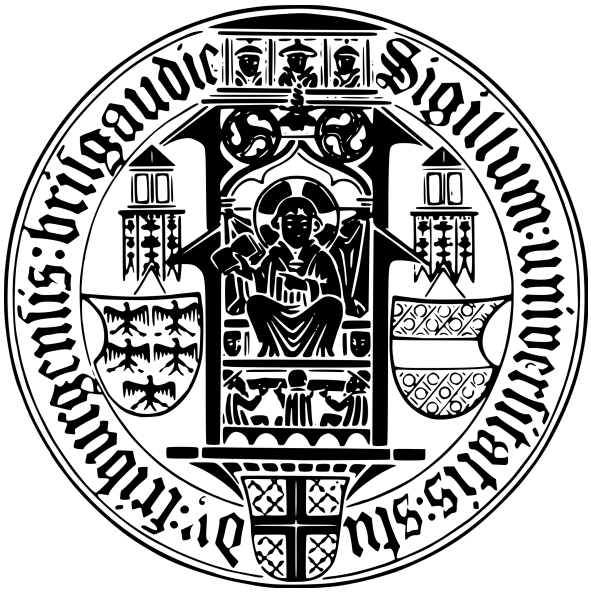
Max Planck Institut für Immunbiologie und Epigenetik
Freiburg im Breisgau



Bioinformatic analyses of MOF-containing complexes in mammals and *Drosophila*

Inaugural-Dissertation

zur Erlangung der Doktorwürde
der Fakultät für Biologie
der Albert-Ludwigs-Universität Freiburg im Breisgau

vorgelegt von
Friederike Dündar

Freiburg im Breisgau
Oktober 2014

Dekan der Fakultät für Biologie: Prof. Dr. Wolfgang Driever
Promotionsvorsitzender: Prof. Dr. Stefan Rotter
Betreuer der Arbeit: Dr. Asifa Akhtar
Referent: Dr. Asifa Akhtar
Koreferent: Prof. Dr. Wolfgang Hess
Drittprüfer: Prof. Dr. Jörn Dengjel
Datum der mündlichen Prüfung:

Acknowledgements

I cannot imagine a place where I would have rather spent my PhD years than the MPI-IE. Representatively, I would like to thank the entire Akhtar Group, the Bioinformatics Group and the Deep Sequencing Unit at the MPI-IE – for endless, fruitful discussions about everything, and for being awesome peers! All my projects were collaborative projects in the best sense and none of them would have been accomplished without the work and insights from everyone involved.

I am especially indebted to:

Asifa who had so much faith in my abilities that did not even exist at the beginning of my PhD, gave me all the support and freedom that I needed to find my way into the bioinformatics community, inspired and motivated me without end with her enthusiasm. **Fidel** who first helped me lose my feaR and then pushed me towards python, eagerly discussed tedious problems and details, enabled me (and many others) to do analyses faster and more impressively than I would have ever imagined, always takes initiative and seeks solutions rather than work-arounds, and makes the most delicious veggie sticks. **Ken** who introduced me to ChIP-seq analysis, the NSL complex and everything else I needed to know, who never ran out of sharp insights, clever questions and new analyses and who truly impressed me with his honest passion for science. I don't know if I had ever become a computational biologist without his help and inspiration. **Sarah** who painlessly introduced me to the UNIX shell and scripting, installed countless stubborn programs, sat with me for hours hunting other people's code bugs, masters the Galaxy, tames the servers and tirelessly fights back the rising chaos of increasing data and people messing with it, and, most importantly, selflessly ensures a constant supply of Luxemburgian coffee pads. **Thomas** who let me become part of the bioinformatics group despite my utter lack of bioinformatics savvy in the beginning, whose support I could always, always count on, whose one-on-one supervision was absolutely essential for the success of my very first bioinformatic steps and without whom I still would have never driven a go-kart. **Tomek** who trusted me with his data, was a constant source of inspiring ideas, made the most beautiful illustrations and always brought exquisite delicacies from Poland and Paris.

I would also like to thank Andrew Pospisilik and Michael Stadler who supported me during TAC meetings and beyond. I couldn't have had a better TAC!

Contents

List of Figures

List of Tables

Abbreviations

ac	acetylation
AT	adenine, thymine
ATP	adenosine triphosphate
bp	base pair
ChIP	chromatin immunoprecipitation
CTD	C-terminal domain
D., d	*Drosophila*
DNA	deoxyribonucleic acid
ESC	embryonic stem cell
G2/M	pre-mitotic phase to mitosis
GC	dinucleotide: guanine, cytosine
h	human
H2B, H3, H4	histones 2B, 3, 4
HAT	histone acetyl transferase
K	lysine
KAT	lysine acetyl transferase
kb	kilo base pair
KMT	lysine methyl transferase
M., m	*Mus*, murine, mouse
me	methylation
modENCODE	(model organism) encyclopedia of DNA elements
NAD	nicotinamide adenine dinucleotide
NPC	neuronal progenitor cell
PCR	polymerase chain reaction
PIC	pre-initiation complex
Pol II	RNA polymerase II
RNA	ribonucleic acid
RNAi	RNA interference
seq	high-throughput DNA sequencing

SES	signal extraction scaling
shRNA	small hairpin RNA
TES	transcription end site
TF	transcription factor
TSS	trancription start site
ub	ubiquitylation
UTR	untranslated region

1. Summary

1.1 Zusammenfassung

Die Histon-Acetyltransferase MOF (males absent on the first) ist das wichtigste Enzym für die Acetylierung von Lysin 16 des Histons H4, wobei die katalytische Spezifität und Effektivität stark von ihren Interaktionspartnern, den MSL (male-specific lethal) und NSL (non-specific lethal) Komplexen, bestimmt wird. Der MSL Komplex ist von herausragender Bedeutung in der Taufliege (*D. melanogaster*), wo er für die transkriptionelle Kompensation der reduzierten Gendosis in X-chromosomal hemizygoten *Drosophila*-Männchen verantwortlich ist. Welche Funktionen der MSL Komplex in Säugetieren erfüllen könnte, war zu Beginn meiner Arbeit kaum bekannt, ebensowenig wie die Rolle des NSL Komplexes, welcher weder in der Fliege noch in Säugetieren umfänglich erforscht war. Das Ziel meines Projekts war es, unser Wissen über beide MOF-Komplexe deutlich zu erweitern.

Die Grundlage meiner Arbeit bildeten Chromatin-Immunpräzipitationsexperimente gefolgt von Hochdurchsatz-DNA-Sequenzierung (ChIP-seq) für deren bioinformatische Analyse zunächst standardisierte Protokolle sowie individuelle Auswertungen etabliert werden mussten. Das daraus resultierende Software-Paket deepTools kann nun für Qualitätskontrollen, Normalisierungen und Datenprozessierung ebenso verwendet werden wie für die bildliche Darstellung der Hochdurchsatz-DNA-Sequenzierungsdaten.

Wir untersuchten die genomweiten Bindestellen des NSL Komplexes in *Drosophila*- (NSL1, NSL3, MCRS2, MBD-R2) und Mauszellen (NSL3, MCRS1) und konnten mittels umfangreicher Charakterisierung der Zielgene und RNAi-basierten Experimenten zeigen, dass der NSL Komplex für die Exprimierung konstitutiv aktiver Gene vonnöten ist, um u.a. eine ausreichende Rekrutierung der RNA Polymerase II innerhalb des Präinitiationskomplexes zu garantieren.

Zusätzlich zum NSL Komplex untersuchten wir auch MOF und den MSL Komplex (MSL1, MSL2) in Mauszellen. In der überwiegenden Mehrzahl der Promoterbindestellen fanden

wir MOF gemeinsam mit dem NSL Komplex vor, in einigen Fällen lag zusätzlich ein Signale des MSL Komplexes vor. Obwohl der MSL Komplex – anders als in der Taufliege – in Mauszellen keine starke Präferenz für das X Chromosom zeigt, stellten sich MSL1 und MSL2 als essentiell für den Erhalt der X-chromosomalen Genexprimierung in embryonalen Stammzellen heraus, da in ihrer Abwesenheit die Transkription von *Tsix* und die damit verbundene Produktionshemmung der X-inaktivierenden *Xist*-RNA stark beeinträchtigt ist. Der NSL Komplex trägt indirekt zur Inhibierung der X-Inaktivierung bei, indem er für die Exprimierung von Pluripotenzfaktoren wie Nanog, Oct4 und Esrrb in embryonalen Stammzellen erfoderlich ist. Darüber hinaus verglichen wir die genomweiten Signale von MSL1 in evolutionär weit voneinander entfernten Spezies (*D. melanogaster, D. virilis* und *M. musculus*), aus welchen wir schlussfolgerten, dass MSL1 geschlechtsunabhängig an Genpromotoren bindet – anders als bislang angenommen, auch in *D. melanogaster*.

Zusammenfassend konnten wir zeigen, dass die mit MOF interagierenden Proteinkomplexe zum Teil deutlich voneinander abgrenzbare Funktionen ausführen: während der NSL Komplex in großem Ausmaß die grundlegende Exprimierung von basalen Haushaltsgenen reguliert, ist der MSL Komplex an hoch spezialisierten, aber ebenfalls lebenswichtigen Prozessen beteiligt.

1.2 Abstract

The histone acetyl transferase males absent on the first (MOF) is responsible for the majority of histone H4 lysine 16 acetylation in *Drosophila* and mammals. Its catalytic specificity and efficiency depend on the interaction with two protein complexes, the male-specific lethal (MSL) complex and the non-specific lethal (NSL) complex. The *Drosophila* MSL complex has been thoroughly examined as it is essential for the transcriptional upregulation of the single male *Drosophila* X chromosome to meet autosomal gene expression levels. Its function in mammals, however, was not clear. Likewise, the role of the NSL complex was poorly understood in both *Drosophila* and mammalian cells. The aim of my project was to further our insights into these two distinct MOF-associated complexes.

All studies presented here were centered around chromatin immunoprecipitation experiments followed by high-throughput DNA sequencing (ChIP-seq). The analyses required the setup of a universal bioinformatic pipeline and customized workflows for downstream analyses and visualizations. These efforts became part of the deepTools software package that allows efficient and reproducible generation of normalized coverage files, offers quality controls and highly customizable visualization of high-throughput sequencing data.

By investigating the genome-wide binding of NSL complex members in *Drosophila* (NSL1, NSL3, MCRS2, MBD-R2) and mouse cells (NSL3, MCRS1) and subsequent extensive characterization of target genes coupled with perturbation experiments, we revealed that the NSL complex is an evolutionarily conserved regulator of housekeeping gene expression that is required for optimal recruitment of the pre-initiation complex.

In mammals, we complemented our study of the NSL complex with genome-wide profiles of MOF, MSL1 and MSL2 in murine embryonic stem cells (ESC) and neuronal progenitor cells (NPC). We determined constant and dynamic binding during differentiation and established the patterns of exclusive and concomitant targeting of both MOF-associated complexes. We found that the NSL complex is the predominant interaction partner of MOF in ESCs and NPCs. While the MSL complex is not specifically enriched on the mouse X chromosome, we could show that MSL1 and MSL2 are important for the maintenance of active X expression in ESCs by maintaining transcription of *Tsix* whose expression inhibits the production of the X-inactivating transcript, *Xist*. The NSL complex indirectly contributes to the repression of X inactivation through expression regulation of key pluripotency factors such as Nanog, Oct4 and Esrrb. Furthermore, the role of MSL1 was elucidated in more detail as the genome-wide profiles from distant species (*D. melanogaster, D. virilis* and *M. musculus*) revealed evolutionarily conserved binding to gene promoters in a sex-independent manner.

In summary, we could show that MOF-associated complexes fulfill distinct vital roles in *Drosophila* and mammals: while the NSL complex is an abundant regulator of basic cellular gene expression, the MSL complex was found to contribute to highly specialized functions.

2. Introduction

Living cells constantly need to monitor their own status as well as their environment while maintaining the means to absorb and utilize energy. The majority of these functions is carried out by a vastly diverse set of amino-acid-based macromolecules which were termed proteins (from the Greek word *proteios* for "being of the first order") to indicate their ubiquitous presence in all organisms[1]. The construction plans for all the proteins of an organism are stored in another biological macromolecule, deoxyribonucleic acid (DNA), that is made up of two different purine-pyrimidine base pairs chained together by sugar-phosphate bonds. The four bases adenosine, cytosine, guanine and thymine (A, C, G, T) form the genetic code that determines the order of amino acids for every protein a cell can make[2].

Not all of the proteins are needed in the same amount or at the same time as many proteins fulfill highly specialized functions, for example in response to environmental stress or developmental cues. Moreover, multi-cellular organisms contain cells with drastically different appearances and functions even though they all share the same genetic information. This wide range of distinct spatio-temporal phenotypes is based on the ability of the individual cells to read (express) those parts of the DNA that they need and to regulate the outcome quantitatively. As shown in Figure 2.1, gene expression is a multi-step process that includes the generation of a ribonucleic acid (RNA) copy, its processing, transport and subsequent translation into a primary amino acid chain that will eventually give rise to a protein. Each step can be influenced and modulated by regulatory mechanisms, but recent reports indicate that the beginning – transcription – is the major means for determining the final protein levels (it is estimated that up to 80% of differences in individual protein abundance are explained solely by differences in transcription[5]).

2.1 Transcription in eukaryotes

Transcription itself is carried out by large protein complexes, most notably RNA Polymerase II (Pol II) that binds to the beginning of genes (promoters) and subsequently produces an RNA copy of the gene. In eukaryotes, i.e. organisms whose cells contain several membrane-enclosed organelles such as the DNA-containing nucleus[6], Pol II does not have

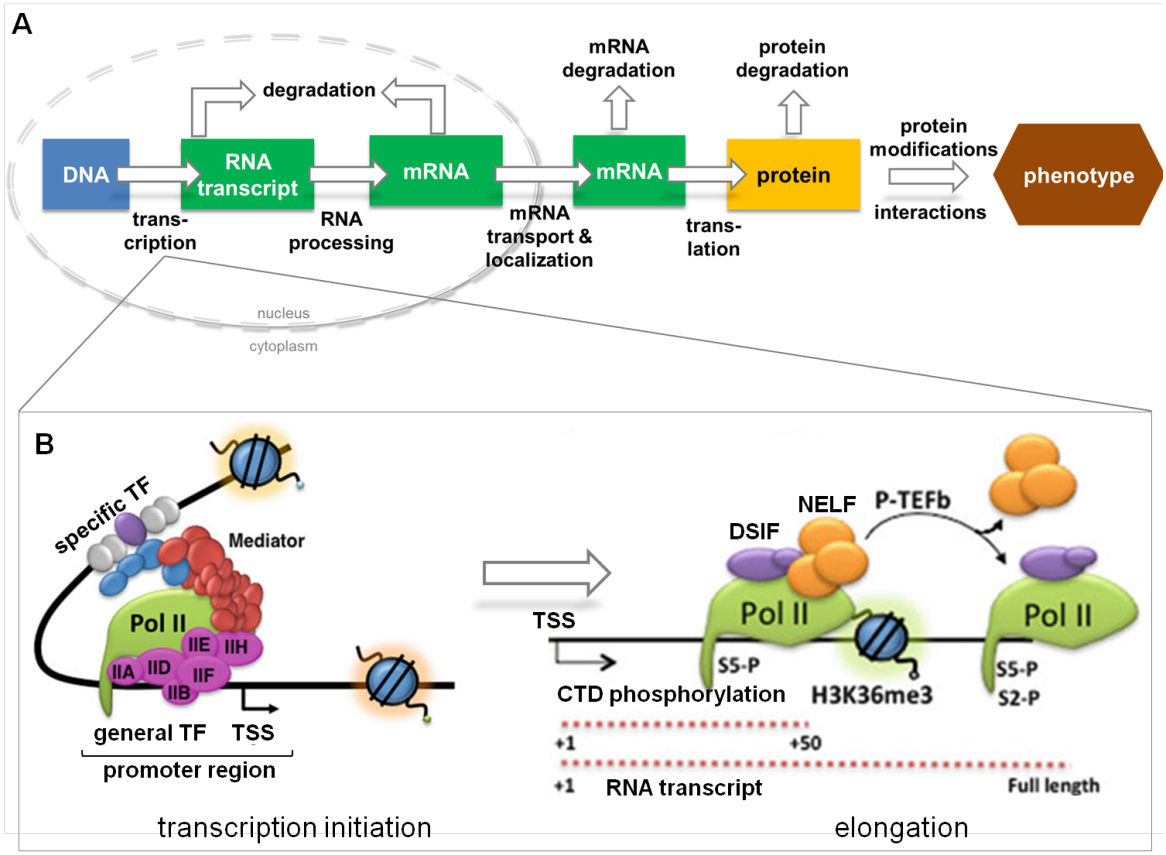


Figure 2.1: Gene expression – the process of generating a protein based on the genetic code – is a multi-step process. **A)** Proteins are the main effector molecules for the phenotype of a cell. DNA-encoded genes are copied into RNA that needs to be processed and exported out of the nucleus into the cytoplasm where the translation of the nucleic acid code into the primary amino acid sequence of a protein takes place. The concept of this image was taken from Alberts et al.[3] **B)** In eukaryotes, the copying of protein-encoding genes into RNA templates is performed by the RNA Polymerase complex II (Pol II) whose action depends on manifold additional proteins. Specific transcription factors (TFs) bind to regulatory DNA regions close to the transcription start site (TSS, promoters) or further away (enhancers) where they i.a. initiate the remodelling of the chromatin to allow for Pol II recruitment. The Mediator complex (shown in red) connects the TFs' activation stimulus with Pol II by direct physical interactions and it is essential for the assembly of the pre-initiation complex that entails Pol II and general transcription factors (GTF, shown in pink). Subsequently, the DNA double-helix is unwound and RNA polymerization is initiated, marked by the phosphorylation of serine 5 (S5-P) of Pol II's C-terminal repeat domain (CTD). For the entire gene's transcription to occur, elongation factors such as the positive transcription elongation factor b (P-TEFb) must replace the GTFs to release the negative elongation factor (NELF) and stimulate phosphorylation of serine 2 (S2-P) of the CTD. The two parts of this image and information conveyed here were taken from Barrero and Malik[4].

high intrinsic affinity for DNA binding. Therefore, protein-protein interactions that promote and stabilize the Pol II-DNA interaction are essential for eukaryotic transcription[7]. Transcription regulation by so-called transcription factors can occur via binding to promoters or TSS-distal regulatory elements (enhancers) that contribute to the initiation of gene expression (see Figure 2.1 for details). Furthermore, transcription factors themselves can be regulated through various means, e.g. by interactions with other proteins and small molecules,

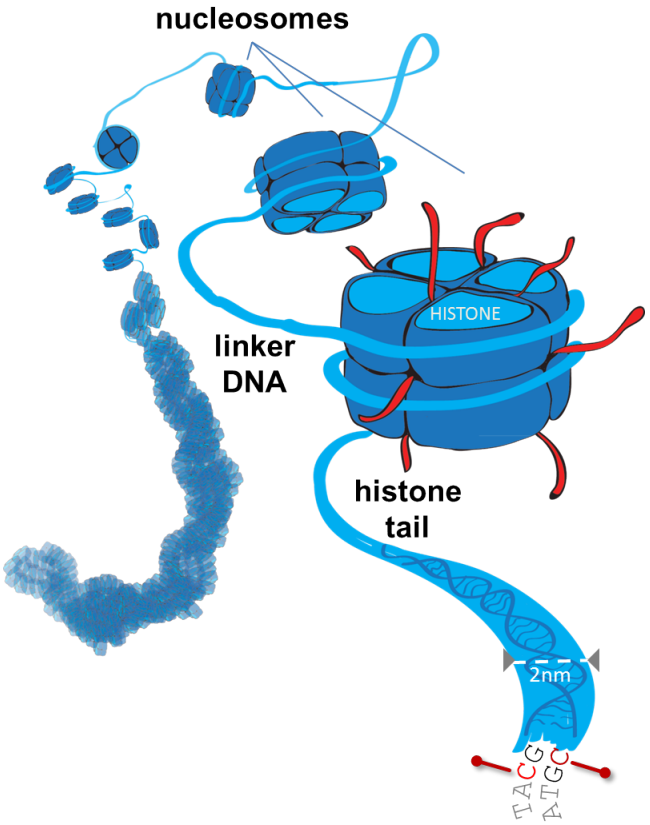


Figure 2.2: In eukaryotic cells, the DNA strand is wrapped around so-called nucleosomes – octamers of the core histone proteins H2A, H2B, H3, H4. The linker DNA between nucleosomes can vary in length (7-100 bp) and is associated with histone H1 (not shown). Nucleosomes are transcription inhibitors as they restrict access to genes, but they can be influenced by physical force and post-translational modifications. The red strands indicate the N-terminal ends of the histones that protrude from the nucleosome and serve as substrates for myriad modifications[9] (Figure 2.3). An additional epigenetic modification that does not alter the DNA code, but influences gene expression, is methylation of cytosines of CpG dinucleotides (red marker on the bottom) that is generally associated with silenced genes. The original image was taken from[10] and modified by Tomasz Chelmicki and myself.

post-translational modifications or by changes of their own expression and degradation.

2.1.1 Chromatin and transcription

In recent years, an additional layer of gene expression regulation has emerged: the packaging of the DNA. In the nuclei of eukaryotes, the DNA strand is generally not readily accessible as it is stored within a compact DNA-protein polymer, termed chromatin. Chromatin is made up of nucleosomes that are composed of protein octamers of four basic, evolutionarily extremely well conserved histone proteins with DNA wrapped around them (approximately 150 nucleotides per nucleosome, Figure 2.2). This enables tremendous compaction of the long DNA strand[8], but it also offers more possibilities for gene expression regulation because nucleosomes sterically hinder transcription and need to be moved before transcription can occur. In fact, chromatin has emerged as a well-suited template for both dynamic, gene-specific effects as well as stable, continuously propagated changes of transcription that persist without changes of the DNA sequence (herein referred to as epigenetics).

The epigenetic influence on gene expression is based on two main properties of chromatin: its physical structure and the potential for myriad post-translational modifications of the histone residues of their N-terminal tails and globular domains (Figure 2.2). The physical properties of nucleosomes can be altered by chromatin remodellers that apply a torsional strain to

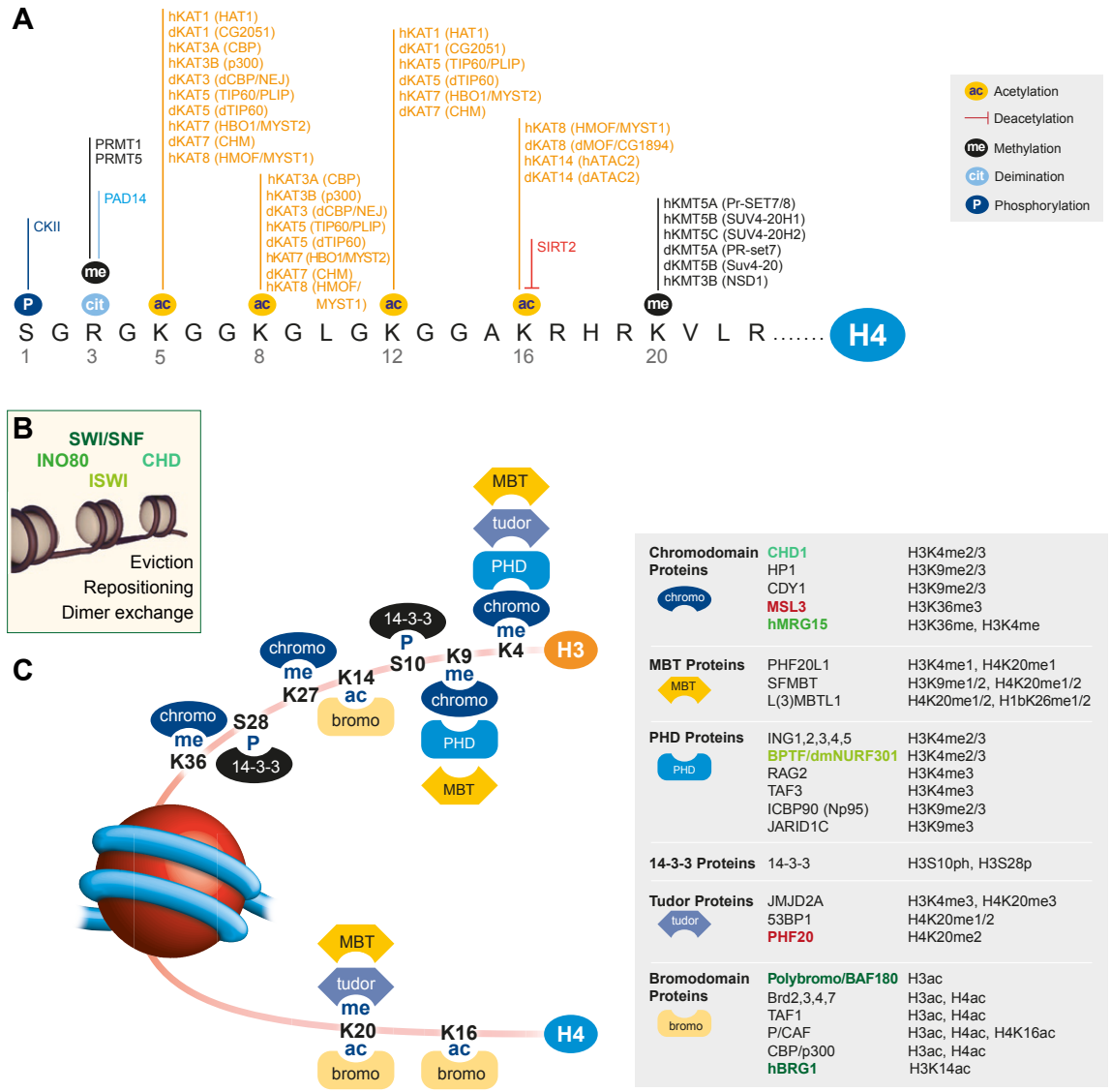


Figure 2.3: Examples of histone modifications, corresponding enzymes and protein domains that recognize them. **A)** Like for the other core histones, the tail of histone 4 (H4) contains numerous residues for covalent post-translational modifications of which some are shown here together with enzymes capable of catalyzing the respective mark in humans (h) and *Drosophila* (d). Note that the different marks may have different biological meanings, e.g. methylation of H4K20 is associated with gene repression while acetylation of H4K16 is associated with gene activation (Figure B.1). Moreover, the residues within the globular domains can also be modified (not shown). CKII = casein kinase II, PRMT = protein arginine methyltransferase, PAD2 = peptidyl arginine deiminiase, KAT = lysine acetyl transferase, KMT = lysine methyl transferase, SIRT2 = NAD-dependent deacetylase sirtuin-2. The image was taken from [11] and modified. See Table 2.1 for details on histone acetyl transferases. **B)** Besides histone-modifying enzymes, chromatin remodellers greatly influence gene accessibility, too. There are four major families of ATP-dependent DNA translocases that can – to varying degrees – reposition, evict and replace nucleosomes. The INO80 family members are particularly important for DNA replication and repair as they can efficiently exchange histones while the multimeric complexes of the SWI/SNF family predominantly slide and evict nucleosomes [12]. **C)** Transcription factors (e.g. the transcription initiation factors TAF1 and 3), chromatin-remodelling complexes (shown in green, see B) and histone-modifying enzymes (e.g. the HATs CDY, p300 and P/CAF and the lysine demethylases JARID1C and JMJD2A) often contain domains that recognize specific histone modifications. Members of MOF-associated complexes are shown in red. The image was taken from [11] and modified.

the DNA strand that generates enough force to change the position of a nucleosome[12]. These ATP-dependent enzymes are required for the maintenance of active as well as silenced genome regions, they can replace the canonical core histones shown in Figure 2.2 with histone variants and they are required for the assembly of chromatin during DNA replication and DNA repair (reviewed by Manelyte and Laengst[12], see Figure 2.3 for the four families of chromatin remodellers).

Numerous histone-modifying enzymes are now known; they catalyze the transfer of small organic groups (e.g. methyl, acetyl and phosphate groups) or small proteins (e.g. ubiquitin) to histone residues, especially to their tail regions (Figure 2.3). The tail domains of the histones are not essential for nucleosome formation, but their modifications contribute to the recruitment of chromatin- and DNA-binding proteins[13] (see Figure 2.3 for histone binding domains) and serve as landmarks of active transcription (euchromatin) as well as silenced regions (heterochromatin). Although more than 100 histone modifications have been described[9,14] and well-defined histone mark combinations are the basis for distinct chromatin states associated with different levels of transcription[15,16] (see Figure B.1), the debate about whether they are cause, consequence or merely a by-product of gene transcription is on-going[17–19]. On the other hand, the impact of chromatin structure on transcription is well established on at least two levels: high nucleosome density interferes with transcription factor binding and gene activation (reviewed by Henikoff and Shilatifard[17]) and higher order chromatin domains separate euchromatin from heterochromatin, presumably to enable efficient transcription and constrain transcriptional inactivation[20]. Proteins that modify histones (e.g. methyl transferases, acetyl transferases) and ATP-dependent nucleosome remodellers have accordingly risen as a new class of transcription-related factors that do not necessarily need to directly interact with DNA to influence gene expression (Figure 2.3).

2.2 The histone acetylase MOF

The histone acetylase (HAT) MOF (males absent on the first) was discovered in a screen for an X-chromosomal factor with essential male-specific function in *Drosophila melanogaster*[22]. HATs catalyze the addition of acetyl groups to histone and non-histone protein residues. Based on the catalytic subunits that show distinct structural features, HATs can be classified into three main families: GCN5-related *N*-acetyl transferases (GNATs) that typically contain bromodomains, the MYST family including MOF that is characterized by chromodomains and an additional group of less conserved proteins with HAT activity (Table 2.1). All HATs strongly depend on the interaction with other proteins to carry out their enzymatic activity in an efficient manner[23]. MOF was originally thought to function predominantly within the male-specific lethal (MSL) complex, more recent studies identified additional interaction partners with essential, sex-independent functions (non-specific lethal

Table 2.1: The main histone acetyl transferase families. The lysine acetyl transferase (KAT) designation indicates the standardized nomenclature suggested for HATs. GNAT = Gcn5-related *N*-acetyl transferase, MYST = MOZ, Ybf2/Sas3, Sas2, Tip60. The table was taken from Berndsen et al. [21]

Enzyme	KAT designation	Histone specificity
GNAT family		
Gcn5	2	H3K9, 14, 36
p/CAF	2B	H3K14
MYST family		
Tip60	5	H4K5, K8, K12 (K16)
MOF	8	H4K5, K8, K16
Sas3	6	H3K14, K23
MOZ	6A	H3K14
p300 and others		
CBP	3A	H2AK5, H2B
p300	3B	H2AK5, H2B
Rtt109	11	H3K56, K9, K23

(NSL) proteins, see Figure 2.5). Members of both complexes are well conserved in mammals, but particularly MOF has attracted considerable biomedical research interest as global loss of H4K16ac was identified as a hallmark of numerous cancers with corresponding misregulation of MOF [24,25] (see Table B.2). Deletion of *Mof* in mice leads to embryonic lethality with profound alterations of the nuclear architecture and DNA damage responses [26,27] and mammalian MOF has been described in a variety of cellular contexts, ranging from spermatogenesis to the protection from myocardial hypertrophy in models of heart stress (Figure 2.4). The underlying mechanisms of these diverse functions have remained largely uncharacterized, but in regard to its role in gene regulation and DNA repair, significant insights have been gained, particularly in the context of its interaction partners (Figure 2.6, Table 2.2, Table 2.3).

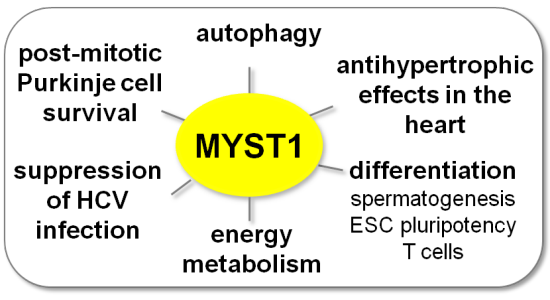


Figure 2.4: The mammalian orthologue of *Drosophila* MOF, MYST1, was shown to be required for longevity of Purkinje cells [28] and the suppression of hepatitis virus C (HCV) infections [29]. It also regulates autophagy [30], protects stressed heart cells from hypertrophy [31] and its depletion from the hypothalamus leads to diet-induced obesity [32]. Additionally, its action is needed for the replacement of histones by protamines during spermatogenesis [33–35] and MYST1 depletion blocks T cell development [36]. Conversely, MYST1 was recently implied in maintaining ESC pluripotency [37].

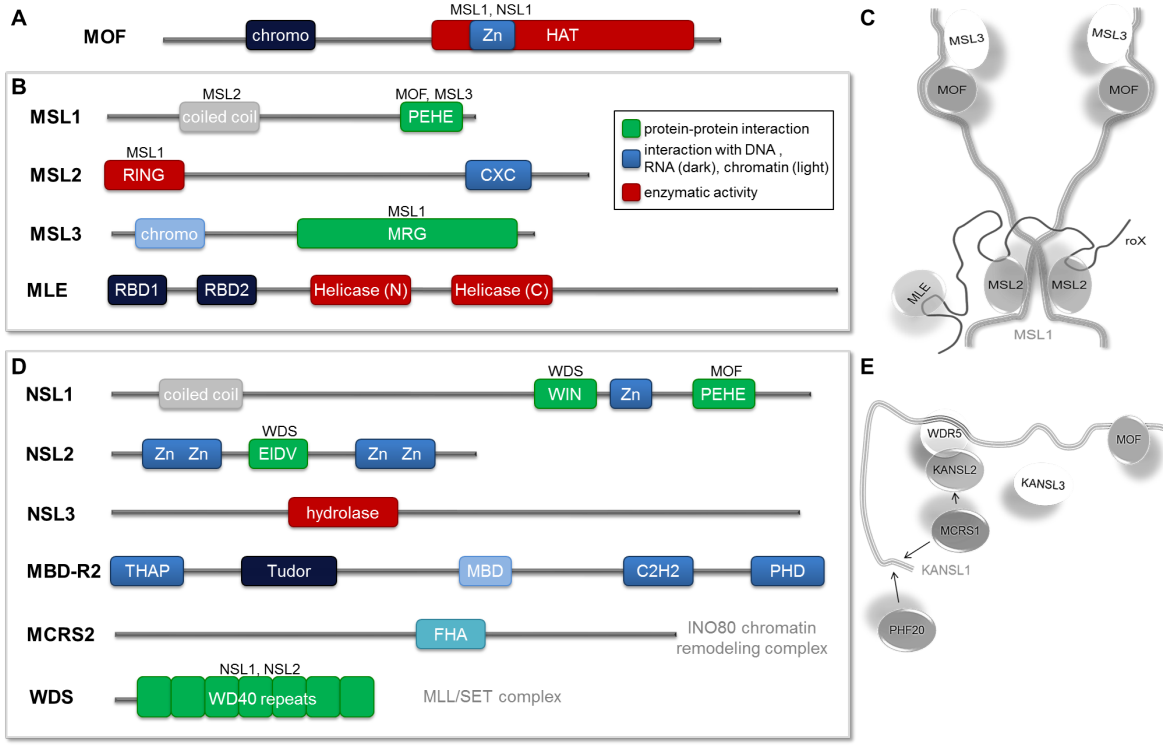


Figure 2.5: Protein domains and interactions of the MSL and NSL complexes. **A)** MOF contains a chromo-barrel domain that can bind DNA and RNA and is essential for histone acetylation[38]. The zinc finger (Zn) motif within the catalytic HAT domain interacts with MSL1, MSL3 and NSL1[39,40]. **B)** MSL complex members: The PEHE domain is specific for MSL1-like proteins[41]. The RING finger domain of MSL2 is required for ubiquitylation and protein-protein interactions[42,43]. MSL3's MRG domain is necessary for the targeting of the male X[44] and preferably binds methylated lysine residues[45] (Figure 2.3). MLE contains N-terminal and C-terminal helicase domains with RNA and DNA remodelling activity[46]. **C)** MSL1 is a largely unstructured scaffold protein that dimerizes with MSL2 and subsequently brings together MSL3 and MOF that both interact with MSL1's PEHE domain[40,47]. *roX* RNA is bound by MSL2 and MLE, completing the complex[48]. **D)** NSL complex members: NSL1 resembles MSL1, but lacks the region C-terminally of PEHE for the interaction with MSL3[49]; instead, like NSL2, it contains an additional WDS-interacting motif (WIN and EIDV, respectively)[50]. MBD-R2 contains multiple chromatin and nucleic acid interaction motifs; the forkhead-associated domain (FHA) of MCRS2 was shown to bind phosphorylated protein residues and was identified as part of the INO80 complex. WDS is a small scaffold protein made up of seven WD40 repeats that is also part of the histone methyl transferase MLL/SET complex[51]. **E)** Analogous to the MOF-MSL1 interaction, MOF is recruited into the NSL complex via an interaction with the C-terminal PEHE domain of NSL1 that also interacts with MCRS2, MBD-R2 and WDS[40,50,52]. WDS directly interacts with NSL1 and NSL2[50]. All images are based on figures and data from[40,50,52–54].

2.2.1 The male-specific lethal (MSL) complex and dosage compensation in *Drosophila*

The first known MOF partners were part of the MSL complex which has been well studied in *D. melanogaster* where its deletion leads to male-specific lethality. Like mammals, male and female fruit flies differ in respect to the number of X chromosomes. In contrast to mammals where the inactivation of one of the two female X chromosomes extends the problem of X monosomy to both sexes[55], *D. melanogaster* has evolved a chromosome-wide transcriptional

upregulation of the single male X chromosome (shown first in 1965 by Mukherjee and Beermann[56]) that equalizes the doses of X-linked and autosomal genes in the heterogametic sex. Owing to its male-specific effects on viability and an extraordinary enrichment along the male fly X chromosome, the MSL complex was proposed as they key component of dosage compensation within *Sophophora* species[57] (reviewed by Laverty et al.[58]).

The exact molecular mechanisms of MSL's X-specific targeting are not completely elucidated up to date, but many details about the assembly of the complex are now known (see Figure 2.5). The MSL complex is a ribonucleoprotein complex composed of five proteins (MOF, MSL1, MSL2, MSL3 (male-specific lethal 1-3), MLE (maleless)) and one long non-coding RNA (roX1 or roX2 (RNA on X), see Figure 2.5 for details and Table B.1 for synonyms and mammalian protein names). In flies with X:A ratios of one, SXL (sex lethal) inhibits the translation of MSL2 by recruiting the RNA-binding protein UNR (upstream of neuroblastoma rat sarcoma) to the 3'-untranslated region (UTR) of the *msl2* transcript[59–61]. Lack of MSL2 suffices to inhibit the formation of the MSL complex in females although all other MSL proteins can in principle be transcribed and translated in both sexes[62,63]. While the recognition of the X chromosome is achieved by the MSL1/MSL2 heterodimer[40,64], accumulation along the chromosome and transcription stimulation depend on the presence of *all* MSL proteins (reviewed by Conrad and Akhtar[65]).

Within the MSL complex, the HAT enzyme MOF has been regarded as the key effector molecule for dosage compensation as the approximately twofold upregulation of X-linked genes in male *Drosophila* seems linked to the long-known hyperacetylation of the male X chromosome[66,67]. It was shown that MOF has an extraordinary preference to acetylate lysine 16 of histone H4 (H4K16ac) if accompanied by the MSL complex[68–71] which is in stark contrast to the rather promiscuous acetylation of various histone residues by other known HATs[23] (Table 2.1). The MSL complex, in turn, is responsible for excessive MOF recruitment to the male X chromosome and the subsequent enrichment of H4K16ac along gene bodies[66,72] that is both stimulated and fine-tuned by the interaction of MOF with MSL complex members[39,40,73].

Since the MSL complex is absent in female flies, it is unlikely that its presence is required for initial activation of X-linked gene transcription; it is instead thought to recognize already expressed genes that are localized on the X chromosome. The current model delineates that the MSL complex is recruited to several high-affinity binding sites on the X chromosome from where it spreads to active genes (i.a. aided by the binding of MSL3 to methylated H3K36, an abundant mark of on-going transcription[74] (Figure B.1)) and subsequently enhances the basal transcription rates by acetylation of H4K16[65].

2.2.2 The effects of H4K16ac on chromatin structure

Acetylation of H3 and H4 tails in general is associated with increased chromatin accessibility as the negative charges of the acetyl group interfere with tight packaging of the poly-anionic DNA strand. In contrast to the acetylation of lysines 5, 8, and 12 of histone H4 that seem to affect the chromatin structure and transcription in a non-specific, cumulative manner, H4K16ac conveys more specific transcriptional effects[75]. One explanation for H4K16's special function is its key role in anchoring neighboring nucleosomes where unmodified H4K16 is required for the formation of strong salt bridges between the H4 tail and the acidic patch on the next histone octamer (reviewed by Kalashnikova et al.[76], Preez and Patterton[77]). Acetylation of H4K16 disrupts these salt bridges and inhibits chromatin compaction which is in line with the inhibition of chromatin fiber formation that was reported for H4K16ac specifically[78,79].

2.2.3 The effects of H4K16ac on transcription

With the exception of yeast, H4K16ac has unanimously been associated with active transcription as open chromatin in general is a key determinant of gene expression (Henikoff and Shilatifard[17], Lee and Workman[23]). In the context of the MSL complex and *Drosophila* dosage compensation, three possible entry points for gene expression enhancement by massive H4K16ac have been proposed:

- **Transcription initiation:** The high accessibility of promoter regions and regulatory elements allows transcription factors to find their cognate target sites more efficiently[80] and might increase the probability of pre-initiation complex assembly[81], thereby elevating local concentrations of regulatory proteins and the transcription machinery itself.
- **Pol II pause release:** In mammals, H4K16ac was shown to recruit Brd4 (Figure 2.3) which in turn recruits P-TEFb, an essential factor for the transition of promoter-bound Pol II into its actively processing state[82] (Figure 2.1). H4K16ac might thus serve as an additional binding site for transcription-related proteins.
- **Elongation of transcription:** Based on the observation that H4K16ac generally covers the entire gene body of dosage-compensated genes with increasing signals towards the 3'-end, it has been suggested that the main effects of H4K16ac for dosage compensation should be on transcription elongation[83]. This is supported by a recent study reporting Pol II to reach the 3'-end of genes more easily on the X chromosome compared to autosomes which was dependent on the presence of MSL2[84].

The fact that H4K16ac stimulates transcription, perhaps via different mechanisms, is not contested[65,85]. However, how the approximately twofold upregulation of male single copy

X-linked genes to the levels of autosomal genes[86,87] is achieved, has not been solved yet. Given the multiple chromatin, protein, DNA and RNA interaction as well as catalytic domains present in the MSL complex (Figure 2.5, Table B.3), it is likely that its effects on transcription are manifold and reach beyond H4K16ac to ensure the modest, but vitally important transcription enhancement of male X-linked genes.

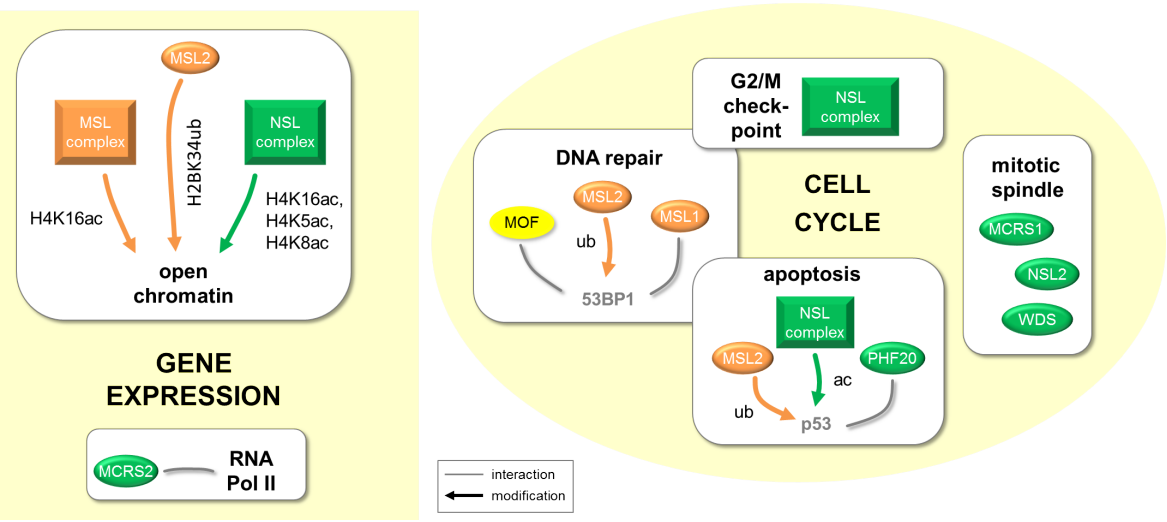


Figure 2.6: The vast majority of functions associated with MOF and its interaction partners are related to general transcriptional activation and the cell cycle. Displayed here are those functions that have either been reported to involve MOF within the context of the NSL and MSL complex or individual complex members. Not shown are the roles of WDR5 (WDS) and MCRS1 (MCRS2) as part of the methyl lysine transferase MLL/SET complex and the nucleosome remodelling INO80 complex, respectively. Grey lines indicate that proteins were found to physically interact, colored arrows indicate post-translational modifications by either MOF (acetylation (ac)) or MSL2 (ubiquitylation (ub)). The details of the function are summarized in Table 2.2 and Table 2.3.

Table 2.2: MOF-associated proteins in transcription activation. Related to Figure 2.6.

Proteins	Function	Biological effect
MOF within the MSL complex (MSL1, MSL2, MSL3, MLE (roX))	acetylation of H4K16ac[68,70,88]	• opening of chromatin (reviewed by Preez and Patterton[77]) • dosage compensation in male *D. melanogaster* that entails the upregulation of the entire X chromosome[65]
MOF within the NSL complex (NSL1, NSL2, NSL3, MCRS2, MBD-R2, WDS)	acetylation of H4K16[70], H4K5, H4K8[71]	• opening of chromatin[77] • housekeeping gene regulation[89,90]
MSL2	ubiquitylation of H2BK34[91]	implicated to stimulate methylation of H3K4[91] and transcription elongation[92]
MCRS2	interaction partner	facilitates Pol II recruitment to target genes[93]

Table 2.3: MOF-associated proteins in cell-cycle-related processes. Related to Figure 2.6

Biological process	Observation
G2/M checkpoint	NSL1, NSL2, NSL3, MCRS2, MBD-R2 and WDS were identified as essential factors for G2/M checkpoint progression following DNA damage in *D. melanogaster*[94]
mitotic spindle	• NSL2 is needed for mitotic spindle assembly[95] • MCRS1 stabilizes the mitotic spindle[96] • WDS was identified in a screen for microtubule-associated proteins[97]
apoptosis	• ubiquitylation of p53 by MSL2 leads to accumulation of p53 in the cytoplasm[43] which is necessary for apoptosis[98] • MOF acetylates p53 in the presence of NSL1[70] which is necessary for apoptosis induction in cells with DNA damages[99,100] • human MBD-R2 (PHF20) stimulates expression of p53 and prevents its degradation via a direct interaction with methylated p53[101,102]
DNA repair	• MOF is generally required for repair of DNA double strand breaks and recruitment of 53BP1 and BRCA[103,104]; its phosphorylated form is particularly important for biasing the cells towards homologous repair during S phase by displacing 53BP1 from the site of the DNA damage[105] • human MSL2 ubiquitylates 53BP1[106] • human MSL1 interacts with 53BP1 that positively stimulates DNA damage repair[107]

2.2.4 Individual functions of MSL1 and MSL2

Evidence for activities of the individual MSL complex members besides MOF predominantly stems from studies on the mammalian orthologues that exist in a sex-independent manner without apparent preference for a particular chromosome (Table 2.2 and 2.3). While this seems to strip the MSL complex in mammals of its exceptional and highly visible role for

dosage compensation, it also opens up a large scope of putative chromatin-related functions that might not depend on MOF. The ubiquitylation of lysine 34 of histone 2B (H2BK34ub) by MSL2, for example, is stimulated by the presence of MSL1, but not MOF[91]. Moreover, MSL1 and MSL2 were reported to directly interact with P-TEFb, thereby aiding the transition of Pol II to elongation[92] (see Figure 2.1).

In addition to direct effects on transcription-related processes, MSL1 and MSL2 have been implied in DNA repair and apoptosis, primarily due to their interaction with the tumor suppressor p53 and the p53-binding protein (53BP1)[43,104,106] (see Figure 2.6 and Table 2.3 for details). These tasks do not necessarily exclude the possibility for MOF interactions as histone acetylation and general chromatin relaxation seem to stimulate the DNA repair pathways[108], but these findings indicate that, like MOF, MSL proteins might also act outside the MSL complex context, especially in non-sophophora species.

2.2.5 The non-specific lethal complex and additional MOF targets

Using mass spectrometry, Mendjan et al.[53] identified additional, sex-unspecific interaction partners of MOF in fly and human cell cultures which were termed non-specific lethal (NSL) proteins owing to the fact that mutations in their genes killed flies of both sexes.

The NSL complex is comprised of seven proteins: MOF, NSL1, NSL2, NSL3 (non-specific lethal 1-3), MCRS2 (microspherule protein), MBD-R2 (methyl-binding domain protein), and WDS (will die slowly; see Table B.1 for synonyms and mammalian protein names). At the beginning of my PhD studies, little was known about the cellular task conveyed by the NSL complex, but the first genome-wide study by Raja et al.[52] had laid the foundation for establishing the NSL complex as a ubiquitous interaction partner that could explain the MSL-independent binding that had been observed for MOF on male as well as female autosomes[52,63,109]. Complementary, biochemical experiments revealed that the interaction with NSL proteins relaxes MOF's substrate specificity towards additional histone residues (H4K5, H4K8)[71]. Furthermore, MOF was reported to be capable of acetylating non-histone targets which predominantly entail proteins that guard the integrity of the genome, such as ATM (Ataxia Telangiectasia mutated), p53, DBC1 (deleted in breast cancer) and NRF2 (nuclear respiratory factor)[80,99,110,111]. Li et al.[70] demonstrated that the interaction with NSL1, but not MSL1, allowed MOF to efficiently acetylate p53 which is important for the induction of apoptosis in cells suffering DNA damage[99,100]. This finding corroborated the notion that the NSL complex might extend the impact of MOF beyond dosage compensation.

2.2.6 Individual functions of NSL complex members

In line with MOF's acetylation of DNA damage response proteins, it is perhaps more than an interesting side note that all NSL complex members were shown to be essential for the

G2/M checkpoint (supplemental material of Kondo and Perrimon[94]) and individual factors have been described in additional cell-cycle- and DNA-repair-related functions (see Figure 2.6, Table 2.3). Besides genome-wide association studies that have implicated mutations in the mammalian orthologues of *Nsl1* and *Nsl2* in different syndromes associated with intellectual disabilities[112–115], NSL1-3 have remained largely uncharacterized. The mammalian counterparts of MCRS2 and WDS (MCRS1/MSP58 and WDR5, Table B.1) have raised more research interest since they were found within additional multimeric chromatin complexes: MCRS1 was shown to interact with the chromatin remodellers INO80, NuRD and SWI/SNF[116–118] (Figure 2.3), while WDR5 is an essential part of the histone methyl transferase MLL[51] (mixed-lineage leukemia) that is responsible for trimethylation of H3K4, a prominent histone mark of active gene promoters (Figure B.1). Furthermore, MCRS1 has been identified as an oncogene and a negative regulator of human telomerase reverse transcriptase (hTERT), directly suggesting a substantial influence on cellular senescence signaling[118,119] (see Table B.2 for cancer-related observations for the individual proteins).

As mentioned previously, most studies reporting functions of the (mammalian) NSL complex members examined the proteins in either isolated or different interaction contexts; insights into the functions of the NSL complex as an entity were lacking at the beginning of my PhD studies. One of my major aims therefore was to investigate the chromatin-associated functions of the NSL complex in *Drosophila* and mouse.

2.3 ChIP-seq for the study of transcription factor binding

The current state-of-the-art method to examine the binding profiles of DNA- and chromatin-interacting proteins *in vivo* is chromatin immunoprecipitation (ChIP) followed by high-throughput NDA sequencing (seq). The endorsement of ChIP-seq as the major means to study transcription factors and histone modifications in a genome-wide fashion was made possible by the steep decline of costs for DNA sequencing during the past decade which was driven by the development of massively parallel sequencing that, in contrast to the traditional Sanger sequencing, yields rather short (35-100 bp), but highly abundant DNA reads.

2.3.1 Chromatin immunoprecipitation

To identify regions of the genome bound by a protein of interest (or marked by a histone modification), the protein-chromatin interactions are usually fixed with formaldehyde before the chromatin is fragmented into pieces of 200-1,000 bp. Then, an antibody against the protein of interest is used to precipitate those fragments to which the protein is bound. After de-crosslinking, the DNA can be purified and sequenced (Table 2.4). Numerous alterations and variations to the basic ChIP protocol exist, such as the use of micrococcal nuclease to

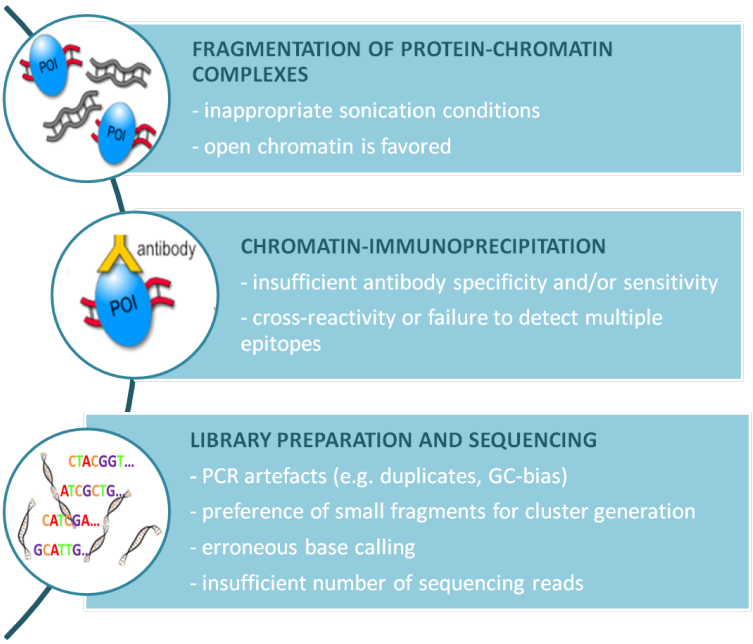


Figure 2.7: Technical issues during ChIP-seq experiments that can interfere with the bioinformatic analysis can occur during every step of the sample preparation. The understanding of the biases introduced by the Illumina sequencing platform has increased profoundly during the past years, but the influence of the fixation and sonication procedures are much less elucidated [120,121]. See Table 2.5 for details of additional biases. The upper two illustrations were taken from [122].

fragment the DNA by digestion rather than sonication [123] and the omission of the formaldehyde fixation (native ChIP) [124]. The resolution of the binding sites can be increased up to the single base level by applying an exonuclease digestion step after the ChIP (ChIP-exo) [125]. Furthermore, Nano-ChIP and linear and *in vitro* transcription (LinDA) have been suggested to enable ChIP-seq with very low cell numbers [126,127].

2.3.2 High-throughput sequencing (Illumina platforms)

Illumina's high-throughput sequencing method requires short DNA fragments that are eventually hybridized to a sophisticated glass slide, the flow cell (see details in Table 2.4). To ensure a signal that will be unambiguously detected, each fragment is massively and clonally amplified using solid-phase PCR to generate clusters of identical molecules. The sequencing of the fragment ends (35–100 bp) itself is based on fluorophore-labelled dNTPs with reversible terminator elements that will become incorporated and excited by a laser one at a time and thereby enable the identification of single bases [128]. For mammalian genomes, it is recommended to sequence at least 20-60 million DNA fragments, depending on the biological question and the nature of the expected signal [129–131].

Table 2.4: The different (non-ChIP-seq-specific) steps of high-throughput DNA sequencing with Illumina platforms.

Library Preparation: Obtain all sequences of interest flanked by adapters	1. DNA fragmentation (150-300 bp) 2. DNA end repair 3. addition of adenosine overhangs 4. adapter ligation 5. purification
Cluster Generation: Local amplification of each sequence of interest	1. hybridization of the adapters of the DNA fragments to the oligonucleotides on the flow cell 2. extension of the hybridized fragments 3. bridge amplification for generation of local clusters of identical DNA fragments 4. flow cell preparation
Sequencing by synthesis: Copying the hybridized DNA fragments	1. 35-100 cycles of: • incorporation of one fluorophore-labelled dNTP containing a reversible terminator • laser excitation of the fluorophore • detection of each cluster's fluorescence • removal of the fluorophore and terminator element from the incorporated dNTP 2. base calling through analysis of the emitted fluorescence spectra

2.3.3 Limitations of ChIP-seq

Eventhough it is the method of choice for genome-wide transcription factor binding and histone mark profiling, ChIP-seq protocols are prone to biases and artifacts and must therefore be highly optimized (Figure 2.7, Table 2.5). There are four factors that determine the success of a ChIP-seq experiment[129,132]: the antibody (which must be rigorously tested to exclude cross-reactivity and ensure specificity and sensitivity[130]), the chromatin extraction (that tends to overrepresent highly transcribed regions[120,133]), the library preparation and sequencing (which can introduce several biases) and the bioinformatic analysis that must be tailored to match the data set's characteristics in order to reveal biologically meaningful insights. It should be noted that all ChIP-seq experiments to date reflect the protein-DNA interactions of a population of cells which suggests that particularly strong signals represent binding sites where the protein of interest is found in the majority of cells.

Table 2.5: Biases and artifacts of ChIP-seq data. Given a rigorously tested antibody, ChIP-seq still suffers from additional technical problems that are due to the sequencing process as well as bioinformatic hurdles.

Problem	Reasons	Solutions
Chromatin context and **transcription**	• euchromatic chromatin is more easily fragmented • formaldehyde fixation does not capture short-lived protein-DNA interactions[121] and preferentially crosslinks proteins with each other[124] • chromatin extraction may unspecifically enrich for highly active genes[120]	• cell-type- and condition-specific input controls[130,134] • if possible, avoid crosslinking[123,124] • optimized chromatin extraction including extensive de-crosslinking, RNase and Proteinase treatments[133] • to identify hyper-ChIPable regions, ChIP against a non-endogenous protein was suggested[120] • immunoprecipitation with immunoglobulin G (IgG, mock IP)[130,135]
Sequencing errors and errors in base calling	• imperfect sequencing chemistry and signal detection • loss of synchronized base-incorporation into the single molecules within one cluster of clonally amplified DNA fragments (phasing and prephasing) (see Table 2.4) • signal intensity decay	• improvement of the sequencing chemistry and detection • optimized software for base calling[136] • computational removal of bases with low base calling scores[137]
GC bias and **duplicate reads**	• GC-rich regions are preferably amplified by PCR • small fragments are preferably hybridized to the flow cell • low number of founder DNA fragments	• optimizing cross-linking, sonication, and the ChIP protocol to ensure that the majority of the genome is present in the sample • limiting PCR cycles during library preparation to a minimum • computational correction for GC content[138,139] and elimination of reads from identical DNA fragments
Copy number variations and mappability	• incomplete genome assemblies • strain-specific differences to the reference assembly may lead to misrepresentation of individual loci • repetitiveness of genomes and shortness of sequencing reads hinder unique read alignment	• increased sequencing depth and control (non-ChIP) sample aid the computational identification of problematic loci[129–131,140,141] • longer sequencing reads • paired-end sequencing[129,140] • exclusion of blacklisted regions that are known to attract artificially high read numbers[142,143] • computational correction for mappability[138] • considering the *effective* genome size[144]

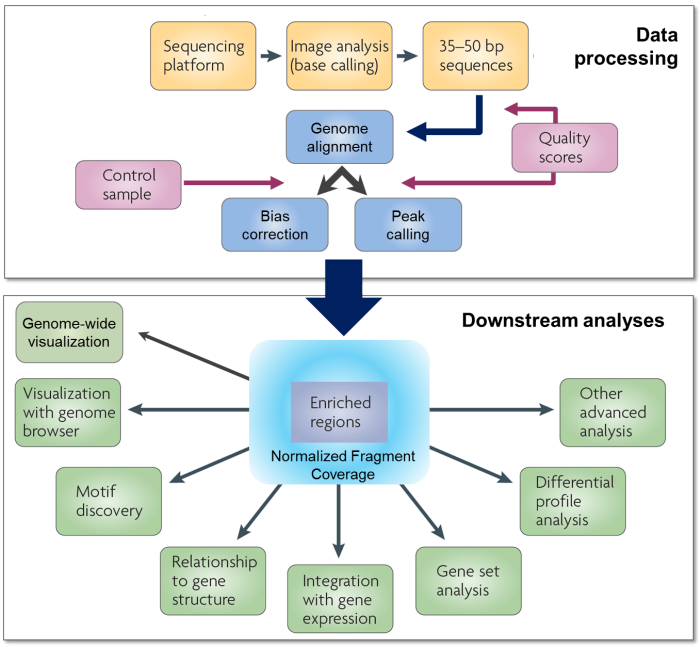


Figure 2.8: Overview of typical computational steps following the completion of high-throughput sequencing. The short DNA reads representing the ends of the DNA fragments hybridized onto the sequencer's flow cell are generated by the vendor-supplied software and first need to be aligned to a reference genome. Identification of significantly enriched binding sites (peak calling) and normalized coverage files are the basis for the vast majority of commonly applied ChIP-seq downstream analyses. For more details, see Table 2.5 and the text. I have significantly modified and complemented the original scheme taken from Park [101].

2.3.4 ChIP-seq data processing

The bioinformatic analysis of ChIP-seq data consists of numerous steps and only the very first tasks are widely standardized: The initial conversion of images of fluorescence into intensity files and ultimately text files that contain the sequencing information is typically done in an automated fashion with vendor-supplied software [136] (yellow boxes in Figure 2.7). The resulting file contains all available information for each DNA read, such as the sequence (represented as a string of A, T, G, C), the read ID (referring to the location of the fragment cluster on the flow cell) and quality scores for every base. Subsequent processing of the sequencing data is aggravated by the large size of the files that require powerful computational infrastructure and efficient handling and manipulation with non-mainstream software tools. UNIX-based operating systems that have traditionally been employed for scientific data analyses provide numerous commands and utilities that tend to perform very specific tasks and are commonly executed through a text-based command interpreter, the UNIX shell. In addition, various programming languages can be used either within the pipelines or for stand-alone scripts (e.g. shell scripts, awk, sed, perl, python, R). Thus, bioinformaticians constantly struggle to find a balance between highly specialized, often improvised solutions and standardized, less flexible programs while trying to ensure reproducibility and transparency through myriad rounds of iteratively adjusted analyses.

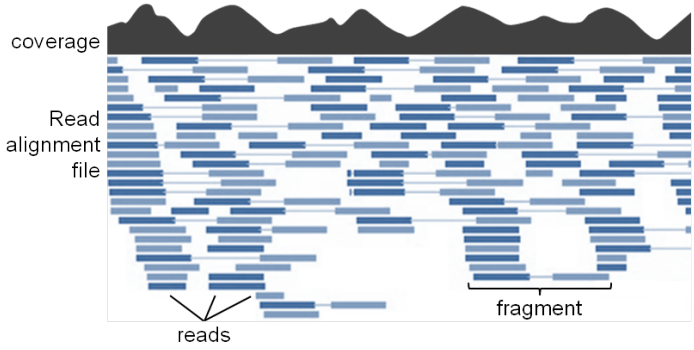


Figure 2.9: Snapshot of a typical visualization of DNA read and coverage files. Shown here are paired-end reads, i.e. both ends of each DNA fragment were sequenced (indicated by dark and light blue boxes) and can now be used to easily reconstruct the precise fragments. The coverage shown on top is based on the number of overlapping fragments.

Genome alignment

The first task of any ChIP-seq analysis is to identify the genomic locus of origin for each DNA read. Due to the large number of reads (several millions) and the genome they originated from (almost three giga base pairs for humans), so-called mapping programs must balance accuracy, speed, computational memory usage and flexibility[145]. Various mapping algorithms exist[146], the two most commonly used programs are bowtie2[147] and BWA[148].

All programs offer manifold options to tune the alignment process to be either faster or more sensitive. Moreover, users can decide how to deal with ambiguous alignments (a read might match more than one region in the genome), gaps and insertions, individual base qualities within a read and mismatches[147,149]. ChIP-seq data does not strongly depend on the exact DNA sequence of every individual DNA read because the final readout is the number of reads overlapping at particular genomic loci compared to background regions (Figure 2.9), which is why mapping results are often accepted with 3-5% of mismatched bases per read[150]. Conversely, ChIP-seq samples can suffer severely from biases that affect the distribution of background reads across the genome which must be assessed and ultimately accounted for[132,138,139] (see Table 2.5).

Control sample

Some biases associated with ChIP-seq data are caused by the genome structure, the chromatin context as well as bioinformatic processing that vitally depends on the state of the genome annotation (Table 2.5). These systematic errors are generally thought to be controlled by the use of a matching input sample[101,129,130,141,144,151,152], i.e. a sample that underwent the same treatment as the ChIP-seq sample (in regard to the cell culture, fixation, lysis, fragmentation (Figure 2.7)) with the exception of the immunoprecipitation step. It is recommended to generate a control for every chromatin preparation, for each cell type and condition and to sequence the control at least as deeply as (or preferably deeper than) the ChIP sample[130,140,152,153]. The goal of the input sample is to have a comprehensive

representation of the background reads that should allow the assessment of under- and over-represented regions due to biological factors (e.g. heterochromatic regions)[134] as well as technical reasons (e.g. unmappable regions and Ultra High Signal regions[143]).

To account for biases introduced by the chromatin precipitation (Figure 2.7, Table 2.5), some researchers prefer to perform a mock immunoprecipitation (IP) with a non-specific immunoglobulin instead of an input sample[130]. Unfortunately, the DNA recovery of mock IP experiments can be exceedingly small which could lead to overrepresentation of sequencing artifacts (Table 2.5). Moreover, it was shown that the use of histone H3 or H4 ChIP-seq does not significantly improve the analysis of ChIP-seq data for histone marks compared to the above mentioned input[154].

Quality controls and metrics

In the past two years, several measures for the assessment of ChIP-seq data quality have been proposed, most notably by the ENCODE (encyclopedia of DNA elements) consortium that generated hundreds of now publicly available ChIP-seq data sets[155]. The quality checks are applied at all stages of the data processing, starting with the determination of contaminant sequences in the raw read files, followed by the quantification of uniquely and non-redundant reads after read alignment and possible overrepresentation of GC-rich regions (Table 2.5). Reads with more than one optimal alignment locus in the genome are usually filtered out in most ChIP-seq analyses as they might increase the risk of false positive binding site detections. However, if the protein of interest is expected to be enriched at repetitive regions of the genome, it would be detrimental to eliminate these ambiguous reads that usually originate from repeats[156]. See Table B.5 for the summary of the most widely used quality metrics including those that assess the reproducibility since two replicates per ChIP-seq and input samples are recommended[130,157].

Once the basic sequencing quality properties are checked, scores for signal-to-noise ratios should give an impression of how well the ChIP worked. It is important to note that the vast majority of the quality controls were optimized for ChIP-seq data of transcription factors and histone marks with abundant and very localized enrichments. ChIP-seq experiments with broad, domain-like enrichments or very few binding sites will very often not meet the recommended quality thresholds due to their inherently reduced signal-to-noise ratios[130,140,158,159] (Table B.5).

Normalized fragment coverages

One part of the ChIP-seq analysis workflow that is often mentioned, but rarely scrutinized in detail is the generation of coverage files for which the read-focused information from the genome alignment is converted into integer counts of the number of reads at a given genomic

locus (Figure 2.8, Figure 2.9). This conversion serves two purposes: a) the possibility to adjust for coverage biases and b) file size reduction which is essential for efficient downstream analysis tools, data sharing and visualization. The aim of the normalization is to make different samples comparable, irrespective of their sequencing depth and additional biases.

Despite its importance, there is no agreement on the optimal generation of coverage files (except for extending the reads to match the expected original DNA fragment size[145,160], Figure 2.9). Numerous strategies for the normalization of differing numbers of total reads per sample exist (e.g. linear scaling factors calculated with various methods[151,157,159], reads per million per kilobase[161], quantile normalization[162], trimmed mean of M-values[163]) and two methods for the correction of GC bias have been presented[138,139].

Identification of binding sites

In addition to coverage files, downstream analyses of ChIP-seq data are usually based on a set of genome regions that represent loci where the immunoprecipitated protein had bound[101,132,140]. These regions should be overrepresented in the ChIP-seq sample compared to the genome-wide background (hence they are commonly referred to as peaks) and numerous algorithms have been developed to separate binding events from background noise (reviewed by Pepke et al.[160]). The two basic steps of peak calling are: 1. identification of regions with large numbers of overlapping reads and 2. the assessment of the significance of the read enrichment. The programs choose slightly different approaches for both steps, influencing the sensitivity, specificity, but most strongly the positional accuracy of the peak region predictions[164]. The reliability of the significance measures absolutely depends on the choice of the statistical model to describe the distribution of background reads which is complicated by the numerous sources of coverage bias mentioned previously and in Table 2.5.

One of the most widely used tools, MACS[144], tries to accommodate for the sample-specific influences of the experimental as well as the bioinformatic procedures by utilizing dynamic Poisson distributions. They are parametrized on the background reads of each sample to capture the genome-wide values as well as local biases surrounding any candidate region taking the input control into consideration. To account for false positive calls, MACS version 1.4 calculates an empirical false discovery rate based on significantly enriched regions in the input sample[144,165].

Peak calling can be tremendously helpful for ChIP-seq analyses, but thorough benchmarking of the performance of the different algorithms is severely hindered by the lack of a comprehensive list of true binding sites that could be used for a fair comparison. In addition to ChIP-qPCR validation of selected peak regions[166], several proxies for peak quality have been used, such as DNA motif enrichments[164] and reproducibility between different peak calling

algorithms[166]. None of these approaches offer a global verification of peak regions, particularly not in regard to their exact position. Therefore, manual inspection is still the method of choice to decide on the success or failure of a peak calling step[167]. This is especially important when dealing with diffuse, domain-like signals for which great inconsistencies between different peak callers were reported[152].

Downstream analyses

Exploratory ChIP-seq analyses strongly depend on visualization and unsupervised clustering of the data to unveil unexpected patterns and correlations that can subsequently be experimentally tested[168]. There is virtually no standardization regarding downstream analyses as they must be tailored to the specific biological questions, but the most widely applied strategies are depicted in Figure 2.8. For example, *de novo* identification of DNA motifs underlying the ChIP-seq peaks is helpful to understand the targeting of a transcription factor[169] and gene ontology terms of putative target genes may give clues to the biological role of the investigated protein[170,171] (see Table B.4 for commonly used programs). Furthermore, additional genome-wide data sets are often integrated to complement the ChIP-seq; transcriptomics, for example, can be used to directly correlate the binding of a protein to transcriptional regulation of possible target genes[172].

2.4 Aims

The goal of my studies was to gain biological insights from the analyses of the binding profiles of MOF-associated proteins: the non-specific lethal complex (NSL) and the male-specific lethal complex (MSL) members. The analyses were based on chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq). Thus, before the biological questions could be addressed, we first needed to establish robust bioinformatic workflows and scripts ranging from data processing to bias identification and correction and manifold customized downstream analyses.

In *Drosophila*, we specifically wanted to address whether the different complex members of the NSL complex would co-occur. Moreover, we wanted to infer and subsequently test new hypotheses about the biological function of the NSL complex in regard to gene expression regulation. Since both MSL and NSL complex had not been previously examined in mammals, we then set out to study their chromatin targeting in mouse cells. To this end, Tomasz Chelmicki generated ChIP-seq profiles of MOF, MSL1, MSL2, NSL3 and MCRS1 in mouse embryonic stem cells and neuronal progenitor cells. The ChIP-seq data study revealed common and different binding principles of the two complexes in pluripotent and differentiated cells which we complemented with transcriptome studies from perturbation experiments and additional genome-wide data sets.

3. Results and discussion

The following chapter summarizes the methods and insights of the four manuscripts that can be found in the appendix of this thesis.

Appendix A.1 corresponds to Lam, Mühlpfordt, Vaquerizas et al. (2012) [90] where we examined the *Drosophila* ChIP-seq profiles of four members of the NSL complex (NSL1, NSL3, MCRS2, MBD-R2; generated by Sunil Raja and Kin Chung Lam) and ChIP-seq data for Pol II from S2 cells depleted of either NSL1 or NSL3 (experiments by Kin Chung Lam).

Appendix A.2 corresponds to Chelmicki, Dündar et al. (2014) [172] for which Tomasz Chelmicki generated ChIP-seq data of MOF, MSL1, MSL2, NSL3, MCRS1 in mouse embryonic stem cells and neuronal progenitor cells. Matthew Turley and Tasneem Khanam generated transcriptome data for cells depleted of individual factors.

Appendix A.3 corresponds to Ramirez, Dündar et al. (2014) [168], the publication of a software suite for quality control, normalization and visualization of high-throughput sequencing data.

Appendix A.4 corresponds to the submitted manuscript by Chlamydas et al. that focuses on the general role of fly and mammalian MSL1 at promoters and makes use of the ChIP-seq profiles of MSL1 in *D. virilis*, *D. melanogaster* and mouse.

3.1 Setting up a ChIP-seq analysis workflow

One of the main goals of my PhD studies was to identify pitfalls and optimized methods of ChIP-seq analyses whose basics had been established in our group by Sarah Diehl and Thomas Manke. In short, there were three main insights:

- Understanding biases of the data is crucial for the analysis as well as for improved experimental procedures.
- Standardized bioinformatic workflows must always be coupled to manual inspections and iterative optimization depending on the quality and nature of the data set and the questions to be answered.

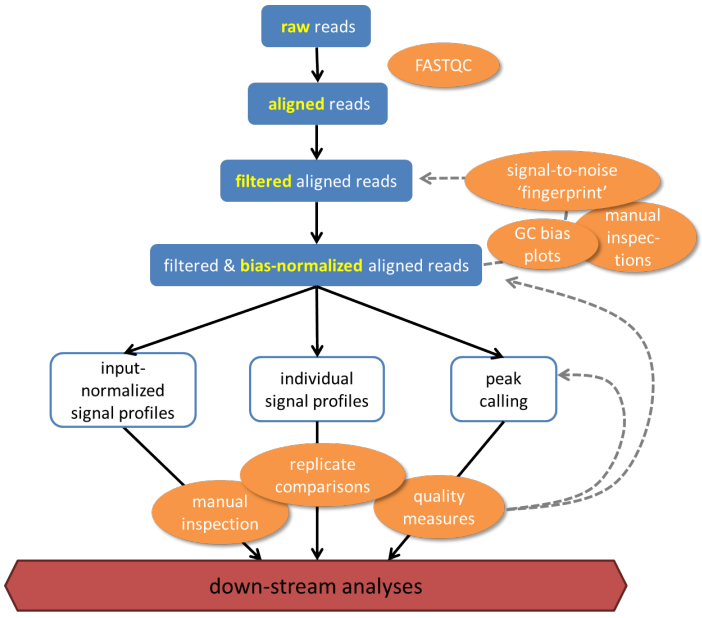


Figure 3.1: Schematic of the general bioinformatics workflow for ChIP-seq analyses. Orange circles depict quality controls, boxes with solid blue fill indicate read-based files (FASTQ and SAM/BAM, for file name explanations see the glossary of Appendix A.3), white boxes indicate files based on genomic intervals. The information of the various quality checks help to adjust all steps to match the samples' properties and the ultimate biological questions to be answered.

- Any software for analysis of high-throughput sequencing data must be inherently designed with some flexibility as the applications, sequencing platforms, file formats, data types and statistical models are constantly evolving.

My current workflow for ChIP-seq analysis is depicted in Figure 3.1. It depends heavily on deepTools, a software package that was developed in collaboration with Fidel Ramírez[168] (Table B.4). The workflow begins with the sequencing information stored as DNA reads in FASTQ format (for details on the file formats, see the glossary within the supplement of Appendix A.3). The reads are aligned using bowtie2[147] with default parameters. I usually discard unmapped, non-uniquely aligned, duplicated and low-quality reads, as well as those mapped to unassembled and mitochondrial chromosomes, to major satellites and other regions with obvious copy number variations. These rather aggressive filtering steps will probably yield non-optimal results for data sets with expected enrichments along repetitive regions and would have to be adjusted accordingly. However, particularly the removal of regions that consistently accumulated extremely high read numbers in both ChIP and input samples, improved the accuracy of scaling factor calculations and peak calling for the data sets that I analyzed (in line, the ENCODE consortium recently published a list of regions whose exclusion from the analysis improved data processing[142,143]). The optimization of the filtering steps was based on various quality controls, e.g. the assessment of the ChIP strength using the method proposed by Diaz et al.[159], manual inspections and GC bias checks.

3.1.1 GC bias normalization

The most prevalent and possibly distorting bias that is introduced by Illumina's PCR-based library preparation and sequencing method (Table 2.4) is the overrepresentation of GC-rich

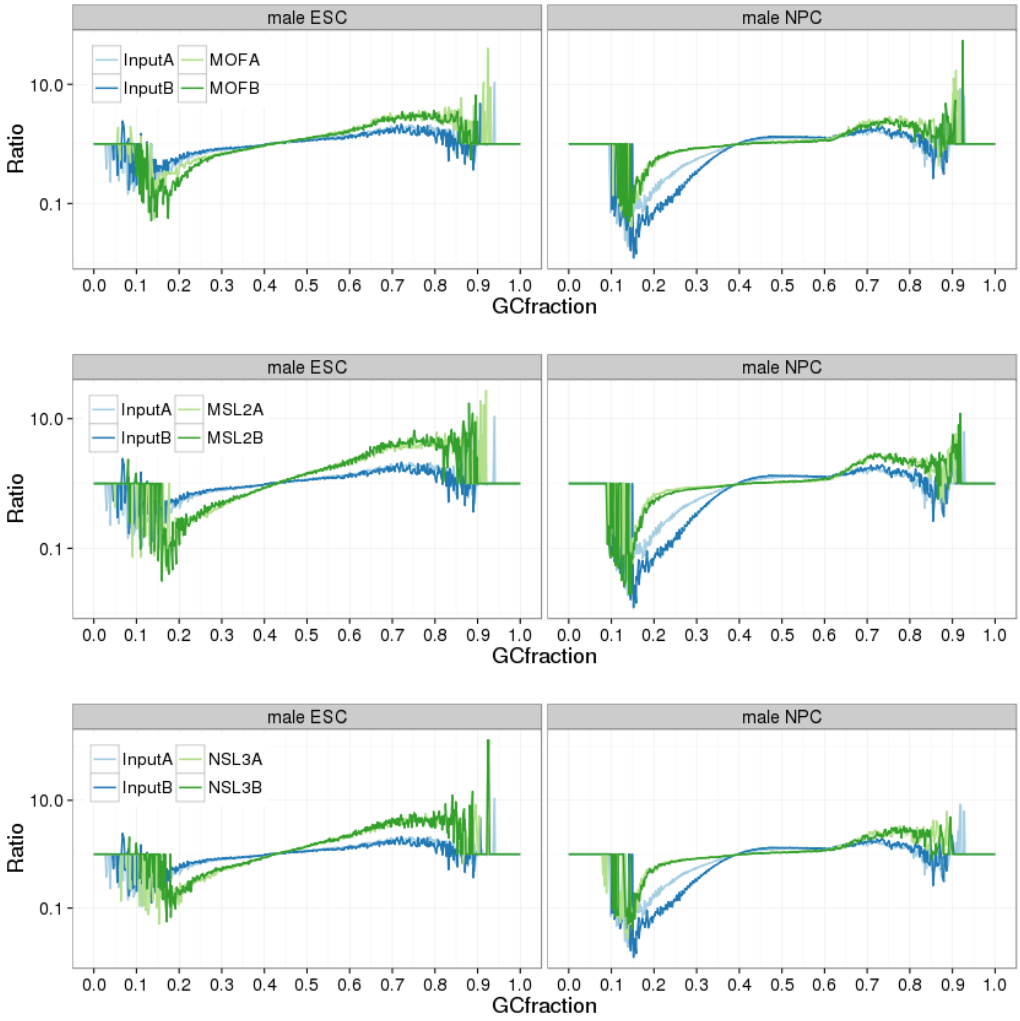


Figure 3.2: Different ratios of observed over expected read counts per GC content bin in ChIP-seq samples from embryonic stem cells (ESC) and neuronal progenitor cells (NPC). Green lines show the ratios for ChIP-seq samples (two replicates each), inputs are shown in blue, note the log scale. Ratios below zero indicate underrepresented regions, ratios above zero overrepresentation. The vast majority of the mouse genome will be covered by regions with 35-60% GC content, thus any bias between 0.35-0.60 (y-axis) should be paid attention to. The sequencing libraries of the ESC samples were prepared with standard Illumina polymerase, while NPC samples were done with fewer PCR cycles and the HighFidelity Polymerase from New England Biolabs. The lack of AT-rich regions in the input samples is most likely due to other experimental steps than PCR (Table 2.5). The values underlying the images were calculated with the `computeGCBias` tool of deepTools[168].

regions. Benjamini and Speed[139] demonstrated that even if the input is handled like the ChIP sample during the experiments, it will unlikely suffice to correct for abundant GC bias because the extent and the regions that are overamplified were shown to be specific for each library preparation. In the light of their insights, we examined the GC bias of the mouse ChIP-seq samples that later on were the basis for Chelmicki, Dündar et al[172]. As shown in Figure 3.2, the different samples indeed had distinct GC bias profiles and the observation that the bias was dramatically reduced when an optimized DNA polymerase and fewer PCR cycles

were used (second column of Figure 3.2) confirmed the claim of Benjamini and Speed [139] that the DNA amplification steps were the major cause of GC bias. Nevertheless, we identified additional issues that need to be taken into account before adjusting the observed read distributions:

1. ChIP-seq samples with strong enrichments at mammalian (typically GC-rich) promoters will have a biologically meaningful overrepresentation of GC-regions. This can, for example, be seen in the NPC ChIP-seq samples in Figure 3.2 where regions with more than 60% GC content are overrepresented. This is due to the preferred binding of the investigated proteins to CpG-rich promoters.

2. In contrast to the claim that the GC bias should be library-specific, replicates of the same experiment showed very similar GC profiles, indicating a dominant ChIP- (and input-) specific effect.

3. The calculation of the expected read distributions is based on the available genome assembly whose quality might influence the outcome, especially if regions with strong sequence composition biases are not included in the assembly.

We addressed the first two issues by a) excluding unmappable as well as significantly enriched regions from the read distribution calculation and b) in the case of the non-optimal (standard Illumina) library preparation, we decided to manually scale the input samples to each ChIP-seq so that the input samples reflected the ChIP-seq-specific GC bias rather than eliminating reads from GC-rich regions.

Perhaps the most important insight from these studies was the fact that optimized ChIP and sequencing protocols that were eventually enforced by our in-house sequencing facility almost completely eliminated GC biases. Individual samples at times still indicate the need for computational correction that should, however, be done with care and full knowledge of possible new artifacts that might be introduced such as the loss of enrichments in GC-rich regions[a]. In contrast to the BEADS package [138] that offers one standardized GC bias correction workflow, deepTools calculates and normalizes the read distributions using two separate tools (`computeGCbias` and `correctGCbias`, Appendix A.3), so that users can explore the expected and observed values first and thereupon decide which normalization strategy might best suit their data set.

Normalization for sequencing depth and input control

Once the read alignments have been thoroughly checked and possibly normalized, the next steps lay the foundation for the majority of ChIP-seq downstream analyses that are typically

[a] GC bias correction is applied on the files containing the aligned reads: reads in overrepresented regions are removed randomly, while underrepresented region will obtain artificially duplicated reads.

not dependent on the DNA sequence of each read, but instead rely on summaries of the number of overlapping reads at each given locus in the genome (Figure 2.9 and white boxes in Figure 3.1). I will usually first generate coverage profiles for each sample individually, normalizing for sequencing depth only:

$$normalized\,bin\,coverage = \frac{observed\,bin\,coverage \times effective\,genome\,size}{mapped\,reads \times fragment\,length}$$

These profiles are useful to confirm that the input-normalized ChIP-seq signals and the peak calling results (see below) match the patterns seen in the simple coverages.

For input normalization, both input and ChIP must be made comparable first before calculating the ChIP enrichments ($log_2 \frac{ChIP}{input}$) for each consecutive genomic bin. If the output of deepTools' `bamFingerprint`[159] indicates a clear ChIP-seq signal compared to the input[158], I tend to use the signal extraction scaling (SES) proposed by Diaz et al.[159]. SES is based on the assumption that the ChIP-seq sample contains both background and immunoprecipitated DNA and that the scaling factor for sequencing depth adjustment should be based on the background signal only. In a plot depicting the cumulative percentage of reads, the input should show a linear increase as each genome region should contain a similar fraction of reads (see the image of `bamFingerprint` in the supplemental material of Appendix A.3). ChIP-seq data, however, will contain a significant fraction of regions with consistently very few reads (= background) and a relatively small number of regions with exceedingly large read numbers (= immunoprecipitated DNA). The SES factor is the ratio of ChIP-seq over input reads at the point where the percentage of input reads maximally exceeds the percentage of ChIP-seq reads[159]. If the separation between background and enrichment signal is not possible, the total number of reads of the less deeply sequenced sample (n) can be used to adjust the sequencing depths:

$$scale\,factor = \frac{n}{mapped\,reads}$$

All above described methods are implemented in `bamCoverage` and `bamCompare` of the deepTools suite (Appendix A.3).

