Facultad de Ciencias

Departamento de Biología Molecular

Identification and Functional Validation of Genomic Boundaries in Mammals

Cristina Vicente García

Madrid, 2014

Facultad de Ciencias

Departamento de Biología Molecular

Identification and Functional Validation of Genomic Boundaries in Mammals

A thesis submitted in fulfillment of the requirements to obtain the PhD degree in Biochemistry, Molecular Biology, Biomedicine and Biotechnology at Universidad Autónoma de Madrid by

Cristina Vicente García

Degree in Biotechnology by Universidad Pablo de Olavide

PhD Director: Dr. **Lluís Montoliu José**

Laboratory "Animal Models by Genetic Manipulation"

Department of Molecular and Cellular Biology

National Centre for Biotechnology (CNB)

State Agency Spanish National Research Council (CSIC)

Campus of Excellence UAM + CSIC

Universidad Autónoma de Madrid

Madrid, 2014

This PhD thesis has been carried out by Cristina Vicente García at the "Animal Models by Genetic Manipulation" laboratory from the Department of Molecular and Cellular Biology at the National Centre for Biotechnology (CNB-CSIC), in Madrid, under the supervision of Dr. Lluís Montoliu José, CSIC Research Scientist, CIBER-ER (Network Center of Research in Biomedicine on Rare Diseases), ISCIII (Instituto de Salud Carlos III) Researcher, and Honorary Professor at the Autonomous University of Madrid (UAM).

The support received from the following Grants and Fellowships has permitted to develop this PhD work with success to its end:

- Functional and structural analysis of gene expression domains in mammalian genomes, Spanish Ministry of Science and Technology, National Plan R+D+I, Basic Biology Program (Molecular Biology), 2006-2009, reference BFU2006-12185. Principal Investigator: Dr. Lluís Montoliu José.
- JAE Introduction to Research Undergraduate Fellowship, CSIC, 2008. Recipient: Cristina Vicente García.
- JAE Predoc PhD Fellowship, CSIC, 2009-2013. Recipient: Cristina Vicente García.
- Identification and functional validation of genomic vertebrate insulators, *in vitro* and *in vivo*, Spanish Ministry of Science and Innovation, National Plan R+D+I, Biotechnology Program, 2009-2012, reference BIO2009-12697. Principal Investigator: Dr. Lluís Montoliu José.
- EMMAservice: Servicing the European Biomedical Research Community: archiving and dissemination of Mouse Models of Human Disease, European Commission, FP7-infrastructures-2008-1, 2009-2012, reference ID 227490. Principal Investigator: Dr. Lluís Montoliu José (ten partners, coordinator: Prof. Dr. Glauco Tocchini-Valentini, CNR-IBC, Monterotondo, Rome, Italy).
- Use of insulator sequences in adenoviruses prepared for gene therapy experimental protocols, CIBER-ER, ISCIII, Spanish Ministry of Health, 2009. Principal Investigator: Dr. Lluís Montoliu José (two partners, coordinator: Dr. Cristina Fillat, CRG-Barcelona).
- Cloning a locus associated with a cone-photoreceptor deficiency in the central retina of mice. A possible new animal model for ARMD, CIBER-ER, ISCIII, Spanish Ministry of Health, 2009, reference INTRA/09/756,1. Principal Investigator: Dr. Lluís Montoliu José

- Short stay grant as a visiting researcher at Prof. Victoria Lunyak's laboratory at the Buck Institute for Research on Aging (Novato, California, USA), JAE Predoc PhD Fellowship, CSIC, 2011. Recipient: Cristina Vicente García.
- Formative short stays at Dr. José Luís Gómez Skarmeta's laboratory at the Andalusian Center for Developmental Biology (CABD), 2011-2013. Funding through BIO2009-12697 and BIO2012-39980. Recipient: Cristina Vicente García.
- Genetic diagnosis and potential therapies for albinism, CIBER-ER, ISCIII, Spanish Ministry of Economy and Competitiveness (MINECO), 2012-2014, reference 13-756/152.09. Principal Investigator: Dr. Lluís Montoliu José, coordinator (one additional group).
- Animal models for investigating diseases of the visual system, R+D Programs in Biomedicine, Autonomous Government of Madrid, 2012-2015, reference S2011/BMD-2439. Principal Investigator: Dr. Lluís Montoliu José, coordinator (four other groups).
- Functional and structural validation of genomic boundaries, MINECO, National Plan R+D+I, Biotechnology Program, 2013-2015, reference BIO2012-39980. Principal Investigator: Dr. Lluís Montoliu José.
- INFRAFRONTIER-I3: Development of mouse mutant resources for functional analysis of human diseases – Enhancing the translation of research into innovation, European Commission, FP7 Capacities, 2013-2016, INFRA-2012-1.1.4, reference 312325. Principal Investigator: Dr. Lluís Montoliu José (23 partners, coordinator: Prof. Dr. Martin Hrabé de Angelis, IEG-HMGU, Munich, Germany).
- International Society for Transgenic Technologies (ISTT), Registration Award to attend the 11th Transgenic Technolology (TT2013) Meeting held in Guangzhou, China, in February 2013. Recipient: Cristina Vicente García.
- Best Oral Presentation Award at the annual International Mammalian Genome Conference of the International Mammalian Genome Society (IMGS), held in Salamanca in September 2013. Recipient: Cristina Vicente García.

To my parents and my brother.

To Miguel.

Acknowledgments

First of all, I am deeply grateful to Lluís Montoliu for trusting me to work on this project. You believed in me and let me find my own way. I have greatly benefited from your advice and knowledge; whereas your patience and words of encouragement helped me in my (many) moments of crisis. Thank you for the opportunities you have created for me along the way.

It is with immense gratitude that I acknowledge the support and help of Almudena, the most patient person in the world. I know it has been hard not to strangle me sometimes! Thank you for caring, for listening and for answering my endless questions. Thank you for your friendship.

Davide, it took a while, but I am glad we learnt to understand each other. Your comments and suggestions were invaluable. I would also like to thank Diego. I was never short of laughter when you were around! Likewise, I would wish to extend my thanks to Marta Cantero, who was always willing to help and listen (excuse me for torturing you with my awful jokes!).

Edu, apologies for being a pain during my first years! I sincerely appreciate your efforts in teaching me, not only experimental techniques, but also a practical way to face life.

I thought I would not miss Óscar teasing me about everything, especially football. But I do… Julia, thank you for that unforgettable ISTT meeting at Guangzhou. There, I confirmed that you possess only goodness in your heart, a gift that seems to abound in the Cryo team: one conversation with Chus is enough to realize that she also has it.

I would not want to miss the opportunity to acknowledge all those other members of the 111 family that contributed with a kind word or insightful comments to the achievement of this PhD thesis: María Tiana ("*Cojo un muelle…*", I just cannot get it out of my head!), Amalia Martínez, Esther Zurita, Soledad Montalbán, Mónica Martínez, Marta Castrillo, Isabel Martín-Dorado, María Barandalla and Almudena Tello.

Acknowledgments

My deepest gratitude goes to José Luis Gómez Skarmeta for offering me the possibility to work with zebrafish. At CABD, I received generous and priceless help from Ana Fernández Miñán… I owe you one!

I would also want to thank Victoria Lunyak for opening the doors of her laboratory to me. You introduced me to the exciting world of stem cells, and taught me to stand strong in adverse situations.

Natalia Jiménez, Joan Segura, you made my life so much easier with aGEM… My infinite gratitude to both of you.

A special thanks is expressed to Laura Barrios for guiding me through the labyrinth of numbers and statistics during my first years of doctoral study. I would have never found my way out without you!

I am indebted to Corina Lorz and Carmen Segrelles for providing me with the COCA cell line, so necessary in the final stretch of this thesis. Thank you.

Jaime Carvajal, thank you for your understanding, support and faith in me, as well as for accepting me in the 211 family: little boss, Robli, Bella, Fita, Raúl and Rosa, I really appreciate your kindness, intelligence and good sense of humor!

Berna, my big sister, you are an endless source of beauty, wisdom and goodness. You made me feel at home from the very first minute we met. You never let me crumble, and I cannot thank you enough for that.

Mum, Dad, thank you. Thank you for raising me under the values of love, honesty, humility and hard work. Thank you for teaching me that, if you try hard enough, the sky is the limit. Most importantly, thank you for not kicking me out of the house when the stress made me unbearable… Manu, thank you for always being there for me.

Words cannot express how grateful I am to Miguel. Thanks for cheering me up in my lowest moments, for making me laugh, for having my back and for making me feel I can fly. Thank you for letting me be in your life.

Summary

Eukaryotic genomes are divided into expression domains, which contain DNA coding sequences together with all the regulatory elements needed for their correct spatio-temporal expression pattern. Genomic boundaries, also known as insulators, flank these domains preventing undesirable crosstalk between the regulatory elements of neighboring domains. They employ various mechanisms and thus, are functionally rather than structurally defined. For this reason, in an attempt to find boundaries in a genome-wide unbiased fashion in mammals, we focused on identifying those loci where the presence of boundary function would be required to satisfy a biological need. For example, we hypothesized that adjacent genes with opposite expression patterns would need to be separated by boundaries to maintain the independency of their different expression domains. Also, boundaries could be found partitioning the chromatin into inactive heterochromatic and active euchromatic domains, impeding the deleterious effects the spread of the former would have on the latter. Finally, boundaries could also bracket clusters of co-expressed genes to ensure their co-regulation and co-expression. Different algorithms, based on the analysis of gene expression data, were developed in order to explore these scenarios. The resulting evolutionarily conserved non-coding putative insulator sequences were functionally validated using a number of assays. Their enhancer-blocking properties were evaluated *in vitro* in human cells in culture, and then *in vivo* by using transgenic zebrafish. Additionally, one of the most powerful elements was further tested for its ability to protect from chromosomal position effects in transgenic mice. The description and characterization of new genomic boundaries would shed some light into the way mammalian genomes are organized, as well as expand the repertoire of genetic tools that can be incorporated in heterologous constructs to improve the gene transfer technologies by preventing chromosomal position effects.

Resumen

Los genomas de eucariotas están divididos en dominios de expresión, que se definen como aquellas porciones del genoma que contienen uno o varios genes y todos los elementos reguladores necesarios para que que se expresen de acuerdo con un patrón espacio-temporal concreto. Los aisladores genómicos, también llamados *insulators*, flanquean estos dominios y los protegen de la influencia no deseada de los elementos reguladores contenidos en los dominios vecinos. Existen diversos mechanismos de aislamiento, por lo que los *insulators* no se definen por una secuencia de ADN concreta, sino porque comparten una misma función. Así, para encontrar aisladores en el genoma de mamíferos de una forma no sesgada, nos propusimos identificar aquellas posiciones del genoma donde se requiere la presencia de función aisladora para satisfacer un problema biológico. Por ejemplo, genes adyacentes con perfiles de expresión completamente distintos deberían estar separados por aisladores que mantuviesen dominios de expresión independientes. Asimismo, cabe esperar la presencia de aisladores entre dominios silentes de heterocromatina y dominios activos de eucromatina. Aquí, impedirían los efectos perjudiciales que el avance de los primeros tendrían sobre los segundos. Finalmente, también podrían encontrarse aisladores flanqueando grupos de genes co-expresados para asegurar su co-regulación y, por tanto, co-expresión. Basándonos en estos escenarios, se desarrollaron diversos algoritmos que usaban datos de expresión génica para predecir la presencia de aisladores. Como resultado de estos algoritmos, se obtuvo una serie de secuencias conservadas evolutivamente y no codificantes que se validaron funcionalmente empleando varios tests. La capacidad de bloqueo de *enhancers* se evaluó mediante ensayos *in vitro* en células humanas en cultivo primero, y luego *in vivo* mediante el uso de peces cebra transgénicos. Además, se analizó la capacidad de uno de los elementos más potentes para proteger de efectos de posición cromosomales en ratones transgénicos. La descripción y caracterización de nuevos aisladores genómicos no sólo sirve para entender mejor cómo se organizan los genomas de mamíferos. También es útil para ampliar el abanico de herramientas disponibles que se pueden usar en construcciones heterólogas para bloquear los efectos de posición cromosomales que se dan comúnmente en experimentos de transferencia genética.

Table of Contents

List of Tables

List of Figures

Abbreviations

3C	Chromatin Conformation Capture
a.m.	*ante meridium*
A.U.	Arbitrary Units
ABA	Allen Brain Atlas
aGEM	anatomic Gene Expression Mapping
ApE	A plasmid Editor
BAC	Bacterial Artificial Chromosome
BioGPS	Gene Portal System
bp	base pair
CABD	*Centro Andaluz de Biología del Desarrollo*
CAR	Cardiac Actin Promoter
CBMSO	*Centro de Biología Molecular Severo Ochoa*
cDNA	complementary DNA
ChIA-PET	Chromatin Interaction Analysis by Paired-End Tag sequencing
ChIP	Chromatin ImmunoPrecipitation
CIEMAT	*Centro de Investigaciones Enérgicas, Medioambientales y Tecnológicas*
CISA	*Centro de Investigación en Sanidad Animal*
CMV	CytoMegaloVirus
CNB	*Centro Nacional de Biotecnología*
CNS	Central Nervous System
COC	Chromosome Organizing Clump
CONSYN	CONstitutive and SYNtenic
CRISPR	Clustered Regulatory Interspaced Short Palindromic Repeat
CRM	*Cis*-Regulatory Module
CSIC	*Consejo Superior de Investigaciones Científicas*
CTS	CTCF Target Site

dATP	deoxyAdenosine TriPhosphate
DAVID	Database for Annotation, Visualization, and Integrated Discovery
dCTP	deoxyCytidine TriPhosphate
dCTP [α-^{32}P]	Phosphorus-32 radiolabeled deoxyCytidine TriPhosphate
DEPC	DiEthyl PyroCarbonate
dGTP	deoxyGuanosine TriPhosphate
DMD	Differentially Methylated Domain
DMEM	Dulbecco's Modified Eagle Medium
DMSO	DiMethyl SulfOxide
DNA	DeoxyriboNucleic Acid
DNaseI	Deoxyribonuclease I
dNTPs	deoxyNucleotides TriPhosphates
dTTP	deoxyThymidine TriPhosphate
EBA	Enhancer-Blocking Assay
ECR	Evolutionarily Conserved Region
EDTA	EthyleneDiamineTetraacetic Acid
EMAGE	Edinburgh Mouse Atlas of Gene Expression
ENCODE	ENCyclopedia Of DNA Elements
ES cells	Embryonic Stem cells
EtBr	Ethidium Bromide
ETC	Extra TFIIIC Site
F_1	First filial generation of transgenic mice
FBS	Fetal Bovine Serum
GENSAT	Gene Expression Nervous System ATlas
GFP	Green Fluorescent Protein
GO	Gene Ontology
GROMIT	Genome Regulatory Organization Mapping with Integrated Transposons
GW	GateWay
GXD	Gene eXpression Database at MGI
HEK 293	Human Embryonic Kidney 293 cells
HEPA	High-Efficiency-Particulate-Air

HEPES	HydroxyEthylPiperazine Ethane Sulfonic acid
hpf	hours post fertilization
HS	Hypersensitive Site (to cleavage by DNase I)
i.e.	*id est*
IBM SPSS	Statistical Package for the Social Sciences by International Business Machines
ICR	Imprinting Control Region
ID	IDentifier
IDT	Integrated DNA Technologies
INIA	*Instituto Nacional de Investigación y tecnología Agraria y Alimentaria*
Kb	Kilo base pair
LAD	Lamin B1-Associated Domains
LB	Luria-Bertani broth
LCR	Locus Control Region
Mb	Mega base pair
MGI	Mouse Genome Informatics
MIR	Mammalian Interspersed Repeats
mQ H_2O	milli Q water
mRNA	messenger RiboNucleic Acid
NCBI	National Center for Biotechnology Information
OD	Optical Density
OMIM	Online Mendelian Inheritance in Man
ONPG	Ortho-NitroPhenyl-β-Galactoside
p.m.	*post meridiem*
PBS	Phosphate Buffered Saline
PCM	*Parque Científico de Madrid*
PCR	Polymerase Chain Reaction
PFA	ParaFolmAldehyde
PMSF	PhenylMethaneSulfonylFluoride
PTU	1-Phenyl-2-ThioUrea
qPCR	quantitative PCR
RABS	Repeat-Associated Binding Sites

RNA	RiboNucleic Acid
RPMI	Roswell Park Memorial Institute
RT	Room Temperature
SAS	Statistical Analysis System
scs, scs'	Specialized Chromatin Structures
SD	Standard Deviation
SDS	Sodium Dodecyl Sulfate
SEM	Standard Error of the Mean
SGAI	*Secretaría General Adjunta de Informática*
SINE	Short Interspersed Nuclear Element
S/MAR	Scaffold/Matrix Attachment Region
SOC	Super Optimal Broth
SPF	Specific Pathogen-Free
SSC	Saline-Sodium Citrate
SV40	Simian Vacuolating virus 40
TAD	Topologically Associating Domain
TAE	Tris-Acetate-EDTA
TALEN	Transcription Activator-Like Effector Nuclease
Taq	*Thermus aquaticus*
TE	Tris-EDTA
Tm	melting Temperature
TR	Thyroid hormone Receptor
TRANSFAC	TRANScription FACtor database
TRE	Thyroid hormone Response Element
trf	tandem repeats
tRNA	transfer RNA
TS	Theiler Stage (for mouse development)
TSS	Transcription Start Site
UAM	*Universidad Autónoma de Madrid*
UAS	Upstream Activation Sequence
UCSC	University of California Santa Cruz

UNAM	*Universidad Nacional Autónoma de México*
UTR	UnTranslated Region
UV	UltraViolet
WT	Wild Type
YAC	Yeast Artificial Chromosome
ZFN	Zinc-Finger Nuclease

1 INTRODUCTION

1.1. Where Does a Gene Begin? Where Does it End? Expression Domains and Boundaries

More than two hundred cell types exist in the human body (Gartner & Hiatt, 2001). This is not surprising given the vast diversity of tasks that our organism carries out to function properly and survive: lung alveolar cells permit the CO_2-O_2 interchange when breathing, intestinal epithelial cells absorb most of the nutrients from the diet, whilst cardiac cells keep our hearts beating. Nevertheless, every cell in our body, regardless of its function, originated from the same fertilized oocyte and thus carries the same genome, the same genetic information[1]. The way they make use of that information is what differentiates them (Splinter & De Laat, 2011).

Before the first drafts of the human genome sequence came to light (Lander et al., 2001; Venter et al., 2001), scientists had predicted the existence of around 100,000 genes. However, this guess was actually found to be a four-fold overestimation. Current estimates indicate that humans possess roughly 21,000 protein-coding genes (genome assembly GRCh37, data from the Genome Reference Consortium: http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/), accounting for only around two per cent of our whole genome. At first, the rest was considered as useless "junk" DNA. On the contrary, it is widely accepted nowadays that most of this non-coding portion of the genome is full of *cis*-regulatory modules that dictate when, where, and to what extent a gene must be expressed (Cecchini et al., 2009; Riethoven, 2010). As a matter of fact, it has been hypothesized that it is this richness in regulatory elements which makes humans such complex organisms (Levine & Tjian, 2003). The same thesis can be applied to mice and other mammals with similar genomes.

In this "regulatory jungle" in which the genomes of higher eukaryotes are immersed (Ruf et al., 2011), how does a specific enhancer or silencer know which gene (or genes) it has to target? For a long time, it has been thought that regulatory elements exert their actions on the closest gene. However, recent data reject this assumption. Dekker et al. (Sanyal et al., 2012) sought to determine all long-range interactions (or interactions between distant loci) between promoters and distal regulatory elements in the 1% of the human genome the ENCODE (ENCyclopedia Of DNA Elements) pilot project has focused on (ENCODE Project Consortium, 2004). They concluded that only 27% of distal elements associates with the closest TSS (Transcription Start Site). This finding adds to earlier observations that hinted that gene regulation processes are more complex than

[1] B and T lymphocytes, which undergo somatic recombination, are the exception.

anticipated. For example, it is recommended to design transgenes that carry, not only the gene of interest, but also a large portion of the endogenous locus, if a faithful recapitulation of the expression program in the transgenic model is desired (Bonifer, 2000). This can be achieved by using transgenes based on bacterial or yeast artificial chromosomes (BACs or YACs, respectively) (Giraldo & Montoliu, 2001; Montoliu, 2002; Montoliu et al., 2009). However, the inclusion of several hundreds of kilobases of flanking sequence is sometimes not enough. This is the case of the murine *Gata-3* locus. Engel and colleagues found that a 120 kb YAC-based *Gata-3* transgene was unable to establish the correct *Gata-3* expression pattern (Lakshmanan et al., 1998), the reason being that additional elements located further away in the genome –even bypassing other genes- were also necessary (Lakshmanan et al., 1999; Hasegawa et al., 2007). Similarly, in other loci, such as in the Sonic hedgehog gene (*Shh*), enhancers have been mapped up to 1 Mb away from the gene they regulate, again, with several unrelated genes in between (Lettice et al., 2003).

These findings, among others, suggest that the concept of gene locus, usually regarded as just the protein-coding sequence and a few kilobases of flanking genome, should be reconsidered. In fact, an alternative model of domain organization was already proposed two decades ago. This model is based on expression domains, defined as the portions of genome that contain protein-coding sequences (one or more genes) and all the regulatory elements needed for their correct expression in time and space (Eissenberg & Elgin, 1991; Giraldo & Montoliu, 2001; Montoliu, 2002; Dillon, 2006).

Spitz and co-workers developed an elegant strategy to readily identify expression domains in the mouse genome. GROMIT (Genome Regulatory Organization Mapping with Integrated Transposons) relies on the mobilization throughout the genome of a *Sleeping Beauty* transposon that carries a *lacZ* reporter gene under a promoter that responds to the enhancers nearby (Ruf et al., 2011). With this strategy, the simple *in vivo* analysis of *lacZ* expression permitted the identification of all the regulatory activities interacting with each integration site. Moreover, the range of action of the regulatory elements involved could also be assessed. For instance, the analysis of the expression of the transposons that had integrated in the same locus –although not in the same exact position- in different transgenic lines revealed that they usually exhibited very similar expression patterns. This observation indicated that all those positions belonged to the same expression domain. However, in some cases, abrupt changes in expression profiles could be observed, disclosing the existence of a transition between domains.

Next, Ren et al. coined the concept of "topological domains" (Dixon et al., 2012). They studied the long-range interaction landscape of the mouse and human genomes, using both embryonic and terminally differentiated cell types. They described megabase-sized domains, the "topological domains", in which extensive long-range interactions between regulatory modules (i.e., enhancer-promoter) occur. As previously suggested by Spitz's work, interactions between elements that belong to different, albeit adjacent, domains rarely take place. Since they were more or less conserved between species and in the different cell types assayed, Ren's group hypothesized that these domains were the basic organizational units of the genome. At the same time, similar domains were described in much more detail in the mouse X chromosome (TADs or Topologically Associating Domains; Nora et al., 2012).

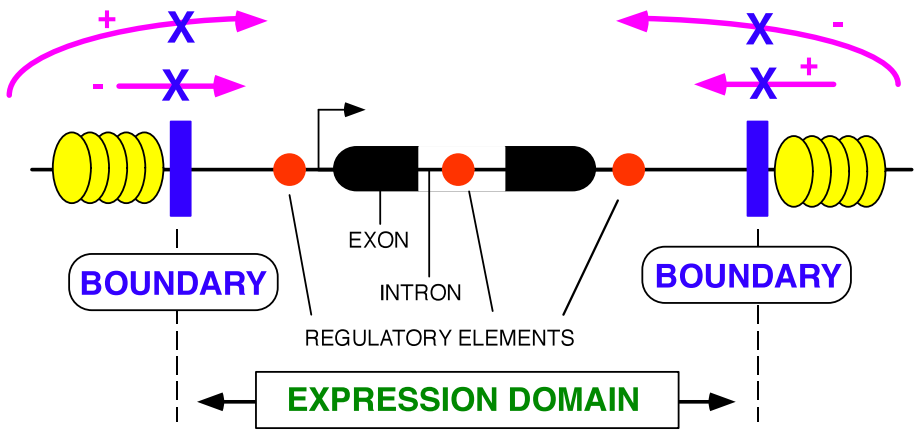


Fig. I 1. The concept of expression domains. An expression domain contains one or more genes and all the regulatory elements that ensure their accurate pattern of expression, including genomic boundaries. Figure from Molto et al., 2011.

The existence of defined expression domains implies the existence of boundaries that separate and render them independent (**Fig. I 1**). The nature of these boundaries is diverse. Genomes continuously evolve: reorganizations, duplications and deletions take place, retrotransposable elements mobilize, and point mutations randomly generate (or erase) protein binding sites. In the midst of these processes, the appearance of any element whose function is beneficial for the establishment of a boundary will be fixed (or, at least, not selected against) throughout evolution at the borders between domains. On many occasions, several boundary-related mechanisms reinforce these borders. Most likely, not all of them are required, but had endured evolution because they are not prejudicial to the cell (Dillon & Sabbattini, 2000). On the contrary, the emergence of a boundary in a place where its function is detrimental will be evolutionarily and rapidly selected against. The work of Dean and colleagues (Hou et al., 2008) supports this

hypothesis. They engineered a transgene that harbored the complete human locus of clustered β-globin genes (including its flanking boundaries) and an additional boundary (the exact same 5' boundary) in between the locus control region (LCR) and the first gene of the cluster. Transgenic mice for this construct showed a reorganization of the three-dimensional structure of the locus that isolated the LCR from the globin genes, preventing their interaction and thus, their expression. Had this boundary materialized there in an organism at some point, it would not have conferred an evolutionary advantage. That is probably the reason why this arrangement of regulatory elements has never been found so far in this locus.

Therefore, it appears reasonable to assume that cells would need boundaries to organize their genomes. But, how do these boundaries work? Some of them tether the chromatin to particular sub-nuclear structures separating two independent domains. Others create a chromatin loop that contains a specific domain, isolating it from the neighbors, by the formation of long-range interactions mediated by proteins at the base of the loop. Also, some boundaries attract a myriad of protein factors (transcription machinery, chromatin remodeling complexes, etc.) that create a physical blockage that restrains any kind of communication between domains (West et al., 2002; Wallace & Felsenfeld, 2007). These and any other pre-existing mechanism in the cell able to establish a boundary would be, in principle, accepted. Hence, genomic boundaries are not associated with a single type of mechanism; they are only defined by their function (Engel & Bartolomei, 2003): they flank expression domains and protect them from undesirable regulatory input from the surroundings (**Fig. I 1**).

1.2. A Little Bit of History and the Chicken cHS4 Insulator

According to the classical view, an element can be considered a genomic boundary or insulator if it possesses enhancer-blocking and/or barrier activities. The former property refers to the fact that some boundaries can block the influence of a distal enhancer on a promoter but only when placed in between them. On the other hand, barrier elements prevent inactive heterochromatin from encroaching on adjacent active euchromatin, thus impeding its silencing (Bell et al., 2001; West et al., 2002; Gaszner & Felsenfeld, 2006) (**Fig. I 2**).

The first boundaries were described in the *Drosophila melanogaster* genome. Specialized Chromatin Structures, *scs* and *scs'*, flank two divergent copies of normally inactive *hsp70* heat-shock genes at the 87A7 chromomere. Upon environmental stress, the

locus decondensates and transcription starts. When the *stimulus* disappears, compaction of the chromatin takes place and the genes become silent again. *Scs* and *scs'* act as barrier insulators that limit the decondensation-compaction processes and serve to define an expression domain at this location (Udvardy et al., 1985). Later, the properties of these elements were explored using transgenic flies in what is now considered the first assay ever developed for testing boundary activity *in vivo*. Usually, transgenes integrate randomly into the genome, and their expression depends, not only on the regulatory elements that were initially placed in the constructs, but also on those present in the genomic locus where they land. For instance, they will remain silent if they integrate into a highly condensed heterochromatic region even if powerful enhancers were included. Also, endogenous enhancers at the insertion site will alter the expression profile of the transgene, making it be expressed in tissues and/or developmental stages different from what was expected. These are some examples of the phenomenon known as chromosomal position effects (Wilson et al., 1990; Giraldo & Montoliu, 2001).

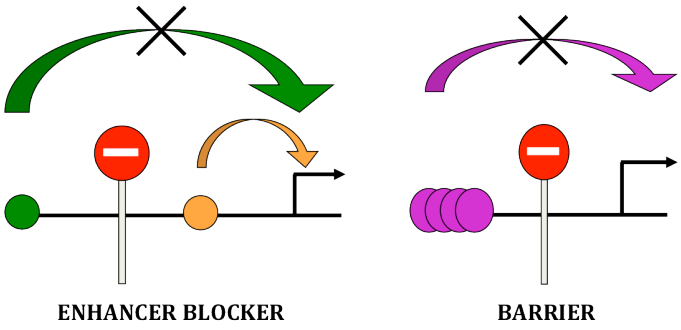


Fig. I 2. Boundary properties. Some boundaries block the action of a distal enhancer on a promoter when placed between the two (enhancer-blocking activity). In addition, others are able to prevent the spread of advancing silencing heterochromatin into an active euchromatic region. Often, boundaries only display one of these properties. Figure from Molto et al., 2011.

Kellum and Schedl demonstrated that the *scs* and *scs'* elements, when shielding a transgenic construct, were able to prevent chromosomal position effects by settling insulated independent domains that prevailed unaffected by the genomic context (Kellum & Schedl, 1991). Specifically, they used constructs with two different versions of the *white* gene: a maxigene that contained all the regulatory elements required to produce wild-type eye color in transgenic flies, and a minigene whose expression was very low and thus, generated flies with pale yellow eyes. They found that when the *white* maxigene was flanked by *scs* and *scs'*, most of the transformants showed wild-type eye color indistinguishable from non-transgenic flies with endogenous *white* expression. These results were not obtained when the maxigene was flanked by random unrelated sequences of the same size as *scs* and *scs'*, or when the original maxigene was used. In these cases, *white* expression was influenced by the genomic context at the site of insertion and hence, was silenced to various levels in the different transgenic lines.

Moreover, the same set of experiments with the *white* minigene revealed that *scs* and *scs'* were also able to maintain the low level of transgene expression in all lines. On the contrary, transgenic flies for the unprotected *white* minigene or with random flanking DNA, showed a wide range of eye colors as a result of the interactions with positive regulatory elements at the integration site. All these experiments suggest that *scs* and *scs'* protect from chromosomal position effects, and hence, they possess barrier activity. Throughout the years, this same test has been applied successfully to identify other genomic boundaries, as in the case of the boundary located upstream of the mouse tyrosinase gene (Giraldo et al., 2003a).

Furthermore, these two scientists devised an additional assay to test the enhancer-blocking properties of these elements. They generated transgenic flies with a construct that contained the *yp-1* enhancer and a *lacZ* cassette under the control of the *hsp70* promoter. They found β-galactosidase activity in the fat body tissue of adult females, an expression pattern driven by the *yp-1* enhancer. However, when they cloned *scs* between the enhancer and the promoter, β-galactosidase activity disappeared. Cloning *scs* upstream from the enhancer restored transgene expression, ruling out the possibility that the element was acting as a silencer, rather than as an insulator, whose function is position dependent (Kellum & Schedl, 1992).

The observation of a mutant phenotype in the fruit fly led to the discovery of a second boundary with enhancer-blocking ability (Geyer & Corces, 1987). The *yellow* gene is responsible for the pigmentation of cuticle structures in the fly. Several enhancers, located upstream from the gene as well as in an intron, regulate its expression. However, the insertion of the *gypsy* retrotransposon in the middle of the regulatory landscape inhibits the expression of the *yellow* gene in a tissue-specific manner by only blocking the action of the upstream enhancers, and not of those proximal to the promoter (Geyer & Corces, 1992).

Since the discovery of these first elements, many more have been described in *Drosophila* (reviewed in Gurudatta & Corces, 2009; i.e., Negre et al., 2010). But insulators have not only been identified in flies. They have also been found in yeast (reviewed in Amouyal, 2010a), sea urchin (i.e., Palla et al., 1997), frog (Robinett et al., 1997), chicken (i.e., Chung et al., 1993; Furlan-Magaril et al., 2011), mouse (reviewed in Molto et al., 2009), goat (i.e., Soulier et al., 2000), human (reviewed in Molto et al., 2009; i.e., Raab et al., 2012), plants (reviewed in Singer et al., 2012) and viruses (i.e., Chen et al., 2007).

The most widely known and best-characterized boundary in vertebrates maps to the 5' end of the *Gallus gallus* β-globin locus (**Fig. I 3**; Chung et al., 1993). Four DNaseI hypersensitive sites (DNaseI HS) conform the LCR that separates the erythroid-specific chicken β-globin gene cluster from a folate receptor gene, located further upstream and also expressed in erythrocytes but earlier in development (Prioleau et al., 1999). Three of these DNaseI HS are erythroid-specific, while the fourth, the most upstream one, is constitutive and functions as a boundary. It is cHS4, the first to be reported in a vertebrate genome. This 1.2-kb insulator exhibits both barrier and enhancer-blocking properties, although they are conveyed through independent mechanisms (Recillas-Targa et al., 2002). A detailed analysis of the region revealed that much of the boundary activity is contained within a 250-bp GC-rich core, in which 5 protein binding site "footprints" were identified (Chung et al., 1997). Although the strongest enhancer-blocking activity falls on footprint II, the combination of footprints II and III recapitulates most of the activity of the core (Recillas-Targa et al., 2002). In fact, four tandem copies of this 90-bp fragment exhibit a more powerful activity than the original 1.2-kb element (Recillas-Targa et al., 1999). The protein CTCF, 'master weaver of the genome' (see below, Phillips & Corces, 2009), binds to footprint II and is fully responsible for the enhancer-blocking activity (Bell et al., 1999; Recillas-Targa et al., 1999). This activity presumably results from the chromatin loops that arise after CTCF tethers the cHS4 insulator to the nucleolar surface through its interaction with nucleophosmin/B23 (Yusufzai & Felsenfeld, 2004a; Yusufzai et al., 2004b).

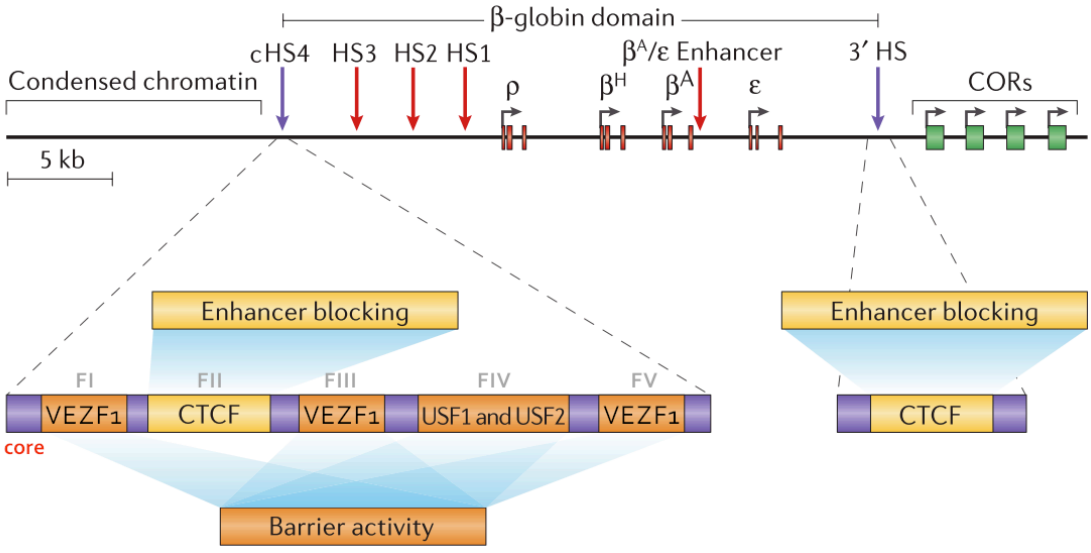


Fig. I 3. Organization of the chicken β-globin domain. The chicken β-globin genes (red boxes) are flanked by a highly condensed heterochromatic region at the 5' end, and by a cluster of olfactory genes CORs (green boxes) at the other end. The 250-bp core (inset) accommodates much of the insulator activity of the cHS4 element (purple arrow), which possesses CTCF-dependent enhancer-blocking activity (FII), as well as barrier activity mediated by VEZF1 (FI, FIII, FV) and USF1/2 proteins (FIV). At the other end of the locus, the 3'-HS insulator (purple arrow) only functions as an enhancer-blocker through its interaction with CTCF. Red arrows depict the DNaseI HS that conform the LCR. Adapted from Gaszner & Felsenfeld, 2006.

The rest of the footprints are involved in the establishment of a barrier that protects the β-globins from the silencing effects of a highly condensed 16-kb region located in between the folate receptor gene and the β-globin cluster (Prioleau et al., 1999). The basic helix-loop-helix leucine zipper transcription factors USF1 and USF2 bind to footprint IV and recruit histone modifying enzymes that act as chain terminators for the propagation of the heterochromatic marks that originate at the condensed 16-kb region (Litt et al., 2001a; Litt et al., 2001b; West et al., 2004). In addition, it has been shown that VEZF1 binds to footprints I, III and V, and maintains low levels of DNA methylation at the locus. The recruitment of this zinc finger protein to the boundary is essential for the formation of a functional barrier (Dickson et al., 2010).

Not only is the 5' end of the β-globin cluster protected, but the 3' end is insulated as well. In this case, a CTCF-dependent enhancer-blocking element, 3'-HS, separates the globins from a cluster of olfactory receptors, which are expressed in the olfactory epithelium and in the brain, but not in erythrocytes (Saitoh et al., 2000).

Unlike the chicken, in mammals there are olfactory genes on both sides of the β-globin cluster, and no heterochromatic domain has been observed upstream of the locus. However, in mice and humans the cluster is also flanked by DNaseI HS that act as CTCF-dependent enhancer-blockers in ectopic constructs (Farrell et al., 2002). Additionally, it has been shown that these sites, together with the LCR, interact in space (Tolhuis et al., 2002; Palstra et al., 2003) forming the so-called "active chromatin hub", which was hypothesized to be required for the correct expression of the genes during erythroid differentiation (Palstra et al., 2003; De Laat & Grosveld, 2003). However, the true relevance of these sites at the endogenous locus is still controversial, since their deletion from the genome (Bender et al., 2006), or the disruption of the CTCF binding site at the 3' boundary (Splinter et al., 2006), at least in mice, does not have a significant negative effect on the expression of the β-globins.

1.3. Mechanisms of Insulator Function

Boundaries employ a wide range of mechanisms to exert their functions. The reason is that cells have exploited the DNA elements and the molecular machinery already present at each locus, and adapted them *ad hoc* to the establishment of a boundary. However, mechanisms of boundary activity can be grouped under the following models: 1) physical obstacle to a processive signal emanating from an enhancer or a heterochromatin

focus ("roadblock model"), 2) sequestering of enhancers or promoters in order to abolish their communication ("decoy model"), 3) formation of chromatin loops that restrict the access of certain regulatory elements to a given promoter ("topological looping model"), or 4) recruitment of chromatin remodellers and histone modifiers that maintain different chromatin states at each side of the boundary (reviewed in Bell et al., 2001; West et al., 2002; Engel & Bartolomei, 2003; Valenzuela & Kamakaka, 2006; Wallace & Felsenfeld, 2007).

The focus of this section will be placed on vertebrate boundaries, with an exception. Early experiments in yeast were fundamental in enlightening how the RNA polymerase transcription machinery is linked with insulation in mammals, so they will be reviewed as well. Likewise, studies in *Drosophila* have greatly expanded our knowledge on boundaries (Gurudatta & Corces, 2009). A very complex picture emerges in this organism, where many insulator proteins have been characterized. Yet, only one of them functions in mammals (CTCF, see below), while the rest are restricted to the *Drosophila* lineage (Schoborg & Labrador, 2010). For this reason, examples in the fruit fly will not be further explored.

1.3.1. The Omnipresent CTCF

The CCCTC-binding factor or CTCF (Filippova, 2008; Ohlsson et al., 2010a; Herold et al., 2012) is a DNA-binding protein implicated in the mechanism of action of most vertebrate insulators described to date. It was independently discovered twice, in both instances as a transcriptional repressor. Years later, it was found that the proteins that regulated the expression of the chicken c-*myc* (Lobanenkov et al., 1990) and *lysozyme* (Kohne et al., 1993) genes were the same one (Burcin et al., 1997).

CTCF is a ubiquitously expressed cell-cycle regulated nuclear 11-Zn-finger protein (Klenova et al., 1993; Klenova et al., 1998; Filippova et al., 1998) that has been strongly conserved throughout evolution. In fact, the global identity in the amino acid sequence of the chicken and human homologs reaches 93%, and attains 100% if only the zinc-finger domain is considered (Filippova et al., 1996). It is also present in the mosquito genome (Gray & Coates, 2005), as well as in other phylogenetically distantly related species such as the fruit fly (Moon et al., 2005), the zebrafish (Pugacheva et al., 2006) and the frog (Burke et al., 2002). These findings, combined with the fact that null mice for this protein display early embryonic lethality (Fedoriw et al., 2004; Heath et al., 2008) indicate that CTCF may be playing crucial roles in eukaryotic organisms.

Indeed, it has been shown that CTCF is involved in a myriad of functions, including transcriptional regulation (i.e., Vostrov & Quitschke, 1997; Awad et al., 1999), X-chromosome inactivation (Chao et al., 2002; Spencer et al., 2011), imprinting (Bell & Felsenfeld, 2000; Hark et al., 2000; Yoon et al., 2005; Hancock et al., 2007; Fitzpatrick et al., 2007), stabilization of trinucleotide repeats (Filippova et al., 2001; Cho et al., 2005; Libby et al., 2008), tumor suppression (Filippova et al., 1998; Rasko 2001), differentiation (Torrano et al., 2005; Delgado-Olguin et al., 2011), V(D)J recombination (Seitan et al., 2012), RNA polymerase II pausing with implications in alternative splicing (Shukla et al., 2011), promoter choice (Guo et al., 2012; Monahan et al., 2012), cellular memory (Burke et al., 2005), insulation (Yang & Corces, 2011b) and organization of the nuclear architecture (Phillips-Cremins & Corces, 2013a), as shall be described later. How can this functional versatility be explained? The answer probably lies in the Zn-finger domain (**Fig. I 4**), which confers on the protein high flexibility regarding the DNA sequences it can recognize, together with the protein partners it can interact with.

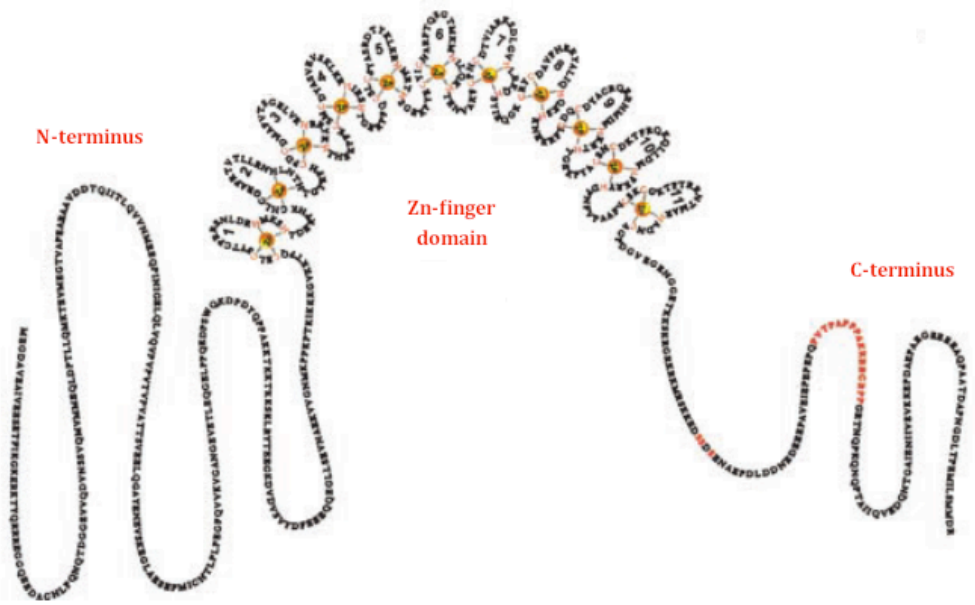


Fig. I 4. Structure of the human CTCF protein. Schematic drawing depicting the NH_2- and COOH- terminal domains of human CTCF, as well as its central 11-Zn-finger DNA binding domain, which consists of 10 Cys_2-Hys_2-class and 1 Cys_2-His-Cys-class Zn fingers (Klenova et al., 1993). Adapted from Filippova et al., 2002.

1.3.1.1. The "CTCF Code"

The exact CTCF binding motif continues to be the subject of debate even after the first characterizations of the CTCF target sites (CTSs) in the chicken (Klenova et al., 1993) and human (Filippova et al., 1996) genomes were conducted two decades ago. Only the fact

that methylation of CpGs within the CTCF binding site abolishes its binding is unreservedly accepted (Bell & Felsenfeld, 2000; Renda et al., 2007).

Many genome wide analyses have sought to determine the consensus sequence (**Fig. I 5**) bound by this "multivalent factor" (Filippova et al., 1996). Early studies, using bioinformatic approaches (Xie et al., 2007) or chromatin-immunoprecipitation-derived technologies like ChIP-on-chip (Kim et al., 2007) or ChIP-Seq (Barski et al., 2007; Jothi et al., 2008; Cuddapah et al., 2009), highlighted similar 19-21 bp core motifs, although quite significant variation in this core could be observed across mammalian genomes, even in the same species. Recent data show that a good fraction of all CTSs are bimodular (Schmidt et al., 2012) or even trimodular (Rhee & Pugh, 2011; Nakahashi et al., 2013): the central 20 bp core would coincide with what had previously been reported, whereas additional motifs found upstream and/or downstream from the core would modulate the strength of the protein-DNA interaction. Of note, not all predicted sites identified in these large screenings are necessarily directly bound by CTCF *in vivo*. Instead, CTCF may be indirectly associated with some of these sites through its interaction with additional proteins (Phillips & Corces, 2009; Nakahashi et al., 2013). This would explain why 5-25% of all CTSs, which is by no means negligible, mismatch the consensus (Kim et al., 2007; Jothi et al., 2008; Cuddapah et al., 2009; Nakahashi et al., 2013).

CTCF, as any other DNA-binding factor, does not always associate with the same strength to the core motif at different loci. Instead, the neighboring base pairs at each

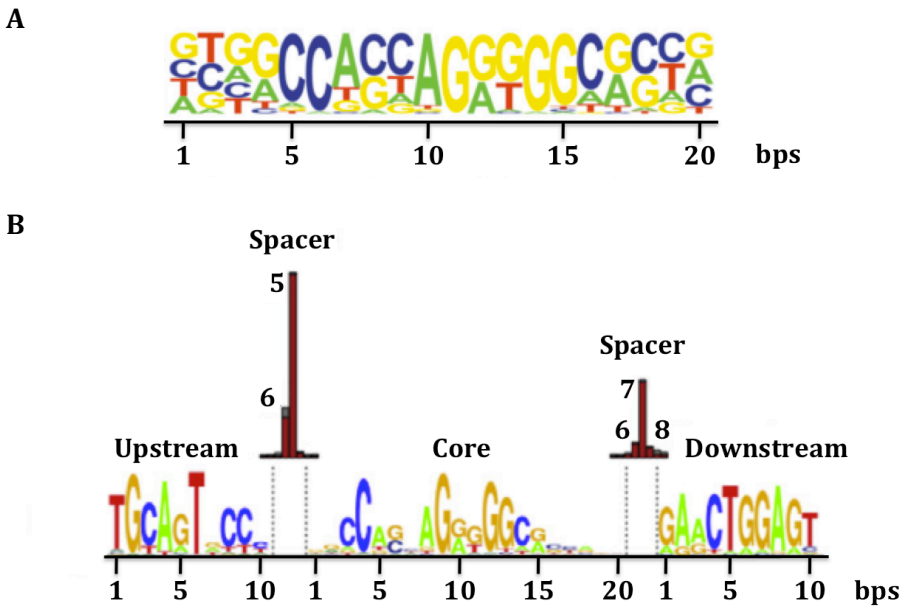


Fig. I 5. Evolution of the discovery of the CTCF binding motif. The first genome-wide mapping of CTSs in human revealed a 20-bp consensus site (A; Kim et al., 2007). It has recently turned out that this motif forms the core of the murine binding site, which is completed by two additional motifs. The upstream and downstream modules are separated from the core by spacers of variable length (B; Nakahashi et al., 2013).

location help in the stabilization of the interaction. They protect and maintain strong interactions even after genetic variation at evolutionarily constrained positions in CTSs (Maurano et al., 2012). This buffering effect may also account for the highly variable CTCF consensus sequence.

According to the "CTCF code" (Ohlsson et al., 2010b), CTCF employs a particular combination of fingers each time, depending on the DNA sequence it is due to bind (see Burcin et al., 1997 for an example). However, not all combinations are possible. Apparently, contiguous fingers (i.e., fingers 1-2 or 9-11) function as indivisible blocks (Nakahashi et al., 2013), with the central block (fingers 4-7) being indispensable for robust binding to DNA (Renda et al., 2007). The fingers left unused are then free to interact with a wide variety of proteins (reviewed in Zlatanova & Caiafa, 2009; Ohlsson et al., 2010b). The final functional outcome would hence depend on its binding partner, which in turn is determined by the underlying genomic sequence.

1.3.1.2. CTCF, "The Master Weaver of the Genome"

CTCF carries an astonishingly large list of functions, and most of them may be occurring simultaneously in the cell. This complicates the assessment of a particular function *in vivo* without interfering with the others.

All of CTCF functions, apart from its role in insulation, are beyond the scope of this PhD thesis and thus will not be discussed, except for two: regulation of transcription and establishment of long-range interactions.

Insulators set boundaries between domains; they are normally neutral elements (Bell et al., 2001), so the proteins that mediate their function are expected to be neutral as well. However, many laboratories have shown that CTCF apparently exerts both positive and negative effects on transcription. Because these observations apparently contradict the expected neutral behavior of an insulator protein, the implication of CTCF in regulating transcription will be considered as well.

Finally, studies about the role of CTCF in establishing long-range interactions have been of utmost importance to solve the CTCF mystery. They have contributed to reconcile all its functions into a single one: the organization of the nuclear architecture. For this reason, they will also be reviewed.

1.3.1.2.1. CTCF Role as a Transcriptional Regulator

Originally, CTCF was described as a negative element that repressed the expression of the chicken *lysozyme* (Burcin et al., 1997) and c-*myc* (Filippova et al., 1996) genes, with the help of YB-1 in the latter case (Chernukhin et al., 2000). Moreover, recent studies implicate CTCF in the silencing of testis-specific human *SPANX* genes in somatic tissues (Kouprina et al., 2007), as well as of the catalytic subunit of telomerase, *hTERT*, in lowly proliferative adult cells (Renaud et al., 2005; Renaud et al., 2007). Apparently, CTCF, in tandem with the transcriptional co-represor SIN3A, recruits histone deacetylases that trigger the silencing cascade (Lutz et al., 2000). At the same time, other laboratories have reported transcriptional activation mediated by this same protein in the human *APP* (Vostrov & Quitschke, 1997) and *IRAK2* (Kuzmin et al., 2005) genes.

Protein truncation experiments were carried out in an attempt to clarify the real behavior of CTCF in transcriptional regulation. The aim of these studies was to identify the functional role of each CTCF domain. Not surprisingly, the experiments revealed both transactivating and repressing domains scattered throughout the protein (Lutz et al., 2000; Defossez & Gilson 2002; Filippova et al., 1996; Kitchen & Schoenherr 2010).

These contradictory data can be easily reconciled under the "CTCF code" hypothesis: it all depends on the protein partner CTCF interacts with at each locus.

1.3.1.2.2. CTCF Role in Insulation

1.3.1.2.2.1. Barrier Activity

Some insulators exhibit both enhancer-blocking and barrier properties, like the chicken cHS4 element (Chung et al., 1993). As discussed above, both activities are separable and depend on different proteins. Particularly, CTCF conveys enhancer-blocking activity at this locus (Recillas-Targa et al., 2002). Ever since this discovery, the general trend in the field was to associate CTCF only with enhancer-blocking, discarding a possible connection with the establishment of barriers (Barkess & West, 2012). Nevertheless, CTCF is actually responsible for both activities at the αEHS-1.4 insulator, a 1.4-kb element that maps to an erythroid specific DNaseI HS upstream from the chicken α-globin domain. Not only is it indispensable for blocking enhancer function in ectopic constructs *in vitro* (Valadez-Graham et al., 2004), but it also confers protection against chromosomal position effects in transgenic mice, possibly in collaboration with additional factors (Furlan-Magaril et al., 2011).

Furthermore, CTSs are enriched at the boundaries between euchromatic and heterochromatic domains in human cells (Barski et al., 2007; Cuddapah et al., 2009; Chen et al., 2012), including the transition regions between silenced condensed chromatin and active escape domains at the inactive X-chromosome (Filippova et al., 2005). Additionally, CTSs have also been found at the borders of lamin B1-associated domains, or LADs, in human (Guelen et al., 2008). These gene-poor regions possess characteristics of heterochromatin, in sharp contrast with the adjacent regions. However, not all borders contained CTSs, and in those that did contain them, CTSs were located at a considerable distance (5 to 10 kb apart). Therefore, CTCF may not be supporting genome-lamina interactions, at least directly, at these borders.

The most investigated function of CTCF is that of bringing together in space distant loci, which can map in the same or even in different chromosomes (Phillips-Cremins & Corces, 2013a). A recent ChIA-PET experiment (Chromatin Interaction Analysis by Paired-End Tag sequencing, Fullwood & Ruan, 2009) uncovered 1,480 intra- and 336 inter-chromosomal long-range interactions mediated by CTCF in murine ES cells (Handoko et al., 2011). More than 70% of all such interactions created looped domains that separated regions in which opposing histone modification marks accumulated (**Fig. I 6**). This result suggests that, indeed, CTCF participates in the erection of domain barriers, probably through the recruitment of other factors like the CHD8 chromodomain helicase. Although the exact mechanism by which CHD8 is required for insulation at certain loci is unknown (Ishihara et al., 2006), it probably relates to its chromatin remodeling capabilities.

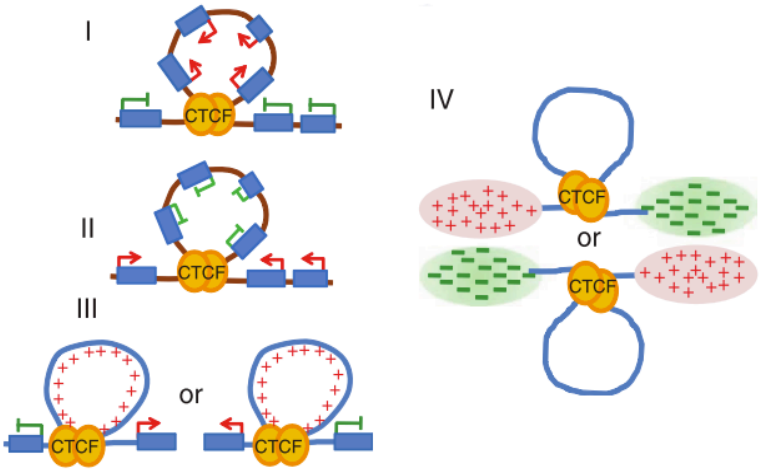


Fig. I 6. CTCF-dependent long-range interactions partition the chromatin into regions with opposite chromatin states. Loops in category I encapsulate active chromatin marks, while leaving repressive ones on the outside. The inversed picture accounts for category II loops. Categories III and IV define loops that separate opposite chromatin profiles at either side of the base: while type III loops harbor active histone marks, type IV loops do not show any particular pattern on the inside. The fifth type of loop (27%) does not fall into any of the previous categories. Adapted from Handoko et al., 2011.

The observation that some CTSs locate between positioned nucleosomes (Filippova et al., 2001; Cuddapah et al., 2009; Chen et al., 2012) only adds more controversy. Weng et al.

demonstrated that CTCF binds in the linker region between perfectly phased nucleosomes, twenty on each side. They proposed that the highly accessible DNA regions between the nucleosomes could be used as landing platforms for other proteins that would prevent the spreading of euchromatin or heterochromatin through the CTS. Thus, CTCF may be indirectly assisting the establishment of barriers genome-wide (Fu et al., 2008). Nevertheless, CTCF clearly does not act as a nucleosome positioning factor at specific genomic positions, such as at the mouse *Igf2/H19* locus (Kanduri et al., 2002). In fact, DNA-bound CTCF is unable to prevent the repositioning of a nucleosome on the CTS, which ultimately leads to CTCF eviction and loss of insulation at the chicken lysozyme locus (Lefevre et al., 2008).

All these data indicate that CTCF is likely functioning as a barrier element at some loci. Alternatively, it may just be cooperating in the formation of one. But it is still possible that some of these CTSs are being carried along passively with other unexplored elements that do convey barrier activity (Phillips & Corces, 2009). For example, the contribution to barrier activity of a Scaffold/Matrix Attachment Region sequence (S/MAR; see below) also present in the αEHS-1.4 insulator, has not been directly addressed yet (Furlan-Magaril et al., 2011).

1.3.1.2.2.2. Enhancer-Blocking Activity

In order to understand how enhancer-blocking elements work, it is essential to first understand how enhancers work (recently reviewed in Bulger & Groudine, 2011). Models of enhancer action can be classified into two groups, depending on whether or not the enhancer needs to contact the promoter directly to effectively influence transcription. Contact models are, by far, the most popular and well-studied ones. According to the "looping model", an enhancer looks for its promoter in space through random collisions, and upon finding it, the contact is somehow stabilized. A modified version of this model posits that enhancers "track" the chromatin fiber until they reach the promoter. In any case, the portion of DNA between both elements has to loop out.

Among the noncontact models, the "spreading" mechanism of enhancer action is the best understood. It postulates that the enhancer serves as a docking platform for the recruitment of chromatin remodeling complexes. These complexes would spread bidirectionally along the DNA decondensing the chromatin, thus enabling the binding of transcriptional activators.

Conversely, two distinct mechanisms of insulation have been proposed to explain the fact that CTCF can block the effects of distal enhancers on promoters when placed between the two: topological looping and interaction models (Engel & Bartolomei, 2003; Valenzuela & Kamakaka, 2006). On the one hand, in the "topological looping model" (**Fig. I 7A**), CTCF would create loops through the establishment of long-range interactions with other proteins, including with itself (Pant et al., 2004) but mainly with the components of the cohesin complex (Rubio et al., 2008; Xiao et al., 2011). Although it has been classically thought that cohesins were only involved in the maintenance of sister chromatids cohesion during mitosis, additional roles in insulation (Parelho et al., 2008; Wendt et al., 2008), transcriptional regulation (Schmidt et al., 2010) and nuclear organization (DeMare et al., 2013) have recently emerged. Alternatively, loops would also arise after CTCF-mediated tethering of the chromatin to certain nuclear structures. Indeed, CTCF associates with the nuclear matrix (Dunn et al., 2003), and with the nucleolus (Yusufzai et al., 2004b). If an enhancer and its targeting promoter locate in different loops, their communication would be hindered, irrespective of the way the enhancer influences the promoter: looping, tracking or spreading. For example, recently it was physically demonstrated that contacts between elements residing in different loops are thwarted in favor of contacts between elements within the same loop, which supports this model of insulation (Mukhopadhyay et al., 2011). Also, both tracking and spreading mechanisms would be hampered at the attachment point with the nuclear structure or at the base of the loop.

On the other hand, "interaction models" can be further subdivided into two. First, in the "decoy model", CTCF would directly interact with the enhancer or the promoter,

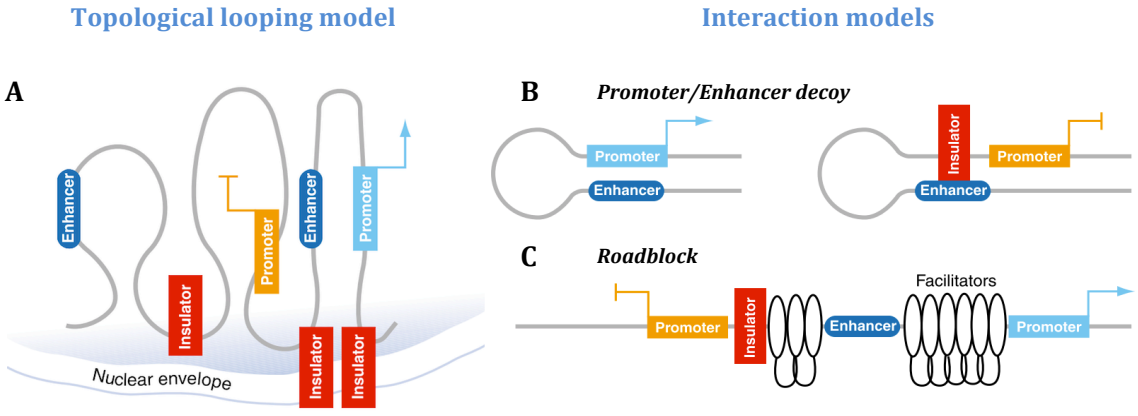


Fig. I 7. Models of insulation mediated by CTCF. A. In the "topological looping model", insulators would abolish enhancer-promoter communication by placing them in different loops. These loops would originate either from the interaction between two insulators or from the tethering of the chromatin to a nuclear structure. **B.** According to the "decoy model", CTCF would sequester the enhancer or the promoter, thus preventing their interaction. **C.** The "roadblock model" posits that CTCF would recruit additional factors in order to create a physical impediment for the advance of any processive activation signal emanating from the enhancer. Adapted from Valenzuela & Kamakaka, 2006.

precluding their communication (**Fig. I 7B**). Support for this model comes from studies demonstrating that CTSs overlaps with enhancers and promoters genome-wide, at least in the mouse genome (Shen et al., 2012). Second, the "roadblock model" suggests that CTCF, with the help of additional proteins, would establish a barrier that physically impedes the advance of any activating signals emitted by an upstream enhancer through the CTS. This model complies well with the position dependency of insulators (they only work when placed in between the enhancer and the promoter), unlike the previous one.

1.3.1.2.3. CTCF Role as an Architectural Protein

Chromatin is not randomly arranged inside the nucleus. Instead, each chromosome occupies its own spot, dubbed territory. The position of the territories is, again, not fortuitous. Gene-rich chromosomes usually group at the center of the nucleus. Here, the chances of gene expression activation are higher than in the periphery, although this does not mean that all genes located at the periphery are transcriptionally inactive (Finlan et al., 2008). Even at such a large scale, the relationship between nuclear position and gene expression is widely accepted (recently reviewed in Gibcus & Dekker, 2013). This relationship becomes more apparent when zooming in on the nuclear architecture. It is being increasingly recognized that chromosomes are organized into megabase-sized topological domains, also known as TADs (see above, Dixon et al., 2012). These domains are largely invariant between different cell types and across species. Long-range interactions between regulatory elements and promoters within each domain are common, unlike inter-domain interactions (**Fig. I 8**). Hence, TADs limit the number of contacts any enhancer can attempt before locating the right promoter. It is unclear how the boundaries between TADs are established. What is clear, however, is the fact that constitutive CTCF binding sites, along with other elements like the cohesin complex, accumulate at these boundaries, as it would have been expected (Li et al., 2013). Seemingly, CTCF may be exploiting its capability to establish long-range interactions at these locations, supporting the demarcation of TADs.

Beyond TADs, chromatin is further organized into subdomains, which are also constitutive and assisted by the combined activity of CTCF and cohesin. At this intermediate scale (100 kb - 1 Mb), subdomains would aid in the establishment of long-range interactions between distal regulatory elements and their target promoters. Here, as in TADs, CTCF would perform a structural role (as in Martin et al., 2011). Finally, long-range interactions between regulatory elements and promoters also occur at the smallest

scale (< 100 kb), and are fixed by the cohesin and mediator complexes (Kagey et al., 2010) and sometimes by CTCF. These interactions define the cellular identity by enabling the initiation of tissue- and development-specific transcriptional programs. Therefore, far from being constitutive, they vary from cell type to cell type (Hou et al., 2010), and even in the same cell type throughout development (Rajapakse et al., 2009) because they respond to cellular transcriptional requirements.

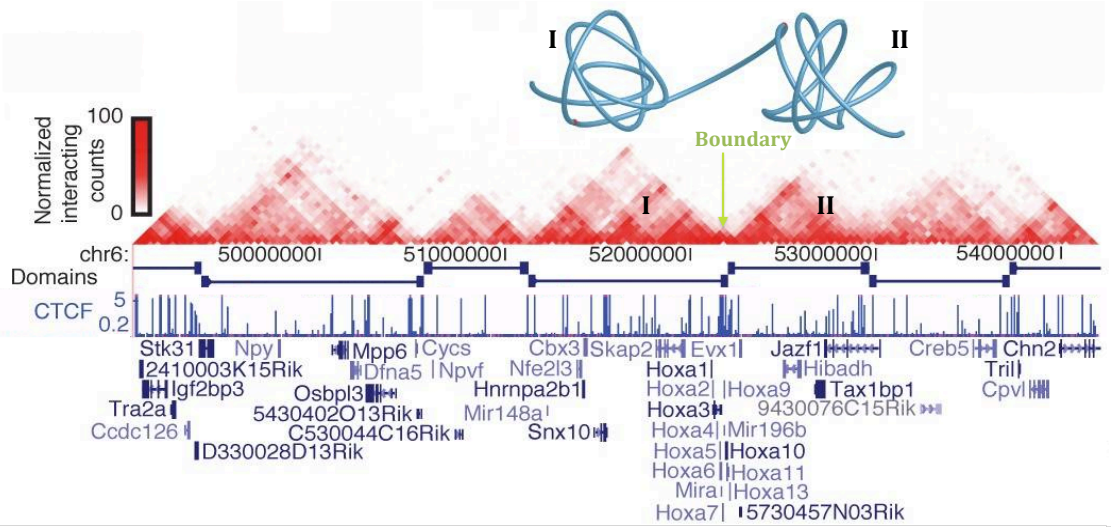


Fig. I 8. Topological domains. The heat map in the center of the figure illustrates normalized interaction frequencies between distant genomic regions in mouse chromosome 6. The map reveals distinct megabase-sized domains (I and II), dubbed "topological domains". Long-range interactions abound within the domains, but are prohibited between them. Boundary elements should exist at the edges of the domains to maintain their independency. CTCF binding sites appear scattered throughout the domains but accumulate at the edges. Adapted from Dixon et al., 2012.

To summarize, CTCF, together with other factors, would be in charge of establishing both structural and tissue-specific long-range interactions, which together serve to organize nuclear architecture. It has been suggested that this is actually the main role of this protein, which has earned CTCF the title "master weaver of the genome" (Phillips & Corces, 2009). Under this hypothesis, all functions ascribed to CTCF would be secondary, measurable byproducts of this main activity.

But, how can the transcriptional regulation and insulator functions be explained under this model? It all depends on the type of interaction promoted at each location. For example, an enhancer would readily find its target promoter if they are trapped within the same CTCF-mediated loop (**Fig. I 9A**). In addition, CTCF could put them directly into contact, given that a significant portion of its binding sites overlaps with enhancers and promoters (**Fig. I 9B**; Shen et al., 2012). This finding was originally taken as evidence in favor of the decoy model of enhancer-blocking activity conferred by CTCF (see above). However, it could be alternatively interpreted as an "enhancer-bridging activity"

(Handoko et al., 2011) that would facilitate gene expression by bringing into contact enhancers and promoters. These findings would turn CTCF into an indirect transcriptional activator at some genomic sites. On the other hand, the inhibition of gene expression could result, for instance, from the tethering of a gene locus to repressive Polycomb bodies by CTCF (as in MacPherson et al., 2009), whereas the relationship between long-range interactions and insulation has already been abundantly described (see "topological looping model" above).

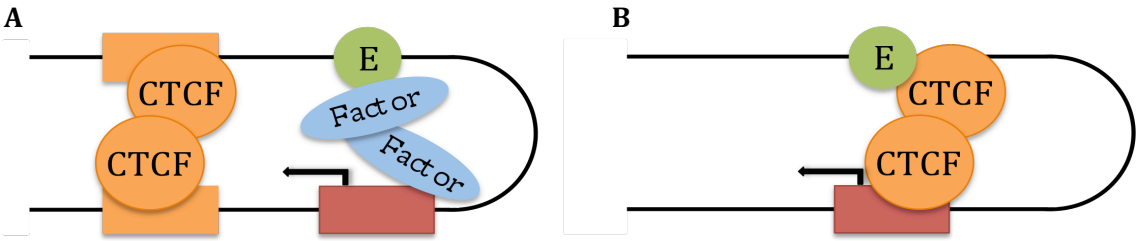


Fig. I 9. Chromatin loops mediated by CTCF promote enhancer-promoter communication. CTCF, alone or in combination with additional proteins, establishes long-range interactions that bring into close proximity an enhancer (E, green circle) with its target promoter (red rectangle). These interactions can occur between insulator sites (A) or directly between enhancer and promoter (B). Adapted from Krivega & Dean, 2012.

