UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE CIENCIAS BIOLÓGICAS

TESIS DOCTORAL

Modelado integrativo de la estructura 3D de macromoléculas

Integrative modeling of the 3D structure of macromolecules

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Ibai Irastorza Azcarate

Director

Damien P. Devos

Madrid 2018

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS BIOLÓGICAS

TESIS DOCTORAL

Modelado integrativo de la estructura 3D de macromoléculas

Integrative modeling of the 3D structure of macromolecules

MEMORIA PARA OPTAR AL GRADO DE DOCTOR
PRESENTADA POR
Ibai Irastorza Azcarate

Director
Damien P. Devos

Madrid, 2018

“Urrunago ikustea lortu badut, erraldoien sobaldetara igo naizelako da. “

Isaac Newton

Acknowledgements

There are many people that should be in this acknowledgements, because, somehow, I feel that all the people that I know and that I have met in these past four years have contributed, even if a little, in helping me become what I am, because this road has not been just about the work that I have done.

I will start with Damien. My sincerest gratitude to you, my mentor and friend, who chose me to start this journey into happiness and despair, despite not knowing what the microbiology platypus was. Thanks a lot for letting me work in what I wanted, for teaching me the differences between bacteria and eukarya, for showing me the Belgium beer world and for being in your office when needed. Thanks a lot for helping me become what I am now.

Many thanks to my labmates, Carlos and Elena, for being great friends and bearing with my mental disorders. Also to David, classmate and labmate for a short time. Thanks a lot to all the good people I met at work, Rafa, both Martas, Rocio, Laura, Sandra, Jose María, etc… for their friendship and all the exploding kitten games. I am very grateful to Silvia and Juan also, wonderful people that will always have a little place in my heart. Jose Luis deserves special thanks, my part time boss and full time friend, for teaching me so much and giving me the opportunity to be part of great projects. Thanks a lot also to all the people that collaborated with me and especially to Oriol, a great scientist and friend. I need to thank also to all the people in the CABD and the ones that on Tuesdays become professional football players. Thanks also to the RSG-Spain, great group of people and friends that are doing a great job for young scientists.

Mi paso por Sevilla ha sido corto pero intenso, y nunca olvidaré lo que he vivido aquí. He conocido gente maravillosa como Curro, que me ha enseñado a ser mejor persona y he hecho muy buenos amigos como Caitlin, Panos y Sole, que han hecho que me sienta en Sevilla, como en casa, que ya es decir, y siempre les recordaré. Quiero agradecer especialmente a Nico, mi compi de trabajo y pareja de yahtzee, pero sobretodo, un gran amigo. Agradecer también a Antonio y Roma, Leo y Toni, Emilio y Lili, Antonio y Jagger, y Pedro, que han contribuido más de lo que creen a que yo haya terminado esta tesis. Les agradezco también a Laura, Joaquín, Joaquinillo y Julián, que son unos amores y les tengo mucho cariño. Estoy muy contento también de haber conocido a David, mi paisano. Una persona maravillosa y un muy buen amigo. Pero a los que de verdad quiero dar las gracias es a Isabel y Nacho. Unas increíbles personas, pero sobre todo amigos. Les debo mucho a ellos. He crecido mucho con ellos, en lo científico y en lo personal y me han enseñado tanto, que estaré toda la vida en deuda. Muchas gracias. Gracias

también a Isabel, Boo, y Juan Carlos, que, aunque ya no esté en Sevilla, siempre le consideraré un buen amigo y le tengo mucho aprecio.

Mención aparte para Sigeco, mis hermanos, que siempre habéis estado ahí para apoyarme a las duras y a las maduras: Guiselle nuestro dios y creador, con permiso de Iñigod; Leander, Felix y Soto, la escisión alemana, mis futuros hospederos; Fran, Rafa y Juanlu, los biomonguers, que más me comprenden; Sebas, mi siempre compañero de piso; Isra, el hipster; Santy, el catalán; Juanin, ese pequeño gran hombre; Roca, el esclavista eta Marks, nire kuttuna. Os debo tanto, que, de hecho, si no fuera por Sigeco, no estaría trabajando en ciencia. Epiccc! Gracias también al resto de gente que he conocido en Madrid, pero sobre todo a Useros y Miguel, Irene, David y Charli gente preciosa y mejores amigos. Muchas gracias también a toda la familia de Carla, especialmente a Gladys y Eduardo, mi segunda familia, que siempre, siempre, me han ayudado y los quiero mucho.

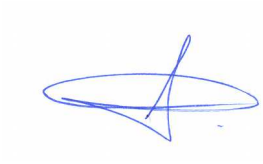
Eta azkenik, gertuen dazenei. Mila esker aita, ama ta Aitor, nire benetako bihotzak ta esker mila nire familixa osuai, edozein lekutara noiela, badakitx beti hor egongo zariela ta, nire onduen, laguntzen. Arrasateko kuadrilia bebai esker pilua. Nahiz ta oporretan bakarrik ikusi, betiko lagunak izen zarie ta hor egon zarie beti. Baster, Hodei, Txak, Arana, Elorza, Ametz, Beixa, Zaba, Martin, Lopez, Txiri, Xabi, Mikel Arana ta falta dienak.

Pero a los que más quiero agradecer es a vosotros, mis verdaderos amores, mis peques, mis bonitas, my moon and stars. Os dedico a vosotras esta tesis. Siempre en mi corazón, Argi, Txula y Carla. Os amo. <3

El Director, Don Damien Paul Devos, Doctor en Ciencias Biológicas y Científico Titular del Centro Andaluz de Biología del Desarrollo (CSIC) y el Tutor, Don Abel Sánchez Jiménez, Doctor en Neurociencias y Profesor en el Departamento de Matemática Aplicada en la Facultad de Biología de la Universidad Complutense de Madrid informan que:

La memoria titulada “Modelado Integrativo de la Estructura 3D de Macromoléculas / Integrative Modeling of the 3D Structure of Macromolecules” que presenta Ibai Irastorza Azcarate, para optar al Grado de Doctor, reune las condiciones exigidas por la legislación vigente y tiene la originalidad, el rigor y la calidad científica necesaria para ser presentada.

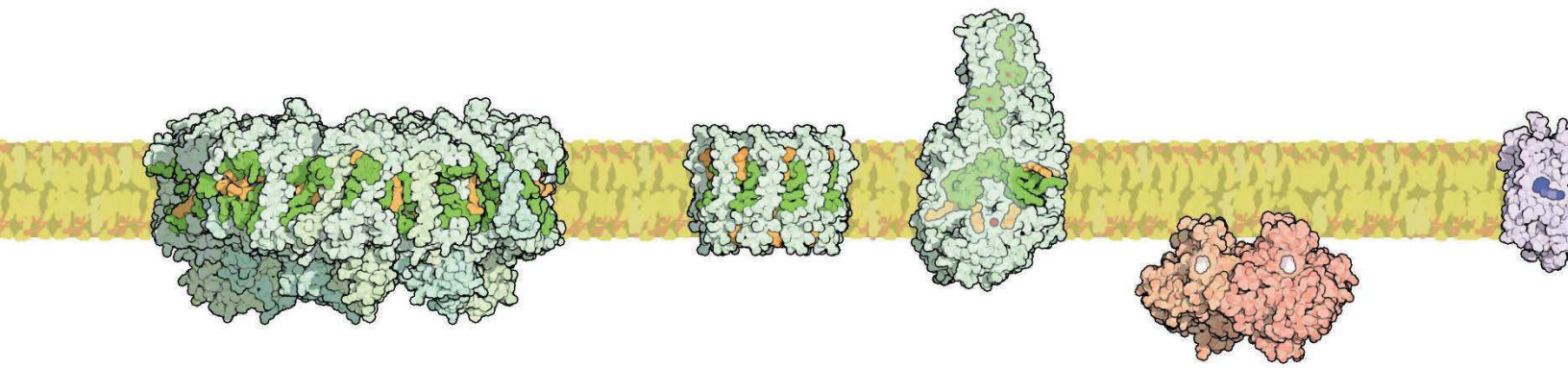
Madrid, 2018



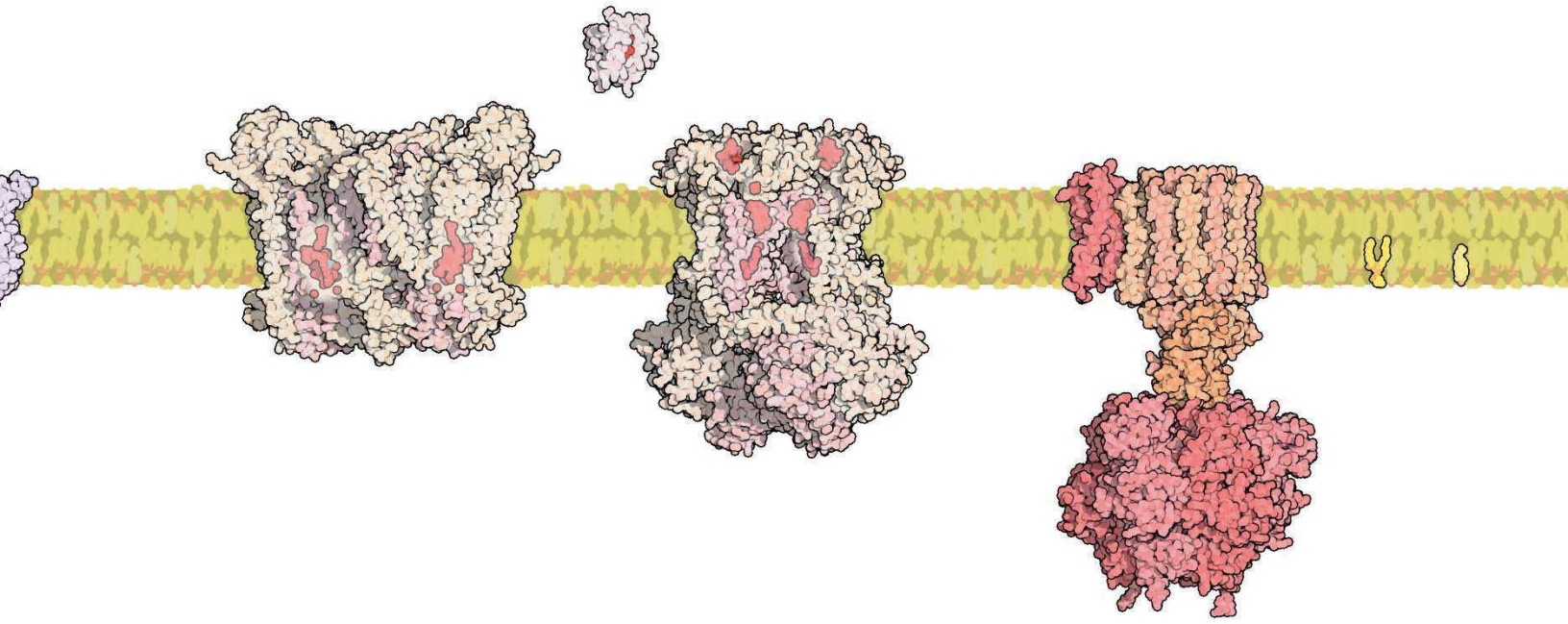
Fdo.: Damien Paul Devos

Fdo.: Abel Sánchez Jiménez

Table of Contents



Resumen/Summary



Resumen

El estudio de la estructura de las biomoléculas es fundamental para entender la vida desde un punto de vista molecular. Los métodos directos usados para resolver la estructura de complejos macromoleculares, como la cristalografía de rayos X, están llegando a su límite y el futuro se está abriendo paso para resolver dichas estructuras mediante la integración de datos provenientes de distintas técnicas.

El objetivo de esta Tesis ha sido desarrollar métodos integrativos para resolver estructuras de complejos macromoleculares. Se han desarrollado dos métodos, uno dirigido a determinar la conformación de complejos multiproteicos y otro para inferir la estructura de la cromatina de distintas regiones del genoma.

El primer método se utilizó para resolver la arquitectura del *exocyst*, un complejo multiproteico compuesto por 8 proteínas de forma alargada, responsable de fusionar vesículas secretoras a la membrana plasmática. El exocyst ha sido difícil de purificar hasta ahora, y, por ello, ha sido difícil resolver su estructura atómica mediante técnicas convencionales como cristalografía de rayos X o resonancia magnética nuclear. El método combina la información estructural de cada subunidad con las distancias entre ellas obtenidas mediante microscopía óptica. Estas distancias, medidas entre distintos fluoróforos fusionados a los extremos amino y carboxilo terminal de cada subunidad, se usan como restricciones espaciales para resolver la estructura. Nuestra herramienta, gracias a algoritmos optimizadores, genera modelos tridimensionales (3D) del exocyst que cumplen las restricciones impuestas. Mediante el análisis de la población de los mejores modelos, se propuso un modelo representativo.

La estructura más representativa del exocyst se asemeja a una mano abierta, donde las subunidades sobresalen desde el núcleo. 7 proteínas de las 8 totales contienen su extremo amino en el núcleo del complejo. El modelo alberga interacciones entre las subunidades que están respaldadas por interacciones directas entre proteínas encontradas en la literatura. La estructura atómica de la subunidad Exo70 encaja con la posición que alberga en nuestro modelo, no solo en tamaño, sino también en forma, validando de manera más sólida nuestro modelo. Adicionalmente, se modeló el exocyst junto con la proteína Sec2, la cual se localiza en la membrana de las vesículas. Nuestro modelo posiciona Sec2 a 50 nm de distancia del exocyst, el radio medio de una vesícula secretora y las subunidades más cercanas a Sec2 son Sec10 y Sec15, de acuerdo con su función, puesto que son las subunidades responsables de la fusión del exocyst con la vesícula. Además, demostramos que el exocyst es un complejo estable e hipotetizamos que varios exocyst trabajan conjuntamente para fusionar la vesícula, debido a que

encontramos una media de 14 copias del exocyst en las zonas de fusión de la membrana plasmática.

Nuestros resultados demuestran que la integración de múltiples datos ayuda a resolver estructuras de complejos multiproteicos y podría usarse en dichos casos donde las técnicas convencionales no son eficientes.

Pero no solo la estructura de las proteínas y complejos proteicos es importante, sino que, recientemente, se ha visto que la estructura juega un papel crítico en la cromatina y está relacionada con la expresión génica. A su vez, la estructura es importante para estudiar sus funciones, su implicación en enfermedades e inferir información evolutiva. Hay muchas formas de estudiar la estructura de la cromatina, pero en este caso hemos utilizado una aproximación similar a la utilizada en el exocyst, mediante datos de 4C-seq (*Circular chromosome conformation capture*). En este caso, desarrollamos una herramienta para predecir la conformación 3D de la cromatina de distintos loci genómicos. Para ello, representamos la cromatina como una concatenación flexible de esferas y calculamos la localización de estas esferas en el espacio 3D gracias a distancias obtenidas mediante 4C-seq, técnica que calcula la frecuencia de interacciones entre distintos fragmentos de ADN. La posición de las esferas se calcula integrando todas las distancias entre estas esferas, como restricciones espaciales. Después, estas restricciones se optimizan y se extraen los modelos 3D que cumplen la mayoría de las restricciones.

El método ha sido respaldado por los distintos casos en los que se ha aplicado. En uno de los casos se comparó la región Hox de vertebrados como pez cebra y ratón y la región Hox del anfioxo, un cordado invertebrado, para saber cómo se formó la configuración bipartita que poseen los vertebrados. La sintenia de la región Hox en diversas especies y nuestro hallazgo mostrando que el anfioxo no tiene una configuración bipartita, demuestran que es una innovación de los vertebrados. El método también fue aplicado en humanos con una translocación de los cromosomas 2 y 13, que genera un gen de fusión denominado PAX3:FOXO1, responsable de una enfermedad llamada rabdomiosarcoma alveolar. Nuestros resultados muestran que los elementos reguladores de la transcripción de FOXO1 se localizan cerca del promotor de PAX3, formando un nuevo paisaje regulador y des-regulando estos genes. Igualmente se empleó el método en la región Shh de ratones, para comparar dos modelos 3D de la región, una silvestre y la otra con una inversión y, por último, se generaron y se compararon modelos 3D de la región Six2/3 en ratón y pez cebra.

Esta herramienta está disponible para la comunidad científica y se ha demostrado que genera modelos 3D de confianza. Es capaz de generar mapas de contacto, similares a los

generados mediante Hi-C (*Chromosome conformation capture* aplicado a todo el genoma) derivados de los modelos 3D (Hi-C virtuales), que pueden ser usados para comparar estructuras entre diferente regiones o especies. Además, esta herramienta es útil a la hora de predecir estructuras 3D de regiones genómicas con variaciones estructurales o incluso cuando no se dispone del mapa de contacto Hi-C de una especie en particular.

Todo este trabajo demuestra que los métodos integrativos no son una alternativa si no un apoyo de gran utilidad cuando los experimentos convencionales no son eficientes.

Summary

Structure is key to a molecular understanding of life. This is why in the past 50 years or so, molecular biologists have solved so many structures of macromolecules. However, direct solving of the structure of macromolecular complexes by conventional methods such as x-ray crystallography, has shown its limits and the future is going towards the integration of diverse information to elucidate such structures.

The aim of these thesis has been to develop integrative methods to elucidate the structure of macromolecular complexes. Two methods have been developed, one aimed at modeling multi-protein structures and the other for chromatin structure.

The first method was applied to uncover the architecture of the exocyst, a multi-protein complex composed by 8 “rod-like” shaped subunits, which is responsible for binding secretory vesicles to the plasma membrane. The exocyst has been difficult to purify, resulting in difficulties to elucidate its atomic structure by mainstream methods such as X-ray crystallography or nuclear magnetic resonance (NMR). The method we developed combines structural information of each subunit with distances between subunits termini derived from light microscopy. These distances were measured between different fluorophores fused to the N and C termini of each subunit, which were afterwards used as spatial restraints towards solving the structure. Our tool then integrates these data and uses optimization algorithms to generate 3D models of the complex. Analysis of the population of models that best fulfill the input data allowed us to propose a representative structure.

The exocyst is shaped like an open hand, were the subunits protrude from the core of the complex to the exterior. Seven out of the eight subunits have the N terminus in the core of the complex. The model is supported by direct protein-protein interactions found in literature that were also found to occur in our model. The crystal structure of the Exo70 subunit fits well in our model, not only by size but also by shape, further supporting the model. We also modelled the exocyst together with Sec2, a protein that is localized at the membrane of the vesicle. Our models locate Sec2 50 nm away from the exocyst, the average radius of a secretory vesicle, and the subunits Sec10 and Sec15 are the ones closer to the vesicle, which is in agreement with their function as responsibles for the vesicle binding. In addition, in agreement with experimental evidence, we showed that the exocyst is a stable complex and we even hypothesize that there could be many exocyst cooperating since we measured an average of 14 exocyst copies on the vesicle fusion sites.

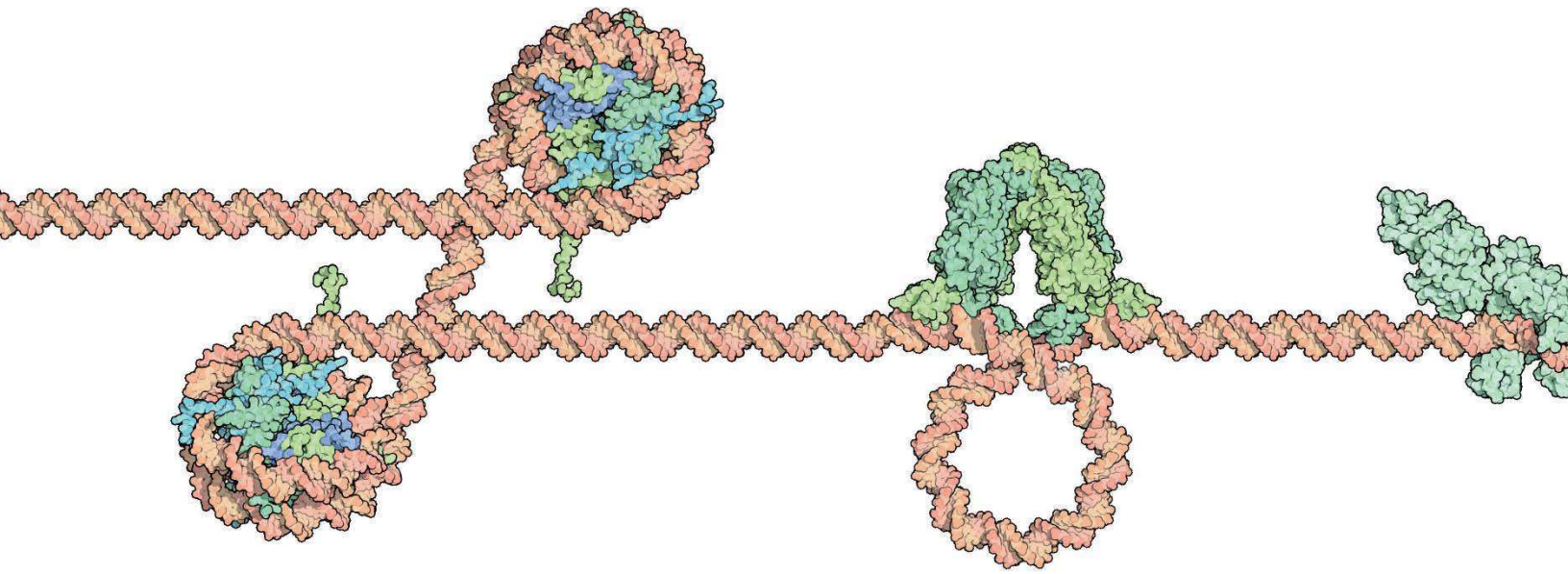
Our results show that the integration of data from different sources can help elucidating the structure of protein complexes and aims to cover the gap generated by main experiments and methods in protein structure resolution.

But not only protein structure is essential. Recently, it has been shown that chromatin structure is as important as protein structure and has an essential role in gene regulation. Chromatin structure is also important to study its function, its role in diseases and to infer evolutionary information. There are many ways to study chromatin structure, but in this case we have used a similar approach to the one used with the exocyst, using 4C-seq (Circular chromosome conformation capture) data. For that purpose, we developed a method to elucidate the three-dimensional (3D) chromatin conformation of different genomic loci. In this tool, the chromatin is represented as a flexible string of beads and distances between them are derived from 4C-seq data, which give the average frequency of contacts between binned fragments of our DNA loci. The localization of the beads is then optimized by integrative considerations of all distances. After an exhaustive optimization, the chromatin models that fulfill most of the restraints are used as representative solutions.

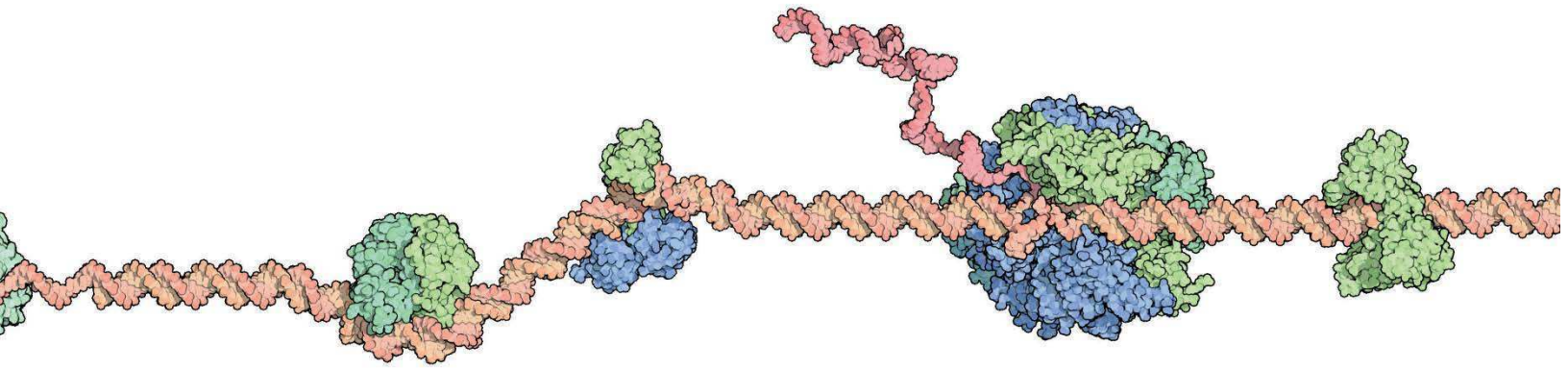
The method has been supported in each of the cases in which it has been applied. These include the case of the Hox cluster where we compared its organization between two vertebrates (mouse and zebrafish) and a non-vertebrate chordate (amphioxus), in order to investigate the evolution of the bipartite configuration of the vertebrate Hox cluster. Gene synteny comparison of other species around the Hox cluster and our findings showing that the amphioxus does not have a bipartite configuration, both indicate that this bipartition is a vertebrate innovation. We also applied the method to a human chromosomal translocation between chromosomes 2 and 13 that generates a fusion gene called PAX3:FOXO1, which is responsible for producing a type of cancer called alveolar rhabdomyosarcoma. Our results show that the enhancers responsible for FOXO1 regulation are now in close contact to the promoter of PAX3, forming a novel regulatory landscape and generating a misregulation of these genes. We further applied the method to generate 3D models of the Shh locus in mouse when an inversion was induced and to generate and compare models of the Six2/3 region in mouse and zebrafish.

The tool is available for the community and has been shown to generate reliable models. A contact map of the region, similar to a Hi-C (Genome wide chromosome conformation capture) map, and called virtual Hi-C, can be derived from the 3D models, which can be used to compare against other species or regions. In addition, it is useful to elucidate the 3D chromatin architecture of particular regions like the ones that have genomic structural variations or when no Hi-C contact maps of the genome of the species are available.

Altogether, our work shows that integrative methods are not just an alternative but a support for structural biologists and can be very useful to overcome particular problems that classical experiments cannot.



Introduction



1 Macromolecules

Life is based on molecules, such as carbohydrates, lipids, amino-acids or nucleotides. These entities are not working in isolation in the cell, but in the forms of multimers or macromolecules, such as proteins or nucleic acids. These macromolecules are essential biological entities for cell functioning.

The sequence of monomers in a macromolecule determines its three-dimensional (3D) structure and it is this 3D structure which is ultimately responsible for its function. The 3D structure that the macromolecule adopts is due to weak bonds between different atoms of the monomers composing its sequence. Similarly, macromolecules can interact with others in macromolecular complexes determining new functions. Thus, it is important to get access to this structure to understand the inner workings of living cells. For instance, the resolution of the 3D structure of the DNA double helix has been transforming for biology, immediately revealing mechanisms of duplication and transmission of the genetic information (Watson and Crick, 1953; Wilkins et al.,1953; Franklin and Gosling, 1953) (Figure 1A). Similarly, the resolution of the 3D structure of the first protein, myoglobin, in 1958, paved the way for a molecular understanding of respiration (Kendrew et al., 1958) (Figure 1B and 1C).

In this work, we focused on the determination of the 3D structure of multi protein complexes and genome loci using bioinformatic integrative approaches. Moreover, we will discuss the importance that the conformation and architecture of these macromolecules have in many aspects of biology.

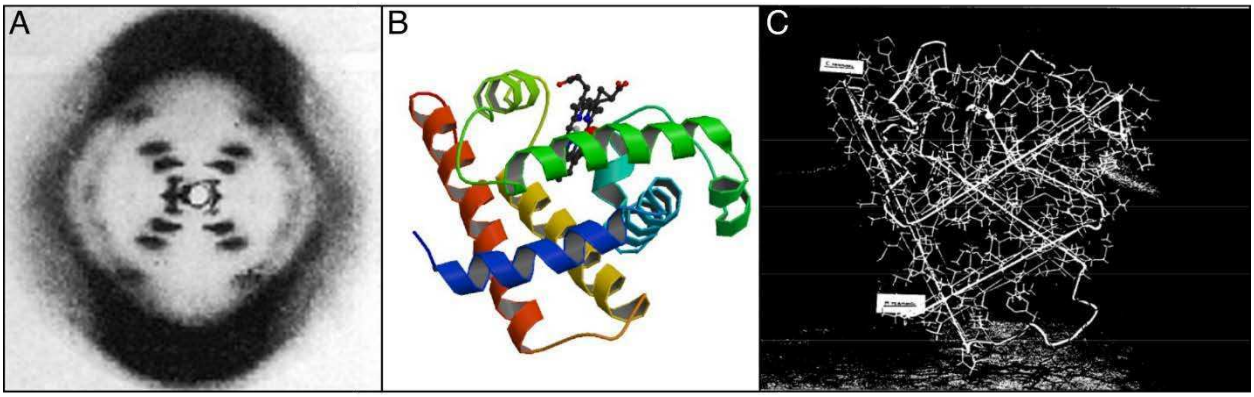


Figure 1. DNA and Myoglobin. (A) X-rays picture that was used to elucidate the structure of the DNA (Franklin and Gosling, 1953). (B) Structure of the Myoglobin (pdb 1MBN). (C) Model of the Myoglobin at 2Å resolution. The white cord follows the course of the polypeptide chain; the iron atom is indicated by a grey sphere, and its associated water molecule by a white sphere (Kendrew et al., 1958).

2 Proteins

2.1 Protein structure

Proteins are molecules composed of many types of smaller building blocks called amino acids. There are 20 different types of natural amino acids and these amino acids are linked together by a strong and stable covalent bond called the peptide bond. The arrangement of these amino acids in the polypeptide chain determines the 3D structure that the protein adopts. There are four levels of organization in the structure of proteins (Figure 2). The first level is the primary structure of a protein, the sequence of amino acid residues in the polypeptide chain. The secondary structure is called the local conformation of this chain, formed by hydrogen bonds that occur between the nearby amino acids. They create simple shapes like helices, strands, turns or loops. The most frequent secondary structures are α helices and β sheets. The polypeptide chain can fold further, compacting and adopting a 3D conformation, called tertiary structure. This happens due to the interactions between amino acids that are far apart in the primary structure. This organization level is generally the most stable, the one that is closest to the minimal free energy. Most proteins are longer than 300 amino acids and portions or fragments of this chain can fold into discrete modules called domains that do not depend on any of the remaining parts of the protein for their stability. The final structural level is the quaternary structure, which is composed by the interaction of various polypeptide chains.

Proteins are essential macromolecules due to their function which is determined by their structure. This is the basic idea of the structure-function paradigm, and, to understand the function of proteins at a molecular level, it is mandatory to determine their 3D structure (Baker & Sali, 2001; Robinson et al., 2007; Sali el al., 2003). But the determination of the 3D structures of proteins can help us in many other areas too. For instance, the structural comparisons of homologous proteins have helped infer evolutionary information and led to the conclusion that structure is more conserved than sequence (Chothia & Lesk, 1986; Sander & Schneider, 1991; Wood & Pearson, 1999). In addition, structural studies have played major roles in evolutionary (Löwe & Amos, 1998; Nogales el al., 1998), medical (Navia et al., 1989) and technological research (Jinek et al., 2014; Kim et al., 1995; Ormo et al., 1996).

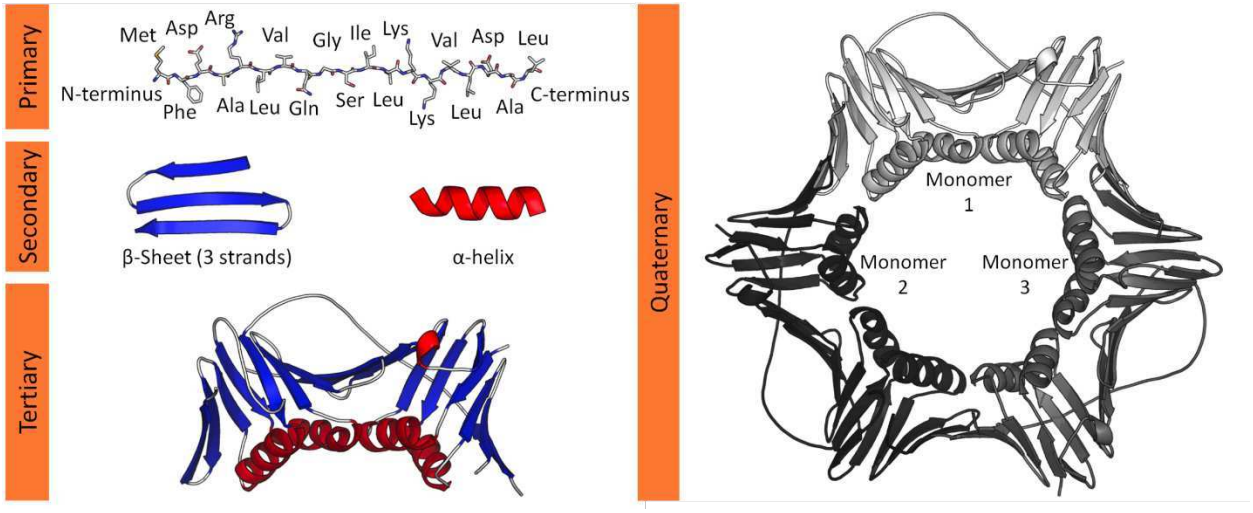


Figure 2. Levels of protein structure. Primary, secondary, tertiary and quaternary organization levels. Source: Wikimedia (commons.wikimedia.org)

2.2 Experimental techniques to determine protein structure

The Myoglobin was the first protein structure to be elucidated back in 1958 by John Kendrew and his colleagues using X-ray crystallography. It was registered in 1976 in the protein database (PDB), repository of all protein structures that has unified the format of these files for further use in science. Many methods are being used nowadays to determine the structure of proteins but the most used ones are derived from X-ray Crystallography (X-ray), Nuclear Magnetic Resonance (NMR) Spectroscopy and Cryo-Electron Microscopy (CryoEM), that together have elucidated more than 99% of the protein structures. These methods provide the experimental data that will be used afterwards to elucidate the 3D structure of the macromolecules. These methods are briefly explained in the next lines:

2.2.1 X-ray Crystallography:

In order to observe proteins at the atomic level, X-ray crystallography uses crystallized samples of purified proteins in high concentrations. These crystals are exposed to X-ray beams which are diffracted into many directions expressing patterns that afterwards can be processed. Analyzing the electron density, the position of the atoms composing the proteins can be derived. The structure can be refined afterwards favoring thermodynamic laws. It is the method that achieves best resolution with no theoretical size limitation. However, crystallography determines a single

conformation of the proteins and we cannot examine their motion which is important given the flexible nature of these. The quality of the final 3D structure is dependent on the generated crystal, which needs to be formed by an accurate and homogeneous alignment of multiple molecules. Nowadays, X-ray crystallography is the most used method to determine structure of proteins, with more than 90% of them determined in the PDB.

2.2.2 NMR Spectroscopy

The first protein structure was determined in 1984 using NMR spectroscopy. Now is the second most used method after X-ray crystallography with the determination of 8% of protein structures. NMR spectroscopy allows structure determination in the solution phase, allowing to examine the dynamics of the molecules, but in contrast, proteins need to have high solubility and not to aggregate. To get the experimental data, the sample is placed inside a powerful magnet that sends radio frequency signals through the sample and then, the absorption of these signals are measured. These information is used to determine the distance between different atom nuclei which will be used next to determine the overall structure. One of it disadvantages is that the method can only be applied to proteins in the size range between 5 and 30 kDa, although there are some cases of solved structures of proteins of around 100 kDa. In contrast, this method provides very good resolution (2-5 Å) and lets us examine disordered and static regions of the protein.

2.2.3 Transmission electron microscopy (TEM) methods: Cryo-Em.

CryoEM is the most used method between the TEM methods. Even though the origin of Cryo-EM dates back to 1984 (Adrian et al., 1984), it is only in the last 5 years that this method has become popular. So far, Cryo-EM has determined the structure of 1% of the proteins, but the number of proteins elucidated with this method has almost doubled between 2015 and 2016. This success is mostly due to the fact that the proteins do not need to be stained or fixed in any way, allowing their examination in their native environment. The aqueous biological sample is frozen rapidly and irradiated with a beam of electrons. Then, a detector senses how the electrons are scattered and, finally, a computer reconstructs the 3D structure of the molecule. Although Cryo-EM only works for samples that have several hundreds of kDas and the resolution is not as high as NMR spectroscopy or X-ray crystallography, it is a very good choice to determine protein complexes and large heterogeneous assemblies because the proteins are in their native state. In addition, these method is improving rapidly and in recent years, protein structures of resolutions below 3 Å have been determined, making this method potentially superior to traditional x-ray crystallography.

2.3 Comparative modeling as a protein structure predictor

Since the creation of PDB, more than 137.000 structures have been deposited and that number is growing every year. By contrast, from the 10,8 millions of proteins that are deposited in UniProt, the most popular protein sequence database, more than 555000 have been manually annotated and reviewed, of which almost 26.000 entries have a pdb structure associated (Figure 3). This gap between the number of protein sequences and the number of determined 3D structures is called the sequence-structure gap. To minimize this gap created by the impossibility to determine the structure of every single protein, scientists started to predict protein structures using computational comparative methods.

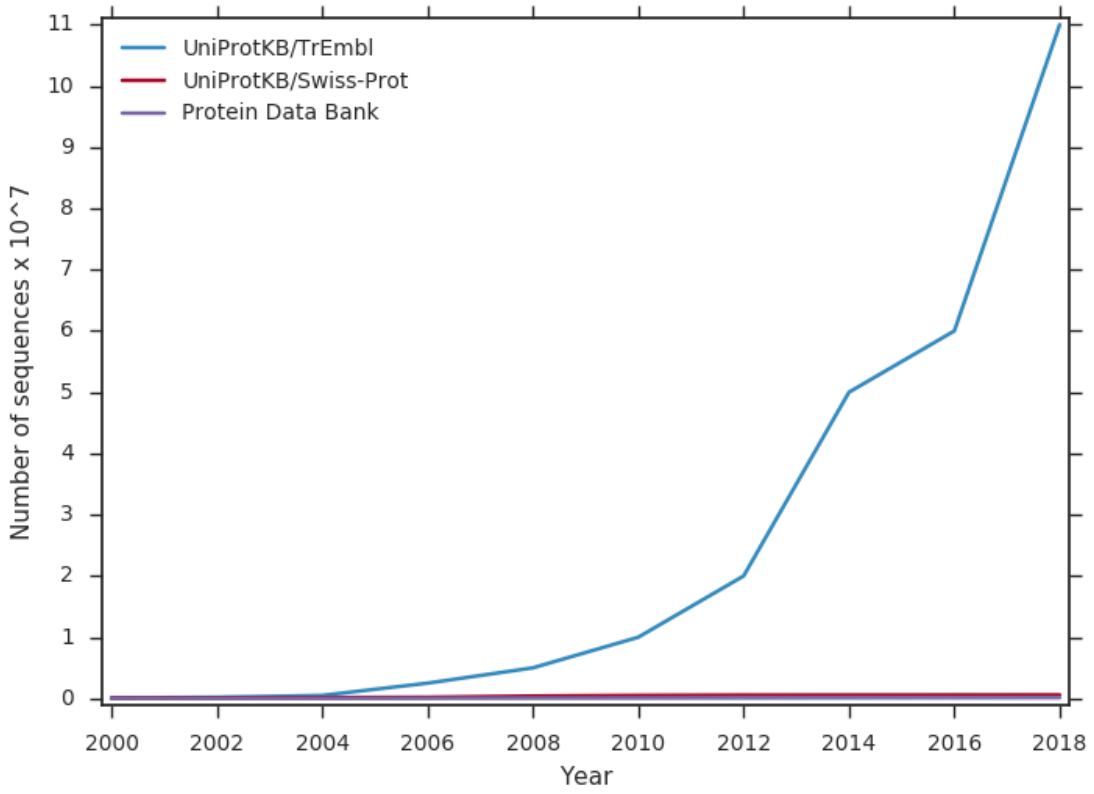


Figure 3. Year growth of protein structures and sequences. In blue, all protein sequences deposited in UniProt; red for manually annotated and reviewed protein sequences deposited in Swiss-Prot and, in purple, protein sequences with resolved structure deposited in the PDB. Data extracted from Uniprot (www.uniprot.org) and PDB (www.rcsb.org).

There are three main methods to predict protein structures: **Comparative modeling**, that uses the already determined structure of some proteins to predict the structure of the protein of interest; **threading and fold recognition** methods, useful when there are no similar proteins with determined structure and **de novo or ab initio** methods, that use bio physical properties of the amino acids to predict its shape, without using any evolutionary information between known structures. Among these three, comparative modeling is the most used method nowadays. The core of this method is the use of homologous proteins of known structure to infer the structure of the protein of interest, although there might be cases where these proteins are not homologous. Comparative modeling is based on the premise that during evolution, structure is more conserved than sequence and thus, evolutionary related proteins have similar 3D structures. The number of folds that a protein can adopt is very limited and, therefore, the space of possible structures is smaller than the space of sequences. On top of that, early and recent comparison of the structures of homologous proteins have shown that 3D structure is rarely affected by small changes in protein sequence. In fact, there are cases where proteins that have a sequence similarity of below 25% have similar structures. Thus, two homologous proteins have similar 3D structure (Figure 4). To predict structures with this method, we first identify a protein of known structure that will be used as a template. Then, we align their sequences and the model is built using the templates as a backbone. Finally the model is evaluated.

2.4 Multi protein complexes

Almost all biological processes depend upon proteins assembling into complexes (Marsh & Teichmann, 2015). Moreover, most proteins interact with other proteins to perform their functions (Alberts, 1998), as in the case of yeast proteins, where 80% of them interact at least with one partner (Gavin et al., 2006). These forms of quaternary structure play important roles in the cell and they are essential for its correct functioning. Some examples of multi protein complexes are the chaperon, the nuclear pore complex, the hemoglobin and the exocyst. The multi protein complexes, together with other molecules, can interact to form molecular machines like the RNA polymerase or the ribosome which are essential for life as we know it. Taking all these factors into account, it is of vital information, not only to determine the structure of single proteins, but also the structure of the complexes that they form.

In recent years, the CryoEM approaches have made enormous advances in the protein structure determination field and are the most suitable experimental techniques to elucidate individual protein complexes (Glaeser, 2016), but many structures of protein complexes are still difficult or challenging due to many intrinsic properties of the complex assemblies.

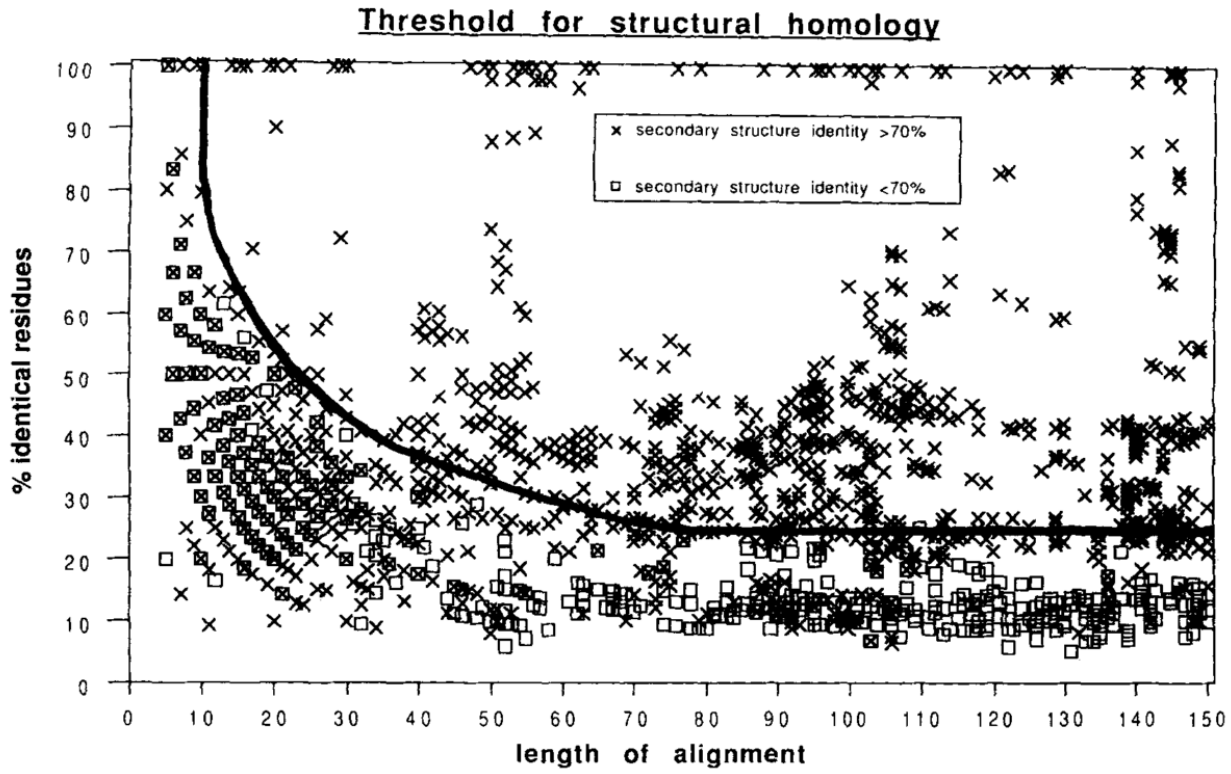


Figure 4. Homology threshold for structurally reliable alignments. Each point represents an alignment between two fragments from proteins of known structure. The threshold (curved line) divides the graph into a region of safe structural homology (upper right) where fragment pairs have good structural similarity (crosses) and a region of homology unknown or unlikely (lower left) where some fragment pairs are structurally similar (crosses) and some are not (squares). Thus, proteins above 25% of identical residues tend to have similar structures and below that is difficult to predict with comparative modeling. Figure extracted from: Sander & Schneider, 1991.

3 DNA

The deoxyribonucleic acid (DNA) is usually composed of two polynucleotide chains twisted around each other, forming a double helix. Nucleotides, which are the building blocks of DNA, are formed by a phosphate group, a sugar and a base, of which there are 4 different types in the DNA: Adenine, Guanine, Cytosine or Thymine (Figure 5). The sequence of these bases defines the message carried by the DNA, all the information necessary to build a cell. Amongst other messages, the sequence of DNA contains genes that codify the sequence of proteins. However important as this is, this is not the only message encoded by DNA. The Encyclopedia of DNA elements (ENCODE) project showed that the DNA contains much more information than was thought before and that the so called “junk” DNA (the one that does not encode proteins) is in fact vital for the organisms (Consortium et al., 2012). Only 1.5% of the human genome consists of exons or protein-coding regions (Wolfsberg et al., 2001), and still, a big part of it participates in biochemical RNA and/or chromatin associated events. Thus, the genomes expression and maintenance are closely regulated by many mechanisms, like non-coding RNA molecules or transcription factors that bind to enhancers (Birney et al., 2007).

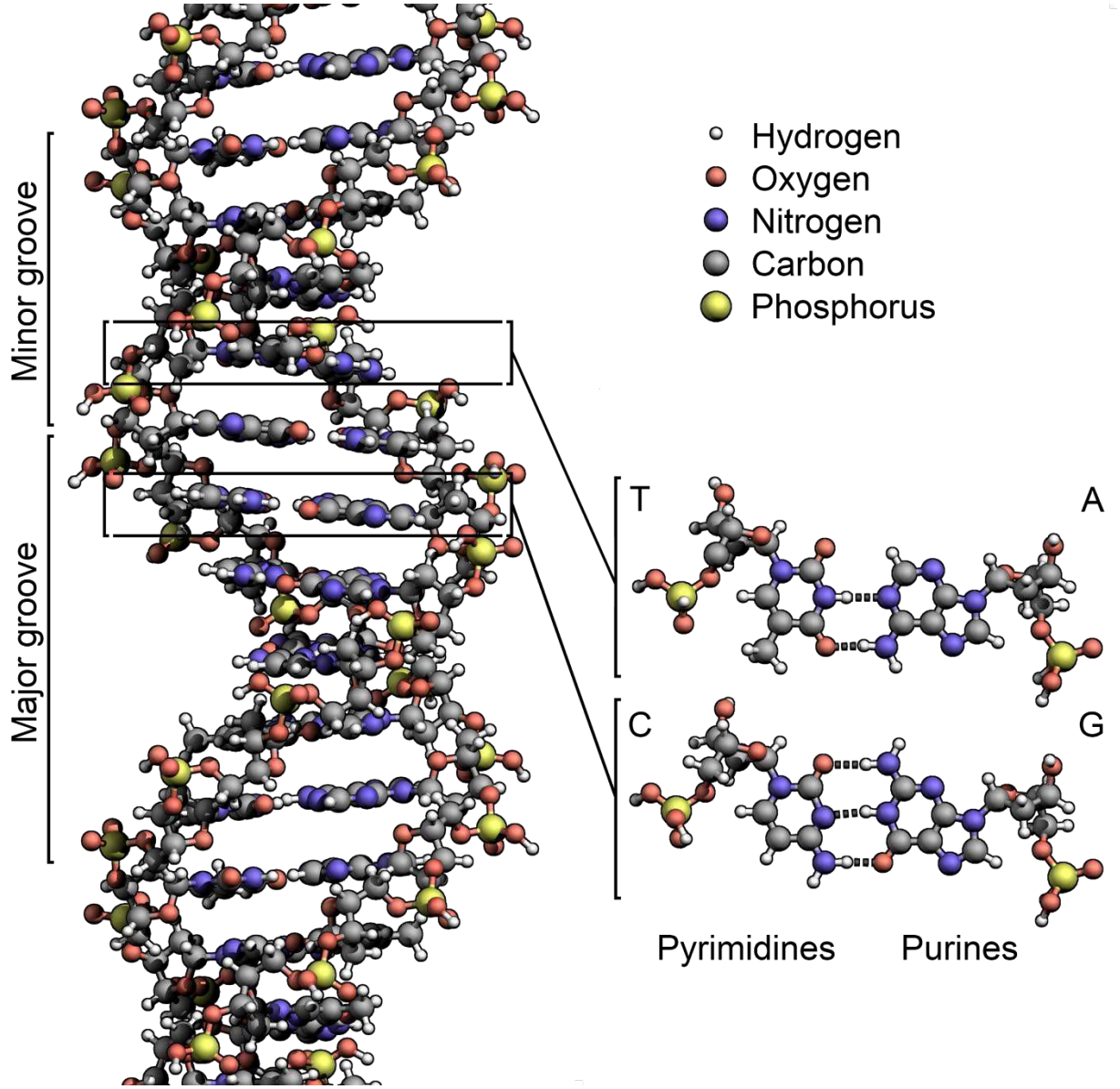


Figure 5. The structure of the DNA double helix. The atoms in the structure are color coded by element and the detailed structures of two base pairs are shown in the bottom right. Source: Wikimedia (commons.wikimedia.org)

3.1 DNA and chromatin structure

The DNA double helix is the first level of organization in the structure of the genome. Two meters of DNA can fit inside a six micrometers cell nucleus. To allow this incredible organizational feature, DNA is tightly compacted inside the nucleus. This compaction is mostly realized by protein complexes called nucleosomes and by other proteins, which, together with the DNA, form a macromolecular complex called chromatin. Nucleosomes are octamers composed of histones and are the main drivers of compaction: each nucleosome is wrapped by 147 base pairs, looping the DNA twice (Figure 6). Their structure and function has been key to study epigenetics, and they have an important role in the accessibility of the DNA.

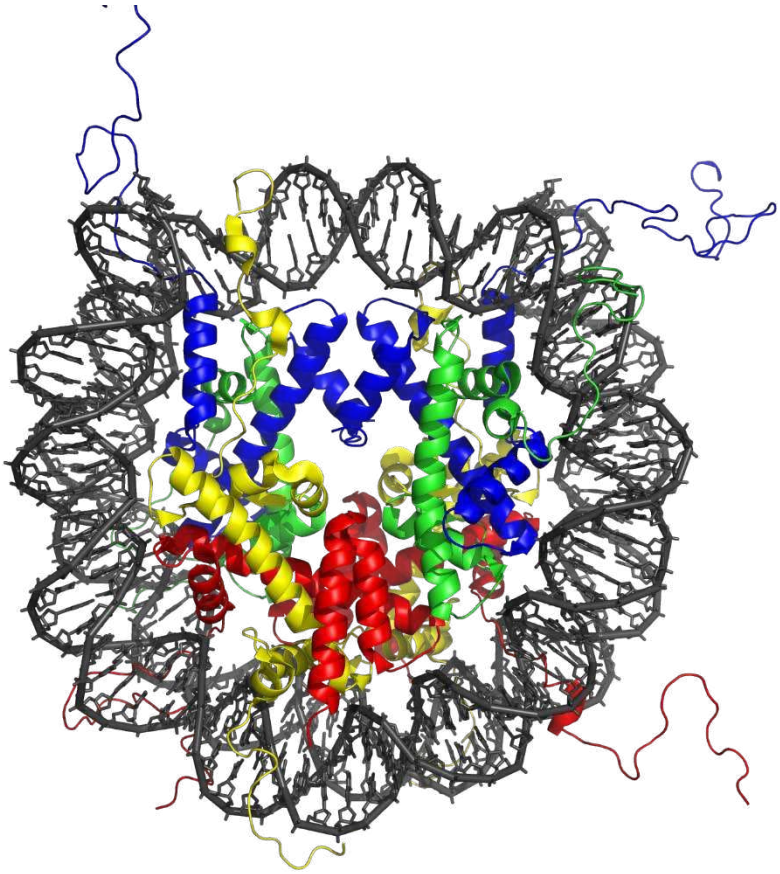


Figure 6. Nucleosome structure. The crystal structure of the nucleosome core particle consisting of 4 core histones (color coded), and DNA. The view is from the top through the super helical axis. Source: Wikimedia (commons.wikimedia.org)

The long standing model is that DNA-nucleosome polymers fold progressively into different levels of discrete higher order chromatin fibers until forming the mitotic chromosomes: the 11 nm DNA and nucleosome assembly, also called “beads on a string” model folds into a 30 nm fiber, which itself folds in a higher order level called chromonema, with 120 nm of width, afterwards in the 300 to 700 nm chromatid and finally in the mitotic chromosome (Figure 7). However, studies on the nucleus with Cryo-EM (Eltsov et al., 2008; McDowall et al., 1986), x-ray scattering (Nishino et al., 2012) and electron spectroscopy imaging (Ahmed et al., 2010; Fussner et al., 2012) have not supported this hierarchical chromatin folding model. Additionally, a recent study has shown that the chromatin is a flexible granular chain with a diameter between 5 nm to 24 nm, packed at different concentration densities in interphase and mitotic chromosomes (Gibcus et al., 2018; Ou et al., 2017), suggesting a different hierarchical compaction of the DNA. In fact, this work showed that the structures proposed in the hierarchical model only appear in *in vitro* studies.

Nevertheless, the packaging of the DNA at different levels has several functions. For instance, the compaction of DNA into chromosomes protects the DNA from damaging. Naked DNA molecules are unstable in cells but chromosomal DNA is extremely stable. Moreover, only the compacted DNA into chromosomes can be transmitted efficiently into daughter cells when the cell divides. Finally, this compacted organization regulates the accessibility of the DNA and, therefore, all the information processing events that involve DNA.

So, although much is known about the structure at the atomic level of the DNA or how the DNA is compacted, many aspects of chromatin structure are still unknown. In the last years, there has been an increasing effort to unravel chromatin structure because it has been shown that it plays a major role in the regulation of gene expression (Andrey et al., 2013; Dixon et al., 2012; Lupiáñez et al., 2015; Nora et al., 2012).

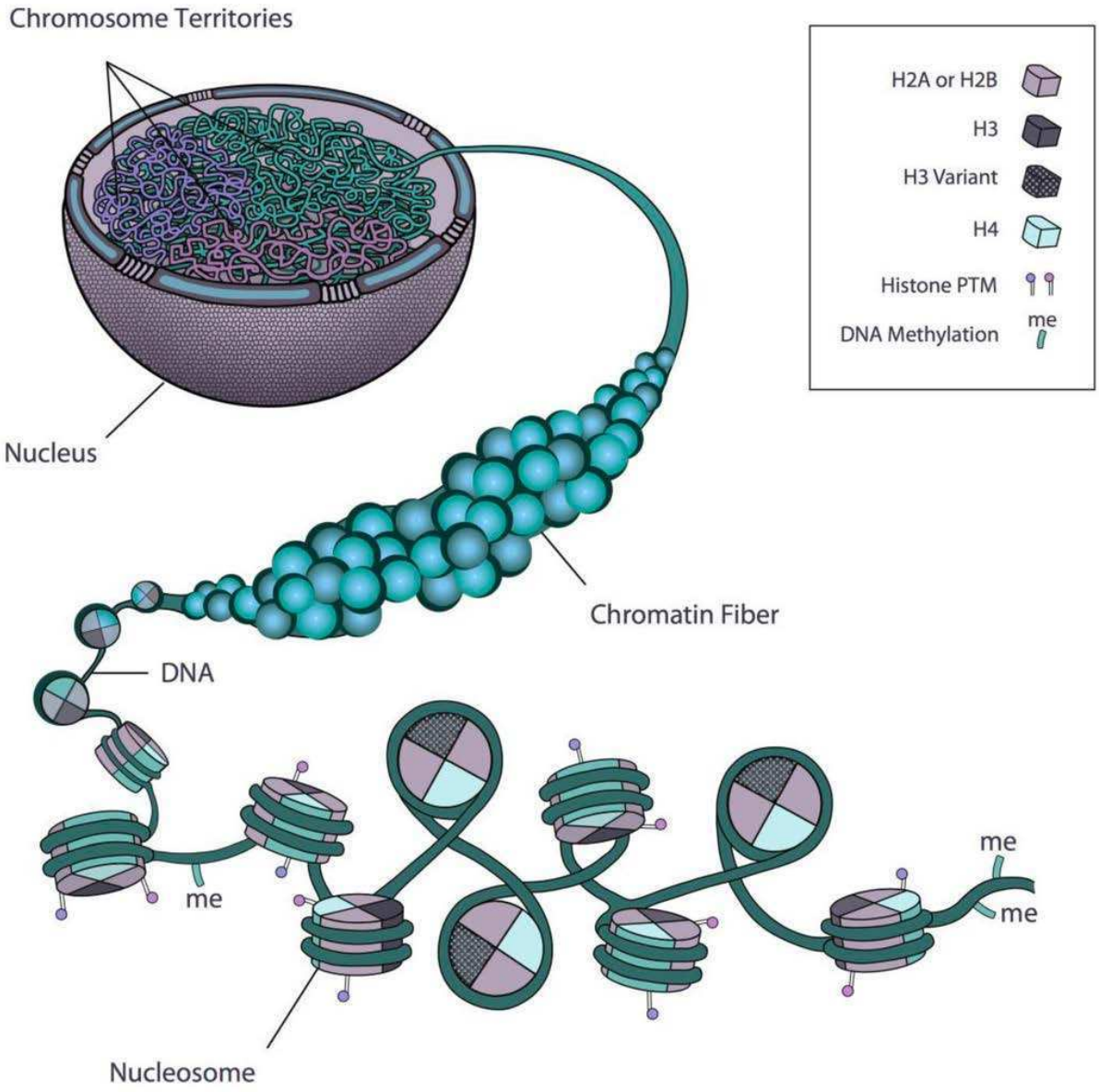


Figure 7. Chromatin compaction levels. DNA wraps nucleosomes forming the chromatin, which is then packed in the nucleus at different concentrations. Figure extracted from S. Rosa & Shaw, 2013.

Chromatin structure is dependent on the phase of its cell cycle: in mitosis and meiosis, the chromatin is packed to facilitate the segregation of the chromosomes in anaphase. In interphase on the contrary, the chromatin is loose to allow transcription and replication. This interphase structure is the least understood and yet, the most thoroughly studied due to the fact that gene regulation occurs at this phase. Recent works involving Hi-C data revealed that the mammalian chromatin is organized and compartmentalized in globular domains called topologically associated domains (TADs) (Dixon et al., 2012; Nora et al., 2012; Sexton et al., 2012) (Figure 8b), although these organized structures have been found also in other animals, plants, yeast and bacteria (Crane et al., 2015; Feng et al., 2014; Hsieh et al., 2015; Imakaev et al., 2013; Mizuguchi et al., 2014). These TADs are the basis for higher level structures referred to as A and B compartments, active and inactive chromatin, which have a role in the partition of the genome inside the nucleus (Dixon et al., 2012; Lieberman-aiden et al., 2009) (Figure 8c). In fact, this partitioning is carried out by compartments of the same type interacting together due to similar epigenetic marks (Rao et al., 2017; Rowley et al., 2017; Schwarzer et al., 2017). This co-localization of same type compartments has been further supported by modeling of 3D nucleus of embryonic stem cells, showing that A compartments are in a ring shape surrounded by B compartments, that locate in the periphery and close to the nucleolus (Stevens et al., 2017). Interestingly, these compartments are not conserved between cell types like TADs and, moreover, TADs can switch between A and B compartments depending on the cell type (Lieberman-aiden et al., 2009; Rao et al., 2014).

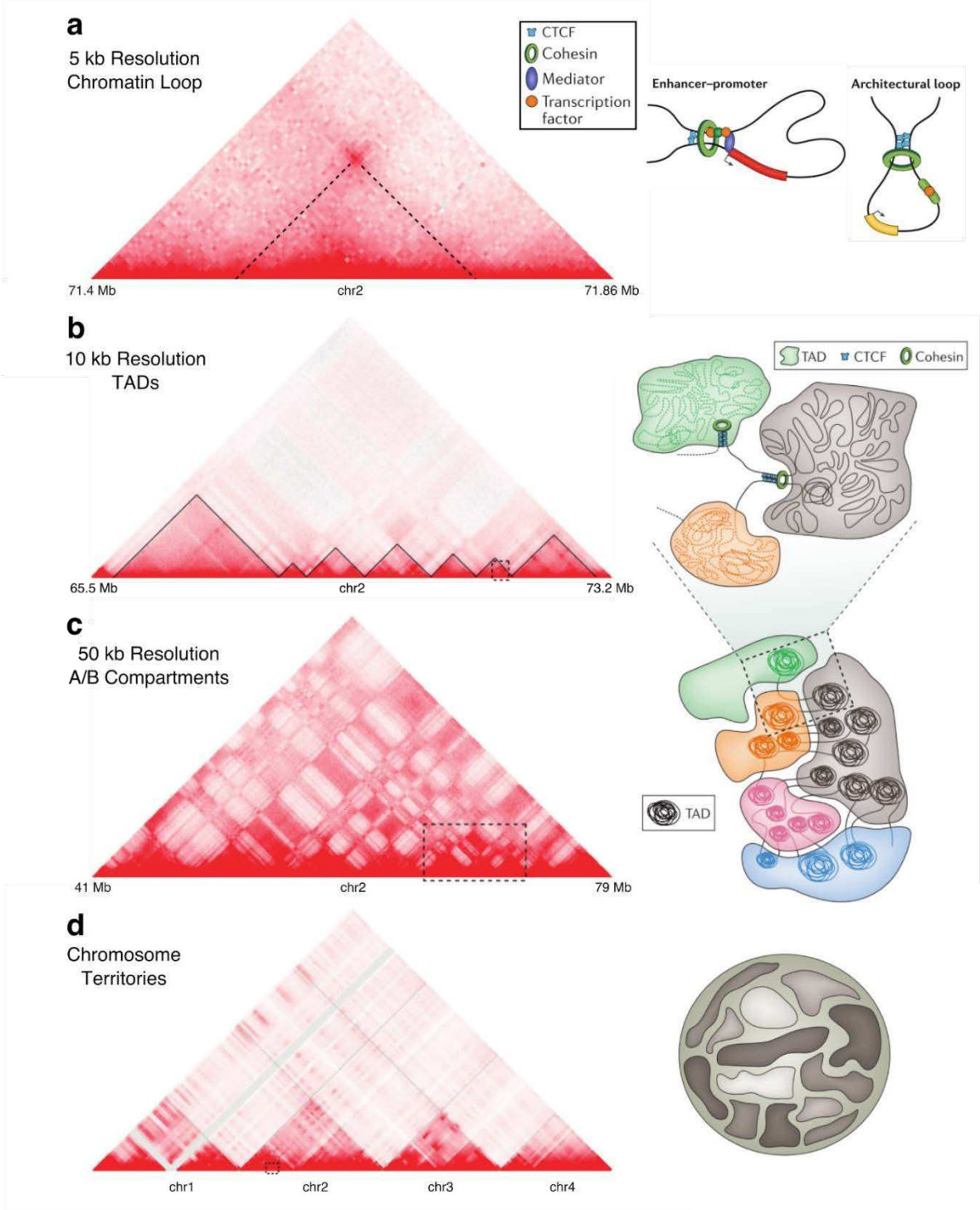

a
5 kb Resolution
Chromatin Loop
CTCF
Cohesin
Mediator
Transcription factor
Enhancer–promoter
Architectural loop
71.4 Mb
chr2
71.86 Mb
b
10 kb Resolution
TADs
TAD
CTCF
Cohesin
65.5 Mb
chr2
73.2 Mb
c
50 kb Resolution
A/B Compartments
TAD
41 Mb
chr2
79 Mb
d
Chromosome
Territories
chr1
chr2
chr3
chr4

Figure 8. Hierarchical organization of chromatin structure. (a) Examples of different types of chromatin loop that can exist within a domain, like enhancer–promoter loop and architectural loop, among others. On the left is an example of an architectural loop as seen in high-resolution Hi-C data. (b) An approximately 8 Mb region containing several TADs. On the right, three different TADs are schematically represented in the 3D space. (c) Different TADs with similar epigenetic signatures are characterized by stronger inter-domain interactions and are organized into compartments. On the right, those compartments are represented as colored regions that contain several TADs. (d) Individual chromosomes (chrs) occupy distinct territories (denoted by irregular shapes) within the nucleus (grey circle). Figure modified from Bonev & Cavalli, 2016.

The existence of TADs has also been supported by other genome wide techniques that do not have most of the limitations that 3C-based approaches have (Beagrie et al., 2017). These domains contain chromatin regions that tend to interact more frequently between themselves than between other regions outside and it has been shown that they often harbor enhancers together with their target genes, having a direct impact in regulation. The average size of a TAD is 185 kb on average, ranging between 40 kb and 3 Mb. Interestingly, they are thought to be conserved between species (54% to 76% between mouse and human) and tissues (50% to 72%) (Dixon et al., 2012; Rao et al., 2014), but the numbers differ depending on the resolution of the studies and also in the definition of TADs. This is due to the fact that these domains are hierarchical domains which can be composed of smaller domains called sub-TADs (Phillips-Cremins et al., 2013; Wijchers et al., 2016). On a larger scale, when TADs interact with each other, they can organize in meta-TADs, when significant inter-TAD interactions occur (Fraser et al., 2015). Both sub and mega-TADs, display a more tissue specific behaviour than TADs (Zhan et al., 2017). But, in the end, the definition of TADs and its boundaries depends on the data resolution and computational algorithms, and, therefore, is somewhat arbitrary.

TAD boundaries are enriched in insulator proteins like CCCTC-binding factor (CTCF), a protein that contains an 11-zinc-finger DNA-binding domain, which was detected at ~76% of all boundaries (Dixon et al., 2012), but only 15% of all mammalian CTCF binding sites are located within a well-defined, well-established boundary. Most of them are located inside TADs and they are thought to be involved in the formation of sub-TADs (Handoko et al., 2011), which suggests that CTCF binding alone may be insufficient for TAD generation. However, CTCF sites at TAD boundaries follow a convergent orientation, suggesting that the directionality of the protein is important for TAD formation (de Wit et al., 2015). Alteration of these sites, such as, an inversion

in a boundary, disrupts its interaction with an upstream convergent CTCF binding site, altering the expression of the neighboring gene. In addition, another study inverting and deleting CTCF motifs showed that domains were destabilized (Sanborn et al., 2015). Both works support the importance of the motif orientation for TAD formation. But transcription could also have an important role. In flies, transcription seems a better predictor of TAD boundaries than CTCF, suggesting that different organisms may use different strategies to establish chromatin domains (Ulianov et al., 2016). However, new studies in flies and mice have shown that TADs arise independently of transcription, since they are established prior to genome transcriptional activation (Flyamer et al., 2017; Hug et al., 2017). In opposing contrast, a recent work generated high resolution Hi-C maps on flies, showing that TAD boundaries are in fact small active domains and not transcriptionally active regions (Rowley et al., 2017). In fact, RNAPII is present in the whole domain, reaching the conclusion that TAD boundaries are defined by segregation of A (active) and B (inactive) compartments.

But it is still not clear how TADs are established and which is the functional difference between them and the compartments. Many believe that CTCF binding along with gene expression are key elements in TAD formation and compartments are formed through attraction and repulsion of individual TADs with similar epigenetic marks. This theory is supported by correlations between chromatin marks in regions within TADs compared against other TADs (Sexton et al., 2012) and TAD boundaries that overlap with compartment transitions (Rao et al., 2014). Further support has been provided by super-resolution microscopy showing differences in spatial interactions between neighboring TADs with different epigenetic marks (Wang et al., 2016), like Polycomb repressed TADs that are highly condensed and repel neighboring domains (Boettiger et al., 2016).

In the last years, many studies have shown that CTCF (Splinter et al., 2006) and cohesin (Hadjur et al., 2009) are essential components of chromatin looping together with the mediator complex (Kagey et al., 2010) which bridges promoters and enhancers (Figure 8a). All these proteins have been proposed to work as architectural proteins in the regulation of genes. In this regard, the loop extrusion model (Alipour & Marko, 2012; Fudenberg et al., 2016; Nasmyth, 2001; Nichols & Corces, 2015; Sanborn et al., 2015) has been proposed as a mechanism to generate TADs and compaction of chromatin (Figure 9). The model explains how cohesin would act as a chromatin extruder while no CTCFs proteins are found. In fact, this model is supported by recent studies in which they show that CTCF or cohesin depletion leads to a loss of TADs, while A/B compartments remain intact, and even reinforced (Nora et al., 2017; Rao et al., 2017; Schwarzer et al., 2017). This results have been further supported by single-cell Hi-C in maternal nuclei, where

despite having loops and TADs, a lack of compartments was shown, implying that these features arise independently (Flyamer et al., 2017).

But not all TADs seem to follow the same patterns. A recent work studied spans of conserved non coding elements (CNE), also known as genomic regulatory blocks (GRB) which can predict the span of some of the strongest and most gene sparse TADs in humans and flies, normally containing developmental genes (Harmston et al., 2017). In addition, other species like Schizosaccharomyces pombe, Caenorhabditis elegans, Saccharomyces cerevisiae and Arabidopsis thaliana have shown TAD like domains but have no CTCF homologs, which is a key player in domain formation (Rowley & Corces, 2016). These suggests that there are many ways to TAD formation.

All these data suggests that the overall process of the chromatin architecture organization is hierarchical. First, dynamic nucleosome contacts form groups that interact with further regions, generating loops. Some of these loops can also be established or stabilized by protein-protein contacts involving architectural proteins (like CTCF or cohesin) and/or regulatory proteins (like transcription factors, Polycomb and heterochromatin proteins) generating TADs. Afterwards, TADs with the same epigenomic marks form compartments, and compartments clutch together to form chromosome territories (Figure 8d). It is clear that these chromatin domains called TADs have an important role in genome function and thus, in cell functioning. So, in this regard, the study of TADs is essential and can even help understand other topics like evolution or diseases, which we will talk about in the next lines.

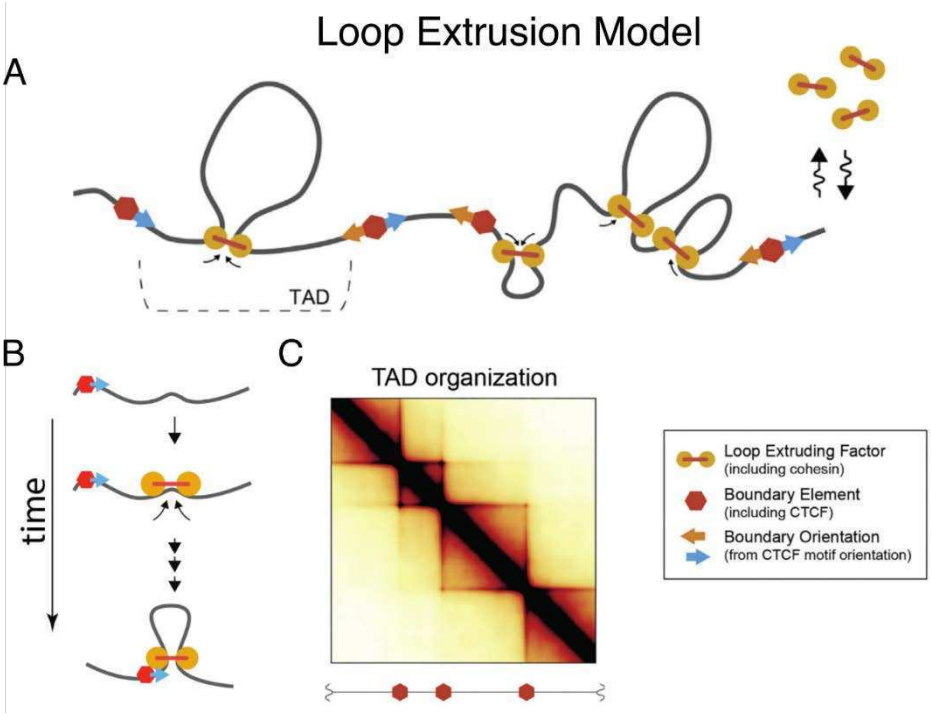


Figure 9. Loop extrusion model. (A) Schematic of loop extrusion model dynamics. (B) Schematic showing loop generation over time. (C) Simulated contact map. Figure modified from Fudenberg et al., 2016; Imakaev et al., 2015.

3.1.1 Understanding chromatin dynamics through TADs

Even though primary TADs and the overall architecture of the chromatin is conserved between species and cell types, the dynamics of the chromatin contribute to the establishment of diverse programs of specific gene expression. Big differences in the overall chromatin architecture have been found between the active and inactive X chromosome in human (Rao et al., 2014) and mouse cells (Deng et al., 2015). While the active X chromosome was composed of regular TADs, large domains called “superdomains” were observed in the inactive chromosome, and, interestingly, the boundaries of the superdomains overlapped with the TAD boundaries in the active chromosomes, suggesting a dynamic reorganization of the architecture. During cell differentiation, a study comparing human ES cells with four different ES-derived lineages that represented early development stages showed that genome organization was mostly unchanged while intra TAD interactions in some domains were strongly altered (Dixon et al., 2015). 3D genome modeling studies done using single-cell data showed also that TAD structures and loops

vary from cell to cell (Stevens et al., 2017). These changes correlate with TADs switching between compartments and with changes in the transcription status of the genes within the TADs, suggesting that transcription is coordinated within TADs. These changes and dynamics within TADs are supported by computational simulations that show how TADs can easily fluctuate between several conformational states in timescales that are much shorter than one cell cycle (Tiana et al., 2016).

It was also thought that long-range interaction could only occur in cells where the target gene is active (Palstra et al., 2003). Nevertheless, studies in mice support that long interactions can happen even when the gene is inactive (Amano et al., 2009). Sonic hedgehog protein (Shh) is regulated by a distal enhancer called ZRS in the posterior region of the limb buds, creating a chromatin loop that contacts both elements. But this loop exists in other cells like in the anterior region of the limb buds, where Shh is not transcribed (Symmons et al., 2016). Similar results were obtained when regulatory elements of the HoxD genes were studied, suggesting that stable long range enhancer-promoter interactions exist (Lonfat & Duboule, 2015). Moreover, interactions and chromatin conformation is also dynamic during cell cycle: G1 cells have mainly short range intra chromosomal interactions while cells in replication are enriched in long range intrachromosomal interactions. G2 and mitosis cell increase in short range contacts since they start to compact (Lazar-Stefanita et al., 2017; Nagano et al., 2017). It has been shown also that SMC complexes (Cohesin and condensin) play a role in the restructuring of the cell, at least in S. Pombe (Kakui et al., 2017) and S. Cerevisiae (Lazar-Stefanita et al., 2017; Schalbetter et al., 2017).

3.1.2 The role of TADs in development and diseases

Understanding TADs is also essential to study development and diseases. Structural variations can affect gene expression and cause pathogenic phenotypes. This has been shown in a study on the human genome, where large scale inversions, deletions and duplications were carried out within the WNT6-IHH-EPHA4-PAX3 locus, affecting the limbs (Lupiáñez et al., 2015). Several TAD boundaries were disrupted leading to ectopic interactions between limb enhancers that normally would be regulating inside the EPHA4 TAD and gene promoters located outside this domain. It was shown that these TAD disruptions were dependent of CTCF-associated boundary elements. This conclusion is supported by another study done in humans (Hnisz et al., 2016) were deletion of CTCF-mediated boundaries were enough to activate proto-oncogenes. In another study, duplications of chromatin regions containing TAD boundaries near the Sox9 locus were able to create neo-TADs, determining their molecular pathology (Franke et al., 2016) due to an over compartmentalization of the region. When the Kcnj2 gene is duplicated in the neo-TAD, a

gain in ectopic contacts of Kcnj2 with the regulatory region of Sox9 generates misexpression of Kcnj2 and a limb malformation phenotype. This findings support the proposal that TADs are genomic regulatory units with a high degree of stability, which restrict the contacts that enhancers establish with their target genes. Thus, TADs enclose the regulatory landscape of the genes that it contains, and it is clear that the disruption of these TADs can produce a deregulation of these genes, which can lead to diseases.

In addition, understanding the function and mechanisms of TADs is also essential to understand development, and a clear example is the HoxD cluster. These genes are located between two TADs and are sequentially activated during development, in an ordered fashion along the chromatin. They are responsible for the anterior-posterior axis development in animals and limb development in vertebrates. In fact, in the first stages of their development, most genes are regulated by enhancers located in the anterior part, but, when fingers are being developed, some of the genes start to be regulated more preferentially by posterior part enhancers, which suggest a topological change of the chromatin during development and a phenomenon that could be difficult to understand without TAD knowledge (Andrey et al., 2013).

TADs are also responsible for bringing together remote enhancers that, in absence of them, would never be able to regulate their target gene. This is the case of the Shh gene, where it was shown that when the ZRS enhancer was located outside the Shh TAD, not being able to contact its target gene, disrupting limb development (Symmons et al., 2016).

3.1.3. TAD conservation in different species

Chromatin architecture is closely bound to gene expression, therefore, it could also be evolutionarily conserved. In fact, domains similar to mammalian TADs have been found in non-mammalian genomes like fruit fly (Sexton et al., 2012) or zebrafish (Gómez-Marín et al., 2015), but recent chromatin-interaction maps have also uncovered domain-like structures in other species. In yeast they have found chromosome interacting domains or CIDs that are much smaller than the megabase scale TADs in mammals and they have an average of 1 to 5 genes within (Hsieh et al., 2015). Self-interacting domains, or SIDs, in fission yeast, with an average size of 50-100 kb were shown to be dependent of cohesin (Mizuguchi et al., 2014) while SIDs with an average size of 1 MB were only found in the X chromosome of Caenorhabditis elegans (Crane et al., 2015). Regarding plants, two studies in Arabidopsis thaliana were carried on, one finding very few and small interactive regions enriched in repressive marks (Grob et al., 2014) and another study observing large domain-like structures called “structural domains” (Feng et al., 2014),

making the existence of these domains in plants debatable. Even similar domains were found in bacteria, which were enriched with active genes at boundaries (Le et al., 2013).

But, not only domains are conserved between some species. Syntenic regions in mouse and human seem to have conserved 3D topology, indicating that not only sequence is conserved (Dixon et al., 2012). This theory was later validated by the analysis of other four different mammalian species (Vietri-Rudan et al., 2015). They showed that this conservation was dependent to CTCF binding sites co-localizing with cohesin, which are big players determining TAD boundaries. It has been also shown that long-range contacts in the Hox loci, which are mediated by polycomb, are conserved between fly species that diverged 40 million years ago (Bantignies et al., 2011).

3.2 Experimental techniques to explore chromatin structure

At the atomic to nanometer scale level, the same techniques that are and were used in the determination of protein structures have been used to elucidate the architecture of the chromatin. For instance, light and electron microscopy can provide direct observation of the structural organization of the chromatin (Flors & Earnshaw, 2011; Rapkin et al., 2012). At a greater scale, FISH studies have revealed that each chromosome occupies a distinct nuclear territory and that the 3D positioning of chromosomes correlate with size and gene density (Bolzer et al., 2005; Branco & Pombo, 2006). Moreover, large and gene poor regions of the chromosomes are located in the periphery, while small and gene rich regions are located in the interior (Bickmore & Van Steensel, 2013).

But a gap has been created at the 1-to-100 nm folding level, since the technology available some years ago was not suited to study this resolution (Figure 10). For instance, even though there has been a case where light microscopy has achieved resolutions of 10-20 nm (Huang, Babcock, & Zhuang, 2010), it usually can't separate objects closer than 200 nm due to diffraction limit, which it makes not suitable to study the chromatin at the 1-to-100 nm resolution.

This folding level of the chromatin is in a scale at which protein and DNA interplay could be studied: how proteins bind the DNA, how genes are repressed or silenced or how DNA folds and unfolds. If we had a clear understanding at this resolution, the gene regulation machinery could be studied with its 3D structure, which would allow for a huge leap forward in many fields like development and disease. In this regard, a recent study elucidated the compaction level of the chromatin fiber using a fluorescent dye in the DNA, enhancing its contrast in electron microscopy, showing that the chromatin folds in itself forming strings of 5 to 24 nm width (Ou et al., 2017), but there is still much to study if we want to understand the chromatin structure at this

level. That is why scientists have started exploring the structure and the architecture of the chromatin using other techniques.

In 2002, the chromatin conformation capture (3C) technique was used to study the conformation of the whole chromosome III in yeast (Dekker et al., 2002) and to show the regulatory interactions of hypersensitive sites in the active β-globin locus, elucidating the regulatory landscape of the region (Tolhuis et al., 2002). This technique shows the frequency of contact between two chromatin regions. Since then, many variants of this technique have been developed, each of them with each own advantages and disadvantages. But the first steps of all these techniques, called 3C-based methods, are the same.

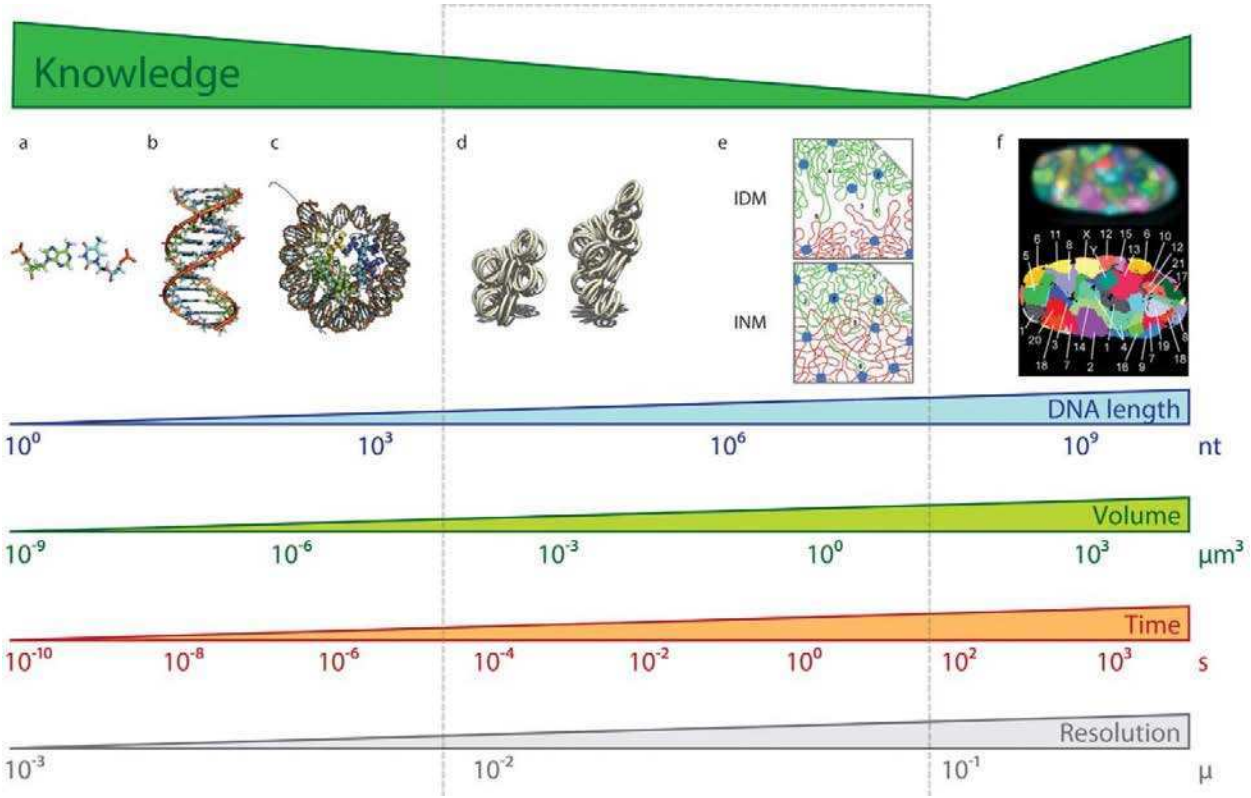


Figure 10. DNA and chromatin have been studied at different resolution scales: (a) nucleotides, (b) DNA structure, (c) nucleosome complex, (d) higher older folding, (e) interchromatin domains and interchromosomal interactions and (f) chromosome territories. More studies are needed in order to provide more insight about the chromatin architecture at the range of resolutions indicated by the dashed rectangle is the. Figure extracted from Marti-Renom & Mirny, 2011.