3.1.2 Peak calling

As described in the introduction, the identification of significantly enriched genome regions is central to the majority of ChIP-seq analysis[140] (Figure 2.8). I chose different peak calling strategies to meet the distinct challenges of each analyzed data set (summarized in Table B.6). All approaches rely on stringent filtering after obtaining the initial lists of significantly enriched regions, but only the Pol II ChIP-seq and the mouse data set allowed for additional selection of peaks found in two replicates.

3.1.3 Exemplary downstream analyses

The analyses presented in Appendix A.1 and Appendix A.2 were driven by two different basic questions: in Lam, Dündar, Vaquerizas et al.[90] we focused on the *similarities* of the binding patterns of the NSL complex members in *D. melanogaster* while our study of both MSL and NSL proteins in mouse cells[172] aimed at identifying the possibly distinct functions of the complexes. I would like to point out three of the most important analysis strategies that formed the basis for almost all subsequent investigations. The details of all downstream analyses, including those related to DNA motif analysis and the integration of numerous published data sets can be found in the methods parts of Appendix A.1 and A.2. Furthermore, Table B.4 gives an overview of the software tools I used.

Binding profiles of MSL and NSL proteins

Both studies were based on the insights from so-called metagene analyses (or more generally: summary plots) and heatmaps. Summary plots depict the average ChIP-seq signal over an arbitrary number of genome regions (see Box 3 in[173]) while heatmaps allow more detailed views of virtually any kind of genome-wide data. Both types of images can be efficiently produced with `profiler` and `heatmapper` of the deepTools suite[168] (see Figures 3E, 4A and Figure supplement 3 of Appendix A.2 for examples).

The summary plots and heatmaps in Figure 3.3 demonstrate that the ChIP-seq enrichments of the individual MSL complex members differ significantly between autosomal and X-linked genes in males as it had been reported before[109,174,175]. We found that the signals of the NSL complex members, however, are not strongly influenced by the chromosome.

In mice where both complexes are expressed and presumably assembled in a sex-independent manner we did not detect the broad enrichments typical of dosage-compensated fly genes, but instead narrow, localized signals, preferably around the TSS of active genes (MOF, NSL3, MCRS1) or at putative regulatory sequences (MSL1, MSL2; see Appendix A.2).

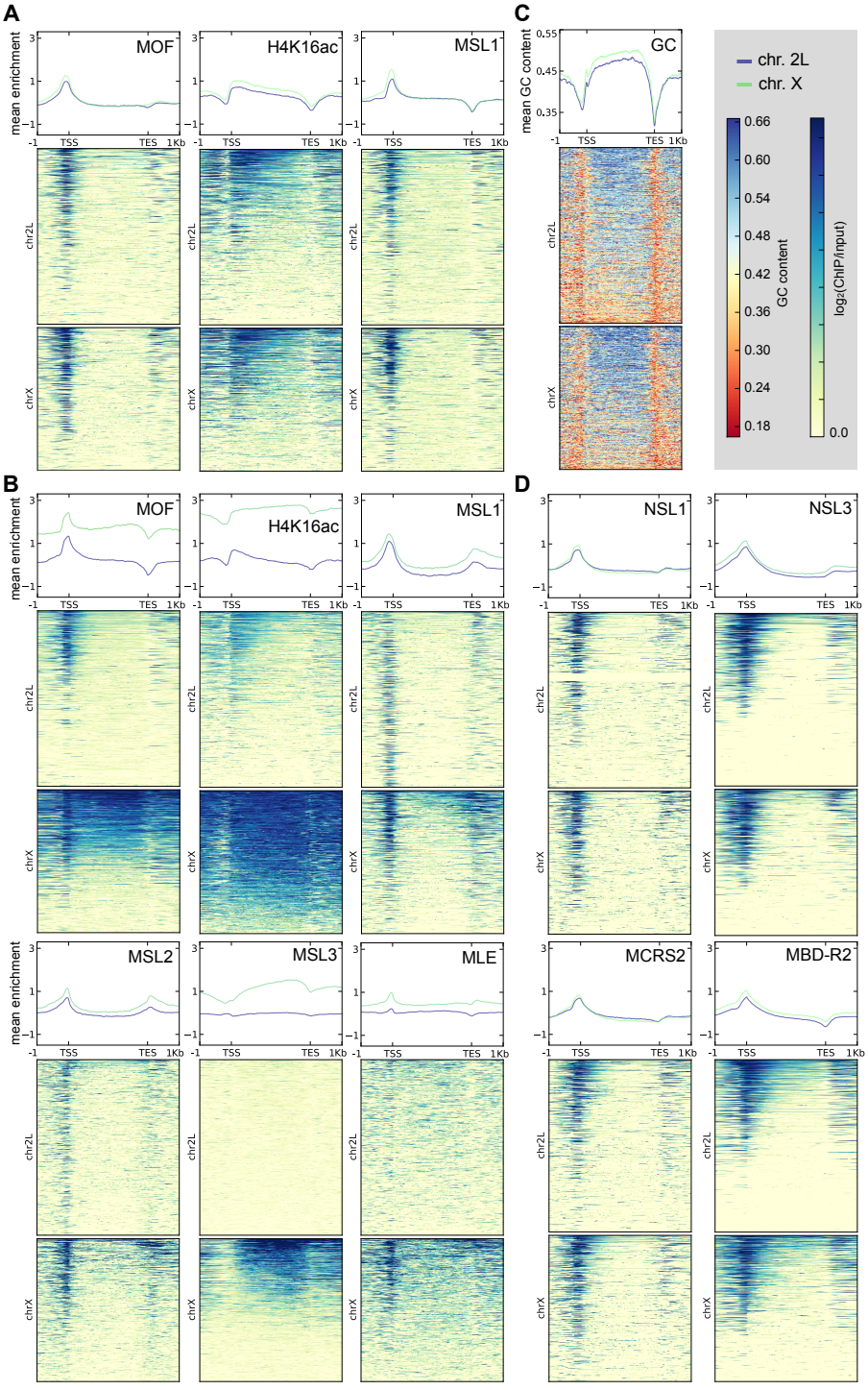


Figure 3.3: Summary plots and heatmaps of various ChIP-seq signal and GC content for *Drosophila* genes on the X chromosome (chrX) and chromosome 2L (chr2L). **A)** MSL complex members in female flies do not distinguish between X and autosomal genes. **B)** MSL complex members in male flies show drastically enriched signals along the majority of X-chromosomal genes (lower panel). **C)** Visualization of the GC content along fly genes that reveals the difference of GC content between the autosomal chromosome 2Land the X chromosome. **D)** NSL complex members in male salivary glands and S2 cells do not distinguish between genes on the X and the autosomal chromosome. All images were generated with `computeMatrix` and `heatmapper` from the deepTools suite (Appendix A.3) using the scale-regions mode to scale gene bodies to 2 kb. See Table B.8 and B.11 for details about the ChIP-seq data sets.

Identifying target genes

In addition to describing the binding patterns of the proteins of interest, the characterization of the target genes is of eminent importance for gaining biological insights. The identification of target genes is usually done on the basis of proximity (e.g. peaks close to the promoter of a gene), but it can be significantly enhanced by taking additional criteria into consideration such as the number and intensity of peaks, expression data from perturbation experiments and binding site conservations[176]. In Lam, Mühlpfordt, Vaquerizas et al.[90], we used very narrow regions to account for the gene-dense nature of the fly genome: Genes were classified as NSL targets if the summit regions of peaks of *all* four profiled proteins (NSL1, NSL3, MCRS2, MBD-R2) overlapped within 200 bp up- or downstream of the transcription start site (TSS). For the mammalian proteins, we identified an initial set of targets if a peak overlapped within 1 kb upstream of a gene's TSS. We then refined these targets by requiring multiple overlapping peaks as well as significantly affected expression after shRNA-mediated depletion of the respective proteins (Appendix A.2).

Identifying target genes that may be regulated via promoter-distal binding in introns or intergenic regions is less straight-forward as no clear rules for enhancer-gene associations have been established although proximity seems to perform reasonably well[177]. Since a profound number of MSL1, MSL2 and NSL3 binding sites in the mammalian cells were located promoter-distally, I made use of GREAT, a web-based program that predicts putative target genes of *cis*-regulatory regions[178] (see Table B.4).

Unsupervised clustering of ChIP-seq signals

While the ChIP-seq studies of the MSL complex (but not the NSL complex) in *D. melanogaster* had been supported by a substantial body of literature on the biological role of the complex[88], very little was known about the functions of the mammalian orthologues of MSL and NSL complexes. Structural and biochemical studies eventually demonstrated that the mammalian proteins, too, were able to form two distinct MOF-containing complexes[40,50,54,179], but whether the proteins would predominantly bind the same target genes or rather carry out independent functions remained elusive. In addition, we wanted to examine whether the binding patterns were changing in murine embryonic stem cells (ESC) compared to neuronal progenitor cells (NPC). To obtain a comprehensive impression of all the enrichments in both cell types, hierarchical clustering proved to be a powerful method to reveal the underlying patterns of exclusive and co-occurring enrichments in an unbiased manner. The method described in Appendix A.2 yielded a robust and insightful result[b] that

[b] In brief, I worked on the union of all peaks from both cell types which were scaled to 1.2 kb; normalized signal values were calculated for 50 bp bins for each ChIP-seq sample. The resulting matrix was rank-transformed, converted into Euclidean distance measures and clustered with the hclust function of R (ward

is shown in Figure 2 of Appendix A.2. The main findings were:

- MOF predominantly associates with the NSL complex at gene promoters.
- There is a small subset of (mostly non-promoter) regions where MOF associates almost exclusively with the MSL complex.
- Particularly MSL2 binds to a significant number of intergenic and intronic binding sites where none of the other proteins investigated in our study were found.

3.2 The roles of MOF within its distinct complexes

3.2.1 MOF within the MSL complex

In *Drosophila*, the function of MOF within the male-specific complex seems well established: it catalyzes the massive acetylation of H4K16 at the male X chromosome, thereby enhancing its transcription. In mammals, the MSL complex has retained its ability to enhance H4K16ac by MOF but besides the general opening of chromatin and transcription enhancement (from promoters as well as enhancers[181]), no specific cellular mechanism that solely depends on H4K16ac could be pin-pointed yet. Of note, the decisive role of MSL1 and MSL2 for the maintenance of X chromosome expression in mouse ESCs (see below) does not depend on MOF.

3.2.2 MOF within the NSL complex

An initially surprising finding from the mammalian study of NSL and MSL complexes was that MOF seemed to almost always be accompanied by a member of the NSL complex which would only sometimes be complemented by signals from the MSL complex. (Interestingly, Figure B.2 reveals that in flies, too, MSL and NSL complexes have largely overlapping target gene sets as all NSL-bound genes contain strong signals of the MSL complex.) The transcriptomes of shRNA-treated cells confirmed the notion that the NSL complex might be the predominant interaction partner in mammalian cells: the effects on gene expression were very similar in MOF- and NSL3-depleted cells, but differed significantly from MSL1 or MSL2 depletions (see Figure 3E-G of Appendix A.2). A remarkable example was the negative effect of MOF and NSL3 depletion on the expression of key pluripotency factors and subsequent loss of pluripotency which was not observed in MSL1- or MSL2-depleted ESCs (Appendix A.2).

linkage)[180]. The rank transformation proved to be the most important step while changing the distance and linkage methods had little effects on the outcome.

Conversely, in flies MOF's presence at NSL-bound promoters does not closely mirror the broader gene enrichments of H4K16ac (Figure 3.3). This could well be due to technical artifacts as the formaldehyde fixation might not capture a transient and less abundant binding of MOF along gene bodies in contrast to the more stable histone modification[121] (Table 2.5). However, several HAT complexes contain bromodomains which could bind H4K16ac at the promoter and perhaps mediate spreading of the signal[182] (Figure 2.3). p300/CBP, for example, was recently shown to be capable of binding to and catalyzing H4K16ac *in vitro*[183,184]. The histone acetyl transferase ATAC2 (a GNAT family HAT, Table 2.1) also contributes to bulk H4K16ac levels in fly embryos and mice[185,186], and interestingly, it strongly interacts with WDS (WDR5 in mammals). Since WDS is part of the NSL complex, one could speculate about ATAC2 recruitment to promoters of NSL targets and subsequent spreading of H4K16ac. It should be noted though that the previously proposed concomitant interaction of WDR5 with both MLL and NSL complexes[187] was recently shown to be of mutually exclusive nature[50], indicating that WDR5 might not be able to act as a bridging factor between different complexes.

Surprisingly, the results of our studies suggest that H4K16ac may not be the major mechanism through which the NSL complex conveys its essential functions (although the NSL1-MOF interaction is very similar to the MSL1-MOF interaction and both increase MOF's HAT activity[40,70]). The ablation of individual NSL complex members in both *Drosophila* and mammalian cells affects H4K16ac levels only moderately (MCRS2[52]) or not at all[179] (NSL1, NSL3, MBD-R2; see Appendix A.1 and A.2). In flies, MSL1 might maintain H4K16ac levels in the absence of NSL proteins since it binds to the same promoters (Figure B.2). Regardless of the NSLs' influence on H4K16ac, however, H4K16ac does not seem to be the most vital function as female flies can survive until adulthood without MOF, although with compromised longevity and fertility[22,38]. In mice, on the other hand, lack of MOF is lethal which indicates that it must play an essential role. The recently discovered ability of MOF to acetylate non-histone proteins may link MOF more directly to vital cellular processes, possibly even in the context of the NSL complex. The regulation of p53-dependent apoptosis induction following DNA damage, for example, relies on the acetylation of p53 by MOF[99,100] and was shown to be enhanced by NSL1, but not MSL1[70].

3.3 The NSL complex regulates housekeeping genes in *D. melanogaster* and *M. musculus*

In *D. melanogaster*, we found that the NSL target genes showed all chromatin and genome properties that had been reported for housekeeping genes: they were strongly enriched for chromatin states of active transcription[16,188] and DNA motifs of non-tissue-specific genes[189],

they predominantly had dispersed transcription initiation patterns[190] and very stable nucleosome positioning around the TSS with consistently depleted -1 nucleosomes (analysis by Juanma Vaquerizas), and, most importantly, the vast majority (>90%) of NSL target genes were moderately, but ubiquitously expressed in a variety of different cell lines and developmental stages[191,192].

In *M. musculus*, the NSL target genes showed similar characteristics: cell-type-independent expression, motifs associated with housekeeping genes (e.g. ELK1, NRF2, E2F[193]) and gene ontology terms associated with basic transcription, translation and metabolism functions (Appendix A.2). Moreover, we found that orthologues of *D. melanogaster* NSL target genes had an increased probability of being an NSL target in mouse cells compared to other gene sets (Figure supplement 2B of Figure 3, Appendix A.2). This suggests that the NSL complex might not recognize one specific DNA motif, but instead identifies its target genes based on the larger chromatin context.

Little is known about the regulation of housekeeping genes that generally show more stable and uniform expression than tightly regulated genes. It is possible that housekeeping genes are maintained in a general non-restrictive chromatin state that allows Pol II to bind in a stochastic manner. This is in line with the previously mentioned properties of housekeeping genes such as well-defined nucleosome-free regions upstream of the TSS and dispersed, multiple transcription starts. How the NSL complex contributes to housekeeping gene expression still needs to be elucidated in detail, but given its members' properties (Figure 2.5, Table B.3), it could contribute to the maintenance of the housekeeping expression environment by multiple means, e.g. through histone modifications or through interactions with nucleosome remodellers as well as the transcription machinery and additional transcription activators.

As discussed previously, MOF does not seem to be essential for the role of the NSL complex. If one assumes that the lethality of mutations in *nsl* genes is caused by the disruption of housekeeping gene expression, there are two possible explanations: lack of MOF might be rescued by the recruitment of a different HAT (such as ATAC2, see above) or transcription activation could mostly be dependent on other NSL complex members and their interaction with the transcription machinery. Indeed, we found that the depletion of the NSL complex (in both *Drosophila* and mammals) has detrimental effects on gene expression with reduced Pol II occupancy for and downregulation of housekeeping genes (Appendix A.1 and A.2). Kin Chung Lam could show that the NSL complex directly influences the recruitment of the pre-initiation complex (Appendix A.1), supporting the notion that the NSL complex is a transcriptional activator even in the absence of MOF[52].

In mouse ESC we observed a specific effect of NSL complex ablation on key pluripotency factors that could not be explained by promoter signals of either NSL3 or MOF[172,179]. One explanation could be the strong enrichment of NSL3 (but not MOF) at ESC super-enhancers that were shown to be important for pluripotency[194]. Alternatively, the decreased expression of pluripotency factors could be the result of the general disturbance of ESC homeostasis when housekeeping gene regulation is tempered with. Indeed, the maintenance of pluripotency and unlimited ESC proliferation was shown to be strongly influenced by several basic cellular mechanisms (e.g. cell cycle, energy metabolism, ribogenesis[195–197]). The importance of the NSL complex for ESC pluripotency might thus reflect the strong need for continuously high levels of cellular building blocks – transcription factors, nucleotides, ribosomes – within the perpetually proliferating ESCs that might be provided by the binding of the NSL complex (including MOF) to promoters of housekeeping genes.

While it is reasonable to assume that the regulation of genes encoding proteins for basic cellular functions is the underlying reason for the non-specific lethality of *nsl* mutations, it should be noted that all NSL complex members except MOF were identified as essential factors of mitosis (Figure 2.6, Table 2.3) which suggests additional, possibly vital and perhaps chromatin-independent functions.

3.4 The specialized tasks of MSL1 and MSL2

As mentioned previously, we found surprisingly few overlaps between MOF and the MSL complex in mammalian cells. Instead of joining MOF at promoters, MSL1 and MSL2 frequently localized to TSS-distal regions.

3.4.1 MSL1 and MSL2 regulate gene expression via TSS-distal binding sites

The TSS-distal binding that we observed in mouse ESCs and NPCs seems reminiscent of the binding to the intronic and intergenic MSL entry sites along the fly X chromosome[198,199]. If that were indeed the case, one would expect that the mammalian TSS-distal targeting should depend on the formation of the MSL1-MSL2 heterodimer[54]. While we did not examine the mammalian TSS-distal binding in molecular detail, we observed that MSL1 and MSL2 strongly affect each other: the depletion of mammalian MSL1 leads to dramatically reduced protein levels of MSL2 and vice versa which mirrors the observations in *Drosophila*[200]. Consequently, the transcriptome changes in MSL1- and MSL2-depleted ESCs are quite similar on a global scale although individual differences do exist (Figure 3E-G and Figure 4E of Appendix A.2). In flies, binding of MSL1-MSL2 is accompanied by the remaining complex

members that are required for the characteristic spreading associated with dosage compensation. It will be interesting to see if the heterodimer is joined by additional proteins at TSS-distal sites in the mammalian genome as well.

The biological importance of the TSS-distal binding sites was supported by our findings in MSL1- or MSL2-depleted cells: genes that had been predicted to be regulated by TSS-distal binding sites were significantly more often and slightly stronger downregulated than promoter targets of MSL1 and MSL2 or putative TSS-distal targets of the NSL complex (Figure 3F, Figure 4E, Figure supplement 4B and 4C of Figure 4 of Appendix A.2). Interestingly, MSL2's ubiquitin ligase activity is enhanced by the presence of MSL1, but independent of MOF[91]. The histone modification set by MSL2, H2BK34ub, was implied to stimulate methylation of H3K4[91] and chromatin recruitment of CDK9, a kinase that is part of the positive elongation factor b (P-TEFb; Figure 2.1)[92]. MSL1 and MSL2 could therefore convey MOF-independent support of transcription.

We noted that several TSS-distal MSL1/MSL2 loci were part of the X inactivation center, the region of the mammalian X chromosome that is necessary and sufficient to drive the random X inactivation in female cells (reviewed by Pollex and Heard[201]). The binding to the intronic minisatellite of *Tsix* was particularly strong and turned out to be a striking example for the eminent biological function of enhancer binding by MSL1 and MSL2 in ESCs. *Tsix* is the rodent-specific antisense transcript of *Xist* that inhibits *Xist* transcription and subsequent X inactivation in female mouse ESCs[201]. Tomasz Chelmicki carefully examined the effects of MSL1, MSL2 and MOF depletion on *Tsix* and *Xist* expression and could show that MSL1 and MSL2 (but not MOF) are required for *Tsix* transcription and efficient *Xist* repression (Appendix A.2). Incidentally, MSL1 and MSL2 therefore secure the expression of the entire mammalian X chromosome by maintaining the transcription of one specific locus in mouse ESCs. The underlying mechanism is very different from the function of the MSL complex in flies, but the overall effects are similar.

3.4.2 The E3 ubiquitin ligase MSL2 is distinctly enriched at SMAD3 motifs

In addition to the profound overlaps between MSL1 and MSL2 (70% of MSL1 binding sites were also enriched with MSL2), we detected a substantial number of MSL2 binding sites without even the trace of a signal by any of the other proteins (Figure 2 of Appendix A.2). These peaks were unusually uniform in size and shape and were found in regions without signs of accessible chromatin or transcription in ESCs. We could not exclude that technical artifacts might be the underlying cause, however, these distinct enrichments were based on signals from two replicates and their number increased even further in the ChIP-seq samples from mouse NPCs (829 solitary peaks in ESCs to 3,635 in NPCs). These mostly intronic

loci are strongly enriched for a $(CAGA)_n$ motif which was described as the binding site for SMAD3, a transcription factor that conveys gene expression changes following stimulation of the transforming growth factor receptor beta (TGF-beta)[202]. Additional experiments are needed to clarify whether these unique signals are signs of novel MSL2 functions or (rather specific) byproducts of the MSL2 ChIP-seq protocol. Unfortunately, even the confirmation with ChIP-qPCR has thus far been hindered by the repetitive nature of these loci.

3.4.3 MSL1 interacts with CDK7

In addition to our findings in mouse cells, structural and biochemical studies of *Drosophila* MSL1 supported the notion from the ChIP-seq profiles (Figure 3.3) that MSL1 was binding to promoters of autosomal and X-linked genes likewise and independently of MSL2[54]. In a very recent study that is about to be published (Appendix A.4), Sarantis Chlamydas investigated the molecular role of MSL1 at promoters. He could show that MSL1 physically interacts with the cyclin-dependent kinase 7 (CDK7) which is responsible for the phosphorylation of the serine residue 5 of Pol II[203] (Figure 2.1). It seems as if MSL1's scaffold ability that it provides within the MSL complex also serves to mediate the interaction of CDK7 with its target promoters, thereby supporting transcription initiation, i.a. at NSL-bound genes. In contrast to the MSL2-dependent binding associated with dosage compensation in *Sophophora*, the promoter binding of MSL1 is conserved in distant species with different dosage compensation strategies. However, the loss of serine 5 phosphorylation upon MSL1 depletion is substantially stronger in flies than in mouse ESCs. This is in line with our observation that MSL1 and MSL2 only bound a subset of promoter regions in ESCs (Appendix A.2). While the MSL1-CDK7 interaction is evolutionarily conserved, mammalian CDK7 does not seem to require MSL1 for serine 5 phosphorylation generally, but for rather specific target genes.

3.5 Conclusion

We showed that the NSL complex has a general role for transcription regulation of housekeeping genes in both mammals and *Drosophila* and that it seems to be the main chromatin interaction partner for MOF, possibly contributing to more than one molecular function. The MSL complex recruits MOF to specific sites where its effects rely on H4K16 acetylation that creates a permissive chromatin environment. The male *Drosophila* X chromosome is an extreme example since an entire chromosome is targeted and affected. The presence of the MSL1-MSL2 at a subset of NSL-bound promoters in mammals suggests that they may enhance transcription in a gene-wise manner.

Our studies have revealed additional facets of both complexes such as the recruitment of the pre-initiation complex by the NSL complex and the evolutionarily conserved, MOF-independent role of MSL1 for the interaction with the transcription initiation factor CDK7 at promoters. Moreover, MSL1 and MSL2 were shown to repress the expression of *Xist* in both male and female mouse ESCs. Since the repression is mediated through the transcription of the rodent-specific antisense transcript of *Xist*, *Tsix*, this is another striking example for the adaption of MSL1 and MSL2 to perform a highly species-specific task and it will be interesting to see how its molecular functions (e.g. the ubiquitin ligase activity of MSL2 or the scaffold purpose of MSL1) predestine this heterodimer to be dedicated to additional particular roles that may be revealed in the future.

A. Publications and manuscripts

A.1 The NSL complex regulates housekeeping genes.

Lam, K. C.*, **Mühlpfordt, F.***, Vaquerizas, J. M.*, Raja, S. J., Holz, H., Luscombe, N. M., Manke, T., Akhtar, A. (2012). *PLoS Genetics*, 8(6), e1002736. doi:10.1371/journal.pgen.1002736
* shared authorship

I performed the majority of the bioinformatic analyses such as peak calling optimization and all further downstream analyses for all ChIP-seq samples. For the ChIP-seq samples of RNA Polymerase II, I additionally performed the read alignment and normalization procedures.

I generated all figures except Figures 1C, 2D, 3B, 5 and Supplementary Figures 3–6.

Together with Ken Lam and Asifa Akhtar I devised, wrote, and revised the manuscript.

OPEN ACCESS Freely available online

PLoS GENETICS

The NSL Complex Regulates Housekeeping Genes in *Drosophila*

Kin Chung Lam[1,2☯], Friederike Mühlpfordt[1,2☯], Juan M. Vaquerizas[3☯], Sunil Jayaramaiah Raja[1], Herbert Holz[1], Nicholas M. Luscombe[3,4], Thomas Manke[1], Asifa Akhtar[1]*

1 Max-Planck Institute of Immunobiology and Epigenetics, Freiburg im Breisgau, Germany, **2** Faculty of Biology, University of Freiburg, Freiburg, Germany, **3** EMBL European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, United Kingdom, **4** Okinawa Institute of Science and Technology, Kunigami-gun, Okinawa, Japan

Abstract

MOF is the major histone H4 lysine 16-specific (H4K16) acetyltransferase in mammals and *Drosophila*. In flies, it is involved in the regulation of X-chromosomal and autosomal genes as part of the MSL and the NSL complexes, respectively. While the function of the MSL complex as a dosage compensation regulator is fairly well understood, the role of the NSL complex in gene regulation is still poorly characterized. Here we report a comprehensive ChIP–seq analysis of four NSL complex members (NSL1, NSL3, MBD-R2, and MCRS2) throughout the *Drosophila melanogaster* genome. Strikingly, the majority (85.5%) of NSL-bound genes are constitutively expressed across different cell types. We find that an increased abundance of the histone modifications H4K16ac, H3K4me2, H3K4me3, and H3K9ac in gene promoter regions is characteristic of NSL-targeted genes. Furthermore, we show that these genes have a well-defined nucleosome free region and broad transcription initiation patterns. Finally, by performing ChIP–seq analyses of RNA polymerase II (Pol II) in NSL1- and NSL3-depleted cells, we demonstrate that both NSL proteins are required for efficient recruitment of Pol II to NSL target gene promoters. The observed Pol II reduction coincides with compromised binding of TBP and TFIIB to target promoters, indicating that the NSL complex is required for optimal recruitment of the pre-initiation complex on target genes. Moreover, genes that undergo the most dramatic loss of Pol II upon NSL knockdowns tend to be enriched in DNA Replication–related Element (DRE). Taken together, our findings show that the MOF-containing NSL complex acts as a major regulator of housekeeping genes in flies by modulating initiation of Pol II transcription.

Citation: Lam KC, Mühlpfordt F, Vaquerizas JM, Raja SJ, Holz H, et al. (2012) The NSL Complex Regulates Housekeeping Genes in *Drosophila*. PLoS Genet 8(6): e1002736. doi:10.1371/journal.pgen.1002736

Editor: Jason Carroll, Cancer Research UK Cambridge Research Institute, United Kingdom

Received October 18, 2011; **Accepted** April 13, 2012; **Published** June 14, 2012

Funding: This work was supported by EU funded ITN "Nucleosome 4D" and DFG funded "SFB746" awarded to AA. JMV acknowledges funding from the ESF Exchange Grant program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: akhtar@immunbio.mpg.de

☯ These authors contributed equally to this work.

Introduction

In the past decade, our understanding of eukaryotic transcriptional regulation has changed from the notion of a "generic entity that functions by a single universal mechanism" [1] to the acknowledgement of diversity in promoter types and functions. Indeed, eukaryotic transcription relies on a complex interplay between DNA binding motifs, covalent histone modifications, higher order chromatin structures and protein-protein interactions. For example, post-translational modifications of histones such as acetylation, methylation, phosphorylation, ubiquitinylation, and sumoylation are prominent mechanisms employed to help modify chromatin structure and are considered to be a prerequisite for the recruitment of general transcription factors (GTFs) (for review see [2,3]). Histone acetylation can impact chromatin structure in several ways: it has been shown that acetylation at different lysine residues can be specifically recognized by distinct protein domains (e.g. bromodomains) [4,5], which in turn recruit chromatin-remodeling factors. Alternatively, acetylation itself may also disrupt interactions between nucleosomes and thus cause chromatin decompaction [6,7]. Both mechanisms can contribute to reduced nucleosome occupancies at transcriptional start sites (TSSs), thereby providing an open chromatin environment for GTF binding [8].

Histone acetyltransferases (HATs) and histone deacetylases (HDACs) work in concert to orchestrate a fine balance of acetylation. HATs can be classified into two predominant families: the GCN5-related *N*-acetyltransferase (GNAT) family (e.g. Gcn5 and p300) [9] and the Moz-Ybf2/Sas3-Sas2-Tip60 (MYST) family (e.g. Tip60 and MOF) [10]. These enzymes often function as part of multi-protein complexes, presumably to increase substrate-specificity and to impose tight regulation of their enzymatic activity. Moreover, mounting evidence suggests that a single HAT can often associate with more than one complex [11]. Gcn5, for example, is a member of both the SAGA and ATAC complexes [12,13] that regulate different sets of inducible genes despite sharing the same HAT [14–18].

Similarly MOF, a MYST-HAT specific for H4K16 acetylation, is also a member of two distinct protein complexes in *Drosophila* and mammals: the Male-Specific Lethal (MSL) and the Non-Specific Lethal (NSL) complexes [19–21]. In *Drosophila*, the MSL complex is targeted to the transcribed regions of X-chromosomal

Author Summary

Housekeeping genes are required to support basic cellular functions and are therefore expressed constitutively in all tissues. Although the homeostasis of housekeeping gene expression is vital for cell survival, most research on the transcription initiation has been focused on TATA-box-containing promoters of inducible and developmental genes, while regulatory mechanisms at the TATA-less promoters of housekeeping genes have remained poorly understood. Using genome-wide chromatin binding profiles, we find that the NSL complex, a histone acetyltransferase-containing complex, is bound to the majority of constitutively active gene promoters. We show that NSL-bound genes display specific sets of DNA motifs, well-defined nucleosome free regions, and broad transcription initiation patterns. In addition, we show that the NSL complex regulates the recruitment of the basal transcription machinery to target promoters; more specifically, we can pinpoint its role to the early steps of Pol II recruitment. Interestingly, we also see that NSL-bound genes are most susceptible to Pol II loss after depletion of NSLs when they contain the DNA Replication–related Element (DRE). Taken together, we provide a genome-wide analysis of a chromatin-modifying complex that is globally involved in the regulation of housekeeping gene expression.

genes where it mediates dosage compensation. The targeting mechanism and modes of action of the MSL complex have been studied extensively (for review see [22–24]). In contrast, details of the NSL complex have only recently started to emerge. Our previous work revealed that the NSL complex is composed of at least seven proteins: NSL1, NSL2, NSL3, MCRS2, MBD-R2, WDS and MOF [20,21]. We have also shown that these proteins are essential for the viability and development of *Drosophila* and that they are required for the recruitment of MOF to the promoters of active genes [21,25]. Using a reporter assay system, Becker and colleagues demonstrated that MOF displays greater potential for transcriptional activation as part of the NSL complex, than in the MSL complex [26]. Additionally, recent reports indicate that in mammals MOF fulfills different functions in the NSL and MSL complex, respectively. It has been shown, for example, that the mammalian NSL1/MOF sub-complex appears to have broader substrate specificity than the MSL1/MOF sub-complex, as it is also able to acetylate non-histone targets [27]. Despite these observations, our understanding of NSL complex targeting and its regulatory function is still limited. Since the complex is conserved from *Drosophila* to mammals [20], unraveling its mechanism of action will be crucial for a better understanding of transcriptional regulation in higher eukaryotes and its evolutionary plasticity.

In order to elucidate the principles that direct NSL targeting, here we have performed a detailed analysis of the NSL binding sites in the genome of *Drosophila melanogaster*. We tested whether the NSL complex binds differently in distinct cell types by comparing ChIP-seq profiles obtained from the salivary glands of third instar larvae and from the Schneider (S2) cell line; our analyses reveal that the repertoire of NSL-bound genes is highly similar between different cell types. Remarkably, by comparing NSL target genes with transcriptome data from 30 distinct developmental stages of *Drosophila*, we find that the NSL complex preferentially targets genes that are constitutively expressed, also referred to as housekeeping genes. Moreover, NSL-bound genes exhibit elevated levels of H3K4me2/3, H3K9ac and H4K16ac and display a distinctive arrangement of the nucleosome free region (NFR) as well as dispersed transcription initiation patterns. Going beyond the study of NSL complex localization, we could furthermore show that the NSL complex is required for optimal recruitment of Pol II and the pre-initiation complex to its target promoters. Finally, using a quantitative model of DNA-protein interaction affinities, we find that the presence of strong DRE motifs in NSL target promoters conveys an increased sensitivity for Pol II loss in cells lacking NSL1 or NSL3. Taken together, our observations reveal a unique promoter configuration that is indicative of NSL binding and establishes the NSL complex as an important transcriptional regulator of constitutively expressed genes in *Drosophila*.

Results/Discussion

NSL complex targets a core set of genes independently of cell type

Genome-wide mappings of two NSL components (NSL1 and MCRS2) were previously performed in the salivary glands of third instar larvae [21]. Here, in addition, we performed chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) analyses of two additional proteins, NSL3 and MBD-R2, using the *Drosophila* embryonic Schneider (S2) cell line. This strategy allowed us to study similarities and differences in DNA binding patterns of the NSL-complex members in tissues of different origins. Moreover, the use of S2 cells offered the possibility to directly compare our results with the publicly available data generated by the modENCODE project that uses the same cell type.

The four proteins displayed significant binding, ranging from 9,409 (NSL3) to 12,234 (MCRS2) genomic regions where peaks were detected (false-discovery rate $<5\%$; statistics for individual proteins are provided in Table S1, for details on data processing see Materials and Methods). As shown in Figure S1A (dark and light blue columns), the majority of ChIP-seq peak summits localize within 800 bp of an annotated Transcription Start Site (TSS). The strongest binding signals (signified by high ChIP-seq tag counts) are observed within 200 bp of TSSs (Figure S1B, Figure 1A). We therefore focused our further analysis on NSL binding in TSS regions.

We defined a gene as NSL target when a ChIP-seq peak summit region (40 bp) was located within $+/-200$ bp of its TSS (see schematic representation, in Figure 1B). Using this criterion, we identified 4,233, 4,265, 5,435, and 5,321 promoters bound by NSL1, MCRS2, NSL3 and MBD-R2, respectively. Particularly in promoter-proximal regions, the binding profiles of NSL1, NSL3, MCRS2 and MBD-R2 are remarkably similar and also significantly overlap with previously published ChIP-chip profiles of WDS and MOF (Figure 1A). Despite the different developmental origins of the tissues used for the analysis of NSL1/MCRS2 and NSL3/MBD-R2, we observe that 78.7% (p-value$<$2.2 e-16; Fisher's exact test) of promoters with significant NSL signals are in common between the samples from S2 cells and larval salivary glands (Figure 1B). We identified a core set of 2,841 genes that are bound by all four NSL complex subunits, suggesting that the NSL proteins mostly operate as a single complex to regulate large numbers of genes in the *Drosophila* genome (Figure 1B, Figure S1C). Furthermore, ChIP followed by quantitative real time PCR (ChIP-qPCR) analysis of eight targets confirmed preferential binding of NSL proteins to the 5′-ends of genes (Figure 1C).

Given the similarity in binding between the subunits, subsequent analyses were based on the stringent core set of 2,841 genes that are bound by all four NSL proteins (thereafter called NSL-bound genes) unless otherwise indicated.

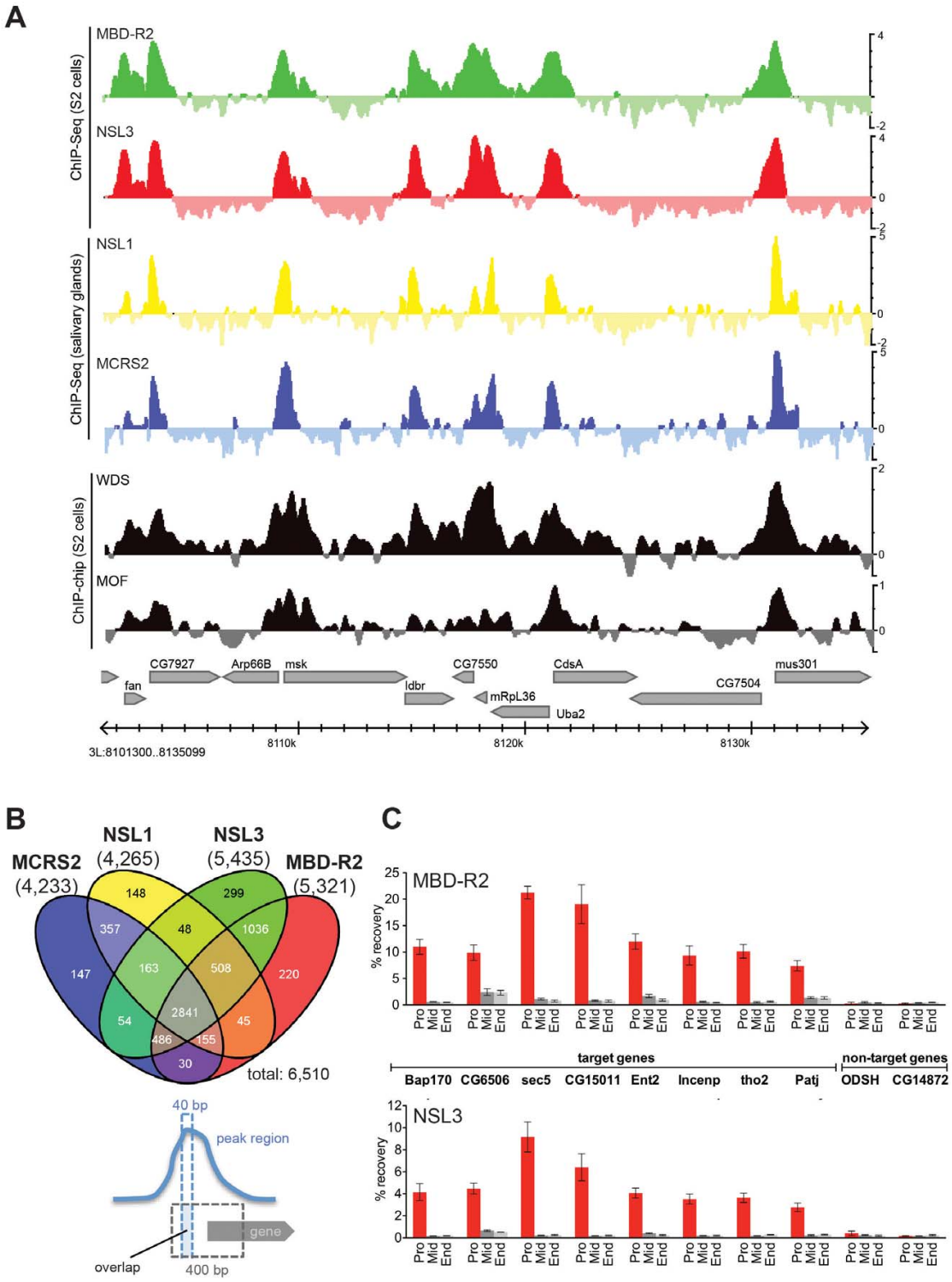


Figure 1. NSL proteins concomitantly bind to 5′ end of genes. (A) Genome Browser snapshot of a gene-rich region on chromosome 3 L. The log_2FCs (ChIP/input) of the newly generated ChIP-seq data of MBD-R2 and NSL3 are compared to those of NSL1, MCRS2 [21], WDS (GEO: GSE20835) and MOF (GEO: GSE27806). (B) The Venn diagram of NSL-bound TSS regions reveals an extensive set of promoters that are concomitantly bound by

all four NSL proteins. As indicated in the cartoon below the Venn diagram, a promoter was called NSL-bound if the 400 bp region surrounding the TSS (gray dashed lines) overlapped with the summit region of a peak determined by MACS and PeakSplitter (dashed blue lines). Using this definition, we identified a total of 6,510 TSSs bound by at least one NSL protein and 2,841 bound by all four. The numbers below the ChIP-ed protein names indicate the numbers of bound TSSs. (C) Chromatin immunoprecipitation followed by quantitative real-time PCR for a set of NSL target genes *(Bap170, CG6506, sec5, CG15011, Ent2, Incenp, tho2, Patj)* and non-target genes *(ODSH, CG14872)* confirm the results of the genome-wide ChIP-seq analyses: NSL proteins predominantly bind to the 5′ end of genes. Primers were designed to target the promoter (Pro), middle (Mid) and end (End) of genes; error bars represent standard deviations obtained from three independent experiments.
doi:10.1371/journal.pgen.1002736.g001

NSL complex targets are defined by an active chromatin state

We find that 68% and 66% of actively transcribed genes in S2 cells (based on expression analysis in [28]) are bound by NSL3 and MBD-R2, respectively (p-value<2.2e-16, Fisher's exact test); similar results were obtained for NSL1 and MCRS2 from salivary glands (Table S1, [21]). To assess the relationship between gene expression, chromatin state and NSL binding, we utilized the large set of histone modification data available from the modENCODE project (see Materials and Methods for accession numbers). Surprisingly, the patterns of histone acetylation and methylation markedly differed among expressed genes depending on the presence or absence of the NSL complex. While hallmarks of transcriptionally active promoters, H3K4me2, H3K4me3, H4K16ac and H3K9ac are present regardless of NSL binding, promoters that are bound by the NSL complex show an even greater enrichment of these marks compared with active promoters that lack NSL binding (Figure 2A). These enrichments of active histone marks cannot be explained by expression level differences between the two groups (Figure S2A).

The increased acetylation of H4K16 among NSL-bound genes is in agreement with the HAT activity of MOF. However, despite a recent report by Conaway and colleagues that showed that the human NSL/MOF complex can also catalyze H4K5 and H4K8 acetylation [19], we did not observe a similar enrichment of these histone marks on NSL-bound genes. One possible explanation is that the NSL/MOF complex in *Drosophila* may have different substrate-specificity for histone residues other than H4K16 compared with humans. Alternatively, since the H4K5 and H4K8 acetylation described above was detected using an *in vitro* system, these modifications may not arise from the primary activity of MOF *in vivo*. In summary, our results indicate that the NSL-complex-bound active genes are enriched for distinct sets of histone modifications when compared with active NSL-non-bound promoters.

To gain a more comprehensive understanding of the combinations of histone modifications found at NSL-bound promoters, we studied the distribution of NSL-bound and -non-bound promoters within the five principal chromatin types (chromatin colors) defined by the location maps of 53 chromatin proteins [29]. Within this model, the chromatin states "yellow" and "red" correspond to active genes, but differ in the combination of histone marks and chromatin binding proteins. Unexpectedly, we found a very significant enrichment of NSL-bound TSSs for the "yellow" chromatin state that is associated specifically with MRG15 and H3K36me3 (87.3% versus 18.2% for NSL-non-bound; p-value<2.2e-16; Fisher's exact test; see Figure 2B), but no comparable enrichment for the "red" chromatin state that is marked by chromatin proteins, such as Brahma, SU(VAR)2–10 and MED31 (9.7% of NSL-bound TSSs versus 8% of NSL-non-bound TSSs for "red"; Figure 2B). Our findings suggest the NSL complex as an additional, previously unknown marker of "yellow" chromatin while genes within "red" chromatin regions are expected to undergo NSL-independent transcriptional regulation.

A similar dominance for one specific state of active chromatin was observed when we repeated the analysis for the 9-chromatin-state model developed by Kharchenko and co-workers [30] (Figure 2C, Figure S2B), supporting the notion of the NSL complex as a regulator of a particular set of actively transcribed genes.

The NSL complex predominantly targets housekeeping genes

The results of the chromatin state analyses and the fact that most NSL binding appears to occur independently of the cell-type, prompted us to examine whether the complex displayed any association with housekeeping genes. To address this question, we defined a set of genes that are constitutively expressed throughout 30 distinct developmental stages of *Drosophila* [31] as our list of housekeeping genes (see Materials and Methods). We then generated heatmaps for Pol II and NSL binding centered on the TSSs of annotated genes [32] that were classified into three classes: constitutively expressed genes (see above), active genes but not expressed throughout all developmental stages of the fly [28,31] and inactive genes. As shown in Figure 2D, the Pol II signal shows the anticipated enrichment downstream of the TSSs of active genes regardless of constitutive or tissue-specific expression. In striking contrast, the NSL binding profiles show a very prominent, almost exclusive enrichment around the TSSs of constitutively expressed genes but not among those active genes that show tissue-specific regulation. Accordingly, 91.6% of NSL3-bound genes, 89.6% of MBD-R2-bound genes and 85.5% of TSSs bound by all four NSLs concomitantly belong to the group of housekeeping genes (Table S1, inlay in Figure 2D). Conversely, out of 5,534 constitutively expressed genes, 4,950 (89.4%; p-value<2.2e-16, Fisher's exact test) were bound by at least one NSL protein (Figure S2C, S2D). This number is likely to be an underestimation as some of the constitutively expressed genes, which are classified as NSL-non-bound according to our strict criteria, also show detectable NSL protein signals (Figure S2E). Taken together, we concluded that the NSL complex preferentially binds to constitutively expressed genes.

NSL-bound promoters have dispersed transcription initiation patterns and distinct nucleosome organization

In addition to expression-based definitions of housekeeping genes, we wanted to test further correlations of NSL binding with characteristics of constitutively expressed genes. Earlier studies have revealed two basic types of *Drosophila* promoters based on the pattern of the transcriptional initiation: broad and peaked [33–36]. While broad promoters preferably belong to housekeeping genes, peaked promoters are associated with tissue-specific expression. Based on data from [35], we found that NSL-bound TSSs are predominantly associated with dispersed transcription initiation patterns (Figure 3A).

We next wanted to investigate whether the NSL-characteristic initiation patterns and histone modifications enrichments also connected to specific structural features of the chromatin.

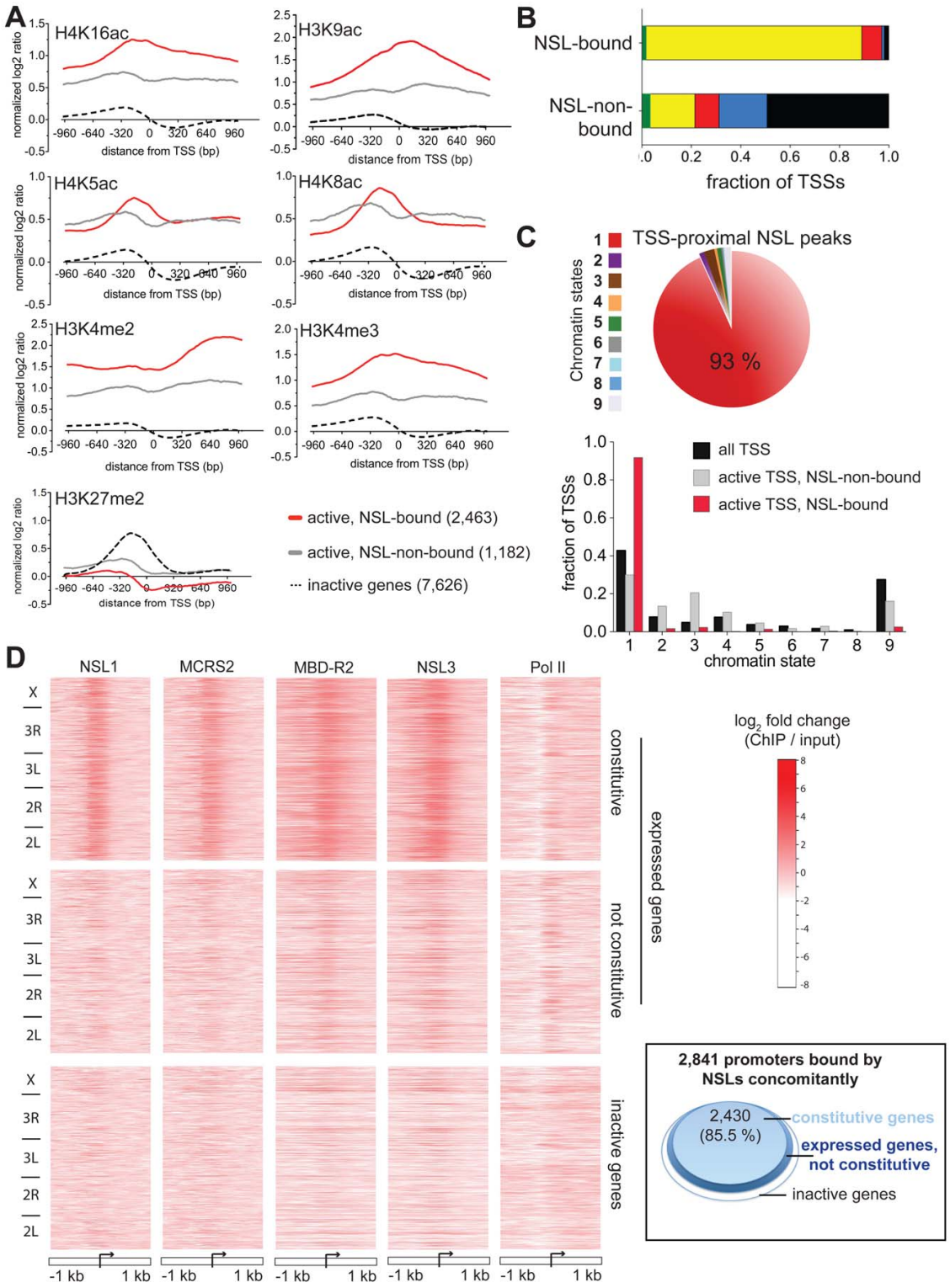

A
H4K16ac
H3K9ac
H4K5ac
H4K8ac
H3K4me2
H3K4me3
H3K27me2
normalized log2 ratio
distance from TSS (bp)
active, NSL-bound (2,463)
active, NSL-non-bound (1,182)
inactive genes (7,626)
B
NSL-bound
NSL-non-bound
fraction of TSSs
C
TSS-proximal NSL peaks
Chromatin states
93 %
all TSS
active TSS, NSL-non-bound
active TSS, NSL-bound
fraction of TSSs
chromatin state
D
NSL1
MCRS2
MBD-R2
NSL3
Pol II
constitutive
not constitutive
inactive genes
expressed genes
log2 fold change
(ChIP / input)
-1 kb
1 kb
2,841 promoters bound by
NSLs concomitantly
2,430
(85.5 %)
constitutive genes
expressed genes,
not constitutive
inactive genes

Figure 2. NSL proteins preferably associate with the promoters of constitutively active genes. (A) Metagene profiles of histone modifications reveal higher ratios of active chromatin marks H3K4me2/3, H4K16ac and H3K9ac for active genes bound by the NSL complex compared to active NSL-non-bound and inactive genes. On the contrary, the repressive mark H3K27me2 is not enriched on gene promoters bound by the NSL complex. Active genes were defined according to the expression data from [28] (see Materials and Methods). The expression levels of NSL-bound and NSL-non-bound active genes are similar (Figure S2A). The $\log_2$ ratios ($=\log_2$FC (ChIP/input)) of the histone modifications were obtained from modENCODE, extracted for 200 bp bins, and normalized to H4 Chip-chip signals. (B) The chromatin color model contains [29] two states of euchromatin: "yellow" and "red". NSL-bound TSSs are predominantly associated with "yellow", but not "red" chromatin. NSL-non-bound genes display chromatin color ratios that resemble the pattern seen by Filion et al. for tissue-specific genes. ("Green" and "blue" correspond to classical and PcG heterochromatin, respectively, while "black" denotes regions of repressive chromatin). (C) For a different model of chromatin states devised by Kharchenko et al., similar results as in Figure 2B were obtained: The pie chart depicts that 93% of all peaks of NSL1, MCRS2, NSL3 and MBD-R2 that localize within +/−200 bp of the nearest TSS associate with regions of chromatin state 1. This is defined as the state of actively transcribed TSSs [30]. Complementary, as shown in the bar chart, NSL-bound TSSs of expressed genes are significantly enriched in chromatin state 1 and depleted of chromatin state 9 (p-values<2.2e-16; binomial test) while NSL-non-bound genes are more equally distributed between states of active TSSs (1) and elongation marks (states 2, 3, 4). (D) Heatmaps of ChIP-seq signals ($\log_2$FCs) demonstrate the strong enrichment of NSL binding around the TSSs of constitutively transcribed genes. In contrast to the Pol II signal that is present in both constitutive and regulatory (not constitutive) active genes, the NSL proteins are predominantly found around the TSSs of constitutively transcribed genes. As indicated on the left-hand side, genes were sorted according to their genomic location. The proteins' binding intensities can be directly compared between the different panels. The inlay (right) illustrates the findings of the heatmap with a focus on genes that are bound by all NSLs concomitantly: 85.5% of NSL-bound promoters are constitutively expressed (light blue area). Active (not constitutive) and inactive genes are represented by dark blue and white areas, respectively.
doi:10.1371/journal.pgen.1002736.g002

Genome-wide analyses of nucleosome-positioning have demonstrated that transcriptionally active genes display a distinct organization, consisting of a precisely located +1 nucleosome around 135 bp downstream of the TSS, a −1 nucleosome that is directly upstream of the TSS and a nucleosome free region (NFR) between them. Additionally, it has been shown that the nucleosome organization can vary quite dramatically depending on the promoter sequences and transcription initiation patterns [37,38].

To assess whether NSL-bound promoters display a specialized nucleosome arrangement, we integrated a recently published map of nucleosome positions in S2 cells [38]. First, we examined the nucleosome occupancy for 4,950 constitutively expressed genes bound by at least one NSL protein, 717 constitutively expressed NSL-non-bound genes, and a set of 6,138 genes with tissue-specific expression (Figure S3; see Materials and Methods). For NSL-bound constitutively expressed genes we observe a well-defined nucleosome organization: Nucleosomes located within 200 bp upstream of the TSSs are strongly depleted while nucleosomes along the gene body are well positioned. In contrast, constitutively expressed genes not bound by the NSL complex (as well as tissue-specific genes) display a very different organization that is characterized by a less pronounced NFR and rather fuzzy positioning of the nucleosomes (Figure S3). This is in line with previous studies where more defined nucleosome positioning was associated with specific promoter sequences [37] and broad transcription initiation patterns [38].

The distinct nucleosome occupancies for NSL-bound genes prompted us to test if the observed difference in nucleosome positioning was related to gene expression levels. The analysis of

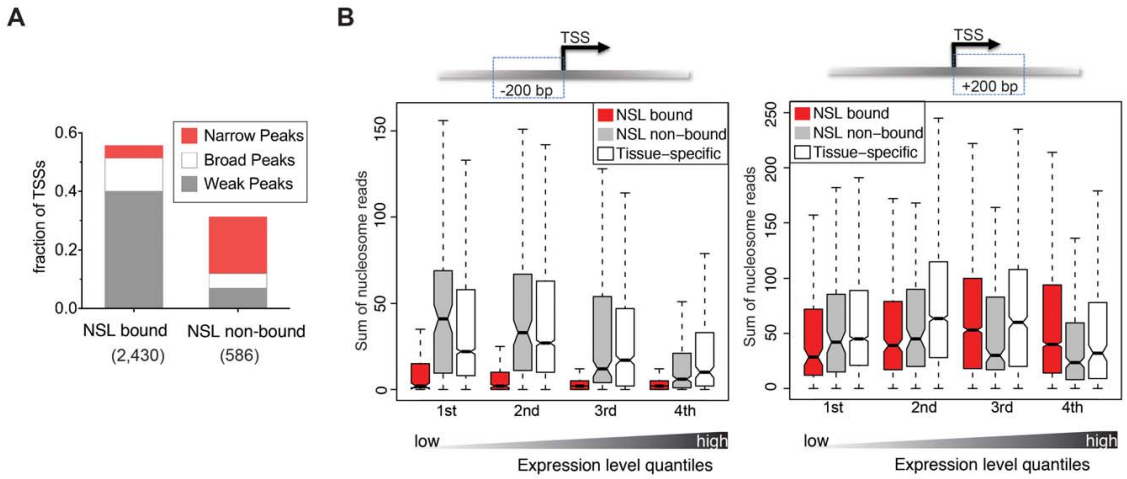


Figure 3. NSL-bound genes display a specific nucleosome organization at their TSS. (A) The TSSs of constitutively active genes, either NSL-bound or –non-bound, were analyzed regarding their reported transcription initiation patterns [35]. NSL-bound TSSs mostly belong to genes with weak and broad transcription initiation peaks (40% and 11.4%) whereas NSL-non-bound TSSs mainly belong to genes with narrow transcription initiation peaks (19.3%). (B) Boxplots of the sum of overlapping nucleosome reads in the regions 200 bp upstream and 200 bp downstream of the TSSs of constitutively expressed NSL-bound genes (red), constitutively expressed NSL-non-bound genes (gray), and tissue-specific genes (white). Genes were stratified based on their gene expression quartile (see Materials and Methods) which demonstrates that the depletion of nucleosomes immediately upstream of the TSS that we observed for NSL-bound housekeeping genes (left side) is independent of expression levels (p-values for −200 bp region <2.2e-16; Wilcoxon test).
doi:10.1371/journal.pgen.1002736.g003

the promoter proximal regions of NSL-bound, NSL-non-bound and tissue-specific genes revealed that the diminished nucleosome occupancy upstream of the TSS is, in fact, independent of the expression levels (Figure 3B).

NSL1 and NSL3 are required for efficient recruitment of Pol II on target promoters

Since the NSL complex predominately targets gene promoters, we next addressed whether its presence is important for the recruitment of RNA Polymerase II (Pol II). For this purpose, we first depleted NSL1, NSL3 and MBD-R2 in S2 cells by dsRNA-mediated depletion. The efficiency of the knockdown was assessed by Western blot analyses of nuclear or cytoplasmic extracts from the relevant cells (Figure S4A). Consistent with previous observations [21], NSL1 depletion had the most severe effect on the stability of NSL2, NSL3 and MCRS2. In contrast, MOF levels remained unaffected or at most showed a modest decrease upon MBD-R2 depletion. Interestingly, in comparison to the severe reduction of overall protein levels for NSL complex members, levels of Pol II, TBP and TFIIB showed almost no or only modest effects upon NSL1, NSL3 and MBD-R2 depletion.

We also assessed the quality of NSL1, NSL3 and MBD-R2 depletion by performing chromatin immunoprecipitation with NSL1, NSL3 and MBD-R2 antibodies in NSL-depleted versus control cells (dsRNA against GFP). Consistent with the Western blot analyses, the ChIP experiments revealed severe depletion of NSL1, NSL3 and MBD-R2 from target promoters (Figure S4B).

Following these quality criteria, we proceeded with genome-wide ChIP-seq analyses of Pol II in NSL1- and NSL3- depleted cells (Figure S5). As shown in Figure 4A, we obtained well-defined enrichments of Pol II binding at both the promoters and along the gene bodies of active genes in the GFP knockdown sample. The accumulation of Pol II at promoters is consistent with previous reports and indicative of widespread Pol II stalling [38]. When examining the global effects of the NSL knockdowns on Pol II levels, we observed a marked decrease in Pol II levels around transcription start sites (Figure 4B, 4C), particularly on genes that we had previously identified as bound by the NSL complex (Figure 4D). The loss of Pol II was even more pronounced in cells lacking NSL1 compared to those lacking NSL3. This effect could have been the consequence of different knockdown efficiencies of dsRNA against NSL1 and NSL3. Additionally, Western blot analyses of the individual NSL proteins revealed different effects of NSL1 and NSL3 depletion on NSL complex stability (see above and Figure S4A). Since protein levels of the other NSL complex members were either mildly affected or unaffected following the knockdown of NSL3, the remaining NSL complex members might have been able to partially continue transcriptional support in the absence of NSL3. This could explain the less severe effects of NSL3 depletion on Pol II binding compared to NSL1 depletion.

Regardless of the difference in the magnitude of Pol II reduction, both knockdowns showed greater effects on NSL-bound genes compared to NSL-non-bound active genes, suggesting that the NSL complex directly promotes the recruitment of Pol II to promoters of its target genes (Figure 4D, 4E). To assess whether the decrease of Pol II signal along the gene body could be attributed to elevated stalling of Pol II at the promoter, we calculated stalling indexes as described in [39] (see Materials and Methods). We could not detect a significant increase in the median stalling index (1.611 in GFP knockdown compared to 1.848 in NSL1 knockdown and 1.649 in NSL3 knockdown samples, p-value$>$0.1 as determined by Wilcoxon rank sum test, see Figure 4F). The unaffected stalling indexes suggest that NSL depletion does not interfere with the transition of Pol II from initiation to elongation. Taken together, these results strongly suggest that the NSL complex is required for efficient recruitment of Pol II at its target promoters.

NSL1, NSL3, and MBD-R2 are required for efficient recruitment of general transcription factors

Pol II recruitment to promoters is a multi-step process requiring the assembly of a functional pre-initiation complex (PIC). In the current model, the TFIID complex (containing TBP) first binds to core promoter regions where it is stabilized by TFIIA and TFIIB. TFIIF and Pol II are subsequently recruited to the core promoter by TFIIB [40,41]. Since we had established a general role of the NSL complex for Pol II recruitment, we now sought to identify the specific initiation step that was affected by NSL depletion.

Our next step was to perform ChIP-qPCR studies of individual NSL target genes following the knockdown of NSL1, NSL3 or MBD-R2. The results revealed that both TBP and TFIIB binding was decreased at promoters, indicating an interruption in the early stage of PIC assembly (Figure 5). In contrast to NSL complex members, TBP and TFIIB protein levels did not show a severe reduction upon NSL1 and NSL3 knockdown (Figure S4A). Consistent with previous observations [21], we did not detect a major difference in H4K16ac levels upon NSL1, NSL3 or MBD-R2 knockdown, possibly due to remaining MOF protein, or slow turnover of H4K16ac or the nucleosomes (Figure S6). Taken together, these data suggests that NSL1, NSL3 and MBD-R2 are required for efficient recruitment of TBP/TFIIB to target promoters presumably for efficient PIC formation.

DRE and motif 1 are associated with Pol II loss caused by NSL depletion

Distinct classes of gene expression patterns, e.g. constitutive or tissue-specific gene expression, are associated with particular promoter DNA motifs. Yet, how the presence or absence of a DNA motif is translated into biological functions often remains elusive. Since the NSL complex preferentially binds housekeeping genes, we wanted to investigate putative underlying DNA motifs and associate them with the effects of NSL depletion on Pol II recruitment.

We first assessed which motifs were enriched in NSL target regions: The unbiased *de novo* motif finder MEME repeatedly identified four known core promoter elements within NSL peak regions: the E-box motif (CAGCTG), DRE (WATCGATW), the reverse complement of a motif resembling DMv2 (TGGYAACR [42]) and motif 1 (YGGTCCACTR [43]; Figure 6A, Figure S7). Applying a quantitative model of transcription factor binding affinities (TRAP) to the 10 well-known *Drosophila* core promoter motifs [43,44], we detect a strong enrichment for DRE and E-box as well as Motifs 1, 6, 7, 8 in NSL-bound promoters compared with non-bound ones (p-values$<$0.0001, Wilcoxon rank sum test; Figure 6B, Figure S8). This is in complete concordance with previous genome-wide studies that suggested a preference of housekeeping genes for these motifs [34,35].

We have shown that the NSL complex is crucial for Pol II recruitment to housekeeping genes. However, Figure 4E reveals variability in the extent of Pol II loss among genes with high NSL binding signals. This is in line with the observation published by Becker and colleagues [45]. One possible explanation could be that different core promoter motifs underlie the variable responses of NSL-bound genes to NSL loss. We thus assessed whether the motif strengths is associated with the impact of NSL depletion on Pol II recruitment. For this purpose, we stratified NSL-bound and –non-bound genes into three subsets according to the magnitude

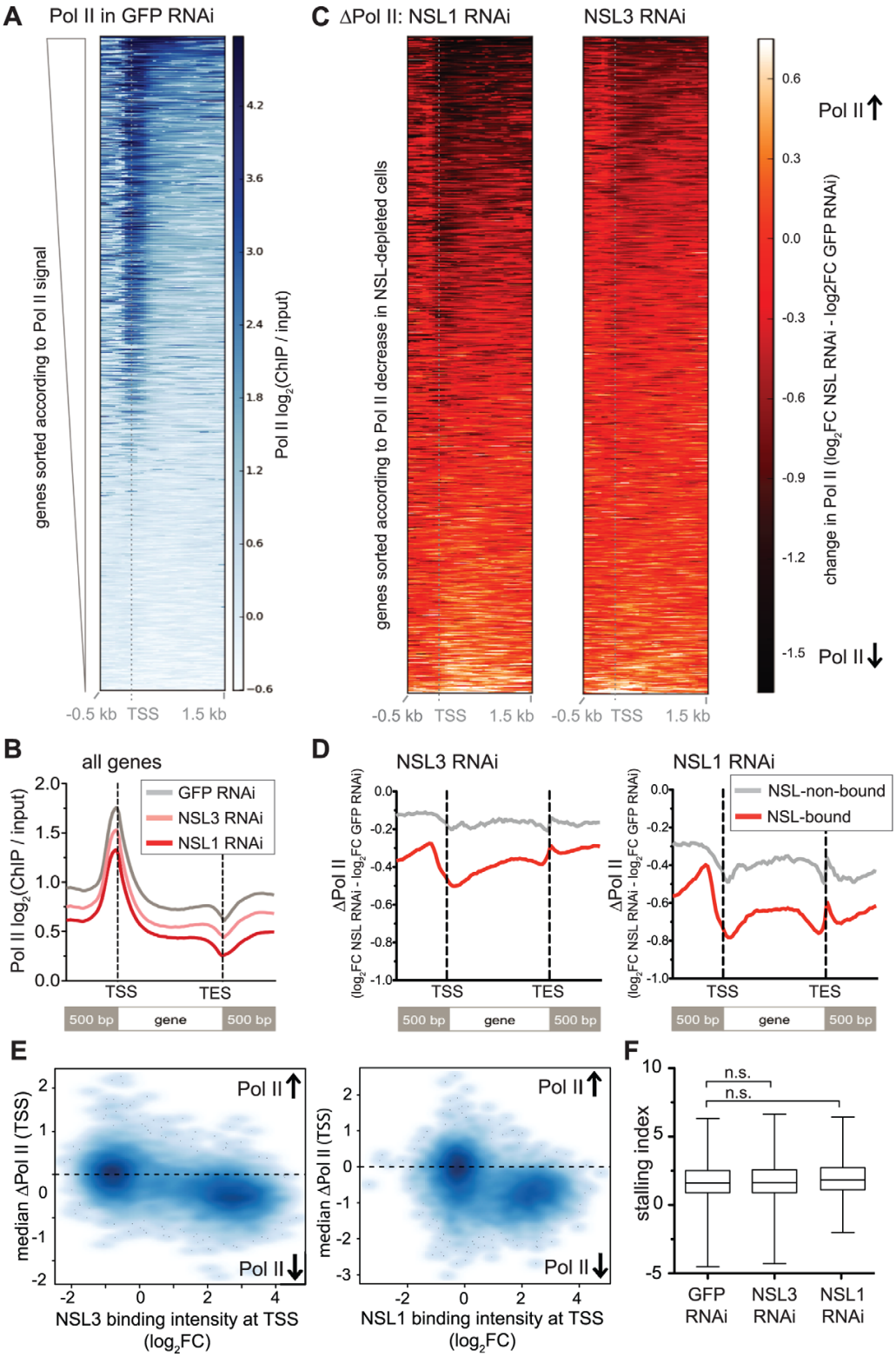

A
Pol II in GFP RNAi
genes sorted according to Pol II signal
Pol II log2(ChIP / input)
4.2
3.6
3.0
2.4
1.8
1.2
0.6
0.0
-0.6
-0.5 kb TSS 1.5 kb
C
ΔPol II: NSL1 RNAi
NSL3 RNAi
genes sorted according to Pol II decrease in NSL-depleted cells
change in Pol II (log2FC NSL RNAi - log2FC GFP RNAi)
Pol II ↑
Pol II ↓
0.6
0.3
0.0
-0.3
-0.6
-0.9
-1.2
-1.5
B
all genes
Pol II log2(ChIP / input)
GFP RNAi
NSL3 RNAi
NSL1 RNAi
TSS TES
500 bp gene 500 bp
D
NSL3 RNAi
NSL1 RNAi
ΔPol II (log2FC NSL RNAi - log2FC GFP RNAi)
NSL-non-bound
NSL-bound
E
median ΔPol II (TSS)
Pol II ↑
Pol II ↓
NSL3 binding intensity at TSS (log2FC)
NSL1 binding intensity at TSS (log2FC)
F
stalling index
n.s.
n.s.
GFP RNAi
NSL3 RNAi
NSL1 RNAi

Figure 4. NSL depletion leads to Pol II loss on target genes of the NSL complex. (A) The heatmap displays input-normalized Pol II binding signals for 5′ ends of *D. melanogaster* genes as captured by ChIP-seq of Rbp3 in S2 cells that had been treated with dsRNA against GFP. Genes were sorted according to the signal strength: Genes with high Pol II binding on promoters as well as along the gene bodies are found in the upper part of the heatmap. They are followed by genes with Pol II binding primarily at the promoter and genes lacking detectable Pol II signals. (B) Metagene profiles of the genome-wide signals of Pol II shows a marked decrease of Pol II binding for cells lacking NSL1 or NSL3 compared to control cells. (C) Here, the change of Pol II binding upon knockdown of NSL1 and NSL3 (ΔPol II) was visualized. The ΔPol II signal is calculated as the difference of normalized Pol II ChIP-seq signal (log_2FC) in NSL-depleted cells and control cells. Genes are ranked according to the change of Pol II in NSL knockdown; genes with greatest Pol II loss are found at the top of the heatmap. Severe reduction of Pol II after NSL depletion is seen around the TSSs and along gene bodies (dark red to black color), but there are also numerous genes that are slightly or not affected (bright red color). (D) Average ΔPol II values were plotted for active genes, separated into NSL-bound and –non-bound ones. The general decrease of Pol II upon NSL knockdown was observed again. In addition, it now becomes more evident that the magnitude of Pol II loss is markedly higher in NSL-bound genes compared to NSL-non-bound genes. (E) To study the association between the loss of Pol II (i.e. negative ΔPol II values) and NSL binding in an unbiased manner, median ΔPol II values at promoters were plotted against the median binding intensities of NSL1 and NSL3 from wild type samples. Genes were filtered for non-overlapping genes and those with significant Pol II binding in the control sample; the promoter region was defined as a 400 bp region centered around the TSS. The scatter plots confirm that genes with substantial NSL signals show markedly lower ΔPol II values than genes without NSL binding (left hand side of the plot). The difference of ΔPol II between NSL-bound and NSL-non-bound genes is statistically highly significant as determined by Wilcoxon rank sum test (p-value<2.2e-16). The observation that the majority of the genes with high NSL binding display a negative ΔPol II value (Pol II loss), suggests the NSL complex as a transcriptional activator whose binding to genes has functional consequences. (F) Stalling indexes for all genes with significant Pol II binding in control and NSL-depleted cells were calculated. Stalling indexes are derived from the ratio of Pol II at the promoter versus Pol II along the gene body (see Materials and Methods); high stalling indexes indicate Pol II accumulation at the promoter and diminished release into transcriptional elongation. No statistically significant difference between the stalling indexes of genes in the three different conditions was observed (median stalling indexes are 1.611 for GFP-RNAi treated cells, 1.649 in NSL3-RNAi treated cells and 1.848 in NSL1-RNAi, p-value>0.1, Wilcoxon rank sum test; n.s. = not significant).
doi:10.1371/journal.pgen.1002736.g004

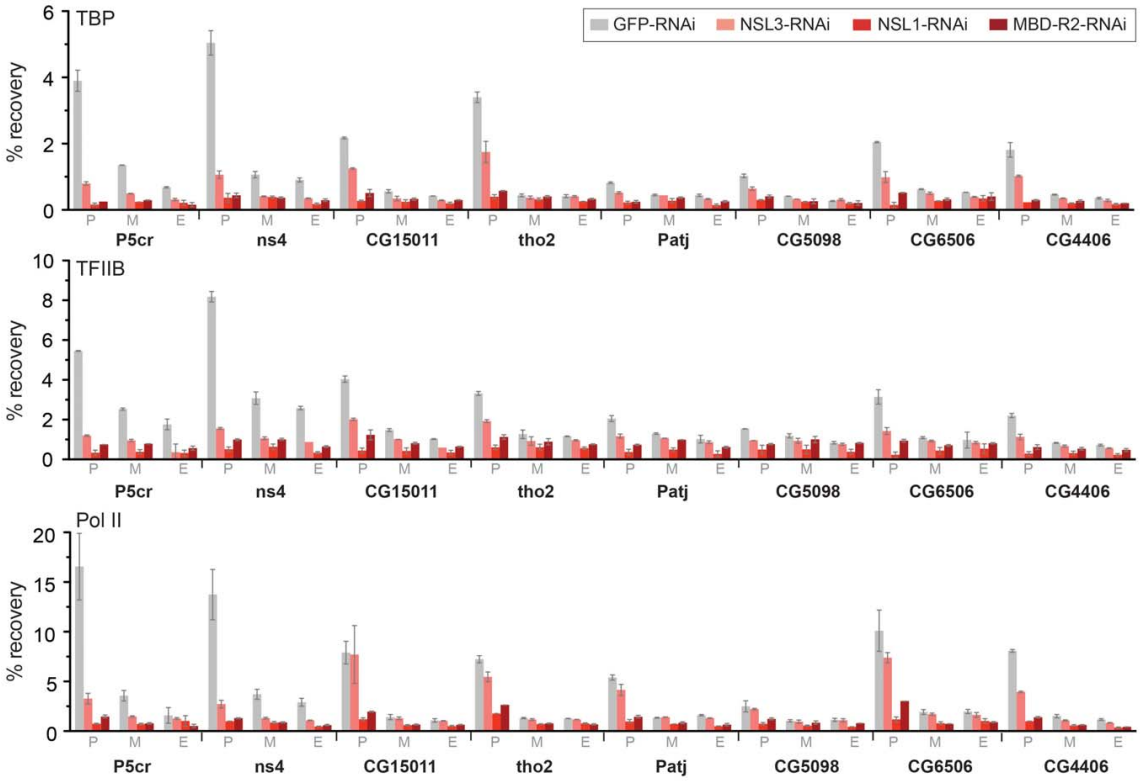


Figure 5. The NSL complex is important for optimal recruitment of the pre-initiation complex. ChIP was performed with antibodies against TBP, TFIIB and Pol II (Rpb3) in NSL1, NSL3 and MBD-R2 depleted S2 cells as well as in GFP knockdown control cells. The quantitative qPCR was performed on six autosomal genes (*P5cr*, *ns4*, *CG15011*, *tho2*, *Patj*, *CG5098*) as well as 2 X-linked genes (*CG6506* and *CG4406*). Primers were positioned at the promoter (P), middle (M) and end (E) of the indicated genes. Percentage recovery is determined as the amount of immunoprecipitated DNA relative to input DNA. Error bars represent the standard deviation between independent experiments.
doi:10.1371/journal.pgen.1002736.g005

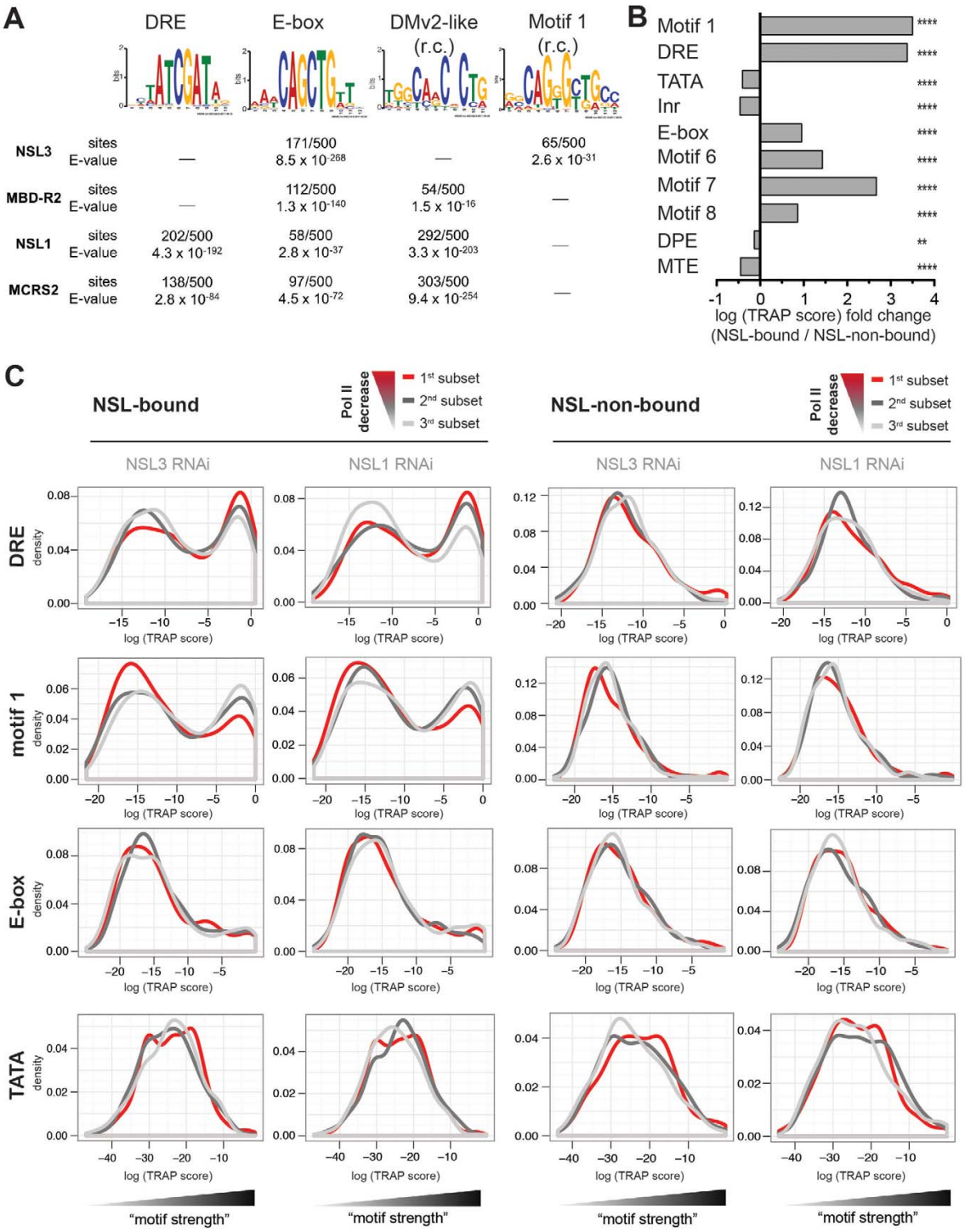


Figure 6. NSL target regions are enriched for housekeeping gene motifs, but only DRE and motif 1 are directly related to Pol II loss upon NSL depletion. (A) Individual *de novo* motif analysis led to the discovery of four non-repetitive DNA motifs that are located within NSL-complex binding sites (r.c. = reverse complement). The analysis was carried out by MEME [62,64] for 100 bp regions around the peak summits. As computational restrictions of MEME allowed only a limited number of base pairs to be analyzed at a time, the results of the 500 highest peak regions

are shown here (for additional peak regions see Figure S7). (B) Motif enrichments were calculated with TRAP [44,63] using the motif matrices for the 10 known core promoter motifs identified by [43]. In our study, the TRAP score can be seen as a measure for the affinity of transcription factors to bind to the DNA regions of interest. We compared the TRAP scores for NSL-bound and –non-bound promoter regions (TSS +/−200 bp) and found Ohler motifs 1, 6, 7, 8 as well as DRE and E-box significantly and selectively enriched in NSL-target regions while TATA box, Inr, DPE and MTE are depleted. The bar plot depicts the fold change between the median TRAP scores of NSL-bound versus –non-bound regions; individual frequency distributions of the motifs' TRAP scores can be seen in Figure S8 (for constitutitve gene promoters). P-values for the comparison of NSL-bound versus –non-bound promoters were calculated with two-sided Wilcoxon rank sum test, **** = P<0.0001, *** = P<0.001, ** = P<0.01, * = P<0.5. (C) To determine the significance of the Ohler motifs for the function of the NSL complex, genes were divided into three classes according to the magnitude of Pol II loss. The 1st subset (red line) corresponds to genes with the most severe Pol II reduction upon NSL knockdown while the 3rd subset (light gray line) contains least affected genes. Density distributions of TRAP scores were then plotted for NSL-bound and –non-bound genes for each Ohler motif individually. For DRE and motif 1 there is a clear distinction between the differently affected NSL-bound genes: NSL targets that lose Pol II binding most dramatically after NSL knockdown (red line) are clearly enriched for high DRE TRAP scores. In contrast, motif 1 shows an inverse pattern compared to DRE: NSL-bound genes with mild Pol II loss (light gray line) tend to contain strong motif 1 sites. This trend is not observed in NSL-non-bound genes. Other motifs such as E-box and TATA box also did not show significant association (also see Figure S9).
doi:10.1371/journal.pgen.1002736.g006

of Pol II loss on promoters and plotted the corresponding distribution densities for each motif's strength (Figure 6C, Figure S9).

Based on the equally strong enrichment of motif 1 and DRE (see Figure 6B) one might have expected a similar importance of these motifs for the function of the NSL complex. Interestingly, when we integrated the genome-wide Pol II binding data, we observed that DRE and motif 1 are associated with Pol II loss upon knockdown of NSL complex members in opposing manners: For the DRE motif we see a positive correlation between the levels of Pol II loss and the abundance of genes with high DRE TRAP scores. Motif 1, on the other hand, is mostly associated with genes that are least sensitive to Pol II loss after NSL depletion (light gray line in Figure 6C). For NSL-non-bound genes, neither DRE nor motif 1 show any enrichment in relation to Pol II loss. Enrichment of E-box and other core promoter motifs (except motif 7, Figure S9) do not exhibit a correlation with the sensitivity to NSL complex depletions.

In conclusion, our analysis demonstrates that NSL-bound promoters are enriched for core promoter motifs DRE, E-box and motif 1, 6, 7, 8 and depleted for TATA, Inr, DPE and MTE sequences. Even more importantly, the presence of DRE motifs is positively associated with the degree of responsiveness of NSL target genes to NSL complex depletion.

Summary

In this study, we have revealed that the majority of the NSL-complex-bound targets are housekeeping genes in *Drosophila*. While chromatin-modifying complexes that regulate tissue-specific genes, such as SAGA, polycomb and trithorax complexes, have been studied extensively, global regulators of housekeeping genes are poorly understood. To our knowledge, the NSL complex is the first identified major regulator of housekeeping genes which is consistent with a recently published study from Becker and colleagues [45].

The promoters of NSL target genes exhibit prominent enrichment of certain histone modifications (H4K16ac, H3K9ac, H3K4me2, H3K4me3) as well as specific core promoter elements (such as DRE, E-box and motif 1). Furthermore, these genes display distinct nucleosome occupancy and dispersed promoter configuration characterized by multiple transcription start sites. The correlation between these promoter characteristics (well-defined chromatin marks, TATA-less DNA sequences and broad initiation patterns) was previously identified for housekeeping genes in mammals and flies [36], but how these promoter features are translated into gene transcription had remained elusive. We now conclusively demonstrate that the NSL complex modulates transcription at the level of transcription initiation by facilitating pre-initiation complex loading onto promoters. Therefore, we propose that the NSL complex is a key *trans*-acting factor that bridges the promoter architecture, defined by the DNA sequence, histone marks and higher chromatin structures with transcription regulation of constitutive genes in *Drosophila* (Figure 7).

Excitingly, the enrichment of DNA motifs on NSL target gene promoters in combination with the genome-wide Pol II binding data has established functional links between the motifs enriched on housekeeping genes and the NSL-dependent Pol II binding to promoters. The abundance of DRE motifs, for example, was found to be positively associated with the magnitude of Pol II loss upon NSL knockdowns. The DRE binding factor (DREF) interacts tightly with TRF2 to modulate the transcription of DRE-containing promoters in a TATA-box-independent fashion [46]. It is tempting to speculate that the NSL complex might also cooperate with the TRF2 complex to facilitate transcription in a specific manner, rendering DRE-containing promoters more sensitive to NSL depletions. As the NSL-bound promoters are associated with a large variety of transcription factors, it will be of great interest to study whether the NSL complex communicates with different transcription regulators, perhaps making use of distinct mechanisms.

