

**IMPACT OF COPY NUMBER VARIATION  
ON CHROMATIN INTERACTIONS AT THE MOUSE  
4E2 CHROMOSOME REGION**

**A Dissertation Presented to  
The Watson School of Biological Sciences  
at Cold Spring Harbor Laboratory**

**in Partial Fulfillment of the Requirements for  
the Degree of Doctor of Philosophy**

**by  
Cinthya Jeanette Zepeda Mendoza**

**Cold Spring Harbor Laboratory**

**September 2014**

## **Acknowledgements**

The completion of my thesis has been possible thanks to the support and help from multiple people, all of whom became great tutors, friends, colleagues, and family.

Above all, I would like to thank David Spector, my supervisor during the past 4 years of graduate studies. David showed me the importance of critical thinking, scientific etiquette, and above all, to pursue problems down to the root. He gave me the freedom to propose experiments and learn from their results, be those priceless data or total failures. His objectivity, perseverance, and constructive scientific thinking are now a solid part of my academic formation, and qualities I aspire to incorporate in my future scientific endeavors. I am honored to have been his student, and to have had his support in the proposal, development, and completion of this thesis work.

I also thank all the members of the Spector lab, past and present. Their useful suggestions and advice provided me with several ideas and a critical way to evaluate my work. I am especially indebted to Jingjing Li, for being the best lab manager one could ask for, and Melanie Eckersley-Maslin, for teaching me everything she knew during my rotation at the lab, for supporting me when experiments seemed to never work out, and for helping me prepare RNA-Seq libraries.

This project brought together collaborations with amazing scientists from around the world. Without their expertise, this research would have never been completed. In chronological order, I would like to first thank Alea Mills (CSHL, USA) for sharing her chromosomally engineered mouse models. Her knowledge of the system, her prompt re-establishment of first generation chimeras, and her availability for discussions made her a key element in the development of this research. I am particularly indebted to her student,

Dongwoo Hwang, for teaching me everything I needed to know about mouse embryonic fibroblasts (MEFs). Next, I want to thank the collaboration of Nathalie Harder, in the laboratory of Karl Rohr at the German Cancer Research Center and University of Heidelberg, GER. The analysis of 3D DNA FISH data would have been impossible without her image analysis skills, her understanding of the microscopy data, tailored scripting, and extensive troubleshooting, which yielded all the required information from thousands of 3D DNA FISH images. In continuing with the microscopy collaborations, I would like to thank Hesus Padilla-Nash in Thomas Ried's laboratory (NIH, USA) for performing the spectral karyotyping analysis of deletion and wild type MEF lines.

On the molecular side of experiments, my special thanks go to Erik Splinter, Elzo de Wit, and Wouter de Laat (Hubrecht Institute, NED). Wouter accepted me in his lab for two weeks to learn 4C and its multiple analysis methods and interpretations. Erik Splinter kindly guided me through the, literally, one whole week protocol, sharing in his tips and tricks for this exquisite yet powerful "C" technique, while Elzo performed some of the initial statistical analysis of my PE-4Cseq data. I want to thank Wouter and all of his lab members for making my visit to the Netherlands a wonderful experience.

The final 4C analysis pipeline developed for this project is the contribution of Swagatam Mukhopadhyay, a computational scientist in the group of Mike Wigler at CSHL, USA. Swagatam not only single handedly tackled a major problem in the field of C data analysis, he was also a great tutor in my endeavor of learning more about chromosome conformation analyses, and how to fairly analyze data without making biased assumptions. His collaboration was instrumental in the completion of this work, and I learnt much from his physics analysis point of view. I thank Swagatam not only for sharing his science, but for

sharing his love for tango, poetry, and his reverence for a good meal.

The second aspect of the molecular analysis was performed in collaboration with Emilie Wong and Paul Flicek (EMBL-EBI). Emilie performed the allele-specific analysis of RNA-Seq data, and it was thanks to her statistics and bioinformatics knowledge that I was able to discover multiple important aspects of the gene expression and chromatin architecture relationship.

I would also like to thank my thesis committee members: my thesis chair, Rob Martienssen, my academic mentor, Tom Gingeras, as well as Alea Mills and Thomas Ried. I specially thank Tom Gingeras for being a great academic mentor, from whom I learnt more about the way big scale data should be approached, the life in the genomics industry, and his advice for continuing in the competitive genomics field. I also thank Jim Lupski for agreeing to be my external examiner, and traveling all the way from Texas to New York.

As part of the CSHL faculty, I would like to thank Leemor Joshua-Tor for being, besides David, my second scientific role model. I also thank Michael Schatz for sharing in his computational knowledge and the introductory lessons to the world of sequencing.

An important part of my life in CSHL and New York are my friends Dongwoo Hwang, Katie Petsch, Raehum Paik, and Patty and Evan Creca. The support and help they provided me through the years, their encouragement, their friendship, and their presence cannot be repaid with anything. Having such wonderful friends shaped much of the person I am today, my vision for the future, and my confidence that, even apart, our bonds will keep growing.

There are two families to whom I am specially indebted, and thankful for all of their support, help, and love. The first is the Soref family, and especially Cathy Soref for being, in

short, my fairy godmother. Throughout these years in New York, Cathy has been a mother, a friend, a counselor, a doctor, among many other titles. She made me feel at home in a time where I missed my family in Mexico the most, and invited me to share in the happiness of her own loving home. There are no words or deeds that can thank Cathy and her family for her support and presence in my life at the lab. The other family I want to thank is the Watsons. Having had the opportunity to live in Ballybung, share in their history, and above all, getting to meet Rufus and know more about him and his life, has been a priceless experience, as a student and as a human being. I can only thank all the lessons learnt, all the new views of the world that scientific and daily life discussions with the Watsons provided me.

Last, but not least, is the gratitude to my family. My parents Blanca and Alberto have always supported me in all of my decisions, from going abroad to study college to pursuing my PhD in another country. I am fortunate to have been born their daughter. Lisandra, even though younger than me, has been both a teacher and a marvelous sister. I am fortunate that she shares in the passion for genomic research, which has allowed us to grow and learn from each other and the very different projects we specialize in. And not yet officially part of my family is my girlfriend Sara Pease, whose loving support and presence made me realize love has no gender, and that despite other people's judgment, the ones who love you will always stay by your side.

## Abstract

The three-dimensional organization of chromatin in eukaryotic cells provides a critical impact toward regulating gene expression and genomic stability. However, little is known about chromatin structure after the occurrence of DNA copy number variants (CNVs). An allele-specific chromosome conformation and gene expression characterization was performed in  $df/+^{Bl6}$  and wild type  $+^{129}/+^{Bl6}$  MEFs.  $df/+^{Bl6}$  is an engineered mouse strain with a 4.3Mb deletion in the 4E2 region, which is syntenic to human 1p36 where CNVs are highly frequent and associated with cancer and mental retardation phenotypes. A new quantitative framework for the analysis of PE-4Cseq data revealed that up to 22% of chromosome 4 sequences display changes in contact probabilities and chromatin compaction between the deletion ( $df$ ) and wild type ( $+^{129}$ ) chromosomes. 3D DNA FISH validations of selected regions showed strong agreement with PE-4Cseq results. RNA-Seq data showed a significant enrichment of differentially expressed (DE) genes contained within differentially interacting regions in  $df$ . A high correlation in DE between 129S5/SvEv<sup>Brd</sup> and C57Bl6/J alleles was discovered, suggesting they are coordinately regulated by a *trans* mechanism. Interestingly, up to ~33% of the  $df$  regions that showed interaction changes are shared with  $+^{Bl6}$ , the wild-type copy of chromosome 4 in  $df/+^{Bl6}$  MEFs. Although many of the  $df$  interaction changes could be explained by *trans* mechanisms affecting both chromosome 4 copies, there exist 659 regions (~23Mb) not shared with  $+^{Bl6}$ , pointing to possible direct effects of CNVs in the underlying chromosome architecture. The present analysis has expanded our understanding of how CNVs alter preferred conformation states of a wild type genomic region, with possible functional consequences which could aid in the study of human disease.

## List of Abbreviations

µm	Micrometer
3C	Chromosome Conformation Capture
3D	Three Dimensional
4C	Chromosome Conformation Capture-on-chip
5C	Carbon Copy Chromosome Conformation Capture
aCGH	Array Comparative Genomic Hybridization
AML	Acute Myeloid Leukemia
BAC	Bacterial Artificial Chromosome
BCP	Biased Contact Probability
bp	Base pairs
ChIA-PET	Chromatin Interaction Analysis by Paired-End Tag Sequencing
ChIP	Chromatin Immunoprecipitation
CML	Chronic Myelogenous Leukemia
CNVs	Copy Number Variants
CTCF	CCCTC-Binding Factor
DNA	Deoxyribonucleic Acid
DE	Differentially Expressed
EM	Electron Microscopy
ER-a	Oestrogen Receptor a
ES	Embryonic Stem cells
FISH	Fluorescence <i>in situ</i> hybridization
FRAP	Fluorescence Recovery After Photobleaching
GO	Gene Ontology
HIV	Human Immunodeficiency Virus
INM	Inner Nuclear Membrane
Kb	Kilobase
KEGG	Kyoto Encyclopedia of Genes and Genomes
LAD	Lamina-Associated Domain
Mb	Megabase

Med1	Mediator Complex Subunit 1
Med12	Mediator Complex Subunit 12
NBs	Nuclear Bodies
NL	Nuclear lamina
NPC	Nuclear Pore Complex, Neural Progenitor Cell
ONM	Outer Nuclear Membrane
P#	Passage # MEFs
P4	Passage 4 MEFs
PCR	Polymerase Chain Reaction
PE	Paired-End Sequencing
PE1	Paired-End Read 1
PE2	Paired-End Read 2
PE-4Cseq	Paired-End Sequencing Chromosome Conformation Capture
PS	Perinuclear Space
qPCR	Quantitative Real Time PCR
RER	Rough Endoplasmic Reticulum
RFLP	Restriction Fragment Length Polymorphism
RM	Repeat Masker
SD	Segmental Duplications
Seq	Sequencing
SER	Smooth Endoplasmic Reticulum
SINE	Short Interspersed Element
Smc1	Structural Maintenance of Chromosomes 1A
SNP	Single Nucleotide Polymorphism
TAD	Topologically Associating Domain



## Contents

Acknowledgements.....	2
Abstract.....	6
List of Abbreviations .....	7
Contents .....	9
List of Figures.....	15
List of Tables .....	18
Chapter 1: An introduction to eukaryotic genome structure.....	21
1.1 The nucleus .....	23
1.1.1 The nuclear envelope and nuclear pore complexes.....	24
1.1.2 The nuclear lamina .....	27
1.1.3 Nuclear bodies .....	29
1.2 Genome packaging.....	31
1.2.1 The 10 nm fiber .....	31
1.2.2. Higher-order structures of chromatin organization .....	34
1.2.2.1 The 30 nm fiber.....	34
1.2.2.2 Topologically associating domains.....	38
1.2.2.3 Chromosome territories .....	40
1.3 Chromosome conformation capture technologies.....	42

1.3.1 3C.....	42
1.3.2 4C.....	45
1.3.3 5C.....	51
1.3.4 Hi-C .....	54
1.3.5 ChIP-based 3C techniques.....	58
1.3.6 C-methodologies discussion.....	60
1.4 Copy number variation in mammalian genomes.....	62
1.5 Characterization of higher-order chromatin organization at the mouse region 4E2.....	66
Chapter 2: CNV mouse models of 4E2.....	71
2.1 Human region 1p36, CNVs, and their roles in disease. ....	71
2.2 Chromosomally engineered $df/+^{Bl6}$ and $dp/+^{Bl6}$ CNV mouse models .....	73
2.3 Genomic characteristics of mouse chromosome 4 and the 4E2 engineered region .....	80
2.4 129S5/SvEv <sup>Brd</sup> and C57Bl6/J chromosome 4 sequence analysis.....	86
2.5 Spectral karyotyping analysis of $df/+^{Bl6}$ and $+^{129}/+^{Bl6}$ MEFs.....	102
2.6 Discussion of CNV mouse models of 4E2.....	104
Chapter 3: Microscopic characterization of higher-order chromatin organization in 4E2 CNVs.....	108
3.1 3D DNA FISH of 4E2 and neighboring regions in $+^{129}/+^{Bl6}$ , $df/+^{Bl6}$ and $dp/+^{Bl6}$ MEFs .....	109

3.2 Development of a dedicated ImageJ plugin for automated analysis of 3D DNA FISH	113
3.3 ImageJ plugin results and validation	116
3.4 Assessing the reproducibility of results derived from 3D DNA FISH experiments...	120
3.5 ImageJ plugin results of 3D DNA FISH of 4E2 and neighboring regions in $+^{129}/+^{Bl6}$ , $df/+^{Bl6}$ and $dp/+^{Bl6}$ MEFs	122
3.6 Discussion of 3D DNA FISH results	132
Chapter 4: Molecular characterization of higher-order chromatin organization in a 4E2 deletion CNV	142
4.1. PE-4Cseq measurement of 4E2 chromatin contacts	142
4.2. PE-4Cseq data filtering and read mapping	146
4.3. Quantitative analysis of PE-4Cseq data	147
4.3.1. Bias correction and data normalization across PE-4Cseq multi-viewpoints	150
4.3.2 Identification of differentially interacting regions in the <i>df</i> chromosome	154
4.4. Changes in local chromatin compaction in the deletion chromosome	161
4.5. Validation of changes in <i>del</i> <sup>129</sup> chromatin interactions by 3D DNA FISH	162
4.6. Protein binding sites inside PE-4Cseq differentially contacting regions	166
4.7. PE-4Cseq results for the <i>del</i> <sup>Bl6</sup> chromosome	175
4.7.1. Contact probability changes for viewpoints surrounding the deletion coordinates	175
4.7.2. Contact probability changes for viewpoints inside deletion CNV coordinates	188

4.8 PE-4Cseq results summary and discussion .....	196
Chapter 5: Gene expression characterization of $df/+^{Bl6}$ and $+^{129}/+^{Bl6}$ MEFs .....	200
5.1 Combined and allele-specific RNA-Seq analysis of $df/+^{Bl6}$ and $+^{129}/+^{Bl6}$ MEFs.....	202
5.2 Enriched DE content within $del^{129}$ and $del^{Bl6}$ differentially interacting regions .....	217
5.3 Enriched DE content within $del^{Bl6}$ differentially interacting regions inside the CNV	232
5.4 DE $df/+^{Bl6}$ genes and Monosomy 1p36.....	237
5.5 Summary of RNA-Seq characterizations of $df/+^{Bl6}$ and $+^{129}/+^{Bl6}$ MEFs .....	242
Chapter 6: Conclusion and Perspectives .....	245
6.1. Summary .....	245
6.1. Discussion .....	247
6.2. Perspectives and future directions.....	255
Chapter 7: Experimental methods.....	259
7.1 Generation of F1 $+^{129}/+^{Bl6}$ and $df/+^{Bl6}$ Embryos.....	259
7.2 MEF Preparation and Cell Culture.....	260
7.3 3D DNA FISH.....	260
7.4 MEF Karyotyping .....	262
7.5 4C Template Preparation.....	263
7.6 PE-4CSeq Viewpoint Amplifications, Sequencing, and Reads Mapping.....	264
7.7 Polymer Physics Analysis of PE-4CSeq data .....	266
7.7.1 Model for bias correction .....	266

7.7.1 Comparison of bias-corrected capture data .....	270
7.8 Allele-specific RNA-Sequencing and Analysis .....	271
Chapter 8: Extended Materials and Methods .....	273
8.1 Protocols, Buffers, and Cell Culture media recipes .....	273
8.1.1 Mouse Tail DNA Isolation .....	273
8.1.2 <i>df</i> PCR genotyping.....	275
8.1.3 <i>dp</i> PCR genotyping.....	277
8.1.4 IACUC Standard Procedure – Mouse Embryonic Fibroblasts (MEFs) .....	279
8.1.5 PI Staining of fixed whole cells.....	284
8.1.6 MEF medium.....	286
8.1.7 Trypsin.....	286
8.1.8 HEPES-buffered saline.....	286
8.1.9 Phosphate Buffered Saline (PBS).....	287
8.1.10 MEF Culture and Splitting .....	288
8.1.11 Nick Translation Protocol.....	289
8.1.12 3D DNA FISH.....	291
8.1.13 RNA isolation.....	294
8.1.14 RNA Sequencing Library Preparation.....	295
8.1.15 cDNA synthesis .....	309
8.1.16 Quantitative RT-PCR .....	309

8.1.17 PE-4CSeq Protocol.....	310
8.1.18 ATP, 100mM solution.....	321
8.1.19 10x Ligation buffer.....	321
8.1.20 Nuclei Isolation.....	322
8.1.21 Ampure XP Protocol 0.9x: .....	324
8.2 Computational Methods.....	325
8.2.1 3D DNA FISH analysis by Correct_and_Measure_3D.class ImageJ plugin .....	325
8.2.2 Custom R, Bash, and Perl scripts for the analysis of Correct_and_Measure_3D.class ImageJ plugin .....	327
8.2.3 PE-4Cseq reads analysis pipeline.....	329
8.2.4 Monte Carlo Simulations for CTCF, Smc1, Med1, and Med12 data.....	330
References.....	332
Author contributions.....	371

## List of Figures

Figure 1.1 Schematic diagram of an animal eukaryotic cell and its different structures .....	26
Figure 1.2 Schematic depictions of chromatin structures .....	37
Figure 1.3 Overview of 3C protocol steps .....	46
Figure 1.4 Outline of the 4C approach.....	48
Figure 1.5 Overview of the 5C methodology .....	53
Figure 1.6 Overview of the Hi-C technique.....	55
Figure 1.7 Schematic of ChIP-based 3C techniques.....	59
Figure 1.8 Selected examples of the different ways in which CNVs can affect chromatin organization and gene expression .....	68
Figure 2.1 Chromosomally engineered <i>df</i> and <i>dp</i> chromosomes.....	76
Figure 2.2 Representative + <sup>129</sup> / <sub>+<sup>Bl6</sup></sub> , <i>df</i> / <sub>+<sup>Bl6</sup></sub> , and <i>dp</i> / <sub>+<sup>Bl6</sup></sub> 13.5 days embryos .....	79
Figure 2.3 Bright-field microscope images of the different MEF genotypes .....	83
Figure 2.4 Confluent MEF FACs profiles.....	85
Figure 2.5 Circular depiction of mouse chromosome 4.....	87
Figure 2.6 Ensembl view of mouse chromosome 4 synteny to human chrs 1,6,8,9.....	89
Figure 2.7 Contig size distributions for chromosome 4 sequence of 129S5/SvEv <sup>Brd</sup> .....	96
Figure 2.8 Mummerplot of nucmer aligned 129S5/SvEv <sup>Brd</sup> assembly chromosome 4 .....	98
Figure 2.9 mummerplot of nucmer aligned 129S5/SvEv <sup>Brd</sup> 4E2 to repeat masked reference C57Bl6/J 4E2 .....	100
Figure 2.10 SNP locations inside the 4E2 region .....	101
Figure 2.11 Abnormal karyotype for + <sup>129</sup> / <sub>+<sup>Bl6</sup></sub> (129S5E117) MEF cell 24 as revealed by SKY.	