The first step is to capture a snapshot of the chromatin. For that, the chromatin of millions of cells is fixed using a fixative agent like formaldehyde. Next, the fixed chromatin is cut with a restriction enzyme. These enzymes are cutters that recognize 6bp such us HindIII, BglII, SacI, BamHI, and EcoRI or alternatively higher frequency cutters can be used like AciI or DpnII, which recognize 4bp. Afterwards, the ends of the cross-linked DNA fragments are religated. In this way, DNA fragments that are in close proximity in the nucleus can be ligated to each other, reflecting the three-dimensional organization of the genome at time of fixation and can therefore be used to infer the chromatin structure. This steps generate the 3C library which is the same for most of the 3C-derived techniques. On the contrary, the measurement of the number of ligation events is different for each of the 3C techniques. The differences between the most popular 3C-based methods are as follow (Figure 12):

Chromatin conformation capture (3C): The 3C experiment quantifies interactions between single pair of genomic fragments (one-vs-one strategy). In this case, ligated fragments are quantified using PCR with known primers.

Circular Chromatin Conformation Capture (4C): 4C technique was the next major advance in the field, which allows identifying all interactions from a specific locus of interest (also called viewpoint) of the genome (Simonis et al., 2006; Zhao et al., 2006) (one-vs-all strategy). It involves a second ligation step to generate self-circularized DNA fragments. Afterwards, inverse PCR is used from the viewpoint to amplify any interaction fragments. Originally the contacts were analyzed using microarrays, but nowadays 4C-seq is used, the version which uses next generation sequencing (NGS) to analyze contacting sequences. 4C has been used to detect interactions between promoters and enhancers, and to show how these change during differentiation and development (Andrey et al., 2013). Very Long range contacts (>10MB) of active genes also have been demonstrated (de Wit et al., 2013). It has also been used in disease mechanisms, demonstrating that chromosomal translocations can bring distal enhancers close to an oncogene leading to malignancy (Gröschel et al., 2014).

Chromosome conformation capture carbon copy (5C): 5C is a technique that detects all interactions between multiple selected sequences (many-vs-many) (Dostie et al., 2006). It relies on multiplexed ligation-mediated amplification (LMA) of a conventional 3C library. Designed 5C primers anneal on either side of all the ligated junctions in the region of interest in a 3C library. These primer pairs are ligated, and can then be amplified using one of two universal sequences

incorporated within them in a single PCR reaction, which can be analyzed by microarray or sequencing. 5C has been used to determine interaction profiles at the pilot regions of the ENCODE project. Massively multiplexed 5C could be used to generate all vs all interaction maps but requires major financial resources.

Hi-C: In contrast to the techniques described above, Hi-C offers the interaction map of all the genome, i.e., all vs all (Lieberman-aiden et al., 2009). NGS methods let it development in 2009. After chromatin is cross-linked and digested, restriction ends are filled in with biotin-labeled nucleotides, slightly changing the 3C library generation. After a blunt-end ligation, DNA is purified and sheared, and a biotin pull-down is performed, so only ligated fragments are selected, ensuring enrichment on ligation junctions that are subsequently sequenced from both ends by paired-end sequencing. The reads are mapped to the genome, outputting a matrix of pairwise interaction frequencies between fragments across the genome. It has also been extensively used to describe principles of chromatin organization (Dixon et al., 2012; Rao et al., 2014), and determine the structure of the chromosome during mitosis (Naumova et al., 2013). Moreover, Hi-C has been used to link trans-interactions with sites associated with chromosomal translocations (Zhang et al., 2012). Hi-C has also been performed in single cells, picking and sequencing single intact nuclei during Hi-C library preparation and have showed the big variability between single cells (Nagano et al., 2013; Ramani et al., 2017).

Since the method's development, many Hi-C maps have been generated: Bacteria like Caulobacter crescentus (Umbarger et al., 2011), Escherichia coli (Cagliero et al., 2013), and in Bacillus subtilis with 30, 10 and 4 kb resolutions (Marbouty et al., 2014, 2015; Wang et al., 2015), fission (Tanizawa et al., 2010) and baker's yeast (Duan et al., 2010), Arabidopsis thaliana (Wang et al., 2015), Drosophila melanogaster (Sexton et al., 2012) and mouse (Dixon et al., 2012), human (Rao et al., 2014), and others mammals (Vietri-Rudan et al., 2015).

ChIA-PET: it combines 3C technology with chromatin immunoprecipitation (ChIP). A specific antibody is used to pull down ligation junctions bound by a protein of interest (Fullwood et al., 2009). It is a "many versus many" approach, as it interrogates contacts between sites bound by a given protein factor. The first experiment was used to identify the interaction network of estrogen receptor α (ERα) (Fullwood et al., 2009). Since then, it has been used to generate interaction data for a number of key chromatin-bound factors like CTCF or cohesin.

Capture-C: Capture-C generates genome-wide interaction maps from a single viewpoint, similar to 4C, but in a high-throughput manner, allowing the interrogation of many viewpoints in a single assay (Hughes et al., 2014). It combines 3C library preparation and NGS with oligonucleotide capture technology. In this technique, the libraries are sonicated, allowing an enrichment of the fragments of interest using biotinylated capture probes, designed for each viewpoint. Finally, these fragments are amplified and sequenced. An improvement of Capture-C called, next-generation (NG) Capture-C (Davies et al., 2015) uses a new oligonucleotide capture process that markedly increases assay sensitivity. Hughes et al. examined about 450 promoters and showed that cis DNA interactions with promoter are most likely within a 600 kb region surrounding it.

All this techniques give us the same type of data: frequency of contact between two DNA regions; and it can be used to generate 3D models of the chromatin or to corroborate models generated by other means. In the next lines, we will talk briefly about these methods.

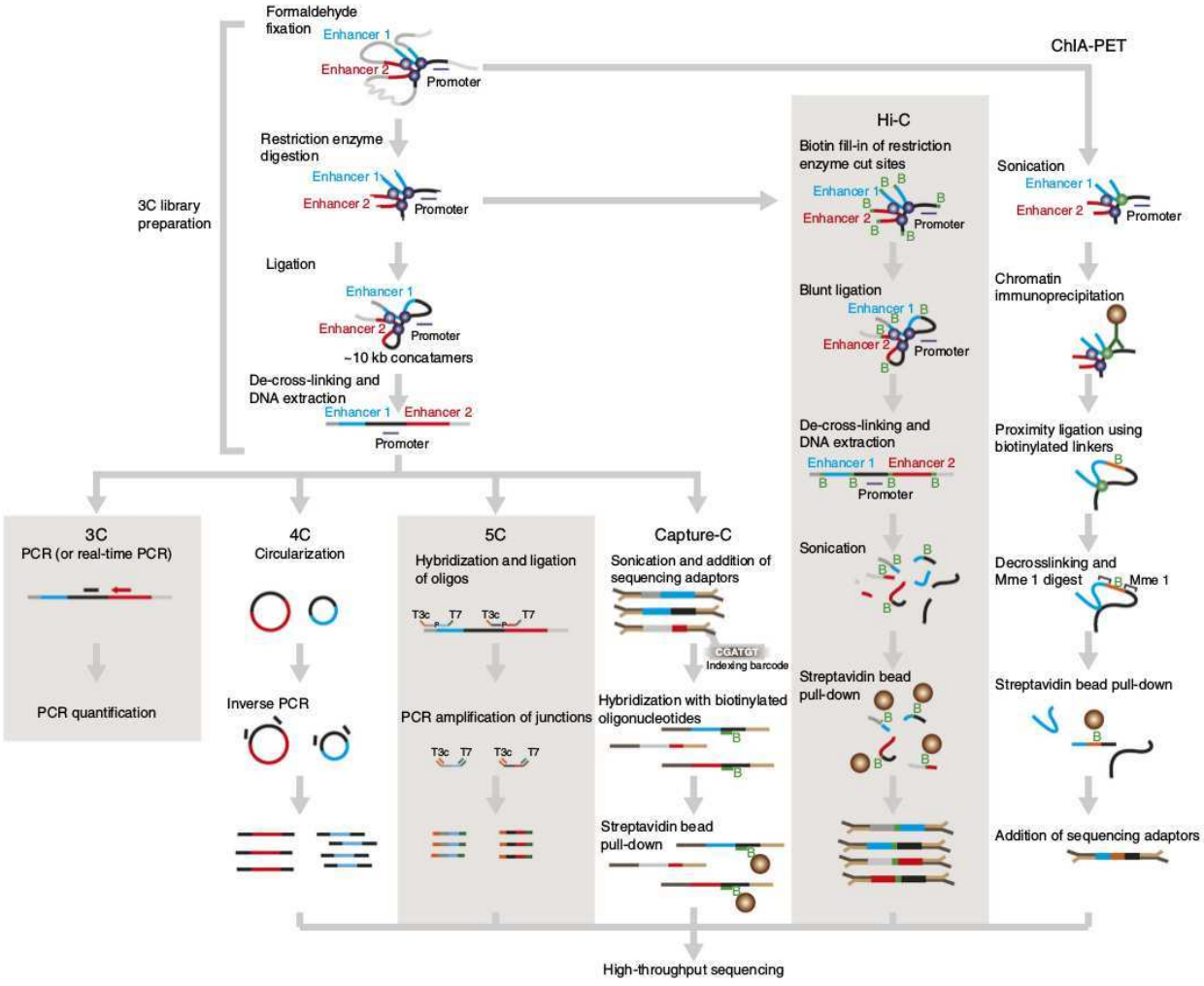


Figure 12. Comparison of different 3C-based methodologies. The 3C library preparation step is shared by all the methodologies. Figure extracted from Davies et al., 2017.

3.3. Chromatin modeling methods

There are many methods to model the 3D structure of the chromatin. We divide them in two main approaches (Imakaev et al., 2015; Serra et al., 2015) (Figure13). The first is known as restraint-based (RB) modeling or data-driven modeling and interprets the 3C-based data as distance restraints between the chromatin fragments which are then used to generate 3D models of the chromatin fiber by satisfaction of these restraints. The second strategy is called thermodynamic-based (TB) modeling or de novo modeling, which is based on the polymer physics principles of the chromatin fiber to generate 3D models that afterwards are corroborated using 3C-based data.

3.3.1. Restraint-based or data driven approaches

Among these methods, there are two main categories, depending on the implementation of restraints:

Consensus structure models or analytical methodologies. These methods use analytical approaches to transform the frequency of contacts into spatial distance between loci (Duan et al., 2010; Hu et al., 2013; Lesne et al., 2014; Segal et al., 2014; Varoquaux et al., 2014; Zhang et al., 2013). Based on this, a set of constraints are imposed and a consensus structure is generated. These approaches are suitable for single-cell based studies because they assume that the 3C-based data only represents a single structure (Nagano et al., 2013), but are not able to describe Hi-C maps because these are derived from a highly variable ensemble of structures. They assume that there are small fluctuations in the average distances between chromatin fragments, but imaging experiments say otherwise, since they show that the variability in the spatial distance between two loci is often similar to their average separation (Giorgetti et al., 2014).

Data-driven ensembles or optimization based methods. These methods try to simulate an ensemble of structures that can explain the experimental 3C-based data (Baù et al., 2011; Giorgetti et al., 2014; Kalhor et al., 2011; Zhang & Wolynes, 2015) The chromatin is described by series of monomers like beads or points that interact due to a set of imposed forces. These forces can impose connectivity or exclusion between monomers or the stiffness of the chromatin. Unlike analytical methodologies, data-driven ensemble approaches use Monte Carlo or Molecular Dynamics to sample the space of possible solutions depending the restraints and generate a set of 3D models that is variable enough to be representative of the 3C-based contact map.

These methods can be further divided into two approaches. The first type are the resampling approaches where each simulation is independent to the others and the variability of the generated models can represent the conformations that the chromatin can adopt. In the end an ensemble of solutions is obtained where each model satisfies the restraints similarly (Benedetti et al., 1988; Halverson et al., 2011; Rosa & Everaers, 2008). The second type is the population-based approach and tries to be loyal to the 3C-based experiment. To accomplish that, each model is generated independently and tries to account for a fraction of the 3C-based data, so that the ensemble of solutions could explain all the variability in the cells. These methods have helped describe many biological features of the chromatin fiber like the presence of compartments and their tendency to aggregate by type (Hu et al., 2013), or the overall genomic organization of the yeast (Tjong et al. 2012), the human (Kalhor et al., 2011) and the Plasmodium genome (Ay et al., 2014).

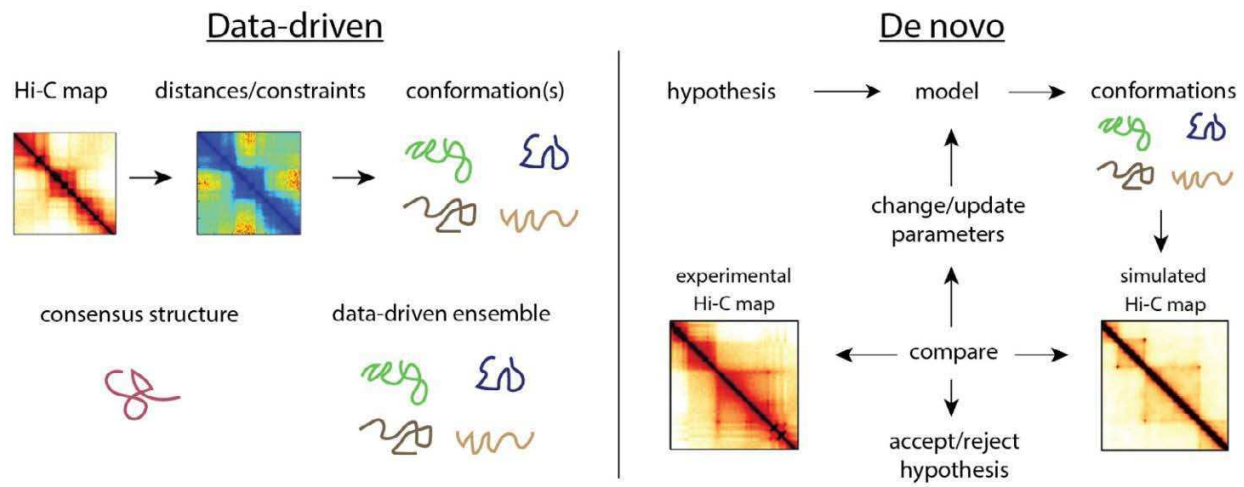


Figure 13. Two main approaches to generate 3D chromatin models. Left: Data-driven approaches (also called restraint based approaches) convert Hi-C maps into distance restraints. Then, a consensus structure or a set of conformations can be generated. Right: De novo approaches generate models based on a hypothesis with basic polymer physics. Contact maps derived from these models are then compared to experimental Hi-C maps which can accept, reject or adjust the hypothesis. Figure extracted from Imakaev et al., 2015.

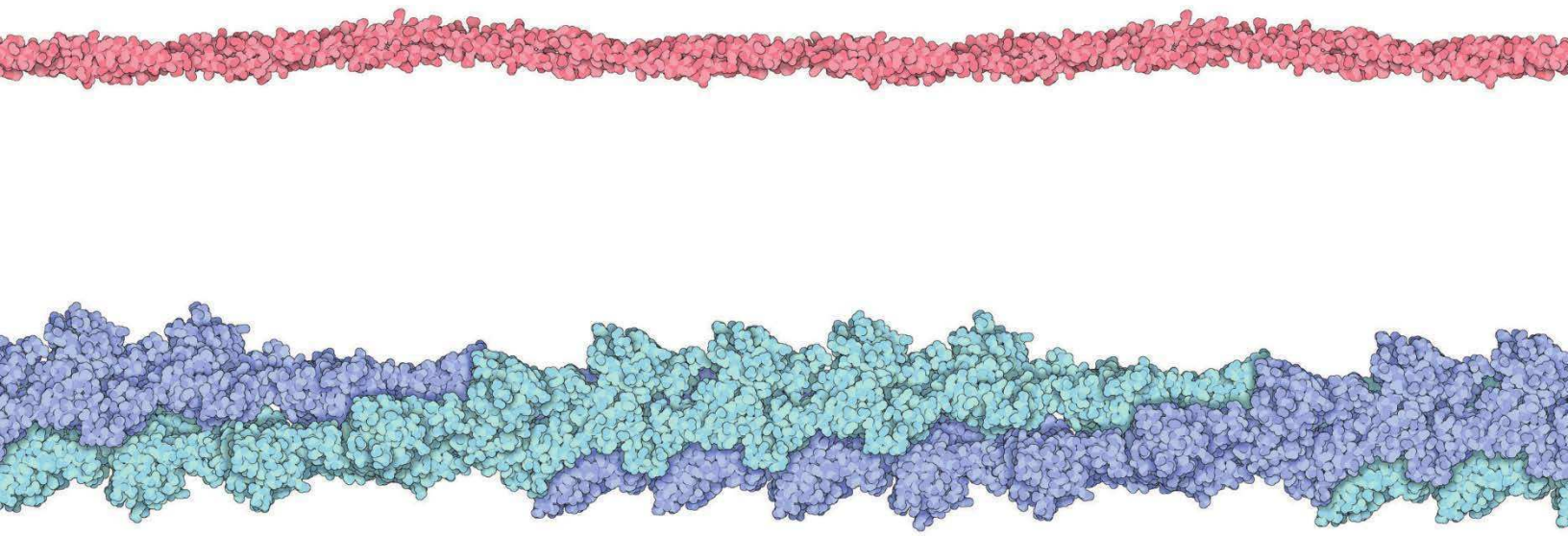
3.3.2 De novo ensembles or thermodynamics-based (TB) modeling

An alternative approach to the ones presented above is to test whether known or hypothesized physical or biological principles are able to generate ensembles that can explain Hi-C maps or key features of them. These models are not generated using the Hi-C maps, on the contrary, the Hi-C maps are used to test the models.

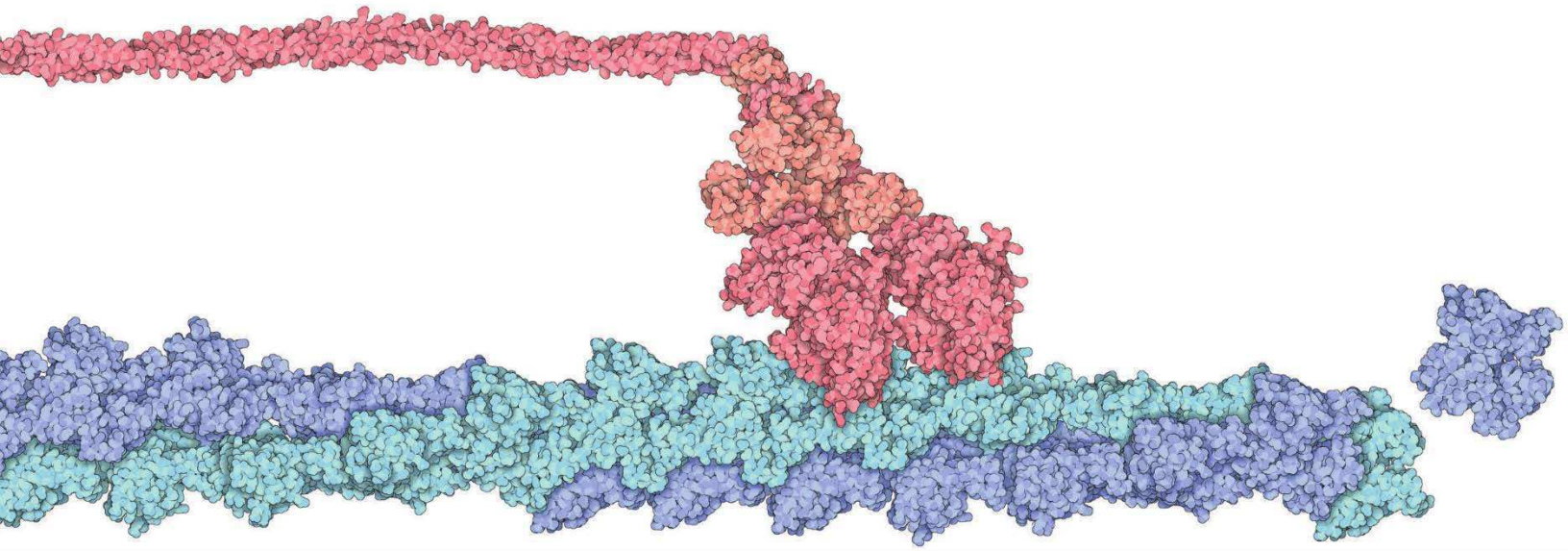
These approaches have been used to interpret the decay of interaction frequencies with the genomic distance (Mirny, 2011), the formation of domains of active and inactive chromatin (Barbieri et al., 2013), epigenetic features like chromatin colors (Jost et al., 2014), chromosome territories (Emanuel et al., 2009; Hahnfeldt et al., 1993; Münkel & Langowski, 1998) and co-expression data (Di Stefano et al., 2013).They can also be divided in two categories, depending on the principles used to generate the models: Mechanistic ensembles, which only use biologically-plausible interactions and structural ensembles.

Among **structural ensembles** we can find polymer ensembles like those generated by random walks (van den Engh et al., 1992), equilibrium globules (Grosberg et al., 1995), melt of polymer rings (Halverson et al., 2011, 2014; Vettorel et al., 2009) and fractal globule (Grosberg et al., 1988; Lieberman-aiden et al., 2009). These ensembles can provide information on chromosomal organization without providing information about mechanisms of folding.

Mechanistic ensembles, on the other hand, test computationally the hypothesis of whether a mechanism or a set of mechanistic constraints could explain a Hi-C map. This approach have been used to test if decondensation from mitosis is able to explain interphase chromosomal organization (Rosa et al., 2010; Rosa & Everaers, 2008) or to test if human mitotic chromosomes could arise from the process of loop extrusion (Naumova et al., 2013). Another study suggested that TADs in mammalian interphase chromosomes could arise from the activity of cis-acting loop extruding factors (Fudenberg et al., 2016).



Objectives



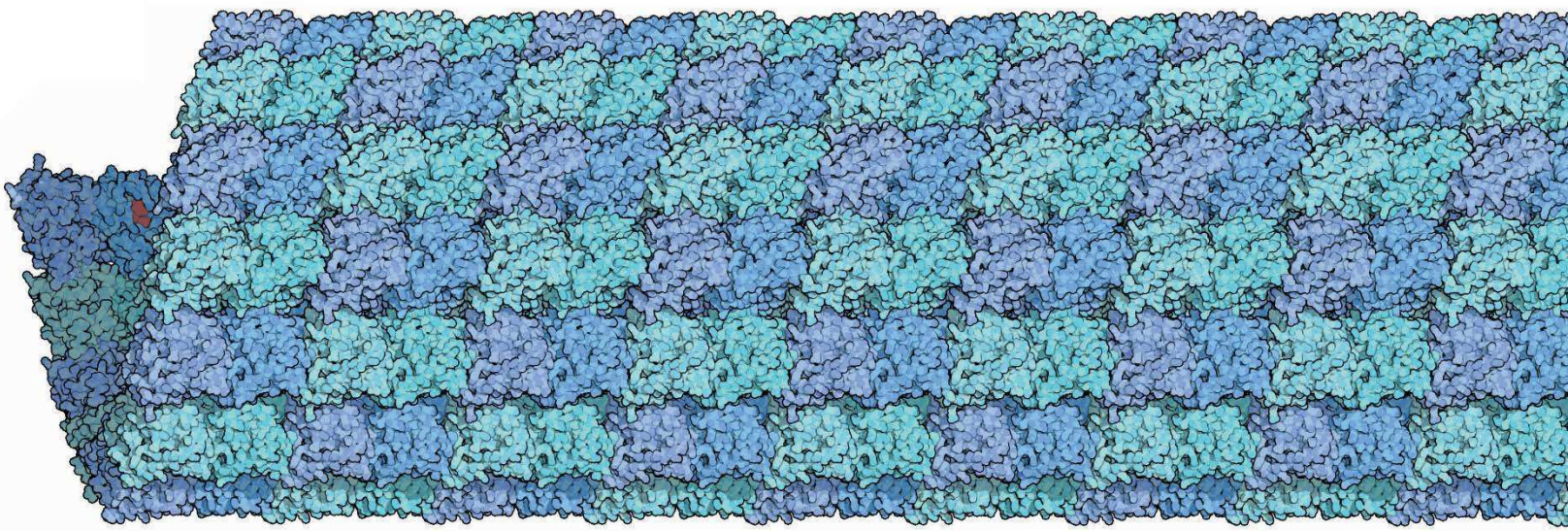
The global objective of this work is to explore the use of integrative modeling methods to determine the structure of macromolecules. The precise objectives of this thesis are the following:

1. Develop a method that integrates distance measurements between proteins to generate 3D models of multi protein complexes.

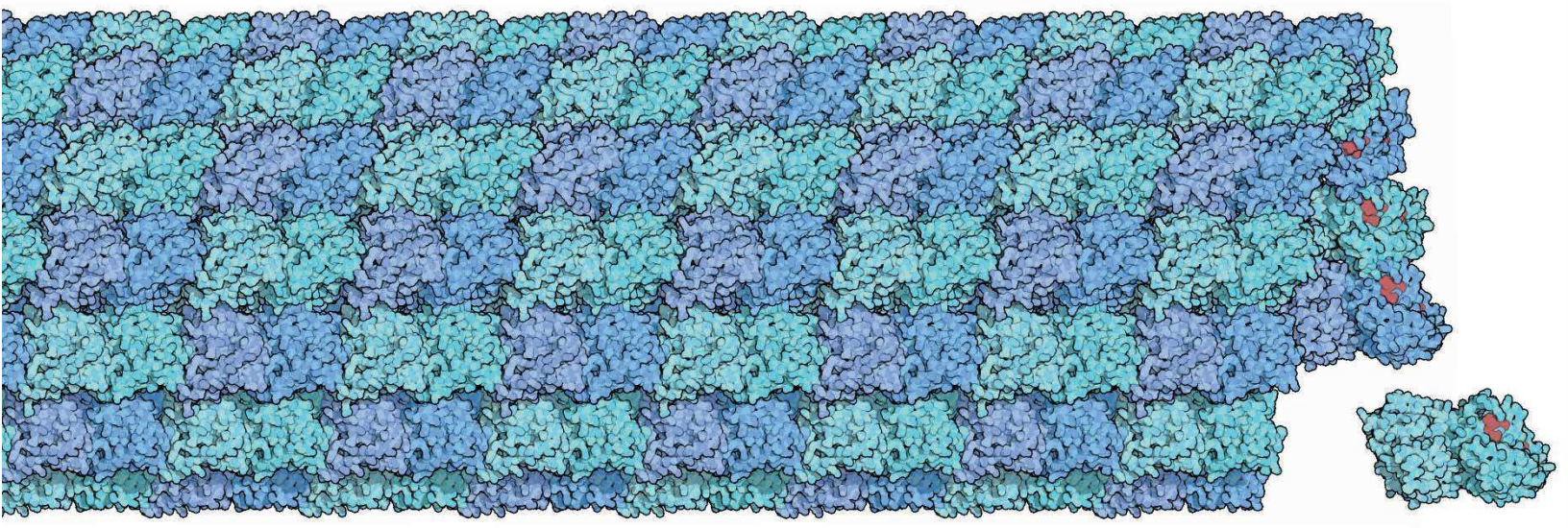
2. Apply the method in objective 1 to the exocyst multi protein complex.

3. Develop a method that infers distance data from 4C-seq data to generate 3D models of chromatin loci.

4. Apply the method in objective 3 to the HoxD locus in zebrafish, mouse and amphioxus, to the PAX3::FOXO1 fusion gene locus in human, Shh locus in wt and mutant mouse and to the Six2/3 locus in mouse and zebrafish.



Results



1

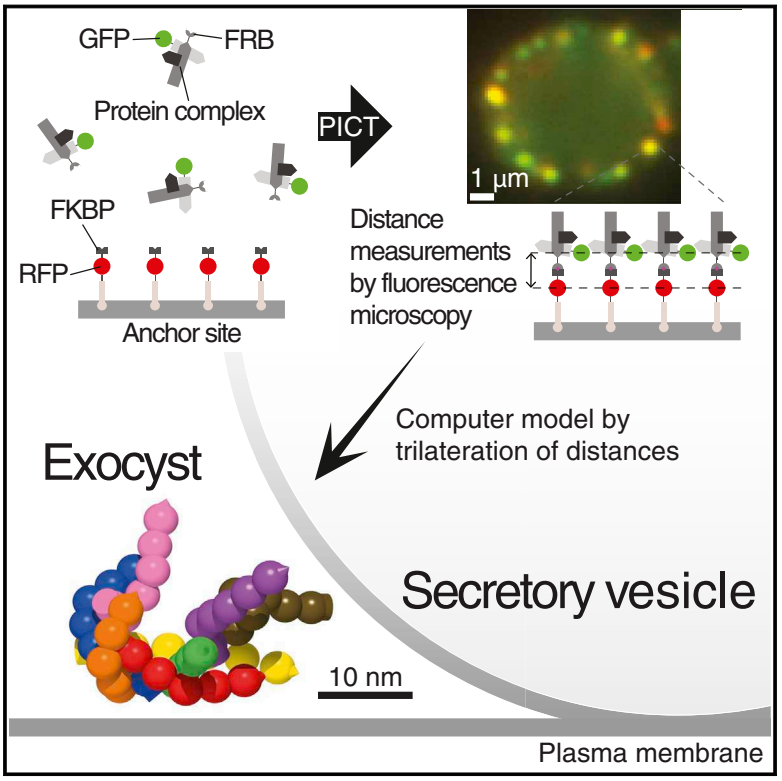
The In Vivo Architecture of the Exocyst Provides Structural Basis for Exocytosis

Article

Cell

The In Vivo Architecture of the Exocyst Provides Structural Basis for Exocytosis

Graphical Abstract



Authors

Andrea Picco, Ibai Irastorza-Azcarate, Tanja Specht, ..., Damien P. Devos, Marko Kaksonen, Oriol Gallego

Correspondence

damienpdevos@gmail.com (D.P.D.), marko.kaksonen@unige.ch (M.K.), oriol.gallego@irbbarcelona.org (O.G.)

In Brief

Microscopy-derived spatial constraints allow modeling of the exocyst structure in vivo.

Highlights

- An integrative approach reconstructs protein complexes in 3D through live-cell imaging
- We use this approach to reconstruct the exocyst complex bound to a vesicle in vivo
- Exocyst is a stable complex and regulatory proteins target its multimerization site
- We model how exocyst binds the vesicle allowing its contact with the plasma membrane

Picco et al., 2017, Cell *168*, 400–412
January 26, 2017
http://dx.doi.org/10.1016/j.cell.2017.01.004

CellPress

Cell

Article

The In Vivo Architecture of the Exocyst Provides Structural Basis for Exocytosis

Andrea Picco,[1,4,5] Ibai Irastorza-Azcarate,[2,5] Tanja Specht,[1] Dominik Böke,[1] Irene Pazos,[3] Anne-Sophie Rivier-Cordey,[4] Damien P. Devos,[2,*] Marko Kaksonen,[1,4,*] and Oriol Gallego[3,6,*]

[1]Cell Biology and Biophysics Unit, European Molecular Biology Laboratory (EMBL), Meyerhofstrasse 1, 69117 Heidelberg, Germany
[2]Centre for Organismal Studies, Heidelberg University, Heidelberg, Germany and Centro Andaluz de Biología del Desarrollo (CABD), Universidad Pablo de Olavide-CSIC, Carretera de Utrera km1, 41013 Sevilla, Spain
[3]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, c/ Baldiri Reixac 10, 08028 Barcelona, Spain
[4]Department of Biochemistry and NCCR Chemical Biology, University of Geneva, Quai Ernest Ansermet 30, 1211 Geneva, Switzerland
[5]Co-first author
[6]Lead Contact
*Correspondence: damienpdevos@gmail.com (D.P.D.), marko.kaksonen@unige.ch (M.K.), oriol.gallego@irbbarcelona.org (O.G.)
http://dx.doi.org/10.1016/j.cell.2017.01.004

SUMMARY

The structural characterization of protein complexes in their native environment is challenging but crucial for understanding the mechanisms that mediate cellular processes. We developed an integrative approach to reconstruct the 3D architecture of protein complexes in vivo. We applied this approach to the exocyst, a hetero-octameric complex of unknown structure that is thought to tether secretory vesicles during exocytosis with a poorly understood mechanism. We engineered yeast cells to anchor the exocyst on defined landmarks and determined the position of its subunit termini at nanometer precision using fluorescence microscopy. We then integrated these positions with the structural properties of the subunits to reconstruct the exocyst together with a vesicle bound to it. The exocyst has an open hand conformation made of rod-shaped subunits that are interlaced in the core. The exocyst architecture explains how the complex can tether secretory vesicles, placing them in direct contact with the plasma membrane.

INTRODUCTION

A mechanistic understanding of the cell requires the structural characterization of its biological complexes. This information is lacking for most cellular complexes due to the limitations of conventional approaches. In addition, most methods require purification of the complex and cannot determine the structure of complexes directly in the cellular environment. Thus, understanding the molecular mechanisms that mediate cellular processes requires the development of innovative hybrid approaches.

Exocytosis delivers cargos in secretory vesicles to the plasma membrane and to the extracellular space. The exocyst complex is responsible for specifically tethering the secretory vesicles to the appropriate plasma membrane sites during exocytosis. Mutations of exocyst subunits result in the accumulation of exocytic vesicles in the cell. This suggests that the exocyst acts before the soluble N-ethylmaleimide-sensitive factor attachment protein receptor (SNAREs) complex mediated fusion of the vesicle with the plasma membrane (Heider and Munson, 2012; Novick et al., 1980). The exocyst is necessary for cell growth, cell polarity, and the correct development of cellular structures (e.g., primary cilia in renal cells) (Heider and Munson, 2012). Exocyst malfunction is associated with many pathologies, such as Polycystic Kidney Disease (Fogelgren et al., 2011) and cancer (Sjöblom et al., 2006; Yamamoto et al., 2013). Despite the importance of the exocyst, the lack of structural information has prevented addressing the molecular mechanisms of exocyst function and regulation.

The exocyst is conserved throughout the eukaryotes and consists of one copy of each of its eight subunits (845 kDa in total): Sec3, Sec5, Sec6, Sec8, Sec10, Sec15, Exo70, and Exo84 (Heider et al., 2016; Hsu et al., 1996; TerBush et al., 1996). Several studies indicate that the exocyst binds to the secretory vesicle with Sec10 and Sec15 (Guo et al., 1999; Roth et al., 1998; Wiederkehr et al., 2004), while Sec3 and Exo70 bind to exocytic sites at the plasma membrane (Boyd et al., 2004; Dong et al., 2005; Finger et al., 1998; Yamashita et al., 2010). However, it is still not known how the exocyst can bind both membranes simultaneously without interfering in the subsequent SNARE-mediated fusion.

The information available on the overall exocyst structure is limited to quick-freeze deep-etch and negative stain electron microscopy (Heider et al., 2016; Hsu et al., 1998). These images showed that the exocyst is constituted of arms of different lengths, ranging from 10–35 nm, that join together at the core of the complex. However, the structure of the complex could not be determined. The crystal structures of the exocyst subunits cover only 26% of the whole complex (Dong et al., 2005; Hamburger et al., 2006; Jin et al., 2005; Sivaram et al., 2006; Wu et al., 2005; Yamashita et al., 2010). The structural data suggest that the subunits are rod shaped and formed by

helical-bundle repeats, with the N and C termini located at the opposite ends of the rod (Dong et al., 2005; Sivaram et al., 2006; Yamashita et al., 2010). However, we do not know how the subunits are organized within the complex and thus where the binding sites for the secretory vesicle and the plasma membrane are positioned in the exocyst.

Here, we describe an approach to determine the 3D architecture of multi-protein complexes in vivo and apply it to the exocyst. Fluorescence microscopy techniques can measure small distances between subunits of protein complexes tagged with fluorophores (Aravamudhan et al., 2014; Churchman et al., 2005, 2006; Clark et al., 2013; Gordon et al., 2004; Huang et al., 2010; Joglekar et al., 2009; Picco et al., 2015; Saffarian and Kirchhausen, 2008; Szymborska et al., 2013; Wan et al., 2009; Yildiz et al., 2003). However, the distances between subunits of a 3D complex are measured from projections onto 2D images, which complicates the interpretation of the measurements. In fact, the implementation of fluorescence microscopy to determine the 3D architecture of protein complexes has not been generalized. We followed up on the PICT (protein interactions from imaging complexes after translocation) technique to engineer yeast cells with immobile anchoring platforms, where the protein complex is recruited in controlled orientation upon inducible translocation (Gallego et al., 2013). We used one fluorophore to tag the anchoring platform, which acts as a landmark and a different fluorophore to tag, one at the time, the termini of each of the exocyst subunits. We analyzed live-cell images as in the SHREC method (single molecule high-resolution colocalization) to estimate the separation between the two fluorophores with a precision below 5 nm (Churchman et al., 2005; 2006). PICT allowed us to reproducibly measure the distances between these two fluorophores for the different orientations with which the complex was recruited. We could thus use these distances as coordinates to position the termini of the subunits in the 3D space by trilateration. The high precision of our measurements allowed us to integrate the subunit positions with the structural information available for each subunit (Russel et al., 2012) to reconstruct the complete 3D molecular architecture of the exocyst in vivo. The architecture of the exocyst provides mechanistic insight into vesicle tethering and raises new questions such as the coordination of several copies of the exocyst during this process.

RESULTS

Positioning the Exocyst Subunits with Respect to a Reference Point

We used fluorescence microscopy to determine the location of the exocyst subunits within the complex. First, we engineered yeast to induce the anchoring of the exocyst to static platforms that we then used as a reference point. We designed the anchoring platforms based on the clathrin adaptor protein Sla2. Sla2 molecules bind tightly to the plasma membrane, and their C-terminal part is exposed in the cytosol (Picco et al., 2015; Skruzny et al., 2012). When endocytosis is blocked by latrunculin A (LatA), Sla2 forms stable and immobile domains on a flat plasma membrane (Kukulski et al., 2012; Skruzny et al., 2012). To recruit the complexes, we used the rapamycin-induced heterodimerization of the FK506-binding protein (FKBP) and the FKBP-rapamycin binding (FRB) domain (Chen et al., 1995; Gallego et al., 2013) (Figure 1A). About 40–50 Sla2 molecules are present at each endocytic site (Picco et al., 2015). We estimated that, on average, about as many exocyst complexes are recruited to the anchor sites. We generated 80 yeast strains expressing the anchor (Sla2 fused to RFP and FKBP: Sla2-RFP-FKBP) and a specific combination of one exocyst subunit tagged with FRB (bait-FRB) and another subunit tagged with GFP at the N or C terminus (prey-GFP) (Figure 1B; Tables S1 and S2). All fusion proteins were expressed from their endogenous loci. All the strains grew normally, except those expressing Sec8 N-terminally tagged with GFP or Sec5-FRB, which were not included in this study. Since deletion of any of the exocyst subunits results in lethality or a severe growth defect (Wiederkehr et al., 2004), this indicates that the tagged proteins were functional.

Upon addition of both rapamycin and LatA, the bait-FRB was bound to the anchor (Figure S1), forcing the recruitment of the entire complex. We imaged yeast cells at their equatorial plane, where we could assume the observed anchoring platforms and the recruited complexes to be planar with the focal plane. Each anchoring platform formed a pair of fluorescent spots resulting from the anchor RFP molecules (red spot) and prey-GFP molecules (green spot) present in the platform. We systematically measured the separation between the centroids of the two spots in at least 60 anchoring platforms. The distance measurements follow a non-Gaussian distribution (Churchman et al., 2006), which we used to estimate the true separation between the RFP and GFP fluorophores with a precision of at least 5 nm (see STAR Methods; Figures 1, 2, and S2). Each bait-FRB used to recruit the complex imposed a specific orientation, with respect to the anchoring platform, depending on its position within the complex (Figure 1 and Figure S1A). This allowed us to image the complex recruited to the anchor with different orientations. For each orientation, we measured the distance from the anchoring platform of different subunit termini. We measured a total of 80 distances (Figure 2).

Determining the 3D Architecture of the Exocyst

We integrated the set of measured distances with the structural features of each subunit to determine the 3D architecture of the exocyst. We used the Integrative Modeling Platform (IMP), a suite of programs that integrates structural constraints derived from diverse experiments to determine the structure of macromolecular complexes (Russel et al., 2012) (Figure 3 and STAR Methods).

We used the 80 distances as restraints to trilaterate in the 3D space the positions of the anchor RFP tag and of the GFP tags fused to the exocyst subunits (Figures 3A and STAR Methods). We repeated the trilateration 10,000 times starting from randomized initial positions. Then, we collected the configurations of fluorophore positions with the best IMP scores and that were compatible with our distance restraints. All of these fluorophore positions clustered in two populations of solutions that are mirror images of each other (Figure S3). Therefore, one of the two mirror groups of solutions is likely to be representative of the positions of the fluorescent tags in the exocyst complex. As our

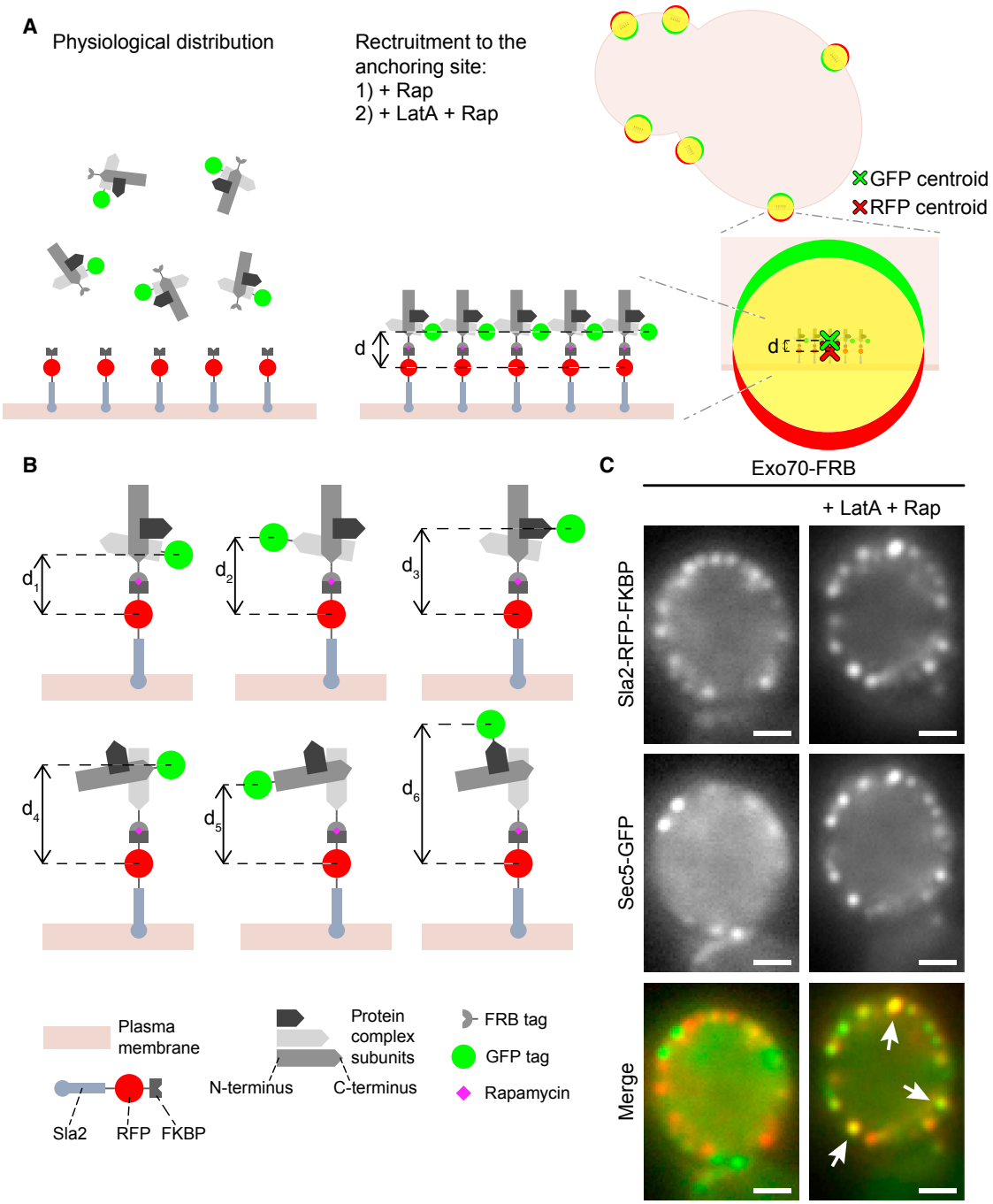


Figure 1. Experimental Setup

(A) Schematic representation of the approach used to measure distances between the fluorophores in the Sla2-RFP-FKBP and the subunits of the complex tagged with GFP. A subunit of the complex is tagged with FRB (bait-FRB). Rapamycin induces the heterodimerization of FKBP and FRB, and the subsequent addition of LatA stabilizes Sla2-RFP-FKBP at the endocytic sites. The complex is thus recruited to immobile Sla2-RFP-FKBP. Sla2-RFP-FKBP and prey-GFP molecules can now be imaged as a pair of diffraction limited spots. We measure the separation "d" between the centroids of the two spots. See also Figures S1A and S2.

(B) Combining different bait-FRBs and prey-GFPs (tagged with GFP either at its N or C terminus) allows measuring the distance between the Sla2-RFP-FKBP and each termini of the subunits when the complex is recruited in different orientations.

(C) Recruitment of Sec5-GFP to Sla2-RFP-FKBP anchoring platforms. Sec5 was fused at the C terminus to GFP and recruited using Exo70-FRB as bait. Arrows point at anchoring platforms, where Sec5-GFP colocalizes with Sla2-RFP-FKBP fluorescent spots upon addition of rapamycin and LatA. Scale bars are 1 μm long.

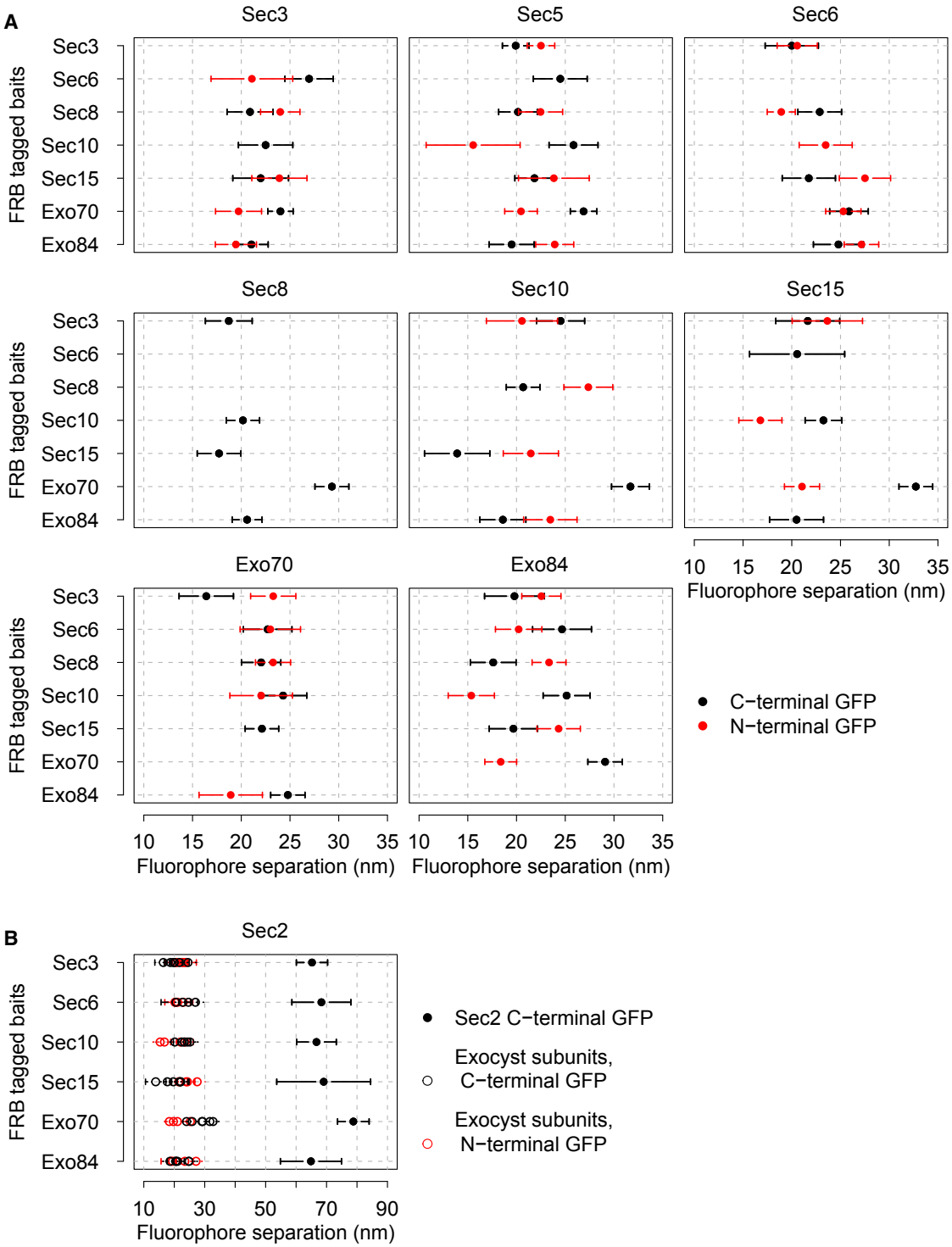


Figure 2. Dataset of Distances between the Anchor-RFP-FKBP and Prey-GFPs

(A) Distances between the fluorophores in the prey-GFP and Sla2-RFP-FKBP for the recruitment mediated by each bait-FRB.

(B) The distance between the GFP tag at the C terminus of Sec2 and Sla2-RFP-FKBP, using the indicated subunits as bait. The empty spots are all the C- and N-terminal (black and red respectively) distances that we measured for all the exocyst subunits (Figure 2A), which are plotted here as a comparison with Sec2-GFP separation from the anchor site (filled spots).

Each box corresponds to the prey-GFP that titles it. Error bars show the SE. For the distance values, see Table S2.

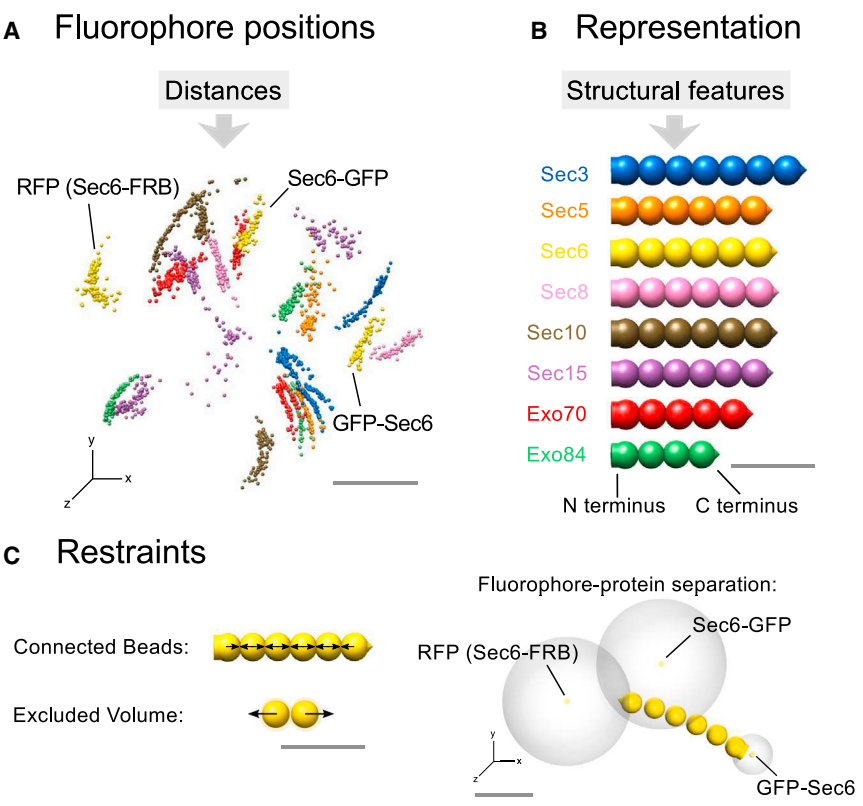


Figure 3. Protocol to Compute the 3D Architecture of the Exocyst

(A) Computed fluorophore positions that are compatible with the distances in Figure 2A used as restraints. As an example, we highlight the positions of Sec6 C-terminal tag (Sec6-GFP), Sec6 N-terminal tag (GFP-Sec6), and the anchor when Sec6 was used as bait (Sec6-FRB). See also Figure S3 and Table S3.

(B) Exocyst subunits were represented as a string of beads according to their structural features. See STAR Methods.

(C) Restraints used to elucidate the exocyst architecture: Connected Beads (i.e., consecutive beads in a string can not be separated by more than a maximum distance), Excluded volume (i.e., two beads can not occupy the same volume), and Fluorophore-protein separation (i.e., N and C termini of exocyst subunits must fall within a maximum distance from the location of the fluorophores they are flagged with. This maximum distance is represented as transparent gray spheres). See also Table S3.

(D) IMP integrates all the data (A, B, and C) to reconstruct the architecture of the exocyst. Among all solutions that satisfy all of our restraints (right, represented by a subset of 100 solutions randomly chosen), we chose the solution with best IMP score to illustrate the main features of the exocyst (left). See also Figures S4 and S5.

Subunits are color-coded in blue (Sec3), orange (Sec5), yellow (Sec6), pink (Sec8), brown (Sec10), purple (Sec15), red (Exo70), and green (Exo84). Scale bars correspond to 10 nm.

data does not allow to distinguish the chirality of the exocyst complex, we arbitrarily chose one of the two populations of fluorophore positions for subsequent analysis.

As atomic structures are not available for most of the exocyst subunits, we used flexible strings of beads to represent each subunit. Each bead was 3.5 nm in diameter, the mean volume of a folded fragment of 120 residues in the PDB (Shen et al., 2005). The size and the structural features of each subunit defined the number of beads used to represent it (Figure 3B and STAR Methods). We constrained consecutive beads in each string to be connected, and we imposed that two beads cannot occupy the same volume (Figure 3C and STAR Methods). To reconstruct the architecture of the exocyst, we then used the fluorophore positions to locate the N- and C-terminal beads of the exocyst subunits they were fused to (Figure 3C and STAR Methods). To locate the N terminus of Sec8, which we could not tag, we used the known interaction between Sec8 and Sec6 (Guo et al., 1999; Sivaram et al., 2006) as a constraint (see STAR Methods and Table S3). We sampled the space of solutions exhaustively by repeating the reconstruction of the exocyst 50,000 times, each time starting from a randomized initial position for each bead.

We selected the 200 solutions that fulfilled all the restraints and had the best IMP score (Figures 3D and 4). The exocyst subunits showed a similar organization in all these solutions (Figures 4A and S3C). Sec8 N terminus and the central part of Sec5 had the highest variability in their location due to the lack of strains with N-terminally GFP-tagged Sec8 or Sec5-FRB. The central region of Sec3 also presented some variability, possibly due to its location near the less-resolved Sec5 and Sec8. The different solutions could be clustered in six groups, which differ only slightly in the organization of Sec3 and Exo70 subunits (Figure S4).

To verify the reliability of our method, we reconstructed the molecular architecture of the conserved oligomeric Golgi (COG) complex for seven of its eight subunits using the same approach that we used to determine the architecture of the exocyst. The molecular architecture of this multisubunit tethering

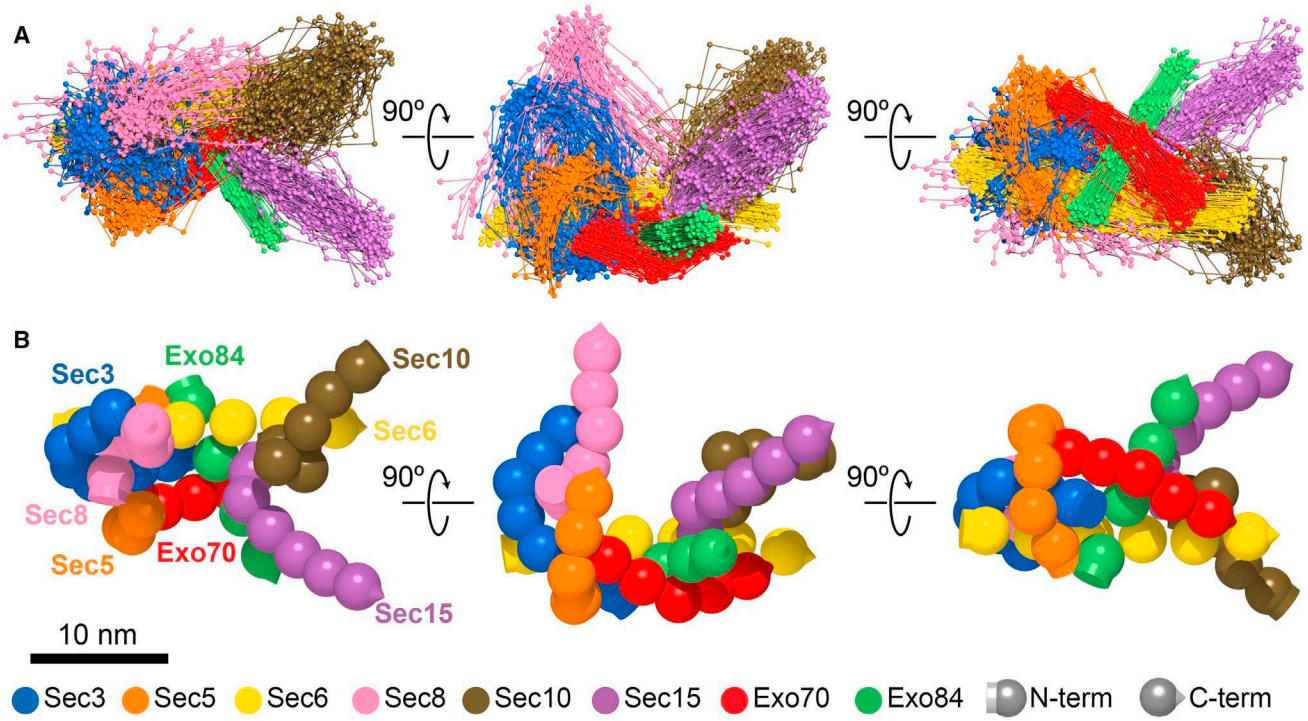


Figure 4. The 3D Architecture of the Exocyst

(A) Different views of 100 randomly selected solutions of the exocyst architecture that satisfy all of the restraints. Exocyst subunits are represented by small beads connected with a line. See also Figures S3, S4, and S6 and Movie S1.

(B) The solution with best IMP score is used as a representative model for the 3D architecture of the exocyst, and it is shown in the same views as (A). The volume of each bead is approximately equivalent to 120 amino acids. See also Movie S2.

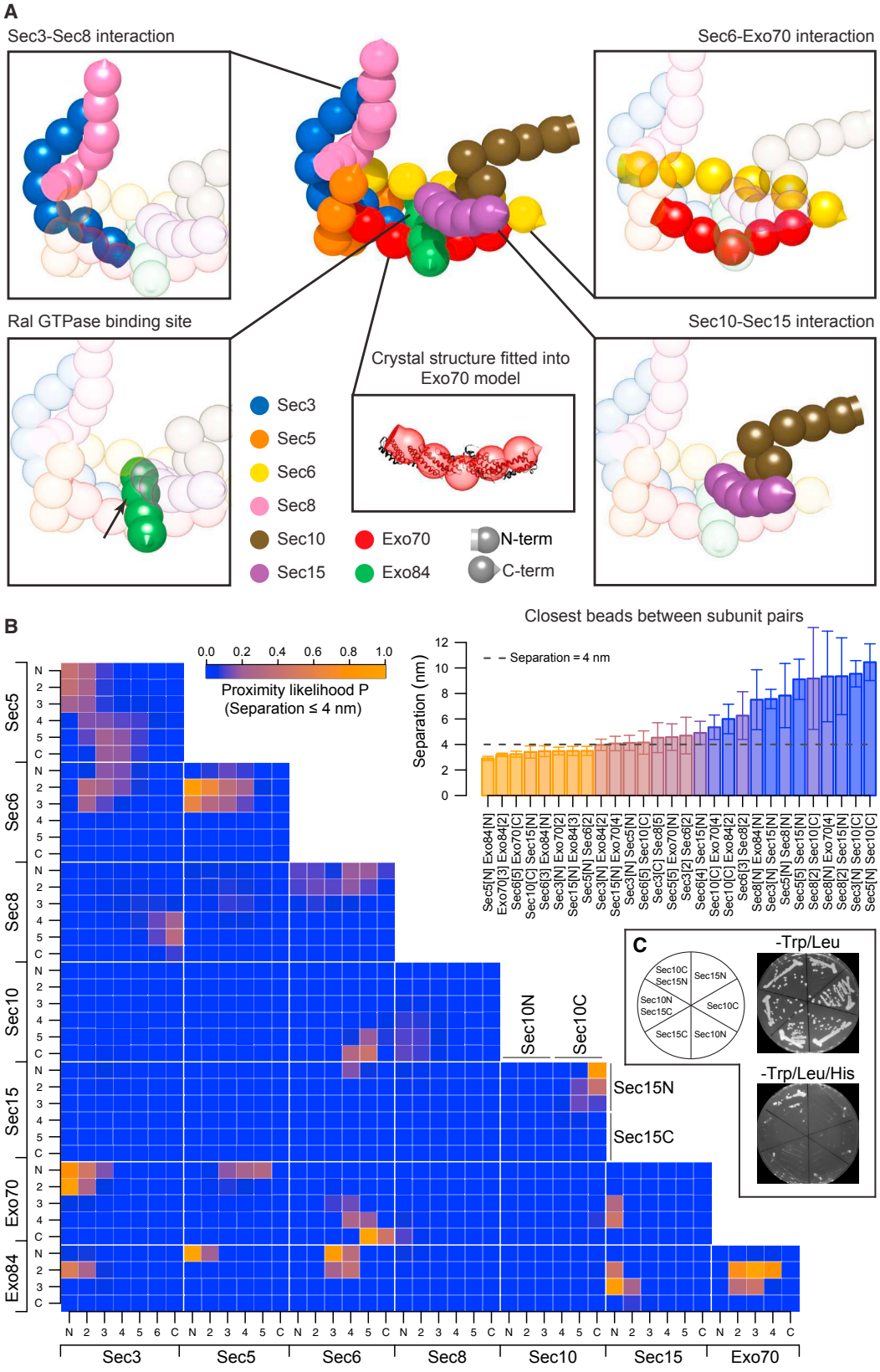
complex was recently determined by negative stain electron microscopy (EM) (Ha et al., 2016). Our 3D reconstruction of the COG complex shows that the spatial arrangement of the subunits is equivalent to the published COG architecture: in both studies, the subunits share the same neighbors, and they are oriented with most of the N termini interlaced at the core of the complex, while the C termini protrude outward. Thus, the COG subunits organize in legs that extend from the core of the complex toward different directions (Figure S5 and STAR Methods).

We next investigated the capability of the computational reconstruction to detect inconsistencies in the dataset of measured distances for the exocyst subunits. We repeated the procedure with our distances randomly assigned to different bait and prey pairs or with a dataset of distances where one of the measurements had been shortened by 18 nm. In both cases, we could not find any solution that satisfied all restraints. When we simulated less-pronounced inconsistencies in the dataset (i.e., swapping two distance measurements between two different pairs of fluorescent tags or shortening a single distance measurement between 6 and 16 nm), the trilateration of the fluorophore positions gave solutions that did not converge in two mirror image solutions and were different from the solutions obtained using the real distances (Figure S3 and STAR Methods). Overall, these results provide confidence in our reconstruction of the exocyst and indicate that the approach is able to efficiently detect inconsistencies in the dataset of 80 distance measurements for the exocyst subunits.

The Molecular Architecture of the Exocyst Complex In Vivo

As a benchmark for the exocyst architecture, we used the Exo70 structure, the only subunit for which the almost complete structure has been solved (Dong et al., 2005; Hamburger et al., 2006). In the cell, Exo70 shows a slightly bent conformation with its N and C termini located 11.3 ± 0.5 nm apart (median ± SE, n = 200; see STAR Methods), which is consistent with the 11.8 nm of separation measured in the Exo70 crystal structure (Dong et al., 2005; Hamburger et al., 2006) (Figure 5A).

Protein-protein interactions among exocyst subunits have been extensively studied in vitro (Table S4). If two subunits interact, they should appear nearby in the exocyst architecture. From the 200 solutions, we calculated the likelihood of the different subunits to be in close proximity (Figure 5 and STAR Methods). Remarkably, 7 out of the 9 reported direct interactions occur between subunits that are in close proximity in our reconstruction (Figure 5; Table S4). Among these interactions, we did not consider those reported for Sec8 because they were used to position Sec8 within the complex. We then investigated the binding between subunits or fragments of them that are in close proximity in the exocyst architecture, but have not been shown to interact. Using yeast two-hybrid assays, we found that the



(legend on next page)

Sec10 C-terminal fragment binds the Sec15 N-terminal fragment, while no interaction was detected between the Sec10 N-terminal and Sec15 C-terminal fragments, which is in agreement with our exocyst reconstruction (Figure 5C). However, we could not detect an interaction of Exo84 with Exo70 and Sec15 N terminus, which have been shown to interact in the human exocyst and appear to be in close proximity in our model (Matern et al., 2001). We also measured the overall volume for the exocyst to be 1,500 nm^3 (STAR Methods), which is similar to the volume estimated from EM data (~1,800–2,200 nm^3) (Heider et al., 2016). In summary, the architecture of the exocyst complex that we determined in vivo recapitulates various published interaction and structural data not used for its determination, providing confidence in the overall accuracy of our results.

Our reconstruction of the in vivo exocyst architecture showed that the exocyst subunits are rod shaped with their N and C termini located at opposite ends of the rod as previously hypothesized based on crystallographic and computational studies (Croteau et al., 2009; Hamburger et al., 2006; Yamashita et al., 2010). All of the subunits except Sec10 are attached to the core of the complex with their N-terminal parts, while their C-terminal ends project outward. Sec10 organization is inverted; its C-terminal part locates in the core, and its N-terminal end projects outward. The exocyst is organized in arms of different lengths, which gather in the central core and project in different directions (Figure 4 and Movies S1 and S2). The distal parts of the arms in the periphery of the complex might change their conformation without affecting the rest of the complex. Our reconstruction is consistent with the EM images of the purified complex and provides a structural basis to explain the flexibility of the exocyst (Heider et al., 2016; Hsu et al., 1998).

The Exo70 and Sec6 subunits form a V-shaped dimer, with their C-terminal parts interacting at the periphery of the complex. Exo70 and Sec6 N termini are separated but are embracing the core of the complex that links them together. Sec3 and Sec8 mirror the Exo70 and Sec6 organization, interacting with their C termini and interlacing their N termini in the core of the complex (Figure 5A). Interestingly, Exo70 is structurally related to Sec3 C terminus, and Sec8 C terminus is structurally related to Sec6 C terminus (see STAR Methods and Figure S6A). When mapped on the exocyst architecture, these fragments present a symmetric distribution in the complex (Figure S6B). Exo84 and Sec5 subunits extend through the core of the complex and are adjacent to both Exo70-Sec6 and Sec3-Sec8 dimers (Figure 5). Sec10 and Sec15 form a sub-complex that is less interlaced with the rest of the complex. Located on top of the core, Sec10-Sec15 subcomplex is proximal to only Exo70, Sec6, and Exo84 (Figure 5).

The exocyst architecture provides mechanistic insight about the regulation of its assembly. Exo84 has been shown to be required for holding the Sec10-Sec15 dimer together with the rest of the complex (Heider et al., 2016). Exo84 is known to be targeted by different signaling pathways that regulate exocyst function and assembly (Jin et al., 2005; Luo et al., 2013; Moskalenko et al., 2003). For instance, in mammals, Ral GTPases bind Exo84 to control exocyst function. The interaction of Ral GTPases governs exocyst function in a broad panel of cellular events such as cell migration (Rossé et al., 2006), autophagy (Bodemann et al., 2011), or postsynaptic membrane growth (Teodoro et al., 2013). Despite the fact that Ral GTPases have not been described in yeast, the Ral GTPase binding domain of Exo84 is conserved and is located where Exo70-Sec6, Sec3-Sec8, and Sec10-Sec15 dimers meet at the core of the complex. This suggests that in mammals, Ral GTPases directly interact with the core of the complex to regulate exocyst assembly (Figure 5A). During mitosis, Cdk1 phosphorylates Exo84 to inhibit the assembly of the exocyst (Luo et al., 2013). Thus, the core of the complex may be a hub for the regulation of exocyst assembly and function.

Exocyst Forms a Stable Complex

The nature of the exocyst assembly in the cell is controversial. Sec3 and Exo70 have been proposed to bind the plasma membrane independently of the exocyst assembly (Boyd et al., 2004). In this model, the full complex forms only when the vesicle arrives at exocytic sites carrying the other six subunits, including Sec5. However, recent biochemical data suggested that the exocyst is a very stable complex (Heider et al., 2016). In the exocyst, each subunit is adjacent to four other proteins on average, mostly in the central core of the complex, suggesting that the subunits are highly interlaced and strongly bound together. We studied the stability of the exocyst complex once it is recruited to the anchoring platforms using FRAP (Gallego et al., 2013). We locally photobleached subunits recruited to an anchoring platform and followed the fluorescence recovery over time. Using Exo70-FRB as bait, 75% of the exocyst showed no disassembly during 7 min, indicating that, at least under the recruitment conditions, the exocyst is a stable complex (Figure 6A). We then used fluorescence cross-correlation spectroscopy (FCCS) to quantify the fraction of Sec3 that is free in the cytosol and the fraction that is associated to Sec5. Interestingly, the vast majority of Sec3 is in complex with Sec5, with a dissociation constant (K_D) of 2.59 nM (Figure 6B), indicating that in the cytosol, the exocyst

Figure 5. Structural Features of the Exocyst

(A) Close-up views of different structural features of the exocyst architecture. Sec3 and Sec8 C termini interact, forming an arm on one side of the complex. Sec6 and Exo70 C termini interact, forming another arm on the opposite side of the exocyst. The vesicle binding subunits Sec10 and Sec15 form a subcomplex. The conformation of Exo70 matches the atomic structure determined by X-ray crystallography (PDB entry 2b1e_A). The arrow indicates the approximate location of the Ral GTPase binding site in Exo84.

(B) The likelihood that beads belonging to pairs of different subunits are in close proximity (i.e., $\leq$ 4 nm). The bar plot shows the average separation between the closest beads among all subunit pairs; the dashed line marks the 4 nm separation; and the error bars represent SD. See also Table S4.

(C) Yeast two-hybrid assay for the interaction between Sec10 C terminus (Sec10C, aa 491-871) and Sec15 N terminus (Sec15N, aa 1-381) and between Sec10 N terminus (Sec10N, aa 1-490) and Sec15 C terminus (Sec15C, aa 382-910).

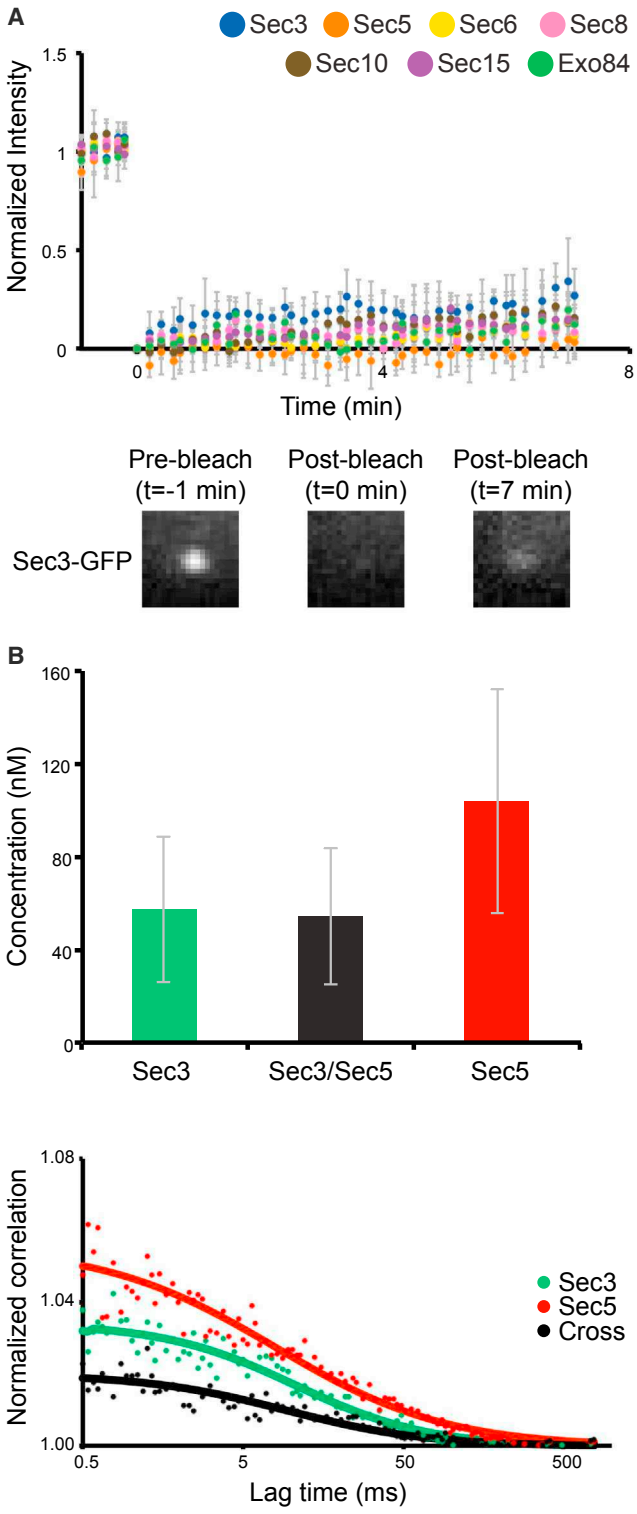


Figure 6. Stability of the Exocyst

(A) FRAP assay to quantify the stability of the exocyst recruited to the anchoring platform using Exo70-FRB as bait. Data from each strain is color coded according to the subunit that was expressed as GFP fusion. The points represent the mean of at least four FRAP experiments, and the error bars indicate the SD. A frame from the GFP channel of a representative FRAP experiment shows Sec3-GFP recruited to an anchoring site before photobleaching, immediately after photobleaching, and at the end of the measurements.

(B) Cytoplasmic FCCS measurement of Sec3-GFP with Sec5-RFP. The top plot shows the average concentration of Sec3 (green), Sec5 (red), and the Sec3 bound to Sec5 (black). Measurements were performed in 15 cells. Error bars indicate SD. FCCS example traces of Sec3-GFP, Sec5-RFP, and the cross-correlation of both proteins is shown at the bottom plot. Correlation curves were normalized.

is strikingly stable, in agreement with biochemical data (Heider et al., 2016).

A Model for the Tethering of Secretory Vesicles

The exocyst has not been purified bound to the secretory vesicle nor has this interaction been reconstituted in vitro. For instance, it is not clear how the exocyst can tether vesicles to the plasma membrane in a way that the complex does not interfere with their fusion. To gain insight into the tethering process, we determined the architecture of the exocyst associated to the secretory vesicle in living cells. We imaged Sec2-GFP, a marker for the secretory vesicles, in six additional strains expressing different exocyst subunits as bait-FRB (Figure S7A and Table S1). Sec2-GFP was efficiently recruited to the anchoring platform in all of the strains, indicating that the exocyst is also capable of binding to vesicles when it is anchored. We measured the distances between Sec2-GFP and the anchoring platform in the six strains and integrated the distances with those that we measured for the N and C termini of the exocyst subunits (Figure 2B and Table S2). We used the same procedure to determine the architecture of the exocyst associated to a secretory vesicle. The architecture of the exocyst is identical to the architecture determined with the subunits only (Figure 7A). Sec2 is 53 nm away from the exocyst, and Sec10 and Sec15 are the closest subunits to Sec2, which is in agreement with the ability of the Sec10-Sec15 sub-complex to associate with the secretory vesicle (Guo et al., 1999; Roth et al., 1998; Wiederkehr et al., 2004) (Figure 7B). Interestingly, Sec10 and Sec15 are located roughly at the opposite side of the exocyst with respect to the N terminus of Sec3 and the C-terminal half of Exo70, which are targeting the exocyst to the plasma membrane (Boyd et al., 2004; Dong et al., 2005; Finger et al., 1998; Yamashita et al., 2010) (Figure 4).

We generated a model for the tethering of secretory vesicles where we approximated the secretory vesicle to a sphere of 50 nm of radius, the average radius of a secretory vesicle (He et al., 2007; Walworth and Novick, 1987). The position determined for Sec2 corresponds to the average location of all Sec2-GFP molecules present on the vesicle. Assuming that Sec2-GFP is on average evenly distributed on its surface, the position determined for Sec2 also defines the position of the vesicle center. We then oriented the exocyst to allow the N-terminal domain of Sec3 and the C-terminal domain of Exo70 to bind the plasma membrane (Boyd et al., 2004; Dong et al., 2005; Guo et al., 1999; Yamashita et al., 2010). Remarkably, the position of the secretory vesicle with respect to the exocyst suggests that the exocyst can bind the vesicle with Sec10-Sec15 and the membrane with Sec3 and Exo70 simultaneously, while the two membranes are in direct contact. The exocyst is thus positioned at the side of

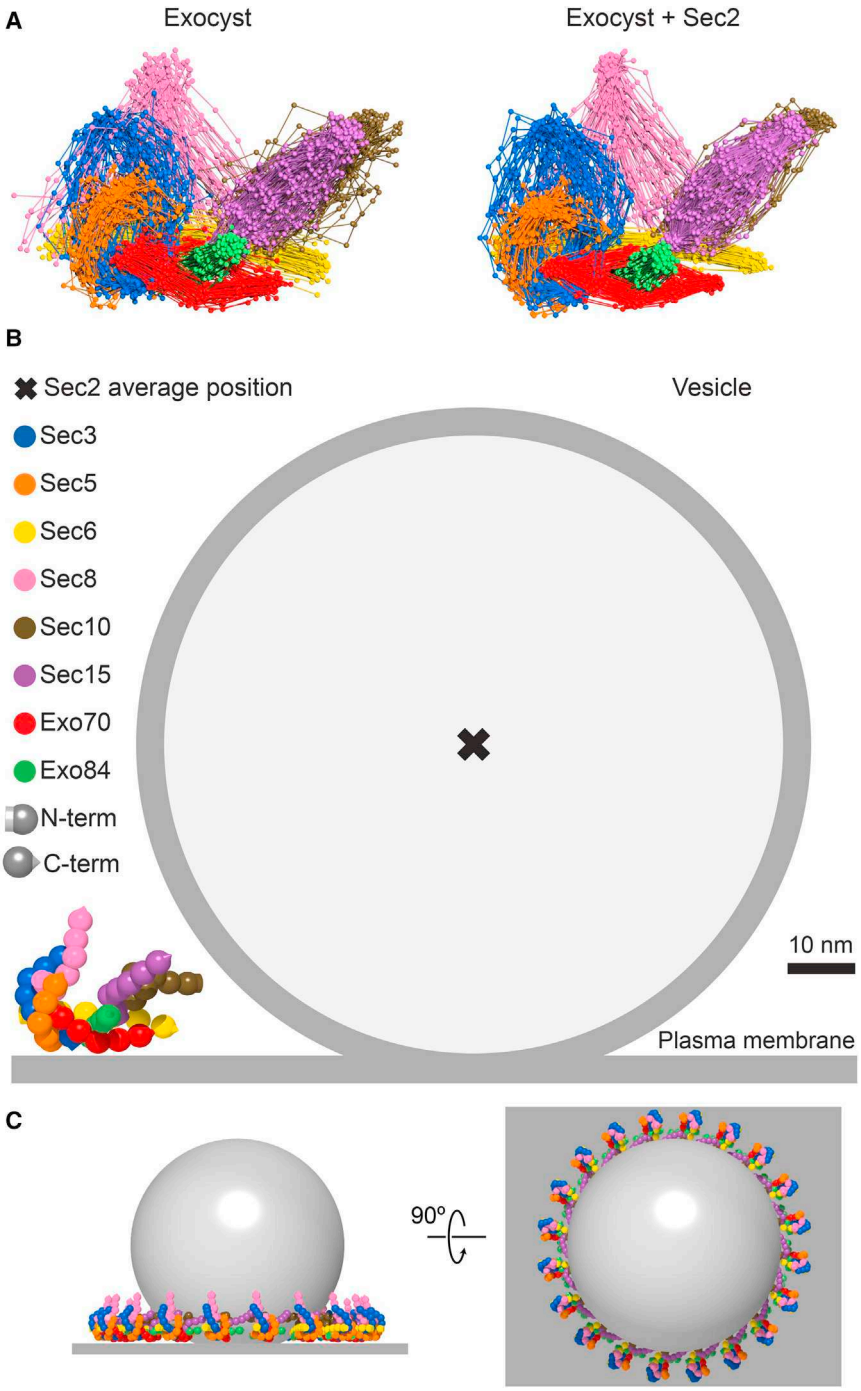


Figure 7. Model for the Vesicle Tethering Mediated by the Exocyst

(A) The 3D architecture of the exocyst reconstructed without Sec2 distances (left) and with Sec2 distances (right).

(B) Schematic representation of the exocyst structure bound to the plasma membrane and a vesicle with a radius of 50 nm. The black cross indicates the average position of the GFP fused to Sec2 C terminus. See also Figure S7.

