

Max Planck Institute of Immunobiology and Epigenetics

Computational epigenomics study of the Male-Specific Lethal complex in flies and mammals

INAUGURAL-DISSERTATION

Author:
Vivek Bhardwaj

Supervisor:
Dr. Asifa Akhtar



to obtain the Doctoral Degree at the

FACULTY OF BIOLOGY

ALBERT-LUDWIGS-UNIVERSITÄT FREIBURG IM BREISGAU

21 September, 2018

Dekan der Fakultät für Biologie: Mr. Decan
Promotionsvorsitzender: Dr. Someone
Betreuer der Arbeit: Dr. Asifa Akhtar
Referent: Dr. Asifa Akhtar
Koreferent: Dr. Thomas Manke
Drittprüfer: Dr. Giorgos Pyrowolakis
Datum der mündlichen Prüfung:

Acknowledgements

This thesis would not be complete without the help and support of people from both the Akhtar Lab and Manke group at the MPI-IE. I am grateful to both my supervisors Dr. Asifa Akhtar and Dr. Thomas Manke for trusting my abilities and allowing me to work with them. Asifa provided me with all the independence I desired to work on my projects and to collaborate with the people in the lab. Discussions with her provided me the motivation and ability to see the significance of the work I was doing. Thomas has always been available for me to discuss all sorts of issues and provided me with the support and encouragement I needed to improve my skills and prepare for the next steps in my career. Opting for this co-supervision was the best choice I could have made for my PhD.

All of my works presented here have been done in collaboration, and would not be completed without the efforts and insights of those involved. I would like to thank :

Fidel, who taught me HiC analysis and mentored me during our collaboration on the HiC project. Due to HiCExplorer and deepTools projects, I was pushed out of my comfort zone to learn Python and to adopt reproducible workflows for my analysis. I gladly cherish all the discussions with Fidel during work as well as during our Sunday morning run. Giuseppe, for initiating the wonderful collaborations on the Drosophila projects, and for all the helpful discussions over the years. Raed, for working with me on the mammalian MSL2 project. Ibrahim and Tugce for initiating the collaboration on the mammalian MLE project, and finally Bilal and Ken for other independent projects. I would like to thank all the people at the Bioinformatics Unit for providing me with such a wonderful and interactive working environment, useful discussions, and fun collaborations. The deep-sequencing unit for producing the great quality data, and people in the Akhtar lab for all the discussions as well as fun activities outside work. Finally I want to thank my TAC members Ritwick Sawarkar and Michael Stadler for the support and discussions during my TAC.

Contents

List of Figures	v
Abbreviations	viii
1 Introduction	1
1.1 Multiple scales of genomic regulation	2
1.1.1 Local scale : DNA elements, histone marks and transcription factors	2
1.1.2 Global scale : Loops, TADs, and compartments	3
1.1.3 Mammalian X-inactivation : an interplay between multiple levels of epigenetic regulation	4
1.2 Regulation of transcription by the MSL complex	5
1.2.1 Dosage compensation in flies via the MSL complex	5
1.2.2 Functions of MOF and MSL complex in mammals	7
1.3 High-throughput techniques to study transcriptional regulation	7
1.3.1 ChIP-Seq : current standards and challenges	8
1.3.2 Analysis of chromosome conformation	9
1.3.3 Transcript profiling methods and analysis	11
1.4 Aims	14
2 Results and Discussion	15
2.1 Insights from high-resolution chromosome conformation analysis in flies	15
2.1.1 Setup of a HiC analysis workflow	16
2.1.2 Relationships between TADs, regulatory elements, and transcription in flies	17
2.1.3 Resources for HiC analysis and visualization	18
2.2 Using promoter-profiling to study MSL-mediated dosage compensation in flies	20
2.2.1 Development of the MAPCap protocol	20
2.2.2 Development of the icetea bioconductor package	21
2.2.3 Transcriptional changes during dosage compensation defects in flies .	22
2.3 Integrating transcriptomic and epigenomic analysis to study MSLs in mammals	24
2.3.1 The role of MSL complex member MLE in mammals	24
2.3.2 The role of MSL complex on active and inactive mammalian X chromosomes	24
2.3.3 A toolkit for integrative epigenomic analysis	27
2.4 Conclusion and Outlook	27
A Publications and Manuscripts	29
A.1 Analysis of chromosome conformation in flies	29

A.2 Galaxy HiCExplorer	67
A.3 Analysis of dosage compensation in flies via promoter-profiling	74
A.4 Interaction of MLE ortholog DHX9 with Alu elements in the human genome	98
A.5 Update of the deepTools toolkit for exploring deep-sequencing data	126
A.6 snakePipes enables reproducible epigenomic analysis	133
B Supplemental information	151
C Academic Vita	153
References	155
C.1 References	155

List of Figures

1.1	Regulation of gene expression by various genetic and epigenetic elements	2
1.2	Proposed functions of TADs in the genome	3
1.3	Functions of the MSL complex in flies and mammals	6
1.4	Visualization of a normalized HiC contact matrix	10
1.5	Overview of popular promoter-profiling methods and the resulting data	13
2.1	A directed acyclic graph (DAG) of the HiC workflow	16
2.2	The HAS interaction network on X-chromosome	18
2.3	Chorogenome navigator	19
2.4	Improvement in TSS detection over different versions of the MAPCap protocol	20
2.5	Effect of RNA composition on TSS enrichment	21
2.6	TSS expression (MAPCap) and H4K16ac levels (ChIP-seq) on male and female embryo brains	23
2.7	Integrative analysis of ChIP-Seq, RNA-Seq and ATAC-Seq in MSL2 knockout (KO) NPCs	25
2.8	A directed acyclic graph (DAG) of the allele-specific RNA-seq analysis workflow	26

Summary

In various species, sex determination is associated with an imbalance in the number of sex chromosomes between males and females. In *Drosophila*, this imbalance is corrected by an chromosome-level epigenetic phenomenon resulting in the upregulation of gene expression on the single male X chromosome. This phenomenon, referred to as dosage compensation, requires the Male-specific lethal (MSL) complex. The 3D conformation of the X-chromosome guides the spreading of the MSL complex, depositing histone (H4K16) acetylation on genes. On the other hand, mammalian dosage compensation occurs via X inactivation in females, and the role of MSL complex in mammals is poorly understood.

The aim of my project was to provide insights into the functions of the MSL complex in flies and mammals through computational epigenomics approach. This involved development of methods and tools for analysis of chromosome conformation and promoter-profiling data, and their integration with other transcriptomic and epigenetic data. Softwares such as HiCExplorer, icetea and deepTools2 were developed to facilitate this, and workflows for reproducible analysis were implemented as part of the snakePipes package. Application of our methods on HiC data in flies revealed that chromatin domains influence gene transcription, and validated previous observation on clustering of MSL2 sites in 3D space. Analysis of the MSL complex member MLE through promoter profiling identified a catalogue of MLE sensitive and insensitive promoters at the male X-chromosome and the difference in MLE action between sexes. Analysis of mammalian ortholog of MLE revealed it's novel function in regulation of Alu transposons independent of the MSL complex, while the analysis of other members of the mammalian MSL complex showed that MSLs are involved in transcriptional regulation of genes on X chromosome in an allele-specific manner.

In summary, we gained new insights into the functions of the MSL complex through analysis and integration of multiple epigenomic and transcriptomic data. This study would motivate further studies utilizing integrative epigenomics to understand the functions of MSLs as well as other regulators of transcription.

Abbreviations

3C	Chromosome conformation capture
3D	3-dimention
4C	Chromosome conformation capture on chip
5C	Chromosome conformation capture carbon copy
ac	acetylation
CAGE	Cap Analysis of Gene Expression
cDNA	complementary DNA
CHART-seq	Capture hybridization analysis of RNA targets and sequencing
ChIP-seq	Chromatin Immunoprecipitation and sequencing
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CTCF	CCCTC-binding factor
CTSS	CAGE Tag Start Sites
DHS-seq	DNase I Hypersensitive Site sequencing
DNMT	DNA MethylTransferase
ENCODE	ENCyclopedia Of DNA Elements
ESC	Embryonic Stem Cells
GAM	Genome Architecture Mapping
GO	Gene Ontology
GRO-seq	Global Run-On sequencing
GUI	Graphical User Interface
HAS	High-Affinity Sites
HMM	Hidden Markov Model
HxKx	Histone (position) Lysine (position)
ICE	Iterative Correction and Eigenvector decomposition
icetea	Integrating Cap Enrichment and Transcript Expression Analysis
ICM	Inner Cell Mass
KO	Knock-Out
lncRNA	long non-coding RNA
MAPCap	Multiplexed Affinity Purification of Capped RNA
me	methylation
MLE	maleless
modENCODE	Model organism ENCODE
MOF	Males-absent On the First
mRNA	messenger RNA
MSL	Male-Specific Lethal
NGS	Next Generation Sequencing
NSL	Non-Specific Lethal

PCA	Principal Component Analysis
PRC	Protein Regulator of Cytokinesis
RAMPAGE	RNA Annotation and Mapping of Promoters for Analysis of Gene Expression
RAP	RNA Affinity Purification
RELACS	Restriction Enzyme-based Labeling of Chromatin in Situ
REs	Restriction Enzymes
RF	Restriction Fragment
RLE	Relative Log-Likelihood
RNA-CLIP	RNA Cross-Linking ImmunoPrecipitation
RNA-seq	RNA sequencing
RNAP-II	RNA polymerase – II
RNAP-III	RNA polymerase – III
SNPs	Single Nucleotide Polymorphisms
SPRITE	Split-Pool Recognition of Interactions by Tag Extension
TADs	Topologically Associating Domain
TFs	Transcription Factors
TMM	Trimmed Mean of M
TSSs	Transcription Start Sites
ub	ubiquitination
UMIs	Unique Molecular Identifiers
UV	Ultra-Violet
UV-CLAP	UV CrossLinking and Affinity Purification
WGBS	Whole-Genome Bisulfite Sequencing
XCI	X-Chromosome Inactivation
Xic	X-Inactivation Center
Xist	X-Inactive Specific Transcript

1. Introduction

Proteins and Nucleic acids (DNA and RNA) are the fundamental building blocks of all life on earth. The famous “central dogma” of life, proposed by Francis Crick ¹ states that there is a sequential flow of information from nucleic acids to proteins, and not backwards. No exceptions to the crick’s definition of central dogma has been observed till date. However, research done before and after the proposal of central dogma has established that this flow of information could be regulated, either through nucleic acids themselves, or through proteins.

A “gene” is the major component of central dogma. The overall definition of a gene has been debated and revised to include protein-coding genes, regulatory elements such as “promoters” and “enhancers”, and regions that produce non-coding RNAs ^{2|| 3}. A protein-coding gene transfers it’s encoded information into the protein product, through “transcription”, and “translation” (Appendix B). DNA elements that regulate this flow of information (or “expression”), such as promoters or enhancers, are stably inherited through generations and contribute to evolution ⁴. Apart from DNA elements, other factors such as semi-stable modifications to DNA elements (eg. methylation), non-coding RNAs (eg. antisense RNA ⁵, micro-RNA ⁶, long non-coding RNA ⁷ etc.), proteins that interact with DNA (histone marks, transcription factors), and proteins that interact with mRNA ⁸ also regulate gene expression. Many of these gene regulatory mechanisms are established either in response to environmental changes or through genetic programming, and in some cases, could be stably maintained through cell cycle, or across multiple generations. Such mechanisms are collectively studied under the term “epigenetics” ⁹.

In order to understand the regulation of gene expression, it’s important to study how various genetic and epigenetic regulators interact with each other. Advances in next generation sequencing (NGS) ^{10,11} have allowed us to study these interactions at multiple scales, ranging from local gene neighborhood to overall organisation of a chromosome. In the following sections, I would discuss the various scales of gene regulation, as well as current state of the art techniques to study them.

1.1 Multiple scales of genomic regulation

1.1.1 Local scale : DNA elements, histone marks and transcription factors

The “selected effect” definition of biological function ¹² implies that all functional elements should be under some sort of selective pressure. The primary evolutionary analysis of mammalian genome therefore proposed ~5% of human-mouse genome to be functional ¹³. This estimate, which was already about 4-times larger than the protein-coding fraction ¹⁴, was further expanded in later analysis ^{15,16}. Genome-wide analysis have also proposed “biochemical function” for many new genetic elements ¹⁷. Most of these non-coding DNA elements such as promoters, enhancers, silencers, insulators and transposons, have been classically known to be involved in regulation of protein-coding genes ¹⁸ (Table-1).

It has become increasingly clear that DNA elements do not act alone, but rather within the influence of the histone code ¹⁹. In fact various evidences suggest that gene regulation happens via a cross-talk between transcription factors, DNA elements, DNA modifications (such as methylation) and histone marks. For example, during embryonic development, pluripotency genes Oct3/4 are switched off in a multi-step process where first a repressor binds to it’s promoter, turning off transcription, followed by recruitment of histone methyltransferase and histone deacetylase enzymes leading to transcriptional repression and H3K9 methylation. This follows binding of DNMT3A and DNMT3B methyltransferases, leading to promoter methylation and ultimately gene suppression ^{20,21}. Apart from this cross-talk, the physical distance between regulatory elements and their targets also influence gene expression. Chromatin marks such as H3K9 methylation could spread into large domains, while TFs are suggested to induce looping of chromatin in order to achieve spatial proximity to their target genes ²². This looping is regulated via insulators, adding additional layer to gene regulation.

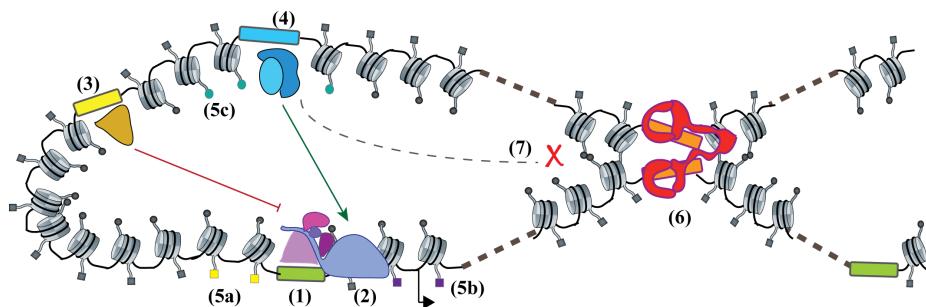


Figure 1.1: Regulation of gene expression by various genetic and epigenetic elements A promoter (1) recruits RNAP-II (2) to perform transcription of target genes. Elements such as repressors (3) or enhancers (4) could recruit transcription factors that either suppress or activate gene expression. All these elements are marked by various histone marks, such as H3K4me3 at promoters (5a), H3K36me3 at gene bodies (5b) or H3K27ac at enhancers (5c), which facilitate gene expression. Insulators (6) recruit proteins that facilitate DNA looping, and block transcription factors from activating genes outside of the established boundaries (7).

1.1.2 Global scale : Loops, TADs, and compartments

Looping of DNA has been shown as the mechanism behind long-range enhancer-promoter interactions in early 1990s²³. Since then, various techniques have been developed to study long-range interactions via chromatin looping (see section on analysis of chromosome conformation). Genome-wide derivation of the chromosome conformation capture (3C) technique, called HiC²⁴ is currently the most popular amongst those. Early HiC studies were limited by their resolution, and therefore discovered higher-order chromatin structures called compartments²⁴ and chromatin domains, referred as topological associated domains (TADs)²⁵. Later studies revealed that the chromatin is hierarchically organized, where the domains with similar chromatin signature are spatially clustered. At the highest level of clustering, genome can be divided into A and B compartments, which separate inactive and active chromatin in the cell²⁴. Compartments could further be segregated into subcompartments that represent clustering of histone marks²⁶. TADs serve as the next (lower) level of this segregation, while enhancer-promoter loops are at the lowest level. Functionally, TADs were shown to act as “regulatory units” or “insulated neighbourhoods”, limiting enhancer-promoter interactions and the spread of chromatin marks²⁷. This property of TADs resembles those of insulators, and in fact several classically studied insulator proteins were found to be associated with TAD boundaries²⁸. Interestingly, unlike loops, which show cell-type specific interactions²⁹, TADs have been indicated to be cell-type invariant and evolutionarily conserved^{25,30}.

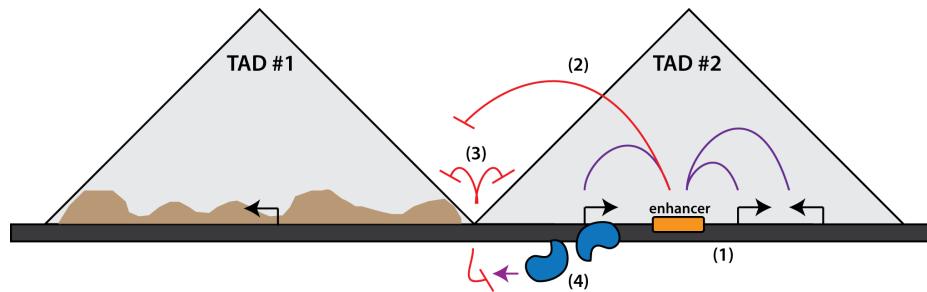


Figure 1.2: **Fig 2. Proposed functions of TADs in the genome.** TADs serve as regulatory units by providing insulated neighbourhoods for enhancers to activate their target genes (1), blocking the enhancers from activating the off-target genes (2), regulating the spread of chromatin marks in cis (3) and blocking the antisense transcription from encroaching into nearby genes (4). Figure inspired from the review by Dixon et al. (2016)²⁷

Due to high-paced technology development, last few years have seen a fascinating advance in our understanding of mechanism of loop and TAD formation. The first study performed using in-situ HiC visualized loops as regions of enriched contacts between long distance loci, and identified the insulator element CTCF, on loop anchors²⁶. The loop anchors associated with cohesin subunits RAD21 and SMC3, along with CTCF motifs oriented in convergent direction corroborating previous evidences on the role of CTCF and cohesin in mediating DNA looping³¹. Several models of loop formation have been proposed³². However, the “extrusion-model” proposing loop formation via DNA loop-extruding smc complexes (cohesin and condensin) has received most evidence, and this model could also explain the 3D genome structure observed with the HiC data^{33–35}. Also, the *in-silico* models produced by the Mirny

lab have proposed loop extrusion as a common mechanism behind formation of loops, TADs and chromosomal compartments^{36–38}. Despite all the *in-silico* evidence, the *in-vivo* evidence linking extrusion to loop formation was lacking, until the studies perturbing loop-associated factors emerged. A depletion of CTCF eliminated CTCF associated loops in the genome³⁹, while the depletion of cohesin eliminated all loop domains⁴⁰. A CRISPR mediated knockout (KO) of the cohesin release factor WAPL showed that the duration of association of cohesin complex (SCC2/SCC4) determines the length of loops⁴¹. These studies reinforced a common mechanism behind loop and TAD formation. However, interestingly they also showed that compartments as well as histone marks remain unaffected upon loop depletion, suggesting an independent mechanism behind segregation of compartments. Another interesting point about studies linking cohesin and loop extrusion is that they don't explain how cohesion physically extrudes chromatin, since no *in-vivo* extrusion activity of cohesin complex has been visualized⁴². Another smc associated protein, condensin, however, has been shown to have the motor activity⁴³ and recently, its *in-vivo* loop extrusion activity has also been visualized⁴⁴.

Single molecule imaging suggests that although cohesin can undergo rapid diffusion along DNA, its mobility is highly restricted by nucleosomes, and the DNA motor proteins can readily push cohesin along DNA⁴⁵. On the same line, transcription has been proposed as a mechanism behind cohesin mediated loop extrusion⁴². HiC studies showed that housekeeping genes are enriched at TAD boundaries^{25,46} providing further evidence linking transcription and TADs. A study last year showed that TADs could emerge in absence of transcription in flies, challenging the speculation of a causal link between transcription and TAD formation⁴⁷, however another recent study showed that transcription could affect 3D genome structure by displacing cohesin from CTCF sites⁴⁸. Therefore, more research to investigate the association between transcription and 3D genome would be required.

1.1.3 Mammalian X-inactivation : an interplay between multiple levels of epigenetic regulation

The process of mammalian X-inactivation (XCI) serves as an excellent example of how multiple levels of epigenetic regulations act in synchrony and influence each other. mammals (as well as *Drosophila*) have XX-XY form of sex determination, leading to an imbalance of X-chromosome gene dosage between sexes. This imbalance is corrected by an epigenetic phenomenon (dosage compensation) where one of the female X-chromosome is randomly inactivated during differentiation⁴⁹. In mouse, XCI happens in two waves, an imprinted paternal XCI is established at 4-8 cell stage, and the paternal X remains inactive in trophectoderm but is reactivated in the inner cell mass of the blastocyst (ICM)⁵⁰. This is followed by the random XCI, which coincides with the down regulation of pluripotency factors such as Nanog, Oct4/Pou5f1 and Sox2⁵¹. The embryonic stem cells (ESCs) derived from the ICM, serve as a good model to study XCI, since XCI can be achieved simply by inducing ESC differentiation⁵².

The X-inactivation center (Xic) is a locus on X-chromosome which is required to trigger XCI⁵³. Xic contains the non-coding RNA Xist, which coats the inactivating X-chromosome during XCI. Xist is negatively regulated by its antisense RNA Tsix, or positively regulated through a ubiquitin ligase Rnf12, in a dosage sensitive manner⁵⁴, both of which are controlled by the pluripotency factors^{55,56}. On Xic, multiple putative regulators of Xist, such as Rnf12,

Jpx, Ftx and Xpr have been shown to be important for specific targeting of X chromosome for the inactivation process (referred as “sensing”) ⁵⁷. The mechanism behind selection of one of the two X-chromosomes for inactivation (referred as “choice”) remains elusive, although in mouse, Tsix has been suggested to be important for the process. One proposed mechanism of Tsix action could be that transient pairing of X-chromosome during ESC differentiation ⁵⁸ somehow leads to an asymmetry in Tsix expression, followed by recruitment of repressive chromatin marks ⁵⁹, which in turn leads to asymmetry in Xist expression.

During XCI, the Xist RNA coats the X-chromosome, which results in depletion of RNAP-II and the activating histone marks, such as H4 acetylation ⁶⁰. The Xist coated chromatin then gets enriched in repressive chromatin marks such as H2AK119ub, H3K9me and H3K27me3 ^{61,62}. It’s been proposed that Xist silencing followed by H2AK119ub could in turn recruit PRC1, which could indirectly recruit PRC2 through Jarid2, bringing in H3K27me3 ⁶³. This rapid change of chromatin state is accompanied by a change in 3D conformation. The X-linked genes initially reside in nuclear periphery, outside the repressive nuclear compartment, and get recruited into it during inactivation ⁶⁴. Ultimately, the inactive-X is condensed to a heterochromatic form, known as the barr body ⁶⁵. The mechanism of Xist spread has been investigated using RNA immunoprecipitation (RAP and CHART-seq) techniques, which suggested that Xist could exploit the chromosome conformation to spread to regions in spatial vicinity ^{66,67}. Due to this intricate interplay of chromatin, non-coding RNA and chromosome conformation, that recent studies studying XCI have turned to multi-assay epigenomic techniques ^{68,69}. It’s evident that similar integrative studies would be required in future to study XCI and other such multi-stage phenomena.

1.2 Regulation of transcription by the MSL complex

1.2.1 Dosage compensation in flies via the MSL complex

In *Drosophila*, dosage compensation is achieved by the upregulation of a single X-chromosome in males ⁷⁰. Males absent on first (MOF), an enzyme that specifically deposits acetylation marks on Histone H4 Lysine 16 (H4K16ac) is required for the process ⁷¹. MOF associates with the male-specific lethal (MSL) complex, containing proteins MSL1, MSL2, MSL3, MLE (maleless) RNA helicase, and non-coding RNAs roX1 and roX2. roX1 and roX2 are lncRNAs containing stem-loop structures which are expressed from the X-chromosome and seem to play a redundant role in dosage compensation ⁷². The E3 ubiquitin ligase MSL2 is only expressed in males and, in association with roX RNAs, provides specificity of the complex towards X-chromosome ^{73,74}. Binding specificity to certain DNA motifs by the MSL2 CXC domain has been shown to bring X-specificity ⁷⁵. MSL1 serves as a scaffold for binding to other proteins in the complex, and plays essential role in the assembly of the complex via its homodimerization and DNA binding property ⁷⁶. MSL3 enhances the acetylation activity of MOF ⁷⁷ and contains a chromodomain which facilitates the spread of the complex towards gene ends ⁷⁸. Finally, the MLE helicase stabilizes the roX RNAs and is important for its incorporation into the complex ⁷⁹.

Studies have shown that following this multi-stage assembly process, the complex is first recruited to certain specific loci in the genome referred to as high-affinity sites (HAS), followed by further spread to nearby loci ⁸⁰. Sites of roX RNA expression serve as strong

high-affinity sites. Further, the complex decorates the X-chromosome and deposits H4K16ac on promoters, gene bodies, as well as intergenic regions. The increase in H4K16ac has been shown to enhance transcription⁸¹. Multiple studies have investigated the precise mechanism of this transcriptional enhancement using genome-wide techniques such as ChIP-seq⁸², GRO-seq⁸³ and nascent RNA-seq⁸⁴. These evidences point out that transcriptional upregulation is a joint result of increased RNAP-II recruitment as well as processivity.

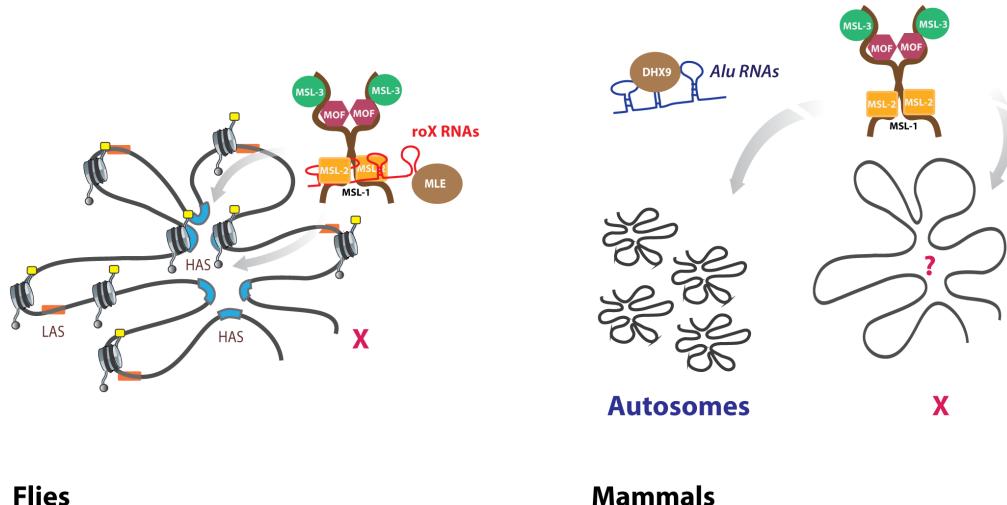


Figure 1.3: Functions of the *MSL* complex in flies and mammals.
In flies, the *MSL* complex assembles in presence of the *roX* RNAs, whose secondary structures are recognized and resolved by the maleless helicase (*MLE*), the complex then targets the high-affinity sites (*HAS*, blue) containing *roX*. The *HAS* cluster in 3D space, allowing efficient targeting by *MSL*s. The complex then spreads to low affinity sites (*LAS*, orange), depositing H4K16ac marks (yellow) on the *X*. In *mammals*, *roX* RNA is absent and the complex is believed to assemble in absence of any RNAs, and targets both *X* and autosomal sites. Specificity towards *X*-chromosome, as well as the role of chromosome conformation is unknown so far. *MLE* homolog *DHX9* associates with a separate class of RNAs and performs independent functions (presented in results). Figure inspired from⁸⁸

There are interesting parallels between *Drosophila* and mammalian dosage compensation systems. Both systems rely on an lncRNA expressed from the X-chromosome, which seem to be important for the recognition of the chromosome (**sensing**), and both of them rely on a global deposition of histone marks which associates with suppression or hyperactivation of genes. In *Drosophila* however, the problem of **choice** seems to be simplified, since the compensation occurs on the single male X-chromosome and the male-specific expression of *MSL2* assures the sex-specificity. Nevertheless, this similarity suggests that questions in one species could be tackled by learning from another. For example, in parallel with the role of chromosome conformation in guiding the spread of *Xist* RNA, research performed in our lab and by others have shown that the spread of the *MSL* complex on the X-chromosome is guided by the 3D spatial proximity of *HAS*^{85,86}. These observations also lead to further interesting questions. For example, studies have shown that despite a global inactivation of X-chromosome, about 15% of genes in humans and a 3% in mice, escape inactivation (referred

to as escapees)⁸⁷. Escapee biology has been heavily studied in mammals, however studies to investigate whether different genes might also respond differently to dosage compensation has been lacking in flies.

1.2.2 Functions of MOF and MSL complex in mammals

Proteins of the MSL complex are conserved from flies to humans, and evidence suggests that MOF is responsible for majority of H4K16ac⁸⁹. Human homologs of MSL1, MSL2, MSL3 and MOF have been co-purified, suggesting that the bulk of the MSL complex stays together. It's compelling to speculate that role of MSL complex along with the role of H4K16ac might have diverged, considering that mammalian dosage compensation works via a different mechanism, and no ortholog of roX RNAs have been discovered on the X-chromosome so far⁹⁰. However, there is evidence that the active X-chromosome in mammals might also be upregulated^{91,92} and a recent study has supported the role of MOF mediated H4K16ac in enhancing transcription on a subset of X-linked genes⁹³. Apart from its association with dosage compensation, MOF mediated H4K16ac have also been implicated in embryogenesis and oncogenesis^{90,94}. However, this should be noted that MOF is also associated with another complex, called the non-specific lethal (NSL) complex, which is conserved from flies to humans, and is involved in depositing H4K16ac on housekeeping and constitutively expressed genes⁹⁵. Therefore some functional outcomes of H4K16ac could relate to the MOF-MSL complex, while others to the MOF-NSL complex^{96,97}.

In flies, the MOF-NSL complex is shown to be responsible for the bulk of H4K16ac on the gene promoters, while the MSL complex deposits the mark on gene bodies of the X-chromosome. A recent study from our lab showed that in mammals, however, MSL and NSL proteins colocalize on promoters of target genes on both X and autosomes⁹⁶. The study also showed that the MSL mediated H4K16ac contributes to the regulation of Tsix expression and the depletion of MSLs before ESC differentiation leads to chaotic XCI. However, the lack of specificity towards the X-chromosome and co-localization with NSL proteins at the promoters raises further questions about the specific genome-wide role of MSL complex, and its role on the mammalian X-chromosome.

Apart from the MSL complex, the human ortholog of MLE helicase, DHX9, is also conserved in mammals, and has been shown to unwind both DNA and double-stranded RNA^{98–100}. However, in absence of it's association with MSL complex, the function of it's RNA helicase activity in mammals had not been studied in detail. A study presented in this thesis investigates the RNA binding partners of DHX9 in mouse and humans, and proposes a novel evolved function of DHX9 in suppression of undesired expression of RNAs¹⁰¹.

1.3 High-throughput techniques to study transcriptional regulation

Due to consistently reducing cost of sequencing, increase in our ability to multiplex many samples, and sequence genetic material from as low as a single cell, a large number of biologists now rely on genome-wide assays for their research. Researchers have quickly realized that the ability to store, process, analyze, interpret and visualize such datasets

would become a major challenge for biology¹⁰². Below I would focus on recent developments in techniques employed in study of protein-DNA interactions, chromosome conformation and transcription.

1.3.1 ChIP-Seq : current standards and challenges

Chromatin Immunoprecipitation followed by sequencing (ChIP-seq) is the broadly used method to detect genome-wide protein-DNA interactions in cells¹⁰³. ChIP-seq allows us to probe genome-wide binding sites for TFs, or study the distribution of histone marks. It has led to numerous discoveries and has been used by large consortia such as ENCODE, to map cell-type specific regulatory elements such as enhancers in the human genome¹⁷. ChIP-seq experiment involves DNA protein cross-linking (using UV or formaldehyde), chromatin fragmentation (using sonication or enzymes), purification of protein-DNA crosslinks with or without antibody (called “ChIP” and “Input” DNA), and high-throughput sequencing of associated DNA. Since the technique involves cross-linking, shearing, and affinity based pull-down, the data obtained from the protocol is directly affected by three major factors : 1) **cross-linking efficiency** of proteins might be affected by its nature of DNA binding (direct vs indirect, transient vs stable) 2) **Stability** of the protein epitope under shearing and washing conditions would affect enrichment and 3) **Antibody cross-reactivity** would result in detection of non-specific interactions^{104,105}. Variations of ChIP-seq protocols as have been developed to tackle biases, such as MNase-ChIP¹⁰⁶, to account for shearing and stability bias, and DamID-seq¹⁰⁷ to account for antibody bias. Further variants such as iChIP¹⁰⁸, Mint-ChIP¹⁰⁹ and RELACS¹¹⁰ aim to increase the multiplexing ability and allow reducing the number of cells. Such methods might receive more popularity in future.

The analysis of ChIP-seq data involves mapping of sequenced DNA to a reference genome, and detection of regions enriched in ChIP DNA (called as “Peaks”) using Input DNA as a negative control. Peaks represent the “detected” binding sites of the TF or histone mark of interest. These peaks are then used to find DNA binding motifs for TFs, detect target genes of interest, and integrate these results with other assays (such as RNA-seq) to understand gene regulation¹¹¹. Apart from the sources of bias described above, the sequencing depth or genomic nucleotide content (GC bias) could lead to sample-specific bias in ChIP-seq, and various “normalization” methods have been developed to account for such biases¹¹²⁻¹¹⁵. Our group previously developed “deepTools”¹¹⁶, a set of tools that allow users to identify biases and perform such normalizations, and visualize the downstream results in a user-friendly manner. However, due to improvements in experimental and analysis methods, new challenges have emerged as previous issues have been resolved. Specifically, increase in multiplexing ability has simplified performing routine large-scale experiments¹¹⁰, and the demand for user-friendly tools for differential ChIP-Seq analysis between conditions¹¹⁷, quantitative normalization using external controls^{118,119}, and large-scale processing and integration of datasets has been increasing. The development of tools that allow multi-sample comparison and integration of quality control and analysis would be widely appreciated. This was the motivation behind our recent update of deepTools¹²⁰.

1.3.2 Analysis of chromosome conformation

Past 10 years have seen lots of development in genome-wide techniques to capture physical interactions between chromatin loci inside the cell. The first popular measure to detect chromatin interactions between loci was called **3C** (chromosome conformation capture). 3C involved cross-linking the chromatin such that the regions in spatial proximity inside the nucleus get linked to each other followed by the digestion of chromatin using restriction enzymes (REs) and the ligation of close-by DNA fragments. These ligated fragments could then be amplified using PCR primers specific to the region of interest, followed by qPCR to measure interaction frequency¹²¹. 3C was extended to 4C, where all regions interacting with a locus of interest (referred to as “bait”) could be detected by inverse PCR using outward facing primers on the 3C-ligated product, followed by hybridization to a microarray chip to detect interaction¹²². The 5C technique introduced later used highly multiplexed ligation-mediated amplification (LMA) to first “copy” and then amplify parts of the 3C library followed by detection via microarrays or sequencing¹²³. Finally the genome-wide version of 3C, called HiC, involved biotin-based pull down of 3C fragments, followed by library preparation and paired-end sequencing of the enriched fragments²⁴. An improved version of HiC protocol performs restriction digestion and ligation of 3C fragments inside intact nuclei, avoiding non-specific interactions. This technique, referred as *in-situ* HiC²⁶ is currently the most popular method for genome-wide detection of chromatin contacts.

All the 3C-based methods however, suffer from two issues : 1) The cross-linking of chromatin makes the technique less quantitative and the proximity ligation introduces additional biases; 2) These methods only allow detection of pair-wise contacts between two genomic bins. Due to its limitations and biases, together with the observed discrepancy between 3C and microscopy-based visualization of interactions, the interpretation of 3C-based proximity as molecular proximity has been criticised^{124–126}. Recent innovations in chromosome conformation techniques, such as GAM¹²⁷, SPRITE¹²⁸ and damC¹²⁹ and have tried to overcome these limitations. SPRITE is particularly interesting since it detects multi-loci chromatin interactions via a relatively simple split-pool barcode tagging of the cross-linked chromatin fragments, and can also detect DNA-RNA interactions¹²⁸. It’s not unlikely that one of these new innovations might replace *in-situ* HiC as a primary method for analysis of 3D genome.

HiC libraries are sequenced as paired-end, where ideally, one read in a pair comes from one of the interacting loci in the genome while the second read comes from another²⁴. However, due to the issues with ligation efficiency, biotinylation and IP of the interacting pairs, many “invalid” HiC pairs also get sequenced. Analysis of HiC data begins with separate **mapping** of the two reads in a pair, followed by pairing and **filtering** of reads. To consider a pair “valid” two reads should come from different fragments which were ligated during the protocol, and therefore should have a restriction site in between them. After filtering of valid pairs, the genome is divided into bins, and the number of paired-end contacts between all bins are calculated to create a HiC matrix.

Initial HiC protocol used whole cell lysis along with REs such as **Hind-III**, with low cutting frequency, leading to low-resolution HiC matrices. In contrast, the use of REs with high cutting frequencies such as **Dpn-II** (4-cutters) combined with other improvements, the recent *in-situ* HiC protocols has allowed us to create higher resolution matrices. In either case, the highest resolution matrices could be obtained by using restriction enzyme cut-sites

as “bins” in the genome^{85,130}.

Several known and unknown factors in the HiC protocol could affect interpretation of the observed data. Normalization methods have been developed to explicitly account for some biases, such as the effective length bias (due to the size of the restriction fragment), GC content bias, and the mappability bias^{131,132}. However since not all source of biases are known in the experiments, the matrix balancing methods, which try to account for both known and unknown biases, have been most popular¹³³. Matrix balancing methods, such as ICE (Iterative Correction and Eigenvector decomposition)¹³⁴, and Knight and Ruiz method^{26,135}, assume uniform visibility of all genomic loci and therefore ensure equal sum of row and column values. Both methods have been shown to be useful in the analysis of HiC data.

HiC interaction matrices reveal that loci which are closer on the linear genome also tend to interact more frequently. In normal interphase cells, a distance dependent decay of interactions is observed, which follows a power law. This interaction is disrupted in mitotic chromosomes, which represent a reduction of short-range and an increase in long-range interactions¹³⁶. It also served as a quality control signal for the comparison of normal samples. However, the interaction frequency could also get exaggerated due to nonspecific proximity ligation effects, and therefore for some analysis, such as detection of chromosomal compartments, a correction of proximity ligation effect is needed in addition to ICE correction.

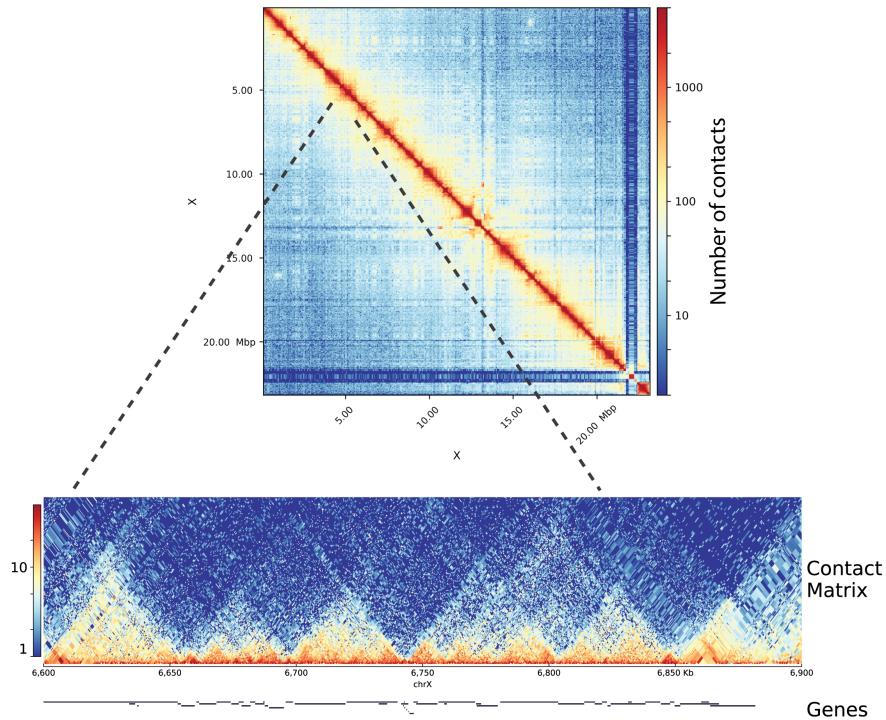


Figure 1.4: **Visualization of a normalized HiC contact matrix.** The HiC matrix (**top**) represents all-to-all contacts between bins in the genome, and is therefore symmetric. After cutting the matrix from diagonal and rotating it (**bottom**), domains (TADs) in shape of triangles become visible.

Normalized HiC matrices could be used for detection of TADs, loops, compartments, modeling of 3D genome and visualization of genomic interactions. Numerous software solutions have been developed over the last few years for the processing and analysis of HiC data that allow these applications (Table-2). Detection of loops and compartments have been formalized better than TADs due to their distinct properties. The original method for detection of chromosomal compartments²⁴ assumes that regions with shared neighbourhood have correlated interaction profiles. It therefore transforms the HiC matrix into a correlation matrix, followed by partitioning of chromosomes into regions of correlated interactions using principal component analysis (PCA). This method has now been implemented in multiple tools (Table-2). The method used for genome-wide detection of loops utilizes high-resolution in-situ HiC data and searches for clusters of matrix entries with enriched interactions relative to a local background. This method referred to as HiCCUPs²⁶, has been implemented in the Juicer pipeline¹³⁷.

Methods for detection of TADs have been evolving fast over the last few years. The first group describing TADs²⁵, observed that within a TAD the bins closer to the boundary have a high interaction bias towards bins in the middle. They developed a metric that quantified the contacts of a bin with a larger region upstream (A) and downstream (B) of that bin and compared it to the contact probability expected under a null distribution of $(A+B)/2$. This directionality index (DI) was used as an observation to infer the true directionality bias via a hidden markov model (HMM), which predicted the topological domains. Later, methods based on maximum likelihood segmentation¹³⁸, breakpoint detection¹³⁹, and hierarchical clustering¹⁴⁰ were developed. In principle, all these methods formalize the observation that TADs are visualized as domains of enriched contacts relative to their neighbourhood on a HiC matrix, and try to produce the most similar set of predictions corresponding to this visual observation. Multiple studies showed that genome organization is hierarchical, and TADs have been speculated to be the level of this hierarchy which are not structurally, but rather functionally privileged^{140,141}. Therefore ideal methodology for TAD detection has still not been settled, and the benchmarking of TAD calling methods rely on visual identification of true positives¹⁴².

1.3.3 Transcript profiling methods and analysis

Transcriptomics (study of all transcripts in the cell), has been one of the most popular fields where NGS is being applied. The most widely used bulk RNA-seq methods (studying a population of cells) are poly-A or ribo-depleted RNA-seq. In **poly-A** RNA-seq, oligos are used to enrich for mRNA (containing poly-A tails) while in **ribo-depleted** RNA-seq, all RNAs (except ribosomal RNA) are sequenced¹⁴³. Ribo-depleted RNA-seq allows for analysis of poly-A along with non poly-A RNA, such as unprocessed transcripts and RNAP-III transcribed RNAs. However, detection of all kinds of RNA requires a high depth of sequencing, the capacity of sequencing library to capture transcripts various lengths, and enough input material from the cells to avoid RNA composition effect (masking of lowly expressed transcripts by abundant RNA). Methods have been developed to enrich for specific RNA species in the cell, such as small (<50 nt) RNAs¹⁴⁴, transiently transcribed RNAs¹⁴⁵, and circular RNAs¹⁴⁶.

The analysis of RNA-seq data¹⁴⁷ includes **trimming** adaptors and low quality bases from single or paired-end reads, followed by **mapping** of the reads to a reference genome.

Genomic mapping of RNA-seq requires splice-aware alignments (allowing for splitting of reads mapping to splice junctions, in order to map across introns). In order to obtain the estimate of transcript numbers in the cells, counting and **summarizing** of reads is performed afterwards, using gene annotations as reference. Usually, RNA-seq is performed on multiple biological **replicates** of two or more groups (control vs knockdown, tissue 1 vs tissue 2 etc.), in order to find genes or transcripts that vary significantly between the groups (called **differential expression** analysis). Differentially expressed genes are used to look for functional enrichments of gene categories using methods like Gene Ontology (GO) analysis¹⁴⁸. Gene level differential expression analysis has been standardised in the field, while methods for quantification and differential expression of transcript isoforms are still under active research¹⁴⁹. A special case of differential expression is where a large number of transcripts are affected, such as knockdown of TFs regulating housekeeping genes, or genome-wide transcriptional activation/repression¹⁵⁰. These cases require normalization using external, non-changing controls referred to as **spike-ins**¹⁵¹. For special applications of RNA-seq such as detection of **circular RNAs**, the mapping and summarization of reads requires an entirely different strategy, and various tools have been developed to facilitate this¹⁵². Another special application of RNAseq is for the identification of **RNA editing**¹⁵³, which require stringent mapping and variant detection followed by multiple rounds of filtering. Due to evolving standards, these applications still rely on custom analysis.

Although RNA-seq has helped with genome-wide annotation of transcripts in the genome, it requires fragmentation of RNA, making it difficult to accurately detect transcription start sites (TSSs) and perform promoter profiling. Techniques for promoter profiling, such as Cap Analysis of Gene Expression (**CAGE**) exploit the fact that RNAP-II adds a methyl-Guanosine cap to the 5'-end of the transcript, and perform the biotinylation and RNase digestion to purify full length cDNA¹⁵⁴. The cDNA is fragmented to later purify a short tag around the 5' start site, followed by sequencing. The sequenced tags, which are short (<50bp) and single-ended for most protocols, are then mapped to reference genome and the promoters are identified using clusters of tags. Even though promoter-profiling protocols capture the expression from the RNAs similar to RNA-seq, the correlation between gene expression estimates obtained from RNA-seq and promoter-profiling methods has historically been low¹⁵⁵. Also, unlike RNA-seq there has been absence of biological replicates from most CAGE experiments, and current CAGE analysis tools do not provide the ability to use replicates for differential expression analysis, limiting the use of this assay in promoter profiling.

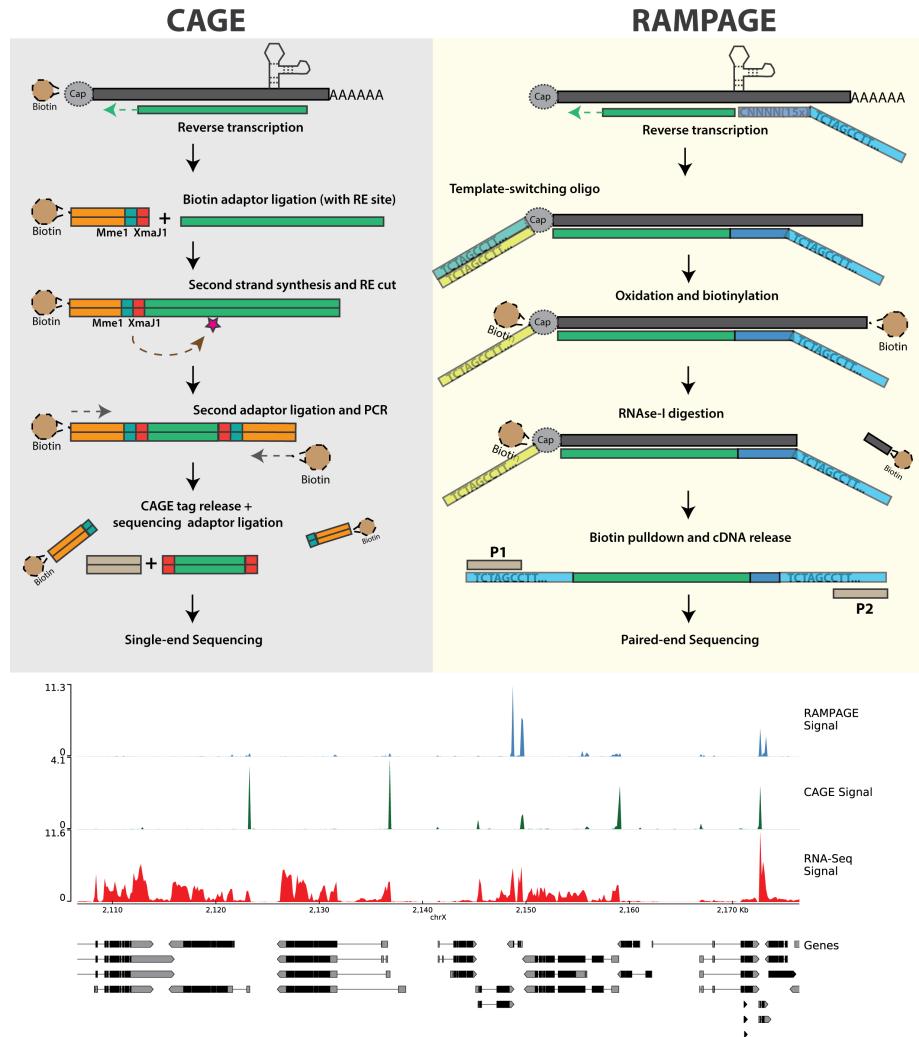


Figure 1.5: *Overview of popular promoter-profiling methods and the resulting data.* **CAGE** begins with biotinylation and pull-down of RNA, followed by reverse transcription and fragmentation of cDNA using restriction enzymes (REs) to produce short (21 bp) tags. **RAMPAGE** on the other hand, uses template-switching oligos containing sample barcodes and PCR primers before biotinylation and pull-down, cutting several steps from the traditional design. Both protocols produce sharp signal at gene promoters (TSS), in contrast to RNA-seq (red) which produces signal all over the gene body. Illustrations inspired by ¹⁵⁶ and ¹⁵⁷.

The earliest methods to identify promoters combined CAGE tags with overlapping 5' start site into a group called **CTSS** (CAGE Tag Start Sites), and the CTSS overlapping with each other were grouped into **tag clusters** ¹⁵⁸. However, this approach (referred as “**distclu**”) used an arbitrary cutoff to avoid overlap between nearby tag clusters, which does not necessarily correspond to cluster shape. Another method (referred to as “**paraclu**”), identifies tag clusters as the segments of genome that maximise the density of CAGE tags per nucleotide within them ¹⁵⁹. Clusters at various scales could be produced by varying the minimum

density (d) parameter, where small values of d produce large, loose clusters, while large values of d produce small, dense clusters. The algorithm finds all clusters for all values of d , and therefore reports subclusters (peaks within peaks) of TSS. Despite the improvements, the method suffers from the issue that it does not correct for the sample-specific biases in the protocols. Issues such as RNA composition effect and capture efficiency of CAGE tags introduce background noise, and affect the detection of robust, reproducible tag clusters. Another method called **reclu** tries to filter detected clusters based on irreproducible discovery rate (IDR) method¹⁶⁰. Nevertheless, methods that allow detection of robust tag clusters by using information from biological replicates would further improve the analysis.

1.4 Aims

The overall goal of my studies was to perform analysis and integration of epigenomics and transcriptomics data, specifically ChIP-seq, HiC, RNA-seq and CAGE, in order to understand the function of the male-specific lethal (MSL) complex in flies and mammals.

In order to achieve the above goal, several bioinformatic challenges discussed before needed to be addressed. The overall goal of the thesis could therefore be divided into three components:

1. Developing methods for high-resolution analysis of HiC data in order to understand the relationships between chromosome conformation and gene regulation in flies that can be applied to understand the mechanism of action of MSLs.
2. Improving existing methods of analysis of promoter profiling in order to perform high-resolution analysis of MSL-mediated dosage compensation on the X-chromosome in flies.
3. Finally, developing workflows for quality control, analysis and integration of data from multiple epigenomic assays, specifically in an allele-specific manner, and utilize them to understand the function of MSLs on the X-chromosome in mammals.

2. Results and Discussion

The following chapter summarizes the methods and insights from the six manuscripts presented in the appendix of this thesis.

Appendix A.1 corresponds to **Ramirez and Bhardwaj et. al (2018)** where we perform high-resolution analysis of online Hi-C datasets in flies and explore the relation of chromosome conformation with regulatory elements and transcription. We also introduce our tools and resources for HiC analysis.

Appendix A.2 corresponds to **Wolf et. al (2018)** which describes implementation of our HiC analysis tools into a Galaxy web-server for end-to-end HiC analysis.

Appendix A.3 corresponds to **Bhardwaj and Semplicio et. al (2018)** where we introduce a new experimental protocol and analysis software for promoter-profiling, and apply them for the analysis of dosage compensation in flies. Experimental data for this manuscript was generated by Giuseppe Semplicio.

Appendix A.4 corresponds to **Aktas and Ilik et. al (2017)** where we describe the function of mammalian ortholog of MLE (DHX9) in controlling genome-wide RNA processing defects mediated by Alu transposons in humans. Experimental data for this manuscript was generated by Tugce Aktas and Ibrahim Ilik.

Appendix A.5 corresponds to **Ramirez and Ryan et. al (2016)** which introduces an upgrade the previously published toolkit from our group “deepTools”, expanding it’s scope to various epigenomic and transcriptomic assays.

Appendix A.6 corresponds to **Bhardwaj and Heyne et. al (2018)** which introduces “snakePipes”, a set of scalable and flexible pipelines that simplify integrative analysis of epigenomic data.

2.1 Insights from high-resolution chromosome conformation analysis in flies

Previous study from our lab investigated the conformation of the male X-chromosome in flies in order to understand the mechanism of targeting of the MSL complex ¹⁶¹. Fidel Ramirez, Thomas Lingg and others generated HiC profiles in *Drosophila* S2 cells using the Hind-III RE, and produced HiC matrices at an average resolution of 4.5 Kb, using the Hind-III cut-sites in the genome as bins. Analysis of this data revealed that MSL2 bound high-affinity sites cluster together on 3D space on the X-chromosome, facilitating the spreading of the

MSL complex. During the study read filtering, matrix generation and normalization using ICE method¹³⁴ was implemented by Fidel in a command line tool called HiCExplorer.

Studies published later^{162,163} performed HiC using the 4-cutter Dpn-II enzyme. This data, combined with the ability of HiCExplorer to create matrices at restriction fragment (RF) resolution, allowed us to study the Fly genome at sub-kilobase resolution (~570 bp) and further understand the relationship between DNA elements, chromosome conformation and transcription¹⁶⁴.

2.1.1 Setup of a HiC analysis workflow

In order to facilitate high-resolution analysis of HiC data, we first improved the TAD calling method described in¹⁶¹ (see details in¹⁶⁴). Specifically, 1) The original matrix is transformed into a z-score matrix before calculating the TAD-separation score 2) For each bin in the genome, the TAD-separation score is calculated for multiple window sizes around the bin and averaged to reduce noise 3) After identification of bins with “local minima” of TAD-separation scores, a Wilcoxon rank-sum test is performed to compare the distribution of that bin with the nearby (left and right) regions, followed by multiple testing correction.

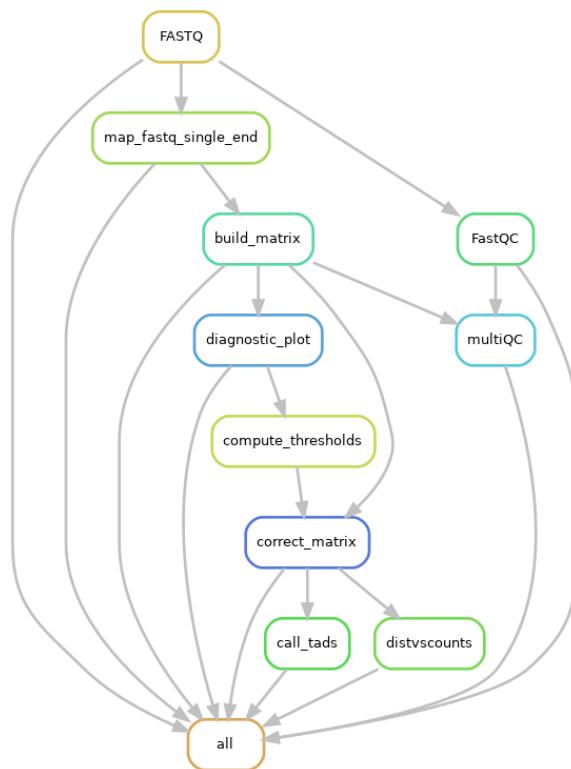


Figure 2.1: *A directed acyclic graph (DAG) of the HiC workflow. It includes FASTQ file quality controls (FastQC) mapping, matrix generation, QC plots, matrix correction and TAD detection. The step marked as “all” collects the final outputs. This workflow has been implemented in snakePipes (see last section).*

I utilized HiCExplorer to create an end-to-end HiC analysis workflow implementing various analysis steps (see Introduction). Computational workflows are like automated protocols that perform a series of analysis steps, ensuring that dependencies between steps are properly resolved (for example, re-running a 5-step workflow with a missing file at step-4 re-runs both step-4 and 5 to ensure that dependent files are updated). Workflows, combined with containerization (each step running in its own virtual environment) ensure the transferability and reproducibility of analysis results¹⁶⁵. This HiC workflow could be utilized for reproducible analysis of both in-house and online HiC datasets (Fig 1). This workflow is now implemented in our epigenomic analysis toolkit called snakePipes, and is routinely used in-house for HiC data analysis.

2.1.2 Relationships between TADs, regulatory elements, and transcription in flies

Using the newly implemented version of HiCExplorer tools, we detected ~2800 boundaries in wild-type female *Drosophila* Kc cells, and found that most (77%) of the TAD boundaries in flies associate with gene promoters. These promoter boundaries associate with active chromatin and have higher DNase hypersensitivity signal (from DHS-seq data) and stronger TAD-separation score than non-promoter boundaries. A carefully performed DNA motif enrichment analysis discovered that promoter-boundaries are enriched with a specific set of core-promoter motifs : motif-1 (M1BP), 2 (Beaf32), 6, 7 (ZIPIC) and 8, while non-promoter boundaries are enriched for some classically known insulator motifs : CTCF, Ibf and Su(Hw). Interestingly, we find that the motifs combinations, rather than the number of motif instances seem to affect the strength of boundaries. Certain motifs combinations, such as Beaf32 + Pita make the boundaries stronger while others, such as Su(Hw) + ibf make them weaker. Our study also challenged the previous studies showing that the number of ChIP-seq peaks of the insulator proteins correlate with boundary strength, and rather emphasised the usefulness of considering motifs at boundaries.

To understand whether there exists a genetic code at boundaries, we used various classification methods, such as logistic regression, random forest and gradient boosting models to predict boundaries using DNA motifs. These methods performed similarly well in identifying boundaries, with a sensitivity and specificity of over 70% on an independent test data. Interestingly the classifiers showed that DNase hypersensitivity is the most important feature in boundary prediction, suggesting that the accessibility of DNA motifs might be a crucial factor for boundary formation.

Using re-analysis of RNA-seq data from modENCODE, we investigated the relationship between our high-resolution TADs and transcription. We find that TADs serve as units of coordinated transcription during *Drosophila* development. Expression of genes with TADs are highly correlated throughout development while genes separated by TAD boundaries show a lack of correlation. We also find that genes at TAD boundaries have expression features of housekeeping genes (high, constitutive expression with low variability) which confirmed previous studies linking housekeeping genes to TAD boundaries.

The catalogue of high-resolution boundaries generated in this study functions as useful resource in understanding the biology of MSLs, as well as other protein complexes in context of 3D chromatin. For example, chromator, a member of MOF-associated NSL (non-specific

lethal) complex was found to be associated with various motifs at boundaries (Fig. S3I of ¹⁶⁴). Further the data produced and analysed during our study allowed us to re-evaluate our previous finding whether the high-affinity sites cluster in 3D space. Using clustering of contact aggregates from the data, we observed that the HAS-HAS clusters emerge after zygotic genome activation (at nuclear cycle 14, together with the establishment of TADs) when MSL2 expression begins ¹⁶⁶. Similar to ¹⁶¹, HAS-HAS clusters were observed in both male and female cells, suggesting that the MSL complex could utilize the pre-existing 3D conformation of the X-chromosome for spreading.

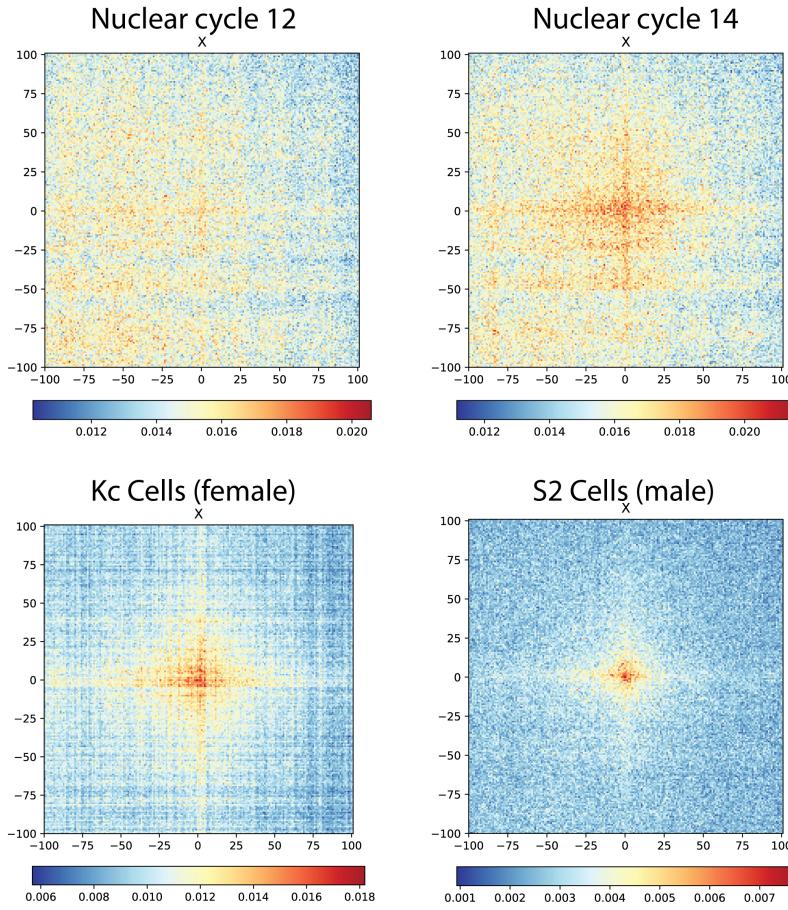


Figure 2.2: **The HAS interaction network on X-chromosome.** An aggregate contact matrix containing mean interactions over all possible contacts within a genomic distance of 0.5 Mb to 15 Mb of HAS loci. Submatrices were extracted for each HAS from the corrected HiC matrix, and the counts were further normalized to total counts of each submatrix. The data from multiple recent *in-situ* HiC studies ^{47,162–164}, analyzed during our study.

2.1.3 Resources for HiC analysis and visualization

During our analysis of chromosome conformation in flies, we improved on the methods developed and presented in the last study from our lab ¹⁶¹. Fidel and I added further functionality to HiCExplorer, such as improved TAD calling, a method to import and

export multiple file formats, and plotting methods. In collaboration with Jose Villaveces (MPI-Biochem), we developed the visualization tool HiCBrowser, which can be used as a standalone tool to browse HiC and other genomic data (ChIP-Seq, RNA-seq etc.). In order to allow biologists to investigate their own gene or region of interest in context of TADs, we developed a web server that contains online HiC data pre-processed using our HiC workflow, along with other genomic data ([<http://chorogenome.ie-freiburg.mpg.de>]). Further in collaboration with Joachim Wolff and Bjoern Gruening, we implemented HiCExplorer along with other tools in galaxy, that facilitates GUI based end-to-end analysis of HiC data for biologists¹⁶⁷. We hope that these resources would facilitate easy and user-friendly HiC analysis for biologists.