This is just a sample of all the possible ways in which CTCF-mediated chromatin loops could regulate gene expression or promote insulation. More scenarios can arise (see Kadauke & Blobel, 2009, and Krivega & Dean, 2012 for additional models of enhancer-promoter communication aided/prevented by chromatin loops and specific examples). Nevertheless, it is important to underline the fact that many functions attributed to CTCF can be unified now under the "weaver hypothesis".

1.3.2. RNA Polymerase Transcription Machinery

Transcription machinery has also been associated with the establishment of genomic boundaries (reviewed in Kirkland et al., 2012). The first evidence came from studies in *Saccharomyces cerevisiae* (**Fig. I 10**). The mating type of this haploid organism, which can be *a* or α, is determined by the composition of the active MAT locus. Throughout the life of the cell, the mating type can be switched by intrachromosomal gene conversion. This process involves the replacement of the genes in the active MAT locus for those in one of the inactive loci that harbor the mating types (HML for the *a* genes, and HMR for the α genes). The silencer elements that flank both HML and HMR loci induce their heterochromatinization and subsequent silencing. Barrier elements bracket these loci, preventing the encroachment of repressive chromatin into adjacent active domains.

Particularly, the right barrier of the HMR locus depends on a tRNA gene (Donze et al., 1999; Donze & Kamakaka, 2001). Other insulators that rely on this type of RNA polymerase III transcribed genes have been described in *Schizosaccharomyces pombe* (Amouyal, 2010a), mice (Lunyak et al., 2007; Ebersole et al., 2011) and humans (Raab et al., 2012). But not all tRNA genes function as boundaries. Apparently, tRNAs flanking sequences play an essential role in insulator function, possibly by helping to stabilize the polymerase III machinery at the promoter (Donze & Kamakaka, 2001).

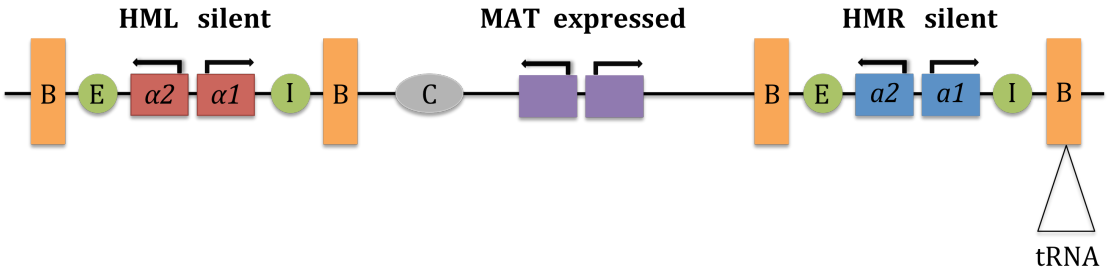


Fig. I 10. Genomic organization of *Saccharomyces cerevisiae* mating-type loci. Sex type is determined by the genetic composition of the active MAT locus. It can turn into *α* or *a* by intrachromosomal gene conversion with the silent HML or HMR genes, respectively. Both HML and HMR loci are bracketed by boundaries (B, orange rectangles) that prevent their inappropriate activation. A tRNA gene establishes the right boundary of the HMR locus. E, enhancer; I, silencer; C, centromere.

In yeast, transcription from the tRNA gene is not required for boundary activity. What is critical is the binding of TFIIIC to an intact B-box in the promoter (Donze & Kamakaka, 2001). In fact, orphan B-boxes that bind TFIIIC but are insufficient to promote transcription (Extra TFIIIC Sites or ETCs in *S. cerevisiae;* Chromosome Organizing Clumps or COCs in *S. pombe*), still function as barriers, especially if multimerized (Valenzuela et al., 2009). In humans, insulator activity of tRNA genes also depends on intact B-boxes. However, up to date, no assays have been carried out to determine if ETCs, also present in the human genome, function as barriers in this organism as well (Raab et al., 2012). In mice, synthetic multimerized B-boxes fail to establish effective barriers. Instead, promoter A-boxes seem to play a bigger part in the game (Ebersole et al., 2011).

Two hypotheses, not mutually exclusive, have been proposed to explain the mechanisms of action of these elements. The first one posits that tRNAs genes may act as passive barriers: the large size of the RNA polymerase III complex would create a physical blockage that prevents the extension of heterochromatin beyond the barrier ("roadblock model"). On the other hand, the most accepted hypothesis suggests that tRNA genes may establish active barriers through the recruitment of chromatin remodeling and histone modifying enzymes that counteract silencing activities (Donze & Kamakaka, 2001). In support of this view, it has been described that barrier activity in yeast depends first on

nucleosome eviction from the insulator mediated by the RSC protein so as to remove the template for heterochromatinization, and second, on histone acetyltransferases that place positive chromatin marks on the surrounding nucleosomes (Donze & Kamakaka, 2001; Valenzuela et al., 2009). The same model could be extended to mammals.

Furthermore, tRNAs convey enhancer-blocking activity as well. For example, in *S. cerevisiae*, tRNA genes and ETCs are able to block an "enhancer-like" element (UAS or Upstream Activation Sequence) from activating a reporter (Simms et al., 2008). Moreover, clusters, but not single copies, of human tRNAs also function as enhancer-blockers *in vitro* (Raab et al., 2012). The "roadblock model" could account for insulator activity at these loci. In both organisms, tRNAs and ETCs cluster in the nucleus through the establishment of long-range interactions (Thompson et al., 2003; Hiraga et al., 2012; Raab et al., 2012). As a result, chromatin loops are created. Enhancer-blocking activity could be explained if the enhancer and the promoter are trapped in different loops ("topological looping model").

Interestingly, unique boundary elements have been described in mice. At the murine growth hormone locus, a tRNA-derived SINE (Short Interspersed Nuclear Element) B2 retrotransposon, with highly conserved convergent RNA polymerase II and III promoters, controls the tissue- and development- specific expression of *Gh*. At the initial stages of development, while the RNA polymerase III machinery is transcribing the SINE B2, the locus is heterochromatic. Then, at embryonic stage 17.5, the SINE B2 element starts being bidirectionally transcribed by both RNA polymerases II and III, favoring the decondensation of the region, its relocalization to a more permissive euchromatic domain, and the subsequent activation of *Gh* transcription by tissue-specific transcription factors in the pituitary gland cells (Lunyak et al., 2007). It is still unclear how RNA polymerase II transcription can trigger these events. It has been suggested that the SINE B2 element could directly recruit histone modifiers to activate the locus. Alternatively, it could promote long-range interactions that prevent the propagation of silencing marks beyond the retrotransposon. In any case, there is still lack of experimental data to support any of these hypotheses (Lunyak, 2008). Of note, not only does this element act as a barrier to the spread of heterochromatin at the endogenous locus, but it also functions as a potent enhancer-blocker in ectopic constructs *in vitro*. This activity relies on intact RNA polymerase II and III promoters (Lunyak et al., 2007).

Another type of retrotransposable element that also meets the two properties of insulators is B1X35S. This is a special type of repetitive element from the SINE B1 family that can be found in the promoters of over 1,300 genes in the mouse genome. B1X35S harbors binding sites for the epithelial-mesenchymal transition regulator Slug/Snai2, and

the dioxin receptor AhR. This tRNA-derived retrotransposon contains functional RNA polymerase II and III promoters that, unlike the SINE B2 described above, transcribe the element in the same direction. In fact, their binding to B1X35S is mutually exclusive. Under basal conditions, transcription by RNA polymerase III normally takes place. However, upon binding of Snai2/Slug and AhR, a polymerase switch occurs: polymerase III is displaced by polymerase II, which starts transcribing B1X35S at a high rate. At the same time, PARylated CTCF is recruited, as well as histone remodellers that initiate the heterochromatinization, and subsequent inactivation, of the downstream genes (Roman et al., 2011a; Roman et al., 2011b). B1X35S greatly differs from other barriers in the sense that it favors the deposition of repressing, not activating, histone marks on the genes it regulates. This is the reason why it was initially described as a silencer (Roman et al., 2008). Hence, B1X35S is considered a barrier that prevents the propagation of the silencing cascade initiated by itself at endogenous loci. Moreover, it exhibits powerful enhancer-blocking activity in ectopic constructs both *in vitro* and *in vivo*, activity that depends on the binding of both Snai2/Slug and AhR (Roman et al., 2011a).

1.3.3. S/MARs

The nuclear matrix is the protein network that supports biological processes through the organization of the chromatin in the nuclear space. Thus, it can be considered the analogous to the cytoskeleton inside the nucleus, although its actual existence and true relevance have been put into question for years (Pederson, 2000). Meanwhile, Scaffold/Matrix Attachment Regions (S/MARs) are AT-rich stretches of DNA that remain bound to the nuclear matrix after removing histones and other proteins (halo-mapping, Mirkovitch et al., 1984). What is more, S/MARs can also reassociate *in vitro* with nuclear matrix preparations in which DNA had previously been removed (Cockerill & Garrard, 1986).

It is believed that S/MARs interact with the nuclear matrix establishing looped domains that favor the expression of the genes encapsulated in the loops (Gombert et al., 2003; Millot et al., 2003). A case example can be found in the mouse *Wap* locus. This gene, expressed exclusively in the mammary gland, is bracketed by two S/MARs that isolate the gene from its widely expressed neighbors *Cpr2* and *Ramp3* (Millot et al., 2003). Several studies have exploited this ability by flanking transgenic constructs by S/MARs in an attempt to protect them from chromosomal position effects, although with only mixed success. Indeed, some of these elements confer position-independent expression of the

transgenes they shield, but not all S/MARs are equally effective (i.e., McKnight et al., 1992; Poljak et al., 1994; Moreno et al., 2011; reviewed in Giraldo & Montoliu, 2001).

A series of experiments aimed at exploring the dynamic features of S/MARs revealed that not all are constantly tethered to the matrix. Instead, two types of elements can be differentiated: constitutive S/MAR are more or less immobile and have a structural role, whereas facultative elements are dynamic and only interact with the matrix upon cellular requirements (Heng et al., 2004). This may explain the differences in performance of the various elements assayed, and the fact they are only successful in stably integrated constructs and not in transient experiments (Phi-Van et al., 1990; Kalos & Fournier, 1995), since only those elements that are in fact associated with the matrix can prevent chromosomal position effects (Krnacik et al., 1995; Heng et al., 2004).

However, in some cases, the genomic boundary activity that had been ascribed to a given S/MAR was later found to be due to the presence of additional true boundaries that were being carried along with it (reviewed in West et al., 2002). For example, the chicken lysozyme 5' S/MAR (Stief et al., 1989) has successfully been used to render transgene expression independent of the integration site even if a species different to the chicken was employed (i.e., Phi-Van et al., 1990; Girod et al., 2005). Nevertheless, the specific subfragment of this S/MAR that binds to the nuclear matrix is not sufficient to provide insulation, which is rather conveyed by additional sequences located in the element (Phi-Van & Stratling, 1996).

The transcriptional augmentation phenomenon (reviewed in Bode et al., 2000) can also explain the beneficial effects the inclusion of S/MARs has on transgenic constructs. According to this hypothesis, these elements would not be neutral, as boundaries are supposed to be. Instead, they would facilitate gene expression, not by targeting transgenes to actively transcribed genomic loci (Goetze et al., 2005), but by exploiting several mechanisms, including the recruitment of the RNA polymerase II machinery through its interaction with the scaffold attachment factor B or SAF-B (Bode et al., 2000). Furthermore, very recent experiments have demonstrated that S/MARs attract histone modifying enzymes that deposit activating marks on the histones at either side of the element (Majocchi et al., 2014). This would be the reason why increases in the transcriptional rates of the genes surrounding a S/MAR occur bidirectionally (Poljak et al., 1994; Goetze et al., 2005).

1.4. Regulation of Insulator Activity

Some insulators are always operative, like those CTCF sites constitutively bound by the protein in all tissues and developmental stages (Martin et al., 2011). However, other insulators are not that static. Instead, they are switched on or off under different circumstances. For instance, on occasions, the expression profile of the genes in a given domain partially overlaps that of the genes in the adjacent domain. In these cases, the existence of a boundary is essential at those tissues/times in which the expression patterns differ, but is dispensable otherwise. This is the case of the SINE B2 element at the mouse *Gh* locus (Lunyak et al., 2007). Early in development, the *Gh* gene is embedded in a highly condensed heterochromatic region. Without further information, a snapshot of the region at this point would lead us to think that it consists of a single uniform domain. Then, at a certain developmental stage, a yet-to-be-described *stimulus* activates the bidirectional transcription of the SINE B2 retrotransposon upstream of *Gh* in the pituitary gland cells. It is at this precise moment when insulation mechanisms come into play, establishing a barrier that splits the region into two independent domains. The *Gh*-containing domain becomes euchromatic and active, while the other remains silenced.

The dynamic nature of some insulators is also reflected in the fact that the majority of barriers are cell-type specific (Cuddapah et al., 2009). This is not surprising given that transcriptional programs vary between cell types. Heart-specific genes surely locate in active euchromatin in the cells of the heart, whereas in the liver, they probably reside in silenced heterochromatin. Since euchromatic and heterochromatic regions are cell-type specific, so should barriers be because, by definition, barriers locate at the boundaries between these regions. This implicitly means that only a subset of all the elements with the potential to act as a barrier in a given cell is actually functioning as such. The rest stays turned off.

As expected, almost everything currently known about the regulation of boundary function comes from studies in CTCF-dependent insulators. The regulation mechanisms of this type of insulators can be classified into two broad groups. While some mechanisms affect CTCF binding to its target genomic sites, others impact directly on its ability to provide insulation at certain regions (Phillips & Corces, 2009).

The *methylation* of the CpG dinucleotides within the CTS is the most widely recognized method to prevent CTCF binding (Wang et al., 2012a). It is the mechanism used by the imprinting machinery to restrict the occupancy of CTSs in an allele-specific manner, as in the *Igf2/H19* locus (Bell & Felsenfeld, 2000; Hark et al., 2000; Kanduri et al., 2000). At

this site, an imprinting control region (or ICR) ensures the monoallelic and parent-of-origin-dependent expression of the genes it separates. CTCF only binds to the unmethylated ICR at the maternal allele. Here, it functions as an enhancer-blocker of the enhancer that maps downstream from *H19*: CTCF confines enhancer action to the maternally-expressed *H19* gene, while preventing its interaction with *Igf2*. In contrast, the methylation of the ICR at the paternal allele abolishes CTCF binding, allowing the enhancer to access and promote the expression of *Igf2* (**Fig. I 11**). However, the picture is much more complex than suggested by this simple description. It involves, for instance, CTCF-mediated long-range interactions between different regulatory elements in the locus (Kurukuti et al., 2006), as well as with other imprinting regions genome-wide (Zhao et al., 2006). In any case, all these interactions disappear upon methylation of the ICR due to the inability of CTCF to bind.

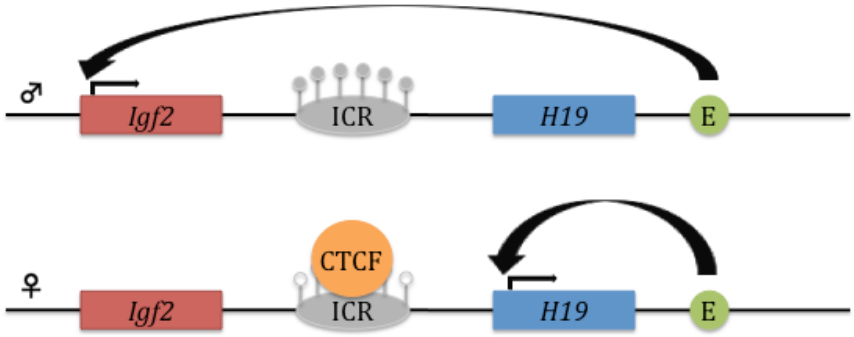


Fig. I 11. Methylation-dependent CTCF binding at its cognate CTS. Linear simplistic illustration of the imprinted mouse *Igf2/H19* domain. Methylation of the imprinting control region (ICR, grey) at the paternal allele prevents CTCF from binding and allows the enhancer (E, green circle) to activate *Igf2* expression. In contrast, CTCF associates with the unmethylated ICR at the maternal allele, restricting the action of the enhancer to *H19*.

The *positioning of nucleosomes* on the CTCF target sites is another means of preventing its binding. In fact, global analyses of CTCF occupancy in two human cell types revealed that those sites that were occupied only in one of them were actually invisible to CTCF in the other cell type because they were blocked by nucleosomes (Cuddapah et al., 2009). At a more local scale, specific examples in which nucleosome repositioning regulates CTCF binding include the chicken lysozyme (Lefevre et al., 2008) and the mouse *Igf2/H19* genes (Kanduri et al., 2002).

On the other hand, it has been shown that CTCF can carry a number of *post-transcriptional modifications* that affect its functions. For example, poly(ADP-ribosyl)ation has proven essential for CTCF-dependent insulator activity (Yu et al., 2004; Farrar et al., 2010). Perhaps, the inability of CTCF to translocate to the nucleolus unless PARylated lies behind this observation (Torrano et al., 2006). Furthermore, SUMOylation supports CTCF

repressive activity on c-*myc*, possibly by favoring the relocalization of the gene to Polycomb bodies (MacPherson et al., 2009). In contrast, phosphorylation by casein kinase II enhances CTCF transactivation potential (Klenova et al., 2001; El-Kady & Klenova, 2005). A direct link between the last two modifications and boundary activity has not been established yet. However, they could be indirectly regulating insulation by dictating which *protein partner* can associate with CTCF at each location.

Interestingly, around 18% of all human CTS harbors Thyroid hormone Response Elements or TREs, regulatory regions bound by the Thyroid hormone Receptor TR. Some of such composite sites possess enhancer-blocking activity, particularly those found in the chicken lysozyme and human c-*MYC* and *ESRRA* genes. In all three cases, insulator activity is relieved in the presence of the TR ligand, which is the T3 hormone (Lutz et al., 2003; Weth et al., 2010). Hence, *hormonal stimuli* can also modulate CTCF function.

1.5. Why Should Boundaries Be Studied?

Boundary elements are remarkable regulatory sequences that deserve further investigation for various reasons. From the point of view of basic research, the study of insulators would serve to shed more light into the organization of the eukaryotic genome inside the nucleus, which is currently regarded as an exciting new layer of gene expression regulation.

Furthermore, from a practical angle, insulators can be employed to improve the gene transfer technologies and their applications (reviewed in Recillas-Targa et al., 2004). One of these applications is gene therapy, a technique aimed at correcting a disease-causing genetic defect. Numerous studies have confirmed the beneficial effects the inclusion of insulators in viral-based vectors exerts on gene therapy (reviewed in Emery, 2011). Importantly, they alleviate two of the major challenges this technology has to face. On the one hand, insulators protect the exogenous constructs from *chromosomal position effects* that usually lead to transgene silencing. On the other hand, boundaries also impede the undesirable influence of the regulatory elements placed within the construct on the endogenous sequences at the integration site. This phenomenon, known as *vector-mediated genotoxicity*, can result, for instance, in malignant cellular transformation through the activation of proto-oncogenes by the exogenous enhancers. However, even if helpful, the elements tested to date fail to completely abrogate these security problems, so the perfect insulator is yet to be found (Molto et al., 2011).

Other outstanding applications in which insulators can make a positive contribution by preventing chromosomal position effects include the generation of transgenic animals (Giraldo et al., 2003b) as biorreactors for the production of recombinant proteins (reviewed in Houdebine, 2000) like human lactoferrin (Cheng et al., 2012), as models for the study of various genetic conditions (i.e., Brenden et al., 2013), or simply as tools to analyze the function and properties of any transgenic piece of DNA of interest in an *in vivo* setting (i.e., Yu et al., 2012). Furthermore, the efforts of many scientists focus on the development of ready-to-use genetic tools with diverse functional elements including insulators; tools that can be used by others to generate their own transgenic animals for any desired application (i.e., Bessa et al., 2009).

1.6. How Are Boundaries Studied?

The search for boundaries follows the typical guidelines commonly used to discover new regulatory elements of any kind (Narlikar & Ovcharenko, 2009; Hardison & Taylor, 2012). Basically, bioinformatic predictions and/or high throughput experimental approaches like ChIP-seq are first applied to highlight potential insulator sites in the genome. However, they can only be considered true insulators after thorough validation under functional assays specifically designed to test for boundary activity.

1.6.1. Looking for Boundaries Genome-Wide

Boundaries are functionally defined, not structurally; that is, the only thing they share is their function. To date, there has not been described a unique DNA motif that can be used to readily identify all of them with a bioinformatic approach. Nor is there yet any antibody that can be employed to pull down all DNA-bound proteins involved in insulation.

In fact, many of the boundaries currently known were not discovered on purpose. Instead, in many instances, they were uncovered by scientists who, upon studying their favorite gene, came across unconventional regulatory modules that functioned as insulators at those particular loci (recent compilations of insulators can be found in Molto et al., 2009 and Herold et al., 2012, as well as in **Table I 1** and **Appendix I 1** in this PhD thesis).

Table I 1. Comprehensive list of insulators described in the mouse genome as in February, 2014. See **Appendix I 1** for more details.

Insulator	Chromosome	5' Gene	3' Gene	References
adHS1	2	*Nr5a1*	*Nr6a1*	Ishihara & Morohashi, 2005
Evx2/Hoxd	2	*Evx2*	*Hoxd13*	Kmita et al., 2002 Yamagishi et al., 2007
Scl/Map17	4	*Pdzk1ip1*	*Cyp4x1*	Follows et al., 2012
Tcrβ/Trypsinogen	6	*Tcrβ*	*Prss2*	Carabana et al., 2011
5'Tyr	7	*Tyr*	*Grm5*	Giraldo et al., 2003a
β-globin 5'HS5	7	*Olfr66*	*Hbb-y*	Farrell et al., 2002 Tolhuis et al., 2002
β-globin 3'HS1	7	*Hbb-b2*	*Olfr68*	Bulger et al., 2003 Bender et al., 2006
PCT12	7	*Mrpl23*	*Nctc1*	Ishihara & Sasaki, 2002
MS/DMD	7	*H19*	*Igf2*	Bell & Felsenfeld, 2000 Hark et al., 2000 Kanduri et al., 2000 Ideraabdullah et al., 2008 Ideraabdullah et al., 2011
KvDMR1	7	*Kcnq1*	-	Kanduri et al., 2002 Fitzpatrick et al., 2007
SP-10	9	*A630095E13Rik*	*Acrv1*	Reddi et al., 2003 Acharya et al., 2006 Abhyankar et al., 2007
Rasgrf1 DMD	9	-	*Rasgrf1*	Yoon et al., 2005
5'Wap	11	*Tbrg4*	*Wap*	Millot et al., 2003
3'Wap	11	*Wap*	*Ramp3*	Montazer-Torbati et al., 2008
Gh	11	*Cd79b*	*Scn4a*	Lunyak et al., 2007
V(D)J rec	12	*Igh*	*hole*	Garrett et al., 2005 Featherstone et al., 2010 Shih & Krangel, 2013
Tcrα/Dad1	14	*Tcrα*	*Dad1*	Zhong & Krangel, 1999 Ortiz et al., 2001 Magdinier et al., 2004 Gomos-Klein et al., 2007
11P	17	*Rxrb*	*Col11a2*	Murai et al., 2008
Eif2s3x sites	X	*Eif2s3x*	*Klhl15*	Filippova et al., 2005
RS14	X	*Xist*	*Tsix*	Spencer et al., 2011
Xist/Tsix	X	*Tsix*	-	Chao et al., 2002
Jarid1c sites	X	*Kiaa0522*	*Jarid1c*	Filippova et al., 2005
B1X35S	-	-	-	Roman et al., 2011a
CONSYN CTCF	-	-	-	Martin et al., 2011
Hox clusters	-	-	-	Srivastava et al., 2013

Still, some attempts have been made to search for boundaries in a genome-wide fashion, but they all revolve around CTCF, either using bioinformatics or ChIP-based experimental approaches (Barski et al., 2007; Kim et al., 2007; Xie et al., 2007; Jothi et al., 2008; Cuddapah et al., 2009; Chen et al., 2012; Nakahashi et al., 2013). Nevertheless, these assays carry a number of problems. Firstly, bioinformatic assays aiming to find all CTCF binding consensus motifs in the genome assume that the mere presence of the protein target site unmistakably correlates with actual CTCF binding, which is certainly not the case (Kim et al., 2007; Chen et al., 2012). Secondly, although ChIP-related experiments do confirm CTCF-DNA associations, the true function of each binding instance is not directly addressed and hence, it would be an error to conclude that all of them are involved in insulation. For example, Dekker and coworkers found that 79% of long-range interactions between distal regulatory elements and promoters bypass CTCF-bound sites (Sanyal et al., 2012), which clearly indicates that, at the endogenous loci, only a small fraction of these sites function as classical enhancer-blockers. More importantly, these approaches focus only on CTCF-dependent insulators, but many more types exist (Molto et al., 2009). In fact, if only CTCF was exploited to flank and define expression domains, then interactions between the enhancers and promoters that reside in the same domain would be expected. However, 38% of such enhancer/promoter pairs do not communicate (Shen et al., 2012), which further supports the existence of additional CTCF-independent insulator mechanisms within these domains, as it has been published (Millot et al., 2003; Lunyak et al., 2007; Roman et al., 2011; Tiana et al., 2012).

Even if insufficient for the study of insulators as a whole, these genome-wide assays have provided a large amount of data about CTCF binding sites and their genomic context, data that are stored under freely available databases for public consultation (CTCFBSDB 2.0; Ziebarth et al., 2013). In addition, they have greatly contributed to deepen our understanding on CTCF properties and behavior, as well as the characteristics of its binding sites. For example, CTSs, which often appear in clusters (Jothi et al., 2008; Chen et al., 2012), can usually be found in high-gene-density regions (Kim et al., 2007; Xie et al., 2007; Chen et al., 2012), unless they contain clusters of coexpressed genes (Kim et al., 2007). In contrast, CTCF binding sites abound within genes with many alternative promoters (Kim et al., 2007). Regarding the genomic distribution, it appears that 45-50% of CTSs are intergenic, and 35-42%, intragenic, while the rest map within promoters (Kim et al., 2007; Jothi et al., 2008; Chen et al., 2012). Finally, several histone modifications are significantly enriched in the nucleosomes surrounding CTSs, although they depend on whether the CTSs are bound by the protein in a constitutive or cell-type specific manner

(Chen et al., 2012). In any case, the H2A.Z histone variant associates with both types of CTCF-bound sites (Barski et al., 2007; Chen et al., 2012).

1.6.2. Functionally Testing Boundaries

Predictions do not always correspond to reality. A DNA sequence, even if predicted to be an insulator, cannot be considered as such unless tested with specific assays designed for that purpose (Molto et al., 2009; Molto et al., 2011; Herold et al., 2012). Usually, these assays seek to determine if a given element possesses any of the two properties classically assigned to insulators: barrier and enhancer-blocking.

Barrier assays evaluate the capacity of a putative boundary to protect a transgenic construct, when flanking it, from chromosomal position effects upon random integration into the genome (**Fig. I 2**; Barkess & West, 2012). These assays can be conducted *in vitro* using mammalian cell lines (i.e., Pikaart et al., 1998; Kim et al., 2009; Raab et al., 2012), or *in vivo* by generating transgenic flies or mice (i.e., Giraldo et al., 2003a; Furlan-Magaril et al., 2011). The precursor of this type of barrier assays can be traced back to the experiment developed by Kellum and Schedl to analyse the *scs* and *scs'* insulators, in which they created transgenic flies with different versions of the *white* gene (Kellum & Schedl, 1991).

In principle, integration into the genome is essential for testing the ability of a sequence to prevent the spread of heterochromatin into the ectopic construct. However, several groups have overcome the necessity of genomic integration by including in their transgenes DNA sequences that attract the silencing machinery to induce the rapid heterochromatinization of their episomal construct (Van der Vlag et al., 2000; Raab et al., 2012). Although faster than conventional assays, these tests only assess the ability of insulators to block negative, but not positive, influences from the surroundings.

Enhancer-blocking assays, as their name states, evaluate if the putative boundary element is able to block the positive effects of a distal enhancer on the expression of a reporter when placed between the two (**Fig. I 2**). Silencers would behave alike when cloned at this same position. Therefore, the element under study is also cloned outside the enhancer-gene unit in order to make sure that it is truly functioning as an insulator, and not as a silencer. At this new location, an insulator would be of no use and the enhancer would boost the reporter gene expression. In contrast, a silencer would eclipse the enhancer irrespective of its position, maintaining low levels of reporter expression. The

enhancer-blocking properties of a given element can be assessed *in vitro* by transfecting different cell lines (Chung et al., 1993; Li & Stamatoyannopoulos, 1994; Kanduri et al., 2000; Lunyak et al., 2007), or *in vivo* through the generation of transgenic flies (Moon et al., 2005) or zebrafish (Bessa et al., 2009). Even large-scale assays in which many putative insulators are tested at the same time *in vitro* have been designed (Mukhopadhyay et al., 2004). Again, all variants of this strategy derive from the original assay conceived by Kellum and Schedl, who evaluated the ability of various sequences to block the *yp-1* enhancer from activating a β-galactosidase reporter in transgenic flies (Kellum & Schedl, 1992).

Of note, both barrier and enhancer-blocking assays utilize ectopic constructs, often in heterologous systems, so care should be taken when interpreting the results: the behavior of a given element in one of these assays does not necessarily correlate with its behavior *in vivo* in its native genomic context (Molto et al., 2009; Phillips & Corces, 2009; Splinter & De Laat, 2011; Barkess & West, 2012).

1.7. Where Would Boundaries Be Expected?

As already stated, most strategies aimed at finding boundaries rely on a genome-wide search of CTCF binding sites. However, the presence of such a sequence at a given locus does not guarantee its functioning as a boundary. Also, boundaries with unknown mechanisms of action are always missed in these sequence-driven approaches. Instead, the strategies that will be addressed here are functionally-driven. Where in the genome would the cell need a boundary to organize the different expression domains? At least three different scenarios arise (**Fig. I 12**; Bell et al., 2001).

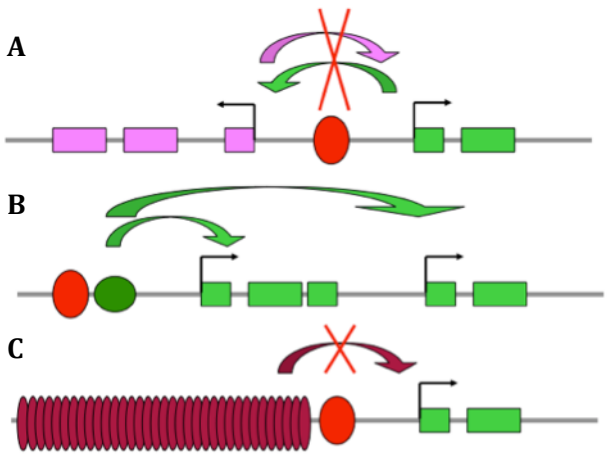


Fig. I 12. Examples of genomic loci that potentially contain boundary elements. **A.** Boundaries could be found separating genes with different expression patterns, as in the mouse *Wap* locus. This gene is only expressed in the mammary gland, whereas its flanking partners, *Cpr2* and *Ramp3* are widely expressed (Millot et al., 2003). **B.** Clusters of co-expressed genes, like the chicken β-globins, should be enclosed and protected by boundaries (Farrell et al., 2002). **C.** Boundaries could also prevent the spread of silencing heterochromatin into an active locus, as in the case of the mouse *Tyr* gene, which needs to be protected from a highly condensed region further upstream (Giraldo et al., 2003a).

First, a boundary could lie between genes whose expression patterns differ in time and/or space in order to maintain independent expression units. There, it would prevent undesirable crosstalk between the regulatory elements that belong to each expression domain. If this hypothesis is true, it would explain why, for instance, ubiquitously expressed genes neighbor genes with more restricted expression patterns without interfering with them (**Fig. I 12A**).

Second, a number of clusters of co-expressed genes exist in the mammalian genomes. Several lines of evidence indicate that such clusters are conserved between species, that the recombination rate within the clusters stays low, and that the expression of the clustered genes co-evolve. These observations suggest that clusters appeared and were fixed throughout evolution because they were convenient for the cells (for review, see Elizondo et al., 2009). Boundaries may flank these clusters of co-expressed genes to protect them from the surroundings and ensure their correct expression patterns, facilitating natural selection (**Fig. I 12B**).

Finally, boundaries could be found at the borders between active euchromatin and inactive heterochromatin. In this case, they would prevent the spreading of silencing signals into loci that need to be active at a certain tissue or developmental stage for the correct functioning of the cell; and *vice versa* (**Fig. I 12C**).

The present PhD thesis describes the development and functional validation of algorithms to detect boundaries genome-wide in loci corresponding to the first two scenarios depicted, as well as the functional validation of insulator sequences derived from an algorithm based on the third scenario but developed elsewhere (see the Materials and Methods section).

2 OBJECTIVES

Several attempts have been made in the past to identify genomic boundaries in mammals. However, they focused solely on the presence of binding sites for the insulator-related CTCF protein, while disregarding the existence of other mechanisms of insulation.

The **main aim** of this PhD thesis was to predict and functionally validate new genomic boundaries in mammals in a genome-wide unbiased manner. To achieve this goal, the following specific objectives were addressed:

1. To develop bioinformatic algorithms to predict boundaries by looking for those genomic loci where cells would require the presence of boundary activity to organize their genomes, namely:
    a. Separating genes with opposite expression patterns, for which gene expression data had to be employed.
    b. Flanking clusters of co-expressed genes, identified by the analysis of gene expression data.
    c. Partitioning the chromatin into active euchromatic and silenced heterochromatic regions, defined by the accumulation of specific chromatin marks.
2. To functionally validate the predicted boundaries with experimental assays specifically designed to test each of the properties that characterize them:
    a. Enhancer-blocking activity by using *in vitro* and *in vivo* assays in human embryonic kidney 293 cells, and transgenic zebrafish, respectively.
    b. Barrier activity or the ability to protect against chromosomal position effects in transgenic mice.
    c. The establishment of long-range interactions by CTCF-dependent potential boundaries by utilizing 3C technology.

3 MATERIALS & METHODS

3.1. Development and *in Silico* Validation of Algorithms to Identify Insulators in the Mouse Genome

3.1.1. Algorithms to Predict the Presence of Boundaries Separating Genes with Different Expression Patterns

3.1.1.1. Genomic and Gene Expression Data Extraction

Gene data information for the murine genome (NCBI mouse assembly 37) were downloaded from Ensembl using the BioMart tool (annotation release 59). Only annotated genes with MGI IDs were considered in this study. Genes were ordered according to their position in the genome, and pairs of adjacent genes were formed. For overlapping genes, multiple pairs were created: a given gene was paired with every overlapping gene, plus the first non-overlapping one.

Next, the web-based tool aGEM v2.0 (Jimenez-Lozano et al., 2009) was used to extract expression data at the adult stage (TS28) for the selected genes. These data served to generate two different algorithms, namely the correlation method and the Euclidean distance method.

3.1.1.2. Correlation Method

The correlation method was already implemented in the aGEM Platform. Briefly, fixing the adult stage (TS28), the Pearson's correlation coefficient was calculated for the expression profiles of a given pair of genes, for all pairs created:

$$r(A,B) = \frac{\sum_{i=1}^{n}((A_i - \bar{x}_A)\cdot(B_i - \bar{x}_B))}{\sqrt{\sum_{i=1}^{n}(A_i - \bar{x}_A)^2 \cdot \sum_{i=1}^{n}(B_i - \bar{x}_B)^2}}$$

where r is the Pearson's correlation coefficient between genes A and B, n is the number of anatomical structures included in the study, A_i and B_i are the expression values for the genes A and B in a given anatomical structure, and $\bar{x}$ represents the mean expression values of genes A and B across all anatomical structures analyzed.

Pairs of significantly negatively correlated genes (correlation value < 0; p-value < 0.05), which potentially contain insulator elements, were selected for further study. Of note, the correlation analysis was only performed when there were gene expression data for a minimum number of common anatomical structures for both genes, specifically, 52.

3.1.1.3. Euclidean Distance Method

For the Euclidean distance method, gene expression data were newly codified as follows: 'zero' for unexpressed genes in a given anatomical structure (equivalent to the 'zero' level in aGEM) and 'one' for expressed ones (as a combination of the 'one' and 'two' levels). In this method, not every anatomical structure was considered; only those with expression information for more than 60% of the genes of each chromosome were taken into account. This implies that, for a given tissue, there was still lack of expression data for some genes (a maximum of 40% of all genes). To solve this problem, missing data points for a gene in a given tissue were replaced with the average gene expression value of the genes flanking it. The Euclidean gene expression distance between two genes was then calculated according to the following formula:

$$D(A,B) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}$$

where D is the gene expression distance between genes A and B, n is the number of anatomical structures included in the study, and A_i and B_i are the expression values for the genes A and B in a given anatomical structure.

An example of such calculations can be found in the following table:

		Expression		
		Gene A	Gene B	$(A_i - B_i)^2$
Tissue	Adipose tissue	1	1	0
	Bone marrow	0	0	0
	Cerebellum	0	0	0
	Epidermis	0	1	1
	Heart	1	1	0
	Hippocampus	1	1	0
	Kidney	0	1	1
	Liver	1	0	1
	Mammary gland	0	0	0
	Ovary	0	0	0
	Pancreas	1	1	0
	Spleen	0	1	1
	Testis	1	0	1
	Vomeronasal organ	1	0	1
			$\sum_{i=1}^{16}(A_i - B_i)^2$	6
			$D(A,B) = \sqrt{\sum_{i=1}^{16}(A_i - B_i)^2}$	**2.45**

Then, a set of distance matrices for all genes of the genome, chromosome by chromosome, was generated: genes were arranged in rows and columns in such a way that the values that form the diagonal were equal to zero, since the distance of a gene from itself is null. In these matrices, each row (or column) represented the distribution of the distances between a gene and all the other genes in the same chromosome. Hence, there were as many distance distributions as genes. The mean and standard deviations of all distributions were calculated. The distance coefficient between two genes, A and B, was considered to be significantly high if it was larger than the mean plus twice the standard error of the mean (SEM) of the distance distributions for both A and B:

$$D > 2 \cdot \frac{s_A}{\sqrt{n}} + \bar{x}_A \qquad \text{and} \qquad D > 2 \cdot \frac{s_B}{\sqrt{n}} + \bar{x}_B \quad ,$$

where D is the gene expression distance between genes A and B, the s and $\bar{x}$ values refer to the standard deviation and mean of the distance distributions of genes A and B for a given chromosome, and n is the number of genes in said chromosome.

Again, only pairs of genes with a distance coefficient significantly high for both members of the pair were considered for further analysis.

The next table collects the three possible scenarios that can arise upon evaluating the statistical significance of the Euclidean distance between the expression profiles of a gene pair across a panel of fourteen tissues (n=14). Only the third gene pair would be retrieved by the algorithm as a hit:

	Gene	$\bar{x}$	s	$2 \cdot \frac{s_{A/B}}{\sqrt{n}} + \bar{x}_{A/B}$	Euclidean Distance	Significantly High Distance for any of the Genes?	Significantly High Distance for the Gene Pair?
Case 1	A	5.1	0.6	5.42	4.1	No	No
	B	3.2	1.0	3.73		Yes	
Case 2	A	4.7	0.8	5.13	3.9	No	No
	B	5.3	1.1	5.89		No	
Case 3	A	3.8	0.7	4.17	5.8	Yes	Yes
	B	4.4	0.3	4.56		Yes	

The distance matrices were generated using the SAS software (*Statistical Analysis System*), whereas the statistical calculations were conducted in IBM SPSS Statistics v19.

This method was developed with the aid of Laura Barrios (*Secretaría General Adjunta de Informática* SGAI, *Consejo Superior de Investigaciones Científicas* CSIC).

3.1.1.4. *In Silico* Validation and Comparison of the Methods

The lists of pairs of genes with different expression patterns obtained from the correlation and Euclidean distance methods were scanned for the presence of previously known insulator elements in order to validate the algorithms (Roman et al., 2011a; Martin et al., 2011; Molto et al., 2009). Enrichment tests were carried out by using the Fisher's exact test function implemented in R (`fisher.test(argument)`).

Also, both lists were compared, and the Venn Diagram Plotter program was used to illustrate the coincidences and discrepancies between the methods.

3.1.1.5. Criteria for the Selection of Sequences to Test for Boundary Activity

The application of functional annotation tools and diverse databases enabled the selection of genes with biologically relevant roles for the cell. In fact, malfunction of some of the genes chosen results in a pathogenic condition in mice and/or their human orthologs. The tools and databases consulted include:

- Functional annotation tools:
    - DAVID Functional Annotation: http://david.abcc.ncifcrf.gov (Huang et al., 2009a; Huang et al., 2009b)
    - BABELOMICS suite (version 4.2), especially FatiGO and Genecodis: http://babelomics.bioinfo.cipf.es/functional.html (Medina et al., 2010)
- Databases:
    - UniProtKB: http://www.uniprot.org/uniprot (The Uniprot Consortium, 2012)
    - Online Mendelian Inheritance in Man (OMIM®): http://www.ncbi.nlm.nih.gov/omim
    - PubMed: http://www.ncbi.nlm.nih.gov/pubmed

Once a pair of genes had been chosen, additional tools facilitated the selection of specific sequences within the locus to test for boundary activity: evolutionary conservation and functional annotation tools.

- Evolutionary conservation analysis:
  - VISTA browser: http://pipeline.lbl.gov/cgi-bin/gateway2 (Frazer et al., 2004)
  - ECR browser: http://ecrbrowser.dcode.org (Ovcharenko et al., 2004)
- Functional annotation of specific DNA sequences:
  - Regulation tracks at genome browsers (i.e. ChIP-seq results for CTCF binding sites in a given cellular type):
    - UCSC: http://genome.ucsc.edu (Kuhn et al., 2013)
    - Ensembl: http://www.ensembl.org (Flicek et al., 2012)

3.1.2. Algorithm to Detect Boundaries Flanking Clusters of Co-Expressed Genes

As a proof-of-concept, only chromosome 18 was examined for the presence of clusters of co-expressed genes, either adjacent or distant in the linear genome.

3.1.2.1. Defining Clusters of Adjacent Co-Expressed Genes

In the Euclidean distance matrices calculated above (section 3.1.1.3), each row (or column) contained the distribution of expression distances between a given gene and the rest of the genes in the same chromosome. Each one of these distributions was unique, defined by specific statistical parameters. In this context, distance values were not absolute; they only acquired meaning when the distribution to which they pertained was considered. Hence, a normalization step was required to enable the direct comparison of the distance values compiled in a matrix.

Euclidean distance values for the genes in chromosome 18 were therefore normalized with respect to the maximum distance value that could be found in the distance distributions of a given pair of genes, A and B. Normalization was conducted taking into account that the normalized Euclidean distance matrix had to be symmetrical:

$$D(A,B) = D(B,A) \quad ,$$

where D is the gene expression distance between genes A and B. That is, the expression distance between gene A and gene B had to be the same as the expression distance between gene B and gene A. One pair of genes was considered at a time; its distance value was divided by the maximum distance value observed in either of the two distributions. For example, the distance matrix for any four genes A, B, C and D can be:

Distance (D)	A	B	C	D
A	0	1	2	3
B	1	0	1	2
C	2	1	0	4
D	3	2	4	0

- Step 1: For genes A and B:

$$Max(A,B) = [MaxD(A/B,C,D); MaxD(B/A,C,D)] \quad ,$$

The distribution of the distances between gene A and the rest of the genes is analyzed in order to find the maximum value *(MaxD(A/B, C, D)*, which is 3. The same is done with the distribution of gene B, *(MaxD(B/A, C, D)*, which takes the value of 2. The larger maximum distance value, which is 3 in this case, is set as the maximum *Max(A,B)*. Then:

$$D_n(A,B) = D_n(B,A) = \frac{D(A,B)}{Max(A,B)} = \frac{D(B,A)}{Max(A,B)} = \frac{1}{3} = 0.33 \quad ,$$

where *D* and D_n are the distance and normalized distance values between genes *A* and *B*, respectively, and *Max(A,B)* is the maximum distance value observed in the distributions of both A and B.

- Step 2: For the rest of the genes:

Step 1 is repeated with all possible combinations between the four genes considered in the example: A and C, A and D, B and C, B and D, C and D.

- Step 3: Matrix creation:

Distance (*D*)	A	B	C	D
A	0	1	2	3
B	1	0	1	2
C	2	1	0	4
D	3	2	4	0

Normalized distance (D_n)	A	B	C	D
A	0	0.33	0.50	0.75
B	0.33	0	0.25	0.50
C	0.50	0.25	0	1
D	0.75	0.50	1	0

Note that the distance value between A and B is the same as between B and C. However, the normalized distance values differ.

To ease the visual interpretation of the results, the normalized Euclidean distance matrix for chromosome 18 was converted into a heat map in Excel, in such a way that high distance values appeared in green, whereas low values were colored in red.

Clusters of adjacent co-expressed genes were found by scrutinizing the normalized matrix and selecting groups of at least 5 consecutive genes with expression distance values below 0.45. BABELOMICs and DAVID functional annotations tools, as well as TRANSFAC® database (http://www.biobase-international.com/product/transcription-factor-binding-sites; Matys et al., 2006) were employed to extract information about possible biological pathways or transcription factor binding sites shared by the members of the same cluster. In addition, Ensembl regulation tracks provided experimental data on the protein factors that bind specific genomic loci.

3.1.2.2. Defining Clusters of Co-Expressed Genes that Lie Far Away from Each Other in the Linear Genome

The genes in chromosome 18 were divided into clusters according to their expression profiles with the *K*-means non-hierarchical clustering algorithm in IBM SPSS Statistics v19. Several trials were made before choosing 30 as the appropriate number of clusters to generate (*K*=30, 15 iterations): a lower *K* yielded clusters with too many components, and *vice versa*. Then, each cluster was analyzed with BABELOMICs and DAVID functional annotation tools, as well as with TRANSFAC®, in an attempt to find common biological functions that could explain why the genes in each cluster display very similar expression profiles.

Finally, COXPRESdb v5.0 (http://coxpresdb.jp/; Obayashi et al., 2013), an online tool that provides networks of co-regulated genes, was used to confirm the stated findings independently.

3.2. Cloning Vectors

3.2.1. Original Vectors

3.2.1.1. *In Vitro* Enhancer-Blocking Assays in HEK 293 Cells

All plasmids used in the *in vitro* enhancer-blocking assays are based on pLuc (4.9 kb) and pELuc (5.5 kb, **Fig. MM 1**). These vectors were kindly provided by Dr. Satoshi

Watanabe (Department of Developmental Biology, National Institute of Agrobiological Sciences, Tsukuba, Japan) (Watanabe et al., 2006). pLuc contains the firefly luciferase reporter gene under the control of human cytomegalovirus minimal promoter (CMV-mP). pELuc included, in addition, the CMV enhancer (CMV-IE). The poly(A) signal of both vectors comes from the SV40 virus.

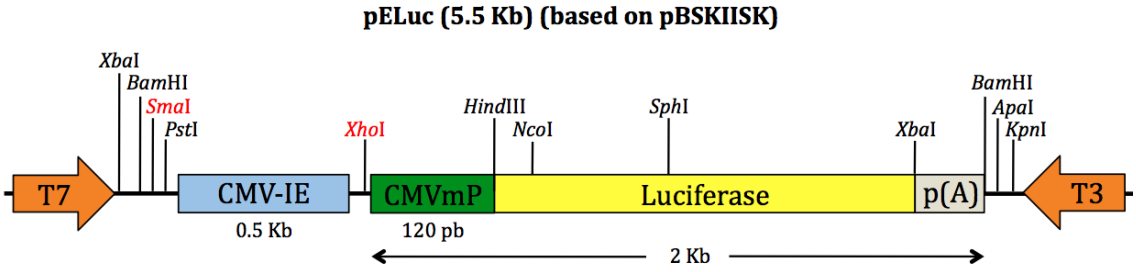


Fig. MM 1. The pELuc plasmid served as the backbone for the *in vitro* enhancer-blocking assays. True insulators should block the influence of the CMV enhancer on the promoter when cloned in between them (at the *Xho*I site), but never at a distal position (*Sma*I site).

A number of control elements were included to monitor the assays: the 1.2-kb insulator from the chicken β-globin locus (cHS4) and its core II/III element (0.2 kb) were employed as positive controls, whereas a mutated version of the CTCF binding site in II/III constituted the negative control II/III Mut (0.2 kb) (Bell et al., 1999; Chung et al., 1993; Recillas-Targa et al., 1999). These elements had been previously cloned in our laboratory, upstream (cHS4E, II/IIIE and II/III Mut E) and downstream (EcHS4, EII/III, EII/III Mut) from the enhancer in pELuc (Lunyak et al., 2007).

In addition, pCMV-lacZ (7.2 kb) was used as a transfection control for normalization purposes (MacGregor & Caskey, 1989).

3.2.1.2. *In Vivo* Enhancer-Blocking Assays in *Danio rerio* (Zebrafish)

For the *in vivo* evaluation of their enhancer-blocking activity, the selected sequences were cloned in the Tol2 transposon-based pminiTol2-Z48-CARGFP vector –p48RCAR hereafter- (5.9 kb, **Fig. MM 2A**). In this vector, the cardiac actin promoter (CAR) from *Xenopus laevis* (Mohun et al., 1986) and a central nervous system enhancer (Z48, also known as Z54390) from *Danio rerio* (De la Calle-Mustienes et al., 2005) drive the expression of a GFP reporter cassette. It also contains the SV40 poly(A) signal. Additionally, p48RCAR had been adapted to the Gateway® Cloning Technology by introducing the LR recombination cassette between the enhancer and the promoter in the direct (p48RCAR GW') or inverse (p48RCAR GW) orientations (7.6 kb, **Fig. MM 2B**).

All of these vectors were a kind gift from Dr. José Luís Gómez Skarmeta (*Centro Andaluz de Biología del Desarrollo*, CABD, Seville, Spain) (Bessa et al., 2009).

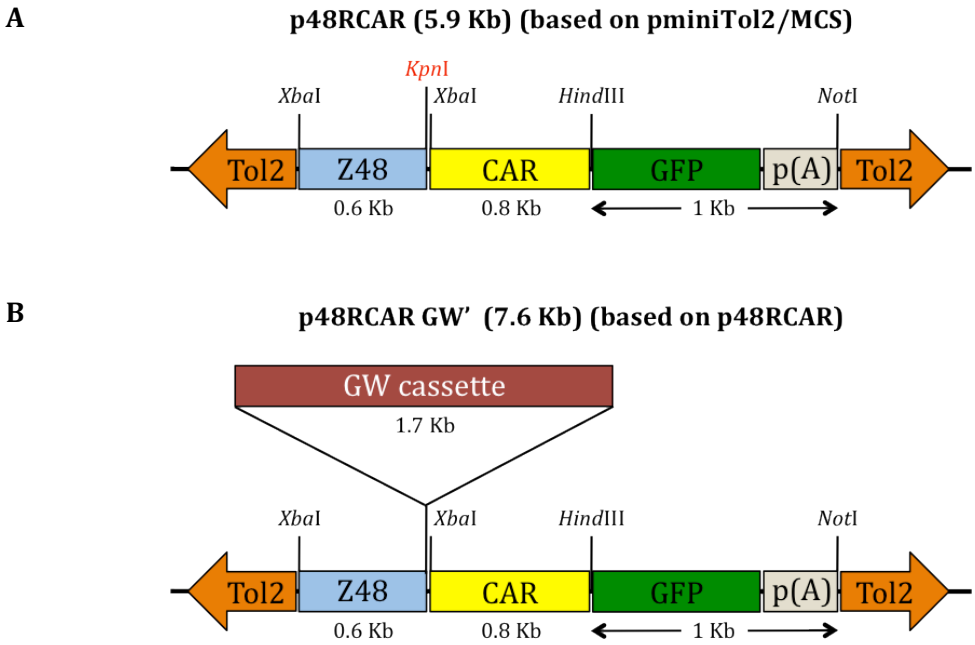


Fig. MM 2. Tol2 transposon-based vectors used in *in vivo* enhancer-blocking assays. A. p48RCAR contains a GFP reporter gene under the control of a central nervous system enhancer (Z48) and a cardiac actin promoter (CAR). These regulatory elements direct GFP expression to the midbrain, and to the heart and somites of transgenic zebrafish embryos, respectively. When an insulator is cloned in the *Kpn*I site (in red), the influence of the enhancer on the reporter is blocked. As a consequence, GFP expression becomes restricted to the somites and heart. **B.** The Gateway® LR recombination cassette had been blunt-cloned in the *Kpn*I site of p48RCAR in both orientations to accelerate the cloning procedure of putative insulator elements. Only the direct orientation, corresponding to p48RCAR GW', is shown.

3.2.1.3. Protection Against Chromosomal Position Effects Assay in Mice

For this experiment, ptrTYR5 (14.1 kb, **Fig. MM 3**) was employed. This plasmid harbors a tyrosinase minigene and was kindly provided by Dr. Schütz (Institute of Cell and Tumor Biology, German Cancer Research Center, Heidelberg, Germany) (Beermann et al., 1991).

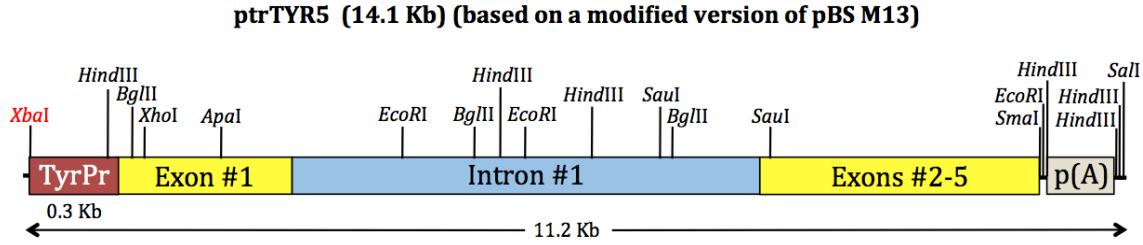


Fig. MM 3. ptrTYR5 plasmid served to test the ability of a DNA sequence to protect from chromosomal position effects in mice. It consists of 270 bp of the 5' flanking sequence (TyrPr, red box downstream from the *Xba*I site), the first exon (first yellow box) and intron (blue box), and a fusion of exons 2 to 5 (second yellow box) of mouse tyrosinase. SV40 splice and polyadenylation signals are also shown (grey box) (Beermann et al., 1991). The elements to test for boundary activity are cloned at the *Xba*I site (in red).

3.2.2. Plasmid Construction

Plasmids were generated either by classical restriction-ligation cloning procedures or by using the Gateway recombination technology, and confirmed by restriction enzyme digestion and sequencing analyses.

3.2.2.1. Bioinformatic Tools

Primers for PCR reactions were designed and checked *in silico* with the help of the following programs:

- Primer design: Primer3, http://frodo.wi.mit.edu/ (Rozen & Skaletsky, 2000)
- Evaluation of primer specificity: Primer-BLAST, http://www.ncbi.nlm.nih.gov/tools/primer-blast/ (Ye et al., 2012)
- Evaluation of primer structural properties (i.e. dimer formation): IDT OligoAnalyzer 3.1, http://www.idtdna.com/analyzer/applications/oligoanalyzer/ (Owczarzy et al., 2008)
- *In silico* PCR at UCSC genome browser: http://genome.ucsc.edu/cgi-bin/hgPcr (Hinrichs et al., 2006)

Restriction enzyme analyses, *in silico* cloning and analyses of sequencing data, including sequence alignments, were conducted using free software:

- BioEdit sequence alignment editor, version 7.0.5.3: version 7.2.0 available for download at http://www.mbio.ncsu.edu/bioedit/page2.html (Hall, 1999)
- ApE plasmid editor, version 2.0.45 (current): available for download at http://biologylabs.utah.edu/jorgensen/wayned/ape/

3.2.2.2. Classical Cloning

In a standard classical cloning procedure (Sambrook et al., 1989; Montoliu, 1997; Ausubel et al., 1999), vectors and inserts were treated separately. First, vectors were usually digested with restriction enzyme(s), dephosphorylated and purified. In parallel, inserts were digested with the same restriction enzyme(s) and purified. Finally, inserts were ligated into the digested vectors and the mixture was transformed into electro- or chemically-competent bacteria. In some cases, a blunting reaction may have been required after the digestion step.

All necessary enzymes were obtained from New England Biolabs (digestion enzymes) and Roche (additional digestion enzymes, T4 DNA polymerase or Klenow fragment for blunting reactions, raPId alkaline phosphatase, T4 DNA ligase). They were used according to the manufacturers' protocols. Purification and bacterial transformation procedures and reagents are described below.

All generated plasmids were confirmed by restriction digestion and sequencing.

3.2.2.3. Gateway-Based Cloning

The Gateway® Cloning Technology (Invitrogen) was used for the cloning of all the elements bound for enhancer-blocking activity testing, either *in vitro* in HEK 293 cells or *in vivo* in zebrafish. It consisted of a three-step cloning system based on the transfer of a sequence of interest from an intermediate entry vector to a final destination vector, taking advantage of the site-specific recombination properties of bacteriophage lambda (Landy, 1989). This implies that the destination vectors, both pELuc and p48RCAR, had to be prepared for the reception of the element of interest: they had to include the LR recombination cassette. p48RCAR had already been adapted to this technology at Dr. Jose Luis Gomez Skarmenta's laboratory (CABD, Seville) (see section 3.2.1.2 in this chapter), unlike pELuc. Hence, in the first step, pELuc was modified with the Gateway® Vector Conversion System (Invitrogen), according to the supplier's specifications (**Fig. MM 4**). Briefly, reading frame A was cloned blunt upstream from the enhancer at a *Sma*I site in the pELuc vector, generating pELuc-OUT (**Fig. MM 4A**). In parallel, reading frame B was cloned blunt at a *Xho*I site between the enhancer and the promoter in pELuc, resulting in pELuc-IN (**Fig. MM 4B**). Both reading frames contained *att*R recombination sites flanking the *ccd*B gene (negative selection) and the chlorampenicol resistance gene (positive selection). Noteworthy, each cassette was cloned in both orientations at both positions.

In the second step, the elements of interest were PCR-amplified from mouse genomic DNA (HsdWin:NMRI) using Expand High Fidelity PCR System (Roche), and A/T cloned in an intermediate entry vector with the Gateway® pCR®8/GW/Topo® TA Cloning® Kit (Invitrogen), according to the supplier. This entry vector hosts *att*L recombination motives flanking the cloning site (**Fig. MM 5**).

Finally, the DNA fragments were transferred to the corresponding destination vectors -pELuc-IN, pELuc-OUT or p48RCAR- by promoting recombination between *att*L and *att*R sites using the LR Clonase™ II enzyme mix (Invitrogen), according to the provider's protocol (**Fig. MM 6**). Of note, PCR products that had been cloned in the direct orientation

in the entry vector were set to recombine with destination vectors with Gateway reading frames in the direct orientation too; the opposite is true for PCR products cloned in the inverse orientation.

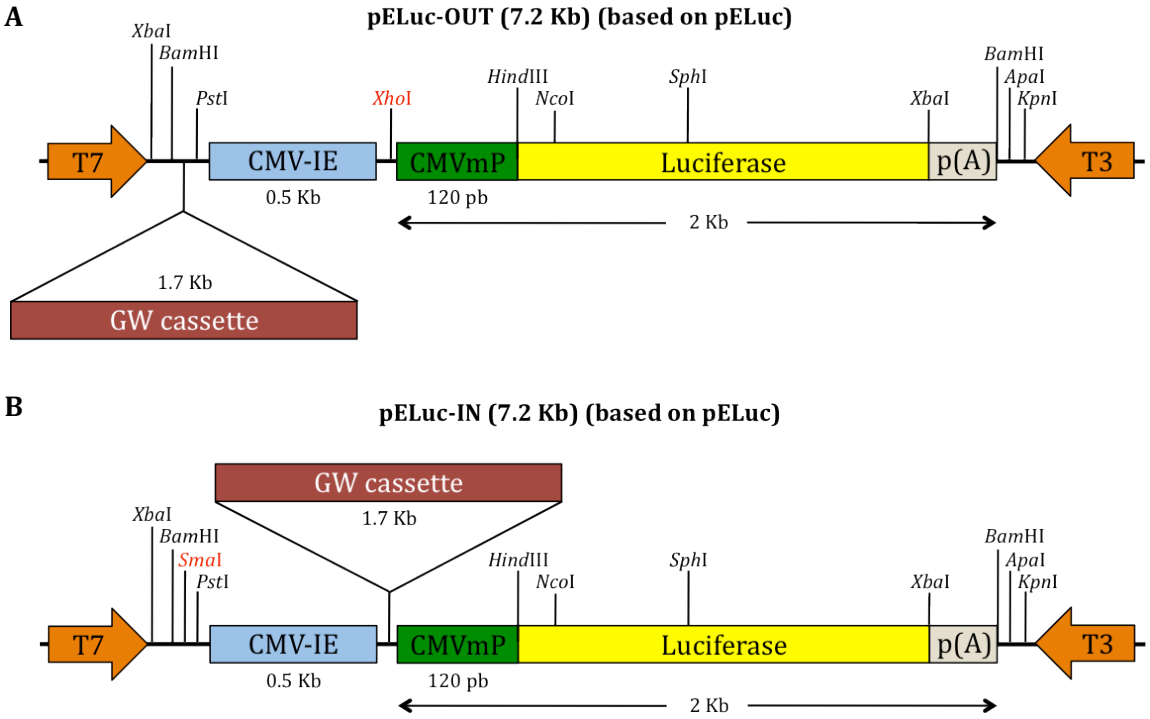


Fig. MM 5. Gateway® Cloning Technology: First step. Conversion of the destination vectors to the Gateway® recombination technology. The Gateway® LR recombination cassette was cloned blunt in the *Sma*I site or *Xho*I site of the pELuc vector, generating pELuc-OUT (A) and pELuc-IN (B), respectively. In each case, the cassette was cloned in both orientations. Only the direct orientation is depicted. p48RCAR had been previously adapted to this technology, and is not shown here.

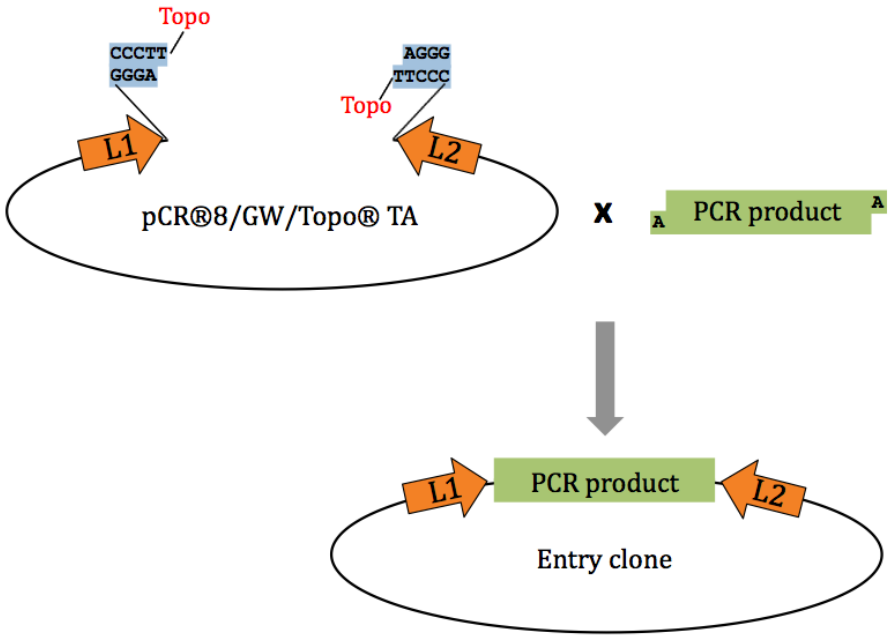


Fig. MM 4. Gateway® Cloning Technology: Second step. TA-cloning of the PCR-amplified desired element into pCR®8/GW/Topo® TA to generate the entry clone.

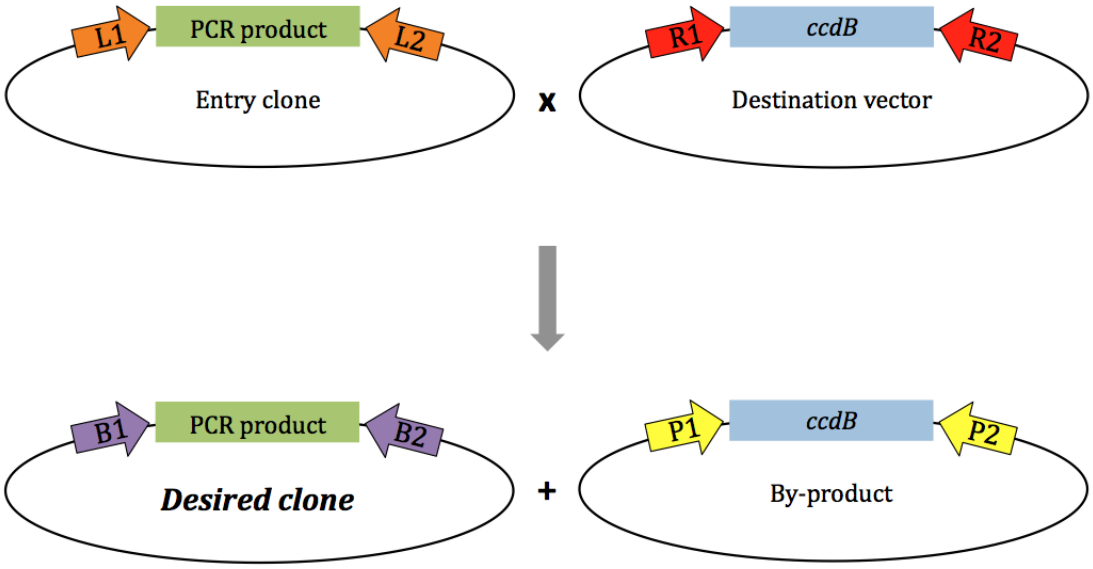


Fig. MM 6. Gateway® Cloning Technology: Third step. LR recombination between the entry clone and the destination vector to create the final desired clone and a by-product. After transforming competent bacteria with the products of the recombination reaction, those bacteria that had acquired the by-product died due to the action of the toxin encoded by the *ccdB* gene.

All generated plasmids were confirmed by restriction digestion and sequence analyses, as well as by PCR with the same primers used in the cloning procedure.

3.2.2.4. DNA Electrophoresis

The separation, identification and analysis of DNA molecules (i.e. PCR products or plasmid fragments after restriction digestion) were carried out by gel electrophoresis. Gels were prepared by melting agarose (Agarose D1 Medium EEO from Pronadisa for routine electrophoresis, and UltraPure™ Agarose from Invitrogen for DNA purification from gel) at different concentrations, ranging from 0.8% to 2%, to separate fragments between 100 bp and 12 kb- in 1x TAE buffer (Tris-Acetate 40 mM, EDTA 2 mM pH 8; Merck) with 0.5 µg/ml of ethidium bromide (EtBr; Sigma-Aldrich). Gels (6x8 or 11x15 cm) were submerged in EtBr-containing 1x TAE buffer in horizontal (8x17 or 16x17 cm) cells (Ecogen), and run at 5 V/cm. Low melting agarose (NuSieve® GTG® Agarose, Lonza) gels were prepared at 3% when small DNA fragments (50 to 200 bp) had to be distinguished.

The molecular weight markers used to estimate the size of the DNA fragments include: 1 kb Plus DNA ladder (Invitrogen, band sizes between 100 pb and 12 kb) and 25 bp DNA ladder (Invitrogen, band sizes between 25 and 500 bp).

The DNA fragments in the agarose gels were visualized under UV light in a gel documentation system (GelDoc 2000, Bio-Rad).

3.2.2.5. Polymerase Chain Reaction (PCR)

Polymerase Chain Reaction or PCR was employed to amplify specific DNA sequences in order to 1) obtain sufficient copies to perform a Gateway-based cloning procedure, 2) check for positive clones, or 3) genotype transgenic animals. The amplification of a DNA piece relies on the enzymatic action of a thermostable DNA polymerase, which generates many copies of a specific DNA sequence by using two oligonucleotides complementary to the target region as primers for amplification.