.....	105
Figure 2.12 Abnormal karyotype for $df/+^{Bl6}$ (129S5E71) MEF cell 11 as revealed by SKY106	
Figure 3.1 3D DNA FISH experiments and analysis .....	111
Figure 3.2 Overview of 3D DNA FISH analysis workflow .....	117
Figure 3.3 An example of 3D DNA FISH segmentation results.....	119
Figure 3.4 Comparison of plugin vs manual distances of chromatin compaction.....	121
Figure 3.5 Chromatin compaction differences between probes bordering the deletion CNV in $df$ chromosomes .....	131
Figure 3.6 Chromatin compaction distributions of BAC sets in control regions.....	134
Figure 3.7 Chromatin compaction distributions of BAC sets 4 and 7 .....	136
Figure 3.8 Nuclei volume differences between the analyzed MEFs.....	138
Figure 4.1 Allelic assignments of chromatin interactions by PE-4Cseq.....	144
Figure 4.2 Circular depiction of mouse of region 147-155.6Mb from chromosome 4 .....	145
Figure 4.3 Raw mapped reads for the $df/+^{Bl6}$ (129S5E71) and $+^{129}/+^{Bl6}$ (129S5E117) first biological PE-4Cseq replicates. ....	149
Figure 4.4 Genuine and physical chromatin contact changes.....	152
Figure 4.5 Bias-correction for $+^{Bl6}$ chromosome from $+^{129}/+^{Bl6}$ for all viewpoints denoted by viewpoint index in $x$ and $y$ axis.....	156
Figure 4.6 Contact probability profiles for the $del^{129}$ and $wt^{129}$ in chromosome 4 .....	159
Figure 4.7 Calculated $v$ per viewpoint for $del^{129}$ vs. the average of $wt^{Bl6}$ , $wt^{129}$ and $del^{Bl6}$ .	164
Figure 4.8 $del^{129}$ 3D DNA FISH validations .....	170
Figure 4.9 Contact probability profiles for the $del^{Bl6}$ and $wt^{Bl6}$ for chromosome 4 .....	180
Figure 4.10 Calculated $v$ per viewpoint for $del^{Bl6}$ vs. $wt^{Bl6}$ .....	181



Figure 4.11 Summary of  $del^{129}$  and  $del^{Bl6}$ , as well as unique and overlapping  $del^{129}$  and  $del^{Bl6}$  differentially interacting regions are shown for region 147-155.6Mb of mouse chromosome 4 ..... 187

Figure 4.12 Comparison of contact probability profiles for the  $del^{Bl6}$  and  $wt^{Bl6}$  for chromosome 4 sequence ..... 191

Figure 5.1 Chromosome 4 depictions of DE genes ..... 210

Figure 5.2 High degree of correlation between log2FoldChange DE values between 129S5/SvEv<sup>Brd</sup> and C57Bl6/J alleles in  $df/+^{Bl6}$  and  $+^{129}/+^{Bl6}$  MEFs .....211

Figure 5.3 qPCR validations for expression values for 9 genes inside the deletion CNV ... 216

Figure 5.4 Rainbow plot of the differential signal surrounding the deletion region in  $del^{129/225}$

Figure 5.5 Graph of CTCF, Med1, Med12, Smc1 protein binding sites along 147-155.Mb of chromosome 4 ..... 227

Figure 5.6 Graph of ECE1 mRNA levels in Monosomy 1p36 derived cell lines and normal karyotypic controls ..... 241

## List of Tables

Table 2.1 Breeding history for MEF generation. ....	78
Table 2.2 Present RepeatMasker classes in mouse chromosome 4 .....	88
Table 2.3 Annotated RefSeq genes inside the 4.3Mb engineered region.....	92
Table 3.1 BACs used as probes for the 3D DNA FISH experiments and their corresponding chromosomal location.....	112
Table 3.2 Descriptive statistics of compaction measurements between 3D DNA FISH of two biological replicates .....	126
Table 3.3 Descriptive statistics of compaction measurements between 3D DNA FISH of 3 different MEF passages.....	128
Table 3.4 Summary of total cells included in the present 3D DNA FISH analysis per genotype and BAC set. ....	129
Table 3.5 Heterochromatin overlap ratios per channel per BAC set and genotype .....	139
Table 4.1 Summary of median magnitude of change, direction, and number of <i>del</i> <sup>129</sup> differentially interacting regions for viewpoints 1, 2, 11, and 12. ....	160
Table 4.2 BACS used for selected PE-4Cseq and chromatin decompaction regions .....	165
Table 4.3 Summary of <i>del</i> <sup>129</sup> differentially interacting regions overlap with CTCF, Mediator, and cohesin binding sites. ....	173
Table 4.4 Summary of Monte Carlo simulations for assessing statistical significance of protein binding overlaps for <i>del</i> <sup>129</sup> differentially interacting regions.....	174
Table 4.5 Summary of median magnitude of change, direction, and number of <i>del</i> <sup>Bl6</sup> differentially interacting regions for viewpoints 1-12 .....	177
Table 4.6 Summary of <i>del</i> <sup>Bl6</sup> differentially interacting regions overlap for viewpoints 1, 2, 11,	

and 12 with CTCF, Mediator, and Smc1 binding sites .....	182
Table 4.7 Summary of Monte Carlo simulations for assessing statistical significance of protein binding overlaps for <i>del<sup>Bl6</sup></i> differentially interacting regions for viewpoints 1, 2, 11, and 12.....	183
Table 4.8 Summary of unique and overlapping <i>del<sup>l29</sup></i> and <i>del<sup>Bl6</sup></i> differentially interacting regions for viewpoints 1, 2, 11, and 12.....	184
Table 4.9 Summary of Monte Carlo simulations for assessing statistical significance of protein binding overlaps for unique and overlapping <i>del<sup>l29</sup></i> and <i>del<sup>Bl6</sup></i> differentially interacting regions for viewpoints 1, 2, 11, and 12.....	185
Table 4.10 Summary of <i>del<sup>Bl6</sup></i> differentially interacting regions overlap for viewpoints 3-10 with CTCF, Mediator, and Smc1 binding sites .....	193
Table 4.11 Summary of Monte Carlo simulations for assessing statistical significance of protein binding overlaps for <i>del<sup>Bl6</sup></i> differentially interacting regions for viewpoints 3-10 ...	195
Table 5.1 RNA-Seq mapping stats per sample .....	204
Table 5.2 DE summary for <i>df/+<sup>Bl6</sup></i> and <i>+<sup>l29</sup>/+<sup>Bl6</sup></i> MEF RNA-Seq data. ....	207
Table 5.3 Selected genes for RNA-Seq validations with their corresponding gene expression .....	213
Table 5.4 <i>del<sup>l29</sup></i> differentially interacting regions overlap with DE 129S5/SvEv <sup>Brd</sup> alleles, combined genes, and total annotated genes in chromosome 4 .....	219
Table 5.5 MC simulations to assess the significance of <i>del<sup>l29</sup></i> and DE and total annotated genes overlap .....	220
Table 5.6 <i>del<sup>Bl6</sup></i> differentially interacting regions overlap with DE C57Bl6/J alleles, combined genes, and total annotated genes in chromosome 4 .....	221

Table 5.7 MC simulations to assess the significance of <i>del</i> <sup>B16</sup> and DE/total annotated genes overlap.....	222
Table 5.8 Unique and shared <i>del</i> <sup>I29</sup> and <i>del</i> <sup>B16</sup> differentially interacting regions overlap for viewpoints 1, 2, 11, and 12 with DE C57Bl6/J alleles, DE 129S5/SvEv <sup>Brd</sup> alleles, combined genes, and total annotated genes in chromosome 4 .....	223
Table 5.9 MC simulations to assess the significance of unique and shared <i>del</i> <sup>I29</sup> and <i>del</i> <sup>B16</sup> differentially interacting regions for viewpoints 1, 2, 11, and 12 and DE/total annotated genes overlap.....	224
Table 5.10 <i>del</i> <sup>I29</sup> and <i>del</i> <sup>B16</sup> , as well as unique and shared <i>del</i> <sup>I29</sup> and <i>del</i> <sup>B16</sup> differentially interacting regions overlap with H3K27ac and H3K4me1 marks .....	229
Table 5.11 MC simulations to assess the significance of <i>del</i> <sup>I29</sup> and <i>del</i> <sup>B16</sup> , as well as unique and shared <i>del</i> <sup>I29</sup> and <i>del</i> <sup>B16</sup> differentially interacting regions for viewpoints 1, 2, 11, and 12 with H3K27ac and H3K4me1 marks genes overlap.....	231
Table 5.12 Normalized reads counts for <i>df/+<sup>B16</sup></i> and <i>+<sup>I29</sup>/+<sup>B16</sup></i> MEFs and their associated <i>df/+<sup>B16</sup></i> over <i>+<sup>I29</sup>/+<sup>B16</sup></i> ratios.....	233
Table 5.13 <i>del</i> <sup>B16</sup> differentially interacting regions overlap for viewpoints 3-10 with DE C57Bl6/J alleles, DE combined genes, and total annotated genes in chromosome 4.....	234
Table 5.14 MC simulations to assess the significance of <i>del</i> <sup>B16</sup> differentially interacting regions for viewpoints 3-10 and their overlaps with DE C57Bl6/J alleles, DE combined genes, and total annotated genes in chromosome 4 .....	236
Table 5.15 Candidate genes associated with different Monosomy 1p36 phenotypes.....	239

## **Chapter 1: An introduction to eukaryotic genome structure**

In the middle of the nineteenth century, Walther Flemming coined the term “chromatin” for naming the stainable material inside eukaryotic nuclei (Flemming, 1877, 1878). What Flemming had observed became later known as “chromosomes.” By careful analysis of their behavior during mitosis and meiosis, Theodor Boveri and Walter Sutton proposed chromosomes to be the carriers of genetic information (Sutton, 1902; Boveri, 1903), the inheritance units whose existence Gregor Mendel and Charles Darwin had previously hypothesized. However, it was not until 1911 that Thomas Hunt Morgan directly linked chromosomal behavior to genetic inheritance, therefore establishing the chromosome theory of heredity (Morgan *et al.*, 1915).

For decades, biological research focused on the study of chromosomal structure and cell cycle dynamics in order to understand how chromosomes express and transmit genetic traits. In 1944, the emphasis on the cytological study of chromosomes changed with the discovery of deoxyribonucleic acid (DNA) as the molecular basis of inheritance, as shown by pioneering experiments of bacterial transformation (Avery, MacLeod, and McCarty, 1944). Through the years, DNA research achieved multiple breakthroughs, including the discovery of its double helical structure (Franklin and Gosling, 1953; Wilkins, Stokes, and Wilson, 1953; Watson and Crick, 1953), the elucidation of its semi-conservative replication mechanism (Meselson and Stahl, 1958), the demonstration of its function as a template for mRNA production (Brenner, Jacob, and Meselson, 1961), and the development of cloning protocols for the characterization of gene function (Jackson *et al.*, 1972; Cohen *et al.*, 1973). A major breakthrough in DNA research was the publication of the first draft of the human

genome sequence (International Human Genome Sequencing Consortium, 2001), opening the doors to high-throughput genomic research.

With the great advances in DNA studies over the past half century, it was realized that the biological complexity of organisms is not dependent on their linear genomic sequences. Developmental control depends on various layers of functional interplay, including epigenetic mechanisms (reviewed in Sasaki and Matsui, 2008), and the spatial organization of regulatory elements scattered across the genome (reviewed in de Laat and Duboule, 2013). The most well-known example of the latter is constituted by critical enhancer looping interactions for the correct activation of genes (Tolhuis *et al.*, 2002; Carter *et al.*, 2002; Murrell *et al.*, 2004; Lanzuolo *et al.*, 2007; Sanyal *et al.*, 2012; Shi *et al.*, 2013).

With the ever increasing need to understand the relationship between chromatin and gene expression for cell functionality, scientists have now returned their attention to a more in-depth structural and physical study of chromosomes inside the nucleus. By the improvement of diverse microscopy approaches (reviewed in Huang *et al.*, 2009), and the development of the chromosome conformation capture technology (3C, Dekker *et al.*, 2002), it has been shown that chromatin has different levels of organization at different length scales, ranging from the typical 10nm chromatin fiber, to the newly identified topologically associated domains (Dixon *et al.*, 2012; Hou *et al.*, 2012; Nora *et al.*, 2012; Sexton *et al.*, 2012), and the fractal globule organization of the genome (Lieberman-Aiden *et al.*, 2009).

In spite of the great advances in our current understanding of chromatin organization inside eukaryotic nuclei, many basic aspects of such structures are still poorly understood. One such question is related to the spatial alteration of chromatin organization upon the occurrence of DNA copy number variation. Genomic copy number variants (CNVs), are

defined as gains (insertions, duplications) or losses (deletions, null genotypes) of at least 1 kilobase (Kb) in size relative to a designated reference genomic sequence (Redon *et al.*, 2006). CNVs are widely observed in mammals and many other organisms, and several of them have been found to influence phenotypic variation and cause disease (reviewed in Weischenfeldt *et al.*, 2013). In fact, much of the current human genomic research is focused on unraveling the associations of CNVs with different disease phenotypes, playing important roles for clinical diagnosis and potential treatment of such conditions.

In an effort to contribute to the CNV and chromatin organization fields, this thesis describes the use of microscopic and molecular techniques for the investigation of the folded structure of mouse chromosome region 4E2, in its wild-type state and after the occurrence of a 4.3 megabases (Mb) DNA deletion or duplication. First, I will walk the reader through an overview of mammalian genome organization, emphasizing the nuclear environment, chromosomal packaging, and its influence on transcription, recombination, and chromosomal stability. I will subsequently expand on the widespread nature of copy number variation, its functional associations, and its importance in clinical genetics. Finally, I will describe my thesis project, which focuses on the analysis of a specific CNV in mouse and its impact on chromatin organization and gene expression.

## **1.1 The nucleus**

Eukaryotic cell nuclei have an average diameter of 10-15 micrometers ( $\mu\text{m}$ ), in which ~2 meters of linear DNA are packaged with proteins that serve in its folding, and for carrying out molecular processes such as DNA replication and transcription. The main function of the

nucleus is to maintain genomic integrity and provide the necessary components for the correct regulation of gene expression, thus constituting the control center of eukaryotic cells [Fig. 1.1A]. Diverse compartments exist in the nucleus, such as the nuclear envelope, nuclear pore complexes, the nuclear lamina, chromosome territories, and a diverse array of nuclear bodies (reviewed in Spector, 1993; Lamond and Earnshaw, 1998; Mao *et al.*, 2011), each having specialized functional tasks.

### **1.1.1 The nuclear envelope and nuclear pore complexes**

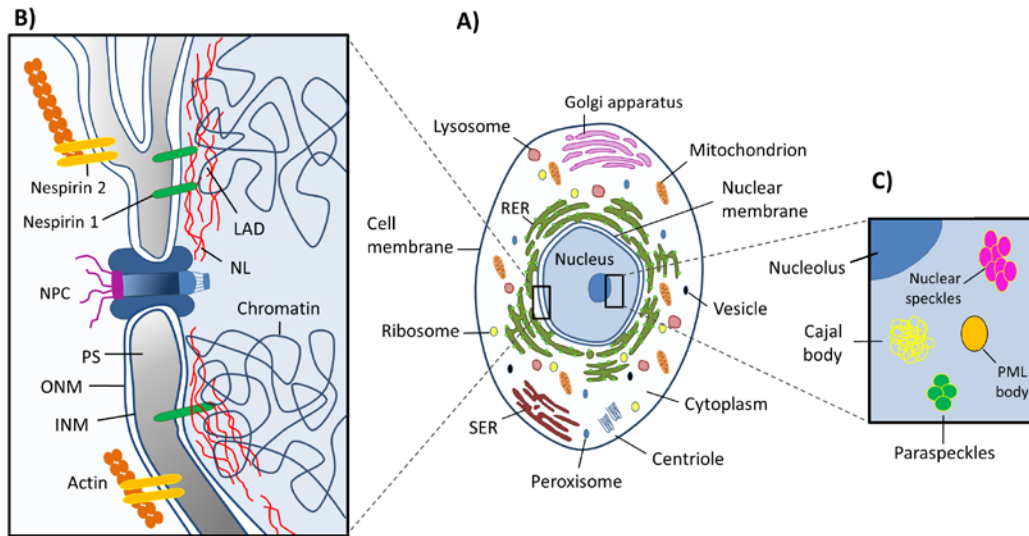
The nucleus is spatially separated from the cytoplasm by two lipid bilayers, the inner nuclear membrane (INM) and outer nuclear membrane (ONM), collectively known as the nuclear envelope [Fig. 1.1B]. The nuclear envelope is perforated by nuclear pore complexes (NPC) (Callan and Tomlin, 1950), highly evolutionary conserved structures shown to regulate nucleocytoplasmic transport (reviewed in Wentz and Rout, 2010), and implicated in genome organization and transcriptional regulation.

Over fifty years ago, NPCs were found to establish connections with chromatin (Engelhardt and Pusa, 1972). With the development of electron microscopy (EM), the nuclear envelope of interphase nuclei was shown to be tightly associated with heterochromatin, while NPCs seemed to be surrounded by decondensed chromatin (Swift, 1959; Watson, 1959; Davies, Murray, and Walmsley, 1974). Very interestingly, in *Saccharomyces cerevisiae* the Nup2p receptor of the nuclear pore complex is involved in the establishment of non-silenced chromatin boundaries (Ishii *et al.*, 2002). In HeLa cells, maintenance of NPC-heterochromatin free regions was shown to involve Translocated Promoter Region protein (TPR), an NPC-associated protein (Krull *et al.*, 2010). Moreover,



the orthologue of TPR in flies binds genomic regions enriched in chromatin marks associated with active transcription (Vaquerizas *et al.*, 2010), therefore arguing for specific roles of NPCs in chromatin organization.

Because decondensed chromatin is associated with active gene transcription, NPCs and their associated proteins could indirectly modulate activation of gene expression. In fact, NPCs have been found to participate in transcriptional activation in several cases, including the association of active genes in yeast with NPCs for robust transcription (Cabal *et al.*, 2006; Dieppois *et al.*, 2006; Taddei *et al.*, 2006; Luthra *et al.*, 2007; Light *et al.*, 2010), the tethering of yeast genes to NPCs by specific “DNA-zip codes” for full transcriptional activation (Ahmed *et al.*, 2010), and the NPC- dosage compensation complex interactions in *Drosophila* for the twofold increase in gene expression of the male X chromosome (Mendjan *et al.*, 2006). Nevertheless, examples of NPC roles in transcriptional silencing have also been reported, such as the silencing of mating-type loci and telomeres in yeast (Stavenhagen and Zakian 1994; Thompson *et al.*, 1994; Maillet *et al.*, 1996; Marcand *et al.*, 1996; Andrulis *et al.*, 1998; Feuerbach *et al.*, 2002), and the discovered associations of NPCs and enriched repressive marks in mammalian cells (Brown *et al.*, 2008). The aforementioned examples suggest that cells within different species have developed specialized functions for the NPCs, taking advantage of their ubiquitous composition and important structural role as a nuclear component (for an extensive review of NPCs and additional functional roles, see Raices and D'Angelo, 2012). Much of the molecular mechanisms guiding these processes are unknown, but are being studied to decipher commonalities between the architectural arrangement of the nucleus and its relationship to diverse gene expression mechanisms.



**Figure 1.1 Schematic diagram of an animal eukaryotic cell and its different structures**

**A)** Different compartments inside animal cells are depicted. RER, rough endoplasmic reticulum. SER, smooth endoplasmic reticulum. **B)** Zoom in view of the nuclear envelope and the nuclear lamina. NPC, nuclear pore complex. PS, perinuclear space. ONM, outer nuclear membrane. INM, inner nuclear membrane. LAD, lamina-associated domain. NL, nuclear lamina. **C)** Zoom in view of major nuclear bodies inside the nucleus.

### 1.1.2 The nuclear lamina

In mammalian cells, a network of intermediate filament proteins named lamins exists between the INM and chromatin, and has been found to connect NPCs to each other (reviewed in Gruenbaum *et al.*, 2005) [Fig. 1.1B]. This structure, better known as the nuclear lamina (NL), is made of lamin polymers and lamin-binding associated proteins, and supports a broad range of biological functions such as nuclear architecture, chromatin organization, and gene expression (reviewed in Goldman *et al.*, 2002).

Lamins are evolutionarily conserved proteins essential for cell viability, and expressed through development (reviewed in Zuela, Bar, and Gruenbaum, 2012). At its basic level, the NL helps maintain the geometry of the nucleus, and proper assembly and disassembly of this structure is required for cell cycle progression (reviewed in Gant and Wilson, 1997; McKeon, 1991). Mutations in genes encoding lamin or other NL component genes cause a wide-range of human diseases, such as muscular dystrophies and laminopathies, found to cause premature aging (reviewed in Worman, 2012), and/or affect several organs like muscle, bone, skin, and the peripheral nervous system. Such mutations highlight the importance of the NL in maintaining proper cell physiology and function.

