TECHNISCHE UNIVERSITÄT MÜNCHEN
Fachgebiet für Bioinformatik

ANALYSIS OF SPATIAL CHROMATIN ORGANIZATION AND ITS EVOLUTIONARY CONSERVATION

STEFANIE KAUFMANN

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. S. Scherer
Prüfer der Dissertation:	
	1. Univ.-Prof. Dr. D. Frischmann
	2. Jun.-Prof. Dr. C. Friedel (Ludwig-Maximilians-Universität München)

Die Dissertation wurde am 10.11.2014 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 16.01.2015 angenommen.

Stefanie Kaufmann: *Analysis of Spatial Chromatin Organization and its Evolutionary Conservation,* A Dissertation in Bioinformatics, © Oktober 2014

SUPERVISOR:
Prof. Dr. Dmitrij Frishman

The greatest phrase in science,
the one that heralds new discovery,
isn't *Eureka!*
but rather *That's funny*.

— Isaac Asimov

Dedicated to my parents Jutta and Helmut Kaufmann.

ABSTRACT

Since Gregor Mendel's first tries to understand Genetics, the knowledge about DNA and its functions has been greatly increased. Today we know that not only its base composition contains valuable information, but also the DNA's structure, both epigenetically in the sense of histone modifications, and epigenomically as in its three-dimensional fold. This thesis aims to analyse the genomes of mammalian organisms with respect to their linear and three-dimensional structure, their relationship and evolution. By investigating a large amount of linear properties such as repeats, also called features, in the context of the genome, we confirm the domain-like structure of the human genome and inter-dependencies between features and eu- or heterochromatin.

In the second part of this work, we successfully transform published Hi-C data from *H. sapiens* and *M. musculus* into a bias-free high-quality inter-chromosomal interaction network. We show that these scale-free contact networks share similar characteristics in both species, such as presumably very flexible, highly interactive regions on chromosome Y, a higher contact density on short, gene-rich chromosomes and a positive association between spatial proximity and functional similarity. However, while intra-chromosomal contacts are largely conserved between human and mouse, individual inter-chromosomal contacts are not, and the feature composition of interacting segments differs vastly between them.

Because understanding genome evolution is important in distinguishing functional from non-functional properties, we focus on linear genome rearrangements in the last part of this thesis. Respecting the hierarchical structure of mammalian genomes, we develop a new tool termed *SyntenyMapper* to identify micro-rearranged regions within large conserved regions ("synteny regions"). We show that our tool delivers more exact results than comparable software and that it can be used to draw both general information on mammalian evolution and to analyse individual genome regions.

Altogether, these different aspects of genome structure and evolution show that it is important to understand all of them in detail and especially their complex interplay. Our results show that the two- and three-dimensional genome organisation of human and mouse is similar only on the functional level, while individual contacts are disrupted due to linear rearrangements. Though linear feature composition and genome fold are highly inter-dependent, this relationship is largely species-specific. It thus appears that the inter-chromosomal interactome is not strongly conserved between mammalian species.

ZUSAMMENFASSUNG

Seit Gregor Mendels ersten Versuchen, Genetik zu verstehen, konnte die wissenschaftliche Gemeinde zahlreiches neues Wissen über DNS und ihre Funktionen zusammentragen. Aktuell wissen wir, dass nicht nur die Basenpaar-Sequenz wertvolle Informationen enthält, sondern auch die Struktur der DNS, sowohl epigenetisch im Sinne von Histonmodifikation, als auch epigenomisch, also die räumliche Faltung. Diese Dissertation hat zum Ziel, die Genome von Säugetieren in Bezug auf ihre lineare und drei-dimensionale Struktur, sowie die Abhängigkeiten zwischen diesen Strukturen und ihre Evolution zu untersuchen. Indem wir eine große Menge linearer Elemente wie Repeats im Genom untersuchen, bestätigen wir die Domänen-Struktur des menschlichen Genoms und die Abhängigkeiten zwischen diesen so-genannten Features und Eu- bzw. Heterochromatin.

Im zweiten Teil gelingt es uns, publizierte Hi-C Daten aus *H. sapiens* und *M. musculus* in unverfälschte hoch-qualitative inter-chromosomale Interaktions-Netzwerke zu übersetzen. Wir zeigen, dass diese skalenfreien Kontakt-Netzwerke in beiden Spezies ähnliche Eigenschaften haben, wie beispielsweise voraussichtlich sehr flexible hochinteraktive Regionen auf Chromosom Y. Weitere Ähnlichkeiten liegen in einer erhöhten Kontakthäufigkeit auf kurzen, gen-reichen Chromosomen und eine schwache positive Korrelation zwischen räumlicher Nähe und funktioneller Ähnlichkeit. Allerdings zeigen wir auch, dass inter-chromosomale Kontakte im Gegensatz zu intra-chromosomalen nicht zwischen Mensch und Maus konserviert sind, ebenso wie die Element-Zusammensetzung von interagierenden Segmenten sich stark zwischen den beiden Spezies unterscheidet.

Weil es nötig ist, Genomevolution zu verstehen um Rückschlüsse auf die funktionalen Eigenschaften zu ziehen, betrachten wir im letzten Teil lineare Genom-Umordnungen. Unter Berücksichtigung der hierarchischen Struktur von Säugetier-Genomen entwickeln wir ein neues Software-Tool namens *SyntenyMapper* um kleine umgeordnete Regionen innerhalb großer konservierter (Synteny-)Regionen zu identifizieren. Wir zeigen, dass unser Tool exakter arbeitet als vergleichbare Software, und dass es sowohl zur allgemeinen Analyse von Säugetier-Evolution verwendet werden kann, als auch zur Untersuchung individueller Genomregionen.

Zusammengenommen zeigen diese verschiedenen Aspekte der Genom-Struktur und -Evolution, dass es wichtig ist, sie vollständig und vor allem ihr komplexes Zusammenspiel zu verstehen. Unsere Ergebnisse zeigen, dass sich die zwei- und drei-dimensionale Genomorganisation von Mensch und Maus nur auf funktioneller Ebene ähnelt, während einzelne inter-chromosomale Kontakte durch Chromosomen-Umordnungen zerstört wurden. Obwohl die lineare Feature-Zusammensetzung und die Faltung des Genoms stark voneinander abhängen, ist dieser Zusammenhang oft spezies-bedingt. Daraus können wir schließen, dass das inter-chromosomale Interaktom von Säugetieren nicht stark konserviert ist.

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications:

S. Kaufmann and D. Frishman. Analysis of micro-rearrangements in 25 eukaryotic species pairs by SyntenyMapper. *PLOS ONE*, accepted, 2014.

S. Kaufmann, C. Fuchs, M. Gonik, EE. Khrameeva, AA. Mironov and D. Frishman. Inter-chromosomal contact networks provide insight into mammalian chromatin organization. *Nucleic Acids Research*, submitted, 2014.

Poster: S. Kaufmann and D. Frishman. Inter-chromosomal contact networks provide insight into mammalian chromatin organization. *Spetses Summer School on Chromatin and Systems Biology*, 2013.

Poster: S. Kaufmann and D. Frishman. Creation and inter-species mapping of physical gene interaction networks. *RECESS Retreat*, 2013.

Poster: S. Kaufmann and D. Frishman. Evolutionary Conservation of Spatial Chromatin Organization in Mouse and Human. *RECESS Retreat*, 2012.

DANKSAGUNG

Eine Promotion zu meistern ist ein langer und oft nicht ganz einfacher Weg, auf dem man oft auf die Hilfe und Unterstützung anderer Menschen angewiesen ist. Auf meinem Weg haben mich viele begleitet, denen ich an dieser Stelle meinen Dank aussprechen möchte.

Eine wichtige Rolle in der Entwicklung meiner fachlichen Fähigkeiten spielte Prof. Caroline Friedel, die mir früh eine Hiwi-Stelle anvertraute und mich direkt an spannenden Projekten mitarbeiten ließ. In dieser Zeit konnte ich meine Programmierfähigkeiten ausbauen und erste Erfahrungen im wissenschaftlichen Arbeiten sammeln, die mir während der Promotion sehr geholfen haben. Auch Prof. Dmitrij Frishman gab mir schon für die Bachelor- und Masterarbeit spannende Projekte, und bot mir auch für die Doktorarbeit ein aktuelles und interessantes Thema. Unter seiner Betreuung bei der Promotion habe ich viel gelernt, und möchte mich an dieser Stelle vor allem für die ausführlichen Korrekturen meiner Paper-Entwürfe bedanken.

Mein Dank gilt auch dem RECESS Graduiertenkolleg und der DFG für die Finanzierung, vor allem aber auch den fachlichen und internationalen Austausch und viele neue Erfahrungen. Besonders danken möchte ich dabei auch Claudia Luksch und Leonie Corry, die mir oft bei bürokratischen Problemen halfen und auch träge Arbeitstage angenehmer gestalteten.

Ohne die fachliche Hilfe meiner Kollegen hätte ich oft sehr viel länger über dem ein oder anderen Problem gebrütet. Hervorheben möchte ich dabei die Hilfe von Jonathan, der sich oft viel Zeit genommen hat und mit seinen wertvollen Tipps nicht wenige Ideen zu meiner Doktorarbeit beigetragen hat. Auch Kerstin stand mir immer gern mit ihrem Rat zur Seite, wenn ich wieder einmal das Whiteboard mit wirren Graphen überzog. Auch die anderen waren immer für Diskussionen bereit, und, auch nicht unwichtig, brachten regelmäßig Kuchen und andere süße Motivationshelferchen, wofür ich mich bedanken möchte.

Auch meine Freunde haben in den letzten drei Jahren eine wichtige Rolle gespielt, und mich mit Spieleabenden und langen Gesprächen gerne abgelenkt, wenn ich mal viel zu tun hatte. Mein Dank geht deshalb an Eva, Steffi, Kerstin, Flo und Robert, die immer gerne Kontrastprogramm zur Promotion boten. Allen voran möchte ich aber Denis danken, der mich nicht nur mit ausgezeichnetem Essen und vielen Scrabble-Partien unterhalten hat, sondern dessen Intellekt und Whiteboard mir immer zur Verfügung standen für abendfüllende Bioinformatik- Diskussionen.

Zuletzt möchte ich meiner Familie danken, die mich immer unterstützt hat, obwohl ein derart langer Ausbildungsweg in unserer Familie ein absolutes Novum darstellte. Zwar wissen sie immer noch nicht so genau, woran ich eigentlich forsche (meine Schwester fasst es recht kompetent zusammen mit 'Irgendwas mit DNA'), aber sie zeigten immer Interesse und hatten nie Zweifel, dass ich es schaffen würde. Deshalb geht an dieser Stelle mein besonderer Dank an meine Eltern und meine Schwestern, die hoffentlich ein bisschen stolz auf mich sind, und auch an meine Großeltern, die mich hoffentlich bald offiziell als 'Frau Doktor' begrüßen dürfen.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

3C Chromatin Conformation Capture

CHIP Chromatin Immunoprecipitation

CTCF CCCTC-Binding Factor

DAM DNA methyltransferase

ESC Embryonic stem cell

FISH Fluorescent In-Situ Hybridization

GO Gene Ontology

LADS Lamina Associated Domains

LINE Long Interspersed Nuclear Element

LCR Locus Control Region

LNCRNA long non-coding RNA

LTR Long Tandem Repeats

NGS Next Generation Sequencing

PCR Polymerase Chain Reaction

RMSE Root Mean Square Error

RTD Replication Timing Domains

SINE Short Interspersed Nuclear Element

SNP Single Nucleotide Polymorphism

SVM Support Vector Machine

TADS Topologically associated domains

TCC Tethered Conformation Capture

TFBS Transcription Factor Binding Sites

WGD Whole Genome Duplication

Part I

INTRODUCTION

Recently, spatial chromatin organization has emerged as another possible level of genome regulation. Hi-C, a high-throughput conformation capture method applicable to whole genomes, has helped uncover a vast net of contacts within as well as between different chromosomes. This introduction gives an overview of the biological background of spatial genome organization and the experimental methods used for its detection.

1 INTRODUCTION

The central dogma of molecular biology, which states that DNA is transcribed into RNA which in turn is translated into protein, has somewhat suffered during the last century. We know now that processes in the cell are more complex and regulated through different machineries.

Chromosomes fold into non-random structures inside the nucleus.

Regulation of gene expression is vital to control cellular reactions to extrinsic or intrinsic factors and even to define differential gene profiles in different tissues. Besides the products of DNA sequences themselves, specifically transcription factors or silencing proteins, it is well known that epigenetic regulations through chemical modifications of nucleotides and histones influence gene regulation (e.g. [16]). Recently, it has become clear that spatial organization of chromosomes reflects a higher order of epigenetic regulation.

Chromosomes are usually viewed as long, linear stretches of DNA with different degrees of compression. However, it is now obvious that chromatin fibres in the nucleus are organized to a great extent. So far, it is not clear whether this spatial organization is a side effect of genome regulation or its cause, but in either case the so-called chromosome interactome might bring new answers to molecular biology.

Recent methodological advances have allowed researchers to investigate the higher-order genome structure globally and in greater detail than previously possible. Among them are chromatin conformation capture methods that can be applied to regions of interest or whole genomes, and applications of these and other methods described in the second part of this introduction (page 15) have led to new knowledge in the field of epigenomics and genome structure. The following sections will give an overview of the general mechanisms behind large-scale spatial chromosome organization, followed by description of small-scale interactions and the processes forming them.

1.1 LARGE SCALE SPATIAL ORGANIZATION OF THE GENOME

To fit into the rather small space of only approximately 6 μm in mammalian nuclei [1], chromosomes must bend and flex. For instance, each human cell comprises DNA of a total length of circa two meters [1] and has to be tightly folded to fit into the nucleus. However, this does not happen in a random fashion like one might expect. Instead we can find some organization regarding the overall folding

of single chromosomes as well as the arrangement of multiple DNA fibres.

Theories and findings on chromatin folding models, higher-order chromosome structures and organization into sub-compartments of the nucleus are discussed in the following sections.

Single chromosomes might fold into a fractal globule structure

Evidence supports the treatment of chromatin fibres as polymers for modelling [120]. Based on this observation, various models for single chromosome folding have been proposed. One of these suggests that smaller parts of chromosomes could fold into the so-called "equilibrium globe" (compare Figure 1 A), which is a compact globular arrangement. However, during the formation of this structure, the chromosome has to form many knots. This configuration appears to be rather impractical, and other models propose organizational forms in which knots are not as frequent. Among these is the "fractal globule", another globular folding which consists of smaller, also globular fractions (compare Figure 1 A). The formation process is suggested to start with the linear polymer which then folds into smaller globules like 'beads on a string', before these monomers arrange to form a large globular form [120, 179].

A fractal globule is formed by the hierarchical process of local regions collapsing in on themselves.

This structure has other advantages besides lack of knots: it is easier to fold into, and certainly easier to unfold again. Since chromosomes do not maintain their structures throughout the entire life cycle of a cell but have to compress into condensed forms during mitosis or decompress during gene activation, unfolding is greatly facilitated if there are no knots in the structure.

Experimental data of intra-chromosomal contacts at a resolution of 1 mega base pair (Mb) support the fractal globule model [120], which is thought to be caused by a set of interactions between genes and regulatory elements in close proximity that initially lead to collapses of chromosomal regions all over the chromatin fibre [179]. However, there might be even more order in this process.

In Chromatin Conformation Capture (3C) experiments, a 1 Mb resolution *means that (sparse) data was summarized over 1 Mb*

Besides short-range interactions along the chromosomes that lead to crumpling of the fibres, there have also been shown to exist weaker long-range interactions that span tens of Mb [181, 188, 120]. Sanyal et al suggest that these weaker interactions lead to an aggregation of active chromatin domains and similar clustering of inactive genes [179]. These clustered regions could then associate with other active and inactive regions along the chromosome, respectively, in a similar fashion like micelles are created through the hydrophobic effect. However, unlike the building of micelles, it is not yet clear what exactly guides these long-range interactions and possible co-localizations of similar chromosomal domains into the globular structures.

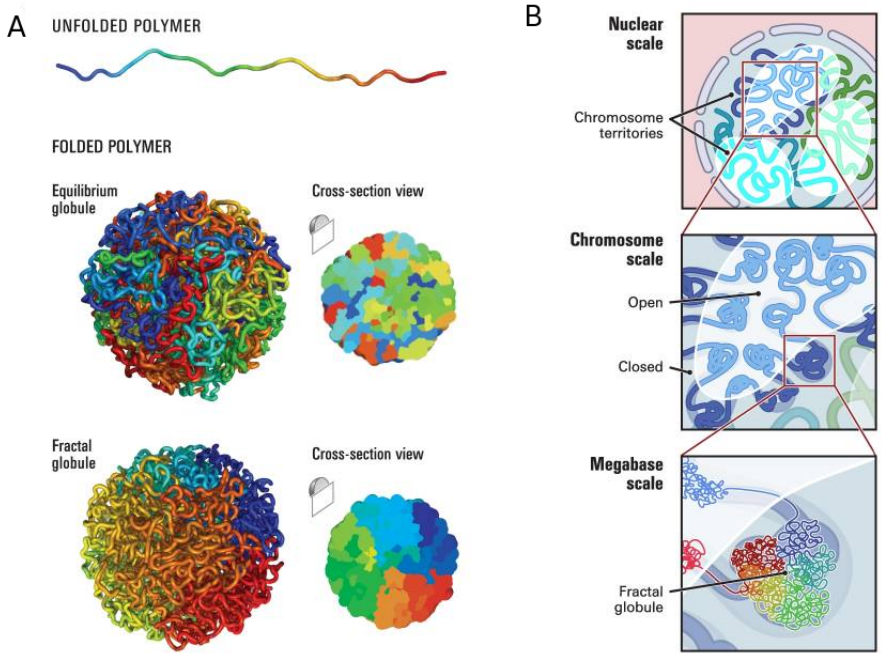


Figure 1: Models for chromosome folding (Taken from Dekker et al. [120]). **A:** Illustration of two models for chromosome folding. Evidence indicates that chromosomes behave like polymers, leading to the suggestion of globular conformations. The presented model of an equilibrium globe where the chromatin fibre is highly entangled and contains many knots is believed to be the less probable version. A fractal globule that is formed by smaller globular monomers built from adjacent chromosomal regions facilitates refolding and unfolding and is considered to be more probable.
B: Large-scale organization of chromosomes in the nucleus. The model suggests that folding into fractal globules happens along all chromosomes and that globules containing active regions naturally aggregate. This leads to larger compartments of active and inactive chromatin.

Nuclear organization of multiple chromosomes reflects clusters of active and inactive genetic loci

If we assume that chromosomes fold into globular structures, we can go on to ask how these are organized in the nuclear space. The most interesting question is whether they lie randomly in the nucleus like a set of yarn balls in a basket or if there is a higher order in their interplay.

In the previous section, we discussed the hypothesis of Sanyal et al. that states smaller active and inactive chromosomal regions might cluster in space. This eventually leads to a higher level structure of active and inactive compartments. Figure 1 B illustrates how fractal globules along each chromosome could co-associate with other such globules of the same or different chromosomes to form regions in the nucleus that are characterized by similar compactness of the chromatin. Experimental data from Lieberman-Aiden et al. confirmed the two-compartment structure of the nucleus [120] (see section 1.3.2.4 for a detailed description). The data show that chromosome regions in the active compartment, termed A, preferentially interact with other active regions, while regions from the inactive compartment, B, also interact mainly with themselves. Dekker et al. also showed that this compartmentalization is cell-type-specific [120].

Chromatin is divided into active and inactive subnuclear clusters

It further appears that chromosomes do not randomly associate with other fibres, but rather have a specific location in the nucleus. These so-called chromosome territories have already been described as early as 1885 [167], though the term itself was not introduced until 1909 [19]. Many studies since then have confirmed the non-random arrangement of chromosomes into stable structures at fixed relative positions in the nucleus [35].

Based on experimental data analysing the interaction of chromatin with the nuclear lamina, a two-wheel model emerges [95]. The main feature of this model is the localization of closed chromatin compartments close to the lamina (described in more detail in section 1.1.3) or the nucleolus, while open and thus active chromosome regions are embedded in between. These active regions are also thought to contain transcription factories and be specialized for expression. The localization seems to further depend on gene-richness of the chromosomes, with gene-dense chromosomes like chromosome 19 being located centrally and gene-poor chromosomes being located closer to the lamina, independent of their size [36]. In cells with flattened nuclei, chromosome size apparently does matter, where small chromosomes can arrange themselves more easily around the nucleolus, while large chromosomes fit more comfortably in the periphery of the nuclear envelope [18].

This localization is not fixed; changes in the peripheral location of chromatin regions in so-called Lamina Associated Domains (LADs)

during spermatogenesis [57] and adipocyte differentiation [111] suggest a dependency between chromosomal radial positioning and expression patterns [95]. Genes activated in the course of differentiation tend to gain distance to the lamina, while previously active and later repressed genes decrease their distance to the nuclear envelope [159]. This dependency can theoretically go either way, with the localization of chromosome regions determining their activity but also vice versa. With the current state of knowledge, neither of these hypotheses can be discarded or claimed to be correct. It is however clear that radial and relational positioning is not random and cell-type specific [155, 156], illustrating the importance of chromatin organization for cell expression or vice versa.

1.1.1 *The role of nuclear sub-compartments*

Besides active and inactive compartments, the nucleus also comprises other so called sub-compartments with specialized roles and properties. Among them are for example nuclear speckles, which are enriched with pre-mRNA splicing factors [172] and the nucleolus. It is well known that the nucleolus recruits nucleolar organizing regions (NORs) for ribosome biogenesis [148]. The question arises whether chromatin organization creates these nuclear sub-compartments, or is maybe even created through recruitments to these sub-compartments, similar to NORs. One can observe the formation of transcription factories throughout the nucleus, small sites of excessive transcription where genes are brought together in cis and in trans, i.e. from the same and from different chromosomes [148]. It has been shown experimentally that transcription factories are preserved under heat shock, even when genes dissociate from them [134], indicating that they really represent nuclear sub-compartments and are not just created through clustering of actively transcribed genes. Different hypotheses state that transcription factories might be specialized through specific transcription factors, or that specialization is only determined by the genes present in it [148].

There are other nuclear bodies that might also influence the chromosomal conformation, such as PML bodies (PB) and Cajal bodies (CB). PBs are stable structures with fixed positions that are involved in cell cycle regulation, apoptosis, and DNA repair and have a suggested role in the organization of p53 responsive genes [148]. CBs supposedly play a role in RNA transcription and processing [148]. Other nuclear bodies are the OPT domain with unknown function and SC35 domains, which are probably involved in storage of splicing factors [148].

1.1.2 *Topologically associated domains (TADs)*

Using whole-genome Hi-C data (experimental method described in section 1.3.2.4), Dixon et al. unveiled the domain-like structure of intra-chromosomal contacts in human and mouse embryonic stem cells and human IMR90 fibroblasts. They observed crisply defined regions in the genome that form many loops and interactions within and only few to other regions, and termed these topological domains or Topologically associated domains (TADs).

The genome consists of megabase-sized local interaction domains

At a size in the mega base pair scale, these relatively small domains are connected by short unorganized linker regions, which are defined as boundary regions and enriched strongly in the known insulator element and transcription factor CTCF (for more on insulators see section 1.2.3). Dixon et al. conclude that boundaries of topological domains correlate with the role of classical insulator and barrier elements. These boundary regions are largely shared between cell types and also conserved between human and mouse. Dixon et al. suggest that domain organization may be stable, while interactions within can be formed more dynamically between cell types.

It is yet unclear how these boundaries are formed. Binding of CTCF alone is not sufficient, as only a small portion of CTCF binding sites (15%) occurs in boundary regions, so other factors enriched in these regions, such as housekeeping genes, tRNA genes and SINE elements, might also play a role.

In general, these observations are in line with the previously described model of the equilibrium globule. Each of the TADs can then be considered as a small more densely packaged unit connected to other such units or small globules by short links, similar to a "beads on a string"-model. Higher-order chromatin structure is formed by folding of these small globular units into more complex structures, possibly an equilibrium globule.

1.1.3 *Lamina associated domains (LADs) [69]*

Large heterochromatic domains associate with the nuclear lamina

As shortly described in section 1.1, genome regions enriched with inactive genes and heterochromatin often reside close to the nuclear lamina. Using a method called DamID and described in detail in section 1.3.3, Guelen et al. experimentally determined these regions genome-wide in human lung fibroblasts [69]. They identified large chromosomal domains with lengths in the megabase-range, similar to TADs, and termed them lamina associated domains or LADs. These domains strongly correlate with gene deserts and generally low gene density, while repressive histone mark H3K7me3 is significantly enriched within LADs. Consequently, Guelen et al. confirm LADs to represent a highly repressive chromatin region and thus also nuclear sub-compartment.

Similar to TADs, CTCF is also enriched in the border regions of these LADs. Its insulator property could be a main factor driving the sharp confine of the domains. Another such factor could be active promoters, which are also enriched close to LAD boundaries and could form barriers to prevent spreading of heterochromatin. Third, density of CpG islands is increased near LAD borders and could also influence LAD formation. However, even CpG islands and CTCF binding sites together mark only 30% of all LAD borders, implying that other, currently unknown elements, also play a role in the formation of these inactive domains.

1.1.4 *Organization of the mitotic chromosome [140]*

Most experimental determinations of the spatial chromatin organization are focused on the interphase chromosomes. Previously described research has shown that these are highly compartmentalized and vary across cell types. Naumova et al. expanded the research to chromosomes during metaphase, and found these to fold very differently [140]. Applying the 5C and Hi-C techniques described in section 1.3 in detail to mitotic chromosomes in HeLa S3 cells, they investigated their structures with respect to different models.

During metaphase, the compartment structure of the genome is resolved and a uniform structure of consecutive loops is formed.

For comparison, experiments were performed at different cell type stages, and Naumova et al. found high correlation of interaction patterns between early-G_1, mid-G_1 and S-phase. During these phases, the cells exhibit the previously described compartmentalization into active and inactive chromatin, also termed A and B compartment, respectively, and enrichment of short-distance interactions over long-distance ones. In Hi-C interaction matrices, these characteristic structures lead to a distinct plaid pattern that represents the two compartments (for an example see Figure 7 on page 22).

Naumova et al. observed a drastic change of spatial chromatin organization in interphase, leading to a complete loss of this characteristic pattern in the interaction matrices. While eigenvector decomposition shows alternating blocks of compartment A and B along the chromosomes in interphase, this compartmentalization vanishes along with cell-type specificity in metaphase. In addition, the previously described TADs are lost during transition into metaphase.

Investigating several polymer models, Naumova et al. determined that a cylindrical loop/scaffold model and scaffold-free model with consecutive loops best fits the data observed in mitotic chromosomes, where no contacts above distances of 10 Mb but frequent contacts below are formed. This model represents a mix of the previously described equilibrium and fractal globule.

As the genome switches between two different structures during the cell cycle progression, the observed mitotic chromosome structure has to be formed from the more complex compartmentalized

interphase chromosome structure. Naumova et al. suggest a two-step process for this switch, where a linear compaction step featuring loop extrusion from so-called SMC complexes is followed by linear ordering of these loop bases through natural axial compression. This second step is caused by the "backbone" that is formed by the loop bases. For an illustration see Figure 2.

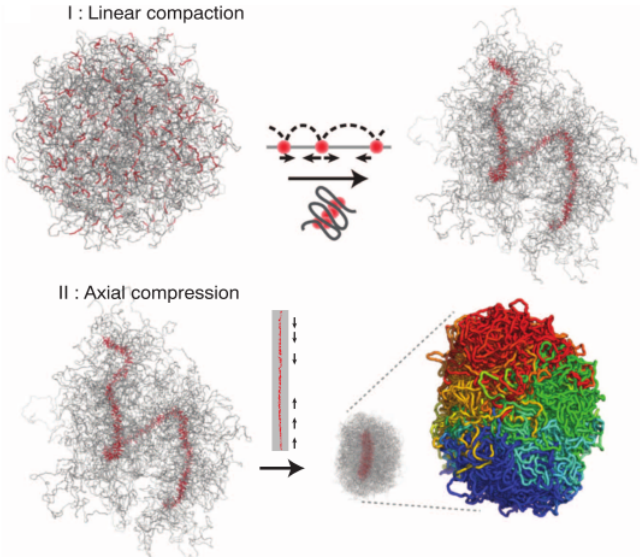


Figure 2: Taken from Naumova et al. [140], this figure illustrates the two-step-formation of the mitotic chromosome structure, which is characterized by consecutive loops. In their model, certain (SMC) complexes first extrude chromosome regions to form loops, serving as loop bases (red points). These bases serve as "backbone", which is naturally compressed in the second step, leading to the coloured structure in the bottom right, where red and blue are the two ends of the polymer.

1.2 SMALL SCALE SPATIAL ORGANIZATION INCLUDES CHROMOSOME LOOPS

Most models for large-scale chromatin folding are based on the formation of a network of short- and long-ranged interactions along chromosomes. These interactions are not only of structural relevance, but also serve a functional purpose. Insulators, enhancers and silencers need to come together in space with their target in order to fulfil their task of insulating, activating or repressing them to directly influence expression. Often, these regulatory elements are in close proximity to their target on the sequence level, but there are many cases where they are located up to 100 kb away [40, 39].

An underlying feature of the mechanisms which lead to a regulator effecting a target are close contacts established through loops of chromatin. In the following sections different types of regulation sys-

tems that rely on chromatin looping and the main molecular players guiding them are described.

1.2.1 *Enhancer-promoter interactions through chromatin looping*

Enhancer-promoter interactions happen not only in close linear distances, but often the regulatory element is far apart from its target on the sequence level. Several experimental studies with 3C techniques (as explained in detail in section 1.3.2) support the looping model, where both elements are brought together through chromosome loops and thus are shape the nuclear organization [42, 24] (compare Figure 3 B).

Chromatin loops may be formed to connect regulatory elements to their targets

One well-studied example for such an interaction is the contact or close proximity between the so called Locus Control Region (LCR) and active globin genes in mice [149, 206]. The chromosomal region containing the inactive β-globin genes loops out during this process, a conformation that is disrupted during differentiation, when newly activated globin genes interact with the LCR. This specific loop thus dynamically rearranges itself with the activation of new genes that come into it, while inactive genes are moved out.

This shows that long-range interactions of enhancers and promoters can be involved in the activation of gene expression.

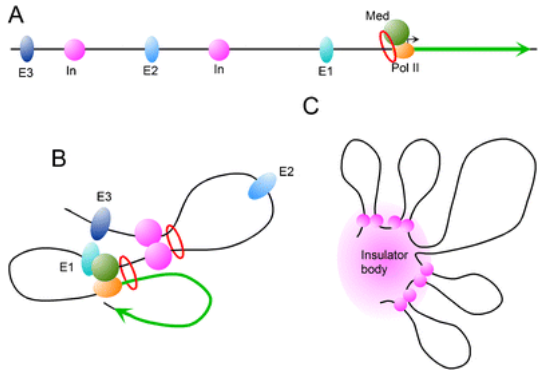


Figure 3: Looping models for genes and regulatory elements (Taken from Hou et al. [87]).
A: Linear representation of a gene (green), its promoter with Polymerase II (PolII) complex and different regulatory elements, with enhancers represented in blue and insulators represented in pink. The red ring illustrates a cohesin complex.
B: Model of looping of chromatin around this gene that leads to an interaction of promoter and terminator, enhancers E1 and E3 and the promoter complex, and isolation of enhancer E2 through an insulator complex such as a CTCF dimer.
C: Model of the formation of insulator bodies through bringing together of multiple insulators and looping out isolated chromatin regions.

1.2.2 *Regulation of transcription by interactions between the promoter and terminator*

Promoter and terminator sites of a gene play an important functional role in its expression, guiding RNA polymerase II (RNAP II) and determining start and end of a gene. Multiple studies on yeast have observed strong physical interactions between these two sites in several genes of varying length [7, 52, 151, 189]. Similar interactions have been detected for the mammalian breast cancer gene BRCA1 [201] and other mammalian genes [147].

These interactions represent short gene loops that bring together start and end site of a gene (compare Figure 3 A, B). Possible reasons for this is efficient recycling of RNAPII and other transcription factors; when the transcription complex is released at the terminator site, it can immediately bind to the promoter due to close proximity. However, no evidence can support this hypothesis at the moment. There is a correlation between formation of these contacts and rapid re-activation of transcription in yeast, indicating a role in transcription memory [202]. If this is the case, these loops could serve as memory gene loops that lead to rapid re-activation of transcription after a transient silencing period. In mammalian cells these loops have not been shown to serve such a purpose, but instead some of them even disappear upon high expression levels [201].

1.2.3 *Insulator-mediated interactions*

Insulators are genomic elements that interfere with the contact of regulatory sequences and their target genes. They can recruit chromatin remodelling enzymes to interfere with the spreading of repressive chromatin (barrier insulators) or block enhancers or silencers through mediation of intra- or inter-chromosomal interactions [87].

Insulators can prevent regulator-target interaction physically

The latter class have been associated with the CTCF protein which can form loops of intervening DNA [88, 232]. CTCF has been shown to co-localize with cohesin [157, 174, 198, 221], a protein that can form a ring around one or two chromatin fibres, and possibly works coordinated with this ring-like protein to steer long-range interactions. It has been observed in human CD4 T-cells that CTCF separates the enhancer and promoter through allocating these to different loops and thus preventing physical contact [72].

Insulators can also come together to form insulator bodies at specific locations in the nucleus [25] (see Figure 3 C). These elements may play an important role in the establishment of the chromatin interactome.

1.2.4 *Polycomb-mediated long-range repressive interactions*

The Polycomb (Pc) complex plays a role in the repression of Hox genes in Drosophila during development. Hox genes are located in different clusters that are more than 10 Mb apart in linear distance. These genes are present and co-regulated in subnuclear structures called Pc bodies that are formed through co-localization of Polycomb group (PcG) proteins and looping out of intermediate chromosome regions [87, 11] (see Figure 4). However, these co-localizations of the Hox genes only happen in tissues where they are repressed, indicating a recruitment to Pc bodies for regulatory purposes.

Further 4C (see section 1.3.2 for a description of this method) experiments showed that there is an extensive interacting network for PcG target genes in the nucleus that mostly contains genes from the same chromosome arm, such as the Hox genes [87].

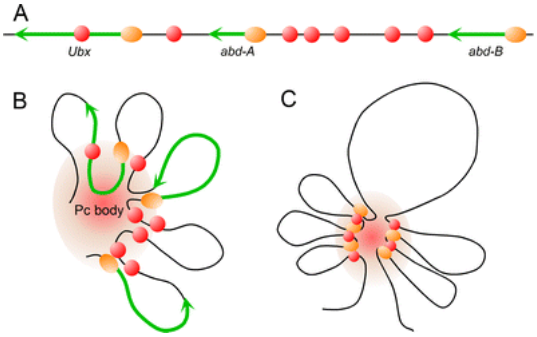


Figure 4: Model of the formation of Pc bodies (Taken from Hou et al. [87]).
A: Linear representation of the *Drosophila bithorax* complex BX-C, with genes represented as green arrows, transcription complexes represented as orange ovals and PREs (Pc response elements) illustrated as red spheres.
B: Formation of a Pc body through looping and specific interactions between PREs and promoters.
C: Multiple Hox loci can be co-repressed in such a Pc body through looping over large distances.

1.2.5 *Long-range interaction and the regulation of imprinted genes*

Co-repressed Hox genes are not the only group of genes that are simultaneously regulated; similarly, co-regulated imprinted genes can form an interaction network. Imprinted genes are expressed dependent on the parent from which the chromosome was inherited; in human, imprinted alleles are silenced so that only the non-imprinted allele from the other parent is expressed [226]. Long-range chromosomal interactions in cis and trans, i.e. on the same and different chromosomes, respectively, have been linked to regulation of gene expression that is dependent on the parent of origin, studied in detail for the *Igf2/H19* locus [112, 137]. For this locus, studies show that the

imprinted expression depends on a specific chromosomal organization that is unique for each allele. The cause for this are varying DNA methylation profiles that lead to different binding patterns for CTCF, a protein that has already been associated with mediation of chromatin loops in section 1.2.3.

1.2.6 *Inter-chromosomal interactions during X-chromosome inactivation in mammals*

X-inactivation in the mammalian female genome of embryonic cells is vital during the development. The decision which of the two copies is inactivated does not happen randomly. Instead, both chromosomes are brought together through unknown mechanisms [227, 9] that appear to depend on CTCF or Oct4 [228, 47]. Both are also essential for the inactivation itself: knockout of CTCF leads to a complete loss of X-inactivation, while silencing of Oct4 leads to inactivation of both copies.

This specific sister chromosome interaction does not involve looping, yet it is dependent on CTCF like some of the chromosome looping mechanisms above described. Another important player in X-inactivation is the long non-coding RNA (lncRNA) Xist, which is described in the following section.

1.2.7 *lncRNA can mediate chromatin state and act across different chromosomes*

Recently, possible roles of lncRNA in shaping the chromatin interactome have emerged [172]. According to new studies, long non-coding RNA might be involved in nuclear organization at many different levels, from forming nuclear bodies such as para- or nuclear speckles [127] and even mediating gene-gene or enhancer-promoter interactions [124], within or even between chromosomes [71, 124].

lncRNAs Xist and Firre exploit and modify spatial chromatin organization

Among the most well known lncRNA is Xist, which plays an important role in X-chromosome inactivation [30, 163, 53]. Its expression in males or autosomes, where the X chromosome is normally not inactivated, is sufficient for the silencing and compaction of the chromosome in a repressed nuclear sub-compartment [30, 163]. The mechanism for this includes recruitment of the polycomb repressive complex 2 (PRC2) and exploiting the three-dimensional chromosome conformation to spread along the X chromosome, mediating gene silencing and the modification of chromatin structure [53, 187].

Another lncRNA that uses and possibly modifies spatial proximity is Firre, which is required for adipogenesis [71]. Though the gene is located on the X chromosome, the *Firre* locus escapes X-inactivation and localizes to genomic regions in cis, but also in trans on chromosomes 2, 9, 15 and 17. These trans-interactions bring together genes

that have regulatory functions in adipogenesis. Firre thus server as a mediator to create a functionally specialized spatial cluster.

This is achieved through a combination of different factors. Firre contains a repeating RNA domain termed RRD that can be bound by the matrix protein hnRNPU, which in turn is also required for trans-localization of Firre. The authors thus suggest a model where Firre is bound by hnRNPU which then connects the locus to other chromosomes by binding of DNA sequence.

1.2.8 *CTCF mediates chromatin domains*

CTCF and its essential role in many processes of higher chromatin order has already been mentioned in previous sections. ChIA-PET experiments aiming to analyse the CTCF-chromatin interaction map on a genome-wide level in mice have been performed by Handoko et al [78]. They observed five distinct chromatin domains that are created through boundaries of linearly arranged active and repressive chromatin regions in which CTCF serves as a boundary marker. It can be concluded that CTCF is not only an insulator, but also an important part of a more general mechanism that brings together regulatory sequences.

1.3 DETERMINING THE CHROMOSOME INTERACTOME: METHODS

Knowledge about chromosomal organization only recently increased, when new and high-throughput methods became available that led to insight into this complex biological process. Based on chromosome conformation capture technologies, Hi-C was developed to gain genome-wide data and better understand the interplay of regions between and along chromosomes. Additionally, DamID proved to be a good method for analysing the nuclear lamina-associated chromosomal regions. In the following sections, these and other methods for the determination of chromatin organization are described.

1.3.1 *Fluorescent In-Situ Hybridization (FISH)*

Before the emergence of DNA sequencing methods, many researchers relied on visual information to gather knowledge about the DNA. In combination with microscopy, the use of stains allowed distinction of different chromosomes or different regions of chromosomes. Fluorescent In-Situ Hybridization (FISH) was used as early as the early 1970s to get a glimpse of spatial chromatin organization in mice [90] with the help of a fluorescent stain. With different probes, different scales of chromatin can be made visible, from single genes to entire chromo-

some territories (CTs). Today, FISH is still used to get an impression of nuclear organization that is easy to understand and interpret.

1.3.1.1 *3D FISH [128]*

Three-dimensional FISH applies the same technique to three-dimensionally preserved cells to get a spatial image of the nucleus. While 2D techniques can only ever illustrate a section of the nucleus, this method allows researchers to generate a holistic image of chromosome organization. Capturing of a single cell in three dimensions is achieved through serial optical sections. The gold standard is obtained with confocal laser scanning microscopy (CLSM), with a resolution of 180-250nm laterally and 500-700nm axially [128, 214]. The spatial chromatin organization needs to be fixated prior to FISH, and pretreatments are necessary (see for example [194]). Multiple colors can be applied to different chromosomes to make them distinguishable [182, 196].

FISH can be used for microscopy-based visual analysis of spatial chromatin organization

Using this approach, Bolzer et al. [18] were able to simultaneously visualize all 46 chromosome territories in human fibroblasts. Figure 5 illustrates a 3D-FISH on an MCF-7 breast cancer cell, taken from Walter et al [214]. With the help of serial confocal sections, researchers can get an impression of the three-dimensional organization in the nucleus by using 2D image generating microscopy. Chromosome territories can be clearly visible in this technique, making it an important tool for the first steps in understanding spatial chromatin organization.

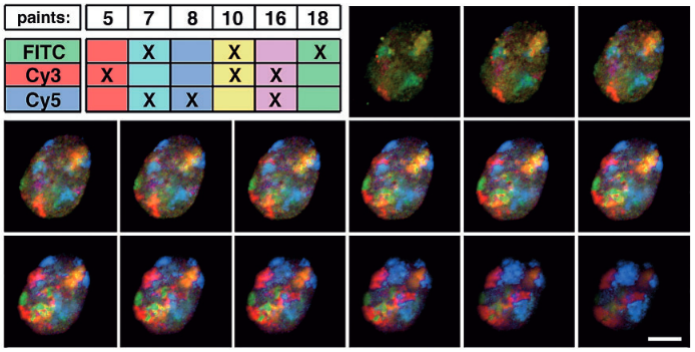


Figure 5: Illustration of microscopy imagery in 3D-FISH on the nucleus of a MCF-7 breast cancer cell, taken from Walter et al [214]. Chromosomes 5, 7, 8, 10, 16 and 18 are stained according to the color legend in the top left corner, FITC, Cy3 and Cy5 are different fluorophores. Each image represents a serial confocal section of the cell, creating a three-dimensional image of the nucleus.

1.3.2 *Intra- and interchromosomal interactions: Chromosome conformation capture (3C)*

While FISH-based methods rely heavily on visual interpretation, the advance of deep sequencing methods has allowed the emergence of a new type of methods. 3C techniques aim to find regions of chromatin that are close in space. The idea behind the technique is simple: after cross-linking such regions with formaldehyde, the DNA is digested with a restriction enzyme like BgI II or Hind III and open ends are joined, leading to a circular or linear strand of DNA that consists of both chromosome regions of the interaction [150, 42, 37]. Subsequent Polymerase Chain Reaction (PCR) or sequencing is used to identify the sequences and map them to the genome (see Figure 6 A, page 20). However, interactions discovered with this method do not necessarily imply a functional contact between two chromosomal regions, but can also be caused by close proximity in transcription bodies, similar nuclear sub-compartments or simply due to chance. In the following sections the chromosome conformation capture method and its evolution are described in more detail.

1.3.2.1 *3C [42]*

3C laid the groundwork for the development of many other chromosome conformation capture methods with different focuses. It was first proposed by Job Dekker's group in 2002 [42] and described as "an approach to detect the frequency of interaction between any two genomic loci" (Dekker et al. 2002 [42]). Compared to FISH, this sequencing-based method has some advantages: it is applicable to genomes of any size and generates a large and very detailed map of the interactome. Even for single genomic loci, many proximal genome regions can be identified. However, this can also be considered a disadvantage, because statistical data processing is necessary before the results can be interpreted. While FISH directly highlights chromosome territories, it takes statistical analysis to detect them with 3C. The method is also more prone to noise that has to be filtered out to detect valid interactions.

3C combines cross-linking with PCR or sequencing to determine the detailed chromatin interactome

The general principle of 3C methods has already been described above. In detail, the procedure is described below as presented in Dekker et al. 2002 [42].

A. Isolation of (intact) nuclei

B. Fixation of higher-order chromatin organization with formaldehyde

C. Digestion of cross-linked DNA with a restriction enzyme

D. Ligation of cross-linked DNA

E. Reversal of cross-linking

F. Quantification with PCR and locus-specific primers

G. Frequency is detected by quantitative PCR reactions

Formaldehyde is able to cross-link proteins and DNA-bound proteins to the corresponding DNA sequence. Since the eukaryotic genomes are packaged and bound to many proteins such as histones, this in effect leads to cross-linking of proximal DNA regions through their bound proteins [42]. After digestion with restriction enzymes, this cross-linking leads to molecules containing two DNA fragments from different loci, which are then favourably ligated in the next step, because intra-molecular ligations are favoured over random ligation events. A control is compared with the quantified PCR products to determine the ratio of observed interactions to expected interactions of two given loci. This ratio should be directly proportional to the interaction frequency and is taken as an approximation of it.

Dekker et al. observed that the frequency of interactions decreases with increasing linear distance. Analysing the yeast genome they also showed that 3C was able to detect the previously known co-localization of telomeres [33, 64] and of centromeres [93, 94] for chromosomes III and IV. Additionally, 3C can also serve as a basis for chromosome modelling. Dekker et al. showed that, according to their 3C experiment, chromosome III of the yeast genome forms a distorted ring [42].

The main limitation of 3C lies in the semi-quantitative [42] or quantitative [197, 224] PCR. Primers need to be designed for each restriction enzyme cutting site of interest. Considering that each locus of interest exists only twice in a diploid cell and that so called 'hairballs' of chromatin, where many fragments aggregate, are common cross-linking side-effects, PCR requires amplification of very rare ligation events. Due to this, the PCR step is very difficult and has to be strictly controlled [41]. Especially for large distances, ligation events are often too infrequent to be detected correctly by PCR. Sequencing or microarray-based methods allow for a more unbiased approach and were developed to improve the 3C methodology. Several modifications exist that focus either on the high-resolution analysis of a region of interest or on genome-wide high-throughput, which are described in the following sections.

1.3.2.2 *Circular chromosome conformation capture (4C)*

To overcome the limitation imposed by PCR through the necessary primers, 4C or circular 3C was developed. Originally, 4C (also termed chromosome conformation capture-on-chip [38]) was used to determine all interaction partners of an interesting genome region by combining 3C with microarrays [237, 188] (Figure 6 B). Here, the interesting sequence is subjected to 3C to form ligated circles with interaction

partners, either naturally or through a second round of digestion by a restriction enzyme and ligation [188]. Inverse PCR and microarrays or Next Generation Sequencing (NGS) methods (termed 4C-seq) are then used to analyse the interaction partners' sequences. The main advantage of this method compared to 3C is that only primers specific to the locus of interest or *viewpoint* are necessary. The circular ligation product allows for amplification of all sequences that are in contact with the viewpoint.

4C is used to find all interactions formed by a region of interest or viewpoint

The first application of the 4C method was to identify interaction partners of the tissue-specific β-globin gene and Rad23a in embryonic mouse cells [188]. Simonis et al. found Rad23a, a housekeeping gene, to interact with other active regions on the same chromosomes in a tissue-unspecific manner. They also found the interaction profile of the β-globin gene to depend on its expression status, interacting with other active loci in erythroid cells, where it is expressed, and with inactive loci in fetal brains, where itself is inactive, too. This example clearly shows the advantages of 3C-based methods over FISH and microscopy, as they allow for easy combination of the data with expression data sets, and can give high resolution information on single loci.

Another modification of this method is called "Adapter ligation" and involves the merging of an adapter sequence to a sticky end at the interaction partner's sequence after a second round of restriction enzyme digestion [122] (Figure 6 C). This sequence can then be used to facilitate PCR by using a primer targeted at it.

1.3.2.3 *Chromosome conformation capture carbon copy (5C)*

Following 3C and 4C, another technique termed 5C (chromosome conformation capture carbon copy, Figure 6 D) was developed based on ligation-mediated amplification [49]. Short primers that target an adjacent region to 3C restriction sites are hybridized to the 3C template to create fusion oligonucleotides that can be identified by sequencing. PCR amplification is again facilitated, since the adapter sequences within the oligonucleotides can be used as PCR primers. The detection of ligation products can again be done by microarrays or NGS.

5C improves 3C by facilitating the PCR amplification

Compared to 4C, the resolution of 5C is lower, because it is limited by the properties of restriction fragment ends. Not every end is suited for the creation of a 5C oligonucleotide [38]. The main advantage compared to 4C is the large-scale approach: 5C creates a matrix of interactions rather than a set of interactions for a given viewpoint. So far, 5C has not been used for a whole-genome analysis, for which Hi-C is better applicable. Still, it can be considered a medium-throughput alternative to 3C and is best used for detection of enhancer-promoter interactions in specific loci or the set of interactions between entire genome regions. Due to an easier protocol, 3C is still more commonly

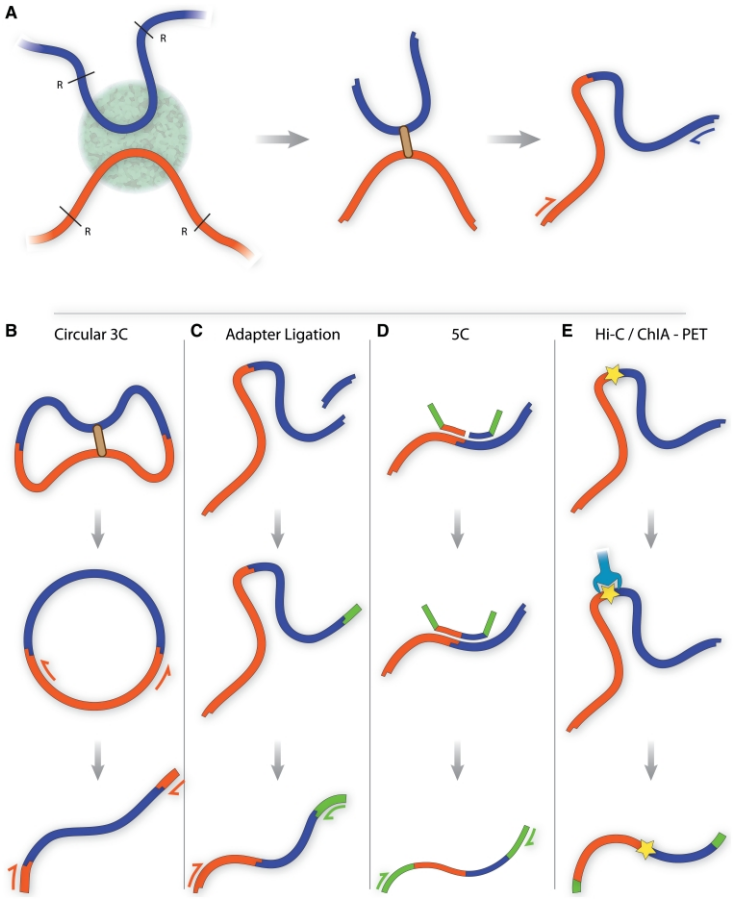


Figure 6: Different chromosome conformation capture techniques (Taken from [150]).
A: Regular 3C [37, 42] starts with cross-linking of interacting chromosome regions with formaldehyde, followed by a digestion step and subsequent joining of restriction enzyme sites. The ligated stretch of DNA can be amplified with PCR and sequenced.
B: Circular 3C or 4C [237, 188] focuses on an interesting region of DNA and includes a second round of restriction and ligation to create circular DNA fragments. No primer for the unknown interaction partner sequence is necessary to amplify all circles and identify the interaction partner.
C: Adapter ligation [122] includes joining of an adapter sequence to the sticky end of the interaction partner's sequence, which can be used for amplification without further knowledge of the target sequence.
D: 5C technique [49] anneals primers based on the restriction enzymes' sites to the ligation site and uses them for amplification.
E: Hi-C [120] uses biotin to label the ligation sites and subsequent isolation with streptavidin beads. Paired-ends sequencing allows sequence identification on a genome-wide level.

used for the first type of experiment, but considering its PCR bias 5C is the better alternative.

As an example for interaction detection in two genome regions, Wang et al. [215] compared the set of interactions between HOX clusters and other genome regions in two cell types, the first of which expressed only 5′ HOXA genes, while the second only expressed 3′ HOXA genes. Their interaction patterns were shown to be diametrically opposed, each cluster forming long-range distances only between active regions [215].

1.3.2.4 *Hi-C*

The previously described methods are mostly limited to analyse interaction partners of specific regions of interest in the genome. To get a genome-wide interaction map, Hi-C was developed by Lieberman-Aiden et al [120]. This "all-vs.-all" method enriches the ligation junctions through integrating biotin as a label (Figure 6 E) after restriction enzyme digestion. After blunt-end ligation, complexes are subsequently purified with streptavidin and sheared, and a pull-down assay separates ligation junctions from other molecules. Massive parallel sequencing is enabled through ligation of the adapter sequences to the library. Through mapping to the genome pairs of sequences from two different loci can be identified, creating a genome-wide matrix of interactions. Resolution of Hi-C is limited and originally lay at 1 Mb [120]. A 10-fold increase in resolution is coupled with a 100-fold increase in sequencing depth due to the quadratic nature of the interaction data. With increase in sequencing depth in the future, the resolution of Hi-C methods promises to increase. In fact, in 2012 Dixon et al. published Hi-C data on mouse and human with a resolution of 20 kilo base pairs (kb) [45].

In Hi-C, the 3C technique is combined with NGS, enabling all-vs-all whole genome analysis

Lieberman-Aiden et al. confirmed the previously observed separation of active and inactive genome regions [188] for the whole human genome. They identified compartments A and B, corresponding to subnuclear locations where genome regions within the active compartment A tend to form interactions mainly within the compartment, and inactive regions in compartment B contact mainly regions within, too. Hi-C data also proved a largely overlapping spatial chromatin organization between two cell types, though many loci resided in different compartments in the two cell types (GM06990 and K562). Smaller genomes like that of *S. cerevisiae* allowed for a higher sequencing resolution. For this species, Hi-C confirmed the clustering of telomeres and centromeres [51] among other properties.

The raw Hi-C interaction matrices are usually subjected to some sort of statistical post-processing to increase the signal. In a first step, the matrices are normalized by division through a matrix of expected interaction counts. These are generated based on the principle that the likelihood of interaction decreases substantially with increasing linear

distance. The normalized Hi-C matrix is then refined by calculating Pearson correlation coefficients for each cell.

A cell $c_{i,j}$ in the normalized matrix corresponds to the interaction frequency of two loci i and j. To refine this value, Lieberman-Aiden et al proposed to calculate spatial proximity values by comparing their entire interaction profiles, i.e. the vector of interaction frequencies to all other loci in the chromosome. These are represented in the matrix by the column i and the row j. A Pearson correlation coefficient is calculated for these two vectors and entered in cell $c_{i,j}$ of a third, spatial proximity, matrix. Contrary to its name, the spatial proximity value of two segments i and j quantifies the similarity of their contact profiles or the propensity to lie in the same compartment, respectively. This matrix clearly shows the so-called plaid pattern caused by compartments A and B in a refined manner, reducing the noise (Figure 7).

The preference of loci for one of the two compartments can also be represented by the first eigenvector of the interaction matrix. The reason for this lies in the fact that the strong compartmentalization of loci causes most of the variance in the data. In addition, the membership of a locus to a compartment is mutual exclusive and interaction profiles of loci in the same compartment were observed to be very similar.

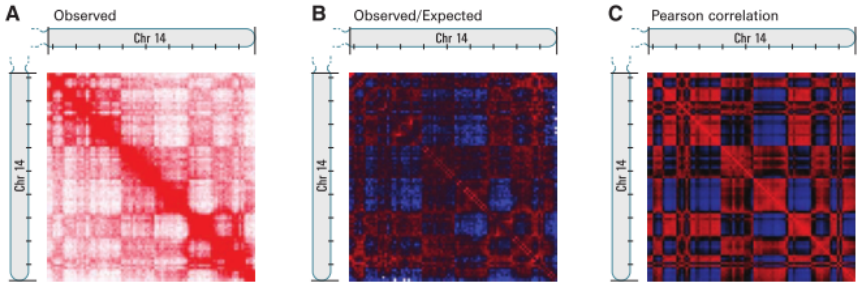


Figure 7: Normalization procedure for Hi-C matrices proposed by Lieberman-Aiden et al, taken from their publication [120]. After normalization of the Hi-C matrix with expected interaction counts, Pearson correlation coefficients are calculated for each row and column and entered in the cell where they overlap. The sharp borders between red and blue squares are called 'plaid pattern' and show, that each locus lies in one of two compartments with many interactions within and little interactions between. Cells with high values in matrix C (red) correspond to segments in the same compartment, while low values (blue) correspond to loci in different compartments.

In summary, Hi-C is a high-throughput method that allows determination of spatial chromatin organization on a genome-wide level, with resolutions depending on the genome size. However, Hi-C is applied to many different cells of the same type, generating a summary of interactions that happen in at least one cell at the time point of the experiment. There are other biases for this method, which are described in detail in part iii. Still, combining chromosome conformation capture techniques with the speed of next generation sequencing

allows us to gain a low-resolution genome-wide map of inter- and intra-chromosomal interactions that can greatly increase our understanding of chromatin organization.

1.3.2.5 *ChIA-PET*

Similar to Hi-C, ChIA-PET (short for Chromatin Interaction Analysis with Paired-End Tag sequencing) also generates a genome-wide library, but achieves this through combination of 3C with Chromatin Immunoprecipitation (ChIP). As such, the focus differs slightly from Hi-C, since ChIA-PET can be applied to all loci bound by a protein of interest [59]. Fragment generation is performed through sonication, and followed by pull-down with an antibody to the protein of interest.

The first application of ChIA-PET was done by Fullwood et al. in 2009 [59] to sites bound by the oestrogen receptor α (ERα). It revealed several thousand intra-chromosomal loops between binding sites, the most prominent of which were reproducible between replicates. However, the technique is not able to detect if this loop formation is dependent on ERα, as immunoprecipitation can only work when the protein is bound. Another disadvantage is that only contacts between loci bound by the same protein can be identified. ChIA-PET is thus more adequate than Hi-C or other methods to identify the interaction network around binding sites of a single protein, but has disadvantages over these techniques in other aspects.

ChIA-PET combines 3C with chromatin immunoprecipitation of DNA-bound proteins

1.3.2.6 *Single-cell Hi-C [138]*

One of the main draw-backs of the Hi-C technique is the averaging of data over millions of nuclei. This is necessary to capture enough information before the sequencing step that a signal can be detected. However, the spatial organization of chromatin is highly flexible because of Brownian motion of chromosome regions without fixed nuclear positions. As a result, the Hi-C average will lead to detection of many mutually exclusive interactions that happened at the time of the experiment in different cells. Single-cell Hi-C aims to eliminate this problem by applying a modified Hi-C technique to one cell at a time.

Single-cell Hi-C is performed on individual nuclei and not averaged over millions of cells

The protocol for single-cell Hi-C by Nagano et al. [138] shares many similarities with regular or 'ensemble' Hi-C, but has some important differences. For one, the cross-linking of DNA and proteins is performed within the nuclei, while ensemble Hi-C performs nuclear lysis first. Nuclei for further analysis are selected visually under the microscope, and the standard Hi-C protocol of cross-link reversal and a biotin pull-down onto streptavidin-coated beads is performed.

Before sequencing, another digestion step using a different restriction enzyme, AluI in the case of the original publication, is applied

and the resulting fragments are ligated to customized and tagged Illumina adapters. Since only two fragments per chromosome locus are present in a single cell, PCR amplification is necessary, followed by paired-end sequencing.

Single-cell Hi-C data lack most mutually exclusive interactions and show what the genome looks like in a small set of cells at a single point in time. However, depending on the agenda of the researcher, averaging of data might still be more useful. If we assume that there are housekeeping interactions that are functional and present in all cells, the averaged Hi-C over multiple cells may allow us to distinguish these frequent interactions from random ones that are caused by Brownian motion. For this reason, single-cell Hi-C is always performed on a set of multiple nuclei.

Nagano et al. compared the results of 60 pooled single-cell experiments in human with normalized [230] ensemble Hi-C data by Lieberman-Aiden et al [120] from approximately 10 million nuclei. They found a strong similarity between both sets, confirming the validity of single-cell Hi-C. Nagano et al. show that the topological domain structure of intra-molecular contacts is conserved between single nuclei, and that inter-domain contacts are highly variable between cells. These differences are not be noticeable in ensemble Hi-C, so that the averaged maps imply a more complex and more inter-connected chromatin network than there actually is. Nagano et al. conclude that each chromosome interacts with a relatively constant number of chromosomes through a limited but constant surface area which is highly variable between cells [138].