(C) Up to 20 copies of the exocyst could participate simultaneously in vesicle tethering, forming a ring that surrounds the interface between the vesicle and the plasma membrane.

the contact surface between the vesicle and the plasma membrane and does not interfere with the subsequent membrane fusion (Figure 7B). Notably, the elongated shape of the Sec10-Sec15 sub-complex keeps the vesicle away from the core of the complex. Sec6 extends its C-terminal part from the core toward the cavity left between the exocyst, the vesicle, and the plasma membrane. Since the Sec6 C-terminal part binds and activates the SNARE complex during vesicle tethering (Dubuke et al., 2015), the architecture of the exocyst suggests that the machinery in charge of fusing the vesicle with the plasma membrane is assembled in this cavity. Indeed, in our model, the cavity between the exocyst, the vesicle, and the plasma membrane is large enough to fit a complex of the three exocytic SNAREs (Figure S7B).

The number of exocyst complexes involved in the tethering of a vesicle is not known. Each exocytic site at the plasma membrane consists of a single vesicle-tethering event at a time (Donovan and Bretscher, 2015). We measured an average of 13 ± 1 molecules of Sec5, 17 ± 2 molecules of Sec6, 15 ± 1 molecules of Sec10, 16 ± 2 molecules of Sec15, and 13 ± 1 molecules of Exo84 (mean ± SE; STAR Methods and Figure S1B) at each exocytic site. We used our model to hypothesize how several exocyst complexes could cooperate during vesicle tethering. The exocyst is displaced from the interface between the vesicle and the plasma membrane. Therefore, many exocyst complexes could bind both the vesicle and the plasma membrane simultaneously. A maximum of ~20 complexes could be accommodated in a ring around the contact area between the two membranes (Figure 7C). One side of this ring would then dock the vesicle, while the other would bind the plasma membrane, allowing the two membranes to establish a direct contact through the central hole of the ring (Figure 7C). This ring organization resembles the organization previously suggested for synaptotagmin during exocytic vesicle fusion (Wang et al., 2014).

DISCUSSION

Understanding the complexity of the cellular machinery requires the development of hybrid approaches. Our approach combines

cell engineering, quantitative fluorescence microscopy, and computational integration of structural data to determine the architecture of protein complexes in living cells. The precision of our measurements is in the nm scale. Therefore, this approach is particularly suited for the study of large multisubunit assemblies. A high number of combinations of baits and preys allows us to generate a large dataset of distances that synergistically constrain the space of possible architectures. In addition, this integrative approach could benefit from the combination of data from EM, crystallography, or crosslinking coupled to mass spectrometry (Erzberger et al., 2014) to increase the precision of the reconstruction. The method requires that the studied complex can be recruited to the anchoring platform in quantities that are large enough to be imaged. This might be a limitation, for instance, for nuclear assemblies or complexes that contain transmembrane proteins. Our approach tolerates, but cannot resolve, the conformational variability of the recruited complex during the imaging time. However, it allows the structural characterization of interactions between the target complex and other cellular components that cannot be purified together or reconstituted in vitro. This information is fundamental to understanding the mode of action of the macromolecular assemblies in the cellular context.

We used this approach to reconstruct the architecture of the exocyst complex directly in living cells. The exocyst is composed of rod-shaped subunits and has a conformation similar to an open hand. Each subunit contributes with one end of the rod to form the core of the complex at the palm of the open hand. The other ends of the rods are exposed at the periphery of the complex, where they form the fingers of the hand, which may be flexible. The flexibility was observed in EM images of the purified exocyst (Heider et al., 2016; Hsu et al., 1998). Exo84 and Sec5 are mostly embracing the core of the complex, and they project little to the periphery of the complex, which would limit their flexibility. Instead, the Sec10-Sec15 subcomplex is situated on top of the core, which may allow them to undergo larger conformational changes. The symmetry we described between Exo70-Sec6 and Sec3-Sec8 supports the idea that the ancestral exocyst complex was composed of multiple copies of fewer proteins that duplicated and diverged (Croteau et al., 2009; Dacks et al., 2008).

We showed that the exocyst subunits form an intricate mesh of interactions at the core of the complex. This explains why all the subunits are critical for exocyst function, although only half of them are directly involved in vesicle and membrane binding. The core of the complex hosts sites that are phosphorylated by Cdk1 during the cell cycle and a conserved domain that is targeted by Ral GTPases in mammals, suggesting that the core is the hub of pathways that regulate the complex assembly. The architecture of the exocyst provides a structural basis to tackle the mechanisms that regulate exocyst function in normal conditions but also under pathological conditions, such as cancer, where these regulatory mechanisms are affected.

The organization of the subunits within the complex allows the exocyst to remain at the side of the interface between the vesicle and the plasma membrane without interfering in their subsequent fusion. We hypothesize that in this manner, the exocyst can induce the assembly of the exocytic SNARE complex in the cavity between the exocyst, the vesicle, and the plasma membrane.

The position of the exocyst next to the contact area between the vesicle and the plasma membrane suggests that several exocyst copies could cooperate during vesicle tethering. Indeed, we measured on average about 14 exocyst complexes at sites of vesicle fusion. Although we cannot say whether they are all actively tethering the vesicle, recent data shows that the residence time at the exocytic site of the exocyst subunits is the same as the time a vesicle remains tethered (Donovan and Bretscher, 2015), suggesting that they must be somehow coordinated. Therefore, it is plausible that several exocyst copies form a ring around the interface between the vesicle and the plasma membrane with an organization similar to the one previously suggested for synaptotagmin (Wang et al., 2014). The cooperation of several exocyst complexes is an exciting possible mechanism to ensure the fidelity in recruiting the vesicle to the appropriate exocytic site on the plasma membrane.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Yeast strains and plasmids
  - Microscope setup
  - Imaging
  - Image processing and distance measurements
  - Exocyst structure determination
  - Controls
  - COG complex structure determination
  - Structurally related fragments, comparative modeling and volume
  - Calculation of Sec2 position
  - Quantification of exocyst subunit abundances, sample preparation
  - FRAP
  - FCCS
  - Yeast two-hybrid
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Image processing and distance measurements
  - Outlier rejection and distance estimation
  - Estimate of the likelihood of subunits being in close proximity
  - Quantification of exocyst subunit abundances, image analysis
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures, five tables, and two movies and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2017.01.004.

AUTHOR CONTRIBUTIONS

O.G., A.P., D.P.D., and M.K. designed the research; O.G., A.P., T.S., D.B., I.P., A.-S.R.-C., I.I.-A., D.P.D., and M.K. conducted the experiments and performed the analysis; and O.G., A.P., I.I.-A., D.P.D., and M.K. discussed results and wrote the manuscript.

ACKNOWLEDGMENTS

We thank Carlos Fernández-Tornero, Christoph W. Müller, Maya Topf, Raúl Méndez, Ignasi Fita, Ori Avinoam, Camilla Godlee, Wanda Kukulski, Bernd Klaus and Wolfgang Huber for critical discussions. Anne-Claude Gavin, Arun Kumar, Marc Abella, Nereida Jiménez, Amy J. Curwin, and Michael Knop for expert help and the sharing of reagents. O.G. was funded by the Ramón y Cajal program (RYC-2011-07967) and a grant from MINECO (BFU2012-36385). IRB Barcelona is the recipient of a Severo Ochoa Award of Excellence from MINECO (Government of Spain). A.P. acknowledges an EIPOD fellowship and M.K. the Swiss National Science Foundation grant (31003A_163267) and NCCR Chemical Biology. D.P.D. and I.I.-A. are financed by the Junta de Andalucía, Captación del Conocimiento para Andalucía (C2A) program.

Received: February 9, 2016
Revised: October 18, 2016
Accepted: January 5, 2017
Published: January 26, 2017

REFERENCES

Aravamudhan, P., Felzer-Kim, I., Gurunathan, K., and Joglekar, A.P. (2014). Assembling the protein architecture of the budding yeast kinetochore-microtubule attachment using FRET. Curr. Biol. *24*, 1437–1446.

Bodemann, B.O., Orvedahl, A., Cheng, T., Ram, R.R., Ou, Y.-H., Formstecher, E., Maiti, M., Hazelett, C.C., Wauson, E.M., Balakireva, M., et al. (2011). RalB and the exocyst mediate the cellular starvation response by direct activation of autophagosome assembly. Cell *144*, 253–267.

Boeke, D., Trautmann, S., Meurer, M., Wachsmuth, M., Godlee, C., Knop, M., and Kaksonen, M. (2014). Quantification of cytosolic interactions identifies Ede1 oligomers as key organizers of endocytosis. Mol. Syst. Biol. *10*, 756.

Boyd, C., Hughes, T., Pypaert, M., and Novick, P. (2004). Vesicles carry most exocyst subunits to exocytic sites marked by the remaining two subunits, Sec3p and Exo70p. J. Cell Biol. *167*, 889–901.

Chen, J., Zheng, X.F., Brown, E.J., and Schreiber, S.L. (1995). Identification of an 11-kDa FKBP12-rapamycin-binding domain within the 289-kDa FKBP12-rapamycin-associated protein and characterization of a critical serine residue. Proc. Natl. Acad. Sci. USA *92*, 4947–4951.

Churchman, L.S., Okten, Z., Rock, R.S., Dawson, J.F., and Spudich, J.A. (2005). Single molecule high-resolution colocalization of Cy3 and Cy5 attached to macromolecules measures intramolecular distances through time. Proc. Natl. Acad. Sci. USA *102*, 1419–1423.

Churchman, L.S., Flyvbjerg, H., and Spudich, J.A. (2006). A non-Gaussian distribution quantifies distances measured with fluorescence localization techniques. Biophys. J. *90*, 668–671.

Clark, A.G., Dierkes, K., and Paluch, E.K. (2013). Monitoring actin cortex thickness in live cells. Biophys. J. *105*, 570–580.

Croteau, N.J., Furgason, M.L.M., Devos, D., and Munson, M. (2009). Conservation of helical bundle structure between the exocyst subunits. PLoS ONE *4*, e4443.

Dacks, J.B., Poon, P.P., and Field, M.C. (2008). Phylogeny of endocytic components yields insight into the process of nonendosymbiotic organelle evolution. Proc. Natl. Acad. Sci. USA *105*, 588–593.

Dong, G., Hutagalung, A.H., Fu, C., Novick, P., and Reinisch, K.M. (2005). The structures of exocyst subunit Exo70p and the Exo84p C-terminal domains reveal a common motif. Nat. Struct. Mol. Biol. *12*, 1094–1100.

Donovan, K.W., and Bretscher, A. (2015). Tracking individual secretory vesicles during exocytosis reveals an ordered and regulated process. J. Cell Biol. *210*, 181–189.

Dosztányi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics *21*, 3433–3434.

Dubuke, M.L., Maniatis, S., Shaffer, S.A., and Munson, M. (2015). The Exocyst Subunit Sec6 Interacts with Assembled Exocytic SNARE Complexes. J. Biol. Chem.

Erzberger, J.P., Stengel, F., Pellarin, R., Zhang, S., Schaefer, T., Aylett, C.H.S., Cimermančič, P., Boehringer, D., Sali, A., Aebersold, R., and Ban, N. (2014). Molecular architecture of the 40S•eIF1•eIF3 translation initiation complex. Cell *158*, 1123–1135.

Finger, F.P., Hughes, T.E., and Novick, P. (1998). Sec3p is a spatial landmark for polarized secretion in budding yeast. Cell *92*, 559–571.

Fogelgren, B., Lin, S.-Y., Zuo, X., Jaffe, K.M., Park, K.M., Reichert, R.J., Bell, P.D., Burdine, R.D., and Lipschutz, J.H. (2011). The exocyst protein Sec10 interacts with Polycystin-2 and knockdown causes PKD-phenotypes. PLoS Genet. *7*, e1001361.

Fotso, P., Koryakina, Y., Pavliv, O., Tsiomenko, A.B., and Lupashin, V.V. (2005). Cog1p plays a central role in the organization of the yeast conserved oligomeric Golgi complex. J. Biol. Chem. *280*, 27613–27623.

Fromont-Racine, M., Rain, J.C., and Legrain, P. (1997). Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. Nat. Genet. *16*, 277–282.

Gallego, O., Specht, T., Brach, T., Kumar, A., Gavin, A.-C., and Kaksonen, M. (2013). Detection and characterization of protein interactions in vivo by a simple live-cell imaging method. PLoS ONE *8*, e62195.

Gordon, M.P., Ha, T., and Selvin, P.R. (2004). Single-molecule high-resolution imaging with photobleaching. Proc. Natl. Acad. Sci. USA *101*, 6462–6465.

Guo, W., Roth, D., Walch-Solimena, C., and Novick, P. (1999). The exocyst is an effector for Sec4p, targeting secretory vesicles to sites of exocytosis. EMBO J. *18*, 1071–1080.

Ha, J.Y., Chou, H.-T., Ungar, D., Yip, C.K., Walz, T., and Hughson, F.M. (2016). Molecular architecture of the complete COG tethering complex. Nat. Struct. Mol. Biol. *23*, 758–760.

Hamburger, Z.A., Hamburger, A.E., West, A.P., Jr., and Weis, W.I. (2006). Crystal structure of the S.cerevisiae exocyst component Exo70p. J. Mol. Biol. *356*, 9–21.

Hartley, D.A. (1993). Cellular interactions in development: A practical approach (Oxford University Press).

He, B., Xi, F., Zhang, J., TerBush, D., Zhang, X., and Guo, W. (2007). Exo70p mediates the secretion of specific exocytic vesicles at early stages of the cell cycle for polarized cell growth. J. Cell Biol. *176*, 771–777.

Heider, M.R., and Munson, M. (2012). Exorcising the exocyst complex. Traffic *13*, 898–907.

Heider, M.R., Gu, M., Duffy, C.M., Mirza, A.M., Marcotte, L.L., Walls, A.C., Farrall, N., Hakhverdyan, Z., Field, M.C., Rout, M.P., et al. (2016). Subunit connectivity, assembly determinants and architecture of the yeast exocyst complex. Nat. Struct. Mol. Biol. *23*, 59–66.

Hsu, S.C., Ting, A.E., Hazuka, C.D., Davanger, S., Kenny, J.W., Kee, Y., and Scheller, R.H. (1996). The mammalian brain rsec6/8 complex. Neuron *17*, 1209–1219.

Hsu, S.C., Hazuka, C.D., Roth, R., Foletti, D.L., Heuser, J., and Scheller, R.H. (1998). Subunit composition, protein interactions, and structures of the mammalian brain sec6/8 complex and septin filaments. Neuron *20*, 1111–1122.

Huang, B., Babcock, H., and Zhuang, X. (2010). Breaking the diffraction barrier: super-resolution imaging of cells. Cell *143*, 1047–1058.

Janke, C., Magiera, M.M., Rathfelder, N., Taxis, C., Reber, S., Maekawa, H., Moreno-Borchart, A., Doenges, G., Schwob, E., Schiebel, E., and Knop, M. (2004). A versatile toolbox for PCR-based tagging of yeast genes: new

fluorescent proteins, more markers and promoter substitution cassettes. Yeast *21*, 947–962.

Jin, R., Junutula, J.R., Matern, H.T., Ervin, K.E., Scheller, R.H., and Brunger, A.T. (2005). Exo84 and Sec5 are competitive regulatory Sec6/8 effectors to the RalA GTPase. EMBO J. *24*, 2064–2074.

Joglekar, A.P., Bouck, D.C., Molk, J.N., Bloom, K.S., and Salmon, E.D. (2006). Molecular architecture of a kinetochore-microtubule attachment site. Nat. Cell Biol. *8*, 581–585.

Joglekar, A.P., Bloom, K., and Salmon, E.D. (2009). In vivo protein architecture of the eukaryotic kinetochore with nanometer scale accuracy. Curr. Biol. *19*, 694–699.

Kaksonen, M., Toret, C.P., and Drubin, D.G. (2005). A modular design for the clathrin- and actin-mediated endocytosis machinery. Cell *123*, 305–320.

Khmelinskii, A., Meurer, M., Duishoev, N., Delhomme, N., and Knop, M. (2011). Seamless gene tagging by endonuclease-driven homologous recombination. PLoS ONE *6*, e23794.

Kukulski, W., Schorb, M., Kaksonen, M., and Briggs, J.A.G. (2012). Plasma membrane reshaping during endocytosis is revealed by time-resolved electron tomography. Cell *150*, 508–520.

Luo, G., Zhang, J., Luca, F.C., and Guo, W. (2013). Mitotic phosphorylation of Exo84 disrupts exocyst assembly and arrests cell growth. J. Cell Biol. *202*, 97–111.

Maeder, C.I., Hink, M.A., Kinkhabwala, A., Mayr, R., Bastiaens, P.I.H., and Knop, M. (2007). Spatial regulation of Fus3 MAP kinase activity through a reaction-diffusion mechanism in yeast pheromone signalling. Nat. Cell Biol. *9*, 1319–1326.

Matern, H.T., Yeaman, C., Nelson, W.J., and Scheller, R.H. (2001). The Sec6/8 complex in mammalian cells: characterization of mammalian Sec3, subunit interactions, and expression of subunits in polarized cells. Proc. Natl. Acad. Sci. USA *98*, 9648–9653.

McGuffin, L.J., Bryson, K., and Jones, D.T. (2000). The PSIPRED protein structure prediction server. Bioinformatics *16*, 404–405.

Moskalenko, S., Tong, C., Rossé, C., Mirey, G., Formstecher, E., Daviet, L., Camonis, J., and White, M.A. (2003). Ral GTPases regulate exocyst assembly through dual subunit interactions. J. Biol. Chem. *278*, 51743–51748.

Novick, P., Field, C., and Schekman, R. (1980). Identification of 23 complementation groups required for post-translational events in the yeast secretory pathway. Cell *21*, 205–215.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera–a visualization system for exploratory research and analysis. J. Comput. Chem. *25*, 1605–1612.

Picco, A., Mund, M., Ries, J., Nédélec, F., and Kaksonen, M. (2015). Visualizing the functional architecture of the endocytic machinery. eLife *4*, e04535.

Rossé, C., Hatzoglou, A., Parrini, M.C., White, M.A., Chavrier, P., and Camonis, J. (2006). RalB mobilizes the exocyst to drive cell migration. Mol. Cell. Biol. *26*, 727–734.

Roth, D., Guo, W., and Novick, P. (1998). Dominant negative alleles of SEC10 reveal distinct domains involved in secretion and morphogenesis in yeast. Mol. Biol. Cell *9*, 1725–1739.

Russel, D., Lasker, K., Webb, B., Velázquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., Peterson, B., and Sali, A. (2012). Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. PLoS Biol. *10*, e1001244.

Saeed, A.I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M., et al. (2003). TM4: a free, open-source system for microarray data management and analysis. Biotechniques *34*, 374–378.

Saffarian, S., and Kirchhausen, T. (2008). Differential evanescence nanometry: live-cell fluorescence measurements with 10-nm axial resolution on the plasma membrane. Biophys. J. *94*, 2333–2342.

Sbalzarini, I.F., and Koumoutsakos, P. (2005). Feature point tracking and trajectory analysis for video imaging in cell biology. J. Struct. Biol. *151*, 182–195.

Schneidman-Duhovny, D., Pellarin, R., and Sali, A. (2014). Uncertainty in integrative structural modeling. Curr. Opin. Struct. Biol. *28*, 96–104.

Sheather, S.J., and Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. J. R. Stat. Soc. *53*, 683–690.

Shen, M.-Y., Davis, F.P., and Sali, A. (2005). The optimal size of a globular protein domain: A simple sphere-packing model. Chem. Phys. Lett. *405*, 224–228.

Sivaram, M.V.S., Furgason, M.L.M., Brewer, D.N., and Munson, M. (2006). The structure of the exocyst subunit Sec6p defines a conserved architecture with diverse roles. Nat. Struct. Mol. Biol. *13*, 555–556.

Sjöblom, T., Jones, S., Wood, L.D., Parsons, D.W., Lin, J., Barber, T.D., Mandelker, D., Leary, R.J., Ptak, J., Silliman, N., et al. (2006). The consensus coding sequences of human breast and colorectal cancers. Science *314*, 268–274.

Skruzny, M., Brach, T., Ciuffa, R., Rybina, S., Wachsmuth, M., and Kaksonen, M. (2012). Molecular basis for coupling the plasma membrane to the actin cytoskeleton during clathrin-mediated endocytosis. Proc. Natl. Acad. Sci. USA *109*, E2533–E2542.

Söding, J. (2005). Protein homology detection by HMM-HMM comparison. Bioinformatics *21*, 951–960.

Söding, J., Biegert, A., and Lupas, A.N. (2005). The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. *33*, W244–W248.

Szymborska, A., de Marco, A., Daigle, N., Cordes, V.C., Briggs, J.A.G., and Ellenberg, J. (2013). Nuclear pore scaffold structure analyzed by super-resolution microscopy and particle averaging. Science *341*, 655–658.

Teodoro, R.O., Pekkurnaz, G., Nasser, A., Higashi-Kovtun, M.E., Balakireva, M., McLachlan, I.G., Camonis, J., and Schwarz, T.L. (2013). Ral mediates activity-dependent growth of postsynaptic membranes via recruitment of the exocyst. EMBO J. *32*, 2039–2055.

TerBush, D.R., Maurice, T., Roth, D., and Novick, P. (1996). The Exocyst is a multiprotein complex required for exocytosis in Saccharomyces cerevisiae. EMBO J. *15*, 6483–6494.

Tong, A.H.Y., and Boone, C. (2006). Synthetic genetic array analysis in Saccharomyces cerevisiae. Methods Mol. Biol. *313*, 171–192.

Walworth, N.C., and Novick, P.J. (1987). Purification and characterization of constitutive secretory vesicles from yeast. J. Cell Biol. *105*, 163–174.

Wan, X., O'Quinn, R.P., Pierce, H.L., Joglekar, A.P., Gall, W.E., DeLuca, J.G., Carroll, C.W., Liu, S.-T., Yen, T.J., McEwen, B.F., et al. (2009). Protein architecture of the human kinetochore microtubule attachment site. Cell *137*, 672–684.

Wang, J., Bello, O., Auclair, S.M., Wang, J., Coleman, J., Pincet, F., Krishnakumar, S.S., Sindelar, C.V., and Rothman, J.E. (2014). Calcium sensitive ring-like oligomers formed by synaptotagmin. Proc. Natl. Acad. Sci. USA *111*, 13966–13971.

Wiederkehr, A., De Craene, J.-O., Ferro-Novick, S., and Novick, P. (2004). Functional specialization within a vesicle tethering complex: bypass of a subset of exocyst deletion mutants by Sec1p or Sec4p. J. Cell Biol. *167*, 875–887.

Wu, S., Mehta, S.Q., Pichaud, F., Bellen, H.J., and Quiocho, F.A. (2005). Sec15 interacts with Rab11 via a novel domain and affects Rab11 localization in vivo. Nat. Struct. Mol. Biol. *12*, 879–885.

Yamamoto, A., Kasamatsu, A., Ishige, S., Koike, K., Saito, K., Kouzu, Y., Koike, H., Sakamoto, Y., Ogawara, K., Shiiba, M., et al. (2013). Exocyst complex component Sec8: a presumed component in the progression of human oral squamous-cell carcinoma by secretion of matrix metalloproteinases. J. Cancer Res. Clin. Oncol. *139*, 533–542.

Yamashita, M., Kurokawa, K., Sato, Y., Yamagata, A., Mimura, H., Yoshikawa, A., Sato, K., Nakano, A., and Fukai, S. (2010). Structural basis for the Rho- and phosphoinositide-dependent localization of the exocyst subunit Sec3. Nat. Struct. Mol. Biol. *17*, 180–186.

Yildiz, A., Forkey, J.N., McKinney, S.A., Ha, T., Goldman, Y.E., and Selvin, P.R. (2003). Myosin V walks hand-over-hand: single fluorophore imaging with 1.5-nm localization. Science *300*, 2061–2065.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, Peptides, and Recombinant Proteins		
5-Fluoroorotic Acid Monohydrate	Formedium	CAS: 5FOA05
Adenine hemisulfate salt	Sigma-Aldrich	Cat#A9126; CAS:321-30-2
Ammonium Sulfate	EMSURE	Cat#101217; CAS: 7783-20-2
Bacteriological Peptone	CONDA Pronadisa	Cat#1616
BglII restriction enzyme	NEB	Cat#R0143S
cloNAT	Werner BioAgents	Cat#5001000
D(+)-Glucose Anhydrous	Formedium	Cat#GLU03
dNTP Mix	Thermo Fisher Scientific	Cat#R0193
Bacteriological Agar	CONDA Pronadisa	Cat#1800
Geneticin Selective Antibiotic (G418 Sulfate)	Thermo Fisher Scientific	Cat#11811031
Glycine	Sigma-Aldrich	Cat#50046; CAS:56-40-6
Hygromycin B	Invivogen	Cat#ant-hm-1; CAS:31282-04-9
Latrunculin A	Enzo Life Sciences	Cat#T119-0500
L-4-Thialysine hydrochloride	Sigma-Aldrich	Cat#A2636; CAS:4099-35-8
L-Alanine	Sigma-Aldrich	Cat#05129; CAS:56-41-7
L-Arginine	Sigma-Aldrich	Cat#A8094; CAS:74-79-3
L-Aspartic Acid	Sigma-Aldrich	Cat#A8949; CAS:56-84-8
L-Asparagine	Sigma-Aldrich	Cat#A0884; CAS:70-47-3
L-Canavanine sulfate salt	Sigma-Aldrich	Cat#C9758; CAS:2219-31-0
L-Cysteine	Sigma-Aldrich	Cat#168149; CAS:52-90-4
L-Glutamic Acid Monosodium Salt Monohydrate	Sigma-Aldrich	Cat#49621; CAS:6106-04-3
L-Glutamine	Sigma-Aldrich	Cat#49419; CAS:56-85-9
L-Histidine	Sigma-Aldrich	Cat#H8000; CAS:71-00-1
L-Isoleucine	Sigma-Aldrich	Cat#I2752; CAS:73-32-5
L-Leucine	Sigma-Aldrich	Cat#L8000; CAS:61-90-5
L-Lysine	Sigma-Aldrich	Cat#L5501; CAS: 56-87-1
L-Methionine	Sigma-Aldrich	Cat#M9625; CAS:63-68-3
L-Phenylalanine	Sigma-Aldrich	Cat#78019; CAS:63-91-2
L-Proline	Sigma-Aldrich	Cat#P5607; CAS:147-85-3
L-Serine	Sigma-Aldrich	Cat#84959; CAS:56-45-1
L-Threonine	Sigma-Aldrich	Cat#T8625; CAS: 72-19-5
L-Tyrosine	Sigma-Aldrich	Cat#T8566; CAS: 60-18-4
L-Valine	Sigma-Aldrich	Cat#V0500; CAS: 72-18-4
myo-Inositol	Sigma-Aldrich	Cat#I5125; CAS: 87-89-8
para-Aminobenzoic Acid	Sigma-Aldrich	Cat#A9878; CAS: 150-13-0
Potassium acetate	Sigma-Aldrich	Cat#60035; CAS:127-08-2
Rapamycin	Sigma-Aldrich	Cat#R8781
SalI restriction enzyme	Thermo Fisher Scientific	Cat#ER0641
ssDNA	Sigma-Aldrich	Cat#D1626
T4 DNA Ligase	NEB	Cat#M0202M
Uracil	Sigma-Aldrich	Cat#U0750; CAS:66-22-8
Velocity DNA Polymerase	Bioline	Cat#BIO-21098
Yeast Extract	CONDA Pronadisa	Cat#1702

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Yeast Nitrogen Base – Low Florescence, without Amino acids, Folic Acid and Riboflavin.	Formedium	Cat#CYN6501
Yeast Nitrogen Base without Amino Acids	Formedium	Cat#CYN0402
Yeast Nitrogen Base without Amino Acids and Ammonium Sulfate	Formedium	Cat#CYN0502
Yeast Synthetic Drop-out Trp-,Ura-,Leu-His-	Formedium	Cat#DSCK1027
Yeast Synthetic Drop-out Trp-,Ura-	Formedium	Cat#DSCK1019
Experimental Models: Organisms/Strains		
S. cerevisiae: BY4741	Invitrogen	Cat#95702
S. cerevisiae: BY4742	Invitrogen	Cat#95702
S. cerevisiae: MKY2128	Gallego et al., 2013	N/A
S. cerevisiae: DDY1102	Kaksonen et al., 2005	N/A
A full list of strains is presented in Table S1	N/A	N/A
Software and Algorithms		
Bash	Free software foundation	ftp://ftp.gnu.org/pub/bash/
HHSearch	Söding, 2005	https://toolkit.tuebingen.mpg.de/hhpred
ImageJ 1.51a	ImageJ developers	http://imagej.net/ImageJ
Particle Tracker v1.5	Sbalzarini and Koumoutsakos, 2005	http://mosaic.mpi-cbg.de/ParticleTracker/
Integrative Modeling Platform	Russel et al., 2012	https://integrativemodeling.org/
MATLAB (R2008a and R2015a) and its Image Analysis suite	MathWorks	https://www.mathworks.com
Metamorph 7.5.5.0	Molecular Devices	https://www.moleculardevices.com/systems/metamorph-research-imaging
Multiexperiment Viewer	Saeed et al., 2003	https://sourceforge.net/p/mev-tm4/discussion/
Perl 5	Larry Wall	https://www.perl.org/
Psi-pred	McGuffin et al., 2000	http://bioinf.cs.ucl.ac.uk/psipred/
Python 2.7.9	Python Software Foundation	https://www.python.org/ [2.7.9]
R 2.9, 2.10 and 3.3.2	R Core Team	https://www.r-project.org/
UCSF Chimera 1.11	Pettersen et al., 2004	https://www.cgl.ucsf.edu/chimera/
IUPred	Dosztányi et al., 2005	http://iupred.enzim.hu/
MATLAB scripts that correct for chromatic aberration; R scripts that compute distances between distinct fluorophores; Python and R scripts that quantify exocyst subunit abundances.	This paper	https://github.com/apicco/exocyst_scripts
Python scripts used to generate the 3D models and files representing exocyst and COG solutions.	This paper	https://github.com/batxes/exocyst_scripts

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for reagents may be directed to the Lead Contact Oriol Gallego (oriol.gallego@irbbarcelona.org).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

All strains used in this study were derivatives of *S. cerevisiae* BY4741/42 (Invitrogen), MKY2128 (Gallego et al., 2013) or DDY1102 (Kaksonen et al., 2005). Construction of strains is described below and a complete list of the strains used in this study is given in Table S1.

METHOD DETAILS

Yeast strains and plasmids

Plasmid pMK0085 coding for GAL1pr-I-SceI was synthesized from pND32 (Khmelinskii et al., 2011) and pFA6a-EGFP-His3MX6 (Janke et al., 2004). Briefly, GAL1pr-I-SceI cassette from pND32 was inserted between the SalI and the BglII sites to replace the eGFP cassette in pFA6a-EGFP-His3MX6. The coding sequence of the *Saccharomyces* cerevisiae Exo84 (aa. 1-753); Sec6 (aa. 1-805) and Sec10 (aa. 1-490) were PCR-amplified and cloned in frame with the Gal4 DNA-binding domain (DBD) into pB43 vector as Bait -Gal4 fusion. The coding sequence of the *Saccharomyces* cerevisiae Sec3 (aa. 601-1336) and Sec10 (aa. 491-871) were PCR-amplified and cloned in frame with the Gal4 DBD into pB66 vector as Gal4-Bait fusion. pB66 and pB43 derive from the original pAS2ΔΔ vector (Fromont-Racine et al., 1997). The empty pB66 and pB43 bait plasmids were used in the control assay. The coding sequence of the *Saccharomyces* cerevisiae Sec15 (aa. 382-910) and Sec8 (aa. 474-1065) were cloned in frame with the Gal4 Activation Domain (AD) into pP7 vector as AD-Prey fusion. The coding sequence of the *Saccharomyces* cerevisiae Exo70 (aa. 1-623) and Sec15 (aa. 1-381) were cloned in frame with the Gal4 Activation Domain (AD) into plasmid pP13 as Prey -AD fusion. The pP7 and pP13 prey plasmids, used in the control assay, are derived from the pP6 plasmid. pP6 derives from the original pGADGH vector (Hartley, 1993). The inserts of the yeast two-hybrid constructs were sequenced. Yeast strains for rapamycin-induced translocation were generated as described in Gallego et al. (Gallego et al., 2013). Yeast genes were tagged or deleted at their genomic loci by PCR-based gene targeting (Janke et al., 2004).

We engineered yeasts cells expressing a C-terminal RFP-FKBP fusion of Sla2 in the strain BY4741 (MKY2436) and in the strain MKY2128 (MKY2129). We use these strains to construct different combinations of bait-FRB and prey-GFP. Overall we generated 174 different strains, which are listed in the Table S1. We followed different strategies according to the characteristics of the tag: We constructed C-terminal GFP fusions of the exocyst subunits, as well as Sec2, with 3xmyeGFP in the parental strain MKY2128 (MKY2548 to MKY2556). The FRB-tagging of the eight subunits of the exocyst (bait-FRB) were performed in the parental strain MKY2436 (MKY2437 to MKY2444). We constructed C-terminal GFP fusions for all the COG subunits, except COG1, with 3xmyeGFP in the parental strain BY4741 (OGY0258 to OGY0264). The FRB-tagging of the eight subunits of the COG (bait-FRB) were performed in the parental strain MKY2129 (OGY0197 to OGY0204). Finally, we used the SGA technology (Tong and Boone, 2006) to mate the strains coding for the bait-FRB and prey-GFP for the exocyst subunits (MKYSGA0009 to MKYSGA0061) and for the COG subunits (OGYSGA4835 to OGYSGA4890).

For the N-terminal GFP fusions, we followed the Endonuclease-driven approach for seamless gene tagging by homologous recombination described in Khmelinskii et al. (Khmelinskii et al., 2011). We introduced the GAL1pr-I-SceI cassette into the *Leu2-3* locus of the BY4742 strain (MKY2558) or the MKY2436 strain (OGY0607) using the forward oligonucleotide mk1171 (TCAAA AAGATCCATGTATAATCTTCATTATTACAGCCCTCTTGACTTATTTCAGGAAAGTTTCGGAGGAG) and the reverse oligonucleotide mk1172 (GTTTCGTCTACCCTATGAACATATTCCATTTTGTAATTTCGTGTCGATCGATGAATTCGAGCTCG). Positive colonies were selected on SD-HIS plates and confirmed by colony-PCR. Using PCR-based gene targeting we tagged N-terminally Sec3, Sec5, Sec6, Sec10 and Sec15 with sfGFP in the MKY2558 parental strain (MKY2660 to MKY2662, MKY2664 and MKY2665) and Cog2, Cog3, Cog4, Cog5, Cog6 and Cog7 in the OGY0607 parental strain (OGY0646 to OGY0651). For the exocyst subunits, strains carrying the bait-FRB (MKY2437 to MKY2444) were initially mated with MKY2557 strain using SGA technology (MKYSGA0062 to MKYSGA0069). The resulting strains were compatible with subsequent automated mating with the strains harboring prey-GFP (MKY2660 to MKY2662, MKY2664 and MKY2665). For the COG subunits, we first swapped the *klURA* cassette in strains expressing the bait-FRB (OGY0197 to OGY0204) by the *kanMX4* cassette (OGY0637 to OGY0644). The resulting strains were compatible with subsequent automated mating with the strains harboring prey-GFP (OGY0646 to OGY0651). Seamless marker excision was confirmed by PCR after cells were sequentially grown in galactose media for 16h and SD plates containing 5-FOA for two days, both for the exocyst subunits (MKYSGA0070 to MKYSGA0098) and for the COG subunits (OGYSGA6274 to OGYSGA6329). We used PCR-based gene targeting (Janke et al., 2004) to successfully tag Exo70 and Exo84 N-terminally to sfGFP, which initially failed with the automated approach. First, the GAL1pr-I-SceI cassette was inserted into the *Leu2-3* locus of the strains MKYSGA0062 to MKYSGA0069 (MKY2895 to MKY2901). Positive colonies were selected on SD-HIS plates and confirmed by colony-PCR. Then, Exo70 and Exo84 sfGFP N-terminal tag was introduced via Endonuclease-driven approach for seamless gene tagging by homologous recombination. Seamless marker excision was confirmed by PCR after cells were sequentially grown in galactose media for 16h and SD plates containing 5-FOA for two days (MKY2923 to MKY2927 for Exo70 and MKY2933, MKY2934, MKY2968, MKY2969 and MKY2971 for Exo84). We were not able to generate yeast strains harboring Sec8 or Cog8 with an sfGFP N-terminal tag. Strains in which Sec5-FRB was combined with N- or C-terminal prey-GFPs could not be generated.

For C-terminal FRB-GFP fusions we tagged exocyst subunits in the parental strain MKY2128 (MKY2540 to MKY2547) using PCR-based gene targeting (Janke et al., 2004). Then we used SGA technology to cross these strains with MKY2436 and incorporate the anchor Sla2-RFP-FKBP (MKYSGA0001 and MKYSGA0003 to MKYSGA0008).

Of the 174 *S. cerevisiae* strains that we generated, we succeeded in measuring the separation between the RFP and GFP tag positions for 168 strains (i.e., 82 strains where COG subunits were used as prey-GFP, 80 strains where exocyst subunits were used as prey-GFP and 6 more strains where Sec2 was tagged with GFP at the C terminus). These data are listed in the Table S2. See Tables S1 and S5 for a complete list of the strains and plasmids used in this study.

Microscope setup

We used an Olympus IX81 microscope equipped with 100x/1.45 objective lens and Hamamatsu Orca-ER camera. Cells were excited with a X-Cite 120Q lamp (Exelitas Technologies) at 100% intensity. Excitation light was reflected with an FF493/574-di01-25x36 (Semrock) dual-edge dichroic beamsplitter. Emission light was filtered with Semrock FF01-520/35-25 BrightLine and with Semrock FF01-624/40-25 BrightLine filters mounted on a filter wheel. All microscope hardware was controlled by Metamorph (Universal Imaging).

Imaging

Exocyst complex recruitment was performed following the principle described for PICT method by Gallego et al. (Gallego et al., 2013). Strains were grown in synthetic defined (SD) medium with appropriate supplements at 25°C overnight and diluted and grown next morning up to exponential phase. Cells were attached to a 35 mm coverslip coated with Concanavalin A and were treated either with vehicle (DMSO) or 10 μM rapamycin (Sigma). After 10 min incubation at room temperature, 200 μM LatA was added to depolymerize actin filaments. All images were acquired within an interval of time between 10 min and 20 min upon LatA addition (20 to 30 min rapamycin treatment). Images of the middle section of cells were acquired with 2.5-3 s of exposure time. The field of view was imaged in the RFP and GFP channels and then it was moved to a different region of the sample, not affected by photobleaching. The preparation and imaging of each strain was duplicated; for each sample we acquired ~10 different fields of view. All acquisitions were done in duplicate for a total of ~20 different fields of view.

To assess the reproducibility of our approach, we repeated on different days the acquisition and the image analysis of a strain harboring Exo70-FRB and Sec5 tagged at the C terminus (MKYSGA0048) or at the N terminus (MKYSGA0090) with GFP. No significant difference was observed between the measurements (Figure S2D).

Image processing and distance measurements

See Image processing and distance measurements in STAR Methods – Quantification and statistical analysis.

Exocyst structure determination

The model of the exocyst architecture was determined with the Integrative Modeling Platform (Russel et al., 2012). The measurements of the separation between fluorophores in the 2D images implicitly contain the information to determine the relative position of the fluorescent tags in the 3D space. First, we used the measured dataset of distances to trilaterate the relative position of each of the fluorescent tags in the 3D space, where the position of each fluorophore depends on the position of all the other fluorophores. Second, we used the position of the fluorescent tags to locate in the 3D space the subunits they were fused to. The interaction between Sec6 and Sec8 was also used as an additional restraint (Guo et al., 1999; Sivaram et al., 2006).

IMP is a suite of programs that integrates information from diverse experiments to determine the structure of macromolecular complexes. IMP assigns a score that quantifies the fulfillment of the restraints (Russel et al., 2012). Each step was divided into four stages: (1) Gathering of data, (2) Representation of tags or subunits and translation of the data into spatial restraints, (3) Optimization and sampling of the space of solutions, and (4) Analysis and assessment of the ensemble of models.

Step I) Localization of the tags

We used the measured distances as restraints to determine the relative spatial location of the fluorescent tags.

1) **Gathering of data.** Distance data (Figure 2A and Table S2) were collected as detailed in STAR Methods—Image Processing and Distance Measurements.

2) **Representation of the tags and translation of data into spatial restraints.** Each fluorescent tag (RFP for the anchor, GFP N-terminal or GFP C-terminal for the exocyst subunits) was represented by a bead (Figure 3A). We encoded all measured distances between the RFP and each GFP as constraints in IMP (Table S2).

3) **Optimization and sampling of the space of solutions.** We used 500 steps of conjugate gradients optimization to determine the position of each of the fluorescent tags, imposing the measured distances as a restraint (Figure 3A and Figure S3). To sample the space of solutions we repeated this procedure 10,000 times with IMP, starting each time from different random initial positions for all tags.

4) **Analysis.** The 200 solutions with the best IMP score fulfilling all restraints were selected and clustered, with Hierarchical Clustering and K-Means clustering, according to their similarity that was measured by their Root Mean Square Deviation (RMSD) using the Multiexperiment Viewer, MeV (Saeed et al., 2003). All models converged into two clusters that are the mirror images of each other (Figure S3). As we could not assess the chirality of the exocyst complex, we randomly selected one of the two populations of tag positions for the subsequent step (Figure 3A) and referred to it as "fluorophore positions."

Step II) Determination of the exocyst 3D architecture

We then used the positions of the fluorescent tags in the 3D space determined previously (Step I) as a scaffold to locate the exocyst subunits. The IMP procedure was similar:

1) **Gathering of data.** To gather structural information from each exocyst subunit we used IUPred (Dosztányi et al., 2005) for disorder prediction, Psi-pred (McGuffin et al., 2000) for secondary structure prediction and HHSearch (Söding, 2005) for comparative modeling template detection (Figure 3B).

One of the solutions of the fluorophore positions was randomly chosen as one of the restrains to locate the subunits. The SD of the positions of the tags in all the solutions computed for the fluorophore positions was associated to each tag position, determined as follows:

$$\sigma = \sqrt{\sigma_x^2 + \sigma_y^2 + \sigma_z^2},$$

where σ_x,σ_x and σ_z are the SDs of the positions of the tags computed along the axis x, y and z respectively.

2) **Representation of subunits and translation of data into spatial restraints.** Given the uncertainty of the modeling process, we focused on giving more importance to the distances gathered through the light microscopy and their variability than to the representation of the subunits (Schneidman-Duhovny et al., 2014). Given the lack of high-resolution structures for the exocyst subunits (less than 26% of the residues have their atomic structure determined experimentally), each subunit was represented by a flexible string of connected beads. The representation is based on a PDB survey where the mean volume of a fragment of 120 residues is a sphere of 3.5 nm of diameter (Shen et al., 2005). The numbers of beads were determined from the structural features of the proteins, including length, unstructured fragments, secondary structure, and tertiary structure (see "Gathering of data"; Figure 3B). We imposed consecutive beads to be connected. For this reason we used a maximum separation restraint between consecutive beads representing each protein, which also varied depending on the structural features of each subunit (see Table S3). Excluded volume restraints were used for all the beads composing the proteins (i.e., two beads can not occupy the same volume being from the same or from different subunits) (Figure 3C). Posteriori validation of the representation is provided by the atomic structure of Exo70 that fits nicely in its corresponding shape in our model (Figure 5A). We empirically varied the size and number of the beads to ensure that our representation is not biasing the results, but the impact on the final reconstruction was negligible (data not shown).

The fluorophore positions determined in Step I ("Localization of the tags") were used to locate the subunits of the complex. To locate the N termini of exocyst subunits we randomly chose one of the solutions among the fluorophore positions (see step I "Localization of the tags"). For each subunit, the bead representing the N terminus was constrained in a sphere centered on the position of the N-terminal GFP in this particular solution. The radius of the sphere was the 95% confidence interval computed from the distribution of the positions for the N-terminal GFP (Figure 3C and Table S3). Since Sec8 could not be tagged with GFP at the N terminus, the direct Sec6-Sec8 protein-protein interaction (Guo et al., 1999; Sivaram et al., 2006) was used as additional constraint to approximately locate the N terminus of the Sec8 subunit within the exocyst architecture.

Exocyst subunits, when used as bait-FRB, were tagged at their C terminus with FRB to be recruited to the anchor. As a result of the recruitment, the RFP tag of the anchor also flagged the C terminus of these subunits. Therefore, to locate the C terminus of exocyst subunits, we used both the position of the C-terminal GFP tag and the position of the RFP that flagged the anchor. The bead representing the C terminus of each subunit was thus constrained at the intersection between two spheres: a sphere centered on the position of the C-terminal GFP, and a sphere centered on the position of the anchor (RFP). The radii for the spheres were half the separation between both GFP C terminus and RFP tags plus the 95% confidence interval computed from the distribution of the fluorophore positions (STAR Methods, Figure 3C, and Table S3). Sec5 could not be tagged with FRB, therefore the Sec5 C terminus location was restrained only inside a sphere centered on Sec5 C-terminal GFP. The radius of the sphere was the 95% confidence interval computed from the distribution of the positions for the C-terminal GFP. All constraints used are summarized in the Table S3.

3) **Optimization and sampling of the space of solutions.** 500 steps of conjugate gradients optimization were used to compute the location of each of the subunits, imposing the tag positions as a restraint (Figure 3D). To sample the space of solutions exhaustively, we computed 50,000 independent optimizations starting from random initial positions for the beads representing the subunits.

4) **Analysis**. The 200 best scoring solutions that satisfied all the input restraints were selected and superimposed using Chimera (Pettersen et al., 2004). To assess the variability in the bead location among the different solutions we measured the SD of the positions as follows:

$$\sigma = \sqrt{\sigma_x^2 + \sigma_y^2 + \sigma_z^2}.$$

Except for the N terminus of Sec8, which we could not tag, all termini are similarly located in the different solutions (Figure S3C). As expected, we detected variability only for the middle region of the proteins, in particular Sec3 and Sec5. The SD of the localization of each bead is shown in Figure S3C.

The population of models was hierarchically clustered, using the RMSD of Exo70 and then Sec3. These subunits were chosen because they showed more than 1 conformation in the superposition of all the models. We obtained 6 sub-populations of models using K-means clustering (Figure S4). To illustrate the structure of the exocyst in the main figures we selected the model with the best IMP score, which is part of the largest cluster of Exo70 solutions.

Controls

As a first control of the computational reconstruction, we repeated the procedure with the same dataset modified by randomly assigning all measured distances to different FRB and GFP pairs. No solution that satisfied all the restraints could be isolated in this case.

As a second control, we repeated the procedure with a dataset where we swapped the distance between two tags that locate at distant sides of the complex (Exo70-FRB with Sec5-GFP) with the distance between two tags that locate close to each other (Sec10-FRB with GFP-Exo84). In this case, we were able to identify solutions that satisfied all restraints, but they clustered in more than two clusters (Figure S3D). This indicates that many configurations of fluorophores positions are compatible with the data, and thus that the switching of only two distances leads to data that are not coherent. In addition, the solutions obtained are different from the location of the fluorescent tags determined with the experimental distance measurements. The RMSDs between the models obtained from this simulated artifact and the experimentally reconstructed location of the fluorophores were always above 15 nm. This is in the upper limit of RMSD between models belonging to different mirror images populations in the original models (Figure S3E).

To measure the robustness of our method, we repeated our procedure various times, shortening the distance measured for Sec3 tagged with GFP at the N terminus when Sec15-FRB was used as bait (23.9 nm) by 2 nm in each iteration. When the distance was reduced by values between 6 (25%) and 16 (67%) nm, the configurations of the fluorescent tags were different from the ones derived with the experimentally measured distances. When the distance was reduced by 18 nm (75%), we were not able to get any model that would fulfill the simulated set of restraints.

Overall this indicates that the reconstruction of protein complexes is sensitive enough to respond to small variations in the measured distances (i.e., the resulting molecular architecture reflects this small differences in the measurements). However, when these variations come from an artifact in the measurements, either the reconstruction does not output any solution that can fulfill all the distance measurements (i.e., when the 80 distances were assigned randomly to different tags pairs or even when a single distance was shortened by more than 16 nm), or the solutions for the location of the fluorescent tags do not converge, reflecting the lack of coherence among the dataset (i.e., after swapping two distances in a dataset of 80 distances or when the shortening of a distance is not more than 16 nm).

COG complex structure determination

We used the same procedure described for the exocyst to determine the molecular architecture of the Cog2, Cog3, Cog4, Cog5, Cog6, Cog7 and Cog8 within the COG complex. Since we could not tag the N terminus of Cog8, in addition of the 82 distance measurements (Table S2) we also used the known interaction between the subunits Cog6 and Cog8 to locate the Cog8 N terminus (Fotso et al., 2005). After repeating the trilateration for each fluorescent tag 10.000 times, we selected the 200 models that had best IMP score and that fulfilled all 82 restraints. These models were clustered based on their RMSD in two populations that were mirror images of each other. One of these clusters was selected and used in the next step to determine the 3D architecture of these 7 subunits within the COG complex. The subunits were represented by strings of beads of 3.5 nm in diameter, taking into account the structural features of each COG subunit. To follow a procedure as close as possible to the one used for the exocyst, the radii of this spheres used to position the beads representing the N and C termini of the subunits were set to be the mean of the radii used to position the N and C termini of the exocyst subunits. Similarly to the exocyst, the bead representing the C terminus of each COG subunit was thus constrained on the intersection between two spheres: a sphere centered on the position of the C-terminal GFP, and a sphere centered on the position of the anchor (RFP). We also imposed an excluded volume restraint, a maximum distance between consecutive beads of the same subunit. All the restraints are listed in Table S3. We computed the reconstruction of the molecular architecture of the 7 COG subunits 50,000 times as for the exocyst. We selected the solution with best IMP score as the representative reconstruction of the 7 subunits of the COG complex (Figure S5).

Structurally related fragments, comparative modeling and volume

The HHpred interactive server for protein homology detection and structure prediction (Söding et al., 2005) was used to identify structurally related fragments. The structure of the residues 60 to 623 of Exo70 has been determined (PDB entry 2b1e_A). Sec3 residues 741 to 1332 were predicted to be similar to this atomic structure with a p value of 2.3E-04. Likewise, the atomic structure for the C-terminal part of Sec6 has been determined (from residues 407 to 805; PDB entry 2fji_1) and the same structure is predicted to represent residues 590 to 997 of Sec8 with a p value of 2.3E-06. Thus, the pair Exo70-Sec6 is structurally related to the pair Sec3-Sec8, at least in fragments of the proteins.

Similarly, the three-dimensional structure of residues 432 to 650 of Sec9, 28 to 264 of Sso1 and 22 to 111 of Snc2 were modeled based on their homology to known structures (PDB entry 3j96_M, PDB entry 4jeh_B and PDB entry 3hd7_A with p values of 2.5E-37, 7.8E-44 and 5.4E-35 respectively).

Chimera (Pettersen et al., 2004) was used to generate a density map of the exocyst architecture solution with best IMP score and to measure its volume.

Calculation of Sec2 position

To locate the position of the C terminus of Sec2 in respect to the exocyst complex, we measured the distance between the recruited Sec2, fused to GFP at the C terminus (prey-GFP), and the Sla2-RFP-FKBP for different bait-FRB (Figure 2B and Table S2). These distances were added as restraints in IMP together with the distances used to compute the fluorophores positions. Using the exact same approach described previously we could thus resolve the average localization of Sec2-GFP. We iterated the optimization 50,000 times starting with different initial positions for the beads of the subunits and using one of the solutions of the fluorophore positions randomly chosen. The fluorophore positions also included the Sec2-GFP fluorescent tag position. We selected the 200

best solutions that fulfilled all the restraints. All of the selected solutions showed the same spatial localization for the C terminus of Sec2 while the position of the exocyst subunits remained unaltered. For simplicity, Sec2 was represented as a single bead marking the average position of its C terminus.

Quantification of exocyst subunit abundances, sample preparation

GFP-Sec5, GFP-Sec6, GFP-Sec10, GFP-Sec15, GFP-Exo84 and Nuf2-GFP cells were tagged with the same sfGFP. Cells were grown overnight on SC-Trp at 25°C till OD_{600} = 0.6. Sec5, Sec6, Sec10, Sec15 or Sec10 cells were mixed together with Nuf2 cells, of the same mating type, with a 1:1 ratio and adhered on a glass coverslip coated with Concanavalin A, with a 10 min incubation at room temperature. Cells were imaged with an Olympus IX81, equipped with a 100x/1.45 NA objective, a FF493/574-di01-25x36 Brightline dual band dichroic (Semrock), an FF01-520/35 Brightline emission filter and a Hamamatsu Orca-ER CCD camera. Cells were excited with a X-Cite 120Q lamp (Exelitas Technologies) at 100% of power for 500 ms and imaged with a z stack of 21 frames separated by 200 nm (Figure S1B).

FRAP

The FRAP experiments were done on a custom-built set-up that focuses a 488 nm laser beam in a 0.5 μm spot on the sample plane. The exocyst complex was recruited to the anchoring platform using Exo70-FRB as bait in cells treated with 10 μM rapamycin. After 10 min incubation at room temperature, 200 μM LatA was added to depolymerize actin filaments. The FRAP experiment were done within an interval of time between 10 min and 20 min upon LatA addition. Images were recorded every 12 s, and exposure time ranged from 900 to 1200 ms, depending on the prey-GFP that was imaged. Before bleaching, we imaged five frames to estimate the initial fluorescence. After bleaching, we followed the fluorescence recovery for 39 frames. Images were background subtracted and all movies were corrected for photobleaching using ImageJ software (https://imagej.nih.gov/ij/). The fluorescence recovery was calculated within a circle of 4 pixels centered on the anchoring platform that was bleached. The average recovery curve was calculated from at least 4 independent experiments aligned to the bleaching time.

FCCS

Fluorescence cross-correlation spectroscopy data were recorded on a Leica SP2 confocal microscope equipped with single-photon counting avalanche photodiodes and a 63x water objective. GFP was excited using a 488 nm argon laser, mCherry was excited by a 561 diode laser. The emitted light was separated by a dichroic mirror (LP560) and then passed into two different detection channels using the filters BP500-550 (GFP) and HQ638DF75 (mCherry). Auto- and crosscorrelation curves, the number of particles in the respective strains, and the K_D of the interaction were calculated as described previously (Boeke et al., 2014; Maeder et al., 2007).

Yeast two-hybrid

Yeast two-hybrid was performed by Hybrigenics, S.A., Paris, France (https://www.hybrigenics-services.com). Bait and prey constructs were transformed in the yeast haploid cells L40deltaGAL4 (MATa) and YHGX13 (Y187 ade2-101::loxP-kanMX-loxP, MATα), respectively. The diploid yeast cells were obtained by mating. These assays are based on the HIS3 reporter gene (growth assay without histidine). Due to the auto-activation properties of Sec3 C terminus (aa 601-1336) we could not assess its interaction with Sec8 C terminus (aa. 474-1065).

QUANTIFICATION AND STATISTICAL ANALYSIS

Image processing and distance measurements

Images were background subtracted and corrected for the uneven cytoplasmic signal by subtracting from the image the median filtered version of the image itself computed with a kernel of 10 pixels. We processed the images to determine the distance between the centroid positions of the RFP tagged to the anchor and of the GFP tagged to exocyst subunits (Figures 2 and Figure S2). For each of the 80 strains we repeated the measurements in 60 to 290 anchoring platforms, distributed in different cells. We only considered RFP and GFP spot pairs resulting from the recruitment of the exocyst to anchoring platforms induced by rapamycin. We discarded regions where individual spot pairs could not be optimally resolved such as the cell neck or cells where the focal plane did not image the middle section of the cell. The image processing was done as follows:

1. Spot detection: Only pairs of spots visible on both the red channel (named W1) and the green channel (named W2) were selected. Spots were detected using Particle Tracker plugin in ImageJ (Sbalzarini and Koumoutsakos, 2005), which tracks the centroid position of each spot pair in the two channels: one centroid for W1 and one centroid for W2.

2. Chromatic aberration correction: Centroid positions were corrected for chromatic aberration using a warping transformation which was computed from images of Tetraspeck beads emitting in both the RFP and GFP channels. The beads' centroids were tracked with ParticleTracker in ImageJ and the warping transformation was computed using a custom written software in MATLAB (Image Processing toolbox; Mathworks).

3. Spot selection: For each spot, we computed the second momentum of brightness and eccentricity. The second momentum of brightness was computed as described by Sbalzarini and Koumoutsakos (Sbalzarini and Koumoutsakos, 2005). The eccentricity e of a spot S was computed as

$$e = \frac{\mu_{20} + \mu_{02} + \sqrt{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2}}{\mu_{20} + \mu_{02} - \sqrt{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2}}$$

where

$$\mu_{ij} = \sum_{x \in S} \sum_{y \in S} x^i y^j I(x, y).$$

$I(x, y)$ is the fluorescence intensity for the pixel in position x and y. We assumed that the spot pairs in focus, whose spots colocalized because of the FRB-FKBP heterodimerization, were the most abundant spots in the cell, and that all these spots shared similar properties in terms of brightness and shape. Therefore, a selection of the spots was performed by identifying the spots that cluster in a 2D space identified by the second momentum of brightness and the eccentricity. The cluster was determined using a 2D binned kernel density estimate (bkde2d function in R, CRAN). We estimated the 2D density distribution that the spot properties obeyed and then we selected those spots that presented a probability of 50% or higher to be found. The bandwidth of the kernel estimate was determined using dpik (Sheather and Jones, 1991). The spots selected for the distance estimate were those that matched the selection criteria in both channels (spots marked in red in Figure S2A, ''Selection of the spot pairs which are in focus'').

4. Quality of Gaussian interpolation: The centroid positions of the selected spots were computed using the center of brightness as in Sbalzarini et al. (Sbalzarini and Koumoutsakos, 2005). The selected spots were then interpolated with a Gaussian function. The quality of the interpolation was assessed by the coefficients of determination R^2_{RFP} and R^2_{GFP} for the red and green channel of each spot pair respectively. We imposed a refinement of the spot selection choosing only those spots that clustered in the space defined by R^2_{RFP} and R^2_{GFP} (Figure S2A, ''Refinement of the spot selection using the goodness of the Gaussian fit'').

5. Angle selection: Spot pairs colocalizing as a result of the FRB-FKBP heterodimerization should on average be organized with the green spot located farther away from the neighboring plasma membrane than the red spot. To increase the likelihood of selecting only those pairs, and not detecting RFP and GFP spot pairs which might be independent endocytic and exocytic sites occurring in close proximity on the plasma membrane, we measured the angle between the centroids of the fluorescent spots and the tangent to the closest point on the plasma membrane. We then chose spots whose angle difference was negligible ($< 0.05\ \pi$). To measure the angle we chose the point on the plasma membrane closest to the centroid of the RFP spot; we then measured the angle between the tangent to the plasma membrane passing through this point and the vectors linking this point to each spot centroid. An efficient recruitment of the prey-GFP to the Sla2-RFP-FKBP anchoring platform had a similar angle between the centroids of the spots and the tangent to their neighbor membrane (Figure S2A, ''Refinement of the spot selection choosing centroid pairs that have the same angle to the plasma membrane'' and the cartoon there).

6. Outlier detection: To exclude the remaining incorrect values among the measured distances we assumed that, given a set of N distance measurements $X = \{x_i\}_{i=1\ldots N}$, the majority of distances are true measurements. Those measurements follow a known non-Gaussian distribution (Churchman et al., 2006)

$$p(x; \mu, \sigma) = \left(\frac{x}{\sigma^2}\right) \exp\left(-\frac{\mu^2 + x^2}{2\sigma^2}\right) I_0\left(\frac{x\mu}{\sigma^2}\right);\ x \in X.$$

μ is the true distance separating the fluorophores, which we want to estimate, and σ is the variance of the distribution. I_0 is the modified Bessel function of integer order 0. μ and σ were determined maximizing the likelihood that we obtained the dataset X:

$$L(\mu, \sigma) = \prod_{x \in X} p(x; \mu, \sigma).$$

We defined an outlier as a contamination in the measurements that is unlikely in respect to the distribution p. If the dataset X has one outlier, this contributes to the likelihood with a low probability, which reduce the overall likelihood value. We thus rejected each measurement x_i, one at the time, and we measured the resulting likelihood for the set of measurements without x_i

$$\tilde{L}(\mu, \sigma) = \prod_{x \in X \setminus \{x_i\}} p(x; \mu, \sigma).$$

μ and σ are the values estimated by maximizing $L(\mu, \sigma)$. A candidate outlier x_{out} is defined as the measurement whose removal gives the highest value for $\tilde{L}(\mu, \sigma)$:

$$\prod_{x \in X \setminus \{x_{out}\}} p(x; \mu, \sigma) > \prod_{x \in X \setminus \{x_i\}} p(x; \mu, \sigma),\ \forall x_i \in X \setminus \{x_{out}\}.$$

We rejected x_{out} and iterated the process computing again a new estimate for μ and σ based on the new set of measurement, without the selected outlier. We then searched for the next outlier in the same way: removing each of the remaining measurement one at the time and computing the likelihood $\tilde{L}$ given the new estimated values of μ and σ.

To stop the outlier rejection we observed that the removal of an evident outlier changed the estimate of μ and σ more than the removal of an outlier that did not deviate much from the predicted distribution. In other words, the Maximum Likelihood Estimate

is very sensitive to strong outliers, which will be the first to be identified. Thus, the more we proceeded with the removal of outliers the smaller the difference between the old and new estimate of μ and σ, with and without the last outlier removed, would become. Let us label each of the iterations with an integer index *l*. Two subsequent outlier rejections will give two values of the distance: μ_l and μ_{l+1}. Their difference

$$\Delta\mu_l = \mu_l - \mu_{l+1}$$

will decrease for increasing values of *l*.

We thus defined a score

$$p_l^{\mu} = \frac{1/\Delta\mu_l}{\sum_j 1/\Delta\mu_j},$$

which is higher when the outlier rejection approximate a dataset of distances well described by $p(x;\mu,\sigma)$. Similarly for σ we could define the score

$$p_l^{\sigma} = \frac{1/\Delta\sigma_l}{\sum_j 1/\Delta\sigma_j}.$$

We then define a scoring function as

$$S\left(p_l^{\mu}, p_l^{\sigma}\right) = p_l^{\mu} \log p_l^{\mu} + p_l^{\sigma} \log p_l^{\sigma}.$$

The selected dataset, candidate to be the closest to the ''true dataset'' without outliers, was the one that maximized the scoring function (Figure S2B–Quantification of the distances between pairs of spots):

$$\max_l \left(S\left(p_l^{\mu}, p_l^{\sigma}\right)\right).$$

In the Table S2 under ''Number of measurements'' we list the number of spot pairs used to estimate the μ (Estimated distance) and the σ (Sigma) in each selected dataset. All error bars in Figure 2 and the Estimated Standard Errors for μ and σ in Table S2 are the Standard Errors estimated from the inverse of the observed Fisher information computed at the maximum of the likelihood.

Outlier rejection and distance estimation

To assess the accuracy of the outlier rejection we generated in silico data from distributions with known true distance and sigma (Churchman et al., 2006). We then contaminated the data with outliers mimicking the range of outliers we encountered experimentally and we run our analysis procedure. We could successfully reject the outliers and get a very accurate estimate of the true distance (Figure S2C).

Estimate of the likelihood of subunits being in close proximity

The exocyst subunits are represented as string of beads. We measured the separation among all bead pairs between different subunits for each of the 200 solutions. For each of the bead pairs we could thus estimate the distribution of all the separations measured in the 200 solutions. We used this distribution to estimate the likelihood that the two beads belonging to different subunits are located closer than 4 nm and we used the likelihood as a score to highlight bead pairs that are likely to be adjacent (Figure 5B).

Quantification of exocyst subunit abundances, image analysis

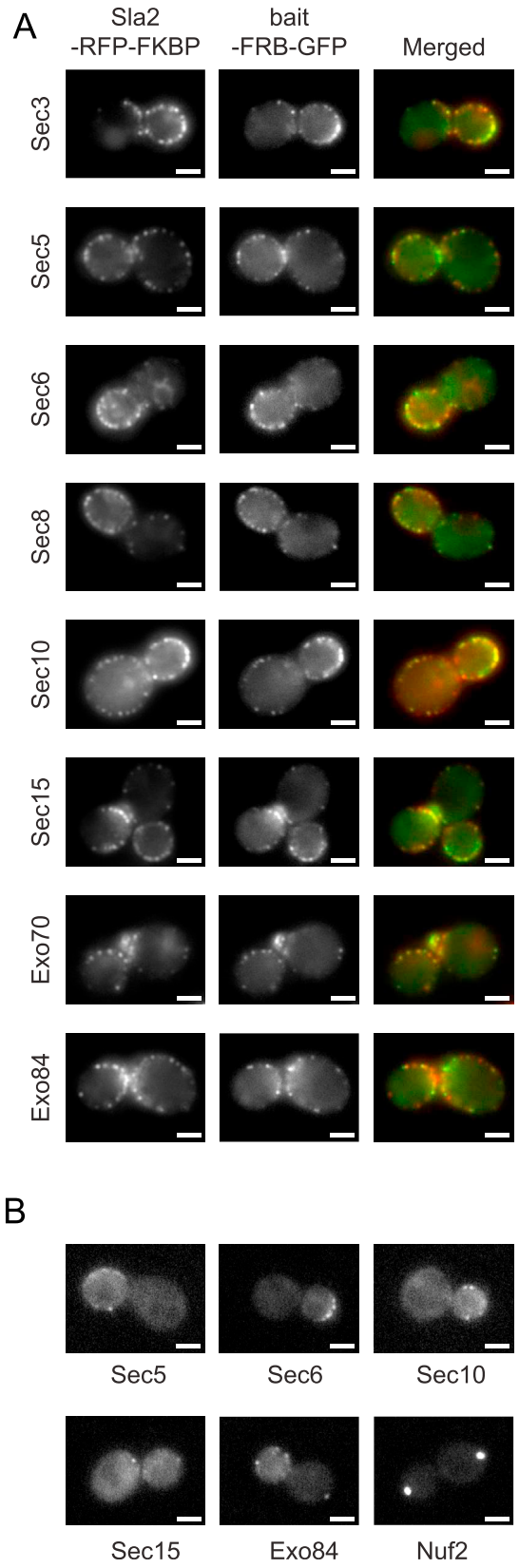
Images were analyzed with a custom software written in Python 2.7 that measured the fluorescence intensity of the fluorescent spots in the images corrected for local background (Joglekar et al., 2006). The number of Nuf2 molecules used to calibrate the fluorescence intensity was 280.6 ± 16.8 molecules each fluorescence spot (Picco et al., 2015).

DATA AND SOFTWARE AVAILABILITY

The collection of MATLAB scripts used to correct for chromatic aberration, R scripts used to compute distances between diffraction limited fluorescent spots of distinct fluorophores and the scripts used to quantify exocyst subunit abundances are available online (https://github.com/apicco/exocyst_scripts).

Files containing the scripts used to generate the 3D models and the actual 3D models of the exocyst and the COG protein complexes generated in this article are available online (https://github.com/batxes/exocyst_scripts).

Cell



(legend on next page)

Figure S1. Recruitment of Exocyst Subunits to the Anchoring Platform, Related to Figure 1

(A) Exocyst subunits were tagged to FRB and GFP at the C terminus (bait-FRB-GFP) in a strain carrying the Sla2-RFP-FKBP as anchor. Upon addition of rapamycin and LatA, cells were imaged for the anchor-RFP-FKBP (left column) and the bait-FRB-GFP (middle column).

(B) Examples of the equatorial plane of the z stacks acquired to quantify the number of Exocyst subunits at sites of exocytosis. Cells expressing Nuf2-GFP were mixed with the samples and used to calibrate the fluorescence intensity (see STAR Methods – Quantification of exocyst subunit abundances).

Scale bars are 2 μm long.

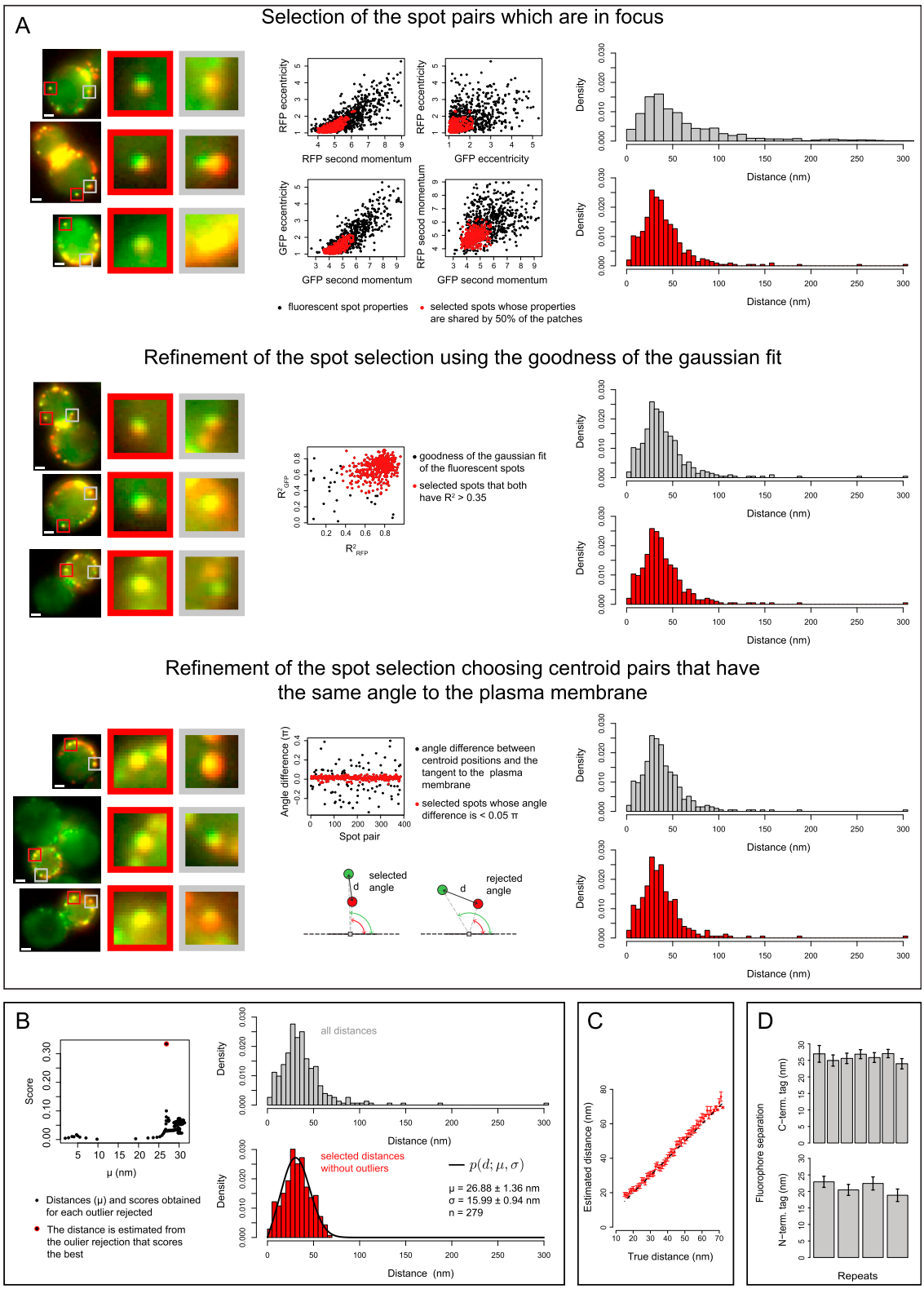


Figure S2. Image Analysis and Distance Estimate, Related to Figure 1

(A) The three steps of the image analysis pipeline: *Selection of the spot pairs which are in focus:* The eccentricity and the second momentum of brightness of the spot pairs were used to select the diffraction limited spots that were nicely in focus and well isolated. On average, 48% ± 7% (Median ± MAD) of the spots were

(legend continued on next page)

selected for further analysis during this step. The images on the left show examples of spots rejected (gray square) and selected (red square). The four central plots show how the selected spots (red) distribute in the space of eccentricities and second momenta. *Refinement of the spot selection using the goodness of gaussian fit:* First round of refinement of the selection of spot pairs. The goodness of the Gaussian fit is used to select the spots that are the nicest in both channels. On average, 88% ± 4% of the spots are selected. The images on the left show examples of the spot rejected (gray square) and selected (red square). The dot plot in the center shows the distribution of the R squared of the selected spots (red dots). *Refinement of the spot selection choosing centroid pairs that have the same angle to the plasma membrane:* Second round of refinement of the spot selection to avoid considering spots that recruited independently on the membrane or whose centroids might be shifted laterally by neighbor spots. The spot pairs whose centroids are at a different angle with respect to the closest tangent to the cell membrane are rejected. On average, 67% ± 5% spots are selected. The images on the left show the spots that are rejected (gray square) and selected (red square). The dot plot shows the difference between the angles of the centroid in the two channels. In red are the spots selected, whose angle difference is less than 0.05 π. For the three steps, the histograms on the right show the distribution of distances before the rejection of the spots (gray) and after the rejection (red); the step in the image analysis pipeline that reduces the most the noise in the dataset of distances is the first one. Scale bars are 2 μm long. The spots taken as examples have been magnified 5 times.

(B) The estimate of the distances: The selected distances are shown in the gray histogram. If outliers are present they are iteratively rejected; for each outlier rejected, we compute a new estimate of the separation between the fluorophores and we associate it to a score. The scores are shown in the dot plot. The spot circled in red marks the dataset whose outlier rejection scored the best and which is taken as the dataset without outliers (see STAR Methods – Image Processing and Distance Measurements section). This is the dataset shown in the red histogram. The true distance between the fluorophores was derived by Maximum Likelihood Estimate using the probability distribution that the distances obey (Churchman et al., 2006).

(C) The accuracy of the distance estimation and outlier rejection. We generated in silico datasets of distance measurements for the range of distance values that we measured experimentally and we contaminated these datasets with outliers mimicking those that we encountered experimentally. We then used these datasets to perform our outlier rejection and estimate the true distances, which were always an accurate estimate of the true distances used to generate the datasets.

(D) Assessment of the reproducibility of our approach. We repeated the sample preparation, sample imaging, data collection, and data analysis for strains expressing Sec5-GFP_C and Exo70-FRB (upper bar plot) or Sec5-GFP_N and Exo70-FRB (lower bar plot). Acquisitions were performed on different days. Error bars show the SE.

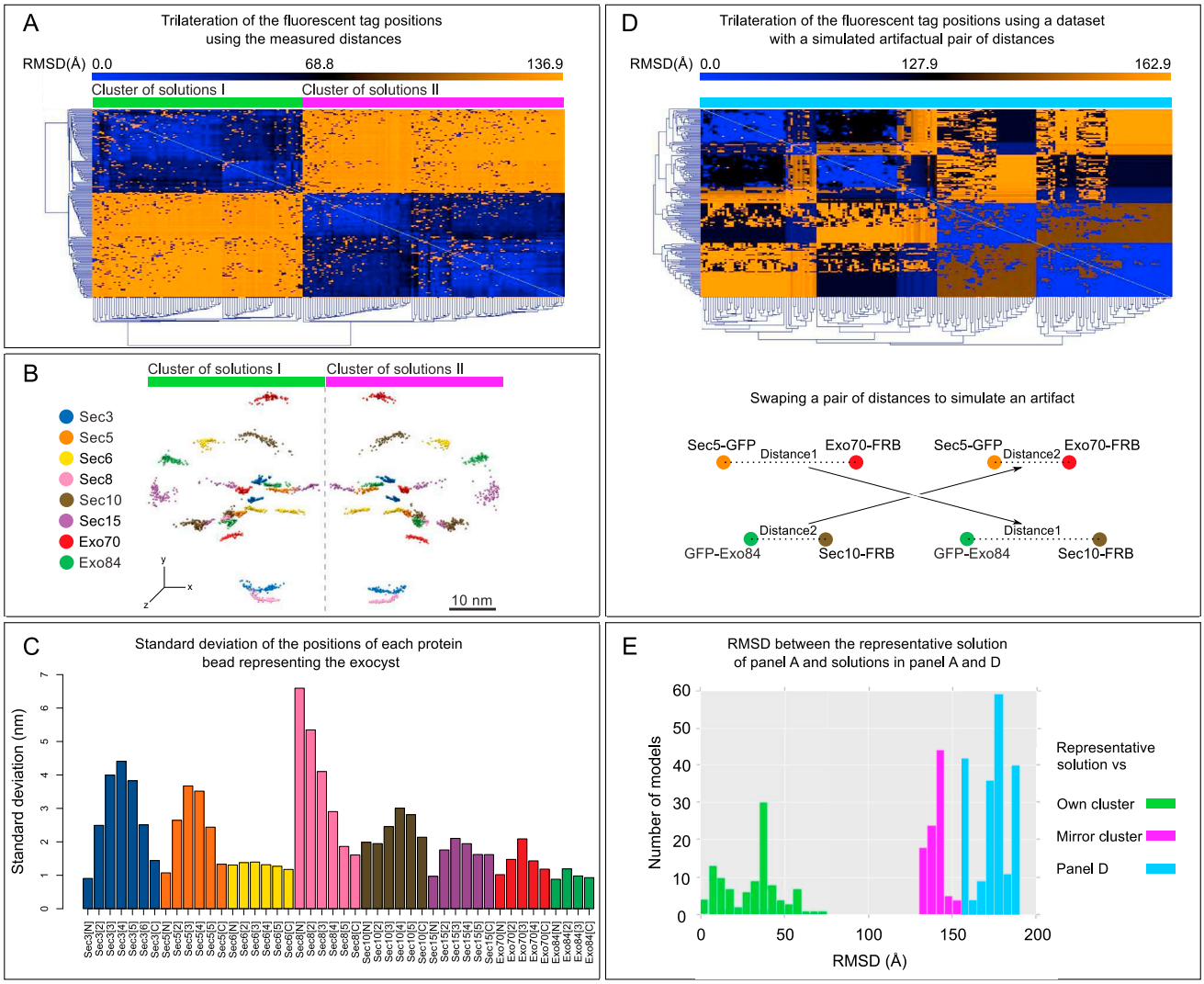


Figure S3. Solutions of the Tag Models, Statistics, and Control, Related to Figures 3 and 4

(A) Heatmap representing the RMSD of each model compared against all others. Low RMSD is colored in blue, high RMSD in orange. Value of RMSD in Å is written next to the colorbar. Two clusters representing two possible solutions are observed. Clusters are labeled with green and pink bars.

(B) Superposition of the fluorophore models of both clusters, which are mirror image of each other. Beads representing the fluorophores are color coded according to the subunit they are tagged to. Clusters are labeled as in (A).

(C) SDs of the positions of individual beads of each subunit, computed from the 200 solutions of the exocyst architecture.

(D) Control: heatmap representing the RMSD of each model compared to all others. Two distances in the generation of these models were swapped as explained in the schematic below. The color code and the units are the same as in (A). Various clusters are observed. The population of models is represented with a teal bar.

(E) Histogram showing the RMSD between the representative fluorophore model and cluster of solutions I (the one of the representative model), cluster of solutions II (mirror image solutions of cluster I) and models in (D) (control). Models belonging to own cluster have values between 0 and 75 Å and models belonging to the mirror image cluster vary between 130 and 155 Å. All models from (D) have values above 155 Å, indicating that they are not similar to the original configurations of fluorophores used for the exocyst modeling.

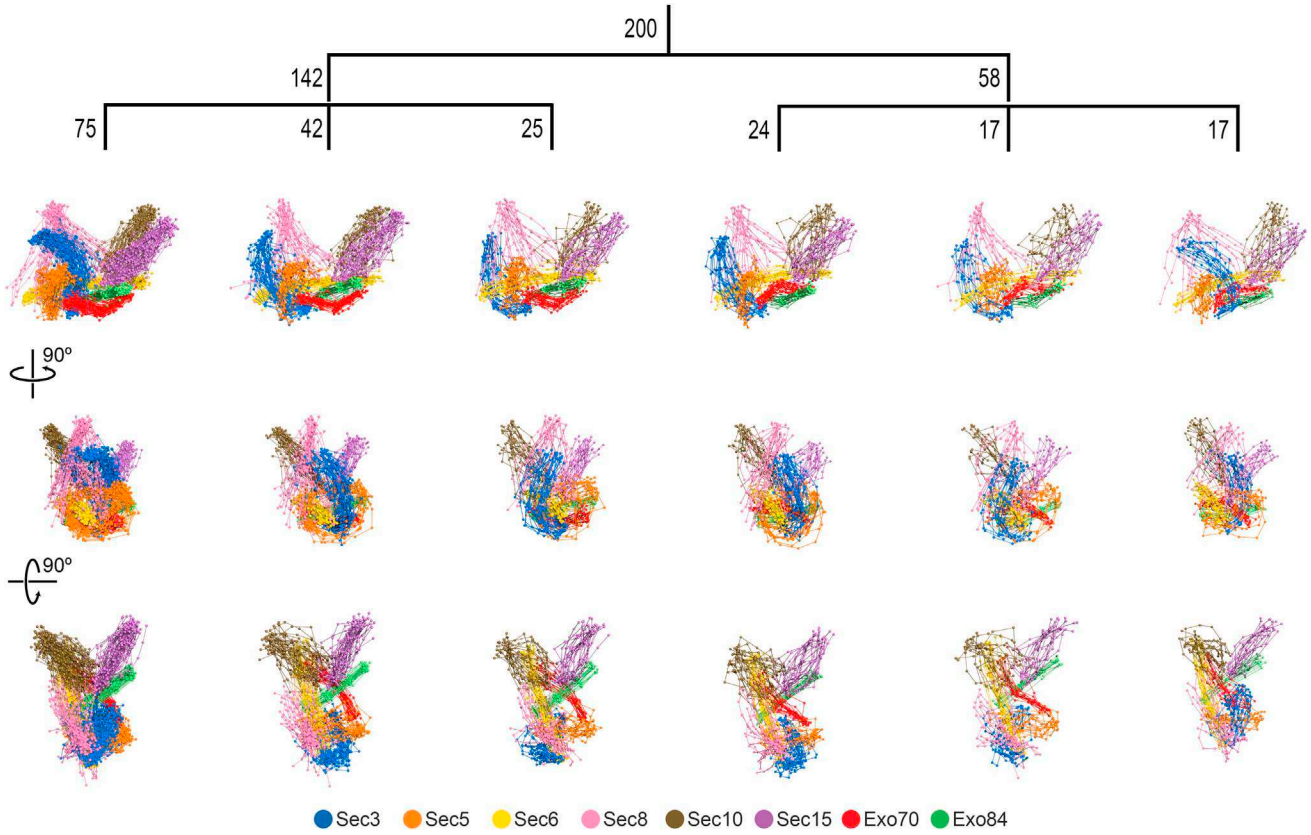


Figure S4. Clustering of All Solutions Obtained for the Computed Exocyst 3D Architecture, Related to Figure 4

The 200 solutions obtained for the exocyst 3D architecture clustered by RMSD (see STAR Methods). Models belonging to each cluster are superposed and shown from different perspectives. The number associated to each branch of the tree (top) indicates the number of models in each cluster. Exocyst subunits are color coded as indicated.