In the majority of analyzed vertebrate cells to date, condensed heterochromatin and late-replicating DNA are generally located toward the nuclear periphery (Rae and Franke, 1972; Fox *et al.*, 1991; Kill *et al.*, 1991; Ferreira *et al.*, 1997). Very recently, thanks to the development of genome-wide mapping techniques, it has been possible to assess the molecular interactions between chromatin and the NL. DamID is a genome-wide application that fuses NL proteins to a DNA adenine methyltransferase (Dam) protein from *Escherichia coli* (van Steensel and Henikoff, 2000; Greil *et al.*, 2006; Vogel *et al.*, 2007). When the

chimeric fusion is expressed in cells, any piece of DNA that is in molecular contact with the NL *in vivo* will be methylated by the tethered Dam, and their identities determined as adenine methylation does not occur endogenously in most eukaryotes.

Through the use of DamID, chromatin-NL interaction maps have been generated for fly, mouse, and human cells (Pickersgill *et al.*, 2006; Guelen *et al.*, 2008; Peric-Hupkes *et al.*, 2010). It was revealed that very large (median size of 500Kb) chromosomal domains engage in interactions with the NL, with mouse and human cells possessing over a thousand of such lamina-associated domains (LADs). Very interestingly, LAD borders seem to be demarcated by sequence-embedded features like CTCF binding sites, CpG islands, and promoters oriented away from LADs. In all three species, LADs were typified by low levels of gene expression, and the lack of active histone marks and RNA polII, indicating that LADs represent a repressive chromatin environment, consistent with the microscopic observations of its association with heterochromatin.

LAD structures can change ~10% during differentiation, and there exists a correlation between transcriptional activation and genes that move away from the NL (Peric-Hupkes *et al.*, 2010). However, many of the genes that relocate to different subnuclear spaces do not exhibit significant changes in gene expression. These results suggest that gene-NL associations are not a determinant of transcriptional activity, in agreement with previous experiments of artificial locus tethering to the NL (Kumaran and Spector, 2008). Interestingly, a recent live-cell study of LAD dynamics showed that only ~30% of LADs are associated with the nuclear periphery, and that upon mitosis LAD positioning is stochastically re-shuffled (Kind *et al.*, 2013). These observations highlight the dynamic nature of nuclear architecture and genomic regulation, and the high degree of heterogeneity

in transcriptional control that can be achieved.

### 1.1.3 Nuclear bodies

Numerous studies have revealed that protein concentrations inside mammalian cell nuclei are not spatially uniform, but rather concentrate in local accumulations known as nuclear bodies (NBs) (reviewed in Dundr and Misteli, 2010; Mao *et al.*, 2011) [Fig. 1.1C]. To date, there are more than ten reported NBs with specialized functions, including:

- The nucleolus, in which ribosomal RNA repeats are transcribed and ribosomes assembled (reviewed in Boisvert *et al.*, 2007);
- Nuclear speckles, which harbor the pre-mRNA splicing machinery (reviewed in Spector and Lamond, 2011);
- Cajal bodies, which contain high concentrations of splicing ribonucleoproteins and implicated in telomerase biogenesis and transport (reviewed in Machyna *et al.*, 2013);
- Promyelocytic leukemia (PML) bodies, whose function is hypothesized to be related to PML partner proteins' modification or degradation (reviewed in Lallemand-Breitenbach and de Thé, 2010).

An explanation for the existence of NBs in the nucleus is to allow for the occurrence of diverse functional processes in the same environment, while putatively increasing the efficiency and modulation of biochemical reactions inside their restricted volumes. They could also serve as storage or assembly sites for proteins.

The assembly of NBs inside the nucleus has drawn much attention, given their lack of

membranous barriers that separate them from the rest of the nucleoplasm (reviewed in Dundr and Misteli, 2010; Mao *et al.*, 2011). Fluorescence recovery after photobleaching (FRAP) experiments have shown rapid and dynamic exchange of major NB components with the nucleoplasm, suggesting a stochastic/ordered assembly of these nuclear sub-organelles (Kruhlak *et al.*, 2000; Phair and Misteli, 2000; Snaar *et al.*, 2000; Chen and Huang, 2001; Weidtkamp-Peters *et al.*, 2008). To date, no specific architectural protein has been identified for the formation of the diverse array of NBs, however, protein-protein and protein-RNA interactions have been identified as the binding forces for their formation and structural maintenance (reviewed in Dundr and Misteli, 2010; Mao *et al.*, 2011; Mao *et al.*, 2011).

Even more interesting is the fact that several NBs have been shown to dynamically sense and respond to cellular changes. Well-known examples of this phenomena are the strictly dependent formation of the nucleolus based on active rRNA transcription (Oakes *et al.*, 1998; Dousset *et al.*, 2000; Olson and Dundr, 2005), the formation of the histone locus body during S-phase in response to histone gene clusters transcriptional activation (Bongiorno-Borbone *et al.*, 2008), the formation of DNA damage repair foci upon DNA-double strand breaks (reviewed in Dellaire and Bazett-Jones, 2007), and the morphological changes of speckles after inhibition of transcription (Spector *et al.*, 1983; Hu *et al.*, 2009).

Despite our limited understanding in NB biogenesis and precise dynamic functions, NBs have been shown to be prominent features of the eukaryotic nuclear landscape that have important roles in nuclear function and cellular responses. Yet, amidst the highly exquisite organization of the nucleus, lies another architectural stratum: the packaging of the genomic sequence.

## **1.2 Genome packaging**

Packaging of DNA inside nuclei is important not only for protecting it against damage, but to ensure coordinated regulation of gene expression and inheritance to daughter cells. Inside the nucleus, DNA forms a complex with numerous proteins that help in its packaging. This DNA-protein complex is called chromatin, the stainable fraction that Walther Flemming observed over a hundred years ago (Flemming, 1877, 1878). Within cells, chromatin is further folded into chromosomes, the basic units of genetic information whose structural configuration changes depending on the cell cycle stage (discussed in section 1.2.2).

Over the years, various studies have uncovered the intricate structure of chromatin into different layers of organization, which we will discuss in the following sections.

### **1.2.1 The 10 nm fiber**

Using a combination of nuclease digestion and careful detergent-based spreading methods, early EM studies revealed that the basic arrangement of chromatin is a structure 10nm in diameter, collectively known as the “10 nm fiber” [Fig. 1.2A,B]. This fiber appears as a “beads on a string” configuration, where DNA is the string while the beads are the arrangement of eight core histone proteins (Olins and Olins, 1974; Kornberg, 1974). This DNA-protein bead unit was subsequently named “nucleosome” (Oudet, Gross-Bellard, and Chambon, 1975).

Further characterization revealed that single nucleosomes consist of ~146 base pairs (bp) of DNA wrapped in 1.65 left-handed superhelical turns around a histone octamer

composed of two copies each of the core histones H2A, H2B, H3, and H4 (Luger *et al.*, 1997). Core histones have been shown to be among the most evolutionary conserved eukaryotic proteins, emphasizing their crucial role in chromatin organization (Malik and Henikoff, 2003). Their positive charge favors the interaction with the negatively charged DNA. A “linker DNA” segment joins adjacent nucleosomes, and its size typically varies from 10 to 80 bp, depending on the organism studied (Felsenfeld and Groudine, 2003). Finally, the linker histone H1 protects the linker DNA from degradation near the nucleosome entry-exit points (Thoma and Koller, 1977), and gives stability to the 10nm fiber for the formation of higher order structures.

The architectural arrangement of DNA into nucleosomes facilitates its functional regulation. DNA wrapped around the surface of the histone octamer would be partially accessible to regulatory proteins, and therefore free to participate in biological processes such as transcription, replication, DNA repair, and recombination. However, it has long been known that *in vitro* transcription is severely impeded by nucleosomal arrays (Huang and Bonner, 1962; Morse, 1989; Laybourn and Kadonaga, 1991; O'Neill, Roberge and Bradbury, 1992). Therefore, nucleosomes have to be moved or modified if DNA is to be accessible for functional processes. In fact, chromatin is subject to nucleosome displacement by dedicated remodeling complexes such as the SWI/SNF, ISWI, CHD, and INO80-SWR1 ATP-dependent families, which participate in all aspects of DNA metabolism (reviewed in Eberharter and Becker, 2004; Workman, 2006).

Histone N-terminal tails protruding from the nucleosome core are subject to various post-translational modifications such as acetylation, phosphorylation, methylation, ubiquitination and ADP-ribosylation, which in turn recruit a myriad of chromatin-remodeling



activities (reviewed in Strahl and Allis, 2000). A “histone code” was proposed to describe the correlations between histone post-translational modifications and functional outputs, for example, H2 lysine acetylation and transcriptional activation, histone H1 and H3 phosphorylation involved in chromosome condensation during mitosis, or the correlation between methylation of lysines H3K9 and H3K27 with transcriptional repression, among others (Jenuwein and Allis, 2001). However, the majority of experimental data on histone post-translational modifications indicates that this “code” is a poor predictor of function at the molecular level, and various examples have been described where canonical modifications are involved in the opposite process from which they were originally associated (reviewed in Sims and Reinberg, 2008).

Besides the modifications of histone N-terminal tails, there exist alternative core histone variants that can be incorporated into nucleosomes for the building of specialized chromatin structures. The most well-known examples are histone H2A.Z, essential in mouse, fly, and frogs and having roles in transcription regulation, DNA repair, heterochromatin formation, chromosome segregation and mitosis (reviewed in Draker and Cheung, 2009); histone variant H3.3, associated with actively transcribed genes but also found in silent loci in pericentric heterochromatin and telomeres (reviewed in Szenker, Ray-Gallet, and Almouzni, 2011), and CENP-A, an H3 variant found at centromeric regions (reviewed in Quénet and Dalal, 2012).

As has been discussed, the 10nm fiber serves as the basic unit of DNA packaging inside the nucleus, and the prime substrate for functional regulation of the genome. The synergistic interplay between nucleosome remodeling, histone variant exchange, histone post-translational modifications, and the chromatin fiber architecture, highlights the

importance of DNA packaging inside the nucleus, and serves as a platform for the building of new layers of chromatin organization, as discussed below.

### **1.2.2. Higher-order structures of chromatin organization**

Overall, nucleosome wrapping results in a compaction of 5-10 fold of the DNA fiber (Kornberg, 1974). However, during mitosis chromosomes form highly condensed structures which can be easily discerned under the microscope, and these structures decondense into differentially stained euchromatin and heterochromatin fractions as soon as cells progress to interphase. The appearance of three different forms of DNA packaging (mitotic chromosomes, euchromatin, and heterochromatin) suggests the existence of higher order structures in which chromatin is organized at different stages of the cell cycle, and with different functional roles.

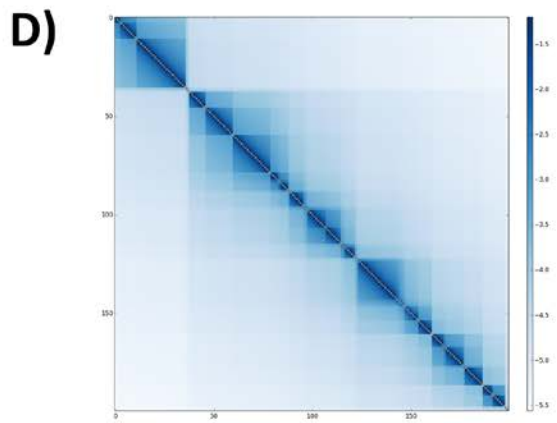
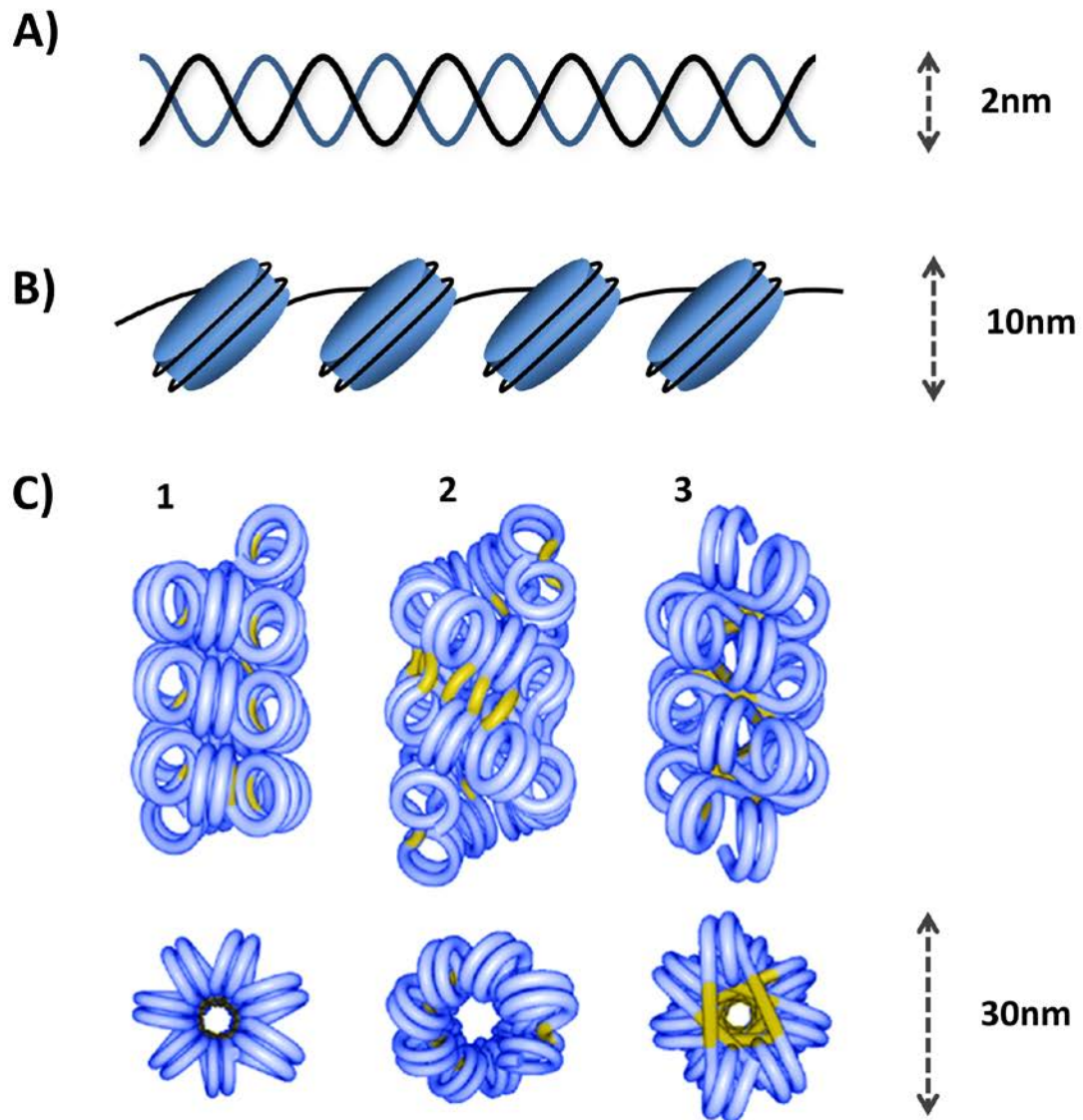
#### **1.2.2.1 The 30 nm fiber**

Prior to the discovery of the 10nm fiber and the identification of nucleosomes as core components of chromatin, the spreading of nucleated amphibian erythrocytes on water surfaces revealed the existence of a chromatin fiber arrangement ~30nm in diameter (Gall, 1966). After this initial observation, the 30nm fiber structure was subsequently detected *in situ* and *in vitro* by several groups, for which different models such as the one-start solenoid helix (Finch and Klug, 1976), zigzag two-start helical ribbon (Woodcock, Frado, and Rattner, 1984), and the interdigitated solenoid (Worcel *et al.*, 1981), were proposed [Fig. 1.2C]. The

heterogeneity of models for describing nucleosome packaging inside 30nm fibers was mainly derived by the different methods by which chromatin was reconstituted *in vitro*, as well as the variability in linker DNA and nucleosome composition used. More recently, a detailed cryo-electron microscopy analysis of reconstituted nucleosomes with *Xenopus laevis* histones lacking post-translational modifications, revealed an H1-dependent double helix twisted by tetranucleosomal units (Song *et al.*, 2014).

*In situ*, 30nm structures have only been detected on the chromatin of nucleated erythrocytes and echinoderm sperm (Grigoryev and Woodcock, 2012). These cell types are highly specialized, and exhibit minimal or largely absent transcriptional activity, presence of highly charged H1-type histones, and longer nucleosome repeat lengths. More recently, mouse retinal rod photoreceptors were shown to harbor arrays of 30nm fibers in the less compact regions surrounding the central mass of heterochromatin (Kizilyaprak *et al.*, 2010). However, no structural signatures for higher-order 30nm folding have been detected in either interphase nuclei or mitotic chromosomes in mammals, casting doubt as to whether this arrangement would be biologically relevant and accommodate basic processes such as transcription, replication, and recombination *in vivo* (reviewed in Fussner *et al.*, 2011).

To date, 30nm fiber arrangements are recognized as a specific type of conformation that chromatin can adopt in different contexts. Given its lack of *in situ* detection in mammalian and other eukaryotic cells, as well as their absence from the most recently derived Hi-C data models, it has been proposed that this arrangement may be utilized solely for specialized cell types such as nucleated erythrocytes, echinoderm sperm, and retinal rod photoreceptors.



## **Figure 1.2 Schematic depictions of chromatin structures**

**A)** The double helical structure of DNA **B)** is further folded into chromatin via its association with nucleosomes, and adopting a beads-on-a-string configuration named 10nm fiber. **C)** Further coiling of the 10nm fiber gives rise to different configurations, collectively known as the 30nm fiber. These include the solenoid (1), zig-zag (2), and interdigitated solenoid (3) arrangements (Obtained from Wu, Bassett, and Travers, 2007, and modified to fit current color scheme). **D)** Representative matrix of defined chromatin interactions for a simulated 60Mb chromatin segment. Topologically associating domain structures are observed as the blue regions surrounding the x-axis, in which contacts are far more frequent compared to the rest of the segment. **E)** Cartoon representation of chromosome territories in mouse fibroblasts. Each color represents a separate chromosome. Notice how each chromosome occupies a defined volume inside the nucleus, but interacts with neighboring chromosome at the periphery of their territories.

Nevertheless, one must not discard the existence of 30nm fiber arrangements which could form spontaneously and dynamically across different genomic regions in interphase nuclei, which could be blind to current microscopy and population-based C methodologies. In-depth analysis of molecular interaction data could shed more light into this elusive, yet intrinsically interesting chromatin conformation structure.

### **1.2.2.2 Topologically associating domains**

Without the accurate detection of 30nm chromatin fibers in mammalian interphase cells, the structural arrangement of chromosomes over the 10nm fiber remained a mystery until the improvement of fluorescence microscopy. Chromatin domains ranging from few hundred Kb to several Mb were first identified microscopically as persistent structural features of chromosomes during interphase (Ma *et al.*, 1998; Cremer and Cremer 2001), which act as replication foci during S phase (Sparvoli, Levi, and Rossi, 1994; Jackson and Pombo, 1998; Zink *et al.*, 1999). However, structural details of chromatin domains were not well understood given the limited resolution of fluorescence microscopes at the time.

With the development of the 3C technique (Dekker *et al.*, 2002, and extensively described in section 1.3), high resolution structural details of these domains were gained. 3C-based methods assess physical interactions among pairs of crosslinked genomic loci, which can give insight into the spatial organization of chromatin at different levels. Unbiased genome-wide (Hi-C) and regional (5C) studies in human, mouse, and fly, revealed the presence of small architectural domains characterized by more frequent associations between their sequences compared to other regions in the genome (Dixon *et al.*, 2012; Hou *et al.*,

2012; Nora *et al.*, 2012; Sexton *et al.*, 2012), similar to the previously microscopically identified domains. These structures are now known as topologically associating domains (TADs) [Fig. 1.2D].