1.3.2.7 *Tethered chromosome conformation capture (TCC) [97]*

TCC increases signal-to-noise ratio compared to Hi-C by immobilizing the DNA fragments on a solid surface

Tethered conformation capture (TCC) is another improved method for the genome-wide detection of DNA interactions. It aims to remove one of the most problematic biases of Hi-C, the low signal-to-noise ratio, by performing the experiments on a solid surface. In Hi-C, random ligations between DNA fragments that are not crosslinked are a main cause of noise. TCC solves this problem with a modified approach. The first steps are similar, as DNA is first cross-linked and then digested by restriction enzymes. However, after this step DNA-bound proteins are cysteine-biotinylated and the fragments are immobilized on streptavidin-coated beads, and ligated subsequently. As a consequence, only crosslinked regions are pulled down. Similar to Hi-C, massively parallel sequencing is applied after purification to the ligation junctions, and a matrix of interactions emerges. Figure 8 gives an overview of the procedure.

Kalhor et al. show that this method accurately reproduces the interaction patterns observed in Hi-C experiments, while reducing the noise to almost half [97]. TCC thus provides a high-quality alternative to Hi-C with reduced need of normalization.

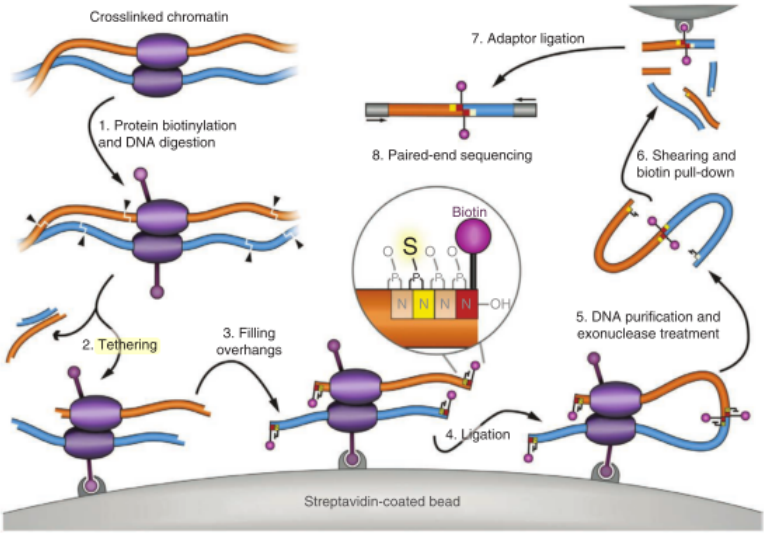


Figure 8: Overview of the tethered conformation capture (TCC) technique, taken from [97]. After cross-linking (1) and tethered (2) to streptavidin-coated beads, blunt ends are created (3) and ligated (4). Cross-linking is then reversed and the DNA is purified (5). The DNA is sheared and only fragments which included a biotinylated nucleotide are pulled down (6) before sequencing (7).

1.3.3 *Interactions with the nuclear lamina: DamID*

DamID is short for *DNA Adenine Methyltransferase IDentification* and is generally used to find binding sites of DNA and chromatin binding proteins. This protein is fused to a DNA methyltransferase (Dam) and expressed in the cell. If it binds to the DNA or chromatin, Adenine bases 50nm close in space to the binding site are methylated, an alteration which does not naturally occur in eukaryotes [210].

DamID allows detection of genome regions that are located in the nuclear periphery

To find chromosomal regions close in space to the nuclear lamina, also termed LADs, it is thus sufficient to create a fusion protein of Lamin B1 and Dam and express it in the cell. The protein will bind to or be in close proximity to DNA, since it is part of the nuclear envelope stabilizing protein mesh, and the methyltransferase can subsequently methylate all Adenines in close regions. A special form of PCR termed mePCR is then carried out to find these regions on sequence level (see Figure 9 for a schematic overview of the process). However, it has to be kept in mind that this method does not prove a direct interaction of chromatin with the lamina, but rather a close proximity.

The aforementioned ChIP is another method to determine interaction between proteins and DNA. However, with DamID no specific antibodies have to be developed. Another advantage for interactions with the lamina is that ChIP only shows the current association of the DNA and the protein, while DamID allows the researcher to capture

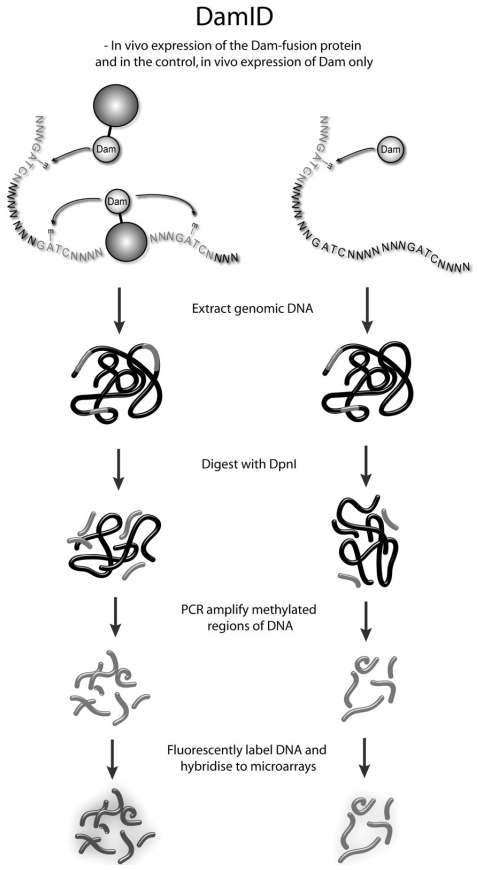


Figure 9: Schematic illustration of the steps of the DamID technique for determination of protein-DNA-contact (Taken from Southall et al [195]). A fusion protein is created consisting of the DNA-binding protein and DNA methyltransferase (Dam) and expressed in the cell, where Dam methylates every Adenine close to the binding site. Genomic DNA is then extracted, digested and sequenced with a special form of PCR that is able to identify methylated regions and microarrays.

everything that is or has been in contact with the lamina. Since chromatin fibres are more flexible and might move to and fro along the nuclear envelope, DamID serves as better method for the analysis of this kind of interaction.

1.4 OUTLOOK

Through development and improvement of chromosome conformation capture techniques, more and more large sets of data on chromatin interactomes in eukaryotic cells become available. In this work, we aim to analyse these data with Bioinformatics tools to gather knowledge about the three-dimensional structure of the genome. We will focus on the evolutionary aspect to find out to which amount this structure is conserved between species, using *Homo sapiens* and *Mus musculus* as model organisms. Looking at the structure of the genome

from different perspectives, this work tries to give a holistic view of what constitutes the chromosome interactome, which factors influence it or are depending from it, how it evolved and to what extent it is functional.

In the second part, the linear genome and its many sequence and structural features and their interplay are the main focus. Three-dimensional inter-chromosomal contacts are the subject of part iii, in which we create a high-confidence interaction network and compare its properties in human and in mouse. The last part lays its focus on linear sequence evolution and rearrangements between chromosomes that led to the different linear structures we see in genomes today.

Altogether, these different perspectives show the inter-dependency of two- and three-dimensional genome organization, and help us better understand its role in the molecular processes of a cell.

Part II

SPATIAL CHROMATIN ORGANIZATION AND GENOMIC FEATURES

Using a large data set of linear and structural genomic features, this part describes the complex interplay of genomic properties and the statistical analysis performed to uncover it.

2 MATERIALS AND METHODS

Eukaryotic genomes are long sequences of nucleotides, coding and non-coding regions, genes, regulatory elements and repeats, often divided into multiple chromosomes. While the three-dimensional structure of these sequences are the main topic of this work, we will first focus on the many other properties genomic sequences have.

For a long time, genes were thought to be the only important elements within the DNA sequence. Today we know that regulatory elements like transcription factor binding sites, histone methylations, repeats or domains such as LADs or Replication Timing Domains (RTD) are important as well, and that they all play together to form the genome. In this part we want to investigate the interplay of these so-called features and their role for the three-dimensional structure of eukaryotic genomes.

2.1 STATISTICAL ANALYSIS OF CHROMOSOMES' PROPERTIES

The basis for our statistical analyses is an in-house database of genomic features in four eukaryotic species. A genomic feature is a sequence- or structure-based property that is distributed along the genome, in the form of elements. An example are repeats, where an element is a single repeat, but the feature track describes the start and end positions of all such repeats in the genome. Other features, such as epigenetic histone modifications, RTD or lamina-proximal regions, also list a score per element that quantifies the (experimental) signal. The database contains feature tracks from UCSC [102, 98] and other sources and was first compiled by Daniel Nasseh (Diploma student, Department of Genome-oriented Bioinformatics, TU München) and subsequently extended by Hongen Xu (PhD student, Department of Genome-oriented Bioinformatics) and myself. In Table 1 you can find a list of most of the features available in this database and their sources.

Sequence-based features like genes or repeats are cell-type independent, but for most other features differentiated cells such as lymphoblasts are the source. In mouse, LADs are available for multiple cell types (mouse embryonic fibroblasts (MEF), Embryonic stem cell (ESC), Astrocytes, neural progenitor cells (NPC)), from which two summary sets termed 'strict' and 'greedy' were created. While 'greedy' refers to the set of all LADs that appear in at least a single cell type, i.e. constitutive LADs, 'strict' contains only those that appear in all of them, i.e. constitutive plus facultative. In human, LADs are only

Table 1: Description of genomic features available in the in-house database.

Feature	Description	Source	Species
DNase I hypersensitivity sites	mark accessible chromatin, collection of cell types	ENCODE, Sabo et al. (2004, 2006) [176, 177], hg18	Human
GC-content	per 1 Mb segment of the genome	UCSC [102] sequences	Human, mouse
Hi-C compartments	In form of Eigenvector, GM06990 lymphoblastoid cells	Lieberman-Aiden et al. (2009) [120]	Human
Histone acetylations	18 acetylations, e.g. H3k9ac, CD4+ T cells	Wang et al. (2008) [217]	Human
Histone methylations	20 methylations, e.g. H3k4me1, CD4+ T cells	Barski et al. (2007) [13]	Human
LADs	Lamina associated domains	Peric-Hupkes et al. (2010) [159]	Human, mouse, fly
lncRNA	lncRNA genes	UCSC [102]	Human, mouse
miRNA	miRNA genes	MirBase version 16 [106, 105, 68, 67, 66]	Human, mouse, worm, fly
Nucleosome occupancy	Predicted	ENCODE, UW (University of Washington), Gupta et al. (2008) [70], hg18	Human
Open chromatin	GM12878 lymphoblastoid cells	ENCODE, Duke/ UNC/ UT-Austin/ EBI [21, 8]	Human, mouse
Repeats	LINE, SINE, LTR and 14 others	RepeatMasker [191]	Human, mouse, worm, fly
RTD	Replication timing domain, Lymphoblasts (Human), ESC (Mouse)	ReplicationDomain DB [219]	Human, mouse
SNPs	Single nucleotide polymorphisms	dbSNP 130 [185]	Human

available for lung fibroblasts. The majority of features in the database are taken from human and mouse. Table 2 lists the assemblies underlying the data collection. If necessary, data between assemblies was lifted using LiftOver [82].

Table 2: Assemblies of species' genomes for which data are available in database, if not stated otherwise.

Species	Assembly
H. sapiens	Mar2006 (NCBI36/hg18)
M. musculus	Jul2007 (NCBI37/mm9)

R [166] was used for simple correlation analyses, display of correlation coefficients in the form of heatmaps and other plots. If not stated otherwise, Pearson correlation is used as a method. For this part and all parts that follow, Java (JDK 7) was used as a programming language if not stated otherwise.

2.2 VISUALIZATION OF CHROMOSOME FEATURE TRACKS

Visualization of feature distribution along the genome is always helpful for interpretation. We aimed to create a plot that not only allows the user to easily see domains, a.k.a. regions where the given feature was enriched or depleted, but also to compare these domains between different features.

Usually, a genomic feature is denoted by a list of genomic coordinates and, possibly, a score. For each chromosome we can transform these tracks of features into a coverage vector by dividing the chromosome into segments of 1 Mb and calculating the percentage basepair overlap between the feature and every segment. If a score is given, the average score of the segment is calculated considering the overlap, i.e. if only 30% of a segment are covered by a feature with a score of 100, the average score will be $100 \cdot 0.3 + 0 \cdot 0.7 = 30$.

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

x Single raw value

μ Mean of all values

σ Standard deviation of all values

For each feature and each chromosome we calculated the z-score or standard score (Equation 1) for the coverage vector and subsequently smoothed it using a moving average with a 3 Mb window.

We implemented a simple method for the identification of domains based on the z-score vector. Due to the nature of this score, domains

can be identified as large regions on the chromosome where almost all values have the same sign. These regions represent domains where the feature is either strongly over- or under-represented with respect to the remaining chromosome. We applied a approach based on a temporary sliding window to identify the boundaries of these domains for each chromosome:

1. Create temporary vector p, with $|p|$ being the length of the z-score vector
2. Slide window of w positions over z-score vector
3. Calculate average value of the sliding window and assign it to the middle position p_{middle} in p
4. Iterating over p, identify approximate domain boundaries as $sign(p_i) \neq sign(p_{i-1})$
5. Identify exact domain boundary in the original z-score vector as the first two adjacent positions in the window with center p_i that have the same sign as p_i

We tested different values for w and settled on $w = 11$. The smaller the window size, the more confined and exact are the domains. Since we were more interested in the overall banding into larger domains than exact division of the feature into positive and negative values, and thus want to allow some small deviations, we decided on a larger window size. Predicted domains are visualized as coloured bands in the feature track plots. Examples can be seen in section 3.0.4 on page 39.

Since visualization of multiple features as genomic context facilitates interpretation of a new feature's distribution, the plotting script per default plots the new feature together with a set of database features (Genes, GC content, LADs, Long Interspersed Nuclear Element (LINE), Short Interspersed Nuclear Element (SINE), Long Tandem Repeats (LTR), Hi-C compartments, RTD and DNase I hypersensitivity sites). The user can choose to calculate and visualize domains for each feature separately, or to transfer the domain banding from the new feature to the plots of others to compare them more easily.

Additionally to the visualization of tracks we used the above defined 1 Mb feature vectors to calculate the Pearson correlation between feature pairs per chromosome.

2.3 LAD BORDER ANALYSIS

Guelen et al. [69] first mapped the interactions of chromatin with the nuclear lamina in human. They also reported specific behaviour of other chromatin features such as genes and histone modifications

around the borders of LADs. Using our broad dataset of features, we performed a similar analysis with additional features on human and mouse, using Peric-Hupkes et al.'s [159] fibroblast dataset of LADs for human and the greedy set of LADs for mouse, which contains all constitutive and facultative LADs. The size of both sets is comparable (see below).

	Mouse	Human
#LADs	1470	1344
Cell type	'greedy'	fibroblast

Similar to Guelen et al. [69], we defined a flanking region of -400bp to +400bp around each LAD border and mirrored the left side onto the right. For a set of features comprising RTD, LINE, SINE, LTR and gene density, we calculated the average coverage (repeats) or average score (RTD) over all LADs for each of these 800 basepairs. Cases where LADs are close to chromosome ends were also taken into account, with the uncovered stretch of the 400bp region being set to a coverage of 0 for that specific LAD.

2.4 SYNTENY-BASED COMPARISON OF LINEAR GENOMIC FEATURES IN HUMAN AND MOUSE

This dissertation aims to analyse the conservation of genome organization in two mammalian species, human and mouse. In a first approach, we compare linear genome features based on ENSEMBL [56, 211] synteny regions. These regions represent long genome sequences in two species that have derived from a common ancestor. In our approach, we re-organize these regions in mouse to mirror the linear organization of the human genome, further termed mosaic chromosomes. Using these comparable genomes we can then calculate Pearson correlation coefficients on different re-organized feature tracks.

A. Calculation of 1 Mb feature coverage vectors for both species and all features

B. Creation of human mosaic chromosome from mouse data using synteny regions (see below)

C. Correlation of features across species, regular approach (tracks and Pearson coefficient)

Creation of human mosaic chromosome from mouse data

Information on synteny regions was downloaded from ENSEMBL Compara version 63 [56, 211] and included in the database. Recre-

ation of humanized mosaic chromosomes with mouse data was done in a straightforward approach:

1. For each synteny region *sr*:
   a) Find all Mb slices that overlap with sr_{human} and sr_{mouse}
   b) Reverse Mb slices from mouse if sr_{mouse} lies on the lagging strand
   c) Stretch or shrink the number of mouse slices to be in accordance with the number of human slices $|mb_{human}|$;
   For this procedure, each feature vector of the slices in sr_{mouse} is enlarged by repeating every element of the vector n times. We decided on $n = 20$ after trying out smaller and larger values, to combine computation speed with exactness. Afterwards, the blown up vector is binned into $|mb_{human}|$ sections, and for each section the average value is calculated.
   d) Plot the newly arranged tracks and calculate Pearson correlation coefficients

We performed the correlation analysis on each chromosome/mosaic chromosome pair from human and mouse and each feature, to get an impression of the conservation of feature distributions. Additionally, the data were visualized using previously described tools.

2.5 DEVELOPMENT OF A PIPELINE FOR INTEGRATION OF NEW FEATURES INTO THE GENOMIC CONTEXT

Based on our database of genomic features in human and other species, we developed a pipeline for the quick integration of a given new feature to set it into the context of the genomic landscape defined by our features. The goal is to interpret the genomic distribution of the new feature and identify similarities or dissimilarities to the distribution of existing features. Since most of these are correlated to some extent, we can draw conclusions from the locations of new elements.

Our pipeline achieves this goal by determining the correlation coefficient between the new feature's distribution and those of a list of database features, and by visualizing these for easy interpretation. The new feature has to be provided in the common .BED format, which comprises three columns defining the genomic position (chromosome, start, end) and an optional score column. The data are then prepared as described in section 2.2 and transformed into a z score vector for each chromosome.

The same pre-processing has already been done for the other genomic features (LINE, LTR and SINE, LADs, Hi-C compartments (i.e. Hi-C Eigenvector), gene density, DNase I hypersensitivity sites and RTD), so a Pearson correlation coefficient can be calculated for each

feature and chromosome. The pipeline generates a heatmap illustrating the similarity of the given feature's distribution with all others per chromosome which can easily be interpreted. As described before, the database features can be grouped into euchromatic (SINE, Hi-C compartment A, high gene density, open chromatin and early replication timing) and heterochromatic (LINE, LTR, LADs). If a new feature is preferentially located in active, euchromatic regions, the heatmap will be clustered into two parts, due to high correlation coefficients with euchromatic features and low coefficients with others. It is thus easy to directly classify a new feature as euchromatic or heterochromatic.

Pearson correlation coefficients are best suited for similarly frequent features. If the new feature is very rare, comparison to a dense database feature will result in a low coefficient. Our pipeline thus also provides visualization of the features distribution along each chromosome (see section 2.2), with a colour-coded domain pattern that is calculated based on abundance and depletion. This domain pattern is directly transferred onto plots of database feature distributions for facilitated comparison.

We have applied this pipeline to a number of new features provided by collaborators or experimental partners.

2.5.1 *LncRNA and their binding sites*

LncRNA are long non-coding (nc)RNA with a role in gene regulation [27]. In collaboration with Svetlana Vinogradova from Moscow State University, we analysed the location of lncRNA genes and binding sites across the human genome. To investigate the binding sites, we used experimental ChiRP-seq data by Chu et al. [34] on two lncRNAs' binding sites (HOTAIR and TERC) in the human genome. This dataset contains 832 binding sites for HOTAIR and 2,198 sites for TERC. We investigate the genomic distribution of both human lncRNAs' sites, focusing on HOTAIR.

ChiRP-seq (Chromatin isolation by RNA purification) identifies genome regions bound by an RNA in question

Additionally, our research partners analysed HOTAIR sequences in multiple genomes of HOTAIR and generated a set of exact GA-rich motifs. Constructing multiple sequence alignments for HOTAIR homologue sequences from human, dog, horse, mouse, rat and cow with ClustalW [65, 205], they were able to identify four different length (23-28 bp) conserved GA-rich motifs. Through scanning of the genome she further identified a total of 1,974 exact matches of these motifs in the human genome. We applied our pipeline to these predicted binding sites as well.

2.5.2 *miRNA*

Data on miRNA genes in the genome were taken from mirBase [106, 105, 68, 67, 66], version 10.0 to fit the hg18 assembly of human. Due

to short size of miRNA genes, we used their number per Mb segment for correlation analyses.

2.5.3 *NUMTS*

During the course of evolution, the once independent mitochondrial genome has been subsequently merged with the nuclear genome in eukaryotic cells. Not only genes moved to the nucleus, but also short fragments of mitochondrial DNA, termed NUMTS (Nuclear mitochondrial sequences) [208]. In a collaboration with the group from Paul Horton [208], we investigated the distribution of these NUMTS along the genome. Data were provided by Paul Horton in form of fasta sequence files and a tabular file containing NUMTS coordinates in the human genome as well as in the mitochondrial genome.

A total of 709 NUMTS spreading a total of 632,224 base pairs of the human genome are given in the data. The highest number of NUMTS is observed for chromosome 2 (78), while chromosome 18, which is rather short, contains the lowest number of only 7 NUMTS.

Additionally to our pipeline steps, we calculated the average overlap with four features (RTD, Hi-C compartments, DNase I hypersensitivity sites, LADs) and NUMTS, normalized by NUMTS length. We performed this analysis on the set of NUMTS and a set of expanded NUMTS which include a 400bp upstream and downstream flank. We created a set of randomized NUMTS by generating a random start position on the same chromosome for each NUMTS, and setting the same length.

3

RESULTS

It has long been known that domains of genomic features are correlated [84]. At the largest scale, chromatin can be classified into open, active or euchromatin and closed, repressed or heterochromatin. As described in the introduction, current models of spatial chromatin organization believe that the nucleus is partitioned into three-dimensional regions where sequences of either of these classes are co-localized. Understanding the complex relationship of chromatin features is thus an important pre-requisite for the work described in this dissertation.

Before analysing the three-dimensional structure of human and mouse in greater detail, we therefore investigated the interplay of known and newly discovered features along the human genome. We studied the correlations between features, visualized domains and compared species. As described in the previous methods section, we have developed a simple visualization tool on the basis of R [166] and Java that allows easy interpretation of the distribution of genomic features.

3.0.4 *Chromosome feature tracks enable visual comparison of domains*

Plots of chromosome tracks provide a good visualization of the distribution of different features along the genome, and it is well known that the structure of chromosomes in R-, G- and C-bands influences or even mirrors other chromosomal features [84, 85, 69]. Table 3 gives a short overview over the relationship between bands and other features, where R-bands are stretches of euchromatin, G-bands consist mainly of heterochromatin like C-bands, which almost exclusively contain satellite repeats.

As described in section 2.2, we applied a simple method for chromosomes' feature visualization. By smoothing the distribution with a sliding window and detecting domains with a sliding window approach, the user can easily identify regions of abundance and depletion of the feature (Figure 10). For comparison, Supplementary Figure S1 shows a straightforward visualization of z score feature tracks (page 191).

Chromosome track plots allow visual comparison of feature domains

Smoothing of the distribution helped to reduce noise from the feature tracks and see the domains more clearly. This is especially true for very varying distributions like gene density or LADs. Also, visualization of domains (domains with mainly positive values, i.e. regions of feature enrichment, highlighted in light pink, and depleted ones

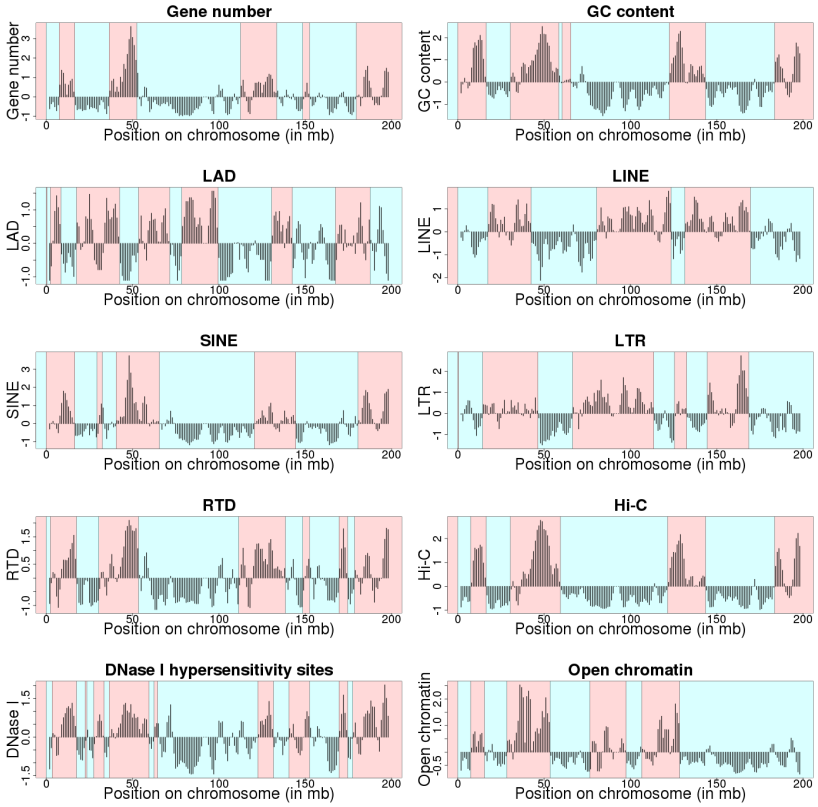


Figure 10: Visualization of z score feature tracks for human chromosome 3 with domain banding calculated for each track. Red bands mark regions where the feature is enriched, blue bands mark depletion. Patterns for correlated features such as SINE, Hi-C, gene number and GC-content are highly alike.

Table 3: Relationships between R-, G- and C-bands and different chromosomal features [84, 85, 69]. **RT:** Replication timing during S-phase

	R-bands	G-bands	C-Bands
Repeats	SINE *Human: Mainly ALU Mouse: Mainly B1*	LINE, LTR	Specific repeats *Very different between species*
Genes	gene-rich	gene-poor	almost no genes
GC-content	high	low	species-dependent, typically AT-rich
RT	early replicating	mid to late	very late
DNA compactness	open	closed	closed
Hi-C	A domain	B domain	not studied
LADs	depleted	enriched	not studied

highlighted in blue) obviously makes interpretation of each track as well as comparison of different ones easier.

For human chromosome 3 (Figure 10) it can be seen immediately that some features' domains, like LADs, tend to be shorter, while others such as Hi-C domains are fewer and often much longer. Similarities between the domain structures of all features are easy to spot: SINE repeats, gene number, GC content and Hi-C domains on chromosome 3 clearly show a domain organization which is highly alike. The same can be said for other features like LINE and LTR. The domain visualization makes spotting these similarities much simpler than the straightforward visualization.

These correlations can also be found in other human chromosomes (data not shown) and confirm the previously known relationships between features (Table 3).

3.1 FEATURES AT LAD BORDERS

Guelen et al. [69] not only conducted the first DamId experiment to detect LADs, genomic regions close to the nuclear lamina, they also showed that chromosome properties change significantly at the borders of these regions. According to their results in human, gene density significantly decreases at the beginning of a LAD and increases after it's end. Similarly, gene expression, Pol II binding and H3k4me2 is higher outside of LADs, while histone modification H3k27me3 is increased within compared to the direct genomic environment. Together, these properties mark LADs as repressive environment. Additionally, the sharp peaks of CTCF binding sites, CpG islands and promoters at the exact border of the LAD are very interesting.

As expected, LADs in human and mouse are positively correlated with LTR and LINE and negatively with gene density, SINE and early replication timing domains (Table 4 and Supplementary Figures S2 and S3 for corresponding heatmaps). In mouse, the correlation is stronger for all features except LTR, which rarely co-localize with LADs in this species. Still, a clear pattern of preferential location in LTR- and LINE-rich regions also emerges for human.

Table 4: Average genome-wide correlation of LAD distribution and other features. In mouse, strict LADs refer to the set of constitutive LADs that are consistent across cell types, while greedy LADs also include facultative elements. RTDs correspond to regions of early replication timing.

Feature	Human LADs	Mouse strict LADs	Mouse greedy LADs
LINE	0.3990	0.6308	0.7588
LTR	0.3398	0.1169	0.1317
SINE	-0.4208	-0.4949	-0.7692
Gene density	-0.2844	-0.4050	-0.4734
RTD	-0.4502	-0.4719	-0.7761

Applying a similar approach as Guelen et al. to human and mouse data, we performed a comparison of feature coverage around the borders of LADs. Figures 11 and 12 illustrate (unsmoothed) coverage curves around the border, with gray areas illustrating the flanking region inside the LAD and white areas for the outer flanks. Figure 11 shows the profiles of ENSEMBL protein coding genes and RTD values around the borders of LADs. We confirm the observation of Guelen et al. [69] that the gene density is higher outside of LADs (compare Supplementary Figure S4, page 194) in human and confirm this observation for mouse.

Gene density decreases at LAD borders in human and mouse

We can also observe a similar profile for the average RTD value, even though data from mouse are restricted to embryonic stem cells for this feature. While there is no clear pattern of enrichment or depletion within the LAD emerging, we can observe two peaks before and after the LAD border, connected by a local minimum located directly at the border. These results imply that the border regions of LADs, which are enriched in many insulating factors (see introduction), rarely coincide with domains of early replication. Instead, these domains tend to lie very close to the border. We can observe a second, more pronounced minimum at a position of +200 in human and +250 in mouse. However, though the profiles share similarities, they are not identical. For example, the average RTD in mouse LADs decreases more strongly than in human and is generally lower in mouse, though this could be caused by cell type differences.

Surprisingly, other than that there are no obvious similarities between the two species (Figure 12). As LINE and LADs both are en-

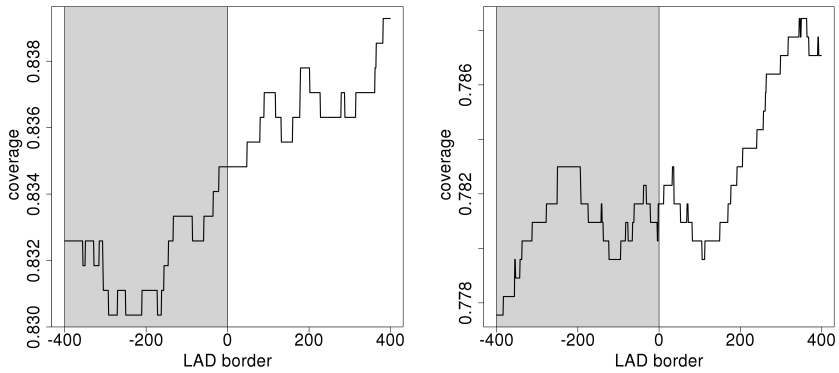
riched in heterochromatin, we expect an enrichment of LINE within LADs and a decrease at the border. In fact, we can observe such a relationship in mouse (Figure 12 (b)), but not in human. There, LINE repeats are frequent before and after the borders of LADs and less so within. In fact, the correlation coefficient for these two features in human is indeed lower than in mouse (0.3990 compared to 0.7588 in mouse, greedy LADs). In this case, while in mouse LINE and LADs appear to coincide more often, there is no such sharp decrease at the border in human.

Since SINE are mainly situated in euchromatic regions, we expect a reversed profile for this repeat. In mouse (Figure 12 (d)), the frequency of SINE increases before and after the LAD. However it appears that SINE frequency increases at the end of the inner flank, while it decreases farther away from the LAD. In human (Figure 12 (c)), the profile is much more pronounced. We can see a similar incidence of SINE repeats at around 370 nucleotides distance from the LAD border in either direction, but a sharp drop at the border itself. Guelen et al have shown that the border itself has a distinct genomic feature profile and is enriched in promoters, CTCF binding sites and CpG islands. Our results suggest that SINE, at least in human, rarely overlap with these highly regulated regions.

Repeats show different border profiles in both species

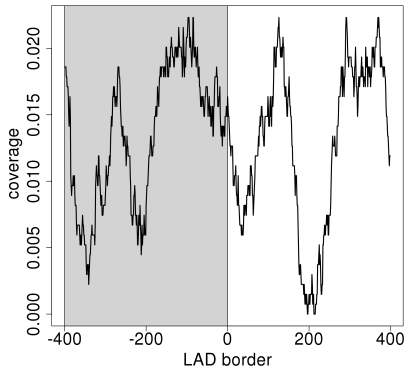
LTR are another class of long repeats that are often found in inactive genome regions. In human, these LTR behave similarly to SINE and we can observe a depletion of these elements at the border. In line with the preferential location in inactive regions, there is also a slightly higher incidence of LTR within the LAD after the trough at the border. As shown before, LTR are correlated with LADs in human (0.3398) in the same range as LINE, indicating a similar amount of co-localization. Again, the border profile in mouse is not similar, but instead rather opposite. We can observe a clear decrease in LTR frequency within the LAD, which is in with the very low correlation coefficients for these two features in mouse (0.1317 for greedy LADs). These results imply that in mouse LTR rarely co-localize with LADs.

We can conclude from this that LAD borders, which have been established to have distinct genomic and chromatin profiles in human, have different such profiles in human and mouse in most cases. While the profiles of gene density and RTD show similarities, repeats behave differently in both species and, with the exception of SINE in human and LTR in mouse, do not show a clear border-specific behaviour. It instead appears as if these repeat classes are not influenced by the domains at the nuclear lamina.

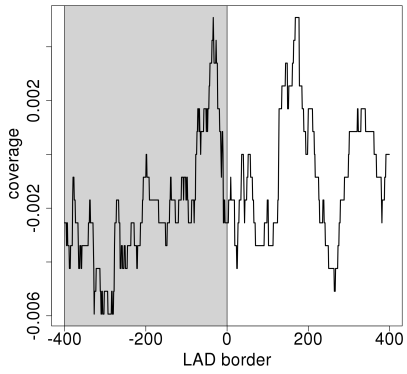


(a) Human genes

(b) Mouse genes

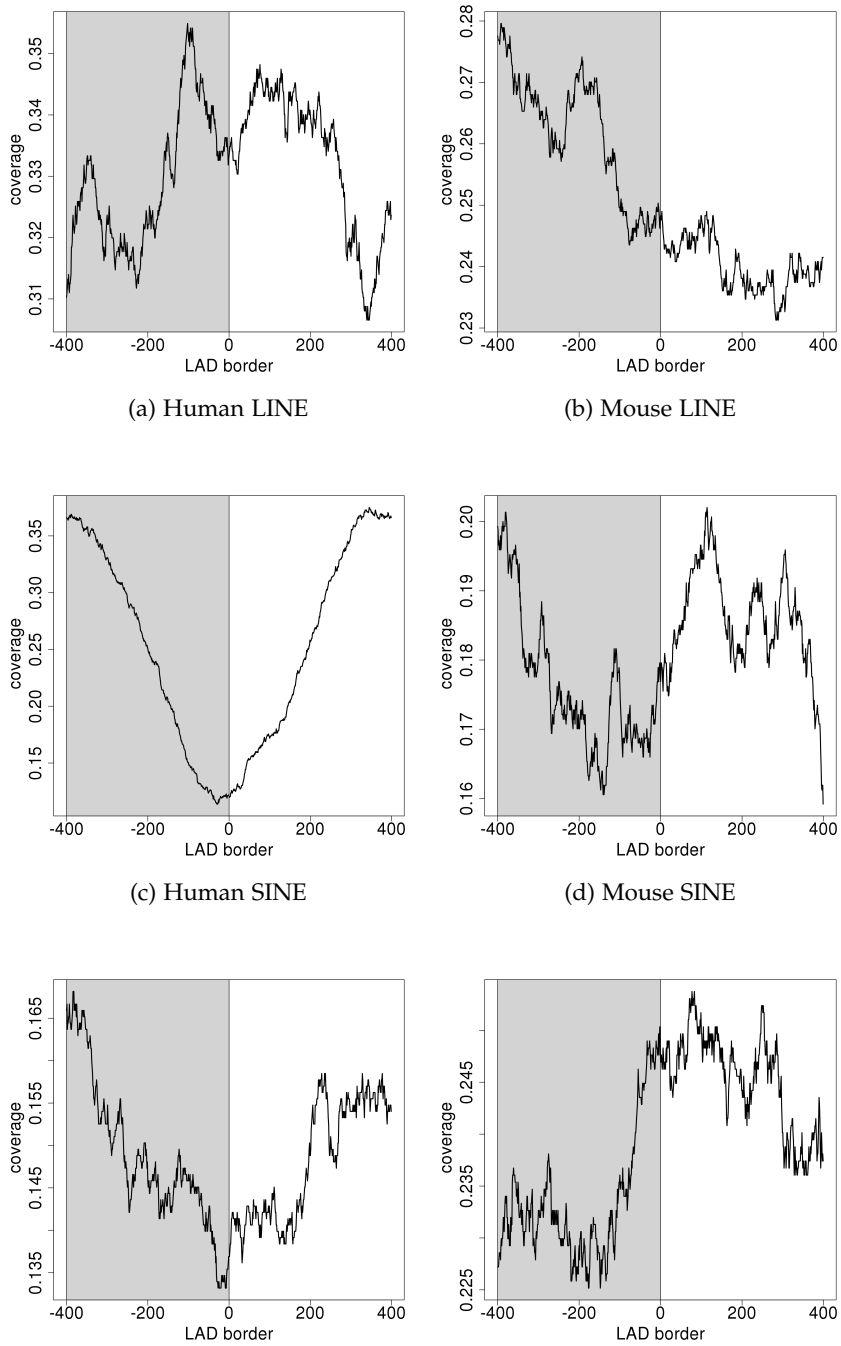


(c) Human RTD



(d) Mouse RTD

Figure 11: Feature density at LAD borders in human and mouse. Right and left border are mirrored, 0 on x-axis is LAD border, gray area shows the 400bp inner flank of the LAD, white area the outer flank of the LAD. Shown is average feature density over all LADs. Displayed is gene density and average RTD value, where high values represent early replication.



(a) Human LINE

(b) Mouse LINE

(c) Human SINE

(d) Mouse SINE

(e) Human LTR

(f) Mouse LTR

Figure 12: Feature density at LAD borders in human and mouse. Right and left border are mirrored, 0 on x-axis is LAD border, gray area shows the 400bp inner flank of the LAD, white area the outer flank of the LAD. Shown is average feature density over all LADs. Displayed is the average basepair density of certain repeat classes.

3.2 CONSERVATION OF LINEAR GENOME FEATURES IN MOUSE AND HUMAN

In the previous sections we have already described some similarities between the human and mouse genomes with respect to genomic features. In both species, generally as heterochromatic considered features localize at the nuclear lamina, while features that are related to gene expression and usually lie within gene-rich regions are located far from this peripheral subnuclear region.

Location of repeats, LADs, RTD, genes and GC content is (weakly) conserved between human and mouse.

We have also shown that, despite these similarities, differences can be found in both species. One example is the behaviour of SINE at LAD borders, which rarely overlap in human, but show no such pattern in mouse. In this section we aim to identify similarities within the distribution of features by re-organizing the mouse genome into a mosaic genome that mirrors human chromosomes. We limited this analysis to features that are directly comparable between the species. A heatmap illustrating the correlation is given in Figure 13.

Nowadays, synteny regards not only the order of genes but also other properties such as sequence similarity, so a positive correlation of gene-related features across species is not a direct consequence of the approach. However, it appears that all features show a weak to medium correlation across most chromosomes in human and mouse. Most similar features are gene number and GC content, with mean correlation coefficients of 0.30 and 0.37, respectively (Y chromosome omitted due to the high percentage of repeats). This is to be expected, because genes are under strong selective pressure and micro-rearrangements of genes rarely happen over multiple megabases (see part iv).

More interesting than features related to gene distribution are those that correspond to chromatin structure, as for example LADs or RTD. Correlation of replication timing domains across human and mouse is evident but low in megabase-scaled mapping, with a mean Pearson coefficient of 0.28. Location of LADs on the other hand are obviously less conserved between species, they even appear to be the least conserved feature together with LTR (mean correlation coefficient for LTR: 0.14, LADs: 0.15). As we have already shown that the mouse genome exhibits only very low correlation between LTR and LADs, contrary to human, it can be expected that either of these two features or both are not well conserved between human and mouse.

Altogether it can be concluded that distributions of gene-related features as well as LINE and SINE feature tracks correlate quite well between human and rearranged mouse data, indicating some level of evolutionary conservation. Features describing chromatin structure, in this case LADs, correlate only weakly between both organisms and thus lead to the conclusion that location of chromosome structural elements is only loosely conserved at the megabase scale.

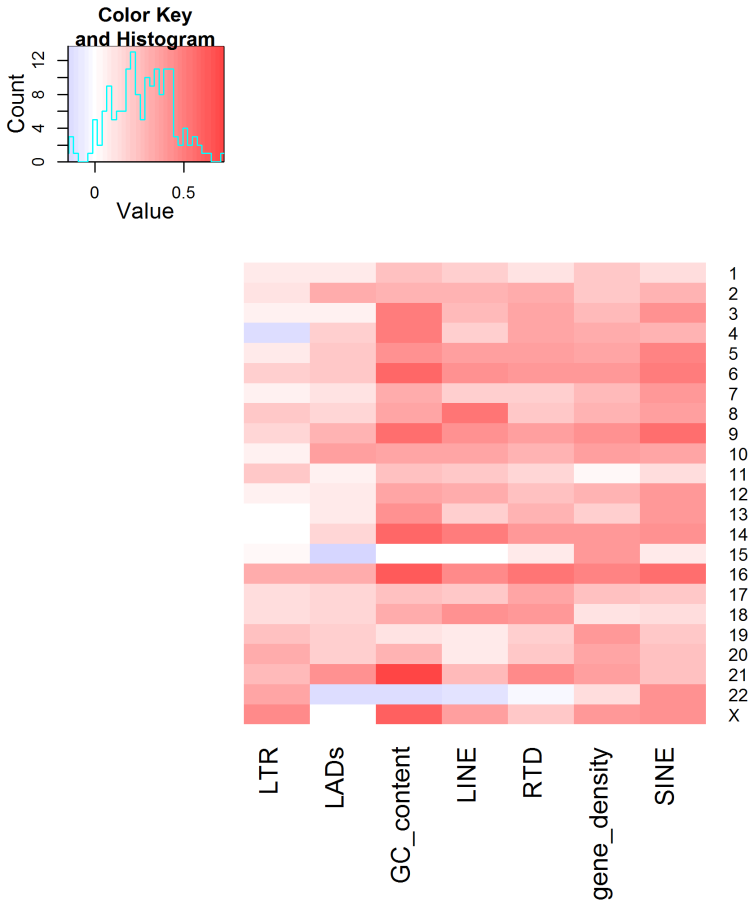


Figure 13: Heatmap of inter-species Pearson correlation coefficients (PCC) for each pair of features from human and mouse, respectively. The mouse genome was rearranged based on synteny information to make it comparable to the human genome. All features show a weak to medium positive correlation over megabase segments, indicating a low level of conservation. Chromosome Y is excluded because of its short length and high percentage of repeats. Chromosome 22 shows less concordance with other chromosomes, probably due to its short length.

However, it has to be kept in mind that the humanized mouse chromosomes do not actually represent chromosomes, but rather parts of them that are rearranged for the purpose of comparison. Some differences in the exact position of LADs can therefore be expected. In the third part of this dissertation we thus analyse similarities of the three-dimensional structure of human and mouse genomes, focusing on functional aspects. Additionally, in the fourth part we explore a different synteny-based mapping that focuses on genes and intergenic regions instead of fixed lengths regions.

3.3 LONG NCRNA AND THEIR CORRELATION TO OTHER FEATURES

Noncoding (nc) RNAs longer than 200 nucleotides are classified as long ncRNA, a very abundant class of transcripts in the mammalian genome [28]. Their conservation across species is considerably lower than conservation of small regulatory ncRNAs, indicating a lack of functional relevance [199]. However, similarly low conservation has been shown for well characterized and functionally important long ncRNA [141], leading to the suggestion that this class of ncRNA is under different selection pressures [154].

Long ncRNAs (lncRNA) have been shown to be involved in multiple cellular processes, among which is epigenetic regulation through imprinting or X-chromosome inactivation (e.g. [225, 100]; see section 1.2.7). Recently, a large number of lncRNA that are directly associated with chromatin modification complexes have been identified [103, 236]. The emerging close relationship of lncRNA and chromatin state and structure led us to analyse the correlation between lncRNA and other chromatin features such as LADs or histone modifications. Of special interest for this comparison are the lncRNA binding sites, a large number of which has recently been experimentally discovered for three lncRNAs in human and drosophila with ChiRP-seq [34]. Because of their regulatory and possibly chromosome structure mediating effect, we focus on these binding sites in the following sections.

3.3.1 *Long ncRNA tend to bind in euchromatic regions*

We investigated the distributions of lncRNA binding sites, since these may directly influence chromatin state. We analysed 3,030 binding sites of TERC and HOTAIR as provided by Chu et al. to find trends in their relationship with other chromosomal features. Correlation coefficients per chromosome can be found visualized as a heatmap in Figure 14.

TERC and HOTAIR bind to euchromatic regions in the human genome

As mentioned before, genomic features can be clustered into those that occur mainly in euchromatic regions and those that lie within heterochromatin. LncRNA binding sites show a medium to strong positive correlation with euchromatic features such as genes, DNase I sites and SINE repeats. For GC-content, Hi-C eigenvectors and RTD, high values are also associated with active chromatin and high gene density, so they can be considered to be markers for euchromatin as well.

LINE and LADs on the other hand are mainly found in heterochromatin. Negative correlation of lncRNA binding sites with these two features confirms a preferred binding of HOTAIR and TERC within gene-rich regions of the genome. Tracks (see Figure 15) confirm this

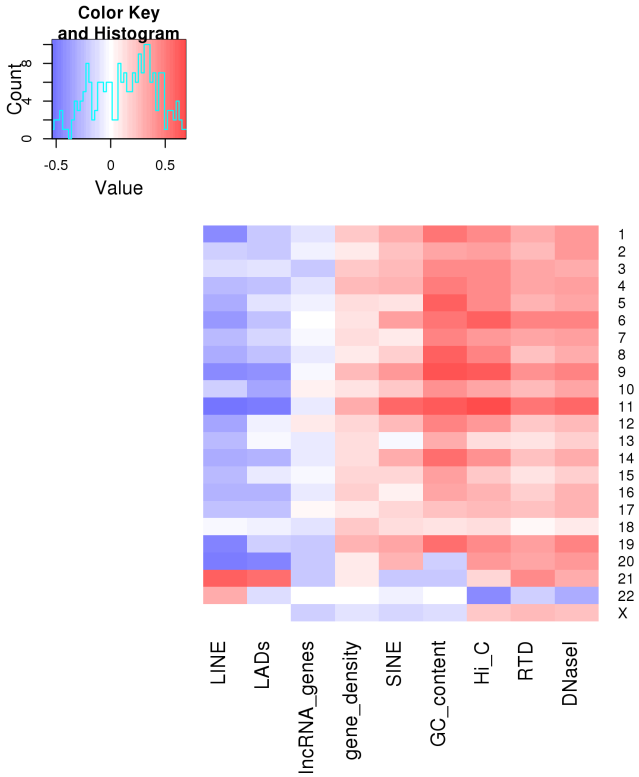


Figure 14: Correlation coefficient values (Pearson) of HOTAIR and TERC (lncRNA) binding sites to other genomic features in human. Again, short chromosomes deviate from the rest.

observation, showing clear domains of enrichment in lncRNA binding sites around the same positions that Hi-C domains, equivalent to Hi-C compartment A, occur (compare Figure 10).

LncRNAs thus appear to target open euchromatic regions with high gene-density, but tend to ignore heterochromatic regions that are close to the nuclear lamina. Chu et al. suggested that lncRNA might act like sequence-specific dictators of chromatin states [34]. They showed co-occupancy of lncRNA HOTAIR and Polycomb domains and hypothesize that the RNA may be involved in recruiting the latter. Polycomb proteins usually invoke epigenetic silencing of genes. Our observations show that HOTAIR and TERC preferentially bind in gene-rich regions, in line with a potential role in gene regulation.

Predicted HOTAIR binding sites are weakly correlated with histone modifications

In a RECESS collaboration with Svetlana Vinogradova (Moscow State University) and based on the suggested ability of lncRNA HOTAIR to directly bind to the DNA molecules through formation of a triple-

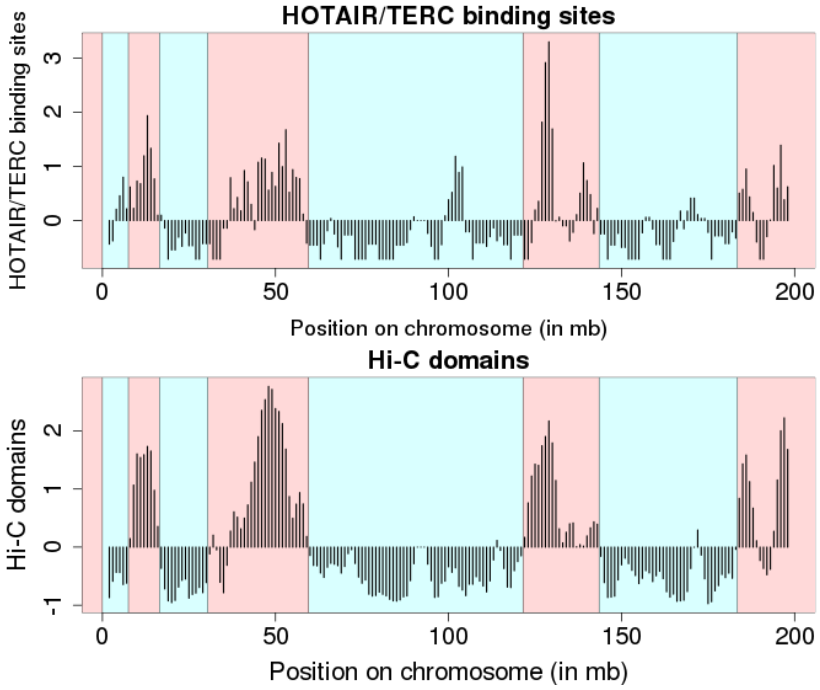


Figure 15: Human lncRNA genes and binding sites on chromosome 3. Red bands represent Hi-C compartment A, which contains predominantly active chromatin, while blue bands represent the complementing compartment B. Most binding sites of HOTAIR and TERC lie within compartment A.

helix structure [207, 135, 55], two datasets consisting of potential HOTAIR binding sites on the human genome were created by her. We aim to complete this analysis through comparison of these predicted binding sites to genomic features that determine or are influenced through chromatin structure. Formation of a DNA-RNA triplex changes the rigidity of the DNA and therefore might participate in chromatin organization.

Predicted HOTAIR binding sites with substitutions have different features than experimental ones

The first dataset, in the following referred to as 'exact motifs', comprises only exact matches of 13bp long GA-rich motifs from HOTAIR's sequence or complementary sequence in the human genome. The second dataset allowed for 6 substitutions in the sequence.

As described before, our genome feature dataset contains a multitude of chromatin organization related tracks for the human genome. For this comparison we added histone modifications H3k27ac, H3k4me3 and H3k4me1 to the previously mentioned features. We have already shown that experimental HOTAIR and TERC binding sites are localized in gene-rich regions of the genome.

With only 1,974 exact matches, the set of predicted lncRNA binding sites is considerably sparse compared to other features. We thus expect to see only weak correlations with dense euchromatic features such as DNase I hypersensitivity sites (970,658). Figure 16 shows that this is, in fact, true for exact HOTAIR motifs. They are on av-

erage weakly positively correlated with euchromatic features and histone modifications, but are rarely located in heterochromatic regions. There are chromosomes for which this trend cannot be observed (e.g. chromosome 16), which is probably caused by the sparsity of the feature.

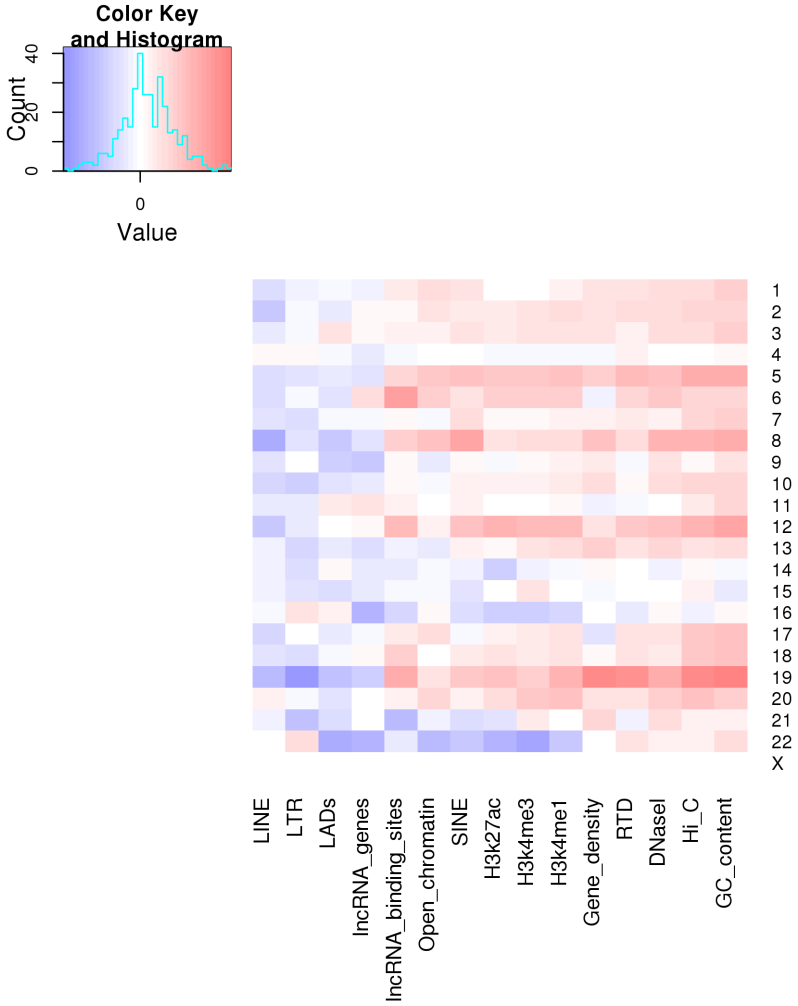


Figure 16: Heatmap of Pearson correlation coefficient between HOTAIR exact binding site motifs on the genome and other genomic tracks. Correlation is weakly positive for most chromosomes for euchromatic features.

When we consider the larger set of 5,566 predicted binding sites with substitutions, we do not find a clear correlation pattern (Supplementary Figure S5). Only for a subset of chromosomes (e.g. chromosome 2) can we observe the distinctive positive and negative correlations with euchromatin and heterochromatin, respectively. Considering the higher density of this dataset we would expect the correlation to become clearer than before. Since this is not the case, we conclude that the predicted motifs with substitutions are no longer functional

HOTAIR motifs, and that instead higher sequence identity is necessary to maintain the function.

3.4 MICRORNA AND THEIR CORRELATION TO OTHER FEATURES

MicroRNAs (miRNAs) are a class of naturally occurring, small non-coding RNA molecules about 21-25 nucleotides in length. They were first described in 1993 by Lee et al. [114] and the term microRNA was introduced in 2001 by Ruvkun et al. [175]. MicroRNAs are partially complementary to one or more messenger RNA (mRNA) molecules and play an important role in the complex network of gene regulation. It is their main function to regulate gene expression in a variety of manners including translational repression, mRNA cleavage, and deadenylation [5, 54]. Each miRNA is thought to regulate multiple genes, and since hundreds of miRNA genes are predicted to be present in higher eukaryotes [121], the potential regulatory circuitry afforded by miRNA is enormous. Several research groups have provided evidence that miRNAs may act as key regulators of processes as diverse as early development [169], cell proliferation and cell death [23], apoptosis and fat metabolism [229], and cell differentiation [48, 31]. Recent studies of miRNA expression implicate miRNAs in brain development [107], chronic lymphocytic leukaemia [26], colonic adenocarcinoma [131], Burkitt's Lymphoma [130], and viral infection [161], suggesting possible links between miRNAs and viral disease, neuro-development and cancer.

The genes encoding miRNAs are much longer than the processed mature miRNA molecule. Often miRNAS are located in introns of their pre-mRNA host genes. They share their regulatory elements, primary transcript and they have a similar expression profile. MicroRNAs are transcribed by RNA polymerase II as large RNA precursors called pri-miRNAs and come complete with a 5' cap and poly-A tail [115]. After further procession, the resulting pre-miRNAs are approximately 70 nucleotides in length and are folded into imperfect stem-loop structures. Once in the cytoplasm, the pre-miRNAs undergo an additional processing step by the RNAse III enzyme Dicer, generating the miRNA, a double-stranded RNA approximately 22 nucleotides in length. Dicer also initiates the formation of the RNA-induced silencing complex (RISC) [76]. RISC is responsible for the gene silencing observed due to miRNA expression and RNA interference [77].

As another level of gene expression regulation, correlation of miRNA origin sites and other genomic features is of interest. One might expect that the often intronic location of miRNAs leads to high correlations with gene density. Besides that we are interested in whether or not miRNA sites tend to lie in readily accessible regions of the chromatin and thus show a negative correlation to LADs.

3.4.1 *There is no direct correlation of miRNAs and genomic features*

Pearson correlation of miRNA number per chromosome slice (1 Mb) and other genomic features is illustrated as a heatmap in Figure 17. Due to the sparsity of the feature, most features show only very weak correlations with miRNA number, that are also inconsistent over the different chromosomes. However, though the correlation coefficients overall are low, a trend for negative correlations with LINE and LADs can be observed, with coefficients up to -0.27 and -0.33, respectively.

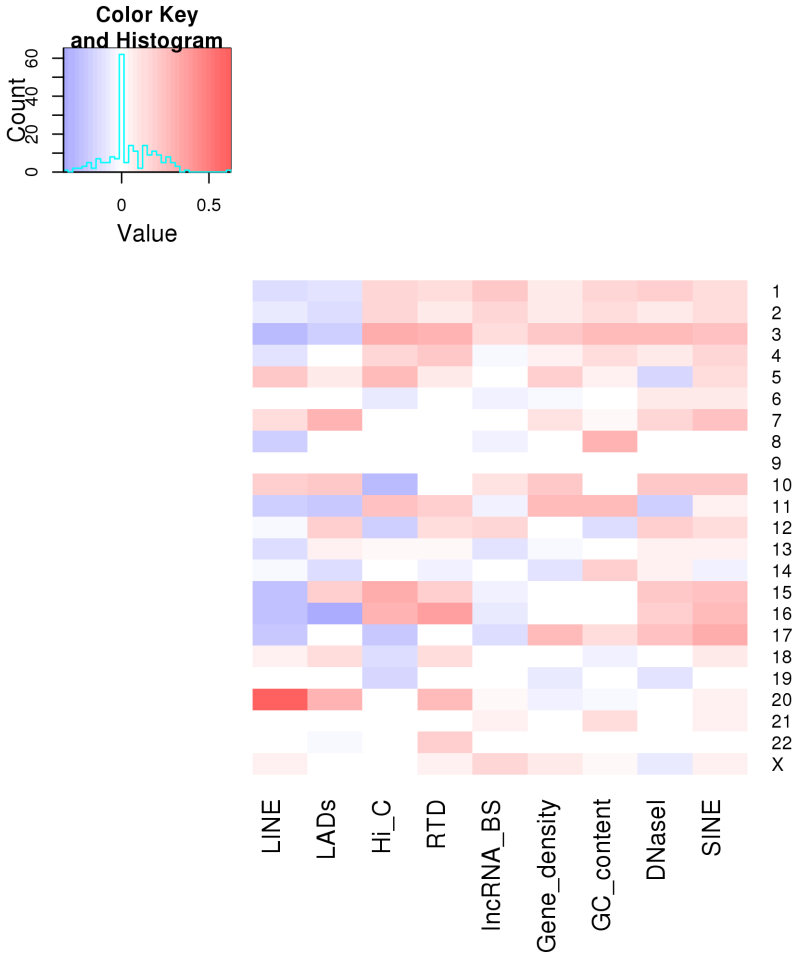


Figure 17: Correlation coefficient values (Pearson) of miRNA gene distribution to other genomic features in human. Due to sparsity of miRNA genes, correlation with other features is weak and a preference for euchromatic regions is only fable.

However, the plot of the miRNA track can only confirm this observation to some extent (compare Figure 18). In direct comparison to the human LADs, miRNA sites appear to be less abundant on this chromosome and only a weak correlation of both tracks over the whole length of the DNA strand can be seen. The negative correla-

tion value of -0.19 for chromosome 3 mirrors this and is caused by the few regions of miRNA abundance mostly but not always falling into inter-LAD regions. Since miRNA genes often are located in introns of genes, and genes have a positive correlation with inter-LAD regions, this result is probably an artefact of the gene-miRNA-relationship.