In contrast to DRE, motif 1 showed an opposing effect on Pol II recruitment to NSL-complex-bound genes as the presence of strong motif 1 sequences was associated with decreased Pol II loss upon NSL depletion. The mechanistic reasons for this remain unclear. However, one can envisage several possible scenarios. It is possible that motif 1 may recruit another transcription factor, which can also function to recruit the transcription machinery. Alternatively, the turnover of the transcription machinery might be slower on promoters containing strong motif 1 sequences. There is precedent for the transcription machinery having various turnover rates on different promoters. For example, in yeast, it has been shown that TBP turnover is faster on TATA-containing than on TATA-less promoters [47]. It is therefore possible that certain levels of the initiation complexes may still be maintained on motif-1-containing promoters, even though the recruitment of the transcription machinery will be compromised in the absence of NSL complex. Further work is required to understand the importance of sequence determinants for NSL complex recruitment and our analysis sets the grounds for targeted experiments in the future.

Taking MOF-mediated H4K16 acetylation into consideration, a putative role of the NSL complex might be to coordinate the opening of promoter architecture by histone acetylation and the assembly of PIC. Coupling of histone acetylation and PIC formation has been described before. For example, TAF1, a component of TFIID, is a histone aceyltransferase [48]. The SAGA complex, which contains Gcn5 and can acetylate H3K9, is reported to interact with TBP and other PIC components to regulate tissue-specific genes [49,50] and the recruitment of P300

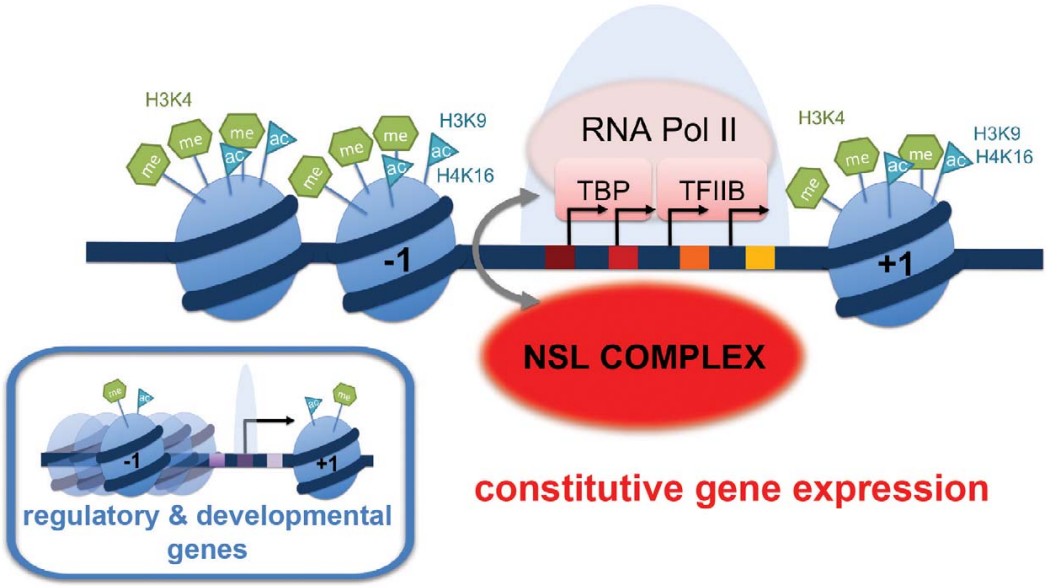


Figure 7. Summary model: NSL-dependent Pol II recruitment to promoters of housekeeping genes. The majority of the NSL-bound targets are constitutively expressed or "housekeeping" genes. These genes are characterized by prominent enrichment of particular histone modifications (H4K16ac, H3K9ac, H3K4me2, H3K4me3) as well as specific core promoter elements (such as DRE, E-box and motif 1; indicated by colored squares). In contrast, tissue-specific or developmentally regulated genes (small inlay) usually contain the TATA-box as the most prominent core promoter element. We propose that the NSL complex acts as a regulator of constitutively expressed genes by facilitating stable recruitment of the pre-initiation complex (PIC) members such as Pol II, TBP and TFIIB on target genes. NSL complex may therefore serve as an important link between specific promoter architecture and PIC assembly.
doi:10.1371/journal.pgen.1002736.g007

to the promoter and H3 acetylation have been shown to proceed binding of TFIID in a coordinated manner [51]. H4K16ac is also well-known for its role in transcription regulation of the male X chromosome, yet how H4K16 acetylation and PIC assembly are coordinated remains elusive. Interestingly, absence of the NSL complex does not severely abolish H4K16ac from target genes. Since the turnover of H4K16ac on target promoter is unknown, it remains possible that H4K16ac could remain for some time at the promoter after the NSL complex is depleted. Further studies will be crucial in unraveling the functional relevance of H4K16 acetylation and NSL complex function on housekeeping genes.

Materials and Methods

Chromatin immunoprecipitation (ChIP) and real-time PCR

Chromatin immunoprecipitation was carried out on S2 cells as previously described [21]. Fixed chromatin was sheared into 200 bp fragments and probed with antibodies against *Drosophila* TBP, TFIIB, Pol II, H4K16ac (sc8662, Santa Cruz), H4 (ab7311, Abcam), NSL1, MCRS2, NSL3 and MBD-R2 [21].

Real-time PCR validation was performed with SYBR-Green PCR master mix (Applied Biosystems) and an ABI7500 real-time PCR thermocycler (Applied Biosystems, Inc.). Recovery was determined as the amount of immunoprecipitated DNA relative to input DNA.

Deep sequencing of ChIP samples

Deep sequencing of NSL3 and MBD-R2 ChIP and input samples was carried out with the Illumina Genome Analyzer II, Pol II ChIP (from GFP-RNAi, NSL1-RNAi, NSL3-RNAi) and respective input samples were deep-sequenced with an Illumina HiSeq2000 machine according to manufacturer's instructions.

Mapping

The sequence reads from our earlier study of NSL binding in salivary glands [21] and the newly generated data from S2 cells were aligned to the *D. melanogaster* genome (dm3) using bowtie [52] with identical parameters. We allowed up to two mismatches and reported only the best alignments which could be aligned uniquely. We obtained 28,335,659 and 27,328,733 uniquely mapped reads for NSL3 and MBD-R2 respectively (input: 24,271,994 reads). The re-mapping of the NSL1 and MCRS2 data resulted in 7,622,096 and 9,405,874 unique reads (input: 6,168,473 reads).

From the samples sequenced with HiSeq 2000, we obtained between 120 to 135 million reads for Pol II ChIPs from S2 cells with knockdowns of NSL1, NSL3, and GFP and 50 to 60 million reads for the corresponding input samples. The correlations between the biological replicates of Pol II ChIP-seq reads from NSL1 and NSL3 knockdowns were excellent (Figure S5).

Peak calling

We used MACS Version 1.4.0rc2 on bed-files of mapped reads from the ChIP-seq experiments of NSL1, MCRS2, NSL3, MBD-R2, and an input control. We employed standard parameters for *D. melanogaster* (including model-building) and a p-value cut off of 10^{-5} [53]. We invoked PeakSplitter [54] as part of the MACS routine to obtain subpeak coordinates. For downstream analyses we used the subpeaks of peaks with a false discovery rate $\leq$5%.

Unless noted otherwise, peak summit regions were henceforth defined as the 40 bp region surrounding a summit identified by PeakSplitter.

Previous ChIP-seq analyses of Pol II have revealed that there are at least two types of Pol II signals: a sharp peak around the TSS of genes that can be either expressed or stalled, and an additional wide-spread region of moderate enrichment over the body of genes which is associated with transcription elongation, pausing and termination. The composite nature of the Pol II signal is not captured optimally by MACS, therefore we normalized the read counts (per 25 bp bins) of the Pol II ChIP-seqs with the BioConductor package DESeq [55], calculated the $\log_2$ fold changes ($\log_2$FC) between library-size-normalized input and ChIP samples and applied a 400 bp sliding window to account for the fragment size obtained after sonication. To determine regions of significant Pol II enrichment we modeled the distribution of $\log_2$FC values based on negative $\log_2$FCs that are assumed to correspond to experimental noise. We calculated the threshold $\log_2$FC values for significant Pol II binding within the three different conditions at an FDR-value cut-off of 0.05 (method described in more detail in [56]). The threshold $\log_2$FCs were 0.64, 0.84 and 1.39 for Pol II signals from GFP-RNAi, NSL3-RNAi, and NSL1-RNAi, respectively.

Lists of genes and associated NSL peaks

The basis of our gene-focused analyses was the list of annotated genes from FlyBase (version 5.30). Genes that are active in S2 cells were obtained from [28]. Data from [31] was used for a list of constitutively active genes: 5,534 genes expressed above a significance threshold (set by [31]) in all 30 developmental stages of *D. melanogaster* were considered constitutively expressed (housekeeping genes). To identify genes that were active in S2 cells, but not constitutively expressed, the gene identifiers of the different lists were adapted with the help of the FlyBase ID converter tool and subsequently subtracted from each other.

Unless indicated otherwise, a TSS was defined NSL-bound when the 400 bp region surrounding the TSS overlapped with an NSL peak summit region. The scripts, BEDTool commands and Galaxy workflows used for these overlaps and analyses are available upon request [57,58].

Calculation of Pol II stalling indexes

For the calculation of the stalling indexes we first applied stringent filters to the genes that were taken into account: we included only non-overlapping genes greater than 1,300 bp and with median Pol II signals above the threshold (see above) at the promoter region. Promoter regions were defined as TSS +/ −200 bp, for the gene body regions we excluded 500 bp after the TSS and 500 bp before the transcription end site (TES) to avoid confounding effects of transcription initiation and termination. Based on previous reports by Muse et al. [39], the stalling index (SI) itself was calculated as follows: $SI = \log_2(r\,(\mathrm{TSS})/r\,(\mathrm{gene\ body}))$ where r is the sum of Pol II ChIP-seq read counts that were adjusted by the input sample and normalized to the region's length.

Calculation of ΔPol II

To assess the change of Pol II upon NSL depletion in comparison to the GFP-RNAi control sample, ΔPol II was calculated as follows: $\Delta\mathrm{Pol\ II} = \log_2\mathrm{FC}\ (\mathrm{NSL\text{-}RNAi}) - \log_2\mathrm{FC}\ (\mathrm{GFP\text{-}RNAi})$.

Graphical representations

We visualized the binding profiles of the NSL complex proteins with our locally installed GBrowser (Version 2.15), uploading normalized $\log_2$FCs and wiggle files from modEncode.

For the summary plots of the histone marks (Figure 2A), we extracted the $\log_2$FCs from publicly available ChIP-chip data: The 2,000 bp TSS regions were split into 100 bins and the average $\log_2$FCs were calculated for each bin and normalized to the corresponding H4 signals.

For the heatmaps shown in Figure 2D we divided the annotated genes (FlyBase version 5.30) into active and inactive in S2 cells [28]. Active genes were further classified as constitutively and not-constitutively transcribed according to [31] (see above). For each gene, we extracted the normalized $\log_2$FCs (ChIP/input) from our ChIP-seq data (NSL1, MCRS2, NSL3, MBD-R2) and published ChIP-chip data of Pol II [32] in 50 bp bins for 1,000 bp up- and downstream of the TSSs. The heatmaps were generated with R using the same scale for every individual image and maintaining the order of the underlying TSS lists to enable direct comparisons between the different binding profiles on the same genes. Mitochondrial genes were excluded.

For the heatmaps of Pol II and ΔPol II (Figure 4A, 4C) we used all *D. melanogater* genes except mitochondrial genes. The $\log_2$FC of Pol II from the GFP-RNAi sample and ΔPol II (see above) for NSL3-RNAi and NSL1-RNAi were extracted in 50 bp bins for the regions 500 bp up- and 1,500 bp downstream of each gene's TSS. Genes were sorted according to the cumulative signal within the displayed region as indicated in the respective figures and legends.

For the metagene profiles of Pol II and ΔPol II signals as shown in Figure 4B and 4D, gene bodies of non-overlapping, size-filtered genes were scaled to the same length; $\log_2$FCs (ChIP/input) were extracted accordingly. Venn diagrams were generated with Venny [59].

Nucleosome occupancy analysis

We measured nucleosome occupancy for constitutively expressed NSL-bound genes, constitutively expressed NSL-non-bound genes and tissue-specific genes in a 200 bp area surrounding their annotated TSSs (4,971, 717 and 6,138 genes respectively). Nucleosome maps for S2 cells were obtained from GEO (accession number: GSE22119 [38]). NSL-bound genes for this analysis were defined as those bound by any of the NSL1, NSL3, MCRS2 or MBD-R2 subunits. Constitutive genes were defined as in [31] (see above). Tissue-specific genes were selected as in [60] on the basis of the 'gene scores' derived from Affymetrix tiling arrays for 25 different cell lines and 30 developmental stages (modENCODE accession number: modENCODE_3305). In order to avoid any bias in nucleosome organization due to differences in gene expression levels, genes were stratified in quartiles according to their expression value (ArrayExpress: E-MEXP-150 [61]). Finally, for each gene, nucleosome occupancy was calculated as the sum of overlapping reads with a 200 bp area up- and downstream its TSS.

Nucleosome metaprofiles were calculated using the average sum of overlapping reads with 25 bp bins spanning the area 500 bp up- and 1000 bp downstream of the TSS of each gene.

MEME

We sorted the peaks identified by MACS and PeakSplitter according to their summits' tag counts and extracted the DNA sequences for a 100 bp region centered around them. The peak

summits were analyzed by MEME [62] in subsequent analyses of 500 sequences each with the following parameters: revcomp, nmotifs = 3, minw = 6, maxw = 12, minsites = 10.

Motif enrichment: TRAP analysis

In addition to the *de novo* motif analysis by MEME, we studied sequence properties of NSL-targets using 10 motif matrices from the supplementary material of [43]. As transcription factors can bind to DNA with a range of affinities, we employed a biophysical model (TRAP [44,63]) that predicts the binding affinity for a motif in a given sequence fragment. We refer to the logarithm of this number as the TRAP score that indicates the strength of the putative protein-DNA interaction for each Ohler motif within a region of interest. The TRAP score enables us to quantitatively assess the corresponding binding affinities, i.e. we do not rely on the binary classification of motif presence or absence. Instead we are able to compare the "protein binding capacity" of different regions of interest.

We applied the TRAP model to NSL binding sites and promoter regions, which we defined as +/−200 bp around the TSS. To assess the localization of the binding signals, more precisely, the TRAP score was calculated for sliding windows of 40 bp over this region. The average TRAP scores for each window were then compared between specific sets of promoters regions (NSL-targets and -non-targets).

To assess the relation between ΔPol II and the TRAP score (Figure 6C, Figure S9), we focused on the promoter regions of non-overlapping genes with median Pol II signal ($\log_2$FC) above the threshold value in GFP-RNAi and ΔPol II below 0 (i.e., loss of Pol II upon NSL knockdown). Tertiles based on ΔPol II were determined with the quantile function of R. We used a 100 bp window around the TSS for TRAP score calculation for all motifs except TATA (40 to 20 bp upstream of the TSS), Inr (TSS +/−20 bp), and DPE (20 to 40 bp downstream of the TSS). Density plots were generated with the R package ggplot2.

Chromatin state associations

We downloaded the bed-files with the genomic coordinates of the 9-state-chromatin model of [30] for S2 cells and the chromatin color model of [29] from modENCODE and identified the number of peak summits (2 bp) or TSSs intersecting with the different states. We also divided the peak summits into three groups according to their overlap with annotated TSSs: proximal (within +/−200 bp), peripheral (between +/−201–800 bp) and distal subpeaks (farther away than 800 bp).

Data from public repositories

For the analysis of histone marks and non-histone chromosome proteins, we downloaded the wiggle-files of ChIP-chip experiments on S2 cells from modEncode/Gene Expression Omnibus.

H3K4me3-S2: GSE20787
H3K4me2-S2: GSE23470
H4K16ac-S2: GSE20799
H3K27me2-TJ.S2: GSE27790
H3K9ac-S2: GSE20790
H4K5ac-S2: GSE20800
H3K18ac-S2. GSE20775
MOF_Q4145.S2: GSE27806
WDS_Q2691.S2: GSE 20835
H4: repset.4620571
Pol II: GSM463297
Nucleosome maps: GSE22119

S2 gene expression data for nucleosome occupancy: E-MEXP-1505 [61]

Primers used for qPCR

L = forward primer, R = reverse primer.
CG6506-pro-L: GCCGATGTTTACCGACAATC
CG6506-pro-R: CATGGTTGGTTATCGGGACT
CG6506-Mid-L: ATCCGTGCCTAATGATACCG
CG6506-Mid-R: ACGGTTGGTGTGAACCAAAT
CG6506-end-L: ACAGTCAGCTCCCAGCAGAT
CG6506-end-R: AAAGTGGCGTGAAAGTTGCT
Sec5-pro-L: GCTGCTCAGCAAGGAGACTT
Sec5-Pro-R: CGGACGAGCATAAAAAGAGC
Sec5-mid-L: GAACTCCCATTGGCGATAAA
Sec5-mid-R: AAATGTCTGGCGAAATGTCC
Sec5-end-L: ATCAACGGCTTCATCTTTCG
Sec5-end-R: GCGTTTTCTTCCATTTTCCA
ODSH-Pro-L: CCCATTTTTCCCACTGACTG
ODSH-Pro-R: GGCGCGTACAAATGAAAAAT
ODSH-Mid-L: AAGATCCGCTAAGCGATGAA
ODSH-Mid-R: GCCAGGAGTTGAAGTTGGTC
ODSH-End-L: AGGCTCTCGTGGGGTAAAAT
ODSH-End-R: GAGCTCACCGATTTGTTTCC
CG15011-pro-L: CAGCCCTGGTATTCGATGTT
CG15011-pro-R: CTCATCTTGGATCGGATCGT
CG15011-Mid-L: CCTGCCACAAGGAACACTTT
CG15011-Mid-R: AGCTGCAACAAGCACAAATG
CG15011-end-L: ACACGGTGTTCTTCCAGTCC
CG15011-end-R: CGCTAAGGAACGTCGAAATC
CG14872_Pro_L: AATCGAGACATTCAGGCACTC
CG14872_Pro_R: TTCCCACACTGAAAAATCCA
CG14872_Mid_L: AAGAGCTTGAACAGCGGAAC
CG14872_Mid_R: GATACGCAAACCGGCATC
CG14872_End_L: TCACGCTCTAAAACCCCAGA
CG14872_End_R: CAGTACGGCATGGGCAAC
Patj_Pro_L: GAGTGCATAGGAGAGGGTAAACA
Patj_Pro_R: GTGGCGTTGGCACACTTT
Patj_Mid_L:CGTCGGTCACCACAATGA
Patj_Mid_R: TTATCCGCCAAGGGTACAAC
Patj_End_L: ACGCGGTTGCTAACTAATGG
Patj_End_R: ACTTCTGGCATCGTTTCTGAC
tho2_Pro_L: CCTCGGATCAGGTGGTACA
tho2_Pro_R: GTCACACTGGCGGAACTAACT
tho2_Mid_L: GGCCACATCCGTGTTTATGT
tho2_Mid_R: GCCAAGACACACTCGTCCA
tho2_End_L: GCTTCACAATGCACGGAAC
tho2_End_R: GAGGAGCGGCAGTACATCA
Ent2_Pro_L: CGTAACGGCACCCCTCAA
Ent2_Pro_R: ACCGCACCGCACTACAAG
Ent2_Mid_L: CCGCCATCCTAGTGCTGCT
Ent2_Mid_R: GCTGCTCCGGCTAATGGT
Ent2_End_L: TCTCGTATCTGGGACCATTTT
Ent2_End_R: TCCCGGAACTGGTATTGAG
Bap170_Pro_L: CCTGCTCGTGAATGCAACT
Bap170_Pro_R: GTGGCGTGAATGGGAAAC
Bap170_Mid_L: ACCCCCAGCATTGTTCCT
Bap170_Mid_R: CTTTCCTCAGACGCCACTTC
Bap170_End_L: ATGAACCGACACACGACTGA
Bap170_End_R: GCCGTAGCCGAGTAGGTGA
Incenp_Pro_L: GTTCTTTCCCTTACCATTTTCC
Incenp_Pro_R: GTTCCCGCCACTACCATCT
Incenp_Mid_L: GAGGACGAGTCGGTGGAG
Incenp_Mid_R: TTGAAAAGCTCATGTGTACGG
Incenp_End_L: GCCACGTAAGGGGAGAGG
Incenp_End_R: GTTCGGGAATATCTGCTTTAGG
ns4_Pro_L: GAGATGCCAACTTGTAGGTGATT
ns4_Pro_R: AAATACATGCAGAGACAGGAGGT

ns4_Mid_L: GCAAGGTGGTCAGCGTTAGT
ns4_Mid_R: GACTAGACCGGGACAATCACA
ns4_End_L: GACAGCGAGGATGAAGACGA
ns4_End_R: CAGCAGAGCAAACACGTTCC
CG5098-Pro-L: GGTCTTGTTTATGGGCGAAA
CG5098-Pro-R: GAGGGAAAGGCGACCTAATC
CG5098-Mid-L: GATGAGCCTCCCAAAAATCA
CG5098-Mid-R: GGCTACTTTGGCTGCTATGC
CG5098-End-L: GGGCATTTCGTAATCCAAGA
CG5098-End-R: TTTGGGGAAGGGAACCTAAC
p5cr_Pro_L: CACACCAAAGCTCAGAGGAGT
p5cr_Pro_R: CCGATTGCATGGGCGTAG
p5cr_Mid_L: GCGAGGGCTGCACTGTTT
p5cr_Mid_R: TGGACTCGGGCACCTGTT
p5cr_End_L: ATGTAATCCCCCGGAACA
p5cr_End_R: GCAAGAAGGATCGGGAATAA
CG4406-pro-R: TATCGACGGTCACACTGCTC
CG4406-mid-L: CCTGGAACTTGAGGAATCCA
CG4406-mid-R: GGCAGCAATGTGCTCATCTA
CG4406-end-L: AGCTCGGAAGGAAACTGTGA
CG4406-end-R: GTGACCAAAAAGCCCTTCAA

RNAi in S2 cells

RNAi of S2 cells was performed as described previously [21]. All knockdown cells were transfected with 10 μg dsRNA against NSL1, NSL3, MBD-R2 or GFP using Lipofectamine RNAiMAX (Invitrogen) and were harvested after 6 days. EGFP control RNAi experiments were performed in parallel.

RNAi sequences used to generate dsRNA for the following genes

NSL1:

T7-NSL1 sense: 5′- TTA ATA CGA CTC ACT ATA GGG AGA ATG GCC CCA GCG CTC ACA-3′

T7-NSL1 antisense: 5′- TTA ATA CGA CTC ACT ATA GGG AGA TGA ACT TGT GGC CAC TGC C-3′

NSL3:

T7-NSL3 sense: 5′- TTA ATA CGA CTC ACT ATA GGG AGA TCC TTG GCG ACT ACC TCA TC-3′

T7-NSL3 antisense: 5′- TTA ATA CGA CTC ACT ATA GGG AGA GTA CCA TTT CGG CCC CTA GTG-3′

MBD-R2:

T7-MBD-R2 sense: 5′- TTA ATA CGA CTC ACT ATA GGG AGA CGC TGG CCA CGT TTA TTA AG-3′

T7-MBD-R2 antisense: 5′- TTA ATA CGA CTC ACT ATA GGG AGA TTG AAG AGA AAA AGC TTG TAC GG-3′

EGFP:

T7-EGFP sense: 5′-TA ATA CGA CTC ACT ATA GGG AGG ATG GTG AGC AAG G

T7-EGFP antisense: 5′-TA ATA CGA CTC ACT ATA GGG AGG ATC GCG CTT CTC G

Accession numbers

All ChIP seq data is available in the ArrayExpress database (http://www.ebi.ac.uk/arrayexpress/) with the accession numbers listed below.

NSL1 and MCRS2 ChIP-Seq from salivary glands: E-MTAB-214

NSL3 and MBD-R2 ChIP-Seq from S2 cells: E-MTAB-1085

Pol II ChIP-Seq from S2 cells (GFP-RNAi, NSL1-RNAi, NSL3-RNAi): E-MTAB-1084

Supporting Information

Figure S1 General characteristics of NSL binding profiles. (A) ChIP-Seq peaks obtained from NSL profiles were classified according to their distance from the nearest annotated TSS. The bar chart shows that the majority of NSL binding events is closely associated with annotated TSSs: 68.7% of NSL3 peaks, 67% of MBD-R2 peaks, 81.5% of NSL1 peaks, and 76.1% of MCRS2 peaks localize within 800 bp up- or downstream of the nearest TSS. The schematic diagram below the bar chart visualizes our definitions: proximal peaks localize within +/−200 bp (dark blue), peripheral peaks between 201–800 bp (light blue) and distal peaks are farther away than 800 bp from a TSS (white). (B) The strongest signals of NSL binding are observed within 200 bp of annotated TSSs. This is shown by the box plot of tag counts of peak summits classified as TSS-proximal, -peripheral, or –distal (whiskers = 2.5–97.5 percentiles). (C) The lack of complete overlap of NSL target genes is mainly due to stringent criteria for defining target genes. In Figure 1B, 1,036 genes were shown as "bound by NSL3 and MBD-R2 only" and 357 genes as "bound by NSL1 and MCRS2 only". We therefore addressed whether these two groups could constitute gene sets that are specific for S2 cells or salivary glands. For this purpose, input-normalized ChIP-seq signals for the promoters for each group of genes were extracted, including those that are bound by all or neither NSL proteins. The box plot shows that the signal of NSL1 and MCRS2 is still significantly higher in those genes that were labeled as "bound by NSL3 and MBD-R2 only" than for those that were defined as NSL-non-bound (p-value<2.2e-16, Wilcoxon test). The same holds true for NSL3 and MBD-R2. Therefore, differences in gene sets are very likely not due to tissue-specific binding, rather to the choice of a very stringent cut-off for the binary decision "bound" or "not-bound". For details about our definition of NSL target genes, see Materials and Methods and Figure 1B.
(PDF)

Figure S2 Assessing the overlaps of NSL signals on gene promoters. (A) Median expression levels between expressed genes that are bound by all four NSLs concomitantly do not differ significantly from expressed genes devoid of NSL binding as shown by the box plot (whisker = 2.5–97.5 percentiles). The expression scores were taken from [28]. (B) The NSL complex preferentially binds to regions of open and actively transcribed chromatin (state 1, [30]) as peak summits intersected with the regions reported by [30] are dramatically enriched for state 1 (regardless of their localization). (C) Overview of TSS-associated NSL binding: 19.25% of annotated TSSs are bound by NSL1, MCRS2, NSL3, and MBD-R2 concomitantly. When looking at the subsets of active and housekeeping genes, the numbers increase to 37.1% (active) and 43.9% (constitutive) that are bound by all four NSLs across different cell types and experiments. To confirm the findings that were based on our own definition of housekeeping genes (see Materials and Methods), we also tested a previously published set of broadly and restrictedly expressed genes [65]. (D) The Venn diagram shows the individual overlaps of the gene promoters bound by the single NSL proteins. The core intersect (2,430) corresponds to the gray bar of "constitutive genes" in Figure S2C, while the total number of 4,950 represents the number of constitutive TSSs bound by at least one NSL. (E) Constitutive genes classified as NSL-non-bound according to our criteria described in Materials and Methods (see Figure 1 for visualization) show slightly, but significantly elevated levels of NSL binding compared to non-constitutively expressed genes. This verifies the preference of the NSL complex for housekeeping genes

and suggests that some constitutive genes classified as NSL-non-bound were missed due to the cut-off we used for all four samples. The boxplot shows the median $\log_2$FCs (ChIP/input) for the 400 bp regions centered around TSSs. The medians were calculated for each gene based on the ChIP-seq tags of all four analyzed NSL proteins.
(PDF)

Figure S3 NSL-bound and NSL-non-bound housekeeping genes display different nucleosome organizations. Nucleosome occupancy metaprofiles for NSL-bound (red), constitutively expressed NSL-non-bound (gray) and tissue-specific (black) genes. Metaprofiles were calculated for each group as the sum of nucleosome reads overlapping 25 bp bins spanning the −500/+1000 bp region centered at the TSS of each gene. The non-shaded white area corresponds to the −200/+200 bp region used for the analysis in Figure 3B.
(PDF)

Figure S4 Depletion of different NSL proteins have distinct effects on the stability of the remaining NSL complex members but not for Pol II machinery components. (A) Western blot analyses of cytoplasmic (C) and nuclear (N) extracts from S2 cells that had been treated with dsRNA against GFP, MBD-R2, NSL1, and NSL3. Depletion of NSL1 greatly affects the stability of other NSL complex proteins namely: NSL2, NSL3, MCRS2, MBD-R2 and WDS. Depletion of NSL3 or MBD-R2 has milder effects on the levels of other NSL proteins. MOF protein levels appear affected upon MBD-R2 depletion but not in NSL1 or NSL3 knockdowns. In contrast, TBP, TFIIB and Pol II are only modestly affected in either knockdown especially when taking into consideration the loading control Nuclear RNA export factor 1 (NXF1). (B) To check whether the dsRNA treatment against NSL3, NSL1, and MBD-R2 efficiently reduced NSL binding to its target regions, ChIP was performed with antibodies against NSL1, NSL3 and MBD-R2 in the respective knockdowns in S2 cells. GFP-RNAi was used as a control. "P", "M", "E" represent promoter, middle and end of gene, respectively. Error bars represent the standard deviation of three independent experiments.
(PDF)

Figure S5 Correlation of biological duplicates for the ChIP-seq of Pol II in knockdowns of NSL1 and NSL3. Correlation plots between the two Pol II ChIP-seq libraries generated from duplicate knockdown experiments for (a) NSL3 and (b) NSL1. Reads were mapped to the genome with bowtie. The read counts plotted here were extracted for 25 bp bins along the entire *D. melanogaster* genome. The Spearman correlations for the biological replicates are excellent (0.96 for NSL3-RNAi samples, 0.97 for NSL1-RNAi samples).
(PDF)

Figure S6 Chromatin immunoprecipitation of H4K16ac in NSL1, NSL3 and MBDR2 depleted cells. ChIP-qPCR was performed using antibodies against H4K16ac and H4 in NSL1, NSL3 or MBD-R2 depleted cells. The H4K16ac signal is normalized against H4 signal from the same region. Consistent with our previous results, H4K16ac is very modestly reduced upon depletion of NSL complex members. The quantitative qPCR was performed on 5 autosomal genes (*P5cr*, *ns4*, *CG15011*, *tho2*, *Patj*, *CG5098*) as well as 2 X-linked genes (*CG6506* and *CG4406*). Primers were positioned at the promoter of the indicated genes. Error bars represent the standard deviation of three independent experiments.
(PDF)

Figure S7 *De novo* motif identification in NSL binding regions. (A) Motifs identified by MEME in peak summit regions of NSL3, MBD-R2, NSL1, and MCRS2. Results of MEME analyses of peaks ranked 501–1,000 (r.c. = reverse complement) confirm the motifs identified in the 500 highest peaks as shown in Figure 6A. (B) Results of MEME analyses of 500 peak summits that were not selected solely according to their height, but also on the basis of their association with constitutively expressed genes. The motifs and their occurrences recapitulate the results from the analysis of the highest intensity peaks (Figure 6A), reinforcing the preference of NSL targeting to genomic regions containing the motifs shown above.
(PDF)

Figure S8 Comparison of motif enrichments for NSL-bound and –non-bound constitutively active TSSs. (A) The bar chart displays the fold change of the core promoter affinities for the sequences of NSL-bound promoters (concomitant binding of NSL1, MCRS2, MBD-R2, NSL3) compared to NSL-non-bound promoters (not bound by any of the NSL proteins). The bar chart shows that even when motif enrichments are calculated within the subset of constitutively active genes, NSL-bound promoters are enriched for motif 1, DRE, E-box, motif 6, 7 and 8 whereas the depletion of TATA box, Inr motif, DPE and MTE becomes less evident. P-values were calculated with two-sided Wilcoxon rank sum test, **** = $P<0.0001$, *** = $P<0.001$, ** = $P<0.01$, * = $P<0.5$, not significant (n.s.) = $P>0.5$. The fold change was calculated as log(median(TRAP score of NSL-bound promoters)/median(TRAP score of NSL-non-bound promoters)). (B) Individual TRAP score [44,63] histograms for the 10 core promoter motifs [43] that underlie the bar chart of S6A. The histograms show the distributions of the motif affinities for NSL-bound and –non-bound promoters of housekeeping genes. The visible shifts towards higher or lower TRAP scores in NSL-bound or –non-bound genes, respectively, represent the fold changes seen in the bar chart.
(PDF)

Figure S9 TRAP score densities of NSL-bound and –non-bound genes. We selected non-overlapping genes that showed significant Pol II binding in control samples and reduced Pol II levels in NSL knockdown conditions and sorted them into three groups according to the magnitude of Pol II loss. The 1st subset (red line) contains genes with the strongest reduction of Pol II in promoter regions; the 3rd subset (gray line) correspondingly contains genes with smallest Pol II loss. TRAP score was calculated as a measure of protein binding affinity towards the known promoter motifs identified by Ohler et al. [43]. Of the motifs shown here, only motif 7 displays a moderate association between the motif's strength and Pol II loss upon NSL depletion (for remaining motifs see Figure 6C).
(PDF)

Table S1 Overview of the numbers of NSL peaks overlapping with annotated TSSs. The numbers of peaks (regions of significant NSL binding signals) determined by MACS [53] are comparable between the different proteins. However, the increased sequencing depth of the NSL3 and MBD-R2 ChIP-seq experiments led the detection of expanded regions of NSL binding signals that is reflected by more widespread peak. The PeakSplitter algorithm [54] divides peaks identified by MACS at sites of local maxima.
(PDF)

Acknowledgments

We thank J. T. Kadonaga for providing the *Drosophila* TBP and TFIIB antibodies. We thank A. G. Ladurner for providing the anti-RBP3 antibody. We thank the members of the lab for helpful discussions and

critical reading of the manuscript; F. M. is particularly indebted to F. Cavalli and F. Ramirez. We are grateful to the EMBL gene core facility and the Next Generation Sequencing Group in the department of Martin Vingron at the Max-Planck-Institute for Molecular Genetics, Berlin, for deep sequencing.

Author Contributions

Conceived and designed the experiments: AA. Performed the experiments: KCL SJR HH. Analyzed the data: KCL FM JMV TM NML AA. Contributed reagents/materials/analysis tools: AA NML TM. Wrote the paper: KCL FM JMV NML TM AA.

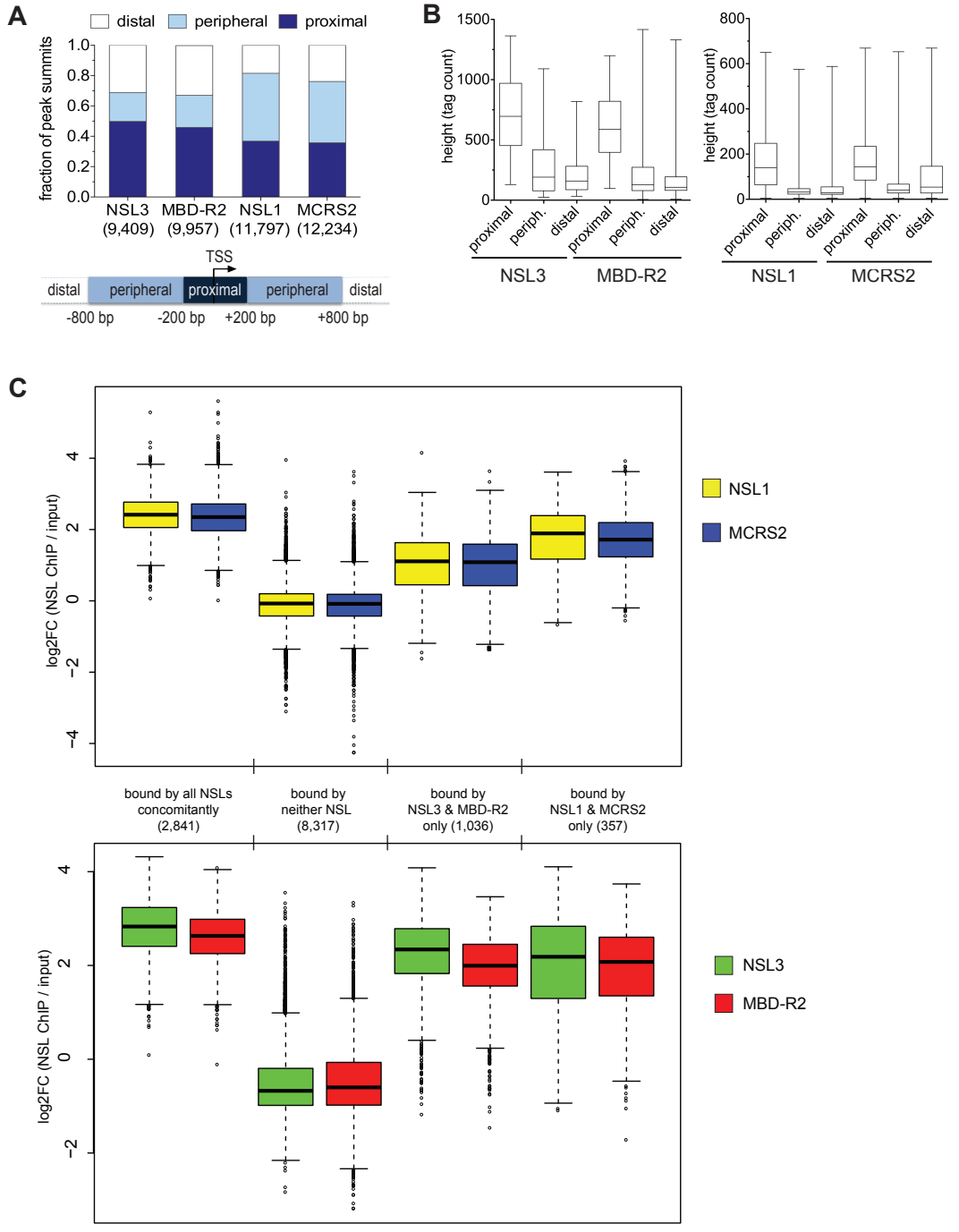
References

1. Juven-Gershon T, Kadonaga JT (2010) Regulation of gene expression via the core promoter and the basal transcriptional machinery. Dev Biol 339: 225–229.
2. Kouzarides T (2007) Chromatin modifications and their function. Cell 128: 693–705.
3. Yang XJ, Seto E (2007) HATs and HDACs: from structure, function and regulation to novel strategies for therapy and prevention. Oncogene 26: 5310–5318.
4. Jacobson RH, Ladurner AG, King DS, Tjian R (2000) Structure and function of a human TAFII250 double bromodomain module. Science 288: 1422–1425.
5. Ruthenburg AJ, Li H, Milne TA, Dewell S, McGinty RK, et al. (2011) Recognition of a mononucleosomal histone modification pattern by BPTF via multivalent interactions. Cell 145: 692–706.
6. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 389: 251–260.
7. Shogren-Knaak M, Ishii H, Sun JM, Pazin MJ, Davie JR, et al. (2006) Histone H4-K16 acetylation controls chromatin structure and protein interactions. Science 311: 844–847.
8. Shahbazian MD, Grunstein M (2007) Functions of site-specific histone acetylation and deacetylation. Annu Rev Biochem 76: 75–100.
9. Vetting MW, LP SdC, Yu M, Hegde SS, Magnet S, et al. (2005) Structure and functions of the GNAT superfamily of acetyltransferases. Arch Biochem Biophys 433: 212–226.
10. Utley RT, Cote J (2003) The MYST family of histone acetyltransferases. Curr Top Microbiol Immunol 274: 203–236.
11. Lee KK, Workman JL (2007) Histone acetyltransferase complexes: one size doesn't fit all. Nat Rev Mol Cell Biol 8: 284–295.
12. Grant PA, Schieltz D, Pray-Grant MG, Steger DJ, Reese JC, et al. (1998) A subset of TAF(II)s are integral components of the SAGA complex required for nucleosome acetylation and transcriptional stimulation. Cell 94: 45–53.
13. Suganuma T, Gutierrez JL, Li B, Florens L, Swanson SK, et al. (2008) ATAC is a double histone acetyltransferase complex that stimulates nucleosome sliding. Nat Struct Mol Biol 15: 364–372.
14. Huisinga KL, Pugh BF (2004) A genome-wide housekeeping role for TFIID and a highly regulated stress-related role for SAGA in Saccharomyces cerevisiae. Mol Cell 13: 573–585.
15. Lebedeva LA, Nabirochkina EN, Kurshakova MM, Robert F, Krasnov AN, et al. (2005) Occupancy of the Drosophila hsp70 promoter by a subset of basal transcription factors diminishes upon transcriptional activation. Proc Natl Acad Sci U S A 102: 18087–18092.
16. Krebs AR, Demmers J, Karmodiya K, Chang NC, Chang AC, et al. (2010) ATAC and Mediator coactivators form a stable complex and regulate a set of non-coding RNA genes. EMBO Rep 11: 541–547.
17. Nagy Z, Riss A, Fujiyama S, Krebs A, Orpinell M, et al. (2010) The metazoan ATAC and SAGA coactivator HAT complexes regulate different sets of inducible target genes. Cell Mol Life Sci 67: 611–628.
18. Suganuma T, Mushegian A, Swanson SK, Abmayr SM, Florens L, et al. (2010) The ATAC acetyltransferase complex coordinates MAP kinases to regulate JNK target genes. Cell 142: 726–736.
19. Cai Y, Jin J, Swanson SK, Cole MD, Choi SH, et al. (2010) Subunit composition and substrate specificity of a MOF-containing histone acetyltransferase distinct from the male-specific lethal (MSL) complex. J Biol Chem 285: 4268–4272.
20. Mendjan S, Taipale M, Kind J, Holz H, Gebhardt P, et al. (2006) Nuclear pore components are involved in the transcriptional regulation of dosage compensation in Drosophila. Mol Cell 21: 811–823.
21. Raja SJ, Charapitsa I, Conrad T, Vaquerizas JM, Gebhardt P, et al. (2010) The nonspecific lethal complex is a transcriptional regulator in Drosophila. Mol Cell 38: 827–841.
22. Straub T, Becker PB (2007) Dosage compensation: the beginning and end of generalization. Nat Rev Genet 8: 47–57.
23. Hallacli E, Akhtar A (2009) X chromosomal regulation in flies: when less is more. Chromosome Res 17: 603–619.
24. Conrad T, Akhtar A (2011) Dosage compensation in Drosophila melanogaster: epigenetic fine-tuning of chromosome-wide transcription. Nat Rev Genet 13: 123–134.
25. Andersen DS, Raja SJ, Colombani J, Shaw RL, Langton PF, et al. (2010) Drosophila MCRS2 associates with RNA polymerase II complexes to regulate transcription. Mol Cell Biol 30: 4744–4755.
26. Prestel M, Feller C, Straub T, Mitlohner H, Becker PB (2010) The activation potential of MOF is constrained for dosage compensation. Mol Cell 38: 815–826.
27. Li X, Wu L, Corsa CA, Kunkel S, Dou Y (2009) Two mammalian MOF complexes regulate transcription activation by distinct mechanisms. Mol Cell 36: 290–301.
28. Cherbas L, Willingham A, Zhang D, Yang L, Zou Y, et al. (2011) The transcriptional diversity of 25 Drosophila cell lines. Genome Res 21: 301–314.
29. Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, et al. (2010) Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell 143: 212–224.
30. Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, et al. (2011) Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. Nature 471: 480–485.
31. Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, et al. (2011) The developmental transcriptome of Drosophila melanogaster. Nature 471: 473–479.
32. Nechaev S, Fargo DC, dos Santos G, Liu L, Gao Y, et al. (2010) Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. Science 327: 335–338.
33. Rach EA, Yuan HY, Majoros WH, Tomancak P, Ohler U (2009) Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome. Genome Biol 10: R73.
34. Hoskins RA, Landolin JM, Brown JB, Sandler JE, Takahashi H, et al. (2011) Genome-wide analysis of promoter architecture in Drosophila melanogaster. Genome Res 21: 182–192.
35. Ni T, Corcoran DL, Rach EA, Song S, Spana EP, et al. (2010) A paired-end sequencing strategy to map the complex landscape of transcription initiation. Nat Methods 7: 521–527.
36. Rach EA, Winter DR, Benjamin AM, Corcoran DL, Ni T, et al. (2011) Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. PLoS Genet 7: e1001274. doi:10.1371/journal.pgen.1001274.
37. Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, et al. (2008) Nucleosome organization in the Drosophila genome. Nature 453: 358–362.
38. Gilchrist DA, Dos Santos G, Fargo DC, Xie B, Gao Y, et al. (2010) Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. Cell 143: 540–551.
39. Muse GW, Gilchrist DA, Nechaev S, Shah R, Parker JS, et al. (2007) RNA polymerase is poised for activation across the genome. Nat Genet 39: 1507–1511.
40. Juven-Gershon T, Hsu JY, Theisen JW, Kadonaga JT (2008) The RNA polymerase II core promoter - the gateway to transcription. Curr Opin Cell Biol 20: 253–259.
41. Pugh BF (1996) Mechanisms of transcription complex assembly. Curr Opin Cell Biol 8: 303–311.
42. FitzGerald PC, Sturgill D, Shyakhtenko A, Oliver B, Vinson C (2006) Comparative genomics of Drosophila and human core promoters. Genome Biol 7: R53.
43. Ohler U, Liao GC, Niemann H, Rubin GM (2002) Computational analysis of core promoters in the Drosophila genome. Genome Biol 3: RESEARCH0087.
44. Thomas-Chollier M, Hufton A, Heinig M, O'Keeffe S, Masri NE, et al. (2011) Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. Nat Protoc 6: 1860–1869.
45. Feller C, Prestel M, Hartmann H, Straub T, Soding J, et al. (2012) The MOF-containing NSL complex associates globally with housekeeping genes, but activates only a defined subset. Nucleic Acids Res 40: 1509–1522.
46. Hochheimer A, Zhou S, Zheng S, Holmes MC, Tjian R (2002) TRF2 associates with DREF and directs promoter-selective gene expression in Drosophila. Nature 420: 439–445.
47. van Werven FJ, van Teeffelen HA, Holstege FC, Timmers HT (2009) Distinct promoter dynamics of the basal transcription factor TBP across the yeast genome. Nat Struct Mol Biol 16: 1043–1048.
48. Mizzen CA, Yang XJ, Kokubo T, Brownell JE, Bannister AJ, et al. (1996) The TAF(II)250 subunit of TFIID has histone acetyltransferase activity. Cell 87: 1261–1270.
49. Sermwittayawong D, Tan S (2006) SAGA binds TBP via its Spt8 subunit in competition with DNA: implications for TBP recruitment. EMBO J 25: 3791–3800.
50. Warfield L, Ranish JA, Hahn S (2004) Positive and negative functions of the SAGA complex mediated through interaction of Spt8 with TBP and the N-terminal domain of TFIIA. Genes Dev 18: 1022–1034.
51. Black JC, Choi JE, Lombardo SR, Carey M (2006) A mechanism for coordinating chromatin modification and preinitiation complex assembly. Mol Cell 23: 809–818.
52. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.
53. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, et al. (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9: R137.

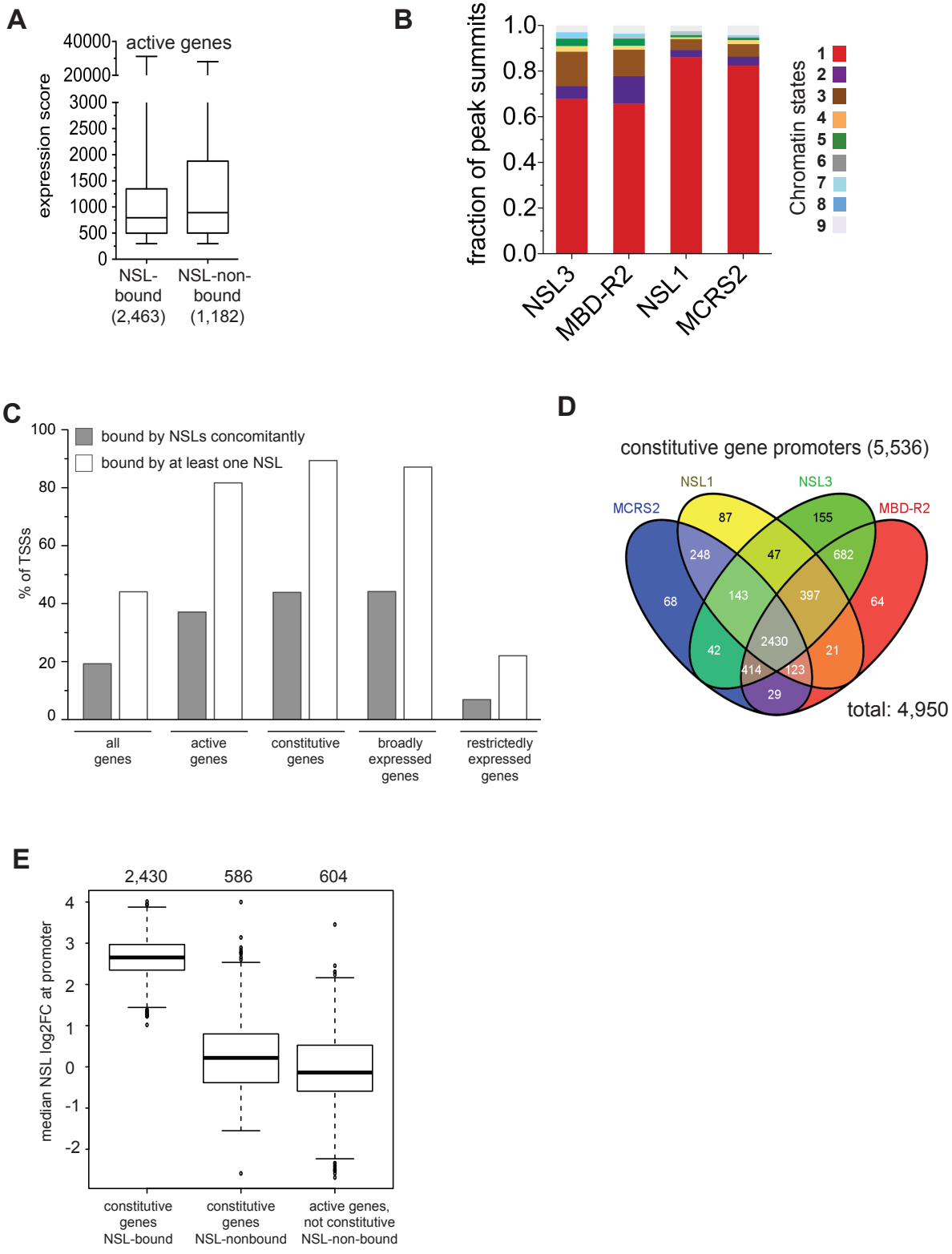
54. Salmon-Divon M, Dvinge H, Tammoja K, Bertone P (2010) PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. BMC Bioinformatics 11: 415.
55. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11: R106.
56. Conrad T, Cavalli FM, Holz H, Hallacli E, Kind J, et al. (2012) The MOF chromobarrel domain controls genome-wide H4K16 acetylation and spreading of the MSL complex. Dev Cell 22: 610–624.
57. Goecks J, Nekrutenko A, Taylor J (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11: R86.
58. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.
59. Oliveros JC (2007) VENNY. An interactive tool for comparing lists with Venn Diagrams.
60. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM (2009) A census of human transcription factors: function, expression and evolution. Nat Rev Genet 10: 252–263.
61. Kind J, Vaquerizas JM, Gebhardt P, Gentzel M, Luscombe NM, et al. (2008) Genome-wide analysis reveals MOF as a key regulator of dosage compensation and gene expression in Drosophila. Cell 133: 813–828.
62. Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res 34: W369–373.
63. Roider HG, Manke T, O'Keeffe S, Vingron M, Haas SA (2009) PASTAA: identifying transcription factors associated with sets of co-regulated genes. Bioinformatics 25: 435–442.
64. Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2: 28–36.
65. Tomancak P, Berman BP, Beaton A, Weiszmann R, Kwan E, et al. (2007) Global analysis of patterns of gene expression during Drosophila embryogenesis. Genome Biol 8: R145.

A.1.1 Supplemental Material

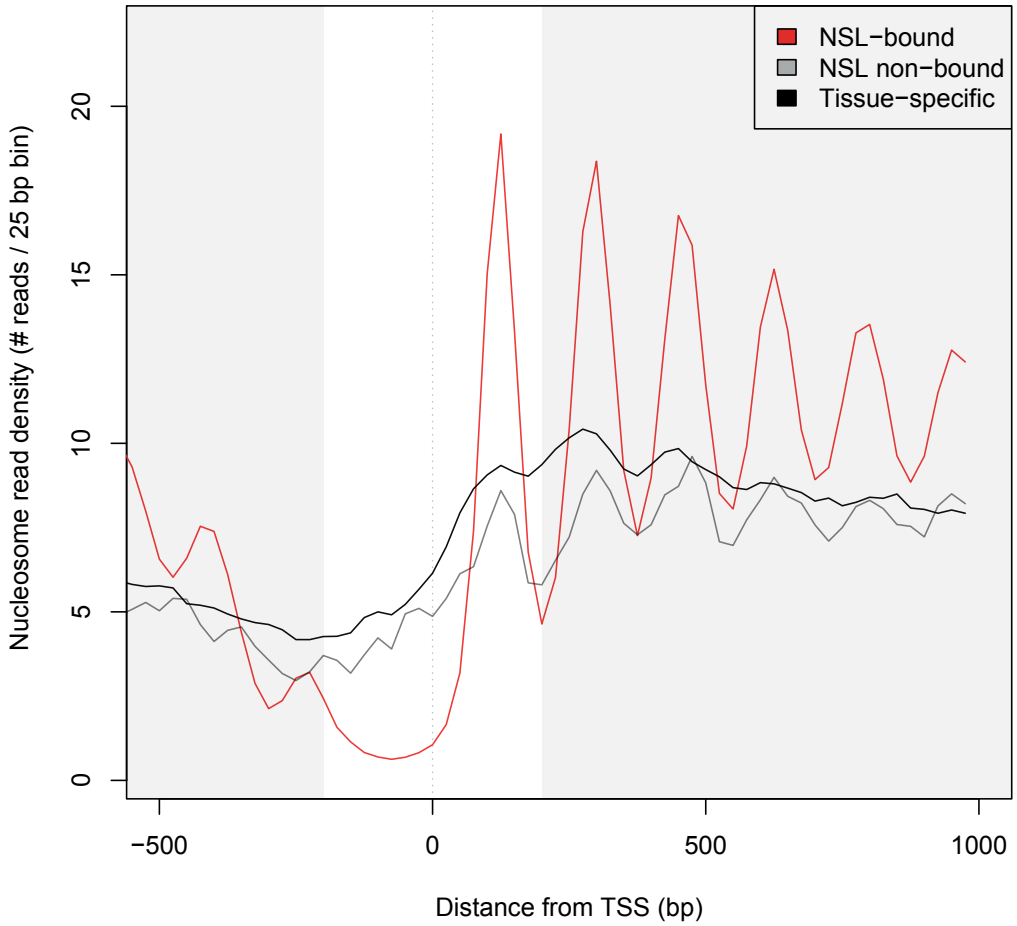
Supplementary Figure 1



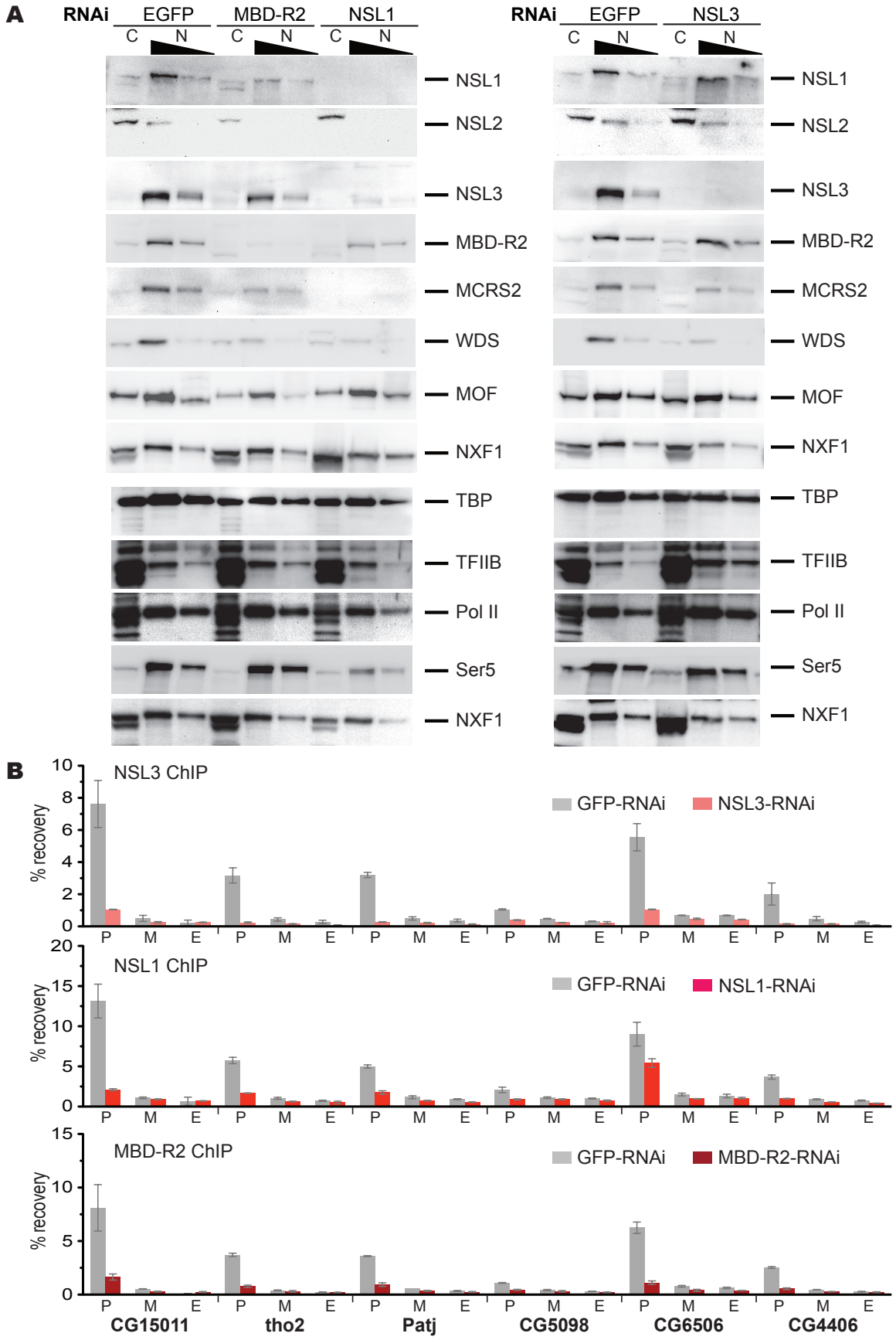
Supplementary Figure 2



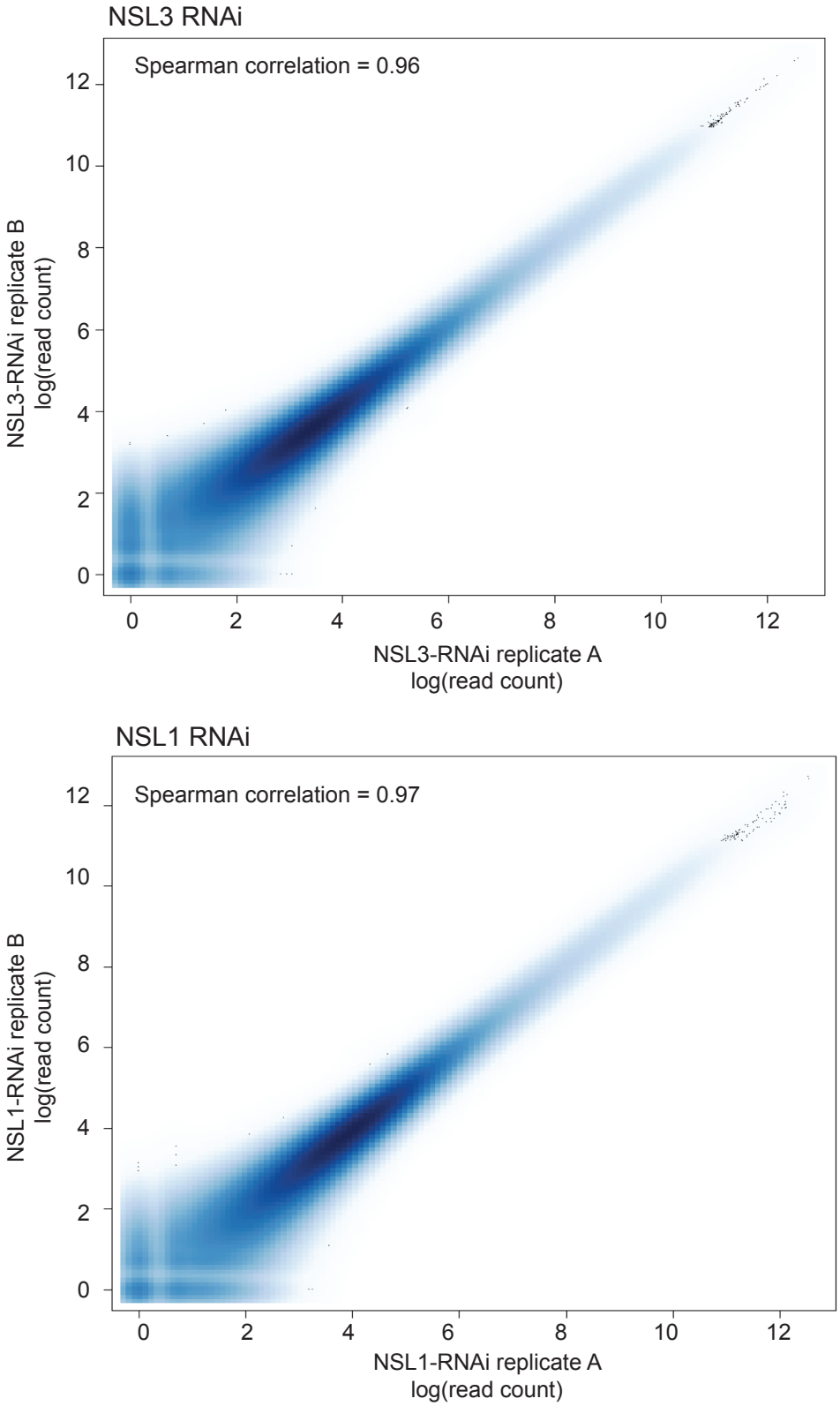
Supplementary Figure 3



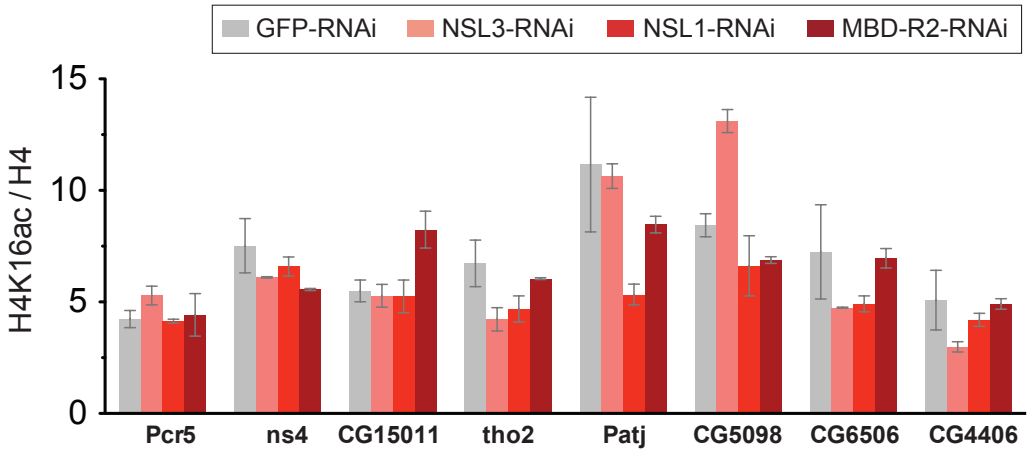
Supplementary Figure 4



Supplementary Figure 5



Supplementary Figure 6



Supplementary Figure 7

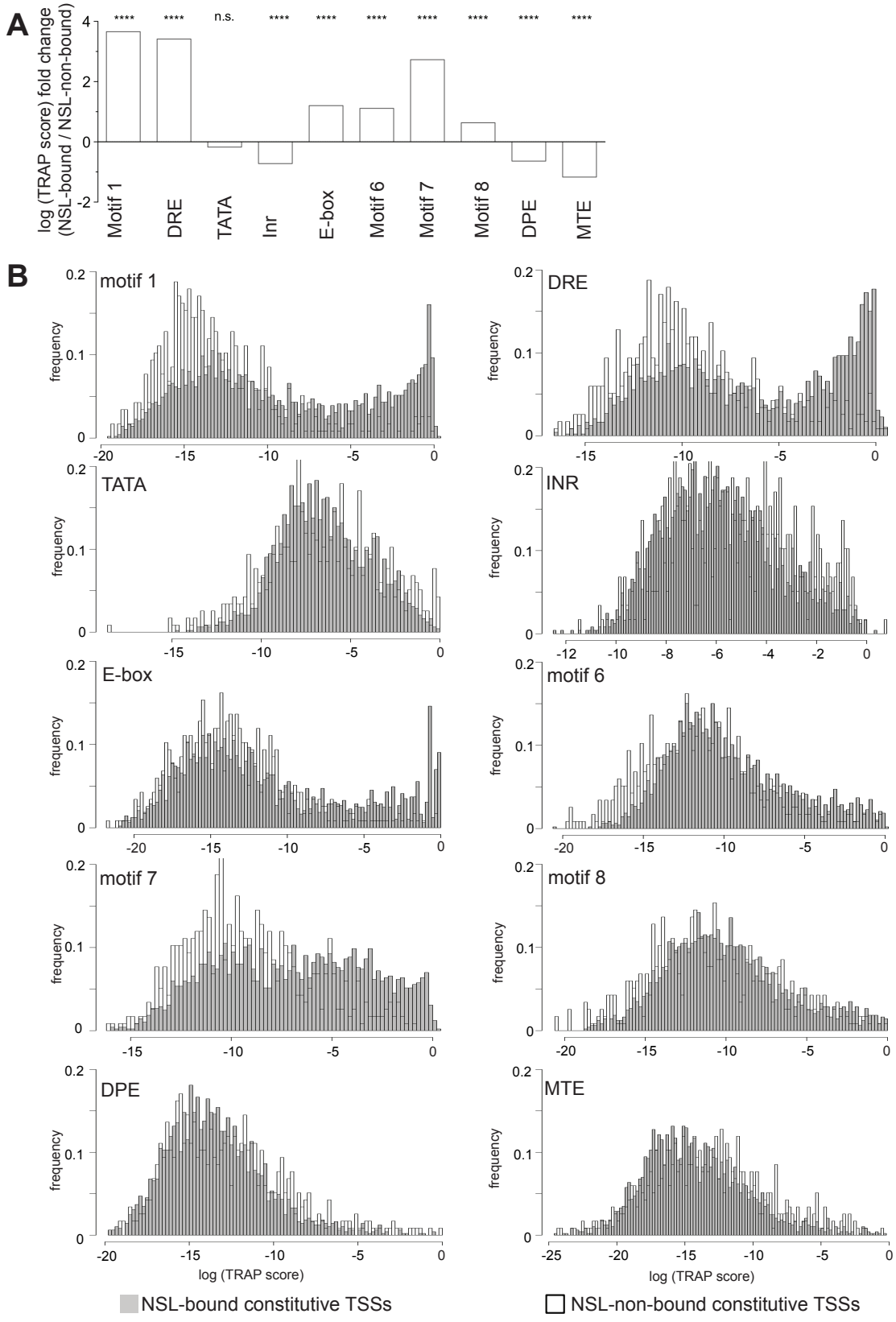
A **Peak summits with tag count ranks 500 - 1,000**

		NSL3	MBD-R2	NSL1	MCRS2
DRE	sites	—	—	183/500	123/500
	E-value			6.5×10^{-188}	3.4×10^{-104}
E-box	sites	117/500	83/500	77/500	90/500
	E-value	3.4×10^{-115}	1.3×10^{-54}	1.7×10^{-61}	8.3×10^{-91}
DMv2-like (r.c.)	sites	—	—	263/500	222/500
	E-value			3.9×10^{-189}	6.4×10^{-146}
Motif 1 (r.c.) (Ohler)	sites	64/500	63/500	—	—
	E-value	4.2×10^{-35}	1.7×10^{-42}		

B **Peak summits of peaks associated with constitutive genes**

		NSL3	MBD-R2	NSL1	MCRS2
DRE	sites	—	—	229/500	184/500
	E-value			3.3×10^{-215}	1.9×10^{-121}
E-box	sites	152/500	103/500	59/500	101/500
	E-value	2.3×10^{-234}	8.6×10^{-109}	2.3×10^{-31}	1.7×10^{-80}
DMv2-like (r.c.)	sites	—	295/500	295/500	337/500
	E-value		1.3×10^{-25}	2.5×10^{-295}	2.6×10^{-276}
Motif 1 (r.c.) (Ohler)	sites	62/500	—	—	—
	E-value	4.3×10^{-33}			

Supplementary Figure 8



Supplementary Figure 9

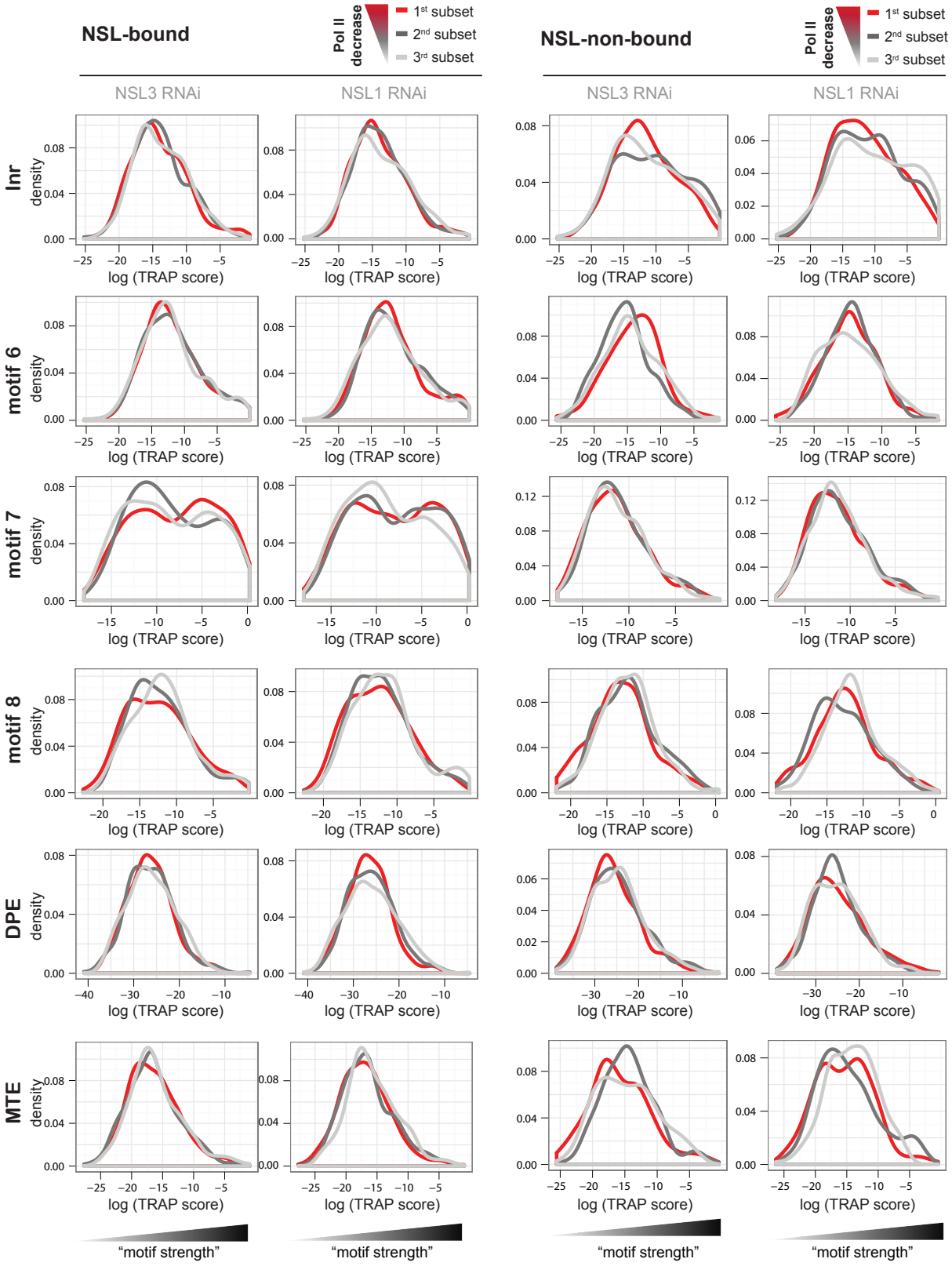


Table S1

		NSL1	MCRS2	NSL3	MBD-R2
peaks (MACS)		3,541	3,733	4,244	3,399
median peak size		1,129 bp	1,314 bp	2,227 bp	2,676 bp
(sub)peaks (PeakSplitter)		11,797	12,234	9,409	9,957
subpeaks per peak (median)		3	3	2	2
median peak size		350 bp	379 bp	1,110 bp	1,040 bp
TSSs bound	general	28.7 %	28.9 %	36.8 %	36.1 %
	active	55.7 %	54.1 %	68.3 %	65.9 %
	constitutive	63.2 %	63.2 %	77.9 %	75.2 %

A.2 MOF-associated complexes ensure stem cell identity and *Xist* repression

Chelmicki, T.*, **Dündar, F*.**, Turley, M.$^{\triangle}$, Khanam, T.$^{\triangle}$, Aktas, T.$^{\triangle}$, Ramírez, F., Gendrel, A. V., Wright, P. R., Videm, R., Backofen, R., Heard, E., Manke, T., Akhtar, A. (2014). eLife. doi:10.7554/eLife.02024
*, $^{\triangle}$ shared authorship

I performed all bioinformatic analyses except the mapping of the RNA-seq data and the determination of differentially expressed genes with DESeq2 which was done by Patrick Wright and Pavan Videm.

I generated all figures except Figures 1, 2G, 3F, 5–7 and the corresponding Supplementary Figures. I contributed to Figure 8.

Together with Tomasz Chelmicki and Asifa Akhtar I devised, wrote, and revised the manuscript.

eLIFE elifesciences.org

RESEARCH ARTICLE

MOF-associated complexes ensure stem cell identity and *Xist* repression

Tomasz Chelmicki[1,2†], Friederike Dündar[1,2,3†], Matthew James Turley[1,2‡], Tasneem Khanam[1‡], Tugce Aktas[1‡], Fidel Ramírez[3], Anne-Valerie Gendrel[4], Patrick Rudolf Wright[5], Pavankumar Videm[5], Rolf Backofen[5,6,7,8], Edith Heard[4], Thomas Manke[3], Asifa Akhtar[1]*

[1]Department of Chromatin Regulation, Max Planck Institute of Immunobiology and Epigenetics, Freiburg, Germany; [2]Faculty of Biology, University of Freiburg, Freiburg, Germany; [3]Bioinformatics Department, Max Planck Institute for Immunobiology and Epigenetics, Freiburg, Germany; [4]Mammalian Developmental Epigenetics Group, Institute Curie, Paris, France; [5]Bioinformatics Group, Department of Computer Science, University of Freiburg, Freiburg, Germany; [6]BIOSS Center for Biological Signalling Studies, University of Freiburg, Freiburg, Germany; [7]Center for Biological Systems Analysis, University of Freiburg, Freiburg, Germany; [8]Center for Non-Coding RNA in Technology and Health, University of Copenhagen, Frederiksberg, Denmark

Abstract Histone acetyl transferases (HATs) play distinct roles in many cellular processes and are frequently misregulated in cancers. Here, we study the regulatory potential of MYST1-(MOF)-containing MSL and NSL complexes in mouse embryonic stem cells (ESCs) and neuronal progenitors. We find that both complexes influence transcription by targeting promoters and TSS-distal enhancers. In contrast to flies, the MSL complex is not exclusively enriched on the X chromosome, yet it is crucial for mammalian X chromosome regulation as it specifically regulates *Tsix*, the major repressor of Xist lncRNA. MSL depletion leads to decreased *Tsix* expression, reduced REX1 recruitment, and consequently, enhanced accumulation of *Xist* and variable numbers of inactivated X chromosomes during early differentiation. The NSL complex provides additional, *Tsix*-independent repression of *Xist* by maintaining pluripotency. MSL and NSL complexes therefore act synergistically by using distinct pathways to ensure a fail-safe mechanism for the repression of X inactivation in ESCs.
DOI: 10.7554/eLife.02024.001

***For correspondence:** akhtar@ie-freiburg.mpg.de

†These authors contributed equally to this work

‡These authors also contributed equally to this work

Competing interests: See page 27

Funding: See page 27

Received: 06 December 2013
Accepted: 11 May 2014
Published: 19 May 2014

Reviewing editor: Danny Reinberg, Howard Hughes Medical Institute, New York University School of Medicine, United States

Introduction

Histone acetyl transferases (HATs) are among the key architects of the cellular epigenetic landscape as the acetylation of histones is unanimously associated with transcriptionally active domains. Many HATs also have the ability to acetylate non-histone proteins extending their influence to diverse cellular pathways inside and outside of the nucleus (reviewed in ***Sapountzi and Cote, 2011***). Based on their catalytic domains, the HATs are classified into two major families, GCN5 N-acetyl transferases (GNATs) and MYST HATs (named after the founding members MOZ, Ybf2/Sas3, Sas2, Tip60), that encompass diverse sets of protein complexes. The individual complex members enhance and modulate the enzymes' activities, guiding the versatile HATs towards specific functions. GCN5, for example, is part of SAGA, ATAC, and SLIK complexes that are associated with distinct histone tail modifications and differential gene regulation (reviewed in ***Lee and Workman, 2007***; ***Nagy et al., 2010***). In contrast, one of the well-known members of the MYST family, MOF (also known as: KAT8, MYST1), is rather substrate-specific for lysine 16 of histone H4 (H4K16) (***Akhtar and Becker, 2000***) and its interaction partners are thought to mainly alter the specificity and extent of MOF's H4K16 acetylation (H4K16ac). As part of the male-specific lethal (MSL) complex (MSL1, MSL2, MSL3, MOF, MLE, roX1 and roX2

eLife digest Gene expression is controlled by a complicated network of mechanisms involving a wide range of enzymes and protein complexes. Many of these mechanisms are identical in males and females, but some are not. Female mammals, for example, carry two X chromosomes, whereas males have one X and one Y chromosome. Since the two X chromosomes in females contain essentially the same set of genes, one of them undergoes silencing to prevent the overproduction of certain proteins. This process, which is called X-inactivation, occurs during different stages of development and it must be tightly controlled.

An enzyme called MOF was originally found in flies in two distinct complexes—the male-specific lethal (MSL) complex, which forms only in males, and the non-specific lethal (NSL) complex, which is ubiquitous in both males and females. These complexes are evolutionary conserved and are also found in mammals. While mammalian MOF is reasonably well understood, the MSL and NSL complexes are not, so Chelmicki, Dündar et al. have used various sequencing techniques, in combination with biochemical experiments, to investigate their roles in embryonic stem cells and neuronal progenitor cells in mice.

These experiments show that MSL and NSL complexes engage in the regulation of thousands of genes. Although the two complexes often show different gene preferences, they often regulate the same cellular processes. The MSL/NSL-dependent regulation of X chromosome inactivation is a prime example of this phenomenon.

The MSL complex reduces the production of an RNA molecule called Xist, which is responsible for the inactivation of one of the two X chromosomes in females. The NSL complex, meanwhile, ensures the production of multiple proteins that are crucial for the development of embryonic stem cells, and are also involved in the repression of X inactivation.

This analysis sheds light on how different complexes can cooperate and complement each other in order to reach the same goal in the cell. The knowledge gained from this study will pave the way towards better understanding of complex processes such as embryonic development, organogenesis and the pathogenesis of disorders like cancer.

DOI: 10.7554/eLife.02024.002

lncRNAs) in *Drosophila melanogaster*, MOF is recruited to the single X chromosome of male flies. The subsequent spreading of H4K16 acetylation results in transcriptional upregulation of the male X chromosome, the major means of *D. melanogaster* dosage compensation (reviewed in ***Conrad and Akhtar, 2011***). In addition to the highly specialized MSL-associated role, MOF is also involved in the more universal and sex-independent regulation of housekeeping genes within the non-specific lethal (NSL) complex (NSL1, NSL2, NSL3, MBD-R2, MCRS2, MOF, WDS) (***Mendjan et al., 2006***; ***Raja et al., 2010***; ***Feller et al., 2012***; ***Lam et al., 2012***).

MOF and most of its interaction partners are conserved in mammals, where MOF is also responsible for the majority of H4K16 acetylation (***Smith et al., 2005***; ***Taipale et al., 2005***). MOF is essential for mammalian embryonic development and unlike the male-specific lethality in *Drosophila*, deletion of *Mof* in mice is lethal for both sexes (***Gupta et al., 2008***; ***Thomas et al., 2008***). More specifically, mammalian MOF is critical for physiological nuclear architecture (***Thomas et al., 2008***), DNA damage repair (***Gupta et al., 2008***), maintenance of stem cell pluripotency (***Li et al., 2012***), differentiation of T cells (***Gupta et al., 2013***), and survival of post-mitotic Purkinje cells (***Kumar et al., 2011***). Compared to MOF, mammalian MSL and NSL complex members are poorly understood. Nevertheless, the individual complex members appear to have important functions in vivo as mutations of the NSL complex member KANSL1 cause the core phenotype of the 17q21.31 microdeletion syndrome (***Koolen et al., 2012***; ***Zollino et al., 2012***) and are common amongst patients with both Down syndrome and myeloid leukemia (***Yoshida et al., 2013***). Another NSL-associated protein, PHF20 has been shown to associate with methylated Lys370 and Lys382 of p53 (***Cui et al., 2012***) and to be required for somatic cell reprogramming (***Zhao et al., 2013a***). WDR5 was shown to be an essential regulator of the core transcription network in embryonic stem cells (***Ang et al., 2011***). The mammalian counterpart of *Drosophila* MSL2 was shown to have the capacity to ubiquitylate p53 (***Kruse and Gu, 2009***) and lysine 34 of histone 2B (***Wu et al., 2011***).

In the study presented here, we set out to dissect the mammalian MOF functions within the MSL and NSL complexes using genome-wide chromatin immunoprecipitation and transcriptome profiles

and biochemical experiments for the core members of MSL and NSL complexes in mouse embryonic stem cells (ESCs) and neuronal progenitor cells (NPCs). We found that the MSL and NSL members possess concurrent, as well as independent functions and that effects generally attributed to MOF are frequently accompanied by the NSL complex. The NSL complex abundantly binds to promoters of broadly expressed genes in ESCs and NPCs. These genes are predominantly downregulated upon depletion of either MOF or KANSL3. In contrast, the MSL complex shows more restricted binding in ESCs, which expands after differentiation, particularly at NPC-specific genes. In addition to promoter-proximal binding, we discover several thousand binding sites of KANSL3 and MSL2 at promoter-distal loci with enhancer-specific epigenetic signatures. The majority of these distal regulatory sites are bound in ESCs, but not in differentiated cells, and genes that are predicted to be targeted by TSS-distal binding of MSL2 are frequently downregulated in sh*Msl2*-treated cells. The distinct, yet synergistic actions of both complexes become very apparent at the X inactivation center (XIC) that encodes numerous non-coding RNAs involved in the silencing of one of the two X chromosomes in differentiating female cells. We show that the MSL but not the NSL complex directly promotes expression of *Tsix*, the inverse transcript and the key murine repressor of *Xist* during early differentiation. Depletion of MSL proteins results in attenuation of *Tsix* transcription, enhanced *Xist* RNA accumulation and 'chaotic' inactivation of variable numbers of X chromosomes during early differentiation. In addition to the very specific effect of MSL1/MSL2-depletion on the XIC genes, we show that MOF together with the NSL complex also influences *Xist* levels, but instead of affecting *Tsix*, MOF and KANSL3 depletion diminish key pluripotency factors involved in repressing *Xist*. Our study provides novel insights into the intricate interplay between MSL and NSL complexes in orchestrating gene expression. Furthermore, we demonstrate how MSLs and NSLs ensure the active state of two X chromosomes in mouse embryonic stem cells via distinct mechanisms.