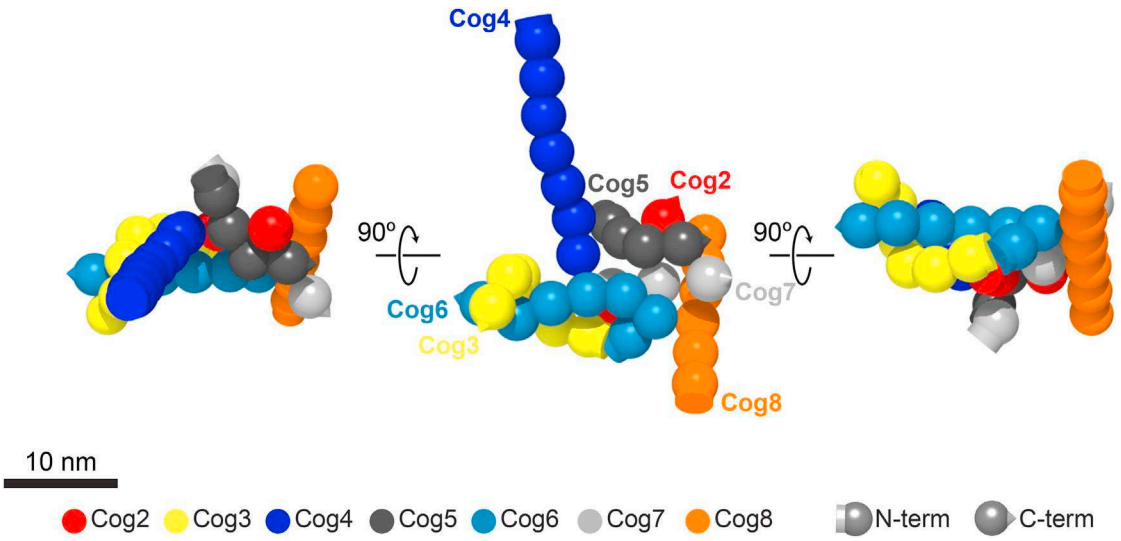


Figure S5. The 3D Architecture of the COG Complex, Related to Figure 3

Different views of the COG complex reconstruction with best IMP score are used to illustrate the 3D architecture of the COG complex. COG subunits are represented with beads. The volume of each bead is approximately equivalent to 120 amino acids folded in a helix-bundle.

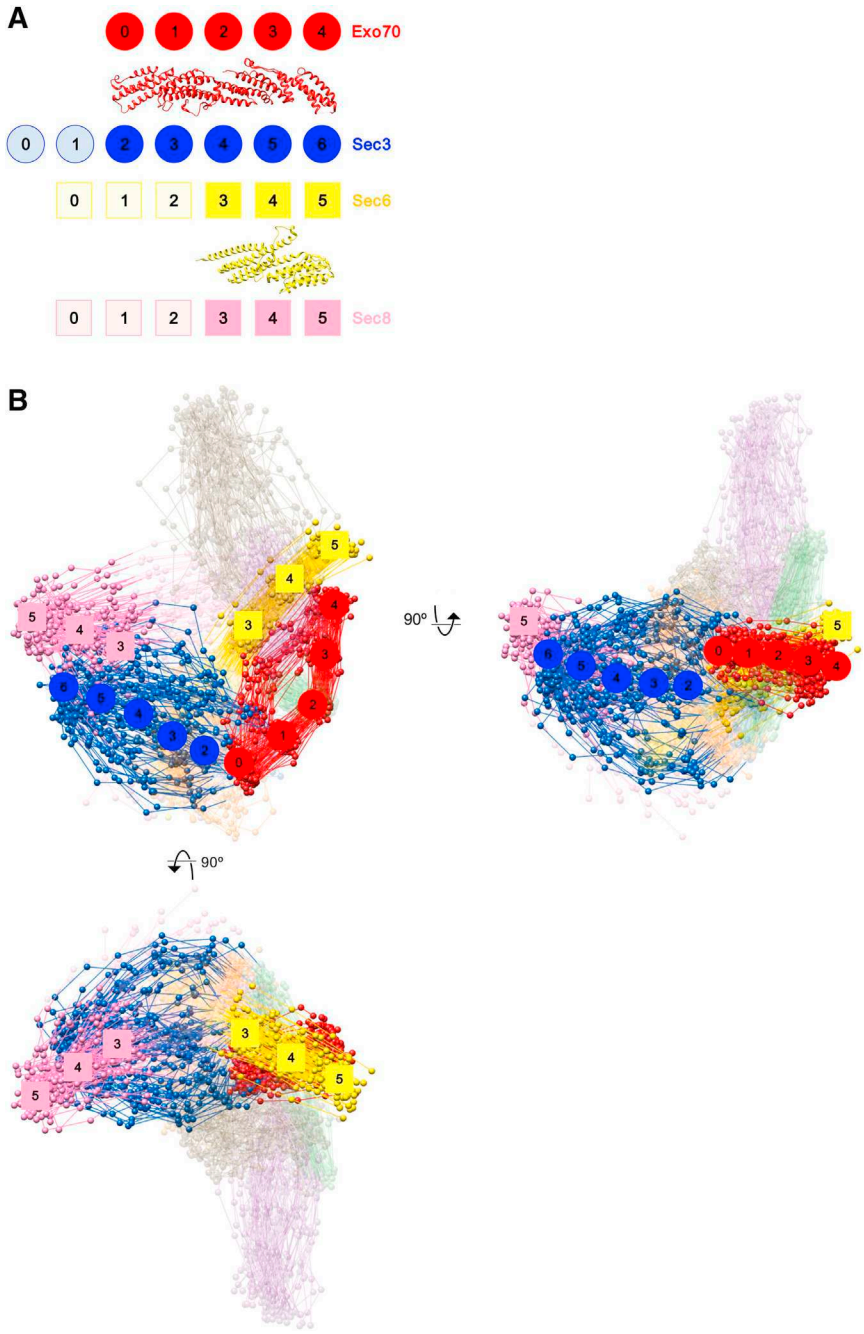


Figure S6. Partial Symmetry in the Exocyst Complex, Related to Figure 4

(A) Bead representation of the Exo70, Sec3, Sec6 and Sec8 subunits are color coded. The atomic structure of Exo70 is displayed in red (PDB entry 2b1e_A) and represented by red circles. Dark blue circles represent the fragment of Sec3 that could be modeled using Exo70 as template. Likewise, the atomic structure of the C-terminal domain of Sec6 is represented in yellow (PDB entry 2fji_1) and the corresponding fragment is highlighted by dark yellow squares. A comparative model of the C-terminal domain of Sec8 based on the C-terminal domain of Sec6 is represented by the last three dark pink squares.

(B) The beads sharing homology are mapped in the 100 best models of the exocyst.

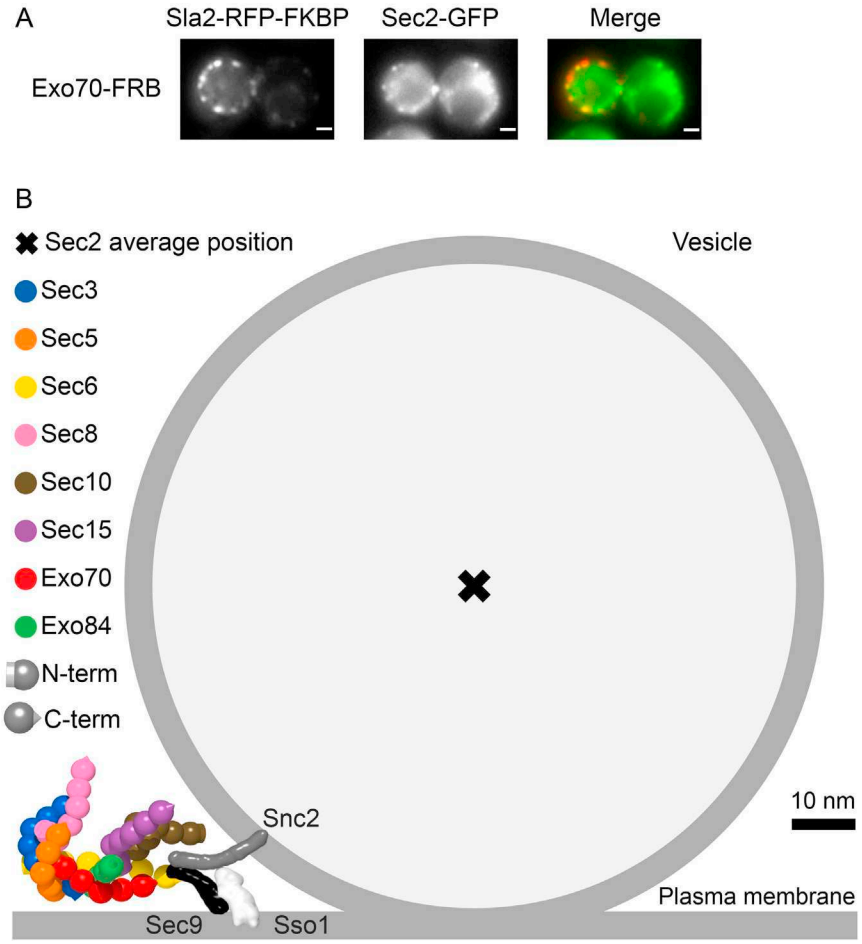


Figure S7. Fitting of the Exocytic SNARE Complex in the Cavity between the Exocyst, the Secretory Vesicle, and the Plasma Membrane, Related to Figure 7

(A) Recruitment of Sec2-GFP to Sla2-RFP-FKBP anchoring platforms. Sec2 was fused at the C terminus to GFP and recruited upon rapamycin and LatA treatment using Exo70-FRB as bait. Scale bars are 1 μm long.

(B) The exocytic SNARE complex was reconstructed using comparative modeling (see STAR Methods) and it is depicted with surface representation (Sec9 in black, Sso1 light gray and Snc2 dark gray). The exocytic SNARE complex is manually positioned taking into account that the C terminus of Snc2 is attached to the vesicle and the C terminus of Sso1 is bound to the plasma membrane.

2

4Cin: a computational pipeline for 3D genome modeling and virtual Hi-C analyses from 4C data

4Cin: a computational pipeline for 3D genome modeling and virtual Hi-C analyses from 4C data

Ibai Irastorza-Azcarate*, Rafael D. Acemel, Juan J. Tena, Ignacio Maeso, José Luis Gómez-Skarmeta* and Damien P. Devos*.

Centro Andaluz de Biología del Desarrollo (CABD), Consejo Superior de Investigaciones Científicas/Universidad Pablo de Olavide, Seville, Spain.

*Corresponding authors: ibai.irastorza@gmail.com, jlgomska@upo.es and damienpdevos@gmail.com

Abstract

The use of 3C-based methods has revealed the importance of the 3D organization of the chromatin for key aspects of genome biology. However, the different caveats of the variants of 3C techniques have limited their scope and the range of scientific fields that could benefit from these approaches. To address these limitations, we present 4Cin, a method to generate 3D models and derive virtual Hi-C (vHi-C) heat maps of genomic loci based on 4C-seq or any kind of 4C-seq-like data, such as those derived from NG Capture-C. 3D genome organization is determined by integrative consideration of the spatial distances derived from as few as four 4C-seq experiments. The 3D models obtained from 4C-seq data, together with their associated vHi-C maps, allow the inference of all chromosomal contacts within a given genomic region, facilitating the identification of Topological Associating Domains (TAD) boundaries. Thus, 4Cin offers a much cheaper, accessible and versatile alternative to other available techniques while providing a comprehensive 3D topological profiling. By studying TAD modifications in genomic structural variants associated to disease phenotypes and performing cross-species evolutionary comparisons of 3D chromatin structures in a quantitative manner, we demonstrate the broad potential and novel range of applications of our method.

Author summary

Chromatin conformation capture (3C) methods have revealed the importance of the 3D organization of the chromatin, which is key to understand many aspects of genome biology. But each of these methods have their own limitations. Here we present 4Cin, a software that generates 3D models of the chromatin from a small number of 4C-seq experiments, a 3C-based method that provides the frequency of contacts between one fragments and the genome (one vs all). These 3D models are used to infer all chromosomal contacts

within a given genomic region (many vs many). The contact maps facilitate the identification of Topological Associating Domains boundaries. Our software offers a much cheaper, accessible and versatile alternative to other available techniques while providing a comprehensive 3D topological profiling. We applied our software to two different loci to study modifications in genomic structural variants associated to disease phenotypes and to compare the chromatin organization in two different species in a quantitative manner.

Introduction

The three-dimensional (3D) architecture of the genome is important for most of its functions, such as gene expression regulation and DNA replication[1–3]. As with proteins, knowledge of the 3D structure of a genomic locus can reveal information not accessible from its primary sequence only. Indeed, the use of chromosome conformation capture (3C) methods together with high-throughput sequencing has profoundly changed our understanding of the 3D nuclear organization, adding a new dimension to the study of genome biology.

Amongst those new key findings is the discovery that the genomes of diverse animal lineages are organized in topologically associating domains (TADs)[4–7], genomic regions that typically span less than one Mbp within which the chromatin has a higher propensity to interact with itself. TADs are broadly preserved in interphase across different cells[4,8], they provide a structural basis to regulatory landscapes[1,9] and their structural perturbation has been linked to diseases[10–12]. Accordingly, TADs are largely conserved across different species[4,13,14].

Despite the growing interest in studying genomic information from a 3D perspective, 3C-based methods are still far from reaching their full potential to investigate a wider range of biological questions, partly because of the inherent limitations of these methods. All 3C technologies are based on similar biochemical principles to capture chromatin interactions, although with important variations (reviewed in [15,16]). They all start by cross-linking chromatin fragments that are located in close proximity in the nuclear space; the genome is then digested and ligated to capture interacting regions. Afterwards, these regions are identified and quantified by PCR or sequencing. Each 3C technique has its own experimental biases, but more importantly, they have different scopes, resolutions, costs, sequencing depths and data processing requirements[15]. Hi-C addresses chromatin contacts between all the regions in the genome and it is currently the only technique that allows the identification of genome-wide, large-scale genomic organizational features. However, this

comes at the cost of losing power to determine fine-scale intra-TAD interactions, which are precisely the ones responsible for the regulation of individual genes and therefore of special interest in a variety of biomedical and genetics fields. This can in principle be overcome by performing Hi-C at the highest possible resolution, but this requires sequencing several billions reads per sample, implying financial costs exceedingly high for the vast majority of laboratories. 4C-seq (Circular Chromosome Conformation Capture) provides a good alternative solution for some of these problems. This technique is able to identify all the interactions of a given region of interest, usually termed 'viewpoint'. With just ~1 million reads, 4C-seq can generate detailed high-resolution interaction profiles for a single locus. This high sensitivity and reduced sequencing cost has made this method particularly suitable for studies comparing multiple samples, between different species, genotypes or developmental stages, where it has been widely used to identify interactions between distal enhancers and gene promoters. Moreover, the recently developed NG Capture-C (next-generation Capture-C) technique[17] yields 4C-seq-like data in a high-throughput manner and of a higher resolution, making it a suitable technique to get detailed information of a certain locus, since multiple probes for multiple viewpoints within the region of interest can be designed.

Notwithstanding these advantages, both 4C-seq and NG Capture-C have also important limitations and provide incomplete information about TAD topology and borders, even when several viewpoints are used. Thus, in the absence of complementary Hi-C information from the same species, it may be difficult to get a complete and integrated picture of the interactions of a certain region. Finally, other technologies such as 5C (Chromosome Conformation Capture Carbon Copy) and Capture Hi-C (when designed to target a particular region using a tiled oligonucleotide capture approach), bridge somehow the gap between Hi-C and 4C-seq, being able to identify the large scale 3D chromatin organization of a given locus together with a high resolution contact map. Furthermore, as in the case of 4C-seq, they require a modest amount of sequencing depth. However, both approaches rely on the use of hundreds to thousands of probes or oligonucleotides from which the interaction profiles are identified and the costs and experimental design to produce these probes are far from trivial.

In sum, currently there is no experimental tool that combines, in a cost-effective manner, high-depth interaction profiles for particular loci with Hi-C-like information on TAD-level organization, hampering the accessibility of C-techniques to a wider number of scientists that will strongly benefit by incorporating 3D chromatin studies in their research.

Integrative modeling methods provide versatile approaches to infer 3D structures, since they are able to consider information derived from different techniques simultaneously. There are several integrative modeling method tools available at the moment that given a matrix of distances between genomic elements inferred from 3C contact frequencies, can compute the localization in the 3D space of these genomic elements[18–21]. These methods mostly use 5C or Hi-C based matrices as input data for the reconstruction of the genome structure, but none of them use 4C-seq-like data[22–25]. We have recently shown that 3D chromatin models can be successfully reconstructed from a small number of 4C-seq interaction profiles[3]. Here, we present 4Cin, a completely automated and easy to use pipeline to generate 3D chromatin models from 4C-seq data. 4Cin can also generate models using 4C-seq-like data coming from recently developed techniques such as NG Capture-C or Capture-C, as long as they are used to capture at least 4 viewpoints within each region(s) of interest. 4Cin also allows the generation of vHi-C maps, the identification of TADs boundaries, the comparison of 3D structures and the integration of 3D structures with different epigenetic features. Here, we show the utility of 4Cin with two detailed case-studies that highlight some of the most important fields of application of our method: the study of genomic loci affected by structural variations causative of aberrant phenotypes using the mouse *Shh* locus, and evolutionary comparisons of 3D chromatin structures across different vertebrate species using the Six gene clusters.

Results

The tool: 4Cin, a 4C-seq to 3D pipeline

4Cin was developed as an alternative to Hi-C to study particular genomic regions. Data from 4C-seq experiments are integrated to obtain 3D models that are represented afterwards as a vHi-C, a Hi-C like matrix of a given genomic locus (Fig 1 and S1). The tool was developed around IMP, the integrative modeling platform[26]. The tool was developed to handle data coming from multiple cells. Thus, the output models are representative of the average conformation of the chromatin in all cells and variability between models has not been shown to be related to chromatin dynamics.

Modeling the chromatin as a string of beads

The genome is represented as a flexible string of beads (Step 1 in Fig 1). The diameter of the beads corresponds to the theoretical length of the portion of straightened chromatin that we are representing, assuming the canonical chromatin width of 30 nm[27,28]. Beads are allowed to inter-penetrate, since we

assume that the chromatin is unlikely to be straightened, occupying the full volume of the bead. We have previously shown that this type of representation generates robust results[3].

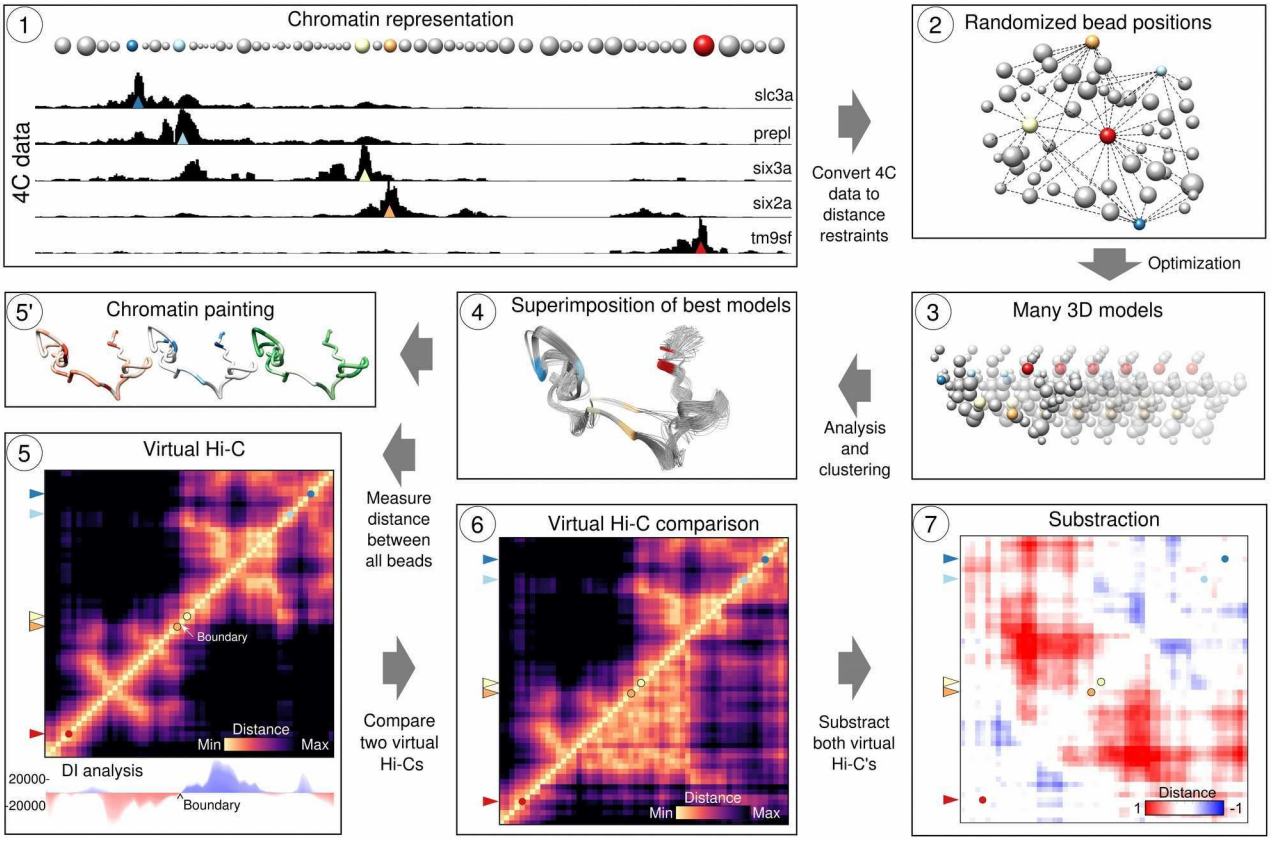


Fig 1. 4Cin pipeline.

(1) A genomic locus is represented as concatenated beads. Beads representing the viewpoints are color coded. The size of the beads is proportional to the size of their corresponding 4C-seq fragments. 4C-seq data is translated into distance restraints that are used in the optimization step. (2) Bead positions are optimized from random start positions. (3) Models that fulfill most of the restraints (i.e. with the best scores) are gathered and clustered based on their RMSD. (4) Models belonging to the most populated cluster are gathered and superimposed. (5') The most representative model can be painted using genomic or epigenomic data. (5) Distance between the beads representing the 3D models is measured from the population of best models and represented as a virtual Hi-C. Directionality index can be calculated to infer TAD boundaries. Two virtual Hi-C's can be compared (6) and subtracted (7).

3D reconstruction of the chromatin: 4C-seq counts as a proxy for distances

The central assumption of all 3C-derived integrative modeling methods is that read counts and physical distances are inversely related; high read counts connecting two DNA fragments imply close proximity

between them whereas low counts imply larger distances. Accordingly, 4Cin uses these distance proxies as restraints (Steps 1 and 2 in Fig 1). Therefore, each 4C-seq experiment includes sequencing data that are interpreted as a pool of distances to the corresponding viewpoint. After various iterations of optimization of the position of the beads and evaluation of their fit with the restraints, a model that fulfills as many of the distance restraints as possible is generated. The optimization procedure combines a Monte Carlo exploration with steps of conjugate gradients as local optimization and simulated annealing. The fulfillment of the restraints is expressed as a score, where a score of 0 represents the fulfillment of all the restraints. The optimization process for each model ends when the score reaches a plateau or reaches 0. The process is repeated many times, generating many (typically 50000) models, in order to explore as completely as possible the variability between the models (Step 3 in Fig 1). A subset of the models that best fits the available data (i.e. those with the best scores) is analyzed afterwards (Step 4 in Fig 1). The end point of 5C or Hi-C experiments is a matrix of contact frequencies represented as a heat map. Hi-C heat map plots show the frequency of interaction between all pairs of DNA fragments which, given the initial 3C assumption, is used as a proxy for spatial proximity. A contact map mimicking a Hi-C heat map, in essence, a 'vHi-C map', can be generated by averaging the distances between all beads in the best 3D models (Step 5 in Fig 1).

To check the robustness of our method, we have generated 3D models of the *six2a-six3a* locus in zebrafish and generated vHi-Cs down-sampling the input 4C data, using a variable percentage of the original 4C-seq read counts to generate the models. The high correlation (Spearman rank correlation $\rho > 0.7$) of the vHi-C's even when only 5% of the original data are used in the modeling, proves that 4Cin is robust to the sequencing depth of the underlying 4C data. We also carried out an unbalanced down-sampling, where three of the five 4C-seq experiments where down-sampled 95% and we also generated models where the raw 4C-seq data was modified, inserting read counts corresponding to the value representing the 95th percentile of the data, as erroneous data, in randomized positions. We generated 3 rounds of modeling, with 1%, 2% and 5% of errors inserted. We were still able to get high correlations (Spearman rank correlation $\rho > 0.7$), supporting even more the robustness of our method (S2 Fig).

Our tool can be parallelized, allowing an acceleration of the process. 50.000 models based on a data set of 5 4C-seq experiments and represented by 56 beads, can be generated in about half an hour, on a computer with 20 cores and CPU's of 2.5GHz. A region with 14 different 4C-seq experiments and 211 beads can be modeled in 7 hours.

Choosing the viewpoints

4Cin modeling is possible with as few as four 4C-seq datasets (distances from four different viewpoints; to be able to position each DNA fragment of the genomic locus in 3D space), but it is important to take into account that in order to leverage the complementarity of the data, the viewpoints should be well distributed along the entire locus. To show the importance of the distribution of the viewpoints, we modeled the *Six2-Six3* locus in mouse (Section 3.3) with three different sets of four viewpoints (S3 Fig). The correlation between the vHi-Cs and the original Hi-C suggests that a small number of viewpoints can generate reliable models, as long as these viewpoints are well distributed along the locus and not focused near the corners. Importantly, we have previously shown with jackknifing experiments that vHi-C maps obtained from 3D models are very robust in terms of the number of viewpoints used, being able to accurately recapitulate original vHi-C results even when 10 out of 14 viewpoints are eliminated (average increase in correlation of 0.12)[3].

Therefore, although the quality of the 3D reconstruction improves by increasing the number of viewpoints provided, this improvement is relatively minor and furthermore, it is paralleled by an increase in computational cost. Thus, based on our experience[3], data coming from between four and ten 4C-seq assays are enough to achieve reliable models of a locus of 2Mbp.

The quality of the data is also important in order to generate reliable models. The tool provides a script to check the quality of the 4C-seq data before starting with the modeling steps (S4 Fig). Kurtosis and skewness values are calculated in order to check the suitability of the data for the modeling[29]: Kurtosis value measures the shape of the distribution, accounting for the central peak and the tails, while skewness value informs about the symmetry of this distribution.

Postprocessing analyses: TAD border calling, vHi-C comparisons and genome painting

TADs are major organizational elements of the chromatin and their organizations are informative about the overall architecture of specific loci. We provide a script that identifies TAD boundaries using the directionality index[4] (Step 5 in Fig. 1). The script calculates the directionality index iteratively, ranging between the biggest (all beads) and smallest (one bead) possible size for a TAD, delivering a set of potential TAD boundaries.

TADs display important structural information, but combining 3D chromatin structure with epigenetic data can also reveal valuable information that is more difficult to observe from a linear perspective. Beads representing the chromatin can be colored with gradients according to genomic and epigenomic data. As

examples, here, we colored the representative chromatin model of the wild type *Shh* locus using CTCF ChiP-seq data (GEO accession: GSM918741)[30]. As expected, the beads with the highest read counts are found near the TAD boundaries[8,14,31,32] and contain high score CTCF binding motifs. We also checked for the orientation of the CTCF binding sites in these peaks and observed the convergent orientation typically found flanking chromatin loops[8] (S5 Fig and S1 Table).

Moreover, two scripts to compare vHi-Cs are provided in the 4Cin package. One allows comparing the organization of homologous loci in different species providing conserved regions, which generates a heat map where each triangle represents a locus (Fig. 4). The other one permits the comparison of different conformations of a region that underwent structural variation or mutation. This one yields a subtraction of both vHi-Cs. Both scripts calculate the correlation between the vHi-Cs that are being compared.

Below we demonstrate the use of the different tools implemented in our 4Cin method studying structural variations as well as evolutionary comparisons of 3D chromatin structures.

Structural variation studies: Disruption of long-range regulation in the *Shh* locus

Genomic mutations that compromise the structural integrity of TADs, such as inversions, duplications and boundary element deletions, have been shown to cause severe transcriptional mis-regulation of their associated genes, leading to the appearance of diverse disease phenotypes[10–12]. To illustrate the utility of 4Cin in understanding the molecular nature and effects of these structural genomic mutations, we focused on the region surrounding the gene sonic hedgehog (*Shh)*, a locus encoding a key diffusible signaling molecule for vertebrate development. *Shh* regulatory landscape spans over 900 kb, comprising several unrelated neighboring genes and multiple long-range enhancers, including one of the most distal enhancers identified so far, the *Shh* limb-specific enhancer known as ZRS. Previous works using 4C-seq data have shown that in mice with genomic mutations in the *Shh*-TAD, such as inversions, deletions and duplications, *Shh* regulatory interactions and expression were impaired, causing severe malformations[33]. In particular, INV(6-C2), a large 600 kb inversion encompassing nearly half of the *Shh*-ZRS TAD, greatly diminished 4C-seq contact frequencies between the ZRS and the *Shh* promoter. By applying 4Cin to these published 4C-seq datasets, we generated 3D models for both the wt *Shh* locus and the INV(6-C2) inversion mutant genotype (Fig. 2).

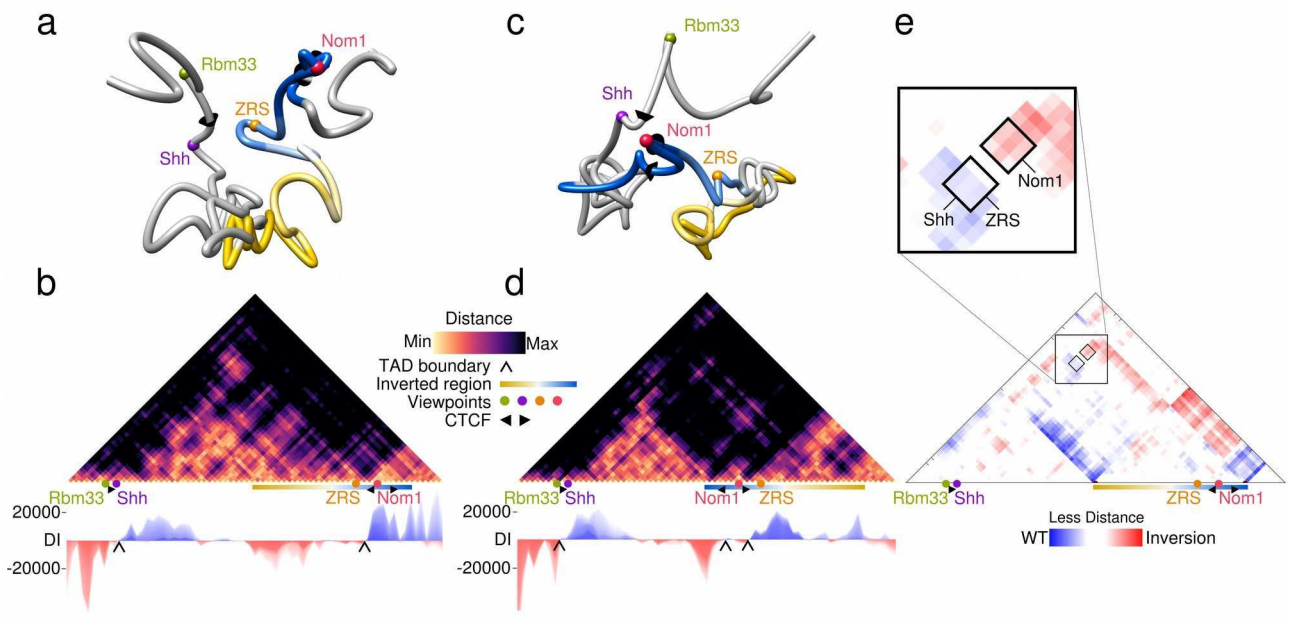


Fig 2. ZRS enhancer lies outside the *Shh*-TAD in mutant mice for the INV(6-C2) inversion.

(a) Representative 3D model of the WT *Shh* region. Viewpoints are depicted as colored beads. CTCF binding sites are represented as oriented cones. The genomic region included in the inversion is colored with a yellow-to-blue gradient. (B) Virtual Hi-C of the WT *Shh* region (Top). Directionality index (Bottom) was applied to call TAD boundaries, showed with black arrows. (C, D) 3D model and virtual Hi-C heat map of the INV(6-C2) mutant, showed as in (A) and (B). (E) Subtraction of heat maps (B and D), blue corresponds to shorter distances in the WT, red to shorter distances in the mutant. The zoom-in shows that in WT mice, *Shh* is close to ZRS and far away from *Nom1* in comparison with the mutant.

This revealed that the two corresponding chromatin topologies are markedly different: whereas in the wt *Shh* and the ZRS lie in close proximity, they are widely separated in the inversion (Fig 2A and 2C). In fact, vHi-C maps derived from these models and subsequent TAD border calling showed that the inversion completely changed the relative locations of some of the TAD boundaries, most likely due to changes in the relative orientations of the CTCF binding sites located next to *Nom1* and ZRS (Figs 2B, 2D and S5).

Thus, in the mutant genotype, the ZRS enhancer together with nearly half of the *Shh* regulatory landscape, are now part of another TAD. This enhancer is therefore isolated from the *Shh* promoter, explaining the reduced contact frequencies observed previously[33]. Indeed, a global quantification of distance changes across the entire locus by comparing the two vHi-Cs contact matrices showed that the distance between ZRS and *Shh* in the 3D models increases in the inversion (Fig 2E).

The topology of the *Shh* locus explains its regulatory organization

Using a large collection of insertions of regulatory sensors at multiple locations within the *Shh* regulatory landscape, the responsiveness to enhancers of different regions within the *Shh*-ZRS TAD was evaluated[33,34]. The results showed that most regions within the TAD were able to respond to at least some of the multiple tissue-specific *Shh* enhancers. However, there were a few insertion locations with no or very little responsiveness. Given that these regulatory "blind spots" did not show any particular location trend in terms of their linear distance to the enhancers (in particular to the ZRS) or local chromatin features such as histone marks or accessibility, the authors hypothesized that the lack of responsiveness may be related to their position within the 3D native structural folding of the locus. To test this hypothesis we mapped the positions of all the insertion sensors to a high resolution 3D chromatin model. We also located the positions of the comprehensive collection of *Shh* regulatory elements so far identified[35–40], which allowed us to define a 3D space containing all known *Shh* enhancers (Fig 3A). We then classified insertion sensors into three groups (high, low and no expression) depending on the level of expression of their associated reporter genes[33,34,41] (S2 Table). Consistent with the proposed hypothesis[33], these different expression activities of the sensors correlated inversely with their average distance to the enhancers (Spearman rank correlation $\rho < 0.05$) (Fig 3C), accordingly, most of the high expression sensors fell inside the enhancer area (Fig 3B). This supports the idea that the low enhancer responsiveness of certain chromatin regions is related to their topological position and their ability to interact with the different enhancers in the locus.

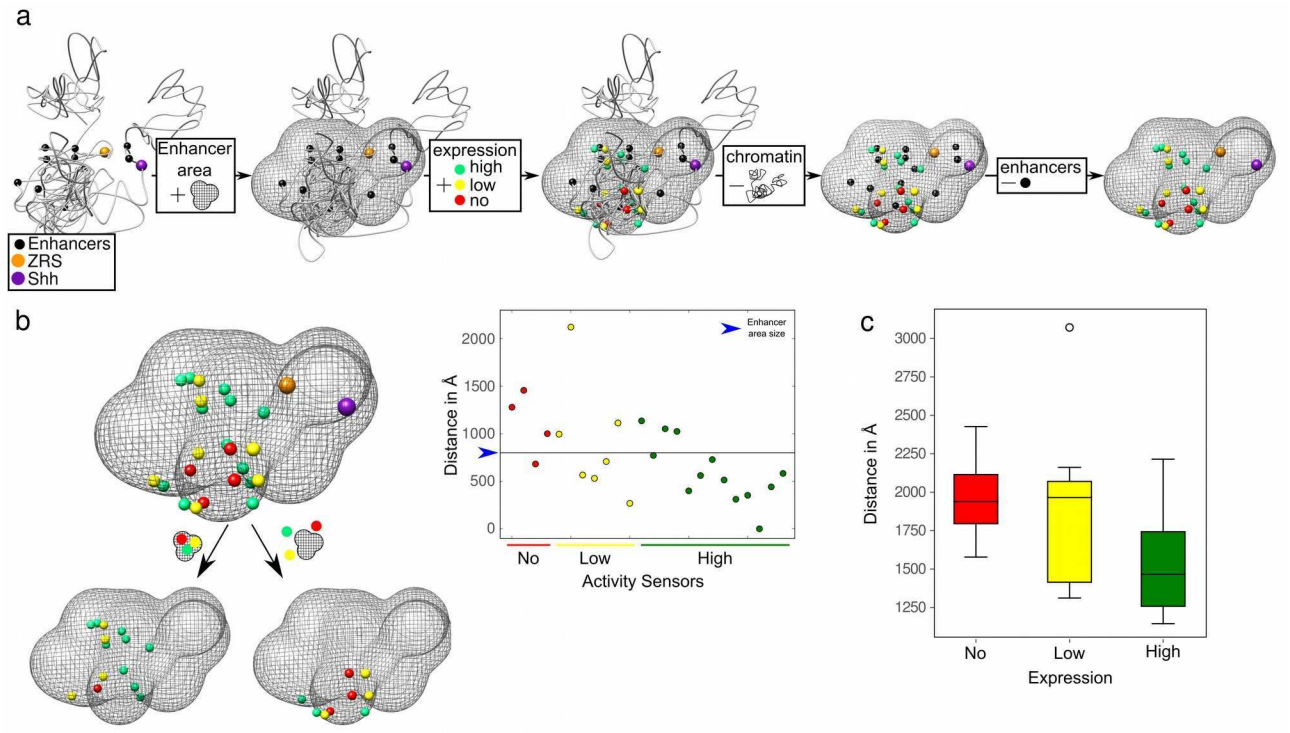


Fig 3. Insertion sensors with high responsiveness are close to enhancers in the 3D space.

(A) Stepwise explanation showing how we obtain top figure in panel B. Beads are color coded to indicate different regions: enhancers (black), *Shh* promoter (purple), ZRS (orange), and the three type of insertions, high, low and no expression (green, yellow and red, respectively). The enhancer area at 75 nm away from the enhancers is shown with a gridded surface. (B) Insertions, *Shh* and ZRS locations relative to the enhancer area. Below, beads that are outside and inside the area are depicted. On the right, barplot showing the distance between enhancers and insertions. (C) Boxplot showing average distance between beads representing the insertions and enhancers.

In conclusion, in comparison with 4C-seq alone and without generating any additional experimental data, the use of 4Cin provides further and deeper insights into the structure and regulatory interactions of a chromatin locus, generating a more complete characterization of the region, with identification of TAD borders, quantitative comparisons between different genetic backgrounds and testing specific hypothesis related to topological interactions.

Comparative genomics and evolution: conserved 3D chromatin structures in the *Six2-Six3* gene cluster of bony vertebrates

The evolution of genome architecture in animals has been traditionally studied using comparative genomics methods that can only consider DNA sequences from a linear perspective. The advent of 3C-based techniques has literally added a new dimension to this field, but so far cross-species comparisons of 3D chromatin structures have been performed only in a handful of species (in particular mammals) and have mostly relied on the use of Hi-C data[4,13]. This situation currently restricts the development of 3D-aware comparative genomic studies, since they would ideally involve the use of evolutionary relevant species for which Hi-C data is either still unavailable or difficult to produce, especially in cases comparing multiple lineages. We applied 4Cin to compare orthologous genomic loci, the *Six2-Six3* gene clusters, from two bony vertebrate species, mouse, a mammal with several published Hi-C data, and zebrafish, a teleost fish for which Hi-C data are still unavailable.

The Six2-Six3 locus is conserved in vertebrates

Six homeobox genes are essential developmental regulators organized in genomic clusters conserved in multiple animal phyla[14,42]. The clusters consist of three subfamilies: Six1/2, Six3/6 and Six4/5. Due to the two rounds of whole genome duplications that happen at the origin of vertebrates, most species within this group have two paralogous copies of the cluster, one containing *Six2* and *Six3* genes, and the other containing *Six1*, *Six6* and *Six4* genes. Teleosts, like zebrafish, have undergone another round of duplication and contain four Six clusters.

Here we use available 4C-seq data to explore the conformation of the cluster containing the *six2a* and *six3a* genes in zebrafish, which has been described to have a bipartite organization that split the regulatory landscapes of each of these genes into two different adjacent TADs[14]. The 3D models of the *six2a-six3a* locus in zebrafish and their derived vHi-C show two TADs with the Six genes located between them (Fig. 4a and 4b), corroborating previous results based on 4C-seq profiles only[14].

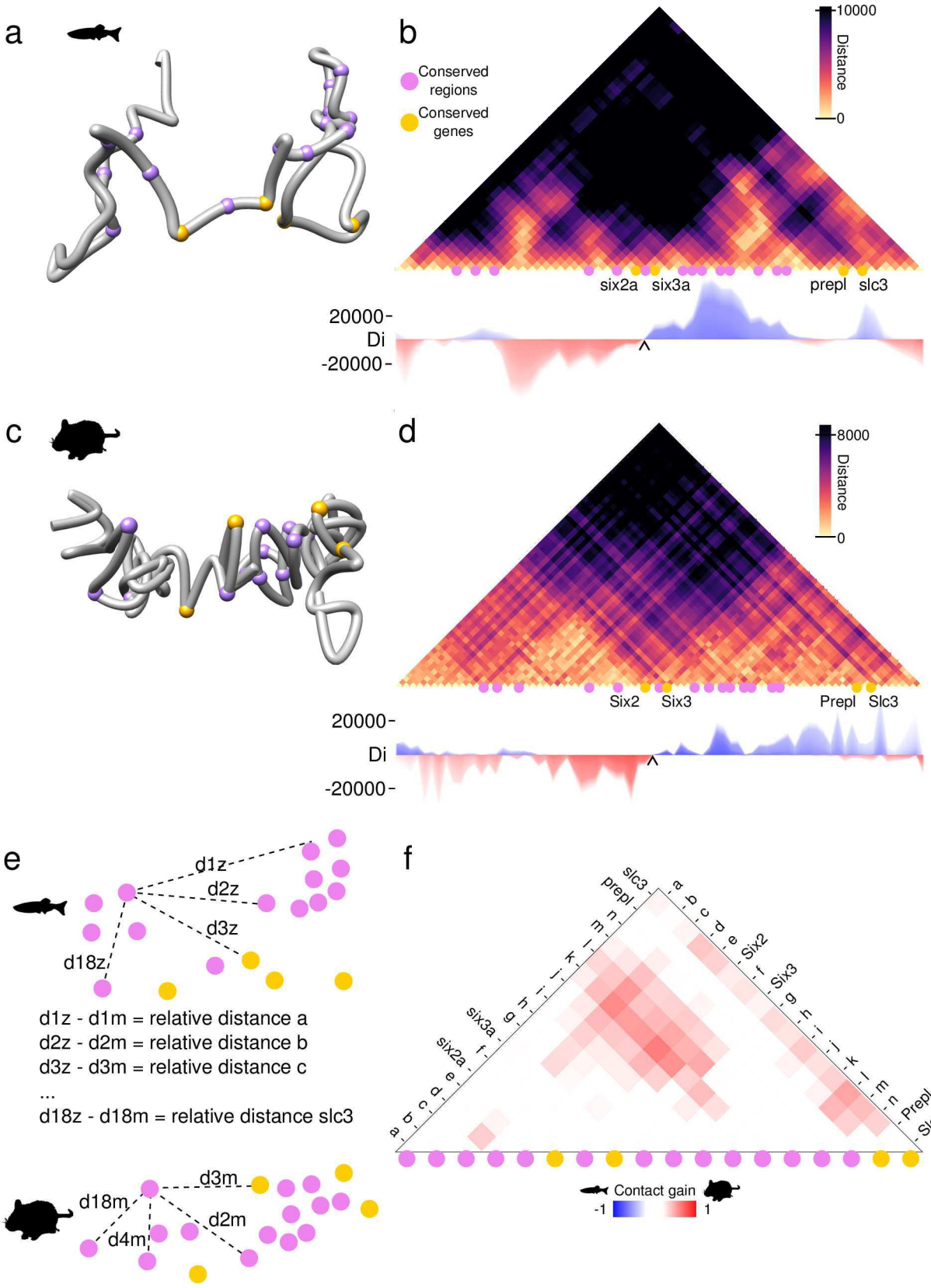

a
b
Conserved regions
Conserved genes
10000
Distance
0
six2a
six3a
prepl
slc3
20000
Di
-20000
c
d
8000
Distance
0
Six2
Six3
Prepl
Slc3
20000
Di
-20000
e
d1z
d2z
d3z
d18z
d1z - d1m = relative distance a
d2z - d2m = relative distance b
d3z - d3m = relative distance c
...
d18z - d18m = relative distance slc3
d3m
d18m
d2m
d4m
f
slc3
prepl
Six2
Six3
six3a
six2a
Prepl
Slc3
Contact gain
-1
1

Fig 4. Mouse and zebrafish *Six2-Six3* clusters have conserved 3D topologies.

(A) Representative 3D model of the *six2a-six3a* gene cluster in zebrafish. Orthologous regions conserved at the sequence level between the two species are depicted as purple beads, genes are indicated with yellow beads. (B) Virtual Hi-C of the zebrafish *six2a-six3a* cluster. Directionality index shows the TAD boundary, represented by a black arrow. (C, D) Representative 3D model of the same cluster and the Virtual Hi-C from mouse, labeled as in (A, B). (E) Changes in relative distances between conserved regions are obtained by subtracting the relative distance values of the two species. (F) Subtraction heat map of the distance changes obtained from (E). Red squares indicate shorter distances in mouse, blue squares indicate shorter distances in zebrafish.

We also generated 3D models of the mouse *Six2-Six3* locus using publicly available mouse Hi-C[4] converted into virtual 4C-seq like data (Figs 4C and 4D). From those 3D models, we derived a vHi-C that shows high correlation with the real Hi-C (Spearman rank correlation ρ = 0.86, S3B Fig), that provides further support to our method, in agreement with our previous observations[3].

In order to quantify the degree of structural similarity of mouse and zebrafish *Six2-Six3* clusters, we focused on a set of 18 regions that are conserved at the sequence level between the two species, comparing the distance heat maps corresponding to these regions (Figs 4E, 4F and S3 Table). The strong correlation observed between these two sets of distances (Spearman rank correlation ρ = 0.81) shows the high degree of topological conservation in the two species. Indeed, the two species have maintained very similar relative distances between these conserved regions, with an average change of just 20% (S6 Fig). Interestingly, the vast majority of distance changes were all in the same direction, decreasing their relative distances in mouse in comparison with zebrafish (red bins, Fig 4F). We hypothesize that the greater compaction in mouse helps compensate for the larger sequence length of this species, maintaining therefore similar 3D structural organizations in the two vertebrate lineages. Nevertheless, we are aware that the differences in the techniques used to model both loci could influence the final modeling.

Directionality index analysis[4] was also applied in these regions to call TAD boundaries (Fig 4B and 4D). A TAD boundary is found between the genes *Six2* and *Six3* in both species, supporting the conserved bipartite configuration of the clusters. Thus, our results show that the evolutionary conservation of gene expression in this cluster is not only due to the presence of conserved regulatory regions but also to a largely constrained 3D chromatin topology along the vertebrate lineage.

Discussion

Thinking about the chromatin as a 3D structure and trying to unravel its spatial organization have become necessary steps to properly understand genetic information in a functionally coherent manner. 3C-based methods can help to achieve this goal, but existing techniques provide different compromises between resolution, scope and costs and can therefore be difficult to implement from economical and technical points of view.

The tool presented here, 4Cin, can generate 3D models and derive vHi-C contact maps from a reduced number of 4C-seq datasets, uniting some of the specific advantages of different 3C techniques in a cost-effective manner. This makes 4Cin particularly useful for a broad range of single-locus studies dealing with multiple samples, conditions or species in which detailed 3D chromatin profiling was until now economically unfeasible. We have illustrated this with detailed examples showing the important biological implications and the multiple possibilities to test specific hypotheses that 4Cin can offer.

In order to generate reliable 3D models, various steps of the process have to be taken with additional caution: Data obtained from different species, tissues, time points or different experiments (like simulating virtual 4C-seq data from Hi-C data) should be carefully harmonized before integration with 4Cin, and, likewise, a proper normalization of these data has to be carried out (Check 4C-seq data processing in Methods section). In addition, the tool expects data to be derived from multiple cells and it is not optimized to be used with single-cell 3C-based experiments.

We believe that 4Cin will expand even further the use, interest and applications of chromatin capture techniques, helping a growing number of researchers to switch the way in which genomic information has been traditionally studied and generating new ideas, hypotheses and methods.

Methods

In this work, we refined and automated our previous algorithm[3] and provide novel scripts to ease the postprocessing analyses of the results and the discovery of biological novelties. This tool generates 3D chromatin models from 4C-seq data. The code is public and available at https://github.com/batxes/4Cin with a GNU GENERAL PUBLIC LICENSE. The usage of the pipeline (Fig 1 and S1 Fig) is also explained in the repository link. The 4Cin pipeline can be deployed pulling the docker image from https://hub.docker.com/r/batxes/4cin_ubuntu/ to avoid the installation of the dependencies. The input data and the final 3D models of all the regions studied in this work are also uploaded in github.

Our method uses the Integrative Modeling Platform (IMP)[26] and is based on a previous work[22]. he 3D models are composed of beads representing chromatin fragments and 4C-seq data is encoded as distance restraints between these beads. IMP tries to fulfill these restraints that are expressed in a single scoring function that the optimization algorithm attempts to minimize.

Chromatin representation

The chromatin is represented as a flexible chain of beads each bead representing a fixed number of consecutive DNA fragments, as previously described[3]. In the *six2a-six3a* locus in zebrafish, 33 DNA fragments are represented as one bead, while for the same region in mouse, each fragment corresponds to one bead, depending on the data resolution. Each bead comprising the Shh locus in mouse, both wild type and the inversion mutant, represents 100 fragments. The size of these beads is proportional to the length of the represented fragments. Assuming a canonical chromatin width of 30 nm (6-7 nucleosomes per 11 nm fiber length[27,28]), the radius (r_i) of these beads is defined as:

$$r_i = 0.0423 * l_i$$

where l_i is the length of the DNA fragments represented in each bead. Our *Six2-Six3* loci models in zebrafish and mouse are represented with 56 and 75 beads, that, at the same time, are representing a region of 1,12 and 1,48 Mbp. The Shh locus is 1,41 Mbp long and is represented by 71 beads.

4C-seq data processing

4C-seq data were analyzed as previously described[43]. Briefly, raw sequencing data were demultiplexed and aligned using mouse July 2007 assembly (mm9) or zebrafish July 2010 (danRer7) as the reference genomes using bowtie[44]. Reads located in fragments flanked by two restriction sites of the same enzyme, or in fragments smaller than 40 bp were filtered out. Mapped reads were then converted to reads-per-first-enzyme-fragment-end units, and smoothed using a 30 fragment mean running window algorithm. To be more consistent, the 4C-seq data corresponding to the Shh region was processed as in Symmons et. Al[33]. For the INV(6-C2) genotype, we mapped the 4C-seq data and did all the subsequent analyses using a custom version of the mouse genome that incorporates the corresponding genomic inversion at the previously described breakpoints[33].

4C-seq data normalization

4C-seq data consists of frequencies of interactions between the viewpoint DNA fragment and the rest of the locus. Our modeling protocol is based on trilateration, so we need at least four distances to locate a bead in 3D space: due to the fact that the 4C-seq method provides information between a DNA fragment and the rest of the fragments, we need at least four 4C-seq experiments to determine the position of a fragment. Each 4C-seq experiment is done in different population of cells and, therefore, the output of each experiment is likely to vary in the number of read counts. Hence, we first adjusted the measured values of each experiment to the same scale, multiplying each read count in each 4C-seq experiment by a factor so that we get the same read counts as the 4C-seq experiment with the biggest number of read counts. For the *Six2-Six3* locus, we used 5 experiments in zebrafish and 10 in mouse, while the Shh locus was modeled with four 4C-seq experiments in both the wild type and the inversion (S4 Table). Afterward, a Z-score is assigned to each bead. The Z-score indicates how many standard deviations separates a datum from the mean, identifying pair of (sets of) fragments (in this case, pair of beads) that interacts more or less than the average interaction frequency. To calculate the Z-score, the data needs to follow a normal distribution. The 4C-seq data does not follow a normal distribution, therefore, read counts are transformed by applying a $\log_{10}$ transformation to achieve a normally distributed data[22]. The Z-score is computed as follows:

The standard score of a raw score x is

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

where μ is the mean of the population and σ is the standard deviation of the population.

We set two thresholds called upper bound Z-score and lower bound Z-score. Contact frequencies that fall between both cutoffs are not used for the modeling as those interaction counts are more likely to happen by chance, since they don't fall in the tails of the normal distribution (S4 Fig). The optimal values for these thresholds were calculated empirically (see **Empirical determination of upper and lower Z-scores**).

Restraints and scoring function

As the chromatin fragments that they represent, consecutive beads were imposed to be connected by the application of harmonic upper bound distance restraints between consecutive beads. These distances are the sum of the radii of both consecutive beads.

We defined the “reach window” as previously described[3]. Briefly, the “reach window” of a 4C-seq experiment is the area between the furthest upstream and downstream fragments with a Z-score above the upper Z-score. Harmonic distance restraints were applied between beads corresponding to the viewpoints and the rest of the beads that were inside the reach window, as long as the Z-scores of the beads were not between the upper and lower Z-scores. Beads outside the reach window were restrained with harmonic lower bound distances. We set as weights the absolute values of the Z-scores of each bead, to give more importance to the beads with lowest and highest read counts.

The conversion from the read counts to the distance restraints is achieved by a linear relationship based on two assumptions: (I) the bead(s) with the maximum number of reads in each experiment will be imposed a harmonic distance restrain of 30 nm[27,28], (ii) the bead(s) with the minimum number of reads or zero reads, will be imposed a harmonic lower bound distance restrain equal to the maximum distance variable (see **Empirical determination of scale** and S4 Fig).

The sum of these restraints is the scoring function, a function that is minimized in each iteration. A scoring function of zero, means that all restraints are fulfilled, thus, this score represents the degree of consistency between the restraints and each 3D structure.

The IMP scoring function is defined as:

$$S(r_i,..,r_N)=\sum_{i=1}^{N-1} U_{conn}(r_i,r_{i+1})+\sum_{j=1}^{M}\left(\sum_{i\in\alpha} U_{irea}(r_i,r_j)+\sum_{i\in\beta} U_{orea}(r_i,r_j)\right)$$

r_i represents the 3D coordinate vector of each bead i, N is the total number of beads in the model, M is the list of assigned numbers to the viewpoint beads, α is the set of beads inside the reach window and β is the list of beads outside the reach window.

Chromatin connectivity restraints (U_{conn}) restrict the beads to be connected within a distance which is the sum of the radii of both consecutive beads with harmonic upper bound distances.

Inside reach window chromatin restraints (U_{irea}) impose harmonic distance restraints between beads inside this window and outside reach window chromatin restraints (U_{orea}) impose harmonic lower bound distances between the rest of beads.

Name	Restraint type	Functional form	distance(nm)	Bead i	weight(k)
U_{conn}	Chromatin connectivity restraint	Harmonic upper bound distance restraint	R_i+R_{i+1}	i = {1..N-1}	1

U_{irea}	Inside reach window chromatin restraint	Harmonic distance restraint	d_{irea}	Any i ∈ β	|z-score|
U_{orea}	outside reach window chromatin restraint	Harmonic lower bound distance restraint	d_{orea}	Any i ∈ α	|z-score|

Empirical determination of scale

Beads containing the lowest number of reads in each experiment will be located at the maximum distance away from the viewpoint. This distance in calculated empirically as follows. Models are generated varying this maximum distance in steps of 1000 (by default) and keeping the upper and lower Z-scores low, in order to take into account most of the distance restraints. Afterwards, the mean length of the models is calculated, summing the distance between consecutive beads in each model. Then, the theoretical length of the chromatin of the region we are modeling is calculated assigning a theoretical length of 0.846nm to each of the nucleotides[27,28]. The maximum distance of the models with a length closer to the theoretical optimum will be set for the final modeling. In the mouse and zebrafish six loci, 11000Å and 13000Å were set as maximum distances. On the other hand, the maximum distance in the Shh locus was of 10000Å for the wild type and 7000Å for the inverted region.

Empirical determination of upper and lower Z-scores

A similar approach as in “Empirical determination of scale” is used to set the upper and lower Z-score parameters. In this case, models are generated with the previously obtained maximum distance fixed, varying the upper and lower Z-score parameters, in bins of 0.1 by default. Then, the distance between the viewpoint beads and the rest of the beads is measured and the mean of these distances is obtained from all the generated models to compare with the raw 4C-seq data (S7 Fig). The Z-scores of the set of models that has the best correlates with the raw data is used in the final modeling. For the zebrafish six locus 0,1 and -0,1 were set as the uZ and lZ, while 0,2 and -0,1 were set for mouse in the same region. Likewise, 0,1 and -0,1 was set in the Shh region for the wild type, and 0.2 and -0.1 were the values for the inverted region.

Optimization

With the maximum distance and upper and lower Z-scores fixed, models are generated starting from entirely random set of positions of the beads. The number of models should be big enough in order to sample the space of solutions thoroughly and allow a reliable analysis afterwards. In this work, 50.000 models were generated in all 4 examples. We have seen that generating less than 10.000 models can lead to very

variable models (S8 Fig). The modeling is carried out with IMP, optimizing the scoring function. The algorithm combines a Monte Carlo exploration with steps of local optimization and simulated annealing. The optimization ends when the score difference between the rounds is below 0.00001 or when the score reaches 0.

Analysis and clustering of models

From the whole population of models, the models with the best score are selected as long as they fulfill most of the distance restraints. The standard deviation and the percentage of restraints fulfillment is used to filter out unreliable models. The method starts from a very low distance and high restraints fulfillment percentage and does many analysis iterations loosening up these cut-offs until 200 models can be retrieved. We selected the 200 models with best score as long as they fulfilled 85% of the restraints for zebrafish and mouse, with an std-dev of 2000Å and 2250Å as a limit of restraint fulfillment. We also selected 200 models with best score for both wild type and inverted Shh locus that fulfilled 85% of the restraints with an std-dev of 1000Å and 960Å respectively.

Then, these models are clustered according to their similarity measured by the Root Mean Square Deviation (RMSD). The goal of this step is to identify mirror image models since we don't have information to discriminate the mirror images. The set of models for both zebrafish and mouse and the wild type and inverted Shh locus were clustered showing two mirror image clusters (S7 Fig).

The number of clusters depends on: 1) the quality of the modeling, but also 2) the structural variability of the genome locus. The high number of cluster could indicate high structural variability, meaning the there is no enough data to filter between them or that the quality if the data is not good.

Representative and superposition of models

4Cin selects the biggest cluster of models for next analyses. From the models in the selected cluster, the most similar model to the average of all the models is used as the representative model (Fig. 3a). The superposition of all the set of final models is shown also to see the variability between them (S9 Fig). In addition, the variability of the beads between the models in the biggest cluster is shown.

Virtual Hi-C generation and comparisons

A contact map is generated resembling a Hi-C heat map plot, that we called virtual Hi-C (vHi-C). For this, the average distance between pair of beads from the best models is calculated (Fig. 2b, 3b). vHi-C's of the

wild type and inverted Shh region were compared and a subtraction of both virtual Hi-C's was done, repositioning the inverted region as in the wild type, in order to compare the change in contacts of each bead. The heat map shows in blue the contacts that were lost in the inversion, and in red the contacts that were gained (Fig 2C). To compare the six loci in both species quantitatively, we measured the distance between beads that represent conserved regions in both mouse and zebrafish (Fig 4, S3 Table). In our models, each bead represents almost the same amount of nucleotides, 20Kbp, making the comparison of conserved regions more reliable.

Generation of the 4C-seq mouse data form Hi-C data

The virtual 4C-seq data representing the viewpoints containing the *six2 and six3* genes and the data of 8 other scattered viewpoints was extracted from the original Hi-C data from Dixon et. al.[4] (Supplementary Table 4). This data was used as 10 single 4C-seq experiments to generate the 3D models and the virtual Hi-C heat map plot.

Directionality index

Directionality index (DI) was calculated in all the vHi-Cs as in ref 3 with slight changes. The DI for the beads at the edges of the vHi-Cs, was calculated by assigning the mean of all the values in the vHi-C for the heat map squares that are not represented in the vHi-C. We also calculated the DI iteratively, ranging the TAD size from 1 (size of TAD = 1 bead) to the total number of beads of the model (size of TAD = N). Afterwards, we overlaped all the DIs to generate the plot (Step 5 in Fig. 1, Fig. 2B, 2D, 4B and 4D) and give a list of all the TAD boundaries called, sorted by the number of times that they were called in each iterations.

Genome Painting

CTCF Chip-seq data used in the Shh region was adquired from Econde https://www.encodeproject.org/experiments/ENCSR000CEB/ and painted in the representative model (mm9 data) with a black-to-white gradient, from high to low score.

3D models manipulation and surface calculation around enhancers

4Cin generates 3D models that can be opened and modified in UCSF Chimera (https://www.cgl.ucsf.edu/chimera/). The molmap command of UCSF Chimera was used to generate a mesh surface of 75 nm in radius around the Shh region enhancers in Fig. 3.

Determination of conserved regions between zebrafish and mouse

To define genomic regions conserved at the sequence level between mouse and zebrafish *Six2-Six3* clusters, we downloaded the corresponding chained alignments available in UCSC. We then manually inspected and curated these aligned regions to verify that their locations and orientations were equivalent in the two species and that they corresponded to bona-fide conserved sequences. Genomic coordinates are provided in S3 Table.

CTCF directionality calculation

Clover (https://zlab.bu.edu/clover/) was used in the Shh region to predict CTCF binding sites and their orientation (S1 Table), using a mouse CTCF position weight matrix (http://cisbp.ccbr.utoronto.ca/TFreport.php?searchTF=T049038_1.02) and setting a threshold of 7.

Mapping of insertion sensors and generation of the enhancer contact area

To map these positions with high accuracy, we generated 3D models of the Shh locus with a fivefold higher resolution. We then selected the best models and used the most representative model to map the enhancers and insertion sensors. The measurements between the enhancers and the sensors were carried out taking into account the whole population of best 3D models.

4C-seq data Down-sampling and erroneous data insertion

Bed files that were mapped in the zebrafish genome (danrer10) were shuffled and then, the first 20%, 40%, 60%, 80%, 90%, and 95% lines of the files were removed to get the down-sampled data. Afterwards, the same procedure as in **4C-seq data processing** was followed. To generate a set of erroneous data, we calculated the value representing the percentile 95 of the data for each experiment, and switched with random read counts of pair of fragments.

Data Availability

Data and Source Code are available at the following Github Repository: https://github.com/batxes/4Cin

References:

1. Nora EP, Dekker J, Heard E. Segmental folding of chromosomes: A basis for structural and regulatory chromosomal neighborhoods? BioEssays. 2013;35: 818–828. doi:10.1002/bies.201300040

2. de Laat W, Duboule D. Topology of mammalian developmental enhancers and their regulatory landscapes. Nature. 2013;502: 499–506. doi:10.1038/nature12753

3. Acemel RD, Tena JJ, Irastorza-Azcarate I, Marlétaz F, Gómez-Marín C, de la Calle-Mustienes E, et al. A single three-dimensional chromatin compartment in amphioxus indicates a stepwise evolution of vertebrate Hox bimodal regulation. Nat Genet. 2016;48: 336–341. doi:10.1038/ng.3497

4. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. Nature Publishing Group; 2012;485: 376–380. doi:10.1038/nature11082

5. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature. 2012;485: 381–385. doi:10.1038/nature11049

6. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell. 2012. pp. 458–472. doi:10.1016/j.cell.2012.01.010

7. Hou C, Li L, Qin ZS, Corces VG. Gene Density, Transcription, and Insulators Contribute to the Partition of the Drosophila Genome into Physical Domains. Mol Cell. 2012;48: 471–484. doi:10.1016/j.molcel.2012.08.031

8. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell. Elsevier Inc.; 2014;159: 1665–1680. doi:10.1016/j.cell.2014.11.021

9. Andrey G, Montavon T, Mascrez B, Gonzalez F, Noordermeer D, Leleu M, et al. A Switch Between Topological Domains Underlies HoxD Genes Collinearity in Mouse Limbs. Science. 2013;340: 1234167–1234167. doi:10.1126/science.1234167

10. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell. 2015;161: 1012–1025. doi:10.1016/j.cell.2015.04.004

11. Hnisz D, Weintraub AS, Day DS, Valton A, Bak RO, Li CH, et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science. 2016;351: 1454–1458. doi:10.1126/science.aad9024

12. Franke M, Ibrahim DM, Andrey G, Schwarzer W, Heinrich V, Schöpflin R, et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature. Nature Publishing Group; 2016;538: 265–269. doi:10.1038/nature19800

13. Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, et al. Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. Cell Rep. The Authors; 2015;10: 1297–1309. doi:10.1016/j.celrep.2015.02.004

14. Gómez-Marín C, Tena JJ, Acemel RD, López-Mayorga M, Naranjo S, de la Calle-Mustienes E, et al. Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. Proc Natl Acad Sci. 2015;112: 7542–7547. doi:10.1073/pnas.1505463112

15. Davies JOJ, Oudelaar AM, Higgs DR, Hughes JR. How best to identify chromosomal interactions: a comparison of approaches. Nat Methods. Nature Publishing Group; 2017;14: 125–134. doi:10.1038/nmeth.4146

16. de Wit E, de Laat W. A decade of 3C technologies: Insights into nuclear organization. Genes Dev. 2012;26: 11–24. doi:10.1101/gad.179804.111

17. Davies JOJ, Telenius JM, McGowan SJ, Roberts NA, Taylor S, Higgs DR, et al. Multiplexed analysis of chromosome conformation at vastly improved sensitivity. Nat Methods. Nature Publishing Group; 2015;13: 74–80. doi:10.1038/nmeth.3664

18. Serra F, Di Stefano M, Spill YG, Cuartero Y, Goodstadt M, Baù D, et al. Restraint-based three-dimensional modeling of genomes and genomic domains. FEBS Lett. 2015;589: 2987–2995. doi:10.1016/j.febslet.2015.05.012

19. Adhikari B, Trieu T, Cheng J. Chromosome3D: reconstructing three-dimensional chromosomal structures from Hi-C interaction frequency data using distance geometry simulated annealing. BMC Genomics. BMC Genomics; 2016;17: 886. doi:10.1186/s12864-016-3210-4

20. Trieu T, Cheng J. 3D genome structure modeling by Lorentzian objective function. Nucleic Acids Res. 2017;45: 1049–1058. doi:10.1093/nar/gkw1155

21. Szalaj P, Michalski PJ, Wróblewski P, Tang Z, Kadlof M, Mazzocco G, et al. 3D-GNOME: an integrated web service for structural modeling of the 3D genome. Nucleic Acids Res. 2016;44: W288–W293. doi:10.1093/nar/gkw437

22. Baù D, Sanyal A, Lajoie BR, Capriotti E, Byron M, Lawrence JB, et al. The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. Nat Struct Mol Biol. Nature Publishing Group; 2011;18: 107–114. doi:10.1038/nsmb.1936

23. Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat Biotechnol. Nature Publishing Group; 2011;30: 90–98. doi:10.1038/nbt.2057

24. Carstens S, Nilges M, Habeck M. Inferential Structure Determination of Chromosomes from Single-Cell Hi-C Data. PLoS Computational Biology. 2016. doi:10.1371/journal.pcbi.1005292

25. Gong K, Tjong H, Zhou XJ, Alber F. Comparative 3D genome structure analysis of the fission and the budding yeast. PLoS One. 2015;10: e0119672. doi:10.1371/journal.pone.0119672

26. Russel D, Lasker K, Webb B, Velázquez-Muriel J, Tjioe E, Schneidman-Duhovny D, et al. Putting the pieces together: Integrative modeling platform software for structure determination of macromolecular assemblies. PLoS Biol. 2012;10: e1001244. doi:10.1371/journal.pbio.1001244

27. Tjong H, Gong K, Chen L, Alber F. Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. Genome Res. 2012;22: 1295–1305. doi:10.1101/gr.129437.111

28. Bystricky K, Heun P, Gehlen L, Langowski J, Gasser SM. Long-range compaction and flexibility of interphase chromatin in budding yeast analyzed by high-resolution imaging techniques. Proc Natl Acad Sci U S A. 2004;101: 16495–500. doi:10.1073/pnas.0402766101

29. Trussart M, Serra F, Bau D, Junier I, Serrano L, Marti-Renom M a., et al. Assessing the limits of restraint-based 3D modeling of genomes and genomic domains. Nucleic Acids Res. 2015;43: 3465–3477. doi:10.1093/nar/gkv221

30. Consortium EP, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis C a, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489: 57–74. doi:10.1038/nature11247

31. Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, et al. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. Cell. Elsevier Inc.; 2015;163: 1611–1627. doi:10.1016/j.cell.2015.11.024

32. Phillips-Cremins JE, Sauria MEG, Sanyal A, Gerasimova TI, Lajoie BR, Bell JSK, et al. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell. 2013;153: 1281–1295. doi:10.1016/j.cell.2013.04.053

33. Symmons O, Pan L, Remeseiro S, Aktas T, Klein F, Huber W, et al. The Shh Topological Domain Facilitates the Action of Remote Enhancers by Reducing the Effects of Genomic Distances. Dev Cell. 2016;39: 529–543. doi:10.1016/j.devcel.2016.10.015

34. Symmons O, Uslu VV, Tsujimura T, Ruf S, Nassari S, Schwarzer W, et al. Functional and topological characteristics of mammalian regulatory domains. Genome Res. 2014;24: 390–400. doi:10.1101/gr.163519.113

35. Bickmore WA, Benabdallah NS, Gautier P, Hekimoglu-Balkan B, Lettice LA, Bhatia S. SBE6: a novel long-range enhancer involved in driving sonic hedgehog expression in neural progenitor cells. Open Biol. 2016;6: 1–11. doi:10.1098/rsob.160197

36. Tsukiji N, Amano T, Shiroishi T. A novel regulatory element for Shh expression in the lung and gut of mouse embryos. Mech Dev. 2014;131: 127–136. doi:10.1016/j.mod.2013.09.003

37. Yao Y, Minor PJ, Zhao Y-T, Jeong Y, Pani AM, King AN, et al. Cis-regulatory architecture of a brain signaling center predates the origin of chordates. Nat Genet. 2016;48: 575–580. doi:10.1038/ng.3542

38. Jeong Y. A functional screen for sonic hedgehog regulatory elements across a 1 Mb interval identifies long-range ventral forebrain enhancers. Development. 2006;133: 761–772. doi:10.1242/dev.02239

39. Epstein DJ, McMahon a P, Joyner a L. Regionalization of Sonic hedgehog transcription along the anteroposterior axis of the mouse central nervous system is regulated by Hnf3-dependent and -independent mechanisms. Development. 1999;126: 281–292. Available: http://dev.biologists.org/content/develop/126/2/281.full.pdf

40. Sagai T, Amano T, Tamura M, Mizushina Y, Sumiyama K, Shiroishi T. A cluster of three long-range enhancers directs regional Shh expression in the epithelial linings. Development. 2009;136: 1665–1674. doi:10.1242/dev.032714

41. Chen C-K, Symmons O, Uslu VV, Tsujimura T, Ruf S, Smedley D, et al. TRACER: a resource to study the regulatory architecture of the mouse genome. BMC Genomics. 2013;14: 215. doi:10.1186/1471-2164-14-215

42. Ferrier DEK. Evolution of Homeobox Gene Clusters in Animals: The Giga-Cluster and Primary vs. Secondary Clustering. Front Ecol Evol. 2016;4: 1–13. doi:10.3389/fevo.2016.00036

43. Noordermeer D, Leleu M, Splinter E, Rougemont J, De Laat W, Duboule D. The Dynamic Architecture of Hox Gene Clusters. Science. 2011;334: 222–225. doi:10.1126/science.1207194

44. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10: R25. doi:10.1186/gb-2009-10-3-r25

Supporting information

S1 Fig. Schematic explanation of 4Cin. (Pipeline) First, the maximum distance, the upper bound Z-score (uZ) and the lower bound Z-score (lZ) are calculated so these parameters are afterwards used in the final modeling. Then, these models are subjected to an analysis to retrieve the best models and clustered based on their RMSD to distinguish between mirror image models. Best models can also be super imposed, to see structural variability. The representative model can be colored depending on genetic or epigenetic data. Finally, best models are used to generate a virtual Hi-C (vHi-C). Additionally, TAD boundaries can be called in the vHi-Cs and other vHi-Cs can be compared using the scripts provided with the pipeline. (Modeling) The modeling process first encodes the 4C-seq data into restraints. Distance restraints are also used to connect beads. These restraints and the representation of the chromatin fragments as beads are taken into account in the optimization process to generate a single model.

S2 Fig. Correlation with down-sampled and erroneous 4C data. Pearson's and Spearman's correlation between the vHi-C derived from six2a-six3a zebrafish locus models and the vHi-C's of the same locus down-sampling and inserting errors in the 4C data.

S3 Fig. Mouse Six2-Six3 cluster topology comparison. (a) Hi-C of the Six2-Six3 cluster (Gene Expression Omnibus (GEO) accession number GSM862722). (b) vHi-C of the Six2-Six3 cluster. (c,d,e) vHi-Cs of the Six2-Six3 cluster generated with different viewpoints. (f) Spearman's correlation between the Hi-C (a) and the vHi-Cs (b,c,d,e).

S4 Fig. Output plot generated by the data_manager.py script. Example corresponds to the Six3 viewpoint in mouse. Top, 4C-seq read counts by bead. Red bar shows the viewpoint. Middle, Z-scores in red corresponding to the read counts from the top panel. Horizontal blue lines indicate the upper bound Z-score and lower bound Z-score. Bottom, Distance restraints encoded from the read counts in the top panel.

S5 Fig. Shh-TAD boundaries are enriched in CTCF sites. (a) vHi-C of the Shh WT region on top. CTCF Chip-seq data corresponding to the region colored in white-to-black gradient, white for low reads, black for high reads. CTCF sites with highest reads are depicted with oriented triangles. (b) Shh WT representative model colored as in panel (a). Yellow beads represent Shh-TAD boundaries. Shh-TAD is encircled in yellow-black. (c and d) vHi-C, CTCF Chip-seq data and representative model depicted as in (a) and (b).

S6 Fig. Zebrafish and mouse topology comparison. Subtraction heat map of the distance changes between conserved regions in the Six2-Six3 cluster in zebrafish and mouse as explained in Fig. 2e. Top triangle

corresponds to zebrafish data and bottom triangle to mouse data. Red squares indicate shorter distances in mouse, blue shorter distances in zebrafish.

S7 Fig. Analysis and clustering of 3D models. (a, c, e, g) Heat maps comparing the raw 4C-seq data and the mean distances between beads of the models with the best parameters (upper bound Z-score, lower bound Z-score and max distance): Shh WT region: 0.1, -0.1, 11000; Shh inverted region: 0.2, -0.1, 10000; Six2-Six3 cluster in zebrafish: 0.1, -0.1, 13000 and in mouse: 0.2, -0.1, 11000. (b, d, f, h) Heat maps showing 2 clusters in the Shh WT and inverted region and the Six2-Six3 cluster in zebrafish and mouse. The clustering was performed based on the RMSD of the 3D models.

S8 Fig. Variability depending on number of models. Average variability of the 3D models depending on the sampling.

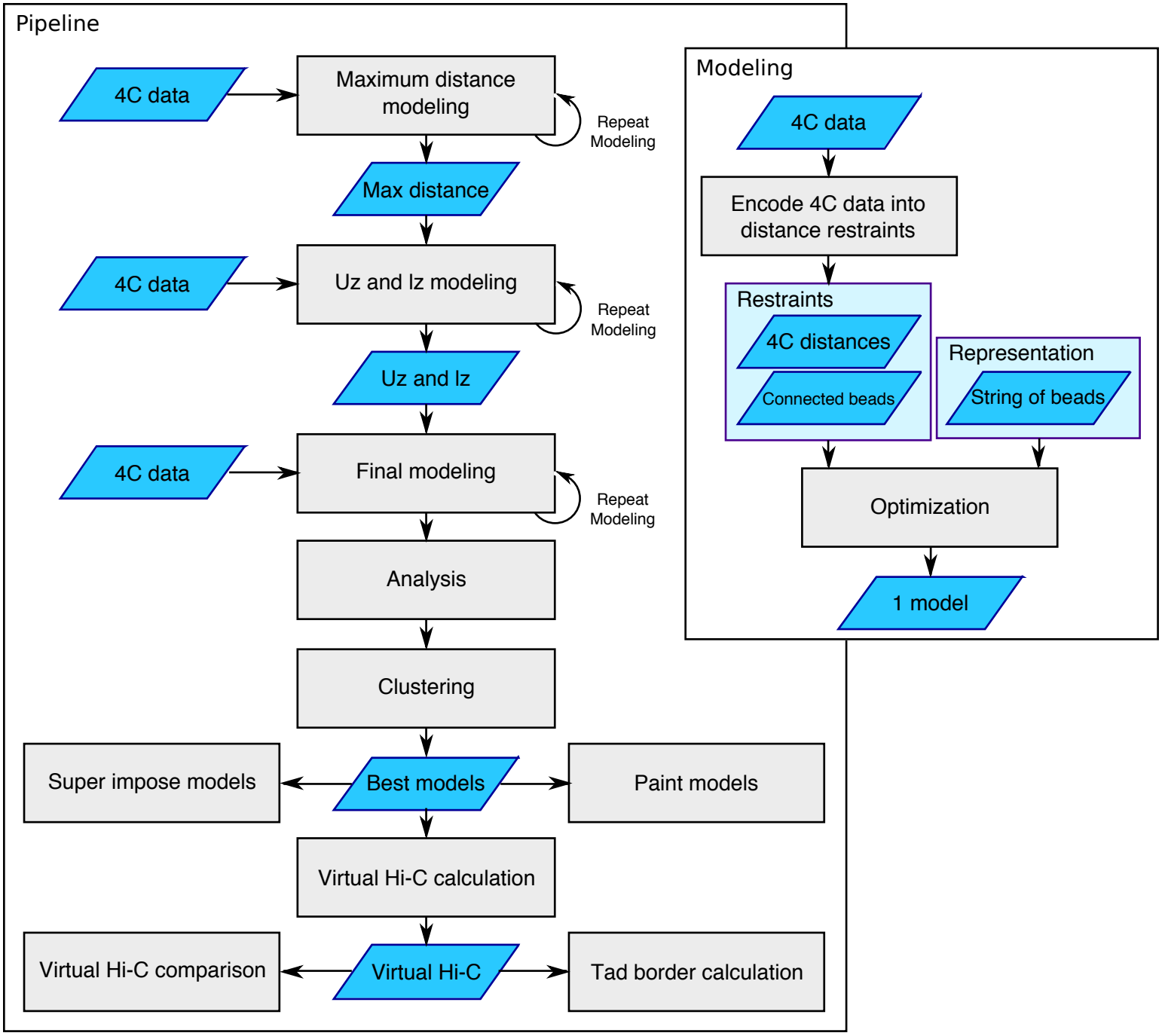
S9 Fig. Superposition of 3D models and their variability. Superposition of 3D models of the biggest cluster after clustering the best models of the Shh WT region and variability of each bead in the cluster showed in standard deviation devided by their maximum distance to show them at scale. (a), Shh inverted region (b), Six2-Six3 cluster in zebrafish (c) and mouse (d).

S1 Table. Viewpoints used. Viewpoints used in the generation of 3D models.

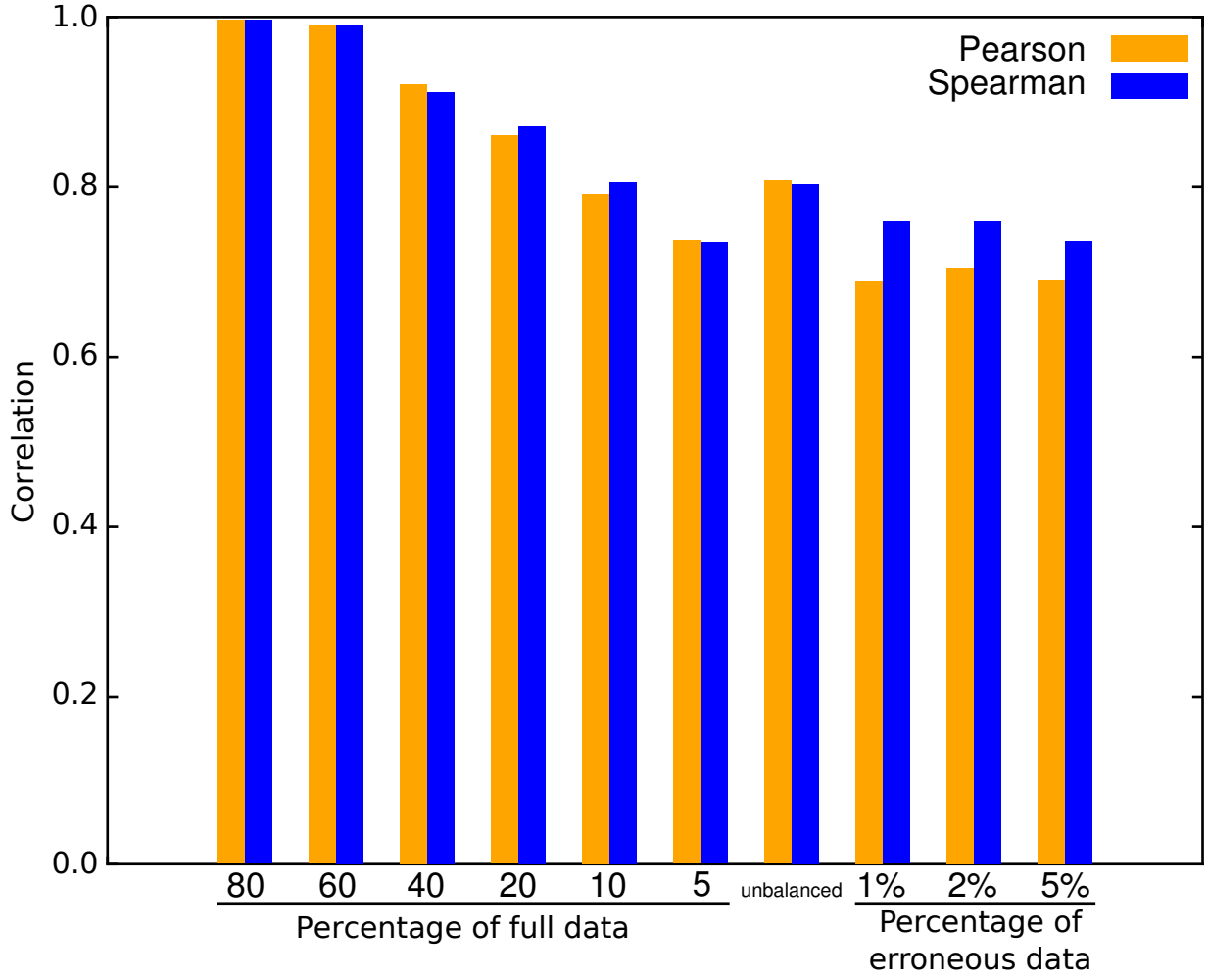
S2 Table. Conserved regions in Six2-Six3. Conserved regions and genes between Zebrafish and Mouse in the Six2-Six3 region

S3 Table. CTCF locations. Location and sign of CTCF binding sites

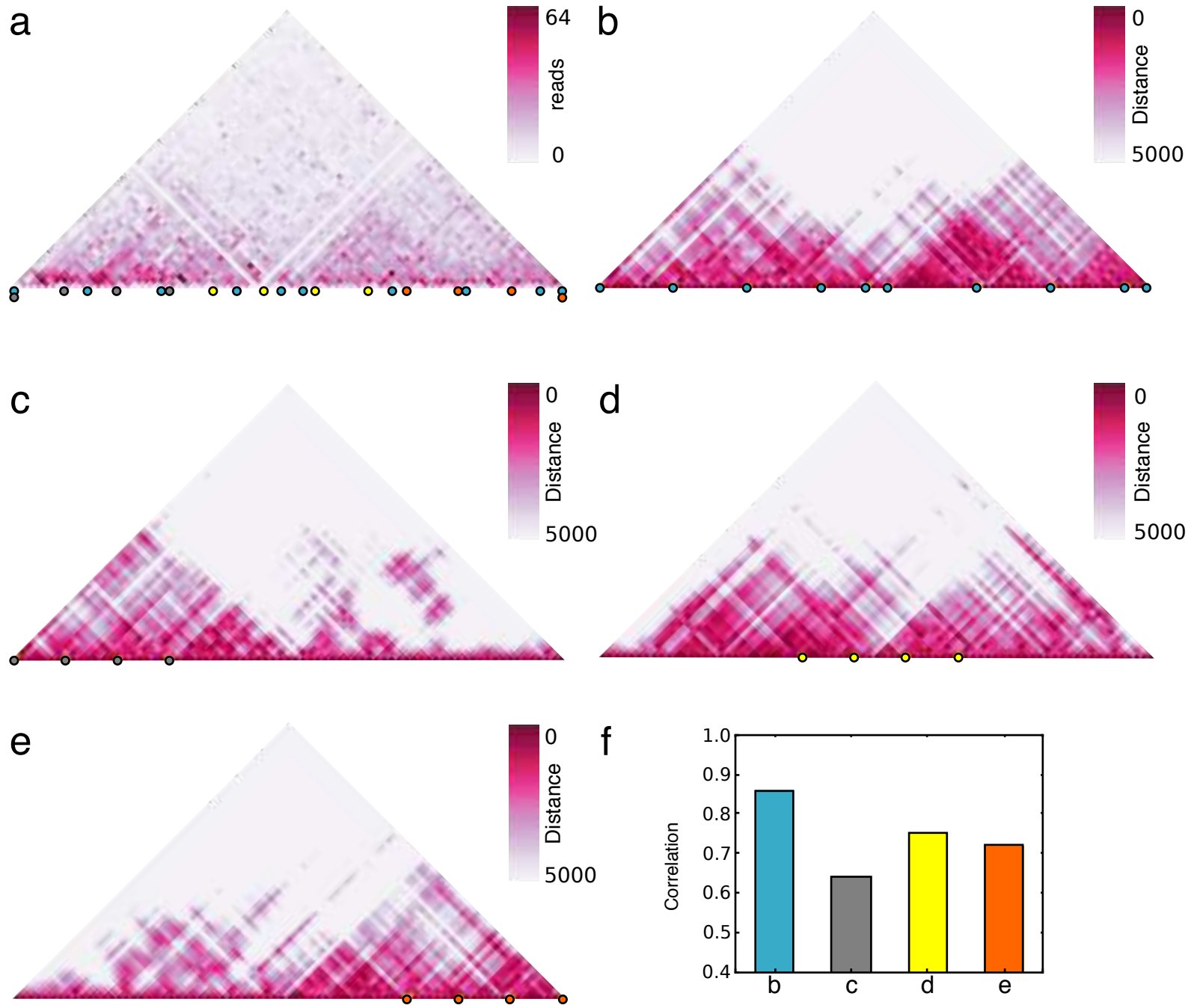
S4 Table. Sensor probes and enhancers. Sensor probes and enhancers in the Shh region.