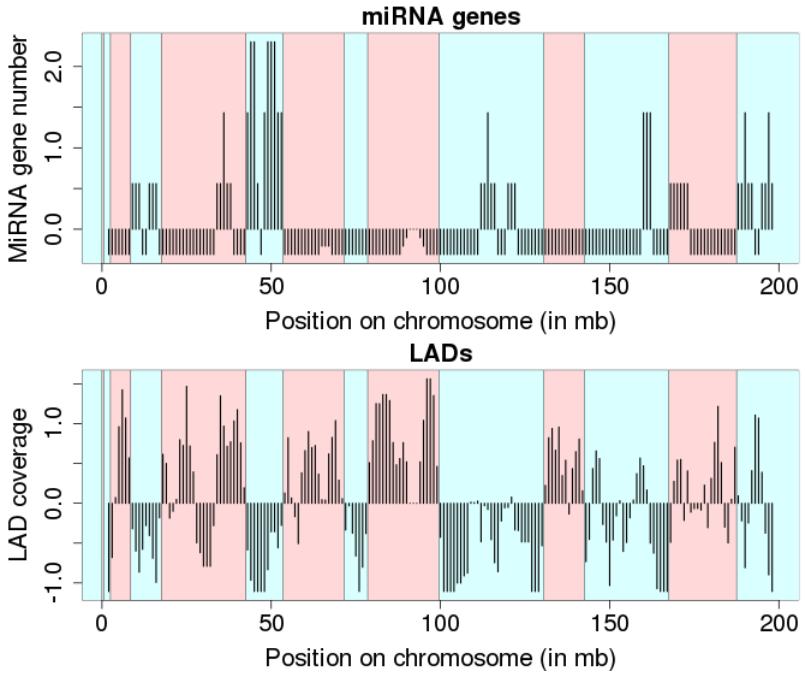


Figure 18: Human miRNA origin sites and LADs on chromosome 3. Red bands signify LADs, blue bands inter-LADs.

It thus appears as if miRNA origin sites do not generally lie in accessible genome regions. The observed slight negative correlation with heterochromatic genome features can be explained by a number of miRNA genes that lie within introns of genes.

3.5 NUCLEAR MITOCHONDRIAL SEQUENCES LIE IN ACCESSIBLE GENOME REGIONS

Despite being organelles in cells of eukaryotic organisms, mitochondria are commonly believed to have evolved from bacterial ancestors, explaining why they have their own small and circular genomes [178]. Only a small proportion of the proteins required for ATP production in mitochondria are encoded by this genome, leading to the hypothesis that many of originally mitochondrial genes were somehow transferred to the nuclear genome during the cause of eukaryotic evolution. Not only genes have migrated into the nucleus and were inserted into chromosomal DNA, but also other mtDNA (mitochondrial DNA) fragments, termed NUMTS (Nuclear MiTochondrial sequences) [50, 80].

Horton et al. conducted a global analysis researching the circumstances of NUMTS insertion into the genome of several eukaryotic organisms [208]. They conclusively showed that this process is not random and that insertion sites tend to have several characteristic properties: Proximity to retrotransposons, but no insertion into these, proximity to regions with high local DNA curvature and regions with high A+T rich oligomers, mainly TAT.

NUMTS: *Nuclear MiTochondrial sequences, non-gene fragments of mtDNA that migrated into the nuclear genome*

Since the insertion mechanism appears to rely on non-homologous end joining repair, Horton et al. hypothesize involvement of L1-EN, an endonuclease. Another explanation for the proximity to retrotransposons could be the limited accessibility of the genome in germ line cells. If only a small proportion of the DNA is accessible to both NUMTS insertion and generation of retrotransposons, these would appear in close proximity very often.

Previous research has shown that NUMTS lie close to retrotransposons

One aspect which was not tackled by Horton et al. are correlations of NUMTS insertion sites to epigenomic or chromosome structure features. We aim to complement their research by conducting correlation analysis to such characteristics, such as LAD coverage, Hi-C compartments or RTD.

3.5.1 *NUMTS distribution along the genome is too scarce for linear correlation analysis*

As for other features before, we calculated NUMTS coverage per 1 Mb segment of each chromosome and performed correlation analysis to other features. Results are given as a heatmap in Supplementary Figure S6.

Both the heatmap and the actual values clearly show that only a very weak correlation exists between the genomic NUMTS distribution and any of the other structural or epigenomic features. Surprisingly, even features shown by Horton et al. to be characteristic for NUMTS insertion, such as GC content or LINE coverage, did not result in a high correlation. One could argue that this is caused by the division of each chromosome into slices of 1 Mb size. Short elements as NUMTS, which cover only 632 kb of the human genome in total in the dataset of Horton et al., can easily be overlooked in such an analysis. Additionally, we are dealing with a total of only 709 NUMTS regions, whereas 1,344 LADs are available for human. A slight negative correlation can thus be expected, since for about half of all LADs no corresponding NUMTS can be present.

This clearly shows that our pipeline is only suited for sufficiently dense features, since otherwise the symmetric method of correlation fails. As a consequence, we performed an additional analysis on this dataset that focuses on the average overlap of NUMTS and other features, compared to randomly distributed elements of the same size.

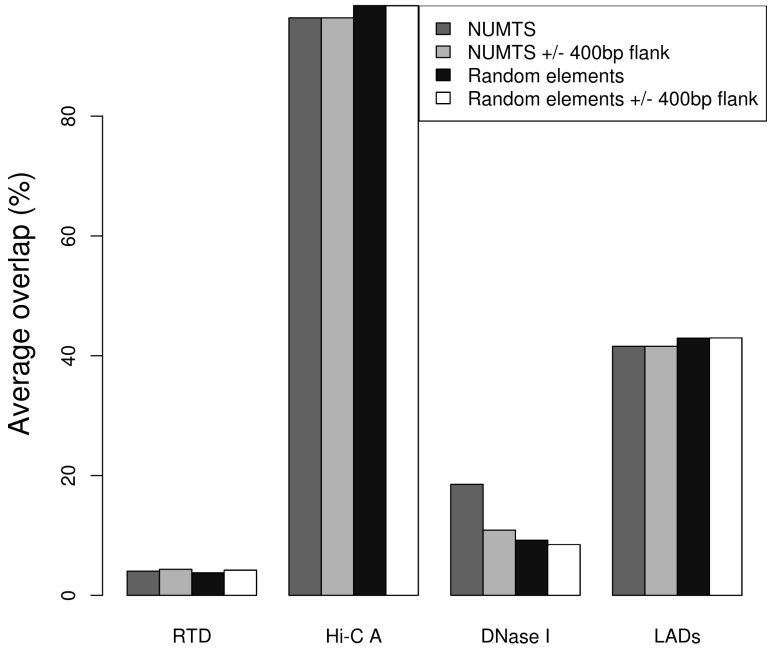


Figure 19: Average percentage of overlap with another genomic feature in NUMTS with or without two 400bp flanks and randomized elements with or without flanks. DNase I hypersensitivity sites are enriched in NUMTS.

Figure 19 illustrates the amount of overlap between NUMTS (with or without 400bp upstream and downstream flanks) and a randomized set of NUMTS (with or without flanks) for comparison. The average overlap with LADs, Hi-C compartments and RTD is very similar in both sets, implying no specific preferential insertion in either active or inactive sites. However, the data clearly show a co-occurrence of DNase I hypersensitivity sites and NUMTS. Horton et al. have not investigated this relationship, though they report a lack of co-occurrence for DNase I hypersensitivity sites and retrotransposons at NUMTS insertion sites. Our results show that NUMTS preferentially lie in highly accessible regions, while their flanks overlap with these regions less often and more closely resemble random data.

NUMTS coincide with DNase I hypersensitivity sites

Horton et al. hypothesized that genome accessibility in the germ line cells could limit both insertion of retrotransposons and NUMTS. Our results confirm that genome accessibility is high in NUMTS. However, since this accessibility decreases in NUMTS flanks, it can be both a prerequisite or a consequence of the insertion. It thus appears as though their first hypothesis of L1-EN as an influencing factor might also play a role.

4

CONCLUSION

In this part we have investigated the role and distribution of different linear genomic features, their interplay and, superficially, their conservation in the human and mouse genomes. To achieve this we have collected data from various sources and created a database of intrinsic genomic features, comprising sequence-based properties such as GC content or repeats, but also domain-like features like LADs or RTD, and elements that describe chromatin structure, mainly histone modifications, DNase I hypersensitivity sites and Hi-C compartment vectors.

We have developed simple methods to analyse the correlation of these features mathematically and visually, which serve well to classify new features with respect to this complex web of elements. Using these, we confirm the previously reported inter-dependency and domain-structure of many genomic properties. In fact, features can be grouped into those preferentially located in euchromatic areas (genes, SINE, Hi-C compartment A, open chromatin, DNase I hypersensitive sites and RTD) and those that prefer heterochromatic areas (LADs, LINE and LTR).

Classifying new, mainly experimentally determined, features into this list, we were able to show that the binding sites of lncRNAs HOTAIR and TERC show a clear preference for active or euchromatic genome regions. miRNA genes, however, are not located in genome regions with specific properties. We have also investigated remnants of mitochondrial DNA (NUMTS)in the human genome that were determined experimentally by Horton et al. [208], and found them to coincide with DNase I hypersensitivity sites, implying a preferential insertion in accessible genome regions. This conclusion is a good example of how the interplay of features can help us better understand biological processes.

As this thesis has a strong focus on the three-dimensional structure of the genome, we investigated domains at the nuclear periphery, termed LADs, and their relationships with other features in more detail. Guelen et al. [69] have already investigated density changes of certain genomic features at LAD borders. It has been shown that regions at the nuclear periphery are mainly inactive, while those in the nuclear center are more active. One would thus expect density of active features to increase at these borders, and Guelen et al. have shown that the border itself has a distinct feature profile. We complement this research and show that, in both human and mouse, the gene density rises at the border of these regions, and RTD borders

tend to coincide with them. However, repeats do not show a distinct density change at LAD borders and also behave differently in both species.

Part III

CONSERVATION OF THE INTER-CHROMOSOMAL SPATIAL CHROMATIN ORGANIZATION

Inspired by biological networks for protein-protein interactions and others, we can transform Hi-C data into an easy to interpret network graph. This part describes the analysis of new Hi-C data for mammalian species *H. sapiens* and *M. musculus* and the interpretation and comparison of their respective inter-chromosomal interactomes in the form of networks.

5

GENERATION OF PHYSICAL SEGMENT AND GENE INTERACTION NETWORKS ON HI-C BASIS

In recent years, the advance of methodology has brought on a large amount of research on the chromatin interactome, with most analyses relying heavily on forms of chromosome conformation capture (see Section 1.3.2) such as the whole-genome experimental method Hi-C. In this method genomic regions close in space are cross-linked, followed by cutting the genome with restriction enzymes and ligating cross-linked fragments before massively parallel sequencing. Using Hi-C, the chromatin interaction maps of various cell types have been reconstructed in the form of interaction probability matrices.

However, it is known that many experimental biases accumulate in Hi-C experiments, among them the non-uniform distribution of restriction enzyme cutting sites in the genome, differences in read mappability and non-specific ligation. Additionally, the data contain a high amount of noise due to random interactions between genomic segments. Since Hi-C data are averaged over millions of cells, these random interactions accumulate and further obscure real information. As many more inter-chromosomal interactions are theoretically possible than within chromosomes, this problem is especially present in distinction of contacts between chromosomes from noise. Current research focuses mainly on intra-chromosomal contacts, mainly due to a better signal, but also because more data are available. Single-cell Hi-C is an option to improve the signal-to-noise ratio, however, only little data from such experiments are available so far.

Hi-C data are usually represented as interaction probability matrices for pairs of chromosomes. A complementing representation was proposed by Lieberman-Aiden et al. [120], who developed the Hi-C method in their lab. According to their publication on the human chromatin interactome, the data can be effectively reduced to the first Eigenvector of a principal component analysis (PCA), which captures the propensity of each genomic segment to lie within one of two nuclear subcompartments. The interactions within these subcompartments are high, while there are sparse interactions between these two compartments, leading to the common plaid pattern shown in Figure 7 on page 22. However, Eigenvector-representation is best suited for intra-chromosomal data, where the signal-to-noise-ratio is higher and less random interactions obscure the two-compartment-pattern. In addition, though Eigenvectors greatly reduce the information from the original data matrices, they are not easy to interpret without pre-

vious knowledge, and information is lost if the compartment structure is not as precise in the data as in the original publication.

Kruse et al. [108] recently proposed a framework in which the inter-chromosomal chromatin interactome is represented as a network of physically interacting genes or genomic segments. Combined with normalization to remove all Hi-C biases, as developed by Yaffe and Tanay [230], they successfully investigated the interactome of *S. cerevisiae* and found a high degree of centromere clustering. Network-based analysis has also greatly improved our understanding of protein-protein interactions and regulatory networks. Its main advantage is easy interpretability and straightforward pattern recognition. For instance, it has long been known that mammalian genes cluster together spatially in transcription factories, which could be easily identified in a network of physical gene interactions. Investigating overlaps with co-regulation and co-expression networks of genes could help us understand the role physical genome organization plays in the regulation of gene expression.

Currently, much is known about the structure of intra-chromosomal gene organization into TADs, but little research has been conducted to investigate the potential functionality of inter-chromosomal contacts. Additionally, Hi-C data are often explored in an isolated manner in single species. We aim to transform the high quality data from Dixon et al. [45] from *H. sapiens* and *M. musculus* embryonic stem cells (hESC and mESC, respectively) into inter-chromosomal gene and segment interaction networks for mammalian organisms, analyse their topologies and possible functionality. We lay our focus on the holistic comparison of two mammalian species' inter-chromosomal interactomes to investigate to which degree the three-dimensional structure of the genome is conserved. As such, this work is the first research conducted on the conservation of inter-chromosomal spatial chromatin organization in mammals.

6

MATERIALS AND METHODS

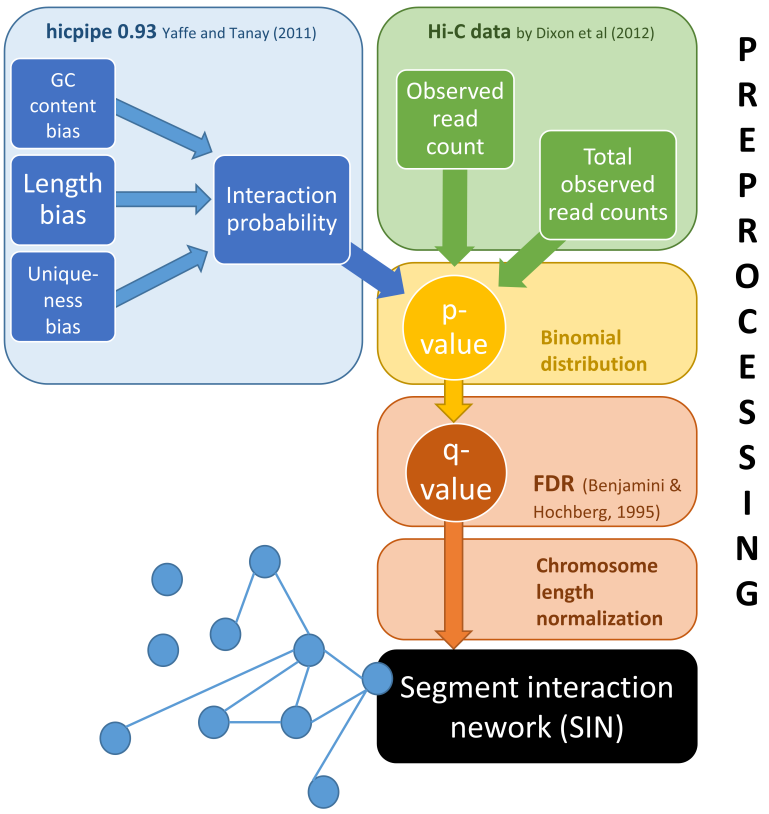


Figure 20: Workflow of network generation approach for Hi-C data adapted from Kruse et al. [108] Only inter-chromosomal Hi-C reads were used as input.

We have adapted the method by Kruse et al. [108] for the generation of segment interaction networks on basis of Hi-C data, for large mammalian genomes. Figure 20 shows the general workflow of the procedure, which includes a normalization method from Yaffe and Tanay [230] (blue box) to calculate interaction probabilities for all segment pairs. This method, called hicpipe, estimates the experimental biases to determine the background probability for interactions. Hi-C data from Dixon et al. [45] are subjected to this normalization and filtered using p- and q-values to create binary contact matrices for both species. Based on these, the segment and gene contact networks are reconstructed and subsequently analysed.

6.1 DATA SOURCE AND PREPARATION

Experimental Hi-C data from Dixon et al [45] for *H.sapiens*, assembly hg18, and *M.musculus*, assembly mm9, are downloaded from GEO (GSE35156). For comparability we used ESC data from both species.

6.1.1 *Normalization*

Due to a high number of known biases influencing the outcomes of Hi-C experiments, normalization of chromosome interaction matrices is essential. As proposed by both Kruse et al. [108] and Dixon et al. [45], we chose the advanced normalization method implemented by Yaffe and Tanay [230], which includes correction of read mappability, elimination of non-specific ligation products and considers length and GC content biases.

Normalization of Hi-C data is necessary to remove known biases

Yaffe and Tanay offer a software solution for this normalization procedure (hicpipe, version 0.93). It requires experimental Hi-C data in a summarized format as well as genomic locations of restriction enzyme cutting sites as input to reconstruct where in the genome the Hi-C fragments arise. Because Hi-C experiments differ in many parameters, the calculation of read mappability and restriction enzyme fragments has to be performed by the user.

In our case, HindIII restriction enzymes were used for Dixon et al.'s Hi-C experiments, and BWA [116] with default parameters to map the resulting Illumina reads back to the genome.

Identification of HindIII fragments in the human and mouse genomes

To find all HindIII restriction enzyme cutting sites in the human and mouse genomes, we used the R bioconductor packages `BSgenome.Hsapiens.UCSC.hg18` and `BSgenome.Mmusculus.UCSC.mm9` [152]. These packages are based on Biostrings [153] and represent the whole genome sequences. Biostrings offers quick text searches with or without mismatches and thus serves perfectly to find short restriction enzyme recognition sequences.

HindIII recognizes the six nucleotide long palindromic sequence `AAGCTT`, cutting after the first `A` on both strands. Using the Biostrings `matchPattern()` function with this sequence on each chromosome, without allowing mismatches, we were able to very quickly find all HindIII cutting sites in the human and mouse genomes.

Hicpipe demands fragment end information as input file, so simple restriction enzyme cut sites do not suffice. Fragment ends are sequence regions that start or end at a restriction enzyme cut site, i.e. the ends of all fragments the genome can be cut into by this particular enzyme. Considering the asymmetric cut after the first nucleotide of the recognition sequence, we could calculate every fragment with two corresponding fragment ends.

We received a total of 1,673,258 fragment ends for human, confirming the number published by Yaffe and Tanay, and 1,646,704 fragment ends for mouse.

Calculating the mappability score for the fragment ends

It is necessary to account for the different so-called mappability of reads to the genome; reads that map at many different positions in the genome should be discarded before continuing with the analysis. The mappability score can thus be considered as a measure of sequence uniqueness. We are calculating the mappability score for each fragment, using a 500 bp cutoff to limit the fragment length. For each HindIII cutting site we are thus left with two fragments that expand 500 bp in either direction.

To calculate the mappability score, we created a high number of artificial reads by breaking each chromosome into overlapping sequence regions of 50 bp length, starting every 10 bp, as suggested by Yaffe and Tanay [230]. These artificial reads were then filtered for information content, discarding every all-N sequence, and transformed into an artificial fastq file, and read quality was set to medium (*I*) for all bases. For the reference genomes we downloaded the `.2bit` versions of the human and mouse genome, assemblies hg18 and mm9, from UCSC [102] and converted them to fasta format. Using the same mapper as Dixon et al., BWA-backtrack samse[117] for single-end reads with default parameters, we first created an index on the genome sequences and then aligned the artificial reads to the genomes. The resulting custom BWA file could then be converted into the alignment format `.sam` and subsequently be converted into `.bam` file format using Samtools sam2bam [118].

Table 5: Statistics on the mapping of artificial 50 bp reads back to the genome from which they were created.

	Human (hg18)	Mouse (mm9)
Total #reads	184,541,039	174,629,674
#Duplicates	0	0
#Mapped	184,430,985 (99.94%)	170,863,902 (97.84%)

Table 5 shows some samtools flagstat [118] statistics on the mapping. The quality of the mapping is naturally high with almost all reads mapped to the genome. Unmapped reads were probably lost due to very low sequence complexity.

To calculate the mappability score, we needed to calculate the percentage of uniquely mapped reads for every 500 bp fragment as determined above. As proposed by Yaffe and Tanay, we used a mapping quality of at least 30 as an indicator for a uniquely mapped read. It-

erating over all fragments, we used samtools view [118] to extract all reads that overlap with the current fragment, and calculated the fraction of reads with a mapping quality > 30. Table 6 and Figure 21 show statistics and distribution of mappability scores for both species. The mappability of reads is generally very high, with a high fraction of fragments having a mappability of 1, indicating that they are covered only by uniquely mapped artificial reads.

The score is used to filter fragments that cannot be mapped to a single genome location due to lack of sequence uniqueness. Yaffe and Tanay suggest to use a mappability threshold of 0.5, which we applied to our data as well. Only a low number of reads (8.09% for H.sapiens, 12.91% for M.musculus) were discarded as invalid.

Table 6: Statistics on the mappability score distribution for human and mouse. Mappability score is calculated as the fraction of uniquely mapped reads per (trimmed) fragment. Validity is defined as a mappability score above 0.5

Species	Mean	Median	StDev	#Valid
H.sapiens	0.89	1	0.24	1,537,959 (91.91%)
M.musculus	0.86	1	0.29	1,434,054 (87.09%)

Calculating GC content for each fragment end

Besides mappability score and length, GC content is also required as input to hicpipe. To determine each fragment ends GC content, we used an in-house sequence extractor tool written by Jonathan Hoser (NGS group, Helmholtz Zentrum München) to parse the respective genome sequences of each fragment from the fasta files. Subsequently, we calculated GC content on the fly and added it to the input. For a distribution of fragment ends' GC content see Figure 21.

Normalization using hicpipe

Using default models for correction parameters, we ran hicpipe version 0.93 on our data. Hicpipe runs 6 steps:

Step 0 Dataset initialization

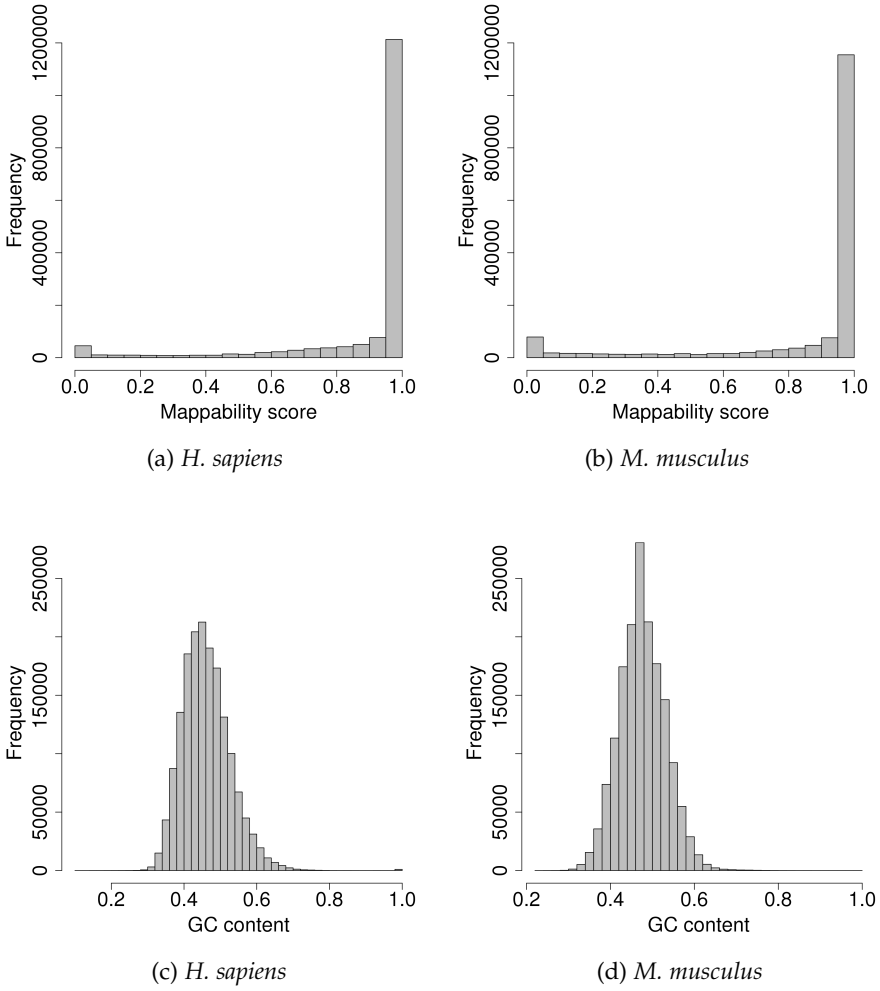
Step 1 Map paired reads to paired fends (=fragment ends)

Step 2 Prepare bias models

Step 3 Learn model

Step 4 Bin by coordinates

Step 5 Compute observed and expected matrices



(a) *H. sapiens* (b) *M. musculus*

(c) *H. sapiens* (d) *M. musculus*

Figure 21: Distribution of mappability scores and GC content of human and mouse fragment ends, illustrating a high amount of fragment reads covered by uniquely mapped reads (Mappability score 1). Reads with a score below 0.5 are discarded, but the distribution shows that only a minor amount is affected. GC content is normally distributed in both species.

The model generation includes the previously calculated mappability score, identification of non-specific ligation products and correction of length and GC-content biases. A non-specific cleavage product is defined as a paired read, i.e. the ligation product of two fragments close in space, for which the sum of the two distances to the next restriction enzyme cutting sites is larger than 500 bp. In such a case the ligated fragments are very long and depend on chromatin compaction, so they are discarded.

To correct the aforementioned biases, correction matrices are initialized. Fragment ends are binned according to length into 20 bins of equal size. The seed matrix is then defined as

$$S_{len}[i,j] = \left(\frac{1}{P_{prior}}\right) \cdot \frac{O_{len}[i,j]}{T_{len}[i,j]} \tag{2}$$

in which P_{prior} describes the prior probability to observe a pair, $O_{len}[i,j]$ is the number of observed pairs for which one fragment end is in bin B_j^{len}, and $T_{len}[i,j]$ is the number of unique pairs with one fragment end in B_i^{len} and the other in B_j^{len}.

The GC content seed matrix S_{gc} is computed accordingly, with binning according to GC-content of the 200 bp region from fragment end toward the fragment. A third matrix is calculated for mappability scores, with binning into steps of 0.1 starting at 0.5.

In step 3 the model is learned, calculating the probability $P(X_{a,b})$ to observe two fragment ends a,b in a paired-end read:

$$P(X_{a,b}) = P_{prior} \cdot F_{len}(a_{len}, b_{len}) \cdot F_{gc}(a_{gc}, b_{gc}) \cdot M(a) \cdot M(b) \tag{3}$$

In this formula $a_{len}, b_{len}, a_{gc}, b_{gc}$ are the length and GC-content bins of the two ends, $F_{len}(a_{len}, b_{len}), F_{gc}(a_{gc}, b_{gc})$ are real valued functions and the M function describes the mappability score. Both F matrices (symmetric matrices with $20 \cdot 20$ parameters) are estimated using maximum likelihood, based on the following likelihood function:

$$\begin{aligned} L(F_{len}, F_{gc}) &= \prod_{\{a,b\} \in I} P(X_{a,b}) \cdot \prod_{\{a,b\} \notin I} (1 - P(X_{a,b})) \\ &= \prod_{c=(a_{len}, a_{gc}, b_{len}, b_{gc})} P(X_{a,b})^{n_c} \cdot [1 - P(X_{a,b})]^{m_c} \end{aligned} \tag{4}$$

in which I is the set of observed fragment end pairs, n_c is the number of observed pairs that match the bin criteria of c, while m_c is the number of unobserved such pairs.

Unfortunately, hicpipe does not create output files containing the probabilities used to determine the expected interaction matrix. No detailed description on the generation of this matrix could be found, but the assumption that $expected = observed \cdot probability$ proved to be

incorrect. We thus were forced to recalculate the probabilities according to equation 3 with output information given by hicpipe.

While we are using only the probabilities determined as described above for network creation, we also require normalized contact matrices for certain analyses. We extracted genome-wide expected matrix counts from hicpipe results and combined the observed matrix O and expected matrix E into the normalized matrix N using the following formula [230]

$$N[i,j] = \frac{O[i,j]}{E[i,j] \cdot N[i] \cdot N[j]} \tag{5}$$

with

$$N[i] = \frac{O[i]}{E[i]}, \text{where} O[i] = \sum_j O[i,j], E[i] = \sum_j E[i,j] \tag{6}$$

Due to large signal strength differences between the main diagonal of genome-wide contact matrices, where most contacts occur, and the matrix peripheries of inter-chromosomal interactions, where relatively few interactions are formed, we applied a logarithmic transformation to the normalized contact matrix. This transformation helps improve the signal strength in inter-chromosomal regions.

6.1.2 *Filtering*

Fragment based filtering

Kruse et al. filtered the normalized Hi-C interaction matrix for *S.cerevisiae* on fragment level. According to Duan et al., filtering should be performed separately for intra- and inter-chromosomal interactions because of the polymer-like properties of chromosomes. This leads to a very strong inverse relationship between intra-chromosomal distance of two fragments and the frequency of their observed interaction, making the calculation fairly complex. Due to this, Kruse et al. focus on inter-chromosomal contact networks only, and since so far no detailed comparison of inter-chromosomal contacts in mammals has been conducted, we chose to do the same.

For the false-discovery rate filtering of interacting segments, the p-value for each inter-chromosomal fragment pair in the original publication is given by assuming binomially distributed fragment pair interactions:

$$P = \sum_{i=k}^{n} \binom{n}{i} m_{norm}^{i} (1 - m_{norm})^{n-i} \tag{7}$$

with m_{norm} being the sum of the normalized interaction probabilities for all four pairs of fragment ends of the two fragments (as calculated in the previous step), k being the observed number of reads for the fragments, and n being the total number of observed reads for inter-chromosomal interacting fragments.

To illustrate the relationship of this p-Value, probability and binomial distribution, we plotted examples for $n = 100$, $k = 10$ and varying probabilities to show the behaviour of the probability mass and the resulting p-Value in Figure 22. Probability values range from $p = 2.5e - 3$ to the counter probability $p = 1 - 2.5e - 3$. k is shown as a dotted black line. The p-Value for each of the given probabilities is calculated as the probability mass that lies to the right of k, shown in shaded areas. Only for low probabilities the majority of the probability mass lies to the left of k, leading to low p-Values. For all other ps the p-Value would be 1 or close to 1. This means that, given a total observed read count of 100, a very low background probability is required for a fragment pair with 10 observed reads to be significant. Mathematically, the p-value signifies the probability of observing at least k reads for a single event, given the background distribution.

To correct for multiple hypothesis errors, false discovery reduction (FDR) is performed to calculate q-values from the p-values, using the method by Benjamini and Hochberg [15] that is integrated into R as `p.adjust(p, method='fdr')`.

Adjustments for mammalian chromosomes

The fragment-based filtering described above is designed for and very well suited for organisms with small genomes and a high sequencing depth in the Hi-C experiments. We considered performing fragment-based filtering for human and mouse data and compared the data properties with the yeast Hi-C data used by Kruse et al. Yeast has a small genome size of 12.2 Mb, and the average read count for a fragment pair in the Hi-C data, which was generated by Duan et al. [51], is 8.22 even for inter-chromosomal contacts. Human and mouse genomes are much larger with 3.3 billion bp and 2.8 billion bp, respectively, and sequencing depth is not high enough to reach similar average read counts. For both genomes, the average number of reads observed for a pair of fragments (intra- and inter-chromosomal) is 2.46 and 2.66, respectively. Additionally, the mammalian genomes introduce complexity into the analysis through the large number of different chromosomes.

In order to receive valuable and interpretable data, we decided to perform filtering not on the level of fragment pairs, but on the level of 500 kb segments. Binning data to overcome low sequencing depth is common practice for the handling of Hi-C data from mammalian genomes. For details on our binning procedure see the next section.

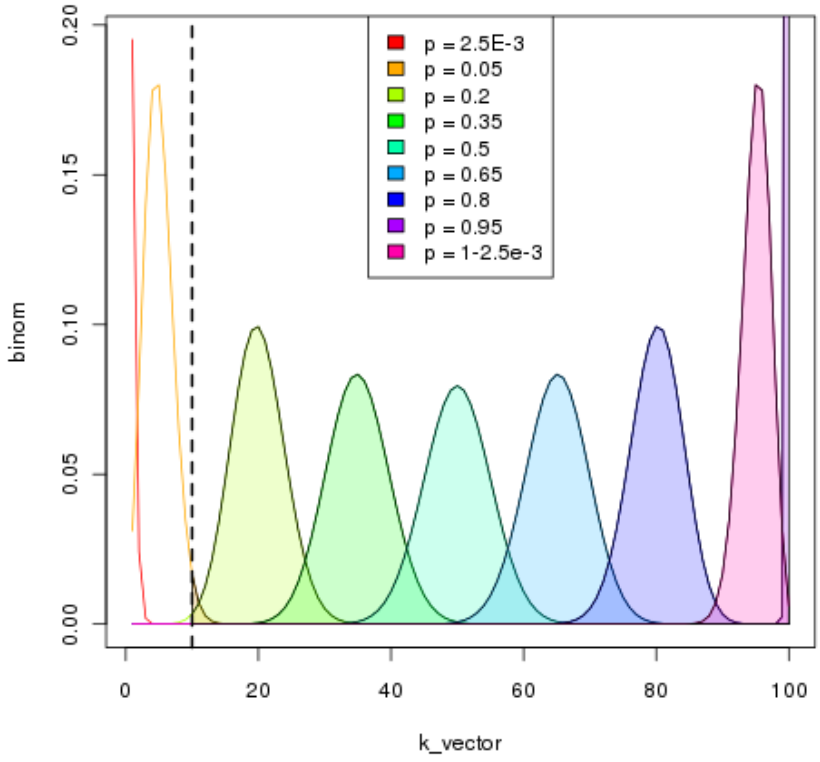


Figure 22: Illustration of binomial distribution density functions for $n = 100$ and varying probabilities. Dotted black line indicates location of $k = 10$. p-value for filtering is calculated as the (shaded) probability mass to the right of k.

While we lose resolution in this process, it is necessary to work with large-scale mammalian data.

In general, the filtering method is implemented in the same manner as above, using a Binomial distribution of segment pairs to calculate the p-value of each inter-chromosomal contact, based on interaction probability, number of observed reads for the segment pair and total number of observed interacting read pairs. However, to account for the complex genome structure of mammals, we calculated the number of observed reads n independently for each pair of chromosomes, resulting in the total number of observed interactions for the current chromosome pair. We summed fragment pairs observed together that fell into the respective 500 kb segments to receive the observed number of reads k for the segment pair. The interaction probability has to be calculated on the level of fragments and fragment ends, because correction factors cannot be sensibly determined for segments of fixed length.

To calculate the background interaction probability of a pair S_1, S_2 of inter-chromosomal 500 kb segments, we collected all fragment ends that lie within one of these two segments and their corresponding fragments. It is important to distinguish between the set of fragment ends and fragments, as not both ends of a given fragment necessarily lie within the same segment. For each pair of fragments $frag_i, frag_j$ for which at least one end lies in segment S_1 and segment S_2, respectively, the interaction probability is calculated as the sum of the interaction probability of all fragment ends (see equation 8).

$$P(frag_i, frag_j) = \sum_{fend_{i'} \in frag_i, fend_{j'} \in frag_j} P'(fend_{i'}, fend_{j'}) \tag{8}$$

The interaction probability for all possible combinations of fragment ends of $frag_i$ and $frag_j$, which are four at most, is calculated only for pairs of fragment ends that both lie within the segments S_1 and S_2, respectively.

$$P'(fend_{i'}, fend_{j'}) = \begin{cases} P(fend_{i'}, fend_{j'}) & \text{if} fend_{i'} \in S_1 \wedge fend_{j'} \in S_2 \\ 0 & \text{else} \end{cases} \tag{9}$$

We calculated the interaction probability of each 500 kb slice pair as the average interaction probability of all possible fragment interactions within. p- and q-values are then calculated as described above.

6.1.3 *Accounting for different chromosome lengths*

Our adaptation of Kruse's approach considers only pairs of chromosomes at a time, and the total number of contacts from each pair of chromosomes is the upper limit for the p-value calculation of contacts on this pair. This might lead to an overestimation of high confidence contacts for shorter chromosomes, which naturally have a lower total number of contacts. Hence we performed a second normalization step after q-value calculation to control for this chromosome length bias.

For each pair of chromosomes chr_a, chr_b, the normalization factor f_{chr_a,chr_b} was calculated as follows:

$$f_{chr_a,chr_b} = \left(\frac{max\_length\_product}{length_{chr_a} \cdot length_{chr_b}} \right) \tag{10}$$

where $length_{chr_a}$ is chr_a's length, and $max\_length\_product$ is the product of the longest and second longest chromosome lengths. Multiplying q-values with f_{chr_a,chr_b} leads to punishment of shorter chromosomes.

6.1.4 *Calculating spatial proximity values*

Besides inter-chromosomal interaction networks, we also converted Dixon et al.'s Hi-C matrices into *spatial proximity values* for further analyses. These values were first introduced by Lieberman-Aiden et al. [120] and represent the contact profile similarity of two DNA segments, in our case of 500 kb length. We modified their approach only slightly to account for the low signal-to-noise ratio in inter-chromosomal data.

First, the normalized interaction matrix N was calculated from the expected matrix E determined with hicpipe, the observed matrix O derived from Hi-C data. As described in section 6.1.1, a normalization formula (equation 5) was used to convert these two matrices into the normalized matrix, and subsequently a logarithmic transformation was applied. The spatial proximity value of two segments i and j is then calculated genome-wide as the Pearson correlation coefficient of rows i and j in matrix N, which are identical to columns i and j due to symmetry. Each column contains the genome-wide normalized contact profile of a given segment, so a high Pearson correlation coefficient of two segments indicates that they tend to interact with and avoid the same regions. The term "spatial proximity value" is thus misleading, as it does not actually capture the distance in space. However, as it has been used in literature before [104], we will use it in this work for the sake of consistency. We only computed inter-

chromosomal spatial proximity values and discarded those for intra-chromosomal segment pairs.

6.2 NETWORK CREATION AND ANALYSIS

Using the confidence q-value calculated for each pair of 500 kb segments, we can apply different thresholds to create binary contact/no-contact matrices. Such matrices can be easily transformed into segment interaction networks (SIN), where nodes are 500 kb segments and edges are introduced between segments that are in contact according to the matrix. In addition, we generated a physical gene interaction network (GIN) for each species. The set of protein-coding genes from ENSEMBL [56] were downloaded using Biomart [99], and liftOver [82] was applied to map their coordinates to the used assemblies hg18 and mm9. Each gene was then mapped to the 500 kb segment where its majority lies, and the GIN was initialized with the set of genes as nodes. For each interacting pair of segments S_i and S_j, we inserted edges between all pairs of genes mapped to S_i and S_j, respectively.

Randomization of segment and gene interaction networks

For validation of the results, we constructed randomized SINs and GINs for comparison. Randomization was performed on the SIN in two steps according to the suggestions of Kruse et al .[108]:

1. Initialization of random contact network
2. Permutation of edges in this network
3. Raise transitivity of the random network

The generation of a random contact network is performed according to Witten and Noble [222]. Let $|S|$ be the number of segments in the genome of question. We first generated random positions of these segments in the three-dimensional space by selecting $|S|$ random points in the three-dimensional space of a cube with side length 1. We assigned a segment to each point by drawing without replacement, and calculated the Euclidean distance (Equation 11) for each pair of points, if the assigned segment pair is inter-chromosomal.

$$dist_{a,b} = \sqrt{\sum_{i=1}^{3}(a_i - b_i)^2} \tag{11}$$

We then calculated the percentage c of confident contacts among all possible inter-chromosomal contacts observed in the human and

mouse SINs, and determined the c% shortest distances and their corresponding segments in the unicube. These pairs of segments are connected with an edge in the random contact network.

To ensure that the randomized network has similar basic properties as the original network, for better comparability, Kruse et al. suggest some additional steps which we also implemented. For one, in step 2 edges are permuted as a second randomization step. Pairs of edges are only rewired if they fit certain criteria: Given four network nodes u, v, s, t and two edges (u, v), (s, t) which were selected uniformly at random from the network, these two edges will be deleted and substituted by (u, t), (s, v) if

a) $u \neq t \wedge s \neq v$

b) (u, t) and (s, v) do not already exist in the network

c) u and t as well as s and v are from different chromosomes

The procedure was repeated $10 \cdot |E|$ times, with $|E|$ being the number of edges in the network. To ensure consistent clustering behaviour between randomized and original network, the transitivity might need to be raised. Transitivity describes the number of observed triangles, i.e. set of three nodes connected by three edges, compared to the number of possible triangles and is calculated as defined by Soffer and Vasquez [192]:

$$\tilde{T} = \frac{\sum_i \delta(i)}{\sum_i \omega(i)}$$

where i is a node in the network, $\delta(i)$ describes the observed number of triangles in the neighbourhood of i, i.e. the number of neighbour pairs of i which themselves share an edge, and $\omega(i)$ describes the maximum possible number of triangles in the neighbourhood of i.

After SIN randomization, we transformed both SINs into random gene interaction networks by mapping the same number of randomly drawn genes to each segment that it originally contained, and adding edges between all pairs of genes from connected segments.

6.2.1 *Basic network analysis*

Using Cytoscape [184], we calculated network statistics. Connected components were extracted using the Java Universal Network/Graph Framework (JUNG) [145]. A connected component is a subgroup of nodes and their edges where a path between each pair of nodes from the subgroup exists, but no paths to any other node in the network. We investigated the following network statistics:

DIAMETER The longest shortest path between any two nodes in the network.

AVERAGE DEGREE Average number of neighbours per node.

CLUSTERING COEFFICIENT [12, 218] The clustering coefficient describes the degree to which the nodes in a network tend to cluster together. It is calculated for each node as follows:

$$C_n = \frac{2e_n}{(k_n(k_n - 1))} \tag{12}$$

n A node in the network

e_n The number of connected pairs between all neighbours of n

k_n The number of n's neighbours.

The network's clustering coefficient is the average clustering coefficient of all nodes, where nodes with less than two neighbours have a $C_n = 0$.

CHARACTERISTIC PATH LENGTH The average length of the shortest paths between all pairs of nodes.

CONNECTIVITY CENTRALIZATION [46] The centralization serves as an index of degree distribution. It is calculated as

$$Centralization = \frac{N}{N-2}\left(\frac{max(k)}{N-1} - \frac{mean(k)}{n-1}\right) \approx \frac{max(k)}{N} - \frac{mean(k)}{n-1} \tag{13}$$

k Connectivity of the network, i.e. the set of degrees of all nodes

N Number of nodes

HETEROGENEITY [46] The heterogeneity is a measure for how different nodes are with respect to their degree and calculated as

$$Heterogeneity = \frac{\sqrt{variance(k)}}{mean(k)} \tag{14}$$

ISOLATED NODES Isolated nodes are those with a degree of 0.

SPATIAL CLUSTERS We defined this term to describe sets of genes that are co-localized in the three-dimensional space of the nucleus. A spatial cluster is formed by all genes from two interacting 500 kb segments.

Analysis of network conservation

We investigated the conservation of GINs in human and mouse based on synteny blocks or regions of conserved gene order. We hypothesize that large genomic rearrangements disrupt the three-dimensional structure, making it not reasonable to explore conservation of individual gene contacts genome-wide. Instead, we focused on blocks of conserved gene order between human and mouse as detected by SyntenyMapper (see part iv). Using a confidence cutoff of 0.05 to capture as many similarities as possible, we extracted the subgraphs for each such synteny block in both human and mouse. Our goal was to find out if genes that lie in conserved genomic regions are in contact with equivalent genes in both species. Thus, we mapped all genes that are in contact with genes from the synteny block in human (creating set H) to their mouse orthologs (H_M), using the syntenic one-to-one ortholog mapping from SyntenyMapper (see Figure 23 for an illustration). We then calculated the overlap between these orthologs and the mouse genes with which genes from the synteny block are in contact (set M). The number was normalized by the smaller number of contact genes from the two species.

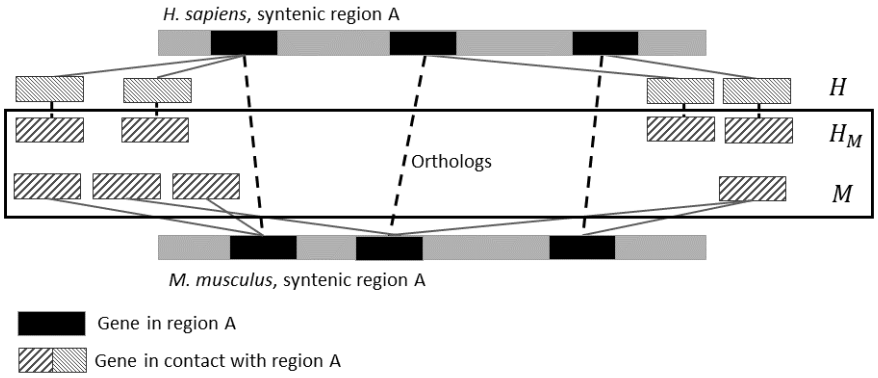


Figure 23: Illustration of how contacts where mapped between human and mouse to determine the degree of conservation. For each region of conserved gene order **A**, all genes that are in contact with genes from **A** in human are extracted and make up the set H. The same is done in mouse, creating the set M. To calculate the overlap between both sets, genes in human are mapped to their mouse orthologs H_M using SyntenyMapper.

For random background we simply shuffled the associations between synteny regions in human and mouse (drawing without replacement). This way, the association between genes and regions is kept intact and no randomization bias is introduced.

Feature enrichment in trans-interacting segments

Using our large database of linear genomic features (see part ii), we calculated average feature overlaps for each autosomal 500 kb segment of the human and mouse genome for the properties listed in Tables 7 and 8. Additional histone modifications were downloaded from the ENCODE project [8] for a better match of cell types.

Table 7: Description of features used for enrichment analysis of trans-interacting segments, *H. sapiens*. Histone modifications were downloaded from ENCODE [8] to match the cell type of stem cells. All histone modifications are active marks.

Feature	Cell type	Source
		H. sapiens
H3k4me3 peaks	hESC	ENCODE [8], Broad Institute
H3k4me1 peaks	hESC	ENCODE [8], Broad Institute
H3k27ac peaks	GM12878	ENCODE [8], Broad Institute
H3k9ac peaks	hESC	ENCODE [8], Broad Institute
LADs	Fibroblasts	NKI, Peric-Hupkes et al. [159]
DNaseI sites	Collection of cell types	ENCODE, Sabo et al. (2004, 2006) [176, 177]
LINE repeats	-	RepeatMasker [191]
LTR repeats	-	RepeatMasker [191]
Nucleosome occupancy	-	ENCODE, UW, Gupta et al. (2008) [70]
Open chromatin	GM12878	ENCODE, Duke/ UNC/ UT-Austin/ EBI [21, 8]
RTD	Lymphoblasts	ReplicationDomain DB [219]
SINE repeats	-	RepeatMasker [191]
SNPs	-	dbSNP [185]

For each segment and feature, we identified overlapping elements and calculated the overlap or, in the case of replication timing domain, the average scores over the length of 500 kb. The feature overlaps of trans-interacting segments were then compared to not interacting segments to test for enrichment and depletion of genomic properties.

6.2.2 *Overlap between trans-interacting segments and transcription factor binding sites*

Given the assumption that co-localization of genes may correlate with their transcription, we analysed if genes in a spatial cluster defined as above showed preferential transcription factor binding sites for certain factors in human. In human, large amounts of high quality data are available from the ENCODE project [8]. We did not perform this

Table 8: Description of features used for enrichment analysis of trans-interacting segments, *M. musculus*.

Feature	Cell type	Source
		M. musculus
H3k4me3 peaks	ES-E14	ENCODE [8]/LICR, Ren et al.
H3k4me1 peaks	ES-E14	ENCODE [8]/LICR, Ren et al.
H3k27ac peaks	ES-E14	ENCODE [8]/LICR, Ren et al.
H3k9ac peaks	ES-E14	ENCODE [8]/LICR, Ren et al.
LADs	mESC	NKI, Peric-Hupkes et al. [159]
DNaseI sites	ES-E14	ENCODE/University of Washington [176, 177]
LINE repeats	-	RepeatMasker [191]
LTR repeats	-	RepeatMasker [191]
Open chromatin	-	ENCODE/Duke/
UNC/UT [21, 8]		
RTD	mESC	ReplicationDomain DB [219]
SINE repeats	-	RepeatMasker [191]
SNPs	-	dbSNP [185]

analysis in mouse due to lack of comparable data. We downloaded a hESC dataset comprising binding site peaks for 55 transcription factors from the ENCODE project (hg19, available at UCSC [102]). For the full list including individual sources, see Supplementary Table S2.

For each spatial cluster, we calculated the percentage of genes that overlap with at least one transcription factor binding site of a given transcription factor. Additionally, the same feature enrichment analysis as described in the previous section was performed to test whether certain transcription factor binding sites are abundant in trans-interacting segments.

6.2.3 *Functional analysis of genes in spatial clusters*

Genes from different chromosomes that come together in a spatial cluster might do so just because of random effects like Brownian motion. However, it is possible that close proximity of genes is, at least partially, functional. We aimed to determine functional similarities of genes within spatial clusters, using the GO functional annotations.

Random interactions can obscure Hi-C based data to the point were signals cannot be found easily. Thus, we performed an analysis suggested by Khrameeva et al. [104], that uses a variance-reducing approach to uncover associations in the data that are obscured by noise. According to the authors, binning of the data according to spatial proximity was necessary to reduce the effect of noise. We evaluate

this procedure in section 6.2.5 and discuss our own modified binning method. After noise reduction, we calculated Pearson correlation coefficients of spatial proximity values (see section 6.1.4) and GO term similarity as determined by the Bioconductor [60] package GOSemSim [231]. GO term similarity was calculated separately for each GO hierarchy ("biological process", "molecular function", "cellular component") and results were combined as average similarity score for each segment pair, ignoring hierarchies for which no GO term was available.

6.2.4 *Comparison with a co-expression network in human*

As described before, large amounts of noise can occlude signals in the data. Khrameeva et al. [104] have shown a correlation between spatial proximity values and co-expression in human fibroblasts. Using a similar method to reduce the effect of noise (see section 6.2.5), we assessed the association strength with the Pearson correlation coefficient and a randomization procedure.

To see if there are any correlations between co-expression and co-localization of genes, we first had to establish co-expression of genes in stem cells. Due to lack of comparably complete data on mouse stem cells, we performed this analysis for human networks only.

We contacted the authors of 'Genome wide profiling of human embryonic stem cells', Liu et al.[123], and they kindly sent us their data on expression of over 20,000 genes in 43 hESC samples. Khrameeva et al. use expression data from a database termed CoexpresDB [144], which contains a large amount of data for somatic cells. We used their definition of a co-expression measure on the data from Liu et al. for better comparability to our data. For each pair of genes from two different but interacting segments, the co-expression measure [104] is calculated as follows:

$$CM(i,j) = \sum_{k=i}^{n} \left(\frac{W_{ki}}{N_i} + \frac{W_{kj}}{N_j} \right) \cdot R_k \tag{15}$$

where

i, j are two genes, with i lying in one segment and j lying in the other

W_{ki}, W_{kj} are the portion of the harbouring 500 kb segments that overlap with these genes

N_i, N_j Number of genes in the corresponding segments

R_k Pearson correlation coefficient of the two genes' expression profiles

6.2.5 *Modified noise reduction procedure*

Khrameeva et al. [104] applied an equal-distance binning method to the spatial proximity value data to reduce the influence of noise in the Hi-C dataset. In this approach, data are binned into intervals with equal absolute length (i.e. $end - start$) according to their spatial proximity value. Correlation is then assessed over the median x (GO term similarity or co-expression measure) and y (spatial proximity value) values of these intervals.

However, consulting several papers and statistics specialists, we see some problems with this approach. Binning data, especially large datasets such as the ones present here, masks variance. In fact, there is absolutely no correlation observable in the raw data (correlation coefficients close to 0), so that binning can only be viewed as a trick to overestimate results and amplify very weak trends in the data. Khrameeva et al. have confirmed that the observed Pearson correlation coefficients in their data often increased from values close to zero to values above 0.9 after binning.

This effect is also termed *correlation inflation* and caused by a statistically not valid binning approach [101]. In short, if the hypothesis is true and there is a correlation in the data, you would not need to resort to binning. In an extreme example, binning of the data into two sets would always result in a perfect Pearson correlation coefficient of 1.0 or -1.0. Of course, such a result does not contain any true information. Being confronted with binned data should always make the reader sceptical. In fact, it has been shown that through use of different interval sizes and equi-distant binning, any correlation ranging from negative to positive can be shown for artificial datasets, and better for larger sets than for smaller ones [212]. A better option that does not distort the data distribution as much is equal-size binning:

1. Ranking of N tuples (x,y) according to x
2. Introduction of 30 bins, each containing $N/30$ entries. Tuples were distributed equally into the bins according to their rank
3. Calculate $mean(x)$ and $mean(y)$ as representations of each bin

While this approach is also problematic from a statistics viewpoint since it strongly decreases the variance in the data, we know that Hi-C data are in fact very noisy because of the averaging over millions of cells and the existence of random contacts due to Brownian motion. We decided to use the correlation assessment on basis of equal-sized binned data, but complement it with a randomization to assess the validity of our results.

- Randomize association of (x, y) 1000 times
- Perform binning, calculate Pearson correlation coefficient

- Fit normal distribution with $n = length(data)$, $m = \mu$ and $sd = \sigma$
- Assess significance of observed value as $1 - CDF(\rho)$, with CDF being the cumulative distribution function and ρ being the observed Pearson correlation coefficient

We are aware that the observed Pearson correlation coefficients on basis of binned data are highly overestimated and do not measure the true strength of the correlation. For this reason, we combine them with the significance value determined through the above described randomization approach. While the given correlation coefficients are overestimated, the given p-value confirms if there is indeed a significant association between x and y present in the data.

Additionally, instead of using medians we chose to use the average x and y values for each bin. Means are more sensitive to outliers, but due to our higher resolution of 500 kb compared to 1 Mb, a large amount of segments without genes leads to a high amount of zeros in the data, strongly influencing the median. These zeros are caused by gene-less segments which automatically have a GO term similarity or Co-expression measure of zero with any other segment. We investigated subsets of the data with exclusion of zeros and found that results were similar, leading to our decision to perform our analyses on the complete inter-chromosomal set. We chose 30 as bin number to make our results comparable to Khrameeva et al. [104].

6.3 PREDICTION OF INTER-CHROMOSOMAL CONTACTS

We investigate the potential predictive power of linear genomic features for prediction of inter-chromosomal contacts in this section. Using properties known to influence the activity state of a genome region, we hope to successfully train a classifier to distinguish between inter-chromosomal contacts and non-contacts.

6.3.1 *Data Preparation*

We calculated overlap percentage and, if the feature has a score or signal weight, the average score per base pair for all 500 kb segments in human and the following features: DNaseI hypersensitivity (score), LADs (score), LINE, LTR, Nucleosome occupancy, open chromatin, RTDs (score), SINE, SNPs, peaks for histone modifications H3k4me1, H3k4me2, H3k4me3, H3k9ac, H3k27me3, H3k36me3, and H3k27ac, gene density and chromosome. We combined the features of two segments and assigned this pair the class 'contact' if the q-value of their interaction was $\leq 1E-3$, in concordance with the network analysis. All other pairs were assigned the class 'no contact'.

With this threshold we are dealing with a vastly imbalanced set, where for each positive instance there are almost 7,000 negative instances. This means that any classifier that labels all instances as 'negative' during training will test well, with accuracies at 99% and very low Root Mean Square Error (RMSE). However, as the positive class is the one of interest, this result is not desired. In order to be able to successfully train a classifier to distinguish the properties of non-contacting from those of contacting segments instead of just assessing their quantity imbalance, we have to remove this bias.

There are multiple ways to deal with such imbalances [165, 92]. For one, not every classification method is equally sensitive to the problem; Support Vector Machine (SVM) are largely unaffected [92]. The efficiency of the applied approach to deal with class imbalance is also dependent on size of the dataset and degree to which the classes are imbalanced. In general, there are five methods [92]:

1. **Random Oversampling**, where class instances from the minority class are copied until the sizes are equal
2. **Focused Oversampling**, where instances with values close to the class boundaries are copied at random from the minority class until the sizes are equal
3. **Random Undersampling**, where n instances from the majority class are drawn randomly, with n being the size of the minority class, to replace the original majority class set
4. **Focused Undersampling**, where n random instances close to the class boundaries are drawn from the majority class, with n being the size of the minority class
5. **Cost-modifying**, where the datasets are not modified, but the misclassification cost for the minority class is increased to match the proportion in the data

We have implemented both random undersampling and used a cost-modifying meta classifier. In artificial tests, random undersampling has been shown to be the least effective correction method [92]. However, this does not necessarily apply to our case, since for us the minority class is the class of interest, with the majority class simply representing 'everything else' and possibly containing a large amount of irrelevant data. As stated by the authors, both cases can lead to a higher efficiency of undersampling. To avoid accidentally introducing a bias in the strongly undersampled negative set, we repeated the procedure $1,000$ times.

For the cost-modifying method we calculated the factor $f = \frac{m_1}{m_2}$, with m_1 being the size of the majority class and m_2 being the size of

the minority class. We created the following cost matrix to use for cost sensitive classification:

$$\begin{pmatrix} TP & FN \\ FP & TN \end{pmatrix} = \begin{pmatrix} 0 & f \\ 1.0 & 0 \end{pmatrix} \tag{16}$$

We performed classification on human and mouse data separately as well as combined. For mouse, fewer features are available, so we had to treat them as missing values for features absent for this species in the combined set (nucleosome occupancy, DNase I hypersensitivity, H3k4me2 and H3k27me3). We also used distinct chromosome identifiers for each species.

6.3.2 *Classification*

For classification itself we used WEKA [74], a platform which can easily be integrated into a Java project and provides a large amount of classification methods. We tried several different classifiers, among them Naïve Bayes and Logistic Regression as baseline models, as well as Neural Networks, Decision Trees and Random Forest. We do not present results on all these models, as early tests showed best results can be achieved with decision trees and Random Forest, so we focus on Random Forest in the following sections.

6.3.2.1 *Random Forest [22]*

Random forest is based on randomly created decision trees of a fixed size and uses the following steps to create a model for classification:

1. A fixed number of trees is started (in our case 10, WEKA's default value)
2. In each tree, at each node, choose a fixed number of random features from the input feature space, that is considerably lower than the input space (in our case 6 out of 37)
3. Perform a split according to these 6 random features
4. Each tree is built up recursively. If a tree is complete, no more nodes are added and leaves are class predictions.

This is done for the training set (see Evaluation below for details on the training and test set). Each entry from the test set is then run through all of the trees, which give a class vote according to the leaf where the query ended up. The class with the highest numbers of votes is taken as predicted class. Random forest is considered to be very robust and not sensible to overfitting [22].

For the cost-modifying procedure described above we used the meta classifier *CostSensitiveClassifier* from Weka on Random Forest.

6.3.2.2 *Feature selection*

To distinguish features with high predictive power from others, we used WEKA's standard feature selection method *CfsSubsetEval* [75] with *BestFirst* and default values. The method itself aims to select a subset of the present attributes that are highly correlated with the class and have low correlation between them, leading to a low redundancy. *BestFirst* is used for the actual feature selection, in our case performing a forward selection by greedy hillclimbing and backtracking.

Hill climbing *is a heuristic algorithm to identify (local) maxima*

We applied feature selection to the imbalanced sets and to 1000 randomly undersampled balanced sets, extracting the cut set of the selected features.

6.3.2.3 *Evaluation*

For classification with cost-modifying procedure, we evaluated the trained classifiers with a 10-fold cross-validation. In order to ensure that our balancing method did not affect the prediction success, we decided to test the accuracy of classifiers trained on the undersampled dataset with a previously determined holdout set containing 10% of the data. This holdout set is not manipulated and contains positive and negative instances in the same, imbalanced proportions as the complete dataset. For the species comprehensive set we combined holdout sets of both organisms. All instances used in the holdout set were naturally removed from the training set before undersampling procedures.