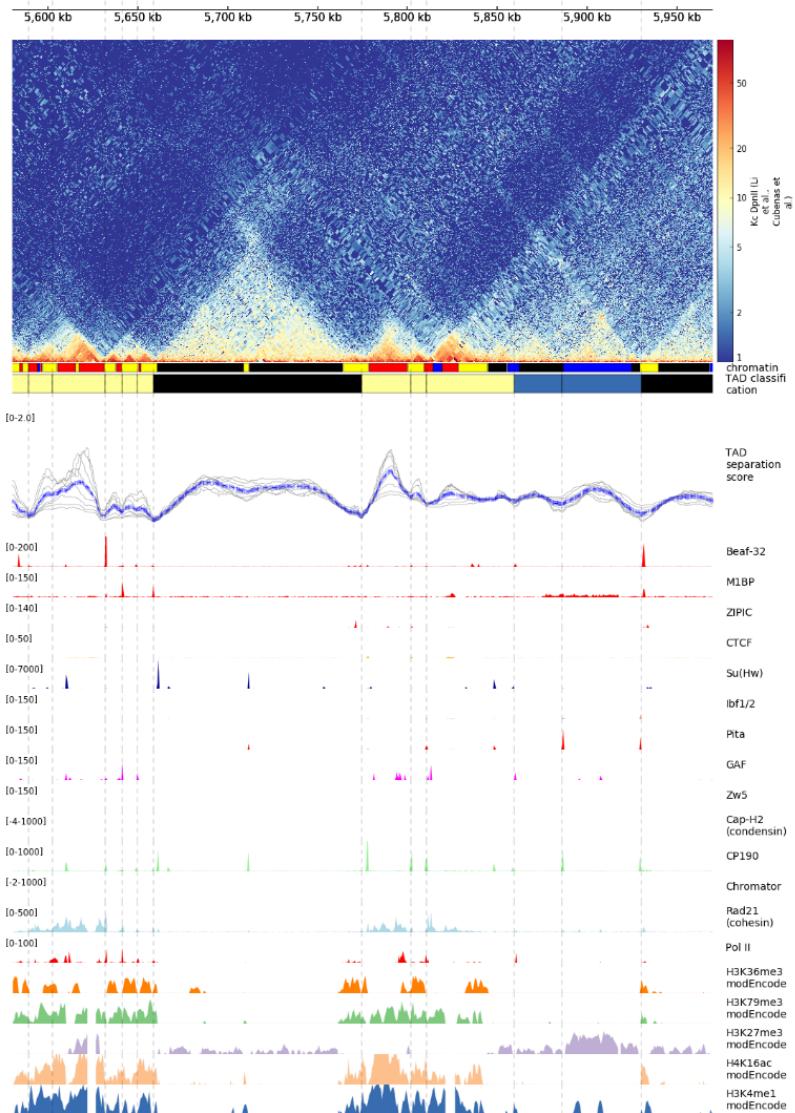


Figure 2.3: ***Chorogenome navigator***. Snapshot of a ~350kb region around MOF in Drosophila* Kc cells, from the chorogenome navigator web server.*

2.2 Using promoter-profiling to study MSL-mediated dosage compensation in flies

2.2.1 Development of the MAPCap protocol

In flies, we wanted to understand how MSL mediated dosage compensation works at the level of individual transcription start sites (TSSs), through promoter-profiling. In collaboration with Giuseppe Semplicio, I contributed to the development of a new experimental protocol for promoter-profiling, termed as MAPCap (Multiplexed Affinity Purification of Capped RNA). MAPCap is performed differently compared to a standard CAGE protocol, in the following major ways :

1. Instead of biotinylation of the 5-mG Cap, MAPCap performs an 5-mG antibody based pull-down of the transcripts.
2. Instead of fragmentation using a restriction site inserted during RT-PCR in CAGE, MAPCap utilizes sonication.
3. RNA is attached to an oligo containing multiplexing barcodes, random barcodes (as UMIs) and sequencing adaptors, allowing us to pool the samples early in the protocol.

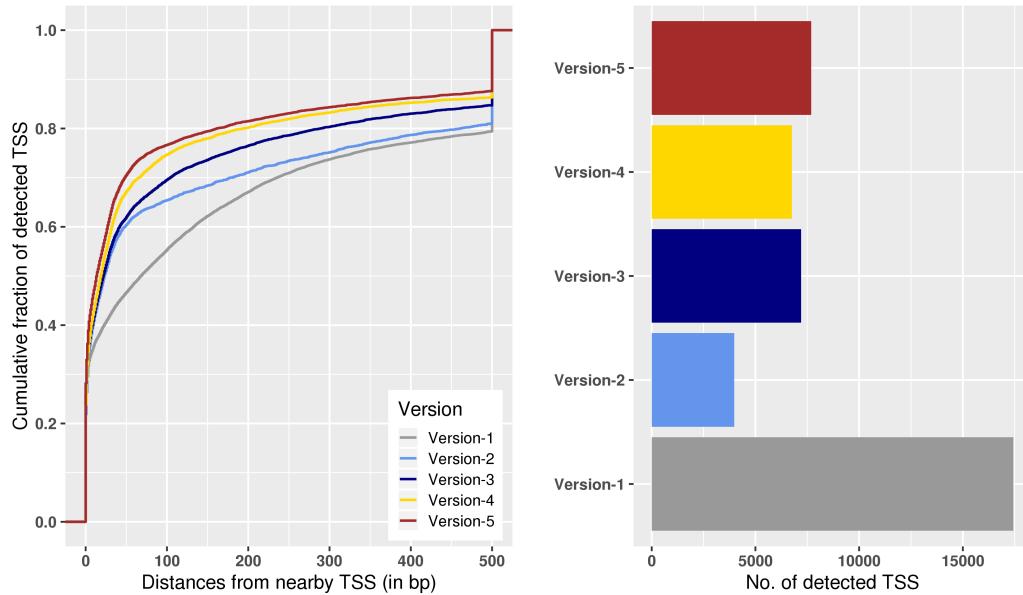


Figure 2.4: *Improvement in TSS detection over different versions of the MAPCap protocol (not completely in chronological order).* **Left :** Distance of detected TSS to an annotated TSS. As the protocol improved, the detected TSS became closer to the annotated TSS in the genome. **Right :** Number of detected TSS for each version. Version 1 has the highest number of detected TSSs, however the precision plot shows that it was mostly due to low signal-to-noise ratio. The number of detected TSSs remains constant for the later versions of the protocol while the precision improved. All samples utilized 5ug of RNA, and TSSs were detected in all samples using the “distclu” method with the same parameters.

Development of MAPCap protocol required multiple rounds of optimizations, where the fragmentation, handling of abundant RNAs, reverse-transcription, and sample multiplexing strategies were changed in response to the insights from the data analysis. **Fig 3** shows how the TSS detection has improved over the various versions of the protocol.

An important insight from the development of the protocol has been the effect of RNA composition bias on the final analysis. In case of protocols like MAPCap, that allow early multiplexing, the effect of composition bias is more pronounced, since there are multiple steps where high abundance RNA can “take over” the overall library composition. Promoter-profiling is usually performed on ribo-depleted RNA, and the RNA composition fluctuates for each run, especially for samples derived from tissues (Fig. 4). This emphasised the need for methods to account for composition bias (such as TMM¹⁶⁸), and the usefulness of replicates in the detection of robust TSSs in presence of noise.

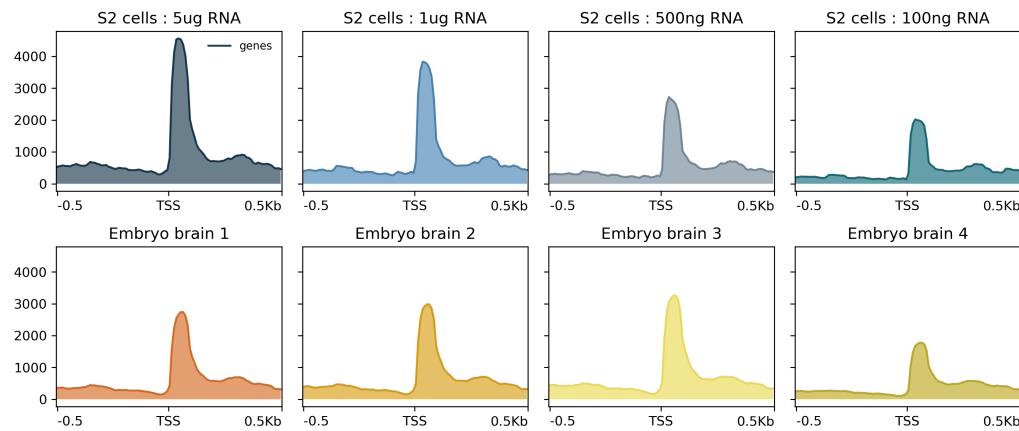


Figure 2.5: ***Effect of RNA composition on TSS enrichment (RPKM normalized signal on TSS).*** **Upper** panel shows the data from a MAPCap run where different input concentrations of RNA were multiplexed. Enrichment decreases as the amount of input RNA gets lower. **Lower** panel shows data from a MAPCap run where total RNA was extracted from different embryos brains. The concentration varies between 200-500 ng per sample. Lower panel represents a more practical MAPCap run, since the RNA composition couldn't be precisely controlled.

2.2.2 Development of the icetea bioconductor package

The new generation promoter-profiling protocols such as MAPCap, and the recently introduced RAMPAGE¹⁵⁷ protocol demand a set of data processing steps different from traditional CAGE. The raw FASTQ files obtained from the protocol contains sample multiplexing barcodes, which are used to separate the samples, and the random barcodes, which could be used to remove PCR duplicates during the analysis. The paired-end data produced by this protocol could be used to improve mapping and annotation of peaks. Also, multiplexing makes it easier for biologists to include biological replicates in their experiments, which could be useful in performing robust TSS detection and downstream analysis (see introduction). Also, a good correlation between expression estimates between MAPCap

and RNA-seq suggested that after appropriate normalization, a MAPCap experiment with biological replicates could also be used for differential TSS expression analysis.

Keeping the above observations in mind, I created a user-friendly R package called **icetea** (**I**ntegrating **C**ap **E**nrichment and **T**ranscript **E**xpression **A**nalysis) that performs processing, TSS detection and differential expression analysis from data such as those from MAPCap and RAMPAGE. **ICetea** implements a new TSS detection approach based on “local enrichment” of CAGE tags, that’s inspired from ChIP-seq analysis. It uses replicates to model robust fold-changes of genomic windows with their local background, and a user-specified fold-change cutoff is used to detect TSS. Further, **icetea** can be used to perform various internal (TMM, RLE etc.) or external (spike-In) normalization of data and detect differentially expressed TSSs between two conditions. **ICetea** is available for use from bioconductor ([\[https://bioconductor.org/packages/icetea\]](https://bioconductor.org/packages/icetea)).

2.2.3 Transcriptional changes during dosage compensation defects in flies

In order to gain insights into transcript-specific dosage compensation in flies, we performed MAPCap, together with RNA-seq, on brains extracted from male and female fly larvae. RNAs were extracted from the mutants of the *maleless* (MLE) gene, which is required for the targeting of the MSL complex to X chromosome through recognition of roX RNAs. Therefore, mutant males are expected to have defects in the upregulation of the X-chromosome. We then deployed a processing pipeline based on icetea for the analysis ([\[https://github.com/vivekbhr/cage/_pipeline\]](https://github.com/vivekbhr/cage/_pipeline)). Analysis showed that most promoters are similarly used between male and female flies. MLE KO leads to a downregulation of majority of TSSs on the X-chromosome in males (>1700 downregulated TSSs), but have almost no effect on females (~23 TSSs differentially expressed). Very few (241) upregulated TSS were observed genome-wide, with only 14 on X-chromosome. Analysis of wild-type ChIP-seq data on the differentially expressed TSS showed that both up-regulated and unchanged TSS had similarly high H4K16ac levels which, while the up-regulated TSS showed low H4K16ac levels in both sexes. We are further generating H4K16ac ChIP-seq in MLE KO to investigate the TSS which remain unchanged upon MLE KO in the MAPCap data in order to understand whether their resistance to change in promoter usage is dependent on their H4K16ac levels.

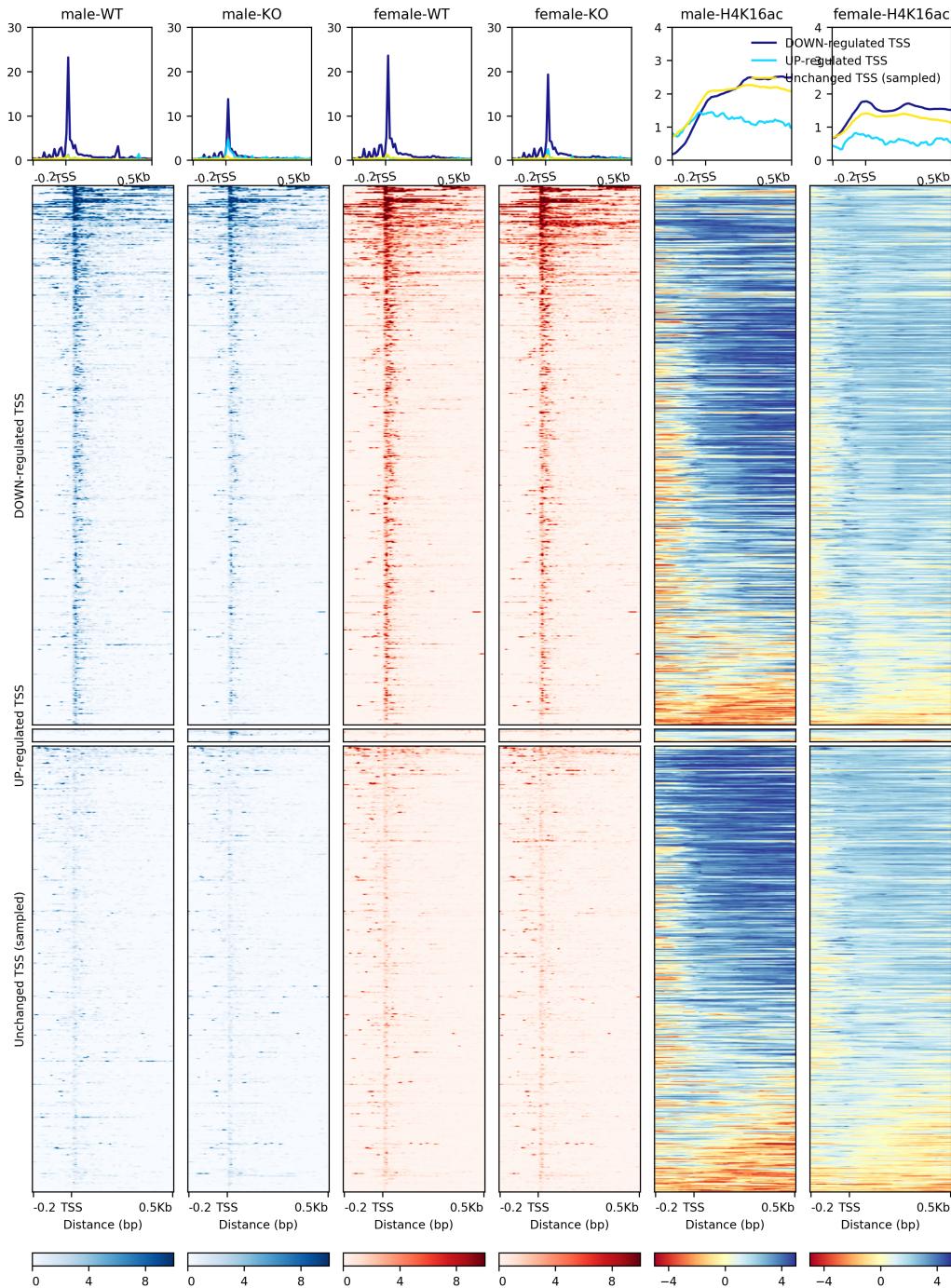


Figure 2.6: **TSS expression (MAPCap) and H4K16ac levels (ChIP-seq) on male and female embryo brains.** Upon KO of MLE, male brains show about ~2-fold downregulation, while female brains do not show a large change. Both downregulated and unchanged TSS have similar H4K16ac levels in wild-type cells (males have higher level than females on the X). The plot is made using a new version of deepTools¹²⁰ that allow creation of coverage files and visualization of data from multiple assays together.

2.3 Integrating transcriptomic and epigenomic analysis to study MSLs in mammals

2.3.1 The role of MSL complex member MLE in mammals

All MSL complex members except the roX RNAs, are conserved from *Drosophila* to mammals (see introduction) and previous study from our lab based on ChIP-seq of the MSLs discovered their genome-wide localization on gene promoters and enhancers⁹⁶. The role of the conserved homolog of MLE, termed as DHX9, was however not investigated in this study. To study the functions of the DHX9 RNA helicase, Tugce Aktas and Ibrahim Ilik in the lab performed an improved version of RNA-CLIP : a method used to determine RNA binding partners of a protein of interest. This method, termed UV-CLAP (Ultraviolet Cross Linking and Immunoprecipitation) produced genome-wide map of DHX9-RNA interactions, and the visual inspection of the data revealed that DHX9 seems to preferentially bind to human Alu elements in the genome.

In collaboration with Tugce and Ibrahim, I performed a genome-wide enrichment analysis of various Alu subfamilies in the DHX9 UV-CLAP data using two independent approaches : 1) consensus repeat mappings, followed by enrichment analysis (adopted from¹⁶⁹) and 2) graph-based clustering of raw DHX9 bound sequences (adopted from¹⁷⁰). Both approaches revealed a significant association of DHX9 with Alu repeats in humans, and their evolutionary relatives : B1 transposons in mice. DHX9 was preferentially enriched on young AluS and AluY families enriched in gene introns. To further investigate the role of DHX9-Alu interaction in introns, Ibrahim performed RNA-seq experiment with poly-A selected and poly-A depleted RNA. My analysis of this data revealed that DHX9 depletion leads to widespread defects in gene expression and splicing, and an increase in production of circular RNA species in the cell. Further, since intronic Alu elements are frequently associated with RNA editing¹⁷¹⁻¹⁷⁴, I developed a workflow for detection of RNA editing from the RNA-seq data ([<https://github.com/vivekbhr/dhx-alu>]). Using the workflow, we identified both an increase and decrease of RNA editing genome-wide, although RNA editing explained only a minor fraction of gene expression changes. Overall, this analysis revealed for the first time, that DHX9 has a major role in regulating RNA processing defects contributed by the Alu elements in the human genome¹⁷⁵.

2.3.2 The role of MSL complex on active and inactive mammalian X chromosomes

Probing the function of the MSL complex in mammals, the previous study from our lab discovered that the differentiation of female mouse embryonic stem cells (ESCs) depleted for MSL complex members (MSL1 and MSL2) leads to chaotic inactivation of the X chromosome. A fraction of these differentiating neural progenitor cells (NPCs) showed both X chromosomes being inactivated⁹⁶. This effect of MSLs was attributed to their regulation of the Tsix promoter through H4K16ac. In order to further investigate the function of MSLs on the X chromosome, we sought to 1) generate genome-wide binding profiles of MSLs in female mouse cell lines 2) Investigate the effect of MSL loss in a system independent of Tsix effects.

Tomsaz Chelmicki therefore generated ChIP-seq profiles of MSLs in female ESCs and NPCs

derived from a mouse strain of hybrid origin, a cross between *Mus musculus* (strain 129Sv) and *Mus castanious* (Cast)¹⁷⁶. To study effects independent of Tsix, Raed Hmadi generated KOs for MSL2 in the NPCs, where X-inactivation has already been established. We then generated RNA-seq profiles of these KO clones, along with the wild-type controls. Further, in collaboration with Laura Arrigoni and Ralf Gilsbach, we generated profiling of various histone marks using RELACS¹⁷⁷, open chromatin profiling using ATAC-seq¹⁷⁸, and whole genome methylation profile using bisulfite-seq (WGBS)¹⁷⁹ on wild-type and KO NPCs.

For analysis of these datasets we needed an approach of mapping and sorting of genomic alignments in an allele-specific manner. I utilized an approach that masks the positions of single nucleotide polymorphisms (SNPs) coming from parental strains in the reference genome before mapping, followed by sorting of unique allele-specific alignments. This method, implemented in the tool SNPsSplit¹⁸⁰, overcomes the “reference bias” : an issue that reads coming from a strain that’s genetically closer to the reference genome, would map better to the reference genome in absence of SNP masking¹⁸¹. Workflows for analysis of epigenomic data that could perform allele-specific sorting were developed (see next section) and applied for the analysis of MSL2 KOs.

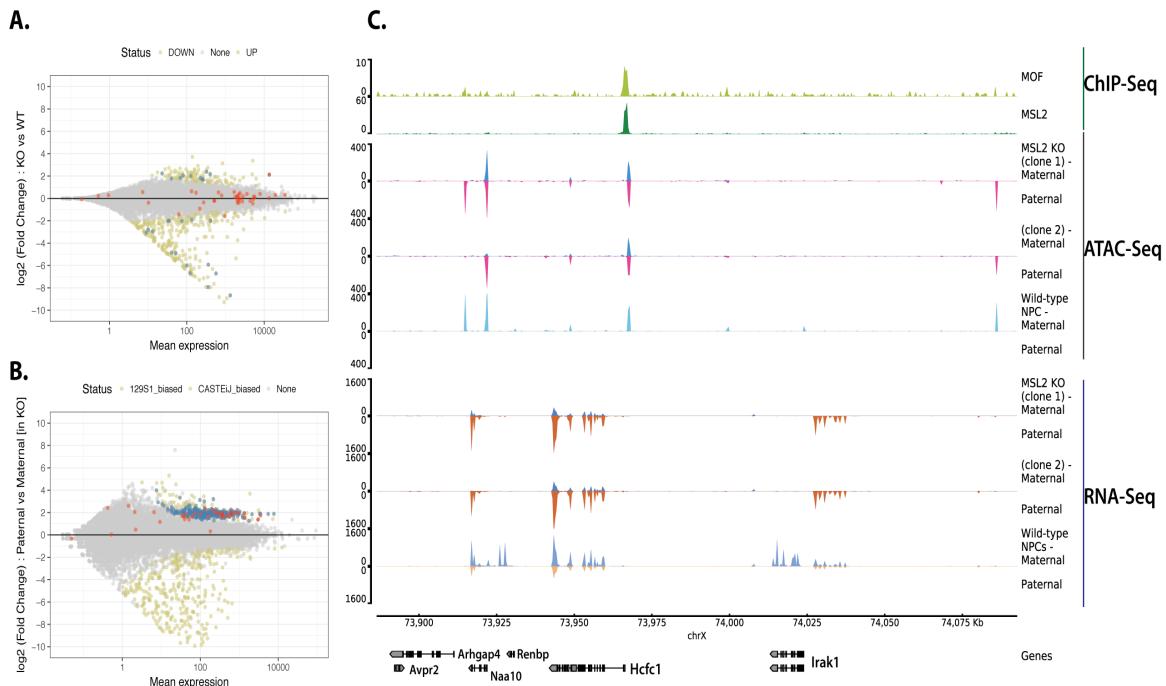


Figure 2.7: Integrative analysis of ChIP-Seq, RNA-Seq and ATAC-Seq in MSL2 knockout (KO) NPCs. **A)** The MSL2 bound genes on X-chromosome (marked in red) seem to be unaffected in normal RNA-seq analysis. Genes differentially expressed between KO and Wild-type cells are marked in green, and affected genes on chrX are marked in blue **B)** Allele-specific analysis of the data reveals that genes on chrX show a clear bias in expression towards the paternal allele (blue). Many of these genes are bound by MSL2 at their promoters (red). **C)** Genome browser track showing a locus with escapees of Xi, MSL2 KO shows activation of paternal allele expression, concordant with open chromatin changes.

Applying our analysis workflow to MSL2 ChIP-seq data revealed that after, but not before differentiation into NPCs, MSL2 binds to significantly more loci on the inactive X-chromosome (Xi), compared to the active X (Xa). Analysis of RNA-seq data revealed an allele-specific downregulation of genes associated with MSL2 binding, upon the loss of MSL2. However the loss on one allele was compensated by a gain in expression from another allele. Analysis of ATAC-seq data in the KOs suggested that this effect is transcriptional, since the promoters associated with loss of expression show a loss of open chromatin and vice-versa (Fig. 5). This analysis suggests that MSL2 might be important for maintaining allelic balance of expression of their target genes. Further, there seems to be an allelic-sensing mechanism on the chromosome that compensates for the loss of MSL2 on the target allele. Further integrative analysis of this data to understand the effects of MSLs on the X chromosome is underway.



Figure 2.8: A directed acyclic graph (DAG) of the allele-specific RNA-seq analysis workflow. The workflow begins with creation of dual-hybrid genomes for user-defined strains of interest, and ends with allele-specific differential expression analysis using a statistical interaction based design. Optional quality-checks can be performed via deepTools¹²⁰ (not displayed here for clarity). This workflow has been implemented in snakePipes.*

2.3.3 A toolkit for integrative epigenomic analysis

The analysis of various epigenomic datasets described in the section above is a massive challenge for a single bioinformatician, and demands a set of uniform processing pipelines. We sought to develop a pipeline that would also allow flexible processing of datasets such that future users are able to change parameters and perform exploratory analysis. This pipeline should be scalable, to run hundreds to thousands of jobs at once, and should be easy to install and use for future. I joined the existing efforts by Steffen Heyne and others in the bioinformatics unit to develop a toolkit that allows analysis of ChIP-seq, RNA-seq, ATAC-seq, WGBS, Hi-C and single cell RNA-seq data. This toolkit, called **snakePipes**, implements various methods and workflows described in the previous sections, in a user-friendly command-line interface. The pipeline also performs allele-specific analysis of data as described in the previous section, upto the point of downstream analysis, such as differential expression of differential peak detection. With snakePipes, we hope that biologists would be able to easily replicate the results from ours and other online studies that present multiple epigenomic assays¹⁸².

2.4 Conclusion and Outlook

Various elements in eukaryotic genome, such as DNA sequences, chromatin and the 3D topological structure act together to facilitate gene regulation. Therefore, Integrative transcriptomic and epigenomic analysis is becoming an important tool to understand genomic regulation. In this project, I performed analysis of various transcriptomic and epigenomic assays in order to understand the function of members of the MSL complex, from flies to humans. To facilitate this, we developed bioinformatic tools and workflows for analysis of transcription, histone marks and 3D conformation of the genome. Application of our methods revealed a catalogue of high-resolution boundaries in flies and their effect on transcription, the role of MSL complex member MLE and H4K16ac in promoter usage in male flies, the newly evolved function of MLE ortholog DHX9 in mammals, and finally an interesting new insight into the role of the mammalian MSL complex on the X-chromosome.

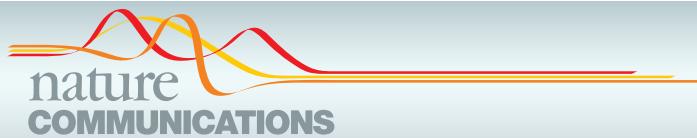
Future directions in studying MSLs would benefit from further integration of data. For example, since most TAD boundaries in flies are associated with promoters, an integration of 3D conformation and promoter-profiling data could reveal the mechanism of how gene-by-gene dosage compensation could be established at early embryogenesis via the spreading of MSL-complex on X. Further, it would be interesting to explore the role of chromosome conformation in other scenarios, such as the recently observed targeting of MSL2 on autosomes¹⁸³. Unlike well studied functions of MSLs in *Drosophila*, we have only recently started to understand the role MSLs in mammals, and in particular, its difference with the role of another MOF associated complex, the NSLs. Rigorously designed ChIP-seq experiments in presence of MSL and NSL knockout controls would be required to avoid antibody non-specificity as well as to study the dependence of MSL targeting on the NSL complex. Our understanding of the role of H4K16ac in mammals would also be improved by studying its relationship with other histone marks. Performing a full histone profiling upon MSL and NSL depletion would be relatively simpler now due to the development of highly multiplexed protocols¹¹⁰, enabling future studies would be able to tackle such question. Further, development of technologies performing low-cell or single cell epigenomics has provided new

scope to study MSL biology in tissues during development and differentiation. Methods for analysis of such assays are yet in their infancy, and therefore would present interesting challenges to work on in the near future.

A. Publications and Manuscripts

A.1 Analysis of chromosome conformation in flies

I contributed to the development of HiCExplorer and HiCBrowser (led by Fidel Ramirez), and developed the Chorogenome Navigator resource. I performed the analysis of motif combinations and boundary strength, prediction of motifs, and relationship of TAD boundary and transcription (Fig 2, S2, 4, S4, 5 and S5). Together with Fidel Ramirez and Thomas Manke I devised, wrote, and revised the manuscript.



ARTICLE

DOI: 10.1038/s41467-017-02525-w OPEN

High-resolution TADs reveal DNA sequences underlying genome organization in flies

Fidel Ramírez¹, Vivek Bhardwaj^{1,2}, Laura Arrigoni¹, Kin Chung Lam¹, Björn A. Grüning^{1,3}, José Villaveces⁴, Bianca Habermann⁴, Asifa Akhtar¹ & Thomas Manke¹

Despite an abundance of new studies about topologically associating domains (TADs), the role of genetic information in TAD formation is still not fully understood. Here we use our software, HiCExplorer (hicexplorer.readthedocs.io) to annotate >2800 high-resolution (570 bp) TAD boundaries in *Drosophila melanogaster*. We identify eight DNA motifs enriched at boundaries, including a motif bound by the M1BP protein, and two new boundary motifs. In contrast to mammals, the CTCF motif is only enriched on a small fraction of boundaries flanking inactive chromatin while most active boundaries contain the motifs bound by the M1BP or Beaf-32 proteins. We demonstrate that boundaries can be accurately predicted using only the motif sequences at open chromatin sites. We propose that DNA sequence guides the genome architecture by allocation of boundary proteins in the genome. Finally, we present an interactive online database to access and explore the spatial organization of fly, mouse and human genomes, available at <http://chorogenome.ie-freiburg.mpg.de>.

¹Max Planck Institute of Immunobiology and Epigenetics, Stübeweg 51, 79108 Freiburg, Germany. ²Faculty of Biology, University of Freiburg, Schänzlestraße 1, 79104 Freiburg, Germany. ³University of Freiburg, Department of Computer Science, Georges-Köhler-Allee 106, 79110 Freiburg, Germany. ⁴Max Planck Institute of Biochemistry and Computational Biology, Am Klopferspitz 18, 82152 Martinsried, Germany. Fidel Ramírez and Vivek Bhardwaj contributed equally to this work. Correspondence and requests for materials should be addressed to T.M. (email: manke@ie-freiburg.mpg.de)

How the DNA packs into the nucleus and coordinates functional activities is a long-standing question in biology. Recent studies have shown that the genome of different organisms is partitioned into chromatin domains, usually called topologically associated domains (TADs), which are invariable between cell types and evolutionary conserved in related species¹.

To understand TAD formation, researchers had focused on the proteins found at TAD boundaries^{2–4}. In mammalian cells, the CCCTC-binding factor (CTCF) protein has been shown to be enriched at chromatin loops, which also demarcate a subset of TAD boundaries (referred to as “loop domains”)⁵. A proposed mechanism, based on the extrusion of DNA by cohesin, suggests that the DNA-binding motif of CTCF and its orientation determine the start and end of the loop^{6,7}. In line with this hypothesis, deletions of the CTCF DNA-motif effectively removed or altered the loop⁶ or caused changes in gene–enhancer interactions that lead to developmental abnormalities in mouse embryos⁸. Additionally, acute depletion of CTCF leads to loss-of-TAD structure on CTCF containing boundaries⁹. However, CTCF-cohesin loops only explain a fraction (<39%) of human TAD boundaries⁵, while plants and bacteria lack CTCF homologs but also show TAD-like structures. Thus, it is possible that additional factors are involved in the formation of TADs.

In contrast to mammals, the genetic manipulation tools available in flies have allowed the characterization of several proteins that, like CTCF, are capable of inhibiting enhancer-promoter interactions. Throughout the manuscript, we will refer to these proteins as “insulator proteins” and their binding motifs as “insulators” or “insulator motifs”. In flies, apart from CTCF, the following DNA-binding insulator proteins have been associated to boundaries^{3,10}: Boundary Element Associated Factor-32 (Beaf-32), Suppressor of Hairy-wing (Su(Hw)), and GAGA factor (GAF). Also, Zest white 5 (Zw5) has been proposed to bind boundaries¹¹. These insulator proteins recruit co-factors critical for their function, such as Centrosomal Protein-190 (CP190) and Mod(mdg4)¹². Recently, novel insulator proteins have been described as binding partners of CP190: the zinc finger protein interacting with CP190 (ZIPIC), Pita¹³ which appear to have human homologs and localizes to TAD boundaries¹¹, and the Insulator binding factors 1 and 2 (Ibf1 and Ibf2)¹⁴. Except for CP190 and Mod(mdg4), all previously characterized boundary associated proteins bind to specific DNA motifs, suggesting that the 3D conformation of chromatin can be encoded by these motifs.

In this study, we sought to identify the DNA encoding behind TAD boundaries in flies. First, we develop software (HiCExplorer) to obtain boundary positions at 0.5 kilobase resolution based on published Hi-C sequencing data from *Drosophila melanogaster* Kc167 cell line^{15,16}. Using these high-resolution TAD boundaries, we identify eight significantly enriched DNA-motifs. Five of these motifs are known to be bound by the insulator proteins: Beaf-32, CTCF, the heterodimer Ibf1 and Ibf2, Su(Hw) and ZIPIC. We find that a large fraction of boundaries contain the motif bound by the motif-1 binding protein (M1BP)¹⁷, a protein associated to constitutively expressed genes. This motif has recently been found at boundaries^{18,19}. The two remaining DNA-motifs have not been associated to boundaries before. Surprisingly, we find that depletion of Beaf-32 has no major effect on chromosome organisation, while the depletion of M1BP leads to cell arrest in M-phase and dramatically affects the Hi-C results. Using machine learning methods based on the acquired DNA-motif information, we could accurately distinguish boundaries from non-boundaries and identify TAD boundaries that were missed when using only Hi-C data. Our results suggest that the genome architecture of flies can be explained predominantly by

the genetic information. We have implemented the methods for Hi-C data processing, TAD calling and visualization into an easy to use tool called HiCExplorer (hicexplorer.readthedocs.io). To facilitate exploration of available Hi-C data, we also provide an interactive online database containing processed high-resolution Hi-C data sets from fly, mouse and human genome, available at <http://chorogenome.ie-freiburg.mpg.de>.

Results

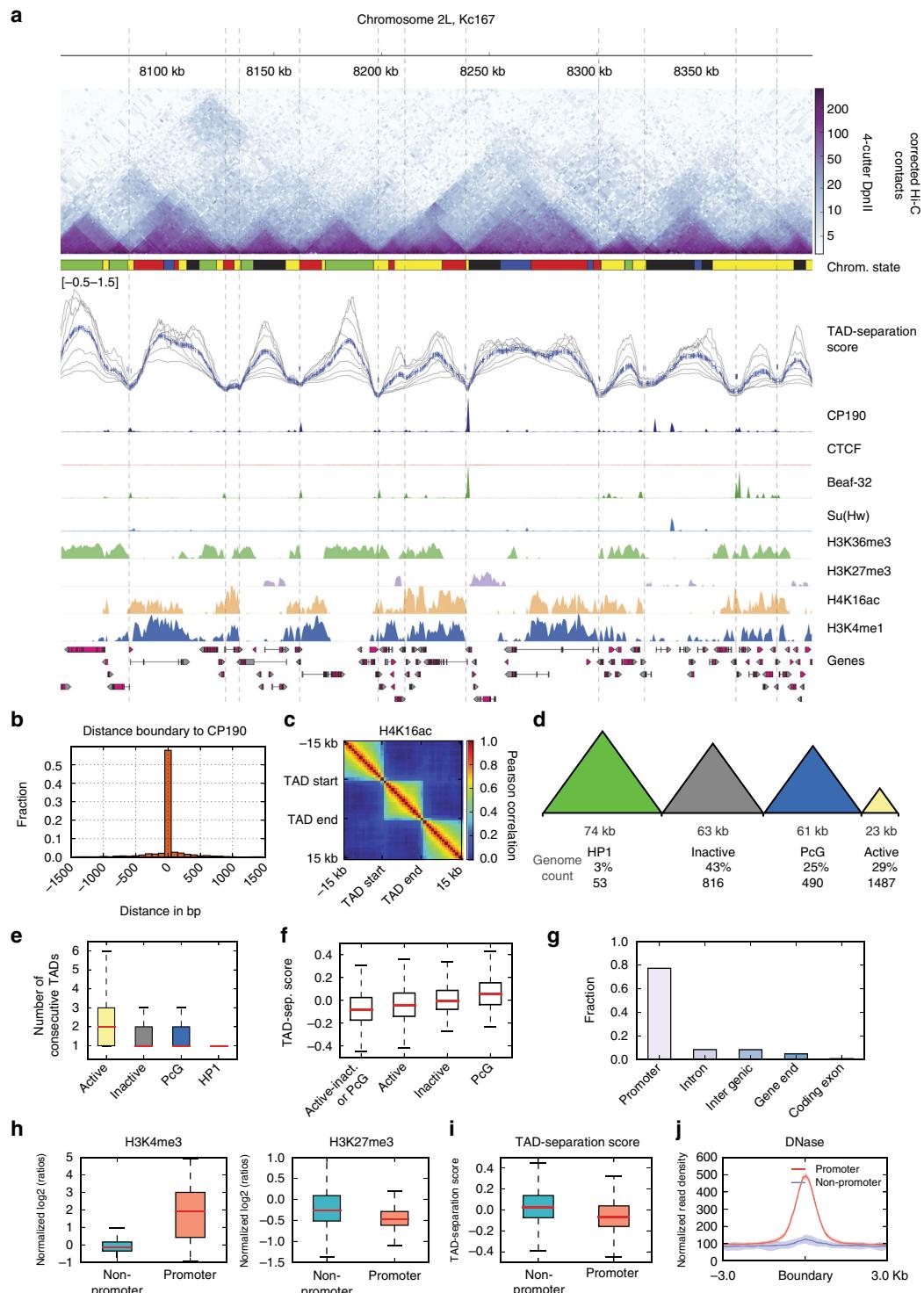
High-resolution TAD boundaries in flies. We obtained Hi-C data for Kc167 cells from two recent studies^{15,16} and processed them to obtain corrected Hi-C contact matrices at restriction fragment resolution (Methods section). These data sets contain the most detailed contact maps in flies, compared to other data sets (Methods section), due to high-sequencing depth (over 246 million valid read pairs) and the use of DpnII, a restriction enzyme with short-restriction fragment size (mean ~570 bp). We found 2852 TADs having a median size of 26 kb (Fig. 1a, Supplementary Figure 1a). We corroborated the precision of our boundaries by comparing their overlap with CP190 peaks (*p*-value = 1.8E – 20, Fisher’s exact test, Fig. 1b) and the separation of histone marks (Fig. 1c, Supplementary Figure 2b). We classified TADs (Supplementary Data 1) using modENCODE histone marks³ as active (enriched for either H3K36me3, H3K4me3, and H4K16ac), polycomb group silenced (PcG) (enriched for H3K27me3), HP1 (enriched for H3K9me3) and inactive (not enriched for any of the marks, see Methods section and Supplementary Figure 1c). A significant fraction of the genome (43%) is covered by large inactive TADs having a mean length of 63 kb (Fig. 1d). In contrast, active chromatin TADs have a mean length of 23 kb and occupy 29% of the genome. PcG chromatin occupies 25% of the genome with TADs that are on average 61 kb. The largest TADs are found for HP1 repressed chromatin, which together occupy 3% of the genome and have a mean length of 74 kb. We also find that active TADs tend to be assembled one after the other due to their higher number (Fig. 1e). Interestingly, the TAD separation score varied significantly (*p*-value < 7.8E – 5, Wilcoxon rank-sum test) between the TAD types (Fig. 1f). The stronger boundaries (low-TAD-separation score) are found between active and inactive or PcG TADs. While the weakest boundaries are found between PcG TADs. Similarly, we find that the TAD-separation score between larger TADs (mostly inactive) is significantly larger than the TAD separation score for smaller TADs (mostly active) (*p*-value = 9.9E – 7, Wilcoxon rank-sum test).

While most of our boundaries overlap with those from previous studies^{3,16}, our method allowed us to identify a larger set of boundaries (Supplementary Figure 1d, e) with increased precision (as measured by distance to CP190, Supplementary Figure 1f-h), which mostly subdivide active TADs from previous studies. We observed that the majority of the boundaries (77%) are located at gene promoters (henceforth referred to as promoter-boundaries, Figure 1g). Promoter-boundaries are different from non-promoter boundaries (23%), since they associate significantly with active chromatin (Fig. 1h, H3K4me3 *p*-value = 4.43E – 144 Wilcoxon rank-sum test, other histone marks can be seen in Supplementary Figure 1i), have lower TAD separation score representing stronger boundaries (*p*-value = 8.5E – 35, Wilcoxon rank-sum test, Fig. 1i), and show higher DNase sensitivity (Fig. 1j).

Specific gene orientation and transcription marks boundaries. Next we correlated our promoter boundaries with gene orientation and transcription. Most promoter-boundaries (70%, *p*-value = 2.5E – 88 Fisher’s exact test), are marked by divergently

oriented gene promoters on either side, while genes in convergent or tandem orientation tend to be inside the TADs. To correlate TADs with gene transcription, we analysed the RNA-Seq data from 14 stages of *Drosophila* development along with the expression in the Kc167 cell line obtained from modENCODE

(Supplementary Figure 2a, Methods section). We found that 95.6% of genes which have a TAD boundary at their promoter are expressed in Kc167 cells (1244 out of 1300) compared to 75.3% of genes which do not have a boundary at their promoter (6892 out of 9149, p -value = 2.19E - 80, Fisher's exact test). Higher



correlation was observed between gene expression inside TADs than between neighboring TADs (Fig. 2a, Supplementary Figure 2c). When we investigated the expression of genes at TAD boundaries we found that these genes show significantly higher expression than the expressed genes which do not have a TAD boundary at their promoters (Fig. 2b, p -value = 6.08E – 21, t -test). Boundary associated genes also show a more stable expression across development than other genes (Fig. 2c, Supplementary Figure 2b), suggesting that these genes are ubiquitously transcribed. Furthermore, we find that for a pair of genes lying next to each other, the variability in their gene expression tend to be correlated during development if these genes are within same TAD, while this correlation is lost if there is a TAD boundary in between (Fig. 2d, Methods section). This is true for gene-pairs in any orientation (convergent, divergent, or tandem, Supplementary Figure 2d).

Taken together, these results suggest that specific gene orientation and level of transcription could be associated with TAD formation.

A comprehensive list of boundary associated DNA motifs. We followed the strategy outlined in (Fig. 3a) to create a comprehensive list of motifs frequently found at boundaries. First, we performed de novo motif calling using MEME-chip²⁰ (Methods section) on our promoter-boundaries and non-promoter boundaries. To filter out motifs that are frequently found at promoters or open chromatin but are not specific to boundaries, we performed an enrichment analysis using two different methods: Ame²¹ and TRAP²². As a second approach, we tested boundaries for the motifs of known insulator proteins and core-promoter motifs from a previous study²³. In contrast to de novo motif detection, searching for known motifs allows additional sensitivity to detect low frequency motifs. After filtering for only consistent results, we could identify 5 motifs enriched at promoter-boundaries and 3 motifs enriched at non-promoter boundaries (Fig. 3b).

The promoter boundary motifs we identified belong to the list of core-promoter motifs 1, 2, 6, 7, and 8 from Ohler et al.²³. Motif-1 is recognized by the recently described ‘motif-1 binding protein’ (M1BP), a protein found at the promoters of transiently paused Pol-II of constitutively expressed genes¹⁷. Motif-2 (also called DRE motif) is recognized by the insulator protein Beaf-32 and DREF²⁴. Motif-7 (also called DMv3) is recognized by the insulator protein ZIPIC. The binding proteins for motif-6 (also known as DMv5) and motif-8 (also known as DMv2) are, to the best of our knowledge, not known. De novo motif calling also identified other core-promoter motifs but they were not found enriched at boundaries. The three motifs that we find enriched on non-promoter boundaries, correspond to the binding sites of Su

(Hw), CTCF and Ibf (Fig. 3b). We could not find an enrichment for motifs of other insulator proteins like GAF, Pita, and Zw5. For clarity, we will refer to the boundary motifs by the name of the insulator protein that binds to them, except for motif-6 and 8.

As an independent validation we find that our three predicted boundary motifs (M1BP motif, motif-6, and motif-8) are also enriched at the binding sites of CP190 and Cap-H2 (condensin II complex) (Supplementary Table 1). We repeated our analysis using the TAD boundaries from previous studies^{3,15,16} and found similar enrichments (Supplementary Table 2).

To better understand the distribution of the motifs on boundaries we performed hierarchical clustering of the binding affinity (TRAP score) for the eight motifs enriched at boundaries (Fig. 3c left panel, Supplementary Figure 3a). We then plotted the ChIP-seq signal of the DNA-binding proteins (Fig. 3c second panel, Supplementary Figure 3b), along with CP190, Cap-H2, Rad21 (Fig. 3c third panel) and RNA Pol-II, over the clusters. The results show that the boundary motifs are usually associated with their corresponding proteins, except for motifs 6 and 8 for which the binding proteins are not known (Fig. 3c second panel). ZIPIC and Ibf show ChIP-seq enrichment at many regions that do not have the motif although their enrichment is higher when the motif is present. This could indicate indirect binding of the proteins as seen for CP190 or could also indicate antibody cross-reactivity or other problems with ChIP-seq experiments^{25,26}. Examples of the boundaries with their motifs and corresponding proteins can be seen in Supplementary Figure 3c–g.

We discover that promoter boundaries are primarily associated with Beaf-32 and M1BP motifs. Promoter boundaries also tend to be associated with condensin II (Cap-H2), RNA polymerase II and housekeeping enhancers. Interestingly, Rad21 (cohesin) ChIP-seq peaks mostly associate with M1BP motif (Fig. 3c–d) and de novo motif calling on Rad21 peaks identified a clear enrichment for M1BP motif (MEME²⁷ E -value = 4.2E – 97) (Supplementary Table 1). Furthermore, M1BP ChIP-seq signal correlates well with Rad21 ChIP-seq (Supplementary Figure 3h). Thus, while Cap-H2 is found together with all boundary promoter motifs, Rad21 is closely associated with M1BP. We found that specific boundaries containing M1BP, as well as ZIPIC and Motif-6 are also associated with paused Pol-II (Fig. 3e) (previously, only M1BP was associated with paused Pol-II¹⁷).

Ibf, CTCF and Su(Hw) are the most common insulator motifs found at non-promoter boundaries, and tend to be associated with enhanced binding of CP190 co-factor (Fig. 3c–d). The ChIP-Seq signals for these proteins follows this observation, showing that promoter and non-promoter boundary proteins are correlated within their group (Supplementary Figure 3h).

We then searched for association of all the transcription factors from modENCODE consortium²⁸, as well as from the

Fig. 1 High-resolution TAD boundaries in flies. **a** Example region of 350 kb showing Hi-C TADs from Kc167 cells. Top panel: Hi-C contact matrix obtained from refs. ^{15,16}. The size of the bins is variable (mean 570 bp) and depends on the genomic location of the DpnII restriction sites. The chromatin state track contains the five classifications from ref. ²⁹: Active chromatin, red and yellow; inactive chromatin, black; Pcg, blue; H3K27ac, green. The TAD separation score track (Methods section) depicts a normalized measure of the contacts between two flanking regions (10–40 kb, depicted by gray line, blue line depict mean score). The boundaries, estimated using the TAD separation score are shown as vertical lines. The following tracks show normalized ChIP-seq coverage for the known boundary proteins CP190, Beaf-32, and Su(Hw) on Kc167³⁹ and CTCF¹⁵. The following tracks contain ChIP-chip data for histone modifications from modEncode²⁸. The image was generated using HiCExplorer. This particular region was selected because many different TADs could be seen; other regions can be browsed at <http://chorogenome.ie-freiburg.mpg.de>. **b** Histogram of the distance of a boundary to the nearest CP190 (common insulator protein co-factor) peak. **c** Correlation of histone marks within and between TADs. Each pixel in the matrix represent the Pearson correlation of the histone mark in all TADs at different distances (Methods section). **d** TAD classification based on histone marks. The numbers below each TAD type represent respectively: mean length, percentage of genome occupied by the TAD and number of TADs of that type. **e** Boxplot of consecutive TAD of each type. **f** TAD-separation score between: active and inactive or Pcg, active-active, inactive-inactive, and Pcg-Pcg. The differences between the groups are all significant (p -value $<= 7.8E - 5$, Wilcoxon rank-sum test). **g** Classification of TAD boundaries. TAD boundaries are classified at promoter if they are within 1000 bp of the annotated TSS. **h** Histone marks at non-promoter and promoter boundaries. Further marks can be seen in Supplementary Figure 1i. **i** TAD-separation score for non-promoter and promoter boundaries. **j** DNase accessibility at non-promoters and promoter boundaries

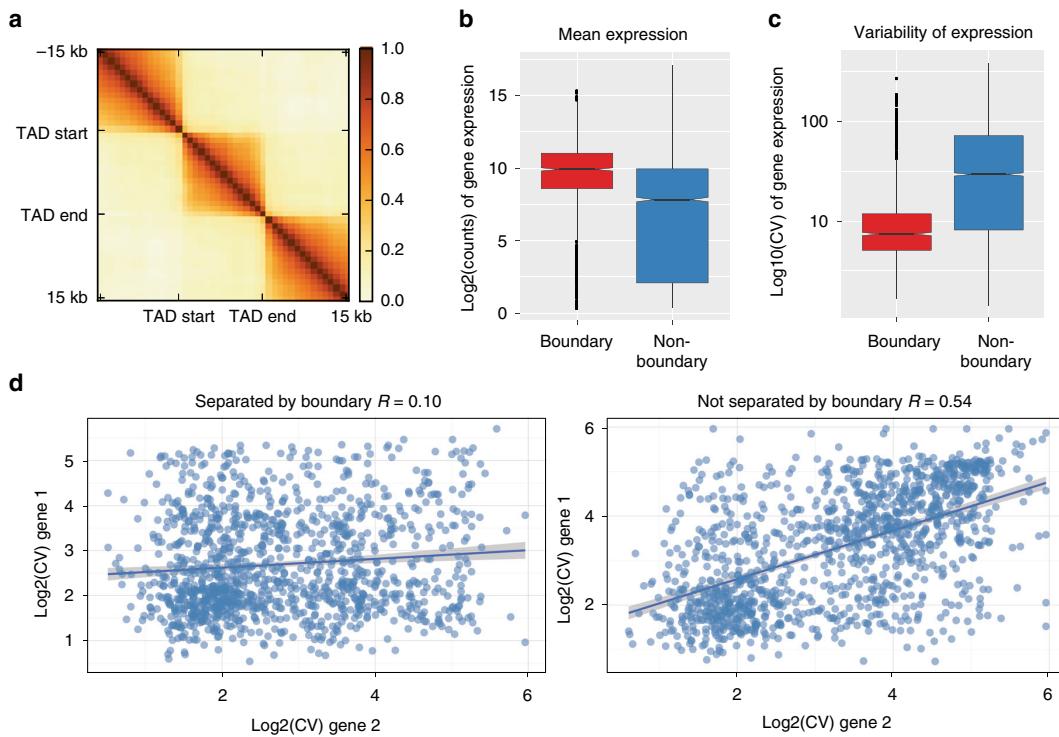


Fig. 2 TAD boundaries are marked by specific gene orientation and transcription. **a** Correlation of mean expression across developmental stages inside TADs vs. outside. Region inside TADs was scaled to 15 kb. Each pixel in the matrix contains the pearson correlation at different distances. **b–c** Mean expression (in Kc167 cells) and variability of expression (during development) for genes whose promoters are at a TAD boundary vs. genes whose promoters are not at boundaries. **d** Coefficient of variation (across developmental stages) between pairs of adjacent genes either separated by a TAD boundary (left) or not separated by a boundary (right). Line shows the linear model fit (shaded region: std. error)

comprehensive collection of ChIP-seq data from refs. ^{15,16} with our boundary motifs. Interestingly, our screen for proteins associated to boundary motifs showed that Nup98, a component of the nuclear pore complex, is associated with motif-6 and Pita, along with CTCF (Fig. 3d and Supplementary Figure 3i). Further validation using de novo motif calling on Nup98 peaks identified motif 6 and Pita motif as the most enriched motifs (MEME²⁷ E -value = 3.4E-4 and 3.3E-9, respectively).

We observed that ChIP-seq peaks of DNA-binding proteins are often found in regions not containing their motif (Fig. 3d). For example, ZIPIC peaks can be seen together with motif 6 or CTCF although ZIPIC motif does not overlap with any of them (Fig. 4a). Similar observations can be made for CTCF ChIP-seq experiments (Supplementary Figure 3j, see discussion), suggesting that motif sequences should be considered along with ChIP-Seq binding sites as more reliable functional predictors of an insulator protein.

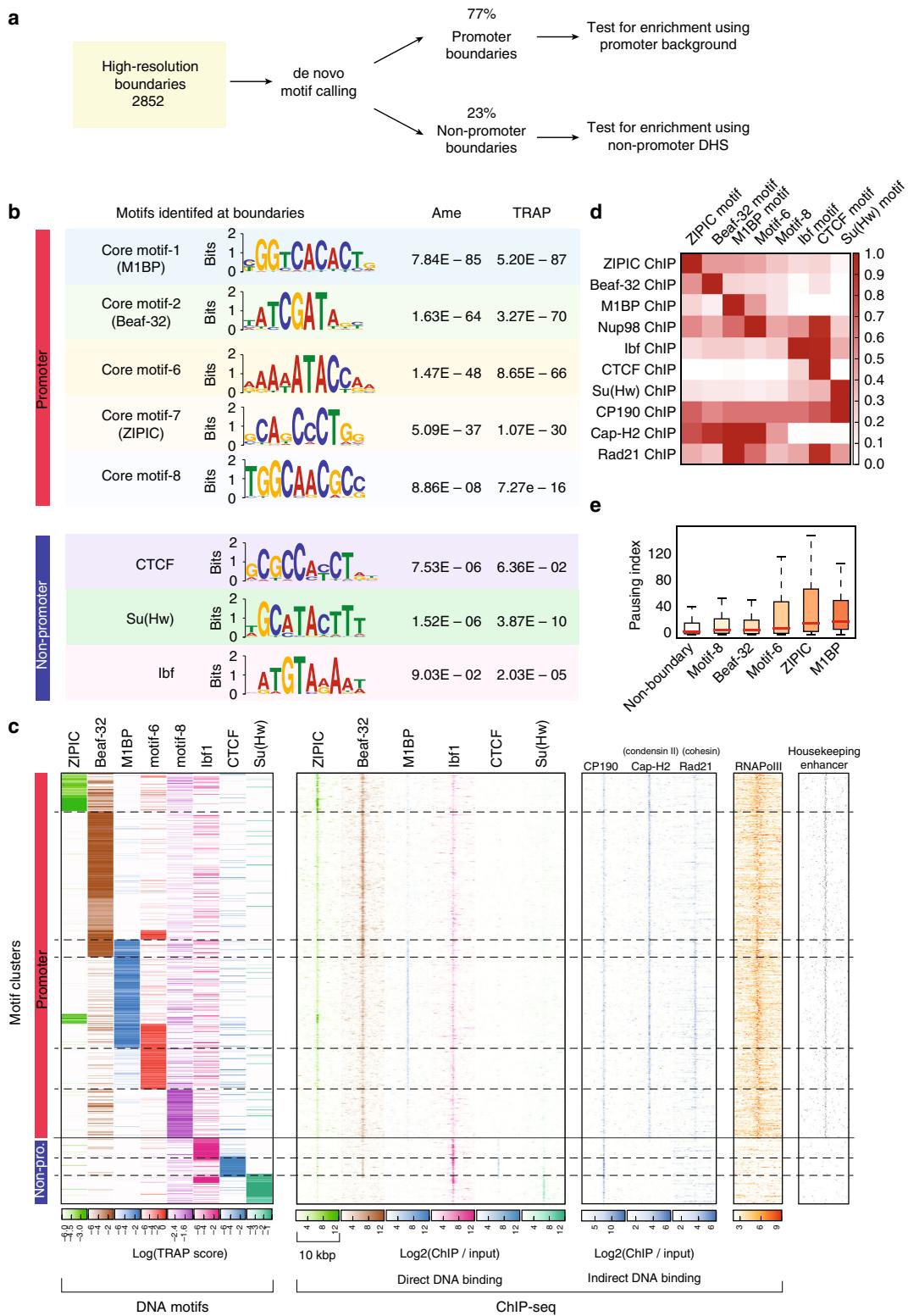
Motif combinations reflect boundary strength. Next we looked at motif combinations at boundaries in relation to boundary strength. Using ChIP-Seq analysis it has been reported that boundary strength increases with the number of proteins bound at the boundary¹⁰. When we looked for motif combinations at boundaries that overlap with their corresponding protein, we did not observe any significant differences in boundary strength between boundaries containing one, two, or three motifs (we only found eight boundaries having more than three motifs, Supplementary Figure 4b), although we could replicate earlier results based on ChIP-seq peaks (Supplementary Figure 4c). We found

that specific motifs or their combinations are associated with boundary strength (Fig. 4a). Boundaries containing the motif for Ibf, Su(Hw) or the combination of the two motifs are weaker than average while the combination of the Beaf-32 motif with either Pita, ZIPIC, or motif-6 result in the strongest boundaries.

We also looked at the association of motifs with active, inactive, P_cG and H_P1 TADs from Fig. 1d. We observe that the promoter-boundary motifs are mostly found between active TADs or between active and inactive (including P_cG) TADs. Conversely, the non-promoter boundary motifs are rarely found between active TADs, and mostly separate active from inactive TADs or are between inactive TADs (Fig. 4b). We find the same trend when analysing the ChIP-chip log₂ ratios of the active histone mark H3K36me3 and the repressive histone mark H3K27me3 surrounding the boundaries (Supplementary Figure 4d). Most promoter-boundary motifs separate active-inactive marks, while most non-promoter boundary motifs lie within inactive marks.

Additionally, we analyzed the chromatin state²⁹ that overlaps with the boundary motifs and found that promoter-boundary motifs lie mostly within active chromatin while non-promoter boundary motifs lie within both active and inactive chromatin (Fig. 4c).

Boundaries can be predicted using motifs. To better characterize boundaries at promoters we used three standard classification methods and ranked features by their relevance to distinguish boundaries from other promoters (Methods section). The features included the TRAP score of all motifs studied along



with other known insulator motifs. We also used DNase hypersensitive sites (DHS) as a feature to identify open promoters, considering that protein binding to a motif requires an accessible promoter. The ranking of feature importance showed that indeed open promoters are required for boundaries (Fig. 5a). For the motifs, the feature importance ranking follows the abundance of the motifs at boundaries: M1BP and Beaf-32 show the highest importance scores, followed by motif-6, ZIPIC, and motif-8. Some features, like GAF and Pita motifs were found to be negatively associated with boundaries at promoters (Supplementary Figure 5a).

Although regularized models used here are less prone to overfitting compared to other machine learning methods, we further protect against overfitting by 10-fold cross validation during training. We then tested model accuracy on an independent test data set. The classifiers showed a sensitivity and specificity over 71% on the test data (Fig. 5b). We also classified open chromatin regions distant from promoters, using only motif scores on these sites as features (Fig. 5c). This resulted in a sensitivity and specificity over 60% (Fig. 5d). The motif binding affinity can also be used as a linear predictor of boundary strength (Supplementary Figure 5b, d). Interestingly, we find the motif bound by GAF, a protein recognized to bind insulators^{30,31}, to be negatively associated with both promoter and non-promoter boundaries (Supplementary Figure 5c, e).

The machine learning predictions can complement the Hi-C derived boundaries at the promoters, as some boundaries predicted by our model are missed by our boundary detection method based on the TAD-separation score (Fig. 5e).

Depletion of Beaf-32 protein does not affect TAD boundaries. Our analysis showed that Beaf-32 and M1BP motifs are enriched at *Drosophila* boundaries. We further explored the effect of depletion of the proteins that bind these motifs on chromosome organisation by performing siRNA mediated depletion of Beaf-32, M1BP or both, followed by *in situ* Hi-C protocol (Methods section). We used the knockdown of GST as a negative control and two biological replicates for each condition. We first observed that while Beaf-32 depletion has no effect on M1BP protein levels, M1BP knockdown leads to downregulation of Beaf-32 (Supplementary Figure 6a). We therefore expected M1BP knockdown to produce either similar, or more exacerbated effects compared to those in Beaf-32 knockdown. Surprisingly, Hi-C analysis showed no global change in chromosome conformation upon knockdown of Beaf-32 compared to GST control (Supplementary Figure 6b). This effect was highly reproducible (Supplementary Figure 6c-f). Depletion of M1BP showed a dramatic effect on the distribution of Hi-C contacts in which short range contacts decrease and inter-chromosomal interactions increase (Supplementary Figure 6d, h). We further found that M1BP knockdown cells were arrested in M-phase (Supplementary Figure 6g), yet the contact distributions are different than those observed for fly mitotic cells

from Hug et. al.¹⁸ (Supplementary Figure 6h). The fraction of inter-chromosomal contacts is often used as a quality measure for Hi-C data³². For healthy cells it had been suggested that this fraction should be <20% of all valid reads. This assumes that chromosomes occupy their own territories which may not be true for perturbed cells. Our data from wild-type cells and Beaf-32 knockdown show an average of 8% inter-chromosomal interactions. The samples from M1BP knockdown and double knockdown show a replicable high fraction of inter-chromosomal contacts (~45%, Supplementary Figure 6d). Since they were identically processed in the same batch and alongside the wild type and Beaf-32 samples, we argue that this is not a technical artifact and the large number of inter-chromosomal contacts probably reflects a biological effect on chromatin.

Resources to explore TADs and associated genomic features.

During our research, we developed processing and analysis tools for chromosome conformation. Our tool-suite, called HiCExplorer, simplifies the Hi-C data pre-processing, quality controls, contact matrix transformation, and TAD calling into a few easy steps (Fig. 6a). HiCExplorer is open source and is available at <https://github.com/deeptools/HiCExplorer/>. Importantly, HiCExplorer can be used with other pipelines and processing tools as we have built-in import/export functions covering commonly used Hi-C data formats. To facilitate analysis, we have integrated HiCExplorer into the Galaxy platform³³. With HiCExplorer, we made available our efforts to create meaningful and accurate visualizations of Hi-C data that can integrate other data sources, examples of which can be seen throughout this manuscript. Further information can be found at the associated documentation (<http://hicexplorer.readthedocs.io>), which includes a full analysis workflow and detailed description of the tools.

Our tools would be beneficial to users that do not routinely perform expensive and technically challenging Hi-C experiments. They enable quick visualization of specific genes or regions of interest in the context of TADs and loops, to understand gene regulation. We provide a resource called the Chorogenome Navigator (Fig. 6b) (<http://chorogenome.ie-freiburg.mpg.de/>), which includes *Drosophila*, Mouse, and Human Hi-C data sets, already processed by HiCExplorer, along with associated gene annotations, histone marks and other TAD/boundary annotations. The underlying program called HiCBrowser (<https://github.com/deeptools/HiCBrowser/>), is also freely available to be used as a standalone browser, where users can include their own genomic tracks. With these resources, we hope to make Hi-C analysis a routine part of genomics workflows.