Results

MOF and its complexes show distinct chromatin binding dynamics during differentiation

To examine the behavior of MSL and NSL proteins in a cell type-specific manner, we derived homogeneous populations of multipotent neuronal progenitor cells (NPCs) from mouse embryonic stem cells (ESCs) (***Conti et al., 2005***; ***Splinter et al., 2011***; ***Gendrel et al., 2014***). We followed the progress of the differentiation process by monitoring cell morphology (***Figure 1A***), as well as protein (***Figure 1B***) and transcript levels of ESC- and NPC-specific markers (***Figure 1—figure supplement 1A–C***). To gain a better understanding of how MOF-associated complexes behave throughout the differentiation process, in parallel to cell type-specific markers, we also monitored the RNA and protein levels of MOF, MSL (MSL1, MSL2), and NSL (KANSL1, KANSL3, MCRS1) complex members (***Figure 1B***, ***Figure 1—figure supplement 1A***). Interestingly, MSL and NSL complex members showed distinct RNA and protein dynamics during the process of differentiation: KANSL1 and KANSL3 protein levels remained unchanged, whereas MSL1, MSL2 and MOF became more abundant in NPCs accompanied by increased H4K16 acetylation (H4K16ac) (***Figure 1B***). These results were confirmed using another ES cell line and its NPC derivative (***Figure 1—figure supplement 1D***). The specificities of the antibodies were confirmed by co-immunoprecipitation assays (***Figure 1—figure supplement 2A–C***), as well as shRNA-mediated knockdowns followed by western blot analyses (for individual knockdowns please see below).

To assess the distinct behaviors of the complexes in more detail, we generated genome-wide chromatin binding profiles for MSL1, MSL2 (MSL complex), KANSL3, MCRS1 (NSL complex), and MOF (MSL and NSL). ChIP-seq experiments in ESCs and NPCs (***Figure 2***) yielded large numbers of high-quality DNA sequence reads and excellent agreements between the biological replicates (***Figure 2—figure supplement 1A***, ***Supplementary file 1A***). Using MACS for peak calling (***Zhang et al., 2008***) and additional stringent filtering ('Materials and methods'), we scored between 1500 and 15,000 regions of significant enrichments for the different proteins (***Supplementary file 1B***).

To uncover patterns of co-occurrence and independent binding, we used unsupervised clustering on the input-normalized signals. This unbiased approach allowed us to determine five main groups of binding distinguished by different combinations of the proteins and cell-type-specific dynamics. As shown in ***Figure 2***, three large clusters of binding sites encompassed regions, where at least 1 of the investigated proteins was present both in ESCs and NPCs (clusters A, B and C). The binding sites of clusters A and B predominantly overlapped with annotated transcription start sites (TSS) in contrast

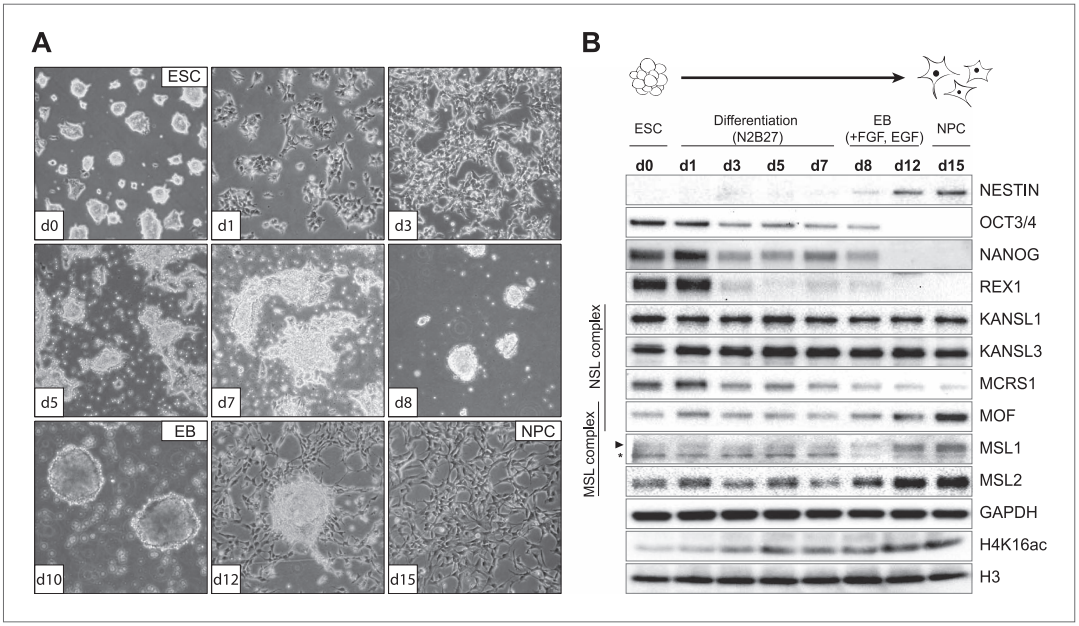


Figure 1. Distinct dynamics of MOF, MSL and NSL complexes during differentiation from ESCs to NPCs. (**A**) We monitored the cell morphology during differentiation of mouse embryonic stem cells into neuronal progenitor cells (NPC) via embryoid body formation (EB) with bright field microscopy. The day of differentiation is indicated in white boxes. (**B**) Western blot analysis for ESC to NPC differentiation. Stages of differentiation together with the day of differentiation (d0–d15) are indicated on top. GAPDH and histone 3 (H3) were used as loading controls. For expression analysis see ***Figure 1—figure supplement 1***.

DOI: 10.7554/eLife.02024.003

The following figure supplements are available for figure 1:

Figure supplement 1. Monitoring RNA and protein levels in ESCs and NPCs.

DOI: 10.7554/eLife.02024.004

Figure supplement 2. Verification of antibodies used in this study.

DOI: 10.7554/eLife.02024.005

to the regions that were bound exclusively in ESCs, which tended to contain inter- and intragenic regions (clusters D and E, ***Figure 2***). The width of the enrichments did not differ profoundly between the groups (cluster E: 836 bp median width, cluster A: 1782 bp median width). We found surprisingly few regions where MOF associated primarily with MSL complex members. Instead, approximately 80% of all MOF peaks displayed strong KANSL3 and MCRS1 signals (cluster B, see ***Figure 2*** and ***Figure 2—figure supplement 1B***), suggesting a predominant role of the NSL complex among MOF-associated complexes and a more specific role for the MSL complex at subsets of promoters and numerous intergenic and intronic regions. As the different clusters showed distinct enrichment patterns and diverse genomic localization, we set out to analyze the individual groups of binding in more detail.

The MSL and NSL complexes co-occur on active promoters of constitutively expressed genes in ESCs and NPCs

We first focused on the characterization of target promoters as the majority of MOF-binding was found around the TSS (mostly clusters A and B in ***Figure 2***, ***Figure 3A***). We identified 8947 TSSs overlapping with ChIP-seq peaks of KANSL3 and/or MCRS1 in ESCs that encompassed virtually all MOF- and MSL-bound TSSs (***Figure 3B***). This pattern did not change substantially in NPCs where TSSs overlapping with MOF peaks almost always (99%) showed significant enrichments of KANSL3 and in 35% of the cases additionally contained a peak of MSL2 (***Figure 3B***, middle panel). Genes that were TSS-bound in ESCs tended to be bound in NPCs as well (***Figure 3B***, middle panel and ***Figure 3—figure supplement 1A***). We next generated RNA-seq data for ESCs and NPCs, determined genes that were expressed in both cell types (FPKM >4) and found that all ChIPed proteins preferably bound to

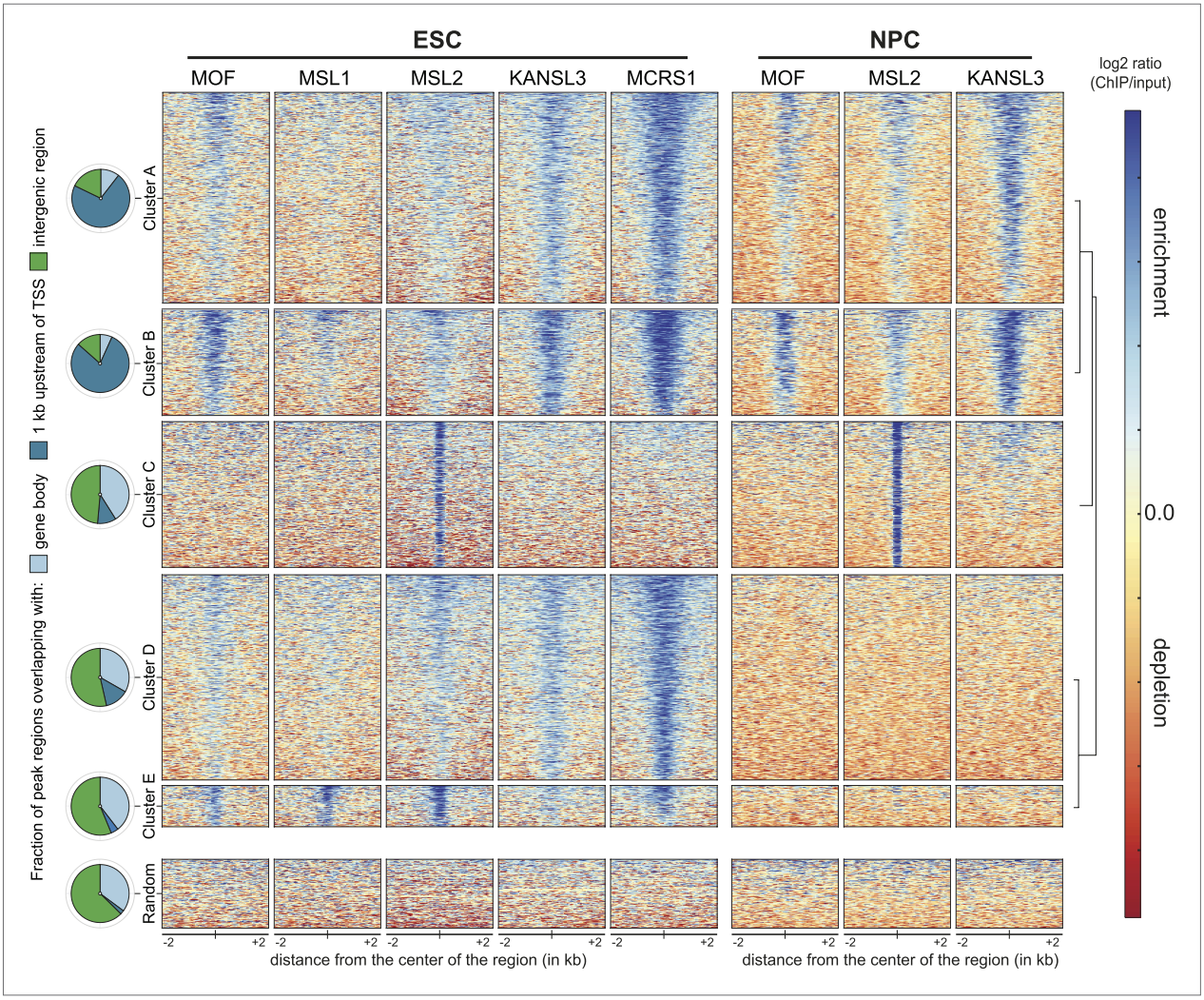


Figure 2. Distinct and shared binding sites of MOF and its complexes in mouse ESCs and NPCs. We applied unsupervised clustering on the union of peaks from all ChIP-seq samples and thereby identified five distinct groups of binding for MOF, MSL and NSL proteins in ESCs and NPCs. Shown here are the input-normalized ChIP signals for each cluster of peaks including a size-matched control set of random genomic regions. The order of the regions is the same for all columns. The pie charts on the left indicate the number of regions from each cluster that overlap with gene bodies, the region 1 kb upstream of genes' TSS or intergenic regions.

DOI: 10.7554/eLife.02024.006

The following figure supplements are available for figure 2:

Figure supplement 1. ChIP-seq quality measures.

DOI: 10.7554/eLife.02024.007

the promoters of active genes (***Figure 3C***). Interestingly, in ESCs, genes whose TSSs were bound by members of both complexes showed higher median expression values than genes bound by only one complex (***Figure 3—figure supplement 1B***). In contrast to the differing expression values, analysis of gene ontology (GO) using DAVID (***Huang da et al., 2009***) revealed basic housekeeping functions for both gene groups, regardless of whether they were bound by the NSL complex only or by both MOF-complexes together (***Figure 3—figure supplement 1C***). Consistently, the promoters of all target gene groups were enriched for motifs associated with broad, non-cell-type-specific expression such as ELK1, YY1, CREB, and E2F (***Xie et al., 2005***; ***Farre et al., 2007***) and showed profound enrichments of

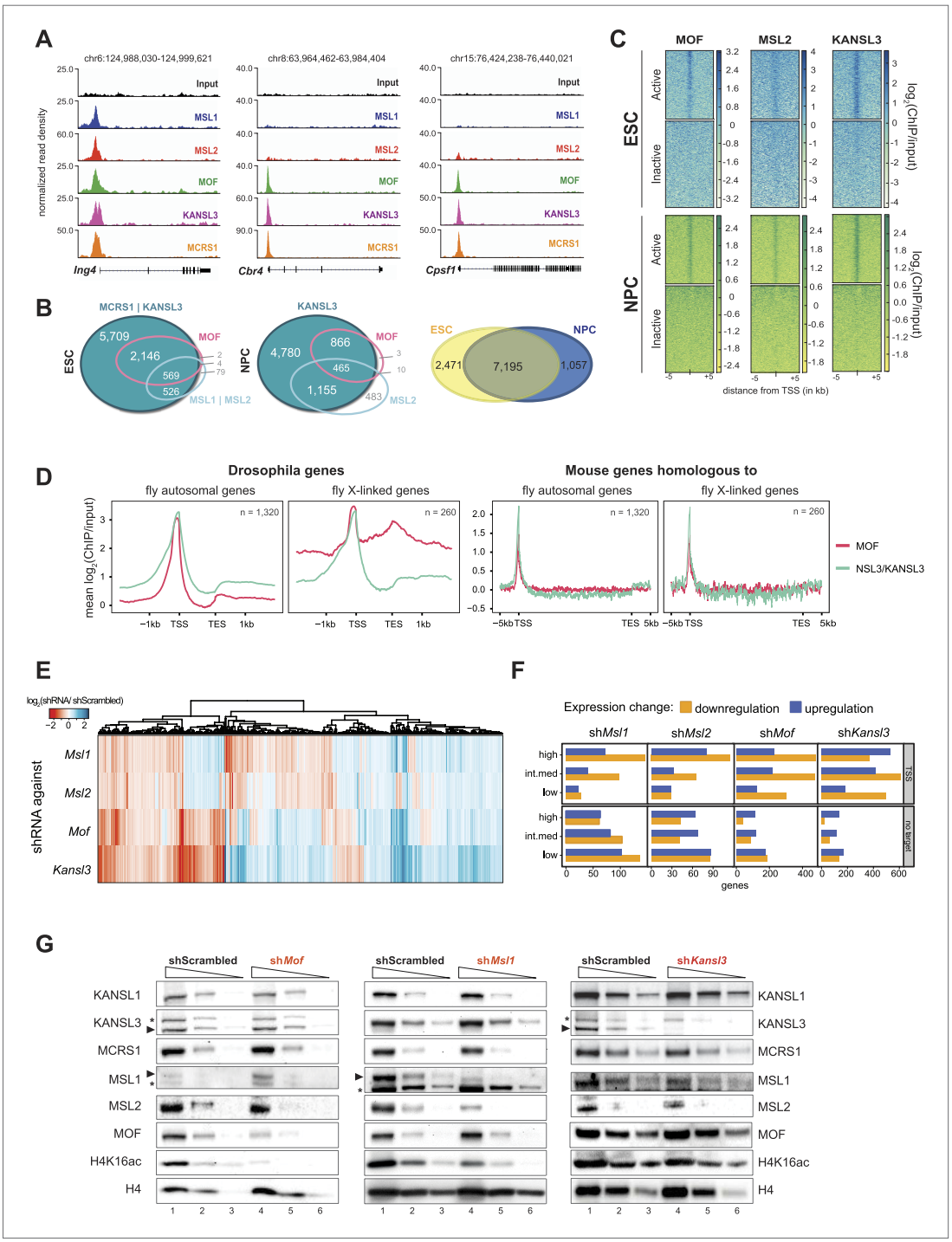


Figure 3. Both MOF-complexes bind to the TSS of broadly expressed genes in mouse ESCs and NPCs. (**A**) Genome browser snapshots of genes targeted by MSL and NSL complexes or by the NSL complex only. Signals were sequencing-depth-normalized and from ESCs. For ChIP-qPCR-based validation of the signals see ***Figure 3—figure supplement 4B***. (**B**) Venn diagrams of genes whose promoter regions (TSS ± 500 bp) overlapped with

Figure 3. Continued on next page

Figure 3. Continued

ChIP-seq peaks of NSL complex members (KANSL3 and/or MCRS1), MOF and MSL complex members (MSL1 and/or MSL2). The right-most panel depicts the overlap of genes bound by at least one factor in ESCs and NPCs. (**C**) The heatmaps display the input-normalized ChIP enrichments of MOF, MSL2 and KANSL3 around the TSS of genes that were active in ESCs as well as NPCs based on RNA-seq data that we generated for both cell types. (**D**) Summary plots of genes bound by the NSL complex in *D. melanogaster* for which mouse homologues were found. The input-normalized ChIP-seq signals around the TSS reveal markedly increased binding of MOF for male X-linked fly genes (left panels) that was not recapitulated in the mouse (right panels; ChIP-seq signals from ESCs). Fly genes were scaled to 1.2 kb and values were extracted from published data sets, mouse genes were scaled to 30 kb. (**E**) Heatmap depicting results of RNA-seq experiments from different shRNA-treated cells. The colors correspond to $\log_2$ fold changes (shRNA-treated cells/scrambled control) for genes whose expression was significantly affected in all knockdown conditions. Values were ordered using hierarchical clustering. (**F**) Bar plot of gene counts for different gene classes. We determined significantly up- and downregulated genes for each knockdown condition and binned them according to their expression strength in wild-type ESCs (high, intermediate, low). Then, for each gene, information about the TSS-targeting was extracted from the corresponding ChIP-seq sample. Non-target genes are neither bound at the promoter nor the gene body and were not predicted to be regulated via TSS-distal binding sites in any of the 5 ChIP-seq ESC samples. For details on the target classification see 'Materials and methods'. (**G**) Western blot analysis of MSL and NSL complex members and H4K16 acetylation in scrambled-, *Mof*-, *Msl1*-, and *Kansl3*-shRNA-treated male ESCs. Three concentrations (100%, 30%, 10%) of RIPA extract were loaded per sample. Asterisks mark the position of unspecific bands; triangles indicate the protein of interest.

DOI: 10.7554/eLife.02024.008

The following figure supplements are available for figure 3:

Figure supplement 1. MSL and NSL complexes target promoters of broadly expressed genes in ESCs and NPCs.

DOI: 10.7554/eLife.02024.009

Figure supplement 2. The NSL-, but not the MSL-binding mode of *D. melanogaster* is present in mammalian cells.

DOI: 10.7554/eLife.02024.010

Figure supplement 3. Effects of shRNA-mediated depletion of MOF, MSL1, MSL2 and KANSL3.

DOI: 10.7554/eLife.02024.011

Figure supplement 4. Assessment of ChIP signals around the TSSs of putative target genes as determined by ChIP-seq.

DOI: 10.7554/eLife.02024.012

CpG islands (***Figure 3—figure supplement 1D***), which is indicative of housekeeping genes (***Landolin et al., 2010***). Interestingly, when we analyzed the subset of genes that gained binding of either KANSL3 or MSL2 in NPCs, we found strong enrichments of GO terms related to embryonic development for KANSL3 targets and cell migration and neuronal development for MSL2 targets.

The TSS-binding of the mouse NSL complex resembles that of the NSL complex in *D. melanogster*

MOF has traditionally been associated with a widespread enrichment along male X-linked genes in flies that is dependent on the MSL proteins (***Figure 3D***, ***Figure 3—figure supplement 2A***). In our mammalian profiles, despite the presence of the MSLs, we could neither detect X-specific enrichments of MOF nor broad domains of binding along gene bodies. Furthermore, promoter-distal binding sites consisted of narrow peaks and no evidence of spreading from intronic or intergenic regions was observed (***Figures 2, 3A,D***).

We then examined whether there was a correlation between NSL complex binding in *D. melanogaster* and mouse cells. Indeed, we found that mouse genes that were homologous to NSL complex targets in *D. melanogaster* had a high probability of being bound by the murine NSL complex as well (Pearson's Chi squared test of independence between NSL binding in the fly and the mouse, p-value <2.2e−16). We additionally observed that mouse genes expressed in ESCs and NPCs, whose fly homologues were NSL targets, showed stronger signals for H3K4me3, MOF, KANSL3, and MCRS1 (but not for MSL1 or MSL2) than the mouse homologues of non-NSL-bound *D. melanogaster* genes (***Figure 3—figure supplement 2B***; lists of NSL-bound and NSL-non-bound fly genes were from ***Lam et al., 2012***). These findings support the notion that the function in housekeeping gene regulation by the *D. melanogaster* NSL complex is evolutionary conserved.

Depletion of MSL and NSL complex members results in genome-wide downregulation of TSS-target genes

To dissect the biological consequences of the gene targeting by the different MSL and NSL proteins in ESCs, we systematically depleted core members of both complexes (MOF, KANSL3, MSL1, MSL2) (***Figure 3—figure supplement 3A***). Interestingly, MOF- or KANSL3-depleted cells showed more

severe proliferation defects than MSL1- and MSL2-depleted cells (***Figure 3—figure supplement 3B***). We subsequently performed RNA-seq experiments from shRNA-treated cells and determined their differential expression against the scrambled control to dissect transcriptional outcomes of the depletions at a global level. We found a striking overlap between the differential expression of MSL1 and MSL2 knockdowns and a higher resemblance of MOF-dependent differential expression to that of KANSL3-depletion (***Figure 3E***). When we specifically focused on genes that we had identified as TSS-bound in our ChIP-seq samples, we found that their transcripts tended to be downregulated in all four knockdowns in comparison to untargeted genes which showed higher fractions of upregulation. These effects were independent of the wild-type expression status of the gene or the chromosome (***Figure 3F***, ***Figure 3—figure supplement 3C***).

TSS-binding of MSL1 and KANSL3 does not require MOF

Turning to the assessment of protein levels in shRNA-treated cells, we detected markedly reduced bulk H4K16 acetylation in MSL1- and MOF-depleted cells and only slight reduction upon KANSL3-depletion. This is consistent with previous reports that indicate MSL1 as the major enhancer of MOF's H4K16 acetylation (***Kadlec et al., 2011***) and demonstrate relaxed substrate specificity for the NSL complex (***Zhao et al., 2013b***). In addition, we found that MSL1-depletion affected the levels of MSL2 but not of NSL complex members while the depletion of KANSL3 moderately decreased protein levels for both complexes (***Figure 3G***). ChIP-qPCR assays in MOF-depleted cells revealed that MSL1 and KANSL3 do not require the presence of MOF to bind to gene promoters, which is in agreement with previous observations in *D. melanogaster* (***Hallacli et al., 2012***; ***Figure 3—figure supplement 4C***).

In summary, our TSS-focused analysis shows that the localized binding of the NSL complex to the promoters of housekeeping genes appears to be a conserved feature between the mammalian and *Drosophila* systems. Unlike in the fly, we do not detect an MSL- and X-chromosome-specific binding mode of MOF in the mouse cell lines. Instead, both complexes narrowly bind to TSSs where their co-occurrence is associated with significantly higher median expression values than those solely bound by the NSL complex. Moreover, we found that MOF is dispensable for the TSS recruitment of its interaction partners and that depletions of the individual proteins predominantly result in the downregulation of TSS-bound genes, further supporting the fact that the promoter-binding of the MSL and the NSL complex is associated with active transcription.

MSL and NSL complex members individually bind to active enhancers in ESCs

In addition to promoter-proximal binding, where both the MSL and NSL complex tend to (co-)occur constitutively in ESCs and NPCs, we identified a large proportion of binding sites where the proteins were present in a dynamic fashion, that is their binding was observed only in ESCs but not in NPCs (***Figure 2***, clusters D and E). In contrast to the binding mode represented by clusters A and B (***Figure 2***), here MSL2, MCRS1, and KANSL3 were predominantly enriched within introns and intergenic regions that underwent significant CpG methylation upon differentiation (e.g., from median 50% CpG methylation in ESCs to more than 80% in NPCs for cluster D; bisulfite sequencing data from ***Stadler et al., 2011***). As shown in ***Figure 4A***, CpG methylation in NPCs was particularly pronounced around the center of the regions with significant ChIP enrichments in ESCs, indicating a correlation between the loss of ChIP-seq signal for MOF, MSL1, MSL2, KANSL3 and MCRS1, and DNA methylation upon differentiation. In addition, the regions of cluster D and, to a lesser extent the MSL1-rich cluster E (***Figure 2***), showed highly localized enrichments of DNase hypersensitivity sites (DNase HS), RNA Polymerase II (Pol II), p300, methylation of histone 3 on lysine 4 (H3K4me1), and acetylation of histone 3 on lysine 27 (H3K27ac) in ESCs (***Figure 4A***), which are characteristic features of enhancer regions. We thus examined whether MOF and its interaction partners were enriched on known enhancer regions, using lists of typical and super enhancers defined by binding sites of the pluripotency factors SOX2, NANOG, and OCT4 (***Whyte et al., 2013***), as well as sets of active and poised enhancers based on histone mark signatures (***Creyghton et al., 2010***).

Interestingly, MSL2, KANSL3 and MCRS1, but not MOF and MSL1, showed profound enrichments for active and poised ESC enhancers (***Figure 4—figure supplement 1A***) as well as along the regions of super enhancers that have been described as being particularly important for maintenance of cell identity (***Whyte et al., 2013***).

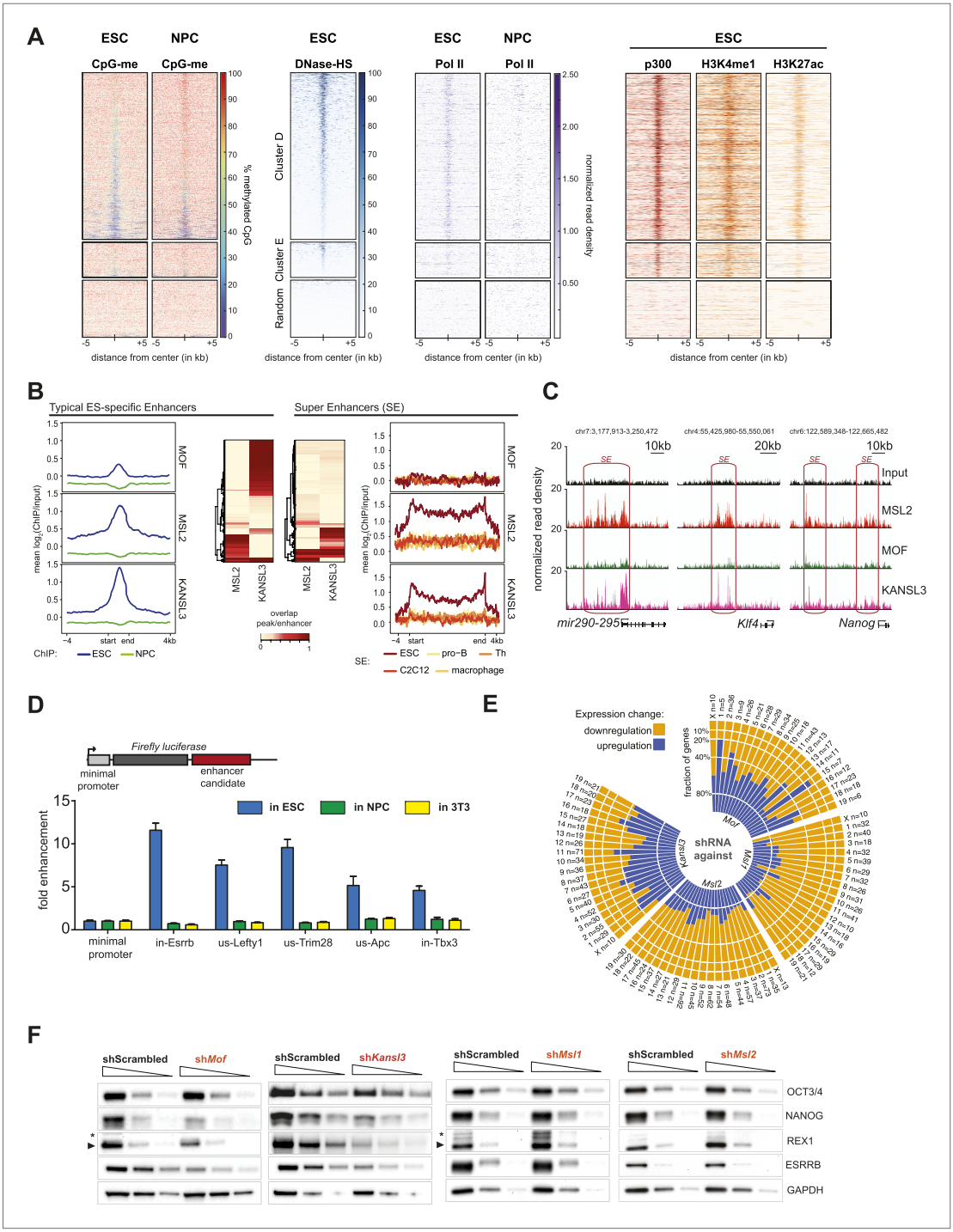


Figure 4. MSL and NSL complex members are enriched at regions with enhancer marks in ESCs. (**A**) Shown here are the fractions of methylated cytosines and ChIP-seq read densities of enhancer markers for regions of ESC-specific enrichments of our proteins of interest. We downloaded the different data from public repositories (see ***Supplementary file 3A*** for details) and calculated the values for the regions of the ESC-specific clusters **D** and **E** and

Figure 4. Continued on next page

Figure 4. Continued

random genomic loci. Most data sets used here were from mouse ESC except one RNA Polymerase II (Pol II) sample from NPC. All heatmaps were sorted according to the DNase hypersensitivity values except for CpG methylation heatmaps which were sorted according to their own values. (**B**) Summary plots of input-normalized ChIP-seq signals along typical (TE) and super enhancers (SE) (***Whyte et al., 2013***). Note that we show the ESC-specific TE only while on the right-hand side we show the signal for SE regions from several cell types. Enhancer regions were scaled to 30 kb (SE) and circa 700 bp (TE). The heatmaps between the summary plots depict how much of each enhancer region overlaps with ChIP-seq peaks of MSL2 or KANSL3. ESC = embryonic stem cells (n = 232), pro-B = progenitor B cells (n = 396), Th = T helper cells (n = 437), C2C12 = myotube cells (n = 536). (**C**) Exemplary genome browser snapshots of annotated super enhancers (SE, pink boxes) for three pluripotency factors displaying the sequencing-depth normalized ESC ChIP-seq signals of MSL2, MOF and KANSL3. See ***Figure 4—figure supplement 4C*** for additional examples. (**D**) Luciferase assays demonstrate the biological activity of regions bound by MOF-associated proteins in ESCs ('in' stands for intronic region, 'us' indicates that the cloned region is upstream of the gene). The firefly luciferase gene was cloned under a minimal promoter together with the putative enhancer region in ESCs, NPCs, and 3T3 cells. The graphs represent at least three independent experiments performed in technical triplicates; error bars represent SEM. (**E**) Bar plots depicting the fraction of significantly up- and downregulated genes per chromosome in the different shRNA-treated cells compared to shScrambled controls (total number of significantly affected genes per sample and chromosome labels are indicated). All genes counted here were classified as TSS-distal target genes in the respective ChIP-seq experiments. See 'Materials and methods' for details of the classifications. (**F**) Western blot analyses of the pluripotency factors in scrambled-, *Mof*-, *Kansl3*-, *Msl1*-, and *Msl2*-shRNA-treated male ESCs. For additional analyses in female ESCs see ***Figure 6C***. The respective dilution (100%, 30%, 10%) of loaded RIPA extract is indicated above each panel. Asterisks mark the position of unspecific bands; triangles indicate the protein of interest. GAPDH was used as the loading control. For antibodies see 'Materials and methods'.

DOI: 10.7554/eLife.02024.013

The following figure supplements are available for figure 4:

Figure supplement 1. MSL2 and KANSL3 show strong enrichments at typical and super enhancers in ESCs.

DOI: 10.7554/eLife.02024.014

Figure supplement 2. MOF is moderately enriched at non-canonical enhancers.

DOI: 10.7554/eLife.02024.015

Figure supplement 3. MSL2 has intergenic binding sites in DNA-hypomethylated regions that are enriched for SMAD3 binding sites.

DOI: 10.7554/eLife.02024.016

Figure supplement 4. Biological significance of the TSS-distal binding sites of the investigated proteins.

DOI: 10.7554/eLife.02024.017

The signals of MSL2 and KANSL3 were specific for ESC enhancers and wide-spread along super enhancer regions (***Figure 4B,C***). We noted that enhancers overlapping with MSL2 ChIP-seq peaks tended to show lower KANSL3 enrichments and vice versa, implying that MSL2 and KANSL3 preferred different enhancer regions (heatmaps in ***Figure 4B***, ***Figure 4—figure supplement 1B***). MOF was not enriched at super enhancers and generally, its binding to TSS-distal sites was much less pronounced than to gene promoters (***Figure 4—figure supplement 1A, 1C***, ***Figure 2***). Like for TSS-specific binding, MOF was not alone (87% of TSS-distal MOF peak regions overlapped with either KANSL3 or MSL2). Since a recent report showed H4K16 acetylation to be present at p300- and H3K27-acetylation-independent enhancer regions (***Taylor et al., 2013***), we analyzed the moderate TSS-distal enrichments of MOF in more detail and observed a slight preference for TSS-distal regions that were not overlapping with previously published ESC enhancer regions (***Figure 4—figure supplement 2A***). In fact, we detected the strongest MOF signals in regions with rather low enrichments of known enhancer marks (see DNase HS, p300, H3K4me1, H3K27ac in ***Figure 4—figure supplement 2B and 2C***), which suggested a preferred binding of MOF outside canonical ESC regulatory regions.

In addition to ESC-specific binding of MSL2 and KANSL3 to predicted enhancers, we also identified a very distinct set of TSS-distal binding sites by MSL2 to introns and intergenic regions without enhancer-associated marks (cluster C in ***Figure 2***). Approximately, 81% of these cluster C regions had solitary MSL2 enrichments without significant signals of any of the other ChIPed proteins. Interestingly, these MSL2 binding sites increased in number and binding strength upon differentiation to NPCs (829 solitary MSL2 peaks in ESCs compared to 3635 in NPCs). In contrast to the previously described binding sites that were characterized by the prevalence of open, active chromatin (***Figures 3 and 4***), here MSL2 was excluded from hypo-methylated DNA regions (***Figure 4—figure supplement 3A***; note the different behavior of KANSL3). When we searched the unique MSL2 binding sites for DNA motifs, we obtained a $(CAGA)_n$ motif (***Figure 4—figure supplement 3B***) that was previously described as a binding site for SMAD3, a transcription factor that translates the TGF-beta receptor response into gene expression regulation (***Zawel et al., 1998***). When we subsequently scanned all the binding sites for the presence of the published, original SMAD3 motif, we found a strikingly specific signal for the center of the solitary MSL2 ChIP-seq peaks (***Figure 4—figure supplement 3C***).

We conclude that MOF, MSL2 and KANSL3 specifically recognize ESC enhancers. In contrast to MSL-MOF-NSL co-occurrence at housekeeping gene promoters, we found evidence for differential and independent binding of the individual proteins to gene bodies and intergenic regions suggesting the potential for distinct tissue-specific regulatory functions of MSL2 and KANSL3. These data reveal a newly evolved function of MSL2 and KANSL3 in mammals, which has not been observed in flies.

Genes associated with TSS-distal binding sites of MSL1 and MSL2 are frequently downregulated in cells lacking MSL1 or MSL2

To study the functional implications of the binding of MSL2 and KANSL3 to putative ESC enhancers, we first tested five different regions located near genes related to pluripotency and self-renewal (***Hu et al., 2009***; ***Young, 2011***). Using luciferase reporter constructs, we found strong transcriptional enhancement for all tested regions in ESCs, but not in NPCs or 3T3 cells which correlated with the presence of MSL2 and/or KANSL3 and MCRS1 in ESCs only (***Figure 4D***, ***Figure 4—figure supplement 4A***).

We then used our RNA-seq data sets from MSL1-, MSL2-, MOF-, and KANSL3-depleted cells to assess the effects on the transcription of those genes that were not bound at promoters, but had been predicted by GREAT (***McLean et al., 2010***) to be regulated by TSS-distal binding sites of the respective protein. As shown in ***Figure 4E***, we again found similar effects for KANSL3- and MOF-depleted cells compared to MSL1- and MSL2-depleted cells with the latter group showing genome-wide downregulation of predicted target genes. In fact, the numbers of TSS-distal targets of MSL1 or MSL2 that were significantly reduced in the respective shRNA-treatments were markedly larger than for genes where MSL1 or MSL2 bound to the promoter (compare ***Figure 3F*** with ***Figure 4—figure supplement 4B***). Moreover, in MSL2-, but not KANSL3-depleted cells, the effects on TSS-distally targeted genes were slightly stronger than for TSS-targets (***Figure 4—figure supplement 4C***).

Depletions of MOF and KANSL3, but not of MSL complex members affect key pluripotency factors

While TSS-binding predominantly occurred at housekeeping genes, we noticed that the majority of enhancer regions associated with key pluripotency factors (e.g., SOX2, ESRRB, MYC, REX1, TBX3, NANOG) were strongly enriched for MSL2 and KANSL3. We thus assessed the effects of the protein depletions on pluripotency factors in ESCs and found strongly reduced levels of NANOG, REX1, and ESRRB in MOF- or KANSL3-depleted cells. Surprisingly, the pluripotency factors remained almost unaffected in cells depleted of MSL1 or MSL2 (***Figure 4F***). These contrasting results were mirrored by decreased levels of alkaline phosphatase (AP) in MOF- and KANSL3-, but not in MSL1- or MSL2-depleted cells (***Figure 4—figure supplement 4D***).

These findings indicate that despite their frequent effects on TSS-distally targeted genes, MSL1 and MSL2 might not show dominant effects at genes that are bound by KANSL3 as well. Therefore, we specifically searched for regions without KANSL3 binding to identify putative MSL-specific functions.

The MSL complex binds multiple loci within the X inactivation center

As described previously, we identified only a small subset of regions in the mouse genome where MSL complex members were enriched exclusively (see cluster E in ***Figure 2***). Strikingly, several of these binding sites fall into a region known as the X inactivation center (XIC). The XIC is the X-chromosomal region necessary and sufficient to control the inactivation of one of the two X chromosomes in females (reviewed in ***Pollex and Heard, 2012***).

The XIC site with the strongest concomitant enrichments of MSL1, MSL2 and MOF was the major promoter (P2) of *Tsix* and its intronic minisatellite—*DXPas34* (***Figure 5A,B***). *DXPas34* is a well-characterized tandem repeat that serves as a binding platform for multiple transcription factors and contains bidirectional enhancing properties essential for the expression of *Tsix*, the antisense transcript of *Xist* (***Debrand et al., 1999***; ***Cohen et al., 2007***; ***Donohoe et al., 2007***; ***Navarro et al., 2010***; ***Gontan et al., 2012***). In rodents, *Tsix* antisense transcription across the *Xist* promoter is required for regulating the levels of *Xist* accumulation. In turn, *DXPas34* deletion impairs the recruitment of Pol II and TFIIB to the major promoter of *Tsix* causing its downregulation (***Vigneau et al., 2006***).

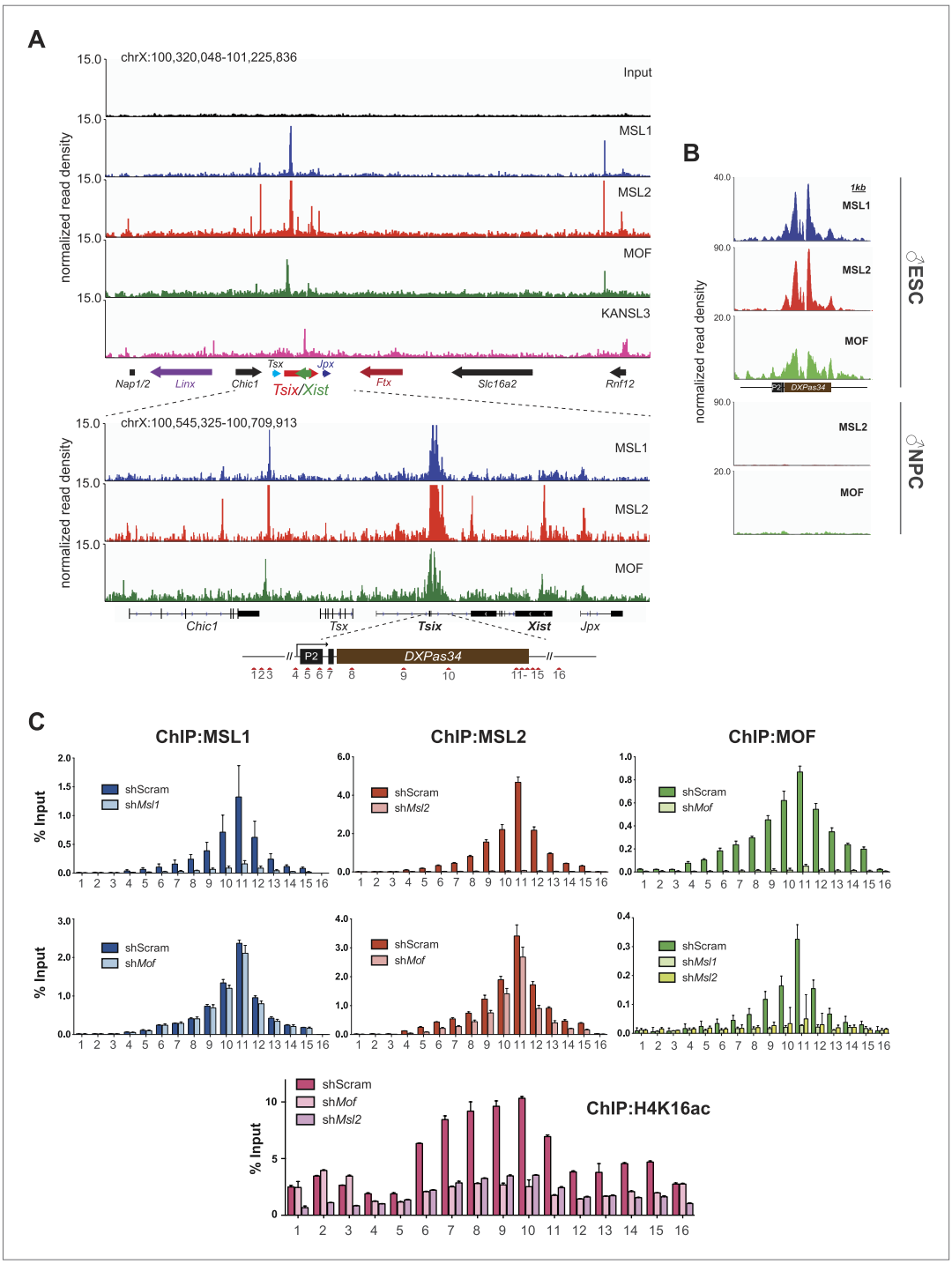


Figure 5. The MSL complex binds multiple loci within the X inactivation center including the *Tsix DXPas34* minisatellite enhancer. (**A**) Genome browser snapshots of the mouse X inactivation center (approximately 0.9 Mb) (upper panel) plus enlargement of the 164 kb region between *Chic1* and *Jpx/Enox* (lower panel). The signals shown are the sequencing-depth normalized profiles for ChIP-seq from ESCs (for corresponding profiles in NPCs see

Figure 5. Continued on next page

Figure 5. Continued

Figure 5—figure supplement 1A); colored arrows indicate genes of lncRNAs. The schematic representation of the *DXPas34* locus depicts the locations of the primer pairs that were used for ChIP-qPCR analyses (***Supplementary file 3B***). (**B**) Genome browser snapshots of the *DXPas34* minisatellite of sequencing-depth normalized ChIP-seq profiles in ESCs and NPCs. (**C**) ChIP-qPCR analyses of MSL1 (blue), MSL2 (red), MOF (green), and H4K16 acetylation (purple) across the *Tsix* major promoter (*P2*) and the *DXPas34* enhancer in male ESCs treated with the indicated shRNAs. For corresponding ChIP-qPCR in female ESCs see ***Figure 5—figure supplement 1C***. Panels in the middle show the effects of MOF depletion on the recruitment of MSL1 and MSL2 to *DXPas34* and vice versa. The bottom panel shows effects of depletion of control (dark pink), MOF (light pink) and MSL2 (purple) on the H4K16 acetylation signal. The labels of the x axes correspond to the arrowheads in (**A**). Results are expressed as mean ± SD of three biological replicates; cells were harvested on day 4 (*Msl1*, *Msl2*) or 5 (*Mof*) after shRNA treatment. For primer pairs see ***Supplementary file 3C***.

DOI: 10.7554/eLife.02024.018

The following figure supplements are available for figure 5:

Figure supplement 1. The MSL proteins bind to multiple loci within the X inactivation center (XIC).

DOI: 10.7554/eLife.02024.019

In addition to the *DXPas34* binding site, we detected MSL peaks on the promoters, gene bodies and intronic regions of other key XIC regulators including the genes of the long non-coding (lnc) RNAs *Xist* and *Jpx*. Additionally, we observed peaks upstream of the *Tsx* gene and both at the TSS and downstream of the *Rnf12* gene (***Figure 5A***). Products of all of these genes were shown to play important roles in orchestrating the process of X inactivation (***Stavropoulos et al., 2001***; ***Shin et al., 2010***; ***Tian et al., 2010***; ***Anguera et al., 2011***; ***Chureau et al., 2011***; ***Gontan et al., 2012***; ***Sun et al., 2013***).

The XIC binding of MSL-MOF was specific to ESCs, as almost all enrichments were abolished upon differentiation, except for some loci upstream of *Xist* where traces of binding could still be detected in NPCs (e.g., *Ftx* and *Jpx* TSS, ***Figure 5—figure supplement 1A***).

We next confirmed the high ChIP-seq enrichments of MSL1, MSL2 and MOF and assessed H4K16 acetylation on the major promoter of *Tsix* and along *DXPas34* with ChIP-qPCR assays covering the entire region in male and female ESCs (***Figure 5C***, ***Figure 5—figure supplement 1B***). Interestingly, the recruitment of MOF was almost completely abolished in both MSL1- and MSL2-depleted cells, whereas the depletion of MOF had no effect on MSL1 and MSL2 binding to the *Tsix* major promoter and *DXPas34* (***Figure 5C***). H4K16 acetylation ChIP signals were severely reduced in both MOF- and MSL2-depleted cells. These results are in agreement with our global observations (***Figure 3G***, ***Figure 3—figure supplement 4C***) and indicate that MSL1 and MSL2 are together necessary and sufficient for the recruitment of MOF and for the deposition of H4K16 acetylation at *DXPas34*.

MSL1 and MSL2 are important for *Tsix* expression

To directly assess the functional outcome of MOF-, MSL1-, and MSL2-depletions, we studied the expression of *Tsix* and *Xist* in shRNA-treated ESCs. Unexpectedly, only MSL1- and MSL2-, but not MOF-depletion led to pronounced downregulation of *Tsix* both in male and female ESCs (***Figure 6A***; note that in our RNA-seq data set for MSL2-depleted cells, *Tsix* was among the five most strongly downregulated genes). Downregulation of *Tsix* was accompanied by moderately elevated *Xist* RNA levels in MSL1- and MSL2-depleted ESCs whereas depletion of MOF yielded the most pronounced (8–15-fold) upregulation of *Xist* without affecting *Tsix*.

To determine the effects on *Tsix* in individual cells, we next performed RNA-FISH with probes against *DXPas34* and *Huwe1* in female ESCs (*Huwe1* was used to mark X chromosomes, for probe references see 'Materials and methods'). The RNA-FISH confirmed the qPCR results as we observed global reduction and in many cases elimination of *DXPas34* signals in MSL1- and MSL2-, but not in MOF-depleted cells (***Figure 6B***, ***Figure 6—figure supplement 1A–C***).

We next wanted to understand the mechanistic differences between the *Tsix*-specific and the *Tsix*-independent effects on *Xist* levels that we found for depletions of MSL1/MSL2 and MOF, respectively. As pluripotency factors are additional regulators of *Xist* (***Navarro et al., 2008***; ***Nesterova et al., 2011***), we assessed the consequences of the different knockdowns on the *Xist*-related pluripotency network in female ESCs. Like for MOF- and KANSL3-depletions in male ESCs (***Figure 4F***), the depletion of MOF (but not of MSL1 or MSL2) in female ESCs resulted in a significant decrease of transcript and protein levels of pluripotency factors that had previously been associated with *Xist* repression (e.g., NANOG and REX1; see ***Figure 6C***, ***Figure 6—figure supplement 1D***).

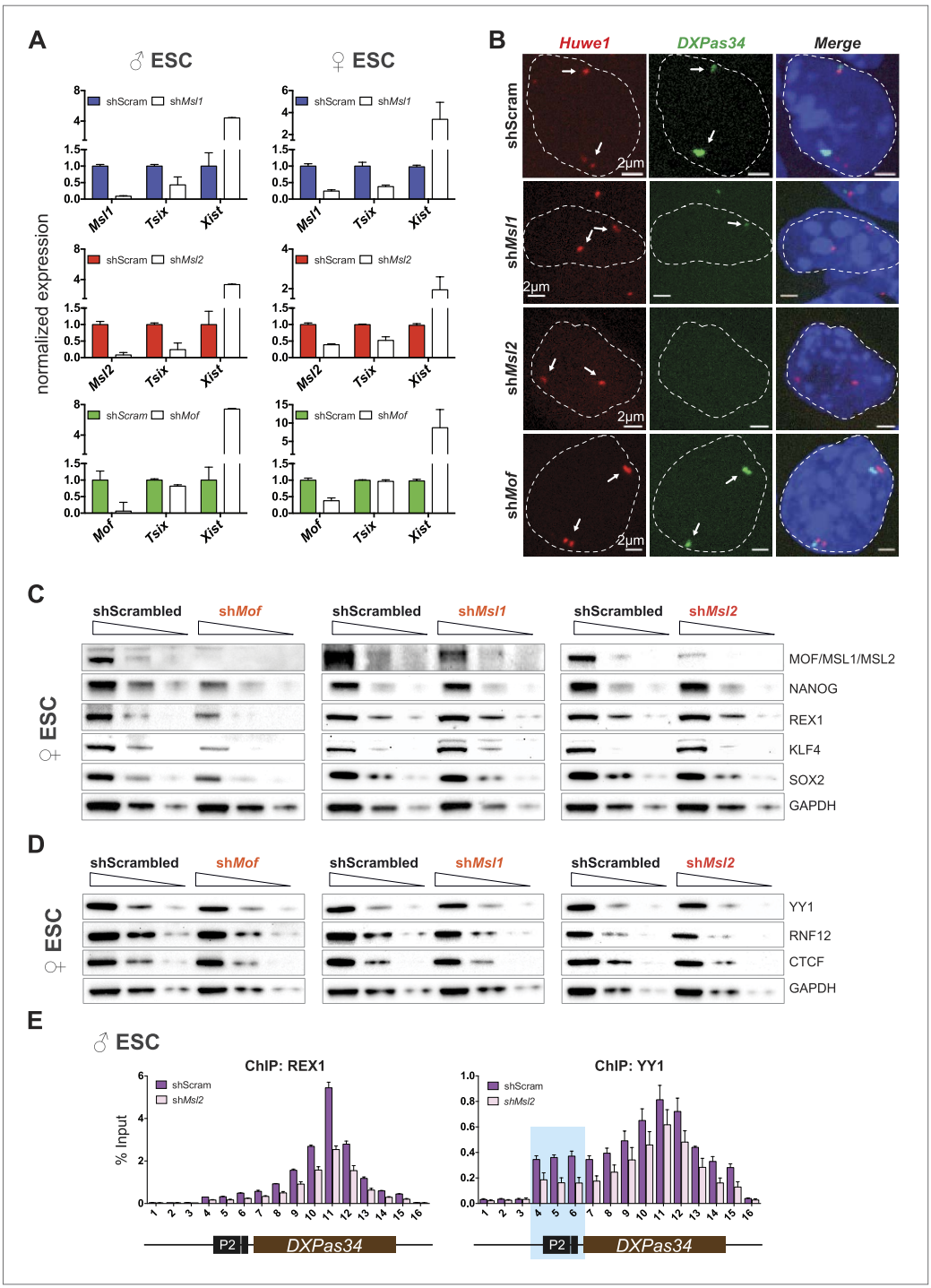


Figure 6. Depletion of MSL1 and MSL2 leads to downregulation of *Tsix* with concomitant upregulation of *Xist*. (**A**) Gene expression analysis for the indicated genes in male and female ESCs treated with scrambled RNA (shScram) or shRNA against *Msl1*, *Msl2*, or *Mof*. All results are represented as relative values normalized to expression levels in shScram (normalized to *Hprt*) and expressed as means ± SD in three biological replicates. (**B**) RNA-FISH

Figure 6. Continued on next page

Figure 6. Continued

for *Huwe1* (red) and *DXPas34* (green) in: scrambled control, sh*Msl1*-, sh*Msl2*-, and sh*Mof*-treated female ESCs. Nuclei were counterstained with DAPI (blue). White arrows denote foci corresponding to *Huwe1* or *Tsix*; dashed lines indicate nuclei borders. For additional images, phenotypes and quantifications see ***Figure 6—figure supplement 1A–C***. For probe references see 'Materials and methods'. (**C**) Western blot analyses of the pluripotency factors in scrambled-, *Mof*-, *Msl1*-, and *Msl2*-shRNA-treated female ESCs. For corresponding expression analyses see ***Figure 6—figure supplement 1D,E***. The respective dilution (100%, 30%, 10%) of loaded RIPA extracts is shown above each panel. GAPDH was used as the loading control. For antibodies see 'Materials and methods'. (**D**) Western blot analyses of the transcription factors involved in regulation of the XIC in scrambled-, *Mof*-, *Msl1*-, and *Msl2*-shRNA-treated female ESCs. The respective dilution (100%, 30%, 10%) of loaded RIPA extracts is shown above each panel. GAPDH was used as the loading control. (**E**) ChIP-qPCR analysis of REX1 (left panel) and YY1 (right panel) across the *Tsix* major promoter (P2) and *DXPas34* in male ESCs treated with the indicated shRNAs. The labels of the x axes correspond to the arrowheads in ***Figure 5A***. For all ChIP experiments, three biological replicates were used; results are expressed as mean ± SD; cells were harvested on day 4 (*Msl2*) or 5 (*Mof*) after shRNA treatment.

DOI: 10.7554/eLife.02024.020

The following figure supplements are available for figure 6:

Figure supplement 1. Cells depleted of MSL1 or MSL2, but not MOF show loss of *DXPas34* foci.

DOI: 10.7554/eLife.02024.021

Taken together, we detect direct binding of MSL complex members to several loci within the X inactivation center including the *Tsix/Xist* locus. Depletion of MSL1 or MSL2, but not MOF led to severe downregulation of *Tsix* expression while depletion of MOF, MSL1, or MSL2 resulted in elevated *Xist* levels. These results indicate a direct regulatory function of MSL1 and MSL2 on the *DXPas34* locus and an indirect NSL-associated MOF effect on *Xist* expression through the pluripotency network.

Depletion of MSL1 and MSL2 leads to impaired recruitment of REX1 and YY1 to regulatory regions of *Tsix*

As loss of MSL1 and MSL2 did not affect the core pluripotency network, we set out to explore what might be the impact of MSL depletion on XIC genes (other than *Tsix* and *Xist*) and transcription factors involved in their regulation. As shown in ***Figure 6—figure supplement 1E***, we observed mild effects on the expression of XIC-encoded genes involved in the regulation of X inactivation. Only depletion of MSL2 led to significant downregulation of *Ftx* and *Jpx* genes whose promoters were bound by MSL1 and/or MSL2 (***Figure 5A***). On the other hand, depletion of MOF led to moderate upregulation of *Linx* lncRNA, which acts synergistically with *Tsix* (***Nora et al., 2012***).

Neither the depletion of MSL1 and MSL2 nor the depletion of MOF significantly influenced protein levels of RNF12, YY1, or CTCF that are known regulators of the XIC (***Figure 6D***; ***Donohoe et al., 2007***, ***2009***; ***Jonkers et al., 2009***; ***Shin et al., 2010***; ***Jeon and Lee, 2011***). Since REX1 and YY1 bind and regulate the *Tsix* locus (***Donohoe et al., 2007***; ***Gontan et al., 2012***), we subsequently tested whether MSL depletion would affect the recruitment of these factors to the *Tsix* major promoter and *DXPas34*. Indeed, the depletion of MSL2 led to significant reduction of REX1 ChIP signals across the *DXPas34* locus whereas the effect on YY1-targeting was less pronounced and restricted to the *Tsix* major promoter (P2) (***Figure 6E***).

Knockdown of *Msl1* and *Msl2* results in enhanced accumulation of *Xist* and X-chromosomal coating in differentiating female ESCs

We next assessed the consequence of MSL-dependent reduction of *Tsix* levels and concomitant upregulation of *Xist* at a cellular level using RNA-FISH for *Xist* upon depletion of individual MSL complex members (for probe reference see 'Materials and methods'). Interestingly, we observed accumulating *Xist* lncRNA and X-chromosomal coating in a small fraction of MSL1- and MSL2-depleted female ESCs (but not MOF-depleted cells; 4–5% of the cell population in sh*Msl1* and sh*Msl2* with comparison to 0.5% in scrambled control, see ***Figure 7A*** and ***Figure 7—figure supplement 1A–C***). These findings suggest that the MSL1- and MSL2-dependent downregulation of *Tsix* is sufficient to cause occasional accumulation of *Xist* lncRNA in undifferentiated female ESCs. The different outcomes following MOF and MSL1/MSL2 depletion on *Xist* confirmed the notion that MOF and MSL1/MSL2 influence the XIC via different mechanisms.

Previous studies have shown that the effects of *Tsix* depletion on *Xist* accumulation and X inactivation become fully apparent after induction of differentiation (***Clerc and Avner, 1998***; ***Debrand et al., 1999***; ***Lee and Lu, 1999***; ***Luikenhuis et al., 2001***; ***Ohhata et al., 2006***; ***Sun et al., 2006***).

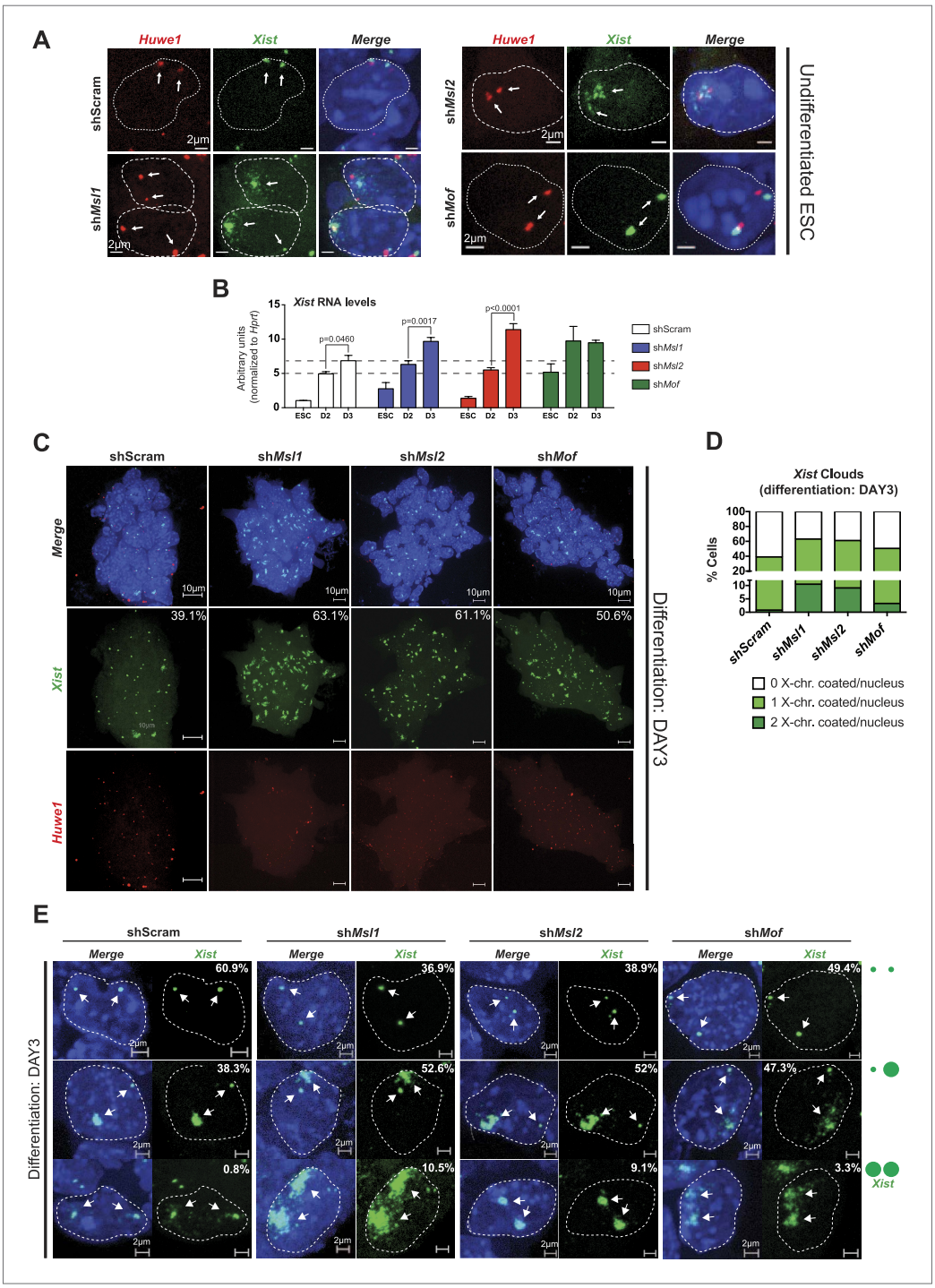


Figure 7. MSL1 and MSL2 depletion leads to enhanced and chaotic *Xist* accumulation in early differentiation. (**A**) RNA-FISH for *Huwe1* (red) and *Xist* (green) in: scrambled control, sh*Msl1*-, sh*Msl2*-, and sh*Mof*-treated female ESCs. Nuclei were counterstained with DAPI (blue). White arrows denote foci corresponding to *Huwe1* or *Xist*; dashed lines indicate nuclei borders. For additional images, phenotypes and quantifications see ***Figure 7—figure supplement 1B–D***.

Figure 7. Continued on next page

Figure 7. Continued

For probe references see 'Materials and methods'. (**B**) Expression analysis for *Xist* in undifferentiated, day 2 (D2) and day 3 (D3) differentiating female ESCs treated with scrambled RNA (shScram) or shRNA against *Mof*, *Msl1*, and *Msl2*. All results are represented as arbitrary units (*Xist* expression in undifferentiated ESCs = 1) normalized to expression levels in shScram (normalized to *Hprt*) and expressed as means ± SD in three biological replicates. p-values for D2-to-D3 expression change were obtained using unpaired *t* test. (**C**) RNA-FISH for *Huwe1* (red) and *Xist* (green) in: scrambled control, sh*Msl1*-, sh*Msl2*-, and sh*Mof*-treated differentiating female ESCs. Nuclei were counterstained with DAPI (blue). RNA-FISH was performed on the sixth day of knockdown (after 72 hr of differentiation). Percentages indicate number of cells with at least one *Xist* cloud for each of the knockdowns. For additional images of multicellular colonies see ***Figure 7—figure supplement 2A***. (**D**) Bar plot summarizing the percentage of *Xist* clouds for individual knockdowns in differentiating (DAY3) female ESCs for individual knockdowns. Cells were divided into three categories: cells carrying no *Xist* clouds (white), single *Xist* cloud (light green), or two *Xist* clouds (dark green). For quantifications, see ***Figure 7—figure supplement 2B***. (**E**) RNA-FISH for *Xist* (green) in: scrambled control, sh*Msl1*-, sh*Msl2*-, and sh*Mof*-treated differentiating (DAY3) female ESCs. Here, we show examples of individual nuclei carrying different patterns of *Xist* accumulation. Percentages correspond to the frequency of the shown *Xist* pattern within the population of cells. White arrows denote *Xist* foci; dashed lines indicate nuclei borders. For quantifications see ***Figure 7—figure supplement 2B***.

DOI: 10.7554/eLife.02024.022

The following figure supplements are available for figure 7:

Figure supplement 1. Depletion of MSL1 and MSL2 leads to occasional accumulation and spreading of *Xist* in undifferentiated ESCs.

DOI: 10.7554/eLife.02024.023

Figure supplement 2. Depletion of MSL1 and MSL2 lead to enhanced *Xist* accumulation in differentiating ESCs.

DOI: 10.7554/eLife.02024.024

We therefore depleted MSL1, MSL2 and MOF, and induced differentiation for 3 days by withdrawing LIF and placing the ESCs in N2B27 media. Consistent with our previous results, the induction of differentiation resulted in a stronger elevation of *Xist* RNA levels in MSL1- and MSL2-depleted cells in comparison to the scrambled control (***Figure 7B***). As *Tsix* expression was not affected in MOF-depleted ESCs and *Xist* levels were already high before induction of differentiation, *Xist* upregulation between day 2 and 3 of differentiation was similar to the scrambled control.

To monitor the effect on the X chromosome more closely, we next performed *Xist* RNA-FISH in MSL1-, MSL2- and MOF-depleted cells after 3 days of differentiation. All three knockdowns resulted in enhanced *Xist* accumulation and X-chromosomal coating (63.1%, 61.1% and 50.6% of all counted cells in sh*Msl1*-, sh*Msl2*-, and sh*Mof*-treated ESCs, respectively, in comparison to scrambled control with 39.1% of counted cells; see ***Figure 7C,D*** and ***Figure 7—figure supplement 2A,B***). Interestingly, we observed that MSL1- and MSL2-depleted differentiating cells contained numerous cells with two inactive X chromosomes. The fraction of cells where both X chromosomes underwent XCI was approximately 10-fold higher in *Msl1* and *Msl2* knockdown compared to the scrambled control (***Figure 7E***). These results are in agreement with previously published data from homozygous *Tsix* mutants that exhibit irregular, 'chaotic' choice for X inactivation (***Lee, 2005***).

Taken together, our data establishes MSL1 and MSL2 among the key regulators of *Tsix* transcription, as the depletion of MSL proteins results in severe downregulation of *Tsix* transcription and enhanced accumulation of *Xist* during early differentiation.

Discussion

We present a thorough characterization of the histone acetyltransferase MOF and its two known complexes in mouse embryonic stem cells (ESCs) and neuronal progenitor cells (NPCs). We determined five basic modes of co-occurrence that revealed cell-type-specific as well as constitutive functions of the different proteins and support the notion that the NSL complex has general, housekeeping functions whereas the MSL complex predominantly performs more specialized tasks. We show that MOF and its associated proteins are involved in gene expression regulation via different means: first, they all target the promoters of housekeeping genes in a cell-type-independent manner and second, members of both complexes occupy different sets of ESC-specific enhancers that are essential for the maintenance of stem cell identity. We demonstrate the distinct and novel functions carried out by the MOF-associated complex members by revealing that both complexes contribute to the repression of X inactivation in ESCs via different means: While we establish the MSL complex as a direct regulator of *Tsix*, MOF and the NSL complex play an important role in the maintenance of pluripotency factors (***Figure 8***).

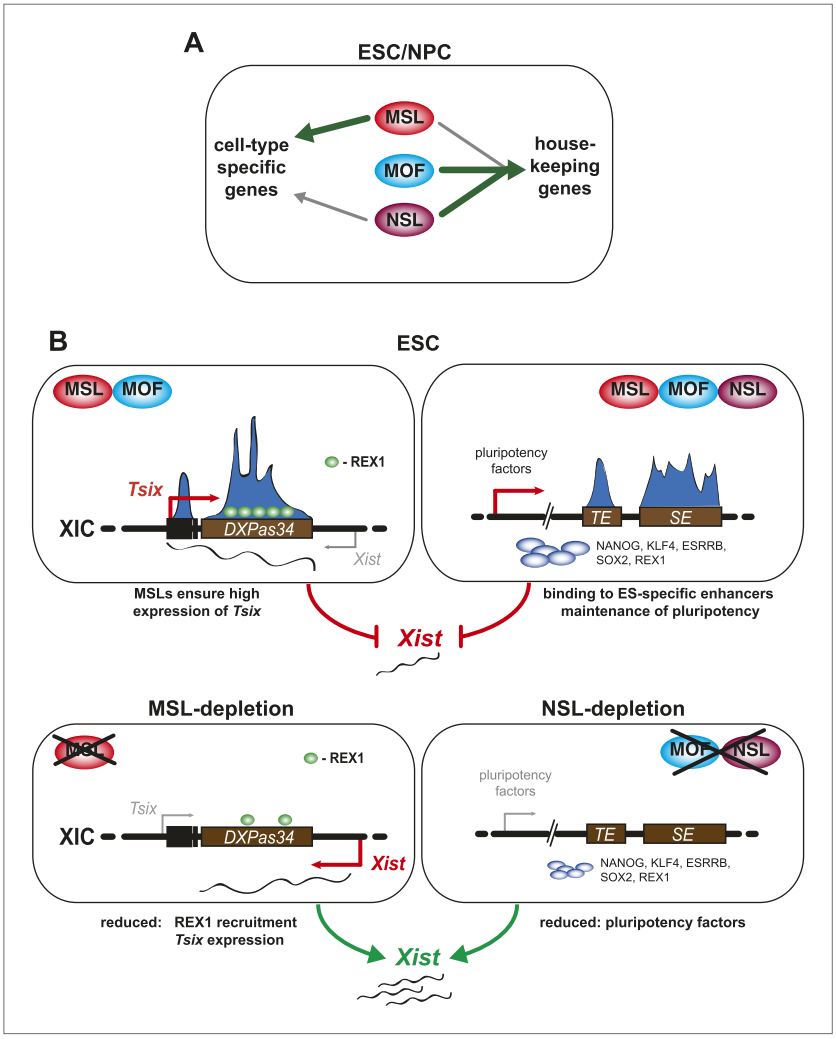


Figure 8. A summary model. Shared and distinct pathways by which MOF, MSLs and NSLs regulate gene expression, pluripotency, and the X inactivation center. (**A**) In this study, we have identified several modes of concurrent and independent binding of mammalian MOF, MSL and NSL proteins. We find that all complexes bind to promoters of housekeeping genes in ESCs and NPCs with NSL complex members occupying the majority of the target genes, while MOF and MSL proteins bind NSL-bound genes in a more restricted manner. Furthermore, we observe that upon differentiation, KANSL3 and MSL2 additionally occupy TSSs of different sets of cell-type-specific genes in the absence of MOF. (**B**) When we studied the functions of MSL and NSL complexes at the murine X inactivation center, we determined two basic mechanisms by which the different proteins affect the maintenance of two active X chromosomes in ESCs. (1) MSLs bind to the promoter and enhancer of *Tsix* whose transcription represses *Xist* expression. Upon depletion of MSLs, *Tsix* expression is compromised, so is REX1 recruitment to the *Tsix* locus. Consequently, *Xist* is increasingly transcribed and can occasionally accumulate. (2) In addition, MOF, MSLs, and NSLs bind to typical enhancers (TE) and super enhancers (SE) in ESCs, and notably those of pluripotency factors. In WT ESCs, the high expression of pluripotency factors is another layer of *Xist* repression. The depletions of MOF or KANSL3, but not of MSL1 or MSL2 reduce the expression of pluripotency factors involved in *Xist* repression causing a *Tsix*-independent increase of *Xist* expression.
DOI: 10.7554/eLife.02024.025

Global effects of MOF are correlated with the NSL complex

Our study sheds light on the interplay between MOF and its complexes in mammals. Despite the fact that the depletion of KANSL3 does not strongly reduce global H4K16 acetylation levels, we

observed strikingly similar protein and transcriptome changes in KANSL3- or MOF-depleted cells (***Figure 3E–G***). On the other hand, MSL1- and MSL2-depletion caused marked decreases of H4K16 acetylation (***Figure 3G***). This is consistent with previous reports that established MSL proteins as the main enhancers of MOF's H4K16 acetylation activity, while the NSL complex was shown to possess broader substrate specificity and can crosstalk with histone methylases (***Cai et al., 2010***; ***Kadlec et al., 2011***; ***Zhao et al., 2013b***). Unexpectedly, we observed remarkably different phenotypic changes in MSL1- or MSL2-depleted cells compared to MOF- and KANSL3-depleted cells (***Figure 3E***, ***Figure 3—figure supplement 3A***). A striking example was the strong reduction of key pluripotency factors in KANSL3- and MOF-depleted cells that remain unaffected in MSL1- and MSL2-knockdowns (***Figure 3G***, ***Figure 4F***). These results support the recent finding that MOF is vital for the maintenance of pluripotency (***Li et al., 2012***), but we furthermore show that this is an NSL- and not MSL-related function of MOF independent of H4K16 acetylation deposition.

Taken together, our data shows that while MOF is the major acetyltransferase for lysine 16 of histone 4 (***Taipale et al., 2005***), MSL-dependent H4K16 acetylation is one of several means through which MOF exerts its crucial biological functions. This notion was further supported by the finding that MOF predominantly binds to promoters of broadly expressed genes as part of the NSL complex and subsequently supports their transcription (***Figure 3A–F***). MSL1 and MSL2, on the other hand, bound to a relatively small subset of broadly expressed MOF-NSL-targeted genes that were significantly stronger expressed than those where MOF was exclusively present with NSL complex members (***Figure 3B***, ***Figure 3—figure supplement 1B***). The additive effects of the complexes on gene expression were intriguing, and whether they influence each other's activity or exert their functions separately should be studied in the future. We propose that the MSL complex fine-tunes MOF's activity and ensures precise regulation of more specific targets—after all, their presence is essential for the recruitment of MOF to NSL-independent targets (***Figure 5B***). Our model is surprisingly similar to the picture that is emerging from *Drosophila* research where the NSL complex regulates housekeeping genes (***Feller et al., 2012***; ***Lam et al., 2012***) while the MSL complex fulfills a highly specialized role on the male X chromosome (reviewed in ***Conrad and Akhtar, 2011***).

MSL2 and KANSL3 can contribute to transcription via enhancer binding

In addition to insights about MOF-related functions of MSL and NSL complexes, we show for the first time additional binding of MSL and NSL proteins to TSS-distal regions with enhancer characteristics. On a global scale, MOF did not yield strong enrichments for canonical enhancers; however, both MSL2 and KANSL3 showed robust signals for TSS-distal regions in ESCs, but not in NPCs, which reflected the transcriptional activity of these regions (***Figure 2***, ***Figure 4A***). This apparent MOF-independent binding of the individual proteins (that tended to prefer different sets of enhancers; ***Figure 4B***) suggests that KANSL3 and MSL2 stimulate transcription even in the absence of the histone acetyltransferase. Both proteins are in principle capable of supporting transcription: the *Drosophila* homologue of KANSL3 can directly activate transcription in vitro (***Raja et al., 2010***) and human MSL2 acts as an E3 ubiquitin ligase at lysine 34 of H2B (H2BK34ub) (***Wu et al., 2011***), which has been suggested to promote methylation of H3K4, and thus gene expression (***Wu et al., 2011***). Indeed, we observed several hundred genes that had been predicted to be regulated by TSS-distal binding sites of MSL2 or KANSL3 to be downregulated in the respective knockdowns with particularly high frequencies in MSL2-depleted cells (***Figure 4E***). It is important to note that the subset of ESC enhancers for key pluripotency factors (e.g., *Klf4*, *Sox2*) were bound concomitantly by KANSL3 and MSL2 and only the depletion of KANSL3, but not of MSL1 or MSL2 diminished protein and transcript levels of these key ESC molecules (see above). It is possible that KANSL3 could rescue loss of MSL2 at certain loci, but the exact mechanisms through which KANSL3 affects transcription via enhancer-binding need to be studied further. Furthermore, the pluripotency network and/or Mediator-related functions at super enhancers may be sufficient and dominant over MSL2 to maintain the expression of the pluripotency factors in the absence of MSL2, but may well be dependent on the function of KANSL3 at these regions.

MSL1 and MSL2 repress X inactivation by regulating *Tsix* expression

When we specifically searched for regions where KANSL3 was not present together with MSL1 and MSL2, we found that the X inactivation center (XIC) showed numerous signals for the MSL complex (***Figure 5***). The XIC, a hot-spot of regulatory lncRNAs, is an X-chromosomal region that contains the main regulators of X chromosome inactivation (XCI). The proper function of XIC-located non-coding

RNAs is influenced by the spatial organization of the XIC and governed by a sophisticated interplay of multiple transcription factors such as pluripotency factors (**Donohoe et al., 2007**; **Navarro et al., 2010**; **Deuve and Avner, 2011**; **Gontan et al., 2012**; **Nora et al., 2012**).

We found that depletion of MSL1 and MSL2 severely reduced *Tsix* expression in male and in female ESCs, moderately increased *Xist* levels (**Figure 6A**), but left pluripotency factors unaffected (**Figure 6C**). In contrast, MOF-depleted cells showed downregulation of pluripotency factors and much higher *Xist* levels. Previous studies demonstrated that in undifferentiated ESCs, where pluripotency factors are highly abundant, even severe downregulation of *Tsix*, or *Tsix*-deletion has almost no effect on *Xist* transcription (**Morey et al., 2001**; **Navarro et al., 2005**; **Nesterova et al., 2011**). Thus, the pronounced *Xist* upregulation seen in MOF-depleted cells seems to be an indirect effect due to the downregulation of pluripotency factors, while the reduction of *Tsix* transcripts in MSL1- and MSL2-depleted cells, where pluripotency factors remain unaffected, has milder consequences on *Xist* levels.

Consequently, we could show that once ESCs are forced to initiate differentiation, the depletion of MOF has mild effects while MSL1- and MSL2-depleted cells, in which *Tsix* expression is prematurely downregulated, indeed suffer from enhanced *Xist* accumulation accompanied by 'chaotic' X inactivation (different numbers of inactivated X chromosomes within a population of cells; **Figure 7B–E**). This is consistent with the notion that the repressive potential of *Tsix* on *Xist* accumulation and the role of *Tsix* and the *DXPas34* locus in the process of counting and choice of XCI (**Lee, 2005**; **Vigneau et al., 2006**) becomes fully apparent during early stages of differentiation where additional repressive factors such as pluripotency factors are downregulated (reviewed in **Rougeulle and Avner, 2004**).

Conclusion

We show that NSL and MSL complex members can function in concert to ensure proper regulation of gene expression, but our findings also strongly imply that members of both complexes have the capacity to act independently. In the case of the X inactivation center, we observe that the MOF-interacting proteins, despite engaging different regulatory means (MSL1, MSL2 through direct regulation of *Tsix*, and MOF-NSL through the pluripotency network) synergize to ensure the proper expression of the X chromosomes in undifferentiated ES cells (**Figure 8**). Our study sets the ground for future research to dissect the intricate interactions and specific functions of MOF and its associated major regulatory proteins in more detail.

Materials and methods

Cell culture

All cell culture was performed in a humidified incubator at 37°C and 5% CO_2. The feeder-dependent mouse female embryonic stem cell line F1-21.6 was cultivated on mitomycin-C-inactivated or irradiated mouse embryonic fibroblasts (MEFs). The feeder-independent mouse male ES cell line WT26, a kind gift from the lab of Thomas Jenuwein, was cultivated on gelatin-coated dishes in ESC culture media KnockOut-DMEM (Gibco, Carlsbad, CA) supplemented with 1% L-glutamine (Gibco), 1% penicillin/streptomycin (Gibco), 1% non-essential amino acids (Gibco), 1% sodium pyruvate, 1% 2-mercaptoethanol. All ESC media contained 15% FBS and 1000 U/ml (for feeder-dependent) or 2000 U/ml (for feeder-) of leukemia inhibitory factor.

Male and female neuronal progenitor cell (NPC) lines were derived from previously mentioned ES cell lines (see below). Mouse 3T3 cells (for luciferase assays) and human HEK293-FT cells (for lentiviral production) were cultivated in DMEM (high glucose, with glutamine, Gibco) supplemented with 10% heat-inactivated serum (PAA Laboratories, North Dartmouth, MA), 1% L-glutamine, 1% penicillin/streptomycin.

NPC differentiation

Mouse ESCs were differentiated into neuronal progenitor cells (NPC) as previously described (**Conti et al., 2005**; **Splinter et al., 2011**). In brief, 1 × 10^6 ESCs (deprived of feeder cells) were plated on 0.1% gelatin-coated dishes in N2B27 medium and cultured for 7 days with daily media changes. The cells were then dissociated from the plate using accutase (Sigma, Germany) and 3 × 10^6 cells were plated on a bacterial petri dish to induce formation of embryoid bodies in N2B27 medium supplemented with 10 ng/ml EGF and FGF2 (Peprotech, Rocky Hill, NJ). After 72 hr, embryoid bodies were transferred to 0.1% gelatin-coated dishes to allow adhesion and expansion of NPCs from the embryoid bodies. NPC lines were maintained in N2B27 medium supplemented with EGF and FGF2

(10 ng/ml each), on 0.1% gelatin-coated flasks. For FISH analysis, F1-21.6 ESCs were grown on gelatin-coated coverslips with a MEF-inactivated monolayer for 24 hr.

Western blot analysis

The Invitrogen precast gel system NuPAGE was used for SDS-PAGE. The 4–12% Bi–Tris gradient gels (for proteins above 20 kDa) or 12% Bis–Tris gels (for histones and histone marks) were loaded with samples supplemented with Roti-Load 1 sample buffer. After blotting, the membranes were blocked in 5% milk with PBS + 0.3% Tween-20 (PBST) mix for at least 1 hr at room temperature. Membranes were then incubated overnight with the primary antibody in 0.5% milk with PBST at 4°C. The next day, membranes were washed three times for 10 min in PBST, incubated with a suitable HRP-coupled secondary antibody for 1 hr at room temperature, washed thrice and proteins were visualized with Lumi-Light Plus Western Blotting Substrate using the Gel Doc XR+ System.

Immunoprecipitation assays (IP and ChIP)

For (co)immunoprecipitation (IP, co-IP) experiments, 1 ml of nuclear extract (0.5 mg/ml) was used. IPs were performed in IP buffer (25 mM HEPES pH 7.6, 150 mM KCl, 5 mM MgCl2, 0.5% Tween20, 0.2 mg/ml BSA, 1× complete protease inhibitors tablet). Extracts were incubated with 5 µg of the respective antibody or normal-rabbit/normal rat serum. For MSL1 15 µl of antibody serum was used. Extracts were incubated with the antibody for 2 hr, rotating at 4°C. Protein-A Sepharose beads (GE Healthcare, United Kingdom), blocked with 1 mg/ml yeast tRNA and 1 mg/ ml BSA (NEB, Ipswich, MA), were used for all ChIP and IP assays.

Chromatin immunoprecipitation (ChIP) assays were performed as previously described (***Pauli, 2010***) with minor changes. Cells were fixed in 1% molecular biology grade formaldehyde (Sigma) 9 min before being quenched with glycine (0.125 M final concentration). Cells were washed twice with ice-cold PBS and lysed on ice for 10 min with 10 ml of Farnham lysis buffer (5 mM PIPES pH 8.0, 85 mM KCl, 0.5% NP-40 + Roche Protease Inhibitor Cocktail Tablet, filtered through 0.2 micron filter unit). Lysates were transferred to a Kontes dounce tissue grinder (K885300-0015, size B) and dounced 15 times in order to break the cells and keep nuclei mostly intact. Crude nuclear prep was transferred to 15-ml falcon tube and nuclei pelleted by centrifugation at 2000 rpm at 4°C for 5 min. Nuclei were resuspended in RIPA lysis buffer (1 × PBS, 1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS + Roche Protease Inhibitor Cocktail Tablet, filtered through 0.2 micron filter unit). The nuclear extract was subjected to chromatin shearing using the Diagenode Bioruptor Plus sonicator (at high setting for a total time of 25 min, 30 s ON, 30 s OFF). The sonicated mixture was centrifuged at 14,000 rpm at 4°C for 5 min and supernatant was collected. Chromatin was supplemented with 5 µg of primary antibody and incubated for 16 hr (antibodies used for ChIP are listed below). After incubation, 50 µl of 50% slurry bead solution was added for another incubation period (2 hr), then beads were washed: four times for 15 min with RIPA lysis buffer, two times for 1 min with LiCl IP wash buffer (250 mM LiCl, 10 mM Tris–HCl pH 8.0, 1 mM EDTA, 0.5% NP-40, 0.5% DOC, filtered through 0.2 micron filter unit), two times for 1 min with TE buffer (1 mM Tris–HCl pH 8.0, 1 mM EDTA, filtered through 0.2 micron filter unit). Washed beads were resuspended in 100 µl of IP elution buffer and subjected to overnight reverse cross-linking (RNase and proteinase K digestions) followed by DNA purification (DNA was purified using Minelute PCR purification kit from Qiagen, Germany). For single IP assay 50 µl of bead solution was used. Purified ChIPed DNA was subjected to qPCR amplification (Applied Biosystems, Carlsbad, CA). Input was used for normalization control. For primer pairs see ***Supplementary file 3***.