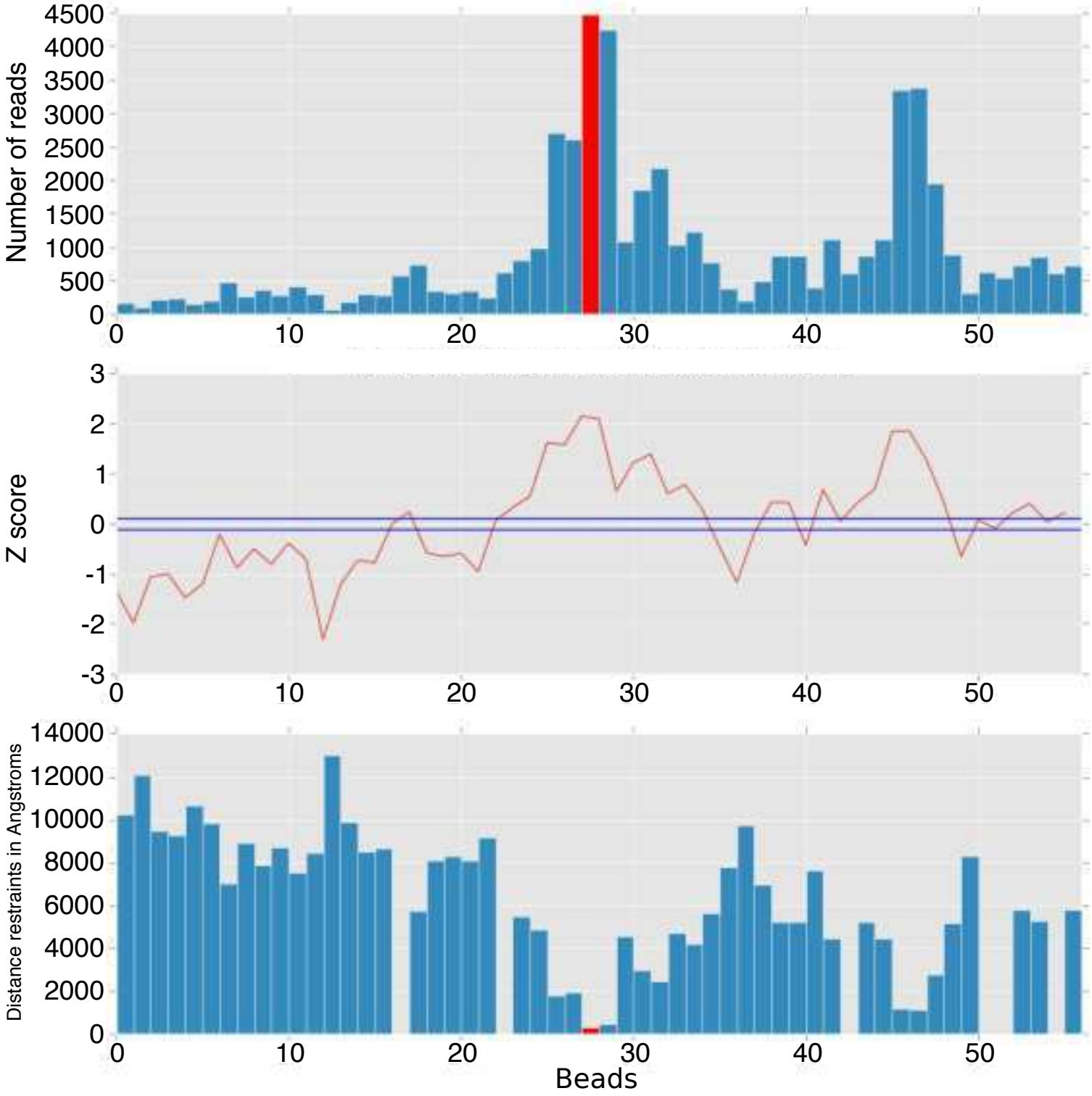
Supplementary Figure 1



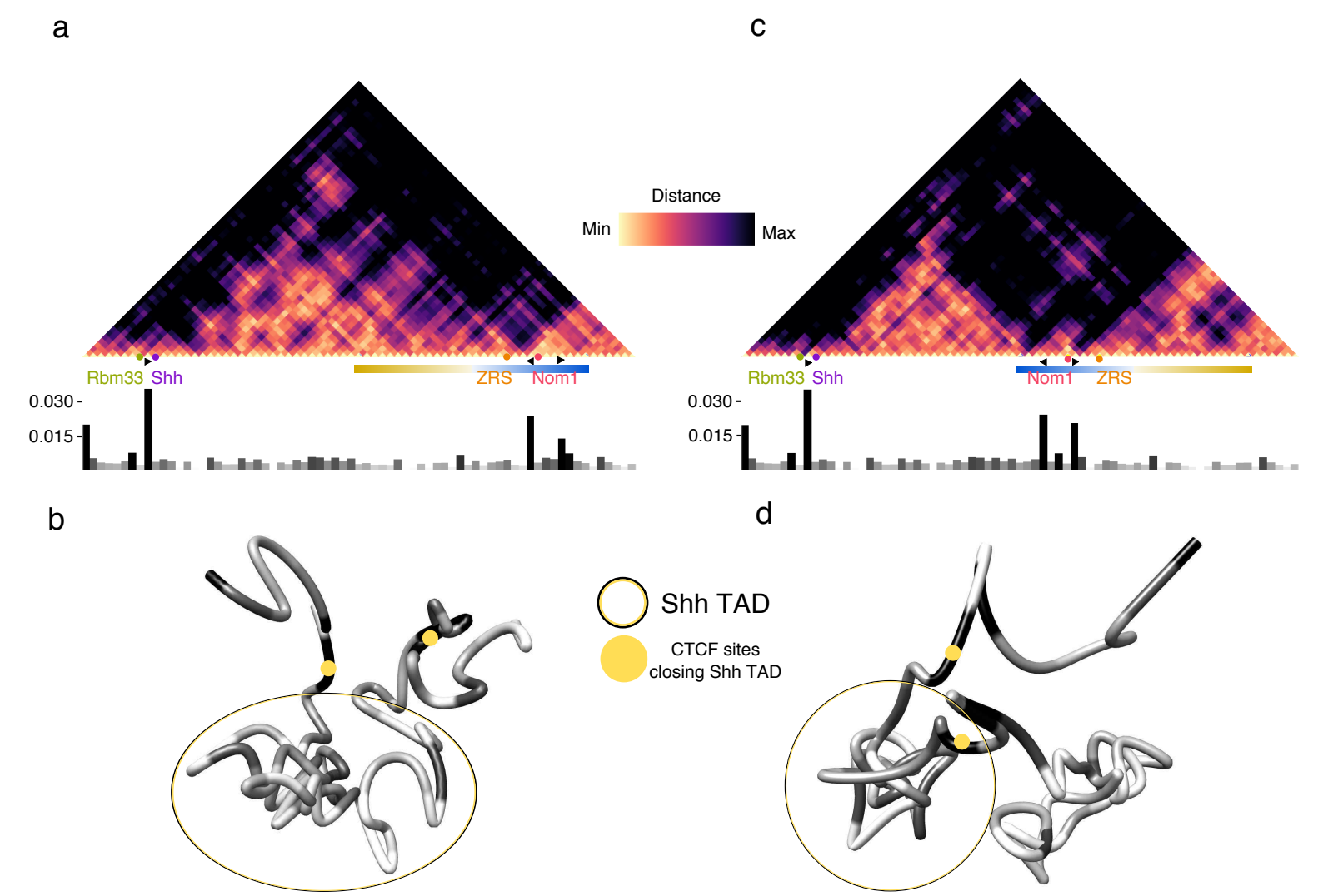
Supplementary Figure 2



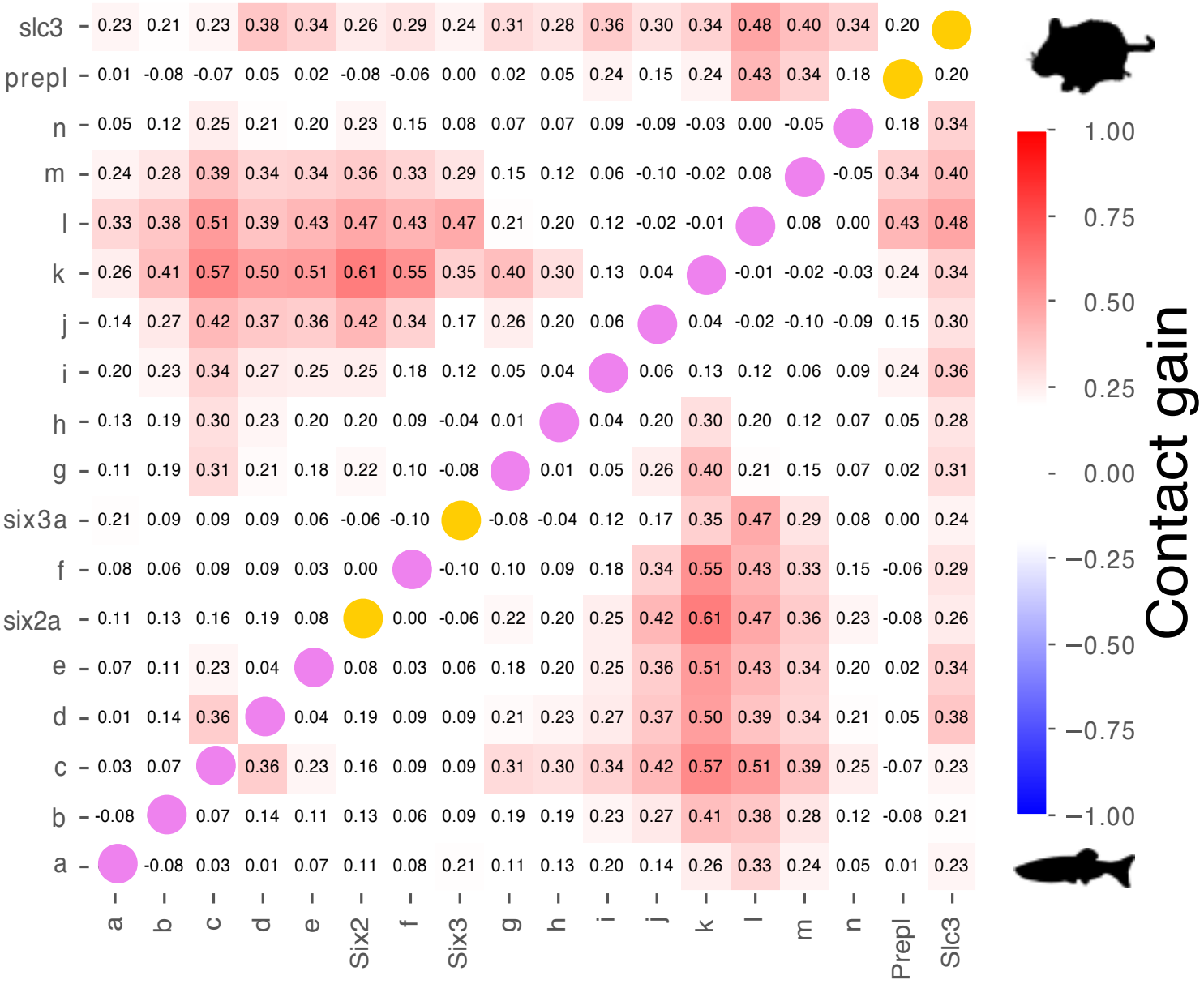
Supplementary Figure 3



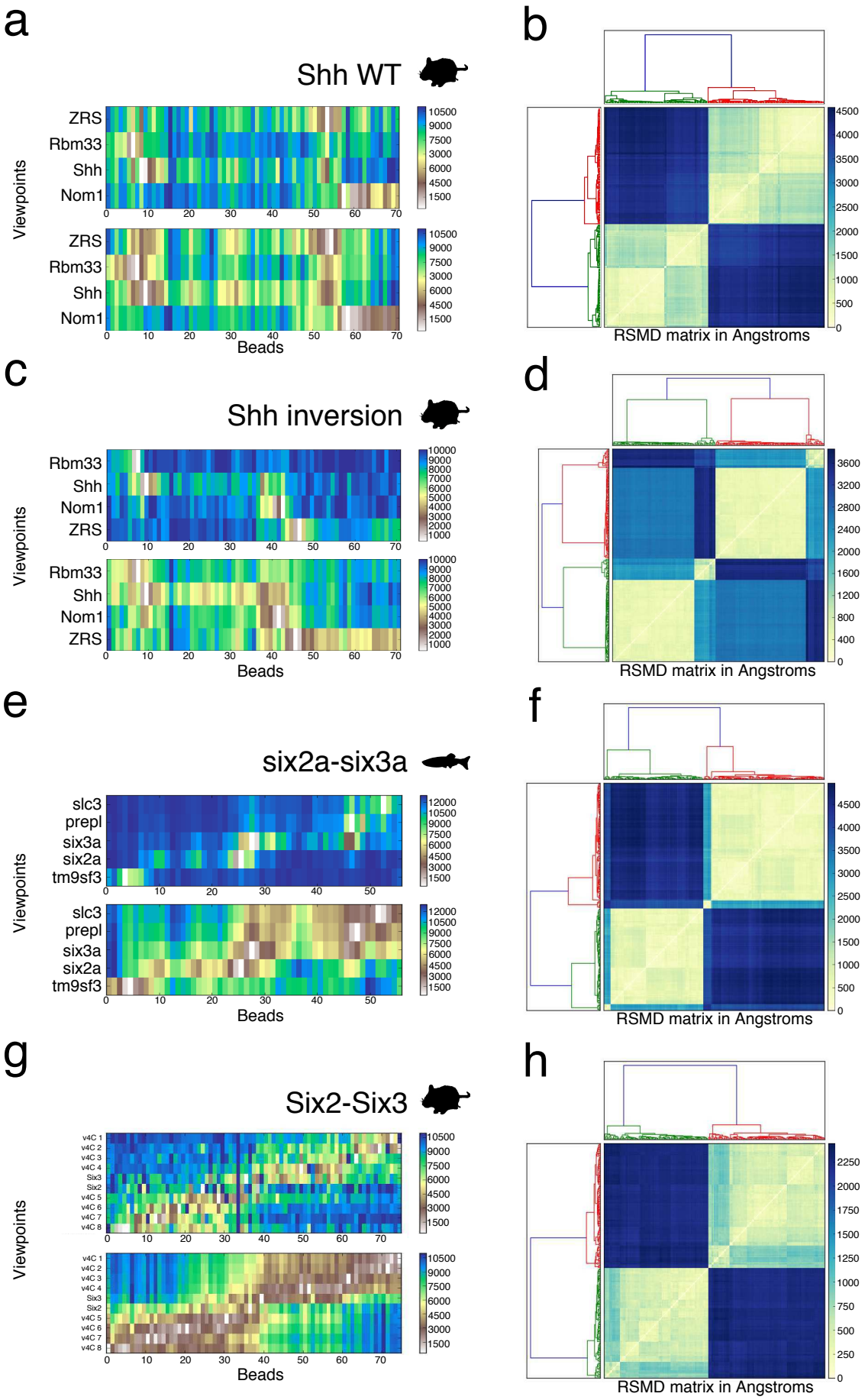
Supplementary Figure 4

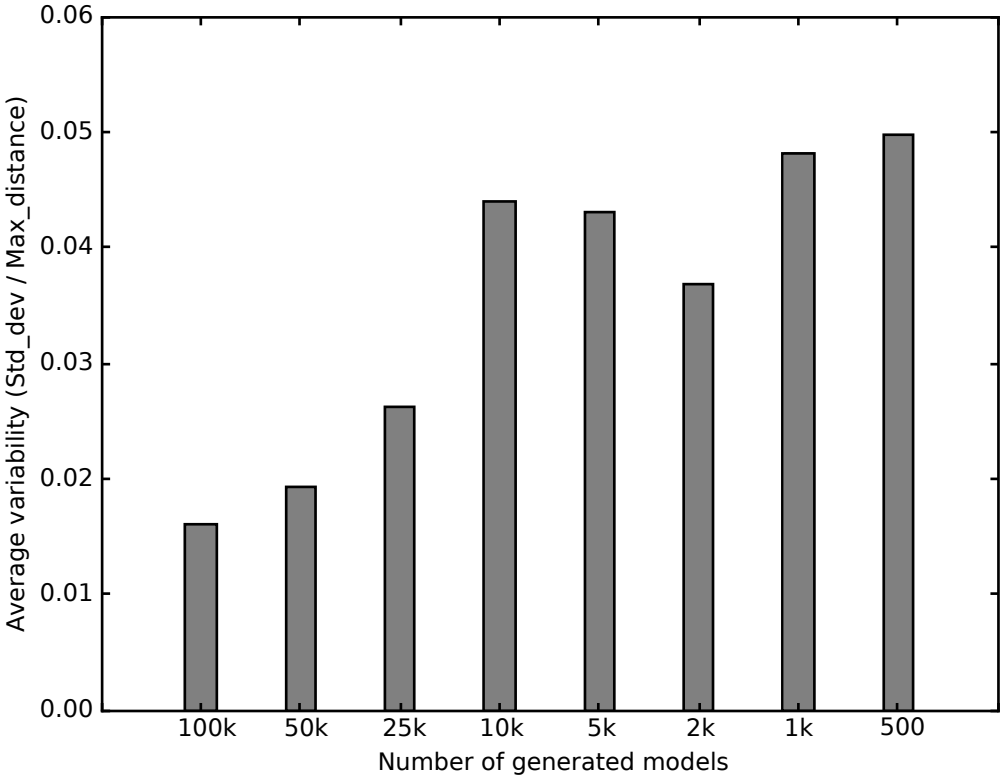

a
c
Distance
Min
Max
Rbm33
Shh
ZRS
Nom1
0.030
0.015
Rbm33
Shh
Nom1
ZRS
0.030
0.015

b
d
Shh TAD
CTCF sites
closing Shh TAD



Supplementary Figure 6

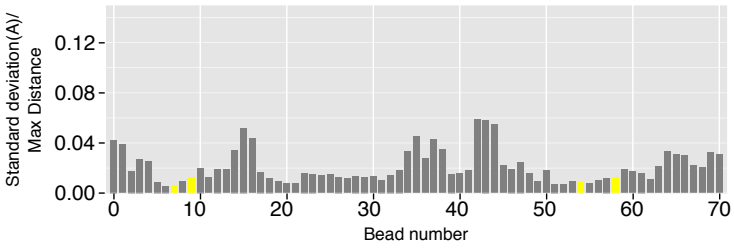

a
Shh WT
ZRS
Rbm33
Shh
Nom1
Viewpoints
Beads
b
RSMD matrix in Angstroms
c
Shh inversion
d
e
six2a-six3a
slc3
prepl
six3a
six2a
tm9sf3
f
g
Six2-Six3
v4C 1
v4C 2
v4C 3
v4C 4
Six3
Six2
v4C 5
v4C 6
v4C 7
v4C 8
h



Supplementary Figure 8

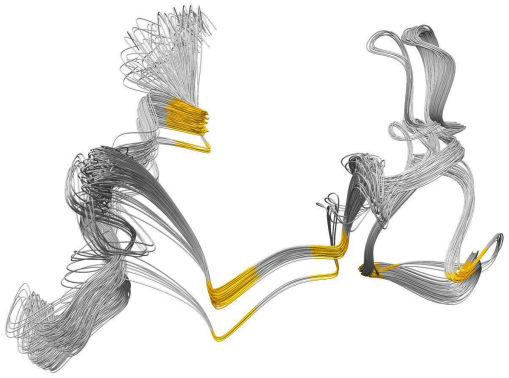
a

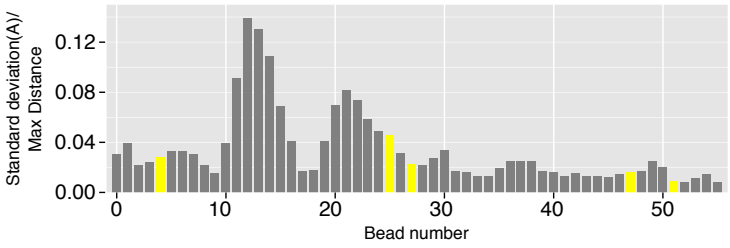
Shh WT



c

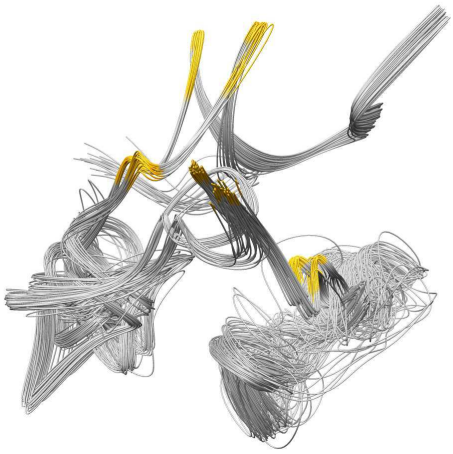
six2a/six3a

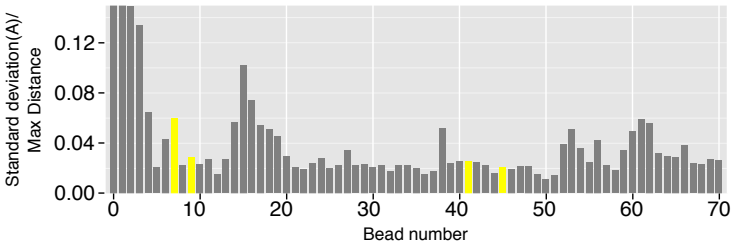




b

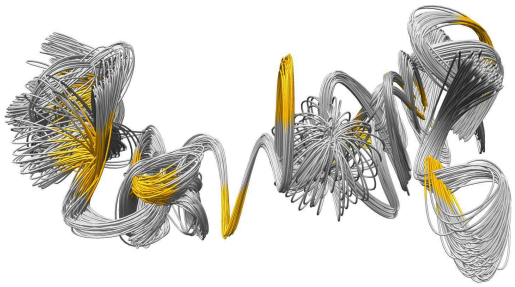
Shh inverted

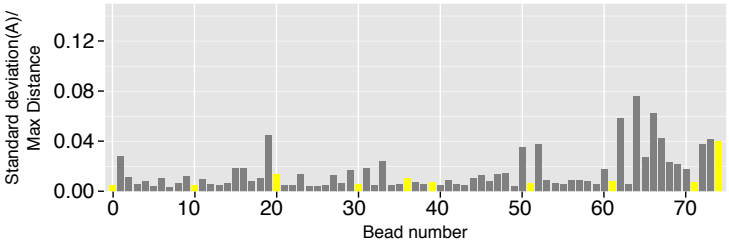




d

Six2/Six3





3

A single three-dimensional chromatin compartment in amphioxus indicates a stepwise evolution of vertebrate Hox bimodal regulation

nature genetics

A single three-dimensional chromatin compartment in amphioxus indicates a stepwise evolution of vertebrate Hox bimodal regulation

Rafael D Acemel[1,5], Juan J Tena[1,5], Ibai Irastorza-Azcarate[1,5], Ferdinand Marlétaz[2,5], Carlos Gómez-Marín[1], Elisa de la Calle-Mustienes[1], Stéphanie Bertrand[3], Sergio G Diaz[1], Daniel Aldea[3], Jean-Marc Aury[4], Sophie Mangenot[4], Peter W H Holland[2], Damien P Devos[1], Ignacio Maeso[1], Hector Escrivá[3] & José Luis Gómez-Skarmeta[1]

The HoxA and HoxD gene clusters of jawed vertebrates are organized into bipartite three-dimensional chromatin structures that separate long-range regulatory inputs coming from the anterior and posterior Hox-neighboring regions[1]. This architecture is instrumental in allowing vertebrate Hox genes to pattern disparate parts of the body, including limbs[2]. Almost nothing is known about how these three-dimensional topologies originated. Here we perform extensive 4C-seq profiling of the Hox cluster in embryos of amphioxus, an invertebrate chordate. We find that, in contrast to the architecture in vertebrates, the amphioxus Hox cluster is organized into a single chromatin interaction domain that includes long-range contacts mostly from the anterior side, bringing distant *cis*-regulatory elements into contact with Hox genes. We infer that the vertebrate Hox bipartite regulatory system is an evolutionary novelty generated by combining ancient long-range regulatory contacts from DNA in the anterior Hox neighborhood with new regulatory inputs from the posterior side.

How the three-dimensional organization of DNA in the nucleus influences regulation of gene expression is a topic of central importance in biology[3]. Despite recent progress in understanding chromatin organization, little is known about how such functional interactions evolve. Here we study the evolutionary pathway leading to the bipartite three-dimensional chromatin architecture regulating vertebrate Hox gene expression. In animals, chromatin is compartmentalized into topological associating domains (TADs)—megabase-scale chromatin regions containing DNA sequences that preferentially interact with one another[4,5]. A paradigmatic example of how TADs organize gene regulatory information is presented by the vertebrate Hox clusters, which contain genes of pivotal importance for animal development[6]. Different chromosome conformation capture techniques have shown that HoxA and HoxD genomic regions are each divided into two main adjacent TADs. These TADs compartmentalize long-range regulatory inputs coming from either side of the clusters into two major domains: enhancers distal to the 3′ flank preferentially contact 'anterior' Hox genes, whereas those beyond the 5′ flank mostly interact with 'posterior' genes (**Fig. 1a**; refs. 2,7–10). This bipartite regulatory topology provides gnathostomes with a versatile bimodal system, allowing Hox genes to pattern multiple structures, including an ancestral role in anteroposterior axis patterning and novel roles in morphological innovations such as paired limbs[1].

To investigate whether the TADs associated with HoxA and HoxD clusters arose independently or have a shared ancestry dating to before the two vertebrate-specific whole-genome duplications (2R WGDs; **Supplementary Fig. 1** and ref. 11), we first studied synteny conservation around Hox clusters between and within species. In mouse, the HoxA- and HoxD-neighboring regions are strikingly different, with many HoxA long-range *cis*-regulatory elements embedded in the introns of neighboring genes, whereas HoxD long-range *cis*-regulatory elements are located in gene deserts (large intergenic regions devoid of coding genes). Data from divergent vertebrates, including elephant shark, indicate the architecture in mouse represents a derived situation and that all vertebrate Hox cluster neighborhoods were originally very similar. What is now a HoxD gene desert in mammals contained copies of HoxA-neighboring genes[12], and the gene-free regions surrounding the other two Hox clusters have also resulted from differential loss of the coding exons of neighboring genes[13] (**Fig. 1b**, **Supplementary Fig. 2** and **Supplementary Note**). Thus, we conclude that differences in the genomic organization of mammalian HoxA and HoxD regulation are derived, not ancestral. This implies that the *cis*-regulatory elements currently engaged in Hox long-range bipartite contacts were primarily intronic and intergenic

[1]Centro Andaluz de Biología del Desarrollo (CABD), Consejo Superior de Investigaciones Científicas/Universidad Pablo de Olavide, Seville, Spain. [2]Department of Zoology, University of Oxford, Oxford, UK. [3]Université Pierre et Marie Curie Université Paris 6, CNRS, UMR 7232, Biologie Integrative des Organismes Marins (BIOM), Observatoire Océanologique de Banyuls-sur-Mer, Banyuls-sur-Mer, France. [4]Commissariat à l'Energie Atomique (CEA), Institut de Génomique (IG), Genoscope, Evry, France. [5]These authors contributed equally to this work. Correspondence should be addressed to J.L.G.-S. (jlgomska@upo.es), H.E. (hescriva@obs-banyuls.fr), I.M. (nacho.maeso@gmail.com) or D.P.D. (damienpdevos@gmail.com).

Received 30 October 2015; accepted 30 December 2015; published online 1 February 2016; doi:10.1038/ng.3497

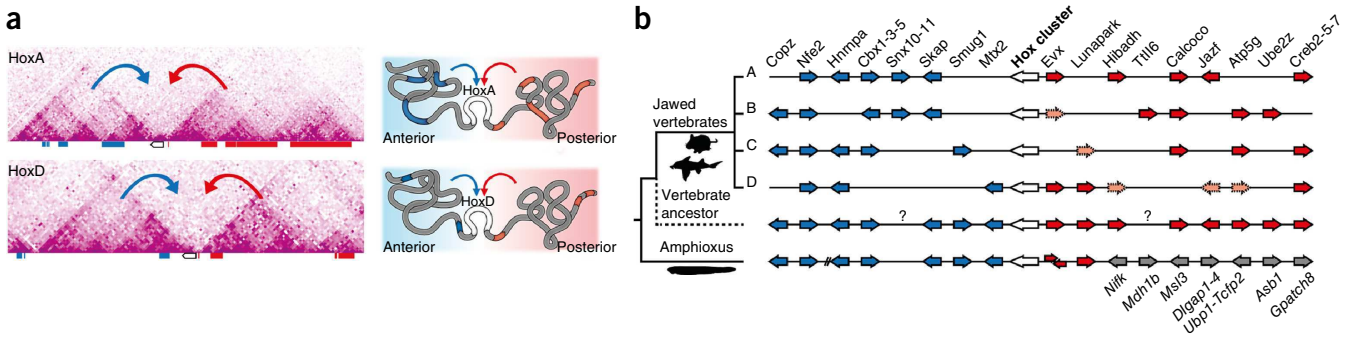


Figure 1 Genomic organization of vertebrate and amphioxus Hox clusters. (**a**) Distribution of TADs (obtained from human Hi-C data sets[20]) and schematics of the chromatin architecture of HoxA and HoxD clusters, showing their similar three-dimensional topologies. Colored bars represent Hox genes (white) and anterior (blue) and posterior (red) neighboring genes. Pink color intensity in the Hi-C plots corresponds to the number of interacting counts between bin pairs. (**b**) Microsynteny arrangements around the Hox clusters of gnathostomes, the pre-WGD vertebrate ancestor and amphioxus. Genes are represented by arrows showing transcriptional orientation (white, Hox clusters; blue, anterior genes; red, posterior genes; gray, genes with non-conserved linkages); those outlined by dashed lines correspond to vertebrate paralogs that have been lost in at least one species. Question marks indicate genes whose status in the vertebrate ancestor could not be inferred. Slashes correspond to the non-conserved amphioxus loci shown in **Supplementary Figure 3**.

within a conserved array of neighboring protein-coding loci before Hox cluster duplications (**Fig. 1b**).

We investigated the ancestry of this arrangement by examining the location of vertebrate Hox-neighboring genes in invertebrate genomes. We find that few of the invertebrate homologs are closely linked to Hox clusters outside chordates and that gene order and orientation are highly variable (for example, vertebrate anterior-linked genes are frequently found on the posterior side in invertebrates and vice versa; **Supplementary Fig. 3**). This shuffling of the Hox syntenic environment suggests that, in the bilaterian ancestor, long-range Hox *cis*-regulatory interactions were either absent or not important enough to constrain microsynteny. In contrast, in amphioxus (a non-vertebrate chordate that retains many ancestral genomic and morphological features; see refs. 14–16), synteny on the anterior side of the Hox cluster is strikingly conserved with vertebrates; gene order and orientation are almost identical to those inferred for the vertebrate ancestor (**Fig. 1b**). On the posterior side, most neighboring genes are different from those in vertebrates: only two immediately adjacent genes, *Evx* and *Lnp*, are conserved in position. The conservation of anterior flanking genes between vertebrates and amphioxus suggests that long-range regulatory interactions from the 3′ side of the cluster had become essential for Hox regulation at the base of the chordate lineage, imposing strong constraints on genomic rearrangements in this region. With regard to the posterior side, given the lack of synteny conservation in non-chordates, at present we cannot discern whether

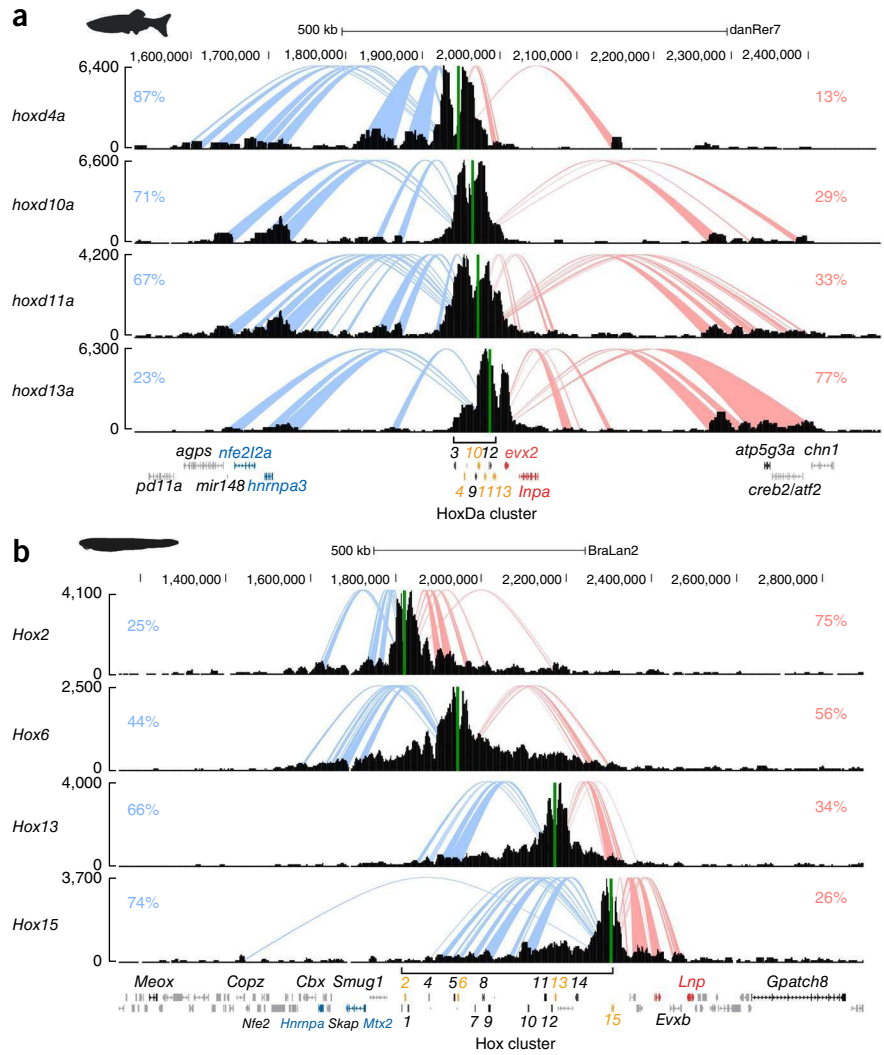


Figure 2 4C-seq interaction profiles of zebrafish and amphioxus Hox clusters. (**a**,**b**) Normalized 4C-seq profiles of several Hox gene promoters in the zebrafish HoxDa locus (**a**) and the amphioxus Hox locus (**b**). Spider plots show statistically significant contacts to the left (blue arcs) and right (red arcs) of each viewpoint. Percentages of reads aligned to statistically significant targets on each side of the viewpoints are indicated in blue (left contacts) and red (right contacts). Units on the *y* axes correspond to normalized interacting counts. Green bars indicate the positions of the viewpoints.

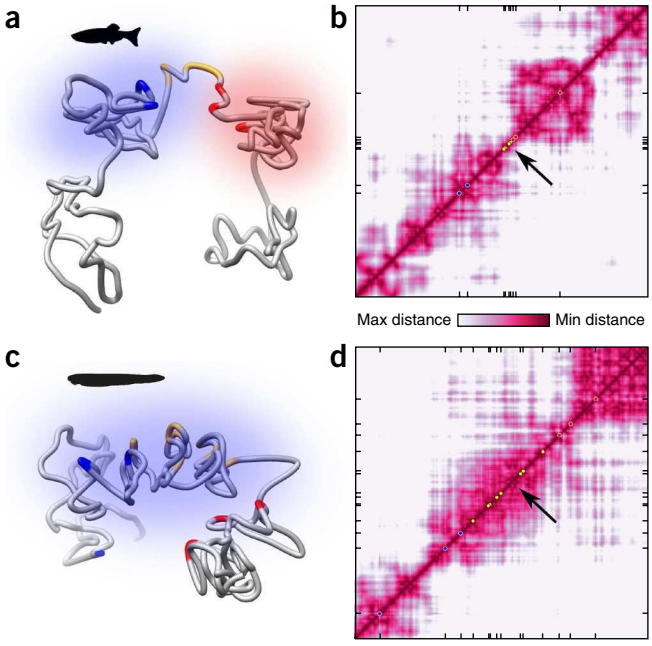


Figure 3 Three-dimensional chromatin architecture of amphioxus and zebrafish Hox clusters. (**a**,**c**) Three-dimensional models of the zebrafish HoxDa (**a**) and amphioxus Hox (**c**) regions. 4C-seq viewpoints are highlighted (blue, anterior genes; yellow, Hox genes; red, posterior genes). (**b**,**d**) Zebrafish and amphioxus virtual Hi-C consensus for all three-dimensional model solutions. 4C-seq viewpoints are represented by circles, using the same color scheme as in **a** and **c**. Arrows indicate the TAD border bisecting the zebrafish Hox cluster (**b**) and the absence of this border in the case of amphioxus (**d**).

amphioxus or vertebrates have diverged the most from the syntenic organization of the chordate ancestor. Whatever the case, beyond *Evx* and *Lnp*, gene synteny has followed different evolutionary routes in these two chordate groups, suggesting that, in stark contrast to the scenario in anterior territories, the regulatory contribution of distant posterior regions was less important or even absent in the last common chordate ancestor.

To evaluate this hypothesis experimentally, we compared Hox chromatin contacts between amphioxus and vertebrate embryos using circular chromosome conformation capture followed by high-throughput sequencing (4C-seq), a method that identifies distal chromatin contacts. Studies in mouse embryonic tissues and whole zebrafish embryos have demonstrated that 4C-seq efficiently resolves the organization of HoxA and HoxD long-range contacts into two adjacent TADs[2,7,8,10,17]. We generated 4C-seq data for 14 gene 'viewpoints' (eight Hox genes and six neighboring genes) in amphioxus embryos and compared these results with previously reported[8] and newly generated zebrafish data (four Hox genes and five neighboring genes). In total, 73 4C-seq data sets were generated, including replicates for all viewpoints and three amphioxus developmental stages (Online Methods).

With these data sets, we first defined target interacting regions for each of the 4C-seq viewpoints (genomic regions showing a statistically significant (P values $<1 \times 10^{-5}$) read enrichment against a randomized background) and quantified the number of reads corresponding to each of these targets (Online Methods). These analyses highlighted the characteristic bipartite distribution of anterior and posterior Hox long-range contacts previously reported in mouse and zebrafish[2,8,10,18] (**Fig. 2a** and **Supplementary Fig. 4**). The zebrafish *hoxd4a* and *hoxd13a* genes showed little contact overlap, with the majority of their interactions mapping to opposing sides of the cluster (83.3% anterior and 76.6% posterior, respectively). In contrast, in amphioxus, Hox genes located at the edges of the cluster showed the opposite trend: most *Hox2* and *Hox15* contacts converged in the same direction, with their interacting regions located primarily within the Hox complex (75.2% and 74.2%, respectively). In fact, regardless of their positions within the cluster, anterior, central and posterior Hox genes exhibited 4C-seq profiles that overlapped extensively, with no signs of a bipartite distribution (**Fig. 2b** and **Supplementary Fig. 5**). Notably, these Hox interaction profiles were developmentally stable, even though the number of active Hox genes in amphioxus changes dramatically from early gastrula to premouth embryo[19] (**Supplementary Fig. 6**). This temporal stability is in line with previous findings in mouse and *Drosophila melanogaster*, where most long-range three-dimensional chromatin interactions are organized similarly across tissues and developmental stages, with only some differences in the intensity of the contacts upon activation of different sets of distal enhancers[7,20,21]. However, despite this temporal uniformity, it is conceivable that the amphioxus TAD structures could be less similar across cell populations with different transcriptional activities than they are in vertebrates; thus, by using whole embryos, we may be missing cell type–specific chromatin interactions.

We then correlated 4C-seq results with synteny data. Consistent with the high level of conservation of anterior neighboring genes, in the majority of amphioxus Hox viewpoints, a significant fraction of contacts mapped to the conserved anterior region (ranging from 14 to 24.8% for the promoters of *Hox2*, *Hox5*, *Hox6*, *Hox7* and *Hox9*; **Supplementary Table 1**). Long-range interactions between Hox genes and anterior territories were even clearer when using 3′ neighboring genes as viewpoints (**Supplementary Fig. 5** and **Supplementary Table 1**). The amphioxus Hox cluster contained 25.5% of *Hnrnpa* interactions, a fraction in a similar range to that of its ortholog in zebrafish (33.4%), and, in the case of amphioxus *Mtx2*, the percentage of contacts corresponding to the Hox complex reached 42.7%. In contrast, on the posterior side, we found striking differences between amphioxus and vertebrates. Hox genes contacted posterior neighboring regions in both chordate lineages; however, the distribution of these 5′ interactions was very different (**Fig. 2a**,**b**). In zebrafish, *hoxd13a* interactions entered into far distant 5′ territories, well beyond the *evx2-lnpa* syntenic region, reaching vertebrate-specific posterior neighboring genes such as *atp5g3a* and *creb2*, consistent with previous reports on the location of zebrafish and mouse 5′ long-range Hox enhancers[7,22,23]. In amphioxus, by contrast, the target interacting regions of the most posterior Hox gene, *Hox15*, were circumscribed to the most proximal neighboring region, with no significant contacts crossing the *Lnp* promoter into the amphioxus-specific territory. Thus, even within the only 5′ region with synteny conservation, interaction profiles are different. In both cases, the Evx-Lnp region contacted Hox genes, but, whereas in amphioxus *Evxa* and *Lnp* showed a clear interaction preference for the Hox cluster (66.1% and 73% of contacts, respectively), zebrafish *evx2* and *lnpa* preferentially contacted vertebrate-specific genomic regions (with only 26.8% and 20.7% of the contacts interacting with the Hox cluster, respectively) (**Supplementary Figs. 4** and **5**, and **Supplementary Table 1**). Taken together, these results suggest that there is an inflexion point for long-range chromatin interactions around the *Evxa-Evxb-Lnp* region in amphioxus, with no significant Hox contacts with 5′ amphioxus-specific genes.

To better characterize vertebrate and amphioxus Hox topologies and identify interaction compartments, we generated virtual

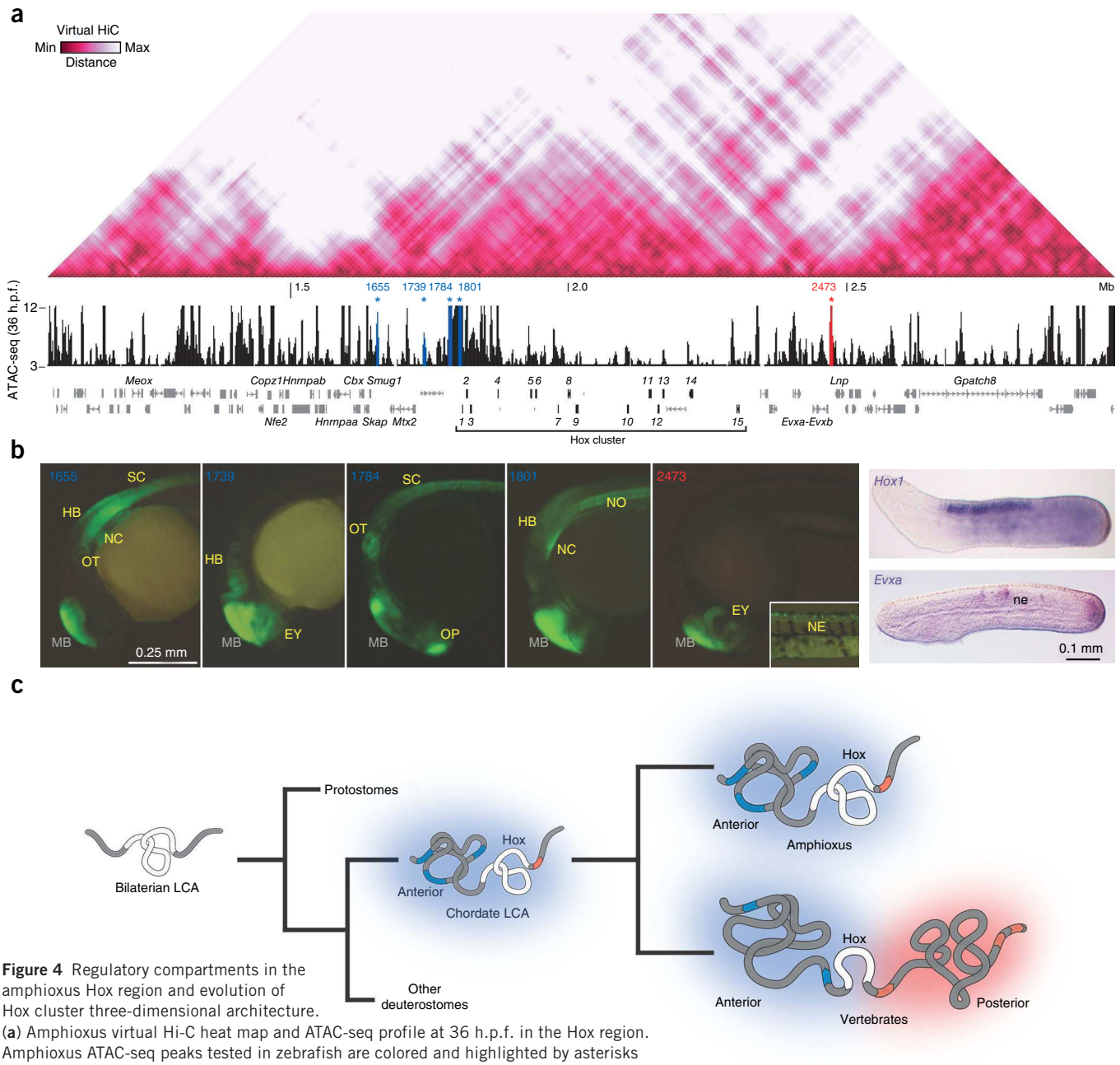


Figure 4 Regulatory compartments in the amphioxus Hox region and evolution of Hox cluster three-dimensional architecture. (**a**) Amphioxus virtual Hi-C heat map and ATAC-seq profile at 36 h.p.f. in the Hox region. Amphioxus ATAC-seq peaks tested in zebrafish are colored and highlighted by asterisks (blue for those in the anterior region, red for those on the posterior side of the cluster). (**b**) Lateral views of embryos from zebrafish transgenic lines at 24 h.p.f. and 48 h.p.f. (inset in 2473) showing GFP expression driven by the amphioxus ATAC-seq peaks (1655, 1739, 1784, 1801 and 2473) highlighted in **a**. Midbrain expression corresponds to the enhancer positive control included in the reporter constructs. Whole-mount *in situ* hybridizations for *Hox1* and *Evxa* in amphioxus embryos at 36 h.p.f. are shown for comparison. Anterior is to the left. EY, eye; HB, hindbrain; MB, midbrain; NC, neural crest cells; NE, neurons; NO, notochord; OT, otic vesicle, OP, olfactory placode, SC, spinal chord. (**c**) Three-dimensional architecture schematics showing an evolutionary scenario for the origin of the bimodal regulatory system of jawed vertebrates. The Hox-only chromatin domain of early bilaterians is first expanded on the anterior side in the chordate ancestor and then on the posterior side at the origin of vertebrates, allowing bipartition of the regulatory topologies of the HoxA and HoxD clusters. LCA, last common ancestor.

three-dimensional chromatin architecture models using the read counts of the 4C-seq signals as a proxy for the distance from each viewpoint (Online Methods and **Supplementary Fig. 7**). As the 4C-seq data correspond to pooled cells from whole embryos, our three-dimensional models provide an average view of chromatin topologies rather than a picture of the dynamic chromatin folding present in each individual cell. These integrative visualizations emphasized how strikingly different the vertebrate and amphioxus three-dimensional chromatin architectures are (**Fig. 3**). In zebrafish, the HoxDa cluster sits between the two separate anterior and posterior chromatin domains, like a hinge on which the two sets of long-range regulatory inputs can swing. In contrast, the amphioxus Hox cluster appears as a large single chromatin domain that contains distant anterior neighboring genes but not posterior ones. To visualize boundaries between chromatin domains, we developed a new approach to transform our three-dimensional modeling data into a heat map of distances (analogous to those obtained by Hi-C, hereafter termed virtual Hi-C; see the Online Methods and **Supplementary Figs. 8–10** for details

on virtual Hi-C validations). As expected, zebrafish virtual Hi-C recovered the bipartite architecture that divides vertebrate HoxD clusters into anterior and posterior TADs (**Fig. 3b**). In contrast, the amphioxus cluster was contained within a single TAD that included the conserved anterior neighboring genes but not the amphioxus-specific posterior genes (such as *Gpatch8*, which has its own interacting compartment) (**Fig. 3d**). Notably, no boundaries bisected the cluster or separated Hox genes from anterior neighboring territories. In the case of *Lnp* and the amphioxus Evx genes, the situation was less clear: although these loci seemed to be part of their own small interaction domain, this region was not completely isolated from its two adjacent compartments (the one containing the Hox cluster and the one including *Gpatch8*). This suggests that the single Hox three-dimensional chromatin domain present in amphioxus has a weaker contact border on its posterior side than in its anterior region and that the *EvxA-EvxB-Lnp* territory can be considered to be an extended boundary region (**Fig. 3d**).

To examine the functional relevance of amphioxus Hox chromatin organization, we searched for putative enhancers active in amphioxus embryos at 36 hours post-fertilization (h.p.f.) (immediately preceding what can be regarded as a pharyngula stage in amphioxus, equivalent to the zebrafish phylotypic stage at 24 h.p.f.) using ATAC-seq (assay for transposase-accessible chromatin using sequencing[24]; **Fig. 4a**). In agreement with the three-dimensional chromatin topologies inferred from the virtual Hi-C results, the distribution of ATAC-seq peaks on either side of the amphioxus Hox gene cluster suggested very different regulatory potentials for the two Hox-neighboring regions (**Supplementary Fig. 11**). Whereas anterior territories were rich in putative distal enhancer regions, the posterior side contained comparatively fewer ATAC-seq peaks. In fact, apart from the peaks tightly associated with the Evx genes or directly overlapping transcriptional start sites and repetitive elements, we only found a single candidate enhancer region, within the intergenic region between *Evxb* and *Lnp*. We then tested four putative enhancer elements from the anterior side of the TAD containing the amphioxus Hox cluster located at different distances from the closest Hox gene (elements 1655, 1739, 1784 and 1801, located 150 kb, 66 kb, 20 kb and 3 kb downstream of *Hox1*, respectively) and the element identified at the posterior side (element 2473, 165 kb upstream of *Hox15*) by generating zebrafish with stable GFP reporter transgenes. All the anterior enhancers promoted expression along the anteroposterior axis, consistent with the expression patterns of amphioxus Hox genes but not with those of neighboring loci (**Fig. 4b** and **Supplementary Fig. 12**; see also ref. 19), suggesting that these regions are amphioxus Hox *cis*-regulatory elements. In contrast, the 2473 posterior element activated GFP expression in isolated neurons in the spinal cord, in a pattern reminiscent of the amphioxus *Evxa* gene (**Fig. 4b**; ref. 25) rather than a Hox gene. These experiments suggest that the three-dimensional organization identified, using 4C-seq and modeling, brings long-range regulatory elements into proximity with amphioxus Hox genes mostly on the anterior side of the cluster (**Fig. 4**).

In summary, our results support a stepwise evolution of the bimodal regulatory machineries of the HoxA and HoxD clusters of jawed vertebrates (**Fig. 4c**). The relatively simple Hox cluster three-dimensional topology of early bilaterian animals, where external, long-range regulation was probably absent, changed profoundly in early chordate evolution, with newly incorporated distal regulatory inputs from anterior neighboring loci becoming a fundamental part of the Hox regulatory architecture. This unipolar topology was further developed in the vertebrate lineage. The acquisition of interactions with distal *cis*-regulatory elements on the posterior side introduced the possibility of a switch between two separate sets of long-range regulatory inputs, allowing an unprecedented plasticity in the developmental usage of the Hox patterning system in vertebrates.

URLs. Sickle (v1.290), https://github.com/najoshi/sickle; RepeatModeler, http://www.repeatmasker.org/RepeatModeler.html; UCSC Genome Browser, http://genome.ucsc.edu/; Ensembl Metazoa, http://metazoa.ensembl.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Data sets presented in this study are available under Gene Expression Omnibus (GEO) accession GSE68737.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We specially thank J. Pascual-Anaya for helping with some figures and helpful discussions. We would also like to thank F. Casares, I. Almudí and J.R. Martínez-Morales for fruitful discussions. Work was funded by grants from the Ministerio de Economía y Competitividad (BFU2013-41322-P to J.L.G.-S.; Juan de la Cierva postdoctoral contract to I.M.; BFU2014-58449-JIN to J.J.T.); the Andalusian government (BIO-396 to J.L.G.-S.; C2A (EE: 2013/2506) to D.P.D. and I.I.-A.); the European Research Council (ERC; grant 268513) to P.W.H.H. and F.M.; a European Molecular Biology Organization (EMBO) short fellowship to I.M.; the Universidad Pablo de Olavide to J.J.T.; and Conicyt 'Becas Chile' to D.A.

AUTHOR CONTRIBUTIONS

R.D.A. carried out the 4C-seq experiments with the help of C.G.-M. and S.G.D. J.J.T. performed the bioinformatic analysis of all the 4C-seq and ATAC-seq data sets. I.I.-A. developed and applied the three-dimensional modeling and virtual Hi-C procedures. F.M. generated the assembly and annotation of the Hox locus in the European amphioxus. J.-M.A. and S.M. ensured sequencing project management at Genoscope. E.d.l.C.-M., R.D.A., J.J.T. and I.M. carried out the zebrafish reporter assays. S.B. and D.A. collected and processed the amphioxus embryonic material and performed *in situ* hybridizations. I.M. completed the amphioxus ATAC-seq experiments and the evolution of synteny analyses. J.L.G.-S., H.E., I.M. and D.P.D. conceived, designed and coordinated the project. J.L.G.-S. and I.M. wrote the manuscript with the help of P.W.H.H. All authors revised and contributed to the final version of the text.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Lonfat, N. & Duboule, D. Structure, function and evolution of topologically associating domains (TADs) at *HOX* loci. *FEBS Lett.* **589**, 2869–2876 (2015).
2. Andrey, G. *et al.* A switch between topological domains underlies *HoxD* genes collinearity in mouse limbs. *Science* **340**, 1234167 (2013).
3. de Laat, W. & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* **502**, 499–506 (2013).
4. Gómez-Díaz, E. & Corces, V.G. Architectural proteins: regulators of 3D genome organization in cell fate. *Trends Cell Biol.* **24**, 703–711 (2014).
5. Ciabrelli, F. & Cavalli, G. Chromatin-driven behavior of topologically associating domains. *J. Mol. Biol.* **427**, 608–625 (2015).
6. Mallo, M. & Alonso, C.R. The regulation of *Hox* gene expression during animal development. *Development* **140**, 3951–3963 (2013).
7. Montavon, T. *et al.* A regulatory archipelago controls *Hox* genes transcription in digits. *Cell* **147**, 1132–1145 (2011).
8. Woltering, J.M., Noordermeer, D., Leleu, M. & Duboule, D. Conservation and divergence of regulatory strategies at *Hox* loci and the origin of tetrapod digits. *PLoS Biol.* **12**, e1001773 (2014).
9. Berlivet, S. *et al.* Clustering of tissue-specific sub-TADs accompanies the regulation of *HoxA* genes in developing limbs. *PLoS Genet.* **9**, e1004018 (2013).
10. Lonfat, N., Montavon, T., Darbellay, F., Gitto, S. & Duboule, D. Convergent evolution of complex regulatory landscapes and pleiotropy at *Hox* loci. *Science* **346**, 1004–1006 (2014).

11. Dehal, P. & Boore, J.L. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* **3**, e314 (2005).
12. Lehoczky, J.A., Williams, M.E. & Innis, J.W. Conserved expression domains for genes upstream and within the *HoxA* and *HoxD* clusters suggests a long-range enhancer existed before cluster duplication. *Evol. Dev.* **6**, 423–430 (2004).
13. Maeso, I. *et al.* An ancient genomic regulatory block conserved across bilaterians and its dismantling in tetrapods by retrogene replacement. *Genome Res.* **22**, 642–655 (2012).
14. Paps, J., Holland, P.W. & Shimeld, S.M. A genome-wide view of transcription factor gene diversity in chordate evolution: less gene loss in amphioxus? *Brief. Funct. Genomics* **11**, 177–186 (2012).
15. Bertrand, S. & Escriva, H. Evolutionary crossroads in developmental biology: amphioxus. *Development* **138**, 4819–4830 (2011).
16. Holland, L.Z. & Onai, T. Early development of cephalochordates (amphioxus). *Wiley Interdiscip. Rev. Dev. Biol.* **1**, 167–183 (2012).
17. Noordermeer, D. *et al.* Temporal dynamics and developmental memory of 3D chromatin architecture at *Hox* gene loci. *eLife* **3**, e02557 (2014).
18. Noordermeer, D. *et al.* The dynamic architecture of *Hox* gene clusters. *Science* **334**, 222–225 (2011).
19. Pascual-Anaya, J. *et al.* Broken colinearity of the amphioxus *Hox* cluster. *Evodevo* **3**, 28 (2012).
20. Dixon, J.R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 2012).
21. Ghavi-Helm, Y. *et al.* Enhancer loops appear stable during development and are associated with paused polymerase. *Nature* **512**, 96–100 (2014).
22. Gonzalez, F., Duboule, D. & Spitz, F. Transgenic analysis of *Hoxd* gene regulation during digit development. *Dev. Biol.* **306**, 847–859 (2007).
23. Gehrke, A.R. *et al.* Deep conservation of wrist and digit enhancers in fish. *Proc. Natl. Acad. Sci. USA* **112**, 803–808 (2015).
24. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
25. Ferrier, D.E., Minguillón, C., Cebrián, C. & Garcia-Fernàndez, J. Amphioxus *Evx* genes: implications for the evolution of the midbrain-hindbrain boundary and the chordate tailbud. *Dev. Biol.* **237**, 270–281 (2001).

ONLINE METHODS

Genome sequencing and assembly. DNA was prepared from a single European amphioxus (*Branchiostoma lanceolatum*) mature male and sequenced using Illumina technology at Genoscope (Centre National de Séquençage, Evry, France). Briefly, two paired-end (180-bp and 700-bp) and six mate-pair (3-, 5- and 8-kb) libraries were generated and sequenced at >200× total coverage.

Reads were quality-trimmed using sickle (v1.290), and errors were corrected using Musket (v1.0.6)[26]; overlapping libraries were merged using Flash[27]. Assembly was carried out using SOAPdenovo (v2.04)[28] with a *k*-mer of 71 for contig generation and of 35 for mapping and scaffolding. Gaps were subsequently filled using GapCloser (v1.2)[28] with an overlap parameter set to 31. The resulting assembly (N50, 649 kb; size, 948.5 Mb) contains allelic copies for most scaffolds (expected genome size of ~500 Mb) that we reconciled using the Haplomerger[29] pipeline, relying on best reciprocal LASTZ alignment after masking repeats using a custom library built with RepeatModeler. The Hox locus was extracted from the final assembly (N50, 1.132 Mb; size, 526.8 Mb) and submitted together with the 4C-seq and ATAC-seq data (GSE68737).

Gene models were built using Evidence Modeler (EVM)[30] on the basis of (i) *de novo* gene prediction obtained using Augustus[31] with custom training based on CEGMA[32] report and (ii) split-aware alignment of human proteins using Exonerate[33] and transcriptome alignment. Models for known genes in the Hox region that were not present in these annotations were added manually. More details regarding *B. lanceolatum* genome assembly and annotation will be provided in a separate publication.

Synteny analyses and genome browsing. Hox-neighboring genes were searched across the different studied species using TBLASTN and BLASTP. We compared the relative orientations and positions of these genes by browsing the genomes of the studied species through the NCBI, UCSC Genome Browser and Ensembl Metazoa webpages, using the following genome versions: elephant shark (*Callorhinchus milii*) 6.1.3, *Lottia gigantea* v1.0, *Mus musculus* Build 38, *Saccoglossus kowalevskii* Build1.1, *Strigamia maritima* Smar1.0 and *Trichoplax adhaerens* v1.0. In the case of the starfish *Acanthaster planci*, no gene annotation or genome browser was available for the published *A. planci* Hox genome scaffold (DF933567.1)[34]. Therefore, we used TBLASTN to search for conserved neighboring genes and Genscan to predict genes *de novo*.

Mouse *Jazf2* pseudogenized exons were detected with VISTA[35] using elephant shark as a reference sequence, LAGAN as the alignment program and the following parameters: 100-bp window and 65% identity in 70 bp.

Amphioxus procurement and culture. *B. lanceolatum* mature adults were collected at the Racou beach in Argelès-sur-Mer (France). Gametes were collected by heat stimulation as previously described[36,37]. Fertilization was undertaken in Petri dishes filled with filtered seawater, and embryos were cultured at 19 °C.

Whole-mount *in situ* hybridization. Partial cDNA for *Gpatch8*, *Nfe2*, *Lnp*, *Slc20*, *Mtx2*, *Hnrnpa* and *Cbx* from *B. lanceolatum* was amplified by RT-PCR and cloned into the pGEM-T Easy vector. DIG-labeled RNA probes were synthesized by *in vitro* translation after plasmid linearization using the appropriate enzymes. Fixation and whole-mount *in situ* hybridization were performed as described in ref. 38. No expression could be detected using whole-mount *in situ* hybridization for *Gpatch8*, *Lnp*, *Slc20* or *Mtx2*; the expression patterns for the rest of the genes are included in **Supplementary Figure 12**.

4C-seq. 4C-seq assays were performed as previously reported[18,39–41]. For each zebrafish biological replicate, 500 embryos at 24 h.p.f. of the Tübingen strain were dechorionated using pronase and deyolked in 1 ml of Ginzburg Fish Ringers (55 mM NaCl, 1.8 mM KCl and 1.25 mM $NaHCO_3$). They were then fixed in 2% formaldehyde in 1× PBS for 15 min at room temperature. For amphioxus biological replicates, embryos (~8,000 at 8 h.p.f. and ~4,000 at 15 and 36 h.p.f.) were concentrated by centrifugation at low speed in 2-ml microtubes. They were fixed for 15 min at room temperature in 1.5 ml of MOPS buffer (0.1 M MOPS pH 7.5, 2 mM $MgSO_2$, 1 mM EGTA and 0.5 M NaCl) containing 1.85% formaldehyde. 155 µl of 10% glycine was added to both species samples to stop fixation, followed by five washes with PBS (NaPBS in the case of amphioxus) at 4 °C. Pellets were frozen in liquid nitrogen and kept at −80 °C. Isolated cells were lysed (lysis buffer: 10 mM Tris-HCl pH 8, 10 mM NaCl, 0.3% Igepal CA-630 (Sigma-Aldrich, I8896) and 1× protease inhibitor cocktail (Complete, Roche, 11697498001)), and the DNA was digested with DpnII (New England BioLabs, R0543M) and Csp6I (Fermentas, Thermo Scientific, FD0214) as primary and secondary enzymes, respectively. T4 DNA ligase (Promega, M1804) was used for both ligation steps. Specific primers were designed around the putative transcriptional start sites of the genes with Primer3 v. 0.4.0 (ref. 42). Illumina adaptors were included in the primer sequences, and eight PCRs were performed with the Expand Long Template PCR System (Roche, 11759060001) and pooled. Two libraries from different biological replicates were generated for each 4C-seq experiment (for each viewpoint and for each developmental stage). These libraries were purified with a High Pure PCR Product Purification kit (Roche, 11732668001), their concentrations were measured using the Quanti-iT PicoGreen dsDNA Assay kit (Invitrogen, P11496) and they were sent for deep sequencing. 4C-seq data were analyzed as previously described[17]. Briefly, raw sequencing data were demultiplexed and aligned using the zebrafish July 2010 assembly (danRer7) and the *B. lanceolatum* reference genomes. Reads located in fragments flanked by two restriction sites of the same enzyme, in fragments smaller than 40 bp or within a window of 10 kb around the viewpoint (indicated by dashed lines in the different figures) were filtered out. Mapped reads were then converted to reads per first enzyme fragment ends and smoothened using a 30-fragment mean running window algorithm. 4C-seq data were normalized by the total weight of reads within the window displayed in the figures.

To calculate statistically significant contacting regions for each viewpoint, an average background level was estimated as previously described[43]. Briefly, fragment distribution in a window of 2 Mb around each viewpoint was randomized, excluding an internal window of 100 kb around the viewpoint to avoid biases due to close contacts. Then, this randomized fragment distribution was smoothened as described above. This randomized profile was then used to calculate the *P* value for each potential target in the observed 4C-seq distribution by means of Poisson probability function. Regions with *P* values below 1×10^{-5} were considered as statistically significant interacting targets.

To calculate the distribution of contacts at each side of the viewpoints, we took into account only those reads overlapping the interacting targets, discarding also those mapped within the 100-kb viewpoint window, as previously reported[8]. The same approach was used to quantify the distribution of contacts in the three windows defined as follows: cluster (from the 5′ UTR of the most 5′ Hox genes (zebrafish, *hoxd13a*; amphioxus, *Hox15*) to the 3′ UTR of the most 3′ Hox genes (zebrafish, *hoxd3a*; amphioxus, *Hox1*)); anterior (downstream of the zebrafish *hoxd3a* and amphioxus *Hox1* genes); and posterior (upstream of the zebrafish *hoxd13a* and amphioxus *Hox15* genes).

Three-dimensional computational modeling and virtual Hi-C. *4C data normalization.* To equalize the amount of reads in all experiments, we normalized the reads for the 4C-seq data sets. We then extracted the data relevant for modeling by calculating the *Z* score (see below on *Z*-score threshold optimization) of those reads as in ref. 44.

Structure determination. The overall approach for determination of genome structure was adapted from a previous work[44] with some variations, using the Integrative Modeling Platform (IMP)[45]. The procedure was divided into three stages:

(1) Representation of the genome locus and translation of the data into spatial restraints. We represented the chromosomal fragment as a flexible string of beads where each bead corresponded to a number of consecutive fragments between ten and 45, depending on the total size of the locus (**Supplementary Fig. 7c**). The size of the beads representing those 20 fragments was proportional to the sum of the sizes of these fragments.

To impose connection between the beads, harmonic upper-bound distance restraints were used between consecutive beads. This distance was the sum of the radii of both beads. Excluded volume restraints were imposed over all the beads so that these would not overlap each other. The reach window of a viewpoint was defined as the area between the furthest upstream and downstream fragments with a *Z* score above the upper *Z* score (u*Z*) (**Supplementary Fig. 13**). Harmonic distance restraints were applied between beads corresponding to the viewpoints and the rest of the beads, as long as the *Z* scores for these beads were above the u*Z* or below the lower *Z* score (l*Z*). We used the absolute

Z score of the reads to give more weight to the most meaningful reads. Beads outside the reach window were restrained with harmonic lower-bound distances, with a weight equal to the absolute *Z* score. With the harmonic lower-bound restraint, we only imposed the criterion that the beads not be closer than their computed distance (**Supplementary Fig. 7**).

(2) Optimization and sampling of the space of solutions. We combined a Monte Carlo exploration with a local optimization of conjugate gradients and simulated annealing. We started with an individual optimization of five steps of conjugate gradients from an entirely random configuration of beads followed by simulated annealing until the score difference between rounds was below 0.00001 or reached 0 (**Supplementary Fig. 7d**). To sample the space of solutions exhaustively, we computed 50,000 independent optimizations for each genome (**Supplementary Fig. 7e**).

(3) Analysis and assessment of the ensemble of models. We gathered the 200 models with the best score. These solutions were then clustered according to their similarity as measured by their root mean square deviation (r.m.s. deviation). We used the Multiexperiment Viewer, MeV[46], with hierarchical clustering and *k*-means clustering. All models were grouped in two clusters that were the mirror image of each other (**Supplementary Fig. 14**). The most representative models (the closest ones to the mean of all solutions within the most populated cluster) are displayed in **Figure 3**. Results were indistinguishable when we used the solutions for the other mirror-image cluster.

Reconstruction of virtual Hi-C data. We used the models from the most populated cluster to generate the heat map plots that were equivalent to Hi-C data. First, we superimposed all the models (**Supplementary Fig. 7f**). To generate virtual Hi-C heat map plots, we measured the distances between all beads in each model and calculated the mean of these distances (**Supplementary Fig. 7g**).

Empirical calculation of the maximum distance, the lZ and the uZ. The calculation of these parameters was carried out as described previously[44] with small variations. The u*Z* score varied between 0.2 and 1.4 in bins of 0.2. The l*Z* score varied in bins of 0.2 between −1.4 and −0.2. The maximum distance varied from 3,000 to 7,000 in bins of 1,000. Because of the heavy computational load, we did not consider narrower bins or higher or lower values.

For each set of parameters, we generated 500 models, calculated the mean distances between the viewpoints and the rest of the fragments, and compared them to the distances that represented each set of 20 fragments of the normalized 4C data (**Supplementary Fig. 15b**,**d**,**f**).

The set of parameters that best fitted the 4C data included 0.2 for u*Z* and −0.2 for l*Z* in amphioxus, zebrafish and mouse. The best maximum distances were different for each species. To allow comparison, we needed to set the same maximum distance for all three. Taking this into account and for the sake of ease of visualization, we settled on the maximum distance of 7,000, whose score was also among the best (**Supplementary Fig. 15a**,**c**,**e**).

Validation of the virtual Hi-C approach. To validate the virtual Hi-C method, we followed two strategies.

(1) Jackknife resampling. We tested the reproducibility and robustness of the virtual Hi-C results by taking advantage of the extensive number of viewpoints available in our amphioxus and zebrafish Hox 4C-seq data. We performed additional modeling experiments by resampling our original data sets using different subsets of 4C data both in zebrafish and amphioxus (**Supplementary Table 2**). We generated 500 models with the same parameters that we used for our initial modeling and reconstructed virtual Hi-C data for each subset. Subsequently, we calculated Spearman's coefficients between the different subsets. This demonstrated that virtual Hi-C results are very reproducible and robust to perturbations, with high correlations even when 60% of the viewpoints were eliminated (**Supplementary Fig. 10** and **Supplementary Table 2**).

(2) Modeling of other loci and shifted calculation of correlations. To validate the models and the virtual Hi-C results derived from them, we generated models for diverse mouse genomic regions using previously published 4C-seq data (from the HoxD locus and two additional loci: *Wnt6-Ihh-Epha4-Pax3* and *Med13l-Tbx3-Tbx5-Rbm19*; refs. 17,47,48). Using these models, we generated the virtual Hi-C results and compared them with previously published experimental Hi-C data[20] (**Supplementary Figs. 8** and **9**). These comparisons were performed shifting the window used for the modeling by 25% of its size in each direction, in steps of 20 kb (**Supplementary Fig. 8**). For each comparison, Spearman's and Pearson's correlations were calculated. Because of the dominance of read counts corresponding to short distances, we calculated these correlations using bins separated by at least 240 kb (HoxD and *Med13l-Tbx3-Tbx5-Rbm19*) or 480 kb (*Wnt6-Ihh-Epha4-Pax3*), to account for the different size of these three loci (~2.12, ~2.48 and ~4.88 Mb, respectively). In all cases, our 4C-seq–derived virtual Hi-C contact matrices accurately recapitulate the TAD organization and borders present in the experimental Hi-C maps, with Spearman's and Pearson's coefficients within the same range (from 0.63 to 0.88) of those typically obtained between different Hi-C experimental conditions (from 0.4 to 0.99; refs. 20,49–51) (**Supplementary Fig. 9** and **Supplementary Table 3**).

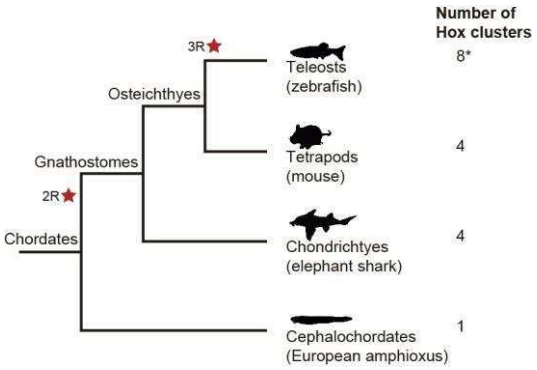
ATAC-seq. ATAC-seq experiments in amphioxus embryos were performed as previously described[23,24]. Approximately 80,000 cells (corresponding to 13 embryos at 36 h.p.f.) were directly lysed in cold lysis buffer (10 mM Tris pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$ and 0.1% Igepal) after removing the seawater by centrifuging briefly. The sample was then incubated for 30 min at 37 °C with TDE1 enzyme and purified with the Qiagen MinElute kit. A PCR reaction was performed with 13 cycles using Ad1F and Ad2.3R primers and KAPA HiFi Hot-Start enzyme (Kapa Biosystems). The resulting library was multiplexed and sequenced in a HiSeq 2000 lane. Reads were aligned using the mentioned *B. lanceolatum* assembly. Duplicated pairs or those separated by more than 2 kb were removed. The enzyme cleavage site was determined as the position −4 (minus strand) or +5 (plus strand) relative to each read start, and this position was extended by 5 bp in both directions for signal visualization. For the zebrafish reporter assays of anterior elements, we selected four regions including ATAC-seq peaks with no overlap with coding exons, transcriptional start sites and repetitive elements. We applied the same criteria to the posterior region, also excluding ATAC-seq peaks tightly associated with amphioxus Evx genes (those located in Evx introns and within 5 kb of Evx transcribed regions). This rendered a single candidate element between *Evxb* and *Lnp* (**Supplementary Fig. 11**).

Transgenesis in zebrafish. Transgenesis assays were performed as previously reported[52]. Putative enhancers were amplified by PCR from amphioxus genomic DNA using the primers listed in **Supplementary Table 4**. The PCR fragments were subcloned into PCR8/GW/TOPO vector and, using Gateway technology (Life Technologies), were shuttled into an enhancer detection vector composed of a *gata2* minimal promoter, an enhanced *GFP* reporter gene and a strong midbrain enhancer (z48) that works as an internal control for transgenesis in zebrafish[23]. Zebrafish transgenic embryos were generated using the Tol2 transposon/transposase method[53], with minor modifications. One-cell embryos were injected with a 2-nl volume containing 25 ng/µl of transposase mRNA, 20 ng/µl of purified constructs and 0.05% phenol red. To ensure the reproducibility of the expression patterns observed in the reporter assays, three or more stable transgenic lines derived from different founders were generated for each construct. All experimental procedures using vertebrates were ethically approved by the Andalusian government.

26. Liu, Y., Schröder, J. & Schmidt, B. Musket: a multistage *k*-mer spectrum–based error corrector for Illumina sequence data. *Bioinformatics* **29**, 308–315 (2013).
27. Magoč , T. & Salzberg, S.L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
28. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* **1**, 18 (2012).
29. Huang, S. *et al.* HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. *Genome Res.* **22**, 1581–1588 (2012).
30. Haas, B.J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
31. Stanke, M. *et al.* AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
32. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
33. Slater, G.S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
34. Baughman, K.W. *et al.* Genomic organization of *Hox* and *ParaHox* clusters in the echinoderm, *Acanthaster planci*. *Genesis* **52**, 952–958 (2014).
35. Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M. & Dubchak, I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* **32**, W273–W279 (2004).

doi:10.1038/ng.3497

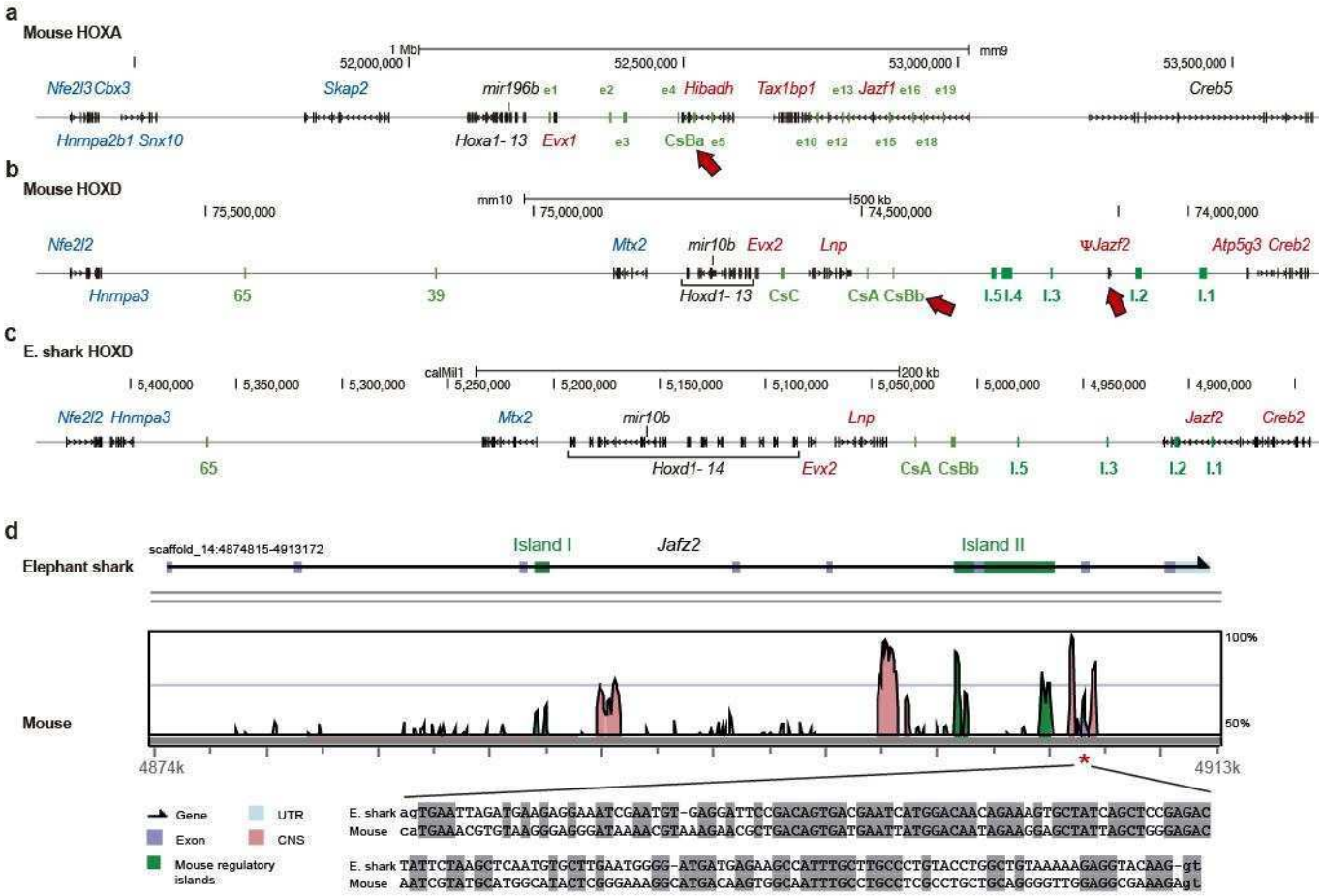
36. Fuentes, M. *et al.* Insights into spawning behavior and development of the European amphioxus (*Branchiostoma lanceolatum*). *J. Exp. Zool. B Mol. Dev. Evol.* **308**, 484–493 (2007).
37. Fuentes, M. *et al.* Preliminary observations on the spawning conditions of the European amphioxus (*Branchiostoma lanceolatum*) in captivity. *J. Exp. Zool. B Mol. Dev. Evol.* **302**, 384–391 (2004).
38. Somorjai, I., Bertrand, S., Camasses, A., Haguenauer, A. & Escriva, H. Evidence for stasis and not genetic piracy in developmental expression patterns of *Branchiostoma lanceolatum* and *Branchiostoma floridae*, two amphioxus species that have evolved independently over the course of 200 Myr. *Dev. Genes Evol.* **218**, 703–713 (2008).
39. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
40. Hagège, H. *et al.* Quantitative analysis of chromosome conformation capture assays (3C-qPCR). *Nat. Protoc.* **2**, 1722–1733 (2007).
41. Splinter, E., de Wit, E., van de Werken, H.J., Klous, P. & de Laat, W. Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: from fixation to computation. *Methods* **58**, 221–230 (2012).
42. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132**, 365–386 (2000).
43. Smemo, S. *et al.* Obesity-associated variants within *FTO* form long-range functional connections with *IRX3*. *Nature* **507**, 371–375 (2014).
44. Baù, D. *et al.* The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nat. Struct. Mol. Biol.* **18**, 107–114 (2011).
45. Russel, D. *et al.* Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol.* **10**, e1001244 (2012).
46. Saeed, A.I. *et al.* TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34**, 374–378 (2003).
47. Lupiáñez, D.G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).
48. van Weerd, J.H. *et al.* A large permissive regulatory domain exclusively controls *Tbx3* expression in the cardiac conduction system. *Circ. Res.* **115**, 432–441 (2014).
49. Vietri Rudan, M. *et al.* Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep.* **10**, 1297–1309 (2015).
50. Hou, C., Li, L., Qin, Z.S. & Corces, V.G. Gene density, transcription, and insulators contribute to the partition of the *Drosophila* genome into physical domains. *Mol. Cell* **48**, 471–484 (2012).
51. Zhang, Y. *et al.* Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell* **148**, 908–921 (2012).
52. Bessa, J. *et al.* Zebrafish enhancer detection (ZED) vector: a new tool to facilitate transgenesis and the functional analysis of *cis*-regulatory regions in zebrafish. *Dev. Dyn.* **238**, 2409–2417 (2009).
53. Kawakami, K. Transgenesis and gene trap methods in zebrafish by using the *Tol2* transposable element. *Methods Cell Biol.* **77**, 201–222 (2004).