Discussion

In this study, we used high resolution (DpnII restriction enzyme) and deeply sequenced (~246 million reads) Hi-C data^{15,16} to map the genomic positions of TAD boundaries within ~600 bp in *D.*

Fig. 3 Eight motifs are enriched at boundaries. **a** Overview of the strategy used to identify de novo motifs. **b** Motifs enriched at promoter and non-promoter boundaries (along with Bonferroni-corrected *p*-values). Two methods were used to estimate enrichment (Methods section): Ame²¹ and TRAP²². **c** Clustering of boundaries by motif binding affinity (Methods section). Each row represents one boundary. Left panel: clustering of motif binding affinity using the TRAP score²². Higher scores indicate stronger predicted binding. Dashed lines delineate the clusters. Following panels: using the motif clustering results, we show the heatmaps corresponding to ChIP-seq enrichments for insulator proteins binding the DNA (second panel), other proteins that bind indirectly (third panel) and RNA Pol-II. Last panel shows housekeeping enhancers from ref. ⁶⁹. For boundaries at promoters, heatmaps are centered at the gene promoter, for non-promoter boundaries, heatmaps are centered at the nearest CP190 peak within 2000 bp. ChIP-seq signal was computed in 50 bp bins for 5000 bp from the center. The scale of each heatmap goes from 1 to 12 for the direct DNA-binding ChIPs and from 1 to the max ChIP-seq value for the indirect binding ChIPs (based on Supplementary Figure 3b). **d** Relationship between motif presence and ChIP-seq peak fold change at boundaries. Each cell in the matrix contains the mean fold change of all respective ChIP-seq peaks having the motif. For each row, the maximum fold change was scaled to 1. **e** Pausing index at different boundary-promoters containing one of the boundary motifs. Non-boundary promoters are plotted as control

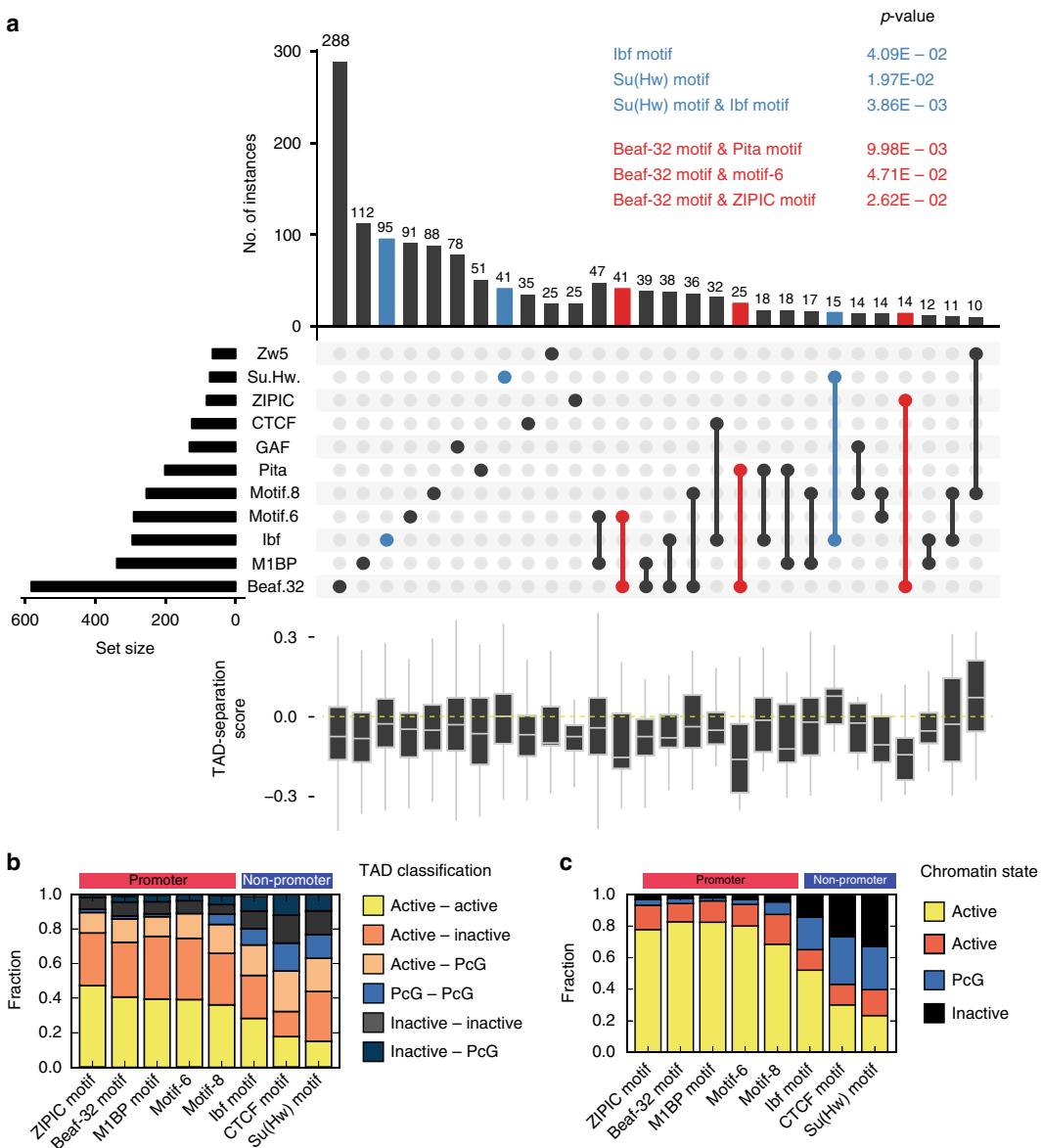


Fig. 4 Promoter and non-promoter boundary motifs show marked differences. **a** TAD-separation score at boundaries grouped by the motif presence. For this analysis we considered a motif to be present if the motif overlaps with a ChIP-seq peak (Methods section). The bars show the overlap between the indicated motifs below. The boxplots show the distribution of the respective TAD-separation score. The sets highlighted in blue have a TAD-separation score distribution significantly larger than the overall TAD-separation score. The p-values (Wilcoxon rank-sum test) are shown above the figure. Similarly, the sets highlighted in red have a distribution significantly smaller. Only motif combinations having >10 instances are shown. Motif combinations with three or more motifs were rare. The intersections were plotted using UpSetR⁷⁰. An overview of the motif overlaps can be seen in Supplementary Figure 4a. **b** Frequency of flanking TAD types (as classified in Fig. 1d) per boundary motif. **c** Frequency of the chromatin state from ref. ²⁹.

melanogaster. Our analysis revealed a larger number of TADs, including many small active TADs (23 kb mean length), that were absent in previous reports^{3,4,34}. We characterized TAD size, boundary strength, chromatin marks, gene orientation, and transcription at the TADs. We perform motif calling at boundaries, validating the presence of known insulators, along with M1BP motif, which recently has also been shown to be associated to boundaries^{18,19} and core promoter motif 6 and motif 8, which have not been associated to boundaries before. Using different machine learning methods, we find that DNA motifs and open

chromatin are sufficient to accurately predict a major fraction of fly boundaries. Finally, we present a set of useful tools and a resource for visualization and annotation of TADs in different organisms.

Our study verifies various properties of fly boundaries indicated in previous publications. We detect that most boundaries associate with promoters and active chromatin (Fig. 1g, h)⁴ and that various known insulator proteins are enriched at boundaries (Fig. 3b,c)^{3,4}. We also detect a comprehensive set of core promoter motifs at boundaries, including the newly discovered

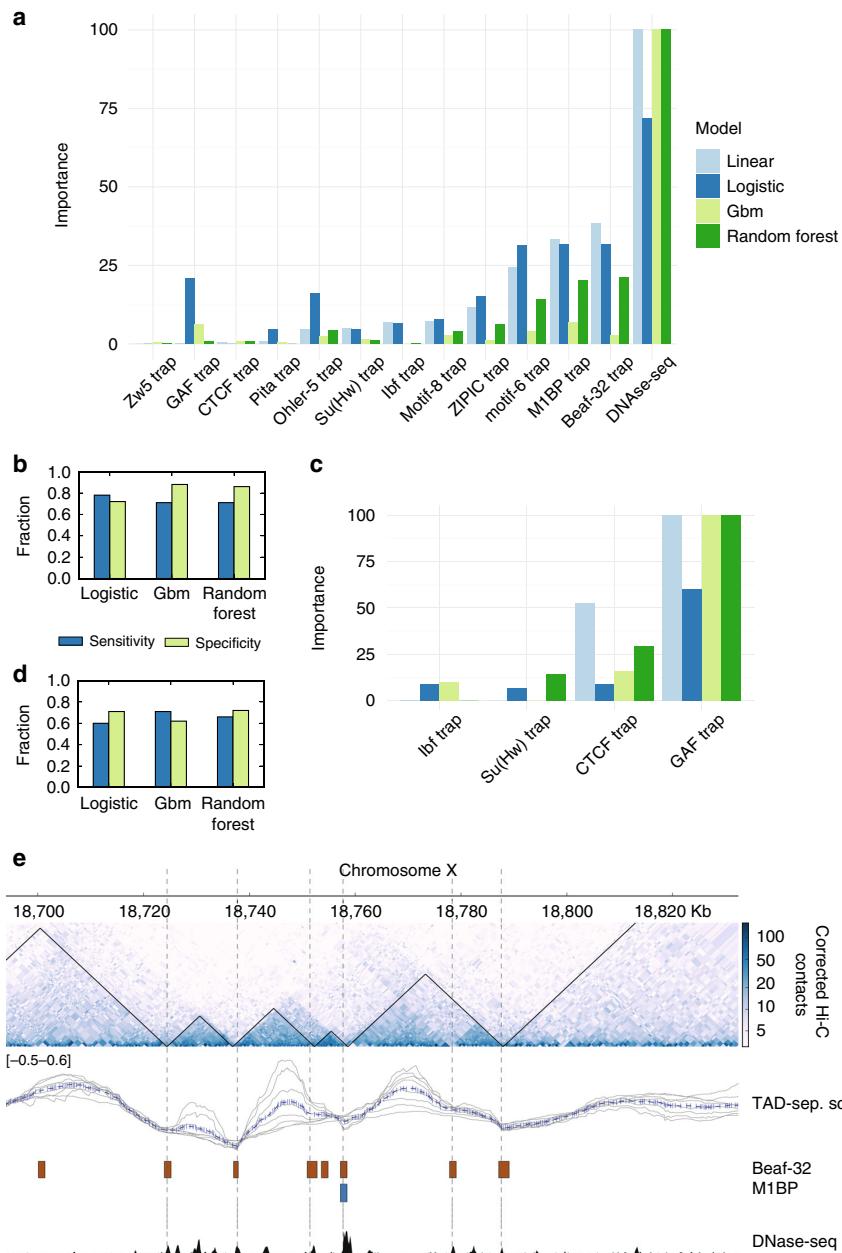


Fig. 5 Prediction of boundaries using machine learning. **a** Feature importance for promoter boundaries computed using four different methods: linear model, logistic regression, gradient boost model (gbm), and random forest. Importance scores of each method were scaled from 0 to 100. Except for DNase-seq, each feature represents the binding affinity (TRAP score) of the respective motif. **b** Sensitivity and specificity for promoter boundaries measured for logistic regression, gradient boost model and random forest. The output of the linear model can be seen in Supplementary Figure 5b. **c** Feature rankings, as in **a**, for non-promoter boundaries. **d** Sensitivity and specificity for non-promoter boundaries. **e** Examples of high-resolution boundaries and predicted boundaries. The high-resolution boundaries (based on the TAD-separation score) are depicted as black triangles on top of the Hi-C heatmap. The predicted boundaries are shown as dotted vertical lines. The tracks below the Hi-C contact map contain the instances of the motifs that overlap with promoters. To aid the visualization of the short motifs, their genomic location was extended by 500 bp in each direction. The last track depicts regions of open chromatin based on DNase-seq from modEncode²⁸.

M1BP motif^{17–19}, and motifs which have been associated to housekeeping gene expression^{35,36}. However, some of our results contradict previous observations. For example, we find that genes at boundaries have higher expression and lower variability of expression throughout fly development (Fig. 2b, c, Supplementary

Figure 2b). This in line with Ulianov et al.^{18,34} and Hug et al.^{18,34} but in contrast with Hou et al.⁴, who suggest that gene density and not the transcriptional state is important for boundary formation. Unlike Hou et al.⁴ we find that genes at boundaries tend to be divergently transcribed. In contrast to various earlier

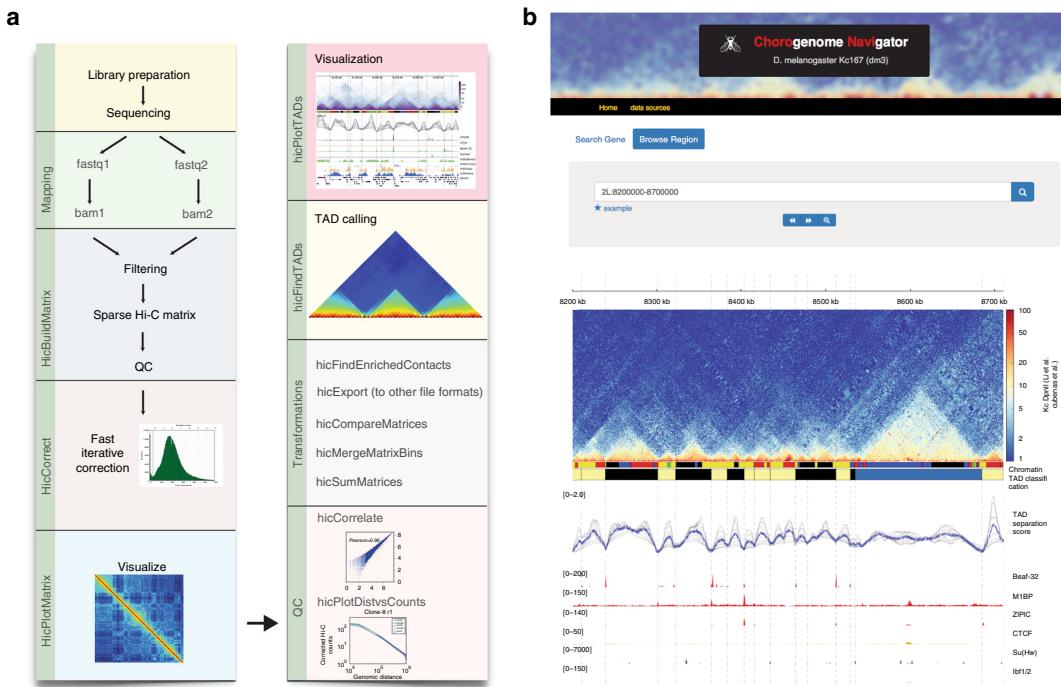


Fig. 6 HiCExplorer and Chorogenome navigator. **a** Pipeline of Hi-C processing and the commands used by HiCExplorer. HiCExplorer is easy to install (`conda install hicexplorer`), easy to use, tested (automatic tests and numerous data sets processed), well documented (<https://hicexplorer.readthedocs.io/>) and is ready to be used within the Galaxy framework³³. **b** The Chorogenome navigator aims to be a repository for available Hi-C data. Currently, we host data from fly, mouse and human. Customized tracks can be created using the HiCBrowser (<https://github.com/deeptools/HiCBrowser>)

studies^{3,4,37}, CTCF does not appear to be a major boundary associated insulator in flies (Figs. 3c, 4a). We also show that the number of insulator motifs at boundaries correlates very little with boundary strength (Supplementary Figure 4b), in contrast to Van Bortle et al.¹⁰.

Most of these differences are due to the increased resolution of detected boundaries (Fig. 1b, Supplementary Figure 1f–h) and the combined analysis of DNA motifs with ChIP-seq data, rather than ChIP-Seq peaks alone. We show that correlating boundaries with ChIP-Seq peaks alone is not a good measure when it comes to determinants of boundary formation. Many DNA-binding proteins show co-localization in ChIP-Seq data without presence of the corresponding DNA motifs (Fig. 3d, Supplementary Figure 3j). This is possible due to cross-linking artifacts and indirect binding, which is, in fact, aggravated at boundaries, which tend to contact each other in 3D space³⁸.

Another argument for considering motifs is the contradicting case of CTCF at boundaries. In contrast to earlier studies based on CTCF ChIP-seq^{3,4}, we find that the CTCF motif is rarely associated to boundaries. This difference is caused by the quality of the ChIP-seq data that can produce spurious peaks. For example, we could observe a significant enrichment of CTCF at boundaries in the ChIP-Seq data from Wood et al.³⁹ but not in the CTCF ChIP-seq data from Li et al.¹⁵. On the other hand, both ChIP-seq data sets show significant enrichment if we only consider ChIP-seq peaks that contain the CTCF motif, (Supplementary Figure 3j). For CTCF, and in general for ChIP-seq experiments in flies, ‘phantom peaks’ are known to occur at active promoters²⁵. Thus, to avoid misleading results our analyses are based on motif presence when possible and for ChIP-Seq data sets, we use significance threshold along with motif binding affinity for analysis (instead of taking a significance cutoff alone).

We observe that boundary strength is associated with the chromatin states of flanking TADs and particular motif combinations, but is not associated with the number of co-occurring boundary motifs. Boundary strength is higher between active and inactive/PcG TADs while is lower at boundaries separating two TADs within the same state (e.g., active–active, inactive–inactive, Fig. 1f). Boundaries containing Beaf-32 are stronger when present together with either motif 6, Pita, or ZIPC motif while weaker with motif 8 (Fig. 4a). Although, the mechanism by which combinations of insulators alter the boundary strength still remains unclear, we observe an association of Nup98 with Pita motif, motif 6, and CTCF (Fig. 3d and Supplementary Figure 3i), suggesting that association with nuclear pore proteins may result in stronger boundaries. Nup98 has now been shown to be functionally important in mediating enhancer-promoter looping in the *Drosophila* genome⁴⁰.

Our results indicate that the two sets of boundary motifs (promoter and non-promoter) participate in the compartmentalization of different types of chromatin. Boundaries containing core promoter motifs are either flanking, or surrounded by active chromatin regions (Fig. 4b). In contrast, the boundaries containing non-promoter motifs tend to be within or at the borders of inactive or repressed chromatin (Fig. 4b). This finding is in line with previous reports showing an enrichment of CTCF at the borders of H3K27me3 domains^{3,37} and an enrichment of Beaf-32 in active chromatin³. This indicates that insulator proteins might serve different functions guided by the DNA sequence. For example, we observe that GAF motif, whose presence is negatively associated with TAD boundaries (Supplementary Figure 5a, c), is rather detected alone at “loop domains”⁵ (Supplementary Figure 5e).

Our analyses indicate that the depletion of the well-studied insulator protein Beaf-32, has no significant effect on the chromosome conformation. However, in *Drosophila melanogaster*, both the Beaf-32 and DREF proteins bind exactly the same DNA motif²⁴. Thus, our results, as well as others^{41–43} point out that DREF, a protein that unlike Beaf-32 is conserved in humans, might have a more prominent role in genome organization than previously thought.

On the other hand, cells under M1BP knockdown grow slower in culture and get arrested in M-Phase, probably because M1BP is a transcription factor of constitutively expressed genes¹⁷. Since M1BP depleted cells show cell cycle defects, it is difficult to separate the direct role of M1BP at boundaries from the indirect effects caused by deregulation of thousands of genes. To study the direct role of M1BP at boundaries, it would be useful to perform either deletion of M1BP motif on boundaries using CRISPR, as shown for CTCF in mammals^{44,45} or through acute and complete depletion of M1BP⁹.

In this study, we present evidence that the DNA sequence contains features that can guide the formation of higher order chromosome organisation. The association of boundary types with a combination of motifs (Fig. 4a), and the fact that we can predict boundaries using DNA sequence alone, in absence of any information about associated protein or histone marks (Fig. 5 and Supplementary Figure 5a–d), leads us to propose a DNA-guided chromatin assembly model. In this model, the boundary elements are recognized by their proteins, which help loading TAD assembly factors onto chromatin. Promoter and non-promoter boundaries can thus have different mechanisms of formation. DNA motifs at inactive regions can attract proteins that may establish TAD domains by setting up barriers for chromatin marks⁴⁶. Although overall barrier activity of insulator proteins have been controversial^{10,43}, it is plausible that the barrier mechanism is active only at a subset of boundaries (like those of inactive TAD domains). DNA motifs at gene promoters can associate with core-promoter proteins which then guide the assembly of Pol-II pre-initiation complex. The pre-initiation complex can then recruit condensins⁴⁷. Once recruited, condensins can perform loop extrusion independent of Pol-II transcriptional activity^{18,48,49}, leading to emergence of TADs^{6,50}. Condensins can also remain associated to chromatin during mitosis, to re-establish TADs after the cell division. In general, our results indicate that active transcription and chromosome conformation are related, in-line with a recent study⁵¹. Future studies investigating the association of Pol-II pre-initiation complex and condensin activity on gene promoters would advance our understanding of mechanism of TAD formation.

Methods

Hi-C processing. Different Hi-C data available for *D. melanogaster* were downloaded from GEO and processed using HiCExplorer (<http://hicexplorer.readthedocs.io/>). The list of data sources can be seen in Supplementary Table 3.

Each mate of the Hi-C sequencing read pairs was aligned separately using bwa mem with parameters ‘-E50 -L0’. The E parameter is the gap extension penalty, which is set high to avoid gapped alignments. This is because a fraction of the reads from a Hi-C experiment contain sequences from two distinct genomic positions. By increasing the gap extension penalty we promote the aligner to map the two parts of the read separately instead of trying to map the read to a single location. The L parameter is the penalty for 5' and 3' clipping which we set to zero to favor such clipping for the same reason as before.

To create the contact matrices, HiCExplorer divides the genome into bins of unequal size demarcated by the genomic positions of the restriction site and a matrix was created having these bins as rows and columns. The mapped reads were processed to count the number of times any two bins were connected by the Hi-C reads pairs. The following reads were discarded: read pairs that were not uniquely mapped or had a mapping score lower than 15, were within 800 bp to each other, were duplicated, contained a dangling end indicative of defective re-ligation or when one of the fragment mates was farther than 1500 bp from the restriction site.

In our processing of the data, we observed that restriction enzymes do not cut with the same efficiency at all sites or sometimes do not cut at all. Because of this, after the creation of the contact matrices, rows, and columns with zero or small total counts were removed. To define a sample-specific lower cutoff, we analyzed the bimodal distribution of total counts per rows, which is a convolution of two distinct distributions. The distribution with lower counts contains all bins with zero reads, mostly from repetitive regions, and also bins with low number of reads, probably from inefficient digestion of restriction sites. We chose as count threshold the position of the local minimum between the two modes. After filtering low count bins, the matrices were corrected following the iterative procedure from ref. ⁵².

For the 4-cutter DpnII restriction enzyme the mean fragment length after removing low coverage bins is 570 bp. For the 6-cutter HindIII the average was 4500 bp.

Identification of boundaries. TAD boundaries were identified using an improved version of TAD-separation score method from ref. ⁵³ which is similar to TopDom⁵⁴. The method works by first transforming the Hi-C contact matrix into a z-score matrix $A = (a_{ij})$. For this, each contact frequency in the matrix is transformed into a z-score based on the distribution of all contacts at the same genomic distance. For a bin l , the contacts between an upstream and downstream region of length w are in the z-score submatrix of $A[\alpha_l, \beta_l]$, such that $\alpha_l \in \{l - w, \dots, l\}$ and $\beta_l \in \{l, \dots, l + w\}$. This submatrix corresponds to the ‘diamond’ seen in Supplementary Fig. 1a. For each matrix bin we compute the TAD-separation score(w) as the mean overall $A[\alpha_l, \beta_l]$ values.

To reduce noise a multi-scale version of the TAD-separation score is computed for different values of w that are averaged per bin, $w \in \{10,000, 12,000, 18,000, 25,000, \text{and } 40,000\}$. Genomic bins with a low-TAD-separation score with respect to neighboring regions (local minima) are indicative of TAD boundaries (Supplementary Figure 1a) and stronger boundaries will have lower scores. To assign a statistical significance to each local minimum we compare the distribution of the z-scores for the submatrices $A[\alpha_l, \beta_l]$ having $l \in \{i, i - v, i + v\}$, where i is the bin of the local minima, and $i - v$ and $i + v$ are the bins at distance $v = 1000$ bp upstream and downstream of i , respectively. We use the Wilcoxon rank-sum test to compare the values of $A[\alpha_l, \beta_l]$ with the values of each of the other two matrices and the highest of the two p -values is used. Finally, we correct the p -values using the Bonferroni method, and report boundaries with $p < 0.001$.

We also used a minimum local minima depth of 0.01. Depth of local minima (referred to as delta) can be considered similar to “fold change” of any minimum, with respect to the neighboring TAD-score average. These parameters were selected by maximizing the AUC for ROC curves in which the positive set contained CP190 binding sites and the negative set contained inactive regions (black chromatin state from ref. ²⁹).

In contrast to other published methods to call TADs, this procedure has several advantages: (i) Each boundary is associated to a TAD-separation score and a p -value, that are useful to characterize strong vs. weak boundaries, (ii) the TAD score can be easily visualized (e.g., as an genome browser track), which is always useful for visual inspection, and (iii) the computation of boundaries takes only minutes, scaling linearly with the length of the genome). Our method differs from the TopDom method in the following aspects: (i) we compute TAD-separation scores using a z-score matrix while TopDom uses the corrected counts matrix, (ii) we use multiple length (w) sizes to compute our TAD-separation score while TopDom uses a single w -value, (iii) we compute p -values using the ‘diamond’ $A[\alpha_l, \beta_l]$ submatrices values in contrast to the ‘diamond and triangle’ distributions used in TopDom. The triangle distribution contain the intra z-score values between bin $l - w$ and l , and the intra z-score values between bin l and $l + w$.

Validation of boundary quality. We used the following functional signatures to validate the quality of our boundaries:

Distance to known insulator co-factor CP190: since all studied insulators proteins bind to CP190^{11,12,14}, a sensible quality measure is the overlap of boundaries with CP190 ChIP-seq peaks. For this, we computed the distance of the boundaries to CP190 peaks using bedtools⁵⁵ closestBed (Fig. 1b, Supplementary Figure 1b). For comparison, we randomized our boundary positions using bedtools shuffleBed (Supplementary Figure 1h) and estimated the new distances to CP190. ShuffleBed simply assigns a new random position for each boundary anywhere in the genome (excluding heterochromatic and unplaced regions). Finally, we computed the background probability of obtaining the observed overlap between CP190 peaks and Hi-C boundaries using bedtools Fisher's exact test.

Separation of histone marks: as boundaries are expected to separate histone marks we used the method described by ref. ⁵ to quantify the correlation of marks within TADs and between TADs. For this, each TAD was scaled to 15 kb, flanked with a 15 kb region and divided into 1 kb bins. For each bin the mean histone ChIP-chip value was recorded, thus generating a matrix of 2852 TADs (rows) and 45 bins (columns). For this we used computeMatrix from deepTools2. The pairwise pearson correlation value of each column was then computed to produce a matrix of size 45×45 .

Classification of TADs. The following histone marks for Kc167 from modEncode²⁸ were used: H3K36me3, H3K4me3, H3K9me3, and H3K27me3. Other marks that correlate closely to these marks were not included. For example,

H3K9me2 correlates closely with H3K9me3, H3K4me2 with H3K4me3, and so on. The average intensity of the marks over the TAD length was computed using multiBigwigSummary from deepTools⁵⁶. The resulting matrix was clustered by computing Euclidean distances between the histone marks and applying hierarchical clustering using the complete method. Five clusters were detected (Supplementary Figure 1c) that correspond to the presence of H3K36me3, H3K9me3, H3K4me3, H3K27me3, or none. Analysis of the TADs containing H3K36me3 and H3K4me1 in the genome revealed that that H3K36me3 is present at exons of active genes while H3K4me1 is mostly present at introns and intergenic regions of active genes and less abundant at exons. Thus, noticing that these two marks are complementary for active regions we classified TADs having predominantly these marks as ‘active’. For the other clusters we used the same categories as²⁹: the cluster of TADs with H3K9me3 was labeled as ‘HPI’ (Heterochromatin Protein 1); the cluster with H3K27me3 was labeled ‘repressed’ or ‘PcG’ (Polycomb group) and the cluster with no mark was labeled as inactive.

Analysis of transcription at boundaries. In order to analyse transcription at boundaries, we downloaded ribo-depleted RNA-Seq data from modENCODE⁵⁷. All data were mapped to the *Drosophila* (dm3) genome using HISAT2 (v2.0.4)⁵⁸ and the reads were summarized per gene using featureCounts (v1.5.0.p1)⁵⁹ using options ‘-p --primary -Q 10’. We used data from Kc167 cells, along with 14 different developmental stages, ranging from embryo to adult. We only used data produced in 2014 in order to avoid batch effects and further confirmed the data quality by clustering the samples by Euclidean distance (Supplementary Figure 2a). We normalized each sample by library size, averaged the counts for replicates, and finally used log transformed counts for all the analysis.

In order to be able to compare all samples from mixed sex, genes on sex chromosomes were excluded for analysis. In total 10,449 autosomal genes were used for expression and variability analysis. Genes were considered expressed if they have a normalized log-count >0. Similar results were obtained using the more stringent cutoff of log counts ≥10, genes with boundary promoters = 45.84% and non-boundary promoter 18.36% (*p*-value = 1.58E – 96, Fisher’s exact test). Variability was assessed using coefficient of variation of a gene across all developmental stages along with Kc167 cells. To measure the effect of boundary on nearby gene expression, we first define pairs of adjacent genes in different orientation (convergent, divergent, or tandem) and record the coefficient of variation for each gene in the pair. Then we obtain the scatterplots for all such gene-pairs and compare the results for two different scenarios: (a) where the pairs are separated by a boundary and (b) where there is no such separation. We further tested whether in Kc167 cells, genes within TADs tend to be more correlated in expression amongst them compared to genes lying in the nearby TADs. For this we used a subset of consecutively arranged TADs that have more than one gene inside them. We then used ANOVA to test whether variability gene expression within TADs is different from variability between TADs. As seen in Supplementary Figure 2c, many TADs pairs are significantly different from each other, while very few TADs are significantly different if we randomly assign genes to TADs.

Identification of boundary motifs. We took the list of boundaries and expanded them by 500 bp on each side. To avoid false positives, repetitive regions from the sequences of those boundaries were replaced by ‘N’s and any region with >10% of ‘N’s was removed. We used MEME-chip²⁰ to identify enriched DNA motifs; MEME-chip internally computes motifs using two methods, DREME⁶⁰ and MEME²⁷. DREME aims to quickly identify short motifs while MEME identifies larger overrepresented sequences (at the expense of significantly longer processing times).

To obtain the position-weight matrices of insulator motifs we ran MEME-chip²⁰ on the peaks called using MACS2⁶¹. We selected the highest scoring motif for each case which invariably corresponded to the motif reported for the protein. ChIP-seq data sources are found in Supplementary Table 4.

Enrichment of motifs using control background. For promoter boundaries a control background composed of all *Drosophila* gene promoters was used to test the enrichment. We downloaded *Drosophila* genes (dm3 assembly) from UCSC table browser⁶² and selected the sequences 200 bp upstream and 50 bp downstream of the transcription start site as core promoter sequences.

We classified these promoter sequences as boundary if they were within 500 bp of a boundary, or non-boundary (control background) if they were farther than 2000 bp from a boundary. Repetitive regions from the sequences of those promoters were replaced by ‘N’s and any region with >10% of ‘N’s was removed. In total, 10,529 background promoters and 1944 boundary promoters were used.

We used two different methods to assess the enrichment of the de novo and known motifs in boundary promoters with respect the control background, namely Ame²¹ from the MEME-suite and a method based on the predicted binding affinity given by TRAP²² that works as follows: for each motif, the log(TRAP score) distribution was computed for both the background and the boundary promoters. The Wilcoxon rank-sum test was then used to test for differences in the distributions. The *p*-values obtained were corrected using FDR. For Ame, we use total hits as scoring method and Fisher’s test for estimating enrichments. We tested

all de novo motifs identified either by MEME or by DREME and all known motifs associated to insulators and CP190 co-factors, as well as all core-promoter motifs.

We also used as control active genes in Kc167. To make this control, we selected those genes that overlapped with the yellow and red chromatin states from ref.²⁹ that are indicative of active chromatin in *Drosophila* Kc cells. The enrichment results were similar to the ones using a more broader list of genes for background.

For the boundaries that are not at promoters we used non-promoter open chromatin sequences obtained from DNase-seq²⁸ as control. In this case, we used 1665 background open chromatin regions and 655 non-promoter boundaries.

Pausing index. Pausing index for all *D. melanogaster* promoters was computed as the ratio of Pol-II ChIP-seq coverage at promoter over coverage at gene body. We used the ChIP-seq data for RNA Pol-II from Li et al.¹⁵. The promoter region was defined as in the previous section (200 bp downstream, 50 bp upstream of transcription start site). The gene body was defined as the region between 50 bp downstream of the transcription start site and the gene end. We used the maximum coverage for the promoters and the median coverage for the gene body.

Processing of ChIP-seq data. The data sources are listed in Supplementary Table 4. For each ChIP-seq data used, we downloaded the respective fastq files and aligned them in the dm3 assembly using Bowtie2⁶³. MACS2 was used to identify peaks for each of the proteins⁶¹. For the respective data sources we downloaded input sequences and aligned them as the ChIP-seq data. bamCompare and bamCoverage from deepTools⁵⁶ were used to create normalized coverage bigwig files.

MEME-chip²⁰ was used to identify motifs based on the MACS2 peaks. The resulting motifs can be seen in Supplementary Table 1.

Clustering of motifs. We used the promoters (200 bp upstream 50 bp downstream) annotated as boundaries and computed the log(TRAP score) for the Beaf-32 motif, motif-1 (M1BP), motif-6, motif-7 (ZIPIC), and motif-8. The scores for each motif were converted to bigwig files and clustered using hierarchical clustering from deepTools²⁹.

All boundaries that were further than 2000 bp of a promoter were centered at the nearest CP190 ChIP-seq peaks within 2000 bp, otherwise the boundary position was not modified. Log(TRAP score) for CTCF, Ibf, and Su(Hw) were computed for these regions and clustered as previously described.

We used hierarchical clustering based on Euclidian distance and the Ward method. The cluster number used was 13 for promoter boundaries and 9 for non-promoter boundaries. In each case, the group composed only of low-TRAP scores was removed. After clustering, the groups were manually ordered to produce the left panel of Fig. 3c. Scale of each heatmap was manually adjusted based on the range of TRAP scores found at the clusters for each motif (Supplementary Figure 3a). The log2 ratio of ChIP-seq/input for the different proteins was used for the center and right panels of Fig. 3c. Each heatmap is centered on the boundary and extended ±5000 bp. Scale of the heatmaps was adjusted based on the log2 ChIP/input for the protein in the respective cluster (Supplementary Figure 3b).

Motif presence. In general, sequence motifs occur with different strengths at different genomic loci, and the notion of presence/absence is largely dependent on arbitrary thresholds. Therefore much of our analysis is based on binding scores of motifs rather than their binary presence, which we invoke only for the purpose of visualization (Fig. 5e, Supplementary Figure 3c) and combinatorial motif analysis (Fig. 4). For Fig. 3d and Supplementary Figure 3, we considered a motif as present at a boundary if the TRAP score was equal or higher than the minimum log(TRAP score) identified for the clusters in Fig. 3c (the distribution of the log(TRAP scores) can be seen in Supplementary Figure 3a). The thresholds used were: ZIPIC motif –4.7, Beaf-32 –5, M1BP –4.5, motif-6 –3, motif-8 –2, Ibf –4, CTCF –4, and Su (Hw) –3. For GAF, Pita and Zw5 motifs we used FIMO⁶⁴ with the following parameters: ‘-max-strand --thresh 1e – 3’. For analysis of motif combinations at boundaries, we also require that the motifs are accompanied by the corresponding ChIP-seq peaks. For motif-6 and motif-8 whose binding proteins are not known, we require that the motif is on an accessible region. For this we use the peaks from the DNase-seq data²⁸.

Boundary prediction and feature ranking. We performed boundary prediction at all *Drosophila* promoters using motif TRAP scores for various transcription factors and DNase-seq signals as features. We utilized methods ranging from simple to complex (linear models, logistic regression, random forest, and stochastic gradient boosting), with the primary purpose to rank the features by importance in boundary prediction. Pre-filtering was done to remove highly correlated features (pearson *R* > 40%). Linear model and random forest was performed using the package Caret⁶⁵, while logistic regression was performed using package glmnet⁶⁶ in R.

Linear model was used with stepwise feature selection algorithm to predict boundary score from features by selecting the combination of features that minimizes the Akaike Information Criteria (AIC). Logistic regression, Random forest, and gradient boosting were used to classify the promoters into boundary and non-boundary, with additional feature selection performed using lasso, for logistic regression. We performed 10-fold cross validation while training all classification models. To evaluate model accuracy, the data were randomly divided

into training (60%) and test (40%) data sets and the sensitivity and specificity was calculated for test predictions. Lasso and gradient boosting models show highly similar sensitivity and specificity when used on an independent test data set, compared to when same data set was used for prediction, suggesting they are robust and less prone to overfitting.

Linear model predicted the boundary scores on the test data set with overall Spearman correlation of 37.6%, while logistic regression and random forest performed predictions with around 73–78% accuracy. After obtaining the best model in each scenario, we ranked the features by their importance in prediction, using the ‘varImp’ function from Caret. ‘varImp’ selects a variable importance predictor based on the model type, which is calculated for each parameter in the model (<https://topepo.github.io/caret/variable-importance.html>). Briefly, the importance score for linear model is the absolute value of the *t*-statistic for the model parameter, for lasso, it is the absolute value of final coefficients, for gbm it is the relative influence score as described in Friedman⁶⁷, and for random forest it is the difference between the classification error-rate for the out-of-bag portion of data and a permuted predictor variable, averaged over all trees and normalized by the standard deviation of the differences (https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#varimp). All importance scores are then scaled between 0 and 100 to compare them together.

Knockdown of M1BP and Beaf-32. S2 cells (obtained from Michael Boutros lab) were cultured in Express Five SFM (Thermo Fisher Scientific) supplemented with glutamax, at 27 °C at a density of 1–16 million/ml. All cell lines are regularly checked for the absence of mycoplasma by PCR detection kit (Jena Bioscience PP-401). dsRNA was generated using HiScribe T7 High Yield RNA Synthesis Kit (NEB) and purified with MEGAclear (Ambion). The dsRNA was heated to 65 °C for 5 min and left to cool to room temperature. Primers used for generating M1BP dsRNA are from ref.¹⁷. Primers used for generating Beaf-32 and GST dsRNABeaf-32_fwd 5'-taatacgaactataggAAATCACGAGGAGCTCACCAA-3'
Beaf-32_rev 5'-taatacgaactataggCTACTCATCCTTGGCAACCG-3'
GST_fwd 5'-taatacgaactataggAGATATTGGATATTAGATACCGT-3'
GST_rev 5'-taatacgaactataggAGATTTGGATATTAGATACCGT-3'

For knockdown experiments, 7.5 million cells were transfected with 100 µg of dsRNA on 10 cm dishes, using Lipofectamine RNAiMAX Reagent (Thermo Fisher Scientific). The M1BP knockdown lasted for 7 days while that of Beaf-32 lasted for 4 days. For M1BP-Beaf-32 double knockdown, cells were first treated with M1BP dsRNA at day 1, followed by treatment with Beaf-32 dsRNA at day 3, and were collected at day 7. All knockdown experiments were performed in two biological replicates and samples were processed in the same batch.

Western blot to determine knockdown efficiency. Protein depletion efficiencies were verified with western blot using specific antibodies against Beaf-32 (1:2000, Creative Diagnostics, CABT-BL422), M1BP (1:1000, gift from David Gilmour, PennState) and Actin (1:2000, Santa Cruz Sc-1616). Blots were incubated overnight at 4 degrees. HRP-coupled goat Anti-Rabbit IgG (1:10,000, GE healthcare) was used as secondary antibody, followed by incubation for 1 h at room temperature. Results were measured using ChemiDoc Imaging system (Bio Rad).

In situ Hi-C of knockdown and control cells. We used a modified version of the in situ Hi-C protocol⁵ to generate our Hi-C maps. All samples were processed in the same batch. S2 cells were fixed with 2% formaldehyde for 10 min at room temperature. Glycine at 125 mM final concentration was added to the plates followed by 5 min incubation. After two washes in PBS, cells were scraped off the plate and pelleted. Each pellet (10–50 million of cells) was resuspended in 1 ml of lysis buffer (10 mM Tris-HCl, pH 8, 10 mM NaCl, 0.2% IGEPAL CA-630) and nuclei were extracted by sonication following the NEXON protocol⁶⁸ (Covaris E220 sonicator, settings: 75 W peak power, 2% duty factor, 200 cycles/burst, for 60–90 s until about 70% of intact nuclei were released). Nuclei were pelleted and resuspended in 0.5% SDS to permeabilize the nuclear membrane and to make chromatin accessible for restriction digestion. After 10 min incubation at room temperature, SDS was quenched adding 1% Triton X-100 (final concentration) and 1X of NEBuffer 3.1 (NEB, B7203S). Nuclei were digested overnight at 37 °C on a rocking platform using DpnII (NEB, R0543M, ~1–5 units per million cells). Prior and after digestion, an aliquot of nuclei was set aside for digestion quality control and DNA quantification. Biotin incorporation was carried out in a final volume of 150 µl using these reaction conditions: 50 mM of each dNTPs, replacing dCTP with biotin-14-dCTP (Life Technologies, 19518-018), 1 U of Klenow (NEB, M0210L) per microgram of DNA, at 25 °C for 1 h to promote fill-in. Ligase mix was added (1X Ligation buffer NEB B0202, 0.8% Triton X-100, 0.1 mg/ml BSA, 2000 U T4 DNA ligase NEB M0202S, final sample volume 1.2 ml) and samples were incubated for 4 h at room temperature under rotation. Nuclei were pelleted, resuspended in SDS buffer (10 mM Tris-HCl, pH 8, 1 mM EDTA, 1% SDS), lysed, de-crosslinked and de-proteinized overnight at 68 °C. DNA was precipitated and sonicated to the size range of 100–600 bp. Biotinylated Hi-C DNA in 1X binding buffer (5 mM Tris-HCl, pH 8, 0.5 mM EDTA, 1 M NaCl) was pulled down using Dynabeads MyOne Streptavidin C1 (Life Technologies, 650.01), using 5 µl of beads per microgram of DNA, pre-washed in 1X binding buffer. Beads were washed twice in Tween wash buffer (5 mM Tris-HCl, pH 8, 0.5 mM EDTA, 1 M NaCl, 0.05% Tween-20) and twice in EB (10 mM Tris-HCl, pH 8). A small aliquot of beads-

bound DNA was kept to measure the DNA concentration, eluting the DNA by incubation at 98 °C for 10 min. Quantity of 10–50 ng of DNA bound to beads was used for library preparation using a modification of the NEBNext Ultra II DNA library preparation workflow (NEB, E7645). DNA bound to beads was then end-repaired, A-tailed, adaptor-ligated using manufacturer's instruction. Beads were reclaimed on a magnet, washed once in EB and eluted at 98 °C for 10 min. DNA was USER-treated to open the adapters in meantime to PCR amplification. Libraries were sequenced paired-end, with a read length of 75 bp, on Illumina HiSeq 3000.

Cell cycle analysis using FACS. Formaldehyde-fixed cells (2% for 10 min) were resuspended in 500 µl of permeabilization buffer (1% BSA, 0.5% Triton X-100 in PBS) and incubated for 15 min at room temperature under mixing. Cells were then incubated overnight at 4 °C with a 1:500 dilution of anti-H3 Ser10 phosphorylated antibody (Abcam, ab5176) and washed twice with permeabilization buffer. Cells were pelleted and resuspended in 400 µl of permeabilization buffer. One microliter of secondary antibody (anti-rabbit conjugated with Alexa 633 fluorophore, A21070, Life Technologies) was added and cells were incubated for 2 h at room temperature (protected from light). After two washes using the permeabilization solution, cells were resuspended in 500 µl of PBS and treated with 5 µl of RNase A for 30 min at room temperature. Nuclei were stained using 1 µg/ml of DAPI prior sorting. Samples were analyzed by flow cytometry using BD LSRII Fortessa instrument. The data have been analyzed using FACSDiva.

Hi-C processing of Beaf-32 and M1BP knockdowns. Hi-C samples for GST control and knockdowns were processed as all other Hi-C samples using HiCExplorer. The bins are based on the restriction fragments for DpnII. The statistics of data processing are shown in Supplementary Figure 6d-f and Supplementary Table 5.

Code availability. HiCExplorer code is available online at: <https://github.com/deeptools/HiCExplorer/>. HiCBrowser code is available at: <https://github.com/deeptools/HiCBrowser>

Data availability. The processed data from ChIP-Seq and Hi-C samples obtained from online sources can be found at: http://chorogenome.ei-freiburg.mpg.de/data_sources.html. Sequencing data for the *in situ* Hi-C experiments described in this article are available at NCBI GEO under accession: GSE97965.

Received: 26 June 2017 Accepted: 6 December 2017

Published online: 15 January 2018

References

- Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin domains: the unit of chromosome organization. *Mol. Cell* **62**, 668–680 (2016).
- Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
- Sexton, T. et al. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458–472 (2012).
- Hou, C., Li, L., Qin, Z. S. & Corces, V. G. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Mol. Cell* **48**, 471–484 (2012).
- Rao, S. S. P. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
- Sanborn, A. L. et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. USA* **112**, E6456–E6465 (2015).
- Nichols, M. H. & Corces, V. G. A CTCF code for 3D genome architecture. *Cell* **162**, 703–705 (2015).
- Lupiáñez, D. G. et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).
- Nora, E. P. et al. Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* **169**, 930–944.e22 (2017).
- Van Bortle, K. et al. Insulator function and topological domain border strength scale with architectural protein occupancy. *Genome Biol.* **15**, R82 (2014).
- Zolotarev, N. et al. Architectural proteins Pita, Zw5, and ZIPIC contain homodimerization domain and support specific long-range interactions in Drosophila. *Nucleic Acids Res.* **44**, 7228–7241 (2016).
- Ong, C.-T. & Corces, V. G. CTCF: an architectural protein bridging genome topology and function. *Nat. Rev. Genet.* **15**, 234–246 (2014).
- Maksimenko, O. et al. Two new insulator proteins, Pita and ZIPIC, target CP190 to chromatin. *Genome Res.* **25**, 89–99 (2015).

14. Cuartero, S., Fresán, U., Reina, O., Planet, E. & Espinàs, M. L. Ibf1 and Ibf2 are novel CP190-interacting proteins required for insulator function. *EMBO J.* **33**, 1–11 (2014).
15. Li, L. et al. Widespread rearrangement of 3D chromatin organization underlies polycomb-mediated stress-induced silencing. *Mol. Cell* **58**, 216–231 (2015).
16. Cubéñas-Potts, C. et al. Different enhancer classes in Drosophila bind distinct architectural proteins and mediate unique chromatin interactions and 3D architecture. *Nucleic Acids Res.* **45**, 1714–1730 (2017).
17. Li, J. & Gilmour, D. S. Distinct mechanisms of transcriptional pausing orchestrated by GAGA factor and M1BP, a novel transcription factor. *EMBO J.* **32**, 1829–1841 (2013).
18. Hug, C. B., Grimaldi, A. G., Kruse, K. & Vaquerizas, J. M. Chromatin architecture emerges during zygotic genome activation independent of transcription. *Cell* **169**, 216–228.e19 (2017).
19. Mourad, R., Li, L. & Cuvier, O. Uncovering direct and indirect molecular determinants of chromatin loops using a computational integrative approach. *PLoS Comput. Biol.* **13**, e1005538 (2017).
20. Ma, W., Noble, W. S. & Bailey, T. L. Motif-based analysis of large nucleotide data sets using MEME-ChIP. *Nat. Protoc.* **9**, 1428–1450 (2014).
21. McLeay, R. C. & Bailey, T. L. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinform.* **11**, 165 (2010).
22. Thomas-Chollier, M. et al. Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. *Nat. Protoc.* **6**, 1860–1869 (2011).
23. Ohler, U., Liao, G.-C., Niemann, H. & Rubin, G. M. Computational analysis of core promoters in the Drosophila genome. *Genome Biol.* **3**, RESEARCH0087 (2002).
24. Gurudatta, B. V., Yang, J., Van Bortle, K., Donlin-Asp, P. G. & Corces, V. G. Dynamic changes in the genomic localization of DNA replication-related element binding factor during the cell cycle. *Cell Cycle* **12**, 1605–1615 (2013).
25. Jain, D., Baldi, S., Zabel, A., Straub, T. & Becker, P. Active promoters give rise to false positive ‘phantom peaks’ in ChIP-seq experiments. *Nucleic Acids Res.* **43**, 6959–6968 (2015).
26. Landt, S. G. et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).
27. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28–36 (1994).
28. Celniker, S. E. et al. Unlocking the secrets of the genome. *Nature* **459**, 927–930 (2009).
29. Filion, G. J. et al. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell* **143**, 212–224 (2010).
30. Belozeroval, V. E., Majumder, P., Shen, P. & Cai, H. N. A novel boundary element may facilitate independent gene regulation in the Antennapedia complex of Drosophila. *EMBO J.* **22**, 3113–3121 (2003).
31. Schweinsberg, S. et al. The enhancer-blocking activity of the Fab-7 boundary from the Drosophila bithorax complex requires GAGA-factor-binding sites. *Genetics* **168**, 1371–1384 (2004).
32. Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. & Chen, L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.* **30**, 90–98 (2011).
33. Goecks, J., Nekrutenko, A. & Taylor, J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11**, R86 (2010).
34. Ulianov, S. V. et al. Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Res.* **26**, 70–84 (2016).
35. Lam, K. C. et al. The NSL complex regulates housekeeping genes in Drosophila. *PLoS Genet.* **8**, e1002736 (2012).
36. Feller, C. et al. The MOF-containing NSL complex associates globally with housekeeping genes, but activates only a defined subset. *Nucleic Acids Res.* **40**, 1509–1522 (2012).
37. Van Bortle, K. et al. Drosophila CTCF tandemly aligns with other insulator proteins at the borders of H3K27me3 domains. *Genome Res.* **22**, 2176–2187 (2012).
38. Liang, J. et al. Chromatin immunoprecipitation indirect peaks highlight long-range interactions of insulator proteins and Pol II pausing. *Mol. Cell* **5**, 1–10 (2014).
39. Wood, A. M. et al. Regulation of chromatin organization and inducible gene expression by a Drosophila insulator. *Mol. Cell* **44**, 29–38 (2011).
40. Pascual-García, P. et al. Metazoan nuclear pores provide a scaffold for poised genes and mediate induced enhancer-promoter contacts. *Mol. Cell* **66**, 63–76, e6 (2017).
41. Emberly, E. et al. BEAF regulates cell-cycle genes through the controlled deposition of H3K9 methylation marks into its conserved dual-core binding sites. *PLoS Biol.* **6**, 2896–2910 (2008).
42. Tue, N. T. et al. DREF plays multiple roles during Drosophila development. *Biochim. Biophys. Acta* **1860**, 705–712 (2017).
43. Schwartz, Y. B. et al. Nature and function of insulator protein binding sites in the Drosophila genome. *Genome Res.* **22**, 2188–2198 (2012).
44. Narendra, V. et al. CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science* **347**, 1017–1021 (2015).
45. Guo, Y. et al. CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. *Cell* **162**, 900–910 (2015).
46. Cuddapah, S. et al. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res.* **19**, 24–32 (2009).
47. Iwasaki, O. et al. Interaction between TBP and condensin drives the organization and faithful segregation of mitotic chromosomes. *Mol. Cell* **59**, 755–767 (2015).
48. Strick, T. R., Kawaguchi, T. & Hirano, T. Real-time detection of single-molecule DNA compaction by condensin I. *Curr. Biol.* **14**, 874–880 (2004).
49. Alipour, E. & Marko, J. F. Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Res.* **40**, 11202–11212 (2012).
50. Fudenberg, G. et al. Formation of chromosomal domains by loop extrusion. *Cell Rep.* **15**, 2038–2049 (2016).
51. Rowley, M. J. et al. Evolutionarily conserved principles predict 3D chromatin organization. *Mol. Cell* **67**, 837–852.e7 (2017).
52. Imaikaev, M. et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* **9**, 999–1003 (2012).
53. Ramírez, F. et al. High-affinity sites form an interaction network to facilitate spreading of the MSL complex across the X chromosome in Drosophila. *Mol. Cell* **60**, 146–162 (2015).
54. Shin, H. et al. TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res.* **44**, e70 (2016).
55. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
56. Ramírez, F. et al. DeepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
57. Gravely, B. R. et al. The developmental transcriptome of Drosophila melanogaster. *Nature* **471**, 473–479 (2011).
58. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016).
59. Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
60. Bailey, T. L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27**, 1653–1659 (2011).
61. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* **7**, 1728–1740 (2012).
62. Karolchik, D. et al. The UCSC table browser data retrieval tool. *Nucleic Acids Res.* **32**, D493–D496 (2004).
63. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
64. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
65. Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw. Artic.* **28**, 1–26 (2008).
66. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
67. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
68. Arrigoni, L. et al. Standardizing chromatin research: a simple and universal method for ChIP-seq. *Nucleic Acids Res.* **44**, e67 (2016).
69. Zabidi, M. A. et al. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**, 556–559 (2015).
70. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph.* **20**, 1983–1992 (2014).
71. Gaszner, M., Vazquez, J. & Schedl, P. The Zw5 protein, a component of the scs chromatin domain boundary, is able to block enhancer-promoter interaction. *Genes Dev.* **13**, 2098–2107 (1999).

Acknowledgements

We would like to thank David Gilmour (PennState University) for providing the M1BP antibody. Also, we would like to thank Ulrike Bönisch and her team at the Deep-Sequencing Facility and Andrea Würch from the FACS facility of the MPI-IE. We further thank Wilhelm Rüsing for IT support, Diana Santacruz for critical reading of the manuscript and Gina Renschler for the help in interpreting FACS results. We also thank Victor Corces Lab whose published data have been instrumental in this research. This work was supported by the German Research Foundation [SFB 992 “Medical Epigenetics”] awarded to A.A. and T.M. and a grant from the Federal Ministry of Education and Research through the German Epigenome Programme DEEP [01KU1216G] awarded to T.M.

Author contributions

F.R. developed HiCExplorer, HiCBrowser and the Chorogenome navigator with support from V.B., B.A.G. and J.V. F.R. and V.B. performed all the analysis and designed the experiments, with inputs from T.M. L.A. and K.C.L. performed the experiments. T.M. supervised all aspects of the analysis, while B.H. and A.A. supervised J.V. and K.C.L., respectively. F.R. and V.B. wrote the manuscript with input from all other authors.

Additional information

Supplementary Information accompanies this paper at <https://doi.org/10.1038/s41467-017-02525-w>.

Competing interests: The authors declare no competing financial interest.

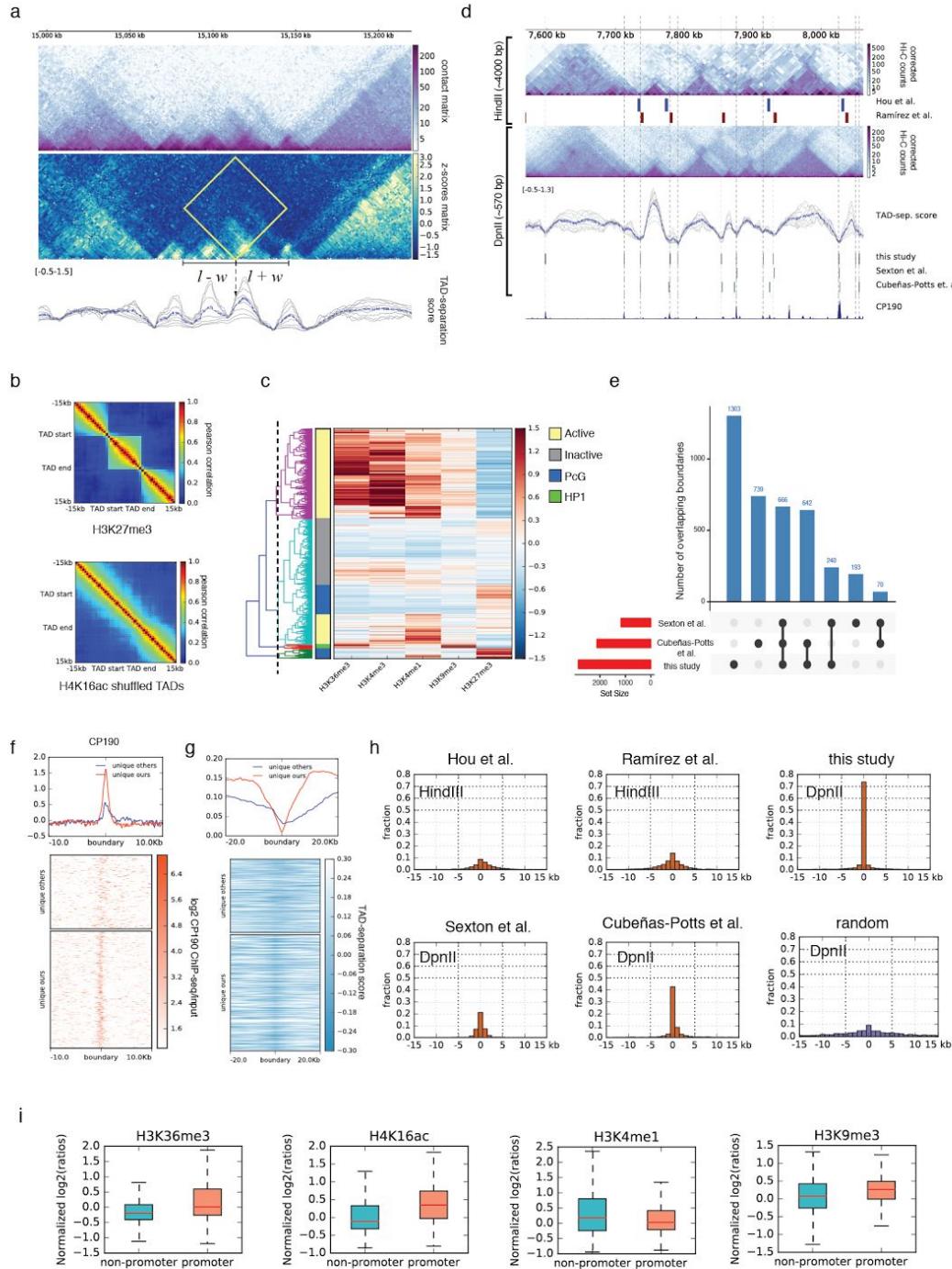
Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

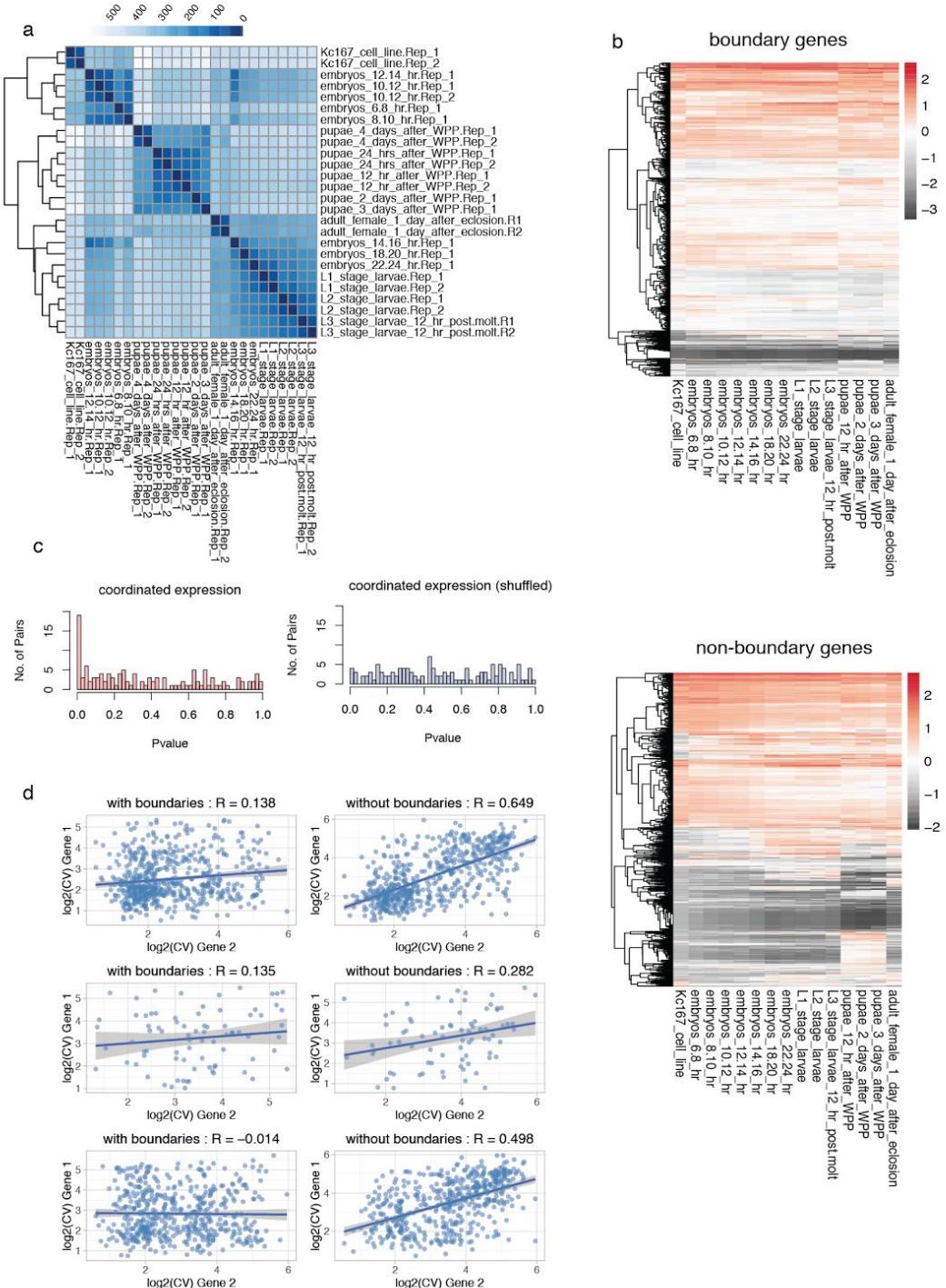


Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

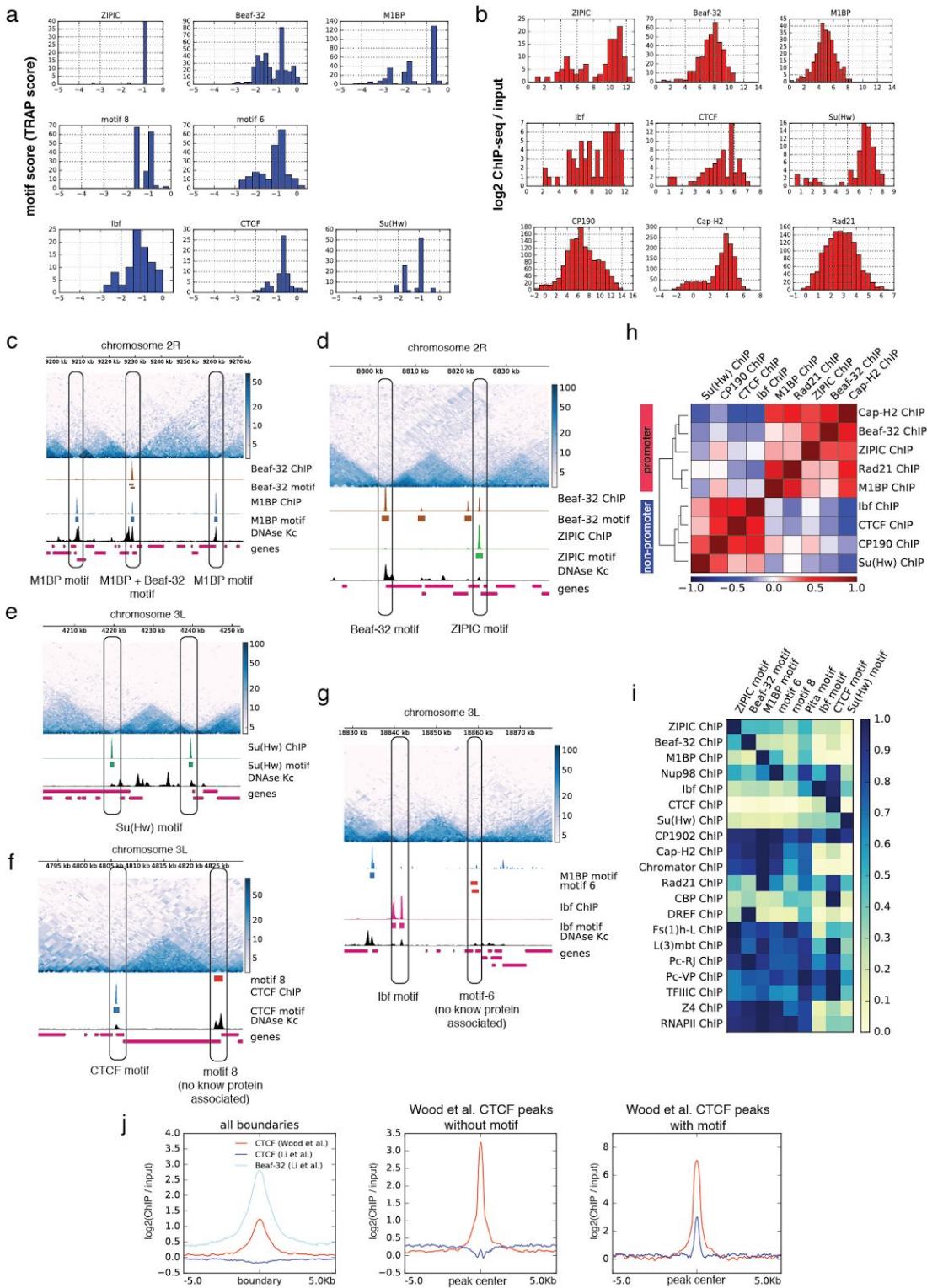
© The Author(s) 2018



Supplementary Figure 1. Detection of TAD boundaries and assessment of their quality. **a.** Method to identify boundaries (see ‘methods’ for details). Top, Hi-C contact matrix counts. Middle, z-score transformed matrix. Last: TAD-separation score at different window lengths (gray lines) and mean score (blue line). The TAD-separation score at bin (l) corresponds to the mean values of the z-scores inside the ‘diamond’, which correspond to the contacts between a region of width w on left and on right. **b.** Similar to Fig. 1c, Pearson correlations of histone marks within and outside TADs (+15kb flanking regions) and for a random placement of boundaries. **c.** Hierarchical clustering of TADs based on modENCODE histone marks for Kc167 cells. Manual annotations for active, inactive, Pcg and HP1 are based on the clustering results (see methods). **d.** Example genomic location comparing published boundaries with the ones generated in this study. Top, HindIII based Hi-C contact matrix counts for S2 cell line¹ and boundaries reported by Hou et al.² (Kc167) and Ramirez et al.¹ (S2). Bottom, DpnII based Hi-C contact matrix and the boundaries from this study, Sexton et al.³ and Cubañas-Potts et al.⁴. Below the Hi-C heatmap is the TAD-separation score as in Fig. 1a. The last track shows CP190 ChIP-seq signal⁵. The vertical lines correspond to the boundaries in this study. **e.** Overlap of boundaries based on Hi-C DpnII experiments. The bars show the overlap between the indicated sets below (black dots). Two boundaries were considered overlapping if they were within 2000 bp from each other. The intersections were plotted using UpSetR¹. **f.** Comparison of unique boundaries in our study with the unique boundaries in previous studies^{2,3} with respect to CP190. Our unique boundaries overlap more frequently with CP190. **g.** Similar to f, but comparing TAD separation score. **h.** Histograms of the distance of our boundaries and other published boundary sets to CP190 peaks. The ‘random’ dataset contains our boundary set randomly shuffled (see methods). **i.** As in Fig. 1h, modENCODE histone marks at non-promoter and promoter boundaries. In all cases the promoter boundaries are associated significantly to the active marks (H3K36me3 and H4K16ac, p -value $\leq 3.106261 \times 10^{-17}$ Wilcoxon rank-sum test).