Antibodies

For MSL1 antibody production, a GST-mMSL1 fusion protein (C-terminal, residues 254–616) was used to immunize rabbits; the final bleed was used in experiments. Antibody specificity was verified with IP and MSL1-specific RNAi followed by Western blot analysis and ChIP assay. We used several commercial antibodies: a-KANSL1 (PAB20355; Abnova, Taiwan), a-KANSL3 (HPA035018; Sigma), a-MCRS1 (11362-1-AP; Proteintech, Chicago, IL), a-MOF (A3000992A; BETHYL Montgomery, TX), a-MSL2 (HPA003413; Sigma), a-NANOG (A300-397A; BETHYL), a-OCT3/4 (sc-5279; Santa-Cruz Dallas, TX), a-REX1 (Ab28141; Abcam, England), a-ESRRB (PP-H6705-00; Perseus Proteomics, Japan), a-KLF4 (Ab72543; Abcam), a-SOX2 (AF2018; R&D Systems, Minneapolis, MN), a-YY1 (A302-779A; BETHYL), a-RNF12/RILM (16121-1-AP; Proteintech) a-GAPDH (A300-639A; BETHYL), a-NESTIN (Ab93666; Abcam), a-CTCF (Ab70303; Abcam), a-H3 (Ab1791; Abcam), a-H4 (Ab10158; Abcam), a-H4K16ac (07-329; Millipore, Billerica, MA).

Luciferase assays

Enhancer candidate regions (see below) were cloned into the firefly luciferase plasmids (pGL4.23; Promega, Witchburg, WI) and transfected into mouse ESCs and 3T3 fibroblasts using Lipofectamine-2000 reagent and into NPCs using LTX-PLUS reagent (Invitrogen). Transfections were performed according to the manufacturer's guidelines except for using a 1:6 DNA to Lipofectamine ratio. Cells were seeded 1 day prior to transfection to achieve 70–80% confluency at the time of transfection. Next, cells were fed with antibiotics-free medium (ES medium with LIF for ESCs and OPTIMEM for NPCs and 3T3s) at least 30 min before transfection and the medium was changed back 6–8 hr after transfection (basal neural medium with FGF and EGF for NPCs). 100 ng of firefly construct with the cloned candidate region was co-transfected with 1 ng of renilla luciferase construct (pRL-TK of Promega) per 96-well and harvested for luciferase assay after 24 hr. Cells were harvested for luciferase assay 24 hr after transfection. The Dual Luciferase Kit (Promega) was used according to the manufacturer's protocol but with reduced substrate volumes of LARII and Stop&Glo reagents (50 µl per well of a 96-well plate with 10 µl cell lysate). Luminescence was measured by using Mithras plate reader (Berthold, Germany).

The transfection efficiency was normalized by firefly counts divided by the renilla counts. The fold enhancement value was calculated by an additional normalization to minimal promoter alone activities in each experiment (the graphs represent at least three independent experiments that were performed in technical triplicates each with error bars representing standard error of the mean). The following enhancer candidate regions were amplified from mouse genomic DNA by PCR and cloned into BamHI-SalI sites (downstream of luciferase gene) of firefly luciferase plasmid pGL4.23 (Promega):

Intron of *Esrrb* (chr12:87,842,537-87,843,719) with primers introducing BamHI and XhoI sites: ATAGGATCCGAAGTAATTGTCTATTGTATCAG (forward), TATCTCGAGAAGAAGAAAGACTGTGTTCAACTCC (reverse).

Upstream of *Lefty* (chr1: 182854617-182855516) with primers introducing BamHI and SalI sites: ATAGGATCCCTTGCGGGGGATATGAGGC (forward), TATGTCGACCTGGGCCTTTCTAAGGC (reverse).

Upstream of *Trim28* (*Kap1*) (chr18: 34309039-34310140) with primer introducing BamHI and SalI sites: ATAGGATCCGAGGACTATTTGAAGGATCTATT (forward), TATGTCGACCTCACTCCCCAACCTCCATTTC (reverse).

Upstream of *Apc* (chr18: 34309039-34310140) with primers introducing BamHI and SalI sites: ATAGGATCCCTGAGCAATGCTCTTCCTCACAAGC(forward),TATGTCGACTTATACTCCAAATAGAATTGTCTG (reverse).

Intron of *Tbx3* (chr5: 120129690- 120130617) with primers introducing BamHI and SalI sites: ATAGGATCCATAAATAAATAAATAAATATCTGATTG (forward), TATGTCGACCGCGAGTCTGGCGATGCCTTGTC (reverse).

RNA extraction followed by cDNA synthesis and quantitative real time PCR

cDNA was synthesized from 500 ng–1 µg of total RNA (extracted from circa 1 million cells using Rneasy kit, Qiagen) with random hexamers using SuperScript-III First Strand Synthesis kit (Invitrogen). The qPCRs were carried in a total reaction volume of 25 µl containing 0.5–1 µl of cDNA, 0.4 µmol of forward and reverse primer mix and 50% 2 × SYBR Green PCR Master Mix (Roche). Gene expression was normalized to multiple controls (*RplP0* or *Hprt*), using the 7500 software V2.0.4 for analysis (Applied Biosystems). For primer pairs used for expression profiling see ***Supplementary file 3C***.

Lentiviral-based RNAi in ESCs

shRNA constructs were either obtained from Sigma in pLKO.1 or designed using Genscript and cloned (please see below for details). For cloning, forward and reverse complimentary DNA oligonucleotides (Eurofins MWG Operon, Germany) designed to produce AgeI (5′) and EcoRI (3′) overhangs were annealed at a final concentration of 2 µM in NEBuffer. The pLKO.1-puro plasmid was digested with AgeI and EcoRI, ligated to the annealed oligonucleotides, and transformed into HB101 competent cells (Promega). Plasmid DNA was purified using the QIAprep Spin Miniprep kit (Qiagen), and the sequence was validated.

For production of lentiviral particles, 70% confluent HEK293FT cells in a 10-cm tissue culture plate were co-transfected with 3.33 µg lentiviral construct, 2.5 µg psPAX2 packaging plasmid and 1 µg pMD2.G envelope plasmid using Lipofectamine-2000 reagent (Invitrogen). To transduce ESCs, either

concentrated or diluted lentiviral particles were used. For concentrated lentivirus, transfections were scaled up and OPTIMEM (Invitrogen) added to the HEK293FT cells following transfection and the lentiviral supernatant collected at 48 and 72 hr post-infection. This was then concentrated using Amicon Ultra-15 centrifugal filter units (Millipore) and added to ESC media supplemented with LIF and 10 μg/ml polybrene (Millipore). For diluted lentivirus, ESC media without LIF was added to the HEK293FT cells and the lentiviral supernatant was collected after 48 hr, filtered through 0.22 μm filters (Whatmann), and added 1:1 with fresh ESC media supplemented with LIF and polybrene to the ESCs. ESCs were then subjected to selection with 1.0 μg/ml puromycin, passaged once, and harvested on day 3, 4, 5 or 6 of knockdown depending on the experiment (the numbers of days are indicated in the corresponding results section).

The following shRNA sequences were used for the knockdowns:

CCGGCCTAAGCACTCTCCCATTAAACTCGAGTTTAATGGGAGAGTGCTTAGGTTTTTG (sh*Msl1*, SIGMA, TRCN0000241378),
CCGGCCCAGTCTCTTAGCCATAATGCTCGAGCATTATGGCTAAGAGACTGGGTTTTTG (sh*Msl2*, SIGMA, TRCN0000243429),
CCGGAAGGCCGAGAAGAATTCTATCTCGAGATAGAATTCTTCTCGGCCTTTTTTTG (sh*Mof*, GENSCRIPT designed),
CCGGCTCCAGTCCTCTTCGTCATTGCTCGAGCAATGACGAAGAGGACTGGAGTTTTTG (sh*Kansl3*, SIGMA, TRCN0000266995),
CCGGAAGTGGCGCCTTAGCAACAACCTCGAGGTTGTTGCTAAGGCGCACTTTTTTTG (sh*Mcrs1*, GENSCRIPT designed),
CCGGCAACAAGATGAAGAGCACCAACTCGAGACAATTCGGAAGAAATCTGAGCTTTTTG (Non-targeting control, SIGMA, SHC002).

Cell proliferation assay

Cells treated with respective shRNAs and scramble control were performed as described earlier in feeder-free W26 mouse ESCs. The cell count was monitored for 6 days post knockdown at 24-hr intervals. In brief, after 4 days of knockdown six sets of 0.4×10^4 cells per well were seeded in triplicates in a 12-well gelatinized plate. The cells were grown in ES cell culture medium supplemented with 2000 U/ml LIF and 1 μg/ml puromycin; the medium was changed every 24 hr. For counting, cells were trypsinized and counted using the Neubauer hemocytometer.

Alkaline phosphatase staining

Detection of alkaline phosphatase, a surface marker and indicator of undifferentiated ESCs, was performed using the following method: feeder-free W26 ESCs were transduced (4 days) with scramble or the shRNAs against the genes of interest. Cells were washed twice with PBS followed by fixation with 4% PFA for 2–3 min. The cells were washed twice with PBS and stained for 20 min with staining solution (25 mM Tris-Maleic acid buffer pH 9.0, 0.4 mg/ml α-Naphthyl Phosphate (Sigma), 1 mg/ml Fast Red TR Salt (Sigma), 8 mM $MgCl_2$, 0.01% Deoxycholate, 0.02% NP40). The reaction was stopped by washing with water followed by two washes with 1 × PBS.

RNA extraction for RNA-seq

Total RNA was extracted from WT26 ESCs and NPCs as biological triplicates using TRIzol Reagent and treated with the TURBO Dnase kit (Ambion).

For RNA-seq of knockdowns, feeder-free WT26 ESCs were transduced with shRNAs specific for *Msl1*, *Msl2*, *Mof*, *Kansl3* and control shRNA as biological triplicates as described above. Briefly, following transduction for 24 hr, cells were washed with PBS thrice to remove the viral supernatant and subjected to puromycin selection (1.5 μg/ml) for 24 hr. In the case of *Msl1/2*, *Mof*, control shRNA the cells were maintained in puromycin selection for 4 days and in case of *Kansl3,* the cells were maintained in puromycin-selection for 84 hr. An additional set of control shRNA was performed alongside with *Kansl3* for 84 hr. Total RNA from all the shRNA-treated cells was extracted using TRIzol Reagent and the samples were treated with DNase using the TURBO DNase kit (Ambion). The quality of the RNA was analyzed using the Bioanalyzer and samples with RIN values between 9 and 10 were used for RNA-seq. For RNA-seq analysis, cDNA libraries were prepared using the Illumina TruSeq Stranded mRNA kit with 3 μg DNase-treated samples.

RNA-FISH

Xist and *Huwe1* probes were described previously (**Chow et al., 2010**). *Tsix* was detected with a *DXPas34* plasmid (**Debrand et al., 1999**). Approximately 1 × 10^5 of F1-21.6 ESCs were plated on gelatin-coated coverslips and incubated for 24 to 48 hr. After fixation and permeabilization, coverslips with cells were washed and stored in 70% EtOH at −20°C. Then the coverslips were dehydrated in 80%, 95%, and 100% EtOH (5 min each) and briefly air-dried. FISH probes were labeled by nick translation (Abbott) with Spectrum Red-dUTP or Spectrum Green-dUTP following the manufacturer's instructions. Labeled probes were precipitated in the presence of salmon sperm (10 μg) and Cot-1 DNA (3 μg), denatured and competed with Cot-1 DNA for 45 min at 37°C. Cells were then directly hybridized with labeled probes at 37°C overnight. Next, coverslips were washed three times in 50% formamide/2 × SSC followed by three washes in 2 × SSC at 42°C. Cells were stained with DAPI (0.2 mg/ml).

Immunofluorescence staining (against NESTIN)

Approximately 1 × 10^5 of male W26 ESCs and NPCs were plated on gelatin-coated coverslips and incubated for 24 hr. The cells were washed twice with PBS and fixed with pre-warmed 4% formaldehyde for 8 min at 37°C. Next, cells were washed thrice with PBS, 5 min each at room temperature and incubated in Permeabilization buffer (1 × PBS, 0.2% Triton X-100) for 5 min at room temperature. After permeabilization cells were incubated in Blocking buffer (1 × PBS, 5% BSA, 0.05% Triton-X100) for 30 min, stained for 1 hr with primary antibody (rabbit polyclonal a-NESTIN, 1:500). Next, cells were washed thrice with Wash buffer (1 × PBS, 0.05% Triton-X100) and incubated in 10% goat normal serum solution (Invitrogen) for 20 min. Secondary antibody (goat anti-rabbit Alexa Fluor-488, 1:1000) was added on coverslips and incubated for 45 min.

Microscopy

We used a spinning disk confocal microscope (Observer 1/Zeiss) with Plan Apochromat 63x1.4-oil objective for magnification. 500 ms exposure time was used for all lasers. Sequential z-axis images were collected in 0.5 μm steps. ZEN Blue software was used for image analysis.

Sequencing

All samples were sequenced by the Deep Sequencing Unit (MPI-IE, Freiburg) using Illumina HiSeq2000. Library preparation was carried out following Illumina standard protocols for paired-end sequencing (50 bp reads). All raw reads can be found in the GEO database under the accession number GSE51746.

RNA-seq data processing

RNA-seq reads were mapped to Ensembl annotation NCBIM37/mm9 using TopHat2 (**Kim et al., 2013**) with the options mate-inner-dist, mate-std-dev and library-type (fr-firststrand). The distance between read mates (mate-inner-dist and mate-std-dev) were assessed individually for each sequenced library based on the output of the sequencer for average fragment size and CV value.

For FPKM value generation, cufflinks (version 2.1.1) was used for each transcript in each condition (three replicates for ESC and NPC) with default parameters; CummeRbund was used for quality checks and data access (**Trapnell et al., 2013**). Based on the distribution of FPKM values, active genes were defined as transcripts with mean FPKM ≥4 (average over the replicates).

Differential gene expression analysis

After mapping of the RNA-seq reads from the shRNA-treated samples (including scrambled control), the reads that mapped to the genome were counted using htseq-count (doi: 10.1101/002824) with the stranded option set to reverse. The annotations present in the *Mus musculus* gtf file from the ENSEMBL release 67 were used as reference for counting.

DESeq2 was used for differential expression analysis (**Anders and Huber, 2010**). In this analysis, all libraries from knockdown cells were compared in a pairwise manner with its corresponding scrambled shRNA samples. Within the DESeq2 workflow, the cooks–cutoff parameter was set to 'FALSE' and the genes with an adjusted p-value ≤0.01 were defined as significantly affected.

ChIP-seq analysis

Read mapping and normalizations

After mapping of the paired-end reads to the mouse genome (mm9) using bowtie version 2 (**Langmead and Salzberg, 2012**), we filtered for duplicate reads, reads with mapping qualities smaller than 2 and

ambiguously mapped reads using samtools (***Li et al., 2009***). We also removed reads mapping to the mitochondrial genome and 'random' chromosomes, as well as known major satellites and duplicated genome regions to avoid coverage biases.

For normalization procedures, several modules of the deepTools suite (https://deeptools.github.io) were used (***Ramirez et al., 2014***). To ensure a fair comparison between all data sets, first, the GC bias of all mapped reads was determined and, if necessary, corrected so that input and ChIP samples had similar GC distributions of their reads (correctGCbias module). In addition, all aligned read files were corrected for sequencing depth using the signal extraction method proposed by ***Diaz et al. (2012)*** and normalized to the cell-type-specific input (bamCompare module).

Peak calling and replicate handling

MACS (version 1.4) was used for peak calling on every sample individually and on the merged files of two replicates (***Zhang et al., 2008***). Only peaks present in both replicates were considered, using the borders and summits defined by peak calling results for the merged replicates. In addition, peaks with $-10\log_{10}$(p-values) lower than 50 and false-discovery rate values greater than 0.1% were excluded from down-stream analyses.

Annotation used for genome-wide analyses

We used the RefSeq gene list for genome version mm9/NCBI37. Unless specified otherwise, alternative transcription start sites were scored as individual TSS in the respective analyses. The list of genes with homologues in different species was downloaded from HomoloGene and subsequently filtered for pairs of mouse and fly genes that belong to the same clusters of homology ID. CpG island information was downloaded from the UCSC Genome Browser (***Wu et al., 2010***), mean observed over expected CpG ratios were extracted for the TSSs ± 0.5 kb using UCSC tools.

Clustering

For ***Figure 2***, a matrix containing the normalized ChIP-seq signals for all peaks was generated as follows: first, the union of peaks was created using mergeBed from the bedtools suite (***Quinlan and Hall, 2010***); then each region was binned to 2 kb and the normalized ChIP values were extracted in 50 bp windows. The ChIP signal values were rank-transformed, converted into euclidean distances using the R function 'dist' and subsequently ordered according to their similarity by the 'hclust' function (using Ward's method). The resulting dendrogram was pruned to 2 to 10 clusters for which the individual ChIP signals for unscaled regions were extracted (***Figure 2***). Visual inspection revealed no striking differences of the binding patterns between the individual clusters for more than 5 clusters.

The 3 clusters displayed in the lower part of ***Figure 4—figure supplement 2B*** were obtained similarly: first, a matrix was generated that contained the normalized ChIP-seq values of MOF, p300, H3K4me1 and DNase hypersensitivity sites for all regions of cluster D that did not overlap with ESC enhancers. The regions were then scaled to 1400 bp and mean values were computed for 50 bp bins using the computeMatrix module of deepTools (***Ramirez et al., 2014***). Further processing was done as described above; the resulting dendrogram was pruned to $k = 3$ and the enrichments of the different factors were computed and visualized for 10 kb regions using the heatmapper module of deepTools.

GO term analysis

For GO term analyses, we used two approaches: the web interface of DAVID (***Huang da et al., 2009***) and GREAT (***McLean et al., 2010***).

For DAVID, we determined genes overlapping with the peaks of the individual ChIP-seq samples (TSS region ± 500 bp) and supplied the corresponding RefSeq-IDs. The background list contained the union of all TSSs bound by at least one ChIPed protein. We used the Functional Annotation Clustering tool, filtered with the option 'high stringency' and manually grouped the returned clusters of gene functions with enrichment scores above 1.3 into even broader terms.

To assess the GO terms of genes that might be regulated by the TSS-distal binding sites of MOF, MSL1, MSL2, KANSL3 and MCRS1, we used GREAT (***McLean et al., 2010***) with the mouse genome as the background data set and default settings. We obtained the top-ranked biological processes of the genes suggested to be *cis*-regulated by the regions combined in cluster D (***Figure 2***).

Analysis of transcription factor binding sites

For the analysis of enriched transcription factor binding sites, we used the R package ChIPEnrich (http://sartorlab.ccmb.med.umich.edu/chip-enrich) and TRAP (***Thomas-Chollier et al., 2011***). The ChIPEnrich package takes peak regions as input and uses a logistic regression approach to test for gene set enrichments while normalizing for mappability and locus length. We supplied the regions belonging to the individual clusters of binding (A–E from ***Figure 2***) and obtained the corresponding enriched transcription factors.

To plot the occurrences of the SMAD3 motif (V$SMAD3_Q6, TRANSFAC name M00701; ***Figure 4—figure supplement 3C***), TRAP was used with the following command to generate a bedgraph file where the log likelihood of a SMAD3 motif occurrence is stored for the entire genome: ANNOTATEv3.04_source/Release/ANNOTATE_v3.04 -s mm9.fa --pssm /transfac.pssm -g 0.5 --ttype balanced -name M00701 -d | awk 'BEGIN{OFS=t}{print $1, $4+7, $4+8, $6}' > SMAD3.pssm.bedgraph.

Heatmap visualizations and summary plots

Heatmaps displaying normalized read densities of ChIP-seq samples, % methylated CpGs and SMAD3 motif score (***Figures 2, 3C, 4A***, ***Figure 4—figure supplement 2B and 3***) were generated with the computeMatrix and heatmapper modules of the deepTools package (***Ramirez et al., 2014***) with 'reference-point' mode. Heatmaps of fractions of overlapped regions as in ***Figure 3—figure supplement 1*** and ***Figure 4B*** as well as $\log_2$ fold changes (knockdown/control) from RNA-seq experiments (***Figure 3E***) were generated with the function 'heatmap.2' from the R gplots package.

The values underlying the summary plots such as the meta-gene and meta-enhancer plots in ***Figures 3D and 4B***, ***Figure 3—figure supplement 2B***, ***Figure 4—figure supplement 1B, 2A,C*** were generated with the computeMatrix module of the deepTools package using either 'reference-point' or 'scale-regions' mode and were visualized with the R package ggplots2.

Working with genomic intervals

For general assessments of overlaps between bed-files and to extract scores for defined regions the bedTools suite (***Quinlan and Hall, 2010***) and UCSC tools (***Kuhn et al., 2013***) were used. The snapshots of the binding profiles were obtained with IGV browser (***Thorvaldsdottir et al., 2013***).

Target definitions

For each knockdown condition for which RNA-seq data had been generated (see above), significantly affected genes were used (adjusted p-value ≤0.01, see above for differential gene expression analysis). Then they were subdivided into TSS- (ChIP-seq peak overlap with TSS ±1 kb), TSS-distal- (ChIP-seq peaks not overlapping with TSS ±1 kb) and non-targets (neither TSS overlap nor part of TSS-distal list). A gene was classified as TSS-distally regulated when at least one of the following criteria was true:

1. TSS-distal peaks overlapped its published super or typical enhancer (***Whyte et al., 2013***)
2. TSS-distal peaks were predicted by GREAT (***McLean et al., 2010***) to regulate the respective gene
3. TSS-distal peaks overlapped with at least one intron

Genes were defined as MSL targets when peaks of MOF and MSL1|MSL2 were overlapping at the TSS ±1 kb or TSS-distal peaks were predicted to regulate the same putative target gene. NSL targets were defined the same way, but with co-occurrences of peaks from MOF and KANSL3|MCRS1.

Acknowledgements

We would like to thank T Jenuwein from the MPI-IE, Freiburg for providing the WT26 ES cell line and U. Riehle from the Deep Sequencing Unit of the MPI-IE, Freiburg for sequencing all samples. We would also like to thank P Kindle (MPI-IE, Freiburg, Imaging facility), S Toscano, and M. Shvedunova for help with imaging. TC is especially grateful to A Chatterjee for providing support and insightful discussions. We thank all members of the labs for helpful discussions. This work was supported by DFG-SFB992 and BIOSSII awarded to AA and RB; DFG-SFB746 awarded to AA. AA is also a member of the EU-NoE-EpiGeneSys.

Additional information

Competing interests

AA: Reviewing editor, *eLife*. The other authors declare that no competing interests exist.

Funding

Funder	Grant reference number	Author
SFB DFG Germany	SFB992	Rolf Backofen, Asifa Akhtar
SFB DFG Germany	SFB746	Asifa Akhtar
DFG Germany	BIOSS II	Rolf Backofen, Asifa Akhtar
EU NoE	EpiGeneSys	Asifa Akhtar

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

TC, FD, Conception and design, Acquisition of data, Analysis and interpretation of data, Drafting or revising the article; MT, TK, TA, FR, Acquisition of data, Analysis and interpretation of data; A-VG, Acquisition of data, Contributed unpublished essential data or reagents; PRW, PV, RB, TM, Analysis and interpretation of data, Drafting or revising the article; EH, Drafting or revising the article, Contributed unpublished essential data or reagents; AA, Conception and design, Analysis and interpretation of data, Drafting or revising the article, Contributed unpublished essential data or reagents

Additional files

Supplementary files

- Supplementary file 1. MOF/MSL/NSL ChIP-seq statistics.

DOI: 10.7554/eLife.02024.026

- Supplementary file 2. Detailed information about publicly available genome-wide resources used in this study.

DOI: 10.7554/eLife.02024.027

- Supplementary file 3. Lists of primer pairs used in ChIP-qPCR and RT-PCR analyses.

DOI: 10.7554/eLife.02024.028

Major dataset

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset ID and/or URL	Database, license, and accessibility information
Duendar F	2013	MOF-associated complexes ensure stem cell identity and Xist repression	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE57701	Publicly available at NCBI GEO.

References

Akhtar A, Becker PB. 2000. Activation of transcription through histone H4 acetylation by MOF, an acetyltransferase essential for dosage compensation in Drosophila. *Molecular Cell* **5**:367–375. doi: 10.1016/S1097-2765(00)80431-1.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biology* **11**:R106. doi: 10.1186/gb-2010-11-10-r106.

Ang YS, Tsai SY, Lee DF, Monk J, Su J, Ratnakumar K, Ding J, Ge Y, Darr H, Chang B, Wang J, Rendl M, Bernstein E, Schaniel C, Lemischka IR. 2011. Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network. *Cell* **145**:183–197. doi: 10.1016/j.cell.2011.03.003.

Anguera MC, Ma W, Clift D, Namekawa S, Kelleher RJ III, Lee JT. 2011. Tsx produces a long noncoding RNA and has general functions in the germline, stem cells, and brain. *PLOS Genetics* **7**:e1002248. doi: 10.1371/journal.pgen.1002248.

Cai Y, Jin J, Swanson SK, Cole MD, Choi SH, Florens L, Washburn MP, Conaway JW, Conaway RC. 2010. Subunit composition and substrate specificity of a MOF-containing histone acetyltransferase distinct from the male-specific lethal (MSL) complex. *The Journal of Biological Chemistry* **285**:4268–4272. doi: 10.1074/jbc.C109.087981.

Chow JC, Ciaudo C, Fazzari MJ, Mise N, Servant N, Glass JL, Attreed M, Avner P, Wutz A, Barillot E, Greally JM, Voinnet O, Heard E. 2010. LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. *Cell* **141**:956–969. doi: 10.1016/j.cell.2010.04.042.

Chureau C, Chantalat S, Romito A, Galvani A, Duret L, Avner P, Rougeulle C. 2011. Ftx is a non-coding RNA which affects Xist expression and chromatin structure within the X-inactivation center region. *Human Molecular Genetics* **20**:705–718. doi: 10.1093/hmg/ddq516.

Clerc P, Avner P. 1998. Role of the region 3' to Xist exon 6 in the counting process of X-chromosome inactivation. *Nature Genetics* **19**:249–253. doi: 10.1038/924.

Cohen DE, Davidow LS, Erwin JA, Xu N, Warshawsky D, Lee JT. 2007. The DXPas34 repeat regulates random and imprinted X inactivation. *Developmental Cell* **12**:57–71. doi: 10.1016/j.devcel.2006.11.014.

Conrad T, Akhtar A. 2011. Dosage compensation in Drosophila melanogaster: epigenetic fine-tuning of chromosome-wide transcription. *Nature Reviews Genetics* **13**:123–134. doi: 10.1038/nrg3124.

Conti L, Pollard SM, Gorba T, Reitano E, Toselli M, Biella G, Sun Y, Sanzone S, Ying QL, Cattaneo E, Smith A. 2005. Niche-independent symmetrical self-renewal of a mammalian tissue stem cell. *PLOS Biology* **3**:e283. doi: 10.1371/journal.pbio.0030283.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA, Young RA, Jaenisch R. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America* **107**:21931–21936. doi: 10.1073/pnas.1016071107.

Cui G, Park S, Badeaux AI, Kim D, Lee J, Thompson JR, Yan F, Kaneko S, Yuan Z, Botuyan MV, Bedford MT, Cheng JQ, Mer G. 2012. PHF20 is an effector protein of p53 double lysine methylation that stabilizes and activates p53. *Nature Structural & Molecular Biology* **19**:916–924. doi: 10.1038/nsmb.2353.

Debrand E, Chureau C, Arnaud D, Avner P, Heard E. 1999. Functional analysis of the DXPas34 locus, a 3' regulator of Xist expression. *Molecular and Cellular Biology* **19**:8513–8525.

Deuve JL, Avner P. 2011. The coupling of X-chromosome inactivation to pluripotency. *Annual Review of Cell and Developmental Biology* **27**:611–629. doi: 10.1146/annurev-cellbio-092910-154020.

Diaz A, Park K, Lim DA, Song JS. 2012. Normalization, bias correction, and peak calling for ChIP-seq. *Statistical Applications in Genetics and Molecular Biology* **11**. Article 9. doi: 10.1515/1544-6115.1750.

Donohoe ME, Silva SS, Pinter SF, Xu N, Lee JT. 2009. The pluripotency factor Oct4 interacts with Ctcf and also controls X-chromosome pairing and counting. *Nature* **460**:128–132. doi: 10.1038/nature08098.

Donohoe ME, Zhang LF, Xu N, Shi Y, Lee JT. 2007. Identification of a Ctcf cofactor, Yy1, for the X chromosome binary switch. *Molecular Cell* **25**:43–56. doi: 10.1016/j.molcel.2006.11.017.

Farre D, Bellora N, Mularoni L, Messeguer X, Alba MM. 2007. Housekeeping genes tend to show reduced upstream sequence conservation. *Genome Biology* **8**:R140. doi: 10.1186/gb-2007-8-7-r140.

Feller C, Prestel M, Hartmann H, Straub T, Soding J, Becker PB. 2012. The MOF-containing NSL complex associates globally with housekeeping genes, but activates only a defined subset. *Nucleic Acids Research* **40**:1509–1522. doi: 10.1093/nar/gkr869.

Gendrel AV, Attia M, Chen CJ, Diabangouaya P, Servant N, Barillot E, Heard E. 2014. Developmental dynamics and disease potential of random monoallelic gene expression. *Developmental Cell* **28**:366–380. doi: 10.1016/j.devcel.2014.01.016.

Gontan C, Achame EM, Demmers J, Barakat TS, Rentmeester E, Van IW, Grootegoed JA, Gribnau J. 2012. RNF12 initiates X-chromosome inactivation by targeting REX1 for degradation. *Nature* **485**:386–390. doi: 10.1038/nature11070.

Gupta A, Guerin-Peyrou TG, Sharma GG, Park C, Agarwal M, Ganju RK, Pandita S, Choi K, Sukumar S, Pandita RK, Ludwig T, Pandita TK. 2008. The mammalian ortholog of Drosophila MOF that acetylates histone H4 lysine 16 is essential for embryogenesis and oncogenesis. *Molecular and Cellular Biology* **28**:397–409. doi: 10.1128/MCB.01045-07.

Gupta A, Hunt CR, Pandita RK, Pae J, Komal K, Singh M, Shay JW, Kumar R, Ariizumi K, Horikoshi N, Hittelman WN, Guha C, Ludwig T, Pandita TK. 2013. T-cell-specific deletion of Mof blocks their differentiation and results in genomic instability in mice. *Mutagenesis* **28**:263–270. doi: 10.1093/mutage/ges080.

Hallacli E, Lipp M, Georgiev P, Spielman C, Cusack S, Akhtar A, Kadlec J. 2012. Msl1-mediated dimerization of the dosage compensation complex is essential for male X-chromosome regulation in Drosophila. *Molecular Cell* **48**:587–600. doi: 10.1016/j.molcel.2012.09.014.

Hu G, Kim J, Xu Q, Leng Y, Orkin SH, Elledge SJ. 2009. A genome-wide RNAi screen identifies a new transcriptional module required for self-renewal. *Genes & Development* **23**:837–848. doi: 10.1101/gad.1769609.

Huang da W, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**:44–57. doi: 10.1038/nprot.2008.211.

Jeon Y, Lee JT. 2011. YY1 tethers Xist RNA to the inactive X nucleation center. *Cell* **146**:119–133. doi: 10.1016/j.cell.2011.06.026.

Jonkers I, Barakat TS, Achame EM, Monkhorst K, Kenter A, Rentmeester E, Grosveld F, Grootegoed JA, Gribnau J. 2009. RNF12 is an X-Encoded dose-dependent activator of X chromosome inactivation. *Cell* **139**:999–1011. doi: 10.1016/j.cell.2009.10.034.

Kadlec J, Hallacli E, Lipp M, Holz H, Sanchez-Weatherby J, Cusack S, Akhtar A. 2011. Structural basis for MOF and MSL3 recruitment into the dosage compensation complex by MSL1. *Nature Structural & Molecular Biology* **18**:142–149. doi: 10.1038/nsmb.1960.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14**:R36. doi: 10.1186/gb-2013-14-4-r36.

Koolen DA, Kramer JM, Neveling K, Nillesen WM, Moore-Barton HL, Elmslie FV, Toutain A, Amiel J, Malan V, Tsai AC, Cheung SW, Gilissen C, Verwiel ET, Martens S, Feuth T, Bongers EM, De Vries P, Scheffer H, Vissers LE, De Brouwer AP, Brunner HG, Veltman JA, Schenck A, Yntema HG, De Vries BB. 2012. Mutations in the chromatin modifier gene KANSL1 cause the 17q21.31 microdeletion syndrome. *Nature Genetics* **44**:639–641. doi: 10.1038/ng.2262.

Kruse JP, Gu W. 2009. MSL2 promotes Mdm2-independent cytoplasmic localization of p53. *The Journal of Biological Chemistry* **284**:3250–3263. doi: 10.1074/jbc.M805658200.

Kuhn RM, Haussler D, Kent WJ. 2013. The UCSC genome browser and associated tools. *Briefings in Bioinformatics* **14**:144–161. doi: 10.1093/bib/bbs038.

Kumar R, Hunt CR, Gupta A, Nannepaga S, Pandita RK, Shay JW, Bachoo R, Ludwig T, Burns DK, Pandita TK. 2011. Purkinje cell-specific males absent on the first (mMof) gene deletion results in an ataxia-telangiectasia-like neurological phenotype and backward walking in mice. *Proceedings of the National Academy of Sciences of the United States of America* **108**:3636–3641. doi: 10.1073/pnas.1016524108.

Lam KC, Mühlpfordt F, Vaquerizas JM, Raja SJ, Holz H, Luscombe NM, Manke T, Akhtar A. 2012. The NSL complex regulates housekeeping genes in Drosophila. *PLOS Genetics* **8**:e1002736. doi: 10.1371/journal.pgen.1002736.

Landolin JM, Johnson DS, Trinklein ND, Aldred SF, Medina C, Shulha H, Weng Z, Myers RM. 2010. Sequence features that drive human promoter function and tissue specificity. *Genome Research* **20**:890–898. doi: 10.1101/gr.100370.109.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**:357–359. doi: 10.1038/nmeth.1923.

Lee JT. 2005. Regulation of X-chromosome counting by Tsix and Xite sequences. *Science* **309**:768–771. doi: 10.1126/science.1113673.

Lee JT, Lu N. 1999. Targeted mutagenesis of Tsix leads to nonrandom X inactivation. *Cell* **99**:47–57. doi: 10.1016/S0092-8674(00)80061-6.

Lee KK, Workman JL. 2007. Histone acetyltransferase complexes: one size doesn't fit all. *Nature Reviews Molecular Cell Biology* **8**:284–295. doi: 10.1038/nrm2145.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing Subgroup. 2009. The sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078–2079. doi: 10.1093/bioinformatics/btp352.

Li X, Li L, Pandey R, Byun JS, Gardner K, Qin Z, Dou Y. 2012. The histone acetyltransferase MOF is a key regulator of the embryonic stem cell core transcriptional network. *Cell Stem Cell* **11**:163–178. doi: 10.1016/j.stem.2012.04.023.

Luikenhuis S, Wutz A, Jaenisch R. 2001. Antisense transcription through the Xist locus mediates Tsix function in embryonic stem cells. *Molecular and Cellular Biology* **21**:8512–8520. doi: 10.1128/MCB.21.24.8512-8520.2001.

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology* **28**:495–501. doi: 10.1038/nbt.1630.

Mendjan S, Taipale M, Kind J, Holz H, Gebhardt P, Schelder M, Vermeulen M, Buscaino A, Duncan K, Mueller J, Wilm M, Stunnenberg HG, Saumweber H, Akhtar A. 2006. Nuclear pore components are involved in the transcriptional regulation of dosage compensation in Drosophila. *Molecular Cell* **21**:811–823. doi: 10.1016/j.molcel.2006.02.007.

Morey C, Arnaud D, Avner P, Clerc P. 2001. Tsix-mediated repression of Xist accumulation is not sufficient for normal random X inactivation. *Human Molecular Genetics* **10**:1403–1411. doi: 10.1093/hmg/10.13.1403.

Nagy Z, Riss A, Fujiyama S, Krebs A, Orpinell M, Jansen P, Cohen A, Stunnenberg HG, Kato S, Tora L. 2010. The metazoan ATAC and SAGA coactivator HAT complexes regulate different sets of inducible target genes. *Cellular and Molecular Life Sciences* **67**:611–628. doi: 10.1007/s00018-009-0199-8.

Navarro P, Chambers I, Karwacki-Neisius V, Chureau C, Morey C, Rougeulle C, Avner P. 2008. Molecular coupling of Xist regulation and pluripotency. *Science* **321**:1693–1695. doi: 10.1126/science.1160952.

Navarro P, Oldfield A, Legoupi J, Festuccia N, Dubois A, Attia M, Schoorlemmer J, Rougeulle C, Chambers I, Avner P. 2010. Molecular coupling of Tsix regulation and pluripotency. *Nature* **468**:457–460. doi: 10.1038/nature09496.

Navarro P, Pichard S, Ciaudo C, Avner P, Rougeulle C. 2005. Tsix transcription across the Xist gene alters chromatin conformation without affecting Xist transcription: implications for X-chromosome inactivation. *Genes & Development* **19**:1474–1484. doi: 10.1101/gad.341105.

Nesterova TB, Senner CE, Schneider J, Alcayna-Stevens T, Tattermusch A, Hemberger M, Brockdorff N. 2011. Pluripotency factor binding and Tsix expression act synergistically to repress Xist in undifferentiated embryonic stem cells. *Epigenetics & Chromatin* **4**:17. doi: 10.1186/1756-8935-4-17.

Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, Van Berkum NL, Meisig J, Sedat J, Gribnau J, Barillot E, Bluthgen N, Dekker J, Heard E. 2012. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**:381–385. doi: 10.1038/nature11049.

Ohhata T, Hoki Y, Sasaki H, Sado T. 2006. Tsix-deficient X chromosome does not undergo inactivation in the embryonic lineage in males: implications for Tsix-independent silencing of Xist. *Cytogenetic and Genome Research* **113**:345–349. doi: 10.1159/000090851.

Pauli F. 2010. Myers Lab ChIP-seq Protocol, v041610.1 and v041610.2. In: Myers R. editor. http://www.hudsonalpha.org/myers-lab

Pollex T, Heard E. 2012. Recent advances in X-chromosome inactivation research. *Current Opinion in Cell Biology* **24**:825–832. doi: 10.1016/j.ceb.2012.10.007.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**:841–842. doi: 10.1093/bioinformatics/btq033.

Raja SJ, Charapitsa I, Conrad T, Vaquerizas JM, Gebhardt P, Holz H, Kadlec J, Fraterman S, Luscombe NM, Akhtar A. 2010. The nonspecific lethal complex is a transcriptional regulator in Drosophila. *Molecular Cell* **38**:827–841. doi: 10.1016/j.molcel.2010.05.021.

Ramirez F, Dundar F, Diehl S, Gruning BA, Manke T. 2014. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Research*. Epub ahead of print.

Rougeulle C, Avner P. 2004. The role of antisense transcription in the regulation of X-inactivation. *Current Topics in Developmental Biology* **63**:61–89. doi: 10.1016/S0070-2153(04)63003-1.

Sapountzi V, Cote J. 2011. MYST-family histone acetyltransferases: beyond chromatin. *Cellular and Molecular Life Sciences* **68**:1147–1156. doi: 10.1007/s00018-010-0599-9.

Shin J, Bossenz M, Chung Y, Ma H, Byron M, Taniguchi-Ishigaki N, Zhu X, Jiao B, Hall LL, Green MR, Jones SN, Hermans-Borgmeyer I, Lawrence JB, Bach I. 2010. Maternal Rnf12/RLIM is required for imprinted X-chromosome inactivation in mice. *Nature* **467**:977–981. doi: 10.1038/nature09457.

Smith ER, Cayrou C, Huang R, Lane WS, Cote J, Lucchesi JC. 2005. A human protein complex homologous to the Drosophila MSL complex is responsible for the majority of histone H4 acetylation at lysine 16. *Molecular and Cellular Biology* **25**:9175–9188. doi: 10.1128/MCB.25.21.9175-9188.2005.

Splinter E, de Wit E, Nora EP, Klous P, Van De Werken HJ, Zhu Y, Kaaij LJ, Van Ijcken W, Gribnau J, Heard E, de Laat W. 2011. The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes & Development* **25**:1371–1383. doi: 10.1101/gad.633311.

Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, Van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D, Tiwari VK, Schubeler D. 2011. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**:490–495. doi: 10.1038/nature10716.

Stavropoulos N, Lu N, Lee JT. 2001. A functional role for Tsix transcription in blocking Xist RNA accumulation but not in X-chromosome choice. *Proceedings of the National Academy of Sciences of the United States of America* **98**:10232–10237. doi: 10.1073/pnas.171243598.

Sun BK, Deaton AM, Lee JT. 2006. A transient heterochromatic state in Xist preempts X inactivation choice without RNA stabilization. *Molecular Cell* **21**:617–628. doi: 10.1016/j.molcel.2006.01.028.

Sun S, Del Rosario BC, Szanto A, Ogawa Y, Jeon Y, Lee JT. 2013. Jpx RNA activates Xist by evicting CTCF. *Cell* **153**:1537–1551. doi: 10.1016/j.cell.2013.05.028.

Taipale M, Rea S, Richter K, Vilar A, Lichter P, Imhof A, Akhtar A. 2005. hMOF histone acetyltransferase is required for histone H4 lysine 16 acetylation in mammalian cells. *Molecular and Cellular Biology* **25**:6798–6810. doi: 10.1128/MCB.25.15.6798-6810.2005.

Taylor G, Eskeland R, Hekimoglu-Balkan B, Pradeepa M, Bickmore WA. 2013. H4K16 acetylation marks active genes and enhancers of embryonic stem cells, but does not alter chromatin compaction. *Genome Research* **23**:2053–2065. doi: 10.1101/gr.155028.113.

Thomas T, Dixon MP, Kueh AJ, Voss AK. 2008. Mof (MYST1 or KAT8) is essential for progression of embryonic development past the blastocyst stage and required for normal chromatin architecture. *Molecular and Cellular Biology* **28**:5093–5105. doi: 10.1128/MCB.02202-07.

Thomas-Chollier M, Hufton A, Heinig M, O'keeffe S, Masri NE, Roider HG, Manke T, Vingron M. 2011. Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. *Nature Protocols* **6**:1860–1869. doi: 10.1038/nprot.2011.409.

Thorvaldsdottir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14**:178–192. doi: 10.1093/bib/bbs017.

Tian D, Sun S, Lee JT. 2010. The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. *Cell* **143**:390–403. doi: 10.1016/j.cell.2010.09.049.

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. 2013. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology* **31**:46–53. doi: 10.1038/nbt.2450.

Vigneau S, Augui S, Navarro P, Avner P, Clerc P. 2006. An essential role for the DXPas34 tandem repeat and Tsix transcription in the counting process of X chromosome inactivation. *Proceedings of the National Academy of Sciences of the United States of America* **103**:7390–7395. doi: 10.1073/pnas.0602381103.

Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA. 2013. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**:307–319. doi: 10.1016/j.cell.2013.03.035.

Wu H, Caffo B, Jaffee HA, Irizarry RA, Feinberg AP. 2010. Redefining CpG islands using hidden Markov models. *Biostatistics* **11**:499–514. doi: 10.1093/biostatistics/kxq005.

Wu L, Zee BM, Wang Y, Garcia BA, Dou Y. 2011. The RING finger protein MSL2 in the MOF complex is an E3 ubiquitin ligase for H2B K34 and is involved in crosstalk with H3 K4 and K79 methylation. *Molecular Cell* **43**:132–144. doi: 10.1016/j.molcel.2011.05.015.

Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. 2005. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434**:338–345. doi: 10.1038/nature03441.

Yoshida K, Toki T, Okuno Y, Kanezaki R, Shiraishi Y, Sato-Otsubo A, Sanada M, Park MJ, Terui K, Suzuki H, Kon A, Nagata Y, Sato Y, Wang R, Shiba N, Chiba K, Tanaka H, Hama A, Muramatsu H, Hasegawa D, Nakamura K, Kanegane H, Tsukamoto K, Adachi S, Kawakami K, Kato K, Nishimura R, Izraeli S, Hayashi Y, Miyano S, Kojima S, Ito E, Ogawa S. 2013. The landscape of somatic mutations in Down syndrome-related myeloid disorders. *Nature Genetics* **45**:1293–1299. doi: 10.1038/ng.2759.

Young RA. 2011. Control of the embryonic stem cell state. *Cell* **144**:940–954. doi: 10.1016/j.cell.2011.01.032.

Zawel L, Dai JL, Buckhaults P, Zhou S, Kinzler KW, Vogelstein B, Kern SE. 1998. Human Smad3 and Smad4 are sequence-specific transcription activators. *Molecular Cell* **1**:611–617. doi: 10.1016/S1097-2765(00)80061-1.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biology* **9**:R137. doi: 10.1186/gb-2008-9-9-r137.

Zhao W, Li Q, Ayers S, Gu Y, Shi Z, Zhu Q, Chen Y, Wang HY, Wang RF. 2013a. Jmjd3 inhibits reprogramming by upregulating expression of INK4a/Arf and targeting PHF20 for ubiquitination. *Cell* **152**:1037–1050. doi: 10.1016/j.cell.2013.02.006.

Zhao X, Su J, Wang F, Liu D, Ding J, Yang Y, Conaway JW, Conaway RC, Cao L, Wu D, Wu M, Cai Y, Jin J. 2013b. Crosstalk between NSL histone acetyltransferase and MLL/SET complexes: NSL complex functions in promoting histone H3K4 di-methylation activity by MLL/SET complexes. *PLOS Genetics* **9**:e1003940. doi: 10.1371/journal.pgen.1003940.

Zollino M, Orteschi D, Murdolo M, Lattante S, Battaglia D, Stefanini C, Mercuri E, Chiurazzi P, Neri G, Marangi G. 2012. Mutations in KANSL1 cause the 17q21.31 microdeletion syndrome phenotype. *Nature Genetics* **44**:636–638. doi: 10.1038/ng.2257.

A.2.1 Supplemental Material

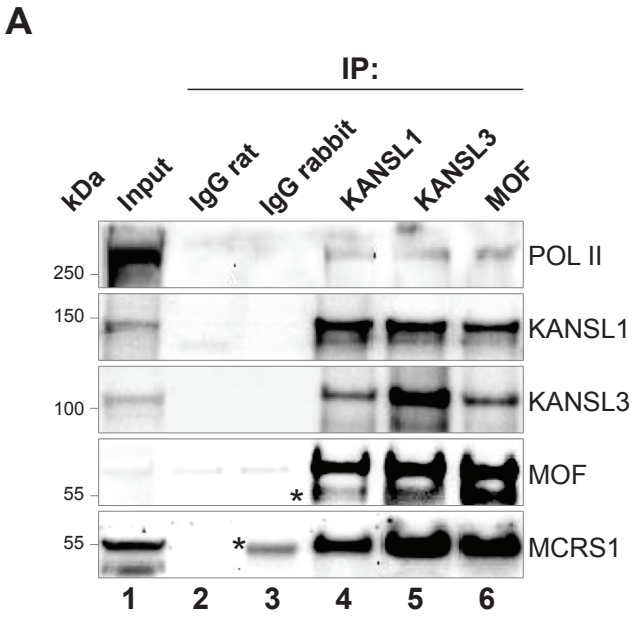
Figure 1—figure supplement 1: Monitoring RNA and protein levels in ESCs and NPCs.

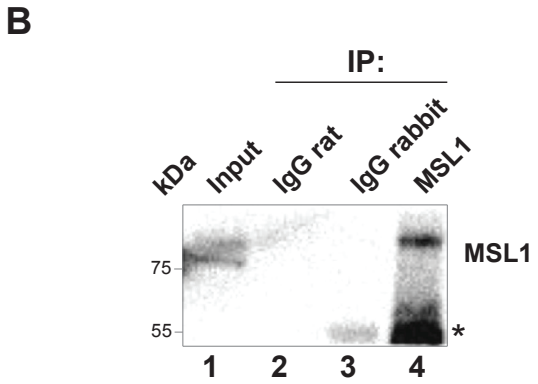
A) We monitored the expression dynamics during ESC differentiation for markers of pluripotency (*Oct4, Nanog, Rex1, Klf4*), embryoid body formation (*Fgf5*), differentiation (*Sox2*), and NPC (*Nestin*). Panels 3 and 4 contain the expression profiles for members of the MSL complex (*Msl1, Msl2*), Mof, and the NSL complex (*Kansl1, Kansl3, Mcrs1*), respectively. All results are represented as relative values individually normalized to Rplp0 expression levels (panel 2) on a given day and to the highest expression level of a given gene during the entire differentiation process (highest expression level of each gene = 1). The x-axes show days of differentiation. All results are expressed as means +/- S.D. for technical replicates. For primers see Supplementary File 3C.

(B) Bright field images illustrate the cell morphology before and after the process of differentiation. The immunofluorescence analysis indicates the specific staining for the NESTIN protein (green) in neuronal progenitors (NPC); DNA is counterstained with DAPI (blue).

(C) Expression changes for selected ESC-specific and NPC-specific markers before and after differentiation of wild-type WT26 cells in using RT-PCR analysis and RNA-seq.

(D) Western blots for proteins from two ES cell lines and their NPC derivatives. Different dilutions were loaded (10%, 30%, 10%) with the order indicated on top of the blots. Anti-GAPDH was used as loading control; arrows indicate the protein of interest.





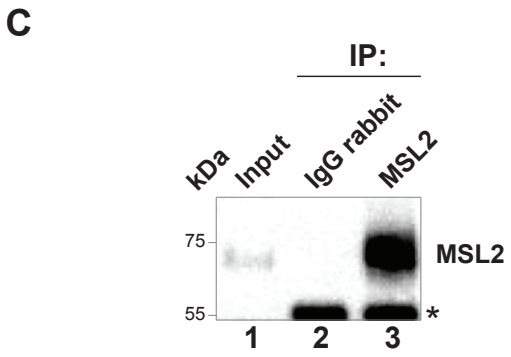


Figure 1—figure supplement 2: Verification of antibodies used in this study.(A) Immunoprecipitations from mouse ESC nuclear extracts with antibodies specific for KANSL1, KANSL3 or MOF and rabbit or rat antisera. The blot was probed with indicated antibodies showing the co-immunoprecipitation of several NSL complex members. Asterisks represent the IgG signal. Pol II = RNA Polymerase II.(B) and (C) same as (A) except that immunoprecipitations were performed with antibodies specific to MSL1 (B) and MSL2 (C). Asterisks represent the IgG signal.

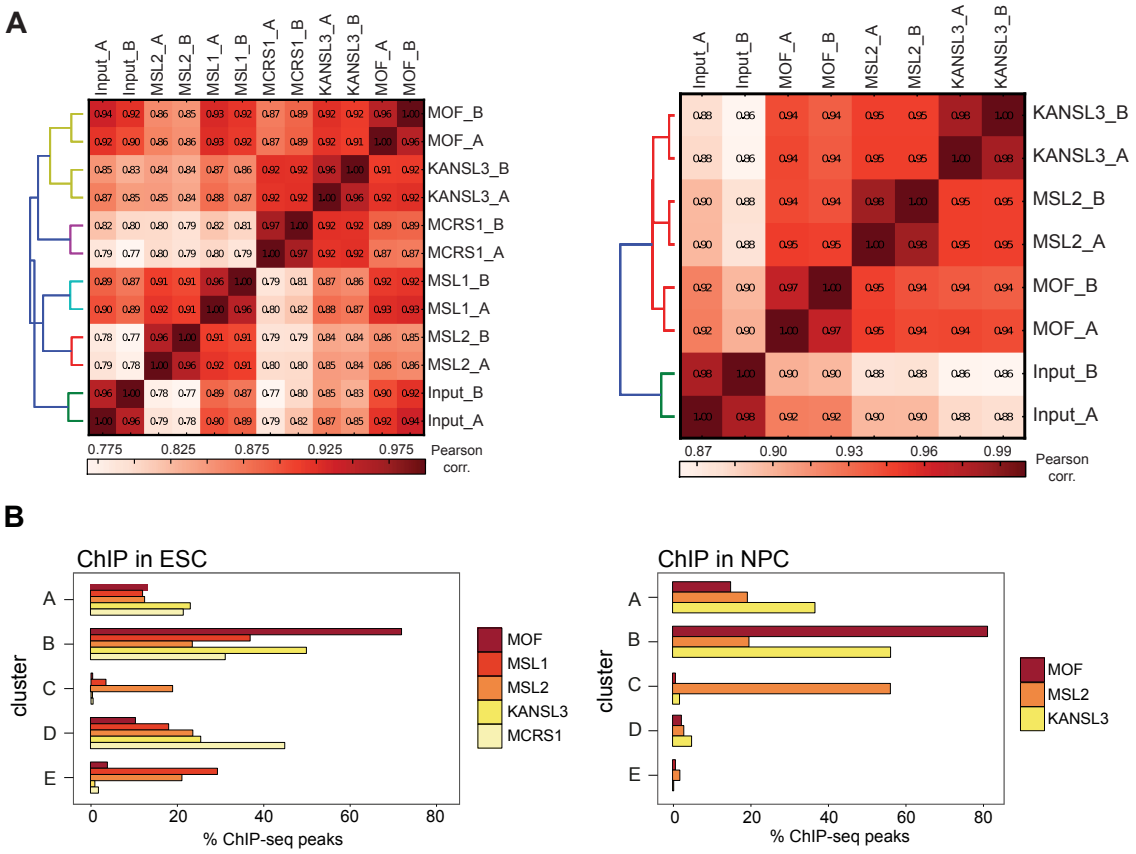


Figure 2—figure supplement: ChIP-seq quality measures.

(A) Correlation plot for all individual ChIP-seq and input samples from ESCs (left) and NPCs. The genome was sampled in windows of 10 kb length; the numbers of reads per bin were counted for each ChIP sample and correlated using Pearson correlation. The calculation and heatmap visualization were done with the bamCorrelate module from the deepTools suite (Ramirez et al., 2014).

(B) The bar chart depicts the fraction of ChIP-seq peaks for each protein that reside within each cluster shown in Figure 2, i.e. approximately 30% of MSL1 peaks in ESCs locate in cluster E. Note that the absolute numbers of peaks differ between the samples (see Supplementary file 1B for absolute peak numbers and Methods and Materials for peak calling details).

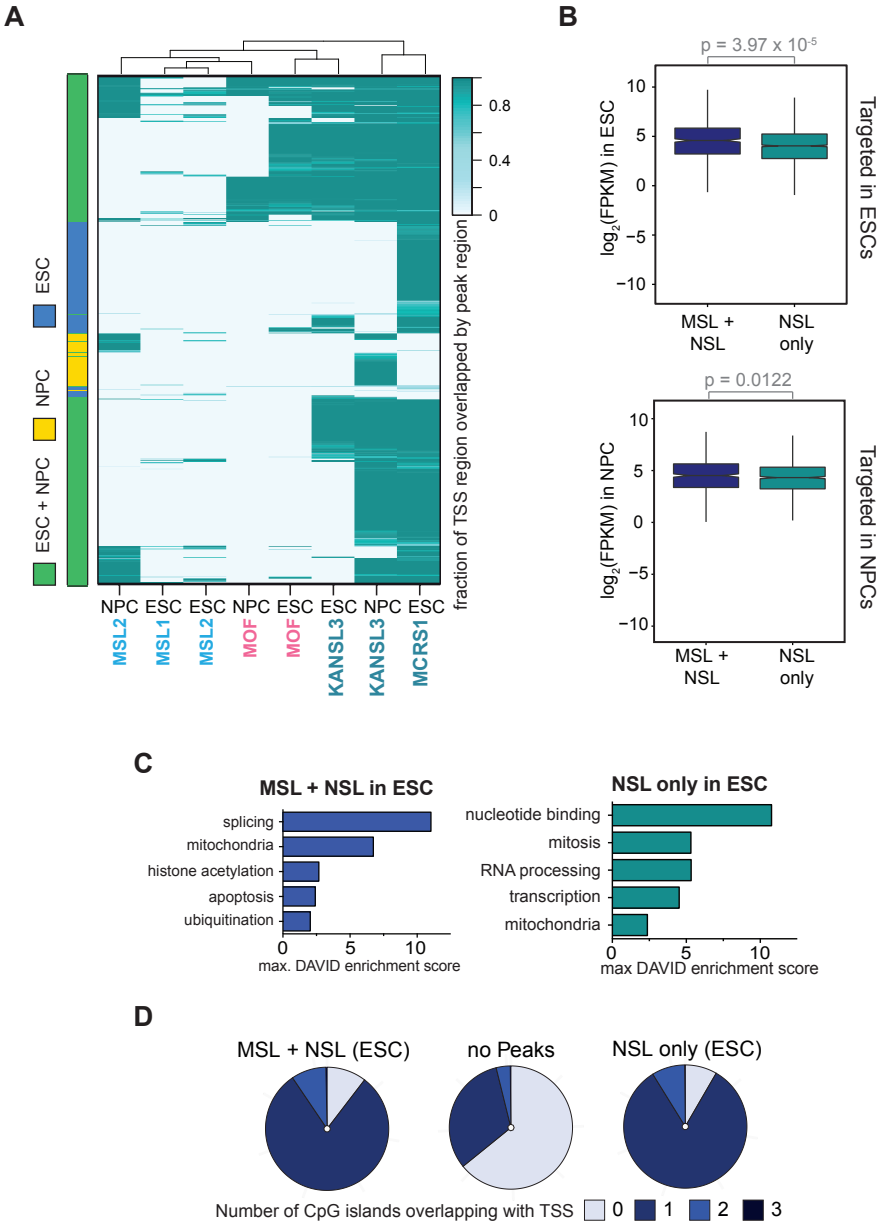


Figure 3—figure supplement 1: MSL and NSL complexes target promoters of broadly expressed genes in ESCs and NPCs.

(A) The heatmap is related to Figure 3B as it is based on all genes that are bound by at least 1 ChIPed factor in ESCs or NPCs. The intensity of the color depicts the fraction of the 1 kb TSS-region that was covered by a binding site of MOF, MSL1, MSL2, KANSL3 or MCRS1. Rows and columns were sorted using hierarchical clustering on the Euclidean distances of the overlap fractions using R. The left color bar indicates which genes are targeted in 1 or both cell types.

(B) Distribution of expression values from RNA-seq data in ESCs and NPCs for genes targeted by MSL and NSL complex members together or by the NSL complex only. P-values were calculated using Welch t-test. (C) Results of the GO term analysis using DAVID (Huang da et al., 2009) on genes that were bound at the TSS in ESCs by NSL complex members only or both MSL and NSL complexes.

(D) The pie charts depict how many times annotated TSSs overlapped with a CpG island. The vast majority of genes that were bound in ESCs by MSL and NSL together or by NSL complex members alone overlapped with at least 1 CpG island (dark and medium blue) while approximately 2/3 of the non-target-TSS did not overlap with any CpG island (light blue for 0 CpG islands within the queried regions).

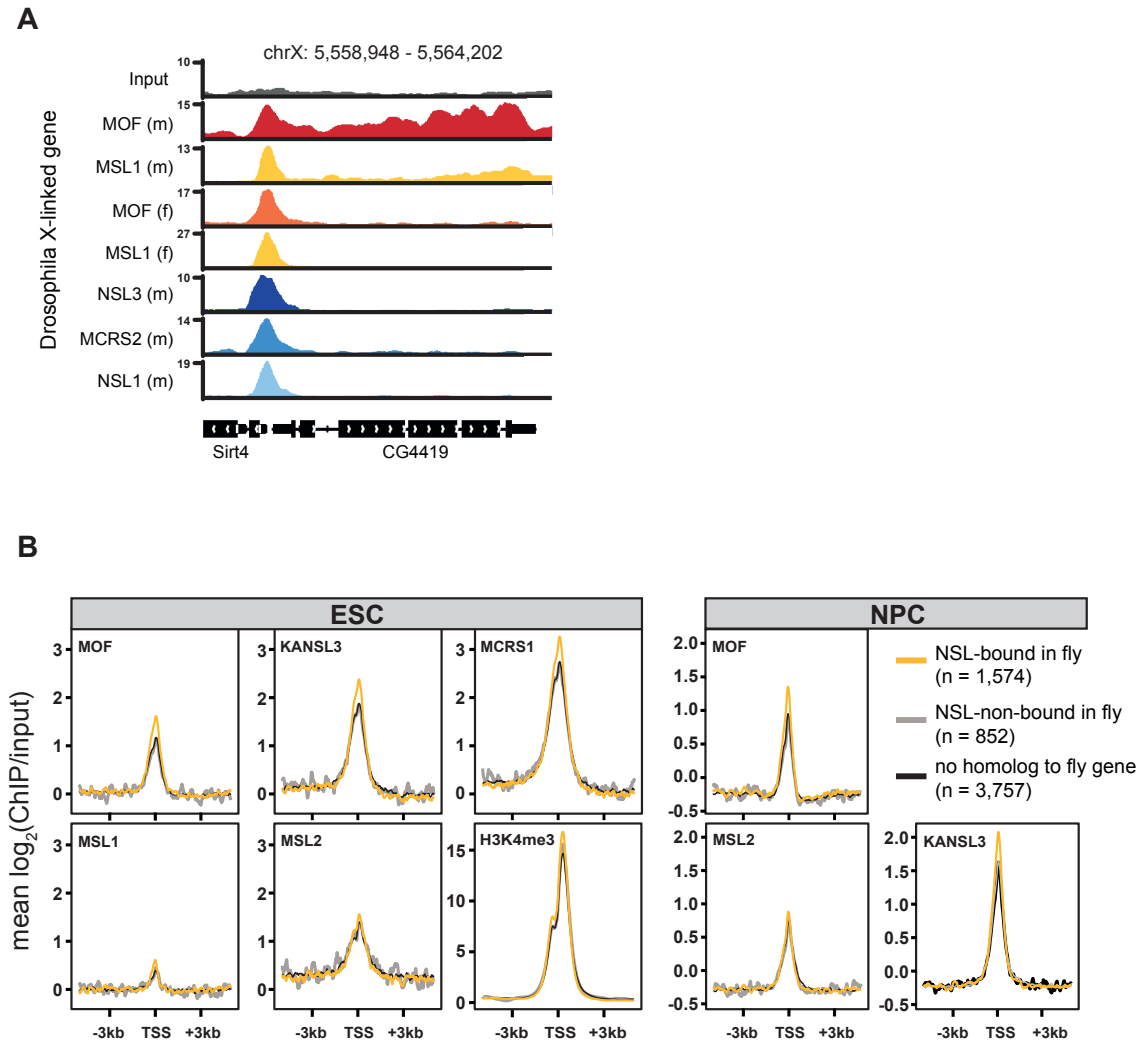


Figure 3—figure supplement 2: The NSL-, but not the MSL-binding mode of D. melanogaster is present in mammalian cells.

(A) Exemplary genome browser snapshots of the X-linked fly gene CG4419. Shown here are the sequencing-depth normalized profiles for ChIP and corresponding input samples, clearly showing a broad enrichment of MOF and MSL1 along the entire gene body in male (m) *D. melanogaster* while all other marks show sharp enrichments around the TSS (including MSL1 and MOF in female (f) *D. melanogaster*) which are similar to those seen for both complexes in mouse cells (Figure 3A and 3D).

(B) Comparison of expressed (FPKM >4) mouse genes whose homologous genes are either bound or not bound by MOF and its complexes in the fly. We extracted the input-normalized ChIP-seq values for 6 kb regions around the TSS using the computeMatrix module of deepTools (Ramirez et al., 2014). H3K4me3 signal is from a published data set, see Supplementary file 2 for the corresponding accession number.

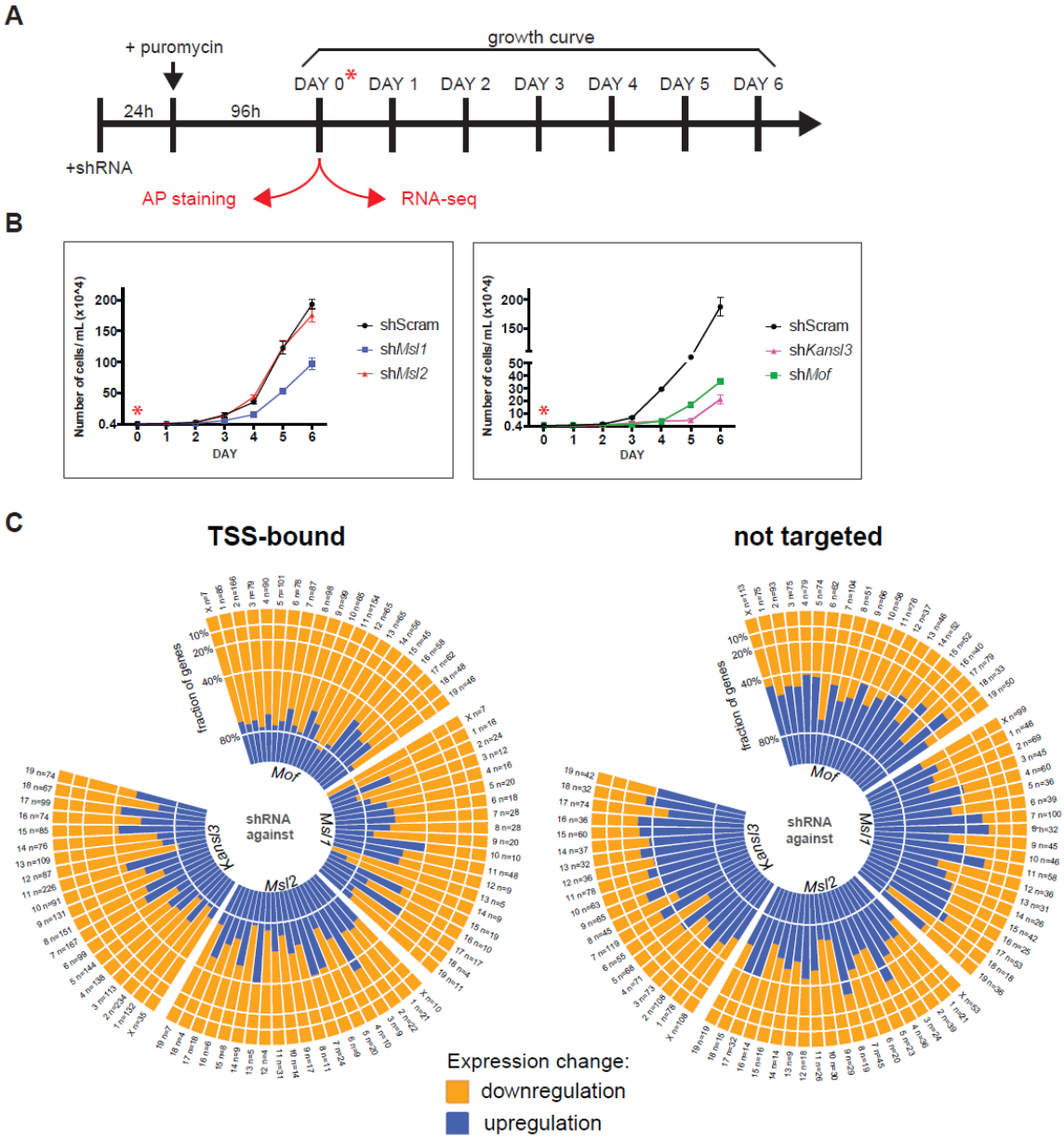


Figure 3—figure supplement 3: Effects of shRNA-mediated depletion of MOF, MSL1, MSL2, and KANSL3.

(A) Time course of knockdown experiments. For experimental details see Methods and Materials. Samples for RNA-sequencing and AP staining (see Figure 4—figure supplement 4) were extracted 4 days after puromycin selection of shRNA-treated cells.

(B) Proliferation assay for shRNA-treated cells, starting at day 4 after puromycin selection (see Figure 3—figure supplement 3A).

(C) Bar plots depicting the fractions of genes (per chromosome) that were significantly up- or downregulated in RNA-seq experiments from shRNA-treated cells. The left plot contains genes which were defined as TSS-targets in the respective ChIP-seq samples, the right plot contains genes that were neither classified as TSS- nor as TSS-distal targets. The labels on each bar indicate the chromosome name and the total number of genes that fulfilled the criteria for this chromosome (significantly affected, TSS-bound or non-targeted). See Methods and Materials for details of the classification.

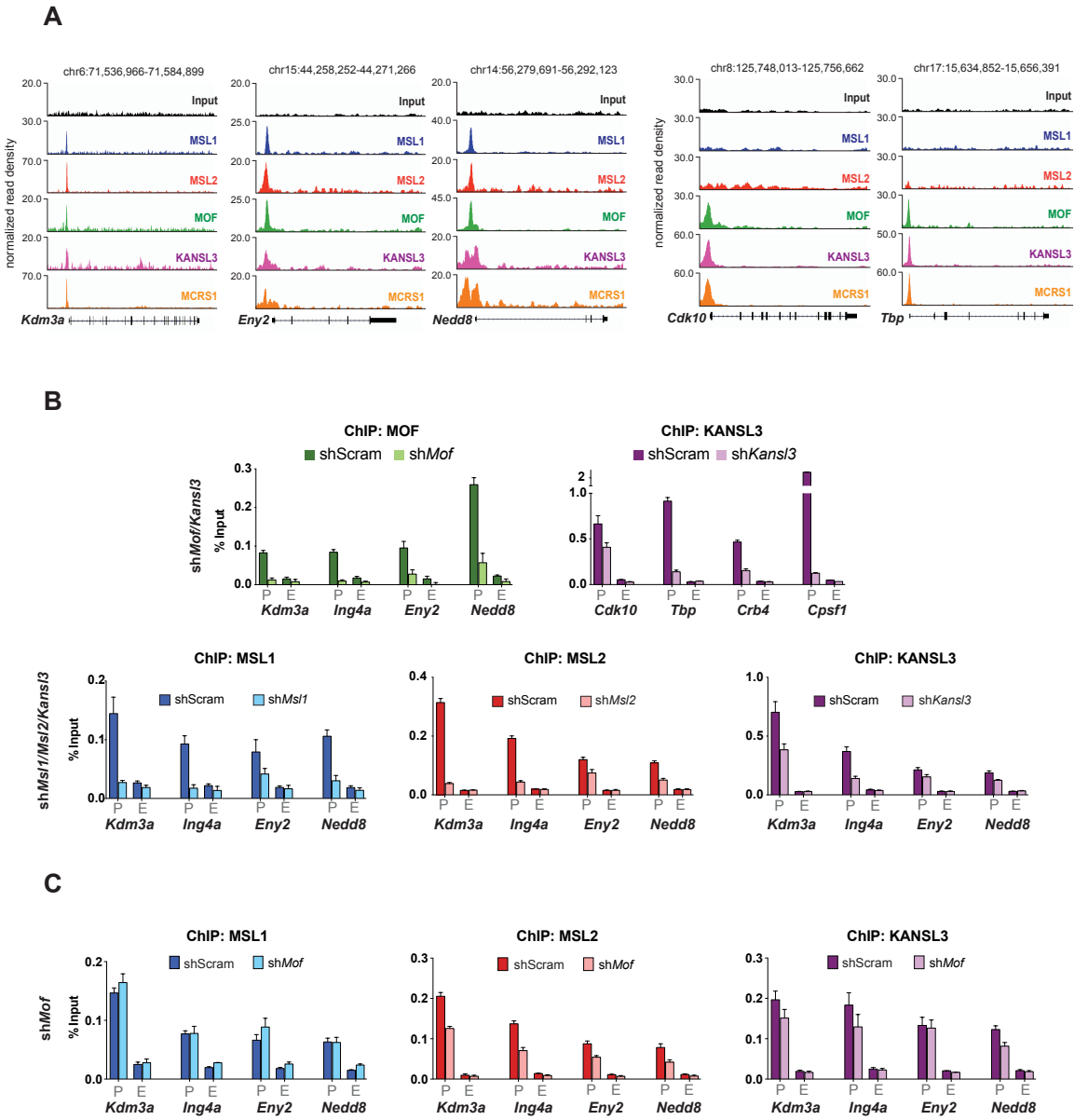


Figure 3—figure supplement 4: Assessment of ChIP signals around the TSSs of putative target genes as determined by ChIP-seq.

(A) Genome Browser snapshots of several MSL/NSL (left) or NSL-only (Visel et al.) target genes and respective sequencing-depth-normalized ChIP-seq and input signals from ESCs. The exact genomic coordinates are indicated on top of each panel. Gene names are indicated on the bottom.

(B) ChIP-qPCR validation for MOF (green) and KANSL3 (purple) signals. Immunoprecipitated DNA was amplified by qPCR with primer sets positioned at the promoter (P) and end (E) of the coding sequence (Supplementary file 3A). Results are expressed as mean +/- S.D. of 3 biological replicates; cells were harvested for experiments on day 4 (*Kansl3*) or 5 (*Mof*) of knockdown.

(C) ChIP-qPCR for MSL1 (blue), MSL2 (red) and KANSL3 (purple) in ESCs treated with sh-RNA (scrambled or against a specific transcript). Signals on genes were evaluated using primers at the promoter (P), and end (E) of the coding sequence. Results are expressed as mean +/- S.D. of 3 biological replicates; cells were harvested for experiments on day 5 of *Mof* knockdown.

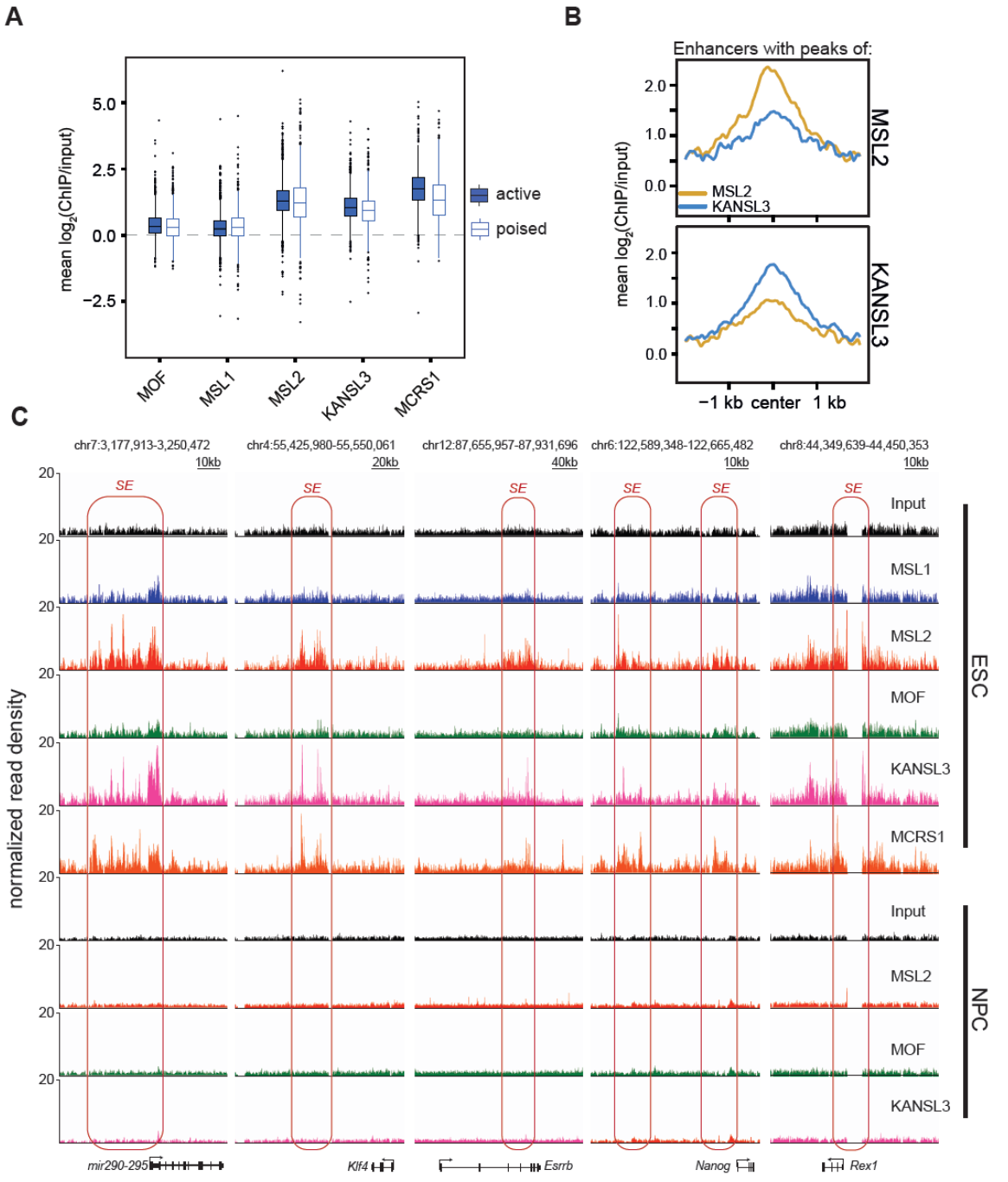


Figure 4—figure supplement 1: MSL2 and KANSL3 show strong enrichments at typical and super enhancers in ESCs.

(A) Boxplots demonstrating the distribution of mean ChIP enrichments for enhancer regions defined by H3K4me1 and H3K27ac marks in ESCs (see Creyghton et al., 2010 for details) that overlap with the clusters of binding defined by our ChIP-seq samples. Mean values were extracting using the UCSCtool bigWigAverageOverBed.

(B) Summary plots for typical enhancer regions (Whyte et al., 2013) that overlapped with either MSL2 (top) or KANSL3 (bottom) peaks. Different colors indicate different ChIP-seq signals. Related to the heatmaps of Figure 4B.

(C) Genome browser snapshots of sequencing-depth normalized ChIP-seq and input profiles for super enhancers of key pluripotency factors.

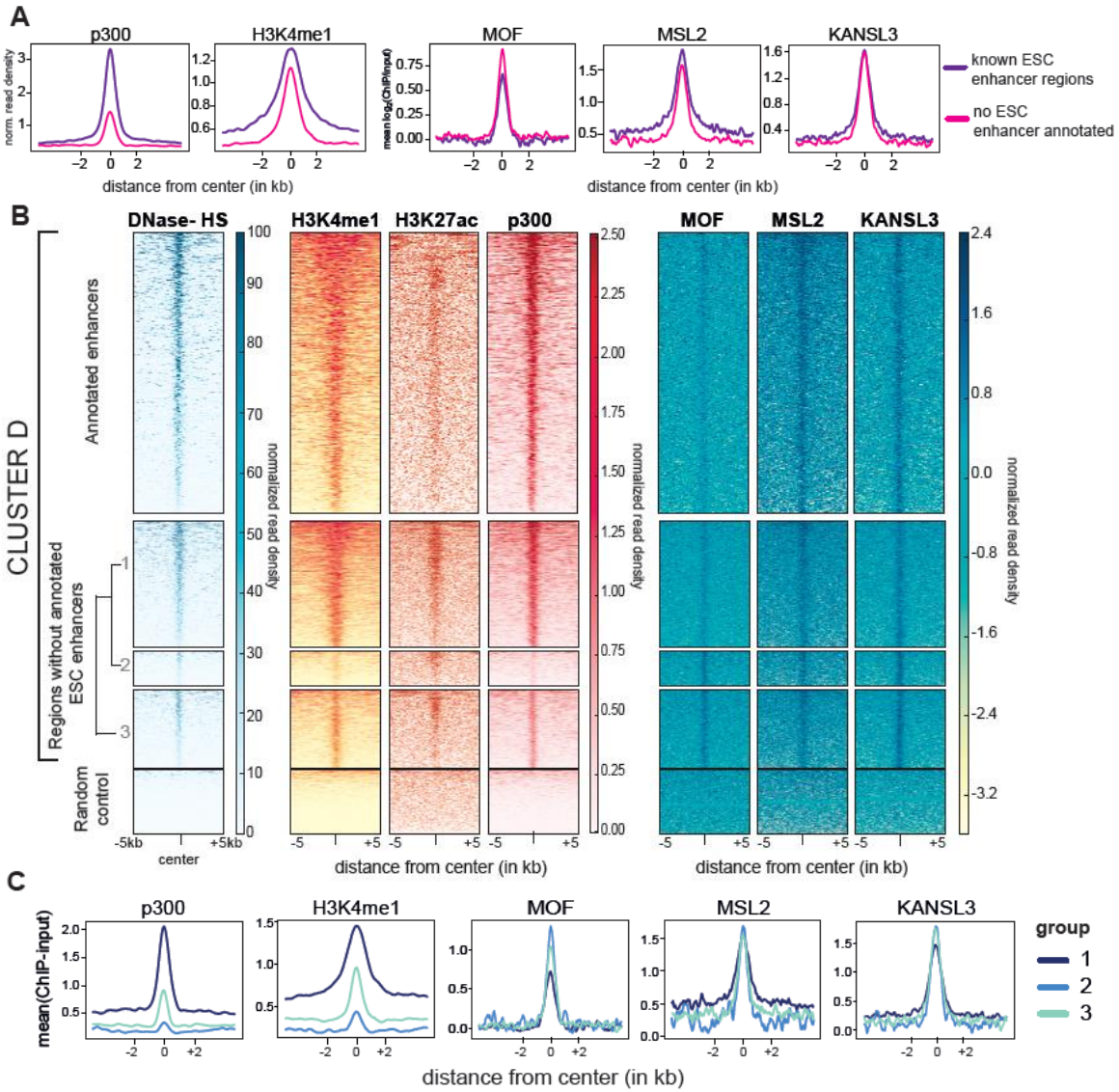


Figure 4—figure supplement 2: MOF is moderately enriched at non-canonical enhancers.

(A) Summary plots of ChIP-seq values for binding sites belonging to cluster D. The regions were divided based on the presence or absence of annotated ESC enhancers (Whyte et al., 2013, Creyghton et al., 2010).

(B) Heatmaps of ChIP-seq read densities of known enhancer markers for the ESC-specific binding sites of our proteins of interest (cluster D, see Figure 2) and random genomic regions. The binding sites of cluster D (excluding regions with TSSs) were divided into 2 basic groups based on the presence or absence of known ESC enhancers (Whyte et al., 2013, Creyghton et al., 2010). The latter group was further divided into 3 (arbitrarily numbered) sub-clusters based on hierarchical clustering of the values from DNase hypersensitivity sites, p300, H3K4me1 and our MOF sample (in ESCs). Heatmaps of the ESC-enhancer-containing regions were sorted according to p300, those of the sub-clustered regions were sorted according to the MOF signal.

(C) Related to (B), shown here are the corresponding summary plots of ChIP-seq values for cluster D binding sites that do not overlap with annotated enhancer regions (bottom part of the heatmaps in the figure above). The 3 indicated groups are based on the hierarchical clustering that was performed on p300, H3K4me1 and MOF values ("Regions without annotated ESC enhancers" in (B)).

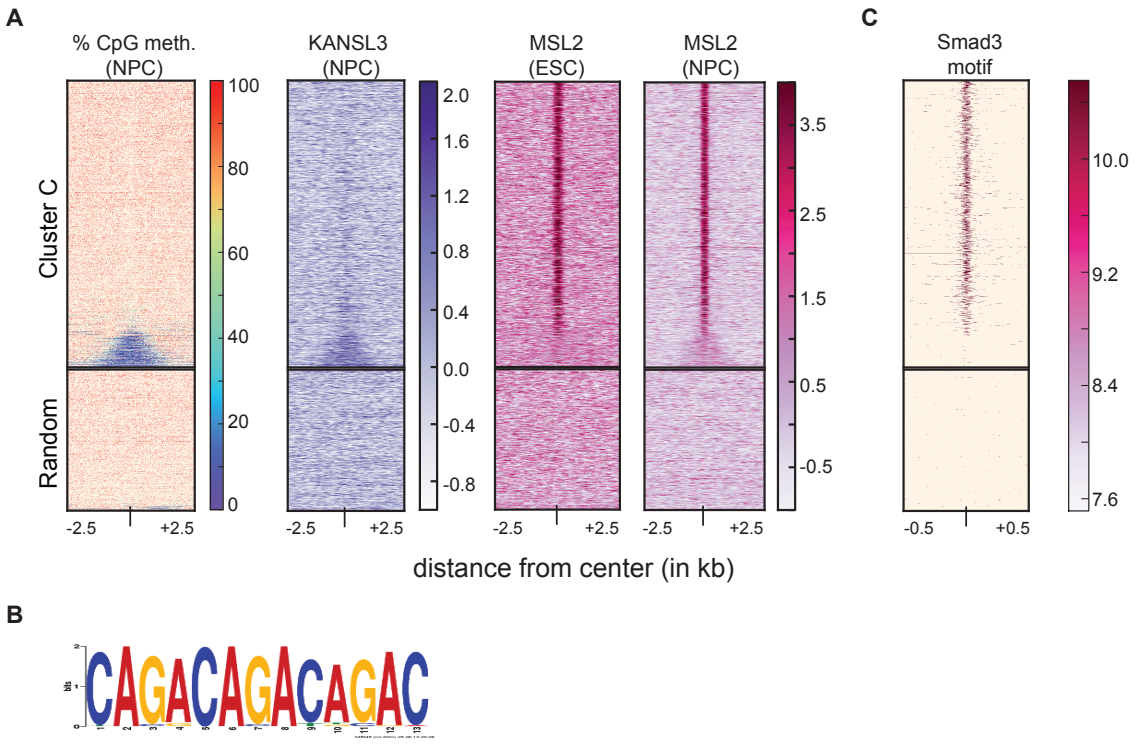


Figure 4—figure supplement 3: MSL2 has intergenic binding sites in DNA-hypomethylated regions that are enriched for SMAD3 binding sites.