We used different measures to assess the prediction accuracy. Accuracy itself was not used, as training on an imbalanced set can lead to high accuracies even if all instances are labelled with the same class. Instead, we focused on the measures described below, where TP stands for the number of true positives, TN the number of true negative predictions, and FP and FN for number of false positives and negatives, respectively. Positives are defined with respect to the target class, in our case the contact class.

PRECISION describes the percentage of instances predicted to be in a certain class that are correctly predicted, also termed positive predictive value (equation 17). As such, there is a precision value for each class, which can be combined using a weighted average that accounts for the number of instances from both classes.

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

RECALL describes the percentage of positive instances that were predicted to be positive, and is also called sensitivity (see equation 18). Like for precision, a weighted average can be calculated for both classes.

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

AREA UNDER PRECISION-RECALL CURVE (AUPRC) is a measure well suited for the evaluation of imbalanced test sets, since it can detect performance difficulties better than AUROC (see below) [20]. The underlying curve is obtained by using different thresholds for the classifier in use, and plotting precision and recall for each such run. Ideally the area under this curve is close to 1, implying a high precision and recall for most thresholds.

AREA UNDER RECEIVER OPERATING CURVE (ROC) or AUROC is calculated similarly to the AUPRC. The underlying curve is calculated for different classifier thresholds, and the points are derived from the recall or true positive rate and false positive rate ($= \frac{FP}{FP+TN}$).

7

RESULTS AND DISCUSSION

7.1 HI-C DATA FROM HUMAN AND MOUSE ESCS

In this thesis we are working with traditional Hi-C data on human and mouse ESCs from Dixon et al [45]. As discussed in the introduction (part i), this experimental method was the first high-throughput approach developed for chromatin conformation capture, and suffers from a low signal-to-noise ratio compared to more recent methods such as Tethered Conformation Capture (TCC) or single cell Hi-C. However, we aim to conduct a holistic comparison of three-dimensional structure in *H. sapiens* and *M. musculus* to analyse the degree of conservation. For this reason and because only little data from TCC and single cell Hi-C experiments is currently available, we decided to use Hi-C data that was derived from both these organisms in comparable cell types in the same experiment.

7.2 NORMALIZATION AND FILTERING

When dealing with high-throughput data, preparation and pre-processing is often equally as important as statistical analysis itself. Signal needs to be separated from noise, which is always present in huge data sets created by imperfect experimental procedures. In the case of Hi-C, some biases related to the experiment setup are known and can be controlled. Among these are unspecific ligation products, fragment length, GC content and read uniqueness biases. We used *hicpipe 0.93*, a method published by Yaffe and Tanay [230], to normalize the raw Hi-C data.

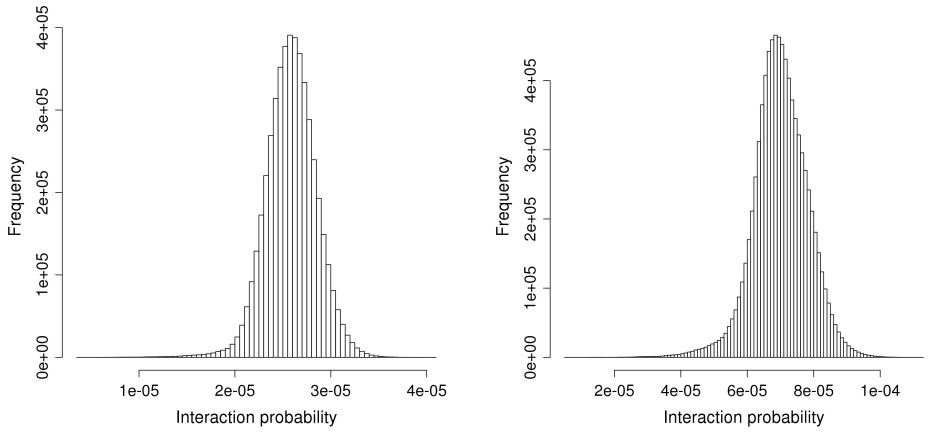
Table 9 summarizes the number of read pairs available in the raw Hi-C data for intra- (within) and inter- (between) chromosomal interactions. Our analysis focuses on inter-chromosomal contacts, which are rare compared to contacts within one chromosome. Previous research has shown that the probability of an interaction decreases linearly with proximity [235], so it is not surprising that the majority of Hi-C reads cover close-range interactions. In the embryonic stem cell data from Dixon et al, 83% (Human) and 89% (Mouse) of paired end reads are formed by fragments from the same chromosome, respectively (Table 9). In the following sections, only inter-chromosomal interactions are considered.

We calculated interaction probability for each pair of 500 kb segments as described in methods (section 6.1.1). Figure 24 shows that this contact probability follows a normal distribution for both human

Table 9: Number of reads covering intra- (within) and inter-chromosomal (between) interactions.

	Intra	Inter
H. sapiens	100,263,614 (83.26%)	20,159,872 (16.74%)
M. musculus	483,760,138 (89.35%)	57,652,738 (10.65%)

and mouse embryonic stem cells. A high value for two loci can be interpreted as a high background probability for these two loci to be reported as spatially close by Hi-C due to their sequence properties, independent of actual proximity. We are thus looking to find pairs of loci with low interaction probability and high observed read counts.



(a) *H. sapiens* (b) *M. musculus*

Figure 24: Distribution of interaction probabilities calculated based on normalization by hicpipe [230] follows a normal distribution for both human and mouse.

To identify these segments, a p-value based on the binomial distribution was calculated to estimate if the read counts observed for a given segment pair can be explained by their background probability (see section 6.1.2). A low p-value therefore indicates a reliable contact. Even though the same method was applied to both species' data and cell types are comparable, the distributions differ drastically (Figure 25). While in *M. musculus* the majority of all segment pairs has a p-value of 1.0, there is far more variance in the p-value distribution of *H. sapiens*, with a high proportion of lower p-values. This implies a higher level of noise in mouse than in human. In fact, the raw Hi-C data comprise around 4.5 times more read pairs for mouse than for human (541,412,877 vs. 120,423,487). If we hypothesize a similar or due to smaller genome size slightly smaller degree of inter-chromosomal connectivity for mouse, this higher read coverage

is caused by a higher amount of noise in the data, which explains the skewed p-value distribution.

The large difference in read numbers also registers in the average read coverage, which is twice as high in mouse than in human (Table 10). Contact probability is in the same range, as seen before in Figure 24. This background probability is independent of the Hi-C data itself, but merely influenced by distance between restriction enzyme binding sites and other genomic features, so noise in the data does not disturb the interaction probability distribution in mouse. P-value and q-value, however, are calculated by combining the background probability with the raw read counts, and the aforementioned large amount of noise in mouse lead to extremely different p-value averages for human (0.41) and mouse (0.97).

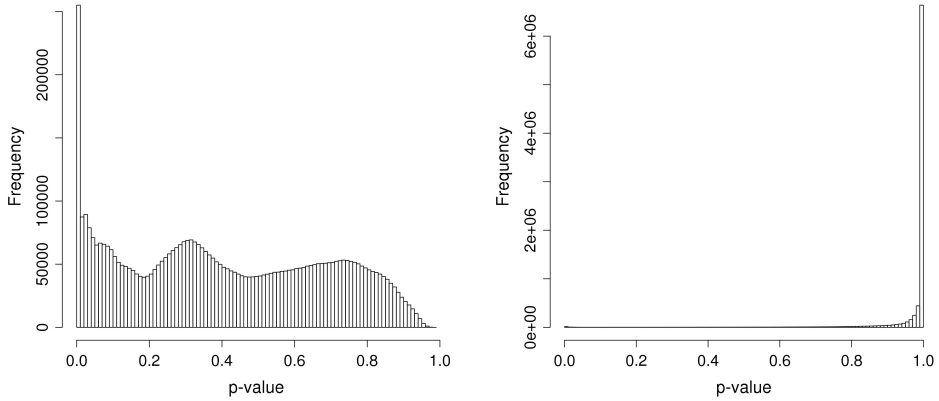
Read coverage in Dixon et al.'s data is twice as high in mouse than in human

Table 10: Statistics on parameters necessary for interaction confidence assessment, inter-chromosomal contacts only. Read coverage and contact probability are the basis for contact p-value calculation. q-value is calculated from p-value distribution for false discovery rate estimation.
SD: standard deviation

Parameter	*H. sapiens*			*M. musculus*		
	Mean	Median	SD	Mean	Median	SD
Read coverage	2.53	2	1.12	4.25	4	3.04
Contact probability	2.6e-5	2.6e-5	2.6e-6	7.0e-5	7.0e-5	8.2e-6
p-value	0.4071	0.3781	0.2741	0.9682	0.9998	0.1124
q-value	0.7052	0.7562	0.2188	0.9989	1	0.0330

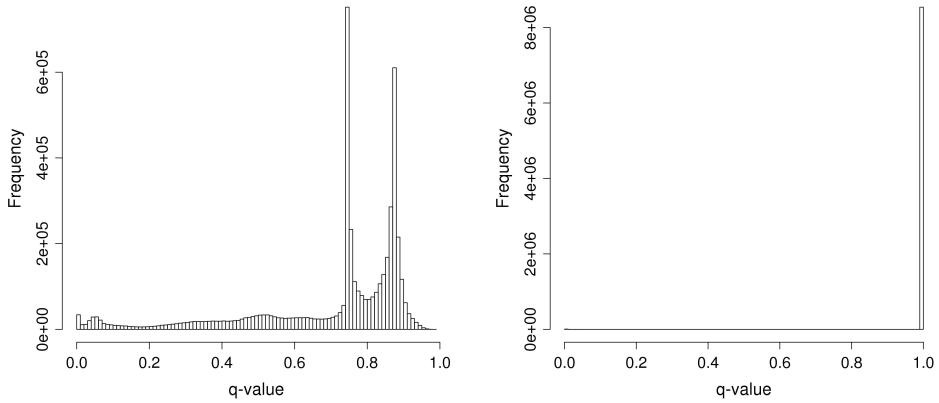
As a consequence, distribution of confidence (q-)values after multiple testing correction also differs between the species (Figure 26). Again, the high amount of noise in mouse leads to a peak at 1.0. In human, two peaks around 0.8 appear, also implying a high percentage of biased reads. During multiple testing correction, p-values are raised to account for the effect that when testing more than one hypothesis, one might be significant by chance. This leads to the majority of locus pairs in mouse having a q-value of 1, and an increased mean of 0.71 compared to p-value distribution in human. However, as it is known that Hi-C experiments produce very noisy data, it can be expected that the filtering procedure leads to a significant loss of false positive data.

We have adapted the p-value based filtering approach to account for large and complex mammalian chromosomes, thereby calculating values separately per chromosome pair. This introduces a length bias, as pairs of short chromosomes with a smaller sum of total observed reads will have increased p-values compared to long chromosomes with larger numbers of reads. To correct for this, we nor-



(a) *H. sapiens* (b) *M. musculus*

Figure 25: Distribution of interaction p-values calculated based on Hi-C bias interaction probabilities and observed read counts for each 500 Kb segment pair. In mouse, large amount of noise leads to a distribution that is heavily skewed towards 1.0, while human segment pairs mostly have lower p-values.



(a) *H. sapiens* (b) *M. musculus*

Figure 26: Distribution of interaction q-values after multiple testing correction of p-values. Again, most mouse segment pairs have a q-value of 1.0 due to noise, while in human there are two peaks around 0.8. Only few contacts have a low q-value and are thus highly confident.

malized q-values by dividing through the maximum combined chromosome length, which is product of the longest and second longest chromosomes' length. Figure 27 shows the number of inter-chromosomal contacts per chromosome before and after normalization for an exemplary q-value cutoff. In human, normalization results are as expected and comprise a stronger decrease in contact numbers for short chromosomes than longer ones. Still, shorter chromosomes such as 21 build more interactions than the very long chromosomes, e.g. 1 or 2.

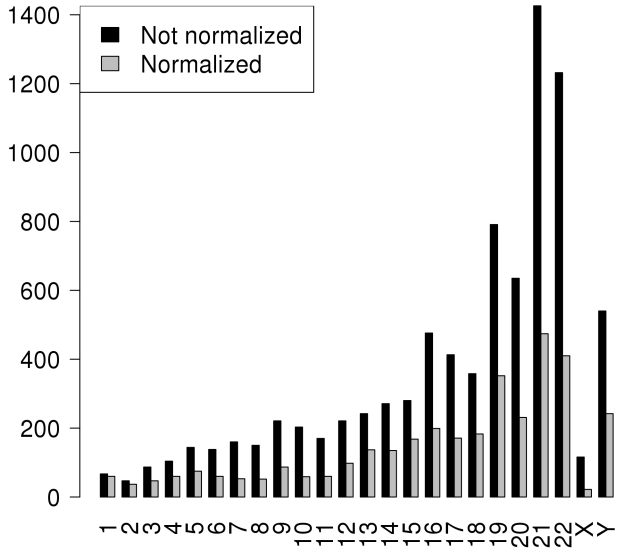
In mouse, the distribution is more uniform, with the exception of outlier chromosomes 11 and Y, which form the majority of contacts. Due to this special genome structure, normalization only slightly reduces contact numbers on each chromosome. Even the very short chromosome Y loses only a small proportion of its contacts, implying that many of these are formed either with strong confidence or to longer chromosomes, as the normalization raises the q-value relative to the length of both involved chromosomes. We will describe the structural properties of these interactomes in detail in the next sections.

7.3 CREATION OF SEGMENT INTERACTION NETWORKS

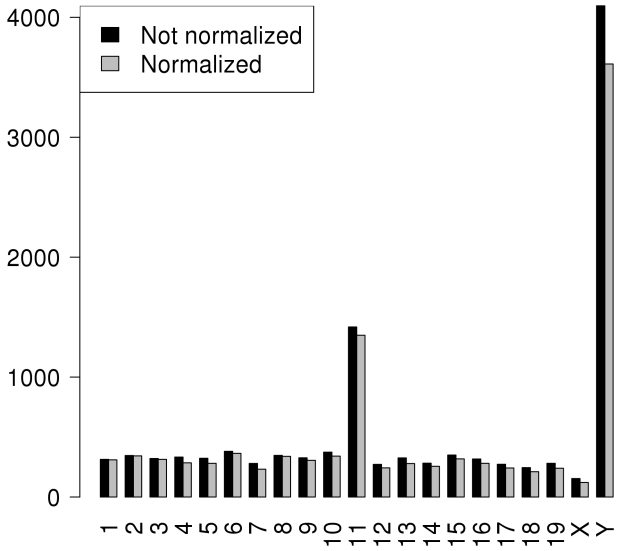
SIN: *(inter-chromosomal) Segment Interaction Network*

In order to convert a matrix of confidence values into a network, one has to choose a q-value threshold and transform them into binary values. It is essential that this threshold is not chosen arbitrary, as a too strict cutoff leads to loss of information, while false positives may be present with a too lax cutoff. A threshold of 0.05 is common practice, so we applied cutoffs between 0.05 and $1e-8$ to investigate network properties before deciding. In the resulting segment interaction networks (SIN) each 500 kb locus is a node and each interaction is an edge. Table 11 summarizes the sizes of the resulting networks. With the less stringent cutoff of 0.05, the majority of all 500 kb segments participate in at least one contact (human: 91.7%, mouse: 86.2%). However, for human this changes rapidly with decreasing q-value threshold. Even at a considerably high threshold of 0.001, less than half the segments form contacts to others. In mouse, we cannot observe this strong correlation. While the number of connected nodes naturally decreases with decreasing cutoff in this species as well, it does so much more slowly. So the mouse SIN (MSIN) still contains almost 80% connected nodes at the above mentioned cutoff of 0.001.

A similar effect appears to influence the degree of connectivity at different q-value thresholds. While in human the number of edges decreases strongly with decreasing cutoff, from 31,401 at 0.05 to 238 at $1e-8$, the difference in connectivity between the least and most stringent cutoffs in mouse is less pronounced (6,731 to 3,953). However, though the percentage of connected nodes is similar in both species at cutoff 0.05, the human SIN (HSIN) has 4.7 times more edges than



(a) *H. sapiens*



(b) *M. musculus*

Figure 27: Number of inter-chromosomal contacts per chromosome before and after chromosome length normalization for an exemplary q-value cutoff of $1e-4$.

the MSIN at the same threshold. These two characteristics imply a higher number of considerably low-confidence interactions in human compared to mouse.

Connected component: A subgraph where each node is connected only to all other nodes in the subgraph by a series of edges

Table 11: Size of segment interaction networks at different q-value cutoffs in human and mouse. 1^{st} component is the first and largest connected component. In both species, clustering is strong and almost all segments are either part of the first component or not connected at all.

Cutoff	#Nodes	#Nodes, connected	#Edges	1^{st} component	
				#Nodes	#Edges
H. sapiens					
0.05	5,732	5,254 (91.66%)	31,401	5,250	31,399 (99.99%)
1E-2		4,128 (72.02%)	13,674	4,077	13,647 (99.81%)
1E-3		2,500 (43.62%)	4,520	2,349	4,435 (98.12%)
1E-4		1,342 (23.41%)	1,736	1,126	1,611 (92.80%)
1E-5		858 (14.97%)	989	692	889 (89.89%)
1E-6		483 (8.43%)	500	294	378 (78.22%)
1E-8		233 (4.07%)	238	90	142 (49.66%)
M. musculus					
0.05	5,093	4,389 (86.18%)	6,731	4,383	6,729 (99.98%)
1E-2		4,363 (85.67%)	6,483	4,357	6,481 (99.97%)
1E-3		4,011 (78.76%)	5,589	4,003	5,586 (99.95%)
1E-4		3,820 (75.01%)	5,133	3,807	5,127 (99.88%)
1E-5		3,772 (74.06%)	4,978		
1E-6		3,452 (67.78%)	4,483	3,428	4,470 (99.71%)
1E-8		3,133 (61.52%)	3,953	3,105	3,938 (99.65%)

The size of the first connected component listed in the last two columns tells us if the network consists of multiple independent sets of connected segments or if the network is strongly clustered and the vast majority of connected loci are part of the same subgraph. In mouse, the latter is the case for all cutoffs, and even at threshold $1e-8$ 99.7% of nodes are part of one large connected component. In

human, this is true only for cutoffs higher than $1e-6$. For $1e-8$, the network starts to decompose into one large inter-connected subgraph which contains around 50% of connected nodes, and multiple small ones with up to 13 nodes. In general, however, both species' networks are highly inter-connected, without significant decomposition into subgraphs.

Our work focuses not only on characteristics of each species' interactome, but also on similarities and differences between them. Both human and mouse are mammals, and Dixon et al. [45] have shown that the intra-chromosomal three-dimensional structure is conserved between them. When deciding on a q-value cutoff to create the SINs for in-depth analyses, we opted for maximized comparability and chose cutoffs at which the connectivity is similar in both organisms. To achieve this, we had to pick different q-value thresholds for human and mouse, namely $1e-3$ and $1e-6$. At this cutoff, both species' SINs contain around 4,500 edges. This way we are able to identify common structural properties of the networks, while the structural differences that influence the network sizes at the different cutoffs are still pronounced, as will be discussed in the following sections. Additionally, we repeated most of the presented analyses on multiple cutoffs to validate that the choice of threshold does not bias the results.

The following sections will describe the main structural features of both the MSIN and HSIN, their differences and similarities. Our goal is to perform a holistic analysis of the inter-chromosomal interactome in mammals. Relationships between inter-chromosomal interactions and well-known genomic features such as repeats or replication timing domain will be discussed, as will correlation between spatial proximity and functional similarity or co-expression of genes.

7.4 THE MOUSE INTER-CHROMOSOMAL CONTACT NETWORK IS STRONGLY SHAPED BY HIGHLY CONNECTED SEGMENTS

A single 500 kb segment on chromosome Y forms the majority of contacts in mouse

The mouse genome contains at confidence value cutoff $1e-6$ approximately the same amount of inter-chromosomal contacts as the HSIN at cutoff $1e-3$. In general, there are more high confidence interactions in mouse, leading to a highly connected network. The Circos [109] plot of contacts allows us to identify another highly influential characteristic of the MSIN (Figure 28): The majority of its connections are formed between either a certain locus on chromosome Y or 11 and other regions distributed all over the genome. Specifically, a 500 kb segment close to the telomere of chromosome 11 (3,000,000 to 3,500,000) forms a high number of inter-chromosomal contacts, mainly to other telomere-proximal regions. Another such segment on chromosome Y (2,500,000 to 3,000,000), which otherwise consists mainly of repeat sequences to which mapping is not possible [2], is in contact with almost the entire genome.

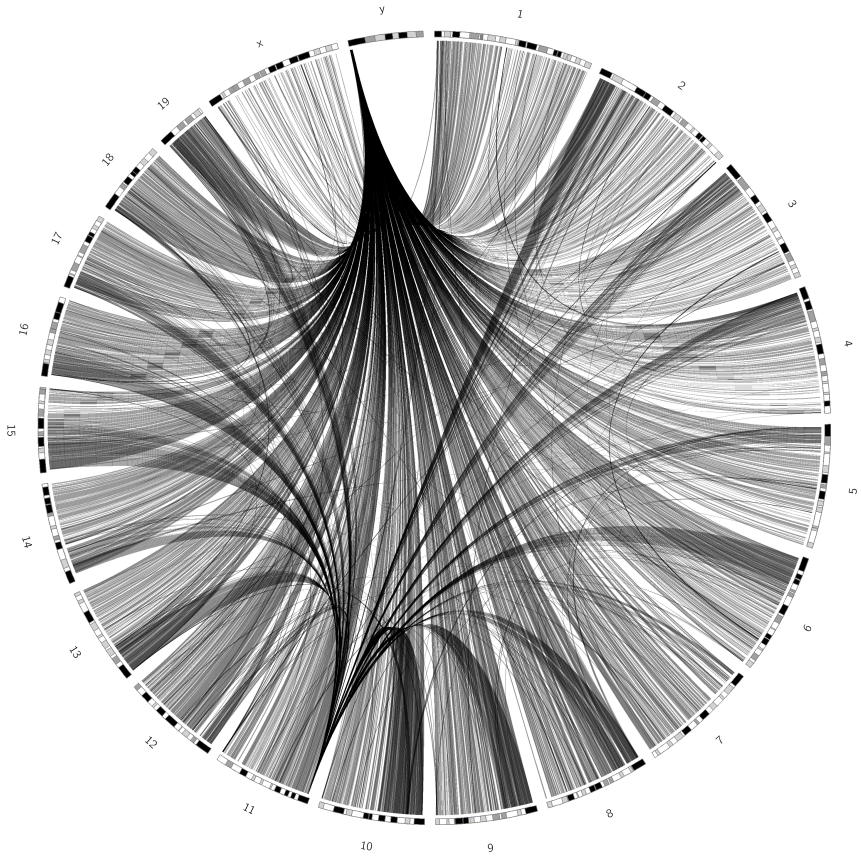


Figure 28: Circos [109] plot of the mouse segment interaction network (MSIN) at a q-value cutoff of $1e-6$. Two highly interactive segments on chromosome 11 and chromosome Y dominate the network.

These two highly interactive segments clearly dominate the entire MSIN. Since the segment on chromosome Y is so interactive and in contact with so many different segments, this feature of the MSIN explains the higher proportion of connected nodes with decreasing q-value cutoffs. Before investigating this property more deeply, the possibility of experiment contamination has to be considered. In a network with 5,093 nodes it is very notable if only two of them participate in the majority of contacts (4,131 of 4,483 edges, 92.2%).

It is theoretically possible that the Hi-C data we work with here contains experimental errors or contaminations. In the original publication, Dixon et al. worked only on intra-chromosomal interactions, so the phenomenon was not covered. Known Hi-C biases are considered and removed during the normalization process and can thus not influence the filtered data. We were able to identify an increase of paired Hi-C reads mapped to the segment on chromosome 11, 3,000,000-3,500,000 ($S1$) and the segment on chromosome Y, 2,500,000-3,000,000 ($S2$). While on average 106,305.3 reads were mapped to each 500 kb segment, the number of paired unfiltered reads mapped to $S1$

is 5.45 times higher, and 4.54 times higher for *S*2. The observed effect is thus already present in the raw data, and even enhanced through normalization, when Hi-C biases are removed.

Both segments contain little to no genes (*S*1: 11, *S*2: 0) and overlap with more repeats than the average segment (*S*1: 23,550 bp overlap = 4.6 times higher, *S*2: 14,660 bp overlap = 2.8 times higher). If reads were mapped to the genome in a non-unique fashion, a high percentage of repetitive sequences in a given segment could lead to overestimation of reads. However, in Dixon et al.'s experiment reads were mapped to the genome uniquely.

Additionally, the high number of different genomic regions both segments and especially *S*2 are in contact with (*S*1: 979, *S*2: 3,152) further proves that this observation is in fact not a contamination. During Hi-C, spatially close regions are crosslinked to each other and fragments resulting from restriction enzyme digestion are then ligated to their spatially close partner. The resulting reads are sequenced as paired end reads. It is highly unlikely that ligation occurs between one fragment and almost 1,000 or even more other fragments if these are not in close proximity when formaldehyde is added to the solution.

It is, however, also improbable that these segments are able to form all these contacts at a single time point. Due to the nature of the Hi-C experiment and in contrast to single-cell Hi-C, read counts are averaged over many different cells and thus provide a summary of all the interactions happening in millions of cells at the time point of the experiment. We hypothesize that the strong interactivity of segments *S*1 and *S*2 is caused by a high flexibility of these regions. If these segments do not have fixed nuclear territories, they are able to move around the nucleus randomly, which leads to contact formation with different regions of the mouse genome in each cell. In the case of segment *S*2 on chromosome Y this is especially pronounced, as it forms contacts with almost the entire genome.

Hi-C data is averaged over millions of cells

Segment *S*1 on chromosome 11 appears to preferentially contact regions close to the telomeres of the remaining chromosomes. Since the mouse genome is telocentric, this strong inter-connectivity of regions close to the centromeres is hinting at the existence of a spatial centromeric cluster in mouse embryonic stem cells (mESC). It is possible that chromosome 11 is located at the center of such a cluster, serving as a scaffold to connect the other centromeres.

Centromere co-localization is a well-known phenomenon that appears in multiple species and different cell types and causes strong clustering behaviour in the published inter-chromosomal contact network of yeast cells [108]. We will further investigate the possibility of centromere clustering in mouse in section 7.6.

In the case of the Y chromosome segment, the flexibility appears to be even more pronounced, as it forms over 3,000 contacts in the

different cells used in the experiment. It is necessary to note that the majority of the mouse Y chromosome contains repeats, specifically internally repetitive 515 kb long units which are repeated 150-200 times [2]. As a consequence, only the first 3 Mb of this chromosome originate from the ancestor autosome pair from which X and Y evolved [2]. Due to the high amount of repeats in the tail of the Y chromosome, only these 3 Mb can be mapped. The highly interactive and gene-less segment thus lies very close to this repeat-rich tail, and we hypothesize that in fact the whole tail is highly interactive due to lack of functionality. We will therefore generalize our assumptions on the whole of chromosome Y, since presumably only few parts of it are not highly interactive.

Since the Y chromosome is absent in females, it cannot have a stabilizing role for the three-dimensional genome structure. It is more likely that it is less incorporated into the nuclear interactome because it is very gene poor and short. We hypothesize that it has a less fixed position in the nucleus than other chromosomes, and is able to move around more freely. In the course of this movement the Y chromosome could be able to build contacts to many different genomic loci, resulting in the interactome we observe.

As mentioned before, there are no genes on chromosome Y, segment *S*2. Segment *S*1, however , contains 11 genes:

Rnf185 Ring finger protein 185 regulates autophagy, a catabolic process required for recycling of cytoplasmic organelles [203].

Pla2g3 Phospholipase A2, group III regulates maturation of mast cells [200].

Inpp5j Inositol polyphosphate 5-phosphatase regulates many different processes [146].

Selm Selenoprotein M has a neuroprotective function due to reduction in reactive oxygen species, and regulation of cytosolic calcium [168].

Smtn Smoothelin is a cytoskeleton-associated protein found in contractile smooth muscle [170].

Drg1 Developmentally regulated GTP binding protein 1 plays a role in differentiation, regulates cell growth under specific conditions and cell cycle arrest [180].

Eif4enif1 Eukaryotic translation initiation factor 4E nuclear import factor 1, involved in translation initiation [190].

Patz1 POZ (BTB) and AT hook containing zinc finger 1 is expressed at early stages of development, knock out leads to severe defects in the CNS and cardiac outflow tract, leading to pre-mature in utero death [209].

Pik3ip1 Phosphoinositide-3-kinase interacting protein 1 negatively regulates PIK3 (regulates cell division, motility, survival) and suppresses development of hepatocellular carcinoma [81].

Limk2 LIM motif-containing protein kinase 2 is involved in cell movement and possible role in inter-neuron migration through the subpallium brain region [6].

Sfi1 SFi1 homolog, spindle assembly associated (yeast) [43].

Considering the hypothesis that co-localization of genes in so called transcription factories serves the purpose of increasing transcription efficiency through reuse of machinery, genes on a segment that co-localizes with many other loci would be expected to be versatile or housekeeping genes. However, there appears to be no pattern in the functions of genes located on segment *S*1, and though many of them perform some sort of regulation, the processes they are involved in are very different and often very specific (e.g. Rnf185).

7.4.1 *The randomized mouse SIN has a uniform contact distribution*

We created a randomized mouse interaction network using randomly in a unicube distributed points as basis (see section 6.2). This randomized MSIN (RMSIN) serves as a basis for the evaluation of significant properties of the real network. We are able to show that the RMSIN has a completely different structure than the observed SIN, starting with a much more uniform distribution of contacts along the chromosomes (Supplementary Figure S10 on page 199). While in the MSIN the typical connected segment has an average degree of 2.60 and there is a high standard deviation of 56.17, the average connected segment degree of the RMSIN is 2.13 with a low standard deviation of only 1.17.

The most distinct feature of the MSIN are the two highly dynamic segments on chromosome 11 and Y, which together form more than 90% of high confidence contacts. The RMSIN does neither contain such segments nor a scale-free like topology with hubs, but instead has a maximum degree of 8, again underlining the uniformity of the network. We can thus conclude that the existence of hubs *S*1 and *S*2 in the MSIN is a non-random property.

7.4.2 *Network properties of the MSIN*

The MSIN is a scale-free network

Biological networks often share some properties. For instance, regulatory networks or protein-protein-interaction networks have a scale-free topology [3, 12], in which few nodes, i.e. genes or proteins, have a high degree, and the majority of nodes have a very low degree. In the case of a regulatory network, a so called hub with a high number of edges could represent an important transcription factor that is involved in the activation of many other genes.

The MSIN's degree distribution also has some scale-free properties, though these are slightly distorted. There are exactly two (mega)

hubs, segments *S*1 and *S*2 which have been described in previous sections. Figure 29 shows the distribution plotted with Cytoscape [184], including a fitted power law distribution. Even though the fit's correlation coefficient is high (0.914), it does not fit the data points well. The main reason for this distortion are the two mega hubs, which fall aside from the remaining points due to their exceptionally high degree. However, the network can still be considered to be scale-free, since exclusion of the two mega hubs would still lead to a power-law degree distribution.

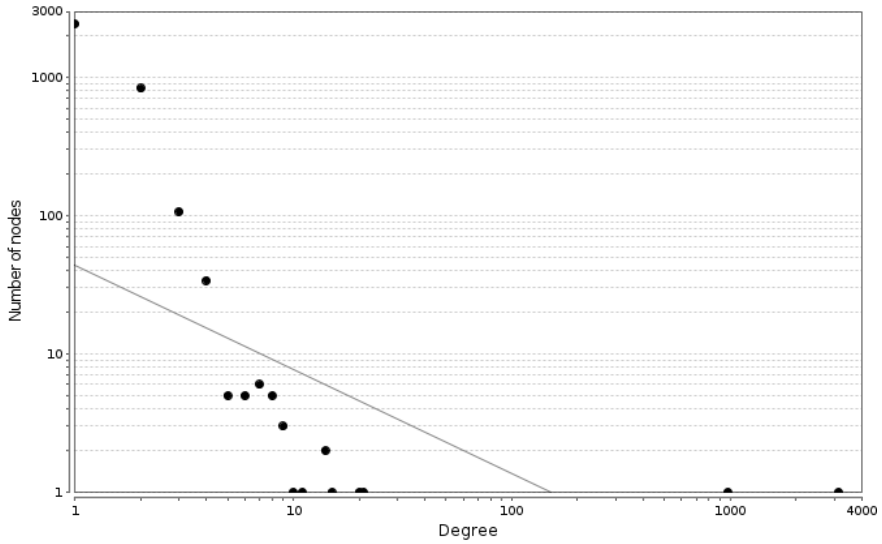


Figure 29: Degree distribution of the MSIN, plotted with Cytoscape [184] in log-log scale. The gray line shows a fitted power law distribution (correlation coefficient 0.914). Even though the fit is high, the two hubs *S*1 and *S*2 disrupt the otherwise relatively scale-free distribution.

Segments *S*1 and *S*2 not only influence the degree distribution but also many other network properties of the MSIN. While the shortest path distribution of the RMSIN is normally distributed (Supplementary Figure S9a), most nodes in the MSIN can be connected by a path of length 2 through one of the two hubs (Supplementary Figure S9b). This is also mirrored in the average shortest paths lengths of 13.2 and 2.1, respectively, given in Table 12.

The average degree over all nodes is the same in both networks due to the same number of edges and nodes. However, all other properties differ: The clustering coefficient of the randomized network is too low for Cytoscape to display it to the necessary decimal place, but slightly higher for the MSIN itself. A network's *clustering coefficient* describes the degree to which the nodes in a network tend to cluster together and is calculated based on different node triangle topologies' frequencies in the network. The reason for the still low clustering coefficient of the MSIN is probably the low number of edges between non-hub nodes. This leads to a low clustering of segments connected to a hub.

Also, the existence of highly connected hubs strongly decreases network diameter compared to the RMSIN, while increasing network centralization. The *diameter* of a network is the maximum shortest path between two nodes and connected to centrality. The *centrality* of each node (i.e. closeness or betweenness centrality) measures the importance of this node for the structure of the network. The global centralization describes how much more important the most important node in the network is compared to all others and is high for the MSIN, because segments *S*1 and *S*2 serve as a bridge connecting many node pairs.

The *heterogeneity* of the MSIN, however, is much higher than that of both the RMSIN and even the HSIN, again caused by the less uniform degree distribution.

Table 12: Basic network properties of human segment interaction network (HSIN) and its randomized version (RHSIN), and the (randomized) mouse segment interaction network ((R)MSIN).

	HSIN	RHSIN	MSIN	RMSIN
Clustering coefficient	0.006	0	0.166	0
Network diameter	13	40	7	31
Network centralization	0.014	0.001	0.619	0.001
Characteristic path length	4.690	16.066	2.137	13.177
Average degree	1.577	1.574	1.761	1.761
Network heterogeneity	2.695	0.805	26.259	0.759
Isolated nodes	3,232	1,201	1,641	882

7.5 SHORT HUMAN CHROMOSOMES FORM MORE TRANS-INTERACTIONS THAN LONG CHROMOSOMES

There is a negative correlation between chromosome length and average degree in human

The human inter-chromosomal segment interaction network contains a similar number of contacts as the previously described MSIN, yet at first glance its overall structure appears to be dramatically different (Figure 30). The main structural feature of this network is the higher abundance of interactions involving at least one short chromosome compared to those between or to longer chromosomes. In fact, there is a strong negative exponential correlation (Pearson correlation coefficient -0.70) with almost no deviation from the regression line between the average degree per chromosome and its length (Figure 31).

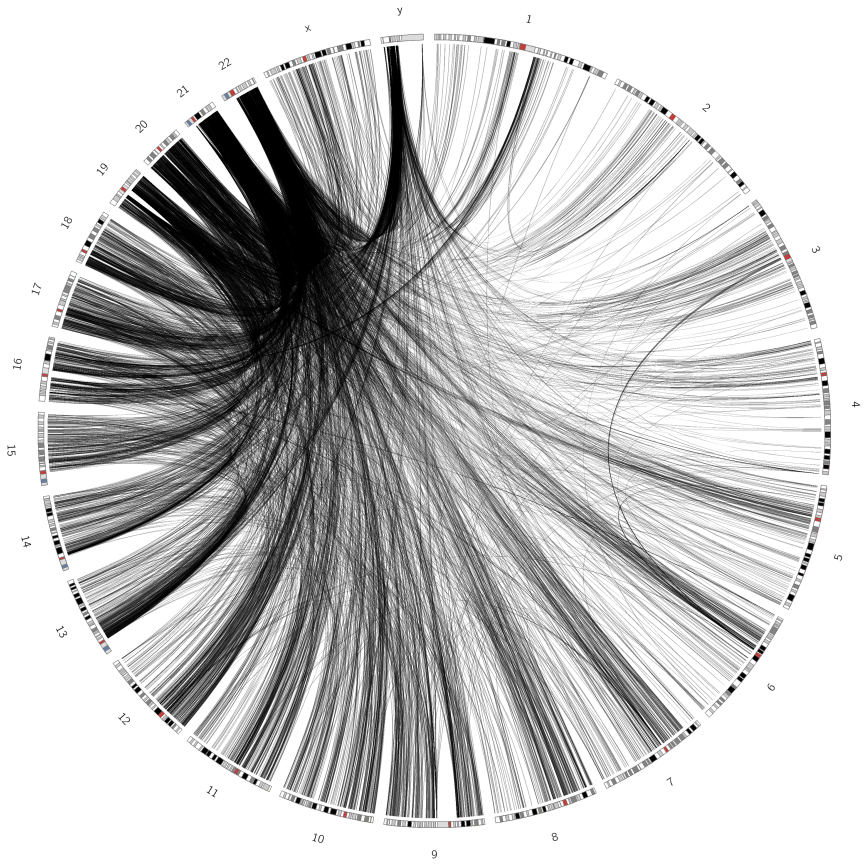


Figure 30: Circos [109] plot of the human segment interaction network (HSIN) at a q-value cutoff of $1e-3$. Short chromosomes form more interactions than longer chromosomes.

Duan et al. have also reported a higher prevalence of interactions between shorter chromosomes in budding yeast, according to Hi-C data [51]. They found that yeast chromosomes interact mostly along their entire length. Thus the so-called Rabl-like orientation, where centromeres are grouped at one pole while telomeres are sorted towards a second pole in the nucleus [167], could cause the preference for interactions between short chromosomes, which are crowded within the set of chromosome arms extending from the centromere cluster to the distal telomeres.

In general, chromosomes have been known to keep to their own territories [35]. Even though there is interweaving and chromatin loops which penetrate other territories, the surface of a short chromosome territory is still larger in relation to its length than that of a long chromosome. This and a central localization could lead to formation of a high number of contacts between short chromosomes due to architectural rather than functional reasons.

However, according to tethered conformation capture (TCC) data, human chromosome territories can be assigned to two main spatial

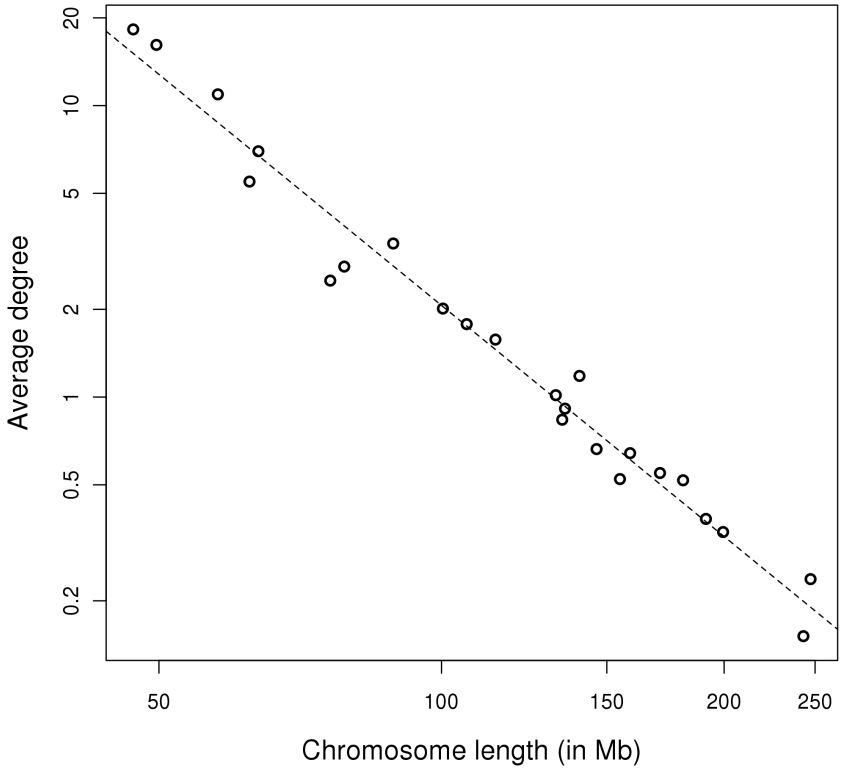


Figure 31: Correlation of average (length-normalized) degree and chromosome length in human; dotted line is the regression line, Pearson correlation coefficient is -0.70. Axes are shown in logarithmic scale.

zones based on distance-based clustering [97]. In concordance with the nuclear architecture model, the first of these groups is located at a central subnuclear region and consists of the chromosomes 1, 11, 14-17 and 19-22, which are relatively gene-rich. Our results confirm high interaction frequency between chromosomes 14-17 and 19-22, and also between these chromosomes and others, which is a logical consequence of their central position. The remaining chromosomes preferentially reside in the nuclear periphery as part of the second group. Overall, our observations of more interactive short chromosomes are in line with this model. However, our data do not enable us to distinguish between the groups for all chromosomes (e.g. chromosome 11 has similar interaction patterns as chromosome 12, even though they are in different groups).

Besides the preference for contact formation involving short, gene-rich chromosomes, the HSIN also shows an increase in contacts near centromeric regions (see Figure 30). While the mouse genome is telocentric, most human chromosomes have a more centrally located centromere. In the Circos plot, these are visualized as small red bands in the chromosome ideograms. Even though reads usually cannot be

mapped to the centromeres due to low sequence complexity, we can observe a higher contact frequency in the regions around them. We will describe the potential causes for this effect in section 7.6.

7.5.1 *The randomized human SIN does not share the HSIN's properties*

A similar effect as observed for the RMSIN can be observed for the randomized HSIN: the distribution of contacts along the chromosomes is much more uniform than real Hi-C data based networks. While the HSIN connected segment has an average degree of 3.62 with a higher standard deviation of 5.84, the RHSIN's regions have on average less contacts (1.99) with a lower standard deviation (1.10), implying similar degrees for all segments. The effect is less dramatic here than in mouse, due to the lack of high-contact regions with close to or even more than 1,000 contacts.

The observed increase in contacts for short chromosomes is thus not present in the randomized HSIN and can be considered a significant property. The maximum degree of the RHSIN is 8, so similar to the RMSIN no hubs exist in this network, either.

7.5.2 *Network properties of the HSIN*

Table 12 on page 100 lists the key network properties for both the HSIN and RHSIN. Since both of these networks lack the mega hubs that dominate the MSIN, their characteristics are very different from the mouse network. For instance, the network diameter of the HSIN is about twice as high, but still significantly lower than that of the RHSIN. In general, the same observation as for the mouse networks holds true here as well; through their scale-free topology the Hi-C networks have shorter paths (characteristic path length HSIN 4.7 vs. RHSIN 16.1) and higher heterogeneity (2.7 vs. 0.8).

The HSIN has a scale-free topology

In fact, a power-law distribution can be fitted to the degree distribution of the HSIN almost perfectly with a correlation coefficient of 0.983 (Figure 32). Because the HSIN is not distorted by mega hubs, this common property of biological networks is clearly recognizable.

Shortest path lengths are distributed almost normally in the HSIN, similar to the RHSIN, but centred around a low mean due to the existence of well-connected nodes that do not exist in the random network (Supplementary Figure S9 on page 198). In general, we can conclude that the HSIN exhibits a clear and non-random scale-free topology.

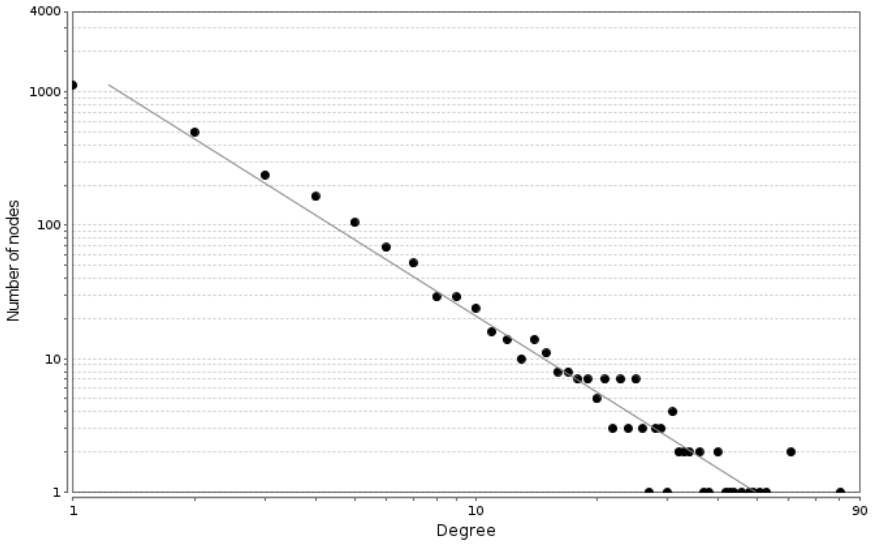


Figure 32: Cytoscape [184] degree distribution of the HSIN shown in log-log scale, power-law distribution fitted to it as a gray line (correlation coefficient: 0.983).

7.6 HUMAN AND MOUSE GENOMES SHOW CENTROMERE CO-LOCALIZATION AND A FLEXIBLE Y CHROMOSOME

While they are similar in size and connectivity (Table 13), we have shown that the defining characteristics of the human and mouse SIN differ. In this section I want to investigate the underlying similarities that might be hidden under the first impressions. I have already pointed out that both SINs have a scale-free topology (sections 7.5.2 and 7.4.2), which influences many of their general properties. Both networks share a relatively low diameter and characteristic path length, and (compared to randomized networks) high heterogeneity and clustering coefficient.

One of the characteristic properties of scale-free topologies is the possibility to reach any node from a given second node through a short path through one of the existing hubs. This effect leads to the observed low average path lengths, and could indicate that the fold of the genomes are dense structures in which any two regions are relatively spatially close. However, due to the nature of Hi-C data, the networks contain more contacts than happen at a single time point, so flexible regions lead to an abundance of contacts. The question remains whether these especially dynamic regions that can form contacts with many different genomic loci do so for functional reasons, e.g. scaffold-like bringing together other loci, or for a lack thereof. In the latter case, no fixed position in the genomic structure could lead to these hubs moving around the nucleus more freely in a diffusion-like manner. While the mega hubs in mouse probably fall in the latter category, it is plausible that hubs with less extreme degrees are centrally located segments in for example transcription factories.

Table 13: Size of **segment interaction networks** and their corresponding largest connected components for human (HSIN), mouse (MSIN) and their randomized versions. The largest component of the SINs contain the majority of genes.

Property	HSIN	RHSIN	MSIN	RMSIN
#Nodes	5,732	5,732	5,093	5,093
#Connected nodes	2,500 (43.61%)	4,531 (79.05%)	3,450 (67.74%)	4,211 (82.68%)
#Edges	4,520	4,517	4,483	4,485
Largest component: #Nodes	2,349	3,601	3,282	3,664
Largest component: #Edges	4,435 (98.12%)	3,894 (86.21%)	4,171 (99.67%)	4,147 (92.46%)

7.6.1 *Both species contain flexible Y-chromosomes*

One example that causes us to believe that the latter hypothesis is true for at least some cases is the role of the Y chromosomes in both species. In both organisms only a small portion of the Y chromosome can be mapped, and in mouse a considerably large part of this portion (17%) forms contacts to almost all other loci in the mouse genome. In human, such a strong interactivity of a Y chromosome region is not observable. However, the short Y chromosome still forms more contacts than the average chromosome (481 vs. 376.67), despite its size being less than half the average length (57.77 Mb vs. an average length of 128.35 Mb) and even less of it can be mapped after sequencing due to repeats (22 Mb).

As mentioned before, we hypothesize that the repeat-rich tails of the Y chromosomes also form many non-specific interactions, which cannot be captured by Hi-C due to the low sequence complexity (see section 7.4.2). Consequently, we believe that the majority of Y chromosomes are highly interactive. Since these short and gene-poor chromosomes cannot form so many contacts simultaneously, we presume them to be very flexible and able to form contacts with many different loci in different cells. This behaviour might be caused by the overall low gene density on chromosome Y and its lower impact on cellular expression compared to other gene-rich chromosomes. We speculate that chromosome Y may be less embedded in the inter-chromosomal contact network and able to move around and form random contacts more freely than other chromosomes in both species, though the effect is stronger in mouse.

7.6.2 *Short chromosomes form more contacts*

The main characteristic of the HSIN is the abundance of contacts between or involving short chromosomes. We have shown that there is a clear negative correlation between average degree and chromosome length in human. In mouse, this property appears to be missing, since the mega hubs on chromosomes Y and 11 dominate the network. However, when these two outlier chromosomes are excluded, a similar characteristic begins to emerge. Figure 33 shows that the remaining chromosomes have an even stronger exponential correlation between degree and length (Pearson correlation coefficient -0.87), though there is more deviation from the regression line. For this analysis only the extremely high-degree chromosomes Y and 11 were excluded, not the edges they form to the remaining chromosomes. We can thus conclude that in mouse as well as in human there is a tendency for short chromosomes to form more contacts.

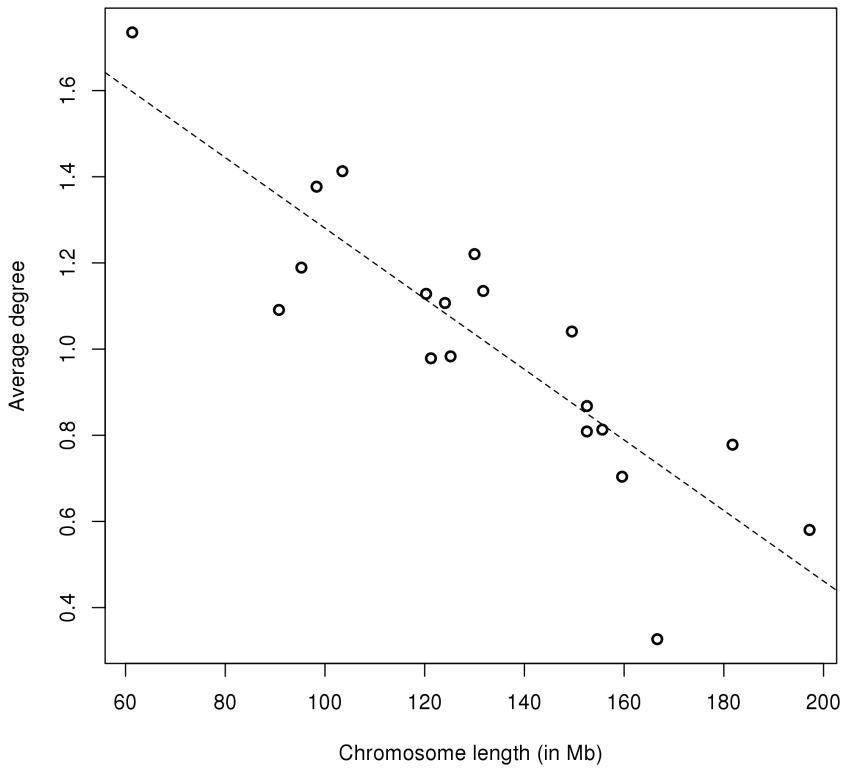


Figure 33: When excluding outlier chromosomes 11 and Y that harbour the MSIN hubs, chromosome length and average degree are negatively correlated in mouse (Pearson correlation coefficient -0.87).

In human, we believe higher gene density of most of the short chromosomes and consequential central location in the nucleus to be the reason for this observation. In mouse, the ten most gene-rich chro-

mosomes are 11, 7, 19, 17, 2, 9, 4, 5, 6 and 8. Only six of the ten chromosomes with the highest degree (excluding 11 and Y) are in this group. It thus appears as if the same conclusion does not hold for mouse. However, this observation could also be a side effect of the observed centromere clustering discussed below. If centromeres are localized spatially close, short chromosomes could be embedded in the extruding chromosomes, thus forming more contacts.

7.6.3 *Centromeres tend to co-localize to some degree*

In both species, we observed co-localization of centromere-proximal regions to different extents. In mouse, a segment close to the telomere of chromosome 11 forms contacts to telomere-close regions of all other chromosomes, potentially serving as a scaffold to bring together centromeres in a spatial cluster. In human, no such segment exists, but abundance of contacts around centromeres is still observable. I have mentioned in previous sections that clustering of centromeres is a known phenomenon in many species. One example is the yeast genome, in which a strong centromeric cluster exists [93, 94] and causes the inter-chromosomal contact network to form a similarly strong cluster [108]. Another example are drosophila polytene chromosomes, which have replicated without cell division and are thus very large, bundle together in the so called chromocenter [29, 142].

Centromere co-localization has been shown in several species, it's function is yet unknown

Before Hi-C or other chromosome conformation capture methods, biologists relied on FISH and visual interpretation to analyse the genome's conformation. With these methods, they were able to detect centromere aggregation in mouse cells as early as 1971 [90]. Hsu et al. found that centromeres aggregate in some but not all mouse cell types. They note that, in mouse, centromere sequences are highly similar, a fact which could explain their coalescence. They also hypothesize that the proximity to the nucleolus could be involved in the clustering, but are unable to prove any of their hypotheses.

Centromere clustering might be a part of the so called Rabl-orientation [167]: in this orientation interphase chromosomes are arranged in a polarized fashion, where centromeres and distal telomeres occupy opposite positions in the nucleus, leading to a certain amount of clustering of each of these. However, it is known that the mouse genome does not share this orientation, and there has only been sporadic evidence for the human genome to behave similarly [93]. Still, centromeric clustering has been shown in several cell types of these two species as well [14, 233].

Jin et al. [93] reported that budding yeast centromeres strongly cluster in different cell types and tissues and investigated whether this is due to a Rabl-orientation. They were able to show with FISH that the centromere clusters lie at the nuclear periphery and a Rabl-like orientation is suggested. They suggest that this clustering could be a

consequence of anaphase chromosome polarization. In fission yeast it has been shown that centromeres cluster adjacent to the spindle pole body and are linked to the anaphase movement of cell division.

In yeast stationary cells, similar to drosophila, the clustering of centromeres is reduced. Since in yeast chromosomes do not assemble at the cell equator, i.e. there is no metaphase plate, the centromere clustering might serve to facilitate the attachment of the chromosomes to the spindle [93].

However, if centromeric clustering close to the spindle pole body (SPB) is merely a relict of cell division chromosome arrangement, Jin et al. argue that this would be randomized by Brownian motion. In a later publication [94] they were able to improve their description of the yeast centromere cluster by adding that they are arranged around the SPB like a rosette, and also show that the clustering can be reconstituted without an anaphase. The dependence of the cluster on the kinetochore protein ndc10 implies active maintenance of the clustering.

Jin et al. suggest that the circular centromere arrangement may be due to the presence of a core bundle of microtubules around which the centromeres form a rosette. They are, however, unable to pinpoint a function of the centromere clustering, which is implied by the active maintenance.

So far, centromere clustering in human and mouse has only been shown for a subset of cell types [14, 233, 90]. The active maintenance and supposed functional role of this structure that has been observed in yeast indicates that similar structures could be conserved in mammalian species, though they might be present only during certain cell cycle phases or in certain cell types. Altogether, our results indicate a certain degree of centromere co-localization in both species, though no strong clustering can be observed.

7.7 TRANS-INTERACTING SEGMENTS ARE ENRICHED IN ACTIVE MARKS IN HUMAN

In the first part of this work we have shown that there is a complex interplay of features along the human and mouse genomes. Some features can be considered active marks that are enriched in euchromatic genome regions, while others appear mainly in heterochromatic regions. To further analyse this dependency, we investigated the overlap between a set of genomic features ranging from histone modifications to LADs (see Tables 7 and 8, page 78) and sets of autosomal 500 kb segments that do interact with other chromosomes (trans-interacting segments) and those that do not, respectively.

Previous research shows that inter-chromosomal contacts in human are enriched in active marks [120, 138]. Our results confirm this relationship (Supplementary Table S1, page 205) for active histone marks

H3k4me1, H3k4me3, H3k9ac, H3k27ac and H3k37me3 obtained from comprehensive ENCODE datasets. While the incidence of these five marks' peaks varies strongly across the human genome (Table 14), we can observe an enrichment in trans-interacting segments that lies between 14.6% (H3k9ac) and 37.3% (H3k4me1) for all of them. Even H3k27ac, for which the data were produced in a different cell type, shows a similar trend.

Table 14: Genome-wide incidence of histone mark peaks in human and mouse, according to ENCODE [8] data, given as percentage of the genome covered by peaks.

Histone modification	*H. sapiens*	*M. musculus*
H3k4me1	15.5%	2.1%
H3k4me3	2.5%	6.6%
H3k9ac	4.8%	2.3%
H3k27ac	4.4%	2.2%
H3k36me3	17.7%	4.7%

In mouse, the picture is quite different. According to the ENCODE data, the incidence of all five types of histone modifications in the genome is lower (Table 14). We are able to detect a similar distribution of histone modification peaks in mouse trans-interacting segments compared with others. Figure 34 summarizes the differences in percentage enrichment and depletion of genomic features in human and mouse.

In mouse, trans-interacting segments have no distinct feature profile

Interestingly, we can observe a different behaviour of human and mouse for most features (Supplementary Table S1, page 205): whereas all features except for the heterochromatic markers LINE and LTR are enriched in human trans-interacting regions, only LADs are clearly enriched in mouse. For all remaining features, the profiles are similar in trans-interacting and other segments. These results imply that in human, inter-chromosomal contacts are mainly formed between active and gene-rich regions. In mouse we cannot observe such a behaviour. The reasons for this are unclear. One possibility is that the mouse ESCs used in the experiment were in a different differentiation stage than the human ESCs. If for instance the mouse cells were in a stagnant phase, we could in theory observe a reduction of contacts between active regions due to this.

Another possible explanation is that the distribution of marks in these species indeed differs slightly. As we have shown multiple structural differences in the genomes of human and mouse so far, this is a valid possibility. This is also supported by the compartment model from Lieberman-Aiden et al. [120], which states that contacts are preferentially formed between regions that are either active or inactive. In this model, no increase of contacts between active segments is assumed, so we hypothesize that the observed enrichment of active

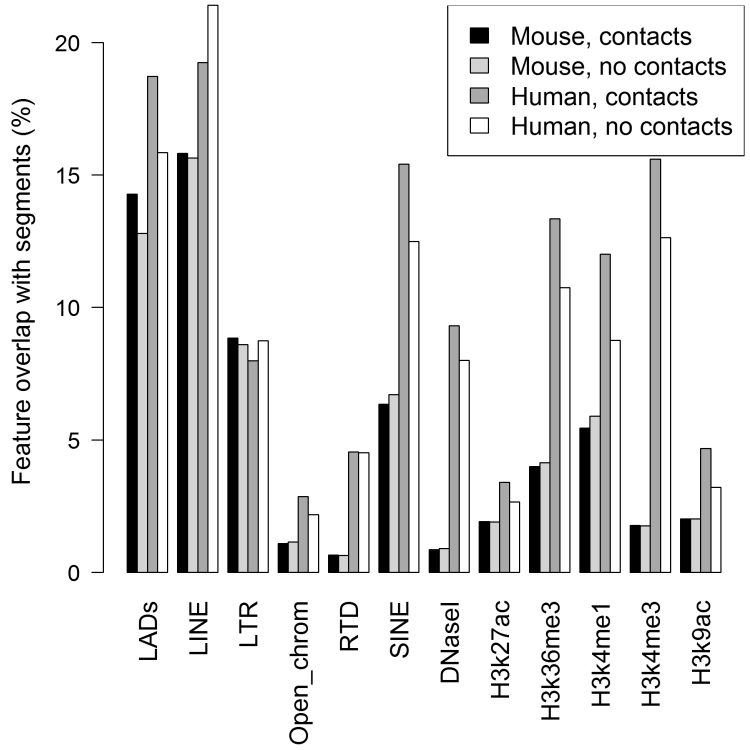


Figure 34: Overlap of features with trans-interacting segments compared to non trans-interacting segments in human and mouse.

marks in human trans-interacting segments mirrors a different feature composition, whereas the lack of this enrichment in mouse does not necessarily indicate a different differentiation stage.