Taq polymerase, from *Termophilus aquaticus BM* (Roche), was utilized for routine amplification of targets up to 3 kb. Standard PCR reactions were performed in aqueous solution containing: 1.5 mM $MgCl_2$, 200 µM of each deoxynucleotide triphosphate (dATP, dCTP, dGTP, dTTP), 0.5 µM of each specific primer, 1x PCR buffer, and 0.25 U of enzyme, in a final volume of 25 µl in 0.2 ml tubes (MJ Research). All reagents were obtained from Roche, except for primers, which were usually produced by Sigma-Aldrich.

PCR reactions were carried out in PTC-100 (MJ Research), MJ Mini (Bio-Rad) or Primus 96 Plus (MWG Biotech) thermocyclers. The program used depended on the DNA target to amplify, but in general, the PCR conditions set were as follows: first, 2 minutes at 94ºC to allow DNA denaturalization; second, 30 to 35 cycles of amplification that consisted of a quick denaturalization step (30 seconds at 94ºC) followed by primer annealing (30 seconds at 5ºC below the melting temperature Tm of the primers) and extension (1 minute per kb to amplify at 72ºC), and third, 10 minutes at 72ºC to ensure full primer extension. PCR reactions were immediately run in agarose gels or kept at 4ºC until further use.

PCR optimization was required in some cases: increasing the annealing temperature, lowering the concentration of primers, DNA or Mg^{2+} (1 mM minimum), or restricting the extension time were usually sufficient to get rid of unspecific product amplification. The contrary was done to increase the yield of specific PCR products.

A high fidelity polymerase (Expand High Fidelity PCR System, Roche) was used when it was necessary to preserve the exact DNA sequence after amplification (i.e. for cloning). Additionally, long DNA targets (up to 20 kb long) were amplified with Expand Long

Template PCR System (Roche), whereas GC-rich PCR System (Roche) was utilized with GC-rich sequences such as repetitive elements, according to the manufacturer's instructions.

3.2.2.6. DNA Purification from Enzymatic Solutions or Agarose Gels

QIAquick PCR Purification kit (QIAGEN) was used to purify DNA fragments up to 10 kb resulting from enzymatic reactions such as PCR, restriction digestions or dephosphorylations, according to the supplier's instructions. Usually, the DNA was eluted in 50 µl of mQ-H_2O or TE pH 8, depending on the downstream applications.

In order to purify specific DNA fragments from a solution that contained fragments of several sizes, an alternative protocol was carried out. First, the DNA fragments were separated by agarose gel electrophoresis (see section 3.2.2.4). Then, the gel was visualized in a long wave (365 nm) UV light transilluminator (UVP), and the desired band was excised with a clean scalpel and kept in a 1.5 ml tube. Finally, the DNA was extracted and purified with the Wizard® SV Gel and PCR Clean-Up System (Promega), according to the manufacturer's protocol.

3.2.2.7. DNA Quantification and Purity Assessment

Concentration and quality of purified DNA were determined by using a NanoDrop™ ND-1000 Spectrophotometer (Thermo Scientific). This device estimates the DNA concentration of a sample by measuring its absorbance at 260 nm, and taking into account that an optical density of one unit correlates to a concentration of 50 µg/ml double-stranded DNA. Furthermore, the quality of the DNA can also be assessed by calculating the following ratios of absorbance: a ratio 260/280 of around 1.8, and a ratio 260/230 in the range of 2.0-2.2 are indicative of good quality DNA. Lower ratios evidence the presence of contaminants that absorb at either 280 nm (i.e. phenol or proteins) or 230 nm (i.e. residual guanidine from DNA purification kits). Finally, for this device, the lower and upper detection limits of double-stranded DNA are 2 and 3700 ng/µl, respectively.

3.2.2.8. DNA Sequencing

DNA sequencing services were provided by *Parque Científico de Madrid* (PCM-UAM, http://www.fpcm.es/es/servicios-a-la-id/servicios/genomica) or by Macrogen Inc. (Seoul, South Korea or Amsterdam, the Netherlands) (http://dna.macrogen.com/eng/).

3.2.2.9. Site-Directed Mutagenesis

Site-directed mutagenesis of the CTCF binding motif in the CorDis-9.2 element was conducted with the GeneArt® Site-Directed Mutagenesis System (Invitrogen), according to the supplier's indications. Briefly, a set of overlapping mutagenic primers with centrally located mutations sites served to PCR-amplify a previously methylated plasmid that contained the sequence to mutate. Recombination between the ends of the PCR product was promoted *in vitro*, and circularized mutated DNA was transformed into DH5α™-T1^R chemically competent bacteria. These bacteria are positive for *McrBC*, an endonuclease that digests methylated DNA. This property allowed the removal of the methylated plasmid with the original wild type sequence, leaving only the mutated product, which was further purified and checked by restriction digestion and sequence analyses.

3.2.2.10. Bacterial Strains and Growth Medium

Recombinant plasmids were usually propagated in *E. coli* TOP10 bacteria with two exceptions. Firstly, plasmids carrying Gateway recombination cassettes (pELuc-IN, pELuc-OUT, p48RCAR GW, p48RCAR GW') contained the *ccd*B "suicidal" gene and could only propagate in *ccd*B resistant bacteria like One Shot® *ccd*B Survival™ 2 T1 Phage-Resistant (T1^R) chemically competent *E. coli*. Secondly, the site-directed mutagenesis procedure (section 3.2.2.9) relied on *McrBC* positive bacteria, specifically, on One Shot® MAX Efficiency® DH5α™-T1^R chemically competent *E. coli*, so all plasmids edited with this method had to be transformed into this strain. All bacterial strains were purchased from Invitrogen.

The complete genotypes of the bacterial strains employed are the following:

- One Shot® TOP10 competent *E. coli:* F-*mcr*A Δ (*mrr-hsd*RMS-*mcr*BC) Φ80*lac*ZΔM15 Δ*lac*X74 *rec*A1 *ara*D139 Δ (*ara-leu*)7697 *gal*U *gal*K *rps*L (Str^R) *end*A1 *nup*G; chemically- (Invitrogen; transformation efficiency of 10^9 cfu/μg plasmid DNA) or electro-competent bacteria (section 3.2.2.11; transformation efficiency of 10^7 cfu/μg plasmid DNA).
- One Shot® *ccd*B Survival™ 2 T1 Phage-Resistant (T1^R) chemically competent *E. coli* (Invitrogen; transformation efficiency of 10^9 cfu/μg plasmid DNA): F-*mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) Φ80*lac*ZΔM15 Δ*lac*X74 *rec*A1 *ara*Δ139 Δ(*ara-leu*)7697 *gal*U *gal*K *rps*L (Str^R) *end*A1 *nup*G *fhu*A::*IS2.*

- One Shot® MAX Efficiency® DH5α™-T1^R chemically competent *E. coli* (Invitrogen; transformation efficiency of 10^9 cfu/µg plasmid DNA): F-Φ80*lac*ZΔM15 Δ(*lac*ZYA-*arg*F)U169 *rec*A1 *end*A1 *hsd*R17(rk-, mk+) *pho*A *sup*E44 *thi*-1 *gyr*A96 *rel*A1 *ton*A.

Plasmid-containing bacteria of any strain were grown in liquid LB medium (Luria-Bertani; 1.0% bacto-tryptone, BD Biosciences; 0.5% yeast extract, Prodanisa; 1.0% NaCl, VWR; pH 7.2), supplemented with the appropriate antibiotic, which depended on the resistance gene each plasmid carried. The antibiotic concentrations applied in each case were: 50 µg/ml ampicillin, 30 µg/ml chloramphenicol or 100 µg/ml spectinomycin. All antibiotics were obtained from Sigma-Aldrich. Liquid cultures were incubated at 37ºC with vigorous shaking at 250 rpm for 8 to 16 hours (Sambrook et al., 1989).

Additionally, bacteria could also be propagated in solid LB medium supplemented with 15 g/L of agar (European Bacteriological Agar, Conda) and the appropriate antibiotic in *Petri* dishes (Sterilin, 10 cm in diameter) at 37ºC for 8 to 16 hours.

3.2.2.11. Preparation and Transformation of Competent *E. coli* Bacteria

TOP10 *E. coli* bacteria were prepared for transformation by electroporation according to the protocol recommended by Invitrogen. A single colony of TOP10 *E. coli* from a fresh LB plate was inoculated in 50 ml of LB broth and incubated at 37ºC in a shaker overnight. On the next morning, this starter culture was transferred to 1 L of LB and incubated at 37ºC until OD_{600} reached 0.5-0.6 units (2-3 hours). At that point, bacterial growth was halted by incubating the culture at 4ºC on ice for a minimum of 30 minutes. Next, bacteria were centrifuged at 2,000 g for 15 minutes at 4ºC, and washed once with 500 ml of ice-cold sterile mQ-H_2O with the same centrifugation conditions. Subsequently, the bacterial pellet was centrifuged twice with 500 ml of ice-cold 10% glycerol in sterile mQ-H_2O at 4,000 g for 15 minutes at 4ºC. Finally, the pellet was resuspended in its own volume, 40 µl aliquots were placed into sterile 1.5 ml tubes, snap-frozen in liquid nitrogen (Air Liquide) and stored at -80ºC until further use.

Electrocompetent bacteria were transformed according to the following protocol. First, an aliquot of bacterial suspension –one per transformation– was thawed on ice for 5 minutes. Second, DNA was added to the bacteria (10 pg to 100 ng of DNA in a maximum volume of 1.5 µl), which were then incubated on ice for 2 minutes. Third, the mixture was transferred to cold transformation cuvettes (2 mm gap width between electrodes, Bio-Rad) and electroporated by using a MicroPulser Electroporator (Bio-Rad) under the

following conditions: U = 2.5 kV, C = 25 µF and R = 200 Ω. Immediately after electroporation, 1 ml of warm LB medium was added to the bacteria, which were then placed in 10 ml sterile tubes at 37ºC for 1 hour with gentle shaking. Finally, two volumes of transformed bacteria (usually 100 and 900 µl) were plated on LB dishes with the appropriate antibiotic (section 3.2.2.10), and incubated overnight at 37ºC.

Commercial chemically competent bacteria (section 3.2.2.10) were transformed according to the manufacturer's protocol (Invitrogen). In short, one vial of competent cells per transformation was first thawed on ice for 5 minutes. The DNA (10 pg to 100 ng) was then added to the bacteria, and the mixture was kept on ice for 30 minutes. Next, the cells were heat-shocked at 42ºC for 30 seconds without shaking and allowed to recover on ice for 2 minutes. The mixture was then incubated in 250 µl of S.O.C. medium (2% tryptone, 0.5% yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$, 20 mM glucose; Invitrogen) in a 10 ml sterile tube at 37ºC for an hour in a shaker. Finally, two volumes (100 and 150 µl) were plated on LB plates with the corresponding antibiotic.

3.2.2.12. Mini- and Maxipreparations of Plasmid DNA from *E. coli*

Plasmid minipreparations (2 ml of starting *E. coli* culture volume) were prepared with Wizard® Plus SV Minipreps DNA Purification System Kit (Promega); whilst maxipreparations (100 or 500 ml of starting *E. coli* culture volume for high or low copy number plasmids, respectively) were carried out using Plasmid Maxi Kit (QIAGEN). Both procedures involved three steps: 1) growth of the bacterial culture, 2) alkaline lysis of the bacteria to release the plasmids, and 3) column-based purification of the DNA. They were performed according to the supplier's protocols.

3.3. Cell Lines and Culture Conditions

Three different adherent cell lines were used in this project: HEK 293 cells (human embryonic kidney cells), L929 cells (murine fibroblasts) and COCA cells (murine keratinocytes). The last-mentioned cell line was gently ceded by Dr. Corina Lorz's laboratory (CIEMAT, Madrid). These cells were expanded and differentiated for 24, 48 and 72 hours at CIEMAT under previously described conditions (Segrelles et al., 2011), whereas the rest of the cell lines were handled at the CNB.

HEK 293 and L929 cell lines were grown in DMEM medium (Dulbecco's Modified Eagle Medium, Gibco) supplemented with sterile-filtered 10% fetal bovine serum (FBS, Sigma-Aldrich), 2 mM L-glutamine (Invitrogen), 10 mM HEPES pH 7.4 (Invitrogen), penicillin (100 U/ml) and streptomycin (100 µg/ml) (penicillin-streptomycin solution; Sigma-Aldrich) under aseptic conditions using a sterile hood (Telstar Bio II Advance).

Cells were cultured at 37ºC, 95% of humidity and 5% CO_2 (CO_2 water-jacketed incubator, Forma Scientific) in plates of different sizes according to the number of cells required for each experiment: 24-well plates (2 cm^2), p150 (151.90 cm^2), p100 (58.10 cm^2) or p60 (21.29 cm^2) dishes were purchased from Falcon.

Usually, cells were grown to nearly 100% confluency before they were detached by incubation with 0.25% trypsin-0.02% EDTA (Gibco) for 5 minutes at 37ºC, and reseeded at 1:8 to 1:10 dilution. Whenever it was necessary to determine the cell number, a cell counting chamber (Sigma) was used according to the manufacturers' specifications.

Cell culture supernatants were regularly checked for the presence of mycoplasma contamination by PCR with primers specific to mycoplasma DNA (see section 3.10). Only cells that were free from biological contaminants were frozen in aliquots of $5 \cdot 10^6$ cells in 1 ml cryotubes (Nunc) and stored in liquid nitrogen tanks. The freezing medium was similar to the regular growth medium of each cell type, but with 10% of DMSO (Merck) as cryoprotectant.

3.4. *In Vitro* Enhancer-Blocking Assays in HEK 293 Cells

3.4.1. Plasmid DNA Transfection into Mammalian Cells

The *in vitro* enhancer-blocking assays (EBA) were performed by transiently transfecting human embryonic kidney 293 (HEK 293) cells with the corresponding experimental and control plasmids in triplicates, in at least two independent assays, as reported (Lunyak et al., 2007).

First of all, maxipreps of pLuc-based plasmids were linearized to avoid bidirectional enhancer activity (Recillas-Targa et al., 1999) using *Asp*718 (Roche), which cut downstream of the luciferase cassette. Alternatively, *Apa*I (Roche) was employed whenever the element to test contained an *Asp*718 site, as in the case of the control constructs pEII/III and pII/IIIE. Also, *Sca*I (Roche) served to linearize the transfection

control pCMV-lacZ. Aliquots of undigested and digested plasmids were always run side by side on an agarose gel (0.8%) to confirm linearization.

Subsequently, 24 hours prior to transfection, $2 \cdot 10^5$ cells were seeded per well in 500 µl of cell growth medium in 24-well plates (Nunc), so that at the time of transfection, cellular density reached 90-95% confluence. Three wells were prepared per plasmid.

Transfection was then carried out using Lipofectamine 2000 reagent (Invitrogen), according to the supplier's protocol. First, 1.98 µg of each experimental plasmid were mixed with 0.42 µg of control pCMV-lacZ in 150 µl of Opti-MEM medium (Invitrogen). In parallel, for each transfection reaction, 6 ul of Lipofectamine 2000 were diluted in 150 µl of Opti-MEM medium and incubated for 5 minutes at room temperature (RT hereafter). The DNA mixtures were then added to the tubes that contained the Lipofectamine 2000 reagent, and incubated for 20 minutes to allow the formation of the transfection complexes. Afterwards, 100 µl of each Lipofectamine 2000/DNA mixture were added drop-wise to each of three wells in the already prepared 24-well plates. In the end, the cells in each well were transfected with a Lipofectamine 2000 : DNA ratio of 2 µl : 0.8 µg in a final volume of 600 µl. Finally, the cells were incubated under normal growth conditions for 24 hours.

3.4.2. Preparation of Cellular Extracts

One day after transfection, cells were washed once with 500 µl of cold PBS (137 mM NaCl, 2.7 mM KCl, 16.2 mM Na_2HPO_4, 1.5 mM KH_2PO_4) and lysed in 125 µl of 1x Reporter Lysis Buffer (Promega) with the aid of a rubber scraper. Cellular lysates were collected in 1.5 ml ice-cold tubes and vortexed for 10 seconds. Next, they were centrifuged at 12,000 g for 5 minutes at 4ºC. Supernatants containing whole-cell protein extracts were transferred to new 1.5 ml ice-cold tubes, and kept at -80ºC until necessary.

3.4.3. β-Galactosidase Activity Measurements in Cellular Extracts

The measurement of β-galactosidase activity was performed by quantifying the hydrolysis of the chromogenic substrate o-nitrophenyl-β-D-galactoside (ONPG, Sigma-Aldrich), as described elsewhere (Hall et al., 1983). Briefly, 100 µl of a 1:100 dilution of each cellular lysate were mixed with 400 µl of Z-buffer (100 mM Na_2HPO_4 pH 7.2, 10 mM KCl, 1 mM $MgSO_4$, 50 mM β-mercaptoethanol freshly added) and 100 µl of ONPG (4 mg/ml

in 100 mM Na_2HPO_4 pH 7.2). Additionally, as a negative control, a tube containing 100 ul of 1x Reporter Lysis Buffer instead of cellular lysate was included. The mixtures were incubated at 37ºC until they turned pale yellow (30 minutes – 1 hour). At that moment, the reactions were stopped with the addition of 250 µl of 1M Na_2CO_3. Then, the absorbance of 200 µl of each solution was read at 414 nm in a microplate reader, using 96-well plates (Nunc) and taking the control tube as the blank reference.

3.4.4. Luciferase Activity Measurements in Cellular Extracts

The luciferase activity of each cellular extract was quantified using 100 µl of the same 1:100 dilutions prepared for the β-galactosidase assay, in opaque-white 96-well plates (Berthold). The Orion Microplate Luminometer (Berthold) was programmed to inject 50 µl of Luciferase Assay Reagent (luciferin, Promega) in each well, with a measurement period of luciferase activity of 10 seconds, and a delay time between wells of 2.05 seconds.

3.4.5. Data Analysis

Relative luciferase activities (A.U./pmol) were obtained by calculating the ratio between luciferase and β-galactosidase activity levels, and normalizing to the amount of transfected construct expressed in pmoles, in each case. Mean relative luciferase activities of the triplicates, along with the corresponding standard deviations, were used to compare the performance of each experimental construct with respect to the controls.

Furthermore, the relative luciferase activity of the control pELuc was divided by that of each construct in order to obtain fold enhancer-blocking activities, which were represented along with the standard deviation of the triplicates.

At least two independent assays were performed per construct. To be able to compare the results among assays, the positive control cHS4 was used as a calibrator. Hence, the fold enhancer-blocking activity of each construct was divided by that of the cHS4 in the same assay. Results were then presented as mean fold enhancer-blocking activity of the replicates, relative to cHS4, and along with the standard error of the mean.

3.5. *In Vivo* Enhancer-Blocking Assays in *Danio rerio* (Zebrafish)

In vivo enhancer-blocking assays were carried out at Dr. José Luís Gómez Skarmeta's laboratory, at *Centro Andaluz de Biología del Desarrollo* CABD, with the help of Dr. Ana Fernández Miñán.

All procedures that required the manipulation of zebrafish individuals met the European Union animal research guidelines (Directives 86/609/CEE and 2010/63/UE), the corresponding Spanish laws (RD1201/2005 and RD53/2013), and the requirements of the CABD and CSIC Ethics Committee.

3.5.1. DNA Purification by Phenol-Chloroform-Isoamyl Alcohol Extraction and Ethanol Precipitation

Plasmid DNA for microinjection was purified using standard protocols (Sambrook et al., 1989). First, 1/10 volume of 3M sodium acetate (pH 5.2) was mixed with the DNA-containing solution in a 1.5 ml phenol-resistant tube. Second, an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1) was added. The contents of the tubes were mixed by inversion until an emulsion formed. The mixture was then centrifuged for 5 minutes at top speed at RT, and the superior aqueous phase was transferred to a new tube. An equal volume of chloroform:isoamyl alcohol (24:1) was added next. Again, the solution was thoroughly mixed by inversion and centrifuged for 5 minutes at top speed at RT. The superior phase was collected in a fresh tube, and 2 volumes of cold 100% ethanol were added. The mixture was then incubated at -20ºC for 1 hour. DNA precipitation was facilitated by centrifuging the samples for 15 minutes at top speed at RT. The pellet was washed with 1 ml of cold 70% ethanol under the same conditions of centrifugation. Finally, the supernatant was decanted carefully and the pellet resuspended in water.

DNA concentrations were determined using Nanodrop™ ND-1000 Spectophotometer (section 3.2.2.7), and the samples were kept at -20ºC until further use.

3.5.2. Fish Husbandry

Zebrafish were maintained under standard conditions (Westerfield, 2007), in a circulating system in which water (E3 medium: 5 mM NaCl, 0.17 mM KCl, 0.4 mM $CaCl_2$, 0.16 mM $MgSO_4$) was continuously filtered and aerated. The animals were kept in 11-liter tanks at 28ºC, with a light cycle of 12:12 hours light:dark (artificial lighting from 9.00 a.m.

to 21.00 p.m.). To increase egg production, fish were fed dry food, rotifers and arthemia, six times a day.

After collection, embryos were always maintained in E3 medium supplemented with 0.01% methylene blue (Sigma-Aldrich) as a fungicide. Additionally, other reagents such as 1-phenyl-2-thiourea (PTU) or tricaine could be incorporated into the medium (see below).

3.5.3. Embryo Collection

The day before a microinjection session, male and female fish from the same tank were separated into two different 8-liter cages, and given extra food. More than one tank could be employed, depending on the number of fertilized eggs that would be needed for microinjection. Several clues helped in the identification of the sex of the zebrafish. First, males are smaller than females. Also, their stripes are darker and have a yellowish cast. On the other hand, females present protruding bellies and are silvery in appearance.

Early in the morning on the microinjection day, male and female fish were rejoined in a mating cage that contained an insert with a wired steel mesh on the bottom. This mesh prevented the fish from eating the freshly-laid eggs, which accumulated at the bottom of the cage typically 5 to 30 minutes later. Eggs were collected by filtration through a plastic tea sieve in a p100 plate with E3 medium, around 10-15 minutes after being laid, to allow time for the *ex vivo* fertilization. Then, they were transferred to a clean and dry plate, and lined up against the wall of a rectangular plastic insert that had been placed inside.

After microinjecting the first set of embryos, more mating individuals could be combined to obtain more embryos.

3.5.4. Microinjection into Zebrafish Embryos

For the *in vivo* evaluation of insulator activity, p48RCAR-based plasmids were microinjected into one-cell stage zebrafish (*Danio rerio*) embryos as reported (Bessa et al., 2009; Kawakami, 2004). Firstly, the constructs were purified by phenol:chloroform:isoamyl alcohol extraction and ethanol precipitation. Secondly, purified constructs were co-injected with Tol2 transposase mRNA in a volume of 5 nanoliters per cell (50 ng/ul of transposase mRNA, 40 ng/ul of purified DNA construct, and 0.05% of phenol red) with the assistance of an IM-300 microinjector (NARISHIGE).

The needles for microinjection (Glass Capillary Filament 1.0 mm x 58 mm, 6'' CIBERTEC 601500) were prepared in a horizontal puller (Sutter instrument model P-67) under the following settings: Heat = ramp test + 30 = 577; Pull = 40; Vel = 50; Time = 50. Before loading the needle (always from the back), the tip was broken off at 1.1 cm from the narrowing of the needle diameter using forceps under the microscope.

The DNA/mRNA mixture was always microinjected into the yolk/cell interphase for several reasons: first, it constitutes an easier target than the cell itself; second, the cellular structure remains intact, and third, the nucleic acids diffuse rapidly into the cell.

After microinjection, embryos were incubated in 30 ml of E3 medium in p100 plates (approximately 50 individuals per plate) at 28ºC. A plate with untreated embryos was kept in parallel as a control: the next day, the proportion of dead embryos (opaque white) in the control plate should be minimum compared to that in the plates with manipulated embryos, in which a high mortality rate was expected due to the microinjection process. Otherwise, the embryos utilized would have been of poor quality. A day after microinjection, dead as well as deformed embryos were removed, and the medium was replaced with fresh E3 containing 0.003% PTU (Sigma-Aldrich) to prevent pigmentation and allow the visualization of GFP fluorescence (Karlsson et al., 2001).

3.5.5. Microscopy and Imaging

Approximately 36 hours post fertilization (hpf), at least 30 transgenic embryos with homogenous bright GFP expression in the somites were selected to evaluate the enhancer-blocking activity of each plasmid, including the basal construct p48RCAR.

These zebrafish were dechorionated in E3 medium supplemented with 0.003% PTU. Firstly, the chorion (external membrane) of each embryo was carefully gripped with a pair of fine-tip tweezers. Secondly, the same point of the chorion was also gripped with a second pair of tweezers. Finally, the tweezers were pulled in opposing directions, releasing the embryo to the medium. The movement of the fish tails facilitated the process.

The platform used to place the embryos for photography consisted of a p60 plate with a base of 1% solid agarose and filled with E3 medium. Several drops of tricaine 0.4% (McFarland & Klontz, 1969; Craig et al., 2006; Sigma-Aldrich) were added to the plate in order to anesthetize the zebrafish so that proper pictures could be taken. A small hole was bored in the agarose base to accommodate the yolk of the animals and ensure that both the midbrain and somites were in focus at the same time.

Fluorescent images were acquired under a fluorescence microscope (Olympus SZX16), using the CellSens Entry 1.6 software (Olympus), always under the same settings.

3.5.6. Image Processing with the LaserPix Software (Bio-Rad)

GFP fluorescence of each transgenic individual was quantified with the LaserPix image analysis software (Bio-Rad) under automatic settings (**Fig. MM 7**). Measurement counts that fell outside the midbrain or somites regions were removed from the analysis. The ratio of fluorescence in somites (driven by the cardiac actin promoter) *versus* fluorescence in the central nervous system (promoted by the Z48 midbrain neuronal enhancer) was calculated and normalized to the ratio achieved by the basal construct p48RCAR. Some individuals did not present any count in the midbrain. In these cases, in order to avoid infinite values for somites to midbrain fluorescence ratios, the measurement count of a single pixel (the minimum possible value) was arbitrarily imputed to the midbrain.

To illustrate the variability of the assay, the enhancer-blocking efficacy of the constructs was represented as boxplots using Prism 6 (GraphPad Software). The non-parametric median test (IBM Statistics SPSS v19) was chosen to statistically compare the results between the control and the experimental constructs.

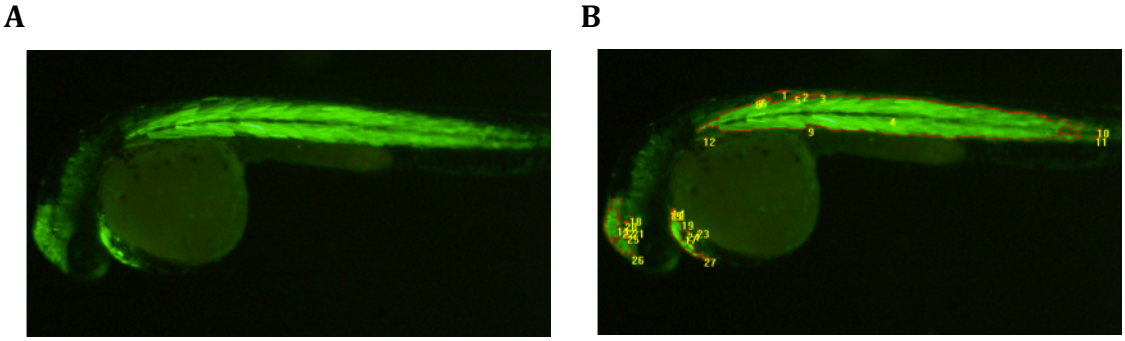


Fig. MM 7. Imaging analysis with the LaserPix software (Bio-Rad). A. Transgenic zebrafish with GFP expression in somites, heart and midbrain (p48RCAR construct). **B.** After processing the image with LaserPix, only the fluorescent pixels in the somites and the midbrain were considered for calculating the *in vivo* enhancer-blocking activity of a given experimental construct.

3.6. Protection Against Chromosomal Position Effects Assay in Mice

All procedures involving the use of mice complied with the European Union animal research legislation (Directives 86/609/CEE and 2010/63/UE), as well as with the

Spanish legislation (RD1201/2005 and RD53/2013). Also, they were reviewed and approved by the corresponding CNB and CSIC Ethics Committees on animal experimentation.

3.6.1. Preparation of Transgenes for Microinjection

Only the CorDis-9.2 element was tested for barrier activity in mice. This element had been previously cloned into the Gateway entry vector (pCR®8/GW/Topo®), from which it was transferred by recombination to the final destination vectors used in the enhancer-blocking assays (both in HEK 293 cells and in zebrafish). Since ptrTYR5 lacked the Gateway recombination cassette, CorDis-9.2 had to be cloned using classical cloning procedures (section 3.2.2.2). Therefore, the element was extracted from the Gateway intermediate vector by digestion with *Eco*RI. The resulting DNA fragments were run on a 0.8% agarose gel. The band corresponding to CorDis-9.2 (1 kb) was purified and cloned blunt at the *Xba*I site that lay just upstream of the mouse tyrosinase minigene in ptrTYR5, generating ptrTYR5-CorDis-9.2.

The 12.2-kb transgene (CorDis-9.2 fused to the tyrosinase minigene) was released from the plasmid by digestion with *Ecl*XI and *Sal*I, and purified by running the sample on a 0.8% agarose gel and extracting the corresponding band. In this case, sterile-filtered (Millex-GS Syringe Filter Unit, 0.22 µm, Millipore) microinjection buffer (10 mM Tris-HCl pH 7.5, 0.1 mM EDTA pH 8.0, in tissue culture tested sterile H_20; Gibco) was used to resuspend the DNA. The sample was further purified by dialysis: it was placed on a dialysis filter disc (Millipore 0.05 µm) that was floating in 40 ml of microinjection buffer in a Petri dish (Sterilin) for three hours. The sample was then carefully recovered and its concentration determined by quantification in a Nanodrop™ ND-1000 Spectophotometer. Additionally, serial dilutions of the sample were run on a 0.8% agarose gel in combination with another transgene of similar size and known concentration in order to confirm spectophotometric measures.

3.6.2. Production of Transgenic Mice by DNA Microinjection

Transgenic mice were produced at the Transgenic Core Facility of the *Centro Nacional de Biotecnología* and the *Centro de Biología Molecular Severo Ochoa* (CNB-CBMSO-CSIC, Madrid) by DNA microinjection (0.5-2 ng/µl) into the pronucleus of albino outbreed

HsdWin:NMRI fertilized mouse oocytes (Harlan Laboratories) using standard procedures (Hogan et al., 1994; Montoliu, 1997).

3.6.3. Mouse Colony Husbandry

Outbred HsdWin:NMRI mice (Harlan Laboratories) were maintained at the animal facility of the *Centro Nacional de Biotecnología* (CNB-CSIC, Madrid). They were bred in cages with easily accessible water and irradiated food pellets (Harlan Laboratories) *ad libitum*, and exposed to light cycles of 12:12 hours light:dark (300 lux of artificial lighting from 8.00 a.m. to 20.00 p.m.). The temperature of the facility was set at 22±1ºC, and the relative humidity at 50±15%; whilst the ventilation system provided high-efficiency-particulate-air filtration (EU13 HEPA) with a ventilation rate of 20 air changes per hour.

Transgenic founder mice were weaned and sexed at 21 days of age. At around eight weeks of age, they were crossed with wild type HsdWin:NMRI individuals in order to determine if the founders were able to transmit the transgene through the germline, and to establish hemicygotic transgenic lines. After weaning, F_1 mice were genotyped and biochemical analyses were conducted on positive transgenic lines (section 3.6.4). In addition, some representative F_1 individuals were anesthetized with 1ml/100g of a mixture of Ketamine (10 mg/ml) and Xylazine (2 mg/ml) and photographed.

Every animal carried a metal earring tag with a code stamped on it (A0001-Z9999; National Band & Tag Company) to allow their identification. Furthermore, mouse colonies were managed with the help of the "*Raton*" database (Montoliu, 2003), freely available upon request at http://www.cnb.csic.es/~montoliu/mouseDB.html.

3.6.4. Transgenic Mice Genotyping and Analysis

3.6.4.1. Genomic DNA Extraction from Tissue Samples

Mouse genomic DNA was extracted from tail biopsies (< 1 cm) obtained at the time of weaning, in a two-step procedure (Montoliu, 1997). First, the tissue was digested overnight with a proteinase K-containing lysis solution (0.5 mg/ml proteinase K, Roche; 100 mM NaCl, 50 nM Tris-HCl pH 8, 100 mM EDTA pH 8, 1% SDS, Merck). In the second step, the DNA was purified by ethanol precipitation, resuspended in TE buffer (10 mM Tris-HCl pH 7.5, 1 mM EDTA pH 8, Merck) and stored at 4ºC until further use.

3.6.4.2. Genotyping by PCR

The presence of the transgene in mouse genomic DNA was determined by PCR. All amplification reactions were carried out using 10 to 40 ng of genomic DNA (1 µl of 1:10 dilutions of the genomic DNA extracted in 3.6.4.1) in a final volume of 25 µl with primers specific for the tyrosinase minigene (**Appendix MM-1**; Beermann et al., 1991). The resulting products were analyzed by DNA electrophoresis.

3.6.4.3. Quantification of Transgene Copy Number and Determination of Integration sites by Southern Blot

Southern blot analyses (Southern, 1975; Montoliu, 1997) complemented the genotyping of transgenic mice. Not only did they provide information about the presence or absence of the transgene, but they were also used to infer the transgene copy number and its integrity. The technique was carried out as previously described (Schedl et al., 1993; Montoliu et al., 1996). First, 5 µg of mouse genomic DNA was digested overnight at 37ºC with 30 U of *Hind*III (Roche) in a final volume of 60 µl (1 x digestion buffer B, Roche; 4 mM spermidin, Sigma-Aldrich). Digestion products were then separated by horizontal electrophoresis in a 0.8% agarose gel (20 x 25 cm) in 1x TAE buffer (Horizon® 20·25, Life Technologies) at 5 V/cm for 4 hours. The 1-kb DNA ladder (Invitrogen, band sizes between 75 bp and 12.2 kb) was labeled with radioactivity (see below) and used as a molecular marker (approximately 12,000 cpm per lane, quantified with a Wallac 1410 liquid scintillation counter).

After electrophoresis, the gel was incubated for 15 minutes in 0.25 N HCl (Merck) to depurinize the DNA and facilitate its transfer to the membrane. This was followed by 2 incubations of 15 minutes in 0.4 N NaOH (Merck) to neutralize and denature the DNA. Afterwards, the DNA was transferred to a nylon membrane (Amersham Hybond™-N, GE Healthcare) by capillarity for 14 to 16 hours in SSC 20x transfer buffer (3 M NaCl, Merck; 0.3 M sodium citrate, Calbiochem). DNA was then fixed to the membrane by UV cross-linking (two pulses of 70 mJ/cm^2 at 254 nm in a CL-1000 Ultraviolet Crosslinker, UVP-Stratagen). At this point, the membrane could be stored at 4ºC until further use.

Alternatively, the membrane was immediately hybridized with the pmTyrE5 probe (Schedl et al., 1992) that revealed polymorphic digested bands between the tyrosinase minigene transgene (3.4 kb) and the endogenous gene (2.2 kb). Before hybridization, 50 ng of probe were labeled with 30 µCi of dCTP [α-^{32}P] (Perkin-Elmer) by using the High

Prime kit (Roche), based on the random primer labeling technique (Feinberg & Vogelstein, 1983). The ProbeQuant™G-50 kit (GE Healthcare) served to purify the labeled probe. Hybridization was carried out at 65ºC for 14 to 16 hours in a rotating oven according to previously described protocols (Montoliu, 1997; Giraldo, 2002). The membrane was finally kept in an exposure cassette (Bio-Rad) at RT in the dark. One to two days later, it was scanned in the Molecular Imager® FX System (Bio-Rad) and the resulting image was analyzed with Quantity One® v4.6.6 software (Bio-Rad).

Furthermore, transgene integration sites in the various transgenic lines were determined using the same procedure with two modifications. First, *Bst*XI, which cut only once in the transgene, was used instead of *Hind*III. Second, the probe specifically hybridized with the transgene: it was targeted to the SV40 poly(A) tail (see primers in **Appendix MM-1**).

3.6.4.4. Quantification of Melanin Content

Cellular melanin content was determined in the eyes of, at least, three representative F_1 individuals from each of the transgenic mouse lines generated in 3.6.2, according to previously described procedures (Donatien & Orlow, 1995; Gimenez et al., 2005). After sacrificing the animals by cervical dislocation, their eyes were extracted, immediately flash-frozen in liquid nitrogen and stored at -80ºC until further use.

For each animal, only one eye was processed. First, it was weighed and then homogenized in 300 µl of PBS with the aid of a Polytron (Ultra-Turrax T8, Ika). A third of this crude protein extract was vigorously mixed in a vortex with 900 µl of a solution that contained 2M NaOH (Merck) and 20% DMSO (Sigma-Aldrich). The samples were incubated on a rotating device protected from light for 14 to 16 hours. Afterwards, the absorbance at 470 nm of the samples was measured in a spectrophotometer (Ultrospec 3100 Pro, Amersham Biosciences). Additionally, 1 µl of each sample was employed to determine the total protein mass in the extracts, using the Protein Assay kit by Bio-Rad, according to the manufacturer's instructions and with serial dilutions of bovine serum albumin (Sigma-Aldrich) as a standard curve.

Finally, average eye melanin content per total protein mass for all the individuals of each transgenic line was calculated and represented, along with the corresponding standard deviations.

3.6.4.5. Quantification of Tyrosinase Expression by TaqMan qPCR

At least three representative F_1 individuals for each transgenic line were sacrificed by cervical dislocation. Their eyes, brain and 1 cm^2 of back skin were extracted and flash-frozen in liquid nitrogen. Then, they were stored at -80ºC until further use.

Tyrosinase expression was assessed using TaqMan Quantitative PCR assays (Applied Biosystems) with a specific mouse tyrosinase probe (*Tyr* Mm00495817_m1). The mouse TATA-binding protein probe (*Tbp* Mm00446973_m1) was included for normalization (Gimenez et al., 2003; Lavado et al., 2006). All experiments were conducted in duplicates and repeated twice, according to the manufacturer's manual in a 7500 Real-Time PCR System (Applied Biosystems) under standard conditions. Gene expression data were analyzed with the 7500 software.

More details regarding RNA extraction from the different animal tissues and cDNA preparation can be found in section 3.8.

3.7. Transient ChIP Assay

Transient chromatin immunoprecipitation (ChIP) assays aim at quantifying the ability of a given protein to bind exogenous targets previously transfected into a cell line.

3.7.1. Plasmid DNA Transfection into Mammalian Cells

Transfection of HEK 293 cells was conducted as previously described in section 3.4.1 with minor differences. In particular, 24 hours prior to transfection, three p60 dishes were seeded with 10^6 cells per dish. Afterwards, 8 µg of each experimental plasmid (CorDis-9.2 or CorDis-9.2-Mut in pELuc-IN) were transfected with 20 µl of Lipofectamine 2000 (Invitrogen). In parallel, the third p60 dish was used as a mock control. All dishes were returned to the incubator for 24 hours before the chromatin immunoprecipation assay.

3.7.2. Chromatin Immunoprecipitation (ChIP)

ChIP assays were performed according to the protocol recommended by Abcam (http://www.abcam.com/ps/pdf/protocols/x_CHip_protocol.pdf).

3.7.2.1. Crosslinking and Cell Harvesting

Cells were crosslinked by adding formaldehyde (37% formaldehyde solution, Merck) drop-wise directly to the dishes, to a final concentration of 0.5% and incubating them with gentle shaking for 5 minutes at RT. The reaction was halted with the addition of glycine (Merck) to a final concentration of 125 mM. The dishes were incubated again with shaking for 5 minutes at RT.

Then, the cells were washed twice with 5 ml of cold PBS supplemented with 0.1 mM PMSF (PhenylMethaneSulfonylFluoride), scraped in 3 ml of the same buffer and transferred to a 15 ml tube (Falcon). The remaining cells were recovered by washing the dishes with 2 ml of cold buffer and transferring the suspensions to the corresponding tubes. The cell suspensions were centrifuged for 5 minutes at 1,000 g at 4ºC: supernatants were discarded and pellets resuspended in 500 µl of cold FA lysis buffer (50 mM HEPES-KOH pH 7.5, 140 mM NaCl, 1 mM EDTA pH 8, 1% Triton X-100, 0.1% sodium deoxycholate, 0.1% SDS and freshly added 0.1 mM PMSF).

3.7.2.2. Sonication

The lysates were sonicated to shear DNA under the following settings: 30% of amplitude and 6 cycles of 10 seconds on / 1 minute off on ice, in a Vibra-Cell™ VC50 Ultrasonic Processor (Sonics and Materials, Inc.) with a 3 mm microtip probe. The resulting fragment sizes ranged between 600 and 1,000 bp.

Cell debris were pelleted by centrifugation at 8,000 g for 30 seconds at 4ºC. An aliquot (50 µl) was taken from the supernatant and kept as the input control. The rest of the sonicated chromatin was snap-frozen in liquid nitrogen (Air Liquide) and stored at -80ºC until further use.

3.7.2.3. Determination of DNA Concentration

The input aliquot was de-crosslinked in order to quantify the DNA concentration of the remaining frozen samples. The reversal of the crosslinks was performed by adding 50 µl of TE buffer (pH 8), 5 µl of 10% SDS, 2 µl of 10 mg/ml proteinase K (Roche) and 2 µl of 10 mg/ml RNase A (Roche), and incubating the samples at 65ºC for 14 to 16 hours. Next, the samples were purified with the QIAquick PCR Purification kit (QIAGEN, section 3.2.2.6) and quantified in a NanoDrop™ ND-1000 Spectrophotometer (Thermo Scientific).

3.7.2.4. Immunoprecipitation

Frozen chromatin samples were thawed on ice for 5 minutes. A pre-clearing step was included to minimize background PCR signals. For each sample, two 1.5 ml tubes were prepared with 1 µg of DNA diluted 1:10 in RIPA buffer (50 mM Tris-HCl pH 8, 150 mM NaCl, 2 mM EDTA pH 8, 1% Igepal CA-630, 0.5% sodium deoxycholate, 0.1% SDS and freshly added 0.1 mM PMSF). Furthermore, 15 µl of Protein A/G Plus-Agarose beads (Santa Cruz Biotechnology, Inc.) were added to the samples, which were then incubated for 14 to 16 hours in a rotating wheel at 4ºC. On the following morning, the samples were centrifuged for 1 minute at 2,000 g at 4ºC and the supernatants were transferred to new tubes.

In one of the two tubes prepared for each sample, 0.5 µl of anti α-CTCF antibody (rabbit polyclonal antibody produced at Dr. Recillas-Targa's laboratory, *Instituto de Fisiología Celular* – UNAM, México D.F.; Furlan-Magaril et al., 2011) was added (ChIP samples). The other samples remained as negative controls. All tubes were incubated for 14 to 16 hours in a rotating wheel at 4ºC.

On the next day, 20 µl of new beads were added to all samples. Again, they were incubated for 2 hours in a rotating wheel at 4ºC and centrifuged at 2,000 g for 1 minute at 4ºC. The beads were washed by centrifugation three times with washing buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA pH 8, 150 mM NaCl, 20 mM Tris-HCl pH 8) and twice with final washing buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA pH 8, 500 mM NaCl, 20 mM Tris-HCl pH 8) under the same conditions.

3.7.2.5. Reversal of the Crosslinks and DNA Elution

The beads were incubated in a rotating wheel at 30ºC for 30 minutes in 120 µl of elution buffer (1% SDS, 100 mM $NaHCO_3$). They were subsequently centrifuged at 2,000 g for 1 minute at RT, and the DNA-containing supernatants were transferred to new 1.5 ml tubes. The reversal of the crosslinks and DNA purification steps were carried out as described in section 3.7.2.5.

3.7.2.6. PCR and Data Analysis

Several PCRs were set up to determine the binding of CTCF to CorDis-9.2 and CorDis-9.2 Mut sequences: experimental PCR with primers that detect both WT and mutant

CorDis-9.2, and positive control PCR for the exon 18 of human *APP*, which indeed binds CTCF (see **Appendix MM-1**).

In the case of the experimental PCR, 1 µl of 1:10 dilutions of ChIP DNA template for all samples (mock transfected or transfected with either plasmid) or 1 µl of input (corresponding to 1% of starting chromatin in each case) were amplified according to the standard protocol. For the positive control PCR, 1 µl of ChIP DNA template (no dilution) was used instead.

PCR products were run on 2% agarose gels and the intensity of the bands was quantified using the Quantity One® v4.6.6 software (Bio-Rad). Each sample was first normalized with respect to its no-antibody negative control, and then to its corresponding positive control (*APP* exon 18). Finally, CTCF binding to CorDis-9.2 Mut was calculated after setting the binding to CorDis-9.2 WT as the maximum possible binding.

3.8. Gene Expression Analysis by Real-Time Quantitative PCR

Real-time quantitative PCRs allow, not only the detection of a specific DNA sequence in a sample, but also its initial concentration. The technique can be divided into two basic steps. First, single-stranded RNA was extracted from the cells under study and reverse transcribed into double-stranded complementary DNA (cDNA). Second, cDNA was PCR-amplified and quantified using an internal control for normalization purposes.

3.8.1. RNA Extraction from Cultured Cells and Animal Tissues

Different methods were employed to extract RNA from biological material. In the case of cultured cells and mouse eyes, the RNeasy Mini kit (QIAGEN) was used, whereas the RNeasy Fibrous Tissue kit (QIAGEN) was preferred to handle mouse skin. Both kits were utilized according to the manufacturer's instructions and using, as starting material, $1 \cdot 10^6$ cultured cells or 30 mg of frozen tissue.

On the other hand, brain RNA was extracted using the classical LiCl-urea protocol. In this case, mouse brains were cut in halves: one to be flash-frozen and kept at -80ºC, and the other to be submerged in 5 ml of buffer H (3M LiCl, 6M urea, in DEPC-treated water) and homogenized in a 15 ml Falcon tube. Afterwards, the samples were kept at 4ºC overnight. On the following morning, the samples were centrifuged at 13,000 g for

15 minutes at 4ºC, and the pellets washed in 5 ml of fresh buffer H under the same conditions of centrifugation. Then, the pellets were resuspended in 3 ml of buffer R (50 mM sodium acetate pH 5.0, 1% SDS, in DEPC-treated water) with the help of a vortex. Finally, RNA was purified by phenol:chloroform:isoamyl alcohol extraction and ethanol precipitation as in step 3.5.1.

Purified RNA samples from either commercial kits or the LiCl-urea protocol were eluted/resuspended in DEPC-treated water and stored at -80ºC until needed.

As a final consideration, the type of tissue being processed also conditioned the type of disruption and homogenization method chosen. Mouse eyes and brain were disrupted and homogenized using a Polytron (Ultra-Turrax T8, Ika). In contrast, skin tissue was disrupted with a mortar and pestle under liquid nitrogen, and subsequently homogenized with a QIAshredder column (QIAGEN).

3.8.2. RNA Quantification and Purity Assessment

RNA quality and quantity were analyzed in a NanoDrop™ ND-1000 Spectrophotometer (Thermo Scientific). The same principles as in the case of DNA quantification (section 3.2.2.7) apply here. However, there are some differences. For example, RNA concentration is estimated considering that an optical density of one unit at 260 nm correlates to a concentration of 40 µg/ml RNA. Also, good quality RNA exhibits a ratio 260/280 of around 2.0 and a ratio 260/230 in the range of 2.0-2.2. Finally, the lower and upper detection limits of RNA for the device are 2 and 3000 ng/µl, respectively.

In addition, aliquots of purified RNA were always run in gel electrophoresis to visually assess their quality.

3.8.3. RNA Reverse Transcription

Usually, 20 µg of purified RNA from each sample were treated with 10 U of DNAse I (Roche) at 37ºC for 20 minutes. The inactivation of the enzyme was achieved by heating the samples at 90ºC for 10 minutes. Then, 2 µg of each product were mixed with 1 µl of dNTPs (10 mM, Roche) and random hexamers (3 µg/µl, Invitrogen; 0.5 µg hexamers/ µg total RNA) in a final volume of 12.5 µl. The mixtures were heated to 65ºC for 5 minutes and cooled on ice for 5 minutes so as to remove secondary structures. Afterwards, 4 µl of

First Strand Buffer 5x (Invitrogen), 2 µl of DTT (0.1 M, Invitrogen) and 1 µl of RNase OUT (Invitrogen) were added to each sample, which were subsequently incubated at 25ºC for 10 minutes and at 37ºC for 2 minutes. Finally, 100 U of SuperScript III reverse transcriptase (Invitrogen) were added in all cases, and the samples were incubated first at 37ºC for 50 minutes, and then at 70ºC for 15 minutes to inactivate the enzyme.

3.8.4. SYBR Green Quantitative PCR

The expression of the cluster of genes coding for the desmosomal proteins that reside in the mouse chromosome 18 was evaluated in COCA cells at several differentiation stages (0h, 24h, 48h and 72h). The L929 cell line, which lacks desmosomes, was used as a non-expressing negative control. The sequence of the primers can be found in **Appendix MM-1**.

In accordance with standard protocols (Gimenez et al., 2003), the concentrations of both forward and reverse primers were optimized to prevent the formation of dimers. In most cases, the best combination was found after testing the following concentrations for each primer: 500 nM, 250 nM and 125 nM in all possible combinations (9 in total). However, in some cases, the concentration of a given primer had to be set as low as 62.5 nM or 31.25 nM.

Furthermore, for all genes to be tested, a standard curve with a serial dilution of cDNA template was generated in order to evaluate the efficiency of the qPCR and to ensure that the working concentrations of the samples fell within the linear range of amplification. qPCR efficiencies between 85 and 115% were accepted.

qPCRs were conducted in duplicates and repeated twice, using SYBR Green PCR Master Mix from Applied Biosystems in a 7500 Real-Time PCR System, according to the provider's manual. The amplification program utilized was as follows: 2 minutes at 50ºC, 10 minutes at 95ºC, and finally, 40 cycles of 15 seconds at 95ºC and 1 minute at 60ºC. The relative standard curve method served to analyze the results. Expression data for all genes were normalized to that of *Gapdh* so as to be able to compare different samples. Normalized results for each gene in each cell line were further relativized to those from undifferentiated COCA cells in order to evaluate the changes in gene expression along the course of differentiation.

3.9. Chromosome Conformation Capture (3C)

Chromosome conformation capture is a molecular biology technique that uncovers long-range interactions between genomic loci that lie far from each other in the linear genome in the same, or even in a different, chromosome. The 3C protocol performed here followed that of Hagege et al., 2007 and Tena et al., 2011.

3.9.1. Crosslinking and Cell Lysis

L929 and COCA cells were cultured in p100 plates, as described in section 3.3. On the first day of the 3C protocol, the plates were washed with room temperature PBS. Then, 3 ml of fresh PBS were carefully added to the plates to prevent the cells from detaching. Next, 3 ml of freshly prepared 4% paraformaldehyde PFA in PBS (2% final concentration; Merck) were added to each plate, which were then incubated at RT for 10 minutes while gently shaking. The crosslinking reaction was quenched by adding cold 1M glycine (125 nM final concentration; Merck) to the plates and incubating them for 5 minutes while gently shaking.

The cells were detached with a scraper, and collected in cold 15 ml Falcon tubes in such a way that each tube contained approximately 10^7 cells. Subsequently, the samples were centrifuged at 1,300 rpm at 4ºC for 8 minutes, and the supernatants removed. The pellets were resuspended in 5 ml of cold lysis buffer (50 mM Tris-HCl pH 8, 150 mM NaCl, 5 mM EDTA, 0.5% NP-40, 1% Triton X-100, 1x complete protease inhibitors from Roche) and placed on ice for 10 minutes to allow the cells to swell. At this point, the cells were transferred to a 15 ml tissue grinder (Fisher Scientific) and homogenized (pestle A) on ice every 10 minutes until the cells were completely lysed, yet with intact nuclei. Lysis efficiency was assessed every 30 minutes by mixing 3 µl of cells with an equal volume of methyl green-pyronin staining (Sigma-Aldrich) on a microscope slide and checking the mixture under the microscope: cytoplasm stained pink whereas nuclei stained blue/green.

Differentiated COCA cells formed very intricate monolayers, difficult to lyse. Before using the grinder, they were disentangled by sonication under the following conditions: 30% of amplitude and 6 cycles of 10 seconds on / 10 seconds off on ice, in a Vibra-Cell™ VC50 Ultrasonic Processor (Sonics and Materials, Inc.) with a 3 mm microtip probe.

Once the majority of the nuclei had been released, the samples were centrifuged for 5 minutes at 1,800 rpm at 4ºC, and the pellets resuspended in 450 µl of mQ-H_2O.

3.9.2. Enzymatic Digestion of Fixed Chromatin

After the addition of 60 µl of 10x digestion buffer B (Roche) and 15 µl of 10% SDS, the samples were incubated at 37ºC for 1 hour while shaking at 900 rpm in a Thermomixer to favor the lysis of the nuclei. Then, 75 µl of 20% Triton X-100 were added to sequester the SDS and allow the subsequent enzymatic digestion. The samples were incubated, again, at 37ºC for 1 hour while shaking at 900 rpm. As the undigested control, 10 µl aliquots were taken from each cell type, and stored at 4ºC until processed in parallel with the digested controls later on. The restriction digestion took place in three consecutive steps. First, 200 U of *Hind*III (Roche) were added to the samples, which were then maintained at 37ºC for 3 hours while shaking. Second, a supplement of 200 more units of enzyme was added to the samples, which were incubated overnight under the same conditions. Finally, 200 U of *Hind*III were added again to the samples the next morning. After 4 hours of incubation at 37ºC while shaking at 900 rpm, aliquots of 10 µl were taken as digested controls.

Digestion efficiency was determined at this point. The control aliquots were taken up to a volume of 95 µl, using 10 mM Tris-HCl pH 7.5. Next, 5 µl of proteinase K (10 mg/ml, Roche) were added to each sample. They were incubated at 65ºC for 1 hour to reverse the crosslinks and the DNA was purified by phenol:chloroform:isoamyl alcohol extraction. The purified undigested and digested controls for each cell type were run on a 0.6% agarose gel and compared. If digestion had taken place, the samples were ready for the next step.

3.9.3. Ligation of Digested Fixed Chromatin

To inactivate the restriction enzyme, the digested samples were heated at 65ºC for 20 minutes and transferred to a 50 ml Falcon tube. A volume of 5.7 ml of mQ-H_2O was added to the samples, together with 700 µl of 10x ligase buffer (300 mM Tris-HCl pH 7.8, 100 mM $MgCl_2$, 100 mM DTT, 10 mM ATP, Promega) and 30 U of T4 DNA ligase (3 U/µl, Promega). They were incubated overnight at 16ºC.

To determine ligation efficiency, aliquots of 100 µl were taken and processed as in the previous section. Purified ligated samples were run along undigested and digested controls and compared. If ligation had occurred as expected, the samples were further processed.

3.9.4. Reversal of the Crosslinks and DNA Purification

The rest of the samples were treated with 30 µl of proteinase K (10 mg/ml, Roche) at 65ºC overnight to reverse the crosslinks. On the following day, RNA was removed from the samples by incubation with 30 µl of RNase A (10 mg/ml, Roche) at 37ºC for 45 minutes. Finally, DNA was purified by phenol:chloroform:isoamyl alcohol extraction followed by ethanol precipitation as in section 3.5.1. The pellets were resuspended in 150 µl of 10 mM Tris-HCl pH 7.5 at 37ºC for 30 minutes followed by an additional step of 14-16 hours at 4ºC. Purified samples were quantified, visually analyzed by gel electrophoresis and stored at -20ºC until needed.

3.9.5. SYBR Green Quantitative PCR of Ligated Products

A set of primers was designed specifically for all *Hind*III fragments that contained CTCF-cohesin sites in the locus of interest. Additional primers targeting genomic regions ± 30 kb from the CTCF-cohesin sites were included as negative controls of interaction (the properties of all primers can be found in **Appendix MM-1**).

The linear range of amplification was determined by performing qPCR on serially diluted BAC control and 3C sample templates, with different primer pairs. Once the working DNA concentration range had been established, qPCRs were conducted, in duplicates, with a fixed primer located on the CTCF-cohesin site #4 and the rest of the primers in the region. Enrichment in a given ligation product was taken as a measure of the interaction frequency between two genomic loci. Hence, relative interaction frequencies were calculated by using a standard curve generated from a serial dilution of the BAC mix control template (section 3.9.6), which contained all possible ligation products in equimolar amounts. To enable the comparison of the results from different 3C samples, the interactions observed in each cell line were normalized to that of the internal *Ercc3*-BAC control (section 3.9.6), a locus that is considered to adopt the same spatial conformation regardless of cell type (De Laat & Grosveld, 2003; Palstra et al., 2003; Drissen et al., 2004).

All experiments were repeated at least twice, using SYBR Green PCR Master Mix from Applied Biosystems in their 7500 Real-Time PCR System, following standard protocols.

3.9.6. BAC Control Template Preparation

A positive control template was needed to normalize the PCR efficiencies of the different primer pairs. This template contained all possible ligation products in equimolar amounts and was generated by randomly digesting and ligating five bacterial artificial chromosomes (BACs) that covered the region of interest in the mouse chromosome 18. The following BACs were obtained from the BACPAC Resources Center (https://bacpac.chori.org/resources.htm) at Children's Hospital Oakland Research Institute in Oakland (California, USA): RP23-471C20 (CTCF-cohesin site #1), RP23-26H5 (CTCF-cohesin site #2), RP23-357D18 (CTCF-cohesin sites #3 and 4), RP23-183F16 (CTCF-cohesin site #5), RP23-44I15 (CTCF-cohesin site #6). An additional BAC that contained the mouse *Ercc3* locus was used to produce an internal control that enabled the comparison of different samples: RP23-148C24.

Upon arrival, BAC clones were streaked in LB plates supplemented with 30 mg/ml chloramphenicol (Sigma-Aldrich) and kept overnight at 37ºC. Up to three colonies per construct were picked and grown overnight in 5 ml of LB supplemented with the same antibiotic. BAC minipreparations were performed (see section 3.9.6.1) and the clones were checked for the presence of the expected sequences by PCR amplification (see the primers used in **Appendix MM-1**). Once the correct clones for each BAC had been identified, maxipreparations were done using the Large-Construct kit from QIAGEN.

All purified BACs (except for that corresponding to the *Ercc3* locus) were carefully quantified by gel densitometry, mixed in equimolar amounts and digested (5 µg of total DNA) in a final volume of 500 µl with 200 U of *Hind*III at 37ºC overnight. In parallel, 1 µg of the BAC containing the *Ercc3* locus was digested under the same conditions. After confirming digestion by gel electrophoresis, the samples were purified by phenol:chloroform:isoamyl acid extraction and ethanol precipitation, and subsequently ligated in a final volume of 100 µl with 6 U of T4 DNA ligase (Promega) at 16ºC overnight. Again, ligation efficiency was estimated by gel electrophoresis and the samples were finally purified as in the previous step. The DNA pellets corresponding to the BAC mix template and the *Ercc3*-BAC template were resuspended in 100 µl of mQ-H_2O.

3.9.6.1. Minipreparations of BACs from *E. coli*

An adequate number of clones per construct were picked and grown in 5 ml of LB medium supplemented with chloramphenicol (30 µg/ml) at 37ºC overnight at 200-

250 rpm. The cells were centrifuged at 10,000 g for 1 minute and the pellets were gently resuspended in 100 µl of prechilled Solution I (50 mM glucose, 25 mM Tris-HCl pH 7.5, 10 mM EDTA pH 8). Lysis was made possible by adding 200 µl of freshly prepared Solution II (0.2N NaOH, 1% SDS) and incubating the samples for up to 5 minutes at RT. Next, lysis was neutralized with 150 µl of Solution III (3M AcK, 11.5% v/v glacial acetic acid) and 10 minutes in ice. Cell debris in each sample were pelleted by centrifugation at top speed for 6 minutes at 4ºC. The DNA-containing supernatants were transferred to new tubes and the DNA was purified by ethanol precipitation. The pellets were finally resuspended in 30 µl of mQ-H_2O.

3.10. Primers

Appendix MM-1 collects all the primers used in this work, including those employed for cloning, genotyping, qPCR, ChIP or 3C experiments.

4 RESULTS

4.1. Development of Algorithms to Predict the Presence of Boundaries Separating Genes with Different Expression Patterns: First Case Scenario

The expression of most genes is tightly regulated by the action of different types of regulatory elements, such as enhancers or silencers. Genes that are very close to each other in the linear genome often exhibit similar expression profiles, simply because they fall within the range of action of the same set of regulatory elements (Hurst et al., 2004). Under our working hypothesis, the fact that two adjacent genes, unlike expected, show completely opposite expression profiles is indicative of the existence of an insulator between them.

4.1.1. Gene Expression Data Retrieval and Analysis

Using the BioMart tool from Ensembl, genomic data (GRCm37; annotation release 59) for all 36,613 mouse genes stored in the database were extracted, including their coordinates and their orientation with respect to the reference sequence. The list was then filtered to retain only those genes that had been curated and registered in MGI, the international database resource for the laboratory mouse, created and maintained by The Jackson Laboratory (http://www.informatics.jax.org/). As a result, a total of 21,161 genes were considered for further study. When the genes were arranged in pairs, chromosome by chromosome, 25,048 pairs of adjacent genes emerged. Note that the number of resulting pairs is larger than expected (number of genes minus one per chromosome). This seeming contradiction stems from the method applied to generate pairs in the cases of overlapping genes (see the Materials and Methods section for more details).

Gene expression data at the mouse adult stage for a total of 425 different tissues were obtained from the aGEM Platform v2.0 (Jimenez-Lozano et al., 2009)[2]. This online tool integrates the information already stored in several gene expression databases, information that mainly comes from *in situ* hybridization techniques and microarrays. The databases included in aGEM v2.0 are: EMAGE (http://www.emouseatlas.org/emage/), GXD (http://www.informatics.jax.org/expression.shtml), GENSAT (http://www.gensat.org/index.html), BioGPS (http://biogps.org/#goto=welcome), ABA (http://mouse.brain-map.org/) and EUREXPRESS (http://www.eurexpress.org/ee/). Of note, GENSAT and ABA only host gene expression data of the central nervous system.

[2] A revised version of the aGEM Platform, v3.1, has been recently released (Jimenez-Lozano et al., 2012).

Each of these databases considers different levels of gene expression (from undetectable to very strong) and codifies them differently, which impedes the direct integration of the data into a single database. However, aGEM solves this problem by combining all the data into three different "expression strength" levels: 'zero' for undetectable or very low gene expression, 'one' for moderate gene expression and 'two' for high gene expression. Importantly, there is a lack of gene expression information available for some genes.

Distance measures are the mathematical parameters used to compare how similar or divergent the expression profiles of two genes are. None of the various measures currently available outperforms the rest in all situations. However, two of them are preferentially used: Pearson's correlation and Euclidean distance (D'haeseleer et al., 2000). Hence, two algorithms, each of them based on one of these distance measures, were developed making use of the gene expression data obtained from aGEM, in order to identify those divergently expressed adjacent gene pairs in the mouse genome.

4.1.1.1. Correlation Method

aGEM already implements Pearson's correlation analyses. This feature enables the comparison of the expression profiles of two genes in two different scenarios: across a panel of anatomical structures for a given developmental stage (**Fig. R 1**), or at a certain developmental stage after having selected a specific anatomical structure.

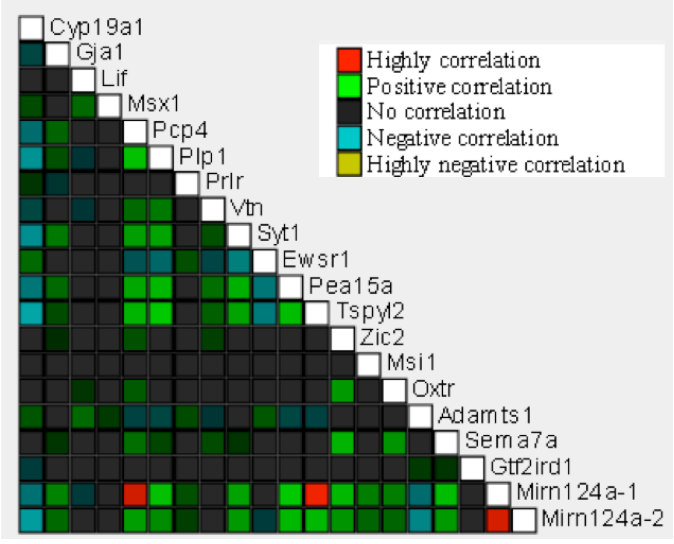


Fig. R 1. Pearson's correlation analysis in aGEM. The expression profiles at the adult stage (Theiler Stage TS28) of a selected set of mouse genes were analyzed in aGEM. Genes with very similar expression patterns were positively correlated (green or red squares for moderate or high positive correlation, respectively), while genes whose expression patterns were opposite showed negative correlation coefficients (blue or yellow squares, depending on the strength of the negative correlation).

In this case, the expression profiles at the adult stage for all pairs of adjacent genes were compared across 425 different tissues (**Appendix R-1**), yielding 1,212 pairs of negatively correlated genes (p-value < 0.05, **Table R-1**). Biologically, this finding means

Table R 1. Top 50 pairs of genes that potentially contain boundary elements derived from the Pearson's correlation method.