TADs have sizes ranging from tens of Kb to several Mb (average size of ~1Mb), and are largely conserved across different mammalian cell types (Dixon *et al.*, 2012). Very interestingly, TAD or TAD-like structures have not been observed in yeast (Duan *et al.*, 2010), bacteria (Umbarger *et al.*, 2011), or plants (Moissiard *et al.*, 2012), suggesting distinct structural organization of specific genomes at the 100Kb-1Mb length scale. Despite their structural conservation in different mammalian cell types, it is still unclear what factors determine TAD boundaries. In general, TAD boundary regions have been found to be enriched in CTCF binding sites, transcription start sites of housekeeping genes, insulator protein binding sites, transfer RNAs, and short interspersed element (SINE) retrotransposons (Dixon *et al.*, 2012). Although the role of these elements in establishing TAD identity needs to be further tested, current evidence suggests that TAD boundaries may be genetically defined. A deletion experiment of a TAD boundary in the X chromosome inactivation center, led to partial fusion of the neighboring TADs (Nora *et al.*, 2012). Definitive proof of the existence of TAD boundary elements would come in a genetic experiment where a specific exogenous TAD boundary is inserted into a larger TAD, with the subsequent observation of the original TAD splitting.

At the functional level, genes within each TAD seem to have coordinated expression during differentiation, arguing for TAD-specific roles in transcriptional regulation. Even more interesting is the fact that TAD boundaries overlap those of DNA replication timing domains (Dixon *et al.*, 2012; Ryba *et al.*, 2010), in agreement with the previously reported

microscopic domains. TAD structures make evident the connection between replication timing and chromatin transcriptional and organizational status in individual chromosomes, highlighting the intrinsic interplay between genomic function and 3D structure.

### **1.2.2.3 Chromosome territories**

Just as TADs possess different arrangements within each chromosome, chromosomes themselves occupy distinct positions inside the nucleus. Knowledge about this organization was first published in 1885, when Carl Rabl reported strikingly similar polarized patterns of chromosome order before and after mitosis in salamander cells (Rabl, 1885). In 1909, Theodor Boveri expanded this theory of chromosome individuality by studying *Ascaris magalocephala* (Boveri, 1909), reproducing Rabl's observations and coining the term “chromosome territories.”

Subsequent cytological investigations using giemsa staining, laser-UV-microirradiation coupled with radioactive labeling, and fluorescence *in situ* hybridization (FISH), have all uncovered the highly organized positioning of chromosomes into defined territories in interphase cells (Stack *et al.*, 1977; Cremer *et al.*, 1982; Guan *et al.*, 1993; reviewed in Cremer and Cremer, 2006) [Fig. 1.2E], with varying levels of interactions between territories that have important consequences for genomic function and stability (Branco and Pombo, 2006; reviewed in Cremer and Cremer, 2010). Even more interesting is the fact that chromosome territories seem to be positioned non-randomly inside the nucleus, with gene-rich chromosomes generally located at the nuclear centroid, and gene-poor ones positioned towards the periphery (reviewed in Cremer and Cremer, 2006).



More recently, the first Hi-C study performed in human cells molecularly confirmed the existence of chromosome territories. By reporting far more frequent interactions between distant sequences located in the same chromosome, compared to any other loci in the rest of the genome, a defined spatial positioning of chromosomes was shown (Lieberman-Aiden *et al.*, 2009). On top of this territorial organization, the study also identified the existence of two classes of genomic compartments, the first one being gene rich, transcriptionally active, and hypersensitive to DNase I digestion, while the second was relatively gene poor, transcriptionally silent, and DNase I insensitive. This is similar to the EM-observed euchromatin and heterochromatin regions in interphase cells. So not only do chromosomes occupy distinct territories, but they fit into an even higher-order chromatin arrangement determined by their transcriptional status.

With the exploration of chromatin structure by the use of diverse microscopy and 3C technologies, it has been possible to elucidate the different organizational units of the chromatin fiber at the sub-chromosomal and chromosomal scale. The current plurality of chromatin conformations highlights the role that 3D architecture plays in the functionality of cells, and reflects the complexity that exists within the nuclei components. Increasing our knowledge of the precise but dynamic organization of chromatin inside eukaryotic nuclei depends on the advances of current technologies for the assessment of chromatin conformation, especially the 3C-based ones, which we will extensively discuss in the following section.

### **1.3 Chromosome conformation capture technologies**

The advance in our molecular understanding of the fine details of chromosome organization has been achieved thanks to the creation and diverse adaptations of the 3C technology. Since its initial publication (Dekker *et al.*, 2002), 3C and its derivative techniques have been the platform by which targeted and genome-wide analyses of chromatin interactions have been performed, revealing new features of chromosomal packing and overall genome architecture. Given its fundamental impact on the studies of chromatin organization, the 3C technique and its several modifications will be extensively described, touching on technical and analysis of results which will become important for the evaluation of the data generated in this project.

#### **1.3.1 3C**

3C is based on the long-time used formaldehyde tissue fixation for the identification of chromatin interacting segments [Fig. 1.3A]. Formaldehyde is a water soluble gas of formula HCHO. Due to its small size, it has rapid penetration into tissue and has therefore been used for a long time as a tissue/cell fixative and embalming agent. In solution, formaldehyde forms methylene hydrate molecules which can react with one another to form polymers. Inside cells, the aldehyde group reacts with nitrogen groups and other protein atoms and forms methylene bridges (-CH<sub>2</sub>-) between proteins in physical proximity. Carbohydrates, lipids, and nucleic acids are thought to be trapped in a matrix of cross-linked proteins, and therefore the original tissue structure is preserved or fixed, depending on the reaction time and conditions of the formaldehyde treatment. Formaldehyde has been

extensively used as a fixative for histology and microscopy, as well as a protein crosslinker for chromatin immunoprecipitation reactions (ChIP).

After genomic crosslinking of a particular cell population, the 3C protocol requires the digestion of DNA by restriction enzymes, typically 6 base pair cutters like *EcoRI*, *XhoI*, *HindIII*, *BglII*, *BamHI*, *KpnI*, *AseI*, *MfeI*, or *NspI* (Dekker *et al.*, 2002; Tolhuis *et al.*, 2002; Palstra *et al.*, 2003; Murrell, Heeson, and Reik, 2004; Spilianakis and Flavell, 2004; Liu and Garrard, 2005) [Fig. 1.3B]. After digestion, chromatin is subject to re-ligation under dilute conditions to favor fusion of fragments held in close spatial proximity [Fig. 1.3C]. As a result, the library of ligated products is the representation of DNA fragments that were physically close together in nuclear space [Fig. 1.3D]. Through the use of specific primers, the frequency of ligation of any selected pair of restriction fragments can be assessed to determine relative spatial proximities in the cell compared to a “control” ligation template. Most 3C control templates were generated through the random ligation of purified genomic DNA, BAC or PAC clones using the same experimental conditions for the assayed template. Initially, 3C protocols used the polymerase chain reaction (PCR) for quantification of interaction frequencies either via product extraction and concentration measurement, or via ethidium bromide gel imaging. These semi-quantitative methods have now been substituted for quantitative PCR measurements (3C-qPCR, Hagège *et al.*, 2007).

3C was initially developed to study the spatial organization of yeast chromosome III (Dekker *et al.*, 2002), but it was subsequently applied to the study of genomic organization and transcriptional regulation. Examples of these include:

- The analysis of long-range looping interactions between the beta-globin locus and its locus control region at specific developmental stages in mouse and human (Tolhuis *et*

- al.*, 2002; Palstra *et al.*, 2003).
- The interactions regulating the timing of the transition between poised and active gene expression in the alpha globin locus (Vernimmen *et al.*, 2007).
  - The partitioning of the imprinted genes *Igf2* and *H19* into parent-specific chromatin loops (Murrell, Heeson, and Reik, 2004).
  - The long-range interactions between actively transcribed *Igκ* alleles and three transcriptional enhancers (Liu and Garrard, 2005).
  - The intrachromosomal contacts among genes in the TH2 cytokine locus (Spilianakis and Flavell, 2004).

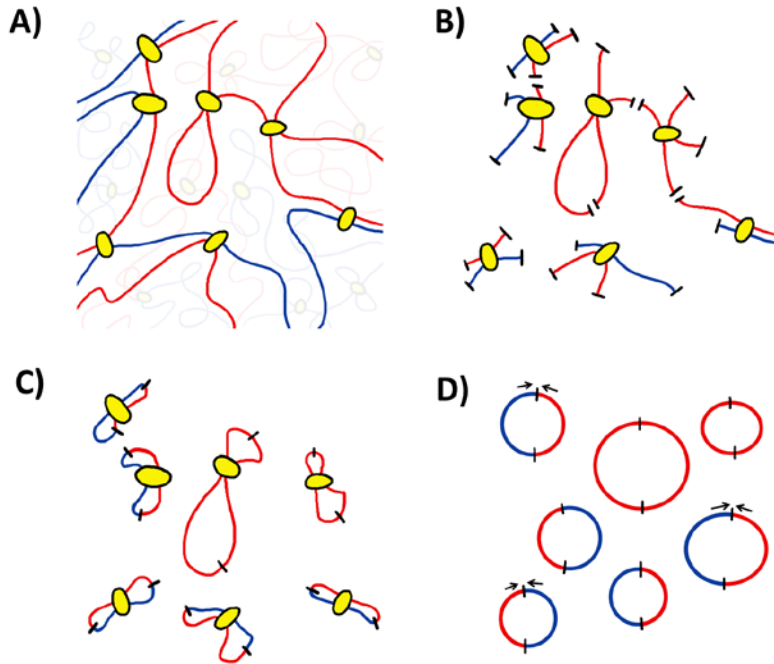
Although labor intensive, the 3C technique has also been used for the identification of gene-specific regulatory elements acting via long-range looping, such as in the case of the cystic fibrosis transmembrane conductase regulator gene (*CFTR*) and its associated enhancers located 20 and 80 kb upstream, and 109 and 203 kb downstream of its promoter (Gheldof *et al.*, 2010).

Nowadays, the standard 3C technique has been substituted for its genome-wide adaptations (discussed in the next sections), however, it remains as an experimental alternative for the assessment of interactions between any two specific DNA fragments, such as enhancer-promoter contacts. A comprehensive discussion of 3C and derived methodologies advantages and technical issues will be presented at the end of this section.

### 1.3.2 4C

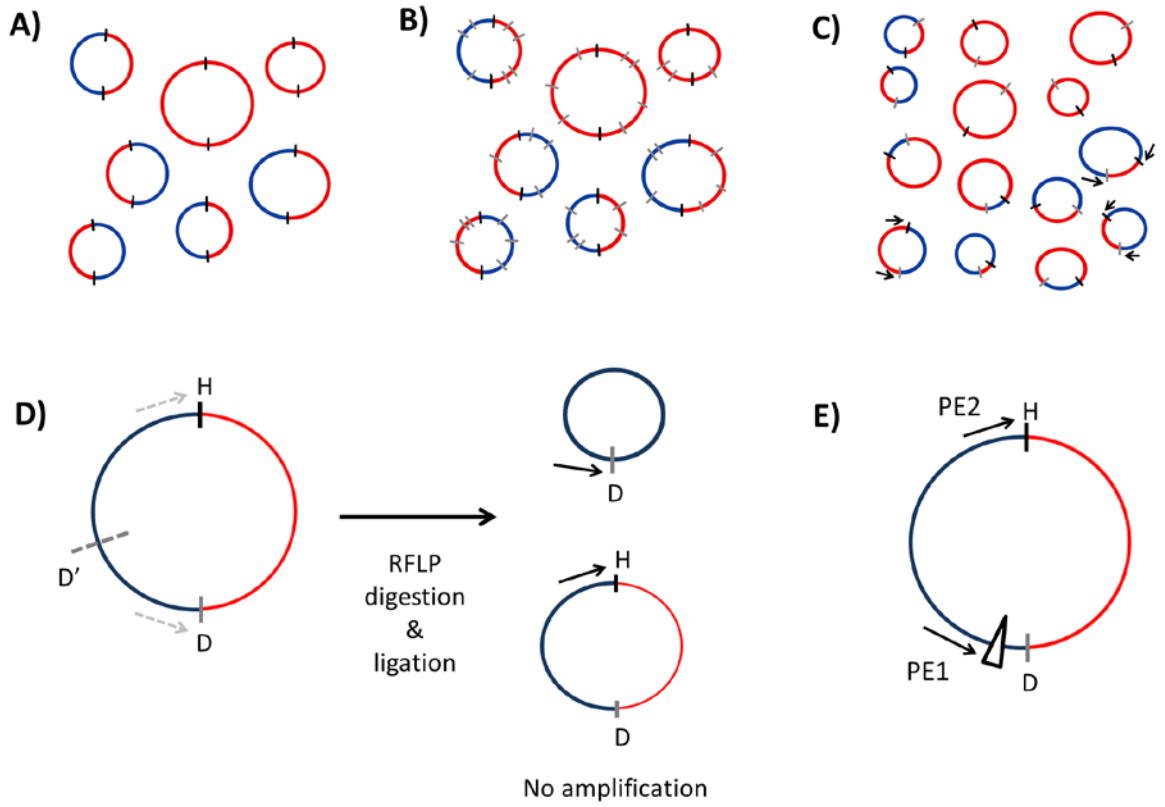
In order to evaluate chromatin interactions at a genome-wide scale, different protocols were developed from the standard 3C technique. The first of them were two methodologies named 4C, but using slightly different protocols.

In the chromosome conformation capture-on-chip (4C) technique, the ligated 3C template is subjected to a second restriction digestion with a frequent cutter enzyme (i.e. 4bp recognition sequence enzymes like *DpnII*, *NlaIII*) [Fig. 1.4A,B], and ligated under dilute conditions to generate small DNA template circles [Fig.1.4C]. By the use of an inverse PCR reaction employing specific primers for a targeted genomic region (known as “bait” or “viewpoint”), interacting sequences (“captures”) can be amplified and their identities determined by the use of DNA microarrays (Simonis *et al.*, 2006). The other methodology, known as circular chromosome conformation capture (4C), employs the same principle of its chip cousin, but does not include a second round of enzymatic restriction digestion and ligation and promotes circle formation through the use of high concentrations of ligase and prolonged incubation times (>1 week, Zhao *et al.*, 2006). To date, only the chip version of 4C has remained actively used in the field of chromatin organization given its several advantages over its homologue. These include improved resolution, less interaction data noise, and heterogeneous template generation through various enzymatic combinations.



**Figure 1.3 Overview of 3C protocol steps**

**A)** Chromatin is cross-linked inside nuclei with formaldehyde. Regions that were near to each other in the 3D space either by protein bridges (yellow dots), or by mere colocalization, will be captured. Most procedures use an average of ten million cells per experiments. **B)** Primary restriction enzyme digestion of crosslinked chromatin. This step is typically performed using a 6bp restriction cutter, such as *HindIII* (restriction sites marked in black lines). **C)** Ligation of digested fragments. This step is performed at very dilute concentrations to favor intra-molecular ligation of fragments. **D)** De-crosslinking and purification of 3C library. The final product of the 3C procedure is a library that represents all sequences that were in physical proximity in the original cell population. Assessment of interactions between any pair of regions is performed through the use of primers targeting these regions (black arrows).



### Figure 1.4 Outline of the 4C approach

**A)** 4C starts with a 3C template generated by chromatin crosslinking and ligation of an enzymatic digestion using a 6bp cutter (restriction sites shown in black). **B)** The 3C template is subjected to a second round of restriction digestion, typically with a 4bp cutter (shown in grey lines). **C)** Dilute ligation of cut DNA results in the generation of a 4C library, where template DNA circles are smaller compared to 3C. Interacting partners (red portion of the circle) of a specific region (blue portion of the circle) are subsequently amplified using specific primers (black arrows). **D)** Schematic outline of the allele-specific 4C approach using restriction fragment length polymorphisms (RFLPs). Presence of a RFLP that inserts a new *DpnII* cutting site (grey dashed line, marked with D'), impedes interacting partner amplification by the generation of two separate circle templates when the ligation is performed. Because of this, only the non-RFLP fragment can be amplified with the original primer set (grey arrows). *HindIII* restriction site marked as H. **E)** PE-4Cseq strategy outline. PE-4Cseq makes use of paired-end sequencing, in which one read (PE1) amplifies a genotyping SNP between the different alleles (white triangle), which is subsequently used to separate the interaction partner reads (PE2).



The identification of interacting partners via DNA chips has been substituted by next generation sequencing, thus giving rise to the 4C-Seq protocol, which enables accurate quantification of chromatin interaction frequencies at higher resolution.

Because of its higher resolution and ability to detect intra-chromosomal (*cis*) as well as inter-chromosomal (*trans*) interactions, 4C has been used to investigate the effects of regulatory control regions and biological processes in the architectural organization of chromatin, and vice versa. Examples of 4C studies include the assessment of developmentally regulated interaction profiles of the beta-globin locus in wider genomic ranges (Simonis *et al.*, 2006); the effects of transcriptional inhibition (Palstra *et al.*, 2008) or activation (Hakim *et al.*, 2011) on chromosomal structure; the impact of an ectopic human locus control region on DNA interactions on a cluster of mouse housekeeping genes (Noordermeer *et al.*, 2008, 2011); and the preferential clustering of polycomb repressed genes in *Drosophila* (Tolhuis *et al.*, 2011; Bantignies *et al.*, 2011); among others.

Originally, the 4C protocol used 6bp sequence recognition restriction enzymes as the primary cutters, and 4bp recognition restriction enzymes as secondary ones (Simonis *et al.*, 2006). This design is referred to as a “6x4” strategy, and identifies the long-range contacts of a viewpoint with larger regions elsewhere on the chromosome and the genome. Most recently, the use of a “4x4” strategy (two 4bp restriction enzymes as primary and secondary cutters), increases interaction profiles resolution by the generation of a higher number of smaller template circles, and therefore a higher coverage of the genomic sequence. This type of amplification makes the 4x4 4C-Seq an excellent approach for the identification of local regulatory elements for any specific gene (van de Werken *et al.*, 2012).

Another technical improvement to the 4C-Seq technique has been its ability to detect allelic conformations. This is achieved by the use of selective enzymatic template digestion (Splinter *et al.*, 2011) [Fig. 1.4D], or by paired-end sequencing, where one read amplifies the interacting partner while the second amplifies a genotyping SNP (PE-4Cseq. Holwerda *et al.*, 2013; de Wit *et al.*, 2013) [Fig. 1.4E].

Similar to 3C, careful analysis must be performed on 4C-generated data before drawing any conclusions. Given the binary nature of contacts present in a 4C profile (i.e. a contact is present or absent in a pool of mapped capture reads), significance of interactions between viewpoints and captures must be assessed based on the enrichment of contacts in the capture vicinity and normalized based on the expected number of mapped background reads and fragment size. Different data normalization and comparison methods for calling significantly interacting regions have been described (Splinter *et al.*, 2011, van de Werken *et al.*, 2012), and I introduce in chapter 4 a novel methodology for quantitative 4C data analysis.

Because of its ability to identify *cis* and *trans* interactions in an allele-specific manner for individual genomic regions, and its capacity to reproduce previously identified Hi-C TAD observations (Amos Tanay, personal communication), I have used the 4C-Seq technology to assess changes in chromatin architecture after the occurrence of copy number variation (see section 1.5 for this thesis project summary, and chapters 2 and 4 for technical details on the protocol and materials used).

### 1.3.3 5C

The carbon copy chromosome conformation capture (5C) was the first high-throughput methodology to report comprehensive interaction profiles between multiple selected chromosomal sequences (Dostie *et al.*, 2006). 5C makes use of primers specially designed to anneal next to each other across ligated junctions of head-to-head interactions present in a standard 3C library [Fig. 1.5A]. Primers that anneal next to each other in the 3C template are ligated, and by including universal tails at the ends of these primers, the ligation products can be amplified [Fig. 1.5B]. As a result, the 5C library is a “carbon copy” of a subset of the original 3C library, determined by the combination of primers used to assess contacts among specific chromosomal regions. Initially, 5C used both microarrays and deep sequencing for the identification of interacting chromatin segments, but nowadays sequencing is mostly used for contact analysis (Dostie *et al.*, 2006; Baù *et al.*, 2011; Umbarger *et al.*, 2011; Nora *et al.*, 2012; Sanyal *et al.*, 2012).