(A) We extracted the percentage of methylated CpGs and the input-normalized ChIP-seq values from KANSL3 and MSL2 and 5 kb surrounding the center of the regions belonging to cluster C (Figure 2) and random genomic control regions. All heatmaps were sorted according to the percentages of methylated CpGs (Stadler et al., 2011).

(B) Motif obtained by MEME analysis on the top 200 MSL2 peaks within cluster C.

(C) Same as for (A), except that the score was the motif hit score for SMAD3 for 1 kb. See Methods and Materials for details.

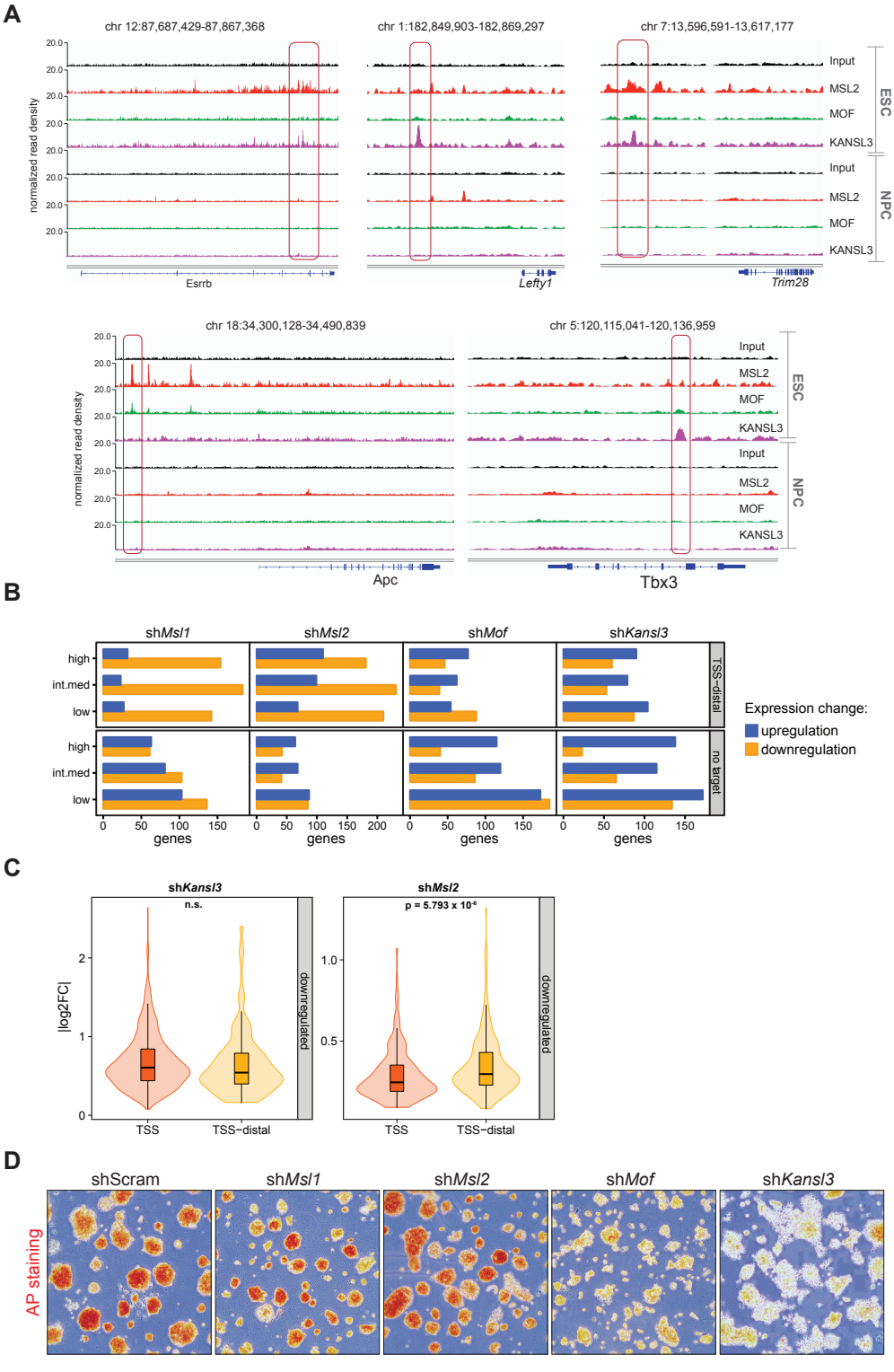


Figure 4—figure supplement 4: Biological significance of the TSS-distal binding sites of the investigated proteins.

(A) Genome browser snapshots of sequencing-depth normalized ChIP-seq and input profiles. Pink boxes mark the regions cloned and transfected into ESCs and NPCs for luciferase assays (Figure 4D).

(B) Genes that were significantly up- or downregulated in the respective shRNA-treatments compared to shScrambled were classified according to ChIP-seq peak overlaps (TSS-distal, no target) and expression strength in wild type ESCs (high, intermediate, low). See Methods and Materials for details of the classifications.

(C) Distribution of absolute log2 fold changes (shKansl3 or shMsl2 compared to shScrambled) for significantly downregulated genes. Different shades of orange indicate different target classes based on ChIP-seq experiments for KANSL3 or MSL2, respectively. P-values were calculated with Welch t-test.

(D) Alkaline phosphatase staining and morphology of ESC colonies in indicated knockdowns after 4 days growth under puromycin selection (see Figure 3—figure supplement 3A). MOF- and KANSL3-depleted cells demonstrate reduced alkaline phosphatase positive colonies with increased differentiation compared with MSL1- and MSL2-depleted cells and scrambled control.

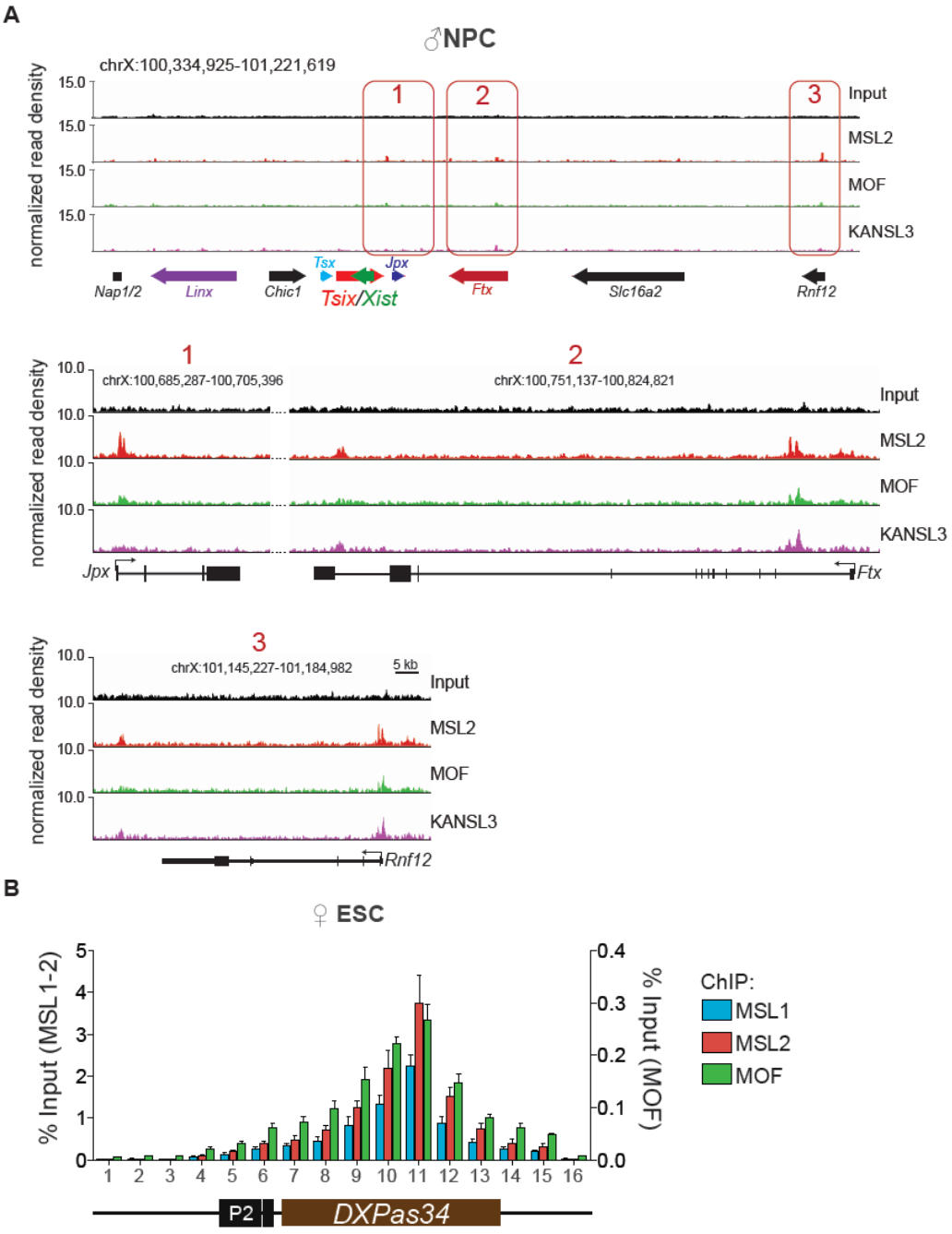
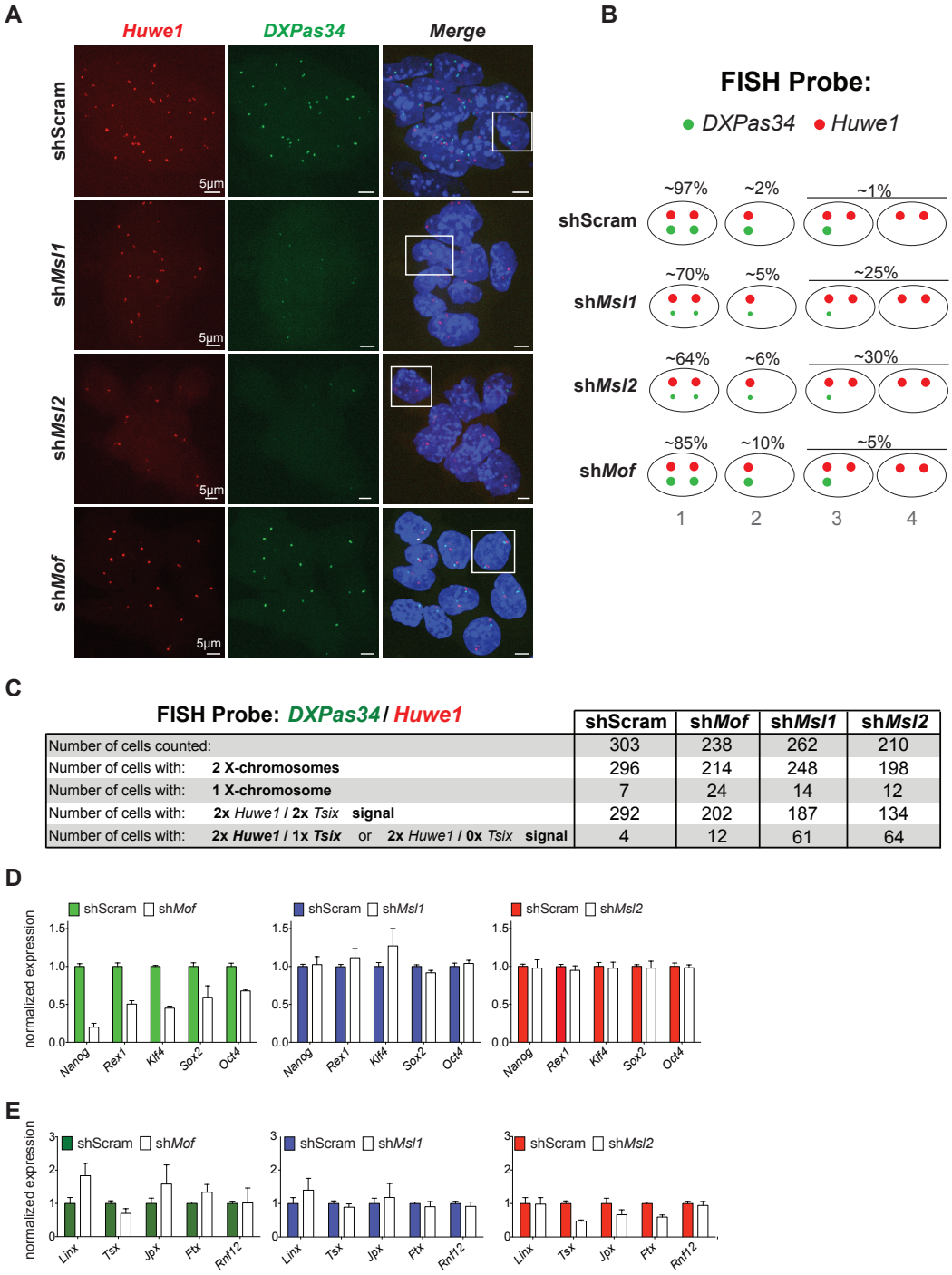


Figure 5—figure supplement 1: The MSL proteins bind to multiple loci within the X inactivation center (XIC).

(A) Genome browser snapshots of the mouse XIC (top panel) with three enlargements on Jpx, Ftx and Rnf12 genes (lower panels). Red boxes with corresponding numbers mark the enlarged regions presented in the lower panels. The exact genomic coordinates are indicated on top of each panel, arrows represent genes. The signals shown are the sequencing-depth normalized ChIP-seq profiles in NPCs.

(B) ChIP analysis of MSL1, MSL2 and MOF across the DXPas34 minisatellite in female ESCs. The x-axis labels indicate the genomic coordinates corresponding to the arrowheads in Figure 5A. The y-axes show the percentage of ChIP recovery for MSL1 and MSL2 (left-hand side) and MOF (right-hand side) normalized to input. For all ChIP experiments, 3 biological replicates were used; all results are expressed as mean +/- S.D.



C

FISH Probe: *DXPas34* / *Huwe1*

	shScram	sh*Mof*	sh*Msl1*	sh*Msl2*
Number of cells counted:	303	238	262	210
Number of cells with: **2 X-chromosomes**	296	214	248	198
Number of cells with: **1 X-chromosome**	7	24	14	12
Number of cells with: **2x *Huwe1* / 2x *Tsix* signal**	292	202	187	134
Number of cells with: **2x *Huwe1* / 1x *Tsix*** or **2x *Huwe1* / 0x *Tsix* signal**	4	12	61	64

Figure 6—figure supplement 1: Cells depleted of MSL1 or MSL2, but not MOF show loss of DXPas34 foci.

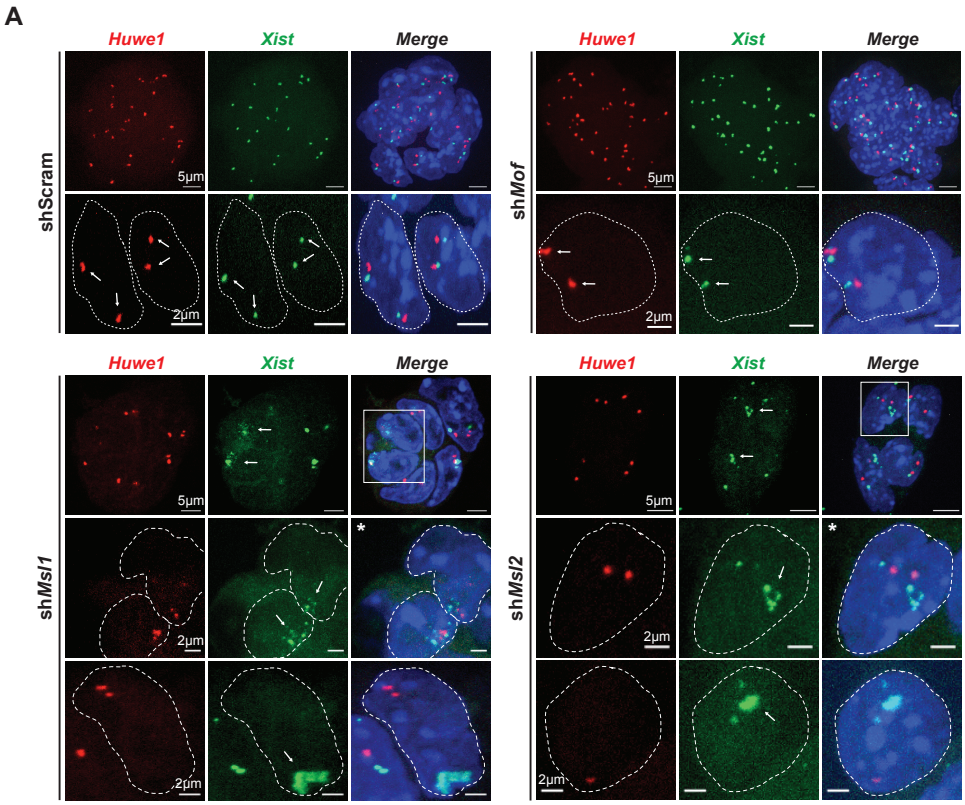
(A) RNA-FISH for *Huwe1* RNA (red) and *DXPas34* RNA (green) in shScrambled-, sh*Msl1*-, sh*Msl2*- and sh*Mof*-treated female ESCs. Shown here are examples of RNA-FISH signals for multicellular colonies and loss of *DXPas34* signal in MSL1- and MSL2-depleted cells. White boxes indicate cells enlarged and resented in Figure 6B. For all experiments, nuclei were counterstained with DAPI (blue).

(B) Summary of RNA-FISH for *DXPas34* and *Huwe1*. Red dots indicate the number of X chromosomes and green dots, *DXPas34* foci (smaller dot = reduced signal). Phenotypes that we observed in knockdowns are categorized into 4 groups containing cells with equal *Huwe1/DXPas34* ratio and with *DXPas34* loss. The percentages indicate how many cells per population showed the respective phenotype.

(C) Corresponding to Figure 6B. Summary of total cell counts from RNA-FISH for (*DXPas34*) and *Huwe1* in MSL1-, MSL2- or MOF-depleted female ESCs.

(D) Gene expression analysis for the indicated genes in female ESCs treated with scrambled RNA (shScram) or shRNA against *Mof*, *Msl1* and *Msl2*. All results are represented as relative values normalized to expression levels in shScram (normalized to *Hprt*) and expressed as means +/- S.D. in 3 biological replicates.

(E) Gene expression analysis for genes of the XIC in female ESCs treated with scrambled RNA or shRNA against *Msl1*, *Msl2* or *Mof*. All results are represented as relative values normalized to expression levels in shScrambled (normalized to *Hprt*) and expressed as means +/- S.D. for 3 biological replicates.



B

FISH Probe:

- *Xist*
- *Xist accumulation*

	1	2	3
shScram	~95%	~5%	~0.5%
sh*Msl1*	~85%	~10%	~5%
sh*Msl2*	~90%	~5%	~4.5%
sh*Mof*	~90%	~10%	~0.8%

C

♀ ESC **FISH Probe: *Xist* / *Huwe1***

	shScram	sh*Mof*	sh*Msl1*	sh*Msl2*
Number of cells counted:	454	245	185	226
Number of cells with: **2 X-chromosomes**	430	219	166	215
Number of cells with: **1 X-chromosome**	24	26	19	11
Number of cells with: **X-chromosomal coating**	2	2	10	10

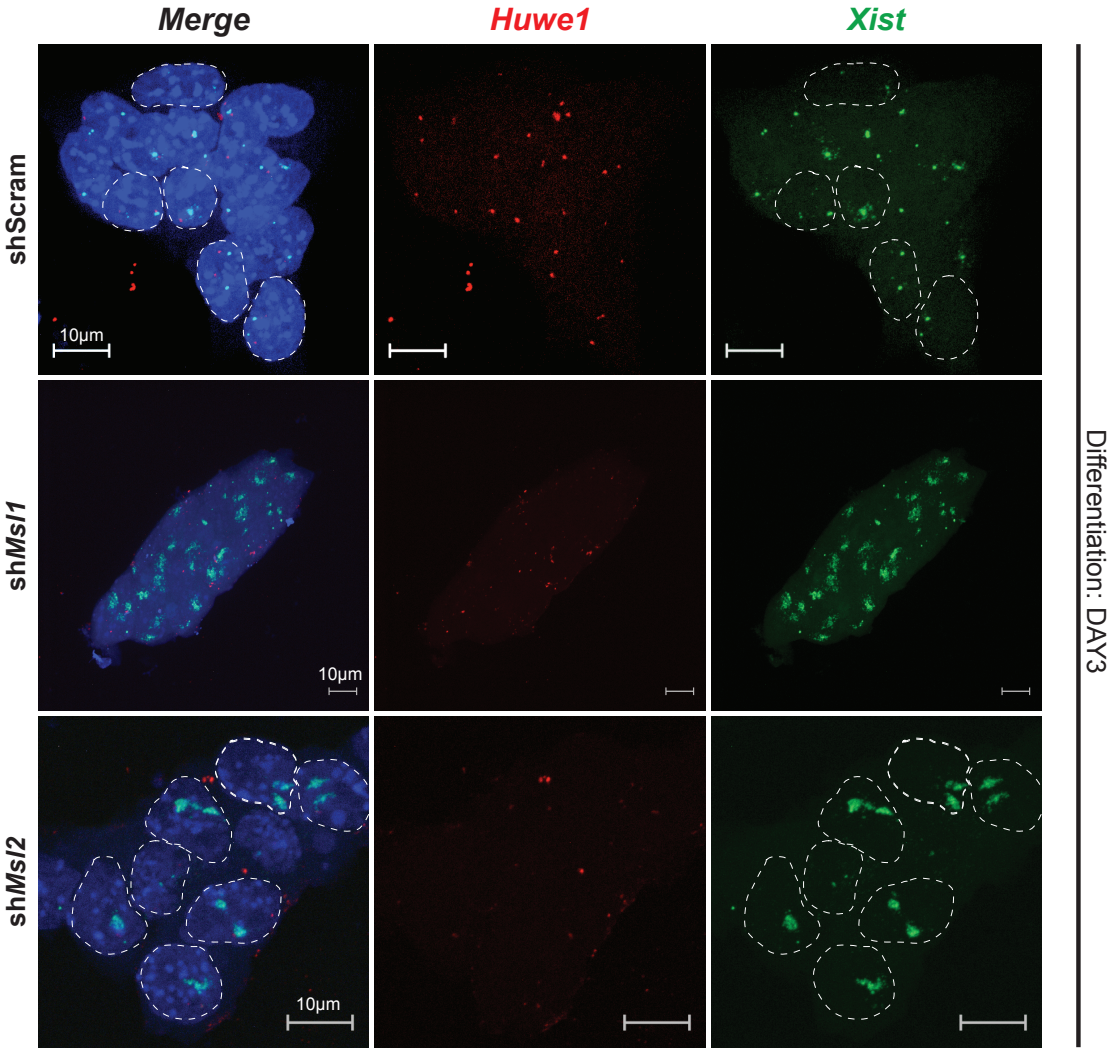
Figure 7—figure supplement 1: Depletion of MSL1 and MSL2 leads to occasional accumulation and spreading of Xist in undifferentiated ESCs.

(A) RNA-FISH for *Huwe1* RNA (red) and *Xist* RNA (green) in shScrambled- (top left) and sh*Mof*- (top right), sh*Msl1*- (bottom left) and sh*Msl2*-treated (bottom right) female ESCs. Shown here are additional examples of RNA-FISH for multicellular colonies and individual cells exhibiting *Xist*-mediated coating (see Figure 7A). White boxes indicate cells enlarged in Figure 7A. White arrows denote *Huwe1* and *Xist* foci. Dashed lines indicate nuclei borders. For all experiments, nuclei were counterstained with DAPI (blue).

(B) Summary of RNA-FISH for *Xist* and *Huwe1*. The number of green dots indicates the number of X chromosomes within the cell while the larger dot indicates *Xist* accumulation. Cells were classified into three phenotypic groups with cells showing sharp, localized *Xist* signals (once or twice) or *Xist* "clouds". The percentages indicate how many cells per population showed the respective phenotype.

(C) Corresponding to Figure 7A. Summary of the total cell counts from *Xist* and *Huwe1* RNA-FISH in indicated knockdowns.

A



B

FISH Probe: *Xist* / *Huwe1*	shScram	sh*Msl1*	sh*Msl2*	sh*Mof*
Number of cells counted:	253	209	285	239
Number of cells with: **0 *Xist* Clouds**	154 (60.9%)	77 (36.9%)	111 (38.9%)	118 (49.4%)
Number of cells with: **1 *Xist* Cloud**	97 (38.3%)	110 (52.6%)	148 (52%)	113 (47.3%)
Number of cells with: **2 *Xsit* Clouds**	2 (0.8%)	22 (10.5%)	26 (9.1%)	8 (3.3%)

Figure 7—figure supplement 2: Depletion of MSL1 and MSL2 lead to enhanced Xist accumulation in differentiating ESCs.

(A) RNA-FISH for *Huwe1* RNA (red) and *Xist* RNA (green) in shScrambled-, sh*Msl1*- and sh*Msl2*-treated differentiating (DAY3) female ESCs. Shown here are additional examples of RNA-FISH for multicellular colonies (see Figure 7C). Dashed lines indicate nuclei borders. For all experiments, nuclei were counterstained with DAPI (blue).

(B) Corresponding to Figure 7C-E. Summary of the total cell counts from *Xist* RNA-FISH in indicated knockdowns. Percentage of cells with respective phenotype indicated in red.

A.3 deepTools: a flexible platform for NGS analysis

Ramírez, F.*, **Dündar, F.***, Diehl, S., Grüning, B. A., Manke, T. (2014).
Nucleic Acids Research. doi:10.1093/nar/gku365
*shared authorship

I contributed to the development and debugging of the software tools (which were programmed to the largest extent by Fidel Ramírez).

I designed and generated the entire content of the deepTools documentation and help pages (`https://github.com/fidelram/deepTools/wiki`) which are still maintained by me.

I initiated the set-up of the public deepTools Galaxy instance `http://deeptools.ie-freiburg.mpg.de/` which was carried out by Sarah Diehl and generated all Galaxy Workflows, information pages (except the video tutorial), and histories.

I contributed to the main figure and devised, wrote, and revised the manuscript together with Fidel Ramírez and Thomas Manke. I generated the supplemental PDF version of the online deepTools documentation that is provided together with the manuscript.

Published online 05 May 2014 *Nucleic Acids Research, 2014, Vol. 12, Web Server issue* ***W187–W191***
doi: 10.1093/nar/gku365

deepTools: a flexible platform for exploring deep-sequencing data

Fidel Ramírez[1,†], Friederike Dündar[1,2,†], Sarah Diehl[1], Björn A. Grüning[3] and Thomas Manke[1,*]

[1]Max Planck Institute of Immunobiology and Epigenetics, Stübeweg 51, 79108 Freiburg, Germany, [2]Faculty of Biology, University of Freiburg, Schänzlestraße 1, 79104 Freiburg, Germany and [3]Department of Computer Science, University of Freiburg, Georges-Köhler-Allee 106, 79110 Freiburg, Germany

Received February 4, 2014; Revised April 5, 2014; Accepted April 15, 2014

ABSTRACT

We present a Galaxy based web server for processing and visualizing deeply sequenced data. The web server's core functionality consists of a suite of newly developed tools, called deepTools, that enable users with little bioinformatic background to explore the results of their sequencing experiments in a standardized setting. Users can upload pre-processed files with continuous data in standard formats and generate heatmaps and summary plots in a straightforward, yet highly customizable manner. In addition, we offer several tools for the analysis of files containing aligned reads and enable efficient and reproducible generation of normalized coverage files. As a modular and open-source platform, deepTools can easily be expanded and customized to future demands and developments. The deepTools webserver is freely available at http://deeptools.ie-freiburg.mpg.de and is accompanied by extensive documentation and tutorials aimed at conveying the principles of deep-sequencing data analysis. The web server can be used without registration. deepTools can be installed locally either stand-alone or as part of Galaxy.

INTRODUCTION

As high-throughput sequencing technologies (also: next-generation sequencing, NGS) continue to become cheaper, faster and more reliable, they are being adapted to address a wide spectrum of biological questions, ranging from transcriptome assessments (RNA-seq) to protein–DNA interactions (ChIP-seq), epigenetic marks (ChIP-seq, BS-seq) and the 3D-structure of the genome (4C, 5C, ChIA-PET, Hi-C). This has led to a widespread adoption of the technology in many laboratories that are now facing the formidable challenge of processing, analyzing and interpreting NGS sequencing data. To add to the burden, researchers are routinely asked to compare their novel experimental results with sequencing data deposited in public repositories like the Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/sra) and the European Nucleotide Archive (ENA, www.ebi.ac.uk/ena). In the past years, many programs have been developed for NGS data processing. The vast majority of these tools, however, require experience with the command-line and often do not provide graphical outputs to guide the interpretation of the results. Additionally, NGS data may suffer from several biases that should be taken into consideration for down-stream analyses. In our experience, the lack of user-friendly software along with comprehensive documentation and explanations of the multiple steps frequently deter biologists from taking part in the processing and analysis of their own data. To address these challenges, we have developed and refined a set of tools, called deepTools, that enable researchers to manage, manipulate and most importantly, explore their NGS data. Our tools are incorporated into the Galaxy framework, one of the most popular analysis platforms for NGS data, that offers easy and intuitive access to numerous bioinformatic applications and strongly supports documentation and reproducibility of analysis steps (1). deepTools provides standardized diagnostic plots for aligned reads, various normalization strategies, extensive support for format conversion and a set of tools for highly customizable meta-analyses and visualizations, such as heatmaps and summary plots. The utilities are easy-to-use as our web server handles the computational complexity and users will encounter the familiar Galaxy environment. The underlying software has been optimized for efficiency and highly parallelized processing, making the tools suitable for routine analysis of large-scale data. Apart from publication-ready images, users can export standardized output files that comply with the formats established by big sequencing consortia (BAM, bigWig, bedGraph, BED). This ensures compatibility with other Galaxy workflows and ex-

*To whom correspondence should be addressed. Tel: +49 0 761 5108 738; Fax: +49 0 761 5108 80738; Email: manke@ie-freiburg.mpg.de
†These authors contributed equally to the work.

W188 *Nucleic Acids Research, 2014, Vol. 12, Web Server issue*

ternal tools such as the IGV browser (2). Importantly, we provide extensive guidance to the tools' usage as well as NGS data analysis in general. Through continuously updated video tutorials, detailed case studies, FAQs and discussion groups we aim to support users throughout their work and lower the barrier for researchers unfamiliar with specific NGS data analyses.

WEB SERVER

The deepTools web server is available at http://deeptools.ie-freiburg.mpg.de. We provide our tools for the analysis of NGS data within the Galaxy framework (1). As depicted in Table 1 and described in more detail in the Supplementary Manual, deepTools modules can be classified into three components: (i) global assessments of aligned reads (quality control), (ii) the generation of normalized coverage files (data extraction and reduction) and (iii) visual exploration and cluster analysis (data interpretation). In addition to the deepTools suite developed by us, our web server includes other tools for data import, for text file manipulations such as filtering and sorting, for operations on genomic intervals (BED files) and for peak calling on ChIP-seq data. These tools have been selected based on the demand by our users and were installed via the Galaxy Tool Shed (4). To facilitate reproducible research, Galaxy will keep track of every operation performed on any given data set and will store the results in the user's history panel. While this is not required, we recommend frequent users of our Galaxy instance to set up an account so that customized workflows can be stored and re-used.

WORKFLOW

Quality Control

Users will typically start by uploading a file of aligned reads (preferably BAM format, but SAM files can also be uploaded and subsequently converted) that they obtained from an NGS facility (Figure 1A). We offer several tools for assessing whether the distribution of aligned reads meets the user's expectation (for a succinct list of all deepTools that are currently available, please see Table 1). One of the most versatile tools is *bamCorrelate* which calculates the correlation of read numbers for two or more files of aligned reads. Based on the correlation measures, *bamCorrelate* generates a clustered heatmap that depicts the distances between the samples (Figure 1B). *bamCorrelate* can thus reveal the similarity between replicates, it can also be used to compare new samples with published data, to identify sample swaps and to generally see whether samples that are expected to show similar read distributions cluster together. In addition to the basic correlation analysis by *bamCorrelate*, the deepTools *computeGCbias* and *correctGCbias* produce diagnostic plots that help detect and correct GC bias using the most recent insights into GC bias properties of NGS samples (5, 6) (Figure 1C). Specifically for ChIP-seq experiments, the *bamFingerprint* module generates simple and informative plots to visually assess the ChIP signal strength as suggested by (7) (see Supplementary Manual for plots and details).

Data processing for downstream analyses

Following the initial assessments of raw read distributions, BAM files are usually processed to decrease their size and to obtain normalized measures of sequencing coverage (Figure 1A). These steps are often perceived as particularly challenging, but deepTools offers two easy-to-use modules (*bamCoverage* and *bamCompare*, see Table 1) that allow for a wide range of normalizations and mathematical operations based on the number of mapped reads covering a genomic region of fixed length (e.g. 25 bp). For example, *bamCoverage* could be used to individually normalize samples with different total read numbers (different sequencing depths) to allow for unbiased comparisons of signal intensities. *bamCompare*, on the other hand, can be used to generate scores based on two BAM files such as differences or ratios [e.g. (sequencing coverage in treatment sample)/(sequencing coverage in control sample)]. The output of both modules is saved in bigWig files describing the position of each genome region and the score associated to them. Due to the significant decrease in size compared to BAM files, the bigWig format is recommended by UCSC for storing and sharing continuous genome-wide sequencing data. These files can be imported into multiple other applications, including genome browsers, and deepTools uses the indexed nature of those files to parallelize operations which significantly speeds up downstream analyses.

Visualization

NGS studies seek to unveil and characterize signal patterns on a global scale. While genome browsers allow for individual snapshots of specific loci, heatmaps and summary plots have become the preferred means to represent data for the simultaneous comparison of numerous and possibly large regions. The deepTools modules *computeMatrix*, *heatmapper* and *profiler* facilitate the creation of such plots. Users must supply a bigWig file of scores (that can be generated with the tools discussed above) and at least one file containing the genomic regions of interest (in BED, INTERVAL or GFF format) for which the values will be extracted and displayed. The tools offer two modes: *reference-point* will center the profile or heatmap around the start, middle or end of each genomic region. This can be used, for example, to create a profile of reads around the transcription start site of genes. The *scale-region* mode will fit all given regions to a user-specified length which is useful to compare read coverage patterns for regions of different lengths such as the bodies of genes (Fig. 1D). In addition to user-supplied groups of regions (Figure 1D, center panel), k-means clustering can be applied to identify regions with similar score distributions in an automated, virtually unbiased fashion that allows for the discovery of unexpected patterns (Figure 1D, right panel). We have separated the calculation of the score matrix from the generation of the image (see tools for visualization in Table 1) because the first step is computationally much more intensive than the latter one. Once the values are calculated with *computeMatrix*, *heatmapper* and *profiler* can quickly produce publication-ready images as they offer a large range of options (e.g. color schemes, labels, titles, format) for optimal data display.

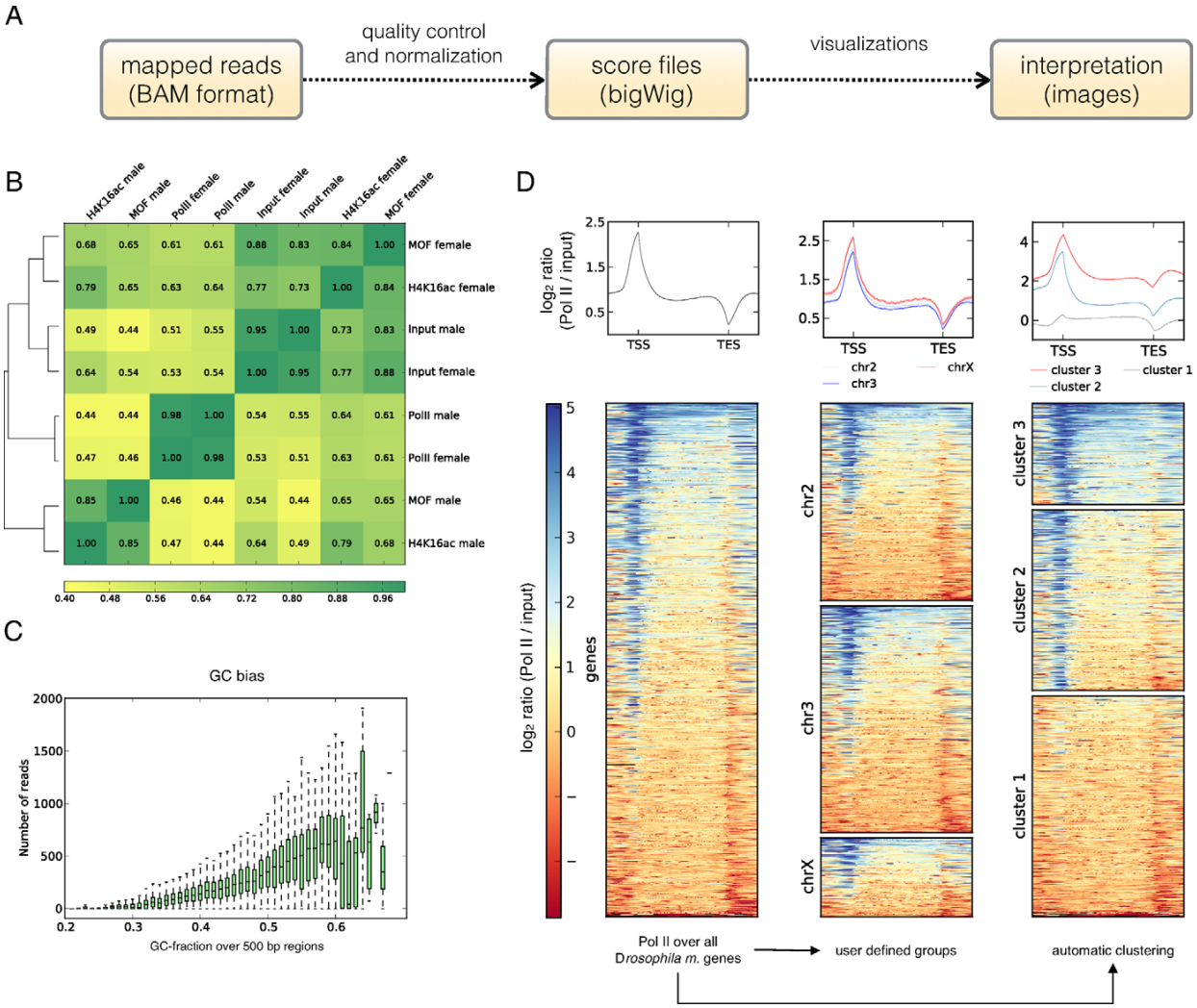


Figure 1. Examples of images created with deepTools. (**A**) Overview of the deepTools workflow that offers tools for visualization and for the intermediary NGS data processing steps (Table 1). Users can either start by directly uploading bigWig files for the generation of heatmap and summary plots, or they may upload BAM files, perform quality controls on them and produce normalized coverage files that can then be used for the visualization steps. (**B**) Clustered heatmap produced by the deepTools *bamCorrelate* module. Shown here are the Pearson correlation coefficients of various ChIP-seq samples; the clustering reveals that the ChIP signals of MOF in male and female cells differ significantly [data from (3), ENA accession: PRJEB3031]. (**C**) Exemplary plot produced by *computeGCbias* to assess the GC distribution of reads within a given BAM file. The sample here shows the typical over-representation of reads with high GC content that is often observed after excessive polymerase chain reaction amplification. An additional plot (not shown here, see Supplementary Materials) takes the genome-specific expectation into consideration. (**D**) Examples of different summary plots and heatmap versions generated by deepTools using normalized read coverages from a ChIP-seq for RNA polymerase II (Pol II) in male *Drosophila melanogaster* cells. *bamCompare* was used to calculate the $\log_2$ ratio of Pol II and the control sample. The resulting bigWig file was supplied together with a BED file containing the gene regions to *computeMatrix* which was used in *scale-region* mode to extract the scores for the genes. The left-most plot shows the subsequent default output of *heatmapper*: The Pol II signal over the body of all genes can be seen and genes are sorted according to the mean score. The summary plot on top of the heatmap indicates that, on average, Pol II is most strongly enriched around the start of genes which is also visible in the heatmaps. The center plot shows the same data, but here we supplied three individual BED files, one per chromosome. The summary plot suggests that the genes on the X chromosome show slightly higher average signals than those on chromosomes 2 and 3 which is consistent with the transcriptional upregulation of the male X chromosome in *Drosophila* (3). Additionally, *heatmapper* allows for the automated clustering of the data as exemplified in the right-most heatmap. Only by indicating the number of clusters to be found, the clustering results in an image where one can clearly differentiate between genes with elevated amounts of Pol II at the promoter and over the gene body (cluster 3) from genes with Pol II primarily at the promoters (cluster 2) and those with very weak Pol II signal (cluster 1). Abbreviations: bp, base pair; chr, chromosome; input, control sample for ChIP-seq experiments; Pol II, RNA polymerase II; TES, transcription end site; TSS, transcription start site.

W190 *Nucleic Acids Research, 2014, Vol. 12, Web Server issue*

Table 1. Overview of currently available deepTools

Tool name	Type	Input files	Main output	Application
bamCorrelate	QC	2 or more BAM	Clustered heatmap of similarity measures	Determine Pearson or Spearman correlations between read distributions
bamFingerprint	QC	2 BAM	Diagnostic plot	Assess enrichment strength of a ChIP-seq sample versus a control
computeGCBias	QC	1 BAM	Diagnostic plots	Compare expected and observed GC distribution of reads
correctGCBias	Normalization	1 BAM	BAM or bigWig	Obtain GC-corrected read (coverage) file
bamCoverage	Normalization	1 BAM	bedGraph or bigWig	Obtain normalized read coverage of a single BAM
bamCompare	Normalization	2 BAM	bedGraph or bigWig	Normalize 2 BAM files to each other with a mathematical operation of Choice (fold change, log2 (ratio), sum, difference)
computeMatrix	Visualization	1 bigWig, min. 1 BED	gzipped table	Calculate the values for heatmaps and summary plots
profiler	Visualization	gzipped table from computeMatrix	xy-plot (summary plot)	Average profiles of read coverage for (groups of) genome regions
heatmapper	Visualization	gzipped table from computeMatrix	(Un)clustered heatmap or read coverages	Identify patterns of read coverages for genome regions

Here, we only indicate the main output files, but every data table underlying any image produced by deepTools can be downloaded and used in subsequent analyses. For a comparison of functionalities with previously published web servers, see Supplementary Table S1.

EXAMPLES AND HELP

Within the web server (http://deeptools.ie-freiburg.mpg.de) we provide a video tutorial and sample data (http://deeptools.ie-freiburg.mpg.de/library) to familiarize every user with the common workflows and various modules of deepTools. The functionality of each module is illustrated with detailed examples from real-life NGS analyses and can be seen once a tool is selected. In addition, we have compiled extensive documentation and tutorials that introduce the Galaxy framework, explain deepTools in more depth and provide step-by-step protocols for typical NGS analyses that can be carried out using our web server. Questions and comments about deepTools can be directed to the deepTools mailing list (deeptools@googlegroups.com) and we regularly update our FAQ section.

IMPLEMENTATION

deepTools is written in Python; the deepTools suite is available as a one-click installation for any local Galaxy instance via the Galaxy Tool Shed (http://toolshed.g2.bx.psu.edu/view/bgruening/deeptools). For technical details of our web server, see Supplementary Table S2. For advanced users and developers, we also offer a stand-alone version for command line usage and free access to the code. More information on the different installation procedures can be found at the code repository (https://deeptools.github.io/).

DISCUSSION AND OUTLOOK

As NGS technologies are advancing inexorably, it has become a key challenge to match the data production rate with our ability to efficiently analyze new data sets. NGS analyses are often characterized by specialized and custom-made scripts, hidden filtering strategies and a subsequent lack of standardization and reproducibility. Now that the wide-spread adoption of sequencing technologies goes beyond large consortia and reaches groups with less bioinformatic support, it is paramount to provide standardized and user-friendly tools for NGS data visualization and interpretation. With deepTools we offer an expandable platform to bridge the gap between the early steps of raw data processing and the iterative data exploration in the search for biological insights. Intuitive usage and seamless integration into the Galaxy platform make deepTools ideally suited for data sharing and reproducible research, for biologists and bioinformaticians alike. New technologies and experimental refinements will bring about their own challenges and specific needs for new data types, normalization and interpretation strategies. Owing to the modular and flexible design of deepTools, additional tools can easily be included in future releases. Moreover, as a platform based on open source code, these tools represent the result of a community effort and have the potential to set standards for the visualization of genome-wide data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENT

The authors would like to thank all users of deepTools, in particular Fabian Kilpert and Lauren Solomon for exten-

sive testing of the web server and the feedback on the tutorial materials.

FUNDING

German Research Foundation [SFB 992, Project Z01]; German Epigenome Programme DEEP [01KU1216G]. Source of Open Access funding: own funds.
Conflict of interest statement. None declared.

REFERENCES

1. Goecks,J., Nekrutenko,A. and Taylor,J. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.
2. Robinson,J.T., Thorvaldsdóttir,H., Winckler,W., Guttman,M., Lander,E.S., Getz,G. and Mesirov,J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.
3. Conrad,T., Cavalli,F.M.G., Vaquerizas,J.M., Luscombe,N.M. and Akhtar,A. (2012) Drosophila dosage compensation involves enhanced Pol II recruitment to male X-linked promoters. *Science*, **337**, 742–746.
4. Blankenberg,D., Von Kuster,G., Bouvier,E., Baker,D., Afgan,E., Stoler,N., Taylor,J. and Nekrutenko,A. (2014) Dissemination of scientific software with Galaxy ToolShed. *Genome Biol.*, **15**, 403.
5. Benjamini,Y. and Speed,T.P. (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.*, **40**, e72.
6. Cheung,M.-S., Down,T.a., Latorre,I. and Ahringer,J. (2011) Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic Acids Res.*, **39**, e103.
7. Diaz,A., Park,K., Lim,D.A. and Song,J.S. (2012) Normalization, bias correction, and peak calling for ChIP-seq. *Stat. Appl. Genet. Mol. Biol.*, **11**, 9.

A.3.1 Supplemental Material

Supplementary table 1. Comparison of available software similar to deepTools

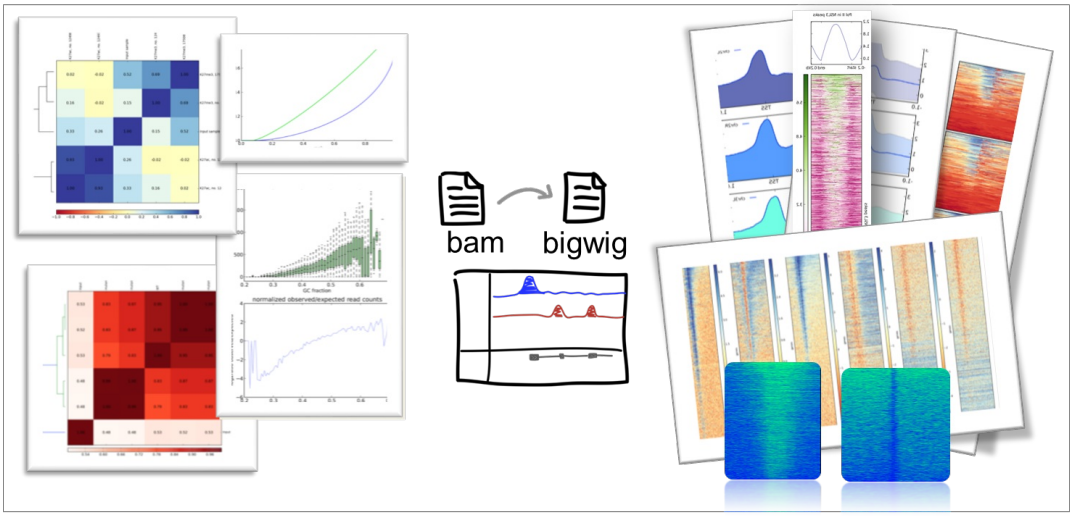
Tool suite	web-server	if there is a webserver, can it be used without login?	Galaxy implemen-tation	Parallelization	QC of BAM files (beyond FASTQC)	processing and normalization of BAM files	generates customizable images	allows for export of data	tutorial	Software	stand alone local installation	url
deepTools	yes	yes	yes	yes	yes	yes	yes	yes	yes	Galaxy & python	yes	http://deeptools.ie-freiburg.mpg.de
seqminer	no	–	no	yes	no	just normalization	yes	yes	website not reachable	Java	yes	http://bips.u-strasbg.fr/seqminer/
spark	yes	no	no	yes	no	no	no	yes	yes	Java	yes	http://www.sparkinsight.org/
CISTROME	yes	no	yes	no	no	no	limited	yes	Galaxy page	Galaxy, R, python	difficult	http://genomebiology.com/2011/12/8/r83
HOMER	no	–	no	no	yes	yes	no	yes	yes	perl	yes	http://homer.salk.edu/homer/chipseq/
NGS plot	no	–	yes	--	no	no	limited	images	no	R, python, perl	yes	https://code.google.com/p/ngsplot/
GeneProf	yes	not for analysis	no	--	no	yes	yes	yes	yes	Java, R, LaTeX	no	http://www.geneprof.org/

Supplementary table 2. Performance details of the deeptools Galaxy server.

Hardware	
processors	16
RAM	141 GB
storage (directly attached)	2 TB
Galaxy Performance Settings	
concurrent users	unlimited
CPU cores per job	6
concurrent jobs per user	1
concurrent jobs	2
queueing system	Galaxy built-in
Exemplary run time	
BAM file	170 M reads
BED file	10000 regions
bamCoverage	10 minutes
computeMatrix	1 minute

deepTools: a flexible platform for exploring deep-sequencing data

MANUAL



Fidel Ramírez, Friederike Dündar, Sarah Diehl, Björn A. Grüning, Thomas Manke

Bioinformatics Group, Max-Planck-Institute of Immunobiology and Epigenetics & Department of Computer Science, University of Freiburg

Web server (incl. sample data): deepTools.ie-freiburg.mpg.de
Code: github.com/fidelram/deepTools

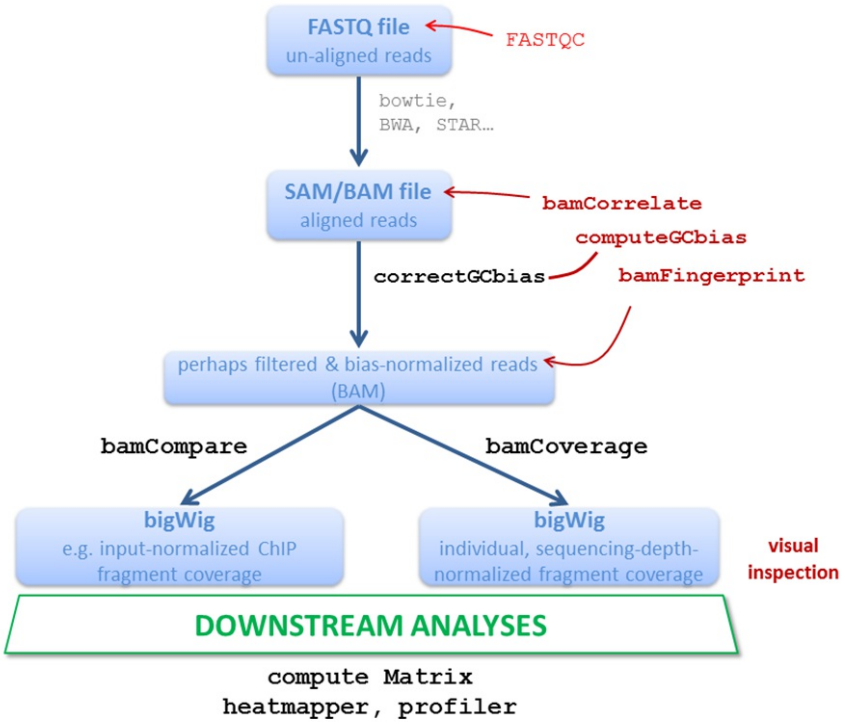
This document contains some chapters of our wiki on deepTools usage for NGS data analysis. For the most updated version of our help site and for **more information about deepTools, a brief introduction into Galaxy** as well as **step-by-step protocols**, please visit: https://github.com/fidelram/deepTools/wiki

Why we built deepTools

The main reason why deepTools was started is the simple fact that in 2011 we could not find tools that met all our needs for NGS data analysis. While there were individual tools for separate tasks, we wanted software that would fulfill *all* of the following criteria:

- **efficiently extract reads from BAM files** and perform various computations on them
- **turn BAM files of aligned reads into bigWig files** using different normalization strategies
- make use of **multiple processors** (speed!)
- generation of **highly customizable images** (change colors, size, labels, file format etc.)
- enable **customized down-stream analyses** which requires that every data set that is being produced can be stored by the user
- **modular approach** - compatibility, flexibility, scalability (i.e. we can add more and more modules making use of established methods)

The flow chart below depicts the different tool modules that are currently available within deepTools (deepTools modules are written in bold red and black font). If you are not familiar with the file names shown here, please see our Glossary at the end of the document for more information.



How we use deepTools

You will find many examples from ChIP-seq analyses in this tutorial, but this does not mean that deepTools is restricted to ChIP-seq data analysis. However, some tools, such as *bamFingerprint* specifically address ChIP-seq-issues. (That being said, we do process quite a bit of RNA-seq, other -seq and genomic sequencing data using deepTools, too, but many normalization issues arose during handling of ChIP-seq data).

As shown in the flow chart above, our work usually begins with one or more FASTQ file(s) of deeply-sequenced samples. After a first quality control using FASTQC, we align the reads to the reference genome, e.g. using bowtie2][].
We then use deepTools to assess the quality of the aligned reads:

1. **Correlation between BAM files** (*bamCorrelate*). This is a very basic test to see whether the sequenced and aligned reads meet your expectations. We use this check to assess the reproducibility - either between replicates and/or between different experiments that might have used the same antibody/the same cell type etc. For instance, replicates should correlate better than differently treated samples.
2. **GC bias check** (*computeGCbias*). Many sequencing protocols require several rounds of PCR-based amplification of the DNA to be sequenced. Unfortunately, most DNA polymerases used for PCR introduce significant GC biases as they prefer to amplify GC-rich templates. Depending on the sample (preparation), the GC bias can vary significantly and we routinely check its extent. In case we need to compare files with different GC biases, we use the *correctGCbias* module to match the GC bias. See the paper by Benjamini and Speed for many insights into this problem.
3. **Assessing the ChIP strength**. We do this quality control do to get a feeling for the signal-to-noise ratio in samples from ChIP-seq experiments. It is based on the insights published by Diaz et al..

Once we are satisfied by the basic quality checks, we normally **convert the large BAM files into a leaner data format, typically bigWig**. bigWig files have several advantages over BAM files that mainly stem from their significantly decreased size:

- useful for data sharing & storage
- intuitive visualization in Genome Browsers (e.g. IGV)
- more efficient downstream analyses are possible

The deepTools modules *bamCompare* and *bamCoverage* do not only allow the simple conversion from BAM to bigWig (or bedGraph for that matter), **the main reason why we developed those tools was that we wanted to be able to *normalize* the read coverages** so that we could compare different samples despite differences in sequencing depth, GC biases and so on.

Finally, once all the files have passed our visual inspections, the fun of downstream analyses with *heatmapper* and *profiler* can begin!

deepTools overview

deepTools consists of a set of modules that can be used independently to work with mapped reads. We have subdivided such tasks into *quality controls* (QC), *normalizations* and *visualizations*.

Here's a concise summary of the tools. In the following pages, you can find more details about the individual tools. We have included many screenshots of our Galaxy deepTools web server to explain the usage of our tools. In addition, we show the commands for the stand-alone usage, as they often indicate the options that one should pay attention to more succinctly.

tool	type	input files	main output file(s)	application
bamCorrelate	QC	2 or more BAM	clustered heatmap	Pearson or Spearman correlation between read distributions
bamFingerprint	QC	2 BAM	1 diagnostic plot	assess enrichment strength of a ChIP sample
computeGCbias	QC	1 BAM	2 diagnostic plots	calculate the expected and observed GC distribution of reads
correctGCbias	QC	1 BAM, output from computeGCbias	1 GC-corrected BAM	obtain a BAM file with reads distributed according to the genome's GC content
bamCoverage	normalization	BAM	bedGraph or bigWig	obtain the normalized read coverage of a single BAM file
bamCompare	normalization	2 BAM	bedGraph or bigWig	normalize 2 BAM files to each other using a mathematical operation of your choice (e.g. log2ratio, difference)
computeMatrix	visualization	1 bigWig, 1 BED	gzipped table, to be used with heatmapper or profiler	compute the values needed for heatmaps and summary plots
heatmapper	visualization	computeMatrix output	heatmap of read coverages	visualize the read coverages for genomic regions
profiler	visualization	computeMatrix output	summary plot	visualize the average read coverages over a group of genomic regions

QC of aligned reads

These tools work on BAM files that contain read-related information (e.g. read DNA sequence, sequencing quality, mapping quality etc.). They are typically generated by read alignment programs such as bowtie2.

The following tools will allow you to inspect your BAM files more closely.

bamCorrelate

This tool is useful to assess the overall similarity of different BAM files. A typical application is to check the correlation between replicates or published data sets, but really, you can apply it to any inquiry that boils down to the question: "How (dis)similar are these BAM files?".

What it does

bamCorrelate computes the overall similarity between **two or more** BAM files based on read coverage (number of reads) within genomic regions, i.e. for each *pair* of BAM files reads overlapping with the same genomic intervals are counted and the counts are correlated. The result is a table of correlation coefficients that will be visualized as a heatmap. The correlation coefficient indicates how "strong" the relationship between the two samples is and it will consist of numbers between -1 and 1. (-1 indicates perfect anticorrelation, 1 perfect correlation.)

We offer two different functions for the correlation computation: Pearson or Spearman. In short, Pearson is an appropriate measure for data that follows a normal distribution, while Spearman does not make this assumption and is generally less driven by outliers, but with the caveat of also being less sensitive.

NOTE: bamCorrelate usually takes a long time to finish, thus it is advisable to first run the tool for a tiny region (using the --region option) to adjust plotting parameters like colors and labels before running the
whole computation.

Important parameters

bamCorrelate can be run in 2 modes: *bins* and *bed*.

In the *bins* mode, the correlation is based on read coverage over consecutive bins of equal size (10k bp by default). This mode is useful to assess the overall similarity of BAM files. The bin size and the distance between bins can be adjusted.

Note that by default, a filtering of extremes is done, when *bins* mode is selected.

In the *BED-file option*, the user supplies a list of genomic regions in BED format in addition to the BAM files. bamCorrelate subsequently uses this list to compare the read coverages for these regions only. This can be used, for example, to compare the ChIP-seq coverages of two different samples for a set of peak regions.

In addition to specifying the regions for which the read numbers should be compared (random regions in the bins mode, selected regions in the BED-file mode), you can also specify what kind of correlation measure you would like to compute: Pearson or Spearman. In short, Pearson is an appropriate measure for data that follows a normal distribution while Spearman does not make this assumption and is generally less driven by outliers. As genome-wide sequencing data very rarely follows a normal distribution and we often encounter few regions that capture extremely high read counts (= outlier), we tend to prefer the Spearman correlation coefficient.

Output files:

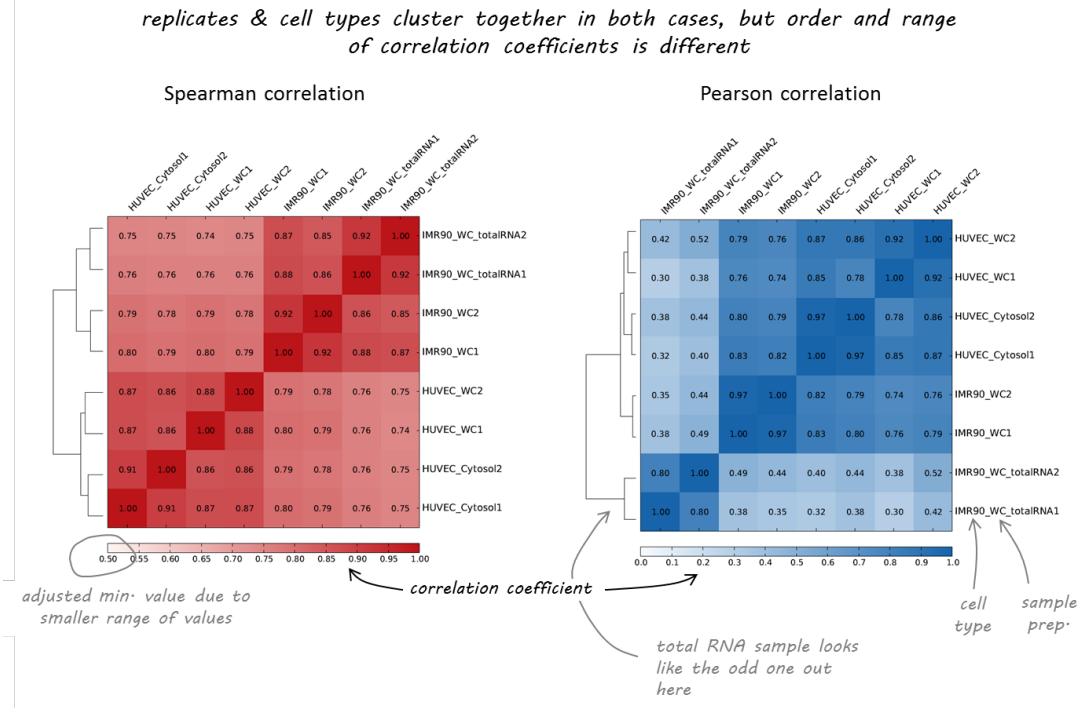
- **diagnostic plot** the plot produced by bamCorrelate is a clustered heatmap displaying the values for each pair-wise correlation, see below for an example
- **data matrix** (optional) in case you want to plot the correlation values using a different program, e.g. R, this matrix can be used

Example Figures

Here is a result of running bamCorrelate: heatmaps where the pairwise correlation coefficients are depicted by varying color intensities and are clustered using hierarchical clustering.

For the two example plots below, we supplied BAM files of RNA-seq data from different human cell lines that we had downloaded from the ENCODE project and a list of genes from RefSeq (Note that you can supply any number of BAM files that you would like to compare. In Galaxy, you just click "Add BAM file", in the command line you simply list all files one after the other, giving meaningful names via the --label option). We then calculated the pair-wise correlations of read numbers for the different genes, once with Spearman correlation, once with Pearson correlation.

You can find the original file at http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/ (just add the file names you see in the command at the end).



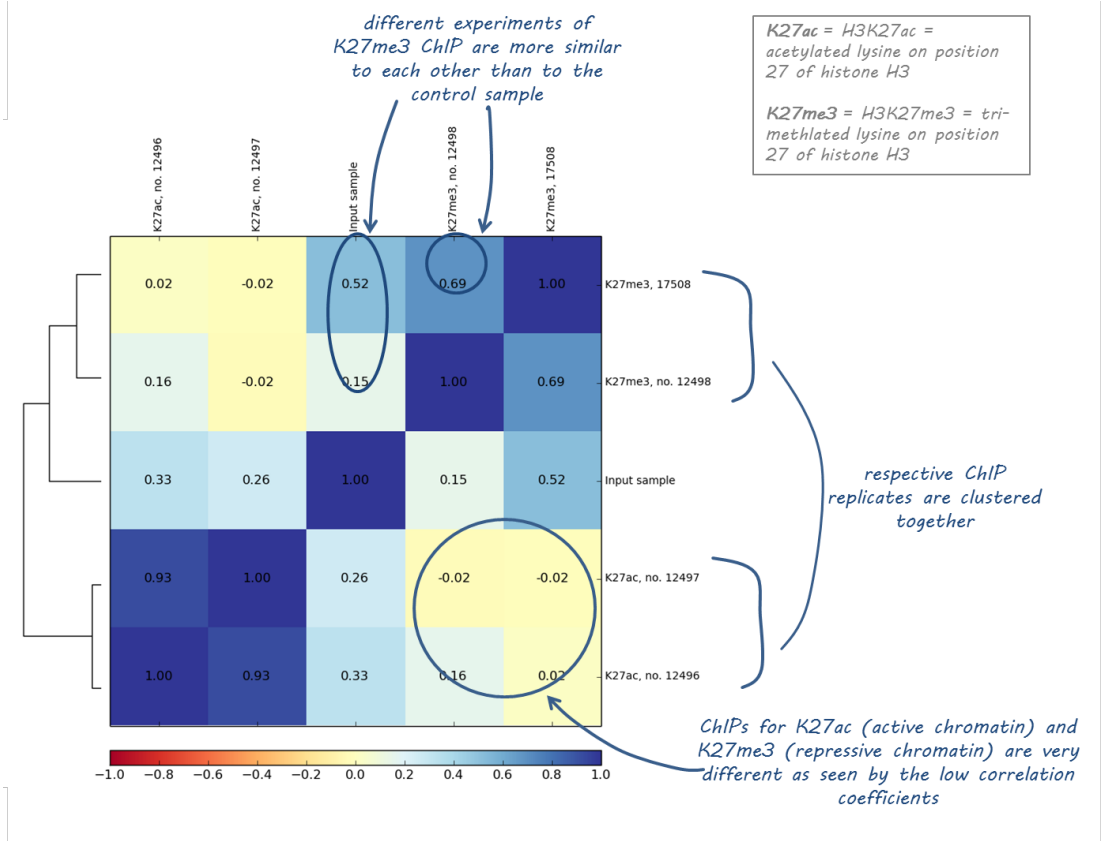
As you can see, both correlation calculations more or less agree which samples are nearly identical (the replicates, indicated by 1 or 2 at the end of the label). The Spearman correlation, however, seems to be more robust and meets our expectations more closely as the two different cell types (HUVEC and IMR90) are clearly separated.

This is the command that was used to generate the plot on the left-hand side:

```
$ deepTools-1.5.7/bin/bamCorrelate BED-file \
--BED RefSeq_Genes.bed \
--bamfiles wgEncodeCshlLongRnaSeqImr90CellPapAlnRep1.bam \
  wgEncodeCshlLongRnaSeqImr90CellPapAlnRep2.bam  \
  wgEncodeCshlLongRnaSeqImr90CellTotalAlnRep1.bam   \
  wgEncodeCshlLongRnaSeqImr90CellTotalAlnRep2.bam \
  wgEncodeCshlLongRnaSeqHuvecCellPapAlnRep1.bam \
  wgEncodeCshlLongRnaSeqHuvecCellPapAlnRep2.bam \
  wgEncodeCshlLongRnaSeqHuvecCytosolPapAlnRep3.bam \
  wgEncodeCshlLongRnaSeqHuvecCytosolPapAlnRep4.bam \
--labels IMR90_WC1 IMR90_WC2 IMR90_WC_totalRNA1 \
IMR90_WC_totalRNA2 HUVEC_WC1 HUVEC_WC2 HUVEC_Cytosol1 HUVEC_Cytosol2 \
--binSize 1000 --corMethod spearman -f 200 \
--colorMap Reds --zMin 0.5 --zMax 1 -o correlation_spearman.pdf
```

Here is another example of ChIP samples for two different histone marks (the histone marks are abbreviated H3K27me3 and H3K27ac and have been shown to mark inactive and active chromatin, respectively. For our example, H3K27ac was ChIPed by the same experimentator for different cell populations while H3K27me3 was performed with the same antibody, but at different times. You can see that the correlation between the H3K27ac replicates is much higher than for the H3K27me3 samples, however, for both histone marks, the ChIP-seq experiments are more similar to each other than to the other ChIP or to the input. In fact, the signals of H3K27ac and H3K27me3 are almost not correlated at all which supports the notion that their

biological function is also quite opposing.



computeGCbias

This tool computes the GC bias using the method proposed by Benjamini and Speed.

What it does

The basic assumption of the GC bias diagnosis is that an ideal sample should show a uniform distribution of sequenced reads across the genome, i.e. all regions of the genome should have similar numbers of reads, regardless of their base-pair composition. In reality, the DNA polymerases used for PCR-based amplifications during the library preparation of the sequencing protocols prefer GC-rich regions. This will influence the outcome of the sequencing as there will be more reads for GC-rich regions just because of the DNA polymerase's preference.

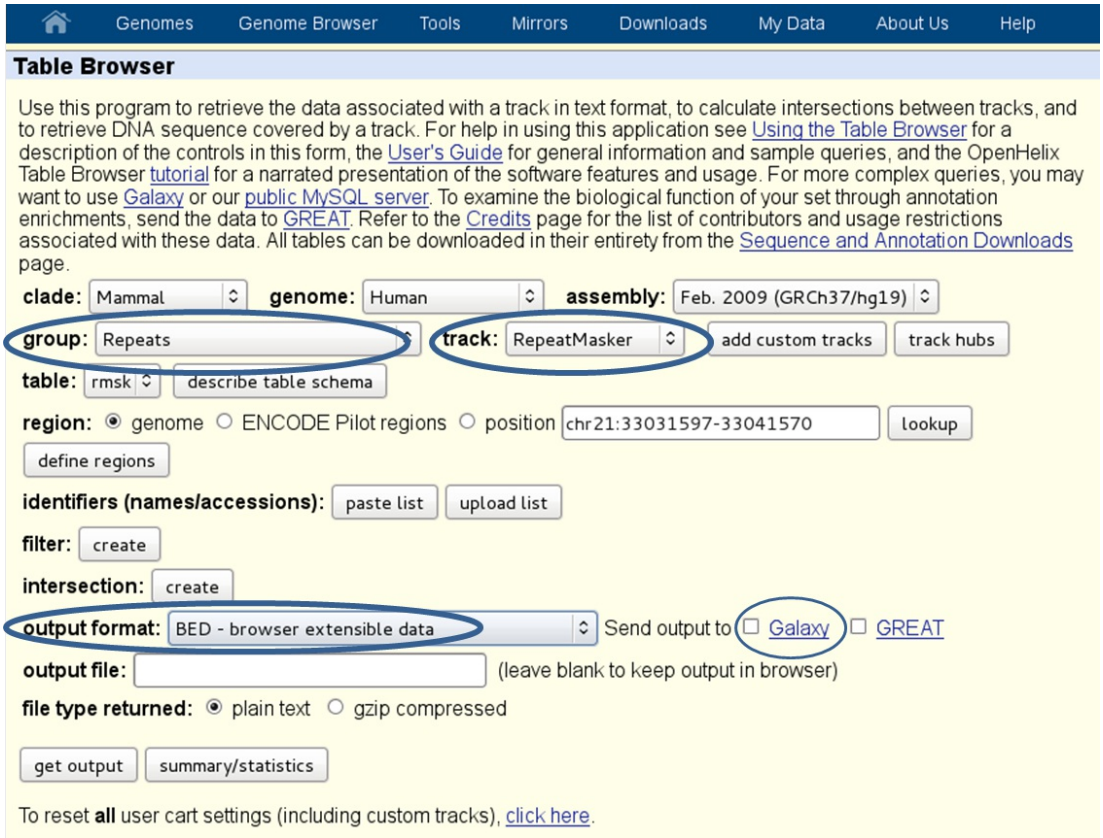
computeGCbias will **first calculate the *expected* GC profile** by counting the number of DNA fragments of a fixed size per GC fraction (GC fraction is defined as the number of G's or C's in a genome region of a given length)(a). This profile is then **compared to the *observed* GC profile** by counting the number of sequenced reads per GC fraction.

(a) The expected GC profile depends on the reference genome as different organisms have very different GC contents. For example, one would expect more fragments with GC fractions between 30% to 60% in mouse samples (GC content of the mouse genome: 45 %) than for genome fragments from Plasmodium falciparum (genome GC content P. falciparum: 20%).

Excluding regions from the read distribution calculation

In some cases, it will make sense to exclude certain regions from the calculation of the read distributions to increase the accuracy of the computation. There are several kinds of regions that are either not expected to show a background read distribution or where the uncertainty of the reference genome might be too big. Please consider the following points:

- **repetitive regions**: if multi-reads (reads that map to more than one genomic position) were excluded from the BAM file, it will help to exclude known repetitive regions. You can get BED files of known repetitive regions from UCSC Table Browser (see the screenshot below for an example of human repetitive elements).



- **regions of low mappability**: these are regions where the mapping of the reads notoriously fails and we recommend to exclude known regions with mappability issues from the GC computation. You can download the mappability tracks for different read lengths from UCSC, e.g. for mouse and human. In the github deepTools folder "scripts", you can find a shell script called *mappabilityBigWig_to_unmappableBed.sh* which will turn the bigWig mappability file from UCSC into a BED file.

- **ChIP-seq peaks**: in ChIP-seq samples it is *expected* that certain regions *should* show more reads than expected based on the background distribution, therefore it makes absolute sense to exclude those regions from the GC bias calculation. We recommend to run a simple, non-conservative peak calling on the uncorrected BAM file first to obtain a BED file of peak regions that should then subsequently be supplied to computeGCbias.

Output files

- **Diagnostic plot**
  - box plot of *absolute* read numbers per genomic GC fraction
  - x-y plot of *observed/expected* read ratios per genomic GC fraction (ideally, ratio should always be 1 (log2(1) = 0))
- **Data matrix**
  - tabular matrix file
  - to be used for GC correction with *correctGCbias*

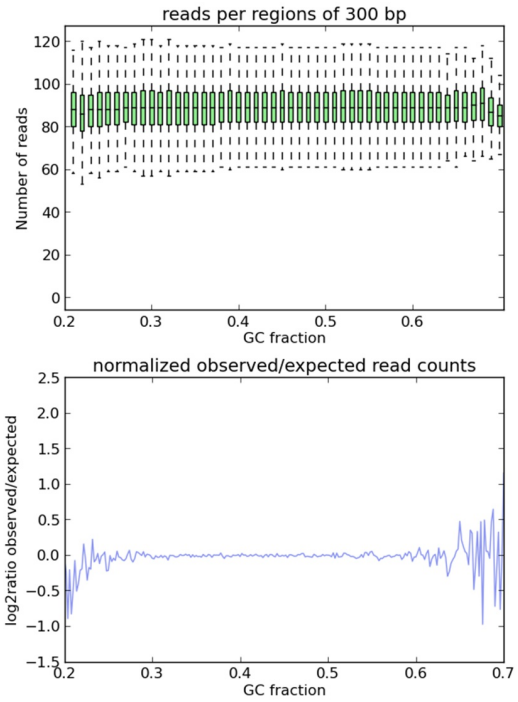
What the plots tell you

In an ideal sample without GC bias, the ratio of observed/expected values should be close to 1 for all GC content bins.

However, due to PCR (over)amplifications, the majority of ChIP samples usually shows a significant bias towards reads with high GC content (>50%) and a depletion of reads from GC-poor regions.

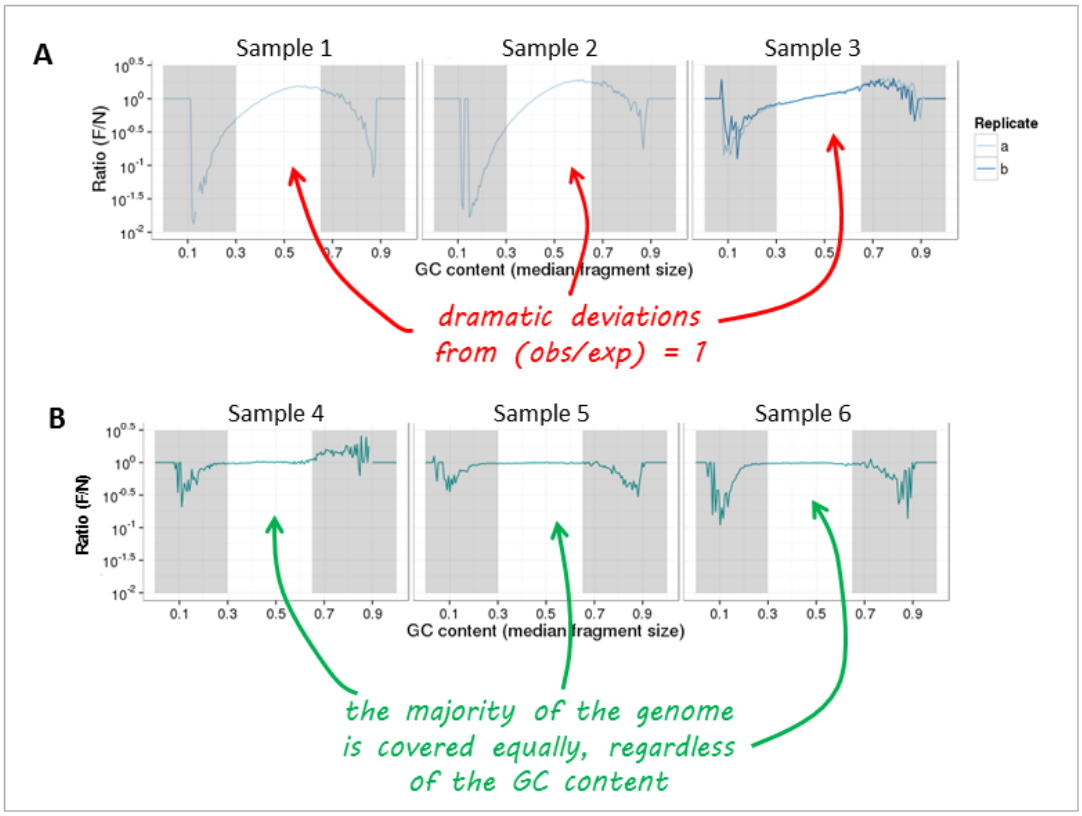
Example figures

Let's start with an ideal case. The following plots were generated with computeGCbias using simulated reads from the *Drosophila* genome.



As you can see, both plots based on **simulated reads** do not show enrichments or depletions for specific GC content bins, there is an almost flat line at log2ratio of 0 (= ratio of 1). The fluctuations on the ends of the x axis are due to the fact that only very, very few regions in the genome have such extreme GC fractions so that the number of fragments that are picked up in the random sampling can vary.

Now, let's have a look at **real-life data** from genomic DNA sequencing. Panels A and B can be clearly distinguished and the major change that took place between the experiments underlying the plots was that the samples in panel A were prepared with too many PCR cycles and a standard polymerase whereas the samples of panel B were subjected to very few rounds of amplification using a high fidelity DNA polymerase.



bamFingerprint

This quality control will most likely be of interest for you if you are dealing with ChIP-seq samples as a pressing question in ChIP-seq experiments is "Did my ChIP work?", i.e. did the antibody-treatment enrich sufficiently so that the ChIP signal can be separated from the background signal? (After all, around 90 % of all DNA fragments in a ChIP experiment will represent the genomic background). We use bamFingerprint routinely to monitor the outcome of ChIP-seq experiments.

What it does

This tool is based on a method developed by Diaz et al. and it determines how well the signal in the ChIP-seq sample can be differentiated from the background distribution of reads in the control sample. For factors that will enrich well-defined, rather narrow regions (e.g. transcription factors such as p300), the resulting plot can be used to assess the strength of a ChIP, but the broader the enrichments are to be expected, the less clear the plot will be. Vice versa, if you do not know what kind of signal to expect, the bamFingerprint plot will give you a straight-forward indication of how careful you will have to be during your downstream analyses to separate biological noise from meaningful signal.

The tool first samples indexed BAM files and counts all reads overlapping a window (bin) of specified length. These counts are then sorted according to their rank and the cumulative sum of read counts is plotted.

Output files:

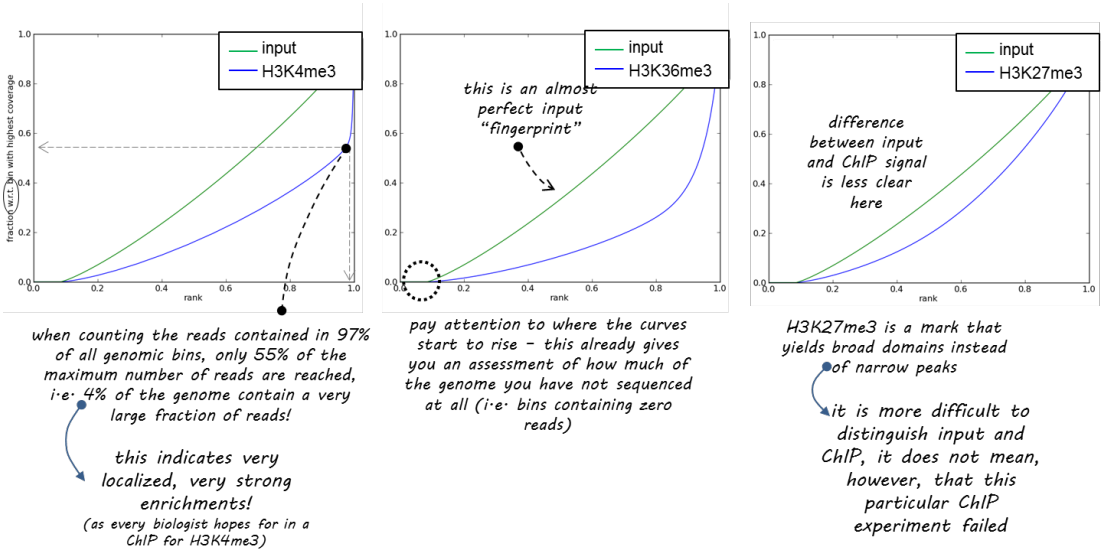
- **Diagnostic plot**
- **Data matrix** of raw counts (optional)

What the plots tell you

An ideal input with perfect uniform distribution of reads along the genome (i.e. without enrichments in open chromatin etc.) should generate a straight diagonal line. A very specific and strong ChIP enrichment will be indicated by a prominent and steep rise of the cumulative sum towards the highest rank. This means that a big chunk
of reads from the ChIP sample is located in few bins which corresponds to high, narrow enrichments seen for transcription factors.

Example figures

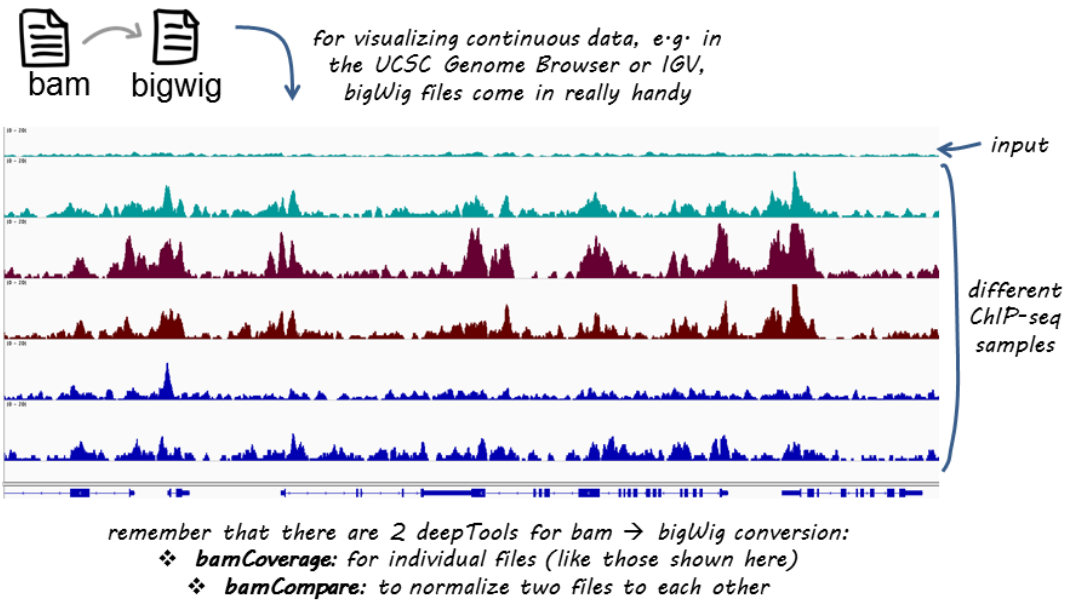
Here you see 3 different fingerprint plots.
We chose these examples to show you how the nature of the ChIP signal (narrow and high vs. wide and not extremely high) is reflected in the "fingerprint" plots. Please note that these plots go by the name of "fingerprints" in our facility because we feel that they help us tremendously in judging individual files, but the idea underlying these plots came from Diaz et al.



Normalization of BAM files

deepTools contains 3 tools for the normalization of BAM files:

1. **correctGCbias**: if you would like to normalize your read distributions to fit the expected GC values, you can use the output from computeGCbias and produce a GC-corrected BAM-file.
2. **bamCoverage**: this tool converts a *single* BAM file into a bigWig file, enabling you to normalize for sequencing depth.
3. **bamCompare**: like bamCoverage, this tool produces a normalized bigWig file, but it takes 2 BAM files, normalizes them for sequencing depth and subsequently performs a mathematical operation of your choice, i.e. it can output the ratio of the read coverages in both files or the like.



correctGCbias

What it does

This tool requires the **output from computeGCBias** to correct a given BAM file according to the method proposed by Benjamini and Speed.

correctGCbias will remove reads from regions with too high coverage compared to the expected values (typically GC-rich regions) and will add reads to regions where too few reads are seen (typically AT-rich regions).

The resulting BAM file can be used in any downstream analyses, but **be aware that you should not filter out duplicates from here on** (duplicate removal would eliminate those reads that were added to reach the expected number of reads for GC-depleted regions).

output

- **GC-normalized BAM file**

bamCoverage

What it does

Given a BAM file, this tool generates a bigWig or bedGraph file of fragment or read coverages. The way the method works is by first calculating all the number of reads (either extended to match the fragment length or not) that overlap each bin in the

genome. Bins with zero counts are skipped, i.e. not added to the output file. The resulting read counts can be normalized using either a given scaling factor, the RPKM formula or to get a 1x depth of coverage (RPGC).