ID	Gene Name 1	Gene Name 2	Intergenic Distance (bp)	Promo-ters[1]	Genomic Coordinates (GRCm37)	Pearson's Coefficient	p-value
MBC0001	*Scn3a*	*Scn2a1*	53230	D	Chr2:65,296,320-65,605,504	-1	0
MBC0002	*Hcn3*	*Clk2*	4599	D	Chr3:88,949,996-88,980,843	-1	0
MBC0003	*Zfp384*	*Ing4*	1895	SO	Chr6:124,959,163-125,001,517	-1	0
MBC0004	*Ing4*	*Lpar5*	16421	SO	Chr6:124,989,778-125,032,490	-1	0
MBC0005	*Nwd1*	*Sin3b*	8558	SO	Chr8:75,170,610-75,281,304	-1	0
MBC0006	*Cd200r1*	*Cd200r2*	72119	SO	Chr16:44,765,849-44,915,953	-1	0
MBC0007	*D17Wsu92e*	*Snrpc*	19416	D	Chr17:27,888,177-27,988,913	-1	0
MBC0008	*Memo1*	*Dpy30*	4665	SO	Chr17:74,600,046-74,715,761	-1	0
MBC0009	*Bcas2*	*Trim33*	100127	SO	Chr3:102,975,578-103,162,691	-0.953	0
MBC0010	*Nanos1*	*Eif3a*	1209	C	Chr19:60,831,890-60,866,546	-0.948	0
MBC0011	*Glg1*	*Rfwd3*	11728	SO	Chr8:113,681,460-113,824,122	-0.946	0
MBC0012	*Nt5dc1*	*Tspyl4*	-121131	D	Chr10:34,008,094-34,021,114	-0.940	0
MBC0013	*BC057079*	*Ptplad2*	-14867	C	Chr4:87,740,533-88,084,832	-0.939	0
MBC0014	*Lpar6*	*Itm2b*	122632	C	Chr14:73,637,807-73,785,096	-0.927	0
MBC0015	*Srgap2*	*Fam72a*	542	D	Chr1:133,181,828-133,436,449	-0.918	0
MBC0016	*Eif3a*	*Fam45a*	20942	D	Chr19:60,837,025-60,912,130	-0.914	0
MBC0017	*Ublcp1*	*Rnf145*	48466	D	Chr11:44,268,073-44,379,022	-0.902	0
MBC0018	*Paics*	*Srp72*	7174	SO	Chr5:77,380,332-77,428,962	-0.901	0
MBC0019	*Arhgef37*	*Csnk1a1*	31347	D	Chr18:61,653,448-61,749,035	-0.890	0
MBC0020	*Cdkl2*	*G3bp2*	11472	SO	Chr5:92,435,100-92,512,684	-0.886	0
MBC0021	*Rpl22*	*Chd5*	4580	SO	Chr4:151,699,851-151,764,303	-0.869	0
MBC0022	*Tsn*	*Mki67ip*	10669	D	Chr1:120,194,732-120,230,399	-0.860	2.22E-16
MBC0023	*Ric8*	*Psmd13*	18238	SO	Chr7:148,042,856-148,084,541	-0.860	2.22E-16
MBC0024	*Phf8*	*Huwe1*	166948	SO	ChrX:147,955,215-148,369,960	-0.884	4.44E-16
MBC0025	*Gtf2h1*	*Ldha*	17675	SO	Chr7:54,051,473-54,110,997	-0.862	4.44E-16
MBC0026	*Thsd4*	*Lrrc49*	46813	SO	Chr9:59,814,738-60,535,965	-0.855	4.44E-16
MBC0027	*Hdlbp*	*Sept2*	33312	D	Chr1:95,304,343-95,406,837	-0.859	6.66E-16
MBC0028	*Rpl35*	*Arpc5l*	2452	D	Chr2:38,857,100-38,871,397	-0.859	6.66E-16
MBC0029	*Sgk3*	*6030422M02Rik*	8790	SO	Chr1:9,788,234-9,932,166	-0.854	8.88E-16
MBC0030	*Csde1*	*Nras*	154	SO	Chr3:102,824,530-102,871,837	-0.866	1.11E-15
MBC0031	*Psmc3*	*Sfpi1*	29727	SO	Chr2:90,894,166-90,955,913	-0.856	1.11E-15
MBC0032	*Zfyve20*	*Trh*	27136	SO	Chr6:92,136,706-92,194,644	-0.850	1.55E-15
MBC0033	*1700034F02Rik*	*Rtn4*	114580	SO	Chr11:29,447,950-29,644,331	-0.875	2.00E-15
MBC0034	*Rpl7a*	*Surf2*	3049	SO	Chr2:26,766,284-26,775,703	-0.857	3.77E-15
MBC0035	*Rab11fip3*	*Decr2*	11802	SO	Chr17:26,125,981-26,227,274	-0.878	5.33E-15
MBC0036	*Atl2*	*Hnrpll*	134220	SO	Chr17:80,247,732-80,461,608	-0.841	2.09E-14
MBC0037	*Mgrn1*	*Nudt16l1*	827	SO	Chr16:4,886,317-4,941,019	-0.840	4.40E-14
MBC0038	*1600027N09Rik*	*Ogfr*	2941	SO	Chr2:180,317,417-180,330,541	-0.824	5.84E-14
MBC0039	*Tmem85*	*2410042D21Rik*	11184	D	Chr2:112,203,168-112,254,397	-0.832	7.24E-14
MBC0040	*AW549877*	*Oxct1*	30637	D	Chr15:3,934,434-4,105,344	-0.845	7.77E-14
MBC0041	*Gm732*	*Brwd3*	788551	SO	ChrX:105,141,074-106,029,711	-0.829	9.93E-14
MBC0042	*9930021D14Rik*	*Caskin1*	13314	SO	Chr17:24,610,829-24,645,850	-0.833	1.14E-13
MBC0043	*Snapin*	*2500003M10Rik*	7923	SO	Chr3:90,291,948-90,313,420	-0.811	3.09E-13
MBC0044	*Rnf14*	*Gnpda1*	9716	C	Chr18:38,456,348-38,498,657	-0.819	3.57E-13
MBC0045	*Tsen34*	*Rps9*	2970	SO	Chr7:3,644,977-3,658,503	-0.829	6.28E-13
MBC0046	*Gorasp2*	*Mettl8*	251925	C	Chr2:70,499,633-70,893,640	-0.814	6.99E-13
MBC0047	*Set*	*Zdhhc12*	18367	C	Chr2:29,912,898-29,949,168	-0.813	7.25E-13
MBC0048	*Cltc*	*Dhx40*	11281	SO	Chr11:86,507,853-86,621,198	-0.873	8.42E-13
MBC0049	*Gar1*	*Cfi*	5343	D	Chr3:129,527,830-129,578,246	-0.809	1.12E-12
MBC0050	*Tcp11l2*	*Polr3b*	8078	SO	Chr10:84,039,371-84,189,922	-0.809	1.16E-12

[1]**Promoters**. Configuration of the promoters of the pair: D, divergent; C, convergent; SO, same orientation.

Highlighted in **light blue**, two of the pairs selected for functional validation (see section 4.1.4.1. below).

that over a thousand of the total number of original pairs possessed opposite expression patterns; that is, whenever one of the genes is expressed, the other is not (**Fig. R 2**). Therefore, this subset of gene pairs potentially contained boundary elements.

Of note, a tissue was only considered in each pairwise comparison, if there was expression information available for both genes of the pair. Hence, the fact that all genes lacked some expression data points prevented from obtaining the correlation analysis across the full panel of tissues (**Appendix R-1**). This implies that the subset of anatomical structures used to conduct the correlation analysis differed in each pairwise comparison, which is one of the main limitations of this approach.

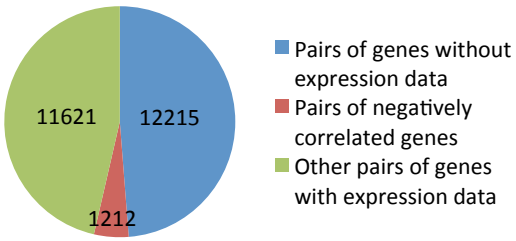


Fig. R 2. Pairs of genes that potentially contain boundary elements derived from the correlation method.

Moreover, there were many genes with expression information for just a few tissues. However, the algorithm required a minimum amount of data to perform the analysis. This requirement was not met for about half of the pairs assayed (12,215 pairs of a total of 25,048; **Fig. R 2**). Therefore, this method represents an underestimation and thus, probably overlooks many pairs of genes with actual inverse expression patterns in which insulators may exist.

4.1.1.2. Euclidean Distance Method

An additional algorithm was developed in order to better exploit the data stored in aGEM: the Euclidean Distance Method. In this case, the list of tissues included in the analysis was simplified after the realization that there were two clearly distinguishable groups within the data. The first group was composed of those tissues with gene expression data for approximately 60% of all the genes. On the contrary, the second group gathered many highly specialized anatomical substructures such as, for instance, the six layers of the cerebral cortex. These tissues only had expression data for a very limited subset of genes: those that have been thoroughly studied in close detail. Even if some genes are indeed expressed divergently in these substructures, our algorithm focuses on finding pairs of genes pertaining to substantially different regulatory pathways (i.e. ubiquitous *versus* brain-specific genes) as our first approach. The identification of gene pairs whose expression profiles only slightly vary due to the fine regulation of the same

pathway is beyond the scope of this work, since the presence of strict boundary elements would probably not be needed at those locations. Hence, this second group of tissues was withdrawn from the analysis. This caused the original list of considered tissues to shrink from 425 to 52 records (**Appendix R-1**).

Still, there were many missing gene expression data points. The Euclidean distance algorithm demanded data for all genes in all tissues examined. The reason is that statistical calculations that help to decide whether the expression patterns of the genes in a pair differ significantly, had to be done considering the entire dataset, and not only the information of the pair in question (see below).

The consequences of this requirement were two-fold. First, the same panel of tissues was used for all the analyses –unlike in the correlation method described above. Second, an *imputation* step was needed to rescue missing data: each lost *datum* in a given anatomical structure was replaced by the average gene expression value of the flanking genes (**Table R 2**). This conservative approach assumed that genes that lie next to each other in the genome exhibit the same expression pattern. Since the algorithm developed here sought to find, precisely, gene pairs that contradict this trend, false positives due to the imputation process were unlikely to present.

Table R 2. Imputation of missing values. An example of imputed values (blue and bold) for five genes of chromosome 18 and three different tissues is shown. 'Ones' and 'zeros' indicate whether a gene is expressed or not in a given tissue, respectively. Values of '0.5' can result from the imputation process.

Gene	Adipose tissue	Adrenal gland	Amygdala
Map3k8	0	0	0
Mtpap	**0.5**	1	0
9430020K01Rik	1	1	0
Gm10556	**1**	**1**	**0.5**
SvilI	1	1	1

Next, the algorithm conducted pairwise comparisons between the expression profiles of any two genes -not necessarily adjacent- using the Euclidean distance as the distance measure (**Fig. R 3**). All of the genes in a given chromosome were compared with all the other genes in that same chromosome. The results were integrated into distance matrices in such a way that each row (or column) contained the distribution of the expression distances between a given gene and all the rest, regardless of their position in the chromosome. Then, the Euclidean distance between two adjacent genes was only considered statistically significant if it was significantly high in the context of the

distributions of each of the genes separately (see the Materials and Methods section for further details).

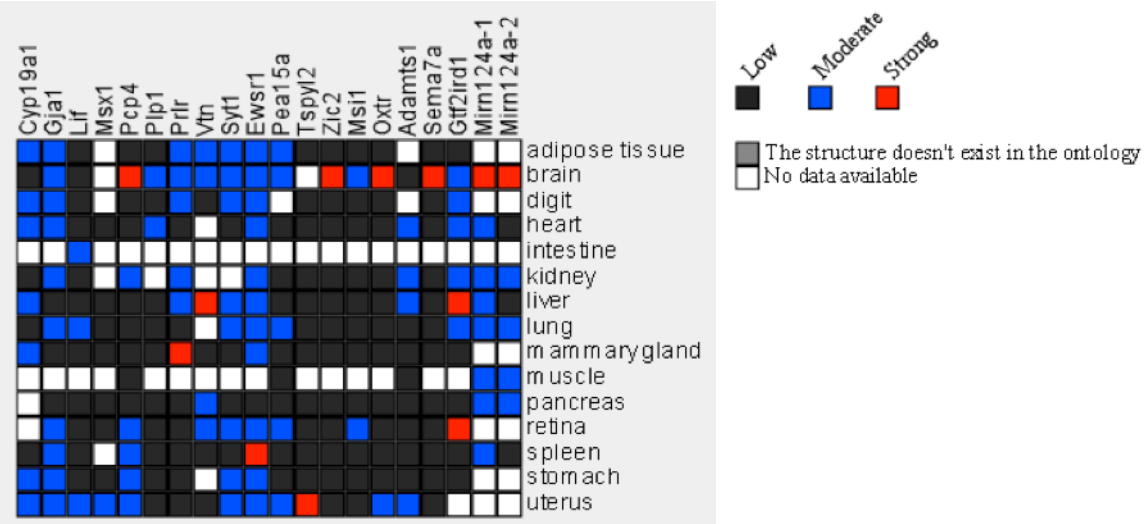


Fig. R 3. Gene expression profiles in aGEM. The color-coded matrix represents the expression profiles of a set of mouse genes (in columns) in a panel of anatomical structures (in rows) at the adult stage (TS28). Black, blue or red squares are assigned to genes with low (or none), moderate or strong expression in a given tissue, respectively.

This method predicted the presence of boundary elements in 6,294 pairs of adjacent genes (**Fig. R 4**; see **Table R 3**). Furthermore, for thirty-six such pairs, the gene expression distance between the members of the pair was maximum; that is, their expression profiles were completely opposite.

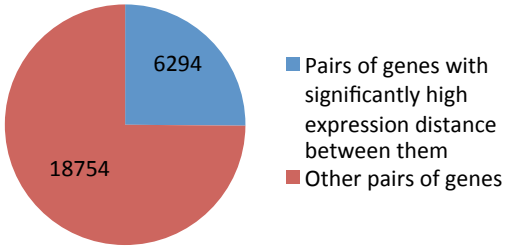


Fig. R 4. Pairs of genes that potentially contain boundary elements derived from the Euclidean distance method.

4.1.2. *In Silico* Validation

The majority of vertebrate insulators described so far contain CTCF binding sites. So as to assess the predictive value of the algorithms, the pairs obtained from each of them were scanned for the presence of previously known insulator-related sequences such as CONSYN CTCF sites (Martin et al., 2011). These sites are a special class of CTCF binding sites that are bound by the protein in all cell types analyzed for a given species (CONstitutive), and occupy syntenic positions in the mouse, human and chicken genomes (SYNtenic).

Table R 3. Top 50 pairs of genes that potentially contain boundary elements derived by the Euclidean distance method.

ID	Gene Name 1	Gene Name 2	Intergenic Distance (bp)	Promo-ters[1]	Genomic Coordinates (GRCm37)	Distance Coeffi-cient	Maximum possible value?[2]
MBD0001	*Eef1b2*	*Gpr1*	2205	C	Chr1:63,223,399-63,261,117	7.211	Max (for both)
MBD0002	*Strc*	*Pdia3*	26607	D	Chr2:121,189,464-121,264,423	7.211	Max (for both)
MBD0003	*Pdia3*	*Serinc4*	372	C	Chr2:121,239,511-121,282,517	7.211	Max (for both)
MBD0004	*Cep152*	*Eid1*	48014	D	Chr2:125,388,824-125,499,837	7.211	Max (for both)
MBD0005	*Shc4*	*Eid1*	-51021	D	Chr2:125,453,183-125,499,837	7.211	Max (for both)
MBD0006	*Tmc2*	*Idh3b*	14864	C	Chr2:130,020,930-130,110,283	7.211	Max (for both)
MBD0007	*9230104L09Rik*	*Cst3*	6684	SO	Chr2:148,672,447-148,701,428	7.211	Max (for both)
MBD0008	*Hdgf*	*BC023814*	6469	C	Chr3:87,710,243-87,734,639	7.211	Max (for both)
MBD0009	*1700022I11Rik*	*Vcp*	-3676	C	Chr4:42,982,818-43,013,379	7.211	Max (for both)
MBD0010	*2210012G02Rik*	*Tmem59*	-717	D	Chr4:106,806,760-106,873,601	7.211	Max (for both)
MBD0011	*Fam167b*	*Eif3i*	13411	SO	Chr4:129,254,059-129,277,892	7.211	Max (for both)
MBD0012	*1810019J16Rik*	*Nudc*	1752	C	Chr4:133,074,878-133,101,911	7.211	Max (for both)
MBD0013	*Wdr8*	*Tprgl*	-9935	C	Chr4:153,516,481-153,534,775	7.211	Max (for both)
MBD0014	*Spon2*	*Ctbp1*	29571	SO	Chr5:33,556,167-33,617,610	7.211	Max (for both)
MBD0015	*Chchd2*	*Scand3*	8253	D	Chr5:130,357,027-130,379,493	7.211	Max (for both)
MBD0016	*Ywhag*	*Srcrb4d*	25595	SO	Chr5:136,384,279-136,450,401	7.211	Max (for both)
MBD0017	*Tmem106b*	*Vwde*	96344	C	Chr6:13,019,759-13,174,965	7.211	Max (for both)
MBD0018	*Plxna4*	*Chchd3*	202784	SO	Chr6:32,095,926-33,010,260	7.211	Max (for both)
MBD0019	*Dnhd1*	*Ilk*	41784	SO	Chr7:112,789,076-112,891,439	7.211	Max (for both)
MBD0020	*Rrp8*	*Ilk*	-784	D	Chr7:112,880,721-112,891,439	7.211	Max (for both)
MBD0021	*Mapk3*	*Gdpd3*	595	SO	Chr7:133,903,115-133,919,157	7.211	Max (for both)
MBD0022	*6330512M04Rik*	*Ctsd*	-2605	SO	Chr7:149,511,742-149,573,943	7.211	Max (for both)
MBD0023	*Ctsd*	*Syt8*	46812	D	Chr7:149,557,053-149,626,301	7.211	Max (for both)
MBD0024	*Yjefn3*	*Ndufa13*	3221	SO	Chr8:72,411,687-72,425,547	7.211	Max (for both)
MBD0025	*Rad23a*	*Calr*	1185	SO	Chr8:87,357,918-87,370,833	7.211	Max (for both)
MBD0026	*Ppib*	*Snx22*	-1445	C	Chr9:65,908,062-65,917,538	7.211	Max (for both)
MBD0027	*Shisa5*	*Atrip*	209	C	Chr9:108,941,079-108,976,638	7.211	Max (for both)
MBD0028	*Rps26*	*Ikzf4*	4343	SO	Chr10:128,061,593-128,083,049	7.211	Max (for both)
MBD0029	*2310033P09Rik*	*Arf1*	676	C	Chr11:59,021,823-59,041,772	7.211	Max (for both)
MBD0030	*Sumo2*	*Nup85*	28158	D	Chr11:115,384,416-115,445,299	7.211	Max (for both)
MBD0031	*Snx31*	*Pabpc1*	40088	SO	Chr15:36,433,817-36,538,728	7.211	Max (for both)
MBD0032	*Tatdn1*	*Ndufb9*	80	D	Chr15:58,721,708-58,771,044	7.211	Max (for both)
MBD0033	*Vmn1r232*	*Ppp2r1a*	30975	D	Chr17:21,050,245-21,102,880	7.211	Max (for both)
MBD0034	*Memo1*	*Dpy30*	4665	SO	Chr17:74,600,046-74,715,761	7.211	Max (for both)
MBD0035	*Tll2*	*Tm9sf3*	7641	SO	Chr19:41,157,243-41,338,461	7.211	Max (for both)
MBD0036	*Morf4l2*	*Glra4*	13984	SO	ChrX:133,267,481-133,314,680	7.211	Max (for both)
MBD0037	*Ccin*	*Clta*	18919	SO	Chr4:43,996,376-44,045,718	7.141	Max (for one)
MBD0038	*Ube1y1*	*Kdm5d*	53564	SO	ChrY:155,092-280,254	4.95	Max (for one)
MBD0039	*Wdr38*	*Arpc5l*	1908	SO	Chr2:38,852,996-38,871,397	7.159	Max (for one)
MBD0040	*Apoa1bp*	*Ttc24*	10914	SO	Chr3:87,860,445-87,882,226	7.159	Max (for one)
MBD0041	*Gm973*	*Sumo1*	4926	C	Chr1:59,573,108-59,727,678	7.141	Max (for one)
MBD0042	*Yme1l1*	*4931423N10Rik*	8216	SO	Chr2:23,011,889-23,122,649	7.141	Max (for one)
MBD0043	*Slc6a17*	*Ubl4b*	36358	SO	Chr3:107,270,467-107,357,860	7.141	Max (for one)
MBD0044	*Psmc2*	*Slc26a5*	6868	C	Chr5:21,291,101-21,371,422	7.141	Max (for one)
MBD0045	*Cox6a1*	*4930430O22Rik*	87094	SO	Chr5:115,795,651-115,886,548	7.141	Max (for one)
MBD0046	*Styxl1*	*Mdh2*	95	D	Chr5:136,223,090-136,266,268	7.141	Max (for one)
MBD0047	*Cycs*	*4921507P07Rik*	6766	SO	Chr6:50,512,562-50,546,589	7.141	Max (for one)
MBD0048	*Arpc4*	*Rpusd3*	24871	C	Chr6:113,328,109-113,369,342	7.141	Max (for one)
MBD0049	*Npas1*	*Grlf1*	17693	SO	Chr7:17,041,071-17,200,342	7.141	Max (for one)
MBD0050	*Cox6b1*	*Etv2*	7499	SO	Chr7:31,401,994-31,421,308	7.141	Max (for one)

[1]**Promoters**. Configuration of the promoters of the pair: D, divergent; C, convergent; SO, same orientation.

[2]**Maximum distance**. The observed distance coefficient assigned to a pair is, sometimes, the maximum possible value. For example, in the case of the first pair MBD0001, the gene whose expression pattern differs the most from *Eef1b2* is actually its neighbor *Gpr1*; and *vice versa*.

Highlighted in **light blue**, four of the pairs selected for functional validation (see section 4.1.4.1. below).

The presence of B1-X35S, a subtype of B1 SINE retrotransposable elements, was also evaluated. These retrotransposons convey insulator activity dependent on the binding of the transcription factors dioxin receptor (AhR) and Slug (Snai2) (Roman et al., 2011a).

As a result, it was found that the pairs of genes obtained from both algorithms contain more of these two types of insulator elements than expected by chance (**Fig R 5**). Of note, more than half of the genes associated with either a CONSYN CTCF site or a B1-X35S retrotransposon appeared as hits of the Euclidean Distance Method (**Table R 4**). Besides, the algorithms also captured other genomic loci in which insulators reside, such as the *Wap* locus (**Table R 5**). However, even if enrichment in already described insulators was observed, the algorithms failed to detect all characterized cases (**Tables R 4** and **5**).

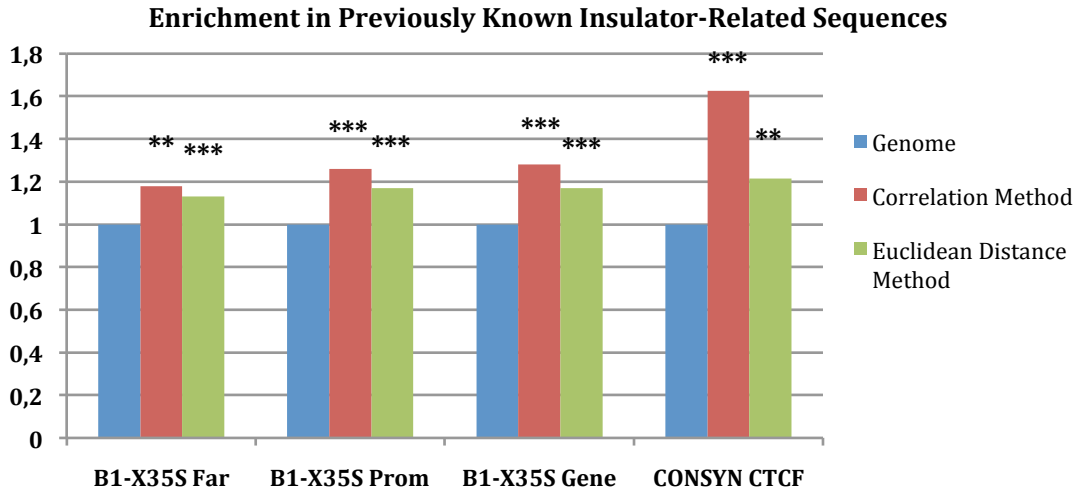


Fig. R 5. *In silico* validation of the quality of the algorithms. Both algorithms contained more previously known insulator-related sequences than expected stochastically.
B1-X35S Far, **Prom** and **Gene** refer to the presence of the B1 SINE retrotransposon B1-X35S near (± 10 kb), in the promoter or inside, at least, one of the genes of each pair (from the whole genome or derived from each of the methods), respectively, as described in (Roman et al., 2011a). **CONSYN CTCF** indicates the presence of constitutive and syntenic CTCF sites inside or near (± 20 kb) each type of pair, as described in (Martin et al., 2011). Data are shown as fold enrichment in each of the elements with respect to the genome. Fisher's Exact Test, ** significant at p-value < 0.01; *** significant at p-value < 0.001.

Table R 4. Percentage of B1-X35S and CONSYN CTCF sites covered by the algorithms.

	Correlation Method	Euclidean Distance Method
B1-X35S Far	12.37 %	51.32 %
B1-X35S Prom	13.65 %	53.13 %
B1-X35S Gene	13.35 %	53.49 %
CONSYN CTCF	21.05 %	57.49 %

Table R 5. Some, but not all, previously described insulators were captured by the algorithms. There was not enough gene expression information (No Data) to draw any conclusion for some pairs.

Insulator	Chromosome	5' Gene	3' Gene	Correlation Method	Euclidean Distance Method
adHS1	2	*Nr5a1*	*Nr6a1*	No	No
Evx2/Hoxd	2	*Evx2*	*Hoxd13*	No	No
Scl/Map17	4	*Pdzk1ip1*	*Cyp4x1*	Yes	Yes
Tcrβ/Trypsinogen	6	*Tcrβ*	*Prss2*	No	No
5'Tyr	7	*Tyr*	*Grm5*	No	No
Bglobin 5'HS5	7	*Olfr66*	*Hbb-y*	No Data	No
Bglobin 3'HS1	7	*Hbb-b2*	*Olfr68*	No Data	No
PCT12	7	*Mrpl23*	*Nctc1*	No	Yes
MS/DMD	7	*H19*	*Igf2*	No	No
KvDMR1	7	*Kcnq1*	-	No	No
PCTs	7	*Nctc1*	*H19*	No	Yes
SP-10	9	*A630095E13Rik*	*Acrv1*	No Data	No
Rasgrf1 DMD	9	-	*Rasgrf1*	No	No
5'Wap	11	*Tbrg4*	*Wap*	No	Yes
3'Wap	11	*Wap*	*Ramp3*	No	Yes
Gh	11	*Cd79b*	*Scn4a*	No	No
V(D)J rec	12	*Igh*	*hole*	No	No
Tcrα/Dad1	14	*Tcra*	*Dad1*	No	No
11P	17	*Rxrb*	*Col11a2*	No	No
Eif2s3x sites	X	*Eif2s3x*	*Klhl15*	No Data	Yes
Xist/Tsix	X	-	*Tsx*	No	Yes
RS14	X	*Tsx*	*Xist*	Yes	Yes
Jarid1c sites	X	*Kiaa0522*	*Jarid1c*	No Data	No

4.1.3. Comparison of the Methods

The number of hits obtained with each algorithm varied considerably. However, there was a noteworthy overlap between them (**Fig. R 6**). The number of pairs common to both algorithms was 690 (**Appendix R-2**), which accounted for around 57% and 11% of the total pairs resulting from the correlation and Euclidean distance methods, respectively.

The correlation method required a minimum amount of data to carry out the analysis. Importantly, it was found that 2,228 hits specific to the Euclidean distance method (40%, approximately) did not meet such requirement and hence, were not analyzed with the first algorithm. This suggests that the imputation process conducted in the second algorithm

enabled the extraction of valuable information from the scant data available for poorly studied genes, data that the correlation method had ignored.

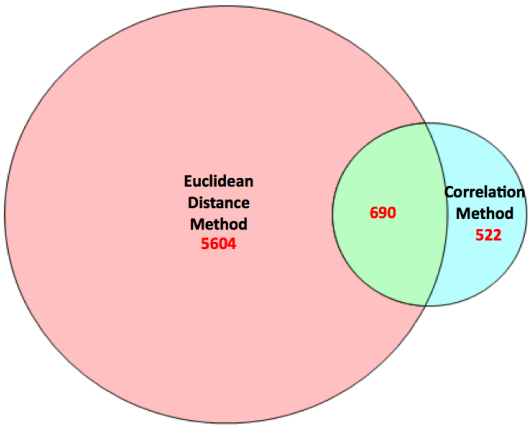


Fig. R 6. Venn diagram showing the overlap of the datasets. The Venn diagram presents the distribution of pairs of genes with significantly different expression profiles obtained from the correlation or Euclidean distance methods.

4.1.4. Functional Validation

4.1.4.1. Selection of Sequences to Test for Boundary Activity

In order to functionally validate both algorithms, several pairs of genes were selected according to diverse criteria (**Table R 6**). First, hits with short intergenic distances -ranging from 60 bp to 6 kb- were considered because it was hypothesized that genes with opposite expression patterns but, at the same time, very close in the linear genome, should be separated by powerful insulator sequences. This would be particularly true if their promoters lay right next to each other in divergent directions. In any case, in an attempt to integrate the diversity found in the pairs, hits with all possible promoter configurations -convergent, divergent or in the same orientation- were taken into account.

Second, pairs derived specifically from each of the algorithms, as well as pairs common to both, were chosen so as to evaluate the performance of each method. It was expected that the chances of finding insulators among the shared pairs would be higher.

Third, some pairs were included in the analysis because they were associated with known human diseases like Parkinson's Disease, and/or because they were involved in crucial cellular processes such as DNA repair or differentiation.

Finally, all pairs had statistically significantly different expression patterns across a panel of tissues; that is, they were obtained by one of the algorithms (or by both). This is the case of all pairs, except for pair number ten, which was analyzed for two reasons. First, as shall be explained below, it contained *Psen1*, and mutations in its human ortholog are the most frequent cause of early onset Alzheimer's Disease (Bekris et al., 2010), so

understanding its regulation may shed some light into the pathogenesis of the disease. Second, the Euclidean expression distance between the two genes bordered statistical significance, so it might have been a false negative.

Table R 6. Selected sequences to functionally validate the algorithms.

Pair	Genes	Derived from...	Chromosome	Intergenic Distance (bp)	Promoters[1]	Elements
1	*Atp2a1-Sh2b1*	Correlation Method	7	3886	SO	Cor-1
2	*Psmc5-Smarcd2*	Correlation Method	11	59	C	Cor-2.1 ; Cor-2.2
3	*Ftsj3-Psmc5*	Correlation Method	11	75	D	Cor-3
4	*Mapk3-Gdpd3*	Distance Method	7	595	SO	Dis-4.1 ; Dis-4.2 ; Dis-4.3 ; Dis-4.4
5	*Tatdn1-Ndufb9*	Distance Method	15	80	D	Dis-5.1 ; Dis-5.2 ; Dis-5.3 ; Dis-5.4 ; Dis-5.5
6	*Shisa5-Trex1*	Distance Method	9	209	C	Dis-6.1 ; Dis-6.2 ; Dis-6.3 ; Dis-6.4
7	*Memo1-Dpy30*	Both Methods	17	4665	SO	CorDis-7.1 ; CorDis-7.2 ; CorDis-7.3
8	*Tsen34-Rps9*	Both Methods	7	2970	SO	CorDis-8.1 ; CorDis-8.2 ; CorDis-8.3
9	*Ddost-Pink1*	Both Methods	4	781	C	CorDis-9.1 ; CorDis-9.2 ; CorDis-9.3
10	*Rbm25-Psen1*	-	12	5079	SO	10.1 ; 10.2 ; 10.3

[1]**Promoters**: Configuration of the promoters of the pair: D, divergent; C, convergent; SO, same orientation.

4.1.4.1.1. Functional Annotation of the Genes Selected for Testing Insulator Function

Atp2a1 and *Sh2b1* formed the first pair (**Fig. R 7A**). Mouse Atp2a1 (MGI:105058) is involved in calcium sequestration in muscular excitation/contraction processes. Its overexpression mitigates muscular dystrophy, and gene therapy strategies with this gene have been proposed as a treatment for the pathology (Goonasekera et al., 2011). Contrastingly, Sh2b1 (MGI:105058) is an adapter protein involved in Janus kinase (JAK) and receptor tyrosine kinase signaling pathways. Its deficiency causes obesity and insulin

resistance in mice (Chua, 2010). Furthermore, mutations in the human orthologs *ATP2A1* and *SH2B1* are associated with Brody myopathy (OMIM 601003; i.e. Karpati et al., 1986), and developmental delay and obesity (OMIM 613444; Bachmann-Gagescu et al., 2010), respectively.

Pairs two and three (**Fig. R 7B**) shared one component, *Psmc5* (MGI:105047), which encodes for a proteosome subunit. The expression profile of this gene significantly differs from both its flanking genes (*Smarcd2* in pair number two, and *Ftsj3* in pair number three), even if the intergenic distances in both cases are very short. In particular, *Psmc5* is highly expressed in the nervous system, while the others are absent or expressed at very low levels in these tissues. Smarcd2 (MGI:1933621) is a chromatin remodeling protein recently associated with the regulation of myogenesis (Goljanek-Whysall et al., 2012), whereas Ftsj3 (MGI:1860295) is thought to be involved in pre-rRNA processing, like its human ortholog (Simabuco et al., 2012).

The Euclidean distance between the genes of pair number four (**Fig. R 7C**), composed of *Mapk3* and *Gdpd3*, reached its maximum possible value. Mapk3 (MGI:1346859) plays an important role in the Mapk/Erk cascade, which is involved in vital biological functions such as cell growth, adhesion, survival and differentiation (Alberts et al., 2002). Hence, its expression should be tightly regulated. On the other hand, *Gdpd3* (MGI:1915866) encodes for a multi-pass membrane protein whose expression is limited to a few tissues (epidermis, eye, kidney and stomach). It may take part in the metabolism of glycerol as it contains a glycerophophodiester phosphodiesterase domain.

The next pair, *Tatdn1-Ndufb9* (**Fig. R 7D**), also scored the maximum possible Euclidean distance, even if their intergenic distance was very short and their promoters were divergent. On the one hand, *Tatdn1* (MGI:1916944) encodes for a very lowly expressed deoxyribonuclease implicated in apoptosis in yeast (Qiu et al., 2005) and *Caenorhabditis elegans* (Parrish & Xue, 2003) and, more recently, in eye development in zebrafish (Yang et al., 2012). Its role in other vertebrates is unknown. On the other hand, Ndufb9 (MGI:1913468) forms part of complex I of the mitochondrial membrane respiratory chain, and abounds in all tissues. Mutations in this gene have been linked with mitochondrial complex I deficiency in human (OMIM 252010; Haack et al., 2012).

Shisa5 and *Trex1* composed pair number six (**Fig. R 7E**), which also attained the maximum distance. Shisa5 (MGI:1915044) is a membrane protein that induces apoptosis upon cellular stress (Bourdon et al., 2002). It is preferentially expressed in the spleen and thymus. Meanwhile, *Trex1* (MGI:1915044) encodes for an exonuclease involved in DNA

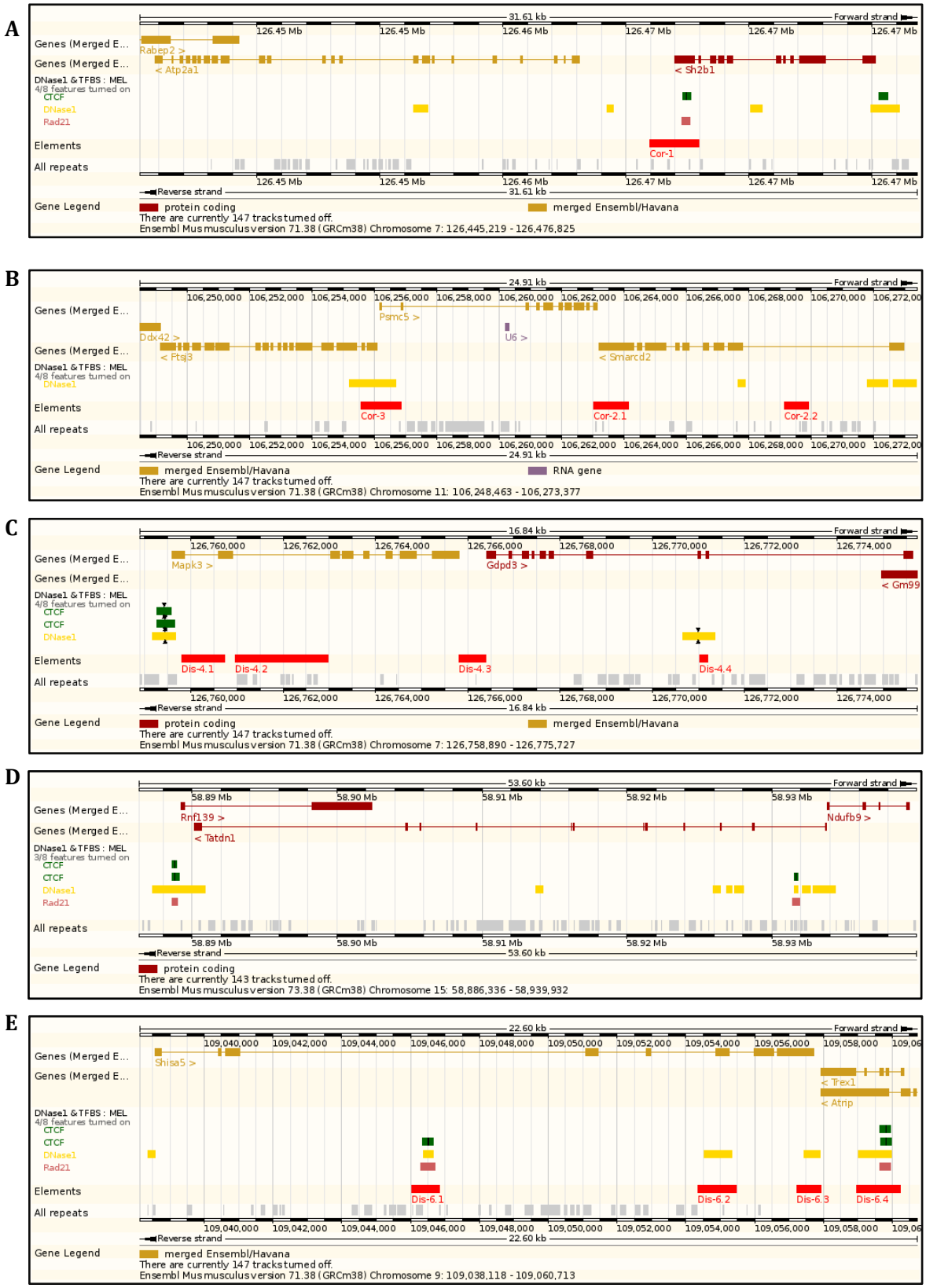


Fig. R 7. Genomic context of the gene pairs selected for functional validation: Pairs one to six. The "Region in detail" page of the Ensembl browser was chosen to depict the genomic context of the six first pairs (A-E). Genes (dark red or orange) are represented as discontinued boxes at the top of each image. The binding sites of CTCF (green) or Rad21 (subdued red) are shown, together with DNase I hypersensitive sites (yellow), at the center of the images. Finally, repeats (grey) and the final selected elements for functional validation of the algorithms (bright red) appear at the bottom of the images. NCBI mouse genome assembly GRCm38.

repair. Its mutations in human are associated with severe diseases, namely Aicardi-Goutieres syndrome 1 (OMIM 225750; i.e. Rice et al., 2007), chilblain lupus (OMIM 610448; i.e. Lee-Kirsch et al., 2007a), susceptibility to systemic lupus erythematosus (OMIM 152700; Lee-Kirsch et al., 2007b), and retinal vasculopathy with cerebral leukodystrophy (OMIM 192315; Richards et al., 2007). On the contrary, *Trex1* null mice develop inflammatory myocarditis and have altered innate immune responses (Morita et al., 2004; Hasan et al., 2013).

The seventh pair (**Fig. R 8A**), formed by *Memo1* and *Dpy30*, was a top hit for both algorithms. Memo1 (MGI:1924140) controls cell migration and is a marker for aggressive metastatic cancers, at least, in humans (Marone et al., 2004). The second member of the pair, Dpy30 (MGI:1924140), is in charge of histone methylation and plays an essential role in differentiation (Jiang et al., 2011).

Tsen34 and *Rps9* was another pair that was obtained from both algorithms (**Fig. R 8B**), and it can be found among the top hits of the Pearson's correlation method. On the one hand, Tsen34 (MGI:1913328) is involved in tRNA processing. Mutations in its human ortholog result in pontocerebellar hypoplasia type 2C (OMIM 612390; Namavar et al., 2011). On the other hand, Rps9 (MGI:1913328) is a ubiquitously expressed ribosomal protein, whose human counterpart is required for normal cell proliferation (Lindstrom & Zhang, 2008).

The first member of the ninth pair (**Fig. R 8C**), *Ddost* (MGI:1194508), encodes for a membrane protein that functions as a glycosyltransferase. Its malfunction in human leads to congenital disorder of glycosylation, type Ir (OMIM 614507; Jones et al., 2012). The second member of the pair is *Pink1* (MGI:1916193). Its activity is essential for the maintenance of mytochondrial function, acting upstream of parkin. Interestingly, mutations in human *PINK1* are the second most frequent cause of familial Parkinson's Disease, right after parkin (OMIM 605909; Kawajiri et al., 2011). Furthermore, *Pink1* knockout mice serve as animal models for the study of the early symptoms of the disease (Glasl et al., 2012).

Finally, the tenth pair was composed of the *Rbm25* and *Psen1* genes (**Fig. R 8D**). Rbm25 (MGI:1914289) is a RNA-binding protein potentially involved in the regulation of apoptosis as its human counterpart (Zhou et al., 2008), and thus, its expression should be tightly controlled. In contrast, *Psen1* (MGI:1914289) encodes for a subunit of the γ-secretase complex (De Strooper, 2003). Mutations in its human ortholog are associated with a number of diseases, namely familial Alzheimer's Disease type 3 (OMIM 607822;

Bekris et al., 2010), familial acne inversa (OMIM 613737; Wang et al., 2010), dilated cardiomyopathy (OMIM 613694; Li et al., 2006), frontotemporal dementia (OMIM 600274; Raux et al., 2000) and Pick's Disease (OMIM 172700; Dermaut et al., 2004).

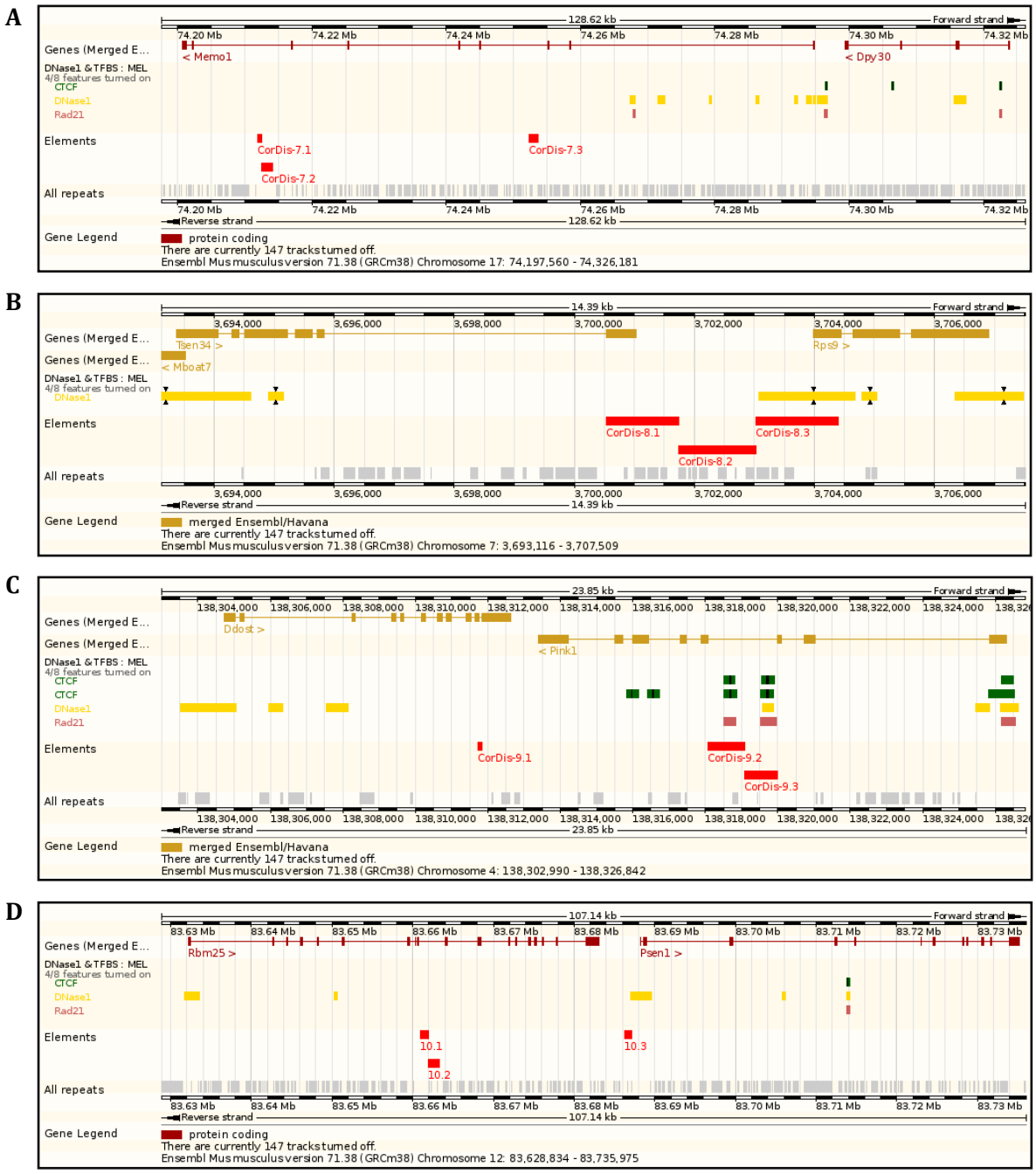


Fig. R 8. Genomic context of the gene pairs selected for functional validation: Pairs seven to ten. "Region in detail" page of the Ensembl browser representing the genomic context of pairs seven to ten (A-D). Genes (dark red or orange) appear at the top of the images as discontinued boxes, whereas the binding sites of CTCF (green) and Rad21 (subdued red), and DNase I hypersensitive sites (yellow) are shown at the center. Repetitive elements (grey) and the elements to be functionally validated for insulator activity (bright red) are depicted at the bottom of the images. NCBI mouse genome assembly GRCm38.

4.1.4.1.2. Criteria for the Selection of Specific Sequences within the Gene Pairs for Testing Insulator Function

Sequences that have been conserved throughout evolution are good candidates to be functionally relevant (Maston et al., 2006; Woolfe et al., 2007; Montoliu et al., 2009). For this reason, by using both VISTA (http://pipeline.lbl.gov/cgi-bin/gateway2; Frazer et al., 2004) and ECR (http://ecrbrowser.dcode.org/; Ovcharenko et al., 2004) browsers, the locus of each selected pair was scanned for the presence of evolutionary conservation, mainly with humans, but also with more distant species like chicken (*Gallus gallus*) or frog (*Xenopus laevis*, only available in the ECR browser) (**Fig. R 9**).

Given that many loci present extensive non-coding evolutionary conservation, the regulation tracks at UCSC and Ensembl genomic browsers were used to aid in the selection of the DNA regions to test for boundary activity. Therefore, the majority of the selected conserved elements contained DNaseI HS alone or in combination with binding sites for different factors like CTCF (insulator-related), Esrrb (estrogen nuclear receptor) or NELFe (involved in transcriptional pausing). In addition, the presence of repetitive elements inside the elements was analyzed because some types of retrotransposons have been associated with insulation (Lunyak et al., 2007; Roman et al., 2011a). Interestingly, boundary activity of sequences without functional annotations or retrotransposable elements may lead to the discovery of new mechanisms of insulation. Some of these elements, without *a priori* clues to their function, were also selected.

However, because not all regulatory sequences are necessarily conserved between species (Bourque et al., 2008; Kunarso et al., 2010), additional elements were included for other reasons. CorDis-9.2 was one of those elements. The human *DDOST-PINK1* locus differs somewhat from its murine homolog, especially when considering the position of the regulatory elements and the presence or absence of specific repetitive elements (**Fig. R 10**). Of note, there are many CTCF binding sites scattered throughout both loci, although they seem to have been reorganized throughout evolution. In spite of this rearrangement, they persist, reinforcing the hypothesis that they must be functionally relevant at this particular genomic location. Three of these shuffled CTCF regulatory regions were chosen for functional validation.

Elements Dis-6.1 (*Shisa5-Trex1*) and 10.3 (*Rbm5-Psen1*) contained CTCF binding sites absent in the human homologs. Importantly, these sites colocalized with SINE B2 retrotransposons which, besides having insulator activity *per se* at some loci (Lunyak et

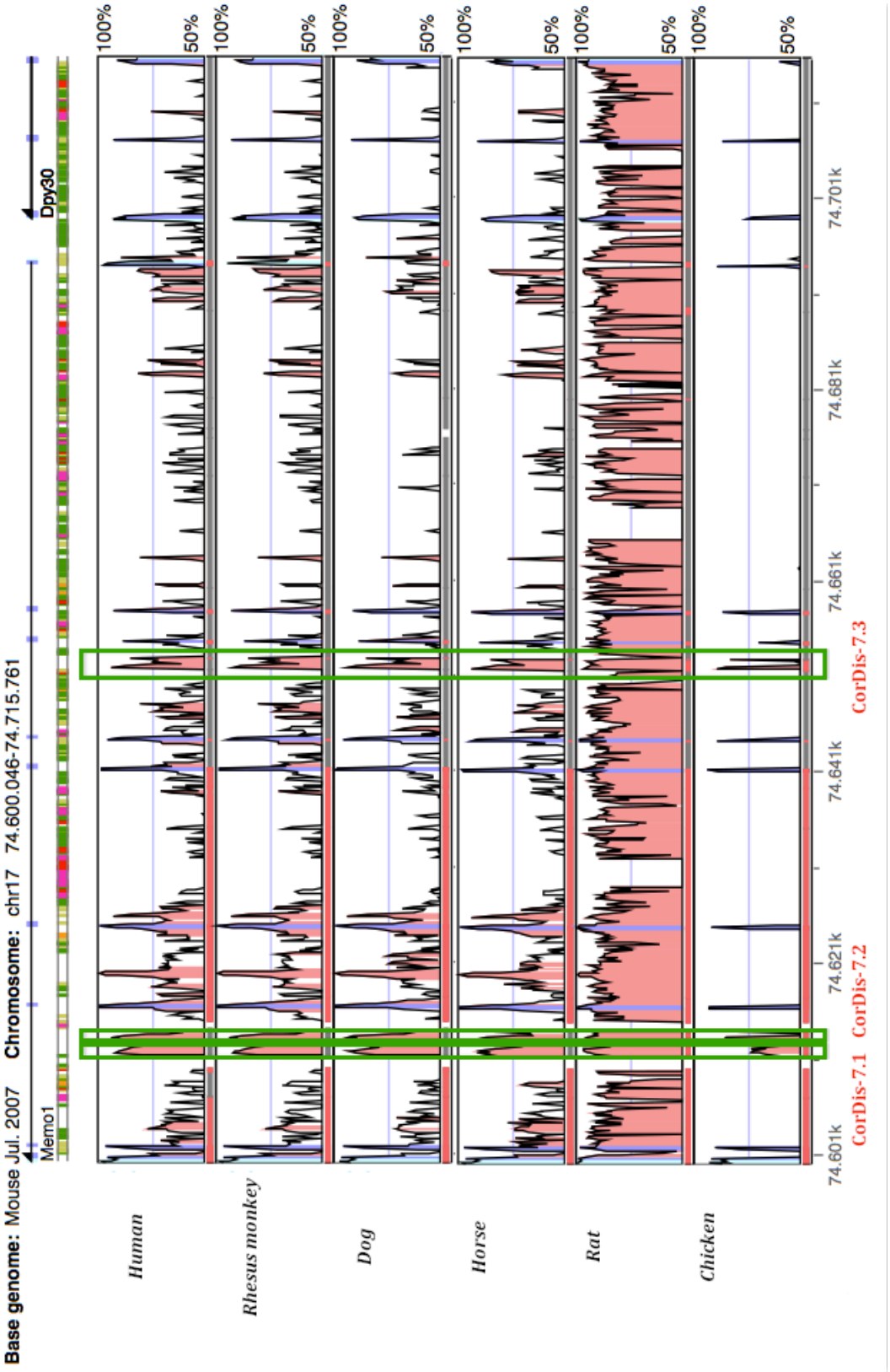


Fig. R 9. Analysis of evolutionarily conserved sequences using the VISTA browser. The *Memo1-Dpy30* locus (CorDis-7) is shown as a case example. The alignment of the mouse locus with that of six other species (human, monkey, dog, horse, rat and chicken, respectively) revealed the presence of several evolutionarily conserved sequences. Green squares frame the three non-coding regions chosen for enhancer-blocking activity testing: CorDis-7.1, CorDis-7.2 and CorDis-7.3. Color code: purple, blue and red mark exons, UTRs and non-coding regions, respectively. Percentage of identity is indicated on the right. Base lines start at 50% identity.

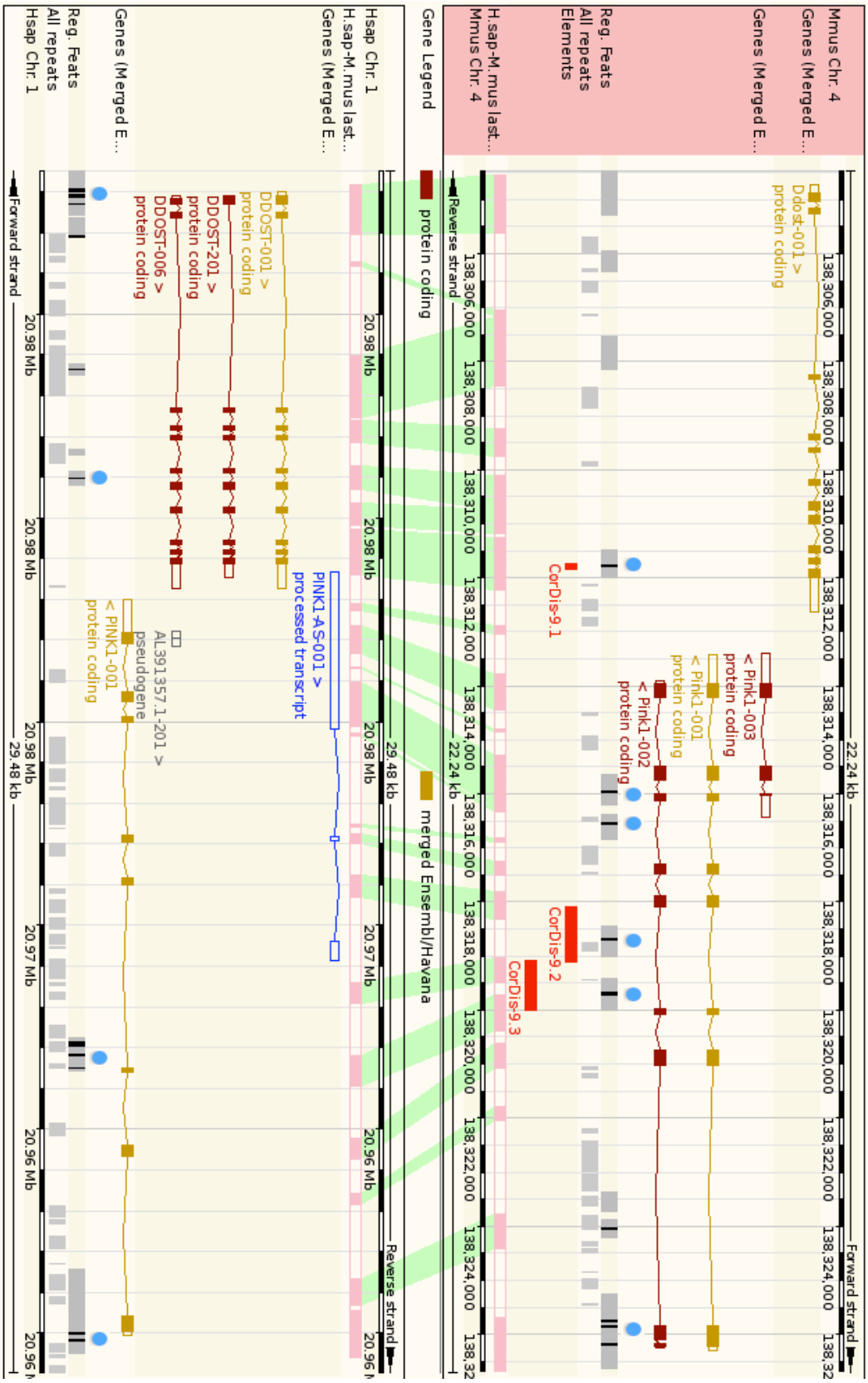


Fig. R 10. Genomic context of the mouse and human *Ddost-Pink1* loci. The mouse *Ddost-Pink1* locus (upper panel) was aligned to its human counterpart (lower panel). Pink boxes and green shadows highlight the conserved sequences. Protein-coding genes are colored in red or orange whereas processed transcripts are depicted in blue. Regulatory features and repetitive elements are shown in gray boxes (see legend on the left). Regulatory features that contain CTCF binding sites are marked with a blue circle. Note that the mouse and human loci map to the forward and reverse strands of their respective reference genomes (GRCm38, GRCh37). Selected sequences (CorDis-9.1, CorDis-9.2 and CorDis-9.3) are presented as red boxes.

al., 2007), have been pointed at as key players in the expansion of, specifically, CTCF sites in the mouse genome (Bourque et al., 2008).

Finally, even if not evolutionarily conserved, CorDis-8.2 was included because it completed the analysis of the intergenic region between *Tsen34* and *Rps9*.

All these elements (**Table R 6**) were tested for *in vitro* enhancer-blocking activity in human embryonic kidney 293 cells (HEK 293 cells) (Lunyak et al., 2007). Only those sequences that performed well in this system were further analyzed for *in vivo* enhancer-blocking activity in zebrafish (*Danio rerio*) (Bessa et al., 2009). Finally, one of the most potent elements in these assays was further evaluated for *in vivo* barrier activity in transgenic mice (*Mus musculus*) (Furlan-Magaril et al., 2011).

4.1.4.2. *In vitro* Enhancer-Blocking Assay in HEK 293 Cells

The *in vitro* enhancer-blocking assay was carried out as previously described using the luciferase-based vector, pELuc (**Fig. R 11B**) (Lunyak et al., 2007). This assay measures the ability of a sequence to block the influence of a distal enhancer on luciferase expression when cloned between the two (IN configuration) (**Fig. R 11C**), upon transient transfection of human HEK 293 cells (Shaw et al., 2002). Putative insulators were also cloned upstream from the enhancer (OUT configuration) (**Fig. R 11D**). This was done as a control to verify that the elements under test were not merely repressive elements that could silence gene expression wherever placed, but real insulators that only function when located in between an enhancer and a promoter. Hence, at this position, luciferase expression should be restored to normal enhanced levels.

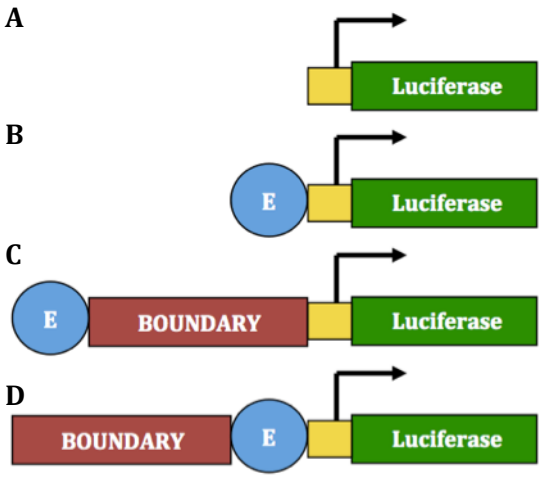


Fig. R 11. Schematic representation of the constructs used in the *in vitro* enhancer-blocking assay. All vectors are based on the pLuc plasmid (A), which provides luciferase basal expression. When the CMV enhancer is included (B), luciferase expression is greatly augmented. The elements to test are cloned in between the enhancer and the promoter (IN configuration) (C) or upstream of the enhancer (out configuration) (D). Insulators will promote a decrease in luciferase expression only in type C constructs, whereas silencers will do so in both C and D vectors.

The Gateway® Cloning Technology was employed to accelerate the cloning procedure of all the elements. This system exploits the recombination properties that bacteriophage lambda uses to integrate into the *E. coli* chromosome (Landy, 1989). Recombination has to occur between specific DNA motives: *att*B in *E. coli* and *att*P in lambda. In order to be able to use the Gateway® Cloning Technology, the destination vector pELuc had to be modified to include an improved version of these *att* sites, *att*R (**Fig. MM 4**). Thus, the Gateway® recombination cassette was cloned between the enhancer and the promoter, generating the pELuc-IN vector. It was also cloned upstream of the enhancer, which originated the pELuc-OUT vector. Then, the elements to test were PCR-amplified and T/A cloned into an intermediate entry vector that already contained *att* sites, *att*L in this case (**Fig. MM 5**). Finally, the recombination between the *att* sites of the entry and destination vectors (both pELuc-IN and pELuc-OUT) was performed, producing the transfer of the element of study from the entry to each of the destination vectors (**Fig. MM 6**).

Figure R 12A illustrates the behavior of the control constructs in this assay. pLuc, which lacked any enhancer element, provided luciferase basal expression. The introduction of an enhancer in pELuc boosted luciferase expression ten-fold. The widely known 1.2-kb insulator from the chicken β-globin locus (cHS4), as well as its core element II/III, were used as positive controls of insulator activity. Also, as a negative control, a mutated version of the II/III core element that does not bind CTCF, was also included (II/III Mut) (Bell et al., 1999; Chung et al., 1993; Recillas-Targa et al., 1999). The cloning of known insulator sequences downstream from the enhancer -but not of the negative control- blocked the influence of the enhancer on the promoter, provoking a dramatic decay in luciferase expression. However, the inclusion of any of the control sequences upstream from the enhancer only moderately affected luciferase expression, indicating that the decrease in gene expression observed in the previous constructs was due to insulator -rather than silencing- activities.

To better visualize the insulator capacity of the elements, the luciferase expression of each of them was normalized to that of the reference pELuc, generating fold enhancer-blocking measurements. As can be seen in **figure R 12B**, the cHS4 element was a potent insulator since it exhibited an 18-fold enhancer-blocking activity; that is, it caused an 18-fold decrease in luciferase expression relative to the control pELuc. On the other hand, a single copy of the II/III core element performed rather poorly in this system. Even so, it was used to set *two* as a threshold for considering that an element conveys insulator activity (Recillas-Targa et al., 1999). Finally, as expected, the mutated version of the II/III

core element failed to block the enhancer and it was used as a negative control. Alternatively, other laboratories use unrelated "neutral" DNA sequences of the same size as the elements under testing as their negative control (Chung et al., 1993). However, these sequences first need to be analyzed in order to confirm the absence of insulator activity within them.

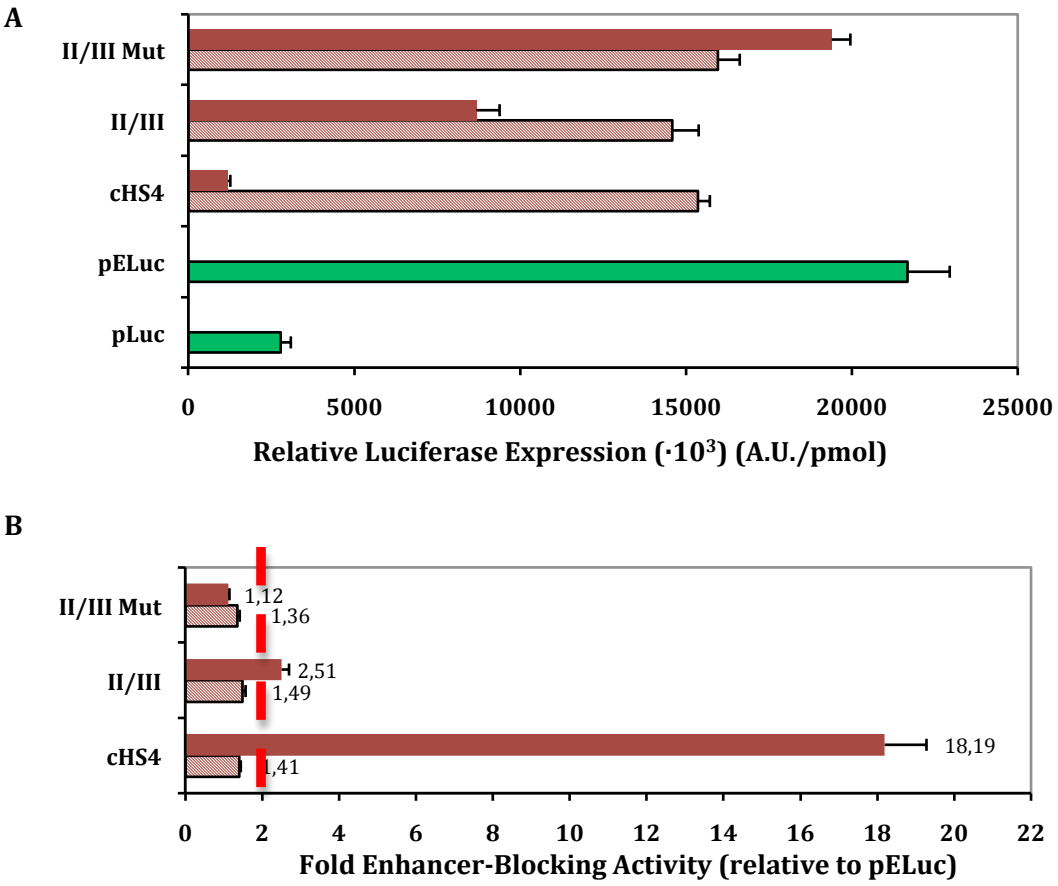


Fig. R 12. ***In vitro*** **enhancer-blocking assay in HEK 293 cells – Control elements. A.** Luciferase expression upon transient transfection with several constructs is shown. pLuc (solid green bar) provided luciferase basal expression. pELuc (solid green bar) contained an enhancer and was used to generate the rest of the constructs. Positive controls included cHS4 and II/III, whereas II/III Mut constituted the negative control. All the control elements were cloned upstream of the enhancer (stripped red bars) or in between the enhancer and the promoter (solid red bars). Data are shown as mean relative luciferase expression (A.U./pmol) + SD of the triplicates for a single independent assay. **B.** The fold enhancer-blocking activity of the control elements was calculated by dividing the luciferase expression of pELuc by that of each element. A sequence was considered to behave as an insulator if its fold enhancer-blocking activity was larger than two (discontinued red line) (Recillas-Targa et al., 1999). Data are shown as mean fold enhancer-blocking activity relative to the control pELuc + SD of the triplicates for a single independent assay.

The selected sequences were grouped into different categories depending on the algorithm they resulted from, and tested for enhancer-blocking activity with this assay.

To be able to compare several sets of experiments, the fold enhancer-blocking activities obtained for each element were normalized to that of the cHS4 insulator. As can

be observed in **figure R 13**, all elements had some insulator activity, and none of them seemed to correspond to repressors. In general, the elements derived from the distance method (either also present in the correlation method or not) exhibited the most potent activities. For example, elements Dis-6.4 and CorDis-9.2 behaved comparably to cHS4 in this assay.

Of note, not all CTCF-site-containing elements performed equally well. Even if the activity of most of them –Cor-2.1, Dis-5.3, Dis-6.1, Dis-6.4, CorDis-8.2, CorDis-9.2, CorDis-9.3, 10.3- almost surpassed two-thirds that of cHS4 (61 to 108%), others barely beat the one-third barrier. This is the case of Cor-1 (35% of cHS4): the large size of this element, 2 kb, may accommodate additional non-annotated regulatory elements that might be counteracting the insulator effect of CTCF. On the contrary, the CorDis-9.1 element, only 100 bp, may be too short to enable the establishment of an enhancer-blocking mechanism (36% of activity with respect to cHS4).

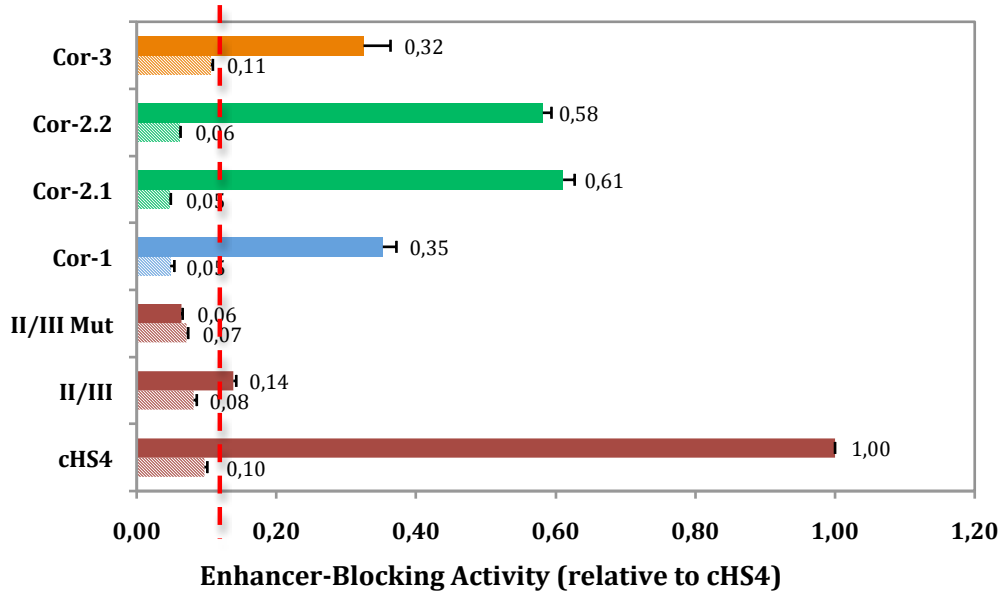
Moreover, many elements without associated functional annotations or repetitive elements like Dis-6.3, CorDis-7.1, CorDis-7.2, CorDis-7.3 and 10.1, exhibited considerable enhancer-blocking activities (44 to 50% with respect to cHS4); unlike 10.2, whose activity was rather weak if compared with that of cHS4 (31%). On the other hand, the substantial activities observed for Cor-2.2, Dis-4.2 and Dis-5.5 (58 to 82% relative to cHS4) may have resulted from the action of retrotransposable elements. In spite of the presence of different types of repeats in Dis-5.2 and CorDis-8.1, their activities only approached one-third that of cHS4 (38 and 28%, respectively).

A number of complex elements were included in the analysis (Cor-3, Dis-5.1, Dis-5.4, Dis-6.2 and CorDis-8.3). They hosted binding sites for many factors such as the transcriptional activator complex nMyc-Max and/or NELFe, which plays a role in RNA II polymerase pausing. In general, these elements behaved moderately to poorly in this *in vitro* assay (18 to 41% with respect to cHS4), possibly due to their complexity, which may have been interfering in the results.