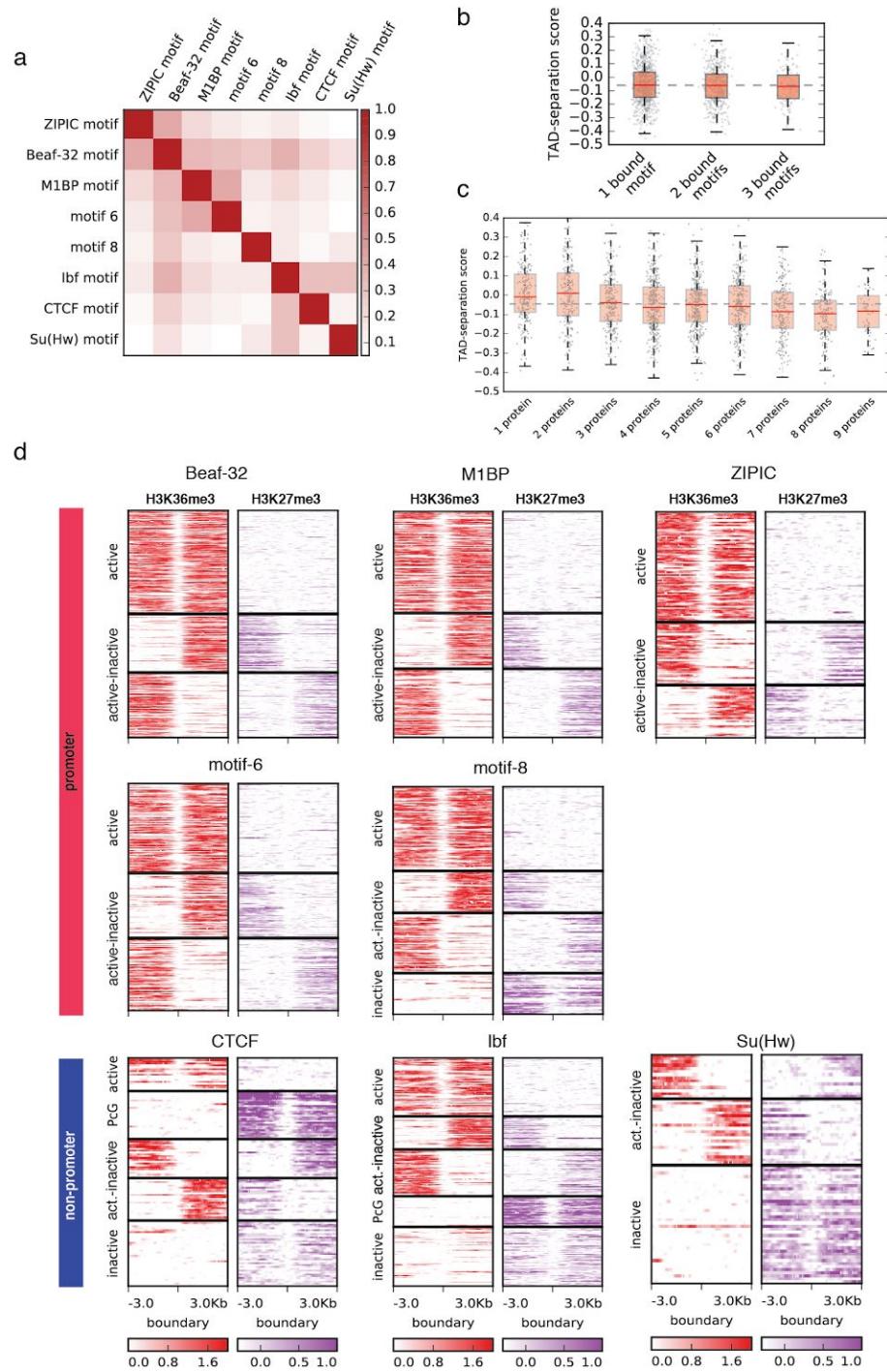


Supplementary Figure 2. Gene expression is coordinated inside TADs. **a.** Euclidean distances between RNA-Seq samples from modENCODE used in this study. Replicates were later merged (mean) for the analysis. **b.** Clustering of genes by expression in Kc167 cells and at different developmental stages. Genes lying on either side of TAD boundaries tend to show consistent expression (top), while genes within TADs show variable expression (bottom) during development (color bar : row-wise z-score). Genes without boundaries were sampled randomly to the same number as genes with boundaries. **c.** P-values from ANOVA between genes within pairs of adjacent TADs (see methods). Expression within TADs is more coordinated (left) compared to genes randomly assigned to TADs (right). **d.** Same as Fig. 2d, here the adjacent gene-pairs are separated by their relative orientation: divergent (top), convergent (middle), and tandem (below) pairs. Gene-pairs without boundaries were sampled randomly to the same number as genes with boundaries. Line shows the linear model fit (shaded region: std. error).

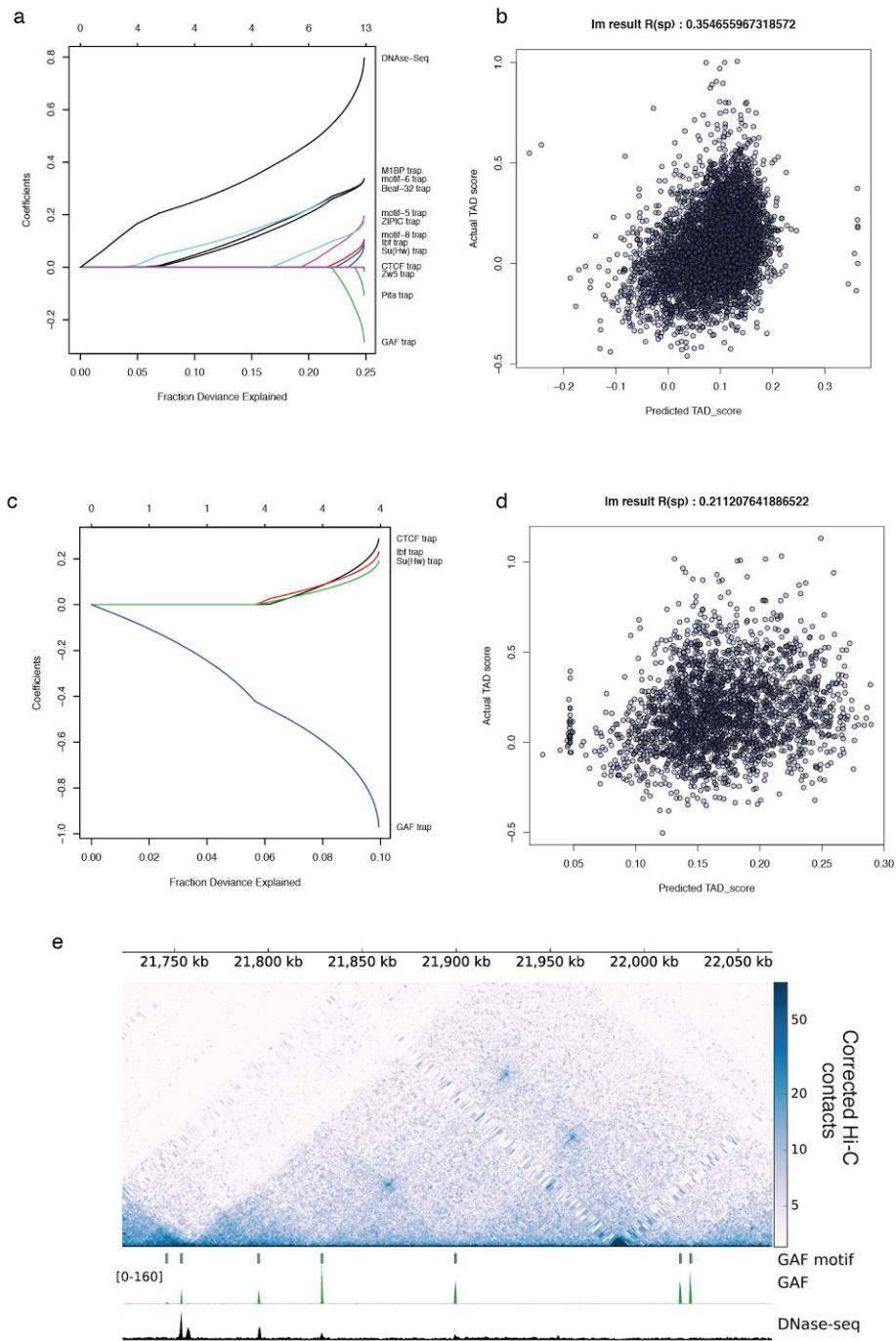


Supplementary Figure 3. Boundary motifs and their relationship to the ChIP-Seq profiles.

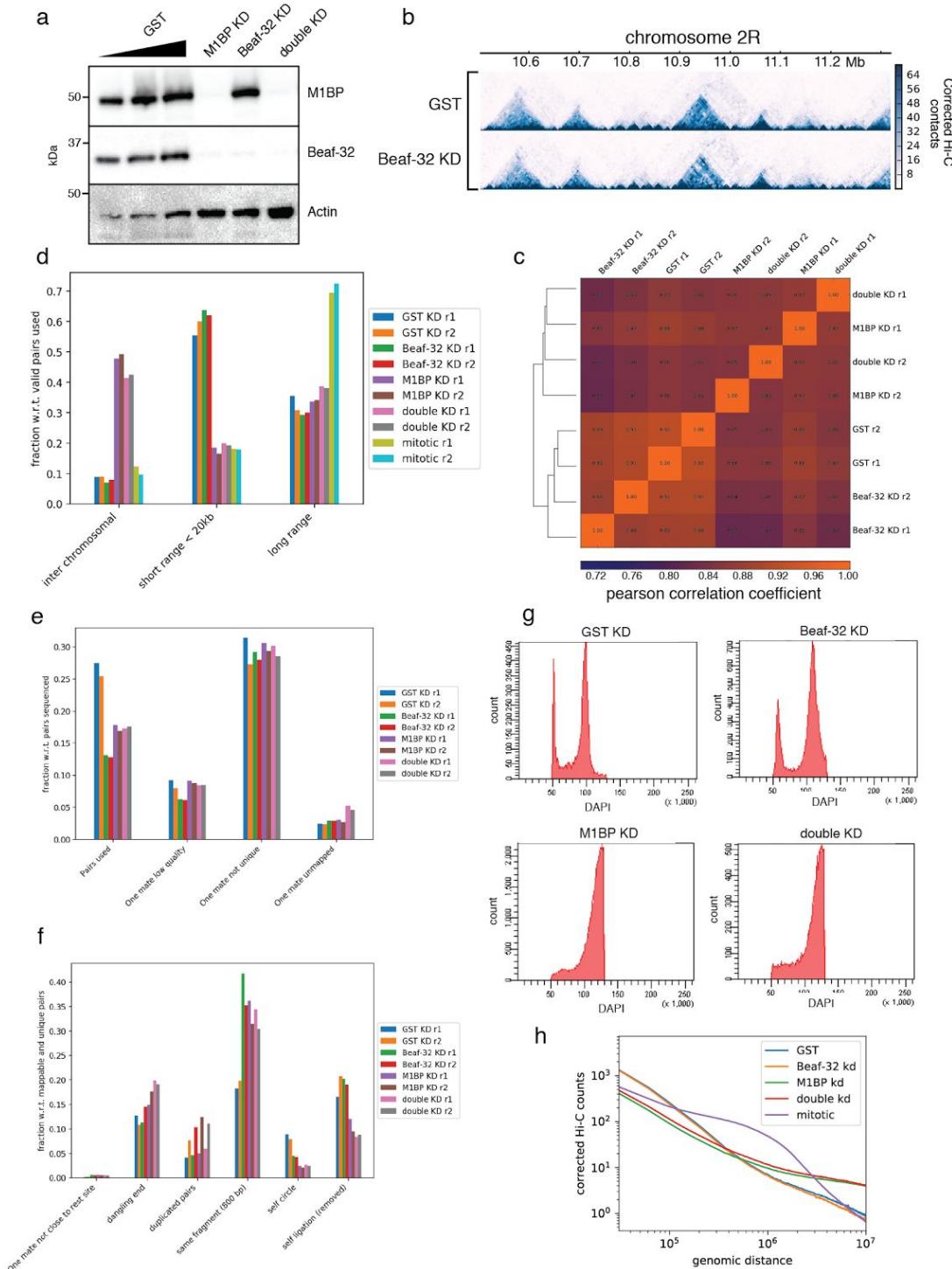
a. TRAP scores for each of the TAD boundary motifs within their cluster. **b.** ChIP-seq log2 ratio (ChIP/input) for the insulator proteins within their cluster. In the case of CP190, Cap-H2 and Rad21 the histogram contains the values over all boundaries. **c-g.** Examples of distinct insulators at boundaries. As in Fig. 1a, the top track shows Hi-C corrected counts. **h.** Heatmap showing Pearson correlations of ChIP-seq log2 ratios (IP / input) measured at TAD boundaries. The *complete* linkage method (also known as furthest neighbour clustering) was used for the hierarchical clustering. **i.** As in Fig. 3d. Each cell in the matrix contains the mean fold change of all respective ChIP-seq peaks having the motif. For each row, the maximum fold change was scaled to 1. **j.** Comparison of CTCF ChIP-seq experiments from Wood et. al.⁴ and Li et. al.⁵. The first panel contains the mean values over all boundaries, the middle panel contains mean values for all CTCF peaks from Wood et. al.⁴ that do not have the CTCF motif, and the last panel contains CTCF peaks from Wood et. al.⁴ that have the CTCF motif. The CTCF ChIP-seq from Li et. al.⁵ only shows enrichment when the CTCF motif is present while the CTCF ChIP-seq from Wood et al.⁶ has unspecific bindings.



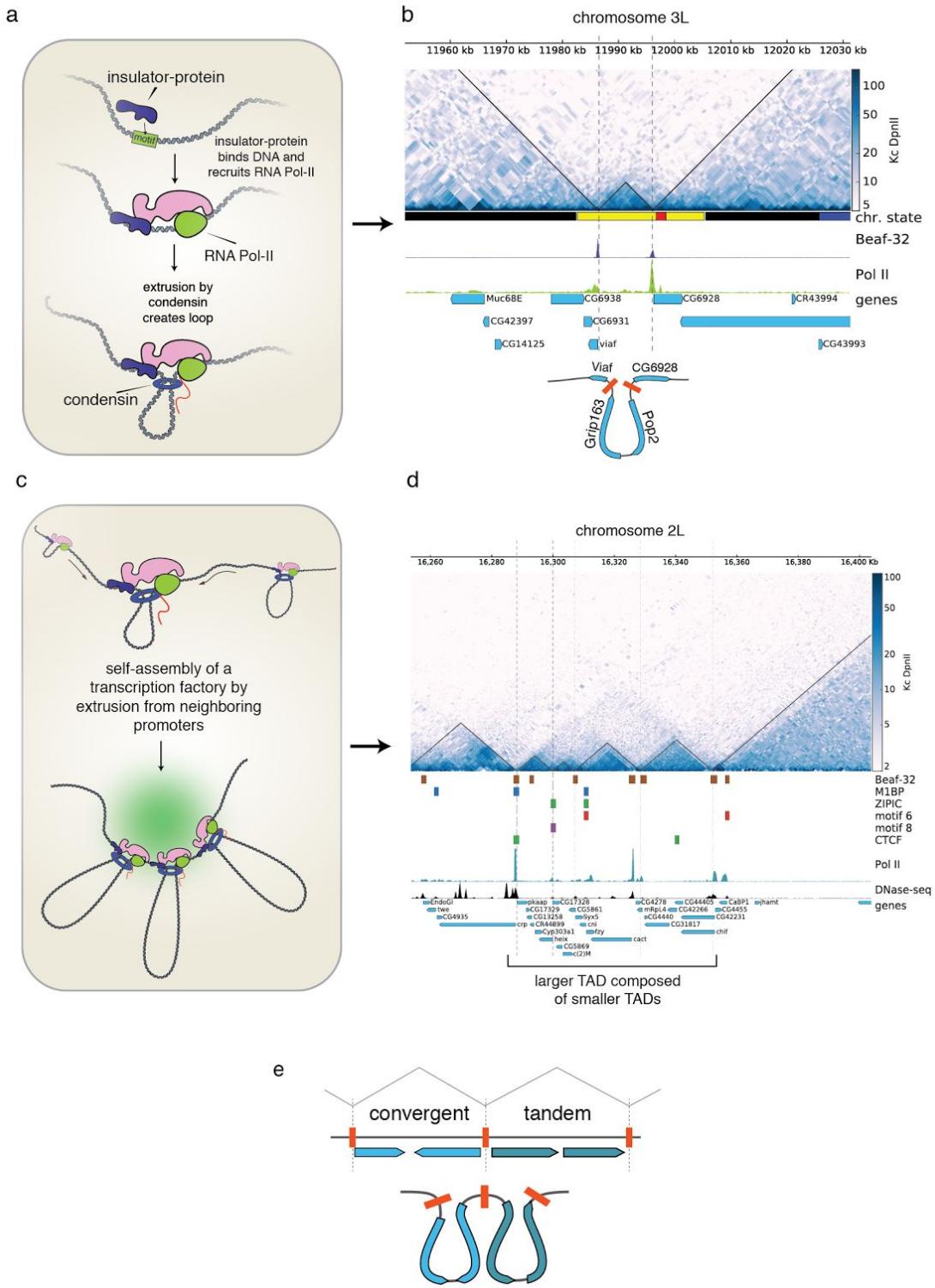
Supplementary Figure 4. Effect of motif combinations on boundary strength and chromatin marks. **a.** Heatmap of the overlap coefficient between boundary motifs. For each pair of motifs, the overlap coefficient is defined as $\text{Overlap-coefficient}(A, B) = A \cap B / \min(|A|, |B|)$ where A, B are the sets of all boundaries containing either motif A or B. **b.** TAD separation at boundaries containing one, two or three insulator motifs. Virtually no difference in boundary strength is observed with the number of insulator motifs present at boundaries. **c.** TAD separation score at boundaries bound by 1 up to 10 proteins known to be associated with boundaries (discarding information about motif enrichment). If we do not consider motif information, some variation can be seen in boundary strength associated to the number of bound proteins, especially between 2 to 3 boundary proteins and between 6 and 7 proteins ($p\text{-value} \leq 0.01$ Wilcoxon rank sum test). **d.** Normalized log₂ ChIP/input at transitions from different histone marks at boundaries for the active chromatin H3K36me3 and the repressive chromatin mark H3K27me3 on a 6 kb region centered at the boundaries. We performed k-means clustering using deepTools⁶. For CTCF and lbf five clusters were used to distinguish boundaries between Polycomb group TADs. The low histone mark values at the boundaries (white color running at the center of the heatmaps) are indicative of nucleosome free regions. The polycomb group repressed chromatin (PcG) is characterized by higher intensities of H3K27me3 compared to the inactive chromatin.



Supplementary Figure 5. Results from lasso and linear model predictions. **a.** Lasso penalized coefficients for promoters. Open chromatin (DNase-seq), followed by Beaf-32, M1BP and motif-6 are the three top predictors to classify promoters as TAD boundaries, while GAF and Pita are negatively associated. **b.** At promoters: TAD score predictions using linear model on an independent test dataset. The predicted scores correlate with the actual TAD-separation scores on promoters. **c.** Lasso penalized coefficients for non-promoter open sites. TRAP score signal for CTCF, Ibf1 and Su(Hw) are positively correlated with boundaries, while GAF shows negative correlation. **d.** At non-promoter open sites : predicted TAD scores from linear model on an independent test dataset. **e.** GAF motif, which is negatively associated with promoter and non-promoter boundaries, can be found alone along with the protein at loop domains.



Supplementary Figure 6. Beaf-32 and M1BP knockdowns Hi-C. **a.** Western blots showing protein levels of M1BP and Beaf-32 (compared to GST control) after RNAi treatment in S2 cells. **b.** Example region showing contacts for all GST and Beaf-32 KD. All matrices were normalized to match the total number of contacts of the smaller matrix. The bin size used was 3kb. **c.** Pearson correlation of corrected Hi-C matrices. The correlation was done between Hi-C bins within 50 kbp. **d.** Fraction of Hi-C pairs classified as inter-chromosomal by *cis* distancer. Mitotic data values were added for comparison from Hug et al.⁷ The total number of reads sequenced are listed in Supplementary Table 5. **e.** Fraction of valid Hi-C pairs used compared to low quality, unmapped and non-unique. **f.** Fraction of Hi-C pairs that are filtered by various reasons. These plots is part of the QC module of HiCExplorer. **g.** FACS analysis using DAPI for GST, Beaf-32, M1BP and M1BP + Beaf-32 knockdowns. M1BP knockdown affects cell growth and causes an arrest in cell-cycle not seen in GST or Beaf-32 KD. **h.** Genomic distance vs. Hi-C counts. Larger matrices were scaled down to match the sum of the smallest matrix. Only replicate 1 of all experiments was used. The ‘mitotic’ data is based on the Hi-C data on mitotic fly cells from Hug et. al. ⁷.



Supplementary Figure 7. Model for the formation of TADs. **a.** Top: Binding of an insulator protein to its cognate motif at a core promoter facilitates RNA Pol-II initiation which in turn recruits condensin complex which initiates extrusion⁸. **b.** As shown by simulations^{9,10} TADs in Hi-C contact maps are consistent with loop extrusion. Here we show a small active region between two large inactive TADs. The active region contains two genes and is flanked by the Beaf-32 insulator. The expected loop from this region is shown below. **c.** Neighboring promoters, each extruding a loop, end up meeting each other to form a rosette reminiscent of proposed transcription factories^{11,12}. Such clusters, are also observed in a recent single-nucleus HiC study¹³. **d.** The TADs demarcated by the insulators motif and RNA Pol-II form a larger TAD. This implies that the smaller TADs remain in closer contact to each other compared to the surrounding inactive chromatin. This hierarchical TAD structure resembles the proposed structure of transcription factories. **e.** Insulators at divergent gene promoters serve as good anchors, by recruiting two Pol-II/Condensin machines in both directions. Red bars represent insulator motifs at promoters and dotted lines show the location of boundaries. The expected loops from this region are shown below.

Supplementary Table 1. Enriched motifs found at ChIP-seq peaks.

MOTIF Protein	Beaf-32	M1BP	Pita	CTCF	Ibf1/2	ZIPIC	Zw5	Su(Hw)	GAF	CP190	Mod (mdg4)	Cap-H2	Chromator	Rad21
Beaf-32 motif	285(600) 3.3e-230	95(600) 4.2e-18	85(600) 1.2e-014			119(600) 5.2e-72				51(300) 2.8e-5		219(600) 3.0e-187	102(300) 5.5e-52	28(300) 9.9e-8
M1BP motif	30(600) 9.5e-7	519(600) 6.0e-101 ²	73(600) 8.4e-054			47(600) 2.3e-6	28(300) 2.1e-10			88(300) 3.3e-42		102(600) 3.4e-98	88(300) 1.2e-074	84(300) 4.2e-97
Pita motif			114(600) 1.1e-093		56 (300) 3.8e-21									
CTCF motif	24(600) 1.1e+3		74(600) 1.1e-53	272(600) 5.0e-562	75(300) 9.3e-92	73(600) 1.2e-22				32(300) 1.0e-20	42(300) 9.4e-44			
Ibf1/2 motif				3.5e-008 (dreme)	195(300) 2.6e-49					2.5e-9 (dreme)				
ZIPIC motif	168(600) 7.6e-55	19(600) 1.4	108(600) 3.3e-108			356(600) 6.1e-256				28(300) 3.6e-4		120(600) 1.2e-61		
Zw5 motif							87(300) 3.1e-98							
Su(Hw) motif								175(300) 3.8e-208		66(300) 7.1e-60	27(300) 9.4e-13		24(300) 5.4e-9	
GAF motif									67(300) 1.8e-152		27(300) 9.4e-13			
Ohler motif 5 motif	40(600) 3.5e+6	27(600) 5.5e-1	48(600) 1.8e-21			71(600) 2.7e-23	27(300) 8.9e-12			26(300) 5.3e-6				
Ohler motif 6 motif		138(600) 9.6e-94				40(600) 3.0e-11				67(300) 4.0e-10		67(600) 4.8e-14		
Ohler motif 8 motif	30(600) 9.5e-7	12(600) 6.1e+8			54(300) 1.0e-7	30(600) 4.7e+3				49(300) 1.6e-2	42(300) 3.4e-9	18(600) 2.4e-12		98(300) 3.0e-7
A-rich repeat									128(300) 6.4e-15				118(300) 2.8e-105	
(CA)n repeat						40(600) 1.0e-11		43(300) 1.7e-42	77(300) 2.5e-276					35(300) 2.0e-22

The table shows number of sites with motifs (total number of sites) and MEME¹⁴ E-value. In some cases the DREME¹⁵ E-value is shown.

Supplementary Table 2. Comparison of motif enrichment using different sets of boundaries.

MOTIF	DE-NOVO					Meme Non promoters	Dreme Non promoters	KNOWN			
	AME all	TRAP all	AME Non promoters	TRAP Non promoters	AME all			AME all	TRAP all	AME Non promoters	TRAP Non promoters
Sexton et al.											
M1BP	6.36E-29	1.49E-22						9.11E-36	2.67E-27		
Beaf-32	1.72E-21	1.87E-23						6.02E-22	4.07E-22		
Motif-6	1.39E-11	5.73E-17						6.50E-16	9.86E-18		
ZIPIC	NA	NA						6.71E-15	1.87E-10		
Motif-8	NA	NA						3.34E-04	2.9E-5		
CTCF	NA	NA	NA		4.00E+10		NA	0.96	5.04E-09	1.4E-4	
Su(Hw)	NA	NA	2.91E-01	0.0313	6.20E+14		NA	0.082	2.49E-02	0.0381	
Ibf1/2	NA	NA	NA				2.22E-04	1.3E-3	NA		6.6E-4
Cubeñas-Potts et al.											
M1BP	1.17E-50	5.90E-34						8.65E-53	6.38E-40		
Beaf-32	1.25E-52	1.04E-50						1.45E-46	1.16E-48		
Motif-6	9.64E-38	1.52E-45						2.23E-38	3.91E-43		
ZIPIC	NA	NA						8.10E-27	1.07E-19		
Motif-8	NA	NA						1.47E-10	9.12E-11		
CTCF	NA	NA	8.10E-11	0.00129	9.17E-06		NA	0.117	7.53E-06	0.0633	
Su(Hw)	NA	NA	NA	6.79E-06			3.69E-02	3.5E-4	1.52E-06	0.03481	
Ibf1/2	NA	NA	NA				1.08E-11	2.75E-09	9.03E-02	6.53E-05	
This study											
M1BP	4.24E-72	9.96E-72						7.84E-85	5.20E-87		
Beaf-32	2.10E-64	1.27E-74						1.63E-64	3.27E-70		
Motif-6	4.83E-45	5.26E-64						1.47E-48	8.65E-66		
ZIPIC	NA	NA						5.09E-37	1.07E-30		
Motif-8	1.58E-04	0.721						8.86E-08	7.27E-16		
CTCF	NA	NA	1.83E-04	0.0511	1.10E-02	1.10E-06	NA	0.095184	7.53E-06	0.0636	
Su(Hw)	NA	0.0987	2.09E-03	3.89E-23	NA		2.10E-04	3.78E-03	4.30E-07	1.52E-06	3.87E-10
Ibf	NA	NA	NA	0.00137	NA		3.16E-07	6.22E-10	9.03E-02	2.03E-05	

Supplementary Table 3. Hi-C data sources

Source	Restriction enzyme	No. of usable reads	GEO accession	Reference
Whole embryos	DpnII	133,483,965	GSE34453	²
Kc167	DpnII	135,274,348	GSE63515	¹⁶
Kc167	DpnII	110,807,526	GSE80701	³
Kc167	HindIII	71,278,991	GSE38468	¹⁷
S2	HindIII	680,121,887	GSE58821	¹⁸
Clone-8	HindIII	131,426,003	GSE58821	¹⁸
third instar larvae salivary glands	HindIII	9,404,794	GSE72512	¹⁹

Supplementary Table 4. ChIP-seq data sources.

Source	GEO accession	Reference
Kc167 Beaf-32	GSM762845	4
Kc167 CP190	GSM762836	4
Kc167 CTCF	GSM1535983	16
Kc167 Su(Hw)	GSM762839	4
Kc167 Cap-H2	GSM1318356	20
Kc167 Chromator	GSM1318357	20
Kc167 Rad21	GSM1318352	20
Kc167 Pita	GSM2133768	3
Kc167 ZIPIC	GSM2133769	3
Kc167 GAF	GSM2133762	3
Kc167 lbf 1	GSM2133766	3
Kc167 lbf 2	GSM2133767	3
Embryo Zw5	GSM2042227	21
S2 M1BP	GSM1208162	22
Kc167 RNA Pol-II	GSM1536014	16
S2 DNase-seq	GSM1000406	23

Supplementary Table 5. Number of sequencing reads for Hi-C control and knockdowns.

	GST rep. A	GST rep. B	Beaf-32 KD rep. A	Beaf-32 KD rep. B	M1BP KD rep. A	M1BP KD rep. B	double KD rep. A	double KD rep. B
Pairs considered	107.8M	118.5M	102.4M	151.0M	298.1M	213.8M	217.3M	180.5M
Pairs used	32.1M	32.5M	14.5M	22.0M	58.7M	40.8M	42.9M	36.6M

References

1. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: Visualization of Intersecting Sets. *IEEE Trans. Vis. Comput. Graph.* **20**, 1983–1992 (2014).
2. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458–472 (2012).
3. Cubeñas-Potts, C. *et al.* Different enhancer classes in Drosophila bind distinct architectural proteins and mediate unique chromatin interactions and 3D architecture. *Nucleic Acids Res.* **45**, 1714–1730 (2017).
4. Wood, A. M. *et al.* Regulation of Chromatin Organization and Inducible Gene Expression by a Drosophila Insulator. *Mol. Cell* **44**, 29–38 (2011).
5. Li, L. *et al.* Widespread rearrangement of 3D chromatin organization underlies polycomb-mediated stress-induced silencing. *Mol. Cell* **58**, 216–231 (2015).
6. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–5 (2016).
7. Hug, C. B., Grimaldi, A. G., Kruse, K. & Vaquerizas, J. M. Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription. *Cell* **169**, 216–228.e19 (2017).
8. Iwasaki, O. *et al.* Interaction between TBP and Condensin Drives the Organization and Faithful Segregation of Mitotic Chromosomes. *Mol. Cell* **59**, 755–767 (2015).
9. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. U. S. A.* **112**,

- E6456–65 (2015).
10. Goloborodko, A., Imakaev, M. V., Marko, J. F. & Mirny, L. Compaction and segregation of sister chromatids via active loop extrusion. *Elife* **5**, (2016).
 11. Cook, P. R. A model for all genomes: the role of transcription factories. *J. Mol. Biol.* **395**, 1–10 (2010).
 12. Sutherland, H. & Bickmore, W. A. Transcription factories: gene expression in unions? *Nat. Rev. Genet.* **10**, 457–466 (2009).
 13. Stevens, T. J. *et al.* 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544**, 59–64 (2017).
 14. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28–36 (1994).
 15. Bailey, T. L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27**, 1653–1659 (2011).
 16. Li, L. *et al.* Widespread Rearrangement of 3D Chromatin Organization Underlies Polycomb-Mediated Stress-Induced Silencing. *Mol. Cell* **58**, 216–231 (2015).
 17. Hou, C., Li, L., Qin, Z. S. & Corces, V. G. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Mol. Cell* **48**, 471–484 (2012).
 18. Ramírez, F. *et al.* High-Affinity Sites Form an Interaction Network to Facilitate Spreading of the MSL Complex across the X Chromosome in Drosophila. *Mol. Cell* **60**, 146–162 (2015).
 19. Eagen, K. P., Hartl, T. A. & Kornberg, R. D. Stable Chromosome Condensation Revealed by Chromosome Conformation Capture. *Cell* **163**, 934–946 (2015).
 20. Van Bortle, K. *et al.* Insulator function and topological domain border strength scale with architectural protein occupancy. *Genome Biol.* **15**, R82 (2014).

21. Zolotarev, N. *et al.* Architectural proteins Pita, Zw5, and ZIPIC contain homodimerization domain and support specific long-range interactions in Drosophila. *Nucleic Acids Res.* **44**, 7228–7241 (2016).
22. Li, J. & Gilmour, D. S. Distinct mechanisms of transcriptional pausing orchestrated by GAGA factor and M1BP, a novel transcription factor. *EMBO J.* **32**, 1829–1841 (2013).
23. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).

A.2 Galaxy HiCExplorer

I contributed to the development of HiCExplorer and designed the template workflow for the galaxy web server. I contributed to the writing and revision of the manuscript along with Joachim Wolff and other authors.

Galaxy HiCExplorer: a web server for reproducible Hi-C data analysis, quality control and visualization

Joachim Wolff¹, Vivek Bhardwaj^{2,6}, Stephan Nothjunge^{5,8}, Gautier Richard^{2,7}, Gina Renschler^{2,6}, Ralf Gilsbach⁵, Thomas Manke², Rolf Backofen^{1,3,4}, Fidel Ramírez^{2,*} and Björn A. Grüning^{1,3,*}

¹Bioinformatics Group, Department of Computer Science, University of Freiburg, Georges-Köhler-Allee 106, 79110 Freiburg, Germany, ²Max Planck Institute of Immunobiology and Epigenetics, Stübeweg 51, 79108 Freiburg im Breisgau, ³Center for Biological Systems Analysis (ZBSA), University of Freiburg, Habsburgerstr. 49, 79104 Freiburg, Germany, ⁴BIOSS Centre for Biological Signaling Studies, University of Freiburg, Schänzlestr. 18, 79104 Freiburg, Germany, ⁵Institute of Experimental and Clinical Pharmacology and Toxicology, Faculty of Medicine, University of Freiburg, Albertstr. 25, 79104 Freiburg, Germany, ⁶Faculty of Biology, University of Freiburg, Schänzlestr. 1, 79104 Freiburg, Germany, ⁷IGEPP, INRA, Agrocampus Ouest, Univ Rennes, 35600 Le Rheu, France and ⁸Hermann Staudinger Graduate School, University of Freiburg, Hebelstrasse 27, 79104 Freiburg, Germany

Received February 23, 2018; Revised April 27, 2018; Editorial Decision May 10, 2018; Accepted May 22, 2018

ABSTRACT

Galaxy HiCExplorer is a web server that facilitates the study of the 3D conformation of chromatin by allowing Hi-C data processing, analysis and visualization. With the Galaxy HiCExplorer web server, users with little bioinformatic background can perform every step of the analysis in one workflow: mapping of the raw sequence data, creation of Hi-C contact matrices, quality assessment, correction of contact matrices and identification of topological associated domains (TADs) and A/B compartments. Users can create publication ready plots of the contact matrix, A/B compartments, and TADs on a selected genomic locus, along with additional information like gene tracks or ChIP-seq signals. Galaxy HiCExplorer is freely usable at: <https://hicexplorer.usegalaxy.eu> and is available as a Docker container: <https://github.com/deeptools/docker-galaxy-hicexplorer>.

INTRODUCTION

Chromosome conformation capture techniques are now widely used to analyse the 3D conformation of chromatin inside the nucleus across a rising number of species, tissues and experimental conditions. In particular, the Hi-C protocol (1) has helped to uncover folding principles of chromatin, demonstrating that the genome is partitioned into active and inactive compartments (called A and B) (1) and that these compartments are further subdivided into topological associated domains (TADs) (2,3). Furthermore, Hi-

C has allowed identification of chromatin loops (4,5), as well as enhancer–promoter interactions (6,7) and their influence on gene expression (8,9).

However, Hi-C data processing requires tabulating hundreds of millions to billions of paired-end reads into large matrices. This poses bioinformatic challenges for efficient processing of the data and subsequent analyses. Here, we introduce Galaxy HiCExplorer, a package that aims to make Hi-C data processing, analysis and visualization available to non-bioinformaticians. Our goal is to provide a software environment able to automate the whole workflow of Hi-C data analyses from raw read mapping, filtering and correction, to the computation of topological associated domains and A/B compartments, and finally to the visualization of contact matrices, along with various other genomic features and omics data. Moreover, Galaxy HiCExplorer is easy to install, maintainable, stable and well documented. The availability of a docker container in conjunction with Bioconda (<http://dx.doi.org/10.1101/207092>), eliminates the need for complex software and dependency installations. Finally, HiCExplorer is transparently developed by a community of collaborators based on best practices (10) for version control, code revisions, manual and automated testing and comprehensive documentation.

COMPREHENSIVE SERVER FOR HI-C ANALYSES

Galaxy HiCExplorer is freely available at <https://hicexplorer.usegalaxy.eu> as well as a Docker container: <https://github.com/deeptools/docker-galaxy-hicexplorer>. Galaxy HiCExplorer was designed to provide an easily accessible data-analysis environment such that

*To whom correspondence should be addressed. Tel: +49 761 2037460; Fax: +49 761 2037462; Email: gruening@informatik.uni-freiburg.de
Correspondence may also be addressed to Fidel Ramírez. Email: ramirez@ie-freiburg.mpg.de

W12 *Nucleic Acids Research*, 2018, Vol. 46, Web Server issue

biomedical researchers can focus on critical research aspects instead of dealing with terminal-based applications that are not user-friendly. It smoothly integrates the HiCExplorer analysis toolset (8) into the Galaxy scientific analysis platform to provide web-based, easy-to-use and thoroughly tested workflows that provide pipelines for the most common Hi-C data processing steps.

In contrast to other available Hi-C analysis software like HiCUP (14), HOMER (15) and TADbit (16) among others (see (17,18) for a comprehensive list of tools), Galaxy HiC-Explorer provides a fully comprehensive analysis pipeline available to much broader community of researchers and is not restricted to a subset of important features. HiC-Pro (19) is one of the few packages that offers a complete pipeline; however, its visualization tools are limited and it is only available as a command line tool. Similarly, Juicer (20) offers a command line tool processing pipeline while Juicebox (21) only provides visualizations. Moreover, the integration of HiCExplorer into Galaxy offers the possibility to process and integrate other data types like ChIP-Seq or RNA-Seq into the analysis using the same interface. None of the aforementioned tools offer web server access except HiFive (22).

A strong advantage of HiCExplorer is that it can take multiple matrix data formats developed by different research groups as input. Thus, it is well integrated in the landscape of Hi-C data analysis algorithms, as Hi-C matrices can be produced by other tools and visualized with HiC-Explorer. Conversely, matrices can be created with HiC-Explorer and then exported to be used by other software. Currently, the Galaxy HiCExplorer supports two major formats: The HiCExplorer specific h5 format and to promote standardization of Hi-C contact matrices the cooler format (23) developed within the 4D nucleome project (24).

GALAXY HiCExplorer TOOLS AND WORKFLOWS

Galaxy HiCExplorer provides a plethora of tools for processing, normalization, analysis, and visualization of Hi-C data (Figure 1A). Apart from HiCExplorer, the <https://hicexplorer.usegalaxy.eu> website and the Docker container also include the genome alignment tools BWA-MEM (25) and Bowtie2 (26), as well as additional tools for text manipulation, data import and quality control. The inclusion of deepTools (27) further facilitates the integration of ChIP-seq, RNA-seq, MNase-seq as well as other kind of datasets with Hi-C data.

The analysis of Hi-C data can be divided into three steps: pre-processing (including quality control), analysis and visualization.

Pre-processing and quality control

hicBuildMatrix. A contact matrix is the main data structure of Hi-C data analysis which is generated from the individual alignment of valid Hi-C paired-end reads. This tool filters out potentially erroneous reads, such as unmappable reads, self-ligated reads, dangling-ends, PCR duplicates or incomplete digestions (4,14) and tabulates the results based on user defined bins (either based on restriction sites or on fixed size bins). Because building the Hi-C matrix is one of

the most time consuming steps in the Hi-C workflow, we developed *hicBuildMatrix* to be multi-processing to significantly reduce running time. A comprehensive quality report is generated as an HTML file. This report includes a number of useful quality measures including: number of valid Hi-C read pairs and the number of filtered reads per category (unmappable and non-unique pairs, duplicates, dangling ends, self-circles, etc.), number of intra-chromosomal, short-range (<20 kb) and long-range contacts, and read pair orientation. Reports from multiple samples can be integrated using MultiQC (28) or using the HiCExplorer tool *hicQC*. Inspection of the *hicBuildMatrix* quality reports helps to identify potential biases or errors in the Hi-C library preparation. For example, a high number of dangling ends is indicative of a problem with the re-ligation step or inefficient removal of dangling ends. The quality report can also be useful to identify differences (long-range versus short-range contacts enrichment for instance) between samples obtained in different conditions.

hicMergeMatrixBins. After a Hi-C contact matrix has been created, lower resolution matrices can be obtained by merging neighboring bins. This is mostly useful for visualization at different zoom levels or to create matrices of lower resolution (larger bin size) in the event of a Hi-C matrix being too poor due to low sequencing depth.

hicCorrelate. This tool computes the correlation between several Hi-C matrices (Figure 1B). *hicCorrelate* can produce a scatter plot or a heatmap using either Pearson or Spearman correlations. The computation of the correlation can be restricted to a range of genomic distances to avoid biasing the correlation results with background contacts. These correlations are useful as a quality control step to compare replicates and to test for differences between various treatments.

hicPlotDistVsCounts. This tool plots the average number of Hi-C contacts at different genomic distances (Figure 1C). It allows the estimation of long-range and short-range contacts from multiple samples at once, and is a useful tool for both quality control and comparison of, for example, treated versus untreated samples that alter chromosome conformation.

hicSumMatrices. After different replicates or similarly obtained Hi-C matrices have been compared using *hicCorrelate*, they can be added up into one single contact matrix with this tool.

hicCorrectMatrix. Allows the removal of biases from the Hi-C matrix using a very fast version of the iterative correction algorithm from Imakaev *et al.* (29). Before the contact matrix is corrected, the right thresholds to prune values need to be selected. The *diagnostic plot* helps users in determining these thresholds.

Analysis

hicFindTADs. This utility can identify TADs from a given corrected contact matrix by first computing a TAD-

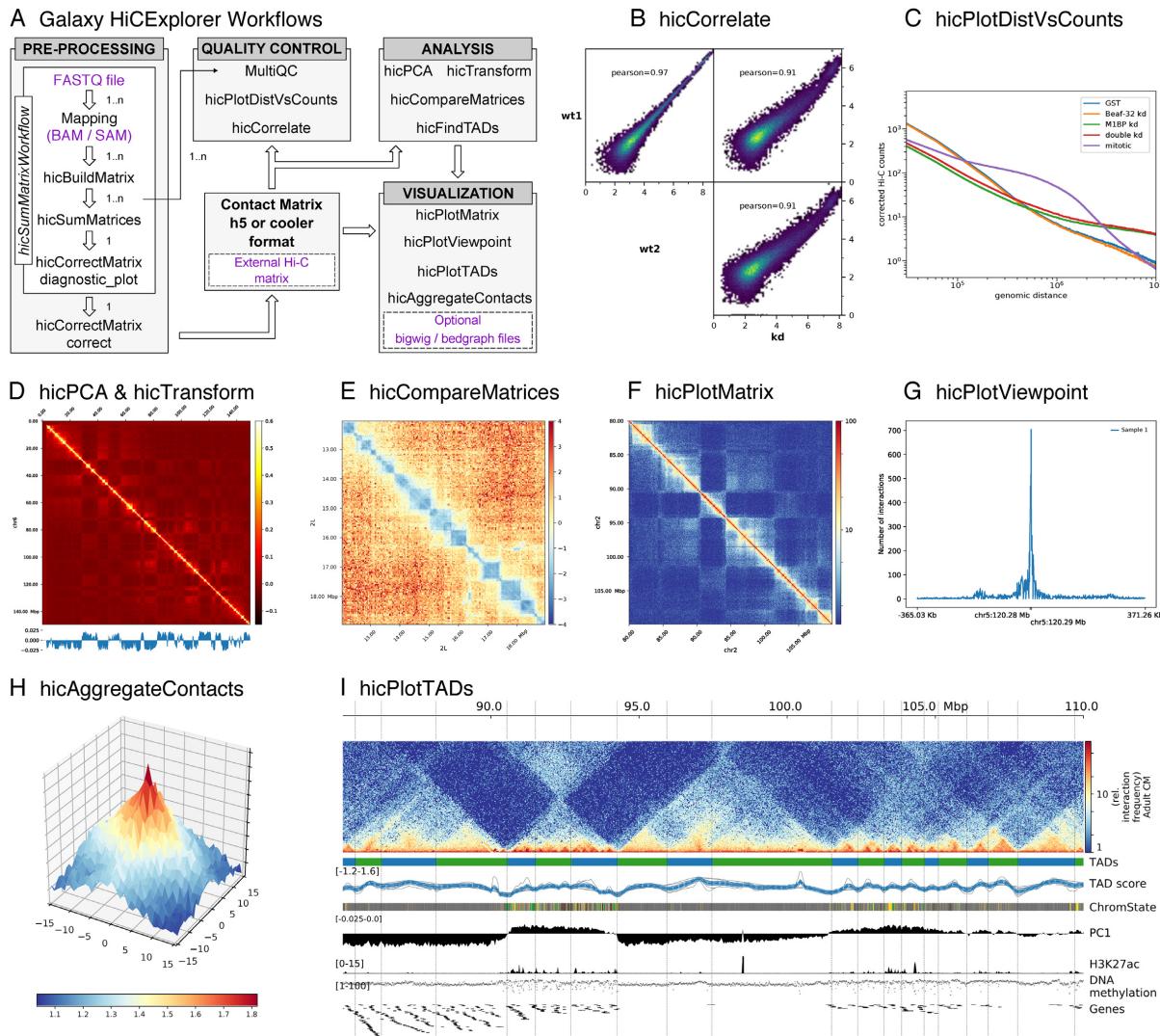


Figure 1. (A) Galaxy HiCExplorer workflows and tools. Entry points for external data are highlighted in purple. **Quality control tools:** (B) Output of *hicCorrelate* comparing two wild types and one knockdown samples. (C) Output of *hicPlotDistVsCounts* that shows changes of the number of contacts for different conditions. **Analysis tools:** (D) *hicPlotMatrix* of the Pearson correlation matrix derived from a contact matrix for chromosome 6 in mouse computed with *hicTransform*. The optional data track at the bottom shows the first eigenvector for A/B compartment obtained using *hicPCA*. (E) The pixel difference between a Hi-C corrected matrix for wild type condition and a knock down was computed using *hicCompareMatrices* and a 7Mb region is visualized using *hicPlotMatrix*. **Visualization tools:** (F) Contact matrix plot of a 80 to 105 Mb region of chromosome 2 in log scale. (G) Example output of *hicPlotViewpoint* showing the corrected number of Hi-C contacts for a single bin in chromosome 5 (output similar to 4C-seq) (11). (H) A Hi-C matrix was converted into an observed vs. expected matrix using *hicTransform* and this matrix, together with the location of high-affinity sites from (12) were used to run *hicAggregateContacts*. (I) 85 Mb to 110 Mb region from human chromosome 2 visualized using *hicPlotTADs*. TADs were computed by *hicFindTADs*. The additional tracks added correspond to: TAD- separation score (as reported by *hicFindTADs*), chromatin state, principal component 1 (A/B compartment) computed using *hicPCA*, ChIP-seq coverage for the H3K27ac mark, DNA methylation, and a gene track. Hi-C data for B, C, E and H from *Drosophila melanogaster* S2 cells from (8). Hi-C data for D, F and I from mouse cardiac myocytes (13). Additional tracks in I from (13).

separation score and then identifying local minima indicative of TAD boundaries (8). In contrast to other TAD identification methods, this tool also returns the TAD-separation score, which can be visualized in a genome browser or using *hicPlotTADs*. The TAD-separation score contains useful information to identify strong and weak boundaries and the density of contacts within TAD and can

be visualized along with the Hi-C matrix (see *hicPlotTADs* tool).

hicPCA. A/B compartments (1) refer to open and closed chromatin that is spatially separated in the cell nucleus (30,31). We compute this using eigenvector decomposition as described by Lieberman-Aiden (1) and using the first

W14 *Nucleic Acids Research*, 2018, Vol. 46, Web Server issue

and second eigenvector. The positive/negative values correspond to open/closed chromatin. A visualization of A/B compartments is shown in Figure 1D.

hicTransform. The three matrices used to compute the A/B compartments (observed/expected, Pearson correlation and covariance matrices) are useful during visualization to achieve a better understanding of the Hi-C data. To enable this, *hicTransform* can compute these three matrices independently of *hicPCA*, and the matrices can then be plotted using the visualization tools.

hicCompareMatrices. *hicCompareMatrices* allows the computation of difference, ratio or log2ratio between two matrices. This is useful to compare replicates or samples from different conditions. It can, for example, help to characterize TAD structure modifications when followed by *hicPlotMatrix* (Figure 1E).

Visualization

hicPlotMatrix. This tool is used to plot contact matrices for a collection of individual chromosomes. It has multiple options to select the matrix colors and the values range. Additionally, bigwig tracks can be attached to plot additional features such as A/B compartments or ChIP-seq data. It is possible to plot a multitude of domains; the entire interaction matrix, individual chromosomes, multiple chromosomes, and various regions of interest (see Figure 1D–F).

hicPlotViewpoint. The viewpoint plot supports a visualization of the number of interactions around a specific reference point or region in the genome, and makes the long-range interactions visible as shown in Figure 1G. The output is comparable to what is obtained using the 4C-seq protocol.

hicAggregateContacts. Facilitates the analysis of long range-contacts by visualizing the average contacts over multiple smaller matrices around a given set of regions (Figure 1H).

hicPlotTADs. To visualize the computed TADs this tool flips the main diagonal of the Hi-C contact matrix by 45° and marks the TADs with triangles. It is possible to plot multiple matrices and add additional data like genes, chromatin states, long-range interactions and any other feature that can be represented as a bigwig or bedgraph file like methylation data, ChIP-seq, or RNA-seq to visually correlate them with TADs and their boundaries. There are multiple options to select the Hi-C matrix layout and colormap, different ways to visualize genes and regions files and also multiple configurations to plot coverage tracks like color, line width, line type, as dots, filled etc. (Figure 1I).

Workflows

Galaxy HiCExplorer provides pre-defined workflows to reduce intermediate steps and to guide a researcher through the different stages. The Galaxy framework offers the possibility to connect tools into workflows called Galaxy workflows. The provided workflows are subdivided into categories depending on the start of the analysis: First, raw

FASTQ files are mapped to generate a contact matrix and its corrected equivalent. Different workflows are provided to cover the case of running many analyses in parallel or whether replicates should be merged to one contact matrix. Second, said contact matrix (or other) is used to compute TADs, A/B compartments and/or to plot them using the provided workflows. All workflows are linked on the homepage of the Galaxy HiCExplorer.

All Galaxy Workflows share a common notion that they should guide the researcher through the analysis, i.e. most parameters in the workflows do not need to be changed. The reference genome needs to be set for the mappers, and a desired bin size as well as the used restriction sites needs to be selected in order to build the contact matrix. Every workflow containing a plotting step needs the region to plot as input.

IMPLEMENTATION

Galaxy HiCExplorer is implemented as a Docker container based on the web-based Galaxy scientific workflow platform (32). HiCExplorer itself is implemented in Python, supporting version 2.7, 3.5 and 3.6, and available as a Bioconda package (<http://dx.doi.org/10.1101/207092>) and as BioContainer (33). This guarantees a fixation of versions and therefore reproducibility of analysis. Galaxy wrappers for HiCExplorer are available at the Galaxy tool shed.

USING HiCExplorer

Installation and usage

The Galaxy HiCExplorer web server can be used by visiting <http://hicexplorer.usegalaxy.eu>, or by installing it on a personal computer or locally (e.g. an institute intranet). For this, pre-configured Docker containers and conda packages are available.

Galaxy HiCExplorer:

Docker :

```
docker run -p 8080:80 quay.io/bgruening/galaxy-hicexplorer
```

hicexplorer.usegalaxy.eu : On <https://hicexplorer.usegalaxy.eu> all HiCExplorer tools and workflows are installed. Use this option if you require high computational resources (e.g. large memory requirements).

HiCExplorer:

The HiCExplorer as a command line tool is available via *conda* or *BioContainers*.

Conda :

BioContainer :

```
docker run quay.io/biocontainers/hicexplorer:latest
```

Training

Training and a documentation are crucial to enable as many scientists as possible to use and understand the Galaxy HiCExplorer. To introduce scientists who are new to Galaxy a guided tour through the Galaxy interface is provided as well as a tour to learn Hi-C data analysis. The tour content is available on the Galaxy Training Network (<http://dx.doi.org/10.1101/225680>) as well and includes example

data hosted on Zenodo. All intermediate files are available in the shared data library of the Galaxy HiCExplorer.

For advanced users a detailed step-by-step tutorial for the analysis of Hi-C data from mouse embryonic stem-cells, as well as a comprehensive API documentation, is hosted at <https://hicexplorer.readthedocs.org>. The how-to describes how to set up the mapping of the reads. It suggests parameter settings for the creation of Hi-C contact matrices and describes the process of merging and threshold determination to remove poor bins prior to correction. The determination of TADs using the separation score is described in detail, including examples on visualization.

DISCUSSION

Galaxy HiCExplorer gives researchers the opportunity to run their Hi-C data analysis in a user-friendly, web browser based environment. The highly configurable framework provided by Galaxy makes this web server extendable to the various needs of researchers. Especially in conjunction with software for other high-throughput analysis protocols like RNA-seq or ChIP-seq, Galaxy HiCExplorer serves as a powerful basis for flexible explorative biomedical research in a high-throughput sequencing data analysis environment.

By combining all the necessary stages of pre-processing and visualization into a single tool, analysis not only becomes easier, but faster, highly reproducible, and more readily exchangeable. Biomedical researchers can focus their efforts on their data analysis without having to concern themselves with the particulars of managing various different software setups and configurations or learning to use command-line tools in an UNIX environment.

ACKNOWLEDGEMENTS

We thank the bioinformatics group at the University of Freiburg and the bioinformatics unit at the Max Planck Institute of Immunobiology and Epigenetics Freiburg.

FUNDING

German Research Foundation for the Collaborative Research Centre 992 Medical Epigenetics [SFB 992/1 2012 and SFB 992/2 2016 awarded to T.M. and R.B. and to Lutz Hein (in support of S.N. and R.G.) and for the DFG project GI 747/2-1 to R.G]; Federal Ministry of Education and Research through the German Epigenome Programme DEEP [01KU1216G awarded to T.M.]; German Federal Ministry of Education and Research [031 A538A de.NBI-RBC awarded to R.B.]; German Federal Ministry of Education and Research [031 L0101C de.NBI-epi awarded to B.G.]. Funding for open access charge: German Federal Ministry of Education and Research.

Conflict of interest statement. None declared.

REFERENCES

- Lieberman-Aiden,E., Van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
- Dixon,J.R., Selvaraj,S., Yue,F., Kim,A., Li,Y., Shen,Y., Hu,M., Liu,J.S. and Ren,B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.
- Nora,E.P., Lajoie,B.R., Schulz,E.G., Giorgetti,L., Okamoto,I., Servant,N., Piolot,T., Van Berkum,N.L., Meisig,J., Sedat,J. *et al.* (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**, 381–385.
- Rao,S.S.P., Huntley,M.H., Durand,N.C. and Stamenova,E.K. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
- Sanborn,A.L., Rao,S.S.P., Huang,S.-C., Durand,N.C., Huntley,M.H., Jewett,A.I., Bochkov,I.D., Chinnappan,D., Cutkosky,A., Li,J. *et al.* (2015) Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E6456–E6465.
- Bonev,B., Mendelson Cohen,N., Szabo,Q., Fritsch,L., Papadopoulos,G.L., Lubling,Y., Xu,X., Lv,X., Hugnot,J.P., Tanay,A. *et al.* (2017) Multiscale 3D genome rewiring during mouse neural development. *Cell*, **171**, 557–572.
- Ron,G., Globerson,Y., Moran,D. and Kaplan,T. (2017) Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains. *Nat. Commun.*, **8**, 2237.
- Ramírez,F., Bhardwaj,V., Arrigoni,L., Lam,K.C., Grüning,B.A., Villalveces,J., Habermann,B., Akhtar,A. and Manke,T. (2018) High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.*, **9**, 189.
- Babaei,S., Mahfouz,A., Hulsman,M., Lelieveldt,B.P., de Ridder,J. and Reinders,M. (2015) Hi-C chromatin interaction networks predict Co-expression in the mouse cortex. *PLoS Comput. Biol.*, **11**, e1004221.
- Jiménez,R.C., Kuzak,M., Alhamdoosh,M., Barker,M., Batut,B., Borg,M., Capella-Gutierrez,S., Chue Hong,N., Cook,M., Corpas,M. *et al.* (2017) Four simple recommendations to encourage best practices in research software. *F1000Research*, **6**, 876.
- Andrey,G., Schöpflin,R., Jerković,I., Heinrich,V., Ibrahim,D.M., Paliou,C., Hochradel,M., Timmermann,B., Haas,S., Vingron,M. *et al.* (2017) Characterization of hundreds of regulatory landscapes in developing limbs reveals two regimes of chromatin folding. *Genome Res.*, **27**, 223–233.
- Ramírez,F., Lingg,T., Toscano,S., Lam,K.C., Georgiev,P., Chung,H.R., Lajoie,B.R., de Wit,E., Zhan,Y., de Laat,W. *et al.* (2015) High-affinity sites form an interaction network to facilitate spreading of the MSL complex across the X chromosome in Drosophila. *Mol. Cell*, **60**, 146–162.
- Nothjunge,S., Nührenberg,T.G., Grüning,B.A., Doppler,S.A., Preissl,S., Schwaderer,M., Rommel,C., Krane,M., Hein,L. and Giltsbach,R. (2017) DNA methylation signatures follow preformed chromatin compartments in cardiac myocytes. *Nat. Commun.*, **8**, 1667.
- Wingett,S., Ewels,P., Furlan-Magaril,M., Nagano,T., Schoenfelder,S., Fraser,P. and Andrews,S. (2015) HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research*, **1310**, 1–12.
- Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
- Serra,F., Bäu,D., Goodstadt,M., Castillo,D., Filion,G. and Marti-Renom,M.A. (2017) Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLoS Comput. Biol.*, **13**, e1005665.
- Schmid,M.W., Grob,S. and Grossniklaus,U. (2015) HiCdat: A fast and easy-to-use Hi-C data analysis tool. *BMC Bioinformatics*, **16**, 277.
- Forcato,M., Nicoletti,C., Pal,K., Livi,C.M., Ferrari,F. and Bicciato,S. (2017) Comparison of computational methods for Hi-C data analysis. *Nat. Methods*, **14**, 679–685.
- Servant,N., Varoquaux,N., Lajoie,B.R., Viara,E., Chen,C.J., Vert,J.P., Heard,E., Dekker,J. and Barillot,E. (2015) HiC-Pro: An optimized and flexible pipeline for Hi-C data processing. *Genome Biology*, **16**, 259.

W16 Nucleic Acids Research, 2018, Vol. 46, Web Server issue

20. Durand,N.C., Shamim,M.S., Machol,I., Rao,S.S., Huntley,M.H., Lander,E.S. and Aiden,E.L. (2016) Juicer provides a One-Click system for analyzing Loop-Resolution Hi-C experiments. *Cell Syst.*, **3**, 95–98.
21. Durand,N.C., Robinson,J.T., Shamim,M.S., Machol,I., Mesirov,J.P., Lander,E.S. and Aiden,E.L. (2016) Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.*, **3**, 99–101.
22. Sauria,M.E., Phillips-Cremins,J.E., Corces,V.G. and Taylor,J. (2015) HiFive: A tool suite for easy and efficient HiC and 5C data analysis. *Genome Biol.*, **16**, 237.
23. Abdennur,N., Goloborodko,A., Imakaev,M. and Mirny,L. (2017) *mirnylab/cooler v0.7.6*. zenodo.org.
24. Dekker,J., Belmont,A.S., Guttman,M., Leshyk,V.O., Lis,J.T., Lomvardas,S., Mirny,L.A., O'Shea,C.C., Park,P.J., Ren,B. *et al.* (2017) The 4D nucleome project. *Nature*, **549**, 219–226.
25. Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arxiv.org*, [arXiv:1303.3997].
26. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
27. Ramírez,F., Ryan,D.P., Grüning,B., Bhardwaj,V., Kilpert,F., Richter,A.S., Heyne,S., Dündar,F. and Manke,T. (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.*, **44**, W160–W165.
28. Ewels,P., Magnusson,M., Lundin,S. and Käller,M. (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**, 3047–3048.
29. Imakaev,M., Fudenberg,G., McCord,R.P., Naumova,N., Goloborodko,A., Lajoie,B.R., Dekker,J. and Mirny,L.A. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, **9**, 999–1003.
30. Stevens,T.J., Lando,D., Basu,S., Atkinson,L.P., Cao,Y., Lee,S.F., Leeb,M., Wölfelhart,K.J., Boucher,W., O'Shaughnessy-Kirwan,A. *et al.* (2017) 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, **544**, 59–64.
31. Dixon,J.R., Jung,I., Selvaraj,S., Shen,Y., Antosiewicz-Bourget,J.E., Lee,A.Y., Ye,Z., Kim,A., Rajagopal,N., Xie,W. *et al.* (2015) Chromatin architecture reorganization during stem cell differentiation. *Nature*, **518**, 331–336.
32. Afgan,E., Baker,D., van den Beek,M., Blankenberg,D., Bouvier,D., Čech,M., Chilton,J., Clements,D., Coraor,N., Eberhard,C. *et al.* (2016) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.*, **44**, W3–W10.
33. da Veiga Leprevost,F., Grüning,B.A., Alves Afifos,S., Röst,H.L., Uszkoreit,J., Barsnes,H., Vaudel,M., Moreno,P., Gatto,L., Weber,J. *et al.* (2017) BioContainers: an open-source and community-driven framework for software standardization. *Bioinformatics*, **33**, 2580–2582.