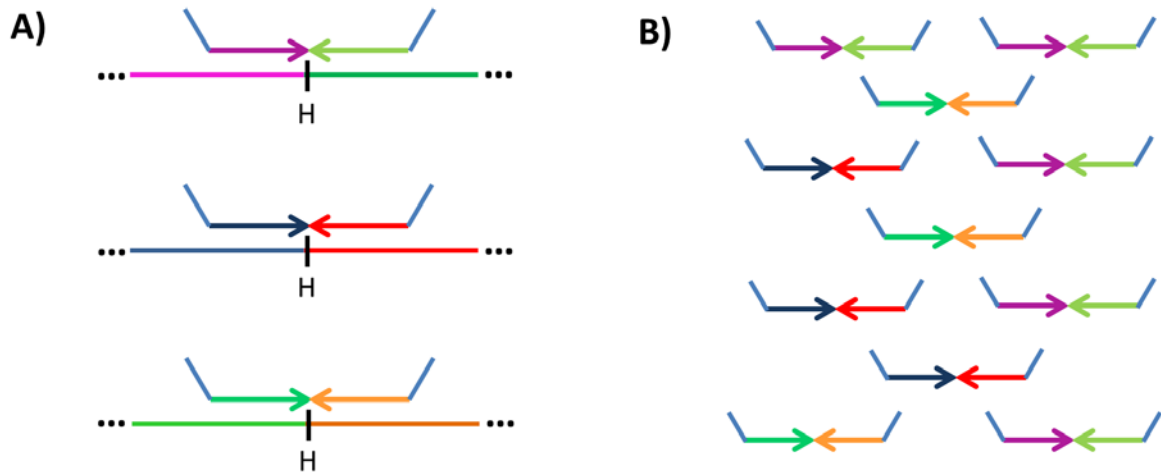
Because of its interaction detection methodology, 5C has been called the “many-versus-many 3C,” in which diverse primer combinations can be multiplexed to question the frequency of ligation of targeted chromosomal regions. 5C data is usually summarized in contact matrices between the assayed fragments for which primers were designed, and these matrices are subsequently subjected to statistical analysis and 3D modeling for discovering specific sub-chromosomal conformations.

5C has extensively contributed to the chromatin organization field with studies that include the discovery of TADs (Baù *et al.*, 2011; Nora *et al.*, 2012), the first report of the 3D architecture of a bacterial genome (Umbarger *et al.*, 2011), the spatial partitioning of the regulatory landscape of the X-inactivation centre (Nora *et al.*, 2012), the three-dimensional

architecture of Hox cluster silencing and activation in humans (Ferraiuolo *et al.*, 2010; Wang *et al.*, 2011), and the systematic usage for identification of contacts between regulatory sequences and gene promoters in the ENCODE pilot project regions (Sanyal *et al.*, 2012).

A major downfall of the 5C methodology lies in the experimental costs. Depending on the size of the selected regions for study, hundreds or thousands of primers need to be designed to cover all the possible ligation products within a 3C library, therefore scaling the expenses for materials, sequencing, and labor time. In fact, most published 5C studies have concentrated on regions <5Mb in size (Dostie *et al.*, 2006; Ferraiuolo *et al.*, 2010; Wang *et al.*, 2011; Baù *et al.*, 2011; Umbarger *et al.*, 2011; Nora *et al.*, 2012). In addition, data interpretation must include several controls both in template generation (similar to 3C control libraries of randomly ligated DNA sequences), and in computational analysis (such as the control of peak calling based on varying genomic distances, Sanyal *et al.*, 2012). Additionally, 5C has not been shown to detect allele-specific conformations, given the difficulty of selective SNP hybridization for the primers used.

All in all, 5C remains as the most suited C technique for the high resolution assessment of interaction profiles of small Mb regions in the genome, whose application has provided great insight into the 50Kb-10Mb scales of chromatin organization.



**Figure 1.5 Overview of the 5C methodology**

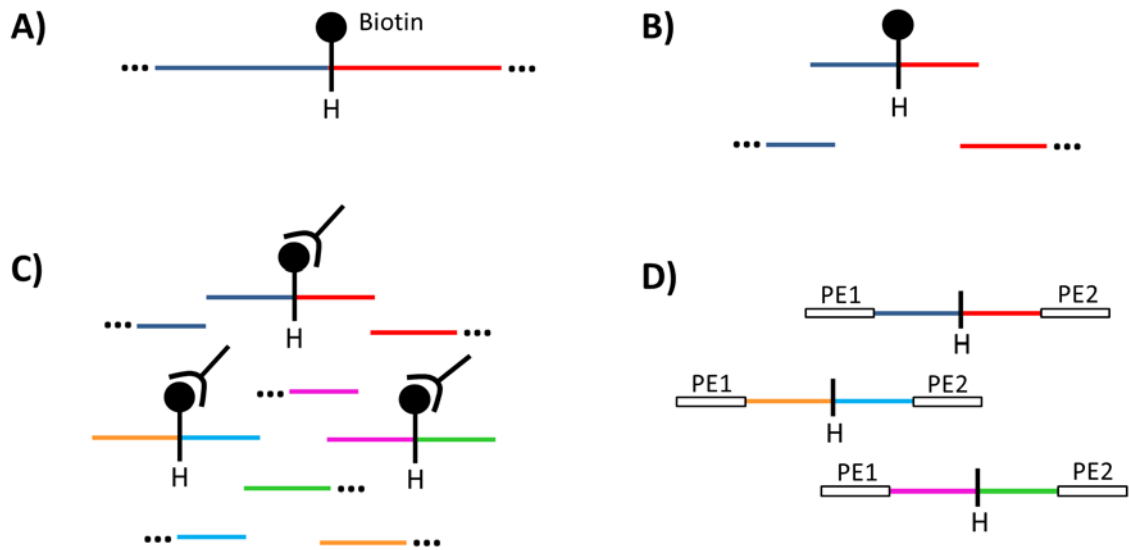
**A)** A standard *HindIII* 3C library is depicted (enzyme restriction sites marked with H). Sequences in the rest of the circle are marked with dots. Different colors represent different genomic sequences captured in the 3C library. Colored arrows represent the different primers that annealed to the borders of each restriction fragment. Annealed primers are subsequently ligated. **B)** Ligated primers are amplified via the adaptors in the tail (depicted as light blue vertical lines), which will facilitate the construction of a sequencing library.

### 1.3.4 Hi-C

Hi-C is the 3C methodology that determines chromatin interactions in an unbiased and genome-wide manner. Hi-C introduces biotin-labeled nucleotides in the restriction ends after the initial enzymatic digestion of crosslinked chromatin [Fig. 1.6A,B], and is followed by ligation, DNA purification, and biotin pull-down [Fig. 1.6C]. The introduced biotin label ensures that only ligation products are selected for further analysis, representing the entirety of interactions present in the 3C template (Lieberman-Aiden *et al.*, 2009; van Berkum *et al.*, 2010) [Fig. 1.6D].

The identity of ligation products is determined by performing paired-end high-throughput sequencing. Resulting reads are mapped to the genome of origin, and interaction matrices are built upon these data. Initial Hi-C experiments performed in human cells produced interaction matrices with a resolution of ~1Mb (Lieberman-Aiden *et al.*, 2009). However, with the advance of sequencing technologies, current Hi-C maps have reached the astounding resolution of 5-10Kb (Jin *et al.*, 2013).

Human Hi-C data confirmed several of the organizational features which had previously been reported by microscopy, such as the territorial organization of chromosomes (reviewed in Cremer and Cremer, 2010), and the compartmentalization of chromatin into active and inactive neighborhoods, similar to the differentially stained euchromatin and heterochromatin in the nucleus (Lieberman-Aiden *et al.*, 2009).



**Figure 1.6 Overview of the Hi-C technique**

A) Right after restriction digestion in the 3C protocol, DNA ends are marked with biotin, followed by blunt-end ligation of crosslinked fragments. B) Labeled fragments are subsequently sheared, in order to select for sizes appropriate for sequencing. C) Sheared ligation junctions are purified from the DNA pool via biotin pull-down by streptavidin magnetic beads. D) Paired-end sequencing adaptors are ligated to the ends of the pulled-down ligation junctions, and libraries made for the examination of global genome interaction profiles.

In addition to confirming previous aspects of chromosome architecture and nuclear organization, Hi-C maps have uncovered new properties of chromatin structure. The first human Hi-C dataset was examined using diverse polymer models, and it was found that a specific model named “fractal globule” allowed for the most biologically relevant properties of the genome, such as easy folding and unfolding of chromosomal sections, chromosome territory separation, and the absence of knots and entanglements which could be detrimental to cellular division (Lieberman-Aiden *et al.*, 2009). The properties of the fractal globule model may be relevant for gene transcriptional regulation, although several other models have been put forward fitting Hi-C data and allowing for basic biological processes to occur (discussed in Barbieri *et al.*, 2013).

Hi-C has been applied to other organisms besides human. These include *Drosophila* (Sexton *et al.*, 2012) and mouse (Dixon *et al.*, 2012). Both studies identified the preferential 3D clustering of genomic regions based on their transcriptional state (similar to the identified compartments in human Hi-C), and uncovered TAD organization for both mouse and fly chromosomes. Very interestingly, it was observed that TAD architecture is stable across different cell types, and highly conserved between mouse and human, indicating that TADs are inherent features of mammalian genomic organization (Dixon *et al.*, 2012). A variation of the Hi-C technology was also used for the high-resolution three dimensional (3D) modeling of the yeast genome (Duan *et al.*, 2010), which confirmed the previously reported Rabl configuration of its chromosomes. Currently, the latest reported uses of Hi-C technology have been the elucidation of the folded structure of the mitotic chromosomes in HeLa cells (Naumova *et al.*, 2013), and the usage of Hi-C data for contig positioning during genome scaffolding and assembly analyses (Kaplan and Dekker, 2013).



The most recent technical advance in the Hi-C protocol has been its coupling to single-cell sequencing. The technique is performed inside the nuclei of permeabilized cells, which are then sorted and subjected to single-cell sequencing to obtain individual maps of chromatin interactions (Nagano *et al.*, 2013). Results have revealed striking levels of cell-to-cell variability in intra- and inter-chromosomal interactions in mouse T helper cells, consistent with previous microscopic observations. In addition, a high degree of variability was observed in internal chromosome organization. However, when the data of several single-cell Hi-C profiles is averaged, it reports previously published TAD boundaries derived from experiments using millions of cells. It remains to be answered whether every TAD is present in each cell of the population, or whether variability is due to reproducible modular domain folding.

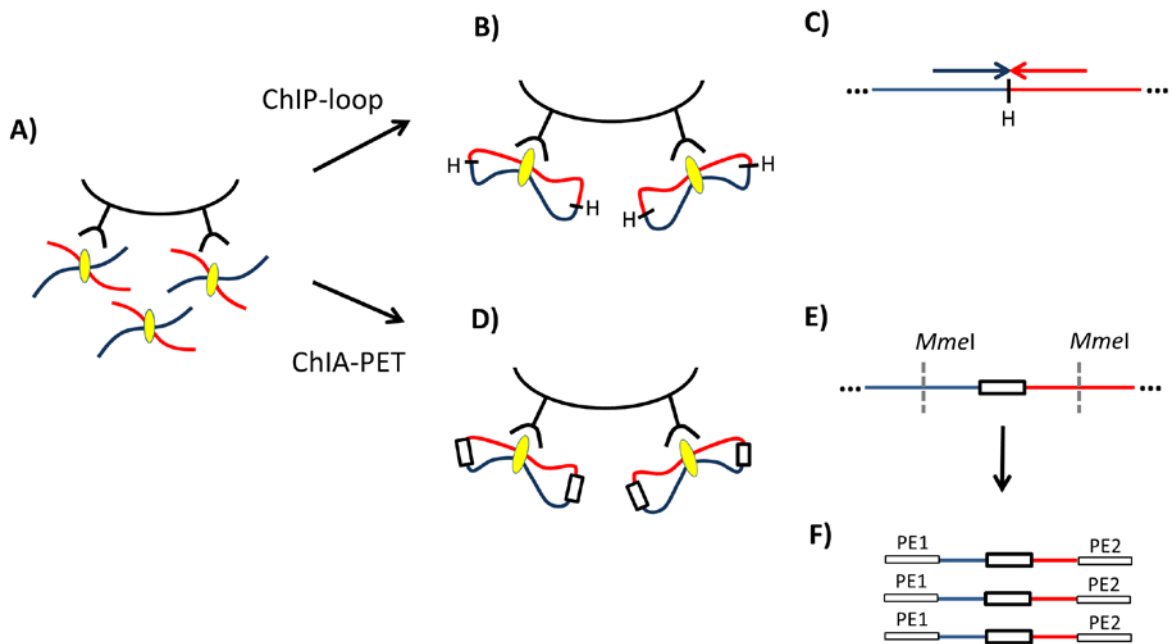
As with any other technology, Hi-C has been found to have certain biases which affect the data analysis, such as ligation fragment sizes, GC content of trimmed ligation junctions, and the uniqueness of the obtained ligation sequences (Yaffe and Tanay, 2011). For this reason, several analysis pipelines have been developed to assess Hi-C datasets (Lieberman-Aiden *et al.*, 2009; Yaffe and Tanay, 2011; Dixon *et al.*, 2012; Jin *et al.*, 2013; Nagano *et al.*, 2013), proposing different models of organization (Barbieri *et al.*, 2013).

Irrespective of the way Hi-C data is preferentially analyzed, several important insights have been derived from its application into different biological questions, providing essential knowledge to our understanding of chromosomal folding at both the Kb and whole genome scales.

### 1.3.5 ChIP-based 3C techniques

A divergent evolution of the 3C methodology involves the use of ChIP for the selection of specific interactions mediated by particular proteins. The ChIP-loop (Horike *et al.*, 2005) and 6C (Tiwari *et al.*, 2008) assays both use standard 3C procedure, however before de-crosslinking, the ligated material is subjected to ChIP using an antibody against a protein of interest [Fig. 1.7A]. ChIP-loop libraries are assayed by standard 3C PCR amplification to detect pair-wise interactions [Fig. 1.7B,C], while 6C includes an additional cloning step and a final PCR contacts assessment.

The genome-wide version of ChIP-loop is ChIA-PET (chromatin interaction analysis by paired-end tag sequencing, Fullwood *et al.*, 2009). ChIA-PET reports contacts between any pair of genomic sites brought together in the nuclear space by a specific protein [Fig. 1.7A,D,E,F]. This technique can be thought of as the ChIP-coupled version of Hi-C. Initial studies using ChIA-PET uncovered the long-range interaction network between estrogen receptor  $\alpha$  (ER- $\alpha$ ) binding sites and gene promoters in human cells (Fullwood *et al.*, 2009), the important role of CTCF in demarcating chromatin-nuclear membrane attachments and potential influence on enhancer-promoter looping in mouse embryonic stem (ES) cells (Handoko *et al.*, 2011), the widespread promoter-centered interactions and clustered aggregation of genes transcribed by RNA polIII (Li *et al.*, 2012), and the very recent mapping of promoter-enhancer interactomes of mouse pluripotent ES cells and differentiated B lymphocytes, with the discovery of wide enhancer usage between tissues (Kieffer-Kwon *et al.*, 2013).



**Figure 1.7 Schematic of ChIP-based 3C techniques**

**A)** Chromatin is crosslinked, sheared, and subjected to ChIP using antibodies against a specific protein. **B)** In the case of the ChIP-loop assay, crosslinked chromatin is cut with a restriction enzyme, and ligated during antibody pull-down. **C)** After ligation, chromatin is decrosslinked, and primers (marked with red and blue arrows) used to amplify interactions between specific sequences. **D)** In ChIA-PET, after chromatin preparation and ChIP, linker sequences are ligated (white rectangles), followed by proximity ligation of the fragments. **E)** Chromatin is decrosslinked, purified, and subjected to *MmeI* restriction digestion to select for appropriately sized fragments for sequencing. **F)** PE adaptors are ligated to the digested *MmeI* fragments, and sequencing libraries made.

ChIA-PET offers yet another level of analysis of chromatin organization: the questioning of the roles of specific proteins in the 3D nuclear space, which provides a suitable way to interrogate chromatin structure and its correlation with genomic functional outputs.

### **1.3.6 C-methodologies discussion**

Taken together, 3C-based studies have revealed the baffling molecular complexity of the 3D organization of eukaryotic genomes, and an integrated view of chromosome architecture at different stages of the cell cycle has started to emerge. From the basic looping interactions to TAD structures, to the compartmentalized segregation of active and inactive chromatin, the discovered networks of short and long-range communications between different elements across the genome makes evident the important role of chromatin organization in the regulation of genomic function, and future exploration of chromosomal folding in different contexts will provide an abundance of new paradigms and insights of nuclear architecture.

Although 3C and derived methods are based on experimentally straightforward steps, the implementation and interpretation of C experiments require careful analysis and planning (Dekker, 2006). 3C cannot *per se* estimate the proportion of cells in which two particular DNA fragments interact, but reports the average patterns of interaction for the analyzed cell populations.

Because 3C methods are based on the principle of formaldehyde fixation and proximity ligation, several unaccounted factors may affect the efficiency of the protocol

(local distribution of restriction sites, fragment sizes, GC fragment content, cohesive ends, genomic repeat content, presence of crosslinked proteins, etc). A recent quantitative investigation of the actual frequencies of ligation between the major beta-globin gene promoter and its distant enhancers revealed that the amount of ligation products does not exceed 1% of all fragments subject to ligation, therefore arguing for a more careful analysis of 3C data in general and additional validations to 3C results (Gavrilov, Golov, and Razin, 2013).

As with any other type of experimental results, 3C data must be corroborated by other complementary means such as 3D DNA FISH, in order to have a comprehensive view of chromatin organization and the frequency of contact occurrence per individual cells in the population. The importance of single-cell analysis in 3C experiments is highlighted by recent observations made using single-cell Hi-C, which revealed a high variability of TAD structures among different cells (Nagano *et al.*, 2013). This is in striking contrast to what had been previously reported on TAD structures and their high evolutionary conservation (Dixon *et al.*, 2012).

Refinement of 3C technologies and future live-cell imaging studies will undoubtedly unify these apparently different results, and take into account the high-variability of genomic loci positioning inside the nucleus (Bolzer *et al.*, 2005; Soutoglou and Misteli, 2007; Chuang and Belmont, 2007).

## 1.4 Copy number variation in mammalian genomes

Since the early 1920's, changes in the karyotypic composition of genomes were known to occur. Work by A.F. Blakeslee in the jimson weed *Datura stramonium* revealed changes in phenotypic characters (leaves and capsule shapes) associated with changes in chromosome number (Blakeslee, 1922). Ten year later, Calvin Bridges reported the duplication of the Bar gene in *Drosophila melanogaster*, which was linked to the reduced-eye mutant phenotype (Bridges, 1936). Subsequent cytogenetic studies in humans ensued, linking specific genetic/genomic disorders and mental retardation syndromes to changes in chromosomal ploidy and DNA duplications and deletions (Jacobs *et al.*, 1959, 1978, 1992; Edwards *et al.*, 1960; Patau *et al.*, 1960; Coco and Penchaszadeh, 1982; reviewed in Lupski, 1998; reviewed in Stankiewicz and Lupski, 2002). Very interestingly, it was also observed that few cases of naturally occurring gene number variations occurred, without major consequences on a person's phenotype (Groot, Mager, and Frants, 1991; Trask *et al.*, 1998; Hollox, Armour, and Barber, 2003).

In 2004, the use of microarray technologies for the detection of DNA aberrations in clinical samples led to the discovery of copy number variation. This phenomenon was simultaneously published by two groups while examining human genome sequences using array comparative genomic hybridization aCGH (Iafrate *et al.*, 2004; Sebat *et al.*, 2004). In these studies, large-scale amplification and deletion differences were detected in genomes of healthy individuals from diverse populations, and these changes were common and present in a wide-range of genomic locations, including coding regions. Although large chromosomal duplications and deletions had been previously detected by cytogenetic observations, their frequency of occurrence was low and mostly related to disease phenotypes. However, further

genomic comparisons using diverse methodologies pointed to copy number alterations as highly frequent, associated with genomic features such as segmental duplications, and shared among several human populations (Tuzun *et al.*, 2005; Sharp *et al.*, 2005).