Supplementary Figure 1

Schematic phylogenetic tree showing the main chordate species used in the present study.

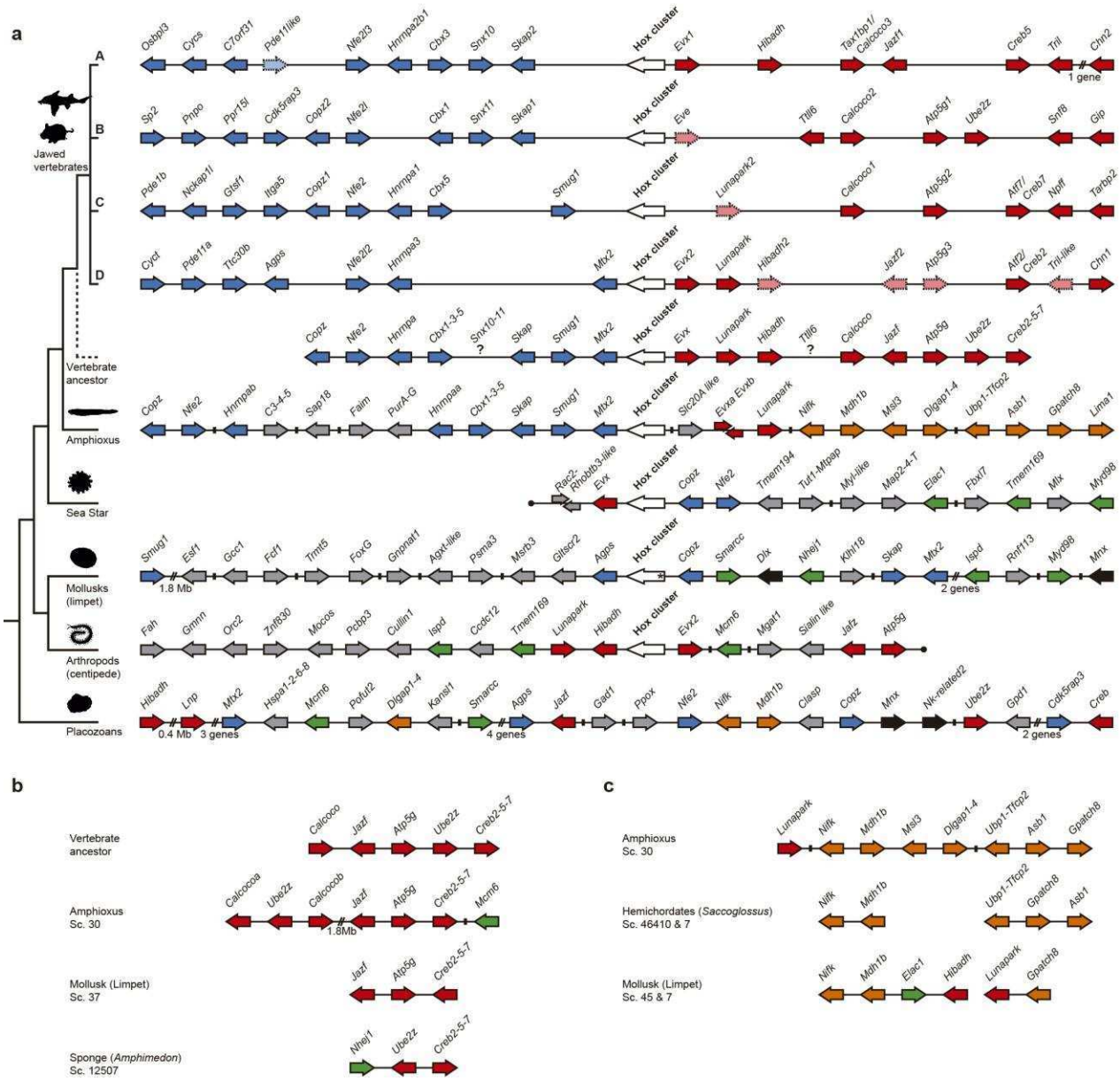
Red stars correspond to the two WGD events that occurred in the vertebrate ancestor (2R) and the extra WGD round that happened at the origin of teleost fish (3R). The asterisk indicates that, in zebrafish in contrast to other teleost species, HoxDb has been secondarily lost and only *mir10* and the anterior and posterior Hox-neighboring genes still remain in this genomic region.



Supplementary Figure 2

'Desertification' of HoxD clusters.

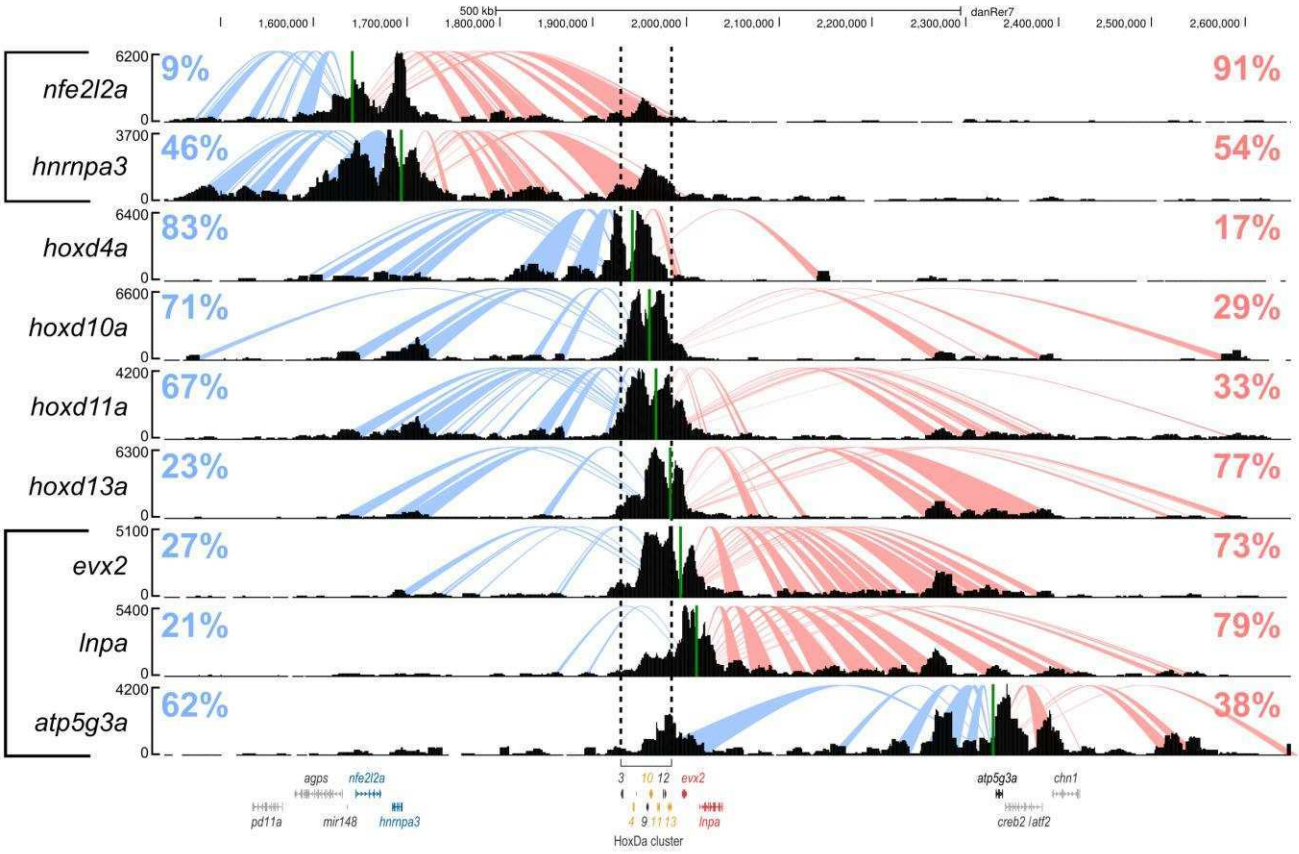
(**a**–**c**) Genomic organization of the mouse HoxA and HoxD and elephant shark HoxD clusters. Symbols for Hox, anterior and posterior neighboring genes are colored in black, blue and red, respectively. Several mouse long-range enhancers and their orthologs in elephant shark are represented by green bars. Red arrows indicate the two paralogous CsB enhancers that demonstrate the loss of a *Hibadh2* gene from HoxA-neighboring regions and the pseudogenized remnant of mouse *Jazf2*. (**d**) VISTA plot of the *Jazf2* genomic region in elephant shark (reference sequence) and mouse, showing the mouse *Jazf2* pseudoexon (red asterisk) and the ancestral intronic location of mouse regulatory islands I and II. VISTA colored peaks (blue, coding; turquoise, UTR; pink, noncoding; green, mouse regulatory islands) indicate regions of at least 70 bp and ≥65% similarity. The alignment below the plot corresponds to the region indicated by the asterisk, showing several mutations in the mouse *Jazf2* sequence, including splice sites (in lower case) and frameshifts.



Supplementary Figure 3

Conservation of microsynteny around Hox-neighboring genes.

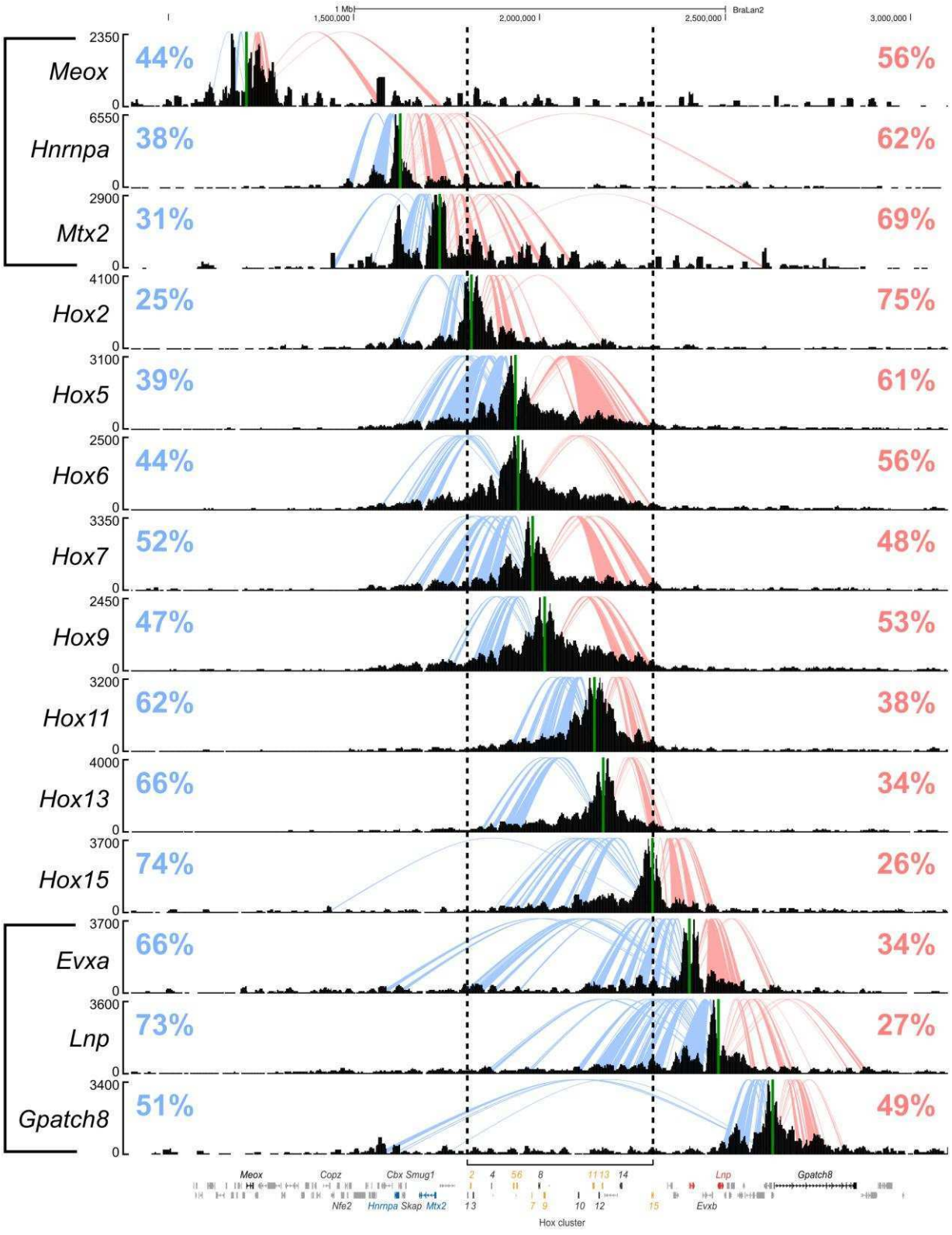
(**a**) Microsynteny arrangements around the Hox clusters of different bilaterian species and the 'ghost' Hox locus of the placozoan *Trichoplax adhaerens*. Note that, because of the lack of synteny conservation, we could not infer a consensus for different vertebrate species beyond the genes included in the vertebrate ancestor reconstruction. Thus, for genes beyond these limits, the information displayed in this figure corresponds mainly to the mouse genome. (**b**) Conserved linkage of vertebrate posterior neighboring genes in amphioxus and non-chordate species. (**c**) Conserved linkage of amphioxus posterior neighboring genes in non-chordate species. Genes are represented by arrows (white, Hox clusters; blue, chordate anterior neighboring genes; red, vertebrate posterior neighboring genes; orange, amphioxus posterior neighboring genes; green, non-chordate neighboring genes linked to Hox genes in at least two species; black, non-Hox ANTP-class homeobox genes). Question marks represent genes whose status in the vertebrate ancestor could not be inferred. Slashes indicate the presence of genes not represented in the figure. Black circles represent the end of the genomic scaffold. Small black rectangles indicate the presence of predicted gene model(s) with no clear orthologs in other species and that in most cases have multiple additional copies in their corresponding genomes. The black asterisk within the Hox cluster arrow of *L. gigantea* indicates the reversed orientation of the last Hox posterior gene in this species.



Supplementary Figure 4

4C-seq interaction profiles of the zebrafish HoxDa region

Normalized 4C-seq profiles of the promoters for several Hox and neighboring genes in the zebrafish HoxDa region (labeled as in **Fig. 2**). The 4C-seq profiles corresponding to neighboring genes are indicated with large brackets at the left margin of the figure. Spider plots are color-coded as in **Fig. 2**. Green lines indicate the positions of the viewpoints. Dotted lines indicate the genomic region containing the HoxDa cluster. Units on the *y* axes correspond to normalized interacting counts.

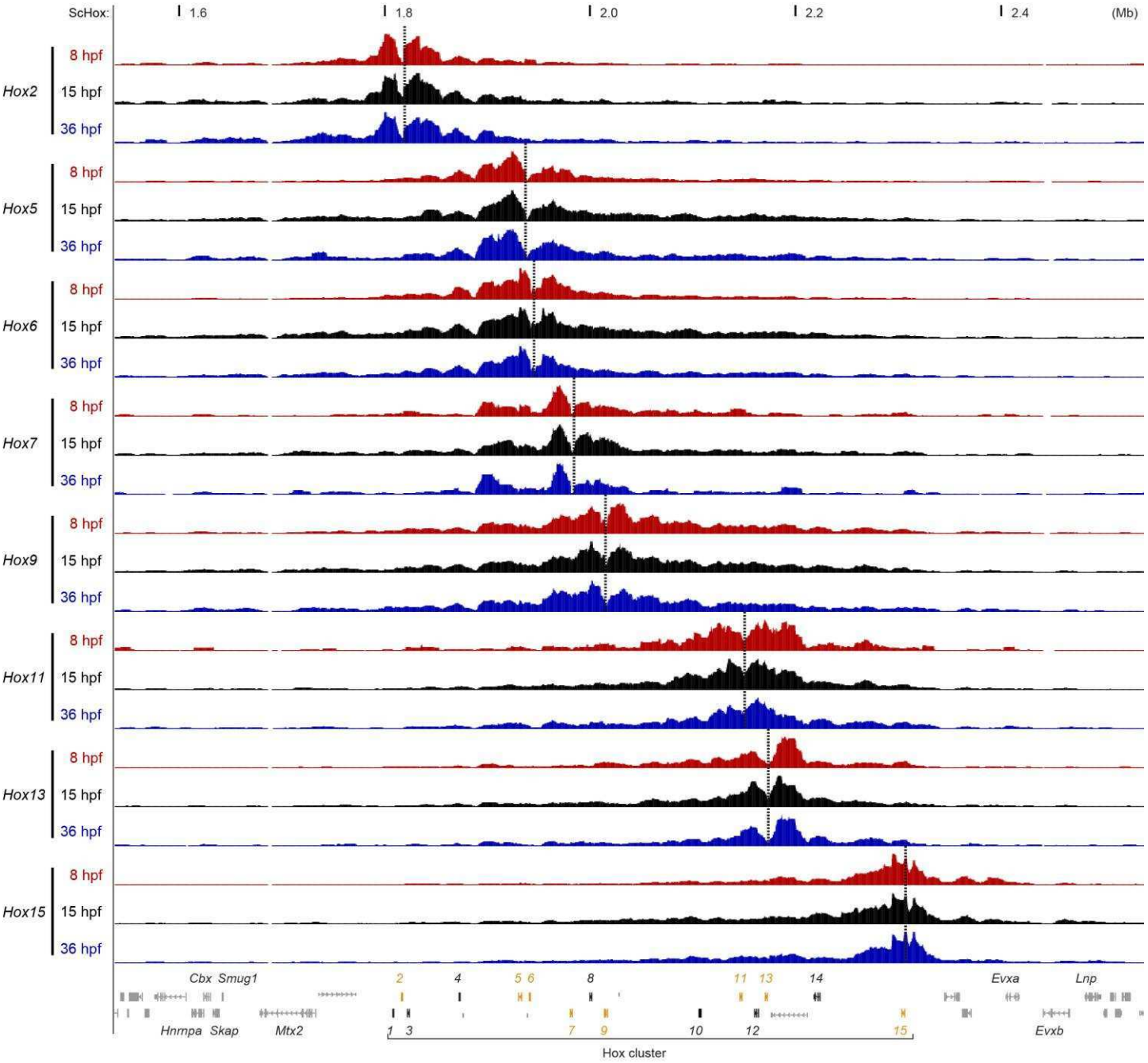


Supplementary Figure 5

4C-seq interaction profiles of amphioxus Hox regions.

Normalized 4C-seq profiles of the promoters for several Hox and neighboring genes in amphioxus Hox region (labeled as in **Fig. 2**). The 4C-seq profiles corresponding to neighboring genes are indicated with large brackets at the left margin of the figure. Spider plots are color-coded as in **Fig. 2**. Green lines indicate the positions of the viewpoints. Dotted lines indicate the genomic region containing the Hox cluster. Units on the *y* axes correspond to normalized interacting counts.

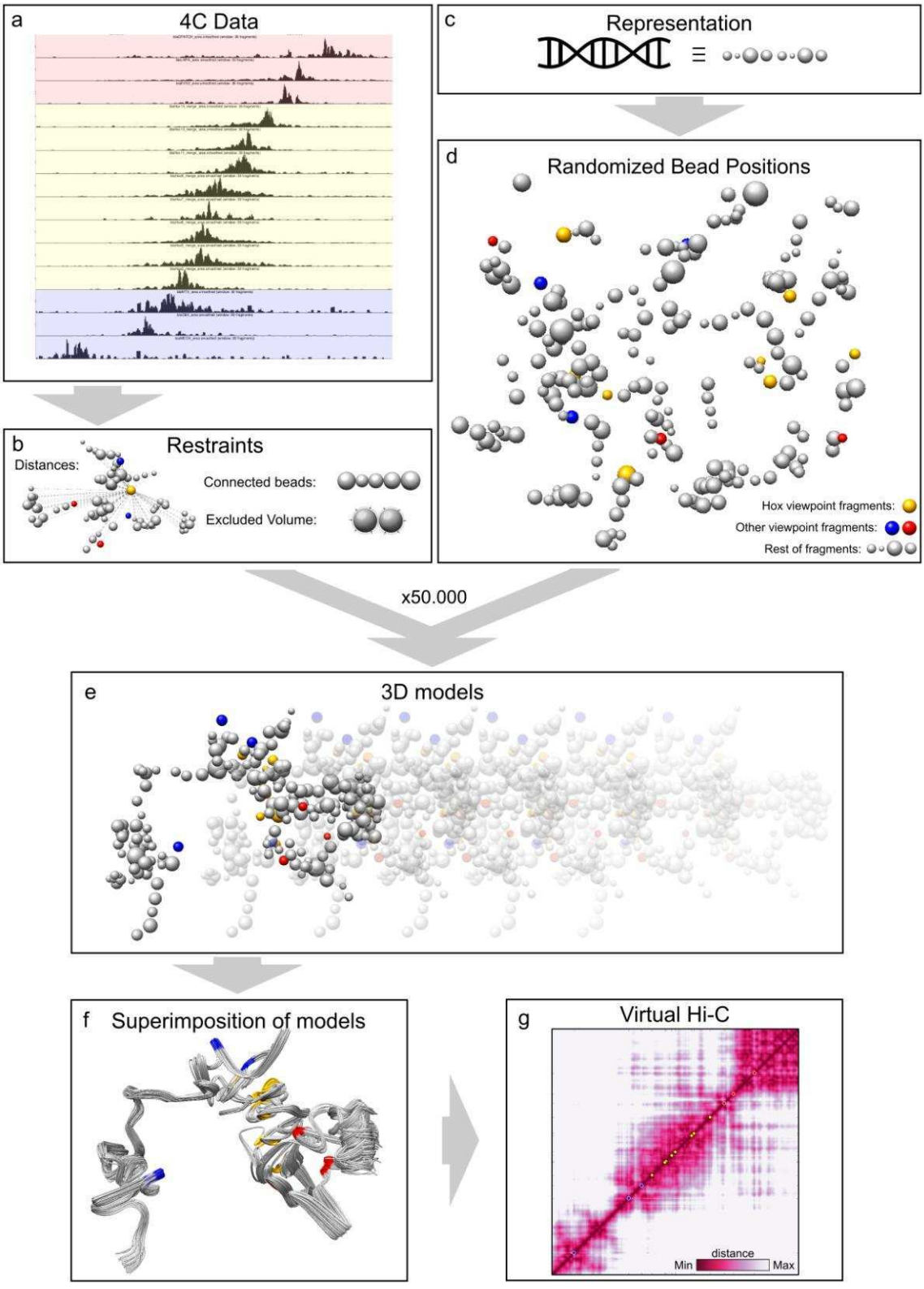
Nature Genetics: doi:10.1038/ng.3497



Supplementary Figure 6

Temporal dynamics of the 4C-seq interaction profiles of amphioxus Hox genes during development.

The amphioxus Hox genomic region showing one replicate for each of the 4C-seq profiles of the promoters of several Hox genes. The three different developmental stages are colored in red (8 h.p.f. gastrula), black (15 h.p.f. early neurula) and blue (36 h.p.f. larva). Dashed lines indicate the positions of the viewpoints.

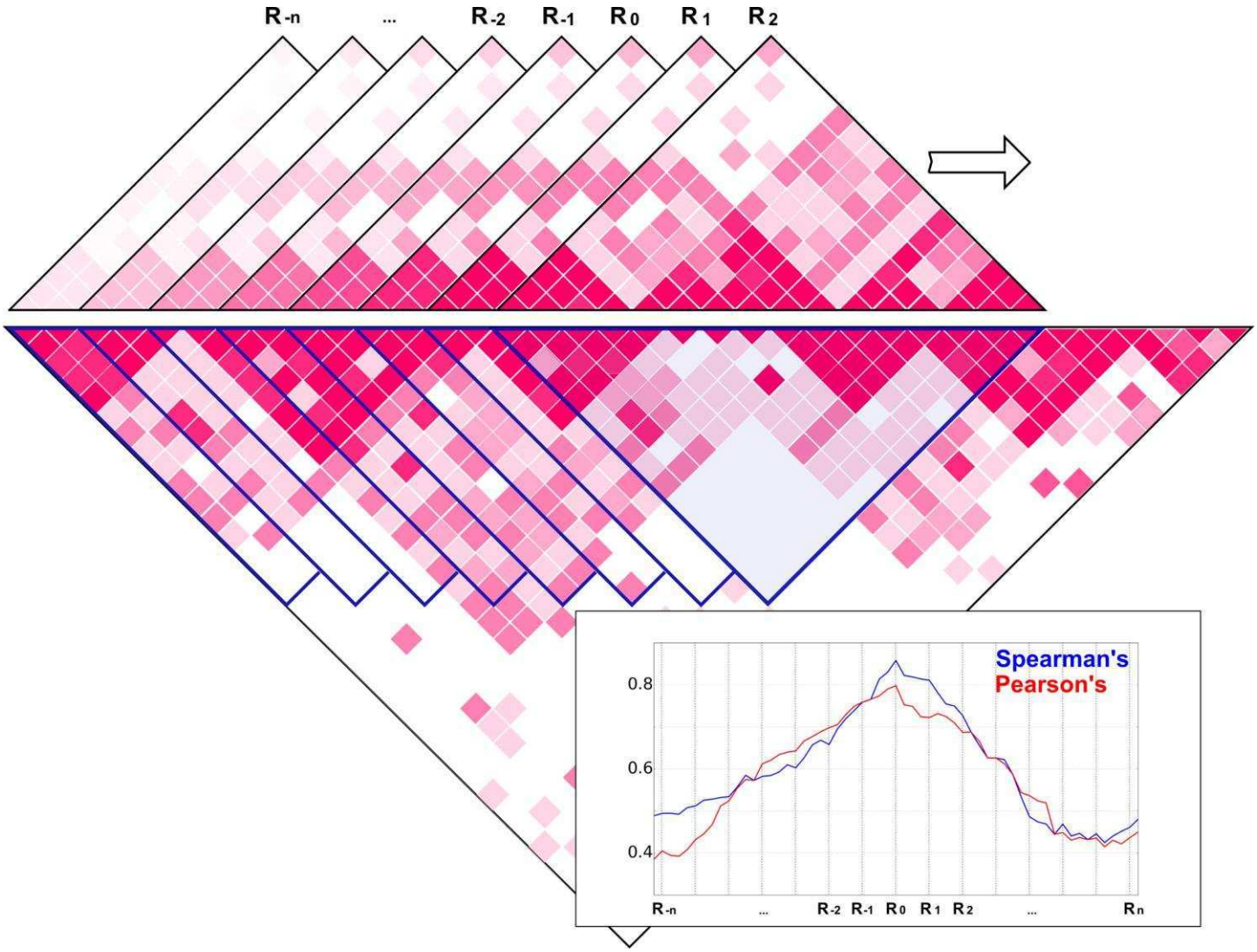


Supplementary Figure 7

Flowchart describing the generation of the models and the virtual Hi-C data.

(**a**,**b**) 4C data (**a**) were translated into distance restraints that were added to the rest of the restraints (**b**). (**c**) The genome was represented as concatenated beads of different size that represented 20 fragments. The size was proportional to the sum of the read counts. (**d**) Models were optimized, starting from randomized bead positions. (**e**) After 50,000 iterations, we selected the 200 models with the best score. These models were clustered on the basis of their RMSD. (**f**,**g**) The models from the most populated mirror image cluster were superimposed (**f**) and the virtual Hi-C heat map was generated by calculating the mean distance between all the beads from all the models (**g**).

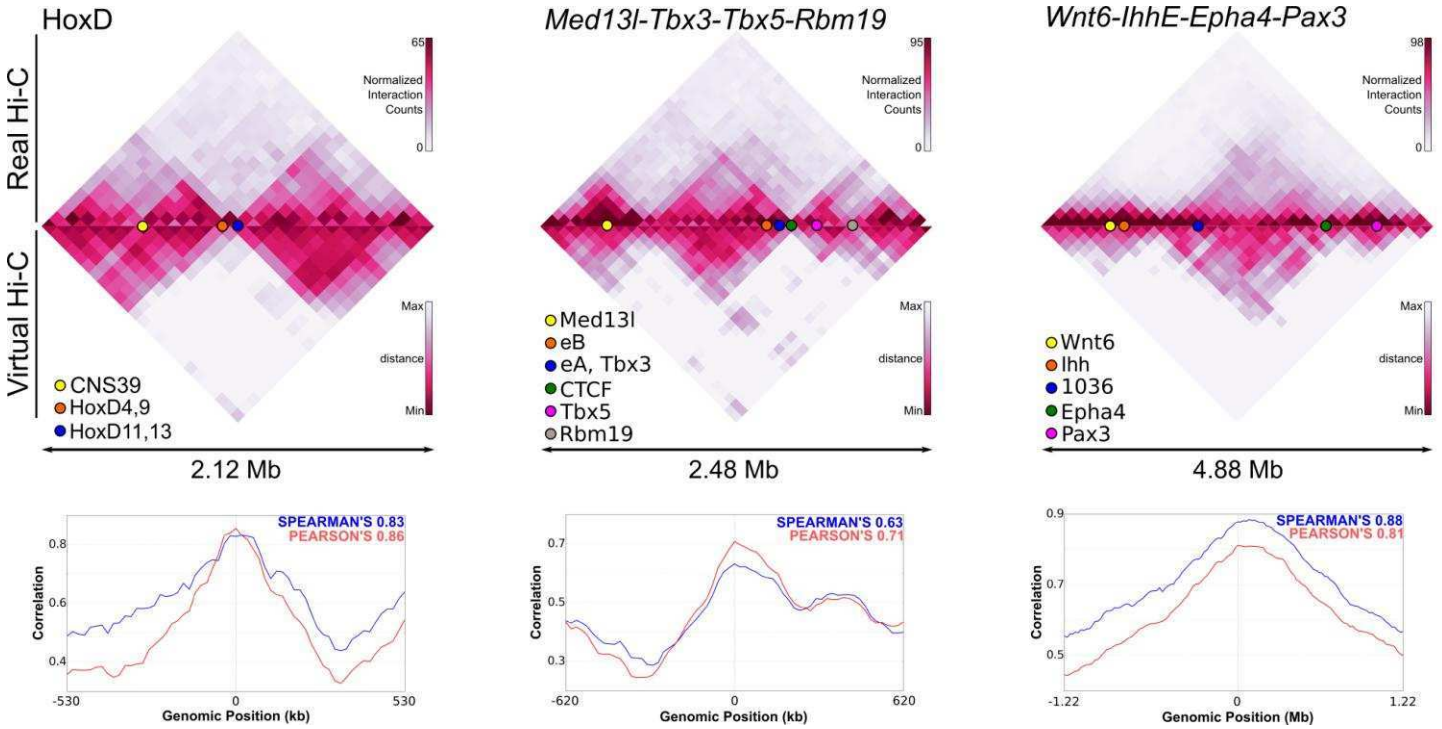
Nature Genetics: doi:10.1038/ng.3497



Supplementary Figure 8

Explanatory cartoon for Hi-C comparisons using the shifting alignment approach.

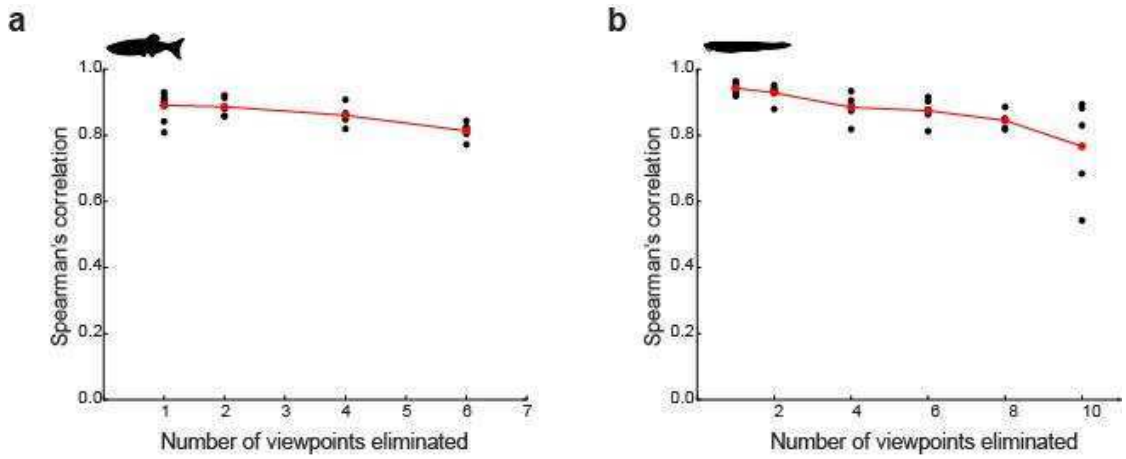
The alignment of the Hi-C matrices being compared is iteratively shifted by bins of 20 kb (reaching ±25% of the total size of the matrix) to obtain a collection of 'mock' coefficients corresponding to misaligned Hi-C maps (from R_{-n} to R_n), together with the coefficient of the correctly aligned comparison (R_0). In this situation, R_0 is expected to be the highest coefficient.



Supplementary Figure 9

Spearman's and Pearson's correlation comparisons between the experimental and virtual Hi-C data of different loci (HoxD; *Med13l-Tbx3-Tbx5-Rbm19*; and *Wnt6-Ihh-Epha4-Pax3*).

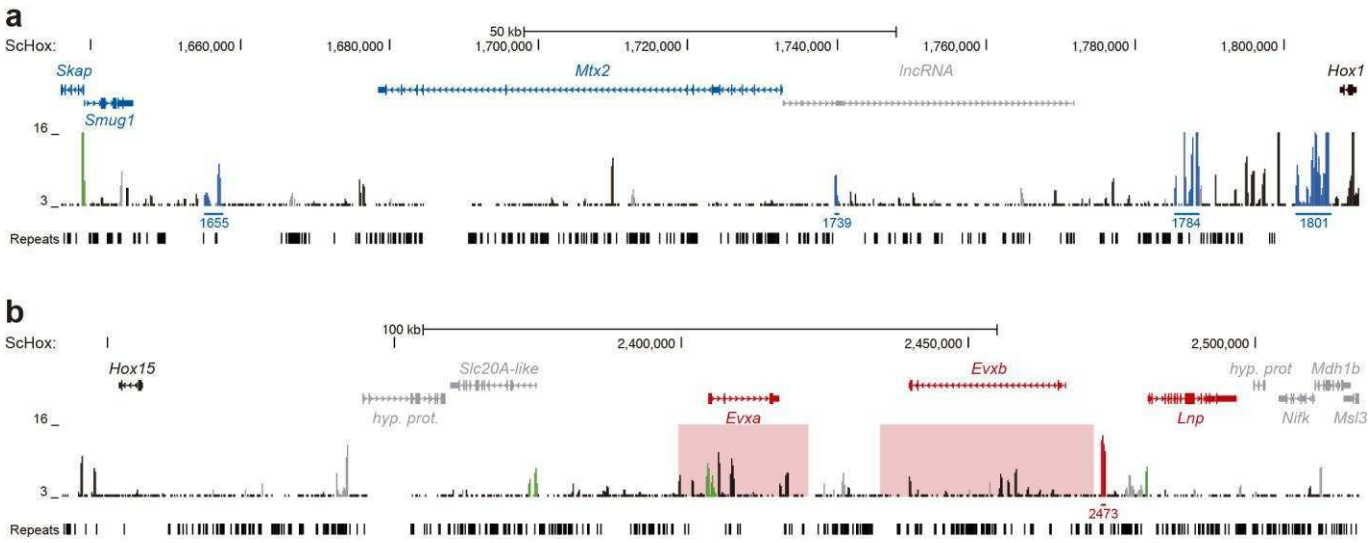
Correlations are gathered shifting the virtual Hi-C across the real Hi-C, as described in **Supplementary Figure 8**. Position 0 corresponds to correct alignment of the Hi-C matrices. The coefficients correspond to the values in the alignment at position 0.



Supplementary Figure 10

Virtual Hi-C jackknife resampling experiments.

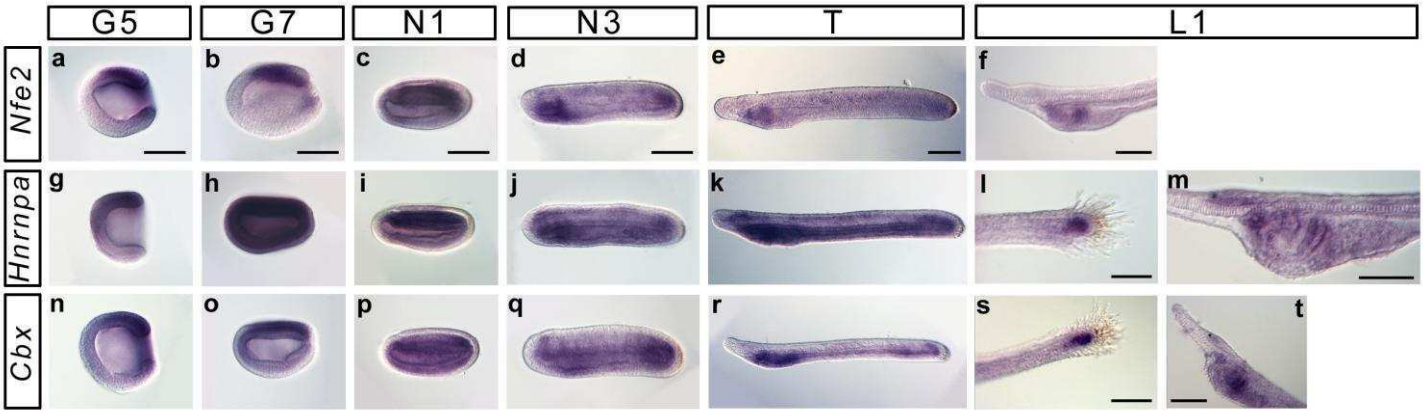
(**a**,**b**) Zebrafish (**a**) and amphioxus (**b**) Spearman's correlation coefficients between the virtual Hi-Cs obtained from the final 3D models and those resultant from the jackknife resampling experiments. For each number of viewpoints eliminated, five different combinations of viewpoint subsets were randomly generated and compared, except in those eliminating a single viewpoint, where all possible combinations were assayed. Black circles correspond to the correlation coefficients of each individual comparison (**Supplementary Table 3**), and red circles indicate the average for each resampling category.



Supplementary Figure 11

ATAC-seq signal distribution in amphioxus Hox-neighboring regions.

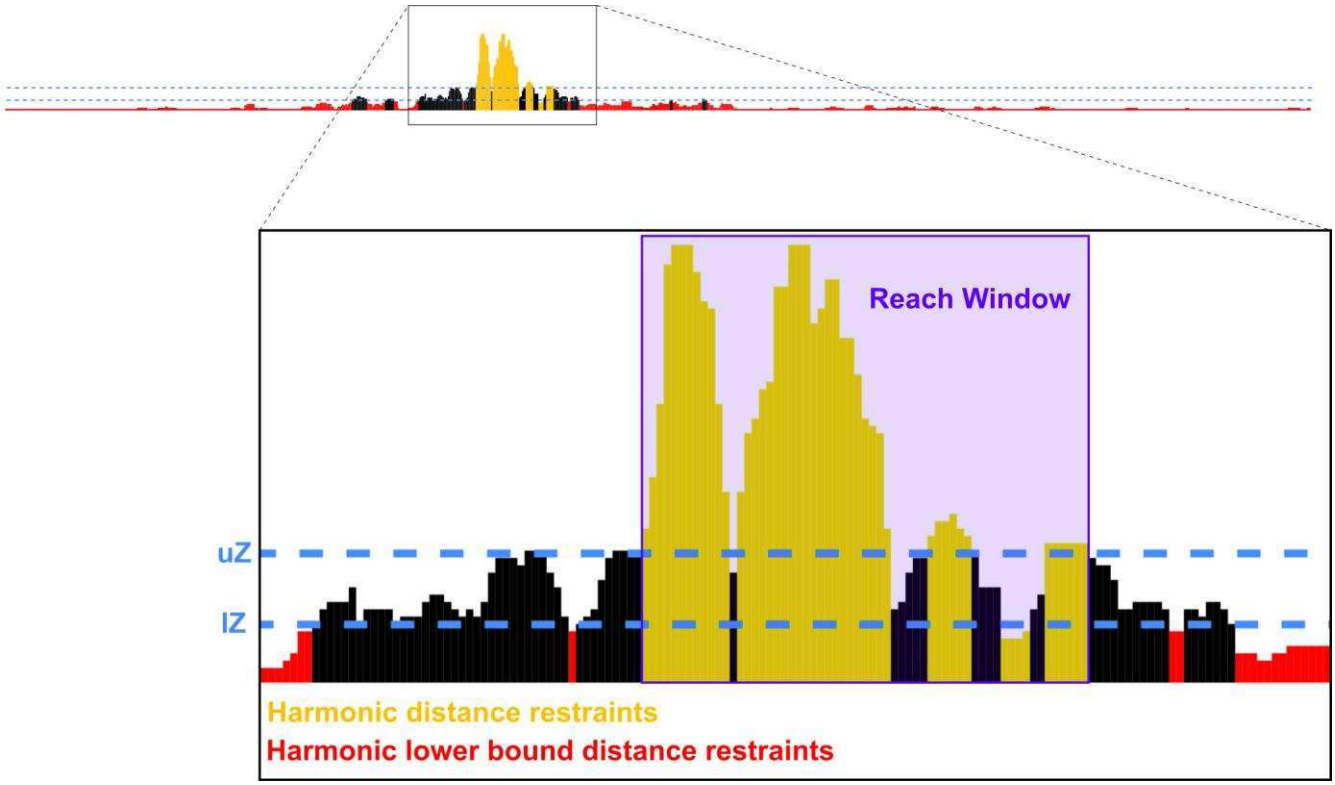
(**a**,**b**) ATAC-seq profiles of the anterior (**a**) and posterior (**b**) regions showing accessible chromatin regions. The ATAC signal is depicted in black, except in cases having overlap with repetitive elements (gray) or transcriptional start sites (green). Elements tested in reporter assays are colored in blue (anterior) and red (posterior). Regions overlapping with the Evx loci ±5 kb are shaded in light red.



Supplementary Figure 12

Whole-mount *in situ* hybridization of amphioxus Hox-neighboring genes.

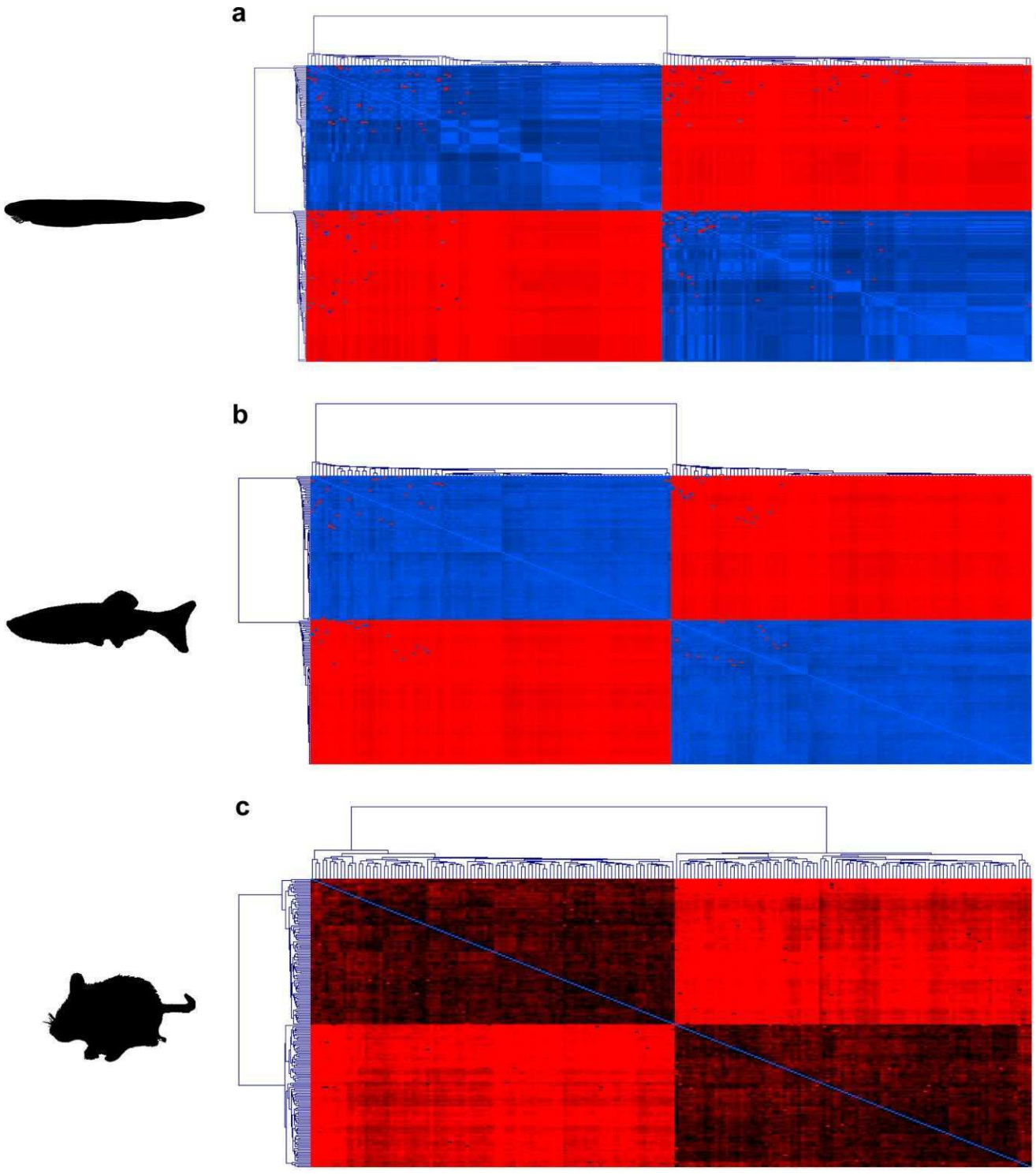
(**a**–**f**) *Nfe2* expression pattern. No regionalized expression could be detected at the 8-cell, blastula and G3 stages. (**a**,**b**) In G5 (**a**) and G7 (**b**) embryos, *Nfe2* is expressed in the mesendoderm of the blastoporal lip and in the presumptive neural plate region. (**c**) At the N1 stage, *Nfe2* is expressed in the mesoderm and neural plate. (**d**) In N3 stage embryos, expression is detected in the endoderm of the pharynx and in the tailbud. (**e**) At the T stage, *Nfe2* is expressed in the endoderm of the pharynx (forming club-shaped gland and preoral pit). (**f**) This expression is maintained in L1 stage larva. (**g**,**h**) *Hnrnpa* expression pattern. No regionalized expression could be detected at the 8-cell, blastula and G3 stages. (**g**) At the G5 stage, *Hnrnpa* is expressed in the dorsal blastoporal lip. (**h**) Then, in G7 embryos, expression is ubiquitous. (**i**) In N1 neurula, expression is detected in the mesoderm and in the neural plate. (**j**) In later N3 neurula stage embryos, expression is ubiquitous in the mesoderm and endoderm but a stronger level of expression is observed in the pharynx and in the tailbud. *Hnrnpa* is also expressed in the cerebral vesicle at this stage. (**k**) At the T stage, expression is observed in the whole gut, in the cerebral vesicle and in some neurons of the neural tube, as well as in the posterior notochord. (**l**,**m**) In L1 larvae, *Hnrnpa* is expressed in the taibud (**l**) and at a lower level in the cerebral vesicle, the club-shaped gland and preoral pit (**m**). (**n**–**t**) *Cbx1-3-5* expression pattern. No regionalized expression could be detected at the 8-cell, blastula and G3 stages. (**n**) In G5 stage gastrulae, expression is detected in the mesendoderm of the blastopral lip and in the dorsal ectoderm. (**o**) At G7 stage, *Cbx1-3-5* is expressed in the mesendoderm and in the presumptive neural plate. (**p**) In N1 neurulae, expression is observed in the mesendoderm and neural plate. (**q**–**t**) From N3 to L1, expression is similar to what is observed for *Hnrnpa*.



Supplementary Figure 13

Translation of the 4C-seq data into distance restraints.

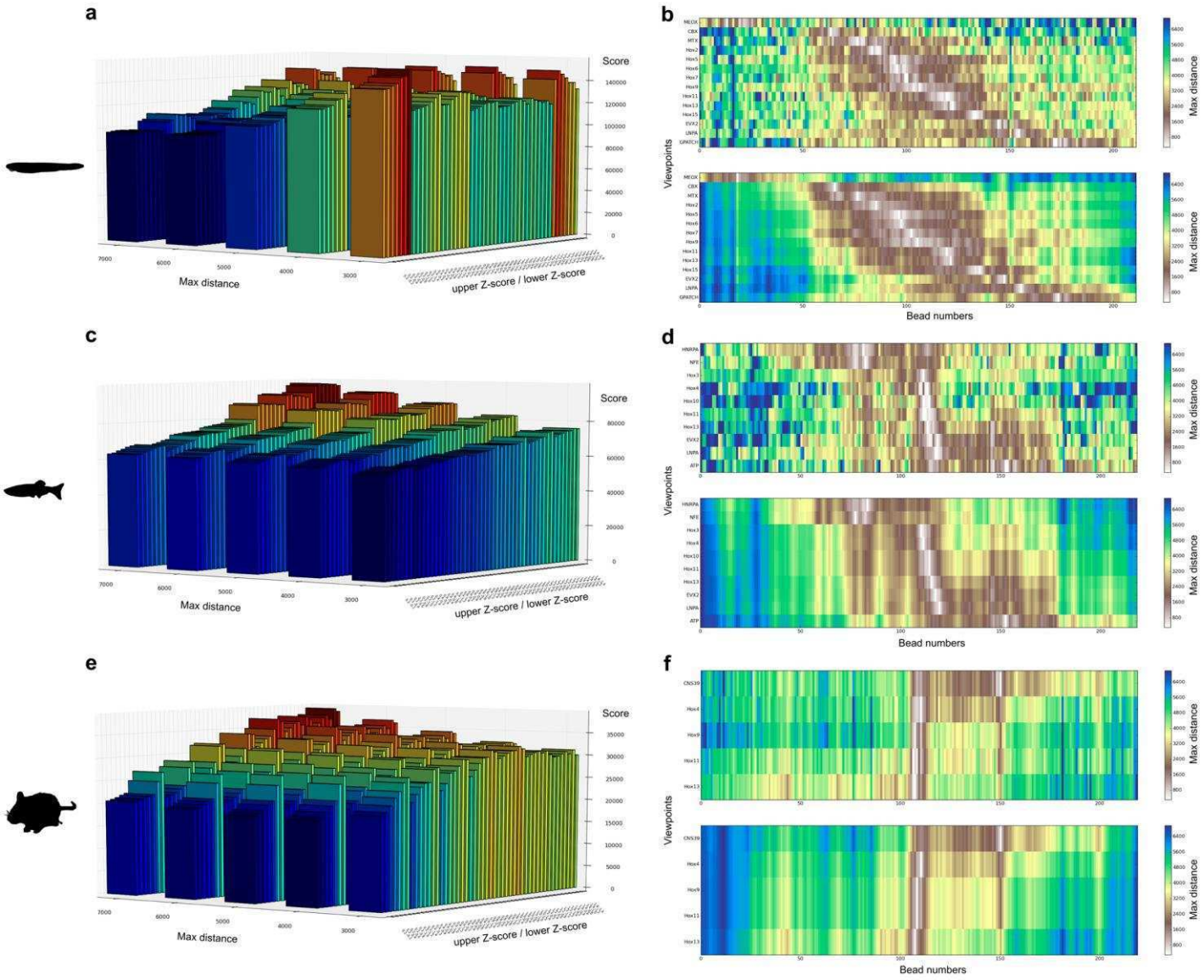
We used the upper *Z* score and lower *Z* score (u*Z* and l*Z*, represented as dashed lines in blue). Statistically significant data were defined as the ones above the u*Z* and below the l*Z*. On the basis of these boundaries, the reach window was established (in purple), an area that covers all fragments between the first (upstream) and last (downstream) fragment above the u*Z*. Those read counts inside the reach window above the u*Z* or below the l*Z* were translated as harmonic distance restraints (yellow), and the rest (red) were translated as harmonic lower-bound distance restraints.



Supplementary Figure 14

Comparisons of amphioxus, zebrafish and mouse models.

(**a**–**c**) Heat map plots showing the RMSD of the 200 models compared between them. All models of each species were clustered in one of the two mirror-image clusters. Blue squares stand for an RMSD of 0 Å. Red squares are for maximum RSMD.



Supplementary Figure 15

Comparisons of different sets of bins for each species.

(**a**,**c**,**e**) 3D bar plots comparing the scores of different sets of bins with the maximum distance, the upper *Z* score and the lower *Z* score as parameters. The gradient of colors depends on the score, from blue (lowest) to red (highest). The score is the mean of the sum of the differences between the calculated distance for each bead from the 4C data and the measured distance in each model. The lower the score, the smaller the difference between the models and the 4C data. (**b**,**d**,**f**) Heat maps comparing the computed distances from the 4C data for each bead and the mean of the measured distances of the models with the best set of bins: 7,000 for maximum distance, 0.2 for u*Z* and –0.2 for the l*Z*.

4

Regulatory landscape fusion in rhabdomyosarcoma through interactions between the PAX3 promoter and FOXO1 regulatory elements

Genome Biology

Regulatory landscape fusion in rhabdomyosarcoma through interactions between the *PAX3* promoter and *FOXO1* regulatory elements

Vicente-García *et al.*

BioMed Central

Vicente-García *et al. Genome Biology* (2017) 18:106
DOI 10.1186/s13059-017-1225-z

Vicente-García et al. Genome Biology (2017) 18:106
DOI 10.1186/s13059-017-1225-z

Genome Biology

RESEARCH Open Access

Regulatory landscape fusion in rhabdomyosarcoma through interactions between the *PAX3* promoter and *FOXO1* regulatory elements

Cristina Vicente-García[1†], Barbara Villarejo-Balcells[2†], Ibai Irastorza-Azcárate[1], Silvia Naranjo[1], Rafael D. Acemel[1], Juan J. Tena[1], Peter W. J. Rigby[2], Damien P. Devos[1], Jose L. Gómez-Skarmeta[1] and Jaime J. Carvajal[1*]

Abstract

Background: The organisation of vertebrate genomes into topologically associating domains (TADs) is believed to facilitate the regulation of the genes located within them. A remaining question is whether TAD organisation is achieved through the interactions of the regulatory elements within them or if these interactions are favoured by the pre-existence of TADs. If the latter is true, the fusion of two independent TADs should result in the rewiring of the transcriptional landscape and the generation of ectopic contacts.

Results: We show that interactions within the *PAX3* and *FOXO1* domains are restricted to their respective TADs in normal conditions, while in a patient-derived alveolar rhabdomyosarcoma cell line, harbouring the diagnostic t(2;13)(q35;q14) translocation that brings together the *PAX3* and *FOXO1* genes, the *PAX3* promoter interacts ectopically with *FOXO1* sequences. Using a combination of 4C-seq datasets, we have modelled the three-dimensional organisation of the fused landscape in alveolar rhabdomyosarcoma.

Conclusions: The chromosomal translocation that leads to alveolar rhabdomyosarcoma development generates a novel TAD that is likely to favour ectopic *PAX3:FOXO1* oncogene activation in non-*PAX3* territories. Rhabdomyosarcomas may therefore arise from cells which do not normally express *PAX3*. The borders of this novel TAD correspond to the original 5'- and 3'- borders of the *PAX3* and *FOXO1* TADs, respectively, suggesting that TAD organisation precedes the formation of regulatory long-range interactions. Our results demonstrate that, upon translocation, novel regulatory landscapes are formed allowing new intra-TAD interactions between the original loci involved.

Keywords: TAD, CTCF, Transcriptional regulation, *FOXO1*, *PAX3*, Alveolar rhabdomyosarcoma, 4C-seq

Background

The advent of chromatin conformation capture technologies (3C and its variants Hi-C, 5C-seq and 4C-seq; reviewed in [1]) has been essential in the identification of megabase-scale chromosomal organisation domains [2–4], which have been termed topologically associating domains (TADs). These are large genome intervals defined by an increased number of long-range chromatin interactions between the loci contained in the same chromosomal domain and a decreased number of interactions with loci in neighbouring domains [5]. Increasing experimental evidence suggests that TADs constitute not only structural but also functional units of the genome. TADs structurally restrain epigenetic domains [2–4], domains that can change coordinately in response to external cues [6]. Furthermore, the genome has been divided into compartments with active or inactive status [7], and during differentiation, regions subject to repositioning from one of these compartments to the other correspond to single or several, consecutive TADs [8, 9]. Therefore, the genes contained within a TAD, as a group, are more or less prone to transcription depending on the epigenetic

* Correspondence: j.carvajal@csic.es
†Equal contributors
[1]Centro Andaluz de Biología del Desarrollo (CABD), CSIC-UPO-JA, Universidad Pablo de Olavide, Carretera de Utrera km1, 41013 Seville, Spain
Full list of author information is available at the end of the article

state of the domain or the nuclear compartment in which they are positioned. In fact, genes within TADs do show gene expression correlation [3, 6], revealing an underlying mechanism of intra-TAD gene regulation, which does not necessarily imply that genes included within a TAD are under the control of the same tissue-specific enhancers.

From an evolutionary point of view, it has been shown that ancestral recombinations leading to loss of synteny occur at TAD borders [10], maintaining their structures and indicating that TADs are under positive selective forces, most likely because the disruption of a TAD has deleterious effects on the regulation of the genes within it. It is still not clear if TADs originate from interactions between enhancers and promoters within the domain or if it is this compartmentalisation that permits and restricts enhancer-promoter contacts [11–13].

The molecular nature of TAD borders is still unclear, although it has been shown that they are enriched in binding sites for the CTCF protein [2, 3], which has been implicated in three-dimensional (3D) chromatin organisation and enhancer-blocking activities [14]. The directionality of the CTCF binding sites seems to be predictive of their loop-forming activity as deletion or inversion of these sites results in the generation of inappropriate enhancer-promoter contacts [15, 16].

A remaining question is how sequence interactions are restricted to individual domains. The borders between adjacent TADs seem to restrict cross-border interactions and thus deletion of these regions results in the misregulation of the genes associated with them. Genome manipulations of the border separating the *Tfap2c* and *Bmp7* loci in the mouse show 'contamination' of the transcriptional landscapes of both genes upon inversion [17], while human disorders such as polydactyly, brachydactyly and F-syndrome have been shown to be related to the deletion, inversion or duplication of borders separating the different TADs containing the *WNT6-IHH/EPHA4/PAX3* loci [18], which leads to otherwise prohibited promoter contacts with enhancer elements located outside their cognate TAD, causing mis-expression of the genes involved. Analyses of various duplications in the proximity of the *SOX9* locus have shown several outcomes depending on the exact nature of the duplication: intra-TAD duplications do not alter overall TAD organisation but may result in increased numbers of intra-TAD contacts and could give rise to a phenotype; and inter-TAD duplications that cross TAD borders generate novel TADs without altering flanking gene expression. In this second case, a phenotype could arise if the novel regulatory landscape created by the duplication includes a coding gene, as it could result in its dysregulation [19].

Thus, the implication is that removal of a border element results in the fusion of adjacent TADs, while the inversion/duplication of a border could allow new regulatory interactions to be formed resulting in inappropriate expression of genes around the inversion/duplication. Importantly, sequences adjacent to the manipulated borders are also rearranged during the process and thus a possible contribution to the observed phenotypes cannot be discarded. Other human chromosomal rearrangements have been shown to result in the dysregulation of gene expression by regulatory elements located in the proximity of the breakpoints (e.g. [20–26]).

Recurrent chromosomal translocations are formed by end-joining of two double-strand chromosomal breaks, which occasionally occur within the introns of individual genes resulting in the generation of a novel chimaeric fusion protein harbouring functional domains from the two proteins and thus new functional properties. In cancer, the formation of novel chimaeric transcription factors, in which the DNA binding domain is encoded by one gene and the transactivation domain is encoded by the other, is common. The *PAX3:FOXO1* fusion gene, arising from the t(2;13)(q35;q14) translocation [27] in the paediatric soft tissue tumour alveolar rhabdomyosarcoma (ARMS), encodes a transcription factor that contains the *PAX3* (paired box 3) DNA-binding domain and the *FOXO1* (forkhead box O1) transactivation domain. This fusion transcription factor dysregulates *PAX3* target genes resulting in gene expression changes that modify pathways involved in proliferation and/or survival, contributing to tumour initiation. Translocations involving *PAX3* (or the closely related *PAX7*) and *FOXO1* are only found in rhabdomyosarcomas. This permits the formulation of two hypotheses: (1) that translocations can occur in multiple cell types but only those expressing the regulatory factors required for the expression of the oncogene give rise to rhabdomyosarcomas; or (2) that the translocations occur in a restricted or unique cell type, usually by means of co-transcription of the two loci involved in the translocation [28, 29]. Even if this second hypothesis turns out to be correct, it is still possible that only those cells that express the correct combination of transcription factors would give rise to tumour cells as the fusion gene will be under the transcriptional control of specific regulatory elements; oncogene activation in a non-*PAX3*-expressing cell type may therefore be essential for the development of the disease. It is thus clear that unravelling the transcriptional regulatory mechanisms of *PAX3*, *FOXO1* and the oncogenic *PAX3:FOXO1* gene should help to identify the elusive cell type of origin for these sarcomas.

Crucially, we show that the t(2;13)(q35;q14) translocation in ARMS not only generates a fusion gene but also a novel fused regulatory landscape that likely controls the expression of the novel gene. The translocation results in the formation of a novel TAD structure that retains the 5' and 3' borders of the wild-type *PAX3* and *FOXO1* TADs,

respectively. Importantly, interactions between the *PAX3* promoter and the *FOXO1* region are similar to those established by the *FOXO1* promoter in its own locus, despite these regulatory regions being in a completely new regulatory landscape. As these interactions are novel, if the establishment of regulatory interactions were to precede TAD formation, we would expect a change in TAD boundaries. Instead, we observe that in the ARMS translocation analysed, the *PAX3* promoter does not interact with sequences downstream of the original *FOXO1* TAD border.

Results

Loss of synteny analyses place the 5' boundary of the *FOXO1/FoxO1* locus in close proximity to its promoter

One of the major unknowns in the study of ARMS is the nature of the cell that originally suffered the *PAX3:FOXO1* chromosomal translocation leading to tumour development. We hypothesised that in the translocated chromosome the fusion gene would be under the control of both *PAX3* and *FOXO1* regulatory elements. For this reason, we first determined the maintenance of synteny surrounding the *FoxO1* locus as an approach to establish the existence of strong constraints on genomic rearrangements as a proxy for the presence of essential *FOXO1* regulatory regions. With the exception of ray-finned fishes, which experienced a whole genome duplication (*D. rerio*, *O. latipes* and *G. aculeatus*; Additional file 1: Figure S1), and rodents (*M. musculus* and *R. rattus*), all species analysed (mammals, birds, amphibians and reptiles) share the same chromosomal structure flanking *FOXO1* (*MRPS31-FOXO1-COG6-LHFP*; Table 1), a structure that has been conserved for at least 450 Mya. The break of synteny upstream of *FoxO1* detected in rodents places the ancestral recombination event in this group between *MRPS31* and *FOXO1* (Fig. 1). Analysis of evolutionarily conserved regions (ECRs) upstream of mouse *FoxO1* shows that a conserved region 47 kb upstream of the gene maps immediately upstream of the human *MAML3* gene on Chr4, while another ECR, located 17 kb upstream of mouse *FoxO1* maps upstream of the human *FOXO1* gene on Chr 13 (Additional file 1: Figure S2). This analysis restricts the ancestral recombination event somewhere in the -17 kb to -47 kb interval upstream of *FOXO1*.

In the case of the *Pax3* locus, the same gene organisation was found in all species analysed: *FARSB-SGPP2-PAX3-EPHA4*. Since no breaks in synteny were observed, no conclusions could be drawn on the span of *Pax3* regulatory elements in the locus but it suggests that strong evolutionary constrains have maintained this syntenic block unaltered.

Hi-C and 4C-seq analyses of the *PAX3/Pax3* and *FOXO1/FoxO1* loci

We then made use of published Hi-C data on human [2] and mouse [5] ES cells, which show that the mouse *FoxO1* gene is included within a single TAD (Fig. 2a), as defined by directionality index analysis (D.I.; 2). Despite the break of synteny immediately upstream of *FOXO1/FoxO1*, the TADs have been maintained in the two species, with similar upstream and downstream borders indicating that the ancestral recombination that gave rise to the

Table 1 Location of genes flanking the *FOXO1* locus in human Chr13 across species

	LHFP	*COG6*	*FOXO1*	*MRPS31*	
Homo sapiens	Chr 13	Chr 13	Chr 13	Chr 13	Human
Macaca mulatta	Chr 17	Chr 17	Chr 17	Chr 17	Macaque
Callithrix jacchus	Chr 5	Chr 5	Chr 5	Chr 5	Marmoset
Canis lupus familiaris	Chr 25	Chr 25	Chr 25	Chr 25	Dog
Monodelphis domestica	Chr 4	Chr 4	Chr 4	Chr 4	Opossum
Mus musculus	Chr 3	Chr 3	Chr 3	Chr 8	Mouse
Rattus norvegicus	Chr 2	Chr 2	Chr 2	Chr 16	Rat
Gallus gallus	Chr 1	Chr 1	Chr 1	Chr 1	Chicken
Alligator mississippiensis	JH731763	JH731763	JH731763	JH731763	American alligator
Xenopus tropicalis	GL172869	GL172869	GL172869	GL172869	Clawed frog
Latimeria chalumnae	JH129255	JH129255	JH127414	JH127414	Coalecanth
Danio rerio	Chr 10/15	Chr 15	Chr 10/15	Chr 5	Zebrafish
Oryzias latipes	Chr 13	Chr 13	Chr 13/14	Chr 14	Medaka
Gasterosteus aculeatus	Group I	Group I	Group I/VII	Group VII	Stickleback
Callorhinchus milli	KI635872	KI635872	KI635872	KI635872	Elephant shark

Gene names are on the top row, animal species on the left column, common names on the right column. In bold, genes mapping to a different syntenic region. The Coelacanth (*L. chalumnae*) genome is fractioned at present and thus it is not possible to ascertain if the *LHFP/COG6* and *FOXO1/MRPS31* scaffolds are contiguous

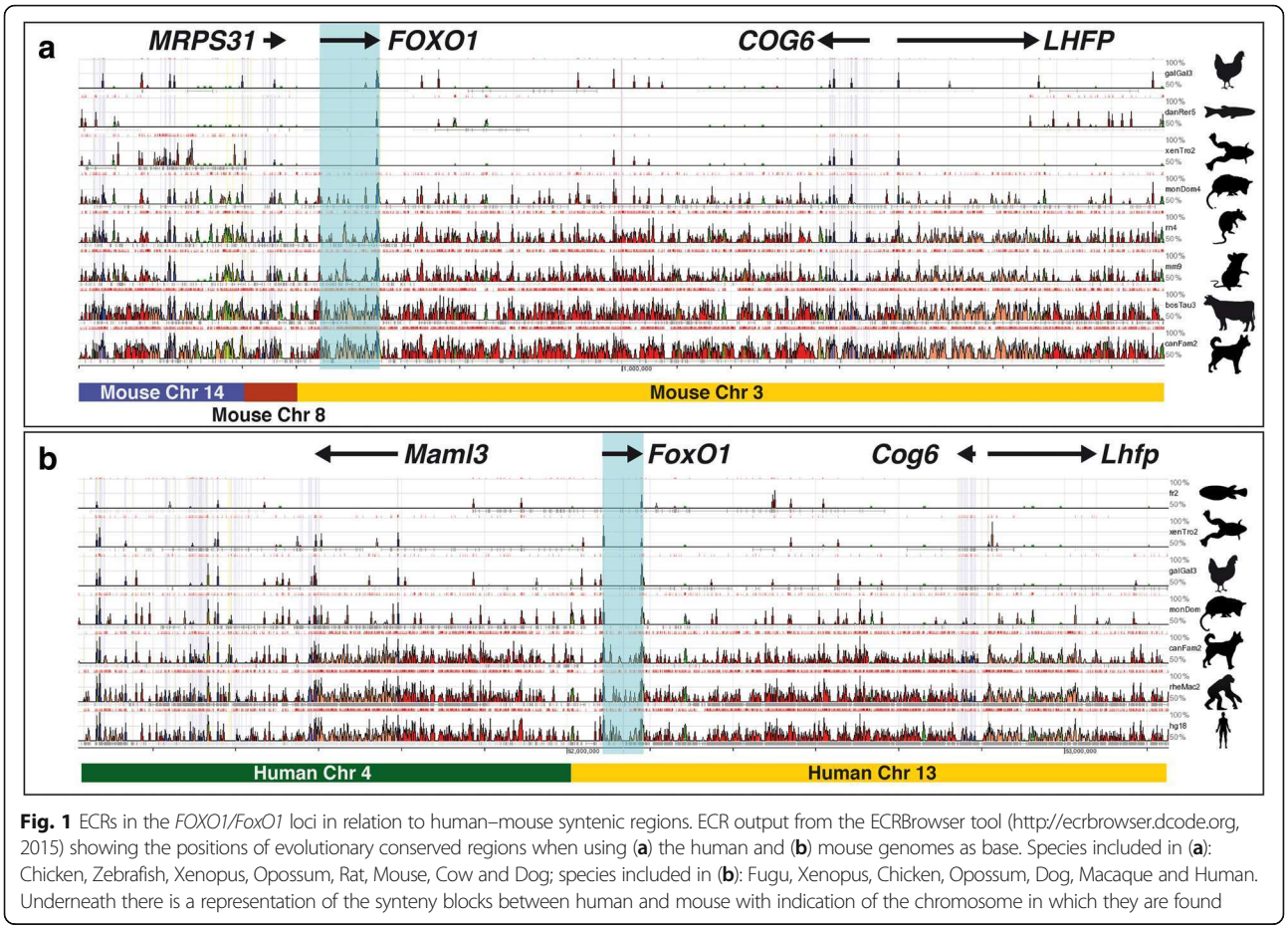


Fig. 1 ECRs in the *FOXO1/FoxO1* loci in relation to human–mouse syntenic regions. ECR output from the ECRBrowser tool (http://ecrbrowser.dcode.org, 2015) showing the positions of evolutionary conserved regions when using (**a**) the human and (**b**) mouse genomes as base. Species included in (**a**): Chicken, Zebrafish, Xenopus, Opossum, Rat, Mouse, Cow and Dog; species included in (**b**): Fugu, Xenopus, Chicken, Opossum, Dog, Macaque and Human. Underneath there is a representation of the synteny blocks between human and mouse with indication of the chromosome in which they are found

synteny break occurred at the TAD border, as shown for other loci [10]. *PAX3/Pax3* are also located in identical TADs in the two species, containing the *SGPP2* and *FARSB* genes and being separated from the *EPHA4* regulatory landscape (Fig. 2b). Our analysis shows the existence of a TAD boundary immediately upstream of *PAX3* in both species. Nevertheless, the Hi-C data reveal extensive contacts between the two domains separated by this putative TAD boundary, suggesting these two domains correspond to sub-TAD structures rather than individual TADs.

We sought to further explore the regulatory landscape for these genes by performing 4C-seq on 9.5 d*pc* (days *post coitum*) whole mouse embryos using the *FoxO1* and *Pax3* promoters as viewpoints. At this developmental stage, both genes are expressed in a variety of progenitor and differentiated cells and thus the 4C-seq data represent an average through different cell types, although overall TAD organisation is mainly invariant across multiple tissues [2, 30]. The data show that interactions of the mouse *Pax3* promoter are almost equally distributed on either side (44% and 56%) and mainly restricted to the TAD that contains it (80.0%), further supporting the hypothesis that the identified boundary immediately upstream of *Pax3* corresponds to a sub-TAD boundary, with *Pax3* regulatory elements being present in both domains. The mouse *FoxO1* promoter interacts preferentially with downstream sequences (67.5%), mainly restricted to the TAD (65.6%); sequences that coincide with H3K27ac active-enhancer marks (Additional file 1: Figure S3) are detected in multiple tissues known to express *FoxO1* [31].

If *FOXO1* enhancer regions are involved in the regulation of the *PAX3:FOXO1* fusion gene, then first we had to gain an insight on the transcriptional regulation of the *FOXO1* gene, identify some of these regions and show that they might be located downstream of the translocation breakpoints.

Identification of translocation breakpoints in different ARMS cell lines

In ARMS, the t(2;13)(q35;q14) translocation occurs between intron 1 of *FOXO1* and intron 7 of *PAX3* [32–34]. In order to determine the contribution of putative enhancer elements translocated to the derivative t(2;13) chromosome towards the new regulatory landscape, we mapped six independent breakpoints in five independent

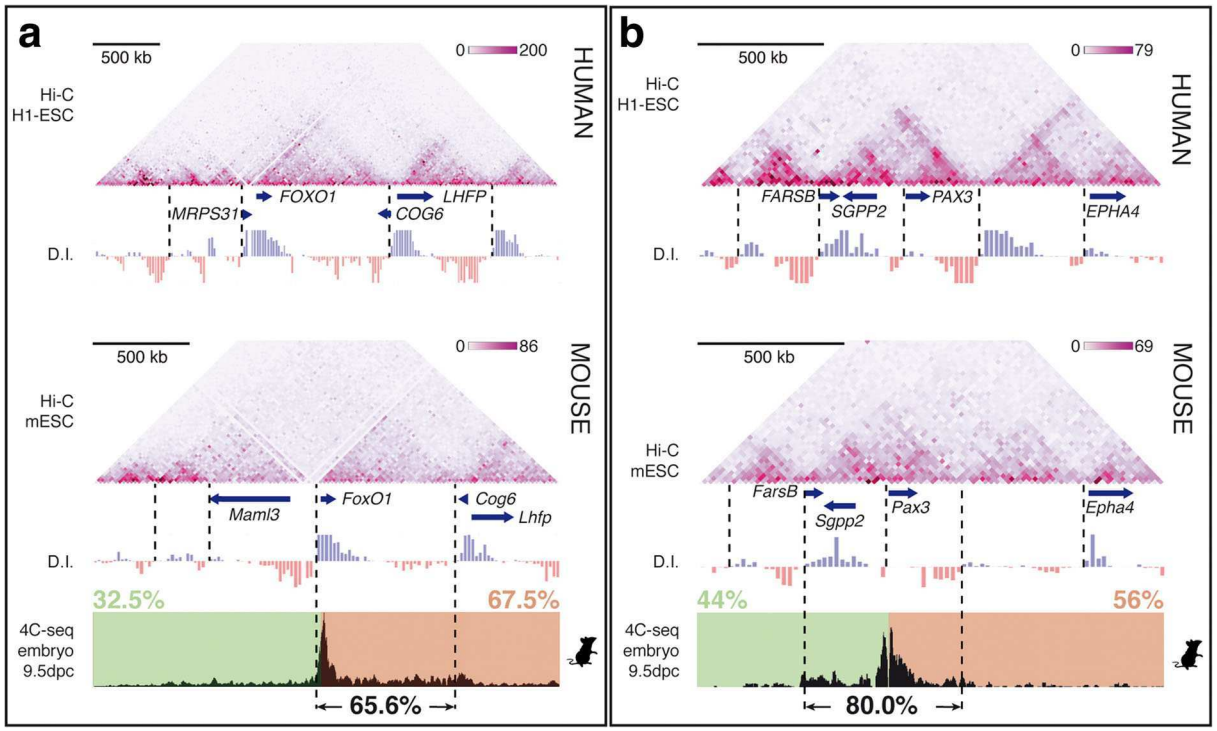


Fig. 2 *FoxO1* and *Pax3* promoter interactions are restricted to their respective TADs. (**a**) *FOXO1/FoxO1* and (**b**) *PAX3/Pax3* Hi-C data (2, 5; 20 kb resolution) coincide with the interactions detected by 4C-seq. The size and position of the *FOXO1* and *FoxO1* TADs, as determined by D.I. analyses, are very similar despite the ancestral recombinations that occurred immediately upstream of the genes. Names and positions of the genes in the different loci are indicated. The transcriptional direction of the different genes is indicated. The percentage of accumulated reads 1 Mb at either side of the *FoxO1* and *Pax3* promoters and within the TADs are shown. In the human genome, the positions of the calculated borders for (**a**) the *FOXO1* locus are (*left* to *right*) Chr13: 41,902,000; Chr13: 41,362,000; Chr13: 40,262,000 and Chr13: 39,502,000 and for (**b**) the *PAX3* locus are (*left* to *right*) Chr2: 223,817,756; Chr2: 223,511,756; Chr2: 223,171,756 (sub-TAD boundary); Chr2: 222,871,756 and Chr2: 222,451,756

ARMS cell lines harbouring this translocation. A series of forward primers around 3 kb apart from each other were designed to span the entire *PAX3* intron 7 (18.7 kb) while a series of reverse primers spaced by ~10 kb was designed to span the entire *FOXO1* intron 1 (104.7 kb) (Additional file 2: Table S1). Forward and reverse primers were used in all possible combinations in a long-distance polymerase chain reaction (LD-PCR) designed to amplify fragments up to 20 kb in length.

Sequence analyses of the SCMC and RH3 breakpoints showed a seamless transition between *PAX3* and *FOXO1* loci (Fig. 3a, b), although the exact point of the RH3 breakpoint cannot be ascertained as it occurs at a region of micro-homology between the two loci (TTA). The sequence of the RH5 breakpoint (Fig. 3c) showed a small amplification of three thymines at the junction between the *PAX3* and *FOXO1* loci. The RMS breakpoint (Fig. 3d) has a 22 bp insertion of a duplicated fragment from chromosome 13 immediately adjacent to the breakpoint. Finally, cell lines RH4 and RH41, derived from the same patient, show the same breakpoint containing a 4.9 kb insertion from chromosome 9 (Fig. 3e). We have previously reported the identification of the RH30 breakpoint [28].

Identification of regulatory regions driving transcription of the *FoxO1* and *Pax3* genes

For *FoxO1*, three overlapping bacterial artificial chromosomes (BACs) were selected from the Children's Hospital Oakland Research Institute (CHORI) library: RP23-66C15 (-116 kb to +104 kb, relative to the *FoxO1* transcriptional start site or TSS), RP24-330H17 (-61 kb to +104 kb) and RP23-96D10 (-38 kb to +148 kb) (Fig. 4a). We introduced a *lacZ* reporter gene at the first coding ATG of *FoxO1* and renamed them according to the lengths of their upstream spans (B116Z-Foxo1, B61Z-Foxo1 and B38Z-Foxo1, respectively). The 5'-end of B38Z-Foxo1 is located within the interval where the loss of synteny occurs and at the TAD border, while B116Z-Foxo1 and B61Z-Foxo1, with almost identical 3'-ends, cross it. We compared the expression patterns driven by these with that of the Foxo1Gt(AD0086)Wtsi gene-trap line ($Foxo1^{Gt\text{-}\beta\text{-}GEO/+}$; [31]).

As expected, all of them fail to recapitulate the complete *FoxO1* expression pattern because none of them contains the full regulatory landscape, which our 4C-seq data indicate spans up to 700 kb downstream of the gene. Interestingly, the B116Z-Foxo1 BAC construct drives ectopic expression in the neural tube

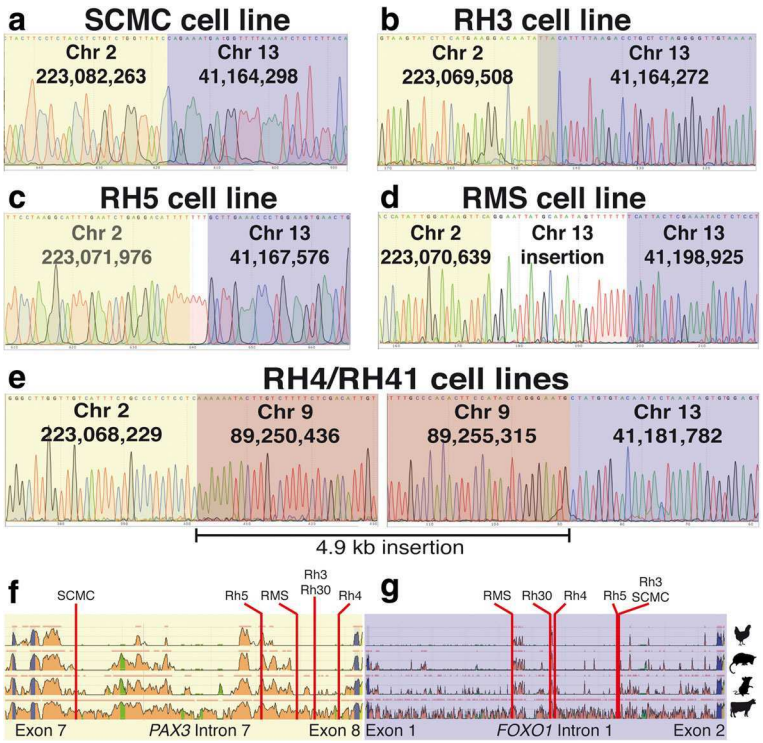


Fig. 3 Mapping ARMS translocations to the base pair level. Sequence tracks of the translocation breakpoints identified in five independent ARMS cell lines: (**a**) SCMC, (**b**) RH3, (**c**) RH5, (**d**) RMS and (**e**) RH41. In three cases, the translocation produces a clean cut between Chr2 (*yellow*) and Chr13 (*purple*) sequences. **e** In the RH41 cell line there is a clean insertion of a 4.9 kb fragment from Chr9 (*red*). The genome positions of the translocation breakpoints are provided (hg19). **f** Detail of the ECR Browser output (Chicken, Opossum, Mouse, Cow; base genome, Human) covering the genomic interval between exons 6 and 8 of *PAX3* showing the precise location of the mapped translocation breakpoints in intron 7. **g** Detail of the ECR Browser output covering the genomic interval between exons 1 and 2 of *FOXO1* showing the location of the mapped translocation breakpoints in intron 1

(Fig. 4b). Unlike B61Z-Foxo1, which also crosses the TAD border, B116Z-Foxo1 contains regions with strong active-enhancer marks in several tissues including some pertaining to the central nervous system. Thus, in this context, the sequence underlying this TAD border does not possess intrinsic transcriptional boundary activity per se because it is unable to block the interactions between regulatory elements and the promoter when placed in between them. Except for this remarkable difference, B61Z-Foxo1 and B116Z-Foxo1 drive very similar expression patterns from 9.5 d*pc* to the adult (note that their 3'-ends are almost identical; compare Additional file 1: Figures S4 and S5). Sites of expression include the myotome, fore-gut and hind-gut diverticula, the stomach, the apical ectodermal ridge (AER), limb, thoracic and facial skeletal muscle, the inner layer of the retina, the posterior wall of the lens vesicle, and the nasal pits. In contrast, the B38Z-Foxo1 construct drives expression from 9.0 d*pc* in vascular precursors throughout the embryo (Fig. 4b and Additional file 1: Figure S6). This finding indicates that a regulatory module for vasculature expression maps in the non-overlapping region between B61Z-Foxo1/B116Z-Foxo1 and B38Z-Foxo1, that is, +104 to +148 kb from the *FoxO1* TSS. Time course analyses of these transgenic lines revealed that all three constructs fail to recapitulate the complete *FoxO1* expression pattern (e.g. no expression is observed in brown adipose tissue -BAT- from 16.5 d*pc* onwards in any of the lines), indicating that the enhancer(s) responsible to drive BAT expression is not contained within these BAC clones.