7.8 GO TERM SIMILARITY IS ASSOCIATED WITH SPATIAL PROXIMITY IN HUMAN AND MOUSE

Co-localization of genes in the nucleus has the potential to be functional, increasing transcription efficiency of co-expressed or functionally related genes in transcription factories. For budding yeast, a correlation between inter-chromosomal contacts and functional similarity has already been shown [86], and, similarly, Khrameeva et al. were able to show a similar correlation for a human lymphoblastoid cell line [104].

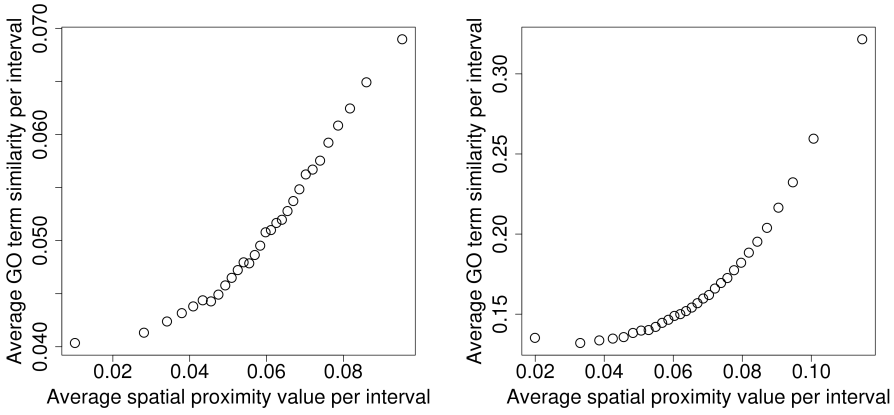
As mentioned before, it is possible that large amount of noise in the data combined with a relatively low sequencing depth hides a relationship between spatial contacts and GO term enrichment. Khrameeva et al. [104] have shown a positive correlation between GO term similarity and spatial proximity for inter-chromosomal contacts of human fibroblasts. Since Hi-C data contain many random contacts, this might lead to underestimation of association with GO term similar-

ity in analysis as described above. We repeated their approach and calculated GO term similarity for all inter-chromosomal pairs of segments, as well as spatial proximity values calculated as described in section 6.1.4.

Binning of the data is necessary to uncover associations

We are able to reproduce Khrameeva et al.'s results for both human and mouse. When grouped into 30 spatial proximity intervals, mean data from mouse and human are strongly correlated with average GO term similarity (Pearson correlation coefficient 0.89 and 0.96, respectively, Figure 35). However, as explained in detail in section 6.2.5, this effect could be due to correlation inflation, as we were unable to observe correlations on the unbinned dataset (-0.03 and 0.09, respectively). Supplementary Figure S11 shows the distribution of Pearson correlation coefficients for 1000 randomized datasets in both human and mouse. In comparison to these, our observed coefficients are significant as assessed with the cumulative distribution function (p-value < 0.01), confirming that there is indeed a positive association present in the data. However, the observed values of 0.96 and 0.89 are obviously overestimations due to variance reduction and should be seen only as a trend indicators of an association. From these results we can conclude that there is a tendency for segments with similar contact profiles to share a functional similarity, but we can make no assumptions regarding the strength of this relationship.

Still, it implies the existence of functional spatial clusters where genes with similar functions are co-localized, such as transcription factories.



(a) Human (b) Mouse

Figure 35: In both human and mouse, average GO term similarity increases with spatial proximity (Pearson correlation coefficients 0.96 and 0.89, respectively). In mouse, this relationship is even exponential (Spearman correlation coefficient 0.99). Data are binned into 30 equal-size spatial proximity intervals (see Section 6.2.5).

7.9 SPATIAL PROXIMITY IN HUMAN IS ASSOCIATED WITH CO-EXPRESSION

Homouz and Kudlicki have also shown a correlation between co-expression and contact frequency for budding yeast [86], which has been confirmed by Khrameeva et al. [104] for human lymphoblasts. Similar to functional enrichment, we analysed whether co-expression is correlated to spatial proximity. We calculated co-expression measures based on stem cell expression profiles from Liu et al. [123] for each pair of segments as proposed by Khrameeva et al. [104], and used a noise-reducing binning method to test for association between these measures and spatial proximity values.

Khrameeva et al. [104] showed that co-expression is correlated with spatial proximity in human fibroblasts. According to the authors and similar to GO term similarity, the correlation is not obvious and noise-reduction is necessary to measure the association. Indeed, we were able to detect a strong correlation between average spatial proximity values and hESC co-expression measures after using the binning method described in section 6.2.5 (Pearson correlation coefficient 0.99, Figure 36). Again, we ensured that this is not caused by correlation inflation by comparison to randomized data (p-value < 0.01, Supplementary Figure S12). This observation confirms the results from Khrameeva et al. for stem cells and suggests that co-localization is functional in many cases, though this association is occluded by random contacts.

Hi-C data noise is probably the reason we were not able to find such correlations in the raw data. While our results clearly show that it is complicated to detect present trends in the data, these could also be used to distinguish functional from random contacts. Using co-expression data sets and GO term similarity, functional gene contacts could be separated from others. However, it has to be kept in mind that there are other functional contacts that do not involve genes on both sides, such as regulator-gene interactions, which would be lost in such an approach.

7.10 HOXB AND HOXC CLUSTERS CO-LOCALIZE IN HUMAN

During GO term enrichment analysis we detected a contact between the HOXC cluster on chromosome 12 and the HOXB cluster on chromosome 17 in human. Homeobox (Hox) genes code for vital transcription factors that are involved in embryo development. It is well known that their expression depends on their order in the genome in many species [113]. In human, four such clusters exist on different chromosomes, but the described physical contact is the only existing one according to our data. The structure of HOX clusters is well con-

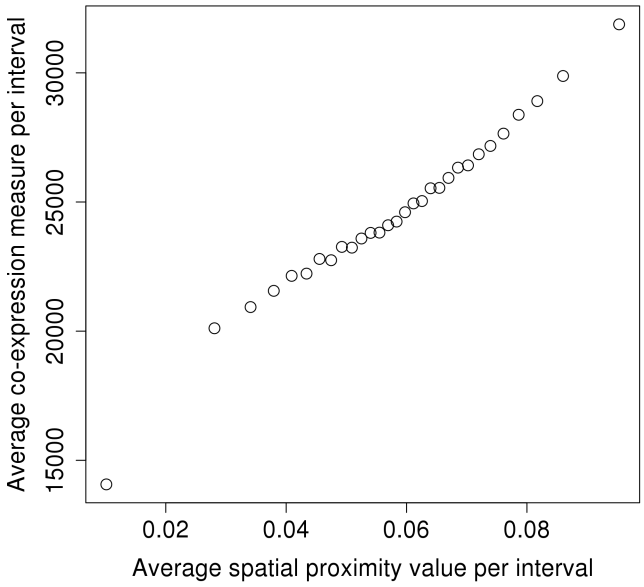


Figure 36: In human, average stem cell co-expression increases with average spatial proximity (Pearson correlation coefficient 0.99). Data are binned into 30 equal size spatial proximity intervals. The observed association is non-random (p-value < 0.01).

served between human and mouse [223], but we were unable to find any similar contacts in mouse.

Considering the important role of HOX genes in embryonic development, the identified contact between HOXB and HOXC in human could be functional. Most HOX clusters in the genome contain copies of different subsets of Hox genes. However, no copy of HOXB genes Hoxb1 and Hoxb2 is present in the HOXC cluster [113], while copies of the HOXC genes Hoxc10, Hoxc11 and Hoxc12 are absent in HOXB. The physical contact could thus extend the HOXC cluster in the three-dimensional space, complimenting the linear HOX clusters with missing genes. In Drosophila, Hox genes are co-regulated and co-localized in so-called Pc bodies when they are repressed. This indicates that we could expect even more contacts in differentiated cell types.

HOX clusters are linear sequences of genes involved in embryo development

The fact that we are unable to find a similar contact in mouse may again be caused by the promiscuous Y chromosome segment, which forms so many contacts to gene-poor regions that such an HOX cluster interaction might simply be lost during q-value filtering.

7.11 CTCF AND RAD1 BIND THE MAJORITY OF GENES IN HSIN SPATIAL CLUSTERS

We analysed TFBS in human using ENCODE [8] data on 50 transcription factors (for a list, see Supplementary Table S2). We searched for TFBS in genes of spatial clusters to identify preferences for certain transcription factors (Figure 37). The heatmap shows the percentage of genes in a spatial cluster (rows) that overlap with at least one binding site for a given transcription factor (columns). Transcription factors are clustered (see dendrogram in Figure 37).

Transcription factors can be grouped into those that are common in many spatial gene clusters, which are USF1, YNF143, CTCF, RAD21, RBBP5, SIN3A, TBP, POLR2A and TAF1, those that bind only few genes in each spatial cluster, which is the largest subset, and a group of transcription factors which bind nearly no genes in spatial clusters. An overview of the average percentage of genes in spatial clusters that has a binding site for a certain transcription factor is also given in Supplementary Table S3 on page 208.

CTCF and RAD21/Cohesin are both known to be involved in spatial chromatin organization

In the small clusters of co-localized genes in human an average of 56% and 63% have binding sites for CTCF and RAD21, respectively. Both these transcription factors are known to be involved in the organization of the genome structure. CCTC-binding factor (CTCF) is a highly conserved protein that is required for long range interactions [89]. Its so far been implicated in many different functions, ranging from insulator activities over imprinting, promoter activation and repression to facilitation of large distance contacts [174, 162, 193].

RAD21 is a subunit of the Cohesin complex and known to be essential for sister chromatid cohesion [132, 213, 73, 183, 193]. Cohesin has a ring-link structure that can hold two strands of DNA close together, therefore being a possible mediator of long-distance DNA contacts. Chromatin conformation studies have already shown that Cohesin is able to form such long-range interactions between its binding sites, and can establish and maintain them even across different chromosomes (see for example [72, 139, 32]).

As most genes in spatial gene clusters have binding sites for the Cohesin subunit RAD21 and CTCF in human, these two proteins could play important roles in establishing the contacts between these genes and/or maintaining them. Both factors often work together; it has been shown that Cohesin can stabilize CTCF binding [89, 193], and conversely CTCF is believed to function as a recruiting factor for Cohesin [221, 89, 193]. The abundance of sites for both factors in co-localized gene clusters is thus not surprising.

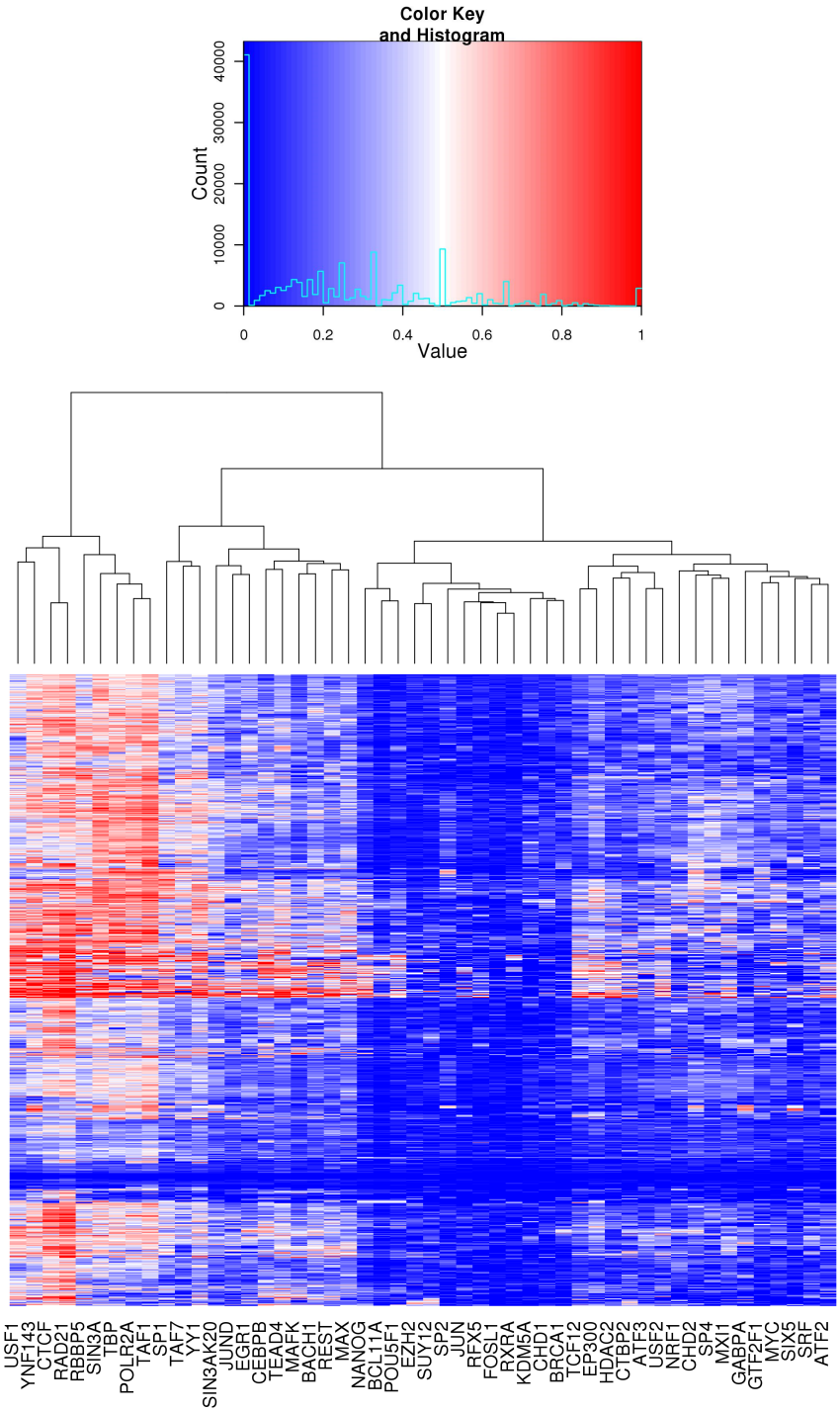


Figure 37: Heatmap showing the percentage of genes in a spatial cluster that overlap with binding sites of a certain transcription factor. CTCF and RAD1 bind most spatially interacting genes.

7.12 INTER-CHROMOSOMAL CONTACTS ARE NOT CONSERVED BETWEEN HUMAN AND MOUSE

Dixon et al. [45] have shown that the domain structure of intra-chromosomal contacts in the form of TADs is conserved between human and mouse. Using the same data, we show that orthologous regions of conserved gene order between both species do not have significantly more conserved contacts than random non-orthologous regions. Out of 3,207 such regions (see part iv for details on the definition of these regions), only 278 form at least one conserved spatial contact, and only 1% of all genes are involved in these conserved interactions.

After random shuffling of association between orthologous regions, we detected conservation of at least one contact for 234 synteny regions. Because of the generally low overlaps in contacts between orthologous regions and the similar results for randomized data, we conclude that the observed overlaps are not biologically significant and that the inter-chromosomal interactomes are not conserved between human and mouse.

Intra-chromosomal contacts are conserved between mouse and human, inter-chromosomal ones are not

Considering the evolutionary history of both genomes, these results are not surprising. A number of large-scale rearrangements, involving regions of multiple megabases, have occurred in both genomes since species separation (also see part iv). As a result, the mouse genome appears to be a mosaic version of a human genome broken apart into its synteny regions, and vice versa. More than 300 such synteny regions exist. For example, the human chromosome 1 contains regions whose orthologs lie in mouse on chromosomes 1, 3, 4, 5, 6, 8, 11 and 13 [211].

Considering this and the fact that chromosomes keep to their own territories [35], it is not surprising that the evolutionary macro-rearrangements that formed the contemporary genomes disrupted the inter-chromosomal interactome of the ancestor genome, while keeping intra-chromosomal contacts over relatively small distances like those in TADs largely constant. In fact, it has been shown that evolutionary chromosome breaks fall preferentially in border regions of TADs, presumably because disruption of the set of interactions within one such domain would be deleterious [143].

While these results suggest a lack of conservation of inter-chromosomal contacts, we have already shown that, on the functional and structural level, both human and mouse genomes have a lot of similarities. Structurally, the abundance of contacts from short chromosomes and flexibility of Y chromosomal regions are present in both species. Functionally, we have shown a similarly strong association between spatial proximity and GO term similarity, indicating that inter-chromosomal transcription factories form in both genomes. After the spatial structure of the chromatin is disrupted due to a large-scale rearrangement, Brownian motion might have led to new ran-

dom contacts between genes with similar functions which then were fixed due to their functional advantage. Though there is no conservation of individual gene contacts, these results imply some degree of conservation of spatial structure and its function.

7.13 COMPARISON TO THE PUBLISHED YEAST HI-C INTERACTION NETWORK

We have adapted the network creation approach proposed by Kruse et al. [108] for budding yeast and made some changes to account for the low coverage of Hi-C data in large genomes. One main difference is that we binned the reads into 500 kb segments and performed p-value calculation separately for pairs of chromosomes, while Kruse et al. worked with Hi-C fragments and calculated p-values for the whole genome.

Though differences in the network structures can be expected due to the significantly different sequencing depths, we compared our SINs' properties to those of Kruse et al.'s yeast segment interaction network. It has also to be kept in mind that the yeast genome is of course much smaller, allowing for a higher number of inter-chromosomal contacts to form relative to intra-chromosomal ones, and is also haploid [238]. The yeast genome is known to cluster at the centromere at a fixed subnuclear position, a fact which dominates the network in terms of clustering behaviour and other parameters [93, 94].

The yeast interactome is strongly clustered due to centromere co-localization

Table 15 gives an overview over network sizes of the inter-chromosomal segment interaction networks for different q-value cutoffs in yeast, human and mouse. Even though the genomes are very different in size, the total number of nodes is comparable between yeast and the mammalian networks. This is caused by a higher sequencing coverage in the Hi-C experiment on the yeast genome, which leads to an increase in resolution and allows use of shorter segments, explaining the similar numbers. Consequently, the number of edges in the yeast segment interaction network at a considerably low q-value cutoff of $1E-3$ is significantly higher than that of HSIN and MSIN (between 20 and 16 times higher). We can observe a strong decline of edges with decreasing cutoff in all three species.

The number of unconnected nodes is dependent on the number of edges and consequently varies between the species, too. Since the yeast network strongly clusters around the centromeres, almost all genes that have at least one connection are part of a single connected component (cutoff $1E-3$: 100% of genes, $1E-10$: 73.74%). We have observed a similar effect in both human and mouse, where at a cutoff of $1E-3$ 99.81% of genes and 99.95% of connected genes are part of this large connected component, respectively. We were also able to show that there is centromeric clustering to some degree in both species, similar to yeast.

Table 15: Size of *S. cerevisiae*, *H. sapiens* and *M. musculus* segment interaction networks at different confidence value cutoffs.

Cutoff	#Nodes	#Singletons	#Edges
S. cerevisiae SIN			
$1E-3$	4,454	284	90,658
$1E-4$	4,454	571	44,720
$1E-6$	4,454	1,749	16,691
$1E-8$	4,454	2,830	8,583
$1E-10$	4,454	3,460	5,218
HSIN			
$1E-2$	5,732	1,604	13,674
$1E-3$	5,732	3,232	4,520
$1E-4$	5,732	4,390	1,736
MSIN			
$1E-2$	5,093	730	6,483
$1E-3$	5,093	1,082	5,589
$1E-4$	5,093	1,273	5,133

We analysed the segments with the highest degrees in the yeast interaction network. At a cutoff of $1E-6$ we found 18 segments with degrees over 90, while the next highest degree is only 9 (see Supplementary Table S4).

In human, there is no such extremely highly connected set of nodes within the core; the highest degree of segments observable is 61 for two segments (at cutoff $1E-3$), and the second highest is 53. Only one of these three segments is located close to the centromeres of human chromosomes (according to UCSC cytobands), but four of them lie on chromosome 21 and two on chromosome Y. As mentioned earlier, the short, gene-rich chromosomes play a major role in the cluster of connected segments in human.

In mouse, the clustering behaviour of the network is more similar to yeast. At a confidence threshold of $1E-6$, there is the previously described highly connected segment from chromosome 11 with 979 contacts, and a segment from chromosome Y which forms 3,152 contacts, exceeding the highest degrees from yeast by far. Due to the telocentric nature of mouse chromosomes, these regions lie within or close to the centromeres. Since all three SINs share a scale-free like topology to different degrees, the yeast genome's network structure falls somewhere between mouse with its extreme hubs and human.

Altogether we can conclude that, like the yeast genome, the human and mouse genomes tend to cluster. The inter-chromosomal contacts

are most dense in certain regions of the genome: in yeast and mouse, it appears to be centromeres that co-localize, while this effect is less strong in human. There, short chromosomes with many contacts are the most distinctive property. The main difference between the yeast network and the human and mouse networks is the higher number of inter-chromosomal interactions that Kruse et al. were able to identify in yeast. This effect is completely circumstantial and caused by the smaller genome size, which leads to a strongly increased average read coverage and consequentially shorter segment size in yeast. This influences the p-value calculation, because more true contacts can be distinguished from the background. Additionally, the shorter size also can lead to proportionally more inter-chromosomal contacts. Chromosomes remain in their own territories [35], and the larger the chromosome, the larger the region that is embedded within such a territory.

Altogether, the yeast segment interaction networks has some general properties that can also be found in the human and mouse segment interaction networks, and differences caused by different sizes of the genomes.

7.14 INTER-CHROMOSOMAL CONTACT PREDICTION SUCCESS IS HIGHLY SPECIES-DEPENDENT

As described in part ii, the linear genome is constituted of many domain-like features which depend on one another and form the linear structure. This feature composition also greatly influences the three-dimensional structure, since it participates in structuring the genome into active and inactive regions which are brought together in the nuclear space through chromatin conformation. In this section we aim to investigate whether this inter-dependency is strong enough to predict inter-chromosomal contacts from easy to obtain linear features.

Hi-C experiments are extensive and costly, so currently only few genomes and cell lines have been analysed. If we are able to show that inter-chromosomal interactions can, to some extent, be predicted from sequence or other more readily available genomic properties, this could pave the way for a computational Hi-C equivalent.

We have trained and tested a Random Forest classifier on feature vectors for 500 kb segment pairs from human, mouse and a combined set. Hi-C networks at the previously described strict thresholds were used as a basis to distinguish contacts (positive class) from non-contacts (negative class). We believe that use of a strict threshold can help reduce the effect of noise and outliers.

7.14.1 *Classification accuracy of contacts is low with cost-sensitive classifier*

A cost-sensitive classifier mis-classifies many instances in the contact class

Our first method to reduce the effect of class imbalance was to apply a cost-sensitive classifier with adapted misclassification penalties for the minority class. Results are summarized in Table 16. High accuracies can be expected even for simple classifiers due to the extreme imbalance of the training and test set. We thus need to focus on class-specific measures, such as area under precision recall curve (AUPRC), precision and recall for the positive class (class 0). Indeed, precision and recall, which are high for the majority class for all sets, reach only medium values on the class of interest (precision: 0.35 to 0.47, recall: 0.28 to 0.62). Similarly, the classifier also achieves good values for the non-contact class in other measures, while recall, AUPRC and AUROC are much worse for the class of contacts.

Table 16: Evaluation results for a cost-sensitive Random Forest classifier that regards class imbalance. Class 0 is the positive class, i.e. 'contact', class 1 is the negative class, i.e. 'no contact'. Precision and recall for the minority class reach only medium values, and AUPRC for the contact class is extremely low in human.

Measure	*H. sapiens*	*M. musculus*	Combined
Accuracy	99.96%	99.96%	99.97%
AUROC	0.74	0.98	0.85
AUPRC, class 0	0.01	0.37	0.28
Precision, class 0	0.35	0.43	0.47
Precision, class 1	1.00	1.00	1.00
Recall, class 0	0.28	0.62	0.52
Recall, class 1	1.00	1.00	1.00

These medium to very low values for the minority class (0) imply that either adjustment of the misclassification error is not a sufficient method to deal with imbalance, or that there is not enough predictive signal in the data to distinguish contacts from other segments. This effect is strongest for human, where the AUPRC is close to 0 and recall and precision reach only 0.28 and 0.35, respectively. For mouse, the predictability of the positive class appears to be higher, with 62% of positive instances being predicted as positive. Still, the precision is low, indicating that many instances are classified false positively. As expected, the results for the combined set lie between human and mouse.

From this we can conclude, that at least with this method to deal with imbalance each species appears to have specific relationships between features and classes. In human, these relationships are less pronounced than in mouse and combination of both species does not

bring a significant improvement. Additionally, in neither species are features predictive enough to achieve a good approximation of filtered Hi-C data.

7.14.2 *Classification on a balanced set fails to achieve good precision on an imbalanced holdout set*

We performed random undersampling of the negative class to create a balanced training set, and evaluated classifier results on an imbalanced holdout set comprising 10% of the original data. Undersampling was performed 1000 times with random subsets of the negative class. Table 17 summarizes the evaluation results for human, mouse and the combined set of both species. Similarly to the cost-sensitive classifier, good results for the class of interest are only achieved for recall (0.93 to 0.97). Again, the worst prediction results are achieved in human, where precision for the contact class is extremely low (0.002), implying a very high number of false positives. Similarly, the AUPRC is very low (0.01), whereas the classifier trained on mouse data reaches a medium value of 0.40. The combined dataset performs significantly worse than the predictor trained on mouse data, reaching a precision of only 0.01 and AUPRC of 0.05 for the minority class.

Table 17: Evaluation results of inter-chromosomal contact prediction, with a Random Forest classifier that was trained on a balanced undersampled training set and tested on an imbalanced holdout set comprising 10% of the data. Though recall reaches good values for the minority class in all sets, a large number of instances are classified false positively.

Measure	*H. sapiens*	*M. musculus*	Combined
AUROC	0.99	1.00	0.98
AUPRC, class 0	0.01	0.40	0.05
Precision, class 0	0.002	0.05	0.01
Precision, class 1	1.00	1.00	1.00
Recall, class 0	0.93	0.97	0.94
Recall, class 1	0.87	0.99	0.94

Prediction precision in the contact class is species-dependent

We conclude that the relationship between features and contact classes is very species-dependent. Additionally, when combining data from both human and mouse, precision and recall of both classes are only slightly improved over the set of human features alone. This implies that the classifiers are unable to properly distinguish positive from negative class instances in the human dataset, and gain only few information from the mouse set when both are combined.

Our goal was to identify if it is possible to use linear sequence and structure features that are easier to obtain than Hi-C data for prediction of inter-chromosomal contacts. Our results show that this is

possible to some extent for mouse, whereas a high number of false positives still remains, but the high species specificity of the evaluation results makes this discovery inapplicable to other species. We can conclude that there is no species-independent subset of features tested here that influences inter-chromosomal contact formation so strongly that it can be used to predict contacts in new species where Hi-C experiments have not yet been conducted. Such a result would be highly unreliable and produce a large amount of false positive predictions, and while it may be close to the truth for some species, there is no way to assess the true prediction accuracy without Hi-C experiments.

7.14.3 *Feature selection confirms lack of predictive power of linear features*

Table 18: Results of Best First feature selection on the complete imbalanced sets and 1000 randomly undersampled sets. Due to the nature of the data as segment *pairs*, each feature appears twice in the list of attributes, once for each segment. Numbers in brackets refer to the segment to which the listed feature belongs.

Set	Imbalanced set	1000 balanced sets
H. sapiens	Gene density (1,2), DNase I (1), SNPs (2), Chromosome (2)	Gene density (1,2), Chromosome (2)
M. musculus	H3K27ac (1), RTD (2), Chromosome (2)	Chromosome (1,2), RTD (2)
Combined	LADs (1,2), LINE (2), RTD (1,2), H3k27ac (1,2), SNPs (2), Gene density (2), Chromosome (1,2)	Gene density (1,2), LINE (2), H3k27ac (2), Chromosome (2)

Feature selection results also confirm that there is no clear correlation between any subset of features and the class. Due to the nature of the data, each linear feature appears as two attributes in the input, one for each segment of the pair. Since both segments stem from the same organism and it was arbitrarily determined which was listed first and which second, we would expect both attributes of a feature to be selected if it turned out to be predictive. However, this is only rarely the case (see Table 18). Selected features are diverse, ranging from the general attributes 'Chromosome' to varying histone modifications, and almost always only one of the two attributes of a single feature is selected. Additionally, the overlap between features selected based on the imbalanced and balanced sets is lower than expected, and there is almost no overlap between human and mouse. These results confirm our previous assumption that the relationship between

features and contact propensity is not only highly species-specific, but also not very strong for the current set of features.

It is possible that consideration of more or other sequence features and/or inclusion of intra-chromosomal contacts leads to better prediction results. However, our results imply that the highly predictive features vary between species, and that even an improved classifier cannot be transferred to another species and achieve the same results. The high flexibility of the inter-chromosomal contact network and a generally lower conservation compared to intra-chromosomal contacts, which also have distinct properties, makes it almost impossible to develop a species-independent contact classification method.

8

CONCLUSION

Using a network-based approach and published ESC Hi-C data, we have analysed the intrinsic characteristics of the inter-chromosomal interactome in two mammalian species, *H. sapiens* and *M. musculus*. We analysed relation to other genomic features, functional and structural aspects and performed a holistic comparison of the two species' genome structures.

We have applied Yaffe and Tanay's [230] normalization approach, followed by the p-value based contact filtering suggested by Kruse et al. [108] with modifications to account for large eukaryotic chromosomes. For both human and mouse we created inter-chromosomal segment and gene interaction networks (SINs and GINs, respectively) of similar sizes, comprising around 4,500 edges. Biological networks often have a scale-free topology, a characteristic we can also observe for SINs but not in their randomized versions. Hub segments with many contacts in a physical contact network can be assumed to either have a very central position in the nucleus that allows them to contact many other segments at once, or be an artefact of the Hi-C method.

At the current state of research, most Hi-C data are captured over millions of cells and then averaged. While there are advances on single-cell Hi-C [138], the data used in this dissertation stem from a conventional Hi-C experiment. A highly connected segment or hub can in this case also be caused by a genome region that is not embedded into the DNA structure in a fixed way, but instead can move around flexibly and contact many different genome regions in different cells. Due to the extremely high amount of contacts of hubs in mouse and general position of these hubs close to centromeric or repeat-rich regions, we conclude that this option is more probable.

When comparing the SINs of human and mouse, we can identify some similarities besides the overall scale-free topology, which is more pronounced in mouse. Both species have contacts on the Y chromosomes which are very interactive and thus, presumably, very flexible. We assume that lower gene-density and thus a lower number of functional contacts allows this chromosome, and especially the regions close to the repeat-rich tail, to move around more freely than others in the nucleus.

Additionally, we show that short chromosomes tend to form more contacts than long ones. This feature is especially pronounced in human, where chromosome length is negatively correlated with (length-normalized) number of contacts. This implies a more central position

of short chromosomes in the nucleus. However, it is possible that this observation is an artefact of gene-richness. Many of the short chromosomes in human are considerably gene-rich, and it has been shown previously that the nucleus organizes chromosomes in two phases, where gene-rich chromosomes are located in the inner and central phase [97].

In other species such as *S. cerevisiae*, centromere co-localization is common [93, 94]. In human and mouse, we can also see a higher abundance for contacts close to the centromeres. In mouse, a highly connected segment on chromosome 11 appears to have a central role in a spatial cluster of centromere close regions, while in human this trend is only weak. Due to the sequencing and mapping steps, centromeres themselves cannot be covered in Hi-C experiments, so it is possible that stronger associations exist.

Besides structural features of the SINs we also investigated the feature composition of trans-interacting segments. As has been reported previously [120, 138], we show that autosomal trans-interacting segments in human are enriched in active histone marks, as are other features such as open chromatin or SINE. In mouse, however, the distribution of these features is very similar between trans-interacting and other segments, with the exception of LADs, which are enriched in both human and mouse contacting segments. We are unsure what causes these differences, but we believe that it could be caused by different differentiation stages in the ESCs of human and mouse used in the Hi-C experimens. If we hypothesize that inter-chromosomal contacts are preferentially formed between active regions,such a difference would explain the observations. However, it is also possible that the observed differences mirror only actual differences in the human and mouse genome. For example, the frequency of histone marks is very different in both genomes to begin with. Additionally, we know that according to the compartment model by Lieberman-Aiden et al. [120] contacts are formed preferentially between regions of the same activity state, but not limited to active regions. It is thus possible that the observed differences are only caused by different feature compositions of both species' genomes.

We have further conducted an analysis of transcription factor binding sites in inter-chromosomal contacts and can confirm the important role of CTCF and RAD21 in either establishing or maintaining these interactions[89, 174, 193, 72, 139, 32].

While investigating the relationship of spatial proximity and functional features, such as co-expression and GO term similarity in inter-chromosomal segment pairs, we discovered that large amounts of noise introduced by random interactions obscures any signal in the raw data directly derived from Hi-C experiments. In previous research, this problem has been circumvented by a complexity reducing binning approach. However, while we show that division of the data

into bins unveils a very strong positive correlation between average functional features and spatial proximity in both human and mouse, we have concluded that this binning method is not statistically valid on its own. It's main function is to reduce variance in the data, uncovering possible underlying trends, but this property can also lead to statistically insignificant results.

However, comparing our results to randomized data we show that the positive correlations between average GO term similarity and spatial proximity in human and mouse, and co-expression (for which no comparable data are available for mouse) and spatial proximity in human, are statistically highly significant. We conclude from this that there is, in fact, a tendency for segments with similar contact profiles to have functional similarities, but have to keep in mind that there is a very large amount of random contacts captured with Hi-C data which obscure this relationship.

We suggest that, in future, noise reduction has to be performed carefully on the Hi-C data to ensure its statistical validity. Additionally, it might be possible to use functional similarity of segments containing genes as an indicator for non-random contacts in filtering.

Since Hi-C and familiar methods are currently expensive and extensive, we evaluated to what extent inter-chromosomal contact formation can be predicted from sequence features. We trained a Random Forest classifier on a set of linear features, comprising repeats, histone modifications and others, to distinguish segments that form inter-chromosomal contacts from those that do not. Extreme class imbalance in favour of the non-contact class made classification more complicated and required measures such as balancing the input set or using cost-sensitive classifiers. We found that the strength of the relationship between these features and the classes is highly species dependent. While in mouse the Random forest classifier trained on a balanced set was able to distinguish positive from negative class instances fairly well, a similar classifier trained on human data misclassified too many negative instances, rendering the method useless. We conclude that, with current state of knowledge, it is not possible to develop a species-independent classification method that uses linear features as input to predict inter-chromosomal contacts without the use of Hi-C.

While we have shown many similarities in human and mouse SINs, we were unable to find conservation of individual inter-chromosomal contacts between genes. It has been shown previously that the intra-chromosomal contact landscape is largely conserved [45], however, due to large-scale chromosome rearrangements (see part iv) and the tendency of chromosomes to keep in their own territory [35], it is almost impossible for the contacts of a certain region to be conserved in both species. However, we believe that functional and structural similarities such as those we have described before, clearly show some

degree of conservation for the properties of the inter-chromosomal interactome. Though, in contrast to the intra-chromosomal interactome, this conservation is less pronounced at the level of individual genes, the overall properties of the network are still similar in human and mouse. However, since many non-specific and presumably non-conserved contacts are also formed, the networks also exhibit some striking differences, as described above.

In future work, it would be interesting to investigate the topological differences and similarities between different cell types. Integration of intra-chromosomal contacts into these networks might also help understand their properties better. Ultimately, a network-based interpretation of the complete human and mouse chromatin interactome at different stages of differentiation would provide a more complete picture of the chromatin organization. Additionally, integration of other mammalian or eukaryotic genomes when new Hi-C data become available could shed more light on the functional and random aspects of these networks. While we have employed a bias-reducing normalization method, Hi-C still struggles with problems such as noise, biases and sequencing depth. The advent of methods with lower signal-to-noise ratio, such as TCC and single cell Hi-C, can also greatly enhance the resolution of these analyses and our understanding of the three-dimensional genome structure.

Part IV

EVOLUTIONARY GENOMICS: SYNTENYMAPPER

Discovery of new genomic features, elements or regulatory relations always comes with the hard task of interpretation. Distinguishing properties with functional value from others is often not easy, but aided by comparison with similar species. Evolutionary genomics is the science of retracing the development of multiple species' genomes and their respective elements. This part presents a novel method, *SyntenyMapper*, for the identification of small genome rearrangements between species pairs.

9

COMPARING GENOMES ON BASIS OF SYNTENY

Comparative genomics is a vast field, yet the most basic step is to find conserved elements (genes, pseudogenes, repeats, regulatory sequences) and regions in the genome. Establishing such equivalent genome regions is a pre-requisite to tracing the evolutionary processes that shaped contemporary genome sequences from a common ancestor. It is also central for the comparison of position-specific functional, structural and evolutionary features measured by modern high-throughput techniques, such as transcription factor binding sites, chromatin accessibility or SNPs. Since evolutionary conservation preserves functionality, researchers can draw conclusions on the biological importance of a feature from its state of conservation.

Existing methods focus on either large-scale or small-scale conserved regions

Identifying equivalent genome regions usually means detecting so-called synteny regions, which represent the longest sequence regions in two genome that share a common evolutionary origin. Usually, these synteny regions contain many conserved segments that are often disrupted by short regions of lower or no similarity [160]. Due to an unknown number of rearrangement events that occurred after the species diverged from the last common ancestor (species separation), the order of these equivalent synteny regions is different in both genomes. Rearrangements are usually (somewhat arbitrarily) divided into two classes based on their size: i) macro-rearrangements, which represent multi-megabase sized intra- and inter-chromosomal relocations of large synteny blocks, and ii) micro-rearrangements [160], or relocations of smaller segments (below 1 Mb) within a synteny region.

This hierarchical structure of evolutionary processes, where large synteny regions are on the one end of the scale and small gene rearrangements on the other, is rarely considered in comparative genomics methods. In addition, each of these tasks themselves are complex. Identification of long genome regions with a common ancestor can only be achieved when gaps are allowed and micro-rearrangements are ignored. Detection of small-scale orthologous regions on the other hand is made complicated by gene duplications and multi-domain proteins, which lead to local similarity hits. The following overview describes the general types of methods applied to the identification of equivalent genome regions (see Figure 38), though this list is not exhaustive.

SEQUENCE-BASED METHODS The most common approaches for the identification of synteny regions as described above are based on whole-genome sequence alignments. A popular example are synteny

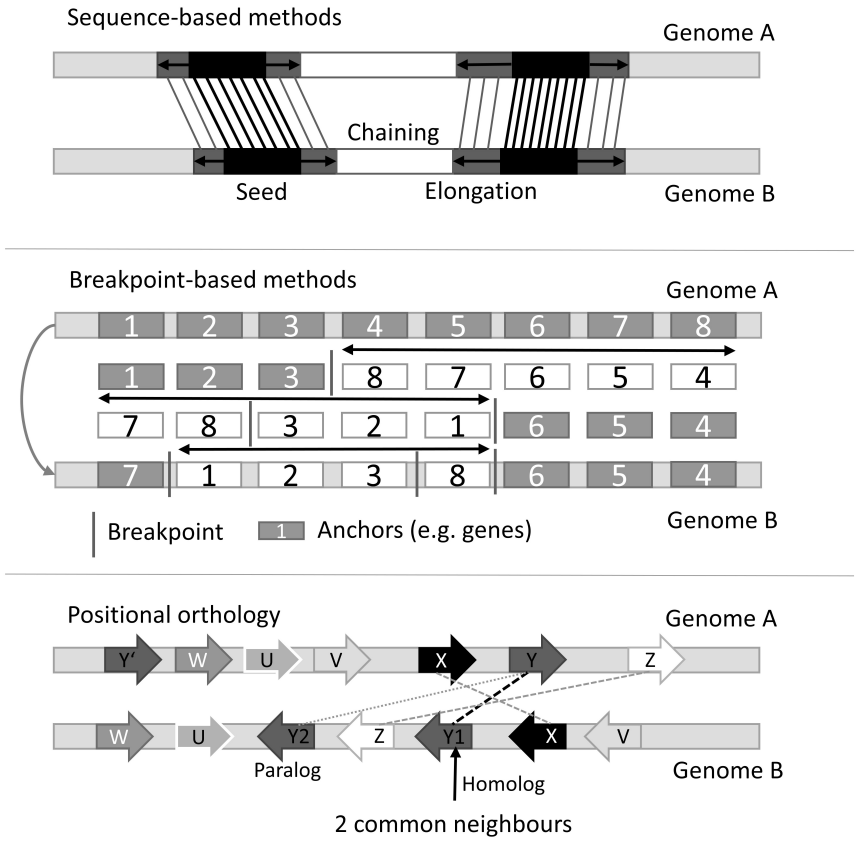


Figure 38: General overview of different approaches for identifying orthologous regions in two genomes. Sequence-based methods (e.g. ENSEMBL Compara) start with short local alignments that are extended to the longest possible alignments over gaps. Breakpoint-based methods use orthologous elements (called 'anchors') to find the minimum number of rearrangements that transforms one genome into the other. Positional orthology tries to distinguish orthologs from paralogs by analysing gene neighbourhoods.

regions from ENSEMBL Compara [56, 211], which extend a short local alignment seed until the similarity score falls below a pre-defined threshold. As a consequence, the process creates a set of medium-length alignments with a low number of gaps and mismatches. In a second elongation step, pairs of such alignments are chained if they are sufficiently close. This way, the procedure ensures that the longest possible regions with common ancestors are found.

Another similar approach based on unique 16-mers instead of raw sequence was published by Liao et al. [119]. Whole-genome alignments are very runtime expensive, so use of k-mers leads to a much faster search for synteny regions. Like ENSEMBL Compara, Liao et al.'s method aims to find the longest conserved regions, making it appropriate for analysing large-scale genome rearrangements, but not for comparisons at the level of genes.

A breakpoint *emerges in a sequence of conserved genes if two genes are neighbours in only one of the two species*

BREAKPOINT-BASED METHODS These methods [160] use breakpoints in the contemporary genomes to reconstruct their evolutionary history as a series of translocations, inversions and duplications. To achieve this, genomes are represented as sequences of homologous elements or anchors, such as genes. An example of a breakpoint-based method is the genome rearrangements Web server GRIMM [204]. Based on Hannenhalli and Pevzner's algorithm [79], it is built upon an initial alignment of orthologous elements and is able to distinguish macro- from micro-rearrangements. However, its focus is the reconstruction of rearrangement series that shaped the genomes, so it focuses on finding macro-rearrangements and the most important micro-rearrangements within them. Small rearrangements are discarded. As a consequence, it is a well-suited method for the identification of macro-rearrangements, but fails to completely reconstruct evolutionary history on the level of genes. Additionally, GRIMM is not able to deal with gene duplications, also called 'word problems' [160].

POSITIONAL ORTHOLOGY Positional orthology can be considered an umbrella term for all ortholog prediction methods that consider on gene neighbourhood for their prediction. These methods, such as localSynteny [96] or MSOAR [58, 186] aim to produce a one-to-one mapping of equivalent genes, while most other methods create many-to-many ortholog groups due to gene duplications. Positional orthology takes the direct neighbours of orthologs into account and, drawing conclusions on the evolutionary time point at which a certain ortholog was created, can map it to its equivalent in the other genome. The resulting pairs have a higher probability to fulfil similar functions in their species than paralogs which could have acquired a new function. The main disadvantage of these methods is that they consider only the neighbourhood composed of the adjacent genes on

both sides of a gene in question. Consideration of larger neighbourhoods could enhance the prediction.

The above described general types of methods all concentrate on only one aspect of the evolutionary links between two species, even though they are highly inter-dependent. We believe that no single approach is able to adequately compare quantitative or qualitative properties along eukaryotic genomes. For example, comparative genomics studies that revolve around genes usually only compare orthologs, even though their immediate environment, composed of regulatory elements in intergenic regions or adjacent genes, can provide additional valuable information. As a consequence, comparison of isolated pairs of orthologs is not sufficient. An example are the Homeobox (Hox) genes, which are expressed in the order in which they lie on the chromosome [158, 220]. Another example are transcription factor binding sites in the intergenic regions or domain-like features such as the previously described LADs, which determine the subnuclear position of the region. Such features are known to correlate with hetero- and euchromatin (see part ii, [69]) and thus can also influence gene expression. Together this stresses the importance of the linear environment in comparative genomics.

It appears as if synteny regions, which were originally defined as regions of conserved gene order in two species, would serve as a perfect basis for the comparison of genes and their neighbourhoods. However, the term has evolved, and since there are only few regions of continuous similarity between mammalian genomes, it now allows for many gaps and mismatches to capture the longest possible regions that derived from a single ancestor sequence. Often, the more general term synteny blocks is often used [160] to describe conserved regions that are interrupted by local micro-rearrangements. Naturally, these regions formed by macro-rearrangements are not suitable for a comparison of genes and their environments. Instead, detection of maximal length blocks of conserved gene order, also termed collinear blocks, is necessary.

We have created a new method, SyntenyMapper, which has already appeared in a publication listed on page vii, that aims to combine both the approaches focused on positional orthology and those that concentrate on detection of macro-rearrangements. Respecting the hierarchical structure of the genome, our method uses pre-calculated synteny regions and orthologous genes to find rearranged regions of conserved gene order within the synteny blocks. Not only does it reconcile macro- and micro-rearrangements this way, it also allows for consideration of genomic properties of orthologous gene neighbourhoods, making it well-suited for a gene-based genome comparison.

SyntenyMapper can be best compared with a class of orthology-based tools developed for the detection of collinear blocks, though these usually work on a genome-wide scale. The most common meth-

ods in this class are Cyntenator [173], MCScanX [216] and i-ADHoRe [164], which use alignment techniques based on orthologous genes to identify regions of conserved gene orders. When applied to synteny regions instead of genomes, these methods can theoretically identify all collinear blocks, making them comparable to our method. However, the genome-wide application of these methods often leads to disregarding of very small rearrangements. In contrast, SyntenyMapper aims to detect *all* micro-rearrangements within predefined synteny blocks, independent of the number of elements they contain and including those of single genes. Cyntenator, i-ADHoRe and MCScanX are less precise, allowing for gaps and mismatches of gene pairs, while our method defines blocks of perfectly conserved order that are ideal for comparison of closely related genomes (see section 11.4). Additionally, the complexity of the task is greatly reduced through the use of predefined synteny regions, leading to a very short runtime for a whole-genome comparison.

SyntenyMapper: finds all *micro-rearrangements in synteny regions & creates a one-to-one ortholog mapping*

In addition to the previously described advantages, SyntenyMapper implements a preprocessing step in order to create a set of syntenic one-to-one orthologs. Therefore, it can be compared to positional orthology methods to some extent. In contrast to these, SyntenyMapper relies on known orthology relationships and then filters many-to-many groups within them using additional information on gene order. The conservation of gene order in a larger segment is thus the main factor for the one-to-one orthology mapping, which is superior to methods only considering direct neighbourhoods.

We have made SyntenyMapper available as stand-alone command line tool on our website[1]. More importantly, we wanted to make it accessible to biologists with little experience in computer science, and have included it into the Galaxy [62, 63, 17] platform as a software repository[2] (repository name 'synteny_mapper'). For a more detailed description of the Galaxy framework see section 10.4.1.

We have applied SyntenyMapper to 25 eukaryotic species pairs for a general analysis of factors driving sequence rearrangements, and the pre-computed results as well as input data from the ENSEMBL Compara database can also be accessed and downloaded from our website.

The following sections describe the SyntenyMapper method in detail, followed by biological applications that show its value for comparative genomics.

1 http://webclu.bio.wzw.tum.de/syntenymapper
2 https://toolshed.g2.bx.psu.edu

10 MATERIAL AND METHODS

10.1 SYNTENYMAPPER

SyntenyMapper is a tool for the refinement of large conserved genome blocks through identification of micro-rearrangements. As input it takes a set of synteny regions, such as ENSEMBL, and orthologous gene pairs.

(ENSEMBL) SYNTENY REGIONS are defined as long genome regions in two species that have evolved from the same sequence in the last common ancestor. They mirror so-called macro-rearrangements, i.e. movements of large genomic blocks to another genomic location that happened in one organism after species separation. For closely related species, the genome structure of one can be reconstructed from the other by re-organization of these synteny regions (e.g. section 2.4).

The challenge in the detection of these regions lies within the correct determination of the borders. Synteny regions could be disrupted in one genome by short or long sequence stretches that have no equivalent in the other species. Multiple methods have been proposed so far (for a detailed description see section 9), and all of them provide suitable input for SyntenyMapper.

ORTHOLOGOUS GENES are pairs of genes in different species that have evolved from one common ancestor. Tools for identification of orthology rely on sequence identity in the simplest cases, but more complex and correct approaches are also available.

Orthology relations are not necessarily pairwise. If there was one or more gene duplication in one organism after species separation, multiple so-called paralogs of this gene exist, and many or all of them might be detected as orthologs of the corresponding single gene in the second organism. This relationship is termed a *one-to-many orthology group*. Similarly, gene duplication before species separation or independent duplication of genes in both species can lead to a *many-to-many* orthology group, where there exist pairwise orthology relations between all pairs of genes.

10.1.1 *Transforming orthology groups into one-to-one orthology pairs*

These complex relationships make it hard to identify gene movements in a genome. It is not clear which genes correspond to each other

in the two species, and due to independent duplication events after species separation there are many cases of genes that have an ortholog partner which is not their evolutionary equivalent. For example, if a gene a_1 in genome A is the ortholog to gene b_1 in genome B, but gets duplicated into gene a'_1, there will be an orthology relationship between a'_1 and b_1, but they will not be equivalents because a'_1 was created only after species separation.

SyntenyMapper thus takes an approach that transforms complex orthology groups into one-to-one pairs of equivalent genes. While it solves this problem during runtime for a maximum efficiency, it can be considered a pre-processing step and is the same for all synteny regions:

A. All genes that lie within the synteny region are grouped into one-to-one, one-to-many and many-to-many ortholog groups

B. Each one-to-many group is reduced to a single one-to-one ortholog pair with the highest sequence identity

C. Asymmetric many-to-many groups (those containing n genes in the genome A and m genes in the genome B, with $n \neq m$) are converted to symmetric many-to-many groups.
If $n = m + \delta$, exactly δ genes will be removed from the genome A based on an ascending ranking according to the average percent sequence identity of each gene to all other genes in the group.

D. Many-to-many groups, all of which are now symmetric, are split into individual one-to-one orthologous pairs. For any many-to-many group consisting of n genes $\{a_1, a_2, ..., a_n\}$ in genome A and m genes $\{b_1, b_2, ..., b_m\}$ in genome B, $n = m$, this is achieved by considering only orthology relationships between the genes with the same sequential number (i.e. a_1 with b_1, a_2 with b_2, etc.). Sequential numbers are given depending on the direction of the synteny region.

Genes without orthologs and the shorter of two overlapping genes are excluded from further analysis. For resolving one-to-many relationships, SyntenyMapper assumes that the corresponding genes share the highest sequence similarity. This is not necessarily true, as after gene duplication the copy might assume the role of the original, which is then under less selective pressure and can accumulate more sequence changes. In such a case it is not possible to detect the original gene, and SyntenyMapper's assumption thus is only an arbitrary solution.

To convert symmetric many-to-many groups into one-to-one pairs, we assume that the order of corresponding genes is the same in both species. This disregards micro-rearrangements of paralogs, but is a

good approximation for the subsequent analysis, as it is not possible to clearly identify corresponding genes.

An illustration of the effects of this pre-processing is shown in Figure 40a, where two synteny regions with their harboured genes and their respective pairwise orthology relationships are shown, obtained from an external source like ENSEMBL. Genes without orthologs (e.g. gene a_2) are excluded from consideration. A one-to-many group is formed by gene a_{10} from genome A and its two orthologs in genome B. The genes a_4, a_8 and a_9 form an asymmetric many-to-many group together with genes b_3 and b_5. All remaining genes are already part of a one-to-one relationship.

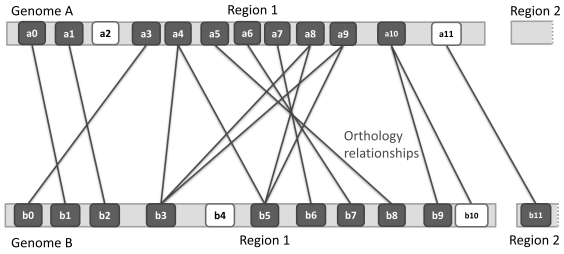
After conducting the pre-processing steps described above, the example synteny regions look as depicted in Figure 40b. Gene a_2 has been excluded and gene a_10 now only has the single ortholog with the highest sequence identity (gene b_9). The asymmetric many-to-many group has first been converted to a symmetric one by removal of gene a_9, and was then split into ortholog pairs according to gene order, resulting in one-to-one orthologous gene pairs $b_3 - a_4$ and $b_5 - a_8$.

After pre-processing, SyntenyMapper uses the resulting set of one-to-one orthologs to identify two types of evolutionary events: translocation and inversion of gene order. *Translocations* are caused by genome regions that break off from their original position and reinsert at another location in one species, causing an apparent disruption of gene order compared to the second organism. During an *inversion*, this region reverses its direction before reinsertion.

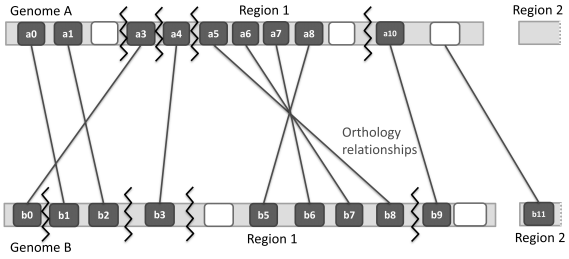
Every time a gene or a set of genes moves around the genome, it leads to cases of unconserved gene neighbourhoods, also called *breakpoints*. In the following section we define genome A to be the reference genome:

A breakpoint $(a_i/b_j, a_{i+1}/b_l)$ is defined by two orthologous gene pairs a_i, b_j and a_{i+1}, b_l if $j \pm 1 \neq l$, where genes a_i, a_{i+1} are from the reference genome A and genes b_j, b_l are from genome B. This can be interpreted as two neighbouring genes from the reference genome whose orthologs are not neighbours.

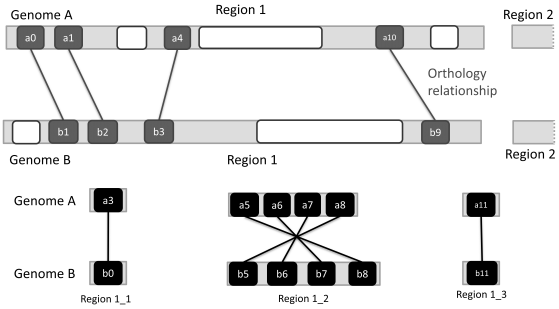
When a rearranging genomic region reinserts into genome B at a new position, *two breakpoints* emerge in A, one at each end, as illustrated in Figure 41a. Consequently, the genes between two subsequent breakpoints in A and their orthologs in B either form a so-called *rearrangement block*, i.e. a set of genes that lie either within a translocated or inversed segment, or belong to a non-rearranged part of the original synteny region that is enclosed between two blocks. To distinguish between these two cases, SyntenyMapper compares the length of subsequent blocks. To achieve this it iterates over breakpoints and looks up the preceding and following breakpoints. If no such points exist, start or end positions of the synteny region are taken as reference points. The identified three breakpoints define two gene segments



(a) Illustration of a synteny region between two species, with numbered boxes representing genes and connecting lines representing orthology relationships. Gene a_{11} and gene b_{11} have no orthologs in their synteny regions, but are orthologous to each other. Genes a_2 and b_4 have no orthologs.



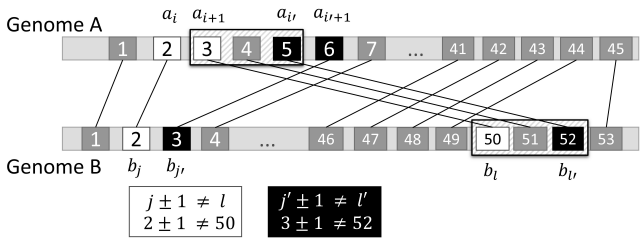
(b) During pre-processing, one-to-many (genes a_{10} and b_9, b_{10}) and asymmetric many-to-many (genes a_4, a_8, a_9 and b_3, b_5), groups are first converted into symmetric groups by excluding genes with the lowest sequence identity to the rest of the group (gene a_9), and subsequently paired as one-to-one orthologs based on gene order. Breakpoints (zig-zag lines) are identified as described in the methods section.



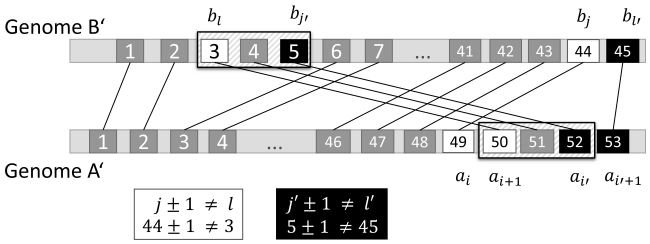
(c) Using breakpoints, SyntenyMapper defines rearranged segments, shown in black, as new synteny regions 1.1 to 1.3 within the long original region.

Figure 39: Illustration of SyntenyMapper pre-processing and result for an example synteny region.

that are adjacent in genome A but not in genome B. The shorter of both is defined as a new rearrangement block and synteny region, which lies within the longer original region.



(a) Illustration of two breakpoints emerging at both ends of a translocated segment a_3, a_4, a_5 in genome A and b_{50}, b_{51}, b_{52} in genome B (hatched box). By definition a breakpoint is constituted by two orthologous gene pairs a_i/b_j and a_{i+1}/b_l if $j \pm l \neq 1$, as shown in the boxes underneath the schema. The second breakpoint is described by a'_i/b'_j and a'_{i+1}/b'_l. White and black boxes mark the four genes forming the first $(a_2/b_2, a_3/b_{50})$ and the second $(a_5/b_{52}, a_6/b_3)$ breakpoint, respectively. A is used as reference genome to define the block formed by a micro-rearrangement as the genes that lie between the adjacent breakpoints in A, in this case a_2, a_3 and a_5, a_6. The genes between these two breakpoints and their orthologs in B form a block.



(b) Effect of reversing the reference genome used in the example shown above. The reference genome used here for the definition of breakpoints is A' (=B in a)). The detected breakpoints are $(a_{49}/b_{44}, a_{50}/b_3)$ (white boxes) and $(a_{52}/b_5, a_{53}/b_{45})$ (black boxes). Based on the adjacent breakpoints in the new reference genome A', the same translocated segment (hatched box) is detected as above.

Figure 40: Illustration of index-based breakpoint detection with different reference genomes.

The choice of a reference genome is arbitrary, as each translocated segment in B with respect to A is also a translocated segment in A with respect to B (see Figure 41b).

SyntenyMapper not only identifies rearrangements within a single synteny region (termed 'internal rearrangements'), but also between two regions ('external rearrangements'). To achieve this, all genes

without orthologs ('orphan genes') in their own synteny regions are collected during the first step of identifying internal rearrangements. In Figure 40b, gene a_{11} in the synteny region 1 of genome A and gene b_{11} in region 2, genome B, do not have orthologs in their own synteny regions, but are orthologous to each other. After all synteny regions are handled, the set of orphan genes are searched for pairs of orthologs, giving priorities to those with maximum sequence similarity in the case of multiple orthologs for one gene. SyntenyMapper then tries to elongate the external rearrangement block by checking if neighbouring orphan genes exist in both species and if they are orthologous as well. Through this the method is able to identify longer translocations between synteny regions.

Figure 40c shows the refinement of synteny regions achieved by SyntenyMapper. As a result of pre-processing, each gene has a single syntenic orthologous partner, and inversed and translocated blocks are identified and redefined as new synteny regions. Genes or groups of genes that were excluded from their synteny region during one of the prior steps, e.g. due to missing orthologs or external rearrangements, are excluded and do not appear in the output. The method generates two output files:

A. Coordinates of all synteny regions, both *original* and newly defined (*internal*/*external*)

B. Syntenic orthologs for each synteny region and their coordinates

The pseudocode in algorithm 1 on page 143 describes the main method of SyntenyMapper.

10.1.2 *Mapping of intergenic regions*

SyntenyMapper is also able to map intergenic regions that can be considered equivalent between two genomes. This application is not available in the published software tool, which focuses on identification of conserved gene order blocks. Mainly developed for our own purpose, this complete mapping of genes and intergenic regions is supposed to be the basis of feature comparison between the genomes of mouse and human already mentioned in section 3.2.

