL expression and cell survival. *J Biol Chem*, 280, 29533-42.
- ZANTON, S. J. & PUGH, B. F. 2006. Full and partial genome-wide assembly and disassembly of the yeast transcription machinery in response to heat shock. *Genes Dev*, 20, 2250-65.
- ZARET, K. S. & CARROLL, J. S. 2011. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev*, 25, 2227-41.
- ZEE, B. M., LEVIN, R. S., XU, B., LEROY, G., WINGREEN, N. S. & GARCIA, B. A. 2010. In vivo residue-specific histone methylation dynamics. *J Biol Chem*, 285, 3341-50.
- ZELIN, E., ZHANG, Y., TOOGUN, O. A., ZHONG, S. & FREEMAN, B. C. 2012. The p23 molecular chaperone and GCN5 acetylase jointly modulate protein-DNA dynamics and open chromatin status. *Mol Cell*, 48, 459-70.
- ZENKLUSEN, D., LARSON, D. R. & SINGER, R. H. 2008. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat Struct Mol Biol*, 15, 1263-71.
- ZHANG, J. & LODISH, H. F. 2007. Endogenous K-ras signaling in erythroid differentiation. *Cell Cycle*, 6, 1970-3.
- ZHANG, J., SOCOLOVSKY, M., GROSS, A. W. & LODISH, H. F. 2003. Role of Ras signaling in erythroid differentiation of mouse fetal liver cells: functional analysis by a flow cytometry-based novel culture system. *Blood*, 102, 3938-46.
- ZHANG, P., BEHRE, G., PAN, J., IWAMA, A., WARA-ASWAPATI, N., RADOMSKA, H. S., AURON, P. E., TENEN, D. G. & SUN, Z. 1999. Negative cross-talk between hematopoietic regulators: GATA proteins repress PU.1. *Proc Natl Acad Sci U S A*, 96, 8705-10.
- ZHANG, P., ZHANG, X., IWAMA, A., YU, C., SMITH, K. A., MUELLER, B. U., NARRAVULA, S., TORBETT, B. E., ORKIN, S. H. & TENEN, D. G. 2000. PU.1 inhibits GATA-1 function and erythroid differentiation by blocking GATA-1 DNA binding. *Blood*, 96, 2641-8.
- ZHANG, W. & BIEKER, J. J. 1998. Acetylation and modulation of erythroid Kruppel-like factor (EKLF) activity by interaction with histone acetyltransferases. *Proc Natl Acad Sci U S A*, 95, 9855-60.
- ZHANG, W., KADAM, S., EMERSON, B. M. & BIEKER, J. J. 2001. Site-specific acetylation by p300 or CREB binding protein regulates erythroid Kruppel-like factor transcriptional activity via its interaction with the SWI-SNF complex. *Mol Cell Biol*, 21, 2413-22.
- ZHANG, Y., LIU, T., MEYER, C. A., EECKHOUTE, J., JOHNSON, D. S., BERNSTEIN, B. E., NUSBAUM, C., MYERS, R. M., BROWN, M., LI, W. & LIU, X. S. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, 9, R137.
- ZHAO, Y., GRANAS, D. & STORMO, G. D. 2009. Inferring binding energies from selected binding sites. *PLoS Comput Biol*, 5, e1000590.

- ZHENG, S., WYRICK, J. J. & REESE, J. C. 2010. Novel trans-tail regulation of H2B ubiquitylation and H3K4 methylation by the N terminus of histone H2A. *Mol Cell Biol*, 30, 3635-45.
- ZHOU, M., OUYANG, W., GONG, Q., KATZ, S. G., WHITE, J. M., ORKIN, S. H. & MURPHY, K. M. 2001. Friend of GATA-1 represses GATA-3-dependent activity in CD4+ T cells. *J Exp Med*, 194, 1461-71.
- ZHOU, V. W., GOREN, A. & BERNSTEIN, B. E. 2011. Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet*, 12, 7-18.
- ZON, L. I., YAMAGUCHI, Y., YEE, K., ALBEE, E. A., KIMURA, A., BENNETT, J. C., ORKIN, S. H. & ACKERMAN, S. J. 1993. Expression of mRNA for the GATA-binding proteins in human eosinophils and basophils: potential role in gene transcription. *Blood*, 81, 3234-41.
- ZYKOVICH, A., KORF, I. & SEGAL, D. J. 2009. Bind-n-Seq: high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. *Nucleic Acids Res*, 37, e151.