A.3 Analysis of dosage compensation in flies via promoter-profiling

I contributed to the development of the MAPCap protocol by providing the analysis input, where all the experiments were performed by Giuseppe Semplicio. I developed the icetea bioconductor package and performed all the analysis presented in the paper, made the figures and wrote the manuscript with input from Giuseppe Semplicio.

Quantitative analysis of dosage compensation in flies using promoter profiling

Vivek Bhardwaj^{1,2*}, Giuseppe Semplicio^{1,2*}, Thomas Manke¹, Asifa Akhtar^{1#}

¹Max Planck Institute for Immunobiology and Epigenetics, 79108, Freiburg, Germany

²Faculty of Biology, University of Freiburg, 79104, Freiburg, Germany

*equal contribution

#corresponding author

akhtar@ie-freiburg.mpg.de

Phone: +49 (0)7615108565

Fax: +49 (0)7615108566

Abstract

Promoter architecture, shape and position of transcription start sites (TSS) play an important role in the regulation of eukaryotic gene expression. Promoter profiling methods like CAGE (Cap Analysis of Gene Expression) are widely used to detect transcription start sites and alternative promoter usage between tissues or during development. However, these methods are rarely used for differential expression analysis. In this study, we describe an approach to combine promoter profiling and differential expression analysis in a single setup, using a fast and simple protocol, MAPCap (Multiplexed Affinity Purification of Capped RNA) along with a new tool “icetea” (<https://bioconductor.org/packages/icetea>). icetea enables detection of TSS at high-resolution after UMI-based removal of PCR duplicates and detects differential gene expression and promoter usage using both internal and external normalization controls. Using MAPCap and icetea, we analyzed TSS expression in the brains of *Drosophila melanogaster* larvae, and observed the effects of knockout of MLE (male-less) helicase on X chromosome dosage compensation at promoters. Our results expand the scope of TSS profiling methods to differential expression analysis.

Introduction

Most genes in eukaryotes express multiple isoforms and transcript isoform expression is a major mechanism behind tissue-specific regulation of gene expression. Isoform diversity could be achieved by the usage of alternative exons, UTRs, transcript start and end sites. Recent analysis of Flies and Human genome has suggested that transcript start and end site selection is a major driver of alternative isoform usage across tissues [1]. Promoter profiling methods, such as CAGE [2], RAMPAGE [3], NanoCAGE [4] and GRO-cap [5] are widely used to identify transcription start sites (TSSs), therefore making them useful for the detection of alternative isoform usage across tissues or developmental stages. A recent comparative analysis of six

such methods identifies CAGE as the overall best method [6] while RAMPAGE [3] comes close second. The amount of time and number of steps required per library was found to be highest for CAGE. RAMPAGE reduces the processing time (to 2 days) and the required input material (up to 5 μ G) therefore providing a suitable alternative. Despite this improvement, the CAGE methods have not gained wide usage besides the detection of new TSS and analysis of promoter architecture.

In this study, we developed a short and easy to perform 5' profiling method, termed as MAPCap (Multiplexed Affinity Purification of Capped RNA), which allows multiplexing of samples and produces paired-end reads. Synthetically designed random barcodes allow removal of PCR duplicates, and external spike-in controls allow accurate relative quantification of TSS expression changes. Further we developed an R/Bioconductor package **icetea** (Integrating Cap Enrichment and Transcript Expression Analysis) (<https://bioconductor.org/packages/icetea>), which allows easy processing and analysis of data obtained from protocols such as MAPCap and RAMPAGE. We performed MAPCap on brains isolated from 3rd instar Fly larvae and quantified the defect of dosage compensation upon knockout of male-less (MLE) gene on X chromosome at transcription start site resolution.

Results

MAPCap enriches Capped RNA from multiplexed samples

A new type of adapter oligo developed previously by our group (called the s-oligo) has dramatically increased the speed to perform transcriptome-wide RNA-protein interaction assays [7]. One main reason is the suppression of so-called “adapter-dimers” which can frequently occur in ligation-mediated clonings of RNA, a widely used approach in RNA library preparation, including promoter-profiling methods. The successful application of the s-oligo in CLIP experiments prompted us to develop a similar approach for promoter-profiling. MAPCap is a method that combines the power of the s-oligo with the easy handling of bead-based affinity purifications (Fig. 1A, see methods). This allows for a fast and reproducible processing of even low RNA input amounts. Abundant RNA species such as sn- and snoRNAs are selectively

degraded by targeted antisense oligos and RNase H increasing the recovery of other capped RNA species. The s-oligo incorporates the sequences of both standard sequencing adapters which omits the usage of an RT-primer and allows for a highly efficient intramolecular ligation. The linear PCR amplification product creates a uniform library with ideal insert sizes of around 150 nt (Fig. S1A).

We performed MAPCap on stage 15 *Drosophila* embryos, in four replicates, and obtained ~10 Million reads per sample after de-multiplexing (Fig. S1B, see methods). For comparison, we analysed the CAGE data downloaded from modENCODE [8] and the RAMPAGE data [3] from embryos corresponding to the same stage. Reads obtained from both CAGE and RAMPAGE protocols show a high 'G' nucleotide bias due to the template-free activity of the reverse-transcriptase during cDNA preparation in RAMPAGE [9], and the design of the attached linkers in CAGE [10]. This demands post-mapping correction and sometimes affects the accuracy of TSS detection [11]. MAPCap, on the other hand, shows no such bias due to the use of s-oligo (Fig. 1B).

In order to evaluate the performance of MAPCap for TSS detection, we first performed TSS calling from MAPCap data using the paraclu method, and compared them to the TSS detected from CAGE and RAMPAGE data using the same method and filtering parameters (see methods). We then evaluated the TSS detection sensitivity, specificity, precision and F1-score between the methods, by comparing them to all annotated TSS present in the *Drosophila* ensembl annotation (release 76) as well as the RNA-Seq data obtained from modENCODE (see methods). CAGE performed the best amongst the three methods, as observed before [6], followed by RAMPAGE and MAPCap (Fig S1C).

We then correlated the depth-normalized counts on 5'-UTRs of known genes between MAPCap, CAGE, RAMPAGE and total RNA-Seq data from the same stage obtained from modENCODE project (see methods). We find that although MAPCap signal shows good correlation with other protocols, CAGE and RAMPAGE show better correlation with each other while MAPCap showed better correlation with RNA-Seq (Fig 1B-C, S1D). An independently performed MAPCap experiment with S2 cells shows the same relationship between the protocols, suggesting that MAPCap produces gene expression estimates closer to RNA-Seq compared to these protocols.

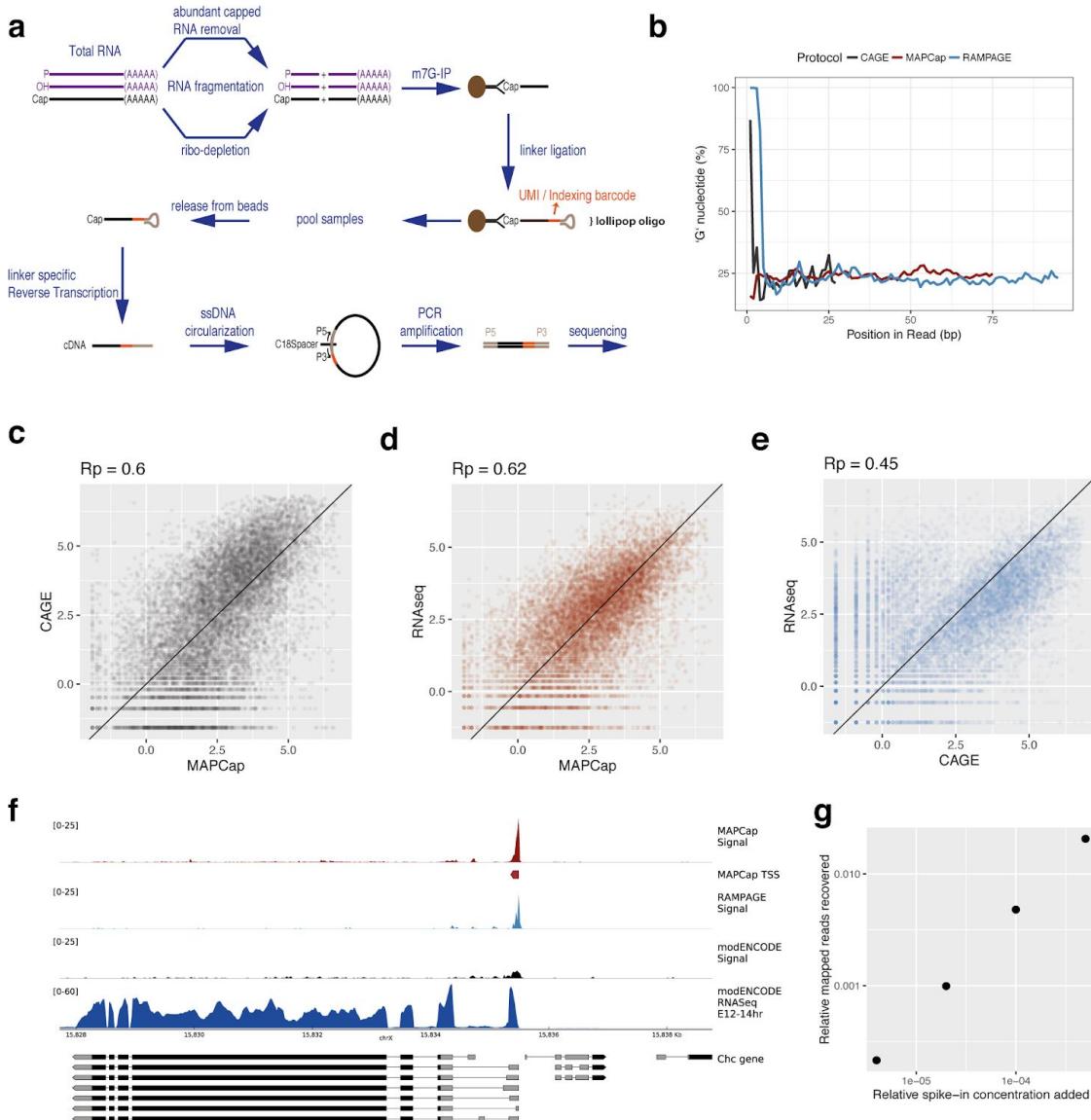
Figure 1

Fig 1. MAPCap protocol and its data quality. **A.** Overview of the MAPCap protocol. After fragmentation and ribo-depletion, the Capped RNA is immunoprecipitated using an antibody, and the s-oligos are attached, afterwards the samples are pooled for PCR and library preparation steps. **B.** Nucleotide content in read positions. CAGE and RAMPAGE show a high

artificial G-bias, due to their cloning steps, while MAPCap shows low bias for any specific nucleotide. **C-E.** *Correlation of MAPCap counts with CAGE, RNA-seq and the correlation of counts between CAGE and RNA-seq, on 3'UTR of genes. MAPCap shows better correlation with RNA-seq than CAGE.* **F.** *Genome track of the de-duplicated counts from MAPCap, RAMPAGE and CAGE on TSS. For MAPCap and RAMPAGE, the de-duplication was performed using 5'-position of the reads and the UMIs, while for CAGE it was performed using only 5'-position. RNA-seq track is shown for comparison.* **G.** *Added relative concentration (w.r.t. total RNA concentration) vs recovered relative counts (w.r.t total read counts) for the embryos. The samples were added with increasing relative concentration of ERCCs.*

The random barcodes present in lollipop oligos allow us to remove PCR duplicates, while preserving the transcript expression signal. Similar de-duplication can be performed for data obtained from the RAMPAGE protocol, where the oligos used as RT-PCR primers also serve as ‘pseudo-random’ barcodes [3]. A comparison of PCR duplicate removed signal from the three protocols show that both MAPCap and RAMPAGE protocols preserve the signal on the TSS, while de-duplication in absence of UMIs lead to near-complete loss of signal from the CAGE protocol (Fig. 1E).

Finally, we tested the sensitivity and relative quantification accuracy of MAPCap protocol for RNAs at different concentration by using external ERCC controls. We prepared spike-in mix containing 10 in-vitro capped ERCC spikes (see methods) in a 2-fold relative concentration ranging from 15.6 pM to 8 nM. We then mixed each replicate of the embryo sample with different concentrations of this spike-in mix (from 0.0004% to 0.05% of isolated RNA), before the beginning of the protocol. Processing of data shows that the relative concentration of spike-ins between sample can be faithfully recovered after sequencing (Fig. 1G). Relative ratio between individual spike-ins within each mix could also be accurately recovered (Fig. S1F), suggesting that MAPCap provides good sensitivity and accuracy to detect original transcript concentrations.

High resolution TSS detection and differential expression analysis using biological replicates

The ease of use and multiplexing ability of MAPCap protocol allows performing biological replicates without adding additional time and effort. We therefore sought to develop analysis methods which could benefit from biological replicates. Popularly used methods for TSS detection (parclu and distclu) are performed on within sample clustering of tags, and don't incorporate biological replicates to improve the performance of TSS detection. We developed a window-based TSS detection method that borrows ideas from window based peak calling methods developed for ChIP-Seq [12,13]. Reads counts are modelled using negative binomial distribution and the TSSs are detected as windows of enrichment in the genome, with respect to a local background (Fig. S2A, see methods). Consecutively enriched windows are then merged to detect both short and long TSSs. We compared peaks obtained from our method with those from paraclu method on the embryos to evaluate sensitivity and specificity. TSSs detected from the new method show higher sensitivity as well as specificity (Fig. 2a).

Apart from the accuracy of TSS detection, the shape of TSSs have also been shown to reflect biologically meaningful signal. Genes with sharp, focussed TSSs are mostly tissue specific and developmentally regulated, while the genes with long TSSs are shown to have housekeeping functions [14]. GO term enrichment analysis of the sharp and broad TSSs obtained from our method confirms previous results (see methods). Genes with sharp TSSs are enriched for processes like morphogenesis and development (Fig. 2B) while genes with broad TSSs are enriched for processes such as protein localization, metabolism and membrane organization (Fig. S2B), suggesting that the new method successfully detects biologically meaningful TSS shapes. Motif enrichment analysis of sharp and broad TSSs further confirm these results, with sharp TSSs being enriched for the Inr element, while broad TSS being enriched for core promoter motif M1BP and others (Fig. 2C, see methods).

Figure 2

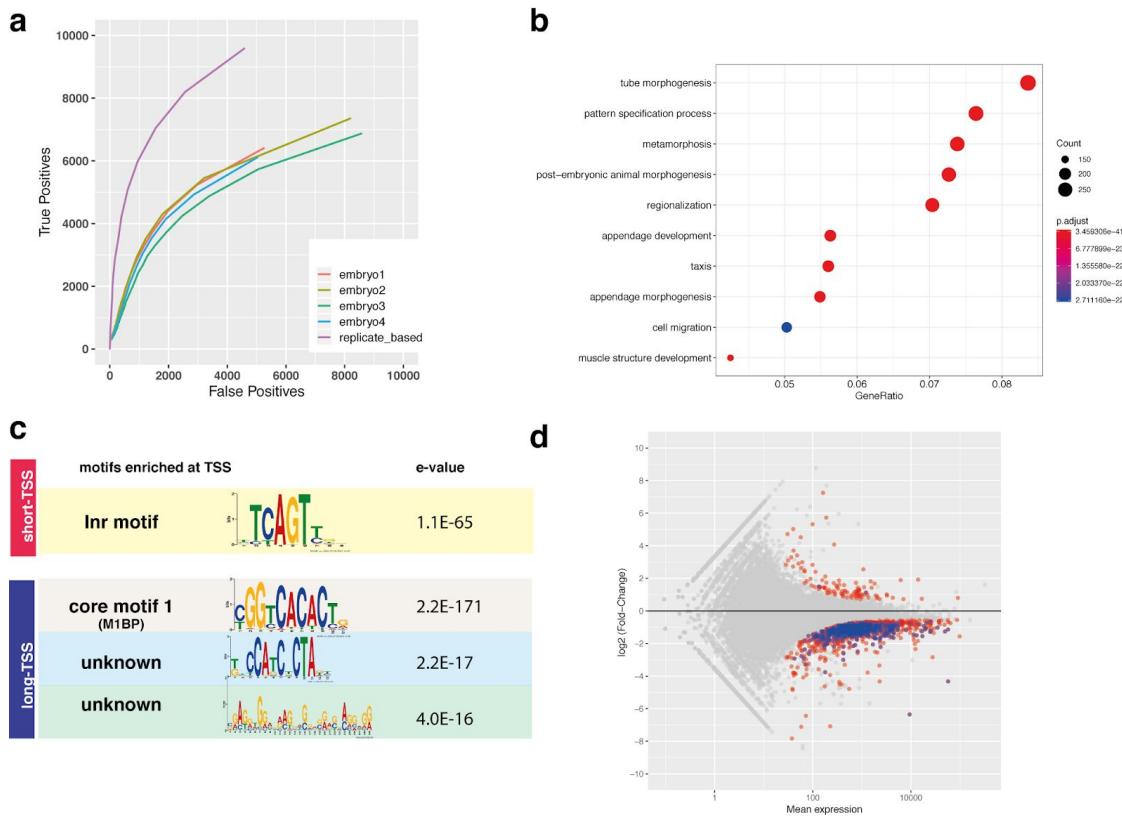


Fig 2. A replicate-based method of TSS detection. **A.** ROC curve of MAPCap data on embryos, samples labelled embryo1 to 4 were processed using *paraclu* method and compared with the new method (replicate based) that uses all 4 samples as replicated for TSS detection. The detection accuracy improves using the new method. **B.** GO enrichment of “sharp” TSSs (<20bp) detected by the new method. **C.** Motif enrichment of the “sharp” (<20bp) and “broad” (>100bp) TSS detected by the method. **D.** Gene-level differential expression estimates obtained from MAPCap data in brains of MLE KO males. Differentially expressed TSSs on X-chromosome are highlighted in blue.

Finally, we also sought to utilize the capped ERCC controls added to the protocol for the analysis of differential TSS usage between groups of samples. To this end, we performed MAPCap on RNA isolated from flies which are knock-outs of *maleless* (MLE) RNA helicase, and compared them with RNA from wild-type flies. The MLE helicase is an important component of the MSL complex which recognizes roX RNAs on the X-chromosome and helps guide the MSL complex to the X, leading to 2-fold upregulation of the male X-chromosome. In absence of MLE,

we therefore expected X chromosome to be downregulated due to failure of dosage compensation. We adopted popularly used method of gene-level differential expression analysis to the MAPCap data and utilized the spike-ins for normalization (see methods). Results show that most genes in MLE KO were downregulated, and most of the downregulated genes were on X chromosome (Fig. 2D). In absence of spike-in normalization, however, we saw more balanced number of up and down-regulated genes, where a bias towards the X-chromosome was not clearly visible (Fig. S2C), suggesting that spike-in normalization provides more useful biological insights.

Effect of MLE KO on dosage compensation of male promoters

After confirming the validity of our experimental and analysis method, we applied MAPCap on total RNA isolated from brains of male and female *Drosophila melanogaster* larvae of both wild-type and MLE KO background (see methods). We then deployed a pipeline that performs de-multiplexing, mapping, de-duplication, TSS detection and annotation of the TSS (Fig. S3A-B). We detected TSSs using our new method, using a fold-change threshold of 4x over the background, followed by comprehensive functional annotation of the detected TSS (see methods). While most of the detected TSS originated from previously annotated TSSs, X% of detected TSS in all samples came from promoter-proximal or intergenic enhancers (Fig 3A).

We then compared the MLE KO and wild-type genotypes in both male and female brains in order to detect differential promoter usage after spike-in normalization. Comparison of wild-type male and female brains showed that most promoters are equally utilized between sexes (Fig. S3C). Similar to our gene-level differential expression analysis, our differential promoter usage analysis revealed a significant downregulation of TSS from X-chromosome in KO males (Fig. 3B), while the females showed almost no effect in promoter usage (Fig. S3D).

Figure 3

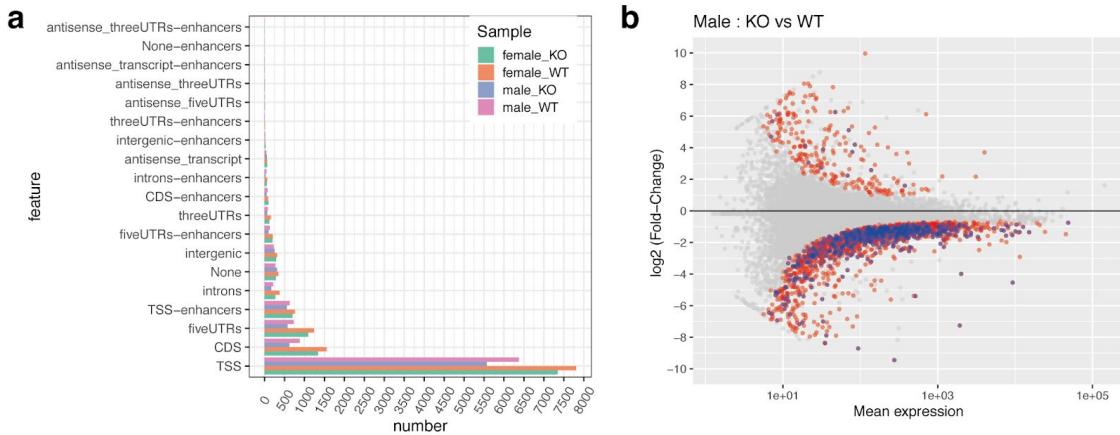


Fig 3. Assessment of promoter usage in the male and female fly brains. A. Annotation of detected TSS into different functional categories. Both wild-type and KO samples show similar enrichment of TSS in each category. **B.** Differential TSS (promoter usage) in male KO, compared to wild-type. Differentially expressed TSSs from X-chromosome are highlighted in blue.

icetea simplifies TSS detection and expression analysis from promoter profiling data

We implemented the processing and analysis methods described in this manuscript in an easy to use R package **icetea** (Integrating Cap Enrichment with Transcript Expression Analysis). **Icetea** performs sample de-multiplexing, PCR de-duplication as well as employs the new TSS detection approach described previously that takes advantage of biological replicates (Fig. 4A). Further functions for quality control (Fig. 4B-C) and quick annotation of detected TSS (Fig. 4D) are also implemented. Differential TSS expression analysis can be performed between group of samples, using either internal or external (spike-in) normalization, allowing accurate quantification of relative gene and isoform expression changes (Fig. 4E). **Icetea** is especially suitable for end-to-end analysis of paired-end 5' profiling techniques such as MAPCap and RAMPAGE, however it can easily be used for analysis of CAGE, GRO-Cap and other promoter profiling protocols. **icetea** is open source and available for use via Bioconductor (<https://bioconductor.org/packages/icetea>) and the source code is available on GitHub (<https://github.com/vivekbhr/icetea>).

Figure 4

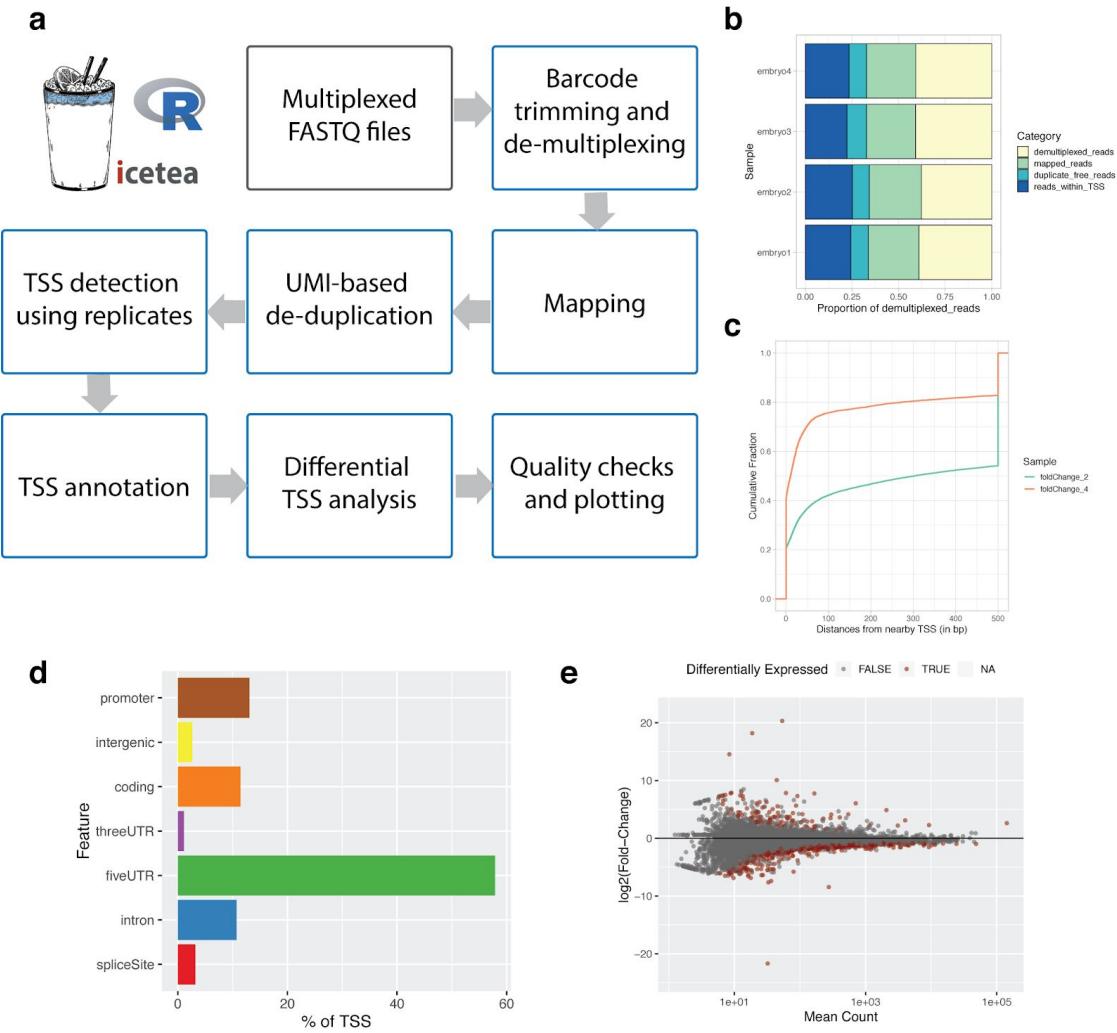


Fig 4. Description of the icetea bioconductor package for analysis of promoter-profiling data. **A.** Steps of data analysis implemented in icetea. Read de-multiplexing and de-duplication based on UMI is supported for MAPCap and RAMPAGE data. **B-E.** Some examples of results obtained from icetea. **B)** Read mapping statistics for embryo data (`plotReadStats`) **C)** TSS distance precision (distance of detected TSS to nearby annotated TSS) at different fold-change cutoffs on embryo data (`plotTSSprecision`) **D)** Annotation of detected TSS for male WT brains (`annotateTSS`). **E)** Differentially expressed TSS between male and female KO samples (`detectDiffTSS`).

Discussion

In this study we introduce an easy to perform promoter profiling technique, along with a new analysis approach which simplifies the integration of TSS discovery and transcript expression analysis. The use of MAPCap protocol along with icetea analysis provides most benefits of CAGE and RNA-seq, at a fraction of total cost and time of performing both the protocols. Unlike CAGE or RAMPAGE which utilize protocol-specific optimization, MAPCap utilizes protocol-agnostic s-oligos, which, apart from promoter profiling, could also be used for iCLIP experiments [7], as well as for RNA-Seq, allowing for wider scope of integrative analysis. We propose that this approach would prove optimal for pilot projects of transcript discovery and gene expression analysis of newly assembled as well as annotated genomes.

Methods

Cells

S2 cells (gift from the butros lab, Heidelberg) were cultured in Express Five SFM media (Thermo Fisher) supplemented with 10% (v/v) Glutamax (Thermo Fisher). Cultures were maintained adherent or in shaking incubators at 27°C at a speed of 80 rpm. Cells were kept at a density of 1-16 million/ml.

Generation of capped ERCC spikes

Ten spikes sequences were chosen from the ERCC spike mix, trimmed to ~500 bp length and ordered as gBlocks from IDT. At the 5' end we inserted a T7 class II promoter ϕ 2.5, which has been shown to create more homogenous 5' ends transcription promoter sequence [15]. Spikes were in vitro transcribed using T7-FLASHScribe Transcription kit (CellScript) according to manufacturer's instructions and purified using MegaClear kit. In vitro capping was performed with Vaccinia capping system. Potentially uncapped RNAs were degraded by treatment of spikes with polyphosphatase and terminator. Samples were cleaned using OCC, concentrations

measured on Qubit. A master mix was created where each subsequent spike was added at half the concentration of the previous spike, starting from 8 fmol/ul.

MAPCap library preparation

RNA from S2 cells was extracted using the Quick RNA Kit (Zymo Research). RNA from dissected brains of embryos was isolated using the DirectZol Kit (Zymo Research). RNA was eluted in 25 ul of RNase-free water. The concentrations are adjusted and capped ERCC spikes and HEK polyA RNA were added at 0.05% of input amount. To remove abundant capped RNAs (snRNA, snoRNAs) as well as rRNA contamination, we added antisense DNA oligos (see table) targeting the RNA species detected from a preliminary MAPCap run. 8ul of oligo mix were added together with 4ul of 10x terminator buffer A (Epicentre). The RNA was heated to 70 °C for 2min followed by an active cooling in the Thermomixer (Eppendorf) to 37 °C. Upon reaching this temperature, 1ul of RNaseH (Life/Invitrogen) was added and incubated at 37 °C for 30 min. The samples were then heated to 70 °C for 2 min, put immediately on ice for 1 min and 1 ul of Terminator exonuclease was added for 1 hr at 30 °C. RNA was purified using RNA clean and concentrator (Zymo Research) and eluted with 100ul TE buffer. The samples were fragmented using a Covaris E220 Ultrasonicator (200 cycles/burst, Duty cycle 5, 175 W, 10%) for 180 s per sample. Fragmented RNA was incubated with 2.5-5 ug of anti-m7G antibody (SYSY, cat no. 201001) pre-coupled to Protein G magnetic beads for 1 hour in IPP buffer (50 mM Tris-HCl pH 7.4, 150 mM NaCl, 0.1% NP-40) rotating at 4 °C. Beads were washed three times with IPP and RNA 3' ends were dephosphorylated using PNK for 30 min at 37 °C. Beads were washed and the s-oligo was ligated using T4 RNA Ligase 1 for 1 hr at 25 °C. S-oligos contain barcode and random nucleotides in the following pattern NNNNNTTTTTNN (N=random nucleotide; T=barcode nucleotide). Excess s-oligos were washed away with IPP buffer and samples were pooled together. After 30 min of treatment with rSAP at 30 °C to dephosphorylate the s-oligo, the RNA was released from the beads using Proteinase K treatment and column purification (Oligo Clean and Concentrator (OCC), Zymo Research). Isolated RNA was reverse-transcribed using SuperScript III (Invitrogen) for 10 consecutive minutes at 42, 50, 55 and 65 °C. After 30 min treatment with RNaseH the cDNA was column-purified using OCC and circularized with CircLigase2 for 2-16 hr. 1 ul of circularized cDNA was taken to determine the amplifications

cycles using qPCR. After PCR amplification the libraries were cleaned up twice using 1x Ampure beads (Beckman Coulter), quantified with Qubit (Thermo Fisher Scientific) and the quality was assessed on Bioanalyzer (Agilent). MAPCap libraries were sequenced on Illumina NextSeq 500, 3000 or HiSeq 2000.

Processing of MAPCap data

Paired-end FASTQ files were trimmed for adaptors using Trimmomatic [16] (v 0.3.7). Samples were de-multiplexed by icetea (v0.99, demultiplexFastq) using provided barcode information, and mapped to the dm6 genome using Rsubread [17] (v 1.22.3, mapping wrapper provided in icetea). For de-duplication, we consider all reads mapping to the same 5'-position and having the same random barcode as duplicates and only keep the first instance of each such alignment (using icetea - filterDuplicates). BigWigs were created using deepTools [18] (v3.0.2) bamCoverage and bamCompare, with the option `--offSet 1 --binSize 1`. Quality control was performed using deepTools and multiQC [19] (v1.3). Genomic regions were plotted using pyGenomeTracks [20] (v2.0). The full MAPCap data processing workflow (described in Fig. S3A) is available at : https://github.com/vivekbhr/cage_pipeline

Comparison with external methods

For evaluation of TSS detection accuracy, we used the *paraclu* method [21] to cluster CAGE tags from CAGE, RAMPAGE and MAPCap data. We used the 12-14hr Sample from modENCODE, and merged the 12, 13 and 14hr samples from RAMPAGE to compare with merged (embryo 1-4) samples from MAPCap data. All samples were then downsampled to 10 million reads and paraclu was run with the parameters : *min_value* = 1, *min_density_rise* = 1, *min_pos_with_data* = 1, *min_sum* = 1, *min_width* = 3, *max_width* = 300 (i.e. all criteria such as minimum reads used for clustering and minimum density of reads per cluster etc. were kept to the lowest, and tag clusters of length 3 to 300 bp were considered for analysis).

The “density score” provided by paraclu for each tag cluster was used to calculate precision and sensitivity. Scores between 10 and 500 were plotted for ROC curve. For the evaluation of true and false positives, we used the RNA-seq data of 12-14 hr embryo from modENCODE and

calculated transcript-level TPMs using Salmon [22]. Transcripts with TPM > 1.0 were considered “expressed” and the TSSs of expressed transcripts which were not detected by the promoter-profiling methods considered “false negatives”. TSSs detected by the methods which did not overlap with a known TSS in dm6 (ensembl-79) annotation were considered “false positives”.

For comparison between replicate-based and paraclu method, we ran paraclu on samples with same criteria, while for our new method we obtained TSSs using a 2-fold local background cutoff. The score per TSS (mean fold-change across enriched windows) was used to evaluate the true and false positives for the analysis. Since the range of scores obtained per TSS is very different between paraclu and our method, there is no comparable cutoff for comparison of precision and sensitivity.

TSS detection and differential TSS usage analysis using replicates

For the detection of transcription start sites using replicates, we first count the 5'-end of reads in 10 bp sliding windows (w) across the genome for all samples (with a slide of 5 bp). For each window, we also calculate all 5'-ends of the reads falling into the corresponding 2 kb background region (b) centered at the window. Counting is done in a strand-specific way, using the *intersectionStrict* mode. We then calculate the fold change (delta) of each window with respect to the background as :

$$\delta = \text{Avg}(\widehat{\omega}) / \text{Avg}(\widehat{b})$$

Where $\text{Avg}(\widehat{\omega})$ and $\text{Avg}(\widehat{b})$ are average logCPM values across replicates, obtained by a fitting single group negative binomial glm :

$$\widehat{Y}_{wi} \sim NB(Mipwj, \phi w)$$

implemented in *mglmOnegroup* function of the edgeR package [23].

For differential TSS usage analysis, strand-specific counting is performed in the same way, on the union of TSSs detected across samples. Library sizes were normalized using the size factors obtained from ERCC counts using median of ratios method from DESeq2. The differential expression analysis was then performed in DESeq2 using *nbinomWaldTest* function.

TSSs with adjusted $p < 0.05$ were considered significantly different between tissues and sexes.

To perform differential gene expression analysis from MAPCap data, we summed the counts obtained from all 3'UTRs of a gene into one, and performed the normalization and differential expression using DESeq2. Spike-in normalization was performed the same way as above.

TSS annotation

For a comprehensive annotation of our detected TSSs, we first created a mutually exclusive set of annotations from dm6 (ensembl-79) GTF file, by first separating genic from intergenic regions, followed by ranking them in this order (5'UTR > CDS > 3'UTR > Introns; and sense > antisense). Further the features were re-annotated by overlapping them with enhancers [24] and repeats (RepBase release 20140131). The annotation pipeline is available as part of the full MAPCap data processing pipeline at : https://github.com/maxplanck-ie/cage_pipeline

Evaluation of promoter width

To evaluate the promoter width distribution obtained from icetea analysis, we divided our 12921 detected TSS into “broad” and “sharp” categories, by taking arbitrary cutoffs : >50 bp (8.1%) and <20 bp (36%), respectively. We performed GO enrichment analysis of the two categories for biological processes (BP) terms and plotted them using the *clusterProfiler* bioconductor package [25] ($p < 0.01$, $q < 0.05$). Further we, extracted the FASTA sequences associated with the two categories from the dm6 genome using the *BSgenome* package and performed de-novo motif enrichment via *meme* [26]. We sampled 10000 3'UTR sequences (100 bp regions) and used them as control for the motif enrichment. *Meme* was then run with the parameters: *-mod zoops -nmotifs 3 -minw 6 -maxw 30 -dna -revcomp*

Acknowledgements

The authors acknowledge the deep-sequencing unit at MPI-IE for data production. AA and TM acknowledge funding from the German Science Foundation (CRC992 “Medical Epigenetics”).

Author Contributions

VB performed the analysis of data with input from GS, developed the icetea bioconductor package, and wrote the manuscript with input from all authors. GS developed the MAPCap protocol with analysis input from VB and performed all the experiments. GS and VB conceived the project with input from AA. TM and AA supervised VB and GS during the project.

Code availability

Icetea is available open source at <https://github.com/vivekbhr/icetea>. All the data presented in the manuscript has been processed via the cage analysis pipeline available at https://github.com/vivekbhr/cage_pipeline.

Conflict of interest

The authors declare no conflict of interest.

References

1. Reyes A, Huber W. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res.* 2018;46:582–92.
2. Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, et al. CAGE: cap analysis of gene expression. *Nat Methods.* 2006;3:211–22.
3. Batut P, Dobin A, Plessy C, Carninci P, Gingeras TR. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res.* 2013;23:169–80.
4. Salimullah, Sakai M, Plessy C, Carninci P. NanoCAGE: A High-Resolution Technique to Discover and Interrogate Cell Transcriptomes. *Cold Spring Harb Protoc.* 2011;2011:db.erratum2011_01.
5. Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet.* 2014;46:1311–20.
6. Adiconis X, Haber AL, Simmons SK, Levy Moonshine A, Ji Z, Busby MA, et al. Comprehensive comparative analysis of 5'-end RNA-sequencing methods. *Nat Methods [Internet].* 2018; Available from: <http://dx.doi.org/10.1038/s41592-018-0014-2>
7. Aktaş T, Avşar İlök İ, Maticzka D, Bhardwaj V, Pessoa Rodrigues C, Mittler G, et al. DHX9 suppresses RNA processing defects originating from the Alu invasion of the human genome. *Nature.* 2017;544:115–9.
8. The modENCODE Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, et al. Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. *Science. American Association for the Advancement of Science;* 2010;330:1787–97.
9. Batut P, Gingeras TR. RAMPAGE: Promoter Activity Profiling by Paired-End Sequencing of 5'-Complete cDNAs. *Current Protocols in Molecular Biology.* John Wiley & Sons, Inc.; 2001.
10. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet.* 2006;38:626–35.
11. Kawaji H, Lizio M, Itoh M, Kanamori-Katayama M, Kaiho A, Nishiyori-Sueki H, et al. Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. *Genome Res.* 2014;24:708–17.
12. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based

- analysis of ChIP-Seq (MACS). *Genome Biol.* 2008;9:R137.
13. Lun ATL, Smyth GK. csaaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Res.* 2016;44:e45.
 14. Lenhard B, Sandelin A, Carninci P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet.* 2012;13:233–45.
 15. Huang F, He J, Zhang Y, Guo Y. Synthesis of biotin-AMP conjugate for 5' biotin labeling of RNA through one-step in vitro transcription. *Nat Protoc.* 2008;3:1848–61.
 16. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30:2114–20.
 17. Liao Y, Smyth GK, Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* 2013;41:e108.
 18. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* 2016;44:W160–5.
 19. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics.* 2016;32:3047–8.
 20. Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, et al. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun.* 2018;9:189.
 21. Frith MC, Valen E, Krogh A, Hayashizaki Y, Carninci P, Sandelin A. A code for transcription initiation in mammalian genomes. *Genome Res.* 2008;18:1–12.
 22. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods.* 2017;14:417–9.
 23. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* 2010;26:139–40.
 24. Kvon EZ, Kazmar T, Stampfel G, Yáñez-Cuna JO, Pagani M, Schernhuber K, et al. Genome-scale functional characterization of *Drosophila* developmental enhancers *in vivo*. *Nature.* 2014;512:91–5.
 25. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS.* 2012;16:284–7.
 26. Bailey TL, Elkan C, Others. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. Department of Computer Science and Engineering, University of California, San Diego; 1994; Available from: http://www.cs.toronto.edu/~brudno/csc2417_15/10.1.1.121.7056.pdf

Supplementary Figures

Figure S1

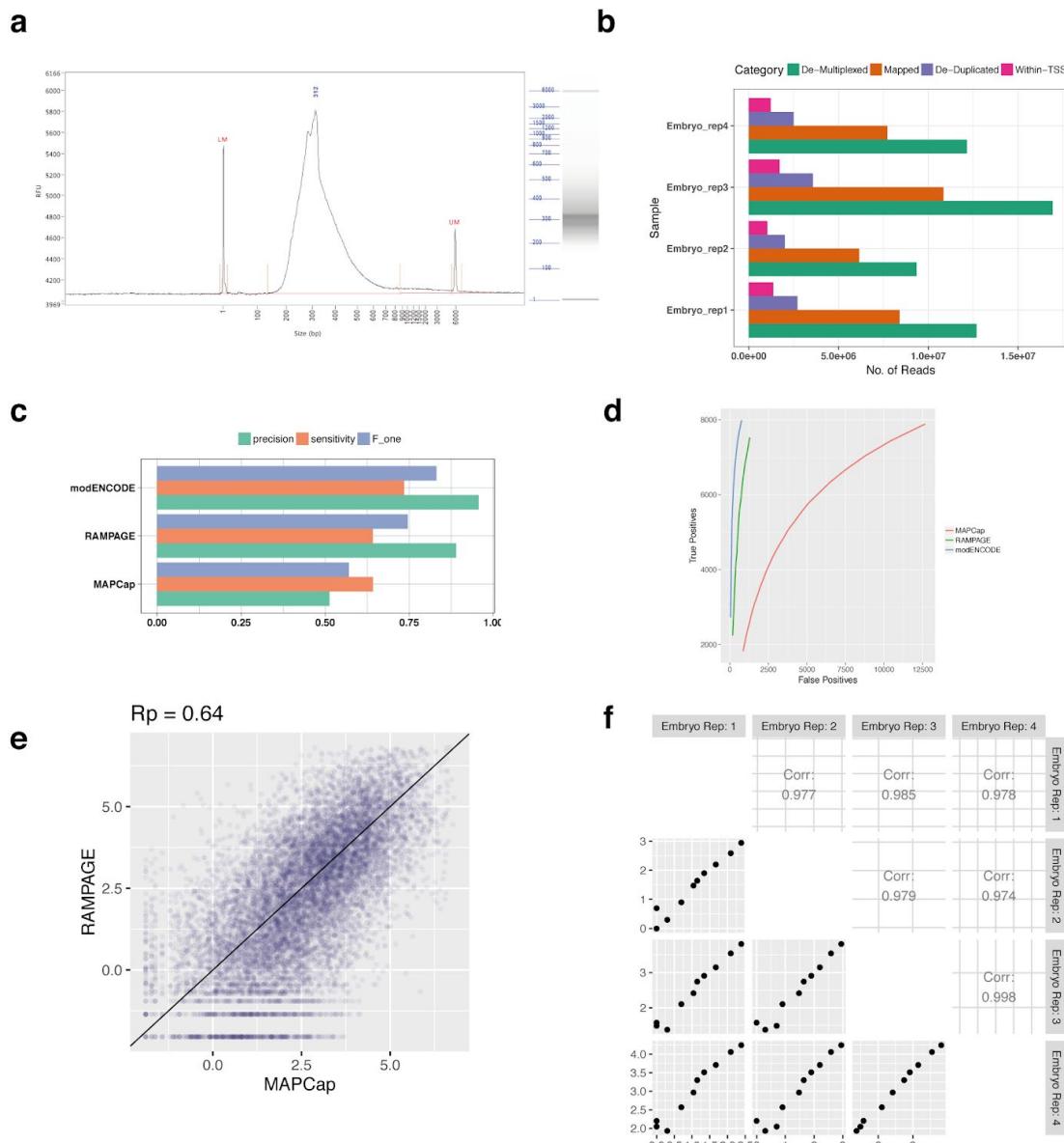


Fig. S1. Assessment of the MAPCap data quality. A. BioAnalyzer profile of the MAPCap library. B. Number of reads kept at each step of MAPCap analysis, for the embryo samples. C.

Precision, Sensitivity and F1-score analysis of MAPCap and other protocols. **D.** ROC curve of the MAPCap protocol compared to other protocols. TSS overlapping with those present in dm6 annotation and also expressed in modENCODE RNAseq data were considered true positives for these analysis. **E.** Same as Fig. 1C; correlation of depth-normalized read counts for MAPCap and RAMPAGE. **F.** Correlation of recovered counts of individual ERCC oligos between samples. Oligo mix was created after 2-fold serial dilution of individual oligos, which is also reflected in the data.

Figure S2

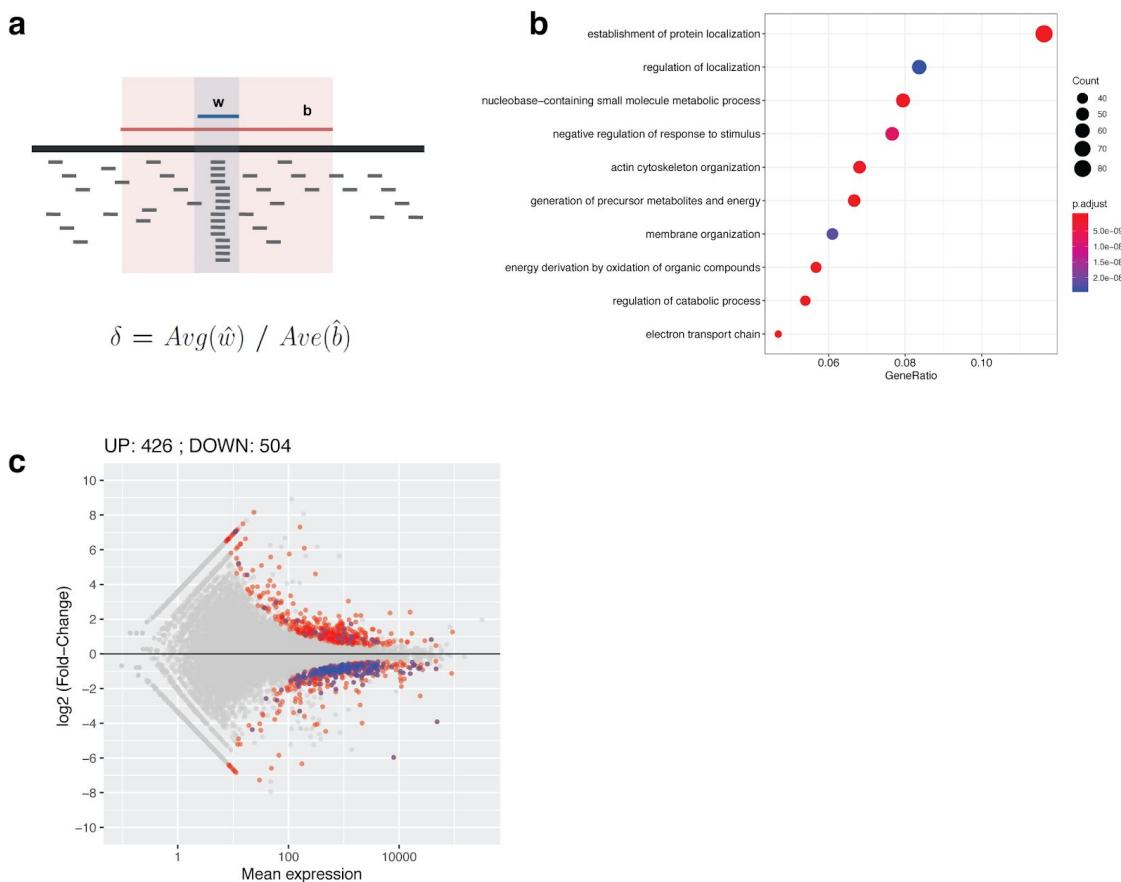


Fig S2. Evaluation of the new TSS detection and differential expression approach. **A.** Schematic diagram of “local enrichment” method, the fold-change (δ) for windows (w) over background (b) is calculated as average fold change of replicates after depth-normalization. **B.** GO enrichment of “broad” (>100bp) TSS detected using our method shows enrichment of basic cellular functions. **C.** MA plot of gene-level differential expression estimates from MAPCap data using internal normalization.

Figure S3

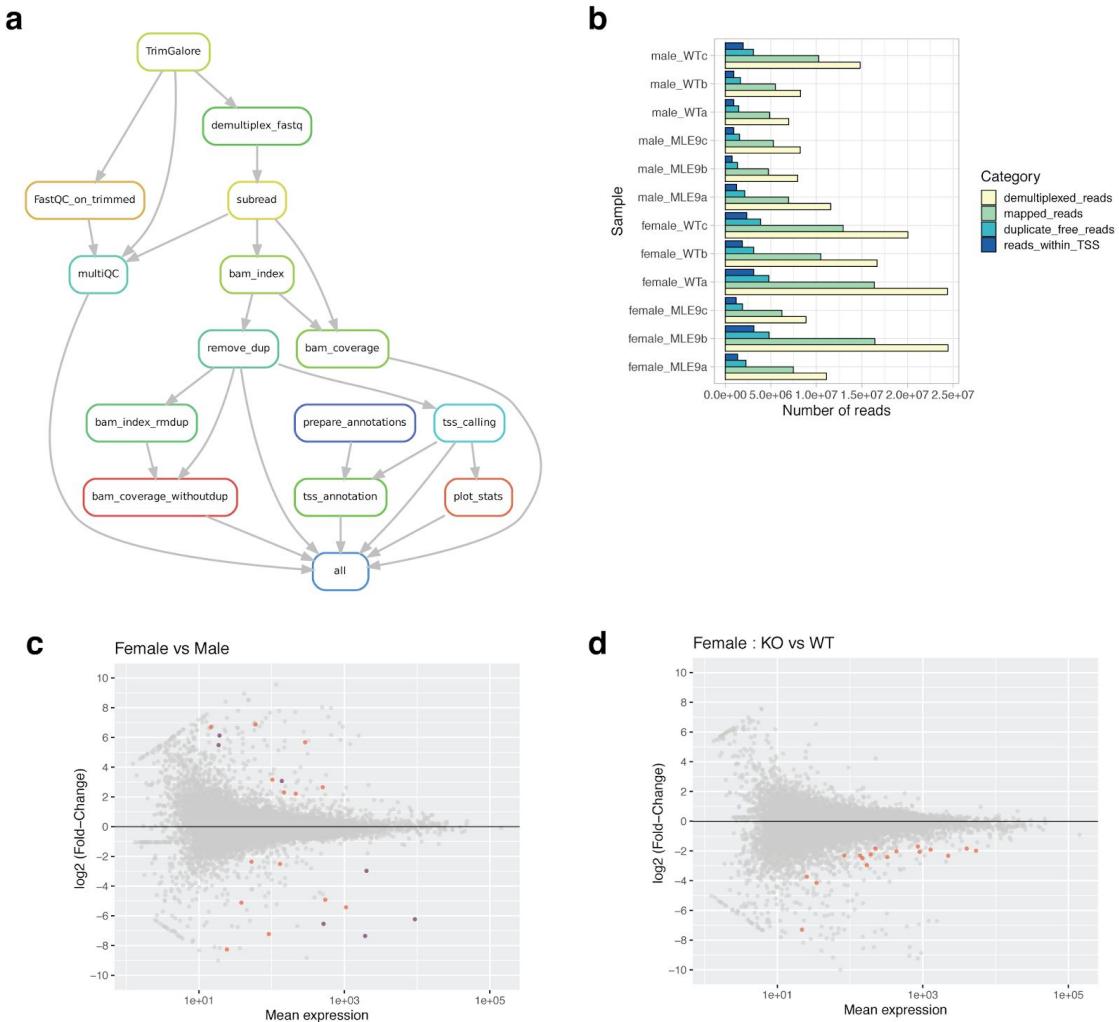


Fig S3. Analysis of dosage compensation defects in MLE KO flies. **A.** Workflow used for the analysis of data. It involves read trimming, demultiplexing using sample barcodes, mapping, duplicate removal (using random barcodes), TSS detection, TSS annotation and generation of coverage files and plots (see methods). **B.** Number of reads kept at each step of the analysis. **C-D.** MA plot of differentially used promoters between wild-type male and female brains, and between MLE KO female brain over wild-type.

Supplementary table 1. List of oligos used for abundant RNA depletion :

CCATAAGGCCGAGAACCGAT
CCTCTACGCCAGGTAAGTAT
TATGCCCTGCGCAAAGAT
TATTGCCACTGCGCAAAGAT
TGATGATCCCCGACACTCGA
CAGTCTACCTCTACTAATGA
TGAAGCGGGGATCGAGACAT
ATCCTGTGAAGTATAAGTCTT
AATTGAAGAGAAACCAGAGT
AGAGAATAAAAATTTCAAT
ACCCAATCGTCACCTCTCGCA
CCAGGACGAGCACCCTTTT
CTCCCCAAGACAAGGAAGGT
TACTCATTAGTTGAGGCAC
TTCATCATATCATCTAGAGA
TGTTCTGCCGAAGCAAGAAC
GCTCTCCTCCAAACACAC
ATGTAATGTTCATCATGTCG

A.4 Interaction of MLE ortholog DHX9 with Alu elements in the human genome

I performed the analysis of UV-CLAP data for estimation of Alu enrichment (Extended data Fig. 2 and 3) and performed the RNA-seq analysis for detection of differential expression, splicing, circular RNAs and RNA editing (Fig. 2a, 2b, 3a, Extended Data Figure 6, 7, 10). I contributed to the writing and revision of the manuscript along with Tugce Aktas, Ibrahim Ilik, Asifa Akhtar and other authors.

LETTER

doi:10.1038/nature21715

DHX9 suppresses RNA processing defects originating from the Alu invasion of the human genome

Tuğçe Aktaş^{1*}, İbrahim Avşar Ilik^{1*}, Daniel Maticzka², Vivek Bhardwaj^{1,3}, Cecilia Pessoa Rodrigues^{1,3}, Gerhard Mittler¹, Thomas Manke¹, Rolf Backofen² & Asifa Akhtar¹

Transposable elements are viewed as ‘selfish genetic elements’, yet they contribute to gene regulation and genome evolution in diverse ways¹. More than half of the human genome consists of transposable elements². *Alu* elements belong to the short interspersed nuclear element (SINE) family of repetitive elements, and with over 1 million insertions they make up more than 10% of the human genome². Despite their abundance and the potential evolutionary advantages they confer, *Alu* elements can be mutagenic to the host as they can act as splice acceptors, inhibit translation of mRNAs and cause genomic instability³. *Alu* elements are the main targets of the RNA-editing enzyme ADAR⁴ and the formation of *Alu* exons is suppressed by the nuclear ribonucleoprotein HNRNPC⁵, but the broad effect of massive secondary structures formed by inverted-repeat *Alu* elements on RNA processing in the nucleus remains unknown. Here we show that DHX9, an abundant⁶ nuclear RNA helicase⁷, binds specifically to inverted-repeat *Alu* elements that are transcribed as parts of genes. Loss of DHX9 leads to an increase in the number of circular-RNA-producing genes and amount of circular RNAs, translational repression of reporters containing inverted-repeat *Alu* elements, and transcriptional rewiring (the creation of mostly nonsensical novel connections between exons) of susceptible loci. Biochemical purifications of DHX9 identify the interferon-inducible isoform of ADAR (p150), but not the constitutively expressed ADAR isoform (p110), as an RNA-independent interaction partner. Co-depletion of ADAR and DHX9 augments the double-stranded RNA accumulation defects, leading to increased circular RNA production, revealing a functional link between these two enzymes. Our work uncovers an evolutionarily conserved function of DHX9. We propose that it acts as a nuclear RNA resolvase that neutralizes the immediate threat posed by transposon insertions and allows these elements to evolve as tools for the post-transcriptional regulation of gene expression.

Mammalian cells react to excessive double-stranded RNA (dsRNA) as a sign of viral infection with an intricate system collectively called the dsRNA response⁸. ADAR³, PKR⁹, Staufen¹⁰, as well as cytoplasmic RNA helicases MDA-5 and RIG-I¹¹, are known to be involved in suppression or activation of the dsRNA response. DHX9 has a unique domain organization that resembles ADAR and PKR (Extended Data Fig. 1a). Surprisingly, despite its essential nature in mouse and human cells^{12,13} and abundance (>1 million copies per cell⁶), little is known about the role of DHX9 in cellular homeostasis. Thus, in order to shed light on the cellular function of DHX9, we first identified its *in vivo* targets by using a UV-crosslinking-based method (uvCLAP: UV crosslinking and affinity purification, Extended Data Fig. 1b, see Methods). Generation of a cell line that expresses DHX9 with a 3×Flag and HBH tag at its C terminus¹⁴ allowed direct comparison with other RNA-binding proteins including QKI-5, QKI-6, KHDRBS1-3, HNRNPK and EIF4A1 (Fig. 1a and Extended Data Figs 1, 2). We observed that human/mouse DHX9 and

its fly orthologue Maleless (Mle) interact mainly with intronic RNA (Extended Data Fig. 2i). Closer inspection revealed that DHX9 peaks are significantly enriched on *Alu* SINEs (Fig. 1a): 60% of DHX9 peaks are on or within 100 nucleotides (nt) of *Alu* repeats; compared to 3–10% for controls (Fig. 1b, all P values $<2.2 \times 10^{-16}$, Fisher’s exact test). Using another UV-crosslinking method (FLASH: fast ligation of RNA after some sort of affinity purification for high-throughput sequencing, see Methods) that can utilize antibodies against endogenously expressed proteins in addition to tagged constructs, we obtained virtually identical DHX9 profiles to those derived with the tagged protein using uvCLAP, confirming the robustness of our observations (Fig. 1a, also see Extended Data Fig. 2a–d). DHX9 peaks obtained by FLASH fall predominantly on *Alu* repeats (Fig. 1b and Extended Data Fig. 1d) and shuffling DHX9 FLASH peaks within introns significantly reduces the percentage of peaks that fall on *Alu* elements (Fig. 1b). Analysis of uvCLAP/FLASH peaks that do not fall directly on *Alu* elements shows that DHX9 peaks tend to be closer to *Alu* elements compared to peaks of other RNA-binding proteins (Fig. 1c and Extended Data Fig. 1c). Shuffling the DHX9 peaks randomly within introns abrogates this closeness to *Alu* elements (Fig. 1c and Extended Data Fig. 1c). Using two genome-mapping-free approaches we further verified that the *Alu* enrichment we score is specific to DHX9 (see Methods and Extended Data Fig. 2e–g).

As *Alu* elements are strictly primate specific-retrotransposons¹⁵, we next tagged endogenous DHX9 with 3×Flag–Avitag¹⁶ in mouse embryonic stem cells and carried out uvCLAP/FLASH experiments (Extended Data Figs 2h and 3a–c). In mice, DHX9 peaks fall predominantly on B1 SINEs (Extended Data Fig. 1f and 3d–f). DHX9 peaks that do not directly fall on B1 elements are close to them compared to shuffled controls (Extended Data Fig. 1g, h). Notably, both *Alu* elements and B1 elements evolved independently from the ancestral 7SL RNA¹⁷ (Extended Data Fig. 1e). Similar to the human data, we used genome-mapping-free methods to verify the specificity of B1 enrichment in DHX9 data (Extended Data Fig. 3d, e).

The common denominator between *Alu*, B1 and roX RNAs (main targets of human, mouse and *Drosophila* DHX9) is their propensity to form structured RNAs^{18,19} that can directly attract DHX9 activity. Notably, however, even a visual inspection of DHX9 binding reveals that not all *Alu* elements are DHX9 targets (Fig. 1a). In order to determine the rules of DHX9 binding, we used a genomic similarity search tool (YASS²⁰) to look for *Alu* elements that can base pair with each other to form exceptionally long dsRNA (Extended Data Fig. 4a, b). We observed that *Alu* elements that are not targeted by DHX9 have a median distance of 1,482 nt ($n = 173,653$) to their closest potential binding partner, whereas *Alu* elements that are targeted by DHX9 have a much shorter median distance of 458 nt ($n = 13,663$) indicating that DHX9 specifically targets long dsRNA formed by base pairing *Alu* elements (Fig. 1d, two-tailed Mann–Whitney U -test, $P < 2.2 \times 10^{-16}$;

¹Max Planck Institute of Immunobiology and Epigenetics, Freiburg, Germany. ²Institute for Informatics, Albert-Ludwigs-University, Freiburg, Germany. ³Faculty of Biology, University of Freiburg, 79104 Freiburg, Germany.

*These authors contributed equally to this work.