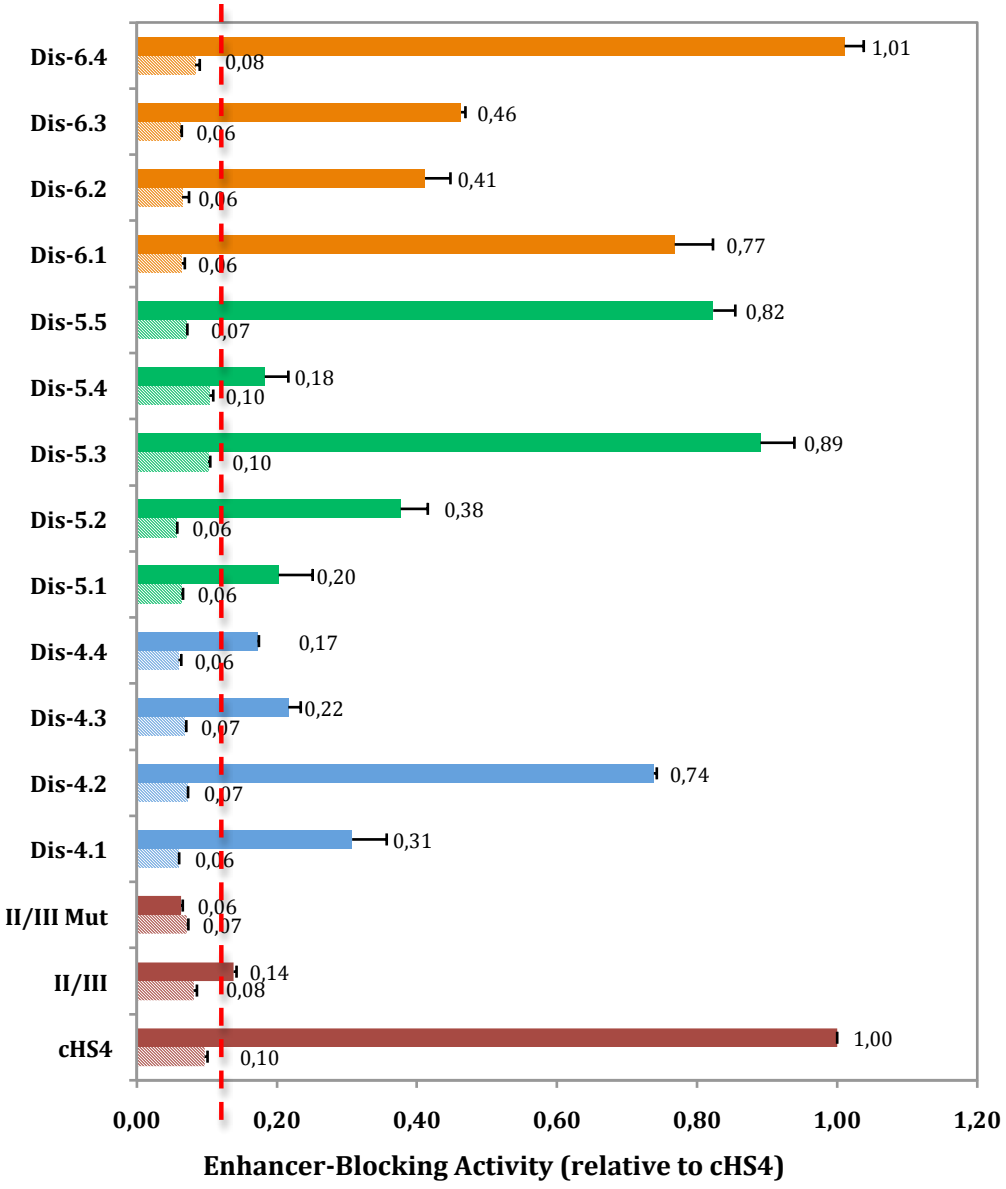
Interestingly, pair number ten also seemed to contain insulators, even if both algorithms overlooked it. As will be seen in the Discussion section, this probably reflects the fact that very restrictive parameters were used to consider a pair of genes as differentially expressed: in an effort to minimize false positives, some true positives were unavoidably missed.

Finally, the rest of the elements, including Dis-4.1, Dis-4.3 and Dis-4.4, conveyed rather weak enhancer-blocking activities (17 to 31% with respect to cHS4).

A



B



(Figure continued)

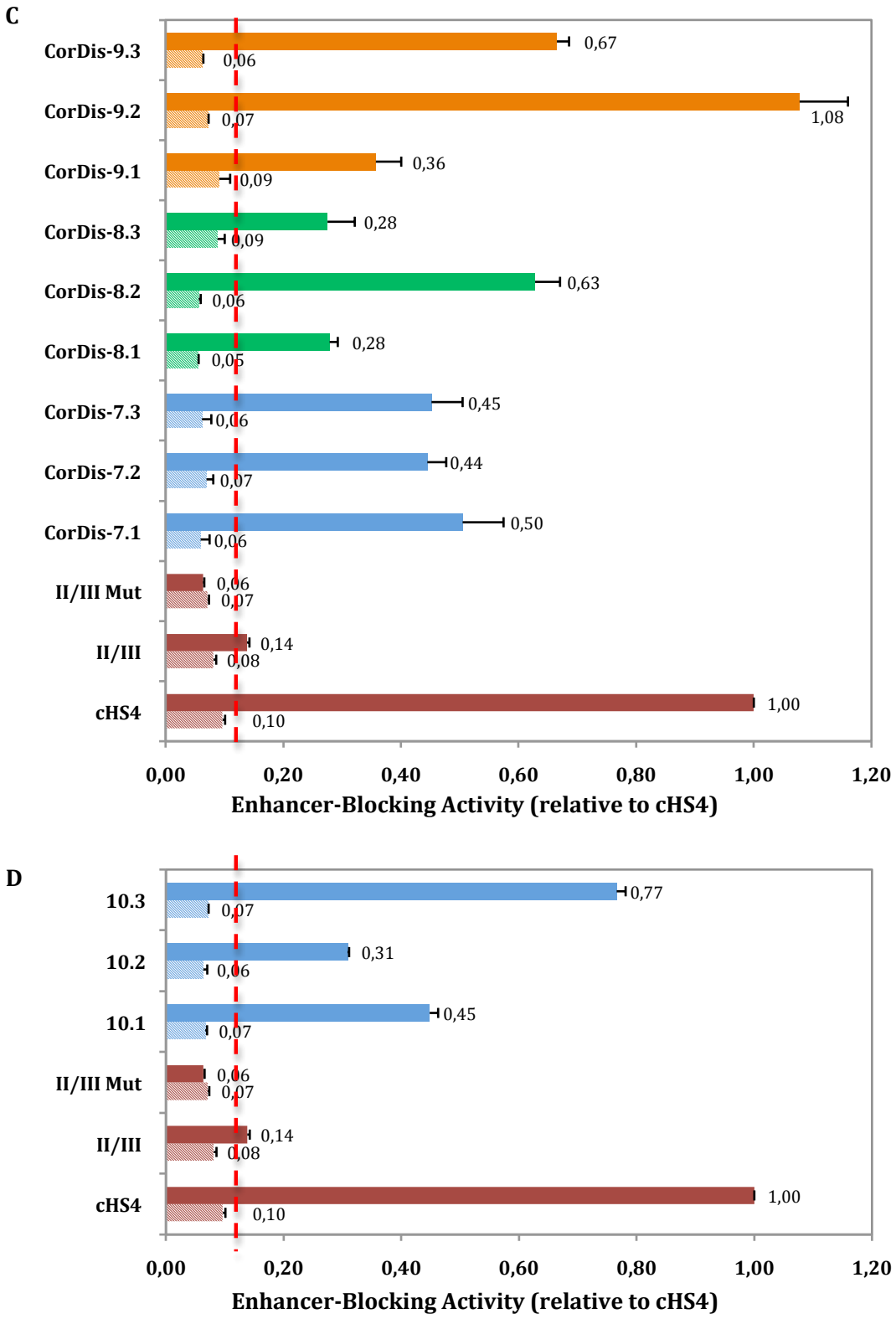


Fig. R 13. *In vitro* enhancer-blocking assay in HEK 293 cells for selected pairs specifically derived from the correlation (A), the Euclidean distance (B) or both (C) methods, as well as an additional pair missed by both methods (D). The fold enhancer-blocking activity for various sequences is shown. The activities have been normalized to that of the cHS4 control to allow for comparison among different sets of experiments. The results for the control elements are shown in red, whereas the segments that belong to the same pair are grouped under the same color. Solid bars represent elements cloned between the enhancer and the promoter, while stripped bars correspond to the same elements but cloned upstream of the enhancer. At this position, only repressors exhibit high scores. By extrapolation, the threshold to consider that an element possesses insulator properties is set at 0,12 (discontinued red line). All experiments were carried out in triplicates in, at least, two independent assays. Data are shown as mean + SEM of the independent assays (n=6).

The introduction of sequences that enlarge the distance between enhancer and promoter could explain the low levels of luciferase expression that we attribute to insulator activity. To rule out this possibility, the size of each element –taken as a measure of the distance between enhancer and promoter– was plotted against its enhancer-blocking activity (**Fig. R 14**). A linear relationship between these parameters would indicate that luciferase expression decays simply because the enhancer influence on the promoter diminishes with distance. However, this was not the case, since no significant correlation was found between element size and enhancer-blocking activity (R^2 = 0.195). For instance, Dis-6.2 failed to reach half the activity of cHS4, even if it is just 100 bp shorter. Nevertheless, the scatterplot in **figure R 14** did suggest that the enhancer-promoter distance may be playing a small role in the effects observed on luciferase expression in the *in vitro* assay. We can conclude that the tested elements actively and significantly block the influence of the enhancer on the promoter and hence contain insulators.

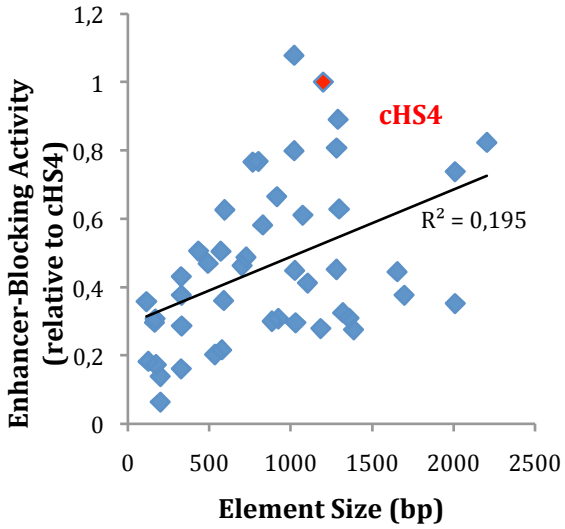


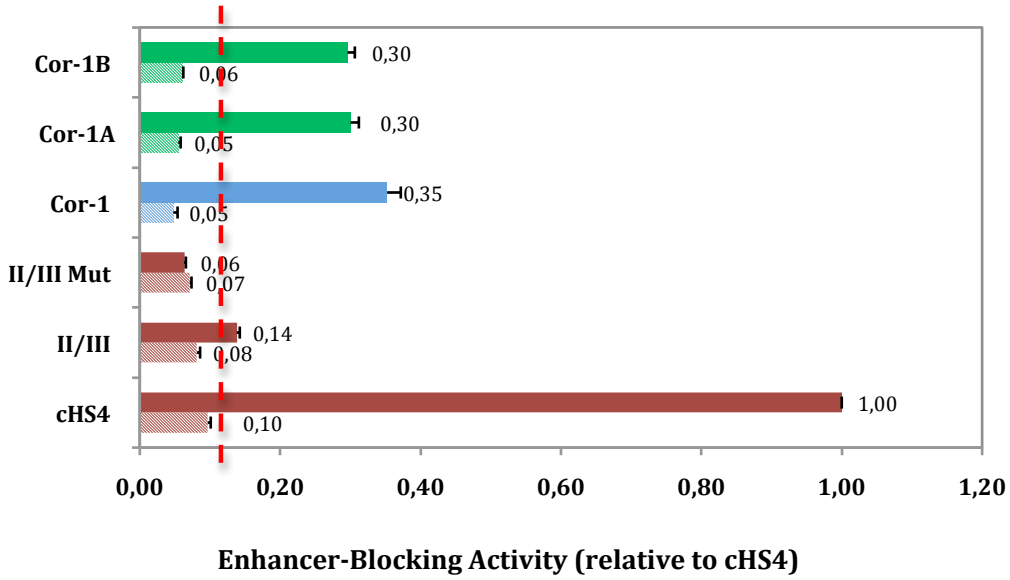
Fig. R 14. The tested elements act as enhancer-blockers *in vitro*. The sizes of the elements were plotted against their enhancer-blocking activity. The data point that corresponds to the cHS4 element is highlighted in red.

4.1.4.2.1. Identification of the Core Insulator Domain for Selected Elements *In Vitro*

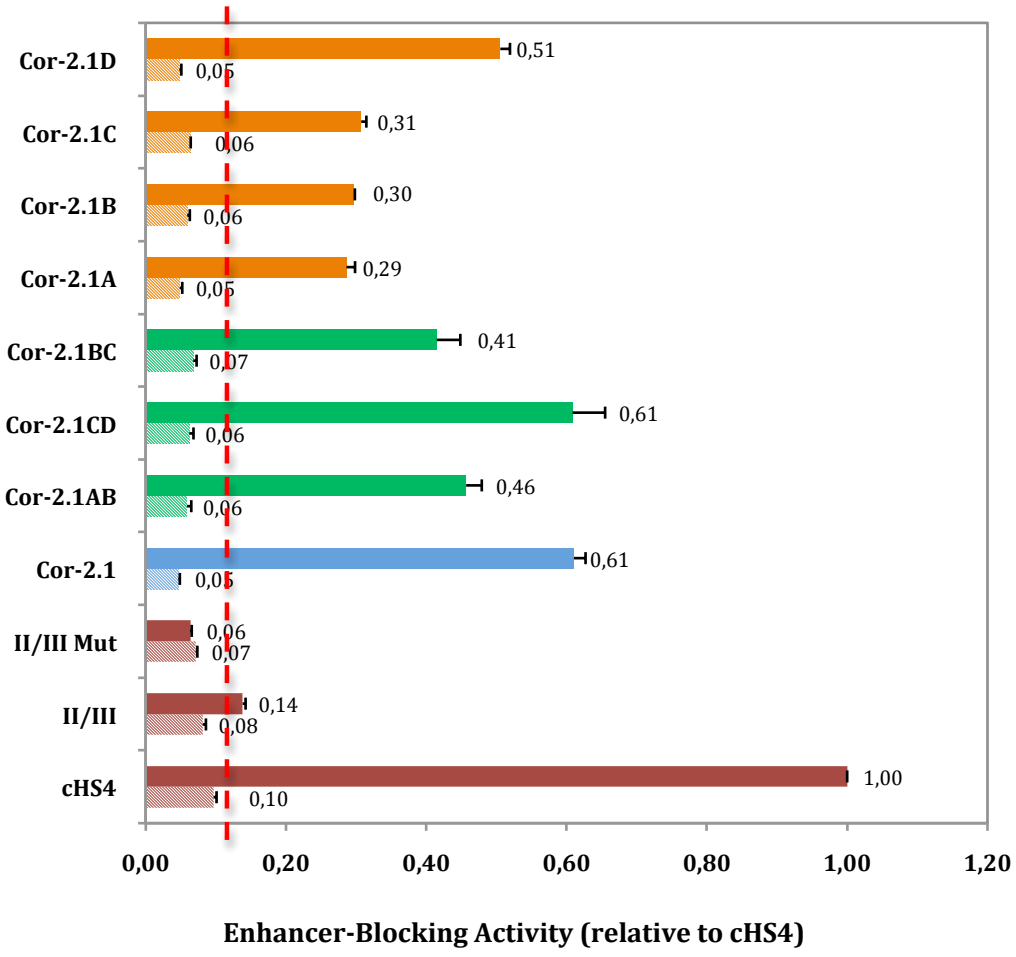
Several elements were further analyzed to try to find their core insulator activity. They were chosen based on an outstanding enhancer-blocking performance *in vitro* and/or the lack of functional genomic annotations that explain why they may be acting as insulators, in an attempt to discover new mechanisms of insulation.

First, Cor-1, which represented the intergenic region between *Atp2a1* and *Sh2b1*, was split in two halves because the original sequence was too long (2 kb). This was thought to be responsible for the moderate enhancer-blocking activity, despite the presence of binding sites for CTCF and the cohesin complex subunit Rad21. These sites were then retained in Cor-1B. However, both fractions A and B exhibited the same activity as the full fragment, so apparently, the Cor-1A half had not been masking the activity of the other half and also contained elements with boundary activity (**Fig. R 15A**).

A



B



(Figure continued)

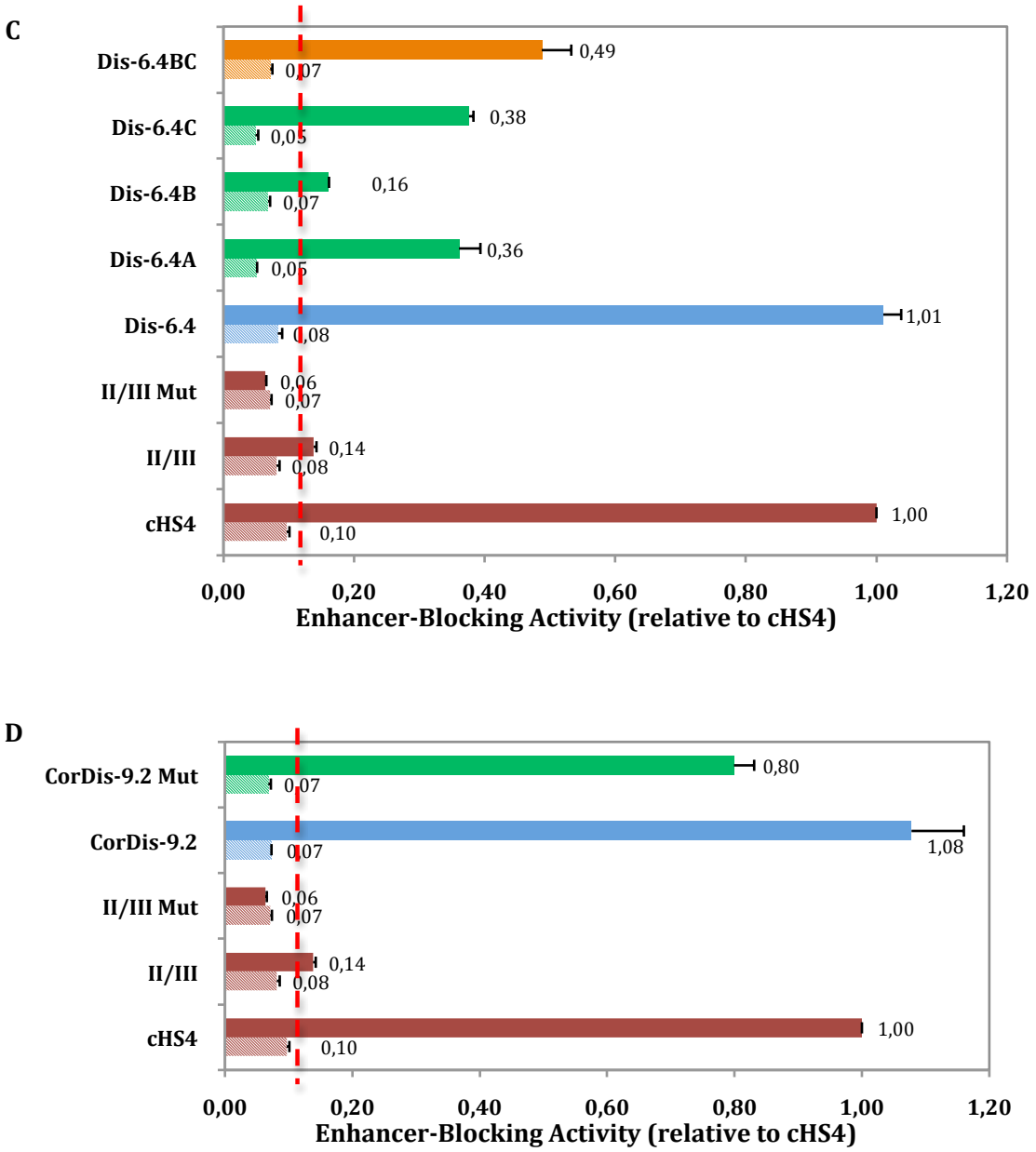


Fig. R 15. Defining the insulator core of selected sequences. Elements Cor-1 (A), Cor-2.1 (B) and Dis-6.4 (C) were chopped into smaller pieces and their *in vitro* enhancer-blocking activity was tested; whereas a mutation in the CTCF binding site was introduced in CorDis-9.2 (D). The results for the control elements are shown in red, for the original elements in blue and for the smaller fragments in green and orange. Solid bars represent elements cloned between the enhancer and the promoter, while stripped bars correspond to the same elements, but cloned upstream of the enhancer. The threshold to consider an element as an insulator was set at 0,12 (discontinued red line). All experiments were carried out in triplicates in, at least, two independent assays. Data are shown as mean + SEM of the independent assays (n=6).

Second, Cor-2.1 was sequentially trimmed until it was found that the core insulator activity resided in its last 400 bp (Cor-2.1D; **Fig. R 15B**). This region, which mapped to the 3'UTR of *Smarcd2*, had not been associated with any DNaseI HS, transcription factor binding site (only very weak CTCF binding in just one tissue: mouse thymus; Mouse ENCODE Consortium et al., 2012) or repetitive element. Yet, it performed half as well as cHS4 even if its size was three times smaller.

Third, Dis-6.4 was initially split in two halves, Dis-6.4A and Dis-6.4BC. The first half contained binding sites for NELFe and the transcriptional activator Zfx, whereas the second hosted the binding of CTCF, Rad21 and Esrrb. Unlike what happened with Cor-1, the effects of both fragments seemed to be additive (**Fig. R 15C**). The BC segment, which contained a portion of *Trex1* 5'UTR, as well as the final exon and 3'UTR of *Atrip* (a gene further upstream *Trex1*), was then divided into the B and C fragments. CTCF-Rad21 binding sites relocated to the B segment, whereas the *Atrip* exon, which mapped to the Esrrb binding site, was excluded from further analyses. Noteworthy, most of the activity of the BC segment was retained in the C portion, which does not bind the CTCF-Rad21 complex.

Finally, a mutation was introduced in the CTCF binding site of CorDis-9.2 (**Fig. R 16A**) to determine if, in this case, this protein was indeed responsible for the observed insulator activity. However, the mutated element still retained 80% of the activity of the wild type (**Fig. R 15D**). There were two possible explanations for this: the mutation may have still been permissive to some CTCF binding, or additional elements in the locus functioned as enhancer-blockers.

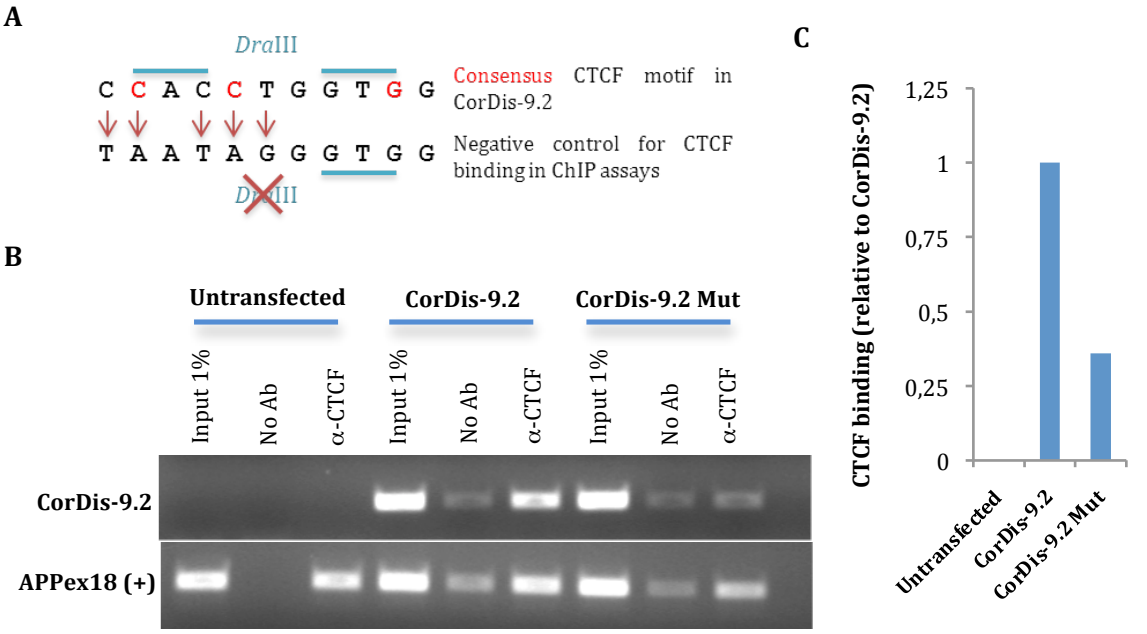


Fig. R 16. The mutation in the CTCF binding site of CorDis-9.2 still allows some CTCF binding. A. The sequence of the CTCF binding site in CorDis-9.2 was converted into a widely used negative control for CTCF binding in ChIP experiments by the introduction of five mutations that affect two of the three most conserved bases of the motif. **B.** A transient-ChIP assay was performed by transfecting HEK 293 cells with either WT or mutant CorDis-9.2, followed by an immunoprecipitation of the DNA-CTCF complexes with a specific antibody. PCRs of the CorDis-9.2 element and of a positive control for CTCF binding (exon 18 from human *APP*) were performed. Untransfected cells were negative for CorDis-9.2 PCR since the primers used were specific for the murine sequence. **C.** The PCR bands were quantified to obtain the relative CTCF binding in the mutant sequence with respect to the wild type.

To rule out the first possibility, a transient chromatin immunoprecipitation experiment (transient-ChIP) was performed. HEK 293 cells were used in this study to reproduce the cellular environment of the *in vitro* enhancer-blocking assay, where both wild type and mutant forms of CorDis-9.2 seem to find all required factors for their insulator activity. Hence, HEK 293 cells were mock-transfected or transfected with CorDis-9.2 in its wild-type or mutant versions. Then, DNA-CTCF complexes were immunoprecipitated with a specific antibody and PCRs were carried out to evaluate the presence of CTCF binding in mutant CorDis-9.2. Exon 18 of human *APP* binds CTCF and was used as a loading control. **Fig. R 16** shows that the mutation did not completely abolish CTCF binding, but reduced it only to 36%. On the other hand, the element also contained a mammalian specific MLT1M retrotransposable element, so none of the possibilities could be discarded.

4.1.4.3. *In Vivo* Enhancer-Blocking Assay in *Danio rerio*

The fact that many of the elements contained enhancer-blocking activity *in vitro* did not necessarily reflect the situation *in vivo*. Therefore, an additional assay was carried out in zebrafish as previously described (Bessa et al., 2009). Briefly, the selected elements were cloned in a Tol2-based transposon between the cardiac actin promoter (CAR) from *Xenopus laevis* (Mohun et al., 1986), and a central nervous system enhancer (Z48) from zebrafish (De la Calle-Mustienes et al., 2005). These elements drove the expression of a GFP cassette to the muscles and midbrain, respectively (**Fig. R 17A**). At this position, insulators should block the influence of the enhancer on the promoter and GFP should only be expressed in muscle cells (somites and heart). On the contrary, innocuous sequences should allow GFP expression also in the midbrain, as happened upon microinjection of the basal construct p48RCAR (**Fig. R 17B**). One-cell stage zebrafish embryos were microinjected up to one hour post fertilization with both the DNA

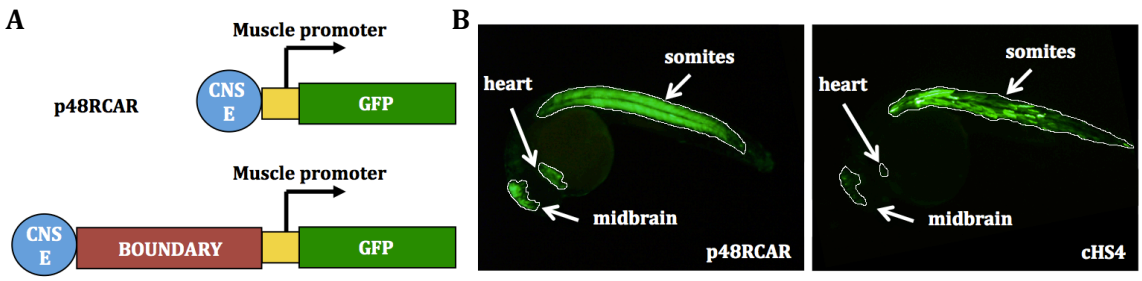


Fig. R 17. *In vivo* enhancer-blocking assay in zebrafish. A. The elements were cloned between a CNS enhancer and a muscle promoter that regulate GFP expression in a Tol2-based transposon. In the basal construct, p48RCAR, the enhancer is adjacent to the promoter. **B.** GFP was expressed in the muscle (somites and heart) and in the midbrain of the zebrafish embryos upon microinjection of the basal construct. However, GFP expression in the midbrain was reduced upon inclusion of the cHS4 insulator.

constructs and Tol2 transposase mRNA. Then, thirty to thirty-six hours later, GFP expression was quantified with the LaserPix software (Bio-Rad) in somites and midbrain and the ratio between them was calculated. If the sequence behaved as an enhancer-blocker, GFP expression in the midbrain would be low or even absent -while that of the muscle would be maintained- giving rise to high muscle to midbrain fluorescence ratios.

Some, but not all, of the elements previously tested *in vitro*, were selected for this *in vivo* assay. At least one element per pair was analyzed. Again, they were grouped according to the algorithm they derived from (**Fig. R 18** and **19**).

Importantly, in transgenesis experiments in zebrafish, even if embryos are microinjected at the one-cell stage, constructs usually integrate into the genome at the multicellular stage, and only in some cells. Thus, these transgenic animals are mosaic. Moreover, integrated constructs are always subject to chromosomal position effects. These are the two reasons why there is a high variability in GFP expression levels and distribution among transgenic individuals produced in the same microinjection session. However, the high number of transgenic animals that are obtained –up to dozens of independent transgenic lines- for each construct, largely compensates for the variability of the data, statistically speaking. In addition, only fish with homogenous reporter gene expression in the somites were considered. In any case, boxplots are used to represent the results since they allow the visualization of the variability of the data.

As anticipated, the *in vitro* enhancer-blocking performance did not perfectly match the situation *in vivo*. For example, elements Cor-2.1 and Cor-2.2 exhibited higher insulator activity than Cor-1 and Cor-3 *in vitro*. However, *in vivo*, the behavior of Cor-2.1 did not significantly differ from that of the negative control, and Cor-2.2 only blocked the enhancer weakly, whereas Cor-1 and Cor-3 possessed potent insulator activities (**Fig. R 18A**).

Mimicking the *in vitro* results, the Euclidean distance method seemed to better predict strong insulators since all selected sequences (except CorDis-8.1) that resulted from this method behaved as powerful enhancer-blockers *in vivo* (**Fig. R 18B** and **C**). In particular, CorDis-7.2 was the most potent insulator of the whole set (note that the median value was the highest among the tested sequences).

In addition, element 10.3 possessed both *in vitro* and *in vivo* enhancer-blocking activity, even if it was ignored by the algorithms (**Fig. R 18D**).

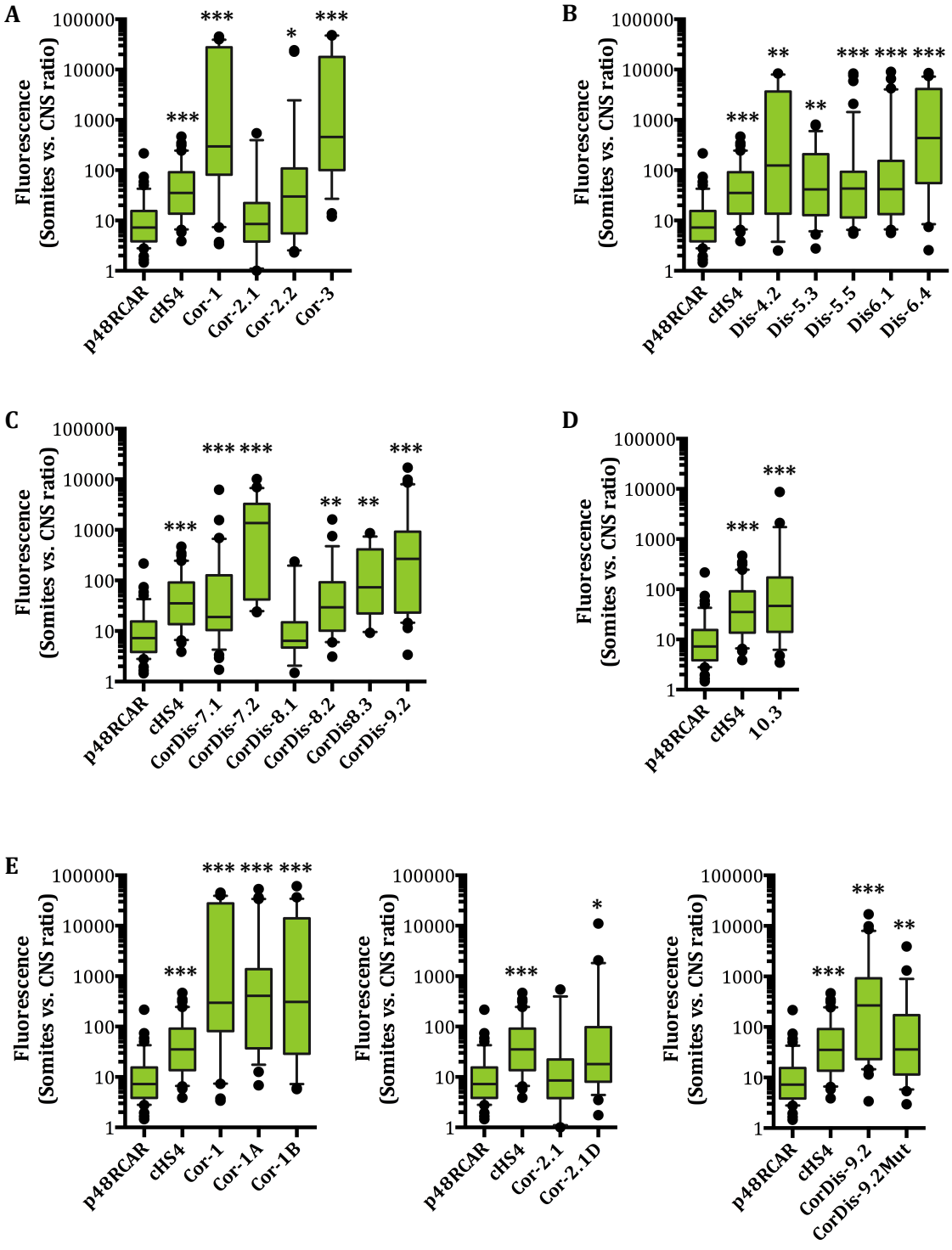


Fig. R 18. *In vivo* enhancer-blocking assay in zebrafish. Quantifications. The effect of the introduction of various sequences between the CNS enhancer and the muscle promoter in p48RCAR is shown. GFP expression was quantified thirty-six hours after microinjection, and the fluorescence ratio between the somites and the CNS was taken as a measure of the enhancer-blocking activity. Figures A, B, C and D show the results for the sequences derived from the correlation, Euclidean distance, or both methods, as well as from the pair that do not belong to any of them, respectively. Figure E illustrates the results of the short segments within Cor-1 and Cor-2.1, as well as the effect of the mutation on the CTCF site in CorDis-9.2. The positive control cHS4 is included in all cases. Data are shown as boxplots that integrate the fold enhancer-blocking activity (fluorescence ratio somites/CNS relative to p48RCAR) for all transgenic individuals for a given construct. The bottom and top of each box represent the first and third quartiles, respectively, whereas the line inside each box represents the median value. The ends of the whiskers indicate the 10th and 90th percentiles. Outliers are depicted as black dots above and below the boxes. Median test: * significant at p-value < 0.05; ** significant at p-value < 0.01; *** significant at p-value < 0.001.

Finally, the *in vivo* assay confirmed that elements Cor-1A and Cor-1B had the same enhancer-blocking activity as the parent Cor-1, that the core insulator from Cor-2.1 failed to provide potent insulation *in vivo*, and that the mutation in the CTCF binding site of CorDis-9.2 is not sufficient to completely abolish its insulator capacity (**Fig. R 18E**).

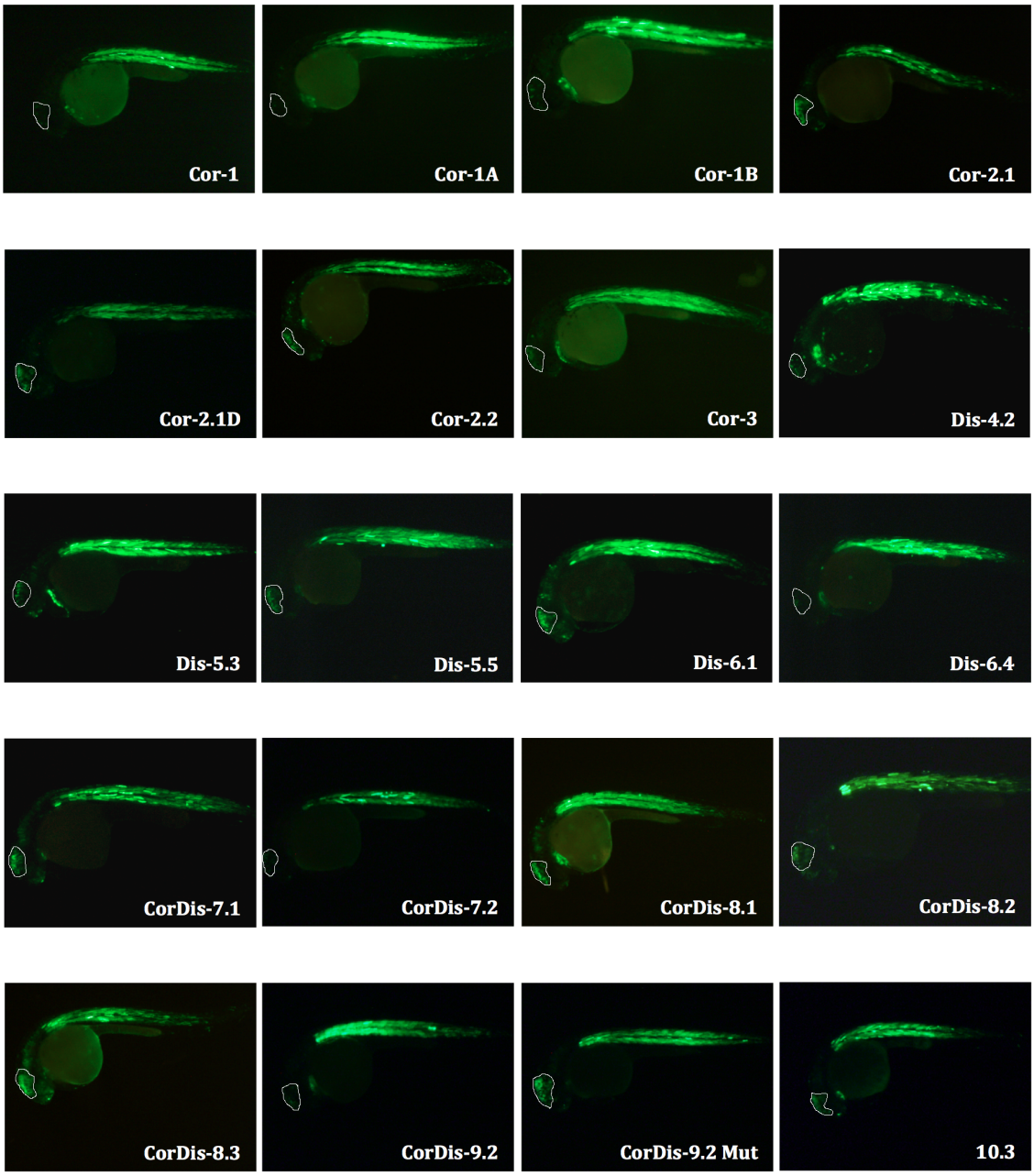


Fig. R 19. Representative transgenic zebrafish individuals obtained as a result of the *in vivo* enhancer-blocking assay. The midbrain of each animal is delimited by white lines.

4.1.4.4. *In Vivo* Testing for Barrier Activity in *Mus musculus*

4.1.4.4.1. Production and Analysis of Transgenic Mouse Lines

Insulator elements are classically defined by their enhancer-blocking and/or barrier activities. Having stated that CorDis-9.2 conveyed potent *in vitro* and *in vivo* enhancer-blocking activity, we sought to determine if it also functioned as a barrier *in vivo* with a previously established transgenic system based on HsdWin:NMRI outbred mice (Furlan-Magaril et al., 2011). In these animals, both copies of the tyrosinase (*Tyr*) gene, which codifies for the key enzyme involved in melanin production, are normally expressed. However, they carry a point mutation that renders the enzyme non-functional, and thus, mice are albino (Jackson & Bennett, 1990). The assay employed here relies on the rescue of this phenotype upon the introduction of a functional tyrosinase minigene by pronuclear microinjection of fertilized oocytes (Beermann et al., 1990). Usually, transgenes integrate into the genome in a tandem multi-copy configuration. The presence of insulators flanking the transgenes should protect them from chromosomal position effects, ensuring a linear relationship between transgene expression and copy number. In this system, tyrosinase expression, and consequently pigmentation levels in the eyes and skin of transgenic animals, should correlate with the number of transgene integration events.

A single copy of CorDis-9.2 was cloned upstream from the tyrosinase mini-gene in ptrTYR5 (Beermann et al., 1991), generating ptrTYR5-CorDis-9.2. Again, the fact that transgenes integrate in tandem multi-copy arrays implies that, eventually, the insulator would shield most copies. Six transgenic founder mice were obtained as a result of nine microinjection sessions, yielding a transgenesis efficiency of 9.5% (six positive transgenic

Table R 7. CorDis-9.2 microinjection data. CorDis-9.2 was cloned in a *Tyrosinase*-containing plasmid (ptrTYR5) and microinjected into albino outbred HsdWin:NMRI fertilized oocytes in nine independent sessions. Six transgenic founder animals were generated. The last row summarizes the results. SO, superovulated female mice. Data from the Transgenic Core Facility of CNB-CBMSO.

SO	SO with Vaginal Plugs	Fertilized oocytes obtained	Microinjected fertilized oocytes	Transferred embryos (% microinjected)	Fosters	Pregnant fosters	Pups	Transgenic pups (% of all pups)
10	8	185	71	29 (40.9)	2	1	2	1 (50.0)
10	5	201	47	25 (53.2)	2	1	5	0 (0.0)
10	4	59	48	38 (79.2)	2	1	0	0 (0.0)
10	5	208	66	33 (50.0)	1	1	0	0 (0.0)
10	5	257	117	72 (61.5)	3	3	18	3 (16.7)
10	10	158	112	75 (67.0)	4	3	16	0 (0.0)
10	3	167	57	30 (52.6)	2	2	13	1 (7.7)
10	3	139	32	30 (93.8)	2	2	9	1 (11.1)
10	2	165	31	22 (71.0)	1	0	0	0 (0.0)
90	45	1539	581	354 (60.9)	19	14	63	6 (9.5)

pups from a total of sixty-three born pups; **Table R 7**). Control transgenic animals for the same construct but without any insulator had been previously produced (Furlan-Magaril et al., 2011). Therefore, they were not generated again for animal welfare reasons.

Founder animals were crossed with wild type NMRI individuals in order to obtain hemizygous transgenic lines. All founders transmitted the transgene through the germline (**Table R 8**). Pigmentation was uniform in the progeny of all of them, although some small patches could be spotted in some mice (i.e. dark patch on the head of mouse #A2700 in **Fig. R 20**). An exception was line #A1683. In this case, mice could clearly be divided into two groups according to their coat colors, beige or dark beige, suggesting two integration events. Interestingly, all the individuals from the #A1743 line, including the founder, were albino, whereas those from line #A1795 only showed pigmentation in the eyes.

Table R 8. All founders transmitted the transgene through the germline.

Founder ID	Sex	Strain	Phenotype	Transmission to F_1
A1683	Male	HsdWin:NMRI	Pigmented (gray)	Yes (27/50)
A1729	Male	HsdWin:NMRI	Pigmented (gray)	Yes (10/13)
A1730	Male	HsdWin:NMRI	Pigmented (mottled gray)	Yes (5/12)
A1743	Female	HsdWin:NMRI	Albino (white fur, red eyes)	Yes (4/14)
A1774	Male	HsdWin:NMRI	Pigmented (beige)	Yes (8/14)
A1795	Female	HsdWin:NMRI	Pigmented (off-white fur, black eyes)	Yes (7/24)

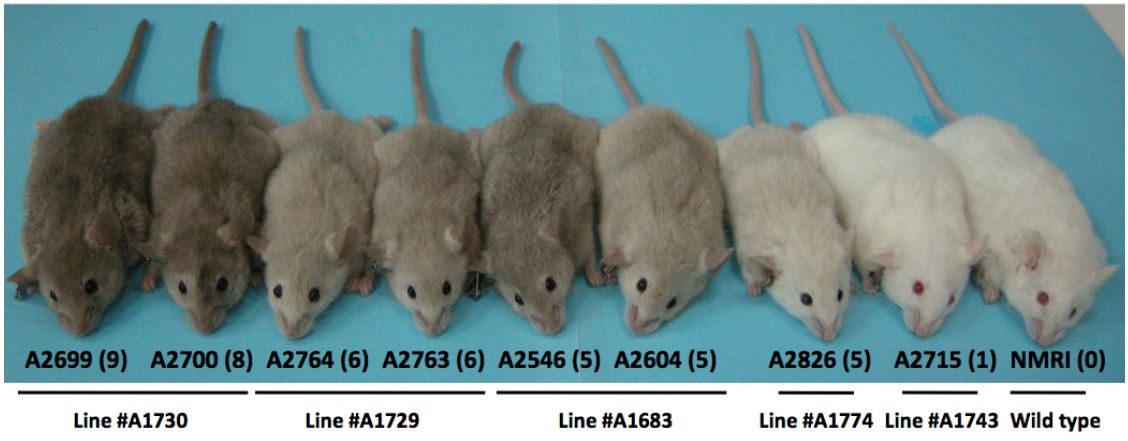


Fig. R 20. Mice became progressively pigmented along with the number of integrated transgene copies. One or two F_1 animals from five of the six transgenic lines generated were anesthetized and photographed together with a wild type NMRI mouse. They were arranged in descending order according to the estimated copy number (in parenthesis).

Three representative animals from five of the six lines (all except for line #A1795) were selected for further analysis. In the case of line #A1683, more individuals were chosen due to the unexpected bimodal phenotype observed. Southern blot analysis of

genomic DNA was performed in order to determine transgene integrity and copy number in these individuals (**Fig. R 21**). The probe used mapped to tyrosinase exon 5 and thus, hybridized with both endogenous and exogenous genes. The transgene lacked several introns, so digestion with *Hind*III generated two fragments of different sizes. Hence, the endogenous gene -present in two copies in the genome- served as an internal control for the quantification of the transgene copy number in each animal (**Fig. R 21** and **Table R 9**).

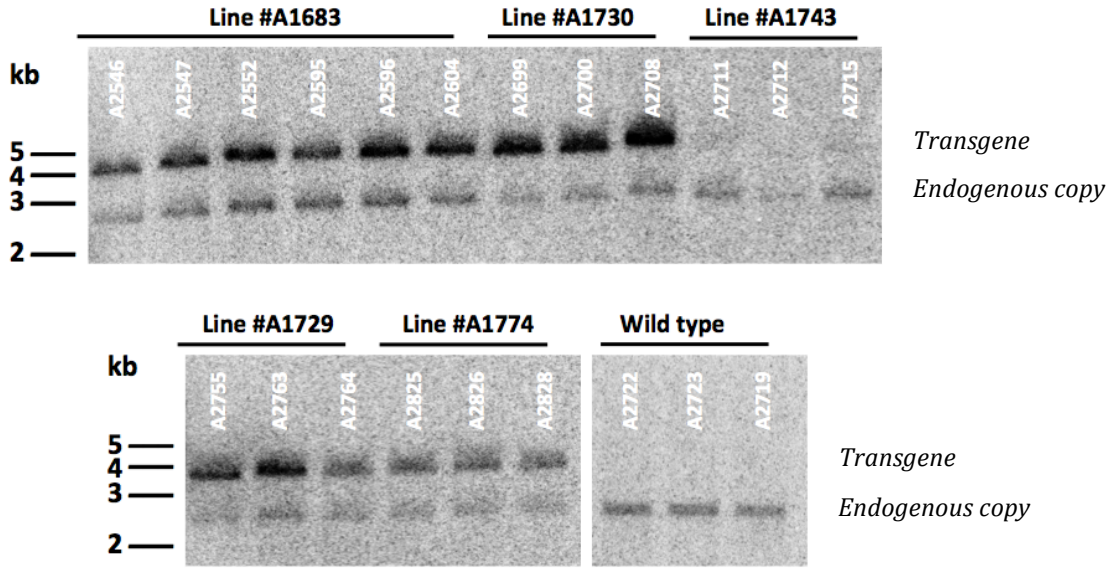


Fig. R 21. Southern blot analysis of transgene integrity and copy number in the different transgenic lines generated. Digestion of genomic DNA with *Hind*III resulted in 3.4 and 2.2 kb bands corresponding to the tyrosinase minigene transgene and the endogenous gene, respectively. Wild type animals were included for reference.

Not surprisingly, line #A1743 contained only one copy of the transgene and hence, only one copy of CorDis-9.2, at the 5' end. Since tyrosinase was not flanked by insulators at both sides, its expression was still subject to chromosomal position effects: the albino phenotype could be a consequence of tyrosinase expression being shut off due to the presence of silencing activity at the integration site.

Whereas the estimated transgene copy number appeared accurate for lines #A1774 and #A1729 (five and six copies, respectively), a high variability in the results could be appreciated for the other two lines, #A1730 (eight copies) and #A1683 (three to five copies). In the latter case, the individuals presented a bimodal phenotype regarding fur pigmentation. It was thus reasonable to think that the transgene may have integrated at several sites in the founder, sites that may have segregated independently in the first generation originating different sub-lines.

Table R 9. Copy number analysis of the F_1 offspring from the transgenic founders. *Hind*III-digested bands were quantified and transgene copy number was calculated using as a reference the intensity of the endogenous *Tyr* gene, which equals to two copies. The phenotype of the animals is included.

Founder	F_1 Individual	Phenotype	Copy Number	Mean Copy Number
A1683	A2546	Pigmented (gray)	4.84	3 to 5
	A2547	Pigmented (gray)	4.17	
	A2552	Pigmented (light gray)	3.72	
	A2595	Pigmented (light gray)	2.89	
	A2596	Pigmented (light gray)	3.99	
	A2604	Pigmented (light gray)	5.14	
A1730	A2699	Pigmented (dark gray)	9.17	8
	A2700	Pigmented (dark gray)	8.03	
	A2708	Pigmented (dark gray)	7.34	
A1743	A2711	Albino (white fur, red eyes)	0.92	1
	A2712	Albino (white fur, red eyes)	1.60	
	A2715	Albino (white fur, red eyes)	0.76	
A1729	A2755	Pigmented (light gray)	5.87	6
	A2763	Pigmented (light gray)	6.13	
	A2764	Pigmented (light gray)	5.86	
A1774	A2825	Pigmented (beige)	4.41	5
	A2826	Pigmented (beige)	5.01	
	A2828	Pigmented (beige)	4.95	

In order to confirm this hypothesis, an additional Southern blot was performed. Instead of *Hind*III, genomic DNA was digested with *Bst*XI, which only cut once inside the transgene. To prevent undesirable interferences with the endogenous *Tyr* locus, a probe that hybridized solely with the transgene (SV40 poly(A) tail) was employed. Given that the transgenes usually integrate in tandem arrays, bands corresponding to head-to-head and/or head-to-tail configurations were expected (**Fig. R 22**). Additional bands of unknown sizes (one per integration event) would also appear.

As anticipated, multiple integration events had occurred in founder A1683, since all three F_1 descendants studied showed different profiles (**Fig. R 23**). Besides, the variability in copy number estimates observed for line #A1730 had surely stemmed from experimental error, because all F_1 descendants did share the same digestion profile. Finally, this analysis confirmed that line #A1743, in which all individuals were phenotypically albino, only contained one copy of the transgene.

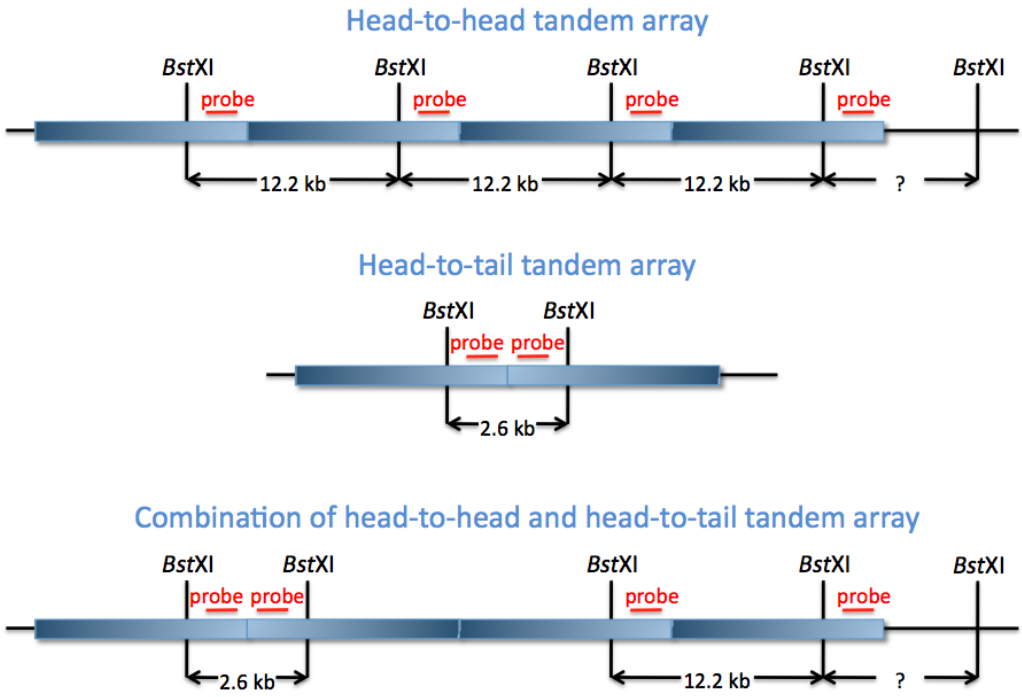


Fig. R 22. Transgene tandem array configurations. Upon digestion with *Bst*XI and subsequent Southern blot analysis (probe in red), head-to-head tandem arrays would be represented by a band of the same size as the transgene. On the contrary, head-to-tail arrays would generate a much shorter band, corresponding to the ends of the transgenes. Combinations of these two types of configurations were also possible.

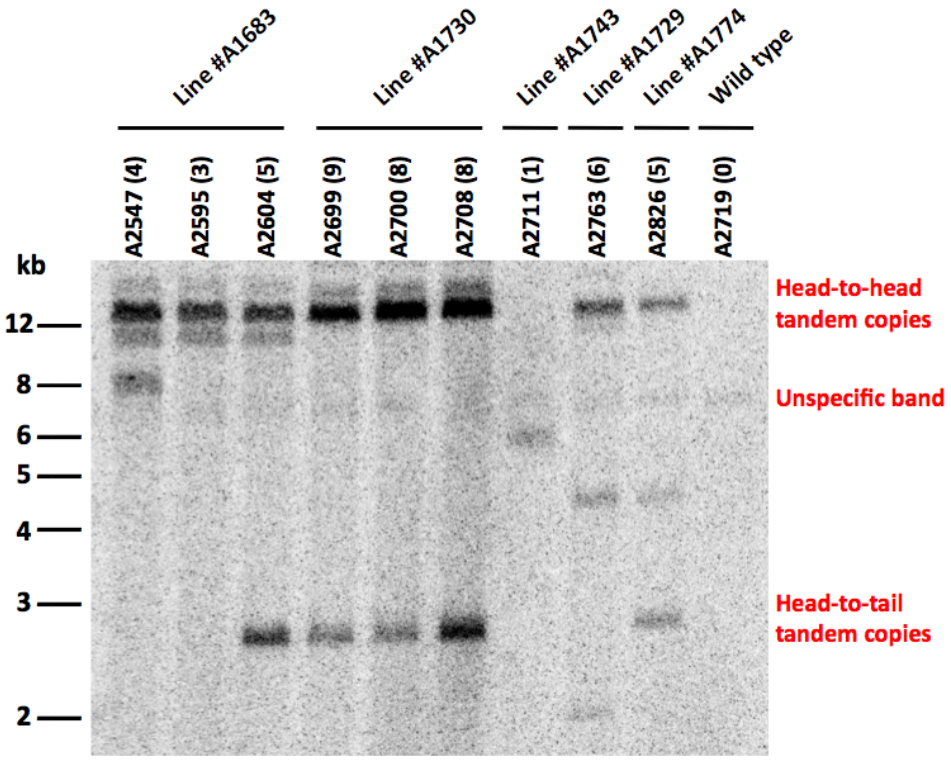


Fig. R 23. Southern blot analysis of transgene integration sites. Genomic DNA from selected F_1 animals of five out of six transgenic lines were digested with *Bst*XI and hybridized with a probe specific for the construct (SV40 poly(A) tail). As a negative control, a wild type NMRI mouse was included. The 12.2 and 2.6 kb bands corresponding to head-to-head and head-to-tail transgene tandem arrays are highlighted.

4.1.4.4.2. Analysis of the Expression of the Transgene

It was noticeable at a glance that the level of pigmentation increased with transgene copy number. Nevertheless, it was also clear that mice with an identical number of copies did not completely share the same coat color (i.e. mice #A2546, #A2604 and #A2826), and that some animals exhibited patches in the fur (i.e. mouse #A2700 in the head).

To unmistakably draw conclusions on the barrier capability of CorDis-9.2, TaqMan real-time PCRs were performed to quantify tyrosinase gene expression in these transgenic mice, as previously described (Furlan-Magaril et al., 2011). The TaqMan probe employed hybridized to both endogenous locus and transgene. In order to get rid of the endogenous contribution, tissues from non-transgenic NRMI individuals were used as calibrator samples. As already stated, these albino mice do synthesize tyrosinase mRNA at the same level as wild type animals, albeit mutant, which generates non-functional proteins (Jackson & Bennett, 1990).

Tyrosinase expression was assessed in the eyes and skin, since these tissues typically express this gene. Additionally, brain was included in the analysis as a non-expressing control (Gimenez et al., 2003) (**Fig. R 24**).

Line A1683 was composed of individuals with different numbers of transgene copies at different integrations sites. Their tyrosinase expression could not simply be averaged and considered representative of the line. Thus, these individuals were removed from further analysis.

Furlan-Magaril et al. demonstrated that, upon random integration into the genome of the backbone vector that lacked any insulator element (ptrTYR5 or simply TYR5), transgene expression was subject to chromosomal position effects. This was true for all tissues examined: transgene expression was relatively uniform in the eye, inconsistent in the skin and nonexistent in the brain, regardless of the transgene copy number (**Fig. R 24**). Furthermore, they proved that the inclusion of an insulator sequence into the construct prevented chromosomal position effects and ensured a linear relationship between transgene expression and copy number. The insulator sequence they investigated was αEHS1.4, an element present upstream from the chicken α-globin domain. Even if the effect in the eye was not evident, it worked perfectly in the skin and, surprisingly, in the brain (**Fig. R 24**).

Transgene expression in the eyes of the mice transgenic for the ptrTYR5-CorDis-9.2 construct was uniform, as in the case of the negative and positive controls. The lack of

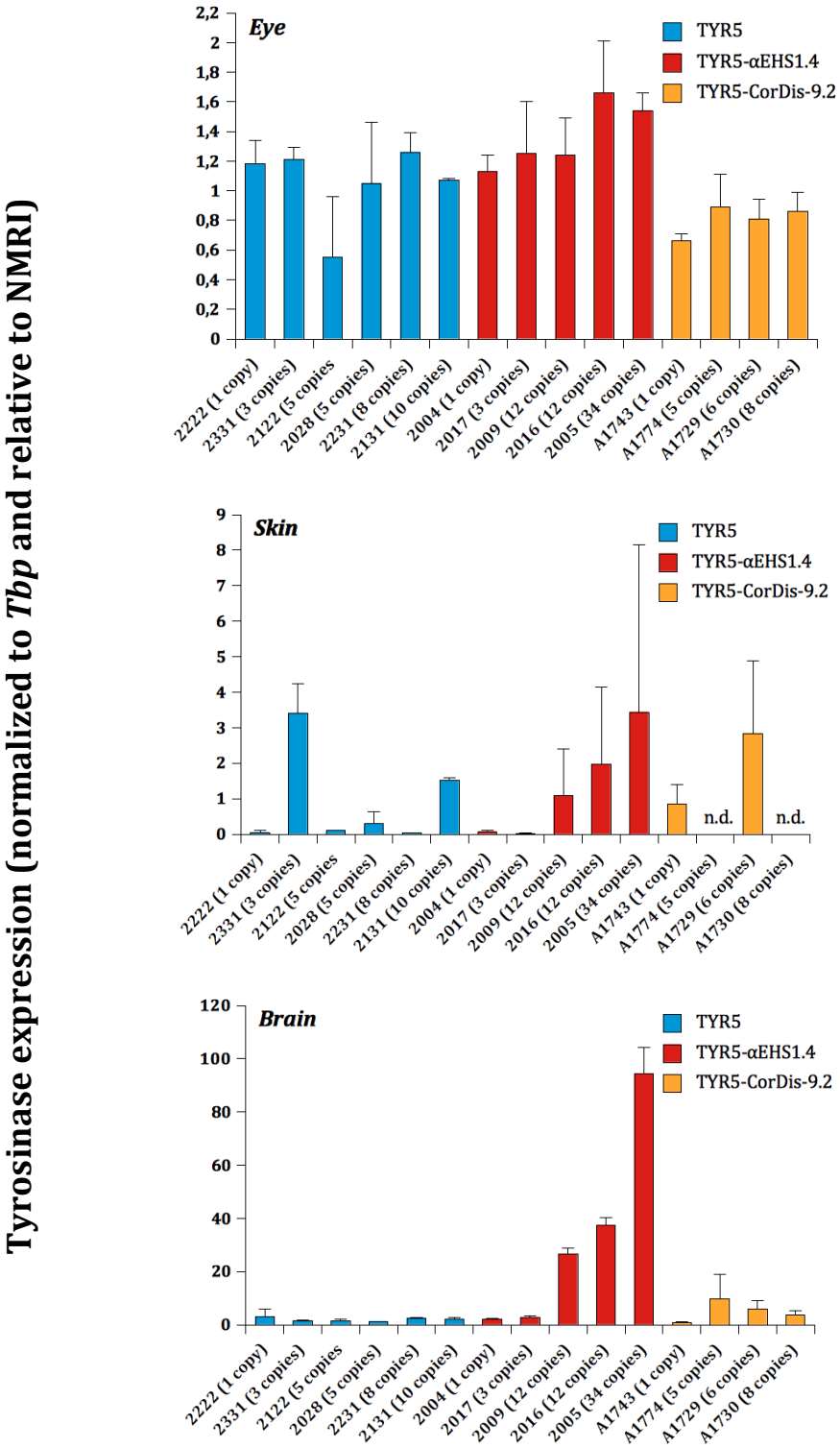


Fig. R 24. *In vivo* testing for barrier activity. Correlation between transgene copy number and tyrosinase expression. The expression of both endogenous and exogenous tyrosinase genes was quantified by TaqMan real time PCR in two tyrosinase-expressing and one non-expressing tissues (eyes, skin and brain, respectively) of several individuals pertaining to different transgenic lines (x-axis). The results were then averaged to obtain mean expression values representative of each line. Data from mice carrying the empty control (TYR5 in the legend), the positive control (TYR5-αEHS1.4) and the element under study (TYR5-CorDis-9.2), are depicted in blue, red and orange, respectively. Wild type NMRI individuals were used as calibrators (black bars). The numbers in parentheses indicate the transgene copy number of each line. All experiments were conducted in duplicates in two independent assays. Data are shown as average tyrosinase expression normalized to the endogenous control *Tbp*, and relative to wild type NMRI + SD.

enough experimental data prevented the accurate analysis of transgene expression in the skin, although animals with a higher number of copies tended to possess higher expression levels of the transgene. Of note, transgene mRNA was present in the brains of some individuals. CorDis-9.2 mapped within an intron of *Pink1*, which is expressed in the central nervous system. Hence, this element could contain additional regulatory elements that may have been functioning in the brain. Nevertheless, transgene mRNA levels did not correlate with transgene copy number in this tissue (**Fig. R 24**).

In all cases, tyrosinase expression in the eyes was homogeneous. Was tyrosinase expression really the same regardless of the transgene introduced or its copy number? Or was there some kind of saturation in the system, that is, a limitation of this method that impeded the observance of any differences? To gain more insight into this issue, the total melanin content in the eyes of all transgenic individuals was quantified (Donatien & Orlow, 1995; Furlan-Magaril et al., 2011) and the average content per transgenic line was calculated (**Fig. R 25**). Apparently, in those mice transgenic for the negative control or for the TYR5-CorDis-9.2 construct, the eyes started to become saturated with pigment with as few as five copies of tyrosinase. On the contrary, the increase in pigment concentration was more gradual when a true insulator, such as αEHS1.4, was included in the vector. Indeed, it seemed there were differences regarding tyrosinase expression in the eyes of all transgenic lines, differences that were unappreciated in the qPCR experiments.

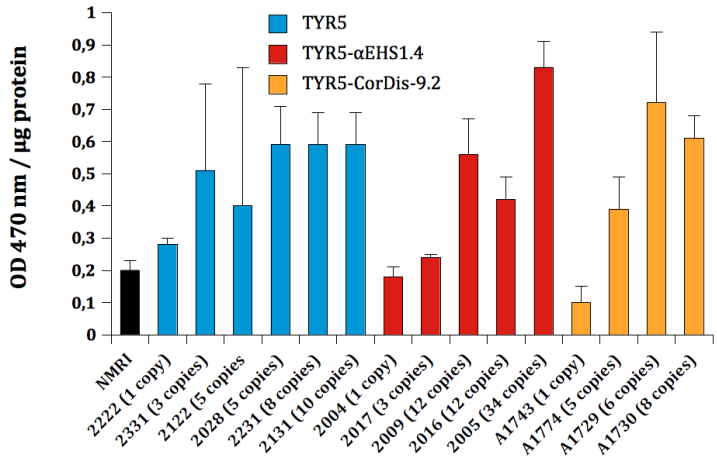


Fig. R 25. *In vivo* testing for barrier activity. Analysis of the total melanin content in the eyes of transgenic mice. The eyes of transgenic individuals representative of each mouse line were extracted and processed to quantify their content in melanin. Average values were calculated and assigned to the corresponding transgenic line (in the x-axis). Transgene copy numbers are indicated in parentheses. Three lines carrying three different constructs are depicted: TYR5 in blue, TYR5-αEHS1.4 in red and TYR5-CorDis-9.2 in orange. An additional non-transgenic control line, NMRI, is shown in black. Data are presented as average melanin content relative to total protein mass + SD.

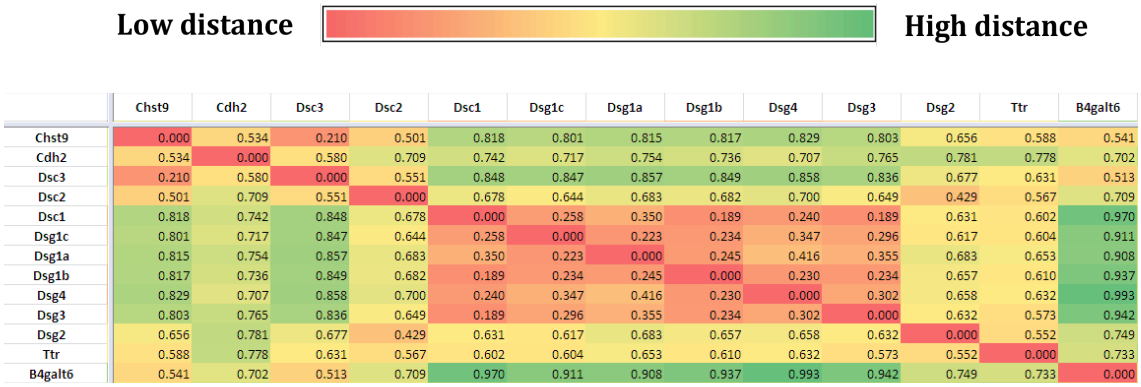
In conclusion, an increase in the level of pigmentation with transgene copy number could be visually observed in albino mice carrying a tyrosinase minigene construct with the CorDis-9.2 element. Nevertheless, the analysis of tyrosinase expression in various tissues of these animals, as well as the melanin content of their eyes, suggested that transgene copy number and expression levels did not strictly correlate. However, a trend could be observed. Hence, CorDis-9.2 did not fully protect from chromosomal position effects in this particular system.