In order to analyse *Pax3* gene expression, several BAC clones were identified from the CHORI library; for this study we selected RP23-260 F1 (end-sequences GeneBank accession numbers: AQ927932 and AQ927929). This BAC carries 30 kb and 135 kb of sequences upstream and downstream of the transcriptional start point of *Pax3*, respectively (Additional file 1: Figure S7a). Thus, the BAC is completely embedded within the TAD although it crosses the putative sub-TAD border. This BAC was modified by the introduction of a *nlacZ*-SV40pA cassette at the translational start point of *Pax3* (construct B30Z-Pax3) and used to generate transgenic lines. The transgene closely follows the endogenous

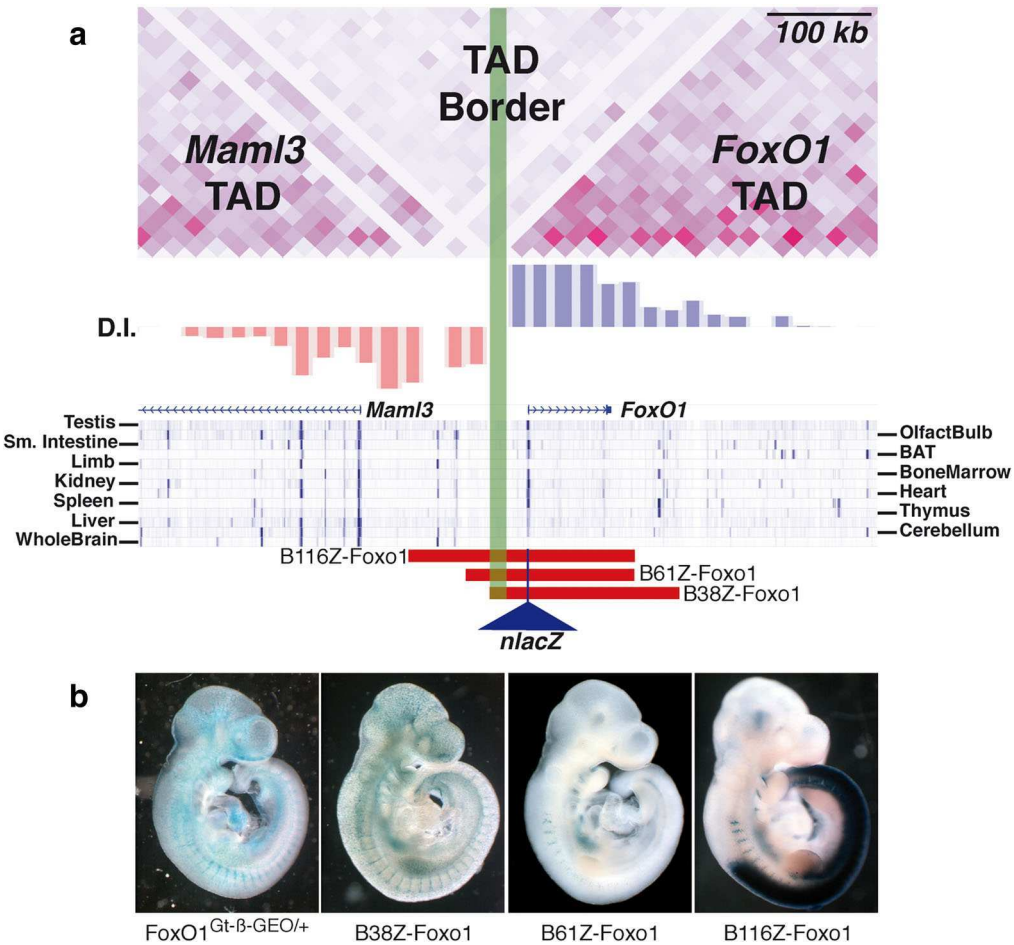


Fig. 4 Crossing the TAD border can drive ectopic transgene expression from enhancers in the adjacent domain. **a** Detail of the Hi-C data from mouse ES cells at the TAD border (*green box*). H3K27ac marks in different mouse tissues are shown underneath, as well as the position of the two coding genes in the region and the relative positions of the three BAC clones used in the study. The 5' ends of the clones cross the TAD border, although in the case of B38, its end maps within the border region. There are strong active-enhancer marks in the non-overlapping region between B116 and B61. **b** Expression patterns of *Foxo1*$^{Gt\text{-}ß\text{-}GEO/+}$, B38Z, B61Z and B116Z at 9.5 d*pc*. B116Z drives strong expression in the CNS, while B61Z does not; B38Z drives strong vascular expression like the gene-trap allele. *D.I.* directionality index

pattern of *Pax3* [35], being expressed in the neural tube, neural crest cells, somites, the hindbrain, the midbrain and forebrain, migrating limb and hyploglossal chord muscle precursors, the pre-somitic mesoderm, trigeminal ganglia and the lateral nasal process (Additional file 1: Figure S7b).

We generated additional lines using another BAC construct carrying 14 kb upstream of the *Pax3* translational start site and 128 kb downstream of it (RP24-235I14). Analysis of transgenic animals carrying B14Z-Pax3 (Additional file 1: Figure S7c) shows an identical pattern of expression to that driven by the B30Z-Pax3 described above. Therefore, the majority of the regulatory elements needed for the correct spatiotemporal expression of *Pax3* during embryonic development are presumably contained within this BAC.

Identification of regulatory regions downstream of the RH30 translocation breakpoint

We wanted to examine the enhancer potential of sequences situated downstream of the translocations in ARMS and for this we generated a new BAC construct in which all sequences downstream of the translocation breakpoint found in the RH30 cell line were deleted (B38Z-*Foxo1*-RH30Δ). We selected this particular breakpoint because the new regulatory landscape generated by the translocation in the RH30 cell line putatively carries more *PAX3* and *FOXO1* regulatory elements than the other cell lines analysed. Comparison of the expression patterns driven by the B38Z-Foxo1, B61Z-Foxo1 and B38Z-Foxo1-RH30Δ (Fig. 5a) in transgenic embryos shows that both the myotomal and embryonic vascular enhancers are located downstream of the RH30 translocation, as B38Z-Foxo1-RH30Δ only drives

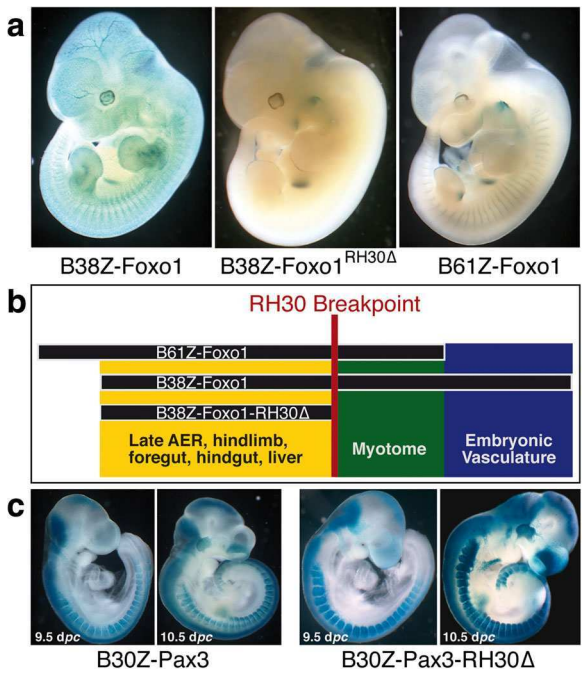


Fig. 5 *Pax3* and *FoxO1* enhancers located downstream of the RH30 breakpoint. **a** At 11.5 d*pc*, B38Z-Foxo1 drives expression in the embryonic vasculature, a hindlimb rostral domain, AER, stomach and gut. Myotomal expression is faint and masked by the vascular expression; B61Z-Foxo1 expresses in all these domains excluding the vasculature; B38Z-Foxo1-RH30Δ is not able to drive vascular or myotomal expression. **b** *Map* showing the position of enhancer regions identified downstream of the RH30 translocation breakpoint. The three BAC clones are represented as *black bars* with their given names. Different *coloured boxes* represent the location of different enhancer regions within the BACs. **c** Deletion of sequences downstream the RH30 breakpoint in B30Z-Pax3 do not result in a pattern change

expression in the AER, the foregut and the stomach. This allows the generation of a preliminary map (Fig. 5b) for the location of enhancer elements in relation to the RH30 translocation, which shows that while the enhancer elements driving expression in the developing fore-gut and hind-gut, the stomach and the AER are located upstream of the RH30 translocation, at least two major enhancers are located downstream of this translocation breakpoint. It is also important to highlight other sites of *FoxO1* expression in the mouse (e.g. brown adipose tissue or BAT), not observed in our transgenic lines but detected in a gene trap mouse strain [31], indicating that the regulatory elements controlling the expression at these other sites are not located within the BACs analysed, but further downstream. Thus, in the translocated chromosome, the *PAX3* promoter is in close proximity, at least in the linear genome, to enhancers active in non-*PAX3* territories (e.g. embryonic vasculature and BAT).

Deletion of the sequences downstream of the RH30 translocation breakpoint from B30Z-Pax3 (construct B30Z-Pax3-RH30) has a very limited effect on the overall expression pattern (Fig. 5c), with some changes in intensity levels at some locations. This result suggests that most, if not all, *PAX3* regulatory modules will be carried by the derivative t(2;13)(q35;q14) chromosome following the translocation event.

Fused regulatory landscape in ARMS

We hypothesised that the translocation event would generate a fusion of the regulatory landscapes defined by the upstream and downstream boundaries of *PAX3* and *FOXO1*, respectively (Fig. 2). This new regulatory landscape would therefore allow the interaction of the *PAX3* promoter with *FOXO1* regulatory sequences and drive the expression of the oncogene in non-*PAX3* territories. To test this, we performed 4C-seq using chromatin from the patient-derived cell line RMS taking viewpoints scattered throughout the *PAX3:FOXO1* fused locus (Fig. 6a). Some of them correspond to CTCF binding sites (VP1, VP2, VP6, VP8 and VP9), while others coincide with ECRs (VP4, VP5, VP7). Specifically, VP4 marks a well-known *PAX3* enhancer that drives neural crest expression [36]. Functional activity of the other two ECRs has not been determined, but they are enriched in active chromatin marks in various tissues. VP3 corresponds to the *PAX3* promoter. 4C-seq data were integrated to create virtual 3D chromatin conformation models (Additional file 3: Movie S1), which were further converted into a virtual Hi-C heatmap (Fig. 6b), as previously described [37]. As an example, one of the virtual models generated is represented in Fig. 6c and d and Additional file 4: Movie S2.

As predicted, the chromosomal rearrangement that takes place in RMS cells generates a new TAD as the result of the fusion of *PAX3* and *FOXO1* regulatory landscapes. Importantly, the borders of this new TAD coincide with those calculated in the wild-type loci (compare the positions of the borders in Figs. 2 and 6). Furthermore, these translocation TAD borders are mainly invariant across a multitude of human tissues (Additional file 2: Table S2), the upstream *PAX3* border and the downstream *FOXO1* border being conserved at a +/− 20 kb resolution in 61.9% and 66.7% of the 21 cell types/tissues analysed, respectively [30]. Thus, the new TAD harbours the *PAX3:FOXO1* fusion gene, as well as *FARSB* and *SGPP2*, while the flanking TADs remain mainly unchanged, with the exception of the boundary at the end of the analysed region, which shows a significant difference. Nevertheless, as this particular predicted boundary is at the end of the analysed region, it may arise as an artefact of the computational approach, which is not reliable at the extremes. Interestingly, the 4C-seq data indicate that these flanking TADs interact with each other (note the rhomboid-like domain above the *PAX3-FOXO1* TAD in Fig. 6b), presumably reinforcing the formation of an isolated highly self-interacting domain in between them. Although the D.I. analysis of the virtual

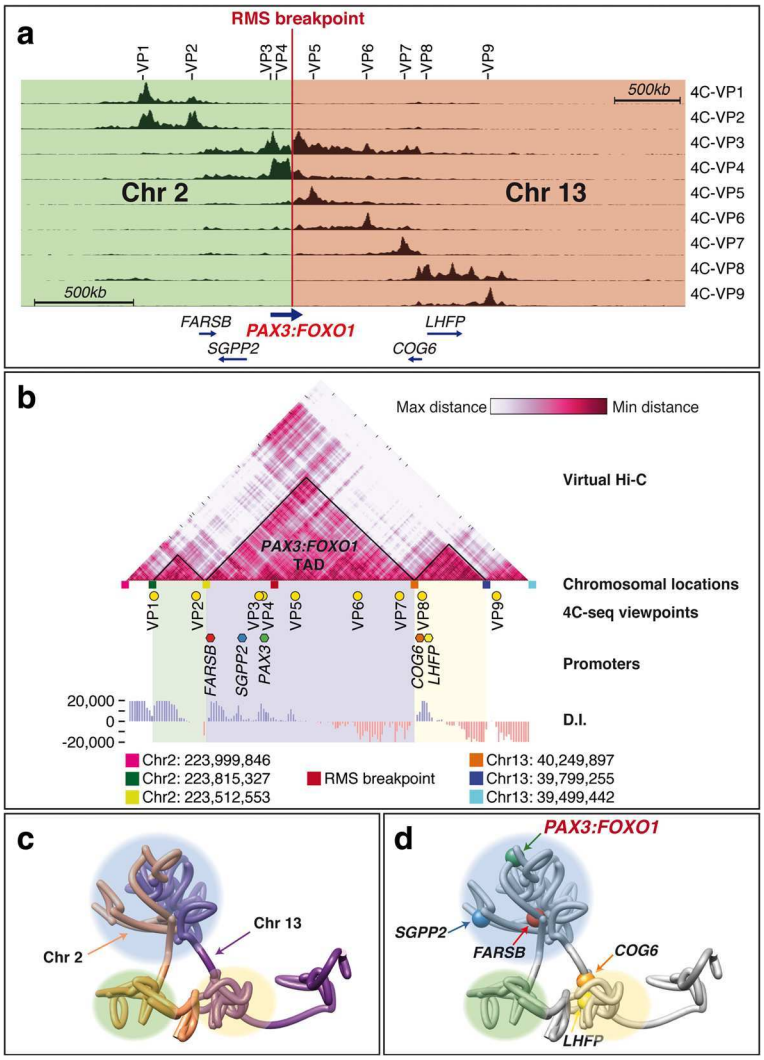


Fig. 6 Virtual-HiC of the *PAX3-FOXO1* locus in RMS cells predicts the generation of a new TAD. **a** 4C-seq profiles using nine different viewpoints (VP1–VP9) spanning 3.5 Mb. The locations of the viewpoints are indicated above the *graph*. The location of the fusion *PAX3:FOXO1* gene and other coding sequences is indicated, as well as the position of the RMS breakpoint. *Green* and *orange boxes* indicate reads mapped to Chr2 or Chr13, respectively. **b** The virtual-Hi-C generated from the 4C-seq data was subjected to a D.I. analysis to determine the location of TAD borders. The upstream and downstream borders thus defined closely match those obtained by D.I. analyses of human Hi-C data while a novel TAD encompassing the *PAX3-FOXO1* fusion locus is predicted. The positions of the viewpoints (*pale green circles*), the promoters of the genes in the region (*coloured hexagons*) and the borders identified by D.I. analysis (*coloured boxes*) are indicated. The chromosomal coordinates of the predicted borders are provided underneath. 3D chromatin architecture model for the locus encompassing the translocation in ARMS, (**c**) showing the contribution of both chromosome regions to the predicted new TAD and (**d**) the location of promoter sequences within the TAD

Hi-C data does not reveal the existence of the predicted sub-TAD containing *SGPP2* (as observed in the Hi-C analyses of wild-type mouse and human loci), the 3D chromatin structure model clearly shows an isolated chromosomal loop that contains the *SGPP2* promoter (Fig. 6d and Additional file 3: Movie S1; Additional file 4: Movie S2).

The human *PAX3* promoter is able to interact with potential *FOXO1* enhancers in RMS cells

Having demonstrated that the *PAX3* promoter lies in the same domain as *FOXO1* regulatory elements in the translocated chromosome, we sought to determine if, indeed, they could interact with each other to drive the expression of the oncogene in *FOXO1*-specific tissues. For this reason, we focused on the 4C-seq data that take the human *PAX3* promoter as a viewpoint and detected strong interactions between the *PAX3* promoter and *FOXO1* regions situated downstream of the identified breakpoint in the RMS cell line (Fig. 7). The first ectopic contacts on the *FOXO1* locus occur immediately downstream of the defined breakpoint, strengthening further our breakpoint mapping strategy.

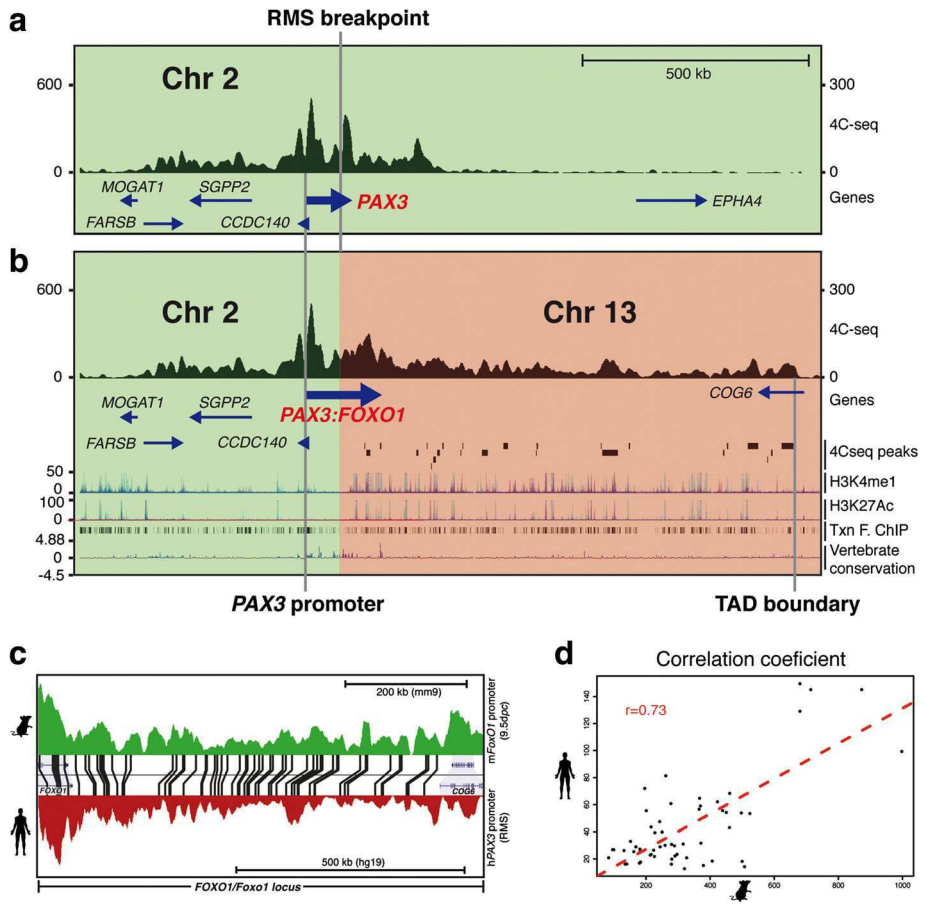


Fig. 7 The *PAX3* promoter interacts with *FOXO1* sequences in a patient-derived ARMS cell line. 4C-seq profiles on (**a**) the *PAX3* and (**b**) the *PAX3-FOXO1* loci obtained using the *PAX3* promoter as a viewpoint in the RMS cell line. The locations of the promoter and the translocation breakpoint in are indicated. In (**b**), the *first row* represents the derivative t(2:13) chromosome 4C-seq profile. *Green* and *orange boxes* indicate reads mapped to Chr2 or Chr13, respectively. The *second row* shows the location of the fusion *PAX3:FOXO1* gene and other coding sequences. The *third row* indicates the locations of 4C-peaks defined using the Peak Calling algorithm; the downstream limit was taken as the TAD boundary defined in the previous experiment. The *fourth* and *fifth rows* show H3K4me1 and H3K27ac marks in different tissues, respectively. The *sixth row* is the transcription factor-ChIP track from UCSC. The *seventh row* indicates vertebrate conservation. **c** Sequence-paired 4C-seq tracks on the *FOXO1/FoxO1* locus from mouse embryos (*green*) and the RMS cell line (*red*) showing the location of ECRs shared between human, mouse and opossum genomes as shown in Additional file 1: Figure S3. **d** Correlation between the 4C-seq signal in the *FOXO1/FoxO1* loci from mouse embryos and RMS human cells. For each conserved element, the 4C-seq signal in human cells corresponding to this region is plotted against the 4C-seq signal from mouse embryos in the orthologous region. The *red dashed line* represents the linear regression line

Furthermore, the span and location of the interactions of the *PAX3* promoter with the *FOXO1* locus in the translocation closely match those detected by 4C-seq in the mouse locus (Fig. 7c and d), suggesting that the *FOXO1* region within the novel TAD is folded in a structure similar to that of the wild-type *FOXO1* locus in chromosome 13; it is within this new chromatin structure that interactions between *FOXO1* regulatory elements and the *PAX3* promoter take place. We then applied a peak-calling algorithm that was able to detect 24 interaction peaks from the translocation point to the TAD border (Additional file 1: Figure S8 and Additional file 2: Table S3). Many of these peaks (16/24) are enriched in active chromatin marks in a variety of tissues known to express *FOXO1*, including skeletal muscle, adipose nuclei and endothelial cells. Also, some of them contain ECRs (5/24), as well as experimentally validated (ChIP-seq) binding sites (14/24) for specific transcription factors (e.g. EP300, MEF2A or CEBPB) or structural proteins such as CTCF and RAD21 (9/24). Together, these data suggest that the *PAX3* promoter engages in interactions with potential *FOXO1* regulatory elements in the translocated chromosome in ARMS tumours, interactions that are restricted to the wild-type 3' TAD border of the *FOXO1* locus.

Discussion

Transcriptional regulation of *FOXO1* and *PAX3*

The transgenic analyses show that *FoxO1* is regulated by individual regulatory regions driving expression of the transgene in different anatomical locations during embryonic development and in the adult. Importantly, we have mapped the enhancer responsible for embryonic vascular expression to the non-overlapping region between B61Z and B38Z (the +104 kb to +148 kb interval), downstream of exon 2 and thus located downstream of all translocation breakpoints in ARMS. None of our constructs is able to direct expression in brown adipose tissue (BAT), a strong site of expression for the endogenous *FoxO1* [31], indicating that this element is located further downstream.

In the case of *Pax3*, differences in the relative intensity of expression between neural tube and somites probably arise from the perdurance of β-galactosidase activity, as noted for other *lacZ* transgenes [38], and the existence of a micro RNA sequence in the 3'UTR of *Pax3* [39, 40] that downregulates somitic expression but cannot act on our *lacZ* construct as it is terminated by the SV40pA sequence. The fact that B14Z-Pax3 contains the 14 kb interval previously described [41] as the only required sequences upstream of the *Pax3* gene and that it can drive most, if not all, of the *Pax3* endogenous pattern during early embryonic development, suggests that most of the embryonic *Pax3* regulatory elements are located downstream of the *Pax3* translational start site.

Structural organisation of the *PAX3:FOXO1* locus in ARMS

Our synteny analysis shows that the chromosomal structure that includes the *FOXO1* locus (*LHFP-COG6-FOXO1-MRPS31*) is highly conserved between species as evolutionary distant as the cartilaginous fish Elephant shark (*Callorhinchus milii*) and humans, revealing that the same gene structure flanking the *FOXO1* gene has been maintained at least over the past 420 Mya. We propose that the localisation of the *FOXO1* promoter in close proximity to the upstream TAD border has been the driving force for the invariant structure of that border. Indeed, a single break of synteny could be identified in all the species covered by our analysis and that arose following a chromosomal rearrangement at the base of the rodents precisely at the interface between the two TAD structures.

Other changes in the genomes of teleosts also took place following the whole genome duplication event at the base of the bony fish group following chromosomal rearrangements. In the three cases analysed, the structure of the syntenic region has also been maintained and the *FOXO1-COG6-LHFP* syntenic group retained. Interestingly, the original upstream structure has also remained on the paralogous gene, indicating the presence of strong constraints for the disaggregation of these genes and their regulatory sequences, even if duplicated.

The study of oncogenic recurrent chromosomal translocations allows investigation of the effects of chromosomal rearrangements on gene expression without the need to resort to the reconstruction of the effect of evolutionary forces upon the process.

We have shown that in ARMS, the *PAX3* promoter interacts strongly with sequences in the *FOXO1* locus, sequences and interactions that are conserved in the wild-type mouse locus and that, in many cases, correlate with the presence of H3K27Ac marks, DNAseI hypersensitive sites, the binding of diverse transcription factors, and ECRs, indicative of active enhancers. This implies that the *PAX3:FOXO1* oncogene is, at least in part, under the control of *FOXO1* regulatory elements. Furthermore, the profile of interactions between the *PAX3* promoter and *FOXO1* sequences correlates with the profile of interactions observed between the mouse *FoxO1* promoter and its regulatory landscape.

The chromatin extrusion model of TAD formation [42, 43] may explain how the borders flanking the fused TAD are conserved after the translocation. According to this model, loop-extruding factors (likely, cohesins) would load randomly onto the DNA forming a small chromatin loop. Then, these factors would slide through the chromatin in opposite directions while still tethered, progressively extruding the DNA between them creating a larger loop. Once they encounter a boundary element (likely, CTCF in a specific orientation), they would be stalled. The new TAD would thus be formed by the interaction between the pre-existing borders creating a new regulatory landscape in which contacts between the *PAX3* promoter and regulatory elements of *FOXO1* take place. We cannot exclude that these interactions may contribute to the formation and/or maintenance of the new TAD, as previously suggested in the case of the *Xist* locus [44].

TADs are composed of and are a consequence of chromatin interactions. However, in the case of the *PAX3:-FOXO1* TAD we argue against a model in which TAD formation is caused by the pre-establishment of specific enhancer-promoter or enhancer-enhancer regulatory interactions. The translocation places the *PAX3* promoter and enhancers from both genes in a new regulatory environment. We would argue that in this new environment the interactions would be significantly different from those established in the wild-type locus and thus if these preceded TAD organisation, a shift of the position of the borders would have been observed.

It has recently been reported that active transcription or gene looping is not required for TAD formation [45]. The authors show conservation of TAD organisation around the *CFTR* locus in five different cancer cell lines,

two of which do not express the gene. Furthermore, looping interactions within the *CFTR*-containing TAD (intra-TAD interactions) were highly specific in those cells that express the gene and absent in those that do not express it. Thus, as previously reported [2, 46], internal TAD organisation is cell-type specific whereas overall TAD structures are mostly conserved, which argues against a model in which TADs are passively formed as a consequence of the establishment of specific regulatory interactions. Additionally, such a model in which the emphasis is placed on the interactions and not on the importance of a border would not explain why the removal of TAD boundaries cause adjacent TADs to merge and a rewiring of regulatory interactions [17–19].

Our analyses also show that while both B61Z- and B116Z-Foxo1 cross the *FoxO1* 5'-TAD border, only B116Z-Foxo1 spans into regions marked by H3K27ac in the whole brain, cerebellum and olfactory tract, which suggest the presence of active neural tissue enhancers. Therefore, the sequence of this TAD border is not sufficient to separate regulatory landscapes, indicating that efficient separation may require interaction between TAD-border sequences, such as convergent CTCF binding sites [15, 16], and other sequences within the TAD domains. In fact, close observation of the mouse Hi-C data reveals that the borders of the *FoxO1*-containing TAD do interact with each other (note the interactions at the peak of the triangle depicting the third TAD at the bottom of Fig. 2a).

Implications for the cell type of origin for ARMS

ARMS tumours appear generally in trunk and extremities [47], but examples of other sites of primary ARMS abound in the literature (e.g. [48–53]), suggesting that they can arise in multiple cell types or in a single cell type found throughout the body, with certain locations such as the extremities being more susceptible than others. ARMS tumours are characterised by the expression of muscle-specific markers (reviewed in [54]), suggesting a possible myogenic origin, although their molecular characteristics are more related to cells that have been committed to the myogenic lineage but are unable to complete terminal differentiation to become skeletal muscle. For example, it has been shown that *MYOD* is activated by the *PAX3-FOXO1* fusion protein while it interferes with its chromatin remodelling functions, inhibiting the expression of the skeletal muscle terminal differentiation factor, *MYOG* [55]. An interesting hypothesis is that dysregulation of PAX3 or PAX7 target genes may result in the activation of the myogenic programme in a non-myogenic lineage, the cells being able to transdifferentiate but unable to fully complete terminal differentiation. It has been shown that ectopic expression of *PAX3* in the lateral plate mesoderm of chick embryos induces the expression of the myogenic regulatory factors *MYF5*, *MYOG* and *MYOD* [56]; expression in mesenchymal stem cells also induces the activation of myogenic markers such as *MYF5*, *MYOD*, *MYOG*, *MCK* and *MHC*, pushing them towards the myogenic lineage, while blocking their osteogenic, chondrogenic or adipogenic potential [57]. It is thus likely that the myogenic-like transcriptome of ARMS tumours [58] is the result of *PAX3:FOXO1* activation rather than a remnant of their lineage origin.

Several cell types have been previously suggested as the origin for ARMS, corresponding to embryonic, postnatal or adult stem cells or adult myofibres [59], both from the myogenic lineage [60–64] or other lineages [65, 66].

Our data reveal a clear set of interactions in the embryo between the *FoxO1* promoter and, in the RMS cell line, the *PAX3* promoter, and far-downstream sequences in the *FOXO1/FoxO1* locus, which presumably correspond to enhancer regions of the gene.

An interesting site of *FoxO1* expression is BAT [31], which can easily transdifferentiate into muscle and vice versa [67–70], while overexpression of a constitutively active *Smoothened* restricted to adipocytes has been shown to give rise to embryonic rhabdomyosarcomas (ERMS) [71] with relative high penetrance.

None of our constructs drive expression in BAT, indicating that the enhancer(s) responsible for this aspect of the expression is located even further downstream. Indeed, epigenetic marks in BAT from 24-week-old mice indicate active sites coincident with downstream regions that interact strongly with both the mouse *FoxO1* and human *PAX3* promoters (Additional file 1: Figure S3), while our data clearly show that the enhancers required for both embryonic and adult vasculature expression are located downstream of all the mapped translocation breakpoints.

Another important site of expression is the developing and adult vasculature, although we have not identified the different cell types associated with this expression. In the embryo, some progenitors for vasculature and skeletal muscle reside in the dermomyotome and their fate decision depends on the ratio between *Pax3* and *Foxc2*, acting as pro-myogenic and pro-angiogenic factors, respectively. Importantly, *Foxc2* expression is repressed both by PAX3 and the PAX3-FOXO1 fusion protein, promoting myogenesis in cells that, under normal circumstances, would not give rise to skeletal muscle [72]. Therefore, we propose the BAT and vasculature cell lineages as new candidates for the cell type of origin for ARMS. As the survival rates for these types of tumour are particularly low (around 70% of patients show recurrent tumour resurgence following current therapies), the final identification of the lineages that can serve as origin for ARMS will provide

further information on the biology of these tumours and the importance of additional activating mutations specific for each lineage, opening new avenues for the development of new targeted therapies based on the transcriptome and epigenome of the individual cell types of origin.

Conclusions

We have shown that novel regulatory landscapes arise as a result of oncogenic human translocations and that these are restricted by the original upstream and downstream TAD boundaries of the genes involved in the translocation, indicating that TAD formation precedes intra-TAD interactions. We have identified several major enhancer regions for *FOXO1* present downstream of all t(2;13) translocations in ARMS and thus potentially able to drive expression of the oncogene in non-*PAX3*-expressing cells. We also indicate that brown adipose tissue and the vasculature should be considered in future studies on cell lineage of origin for ARMS. Ectopic oncogene activation may be an essential step in the tumorigenic process, as expression in a particular cell type, the often-elusive cell of origin, may be required for disease development.

Methods

Integration of a *LacZ* reporter gene into BAC clones

To target the *FoxO1* BACs, homology arms were synthesised by standard PCR methods using the oligonucleotide primers pFoxHAF + *Apa*I/pFoxHAR + *Apa*I (Additional file 2: Table S1) which generate a 410 bp fragment spanning 204 bp and 206 bp upstream and downstream of the first coding ATG of *FoxO1*, respectively. We then used the single *Nco*I site at position -1 to insert a linker sequence (Additional file 2: Table S1). Into the single *Bgl*II of the linker we then cloned a *galK* selectable marker [73] or a ~3 kb *Bam*HI fragment from our standard construct #1 [74] containing a nuclear-localised *lacZ* reporter gene and a SV40 polyadenylation signal. To target the *Pax3* BACs, homology arms were synthesised by standard PCR methods using the oligonucleotide primer pairs pPax3_5HAF + *Eag*I/pPax3_5HAR + Link and pPax3_3-HAF + Link/pPax3_3HAR + *Eag*I (Additional file 2: Table S1) and then joined by PCR. This generates a 950 bp fragment spanning 461 bp and 468 bp upstream and downstream of the first coding ATG of *Pax3*, respectively, and introduces a small polylinker immediately upstream of the gene. We then used the single *Bgl*II site at position -1 to insert the *galK* selectable marker or the nuclear-localised *lacZ* reporter gene and a SV40-polyA. These constituted the targeting cassettes. The B116-Foxo1, B61Z-Foxo1, B38Z-Foxo1, B14-Pax3 and B30-Pax3 BAC constructs were then modified by two-step *galK* recombineering [73] with modifications as previously described [75]. All positive clones were checked for integrity by multiple restriction digests and inserts sequenced prior to pronuclear injection. The number of independent transgenic lines showing similar expression patterns for each construct is as follows: B38Z-Foxo1: four lines; B61Z-Foxo1: three lines; B116Z-Foxo1: three lines; B14Z-Pax3: two lines; B30Z-Pax3: four lines.

RH30 deletion in BAC clones

To generate the deletions at the RH30 breakpoint sequence in mouse BACs, we made homology cassettes (Additional file 2: Table S1) with ~75 bp of homology at either side of the mouse sequence corresponding to the breakpoint in the RH30 cell line and containing a LoxP511 site in the same orientation as the one in the BAC vector-backbone (pBACe3.6). The cassettes were then inserted by single-step recombineering [73] in B38Z-Foxo1 and B30Z-Pax3. Positive clones were sequenced and transferred into the SW106 *E. coli* bacterial strain [73] that carries an Arabinose-inducible *Cre* gene for the excision of the intervening fragments. Following induction of *Cre* expression, positive clones were identified and checked for integrity by multiple restriction digests; deletions were confirmed by sequencing prior to pronuclear injection. The number of independent transgenic lines showing similar expression patterns for each construct is as follows: B38Z-Foxo1-RH30Δ: three lines; B30Z-Pax3-RH30Δ: two lines.

Generation of transgenic mice and embryo analyses

BAC DNA was prepared using the QIAgen maxiprep kit (QIAGEN Ltd., UK) as previously described [75]. After dialysis against microinjection buffer (10 mM Tris-HCl pH 7.5, 0.1 mM EDTA pH 8.0 and 100 mM NaCl), DNA was diluted to 1.6–1.8 ng/mL in microinjection buffer and used for pronuclear injection of fertilised mouse eggs from B6CBAF1/OlaHsd crosses using standard techniques. Embryo β-galactosidase staining was performed as previously described [75]. Embryo pictures were obtained using a Nikon SMZ1500 microscope and a JVC KY-F55B 3-CCD camera connected to a Scion Series 7 card. Images were imported into AdobePhotoshop (v12.0 x64) and whole image correction applied using the 'AutoLevels' tool.

Identification of breakpoints in ARMS cell lines

The RH3, RH28 and RH41 cell lines were obtained from Dr Peter Houghton (St Jude Children's Research Hospital, Memphis, TN, USA); the RMS, SCMC and RH30 cell lines were a kind gift from Dr Janet Shipley (The Institute of Cancer Research, Sutton, UK). Cells were grown in Dulbecco's Modified Eagle's Medium (DMEM, SIGMA UK) supplemented with 10% (v/v) fetal calf serum, 60 mg/mL Benzylpenicillin and 100 mg/mL Streptomycin sulphate. Cells were isolated from two

75 cm^2 flasks (Nunc) at 80% confluency by standard methods and genomic DNA extracted as previously described [76]. LD-PCR was used to amplify the genomic DNA from the different cell lines using all possible combinations from 11 oligonucleotides evenly spaced over ~110 kb and covering intron 1 of *FOXO1* (Foxo1-LD primers) and seven oligonucleotides evenly spaced over ~27 kb and covering intron 7 of *PAX3* (Pax3-LD primers) (Additional file 2: Table S1). LD-PCR was performed using the Expand Long Template PCR kit (Roche), using Buffer 3, as instructed by the manufacturers. The SCMC breakpoint was amplified with the Foxo1-LD8/Pax3-LD6 primer pair (3.1 kb); the RH3 breakpoint was amplified with the Foxo1-LD8/Pax3-LD2 primer pair (1.3 kb fragment); the RH5 breakpoint was amplified using the Foxo1-LD8/Pax3-LD3 primer pair (5.3 kb); the RMS breakpoint was amplified using the Foxo1-LD5/Pax3-LD3 primer pair (7.8 kb); the RH4/RH41 breakpoint was amplified using the Foxo1-LD7/Pax3-LD3 primer pair (12.8 kb fragment). Products were cloned into pCR2.1-TOPO (Invitrogen) and sequenced. We have previously reported the sequence of the RH30 translocation breakpoint [28].

4C-seq analyses

4C-seq assays were performed as previously reported [77–80]. Briefly, hybrid CBA/C57Bl6 mouse embryos at the desired stage were disrupted using 1X PBS/0.125% (w/v) collagenase (Sigma-Aldrich). 10^7 individual cells were fixed in 1X PBS/2% (w/v) formaldehyde for 15 min at room temperature. A total of 155 μl of 10% (w/v) Glycine were added to stop the fixation, followed by a wash by centrifugation with 1X PBS at 4 °C. Pellets were frozen in liquid nitrogen and kept at -80 °C. Isolated cells were lysed (lysis buffer: 10 mM Tris-HCl pH 8, 10 mM NaCl, 0.3% (v/v) IGEPAL CA-630 [Sigma-Aldrich]), 1X protease inhibitor cocktail (cOmplete, Roche) was added and the DNA digested with *Dpn*II and *Csp*6I as primary and secondary enzymes, respectively. T4 DNA ligase was used for both ligation steps. Specific primers were designed at the genes promoters 4C-mPax3 (mouse *Pax3* promoter), 4C-hPAX3 (human *PAX3* promoter) and 4C-mFoxo1 (mouse *FoxO1* promoter), as well as for the rest of the viewpoints (VP1–VP9) (Additional file 2: Table S1) with Primer3 (v. 0.4.0) [81]. Illumina adaptors were included in the primer sequences. Eight separate PCRs were performed for each viewpoint with Expand Long Template PCR System (Roche) and pooled together. The libraries were purified with a High Pure PCR Product Purification Kit (Roche), concentrations measured using the Quanti-iTTM PicoGreen dsDNA Assay Kit (Invitrogen) and sent for deep sequencing.

4C-seq data analyses and 3D chromatin modelling

4C-seq data were analysed as previously described [79]. Briefly, raw sequencing data were de-multiplexed and aligned using mouse July 2007 assembly (mm9) or human February 2009 (hg19) as the reference genomes. Reads located in fragments flanked by two restriction sites of the same enzyme, or in fragments smaller than 40 bp, were filtered out. Mapped reads were then converted to reads-per-first-enzyme-fragment-end units and smoothed using a 30 fragment mean running window algorithm, uploaded to the UCSC genome browser [82] (http://genome.ucsc.edu/, 2015) and subjected to a five-pixel smoothing window. In Fig. 7, as reads upstream of the breakpoint come from both the intact and translocated *PAX3* locus and downstream reads map to *PAX3* or *FOXO1*, 4C-seq scales have been adjusted to normalise reads at either side of the translocation.

The protocol of the chromatin modelling based on 4C-seq data was applied as previously described [37]. Briefly, 4C-seq data were used as a proxy of distance between individual viewpoints and the rest of the DNA fragments under the assumption that 4C-seq reads are inversely proportional to their spatial distance. These distances were used as restraint coordinates to locate the position of DNA fragments in the 3D space. The Integrative Modelling Platform (IMP) [83] was used for the generation of chromatin 3D models. The 200 top-scoring models were selected out of 50,000 and then clustered in two populations that were mirror image of each other. The most populated cluster was selected and used for the calculation of the Virtual Hi-C, as previously described [37].

4C-seq reads corresponding to the derivative t(2:13) chromosome were duplicated in order to compensate the theoretical quantity of whole chromosomes depending on the viewpoint used. Reads were then normalised and the Z-scores calculated as previously described [37] to filter out the non-significant data. For peak calling of 4C-seq data, interaction calling was carried out using as a background a two-sided monotonic regression calculated using the Pool Adjacent Violators Algorithm (PAVA) from the R-package isotone [84]. With this background, we computed the distribution of residuals (differences between observed and expected values for each fragment) and defined as peaks those fragments with residuals that were above the third quartile plus 1.5 × IQR, IQR being the interquartile range [85]. Peaks less than 500 bp apart were merged together in a single unit.

Directionality index and boundary calling

Boundary calling was carried out using the D.I. [2]. The D.I. at each position is based on fragments contacts for both sides, but we only used data limited to these regions of interest. Thus, we are missing data for the fragments located at the borders. We simulated the

missing data for the fragments in the borders by taking the mean value of the complete dataset as reference. We calculated the D.I. of the Hi-C's for both loci in the two species iteratively, changing the expected TAD size variable in each iteration (Additional file 2: Table S4). We selected the boundaries that appeared in all the iterations. We used the same approach for the virtual Hi-C of the truncated locus but we selected the top two boundaries which appeared in 96% of the iterations (Additional file 2: Table S4). Hi-C data were taken from the Epigenome Browser (http://egg.wustl.edu/d/; 2016); the datasets used for these calculations were: MM9: Esc_20kb_hindIII_rep1_mouse and HG19: Esc_20kb_hindIII_rep2_human.

Additional files

Additional file 1: Figure S1. Orthologous pairwise clusters involving the *FoxO1* gene. **Figure S2.** Conservation analysis across the *FoxO1-Maml3* intergenic region. **Figure S3.** ECRs identified in the *FoxO1* region downstream of the RMS breakpoint and associated H3K27ac marks. **Figure S4.** Time-course of embryos carrying the B116Z-Foxo1 reporter construct. **Figure S5.** Time-course of embryos carrying the B61Z-Foxo1 reporter construct. **Figure S6.** Time-course of embryos carrying the B38Z-Foxo1 reporter construct. **Figure S7.** Recapitulation of Pax3 endogenous expression pattern by a BAC carrying 30 kb of upstream sequences. **Figure S8** Peaks of interaction established by the *PAX3* promoter at the *FOXO1* locus in RMS cells. (PDF 3020 kb)

Additional file 2: Table S1. Oligonucleotides used in this work. **Table S2.** Comparison of the TAD borders called at the *PAX3* and *FOXO1* human loci. **Table S3.** Interaction peaks between the *PAX3* promoter and regions within the *FOXO1* locus in RMS cells. **Table S4.** Number of times the defined TAD boundaries appeared in the iteration. (PDF 296 kb)

Additional file 3: Movie S1. PAX3:FOXO1 3D superposition model. (MP4 75352 kb)

Additional file 4: Movie S2. PAX3:FOXO1 3D chromatin model. (MP4 9598 kb)

Acknowledgements
We are grateful to all members of the Carvajal and Gómez-Skarmeta laboratories for helpful discussions; to J.R. Martínez-Morales for critical reading of the manuscript; to Ana Jesús Franco Gómez and Cándida Mateos Orozco, at the Animal House facility at CABD, for expert animal husbandry; and to Ana Fernández Miñán, at the Functional Genomics Platform from the CABD, for her support on the design and completion of 4C-seq experiments.

Funding
This work was supported by The Institute of Cancer Research and Cancer Research UK (Grant C1178/A4520). JJC was funded by grants from Spanish Ministerio de Ciencia e Innovación (BFU2011-22928) and the European Commission (PCIG10-GA-2011-303904). JLGS was funded by grants from Spanish Ministerio de Economía y Competitividad (BFU2013-41322-P) and the Andalusian Government (BIO-396). DPD and IIA were funded by grants from the Spanish Ministerio de Economia y Competitividad (BFU2013-40866-P) and from the Junta de Andalucia (C2A). BV-B held a studentship from The Institute of Cancer Research. We acknowledge support of the publication fee by the CSIC Open Access Publication Support Initiative through its Unit of Information Resources for Research (URICI).

Availability of data and materials
The 4C-seq datasets supporting the results of this article have been deposited in the GEO database under accession number GSE69439 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=cvujaoowdnatbuv&acc=GSE69439). In addition, all Additional data are available in the figShare repository (https://figshare.com/articles/Vicente-Garc_a_2017_Additional_Figures/4880396).

Authors' contributions
CV-G: Drafting of the manuscript; acquisition, analysis and interpretation of 4C-seq data; analysis and interpretation of all data; critical revision of the manuscript. BV-B: Acquisition, analysis and interpretation of transgenic data; mapping translocations. II-A: Acquisition, analysis and interpretation of Virtual HiC data. SN: Acquisition, analysis and interpretation of 4C-seq data. RDA: Acquisition, analysis and interpretation of 4C-seq data. JJT: Critical revision of the manuscript; Analysis and interpretation of 4C-seq data. PWJR: Study concept and design; drafting of the manuscript; study supervision. DPD: Analysis and interpretation of Virtual HiC data; study supervision. JLG-S: Analysis and interpretation of 4C-seq data; study supervision. JJC: Study concept and design; transgenic generation; acquisition of transgenic data; analysis and interpretation of all data; study supervision; drafting of the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Ethics approval and consent to participate
All animal experimentation was performed using protocols approved by the Universidad Pablo de Olavide Ethical Committee (Seville, Spain) and The Institute of Cancer Research Ethical Committee in accordance with Spanish Royal Decree 53/2013, European Directive 2010/63/EU, the United Kingdom Animals (Scientific Procedures) Act 1986, and other relevant guidelines.

Publisher's Note

Author details
[1]Centro Andaluz de Biología del Desarrollo (CABD), CSIC-UPO-JA, Universidad Pablo de Olavide, Carretera de Utrera km1, 41013 Seville, Spain. [2]Division of Cancer Biology, The Institute of Cancer Research, Chester Beatty Laboratories, 237 Fulham Road, London SW3 6JB, UK.

Received: 15 December 2016 Accepted: 28 April 2017
Published online: 14 June 2017

References
1. Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet. 2013;14:390–403.
2. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485:376–80.
3. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature. 2012;485:381–5.
4. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-dimensional folding and functional organisation principles of the Drosophila genome. Cell. 2012;148:458–72.
5. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, et al. A map of the cis-regulatory sequences in the mouse genome. Nature. 2012;488:116–20.
6. Le Dily F, Baù D, Pohl A, Vincent GP, Serra F, Soronellas D, et al. Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. Genes Dev. 2014;28:2151–62.
7. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326:289–93.
8. Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, et al. Chromatin architecture reorganization during stem cell differentiation. Nature. 2015;518:331–6.
9. Fraser J, Ferrai C, Chiariello AM, Schueler M, Rito T, Laudanno G, et al. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. Mol Syst Biol. 2015;11:1–14.
10. Vietri-Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, Hadjur S. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Rep. 2015;10:1297–309.

11. de Laat W, Duboule D. Topology of mammalian developmental enhancers and their regulatory landscapes. Nature. 2013;502:499–506.
12. Ghavi-Helm Y, Klein FA, Pakozdi T, Ciglar L, Noordermeer D, Huber W, et al. Enhancer loops appear stable during development and are associated with paused polymerase. Nature. 2014;512:96–100.
13. Lonfat N, Duboule D. Structure, function and evolution of topologically associating domains (TADs) at HOX loci. FEBS Lett. 2015;589:2869–76.
14. Phillips JE, Corces VG. CTCF: master weaver of the genome. Cell. 2009;137: 1194–211.
15. Gómez-Marín C, Tena JJ, Acemel RD, López-Mayorga M, Naranjo S, de la Calle-Mustienes E, et al. Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. Proc Natl Acad Sci U S A. 2015;112:7542–7.
16. Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, et al. CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. Cell. 2015;162:900–10.
17. Tsujimura T, Klein FA, Langenfeld K, Glaser J, Huber W, Spitz F. A discrete transition zone organizes the topological and regulatory autonomy of the adjacent tfap2c and bmp7 genes. PLoS Genet. 2015;11:e1004897.
18. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell. 2015;161:1012–25.
19. Franke M, Ibrahim DM, Andrey G, Schwarzer W, Heinrich V, Schöpflin R, et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature. 2016;538:265–9.
20. Bakhshi A, Jensen JP, Goldman P, Wright JJ, McBride OW, Epstein AL, et al. Cloning the chromosomal breakpoint of t(14;18) human lymphomas: clustering around JH on chromosome 14 and near a transcriptional unit on 18. Cell. 1985;41:899–906.
21. Gröschel S, Sanders MA, Hoogenboezem R, Zeilemaker A, Havermans M, Erpelinck C, et al. Mutational spectrum of myeloid malignancies with inv(3)/t(3;3) reveals a predominant involvement of RAS/RTK signaling pathways. Blood. 2015;125:133–9.
22. Kovalchuk AL, Ansarah-Sobrinho C, Hakim O, Resch W, Tolarová H, Dubois W, et al. Mouse model of endemic Burkitt translocations reveals the long-range boundaries of Ig-mediated oncogene deregulation. Proc Natl Acad Sci U S A. 2012;109:10972–7.
23. Polack A, Feederle R, Klobeck G, Hörtnagel K. Regulatory elements in the immunoglobulin kappa locus induce c-myc activation and the promoter shift in Burkitt's lymphoma cells. EMBO J. 1993;12:3913–20.
24. Ryan RJ, Drier Y, Whitton H, Cotton MJ, Kaur J, Issner R, et al. Detection of enhancer-associated rearrangements reveals mechanisms of oncogene dysregulation in B-cell lymphoma. Cancer Discov. 2015;5:1058–71.
25. Tsujimoto Y, Cossman J, Jaffe E, Croce CM. Involvement of the bcl-2 gene in human follicular lymphoma. Science. 1985;228:1440–3.
26. Yamazaki H, Suzuki M, Otsuki A, Shimizu R, Bresnick EH, Engel JD, et al. A remote GATA2 hematopoietic enhancer drives leukemogenesis in inv(3)(q21;q26) by activating EVI1 expression. Cancer Cell. 2014;25:415–27.
27. Douglass EC, Valentine M, Etcubanas E, Parham D, Webber BL, Houghton PJ, et al. A specific chromosomal abnormality in rhabdomyosarcoma. Cytogenet Cell Genet. 1987;45:148–55.
28. Lagutina IV, Valentine V, Picchione F, Harwood F, Valentine MB, Villarejo-Balcells B, et al. Modeling of the human alveolar rhabdomyosarcoma Pax3-Foxo1 chromosome translocation in mouse myoblasts using CRISPR-Cas9 nuclease. PLoS Genet. 2015;11:e1004951.
29. Meaburn KJ, Misteli T, Soutoglou E. Spatial genome organization in the formation of chromosomal translocations. Semin Cancer Biol. 2007;17:80–90.
30. Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, et al. Compendium of chromatin contact maps reveals spatially active regions in the human genome. Cell Rep. 2016;17:2042–59.
31. Villarejo-Balcells B, Guichard S, Rigby PW, Carvajal JJ. Expression pattern of the Foxo1 gene during embryonic development. Gene Expr Patterns. 2011; 11:299–308.
32. Galili N, Davis RJ, Fredericks WJ, Mukhopadhyay S, Rauscher 3rd FJ, Emanuel BS, et al. Fusion of a fork head domain gene to PAX3 in the solid tumour alveolar rhabdomyosarcoma. Nat Genet. 1993;5:230–6.
33. Shapiro DN, Sublett JE, Li B, Downing JR, Naeve CW. Fusion of PAX3 to a member of the forkhead family of transcription factors in human alveolar rhabdomyosarcoma. Cancer Res. 1993;53:5108–12.
34. Xia SJ, Barr FG. Analysis of the transforming and growth suppressive activities of the PAX3-FKHR oncoprotein. Oncogene. 2004;23:6864–71.
35. Goulding MD, Chalepakis G, Deutsch U, Erselius JR, Gruss P. Pax-3, a novel murine DNA binding protein expressed during early neurogenesis. EMBO J. 1991;10:1135–47.
36. Degenhardt KR, Milewski RC, Padmanabhan A, Miller M, Singh MK, Lang D, et al. Distinct enhancers at the Pax3 locus can function redundantly to regulate neural tube and neural crest expressions. Dev Biol. 2010;339:519–27.
37. Acemel RD, Tena JJ, Irastorza-Azcarate I, Marlétaz F, Gómez-Marín C, de la Calle-Mustienes E, et al. A single three-dimensional chromatin compartment in amphioxus indicates a stepwise evolution of vertebrate Hox bimodal regulation. Nat Genet. 2016;48:336–41.
38. Echelard Y, Vassileva G, McMahon AP. Cis-acting regulatory sequences governing Wnt-1 expression in the developing mouse CNS. Development. 1994;120:2213–24.
39. Crist CG, Montarras D, Pallafacchina G, Rocancourt D, Cumano A, Conway SJ, et al. Muscle stem cell behavior is modified by microRNA-27 regulation of Pax3 expression. Proc Natl Acad Sci U S A. 2009;106:13383–7.
40. Goljanek-Whysall K, Sweetman D, Abu-Elmagd M, Chapnik E, Dalmay T, Hornstein E, et al. MicroRNA regulation of the paired-box transcription factor Pax3 confers robustness to developmental timing of myogenesis. Proc Natl Acad Sci U S A. 2011;108:11936–41.
41. Natoli TA, Ellsworth MK, Wu C, Gross KW, Pruitt SC. Positive and negative DNA sequence elements are required to establish the pattern of Pax3 expression. Development. 1997;124:617–26.
42. Sanborn AL, Rao SS, Huang SC, Durand NC, Huntley MH, Jewett AI, et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc Natl Acad Sci U S A. 2015; 112:E6456–65.
43. Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny LA. Formation of chromosomal domains by loop article formation of chromosomal domains by loop extrusion. Cell Rep. 2016;15:2038–49.
44. Giorgetti L, Galupa R, Nora EP, Piolot T, Lam F, Dekker J, Tiana G, et al. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. Cell. 2014;157:950–63.
45. Smith EM, Lajoie BR, Jain G, Dekker J. Invariant TAD boundaries constrain Cell-type-specific looping interactions between promoters and distal elements around the CFTR locus. Am J Hum Genet. 2016;98:185–201.
46. Phillips-Cremins JE, Sauria ME, Sanyal A, Gerasimova TI, Lajoie BR, Bell JS, et al. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell. 2013;153:1281–95.
47. Tsokos M, Webber BL, Parham DM, Wesley RA, Miser A, Miser JS, et al. Rhabdomyosarcoma. A new classification scheme related to prognosis. Arch Pathol Lab Med. 1992;116:847–55.
48. Chiarle R, Godio L, Fusi D, Soldati T, Palestro G. Pure alveolar rhabdomyosarcoma of the corpus uteri: description of a case with increased serum level of CA-125. Gynecol Oncol. 1997;66:320–3.
49. Chikhalkar S, Gutte R, Holmukhe S, Khopkar U, Desai S, Gupta S. Alveolar rhabdomyosarcoma arising in a giant congenital melanocytic nevus in an adult–case report with review of literature. Int J Dermatol. 2013;52:1372–5.
50. Haerr RW, Turalba CI, el-Mahdi AM, Brown KL. Alveolar rhabdomyosarcoma of the larynx: case report and literature review. Laryngoscope. 1987;97:339–44.
51. Nunez AL, Elgin JN, Fatima H. Fine-needle aspiration biopsy of alveolar rhabdomyosarcoma of Stensen's duct: a case report and review of the literature. Diagn Cytopathol. 2014;42:1069–74.
52. Tailor IK, Motabi I, Alshehry N, Zaidi S, Algwaiz L. Alveolar rhabdomyosarcoma masquerading as Burkitt's lymphoma in bone marrow. Hematol Oncol Stem Cell Ther. 2015;8:38–9.
53. Valencerina-Gopez E, Dauterman J, Layfield LJ. Fine-needle aspiration biopsy of alveolar rhabdomyosarcoma of the parotid: a case report and review of the literature. Diagn Cytopathol. 2001;24:249–52.
54. Barr FG. Gene fusions involving PAX and FOX family members in alveolar rhabdomyosarcoma. Oncogene. 2001;20:5736–46.
55. Calhabeu F, Hayashi S, Morgan JE, Relaix F, Zammit PS. Alveolar rhabdomyosarcoma-associated proteins PAX3/FOXO1A and PAX7/FOXO1A suppress the transcriptional activity of MyoD-target genes in muscle stem cells. Oncogene. 2013;32:651–62.
56. Maroto M, Reshef R, Münsterberg AE, Koester S, Goulding M, Lassar AB. Ectopic Pax-3 activates MyoD and Myf-5 expression in embryonic mesoderm and neural tissue. Cell. 1997;89:139–48.
57. Gang EJ, Bosnakovski D, Simsek T, To K, Perlingeiro RC. PAX3 activation promotes the differentiation of mesenchymal stem cells toward the myogenic lineage. Exp Cell Res. 2008;314:1721–33.

58. Davicioni E, Finckenstein FG, Shahbazian V, Buckley JD, Triche TJ, Anderson MJ. Identification of a PAX-FKHR gene expression signature that defines molecular classes and determines the prognosis of alveolar rhabdomyosarcomas. Cancer Res. 2006;66:6936–46.
59. Keller C, Capecchi MR. New genetic tactics to model alveolar rhabdomyosarcoma in the mouse. Cancer Res. 2005;65:7530–2.
60. Keller C, Hansen MS, Coffin CM, Capecchi MR. Pax3:Fkhr interferes with embryonic Pax3 and Pax7 function: implications for alveolar rhabdomyosarcoma cell of origin. Genes Dev. 2004;18:2608–13.
61. Keller C, Arenkiel BR, Coffin CM, El-Bardeesy N, DePinho RA, Capecchi MR. Alveolar rhabdomyosarcomas in conditional Pax3: Fkhr mice: cooperativity of Ink4a/ARF and Trp53 loss of function. Genes Dev. 2004;18:2614–26.
62. Schaaf GJ, Ruijter JM, van Ruissen F, Zwijnenburg DA, Waaijer R, Valentijn LJ, et al. Full transcriptome analysis of rhabdomyosarcoma, normal, and fetal skeletal muscle: statistical comparison of multiple SAGE libraries. FASEB J. 2005;19:404–6.
63. Tiffin N, Williams RD, Shipley J, Pritchard-Jones K. PAX7 expression in embryonal rhabdomyosarcoma suggests an origin in muscle satellite cells. Br J Cancer. 2003;89:327–32.
64. Zhang M, Truscott J, Davie J. Loss of MEF2D expression inhibits differentiation and contributes to oncogenesis in rhabdomyosarcoma cells. Mol Cancer. 2013;12:150.
65. Abraham J, Nuñez-Álvarez Y, Hettmer S, Carrió E, Chen HI, Nishijo K, et al. Lineage of origin in rhabdomyosarcoma informs pharmacological response. Genes Dev. 2014;28:1578–91.
66. Ren YX, Finckenstein FG, Abdueva DA, Shahbazian V, Chung B, Weinberg KI, et al. Mouse mesenchymal stem cells expressing PAX-FKHR form alveolar rhabdomyosarcomas by cooperating with secondary mutations. Cancer Res. 2008;68:6587–97.
67. Grimaldi PA, Teboul L, Inadera H, Gaillard D, Amri EZ. Trans-differentiation of myoblasts to adipoblasts: triggering effects of fatty acids and thiazolidinediones. Prostaglandins Leukot Essent Fatty Acids. 1997;57:71–5.
68. Hu E, Tontonoz P, Spiegelman BM. Transdifferentiation of myoblasts by the adipogenic transcription factors PPAR gamma and C/EBP alpha. Proc Natl Acad Sci U S A. 1995;92:9856–60.
69. Jumabay M, Abdmaulen R, Ly A, Cubberly MR, Shahmirian LJ, Heydarkhan-Hagvall S, et al. Pluripotent stem cells derived from mouse and human white mature adipocytes. Stem Cells Transl Med. 2014;3:161–71.
70. Kazama T, Fujie M, Endo T, Kano K. Mature adipocyte-derived dedifferentiated fat cells can transdifferentiate into skeletal myocytes in vitro. Biochem Biophys Res Commun. 2008;377:780–5.
71. Hatley ME, Tang W, Garcia MR, Finkelstein D, Millay DP, Liu N, et al. A mouse model of rhabdomyosarcoma originating from the adipocyte lineage. Cancer Cell. 2012;22:536–46.
72. Lagha M, Brunelli S, Messina G, Cumano A, Kume T, Relaix F, et al. Pax3: Foxc2 Reciprocal repression in the somite modulates muscular versus vascular cell fate choice in multipotent progenitors. Dev Cell. 2009;17:892–9.
73. Warming S, Costantino N, Court DL, Jenkins NA, Copeland NG. Simple and highly efficient BAC recombineering using galK selection. Nucleic Acids Res. 2005;33:e36.
74. Summerbell D, Ashby PR, Coutelle O, Cox D, Yee S, Rigby PW. The expression of Myf5 in the developing mouse embryo is controlled by discrete and dispersed enhancers specific for particular populations of skeletal muscle precursors. Development. 2000;127:3745–57.
75. Carvajal JJ, Keith A, Rigby PW. Global transcriptional regulation of the locus encoding the skeletal muscle determination genes Mrf4 and Myf5. Genes Dev. 2008;22:265–76.
76. Sambrook J, Fritsch EF, Maniatis T. Molecular cloning, a laboratory manual. 2nd ed. New York: Cold Spring Harbor Laboratory Press; 1989.
77. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science. 2002;295:1306–11.
78. Hagège H, Klous P, Braem C, Splinter E, Dekker J, Cathala G, et al. Quantitative analysis of chromosome conformation capture assays (3C-qPCR). Nat Protoc. 2007;2:1722–33.
79. Noordermeer D, Leleu M, Splinter E, Rougemont J, De Laat W, Duboule D. The dynamic architecture of Hox gene clusters. Science. 2011;334:222–5.
80. Splinter E, de Wit E, van de Werken HJ, Klous P, de Laat W. Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: from fixation to computation. Methods. 2012;58:221–30.
81. Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol. 2000;132:365–86.
82. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12:996–1006.
83. Russel D, Lasker K, Webb B, Velázquez-Muriel J, Tjioe E, Schneidman-Duhovny D, et al. Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. PLoS Biol. 2012;10:e1001244.
84. de Wit E, Vos ES, Holwerda SJ, Valdes-Quezada C, Verstegen MJ, Teunissen H, et al. CTCF binding polarity determines chromatin looping. Mol Cell. 2015;60:676–84.
85. Kaaij LJ, Mokry M, Zhou M, Musheev M, Geeven G, Melquiond AS, et al. Enhancers reside in a unique epigenetic environment during early zebrafish development. Genome Biol. 2016;17:146.

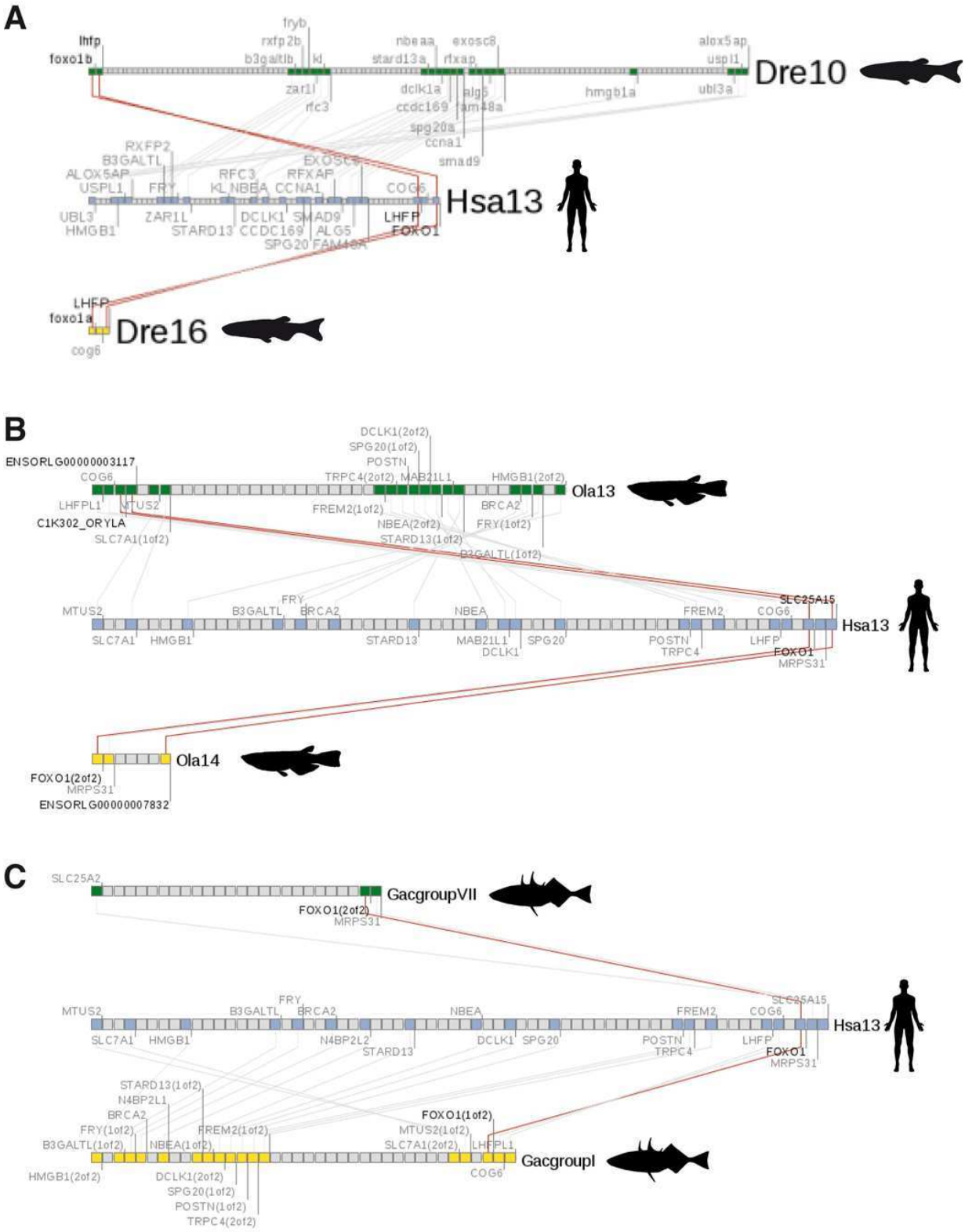


Figure S1: Orthologous pairwise clusters involving the *Foxo1* gene. Syntenic cluster output generated using the Synteny Database using the (A) *Danio rerio* (Zebrafish), (B) *Oryzias latipes* (Medaka) or (C) *Gasterosteus aculeatus* (Stickleback) as source genomes and the Human as outgroup. Sliding window size: 50 genes.

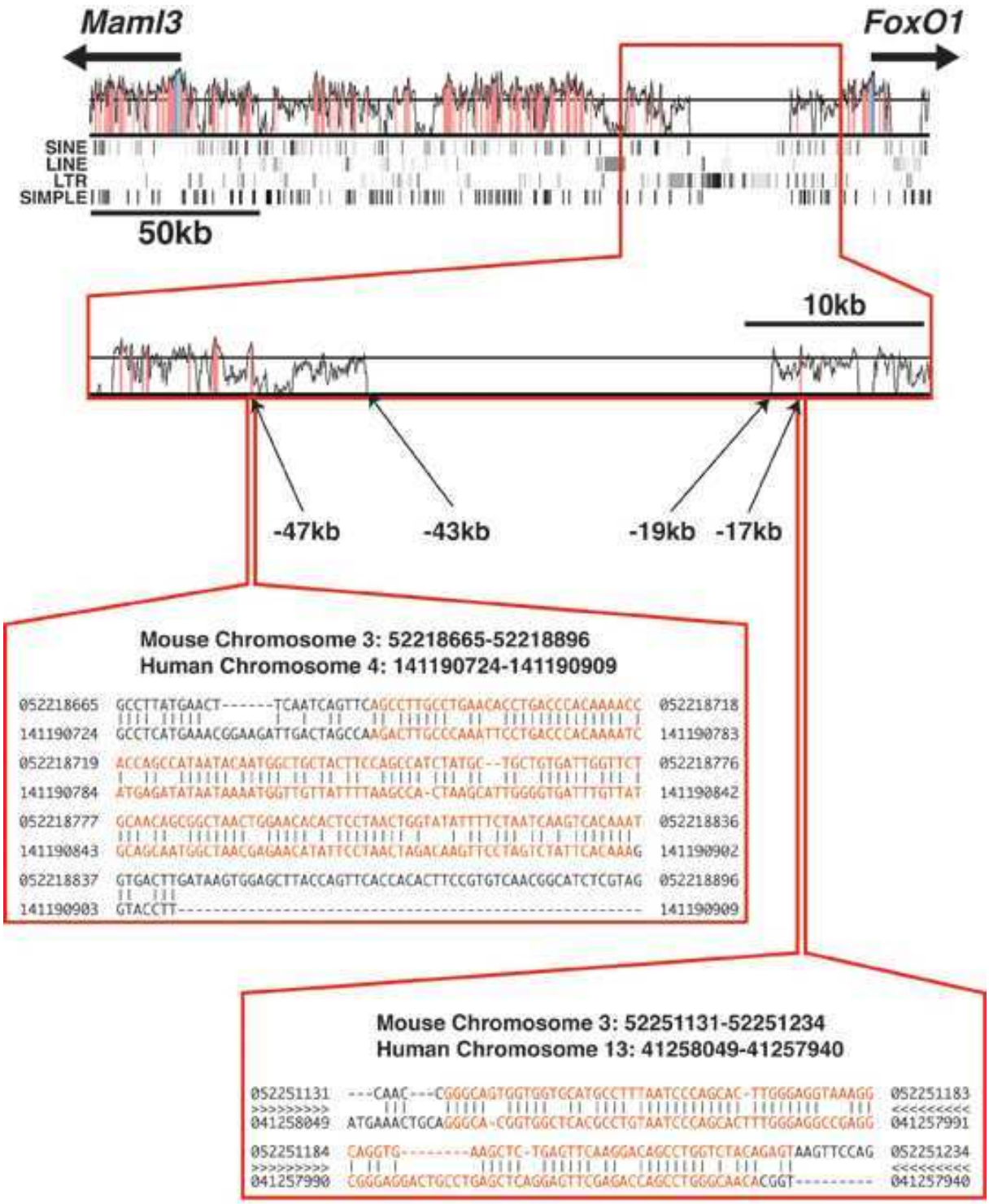


Figure S2: Conservation analysis across the *FoxO1-Maml3* intergenic region. VISTA BROWSER (http://pipeline.lbl.gov; 2017) analysis of the region of interest. The base genome is mouse (mm10), compared to human (hg19). Red peaks are ECRs based on standard parameters (70% identity over 100bp; 100bp sliding window). Underneath, the output from the UCSC Genome Browser showing the location repeats. Note the large LTR in the region without homology. To the left of the LTR, all peaks correspond to sequences in human chr4, while peaks situated to the right have are homologous to human sequences on chr13. The first of such sequences is shown in the alignments underneath, indicating the chromosomal positions; in red, conserved bases. The positions of the furthermost conserved sequences in relation to the *FoxO1* TSS are indicated.

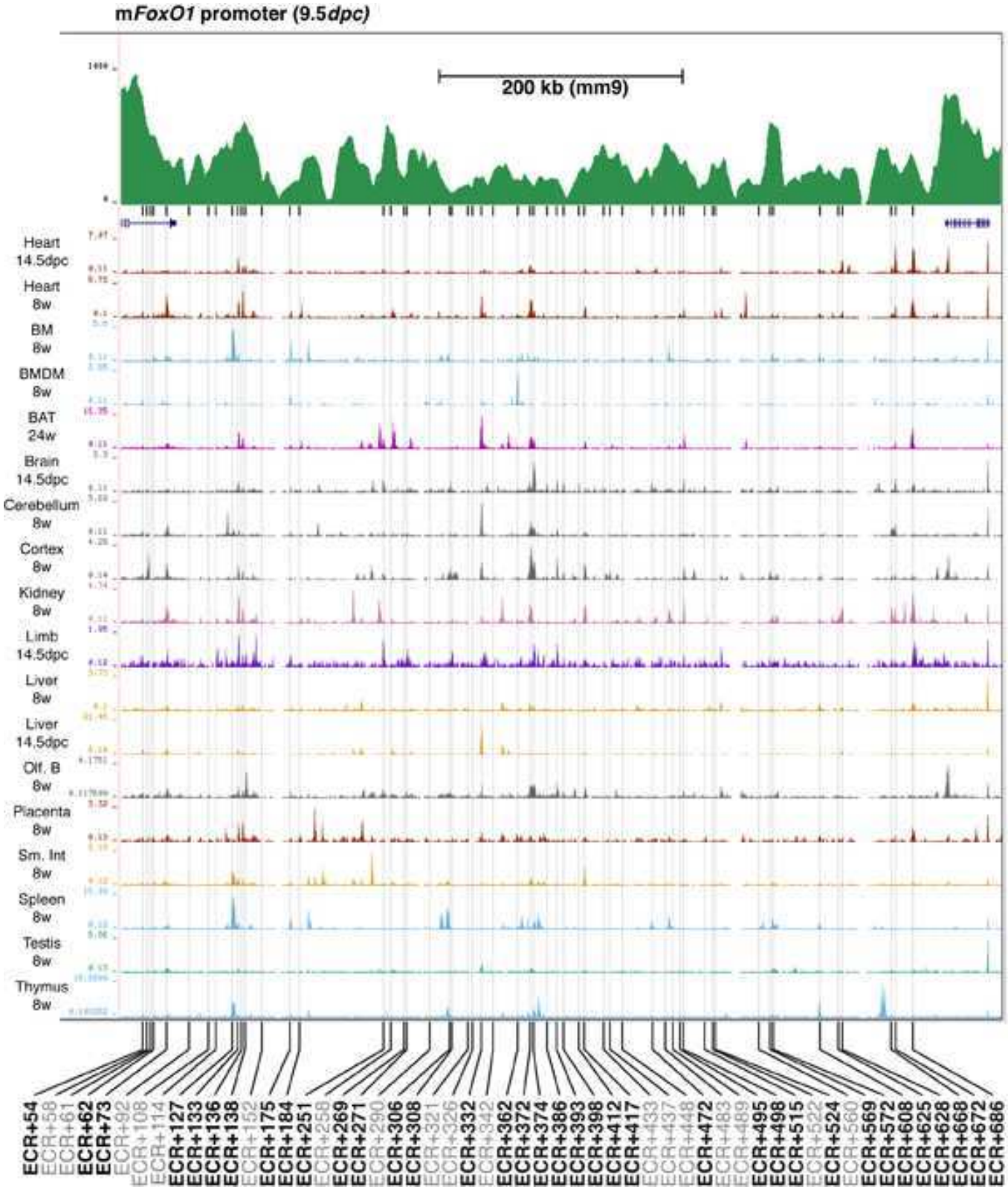


Figure S3: ECRs identified in the *Foxo1* region downstream of the RMS breakpoint and associated H3K27ac marks. In green, the 4C-seq profile when using the *Foxo1* promoter as a viewpoint on chromatin obtained from 9.5 d*pc* mouse embryos. Note how most peaks co-localise with ECRs throughout the landscape. H3K27ac marks in different mouse tissues are also included. Those ECRs co-localising with strong marks are indicated in bold.

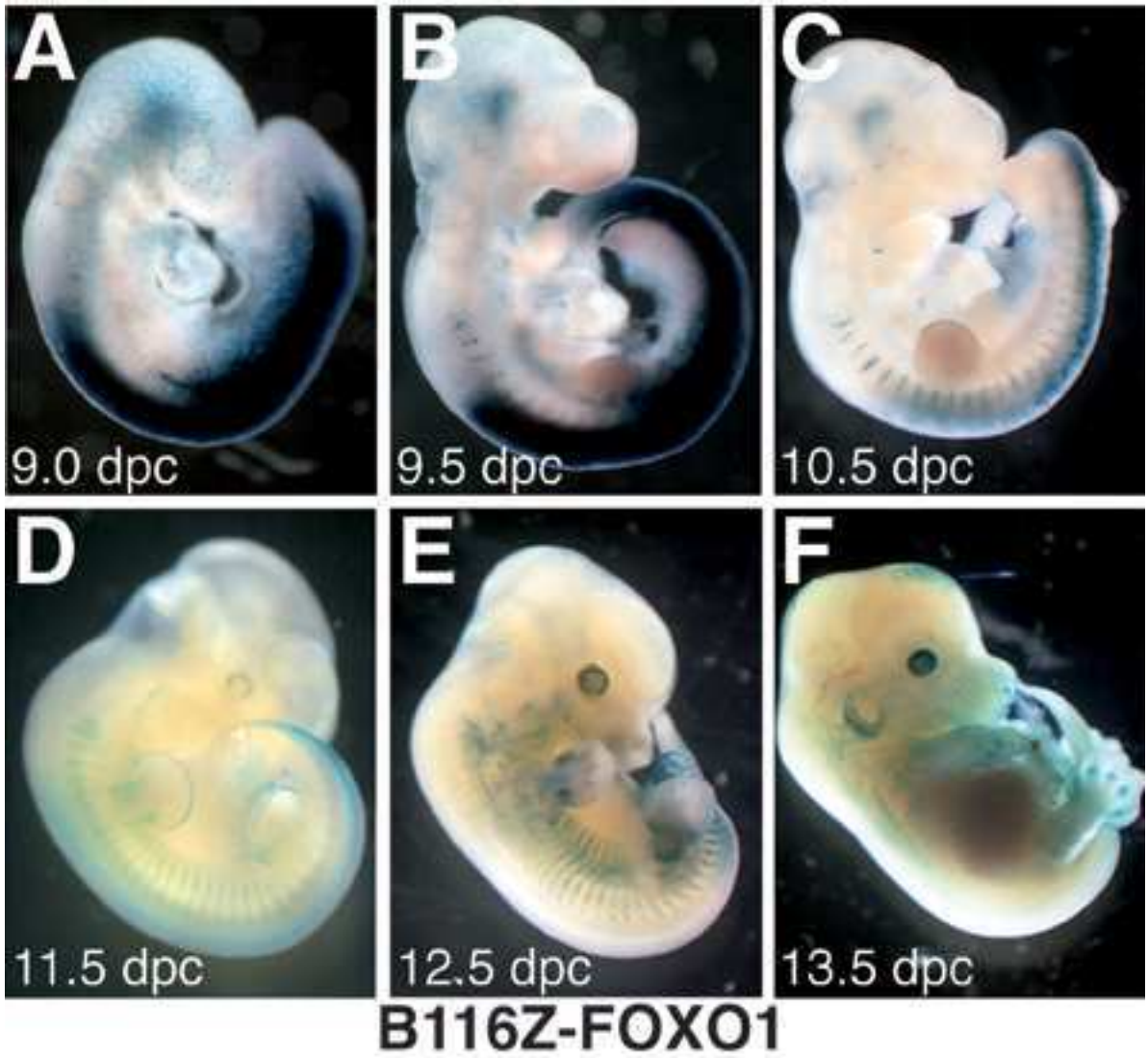


Figure S4: Time-course of embryos carrying the B116Z-Foxo1 reporter construct. Expression starts before 9.0 d*pc* (A) in the neural tube, neural crest and migrating neural crest. At 9.5 d*pc* (B), high levels of transgene expression are detected in the neural tube, the mesonephros, and the vitelin vein, central myotome, head neural crest cells and cells migrating into the forelimb. At 10.5 d*pc* (C), neural tube and mesonephros expression is downregulated, maintained in the foregut and the myotome of cervical and thoracic somites and activated in the AER. At 11.5 d*pc* (D), expression is detected in the myotome, AER, pharyngeal region of the foregut and the posterior half of the neural tube. At 12.5 d*pc* (E), expression is mainly restricted to skeletal musculature, with activation in retina, lens vesicle, pre-cartilage primordia of forelimbs, umbilical cord and neural tube in the tail region. At 13.5 d*pc* (F), the transgene is downregulated in all skeletal muscles, maintained in pre-cartilage primordia of phalangeal bones, and activated in the nasal pits, head epidermis, and follicles of the vibrissae and sinus of sensory facial hairs.