Nomenclature for identification of these changes was standardized, and the identified regions were called copy number variants (CNVs), defined as segments >1kb in size and present at variable copy numbers compared to a reference genome (Feuk *et al.*, 2006). CNVs comprise, together with insertions, inversions, and translocations, the “structural variation” of the human genome, whose contribution to sequence heterogeneity makes them important components of human genetic diversity and disease susceptibility.

After their initial discovery, several studies focusing on the characterization of CNVs in diverse human populations ensued (Conrad *et al.*, 2006; McCarroll *et al.*, 2006; Hinds *et al.*, 2006), including a comprehensive analysis of 270 individuals from the HapMap project (Redon *et al.*, 2006; Conrad *et al.*, 2010). Additionally, with the peak of usage of sequencing technologies, several groups further expanded the thus available catalogue of CNVs in humans at a much higher resolution (Korbel *et al.*, 2007; Alkan *et al.*, 2009; Chen *et al.*, 2009; Hormozdiari *et al.*, 2009; McKernan *et al.*, 2009; Sudmant *et al.*, 2010; International HapMap 3 Consortium, 2010; 1000 Genomes Project, 2011).

It was observed that genomic CNVs arise by various mechanisms, including homologous and non-homologous recombination coupled to replicative and non-replicative DNA processes (reviewed in Hastings *et al.*, 2009). Because recombination is a basic molecular mechanism, CNVs were presumed to be important players in eukaryotic evolution and originators of phenotypic variation. In fact, besides humans, CNVs have also been detected in *Drosophila* (Dopman and Hartl, 2007), mouse (Egan *et al.*, 2007; Graubert *et al.*,

2007; She *et al.*, 2008; Cahan *et al.*, 2009), rat (Guryev *et al.*, 2008), dogs (Chen *et al.*, 2009), pigs (Ramayo Caldas *et al.*, 2010), goats (Fontanesi *et al.*, 2010), rhesus macaque (Lee *et al.*, 2008; Gokcumen *et al.*, 2011; Iskow *et al.*, 2012; Gokcumen *et al.*, 2013), chimpanzee (Perry *et al.*, 2006; Perry *et al.*, 2008; Gazave *et al.*, 2011; Gokcumen *et al.*, 2011; Iskow *et al.*, 2012; Gokcumen *et al.*, 2013), orangutan (Gazave *et al.*, 2011; Gokcumen *et al.*, 2013), bonobo and gorilla (Gazave *et al.*, 2011), and plants (Schnable *et al.*, 2009; DeBolt, 2010; McHale *et al.*, 2012; Muñoz-Amatriaín *et al.*, 2013), with diverse species-specific phenotypic associations. The importance of CNVs in primate evolution has been highlighted, due to their roles in the adaptive phenotypic differences between humans and apes by alterations of gene families and gene expression phenotypes (Perry *et al.*, 2008; McLean *et al.*, 2011; Iskow *et al.*, 2012; Gokcumen *et al.*, 2013).

To date, thousands of CNVs have been identified in human (consult the Database of Genomic Variants for a comprehensive list of available CNVs), containing hundreds of genes and disease loci, segmental duplications, and revealing population-specific CNVs and genetic linkage disequilibrium, therefore making CNVs an important resource for genetic disease studies.

In addition to the already characterized repertoire of deletion/duplication syndromes (reviewed in Lupski, 1998, and Stankiewicz and Lupski, 2002), dozens of human diseases have been linked to CNVs, either inherited or arising by *de novo* germline/somatic mutations. Complex diseases such as autism spectrum disorders (Sebat *et al.*, 2007; Pinto *et al.*, 2010; Sanders *et al.*, 2011; Levy *et al.*, 2011; Gilman *et al.*, 2011), schizophrenia (Stefansson *et al.*, 2008; McCarthy *et al.*, 2009; reviewed in Hosak, 2013), Crohn's disease (McCarroll *et al.*, 2008; Craddock *et al.*, 2010), rheumatoid arthritis and types 1 and 2 diabetes (Craddock *et*



*al.*, 2010), psoriasis (de Cid *et al.*, 2009), osteoporosis (Yang *et al.*, 2008), glomerulonephritis (Aitman *et al.*, 2006), as well as a myriad of different cancer types (Greenman *et al.*, 2007; Stephens *et al.*, 2009; Campbell *et al.*, 2010; Lee *et al.*, 2010; Pleasance *et al.*, 2010; Berger *et al.*, 2011; Hillmer *et al.*, 2011; Khurana *et al.*, 2013; Yang *et al.*, 2013; reviewed in Meyerson, Gabriel, and Getz, 2010; reviewed in Hanahan and Weinberg, 2011) have been associated to CNVs. Additionally, various CNVs have been shown to play roles in normal phenotypic variability, like male testosterone metabolism (Jakobsson *et al.*, 2006), reduced susceptibility to human immunodeficiency virus (HIV) infection (Gonzalez *et al.*, 2005), and amylase copy-number correlations to starch diet (Perry *et al.*, 2007).

CNVs can give rise to different phenotypes through several mechanisms. For example, CNVs can alter gene dosage (gene deletion/duplication), unmask recessive alleles or functional SNPs, disrupt gene promoters and regulatory elements associations, promote gene fusions, among others. Early studies of CNVs impact on genome-wide expression revealed a positive correlation with transcription, however, for 20% of the assayed regions the correlation went in the opposite direction (Stranger *et al.*, 2007), revealing the heterogeneous impact of CNVs presence on genomic function. Interestingly, two subsequent studies reported that CNVs and gene dosage relationships largely deviate from their expected linear ratios compared to wild-type genotypes (Schuster-Böckler, Conrad, and Bateman, 2010; Schlattl *et al.*, 2011), suggesting dosage compensation, thus adding a new layer of complexity in CNV-transcription analysis.

To date, few studies have carefully assessed CNV-gene expression relationships in human (Stranger *et al.*, 2007; Schuster-Böckler, Conrad, and Bateman, 2010; Schlattl *et al.*,

2011), human-ape comparisons (Iskow *et al.*, 2012; Gokcumen *et al.*, 2013), and mice (Cahan *et al.*, 2009; Orozco *et al.*, 2009). However, while these studies have described genome-wide or gene-specific expression-CNV correlations, none has addressed the alteration of chromatin structure and subsequent transcriptional impact after DNA deletion or amplification events. Based on this information, we set out to molecularly and microscopically characterize a common CNV and its impact in both chromosome architecture and transcriptional output.

## **1.5 Characterization of higher-order chromatin organization at the mouse region 4E2**

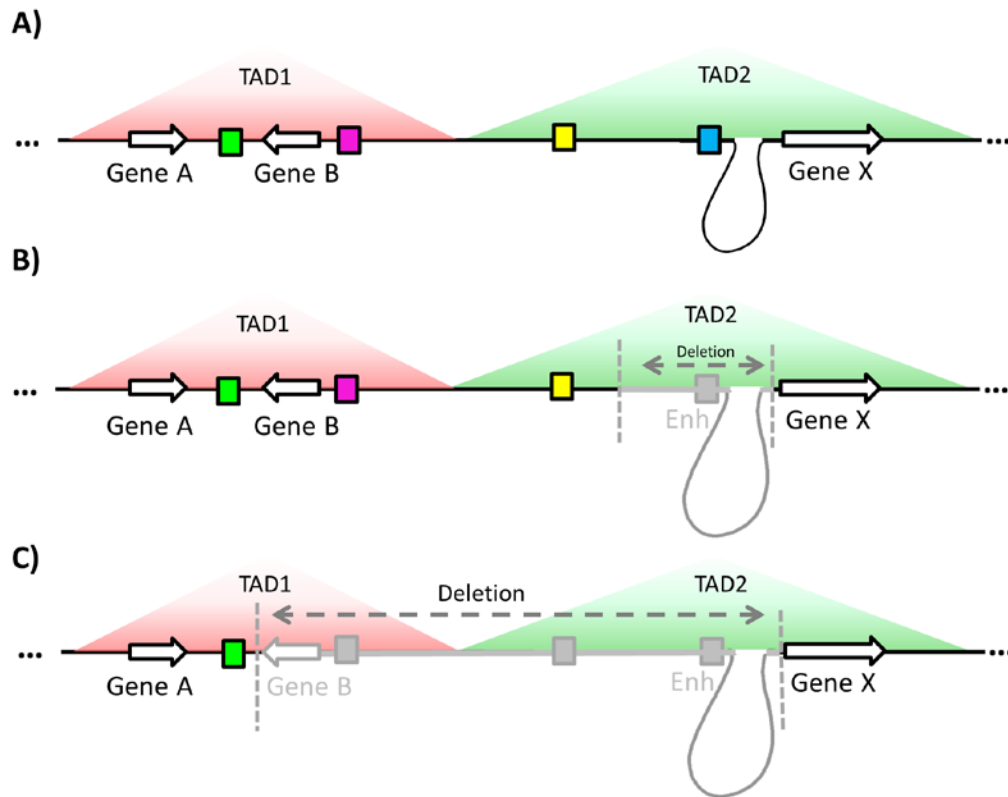
As I have described through the first sections of this chapter, chromatin organization in eukaryotic cells is an important feature in large-scale condensation of the genome, genomic stability, and gene expression regulation. Recent technological advances have allowed a more detailed study of chromosome structure. The development of the 3C approach, a technique that allows the detection of physical chromatin interactions between genomic elements, and all 3C-derived methodologies, have provided a clearer picture of overall and locus-specific genome conformations and the importance of chromatin interactions in quantitatively and temporally controlling gene expression.

Because of its role as a mode of transcriptional control, disruption of regulatory chromatin interactions due to genomic recombination can have pathological implications by altering gene expression patterns of genes surrounding the rearrangement. Genome-wide studies performed on cell lines from the HapMap project revealed widespread genetic associations of CNVs and gene expression changes in *cis* over large genomic distances

(Stranger *et al.*, 2007). Two other studies have also reported altered expression of diploid genes up to half or 6.5 Mb away from the breakpoints of deletions that cause Williams-Beuren syndrome (Merla *et al.*, 2006), and Smith-Magenis and Potocki-Lupski syndromes (Ricard *et al.*, 2010). These observations have led to the hypothesis that CNVs have a complex effect on gene transcription that might involve altered long-range chromatin organization.

At the chromatin level, CNVs can potentially disrupt associations of gene promoters and their regulatory elements [Fig. 1.8A,B], affect the positioning of preferred regulatory elements of genes, or affect TAD boundaries and fuse two differentially regulated chromatin regions [Fig. 1.8C], which could have many important functional and pathological implications. Therefore, a wider understanding of changes in chromatin architecture upon recombination will provide more insights into the basic principles of chromosome conformation, its alteration upon sequence disruption, and its functional impact on cellular transcriptional status. For that reason, it is the purpose of my thesis research to characterize in detail the higher-order chromatin organization of a genomic region associated with recurrent recombination in its diploid state and upon copy-number variations.

To this end, I selected the mouse 4C6-E2 region for CNV-chromatin organization studies. Mouse 4C6-E2 bands are syntenic to human 1p32.1-36 bands. 1p36 deletions are relatively common CNVs in the human genome, often present in a wide variety of cancers (reviewed in Bagchi and Mills, 2008), and originating a mental retardation syndrome known as “Monosomy 1p36” (reviewed in Slavotinek, Shaffer, and Shapira, 1999).



**Figure 1.8 Selected examples of the different ways in which CNVs can affect chromatin organization and gene expression**

**A)** Structure of a representative genomic region in which there exist 2 TADs (marked with the red and green triangles), 3 genes (marked with empty arrows), and several regulatory elements (colored rectangles). There is a looping interaction between Gene X and a regulatory element marked by the blue box. **B)** After the occurrence of a deletion CNV (grey arrow), the looping interaction is lost due to the loss of the DNA sequence, disrupting the association of Gene X's promoter with its regulatory element. **C)** In the case of the occurrence of a bigger deletion, the included TAD boundary between TADs 1 and 2 is lost, fusing two differentially regulated chromatin regions. Although only deletions were discussed in this figure, duplications can have the same effects.

I used two chromosome-engineered CNV models of the 4E2 bands, provided by Alea Mills, CSHL. These are  $df/+^{Bl6}$ , a heterozygous mouse strain with a 4.3Mb deletion in the 4E22 band, and  $dp/+^{Bl6}$ , harboring a duplication of the same region (Bagchi *et al.*, 2007) (see Chapter 2 for an extensive description of the mouse models used).

The present thesis research focuses on answering specific questions related to chromatin organization and genomic transcriptional state on 4E2 CNV engineered mouse models, divided into three experimental aims:

**1. Microscopic characterization of higher-order chromatin organization of 4E22 and neighboring regions in  $+^{129}/+^{Bl6}$ ,  $df/+^{Bl6}$ , and  $dp/+^{Bl6}$  MEFs.** What are the chromatin compaction status, nuclear positioning, and overlap with heterochromatin foci of a DNA region after the occurrence of a CNV compared to its WT state?

**2. Molecular characterization of higher-order chromatin organization of 4E22 and neighboring regions in  $+^{129}/+^{Bl6}$ ,  $df/+^{Bl6}$ , and  $dp/+^{Bl6}$  MEFs.** What is the chromatin conformation status of a unique region of DNA after the occurrence of a CNV compared to its WT state? What is the chromatin conformation status of neighboring regions of a CNV compared to its WT states? To what extent has chromatin architecture changed in regions bordering CNVs compared to WT regions? Can the observed variation be modeled using polymer physics and/or fitted into current genome conformation frameworks?

**3. Characterization of gene expression states in  $+^{129}/+^{Bl6}$ ,  $df/+^{Bl6}$ , and  $dp/+^{Bl6}$  MEFs.** What are the overall and allele-specific changes in expression after the occurrence of CNVs

compared to WT? Are changes in chromatin organization associated with gene expression differences? Is dosage compensation detected for CNV-associated genes?

Altogether, analysis of chromatin organization in the 4E2 region allowed us to investigate the state of gene-gene interactions in a WT 4E2 region compared to CNV zones, survey how the presence of CNVs altered preferred conformation states of 4E2 neighboring genes, and determine whether chromatin re-organization played a role in differential expression events in 4E2 CNV regions. Results will be presented and further discussed in the following chapters.

## Chapter 2: CNV mouse models of 4E2

The study of chromatin organization after copy number alterations calls for the use of a specific system in which the start and end of the CNVs are known for the design of both microscopic and molecular experiments. Additionally, selecting a frequent human CNV could potentially yield biological insights into its functional impact on genomic function.

As stated in the introduction, we analyzed the mouse syntenic region of human 1p36. There are two main reasons for choosing this region for our particular study: firstly, deletions of 1p36 are a relatively common chromosome abnormality (Heilstedt *et al.*, 2003; Bagchi and Mills, 2008, and references therein), and secondly, because of the availability of a chromosomally-engineered deletion and duplication (*df/dp*) 4E2 mouse strain (Bagchi *et al.*, 2007). The *df/dp* and derived heterozygote lines had been previously characterized and successfully used for the discovery of novel tumor suppressors in the region (Bagchi *et al.*, 2007). Such well-studied systems were therefore best suited for our analyses, and especially interesting given the high frequency of 1p36 CNVs.

### 2.1 Human region 1p36, CNVs, and their roles in disease.

Deletions of the region 36 on the short arm of chromosome 1 are common chromosome abnormalities in the human genome (Heilstedt *et al.*, 2003; Bagchi and Mills, 2008, and references therein). They are often present in a wide variety of cancers, including acute myeloid leukemia (AML), chronic myelogenous leukemia (CML), melanoma, pheochromocytoma, oligodendroglioma, neuroblastoma, meningioma, and non-Hodgkinis lymphoma, as well as thyroid, colorectal, breast, and cervical cancers (Li *et al.*, 2001; Bagchi

*et al.*, 2007; Midorikawa *et al.*, 2009; Zhang *et al.*, 2010; reviewed in Bagchi and Mills, 2008).

In addition to being a persistent cancer CNV, 1p36 deletions generate a syndrome known as “Monosomy 1p36,” a congenital genetic disorder characterized by mental retardation, developmental delay, hypotonia, and dysmorphic facial features (reviewed in Slavotinek, Shaffer, and Shapira, 1999). The syndrome incidence is high, about 1 case in every 5,000 to 10,000 births, making it one of the most common *de novo* deletion syndromes (Heilstedt *et al.*, 2003; Rosenfeld *et al.*, 2010).

Deletions causing Monosomy 1p36 do not have common breakpoints or sizes, but are usually located towards the terminal part of 1p36. Interstitial deletions and other complex rearrangements have also been observed (Rosenfeld *et al.*, 2010). Detailed analyses of several of these deletion CNVs have pinpointed the critical regions for some features of the syndrome (Zhu *et al.*, 2013; Arndt *et al.*, 2013; Kim *et al.*, 2013), however, analyses using array comparative genomic hybridization (aCGH) revealed two patients with different deletion sizes and positions who shared the same clinical features (Redon *et al.*, 2005). Although these phenotypes could be caused by microdeletions not detected by the then available aCGH technology, this observation suggests that 1p36 deletions might cause the disease features by their positional effects rather than by the contiguous gene deletions themselves. It is thus likely that besides altering the dosage of genes present in the region, the deletion is disrupting long-range chromosomal interactions that might be playing a role in gene expression regulation that may cause the different disease phenotypes.

Finally, 1p36 deletions have also been associated with other developmental delay phenotypes (Cooper *et al.*, 2011), and duplications with schizophrenia (Rees *et al.*, 2013).



Despite the frequency of 1p36 deletions in human disease, chromatin organization of the region after the occurrence of copy-number variation has not been characterized. It is likely that long-range chromatin interactions could influence the modulation of gene expression in the 1p36 region, especially that of developmental and tumor suppressor candidates. Such detailed analysis is possible in mice, given the availability of chromosomally-engineered strains targeting the syntenic region of 1p36.

## 2.2 Chromosomally engineered $df/+^{Bl6}$ and $dp/+^{Bl6}$ CNV mouse models

Mouse 4C6-E2 bands are syntenic to the 1p32.1-36 region in humans, and are approximately 60Mb in size. In order to examine 4E2 chromatin organization changes upon copy number variation, I used two chromosome-engineered CNV models of the 4E2 band. These are  $df$ , an engineered chromosome 4 harboring a 4.3Mb deletion in 4E22 (150-154.3Mb), and  $dp$ , harboring a duplication of the same region [Fig. 2.1A]. These chromosomes were originally described in Bagchi *et al.*, 2007, where the engineered  $df/dp$  strain and derived progeny were characterized in the context of cancer studies. The  $df/dp$  strain was kindly provided by Alea Mills, CSHL.