4.2. Development of an Algorithm to Detect Boundaries Flanking Clusters of Co-Expressed Genes: Second Case Scenario

Clusters of genes with shared expression profiles are common in mammalian genomes. Genomic boundaries may flank these clusters, helping to preserve their functional integrity. In fact, boundaries have been found shielding the chicken, mouse and human β-globin loci (reviewed in Amouyal, 2010b), at the 5' end of the mouse *Hoxd* gene cluster (Kmita et al., 2002; Yamagishi et al., 2007) and at the 3' end of the mouse *Igh* locus (Garrett et al., 2005).

4.2.1. Defining Clusters of Adjacent Co-Expressed Genes

The Euclidean distance method generated a set of distance matrices that harbored the distributions of gene expression distances between each gene and the rest of the genes in a given chromosome. There were as many distributions as genes, and they all differed. In order to compare distance values that come from different distributions, the matrices were normalized (see the Materials and Methods section) and converted into heat maps. The representation of the data as heat maps facilitated the visualization of the results (**Fig. R 26**). For example, co-expressed genes would exhibit low expression distances among them, and if they lay next to each other in the linear genome forming a cluster, a group of low distance values would sit on the diagonal. That was the case of the cluster of genes that codifies for the proteins engaged in the formation of the desmosomes: desmogleins (*Dsgs*) and desmocollins (*Dscs*). These structures are intercellular anchoring junctions abundant in the heart and the skin (Alberts et al., 2002). Even though there are nine genes in the cluster, the outermost genes *Dsc3*, *Dsc2* and *Dsg2* were left out of our predictions.



	Chst9	Cdh2	Dsc3	Dsc2	Dsc1	Dsg1c	Dsg1a	Dsg1b	Dsg4	Dsg3	Dsg2	Ttr	B4galt6
Chst9	0.000	0.534	0.210	0.501	0.818	0.801	0.815	0.817	0.829	0.803	0.656	0.588	0.541
Cdh2	0.534	0.000	0.580	0.709	0.742	0.717	0.754	0.736	0.707	0.765	0.781	0.778	0.702
Dsc3	0.210	0.580	0.000	0.551	0.848	0.847	0.857	0.849	0.858	0.836	0.677	0.631	0.513
Dsc2	0.501	0.709	0.551	0.000	0.678	0.644	0.683	0.682	0.700	0.649	0.429	0.567	0.709
Dsc1	0.818	0.742	0.848	0.678	0.000	0.258	0.350	0.189	0.240	0.189	0.631	0.602	0.970
Dsg1c	0.801	0.717	0.847	0.644	0.258	0.000	0.223	0.234	0.347	0.296	0.617	0.604	0.911
Dsg1a	0.815	0.754	0.857	0.683	0.350	0.223	0.000	0.245	0.416	0.355	0.683	0.653	0.908
Dsg1b	0.817	0.736	0.849	0.682	0.189	0.234	0.245	0.000	0.230	0.234	0.657	0.610	0.937
Dsg4	0.829	0.707	0.858	0.700	0.240	0.347	0.416	0.230	0.000	0.302	0.658	0.632	0.993
Dsg3	0.803	0.765	0.836	0.649	0.189	0.296	0.355	0.234	0.302	0.000	0.632	0.573	0.942
Dsg2	0.656	0.781	0.677	0.429	0.631	0.617	0.683	0.657	0.658	0.632	0.000	0.552	0.749
Ttr	0.588	0.778	0.631	0.567	0.602	0.604	0.653	0.610	0.632	0.573	0.552	0.000	0.733
B4galt6	0.541	0.702	0.513	0.709	0.970	0.911	0.908	0.937	0.993	0.942	0.749	0.733	0.000

Fig. R 26. Heat map of a portion of the normalized distance matrix for chromosome 18. Genes are arranged in rows and columns in the same order as they appear in the genome. Expression distance values range from 'zero' (red) to 'one' (green). The diagonal represents the distance between each gene with itself and therefore takes the minimum value 'zero'. A group of low distance values over the diagonal are easily detected and correspond to clusters of adjacent co-expresed genes.

A closer look at the genomic context of this cluster revealed numerous binding sites for CTCF and the cohesin subunit Rad21 (**Fig. R 27**). As described in the Introduction section, CTCF-cohesin complexes are involved in the establishment of long-range interactions that help in the organization of the nuclear architecture. The fact that they were found flanking the predicted cluster of co-expression suggested that they might participate in the formation of a three-dimensional structure that organizes and protects the locus. However, this was only a hypothesis that needed to be functionally validated.

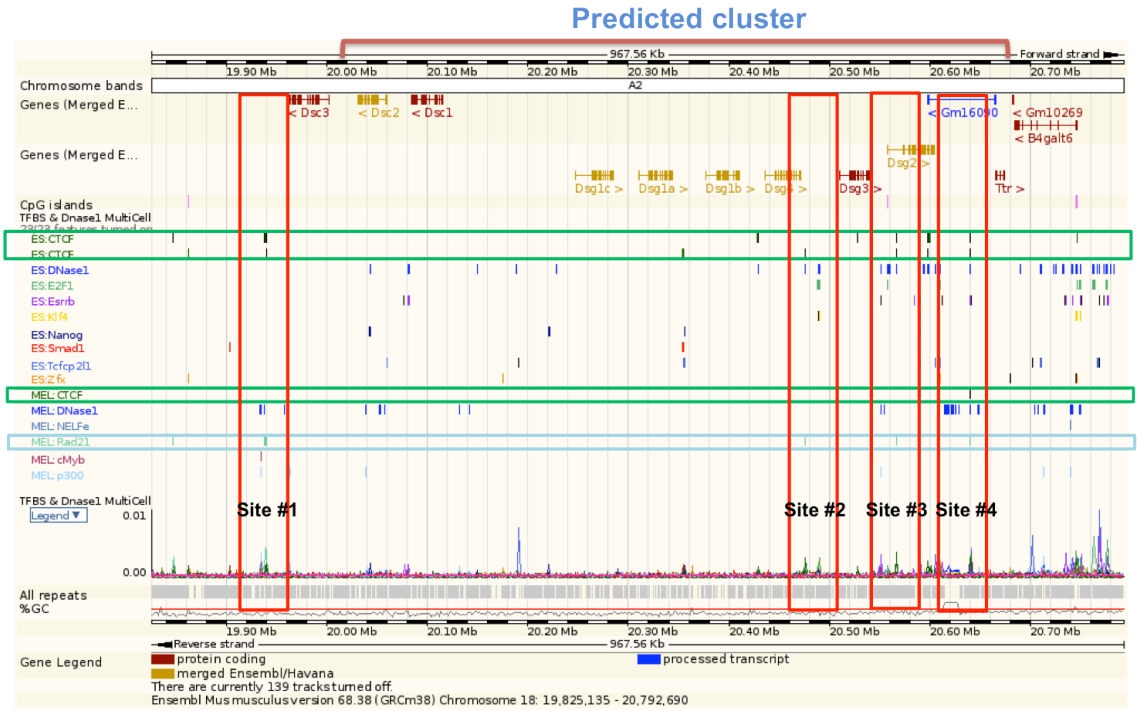


Fig. R 27. Genomic context of the cluster of desmogleins and desmocollins in chromosome 18. Genes are colored in red, orange or blue. The binding sites for different factors are shown in boxes of different colors (see legend on the left). CTCF and Rad21 binding sites tracks are framed in green and light blue, respectively. Candidate docking sites for loop formation are composed of CTCF and Rad21 binding sites (red rectangles, site numbers one to four). The nearest gene upstream of *Dsc3* is 3 Mb away (*Cdh2*).

The algorithm also highlighted the α cluster of protocadherin (*Pcdh*) genes. This locus is plagued with binding sites for CTCF and cohesin. It has been shown elsewhere that, in this case, these proteins participate in the establishment of DNA loops within the cluster. The loops favor the interactions between downstream enhancers and individual promoters, playing a fundamental role in *Pcdh*α promoter choice (Monahan et al., 2012; Guo et al., 2012).

Hence, the algorithm was proven successful in identifying the well-established *Pcdh*α cluster, truly organized by insulator-related proteins. Functional studies in the locus of the desmosomal proteins would determine if the algorithm could also pinpoint new clusters of co-expressed genes and the genomic boundaries associated with them.

4.2.2. Co-Expressed Genes also Cluster in Space

Accumulating evidence suggests the existence of transcription factories in the nucleus (Iborra et al., 1996). RNA polymerases and other protein complexes necessary for transcription are abundant in these regions. Some authors postulate that cell-type specific genes or genes that belong to the same biological pathway come together in space and cluster in specialized transcription factories that contain a high concentration of specific transcription factors (Schoenfelder et al., 2010; Li et al., 2012).

In an attempt to predict intra-chromosomal long-range interactions based on gene expression data, the genes in chromosome 18 were clustered according to their expression profiles using the *K*-means algorithm. Two functional annotation tools, FatiGO (http://bioinfo.cipf.es/babelomicswiki/tool:fatigo; Medina et al., 2010) and DAVID (http://david.abcc.ncifcrf.gov/; Huang et al., 2009a; Huang et al., 2009b), permitted the search of biological pathways significantly enriched in each cluster. Common biological pathways would explain why the genes of each group displayed very similar expression profiles. Special attention was paid to the cluster that hosted most of the desmosomal genes (**Table R 10**). For this cluster, two groups of significantly enriched functional annotations were found: cell-cell adhesion (GO:0016337) and related, and gamete generation (GO:0007276) and associated terms. As expected, genes that coded for the desmocollins and desmogleins (*Dsc1*, *Dsg1a*, *Dsg1b*, *Dsg1c*, *Dsg3*) were associated with cell-cell-adhesion-related GO terms. Interestingly, an additional gene was included in this subset: *Spink5*. This protease inhibits other proteases in charge of degrading desmosomes and hence, is a positive regulator of the route (reviewed in Ovaere et al., 2009).

Table R 10. Representative cluster of co-expressed genes in chromosome 18. Genes from chromosome 18 were grouped in thirty clusters according to their expression profiles. Genes associated with cell-cell adhesion processes (GO:0016337) are highlighted in blue and bold.

Example. Cluster of co-expressed genes in chromosome 18				
1700034E13Rik	*Cabyr*	*F730048M01Rik*	*Lims2*	*Rnf138*
1700065I17Rik	*Ccdc11*	*Fam170a*	*Lyzl1*	*Slc23a1*
2010001M09Rik	*Cetn1*	*Fbn2*	*Mospd4*	*Slc25a2*
4833403I15Rik	*Ctdp1*	*Fbxo15*	*Mppe1*	*Spata24*
4921524L21Rik	*Cxxc1*	*Ftmt*	*Nme5*	*Spink12*
4921528I01Rik	*Diap1*	*Gm10265*	*Osbpl1a*	***Spink5***
4930503L19Rik	*Dnd1*	*Gm1614*	*Pabpc2*	*Spry4*
Afap1l1	***Dsc1***	*Gm4841*	*Pard6g*	*Stard6*
Alpk2	***Dsg1a***	*Gm94*	*Plac8l1*	*Stk32a*
Ankrd29	***Dsg1b***	*Gykl1*	*Poli*	*Taf4b*
Arsi	***Dsg1c***	*Hbegf*	*Polr2d*	*Trim36*
AW554918	***Dsg3***	*Hdhd1a*	*Prdm6*	*Zfp35*
Bambi	*Eif1a*	*Iigp1*	*Psma8*	*Zfp474*

An analysis of the enrichment in transcription factor binding sites in the promoters of the genes of the desmosome cluster (including *Spink5*) yielded no significant results. For this reason, the promoters of each gene were analyzed, one by one, using TRANSFAC® database (http://www.biobase-international.com/product/transcription-factor-binding-sites; Matys et al., 2006). Only the promoters (± 1 Kb from the TSS) of *Dsc3* and *Dsc1* hosted binding sites for transcription factors, particularly, from the C/EBP family: C/EBPbeta appeared to bind both promoters, whereas C/EBPdelta would only bind that of *Dsc3*. Not surprisingly, C/EBPbeta plays an important role in the regulation of the differentiation of keratinocytes (Zhu et al., 1999), where desmosomes are most abundant. Additionally, other transcription factors associated with the regulation of epithelial-related processes would also bind in the vicinity of these genes (**Table R 11**). This would explain why their expression profiles are so similar.

COXPRESDB (http://coxpresdb.jp/; Obayashi et al., 2013) is yet another database that integrates gene expression information from a variety of sources and provides a list of genes that co-express with the query. Importantly, the search is not limited to one chromosome in particular. Instead, all the genes in the genome are put into question. Using one of the genes that code for the desmogleins as the query, the algorithm confirmed the co-expression of several genes from the genomic cluster involved in the

formation of desmosomes, as well as of *Spink5*, in a manner independent of the expression data used to derive our conclusions (**Fig. R 28**).

Table R 11. Transcription factors binding in the vicinity of the genes involved in the formation of desmosomes. TRANSFAC functional analysis revealed that several transcription factors would bind near (± 150 Kb) –but not within- the promoters of the genes in the desmosome cluster (including *Spink5*). Only transcription factors that regulate epithelial-related processes are shown.

Transcription factor…	…binds in the vicinity of…	Regulation of epithelial-related processes
Brg1 (Smarca4)	*Dsc3, Dsc2, Dsc1, Dsg4, Dsg2*	Keratinocyte terminal differentiation (Indra et al., 2005)
Cdx-2	*Dsc3, Dsc2, Dsc1, Dsg1c, Dsg1b, Dsg1a, Dsg4, Dsg3, Dsg2, Spink5*	Development and differentiation of epithelial intestinal cells (Aoki et al., 2011)
Foxa2 (HNF-3β)	*Dsc2, Dsc1, Dsg1c, Dsg3, Dsg2, Spink5*	Differentiation of lung epithelial cells (Wan et al., 2005)
Tcf7l1 (TCF-3)	*Dsc2, Dsg4, Spink5*	Epidermal terminal differentiation (Merrill et al., 2001)
Sox2	*Dsc1*	Differentiation of lung epithelial cells (Gontan et al., 2008)
Stat3	*Dsc1, Dsg1c, Dsg2*	Keratinocyte proliferation and migration (Sano et al., 2000; Gartsbein et al., 2006)
C/EBPalpha	*Dsg1c, Spink5*	Development and proliferation of lung epithelial cells (Berg et al., 2006; Yang et al., 2011a)
Klf4 (Gklf)	*Dsg4, Dsg2*	Proliferation, migration, differentiation and positioning of intestinal epithelial cells (Ghaleb et al., 2011) Epidermis development (Segre et al., 1999)

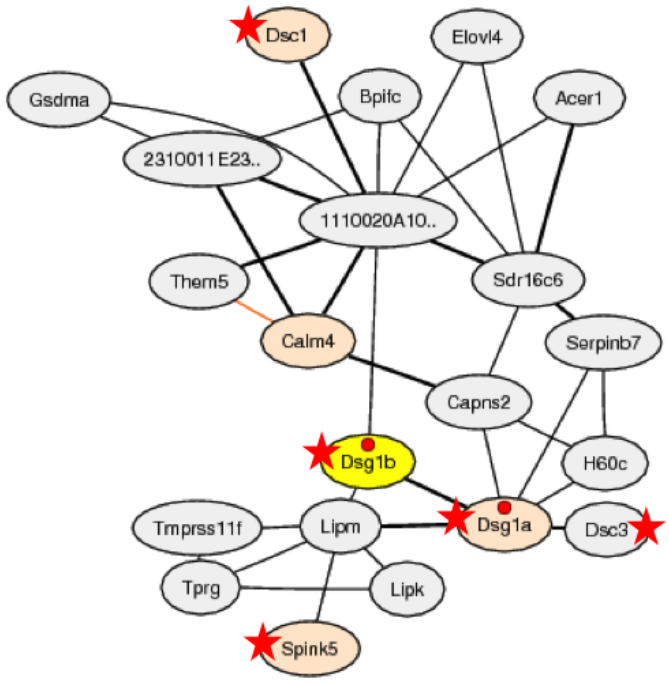


Fig. R 28. Cluster of genes that co-express with *Dsg1b*. The diagram, obtained from COXPRESDB, depicts the set of genes with the same expression profile as *Dsg1b* (yellow node). Orange nodes and edges indicate conserved co-expression between human and mouse orthologs. Red stars mark the genes involved in the formation of desmosomes.

4.2.3. Functional Validation

A more in-depth analysis of the Euclidean distance matrix generated from expression data for chromosome 18 revealed the presence of a cluster of co-expressed genes engaged in the formation of desmosomes. Binding sites for the insulator-related proteins CTCF and Rad21 flanked this cluster. Also, it was found that *Spink5*, a gene that resides 23 Mb further downstream in the chromosome, shared the same expression pattern as the desmosomal proteins. Do the CTCF-cohesin complexes that bind at both ends of the cluster interact with each other? Do they create a loop that isolates the desmosomal genes, and favors an optimal environment for coordinated gene expression? Moreover, is there any interaction between the cluster and *Spink5*?

Long-range interactions can be structural and thus conserved in the majority of the cell types of the organism. Also, they can be dynamic and responsive to transcriptional needs (Hou & Corces, 2012; Phillips-Cremins et al., 2013b). For example, the interaction between a promoter and a distal enhancer that drives gene expression only early in development will be discontinued at a later stage. Furthermore, the gene expression program of each cell type is specific to that cell type, and therefore so are the enhancer-promoter interactions they establish.

Chromosome conformation capture (3C) and related methods constitute very popular tools to evaluate long-range interactions between distant genomic locations (reviewed in Simonis et al., 2007). Obviously, they have to be applied in those cells in which interactions are expected. Desmogleins and desmocollins make desmosomes mainly in keratinocytes. For this reason, the murine epidermal keratinocyte derived COCA cell line was used in the analysis (Segrelles et al., 2011).

Undifferentiated COCA cells do not establish desmosomal junctions. Hence, a cocktail of factors was added to the cells in culture to induce their differentiation. Samples at zero, twenty-four, forty-eight and seventy-two hours post induction were taken so as to measure the level of expression of all the genes under study by SYBR green real-time PCR. The purpose of this experiment was two-fold: to ensure that the desmosomal genes were indeed expressed in these cells, and to choose the time point at which most, if not all, of the genes were expressed in order to prepare the cells at the right stage for the 3C assay. Mouse L929 fibroblastic cell line was included as a non-expressing negative control (**Fig. R 29**).

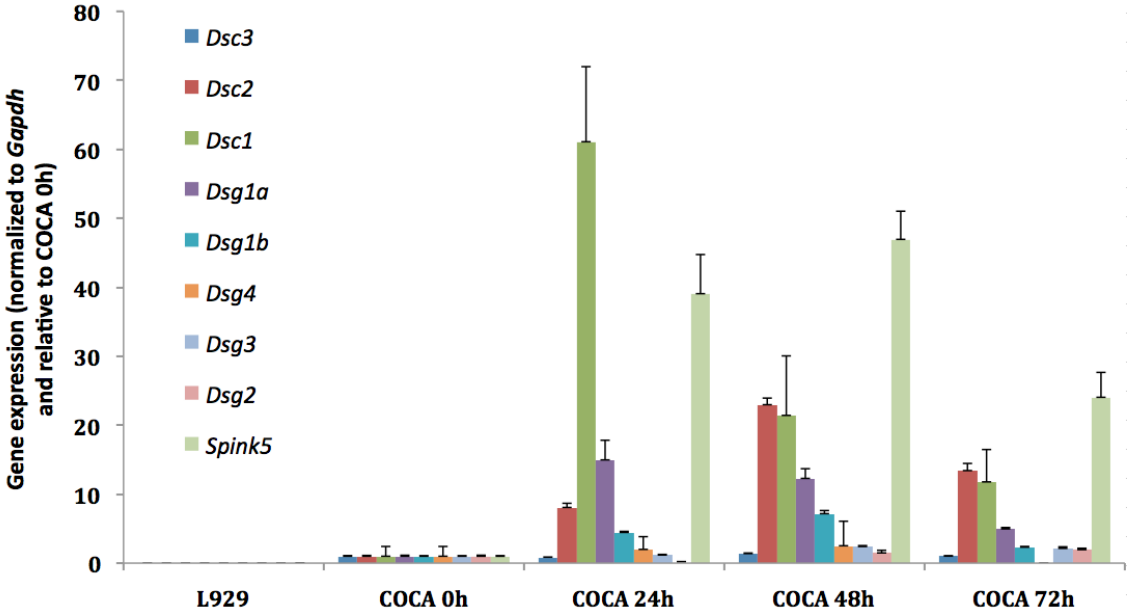


Fig. R 29. Expression profile of the desmosomal genes in the COCA cell line. The expression of the clustered genes involved in the formation of desmosomes and *Spink5* was quantified by SYBR green real-time PCR in the COCA cell line at various differentiation time points (zero, twenty-four, forty-eight and seventy-two hours post induction). *Dsg1c* mRNA was not detected in any sample and thus, does not appear in the graphic. L929 cells were included in the analysis as a non-expressing negative control. Samples were measured in duplicates in two independent experiments. Data are shown as gene expression normalized to the endogenous control *Gapdh* and relative to undifferentiated COCA cells (zero hours) + SD of the duplicates.

As expected, mRNA coding for the desmosomal proteins was absent in L929 cells. At twenty-four hours post induction, COCA cells expressed most of the genes at a moderate to high level. Expression became more abundant at forty-eight hours and started disappearing at seventy-two hours. *Dsg1c* mRNA was not detected in any case, even if two pairs of primers were tested (one of them was successfully used in Whittock, 2003).

Most desmosomal genes were being expressed in COCA cells at twenty-four hours post induction of differentiation, which suggested that the long-range interactions between the regulatory elements in the locus had already been established. This was the time point chosen to carry out the 3C-qPCR assay.

As described above, the locus contained four binding sites for the CTCF-cohesin complex (**Fig. R27**). When using an anchor primer mapping to site #4, a moderate interaction with site number one was detected in both cell types (**Fig. R 30**). Note that these regions are ~700 kb apart, making it highly improbable that this observation was simply caused by random collision events. In contrast, sites number three and four are rather close to each other (~73 kb). Yet, no interactions were revealed between them. The fact that this interaction seemed to exist in expressing and non-expressing cells suggested

that it had a structural role and participated in the organization of the chromatin inside the nucleus. This possibility will be further discussed in the following section.

It could be argued that the weak interaction established between sites number three and four in COCA cells resulted from the proximity of both sites. However, the relative interaction frequency between the anchor and the control primer that mapped downstream from site number three is a little lower, indicating that an interaction between those sites may actually be established.

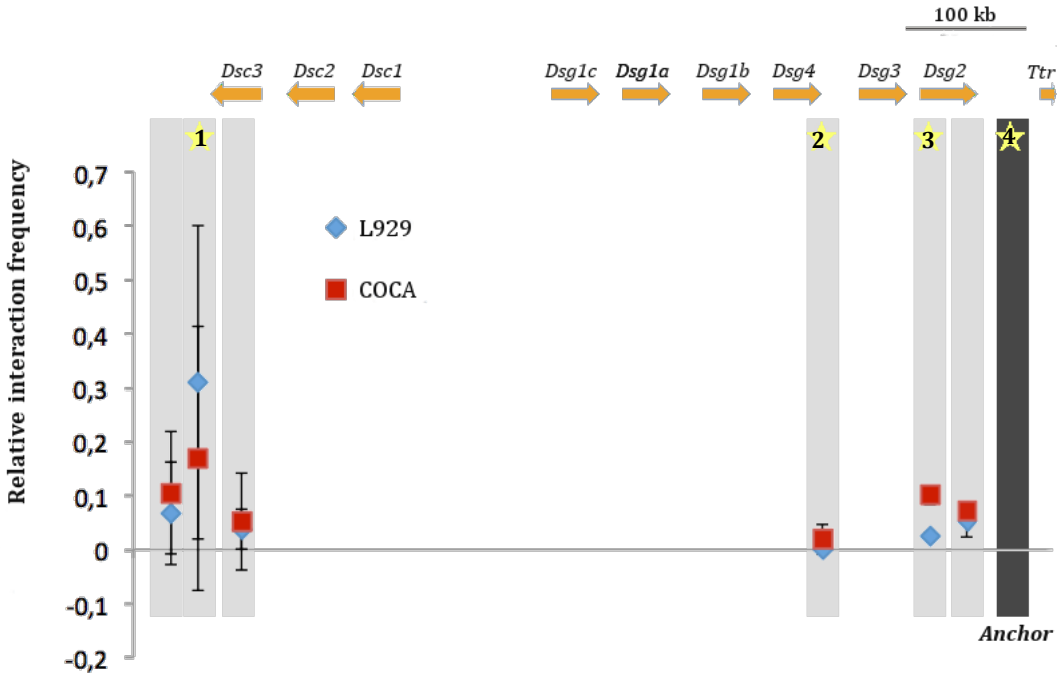


Fig. R 30. 3C-qPCR analysis of long-range interactions in the cluster of desmosomal genes. The interactions between the four CTCF-cohesin sites (yellow stars) in the desmosomal clustered genes (upper panel) were questioned by a 3C-qPCR assay in L929 (blue) and COCA (red) cells. The anchor primer (black bar) was set in the *Hind*III fragment that contained site number four. Light gray bars highlight the rest of the primers. Control primers were designed ± 30 kb from the fragments that contained the CTCF-cohesin binding sites. Quantification of the interactions was carried out by SYBR green real-time PCR, using triplicates for each sample in, at least, two independent experiments. Results are presented as mean interaction frequencies relative to those at the control *Ercc3* locus ± SD of the triplicates.

Additional assays failed to detect any long-range interactions between site number four and two other sites present in the *Spink5* locus. More experiments with a different anchor primer (or even another 3C-related method such as 4C) may be needed to draw a definite conclusion on this issue.

These data suggested that the CTCF-cohesin sites that flanked the cluster of the desmosomal genes interact with each other. This means that the algorithm also succeeded in finding new boundaries that shield a cluster of co-expressed genes, which was the objective of the study.

4.3. Functional Validation of an Algorithm that Identifies Boundaries Partitioning the Chromatin into Active and Silenced Domains: Third Case Scenario

Chromatin is structured into highly condensed silenced heterochromatin and more transcriptionally open euchromatin. This organization is crucial for the correct expression patterning of the cells: upon differentiation, once the cell has committed to a given lineage, the genes that are not necessary anymore will be silenced and confined in heterochromatin domains, whereas tissue-specific genes will remain turned on and in euchromatin regions. The borders between heterochromatin and euchromatin need to be perfectly controlled to guarantee the correct functioning of the cell, and here is where genomic boundaries or insulators may come into play.

4.3.1. Description of the Algorithm

King Jordan and collaborators developed an algorithm that sought to evaluate whether MIR elements functioned as genomic boundaries in the human genome. MIRs (Mammalian-wide Interspersed Repeats) belong to the SINE family of retrotransposable elements and contain B-box promoter elements, which have been associated with insulator activity in yeast, mice and humans (for review see Kirkland et al., 2012). They used human $CD4^+$ cells as a model because of the abundant functional data available for this particular cell type.

The original pool of 324,863 MIR elements with intact B-boxes in the human genome was sequentially filtered until only 1,178 elements remained. This subset was bound by RNA polymerase III, partitioned the chromatin into active *versus* repressive domains

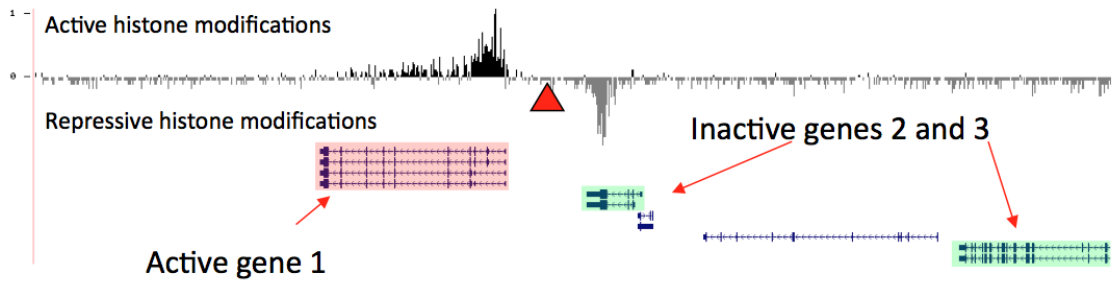


Fig. R 31. MIR elements partition the chromatin into active *versus* silenced domains. In human $CD4^+$ cells, MIR elements (red triangle) are found at the boundaries between euchromatin and heterochromatin domains, which are defined by the presence of clusters of active or repressive histone modifications, respectively. Genes located in the euchromatic region remain active (red), whereas genes at the repressive side remain silent (green). Figure from King Jordan (unpublished data).

based on the analysis of the histone marks present at both sides of the elements, and segregated active *versus* silent genes (**Fig. R 31**). In addition, CTCF binding sites were enriched in the vicinity of these elements. Hence, these 1,178 MIRs may be playing a role in the establishment of genomic boundaries in human CD4+ cells.

4.3.2. Functional Validation

In order to validate the insulator activity of the elements unraveled by the algorithm, *in vitro* and *in vivo* enhancer-blocking assays were performed with four of the 1,178 elements described. The elements were named according to the chromosome they belonged to (chromosomes 1, 2, 4 and 11).

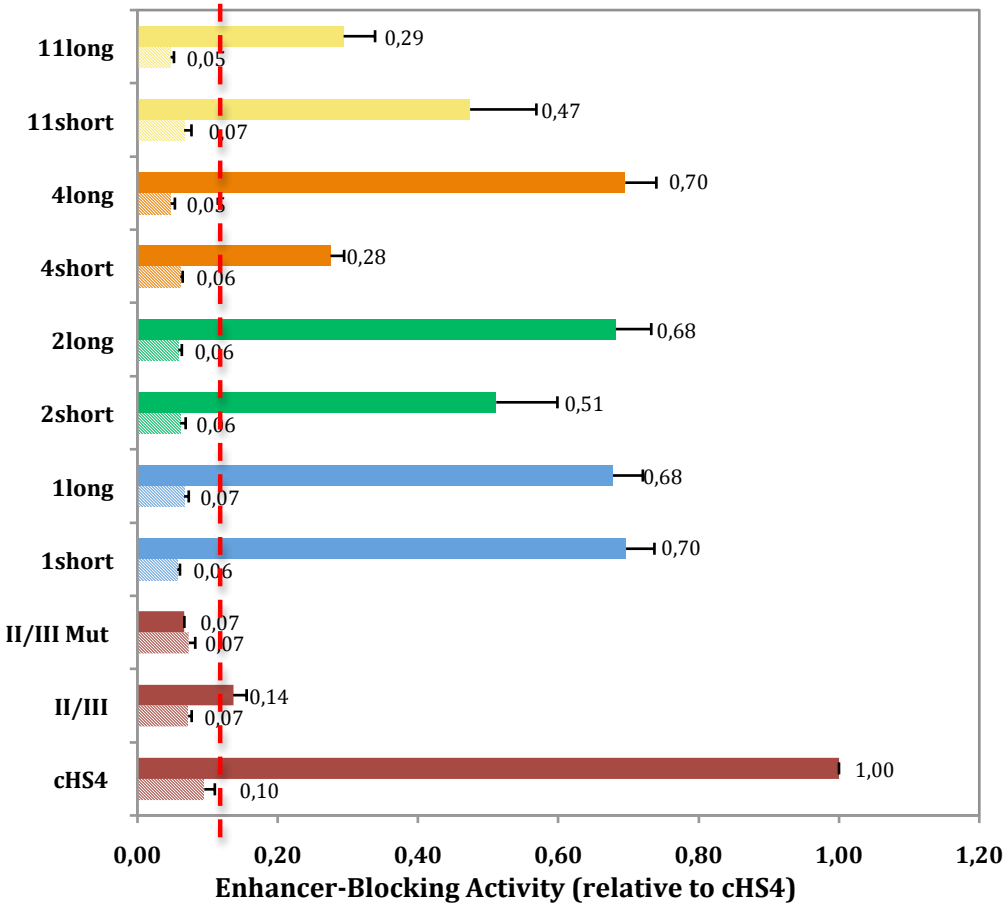


Fig. R 32. Evaluation of the *in vitro* enhancer-blocking activity of selected human MIR retrotransposable elements. The fold enhancer-blocking activity of short and long versions of human genomic loci that contained MIRs is shown. Normalization of the activities to that of the cHS4 control was conducted. Control elements are shown in red, whereas the elements that belong to the same retrotransposon are grouped under the same color. Solid bars represent elements cloned between the enhancer and the promoter; stripped bars correspond to the same elements but cloned upstream of the enhancer. At this position, only silencing elements shut down luciferase expression. A sequence conveys insulator activity if its fold enhancer-blocking activity is larger than 0,12 (discontinued red line). All experiments were carried out in triplicate in three independent assays. Data are shown as mean + SEM of the independent assays (n=6).

Because it was possible that insulator function depended not only on the MIR retrotransposon but also on adjacent sequences, both short and long versions of each element were tested. The short versions (approximately 200 bp long) referred exclusively to the MIR elements, whereas the longer versions included 800 bp around the elements.

In vitro enhancer-blocking assays in HEK 293 cells revealed that all eight sequences had a high insulator activity and did not simply act as silencing elements (**Fig. R 32**). Both short and long versions of elements 1 and 2 had similar enhancer-blocking activities. However, 4short performed far worse than its longer version, suggesting that the insulator activity was dependent not solely on the MIR element, but also on the adjacent sequences. The opposite appeared to be occurring with element 11, though: the enhancer-blocking activity of the short version almost doubled that of the longer one, possibly implying the existence of other regulatory sequences that were counteracting the insulator effect of the retrotransposon at this particular locus.

To further validate these MIR-containing sequences as insulators, *in vivo* enhancer-blocking assays were carried out in zebrafish (**Fig. R 33** and **34**). Surprisingly, neither the long, nor the short version of element 1 exhibited insulator activity, even if they were proven *in vitro*. Regarding the rest of the elements, only 2short and 11long showed statistically significant enhancer-blocking activity. These results apparently contradicted the *in vitro* assays. However, as shall be discussed below, *in vivo* testing in zebrafish relies on the assumption that all necessary elements for the establishment of these genomic boundaries are conserved from zebrafish to human, and that may not be true in all cases.

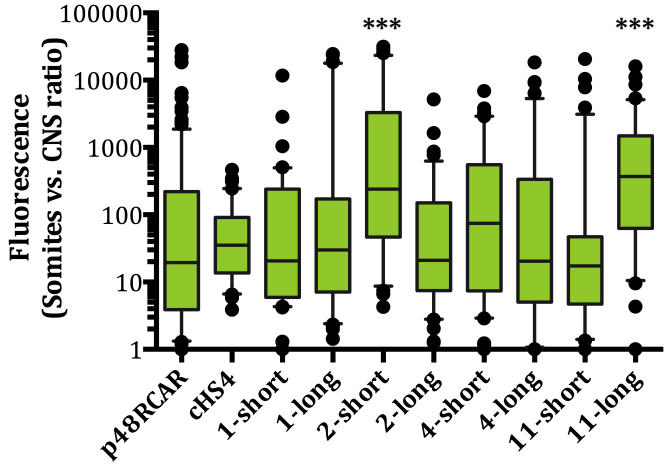


Fig. R 33. *In vivo* enhancer-blocking assay in zebrafish. MIR elements. Selected MIR-containing sequences were cloned between the CNS enhancer and the muscle promoter in p48RCAR. **A.** GFP fluorescence ratio between the somites and the midbrain (delimited by a white line) indicates the enhancer-blocking activity. Data are shown as box-plots that integrate the fold enhancer-blocking activity (fluorescence ratio somites/CNS relative to the basal level represented by p48RCAR) for all transgenic individuals for a given construct. Median test; * significant at p-value < 0.05; ** significant at p-value < 0.01; *** significant at p-value < 0.001.

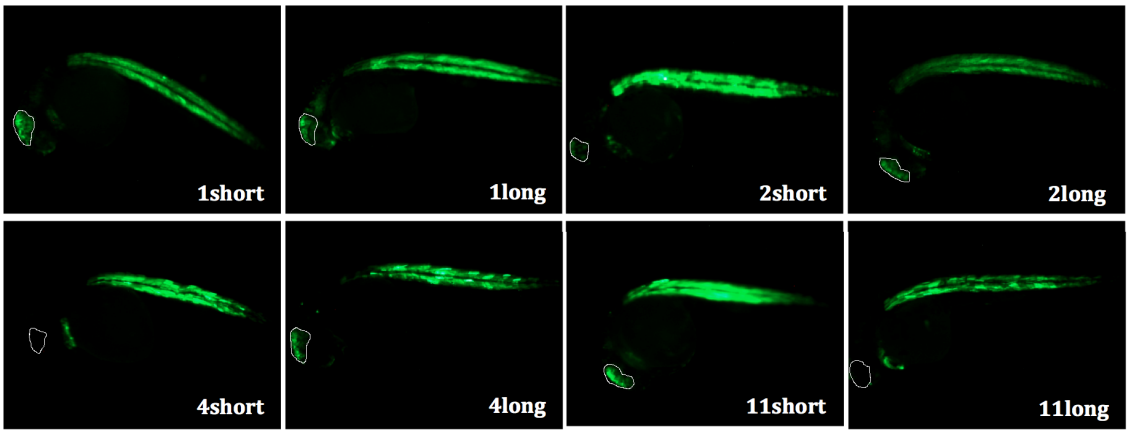


Fig. R 34. Representative transgenic zebrafish individuals obtained in the analysis of the *in vivo* enhancer-blocking activity of human MIR elements.

Therefore, this new algorithm, aimed at finding insulators that define the boundaries between active and silenced chromatin, was able to predict a new type of insulator element in human based on MIR retrotransposons (Wang et al., submitted).

5 DISCUSSION

Enhancers, silencers and LCRs are the most widely recognized and studied regulatory elements. Insulators, however, are not equally well-known. Yet, they are fundamental since they delimit expression domains, ensuring the correct spatio-temporal expression patterning of the genes embedded within them (Engel & Bartolomei, 2003; Molto et al., 2011). Moreover, insulators can be used as genetic tools to alleviate chromosomal position effects in ectopic constructs intended for gene transfer applications (i.e., production of transgenic animals or gene therapy) (Recillas-Targa et al., 2004; Molto et al., 2011). But few insulators have been described and well characterized so far. Among them, special attention has been paid to the 1.2-kb element from the chicken β-globin locus, whose properties have been extensively exploited with variable success (reviewed in Giraldo et al., 2003b; Molto et al., 2011; i.e., Puthenveetil et al., 2004; Klopstock et al., 2007).

For these reasons, the work developed here aimed at, and succeeded in, describing new functional insulators in mammalian genomes in order to broaden our knowledge on gene expression regulation, as well as to increase the repertoire of tools that can be used to produce more efficient transgenic constructs. This was achieved by following a two-step strategy: first, by developing several algorithms to find putative insulators in a genome-wide unbiased fashion in mice and second, by functionally validating them, along with other elements found in the human genome elsewhere (Wang et al., unpublished results), using both *in vitro* and *in vivo* experimental approaches.

5.1. Where Would Boundaries Be Expected?

As stated in the Introduction section, there are different mechanisms of insulation that rely on different proteins and DNA motifs. The only thing that all boundaries have in common is their function, and that is the reason why boundaries are considered to be functionally rather than structurally defined (Engel & Bartolomei, 2003).

Previous genome-wide screens tried to find insulators by taking advantage of structural features such as CTCF binding (Barski et al., 2007; Kim et al., 2007; Xie et al., 2007; Jothi et al., 2008; Cuddapah et al., 2009; Chen et al., 2012; Nakahashi et al., 2013). However, this type of approach is not thorough for two reasons. First, not only does it miss already described CTCF-independent insulators like S/MARs (Millot et al., 2003) or tRNAs (Lunyak et al., 2007; Roman et al., 2011), but it also fails to discover new mechanisms of insulation. Second, not all CTCF binding sites -whether bioinformatically predicted or

functionally identified by chromatin immunoprecipitation techniques- behave as boundaries (Sanyal et al., 2012). In contrast to these early screens, the work presented here is unbiased regarding structure, because it does not impose any *a priori* condition like CTCF binding. Instead, the focus was placed on searching for those genomic locations where the presence of boundary function is expected because there is a biological need for such function. Thus, it was hypothesized that boundary activity could be found separating genes with opposite expression patterns to create independent expression domains, flanking clusters of co-expressed genes to ensure their coordinated regulation and expression, or establishing a barrier between open euchromatic and condensed heterochromatic domains to maintain separate independent active and silenced domains.

5.1.1. Separating Genes with Opposite Expression Patterns

As extensively reviewed elsewhere (Hurst et al., 2004), adjacent genes are usually co-expressed (i.e., Cohen et al., 2000). One of the possible explanations for this observation could be that these genes are under the influence of (or co-regulated by) the same set of transcription factors that bind in their vicinity, causing their co-expression. The fact that some adjacent genes deviate from the norm and exhibit opposite expression patterns suggests the existence of some sort of mechanism that restricts the interactions between the regulatory elements and the promoters at those loci. The enhancer-blocking activity of some insulators may be responsible for this phenomenon in some cases.

Two genes differ in their expression patterns if they are expressed in different tissues at a given developmental stage, or conversely, in the same tissue but at different developmental stages. The majority of gene expression experiments are conducted at a single developmental stage -usually the adult stage- whereas gene expression data throughout development is scant and solely available for a limited set of genes. Thus, the focus of this work was only placed on finding the first type of pairs; that is, those adjacent pairs of genes whose patterns of expression are opposite at the adult stage.

Therefore, the expression profiles at the adult stage of all the genes in the mouse genome were evaluated using two algorithms that employ different distance measures: Pearson's correlation or Euclidean distance. These measures are the most commonly used by gene expression clustering algorithms (D'haeseleer, 2005). Pearson's correlation serves to reveal groups of genes whose expression patterns follow the same trend across a panel of experimental samples –tissues in this work. For example, two genes would be positively correlated if they are expressed in the same set of tissues, but their actual expression

levels –whether high or low– are irrelevant (genes B and C in **Fig. D 1**). Euclidean distance, however, places more emphasis on absolute differences in expression levels than on trends (genes A and B in **Fig. D 1**).

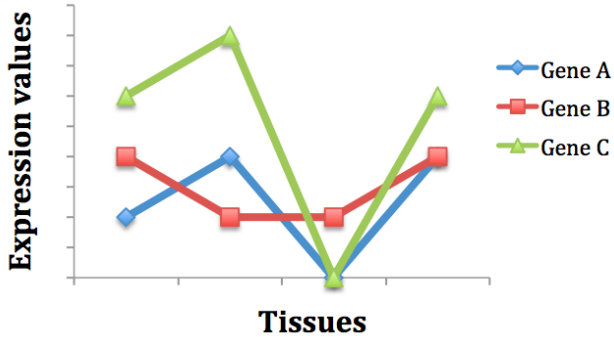


Fig. D 1. Pearson's correlation and Euclidean distance measures. The expression patterns of three genes (A-C) are considered across a panel of tissues. Pearson's correlation would indicate that genes B and C conform the most similar pair because they share the same expression trend, whereas Euclidean distance would argue that gene B is closer to gene A than to gene C because their absolute expression levels differ less.

The aGEM platform, from which the information was retrieved, integrates expression data from various sources and codifies it into three levels: 'zero', 'one' or 'two' corresponding to 'no', 'low' or 'high' gene expression in 425 different anatomical structures at a given developmental stage. The correlation analysis was performed by taking into account these three levels, whereas data were further simplified to only two levels ('yes' or 'no' expression) in the case of the algorithm that used Euclidean distances. Hence, due to the narrow range of gene expression values available in this particular study, both algorithms were expected to yield similar results. Indeed, there was substantial overlap in the results between the two algorithms (**Fig. R 6**). Nevertheless, the Euclidean distance method produced many more pairs than the Pearson's correlation method (6,294 vs. 1,212). Probably, the imputation of missing values conducted in the former was able to rescue some true positives that had been overlooked by the correlation algorithm simply because there were not enough data to carry out the analysis. At the same time, however, because the imputation process was conservative, it is possible that the true experimentally measured differences in the expression patterns of some gene pairs were diluted by the artificial addition of similarities (see section 4.1.1.2 for more details on the imputation process). Hence, the Euclidean distance algorithm probably produced a number of false negatives that might have been conversely recovered by the correlation algorithm method. In any case, the major source of false negatives in any of the methods was probably the lack of expression data for many genes, which is the reason why imputation was necessary in the first place. This limitation can presumably be overcome in the long run after deeper expression analyses of all the genes in the mouse genome.

Importantly, the loci of the gene pairs selected as hits by both algorithms were significantly enriched in insulator-related sequences (**Fig. R 5**), and contained well-known insulators (**Table R 5**), but not all of them. Other reasons apart from the massive lack of gene expression data available in the databases may explain this fact. First, those insulators that regulate imprinting, such as the one at the *Igf2/H19* locus (see **Table R 5** and **Appendix I-1** for specific examples), would be missed by these methods. Imprinting provokes the parent-of-origin-specific expression of each gene in a pair in such a way that one gene will only be expressed from the maternal allele, whereas the other will only be expressed from the paternal one (Ideraabdullah et al., 2008). However, both genes are usually expressed in the same set of tissues –though from different parent-of-origin alleles– so the algorithms would fail to detect them.

The second reason only applies to the correlation method, which only selected genes with strictly opposite expression patterns. However, insulators may also be found between genes whose expression patterns are independent, that is, coincident in some tissues and divergent in others. This picture emerges more difficult to resolve in a large-scale bioinformatic screening and was purposely left unexplored to avoid filling the hit lists with many false positives. In this approach, specificity was prioritized over sensitivity.

Finally, the third reason is specific to the Euclidean distance method. Here, pairs formed by genes with highly restricted expression patterns, such as *Rxrb* and *Col11a2* (**Table R 5** and **Appendix I-1**), were probably missed because of the way distances were calculated: these genes are more alike than different because they are not expressed in the majority of tissues (which count as coincidences and do not contribute positively to the distance measurement) and only differ in a few cases (**Fig. D 2**).

Nevertheless, both algorithms were able to capture many currently known mouse insulators in a genome-wide and unbiased fashion, indicating that they were promising tools for the discovery of new elements.

In order to functionally validate the algorithm, several gene pairs were selected according to diverse criteria: intergenic distance, promoter orientation and biological/biomedical significance (see the section 4.1.4.1). Nine out of the ten pairs used were chosen among the top hits obtained by either or both of the algorithms. The tenth pair, formed by the *Rbm25* and *Psen1* genes, was additionally included because mutations in human *PSEN1* have been associated with early onset familial Alzheimer's Disease (OMIM ID: 104311). Thus, any information regarding its regulation, even if it comes from its mouse ortholog, may contribute to the understanding of the molecular pathogenesis of

the disease. Also, the Euclidean distance value between *Rbm25* and *Psen1* bordered the significance threshold established for this method. As seen in **figure R 13D**, this locus contains insulator elements and thus, it could be considered as a false negative.

A

$$D(A,B) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}$$

D: Gene expression distance between genes A and B
n: number of anatomical structures
A_i, B_i: Gene expression value for genes A and B in each tissue

B

Case 1

	Bone Marrow	Epidermis	Heart	Kidney	Liver	Lung	Ovary	Spleen
Gene A	No	No	No	No	No	No	No	Yes
Gene B	Yes	No	No	No	No	No	No	No
Distance	1	0	0	0	0	0	0	1

Overall distance $\sqrt{2}$

C

Case 2

	Bone Marrow	Epidermis	Heart	Kidney	Liver	Lung	Ovary	Spleen
Gene C	No	Yes	No	Yes	No	Yes	No	Yes
Gene D	Yes	No	Yes	No	Yes	No	Yes	No
Distance	1	1	1	1	1	1	1	1

Overall distance $\sqrt{8}$

Fig. D 2. Euclidean distance algorithm. A. Euclidean distance formula. Tables **B** and **C** exemplify two cases of gene pairs (in rows) with very different expression profiles across a panel of tissues (in columns). The overall distance in case 1 is much lower than in case 2. The first pair would probably be missed by the algorithm, resulting in a false negative.

The loci of these gene pairs were bioinformatically analyzed in the search for DNA sequences that may be exerting insulator function. Bioinformatic predictions of new *cis*-regulatory modules (CRMs) usually consider two major factors: evolutionary conservation and the presence of the consensus binding site for the regulatory protein under study (Narlikar & Ovcharenko, 2009; Hardison & Taylor, 2012).

On the one hand, comparative genomics approaches assume that those DNA sequences that are serving a regulatory function are under positive selection and will be conserved throughout evolution (Maston et al., 2006). These approaches have been successfully used on many occasions (reviewed in Boffelli et al., 2004). However, they are not infallible. Only those sequences under strong purifying selection would be uncovered. In fact, it has been estimated that, in general, one third of all human CRMs are not conserved in rodents (Dermitzakis & Clark, 2002), and the number would presumably rise

exponentially when considering more distant species. This means that many regulatory elements would be overlooked if employing these approaches alone.

Several reasons may explain why certain CRMs are not conserved (Maston et al., 2006; Hiller et al., 2012). First, transcription factor binding sites often vary to various extents between species. This sequence degeneracy can lead to many false negatives if strict parameters are set in the comparison analyses. Second, often there exists a redundancy of regulatory elements with the same function at the same genomic spot, so the gain or loss of a single element at a given time in evolution would not have a profound effect on gene expression. Third, the gene under the influence of a given CRM may simply disappear. Therefore, the CRM will cease to be under positive selection. Fourth, the expression of a given gene may evolve in such a way that the function of a specific CRM may not be necessary anymore. Fifth, the complexity of higher eukaryotes lies, precisely, in their regulatory networks. Hence, throughout evolution, new CRMs appear in higher eukaryotes, and these will, by definition, not exist at a lower evolutionary scale. Finally, a number of reports have shown the existence of Repeat-Associated Binding Sites or RABS, transposable elements that harbor binding sites for certain transcription factors (Bourque et al., 2008; Kunarso et al., 2010; Schmidt et al., 2012). When these RABS mobilize to different genomic locations, new transcription factor binding sites emerge. Importantly, RABS are lineage-specific, which means that the repertoire of new CRMs created by their mobilization is unique to a given species. As a consequence, approaches based solely on genetic conservation will fail to unveil them. Of note, around 30% of all CTCF binding sites occupied in mouse embryonic stem cells reside in RABS, being the SINE B2 family of retrotransposons the major contributor to this phenomenon (Bourque et al., 2008; Kunarso et al., 2010). Moreover, more than 10% of all CTCF-bound sites in human embryonic stem cells are also contained within RABS, although no family of transposons is particularly enriched in this case (Kunarso et al., 2010). In any case, evolutionary conservation is a key feature that needs to be taken into account when searching for CRMs, including insulators, although it should be noted that lack of conservation does not necessarily imply lack of functional relevance.

On the other hand, the fastest way of bioinformatically predicting all genomic sites with which a protein interacts is scanning the genome for its consensus binding motif. However, this approach suffers from two major drawbacks. First, all binding sites may not be functional; that is, the protein under study may not necessarily bind all its target sites *in vivo* at all times. For example, many CTCF target sites are occupied in a cell-type specific fashion, which implies that only a fraction of all CTSs is actually functional in each cell type

(Chen et al., 2012). Second, it is obviously absolutely necessary to know the consensus motif in advance. This poses a major problem in the context of insulator-related sequences: since boundaries are functionally, and not structurally, defined, there is not a unique sequence motif that can be used to identify all of them genome-wide. Even if there were, insulators working through mechanisms yet unknown would be missed.

Hence, the decision about the sequences to test was made after taking into consideration both factors: evolutionary conservation and experimentally validated binding of specific proteins involved in insulation such as CTCF or Rad21 (**Table D 1**). In addition, the presence of repeats was considered because some of them can function as insulators (Lunyak et al., 2007; Roman et al., 2011). In an attempt to unveil new mechanisms of insulation, some of the sequences selected, even if evolutionarily conserved, had not been associated, at the time, with any particular feature. At most, some of them contained a DNaseI hypersensitive site, presumably indicative of the binding of a protein. Interestingly, the ENCODE project (Mouse ENCODE Consortium et al., 2012) revealed that some of these sites, many of which convey considerable insulator activity, do bind certain proteins, including CTCF (**Table D 1**). This fact highlights the predictive potential of the algorithms. Additionally, only non-coding regions were selected, although it has been shown that coding regions may have regulatory potential (i.e., Neznanov et al., 1997; Birnbaum et al., 2012).

All the gene pairs selected for functional validation contained, at least, one element able to block the influence of a distal enhancer on a promoter when cloned in between the two in ectopic constructs *in vitro* (Lunyak et al., 2007; **Fig. R 11** & **Table D 1**). This blockage presumably resulted from the active action of insulation mechanisms present in the elements tested, and not merely from an increase in the distance between enhancer and promoter (**Fig. R 14**). In addition, the elements were only evaluated in one orientation (the same as the mouse reference genome). It is possible that the behavior of some of the sequences changes in the opposite orientation, especially those that host binding sites for many different factors, as described elsewhere (i.e. Bell & Felsenfeld, 2000; reviewed in West et al., 2002).

Importantly, two elements, Dis-6.4 and CorDis-9.2, behaved comparably to the chicken β-globin insulator (**Fig. R 13**). These elements need to be further explored to gain more insight into their functioning, in order to better exploit their properties in the construction of vectors for transgenesis. While Dis-6.4 probably depends on CTCF, the role of this protein in the mechanism of action of CorDis-9.2 remains unclear: mutation of its binding site in this sequence only resulted in a 20% reduction of enhancer-blocking

Table D 1. Properties of the elements tested and their performance in the enhancer-blocking assays. The features annotated by the ENCODE project after selection of the sequences are highlighted in red.

Pair	Elements	Size (bp)	Annotated Features	Repeats	Enhancer-Blocking Activity	
					In vitro[1]	*In vivo*[2]
1: *Atp2a1-Sh2b1*	Cor-1	2009	DNaseI HS, CTCF, Rad21, **PolII, SMC3**	SINE (PB1D10), simple	+	***
2: *Psmc5-Smarcd2*	Cor-2.1	1074	**CTCF, PolII, Pol2S2**	simple	++	-
	Cor-2.2	828	DNaseI HS	SINE (B1_Mm, B2_Mm1t), simple	++	*
3: *Ftsj3-Psmc5*	Cor-3	1320	DNaseI HS, E2F1, nMyc, Max, NELFe, **CTCF, PolII, TBP…**	simple	+	***
4: *Mapk3-Gdpd3*	Dis-4.1	925	DNaseI HS, Zfx, **PolII, p300**	-	+	**N/A**
	Dis-4.2	2010	DNaseI HS, Esrrb	SINE (MIR, MIRb, B1F2, PB1D9)	++	**
	Dis-4.3	576	Klf4	-	+/-	**N/A**
	Dis-4.4	173	DNaseI HS, **PolII, TBP**	-	+/-	**N/A**
5: *Tatdn1-Ndufb9*	Dis-5.1	700	DNaseI HS, p300, **Mafk, COREST, p300, ETS1, GATA1**	SINE (B3A), simple	+/-	**N/A**
	Dis-5.2	1697	-	DNA (Charlie1b), simple	+	**N/A**
	Dis-5.3	1290	DNaseI HS, CTCF, Rad21, **COREST, SIN3A, SMC3, GATA1**	SINE (B1_Mus1, B3A, B4A), low complexity	+++	**
	Dis-5.4	124	DNaseI HS, E2F1, Max, NELFe, Esrrb, Zfx, PolII, **TBP, p300**	-	+/-	**N/A**
	Dis-5.5	2204	**DNaseI HS**	SINE (B4A), low complexity	+++	***
6: *Shisa5-Trex1*	Dis-6.1	800	DNaseI HS, CTCF, Rad21, **Max, cJun, SMC3, SIN3A…**	SINE (B1_Mur2, B2_Mm1t), LINE (L1_Mus3), low complexity, simple	+++	***
	Dis-6.2	1104	DNaseI HS, NELFe, PolII, **BHLHE40, ETS1, GCN5, cJun…**	simple	+	**N/A**
	Dis-6.3	700	DNaseI HS, **PolII, GATA1, ETS1**	-	+	**N/A**
	Dis-6.4	1281	DNaseI HS, CTCF, Rad21, Zfx, NELFe, Esrrb, PolII, **COREST, SIN3A, Max, Mxi1…**	-	++++	***
7: *Memo1-Dpy30*	CorDis-7.1	572	-	-	++	***
	CorDis-7.2	1654	-	-	+	***
	CorDis-7.3	1281	-	-	+	**N/A**
8: *Tsen34-Rps9*	CorDis-8.1	1185	-	SINE (MIRb, RSINE1, PB1D10, B3)	+	-
	CorDis-8.2	1297	**CTCF, Rad21, PolII**	SINE (B1_Mus1, B1_Mus2, B1_Mur1, B1_Mur2, B2_Mm2, B3), simple	++	**
	CorDis-8.3	1387	DNaseI HS, E2F1, Zfx, Max, NELFe, cMyb, PolII, **CTCF**	SINE (B1_Mur2, B1F2, B2_Mm2), LTR (MTD)	+	**
9: *Ddost-Pink1*	CorDis-9.1	115	DNaseI HS, CTCF, **Rad21, SMC3, PolII**	-	+	**N/A**
	CorDis-9.2	1021	DNaseI HS, CTCF, Rad21, **SMC3**	SINE (MIRm), LTR (MLT1M)	++++	***
	CorDis-9.3	915	DNaseI HS, CTCF, Rad21, **NELFe, SMC3**	-	++	-
10: *Rbm25-Psen1*	10.1	1025	**Pol2S2**	-	+	**N/A**
	10.2	1360	**Pol2S2**	-	+	**N/A**
	10.3	768	CTCF, **PolII, Pol2S2**	SINE (B2_Mm1t)	+++	***

[1]***In vitro* enhancer-blocking activity** relative to cHS4. **++++**: 100-108%; **+++**: 75-99%; **++**: 50-74%; **+**: 25-49%; **+/-**: 14-24%.

[2]***In vivo* enhancer-blocking activity.** ***: significant at p-value < 0.001; **: significant at p-value < 0.01; *: significant at p-value < 0.05; -: not significant; N/A: not assayed.

activity (**Fig. R 15D**). This could be explained by the fact that the mutations introduced failed to completely abrogate CTCF binding (**Fig. R 16**). This is not surprising, considering that many reports have highlighted the difficulty in mutating CTCF binding sites (i.e., Burcin et al., 1997; Renaud et al., 2005; Yoon et al., 2005; Fitzpatrick et al., 2007; Shukla et al., 2011). Two reasons account for this difficulty. First, CTCF is able to recognize a variety of DNA sequences through the combinatorial use of its 11 Zn fingers. In fact, the exact binding motif remains unknown (see section 1.3.1.1.). Second, 37% of the sequences known to bind CTCF contain more than one motif (Kim et al., 2007), so the identification and mutation of all of them is necessary for complete loss of protein binding. Alternatively, CorDis-9.2 also encloses a mammalian specific MLT1M retrotransposable element, which may be playing a role in insulation at this locus, a possibility yet to be explored.

Several of the elements, including Dis-6.4, were trimmed down in an attempt to define their core insulator region and discover the mechanism of insulation (**Fig. R 15**). In addition, small but powerful insulator elements are currently demanded to improve gene transfer vectors, especially those viral vectors with size limitations used in gene therapy applications (Emery, 2011). Of note, a fragment less than half the size of the original Cor-2.1 element, Cor-2.1D, was able to almost fully recapitulate the activity of the parent fragment. Yet, it only hosts a weak binding site for CTCF in the mouse thymus (Mouse ENCODE Consortium et al., 2012). This binding site has only been found, and weakly, in one tissue, which may indicate that it is a false positive. Alternatively, CTCF may actually bind there in a tissue-specific manner, probably exerting regulatory functions. In any case, the enhancer-blocking assay was carried out in human embryonic kidney cells, presumably neuronal in origin (Shaw et al., 2002) and thus, non-homologous to the mouse thymus. Therefore, it is highly improbable that CTCF may be responsible for the *in vitro* enhancer-blocking behavior of Cor-2.1D. Additional experiments are required to confirm the absence of CTCF binding and to elucidate if Cor-2.1D harbors an unknown mechanism of insulation.

The *in vitro* enhancer-blocking assay employed here is a fast, highly reproducible and reliable (based on good positive and negative controls, **Fig. R 11** & **12**) method to functionally validate insulators. It has been successfully used in the past for the evaluation of several boundaries, including the SINE B2 element of the murine growth hormone locus (i.e., Lunyak et al., 2007) and the element that separates the ubiquitously expressed human *RUVBL2* gene and the hypoxia-inducible gene *GYS1* (Tiana et al., 2012). However, it poses some drawbacks. First, human cells are used to test mouse DNA sequences. In this heterologous system, it is assumed that all the regulatory elements needed for the putative

mouse insulator to function are present in human, although this may not be true in all cases. Second, some insulators are only active at a given tissue or developmental stage (Lunyak et al., 2007), so HEK 293 cells may not be always appropriate. Third, we performed transient transfections. Some constructs probably integrated into the genome but the majority of them surely remained episomal. Thus, the effects the chromatin environment may exert over the sequences under testing were not really considered. Finally, the elements were evaluated in ectopic constructs *in vitro*, which may not be a reflection of the situation at the endogenous loci (West et al., 2002; Phillips & Corces, 2009; Splinter & De Laat, 2011). For example, deletion (Bender et al., 2006) or mutation (Splinter et al., 2006) of the DNaseI hypersensitive sites that flank the mouse β-globin genes and that act as enhancer-blockers *in vitro* (Farrell et al., 2002) do not provoke changes in gene expression or global heterochromatinization of the locus. Hence, the regulatory function of a particular DNA sequence should ideally be assessed in its native context (Barkess & West, 2012). Recently developed tools like Zinc-Finger Nucleases (ZFNs), Transcription Activator-Like Effector Nucleases (TALENs) or Clustered Regulatory Interspaced Short Palindromic Repeat (CRISPR)/Cas-based RNA-guided DNA endonucleases may facilitate this task (reviewed in Gaj et al., 2013).

Some of the most prominent enhancer-blockers *in vitro*, at least one per gene pair, were interrogated again in an *in vivo* setting in transgenic zebrafish (Bessa et al., 2009). Again, this is a heterologous system. If a given element displays enhancer-blocking activity in zebrafish, having displayed it *in vitro* in human cells as well, it is reasonable to assume that it will also do so in mice. Conversely, elements scoring negatively in this test do not necessarily lack enhancer-blocking activity. It is possible that they contain a mechanism of insulation that appeared later in evolution and that is not present in zebrafish. In any case, care should be taken when extrapolating results from one species to another. For example, the activity of 83% of human enhancers differs in transgenic mice and zebrafish, mainly due to differences in the expression, activity or specificity of the transcription factors involved (Ariza-Cosano et al., 2012). It is likely that similar results would be obtained if repeating the experiment with insulator sequences: the elements needed for the establishment of enhancer-blocking function may either not be conserved between mammals and zebrafish, or have diverged throughout evolution.

Microinjection of any construct into one-cell stage embryos always generates mosaic animals because it is highly improbable that the transgene integrates into the genome at this stage. Instead, it usually integrates into a limited number of cells when the embryo reaches the multicellular stage. This circumstance, in conjunction with the fact that

integrated transgenes are subject to chromosomal position effects, accounts for the high variability observed in GFP expression levels and distribution among individuals transgenic for the same construct, which is another important caveat of this method. However, in our case, many transgenic individuals were produced for each of the elements under testing in order to be sure of the phenotype and to be able to carry out accurate and informative statistical tests. The use of the first generation offspring of these transgenic founders would help to solve the variability problem, but this would ruin one of the main advantages of the system: its rapidity.

CorDis-9.2 acted as a potent enhancer-blocker both *in vitro* and *in vivo*. For this reason, it was further tested in transgenic mice in an assay aimed at deciphering if it was able to protect from chromosomal position effects as well (Furlan-Magaril et al., 2011); that is, if it also possessed the second property of insulators: barrier activity. Even if CorDis-9.2 increased the probability of transgene expression, it was unable to completely abrogate chromosomal position effects (**Fig. R 24**). This is not surprising, given that most of the insulators described so far only have one of the two main properties, enhancer-blocking or barrier activity. The chicken β-globin cHS4 boundary constitutes a remarkable exception (Chung et al., 1993; reviewed in Giraldo et al., 2003b). Furthermore, some insulators appear to be context-specific and do not perform equally well in all experimental settings (Molto et al., 2009).

5.1.2. Flanking Clusters of Co-Expressed Genes

Co-regulation of a particular subset of genes that participate in the same biological pathway constitutes a means by which cells ensure that all components will be expressed at the same time, thus allowing the correct functioning of the pathway. The arrangement of the genes to be co-regulated next to each other in the linear genome is one mechanism of co-regulation (reviewed in Hurst et al., 2004). The simplest case consists of bidirectional promoters, which are able to co-regulate the expression of the genes that lie at either side (Trinklein et al., 2004). In addition, clusters of adjacent co-expressed genes (also known as positional clusters) exist in several species (Cohen et al., 2000; Hurst et al., 2004; Woo et al., 2010; Szczepinska & Pawłowski, 2013). It has been proposed that these clusters are contained within the same domain, a domain that would need to be flanked by boundaries to preserve its integrity (Bell et al., 2001; Hurst et al., 2004).

For this reason, we sought to detect clusters of regionally co-expressed genes in order to uncover the boundaries that may be bracketing them. By analyzing one of the

expression distance matrices calculated in the Euclidean distance method developed here as a case example, several such clusters were identified, including the already described α cluster of protocadherin (*Pcdh*) genes (Monahan et al., 2012; Guo et al., 2012). Functional analyses revealed that one of the clusters was highly enriched in genes devoted to the formation of desmosomal junctions. CTCF sites, together with sites for the cohesin subunit Rad21, were found demarcating the cluster (**Fig. R 27**), as observed in other clusters (Xie et al., 2007). In addition, they were depleted within the locus, as expected (Kim et al., 2007).

3C experiments were carried out to investigate whether these CTCF-Rad21 sites interacted with each other creating a looped domain in which co-regulation mediated by transcription factors implied in epithelial-related processes would be facilitated. The proximity in space of these genes would explain why only the promoters of some of them host binding sites for these transcription factors (**Table R 11**), as shown before in other clusters (Cohen et al., 2000). Indeed, we observed a higher than expected frequency of contacts between the CTCF-cohesin sites that flank the cluster, indicating that they possibly interact. This finding was obtained in keratinocyte cells (COCA cell line) that express these genes, as well as in non-expressing cells (L929 fibroblasts). This result suggests that this 1-Mb region constitutes a structural domain or TAD and, as such, it would be mostly conserved across tissues (Dixon et al., 2012). In fact, it is defined as a TAD in other non-expressing cells such as ES cells or those derived from the brain cortex (**Fig. D 3**). Therefore, the desmosomal genes would be enclosed in a structural looped domain regardless of the cell type, which means that transcriptional activity is probably not required for its formation.

The locus also contains other CTCF binding sites not associated with additional proteins (**Fig. R 27**). Unlike structural constitutive sites, these may play a role in the regulation of gene expression through the establishment of long-range interactions between promoters and regulatory elements and thus, would be tissue-specific (Phillips-Cremins & Corces, 2013a). However, this possibility remains unexplored.