As SyntenyMapper identifies blocks of conserved gene regions, the main idea behind this mapping is that the intergenic region between two genes in such a block can be considered equivalent in both genomes. This does not imply that there is sequence conservation in this region. Instead, intergenic regions might harbour specific regulatory elements or have a specific GC content to allow for more or less flexibility of the DNA at this position. These features often influence nearby genes and, if functionally relevant, need to be conserved

input : ENSEMBL Synteny regions SR, one-to-one orthologs
output : Maximum length blocks of conserved gene order
function j(i): *returns index of gene i's ortholog in genome B*

for *all syntenic regions* $sr \in SR$ **do**
  I: *List of genes in sr, genome A, ordered by location, after pre-processing*
  for *all genes* $i \in I$ **do**
  if $j(i-1) \pm 1 \neq j(i)$ **then**
    // $(a_{(i-1)}/b_{j(i-1)}, a_i/b_{j(i)})$ is breakpoint
    // Find following breakpoint $(a_{i'}/b_{j(i')}, a_{i'+1}/b_{j(i'+1)})$:
    find i' *where* $j(i') \pm 1 \neq j(i'+1)$;
    // Find preceding breakpoint
      $(a_{i''}/b_{j(i'')}, a_{i''-1}/b_{j(i''-1)})$:
    $i'' :=$ *previous breakpoint*;
    // Three breakpoints enclose two blocks
    if $||(i'-1)-i|| > ||(i-1)-i''||$ **then return**
    new block: $(i'' \rightarrow (i-1))$;
    ; // first block is shorter
    else return *new block:* $(i \rightarrow (i'-1))$;
    ; // second block is shorter
  end
  ;
end

Algorithm 1 : Pseudocode of SyntenyMapper's main method. Special cases like inversions or overlaps are not described for the sake of simplicity.

if the genomic structure of the gene location is also conserved. SyntenyMapper thus maps intergenic regions between conserved genes as well, providing a basis for the TrackMapper tool described in section 10.3.

This mapping is straightforward in general, however, sometimes the presence of genes without orthologs or duplicates excluded in a previous step make it more difficult. In such cases, the number of intergenic regions between the orthologous genes in conserved order is different in both species. We applied a simple method to solve this problem, splitting the single intergenic region on one genome into equal sized regions of the same number as there are in the second genome. If the number of intergenic regions to be mapped is different yet higher than 1 in both species, we used a greedy approach and mapped as many intergenic regions pairwise as possible following the order along the chromosomes. For the remaining regions we performed the same splitting procedure as above, mapping the larger number of intergenic regions from one species to the last such region in the other.

10.1.3 *Implementation*

SyntenyMapper is implemented in Java and integrated in the Galaxy platform [62, 63, 17] as repository 'synteny_mapper', for easy use and accessibility. It provides the option to download synteny regions and orthologs directly from ENSEMBL or upload own data in a specific format. Additionally, pre-computed results and input data for ENSEMBL Compara synteny regions can be downloaded from our website[1]. SyntenyMapper computes micro-rearrangements fast, since each region is treated separately and no all-vs-all comparison for two whole genomes is necessary. It is able to analyse the human and mouse genomes in under one minute on a standard Linux workstation.

Additionally included in the 'synteny_mapper' repository is a Circos-based [110] tool for visualization and *TrackMapper*, a tool for comparison of UCSC-style feature tracks for two species on basis of their syntenic orthologs.

10.2 CIRCOS VISUALIZATION

Circos [110] is a visualization software that is suited well for genomic representations. It relies on the circular arrangement of elements, termed *ideograms*, which can be either genomic regions, chromosomes, or other. In addition to that, it can add feature plots, e.g. a graph that shows gene density, or rectangular elements such as transcription factor binding sites or genes. The most striking feature that

1 http://webclu.bio.wzw.tum.de/syntenymapper

uses the circular arrangement is the possible addition of links or ribbons, which can connect any point in one genomic region or ideogram to another.

These features make it a good visualization tool for our purposes. We use Circos to visualize a synteny region in two genomes as ideograms, add genes as feature elements and connect syntenic orthologs. To easily identify rearrangements within a region, we apply different colors to in-order ortholog pairs and rearranged genes. As a result, the internal rearrangements are silhouetted against the remainder of the synteny region.

If there is an external rearrangement in the region, our Circos adaptation also includes the second synteny region in a quarter of the circle, and adds links between externally translocated genes. All of this is achieved by feeding Circos a prepared configuration file that is adapted for each synteny region based on SyntenyMapper's output.

If the users prefers a linear visualization over Circos plots, our Galaxy tool also provides this option. We have implemented this second graph with R [166], again showing genes as coloured boxes with lines representing orthology, and rearrangements marked as different colors. Both Circos and R are automatically downloaded and installed upon installation of the SyntenyMapper Galaxy tool.

10.3 TRACKMAPPER

TrackMapper provides a direct application of SyntenyMapper's results, allowing the user to directly compare so called feature tracks from two species. A feature track can be provided in a BED format file, containing a set of genomic elements and their genomic positions (chromosome number, start and end coordinates) as well as an optional score in columns 1-4, respectively. Genomic features can be anything from regulatory elements such as transcription factor binding sites, in which case the score could refer to the binding strength, to sequence-inherent features like repeats.

Using SyntenyMapper's one-to-one gene mapping, TrackMapper can compare the overlap of the given feature with each gene, and compare this number between species. It does this by calculating average coverage, i.e. the percentage of the gene's base pairs overlapping with the feature (e.g. a LINE repeat) if no signal weight in the form of a score is given, or the average value of the feature if a signal weight is given (e.g. values representing each base pair's relative time of replication).

TrackMapper normalizes the resulting list or vector of average coverage values for each gene by conversion into z scores, through subtracting the mean and dividing by standard deviation. Let z_A and z_B be the vectors of feature z scores for syntenic one-to-one orthologs found by SyntenyMapper in genomes A and B, respectively. The mea-

sure of similarity between any two feature tracks, calculated as $z = |z_A - z_B|$, can be downloaded or further analysed with other Galaxy Tools, e.g. for the plotting of histograms. Additionally, TrackMapper provides the vector mean as a compact similarity measure between the two tracks over all genes.

Currently, a widely used tool for mapping of tracks between species is LiftOver [83], which was not designed for this purpose and is also asymmetrical, in that it converts a feature track from one reference species to the other. By contrast, TrackMapper is able to directly compare feature tracks from two species on the gene level without defining one genome as a source and the other as a target.

10.4 INTEGRATION OF SYNTENYMAPPER INTO GALAXY

10.4.1 *The Galaxy Platform*

Galaxy is an (online) platform and framework for Bioinformatic tools aimed at easy use and re-use. The user does not have to install the platform to use common tools (e.g. liftOver, filtering, many NGS applications) through well-known HTML forms. For more specific software that has been integrated into the platform or for more performance-demanding tasks, a local Galaxy installation can be set up easily. Galaxy does not handle data in form of flat-files, but rather as history objects. The user can upload a data object from a file on his computer or directly load it from UCSC through Galaxy's own interface. The history window that is always at the right margin of the browser window contains all data the user has loaded, as well as all tools he has run on it. It is thus easy to recreate results from other groups by rerunning their shared history objects, often combined into a workflow.

Sharing is a very important concept within Galaxy, since users are encouraged to share their experiments/workflows and data with the community, making an effort in making Bioinformatic analyses more reproducible than they are now. The users do not have to understand every single step to be able to use a workflow on their own data, since all they have to do is change the input.

Galaxy lets the user work with very large datasets, since there is a lot of computer power behind the project. Loading of large data and running of complex analyses takes its time, but the user can continue working on other things while the history objects turn from 'running' (marked yellow) to 'done' (green) or 'error' (red). Every output and input file can be viewed in the browser (see Figure 41) and downloaded.

From the developer's side, Galaxy makes it fairly easy to integrate a new tool into the platform. It demands a command line-ran tool and an XML-file that contains all information on input, output, files

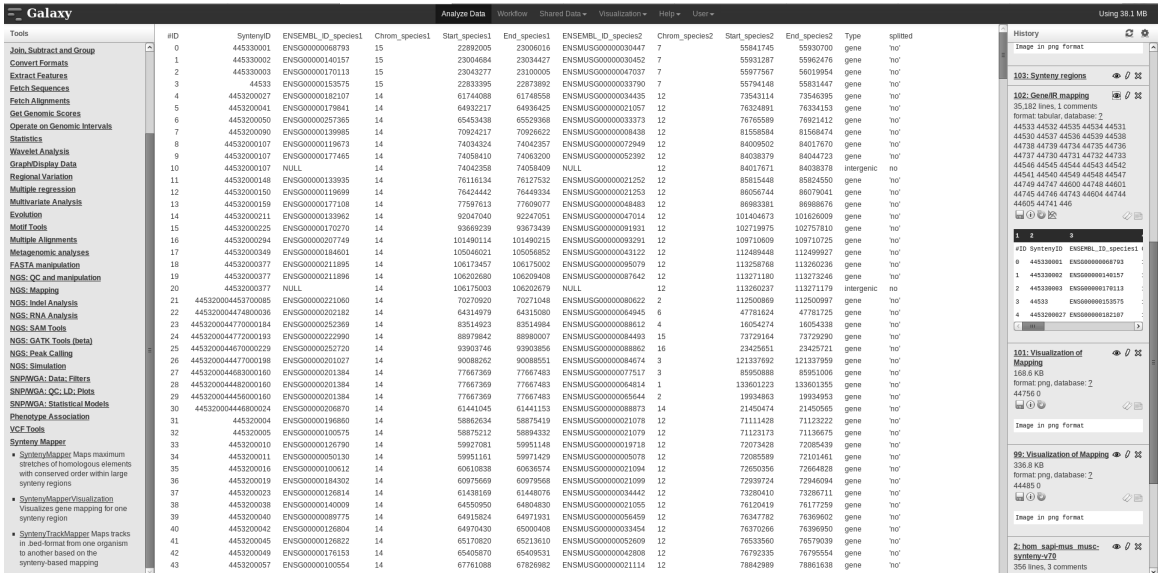


Figure 41: The Galaxy Browser. The panel on the left contains available tools, the panel on the right contains history objects. In the main panel, data from one of these objects are displayed.

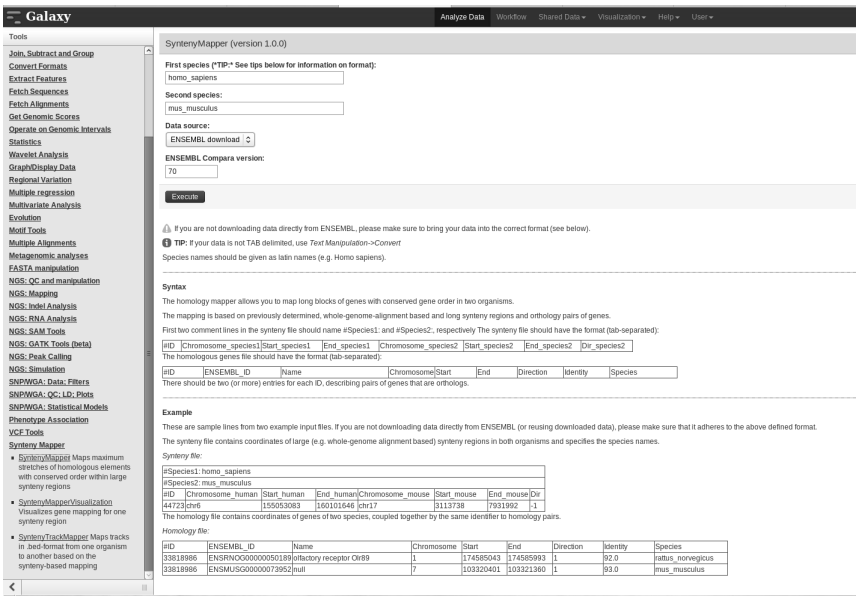


Figure 42: The SyntenyMapper tool integrated into Galaxy. In the tool panel (left), the package containing its three sub tools is listed. In the main panel, input options and detailed help information is given. The history panel is not shown.

for testing and help. The latter is again very important, since Galaxy targets biologists with poor knowledge of computer science, so it is vital to explain well to the audience what the tool does and how it has to be used. The XML format itself is very powerful and reaches from simple text box-style input fields to conditional fields and input drop down lists that are read from a column of an input file. Unfortunately, the documentation relies mainly on examples, making it often hard to understand how to achieve a certain result.

Before upload into the so-called tool shed that hosts all software developed for Galaxy, the user has to download a local version, which works identically to the web server. Here one can try and test all tools, going on to testing them in the online test toolshed before finally publishing them in the official tool shed, where users can access them. In order for your software to become tested, you have to supply test files and corresponding outputs, so that the platform can assess whether your tool works correctly.

Altogether Galaxy is an online framework for easy access to tools that enables (mostly) easy integration of newly developed tools. SyntenyMapper is Java-based and integrated into Galaxy as a command line runnable .jar package. Input and output files are defined as follows:

Input (Figure 42)

- Latin name of the first species

- Latin name of the second species
- Option 1: Download input files from ENSEMBL directly
  - Choose ENSEMBL version (default: current version 70)
- Option 2: Use already downloaded input files from the history panel
  - File containing ENSEMBL synteny regions in a certain syntax, tabular
  - File containing ENSEMBL orthologous genes in a certain syntax, tabular

Option 2 requires files in a very specific format and is aimed at users who have already downloaded the input files in a previous step and want to reuse them. For more advanced users who want to use their own synteny regions and orthologs, a detailed definition of the files' syntax is given in the help information below. If the user wants to load the data from ENSEMBL, it is automatically put into the correct format and will directly be used by the tool itself.

Output

- File containing redefined synteny regions, tabular
- File containing the one-to-one gene mapping, tabular
- When Option 1 was chosen:
  - File containing ENSEMBL synteny regions, tabular
  - File containing ENSEMBL orthologous genes, tabular

The first output file contains all ENSEMBL synteny regions in a new format, adding those that were newly defined during the tool's runtime and a status that distinguishes them from original regions. The second output file contains the one-to-one mapping of genes, listing a reference identifier to the corresponding synteny region, name, chromosome, start and end for both organisms, direction in the second organism.).

Output files 3 and 4 are only generated if download from ENSEMBL was chosen and can be used to rerun the analysis very fast.

Performance

Overall runtime of the tool itself is very short and takes less than a minute on a standard Linux workstation. Only download of data from ENSEMBL takes longer, but this has to be done only once for each species pair. It is not uncommon for Galaxy that loading of data takes a long time, so this is not a real drawback.

10.4.2 *Visualization*

The SyntenyMapper tool package not only contains the mapping tool itself, but also the other earlier described tools. This section describes the make-up of the visualization tool in Galaxy, which installs its own Circos and R version upon installing. Usage is very easy after running of SyntenyMapper.

Input

- Output file 1 from SyntenyMapper
- Output file 2 from SyntenyMapper
- A chosen synteny region from a drop down list of available regions
- Linear version check box

Output

- A PDF graphic of the visualization

10.4.3 *TrackMapper*

Similar to the visualization tool, TrackMapper requires only the SyntenyMapper output and a BED file containing a feature track. The output is a tabular file identical to the input gene mapping with an additional column for the similarity measure and a comment line including the summarized measure.

10.5 ENSEMBL COMPARA TEST DATA

We performed a whole-genome mapping with SyntenyMapper on *H. sapiens* and *M. musculus* (assemblies hg19 and mm10, respectively), using data on synteny regions and orthologs from ENSEMBL Compara [56, 211] version 73. A total of 356 synteny regions with a mean length of 7.63 Mb and 6.89 Mb in human and mouse, respectively, were obtained. The complete set of ENSEMBL protein-coding genes, containing 23,618 and 22,796 unique genes in human and mouse with mean lengths of 59.8 kb and 44.3 kb, respectively, was used to map genomic coordinates to pairwise orthologs. ENSEMBL Compara provides 27,453 pairwise orthology relationships between protein-coding genes. In human, 55.10 genes on average lie within each synteny region, while mouse synteny regions contain an average of 58.40 genes.

10.6 GLOBAL COMPARISON OF 25 EUKARYOTIC SPECIES PAIRS

ENSEMBL Compara offers synteny regions only for a limited set of species pairs, most of them involving human. The complete list can be found in Table 19. We downloaded synteny regions and orthologs for all provided pairs, and performed a general comparison of these eukaryotic species.

Table 19: List of species pairs for which ENSEMBL Compara synteny regions are available.

C. familiaris (Dog)	*E. caballus* (Horse)
G. gallus (Chicken)	*A. carolinensis* (Lizard) *M. gallopavo* (Wild turkey)
H. sapiens	*B. taurus* (Cow) *C. jacchus* (Marmoset) *C. familiaris* (Dog) *E. caballus* (Horse) *F. catus* (Cat) *G. gallus* (Chicken) *G. gorilla* (Gorilla) *M. macaca* (Macaque) *M. domestica* (Opossum) *M. musculus* (Mouse) *O. anatinus* (Platypus) *O. cuniculus* (Rabbit) *P. troglodytes* (Chimp) *P. abelii* (Orang-Utan) *R. norvegicus* (Rat) *S. scrofa* (Pig)
M. musculus (Mouse)	*B. taurus* (Cow) *G. gallus* (Chicken) *C. familiaris* (Dog) *O. anatinus* (Platypus) *S. scrofa* (Pig) *R. norvegicus* (Rat)

10.6.1 *Calculation of sequence similarity between synteny regions*

We calculated the sequence similarity of ENSEMBL Compara synteny regions with our own implementation of the sequence comparison algorithm proposed by Rieck and Laskov [171]. This trie-based algorithm is able to compute in linear runtime. A trie is a tree struc-

ture similar to a suffix tree, with the distinction that for a finite set of strings or words, every string is represented in the path from the root to one of the leaves, instead of the suffixes of a single string. Long sequences such as synteny regions are converted into finite sets of words by splitting into overlapping k-mers, in our case $k = 6$. These are then inserted into a new trie by iterating over each 6-mers' characters and, starting from the root of the trie, following the edge with the corresponding character to the next node. If no such edge exists, one or more new edges need to be inserted. This ensures that the building of the trie happens in a runtime of $O(k \times n)$, with n being the length of the sequence. The number of times a string exists in the sequence can be counted for each node during runtime.

The algorithm then makes use of the trie's maximum depth of k by performing a parallel depth-first search on the tries of the two sequences. A distance score is calculated for each node using the *inner function* m, which compares the number of occurrences of the string represented by that node in both sequences. The score is accumulated over the whole trie using the outer function $\oplus$. In our implementation, we used the sum $\sum$ as outer function $\oplus$ and the Manhattan distance as inner function m (Equation 19).

$$m = |\phi(x) - \phi(y)| \tag{19}$$

x, y nodes in the trie

$\phi(x)$ returns number of occurrences of the string that is readable from the root of the trie to node x in the sequence

10.7 COMPARISON OF SYNTENYMAPPER WITH CYNTENATOR, I-ADHORE AND MCSCANX

SyntenyMapper's main objective to refine regions obtained by macro-rearrangements through identification of conserved gene order blocks is unique. However, similar results can be achieved by applying software tools for the detection of collinear blocks, i.e. blocks of conserved gene order, to ENSEMBL syntenic regions instead of whole genomes. This way, they can identify micro-rearranged regions within the syntenic regions and with a similar computation speed to SyntenyMapper. We thus compared our results for human and mouse, assemblies hg19 and mm10, to those from CYNTENATOR [173], i-ADHoRe [164] and MCScanX [216].

We applied the three methods to synteny regions of human and mouse instead of the whole genomes, to achieve a comparability with SyntenyMapper results. In all methods, genomes are represented by ordered lists of genes, and (symmetrical) orthology information is given in a table, including sequence identities if possible. CYNTENA-

TOR was run with default parameters and the homology type option 'BLAST' to consider these sequence identities. The necessary guide tree was simple (*(human mouse)*), since it only had to include two species. i-ADHoRe contains a large amount of parameters that can be set individually. We used values suggested in the documentation, including gap size and cluster gap size of 15, probability cutoff 0.001, q-value 0.9 and three anchor points. We ran MCScanX_h with default parameters.

Unfortunately, there is no gold standard in this field of comparative genomics, and genome simulation is fairly complex. We thus performed a qualitative comparison of the three methods. As visual guideline for the interpretation, we used the Circos visualization of SyntenyMapper results. This tool draws all genes that are mapped to each region and a subset of the orthology relationships, both downloaded from ENSEMBL. SyntenyMapper results are mainly visualized through colouring. We thus believe that this visualization can help us identify reasons for discrepancies and incorrect assignments of all three methods, including SyntenyMapper itself. Additionally, we performed manual comparison of result files in certain cases.

Additionally to qualitative comparisons, we applied an approach recently suggested by Ghiurcuta and Moret [61] to measure the quality of the detected collinear blocks. In their publication they write up a formal definition of *syntenic blocks (SBs)* to enable measurement of concordance with this definition and, as such, quality of a set of syntenic regions. They require of each set of SBs in multiple genomes, or SB families (SBF) that they are connected by homology relationships between markers such as genes. Collinearity or conserved gene order is not required, but defined as well in subsequent definitions. Genes without orthologs are excluded from the quality assessment. For a detailed description of the formal definitions see the original publication.

Using these definitions, Ghiurcuta and Moret are able to suggest a number of measures that allow comparison between synteny detection methods and quality assessment. In the below described measures, markers are genes and those without any homologs are ignored.

- *SBFs* gives the number of SBs (two organisms) or SBFs (more than two organisms)
- *w/o homologs in the SBF* lists the number of SBFs that contain at least one marker without a homolog in the SBF, but elsewhere
- *Content overlap* lists the number of SBFs that contain at least one marker that also appears in another SBF
- *Selective content* gives the number of non-collinear blocks

- *Block incompleteness* is denoted as $\frac{E(X)}{E'(X)}$, with $E(X)$ being the number of markers lying in the SBF according to the used tool's output, and $E'(X)$ being the total number of markers lying in the SBF's region
- *Relaxed score* counts the markers in an SBF that have at least one homolog within the SBF, and divides it by total number of markers in the SBF

We calculated these measures for SyntenyMapper's, Cyntenator's and i-ADHoRe's results for the collinear block detection between human and mouse (assemblies hg19 and mm10, respectively).

10.8 MAPPING OF FEATURE COMPOSITION IN THE HUMAN AND MOUSE GENOME

We applied TrackMapper to the following genomic features in human and mouse (assemblies hg19 and mm10, respectively): LADs, LINE, LTR, SINE, Open chromatin, RTD and Single Nucleotide Polymorphism (SNP). For a detailed description of these features see section 2.1. To match the assemblies, we used liftOver to update the coordinates of the BED files before applying TrackMapper. Additionally, we created randomized versions of all 14 human BED files by shuffling the coordinates for each element, while maintaining the chromosome. Using TrackMapper, we calculated difference measures for all one-to-one orthologs as identified with SyntenyMapper, and for the inferred orthologous intergenic regions, and compared observed to randomized data with the Kolmogorov-Smirnov test [129] and the Rank sum test [126].

KOLMOGOROV-SMIRNOV TEST is a statistical method to test if two sample sets were drawn from the same distribution. It can be applied either to two variates or a single variate which is compared to a previously defined probability distribution. In our case, we tested if the distributions of difference measures in the observed and random data were significantly different with respect to their *shape*.

RANK SUM TEST or Wilcoxon-Mann-Whitney test also investigates the significance of two distributions' difference. It's null hypothesis is that the two distributions are the same. Contrary to the T-test it works well on non-normal distributions. While the Kolmogorov-Smirnov test compares two distributions' shape, the rank sum test compares their *location*.

11 RESULTS

11.1 SYNTENYMAPPER: A NEW TOOL FOR REFINING SYNTENIC ORTHOLOGS

The previous chapter describes the algorithm of SyntenyMapper and its associated tools. This section shows different applications of the method in the field of evolutionary genomics. The main task of SyntenyMapper is to identify so-called micro-rearrangements, *i.e.* small-scale translocations or inversions of genes or groups of genes. Other methods for the reconstruction of evolutionary history focus mainly on large-scale developments. Many of these methods are available so far and provide high-quality synteny regions, which mirror the macro-rearrangements that happened after species separation. SyntenyMapper uses these synteny regions and performs a refinement by detecting small-scale evolutionary events within and between them.

It was developed to use synteny regions from ENSEMBL Compara [56, 211], but can perform on any other method's synteny regions if they match the required format. SyntenyMapper complements the set of one-to-one orthologs from ENSEMBL by finding syntenic one-to-one orthologs among one-to-many/many-to-many orthologous groups. It subsequently uses these to identify deletions, inversions, local and distant translocations within ENSEMBL synteny regions, further refining the definition of these regions. As a result, SyntenyMapper provides the user with a set of evolutionary building blocks with completely conserved gene order between two species.

This insight into the small-scale evolutionary history of two genomes can enhance the general knowledge in the field of evolutionary genomics. Additionally, it can help to measure importance of the immediate gene environment. An example for the importance of gene neighbourhood are the Homeobox (Hox) genes, the expression of which is determined by their order on the chromosome [158, 220]. SyntenyMapper allows the user to identify blocks of conserved gene order, which he can then subject to GO enrichment analysis to detect similar cases of neighbouring sets of genes with related functions. If the order and/or neighbourhood of genes is vital for their function and expression profile, there is a high selection pressure against the disruption of this group of genes. There are also intergenic features that influence gene expression, such as LADs which can control the subnuclear localization of the entire genomic region. Using the blocks of conserved order identified by SyntenyMapper, scientists can compare the location of such elements in corresponding genomic regions

not only confined to the neighbourhood of a single gene or a very vague position within a synteny region.

Through the integration into the Galaxy platform the tool is easily accessible and can be used directly on data downloaded from ENSEMBL Compara or user-supplied orthologs and synteny regions. The resulting refined regions and their associated annotation tracks can be visualized using Circos [110] and analysed with our own track-comparison tool TrackMapper as described in the previous sections. In the following sections we demonstrate the merits of our method by applying it to pairs of eukaryotic genomes.

11.2 DETECTION OF MICRO-REARRANGEMENTS BETWEEN THE HUMAN AND MOUSE GENOMES

We applied SyntenyMapper to synteny regions and orthologs between the genomes of *H. sapiens* (hg19) and *M. musculus* (mm10), using data obtained from the ENSEMBL Compara database. Human and mouse, though not very closely related, share a similar genome. Since species separation the genome has been rearranged in both species, leading to a high number of synteny regions which are distributed differently in the two organisms' genomes. As a result of the high similarity within the regions, synteny region length is highly correlated between the species (Figure 43).

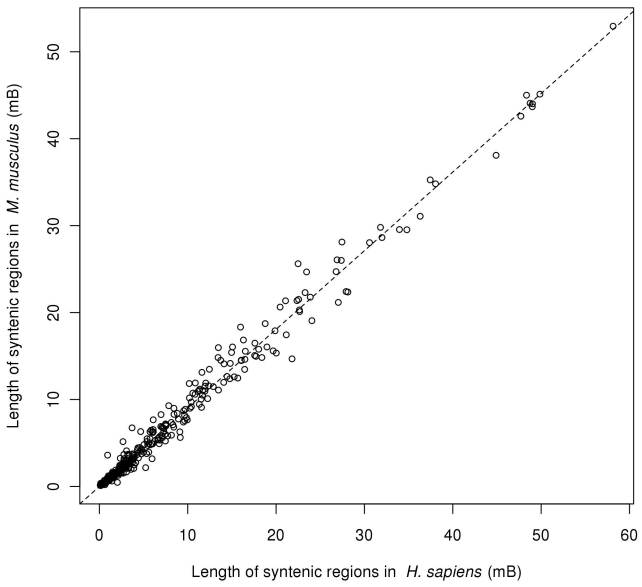


Figure 43: Comparison of syntenic region length in human and mouse. The regression line is shown as dotted line (Pearson correlation coefficient 0.9932).

Table 20 holds an overview of the frequency of different types of orthology relationships in human and mouse. The majority of protein-coding ENSEMBL genes (70.07%) have exactly one ortholog within their synteny regions, and are thus already one-to-one ortholog pairs. The second largest subset (12.30%) is constituted by the genes that do not have any orthologs and are therefore excluded by SyntenyMapper.

Human and mouse share 356 ENSEMBL synteny regions of similar length

Genes in many-to-many groups within a single synteny region make up only 2.30% of all genes, and most of these groups are asymmetric. Finally, genes that only have orthologs in other synteny regions (external orthologs) make up a mere 1.27% of the data. All other types of orthology relationships, such as genes with orthologs in regions not covered by synteny regions, represent less than 2% of all cases and are ignored by the current version of SyntenyMapper.

During the pre-processing step, SyntenyMapper converts (asymmetric) many-to-many ortholog groups into syntenic ortholog pairs. A total of 941 genes (2.1% of all human and mouse genes) are excluded due to this process as non-syntenic, and a total of 10,840 non-syntenic relationships from the original set of 27,453 ENSEMBL Compara orthologs are eliminated with them. SyntenyMapper is left with 16,613 syntenic ortholog pairs between human and mouse, which it uses to identify 2,898 new synteny regions. The resulting set of synteny regions comprises the 356 original ENSEMBL Compara synteny regions (10.94%) as well as the newly detected blocks of genes that were subject to micro-rearrangements within (2,817, 86.57%) or between them (81, 2.49%).

Internal micro-rearrangement are thus much more frequent than external ones or even the macro-rearrangements that formed the original synteny regions. However, the latter are of course significantly longer, containing 73.68% of all genes compared to blocks subject to internal rearrangements, which harbour only 1.52 genes on average. The order of genes in long synteny regions is thus mostly disturbed by very short blocks of one or two genes.

The majority of micro-rearrangements occur within synteny regions

Distant translocations of genes between different synteny regions are rare and, in the specific case of the human/mouse comparison presented here, almost always contain only a single gene. The reason for the difference in frequency of distant and internal rearrangements is probably the proximity itself. Within a single synteny regions, genes lie within a linear distance of 7 to 8 Mb to each other, making it easier for a genomic region that broke off to re-insert closeby. It is also known that spatial proximity is one of the triggers for rearrangements [125], which correlates with linear proximity [120].

Identification of these micro-rearrangements can greatly improve our understanding of evolutionary events shaping extant genomes and is instrumental for comparing the properties of regions of conserved gene order between two species. The following examples il-

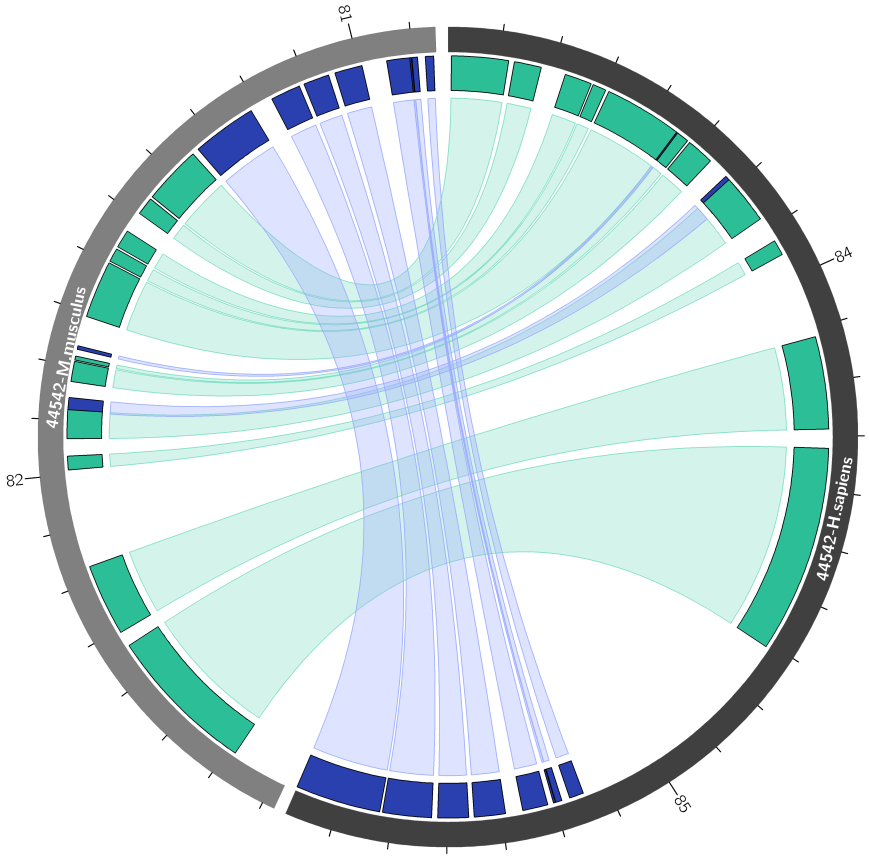
Table 20: Frequency of different cases of orthologous relationships for a given gene in a syntenic region, human vs. mouse. The first five cases are covered by SyntenyMapper, the remainder of cases make up less than 2% of all genes. *Internal orthologs:* Orthologous genes that lie in the same synteny region. *External orthologs:* Orthologous genes that lie in different synteny regions *Synteny-block-free region:* Genomic region that is not covered by ENSEMBL synteny region.

Type of orthology relationship	# Genes
No ortholog	5,705 (12.30%)
One internal ortholog	32,505 (70.07%)
Many internal orthologs	1,062 (2.29%)
One external ortholog	487 (1.05%)
Many external orthologs	105 (0.23%)
Many in- and external orthologs	204 (0.44%)
One ortholog in a synteny-block-free region	99 (0.21%)
Many orthologs in a synteny-block-free region	15 (0.03%)
Many orthologs: internal, external, and in synteny-block-free regions	166 (0.36%)

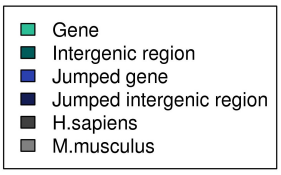
lustrate how SyntenyMapper complements the pre-calculated ENSEMBL synteny regions, which are created in such a way that their length is maximal, by identifying the previously ignored small rearrangements within them.

Figure 44 shows a medium sized synteny region that has the same orientation in human and mouse. Most of the genes within share one-to-one orthology relationships. However, the order of genes is obviously disrupted by the translocation of a large block of seven genes, which is located at the beginning of the region in human and at the end in mouse. The order and direction of genes within this block is preserved. Interestingly, there is a sizeable gap between syntenic orthologs and translocated genes on the human chromosome, which can imply that the translocation happened in *H. sapiens* after human-mouse divergence. Discoveries like this are informative on their own, since they shed light on the processes of evolutionary genomics within different organisms. Additionally, they can serve as the basis for functional analyses, considering not only single genes but their conserved neighbourhood as well.

This example illustrates how easy it is with SyntenyMapper and the Circos-based visualization tool to identify recent genome rearrangements that make any linear comparison impossible. The analysis is not restricted to translocations within one region: Figure 45 shows the relationship between two different synteny regions that harbour an external ortholog pair. These rare external translocations (1.27% of



(a)



(b)

Figure 44: Visualization of SyntenyMapper results for a syntenic region (ENSEMBL identifier 44542) in human (dark grey ideogram, right) and mouse (light grey ideogram, left). Ticks are placed at 100 kb distance and the numbers represent positions in Mb on the human and mouse chromosomes 15 and 7, respectively. The Circos circular plot illustrates the positions of genes/intergenic regions for one syntenic region in both species and the correspondence between them. Micro-rearrangements are illustrated by color-coding, with syntenic orthologs and out-of-order genes shown in light green and blue, respectively. A large block of seven genes (blue) was translocated in either human or mouse. In the Galaxy version of the plots, gene annotations are given as labels and as direct links to ENSEMBL through clicks onto the gene track.

all genes) almost always involve single genes in human and mouse. We hypothesize that translocations of single genes over large linear distances are favoured by short spatial distances between the chromosomal regions involved, as has been shown for cancer cells [125]. As a consequence, these events could help infer knowledge about the three-dimensional structure of the genome and the flexibility of the chromosome regions involved in the external rearrangement. All micro-rearrangements illustrated in this section were not annotated by ENSEMBL Compara and cannot be directly inferred from orthology relationships, stressing the value of SyntenyMapper for comparative genomics.

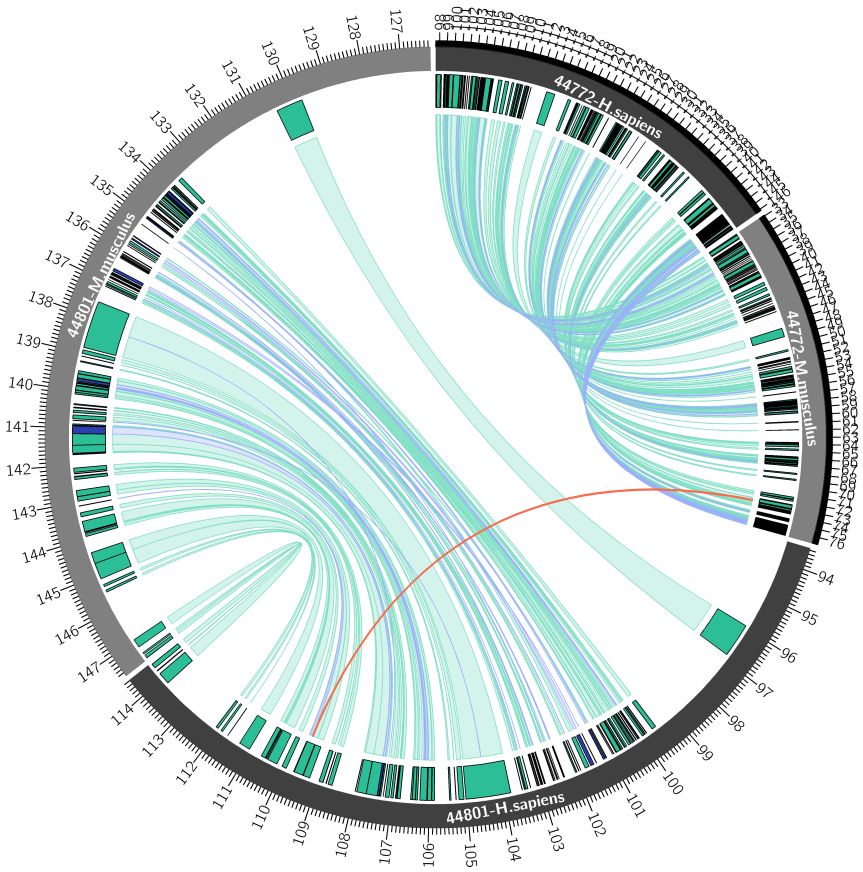


Figure 45: Translocation of a single gene from the human region 44801 to the mouse region 44598, shown with red line. Ticks are placed at 100 kb distance and the numbers show the positions in Mb on chromosomes X in human and mouse (region 44801) as well as on chromosomes 19 in human and 7 in mouse (region 44598). For colour legend see Figure 44 (b).

11.3 SYNTENYMAPPER GENOME COMPARISONS FOR 25 EUKARYOTIC SPECIES PAIRS

ENSEMBL Compara [56, 211] provides synteny regions and orthologs for 25 pairs of eukaryotic species, 16 of them involving human. Using SyntenyMapper, we conducted a large-scale analysis of micro-rearrangements for these organisms. We performed research on correlations between evolutionary distance and size and number of synteny regions, internal and external micro-rearranged regions (see Table 21).

Table 21: Statistics of pre-computed synteny mapping for ENSEMBL Compara (version 73).
SR: ENSEMBL Compara synteny regions.

		Regular SR		Internal SR		External SR	
Species pair	# Orthologs (total)	Nr.	Avg. gene Nr.	Nr.	Avg. gene Nr.	Nr.	Avg. gene Nr.
Dog - Horse	16,154	201	67.02	1,487	1.50	395	1.00
Chicken - Lizard	5,875	261	18.94	456	1.94	43	1.16
Chicken - Wild Turkey	11,948	114	88.40	1,031	1.65	130	1.28
Human - Cow	16,673	388	32.58	2,627	1.45	209	1.03
Human - Marmoset	16,663	266	48.67	2,822	1.19	358	1.01
Human - Dog	16,357	308	38.62	2,641	1.60	223	1.03
Human - Horse	16,360	246	50.78	2,586	1.45	121	1.03
Human - Cat	16,198	265	47.70	2,524	1.35	142	1.01
Human - Chicken	10,926	423	17.92	1,605	1.94	191	1.18
Human - Gorilla	17,582	84	173.99	2,689	1.03	195	1.0
Human - Macaque	16,911	219	62.98	2,520	1.11	327	1.02
Human - Opossum	12,280	496	19.09	2,113	1.25	173	1.01
Human - Mouse	16,613	356	34.38	2,817	1.52	81	1.03

Continued on next page

Statistics of synteny mapping for ENSEMBL Compara // *continued*

Species pair	# Orthologs (total)	Regular SR		Internal SR		External SR	
		Nr.	Avg. gene Nr.	Nr.	Avg. gene Nr.	Nr.	Avg. gene Nr.
Human - Platypus	1,910	210	7.24	293	1.30	7	1.14
Human - Rabbit	11,501	229	37.35	1,749	1.60	156	1.02
Human - Chimp	17,249	139	94.74	2,425	1.63	115	1.02
Human - Orang-Utan	16,778	150	91.36	2,617	1.13	113	1.02
Human - Rat	16,314	546	21.64	2,577	1.64	264	1.05
Human - Pig	16,314	357	25.33	3,655	1.44	115	1.08
Mouse - Cow	16,427	439	31.01	1,967	1.33	195	1.02
Mouse - Chicken	10,786	502	15.59	1,280	2.12	231	1.10
Mouse - Dog	16,060	364	35.53	1,977	1.44	285	1.0
Mouse - Platypus	1,860	235	6.63	223	1.31	9	1.0
Mouse - Pig	14,045	397	22.81	3,180	1.52	137	1.04
Mouse - Rat	18,741	554	26.74	2,205	1.49	616	1.03

To better understand relationships between evolutionary distance and evolutionary events, we did a general analysis on the data from ENSEMBL Compara. Among the most closely related species pairs in the set are human *vs* chimp (*P. troglodytes*) and mouse *vs* rat (*R. norvegicus*). Both of them share a high number of orthologs, however, the number of ENSEMBL synteny regions differs vastly: while human and chimp share only 139 synteny regions, there are 554 between mouse and rat. Because the average synteny region length between human and chimp is about four times the size of synteny regions in mouse and rat (22.0 Mb and 4.8 Mb, respectively), the percentage of the genomes covered is similar (human *vs* chimp: 87.76%, mouse *vs* rat: 94.64%).

Using branch lengths of the UCSC species tree [133] (available for all species pairs considered in this study except for those involving gorilla, pig, orang-utan, common marmoset and turkey) we analysed the correlation between the evolutionary distance and micro- and macro-rearrangement related genome features. In general, one would expect more closely related species to share a low number of very long synteny regions, and an increase in the number and decrease in the average length of regions with growing evolutionary distance. Indeed, for species pairs separated by short or medium evolutionary distance, this expectation yields true and the average synteny region length and their number exhibit a negative exponential correlation (Figure 46).

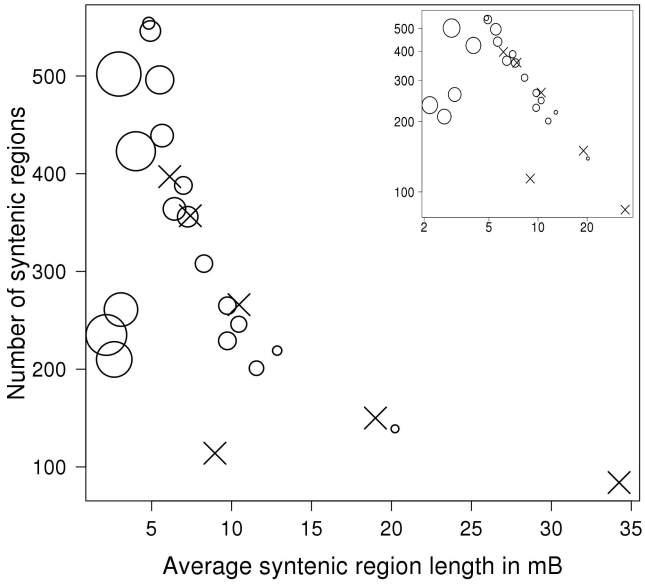


Figure 46: Dependence of synteny features on the average synteny region length (x axis) and evolutionary distance (circle size, inferred from branch lengths in Miller et al [133], calculated as the average number of substitutions per site).
Negative correlation between the number of synteny regions and their average length (Inset: logarithmic axis scales). Closely related species (small circles) tend to have fewer, longer synteny regions. Distant species (large circles) tend to have high numbers of very short syntenic regions. Crosses correspond to the species pairs with no distance information available.

Closely related species have few long synteny regions

However, more distantly related species (*e.g.* human *vs* platypus) share fewer synteny regions than would be expected based on this exponential correlation pattern. The deviation is caused by very low sequence similarity between species in the cases of human *vs* platypus, mouse *vs* platypus and lizard *vs* chicken. The overall dissimilarity of the genome caused by the large evolutionary distance leads to a large portion of the genome that is not covered by synteny regions

because the sequences have mutated too much. In the case of human *vs* platypus, only 21% of the genome is covered by synteny regions (see Table 22).

Table 22: Evolutionary distance and genome coverage by synteny regions for all species pairs.

Species pair	Evolutionary distance	Percentage of longer genome covered by ENSEMBL synteny regions
Dog - Horse	0.25	94.84%
Chicken - Lizard	0.91	53.98%
Chicken - Wild Turkey	-	94.48%
Human - Cow	0.36	88.66%
Human - Marmoset	-	89.82%
Human - Dog	0.35	89.29%
Human - Horse	0.30	89.21%
Human - Cat	0.35	89.34%
Human - Chicken	1.10	77.82%
Human - Gorilla	-	91.95%
Human - Macaque	0.07	90.77%
Human - Opossum	0.72	76.70%
Human - Mouse	0.46	87.52%
Human - Platypus	0.98	20.43%
Human - Rabbit	0.36	73.13%
Human - Chimp	0.02	85.35%
Human - Orang-Utan	-	83.73%
Human - Rat	0.46	87.11%
Human - Pig	-	87.85%
Mouse - Cow	0.53	88.42%
Mouse - Chicken	1.28	73.68%
Mouse - Dog	0.53	88.42%
Mouse - Platypus	1.16	19.91%
Mouse - Pig	-	87.08%
Mouse - Rat	0.16	93.58%

Internal *micro-rearrangement: Within a synteny region*
External *micro-rearrangement: Between two synteny regions*

There are two other distant species pairs (human *vs* chicken, mouse *vs* chicken) that do not deviate from the general trend as much, but still show a low genome synteny coverage of the respective longer genomes due to significant size differences (human genome 3.1 Gb, mouse genome 2.7 Gb, chicken genome 1.1 Gb).

SyntenyMapper is a gene-based approach and as such dependent on the number of orthologs, which is naturally higher for closely related species. Figure 48a shows that the number of internal micro-rearrangements per synteny region, ranging between 0 and 30, increases

with the increasing average synteny region length and decreasing evolutionary distance of pairs. The number of micro-rearrangements within a region thus depends on the size of this region, which is in turn correlated with the evolutionary distance, as illustrated in Figure 46.

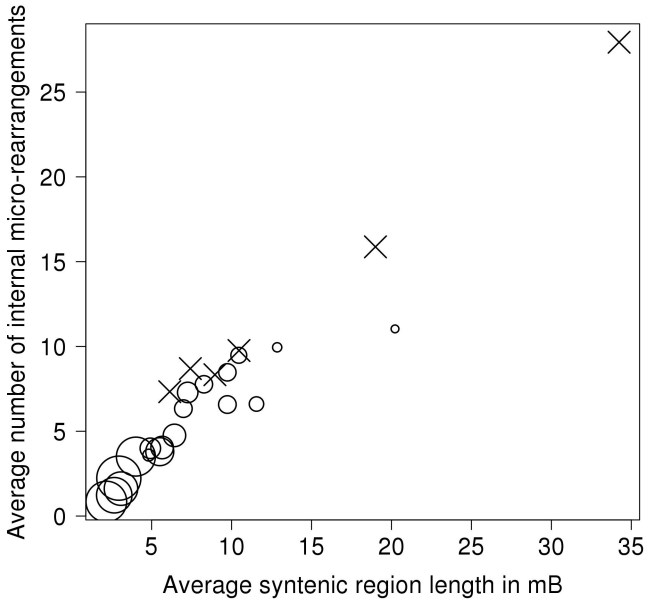
Even though the total number of micro-rearrangements indirectly depends on the degree of relationship between the genomes, there is no correlation between the density of micro-rearrangements (internal or external) and evolutionary distance (Figure 48). From this we can conclude that the driving factor for the formation of micro-rearrangements is the harbouring region's length, while they are not directly dependent on the evolutionary relationship. Additionally, we found only a slight positive correlation between sequence distance, determined as described in section 10.6.1, and the number of internal micro-rearrangements (Pearson correlation coefficient 0.22). In particular, low sequence similarity does not necessarily lead to a high number of micro-rearrangements (see Supplementary Figure S13 in the appendix).

Number of internal micro-rearrangements depends only on size of the harbouring region

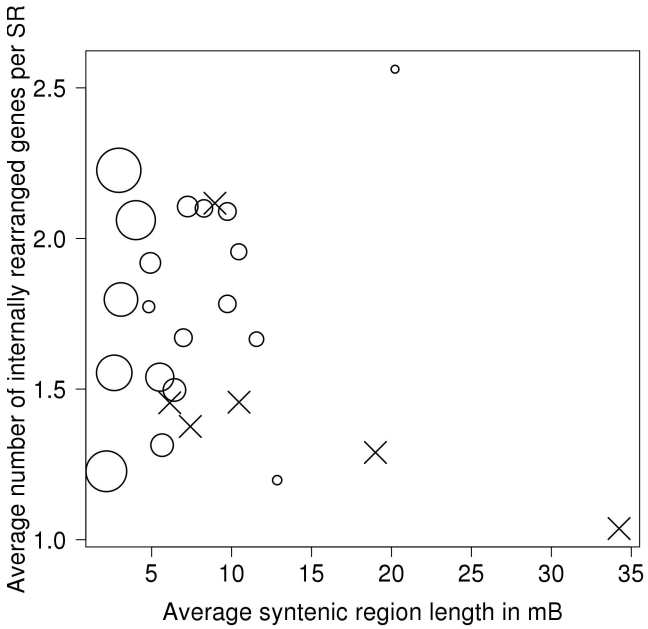
Therefore, higher numbers of micro-rearranged regions per synteny region in more closely related species are mainly due to the greater length of their synteny regions. By contrast, the size of the rearranged regions (*i.e.* the number of genes they contain) does not show any dependence on the synteny region length or evolutionary distance (Figure 48b), because the mechanisms of transposition, which is the main cause of small-scale translocations, are the same in all species.

While complex processes involving double strand breaks of chromosomes cause macro-rearrangements [125], smaller rearrangements are most often caused by the cut-and-paste mechanism of DNA transposition [136]. Our results show that there are constraints that limit the total length of the translocated genome region. In biotechnology, DNA transposons are often used as vector elements, and it is known that transposition efficiency decreases with increasing size of the cargo [91, 44, 10]. We show that the average cargo of transposons in higher eukaryotic genomes comprises between 1 and 2.5 genes for translocations over short linear distances, and between 1 and 1.2 genes for more distant translocations. This indicates that the length constraints of transposable elements also depend on the linear distance between the source and the target position in the genome. This insight could be valuable for biotechnology, where ways to overcome the length limitations in transposons are needed [10, 234].

External translocations between different synteny regions are less common than internal rearrangements, with an average number of 0 to 4 per synteny region. Overall there is only a slight tendency for the longer synteny regions of more closely related species to contain a higher number of such external micro-rearrangements (Figure 50a). In the majority of cases, only a single gene is subject to an



(a)



(b)

Figure 47: a) The average number of internal micro-rearrangements per synteny region strongly correlates with synteny region length and evolutionary distance. Evolutionary distant pairs of species share short synteny regions with few internal micro-rearrangements.
b) The size of the internal micro-rearrangement (average number of genes involved) does not correlate with the synteny region length and evolutionary distance.

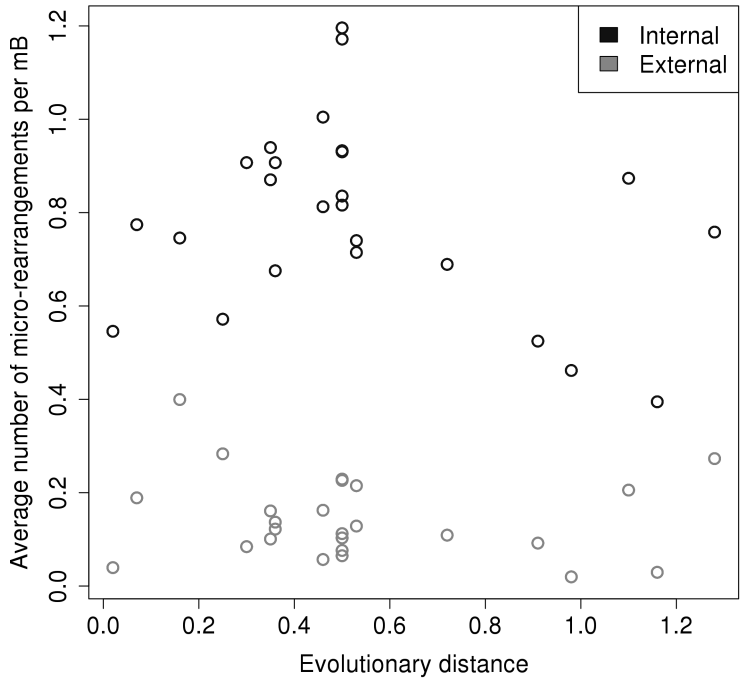
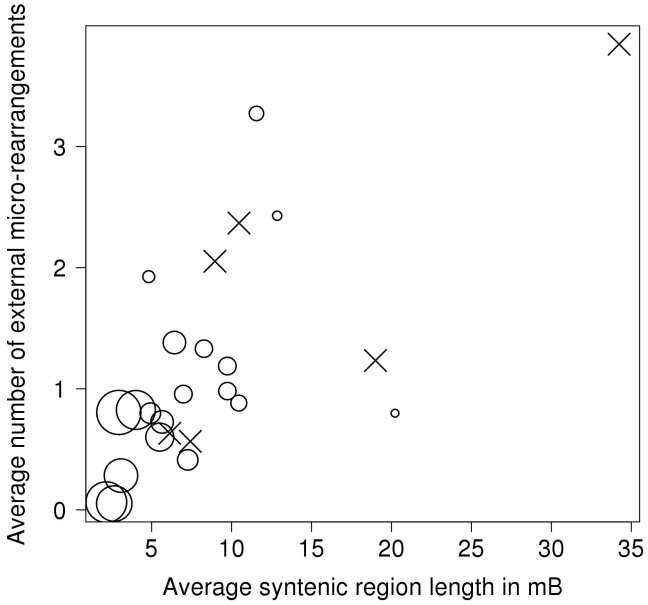


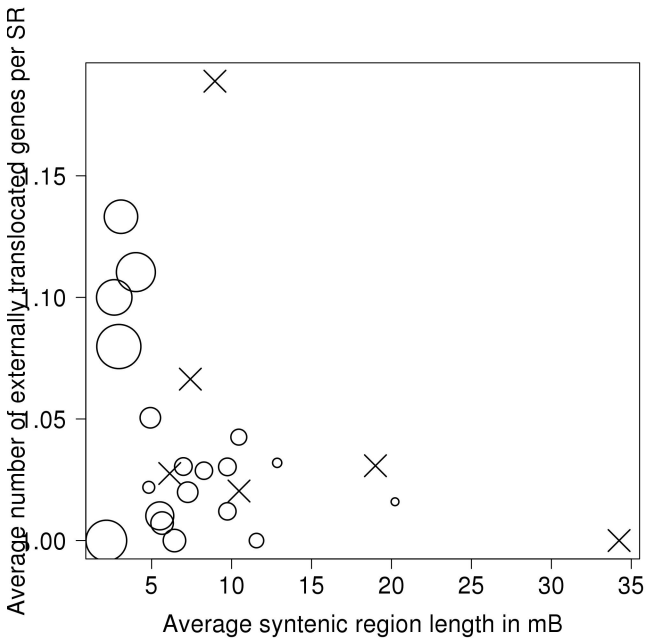
Figure 48: Average number of micro-rearrangements (internal and external) per megabase of synteny regions versus evolutionary distance. No correlation can be observed.

external translocation, regardless of the size of the harbouring region or the evolutionary distance (Figure 50b). However, with the exception of mouse and platypus, which harbour only 1.0 external micro-rearranged gene on average, all distantly related species pairs (mouse *vs* chicken, human *vs* platypus, human *vs* chicken, chicken *vs* lizard) have somewhat longer external translocations than other species pairs (average > 1.05).

Such larger external micro-rearrangements are especially common between chicken (*G. gallus*) and wild turkey (*M. gallopavo*). In particular, SyntenyMapper was able to detect a group of six consecutive genes that was translocated from one synteny region to another since species separation (see Figure 50). As the only closely related species pair with common larger external translocations, chicken and turkey genomes behave very differently from the others in this aspect. Further research could be able to identify what triggers these differences and gather more knowledge on evolutionary mechanisms.



(a)



(b)

Figure 49: a) No clear correlation between synteny region length and rearrangement number can be observed for external micro-rearrangements.
b) Similarly, the number of genes involved in external translocations is generally independent of the syntenic region length. However, with the exception of mouse and platypus (average number of genes in externally translocated regions: 1.0), distant species pairs (large circles) tend to have somewhat longer externally translocated regions (average > 1.05).

Figure 50: A translocation involving six consecutive genes between two synteny regions in chicken (G. gallus, dark grey) and wild turkey (M. gallopavo, light grey), marked by red lines. Ticks are placed at 100 kb distance and the numbers show the positions in Mb on chromosomes 6 in chicken and 8 in turkey (region 47198) as well as on chromosomes 17 in chicken and 19 in turkey (region 47135). For colour legend see Figure 44 (b).

11.4 COMPARISON OF SYNTENYMAPPER WITH CYNTENATOR, I-ADHORE AND MCSCANX

As described in section 10.7, we applied three widely used collinear block detection tools termed Cyntenator, i-ADHoRe and MCScanX as well as our own to ENSEMBL synteny regions and compared the results both qualitatively and quantitatively according to measures proposed by Ghiurcuta and Moret [61]. SyntenyMapper is the only tool developed to regard the hierarchical structure of synteny regions in eukaryotic genomes and detect all minor rearrangements in macro-rearranged regions. However, tools for collinear block detection such as Cyntenator [173], i-ADHoRe [164] or MCScanX [216] can also be applied to synteny regions instead of whole genomes to achieve a similar effect. While this approach misses rearrangements between synteny regions, these are very rare. We discarded single-gene rearrangements from SyntenyMapper's results to obtain better comparability between the methods.

11.4.1 *Cyntenator is unable to detect inversions*

Cyntenator uses the Smith-Waterman Algorithm usually applied to genomic sequences, and applies it to genomes which are represented as sequences of genes. Homology relationships can be obtained previously with BLASTP [4]. Through this easy and straightforward approach it is able to identify local alignments of genes. These local alignments represent blocks of conserved gene order in both species. In contrast to our method, Cyntenator can handle multiple genomes when provided with a phylogenetic guide tree that can be obtained from other sources. Starting with pairwise alignments for neighbouring leaves, it uses a progressive alignment procedure to integrate other species. A phylogenetic distance-dependent penalty is used to ensure that homologous genes from closely related species are included with priority over those from distantly related species. For a genome with n genes, Cyntenator performs with a runtime complexity of $O(n^3)$ and space complexity of $O(n^2)$ [173].

We have identified a set of reasons for discrepancies between results from Cyntenator and SyntenyMapper. All examples given are taken from the human and mouse comparison. Due to design of the experiment, Cyntenator cannot detect external micro-rearrangements, leading to different numbers of identified collinear blocks in only 8 of 356 synteny regions (2.25%). While this small discrepancy can be neglected, another methodological difference leads to a larger number of missed collinear blocks: Cyntenator is, in contrast to our method, unable to detect inversed regions of conserved gene order of any size. Such inversions are common and occur in 170 out of 356 (47.75%) of all synteny regions. There are also cases where the complete region

is inversed in the second species (see Figure 51). Since these cases are apparently not accounted for in the Cyntenator method, it is unable to detect any of them as collinear blocks.