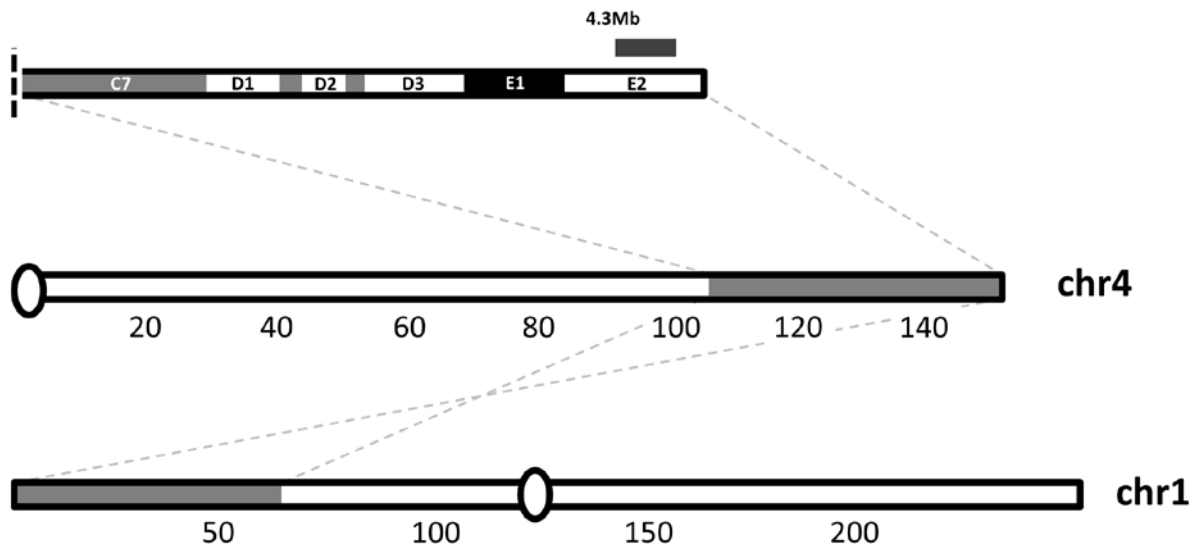
The  $df/dp$  strain described in Bagchi *et al.*, 2007, was engineered using 129S5/SvEv<sup>Brd</sup>-derived ES cells. By the time of the beginning of this project, the  $df/dp$  line had been crossed to the C57Bl6/J mouse strain for maintenance. With this breeding scheme, the resulting engineered chromosomes were meiotic recombination products of the 129S5 and C57Bl6/J lines, therefore confounding potentially useful SNPs for subsequent genotyping for the molecular 4C experiments (see Chapter 4). Moreover, as Monosomy 1p36

patients are heterozygous for this region (Heilstedt *et al.*, 2003), and heterozygous deletions in 1p36 are associated with cancer progression/maintenance (Bagchi and Mills, 2008, and references therein), there is a compelling need for the correct identification of the altered chromosome from its WT homologue to study CNVs in a functionally relevant scenario.

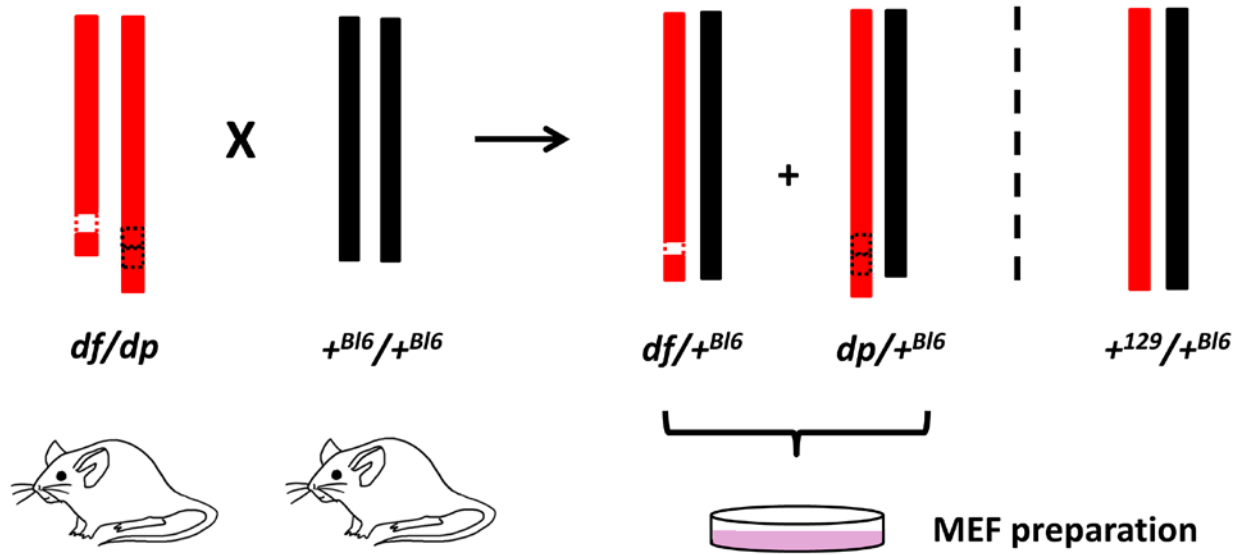
In order to obtain genetically homogeneous chromosomes for subsequent studies, new *df/dp* chimeras were generated for this project by injection of chromosomally engineered *df/dp* stem cells into C57Bl6/J blastocysts as described in Bagchi *et al.*, 2007. Thirteen males born from such clone injections were selected and mated with C57Bl6/J females to assess germline transmission of the engineered chromosomes. Of these, four chimeras had germline transmission, as genotyped by PCR of their offspring's tail DNA [Table 2.1]. A second generation of *df/dp* chimeras ensued for strain preservation, in which engineered *df/dp* stem cells were injected into C57Bl6/N blastocysts (albino). Two of these chimeras had confirmed germline transmission [Table 2.1].

To succeed in differentiating the CNV engineered chromosomes from their WT homologues to study chromatin organization and CNVs in their heterozygote state, germline transmitting chimeras were mated with C57Bl6/J females to obtain F1 *df/+<sup>Bl6</sup>* and *dp/+<sup>Bl6</sup>* embryos, where there is equal chromosome contribution of the 129S5/SvEv<sup>Brd</sup> and C57Bl6/J strains [Table 2.1 and Fig. 2.1B]. From this section and throughout the rest of the chapters, *+<sup>Bl6</sup>* corresponds to WT chromosome 4 from the C57Bl6/J background, and *+<sup>129</sup>* is WT chromosome 4 from 129S5/SvEv<sup>Brd</sup>.

A)



B)



**Figure 2.1 Engineered *df* and *dp* chromosomes**

**A)** Schematic depiction of Giemsa banding scheme of the location of the engineered 4.3Mb segment in mouse chromosome 4, and its correspondence in human chromosome 1. **B)** Breeding scheme for *df/dp* chimeras and the generation of *df/+<sup>Bl6</sup>*, and *dp/+<sup>Bl6</sup>* embryos.

F1 crosses between 129S5/SvEv<sup>Brd</sup> and C57Bl6/J mice ( $+^{129}/+^{Bl6}$ ) were used as WT controls for all experiments performed. The 129S5/SvEv<sup>Brd</sup> inbred mouse strain was obtained from the laboratory of Allan Bradley at the Wellcome Trust Sanger Institute, UK (4 females and 4 males received in 6/29/11, D.O.B. 5/14/11), while C57Bl6/J females ~6 weeks old were purchased from Taconic as needed for breeding.

In agreement to what had been previously reported,  $df/df$ ,  $dp/dp$ , and  $dp/+^{Bl6}$  genotypes were embryonic lethal (Bagchi *et al*, 2007), and pups from these genotypes were never observed in our breeding history. In addition,  $dp/+^{Bl6}$  embryos had developmental defects compared to  $df/+^{Bl6}$  and  $+^{129}/+^{Bl6}$  [Fig. 2.2]. Very interestingly, mating  $df/dp$  chimeras to 129S5/SvEv<sup>Brd</sup> females only produced a single  $df/+^{129}$  mouse in almost two years of continuous breeding. Even after female super-ovulation treatments, no other  $df/+^{129}$  or  $dp/+^{129}$  embryos were obtained. Curiously, the single  $df/+^{129}$  mouse that was obtained had to be sacrificed due to various phenotypic abnormalities (hunched back, crisped hair, small size, malocclusion, and mild conjunctivitis), suggesting strain-specific genetic background dependencies.

MEFs were derived from 13.5 day embryos of  $+^{129}/+^{Bl6}$ ,  $df/+^{Bl6}$ , and  $dp/+^{Bl6}$  genotypes, and used for the whole study (see protocol details in Chapter 8). Morphologically,  $dp/+^{Bl6}$  MEFs tend to have larger nuclear volumes (~260 $\mu$ m<sup>3</sup> difference) compared to  $df/+^{Bl6}$  and  $+^{129}/+^{Bl6}$  [Fig. 2.3], and halt growth after passage 5 in culture. To avoid issues with cellular senescence and the use of apoptotic cells, passage 4 MEFs (P4) were used for all experiments.

Embryo number	Parents		df/+	dp/+	male D.O.B.	female D.O.B.
	Male	Female				
From 1 to 6	54 Cre H3.3 ( <i>df/dp</i> )	129S5 WT	0	0	6/30/2011	9/8/2011
From 7 to 13	54 Cre H8 2A ( <i>df/dp</i> )	129S5 WT	0	0	6/30/2011	9/8/2011
From 14 to 18	129S5 (WT)	129S5 WT	0	0	5/14/2011	9/8/2011
From 19 to 26	129S5 (WT)	C57Bl6/J WT	0	0	9/8/2011	12/26/2011
From 27 to 31	54 Cre H3.3 ( <i>df/dp</i> )	C57Bl6/J WT	0	129S5E29	6/30/2011	12/12/2011
From 32 to 41	54 Cre H3.3 ( <i>df/dp</i> )	C57Bl6/J WT	129S5E36	129S5E32,129S5E39	6/30/2011	12/12/2011
From 42 to 49	54 Cre H3.3 ( <i>df/dp</i> )	C57Bl6/J WT	0	0	6/30/2011	2/27/2012
From 50 to 57	54 Cre H8 2A ( <i>df/dp</i> )	C57Bl6/J WT	129S5E56	0	6/30/2011	12/12/2011
From 58 to 66	54 Cre H8 2A ( <i>df/dp</i> )	C57Bl6/J WT	0	129S5E60, 129S5E61	6/30/2011	3/5/2012
From 67 to 74	54 Cre H2 3.2 ( <i>df/dp</i> )	C57Bl6/J WT	129S5E71	0	6/30/2011	3/5/2012
From 75 to 86	54 Cre H2 3.2 ( <i>df/dp</i> )	C57Bl6/J WT		129S5E77,80,81	6/30/2011	3/5/2012
From 87 to 96	129S5 (WT)	C57Bl6/J WT	0	0	9/8/2011	5/5/2012
From 97 to 103	54 Cre H2 3.2 ( <i>df/dp</i> )	C57Bl6/J WT	129S5E98	129S5E97,99	6/30/2011	4/23/2012
From 104 to 113	54 Cre H2 3.2 ( <i>df/dp</i> )	C57Bl6/J WT	0	0	6/30/2011	5/7/2012
From 114 to 120	129S5 (WT)	C57Bl6/J WT	0	0	3/28/2012	3/16/2012
From 121 to 126	54 Cre H3.3 ( <i>df/dp</i> )	C57Bl6/J WT	0	0	6/30/2011	3/5/2012

**Table 2.1 Breeding history for MEF generation.**

Mating history of three germline *df/dp* transmission chimeras with C57Bl6/J females and derived heterozygote progeny



**Figure 2.2 Representative  $+^{129}/+^{Bl6}$ ,  $df/+^{Bl6}$ , and  $dp/+^{Bl6}$  13.5 days embryos**

These embryos were derived from a mating between a *df/dp* chimeric male and a WT C57Bl6/J female. Notice the limbs, cranial, and overall developmental size abnormalities of *dp/+<sup>Bl6</sup>* embryos compared to  $+^{129}/+^{Bl6}$  and  $df/+^{Bl6}$  (for detailed information on  $dp/+^{Bl6}$  embryo phenotypes, see Bagchi *et al.*, 2007).

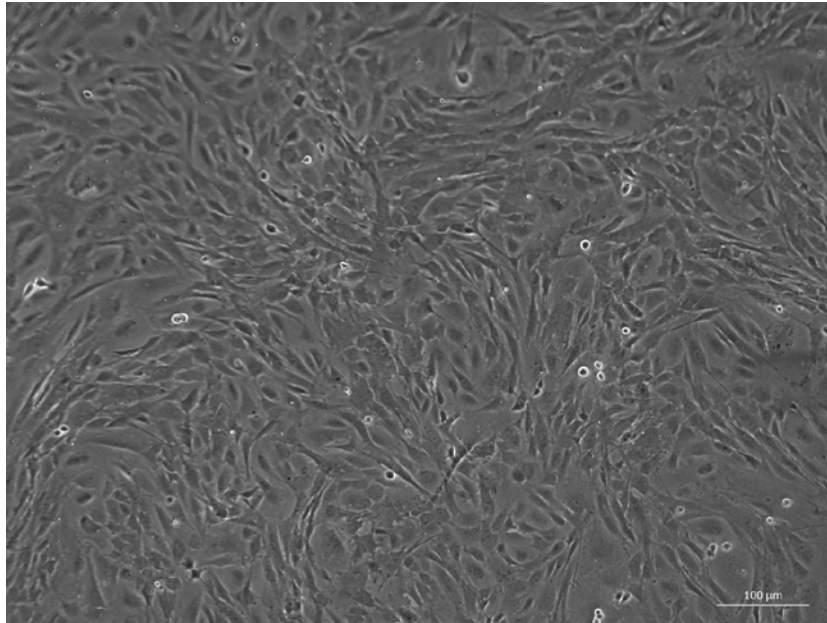
In order to avoid biases originated by differential cell cycle stages of the cultures, MEF plates were used after they had reached confluency. We measured cell cycle stage through DNA by flow cytometry in confluent P4 plates for all genotypes [Fig. 2.4; see protocol details in Chapter 8), and decided to use cells 10 hours after they had reached confluency for all experiments to allow remaining dividing cells to finish their cell cycle (final >90% cells in G0/G1 phase).

### **2.3 Genomic characteristics of mouse chromosome 4 and the 4E2 engineered region**

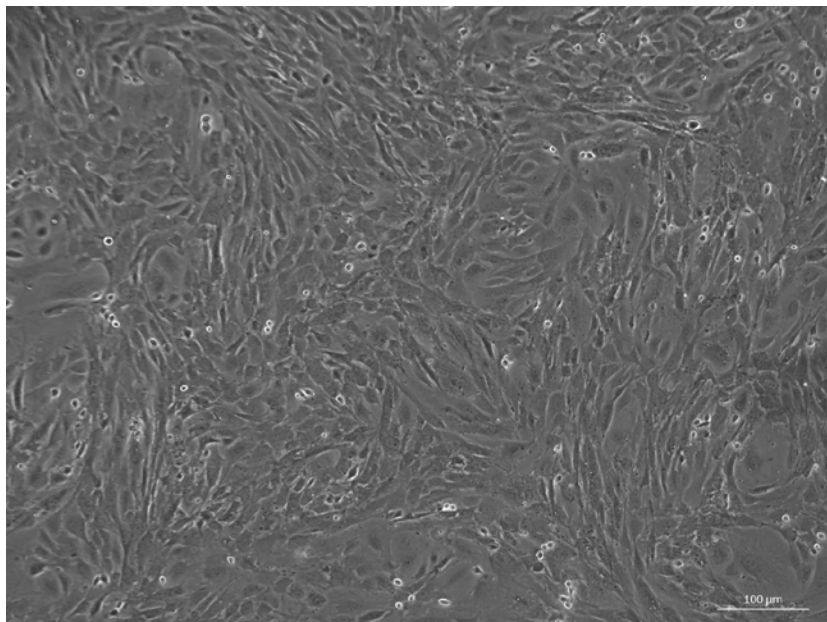
The UCSC mm9 genome assembly was used for all genomic analyses in this project. Under this genome version, the reference C57Bl6/J mouse chromosome 4 has a size of 155,630,120 bp. 2,027 RefSeq genes have been annotated in this chromosome, covering 58,602,429 bp (~38%) of its sequence, while Ensembl annotation found 2,383 genes, covering 60,451,615bp (~38%) of chromosome 4 sequence [Fig. 2.5]. A total of 9,995 segmental duplications (SDs) are located in this chromosome (~6% sequence), while RepeatMasker elements are 299,884 in number and comprise 68,045,614 bp (~44% of chromosome sequence) [Fig. 2.5. Table 2.2]. Overall, mouse chr4 is syntenic to diverse tracks in human chromosomes 1, 6, 8, and 9 [Fig. 2.6]. The engineered region in 4E2 spans ~4.3Mb in total, starting at 150,078,960 bp and ending in 154,420,125 bp. This region is syntenic to human chr1: 2336241-8086393 as defined by the boundary genes *PEX10* and *ERRF1* located in 1p36 in the hg19 human genome version. The engineered region has 53 annotated RefSeq genes (2,314,508 bp) [Table 2.3], 4 SDs (15,194 bp), and 7,669 RepeatMasker elements (1,208,279 bp) [Fig. 2.5].



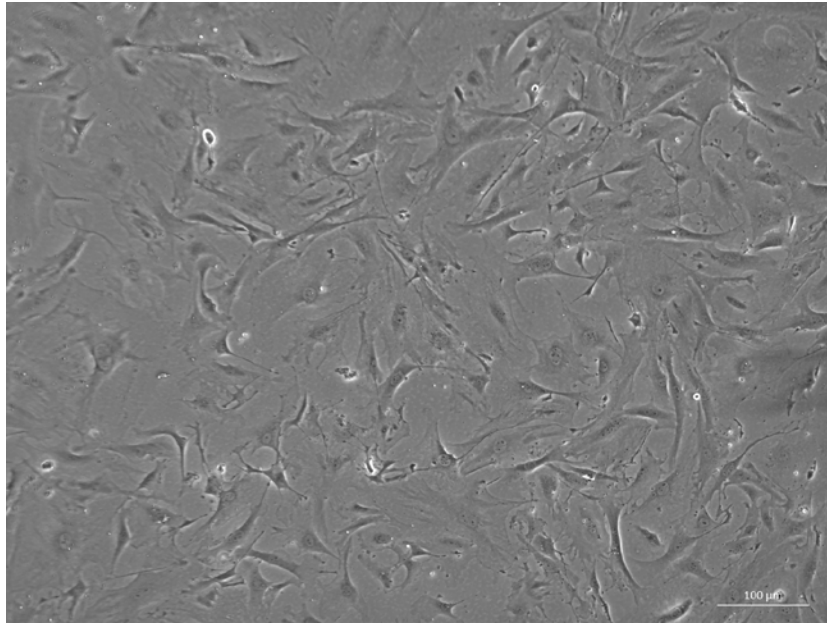
**A)**



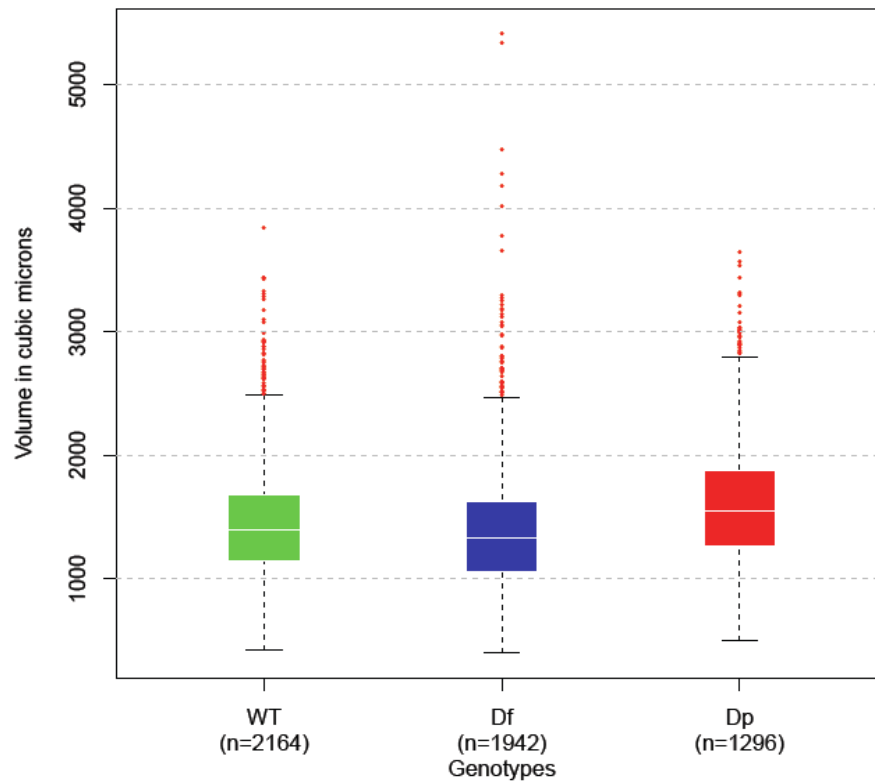
**B)**



C)