RESEARCH LETTER

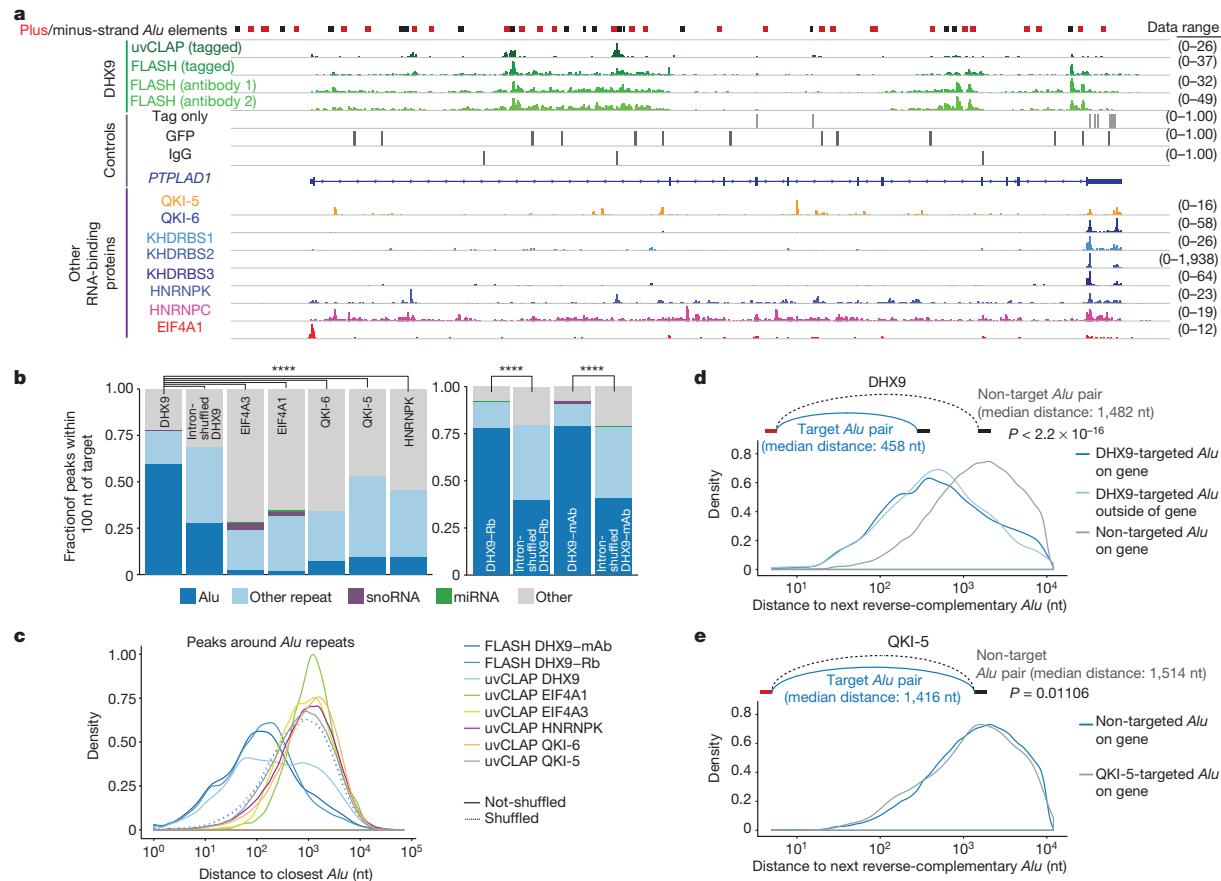


Figure 1 | uvCLAP and FLASH reveal DHX9 binding on *Alu* elements. **a**, IGV (Integrative Genomics Viewer) snapshot showing tagged (uvCLAP) and endogenous (FLASH) DHX9 binding on *Alu* elements of PTPLAD1. Additional tracks include control experiments and other RNA-binding proteins (tracks show merged biological duplicates). ‘Data range’ shows the coverage of uniquely mapped alignments for each profile. **b**, Enrichment of uvCLAP (left) and FLASH (right) peaks on *Alu* (all P values $<2.2 \times 10^{-16}$, Fisher’s exact test) or other elements in humans.

miRNA, microRNA; snoRNA, small nucleolar RNA. **c**, Distance of uvCLAP/FLASH peaks to the nearest *Alu* element. DHX9 peaks are closer to *Alu* elements, compared to shuffled peaks and other RNA-binding proteins (all P values $<2.2 \times 10^{-16}$, one-tailed Mann–Whitney U -test). **d**, **e**, Distance distributions of target, or non-target *Alu* pairs in DHX9 (**d**) or QKI-5 (**e**) uvCLAP peaks (P values calculated with two-tailed Mann–Whitney U -test, see Supplementary Table 3).

Extended Data Fig. 4c–h). DHX9 FLASH yields virtually identical results while other RNA-binding proteins appear to be agnostic to the positioning of *Alu* elements in and around their binding sites (Fig. 1e, Extended Data Fig. 4c–h and Supplementary Table 3).

Presence of *Alu* repeats and complementary sequences around exons correlate with circular RNA (circRNA) formation^{21,22}. To determine whether DHX9 plays a role in this process, we first checked the abundance of previously reported circRNAs²³ using quantitative reverse-transcription PCR (RT-qPCR) in DHX9-depleted cells and observed a robust increase in circRNA levels (Extended Data Fig. 5a). Next, we generated sequencing libraries from poly(A)⁺-depleted total RNA and indeed observed a clear, reproducible and global increase in the number of unique circRNAs in DHX9-depleted cells (Fig. 2a). In total, we detected 25,658 unique circRNAs in control short interfering RNA (siRNA)-treated samples and 50,355 in DHX9-siRNA-treated samples. DHX9-specific circRNAs contain significantly more *Alu* elements in flanking introns compared to control-specific circRNAs (90.65% in at least one flanking intron, 71.78% in both, $P < 2.2 \times 10^{-16}$, Fisher’s exact test). Moreover, DHX9-depleted cells showed more than a fourfold increase in circRNA-producing genes compared to controls (Fig. 2b). We independently verified the upregulation and RNaseR insensitivity of the ten most upregulated circRNAs by RT-qPCR (Fig. 2c and Extended Data Fig. 5b).

In order to determine whether DHX9 has any effect on mRNAs that contain inverted-repeat *Alu* elements in their 3' UTRs, we used a luciferase reporter system comparing constructs that have inverted-repeat-*Alu*-containing 3' UTRs to *Alu*-free 3' UTR constructs (Extended Data Fig. 5c, d). We observed a significant downregulation of luciferase activity for constructs that contain 3' UTR inverted-repeat *Alu* elements upon DHX9 depletion (Fig. 2d, Extended Data Fig. 5e–g). This effect could be rescued by overexpressing a wild-type DHX9 transgene but not a ‘helicase-dead’ mutant²⁴ (DHX9(GET)), suggesting that DHX9 resolves rather than passively coats inverted-repeat *Alu* elements, as previously suggested for Staufen proteins²⁵, in order to de-repress translation of mRNAs with inverted-repeat *Alu* elements in their 3' UTRs (Fig. 2d, Extended Data Fig. 5c–g).

In addition to increasing circRNA formation, *Alu* elements can potentially disrupt normal RNA biogenesis by masking primary RNA sequences for splicing and transcript termination. Analysis of sequencing libraries prepared with poly(A)⁺ RNA from DHX9-depleted cells revealed significant changes in the expression of 5,890 genes (Extended Data Fig. 6a–g). Further analysis showed 9,146 differentially spliced exons (at FDR-adjusted $P < 0.01$) affecting more than 4,400 genes in total (Extended Data Fig. 6e, f). Notably, CCL25, a cytokine that is almost exclusively expressed in the intestines and tissues involved in T-cell development²⁶ was among the most highly upregulated genes

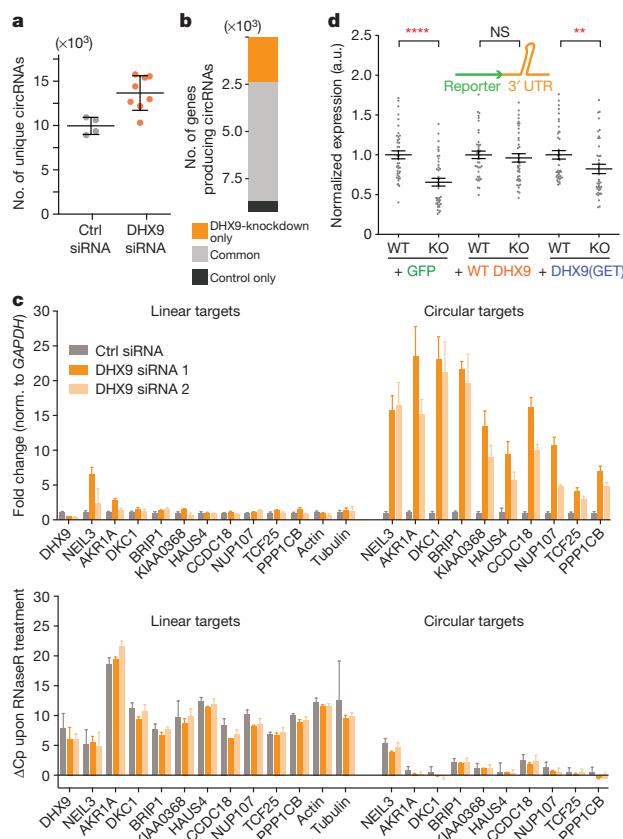


Figure 2 | Loss of DHX9 protein causes increased circular RNA generation and impaired translation. **a**, Number of unique circRNAs detected in control (grey) or DHX9-siRNA-treated samples (orange). Error bars represent s.d. of biological quadruplicates. **b**, Number of genes producing circRNA species. Orange, DHX9-knockdown-specific; black, control-specific; grey, common. **c**, Top, RT-qPCR assays quantifying changes in linear RNAs versus circRNAs. Bottom, sensitivity of linear/circRNA to RNaseR represented as the change in the value of crossing points (ΔC_p). Error bars represent s.e.m. of four biological replicates. **d**, Luciferase assays show that DHX9 depletion significantly reduces expression of inverted-repeat-Alu-containing 3' UTR elements, which can be alleviated with wild-type (WT) DHX9 overexpression and only partially with DHX9(GET). Mann-Whitney *U*-test was used to calculate *P* values. ****P* < 0.0001; ***P* = 0.0032; wild-type DHX9, *P* = 0.4278. Error bars represent s.d. of two biological replicates.

(Extended Data Fig. 6h). Closer inspection of this locus revealed that upregulation of *CCL25* is probably caused by a splicing/termination failure during the processing of an *ELAVL1* anti-sense transcript (*ELAVL1^{AS}*) that is bound by DHX9 *in vivo* (Extended Data Fig. 6i). Sashimi plots show that upon DHX9 depletion, the first exon of *ELAVL1^{AS}* is spliced to a cryptic splice-acceptor site in the first exon of *CCL25* (approximately 40 kb downstream of *ELAVL1^{AS}*), leading to around a 60-fold upregulation of *CCL25*. Similar defects were observed in other highly upregulated genes (see Extended Data Fig. 7a–d).

We next investigated the consequences of DHX9 depletion on cell morphology and observed the formation of giant cells containing numerous small, pleomorphic nuclei (Extended Data Fig. 8a–c, Supplementary Videos 1–7). Live cell imaging revealed that cells lose the ability to align their chromosomes at the metaphase plate and eventually disintegrate into small and large spherical pieces that rapidly fuse to form a large cell containing a variable number of small nuclei (as many as 23 nuclei per cell, Extended Data Fig. 8c; also compare Supplementary Video 2 with Supplementary Video 7). A similar

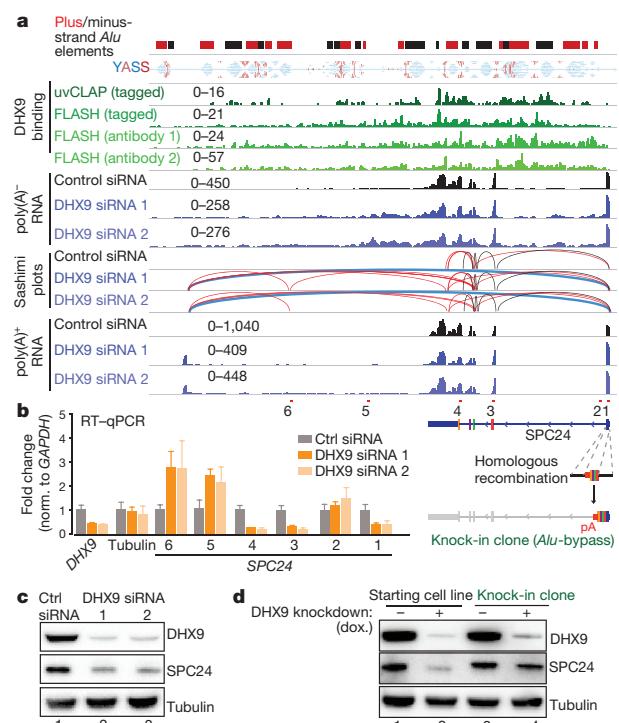


Figure 3 | Loss of DHX9 leads to RNA processing defects. **a**, IGV plots of DHX9-binding (green) and poly(A)⁺/poly(A)⁻ RNA sequencing (blue) data and Sashimi plots depicting exon-exon connectivities (red, non-canonical splicing events; blue, connection between the first and the newly emerging terminal exon of SPC24). For visual clarity, only the reads on the same strand as SPC24 are depicted. Data range shows the coverage of uniquely mapped alignments for each profile. **b**, Left, RT-qPCR validation of SPC24 reduction following DHX9 depletion (qPCR amplicons (1–6) are depicted in a, bottom). Error bars represent s.e.m. of biological quadruples. Right, schematic representation of the Alu-bypass allele. **c**, SPC24 levels in HeLa cells upon DHX9 knockdown with two independent siRNAs compared to control siRNA. **d**, SPC24 levels are shown before (lanes 1, 3) or after (lanes 2, 4) DHX9 depletion, in parental (left) or Alu-bypass (right) cell line. Representative of three experiments.

phenotype has been reported by the MitoCheck consortium²⁷ under the ‘Grape’ category. Among the 153 Grape-phenotype-causing genes that were expressed in cells in our study, SPC24 stands out as the most downregulated, and is one of four genes that has inverted *Alu* repeats at its 3' UTR (Fig. 3a, b, left, and Extended Data Fig. 8d, left). Consistent with the chromosomal alignment defects, SPC24 is part of the Ndc80 complex, which acts as a tether between microtubules and chromosomes, forming an essential part of the kinetochore²⁸ (Extended Data Fig. 8d, right). Transcriptional reduction of SPC24 also coincides with a reduction in its protein levels (Fig. 3c). Notably, both the 3' UTR of SPC24 and a downstream, *Alu*-rich intergenic domain are bound almost end-to-end by DHX9 (Fig. 3a). Similar to *CCL25*, Sashimi plots reveal that the first exon of SPC24 is frequently spliced to a newly emerging exon around 15 kb downstream of the original terminal exon (blue lines in Fig. 3a). To rescue this defect, we created an ‘Alu-bypass allele’ of SPC24 by knocking-in a construct that contains the coding region of SPC24 (amino acids 54–197), together with a polyadenylation site, immediately downstream of the first exon of SPC24 that codes for the first 53 amino acids (Fig. 3b, right, and Extended Data Fig. 8e). Unlike the wild-type allele, the *Alu*-bypass allele is expressed independently of DHX9, indicating that the presence of *Alu* elements interferes with SPC24 expression in the absence of DHX9 (Fig. 3d).

RESEARCH LETTER

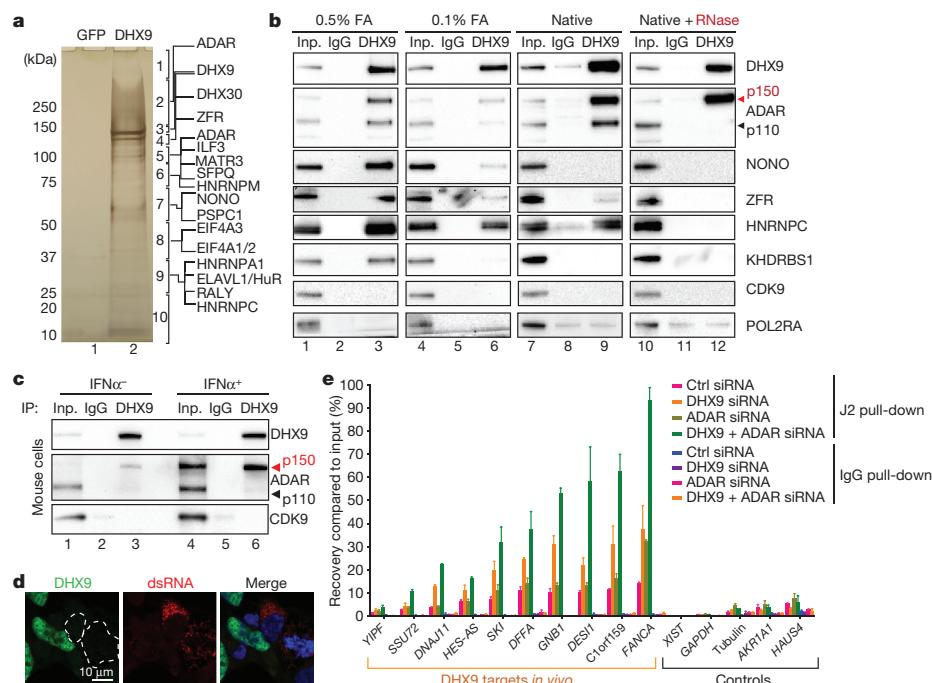


Figure 4 | RNA-independent interaction of ADAR p^{150} with DHX9 is conserved between mouse and human. a, a, Silver-stained gel image for the tandem-affinity purification experiment (see Extended Data Fig. 9a–f). **b,** Immunoprecipitation of the endogenous DHX9 protein using formaldehyde-crosslinked nuclei or native whole-cell lysates. Lanes 1, 4: inputs for 0.5%, 0.1% formaldehyde-crosslinked nuclei, respectively; lanes 7, 10: inputs for the native lysates without and with RNaseA treatment, respectively; lanes 2, 5, 8, and 11: IgG controls; lanes 3, 6, 9, 12: DHX9

immunoprecipitations. For gel source data, see Supplementary Fig. 1. **c,** Mouse DHX9 interacts with nuclear ADAR(p150) (lanes 3 and 6) (Also see Extended Data Fig. 9h–l). **d,** DHX9 depletion leads to accumulation of dsRNA (green, DHX9; red, dsRNA (J2); blue, DNA (Hoechst)). Representative of two independent experiments. **e,** RIP-qPCR (RNA immunoprecipitation quantitative PCR) with J2 antibody. Error bars show s.d. of two independent experiments.

Next, we used a stringent, SILAC (stable isotope labelling with amino acids in cell culture)-based purification strategy to determine the DHX9 interactome in human cells (Fig. 4a, Extended Data Fig. 9a–f and Methods). We identified 80 reproducible DHX9 interactors showing at least tenfold enrichment over background, (Supplementary Table 1) of which approximately 94% (75/80) are classified as ‘RNA-binding’ (UniProt 29). Purifications performed under native conditions with or without RNase treatment revealed that the majority of these interactions are RNA-bridged (Extended Data Fig. 9d–f and Fig. 4b). Notably, while we detected two major isoforms of ADAR (ADAR(p150) and ADAR(p110); interferon-inducible and constitutive isoforms of ADAR, respectively) co-immunoprecipitating with DHX9 under cross-linked and native conditions, only ADAR(p150)-interaction resisted RNase-treatment (Fig. 4b, lanes 9 and 12). We further verified this interaction by expressing Flag-tagged ADAR(p150) in HEK293 cells (Extended Data Fig. 9g). Despite the evolutionary divergence of SINEs, the DHX9–ADAR(p150) interaction is conserved in mice (Fig. 4c) and is most pronounced in the nucleus, even though ADAR(p150) is equally distributed between the nucleus and cytoplasm (Extended Data Fig. 9h–l).

Finally, we observed that DHX9-depletion leads to appearance of J2 mAb-positive speckles, indicative of accumulating dsRNA in cells (Fig. 4d and Extended Data Fig. 10a). RNA immunoprecipitation–qPCR experiments using the same antibody revealed that DHX9 depletion leads to a two- to threefold increase in J2-recovered DHX9 targets, while ADAR depletion has little effect (Fig. 4e). Surprisingly, although we did not observe major changes in RNA-editing upon DHX9 depletion (Extended Data Fig. 10c), co-depletion of DHX9 and ADAR led to an augmented phenotype in all target regions both in immunoprecipitated dsRNA (Fig. 4e)

and circRNA production (Extended Data Figs 10b, d), suggesting a synergistic interaction.

In summary, DHX9 binds to independently evolved SINEs in humans and mice and interacts specifically with ADAR(p150), underscoring its role as a nuclear dsRNA resolvase both under normal conditions and probably during viral invasion. We propose that elevated levels of DHX9 as a potential pre-emptive measure against viral invasions can increase the tolerance of the host genome for a higher number of SINE insertions over the course of evolution which may enhance SINE/LINE-mediated genomic or transcriptomic innovation in evolutionarily complex organisms (Extended Data Fig. 10e).

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 5 August 2016; accepted 23 February 2017.

Published online 29 March 2017.

1. Kazazian, H. H., Jr. Mobile elements: drivers of genome evolution. *Science* **303**, 1626–1632 (2004).
2. de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A. & Pollock, D. D. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* **7**, e1002384 (2011).
3. Elbarbary, R. A., Lucas, B. A. & Maquat, L. E. Retrotransposons as regulators of gene expression. *Science* **351**, aac7247 (2016).
4. Kim, D. D. Y. *et al.* Widespread RNA editing of embedded alu elements in the human transcriptome. *Genome Res.* **14**, 1719–1725 (2004).
5. Zarnack, K. *et al.* Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell* **152**, 453–466 (2013).
6. Hein, M. Y. *et al.* A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* **163**, 712–723 (2015).

7. Koh, H. R., Xing, L., Kleiman, L. & Myong, S. Repetitive RNA unwinding by RNA helicase A facilitates RNA annealing. *Nucleic Acids Res.* **42**, 8556–8564 (2014).
8. Gantier, M. P. & Williams, B. R. G. The response of mammalian cells to double-stranded RNA. *Cytokine Growth Factor Rev.* **18**, 363–371 (2007).
9. Lemaire, P. A., Anderson, E., Lary, J. & Cole, J. L. Mechanism of PKR Activation by dsRNA. *J. Mol. Biol.* **381**, 351–360 (2008).
10. Elbarbary, R. A., Li, W., Tian, B. & Maquat, L. E. STAU1 binding 3' UTR IRAIus complements nuclear retention to protect cells from PKR-mediated translational shutdown. *Genes Dev.* **27**, 1495–1510 (2013).
11. Oshiumi, H., Kouwaki, T. & Seya, T. Accessory factors of cytoplasmic viral RNA sensors required for antiviral innate immune response. *Front. Immunol.* **7**, 200 (2016).
12. Lee, C. G. et al. RNA helicase A is essential for normal gastrulation. *Proc. Natl Acad. Sci. USA* **95**, 13709–13713 (1998).
13. Wang, T. et al. Identification and characterization of essential genes in the human genome. *Science* **350**, 1096–1101 (2015).
14. Tagwerker, C. et al. A tandem affinity tag for two-step purification under fully denaturing conditions: application in ubiquitin profiling and protein complex identification combined with *in vivo* cross-linking. *Mol. Cell. Proteomics* **5**, 737–748 (2006).
15. Batzer, M. A. & Deininger, P. L. Alu repeats and human genomic diversity. *Nat. Rev. Genet.* **3**, 370–379 (2002).
16. Kim, J., Cantor, A. B., Orkin, S. H. & Wang, J. Use of *in vivo* biotinylation to study protein–protein and protein–DNA interactions in mouse embryonic stem cells. *Nat. Protocols* **4**, 506–517 (2009).
17. Kramerov, D. A. & Vassetzky, N. S. Origin and evolution of SINEs in eukaryotic genomes. *Heredity* **107**, 487–495 (2011).
18. Ilik, I. A. et al. Tandem stem-loops in roX RNAs act together to mediate X chromosome dosage compensation in *Drosophila*. *Mol. Cell* **51**, 156–173 (2013).
19. Quinn, J. J. et al. Revealing long noncoding RNA architecture and functions using domain-specific chromatin isolation by RNA purification. *Nat. Biotechnol.* **32**, 933–940 (2014).
20. Noé, L. & Kucherov, G. YASS: enhancing the sensitivity of DNA similarity search. *Nucleic Acids Res.* **33**, W540–W543 (2005).
21. Jeck, W. R. et al. Circular RNAs are abundant, conserved, and associated with AlU repeats. *RNA* **19**, 141–157 (2013).
22. Zhang, X.-O. et al. Complementary sequence-mediated exon circularization. *Cell* **159**, 134–147 (2014).
23. Memczak, S. et al. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495**, 333–338 (2013).
24. Lee, C. G., Chang, K. A., Kuroda, M. I. & Hurwitz, J. The NTPase/helicase activities of *Drosophila* maleless, an essential factor in dosage compensation. *EMBO J.* **16**, 2671–2681 (1997).
25. Ricci, E. P. et al. Staufen1 senses overall transcript secondary structure to regulate translation. *Nat. Struct. Mol. Biol.* **21**, 26–35 (2014).
26. Svensson, M. & Agace, W. W. Role of CCL25/CCR9 in immune homeostasis and disease. *Expert Rev. Clin. Immunol.* **2**, 759–773 (2006).
27. Neumann, B. et al. Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature* **464**, 721–727 (2010).
28. McCleland, M. L. et al. The highly conserved Ndc80 complex is required for kinetochore assembly, chromosome congression, and spindle checkpoint activity. *Genes Dev.* **17**, 101–114 (2003).
29. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204–D212 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank members of the Akhtar laboratory especially B. Sheikh, C. Keller and K. Lam for critical reading of the manuscript and helpful discussions. We also thank N. Ioving and R. Sawarkar for critical reading of the manuscript. We also thank the members of the Deep Sequencing, Imaging, FACS and the Proteomics Facilities for their support. We acknowledge the support of the Freiburg Galaxy Team: B. Grüning and T. Houwaart, Bioinformatics, University of Freiburg. This work was supported by CRC 992/2016 awarded to A.A., T.M. and R.B., and CRC 746 and CRC 1140 awarded to A.A. R.B. is also funded by BA2168/11-1, SPP 1738 and CRC TRR 167.

Author Contributions T.A., I.A.I. and A.A. designed the experiments, T.A. and I.A.I. performed most of the experiments except live-cell imaging and mass-spectrometry, which were carried out by C.P.R. and G.M., respectively; I.A.I. and T.A. developed uvCLAP and FLASH methods with input from D.M., R.B. and A.A.; D.M. analysed uvCLAP and FLASH data and was supervised by R.B.; V.B. analysed the RNA-seq data, carried out repeat-enrichment and circRNA analysis and was supervised by T.M. and A.A.; I.A.I., T.A. and A.A. wrote the manuscript with input from D.M. and V.B.; R.B. and A.A. acquired funding; A.A. supervised all aspects of the study. All authors reviewed, edited and approved the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to A.A. (akhtar@ie-freiburg.mpg.de).

METHODS

No statistical methods were used to predetermine sample size. $P < 0.01$ was considered statistically significant. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Cell culture and generation of stable cell lines. HEK293FT (gift from the Jenuwein laboratory), HeLa (gift from the Ellenberg laboratory) and FLPin Trex HEK293 (Thermo Fisher Scientific, R78007) cells are maintained with DMEM-Glutamax supplemented with sodium pyruvate, glucose and 10% FBS. In addition to those supplements, FLPin Trex HEK293 cells were maintained in zeocin- and blasticidin-containing medium according to manufacturer's protocol and the zeocin selection is exchanged with hygromycin upon transgene transfection. All the transgenes were cloned into pCDNA5-FRT/To (Thermo Fisher Scientific catalogue no. V6520-20) with a C-terminal 3×Flag–HBH tag and were co-transfected with pOG44 plasmid with a 1:9 DNA concentration ratio. Cells were re-plated in different dilutions 24 h after transfection and 150 µg ml⁻¹ hygromycin selection was initiated 48 h after transfection. Cell lines were maintained with blasticidin and hygromycin at all times and the transgenes were induced with either with 0.1 µg ml⁻¹ (for uvCLAP) or 1.0 µg ml⁻¹ (for purifications) doxycyclin.

CRISPR–Cas9 facilitated endogenous GFP tagging of the human DHX9 was performed in FLPinTrex HEK293 cell line carrying the inducible SPOP-GFP-nanobody³⁰. The repair template was co-transfected in a 3:1 ratio with SpCas9 (pX459, Addgene no. 62988) vector carrying the guide RNA #1 (Extended Data Fig. 5d). Cells were selected 48 h after the initial selection with 1 µg ml⁻¹ puromycin for 4 days total. Single-cell sorting into 96-well plates was carried out by gating for the cells with the strongest GFP signals (9.42% of all events). Colonies were expanded and screened for homozygously tagged DHX9 alleles using western blots. The knock-in of resistant allele at the SPC24 locus was performed in a homozygously GFP-tagged DHX9 clone using same conditions as described above (see Supplementary Table 2 for the sequences of guide RNAs, repair templates and screening oligos).

Mouse ES cells were maintained with 15% FBS, 2,000 U ml⁻¹ LIF, sodium pyruvate, NEAA, Glutamax and 0.1 nM β-mercaptoethanol supplemented into DMEM. CRISPR–Cas9 facilitated endogenous tagging of the mouse DHX9 was performed in a mouse ES cell line (WT26 male ES cell line was a gift from the Jenuwein laboratory) that we modified by adding a BirA ligase (plasmid provided by the Orkin laboratory¹⁶). Endogenous tagging was achieved by a co-transfection of ssODN (sequences of ssODN and screening oligos are listed in Supplementary Table 2) that provided the repair template with a 3×Flag–Avitag tag and the Cas9 plasmid that provided the guide RNA (based on pX459; Zhang laboratory) according to an earlier published protocol³¹. Initial screening of the clones was performed by PCR amplification flanking the tagged locus and scoring a size shift in PCR product. Positive clones from the first screen are then validated for the expression of Flag-tagged DHX9 protein (see Supplementary Table 2 for the sequences of guide RNAs, repair templates and screening oligos). ES cells were differentiated into NPCs as described previously³². FLP-in-3T3 cells were purchased from Thermo Fisher (R76107) and maintained in DMEM-Glutamax supplemented with sodium pyruvate, glucose and 10% FBS and zeocin according to the manufacturer's protocol. All cell lines are regularly checked for the absence of mycoplasma by PCR detection kit (Jena Bioscience PP-401).

uvCLAP. Stable cell line derived from FLPin Trex HEK293 cells are induced with 0.1 µg ml⁻¹ doxycycline for ~16 h; as controls, cell lines expressing either the tag alone or eGFP with the same tag were used. Cells are rinsed with PBS and crosslinked with 0.15 mJ cm⁻² UVC light. After crosslinking, cells are pelleted by centrifugation, snap-frozen in liquid nitrogen and kept at -80 °C until use. Cells are lysed with 0.5 ml of lysis buffer (50 mM Tris-Cl, pH 7.4; 140 mM NaCl, 1 mM EDTA, 1% Igepal CA-630, 0.1% SDS, 0.1% DOC), mildly sonicated and immunoprecipitated with anti-Flag beads for 1 h at 4 °C. Beads are washed with lysis buffer and bound material is eluted with 3×Flag peptide (250 µg ml⁻¹). The eluate is then incubated with MyONEC1 beads to collect biotinylated target protein, after which the beads are washed with high-stringency buffers (0.1% SDS, 1 M NaCl, 0.5% LiDS, 0.5 M LiCl and 1% SDS, 0.5 M LiCl) to remove non-specific interactors. 3' linkers are ligated with T4 RNA Ligase 1, excess adapters are washed away and 3'-tagged, cross-linked RNA is released with proteinase K digestion and column purification (Zymo DNA Clean & Concentrator). Reverse transcription is carried out with SuperScript III and barcoded reverse-transcription primers. After reverse transcription, relevant samples are mixed and the cDNA is separated on a 6% 6M Urea PAA gel. Size-fractionated cDNA is then processed essentially as described in the iCLIP protocol to generate sequencing libraries³³.

uvCLAP employs a tri-barcode approach that combines barcoded ScriptSeq Index PCR Primers and two custom barcodes located adjacent to the 5' and 3' adapters. Barcodes located at the 5' end of the inserts were designed using edittag³⁴.

Barcodes located at the 3' end of the inserts were created according to the semi-random patterns DRYYR and DYRRY (where D = not C; R = purine; Y = pyrimidine) and allowed to encode two mutually exclusive experimental conditions. uvCLAP uses a random tag strategy³⁵ to determine individual crosslinking events. uvCLAP primers contain 5 random nucleotides that are interleaved with the 5' barcodes according to the pattern NNNT₁T₂T₃T₄T₅NN (N = random nucleotide; T = barcode nucleotide). In combination with the semi-random 3' barcodes, this allows us to distinguish at least 49,152 ($4^5 \cdot 3 \cdot 2^4$) crosslinking events per nucleotide.

Size fractions were tagged via indexed PCR primers, pull-down conditions and biological replicates were tagged via 5' and 3' barcodes. Barcodes and random tags were extracted using custom scripts. Adapters were trimmed and the library was demultiplexed using Flexbar (v2.32)³⁶. Possible readthroughs into the barcoded regions were removed by clipping 10 and 5 nt from the 3' ends of first and second mate reads. Reads were mapped to the reference genomes (hg19, mm10, dm3) using bowtie2 (v2.2.0)³⁷ with parameters: --very-sensitive -end-to-end -no-mixed -no-discordant -maxins 200. We excluded all reads for which bowtie2 could identify multiple distinct alignments as indicated by the 'XS:i' flag and used the alignments of the remaining uniquely mapped reads to determine crosslinking events and crosslinked nucleotides as previously described¹⁸. Crosslinking events of the size fractions and biological replicates were combined.

uvCLAP peaks were called on crosslinked nucleotides with JAMM (version 1.0.7rev1, parameters: -d y -t paired -b 50 -w 1 -m normal)³⁸ using the two replicates of the respective pull-down condition as foreground and the two replicates of the corresponding control pull-down condition as background (Extended Data Fig. 2a-d, h). The peaks of all experiments were merged and used to calculate pairwise Spearman correlations among all replicates based on the number of events falling on each peak region, showing increased pairwise correlations between biological replicates within each condition over replicates of differing experimental conditions (deeptools version 2.3.5)³⁹ (see Extended Data Fig. 2a-d).

FLASH. Cells were crosslinked with 0.15 mJ cm⁻² UVC irradiation, lysed with 1× NLB buffer (1× PBS, 0.3 M NaCl, 1% Triton-X, 0.1% Tween-20) and homogenized by water-bath sonication. The target protein of interest was pulled-down with paramagnetic beads pre-coupled to antibodies against the protein of interest. After a brief RNase I-digestion, RNA 3' ends were healed with T4 PNK. Custom-made, barcoded adapters were ligated using T4 RNA Ligase 1 or T4 RNA Ligase 2KQ (if pre-adenylated adapters are used) for 1 h at 25 °C. Custom FLASH adapters contained two barcodes and random nucleotides adjacent to the 3'-adapters according to the pattern NNB₁B₂NT₁T₂T₃T₄T₅T₆NN (N = random tag nucleotide; T = tag nucleotide; B = RY-space tag nucleotide). Random tags were used to merge PCR-duplicates, regular tags were used to specify the pull-down condition, and semi-random RY-space tags were used to distinguish the biological replicates. Excess adapters were washed away, negative controls were mixed with experimental controls and RNA was isolated with Proteinase K treatment and column purification. Isolated RNA was reverse-transcribed and treated with RNase H. cDNA was column-purified and circularized with CircLigase for 2–16 h. Circularized cDNA was directly PCR amplified, quantified with Qubit/Bioanalyzer and sequenced on Illumina NextSeq 500 in paired-end mode.

FLASH read processing and mapping was performed using the Galaxy platform⁴⁰. Adapters were trimmed using Flexbar (v2.5). Libraries were demultiplexed using bctools (<https://github.com/dimaticzka/bctools>, v0.2.0) and Flexbar (v2.5). Possible readthroughs into the barcoded regions were removed by clipping 13 nt from the 3' ends of first mate reads. Reads were mapped to the reference genomes (hg19, mm10) using bowtie2 (v2.2.6) with parameters: --very-sensitive -end-to-end -no-mixed -no-discordant -maxins 500. We excluded all reads for which bowtie2 could identify multiple distinct alignments as indicated by the XS:i flag and used the alignments of the remaining uniquely mapped reads to determine crosslinking events as previously described¹⁸.

FLASH peaks were called on alignments of crosslinking events with PEAKachu (<https://github.com/tbischler/PEAKachu/releases/tag/0.0.1alpha2>, version 0.0.1alpha2, parameters: -pairwise_replicates -m 0 -n manual -size_factors 1 1 0.75 0.75)⁴¹, using the two replicates of the respective pulldown condition as foreground and the two replicates of the corresponding control pull-down condition as background (Supplementary Table 3). Pairwise Spearman correlations were calculated as for uvCLAP (Extended Data Fig. 2a-d, h).

Graph based clustering and repeat enrichment of uvCLAP data. We followed a strategy described before to assess the enrichment of different repeat families in uvCLAP samples⁴². Briefly, we created a library of canonical repeat sequences and repeat instances from the corresponding human (hg38) and mouse (mm10) genome, from repeatmasker (RepeatMasker open-4.0.5). We then mapped the sequencing reads to this library, and the uniquely mapping reads were used to calculate the maximum likelihood estimate of enrichment (MLE) for each repeat family, for the test (eg. DHX9) samples over the corresponding control (empty vector)

sample. We then performed hierarchical clustering using the MLE scores to visualize uniquely enriched repeat families for each sample.

As an alternative approach, we followed a mapping-free, graph-based sequence clustering approach described previously for plant genomes⁴³. In brief, we first sampled 100,000 reads from DHX9, EIF4A3, QKI-6 and QKI-5 datasets, followed by a clustering pipeline⁴⁴ which performs an all-to-all pairwise comparison between the reads in order to construct a graph, containing reads as nodes and the overlap between reads as edges. The vertices in the graph correspond to the reads and the edge weights represent the similarity score between the reads. The graph is then divided into clusters representing different connected communities. We then annotated the reads belonging to each cluster using RepeatMasker for the presence of different repeat families. Top clusters (with highest number of reads) for each sample are then visualized to assess the composition of major *Alu* families and other repeats. Increasing the sample size or multiple rounds of sampling had no effect on the composition of top clusters. Analysis code is publicly available at: <https://github.com/vivekblr/dhx-alu>.

Detection of paired *Alu* elements. We performed a similarity search (YASS²⁰ v1.4) to determine pairs of sequences reverse-complementary to each other within each gene (Ensembl 75). For each experiment, we selected all genes containing at least one peak and three annotated *Alu* elements (UCSC RepeatMasker track). Within the selected genes we determined *Alu* pairs based on all reverse-complementary YASS pairs with distances below 10,000 nt. Two *Alu* elements were considered paired if at least 75% of the first *Alu* was overlapped by one part of the YASS alignment and at least 75% of the second *Alu* was overlapped by the corresponding reverse-complementary part of the YASS alignment. For each paired *Alu* we then determined the distance to its nearest partner. *Alu* elements were categorized as targeted if a peak was located within 100 nt, otherwise as untargeted. For DHX9 we repeated this analysis for targeted *Alu* elements in extragenic regions.

Sample preparation for qPCR detection of circular RNAs and RNA-seq libraries. The siRNA knockdowns (5nM of Silencer Select (Ambion) control siRNA 4390846 or s4019 and s4020 siRNA for DHX9 and s1009 for ADAR) were performed on FLPinTrex HEK293 cells for 3 days in quadruplicates using RNAiMAX (Thermo Fisher Scientific). Total RNA extracts from siRNA transfected cells was prepared by Zymo Quick-RNA Kit and first-strand cDNA synthesis was carried out with SuperScriptIII. Quantitative real-time PCR is performed using the oligos listed in Supplementary Table 2 with FastStart SYBR Green Master (Roche 04 913 914 001). For RNase R treated samples: 500 ng of total RNA was incubated with 20 U of RNase R (Epicentre/Illumina) for 30 min at 30 °C after which the reactions were cleaned up with RNA Clean and Concentrator columns (Zymo) before proceeding with reverse transcription. For the poly(A)⁺ libraries, TruSeq Stranded mRNA Library Prep Kit (Illumina) was used as per manufacturer's instructions. For the poly(A)⁻ libraries, poly(A)⁺ RNA was depleted using paramagnetic oligo-d(T) beads (Thermo Fisher Scientific) twice. The remaining, poly(A)⁺-depleted RNA is processed with the TruSeq Stranded Total RNA Library Prep Kit (Illumina).

RNA sequencing and analysis. Both poly(A)⁺ and poly(A)⁻ reads were trimmed using Trim Galore to remove adaptors and mapped to GRCh38 (hg38) genome sequence using TopHat2 (v2.0.13)⁴⁵. Poly(A) reads were assigned to features (Gencode 21) using featurecounts (v1.5.1) (counting properly paired primary alignments above mapping quality of 10)⁴⁶, and differential expression analysis was performed using edgeR (v3.12.0)⁴⁷. We identified genes as differentially expressed if they are called significant in two independent comparisons (control vs siRNA1-s4019 and control vs siRNA2-s4020, FDR < 0.05), with expression fold change in same direction, in both comparisons. For analysis of differentially spliced exons, the gene models were flattened and differential expression was performed at exon level using DEXSeq (v1.18)⁴⁸. Exons were called differentially spliced at FDR adjusted $P < 0.01$. For the poly(A)⁻ library, the reads were mapped to the genome using STAR (v2.4.2)⁴⁹ (using option --chimSegmentMin 20) and the circular RNAs were annotated using circExplorer (v2.1)²². Depth of sequencing was not significantly different between samples, $P = 0.9482$, Welch *t*-test.

Luciferase assays. The siRNA knock-downs (5 nM of Silencer Select (Ambion) control siRNA 4390846 or s4019 siRNA for DHX9) were performed on HEK293FT or HeLa cells and the reporter constructs (50 ng of pmirGlo with 3' UTR inserts) were transfected on the 48th hour of knockdown. All luciferase assays were performed on the 72nd hour of knockdown. The knockout clone was generated by co-transfected the two guide RNAs; 1 and 2 (Extended Data Fig. 5d) co-expressing wild-type SpCas9 (pX459, Addgene no. 62988) and selected with 1 $\mu\text{g ml}^{-1}$ puromycin for 4 days. Colonies were picked, screened and expanded. Rescue constructs (GFP, wild-type DHX9, and DHX9(GET)) were transfected into wild-type (only Cas9-expressing) and knockout clones and the transfection of the luciferase constructs was performed 24 h after the rescue construct transfection in technical quadruplicates (see Extended Data Fig. 5e for the expression validation of the rescue constructs). All luciferase reads were performed on the 48th hour of the

primary transfection. The data points in Fig. 2d (and in Extended Data Fig. 5f) show all technical quadruplicates of the two biological replicates (that is, eight data points per 3' UTR construct).

Live cell imaging. All live cell imaging experiments were performed using monoclonal reporter cell lines that were generated as previously described⁵⁰. In brief, HeLa 'Kyoto' cell line was transfected with pH2V-mpH2B-mCherry-IRES-neo3 and pmEGFP- α -tubulin-IRES-puro2b plasmids. Thus, in the presence of neomycin and puromycin cells are H2B-mCherry-tubulin-EGFP stable reporter cells. Cells were cultured in Dulbecco's modified eagle medium (DMEM; GIBCO) supplemented with 10% (v/v) and treated either with siRNA control or siRNA against DHX9 in eight-well coverslip-bottomed dishes (ibidi GmbH cat. no. 80826). Six hours following transfection, medium was changed and supplemented with 1% penicillin/streptomycin, puromycin (1 $\mu\text{g ml}^{-1}$) and genetin/neomycin (200 $\mu\text{g ml}^{-1}$) (Invitrogen, Life Technologies, cat. no. 21810031). 24, 48 or 72 h post transfection, dishes were transferred to a Tokai Hit stage incubation unit, wherein cells were maintained at 37 °C and humidified atmosphere of 5% CO₂ throughout the experiment. Cells were imaged using a Zeiss Observer.Z1 with the CSU-X1 spinning disk head (Yokogawa) and the QuantEM:512SC camera (Photometrics) and were visualized using an air 40 \times objective. Z-stack parameters were taken every 5 min for at least 16 h and a maximum of 24 h. Laser power and exposure times were kept to a minimum condition (<9%). For analysis cells were manually counted and evaluated at the latest time point using Zen lite-Blue software (ZEISS). Nuclei numbers (Extended Data Fig. 8c) come from 15 fields of each condition at 24 h knockdown (in total 21 cells for wild type, 21 cells for DHX9 knockdown are included in the plot), from 5 fields for wild type and 15 fields for knockdown at 48 h knock-down (in total 28 cells for wild type and 40 cells for DHX9 knockdown are included in the plot), and from 10 fields for wild type and 15 fields for DHX9 knockdown at 72 h knockdown (in total 30 cells for wild type, 32 cells for DHX9 knockdown are included in the plot).

SILAC labelling of the cells and the biochemical purification of DHX9. Cells were adapted to SILAC medium (with K8R10 (1:1500 each) for Heavy or with K0R0 for Light) for 7 days and then were expanded to larger cell culture plates for another 7–10 days. Crosslinked purification was performed on 5–10 mg of 0.5% formaldehyde crosslinked nuclear extracts that were solubilized by mechanical shearing using sonication. Flag IP (Flag-M2 Agarose resin) was performed overnight at 4 °C and the proteins were eluted by 6 M urea-containing buffer (6 M Urea, 0.5 M NaCl, 0.5% NP-40, 50 mM Na₂HPO₄, 50 mM Tris pH 8.0, 0.05% Tween-20). Streptavidin (using MyOneC1 magnetic beads) pull-down was performed on the diluted (down to 1 M urea) eluate for 3 h at 4 °C. Streptavidin-bound proteins were sequentially washed with buffers containing 1 M and 6 M Urea. For the in-gel analysis proteins were eluted from beads by boiling and for the in-solution analysis beads were rinsed three times in *N*-octyl- β -glucopyranoside containing Tris (pH 7.4) buffer before trypsinization. Native purifications were performed on whole-cell extracts that were prepared in 450 mM NaCl and 1% Triton-X containing buffer. Silver Gel staining was performed using the Silver Quest Silver Staining Kit (Thermo Fisher LC6070).

Immunoprecipitation. Native whole-cell extracts or crosslinked nuclear extracts were incubated with 1 μg of DHX9 (polyclonal rabbit, see 'Antibodies' section) or of IgG (rabbit Santa Cruz sc-2017) antibody overnight at 4 °C and coupled to the ProtG Dynabeads (Life Technologies 10004D) for 1 h. Beads were washed in the IP buffer three times for 5 min each. Elution from the beads was performed in 2 \times protein loading dye by incubating for 10 min at 70 °C (or 95 °C for crosslinked samples) with shaking. DHX9 was detected with monoclonal mouse antibody (for the rest of the detected proteins see 'Antibodies' section).

Immunofluorescence. Cells were crosslinked with 4% methanol-free formaldehyde in PBS at room temperature for 10 min and permeabilized with 0.1% Triton-X and 1% BSA in PBS for 30 min at room temperature. Primary antibodies (see details in the Antibody section) were diluted (1:1000 for the DHX9 pRb, 1:500 for the SC-35 (SRSF2) and J2 antibodies) in PBS with 0.1% Triton-X and 1% BSA and incubated with fixed cells at 4 °C for ~16 h. Fluorescently labelled secondary antibodies with the appropriate serotype were used reveal target proteins. Hoechst 33342 was used to stain DNA. Imaging was performed with a Leica SP5 confocal microscope.

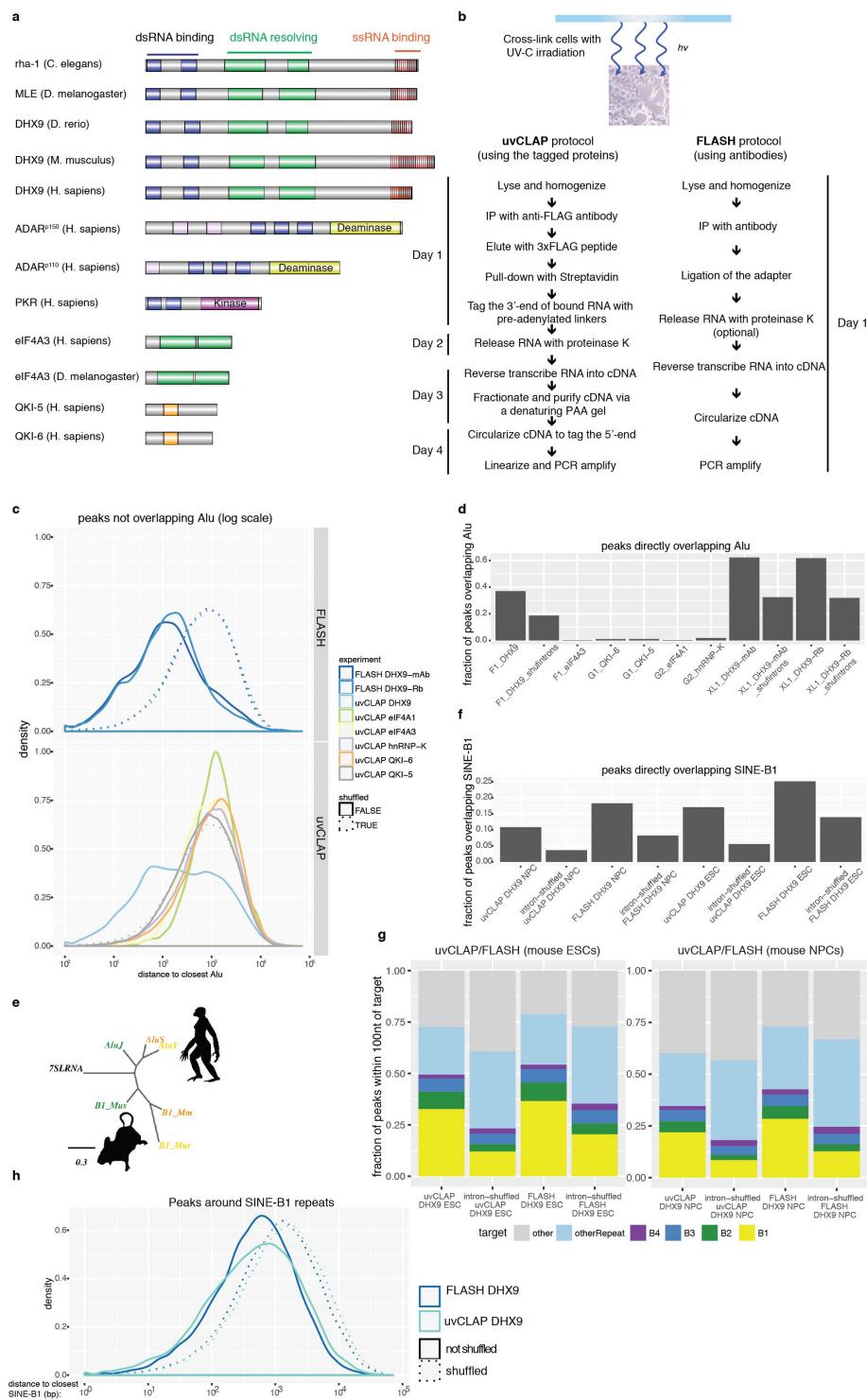
J2 (dsRNA) pull-down. Cells were harvested on the third day of siRNA treatment (siRNA transfections were carried out as described above) in RIP buffer (25 mM HEPES pH 7.2, 75 mM NaCl, 5 mM MgCl₂, 0.1% Igepal CA-630, 1 U μl^{-1} RNasin), sonicated (Bioruptor, 5 min, 30 s ON, 30 s OFF) and centrifuged at 20,000g for 20 min at 4 °C. Clarified lysates were incubated with 25 μl of ProtG Dynabeads pre-coupled to 2 μg of J2 mAb (Scicons) or mouse-IgG (Santa Cruz) for 3 h at 4 °C. The beads were washed three times with the RIP buffer and bound RNA was isolated using the RNA Clean and Concentrator columns (Zymo). Eluted RNA was reverse transcribed using SuperScript III with the following modifications:

RNA was incubated at 70°C for 3 min to remove secondary structures, which is followed by reverse transcription at 50°C for 15 min, 60°C for 10 min and 70°C for another 10 min. Quantitative real-time PCR was performed using the oligos listed in Supplementary Table 2 with FastStart SYBR Green Master (Roche 04 913 914 001), enrichments were calculated using the $2^{-\Delta\Delta C_t}$ method using 10% of input material as a reference for each immunoprecipitation.

Antibodies. DHX9 (monoclonal mouse, SantaCruz sc13183, polyclonal rabbit Abcam ab26271, monoclonal rabbit Abcam ab183731), ADAR (for the detection of human protein: Sigma SAB1401004, for the detection of mouse protein: Santa Cruz sc-73408), J2 antibody (Scicons, English and Scientific Consulting, Hungary), NONO (Sigma AV40715), ZFR (Sigma SAB2104153), HNRNPC (Santa Cruz sc-32308), KHDRBS1 (Sigma S9575), POL2 (4H8 Active Motif 101307, and 8WG16 Abcam ab817), CDK9 (Santa Cruz sc484), Flag-HRP (Sigma A8592), streptavidin-HRP (Pierce 21130), SPC24 (Bethyl, A304-260A), SRSF2/SC35 (Sigma S4045), MSL2 (Sigma HPA003413), MSL1 (polyclonal rabbit antibody generated by the Akhtar laboratory, previously described in ref. 32), MOF (Bethyl A300-992A).

Data accessibility. All the uvCLAP, FLASH and RNA-seq data in this study have been deposited to the Gene Expression Omnibus database under the accession number GSE85164.

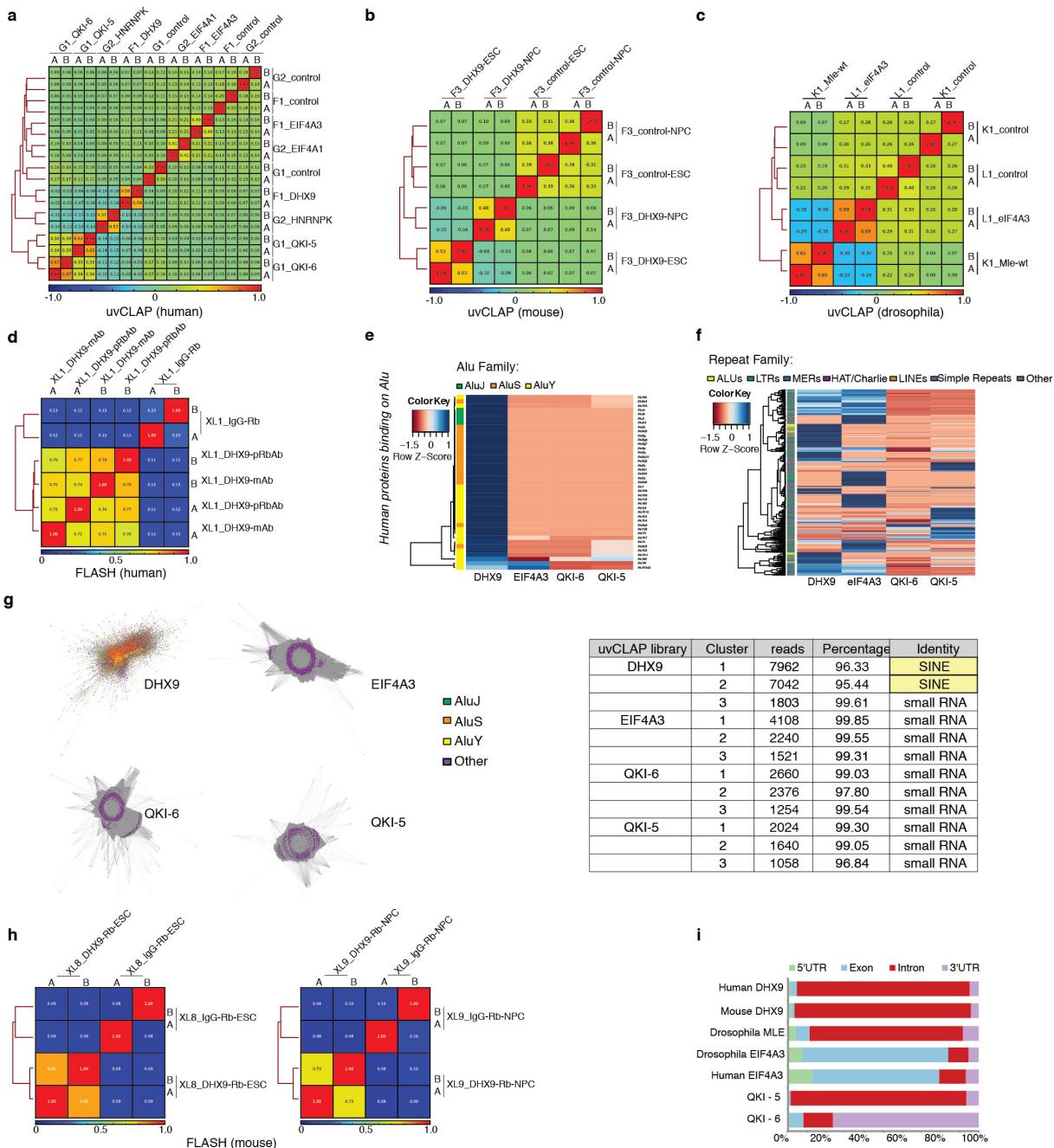
30. Shin, Y. J. *et al.* Nanobody-targeted E3-ubiquitin ligase complex degrades nuclear proteins. *Sci. Rep.* **5**, 14269 (2015).
31. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protocols* **8**, 2281–2308 (2013).
32. Chelmicki, T. *et al.* MOF-associated complexes ensure stem cell identity and Xist repression. *eLife* **3**, e02024 (2014).
33. Konig, J. *et al.* iCLIP–transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution. *J. Vis. Exp.* **50**, 2638 (2011).
34. Faircloth, B. C. & Glenn, T. C. Not all sequence tags are created equal: designing and validating sequence identification tags robust to indels. *PLoS One* **7**, e42543 (2012).
35. König, J. *et al.* iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.* **17**, 909–915 (2010).
36. Dodt, M., Roehr, J. T., Ahmed, R. & Dieterich, C. FLEXBAR—flexible barcode and adapter processing for next-generation sequencing platforms. *Biology (Basel)* **1**, 895–905 (2012).
37. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
38. Ibrahim, M. M., Lacadie, S. A. & Ohler, U. JAMM: a peak finder for joint analysis of NGS replicates. *Bioinformatics* **31**, 48–55 (2015).
39. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44** (W1), W160–W165 (2016).
40. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* **44** (W1), W3–W10 (2016).
41. Holmqvist, E. *et al.* Global RNA recognition patterns of post-transcriptional regulators Hfq and CsrA revealed by UV crosslinking *in vivo*. *EMBO J.* **35**, 991–1011 (2016).
42. Day, D. *et al.* Estimating enrichment of repetitive elements from high-throughput sequence data. *Genome Biol.* **11**, R69 (2010).
43. Novák, P., Neumann, P. & Macas, J. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics* **11**, 378 (2010).
44. Novák, P., Neumann, P., Pech, J., Steinhaisl, J. & Macas, J. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* **29**, 792–793 (2013).
45. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
46. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
47. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
48. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22**, 2008–2017 (2012).
49. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
50. Steigemann, P. *et al.* Aurora B-mediated abscission checkpoint protects against tetraploidization. *Cell* **136**, 473–484 (2009).
51. Picardi, E. *et al.* REDIporta: a comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Res.* **45**, D750–D757 (2017).



Extended Data Figure 1 | See next page for caption.

Extended Data Figure 1 | uvCLAP/FLASH identifies DHX9 targets in humans and mice. **a**, The domain organization of *Drosophila*, human and mouse Mle/DHX9 is shown side by side with the other RNA-binding proteins, EIF4A3, QKI-5/6 used in uvCLAP, and the two other enzymatic-domain-containing dsRNA-binding proteins PKR and ADAR. Blue boxes show dsRNA-binding domains, green boxes show helicase-N and helicase-C domains, orange boxes show KH domains, pink box shows the Z-DNA-interacting domains of ADAR. Orange boxes with vertical lines show the RGG-repeats at the C-term of DHX9/Mle/RHA proteins that interact with single-stranded nucleic acids. **b**, The schematic representation of the two methods (uvCLAP and FLASH) developed for the identification of *in vivo* targets of RNA-binding proteins. **c**, The distance between peaks which do not directly overlap with an *Alu* repeat to the nearest *Alu* element for RNA-binding proteins EIF4A3, EIF4A1, HNRNPK, QKI-5 and QKI-6 is significantly further than similar peaks for DHX9. uvCLAP and FLASH data are shown separately which are put together in Fig. 1c. **d**, The fraction of peaks which are directly overlapping with an *Alu* element is depicted for DHX9 in comparison to peaks

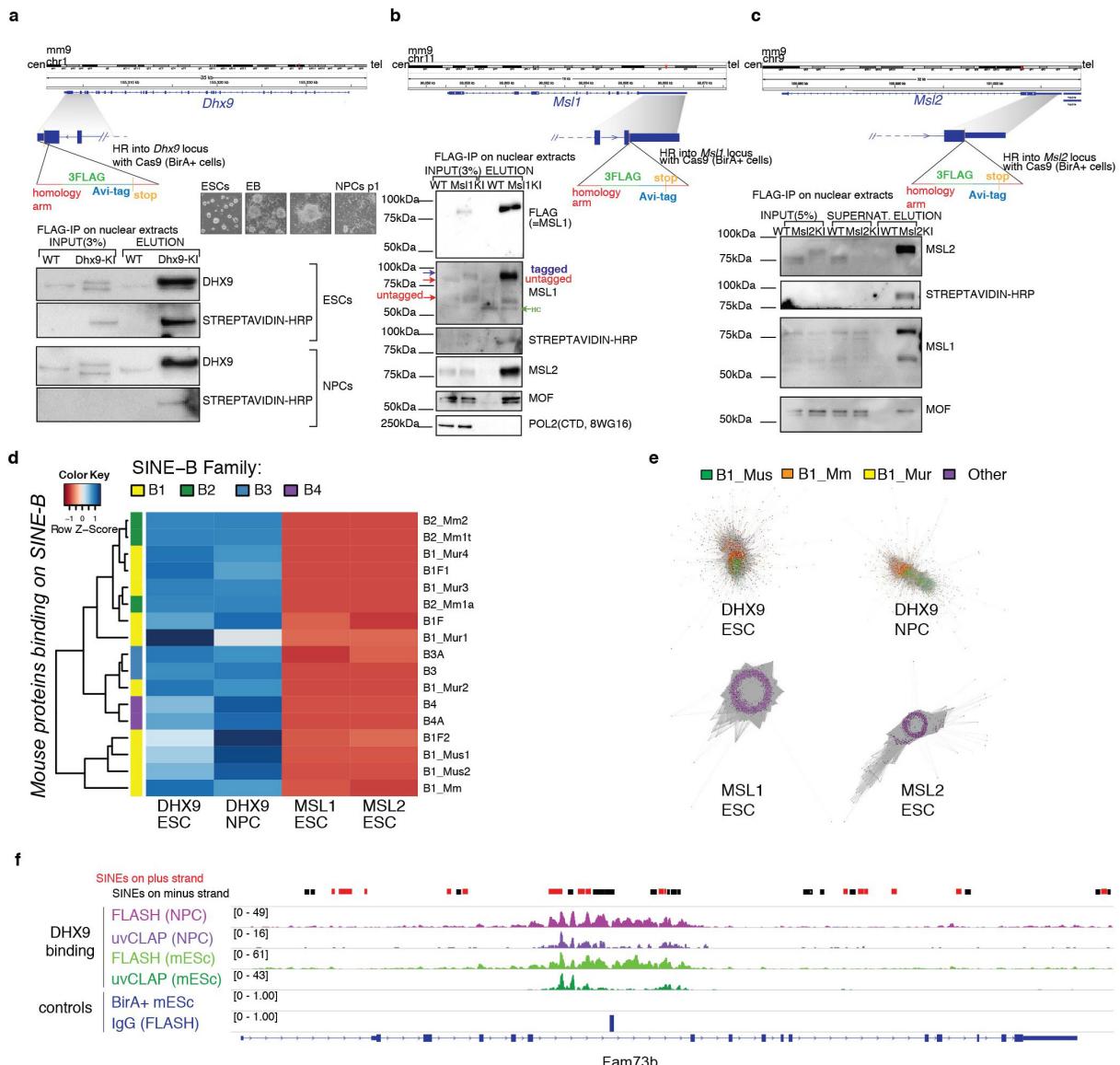
randomly placed in targeted regions within a sequencing library (shown as F1, G1, G2) and for EIF4A3, QKI-6, QKI-5, EIF4A1 and HNRNPK in uvCLAP and FLASH experiments (sequencing library XL1). **e**, The evolutionary divergence of primate and murine SINEs from the 7SL RNA. Scale bar, genetic distance (arbitrary units). **f**, The fraction of peaks which are directly overlapping with a SINE B1 repeat is depicted for DHX9 in comparison to peaks randomly placed in targeted regions. **g**, Fraction of uvCLAP and FLASH peaks in mouse embryonic stem (mES) cells and neural progenitor cells (NPCs) with a distance of at most 100 nucleotides from SINE B repeats is significantly enriched ($P < 2.2 \times 10^{-16}$; Fisher's exact test) for DHX9 in comparison to shuffled intron controls. **h**, The distances of DHX9 peaks in mES cells, which do not directly overlap with B1 repeats, to the nearest B1 repeats are shown. The distance is significantly smaller than the corresponding shuffled controls with median distances of 508 ($n = 33,686$) vs 1,312.5 ($n = 38,436$) nucleotides for uvCLAP and 514 ($n = 44,557$) vs 1,108.5 ($n = 51,272$) nucleotides for FLASH (all $P < 2.2 \times 10^{-16}$; one-tailed Mann–Whitney *U*-test). n = number of uvCLAP/FLASH peaks.