Experiments that combine gene expression and long-range interaction data have evidenced that co-expressed genes also group in the nuclear space, even if they map far away from each other in the linear genome with several unrelated genes in between them (Dong et al., 2010; Woo et al., 2010). For example, it has been shown that a CTCF-dependent long-range interaction couples transcription of the human insulin gene in pancreatic β-cells with that of *SYT8*, a gene located over 300 kb away and that is involved in insulin secretion (Xu et al., 2011). The authors hypothesized that this interaction, which

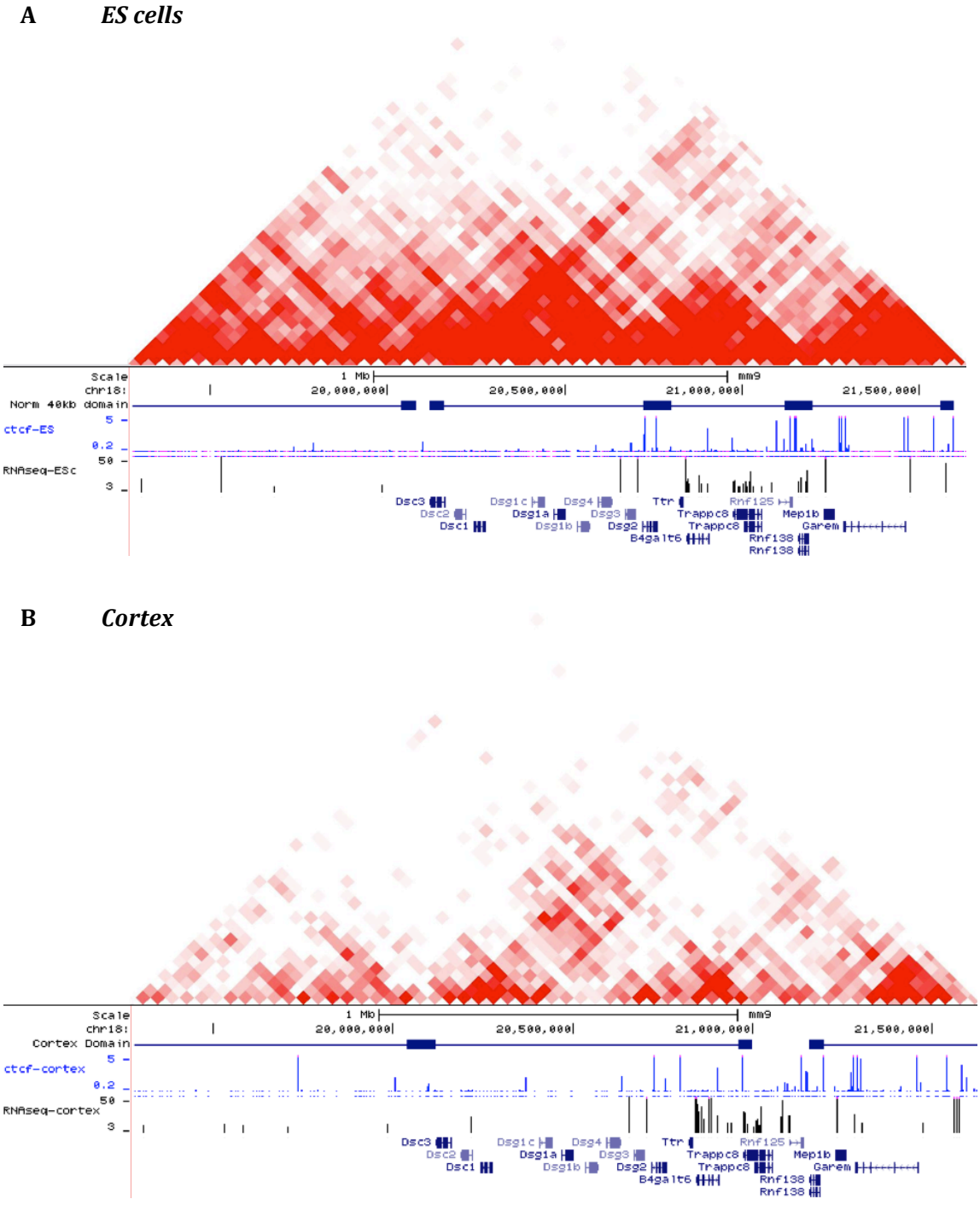


Fig. D 3. The cluster of genes involved in the formation of desmosomes are contained within a topologically associating domain in ES cells (A) and in cells derived from the brain cortex (B). Each panel shows the heat map of contact frequencies between any two loci assayed (top), as well as the corresponding genomic region extracted from the UCSC genome browser (bottom) with the following information: the resulting topologically associating domains, CTCF binding sites in the specified cell types, RNA-seq data, and the genes embedded in the domain. Figure from the online resource that enables the visualization of long-range interactions detected by Hi-C technology developed at Bing Ren's laboratory (http://yuelab.org/hi-c/index.html; Dixon et al., 2012).

was increased by glucose, promoted *SYT8* transcription by either recruiting the active histone marks present at the insulin promoter to the *SYT8* promoter, or by allowing the transcription factors that bind the insulin promoter to regulate the *SYT8* as well.

Alternatively, spatial clusters of co-expressed genes may compose specialized transcription factories in which co-regulation is thought to be facilitated by the accumulation of a set of transcription factors destined to trigger specific biological pathways (Schoenfelder et al., 2010; Li et al., 2012). This phenomenon has been baptized as *chroperon* or "chromatin-based operon mechanism for spatiotemporal regulation of gene transcription in eukaryotes" (Li et al., 2012).

So as to explore the role of the insulator protein CTCF in the establishment of this type of long-range interactions, the search of co-expressed genes was extended to find groups of genes with similar expression profiles, but not necessarily adjacent in the genomic sequence. As expected, one of the resulting clusters contained the desmosomal genes identified in the previous screen. Another gene from the formation pathway of desmosomes, *Spink5*, was included in this cluster as well. Of note, *Spink5* maps 23 Mb downstream of the positional cluster in the same chromosome. Unfortunately, 3C experiments failed to reveal any interaction between *Spink5* and the positional cluster. Nevertheless, only two of the various CTCF sites present at the *Spink5* locus were interrogated; the interaction, if existing, may be established with a different unexplored site.

The analysis of the frequency of contacts between distant loci has been extensively used to reconstruct the three-dimensional organization of chromosomes (i.e., Bau et al., 2011; Ben-Elazar et al., 2013). However, this study shows that it may also be possible to use gene expression data to model genomic structure or, at least, to identify expression domains and the boundaries that delimit them.

5.1.3. Establishing a Barrier between Euchromatic and Heterochromatic Domains

Eukaryotic genomes are divided into open active euchromatic and condensed silenced heterochromatic domains. Each type of domain is enriched in specific histone chromatin marks and variants that can be used to map them in ChIP-based experiments (Barski, et al., 2007). In some cases, the transition points between distinct domains are fuzzy, and characterized by a *gradient of histone modification*. These "negotiable" borders

result from a balance between two conflicting chromatin remodeling activities (i.e., histone acetylation *versus* deacetylation) that operate at either side. In contrast, other borders are sharp, suggesting the existence of particular barrier elements located at those borders that establish *walls* to actively partition the chromatin into different states (Kimura & Horikoshi, 2004). Hence, in order to discover barrier insulators genome-wide in an unbiased manner, it would be necessary to first map the different chromatin domains, and then analyze these "fixed" borders to find the mechanisms responsible for insulation. Jordan and collaborators followed this strategy in human $CD4^+$ T cells using available ChIP-seq data for several histone chromatin marks and variants (Wang et al., 2012b). They observed that MIRs (Mammalian-wide Interspersed Repeats), retrotransposable elements from the SINE family, were highly enriched at the transition points between heterochromatic and euchromatic regions, and re-conducted the whole chromatin analysis focusing on these elements (unpublished results). First, they identified all human MIRs with intact B-boxes -a feature associated with insulator activity in several species as already stated. Then, they filtered the list to obtain only those elements that bind the RNA polymerase III machinery, and that partition the chromatin into transcriptionally active *versus* repressed domains (according to histone modification profiles and RNA-seq data available for this particular cell type). The final subset of MIR elements was hypothesized to harbor good candidates to convey barrier activity at the chromatin borders of human $CD4^+$ T cells. The fact that only MIRs were further studied does not exclude the possibility that additional factors located at the borders (such as CTCF binding sites) contribute to barrier formation.

Indeed, some of the elements tested conveyed enhancer-blocking activity *in vitro* in HEK 293 cells (**Fig. R 32**) and *in vivo* in transgenic zebrafish (**Fig. R 33**). More experiments are needed to elucidate the precise mechanism of insulation. The ability of these elements to protect from chromosomal position effects in transgenic animals, that is, their ability to truly function as barriers, remains unexplored.

There are two types of heterochromatic regions: constitutive and facultative. Constitutive heterochromatin is composed of repetitive elements (telomeres and centromeres), and plays a role in genomic stability. On the other hand, facultative heterochromatin contains genes that are silent in a given cell type and/or developmental stage, but that become active elsewhere (Oberdoerffer & Sinclair, 2007). For example, genes involved in muscle function will pertain to the euchromatic portion of the genome in myocytes, whereas they will probably be embedded in (facultative) heterochromatin in lymphocytes. Thus, this type of heterochromatin responds to cellular transcriptional

needs. Since euchromatic and heterochromatic domains are cell-type and -stage specific, so should the barriers that separate them be (Cuddapah et al., 2009; Chen et al., 2012). Hence, the aforementioned strategy to find barrier insulators using histone modification profiles will only disclose those elements operating in the tissue and developmental time point chosen for study (in $CD4^+$ cells in this case). Care must be taken when extrapolating the results to other conditions.

The fact that the positions of the barriers change throughout the life of a cell as differentiation proceeds, makes barriers *dynamic in nature*. This is exemplified by the SINE B2 element at the murine *Gh* locus, which only becomes active as a barrier at embryonic stage 17.5 in the pituitary gland cells. As a result, *Gh* is protected from the advance of silencing epigenetic marks that emanate from a heterochromatinization focus upstream of the SINE B2 retrotransposon, and starts being expressed (Lunyak et al., 2007).

Unlike the actual position of the barriers, the mechanisms employed to establish them do not depend on the cell type and developmental stage. This means that MIR elements, as a mechanism of insulation, are not necessarily specific to $CD4^+$ cells, and may function in other cell types as well. What is specific is the particular subset of MIRs disclosed by the algorithm. Therefore, the algorithm was able to uncover new mechanisms of insulation in an unbiased manner (Wang et al., submitted).

5.2. Relevance of the Study

The concept of genomic boundary groups diverse types of elements of different nature. What unites them is their function as delimiters of expression domains. The novelty of the approach followed here lies in the fact that it was boundary function, and not boundary-associated sequences, that was used to predict the presence of boundaries in a genome-wide fashion in mammals. Unlike previous screens, this approach allows the discovery of new mechanisms of insulation. Indeed, we have been able to describe a new type of insulator element based on MIR repeats in human. It is possible that additional mechanisms emerge after deeper analyses of other insulator sequences identified here, particularly of those not associated with any particular feature.

Traditionally, the protein-coding genome has received more attention than its non-coding counterpart, presumably because it is easier to work with: genes are transcribed into mRNAs, which are further translated into proteins, and both these products are

tangible and measurable, directly. Yet, the characterization of regulatory elements is fundamental for the understanding of why, how, where and when genes are expressed.

Most of the scientific research aimed at finding the cause of human or animal diseases focuses on finding defects in protein-coding genes. However, it is being increasingly recognized that the malfunction of regulatory elements can also derive in a pathogenic condition (Maston et al., 2006; Spielmann & Mundlos, 2013). For example, the deletion of a barrier element (or specialized silencer) between human *H2AFY* and *PITX1* enables the influence of a fore- and hind limb enhancer on the promoter of *PITX1*, only expressed in the hind limb in normal conditions (Spielmann et al., 2012; Spielmann & Mundlos, 2013). This causes the so-called Liebenberg syndrome, characterized by a partial transformation of the upper extremities to lower extremities (Liebenberg, 1973). This study has shown that many boundaries reside in the loci of several genes with vital cellular functions, including some that cause disease when mutated (i.e., *Trex1, Psen1* or *Pink1*). It is possible that mutations of some of these boundaries contribute to the development of disease. More in depth analyses of these elements may provide valuable information in that regard.

Finally, from a practical point of view, the insulators described here could be incorporated in constructs to alleviate the chromosomal position effects gene transfer technologies suffer (Giraldo et al., 2003b; Recillas-Targa et al., 2004; Molto et al., 2009; Emery, 2011).

6 CONCLUSIONS

1. Two algorithms were developed to predict the existence of boundaries separating genes with opposite expression profiles in a genome-wide fashion in mice, using the gene expression data stored in the online platform aGEM. The algorithms differed in the parameters employed in each case: Pearson's correlation or Euclidean distance.

2. The algorithms predicted the presence of boundaries in 6,816 pairs of adjacent genes in the mouse genome. These pairs were enriched in previously identified insulator-related sequences.

3. The loci of ten pairs, nine of them obtained by either or both of the algorithms, were analyzed. Within them, several non-coding evolutionarily conserved sequences, some of them associated with functional features such as CTCF binding, were chosen for functional validation *in vitro* in human embryonic kidney 293 cells, and *in vivo* in transgenic zebrafish. All of the gene pairs selected contained, at least, one element with enhancer-blocking activity in these assays.

4. The locus of the gene pair formed by *Ddost* and the Parkinson's-associated gene *Pink1* harbors a potent *in vitro* and *in vivo* enhancer-blocking element, CorDis-9.2. Its activity partially depends on the binding of CTCF. However, this element, while increasing the probability of expression of the associated reporter construct, did not fully protect from chromosomal position effects in transgenic mice.

5. The Euclidean distance algorithm served as a base to develop an additional algorithm that predicts the presence of genomic boundaries flanking positional clusters of co-expressed genes in the mouse chromosome 18. One such cluster, composed of the genes involved in the formation of desmosomes, was identified. In addition, the algorithm suggested the existence of long-range interactions between this cluster and *Spink5*, also associated with the same pathway but mapping 23 Mb downstream of the positional cluster.

6. 3C experiments carried out in cells that express the desmosomal genes (epidermal keratinocyte derived COCA cells), as well as in non-expressing cells (L929 fibroblasts), demonstrated long-range interactions between the CTCF-dependent elements that bracket this positional cluster. Long-range interactions between this locus and *Spink5* were not detected.

7. Mammalian-wide Interspersed Repeats or MIRs found partitioning the chromatin into open active euchromatin and condensed silenced heterochromatin in human $CD4^+$ cells, possessed *in vitro* enhancer-blocking activity in human embryonic kidney 293 cells. Some of them also functioned as enhancer-blockers *in vivo* in transgenic zebrafish. Thus, this study succeeded in describing a new mechanism of insulation based on human MIR elements.

Main conclusion

We have successfully predicted and functionally validated new genomic boundaries in mammalian genomes focusing, not on structural features such as CTCF-binding, but on finding those loci where the presence of boundary function is necessary for the organization of the genome and the correct functioning of the cells. Furthermore, we have been able to analyze the two properties that characterize insulators in general, namely enhancer-blocking and barrier activities, as well as a third property associated with CTCF-dependent insulators in particular, that is, the establishment of long-range interactions. Not only do the insulators identified here provide information about the regulation and organization of specific loci, but they also could be used as genetic tools to improve gene transfer technologies.

7 CONCLUSIONES

1. Se desarrollaron dos algoritmos para predecir la presencia de aisladores genómicos separando genes con perfiles de expresión opuestos en el genoma de ratón. Para ello, se analizaron los datos de expresión génica almacenados en la plataforma en red aGEM. Los algoritmos difieren en los parámetros usados en cada caso: correlación de Pearson o distancia Euclídea.

2. Los algoritmos predijeron la presencia de aisladores en 6.816 parejas de genes adyacentes en el genoma de ratón. Estas parejas estaban enriquecidas en secuencias previamente asociadas a aisladores.

3. Se examinaron los loci de diez parejas de genes, nueve de las cuales fueron obtenidas por uno o ambos algoritmos. Dentro de cada locus, se seleccionaron varias regiones no codificantes pero conservadas evolutivamente. Algunas de ellas se encontraban asociadas a determinados rasgos funcionales, como la presencia de sitios de unión para la proteína CTCF. Estas secuencias se evaluaron funcionalmente mediante ensayos *in vitro* en células embrionarias humanas de riñón (HEK 293), e *in vivo* en peces cebra transgénicos. Estos ensayos revelaron que todas las parejas seleccionadas contenían al menos un elemento con actividad bloqueante sobre elementos potenciadores (*enhancers,* en inglés) de la expresión génica.

4. La región genómica que engloba a la pareja de genes *Ddost* y *Pink1*, este último asociado a la enfermedad de Parkinson, contiene un elemento con gran capacidad de bloqueo de potenciadores de la expresión génica, tanto *in vitro* como *in vivo*. La actividad de este elemento, denominado CorDis-9.2, depende parcialmente de la unión de CTCF. En ratones transgénicos, este elemento provoca un aumento de la probabilidad de expresión de la construcción indicadora asociada, si bien no es capaz de proteger completamente de los efectos de posición cromosomales.

5. El algoritmo basado en distancias Euclídeas sirvió de punto de partida para el desarrollo de un algoritmo adicional para predecir la presencia de aisladores flanqueando grupos de genes adyacentes con el mismo perfil de expresión, en el cromosoma 18 de ratón. Entre otros, se identificó un grupo de genes implicados en la formación de desmosomas. Además, el algoritmo apuntó la posible existencia de interacciones de largo alcance entre este grupo de genes y *Spink5*, gen asociado a la misma ruta biológica pero localizado 23 Mb aguas abajo.

6. Los experimentos de 3C llevados a cabo tanto en células murinas que expresan los genes implicados en la formación de desmosomas (células derivadas de queratinocitos epidérmicos COCA), como en células que no los expresan (fibroblastos L929), mostraron la existencia de interacciones de largo alcance entre las regiones de unión de la proteína CTCF que flanquean este grupo de genes. No se detectaron este tipo de interacciones entre dicho locus y *Spink5*.

7. Se identificaron diversas repeticiones dispersas ampliamente distribuidas en mamíferos (o MIRs) localizadas separando regiones de cromatina abierta y activa o eucromatina, y regiones de cromatina compacta y silente o heterocromatina, en células $CD4^+$ humanas. Estos elementos mostraron actividad de bloqueo de potenciadores de la expresión génica en ensayos *in vitro* en células HEK 293. Algunos de ellos también mostraban dicha actividad *in vivo* en peces cebra transgénicos. Por tanto, con este estudio se ha podido identificar un nuevo mecanismo de aislamiento basado en MIRs en humanos.

Conclusión principal

Con esta tesis doctoral se han podido predecir y validar funcionalmente nuevos aisladores genómicos en mamíferos, mediante una estrategia basada en la búsqueda de aquellas regiones donde la existencia de función aisladora es necesaria para la organización del genoma y el correcto funcionamiento celular. Así, este estudio difiere de otros anteriores enfocados en la persecución de la presencia de determinados rasgos estructurales, como la unión de la proteína CTCF a una determinada región. Además, se han analizado con éxito las dos propiedades que caracterizan a los aisladores en general, esto es, las actividades de barrera y de bloqueo de potenciadores de la actividad génica, así como una tercera propiedad asociada especialmente a aisladores dependientes de CTCF: el establecimiento de interacciones de largo alcance. Los aisladores genómicos identificados en este estudio no sólo proporcionan información acerca de la regulación y la organización de loci específicos, sino que también podrían usarse como herramientas genéticas para mejorar las tecnologías de transferencia de genes.

8 REFERENCES

Abhyankar MM, Urekar C, Reddi PP (2007) A novel CpG-free vertebrate insulator silences the testis-specific SP-10 gene in somatic tissues: role for TDP-43 in insulator function. *J Biol Chem* 282(50):36143–54.

Acharya KK, Govind CK, Shore AN, Stoler MH, Reddi PP (2006) Cis-requirement for the maintenance of round spermatid-specific transcription. *Dev Biol* 295(2):781–90.

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2002) In: *Molecular Biology of the Cell*, 4th Edition, Garland Science Publishing, New York.

Amouyal M (2010a) Gene insulation. Part I: natural strategies in yeast and Drosophila. *Biochem Cell Biol* 88(6):875–84.

Amouyal M (2010b) Gene insulation. Part II: natural strategies in vertebrates. *Biochem Cell Biol* 88(6):885-98.

Aoki K, Kakizaki F, Sakashita H, Manabe T, Aoki M, Taketo MM (2011) Suppression of colonic polyposis by homeoprotein CDX2 through its nontranscriptional function that stabilizes p27Kip1. *Cancer Res* 71(2):593–602.

Ariza-Cosano A, Visel A, Pennacchio LA, Fraser HB, Gomez-Skarmeta JL, Irimia M, Bessa J (2012) Differences in enhancer activity in mouse and zebrafish reporter assays are often associated with changes in gene expression. *BMC Genomics* 13:713.

Ausubel FA, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, Struhl K (eds.) (1999) In: *Current Protocols in Molecular Biology,* John Wiley & Sons, New York.

Awad TA, Bigler J, Ulmer JE, Hu YJ, Moore JM, Lutz M, Neiman PE, et al. (1999) Negative transcriptional regulation mediated by thyroid hormone response element 144 requires binding of the multivalent factor CTCF to a novel target DNA sequence. *J Biol Chem* 274(38):27092–8.

Bachmann-Gagescu R, Mefford HC, Cowan C, Glew GM, Hing AV, Wallace S, Bader PI, et al. (2010) Recurrent 200-kb deletions of 16p11.2 that include the SH2B1 gene are associated with developmental delay and obesity. *Genet Med* 12(10):641-7.

Barkess G, West AG (2012) Chromatin insulator elements: establishing barriers to set heterochromatin boundaries. *Epigenomics* 4(1):67–80.

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, et al. (2007) High-resolution profiling of histone methylations in the human genome. *Cell* 129(4):823–37.

Bau D, Sanyal A, Lajoie BR, Capriotti E, Byron M, Lawrence JB, Dekker J, Marti-Renom MA (2011) The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nat Struct Mol Biol* 18(1):107-14.

Beermann F, Ruppert S, Hummler E, Bosch FX, Muller G, Ruther U, Schutz G (1990) Rescue of the albino phenotype by introduction of a functional tyrosinase gene into mice. *EMBO J* 9(9):2819–26.

Beermann F, Ruppert S, Hummler E, Schutz G (1991) Tyrosinase as a marker for transgenic mice. *Nucleic Acids Res* 19(4):958.

Bekris LM, Yu CE, Bird TD, Tsuang DW (2010) Genetics of Alzheimer disease. *J Geriatr Psychiatry Neurol* 23(4):213-27.

Bell AC, Felsenfeld G (2000) Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* 405(6785):482–5.

Bell AC, West AG, Felsenfeld G (1999) The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* 98(3):387–96.

Bell AC, West AG, Felsenfeld G (2001) Insulators and boundaries: versatile regulatory elements in the eukaryotic genome. *Science* 291(5503):447–50.

Ben-Elazar S, Yakhini Z, Yanai I (2013) Spatial localization of co-regulated genes exceeds genomic gene clustering in the Saccharomyces cerevisiae genome. *Nucleic Acids Res* 41(4):2191-201.

Bender MA, Byron R, Ragoczy T, Telling A, Bulger M, Groudine M (2006) Flanking HS-62.5 and 3' HS1, and regions upstream of the LCR, are not required for beta-globin transcription. *Blood* 108(4):1395–401.

Berg T, Didon L, Nord M (2006) Ectopic expression of C/EBPalpha in the lung epithelium disrupts late lung development. *Am J Physiol-Lung C* 291(4):L683–93.

Bessa J, Tena JJ, De la Calle-Mustienes E, Fernandez-Miñan A, Naranjo S, Fernandez A, Montoliu L, et al. (2009) Zebrafish enhancer detection (ZED) vector: a new tool to facilitate transgenesis and the functional analysis of cis-regulatory regions in zebrafish. *Dev Dynam* 238(9):2409–17.

Birnbaum RY, Clowney EJ, Agamy O, Kim MJ, Zhao J, Yamanaka T, Pappalardo Z, et al. (2012) Coding exons function as tissue-specific enhancers of nearby genes. *Genome Res* 22(6):1059-68.

Bode J, Benham C, Knopp A, Mielke C (2000) Transcriptional augmentation: modulation of gene expression by scaffold/matrix-attached regions (S/MAR elements). *Crit Rev Eukar Gene* 10(1):73–90.

Boffelli D, Nobrega MA, Rubin EM (2004) Comparative genomics at the vertebrate extremes. *Nat Rev Genet* 5(6):456–65.

Bonifer C (2000) Developmental regulation of eukaryotic gene loci: which cis-regulatory information is required? *Trends Genet* 16(7):310–5.

Bourdon JC, Renzing J, Robertson PL, Fernandes KN, Lane DP (2002) Scotin, a novel p53-inducible proapoptotic protein located in the ER and the nuclear membrane. *J Cell Biol* 158:235-46.

Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew J, et al. (2008) Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res* 18(11):1752–62.

Brenden N, Madeyski-Bengtson K, Martinsson K, Svard R, Albery-Larsdotter S, Granath B, Lundgren H, et al. (2013) A triple-transgenic immunotolerant mouse model. *J Pharm Sci* 102(3):1116–24.

Bulger M, Groudine M (2011) Functional and mechanistic diversity of distal transcription enhancers. *Cell* 144(3):327–39.

Bulger M, Schubeler D, Bender MA, Hamilton J, Farrell CM, Hardison RC, Groudine M (2003) A complex chromatin landscape revealed by patterns of nuclease sensitivity and histone modification within the mouse beta-globin locus. *Mol Cell Biol* 23(15):5234–44.

Burcin M, Arnold R, Lutz M, Kaiser B, Runge D, Lottspeich F, Filippova GN, et al. (1997) Negative protein 1, which is required for function of the chicken lysozyme gene silencer in conjunction

with hormone receptors, is identical to the multivalent zinc finger repressor CTCF. *Mol Cell Biol* 17(3):1281–8.

Burke LJ, Hollemann T, Pieler T, Renkawitz R (2002) Molecular cloning and expression of the chromatin insulator protein CTCF in Xenopus laevis. *Mech Develop* 113(1):95–8.

Burke LJ, Zhang R, Bartkuhn M, Tiwari VK, Tavoosidana G, Kurukuti S, Weth C, et al. (2005) CTCF binding and higher order chromatin structure of the H19 locus are maintained in mitotic chromatin. *EMBO J* 24(18):3291–300.

Carabana J, Watanabe A, Hao B, Krangel MS (2011) A barrier-type insulator forms a boundary between active and inactive chromatin at the murine TCRβ locus. *J Immunol* 186(6):3556–62.

Cecchini KR, Raja Banerjee A, Kim TH (2009) Towards a genome-wide reconstruction of cis-regulatory networks in the human genome. *Semin Cell Dev Biol* 20(7):842–8.

Chao W, Huynh KD, Spencer RJ, Davidow LS, Lee JT (2002) CTCF, a candidate trans-acting factor for X-inactivation choice. *Science* 295(5553):345–7.

Chen H, Tian Y, Shu W, Bo X, Wang S (2012) Comprehensive identification and annotation of cell type-specific and ubiquitous CTCF-binding sites in the human genome. *PloS one* 7(7):e41374.

Chen Q, Lin L, Smith S, Huang J, Berger SL, Zhou J (2007) CTCF-dependent chromatin boundary element between the latency-associated transcript and ICP0 promoters in the herpes simplex virus type 1 genome. *J Virol* 81(10):5192–201.

Cheng Y, An LY, Yuan YG, Wang Y, Du FL, Yu BL, Zhang ZH, et al. (2012) Hybrid expression cassettes consisting of a milk protein promoter and a cytomegalovirus enhancer significantly increase mammary-specific expression of human lactoferrin in transgenic mice. *Mol Reprod Dev* 79(8):573–85.

Chernukhin IV, Shamsuddin S, Robinson AF, Carne AF, Paul A, El-Kady AI, Lobanenkov VV, et al. (2000) Physical and functional interaction between two pluripotent proteins, the Y-box DNA/RNA-binding factor, YB-1, and the multivalent zinc finger factor, CTCF. *J Biol Chem* 275(38):29915–21.

Cho DH, Thienes CP, Mahoney SE, Analau E, Filippova GN, Tapscott SJ (2005) Antisense transcription and heterochromatin at the DM1 CTG repeats are constrained by CTCF. *Mol Cell* 20(3):483–9.

Chua S (2010) SH2B1--the adaptor protein that could. *Endocrinology* 151(9):4100–2.

Chung JH, Bell AC, Felsenfeld G (1997) Characterization of the chicken beta-globin insulator. *P Natl Acad Sci USA* 94(2):575–80.

Chung JH, Whiteley M, Felsenfeld G (1993) A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila. *Cell* 74(3):505–14.

Cockerill PN, Garrard WT (1986) Chromosomal loop anchorage of the kappa immunoglobulin gene occurs next to the enhancer in a region containing topoisomerase II sites. *Cell* 44(2):273–82.

Cohen BA, Mitra RD, Hughes JD, Church GM (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet* 26(2):183-6.

Craig MP, Gilday SD, Hove JR (2006) Dose-dependent effects of chemical immobilization on the heart rate of embryonic zebrafish. *Lab Animal* 35(9):41–7.

Cuddapah S, Jothi R, Schones DE, Roh T, Cui K, Zhao K (2009) Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res* 19(1):24–32.

D'haeseleer P (2005) How does gene expression clustering work? *Nat Biotechnol* 23(12):1499–1501.

D'haeseleer P, Liang S, Somogyi R (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16(8):707–26.

De la Calle-Mustienes E, Feijoo CG, Manzanares M, Tena JJ, Rodriguez-Seguel E, Letizia A, Allende ML, Gomez-Skarmeta JL (2005) A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. *Genome Res* 15:1061-72.

De Laat W, Grosveld F (2003) Spatial organization of gene expression: the active chromatin hub. *Chromosome Res* 11(5):447–59.

De Strooper B (2003) Aph-1, Pen-2, and Nicastrin with Presenilin generate an active gamma-secretase complex. *Neuron* 38(1):9-12.

Defossez PA, Gilson E (2002) The vertebrate protein CTCF functions as an insulator in Saccharomyces cerevisiae. *Nucleic Acids Res* 30(23):5136–41.

Delgado-Olguin P, Brand-Arzamendi K, Scott IC, Jungblut B, Stainier DY, Bruneau BG, Recillas-Targa F (2011) CTCF promotes muscle differentiation by modulating the activity of myogenic regulatory factors. *J Biol Chem* 286(14):12483–94.

DeMare LE, Leng J, Cotney J, Reilly SK, Yin J, Sarro R, Noonan JP (2013) The genomic landscape of cohesin-associated chromatin interactions. *Genome Res* 23(8):1224–34.

Dermaut B, Kumar-Singh S, Engelborghs S, Theuns J, Rademakers R, Saerens J, Pickut BA, et al. (2004) A novel presenilin 1 mutation associated with Pick's disease but not beta-amyloid plaques. *Ann Neurol* 55(5):617-26.

Dermitzakis ET, Clark AG (2002) Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol* 19(7):1114–21.

Dickson J, Gowher H, Strogantsev R, Gaszner M, Hair A, Felsenfeld G, West AG (2010) VEZF1 elements mediate protection from DNA methylation. *PLoS Genet*, 6(1):e1000804.

Dillon N (2006) Gene regulation and large-scale chromatin organization in the nucleus. *Chromosome Res* 14(1):117–26.

Dillon N, Sabbattini P (2000) Functional gene expression domains: defining the functional unit of eukaryotic gene regulation. *BioEssays* 22(7):657–65.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, et al. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485(7398):376–80.

Donatien PD, Orlow SJ (1995) Interaction of melanosomal proteins with melanin. *Eur J Biochem* 232(1):159–64.

Dong X, Li C, Chen Y, Ding G, Li Y (2010) Human transcriptional interactome of chromatin contribute to gene co-expression. *BMC Genomics* 11:704.

Donze D, Adams CR, Rine J, Kamakaka RT (1999) The boundaries of the silenced HMR domain in Saccharomyces cerevisiae. *Genes Dev* 13(6):698–708.

Donze D, Kamakaka RT (2001) RNA polymerase III and RNA polymerase II promoter complexes are heterochromatin barriers in Saccharomyces cerevisiae. *EMBO J* 20(3):520–31.

Drissen R, Palstra RJ, Gillemans N, Splinter E, Grosveld F, Philipsen S, De Laat W (2004) The active spatial organization of the beta-globin locus requires the transcription factor EKLF. *Genes Dev* 18(20):2485–90.

Dunn KL, Zhao H, Davie JR (2003) The insulator binding protein CTCF associates with the nuclear matrix. *Exp Cell Res* 288(1):218–23.

Ebersole T, Kim JH, Samoshkin A, Kouprina N, Pavlicek A, White RJ, Larionov V (2011) tRNA genes protect a reporter gene from epigenetic silencing in mouse cells. *Cell Cycle* 10(16):2779–91.

Eissenberg JC, Elgin SC (1991) Boundary functions in the control of gene expression. *Trends Genet* 7(10):335–40.

El-Kady A, Klenova E (2005) Regulation of the transcription factor, CTCF, by phosphorylation with protein kinase CK2. *FEBS Lett* 579(6):1424–34.

Elizondo LI, Jafar-Nejad P, Clewing JM, Boerkoel CF (2009) Gene clusters, molecular evolution and disease: a speculation. *Curr Genomics* 10(1):64–75.

Emery DW (2011) The use of chromatin insulators to improve the expression and safety of integrating gene transfer vectors. *Hum Gene Ther* 22(6):761–74.

ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306(5696):636-40.

Engel N, Bartolomei MS (2003) Mechanisms of insulator function in gene regulation and genomic imprinting. *Int Rev Cytol* 232:89–127.

Farrar D, Rai S, Chernukhin I, Jagodic M, Ito Y, Yammine S, Ohlsson R, et al. (2010) Mutational analysis of the poly(ADP-ribosyl)ation sites of the transcription factor CTCF provides an insight into the mechanism of its regulation by poly(ADP-ribosyl)ation. *Mol Cell Biol* 30(5):1199–216.

Farrell CM, West AG, Felsenfeld G (2002) Conserved CTCF insulator elements flank the mouse and human beta-globin loci. *Mol Cell Biol* 22(11):3820–31.

Featherstone K, Wood AL, Bowen AJ, Corcoran AE (2010) The mouse immunoglobulin heavy chain V-D intergenic sequence contains insulators that may regulate ordered V(D)J recombination. *J Biol Chem* 285(13):9327–38.

Fedoriw AM, Stein P, Svoboda P, Schultz RM, Bartolomei MS (2004) Transgenic RNAi reveals essential function for CTCF in H19 gene imprinting. *Science*, 303(5655):238–40.

Feinberg AP, Vogelstein B (1983) A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. *Anal Biochem* 132(1):6–13.

Filippova GN (2008) Genetics and epigenetics of the multifunctional protein CTCF. *Curr Top Dev Biol* 80(07):337–60.

Filippova GN, Cheng MK, Moore JM, Truong JP, Hu YJ, Nguyen DK, Tsuchiya KD, et al. (2005) Boundaries between chromosomal domains of X inactivation and escape bind CTCF and lack CpG methylation during early development. *Dev Cell* 8(1):31–42.

Filippova GN, Fagerlie S, Klenova EM, Myers C, Dehner Y, Goodwin G, Neiman PE, et al. (1996) An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Mol Cell Biol* 16(6):2802–13.

Filippova GN, Lindblom A, Meincke LJ, Klenova EM, Neiman PE, Collins SJ, Doggett NA, et al. (1998) A widely expressed transcription factor with multiple DNA sequence specificity, CTCF, is localized at chromosome segment 16q22.1 within one of the smallest regions of overlap for common deletions in breast and prostate cancers. *Gene Chromosome Canc* 22(1):26–36.

Filippova GN, Qi C, Ulmer JE, Moore JM, Ward MD, Hu YJ, Loukinov DI, et al. (2002) Tumor-associated zinc finger mutations in the CTCF transcription factor selectively alter tts DNA-binding specificity. *Cancer Res* (1):48–52.

Filippova GN, Thienes CP, Penn BH, Cho DH, Hu YJ, Moore JM, Klesert TR, et al. (2001) CTCF-binding sites flank CTG/CAG repeats and form a methylation-sensitive insulator at the DM1 locus. *Nat Genet* 28(4):335–43.

Finlan LE, Sproul D, Thomson I, Boyle S, Kerr E, Perry P, Yistra B, et al. (2008). Recruitment to the nuclear periphery can alter expression of genes in human cells. *PLos Genet* 4(3):e10000039.

Fitzpatrick GV, Pugacheva EM, Shin JY, Abdullaev Z, Yang Y, Khatod K, Lobanenkov VV, et al. (2007) Allele-specific binding of CTCF to the multipartite imprinting control region KvDMR1. *Mol Cell Biol* 27(7):2636–47.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, et al. (2012) Ensembl 2012. *Nucleic Acids Res* 40(Database issue):D84–90.

Follows GA, Ferreira R, Janes ME, Spensberger D, Cambuli F, Chaney AF, Kinston SJ, et al. (2012) Mapping and functional characterisation of a CTCF-dependent insulator element at the 3' border of the murine Scl transcriptional domain. *PloS one* 7(3):e31484.

Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I (2004) VISTA: computational tools for comparative genomics. *Nucleic Acids Res* 32(Web Server issue):W273–9.

Fu Y, Sinha M, Peterson CL, Weng Z (2008) The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet* 4(7):e1000138.

Fullwood MJ, Ruan Y (2009) ChIP-based methods for the identification of long-range chromatin interactions. *J Cell Biochem* 107(1):30–9.

Furlan-Magaril M, Rebollar E, Guerrero G, Fernandez A, Molto E, Gonzalez-Buendia E, Cantero M, et al. (2011) An insulator embedded in the chicken α-globin locus regulates chromatin domain configuration and differential gene expression. *Nucleic Acids Res* 39(1):89–103.

Gaj T, Gersbach CA, Barbas CF 3rd (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol* 31(7):397-405.

Garrett FE, Emelyanov AV, Sepulveda MA, Flanagan P, Volpi S, Li F, Loukinov D, et al. (2005) Chromatin architecture near a potential 3' end of the igh locus involves modular regulation of histone modifications during B-Cell development and in vivo occupancy at CTCF sites. *Mol Cell Biol* 25(4):1511–25.

Gartner LP, Hiatt JL (2001) In: *Color Textbook of Histology*, WB Saunders Company Philadelphia.

Gartsbein M, Alt A, Hashimoto K, Nakajima K, Kuroki T, Tennenbaum T (2006) The role of protein kinase C delta activation and STAT3 Ser727 phosphorylation in insulin-induced keratinocyte proliferation. *J Cell Sci* 119(Pt 3):470–81.

Gaszner M, Felsenfeld G (2006) Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet* 7(9):703–13.

Geyer PK, Corces VG (1987) Separate regulatory elements are responsible for the complex pattern of tissue-specific and developmental transcription of the yellow locus in Drosophila melanogaster. *Genes Dev* 1(9):996–1004.

Geyer PK, Corces VG (1992) DNA position-specific repression of transcription by a Drosophila zinc finger protein. *Genes Dev* 6(10):1865–73.

Ghaleb AM, McConnell BB, Kaestner KH, Yang VW (2011) Altered intestinal epithelial homeostasis in mice with intestine-specific deletion of the Kruppel-like factor 4 gene. *Dev Biol* 349(2):310–20.

Gibcus JH, Dekker J (2013) The hierarchy of the 3D genome. *Mol Cell* 49(5):773–82.

Gimenez E, Lavado A, Giraldo P, Montoliu L (2003) Tyrosinase gene expression is not detected in mouse brain outside the retinal pigment epithelium cells. *Eur J Neurosci* 18(9):2673-6.

Gimenez E, Lavado A, Jeffery G, Montoliu L (2005) Regional abnormalities in retinal development are associated with local ocular hypopigmentation. *J Comp Neurol* 485(4):338-47.

Giraldo P (2002) Doctoral Thesis: *Análisis funcional y estructural de la región controladora de locus del gen de la tirosinasa de ratón*. Universidad Autónoma de Madrid, Madrid.

Giraldo P, Martinez A, Regales L, Lavado A, Garcia-Diaz A, Alonso A, Busturia A, et al. (2003a) Functional dissection of the mouse tyrosinase locus control region identifies a new putative boundary activity. *Nucleic Acids Res* 31(21):6290–305.

Giraldo P, Montoliu L (2001) Size matters: use of YACs, BACs and PACs in transgenic animals. *Transgenic Res* 10(2):83–103.

Giraldo P, Rival-Gervier S, Houdebine LM, Montoliu L (2003b) The potencial benefits of insulators on heterologous constructs in transgenic animals. *Transgenic Res* 12(6):751-5.

Girod PA, Zahn-Zabal M, Mermod N (2005) Use of the chicken lysozyme 5' matrix attachment region to generate high producer CHO cell lines. *Biotechnol Bioeng* 91(1):1–11.

Glasl L, Kloos K, Giesert F, Roethig A, Di Benedetto B, Kühn R, Zhang J, et al. (2012) Pink1-deficiency in mice impairs gait, olfaction and serotonergic innervation of the olfactory bulb. *Exp Neurol* 235(1):214–27.

Goetze S, Baer A, Winkelmann S, Nehlsen K, Seibler J, Maass K, Bode J (2005) Performance of genomic bordering elements at predefined genomic loci. *Mol Cell Biol* 25(6):2260–72.

Goljanek-Whysall K, Pais H, Rathjen T, Sweetman D, Dalmay T, Munsterberg A (2012) Regulation of multiple target genes by miR-1 and miR-206 is pivotal for C2C12 myoblast differentiation. *J Cell Sci* 125(Pt 15):3590-600.

Gombert WM, Farris SD, Rubio ED, Morey-Rosler KM, Schubach WH, Krumm A (2003) The c-myc insulator element and matrix attachment regions define the c-myc chromosomal domain. *Mol Cell Biol* 23(24):9338–48.

Gomos-Klein J, Harrow F, Alarcon J, Ortiz BD (2007) CTCF-independent, but not CTCF-dependent, elements significantly contribute to TCR-alpha locus control region activity. *J Immunol* 179(2):1088–95.

Gontan C, de Munck A, Vermeij M, Grosveld F, Tibboel D, Rottier R (2008) Sox2 is important for two crucial processes in lung development: Branching morphogenesis and epithelial cell differentiation. *Dev Biol* 317(1):296–309.

Goonasekera SA, Lam CK, Millay DP, Sargent MA, Hajjar RJ, Kranias EG, Molkentin JD (2011) Mitigation of muscular dystrophy in mice by SERCA overexpression in skeletal muscle. *J Clin Invest* 121(3):1044–52.

Gray CE, Coates CJ (2005) Cloning and characterization of cDNAs encoding putative CTCFs in the mosquitoes, Aedes aegypti and Anopheles gambiae. *BMC Mol Biol* 6:16.

Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, Eussen BH, et al. (2008) Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* 453(7197):948–51.

Guo Y, Monahan K, Wu H, Gertz J, Varley KE, Li W, Myers RM, et al. (2012) CTCF/cohesin-mediated DNA looping is required for protocadherin α promoter choice. *P Natl Acad Sci USA* 109(51):21081–6.

Gurudatta BV, Corces VG (2009) Chromatin insulators: lessons from the fly. *Brief Funct Genomic Proteomics* 8(4):276–82.

Haack TB, Madignier F, Herzer M, Lamantea E, Danhauser K, Invernizzi F, Koch J, et al. (2012) Mutation screening of 75 candidate genes in 152 complex I deficiency cases identifies pathogenic variants in 16 genes including NDUFB9. *Genet* 49:83-9.

Hagege H, Klous P, Braem C, Splinter E, Dekker J, Cathala G, De Laat W, et al. (2007) Quantitative analysis of chromosome conformation capture assays (3C-qPCR). *Nat Protoc* 2(7):1722–33.

Hall CV, Jacob PE, Ringold GM, Lee F (1983) Expression and regulation of Escherichia coli lacZ gene fusions in mammalian cells. *J Mol Appl Genet* 2(1):101–9.

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser* 41:95–8.

Hancock AL, Brown KW, Moorwood K, Moon H, Holmgren C, Mardikar SH, Dallosso AR, et al. (2007) A CTCF-binding silencer regulates the imprinted genes AWT1 and WT1-AS and exhibits sequential epigenetic defects during Wilms' tumourigenesis. *Hum Mol Genet,* 16(3):343–54.

Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CW, et al. (2011) CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet* 43(7):630–8.

Hardison RC, Taylor J (2012) Genomic approaches towards finding cis-regulatory modules in animals. *Nat Rev Genet* 13(7):469–83.

Hark AT, Schoenherr CJ, Katz DJ, Ingram RS, Levorse JM, Tilghman SM (2000) CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* 405(6785):486–9.

Hasan M, Koch J, Rakheja D, Pattnaik AK, Brugarolas J, Dozmorov I, Levine B, et al. (2013) Trex1 regulates lysosomal biogenesis and interferon-independent activation of antiviral genes. *Nat Immunol* 14(1):61–71.

Hasegawa SL, Moriguchi T, Rao A, Kuroha T, Engel JD, Lim KC (2007) Dosage-dependent rescue of definitive nephrogenesis by a distant Gata3 enhancer. *Dev Biol* 301(2):568–77.

Heath H, Ribeiro de Almeida C, Sleutels F, Dingjan G, Van de Nobelen S, Jonkers I, Ling KW, et al. (2008) CTCF regulates cell cycle progression of alphabeta T cells in the thymus. *EMBO J* 27(21):2839–50.

Heng HH, Goetze S, Ye CJ, Liu G, Stevens JB, Bremer SW, Wykes SM, et al. (2004) Chromatin loops are selectively anchored using scaffold/matrix-attachment regions. *J Cell Sci* 117(Pt 7):999–1008.

Herold M, Bartkuhn M, Renkawitz R (2012) CTCF: insights into insulator function during development. *Development* 139(6):1045–57.

Hiller M, Schaar BT, Bejerano G (2012) Hundreds of conserved non-coding genomic regions are independently lost in mammals. *Nucleic Acids Res* 40(22):11463–76.

Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, et al. (2006) The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* 34(Database issue):D590–8.

Hiraga S, Botsios S, Donze D, Donaldson AD (2012) TFIIIC localizes budding yeast ETC sites to the nuclear periphery. *Mol Biol Cell* 23(14):2741–54.

Hogan B, Beddington R, Costantini F, Lacy E (1994) In: *Manipulating the Mouse Embryo*, Cold Spring Harbor Laboratory Press, New York.

Hou C, Corces VG (2012) Throwing transcription for a loop: expression of the genome in the 3D nucleus. *Chromosoma* 121(2):107–16.

Hou C, Dale R, Dean A (2010) Cell type specificity of chromatin organization mediated by CTCF and cohesin. *P Natl Acad Sci USA* 107(8):3651–6.

Hou C, Zhao H, Tanimoto K, Dean A (2008) CTCF-dependent enhancer-blocking by alternative chromatin loop formation. *P Natl Acad Sci USA* 105(51):20398–403.

Houdebine LM (2000) Transgenic animal bioreactors. *Transgenic Res* 9(4-5):305–20.

Huang DW, Sherman BT, Lempicki RA (2009a) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37(1):1–13.

Huang DW, Sherman BT, Lempicki RA (2009b) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4(1):44–57.

Hurst LD, Pal C, Lercher MJ (2004) The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet* 5(4):299-310.

Iborra FJ, Pombo A, Jackson DA, Cook PR (1996) Active RNA polymerases are localized within discrete transcription " factories " in human nuclei. *J Cell Sci* 109(Pt 6):1427–36.

Ideraabdullah FY, Abramowitz LK, Thorvaldsen JL, Krapp C, Wen SC, Engel N, Bartolomei MS (2011) Novel cis-regulatory function in ICR-mediated imprinted repression of H19. *Dev Biol* 355(2):349–57.

Ideraabdullah FY, Vigneau S, Bartolomei MS (2008) Genomic imprinting mechanisms in mammals. *Mutat Res* 647(1-2):77–85.

Indra AK, Dupe V, Bornert J, Messaddeq N, Yaniv M, Mark M, Chambon P, et al. (2005) Temporally controlled targeted somatic mutagenesis in embryonic surface ectoderm and fetal epidermal keratinocytes unveils two distinct developmental functions of BRG1 in limb morphogenesis and skin barrier formation. *Development* 132(20):4533–44.

Ishihara K, Oshimura M, Nakao M (2006) CTCF-dependent chromatin insulator is linked to epigenetic remodeling. *Mol Cell* 23(5):733–42.

Ishihara K, Sasaki H (2002) An evolutionarily conserved putative insulator element near the 3' boundary of the imprinted Igf2/H19 domain. *Hum Mol Genet* 11(14):1627–36.

Ishihara SL, Morohashi K (2005) A boundary for histone acetylation allows distinct expression patterns of the Ad4BP/SF-1 and GCNF loci in adrenal cortex cells. *Biochem Biophys Res Commun* 329(2):554–62.

Jackson IJ, Bennett DC (1990) Identification of the albino mutation of mouse tyrosinase by analysis of an in vitro revertant. *Proc Natl Acad Sci USA* 87(18):7010-4.

Jiang H, Shukla A, Wang X, Chen WY, Bernstein BE, Roeder RG (2011) Role for Dpy-30 in ES cell-fate specification by regulation of H3K4 methylation within bivalente domains. *Cell* 144(4):513-25.

Jimenez-Lozano N, Segura J, Macias JR, Vega J, Carazo JM (2009) aGEM: an integrative system for analyzing spatial-temporal gene-expression information. *Bioinformatics* 25(19):2566–72.

Jimenez-Lozano N, Segura J, Macias JR, Vega J, Carazo JM (2012) Integrating human and murine anatomical gene expressaion data for improved comparisons. *Bioinformatics* 28(3):397–402.

Jones MA, Ng BG, Bhide S, Chin E, Rhodenizer D, He P, Losfeld ME, et al. (2012) DDOST mutations identified by whole-exome sequencing are implicated in congenital disorders of glycosylation. *Am J Hum Genet* 90(2):363-8.

Jothi R, Cuddapah S, Barski A, Cui K, Zhao K (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res* 36(16):5221–31.

Kadauke S, Blobel GA (2009) Chromatin loops in gene regulation. *Biochim Biophys Acta,* 1789(1):17–25.

Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, Van Berkum NL, Ebmeier CC, et al. (2010) Mediator and cohesin connect gene expression and chromatin architecture. *Nature* 467(7314):430–5.

Kalos M, Fournier RE (1995) Position-independent transgene expression mediated by boundary elements from the apolipoprotein B chromatin domain. *Mol Cell Biol* 15(1):198–207.

Kanduri C, Pant V, Loukinov D, Pugacheva E, Qi CF, Wolffe A, Ohlsson R, et al. (2000) Functional association of CTCF with the insulator upstream of the H19 gene is parent of origin-specific and methylation-sensitive. *Curr Biol* 10(14):853–6.

Kanduri M, Kanduri C, Mariano P, Vostrov AA, Quitschke W, Lobanenkov V, Ohlsson R (2002) Multiple nucleosome positioning sites regulate the CTCF-mediated insulator function of the H19 imprinting control region. *Mol Cell Biol,* 22(10):3339–44.

Karlsson J, Von Hofsten J, Olsson PE (2001) Generating transparent zebrafish: a refined method to improve detection of gene expression during embryonic development. *Mar Biotechnol (NY)* 3(6):522–7.

Karpati G, Charuk J, Carpenter S, Jablecki C, Holland P (1986) Myopathy caused by a deficiency of Ca(2+)-adenosine triphosphate in sarcoplasmic reticulum (Brody's disease). *Am Neurol* 20:38-49.

Kawajiri S, Saiki S, Sato S, Hattori N (2011) Genetic mutations and functions of PINK1. *Trends Pharmacol Sci* 32(10):573–80.

Kawakami K (2004) Transgenesis and gene trap methods in zebrafish by using the Tol2 transposable element. *Methods Cell biol* 77:201–22.

Kellum R, Schedl P (1991) A position-effect assay for boundaries of higher order chromosomal domains. *Cell* 64(5):941–50.

Kellum R, Schedl P (1992) A group of scs elements function as domain boundaries in an enhancer-blocking assay. *Mol Cell Biol* 12(5):2424–31.

Kim J, Ebersole T, Kouprina N, Noskov VN, Ohzeki J, Masumoto H, Mravinac B, et al. (2009) Human gamma-satellite DNA maintains open chromatin structure and protects a transgene from epigenetic silencing. *Genome Res* 19(4)533–44.

Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, et al. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* 128(6):1231–45.

Kimura A, Horikoshi M (2004) Partition of distinct chromosomal regions: negotiable border and fixed border. *Genes Cells* 9(6):499–508.

Kirkland JG, Raab JR, Kamakaka RT (2012) TFIIIC bound DNA elements in nuclear organization and insulation. *Biochim Biophys Acta* 1829(3-4):418–24.

Kitchen NS, Schoenherr CJ (2010) Sumoylation modulates a domain in CTCF that activates transcription and decondenses chromatin. *J Cell Biochem* 111(3):665–75.

Klenova EM, Chernukhin IV, El-Kady A, Lee RE, Pugacheva EM, Loukinov DI, Goodwin GH, et al. (2001) Functional phosphorylation sites in the C-terminal region of the multivalent multifunctional transcriptional factor CTCF. *Mol Cell Biol* 21(6):2221–34.

Klenova EM, Fagerlie S, Filippova GN, Kretzner L, Goodwin GH, Loring G, Neiman PE, et al. (1998) Characterization of the chicken CTCF genomic locus, and initial study of the cell cycle-regulated promoter of the gene. *J Biol Chem* 273(41):26571–9.

Klenova EM, Nicolas RH, Paterson HF, Carne AF, Heath CM, Goodwin GH, Neiman PE, et al. (1993) CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms. *Mol Cell Biol* 13(12):7612–24.

Klopstock N, Levy C, Olam D, Galun E, Goldenberg D (2007) Testing transgenic regulatory elements through live mouse imaging. *FEBS Lett* 581(21):3986-90.

Kmita M, Tarchini B, Duboule D, Herault Y (2002) Evolutionary conserved sequences are required for the insulation of the vertebrate Hoxd complex in neural cells. *Development* 129(23):5521–8.

Kohne AC, Baniahmad A, Renkawitz R (1993) NeP1. A ubiquitous transcription factor synergizes with v-ERBA in transcriptional silencing. *J Mol Biol* 232(3):747–55.

Kouprina N, Noskov VN, Pavlicek A, Collins NK, Schoppee Bortz PD, Ottolenghi C, Loukinov D, et al. (2007) Evolutionary diversification of SPANX-N sperm protein gene structure and expression. *PloS one* 2(4):e359.

Krivega I, Dean A (2012) Enhancer and promoter interactions-long distance calls. *Curr Opin Genet Dev 22*(2):79–85.

Krnacik MJ, Li S, Liao J, Rosen JM (1995) Position-independent expression of whey acidic protein transgenes. *J Biol Chem* 270(19):11119–29.

Kuhn RM, Haussler D, Kent WJ (2013) The UCSC genome browser and associated tools. *Brief Bioinform* 14(2):144–61.

Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, et al. (2010) Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* 42(7):631–4.

Kurukuti S, Tiwari VK, Tavoosidana G, Pugacheva E, Murrell A, Zhao Z, Lobanenkov V, et al. (2006) CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. *P Natl Acad Sci USA* 103(28):10684–9.

Kuzmin I, Geil L, Gibson L, Cavinato T, Loukinov D, Lobanenkov V, Lerman MI (2005) Transcriptional regulator CTCF controls human interleukin 1 receptor-associated kinase 2 promoter. *J Mol Biol* 346(2):411–22.

Lakshmanan G, Lieuw KH, Grosveld F, Engel JD (1998) Partial rescue of GATA-3 by yeast artificial chromosome transgenes. *Dev Biol* 204(2):451–63.

Lakshmanan G, Lieuw KH, Lim KC, Gu Y, Grosveld F, Engel JD, Karis A (1999) Localization of distant urogenital system-, central nervous system-, and endocardium-specific transcriptional regulatory elements in the GATA-3 locus. *Mol Cell Biol* 19(2):1558–68.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921.

Landy A (1989) Dynamic, structural, and regulatory aspects of lambda site-specific recombination. *Annu Rev Biochem* 58:913–49.

Lavado A, Jeffery G, Tovar V, de la Villa P, Montoliu L (2006) Ectopic expression of tyrosinase hydroxylase in the pigmented epithelium rescues the retinal abnormalities and visual function common in albinos in the absence of melanin. *J Neurochem* 96(4):1201-11.

Lee-Kirsch MA, Chowdhury D, Harvey S, Gong M, Senenko L, Engel K, Pfeiffer C, et al. (2007a) A mutation in TREX1 that impairs susceptibility to granzyme A-mediated cell death underlies familial chilblain lupus. *J Mol Med* 85:531-7.

Lee-Kirsch MA, Gong M, Chowdhury D, Senenko L, Engel K, Lee YA, de Silva U, et al. (2007b) Mutations in the gene encoding the 3-prime-5-prime DNA exonuclease TREX1 are associated with systemic lupus erythematosus. *Nat Genet* 39:1065-7.

Lefevre P, Witham J, Lacroix CE, Cockerill PN, Bonifer C (2008) The LPS-induced transcriptional upregulation of the chicken lysozyme locus involves CTCF eviction and noncoding RNA transcription. *Mol Cell* 32(1):129–39.

Lettice LA, Heaney SJH, Purdie LA, Li L, De Beer P, Oostra BA, Goode D, et al. (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* 12(14):1725–35.

Levine M, Tjian R (2003) Transcription regulation and animal diversity. *Nature* 424(6945):147–51.

Li D, Parks SB, Kushner JD, Nauman D, Burgess D, Ludwigsen S, Partain J, et al. (2006) Mutations of presenilin genes in dilated cardiomyopathy and heart failure. *Am J Hum Genet* 79(6):1030-9.

Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, et al. (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* 148(1-2):84–98.

Li Q, Stamatoyannopoulos G (1994) Hypersensitive site 5 of the human beta locus control region functions as a chromatin insulator. *Blood* 84(5):1399–401.

Li Y, Huang W, Niu L, Umbach DM, Covo S, Li L (2013) Characterization of constitutive CTCF/cohesin loci: a possible role in establishing topological domains in mammalian genomes. *BMC Genomics* 14(1):553.

Libby RT, Hagerman KA, Pineda VV, Lau R, Cho DH, Baccam SL, Axford MM, et al. (2008) CTCF cis-regulates trinucleotide repeat instability in an epigenetic manner: a novel basis for mutational hot spot determination. *PLoS Genet* 4(11):e1000257.

Liebenberg F (1973) A pedigree with unusual anomalies of the elbows, wrists and hands in five generations. *S Afr Med J* 47(17):745-8.

Lindstrom MS, Zhang Y (2008) Ribosomal protein S9 is a novel B23/NPM-binding protein required for normal cell proliferation. *J Biol Chem* 283(23):15568-76.

Litt MD, Simpson M, Gaszner M, Allis CD, Felsenfeld G (2001a) Correlation between histone lysine methylation and developmental changes at the chicken beta-globin locus. *Science* 293(5539):2453–5.

Litt MD, Simpson M, Recillas-Targa F, Prioleau MN, Felsenfeld G (2001b) Transitions in histone acetylation reveal boundaries of three separately regulated neighboring loci. *EMBO J* 20(9):2224–35.

Lobanenkov VV, Nicolas RH, Adler VV, Paterson H, Klenova EM, Polotskaja AV, Goodwin GH (1990) A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. *Oncogene* 5(12):1743–53.

Lunyak VV (2008) Boundaries. Boundaries...Boundaries??? *Curr Opin Cell Biol* 20(3):281–7.

Lunyak VV, Prefontaine GG, Nuñez E, Cramer T, Ju BG, Ohgi KA, Hutt K, et al. (2007) Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science* 317(5835):248–51.

Lutz M, Burke LJ, Barreto G, Goeman F, Greb H, Arnold R, Schultheiss H, et al. (2000) Transcriptional repression by the insulator protein CTCF involves histone deacetylases. *Nucleic Acids Res* 28(8):1707–13.

Lutz M, Burke LJ, LeFevre P, Myers FA, Thorne AW, Crane-Robinson C, Bonifer C, et al. (2003) Thyroid hormone-regulated enhancer blocking: cooperation of CTCF and thyroid hormone receptor. *EMBO J* 22(7):1579–87.

MacGregor GR, Caskey CT (1989) Construction of plasmids that express E. coli beta-galactosidase in mammalian cells. *Nucleic Acids Res* 17(6):2365.

MacPherson MJ, Beatty LG, Zhou W, Du M, Sadowski PD (2009) The CTCF insulator protein is posttranslationally modified by SUMO. *Mol Cell Biol* 29(3):714–25.

Magdinier F, Yusufzai TM, Felsenfeld G (2004) Both CTCF-dependent and -independent insulators are found between the mouse T cell receptor alpha and Dad1 genes. *J Biol Chem* 279(24):25381–9.

Majocchi S, Aritonovska E, Mermod N (2014) Epigenetic regulatory elements associate with specific histone modifications to prevent silencing of telomeric genes. *Nucleic Acids Res* 42(1):193–204.

Marone R, Hess D, Dankort D, Muller WJ, Hynes NE, Badache A (2004) Memo mediates ErbB2-driven cell motility. *Nat Cell Biol* 6(6):515-22.

Martin D, Pantoja C, Fernandez Miñan A, Valdes-Quezada C, Molto E, Matesanz F, Bogdanovic O, et al. (2011) Genome-wide CTCF distribution in vertebrates defines equivalent sites that aid the identification of disease-associated genes. *Nat Struct Mol Biol* 18(6):708–14.

Maston GA, Evans SK, Green MR (2006) Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* 7:29–59.

Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, et al. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 34(Database issue):D108–10.

Maurano MT, Wang H, Kutyavin T, Stamatoyannopoulos JA (2012) Widespread site-dependent buffering of human regulatory polymorphism. *PLoS Genet* 8(3):e1002599.

McFarland WN, Klontz GW (1969) Anesthesia in fishes. *Fed Proc* 28(4):1535–40.

McKnight RA, Shamay A, Sankaran L, Wall RJ, Hennighausen L (1992) Matrix-attachment regions can impart position-independent regulation of a tissue-specific gene in transgenic mice. *P Natl Acad Sci USA* 89(15):6943–7.

Medina I, Carbonell J, Pulido L, Madeira SC, Goetz S, Conesa A, Tarraga J, et al. (2010) Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic Acids Res* 38(Web Server issue):W210–3.

Merrill BJ, Gat U, Dasgupta R, Fuchs E (2001) Tcf3 and Lef1 regulate lineage differentiation of multipotent stem cells in skin. *Genes Dev* 15(13):1688–705.

Millot B, Montoliu L, Fontaine M, Mata T, Devinoy E (2003) Hormone-induced modifications of the chromatin structure surrounding upstream regulatory regions conserved between the mouse and rabbit whey acidic protein genes. *Biochem J* 372(Pt 1):41–52.

Mirkovitch J, Mirault ME, Laemmli UK (1984) Organization of the higher-order chromatin loop: specific DNA attachment sites on nuclear scaffold. *Cell* 39(1):223–32.

Mohun TJ, Garret N, Gurdon JB (1986) Upstream sequences required for tissue-specific activation of the cardiac actin gene in Xenopus laevis embryos. *EMBO J* 5:3185-93.

Molto E, Fernandez A, Montoliu L (2009) Boundaries in vertebrate genomes: different solutions to adequately insulate gene expression domains. *Brief Funct Genomic Proteomic* 8(4):283–96.

Molto E, Vicente-Garcia C, Fernandez A, Montoliu L (2011) Genomic insulators in transgenic animals. Brakebusch C, Pihlajaniemi T (eds.) In: *Mouse as a model organism – from animals to cells,* Springer.

Monahan K, Rudnick ND, Kehayova PD, Pauli F, Newberry KM, Myers RM, Maniatis T (2012) Role of CCCTC binding factor (CTCF) and cohesin in the generation of single-cell diversity of protocadherin-α gene expression. *P Natl Acad Sci USA* 109(23):9125–30.

Montazer-Torbati MB, Hue-Beauvais C, Droineau S, Ballester M, Coant N, Aujean E, Petitbarat M, et al. (2008) Epigenetic modifications and chromatin loop organization explain the different expression profiles of the Tbrg4, WAP and Ramp3 genes. *Exp Cell Res* 314(5):975–87.

Montoliu L (1997) In: *Generation of transgenic mice: A laboratory manual*, Heidelberg: German Cancer Research Center.

Montoliu L (2002) Gene transfer strategies in animal transgenesis. *Cloning Stem Cells* 4(1):39-46.

Montoliu L (2003) Simple databases to monitor the generation and organisation of transgenic mouse colonies. *Transgenic Res* 12(2):251–3.

Montoliu L, Roy R, Regales L, Garcia-Diaz A (2009) Design of vectors for transgene expression: The use of genomic comparative approaches. *Comp Immunol Microbiol Infect Dis* 32(2):81–90.

Montoliu L, Umland T, Schutz G (1996) A locus control region at -12 kb of the tyrosinase gene. *EMBO J* 15(22):6026–34.

Moon H, Filippova G, Loukinov D, Pugacheva E, Chen Q, Smith ST, Munhall A, et al. (2005) CTCF is conserved from Drosophila to humans and confers enhancer blocking of the Fab-8 insulator. *EMBO Rep* 6(2):165–70.

Moreno R, Martinez I, Petriz J, Nadal M, Tintore X, Gonzalez JR, Gratacos E, et al. (2011) The β-interferon scaffold attachment region confers high-level transgene expression and avoids extinction by epigenetic modifications of integrated provirus in adipose tissue-derived human mesenchymal stem cells. *Tissue Eng Part C Methods* 17(3):275–87.

Morita M, Stamp G, Robins P, Dulic A, Rosewell I, Hrivnak G, Daly G, et al. (2004) Gene-targeted mice lacking the Trex1 (DNase III) 3'-->5' DNA exonuclease develop inflammatory myocarditis. *Mol Cell Biol* 24(15):6719–27.

Mouse ENCODE Consortium, Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, et al. (2012) An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol* 13(8):418.

Mukhopadhyay R, Yu W, Whitehead J, Xu J, Lezcano M, Pack S, Kanduri C, et al. (2004) The binding sites for the chromatin insulator protein CTCF map to DNA methylation-free domains genome-wide. *Genome Res* 14(8):1594–602.

Mukhopadhyay S, Schedl P, Studitsky VM, Sengupta AM (2011) Theoretical analysis of the role of chromatin interactions in long-range action of enhancers and insulators. *P Natl Acad Sci USA* 108(50):19919–24.

Murai J, Ikegami D, Okamoto M, Yoshikawa H, Tsumaki N (2008) Insulation of the ubiquitous Rxrb promoter from the cartilage-specific adjacent gene, Col11a2. *J Biol Chem* 283(41):27677–87.

Nakahashi H, Kwon KRK, Resch W, Vian L, Dose M, Stavreva D, Hakim O, et al. (2013) A Genome-wide Map of CTCF Multivalency Redefines the CTCF Code. *Cell Rep* 3(5):1678–89.

Namavar Y, Barth PG, Kasher PR, van Ruissen F, Brockmann K, Bernert G, Writzl K, et al. (2011) Clinical, neuroradiological and genetic findings in pontocerebellar hypoplasia. *Brain* 134(Pt 1):143-56.

Narlikar L, Ovcharenko I (2009) Identifying regulatory elements in eukaryotic genomes. *Brief Funct Genomics Proteomic* 8(4):215–30.

Negre N, Brown CD, Shah PK, Kheradpour P, Morrison CA, Henikoff JG, Feng X, et al. (2010) A comprehensive map of insulator elements for the Drosophila genome. *PLoS Genet* 6(1):e1000814.

Neznanov N, Umezawa A, Oshima RG (1997) A regulatory element within a coding exon modulates keratin 18 gene expression in transgenic mice. *J Biol Chem* 272(44):27549-57.

Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, et al. (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 485(7398):381–5.

Obayashi T, Okamura Y, Ito S, Tadaka S, Motoike IN, Kinoshita K (2013) COXPRESdb: a database of comparative gene coexpression networks of eleven species for mammals. *Nucleic Acids Res* 41(Database issue):D1014–20.

Oberdoerffer P, Sinclair DA (2007) The role of nuclear architecture in genomic instability and ageing. *Nat Rev Mol Cell Biol* 8(9):692–702.

Ohlsson R, Bartkuhn M, Renkawitz R (2010a) CTCF shapes chromatin by multiple mechanisms: the impact of 20 years of CTCF research on understanding the workings of chromatin. *Chromosoma* 119(4):351–60.

Ohlsson R, Lobanenkov V, Klenova E (2010b) Does CTCF mediate between nuclear organization and gene expression? *BioEssays* 32(1):37–50.

Ortiz BD, Harrow F, Cado D, Santoso B, Winoto A (2001) Function and factor interactions of a locus control region element in the mouse T cell receptor-alpha/Dad1 gene locus. *J Immunol* 167(7):3836–45.

Ovaere P, Lippens S, Vandenabeele P, Declercq W (2009) The emerging roles of serine protease cascades in the epidermis. *Trends Biochem Sci* 34(9):453–63.

Ovcharenko I, Nobrega MA, Loots GG, Stubbs L (2004) ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucleic Acids Res* 32(Web Server issue):W280–6.