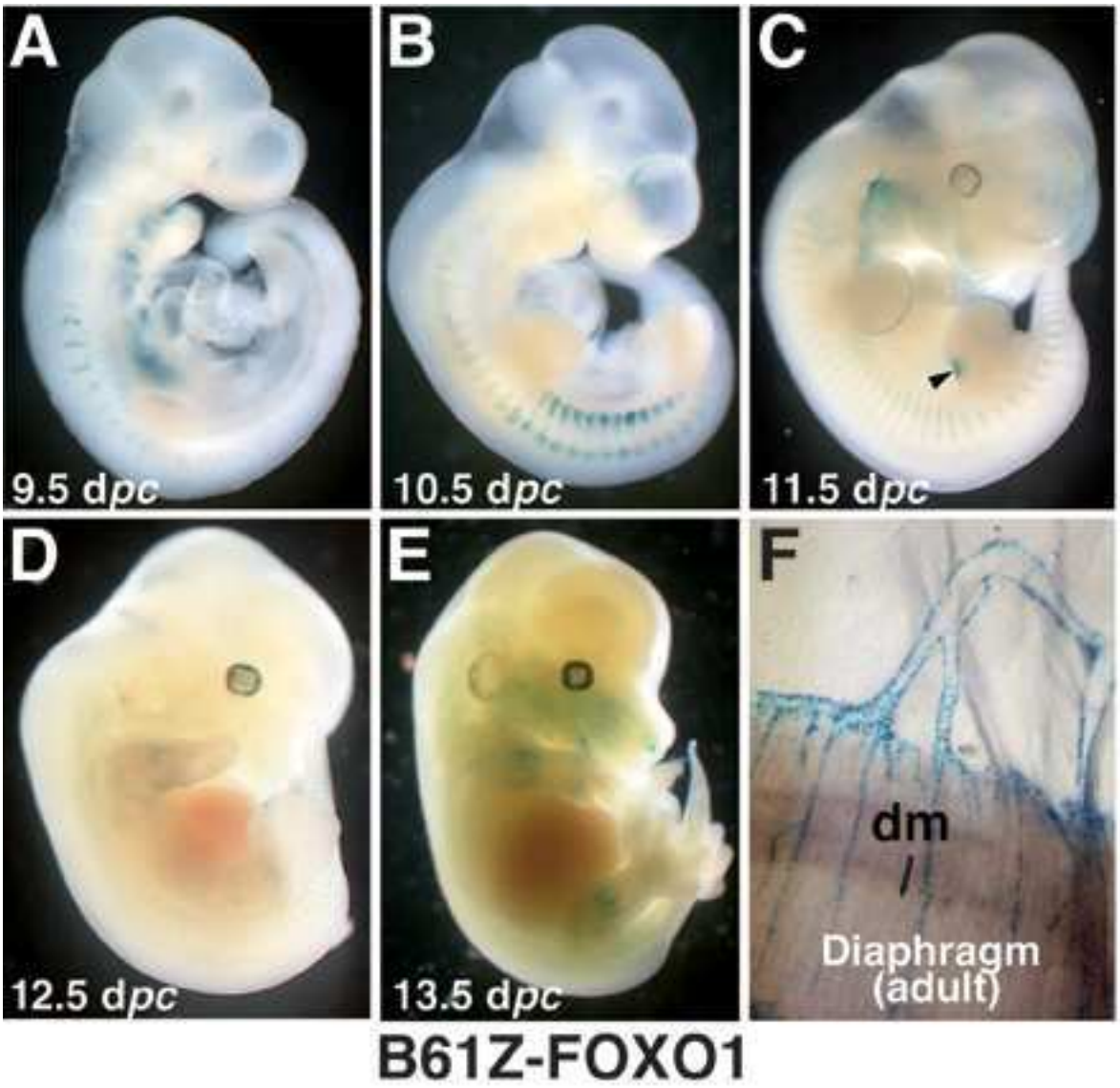


Figure S5: Time-course of embryos carrying the B61Z-Foxo1 reporter construct. Expression is observed at 9.5 d*pc* (A) in the myotome of cervical somites, and fore- and hind-gut, at 10.5 d*pc* (B), in cervical and thoracic somites, gut, vitelin vein and AER, at 11.5 d*pc* (C), in myotome, AER, pharyngeal region, and a hindlimb rostral domain (arrowhead). By 12.5 d*pc* (D), the transgene is downregulated. At 13.5 d*pc* (E), expression corresponds to skeletal muscle, retina, lens vesicle, and nasal pits. In the adult (F), there is strong vasculature expression.

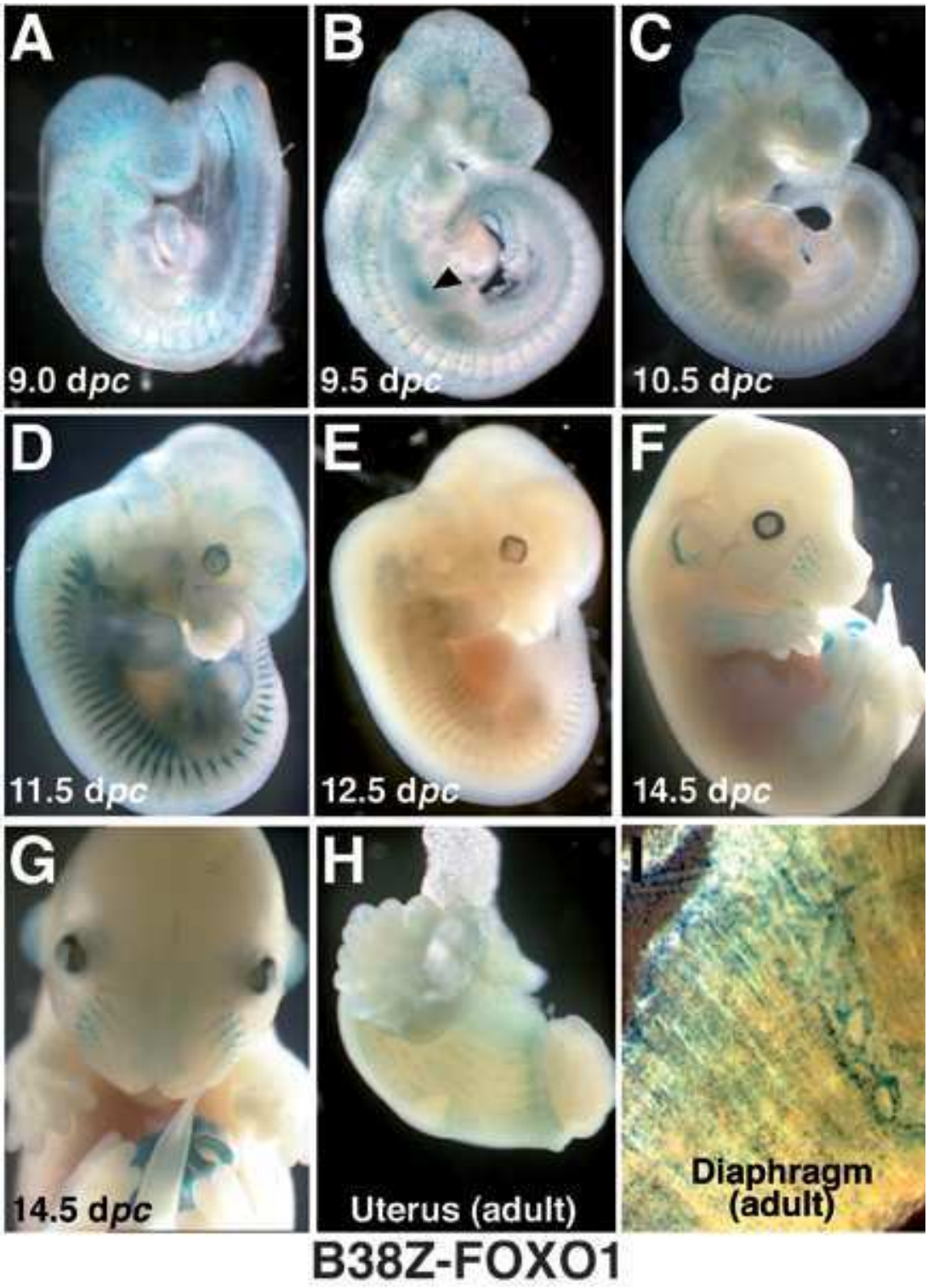


Figure S6: Time-course of embryos carrying the B38Z-Foxo1 reporter construct. Expression is observed at 9.0 d*pc* (A) in vascular precursors, at 9.5 d*pc* (B) in all the vasculature and the foregut (arrowhead). At 10.5 d*pc* (C), expression is maintained in vascular precursors and foregut. At 11.5 d*pc* (D), vascular expression downregulates and myotomal expression is upregulated. At 12.5 d*pc* (E) expression is mainly in skeletal muscle lineage. At 14.5 d*pc* (F), expression is faintly maintained in limb musculature, upregulated in ear cartilage, nasal pits, vibrissae, sensory facial hair follicles, and umbilical cord (G). In the adult, expression is observed in smooth muscle (H), skeletal muscle and vasculature (I). *dm*: diaphragm muscle.

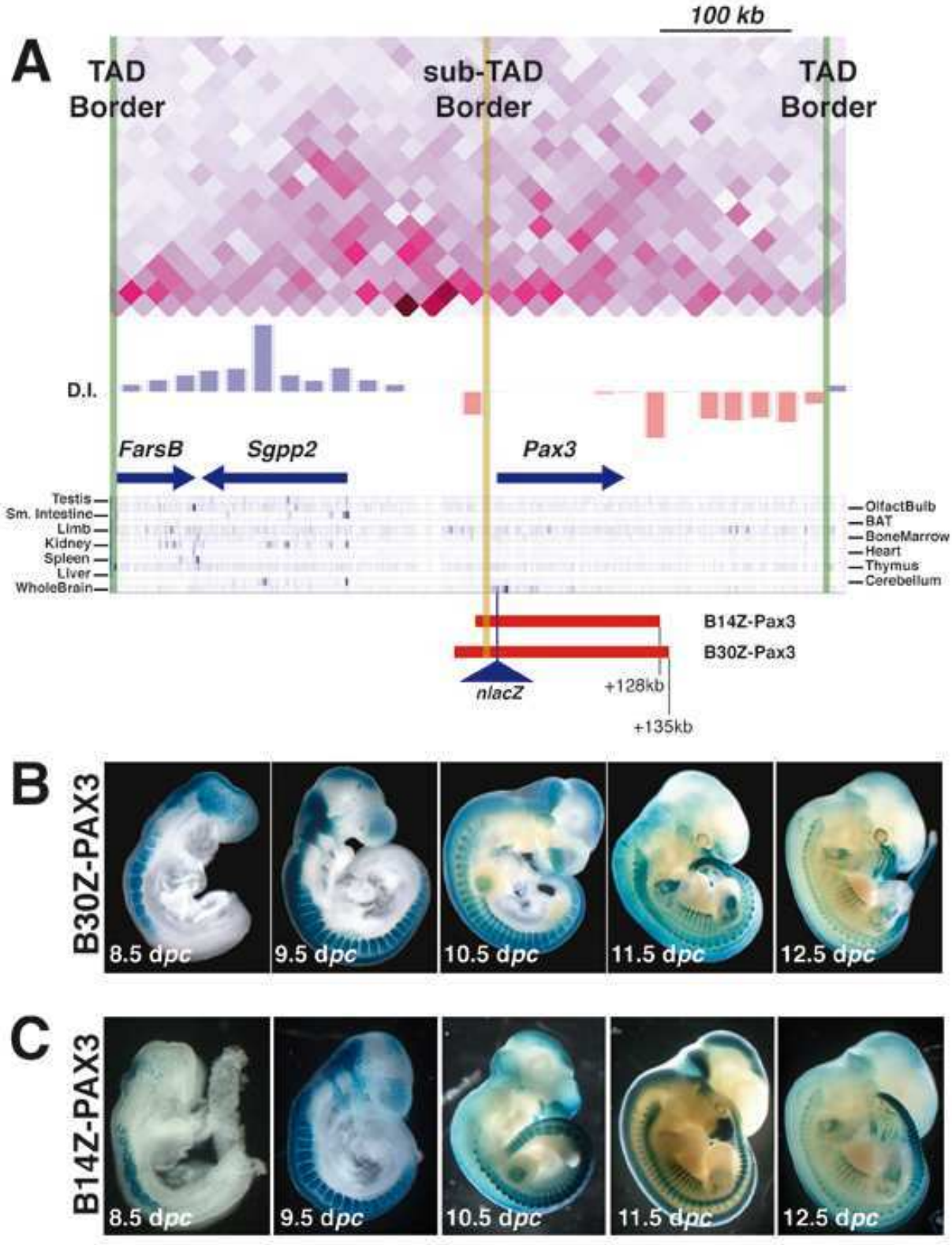


Figure S7. Recapitulation of *Pax3* endogenous expression pattern by a BAC carrying 30kb of upstream sequences. (A) Detail of the Hi-C data from mouse ES cells between the TAD borders (green boxes) and showing the position of the subTAD border and the D.I. analysis output. H3K27ac marks in different mouse tissues are shown underneath, as well as the position of the three coding genes in the region and the relative positions of the two BAC clones used in the study. The 5' ends of the clones cross the subTAD border. (B) Expression patterns of B30Z-PAX3 and (C) B14Z-PAX3 from 8.5 d*pc* to 12.5 d*pc*.

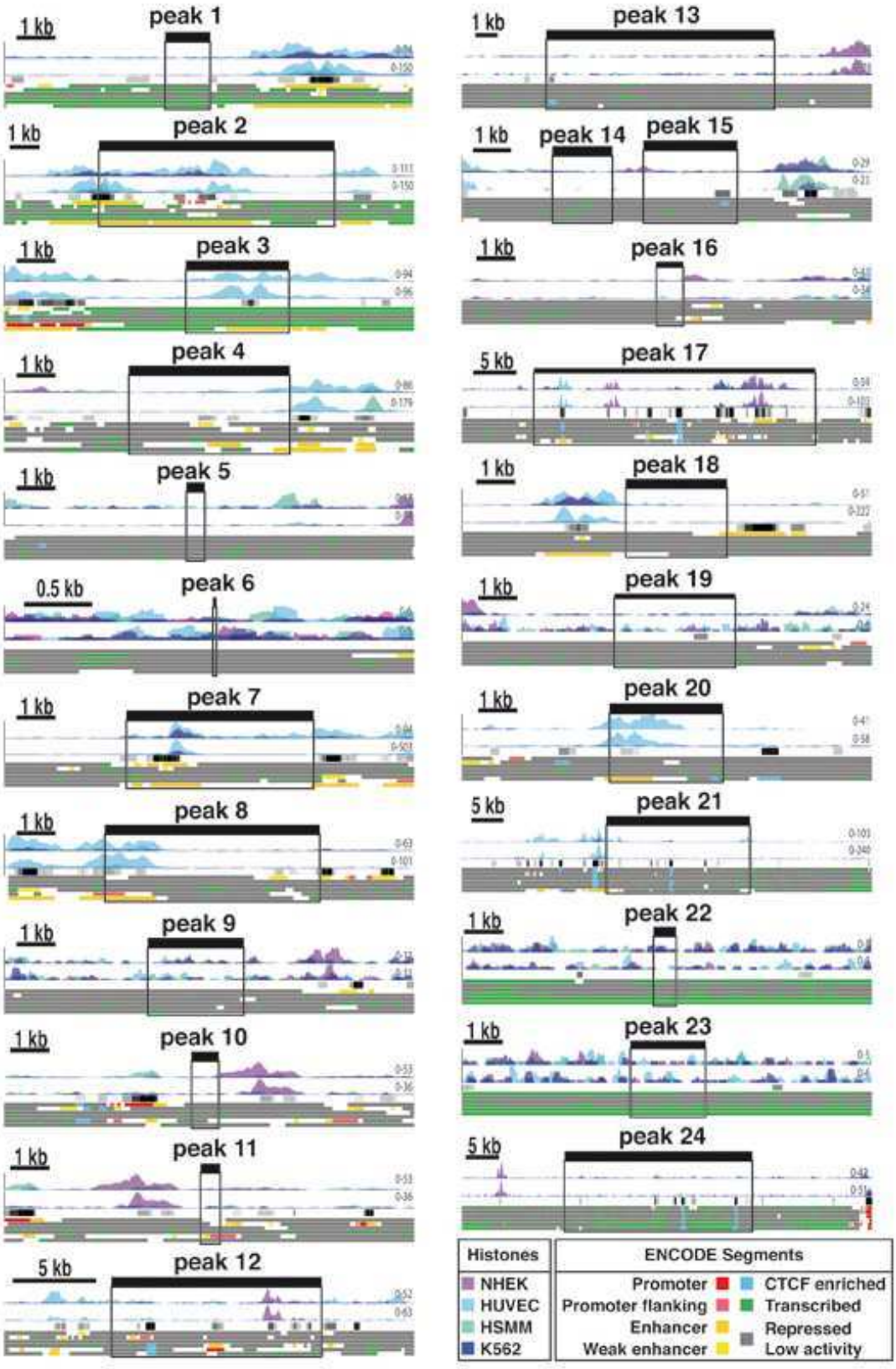
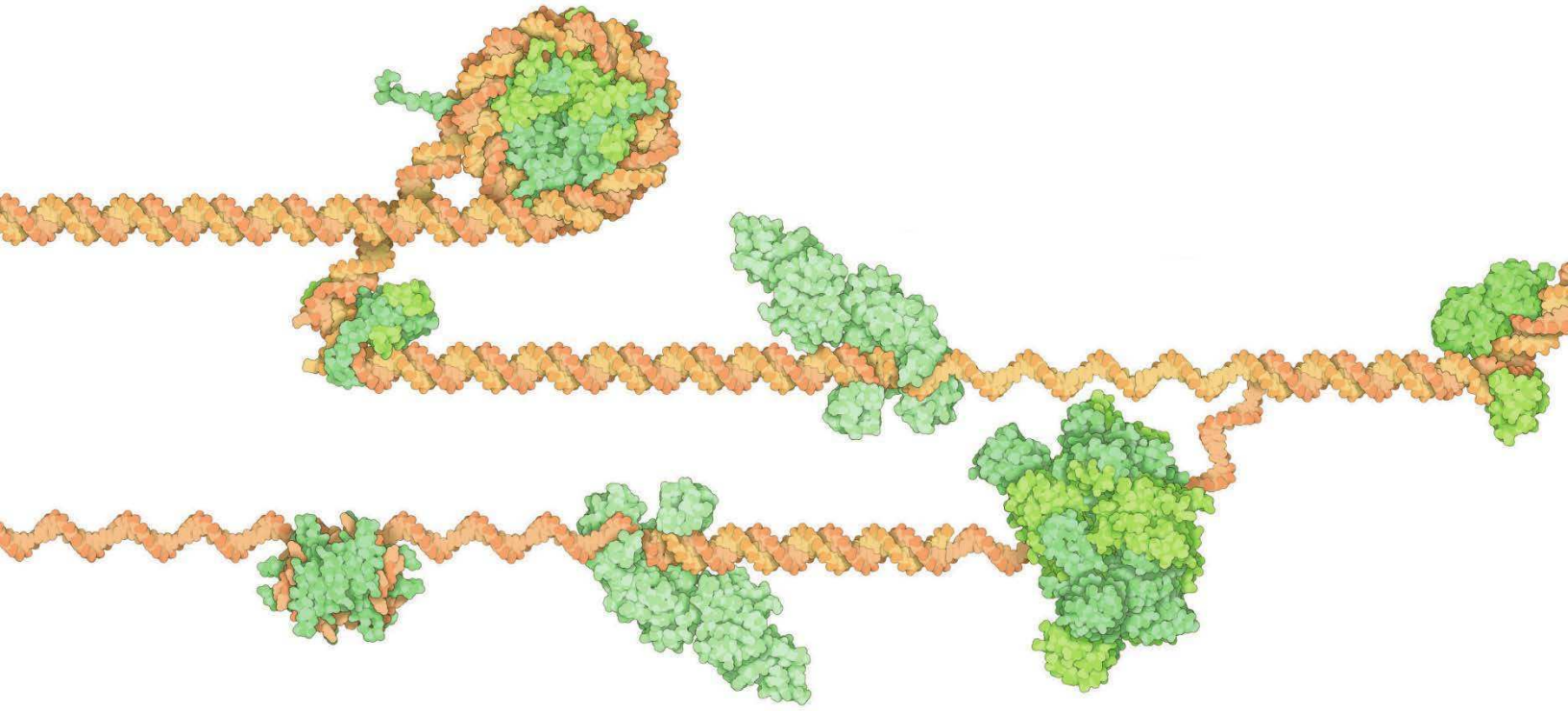
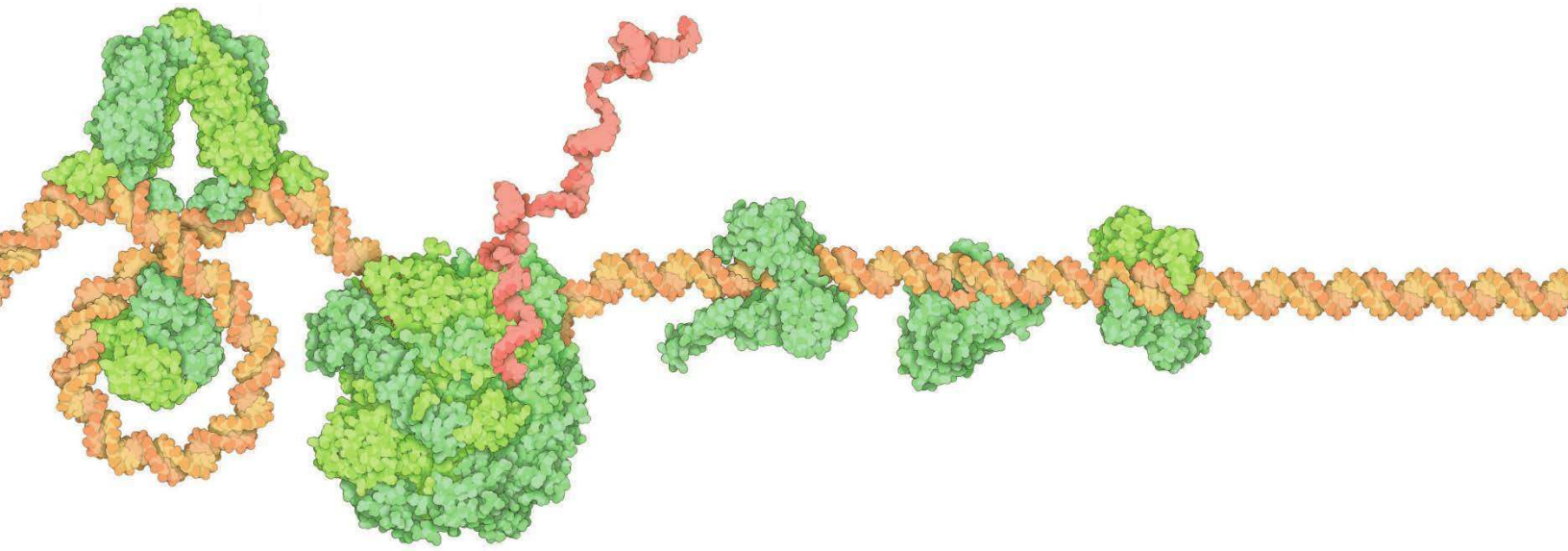


Figure S8. Peaks of interaction established by the *PAX3* promoter at the *FOXO1* locus in RMS cells. Graphical representation of the 4C-seq contacts identified by the Peak Calling algorithm. The peaks are outlined as a box; underneath are representations of H3K4me1, H3K27ac, Transcription Factor ChIP-seq and predicted ENCODE regions. A colour key in included.



Discussion



1 Challenges in the elucidation of macromolecule structures

Most macromolecules in the cell function by forming large assemblies, and thus, information on these big macromolecular complexes is vital to understand the mechanisms underlying their function, but their structural characterization is even more difficult (Ahnert et al., 2015; Nooren & Thornton, 2003; Whitty, 2008). Many biological macromolecules are not compliant to mainstream structural biology methods like X-ray crystallography, NMR, or EM for many reasons: they are unstable, insoluble, too large or small or not of adequate purity, for example. Recent advances in X-ray crystallography and EM techniques have allowed the elucidation of high resolution structures of large protein complexes, but progresses are slow and the structure of a big number of dynamic protein assemblies remain unsolved. These complexes usually have fast association-dissociation kinetics and flexible subunits like unfolded or intrinsically disordered ones. Complexes that can undergo rapid conformational rearrangement are also included in this group (Boehr et al., 2009; Chen et al., 2011; Kalodimos, 2012; Marsh et al., 2012; Mittag et al., 2010; Tompa & Fuxreiter, 2008; Tzeng & Kalodimos, 2011). NMR spectroscopy is the best technique so far to study such dynamic complexes but, until recently, it has been limited to low molecular weight ones (up to approximately 35 kDa).

To overcome these problems, hybrid and integrative methods have stepped up (Alber et al., 2007; C. V. Robinson et al., 2007; Russel et al., 2012; Sali et al., 2003; Schneidman-duhovny et al., 2014; Stengel et al., 2012) elucidating many 3D structure of protein complexes that would be difficult to resolve with conventional methods (Alber et al., 2007; Kosinski et al., 2016; Lasker et al., 2012; Robinson et al., 2015).

Like in the case of proteins, resolving chromatin structure has proven difficult although the reasons are different. The main problem of its structure is its highly dynamic nature. Single-cell studies have revealed that chromatin structure varies between cells and in time (Nagano et al., 2013; Stevens et al., 2017). Thus, our main goal is not to elucidate its structure but rather to unravel the average organization and dynamics of the different chromatin domains. For this purpose it is required to understand the role of the proteins and RNAs that are related to chromatin functions as well as their interactions and involvement in the compaction of the chromatin. In addition, some of these players implicated in its organization can vary across different species, difficulting our understanding of its mechanisms and broadening its study.

In this regard, integrative approaches are also emerging for 3D chromatin modeling. These approaches usually define a set of distance restraints from 3C based and live cell imaging data that are afterwards satisfied using optimization algorithms (Ay et al., 2014; Baù et al., 2011;

Duan et al., 2010; Giorgetti et al., 2014; Kalhor et al., 2011; Umbarger et al., 2011; Wang et al., 2015).

2 What did we learn from the exocyst structure?

In this work, we have provided coarse grained 3D models for the yeast exocyst *in vivo*, using a multidisciplinary and integrative approach that combines cell engineering, quantitative fluorescence microscopy, and bioinformatics. The method we developed uses distances between tags located at each termini of each proteins measured at great precision, which are integrated with the structural features of the protein subunits.

We have shown that the exocyst is a stable complex with a hand-like shape, composed of rod-shaped subunits that are interlaced by their N-termini in the core of the complex and with their C-termini facing outward (except for Sec10). The exocyst structure also provides evolutionary insights on its origin, since the symmetry found between Exo70-Sec6 and Sec3-Sec8 dimers supports the idea that the ancestral exocyst complex was composed of fewer subunits that duplicated and diverged (Croteau et al. 2009; Dacks et al., 2008). In addition, we proposed a model of how the exocyst is able to tether the vesicle to the plasma membrane. The model we propose could be used as a base for future experiments since it is in agreement with most of the previous studies but it also raises new questions. We propose that a maximum of approximately 20 exocysts could be cooperating in the tethering process, but the molecular mechanisms to coordinate them are still unknown. The structure and the proposed tethering model also helps in the design of future experiments to characterize binding domains, not only intra complex but also with other molecules.

Furthermore, the hybrid approach that we have used to elucidate the 3D architecture of the exocyst could help to obtain the structure of other large multi-subunit assemblies *in vivo*, as we also resolved the 3D architecture of the conserved oligomeric Golgi (COG) complex, a related multi-subunit protein complex involved also in intracellular transport. Our method allows to elucidate structures of complexes that are difficult to purify. It is also useful for those that are interacting with other cellular components and cannot be reconstituted *in vitro*. Moreover, the capabilities of getting a reliable 3D structure by this approach could be enhanced combining data from other experiments like EM, X-ray crystallography or cross-linking coupled to mass spectrometry.

The precision of our measurements is in the nanometer scale, making the approach particularly suitable for the study of large multisubunit assemblies, however, the number and type

of subunits is a potential limitation of our method. Indeed, reasonable number of subunits are needed to get enough distances to trilaterate the position of the fluorophores fused to their termini. At least the measurement of four distances from different fluorophores to a particular fluorophore are necessary to pinpoint its position in the 3D space. But the lack of distances to locate a particular fluorophore can be overcome using other different type of data or strategies. For instance, due to lack of distance restraints, we could have two populations of 3D models of a complex where a particular subunit has a different localization in each population, and one of them is unbound to the assembly. We could make the system discard those conformations, setting a restrain that penalizes the unbound subunits. In fact, our approach is also designed for hetero-multimer protein complexes. Complexes containing multiple copies of the same subunits should be handled differently, as has been done with the nuclear pore complex (Alber et al., 2007).

This approach has been used to study the structure of the exocyst and the COG, two multi protein complexes of the CATCHR (Complex Associated with Tethering Containing Helical Rods) family that are known to be composed of rod-shaped proteins that protrude from the core of the complex. Both complexes have an open organization and the fluorophores fused to the N and C termini of the subunits did not prevent the exocyst from assembling. Furthermore, our method requires that the complex can be recruited to the anchoring platform in quantities that are large enough to be imaged, making the study of nuclear complexes or assemblies that contain transmembrane proteins difficult, but not impossible. Our method could also have some problems if the fused fluorophores obstruct the assembly of the protein complex to study, but, on the contrary, it could be helpful elucidating the structure of complexes with open conformations. This is the case of another member of the CATCHR family, the Golgi-associated retrograde protein (GARP). It is composed of 4 rod-like shaped subunits and their atomic structure is unknown, although some parts have been crystallized (Fridmann-Sirkis et al., 2006; Pérez-Victoria et al., 2010; Vasan et al., 2010). The method could be applied in this case but the fact that it is composed of only 4 subunits could be problematic when trilaterating the position of the N and C tagged fluorophores.

A recent article has generated EM images of the exocyst at a high resolution and fitted atomic structure of the subunits in these images (Mei et al., 2018). They used the already crystallized fragments of the subunits and comparative modeling to predict the structure of the missing subunits. This study, which is in good agreement with our work, is supported by a previous work (Heider et al., 2015), showing the exocyst in a closed conformation, where all the subunits are piled together and meet together in the core of the complex. But the exocyst could adopt many conformations (Hsu et al., 1998) and our work could be showing the active conformation of the

complex. So, a straightforward work would be to compare and reevaluate our model with these works, and see if a transition from the closed to the open conformation could be possible.

3 Contribution of 4C-seq data to the chromatin structure problem through 3D models.

The interpretation of the 3C-based methods, including 4C-seq data, is difficult (Dekker et al., 2013). When many cells are used, they provide a frequency of contacts between fragments, an average, and there is no direct and verified way to transform these frequency of contacts into distances. To minimize the possible biases of these methods, the integrative approaches can be useful. They try to fulfill most of these distances and, in the process, they are able to single out incoherent or wrong data generated due to the experiment. But all this filtering is, at some point, dependent on the representation of the chromatin. Many methods have considered that the chromatin is in the 30 nm state on average but recent works have shown that this width might be smaller (Ou et al., 2017). The accurate representation of the chromatin is important in order to get reliable 3D models, but, in this regard, data coming from different techniques could be used. Techniques like ATAC-seq show the accessibility of the chromatin at high resolution and could potentially be used in the future to specifically represent the chromatin at the bp level. In the same way, some epigenetic marks are informative of the level of compaction that the chromatin can have, and these data could be used as a proxy of the volume that certain chromatin fragments are occupying. There is not a unique way of addressing this problem, but the integration of multiple type of data can definitely help. In this regard, we developed a tool that the scientific community can use to generate chromatin structures of particular regions.

The computational method we developed for the 3D modeling of the exocyst is, in the end, based in the distances between points that are used as restraints. The goal is to define the position of those points that best fits with the input data after optimization iterations. We have adapted the same approach to elucidate the architecture of chromatin fragments, using 3C-based technologies data. As explained earlier, 3C-based methods provide the frequency of contacts between different regions of the chromatin, and this data can be interpreted as spatial distance between those regions.The 4C-seq data provide the distances between the viewpoints and the rest of the fragments, and, as a result, we need a small number of 4C-seq experiments to have enough distances to each fragment (represented as beads) to be able to trilaterate in the 3D space.

Taking this into consideration, we developed a software, 4Cin, that uses this principle and 4C-seq data to generate 3D chromatin structures. The idea of integrating distances derived from 3C-based data is not new, and as mentioned in the introduction, there are many algorithms that generate 3D models of the chromatin using Hi-C or 5C data. 4C-seq data is more accessible in many cases, cheaper and easier to analyze (Results section 2) than most of the other 3C-based data. Exploiting these benefits, we applied 4Cin to study evolution (Results section 2 & 3), diseases (Results section 4) and structural variations (Results section 2 & 4). In addition, 4Cin is able to generate a Hi-C like contact maps of our region of interest, called virtual Hi-C, using 4C-seq data. In all these cases, we provided a set of 3D chromatin models and, for illustration purposes, a representative model was extracted from the dataset.

3.1 Studying the chromatin through the 3D models

Our results do not show a highly variable population of chromatin structures, which was not expected, as it is known that the chromatin is a very dynamic macromolecular complex (Nagano et al., 2013; Stevens et al., 2017). But, interestingly, the models generated through our pipeline show the variable nature of the chromatin structure, since we cannot find 3D chromatin models where all restraints are satisfied, even when a loose/permissible cut off is used. This means that the data are explaining many different models. The opposite happens in protein structure modeling, where only a model (or a few of them) can explain all the data (Erzberger et al., 2014; Fernandez-Martinez et al., 2016; Kosinski et al., 2016). Ultimately, 3C-based experiments that are not done in single cells, could be showing the probability of all chromatin fragment pairs to be in close contact between each other in millions of cells. The fact that we cannot satisfy all restraints in the optimization process, suggests that the structure of the chromatin in those cells is very different between them, as already shown (Giorgetti et al., 2014). But approaches that use data coming from many cells are also necessary in case we are studying long range contacts, because we cannot determine if these contacts are happening in the majority of cells, unless many single-cell Hi-C's are generated. Due to its integrative nature, our algorithm shows the average of all this variability, filtering out the less probable contacts and showing the overall structure, or, in other words, the 3D chromatin organization that most of the cells have most of the time. Since we are studying chromatin loci with lengths comprised between 0.5 to 3 Mb, we are able to predict the average TAD organization of these regions. So, our approach, that shows the average organization of the chromatin, is very useful since we are seeing TADs as the "fenced playgrounds" of promoters and enhancers, and not as a physical and structural entity. In this regard, and with the resolutions that we can reach with our tool, we could try to find out if the

structure of the chromatin at the Mbp level is important in evolution, comparing conserved GRB in a genome of a particular species or between different species. If certain genes are regulated by the same enhancers in different species, the structure of the chromatin for both species should be similar, and we could study the relationship between these regulatory landscapes and chromatin structure. In order to compare these particular and similar chromatin structures, tools and approaches used to study protein structures could be used, but for that, we need to think of the chromatin as a static macromolecule, or at least, take the average conformation of the chromatin as a static structure. We could even study if the structure of the chromatin could had been the driver of certain enhancers' development, due to spatial proximity of certain regions, widening the regulatory landscape of some genes.

As we already mentioned in the introduction, TADs are quite conserved between species or tissues, therefore, genes and their regulatory landscapes are conserved. But how did the whole regulatory landscape conserve during evolution, as a unit?

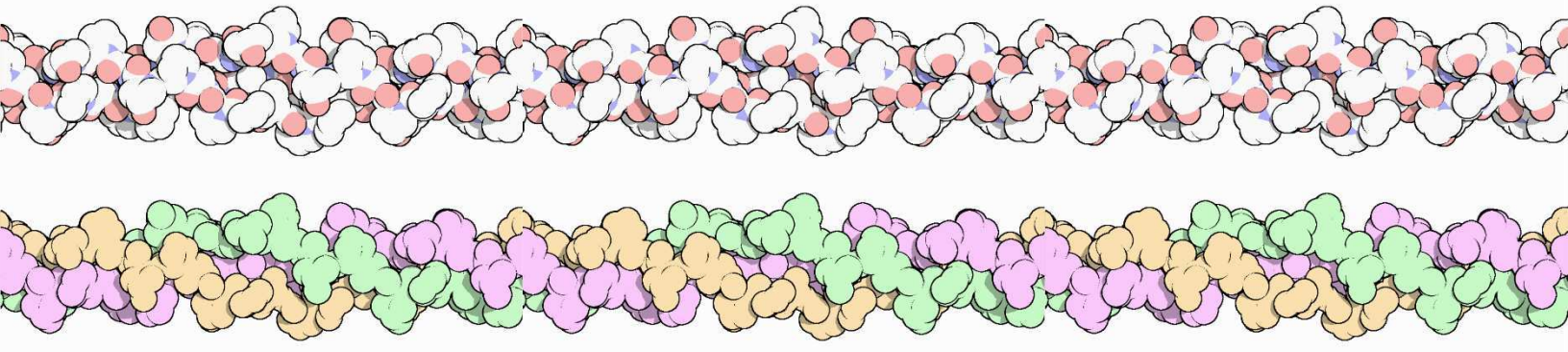
The loop extrusion model explains how cohesin is extruding the chromatin until it meets CTCF proteins. This extrusion entangles the chromatin in these loop anchors composed of cohesin and CTCF, generating torsional stress. A recent work (Canela et al., 2017) showed that those entanglements are vulnerable to DNA breaks mediated by the topoisomerase 2 enzymes (TOP2) which is expected to relieve the stress. These TOP2 mediated double strand breaks (DBS) can occur simultaneously in many regions of the chromosome and even in different chromosomes. Therefore, these intra or inter-chromosomal translocations, could occur in TAD boundaries and contribute to evolution, reshuffling these TADs in the genome. In fact, Hi-C maps of orthologous mouse and dog genomes showed that there are many insertions and deletions in CTCF/cohesin loop anchors (Vietri-Rudan et al., 2015), suggesting that TOP2 could have contributed to these rearrangements. Many syntenic regions are conserved also between different species that are located in different chromosomes (Irimia et al., 2012, 2013) and could also have been rearranged by this mechanism. This could explain why TADs are conserved between species, not only the genes, but also their enhancers. Canela and colleagues also showed that translocation breakpoint regions in many cancers are enriched in CTCF loop anchors. In our work (Results section 4) we showed how a chromosomal translocation leads to alveolar rhabdomyosarcoma and how the translocation breakpoint occurs in the middle of two genes that are localized near TAD borders. It is well known that TAD borders are enriched in CTCF, which suggests that this cancer could also be developed due to TOP2 mediated DBS.

4 Integrative approaches are necessary to understand the chromatin

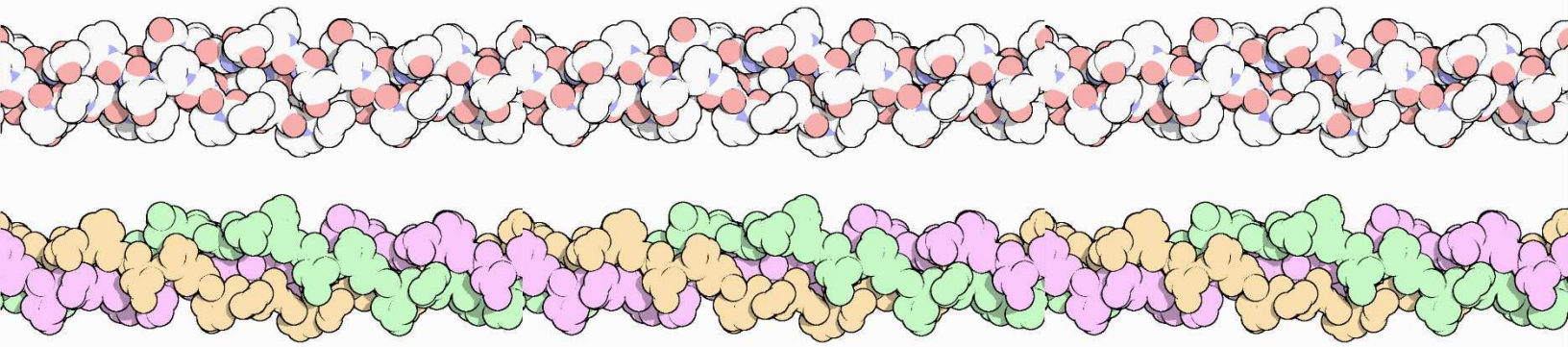
Integrative approaches could be helpful to elucidate some of the open questions on the dynamics of the chromatin. For instance, TAD formation mechanisms (or even TAD calling) are still a matter of debate since there are many components participating. A recent work suggested that architectural proteins do not form strong domains by themselves, but A/B compartment segregation is the main driver (Rowley et al., 2017). Still, they do not rule out the possibility that architectural proteins could be playing a role in conjunction with transcription. Moreover, another study showed that some TADs could be predicted with GRBs, suggesting that they could also be driven by additional components (Harmston et al., 2017).These GRBs and their constituent CNEs are defined by sequence conservation exclusively and this is stable between cell types, being a very good predictor of some TADs.

In conclusion, the folding of the chromatin is driven by epigenetics, architectural proteins, transcription, gene promoters, their enhancers and even CNEs acting as enhancers, and all of these factors are important, at least in higher eukaryotes. In this regard, integrative approaches are best suited, because many different type of data needs to be taken into account, since all factors are contributing to the chromatin structure.

We need many different approaches and more importantly, the integration of these approaches to reach solid conclusions about the chromatin organization and its dynamics in all their different scales, from cis regulatory dynamics to inter-chromosome scale. As a matter of fact, we need to take into account all the information, in each level of chromatin organization, in different species and tissues, to be able to explain the bigger picture, and we need to do it integratively. The integration of all components has the potential to explain the chromatin folding. Single cell analyses, new techniques, microscopy imaging, EM snapshots… all have been very useful in this regard, but most of the recent works combine many approaches, suggesting that the integrative approaches are more than necessary to understand all about the chromatin.



Conclusions



1) Integrative approaches are able to reconstruct the architecture of multi protein complexes which are difficult to elucidate by other means.

2) The exocyst has an open conformation, with its subunits protruding from the core of the complex to the exterior and can work together with other exocyst complexes binding the vesicle and allowing to contact with the plasma membrane.

3) Integrative approaches are able to generate reliable 3D chromatin models using data derived from 3C-based methods; in the case of 4Cin, it uses 4C-seq data integratively to predict the 3D folding of the chromatin in an efficient and reliable way.

4) Using 3D chromatin models generated by 4Cin, we have shown that the bipartite chromatin organization of Hox clusters is a vertebrate novelty, suggesting that changes in TAD architecture could have played a fundamental role in the evolution of gene regulation and developmental mechanisms in animals and that integrative approaches are useful to study the evolution of genome organization.

5) Chromosomal translocation can lead to the generation of neo TADs, disrupting regulatory domains and generating diseases.

References

Adrian, M., Dubochet, J., Lepault, J., & McDowall, A. W. (1984). Cryo-electron microscopy of viruses. *Nature*, *308*(5954), 32–36. http://doi.org/10.1038/308032a0

Ahmed, K., Dehghani, H., Rugg-Gunn, P., Fussner, E., Rossant, J., & Bazett-Jones, D. P. (2010). Global chromatin architecture reflects pluripotency and lineage commitment in the early mouse embryo. *PLoS ONE*, *5*(5). http://doi.org/10.1371/journal.pone.0010531

Ahnert, S. E., Marsh, J. A., Hernández, H., Robinson, C. V, & Teichmann, S. A. (2015). Periodic Table of Protein Complexes. *Science*. http://doi.org/10.1126/science.aaa2245

Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., … Rout, M. P. (2007). The molecular architecture of the nuclear pore complex. *Nature*, *450*(7170), 695–701. http://doi.org/10.1038/nature06405

Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., … Sali, A. (2007). Determining the architectures of macromolecular assemblies. *Nature*, *450*(7170), 683–694. http://doi.org/10.1038/nature06404

Alberts, B. (1998). The Cell as a Collection Overview of Protein Machines: Preparing theNext Generation of Molecular Biologists. *Cell*, *92*(3), 1–4. http://doi.org/10.1016/S0092-8674(00)80922-8

Alipour, E., & Marko, J. F. (2012). Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Research*, *40*(22), 11202–11212. http://doi.org/10.1093/nar/gks925

Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H., & Shiroishi, T. (2009). Chromosomal Dynamics at the Shh Locus: Limb Bud-Specific Differential Regulation of Competence and Active Transcription. *Developmental Cell*, *16*(1), 47–57. http://doi.org/10.1016/j.devcel.2008.11.011

Andrey, G., Montavon, T., Mascrez, B., Gonzalez, F., Noordermeer, D., Leleu, M., … Duboule, D. (2013). A Switch Between Topological Domains Underlies HoxD Genes Collinearity in Mouse Limbs. *Science*, *340*(6137), 1234167–1234167. http://doi.org/10.1126/science.1234167

Ay, F., Bunnik, E. M., Varoquaux, N., Bol, S. M., Prudhomme, J., Vert, J. P., … Le Roch, K. G. (2014). Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. *Genome Research*, *24*(6), 974–988. http://doi.org/10.1101/gr.169417.113

Baker, D., & Sali, A. (2001). Protein structure prediction and structural genomics. *Science*, *294*(5540), 93–96. http://doi.org/10.1126/science.1065659

Bantignies, F., Roure, V., Comet, I., Leblanc, B., Schuettengruber, B., Bonnet, J., … Cavalli, G. (2011). Polycomb-dependent regulatory contacts between distant hox loci in drosophila. *Cell*, *144*(2), 214–226. http://doi.org/10.1016/j.cell.2010.12.026

Barbieri, M., Chotalia, M., Fraser, J., Lavitas, L.-M., Dostie, J., Pombo, A., & Nicodemi, M. (2013). A model of the large-scale organization of chromatin. *Biochemical Society Transactions*, *41*(2), 508–512. http://doi.org/10.1042/BST20120238

Baù, D., Sanyal, A., Lajoie, B. R., Capriotti, E., Byron, M., Lawrence, J. B., … Marti-Renom, M. a. (2011). The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nature Structural & Molecular Biology*, *18*(1), 107–114. http://doi.org/10.1038/nsmb.1936

Beagrie, R. A., Scialdone, A., Schueler, M., Kraemer, D. C. A., Chotalia, M., Xie, S. Q., … Pombo, A. (2017). Complex multi-enhancer contacts captured by genome architecture mapping. *Nature*, *543*(7646), 519–524. http://doi.org/10.1038/nature21411

Benedetti, F., Dorier, J., Burnier, Y., & Stasiak, A. (2014). Models that include supercoiling of topological domains reproduce several known features of interphase chromosomes. *Nucleic Acids Research*, *42*(5), 2848–2855. http://doi.org/10.1093/nar/gkt1353

Bickmore, W. A., & Van Steensel, B. (2013). Genome architecture: Domain organization of interphase chromosomes. *Cell*. http://doi.org/10.1016/j.cell.2013.02.001

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., … De Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, *447*(7146), 799–816. http://doi.org/10.1038/nature05874

Boehr, D. D., Nussinov, R., & Wright, P. E. (2009). The role of dynamic conformational ensembles in biomolecular recognition. *Nature Chemical Biology*, *5*(11), 789–796. http://doi.org/10.1038/nchembio.232

Boettiger, A. N., Bintu, B., Moffitt, J. R., Wang, S., Beliveau, B. J., Fudenberg, G., … Zhuang, X. (2016). Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature*, *529*(7586), 418–422. http://doi.org/10.1038/nature16496

Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., … Cremer, T. (2005). Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biology*, *3*(5), 0826–0842. http://doi.org/10.1371/journal.pbio.0030157

Bonev, B., & Cavalli, G. (2016). Organization and function of the 3D genome. *Nature Reviews Genetics*, *17*(11), 661–678. http://doi.org/10.1038/nrg.2016.112

Branco, M. R., & Pombo, A. (2006). Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biology*, *4*(5), 780–788. http://doi.org/10.1371/journal.pbio.0040138

Cagliero, C., Grand, R. S., Jones, M. B., Jin, D. J., & O'Sullivan, J. M. (2013). Genome conformation capture reveals that the Escherichia coli chromosome is organized by replication and transcription. *Nucleic Acids Research*, *41*(12), 6058–6071. http://doi.org/10.1093/nar/gkt325

Canela, A., Maman, Y., Jung, S., Wong, N., Callen, E., Day, A., … Nussenzweig, A. (2017). Genome Organization Drives Chromosome Fragility. *Cell*, *170*(3), 507–521.e18. http://doi.org/10.1016/j.cell.2017.06.034

Chen, L., Balabanidou, V., Remeta, D. P., Minetti, C. A. S. A., Portaliou, A. G., Economou, A., & Kalodimos, C. G. (2011). Structural instability tuning as a regulatory mechanism in protein-protein interactions. *Molecular Cell*, *44*(5), 734–744. http://doi.org/10.1016/j.molcel.2011.09.022

Chothia, C., & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, *5*(4), 823–6. http://doi.org/060 fehlt

Consortium, E. P., Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. a, … Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. http://doi.org/10.1038/nature11247

Crane, E., Bian, Q., McCord, R. P., Lajoie, B. R., Wheeler, B. S., Ralston, E. J., … Meyer, B. J. (2015). Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*, *523*(7559), 240–244. http://doi.org/10.1038/nature14450

Croteau, N. J., Furgason, M. L. M., Devos, D., & Munson, M. (2009). Conservation of helical bundle structure between the exocyst subunits. *PLoS ONE*, *4*(2). http://doi.org/10.1371/journal.pone.0004443

Dacks, J. B., Poon, P. P., & Field, M. C. (2008). Phylogeny of endocytic components yields insight into the process of nonendosymbiotic organelle evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(2), 588–593. http://doi.org/10.1073/pnas.0707318105

Davies, J. O. J., Oudelaar, A. M., Higgs, D. R., & Hughes, J. R. (2017). How best to identify chromosomal interactions: a comparison of approaches. *Nature Methods*, *14*(2), 125–134. http://doi.org/10.1038/nmeth.4146

Davies, J. O. J., Telenius, J. M., McGowan, S. J., Roberts, N. A., Taylor, S., Higgs, D. R., & Hughes, J. R. (2015). Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nature Methods*, *13*(1), 74–80. http://doi.org/10.1038/nmeth.3664

de Wit, E., Bouwman, B. A. M., Zhu, Y., Klous, P., Splinter, E., Verstegen, M. J. A. M., … De Laat, W. (2013). The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature*, *501*(7466), 227–231. http://doi.org/10.1038/nature12420

de Wit, E., Vos, E. S. M., Holwerda, S. J. B., Valdes-Quezada, C., Verstegen, M. J. A. M., Teunissen, H., … de Laat, W. (2015). CTCF Binding Polarity Determines Chromatin Looping. *Molecular Cell*, *60*(4), 676–684. http://doi.org/10.1016/j.molcel.2015.09.023

Dekker, J., Marti-Renom, M. a, & Mirny, L. a. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews. Genetics*, *14*(6), 390–403. http://doi.org/10.1038/nrg3454

Dekker, J., Rippe, K., Dekker, M., Kleckner, N., Woodcock, C. L., Dimitrov, S., … Bustamante, C. (2002). Capturing chromosome conformation. *Science (New York, N.Y.)*, *295*(5558), 1306–11. http://doi.org/10.1126/science.1067799

Deng, X., Ma, W., Ramani, V., Hill, A., Yang, F., Ay, F., … Disteche, C. M. (2015). Bipartite structure of the inactive mouse X chromosome. *Genome Biology*, *16*(1), 1–21. http://doi.org/10.1186/s13059-015-0728-8

Di Stefano, M., Rosa, A., Belcastro, V., di Bernardo, D., & Micheletti, C. (2013). Colocalization of Coregulated Genes: A Steered Molecular Dynamics Study of Human Chromosome 19. *PLoS Computational Biology*, *9*(3), 1–13. http://doi.org/10.1371/journal.pcbi.1003019

Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., … Ren, B. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature*, *518*(7539), 331–336. http://doi.org/10.1038/nature14222

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., … Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376–380. http://doi.org/10.1038/nature11082

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., … Dekker, J. (2006). Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Research*, *16*(10), 1299–1309. http://doi.org/10.1101/gr.5571506

Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y. J., Lee, C., … Noble, W. S. (2010). A three-dimensional model of the yeast genome. *Nature*, *465*(7296), 363–367. http://doi.org/10.1038/nature08973

Eltsov, M., MacLellan, K. M., Maeshima, K., Frangakis, A. S., & Dubochet, J. (2008). Analysis of cryo-electron microscopy images does not support the existence of 30-nm chromatin fibers in mitotic chromosomes in situ. *Proceedings of the National Academy of Sciences*, *105*(50), 19732–19737. http://doi.org/10.1073/pnas.0810057105

Emanuel, M., Radja, N. H., Henriksson, A., & Schiessel, H. (2009). The physics behind the larger scale organization of DNA in eukaryotes. *Physical Biology*, *6*(2). http://doi.org/10.1088/1478-3975/6/2/025008

Erzberger, J. P., Stengel, F., Pellarin, R., Zhang, S., Schaefer, T., Aylett, C. H. S., … Ban, N. (2014). Molecular Architecture of the 40S□eIF1□eIF3 Translation Initiation Complex. *Cell*, *158*(5), 1123–1135. http://doi.org/10.1016/j.cell.2014.07.044

Feng, S., Cokus, S. J., Schubert, V., Zhai, J., Pellegrini, M., & Jacobsen, S. E. (2014). Genome-wide Hi-C Analyses in Wild-Type and Mutants Reveal High-Resolution Chromatin Interactions in Arabidopsis. *Molecular Cell*, *55*(5), 694–707. http://doi.org/10.1016/j.molcel.2014.07.008

Fernandez-Martinez, J., Kim, S. J., Shi, Y., Upla, P., Pellarin, R., Gagnon, M., … Rout, M. P. (2016). Structure and Function of the Nuclear Pore Complex Cytoplasmic mRNA Export Platform. *Cell*, *167*(5), 1215–1228.e25. http://doi.org/10.1016/j.cell.2016.10.028

Flors, C., & Earnshaw, W. C. (2011). Super-resolution fluorescence microscopy as a tool to study the nanoscale organization of chromosomes. *Current Opinion in Chemical Biology*, *15*(6), 839–844. http://doi.org/10.1016/j.cbpa.2011.10.004

Flyamer, I. M., Gassler, J., Imakaev, M., Ulyanov, S. V., Abdennur, N., Razin, S. V., … Tachibana-Konwalski, K. (2017). Single-cell Hi-C reveals unique chromatin reorganization at oocyte-tozygote transition. *Nature Publishing Group*, 1–17. http://doi.org/10.1038/nature21711

Franke, M., Ibrahim, D. M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, R., … Mundlos, S. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, *538*(7624), 265–269. http://doi.org/10.1038/nature19800

Franklin, R. E., and Gosling, R., 1953a, Molecular configuration in sodium thymonucleate. Nature, Lond. 171:740-741.

Fraser, J., Ferrai, C., Chiariello, A. M., Schueler, M., Rito, T., Laudanno, G., … Nicodemi, M. (2015). Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol Syst Biol*, *11*, 1–14. http://doi.org/10.15252/msb

Fridmann-Sirkis, Y., Kent, H. M., Lewis, M. J., Evans, P. R., & Pelham, H. R. B. (2006). Structural analysis of the interaction between the SNARE Tlg1 and Vps51. *Traffic*, *7*(2), 182–190. http://doi.org/10.1111/j.1600-0854.2005.00374.x

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., & Mirny, L. A. (2016). Formation of Chromosomal Domains by Loop Extrusion. *Cell Reports*, *15*(9), 2038–2049. http://doi.org/10.1016/j.celrep.2016.04.085

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. Bin, … Ruan, Y. (2009). An oestrogen-receptor-α-bound human chromatin interactome. *Nature*, *462*(7269), 58–64. http://doi.org/10.1038/nature08497

Fussner, E., Strauss, M., Djuric, U., Li, R., Ahmed, K., Hart, M., … Bazett-Jones, D. P. (2012). Open and closed domains in the mouse genome are configured as 10-nm chromatin fibres. *EMBO Reports*, *13*(11), 992–996. http://doi.org/10.1038/embor.2012.139

Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., … Superti-Furga, G. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature*, *440*(7084), 631–636. http://doi.org/10.1038/nature04532

Gibcus, J. H., Samejima, K., Goloborodko, A., Samejima, I., Naumova, N., Nuebler, J., … Dekker, J. (2018). A pathway for mitotic chromosome formation. *Science*, *6135*(January), eaao6135. http://doi.org/10.1126/science.aao6135

Giorgetti, L., Galupa, R., Nora, E. P., Piolot, T., Lam, F., Dekker, J., … Heard, E. (2014). Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell*, *157*(4), 950–963. http://doi.org/10.1016/j.cell.2014.03.025

Glaeser, R. M. (2016). Protein complexes in focus. *eLife*. http://doi.org/10.7554/eLife.13046

Gómez-Marín, C., Tena, J. J., Acemel, R. D., López-Mayorga, M., Naranjo, S., de la Calle-Mustienes, E., … Gómez-Skarmeta, J. L. (2015). Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. *Proceedings of the National Academy of Sciences*, *112*(24), 7542–7547. http://doi.org/10.1073/pnas.1505463112

Grob, S., Schmid, M. W., & Grossniklaus, U. (2014). Hi-C Analysis in Arabidopsis Identifies the KNOT, a Structure with Similarities to the flamenco Locus of Drosophila. *Molecular Cell*, *55*(5), 678–693. http://doi.org/10.1016/j.molcel.2014.07.009

Grosberg, A. Y., Khokhlov, A. R., Stanley, H. E., Mallinckrodt, A. J., & McKay, S. (1995). Statistical Physics of Macromolecules. *Computers in Physics*, *9*(2), 171. http://doi.org/10.1063/1.4823390

Grosberg, A. Y., Nechaev, S. K., & Shakhnovich, E. I. (1988). The role of topological constraints in the kinetics of collapse of macromolecules. *Journal de Physique*, *49*(12), 2095–2100. http://doi.org/10.1051/jphys:0198800490120209500

Gröschel, S., Sanders, M. A., Hoogenboezem, R., De Wit, E., Bouwman, B. A. M., Erpelinck, C., … Delwel, R. (2014). A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in Leukemia. *Cell*, *157*(2), 369–381. http://doi.org/10.1016/j.cell.2014.02.019

Hadjur, S., Williams, L. M., Ryan, N. K., Cobb, B. S., Sexton, T., Fraser, P., … Merkenschlager, M. (2009). Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. *Nature*, *460*(7253), 410–413. http://doi.org/10.1038/nature08079

Hahnfeldt, P., Hearst, J. E., Brenner, D. J., Sachs, R. K., & Hlatky, L. R. (1993). Polymer models for interphase chromosomes. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(16), 7854–7858. http://doi.org/10.1073/pnas.90.16.7854

Halverson, J. D., Lee, W. B., Grest, G. S., Grosberg, A. Y., & Kremer, K. (2011). Molecular dynamics simulation study of nonconcatenated ring polymers in a melt. I. Statics. *Journal of Chemical Physics*, *134*(20). http://doi.org/10.1063/1.3587137

Halverson, J. D., Smrek, J., Kremer, K., & Grosberg, A. Y. (2014). From a melt of rings to chromosome territories: The role of topological constraints in genome folding. *Reports on Progress in Physics*, *77*(2). http://doi.org/10.1088/0034-4885/77/2/022601

Handoko, L., Xu, H., Li, G., Ngan, C. Y., Chew, E., Schnapp, M., … Wei, C. L. (2011). CTCF-mediated functional chromatin interactome in pluripotent cells. *Nature Genetics*, *43*(7), 630–638. http://doi.org/10.1038/ng.857

Harmston, N., Ing-Simmons, E., Tan, G., Perry, M., Merkenschlager, M., & Lenhard, B. (2017). Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. *Nature Communications*, *8*(1). http://doi.org/10.1038/s41467-017-00524-5

Heider, M. R., Gu, M., Duffy, C. M., Mirza, A. M., Marcotte, L. L., Walls, A. C., … Munson, M. (2015). Subunit connectivity, assembly determinants and architecture of the yeast exocyst complex. *Nature Structural & Molecular Biology*, (December), 1–10. http://doi.org/10.1038/nsmb.3146

Hnisz, D., Weintraub, A. S., Day, D. S., Valton, A., Bak, R. O., Li, C. H., … Young, R. A. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, *351*(6280), 1454–1458. http://doi.org/10.1126/science.aad9024

Hsieh, T. H. S., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., & Rando, O. J. (2015). Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell*, *162*(1), 108–119. http://doi.org/10.1016/j.cell.2015.05.048

Hsu, S. C., Hazuka, C. D., Roth, R., Foletti, D. L., Heuser, J., & Scheller, R. H. (1998). Subunit composition, protein interactions, and structures of the mammalian brain sec6/8 complex and septin filaments. *Neuron*, *20*(6), 1111–1122. http://doi.org/10.1016/S0896-6273(00)80493-6

Hu, M., Deng, K., Qin, Z., Dixon, J., Selvaraj, S., Fang, J., … Liu, J. S. (2013). Bayesian Inference of Spatial Organizations of Chromosomes. *PLoS Computational Biology*, *9*(1), e1002893. http://doi.org/10.1371/journal.pcbi.1002893

Huang, B., Babcock, H., & Zhuang, X. (2010). Breaking the diffraction barrier: Super-resolution imaging of cells. *Cell*, *143*(7), 1047–1058. http://doi.org/10.1016/j.cell.2010.12.002

Hug, C. B., Grimaldi, A. G., Kruse, K., & Vaquerizas, J. M. (2017). Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription. *Cell*, *169*(2), 216–228.e19. http://doi.org/10.1016/j.cell.2017.03.024

Hughes, J. R., Roberts, N., McGowan, S., Hay, D., Giannoulatou, E., Lynch, M., … Higgs, D. R. (2014). Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nature Genetics*, *46*(2), 205–12. http://doi.org/10.1038/ng.2871

Imakaev, M. V, Fudenberg, G., & Mirny, L. A. (2015). Modeling chromosomes: Beyond pretty pictures. *FEBS Letters*. Federation of European Biochemical Societies. http://doi.org/10.1016/j.febslet.2015.09.004

Irimia, M., Maeso, I., Roy, S. W., & Fraser, H. B. (2013). Ancient cis-regulatory constraints and the evolution of genome architecture. *Trends in Genetics*, *29*(9), 521–528. http://doi.org/10.1016/j.tig.2013.05.008

Irimia, M., Tena, J. J., Alexis, M. S., Fernandez-Miñan, A., Maeso, I., Bogdanović, O., … Fraser, H. B. (2012). Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. *Genome Research*, *22*(12), 2356–2367. http://doi.org/10.1101/gr.139725.112

Jinek, M., Jiang, F., Taylor, D. W., Sternberg, S. H., Kaya, E., Ma, E., … Doudna, J. A. (2014). Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science*, *343*(6176). http://doi.org/10.1126/science.1247997

Jost, D., Carrivain, P., Cavalli, G., & Vaillant, C. (2014). Modeling epigenome folding: Formation and dynamics of topologically associated chromatin domains. *Nucleic Acids Research*, *42*(15), 9553–9561. http://doi.org/10.1093/nar/gku698

Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., van Berkum, N. L., … Young, R. A. (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, *467*(7314), 430–435. http://doi.org/10.1038/nature09380

Kakui, Y., Rabinowitz, A., Barry, D. J., & Uhlmann, F. (2017). Condensin-mediated remodeling of the mitotic chromatin landscape in fission yeast. *Nature Genetics*, *49*(10), 1553–1557. http://doi.org/10.1038/ng.3938

Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., & Chen, L. (2011). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature Biotechnology*, *30*(1), 90–98. http://doi.org/10.1038/nbt.2057

Kalodimos, C. G. (2012). Protein function and allostery: A dynamic relationship. *Annals of the New York Academy of Sciences*, *1260*(1), 81–86. http://doi.org/10.1111/j.1749-6632.2011.06319.x

Kendrew, JC, Bodo, G, Dintzis, HM, Parrish, RG,Wyckoff, H, and Phillips, DC (1958). “A three-dimensional model of the myoglobin molecule obtained by x-ray analysis.” Nature (London) 181(4610), 662–666.

Kim, Y., Eom, S. H., Wang, J., Lee, D.-S., & et al. (1995). Crystal structure of Thermus aquaticus DNA polymerase. *Nature*. http://doi.org/10.1038/376612a0

Kosinski, J., Mosalaganti, S., von Appen, A., Teimer, R., DiGuilio, A. L., Wan, W., … Beck, M. (2016). Molecular architecture of the inner ring scaffold of the human nuclear pore complex. *Science*, *352*(6283), 363–365. http://doi.org/10.1126/science.aaf0643

Lasker, K., Forster, F., Bohn, S., Walzthoeni, T., Villa, E., Unverdorben, P., … Baumeister, W. (2012). Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach.

Proceedings of the National Academy of Sciences, *109*(5), 1380–1387. http://doi.org/10.1073/pnas.1120559109

Lazar-Stefanita, L., Scolari, V. F., Mercy, G., Muller, H., Guérin, T. M., Thierry, A., … Koszul, R. (2017). Cohesins and condensins orchestrate the 4D dynamics of yeast chromosomes during the cell cycle. *The EMBO Journal*, *36*(18), e201797342. http://doi.org/10.15252/embj.201797342

Le, T. B. K., Imakaev, M. V., Mirny, L. A., & Laub, M. T. (2013). High-Resolution Mapping of the Spatial Organization of a Bacterial Chromosome. *Science*, *342*(6159), 731–734. http://doi.org/10.1126/science.1242059

Lesne, A., Riposo, J., Roger, P., Cournac, A., & Mozziconacci, J. (2014). 3D genome reconstruction from chromosomal contacts. *Nature Methods*, *11*(11), 1141–1143. http://doi.org/10.1038/nmeth.3104

Lieberman-aiden, E., Berkum, N. L. Van, Williams, L., Imakaev, M., Ragoczy, T., Telling, A., … Mirny, L. A. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, *326*(9 October 2009), 289–293. http://doi.org/10.1038/nature08398

Lonfat, N., & Duboule, D. (2015). Structure, function and evolution of topologically associating domains (TADs) at HOX loci. *FEBS Letters*, *589*(20), 2869–2876. http://doi.org/10.1016/j.febslet.2015.04.024

Löwe, J., & Amos, L. A. (1998). Crystal structure of the bacterial cell division protein FtsZ. *Nature*, *391*(1996), 203–206.

Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., … Mundlos, S. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, *161*(5), 1012–1025. http://doi.org/10.1016/j.cell.2015.04.004

Marbouty, M., Cournac, A., Flot, J. F., Marie-Nelly, H., Mozziconacci, J., & Koszul, R. (2014). Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. *eLife*, *3*, e03318. http://doi.org/10.7554/eLife.03318

Marbouty, M., Le Gall, A., Cattoni, D. I., Cournac, A., Koh, A., Fiche, J. B., … Nollmann, M. (2015). Condensin- and Replication-Mediated Bacterial Chromosome Folding and Origin Condensation Revealed by Hi-C and Super-resolution Imaging. *Molecular Cell*, *59*(4), 588–602. http://doi.org/10.1016/j.molcel.2015.07.020

Marsh, J. A., & Teichmann, S. A. (2015). Structure, Dynamics, Assembly, and Evolution of Protein Complexes. *Annual Review of Biochemistry*, *84*(1), 551–575. http://doi.org/10.1146/annurev-biochem-060614-034142

Marsh, J. A., Teichmann, S. A., & Forman-Kay, J. D. (2012). Probing the diverse landscape of protein flexibility and binding. *Current Opinion in Structural Biology*, *22*(5), 643–650. http://doi.org/10.1016/j.sbi.2012.08.008

Marti-Renom, M. a., & Mirny, L. a. (2011). Bridging the resolution gap in structural modeling of 3D genome organization. *PLoS Computational Biology*, *7*(7), 1–6. http://doi.org/10.1371/journal.pcbi.1002125

McDowall, a W., Smith, J. M., & Dubochet, J. (1986). Cryo-electron microscopy of vitrified chromosomes in situ. *The EMBO Journal*, *5*(6), 1395–1402. http://doi.org/10.1038/emboj.2009.340

Mei, K., Li, Y., Wang, S., Shao, G., Wang, J., Ding, Y., … Guo, W. (2018). Cryo-EM structure of the exocyst complex. *Nature Structural & Molecular Biology*. http://doi.org/10.1038/s41594-017-0016-2

Mirny, L. A. (2011). The fractal globule as a model of chromatin architecture in the cell. *Chromosome Research*, *19*(1), 37–51. http://doi.org/10.1007/s10577-010-9177-0

Mittag, T., Kay, L. E., & Forman-Kaya, J. D. (2010). Protein dynamics and conformational disorder in molecular recognition. *Journal of Molecular Recognition*, *23*(2), 105–116. http://doi.org/10.1002/jmr.961

Mizuguchi, T., Fudenberg, G., Mehta, S., Belton, J.-M., Taneja, N., Folco, H. D., … Grewal, S. I. S. (2014). Cohesin-dependent globules and heterochromatin shape 3D genome architecture in S. pombe. *Nature*, *516*(7531), 432–5. http://doi.org/10.1038/nature13833

Münkel, C., & Langowski, J. (1998). Chromosome structure described by a polymer model. *Physical Review E*, *57*(5–B), 5888–5896. Retrieved from http://macromol.dkfz-heidelberg.de/pdf-files/CHR/CHR-1998-6386.pdf

Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., … Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, *502*(7469), 59–64. http://doi.org/10.1038/nature12593

Nagano, T., Lubling, Y., Várnai, C., Dudley, C., Leung, W., Baran, Y., … Tanay, A. (2017). Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*, *547*(7661), 61–67. http://doi.org/10.1038/nature23001

Nasmyth, K. (2001). Disseminating the Genome: Joining, Resolving, and Separating Sister Chromatids During Mitosis and Meiosis. *Annual Review of Genetics*, *35*(1), 673–745. http://doi.org/10.1146/annurev.genet.35.102401.091334

Naumova, N., Imakaev, M., Fudenberg, G., Zhan, Y., Lajoie, B. R., Mirny, L. A., & Dekker, J. (2013). Organization of the Mitotic Chromosome. *Science*, *342*(6161), 948–953. http://doi.org/10.1126/science.1236083

Navia, M. A., Fitzgerald, P. M., McKeever, B. M., Leu, C. T., Heimbach, J. C., Herber, W. K., … Springer, J. P. (1989). Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1. *Nature*. http://doi.org/10.1038/337615a0

Nichols, M. H., & Corces, V. G. (2015). A CTCF Code for 3D Genome Architecture. *Cell*, *162*(4), 703–705. http://doi.org/10.1016/j.cell.2015.07.053

Nishino, Y., Eltsov, M., Joti, Y., Ito, K., Takata, H., Takahashi, Y., … Maeshima, K. (2012). Human mitotic chromosomes consist predominantly of irregularly folded nucleosome fibres without a 30-nm chromatin structure. *EMBO Journal*, *31*(7), 1644–1653. http://doi.org/10.1038/emboj.2012.35

Nogales, E., Wolf, S. G., & Downing, K. H. (1998). Electron Crystallography. *Nature*, *391*(January), 199–204. http://doi.org/10.1007/1-4020-3920-4

Nooren, I. M. A., & Thornton, J. M. (2003). Diversity of protein-protein interactions. *EMBO Journal*, *22*(14), 3486–3492. http://doi.org/10.1093/emboj/cdg359

Nora, E. P., Goloborodko, A., Valton, A.-L., Gibcus, J. H., Uebersohn, A., Abdennur, N., … Bruneau, B. G. (2017). Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell*, *169*(5), 930–944.e22. http://doi.org/10.1016/j.cell.2017.05.004

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., … Heard, E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, *485*, 381–385. http://doi.org/10.1038/nature11049

Ou, H. D., Phan, S., Deerinck, T. J., Thor, A., Ellisman, M. H., & O'Shea, C. C. (2017). ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science*, *357*(6349), eaag0025. http://doi.org/10.1126/science.aag0025

Palstra, R. J., Tolhuis, B., Splinter, E., Nijmeijer, R., Grosveld, F., & De Laat, W. (2003). The β-globin nuclear compartment in development and erythroid differentiation. *Nature Genetics*, *35*(2), 190–194. http://doi.org/10.1038/ng1244

Pérez-Victoria, F. J., Abascal-Palacios, G., Tascón, I., Kajava, A., Magadán, J. G., Pioro, E. P., … Hierro, A. (2010). Structural basis for the wobbler mouse neurodegenerative disorder caused by mutation in the Vps54 subunit of the GARP complex. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(29), 12860–5. http://doi.org/10.1073/pnas.1004756107

Phillips-Cremins, J. E., Sauria, M. E. G., Sanyal, A., Gerasimova, T. I., Lajoie, B. R., Bell, J. S. K., … Corces, V. G. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, *153*(6), 1281–1295. http://doi.org/10.1016/j.cell.2013.04.053

Ramani, V., Deng, X., Qiu, R., Gunderson, K. L., Steemers, F. J., Disteche, C. M., … Shendure, J. (2017). Massively multiplex single-cell Hi-C. *Nature Methods*, *14*(3), 263–266. http://doi.org/10.1038/nmeth.4155

Rao, S. S. P., Huang, S. C., Glenn St Hilaire, B., Engreitz, J. M., Perez, E. M., Kieffer-Kwon, K. R., … Aiden, E. L. (2017). Cohesin Loss Eliminates All Loop Domains. *Cell*, *171*(2), 305–320.e24. http://doi.org/10.1016/j.cell.2017.09.026

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., … Aiden, E. L. L. (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, *159*(7), 1665–1680. http://doi.org/10.1016/j.cell.2014.11.021

Rapkin, L. M., Anchel, D. R. P., Li, R., & Bazett-Jones, D. P. (2012). A view of the chromatin landscape. *Micron*, *43*(2–3), 150–158. http://doi.org/10.1016/j.micron.2011.11.007

Robinson, P. J., Trnka, M. J., Pellarin, R., Greenberg, C. H., Bushnell, D. A., Davis, R., … Unit, S. B. (2015). Molecular architecture of the yeast Mediator complex, 1–29. http://doi.org/10.7554/eLife.08719

Robinson, C. V., Sali, A., & Baumeister, W. (2007). The molecular sociology of the cell. *Nature*, *450*(7172), 973–982. http://doi.org/10.1038/nature06523

Ormo, M, Cubitt, A. B., Kallio K., Gross L. A., Tsien R. Y., Remingtont S. J. (1996). Green fluorescent protein. *Science*, *273*(September), 246–252.

Rosa, A., Becker, N. B., & Everaers, R. (2010). Looping probabilities in model interphase chromosomes. *Biophysical Journal*, *98*(11), 2410–2419. http://doi.org/10.1016/j.bpj.2010.01.054

Rosa, A., & Everaers, R. (2008). Structure and dynamics of interphase chromosomes. *PLoS Computational Biology*, *4*(8). http://doi.org/10.1371/journal.pcbi.1000153

Rosa, S., & Shaw, P. (2013). Insights into Chromatin Structure and Dynamics in Plants. *Biology*, *2*(4), 1378–1410. http://doi.org/10.3390/biology2041378

Rowley, M. J., & Corces, V. G. (2016). The three-dimensional genome: Principles and roles of long-distance interactions. *Current Opinion in Cell Biology*, *40*, 8–14. http://doi.org/10.1016/j.ceb.2016.01.009

Rowley, M. J., Nichols, M. H., Lyu, X., Ando-Kuri, M., Rivera, I. S. M., Hermetz, K., … Corces, V. G. (2017). Evolutionarily Conserved Principles Predict 3D Chromatin Organization. *Molecular Cell*, *67*(5), 837–852.e7. http://doi.org/10.1016/j.molcel.2017.07.022

Russel, D., Lasker, K., Webb, B., Velázquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., … Sali, A. (2012). Putting the pieces together: Integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biology*, *10*(1), e1001244. http://doi.org/10.1371/journal.pbio.1001244

Sali, A., Glaeser, R., Earnest, T., & Baumeister, W. (2003). From words to literature in structural proteomics. *Nature*, *422*(6928), 216–225. http://doi.org/10.1038/nature01513

Sanborn, A. L., Rao, S. S. P., Huang, S.-C., Durand, N. C., Huntley, M. H., Jewett, A. I., … Aiden, E. L. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences* . http://doi.org/10.1073/pnas.1518552112

Sander, C., & Schneider, R. (1991). Database of homology???derived protein structures and the structural meaning of sequence alignment. *Proteins: Structure, Function, and Bioinformatics*, *9*(1), 56–68. http://doi.org/10.1002/prot.340090107

Schalbetter, S. A., Goloborodko, A., Fudenberg, G., Belton, J. M., Miles, C., Yu, M., … Baxter, J. (2017). SMC complexes differentially compact mitotic chromosomes according to genomic context. *Nature Cell Biology*, *19*(9), 1071–1080. http://doi.org/10.1038/ncb3594

Schneidman-duhovny, D., Pellarin, R., & Sali, A. (2014). Uncertainty in integrative structural modeling. *Current Opinion in Structural Biology*, *28*, 96–104. http://doi.org/10.1016/j.sbi.2014.08.001

Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., … Spitz, F. (2017). Two independent modes of chromatin organization revealed by cohesin removal. *Nature*, *551*(7678), 51–56. http://doi.org/10.1038/nature24281

Segal, M. R., Xiong, H., Capurso, D., Vazquez, M., & Arsuaga, J. (2014). Reproducibility of 3D chromatin configuration reconstructions. *Biostatistics*, *15*(3), 442–456. http://doi.org/10.1093/biostatistics/kxu003

Serra, F., Di Stefano, M., Spill, Y. G., Cuartero, Y., Goodstadt, M., Baù, D., & Marti-Renom, M. A. (2015). Restraint-based three-dimensional modeling of genomes and genomic domains. *FEBS Letters*, *589*(20), 2987–2995. http://doi.org/10.1016/j.febslet.2015.05.012

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., … Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*. http://doi.org/10.1016/j.cell.2012.01.010

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., … de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature Genetics*, *38*(11), 1348–1354. http://doi.org/10.1038/ng1896

Splinter, E., Heath, H., Kooren, J., Palstra, R. J., Klous, P., Grosveld, F., … De Laat, W. (2006). CTCF mediates long-range chromatin looping and local histone modification in the ??-globin locus. *Genes and Development*, *20*(17), 2349–2354. http://doi.org/10.1101/gad.399506

Stengel, F., Aebersold, R., & Robinson, C. V. (2012). Joining Forces: Integrating Proteomics and Cross-linking with the Mass Spectrometry of Intact Complexes. *Molecular & Cellular Proteomics*, *11*(3), R111.014027. http://doi.org/10.1074/mcp.R111.014027

Stevens, T. J., Lando, D., Basu, S., Liam, P., Cao, Y., Lee, S. F., … Hendrich, B. (2017). 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, 1–21. http://doi.org/10.1038/nature21429

Symmons, O., Pan, L., Remeseiro, S., Aktas, T., Klein, F., Huber, W., & Spitz, F. (2016). The Shh Topological Domain Facilitates the Action of Remote Enhancers by Reducing the Effects of Genomic Distances. *Developmental Cell*, *0*(0), 47–57. http://doi.org/10.1016/j.devcel.2016.10.015

Tanizawa, H., Iwasaki, O., Tanaka, A., Capizzi, J. R., Wickramasinghe, P., Lee, M., … Noma, K. I. (2010). Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Research*, *38*(22), 8164–8177. http://doi.org/10.1093/nar/gkq955

Tiana, G., Amitai, A., Pollex, T., Piolot, T., Holcman, D., Heard, E., & Giorgetti, L. (2016). Structural Fluctuations of the Chromatin Fiber within Topologically Associating Domains. *Biophysical Journal*, *110*(6), 1234–1245. http://doi.org/10.1016/j.bpj.2016.02.003

Tjong, H., Gong, K., Chen, L., & Alber, F. (2012). Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. *Genome Research*, *22*(7), 1295–1305. http://doi.org/10.1101/gr.129437.111

Tolhuis, B., Palstra, R. J., Splinter, E., Grosveld, F., & De Laat, W. (2002). Looping and interaction between hypersensitive sites in the active β-globin locus. *Molecular Cell*, *10*(6), 1453–1465. http://doi.org/10.1016/S1097-2765(02)00781-5

Tompa, P., & Fuxreiter, M. (2008). Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends in Biochemical Sciences*, *33*(1), 2–8. http://doi.org/10.1016/j.tibs.2007.10.003

Tzeng, S. R., & Kalodimos, C. G. (2011). Protein dynamics and allostery: An NMR view. *Current Opinion in Structural Biology*, *21*(1), 62–67. http://doi.org/10.1016/j.sbi.2010.10.007

Ulianov, S. V, Khrameeva, E. E., Gavrilov, A. A., Flyamer, I. M., Kos, P., Mikhaleva, E. A., … Razin, S. V. (2016). Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains, 70–84. http://doi.org/10.1101/gr.196006.115.12

Umbarger, M. a., Toro, E., Wright, M. a., Porreca, G. J., Baù, D., Hong, S. H., … Church, G. M. (2011). The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation. *Molecular Cell*, *44*(2), 252–264. http://doi.org/10.1016/j.molcel.2011.09.010

van den Engh, G., Sachs, R., & Trask, B. J. (1992). Estimating genomic distance from DNA sequence location in cell nuclei by a random walk model. *Science (New York, N.Y.)*, *257*(5075), 1410–1412. http://doi.org/10.1126/science.1388286

Varoquaux, N., Ay, F., Noble, W. S., & Vert, J. P. (2014). A statistical approach for inferring the 3D structure of the genome. *Bioinformatics*, *30*(12), i26–i33. http://doi.org/10.1093/bioinformatics/btu268

Vasan, N., Hutagalung, A., Novick, P., & Reinisch, K. M. (2010). Structure of a C-terminal fragment of its Vps53 subunit suggests similarity of Golgi-associated retrograde protein (GARP) complex to a family of tethering complexes. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(32), 14176–81. http://doi.org/10.1073/pnas.1009419107

Vettorel, T., Grosberg, A. Y., & Kremer, K. (2009). Statistics of polymer rings in the melt: A numerical simulation study. *Physical Biology*, *6*(2). http://doi.org/10.1088/1478-3975/6/2/025013

Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D. T., Tanay, A., & Hadjur, S. (2015). Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. *Cell Reports*, *10*(8), 1297–1309. http://doi.org/10.1016/j.celrep.2015.02.004

Wang, C., Liu, C., Roqueiro, D., Grimm, D., Schwab, R., Becker, C., … Weigel, D. (2015). Genome-wide analysis of local chromatin packing in Arabidopsis thaliana. *Genome Research*, *25*(2), 246–256. http://doi.org/10.1101/gr.170332.113

Wang, S., Wang, S., Su, J., Beliveau, B. J., Bintu, B., & Moffitt, J. R. (2016). Spatial organization of chromatin domains and compartments in single chromosomes, *8084*.

Wang, S., Xu, J., & Zeng, J. (2015). Inferential modeling of 3D chromatin structure. *Nucleic Acids Research*, *43*(8), e54. http://doi.org/10.1093/nar/gkv100

Wang, X., Le, T. B. K., Lajoie, B. R., Dekker, J., Laub, M. T., & Rudner, D. Z. (2015). Condensin promotes the juxtaposition of dna flanking its loading site in Bacillus subtilis. *Genes and Development*, *29*(15), 1661–1675. http://doi.org/10.1101/gad.265876.115

Watson, J.D., and Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171, 737–738.

Whitty, A. (2008). Cooperativity and biological complexity. *Nature Chemical Biology*, *4*(8), 435–439. http://doi.org/10.1038/nchembio0808-435

Wilkins, M. H. F., Stokes, A. R., and Wilson, H. R., 1953, Molecular structure of desoxypentose nucleic acids. Nature, Lond. I71:738-740.

Wijchers, P. J., Krijger, P. H. L., Geeven, G., Zhu, Y., Denker, A., Verstegen, M. J. A. M., … de Laat, W. (2016). Cause and Consequence of Tethering a SubTAD to Different Nuclear Compartments. *Molecular Cell*, *61*(3), 461–473. http://doi.org/10.1016/j.molcel.2016.01.001

Wolfsberg, T. G., McEntyre, J., & Schuler, G. D. (2001). Guide to the draft human genome. *Nature*, *409*(6822), 824–826. http://doi.org/10.1038/35057000

Wood, T. C., & Pearson, W. R. (1999). Evolution of protein sequences and structures. *Journal of Molecular Biology*, *291*(4), 977–995. http://doi.org/10.1006/jmbi.1999.2972

Zhan, Y., Mariani, L., Barozzi, I., Schulz, E. G., Bluthgen, N., Stadler, M., … Giorgetti, L. (2017). Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. *Genome Research*, gr.212803.116. http://doi.org/10.1101/gr.212803.116

Zhang, B., & Wolynes, P. G. (2015). Topology, structures, and energy landscapes of human chromosomes. *Proceedings of the National Academy of Sciences*, *112*(19), 6062–6067. http://doi.org/10.1073/pnas.1506257112

Zhang, Y., McCord, R. P., Ho, Y. J., Lajoie, B. R., Hildebrand, D. G., Simon, A. C., … Dekker, J. (2012). Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell*, *148*(5), 908–921. http://doi.org/10.1016/j.cell.2012.02.002

Zhang, Z., Li, G., Toh, K. C., & Sung, W. K. (2013). Inference of spatial organizations of chromosomes using semi-definite embedding approach and Hi-C data. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7821 LNBI*, 317–332. http://doi.org/10.1007/978-3-642-37195-0_31

Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., … Ohlsson, R. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature Genetics*, *38*(11), 1341–1347. http://doi.org/10.1038/ng1891