- RPKM:
  - reads per kilobase per million reads
  - The formula is: RPKM (per bin) = number of reads per bin / (number of mapped reads (in millions) * bin length (kp))
- RPGC:
  - reads per genomic content
  - used to normalize reads to 1x depth of coverage
  - sequencing depth is defined as: (total number of mapped reads * fragment length) / effective genome size

output

- **coverage file** either in bigWig or bedGraph format

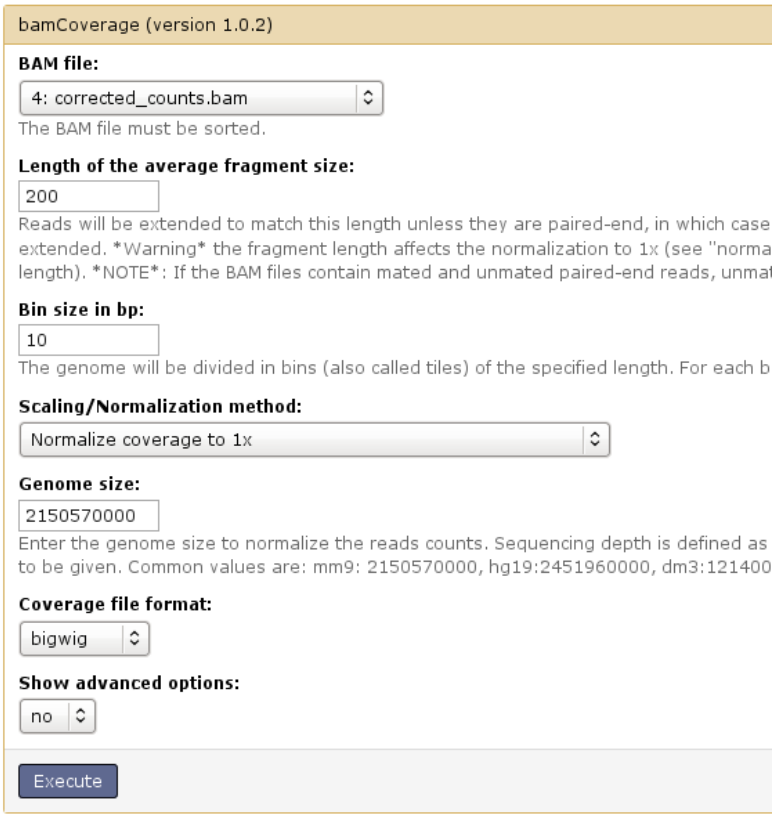
Usage

Here's an exemplary command to generate a single bigWig file out of a single BAM file via the command line:

```
$/deepTools-1.5/bin/bamCoverage --bam corrected_counts.bam \
--binSize 10 --normalizeTo1x 2150570000 --fragmentLength 200 \
-o Coverage.GCcorrected.SeqDepthNorm.bw --ignoreForNormalization chrX
```

- The bin size **(-bs)** can be chosen completely to your liking. The smaller it is, the bigger your file will be.
- This was a mouse sample, therefore the effective genome size for mouse had to be indicated once it was decided that the file should be normalize to 1x coverage.
- Chromosome X was excluded from sampling the regions for normalization as the sample was from a male mouse that therefore contained pairs of autosome, but only a single X chromosome.
- The fragment length of 200 bp is only the fall-back option of bamCoverage as the sample provided here was done with paired-end sequencing. Only in case of singletons will bamCoverage resort to the user-specified fragment length.
- --ignoreDuplicates - important! in case where you normalized for GC bias using correctGCbias, you should absolutely **NOT** set this parameter

Using deepTools Galaxy, this is what you would have done (pay attention to the hints on the command line as well!):



bamCompare

What it does

This tool compares **two** BAM files based on the number of mapped reads. To compare the BAM files, the genome is partitioned into bins of equal size, the reads are counted for each bin and each BAM file and finally, a summarizing value is reported.This value can be the ratio of the number of reads per bin, the log2 of the ratio or the difference. This tool can normalize the number of reads on each BAM file using the SES method proposed by Diaz et al. Normalization based on read counts is also available. If paired-end reads are present, the fragment length reported in the BAM file is used by default.

output file

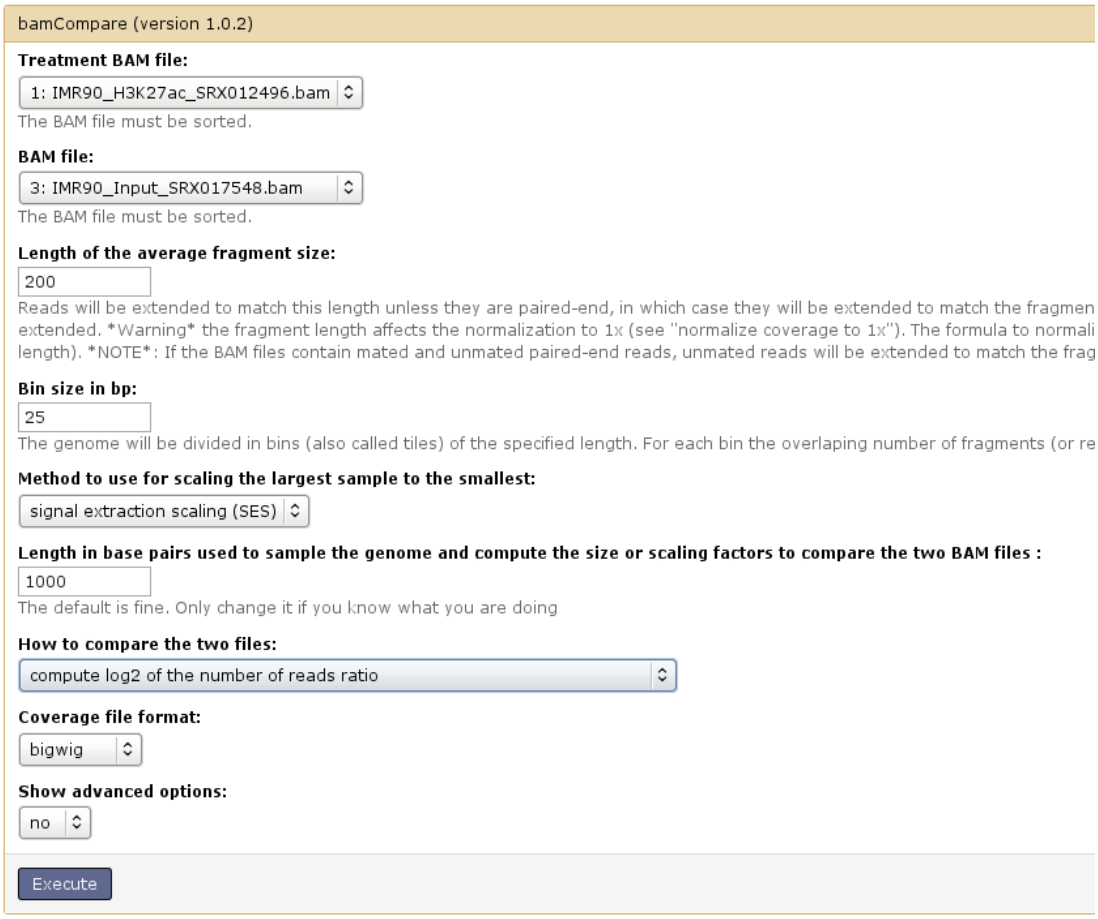
- same as for bamCoverage, except that you now obtain **1** coverage file that is based on **2** BAM files.

Usage

Here's an example command that generated the log2(ChIP/Input) values via the command line.

```
$ /deepTools-1.5/bin/bamCompare --bamfile1 ChIP.bam -bamfile2 Input.bam \
--binSize 25 --fragmentLength 200 --missingDataAsZero no \
--ratio log2 --scaleFactorsMethod SES -o log2ratio_ChIP_vs_Input.bw
```

The Galaxy equivalent:



Note that the option "missing Data As Zero" can be found within the "advanced options" (default: no).

- like for bamCoverage, the bin size is completely up to the user
- the fragment size (-f) will only be taken into consideration for reads without mates
- the SES method (see below) was used for normalization as the ChIP sample was done for a histone mark with highly localized enrichments (similar to the left-most plot of the bamFingerprint-examples)

Some (more) parameters to pay special attention to

--scaleFactorsMethod (in Galaxy: "Method to use for scaling the largest sample to the smallest")

Here, you can choose how you would like to normalize to account for variation in sequencing depths. We provide:

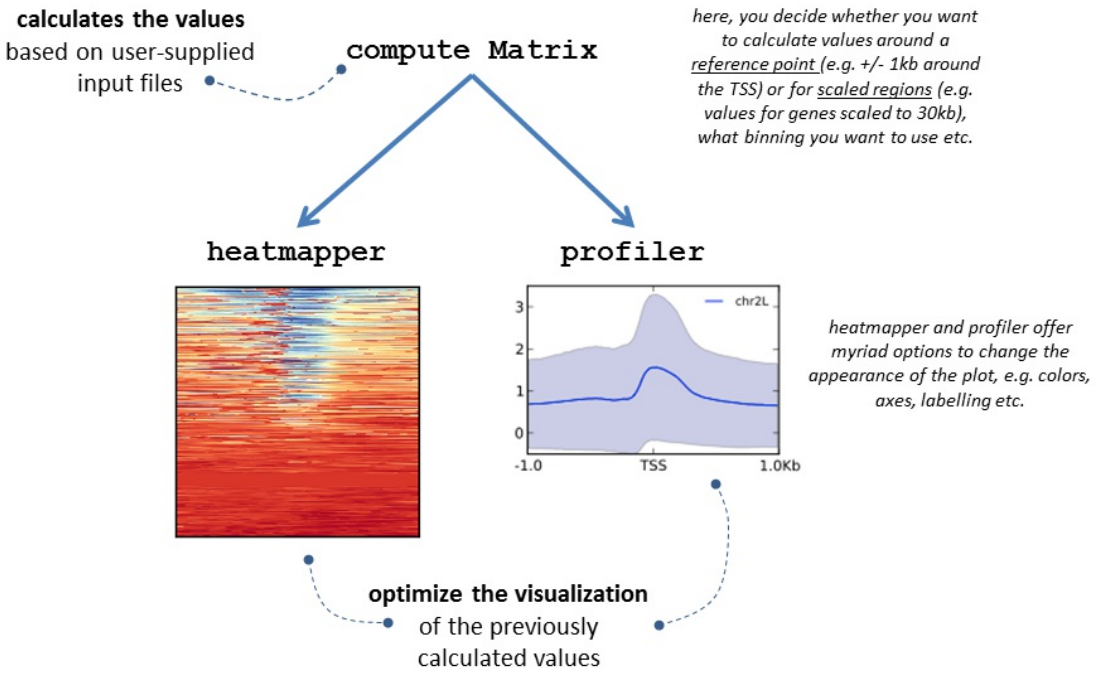
- the simple normalization **total read count**
- the more sophisticated signal extraction (SES) method proposed by Diaz et al. for the normalization of ChIP-seq samples. **We recommend to use SES only for those cases where the distinction between input and ChIP is very clear in the bamFingerprint plots**. This is usually the case for transcription factors and sharply defined histone marks such as H3K4me3.

--ratio (in Galaxy: "How to compare the two files")

Here, you get to choose how you want the two input files to be compared, e.g. by taking the ratio or by subtracting the second BAM file from the first BAM file etc. In case you do want to subtract one sample from the other, you will have to choose whether you want to normalize to 1x coverage (--normalizeTo1x) or to **r**eads **p**er **k**ilobase (--normalizeUsingRPKM; similar to RNA-seq normalization schemes).

Visualization

The modules for visualizing scores contained in bigWig files are separated into 1 tool that calculates the values (*computeMatrix*) and 2 tools that contain many, many options to fine-tune the plots (*heatmapper* and *profiler*). In other words: computeMatrix generates the values that are the basis for heatmapper and profiler.



computeMatrix

This tool summarizes and prepares an intermediary file containing scores associated with genomic regions that can be used afterwards to plot a heatmap or a profile.

Genomic regions can really be anything - genes, parts of genes, ChIP-seq peaks, favorite genome regions... as long as you provide a proper file in BED or INTERVAL format. This tool can also be used to filter and sort regions according to their score.

As indicated in the plot above, computeMatrix can be run with either one of the two modes: **scaled regions** or **reference point**.

Please see the example figures down below for explanations of parameters and options.

Output files

- **obligatory**: zipped matrix of values to be used with *heatmapper* and/or *profiler*
- **optional** (can also be generated with heatmapper or profiler in case you forgot to produce them in the beginning):
  - BED-file of the regions sorted according to the calculated values
  - list of average values per genomic bin
  - matrix of values per genomic bin per genomic interval

heatmapper

The heatmapper depicts values extracted from the bigWig file for each genomic region individually.
It requires the output from computeMatrix and most of its options are related to tweaking the visualization only. The values calculated by computeMatrix are not changed.

Definitely check the example at the bottom of the page to get a feeling for how many things you can tune.

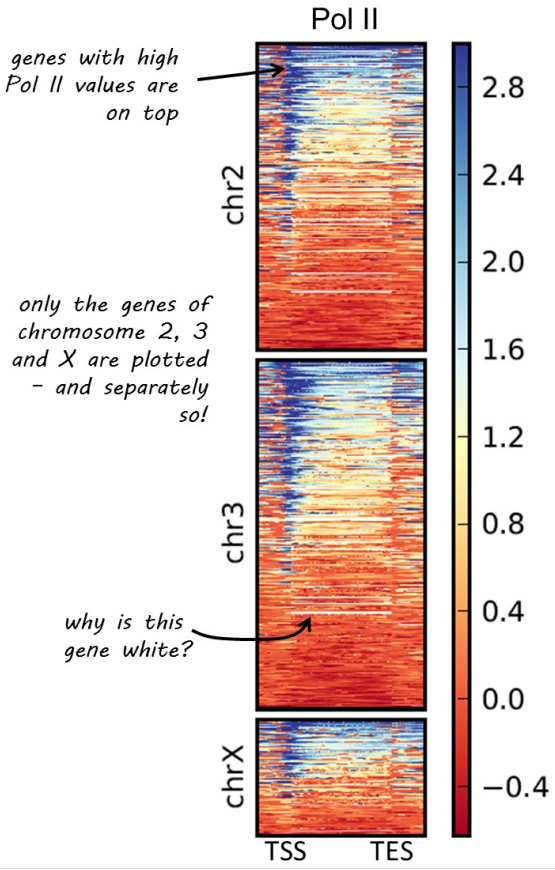
profiler

This tool plots the average enrichments over all genomic regions supplied to computeMarix. It is a very useful complement to the heatmapper, especially in cases when you want to compare the scores for many different groups. Like heatmapper, profiler does not change the values that were compute by computeMatrix, but you can choose between many different ways to color and display the plots.

Example figures

Here you see a typical, not too pretty example of a heatmap. We will use this example to explain several features of computeMatrix and heatmapper, so do take a closer look.

1st example: Heatmap with all genes scaled to the one size and user-specified groups of genes



As you can see, all genes have been scaled to the same size and the (mean) values per bin size (10 bp) are colored accordingly. In addition to the gene bodies, we added 500 bp up- and down-stream of the genes.

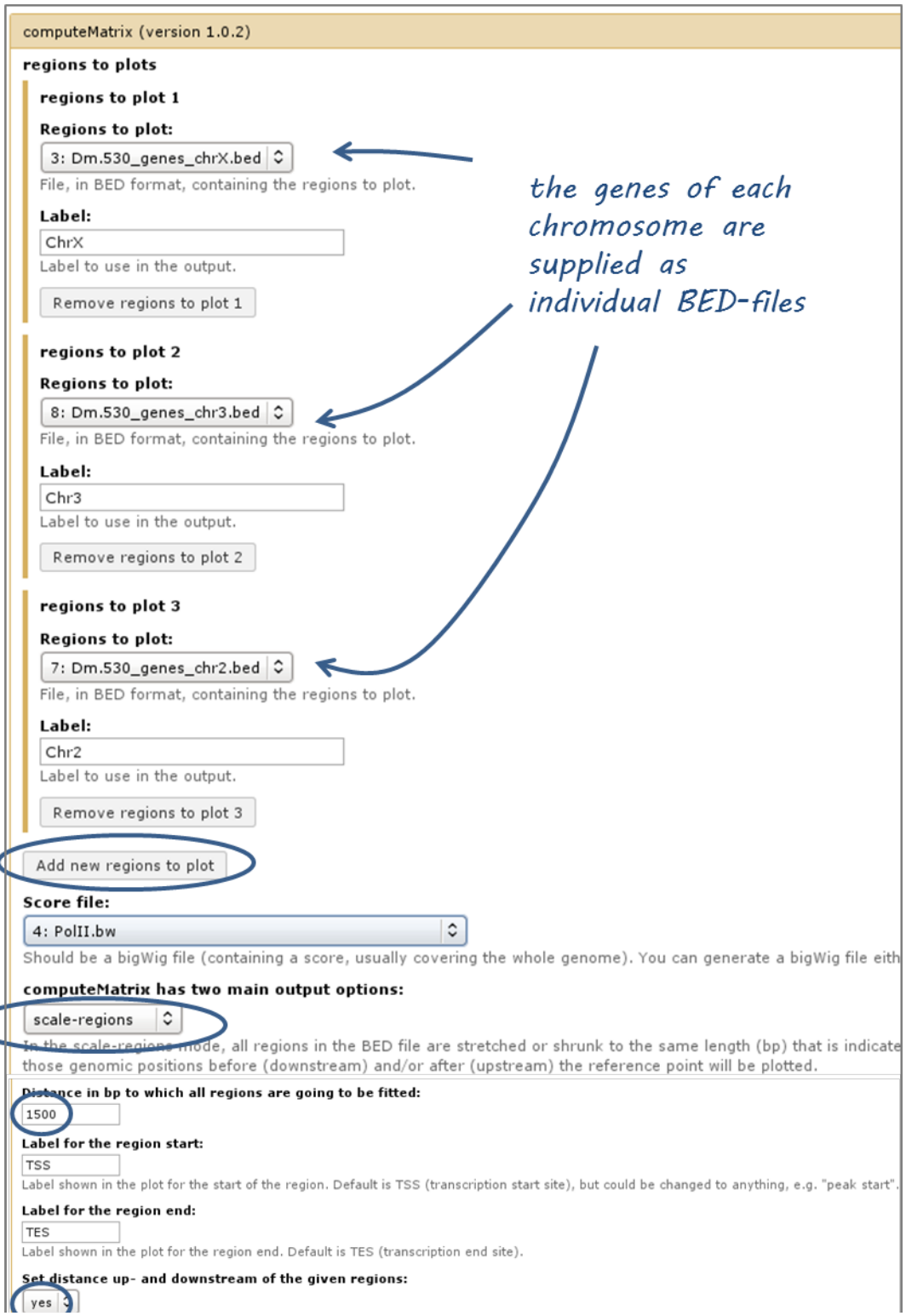
The plot was produced with the following commands:

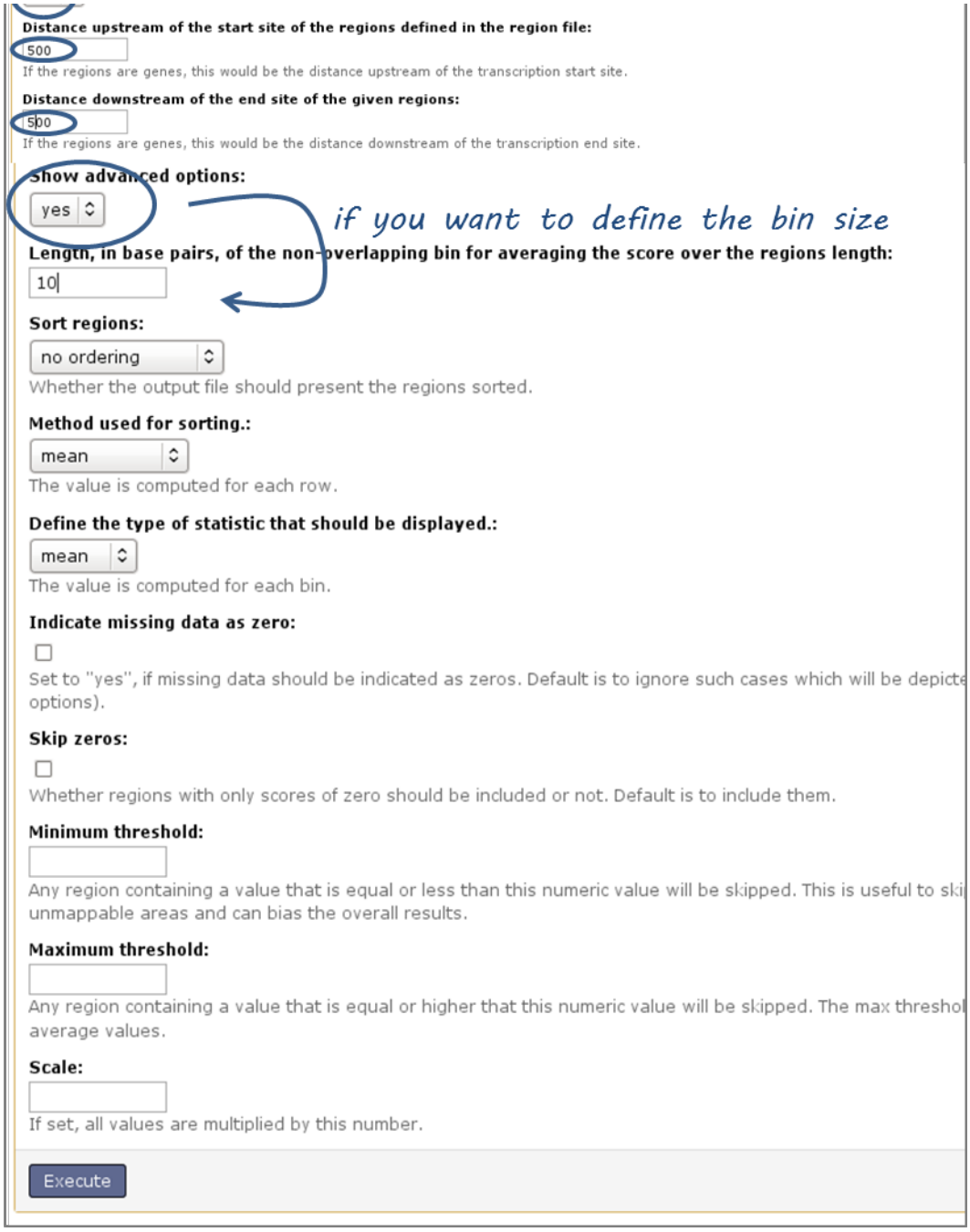
```
$ /deepTools-1.5.2/bin/computeMatrix scale-regions --regionsFileName Dm.genes.indChromLabeled.bed \
--scoreFileName PolII.bw --beforeRegionStartLength 500 --afterRegionStartLength 500 \
--regionBodyLength 1500 --binSize 10 \
--outFileName PolII_matrix_scaledGenes --sortRegions no

$ /deepTools-1.5.2/bin/heatmapper --matrixFile PolII_matrix_scaledGenes \
```

```
--outFileName PolII_indChr_scaledGenes.pdf \
--plotTitle "Pol II" --whatToShow "heatmap and colorbar"
```

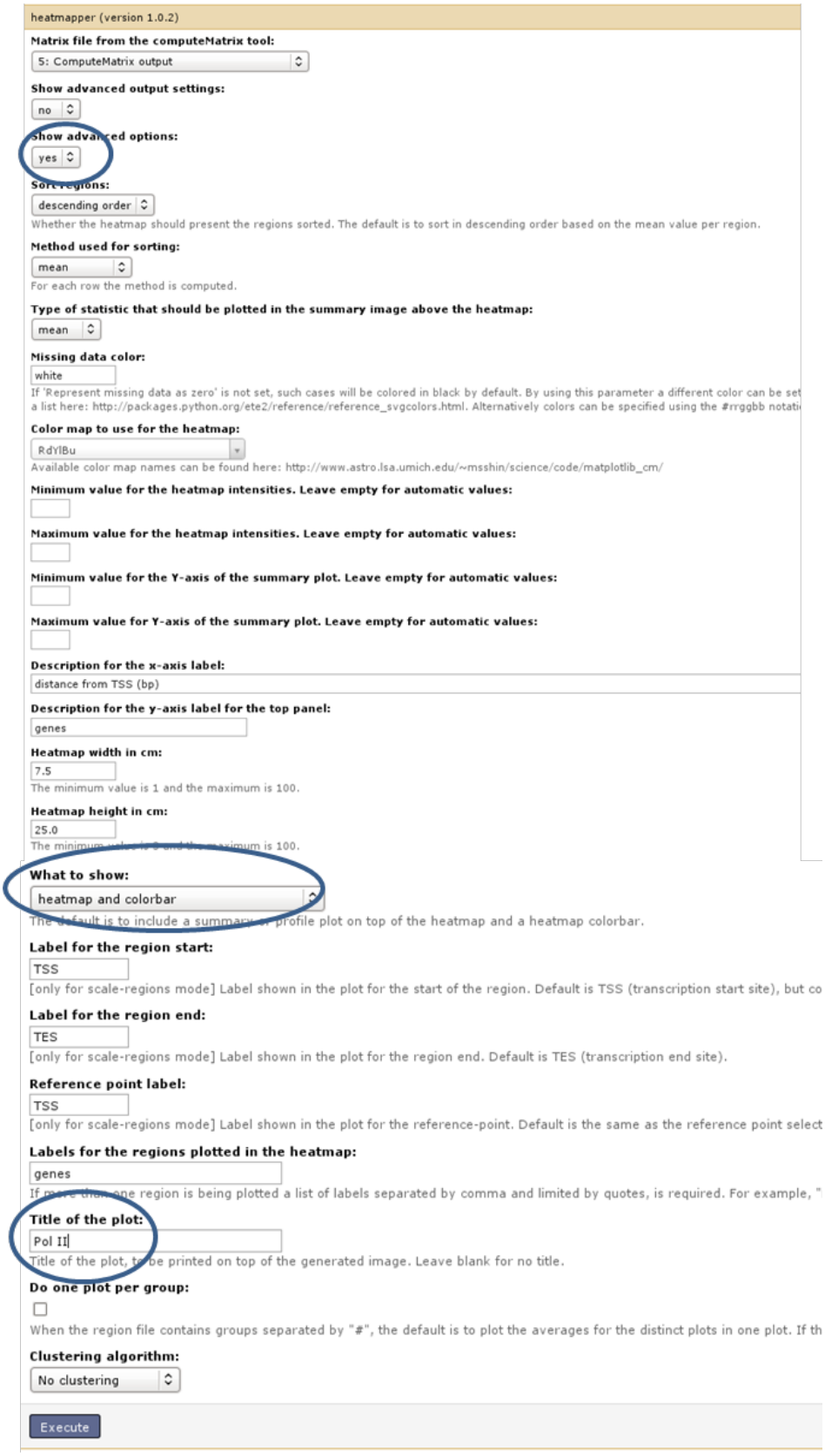
This is what you would have to select to achieve the same result within Galaxy (pay attention to the fact that you will have to use two tools, computeMatrix and heatmapper):

computeMatrix




Distance upstream of the start site of the regions defined in the region file:
500
If the regions are genes, this would be the distance upstream of the transcription start site.
Distance downstream of the end site of the given regions:
500
If the regions are genes, this would be the distance downstream of the transcription end site.
Show advanced options:
yes
if you want to define the bin size
Length, in base pairs, of the non-overlapping bin for averaging the score over the regions length:
10
Sort regions:
no ordering
Whether the output file should present the regions sorted.
Method used for sorting.:
mean
The value is computed for each row.
Define the type of statistic that should be displayed.:
mean
The value is computed for each bin.
Indicate missing data as zero:
Set to "yes", if missing data should be indicated as zeros. Default is to ignore such cases which will be depicte
options).
Skip zeros:
Whether regions with only scores of zero should be included or not. Default is to include them.
Minimum threshold:
Any region containing a value that is equal or less than this numeric value will be skipped. This is useful to ski
unmappable areas and can bias the overall results.
Maximum threshold:
Any region containing a value that is equal or higher that this numeric value will be skipped. The max thresho
average values.
Scale:
If set, all values are multiplied by this number.
Execute

heatmapper



The main difference between computeMatrix usage on the command line and Galaxy: the input of the regions file (BED)

Note that we supplied just *one* BED-file via the command line whereas in Galaxy we indicated three different files (one per chromosome).

On the command line, the program expects a BED file where different groups of genomic regions are concatenated into one file, where the beginning of each group should be indicated by "#group name".
The BED-file that was used here, contained 3 such lines and could be prepared as follows:

```
$ grep ^chr2 AllGenes.bed > Dm.genes.indChromLabeled.bed
$ echo "#chr2" >> Dm.genes.indChromLabeled.bed
$ grep ^chr3 AllGenes.bed >> Dm.genes.indChromLabeled.bed
$ echo "#chr3" >> Dm.genes.indChromLabeled.bed
$ grep ^chrX AllGenes.bed >> Dm.genes.indChromLabeled.bed
$ echo "#chrX" >> Dm.genes.indChromLabeled.bed
```

In Galaxy, you can simply generate three different data sets starting from a whole genome list of *Drosophila melanogaster* genes by using the "Filter" tool ("Filter and Sort" --> "Filter") on the entries in the first column three times:

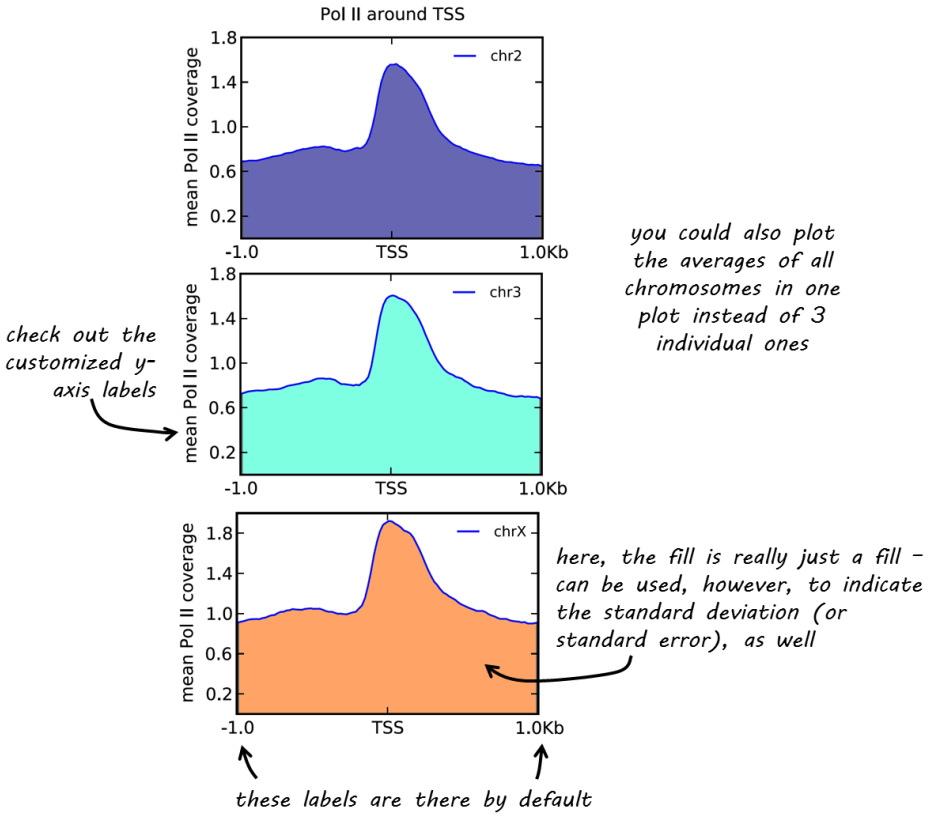
1. c1=="chr2" --> Dm.genes.chr2.bed
2. c1=="chr3" --> Dm.genes.chr3.bed
3. c1=="chrX" --> Dm.genes.chrX.bed

Important parameters for optimizing the visualization

1. **sorting of the regions**: The default of heatmapper is to sort the values in descending order. You can change that to ascending, no sorting at all or according to the size of the region (Using the `--sort` option on the command line or advanced options in Galaxy). We strongly recommend to leave the sorting option at "no sorting" for the initial computeMatrix step.
2. **coloring**: The default coloring by heatmapper is done using the python color map "RdYlBu", but this can be changed (--colorMap on the command line, advanced options within Galaxy).
3. **dealing with missing data**: You have certainly noticed that some gene bodies are depicted as white lines within the otherwise colorful mass of genes. Those regions are due to genes that, for whatever reason, did not have any read coverage in the bigWig file. There are several ways to handle these cases:
    - **--skipZeros** this is useful when your data actually has a quite nice coverage, but there are 2 or 3 regions where you deliberately filtered out reads or you don't expect any coverage (e.g. hardly mapable regions). This will only work if the entire region does not contain a single value.
    - **--missingDataAsZero** this option allows computeMatrix do interpret missing data points as zeroes. Be aware of the changes to the average values that this might cause.
    - **--missingDataColor** this is in case you have very sparse data or were missing values make sense (e.g. when plotting methylated CpGs - half the genome should have no value). This option then allows you to pick out your favorite color for those regions. The default is black (was white when the above shown image was produced).

2nd example: Summary plots with all genes scaled to the one size and user-specified groups of genes

Here's the **profiler** plot corresponding to the heatmap above. There's one major difference though - do you spot it?



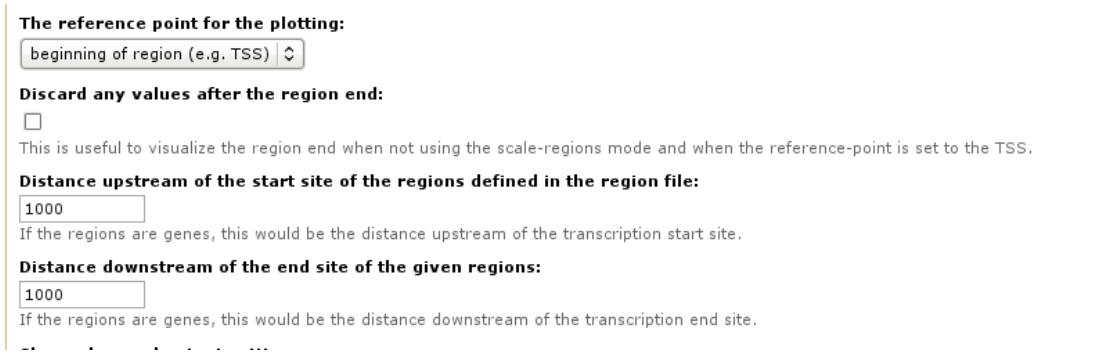
We used the same BED file(s) as for the heatmap, hence the 3 different groups (1 per chromosome). However, this time we used computeMatrix not with *scale-regions* but with *reference-point* mode.

```
$ /deepTools-1.5.2/bin/computeMatrix reference-point --referencePoint TSS \
--regionsFileName Dm.genes.indChromLabeled.bed --scoreFileName PolII.bw \
--beforeRegionStartLength 1000 --afterRegionStartLength 1000 \
--binSize 10 --outFileName PolII_matrix_indChr_refPoint \
--missingDataAsZero --sortRegions no

$ /deepTools-1.5.2/bin/profiler --matrixFile PolII_matrix_indChr_refPoint \
--outFileName profile_PolII_indChr_refPoint.pdf
--plotType fill --startLabel "TSS" \
--plotTitle "Pol II around TSS" --yAxisLabel "mean Pol II coverage" \
--onePlotPerGroup
```

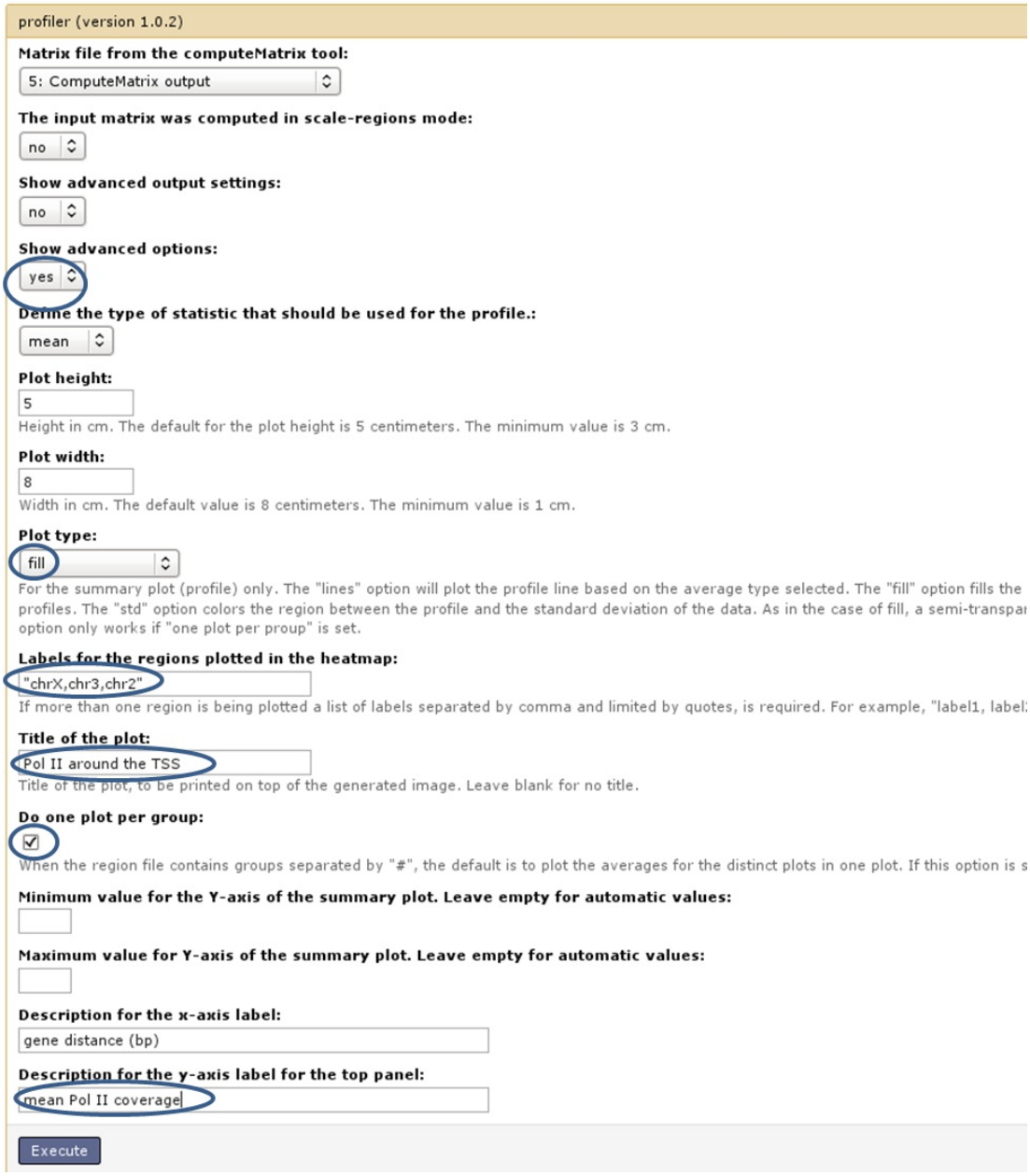
When you compare the profiler commands with the heatmapper commands, you also notice that we made use of many more labeling options here, e.g. `--yAxisLabel` and a more specific title via `-T`

This is how you would have obtained this plot in Galaxy (only the part that's *different* from the above shown command for the scale-regions version is shown):

computeMatrix


The reference point for the plotting:
beginning of region (e.g. TSS)
Discard any values after the region end:
This is useful to visualize the region end when not using the scale-regions mode and when the reference-point is set to the TSS.
Distance upstream of the start site of the regions defined in the region file:
1000
If the regions are genes, this would be the distance upstream of the transcription start site.
Distance downstream of the end site of the given regions:
1000
If the regions are genes, this would be the distance downstream of the transcription end site.

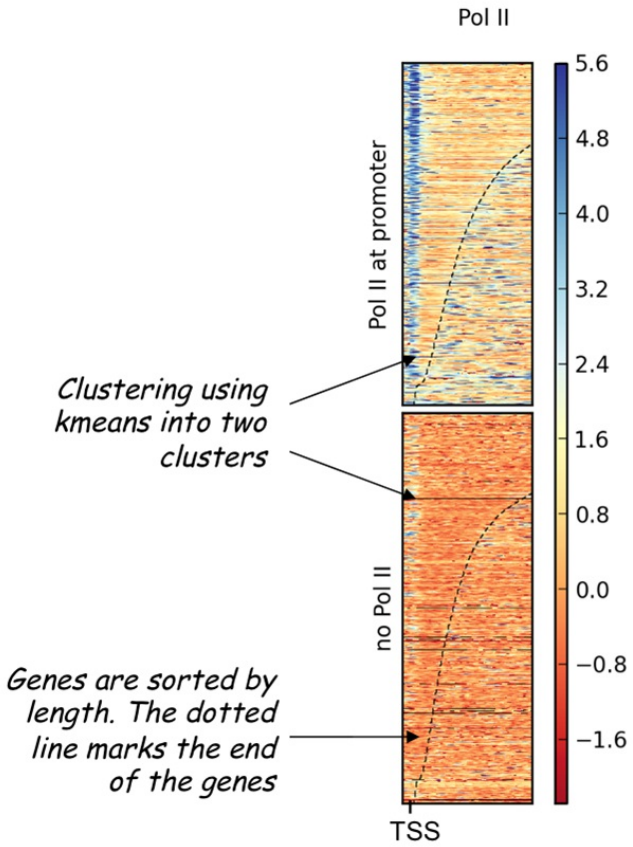
profiler


profiler (version 1.0.2)
Matrix file from the computeMatrix tool:
5: ComputeMatrix output
The input matrix was computed in scale-regions mode:
no
Show advanced output settings:
no
Show advanced options:
yes
Define the type of statistic that should be used for the profile.:
mean
Plot height:
5
Height in cm. The default for the plot height is 5 centimeters. The minimum value is 3 cm.
Plot width:
8
Width in cm. The default value is 8 centimeters. The minimum value is 1 cm.
Plot type:
fill
For the summary plot (profile) only. The "lines" option will plot the profile line based on the average type selected. The "fill" option fills the
profiles. The "std" option colors the region between the profile and the standard deviation of the data. As in the case of fill, a semi-transpar
option only works if "one plot per proup" is set.
Labels for the regions plotted in the heatmap:
"chrX,chr3,chr2"
If more than one region is being plotted a list of labels separated by comma and limited by quotes, is required. For example, "label1, label
Title of the plot:
Pol II around the TSS
Title of the plot, to be printed on top of the generated image. Leave blank for no title.
Do one plot per group:
When the region file contains groups separated by "#", the default is to plot the averages for the distinct plots in one plot. If this option is s
Minimum value for the Y-axis of the summary plot. Leave empty for automatic values:
Maximum value for Y-axis of the summary plot. Leave empty for automatic values:
Description for the x-axis label:
gene distance (bp)
Description for the y-axis label for the top panel:
mean Pol II coverage
Execute

3rd example: Heatmap with all genes scaled to the one size and kmeans clustering

Instead of supplying groups of regions on your own, you can use the clustering function of heatmapper to get a first impression whether the signal of your experiment can be easily clustered into two or more groups of similar signal distribution.

Have a look at this example with two clusters. The values correspond to log2ratios(ChIP/input) from a ChIP-seq experiment for RNA Polymerase II in *Drosophila melanogaster*:

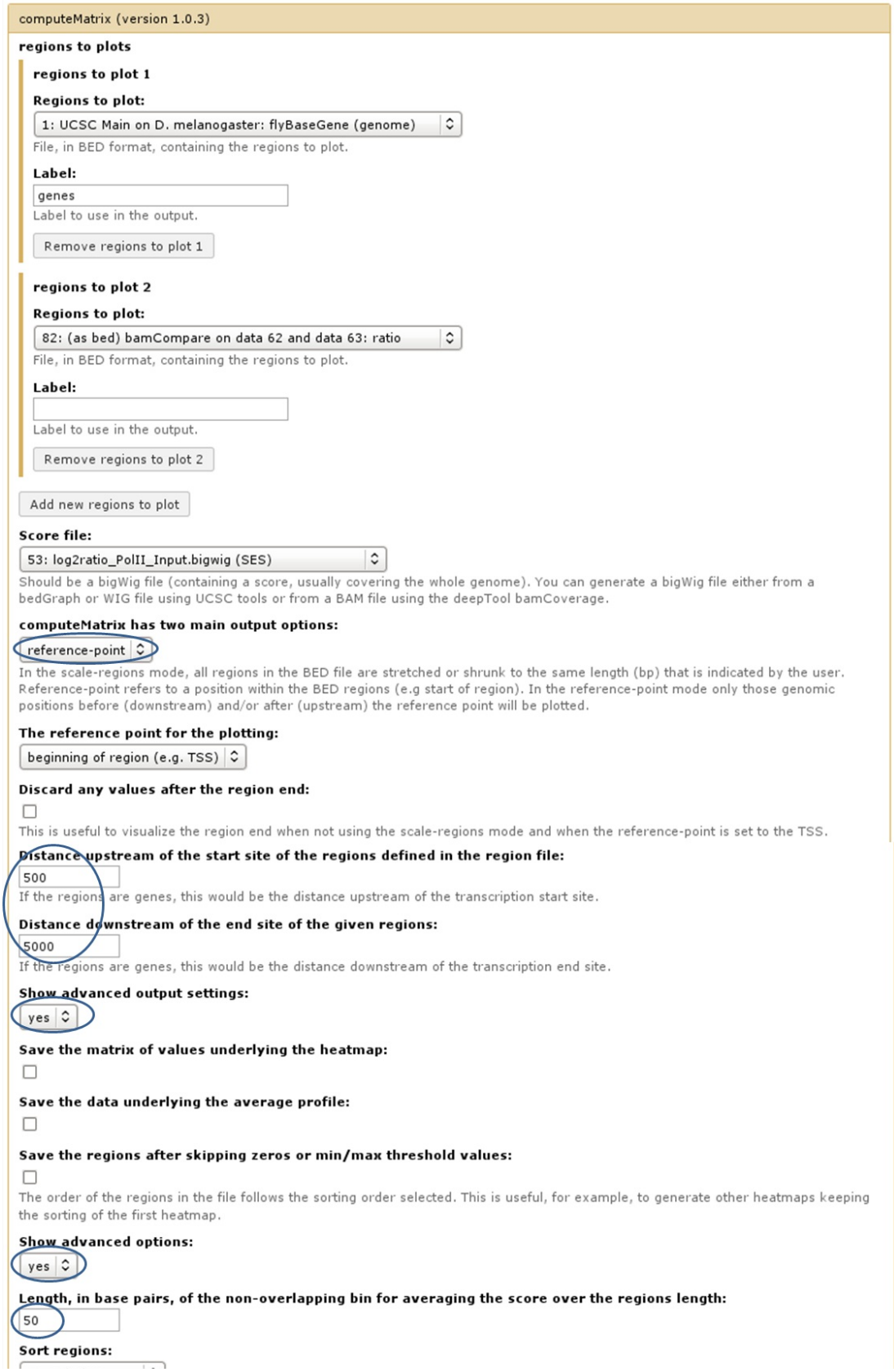


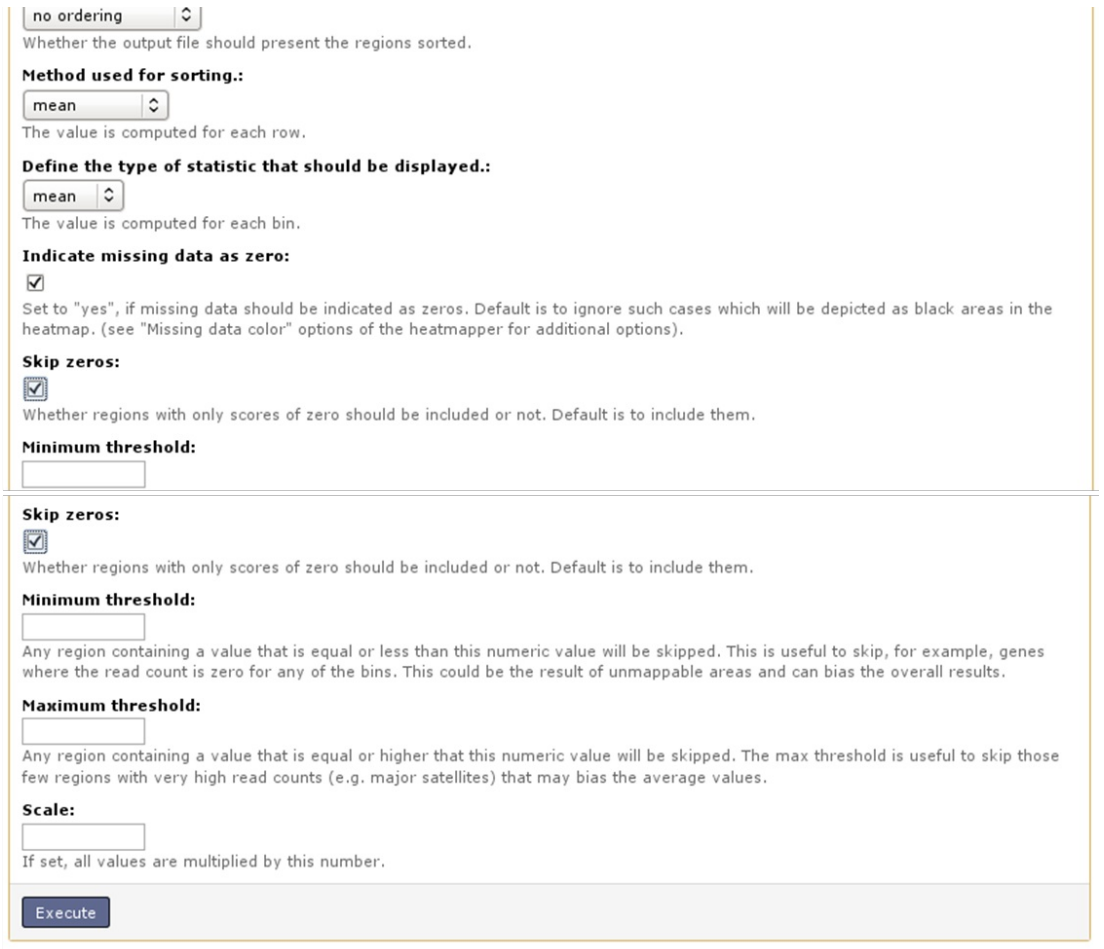
The plot was produced with the following commands:

```
$ /deepTools-1.5.2/bin/computeMatrix reference-point \
--regionsFilenName Dm.genes.indChromLabeled.bed \
--scoreFileName PolII.bw \
--beforeRegionStartLength 500 --afterRegionStartLength 5000 \
--binSize 50 \
--outFileName PolII_matrix_TSS

$ /deepTools-1.5.2/bin/heatmapper --matrixFile PolII_matrix_TSS \
--kmeans 2 \
--sortUsing region_length \
--outFileName PolII_two_clusters.pdf \
--plotTitle "Pol II" --whatToShow "heatmap and colorbar"
```

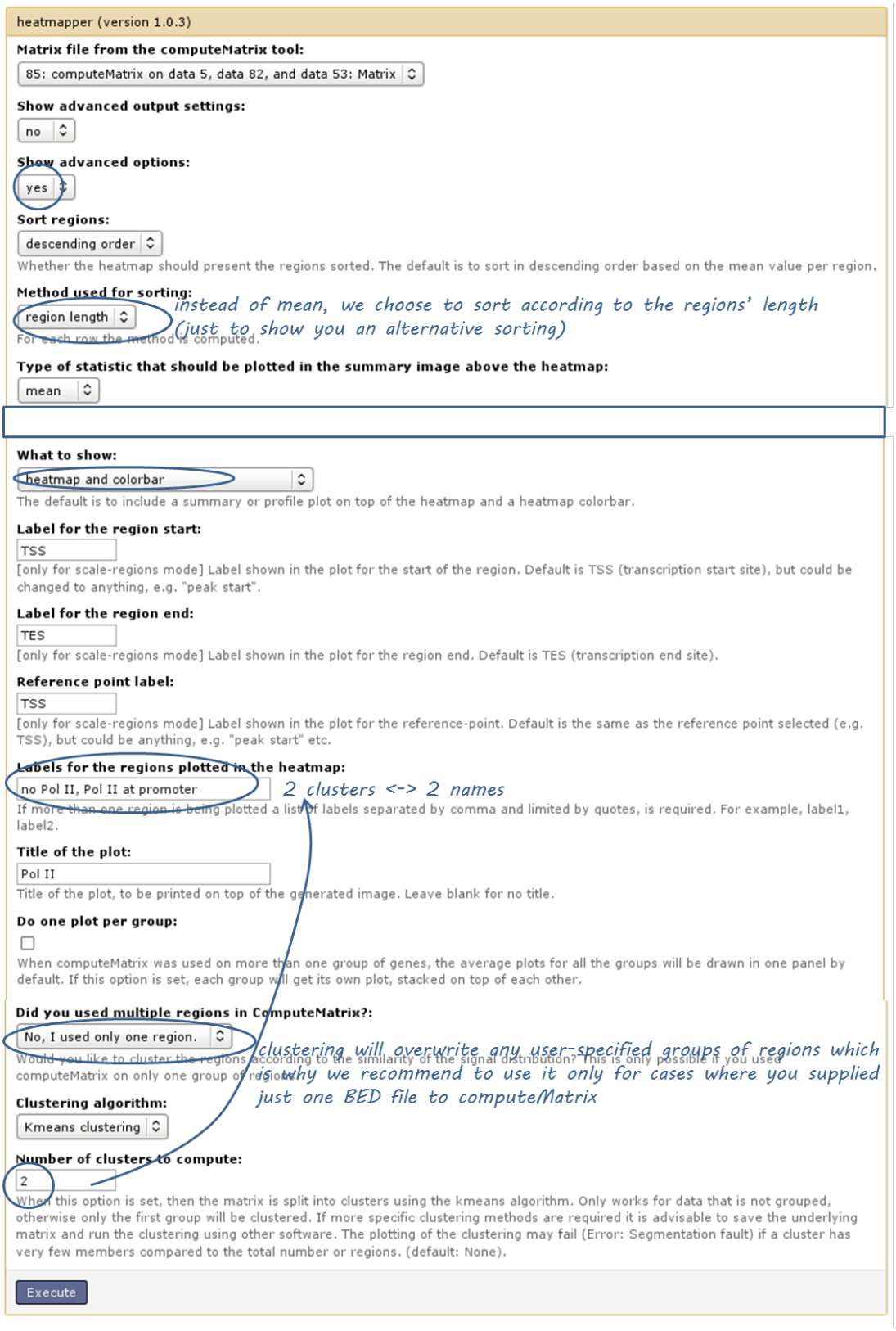
In Galaxy, these are the screenshots from the commands for computeMatrix and heatmapper:

computeMatrix




no ordering
Whether the output file should present the regions sorted.
Method used for sorting.:
mean
The value is computed for each row.
Define the type of statistic that should be displayed.:
mean
The value is computed for each bin.
Indicate missing data as zero:
Set to "yes", if missing data should be indicated as zeros. Default is to ignore such cases which will be depicted as black areas in the heatmap. (see "Missing data color" options of the heatmapper for additional options).
Skip zeros:
Whether regions with only scores of zero should be included or not. Default is to include them.
Minimum threshold:
Skip zeros:
Whether regions with only scores of zero should be included or not. Default is to include them.
Minimum threshold:
Any region containing a value that is equal or less than this numeric value will be skipped. This is useful to skip, for example, genes where the read count is zero for any of the bins. This could be the result of unmappable areas and can bias the overall results.
Maximum threshold:
Any region containing a value that is equal or higher that this numeric value will be skipped. The max threshold is useful to skip those few regions with very high read counts (e.g. major satellites) that may bias the average values.
Scale:
If set, all values are multiplied by this number.
Execute

heatmapper


heatmapper (version 1.0.3)
Matrix file from the computeMatrix tool:
85: computeMatrix on data 5, data 82, and data 53: Matrix
Show advanced output settings:
no
Show advanced options:
yes
Sort regions:
descending order
Whether the heatmap should present the regions sorted. The default is to sort in descending order based on the mean value per region.
Method used for sorting:
region length
instead of mean, we choose to sort according to the regions' length (just to show you an alternative sorting)
For each row the method is computed.
Type of statistic that should be plotted in the summary image above the heatmap:
mean
What to show:
heatmap and colorbar
The default is to include a summary or profile plot on top of the heatmap and a heatmap colorbar.
Label for the region start:
TSS
[only for scale-regions mode] Label shown in the plot for the start of the region. Default is TSS (transcription start site), but could be changed to anything, e.g. "peak start".
Label for the region end:
TES
[only for scale-regions mode] Label shown in the plot for the region end. Default is TES (transcription end site).
Reference point label:
TSS
[only for scale-regions mode] Label shown in the plot for the reference-point. Default is the same as the reference point selected (e.g. TSS), but could be anything, e.g. "peak start" etc.
Labels for the regions plotted in the heatmap:
no Pol II, Pol II at promoter
2 clusters <-> 2 names
If more than one region is being plotted a list of labels separated by comma and limited by quotes, is required. For example, label1, label2.
Title of the plot:
Pol II
Title of the plot, to be printed on top of the generated image. Leave blank for no title.
Do one plot per group:
When computeMatrix was used on more than one group of genes, the average plots for all the groups will be drawn in one panel by default. If this option is set, each group will get its own plot, stacked on top of each other.
Did you used multiple regions in ComputeMatrix?:
No, I used only one region.
clustering will overwrite any user-specified groups of regions which is why we recommend to use it only for cases where you supplied just one BED file to computeMatrix
Would you like to cluster the regions according to the similarity of the signal distribution? This is only possible if you used computeMatrix on only one group of regions.
Clustering algorithm:
Kmeans clustering
Number of clusters to compute:
2
When this option is set, then the matrix is split into clusters using the kmeans algorithm. Only works for data that is not grouped, otherwise only the first group will be clustered. If more specific clustering methods are required it is advisable to save the underlying matrix and run the clustering using other software. The plotting of the clustering may fail (Error: Segmentation fault) if a cluster has very few members compared to the total number or regions. (default: None).
Execute

When the `--kmeans` option is chosen and more than 0 clusters are specified, heatmapper will run the k-means clustering algorithm. In this example, we wanted to divide *Drosophila melanogaster* genes into two clusters. As you can see above, the algorithm nicely identified two groups - one with mostly those genes with lots of Pol II at the promoter region (top) from those genes without PolI II at the promoter (bottom).
Please note that the clustering will only work if the initial BED-file used with computeMatrix contained only *one* group of genes.

The genes belonging to each cluster can be obtained by via `--outFileSortedRegions` on the command line and "advanced output options in Galaxy". On the command line, this will result in a BED file where the groups are separated by a hash tag. In Galaxy, you will obtain individual data sets per cluster.

To have a better control on the clustering it is recommended to load the matrix raw data into **specialized software like cluster3 or R**. You can obtain the matrix via the option `--outFileNameMatrix` on the command line and by the "advanced output options" in Galaxy. The order of the rows is the same as in the output of the `--outFileSortedRegions` BED file.

Glossary

Like most specialized fields, next-generation sequencing has inspired many an acronym. We are trying to keep track of those abbreviations that we heavily use. Do make us aware if something is unclear: deeptools@googlegroups.com

If you are unfamiliar with the file formats of next-generation sequencing data, do have a look on the next page.

Abbreviations

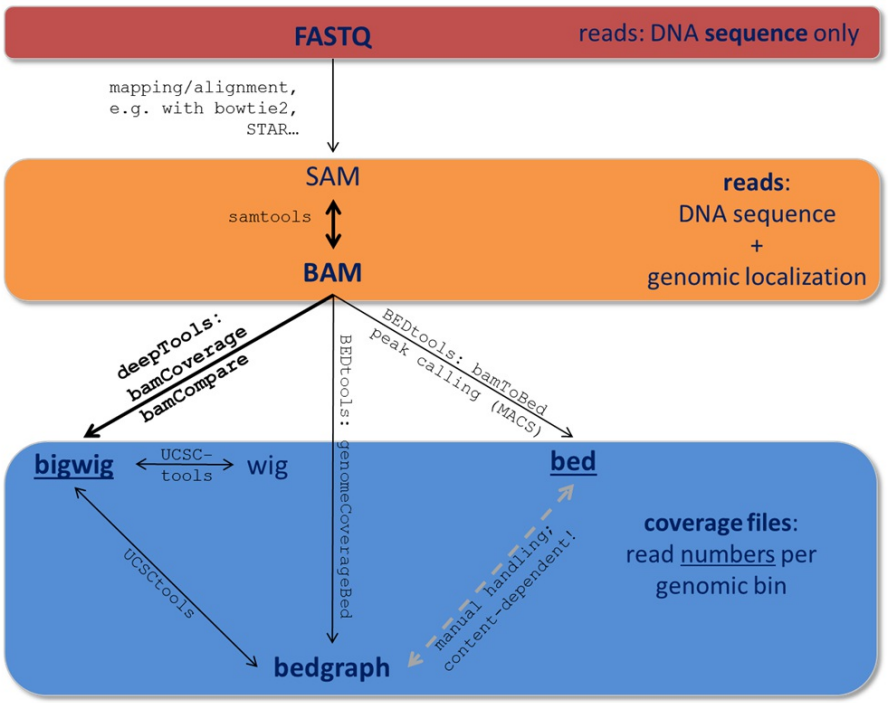
Acronym	full phrase	Synonyms/Explanation
-seq	-sequencing	indicates that an experiment was completed by DNA sequencing using NGS
ChIP-seq	chromatin immunoprecipitation sequencing	NGS technique for detecting transcription factor binding sites and histone modifications (see entry "Input" for more information)
DNase	deoxyribonuclease	micrococcal nuclease
HTS	high-throughput sequencing	next-generation sequencing, massive parallel short read sequencing, deep sequencing
Input	--	control experiment typically done for ChIP-seq experiments (see above) - while ChIP-seq relies on antibodies to enrich for DNA fragments bound to a certain protein, the input sample should be processed exactly the same way, excluding the antibody. This way, one hopes to account for biases introduced by the sample handling and the general chromatin structure of the cells
MNase	micrococcal nuclease	DNase
NGS	next-generation sequencing	high-throughput (DNA) sequencing, massive parallel short read sequencing, deep sequencing
RPGC	reads per genomic content	used to normalize read numbers (also: normalize to 1x sequencing depth), sequencing depth is defined as: (total number of mapped reads * fragment length) / effective genome size.
RPKM	reads per kilobase per million reads	used to normalize read numbers, the following formula is used by bamCoverage: RPKM (per bin) = number of reads per bin / (number of mapped reads (in millions) * bin length (kb))

File Formats

Data obtained from next-generation sequencing data must be processed several times. Most of the processing steps are aimed at extracting only those information that are truly needed for a specific down-stream analysis and to discard all the redundant entries. Therefore, **specific data formats are often associated with different steps of a data processing pipeline**. These associations, however, are by no means binding, but you should understand what kind of data is represented in which data format - this will help you to select the correct tools further down the road.

Here, we just want to give very brief key descriptions of the file, for elaborate information we will link to external websites (links to be found in our online wiki). Be aware, that the file name sorting here is purely alphabetically, not according to their usage within an analysis pipeline that is depicted here.
For more information on the different tool collections mentioned in the figure, please check the following links:

- deepTools wiki: http://github.com/fidelram/deepTools/wiki
- samtools: http://samtools.sourceforge.net/http://samtools.sourceforge.net/
- UCSCtools download: http://hgdownload.cse.ucsc.edu/admin/exe/
- BEDtools: http://bedtools.readthedocs.org/en/latest/



2bit

- compressed, binary version of genome sequences that are often stored in [FASTA][]
- most genomes in 2bit format can be found at UCSC]
- [FASTA][] files can be converted to 2bit using the UCSC programm **faToTwoBit** available for different platforms at UCSC
- more information can be found here or from UCSC

BAM

- typical file extension: .bam

- *binary* file format (complement to SAM)
- contains information about sequenced reads *after alignment* to a reference genome
- each line = 1 mapped read, with information about:
  - its mapping quality (how certain is the read alignment to this particular genome locus?)
  - its sequencing quality (how well was each base pair detected during sequencing?)
  - its DNA sequence
  - its location in the genome
  - etc.
- highly recommended format for storing data
- to make a BAM file human-readable, one can, for example, use the program samtools view (from UCSC tools)
- for more information, see below for the definition of SAM files

bed

- typical file extension: .bed
- text file
- used for genomic intervals, e.g. genes, peak regions etc.
- actually, there is a rather strict definition of the format that can be found at UCSC
- for deepTools, the first 3 columns are important: chromosome, start position of the region, end position of the genome
- do not confuse it with the bedGraph format (eventhough they are quite similar)
- example lines from a BED file of mouse genes (note that the start position is 0-based, the end-position 1-based, following UCSC conventions for BED files):

```
chr1    3204562 3661579 NM_001011874    Xkr4    -
chr1    4481008 4486494 NM_011441   Sox17   -
chr1    4763278 4775807 NM_001177658    Mrpl15  -
chr1    4797973 4836816 NM_008866   Lypla1  +
```

bedGraph

- typical file extension: .bg, .bedgraph
- text file
- similar to BED file (not the same!), it can ONLY contain 4 columns and 4th column MUST be a score
- again, read the UCSC description for more details
- 4 exemplary lines from a bedGraph file (like BED files following the UCSC convention, the start position is 0-based, the end-position 1-based in bedGraph files):

```
chr1 10 20 1.5
chr1 20 30 1.7
chr1 30 40 2.0
chr1 40 50 1.8
```

bigWig

- typical file extension: .bw, .bigwig
- *binary* version of a bedGraph file
- usually contains 4 columns: chromosome, start of genomic bin, end of genomic bin, score
- the score can be anything, e.g. an average read coverage
- UCSC description for more details

FASTA

- typical file extension: .fasta
- text file, often gzipped (--> .fasta.gz)
- very simple format for **DNA/RNA** or **protein** sequences, this can be anything from small pieces of DNA or proteins to entire genome information (most likely, you will get the genome sequence of your organism of interest in fasta format)
- see the 2bit file format entry for a compressed alternative of the fasta format
- example from wikipedia showing exactly one sequence:

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
```

```
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

FASTQ

- typical file extension: .fastq, fq
- text file, often gzipped (--> .fastq.gz)
- contains raw read information (e.g. base calls, sequencing quality measures etc.), but not information about where in the genome the read originated from
- example from the wikipedia page

```
A FASTQ file containing a single sequence might look like this:
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65

The character '!' represents the lowest quality while '~' is the highest.
```

SAM

- typical file extension: .sam
- should be the result of an alignment of sequenced reads to a reference genome
- each line = 1 mapped read, with information about its mapping quality, its sequence, its location in the genome etc.
- it is recommended to generate the binary (compressed) version of this file format: BAM
- for more information, see the SAM specification
- two exemplary lines
  - each one corresponds to one read (named r001 and r002 here)
  - the different columns contain various information about each read, e.g. which chromosome they were mapped to (here: chr1) and the left-most mapping position in the genome (here: 7 and 9 on chr1); the *flag* in the second column summarizes multiple information about each single read at once (in hexadecimal encoding) (see below for more information on the flag)

```
r001 163 chr1 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 chr1 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
```

- the flag contains the answer to several yes/no assessments that are encoded in a single number. The questions are the following ones:
  - template having multiple segments in sequencing = Is the read part of a read pair?
  - each segment properly aligned according to the aligner = Was the read properly paired?
  - segment unmapped = Is the read unmapped?
  - next segment in the template unmapped = Is the mate unmapped?
  - reverse complemented = Did the read map to the reverse strand?
  - next segment in the template is reversed = Did the mate map to the reverse strand?
  - the first segment in the template = Is this read the first one in the pair?
  - the last segment in the template = Is this read the second one in the pair?
  - secondary alignment = Is this not the primary (i.e. unique optimal) alignment for the read?
  - not passing quality controls = Did the read not pass the quality control?
  - PCR or optical duplicate = Was this read a PCR or optical duplicate?
- for more details on the flag, see this thorough explanation or this more technical explanation

Links and references

Literature

Benjamini and Speed, Nucleic Acids Research (2012): http://nar.oxfordjournals.org/content/40/10/e72
Diaz et al., Stat. Appl. Gen. Mol. Biol. (2012): http://www.degruyter.com/view/j/sagmb.2012.11.issue-3/1544-6115.1750/1544-6115.1750.xml

For more NGS-related literature, see our collection at the deepTools web server: http://deeptools.ie-freiburg.mpg.de/u/fduendar/p/useful-bioinfo-literature

Additional bioinformatic tools

bowtie2: http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml
cluster3: http://bonsai.hgc.jp/~mdehoon/software/cluster/
IGV (Integrative Genome Browser): http://www.broadinstitute.org/igv/
k-means clustering: http://en.wikipedia.org/wiki/K-means_clustering
R: http://www.r-project.org/

File format information

SAM file specification: http://samtools.sourceforge.net/SAMv1.pdf
File formats explained at UCSC: http://genome.ucsc.edu/FAQ/FAQformat.html#format1

Fidel Ramírez, Friederike Dündar, Sarah Diehl, Björn A. Grüning, Thomas Manke

Bioinformatics Group, Max-Planck-Institute of Immunobiology and Epigenetics, Freiburg & Department of Computer Science, University of Freiburg

Web server (incl. sample data): http://deepTools.ie-freiburg.mpg.de
Code: https://github.com/fidelram/deepTools
Wiki & Tutorials: https://github.com/fidelram/deepTools/wiki

A.4 A regulatory feedback loop between MSL1 and CDK7 controls RNA polymerase II Serine 5 phosphorylation

Chlamydas, S., Holz, H., Pelechano, V., Chelmicki, T., Georgiev, P., **Dündar, F.**, Dasmeh, P., Mittler, G., Tavares Cadete, F., Ramírez, F., Conrad, T., Wei, W., Raja, S., Manke, T., Luscombe, N. M., Steinmetz, L. M., Akhtar, A. *Submitted for review.*

I generated Figure 1A and Figure S1C-E using data processed by Fidel Ramírez (*D. melanogaster*) and myself (*D. virilis, M. musculus*) and contributed to the revision of the manuscript.

Abstract:

Proper gene expression requires the coordinated interplay between transcriptional co-activators, transcription factors and the general transcription machinery. We find that MSL1, a component of the *Drosophila* dosage compensation complex, functions as a transcriptional co-activator with CDK7, a subunit of the Cdk-activating kinase (CAK) complex, which phosphorylates Ser5 of the carboxy-terminal domain of RNA polymerase II. The MSL1/CDK7 regulatory feedback loop is evolutionarily conserved and regulates both autosomal and X-linked target gene expression. MSL1 promotes the efficient recruitment and catalytic activity of CAK at core promoters. CDK7- mediated phosphorylation of MSL1 is essential for its proper targeting to specific autosomal sites and for X chromosome-enrichment of the entire MSL complex in male flies. In addition, MSL1 is required for proper mRNA 5' capping, bridging transcription with posttranscriptional regulatory events. Thus, the MSL1/CDK7 feedback loop promotes robust target gene expression genome-wide and is distinctly important for dosage compensation in *Drosophila*.

B. Supplemental Information

B.1 Supplemental material related to the MSL and NSL complexes

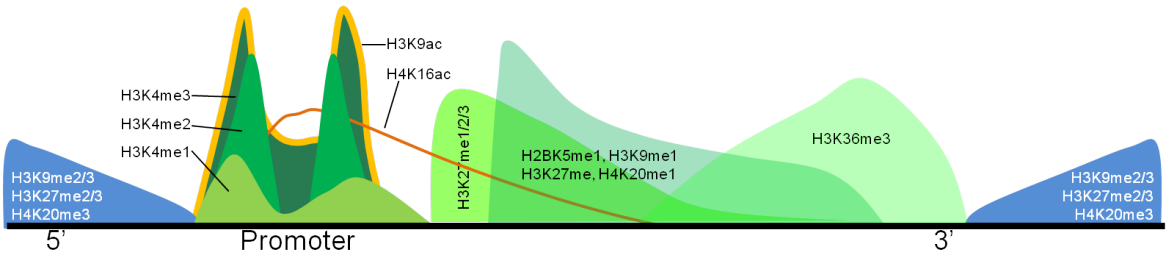


Figure B.1: Schematic patterns of selected histone marks along an active gene. Histone modifications associated with active transcription are shown in green (methylation) and orange (acetylation), repressive histone marks are shown in blue. Active promoters are strongly labelled with H3K9ac and H3K4me3, surrounded by H3K4me2/1. Inactive genes are typically covered with H3K9me2/3, H4K20me3 and H3K27me3 that also peaks around the silenced promoters. The image is based on[204]. Note that the signal of H4K16ac for dosage compensated genes in *Drosophila* inreases towards the 3'-end.

Table B.1: Protein names of MSL- and NSL complex members in *D. melanogaster*, mouse and humans. Bold font indicates the name that I used throughout most of the manuscript unless noted otherwise. MSL = male-specific lethal, NSL = non-specific lethal

Drosophila	**Mammals**
MOF (males absent on first)	MYST1, hMOF, KAT8 (lysine acetyl transferase 8)
MSL1	hMSL1
MSL2	hMSL2, Msl2l1, Rnf184 (ring finger 184)
MSL3	hMSL3, Msl31
MLE (maleless)	DHX9 (aspartic-acid-glutamine-alanine-histidin (DEAH) box helicase 9)
NSL1, waharan	hMSL1v1, Kansl1 (KAT8 regulatory NSL complex subunit 1)
NSL2, dgt1 (dimmed gamma tubulin 1)	Kansl2 (KAT8 regulatory NSL complex subunit 2)

Continued on next page

Table B.1 – *Continued from previous page*

Drosophila	**Mammals**
NSL3, Rcd1 (reduction in centrosomin dots)	Kansl3 (KAT8 regulatory NSL complex subunit 3)
MBD-R2 (methyl-binding domain protein)	PHF20 (plant homeo domain (PHD) finger protein 20), GLEA2 (glioma-expressed antigen 2)
MCRS2, dMCRS1, p78, Rcd5 (reduction in centrosomin dots)	p78, MSP58 (microspherule protein 58)
WDS (will die slowly)	WDR5 (tryptophan-aspartic acid (WD) repeat domain 5)

Table B.2: Association of MSL and NSL complex members with human cancers. Note that I used the mammalian protein names, see Table B.1 for synonyms and *Drosophila* names.

Protein	**Observation**
MYST1	• downregulated in colorectal, gastric, ovarian, hepatocellular, breast cancer and renal cell carcinoma[25,205–209], loss of H4K16ac is a hallmark of numerous cancers[24] • upregulated in human non-small-cell lung cancer[80]
MSL1	single nucleotide polymorphism in *MSL1* is associated with decreased risk of invasive serous ovarian cancer[210]
MSL3	might enhance the proliferation of hematopoetic cells in acute myeloid leukemia[211]
PHF20	• auto-antibodies against PHF20 are positively correlated with the survival rate of neuroblastoma patients[212,213] • high expression in non-small-cell lung cancer[214], genetic alteration of *PHF20* are associated with its progression[215]
MCRS1	• oncogene with the potential for malignant cell transformation[118,216] • upregulated in all investigated cancer types[217–220]
WDR5	implicated in the establishment and progression of leukemia due to its interaction with the methyl transferase MLL (mixed-lineage leukemia; reviewed by Wu and Shu[221])

Table B.3: Enzymes within the MSL and NSL complexes.

Complex	**Protein**	**Function**
MSL, NSL	MOF	histone acetyl transferase
MSL	MSL2	E3 ubiquitin ligase
	MLE	DNA and RNA helicase
NSL	NSL3	putative hydrolase function

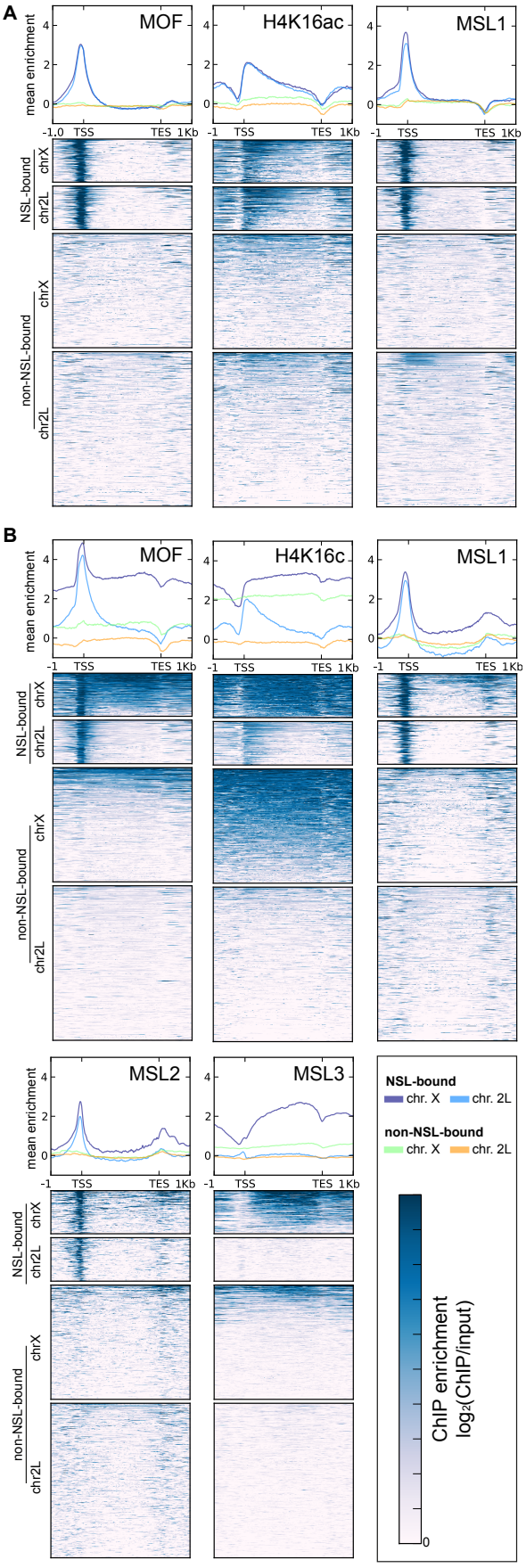


Figure B.2: ChIP-seq signals of MSL complex members for NSL targets and NSL-non-bound genes in *Drosophila*. All images were generated with `computeMatrix` and `heatmapper` from the deepTools suite[168] using the scale-regions mode to scale gene bodies to 2 kb. See Table B.8 and B.11 for details about the ChIP-seq data sets. **A)** There is no difference in the ChIP-seq signals of MOF, H4K16ac and MSL1 from female larva for autosomal (chromosome 2L, chr2L) and X-linked (chrX) genes, but there are almost exclusively strong enrichments for NSL targets. **B)** In males, the signals of the MSL complex are visibly stronger for X-linked NSL targets, but only MSL3 and H4K16ac show significant enrichments for non-NSL-bound genes.

B.2 Supplemental bioinformatics-related tables

Table B.4: Bioinformatic tools used for analyses presented here (alphabetical order; if available, I indicated the versions of the tools that I used). For explanations of the file formats mentioned here, please see the glossary within the supplement of Appendix A.3.

Name	Application	Website	Reference
bedtools (2.10. to 2.17)	working with genomic intervals, e.g. intersecting two files with different peak regions	`http://bedtools.readthedocs.org/en/latest/`	Quinlan and Hall [222]
bowtie bowtie-1.0.0 (Appendix A.1), bowtie2-2.2.2 (Appendix A.2)	alignment of reads to the reference genome	`http://bowtie-bio.sourceforge.net/index.shtml`	Langmead and Salzberg [147]
ChIPEnrich	gene ontology enrichments for target genes identified by ChIP-seq	`http://sartorlab.ccmb.med.umich.edu/chip-enrich`	Welch et al. [171]
DAVID	gene identifier mapping and gene ontology term enrichment analyses	`http://david.abcc.ncifcrf.gov/`	Huang et al. [170]
deepTools (up to 1.5.7)	quality controls of BAM files, normalizations, coverage file generation, visualizations with heatmaps and summary plots	`http://deeptools.github.io/`	Ramirez, Dündar et al. [168]
DESeq (1.10.1)	calculating normalized fold change values for Pol II ChIP-seq data set [223]	`http://www-huber.embl.de/users/anders/DESeq/`	Anders and Huber [224]
fastqc	quality control of FASTQ files	`http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`	not available
Galaxy	vast collection of tools for manifold tasks including working with genomic intervals, joining of lists, motif search etc.	in-house installation	Goecks et al. [225]
ggplot2 (0.9.3.1)	R package for highly customizable plots (e.g. boxplots, x-y plots, bar charts etc.)	`http://ggplot2.org/`	Wickham [226]
GREAT (2.0)	web-based tool for target gene prediction	`http://bejerano.stanford.edu/great/public/html/`	McLean et al. [178]
Integrative Genomics Viewer (up to 2.3.32)	genome browser for visualization of BAM, BED, bigWig and bedGraph files	`http://www.broadinstitute.org/igv/`	Thorvaldsdóttir et al. [227]
liftover (2013)	conversion of sequence coordinates from mm8 assembly to mm9	`https://genome.ucsc.edu/cgi-bin/hgLiftOver`	not available
MACS (1.4.2)	identification of significantly enriched ChIP-seq regions	`http://liulab.dfci.harvard.edu/MACS/`	Zhang et al. [144]

Continued on next page

Table B.4 – *Continued from previous page*

Name	Application	Website	Reference
MEME (4.9.0)	*de novo* DNA motif identification	`http://meme.nbcr.net/meme/`	Machanick and Bailey [169], Bailey and Elkan [228]
PeakSplitter (1.0)	splitting peak regions predicted by MACS into separate regions based on local minima detection	`http://www.ebi.ac.uk/research/bertone/software`	Salmon-Divon et al. [229]
samtools (0.1.19)	SAM and BAM file handling (indexing, number of mapped/unmapped and uniquely reads etc.)	`http://samtools.sourceforge.net/`	Li et al. [70]
TRAP (Annotate v3.04)	transcription factor binding affinity calculation	`http://trap.molgen.mpg.de/PASTAA.htm`	Roider et al. [230]
UCSC tools (2013)	format conversion (bedGraph to bigWig)	`http://www.broadinstitute.org/igv/`	Kent et al. [231]
Venny (2007)	generation of Venn diagrams	`http://bioinfogp.cnb.csic.es/tools/venny/`	not available

Table B.5: Quality metrics of ChIP-seq experiments. The sequencing is strongly influenced by the success of the sample and library preparation (see Table 2.5). The metrics shown here should guide the data processing and raise awareness for possible problems during downstream analyses. The failure of a ChIP-seq experiment to meet the recommended criteria does not immediately imply that no biologically meaningful information could be derived from it as ChIP-seq experiments with very few or very diffuse binding sites will always perform worse for genome-wide measures of signal-to-noise-ratios.