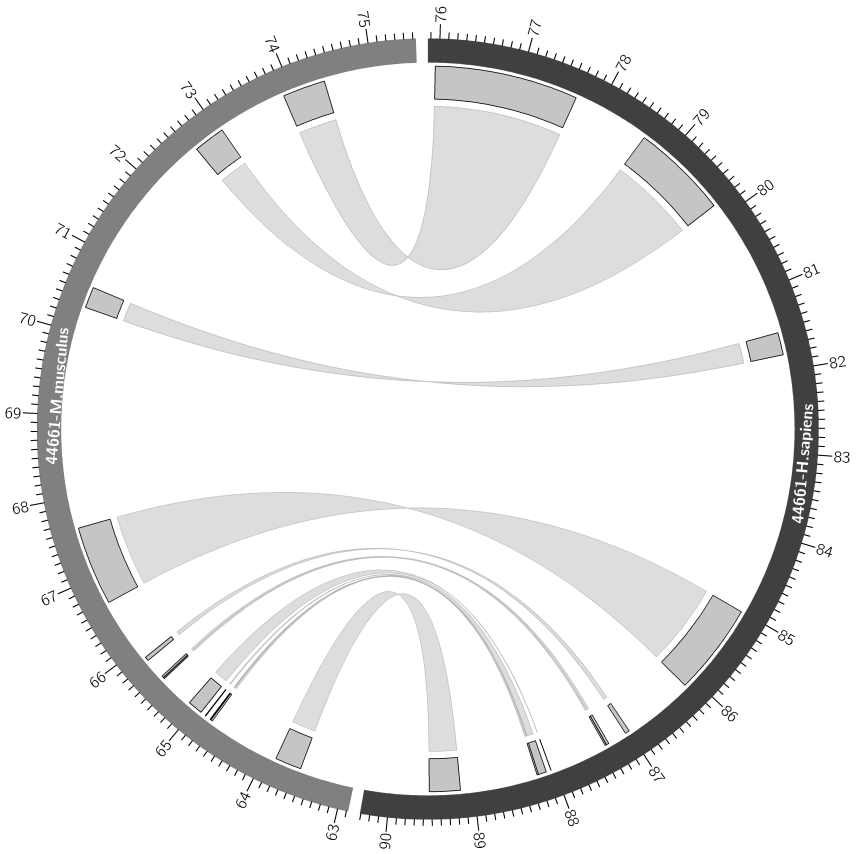


Figure 51: Example of a completely inversed synteny region. Cyntenator does not detect any collinear block for this region (ENSEMBL identifier 44661). Light grey blocks are genes, connecting ribbons illustrate orthology relationships.

In contrast, we have identified only one synteny region where our method does not detect a collinear block of four genes that is found by Cyntenator, because two genes in this block overlap and are consequently excluded.

Other minor differences in the number of detected collinear blocks are caused by SyntenyMapper's hierarchical definition. If an ENSEMBL syntenic region contains one rearranged region, SyntenyMapper will define this one region as embedded in the original region. Other tools, however, will define three consecutive collinear blocks. Besides this, methodological differences also lead to minor length differences in block definition. Cyntenator applies the Smith-Waterman alignment algorithm and thus can mismatch non-orthologous genes and include gaps when genes cannot be mapped to an ortholog. While SyntenyMapper also allows for the presence of genes without

orthologs, it will fragment a collinear block that is disrupted by a pair of non-orthologs into two blocks. Additionally, overlapping genes are excluded by our method. When we compare results from Cyntenator and SyntenyMapper, we thus often find apparently shorter but equivalent collinear blocks in the results from our method.

We have also identified some cases where we could find no reason for Cyntenator's decision to fragment two consecutive collinear blocks (Supplementary Figure S14). In conclusion, SyntenyMapper is better suited for the identification of all collinear blocks in a synteny region, including inversions.

11.4.2 *i-ADHoRe allows mismatches of orthologs*

Similar to Cyntenator, i-ADHoRe creates a gene alignment with either the Needleman-Wunsch Algorithm, a greedy graph algorithm or a new algorithm named GG2, using a *gene homology matrix* as basis. The columns of this matrix correspond to genes from one organism, and rows to genes from the other. The cells contain binary values that indicate presence of an orthology relationship. Like Cyntenator, i-ADHoRe also includes gaps and/or mismatches to create a collinear block with as many orthology pairs as possible. A progressive approach is implemented to allow for multi-species comparison.

i-ADHoRe is another common tool for the detection of collinear blocks. The main difference between our method and this approach is the usage of gaps and mismatches to create longer alignments in i-ADHoRe. SyntenyMapper, on the other hand, ignores only genes without orthologs and does not allow mismatches, but instead disrupts a collinear block if it is interrupted by non-orthologous genes. In general, both methods largely agree on the number of collinear blocks present in a synteny region, even though the sizes differ due to mismatches and gaps. For 222 (62.36%) synteny regions, both tools detect the same number of blocks. The majority of the remaining regions (68, 19.10%) contain less such blocks according to i-ADHoRe than to SyntenyMapper, caused mainly by mismatches and gaps which lead to disruption in SyntenyMapper. An example is shown in Supplementary Figure S15, where SyntenyMapper correctly identifies a micro-rearrangement of three genes within a block of conserved order of eight genes, while i-ADHoRe detects only one block with all eleven genes.

Because i-ADHoRe allows mismatches, it produces less exact results

The remainder of cases (43, 12.08%) are synteny regions where SyntenyMapper identifies less collinear blocks than the other method. We have already described in the comparison to Cyntenator that the hierarchical definition, where a set of genes is defined as the conserved order backbone of the synteny region and micro-rearrangements are defined as embedded within this set, leads to smaller numbers of detected blocks in some cases. Additionally, we have discovered that

longer stretches of genes without orthologs cause disruption of blocks in i-ADHoRe results. Our tool, however, discards any genes without orthologs in a very early step and disregards them later. Since they have no equivalent in the other genome, we do not consider them to disrupt conserved gene order blocks.

In general, we can conclude that i-ADHoRe, similar to Cyntenator, is a good tool for a less exact detection of micro-rearrangements. However, the effort to create long collinear blocks leads to mismatch pairing of non-orthologous genes, and, consequently, to ill-defined blocks. SyntenyMapper is thus better suited if a comprehensive analysis of all exact micro-rearrangements is to be conducted. Another advantage is the hierarchical approach for which it was designed, setting micro-rearrangements within the context of large synteny regions. While other methods can also be used for this purpose, SyntenyMapper simplifies the usage by downloading the necessary synteny regions itself, and also regards the hierarchy in the definition of micro-rearrangements. i-ADHoRe and Cyntenator, on the other hand, are applicable to more than two genomes and can thus be used for more complex evolutionary analyses.

11.4.3 *MCScanX applies no pre-processing to take care of many-to-many ortholog groups*

We also aimed to compare our results with MCScanX, a software package that includes not only methods for the identification of collinear blocks, but also many tools for downstream analysis and visualization. Like Cyntenator, it is based on a list of orthologous genes, for example obtained by BLASTP [4], and uses a dynamic programming approach to find chains of collinear gene pairs in the two genomes. Its main advantage is the large set of downstream analysis tools, which includes four different visualizations of the results, classification of duplicated genes into specific classes or detection of Whole Genome Duplication (WGD) events.

However, during comparison between SyntenyMapper and MCScanX it became apparent that this program fails to handle duplicates correctly. We have found many cases where MCSCanX splits a one-to-many ortholog group into multiple overlapping collinear blocks, leading to a single gene appearing in more than one such block and being paired with different orthologs. The reason for this is probably a lack of pre-processing to filter one-to-many and many-to-many ortholog groups, so that for these genes more than one orthology relationship is registered. SyntenyMapper transforms these relationships into one-to-one gene pairs before applying its collinear block detection method, thereby effectively taking care of duplicates. As MCScanX does not do this, a gene with more than one ortholog can end up in more than one collinear block.

Per definition, a single gene cannot be part of two collinear blocks unless it sits at the very edge, and never of more than two. From these results we conclude that MCScanX cannot provide the same functionality as SyntenyMapper or Cyntenator, as this behaviour leads to significant overlapping of the detected collinear blocks.

11.4.4 *Quantitative comparison of SyntenyMapper, i-ADHoRe and Cyntenator*

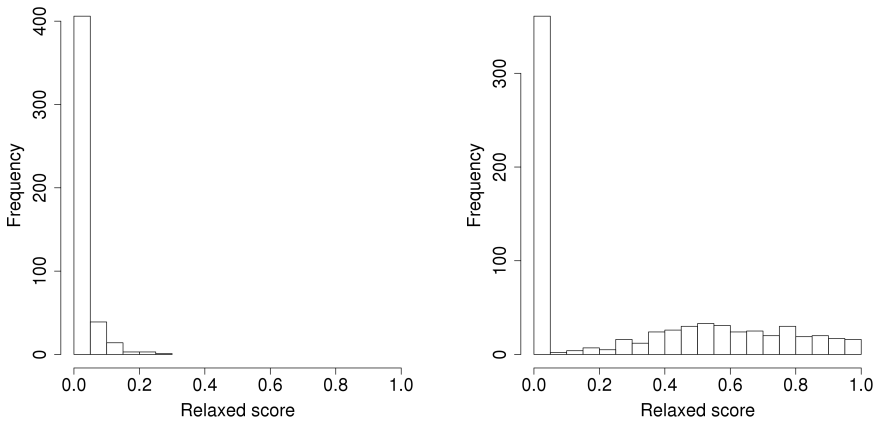
Besides the above described qualitative comparison of results, we applied quality measures proposed by Ghiurcuta and Moret [61] to assess the overall performance of both methods. These measures are based off a formal definition of *syntenic blocks* (SB), and the concordance of method results with this definition. Table 23 summarizes the measures for SyntenyMapper, i-ADHoRe and Cyntenator. Though SyntenyMapper also includes original synteny regions in the output, we calculated measures only for the set of rearranged collinear blocks detected by it.

Table 23: Measures and properties for evaluation of syntenic block detection methods, and results for SyntenyMapper, i-ADHoRe and Cyntenator. For details on the measures see methods section 10.7. Maximum relaxed score is 1.0.

Measure	SyntenyMapper micro-rearrangements	i-ADHoRe	Cyntenator
SBFs	2,898	722	466
Content overlap	19	361	40
w/o homologs in the SBF	0	361	40
Selective content	0	0	0
Mean relaxed score	0.83	0.30	0.02
Median relaxed score	1.0	0.04	0.009

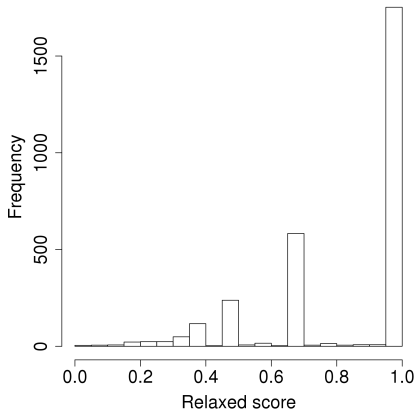
As all methods compared here are tools for detection of collinear blocks, the selective content which measures the number of non-collinear blocks is 0 for all. The number of SBFs or syntenic block families is significantly higher in SyntenyMapper's output, due to it's exact approach. Still, the number of SBFs that contain genes that are also part of other families, or content overlap, is much lower in SyntenyMapper's micro-rearrangements compared to Cyntenator's collinear blocks (0.7% vs. 8.6%) and caused by overlapping ENSEMBL synteny regions. In fact, SyntenyMapper identifies all overlapping genes by design and excludes them from the current block. In i-ADHoRe even half of all SBFs overlap.

More important than content overlap in terms of SB quality is the number of SBF with at least one marker that has no homologs in the SBF. These markers basically represent externally translocated genes, since genes without homologs are excluded for the analysis. SyntenyMapper not only treats these separately, but it also disrupts internal rearrangements when such a case appears. As a consequence, the measure is 0, while for Cyntenator 8.6% and for i-ADHore 50% of SBFs contain such cases. So far we can conclude that through its exact treatment of many exceptional cases and lack of size limitation, SyntenyMapper adheres better to the formal description of SBs than other methods.



(a) Cyntenator

(b) i-ADHoRe



(c) SyntenyMapper

Figure 52: Relaxed scores of collinearity detection for three methods applied to the human and mouse genomes. The relaxed score is a measure proposed by Ghiurcuta and Moret [61] that quantifies the percentage of genes in a collinear block that have no orthologs aside from this block. The best score is 1.0.

This can also be seen in the distribution of the proposed relaxed scoring for all SBs, as shown in Figure 52. Since we are evaluating

all tools only for pairwise genome comparison, the weighted scoring does not provide additional information. Additionally, the block incompleteness measure is identical to the relaxed scoring for SyntenyMapper, as all markers available in the SB according to the output also have homologs within the SB. We thus focus on the relaxed score alone for comparison. An ideal relaxed score equals 1.0 and signifies that all genes in the region also have a homolog in the the same region, disregarding genes without homologs.

SyntenyMapper outperforms Cyntenator and i-ADHoRe according to quantitative measures

For Cyntenator and i-ADHoRe, the majority of SBs has a low relaxed score, with a median of 0.009 and 0.04, respectively. Our own tool, on the other hand, has a median relaxed score of 1.0 and clearly shows a distribution heavily skewed to the right. In fact, only removed overlapping genes cause deviances from the perfect score of 1.0. From these results we can clearly see that SyntenyMapper preserves the formal definition of SBs much better when detecting collinear blocks than Cyntenator or i-ADHoRe. The reason for this lies in the implementation of SyntenyMapper, which strives to consider every possible exception to create perfect collinear blocks.

Ghiurcuta and Moret mention the hierarchical structure of SBs in the genome that is often ignored in synteny region detection tools. While SyntenyMapper is not scalable to output different granularity levels, it accounts for the existence of both macro- and microrearrangements, which Cyntenator and i-ADHoRe do not. With standard use, both methods find small collinear blocks along the entire genome. SyntenyMapper's hierarchical approach thus is another advantage over the other methods.

11.5 COMPARISON OF FEATURE COMPOSITION IN THE HUMAN AND MOUSE GENOME

SyntenyMapper was first developed with the goal of creating a mapping of genes and intergenic regions that lie at equivalent positions in the human and mouse genome. When comparing the feature composition of these two species' genomes on a megabase-scale level, using ENSEMBL synteny regions as guide (see section 3.2), we felt that a gene-based mapping would be more appropriate. The main problem with comparing regions of a fixed length between two genomes is the lack of selection pressure on large portions of intergenic regions. These regions can be subject to copy number variations or other insertions/deletions that strongly influence their length.

If we consider a genome as a sequence of functional elements, fixed length regions at seemingly equivalent genomic positions might contain a long intergenic region in one species and many genes in another, caused by for example copy number variations. We thus decided to develop a mapping software that identified pairs of genes and pairs of intergenic regions that lie at equivalent genomic posi-

tions, by detection blocks of conserved gene order. While the main application of SyntenyMapper as it evolved is the comparison of synteny regions and the micro-rearrangements that accumulated since species separation, it serves well in creating such a mapping.

We applied it to the genomes of human and mouse, generating one-to-one gene pairs and one-to-one intergenic region pairs determined as described in section 10.1.2. This scaffold was then used to calculate difference measures for ten genomic features (LINE, SINE, LTR, LADs, RTD, SNPs, open chromatin and histone modifications H3k4me1, H3k4me3, H3k9ac) using TrackMapper. Figure 53 shows the distributions for observed and randomized feature coverage differences.

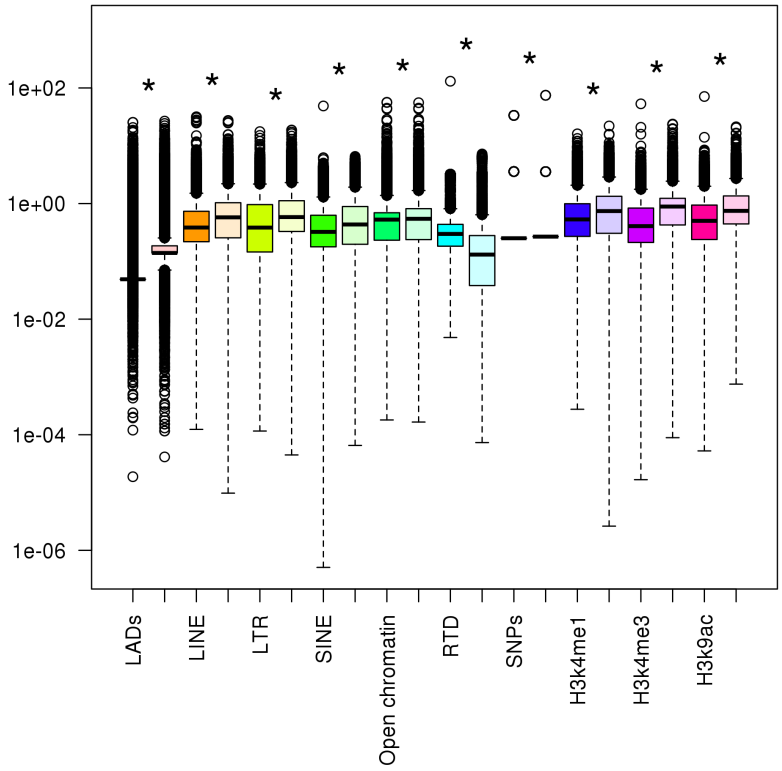


Figure 53: Distributions of difference measures for ten genomic features between all equivalent gene and intergenic region pairs for human and mouse, identified by SyntenyMapper. Light coloured boxplots (right side of each feature) are results for randomized features. Stars mark features where the difference between observed and random results is significant according to Kolmogorov-Smirnov and Rank sum test. Y axis is to log scale.

Kolmogorov-Smirnov tests and Rank sum tests were performed, revealing that all distributions are significantly different from those based on randomized features (all p-values $< 2.2 \cdot 10^{-16}$). For all

repeats (LINE, SINE, LTR), LADs, SNPs, histone modifications and open chromatin tracks, the difference between genes and intergenic regions in human and mouse is lower than for the random set. We can conclude from this that these features are conserved in the equivalent genome positions. Our previous analysis (section 3.2) also showed weak conservation for repeats, LADs and also RTD on a megabase-scale. Gene density and GC content are naturally correlated between genes and are not shown here.

The largest difference between medians of real and random distributions, implying the strongest conservation, can be observed for Histone modifications (median differences ranging from 0.21, H3k4me1, to 0.48, H3k4me3). Since these modifications mark active or inactive chromatin, it is not surprising that they are conserved for genes and intergenic regions, respectively. Interestingly, RTD density is not conserved between genes and intergenic regions in human and mouse. Instead, the difference measures are lower for randomized RTDs, indicating that RTDs are less conserved than expected by chance. RTDs are domains along the genome that are replicated at similar time points. Conservation that is lower than expected at random could indicate that, through genome reorganization in the form of macro-rearrangements, this structure was disrupted and had to be reformed independently on both species' genomes. However, replication time is partially linked with gene expression, e.g. housekeeping genes are replicated first, and replication domains represent the starting regions around replication origins. If only genes are considered, RTDs are in fact conserved (data not shown), implying a lack of conservation only for intergenic regions.

Feature distributions in human and mouse are significantly less different than expected by chance

With the exception of RTDs, all reviewed features show evidence of conservation between equivalent genes and intergenic regions in mouse. Not only sequence features were reviewed, but also structural ones such as location at the nuclear periphery and accessibility of chromatin. Thus, we can conclude that there is some degree of structure conservation between human and mouse. Combined with the results from the previous part, we show that the human and mouse genomes are similar in many structural and sequence aspects, even though the three-dimensional folds are different as a result of macro-rearrangements during evolution.

12

CONCLUSION

SyntenyMapper is a new fast comparative genomics tool for the detection of micro-rearrangements, positional orthology, and direct comparison of genomic features between two species. Most existing methods in comparative genomics focus on finding regions caused by either macro- or micro-rearrangements on a whole-genome scale, using only whole-genome alignments or homologous elements for their definition. By contrast, our method integrates both types of approaches and detects micro-rearrangements within large synteny regions by identifying blocks of conserved gene order, which can be seen as the smallest evolutionary building blocks within largely conserved regions. Though there are other collinearity detection methods which are comparable to SyntenyMapper, we show that they perform inferior to our method and lead to less exact results. SyntenyMapper provides a set of high quality exact collinear blocks, which can be used to gather insights into evolutionary history.

We have applied our method to 25 eukaryotic species pairs based on synteny regions and ortholog sets from ENSEMBL Compara [56, 211]. As expected, evolutionary distance and synteny region number and length are proportional, in the sense that closely related species contain few but long synteny regions, compared to a high number of short regions in distant pairs. The number of internal micro-rearrangements is correlated to the size of the harbouring synteny region, which in turn depends on the evolutionary distance between the genomes. However, the density of micro-rearrangements, i.e. the number of micro-rearrangements per Mb, shows no correlation with evolutionary distance. In general, regions with high sequence similarity tend to have fewer micro-rearrangements, but the correlation is weak. The main factor determining the number of micro-rearrangements in a synteny region is thus length of this region.

In line with this is the observation that more distant species pairs tend to have short to medium synteny regions harbouring between 0 and 10 internal micro-rearrangements, while more closely related organisms with long synteny regions (over 10 Mb length) contain between 10 and 25 micro-rearrangements. External rearrangements, i.e. regions translocating between synteny regions, are rare, with only 0 to 4 such translocations per synteny region in all genome pairs, and show only a slight trend for a higher number in longer regions.

The size of internal translocations is not correlated to synteny region length or evolutionary distance, implying that the transposition mechanism acts regardless of evolutionary history. It is known that

transposition efficiency decreases with the length of the translocated sequence. Our results confirm this and show that internal translocations are in general longer than external translocations, implying that long-distance translocations have stricter length constraints. External translocations are also not dependent on synteny region length or evolutionary relationship, though longer externally translocated regions appear to occur only in distant species pairs.

The following enumeration summarizes the main discoveries made in the global genome comparison of 25 eukaryotic species pairs.

A. Closely related species share few long synteny regions

B. Distant species share many short synteny regions

C. Density of micro-rearrangements is independent of evolutionary distance

D. The number of internal micro-rearrangements is mainly influenced by length of the harbouring region

E. Size of micro-rearrangements (internal and external) is independent on synteny region size and evolutionary distance

F. External translocations are rare and small

While most of these results can be expected, SyntenyMapper constructed a detailed summary of the relationships of macro/micro-rearrangement length, number and evolutionary distance, and confirmed for the first time that these expected correlations are indeed true.

Additionally, we show that our tool TrackMapper, which is based on the syntenic one-to-one mapping of genes created by SyntenyMapper, can directly compare quantitative genomic features between two species. We apply it to 10 genomic features, comprising repeats, histone modifications and others, and show that all these are more similar than expected between human and mouse genes and intergenic regions, with the exception of RTD. These results imply that the feature structure of both genomes is largely conserved, in line with results from the previous parts.

Part V

SUMMARY

13

SUMMARY

This dissertation examines the structural properties of mammalian genomes on different levels, focusing on the human genome and its evolutionary development. It covers the sequential and structural features, the three-dimensional structure, and the linear structure in the course of evolution.

In the first part of this work we discuss multiple, often inter-dependent, genomic features that are distributed along mammalian genomes and influence many processes, from gene expression to three-dimensional folding. Together, these linear features make up a complex structure that can allow us to better understand cellular processes.

We compiled a database of many different genomic features, including sequence-based properties such as repeats (LINE, LTR, SINE) or SNPs, epigenetic (histone acetylations and methylations, regions of open chromatin), structural (Hi-C compartments, lamina associated domains, DNase I hypersensitivity sites), and other features (replication timing domains). Since these often domain-like features can be clustered into euchromatic features which mark active genome regions and heterochromatic, inactive markers (mainly LADs and long repeats), we are able to classify new, experimentally determined features with respect to their genomic distribution.

For this purpose we have developed a pipeline that determines the correlation between a new feature's distribution along each chromosome and our database features. On top of that, visualization of the chromosomal domains where it is enriched and depleted allow for easy interpretation of preferential locations. For example, we were able to prove that lncRNAs HOTAIR and TERC preferentially bind in regions which are characterized as euchromatic or active due to their abundance of genes and lack of long repeats. We also show that mitochondrial sequences tend to integrate into the human genome at locations that are highly accessible, as marked by DNase I hypersensitivity sites. Altogether our results show that the genome consists of a complex net of properties that are inter-dependent and can help understand new features.

In fact, when investigating the feature distribution on human and mouse chromosomes and comparing them using ENSEMBL Compara [56, 211] synteny regions, they are weakly conserved. Human and mouse are considerably distant species in the phylogenetic tree of life, and have accumulated many so-called macro-rearrangements where large regions relocate in the genome due to double strand

breaks. As a consequence, their sequential structure differs, but the genome of one species can be reorganized according to synteny regions to make it comparable to the other. Using this approach we were able to show the strongest correlation for GC content per 1 Mb segment, but gene density, SINE, LINE and RTD coverage are also slightly conserved. The heterochromatic features LTR and LADs are only very weakly correlated.

As this thesis focuses on chromosome structures, we additionally investigated lamina-associated domains (LADs) and the behaviour of other features at the borders of these repressive environments. Confirming previous research [69] we show that gene density decreases at LAD borders, not only in human but also in mouse. Similarly, replication timing domains show a comparable profile in both species. However, other features behave species-specific: SINE and LTR appear to rarely overlap with LAD border regions in human, while showing no such distinction in mouse.

Genome-wide we were able to show a positive correlation between LADs and LINE as well as LTR, and negative correlations with euchromatic features SINE, gene density and RTD. Though these relationships are the same for both species, their strength varies. Together, these results imply that the relationship between different genomic elements and the borders of LADs is not entirely conserved, though the overall trends are similar.

In summary, we have confirmed the inter-dependency and domain-like structures of many genomic features, sequence-based as well as structural. We make use of these relationships by classifying new features as eu- or heterochromatic based on their correlation with known features. Additionally, we show that the positions of repeats, RTD, GC content, LADs and gene density are weakly conserved in the genomes of human and mouse.

The second part of this work focuses mainly on the three-dimensional structure of human and mouse genomes as derived by Hi-C experiments. These structures do not exist independently of linear features and overall processes. For example, chromosome loops can be formed to bring together genes with similar functions to be co-expressed in transcription factories. It is thus of great importance to fully understand the chromosome interactome and its relevance for biological processes.

While previous research focuses mainly on intra-chromosomal interactions, we complement it by concentrating on inter-chromosomal contacts in human and mouse embryonic stem cells. Using published data from Dixon et al. [45] and normalizing it, we constructed a high-confidence contact network for 500 kb segments.

Similar to other biological networks, segment interaction networks (SIN) have a scale-free topology in both species. Other similarities include an increased contact density for short, gene-rich chromosomes,

which are located centrally in the nucleus, enabling them to form more contacts. The repeat-rich and only partially determinable by sequencing Y chromosomes have a high number of interactions in both species, though this effect is much more pronounced in mouse. We assume that a less fixed position due to its low gene content enables the Y chromosomes to move around more freely. This flexibility would allow it to form contacts with many different genome regions in the millions of cells over which Hi-C data is collected.

The mouse genome also shows evidence of centromere-co-localization, with a central position of chromosome 11 in the centromere cluster. We can only faintly observe a similar trend in human, showing that there are further structural differences than those observed in part ii. However, investigating the correlation between inter-chromosomal spatial proximity and functional aspects, we show that in both species spatial proximity correlates with GO term similarity, though this trend is obscured by noise. In line with this result is the positive correlation between proximity and co-expression in human. For mouse, no comparable data set was available.

While these results show that there are functional similarities in the inter-chromosomal genome structures of human and mouse, there are also properties that strongly differ. In human, interactive segments are enriched in active histone marks and euchromatic features. This enrichment can not be observed in mouse, possibly caused by differences in the differentiation stages of the underlying cells, or by differing feature compositions in human and mouse.

Previous research has shown that intra-chromosomal contacts are conserved between human and mouse [45]. This is not the case for the inter-chromosomal networks, which share no more contacts than expected by chance. We believe that the disruption of chromosomes through large macro-rearrangements caused the three-dimensional structure to reform, while contacts over short linear distances could be maintained. Subsequently, new inter-chromosomal contacts would form to uphold the functional purposes of the interactome. As a consequence we can observe the previously described correlation between LADs in both species, and the correlation between spatial proximity and functional similarity.

Regarding the formation and maintenance of the three-dimensional structure, we confirm the important role of CTCF and RAD21 using ENCODE [8] data on human. Binding sites of both these transcription factors are enriched in trans-interacting segments.

Currently, Hi-C data is associated with a complex experimental procedure. As a consequence, new data becomes only slowly available. We investigated whether the relationship between linear features as described above and contact propensity is strong enough to allow prediction of inter-chromosomal interactions. We trained and tested a random forest classifier on a set of 36 features from two seg-

ments, employing data preparation methods to deal with the extreme class imbalance that is caused by the high number of non-contacting segment pairs. We show that the success of this method is highly species-dependent, with promising results only achieved for mouse. We thus conclude that at the current state of research no species-independent classification method can be developed to predict Hi-C data.

Altogether the network-based analysis shows that the human and mouse inter-chromosomal interactomes have mainly functional and structural similarities, and that individual contacts are not conserved.

While the first two parts of this thesis focus on linear and three-dimensional structural features, we investigate the evolutionary history of the linear genome in the last part. Two- and three-dimensional structures are obviously highly connected, so it is necessary to understand both in order to draw conclusions from either one. During mammalian evolution, large chromosome rearrangements were common, thereby disrupting both the linear and higher-order structures. Besides these, a large amount of smaller rearrangements was able to accumulate if they did not cause deleterious effects. We have developed a new tool, *SyntenyMapper*, to investigate the history of small scale rearrangements between two genomes in regions of large scale rearrangements, thus regarding the hierarchical structure. This tool outperforms comparable software and is thus a valuable addition to comparative genomics.

We applied SyntenyMapper to 25 eukaryotic species pairs and confirm that the number and length of large-scale rearrangements or synteny regions is dependent on evolutionary distance. Closely related species tend to share a small number of very long such regions. The number and size of small-scale rearrangements within synteny regions, however, is only dependent on the embedding region's size.

We show that SyntenyMapper is superior to other methods for collinear block detection using a newly proposed quantitative comparison by Ghiurcuta and Moret [61]. Additionally, it creates a visualization that is easy to interpret and facilitates the analysis of linear genome evolution, and provides a tool for the comparison of feature tracks termed *TrackMapper*.

Using feature data compiled in part ii, we applied TrackMapper to the human and mouse genome for a gene- and intergenic-region-based conservation analysis. We have already shown that these features are weakly conserved over segments of 1 Mb. In this part we confirm this conservation at the level of genes.

To sum up these results, we show that genome organization, be it linear or higher-order, has many similarities between eukaryotic organisms, mainly human and mouse. Though many rearrangements have occurred in these genomes since species separation, distribution of linear sequence and structural features is conserved to some ex-

tent. Similarly, the three-dimensional inter-chromosomal interactome exhibits structural similarities between both species, even though no conservation of individual contacts can be observed.

Altogether, this thesis provides a view at different aspects of genome organization, two- and three-dimensional, and highlights its evolutionary development.

Part VI

APPENDIX

A

SUPPLEMENTARY FIGURES

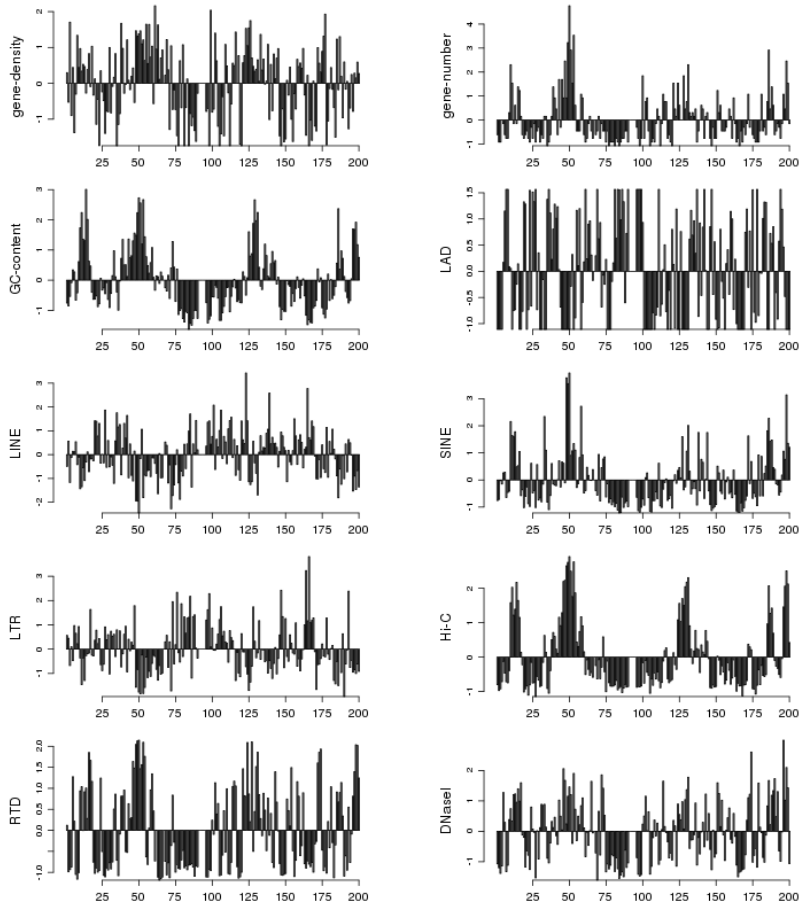


Figure S1: Simple visualization of feature tracks for human chromosome 3, implemented by Daniel Nasseh.

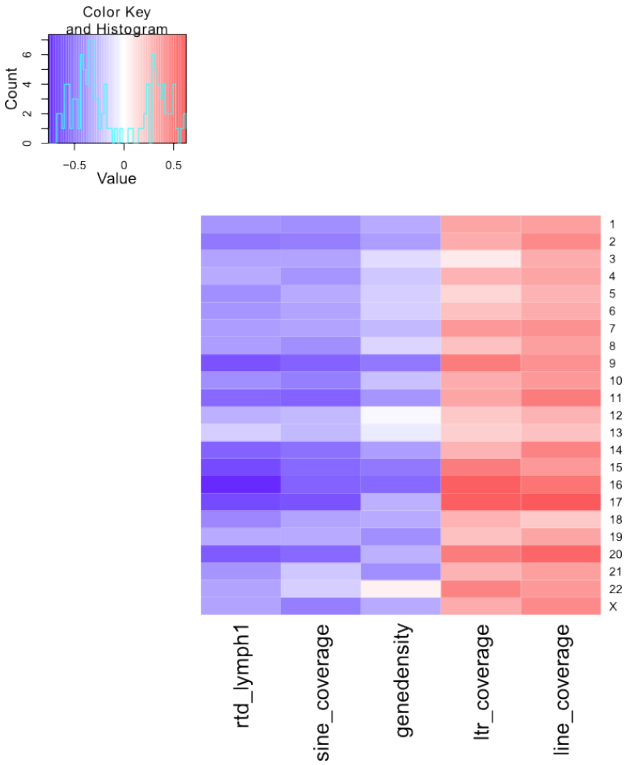
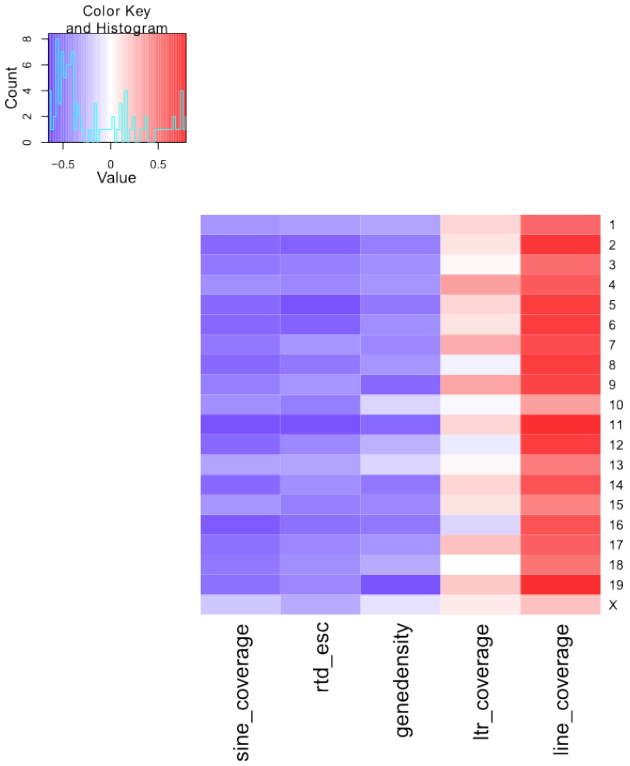
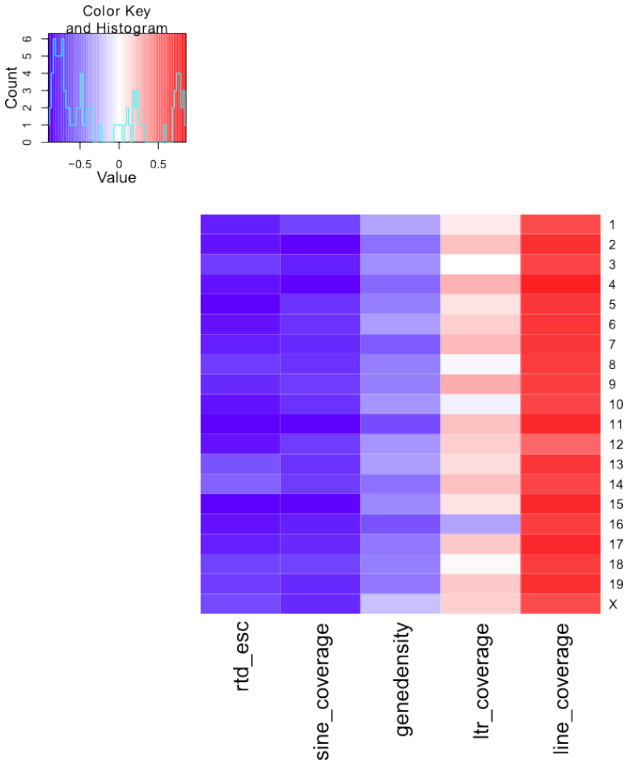


Figure S2: Heatmaps illustrating correlation of LAD distribution to other genomic features in human.



(a) Mouse, strict



(b) Mouse, greedy

Figure S3: Heatmaps illustrating correlation of LAD distribution to other genomic features in mouse. Strict refers to the set of constitutive LADs which are consistent across cell types, while the greedy set also includes all facultative LADs.

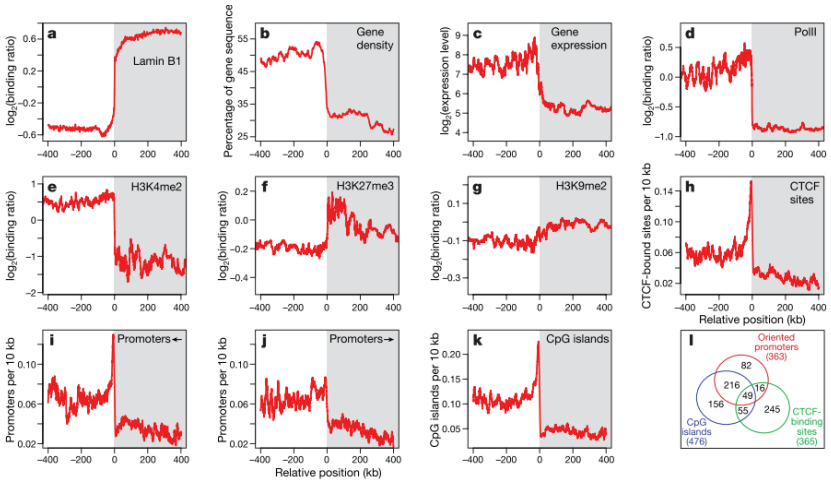


Figure S4: Profiles of genomic and chromatin features around LAD borders, taken from Guelen et al. [69]

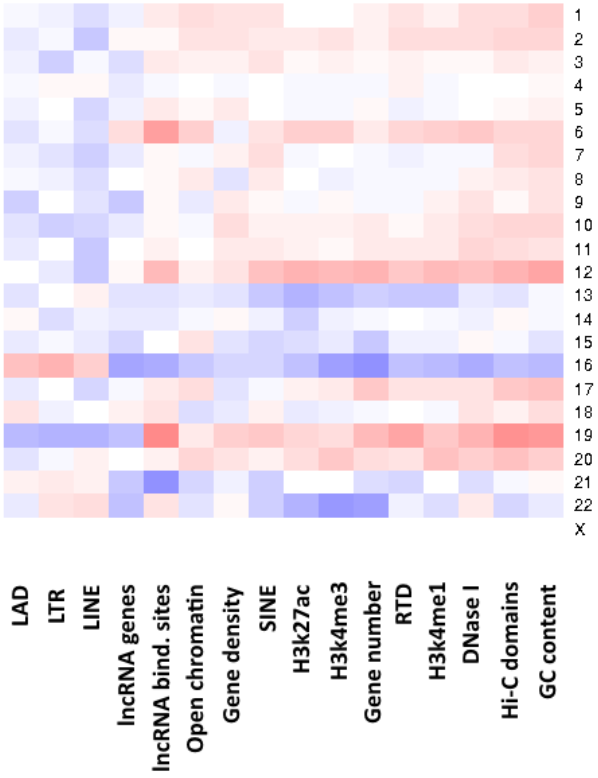


Figure S5: Heatmap of Pearson correlation coefficient between HOTAIR binding site motifs with substitutions on the genome and other genomic tracks. No clear correlation pattern is emerging.

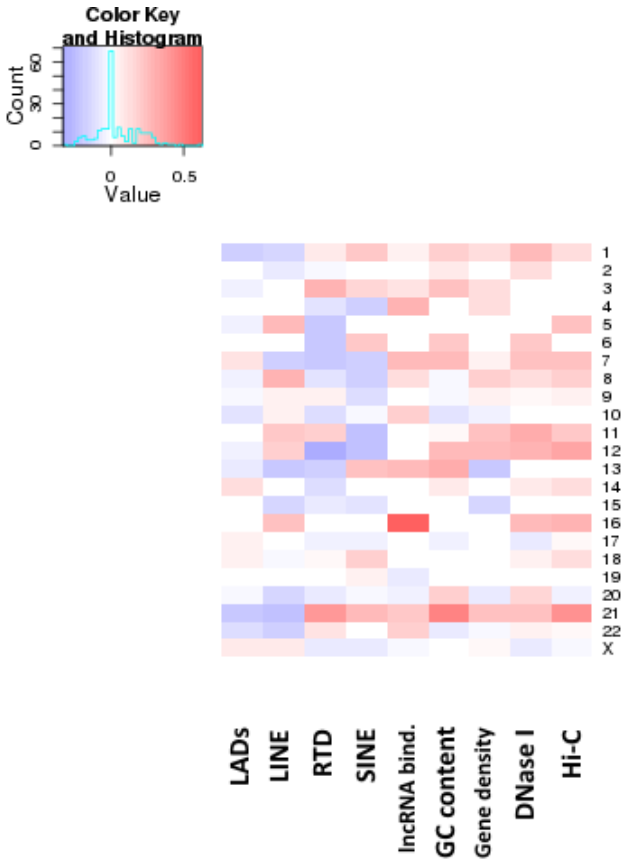


Figure S6: Heatmap of correlation coefficient values (Pearson) between NUMTS distribution and other genomic features. NUMTS coverage per 1 Mb slice was compared to human LAD coverage, LINE and SINE coverage, GC content, gene density, DNase hypersensitivity sites as well as RTD data for two cell lines and Hi-C domains. There are only very weak correlations, which are often negligible.

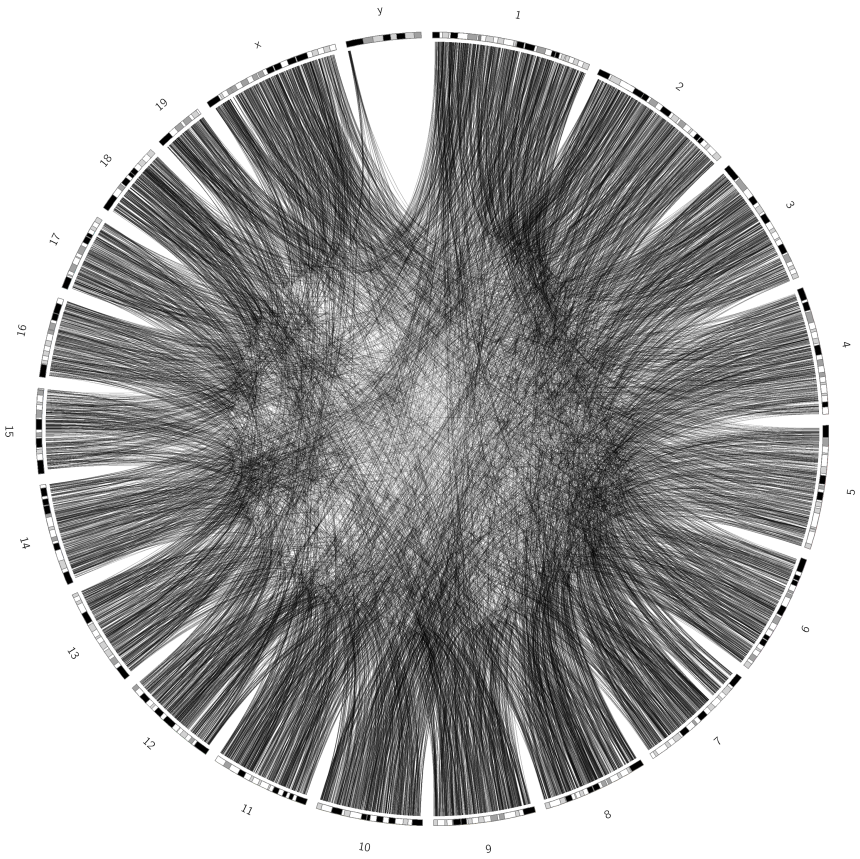
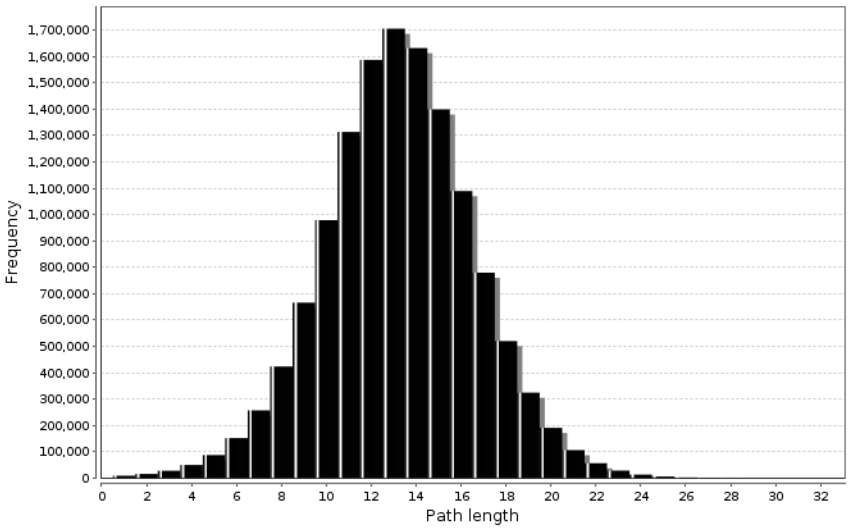
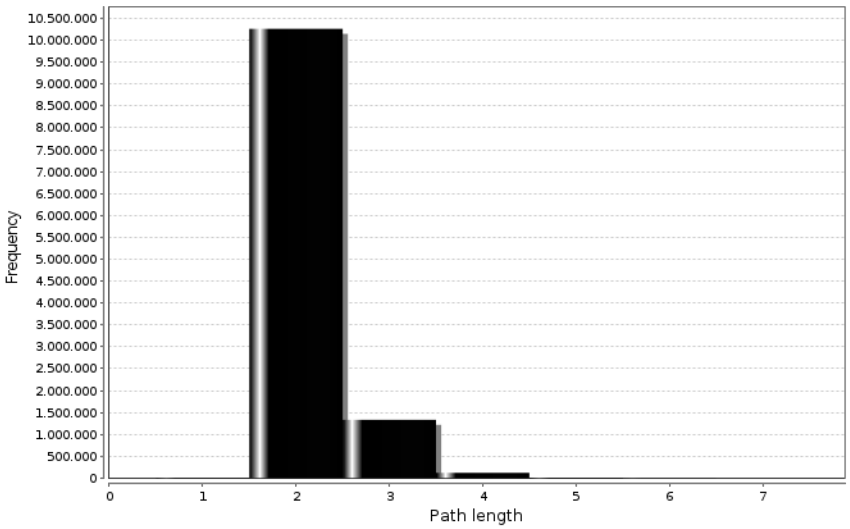


Figure S7: Illustration of inter-chromosomal contacts in the randomized mouse segment interaction network. Banded ideograms represent chromosomes 1 to Y, black lines connecting them are spatial contacts at confidence level cutoff $1e-6$.

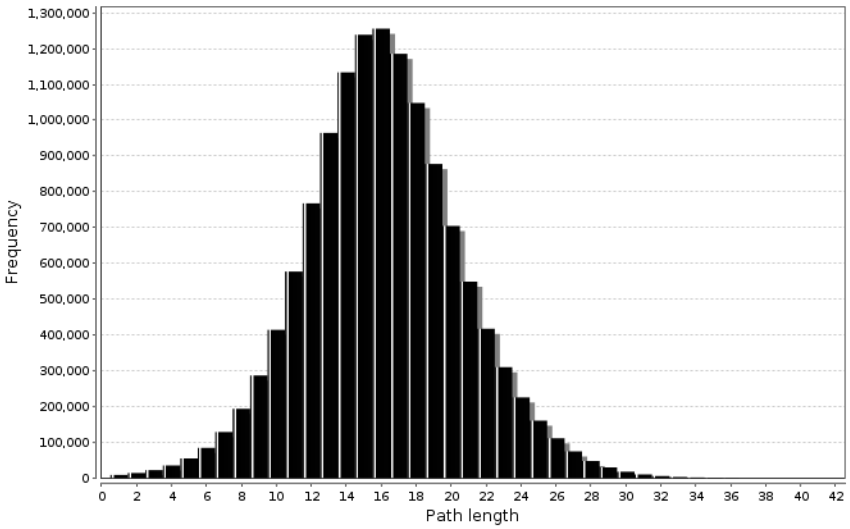


(a) RMSIN

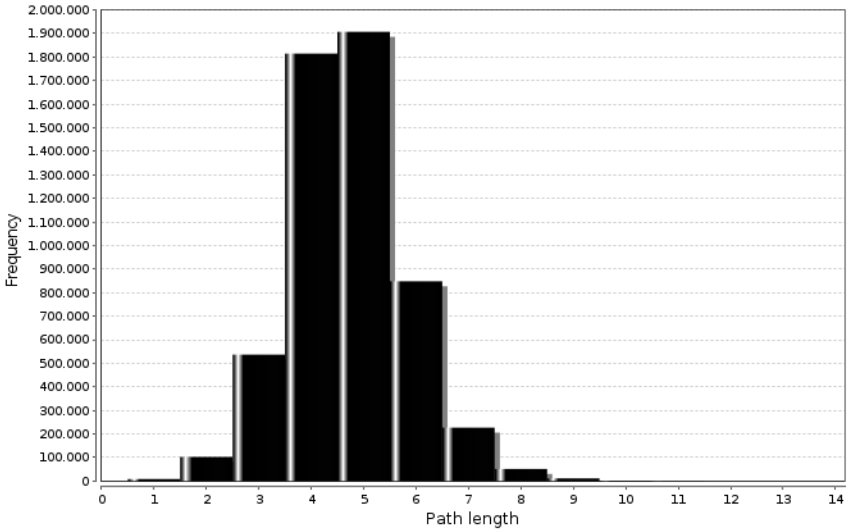


(b) MSIN

Figure S8: Distribution of shortest path lengths in the RMSIN and MSIN (cutoff $1e-6$), plotted with Cytoscape [184].



(a) RHSIN



(b) HSIN

Figure S9: Distribution of shortest path lengths in the RHSIN and HSIN (cutoff $1e-3$), plotted with Cytoscape [184].

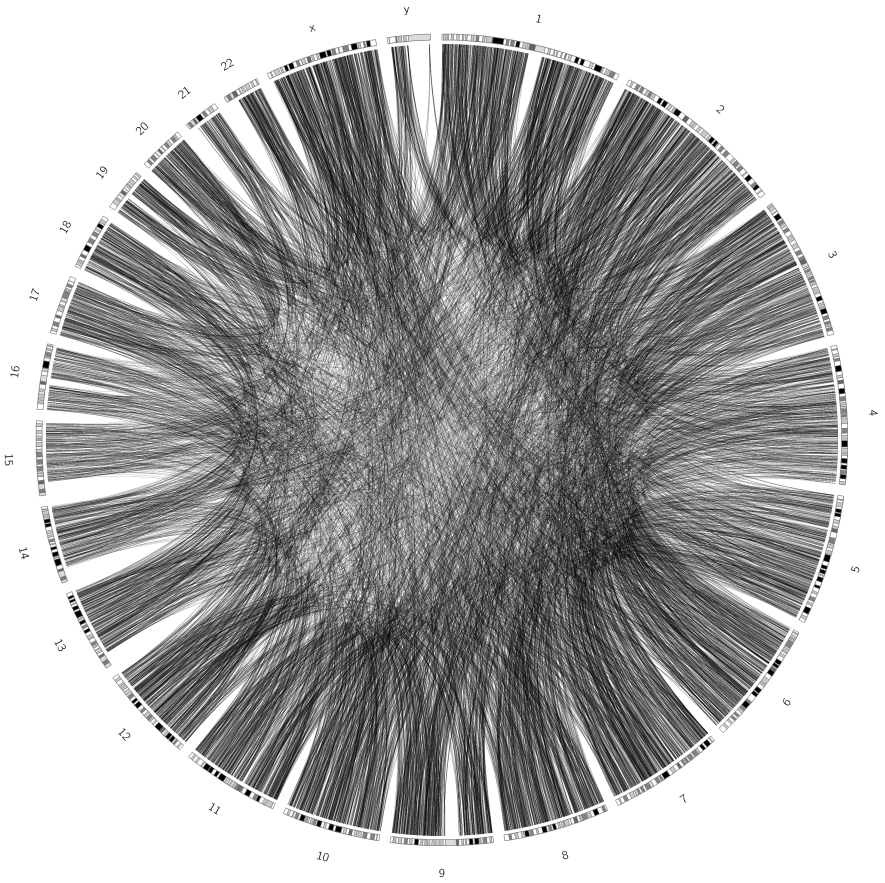
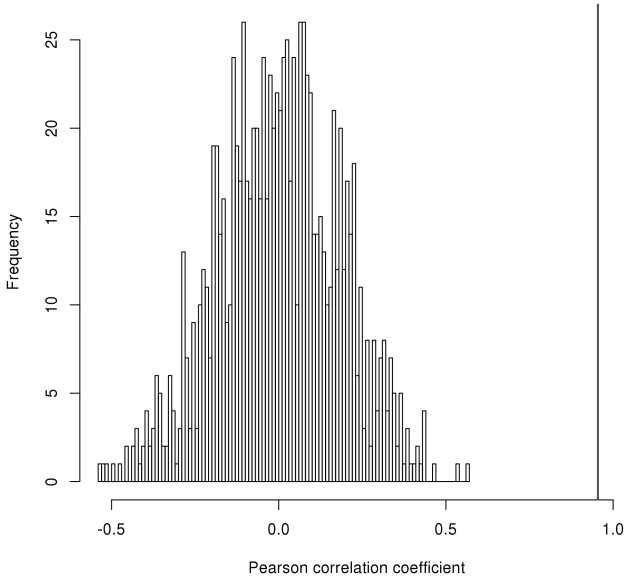
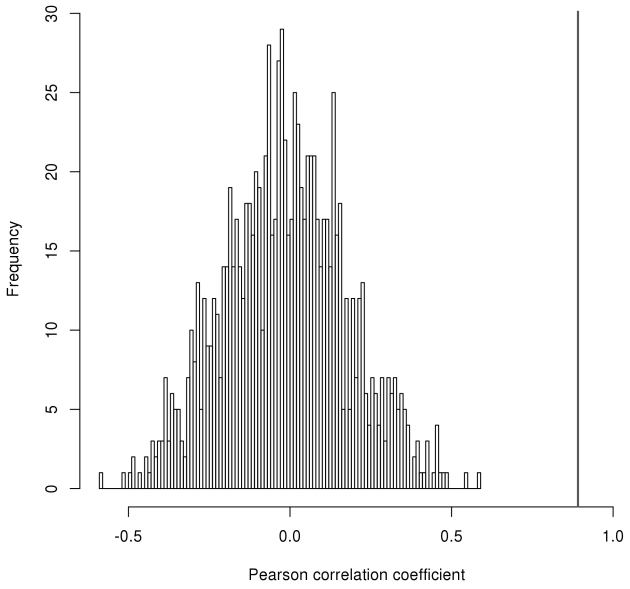


Figure S10: Illustration of inter-chromosomal contacts in the randomized human segment interaction network. Banded ideograms represent chromosomes 1 to Y, black lines connecting them are spatial contacts at confidence level cutoff $1e-3$.



(a) *H. sapiens*



(b) *M. musculus*

Figure S11: Validation of correlation between spatial proximity and GO term similarity through comparison with randomized data (section 6.2.5). p-value was assessed with the cumulative distribution function and is 0.

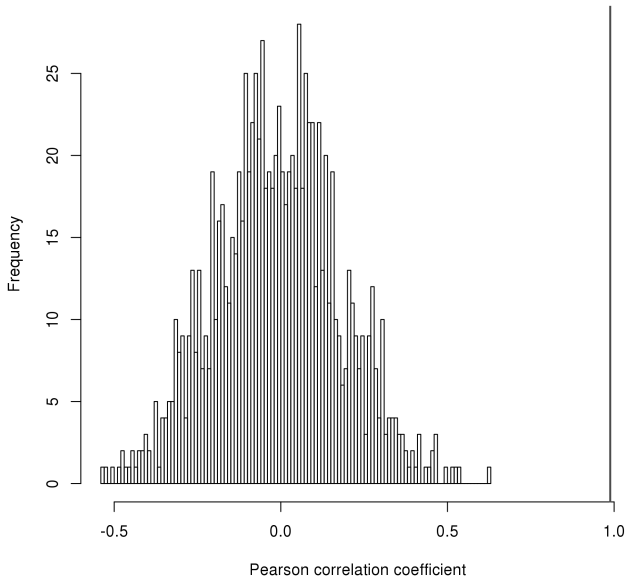


Figure S12: Validation of correlation between spatial proximity and co-expression in human through comparison with randomized data (section 6.2.5). p-value was assessed with the cumulative distribution function and is 0.

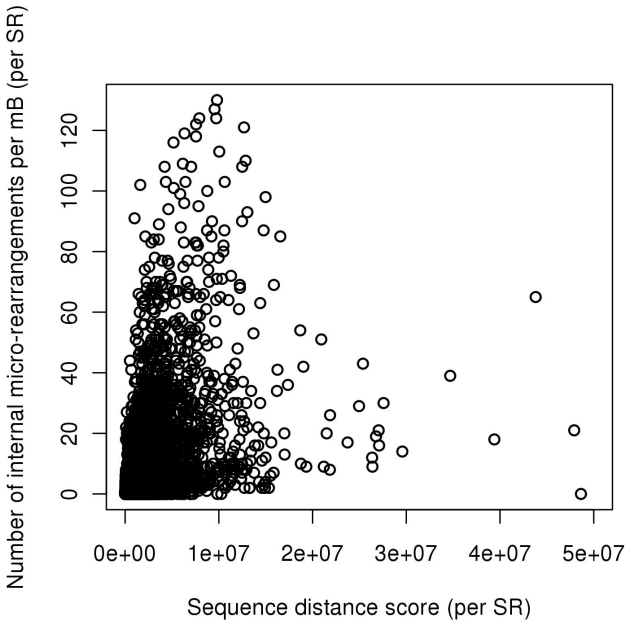


Figure S13: Synteny regions (SR) with low sequence similarity (high distance score) show a trend to contain more internal micro-rearrangements per megabase (Pearson correlation coefficient 0.22). Outliers with extremely high sequence distance are not shown.

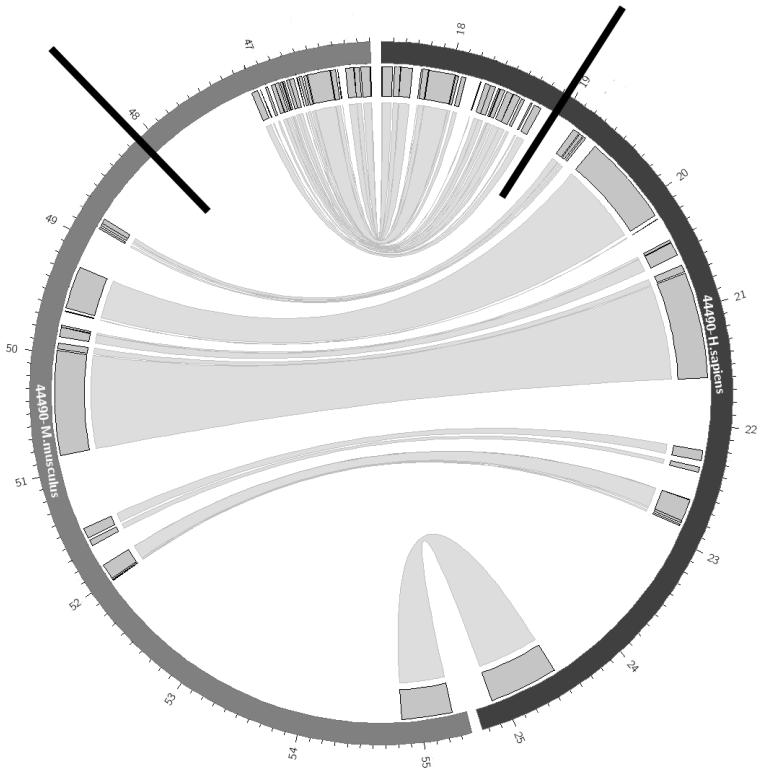


Figure S14: Example of a synteny region where Cyntenator fragments a collinear block for no apparent reason (ENSEMBL identifier 44490) at the indicated black lines. Grey boxes represent genes, connecting ribbons are orthology relationships.