Extended Data Figure 2 | See next page for caption.

Extended Data Figure 2 | Reproducibility of the replicates for each uvCLAP/FLASH library and enrichment of *Alu* repeat binding in human cells. **a**, Replicates (shown as A and B) in human uvCLAP sequencing libraries F1, G1 and G2 are compared to each other using Spearman correlation and similarity is represented as a heat map. Except for replicates the only other two profiles which correlate are of QKI isoforms. **b**, Replicates (shown as A and B) in mouse uvCLAP sequencing library F3, which is composed of DHX9 knock-in mES cells and NPCs, and BirA ligase control cell line, are compared to each other using Spearman correlation and similarity is represented as a heat map. Replicates of each experiment cluster together. **c**, Replicates (shown as A and B) in *Drosophila melanogaster* uvCLAP sequencing libraries K1 and L1, which are composed of Mle and Eif4a3 profiles, are compared to each other using Spearman correlation and similarity is represented as a heat map. Replicates of each experiment cluster together. **d**, Replicates (shown as A and B) in human FLASH sequencing library XL1 are compared to each other by using Spearman correlation. DHX9 library replicates as well as different antibody used in two independent library replicates cluster together, whereas IgG control is separated. **e**, Enrichments of *Alu*

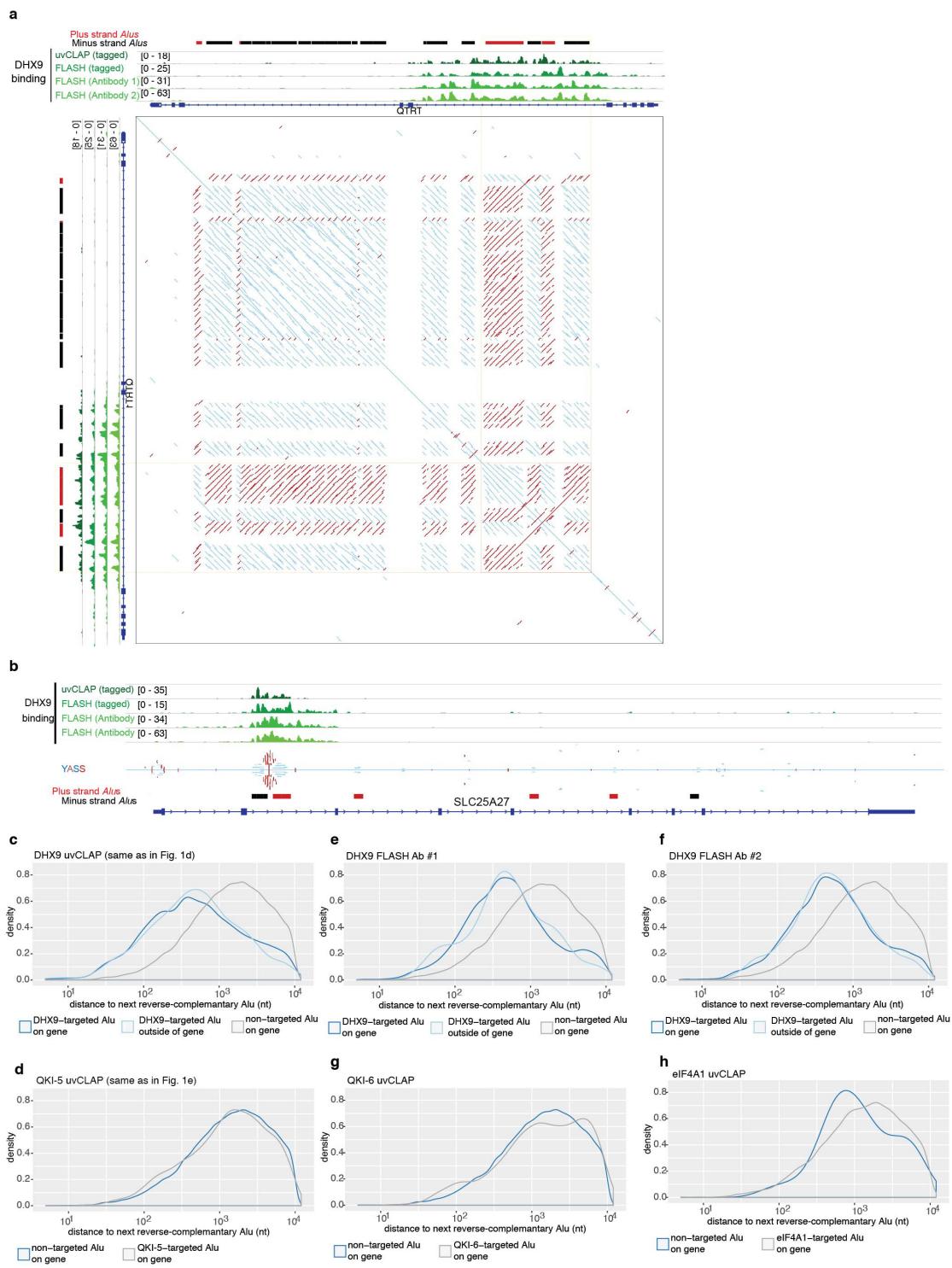
repeat subfamilies in uvCLAP data (see Methods). All subfamilies of *Alu* elements are highly enriched in the DHX9 uvCLAP experiment, compared to EIF4A3, QKI-5 and QKI-6. **f**, Enrichments of all human repeat families in uvCLAP data (see Methods). *Alu* elements are highly enriched in the DHX9 uvCLAP experiment, compared to EIF4A3, QKI-5 and QKI-6. **g**, Mapping-free clustering (see Methods) of repeated reads within uvCLAP libraries for human DHX9, EIF4A3, QKI-6 and QKI-5 reveals that only the top cluster of DHX9 library is composed of *Alu* elements. Right, a table view for the identity of the top three clusters (maximum reads) produced by graph based clustering of raw sequencing reads from different uvCLAP experiments. *Alu* elements bound by DHX9 form two distinct clusters (as depicted in the table) while the top three clusters in other samples (QKI-6, QKI-5 and eIF4A3) contain various small RNA families. **h**, Right, replicates (shown as A and B) in mouse FLASH sequencing library XL9 are compared to each other using Spearman correlation (see Supplementary Table 3 for mapping statistics and number of peaks). **i**, Binding frequencies of DHX9, EIF4A3 and QKI isoform crosslinking events on mRNA targets shown for UTRs, exons and introns.



Extended Data Figure 3 | Mapping for DHX9-bound RNAs onto repeats reveals an enrichment of SINE B binding in mouse cells. **a**, Top, endogenous knock-in strategy is shown for the mouse *Ddx9* gene. Right, differentiation snapshots of the tagged mES cell line into NPCs. Bottom, the identity of the tagged protein and its biotinylation is shown by DHX9 and streptavidin blots from Flag immunoprecipitation (IP) carried in nuclear extracts both for mES cells and NPCs. **b**, Endogenous knock-in strategy is shown for the mouse *Msl1* gene. The integrity of the protein and its biotinylation is shown by MSL1 and streptavidin blots from Flag IP carried in nuclear extracts. The interaction ability of the tagged MSL1 protein with its known interaction partners (within MSL complex) is validated by this Flag IP (performed on soluble nuclear extracts) by showing the co-IP for MOF (MYST1) and MSL1. **c**, Endogenous knock-in strategy is shown for the mouse *Msl2* gene. The integrity of the protein and its biotinylation is shown by MSL2 and streptavidin blots from Flag IP carried in nuclear

extracts. The interaction ability of the tagged MSL2 protein with its known interaction partners (within MSL complex) is validated by this Flag IP (performed on soluble nuclear extracts) by showing the co-IP for MOF (MYST1) and MSL2. **d**, Enrichments of SINE repeat subfamilies in uvCLAP data (see Methods). All SINE B subfamilies are highly enriched in DHX9 uvCLAP experiment, both in mES cells and NPCs, compared to MSL1 and MSL2. **e**, Mapping-free clustering of reads (see Methods) within uvCLAP libraries (mES cells and NPCs) reveals that only the top cluster of mouse DHX9 library is composed of SINE B repeats. **f**, Snapshots of uvCLAP and FLASH binding events (both in mES cells and NPCs) within *Fam73b* gene show that the crosslinking sites of DHX9 reside on SINE B repeats on opposing strands. SINEs on the plus strand are shown with red boxes, SINEs on the minus strand are shown with black boxes. Biological replicates were merged in these representations (also see Extended Data Fig. 2 for reproducibility between biological replicates).

RESEARCH LETTER

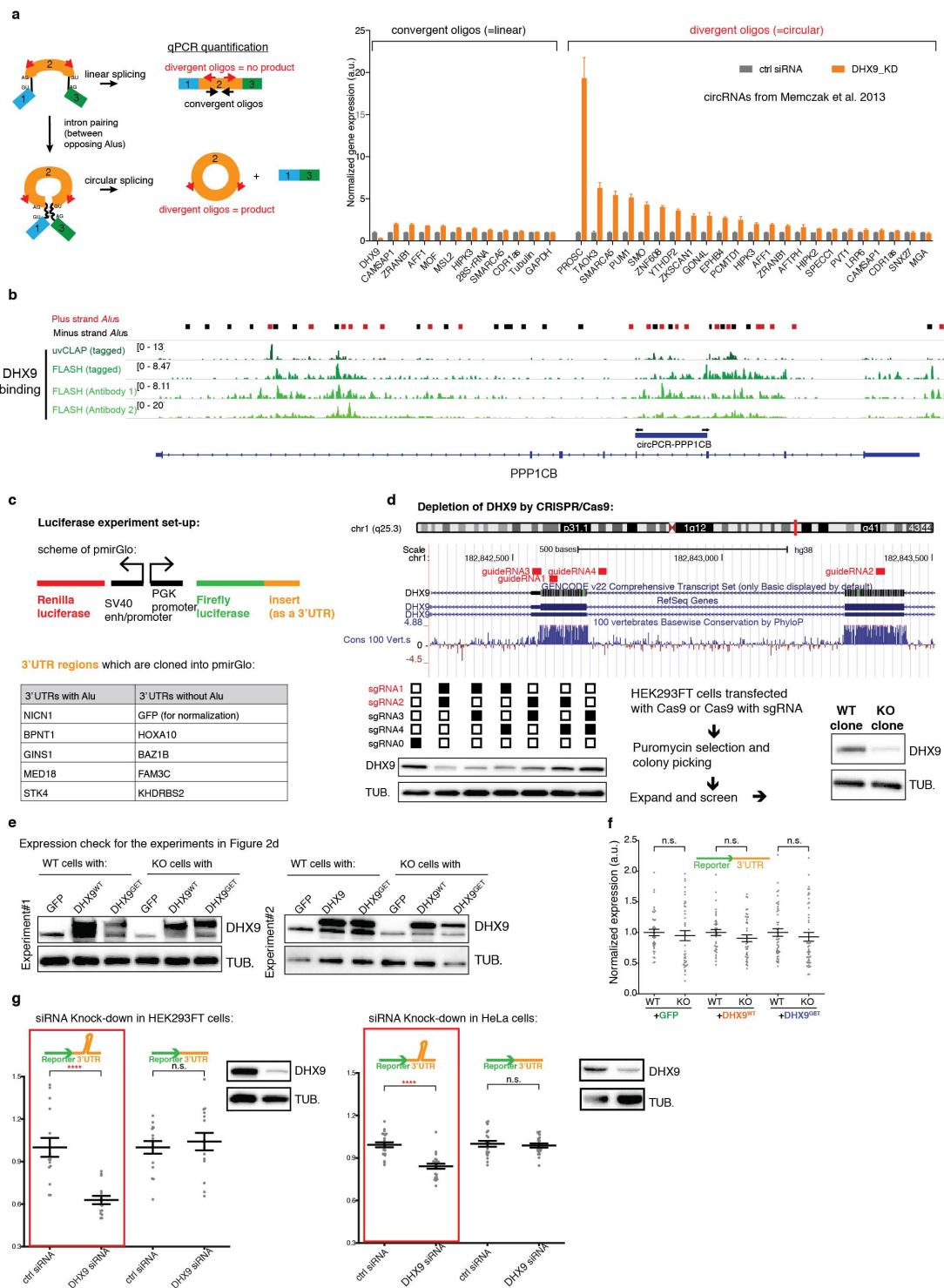


Extended Data Figure 4 | See next page for caption.

Extended Data Figure 4 | uvCLAP and FLASH reveal DHX9 binding on inverted-repeat *Alu* elements. **a**, A snapshot of DHX9 binding in the intron of the *QTRT1* gene. *Alu* elements are depicted with red (plus strand) or black (minus strand) boxes. Dot-plot generated by YASS (see Methods) showing that DHX9 binds to inverted repeats (structure-forming), shown by red lines, but not to direct repeats, shown by blue lines. **b**, A snapshot of DHX9 binding in the intron of the *SLC25A27* gene where there are *Alu* elements on opposite strands. YASS-generated dot-plots as in **a** (for simplicity only the diagonal of the YASS dot-plot is shown). **c**, Genome-wide analysis of uvCLAP DHX9 binding near paired *Alu* elements shows preferential targeting of *Alu* elements with nearby binding partners in both gene (dark blue) and outside gene (light blue) bound regions. **d**, Genome-wide analysis of uvCLAP QKI-5 binding near

paired *Alu* elements (binding enriched in introns) shows no significant difference for closeness to nearest potential binding partner. **e**, Genome-wide analysis of FLASH DHX9 binding near paired *Alu* elements (Ab#1) shows a preferential targeting of *Alu* elements with nearby binding partners in both gene (dark blue) and outside gene (light blue) bound regions. **f**, Genome-wide analysis of FLASH DHX9 (Ab#2) binding near paired *Alu* elements shows a preferential targeting of *Alu* elements with nearby binding partners in both gene (dark blue) and outside gene (light blue) bound regions. **g**, Genome-wide analysis of uvCLAP QKI-6 binding near paired *Alu* elements shows no significant difference for closeness to nearest potential binding partner. **h**, Genome-wide analysis of uvCLAP EIF4A1 binding near paired *Alu* elements shows no significant difference for closeness to nearest potential binding partner.

RESEARCH LETTER

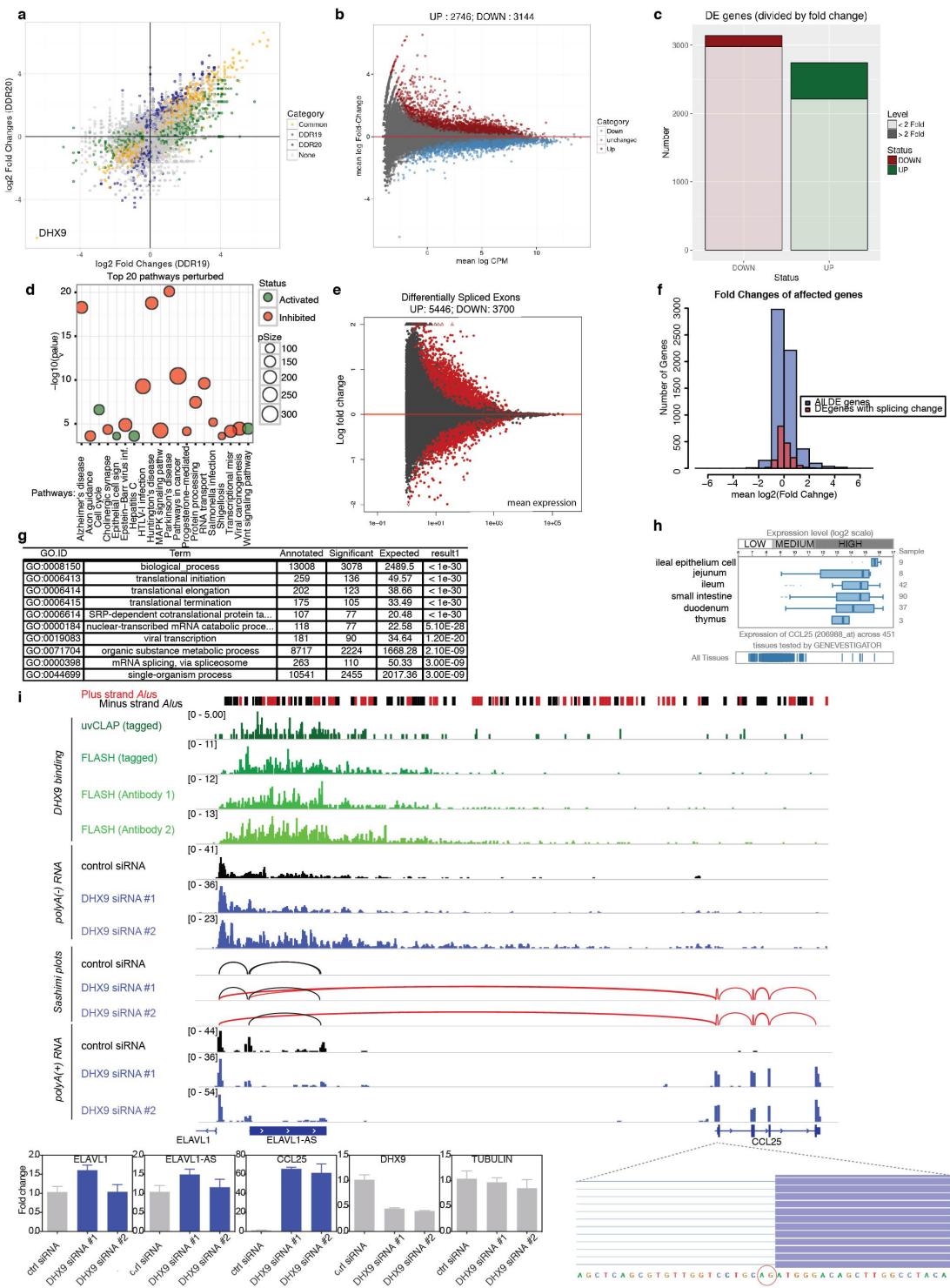


Extended Data Figure 5 | See next page for caption.

Extended Data Figure 5 | DHX9 depletion increases circular RNA formation and represses translation of reporters containing inverted-repeat *Alu* elements within the 3' UTR. **a**, Left, the qPCR approach for detecting linear/circular RNAs with two sets of oligos indicated as divergent and convergent used for the detection of circular and linear RNAs, respectively. Right, quantitative real-time PCR analysis of previously reported circRNAs²³ in DHX9-siRNA-treated HEK293FT cells (error bars represent standard deviation between biological quadruplicates). **b**, Snapshot of a circRNA generated locus, PPP1CB with DHX9 crosslinking sites on inverted *Alu* repeats. Blue bar depicted as 'circPCR-PPP1CB' represents the qPCR scored region in (Fig. 2c). Biological replicates were merged in these representations (also see Extended Data Fig. 2 for reproducibility between biological replicates). **c**, Top, schematic drawing of pmirGlo insert cloning is shown. Bottom, 3' UTRs used as inserts in luciferase assays. **d**, Top, genomic position of tested guide RNAs in CRISPR/Cas9 depletion of DHX9 is shown

(see Supplementary Table 2 for guide RNA sequences). Bottom left, efficiency of guide RNA pairs is shown by DHX9 and tubulin western blots. Bottom right, description for the making of the constitutive DHX9-depleted clone. **e**, Expression check by western blot analysis of rescue cells used on two independent (different days) experiments. **f**, Luciferase assays show that DHX9-depletion does not alter the expression of 3' UTR elements without inverted-repeat *Alu* elements. **g**, Luciferase assay results are shown carried in HEK293FT and HeLa cells upon siRNA knockdown of DHX9. Knockdown efficiency is validated by western blot analysis for DHX9 and tubulin. Similar to what is shown in Fig. 2d reporter expression from the constructs with an inverted-repeat *Alu* element in 3' UTRs are affected upon DHX9 depletion. Error bars represent standard deviation of a total of 20 data points that come from one experiment carried out with biological quadruplicates for each cloned 3' UTR insert (5 with and 5 without inverted-repeat *Alu* elements).

RESEARCH LETTER

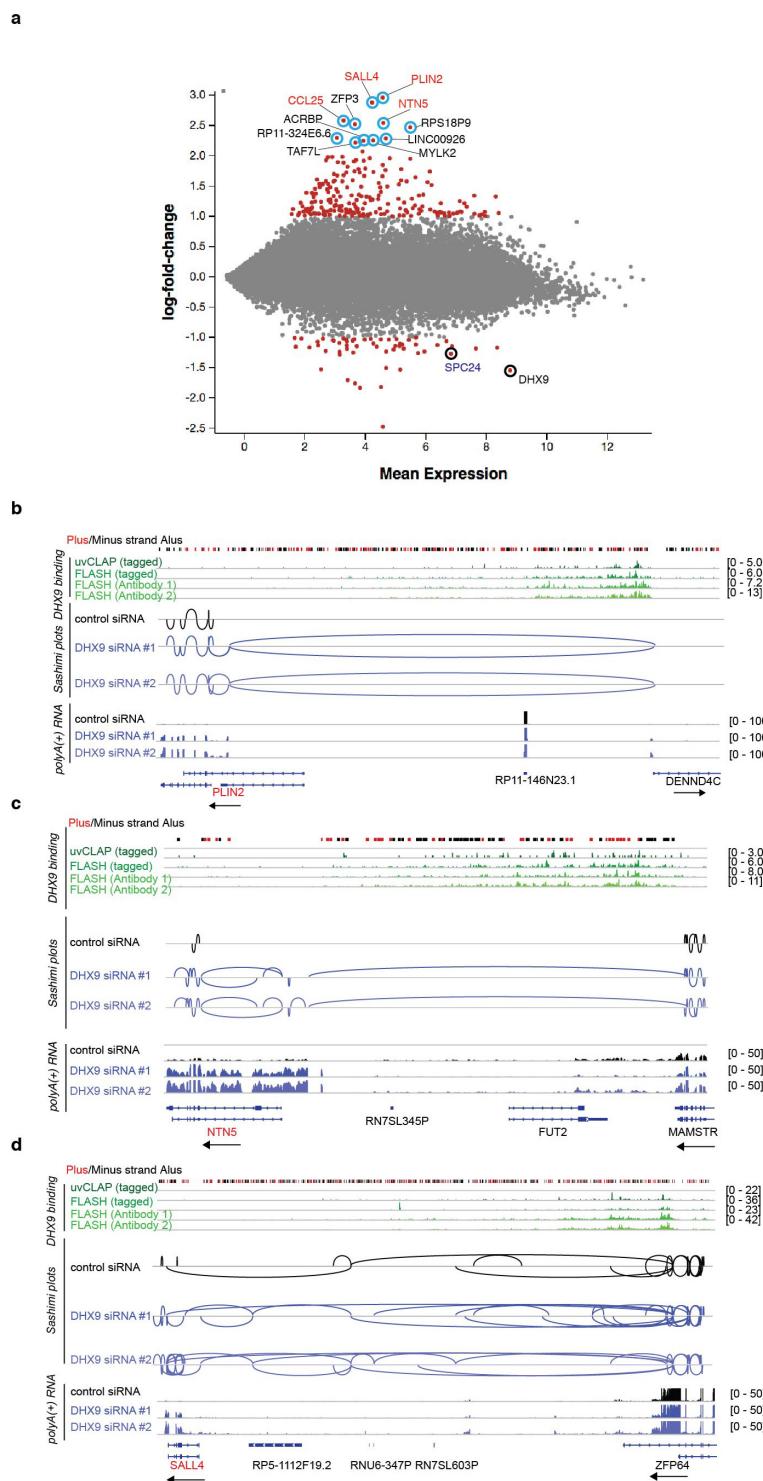


Extended Data Figure 6 | See next page for caption.

Extended Data Figure 6 | DHX9 depletion leads to changes in gene expression and exon usage. **a**, Reproducibility of gene-expression changes between cells treated with siRNA 1 (green) and siRNA 2 (blue) (both show that DHX9 is the most severely downregulated gene in these datasets). Genes with a significant change in their expression levels in both experiments are shown with yellow points. **b**, MA plot of genes that show a reproducible change in both RNAi experiments (yellow dots in **a**). **c**, The extent of gene expression changes in the reproducibly misregulated genes in **b**. **d**, Pathway perturbation analysis using the data in **b**. **e**, Differentially spliced exons that show a reproducible change in DHX9-depletion experiments (siRNA 1 and siRNA 2). **f**, Differentially expressed genes (DEgenes) with respect to their detected splicing changes shows 55% of the mis-spliced genes are downregulated. **g**, GO term analysis using data in **e**. **h**, Tissue expression database (Genevestigator) results for *CCL25* gene show its expression is specific to the intestine and thymus tissues. *CCL25* is expressed at low levels every other tissue (in total 451 tissues). **i**, DHX9

binding in the *ELAVL1-CCL25* locus is shown together with the poly(A)⁻ and poly(A)⁺ RNA-seq data. Sashimi plots generated from control and DHX9-siRNA-treated samples show a new exon–exon junction from *ELAVL1^{AS}* to *CCL25* (depicted with red lines). For clarity, only plus strand data are shown. Biological replicates were merged in these representations (also see Extended Data Fig. 2 for reproducibility between biological replicates). Poly(A)⁺ RNA-seq data show that an anti-sense transcript in the opposite direction to *ELAVL1* is now connected via a cryptic splice acceptor site to the first exon of *CCL25* upon DHX9 depletion (two independent siRNAs, four biological replicates each). Bottom, qPCR validation of the *CCL25* upregulation upon RNA processing defects in DHX9 knockdown samples. We observe that *CCL25* is ~60-fold upregulated in DHX9-depleted cells, while the expression of neither *ELAVL1* nor its accompanying anti-sense transcript change significantly. The cryptic splice acceptor site is enlarged on the bottom right.

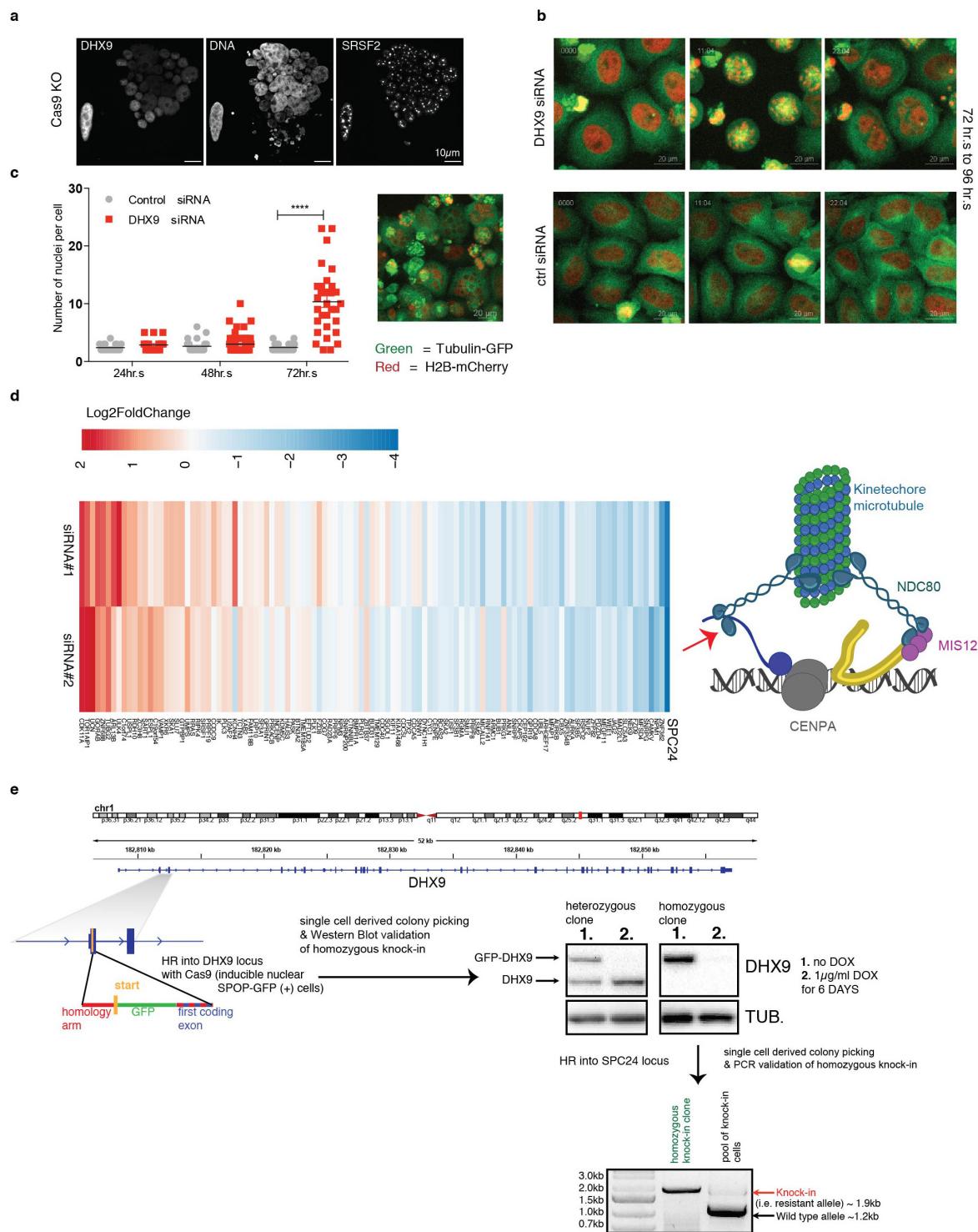
RESEARCH LETTER



Extended Data Figure 7 | See next page for caption.

Extended Data Figure 7 | Many upregulated genes in DHX9 knockdown show RNA processing defects in their gene locus. **a**, Genes that were more than twofold up- or downregulated are marked with red dots. Top 11 upregulated genes, which show signs of transcription bleed-through from upstream genes are highlighted with blue circles. *CCL25*, *SALL4*, *PLIN2* and *NTN5* are further highlighted. **b**, DHX9 binding in the *PLIN2-RP11-146N23.1* gene locus on the nascent RNA originating from the antisense transcription of the *DENND4C* promoter and splicing defect upon DHX9 knockdown. Sashimi plots depicting the exon junctions of these genes are shown in blue for the DHX9 knockdown, whereas the control sample (shown with black lines) do not display such a joining event (threshold is set to one read). Biological replicates were merged in these representations (also see Extended Data Fig. 2 for reproducibility between biological replicates). For clarity, only minus strand data are shown. **c**, DHX9 binding in the *NTN5-MAMSTR* gene locus on the nascent

(bleed-through) RNA and splicing defect upon DHX9 knockdown. Sashimi plots depicting the exon junctions of these genes are shown in blue for the DHX9 knockdown, whereas the control sample (shown with black lines) do not display such a joining event (threshold is set to one read). Biological replicates were merged in these representations (also see Extended Data Fig. 2 for reproducibility between biological replicates). For clarity, only minus strand data are shown. **d**, DHX9 binding in the *SALL4-ZFP64* gene locus on the nascent (bleed-through) RNA and splicing defect upon DHX9 knockdown. Sashimi plots depicting the exon junctions of these genes are shown in blue for the DHX9 knockdown whereas the control sample (shown with black lines) displays a lower frequency of such joining events (threshold is set to 1 reads). Biological replicates were merged in these representations (also see Extended Data Fig. 2 for reproducibility between biological replicates). For clarity, only minus strand data are shown.

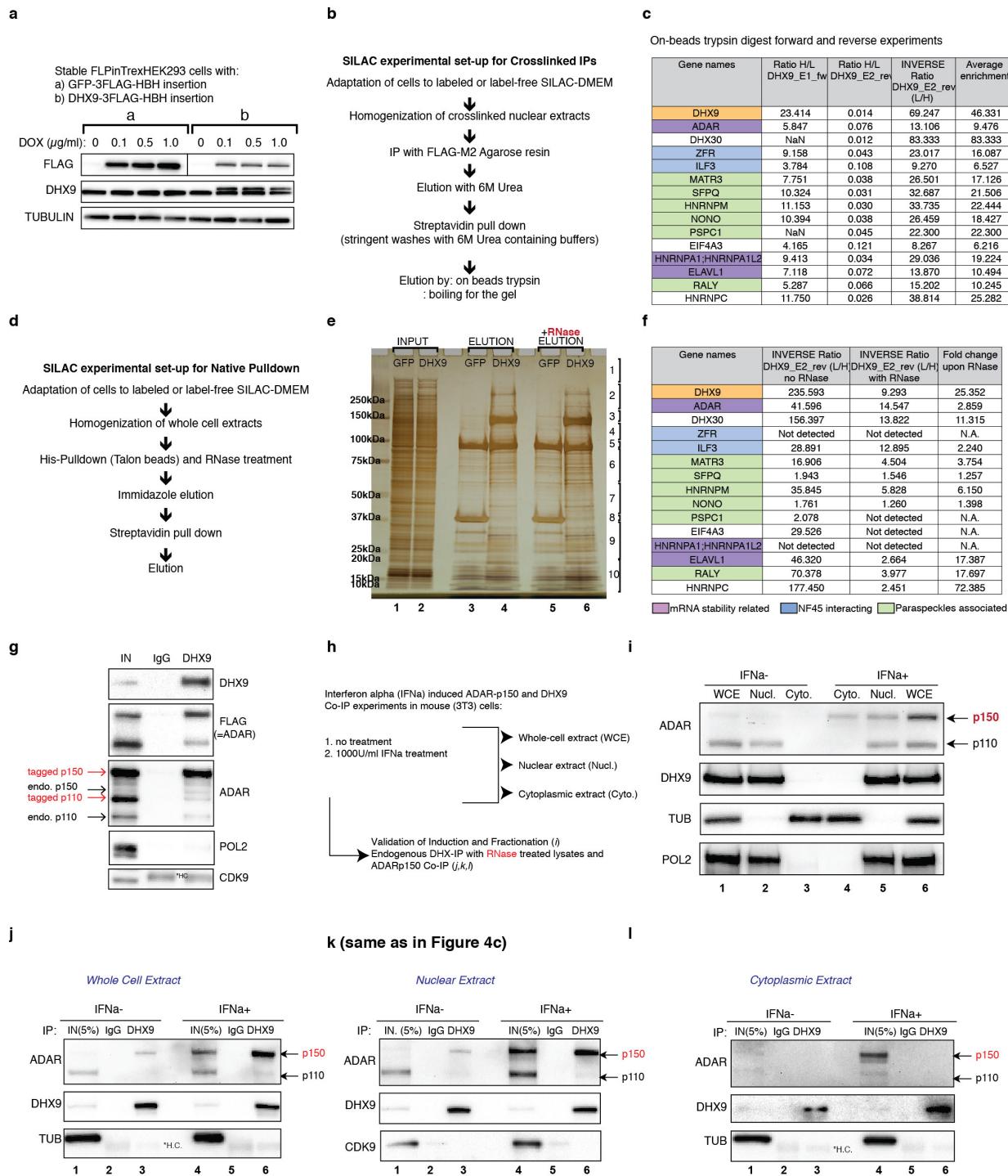
RESEARCH LETTER


Extended Data Figure 8 | See next page for caption.

Extended Data Figure 8 | DHX9 depletion leads to a catastrophic disintegration of the nucleus. **a**, CRISPR–Cas9 DHX9-knockout cell (on the right side of the image) displays a multinucleated cell phenotype in comparison to a DHX9-expressing cell (on the left side of the image). Scale bar, 10 μ m. **b**, Three still images from Supplementary Video 1 (control) and Supplementary Video 3 (knockdown) representing the start, middle and end point of 22-h-long live imaging for control or DHX9-siRNA-treated HeLa H2B–mCherry, tubulin–GFP cells. **c**, Quantification of the number of nuclei at the end point of live cell imaging (\sim 24 h imaging duration). Time points on the graph represent the start point of the imaging. All the imaged fields are taken into consideration and only the cells with more than one nucleus are included in the analysis. Statistical analysis performed with Kruskal–Wallis test shows that only the 72–96 h time point of DHX9-knockdown cells significantly differs from the rest with a $P < 0.0001$ (Kruskal–Wallis). Also see Supplementary Videos 2 and 7 for full field views of imaging and Supplementary Videos 5 and 6 for a

second siRNA knockdown of DHX9. Right, multinucleated cells at the end point of imaging (at 96 h of DHX9 knockdown) in a different field from Supplementary Video 7. **d**, Left, expression levels for the Mitochondria Consortium Grape phenotype category genes (117 are expressed in the RNA-seq out of the 153 in total, mean counts per million > 0). Right, schematic view of the Ndc80 complex, red arrow points at the SPC24 protein. **e**, Left, experimental design of the homologous recombination of the GFP tag at the DHX9 locus is shown. Single-cell-derived heterozygously or homozygously GFP-tagged clones successfully deplete the GFP-DHX9 protein upon doxycycline induction of SPOP-GFP. Right, PCR-based screening of the knock-in allele is shown for pool of cells before the colony picking and the homozygous knock-in *Alu*-bypass SPC24 clone after isolation. Knock-in causes a size shift of the PCR product (from 1.2 to 1.9 kb). Correct insertion of the repair construct into the endogenous locus is further validated by Sanger sequencing (not shown).

RESEARCH LETTER

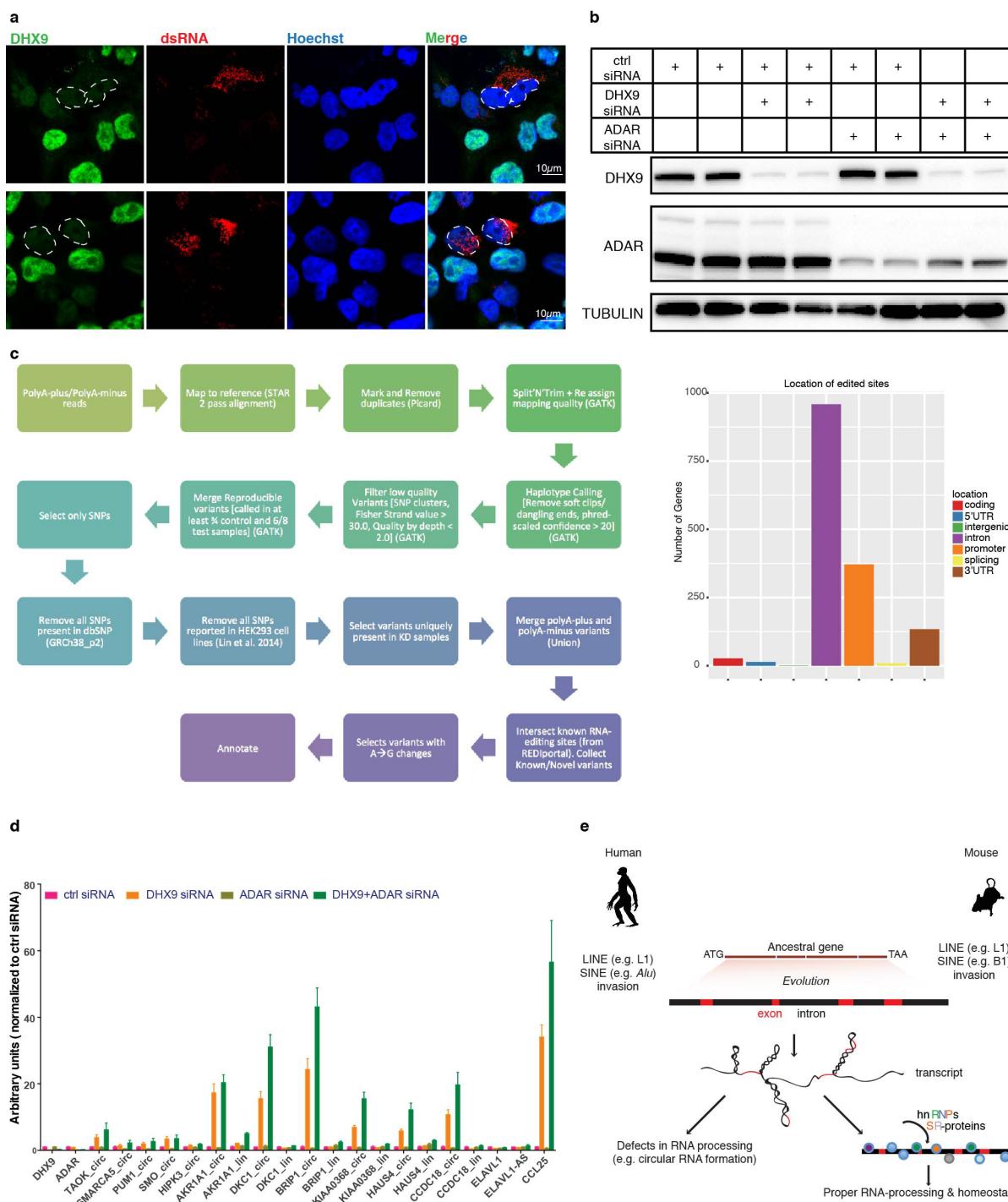


Extended Data Figure 9 | See next page for caption.

Extended Data Figure 9 | Tandem affinity purification of DHX9 from crosslinked nuclear extracts or native whole-cell extracts identifies a set of RNA-binding proteins that interact with DHX9.

a, Induction test for the stable FLPinTrex HEK293 cell line with a single-copy C-terminal $3 \times$ Flag– $6 \times$ His–biotin– $6 \times$ His-tagged GFP or DHX9 protein. **b**, Experimental set-up for the SILAC quantified formaldehyde crosslinked DHX9 interactome (either loaded on gel or eluted from beads by trypsinization) obtained by tandem affinity purification. **c**, Fold enrichment values of on bead digested proteins from forward and reverse SILAC experiments (forward: DHX9 cells are grown in heavy isotope labelled SILAC medium and GFP cells are grown in unlabelled medium; reverse: DHX9 cells are grown in unlabelled medium and GFP cells are grown in heavy isotope labelled SILAC medium) for those proteins highlighted in Fig. 4a. **d**, Experimental set-up for the affinity tag pull-downs employing native whole-cell extracts that were used for in-gel digestion LC–MS. **e**, Silver-stain gel of native pull-down experiments, where lanes 1 and 2 show the input from GFP and DHX9 cell lines (DHX9 cells are grown in unlabelled medium and GFP cells are grown in heavy

isotope labelled SILAC medium), lanes 3 and 4 are the final eluates of His-tag pull-down followed by streptavidin pull-downs from GFP and DHX9 cell lines, and lanes 5 and 6 are RNaseA-treated final eluates. The gel is cut into 10 slices as shown and both GFP and DHX9 lanes are combined (3 and 4 together; 5 and 6 together) prior to mass-spectrometry analysis. **f**, Fold enrichment values in the native purifications for the proteins highlighted in Fig. 4a, showing a depletion for most of these proteins upon RNase digestion. **g**, Validation of DHX9 interaction with ADAR is shown in a stable cell line for ADAR that can express both p110 and p150 isoforms. The doublet is detected with a blot for Flag after DHX9 IP in this stable cell line. POL2 and CDK9 blots serve as loading controls. **h**, The description of the experimental set-up for IFN α induction and DHX9–ADAR co-IP in different fractions of the mouse 3T3 cell line. **i**, Validation of fractionation with tubulin (lanes 3, 4) and RNA–POL2 blots (lanes 2, 5) and IFN α induction of ADAR(p150) (lanes 5 and 6). **j–l**, Mouse DHX9 interacts with ADAR(p150) in the whole cell (**j**, lanes 3 and 6) and nuclear extracts (**k**, lanes 3 and 6) but not in the cytoplasmic extracts (**l**, lanes 3 and 6).

RESEARCH LETTER


Extended Data Figure 10 | See next page for caption.

Extended Data Figure 10 | DHX9 and ADAR cooperatively work on large RNA secondary structures. **a**, Anti-DHX9 and J2 (dsRNA) antibody staining are shown for a mixed population of cells treated with control siRNA (DHX9 staining present, green) and DHX9 siRNA (DHX9 staining absent, for example, marked nuclei). DHX9-depleted cells accumulate dsRNA. **b**, The efficiency of knockdown in replicate experiments for J2 pull-down experiment displayed in Fig. 4e are shown by western blot. **c**, Left, analysis pipeline for detecting differential editing between control and knockdown RNA-seq libraries (both poly(A)⁺ and poly(A)⁻ considered). Right, distribution of differential A-to-I editing sites detected in DHX9 knockdown is shown. After filtering for known SNPs from dbSNP and HEK293 genome, 1,244 genes with 1,807 potential A-to-I RNA editing changes in DHX9-siRNA-treated samples were detected which were absent in controls. 48% of these sites have been reported previously⁵¹. 77% of genes showing editing changes have these changes within introns. Additionally, we observed that 21% of genes with intronic editing also showed splicing defects and 28% of genes with overall editing showed misregulation (differentially expressed) upon DHX9 knockdown. These genes are a small fraction of all the genes with splicing defects (4.5%) or expression changes (6%) observed in DHX9-depleted cells. **d**, Effect of

DHX9, ADAR or double knockdown of DHX9 and ADAR on circular RNA generation and CCL25 upregulation is shown. ADAR depletion alone does not have an appreciable impact on circular RNA generation, whereas the effect of DHX9 depletion is augmented when it is combined with ADAR depletion. **e**, A summary model. *Alu* elements are the most abundant transposable elements making up >10% of our own genomes. These elements are potentially harmful to their host and are depleted entirely from 5' UTRs of genes and developmentally important loci, such as the HOXA-D clusters but are found in abundance in intronic and intergenic regions. We show that DHX9, a highly conserved, very abundant nuclear RNA helicase, directly interacts with *Alu* elements *in vivo* and suppresses circular RNA formation, which is probably a symptom of *Alu*-mediated splicing defects. We propose that our cells have developed a dependency on DHX9 to remove strong secondary structures, such as the ones originating from *Alu* insertions in the transcribed parts of our genome. We also show that in mice, DHX9 interacts with SINE B elements that are the murine equivalent of *Alu* elements, underscoring the flexibility of DHX9-mediated control of retrotransposon toxicity throughout evolution.

A.5 Update of the deepTools toolkit for exploring deep-sequencing data

I contributed to the development of deepTools2 (led by Fidel Ramirez and Devon Ryan) through features and bugfixes, and to the testing and update of the documentation and the galaxy server. I helped with the writing and revision of the manuscript by Fidel Ramirez and other authors.

deepTools2: a next generation web server for deep-sequencing data analysis

Fidel Ramírez^{1,†}, Devon P Ryan^{1,†}, Björn Grüning², Vivek Bhardwaj^{1,3}, Fabian Kilpert¹, Andreas S Richter¹, Steffen Heyne¹, Friederike Dündar⁴ and Thomas Manke^{1,*}

¹Max Planck Institute of Immunobiology and Epigenetics, 79108 Freiburg, Germany, ²University of Freiburg, Department of Computer Science, 79110 Freiburg, Germany, ³Faculty of Biology, University of Freiburg, 79104 Freiburg, Germany and ⁴Weill Cornell Medical College, Applied Bioinformatics Core, Department of Physiology and Biophysics, New York, NY 10065, USA

Received February 02, 2016; Revised March 22, 2016; Accepted April 02, 2016

ABSTRACT

We present an update to our Galaxy-based web server for processing and visualizing deeply sequenced data. Its core tool set, deepTools, allows users to perform complete bioinformatic workflows ranging from quality controls and normalizations of aligned reads to integrative analyses, including clustering and visualization approaches. Since we first described our deepTools Galaxy server in 2014, we have implemented new solutions for many requests from the community and our users. Here, we introduce significant enhancements and new tools to further improve data visualization and interpretation. deepTools continue to be open to all users and freely available as a web service at deepTools.ie-freiburg.mpg.de. The new deepTools2 suite can be easily deployed within any Galaxy framework via the toolshed repository, and we also provide source code for command line usage under Linux and Mac OS X. A public and documented API for access to deepTools functionality is also available.

INTRODUCTION

The analysis of data from high-throughput DNA sequencing experiments continues to be a major challenge for many researchers. The rapidly increasing diversity of experimental assays using high-throughput sequencing has led to a concomitant increase in the number of analysis packages that allow for insightful visualization and downstream analyses (e.g. ChAsE (1), the ChIP-seq web server (<http://ccg.vital-it.ch/chipseq>), Genomation (2), Homer (3), ngs.plot (4)). Many of these tools require command line experience and input data, which is already quality controlled and properly normalized. As a result, many research groups still

do not have the capacity to process and analyse the data generated with deep sequencing technologies. With this in mind, we developed deepTools, a modular suite of fast and user-friendly tools that we implemented within a publicly accessible Galaxy instance (5). deepTools support a wide range of functions, such as various quality controls, different normalization schemes and genome-wide visualizations. Since the original publication of deepTools, and motivated by frequent requests from biologists and bioinformaticians, we have published 10 releases with numerous performance improvements and expansions of scope. In the process, we rewrote large sections of the code, revised the documentation, and updated our server hardware. We added new ways to process and filter deep sequencing data, provided new tools for quality control and analysis, and updated our documentation and examples. Moreover, we have significantly increased the computational speed (sometimes up to 100-fold), extended automated testing for most components, and created an API allowing others to seamlessly incorporate deepTools' functionality into their own programs. Here, we present our latest release (deepTools2) and its associated web server, which is based on the well-known Galaxy platform (6). This ensures that users lacking access to specialized resources and servers can still benefit from a simple and flexible framework for reproducible analysis.

POWERFUL WEB SERVER FOR DEEP SEQUENCING ANALYSES

Our deepTools Galaxy server is freely available at <http://deepTools.ie-freiburg.mpg.de>. The deepTools suite supports four common tasks: (i) quality control, (ii) data processing and normalization, (iii) data integration and (iv) visualization (see Table 1). The package has been designed to take as input files some of the most established formats in the deep sequencing field, such as BAM for aligned reads, bigWig for scores (e.g. normalized read coverages) associated with genomic regions, and BED for coordinates of genome

*To whom correspondence should be addressed. Tel: +49 761 5108738; Fax: +49 761 5108 80738; Email: manke@ie-freiburg.mpg.de

†These authors contributed equally to the paper as first authors.

regions. All tools come with a large number of options that can be used to fine-tune analyses or filter input datasets for faster data exploration. The visualizations and output files encompass highly customizable, publication-ready images as well as genome-wide scores in bigWig or bedGraph format and tab-separated summaries, e.g. of pairwise correlation metrics that can readily be used for further downstream analyses.

Users can upload data from their computers, e.g. via FTP, and can download additional files from the UCSC table browser (7). In addition, we have added more commonly used reference genomes and annotations to the web server's data library. Moreover, comprehensive test data is provided to allow users to easily explore the server's functionality without the need to upload their own data. Registered users can share data, histories and workflows with collaborators and reviewers. To support the increasing number of users, we have made substantial upgrades to our hardware, both in terms of the number of CPUs and available storage.

ADDITIONAL QUALITY CONTROLS

The first step of all analyses is quality control. In addition to the previously established GC-bias detection (*computeGCbias*) and the ChIP-specific assessment of enrichment (*plotFingerprint*), we have expanded the diagnostic capabilities of deepTools (Figure 1A). First, we have added a new program, *plotCoverage*, to inspect the genome-wide distribution of fragment coverage for several samples. This is crucial information for deciding whether the basic goals of the experimental design were met, particularly regarding the desired sequencing depth. For paired-end data, the new program *bamPEFragmentSize* provides a very sensitive quality check of whether the size distribution of sequenced fragments corresponds to the expectations based on the library preparation. The average fragment size of a sample is also an important parameter for many downstream analyses, such as peak calling (8,9).

Finally, we have enhanced the capacity for comparative quality control and replicate analysis over multiple samples. Users can now perform principal component analysis (PCA) of multiple samples and generate the corresponding plots with the program *plotPCA*. This tool takes as input a matrix, in which each column represents a sample-specific vector of fragment counts or any other genome-wide score. It can unveil unexpected patterns, such as batch effects, outliers or sample swaps. In a similar vein, *plotCorrelation* has been significantly updated to fine-tune the joint analysis of samples and visualization of the correlation structure. It is now also able to produce scatterplots. Previously, the correlation analysis was performed together with the data collection from BAM files by one tool, *bamCorrelate*. To improve both speed and flexibility, we have replaced this tool and separated the computationally demanding data retrieval from the actual visualization tasks (*plotCorrelation* and *plotPCA*). The data collection step can now be done using *multiBamSummary* or *multiBigwigSummary*, as described below.

ENHANCED PROCESSING FLEXIBILITY, SCOPE AND SPEED

Once the general quality controls have been carried out, researchers are faced with the actual processing of data, such as filtering, normalization and format conversions. Due to the sheer size of the aligned reads files, this is a non-trivial and cumbersome aspect of all deep sequencing workflows. Our web server supports this challenging part in two ways: (i) the Galaxy environment automatically keeps track of every analysis step, so that users can always and indefinitely identify the tools and parameters that were used to generate a file; (ii) deepTools can perform combined filtering, normalization and format conversion in a single framework, which has contributed to the appeal of deepTools even outside the web server. An important enhancement in this regard is the general support of all deepTools components for flexible filtering of alignments files (BAM) based on the SAM flag (10), in addition to the optional exclusion of duplicates and reads with low alignment scores. This is designed to prevent accumulation of large and unnecessary intermediate files that often occur when different filtering strategies are compared.

We have also enhanced the capability of deepTools to recognize and process new sequencing read types. For example, deepTools now parses CIGAR strings and spliced-read alignments. This is particularly important for *bamCoverage*, which can now properly handle spliced reads from strand-specific RNA-seq data and convert them into meaningful coverage tracks (Figure 1B). The same tool was also enhanced to accommodate MNase-seq data, which results in very high-resolution summaries for nucleosome positions.

Frequently, genome-wide data is available from large consortia and data portals, such as IHEC (<http://ihec-epigenomes.org>), ENCODE (<https://www.encodeproject.org>) and BLUEPRINT (<http://www.blueprint-epigenome.eu>). In most cases, the provided data has already been processed and normalized. Since this information is most commonly stored as bigWig files, most deepTools components now have the capacity to handle data stored in bigWig format. Importantly, the processing of bigWig files was sped up by up to 100-fold by using the new Python *pyBigWig* package (<http://dx.doi.org/10.5281/zenodo.45238>) and the *libBigWig* library written in C (<http://dx.doi.org/10.5281/zenodo.45278>), which were developed simultaneously as side projects. The command line versions of the tools additionally allow users to directly use remotely stored files without downloading them beforehand.

IMPROVED INTEGRATIVE ANALYSES AND VISUALIZATIONS

Following the normalization and aggregation of aligned reads, visualization of the data is a vital part of almost every bioinformatics analysis as it allows users to explore their data and generate hypotheses. In addition, downstream analyses and data interpretation often depend on the integration of multiple samples. One of the major changes for deepTools2 is the capability to process numerous signal (bigWig) and region files (BED) in a joint analysis, particularly for summarizing fragment coverages or other scores

Table 1. Typical applications of deepTools components and a summary of their main inputs and outputs

tool	application	input files	main output file(s)
quality control			
<i>plotCorrelation</i>	compute and visualize correlations between multiple samples	output from <i>multiBamSummary</i> or <i>multiBigwigSummary</i>	heatmap of correlation coefficients, pairwise scatterplots
<i>plotPCA</i>	compute and visualize the principal component analysis	output from <i>multiBamSummary</i> or <i>multiBigwigSummary</i>	PCA plot, scree plot
<i>plotFingerprint</i> (prev.: <i>bamFingerprint</i>)	assess enrichment strength of a ChIP-seq experiment	2 or more BAM	diagnostic plot
<i>computeGCBias</i>	calculate the expected and observed GC distribution of reads	1 BAM and one 2bit, optional: 1 BED	diagnostic plots, tabular output for <i>correctGCBias</i>
<i>plotCoverage</i>	compute the coverage distribution	1 or more BAM	diagnostic plot and tabular output
<i>bamPEFragmentSize</i>	compute distribution of fragment lengths for paired-end alignments	1 BAM	distribution plot and summary statistics
data processing and normalization			
<i>multiBamSummary</i>	count the number of overlapping reads per bin or genomic region across multiple samples	2 or more BAM, optional: 1 BED	compressed matrix data
<i>multiBigwigSummary</i>	calculate binned, genome-wide scores across multiple samples	2 or more bigWig, optional: 1 BED	compressed matrix data
<i>bamCoverage</i>	obtain the sequencing depth normalized read coverage of a single sample	1 BAM	bedGraph or bigWig
<i>bamCompare</i>	normalize the read coverage of 2 samples with a specific operation (e.g. log2ratio, difference)	2 BAM	bedGraph or bigWig
<i>correctGCBias</i>	obtain a BAM file with reads distributed according to the genome's GC content	1 BAM and tabular output from <i>computeGCBias</i>	GC bias-corrected BAM
<i>computeMatrix</i>	compute distribution of scores over aligned genomic regions	1 or more bigWig, 1 or more BED	compressed matrix data
heatmaps and summary plots			
<i>plotHeatmap</i> (prev.: <i>heatmapper</i>)	visualize pre-computed scores per genomic region	<i>computeMatrix</i> output	heatmap plot
<i>plotProfile</i> (prev.: <i>profiler</i>)	visualize score summaries over groups of genomic regions	<i>computeMatrix</i> output	summary plot ("meta-profile")
<i>plotCoverage</i>	assess the sequencing depth by way of calculating the frequencies of read coverages	1 or more BAM	diagnostic plots

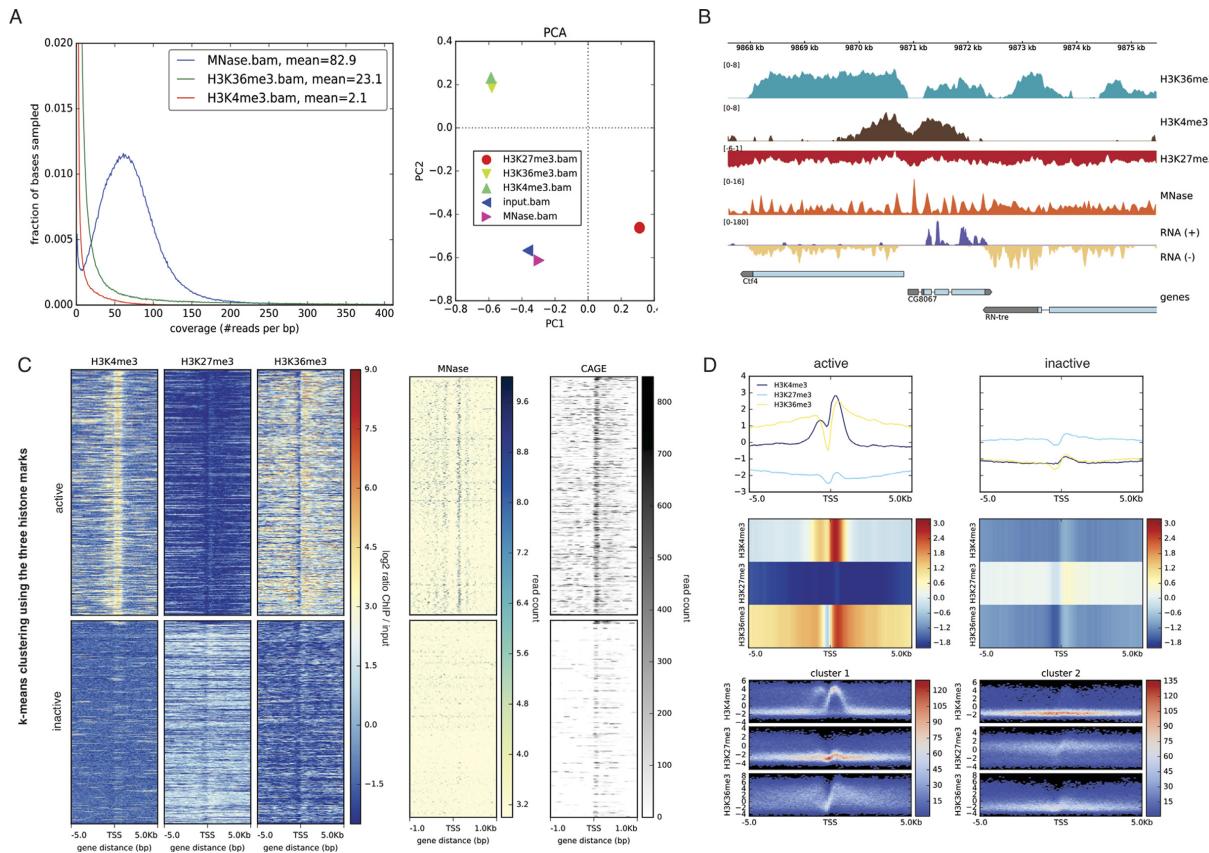


Figure 1. deepTools2 has enhanced features for quality control, data reduction, normalization and visualization of deeply sequenced data. **(A)** deepTools2 comes with two new quality control tools: *plotCoverage*, which displays sequencing depth distribution; and *plotPCA*, which plots the results of principal component analysis (*PCA*) on BAM or bigWig files. **(B)** Signal visualization by means of normalized bigWig tracks, produced by deepTools' *bamCompare* (ChIP-seq samples) and *bamCoverage*, which can now handle MNase- and strand-specific RNA-seq samples. In the genes track, the gray color represents untranslated regions and the thin lines represent introns. **(C)** *computeMatrix* and *plotHeatmap* were used to summarize and cluster multiple bigWig scores over genomic intervals. The image shows the resulting *k*-means clustering of ChIP-seq signals of three histone marks around the transcription start site of genes (with $k = 2$). Subsequently, the clustered regions were used to plot MNase-seq and CAGE read coverages. In the image, each row corresponds to the same genomic region. **(D)** *plotProfile* now offers additional means for visualization. The top panel shows the average signal for different samples and different cluster of regions (left and right plot). In the middle, the same profiles are represented as heatmaps. The bottom panel shows summary profiles where colors correspond to the observed frequency of the signal within each cluster. All data are from the *Drosophila melanogaster* S2 cell line and are publicly available (see Section Accession numbers).

across genomic regions. The new tools *multiBamSummary* and *multiBigwigSummary* summarize the information from multiple samples into a single score for each region per sample. The resulting matrix can be used for correlation analyses, pairwise scatter plots and PCA (see above), or other downstream analyses.

Similarly, *computeMatrix* calculates the distribution of scores over selected genomic regions. Typical applications include the computation of signals around promoters or across gene bodies that are scaled to the same length. The primary purpose is to produce a matrix that can later be visualized. One of the most frequently requested features of deepTools was the extension of *computeMatrix* to allow for the combined analysis of multiple samples. Now, *computeMatrix* can aggregate regional scores from many different samples into a single matrix for subsequent visualization. Figure 1C illustrates this capability for 5 different

signal files (H3K4me3, H3K27me3, H3K36me3, MNase, CAGE). Here, genome-wide scores were provided as five bigWig files, and a single BED file of annotated genes. From these inputs, *computeMatrix* generated a matrix file that was then visualized using *plotHeatmap* and *plotProfile*. Those tools have been extended to produce composite images of multiple heatmaps and summary profiles (Figure 1C and D). They also include options for unsupervised data analysis (*k*-means and hierarchical clustering), which are most useful if no prior grouping of regions is available. In the example of Figure 1C, two groups of regions were determined by *k*-means clustering of the signal profiles with $k = 2$. Alternatively, the user can provide multiple groups of pre-defined regions (multiple BED files) to *computeMatrix*. The tool will then output a structured matrix in which the rows are grouped according to the pre-defined regions, e.g. different sets of genes. Subsequent visualization corresponds

to a supervised analysis, where separate heatmaps and summary profiles are generated for each of the previously defined groups of interest. Both *plotHeatmap* and *plotProfile* now also offer many new options for customizing their output (Figure 1D).

IMPLEMENTATION AND ACCESS

Our deepTools web server has been implemented within the Galaxy framework (6). The software has been virtualized as a Docker container, which currently has access to 24 cores, 141GB memory and more than 8TB of disk storage. Our web server offers all deepTools' functionality for free and without registration. Users with more challenging requirements can access deepTools through multiple additional means:

Toolshed

For already existing local Galaxy instances, deepTools can be easily installed through the Galaxy toolshed (11).