Property	Rationale	Recommendation	Limitation
		Library and sequencing quality	
Sequencing depth	measure of how many times the genome is covered with DNA reads (on average): $Coverage = reads \times fragment\ length / genome\ size$	• majority of ChIP-seq studies: coverage between 1-5[232] • should always be complemented by the number of bases with zero coverage	average coverage does not take gaps and uneven read distributions into account[232]
Overrepresented sequences	useful for the identification of: • adapter and foreign DNA contamination • suggestion of GC bias	• can be assessed with FASTQC[233] • contamination should be removed, e.g. with Trimmomatic[234] • GC bias should be re-assessed after read alignment and filtering, e.g. with deepTools[168]	• FASTQC will always work on a sample of sequences, not the full data set • contamination removal is computationally intensive
PCR bottleneck coefficient (PBC)	suggested metric for sequencing library complexity: ratio of genomic locations with *only one* aligned read to the number of loci with *at least* one aligned read[235]	• can be calculated with ENCODE tools[236] • 0-0.5: severe bias, 0.9-1.0: no bias	• the more deeply a sample is sequenced, the more likely it is that duplicated reads could correspond to different DNA fragments[130,144,237] • duplication ratios might be over-estimated for ChIP-seq samples with very localized and very strong enrichments[129]
GC bias	after read alignment, the observed read numbers per GC content can be compared to the expected values based on the reference genome sequence (Appendix A.3)	• the majority of the genome should not show significant deviation from $observed/expected = 1$[158] • deepTools can be used to computationally adjust the ratio to 1 (Appendix A.3)	• ChIP-seq of factors binding GC-rich regions • correction for lack of AT-rich sequences leads to artificial introduction of duplicate reads[158]

Continued on next page

Table B.5 – *Continued from previous page*

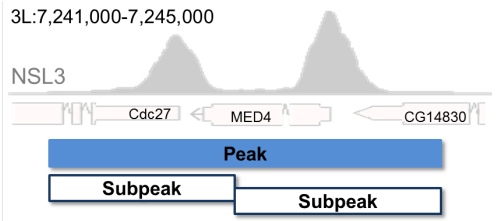
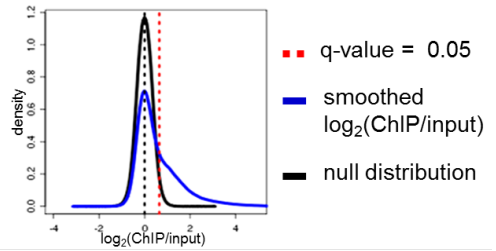
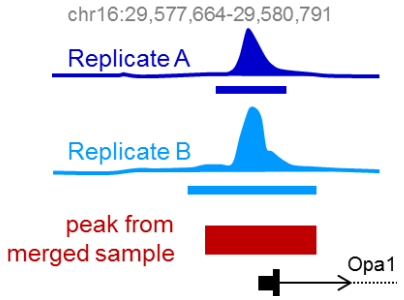
Metric	Rationale	Recommendation	Limitation
		REPRODUCIBILITY	
Correlation of read coverages	• two replicates are recommended for each ChIP-seq and input experiment[130,157] • corresponding replicates should show similar (highly correlating) read coverage distributions[150]	• correlations can be visually represented with multi-dimensional scaling plots[238] or clustered heatmaps (Appendix A.3) • replicates should cluster together	Pearson correlation is sensitive to outliers which might inflate the correlation coefficient[140]
Irreproducibility Discovery Rate (IDR)	• metric for consistency between two replicates[239] • based on the comparison of the ranks of peak regions identified in both replicates	• all measures output by the IDR package should be within a factor of two[130]	• strongly depends on the results of the peak calling step (regions as well as significance measures); requires very relaxed peak calling parameters[140,239] • results are dominated by the weakest replicate[130] • not recommended for broad enrichments[240]
		SUCCESS OF THE IMMUNOPRECIPITATION	
Standardized Standard Deviation (SSD)	describes the variation in signal depth across the genome[142]: $ssd = 1000 \times sd/\sqrt{n}$[238]	ChIP and input samples should show different SSD values[238]	sensitive to outlier regions with artificially high coverage in both sample types[142]
Cumulative percentages of read counts	• an ideal, strong ChIP-seq samples should have relatively few regions that contain a large fraction of DNA reads (= enrichments)[241]	ideally, input and ChIP-seq sample should show clearly different cumulative percentages[158,159,241] (Appendix A.3)	ChIP-seq samples with broad, domain-like enrichments are not well represented

Continued on next page

Table B.5 – *Continued from previous page*

Metric	Rationale	Recommendation	Limitation
Normalized and relative strand cross-correlation (NSC, RSC)	• DNA reads are counted separately for forward and reverse strand, then the cross-correlation (cc) is calculated for incremental distances between the strand-specific coverages[130,140,235] • $nsc = cc_{fragmentLength}/cc_{background}$ • $rsc = cc_{fragmentLength}/cc_{readLength}$	NSC $\geq$ 1.05 and RSC $\geq$ 0.8[130]	broad enrichments and factors with few binding sites will meet the suggested threshold[130,140]
Fraction of reads in peaks (FRiP)	simple proxy for the success of an IP: $frip = reads\ in\ peaks/total\ reads$[130]	FRiP $\geq$ 1%	• requires peak calling • depends strongly on the nature of the ChIP-seq signal (width, strength, number of peak regions) and should therefore not be used to compare ChIP-seq experiments for different proteins of interest[130]

Table B.6: Peak calling strategies adjusted for the different ChIP-seq sample characteristics.

ChIP-seq sample	Issues	Strategy	
NSL1, NSL3, MCRS2, MBD-R2 in *D. melanogaster*	peak regions often contained more than one local maximum due to the gene-dense nature of the *Drosophila* genome	1. peak calling with MACS 1.4.2 with default parameters[165] 2. peaks with FDR $\leq$ 5% 3. identification of smaller *subpeaks* with PeakSplitter[229]	
Pol II in *D. melanogaster* cells (depleted of NSL1 or NSL3)	the mixed signal of Pol II (localized, TF-like enrichments around TSS, broad, shallow enrichments along gene bodies) was not well captured by MACS	1. a symmetric null distribution was fitted to regions with negative enrichments, i.e. where the input signal exceeded the ChIP signal[223] (black line) 2. genomic bins that deviated from the expectation (compare the blue with the black line) were determined as significantly enriched (threshold: q-value $\leq$ 0.05)	
MOF, MSL1, MSL2, NSL3, MCRS1 in mouse ESCs and NPCs	non-optimal signal-noise ratios, GC bias	1. in ESCs: adjustment of the GC bias in the input sample to match each ChIP-seq sample's GC bias 2. peak calling with MACS 1.4.2 with default parameters[165] for each ChIP-seq replicate (blue boxes) and the merged file (red box) 3. only peaks that were present in both replicates and met the FDR threshold of $\leq$ 1% were used	

B.2.1 Datasets

All analyses were enriched by the integration of previously published, publicly available high-throughput sequencing data and annotation.

Table B.7: Publicly available data bases that were used.

Name	Application	Website
ArrayExpress	up- and download of high-throughput sequencing data (only raw read files)	`http://www.ebi.ac.uk/arrayexpress/`
BioMart	mapping of gene and transcripts between different annotations for genome version mm9, e.g. Ensembl and RefSeq; download of non-protein-coding transcripts	`http://may2012.archive.ensembl.org/biomart/martview/`
FlyBase	download of gene lists and genome sequence	`http://flybase.org/`
GEO	up- and download of high-throughput sequencing data (raw and processed data)	`http://www.ncbi.nlm.nih.gov/geo/`
HomoloGene	list of orthologous genes in *Drosophila*, mice and humans	`http://www.ncbi.nlm.nih.gov/homologene`
UCSC Table Browser	RefSeq gene lists, mappability tracks, GC content tracks, mouse DNase hypersensitivity tracks	`https://genome.ucsc.edu/`

All samples generated by members of the Akhtar lab were uploaded to public repositories. Accession numbers starting with G indicate the GEO database, accession numbers starting with E indicate ArrayExpress (see Table B.7). The DNA read numbers are based on filtered reads. Read and accession numbers refer to all available replicates of each experiment.

Drosophila datasets

The subsequently listed data was used for Lam, Mühlpfordt, Vaquerizas et al.[90] (Appendix A.1) and Chlamydas et al.

Table B.8: In-house generated ChIP-seq samples from *D. melanogaster* larva and cell culture (Schneider S2 cells[242]).

Sample	Generated by	Cells/Tissue	Reads ($\times 10^6$)	Accession number
Input	Sunil J. Raja	larval salivary glands	6.1	E-MTAB-214
NSL1	Sunil J. Raja	larval salivary glands	7.6	E-MTAB-214
MCRS2	Sunil J. Raja	larval salivary glands	9.4	E-MTAB-214
Input	Ibrahim Ilik	S2 cells	21.2	E-MTAB-1085
NSL3	Kin C. Lam	S2 cells	28.3	E-MTAB-1085

Continued on next page

Table B.8 – *Continued from previous page*

Sample	Generated by	Cells/Tissue	Reads ($\times 10^6$)	Accession number
MBD-R2	Kin C. Lam	S2 cells	27.3	E-MTAB-1085
Input	Kin C. Lam	S2 cells (NSL1 RNAi)	47.1	E-MTAB-1084
Pol II (Rbp3)	Kin C. Lam	S2 cells (NSL1 RNAi)	132; 132.1	E-MTAB-1084
Input	Kin C. Lam	S2 cells (NSL3 RNAi)	57.2	E-MTAB-1084
Pol II (Rbp3)	Kin C. Lam	S2 cells (NSL3 RNAi)	120; 134.9	E-MTAB-1084
Input	Kin C. Lam	S2 cells (GFP RNAi)	50; 60	E-MTAB-1084
Pol II (Rbp3)	Kin C. Lam	S2 cells (GFP RNAi)	134	E-MTAB-1084
Input	Thomas Conrad	larval salivary glands (females)	14.8	E-MTAB-911
Input	Thomas Conrad	larval salivary glands (males)	6.6	E-MTAB-911
H4K16ac	Thomas Conrad	larval salivary glands (females)	11.8	E-MTAB-911
H4K16ac	Thomas Conrad	larval salivary glands (males)	9.4	E-MTAB-911
MOF	Thomas Conrad	larval salivary glands (females)	7.5	E-MTAB-911
MOF	Thomas Conrad	larval salivary glands (males)	9.2	E-MTAB-911
MSL1	Thomas Conrad	larval salivary glands (males)	6.7	GSM1502677
MSL1	Sunil J. Raja	larval salivary glands (females)	10.7	GSM1502675

Table B.9: In-house generated ChIP-seq samples from *D. virilis* larva that were used for Chlamydas et al. (Appendix A.4)

Sample	Generated by	Cells/Tissue	Reads ($\times 10^6$)	Accession number
Input	Sarantis Chlamydas	larval salivary glands (females)	224.3; 230	GSM1502682, GSM1502684
MSL1	Sarantis Chlamydas	larval salivary glands (males)	210; 225	GSM1502681, GSM1502683

Table B.10: ChIP-chip modENCODE data sets for which we downloaded the processed coverage files from GEO. All ChIP experiments were done in S2 cells.

Sample	Downloaded data	Accession number
H3	.bedgraph	GSM93208
H3K4me3	.bedgraph	GSE20787
H3K36me3	.bedgraph	GSE20785
H3K4me2	.bedgraph	GSE23470
H4K5ac	.bedgraph	GSE20800
H3K9ac	.bedgraph	GSE20790
H4K8ac	.bedgraph	GSE20801
H3K9me3	.bedgraph	GSE20794
H3K9me3	.bedgraph	GSE20794

Table B.11: ChIP-seq data of MSL complex members, downloaded and processed by Fidel Ramírez and used for Figure 3.3.

Sample	Accession number
MSL2	GSE37864
MSL3	GSE37864
MLE	SRX111555

Mouse datasets

The following data was used for Chelmicki, Dündar et al.[172] (Appendix A.2). MSL1 and input samples were also used for Chlamydas et al. (Appendix A.4).

Table B.12: ChIP-seq samples from murine embryonic stem cells (mESC) and neuronal progenitor cells (mNPC) generated by Tomasz Chelmicki and the Deep Sequencing Facility at the Max Planck Institute of Immunobiology and Epigenetics.

Sample	Generated by	Cells/Tissue	Reads ($\times 10^6$)	Accession number
Input	Tomasz Chelmicki	mESC	62; 27	GSM1251941, GSM1251942
MOF	Tomasz Chelmicki	mESC	23.3; 34.9	GSM1251945, GSM1251946
MSL1	Tomasz Chelmicki	mESC	28.3; 20.3	GSM1251947, GSM1251948
MSL2	Tomasz Chelmicki	mESC	15.8; 11	GSM1251949, GSM1251950
KANSL3	Tomasz Chelmicki	mESC	21.6; 15.3	GSM1251951, GSM1251952
MCRS1	Tomasz Chelmicki	mESC	19.7; 20.7	GSM1251943, GSM1251944
Input	Tomasz Chelmicki	mNPC	106; 102	GSM1251953, GSM1251954
MOF	Tomasz Chelmicki	mNPC	65; 75	GSM1251955, GSM1251956
MSL2	Tomasz Chelmicki	mNPC	116; 112	GSM1251957, GSM1251958

Continued on next page

Table B.12 – *Continued from previous page*

Sample	Generated by	Cells/Tissue	Reads ($\times 10^6$)	Accession number
KANSL3	Tomasz Chelmicki	mNPC	82.8; 93.6	GSM1251957, GSM1251960

Table B.13: In-house generated RNA-seq samples from murine embryonic stem cells (mESC) that were processed by Patrick Wright and Pavan Videm. All samples were treated with lentiviral vectors carrying shRNA constructs before RNA extraction and cDNA generation.

Sample	Generated by	Cells/Tissue	Accession number
Scrambled 1	Tasneem Khanam	mESC	GSM1386921, GSM1386922, GSM1386923
Scrambled 2	Matthew Turley	mESC	GSM1386924, GSM1386925, GSM1386926
Scrambled 3	Tasneem Khanam	mESC	GSM1386927, GSM1386928
MOF RNAi	Tasneem Khanam	mESC	GSM1386909, GSM1386910, GSM1386911
NSL3 RNAi	Tasneem Khanam	mESC	GSM1386918, GSM1386919, GSM1386920
MSL1 RNAi	Matthew Turley	mESC	GSM1386912, GSM1386913, GSM1386914
MSL2 RNAi	Matthew Turley	mESC	GSM1386915, GSM1386916, GSM1386917

Table B.14: Publicly available mouse data sets and annotation that were used for Chelmicki, Dündar et al.[172] (Appendix A.2)

Sample	Downloaded data	Source
Enhancers defined by histone marks	.bed-file of peaks	supplement of Creyghton et al.[243]
Super/typical enhancers	.bed-file of regions	supplement of Whyte et al.[194]
RNA Pol II ChIP-seq (mESC)	.wig-file of signal	GSM632040
RNA Pol II ChIP-seq (mNPC)	.wig-file of signal	GSM632050
p300 ChIP-seq (mESC)	.wig-file of signal	GSM723018
H3K4me1 ChIP-seq (mESC)	.wig-file of signal	GSM723016
H3K27ac ChIP-seq (mESC)	FASTQ file	GSM1005503
CpG methylation (mESC)	.tsv-file of counts	GSE30202
CpG methylation (mNPC)	.tsv-file of counts	GSE30202
DNase hypersensitivity (mESC)	.wig-file of signal	from the UCSC Genome Browser: wgEncodeUwDnaseEscj7S129ME0SigRep1

C. Academic vita

Education

since May 2011 PhD Candidate	International Max Planck Research School for Molecular and Cellular Biology, Faculty of Biology, University of Freiburg
2010 Master of Science (Biomedicine)	University of Würzburg Thesis: *Molecular Studies on BAR domain proteins in murine hematopoietic cells*
2008 Bachelor of Science (Biomedicine)	University of Würzburg Thesis: *Effects of NAD(P)H Oxidase Inhibitors on Thrombocyte Function and Haemostasis*
2006-2011 Stipend	German National Merit Foundation
2005 Abitur	Neideck Gymnasium Arnstadt
2003 High School Diploma	Vallivue High School, Caldwell, ID, USA
2002 Oberstufenabschluss Kammermusik	Kreismusikschule Arnstadt-Ilmenau

Publications

Chelmicki T and **Dündar F*** et al. MOF complexes use short and long-range interactions to ensure stem cell identity and Xist repression. *eLife* 2014 May 19;3:e02024.

Ramirez F and **Dündar F*** et al. deepTools: Analyzing and visualizing high-throughput sequencing data. *Nucleic Acids Research* 2014 Jul;42(Web Server issue):W187-91.

Lam KC, **Mühlpfordt F*** and Vaquerizas JM et al. The NSL complex regulates housekeeping genes in Drosophila. *PLoS Genetics* 2012 May;8(6):e1002736.

Bobak N et al.** Volume regulation of murine T lymphocytes relies on voltage-dependent and two-pore domain potassium channels. *Biochim Biophys Acta.* 2011 Aug;1808(8):2036-44.

* shared first authorship, ** author on 5th position

Bibliography

[1] Vickery, H. The origin of the word protein. *The Yale journal of biology and medicine*, (June 1949), 1950.

[2] Campbell, Neil A; Reece, J. B. Die struktur und funktion biologischer makromolekuele. In Markl, J., editor, *Biologie*, pages 75–102. Spektrum Akademischer Verlag, 2003.

[3] Alberts, B., Johnson, A., and Lewis, J. An Overview of Gene Control. In *Molecular Biology of the Cell*. Garland Science, New York, USA, 4th edition, 2002.

[4] Barrero, M. J. and Malik, S. The RNA Polymerase II Transcriptional Machinery and Its Epigenetic Context. In Kundu, T. K., editor, *Epigenetics: Development and Disease*, volume 61 of *Subcellular Biochemistry*, chapter 11, pages 237–259. Springer Netherlands, Dordrecht, 2013. ISBN 978-94-007-4524-7. doi: 10.1007/978-94-007-4525-4.

[5] Li, J. J., Bickel, P. J., and Biggin, M. D. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ*, 2:e270, 2014. doi: 10.7717/peerj.270.

[6] Campbell, Neil A; Reece, J. B. Die zelle: Ein panoramablick. In Markl, J., editor, *Biologie*, page 133. Spektrum Akademischer Verlag, 2003.

[7] Brown, T. A. Assembly of the Transcription Initiation Complex. In *Genomes*. Wiley-Liss, 2002.

[8] Woodcock, C. L. and Ghosh, R. P. Chromatin higher-order structure and dynamics. *Cold Spring Harbor Perspectives in Biology*, 2(5):a000596, May 2010. doi: 10.1101/cshperspect.a000596.

[9] Kouzarides, T. Chromatin modifications and their function. *Cell*, 128(4):693–705, Feb 2007. doi: 10.1016/j.cell.2007.02.005.

[10] Resverlogix,. Epigenetics, 2014. URL `http://www.resverlogix.com/programs/epiGenetics`. [Online; accessed February-2014].

[11] Tony, K. and Bannister, A. Epigenetic marks and binding proteins — Guide to epigenetic marks. URL `http://www.abcam.com/index.html?pageconfig=resource&rid=11924`. [Online; accessed 20-09-2014].

[12] Manelyte, L. and Laengst, G. Chromatin Remodelers and Their Way of Action. In Radzioch, D., editor, *Chromatin Remodellers*. InTech, April 2013. doi: 10.5772/50815. URL `http://www.intechopen.com/books/chromatin-remodelling/chromatin-remodelers-and-their-way-of-action`.

[13] Turner, B. M. Nucleosome signalling; an evolving concept. *Biochimica et Biophysica Acta*, 1839(8):623–626, Aug 2014. doi: 10.1016/j.bbagrm.2014.01.001.

[14] Tan, M., Luo, H., Lee, S., Jin, F., Yang, J. S., Montellier, E., Buchou, T., Cheng, Z., Rousseaux, S., Rajagopal, N., Lu, Z., Ye, Z., Zhu, Q., Wysocka, J., Ye, Y., Khochbin, S., Ren, B., and Zhao, Y. Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell*, 146(6):1016–1028, Sep 2011. doi: 10.1016/j.cell.2011.08.008.

[15] Ernst, J. and Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnoly*, 28(8):817–825, Aug 2010. doi: 10.1038/nbt.1662.

[16] Kharchenko, P. V., Alekseyenko, A. A., Schwartz, Y. B., Minoda, A., Riddle, N. C., Ernst, J., Sabo, P. J., Larschan, E., Gorchakov, A. A., Gu, T., Linder-Basso, D., Plachetka, A., Shanower, G., Tolstorukov, M. Y., Luquette, L. J., Xi, R., Jung, Y. L., Park, R. W., Bishop, E. P., Canfield, T. K., Sandstrom, R., Thurman, R. E., MacAlpine, D. M., Stamatoyannopoulos, J. A., Kellis, M., Elgin, S. C. R., Kuroda, M. I., Pirrotta, V., Karpen, G. H., and Park, P. J. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature*, 471(7339):480–485, Mar 2011. doi: 10.1038/nature09725.

[17] Henikoff, S. and Shilatifard, A. Histone modification: cause or cog? *Trends in Genetics*, 27(10):389–96, October 2011. doi: 10.1016/j.tig.2011.06.006.

[18] Ptashne, M. Faddish stuff: epigenetics and the inheritance of acquired characteristics. *FASEB journal*, 27(1):1–2, January 2013. doi: 10.1096/fj.13-0101ufm.

[19] Whitehouse, I. and Smith, D. J. Chromatin dynamics at the replication fork: there's more to life than histones. *Current Opinion in Genetics and Development*, 23(2):140–146, Apr 2013. doi: 10.1016/j.gde.2012.12.007.

[20] Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, May 2012. doi: 10.1038/nature11082.

[21] Berndsen, C. E., Tsubota, T., Lindner, S. E., Lee, S., Holton, J. M., Kaufman, P. D., Keck, J. L., and Denu, J. M. Molecular functions of the histone acetyltransferase chaperone complex Rtt109-Vps75. *Nature Structural and Molecular Biology*, 15(9):948–956, Sep 2008.

[22] Hilfiker, A., Hilfiker-Kleiner, D., Pannuti, A., and Lucchesi, J. C. mof, a putative acetyl transferase gene related to the Tip60 and MOZ human genes and to the SAS genes of yeast, is required for dosage compensation in Drosophila. *EMBO Journal*, 16:2054–2060, 1997. doi: 10.1093/emboj/16.8.2054.

[23] Lee, K. K. and Workman, J. L. Histone acetyltransferase complexes: one size doesn't fit all. *Nature Reviews Molecular Cell Biology*, 8:284–295, 2007.

[24] Fraga, M. F., Ballestar, E., Villar-Garea, A., Boix-Chornet, M., Espada, J., Schotta, G., Bonaldi, T., Haydon, C., Ropero, S., Petrie, K., Iyer, N. G., Pérez-Rosado, A., Calvo, E., Lopez, J. A., Cano, A., Calasanz, M. J., Colomer, D., Piris, M. A., Ahn, N., Imhof, A., Caldas, C., Jenuwein, T., and Esteller, M. Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. *Nature Genetics*, 37:391–400, 2005. doi: 10.1038/ng1531.

[25] Pfister, S., Rea, S., Taipale, M., Mendrzyk, F., Straub, B., Ittrich, C., Thuerigen, O., Sinn, H. P., Akhtar, A., and Lichter, P. The histone acetyltransferase hMOF is frequently downregulated in primary breast carcinoma and medulloblastoma and constitutes a biomarker for clinical outcome in medulloblastoma. *International Journal of Cancer*, 122:1207–1213, 2008.

[26] Thomas, T., Dixon, M. P., Kueh, A. J., and Voss, A. K. Mof (MYST1 or KAT8) is essential for progression of embryonic development past the blastocyst stage and required for normal chromatin architecture. *Molecular and Cellular Biology*, 28:5093–5105, 2008. doi: 10.1128/MCB.02202-07.

[27] Gupta, A., Guerin-Peyrou, T. G., Sharma, G. G., Park, C., Agarwal, M., Ganju, R. K., Pandita, S., Choi, K., Sukumar, S., Pandita, R. K., Ludwig, T., and Pandita, T. K. The mammalian ortholog of Drosophila MOF that acetylates histone H4 lysine 16 is essential for embryogenesis and oncogenesis. *Molecular and Cellular Biology*, 28(1):397–409, January 2008. doi: 10.1128/MCB.01045-07.

[28] Kumar, R., Hunt, C. R., Gupta, A., Nannepaga, S., Pandita, R. K., Shay, J. W., Bachoo, R., Ludwig, T., Burns, D. K., and Pandita, T. K. Purkinje cell-specific males absent on the first (mMof) gene deletion results in an ataxia-telangiectasia-like neurological phenotype and backward walking in mice. *Proceedings of the National Academy of Sciences of the United States of America*, 108:3636–3641, 2011. doi: 10.1073/pnas.1016524108.

[29] Fusco, D. N., Brisac, C., John, S. P., Huang, Y.-W., Chin, C. R., Xie, T., Zhao, H., Jilg, N., Zhang, L., Chevaliez, S., Wambua, D., Lin, W., Peng, L., Chung, R. T., and Brass, A. L. A genetic screen identifies interferon-α effector genes required to suppress hepatitis C virus replication. *Gastroenterology*, 144(7):1438–49, 1449.e1–9, June 2013. doi: 10.1053/j.gastro.2013.02.026.

[30] Füllgrabe, J., Lynch-Day, M. A., Heldring, N., Li, W., Struijk, R. B., Ma, Q., Hermanson, O., Rosenfeld, M. G., Klionsky, D. J., and Joseph, B. The histone H4 lysine 16 acetyltransferase hMOF regulates the outcome of autophagy. *Nature*, advance on, July 2013. doi: 10.1038/nature12313.

[31] Qiao, W., Zhang, W., Gai, Y., Zhao, L., and Fan, J. The histone acetyltransferase MOF overexpression blunts cardiac hypertrophy by targeting ROS in mice. *Biochemical and Biophysical Research Communications*, May 2014. doi: 10.1016/j.bbrc.2014.04.112.

[32] Brenachot, X., Rigault, C., Nédélec, E., Laderrière, A., Khanam, T., Gouazé, A., Chaudy, S., Lemoine, A., Datiche, F., Gascuel, J., Pénicaud, L., and Benani, A. The Histone acetyltransferase MOF activates hypothalamic polysialylation to prevent diet-induced obesity in mice. *Molecular Metabolism*, June 2014. doi: 10.1016/j.molmet.2014.05.006.

[33] Meistrich, M. L., Trostle-Weige, P. K., Lin, R., Bhatnagar, Y. M., and Allis, C. D. Highly acetylated H4 is associated with histone displacement in rat spermatids. *Molecular Reproduction and Development*, 31(3):170–181, Mar 1992. doi: 10.1002/mrd.1080310303.

[34] Thomas, T., Loveland, K. L., and Voss, A. K. The genes coding for the MYST family histone acetyltransferases, Tip60 and Mof, are expressed at high levels during sperm development. *Gene Expression Patterns*, 7(6):657–665, Jun 2007. doi: 10.1016/j.modgep.2007.03.005.

[35] Lu, L.-Y., Wu, J., Ye, L., Gavrilina, G. B., Saunders, T. L., and Yu, X. RNF8-dependent histone modifications regulate nucleosome removal during spermatogenesis. *Developmental Cell*, 18(3):371–384, Mar 2010. doi:

10.1016/j.devcel.2010.01.010.

[36] Gupta, A., Hunt, C. R., Pandita, R. K., Pae, J., Komal, K., Singh, M., Shay, J. W., Kumar, R., Ariizumi, K., Horikoshi, N., Hittelman, W. N., Guha, C., Ludwig, T., and Pandita, T. K. T-cell-specific deletion of Mof blocks their differentiation and results in genomic instability in mice. *Mutagenesis*, 28(3):263–70, May 2013. doi: 10.1093/mutage/ges080.

[37] Li, X., Li, L., Pandey, R., Byun, J., Gardner, K., Qin, Z., and Dou, Y. The Histone Acetyltransferase MOF Is a Key Regulator of the Embryonic Stem Cell Core Transcriptional Network. *Cell Stem Cell*, 11(2):163–178, 2012.

[38] Conrad, T., Cavalli, F. M. G., Holz, H., Hallacli, E., Kind, J., Ilik, I., Vaquerizas, J. M., Luscombe, N. M., and Akhtar, A. The MOF Chromobarrel Domain Controls Genome-wide H4K16 Acetylation and Spreading of the MSL Complex. *Developmental Cell*, 22(3):610–24, March 2012. doi: 10.1016/j.devcel.2011.12.016.

[39] Morales, V., Straub, T., Neumann, M. F., Mengus, G., Akhtar, A., and Becker, P. B. Functional integration of the histone acetyltransferase MOF into the dosage compensation complex. *EMBO Journal*, 23:2258–2268, 2004. doi: 10.1038/sj.emboj.7600235.

[40] Kadlec, J., Hallacli, E., Lipp, M., Holz, H., Sanchez-Weatherby, J., Cusack, S., and Akhtar, A. Structural basis for MOF and MSL3 recruitment into the dosage compensation complex by MSL1. *Nature Structural and Molecular Biology*, 18(2):142–9, February 2011. doi: 10.1038/nsmb.1960.

[41] Marín, I. Evolution of chromatin-remodeling complexes: comparative genomics reveals the ancient origin of "novel" compensasome genes. *Journal of Molecular Evolution*, 56(5):527–39, May 2003. doi: 10.1007/s00239-002-2422-1.

[42] Copps, K., Richman, R., Lyman, L. M., Chang, K. A., Rampersad-Ammons, J., and Kuroda, M. I. Complex formation by the Drosophila MSL proteins: role of the MSL2 RING finger in protein complex assembly. *EMBO Journal*, 17(18):5409–5417, Sep 1998. doi: 10.1093/emboj/17.18.5409.

[43] Kruse, J.-P. and Gu, W. MSL2 promotes Mdm2-independent cytoplasmic localization of p53. *The Journal of Biological Chemistry*, 284(5):3250–63, January 2009. doi: 10.1074/jbc.M805658200.

[44] Buscaino, A., Köcher, T., Kind, J. H., Holz, H., Taipale, M., Wagner, K., Wilm, M., and Akhtar, A. MOF-regulated acetylation of MSL-3 in the Drosophila dosage compensation complex. *Molecular Cell*, 11:1265–1277, 2003. doi: 10.1016/S1097-2765(03)00140-0.

[45] Moore, S. A., Ferhatoglu, Y., Jia, Y., Al-Jiab, R. A., and Scott, M. J. Structural and biochemical studies on the chromo-barrel domain of male specific lethal 3 (MSL3) reveal a binding preference for mono- or dimethyllysine 20 on histone H4. *The Journal of Biological Chemistry*, 285(52):40879–90, December 2010. doi: 10.1074/jbc.M110.134312.

[46] Lee, C. G., Chang, K. A., Kuroda, M. I., and Hurwitz, J. The NTPase/helicase activities of Drosophila maleless, an essential factor in dosage compensation. *EMBO Journal*, 16:2671–2681, 1997. doi: 10.1093/emboj/16.10.2671.

[47] Scott, M. J., Pan, L. L., Cleland, S. B., Knox, A. L., and Heinrich, J. MSL1 plays a central role in assembly of the MSL complex, essential for dosage compensation in Drosophila. *EMBO Journal*, 19:144–155, 2000. doi: 10.1093/emboj/19.1.144.

[48] Ilik, I. A., Quinn, J. J., Georgiev, P., Tavares-Cadete, F., Maticzka, D., Toscano, S., Wan, Y., Spitale, R. C., Luscombe, N., Backofen, R., Chang, H. Y., and Akhtar, A. Tandem stem-loops in roX RNAs act together to mediate X chromosome dosage compensation in Drosophila. *Molecular Cell*, 51(2):156–73, July 2013. doi: 10.1016/j.molcel.2013.07.001.

[49] Smith, E. R., Cayrou, C., Huang, R., Lane, W. S., Côté, J., and Lucchesi, J. C. A human protein complex homologous to the Drosophila MSL complex is responsible for the majority of histone H4 acetylation at lysine 16. *Molecular and Cellular Biology*, 25(21):9175–88, November 2005. doi: 10.1128/MCB.25.21.9175-9188.2005.

[50] Dias, J., Van Nguyen, N., Georgiev, P., Gaub, A., Brettschneider, J., Cusack, S., Kadlec, J., and Akhtar, A. Structural analysis of the KANSL1/WDR5/KANSL2 complex reveals that WDR5 is required for efficient assembly and chromatin targeting of the NSL complex. *Genes & Development*, 28(9):929–42, May 2014. doi: 10.1101/gad.240200.114.

[51] Trievel, R. C. and Shilatifard, A. WDR5, a complexed protein. *Nature Structural & Molecular Biology*, 16(7):678–80, July 2009. doi: 10.1038/nsmb0709-678.

[52] Raja, S. J., Charapitsa, I., Conrad, T., Vaquerizas, J. M., Gebhardt, P., Holz, H., Kadlec, J., Fraterman, S., Luscombe, N. M., and Akhtar, A. The nonspecific lethal complex is a transcriptional regulator in Drosophila. *Molecular Cell*, 38(6):827–41, June 2010. doi: 10.1016/j.molcel.2010.05.021.

[53] Mendjan, S., Taipale, M., Kind, J., Holz, H., Gebhardt, P., Schelder, M., Vermeulen, M., Buscaino, A., Duncan, K., Mueller, J., Wilm, M., Stunnenberg, H. G., Saumweber, H., and Akhtar, A. Nuclear pore

components are involved in the transcriptional regulation of dosage compensation in Drosophila. *Molecular Cell*, 21:811–823, 2006.

[54] Hallacli, E., Lipp, M., Georgiev, P., Spielman, C., Cusack, S., Akhtar, A., and Kadlec, J. Msl1-mediated dimerization of the dosage compensation complex is essential for male X-chromosome regulation in Drosophila. *Molecular Cell*, 48(4):587–600, November 2012. doi: 10.1016/j.molcel.2012.09.014.

[55] Wright, A. E. and Mank, J. E. Battle of the sexes: conflict over dosage-sensitive genes and the origin of X chromosome inactivation. *Proceedings of the National Academy of Sciences of the United States of America*, 109(14):5144–5, April 2012. doi: 10.1073/pnas.1202905109.

[56] Mukherjee, A. S. and Beermann, W. Synthesis of ribonucleic acid by the X-chromosomes of Drosophila melanogaster and the problem of dosage compensation. *Nature*, 207(998):785–786, Aug 1965.

[57] Ruiz, M. F., Esteban, M. R., Doñoro, C., Goday, C., and Sánchez, L. Evolution of dosage compensation in Diptera: the gene maleless implements dosage compensation in Drosophila (Brachycera suborder) but its homolog in Sciara (Nematocera suborder) appears to play no role in dosage compensation. *Genetics*, 156 (4):1853–1865, Dec 2000.

[58] Laverty, C., Lucci, J., and Akhtar, A. The MSL complex: X chromosome and beyond. *Current Opinion in Genetics and Development*, 20(2):171–178, Apr 2010. doi: 10.1016/j.gde.2010.01.007.

[59] Abaza, I., Coll, O., Patalano, S., and Gebauer, F. Drosophila UNR is required for translational repression of male-specific lethal 2 mRNA during regulation of X-chromosome dosage compensation. *Genes & Development*, 20(3):380–389, Feb 2006. doi: 10.1101/gad.371906.

[60] Duncan, K., Grskovic, M., Strein, C., Beckmann, K., Niggeweg, R., Abaza, I., Gebauer, F., Wilm, M., and Hentze, M. W. Sex-lethal imparts a sex-specific function to UNR by recruiting it to the msl-2 mRNA 3' UTR: translational repression for dosage compensation. *Genes & Development*, 20(3):368–379, Feb 2006. doi: 10.1101/gad.371406.

[61] Hennig, J., Militti, C., Popowicz, G. M., Wang, I., Sonntag, M., Geerlof, A., Gabel, F., Gebauer, F., and Sattler, M. Structural basis for the assembly of the Sxl-Unr translation regulatory complex. *Nature*, Sep 2014. doi: 10.1038/nature13693.

[62] Gorman, M., Franke, A., and Baker, B. S. Molecular characterization of the male-specific lethal-3 gene and investigations of the regulation of dosage compensation in Drosophila. *Development*, 121(2):463–475, Feb 1995.

[63] Bhadra, U., Pal-Bhadra, M., and Birchler, J. A. Role of the male specific lethal (msl) genes in modifying the effects of sex chromosomal dosage in Drosophila. *Genetics*, 152(1):249–268, May 1999.

[64] Lyman, L. M., Copps, K., Rastelli, L., Kelley, R. L., and Kuroda, M. I. Drosophila male-specific lethal-2 protein: structure/function analysis and dependence on MSL-1 for chromosome association. *Genetics*, 147: 1743–1753, 1997.

[65] Conrad, T. and Akhtar, A. Dosage compensation in Drosophila melanogaster: epigenetic fine-tuning of chromosome-wide transcription. *Nature Reviews Genetics*, 13(2):123–34, February 2011. doi: 10.1038/nrg3124.

[66] Turner, B. M., Birley, A. J., and Lavender, J. Histone H4 isoforms acetylated at specific lysine residues define individual chromosomes and chromatin domains in Drosophila polytene nuclei. *Cell*, 69:375–384, 1992. doi: 10.1016/0092-8674(92)90417-B.

[67] Bone, J. R., Lavender, J., Richman, R., Palmer, M. J., Turner, B. M., and Kuroda, M. I. Acetylated histone H4 on the male X chromosome is associated with dosage compensation in Drosophila. *Genes & Development*, 8:96–104, 1994. doi: 10.1101/gad.8.1.96.

[68] Akhtar, A. and Becker, P. B. Activation of transcription through histone H4 acetylation by MOF, an acetyltransferase essential for dosage compensation in Drosophila. *Molecular Cell*, 5:367–375, 2000. doi: 10.1016/S1097-2765(00)80431-1.

[69] Smith, E. R., Pannuti, A., Gu, W., Steurnagel, A., Cook, R. G., Allis, C. D., and Lucchesi, J. C. The drosophila MSL complex acetylates histone H4 at lysine 16, a chromatin modification linked to dosage compensation. *Molecular and Cellular Biology*, 20(1):312–318, Jan 2000.

[70] Li, X., Wu, L., Corsa, C. A. S., Kunkel, S., and Dou, Y. Two mammalian MOF complexes regulate transcription activation by distinct mechanisms. *Molecular Cell*, 36:290–301, 2009.

[71] Cai, Y., Jin, J., Swanson, S. K., Cole, M. D., Choi, S. H., Florens, L., Washburn, M. P., Conaway, J. W., and Conaway, R. C. Subunit composition and substrate specificity of a MOF-containing histone acetyltransferase distinct from the male-specific lethal (MSL) complex. *The Journal of Biological Chemistry*, 285(7):4268–72, February 2010. doi: 10.1074/jbc.C109.087981.

[72] Gelbart, M. E., Larschan, E., Peng, S., Park, P. J., and Kuroda, M. I. Drosophila MSL complex globally acetylates H4K16 on the male X chromosome for dosage compensation. *Nature Structural and Molecular*

Biology, 16(8):825–32, August 2009. doi: 10.1038/nsmb.1644.

[73] Prestel, M., Feller, C., Straub, T., Mitlöhner, H., and Becker, P. B. The activation potential of MOF is constrained for dosage compensation. *Molecular Cell*, 38(6):815–26, June 2010. doi: 10.1016/j.molcel.2010.05.022.

[74] Larschan, E., Alekseyenko, A. a., Gortchakov, A. a., Peng, S., Li, B., Yang, P., Workman, J. L., Park, P. J., and Kuroda, M. I. MSL complex is attracted to genes marked by H3K36 trimethylation using a sequence-independent mechanism. *Molecular Cell*, 28(1):121–33, October 2007. doi: 10.1016/j.molcel.2007.08.011.

[75] Dion, M. F., Altschuler, S. J., Wu, L. F., and Rando, O. J. Genomic characterization reveals a simple histone H4 acetylation code. *Proceedings of the National Academy of Sciences of the United States of America*, 102(15):5501–6, April 2005. doi: 10.1073/pnas.0500136102.

[76] Kalashnikova, A. A., Porter-Goff, M. E., Muthurajan, U. M., Luger, K., and Hansen, J. C. The role of the nucleosome acidic patch in modulating higher order chromatin structure. *Journal of the Royal Society*, 10 (82):20121022, May 2013. doi: 10.1098/rsif.2012.1022.

[77] Preez, L. L. and Patterton, H.-g. Epigenetics: Development and Disease. In Kundu, T. K., editor, *Epigenetics: Development and Disease*, volume 61 of *Subcellular Biochemistry*, pages 37–55. Springer Netherlands, Dordrecht, 2013. doi: 10.1007/978-94-007-4525-4.

[78] Shogren-Knaak, M., Ishii, H., Sun, J.-M., Pazin, M. J., Davie, J. R., and Peterson, C. L. Histone H4-K16 acetylation controls chromatin structure and protein interactions. *Science*, 311:844–847, 2006. doi: 10.1126/science.1124000.

[79] Robinson, P. J. J., An, W., Routh, A., Martino, F., Chapman, L., Roeder, R. G., and Rhodes, D. 30 nm chromatin fibre decompaction requires both H4-K16 acetylation and linker histone eviction. *Journal of Molecular Biology*, 381(4):816–825, Sep 2008. doi: 10.1016/j.jmb.2008.04.050.

[80] Chen, Z., Ye, X., Tang, N., Shen, S., Li, Z., Niu, X., Lu, S., and Xu, L. The histone acetylranseferase hMOF acetylates Nrf2 and regulates anti-drug responses in human non-small cell lung cancer. *British Journal of Pharmacology*, 171(13):3196–211, July 2014. doi: 10.1111/bph.12661.

[81] Conrad, T., Cavalli, F. M. G., Vaquerizas, J. M., Luscombe, N. M., and Akhtar, A. Drosophila dosage compensation involves enhanced Pol II recruitment to male X-linked promoters. *Science*, 337(6095):742–6, August 2012. doi: 10.1126/science.1221428.

[82] Zippo, A., Serafini, R., Rocchigiani, M., Pennacchini, S., Krepelova, A., and Oliviero, S. Histone crosstalk between H3S10ph and H4K16ac generates a histone code that mediates transcription elongation. *Cell*, 138 (6):1122–1136, Sep 2009. doi: 10.1016/j.cell.2009.07.031.

[83] Smith, E. R., Allis, C. D., and Lucchesi, J. C. Linking global histone acetylation to the transcription enhancement of X-chromosomal genes in Drosophila males. *The Journal of Biological Chemistry*, 276(34): 31483–6, August 2001. doi: 10.1074/jbc.C100351200.

[84] Larschan, E., Bishop, E. P., Kharchenko, P. V., Core, L. J., Lis, J. T., Park, P. J., and Kuroda, M. I. X chromosome dosage compensation via enhanced transcriptional elongation in Drosophila. *Nature*, 471 (7336):115–118, Mar 2011. doi: 10.1038/nature09757.

[85] Birchler, J. A., Pal-Bhadra, M., and Bhadra, U. Dosage dependent gene regulation and the compensation of the X chromosome in Drosophila males. *Genetica*, 117(2-3):179–190, Mar 2003.

[86] Hamada, F. N., Park, P. J., Gordadze, P. R., and Kuroda, M. I. Global regulation of X chromosomal genes by the MSL complex in Drosophila melanogaster. *Genes & Development*, 19(19):2289–94, October 2005. doi: 10.1101/gad.1343705.

[87] Straub, T., Gilfillan, G. D., Maier, V. K., and Becker, P. B. The Drosophila MSL complex activates the transcription of target genes. *Genes & Development*, 19(19):2284–2288, Oct 2005. doi: 10.1101/gad.1343105.

[88] Taipale, M., Rea, S., Richter, K., Vilar, A., Lichter, P., Imhof, A., and Akhtar, A. hMOF histone acetyltransferase is required for histone H4 lysine 16 acetylation in mammalian cells. *Molecular and Cellular Biology*, 25:6798–6810, 2005. doi: 10.1128/MCB.25.15.6798-6810.2005.

[89] Feller, C., Prestel, M., Hartmann, H., Straub, T., Söding, J., and Becker, P. B. The MOF-containing NSL complex associates globally with housekeeping genes, but activates only a defined subset. *Nucleic Acids Research*, 40(4):1509–22, February 2012. doi: 10.1093/nar/gkr869.

[90] Lam, K. C., Mühlpfordt, F., Vaquerizas, J. M., Raja, S. J., Holz, H., Luscombe, N. M., Manke, T., and Akhtar, A. The NSL complex regulates housekeeping genes in Drosophila. *PLoS Genetics*, 8(6):e1002736, January 2012. doi: 10.1371/journal.pgen.1002736.

[91] Wu, L., Zee, B. M., Wang, Y., Garcia, B. A., and Dou, Y. The RING Finger Protein MSL2 in the MOF Complex Is an E3 Ubiquitin Ligase for H2B K34 and Is Involved in Crosstalk with H3 K4 and K79 Methylation. *Molecular Cell*, 43(1):132–144, 2011.

[92] Wu, L., Li, L., Zhou, B., Qin, Z., and Dou, Y. H2B Ubiquitylation Promotes RNA Pol II Processivity via PAF1 and pTEFb. *Molecular Cell*, pages 1–12, May 2014. doi: 10.1016/j.molcel.2014.04.013.

[93] Andersen, D. S., Raja, S. J., Colombani, J., Shaw, R. L., Langton, P. F., Akhtar, A., and Tapon, N. Drosophila MCRS2 associates with RNA polymerase II complexes to regulate transcription. *Molecular and Cellular Biology*, 30(19):4744–55, October 2010. doi: 10.1128/MCB.01586-09.

[94] Kondo, S. and Perrimon, N. A genome-wide RNAi screen identifies core components of the G2-M DNA damage checkpoint. *Science Signaling*, 4:rs1, 2011. doi: 10.1126/scisignal.2001350.

[95] Goshima, G., Wollman, R., Goodwin, S. S., Zhang, N., Scholey, J. M., Vale, R. D., and Stuurman, N. Genes required for mitotic spindle assembly in Drosophila S2 cells. *Science*, 316:417–421, 2007. doi: 10.1126/science.1141314.

[96] Meunier, S. and Vernos, I. K-fibre minus ends are stabilized by a RanGTP-dependent mechanism essential for functional spindle assembly. *Nature Cell Biology*, 13(12):1406–14, December 2011. doi: 10.1038/ncb2372.

[97] Hughes, J. R., Meireles, A. M., Fisher, K. H., Garcia, A., Antrobus, P. R., Wainman, A., Zitzmann, N., Deane, C., Ohkura, H., and Wakefield, J. G. A microtubule interactome: complexes with roles in cell cycle and mitosis. *PLoS Biology*, 6(4):e98, April 2008. ISSN 1545-7885. doi: 10.1371/journal.pbio.0060098.

[98] Muscolini, M., Montagni, E., Palermo, V., Di Agostino, S., Gu, W., Abdelmoula-Souissi, S., Mazzoni, C., Blandino, G., and Tuosto, L. The cancer-associated K351N mutation affects the ubiquitination and the translocation to mitochondria of p53 protein. *The Journal of Biological Chemistry*, 286(46):39693–702, November 2011. doi: 10.1074/jbc.M111.279539.

[99] Sykes, S. M., Mellert, H. S., Holbert, M. A., Li, K., Marmorstein, R., Lane, W. S., and McMahon, S. B. Acetylation of the p53 DNA-Binding Domain Regulates Apoptosis Induction. *Molecular Cell*, 24:841–851, 2006. doi: 10.1016/j.molcel.2006.11.026.

[100] Sykes, S. M., Stanek, T. J., Frank, A., Murphy, M. E., and McMahon, S. B. Acetylation of the DNA binding domain regulates transcription-independent apoptosis by p53. *The Journal of Biological Chemistry*, 284(30): 20197–205, July 2009. doi: 10.1074/jbc.M109.026096.

[101] Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10 (10):669–680, Oct 2009. doi: 10.1038/nrg2641.

[102] Cui, G., Park, S., Badeaux, A. I., Kim, D., Lee, J., Thompson, J. R., Yan, F., Kaneko, S., Yuan, Z., Botuyan, M. V., Bedford, M. T., Cheng, J. Q., and Mer, G. PHF20 is an effector protein of p53 double lysine methylation that stabilizes and activates p53. *Nature Structural and Molecular Biology*, 19(9):916–24, September 2012. doi: 10.1038/nsmb.2353.

[103] Sharma, G. G., So, S., Gupta, A., Kumar, R., Cayrou, C., and Avvakumov, N. MOF and Histone H4 Acetylation at Lysine 16 Are Critical for DNA Damage Response and Double-Strand Break Repair. *Molecular and Cellular Biology*, 30:3582–3595, 2010.

[104] Li, X. and Dou, Y. New perspectives for the regulation of acetyltransferase MOF. *Epigenetics : official journal of the DNA Methylation Society*, 5(3):185–188, April 2010.

[105] Gupta, A., Hunt, C. R., Hegde, M. L., Chakraborty, S., Udayakumar, D., Horikoshi, N., Singh, M., Ramnarain, D. B., Hittelman, W. N., Namjoshi, S., Asaithamby, A., Hazra, T. K., Ludwig, T., Pandita, R. K., Tyler, J. K., and Pandita, T. K. MOF phosphorylation by ATM regulates 53BP1-mediated double-strand break repair pathway choice. *Cell Reports*, 8(1):177–189, 2014.

[106] Lai, Z., Moravcová, S., Canitrot, Y., Andrzejewski, L. P., Walshe, D. M., and Rea, S. Msl2 is a novel component of the vertebrate DNA damage response. *PLoS One*, 8(7):e68549, January 2013. doi: 10.1371/journal.pone.0068549.

[107] Gironella, M., Malicet, C., Cano, C., Sandi, M. J., Hamidi, T., Tauil, R. M. N., Baston, M., Valaco, P., Moreno, S., Lopez, F., Neira, J. L., Dagorn, J. C., and Iovanna, J. L. p8/nupr1 regulates DNA-repair activity after double-strand gamma irradiation-induced DNA damage. *Journal of Cellular Physiology*, 221 (3):594–602, December 2009. doi: 10.1002/jcp.21889.

[108] Altmeyer, M. and Lukas, J. To spread or not to spread–chromatin modifications in response to DNA damage. *Curr Opin Genet Dev*, 23(2):156–165, Apr 2013. doi: 10.1016/j.gde.2012.11.001.

[109] Kind, J., Vaquerizas, J. M., Gebhardt, P., Gentzel, M., Luscombe, N. M., Bertone, P., and Akhtar, A. Genome-wide analysis reveals MOF as a key regulator of dosage compensation and gene expression in Drosophila. *Cell*, 133(5):813–28, May 2008. doi: 10.1016/j.cell.2008.04.036.

[110] Gupta, A., Sharma, G. G., Young, C. S. H., Agarwal, M., Smith, E. R., Paull, T. T., Lucchesi, J. C., Khanna, K. K., Ludwig, T., and Pandita, T. K. Involvement of human MOF in ATM function. *Molecular and Cellular Biology*, 25:5292–305, 2005.

[111] Zheng, H., Yang, L., Peng, L., Izumi, V., Koomen, J., Seto, E., and Chen, J. hMOF acetylation of DBC1/CCAR2 prevents binding and inhibition of SirT1. *Molecular and Cellular Biology*, 33(24):4960–4970,

Dec 2013. doi: 10.1128/MCB.00874-13.

[112] Itsara, A., Vissers, L. E. L. M., Steinberg, K. M., Meyer, K. J., Zody, M. C., Koolen, D. A., de Ligt, J., Cuppen, E., Baker, C., Lee, C., Graves, T. A., Wilson, R. K., Jenkins, R. B., Veltman, J. A., and Eichler, E. E. Resolving the breakpoints of the 17q21.31 microdeletion syndrome with next-generation sequencing. *American Journal of Human Genetics*, 90(4):599–613, April 2012. doi: 10.1016/j.ajhg.2012.02.013.

[113] Koolen, D. A., Kramer, J. M., Neveling, K., Nillesen, W. M., Moore-Barton, H. L., Elmslie, F. V., Toutain, A., Amiel, J., Malan, V., Tsai, A. C.-H., Cheung, S. W., Gilissen, C., Verwiel, E. T. P., Martens, S., Feuth, T., Bongers, E. M. H. F., de Vries, P., Scheffer, H., Vissers, L. E. L. M., de Brouwer, A. P. M., Brunner, H. G., Veltman, J. A., Schenck, A., Yntema, H. G., and de Vries, B. B. A. Mutations in the chromatin modifier gene KANSL1 cause the 17q21.31 microdeletion syndrome. *Nature Genetics*, 44(6):639–641, Jun 2012. doi: 10.1038/ng.2262.

[114] Zollino, M., Orteschi, D., Murdolo, M., Lattante, S., Battaglia, D., Stefanini, C., Mercuri, E., Chiurazzi, P., Neri, G., and Marangi, G. Mutations in KANSL1 cause the 17q21.31 microdeletion syndrome phenotype. *Nature Genetics*, 44(6):636–8, June 2012. doi: 10.1038/ng.2257.

[115] Gilissen, C., Hehir-Kwa, J. Y., Thung, D. T., van de Vorst, M., van Bon, B. W. M., Willemsen, M. H., Kwint, M., Janssen, I. M., Hoischen, A., Schenck, A., Leach, R., Klein, R., Tearle, R., Bo, T., Pfundt, R., Yntema, H. G., de Vries, B. B. a., Kleefstra, T., Brunner, H. G., Vissers, L. E. L. M., and Veltman, J. a. Genome sequencing identifies major causes of severe intellectual disability. *Nature*, 511:344–347, 2014. doi: 10.1038/nature13394.

[116] Jin, J., Cai, Y., Yao, T., Gottschalk, A. J., Florens, L., Swanson, S. K., Gutiérrez, J. L., Coleman, M. K., Workman, J. L., Mushegian, A., Washburn, M. P., Conaway, R. C., and Conaway, J. W. A mammalian chromatin remodeling complex with similarities to the yeast INO80 complex. *The Journal of Biological Chemistry*, 280(50):41207–12, December 2005. doi: 10.1074/jbc.M509128200.

[117] Shimono, K., Shimono, Y., Shimokata, K., Ishiguro, N., and Takahashi, M. Microspherule protein 1, Mi-2beta, and RET finger protein associate in the nucleolus and up-regulate ribosomal gene transcription. *The Journal of Biological Chemistry*, 280(47):39436–47, November 2005. doi: 10.1074/jbc.M507356200.

[118] Hsu, C.-C., Lee, Y.-C., Yeh, S.-H., Chen, C.-H., Wu, C.-C., Wang, T.-Y., Chen, Y.-N., Hung, L.-Y., Liu, Y.-W., Chen, H.-K., Hsiao, Y.-T., Wang, W.-S., Tsou, J.-H., Tsou, Y.-H., Wu, M.-H., Chang, W.-C., and Lin, D.-Y. 58-kDa microspherule protein (MSP58) is novel Brahma-related gene 1 (BRG1)-associated protein that modulates p53/p21 senescence pathway. *The Journal of Biological Chemistry*, 287(27):22533–48, June 2012. doi: 10.1074/jbc.M111.335331.

[119] Hsu, C.-C., Chen, C.-H., Hsu, T.-I., Hung, J.-J., Ko, J.-L., Zhang, B., Lee, Y.-C., Chen, H.-K., Chang, W.-C., and Lin, D.-Y. The 58-kda microspherule protein (MSP58) represses human telomerase reverse transcriptase (hTERT) gene expression and cell proliferation by interacting with telomerase transcriptional element-interacting factor (TEIF). *Biochimica et Biophysica Acta*, 1843(3):565–79, March 2014. doi: 10.1016/j.bbamcr.2013.12.004.

[120] Teytelman, L., Thurtle, D. M., Rine, J., and van Oudenaarden, A. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 110(46):18602–7, November 2013. doi: 10.1073/pnas.1316064110.

[121] Gavrilov, A., Razin, S. V., and Cavalli, G. In vivo formaldehyde cross-linking: it is time for black box analysis. *Briefings in Functional Genomics*, Sep 2014. doi: 10.1093/bfgp/elu037.

[122] Wikipedia,. ChIP-on-chip — Wikipedia, The Free Encyclopedia, 2014. URL `http://en.wikipedia.org/wiki/ChIP-on-chip`. [Online; accessed 2012].

[123] Kasinathan, S., Orsi, G. A., Zentner, G. E., Ahmad, K., and Henikoff, S. High-resolution mapping of transcription factor binding sites on native chromatin. *Nature Methods*, 11(2):203–9, February 2014. doi: 10.1038/nmeth.2766.

[124] O'Neill, L. P. and Turner, B. M. Immunoprecipitation of native chromatin: NChIP. *Methods*, 31(1):76–82, Sep 2003.

[125] Rhee, H. S. and Pugh, B. F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, 147(6):1408–1419, Dec 2011. doi: 10.1016/j.cell.2011.11.013.

[126] Adli, M. and Bernstein, B. E. Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq. *Nature Protocols*, 6(10):1656–1668, Oct 2011. doi: 10.1038/nprot.2011.402.

[127] Shankaranarayanan, P., Mendoza-Parra, M.-A., Walia, M., Wang, L., Li, N., Trindade, L. M., and Gronemeyer, H. Single-tube linear DNA amplification (LinDA) for robust ChIP-seq. *Nature Methods*, 8(7):565–567, Jul 2011. doi: 10.1038/nmeth.1626.

[128] Techspotlight: Illumina Sequencing Technology. URL `http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf`.

[129] Chen, Y., Negre, N., Li, Q., Mieczkowska, J. O., Slattery, M., Liu, T., Zhang, Y., Kim, T.-K., He, H. H., Zieba, J., Ruan, Y., Bickel, P. J., Myers, R. M., Wold, B. J., White, K. P., Lieb, J. D., and Liu, X. S. Systematic evaluation of factors influencing ChIP-seq fidelity. *Nature Methods*, 9(6):609–614, Jun 2012. doi: 10.1038/nmeth.1985.

[130] Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K. I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A. J., Hoffman, M. M., Iyer, V. R., Jung, Y. L., Karmakar, S., Kellis, M., Kharchenko, P. V., Li, Q., Liu, T., Liu, X. S., Ma, L., Milosavljevic, A., Myers, R. M., Park, P. J., Pazin, M. J., Perry, M. D., Raha, D., Reddy, T. E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J. A., Tolstorukov, M. Y., White, K. P., Xi, S., Farnham, P. J., Lieb, J. D., Wold, B. J., and Snyder, M. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, 22(9):1813–1831, Sep 2012. doi: 10.1101/gr.136184.111.

[131] Jung, Y. L., Luquette, L. J., Ho, J. W. K., Ferrari, F., Tolstorukov, M., Minoda, A., Issner, R., Epstein, C. B., Karpen, G. H., Kuroda, M. I., and Park, P. J. Impact of sequencing depth in ChIP-seq experiments. *Nucleic Acids Research*, 42(9):e74, 2014. doi: 10.1093/nar/gku178.

[132] Liu, E. T., Pott, S., and Huss, M. Q&A: ChIP-seq technologies and the study of gene regulation. *BMC Biology*, 8:56, 2010. doi: 10.1186/1741-7007-8-56.

[133] Waldminghaus, T. and Skarstad, K. ChIP on Chip: surprising results are often artifacts. *BMC Genomics*, 11:414, 2010. doi: 10.1186/1471-2164-11-414.

[134] Vega, V. B., Cheung, E., Palanisamy, N., and Sung, W.-K. Inherent signals in sequencing-based Chromatin-ImmunoPrecipitation control libraries. *PLoS One*, 4(4):e5241, 2009. doi: 10.1371/journal.pone.0005241.

[135] Marinov, G. K., Kundaje, A., Park, P. J., and Wold, B. J. Large-scale quality analysis of published ChIP-seq data. *G3 (Bethesda)*, 4(2):209–223, Feb 2014. doi: 10.1534/g3.113.008680.

[136] Ledergerber, C. and Dessimoz, C. Base-calling for next-generation sequencing platforms. *Briefings in Bioinformatics*, 12(5):489–497, Sep 2011. doi: 10.1093/bib/bbq077.

[137] Minoche, A. E., Dohm, J. C., and Himmelbauer, H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biology*, 12(11):R112, 2011. doi: 10.1186/gb-2011-12-11-r112.

[138] Cheung, M.-S., Down, T. A., Latorre, I., and Ahringer, J. Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic Acids Research*, 39(15):e103, Aug 2011. doi: 10.1093/nar/gkr425.

[139] Benjamini, Y. and Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, 40(10):e72, May 2012. doi: 10.1093/nar/gks001.

[140] Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Madrigal, P., Taslim, C., and Zhang, J. Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Computational Biology*, 9(11): e1003326, 2013. doi: 10.1371/journal.pcbi.1003326.

[141] Kidder, B. L., Hu, G., and Zhao, K. ChIP-Seq: technical considerations for obtaining high-quality data. *Nature Immunology*, 12(10):918–922, Oct 2011. doi: 10.1038/ni.2117.

[142] Carroll, T. S., Liang, Z., Salama, R., Stark, R., and de Santiago, I. Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. *Frontiers in Genetics*, 5:75, 2014. doi: 10.3389/fgene.2014.00075.

[143] Kundaje, A. Blacklisted genomic regions for functional genomics analysis, 2014. URL `https://sites.google.com/site/anshulkundaje/projects/blacklists`. accessed 18/09/2014s.

[144] Zhang, Z. D., Rozowsky, J., Snyder, M., Chang, J., and Gerstein, M. Modeling ChIP sequencing in silico with applications. *PLoS Computational Biology*, 4(8):e1000158, 2008. doi: 10.1371/journal.pcbi.1000158.

[145] Leleu, M., Lefebvre, G., and Rougemont, J. Processing and analyzing ChIP-seq data: from short reads to regulatory interactions. *Briefings in Functional Genomics*, 9(5-6):466–476, Dec 2010. doi: 10.1093/bfgp/elq022.

[146] Flicek, P. and Birney, E. Sense from sequence reads: methods for alignment and assembly. *Nature Methods*, 6(11 Suppl):S6–S12, Nov 2009. doi: 10.1038/nmeth.1376.

[147] Langmead, B. and Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4): 357–359, Apr 2012. doi: 10.1038/nmeth.1923.

[148] Li, H. and Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–595, Mar 2010. doi: 10.1093/bioinformatics/btp698.

[149] Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009. doi: 10.1186/gb-2009-10-3-r25.

[150] Bardet, A. F., He, Q., Zeitlinger, J., and Stark, A. A computational pipeline for comparative ChIP-seq analyses. *Nature Protocols*, 7(1):45–61, Jan 2012. doi: 10.1038/nprot.2011.420.

[151] Kharchenko, P. V., Tolstorukov, M. Y., and Park, P. J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature Biotechnology*, 26(12):1351–1359, Dec 2008. doi: 10.1038/nbt.1508.

[152] Ho, J. W. K., Bishop, E., Karchenko, P. V., Nègre, N., White, K. P., and Park, P. J. ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. *BMC Genomics*, 12:134, 2011. doi: 10.1186/1471-2164-12-134.

[153] Tuteja, G., White, P., Schug, J., and Kaestner, K. H. Extracting transcription factor targets from ChIP-Seq data. *Nucleic Acids Research*, 37(17):e113, Sep 2009. doi: 10.1093/nar/gkp536.

[154] Flensburg, C., Kinkel, S. A., Keniry, A., Blewitt, M., and Oshlack, A. A comparison of control samples for ChIP-seq of histone modifications. *bioRxiv*, page 007609, August 2014. doi: 10.1101/007609.

[155] Consortium, E. N. C. O. D. E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012. doi: 10.1038/nature11247.

[156] Treangen, T. J. and Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13(1):36–46, Jan 2012. doi: 10.1038/nrg3117.

[157] Rozowsky, J., Euskirchen, G., Auerbach, R. K., Zhang, Z. D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., and Gerstein, M. B. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature Biotechnology*, 27(1):66–75, January 2009. doi: 10.1038/nbt.1518.

[158] Dündar, F. Quality controls of deeply sequenced reads — deepTools Wiki, 2014. URL `https://github.com/fidelram/deepTools/wiki/QC`. [Online; accessed 20-September-2014].

[159] Diaz, A., Nellore, A., and Song, J. S. CHANCE: comprehensive software for quality control and validation of ChIP-seq data. *Genome Biology*, 13(10):R98, 2012. doi: 10.1186/gb-2012-13-10-r98.

[160] Pepke, S., Wold, B., and Mortazavi, A. Computation for ChIP-seq and RNA-seq studies. *Nature Methods*, 6(11 Suppl):S22–S32, Nov 2009. doi: 10.1038/nmeth.1371.

[161] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, Jul 2008. doi: 10.1038/nmeth.1226.

[162] Mendoza-Parra, M. A., Sankar, M., Walia, M., and Gronemeyer, H. POLYPHEMUS: R package for comparative analysis of RNA polymerase II ChIP-seq profiles by non-linear normalization. *Nucleic Acids Research*, 40(4):e30, Feb 2012. doi: 10.1093/nar/gkr1205.

[163] Robinson, M. D. and Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25, 2010. doi: 10.1186/gb-2010-11-3-r25.

[164] Wilbanks, E. G. and Facciotti, M. T. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS One*, 5(7):e11471, 2010. doi: 10.1371/journal.pone.0011471.

[165] Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nature Protocols*, 7(9):1728–1740, Sep 2012. doi: 10.1038/nprot.2012.101.

[166] Laajala, T. D., Raghav, S., Tuomela, S., Lahesmaa, R., Aittokallio, T., and Elo, L. L. A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. *BMC Genomics*, 10:618, 2009. doi: 10.1186/1471-2164-10-618.

[167] Rye, M. B., Sætrom, P., and Drabløs, F. A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs. *Nucleic Acids Research*, 39(4):e25, Mar 2011. doi: 10.1093/nar/gkq1187.

[168] Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A., and Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Research*, 42(Web Server issue):W187–W191, Jul 2014. doi: 10.1093/nar/gku365.

[169] Machanick, P. and Bailey, T. L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, 27(12):1696–1697, Jun 2011. doi: 10.1093/bioinformatics/btr189.

[170] Huang, D. W., Sherman, B. T., and Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, 2009. doi: 10.1038/nprot.2008.211.

[171] Welch, R. P., Lee, C., Imbriano, P. M., Patil, S., Weymouth, T. E., Smith, R. A., Scott, L. J., and Sartor, M. A. ChIP-Enrich: gene set enrichment testing for ChIP-seq data. *Nucleic Acids Research*, 42(13):e105, 2014. doi: 10.1093/nar/gku463.

[172] Chelmicki, T., Dündar, F., Turley, M., Khanam, T., Aktas, T,, Ramirez, F., Gendrel, A.-V., Wright, P. R., Videm, P., Backofen, R., Heard, E., Manke, T., and Akhtar, A. MOF-associated complexes ensure stem cell identity and Xist repression. *eLife*, 2014. doi: 10.7554/eLife.02024.

[173] Ferrari, F., Alekseyenko, A. A., Park, P. J., and Kuroda, M. I. Transcriptional control of a whole chromosome: emerging models for dosage compensation. *Nature Structural and Molecular Biology*, 21(2):118–125, Feb 2014. doi: 10.1038/nsmb.2763.

[174] Alekseyenko, A. A., Larschan, E., Lai, W. R., Park, P. J., and Kuroda, M. I. High-resolution ChIP-chip analysis reveals that the Drosophila MSL complex selectively identifies active genes on the male X chromosome. *Genes & Development*, 20:848–857, 2006. doi: 10.1101/gad.1400206.

[175] Straub, T., Zabel, A., Gilfillan, G. D., Feller, C., and Becker, P. B. Different chromatin interfaces of the Drosophila dosage compensation complex revealed by high-shear ChIP-seq. *Genome Research*, 23(3):473–85, March 2013. doi: 10.1101/gr.146407.112.

[176] Sikora-Wohlfeld, W., Ackermann, M., Christodoulou, E. G., Singaravelu, K., and Beyer, A. Assessing computational methods for transcription factor target gene identification based on ChIP-seq data. *PLoS Computational Biology*, 9(11):e1003342, 2013. doi: 10.1371/journal.pcbi.1003342.

[177] Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V. V., and Ren, B. A map of the cis-regulatory sequences in the mouse genome. *Nature*, 488(7409):116–120, Aug 2012. doi: 10.1038/nature11243.

[178] McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., and Bejerano, G. GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, 28 (5):495–501, May 2010. doi: 10.1038/nbt.1630.

[179] Ravens, S., Fournier, M., Ye, T., Stierle, M., Dembele, D., Chavant, V., and Tora, L. MOF-associated complexes have overlapping and unique roles in regulating pluripotency in embryonic stem cells and during differentiation. *eLife*, page e02104, June 2014. doi: 10.7554/eLife.02104.

[180] R Core Team,. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org/`.

[181] Taylor, G., Eskeland, R., Hekimoglu-Balkan, B., Pradeepa, M., and Bickmore, W. A. H4K16 acetylation marks active genes and enhancers of embryonic stem cells, but does not alter chromatin compaction. *Genome Research*, August 2013. doi: 10.1101/gr.155028.113.

[182] Forsberg, E. C. and Bresnick, E. H. Histone acetylation beyond promoters: long-range acetylation patterns in the chromatin world. *Bioessays*, 23(9):820–830, Sep 2001. doi: 10.1002/bies.1117.

[183] Filippakopoulos, P., Picaud, S., Mangos, M., Keates, T., Lambert, J.-P., Barsyte-Lovejoy, D., Felletar, I., Volkmer, R., Müller, S., Pawson, T., Gingras, A.-C., Arrowsmith, C. H., and Knapp, S. Histone recognition and large-scale structural analysis of the human bromodomain family. *Cell*, 149(1):214–231, Mar 2012. doi: 10.1016/j.cell.2012.02.013.

[184] Henry, R. A., Kuo, Y.-M., and Andrews, A. J. Differences in specificity and selectivity between CBP and p300 acetylation of histone H3 and H3/H4. *Biochemistry*, 52(34):5746–5759, Aug 2013. doi: 10.1021/bi400684q.

[185] Suganuma, T., Gutiérrez, J. L., Li, B., Florens, L., Swanson, S. K., Washburn, M. P., Abmayr, S. M., and Workman, J. L. ATAC is a double histone acetyltransferase complex that stimulates nucleosome sliding. *Nature Structural & Molecular Biology*, 15:364–372, 2008. doi: 10.1038/nsmb.1397.

[186] Guelman, S., Kozuka, K., Mao, Y., Pham, V., Solloway, M. J., Wang, J., Wu, J., Lill, J. R., and Zha, J. The double-histone-acetyltransferase complex ATAC is essential for mammalian development. *Molecular and Cellular Biology*, 29(5):1176–1188, Mar 2009. doi: 10.1128/MCB.01599-08.

[187] Dou, Y., Milne, T. A., Tackett, A. J., Smith, E. R., Fukuda, A., Wysocka, J., Allis, C. D., Chait, B. T., Hess, J. L., and Roeder, R. G. Physical association and coordinate function of the H3 K4 methyltransferase MLL1 and the H4 K16 acetyltransferase MOF. *Cell*, 121(6):873–85, June 2005. doi: 10.1016/j.cell.2005.04.031.

[188] Filion, G. J., van Bemmel, J. G., Braunschweig, U., Talhout, W., Kind, J., Ward, L. D., Brugman, W., de Castro, I. J., Kerkhoven, R. M., Bussemaker, H. J., and van Steensel, B. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell*, 143(2):212–224, Oct 2010. doi: 10.1016/j.cell.2010.09.009.

[189] Ohler, U., chun Liao, G., Niemann, H., and Rubin, G. M. Computational analysis of core promoters in the Drosophila genome. *Genome Biology*, 3(12):RESEARCH0087, 2002.

[190] Hoskins, R. A., Landolin, J. M., Brown, J. B., Sandler, J. E., Takahashi, H., Lassmann, T., Yu, C., Booth, B. W., Zhang, D., Wan, K. H., Yang, L., Boley, N., Andrews, J., Kaufman, T. C., Graveley, B. R., Bickel, P. J., Carninci, P., Carlson, J. W., and Celniker, S. E. Genome-wide analysis of promoter architecture in Drosophila melanogaster. *Genome Research*, 21(2):182–192, Feb 2011. doi: 10.1101/gr.112466.110.

[191] Cherbas, L., Willingham, A., Zhang, D., Yang, L., Zou, Y., Eads, B. D., Carlson, J. W., Landolin, J. M., Kapranov, P., Dumais, J., Samsonova, A., Choi, J.-H., Roberts, J., Davis, C. A., Tang, H., van Baren, M. J., Ghosh, S., Dobin, A., Bell, K., Lin, W., Langton, L., Duff, M. O., Tenney, A. E., Zaleski, C., Brent, M. R., Hoskins, R. A., Kaufman, T. C., Andrews, J., Graveley, B. R., Perrimon, N., Celniker, S. E., Gingeras, T. R., and Cherbas, P. The transcriptional diversity of 25 Drosophila cell lines. *Genome Research*, 21(2):301–314, Feb 2011. doi: 10.1101/gr.112961.110.

[192] Graveley, B. R., Brooks, A. N., Carlson, J. W., Duff, M. O., Landolin, J. M., Yang, L., Artieri, C. G., van Baren, M. J., Boley, N., Booth, B. W., Brown, J. B., Cherbas, L., Davis, C. A., Dobin, A., Li, R., Lin, W., Malone, J. H., Mattiuzzo, N. R., Miller, D., Sturgill, D., Tuch, B. B., Zaleski, C., Zhang, D., Blanchette, M., Dudoit, S., Eads, B., Green, R. E., Hammonds, A., Jiang, L., Kapranov, P., Langton, L., Perrimon, N., Sandler, J. E., Wan, K. H., Willingham, A., Zhang, Y., Zou, Y., Andrews, J., Bickel, P. J., Brenner, S. E., Brent, M. R., Cherbas, P., Gingeras, T. R., Hoskins, R. A., Kaufman, T. C., Oliver, B., and Celniker, S. E. The developmental transcriptome of Drosophila melanogaster. *Nature*, 471(7339):473–479, Mar 2011. doi: 10.1038/nature09715.

[193] Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., and Kellis, M. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434(7031):338–345, Mar 2005. doi: 10.1038/nature03441.

[194] Whyte, W. a., Orlando, D. a., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I., and Young, R. a. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2):307–19, April 2013. doi: 10.1016/j.cell.2013.03.035.

[195] Folmes, C. D. L., Dzeja, P. P., Nelson, T. J., and Terzic, A. Metabolic plasticity in stem cell homeostasis and differentiation. *Cell Stem Cell*, 11(5):596–606, Nov 2012. doi: 10.1016/j.stem.2012.10.002.

[196] Koledova, Z., Krämer, A., Kafkova, L. R., and Divoky, V. Cell-cycle regulation in embryonic stem cells: centrosomal decisions on self-renewal. *Stem Cells and Development*, 19(11):1663–1678, Nov 2010. doi: 10.1089/scd.2010.0136.

[197] Watanabe-Susaki, K., Takada, H., Enomoto, K., Miwata, K., Ishimine, H., Intoh, A., Ohtaka, M., Nakanishi, M., Sugino, H., Asashima, M., and Kurisaki, A. Biosynthesis of Ribosomal RNA in Nucleoli Regulates Pluripotency and Differentiation Ability of Pluripotent Stem Cells. *Stem Cells*, Sep 2014. doi: 10.1002/stem.1825.

[198] Alekseyenko, A. A., Peng, S., Larschan, E., Gorchakov, A. A., Lee, O.-K., Kharchenko, P., McGrath, S. D., Wang, C. I., Mardis, E. R., Park, P. J., and Kuroda, M. I. A sequence motif within chromatin entry sites directs MSL establishment on the Drosophila X chromosome. *Cell*, 134(4):599–609, August 2008. doi: 10.1016/j.cell.2008.06.033.

[199] Straub, T., Grimaud, C., Gilfillan, G. D., Mitterweger, A., and Becker, P. B. The chromosomal high-affinity binding sites for the Drosophila dosage compensation complex. *PLoS Genetics*, 4(12):e1000302, December 2008. doi: 10.1371/journal.pgen.1000302.

[200] Kelley, R. L., Solovyeva, I., Lyman, L. M., Richman, R., Solovyev, V., and Kuroda, M. I. Expression of Msl-2 causes assembly of dosage compensation regulators on the X chromosomes and female lethality in Drosophila. *Cell*, 81(6):867–877, June 1995. doi: 10.1016/0092-8674(95)90007-1.

[201] Pollex, T. and Heard, E. Recent advances in X-chromosome inactivation research. *Current Opinion in Cell Biology*, 24(6):825–32, December 2012. doi: 10.1016/j.ceb.2012.10.007.

[202] Zawel, L., Dai, J. L., Buckhaults, P., Zhou, S., Kinzler, K. W., Vogelstein, B., and Kern, S. E. Human Smad3 and Smad4 are sequence-specific transcription activators. *Molecular Cell*, 1(4):611–617, Mar 1998.

[203] Sansó, M. and Fisher, R. P. Pause, play, repeat: CDKs push RNAP II's buttons. *Transcription*, 4(4):146–152, 2013.

[204] Barth, T. K. and Imhof, A. Fast signals and slow marks: the dynamics of histone modifications. *Trends Biochem Sci*, 35(11):618–626, Nov 2010. doi: 10.1016/j.tibs.2010.05.006. URL `http://dx.doi.org/10.1016/j.tibs.2010.05.006`.

[205] Giampieri, R., Scartozzi, M., Loretelli, C., Piva, F., Mandolesi, A., Lezoche, G., Prete, M. D., Bittoni, A., Faloppi, L., Bianconi, M., Cecchini, L., Guerrieri, M., Bearzi, I., and Cascinu, S. Cancer stem cell gene profile as predictor of relapse in high risk stage II and stage III, radically resected colon cancer patients. *PLoS One*, 8(9):e72843, 2013. doi: 10.1371/journal.pone.0072843.

[206] Liu, N., Zhang, R., Zhao, X., Su, J., Bian, X., Ni, J., Yue, Y., Cai, Y., and Jin, J. A potential diagnostic marker for ovarian cancer: Involvement of the histone acetyltransferase, human males absent on the first. *Oncology Letters*, 6(2):393–400, Aug 2013. doi: 10.3892/ol.2013.1380.

[207] Wang, Y., Zhang, R., Wu, D., Lu, Z., Sun, W., Cai, Y., Wang, C., and Jin, J. Epigenetic change in kidney tumor: downregulation of histone acetyltransferase MYST1 in human renal cell carcinoma. *Journal of Experimental and Clinical Cancer Research*, 32:8, 2013. doi: 10.1186/1756-9966-32-8.

[208] Cao, L., Zhu, L., Yang, J., Su, J., Ni, J., Du, Y., Liu, D., Wang, Y., Wang, F., Jin, J., and Cai, Y. Correlation of low expression of hMOF with clinicopathological features of colorectal carcinoma, gastric cancer and renal cell carcinoma. *International Journal of Oncology*, 44(4):1207–1214, Apr 2014. doi: 10.3892/ijo.2014.2266.

[209] Zhang, J., Liu, H., Pan, H., Yang, Y., Huang, G., Yang, Y., Zhou, W.-P., and Pan, Z.-Y. The histone acetyltransferase hMOF suppresses hepatocellular carcinoma growth. *Biochemical Biophysical Research*

Communication, Aug 2014. doi: 10.1016/j.bbrc.2014.08.122.

[210] Peedicayil, A., Vierkant, R. A., Hartmann, L. C., Fridley, B. L., Fredericksen, Z. S., White, K. L., Elliott, E. A., Phelan, C. M., Tsai, Y.-Y., Berchuck, A., Iversen, E. S., Couch, F. J., Peethamabaran, P., Larson, M. C., Kalli, K. R., Kosel, M. L., Shridhar, V., Rider, D. N., Liebow, M., Cunningham, J. M., Schildkraut, J. M., Sellers, T. A., and Goode, E. L. Risk of ovarian cancer and inherited variants in relapse-associated genes. *PloS One*, 5(1):e8884, January 2010. doi: 10.1371/journal.pone.0008884.

[211] Sinenko, S. A., Hung, T., Moroz, T., Tran, Q.-M., Sidhu, S., Cheney, M. D., Speck, N. A., and Banerjee, U. Genetic manipulation of AML1-ETO-induced expansion of hematopoietic precursors in a Drosophila model. *Blood*, 116(22):4612–20, November 2010. doi: 10.1182/blood-2010-03-276998.

[212] Fischer, U., Struss, A. K., Hemmer, D., Pallasch, C. P., Steudel, W. I., and Meese, E. Glioma-expressed antigen 2 (GLEA2): a novel protein that can elicit immune responses in glioblastoma patients and some controls. *Clinical and Experimental Immunology*, 126(2):206–213, Nov 2001.

[213] Pallasch, C. P., Struss, A.-K., Munnia, A., König, J., Steudel, W.-I., Fischer, U., and Meese, E. Autoantibodies against GLEA2 and PHF3 in glioblastoma: tumor-associated autoantibodies correlated with prolonged survival. *International Journal of Cancer*, 117(3):456–459, Nov 2005. doi: 10.1002/ijc.20929.

[214] Taniwaki, M., Daigo, Y., Ishikawa, N., Takano, A., Tsunoda, T., Yasui, W., Inai, K., Kohno, N., and Nakamura, Y. Gene expression profiles of small-cell lung cancers: Molecular signatures of lung cancer. *International Journal of Oncology*, 29(3):567–575, 2006.

[215] Bankovic, J., Stojsic, J., Jovanovic, D., Andjelkovic, T., Milinkovic, V., Ruzdijic, S., and Tanic, N. Identification of genes associated with non-small-cell lung cancer promotion and progression. *Lung cancer (Amsterdam, Netherlands)*, 67(2):151–9, March 2010. doi: 10.1016/j.lungcan.2009.04.010.

[216] Okumura, K., Zhao, M., Depinho, R. A., Furnari, F. B., and Cavenee, W. K. Cellular transformation by the MSP58 oncogene is inhibited by its physical interaction with the PTEN tumor suppressor. *Proceedings of the National Academy of Sciences of the United States of America*, 102(8):2703–6, February 2005. doi: 10.1073/pnas.0409370102.

[217] Shi, H., Li, S.-J., Zhang, B., Liu, H.-L., and Chen, C.-S. Expression of MSP58 in human colorectal cancer and its correlation with prognosis. *Medical Oncology*, 29(5):3136–42, December 2012. doi: 10.1007/s12032-012-0284-y.

[218] Wu, L., guo Zhang, Z., zhou Qin, H., Zhang, J., dong Gao, G., Lin, W., Wang, J., and Zhang, J. Down-regulation of MSP58 suppresses cell proliferation in neuroblastoma cell lines. *Neuroreport*, 23(16):932–936, Nov 2012. doi: 10.1097/WNR.0b013e328359566e.

[219] Lin, W., Li, X.-M., Zhang, J., Huang, Y., Wang, J., Zhang, J., Jiang, X.-F., and Fei, Z. Increased expression of the 58-kD microspherule protein (MSP58) is correlated with poor prognosis in glioma patients. *Medical Oncology*, 30(4):677, December 2013. doi: 10.1007/s12032-013-0677-6.

[220] Zhong, M., Zhang, X., Li, B., Chen, C.-s., Ji, G.-l., Li, S.-x., Bi, D.-q., Zhao, Q.-c., and Shi, H. Expression of MSP58 in hepatocellular carcinoma. *Medical Oncology*, 30(2):539, June 2013. doi: 10.1007/s12032-013-0539-2.

[221] Wu, M. and Shu, H.-B. MLL1/WDR5 complex in leukemogenesis and epigenetic regulation. *Chinese Journal of Cancer*, 30(4):240–6, April 2011.

[222] Quinlan, A. R. and Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, Mar 2010. doi: 10.1093/bioinformatics/btq033.

[223] Ali, R., Cavalli, F. M., Vaquerizas, J. M., and Luscombe, N. M. A pipeline for ChIP-seq data analysis — Epigenesys Protocols, 2012. URL `http://www.epigenesys.eu/images/stories/protocols/pdf/20120529112503_p56.pdf`. [Online; accessed 17-06-2014].

[224] Anders, S. and Huber, W. Differential expression analysis for sequence count data. *Genome Biology*, 11 (10):R106, 2010. doi: 10.1186/gb-2010-11-10-r106.

[225] Goecks, J., Nekrutenko, A., Taylor, J., and Team, G. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11 (8):R86, 2010. doi: 10.1186/gb-2010-11-8-r86.

[226] Wickham, H. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009. ISBN 978-0-387-98140-6.

[227] Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192, Mar 2013. doi: 10.1093/bib/bbs017.

[228] Bailey, T. L. and Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biology*, 2:28–36, 1994.

[229] Salmon-Divon, M., Dvinge, H., Tammoja, K., and Bertone, P. PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. *BMC Bioinformatics*, 11:415, 2010. doi: 10.1186/1471-2105-11-415.

[230] Roider, H. G., Kanhere, A., Manke, T., and Vingron, M. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, 23(2):134–141, Jan 2007. doi: 10.1093/bioinformatics/btl565.

[231] Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., and Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, 26(17):2204–2207, Sep 2010. doi: 10.1093/bioinformatics/btq351.

[232] Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2):121–132, Feb 2014. doi: 10.1038/nrg3642. URL `http://dx.doi.org/10.1038/nrg3642`.

[233] Andrews, S. FASTQC – A quality control tool for high throughput sequence data, 2014. URL `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`. [Online; accessed 28-September-2014].

[234] Bolger, A. M., Lohse, M., and Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, Aug 2014. doi: 10.1093/bioinformatics/btu170.

[235] Quality Metrics, 2012. URL `https://genome.ucsc.edu/ENCODE/qualityMetrics.html#definitions`. [Online; accessed 28-September-2014].

[236] Kundaje, A., Jung, L. Y., Kharchenko, P., Wold, B., Sidow, A., Batzoglou, S., and Park, P. ENCODE tools, 2014. URL `https://code.google.com/p/phantompeakqualtools/`. [Online; accessed 28-September-2014].

[237] Yan, H., Evans, J., Kalmbach, M., Moore, R., Middha, S., Luban, S., Wang, L., Bhagwate, A., Li, Y., Sun, Z., Chen, X., and Kocher, J.-P. A. HiChIP: a high-throughput pipeline for integrative analysis of ChIP-Seq data. *BMC Bioinformatics*, 15(1):280, 2014. doi: 10.1186/1471-2105-15-280.

[238] Planet, E., Attolini, C. S.-O., Reina, O., Flores, O., and Rossell, D. htSeqTools: high-throughput sequencing quality control, processing and visualization in R. *Bioinformatics*, 28(4):589–590, Feb 2012. doi: 10.1093/bioinformatics/btr700. URL `http://dx.doi.org/10.1093/bioinformatics/btr700`.

[239] Li, Q., Brown, J. B., Huang, H., and Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Annals of Applied Statistics*, (5):1752–1779, 2011.

[240] Kundaje, A. TF ChIP-seq peak calling using the Irreproducibility Discovery Rate (IDR) framework, 2013. URL `https://sites.google.com/site/anshulkundaje/projects/idr`. [Online; accessed August 2013].

[241] Diaz, A., Park, K., Lim, D. A., and Song, J. S. Normalization, bias correction, and peak calling for ChIP-seq. *Statistical Applications for Genetics and Molecular Biology*, 11(3):Article 9, 2012. doi: 10.1515/1544-6115.1750.

[242] Schneider, I. Cell lines derived from late embryonic stages of Drosophila melanogaster. *Journal of Embryology and Experimental Morphology*, 27:353–365, 1972. ISSN 0022-0752. doi: VL-27.

[243] Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. a., Frampton, G. M., Sharp, P. a., Boyer, L. a., Young, R. a., and Jaenisch, R. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America*, 107(50):21931–6, December 2010. doi: 10.1073/pnas.1016071107.