Owczarzy R, Tataurov AV, Wu Y, Manthey JA, McQuisten KA, Almabrazi HG, Pedersen KF, et al. (2008) IDT SciTools: a suite for analysis and design of nucleic acid oligomers. *Nucleic Acids Res* 36(Web Server issue):W163–9.

Palla F, Melfi R, Anello L, Di Bernardo M, Spinelli G (1997) Enhancer blocking activity located near the 3' end of the sea urchin early H2A histone gene. *P Natl Acad Sci USA* 94(6):2272–7.

Palstra RJ, Tolhuis B, Splinter E, Nijmeijer R, Grosveld F, De Laat W (2003) The beta-globin nuclear compartment in development and erythroid differentiation. *Nat Genet* 35(2):190–4.

Pant V, Kurukuti S, Pugacheva E, Shamsuddin S, Mariano P, Renkawitz R, Klenova E, et al. (2004) Mutation of a single CTCF target site within the H19 imprinting control region leads to loss of Igf2 imprinting and complex patterns of de novo methylation upon maternal inheritance. *Mol Cell Biol* 24(8):3497–504.

Parelho V, Hadjur S, Spivakov M, Leleu M, Sauer S, Gregson HC, Jarmuz A, et al. (2008) Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* 132(3):422–33.

Parrish JZ, Xue D (2003) Functional genomic analysis of apoptotic DNA degradation in C. elegans. *Mol Cell* 11(4):987-96.

Pederson T (2000) Half a century of "the nuclear matrix". *Mol Biol Cell* 11(3):799–805.

Phi-Van L, Stratling WH (1996) Dissection of the ability of the chicken lysozyme gene 5' matrix attachment region to stimulate transgene expression and to dampen position effects. *Biochemistry* 35(33):10735–42.

Phi-Van L, Von Kries JP, Ostertag W, Stratling WH (1990) The chicken lysozyme 5' matrix attachment region increases transcription from a heterologous promoter in heterologous cells and dampens position effects on the expression of transfected genes. *Mol Cell Biol* 10(5):2302–7.

Phillips JE, Corces VG (2009) CTCF: master weaver of the genome. *Cell* 137(7):1194–211.

Phillips-Cremins JE, Corces VG (2013a) Chromatin insulators: linking genome organization to cellular function. *Mol Cell* 50(4):461–74.

Phillips-Cremins JE, Sauria MEG, Sanyal A, Gerasimova TI, Lajoie BR, Bell JSK, Ong CT, et al. (2013b) Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage Commitment. *Cell* 153(6):1281–95.

Pikaart MJ, Recillas-Targa F, Felsenfeld G (1998) Loss of transcriptional activity of a transgene is accompanied by DNA methylation and histone deacetylation and is prevented by insulators. *Genes Dev* 12(18):2852–62.

Poljak L, Seum C, Mattioni T, Laemmli UK (1994) SARs stimulate but do not confer position independent gene expression. *Nucleic Acids Res* 22(21):4386–94.

Prioleau MN, Nony P, Simpson M, Felsenfeld G (1999) An insulator element and condensed chromatin region separate the chicken beta-globin locus from an independently regulated erythroid-specific folate receptor gene. *EMBO J* 18(14):4035–48.

Pugacheva EM, Kwon YW, Hukriede NA, Pack S, Flanagan PT, Ahn JC, Park JA, et al. (2006) Cloning and characterization of zebrafish CTCF: Developmental expression patterns, regulation of the promoter region, and evolutionary aspects of gene organization. *Gene* 375:26–36.

Puthenveetil G, Scholes J, Carbonell D, Qureshi N, Xia P, Zeng L, Li S, et al. (2004) Successful correction of the human beta-thalassemia major phenotype using a lentiviral vector. *Blood* 104(12):3445–53.

Qiu J, Yoon JH, Shen B (2005) Search for apoptotic nucleases in yeast: role of Tat-D nuclease in apoptotic DNA degradation. *J Biol Chem* 280(15):15370-9.

Raab JR, Chiu J, Zhu J, Katzman S, Kurukuti S, Wade PA, Haussler D, et al. (2012) Human tRNA genes function as chromatin insulators. *EMBO J* 31(2):330–50.

Rajapakse I, Perlman MD, Scalzo D, Kooperberg C, Groudine M, Kosak ST (2009) The emergence of lineage-specific chromosomal topologies from coordinate gene regulation. *P Natl Acad Sci USA* 106(16):6679–84.

Rasko JE, Klenova EM, Leon J, Filippova GN, Loukinov DI, Vatolin S, Robinson AF, et al. (2001) Cell growth inhibition by the multifunctional multivalent zinc-finger factor CTCF. *Cancer Res* 61(16):6002–7.

Raux G, Gantier R, Thomas-Anterion C, Boulliat J, Verpillat P, Hannequin D, Brice A, et al. (2000) Dementia with prominente frontotemporal features associated with L113P presenilin 1 mutation. *Neurology* 55(10):1577-8.

Recillas-Targa F, Bell AC, Felsenfeld G (1999) Positional enhancer-blocking activity of the chicken beta-globin insulator in transiently transfected cells. *P Natl Acad Sci USA* 96(25):14354–9.

Recillas-Targa F, Pikaart MJ, Burgess-Beusse B, Bell AC, Litt MD, West AG, Gaszner M, et al. (2002) Position-effect protection and enhancer blocking by the chicken beta-globin insulator are separable activities. *P Natl Acad Sci USA* 99(10):6883–8.

Recillas-Targa F, Valadez-Graham V, Farrell CM (2004) Prospects and implications of using chromatin insulators in gene therapy and transgenesis. *Bioessays* 26(7):796-807.

Reddi PP, Shore AN, Shapiro JA, Anderson A, Stoler MH, Acharya KK (2003) Spermatid-specific promoter of the SP-10 gene functions as an insulator in somatic cells. *Dev Biol* 262(1):173–82.

Renaud S, Loukinov D, Abdullaev Z, Guilleret I, Bosman FT, Lobanenkov V, Benhattar J (2007) Dual role of DNA methylation inside and outside of CTCF-binding regions in the transcriptional regulation of the telomerase hTERT gene. *Nucleic Acids Res* 35(4):1245–56.

Renaud S, Loukinov D, Bosman FT, Lobanenkov V, Benhattar J (2005) CTCF binds the proximal exonic region of hTERT and inhibits its transcription. *Nucleic Acids Res* 33(21):6850–60.

Renda M, Baglivo I, Burgess-Beusse B, Esposito S, Fattorusso R, Felsenfeld G, Pedone PV (2007) Critical DNA binding interactions of the insulator protein CTCF: a small number of zinc fingers mediate strong binding, and a single finger-DNA interaction controls binding at imprinted loci. *J Biol Chem* 282(46):33336–45.

Rhee HS, Pugh BF (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* 147(6):1408–19.

Rice G, Patrick T, Parmar R, Taylor CF, Aeby A, Aicardi J, Artuch R, et al. (2007) Clinical and molecular phenotype of Aicardi-Goutieres syndrome. *Am J Hum Genet* 81:713-25.

Richards A, van den Maagdenberg AMJM, Jen JC, Kavanagh D, Bertram P, Spitzer D, Liszewski MK, et al. (2007) C-terminal truncations in human 3-prime-5-prime DNA exonuclease TREX1 cause autosomal dominant retinal vasculopathy with cerebral leukodystrophy. *Nat Genet* 39:1068-70.

Riethoven JM (2010) Regulatory regions in DNA: promoters, enhancers, silencers, and insulators. *Methods Mol Biol* 674(1):33–42.

Robinett CC, O'Connor A, Dunaway M (1997) The repeat organizer, a specialized insulator element within the intergenic spacer of the Xenopus rRNA genes. *Mol Cell Biol* 17(5):2866–75.

Roman AC, Benitez DA, Carvajal-Gonzalez JM, Fernandez-Salguero PM (2008) Genome-wide B1 retrotransposon binds the transcription factors dioxin receptor and Slug and regulates gene expression in vivo. *P Natl Acad Sci USA* 105(5):1632–7.

Roman AC, Gonzalez-Rico FJ, Fernandez-Salguero PM (2011b) B1-SINE retrotransposons: Establishing genomic insulatory networks. *Mob Genet Elements* 1(1):66–70.

Roman AC, Gonzalez-Rico FJ, Molto E, Hernando H, Neto A, Vicente-Garcia C, Ballestar E, et al. (2011a) Dioxin receptor and SLUG transcription factors regulate the insulator activity of B1 SINE retrotransposons via an RNA polymerase switch. *Genome Res* 21(3):422–32.

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 132:365–86.

Rubio ED, Reiss DJ, Welcsh PL, Disteche CM, Filippova GN, Baliga NS, Aebersold R, et al. (2008) CTCF physically links cohesin to chromatin. *P Natl Acad Sci USA* 105(24):8309-14.

Ruf S, Symmons O, Uslu VV, Dolle D, Hot C, Ettwiller L, Spitz F (2011) Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor. *Nat Genet* 43(4):379–86.

Saitoh N, Bell AC, Recillas-Targa F, West AG, Simpson M, Pikaart M, Felsenfeld G (2000) Structural and functional conservation at the boundaries of the chicken beta-globin domain. *EMBO J* 19(10):2315–22.

Sambrook J, Fritsch EF, Maniatis T (1989) In: *Molecular Cloning. A laboratory manual,* Cold Spring Harbor Laboratory Press, New York.

Sano S, Kira M, Takagi S, Yoshikawa K, Takeda J, Itami S (2000) Two distinct signaling pathways in hair cycle induction: Stat3-dependent and -independent pathways. *P Natl Acad Sci USA* 97(25):13824–9.

Sanyal A, Lajoie BR, Jain G, Dekker J (2012) The long-range interaction landscape of gene promoters. *Nature* 489(7414):109–13.

Schedl A, Beermann F, Thies E, Montoliu L, Kelsey G, Schutz G (1992) Transgenic mice generated by pronuclear injection of a yeast artificial chromosome. *Nucleic Acids Res* 20(12):3073–7.

Schedl A, Montoliu L, Kelsey G, Schutz G (1993) A yeast artificial chromosome covering the tyrosinase gene confers copy number-dependent expression in transgenic mice. *Nature* 362(6417):258–61.

Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonçalves A, Kutter C, Brown GD, et al. (2012) Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* 148(1-2):335–48.

Schoborg TA, Labrador M (2010) The phylogenetic distribution of non-CTCF insulator proteins is limited to insects and reveals that BEAF-32 is Drosophila lineage specific. *J Mol Evol* 70(1):74–84.

Schoenfelder S, Sexton T, Chakalova L, Cope NF, Horton A, Andrews S, Kurukuti S, et al. (2010) Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat Genet* 42(1):53–61.

Segre JA, Bauer C, Fuchs E (1999) Klf4 is a transcription factor required for establishing the barrier function of the skin. *Nat Genet* 22(4):356–60.

Segrelles C, Holguin A, Hernandez P, Ariza JM, Paramio JM, Lorz C (2011) Establishment of a murine epidermal cell line suitable for in vitro and in vivo skin modelling. *BMC Dermatol* 11(1):9.

Seitan VC, Krangel MS, Merkenschlager M (2012) Cohesin, CTCF and lymphocyte antigen receptor locus rearrangement. *Trends Immunol* 33(4):153–59.

Shaw G, Morse S, Ararat M, Graham FL (2002) Preferential transformation of human neuronal cells by human adenoviruses and the origino f HEK 293 cells. *FASEB J* 16(8):869-71.

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, et al. (2012) A map of the cis-regulatory sequences in the mouse genome. *Nature* 488(7409):116–20.

Shih HY, Krangel MS (2013) Chromatin Architecture, CCCTC-Binding Factor, and V(D)J Recombination: Managing Long-Distance Relationships at Antigen Receptor Loci. *J Immunol* 190(10):4915–21.

Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, Oberdoerffer P, et al. (2011) CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* 479(7371):74–9.

Simabuco FM, Morello LG, Aragao AZ, Paes Leme AF, Zanchin NI (2012) Proteomic characterization of the human FTSJ3 preribosomal complexes. *J Proteome Res* 11(6):3112-26.

Simms TA, Dugas SL, Gremillion JC, Ibos ME, Dandurand MN, Toliver TT, Edwards DJ, et al. (2008) TFIIIC binding sites function as both heterochromatin barriers and chromatin insulators in Saccharomyces cerevisiae. *Eukaryot Cell* 7(12):2078–86.

Simonis M, Kooren J, De Laat W (2007) An evaluation of 3C-based methods to capture DNA interactions. *Nat Methods* 4(11):895–901.

Singer SD, Liu Z, Cox KD (2012) Minimizing the unpredictability of transgene expression in plants: the role of genetic insulators. *Plant Cell Rep* 31(1):13–25.

Soulier S, Stinnakre MG, Da Silva JC, Lepourry L, Mata X, Besnard N, Vilotte JL (2000) Distal element(s) is(are) required for position-independent expression of the goat alpha-lactalbumin gene in transgenic mice. Potential relationship with the location of the cyclin T1 locus. *Genet Sel Evol* 32(6):621–30.

Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol* 98(3):503–17.

Spencer RJ, Del Rosario BC, Pinter SF, Lessing D, Sadreyev RI, Lee JT (2011) A boundary element between Tsix and Xist binds the chromatin insulator Ctcf and contributes to initiation of X-chromosome inactivation. *Genetics* 189(2):441–54.

Spielmann M, Brancati F, Krawitz PM, Robinson PN, Ibrahim DM, Franke M, Hecht J, et al. (2012) Homeotic arm-to-leg transformation associated with genomic rearrangements at the PITX1 locus. *Am J Hum Genet* 91(4):629-35.

Spielmann M, Mundlos S (2013) Structural variations, the regulatory landscape of the genome and their alteration in human disease. *Bioessays* 35(6):533-43.

Splinter E, De Laat W (2011) The complex transcription regulatory landscape of our genome: control in three dimensions. *EMBO J* 30(21):4345–55.

Splinter E, Heath H, Kooren J, Palstra RJ, Klous P, Grosveld F, Galjart N, De Laat W (2006) CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes Dev* 20(17):2349-54.

Srivastava S, Puri D, Garapati HS, Dhawan J, Mishra RK (2013) Vertebrate GAGA factor associated insulator elements demarcate homeotic genes in the HOX clusters. *Epigenetics Chromatin* 6(1):8.

Stief A, Winter DM, Stratling WH, Sippel AE (1989) A nuclear DNA attachment element mediates elevated and position-independent gene activity. *Nature* 341(6240):343–5.

Szczepinska T, Pawłowski K (2013) Genomic positions of co-expressed genes: echoes of chromosome organization in gene expression data. *BMC Res Notes* 6:229.

Tena JJ, Alonso ME, de la Calle-Mustienes E, Splinter E, De Laat W, Manzanares M, Gomez-Skarmeta JL (2011) An evolutionarily conserved three-dimensional structure in the vertebrate Irx clusters facilitates enhancer sharing and co-regulation. *Nat Commun* 2:310.

The Uniprot Consortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res* 40(Database issue):D71–5.

Thompson M, Haeusler RA, Good PD, Engelke DR (2003) Nucleolar clustering of dispersed tRNA genes. *Science* 302(5649):1399–401.

Tiana M, Villar D, Perez-Guijarro E, Gomez-Maldonado L, Molto E, Fernandez-Miñan A, Gomez-Skarmeta JL, et al. (2012) A role for insulator elements in the regulation of gene expression response to hypoxia. *Nucleic Acids Res* 40(5):1916-27.

Tolhuis B, Palstra RJ, Splinter E, Grosveld F, De Laat W (2002) Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol Cell* 10(6):1453–65.

Torrano V, Chernukhin I, Docquier F, D'Arcy V, Leon J, Klenova E, Delgado MD (2005) CTCF regulates growth and erythroid differentiation of human myeloid leukemia cells. *J Biol Chem* 280(30):28152–61.

Torrano V, Navascues J, Docquier F, Zhang R, Burke LJ, Chernukhin I, Farrar D, et al. (2006) Targeting of CTCF to the nucleolus inhibits nucleolar transcription through a poly(ADP-ribosyl)ation-dependent mechanism. *J Cell Sci* 119(Pt 9):1746–59.

Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM (2004) An abundance of bidirectional promoters in the human genome. *Genome Res* 14(1):62-6.

Udvardy A, Maine E, Schedl P (1985) The 87A7 chromomere. Identification of novel chromatin structures flanking the heat shock locus that may define the boundaries of higher order domains. *J Mol Biol* 185(2):341–58.

Valadez-Graham V, Razin SV, Recillas-Targa F (2004) CTCF-dependent enhancer blockers at the upstream region of the chicken alpha-globin gene domain. *Nucleic Acids Res* 32(4):1354–62.

Valenzuela L, Dhillon N, Kamakaka RT (2009) Transcription independent insulation at TFIIIC-dependent insulators. *Genetics* 183(1):131–48.

Valenzuela L, Kamakaka RT (2006) Chromatin insulators. *Annu Rev Genet* 40:107–38.

Van der Vlag J, Den Blaauwen JL, Sewalt RG, Van Driel R, Otte AP (2000) Transcriptional repression mediated by polycomb group proteins and other chromatin-associated repressors is selectively blocked by insulators. *J Biol Chem* 275(1):697–704.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, et al. (2001) The sequence of the human genome. *Science* 291(5507):1304–51.

Vostrov AA, Quitschke WW (1997) The zinc finger protein CTCF binds to the APBbeta domain of the amyloid beta-protein precursor promoter. Evidence for a role in transcriptional activation. *J Biol Chem* 272(52):33353–9.

Wallace JA, Felsenfeld G (2007) We gather together: insulators and genome organization. *Curr Opin Genet Dev* 17(5):400–7.

Wan H, Dingle S, Xu Y, Besnard V, Kaestner KH, Ang S, Wert S, et al. (2005) Compensatory roles of Foxa1 and Foxa2 during lung morphogenesis. *J Biol Chem* 280(14):13809–16.

Wang B, Yang W, Wen W, Sun J, Su B, Liu B, Ma D, et al. (2010) Gamma-secretase gene mutations in familial acne inversa. *Science* 330(6007):1065.

Wang H, Maurano MT, Qu H, Varley KE, Gertz J, Pauli F, Lee K, et al. (2012) Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res* 22(9):1680–8.

Wang J, Lunyak VV, Jordan IK (2012) Genome-wide prediction and analysis of human chromatin boundary elements. *Nucleic Acids Res* 40(2):511–29.

Wang J, Vicente-Garcia C, Molto E, Fernandez-Miñan A, Neto A, Gomez-Skarmeta JL, Montoliu L, et al. (*Submitted*) MIR retrotransposon sequences provide insulators to the human genome.

Watanabe S, Watanabe S, Sakamoto N, Sato M, Akasaka K (2006) Functional analysis of the sea urchin-derived arylsulfatase (Ars)-element in mammalian cells. *Genes Cells* 11(9):1009–21.

Wendt KS, Yoshida K, Itoh T, Bando M, Koch B, Schirghuber E, Tsutsumi S, et al. (2008) Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* 451(7180):796–801.

West AG, Gaszner M, Felsenfeld G (2002) Insulators: many functions, many mechanisms. *Genes Dev* 16(3):271–88.

West AG, Huang S, Gaszner M, Litt MD, Felsenfeld G (2004) Recruitment of histone modifications by USF proteins at a vertebrate barrier element. *Mol Cell* 16(3):453–63.

Westerfield M (2007) In: *The Zebrafish Book. A guide for the laboratory use of zebrafish (Danio rerio)*, 5th Edition, Eugene, University of Oregon Press.

Weth O, Weth C, Bartkuhn M, Leers J, Uhle F, Renkawitz R (2010) Modular insulators: genome wide search for composite CTCF/thyroid hormone receptor binding sites. *PloS one* 5(4):e10119.

Whittock NV (2003) Genomic sequence analysis of the mouse desmoglein cluster reveals evidence for six distinct genes: characterization of mouse DSG4, DSG5, and DSG6. *J Invest Dermatol* 120(6):970–80.

Wilson C, Bellen HJ, Gehring WJ (1990) Position effects on eukaryotic gene expression. *Annu Rev Cell Biol* 6:679-714.

Woo YH, Walker M, Churchill GA (2010) Coordinated expression domains in mammalian genomes. *PLoS One* 5(8):e12158.

Woolfe A, Goode DK, Cooke J, Callaway H, Smith S, Snell P, McEwen GK, et al. (2007) CONDOR: a database resource of developmentally associated conserved non-coding elements. *BMC Dev Biol* 7:100.

Xiao T, Wallace J, Felsenfeld G (2011) Specific sites in the C terminus of CTCF interact with the SA2 subunit of the cohesin complex and are required for cohesin-dependent insulation activity. *Mol Cell Biol* 31(11):2174–83.

Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, Lander ES (2007) Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *P Natl Acad Sci USA* 104(17):7145–50.

Xu Z, Wei G, Chepelev I, Zhao K, Felsenfeld G (2011) Mapping of INS promoter interactions reveals its role in long-range regulation of SYT8 transcription. *Nat Struct Mol Biol* 18(3):372-8.

Yamagishi T, Ozawa M, Ohtsuka C, Ohyama-Goto R, Kondo T (2007) Evx2-Hoxd13 intergenic region restricts enhancer association to Hoxd13 promoter. *PloS one* 2(1):e175.

Yang G, Hinson MD, Bordner JE, Lin QS, Fernando AP, La P, Wright CJ, et al. (2011a) Silencing hyperoxia-induced C/EBPα in neonatal mice improves lung architecture via enhanced proliferation of alveolar epithelial cells. *Am J Physiol Lung Cell Mol Physiol* 301(2):L187–96.

Yang H, Liu C, Jamsen J, Wu Z, Wang Y, Chen J, Zheng L, Shen B (2012) The DNase domain-containing protein TATDN1 plays an important role in chromosomal segregation and cell cycle progression during zebrafish eye development. *Cell Cycle* 11(24):4626-32.

Yang J, Corces VG (2011b) Chromatin insulators: a role in nuclear organization and gene expression. *Adv Cancer Res* 110:43–76.

Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL (2012) Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* 13:134.

Yoon B, Herman H, Hu B, Park YJ, Lindroth A, Bell A, West AG, et al. (2005) Rasgrf1 imprinting is regulated by a CTCF-dependent methylation-sensitive enhancer blocker. *Mol Cell Biol* 25(24):11184–90.

Yu S, Zhou X, Hsiao JJ, Yu D, Saunders TL, Xue HH (2012) Fidelity of a BAC-EGFP transgene in reporting dynamic expression of IL-7Rα in T cells. *Transgenic Res* 21(1):201–15.

Yu W, Ginjala V, Pant V, Chernukhin I, Whitehead J, Docquier F, Farrar D, et al. (2004) Poly(ADP-ribosyl)ation regulates CTCF-dependent chromatin insulation. *Nat Genet* 36(10):1105–10.

Yusufzai TM, Felsenfeld G (2004a) The 5'-HS4 chicken beta-globin insulator is a CTCF-dependent nuclear matrix-associated element. *P Natl Acad Sci USA* 101(23):8620–4.

Yusufzai TM, Tagami H, Nakatani Y, Felsenfeld G (2004b) CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. *Mol Cell* 13(2):291–8.

Zhao Z, Tavoosidana G, Sjolinder M, Gondor A, Mariano P, Wang S, Kanduri C, et al. (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet* 38(11):1341–7.

Zhong XP, Krangel MS (1999) Enhancer-blocking activity within the DNase I hypersensitive site 2 to 6 region between the TCR alpha and Dad1 genes. *J Immunol* 163(1):295–300.

Zhou A, Ou AC, Cho A, Benz EJ Jr, Huang SC (2008) Novel splicing factor RBM25 modulates Bcl-x pre-mRNA 5'splice site selection. *Mol Cell Biol* 28(19):5924-36.

Zhu S, Oh H, Shim M, Sterneck E, Johnson PF, Smart RC (1999) C / EBPbeta Modulates the Early Events of Keratinocyte Differentiation Involving Growth Arrest and Keratin 1 and Keratin 10 Expression. *Mol Cell Biol* 19(10):7181–90.

Ziebarth JD, Bhattacharya A, Cui Y (2013) CTCFBSDB 2.0: a database for CTCF-binding sites and genome organization. *Nucleic Acids Res* 41(Database issue):D188–94.

Zlatanova J, Caiafa P (2009) CTCF and its protein partners: divide and rule*? J Cell Sci* 122(Pt 9):1275–84.

APPENDICES

Insulator	Chromosome	5' Gene	3' Gene	Properties
adHS1	2	*Nr5a1*	*Nr6a1*	Three DNaseI HS exist between *Nr5a1* (*Ad4BP/SF-1*) and *Nr6a1* (*GCNF*), whose expression patterns greatly differ: one binds CTCF and the others are MARs. There is a clear histone acetylation boundary in between the genes (Ishihara & Morohashi, 2005).
Evx2/Hoxd	2	*Evx2*	*Hoxd13*	An insulator element exists in the intergenic region between *Evx2* and the *Hoxd* gene cluster. It restrains central nervous system enhancers present in the locus to the *Evx2* gene, protecting the *Hox* developmental genes from undesired regulatory input. Also, it prevents chromosomal position effects in transgenic events. However, its activity is not constitutive: it depends on an adjacent regulator (Kmita et al., 2002; Yamagishi et al., 2007).
Scl/Map17	4	*Pdzk1ip1*	*Cyp4x1*	A DNaseI HS that binds CTCF separates the haematopoietic-specific *Scl/Map17* (*Pdzk1ip1*) domain from the liver-specific *Cyp4x1* gene. It possesses *in vitro* and *in vivo* insulator properties (Follows et al., 2012).
Tcrβ/ Trypsinogen	6	*Tcrβ*	*Prss2*	A retroviral LTR (LTR BglII family ERVK) acts as a chromatin barrier that prevents the spreading of silencing heterochromatin from the inactive trypsinogen genes into the active *Tcrβ* locus in thymocytes (Carabana et al., 2011).
5'Tyr	7	*Tyr*	*Grm5*	A CTCF-dependent insulator localizes upstream the LCR element that controls the expression of *Tyr*. It prevents from chromosomal position effects *in vivo* and is thought to protect *Tyr* from a highly condensed region further upstream (Giraldo et al., 2003a).
Bglobin 5'HS5	7	*Olfr66*	*Hbb-y*	Several DNaseI HS flank the mouse β-globin *locus* and isolate it from the adjacent odorant receptor genes. Site #5 at the 5' end and the only site at the 3' end, bind CTCF and convey enhancer-blocking (Farrell et al., 2002), but not barrier activity (Bulger et al., 2003). Both sites interact in erythroid cells, where the globins are expressed, but not in non-expressing tissue (Tolhuis et al., 2002). Despite initial beliefs, these sites are dispensable for the regulation of the genes in the *locus* (Bender et al., 2006).
Bglobin 3'HS1	7	*Hbb-b2*	*Olfr68*	
PCT12	7	*Mrpl23*	*Nctc1*	Between *Mrpl23* and *Nctc1,* there is a CTCF-dependent enhancer-blocking element that defines the 5' end of the *H19/Igf2* imprinted domain, which maps upstream of the first gene of the pair (Ishihara & Sasaki, 2002).

(Table continued)

Insulator	Chromosome	5' Gene	3' Gene	Properties
MS/DMD	7	*H19*	*Igf2*	Two DNaseI HS with two CTCF-binding sites each, map to the ICR that separates the imprinted *Igf2* and *H19* genes, whose expression is monoallelic and parent-of-origen-dependent. These sites exhibit methylation-dependent enhancer-blocking activity *in vitro* and *in vivo* (Bell & Felsenfeld, 2000; Hark et al., 2000; Kanduri et al., 2000). Moreover, CTCF-cohesin long-range interactions (Kurukuti et al., 2006), as well as other elements in *cis* (ICR size and CpG density), mediate in the establishment of proper imprinting and in the regulation of gene expression (Ideraabdullah et al., 2008; Ideraabdullah et al., 2011).
KvDMR1	7	*Kcnq1*	-	An ICR element located in an intron of *Kcnq1* regulates the imprinting of the well-characterized imprinted gene cluster of the region. The ICR binds CTCF and consists of, at least, an enhancer and an enhancer-blocker (Kanduri et al., 2002; Fitzpatrick et al., 2007)
SP-10	9	*A630095E13Rik*	*Acrv1*	The proximal promoter of *Acrv1* or *SP-10*, a testis-specific gene, conveys *in vitro* and *in vivo* enhancer-blocking properties. It has been proposed that TDP-43 binds to this region and tethers the DNA to the nuclear matrix, but only in tissues other than the testes (Reddi et al., 2003; Acharya et al., 2006; Abhyankar et al., 2007).
Rasgrf1 DMD	9	-	*Rasgrf1*	The imprinting of *Rasgrf1* is controlled by a DMD (differentially methylated domain) that maps upstream the gene. It also works as a CTCF-dependent enhancer-blocker (Yoon et al., 2005).
5'Wap	11	*Tbrg4*	*Wap*	The S/MARs that flank the *Wap* gene prevent the influence of the regulatory elements from the adjacent genes and restrict *Wap* expression to the mammary gland (Millot et al., 2003; Montazer-Torbati et al., 2008).
3'Wap	11	*Wap*	*Ramp3*	
Gh	11	*Cd79b*	*Scn4a*	A SINE B2 retrotransposable element resides between *Cd79b* and *Scn4a*. At a given developmental stage and in certain tissue types, the element gets transcribed from convergent RNA polymerases II and III promoters, establishing a barrier that permits the expression of the growth hormone gene upstream *Cd79b*. This element also conveys enhancer-blocking activity *in vitro* (Lunyak et al., 2007).
V(D)J rec	12	*Igh*	*hole*	V and D gene segments are separated by IGCR1 (InterGenic Control Region 1), a CTCF-dependent insulator with *in vitro* enhancer-blocking activity that plays an important role in V(D)J recombination in B lymphocytes precursors through the establishment of long-range interactions with other regions of the locus (Featherstone et al., 2010; Shih & Krangel, 2013). In addition, the complex regulation of the Igh locus does not affect the genes further downstream because of the presence of enhancer-blockers in the intergenic region. These elements bind CTCF in a developmentally regulated fashion (Garrett et al., 2005).

(Table continued)

Insulator	Chromosome	5' Gene	3' Gene	Properties
11P	17	*Rxrb*	*Col11a2*	While the expression of *Col11a2* is very much restricted, *Rxrb* is present in a variety of tissues. This is possible due to the presence of an insulator that blocks the cartilage-specific enhancer of *Col11a2,* impeding its interaction with the *Rxrb* promoter. The insulator does not bind CTCF *in vivo* (Murai et al., 2008)
Eif2s3x sites	X	*Eif2s3x*	*Klhl15*	CTCF binds in the intergenic region between *Eif2s3x* and *Klhl15*, establishing a barrier that permits the escape of *Eif2s3x* from X chromosome inactivation (Filippova et al., 2005).
Xist/Tsix	X	-	*Tsx*	CTCF binds to the 5' end of *Tsix* marking the X chromosome that needs to remain active. This region displays orientation-dependent enhancer-blocking activity *in vitro* (Chao et al., 2002).
RS14	X	*Tsx*	*Xist*	A CTCF-dependent enhancer-blocker separates the convergently expressed *Xist* and *Tsix* in the X chromosome. Malfunction of this element favors *Tsix* over *Xist* expression, impeding the inactivation of the chosen chromosome and leading to defects in cell differentiation and growth (Spencer et al., 2011).
Jarid1c sites	X	*Kiaa0522*	*Jarid1c*	*Jarid1c* escapes silencing and remains active in the condensed X chromosome in females, unlike the adjacent gene *Kiaa0522*. A CTCF-dependent boundary locates in the intergenic region and displays enhancer-blocking and barrier activities (Filippova et al., 2005)
B1X35S	-	-	-	B1X35S is a retrotransposable element from the SINE B1 family. It functions as a powerful enhancer-blocking element *in vitro* and *in vivo.* These activities depend on the binding of Ahr and Slug transcription factors, which, in turn, promote a switch from polymerase III- to polymerase II-dependent transcription of the retrotransposon (Roman et al., 2008; Roman et al., 2011).
CONSYN CTCF	-	-	-	A subset of all CTCF binding-sites in the murine genome is occupied in all cell types (CONstitutive) and in the same orthologous positions (SYNtenic) in the chicken and human genomes. These 247 CONSYN CTCF sites exhibit both *in vitro* and *in vivo* enhancer-blocking activities (Martin et al., 2011).
Hox clusters	-	-	-	Over 20 histone-depleted regions from all four *Hox* gene clusters function as enhancer-blockers *in vitro*. Most of these regions bind GAGA, which may be involved in the mechanism of insulation by recruiting chromatin remodeling complexes (Srivastava et al., 2013).

APPENDIX MM-1: Primers Used in this Work

Primers used to amplify, clone and test selected sequences for insulator activity. The name and sequence (5' to 3') of the primers are shown. Each pair of forward (F) and reverse (R) primers served to amplify the specified human (H) or mouse (M) element. The genomic coordinates of each element are also listed (genome assemblies GRCm37 and GRCh37 for mouse and human sequences, respectively). In some cases, the recognition motif of certain restriction enzymes (underlined) was included in the primer to facilitate cloning.

Name	Sequence (5' to 3')	Element (H/M)	Genomic Coordinates
1_235 short-F 1_235 short-R	ATACACTCGAGATGCATGATATGGCCC AGTGATGGTC ATACACTCGAGATGCATAGTCATGCCC ATACCACCTC	MIR 1short (H)	Chr1:23,683,272-23,683,501
1_235 long-F 1_235 long-R	ATACACTCGAGATGCATTGATTGGGAT AAACCCAGGA ATACACTCGAGATGCATTCCCATTGCA TGATCTGTTT	MIR 1long (H)	Chr1:23,682,851-23,683,934
2_979 short-F 2_979 short-R	ATACACTCGAGCTGCAGTGAACATAGG AGGGGAGGTG ATACACTCGAGCTGCAGAAGATGATCC ACCCTGCAAT	MIR 2short (H)	Chr2:98,633,063-98,633,436
2_979 long-F 2_979 long-R	ATACACTCGAGCTGCAGAGGAGCCAGT CACAGAAGGA ATACACTCGAGCTGCAGTGCTTTGAAA CCCTTTACGC	MIR 2long (H)	Chr2:98,632,677-98,633,820
4_130 short-F 4_130 short-R	ATACAGTCGACCTGCAGGGATATTCCC ACTGCAAACC ATACAGTCGACCTGCAGCGGGGAAGTT AGGAAAAAGG	MIR 4short (H)	Chr4:130,012,748-130,013,106
4_130 long-F 4_130 long-R	ATACAGTCGACCTGCAGTCATGCAAAC GCTACCATTC ATACAGTCGACCTGCAGGCAATACGGC TGGTTTGAGT	MIR 4long (H)	Chr4:130,012,383-130,013,468
11_822 short-F 11_822 short-R	ATACACTCGAGATGCATAACGGCAATA ACAGCTACCA ATACACTCGAGATGCATTAGGGAGTGG TTAGGCTCCA	MIR 11short (H)	Chr11:82,611,902-82,612,195
11_822 long-F 11_822 long-R	ATACACTCGAGATGCATCAGAAGCGCA CAGGCTAAG ATACACTCGAGATGCATAGTCTTTCTC CCCGACAGGT	MIR 11long (H)	Chr11:82,611,495-82,612,630
Test3-1_F Test3-1_y_3-3_R	ACTGGAAGGTGGCAGGGGGAA CCACCCTCCTGCCCTTCGGA	Cor-1 (M)	Chr7:133,609,489-133,611,497
Test3-2_F Test3-2_R	AGGCCTTCCCACAGAGGTGCT GTGAGGCCTGCTGCTGCCTC	Cor-1A (M)	Chr7:133,609,116-133,610,000
Test3-3_F Test3-1_y_3-3_R	GCCCATCCCAGGTCCTACTCCA CCACCCTCCTGCCCTTCGGA	Cor-1B (M)	Chr7:133,610,468-133,611,497
Test2-1_F Test2-1_R	GCTATGGAAGTGAGGCTGTGTCCT TCAGGTGCAGCAGCGAAGGC	Cor-2.1 (M)	Chr11:106,124,364-106,125,437

(Table continued)

Name	Sequence (5' to 3')	Element (H/M)	Genomic Coordinates
EBA_8-1_F EBA_9-1_R	ATGGAAGTGAGGCTGTGTCC GGGAGGTGGAATCTCCTGA	Cor-2.1A (M)	Chr11:106,124,367-106,124,696
EBA_8-3_F EBA_8-1_R	CCCTGCCCAAGCAGCTAT TTCTATGTGGGAGCCACTTTTT	Cor-2.1B (M)	Chr11:106,124,696-106,124,859
EBA_8-2_F EBA_8-3_R	GAAAAACCTGCTCCCCAAAT CTAGGGCCAGTGCCAGAG	Cor-2.1C (M)	Chr11:106,124,857-106,125,024
EBA_9-4_F EBA_8-2_R	CCTAGGGTGGAGCCTGTAAA GCTTTGTCGTCTTCTTCAGG	Cor-2.1D (M)	Chr11:106,125,020-106,125,452
EBA_8-1_F EBA_8-1_R	ATGGAAGTGAGGCTGTGTCC TTCTATGTGGGAGCCACTTTTT	Cor-2.1AB (M)	Chr11:106,124,367-106,124,859
EBA_8-3_F EBA_8-3_R	CCCTGCCCAAGCAGCTAT CTAGGGCCAGTGCCAGAG	Cor-2.1BC (M)	Chr11:106,124,696-106,125,024
EBA_8-2_F EBA_8-2_R	GAAAAACCTGCTCCCCAAAT GCTTTGTCGTCTTCTTCAGG	Cor-2.1CD (M)	Chr11:106,124,857-106,125,452
Test2-2_F Test2-2_R	ACGGCTGAGCCCTCCAGGAA GGACCAGGGCTACAGATTTGGAGG	Cor-2.2 (M)	Chr11:106,130,434-106,131,261
Test1-1_F Test1-1_R	CAGCGCGAACCCACTCACCA ACGGAGTCCACTGCCTGCCT	Cor-3 (M)	Chr11:106,116,881-106,118,200
EBA_12-2_F EBA_12-2_R	GTACGGCATGGTCAGGTGAG GAGAGGGCCAGAGAGATGTG	Dis-4.1 (M)	Chr7:133,903,318-133,904,242
EBA_12-3_F EBA_12-3_R	GGCTCTGTTGTCTGTGGTCA CTCTGACGAGGAGCTCAGGT	Dis-4.2 (M)	Chr7:133,904,474-133,906,483
EBA_12-1_F EBA_12-1_R	ATGAGCCTGCCTCTGTCTGT GCCTGCTCAGCCAGTGTAG	Dis-4.3 (M)	Chr7:133,909,330-133,909,905
EBA_12-4_F EBA_12-4_R	AACCAACTGTCGGCTACCAT TGTCGGATCAGACTTTTTCTCA	Dis-4.4 (M)	Chr7:133,914,533-133,914,705
10-5-F-Alt 10-5-R-Alt	TGTGTGTGTTGTTTCACAGGTC ACGGTCTCAGTCCTTTCTGC	Dis-5.1 (M)	Chr15:58,745,240-58,745,773
EBA_10-4_F EBA_10-4_R	CTGACTTCCAGGGACCAAAC CCCCATGTCAGCAGGAGTA	Dis-5.2 (M)	Chr15:58,755,549-58,757,245
EBA_10-3_F EBA_10-3_R	TGGCAACCCTGCTATTTAGG AAGCGGAATGAAGGGTTTTT	Dis-5.3 (M)	Chr15:58,762,386-58,763,675
EBA_10-1_F EBA_10-1_R	CTCCTCCCAGGGAAAAGC TAGCGTCCGGCTCACCTT	Dis-5.4 (M)	Chr15:58,765,270-58,765,393
EBA_10-2_F EBA_10-2_R	ACTTGGGGGACACAGTGAAG ATCAATGCATGCCAGATGAG	Dis-5.5 (M)	Chr15:58,765,561-58,767,764
EBA_11-3_F EBA_11-3_R	GGTCCTCTCCTCCCATTGAT GATATGGAGATTAGAACGGGTGT	Dis-6.1 (M)	Chr9:108,948,544-108,949,343

(Table continued)

Name	Sequence (5' to 3')	Element (H/M)	Genomic Coordinates
EBA_11-2_F EBA_11-2_R	TTTACTGCCTGGTTAGGCAAA GACTGGCAGATCACCCAAAT	Dis-6.2 (M)	Chr9:108,956,869-108,957,972
EBA_11-1_F EBA_11-1_R	AGGGCCACATGTGTTCACTA CTTCTCAGCCTGGGCCTCT	Dis-6.3 (M)	Chr9:108,959,731-108,960,430
EBA_11-4_F EBA_11-4_R	AGCAGCAGGAAGGTCTCAAG TCGTGTGCCTCTGTCAGACT	Dis-6.4 (M)	Chr9:108,961,468-108,962,748
EBA_11-4_F 11_4_A_R	AGCAGCAGGAAGGTCTCAAG AAGGAGGCAGCCCTTTTTAG	Dis-6.4A (M)	Chr9:108,961,468-108,962,056
11_4_B_F 11_4_C_R	TGATCCAGGCCTTCCTAAAA CTGAGCATCCAGCCACAAG	Dis-6.4B (M)	Chr9:108,962,023-108,962,350
11_4_D_F EBA_11-4_R	CCTCTGAGCCAAAAGGAAGA TCGTGTGCCTCTGTCAGACT	Dis-6.4C (M)	Chr9:108,962,420-108,962,748
11_4_B_F EBA_11-4_R	TGATCCAGGCCTTCCTAAAA TCGTGTGCCTCTGTCAGACT	Dis-6.4BC (M)	Chr9:108,962,023-108,962,748
EBA_7-1_F EBA_7-1_R	TCAGGGCATCTGTTATGCAC GCTAGCACTGCTGTTGTAATGC	CorDis-7.1 (M)	Chr17:74,611,278-74,611,849
EBA_7-2_F EBA_7-2_R	TCTGTGCATTACAACAGCAGTG CTGGAGGTTGGCCCAAGT	CorDis-7.2 (M)	Chr17:74,611,823-74,613,476
EBA_7-3_F EBA_7-3_R	GGGGACTTCGGTAAGCATCT TTTTCTCCTTAGAAATTCCGATTG	CorDis-7.3 (M)	Chr17:74,651,739-74,653,019
Test4-1_F Test4-1_R	CCCTGCGGAGGGGAACTGCTA AGAGGGCAGAATAGGGCACCAA	CorDis-8.1 (M)	Chr7:3,652,089-3,653,273
Test4-2_F Test4-2_R	AGTGTTCAGGTGAAGGCAGGGGA ACATGCACTACCACAGCCAGCT	CorDis-8.2 (M)	Chr7:3,653,368-3,654,664
Test4-3_F Test4-3_R	CAGGACTCATAGACCCCATTTCCCT TGGCGACCGGCATGTTGACC	CorDis-8.3 (M)	Chr7:3,654,580-3,655,966
EBA_5-3_F EBA_5-3_R	CTCTTCCACCCAGGTAAGTGA TGGCCTCACTGACACCTGTA	CorDis-9.1 (M)	Chr4:137,867,649-137,867,763
EBA_5-2_F EBA_5-2_R	TTCTGTTCAGGCAAATGGAA TAGTGCAAGGACCCACTCCT	CorDis-9.2 (M)	Chr4:137,873,992-137,875,012
EBA_5-1_F EBA_5-1_R	AGGAGTGGGTCCTTGCACTA ACAGGTGAGCACTCCTTTGC	CorDis-9.3 (M)	Chr4:137,874,993-137,875,907
EBA_6-1_F EBA_6-1_R	GGGGCAGCAGGATAACATAC TGACAAACCCCTAAAATGCTG	10.1 (M)	Chr12:85,001,913-85,002,937
EBA_6-2_F EBA_6-2_R	CAGCATTTTAGGGGTTTGTCA AGCAGAAGTTATGCCCACCA	10.2 (M)	Chr12:85,002,917-85,004,276
EBA_6-3_F EBA_6-3_R	GCCCGAGTGGGATACATTAG CAATCATTTACTTATATGGGAAGAAGG	10.3 (M)	Chr12:85,027,284-85,028,051

Primers used to determine the expression of the mouse genes coding for the desmosomal proteins (desmogleins and desmocollins) by quantitative PCR. The name and sequence (5' to 3') of each primer, together with the gene they amplify and the genomic corresponding coordinates of the amplicons (mouse assembly GRCm37), are listed. Note that quantitative PCRs are conducted using cDNA, which lacks any introns, as a template. However, the coordinates indicated here usually encompass, at least, one intron.

Name	Sequence (5' to 3')	Gene	Genomic Coordinates
qPCR_Dsc3_F Dsc3_R	TCACCACCGTCTCTCACTACA TCTGAATGTGGGTGCATTGT	*Dsc3*	Chr18:20,139,629-20,142,068
qPCR_ Dsc2_F Dsc2_R	GGCTCCTGGAGATGACAAAG AGTTCCCATGGTGGTGAAAG	*Dsc2*	Chr18:20,191,752-20,193,107
Dsc1_F qPCR_Dsc1_R	GCAACAACTGCTGATGGCTA CACTTGTCCCACTGAAGTTCC	*Dsc1*	Chr18:20,257,659-20,260,457
qPCR_Dsg1c_F Dsg1c_R	TGGAACAATACGAAAGGCTTA GGTTTCTCAGTGGATGTTGTG	*Dsg1c*	Chr18:20,433,754-20,435,485
qPCR_Dsg6_F qPCR_Dsg6_R	CAGGTCAAGCTACAAACAAG GTACCATGATGATTGTCCCTG	*Dsg1c* (Whittock, 2003)	Chr18:20,433,340-20,435,433
qPCR_Dsg1a_F qPCR_Dsg1a_R	GGCACTCGCCCTAACACTAA GGGTCTCAAGGGTTTCATCA	*Dsg1a*	Chr18:20,492,178-20,492,235
qPCR_Dsg1b_F qPCR_Dsg1b_R	CTTCAAGTGGTGGAGGCAGT CCGGTTCGTCTAAGGGATTT	*Dsg1b*	Chr18:20,557,990-20,558,055
Dsg4_F qPCR_Dsg4_R	GCACGATGTCCAATTTCCTT ATCTGAGCCCCTCACCAGTA	*Dsg4*	Chr18:20,610,387-20,611,478
qPCR_Dsg3_F Dsg3_R	AGGCATCCTGAAGGTGGTTA GATTCCCTCTCGGACATCAA	*Dsg3*	Chr18:20,683,812-20,686,249
qPCR_Dsg2_F qPCR_Dsg2_R	TTTCTGCTGACAGGCTATGC CATCTGTGGCGGTGATTTTC	*Dsg2*	Chr18:20,737,829-20,739,045
qPCR_Spink5_F Spink5_R	CCTGTCTGTGGTGATGATGG TTTCAGGTTGCTGCTTTCCT	*Spink5*	Chr18:44,176,091-44,178,418
gapdhF gapdhR	ATGTTTGTGATGGGTGTGAA ATGCCAAAGTTGTCATGGAT	*Gapdh*	Chr11:26,687,014-26,687,135

Primers used in the 3C assay. The name, sequence (5' to 3'), genomic coordinates (mouse assembly GRCm37) and genomic target or utility of each primer are shown.

Name	Sequence (5' to 3')	Target / Utility	Genomic Coordinates
3C_H_2,3Up_F	TGCAGAATTTCAAAGCATGG	-30 kb upstream CTCF-cohesin site #1	Chr18:20,082,305-20,082,324
3C_H_2,3_F	CCAGTTCCAGAGATGGATGG	CTCF-cohesin site #1	Chr18:20,112,673-20,112,692
3C_H_2,3Down_F	AAGGCTGTGATGAATAGATACCG	+30 kb downstream CTCF-cohesin site #1	Chr18:20,147,433-20,147,455
3C_H_4_F	GTTGAGCACACCAGAGATGC	CTCF-cohesin site #2	Chr18:20,636,432-20,636,451
3C_H_5_F	TCTTTCATCACCCACAGAGC	CTCF-cohesin site #3	Chr18:20,721,811-20,721,830
3C_H_5Down_F	GTCTGCTATGTGGGCAGAGG	+30 kb downstream CTCF-cohesin site #3	Chr18:20,762,632-20,762,651
3C_H_6_F	TGTGAGGTGCTTTAACATTGC	CTCF-cohesin site #4 (fixed primer)	Chr18:20,798,754-20,798,774
3C_H_7_F	AATTAGCACCATGCACTCTGG	CTCF-cohesin site #5	Chr18:44,002,920-44,002,940
3C_H_8_F	GGCTGCTGAACTTGGTTAGG	CTCF-cohesin site #6	Chr18:44,190,732-44,190,751
XPB1	TGACCTCCACACTCCTGAC	*Ercc3* locus (control primer #1)	Chr18:32,412,555-32,412,575
XPB2	ATGCGCAATTAGAAACTGC	*Ercc3* locus (control primer #2)	Chr18:32,420,893-32,420,913
2-3F 2-3R	AGGTGGAAATAGATGCCCCACCCCAACA CCTGAGTTGAGCATGCACCTTCTGTCTGC	PCR primers to check the presence of CTCF-cohesin site #1 on BAC clones	Chr18:20,096,124-20,096,343
4F 4R	CAGACTGTTTTGGAAAAGCCTTGGGTGT TGAATCCATGTCCTGAATGGAACATGTAAGA	PCR primers to check the presence of CTCF-cohesin site #2 on BAC clones	Chr18:20,633,490-20,633,817
5F 5R	GGCACCAAATACCTGCACCACTGAGCTG TCCTTTGTGCTCTTGAGTTTGGGGGAGTTG	PCR primers to check the presence of CTCF-cohesin site #3 on BAC clones	Chr18:20,721,396-20,721,722

(Table continued)

Name	Sequence (5' to 3')	Target / Utility	Genomic Coordinates
6-6downF 6-6downR	AAGGTTGGGGGTCACGGGGCACTAGAAT AGGGGTGGGCGATGACAGGGACATTAAG	PCR primers to check the presence of CTCF-cohesin site #4 on BAC clones	Chr18:20,799,519-20,799,681
8F 8R	CATCACTCCATGGTCTGCATCAGCTCCT CAAGGAAATTGGTGCCTGGGTGCTCAGT	PCR primers to check the presence of CTCF-cohesin site #5 on BAC clones	Chr18:43,999,760-44,000,079
10F 10R	GTAGATATGTGTGAAGAGGGAAAAGAGATCAG GGAAGGAGTCTCTGTGGGCAGGAAGGAA	PCR primers to check the presence of CTCF-cohesin site #6 on BAC clones	Chr18:44,196,265-44,196,453
Ercc3-1Fw Ercc3-1Rv	TGACTCATGTGTGTGAGAGAGGGGTGGTGA TCAGTAACACCTGGCTTGCCTGCCTTTCC	PCR primers to check the presence of the *Ercc3* gene on BAC clones	Chr18:32,412,298-32,412,448

Primers used for additional purposes. The name, sequence (5' to 3'), genomic coordinates (when applicable) and utility of each primer are shown.

Name	Sequence (5' to 3')	Purpose // Genomic Coordinates (Assembly)
gateway-p2	CAGTGTGCCGGTCTCCGTTATCG	Sequencing primer from the Gateway vector conversion system kit
GW1	GTTGCAACAAATTGATGAGCAATGC	Sequencing primer from the Gateway TA cloning kit
GW2	GTTGCAACAAATTGATGAGCAATTA	Sequencing primer from the Gateway TA cloning kit
R-CMV-IE	GCGGGGGTCGTTGGGCGGTC	Reverse primer to check cloning in pELuc-OUT by PCR
R-CMV-mP	AGGCCTCCCACCGTACACGCC	Reverse primer to check cloning in pELuc-IN by PCR
seq-48RCar	CTACACACACGGGTTAGACAGAGAT	Sequencing primer to check cloning in p48RCAR
5-2-seq3'-tyr5	CTGGGGCCTAGTTCATGTGT	Sequencing primer to check cloning in ptrTYR5
ex1-seq-tyr5	CAGCCAAGAACATTTTCTCCTT	Sequencing primer to check cloning in ptrTYR5
5-2-seq5'-tyr5	ATGGCAACTCTGCCTGAATG	Sequencing primer to check cloning in ptrTYR5
Tyr5A Tyr5B	GAGCCTTACTTGGAACAAGCC CTGCTCCCATTCATCAGTTCC	Primers for genotyping mice transgenic for the ptrTYR5-CorDis-9.2 by PCR
SV40-F SV40-R	ATCTAGTGATGATGAGGCTACTGCT TAATAGCAGACACTCTATGCCTGTG	Primers for the SV40 probe used to determine the integration sites of the ptrTYR5-CorDis-9.2 transgene by Southern Blot
MYA MYB	GGCGAATGGGTGAGTAACACG CGGATAACGCTTGCGACCTATG	Primers for testing the presence of mycoplasma contamination in cell culture by PCR
1F_ChIP_B52 2R_ChIP_B52	GCATCCTCTTTCCCAAATGA CACCCTGACCTTGGTTCCTA	ChIP primers to PCR-amplify CorDis-9.2 // Chr4:137,874,548-137,874,738 (mouse assembly GRCm37)
Exon18APP_F Exon18APP_R	TCAGCTCTCCTCTTGTTTTTCA ACAGCACAGCTGTCAAAAGG	ChIP primers to PCR-amplify the positive control (*APP* exon 18) // Chr21:27,253,841-27,254,105 (human assembly GRCh37)
CTCF-MUT-52-F CTCF-MUT-52-R	TCGCTGTAGTGCTGtaatagGGTGGCCCTT GTTCACCCCT GAACAAGGGCCACCctattaCAGCACTACA GCGAACTGCA	Site-directed mutagenesis of the CTCF site in CorDis-9.2

APPENDIX R-1: Anatomical Structures Considered in this Study

Anatomical Structures in aGEM			
accumbens nucleus	central nervous system	endothelium;cornea	hippocampus
adipose tissue	cerebellar cortex	enteric nervous system	hippocampus region*
adrenal gland	cerebellum	entorhinal cortex	hypoglossal XII nucleus
adrenal gland cortex	cerebellum ext. granule cell layer	epidermis	hypothalamus
adrenal gland medulla	cerebellum granule cell layer	epidermis stratum basale	ileum
adrenal gland zona glomerulosa	cerebellum int. granule cell layer	epididymis	incisor
alpha cell	cerebellum molecular layer	epithalamus	inferior colliculus
alveolar smooth muscle	cerebellum purkinje cell layer	epithelium;cornea	inferior olive
alveolus	cerebellum white matter	erythrocyte	inguinal lymph node
alveolus epithelium	cerebral cortex	esophagus	inner ear
ampullary gland	cerebral cortex layer	esophagus epithelium	inner nuclear layer
amygdala	cerebral white matter	esophagus smooth muscle	inner plexiform layer
anterior commissure	chondrocranium	extraocular skeletal muscle	inner segment;photoreceptor layer
anterior olfactory nucleus	choroid	eye	interalveolar septum
anterior thalamic group	choroid plexus	female germ cell	intestine
aorta	ciliary body	female;reproductive system	intestine epithelium
aorta tunica media	ciliary epithelium	femoral nerve	iris
apex of caecum	cingulate cortex	femur	jaw
arcuate nucleus	coagulating gland	fimbria hippocampus	jejunum
area postrema	coat hair	foot skin	kidney
arterial system	coat hair bulb	forebrain	kidney collecting duct
artery	cochlea	forelimb skeletal muscle	kidney cortex
ascending colon	cochlear duct	fourth ventricle	kidney interstitium
auricle	cochlear VIII nucleus	frontal cortex*	kidney medulla
basal ganglia	colon	fundus;stomach	kidney pelvis
basolateral amygdaloid nucleus	connective tissue	gall bladder	large intestine
bile duct	cornea	ganglion cell layer;retina layer	large intestine mucosa
blood	coronary artery	gastrocnemius;hindlimb skeletal muscle	larynx
blood vessel	corpus callosum	gastrointestinal system	larynx cartilage
blood vessel endothelium	corpus striatum	geniculate thalamic group	larynx epithelium
blood vessel smooth muscle	cranial ganglion	gigantocellular reticular nucleus	larynx submucosa gland
body	cranial nerve	globus pallidus	lateral geniculate nucleus
bone	cranium	glomerulus	lateral habenular nucleus
bone marrow	cumulus oophorus	gonad	lateral septal complex*
brain	decidua	granule cell layer;dentate gyrus	lateral septal nucleus
brain ependyma	deep cerebellar nucleus	granule cell layer;hippocampus	lateral ventricle
brain grey matter	dentate gyrus	gut	layer I;cerebral cortex layer
brain ventricle	dermis	habenula	layer II;cerebral cortex layer
brainstem	descending colon	hair follicle	layer III;cerebral cortex layer
brainstem nucleus	diaphragm	hair outer root sheath	layer IV;cerebral cortex layer
bronchiole	diencephalon	hand	layer V;cerebral cortex layer
bronchiole epithelium	digit*	harderian gland	layer VI;cerebral cortex layer
bronchus	dorsal raphe nucleus	head	lens
bronchus epithelium	dorsal root ganglion	heart	lens epithelium
brown fat;adipose tissue	dorsomedial hypothalamic nucleus	heart atrium	lens fiber
CA1 pyramidal cell layer	ductus arteriosus	heart left ventricle	limb
CA1;hippocampus region	duodenum	heart right atrium	limb long bone
CA2;hippocampus region	ear	heart valve	limb skeletal muscle
CA3;hippocampus region	ear skin	heart ventricle	liver
CA4;hippocampus region	endometrium	hilus;dentate gyrus	liver parenchyma
caecum	endometrium epithelium	hindbrain	locus coeruleus
cardiac muscle	endometrium glandular epithelium	hindlimb	lower jaw incisor
cartilage	endometrium luminal epithelium	hindlimb skeletal muscle	lower jaw molar
caudate-putamen	endometrium stroma	hippocampal formation*	lower leg
lung	ovary stratum granulosum	sclera	tail
lung blood vessel	ovary theca	sebaceous gland	tegmentum

(Table continued)

Anatomical Structures in aGEM			
lung connective tissue	oviduct	secondary oocyte	telencephalon
lung epithelium	pallidum*	seminal vesicle	temporal bone
lymph node*	pancreas*	seminiferous tubule	tendon
lymph node follicle	pancreatic acinus	seminiferous tubule epithelium	testis
lymphocyte	pancreatic duct	septal cortex	testis interstitial tissue
macula of utricle	pancreatic islet	septal olfactory organ	thalamus
male germ cell	parabrachial nucleus	septal organ of Gruneberg	third ventricle
male;reproductive system	parathyroid gland	skeletal muscle	thymus
mammary gland	paraventricular hypothalamic nucleus	skeleton	thymus cortex
mammillary body	parotid gland	skin	thymus medulla
mandible	pars anterior;adenohypophysis	small intestine*	thyroid gland
mast cell	pars intermedia;adenohypophysis	small intestine crypt of lieberkuhn	tibia
medial amygdaloid nucleus	pars posterior;neurohypophysis	small intestine epithelium	tibialis anterior
medial habenular nucleus	penis	small intestine villus	tongue
medial preoptic region	peyer's patch	smooth muscle	tongue epithelium
medulla oblongata*	peyer's patch germinal center	soleus	tooth
megakaryocyte	pharynx	solitary tract nucleus	tooth enamel organ
meninges	photoreceptor layer	spermatid	trachea
midbrain	pigmented retinal epithelium	spermatocyte	trachea cartilage
midbrain periaqueductal grey	piriform cortex	spermatogonium	trachea epithelium
modiolus	pituitary gland	spermatozoon	trachea smooth muscle
molar	placenta	spinal cord	trigeminal V ganglion
muscle	platelet	spinal cord central canal	trigeminal V nucleus
myelencephalon	pons*	spinal cord dorsal horn	trigeminal V spinal sensory nucleus
myenteric nerve plexus	pontine nucleus	spinal cord grey matter	unfertilized egg
myometrium	pretectal region	spinal cord ventral horn	urethra
nasal cavity epithelium	primary oocyte	spinal cord white matter	urinary bladder
neocortex	primary spermatocyte	spiral organ	urinary bladder lamina propria
nerve	prostate gland	spleen	urinary bladder mucosa
nervous system	prostate gland anterior lobe	spleen capsule	urinary bladder muscularis mucosa
neural retinal epithelium	prostate gland dorsolateral lobe	spleen red pulp	urinary bladder smooth muscle
nucleus of lateral olfactory tract	prostate gland epithelium	spleen trabeculum	urinary bladder urothelium
olfactory bulb*	prostate gland smooth muscle	spleen white pulp	uterine cervix
olfactory bulb ext. plexiform layer	prostate gland ventral lobe	stomach*	uterine cervix epithelium
olfactory bulb glomerular layer	pulmonary artery	stomach cardiac region	uterus
olfactory bulb granule cell layer	pyloric antrum	stomach epithelium	utricle
olfactory bulb int. plexiform layer	pyramidal cell layer;hippocampus layer	stomach glandular region mucosa	vagina
olfactory bulb mitral cell layer	raphe nucleus	stomach pyloric region	vagina smooth muscle
olfactory cortex*	renal cortex collecting duct	stratum lacunosum;hippocampus layer	vagus X ganglion
olfactory epithelium*	renal cortex tubule	subiculum	vas deferens
oocyte	renal papilla	sublingual gland	ventral tegmental area
optic nerve	renal proximal tubule	submandibular gland	ventral thalamus
oral region;gastrointestinal system	renal tubule	substantia nigra	ventricular zone;brain
outer nuclear layer	reticular thalamic nucleus	substantia nigra pars compacta	ventromedial hypothalamic nucleus
outer plexiform layer	retina	subventricular zone;brain	vestibular apparatus
outer segment;photoreceptor layer	retina layer	superior cervical ganglion	vibrissa follicle
ovary	retrohippocampal cortex*	superior colliculus	vomeronasal organ
ovary antral follicle	rib	superior olivary nucleus	white fat;adipose tissue
ovary follicle	salivary gland	suprachiasmatic nucleus	white matter
ovary primary follicle	sciatic nerve	sympathetic ganglion	zona incerta
ovary secondary follicle			

The tissues considered in the Euclidean Distance Method (highlighted in blue) correspond to 12% of all tissues stored in aGEM. They contain gene expression data for, at least, 60% of the genes except for those in chromosome Y (marked with *), in which the tissues “Brain” and “Muscle” are additionally included.

APPENDIX R-2: Top 50 Pairs of Genes that Potentially Contain Boundary Elements Commonly Obtained by Both Algorithms

Gene Name 1	Gene Name 2	Intergenic Distance (bp)	Promo-ters[1]	Genomic Coordinates (GRCm37)	ID (Distance Method)	ID (Correlation Method)
Zfp384	*Ing4*	1895	SO	Chr6:124,959,163-125,001,517	MBD02043	MB00003
Ing4	*Lpar5*	16421	SO	Chr6:124,989,778-125,032,490	MBD03506	MB00004
Cd200r1	*Cd200r2*	72119	SO	Chr16:44,765,849-44,915,953	MBD00175	MB00006
Memo1	*Dpy30*	4665	SO	Chr17:74,600,046-74,715,761	MBD00034	MB00008
Nt5dc1	*Tspyl4*	-121131	D	Chr10:34,008,094-34,021,114	MBD00927	MB00012
BC057079	*Ptplad2*	-14867	C	Chr4:87,740,533-88,084,832	MBD00068	MB00013
Srgap2	*Fam72a*	542	D	Chr1:133,181,828-133,436,449	MBD00114	MB00015
Rpl22	*Chd5*	4580	SO	Chr4:151,699,851-151,764,303	MBD00069	MB00021
Ric8	*Psmd13*	18238	SO	Chr7:148,042,856-148,084,541	MBD00776	MB00023
Thsd4	*Lrrc49*	46813	SO	Chr9:59,814,738-60,535,965	MBD00082	MB00026
Sgk3	*6030422M02Rik*	8790	SO	Chr1:9,788,234-9,932,166	MBD00076	MB00029
1600027N09Rik	*Ogfr*	2941	SO	Chr2:180,317,417-180,330,541	MBD00520	MB00038
9930021D14Rik	*Caskin1*	13314	SO	Chr17:24,610,829-24,645,850	MBD00344	MB00042
Snapin	*2500003M10Rik*	7923	SO	Chr3:90,291,948-90,313,420	MBD00324	MB00043
Tsen34	*Rps9*	2970	SO	Chr7:3,644,977-3,658,503	MBD01439	MB00045
Tcp11l2	*Polr3b*	8078	SO	Chr10:84,039,371-84,189,922	MBD00557	MB00050
Fam60a	*Dennd5b*	41604	SO	Chr6:148,869,557-149,050,202	MBD00318	MB00053
Ier3ip1	*Katnal2*	35543	C	Chr18:77,168,756-77,286,047	MBD00345	MB00055
Creld1	*Prrt3*	585	C	Chr6:113,433,291-113,451,925	MBD00815	MB00056
Oprd1	*Ythdf2*	40426	SO	Chr4:131,666,641-131,768,218	MBD00585	MB00058
Zfp366	*Ptcd2*	72613	C	Chr13:99,954,778-100,114,624	MBD00474	MB00059
Txndc9	*Rev1*	55441	SO	Chr1:38,040,712-38,186,507	MBD00737	MB00060
Acer2	*Slc24a2*	48302	C	Chr4:86,520,300-86,876,444	MBD00544	MB00070
Atrx	*Magt1*	38687	SO	ChrX:102,992,954-103,207,245	MBD00241	MB00071
Zcchc8	*Rsrc2*	7490	SO	Chr5:124,149,808-124,199,421	MBD00326	MB00072
Sh3bgrl	*Pou3f4*	1616549	SO	ChrX:106,290,704-108,012,545	MBD01338	MB00073
Frs2	*Yeats4*	66747	SO	Chr10:116,507,185-116,661,546	MBD01029	MB00074
Fbxw2	*Psmd5*	25777	SO	Chr2:34,660,034-34,726,459	MBD00147	MB00076
2310037I24Rik	*Ccnt1*	8954	SO	Chr15:98,349,249-98,401,354	MBD00829	MB00079
Pdxp	*Lgals1*	7225	SO	Chr15:78,744,387-78,760,622	MBD00976	MB00080
Rasa2	*Zbtb38*	51146	SO	Chr9:96,439,719-96,653,250	MBD00412	MB00082
Rpl5	*Fam69a*	-952	C	Chr5:108,329,521-108,416,104	MBD01573	MB00087
Grip2	*C130022K22Rik*	50803	D	Chr6:91,711,503-91,849,837	MBD00638	MB00088
Ablim2	*Afap1*	8346	SO	Chr5:36,100,529-36,346,572	MBD00550	MB00090
Map3k12	*Tarbp2*	1380	D	Chr15:102,328,080-102,354,107	MBD00558	MB00091
Fgfr2	*Etos1*	-2399660	SO	Chr7:137,305,965-137,915,825	MBD00784	MB00095
Usp48	*Rap1gap*	6189	SO	Chr4:137,149,667-137,285,782	MBD01782	MB00096
Ddx24	*Ifi27l1*	8364	D	Chr12:104,646,194-104,678,449	MBD00952	MB00099
Ggt7	*Gss*	44944	SO	Chr2:155,316,115-155,418,546	MBD01087	MB00100
Trappc4	*Rps25*	166	D	Chr9:44,211,844-44,218,272	MBD00669	MB00104
Golph3	*Pdzd2*	5789	C	Chr15:12,251,205-12,669,679	MBD00343	MB00107
Narf	*Wdr45l*	71368	C	Chr11:121,098,567-121,215,759	MBD00132	MB00108
Sec13	*Atp2b2*	4926	SO	Chr6:113,678,061-113,992,607	MBD00852	MB00109
Pfkfb1	*Tro*	1426	C	ChrX:147,022,772-147,092,126	MBD01107	MB00113
Clk2	*Scamp3*	552	SO	Chr3:88,968,717-88,986,687	MBD01165	MB00117
Rnf41	*Smarcc2*	19877	SO	Chr10:127,848,771-127,927,228	MBD00626	MB00119
Secisbp2	*Sema4d*	17202	C	Chr13:51,747,066-51,889,116	MBD00780	MB00120
Txndc16	*Ero1l*	62876	SO	Chr14:45,754,123-45,938,237	MBD00119	MB00125
BC017643	*Narf*	7931	D	Chr11:121,083,902-121,117,170	MBD00621	MB00126
Ttl	*Polr1b*	4712	SO	Chr2:128,891,678-128,952,330	MBD00319	MB00129

[1]**Promoters**. Configuration of the promoters of the pair: D, divergent; C, convergent; SO, same orientation.

Highlighted in **light blue**, two of the pairs selected for functional validation.