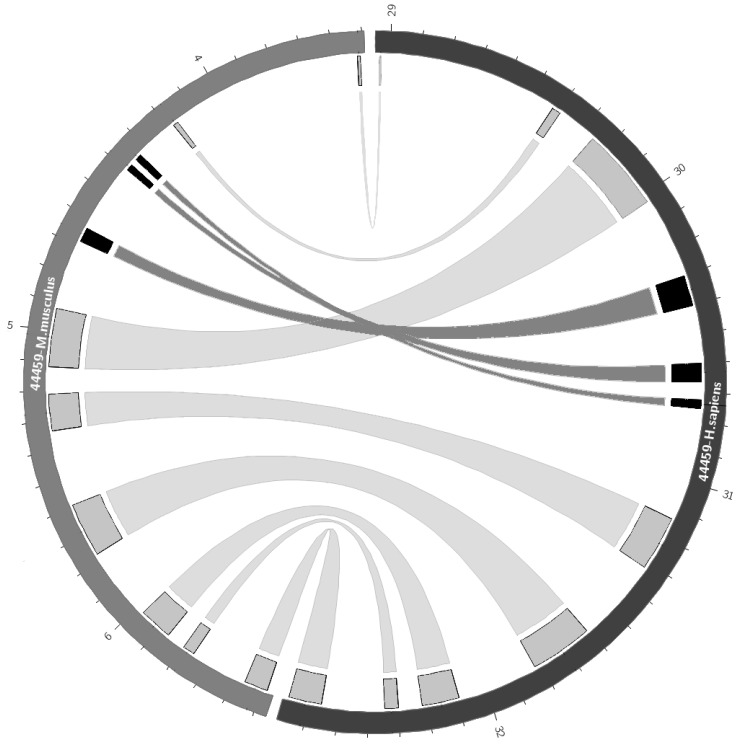


Figure S15: Illustration of a synteny region (ENSEMBL identifier 44459) where SyntenyMapper correctly identifies a micro-rearrangement containing three genes (marked in black). i-ADHoRe fails to recognize this micro-rearrangement due to gaps and mismatches and instead defines a single collinear block of eleven genes. Grey boxes represent genes, connecting ribbons are orthology relationships.

B

SUPPLEMENTARY TABLES

Table S1: Overlap of genomic features with trans-interacting segments and other segments in human and mouse.

	Total overlap with trans-interacting segments (%)	Total overlap with not trans-interacting segments (%)
	H. sapiens	
H3k4me1	12.01	8.75
H3k4me3	15.60	12.64
H3k9ac	3.68	3.21
H3k27ac	3.39	2.66
H3k36me3	13.34	10.75
LADs	18.72	15.85
DNase I sites	9.31	8.00
LINE repeats	19.25	21.40
LTR repeats	7.98	8.74
Open chromatin	2.86	2.18
RTD	4.55	4.51
SINE repeats	15.41	12.49
	M. musculus	
H3k4me1	5.45	5.89
H3k4me3	1.77	1.76
H3k9ac	2.01	2.02
H3k27ac	1.91	1.90
H3k36me3	3.99	4.14
LADs	14.27	12.80
DNase I sites	0.86	0.89
LINE repeats	15.82	15.64
LTR repeats	8.84	8.60
Open chromatin	1.09	1.14
RTD	0.65	0.64
SINE repeats	6.34	6.71

Table S2: A total list of all transcription factors from the ENCODE [8] transcription factor binding site set and their respective sources. The dataset can be accessed on UCSC [102] as the ENCODE TFBS uniform track for hg19, cell line hESC.

Transcription factor	Source
ATF2	HudsonAlpha
ATF3	HudsonAlpha
BACH1	Stanford
BCL11A	HudsonAlpha
BRCA1	Stanford
CEBPB	Stanford
CHD1	Stanford
CHD2	Stanford
CTBP2	USC
CTCF	HudsonAlpha
EGR1	HudsonAlpha
EP300	HudsonAlpha
EZH2	Broad Institute
FOSL1	HudsonAlpha
GABPA	HudsonAlpha
GTF2F1	Stanford
HDAC2	HudsonAlpha
JUN	Stanford
JUND	HudsonAlpha
KDM5A	Broad Institute
MAFK	Stanford
MAX	USC
MXI1	Stanford
MYC	Stanford
NANOG	HudsonAlpha
NRF1	Stanford
POLR2A	HudsonAlpha
POU5F1	HudsonAlpha
RAD21	HudsonAlpha
RBBP5	Broad Institute
REST	HudsonAlpha
RFX5	Stanford
RXRA	HudsonAlpha
SIN3A	Stanford
SIN3AK20	HudsonAlpha

Continued on next page

Transcription factors in ENCODE TFBS set // *continued*

Transcription factor	Source
SIX5	HudsonAlpha
SP1	HudsonAlpha
SP2	HudsonAlpha
SP4	HudsonAlpha
SRF	HudsonAlpha
SUY12	USC
TAF1	HudsonAlpha
TAF7	HudsonAlpha
TBP	Stanford
TCF12	HudsonAlpha
TEAD4	HudsonAlpha
USF1	HudsonAlpha
USF2	Stanford
YY1	HudsonAlpha
YNF143	Stanford

Table S3: Average percentage of genes in a spatial cluster with at least one TFBS of the given transcription factor. Highest percentage is reached for CTCF and RAD21.

ATF2	ATF3	BACH1	BCL11A	BRCA1
8.10%	8.60%	21.54%	2.97%	2.31%
CEBPB	CHD1	CHD2	CTBP2	**CTFC**
23.13%		16.62%	9.05%	55.73%
EGR1	EP300	EZH2	FOSL1	GABPA
21.01%	16.15%	3.16%	0.70%	11.49%
GTF2F1	HDAC2	JUN	JUND	KDM5A
7.32%	9.18%	2.51%	17.00%	2.41%
MAFK	MAX	MXI1	MYC	NANOG
15.04%	21.84%	15.53%	8.65%	8.16%
NRF1	POLR2A	POU5F1	**RAD21**	RBBP5
9.77%	50.56%	5.65%	63.33%	45.27%
REST	RFX5	RXRA	SIN3A	SIN3AK20
19.98%	2.17%	1.62%	53.05%	23.16%
SIX5	SP1	SP2	SP4	SRF
5.94%	34.67%	3.39%	14.52%	6.16%
SUY12	TAF1	TAF7	TBP	TCF12
3.12%	57.41%	30.94%	51.22%	10.92%
TEAD4	USF1	USF2	YY1	YNF143
27.17%	41.98%	11.98%	40.59%	50.80%

Table S4: Highly connected segments in the yeast segment interaction network at a cutoff of $1E-6$. These segments have a degree of over 90 and are thus the hubs of the network.

Chromosome	From	To
1	146,576	149,006
2	225,664	228,417
2	238,487	239,853
7	493,253	497,799
7	517,092	519,491
5	140,892	142,508
8	97,163	112,845
12	132,031	136,188
12	142,649	143,562
13	266,152	273,100
13	283,061	285,477
13	278,976	279,474
14	612,371	614,754
15	340,020	343,385
16	541,718	542,314
16	551,910	553,022
16	567,874	572,852
16	572,853	573,737

BIBLIOGRAPHY

[1] *Molecular Biology of the Cell*, volume 4th ed., chapter Chapter 4, pages 191–234. Garland Science, 2002.

[2] *Sequence of the Mouse Y Chromosome*, chapter The Mouse Y: The rapid expansion of a degenerating chromosome, pages 41–84. Massachusetts Institute of Technology, 2008.

[3] R. Albert. Scale-free networks in cell biology. *J. Cell. Sci.*, 118 (Pt 21):4947–4957, Nov 2005.

[4] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410, Oct 1990.

[5] S. L. Ameres and P. D. Zamore. Diversifying microRNA sequence and function. *Nat. Rev. Mol. Cell Biol.*, 14(8):475–488, Aug 2013.

[6] W. D. Andrews, A. Zito, F. Memi, G. Jones, N. Tamamaki, and J. G. Parnavelas. Limk2 mediates semaphorin signalling in cortical interneurons migrating through the subpallium. *Biol Open*, 2(3):277–282, Mar 2013.

[7] A. Ansari and M. Hampsey. A role for the CPF 3'-end processing machinery in RNAP II-dependent gene looping. *Genes Dev.*, 19:2969–2978, Dec 2005.

[8] No authors listed. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306(5696):636–640, Oct 2004.

[9] C. P. Bacher, M. Guggiari, B. Brors, S. Augui, P. Clerc, P. Avner, R. Eils, and E. Heard. Transient colocalization of X-inactivation centres accompanies the initiation of X inactivation. *Nat. Cell Biol.*, 8:293–299, Mar 2006.

[10] D. Balciunas, K. J. Wangensteen, A. Wilber, J. Bell, A. Geurts, S. Sivasubbu, X. Wang, P. B. Hackett, D. A. Largaespada, R. S. McIvor, and S. C. Ekker. Harnessing a high cargo-capacity transposon for genetic applications in vertebrates. *PLoS Genet.*, 2(11):e169, Nov 2006.

[11] F. Bantignies and G. Cavalli. Polycomb group proteins: repression in 3D. *Trends Genet.*, 27:454–464, Nov 2011.

[12] A. L. Barabasi and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, 5(2):101–113, Feb 2004.

[13] A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, May 2007.

[14] M. F. Bartholdi. Nuclear distribution of centromeres during the cell cycle of human diploid fibroblasts. *J. Cell. Sci.*, 99 (Pt 2): 255–263, Jun 1991.

[15] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57, 1995.

[16] M. Biel, V. Wascholowski, and A. Giannis. Epigenetics–an epicenter of gene regulation: histones and histone-modifying enzymes. *Angew. Chem. Int. Ed. Engl.*, 44(21):3186–3216, May 2005.

[17] D. Blankenberg, G. Von Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, and J. Taylor. Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol*, Chapter 19:1–21, Jan 2010.

[18] A. Bolzer, G. Kreth, I. Solovei, D. Koehler, K. Saracoglu, C. Fauth, S. Muller, R. Eils, C. Cremer, M. R. Speicher, and T. Cremer. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol.*, 3(5):e157, May 2005.

[19] T. Boveri. Die Blastomerenkerne von *Ascaris megalocephala* und die Theorie der Chromosomenindividualität. *Arch Zellforsch*, 3: 181–268, 1909.

[20] K. Boyd, K. H. Eng, and D. C. Page. *Machine Learning and Knowledge Discovery in Databases*, chapter Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals, pages 451–466. Springer Berlin Heidelberg, 2013.

[21] A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322, Jan 2008.

[22] L. Breiman. Random Forests. *Machine Learning*, 45:5–32.

[23] J. Brennecke, D. R. Hipfner, A. Stark, R. B. Russell, and S. M. Cohen. bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. *Cell*, 113:25–36, Apr 2003.

[24] M. Bulger and M. Groudine. Functional and mechanistic diversity of distal transcription enhancers. *Cell*, 144:327–339, Feb 2011.

[25] A. M. Bushey, E. R. Dorman, and V. G. Corces. Chromatin insulators: regulatory mechanisms and epigenetic inheritance. *Mol. Cell*, 32:1–9, Oct 2008.

[26] G. A. Calin, C. Sevignani, C. D. Dumitru, T. Hyslop, E. Noch, S. Yendamuri, M. Shimizu, S. Rattan, F. Bullrich, M. Negrini, and C. M. Croce. Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc. Natl. Acad. Sci. U.S.A.*, 101:2999–3004, Mar 2004.

[27] J. Cao. The functional role of long non-coding RNAs and epigenetics. *Biol Proced Online*, 16:11, 2014.

[28] P. Carninci et al. The transcriptional landscape of the mammalian genome. *Science*, 309:1559–1563, Sep 2005.

[29] A. T. Carpenter. Electron microscopy of meiosis in Drosophila melanogaster females. I. Structure, arrangement, and temporal change of the synaptonemal complex in wild-type. *Chromosoma*, 51(2):157–182, 1975.

[30] J. Chaumeil, P. Le Baccon, A. Wutz, and E. Heard. A novel role for Xist RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced. *Genes Dev.*, 20(16):2223–2237, Aug 2006.

[31] X. Chen. A microRNA as a translational repressor of APETALA2 in Arabidopsis flower development. *Science*, 303: 2022–2025, Mar 2004.

[32] R. Chien, W. Zeng, S. Kawauchi, M.A. Bender, R. Santos, H.C. Gregson, J.A. Schmiesing, D.A. Newkirk, X. Kong, and A.R. Ball. Cohesin mediates chromatin interactions that regulate mammalian β-globin expression. *J Biol Chem*, 286:17870–17878, 2011.

[33] Y. Chikashige, D. Q. Ding, Y. Imai, M. Yamamoto, T. Haraguchi, and Y. Hiraoka. Meiotic nuclear reorganization: switching the position of centromeres and telomeres in the fission yeast Schizosaccharomyces pombe. *EMBO J.*, 16(1):193–202, Jan 1997.

[34] C. Chu, K. Qu, F. L. Zhong, S. E. Artandi, and H. Y. Chang. Genomic Maps of Long Noncoding RNA Occupancy Reveal Principles of RNA-Chromatin Interactions. *Mol Cell*, Sep 2011.

[35] T. Cremer, M. Cremer, S. Dietzel, S. Muller, I. Solovei, and S. Fakan. Chromosome territories–a functional nuclear landscape. *Curr. Opin. Cell Biol.*, 18(3):307–316, Jun 2006.

[36] J. A. Croft, J. M. Bridger, S. Boyle, P. Perry, P. Teague, and W. A. Bickmore. Differences in the localization and morphology of chromosomes in the human nucleus. *J. Cell Biol.*, 145:1119–1131, Jun 1999.

[37] K. E. Cullen, M. P. Kladde, and M. A. Seyfred. Interaction between transcription regulatory regions of prolactin chromatin. *Science*, 261:203–206, Jul 1993.

[38] E. de Wit and W. de Laat. A decade of 3C technologies: insights into nuclear organization. *Genes Dev.*, 26(1):11–24, Jan 2012.

[39] A. Dean. On a chromosome far, far away: LCRs and gene expression. *Trends Genet.*, 22(1):38–45, Jan 2006.

[40] A. Dean. In the loop: long range chromatin interactions and gene regulation. *Brief Funct Genomics*, 10:3–10, Jan 2011.

[41] J. Dekker. The three 'C' s of chromosome conformation capture: controls, controls, controls. *Nat. Methods*, 3(1):17–21, Jan 2006.

[42] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner. Capturing chromosome conformation. *Science*, 295:1306–1311, Feb 2002.

[43] G. Diez-Roux, S. Banfi, M. Sultan, L. Geffers, S. Anand, D. Rozado, A. Magen, E. Canidio, M. Pagani, I. Peluso, N. Lin-Marq, M. Koch, M. Bilio, I. Cantiello, R. Verde, C. De Masi, S. A. Bianchi, J. Cicchini, E. Perroud, S. Mehmeti, E. Dagand, S. Schrinner, A. Nurnberger, K. Schmidt, K. Metz, C. Zwingmann, N. Brieske, C. Springer, A. M. Hernandez, S. Herzog, F. Grabbe, C. Sieverding, B. Fischer, K. Schrader, M. Brockmeyer, S. Dettmer, C. Helbig, V. Alunni, M. A. Battaini, C. Mura, C. N. Henrichsen, R. Garcia-Lopez, D. Echevarria, E. Puelles, E. Garcia-Calero, S. Kruse, M. Uhr, C. Kauck, G. Feng, N. Milyaev, C. K. Ong, L. Kumar, M. Lam, C. A. Semple, A. Gyenesei, S. Mundlos, U. Radelof, H. Lehrach, P. Sarmientos, A. Reymond, D. R. Davidson, P. Dolle, S. E. Antonarakis, M. L. Yaspo, S. Martinez, R. A. Baldock, G. Eichele, and A. Ballabio. A high-resolution anatomical atlas of the transcriptome in the mouse embryo. *PLoS Biol.*, 9(1):e1000582, 2011.

[44] S. Ding, X. Wu, G. Li, M. Han, Y. Zhuang, and T. Xu. Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. *Cell*, 122(3):473–483, Aug 2005.

[45] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, May 2012.

[46] J. Dong and S. Horvath. Understanding network concepts in modules. *BMC Syst Biol*, 1:24, 2007.

[47] M. E. Donohoe, S. S. Silva, S. F. Pinter, N. Xu, and J. T. Lee. The pluripotency factor Oct4 interacts with Ctcf and also controls X-chromosome pairing and counting. *Nature*, 460:128–132, Jul 2009.

[48] J. Dostie, Z. Mourelatos, M. Yang, A. Sharma, and G. Dreyfuss. Numerous microRNPs in neuronal cells containing novel microRNAs. *RNA*, 9:180–186, Feb 2003.

[49] J. Dostie, T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R. D. Green, and J. Dekker. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.*, 16:1299–1309, Oct 2006.

[50] H. G. du Buy and F. L. Riley. Hybridization between the nuclear and kinetoplastDNA's of Leishmania enriettii and between nuclear and mitochondrial DNA's of mouse liver. *Proc. Natl. Acad. Sci. U.S.A.*, 57:790–797, Mar 1967.

[51] Z. Duan, M. Andronescu, K. Schutz, S. McIlwain, Y. J. Kim, C. Lee, J. Shendure, S. Fields, C. A. Blau, and W. S. Noble. A three-dimensional model of the yeast genome. *Nature*, 465 (7296):363–367, May 2010.

[52] B. El Kaderi, S. Medler, S. Raghunayakula, and A. Ansari. Gene looping is conferred by activator-dependent interaction of transcription initiation and termination machineries. *J. Biol. Chem.*, 284:25015–25025, Sep 2009.

[53] J. M. Engreitz, A. Pandya-Jones, P. McDonel, A. Shishkin, K. Sirokman, C. Surka, S. Kadri, J. Xing, A. Goren, E. S. Lander, K. Plath, and M. Guttman. The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science*, 341(6147):1237973, Aug 2013.

[54] A. Eulalio, E. Huntzinger, T. Nishihara, J. Rehwinkel, M. Fauser, and E. Izaurralde. Deadenylation is a widespread effect of miRNA regulation. *RNA*, 15(1):21–32, Jan 2009.

[55] G. Felsenfeld, D.R. Davies, and A. Rich. Formation of a three-stranded polynucleotide molecule. *Am. Chem. Soc.*, 79:2023–4, 1957.

[56] P. Flicek, I. Ahmed, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, C. Garcia-Giron, L. Gordon, T. Hourlier, S. Hunt,

T. Juettemann, A. K. Kahari, S. Keenan, M. Komorowska, E. Kulesha, I. Longden, T. Maurel, W. M. McLaren, M. Muffato, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, G. R. Ritchie, M. Ruffier, M. Schuster, D. Sheppard, D. Sobral, K. Taylor, A. Thormann, S. Trevanion, S. White, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, J. Harrow, J. Herrero, T. J. Hubbard, N. Johnson, R. Kinsella, A. Parker, G. Spudich, A. Yates, A. Zadissa, and S. M. Searle. Ensembl 2013. *Nucleic Acids Res.*, 41(Database issue):48–55, Jan 2013.

[57] H. A. Foster, L. R. Abeydeera, D. K. Griffin, and J. M. Bridger. Non-random chromosome positioning in mammalian sperm nuclei, with migration of the sex chromosomes during late spermatogenesis. *J. Cell. Sci.*, 118:1811–1820, May 2005.

[58] Z. Fu, X. Chen, V. Vacic, P. Nan, Y. Zhong, and T. Jiang. MSOAR: a high-throughput ortholog assignment system based on genome rearrangement. *J. Comput. Biol.*, 14(9):1160–1175, Nov 2007.

[59] M. J. Fullwood, M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei, E. G. Chew, P. Y. Huang, W. J. Welboren, Y. Han, H. S. Ooi, P. N. Ariyaratne, V. B. Vega, Y. Luo, P. Y. Tan, P. Y. Choy, K. D. Wansa, B. Zhao, K. S. Lim, S. C. Leow, J. S. Yow, R. Joseph, H. Li, K. V. Desai, J. S. Thomsen, Y. K. Lee, R. K. Karuturi, T. Herve, G. Bourque, H. G. Stunnenberg, X. Ruan, V. Cacheux-Rataboul, W. K. Sung, E. T. Liu, C. L. Wei, E. Cheung, and Y. Ruan. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, 462:58–64, Nov 2009.

[60] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, 5(10):R80, 2004.

[61] C. G. Ghiurcuta and B. M. Moret. Evaluating synteny for improved comparative studies. *Bioinformatics*, 30(12):9–18, Jun 2014.

[62] B. Giardine, C. Riemer, R. C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, W. J. Kent, and A. Nekrutenko. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, 15(10):1451–1455, Oct 2005.

[63] J. Goecks, A. Nekrutenko, J. Taylor, E. Afgan, G. Ananda, D. Baker, D. Blankenberg, R. Chakrabarty, N. Coraor, J. Goecks, G. Von Kuster, R. Lazarus, K. Li, A. Nekrutenko, J. Taylor, and K. Vincent. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, 11(8):R86, 2010.

[64] M. Gotta, T. Laroche, A. Formenton, L. Maillet, H. Scherthan, and S. M. Gasser. The clustering of telomeres and colocalization with Rap1, Sir3, and Sir4 proteins in wild-type Saccharomyces cerevisiae. *J. Cell Biol.*, 134(6):1349–1363, Sep 1996.

[65] M. Goujon, H. McWilliam, W. Li, F. Valentin, S. Squizzato, J. Paern, and R. Lopez. A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.*, 38(Web Server issue):W695–699, Jul 2010.

[66] S. Griffiths-Jones. The microRNA Registry. *Nucleic Acids Res.*, 32(Database issue):D109–111, Jan 2004.

[67] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, 34(Database issue):D140–144, Jan 2006.

[68] S. Griffiths-Jones, H. K. Saini, S. van Dongen, and A. J. Enright. miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, 36 (Database issue):D154–158, Jan 2008.

[69] L. Guelen, L. Pagie, E. Brasset, W. Meuleman, M. B. Faza, W. Talhout, B. H. Eussen, A. de Klein, L. Wessels, W. de Laat, and B. van Steensel. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453 (7197):948–951, Jun 2008.

[70] S. Gupta, J. Dennis, R. E. Thurman, R. Kingston, J. A. Stamatoyannopoulos, and W. S. Noble. Predicting human nucleosome occupancy from primary sequence. *PLoS Comput. Biol.*, 4(8): e1000134, 2008.

[71] E. Hacisuleyman, L. A. Goff, C. Trapnell, A. Williams, J. Henao-Mejia, L. Sun, P. McClanahan, D. G. Hendrickson, M. Sauvageau, D. R. Kelley, M. Morse, J. Engreitz, E. S. Lander, M. Guttman, H. F. Lodish, R. Flavell, A. Raj, and J. L. Rinn. Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nat. Struct. Mol. Biol.*, 21 (2):198–206, Feb 2014.

[72] S. Hadjur, L. M. Williams, N. K. Ryan, B. S. Cobb, T. Sexton, P. Fraser, A. G. Fisher, and M. Merkenschlager. Cohesins form

chromosomal cis-interactions at the developmentally regulated IFNG locus. *Nature*, 460:410–413, Jul 2009.

[73] M.A. Hakimi, D.A. Bochar, J.A. Schmiesing, Y. Dong, O.G. Barak, D.W. Speicher, K. Yokomori, and R. Shiekhattar. A chromatin remodelling complex that loads cohesin onto human chromosomes. *Nature*, 418:994–998, 2002.

[74] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11.

[75] M. A. Hall. *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand, 1998.

[76] S. M. Hammond. Dicing and slicing: the core machinery of the RNA interference pathway. *FEBS Lett.*, 579:5822–5829, Oct 2005.

[77] S. M. Hammond, E. Bernstein, D. Beach, and G. J. Hannon. An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. *Nature*, 404:293–296, Mar 2000.

[78] L. Handoko, H. Xu, G. Li, C. Y. Ngan, E. Chew, M. Schnapp, C. W. Lee, C. Ye, J. L. Ping, F. Mulawadi, E. Wong, J. Sheng, Y. Zhang, T. Poh, C. S. Chan, G. Kunarso, A. Shahab, G. Bourque, V. Cacheux-Rataboul, W. K. Sung, Y. Ruan, and C. L. Wei. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat. Genet.*, 43:630–638, Jul 2011.

[79] S. Hannenhalli and P. A. Pevzner. Transforming men into mice (polynomial algorithm for genomic distance problem). In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, FOCS '95, pages 581–, Washington, DC, USA, 1995. IEEE Computer Society. ISBN 0-8186-7183-1. URL `http://dl.acm.org/citation.cfm?id=795662.796277`.

[80] E. Hazkani-Covo, R. M. Zeller, and W. Martin. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet.*, 6:e1000834, Feb 2010.

[81] X. He, Z. Zhu, C. Johnson, J. Stoops, A. E. Eaker, W. Bowen, and M. C. DeFrances. PIK3IP1, a negative regulator of PI3K, suppresses the development of hepatocellular carcinoma. *Cancer Res.*, 68(14):5591–5598, Jul 2008.

[82] A. S. Hinrichs, D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey, R. A. Harte, F. Hsu, J. Hillman-Jackson, R. M. Kuhn, J. S. Pedersen, A. Pohl, B. J. Raney, K. R. Rosenbloom, A. Siepel, K. E. Smith, C. W. Sugnet,

A. Sultan-Qurraie, D. J. Thomas, H. Trumbower, R. J. Weber, M. Weirauch, A. S. Zweig, D. Haussler, and W. J. Kent. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.*, 34(Database issue):D590–598, Jan 2006.

[83] A. S. Hinrichs, D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey, R. A. Harte, F. Hsu, J. Hillman-Jackson, R. M. Kuhn, J. S. Pedersen, A. Pohl, B. J. Raney, K. R. Rosenbloom, A. Siepel, K. E. Smith, C. W. Sugnet, A. Sultan-Qurraie, D. J. Thomas, H. Trumbower, R. J. Weber, M. Weirauch, A. S. Zweig, D. Haussler, and W. J. Kent. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.*, 34(Database issue):D590–598, Jan 2006.

[84] G. P. Holmquist. Chromosome bands, their chromatin flavors, and their functional features. *Am. J. Hum. Genet.*, 51(1):17–37, Jul 1992.

[85] G.P. Holmquist. Chromosomal Bands and Sequence Features. *eLS. John Wiley & Sons Ltd*, 2005.

[86] D. Homouz and A. S. Kudlicki. The 3D organization of the yeast genome correlates with co-expression and reflects functional relations between genes. *PLoS ONE*, 8(1):e54699, 2013.

[87] C. Hou and V. G. Corces. Throwing transcription for a loop: expression of the genome in the 3D nucleus. *Chromosoma*, Nov 2011.

[88] C. Hou, H. Zhao, K. Tanimoto, and A. Dean. CTCF-dependent enhancer-blocking by alternative chromatin loop formation. *Proc. Natl. Acad. Sci. U.S.A.*, 105:20398–20403, Dec 2008.

[89] C. Hou, R. Dale, and A. Dean. Cell type specificity of chromatin organization mediated by CTCF and cohesin. *Proc Natl Acad Sci*, 107:3651–3656, 2010.

[90] T. C. Hsu, J. E. Cooper, M. L. Mace, and B. R. Brinkley. Arrangement of centromeres in mouse cells. *Chromosoma*, 34(1):73–87, 1971.

[91] Z. Izsvak, Z. Ivics, and R. H. Plasterk. Sleeping Beauty, a wide host-range transposon vector for genetic transformation in vertebrates. *J. Mol. Biol.*, 302(1):93–102, Sep 2000.

[92] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6:429–49, 2002.

[93] Q. Jin, E. Trelles-Sticken, H. Scherthan, and J. Loidl. Yeast nuclei display prominent centromere clustering that is reduced in nondividing cells and in meiotic prophase. *J. Cell Biol.*, 141(1): 21–29, Apr 1998.

[94] Q. W. Jin, J. Fuchs, and J. Loidl. Centromere clustering is a major determinant of yeast interphase nuclear organization. *J. Cell. Sci.*, 113 (Pt 11):1903–1912, Jun 2000.

[95] B. Joffe, H. Leonhardt, and I. Solovei. Differentiation and large scale spatial organization of the genome. *Curr. Opin. Genet. Dev.*, 20:562–569, Oct 2010.

[96] J. Jun, I. I. Mandoiu, and C. E. Nelson. Identification of mammalian orthologs using local synteny. *BMC Genomics*, 10:630, 2009.

[97] R. Kalhor, H. Tjong, N. Jayathilaka, F. Alber, and L. Chen. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.*, 30(1):90–98, Jan 2012.

[98] D. Karolchik, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, and W. J. Kent. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, 32(Database issue):D493–496, Jan 2004.

[99] A. Kasprzyk. BioMart: driving a paradigm change in biological data management. *Database (Oxford)*, 2011:bar049, 2011.

[100] S. Katayama, Y. Tomaru, T. Kasukawa, K. Waki, M. Nakanishi, M. Nakamura, H. Nishida, C. C. Yap, M. Suzuki, J. Kawai, H. Suzuki, P. Carninci, Y. Hayashizaki, C. Wells, M. Frith, T. Ravasi, K. C. Pang, J. Hallinan, J. Mattick, D. A. Hume, L. Lipovich, S. Batalov, P. G. Engstrom, Y. Mizuno, M. A. Faghihi, A. Sandelin, A. M. Chalk, S. Mottagui-Tabar, Z. Liang, B. Lenhard, and C. Wahlestedt. Antisense transcription in the mammalian transcriptome. *Science*, 309:1564–1566, Sep 2005.

[101] P.W. Kenny and C.A. Montanari. Inflation of correlation in the pursuit of drug-likeness. *J Comput Aided Mol Des*, 27:1–13, 2013.

[102] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at UCSC. *Genome Res.*, 12(6):996–1006, Jun 2002.

[103] A. M. Khalil, M. Guttman, M. Huarte, M. Garber, A. Raj, D. Rivea Morales, K. Thomas, A. Presser, B. E. Bernstein, A. van Oudenaarden, A. Regev, E. S. Lander, and J. L. Rinn. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, 106:11667–11672, Jul 2009.

[104] E.E. Khrameeva, A.A. Mironov, G.G. Fedonin, P. Khaitovich, and M.S. Gelfand. Spatial proximity and similarity of the epigenetic state of genome domains. *PLoS ONE*, 7(4), 2012.

[105] A. Kozomara and S. Griffiths-Jones. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, 39(Database issue):D152–157, Jan 2011.

[106] A. Kozomara and S. Griffiths-Jones. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, 42(Database issue):68–73, Jan 2014.

[107] A. M. Krichevsky, K. S. King, C. P. Donahue, K. Khrapko, and K. S. Kosik. A microRNA array reveals extensive regulation of microRNAs during brain development. *RNA*, 9:1274–1281, Oct 2003.

[108] K. Kruse, S. Sewitz, and M. M. Babu. A complex network framework for unbiased statistical analyses of DNA-DNA contact maps. *Nucleic Acids Res.*, 41(2):701–710, Jan 2013.

[109] Martin I Krzywinski, Jacqueline E Schein, Inanc Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: An information aesthetic for comparative genomics. *Genome Research*, 2009. doi: 10.1101/gr.092759.109. URL `http://genome.cshlp.org/content/early/2009/06/15/gr.092759.109.abstract`.

[110] Martin I Krzywinski, Jacqueline E Schein, Inanc Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: An information aesthetic for comparative genomics. *Genome Research*, 2009. doi: 10.1101/gr.092759.109. URL `http://genome.cshlp.org/content/early/2009/06/15/gr.092759.109.abstract`.

[111] M. Kuroda, H. Tanabe, K. Yoshida, K. Oikawa, A. Saito, T. Kiyuna, H. Mizusawa, and K. Mukai. Alteration of chromosome positioning during adipocyte differentiation. *J. Cell. Sci.*, 117:5897–5903, Nov 2004.

[112] S. Kurukuti, V. K. Tiwari, G. Tavoosidana, E. Pugacheva, A. Murrell, Z. Zhao, V. Lobanenkov, W. Reik, and R. Ohlsson. CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. *Proc. Natl. Acad. Sci. U.S.A.*, 103:10684–10689, Jul 2006.

[113] T.R.J. Lappin, D. G. Grier, A. Thompson, and H.L. Halliday. HOX GENES: Seductive Sience, Mysterious Mechanisms. *Ulster Med J.*, 75:23–31, 2006.

[114] R. C. Lee, R. L. Feinbaum, and V. Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75:843–854, Dec 1993.

[115] Y. Lee, M. Kim, J. Han, K. H. Yeom, S. Lee, S. H. Baek, and V. N. Kim. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J.*, 23:4051–4060, Oct 2004.

[116] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, Jul 2009.

[117] H. Li and R. Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–595, Mar 2010.

[118] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug 2009.

[119] B. Y. Liao, Y. J. Chang, J. M. Ho, and M. J. Hwang. The UniMarker (UM) method for synteny mapping of large genomes. *Bioinformatics*, 20(17):3156–3165, Nov 2004.

[120] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326:289–293, Oct 2009.

[121] L. P. Lim, M. E. Glasner, S. Yekta, C. B. Burge, and D. P. Bartel. Vertebrate microRNA genes. *Science*, 299:1540, Mar 2003.

[122] J. Q. Ling, T. Li, J. F. Hu, T. H. Vu, H. L. Chen, X. W. Qiu, A. M. Cherry, and A. R. Hoffman. CTCF mediates interchromosomal colocalization between Igf2/H19 and Wsb1/Nf1. *Science*, 312:269–272, Apr 2006.

[123] Y. Liu, S. Shin, X. Zeng, M. Zhan, R. Gonzalez, F. J. Mueller, C. M. Schwartz, H. Xue, H. Li, S. C. Baker, E. Chudin, D. L. Barker, T. K. McDaniel, S. Oeser, J. F. Loring, M. P. Mattson, and M. S. Rao. Genome wide profiling of human embryonic stem cells (hESCs), their derivatives and embryonal carcinoma cells to develop base profiles of U.S. Federal government approved hESC lines. *BMC Dev. Biol.*, 6:20, 2006.

[124] P. G. Maass, A. Rump, H. Schulz, S. Stricker, L. Schulze, K. Platzer, A. Aydin, S. Tinschert, M. B. Goldring, F. C. Luft, and S. Bahring. A misplaced lncRNA causes brachydactyly in humans. *J. Clin. Invest.*, 122(11):3990–4002, Nov 2012.

[125] R. S. Mani and A. M. Chinnaiyan. Triggers for genomic rearrangements: insights into genomic, cellular and environmental influences. *Nat. Rev. Genet.*, 11(12):819–829, Dec 2010.

[126] H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Annals of Mathematical Statistics*, 18:50–60.

[127] Y. S. Mao, H. Sunwoo, B. Zhang, and D. L. Spector. Direct visualization of the co-transcriptional assembly of a nuclear body by noncoding RNAs. *Nat. Cell Biol.*, 13(1):95–101, Jan 2011.

[128] Y. Markaki, D. Smeets, S. Fiedler, V. J. Schmid, L. Schermelleh, T. Cremer, and M. Cremer. The potential of 3D-FISH and superresolution structured illumination microscopy for studies of 3D nuclear architecture: 3D structured illumination microscopy of defined chromosomal structures visualized by 3D (immuno)-FISH opens new perspectives for studies of nuclear architecture. *Bioessays*, 34(5):412–426, May 2012.

[129] F.J. Massey. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46:68–78, 1951.

[130] M. Metzler, M. Wilda, K. Busch, S. Viehmann, and A. Borkhardt. High expression of precursor microRNA-155/BIC RNA in children with Burkitt lymphoma. *Genes Chromosomes Cancer*, 39: 167–169, Feb 2004.

[131] M. Z. Michael, S. M. O' Connor, N. G. van Holst Pellekaan, G. P. Young, and R. J. James. Reduced accumulation of specific microRNAs in colorectal neoplasia. *Mol. Cancer Res.*, 1:882–891, Oct 2003.

[132] C. Michaelis, R. Ciosk, and K. Nasmyth. Cohesins: chromosomal proteins that prevent premature separation of sister chromatids. *Cell*, 91:35–45, 1997.

[133] W. Miller, K. Rosenbloom, R. C. Hardison, M. Hou, J. Taylor, B. Raney, R. Burhans, D. C. King, R. Baertsch, D. Blankenberg, S. L. Kosakovsky Pond, A. Nekrutenko, B. Giardine, R. S. Harris, S. Tyekucheva, M. Diekhans, T. H. Pringle, W. J. Murphy, A. Lesk, G. M. Weinstock, K. Lindblad-Toh, R. A. Gibbs, E. S. Lander, A. Siepel, D. Haussler, and W. J. Kent. 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res.*, 17(12):1797–1808, Dec 2007.

[134] J. A. Mitchell and P. Fraser. Transcription factories are nuclear subcompartments that remain in the absence of transcription. *Genes Dev.*, 22:20–25, Jan 2008.

[135] A. R. Morgan and R. D. Wells. Specificity of the three-stranded complex formation between double-stranded DNA and single-stranded RNA containing repeating nucleotide sequences. *J. Mol. Biol.*, 37(1):63–80, Oct 1968.

[136] M. Munoz-Lopez and J. L. Garcia-Perez. DNA transposons: nature and applications in genomics. *Curr. Genomics*, 11(2):115–128, Apr 2010.

[137] A. Murrell, S. Heeson, and W. Reik. Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops. *Nat. Genet.*, 36: 889–893, Aug 2004.

[138] T. Nagano, Y. Lubling, T. J. Stevens, S. Schoenfelder, E. Yaffe, W. Dean, E. D. Laue, A. Tanay, and P. Fraser. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, Oct 2013.

[139] R. Nativio, K.S. Wendt, Y. Ito, J.E. Huddleston, S. Uribe-Lewis, K. Woodfine, C. Krueger, W. Reik, J.M. Peters, and A. Murrel. Cohesin is required for higher-order chromatin conformation at the imprinted IGF2-H19 locus. *PLoS Genet.*, 5, 2009.

[140] N. Naumova, M. Imakaev, G. Fudenberg, Y. Zhan, B. R. Lajoie, L. A. Mirny, and J. Dekker. Organization of the mitotic chromosome. *Science*, 342(6161):948–953, Nov 2013.

[141] T. B. Nesterova, S. C. Barton, M. A. Surani, and N. Brockdorff. Loss of Xist imprinting in diploid parthenogenetic preimplantation embryos. *Dev. Biol.*, 235:343–350, Jul 2001.

[142] S. Nokkala and J. Puro. Cytological evidence for a chromocenter in Drosophila melanogaster oocytes. *Hereditas*, 83(2):265–268, 1976.

[143] Elphège P Nora, Job Dekker, and Edith Heard. Segmental folding of chromosomes: a basis for structural and regulatory chromosomal neighborhoods? *BioEssays : news and reviews in molecular, cellular and developmental biology*, 35(9):818–28, September 2013. ISSN 1521-1878. doi: 10.1002/bies.201300040. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3874840&tool=pmcentrez&rendertype=abstract`.

[144] T. Obayashi, S. Hayashi, M. Shibaoka, M. Saeki, H. Ohta, and K. Kinoshita. COXPRESdb: a database of coexpressed gene networks in mammals. *Nucleic Acids Res.*, 36(Database issue): 77–82, Jan 2008.

[145] Joshua O'Madadhain, Danyel Fisher, Scott White, Padhraic Smyth, and Yan biao Boey. Analysis and visualization of network data using jung.

[146] L. M. Ooms, K. A. Horan, P. Rahman, G. Seaton, R. Gurung, D. S. Kethesparan, and C. A. Mitchell. The role of the inositol polyphosphate 5-phosphatases in cellular function and human disease. *Biochem. J.*, 419(1):29–49, Apr 2009.

[147] D. O'Reilly and D. R. Greaves. Cell-type-specific expression of the human CD68 gene is associated with changes in Pol II phosphorylation and short-range intrachromosomal gene looping. *Genomics*, 90:407–415, Sep 2007.

[148] C. S. Osborne and C. H. Eskiw. Where shall we meet? A role for genome organisation and nuclear sub-compartments in mediating interchromosomal interactions. *J. Cell. Biochem.*, 104:1553–1561, Aug 2008.

[149] C. S. Osborne, L. Chakalova, K. E. Brown, D. Carter, A. Horton, E. Debrand, B. Goyenechea, J. A. Mitchell, S. Lopes, W. Reik, and P. Fraser. Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat. Genet.*, 36:1065–1071, Oct 2004.

[150] C. S. Osborne, P. A. Ewels, and A. N. Young. Meet the neighbours: tools to dissect nuclear structure and function. *Brief Funct Genomics*, 10:11–17, Jan 2011.

[151] J. M. O'Sullivan, S. M. Tan-Wong, A. Morillon, B. Lee, J. Coles, J. Mellor, and N. J. Proudfoot. Gene loops juxtapose promoters and terminators in yeast. *Nat. Genet.*, 36:1014–1018, Sep 2004.

[152] H. Pages. *BSgenome: Infrastructure for Biostrings-based genome data packages*. R package version 1.28.0.

[153] H. Pages, P. Aboyoun, R. Gentleman, and S. DebRoy. *Biostrings: String objects representing biological sequences, and matching algorithms*. R package version 2.28.0.

[154] K. C. Pang, M. C. Frith, and J. S. Mattick. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet.*, 22:1–5, Jan 2006.

[155] L. A. Parada, P. G. McQueen, P. J. Munson, and T. Misteli. Conservation of relative chromosome positioning in normal and cancer cells. *Curr. Biol.*, 12:1692–1697, Oct 2002.

[156] L. A. Parada, P. G. McQueen, and T. Misteli. Tissue-specific spatial organization of genomes. *Genome Biol.*, 5:R44, 2004.

[157] V. Parelho, S. Hadjur, M. Spivakov, M. Leleu, S. Sauer, H. C. Gregson, A. Jarmuz, C. Canzonetta, Z. Webster, T. Nesterova, B. S. Cobb, K. Yokomori, N. Dillon, L. Aragon, A. G. Fisher, and M. Merkenschlager. Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell*, 132:422–433, Feb 2008.

[158] J. C. Pearson, D. Lemons, and W. McGinnis. Modulating Hox gene functions during animal body patterning. *Nat. Rev. Genet.*, 6(12):893–904, Dec 2005.

[159] D. Peric-Hupkes, W. Meuleman, L. Pagie, S. W. Bruggeman, I. Solovei, W. Brugman, S. Graf, P. Flicek, R. M. Kerkhoven, M. van Lohuizen, M. Reinders, L. Wessels, and B. van Steensel. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol. Cell*, 38:603–613, May 2010.

[160] P. Pevzner and G. Tesler. Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res.*, 13(1):37–45, Jan 2003.

[161] S. Pfeffer, M. Zavolan, F. A. Grasser, M. Chien, J. J. Russo, J. Ju, B. John, A. J. Enright, D. Marks, C. Sander, and T. Tuschl. Identification of virus-encoded microRNAs. *Science*, 304:734–736, Apr 2004.

[162] J.E. Philips and V.G. Corces. CTCF: Master Weaver of the Genome. *Cell*, 137:1194–1211, 2009.

[163] K. Plath, S. Mlynarczyk-Evans, D. A. Nusinow, and B. Panning. Xist RNA and the mechanism of X chromosome inactivation. *Annu. Rev. Genet.*, 36:233–278, 2002.

[164] S. Proost, J. Fostier, D. De Witte, B. Dhoedt, P. Demeester, Y. Van de Peer, and K. Vandepoele. i-ADHoRe 3.0–fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res.*, 40(2):e11, Jan 2012.

[165] F. Provost. Machine Learning from Imbalanced Data Sets 101. *Invited paper for the AAAI 2000 Workshop on Imblanced Data Sets*, 2000.

[166] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org/`.

[167] C. Rabl. Über Zelltheilung. *Morphol. Jahrb.*, 10:214–330, 1885.

[168] M. A. Reeves, F. P. Bellinger, and M. J. Berry. The neuroprotective functions of selenoprotein M and its role in cytosolic

calcium regulation. *Antioxid. Redox Signal.*, 12(7):809–818, Apr 2010.

[169] B. J. Reinhart, F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horvitz, and G. Ruvkun. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, 403:901–906, Feb 2000.

[170] S. Rensen, G. Merkx, P. Doevendans, A. Geurts Van Kessel, and G. van Eys. Structure and chromosome location of Smtn, the mouse smoothelin gene. *Cytogenet. Cell Genet.*, 89(3-4):225–229, 2000.

[171] K. Rieck and P. Laskov. Linear-Time Computation of Similarity Measures for Sequential Data. *J Mach Learn Res*, 9, 2008.

[172] J. Rinn and M. Guttman. RNA Function. RNA and dynamic nuclear organization. *Science*, 345(6202):1240–1241, Sep 2014.

[173] C. Rodelsperger and C. Dieterich. CYNTENATOR: progressive gene order alignment of 17 vertebrate genomes. *PLoS ONE*, 5 (1):e8861, 2010.

[174] E. D. Rubio, D. J. Reiss, P. L. Welcsh, C. M. Disteche, G. N. Filippova, N. S. Baliga, R. Aebersold, J. A. Ranish, and A. Krumm. CTCF physically links cohesin to chromatin. *Proc. Natl. Acad. Sci. U.S.A.*, 105:8309–8314, Jun 2008.

[175] G. Ruvkun. Molecular biology. Glimpses of a tiny RNA world. *Science*, 294:797–799, Oct 2001.

[176] P. J. Sabo, M. Hawrylycz, J. C. Wallace, R. Humbert, M. Yu, A. Shafer, J. Kawamoto, R. Hall, J. Mack, M. O. Dorschner, M. McArthur, and J. A. Stamatoyannopoulos. Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proc. Natl. Acad. Sci. U.S.A.*, 101(48):16837–16842, Nov 2004.

[177] P. J. Sabo, M. S. Kuehn, R. Thurman, B. E. Johnson, E. M. Johnson, H. Cao, M. Yu, E. Rosenzweig, J. Goldy, A. Haydock, M. Weaver, A. Shafer, K. Lee, F. Neri, R. Humbert, M. A. Singer, T. A. Richmond, M. O. Dorschner, M. McArthur, M. Hawrylycz, R. D. Green, P. A. Navas, W. S. Noble, and J. A. Stamatoyannopoulos. Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat. Methods*, 3(7):511–518, Jul 2006.

[178] L. Sagan. On the origin of mitosing cells. *J. Theor. Biol.*, 14: 255–274, Mar 1967.

[179] A. Sanyal, D. Bau, M. A. Marti-Renom, and J. Dekker. Chromatin globules: a common motif of higher order chromosome structure? *Curr. Opin. Cell Biol.*, 23:325–331, Jun 2011.

[180] T. Sazuka, Y. Tomooka, Y. Ikawa, M. Noda, and S. Kumar. DRG: a novel developmentally regulated GTP-binding protein. *Biochem. Biophys. Res. Commun.*, 189(1):363–370, Nov 1992.

[181] S. Schoenfelder, T. Sexton, L. Chakalova, N. F. Cope, A. Horton, S. Andrews, S. Kurukuti, J. A. Mitchell, D. Umlauf, D. S. Dimitrova, C. H. Eskiw, Y. Luo, C. L. Wei, Y. Ruan, J. J. Bieker, and P. Fraser. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat. Genet.*, 42:53–61, Jan 2010.

[182] E. Schrock, S. du Manoir, T. Veldman, B. Schoell, J. Wienberg, M. A. Ferguson-Smith, Y. Ning, D. H. Ledbetter, I. Bar-Am, D. Soenksen, Y. Garini, and T. Ried. Multicolor spectral karyotyping of human chromosomes. *Science*, 273(5274):494–497, Jul 1996.

[183] V.C. Seitan and M. Merkenschlager. Cohesin and chromatin organisation. *Current opinion in genetics & development*, 22(2): 93–100, 2011.

[184] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11):2498–2504, Nov 2003.

[185] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, 29(1):308–311, Jan 2001.

[186] G. Shi, L. Zhang, and T. Jiang. MSOAR 2.0: Incorporating tandem duplications into ortholog assignment based on genome rearrangement. *BMC Bioinformatics*, 11:10, 2010.

[187] M. D. Simon, S. F. Pinter, R. Fang, K. Sarma, M. Rutenberg-Schoenberg, S. K. Bowman, B. A. Kesner, V.K. Maier, R.E. Kingston, and J. T. Lee. High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. *Nature*, 504:465–9, 2014.

[188] M. Simonis, P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel, and W. de Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.*, 38:1348–1354, Nov 2006.

[189] B. N. Singh and M. Hampsey. A transcription-independent role for TFIIB in gene looping. *Mol. Cell*, 27:806–816, Sep 2007.

[190] W. C. Skarnes, B. Rosen, A. P. West, M. Koutsourakis, W. Bushell, V. Iyer, A. O. Mujica, M. Thomas, J. Harrow, T. Cox, D. Jackson, J. Severin, P. Biggs, J. Fu, M. Nefedov, P. J. de Jong, A. F. Stewart, and A. Bradley. A conditional knockout resource for the genome-wide study of mouse gene function. *Nature*, 474 (7351):337–342, Jun 2011.

[191] A.F.A. Smit, R. Hubley, and P. Green. Repeatmasker open - 3.0, 1996-2010. `http://www.repeatmasker.org`.

[192] S.N. Soffer and A. Vazquez. Clustering coefficient without degree correlations biases. *Phys. Rev.*, 71, 2005.

[193] S. Sofueva and S. Hadjur. Cohesin-mediated chromatin interactions–into the third dimension of gene regulation. *Briefings in Functional Genomics*, 11:205–216, 2012.

[194] I. Solovei, A. Cavallo, L. Schermelleh, F. Jaunin, C. Scasselati, D. Cmarko, S. Fakan, and T. Cremer. Spatial preservation of nuclear chromatin architecture during three-dimensional fluorescence in situ hybridization (3D-FISH). *Exp Cell Res*, 276:10–23, 2002.

[195] T. D. Southall and A. H. Brand. Chromatin profiling in model organisms. *Brief Funct Genomic Proteomic*, 6:133–140, Jun 2007.

[196] M. R. Speicher and N. P. Carter. The new cytogenetics: blurring the boundaries with molecular biology. *Nat. Rev. Genet.*, 6(10): 782–792, Oct 2005.

[197] E. Splinter, H. Heath, J. Kooren, R. J. Palstra, P. Klous, F. Grosveld, N. Galjart, and W. de Laat. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes Dev.*, 20(17):2349–2354, Sep 2006.

[198] W. Stedman, H. Kang, S. Lin, J. L. Kissil, M. S. Bartolomei, and P. M. Lieberman. Cohesins localize with CTCF at the KSHV latency control region and at cellular c-myc and H19/Igf2 insulators. *EMBO J.*, 27:654–666, Feb 2008.

[199] K. Struhl. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat. Struct. Mol. Biol.*, 14:103–105, Feb 2007.

[200] Y. Taketomi, N. Ueno, T. Kojima, H. Sato, R. Murase, K. Yamamoto, S. Tanaka, M. Sakanaka, M. Nakamura, Y. Nishito, M. Kawana, N. Kambe, K. Ikeda, R. Taguchi, S. Nakamizo, K. Kabashima, M. H. Gelb, M. Arita, T. Yokomizo, M. Nakamura, K. Watanabe, H. Hirai, M. Nakamura, Y. Okayama, C. Ra,

K. Aritake, Y. Urade, K. Morimoto, Y. Sugimoto, T. Shimizu, S. Narumiya, S. Hara, and M. Murakami. Mast cell maturation is driven via a group III phospholipase A2-prostaglandin D2-DP1 receptor paracrine axis. *Nat. Immunol.*, 14(6):554–563, Jun 2013.

[201] S. M. Tan-Wong, J. D. French, N. J. Proudfoot, and M. A. Brown. Dynamic interactions between the promoter and terminator regions of the mammalian BRCA1 gene. *Proc. Natl. Acad. Sci. U.S.A.*, 105:5160–5165, Apr 2008.

[202] S. M. Tan-Wong, H. D. Wijayatilake, and N. J. Proudfoot. Gene loops function to maintain transcriptional memory through interaction with the nuclear pore complex. *Genes Dev.*, 23:2610–2624, Nov 2009.

[203] F. Tang, B. Wang, N. Li, Y. Wu, J. Jia, T. Suo, Q. Chen, Y. J. Liu, and J. Tang. RNF185, a novel mitochondrial ubiquitin E3 ligase, regulates autophagy through interaction with BNIP1. *PLoS ONE*, 6(9):e24367, 2011.

[204] G. Tesler. GRIMM: genome rearrangements web server. *Bioinformatics*, 18(3):492–493, Mar 2002.

[205] J. D. Thompson, T. J. Gibson, and D. G. Higgins. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics*, Chapter 2:Unit 2.3, Aug 2002.

[206] B. Tolhuis, R. J. Palstra, E. Splinter, F. Grosveld, and W. de Laat. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol. Cell*, 10:1453–1465, Dec 2002.

[207] M. C. Tsai, O. Manor, Y. Wan, N. Mosammaparast, J. K. Wang, F. Lan, Y. Shi, E. Segal, and H. Y. Chang. Long noncoding RNA as modular scaffold of histone modification complexes. *Science*, 329(5992):689–693, Aug 2010.

[208] J. Tsuji, M. C. Frith, K. Tomii, and P. Horton. Mammalian NUMT insertion is non-random. *Nucleic Acids Res.*, 40(18):9073–9088, Oct 2012.

[209] T. Valentino, D. Palmieri, M. Vitiello, A. Simeone, G. Palma, C. Arra, P. Chieffi, L. Chiariotti, A. Fusco, and M. Fedele. Embryonic defects and growth alteration in mice with homozygous disruption of the Patz1 gene. *J. Cell. Physiol.*, 228(3):646–653, Mar 2013.

[210] B. van Steensel and S. Henikoff. Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat. Biotechnol.*, 18:424–428, Apr 2000.

[211] A. J. Vilella, J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, and E. Birney. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, 19(2):327–335, Feb 2009.

[212] H. Wainer, M. Gessaroli, and M. Verdi. Finding What Is Not There through the Unfortunate Binning of Results: The Mendel Effect. *Visual Revelations, CHANCE*, 19(1):49–52, 2006.

[213] I.C. Waizenegger, S. Hauf, A. Meinke, and J.M. Peters. Two distinct pathways remove mammalian cohesin from chromosome arms in prophase and from centromeres in anaphase. *Cell*, 103: 399–410, 2000.

[214] J. Walter, B. Joffe, A. Bolzer, H. Albiez, P. A. Benedetti, S. Muller, M. R. Speicher, T. Cremer, M. Cremer, and I. Solovei. Towards many colors in FISH on 3D-preserved interphase nuclei. *Cytogenet. Genome Res.*, 114(3-4):367–378, 2006.

[215] K. C. Wang, Y. W. Yang, B. Liu, A. Sanyal, R. Corces-Zimmerman, Y. Chen, B. R. Lajoie, A. Protacio, R. A. Flynn, R. A. Gupta, J. Wysocka, M. Lei, J. Dekker, J. A. Helms, and H. Y. Chang. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature*, 472(7341): 120–124, Apr 2011.

[216] Y. Wang, H. Tang, J. D. Debarry, X. Tan, J. Li, X. Wang, T. H. Lee, H. Jin, B. Marler, H. Guo, J. C. Kissinger, and A. H. Paterson. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.*, 40(7):e49, Apr 2012.

[217] Z. Wang, C. Zang, J. A. Rosenfeld, D. E. Schones, A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, W. Peng, M. Q. Zhang, and K. Zhao. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.*, 40(7):897–903, Jul 2008.

[218] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, Jun 1998.

[219] N. Weddington, A. Stuy, I. Hiratani, T. Ryba, T. Yokochi, and D. M. Gilbert. ReplicationDomain: a visualization tool and comparative database for genome-wide replication timing data. *BMC Bioinformatics*, 9:530, 2008.

[220] D. M. Wellik. Hox patterning of the vertebrate axial skeleton. *Dev. Dyn.*, 236(9):2454–2463, Sep 2007.

[221] K. S. Wendt, K. Yoshida, T. Itoh, M. Bando, B. Koch, E. Schirghuber, S. Tsutsumi, G. Nagae, K. Ishihara, T. Mishiro, K. Yahata, F. Imamoto, H. Aburatani, M. Nakao, N. Imamoto, K. Maeshima, K. Shirahige, and J. M. Peters. Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature*, 451:796–801, Feb 2008.

[222] D. M. Witten and W. S. Noble. On the assessment of statistical significance of three-dimensional colocalization of sets of genomic elements. *Nucleic Acids Res.*, 40(9):3849–3855, May 2012.

[223] L. "Wolpert, R. Beddington, J. Brockes, T.M. Jessel, P.A. Lawrence, and E." Meyerowitz. *Principles of Development*. Oxford Univ. Press, Oxford, England, 1998.

[224] H. Wurtele and P. Chartrand. Genome-wide scanning of HoxB1-associated loci in mouse ES cells using an open-ended Chromosome Conformation Capture methodology. *Chromosome Res.*, 14(5):477–495, 2006.

[225] A. Wutz and J. Gribnau. X inactivation Xplained. *Curr. Opin. Genet. Dev.*, 17:387–393, Oct 2007.

[226] A. Wutz, O. W. Smrzka, N. Schweifer, K. Schellander, E. F. Wagner, and D. P. Barlow. Imprinted expression of the Igf2r gene depends on an intronic CpG island. *Nature*, 389(6652):745–749, Oct 1997.

[227] N. Xu, C. L. Tsai, and J. T. Lee. Transient homologous chromosome pairing marks the onset of X inactivation. *Science*, 311: 1149–1152, Feb 2006.

[228] N. Xu, M. E. Donohoe, S. S. Silva, and J. T. Lee. Evidence that homologous X-chromosome pairing requires transcription and Ctcf protein. *Nat. Genet.*, 39:1390–1396, Nov 2007.

[229] P. Xu, S. Y. Vernooy, M. Guo, and B. A. Hay. The Drosophila microRNA Mir-14 suppresses cell death and is required for normal fat metabolism. *Curr. Biol.*, 13:790–795, Apr 2003.

[230] E. Yaffe and A. Tanay. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.*, 43(11):1059–1065, Nov 2011.

[231] Guangchuang Yu, Fei Li, Yide Qin, Xiaochen Bo, Yibo Wu, and Shengqi Wang. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26(7), 2010.

[232] T. M. Yusufzai, H. Tagami, Y. Nakatani, and G. Felsenfeld. CTCF tethers an insulator to subnuclear sites, suggesting

shared insulator mechanisms across species. *Mol. Cell*, 13:291–298, Jan 2004.

[233] A. O. Zalensky, M. J. Allen, A. Kobayashi, I. A. Zalenskaya, R. Balhorn, and E. M. Bradbury. Well-defined genome architecture in the human sperm nucleus. *Chromosoma*, 103(9):577–590, May 1995.

[234] H. Zayed, Z. Izsvak, O. Walisko, and Z. Ivics. Development of hyperactive sleeping beauty transposon vectors by mutational analysis. *Mol. Ther.*, 9(2):292–304, Feb 2004.

[235] Y. Zhang, R. P. McCord, Y. J. Ho, B. R. Lajoie, D. G. Hildebrand, A. C. Simon, M. S. Becker, F. W. Alt, and J. Dekker. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell*, 148(5):908–921, Mar 2012.

[236] J. Zhao, T. K. Ohsumi, J. T. Kung, Y. Ogawa, D. J. Grau, K. Sarma, J. J. Song, R. E. Kingston, M. Borowsky, and J. T. Lee. Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol. Cell*, 40:939–953, Dec 2010.

[237] Z. Zhao, G. Tavoosidana, M. Sjolinder, A. Gondor, P. Mariano, S. Wang, C. Kanduri, M. Lezcano, K. S. Sandhu, U. Singh, V. Pant, V. Tiwari, S. Kurukuti, and R. Ohlsson. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.*, 38:1341–1347, Nov 2006.

[238] C. Zimmer and E. Fabre. Principles of chromosomal organization: lessons from yeast. *J. Cell Biol.*, 192(5):723–733, Mar 2011.