Docker Image

For those wishing to use deepTools locally within the Galaxy framework, we have significantly simplified the way in which a deepTools Galaxy server can be deployed. Through incorporating everything in a single Docker container, users can now install a deepTools Galaxy web server on a desktop machine with a single command. Further details are available in our online documentation (<http://deeptools.readthedocs.org/en/latest/content/installation.html>).

Stand alone

deepTools can be installed through the Python package index (PyPi, <https://pypi.python.org/pypi/deepTools>), through GitHub (<https://github.com/fidelram/deepTools>) and in the bioconda channel of Anaconda (<https://anaconda.org/bioconda/deeptools>). More information can be found at <http://deeptools.readthedocs.org/en/latest/content/installation.html>.

API

Finally, deepTools are now fully available as a Python package. Numerous functions, such as the function for summarizing read counts from a BAM file per genomic region (*countReadsPerBin*), can easily be accessed in any Python script that imports the deepTools package. Usage details can be found at <http://deeptools.readthedocs.org/en/latest/content/api.html>.

EXTENSIVE TESTING, DOCUMENTATION AND COMMUNITY SUPPORT

deepTools have greatly benefited from transparent development as they are hosted on [github.org](https://github.com), where users and collaborators can directly post issues, fork their own versions,

and follow the changes to the code. All tools contain comprehensive and automatic tests that evaluate proper functioning after any modification of the code. We continuously update our documentation, and respond to questions sent to our mailing list (deeptools@googlegroups.com) and through the Galaxy bug report system. For this major update, we have reorganized the documentation with updated examples of tools usage, step-by-step protocols and FAQs. Also, detailed usage descriptions have been added to the Galaxy wrappers. To further improve our documentation, we migrated it to readthedocs.org, where we can offer versioned documentation and an easy way to export it as a composite PDF file. The most up-to-date documentation can now always be found at <http://deeptools.readthedocs.org/en/latest/>.

DISCUSSION AND OUTLOOK

The adoption of deep sequencing in many laboratories has created the need for accessible, efficient and transparent methods to process the data directly by the researchers. However, the ever-increasing throughput of deep sequencing technologies requires sophisticated bioinformatics solutions that are not always available to every lab. Via the deepTools web server, researchers can access a set of easy to use, yet powerful programs that cover quality control, normalization, integration and visualization of the data. Since deepTools employ a high level of parallelization for the computationally most expensive tasks, they are well suited to work with a large number of samples emerging from large-scale data production centers (12,13) or single-cell sequencing (14). The Galaxy framework is frequently used for establishing reproducible and standardized analysis workflows, even for groups where bioinformatics support is otherwise scarce. We have also found it to be a user-friendly environment that most biomedical scientists are comfortable with. This makes it a versatile platform for training workshops, as well as for sharing data and workflows among collaborators.

The modular design of deepTools has allowed us to significantly expand the scope of the analyses compared to its first release. More functionality can be added in the future, and since the software development is completely open and transparent, it will benefit from contributions from the wider community to incorporate new standards and best practices in this rapidly evolving field.

ACCESSION NUMBERS

The datasets used in Figure 1 were: H3K4me3 and H3K26me3 from GEO:GSE41440 (15), H3K36me3 from GEO:GSE27679 (16), MNase-seq from GEO:GSE58821 (17), CAGE from GEO:GSE52884 (18).

ACKNOWLEDGEMENTS

The authors would like to thank all users of deepTools for their constructive feedback and suggestions. We would further like to thank specific users for their valuable input: Peter Ebert, Sebastian Preissl and Ralf Gilsbach. We would also like to thank Abdullah Sahyoun for insightful discussions.

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

- Younesy,H., Nielsen,C.B., Möller,T., Alder,O., Cullum,R., Lorincz,M.C., Karimi,M.M. and Jones,S.J.M. (2013) An interactive analysis and exploration tool for epigenomic data. *Comput. Graph. Forum*, **32**, 91–100.
- Akalin,A., Franke,V., Vlahoviček,K., Mason,C.E. and Schübeler,D. (2015) Genomation: a toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics*, **31**, 1127–1129.
- Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
- Shen,L., Shao,N., Liu,X. and Nestler,E. (2014) ngs.plot: quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics*, **15**, 284.
- Ramírez,F., Dündar,F., Diehl,S., Grüning,B.A. and Manke,T. (2014) deepTools: a flexible platform for exploring deep-sequencing data. **42**, W187–W191.
- Hillman-Jackson,J., Clements,D., Blankenberg,D., Taylor,J. and Nekrutenko,A. Galaxy Team (2012) Using Galaxy to perform large-scale interactive data analyses. *Curr. Protoc. Bioinformatics*, Chapter 10, Unit 10.5.47.
- Speir,M.L., Zweig,A.S., Rosenbloom,K.R., Raney,B.J., Paten,B., Nejad,P., Lee,B.T., Learned,K., Karolchik,D., Hinrichs,A.S. et al. (2016) The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res.*, **44**, D717–D725.
- Feng,J., Liu,T., Qin,B., Zhang,Y. and Liu,X.S. (2012) Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.*, **7**, 1728–1740.
- Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P. et al. (2012) ChIP-seq guidelines and practices of the ENCODE and modeENCODE consortia. *Genome Res.*, **22**, 1813–1831.
- Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
- Blankenberg,D., Von Kuster,G., Bouvier,E., Baker,D., Afgan,E., Stoler,N., Taylor,J., Nekrutenko,A. and Galaxy Team (2014) Dissemination of scientific software with Galaxy ToolShed. *Genome Biol.*, **15**, 403.
- Bernstein,B.E., Birney,E., Dunham,I., Green,E.D., Gunter,C. and Snyder,M. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
- Consortium,R.E., Kundaje,A., Meuleman,W., Ernst,J., Bilelky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. et al. (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
- Klein,A.M., Mazutis,L., Akartuna,I., Tallapragada,N., Veres,A., Li,V., Peshkin,L., Weitz,D.A. and Kirschner,M.W. (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, **161**, 1187–1201.
- Herz,H.-M., Mohan,M., Garruss,A.S., Liang,K., Takahashi,Y.-H., Mickey,K., Voets,O., Verrijzer,C.P. and Shilatifard,A. (2012) Enhancer-associated H3K4 monomethylation by Trithorax-related, the Drosophila homolog of mammalian Mll3/Mll4. *Genes Dev.*, **26**, 2604–2620.
- Chen,Y., Negre,N., Li,Q., Mieczkowska,J.O., Slattery,M., Liu,T., Zhang,Y., Kim,T.-K., He,H.H., Zieba,J. et al. (2012) Systematic evaluation of factors influencing ChIP-seq fidelity. *Nat. Methods*, **9**, 609–614.
- Ramírez,F., Lingg,T., Toscano,S., Lam,K.C., Georgiev,P., Chung,H.-R., Lajoie,B.R., de Wit,E., Zhan,Y., de Laat,W. et al. (2015) High-affinity sites form an interaction network to facilitate spreading of the MSL complex across the X chromosome in Drosophila. *Mol. Cell*, **60**, 146–162.
- Liang,J., Lacroix,L., Gamot,A., Cuddapah,S., Queille,S., Lhoumaud,P., Lepetit,P., Martin,P.G.P., Vogelmann,J., Court,F. et al. (2014) Chromatin immunoprecipitation indirect peaks highlight long-range interactions of insulator proteins and Pol II pausing. *Mol. Cell*, **53**, 672–681.

A.6 snakePipes enables reproducible epigenomic analysis

I developed the allele-specific and HiC workflows and contributed to DNA-mapping, ChIP-seq, ATAC-seq and RNA-seq workflows and documentation. I performed the analysis, prepared the figures and wrote the manuscript with input from all authors.

snakePipes enable flexible, scalable and integrative epigenomic analysis

Vivek Bhardwaj^{*1,2}, Steffen Heyne^{*1}, Katarzyna Sikora¹, Leily Rabbani¹, Michael Rauer¹, Fabian Kilpert³, Andreas S Richter⁴, Devon P Ryan^{1#}, Thomas Manke^{1#}

¹Max Planck Institute of Immunobiology and Epigenetics. Stübeweg 51, 79108 Freiburg
Germany

²Faculty of Biology, University of Freiburg, Schänzlestraße 1, 79104 Freiburg, Germany

³Institutes of Neurogenetics & Cardiogenetics, University of Lübeck, Maria-Goeppert-Str. 1,
23562 Lübeck, Germany

⁴Genedata AG, Margarethenstrasse 38, 4053 Basel, Switzerland

* co-first authors (arranged alphabetically)

Corresponding author

manke@ie-freiburg.mpg.de

+49 (0)761 5108738

ryan@ie-freiburg.mpg.de

Abstract

The scale and diversity of epigenomics data has been rapidly increasing and ever more studies now present analyses of data from multiple epigenomic techniques. Performing such integrative analysis is time-consuming, especially for exploratory research, since there are currently no pipelines available that allow fast processing of datasets from multiple epigenomic assays while also allow for flexibility in running or upgrading the workflows. Here we present a solution to this problem : *snakePipes*, which can process and perform downstream analysis of data from all common epigenomic techniques (ChIP-seq, RNA-seq, Bisulfite-seq, ATAC-seq, Hi-C and single-cell RNA-seq) in a single package. We demonstrate how *snakePipes* can simplify integrative analysis by reproducing and extending the results from a recently published large-scale epigenomics study with a few simple commands. *snakePipes* are available under an open-source license at <https://github.com/maxplanck-ie/snakepipes>.

Main

Epigenomics is a fast growing field, and due to the consistent fall in the price of sequencing, increase in multiplexing abilities, and multiple innovations in laboratory protocols, it has become increasingly convenient to perform multiple epigenomic assays within a project. However, a major bottleneck on the way to process and analyse this data in a reproducible way, particularly for novice analysts, is the availability of analysis pipelines. Next-generation sequencing (NGS) analysis pipelines are composed of a series of data processing steps, employ standardised processing parameters, and are usually scalable to large number of samples ¹. Due to such properties, most pipelines are currently developed and deployed for settings where standardized, large scale analysis is required. Examples are RNA-seq variant-calling pipelines deployed in clinical settings ², or processing pipelines developed for large-scale consortia ^{3,4}.

However, in a typical basic science research setting, researchers also seek to modify parameters, update tool versions or extend the workflows, while maintaining their scalability and ease-of-use. Conventional NGS pipelines, although scalable, do not allow for this flexibility. Options for exploratory and downstream analysis have been limited, resulting in various expert users developing their own custom pipelines suited to their needs. Computational frameworks such as Galaxy⁵ and Nextflow⁶ exist, but they still demand novice users to be trained and implement their workflows themselves from scratch. This leads to a conundrum, how can we provide a set of workflows following best-practices that are easy to install and run for the novice users, while still providing the flexibility of extending and upgrading the workflows to the expert users?

We developed snakePipes to address such requirements. snakePipes provide flexible processing as well as downstream analysis of data from the most common assays used in epigenomic studies: ChIP-seq, RNA-seq, whole-genome bisulfite-seq (WGBS), ATAC-seq, Hi-C and single-cell RNA-seq in a single package (Fig. 1b, Implementation Details). It employs snakemake⁷ as a workflow language, which benefits from easy readability of the code, widespread adoption, and offers scalability using most cluster and cloud computing platforms. snakePipes also makes use of conda environments and the bioconda platform⁸, which allow hassle-free installation and upgrade of tools (Fig. 1a, Implementation Details). Conda environments allow execution of tools avoiding dependency conflicts, and do not require root permissions to run. Due to a modular architecture, various tools are shared between workflows, which simplifies data integration since data from multiple technologies are processed using identical tool versions and genome annotations. The genome annotations and indices are shared by all workflows, and can also be generated directly via snakePipes, facilitating easy setup as well as integrative analysis. Finally, workflows in snakePipes employ extensive quality-checks and also produce reports using multiQC⁹ and R, that inform the user of processing and analysis results.

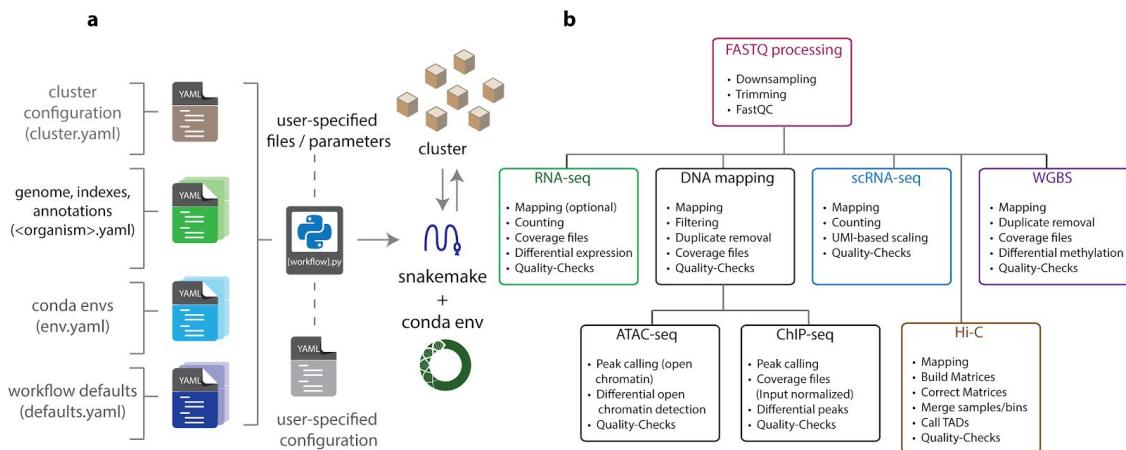
Figure 1

Figure 1. General architecture and available workflows in snakePipes. **a. Configuration.** All configurable parameters for snakepipes are defined as YAML files. Organism-specific YAML files need to be configured once during setup, while the configuration of cluster, conda envs and workflow defaults is optional. After installation, the location of these files can be revealed by the snakePipes info command. Workflows are all executed using their command line wrappers. During execution user parameters and/or a user-provided YAML file can be used to override the workflow defaults and add flexibility to the processing/analysis. **b. Execution.** FASTQ files provided by the user first goes through FASTQ processing, followed by one of the workflows. Outputs from DNA mapping workflow can be further used as input for the ChIP-seq and ATAC-seq workflows. Only general processing steps are listed and each workflow includes workflow-specific quality-checks.

Apart from conventional processing steps such as mapping, counting and peak calling, workflows in snakePipes also include various downstream analysis. All workflows (except scRNA-seq workflow) optionally accept a sample information (tab-separated) file that can be used to define groups of sample. This allows comparative analysis such as differential gene or transcript expression analysis for the RNA-seq workflow, differential peak calling for ChIP-Seq workflow, differential open chromatin detection for ATAC-seq workflow, and detection of differentially methylated regions (either de-novo or on user-specified regions) for WGBS workflow. The HiC workflow uses sample information to merge groups and can perform TAD calling with parameters adapted to the resolution of produced matrix (using HiCExplorer¹⁰). This preliminary analysis, combined with visualization-ready bed and bigwig files, allows users to quickly interpret their data.

Most workflows in snakePipes also allow processing and downstream analysis of data in an allele-specific manner. DNA-mapping and RNA-seq workflows utilize SNPSplit¹¹ to map data to a single or dual-hybrid genome, for samples coming from inbred mouse strains or from *Drosophila* SNP lines using the “allelic-mapping” mode. Single/dual hybrid genome indices can be provided externally, or can be created on the fly by simply providing a VCF file¹² and the name of desired strains. The workflows also generate allele-specific coverage files, QC reports, and performs allele-specific differential analysis using appropriate statistical design, without requiring any additional user intervention.

In order to demonstrate how snakePipes can simplify analysis and interpretation of data from multiple epigenomic assays, we downloaded and processed data from a recently published study that investigated the role of Smchd1 protein on the mammalian X-chromosome¹³. The knock-out of Smchd1 in mouse neural progenitor cells (NPCs) affects the organization of inactive X-chromosome and leads to a loss of H3K27me3 domains along with a gain of H3K4me3 on various genes on the inactive X-chromosome. This results in de-repression of genes, seen as up-regulation in RNA-seq data. We observed the same effects upon re-analysis of the ChIP-Seq, RNA-seq and Hi-C data from the study, directly from the output of snakePipes, without further downstream analysis (Fig 2, Fig S1a-b). Further, we integrated this information with ATAC-seq data available online¹⁴ to discover that the genes de-repressed upon Smchd1 knock-out display significantly higher open chromatin signal at their promoters in the wild-type NPCs, compared to down-regulated or unchanged, lowly expressed genes (Fig. S1c). Whole-genome bisulfite data obtained from another study (GSE101090) suggested that under wild-type conditions, promoters of the de-repressed genes show methylation levels slightly higher than the downregulated genes but significantly lower than unchanged, lowly expressed genes (Fig. S1d), corroborating previous¹⁵ and recent¹⁶ links between promoter CpG methylation and gene repression.

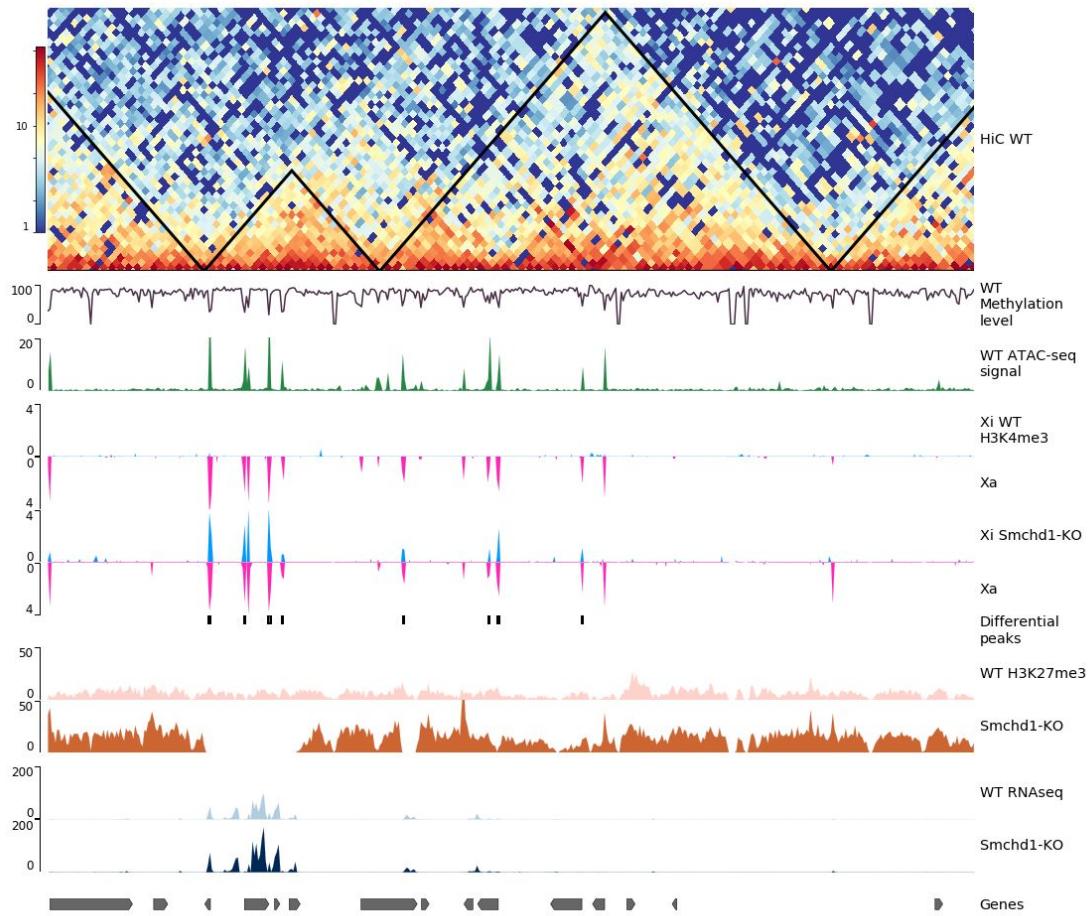


Figure 2. Outputs from *snakePipes* workflows quickly provide meaningful biological interpretation. Genome tracks plotted using pyGenomeTracks¹⁰. **Top track:** output of HiC workflow with TADs (black triangles) on wild-type (WT) NPCs. **Tracks 2:** Methylation signal reported from the whole-genome bisulfite-seq (WGBS) workflow on wild-type NPCs. Gene promoters show demethylation. **Track 3:** Open chromatin signal from ATAC-seq workflow on wild-type NPCs. **Track 4-7 :** output of allele-specific ChIP-Seq workflow, Inactive X chromosome (Xi) tracks are in Blue while the active X chromosome (Xa) is shown in Pink. The knock-out (KO) of Smchd1 shows increase in H3K4me3 on Xi. **Track 8 :** Allele-specific differential peaks detected by the ChIP-Seq workflow in KO, over WT. **Track 9-10 :** H3K27me3 signal from the ChIP-Seq workflow. KO shows loss of H3K27me3 domains on certain genes. **Track 11-12 :** RPKM signal from RNA-seq workflow. Genes which loose H3K27me3 show up-regulation in KO. **Bottom track :** Genes.

Our results demonstrate how a multi-assay epigenomic analysis toolkit like snakePipes could simplify data processing, reproduce previously published results, and allow new biological interpretations with minimal effort. snakePipes are under active development, open source and available via conda : `conda install -c mpi-ie -c bioconda -c conda-forge snakePipes`. Source code is available at <https://github.com/maxplanck-ie/snakepipes> and the documentation is hosted at <https://snakemake.readthedocs.io/en/latest/>.

Implementation Details

General architecture of snakePipes

The general architecture of snakePipes is summarized in Fig. 1a. snakePipes utilizes conda and bioconda for setup and execution of workflows. All information required for workflow execution is stored in easy-to-edit YAML (Yet Another Markup Language) files. The **cluster.yaml** file defines the command used to execute each process on a cluster (assuming HPC cluster infrastructure like Slurm or SGE has been set up). For each organism of interest, the **<organism>.yaml** file describes the location of genome fasta, (mapping) indices and annotations. This allows running different workflows on exactly the same version of genome and annotations. The conda **env.yaml** file is used to specify version number of required tools for each workflow. The required tools are then fetched and setup automatically via the conda and bioconda repositories either during workflow setup, or execution. The **defaults.yaml** file specifies reasonable default parameter for each workflow according to the best practices for the most common sequencing protocols. These files are used by the command line wrappers, that use snakemake to execute the workflows on the cluster. In the absence of cluster setup, workflows can also be executed locally. Each step of a workflow is defined as a snakemake “rule”, which is executed in its own virtual environment, avoiding conflicts between tools. All log files, including user-supplied commands are written in the working directory, along with (optionally) a graph of executed steps. This allows users to easily reproduce and communicate their analysis in the future.

Running and testing the workflows

Comprehensive documentation for snakePipes can be found online: <https://snakemake.readthedocs.io/> and the test datasets are available on zenodo : <https://zenodo.org/record/1346303>. snakePipes also provide a `createIndices` workflow that creates genome indices and annotations (contents of <organism>.yaml) from a user-specified genome fasta file or URL.

Workflows in snakePipes

snakePipes provide DNA-mapping, ChIP-seq, ATAC-seq, RNA-seq, whole-genome bisulfite-seq (WGBS), HiC and single-cell RNA-seq workflows. All workflows make use of the common fastq downsampling (via seqtk (<https://github.com/lh3/seqtk>)) and trimming (via cutadapt¹⁷ and Trim Galore! (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/)) module. All workflows also produce an interactive report using multiQC⁹ that summarizes outputs from multiple workflow steps and samples.

In the **DNA-mapping** workflow, the fastq files are aligned to the genome via Bowtie2¹⁸ and filtering can be performed via samtools¹⁹ using user-provided parameters. Various quality-checks are performed via SamBamba²⁰, Picard²¹, deepTools²² and (optionally) qualimap²³. Coverage files (bigwigs) are generated via deepTools. The output of DNA-mapping workflow can then be used for ChIP-Seq or ATAC-seq workflows. DNA-mapping workflow handles both single and paired-end files and could also be used for whole-genome alignments.

The **ChIP-seq** workflow takes information about the samples (corresponding input controls, expected broad/sharp mark) using a yaml file, and performs ChIP-specific quality-checks via deepTools. It also generates input-normalized bigwig files and performs peak calling for both sharp (via MACS2²⁴) and broad (via histoneHMM²⁵) marks. The **ATAC-seq** workflow takes paired-end DNA mapping output and performs quality-checks and filtering useful for ATAC-seq samples. It then performs detection of open chromatin using MACS2. Both ChIP-Seq and ATAC-Seq workflows can perform detection of differential peaks or differential open chromatin regions between groups of samples using CSAW²⁶, if a sample sheet is provided.

The **RNA-seq** workflow can be run in “alignment” or “alignment-free” mode. In the alignment mode, fastq files are aligned to the genome via user-selected aligner (STAR²⁷ or HISAT2²⁸) and high-quality primary alignments are counted via featureCounts²⁹. In the alignment-free mode, the transcripts are directly quantified via Salmon³⁰. Transcripts can be filtered for various features before quantification. Additionally, the “*deepTools_qc*” mode can be added, which performs various quality-checks via deepTools and produces normal and depth-normalized (RPKM) coverage files. Differential gene and transcript expression analysis could then be performed using DESeq2³¹, wasabi (<https://github.com/COMBINE-lab/wasabi>) and Sleuth³².

The **scRNA-seq** workflow performs mapping and counting of data obtained from the CEL-Seq2 protocol³³. Fastq files are first preprocessed by moving cell barcodes and unique molecular indices (UMIs) to read headers and then mapped using STAR (Dobin et al.). Quantification is then performed per-cell by accounting for UMIs and the resulting counts corrected for Poisson sampling. A variety of quality control steps are also taken, such as computing a heatmap per well-plate of obtained transcript counts and the correlation between reads and UMIs. After the workflow is finished the resulting counts files are ready for custom downstream analysis (e.g., clustering or differential expression).

The **Hi-C** workflow performs read mapping of paired-end HiC data using BWA³⁴. It then uses HiCExplorer¹⁰ to build and correct the Hi-C matrices using the iterative correction (ICE) method³⁵. Matrices can be built at a user-specified resolution or at restriction fragment resolution by simply specifying the name of the restriction enzyme. The corrected HiC matrices can then be used for detection and visualization of topologically associated domains (TADs)³⁶. Quality reports are produced using HiCExplorer and are then summarized by MultiQC for comparison of samples.

The **WGBS** (whole-genome bisulfite-seq) workflow performs mapping of paired-end WGBS-seq data on a bisulfite converted genome using bwa-meth³⁷. To help assess the quality of the experiment, several metrics, including bisulfite conversion rate and coverage of random CpGs, are collected in a report. Counting of reads supporting methylated and unmethylated cytosines is performed with MethylDackel (<https://github.com/dpryan79/MethylDackel>). De novo discovery

of differentially methylated regions (DMRs) can also be performed using Metilene³⁸ if a sample sheet is provided.

Methods

Processing of Online data

HiC, ChIP-Seq, and RNA-seq data for Smchd KO and wild-type Neural Progenitor Cells (NPCs) was downloaded from GSE99991. ATAC-Seq data for wild-type NPCs was downloaded from GSE71156 and WGBS data for wild-type NPCs was downloaded from GSE101090. All data was processed with snakePipes (version 1.0.0alpha5) on mouse genome GRCm38 (mm10).

The parameters specified for processing the data are as follows:

ATAC-Seq

DNA-mapping performed on mouse genome with parameters : `*-m allelic-mapping -j 30 --gcbias --mapq 5 --dedup --fastqc --trim --properpairs*

ATAC-seq performed on DNA-mapping output with parameters : `*--bw-binsize 10*`. By default, fragments longer than 150 nt are removed from peak calling `*--atac-fragment-cutoff 150*`.

ChIP-Seq

DNA-mapping performed on a dual hybrid (129S1/CAST) mouse genome with parameters : `*-m allelic-mapping --trim --fastqc --bw-binsize 10 --plotFormat pdf --dedup --mapq 10 --SNPfile <snp_positions.txt> --Nmasked_index <bowtie2_index.bt2>* `

ChIP-Seq performed on DNA-mapping output with parameters : `*--bw-binsize 10*` with *chip_sampleInfo.yaml* file which specified the corresponding input controls and peak-type.

H3K4me3 samples were specified as `sharp` while the H3K27me3 samples were described as `broad`

RNA-Seq

RNA-seq workflow was run with parameters : `--fastqc --trim -m alignment,deepTools_qc --DE sampleinfo.tsv` where sampleinfo.tsv file defined groups with replicates (5 control and 5 knock-out samples). One of the knock-out samples was removed after inspecting PCA output (Fig. S1A) and workflow was re-run to obtain differentially expressed genes (Fig. S1B).

Hi-C

Hi-C workflow was run with parameters : `--merge_samples --sampleInfo sampleinfo.tsv --distVsCount --bin_size 10000 --trim --fastqc` where sampleinfo.tsv was used to define the two Hi-C replicates.

WGBS

WGBS workflow was run with all default parameters.

Downstream analysis

Plotting of ATAC-seq signal on up-regulated, down-regulated and 500 lowly expressed, un-affected genes was performed by subsetting gene sets from DESeq2 output of snakePipes RNA-seq module, followed by deepTools computeMatrix with options : *-a 5000 -b 5000*, and plotHeatmap with option *--plotType se*. For WGBS data, bedgraph output of CpG methylation signal from WGBS pipeline was converted to bigWig using *UCSCtools bedGraphToBigWig* and plotting using computeMatrix (*--binsize 100*) and plotHeatmap.

Code availability

snakePipes is open-source and freely available on GitHub :
<https://github.com/maxplanck-ie/snakepipes>

Acknowledgements

Authors acknowledge the contribution of Gina Renschler and Jana Böhm on testing of workflows. We also thank Chen-Yu Wang for useful comments on our preprint. TM acknowledges funding from the German Science Foundation (CRC992 “Medical Epigenetics”).

Availability of data and materials

Online datasets re-analysed during this study are available in GEO with accession numbers : GSE99991, GSE71156 and GSE101090

Competing interests

The authors declare no competing interests.

Authors' contributions

VB developed the allele-specific and HiC workflows and contributed to DNA-mapping, ChIP-seq, ATAC-seq and RNA-seq workflows and documentation. SH developed the scRNA-seq workflow and documentation and contributed to DNA-mapping, ChIP-seq and RNA-seq workflows. DPR improved the wrapper design, integrated installation and conda support, contributed to the documentation and bug fixes to various workflows. KS developed the WGBS workflow and documentation. LR contributed to HiC workflow and documentation. MR developed ATAC-seq workflow. FK contributed to RNA-seq workflow. AR contributed to DNA-mapping and ChIP-seq workflow. FK, SH and AR contributed to the general design of snakePipes and wrote the early version of the wrappers. VB performed the analysis and wrote the manuscript with input from all authors. TM conceived the project and supervised the development of snakePipes.

References

1. Leipzig, J. A review of bioinformatic pipeline frameworks. *Brief. Bioinform.* **18**, 530–536 (2017).
2. Blanca, J. M., Pascual, L., Ziarsolo, P., Nuez, F. & Cañizares, J. ngs_backbone: a pipeline for read cleaning, mapping and SNP calling using next generation sequence. *BMC Genomics* **12**, 285 (2011).
3. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
4. Fernández, J. M. *et al.* The BLUEPRINT Data Analysis Portal. *Cell Syst* **3**, 491–495.e5 (2016).
5. Goecks, J., Nekrutenko, A., Taylor, J. & Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11**, R86 (2010).
6. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
7. Köster, J. & Rahmann, S. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
8. Grüning, B. *et al.* Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods* **15**, 475–476 (2018).
9. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for

- multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
10. Ramírez, F. *et al.* High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.* **9**, 189 (2018).
 11. Krueger, F. & Andrews, S. R. SNPsplits: Allele-specific splitting of alignments between genomes with known SNP genotypes. *F1000Res.* **5**, 1479 (2016).
 12. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
 13. Wang, C.-Y., Jégu, T., Chu, H.-P., Oh, H. J. & Lee, J. T. SMCHD1 Merges Chromosome Compartments and Assists Formation of Super-Structures on the Inactive X. *Cell* **174**, 406–421.e25 (2018).
 14. Giorgetti, L. *et al.* Structural organization of the inactive X chromosome in the mouse. *Nature* **535**, 575–579 (2016).
 15. Schübeler, D. Function and information content of DNA methylation. *Nature* **517**, 321–326 (2015).
 16. Korthauer, K. & Irizarry, R. A. Genome-wide repressive capacity of promoter DNA methylation is revealed through epigenomic manipulation. *bioRxiv* 381145 (2018). doi:10.1101/381145
 17. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
 18. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
 19. Li, H. *et al.* 692 (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2093
 20. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing

- of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
21. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
 22. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–5 (2016).
 23. García-Alcalde, F. *et al.* Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* **28**, 2678–2679 (2012).
 24. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* **7**, 1728–1740 (2012).
 25. Heinig, M. *et al.* histoneHMM: Differential analysis of histone modifications with broad genomic footprints. *BMC Bioinformatics* **16**, 60 (2015).
 26. Lun, A. T. L. & Smyth, G. K. csaW: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Res.* **44**, e45 (2016).
 27. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
 28. Kim, D., Langmead, B. & Salzberg, S. HISAT2: graph-based alignment of next-generation sequencing reads to a population of genomes. (2017).
 29. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
 30. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
 31. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
 32. Pimentel, H., Bray, N. L., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat. Methods* **14**, 687–690 (2017).

33. Hashimshony, T. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).
34. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).
35. Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* **9**, 999–1003 (2012).
36. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
37. Pedersen, B. S., Eyring, K., De, S., Yang, I. V. & Schwartz, D. A. Fast and accurate alignment of long bisulfite-seq reads. *arXiv [q-bio.GN]* (2014).
38. Jühling, F. *et al.* metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome Res.* **26**, 256–262 (2016).

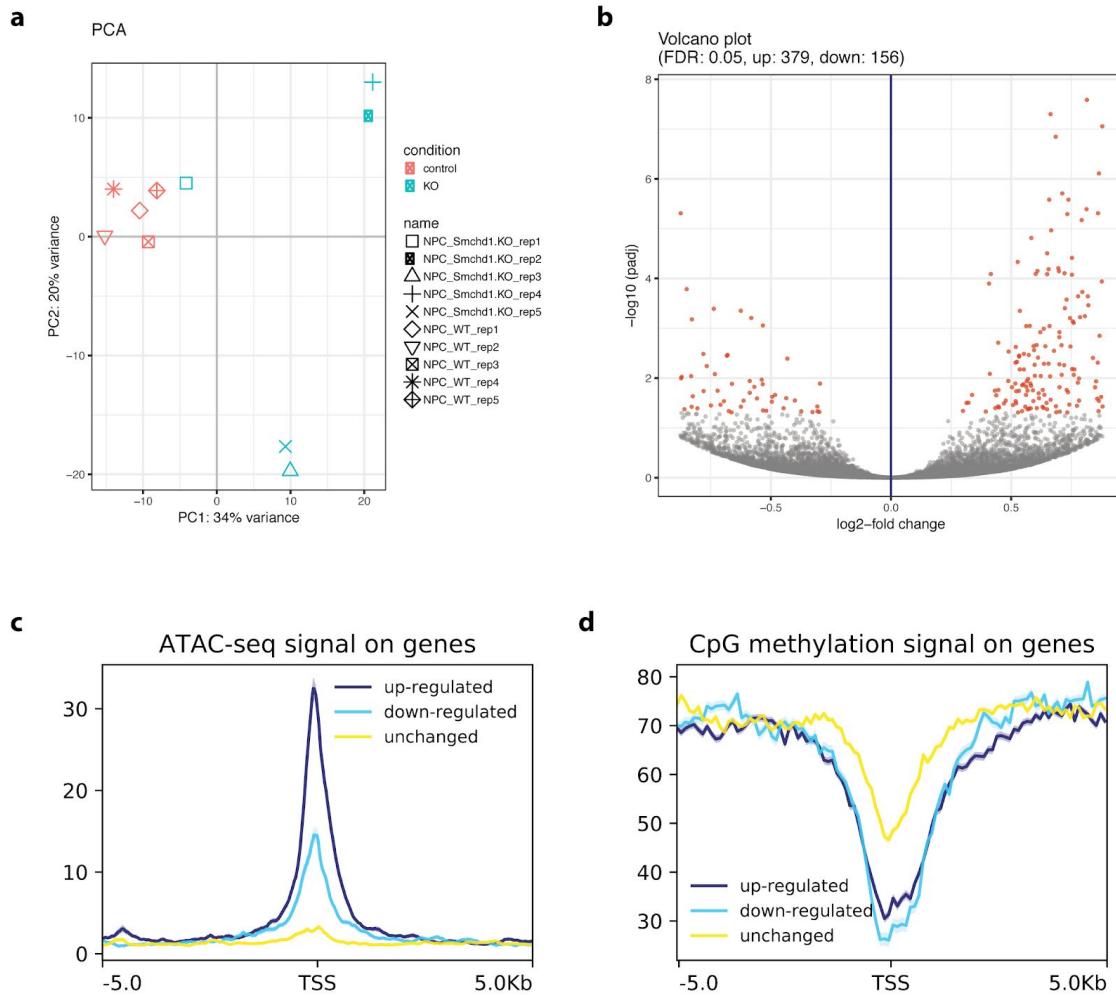
Figure S1

Figure S1. Analysis of de-repressed genes upon *Schmd1* knock-out. **A.** PCA output from snakepipes suggested that one knock-out sample (replicate1) behaves differently. This sample was later revealed to be the XO clone which lost its inactive X chromosome. The sample was removed for DESeq2 analysis and the workflow was re-run. **B.** Volcano plot for DESeq2 output from snakePipes (knock-out replicate 1 excluded), shows an increase in up-regulated genes, indicating de-repression upon knock-out. **C.** Wild-type ATAC-seq signal on UP, DOWN and unchanged (NONE) genes, gene lists were extracted from DESeq2 output of RNA-seq workflow and depth-normalized bigwigs from ATAC-seq workflow was used for plotting. **D.** Wild-type methylation level reported by the WGBS workflow on UP, DOWN and unchanged (NONE) genes. (TSS = Transcription Start Site)

B. Supplemental information

C. Academic Vita

References

C.1 References

1. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561–563 (1970).
2. Epp, C. D. Definition of a gene. *Nature* **389**, 537 (1997).
3. Gerstein, M. B. *et al.* What is a gene, post-ENCODE? History and updated definition. *Genome Res.* **17**, 669–681 (2007).
4. Berthelot, C., Villar, D., Horvath, J. E., Odom, D. T. & Flieck, P. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nat Ecol Evol* **2**, 152–163 (2018).
5. Villegas, V. E. & Zaphiropoulos, P. G. Neighboring gene regulation by antisense long non-coding RNAs. *Int. J. Mol. Sci.* **16**, 3251–3266 (2015).
6. Catalanotto, C., Cogoni, C. & Zardo, G. MicroRNA in Control of Gene Expression: An Overview of Nuclear Functions. *Int. J. Mol. Sci.* **17**, (2016).
7. Engreitz, J. M., Ollikainen, N. & Guttman, M. Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression. *Nat. Rev. Mol. Cell Biol.* **17**, 756–770 (2016).
8. Siomi, H. & Dreyfuss, G. RNA-binding proteins as regulators of gene expression. *Curr. Opin. Genet. Dev.* **7**, 345–353 (1997).
9. Berger, S. L., Kouzarides, T., Shiekhattar, R. & Shilatifard, A. An operational definition of epigenetics. *Genes Dev.* **23**, 781–783 (2009).
10. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
11. Schuster, S. C. Next-generation sequencing transforms today's biology. *Nat. Methods* **5**, 16–18 (2008).
12. Griffiths, P. E. In what sense does ‘nothing make sense except in the light of evolution’? *Acta Biotheor.* **57**, 11–32 (2009).
13. Mouse Genome Sequencing Consortium *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).

14. Church, D. M. *et al.* Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol.* **7**, e1000112 (2009).
15. Rands, C. M., Meader, S., Ponting, C. P. & Lunter, G. 8.2% of the Human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLoS Genet.* **10**, e1004525 (2014).
16. Graur, D. An Upper Limit on the Functional Fraction of the Human Genome. *Genome Biol. Evol.* **9**, 1880–1885 (2017).
17. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
18. Riethoven, J.-J. M. Regulatory Regions in DNA: Promoters, Enhancers, Silencers, and Insulators. in *Computational Biology of Transcription Factor Binding* (ed. Ladunga, I.) 33–42 (Humana Press, 2010).
19. Huang, H., Sabari, B. R., Garcia, B. A., Allis, C. D. & Zhao, Y. SnapShot: histone modifications. *Cell* **159**, 458–458.e1 (2014).
20. Fuhrmann, G. *et al.* Mouse germline restriction of Oct4 expression by germ cell nuclear factor. *Dev. Cell* **1**, 377–387 (2001).
21. Feldman, N. *et al.* G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis. *Nat. Cell Biol.* **8**, 188–194 (2006).
22. Smallwood, A. & Ren, B. Genome organization and long-range regulation of gene expression by enhancers. *Curr. Opin. Cell Biol.* **25**, 387–394 (2013).
23. Cullen, K. E., Kladde, M. P. & Seyfred, M. A. Interaction between transcription regulatory regions of prolactin chromatin. *Science* **261**, 203–206 (1993).
24. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
25. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
26. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
27. Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin Domains: The Unit of Chromosome Organization. *Mol. Cell* **62**, 668–680 (2016).
28. Phillips-Cremins, J. E. & Corces, V. G. Chromatin insulators: linking genome organization to cellular function. *Mol. Cell* **50**, 461–474 (2013).
29. de Laat, W. & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* **502**, 499–506 (2013).
30. Vietri Rudan, M. *et al.* Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep.* **10**, 1297–1309 (2015).
31. Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137**, 1194–1211 (2009).

32. Nicodemi, M. & Pombo, A. Models of chromosome structure. *Curr. Opin. Cell Biol.* **28**, 90–95 (2014).
33. Alipour, E. & Marko, J. F. Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Res.* **40**, 11202–11212 (2012).
34. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E6456–65 (2015).
35. Yardimci, G. G. & Noble, W. S. Predictive model of 3D domain formation via CTCF-mediated extrusion. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 14404–14405 (2015).
36. Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep.* **15**, 2038–2049 (2016).
37. Goloborodko, A., Marko, J. F. & Mirny, L. A. Chromosome Compaction by Active Loop Extrusion. *Biophys. J.* **110**, 2162–2168 (2016).
38. Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N. & Mirny, L. A. Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E6697–E6706 (2018).
39. Nora, E. P. *et al.* Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169**, 930–944.e22 (2017).
40. Rao, S. S. P. *et al.* Cohesin Loss Eliminates All Loop Domains. *Cell* **171**, 305–320.e24 (2017).
41. Haarhuis, J. H. I. *et al.* The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. *Cell* **169**, 693–707.e14 (2017).
42. Richterova, J., Huraiova, B. & Gregan, J. Genome Organization: Cohesin on the Move. *Mol. Cell* **66**, 444–445 (2017).
43. Terakawa, T. *et al.* The condensin complex is a mechanochemical motor that translocates along DNA. *Science* **358**, 672–676 (2017).
44. Ganji, M. *et al.* Real-time imaging of DNA loop extrusion by condensin. *Science* **360**, 102–105 (2018).
45. Stigler, J., Çamdere, G. Ö., Koshland, D. E. & Greene, E. C. Single-Molecule Imaging Reveals a Collapsed Conformational State for DNA-Bound Cohesin. *Cell Rep.* **15**, 988–998 (2016).
46. Ulianov, S. V. *et al.* Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Res.* **26**, 70–84 (2016).
47. Hug, C. B., Grimaldi, A. G., Kruse, K. & Vaquerizas, J. M. Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription. *Cell* **169**, 216–228.e19 (2017).
48. Heinz, S. *et al.* Transcription Elongation Can Affect Genome 3D Structure. *Cell* **174**, 1522–1536.e22 (2018).

49. Graves, J. A. M. Evolution of vertebrate sex chromosomes and dosage compensation. *Nat. Rev. Genet.* **17**, 33–46 (2016).
50. Okamoto, I., Otte, A. P., Allis, C. D., Reinberg, D. & Heard, E. Epigenetic dynamics of imprinted X inactivation during early mouse development. *Science* **303**, 644–649 (2004).
51. Escamilla-Del-Arenal, M., da Rocha, S. T. & Heard, E. Evolutionary diversity and developmental regulation of X-chromosome inactivation. *Hum. Genet.* **130**, 307–327 (2011).
52. Martin, G. R. *et al.* X-chromosome inactivation during differentiation of female teratocarcinoma stem cells in vitro. *Nature* **271**, 329 (1978).
53. Augui, S. *et al.* Sensing X chromosome pairs before X inactivation via a novel X-pairing region of the Xic. *Science* **318**, 1632–1636 (2007).
54. Barakat, T. S. *et al.* RNF12 activates Xist and is essential for X chromosome inactivation. *PLoS Genet.* **7**, e1002001 (2011).
55. Navarro, P. *et al.* Molecular coupling of Tsix regulation and pluripotency. *Nature* **468**, 457–460 (2010).
56. Navarro, P., Moffat, M., Mullin, N. P. & Chambers, I. The X-inactivation trans-activator Rnf12 is negatively regulated by pluripotency factors in embryonic stem cells. *Hum. Genet.* **130**, 255–264 (2011).
57. Galupa, R. & Heard, E. X-chromosome inactivation: new insights into cis and trans regulation. *Curr. Opin. Genet. Dev.* **31**, 57–66 (2015).
58. Masui, O. *et al.* Live-cell chromosome dynamics and outcome of X chromosome pairing events during ES cell differentiation. *Cell* **145**, 447–458 (2011).
59. Navarro, P., Pichard, S., Ciaudo, C., Avner, P. & Rougeulle, C. Tsix transcription across the Xist gene alters chromatin conformation without affecting Xist transcription: implications for X-chromosome inactivation. *Genes Dev.* **19**, 1474–1484 (2005).
60. Jeppesen, P. & Turner, B. M. The inactive X chromosome in female mammals is distinguished by a lack of histone H4 acetylation, a cytogenetic marker for gene expression. *Cell* **74**, 281–289 (1993).
61. Heard, E. *et al.* Methylation of histone H3 at Lys-9 is an early mark on the X chromosome during X inactivation. *Cell* **107**, 727–738 (2001).
62. Chaumeil, J., Okamoto, I., Guggiari, M. & Heard, E. Integrated kinetics of X chromosome inactivation in differentiating embryonic stem cells. *Cytogenet. Genome Res.* **99**, 75–84 (2002).
63. da Rocha, S. T. *et al.* Jarid2 Is Implicated in the Initial Xist-Induced Targeting of PRC2 to the Inactive X Chromosome. *Mol. Cell* **53**, 301–316 (2014).
64. Chaumeil, J., Le Baccon, P., Wutz, A. & Heard, E. A novel role for Xist RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced. *Genes Dev.* **20**, 2223–2237 (2006).

65. Rego, A., Sinclair, P. B., Tao, W., Kireev, I. & Belmont, A. S. The facultative heterochromatin of the inactive X chromosome has a distinctive condensed ultrastructure. *J. Cell Sci.* **121**, 1119–1127 (2008).
66. Engreitz, J. M. *et al.* The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science* **341**, 1237973 (2013).
67. Simon, M. D. *et al.* High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. *Nature* **504**, 465–469 (2013).
68. Giorgetti, L. *et al.* Structural organization of the inactive X chromosome in the mouse. *Nature* **535**, 575–579 (2016).
69. Wang, C.-Y., Jégu, T., Chu, H.-P., Oh, H. J. & Lee, J. T. SMCHD1 Merges Chromosome Compartments and Assists Formation of Super-Structures on the Inactive X. *Cell* **174**, 406–421.e25 (2018).
70. Lucchesi, J. C. & Kuroda, M. I. Dosage compensation in Drosophila. *Cold Spring Harb. Perspect. Biol.* **7**, (2015).
71. Hilfiker, A., Hilfiker-Kleiner, D., Pannuti, A. & 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 J.* **16**, 2054–2060 (1997).
72. Franke, A. & Baker, B. S. The rox1 and rox2 RNAs are essential components of the compensasome, which mediates dosage compensation in Drosophila. *Mol. Cell* **4**, 117–122 (1999).
73. Li, F., Schiemann, A. H. & Scott, M. J. Incorporation of the noncoding roX RNAs alters the chromatin-binding specificity of the Drosophila MSL1/MSL2 complex. *Mol. Cell. Biol.* **28**, 1252–1264 (2008).
74. Fauth, T., Müller-Planitz, F., König, C., Straub, T. & Becker, P. B. The DNA binding CXC domain of MSL2 is required for faithful targeting the Dosage Compensation Complex to the X chromosome. *Nucleic Acids Res.* **38**, 3209–3221 (2010).
75. Villa, R., Schauer, T., Smialowski, P., Straub, T. & Becker, P. B. PionX sites mark the X chromosome for dosage compensation. *Nature* **537**, 244–248 (2016).
76. Hallacli, E. *et al.* Msl1-mediated dimerization of the dosage compensation complex is essential for male X-chromosome regulation in Drosophila. *Mol. Cell* **48**, 587–600 (2012).
77. Buscaino, A. *et al.* MOF-regulated acetylation of MSL-3 in the Drosophila dosage compensation complex. *Mol. Cell* **11**, 1265–1277 (2003).
78. Sural, T. H. *et al.* The MSL3 chromodomain directs a key targeting step for dosage compensation of the Drosophila melanogaster X chromosome. *Nat. Struct. Mol. Biol.* **15**, 1318–1325 (2008).
79. Meller, V. H. *et al.* Ordered assembly of roX RNAs into MSL complexes on the dosage-compensated X chromosome in Drosophila. *Curr. Biol.* **10**, 136–143 (2000).
80. Kelley, R. L. *et al.* Epigenetic spreading of the Drosophila dosage compensation complex from roX RNA genes into flanking chromatin. *Cell* **98**, 513–522 (1999).

81. Akhtar, A. & Becker, P. B. Activation of transcription through histone H4 acetylation by MOF, an acetyltransferase essential for dosage compensation in Drosophila. *Mol. Cell* **5**, 367–375 (2000).
82. Conrad, T., Cavalli, F. M. G., Vaquerizas, J. M., Luscombe, N. M. & Akhtar, A. Drosophila dosage compensation involves enhanced Pol II recruitment to male X-linked promoters. *Science* **337**, 742–746 (2012).
83. Larschan, E. *et al.* X chromosome dosage compensation via enhanced transcriptional elongation in Drosophila. *Nature* **471**, 115–118 (2011).
84. Ferrari, F. *et al.* ‘Jump Start and Gain’ Model for Dosage Compensation in Drosophila Based on Direct Sequencing of Nascent Transcripts. *Cell Rep.* **5**, 629–636 (2013).
85. Ramírez, F. *et al.* High-Affinity Sites Form an Interaction Network to Facilitate Spreading of the MSL Complex across the X Chromosome in Drosophila. *Mol. Cell* **60**, 146–162 (2015).
86. Schauer, T. *et al.* Chromosome topology guides the Drosophila Dosage Compensation Complex for target gene activation. *EMBO Rep.* (2017). doi:10.15252/embr.201744292
87. Berletch, J. B., Yang, F., Xu, J., Carrel, L. & Disteche, C. M. Genes that escape from X inactivation. *Hum. Genet.* **130**, 237–245 (2011).
88. Samata, M. & Akhtar, A. Dosage Compensation of the X Chromosome: A Complex Epigenetic Assignment Involving Chromatin Regulators and Long Noncoding RNAs. *Annu. Rev. Biochem.* **87**, 323–350 (2018).
89. Smith, E. R. *et al.* A human protein complex homologous to the Drosophila MSL complex is responsible for the majority of histone H4 acetylation at lysine 16. *Mol. Cell. Biol.* **25**, 9175–9188 (2005).
90. Rea, S., Xouri, G. & Akhtar, A. Males absent on the first (MOF): from flies to humans. *Oncogene* **26**, 5385–5394 (2007).
91. Nguyen, D. K. & Disteche, C. M. Dosage compensation of the active X chromosome in mammals. *Nat. Genet.* **38**, 47–53 (2006).
92. Deng, X. *et al.* Evidence for compensatory upregulation of expressed X-linked genes in mammals, *Caenorhabditis elegans* and *Drosophila melanogaster*. *Nat. Genet.* **43**, 1179–1185 (2011).
93. Deng, X. *et al.* Mammalian X upregulation is associated with enhanced transcription initiation, RNA half-life, and MOF-mediated H4K16 acetylation. *Dev. Cell* **25**, 55–68 (2013).
94. Gupta, A. *et al.* The mammalian ortholog of Drosophila MOF that acetylates histone H4 lysine 16 is essential for embryogenesis and oncogenesis. *Mol. Cell. Biol.* **28**, 397–409 (2008).
95. Lam, K. C. *et al.* The NSL complex regulates housekeeping genes in Drosophila. *PLoS Genet.* **8**, e1002736 (2012).

96. Chelmicki, T. *et al.* MOF-associated complexes ensure stem cell identity and Xist repression. *Elife* **3**, e02024 (2014).
97. Sheikh, B. N. *et al.* MOF maintains transcriptional programs regulating cellular stress response. *Oncogene* **35**, 2698–2710 (2016).
98. Schütz, P. *et al.* Crystal structure of human RNA helicase A (DHX9): structural basis for unselective nucleotide base binding in a DEAD-box variant protein. *J. Mol. Biol.* **400**, 768–782 (2010).
99. Chakraborty, P. & Grosse, F. Human DHX9 helicase preferentially unwinds RNA-containing displacement loops (R-loops) and G-quadruplexes. *DNA Repair* **10**, 654–665 (2011).
100. Jain, A. *et al.* DHX9 helicase is involved in preventing genomic instability induced by alternatively structured DNA in human cells. *Nucleic Acids Res.* **41**, 10345–10357 (2013).
101. Aktaş, T. *et al.* DHX9 suppresses RNA processing defects originating from the Alu invasion of the human genome. *Nature* **544**, 115–119 (2017).
102. Pop, M. & Salzberg, S. L. Bioinformatics challenges of new sequencing technology. *Trends Genet.* **24**, 142–149 (2008).
103. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
104. Egelhofer, T. A. *et al.* An assessment of histone-modification antibody quality. *Nat. Struct. Mol. Biol.* **18**, 91–93 (2011).
105. Rothbart, S. B. *et al.* An Interactive Database for the Assessment of Histone Antibody Specificity. *Mol. Cell* **59**, 502–511 (2015).
106. Skene, P. J. & Henikoff, S. A simple method for generating high-resolution maps of genome-wide protein binding. *Elife* **4**, e09225 (2015).
107. Marshall, O. J., Southall, T. D., Cheetham, S. W. & Brand, A. H. Cell-type-specific profiling of protein-DNA interactions without cell isolation using targeted DamID with next-generation sequencing. *Nat. Protoc.* **11**, 1586–1598 (2016).
108. Lara-Astiaso, D. *et al.* Immunogenetics. Chromatin state dynamics during blood formation. *Science* **345**, 943–949 (2014).
109. van Galen, P. *et al.* A Multiplexed System for Quantitative Comparisons of Chromatin Landscapes. *Mol. Cell* **61**, 170–180 (2016).
110. Arrigoni, L. *et al.* Ultra-parallel ChIP-seq by barcoding of intact nuclei. *bioRxiv* 276469 (2018). doi:10.1101/276469
111. Nakato, R. & Shirahige, K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. *Brief. Bioinform.* **18**, 279–290 (2017).
112. Rozowsky, J. *et al.* PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.* **27**, 66–75 (2009).

113. Orlando, D. A. *et al.* Quantitative ChIP-Seq normalization reveals global modulation of the epigenome. *Cell Rep.* **9**, 1163–1170 (2014).
114. Liang, K. & Keleş, S. Normalization of ChIP-seq data with control. *BMC Bioinformatics* **13**, 199 (2012).
115. Shao, Z., Zhang, Y., Yuan, G.-C., Orkin, S. H. & Waxman, D. J. MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. *Genome Biol.* **13**, R16 (2012).
116. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–91 (2014).
117. Steinhauser, S., Kurzawa, N., Eils, R. & Herrmann, C. A comprehensive comparison of tools for differential ChIP-seq analysis. *Brief. Bioinform.* **17**, 953–966 (2016).
118. Bonhoure, N. *et al.* Quantifying ChIP-seq data: a spiking method providing an internal reference for sample-to-sample normalization. *Genome Res.* **24**, 1157–1168 (2014).
119. Orlando, D. A. *et al.* Quantitative ChIP-Seq normalization reveals global modulation of the epigenome. *Cell Rep.* **9**, 1163–1170 (2014).14
120. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–5 (2016).
121. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
122. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.* **38**, 1348–1354 (2006).
123. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16**, 1299–1309 (2006).
124. Belmont, A. S. Large-scale chromatin organization: the good, the surprising, and the still perplexing. *Curr. Opin. Cell Biol.* **26**, 69–78 (2014).
125. Williamson, I. *et al.* Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. *Genes Dev.* **28**, 2778–2791 (2014).
126. Giorgetti, L. & Heard, E. Closing the loop: 3C versus DNA FISH. *Genome Biol.* **17**, 215 (2016).
127. Beagrie, R. A. *et al.* Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543**, 519–524 (2017).
128. Quinodoz, S. A. *et al.* Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell* **174**, 744–757.e24 (2018).
129. Redolfi, J. *et al.* Modeling of DNA methylation in cis reveals principles of chromatin folding in vivo in the absence of crosslinking and ligation. *bioRxiv* 407031 (2018). doi: 10.1101/407031

130. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458–472 (2012).
131. Hu, M. *et al.* HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics* **28**, 3131–3133 (2012).
132. Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.* **43**, 1059–1065 (2011).
133. Schmitt, A. D., Hu, M. & Ren, B. Genome-wide mapping and analysis of chromosome architecture. *Nat. Rev. Mol. Cell Biol.* **17**, 743–755 (2016).
134. Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* **9**, 999–1003 (2012).
135. Knight, P. A. & Ruiz, D. A fast algorithm for matrix balancing. *IMA J. Numer. Anal.* **33**, 1029–1047 (2013).
136. Naumova, N. *et al.* Organization of the mitotic chromosome. *Science* **342**, 948–953 (2013).
137. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95–98 (2016).
138. Lévy-Leduc, C., Delattre, M., Mary-Huard, T. & Robin, S. Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics* **30**, i386–92 (2014).
139. Shin, H. *et al.* TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res.* **44**, e70 (2016).
140. Filippova, D., Patro, R., Duggal, G. & Kingsford, C. Identification of alternative topological domains in chromatin. *Algorithms Mol. Biol.* **9**, 14 (2014).
141. Zhan, Y. *et al.* Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. *Genome Res.* **27**, 479–490 (2017).
142. Dali, R. & Blanchette, M. A critical assessment of topologically associating domain prediction tools. *Nucleic Acids Res.* **45**, 2994–3005 (2017).
143. Hrdlickova, R., Toloue, M. & Tian, B. RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip. Rev. RNA* **8**, (2017).
144. Pritchard, C. C., Cheng, H. H. & Tewari, M. MicroRNA profiling: approaches and considerations. *Nat. Rev. Genet.* **13**, 358–369 (2012).
145. Schwalb, B. *et al.* TT-seq maps the human transient transcriptome. *Science* **352**, 1225–1228 (2016).
146. Memczak, S. *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495**, 333–338 (2013).
147. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).

148. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
149. Finotello, F. & Di Camillo, B. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Brief. Funct. Genomics* **14**, 130–142 (2015).
150. Lovén, J. *et al.* Revisiting global gene expression analysis. *Cell* **151**, 476–482 (2012).
151. Jiang, L. *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21**, 1543–1551 (2011).
152. Gao, Y. & Zhao, F. Computational Strategies for Exploring Circular RNAs. *Trends Genet.* **34**, 389–400 (2018).
153. Bass, B. L. RNA editing by adenosine deaminases that act on RNA. *Annu. Rev. Biochem.* **71**, 817–846 (2002).
154. Carninci, P. *et al.* High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics* **37**, 327–336 (1996).
155. Kawaji, H. *et al.* Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. *Genome Res.* **24**, 708–717 (2014).
156. Kodzius, R. *et al.* CAGE: cap analysis of gene expression. *Nat. Methods* **3**, 211–222 (2006).
157. Batut, P., Dobin, A., Plessy, C., Carninci, P. & Gingeras, T. R. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res.* **23**, 169–180 (2013).
158. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38**, 626–635 (2006).
159. Frith, M. C. *et al.* A code for transcription initiation in mammalian genomes. *Genome Res.* **18**, 1–12 (2008).
160. Ohmiya, H. *et al.* RECLU: a pipeline to discover reproducible transcriptional start sites and their alternative regulation using capped analysis of gene expression (CAGE). *BMC Genomics* **15**, 269 (2014).
161. Ramírez, F. *et al.* High-Affinity Sites Form an Interaction Network to Facilitate Spreading of the MSL Complex across the X Chromosome in Drosophila. *Mol. Cell* **60**, 146–162 (2015).15
162. Li, L. *et al.* Widespread rearrangement of 3D chromatin organization underlies polycomb-mediated stress-induced silencing. *Mol. Cell* **58**, 216–231 (2015).
163. Cubeñas-Potts, C. *et al.* Different enhancer classes in Drosophila bind distinct architectural proteins and mediate unique chromatin interactions and 3D architecture. *Nucleic Acids Res.* (2016). doi:10.1093/nar/gkw1114
164. Ramírez, F. *et al.* High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.* **9**, 189 (2018).

165. da Veiga Leprevost, F. *et al.* BioContainers: an open-source and community-driven framework for software standardization. *Bioinformatics* **33**, 2580–2582 (2017).
166. Lott, S. E. *et al.* Noncanonical compensation of zygotic X transcription in early *Drosophila melanogaster* development revealed through single-embryo RNA-seq. *PLoS Biol.* **9**, e1000590 (2011).
167. Wolff, J. *et al.* Galaxy HiCExplorer: a web server for reproducible Hi-C data analysis, quality control and visualization. *Nucleic Acids Res.* **46**, W11–W16 (2018).
168. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
169. Day, D. S., Luquette, L. J., Park, P. J. & Kharchenko, P. V. Estimating enrichment of repetitive elements from high-throughput sequence data. *Genome Biol.* **11**, R69 (2010).
170. Novák, P., Neumann, P., Pech, J., Steinhaisl, J. & Macas, J. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* **29**, 792–793 (2013).
171. Athanasiadis, A., Rich, A. & Maas, S. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biol.* **2**, e391 (2004).
172. Kim, D. D. Y. *et al.* Widespread RNA editing of embedded alu elements in the human transcriptome. *Genome Res.* **14**, 1719–1725 (2004).
173. Bazak, L. *et al.* A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. *Genome Res.* **24**, 365–376 (2014).
174. Daniel, C., Silberberg, G., Behm, M. & Öhman, M. Alu elements shape the primate transcriptome by cis-regulation of RNA editing. *Genome Biol.* **15**, R28 (2014).
175. Aktaş, T. *et al.* DHX9 suppresses RNA processing defects originating from the Alu invasion of the human genome. *Nature* **544**, 115–119 (2017).¹⁶
176. Conti, L. *et al.* Niche-independent symmetrical self-renewal of a mammalian tissue stem cell. *PLoS Biol.* **3**, e283 (2005).
177. Arrigoni, L. *et al.* Ultra-parallel ChIP-seq by barcoding of intact nuclei. *bioRxiv* 276469 (2018). doi:1710.1101/276469
178. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
179. Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 1827–1831 (1992).
180. Krueger, F. & Andrews, S. R. *SNPsplit*: Allele-specific splitting of alignments between genomes with known SNP genotypes. *F1000Res.* **5**, 1479 (2016).
181. Degner, J. F. *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**, 3207–3212 (2009).

182. Bhardwaj, V. *et al.* snakePipes enable flexible, scalable and integrative epigenomic analysis. *bioRxiv* 407312 (2018). doi:10.1101/407312

183. Valsecchi, C. I. K. *et al.* Facultative dosage compensation of developmental genes on autosomes in Drosophila and mouse embryonic stem cells. *Nat. Commun.* **9**, 3626 (2018).