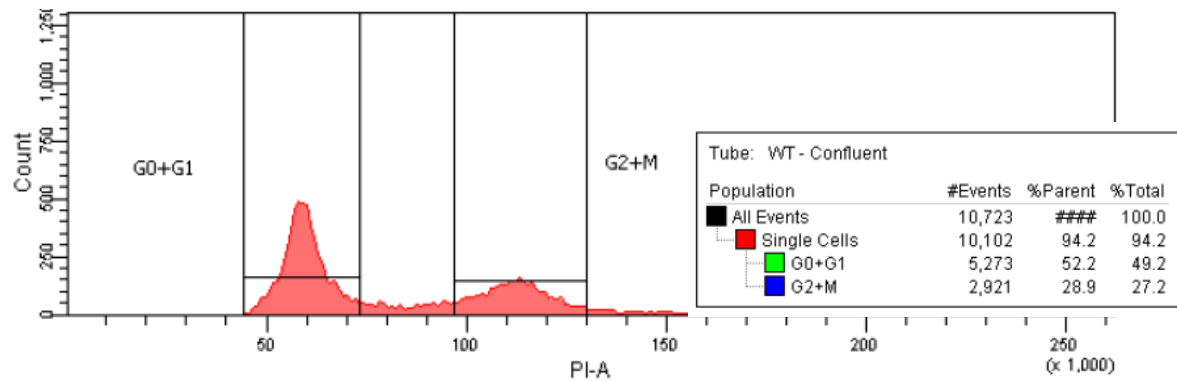
D)



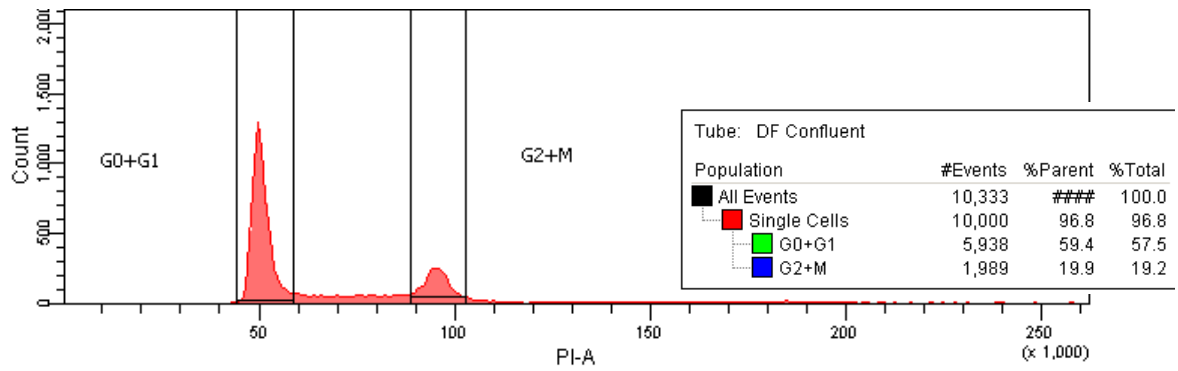
**Figure 2.3 Bright-field microscope images of the different MEF genotypes**

**A)**  $+^{129}/+^{Bl6}$ , **B)**  $df/+^{Bl6}$ , and **C)**  $dp/+^{Bl6}$ . **D)** Quantitation of nuclear volumes for  $+^{129}/+^{Bl6}$ ,  $df/+^{Bl6}$ , and  $dp/+^{Bl6}$  MEFs with measurements derived from 3D DNA FISH using an automated image analysis. For an extensive analysis of nuclear volume as well as other nuclear measurements, see Chapter 3.

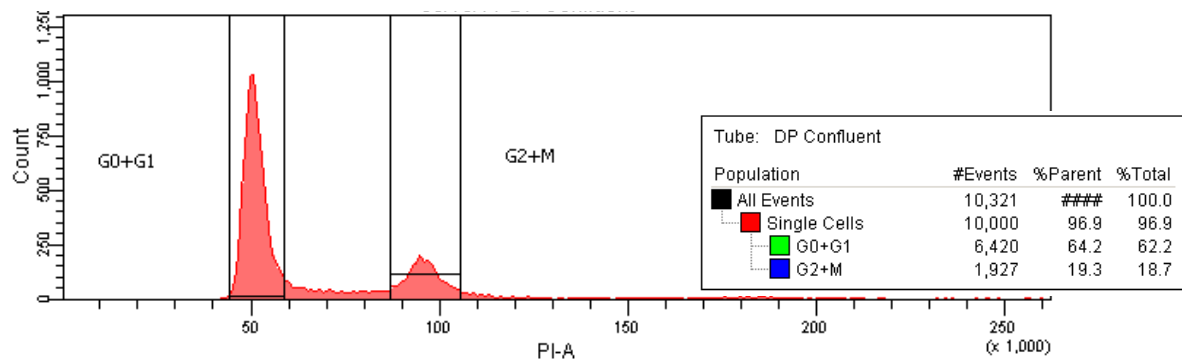
**A)**



**B)**



**C)**



**Figure 2.4 Confluent MEF FACs profiles**

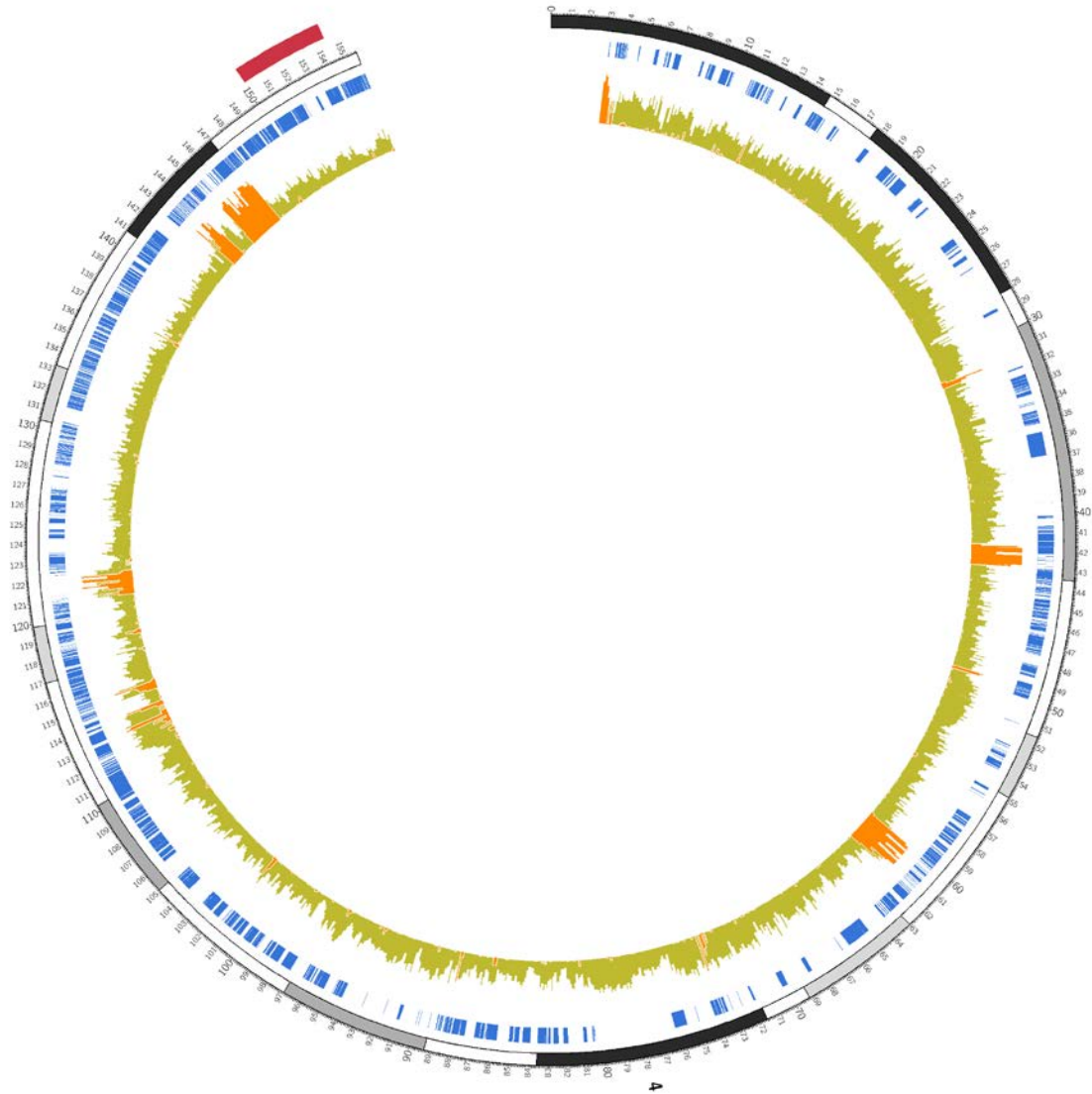
**A)**  $+^{129}/+^{Bl6}$ , **B)**  $df/+^{Bl6}$ , and **C)**  $dp/+^{Bl6}$  MEFs. To increase the number of cells at G0+G1 states, confluent plates were used ~10 hours after reaching confluency to allow remaining dividing cells to finish M or S phase.

## 2.4 129S5/SvEv<sup>Brd</sup> and C57Bl6/J chromosome 4 sequence analysis

The major scope of this thesis research is to determine changes in chromosomal architecture upon CNV events. As we are using two different mouse strains for the allele-specific identification of these conformations, an in-depth understanding is necessary regarding the chromosome 4 sequences of both 129S5/SvEv<sup>Brd</sup> and C57Bl6/J strains.

The Sanger Mouse Sequencing Consortium sequenced the 129S5/SvEv<sup>Brd</sup> strain genome with a 19X coverage (Keane *et al.*, 2011). 26,315 contigs were reported for the 129S5/SvEv<sup>Brd</sup> chromosome 4, where the smallest contig was 19 bp in size, the biggest was 211,300 bp, median size was 759bp, and the average size was 5,699 bp. Total contig bp sum for chr4 is 149,956,862 bp [Fig. 2.7].

To assess the accuracy of contig positions for 129S5/SvEv<sup>Brd</sup> chromosome 4 as reported by the Sanger Mouse Sequencing Consortium, an “assembled” 129S5/SvEv<sup>Brd</sup> chromosome 4 was constructed using the coordinates given for each contig. The total assembled length of 129S5/SvEv<sup>Brd</sup> chromosome 4 is 158,463,221bp. We generated pairwise alignments between the assembled 129S5/SvEv<sup>Brd</sup> chromosome 4 and the reference C57Bl6/J chromosome 4. We performed these alignments using nucmer (Kurtz *et al.*, 2004), a MUMmer 3.0 package program which allows the alignment of multiple reference and query sequences. To derive the maximally matched extended alignments, nucmer was run using a minimum cluster length of 500 bp, and using only anchor matches unique in the reference sequence after repeat masking the sequence.



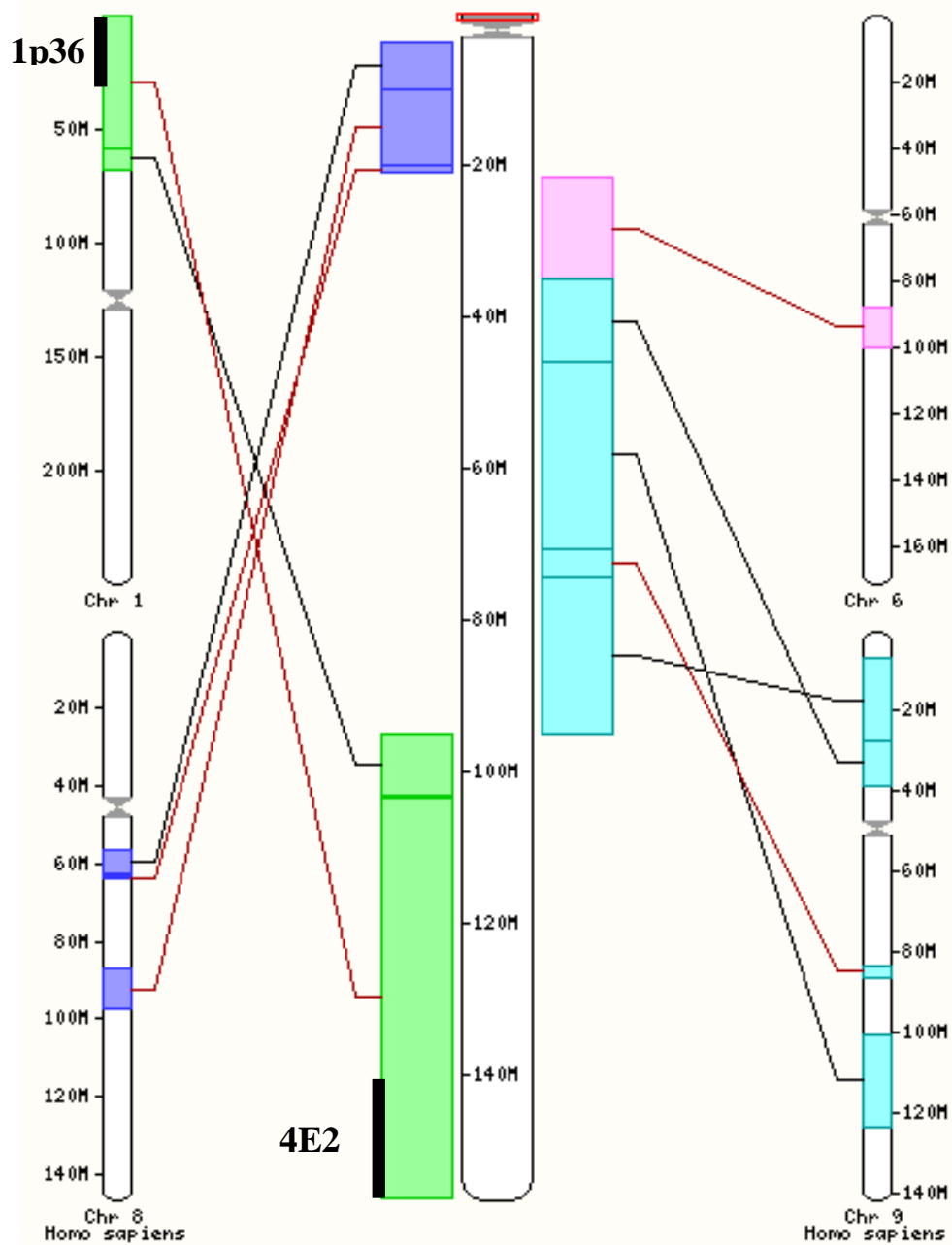
**Figure 2.5 Circular depiction of mouse chromosome 4**

Outer red box corresponds to the deletion CNV. This deletion spans ~4.3Mb in size, starting at 150,078,960 bp and ending in 154,420,125 bp (genome version mm9). Outer to innermost circles: Blue circle are annotated RefSeq genes. Dark green histogram represents the density in bp of RepeatMasker elements, and orange histogram represents SDs sequence density [Supp. Table 2.1].

Repeat class	Number of occurrences
snRNA	190
Other	1,125
rRNA	74
tRNA	278
LINE	55,996
scRNA	552
SINE	101,844
DNA	8,517
RNA	41
srpRNA	21
Low_complexity	21,205
Simple_repeat	61,043
Satellite	85
LTR	48,464
Unknown	419
RC	30
<b>Total</b>	<b>299,884</b>

**Table 2.2 Present RepeatMasker classes in mouse chromosome 4**





**Figure 2.6 Ensembl view of mouse chromosome 4 synteny to human chrs 1,6,8,9**

The 1p36 and 4E2 regions are marked with black bars. Image obtained from ensemble.org

Chr	Gene Start	Gene End	MGI symbol	WikiGene description
4	150228028	150243001	<i>Errfi1</i>	ERBB receptor feedback inhibitor 1
4	150271242	150288546	<i>Park7</i>	Parkinson disease (autosomal recessive, early onset) 7
4	150294299	150320211	<i>Tnfrsf9</i>	tumor necrosis factor receptor superfamily, member 9
4	150371206	150375919	<i>Uts2</i>	urotensin 2
4	150378821	150418698	<i>Per3</i>	period homolog 3 (Drosophila)
4	150421414	150432050	<i>Vamp3</i>	vesicle-associated membrane protein 3
4	150433634	151235985	<i>Camta1</i>	calmodulin binding transcription activator 1
4	151236299	151237413	<i>9230110K08Rik</i>	RIKEN cDNA 9230110K08 gene
4	151307840	151356062	<i>Dnajc11</i>	DnaJ (Hsp40) homolog, subfamily C, member 11
4	151356748	151363106	<i>Thap3</i>	THAP domain containing, apoptosis associated protein 3
4	151365257	151370292	<i>Phf13</i>	PHD finger protein 13
4	151383026	151391785	<i>Klhl21</i>	kelch-like 21 (Drosophila)
4	151393885	151401780	<i>Zbtb48</i>	zinc finger and BTB domain containing 48
4	151402023	151412677	<i>Tas1r1</i>	taste receptor, type 1, member 1
4	151413441	151435603	<i>Nol9</i>	nucleolar protein 9
4	151446607	151489509	<i>Plekhg5</i>	pleckstrin homology domain containing, family G member 5

4	151490188	151494219	<i>Tnfrsf25</i>	tumor necrosis factor receptor superfamily, member 25
4	151494977	151526331	<i>Espn</i>	espin
4	151532976	151536578	<i>Hes2</i>	hairy and enhancer of split 2 (Drosophila)
4	151552243	151645956	<i>Acot7</i>	acyl-CoA thioesterase 7
4	151648341	151659446	<i>Gpr153</i>	G protein-coupled receptor 153
4	151660081	151665771	<i>Hes3</i>	hairy and enhancer of split 3 (Drosophila)
4	151671393	151678137	<i>Icmt</i>	isoprenylcysteine carboxyl methyltransferase
4	151681132	151692717	<i>Rnf207</i>	ring finger protein 207
4	151699971	151706585	<i>Rpl22</i>	ribosomal protein L22 pseudogene
4	151712760	151764303	<i>Chd5</i>	chromodomain helicase DNA binding protein 5
4	151764853	151851589	<i>Kcnab2</i>	potassium voltage-gated channel, beta member 2
4	151852251	151937292	<i>Nphp4</i>	nephronophthisis 4 (juvenile) homolog (human)
4	152071772	152073211	<i>Gm833</i>	hypothetical protein LOC100044224
4	152747330	152856939	<i>Ajap1</i>	adherens junction associated protein 1
4	153331016	153331153	<i>BC049688</i>	
4	153331346	153336023	<i>A430005L14Rik</i>	RIKEN cDNA A430005L14 gene
4	153338564	153349235	<i>Dffb</i>	DNA fragmentation factor, beta subunit
4	153349320	153381922	<i>BC046331</i>	cDNA sequence BC046331

4	153385922	153395619	<i>Lrrc47</i>	similar to leucine rich repeat containing 47
4	153396778	153397095	<i>1190007F08Rik</i>	
4	153400753	153416786	<i>Ccdc27</i>	coiled-coil domain containing 27
4	153432953	153514317	<i>Trp73</i>	transformation related protein 73
4	153516483	153530924	<i>Wdr8</i>	WD repeat domain 8
4	153531597	153534775	<i>Tprgl</i>	transformation related protein 63 regulated like
4	153545133	153648558	<i>Megf6</i>	multiple EGF-like-domains 6
4	153652579	153674163	<i>Arhgef16</i>	Rho guanine nucleotide exchange factor (GEF) 16
4	153690234	154010982	<i>Prdm16</i>	PR domain containing 16
4	154040542	154041976	<i>Actrt2</i>	actin-related protein T2
4	154230336	154241234	<i>B230396O12Rik</i>	RIKEN cDNA B230396O12 gene
4	154243745	154269637	<i>Mme11</i>	membrane metallo-endopeptidase-like 1
4	154270539	154273152	<i>2810405K02Rik</i>	RIKEN cDNA 2810405K02 gene
4	154296319	154302186	<i>Tnfrsf14</i>	tumor necrosis factor receptor superfamily, member 14
4	154335030	154336477	<i>Hes5</i>	hairy and enhancer of split 5 (Drosophila)
4	154338242	154355047	<i>Pank4</i>	pantothenate kinase 4
4	154357235	154393351	<i>Plch2</i>	phospholipase C, eta 2

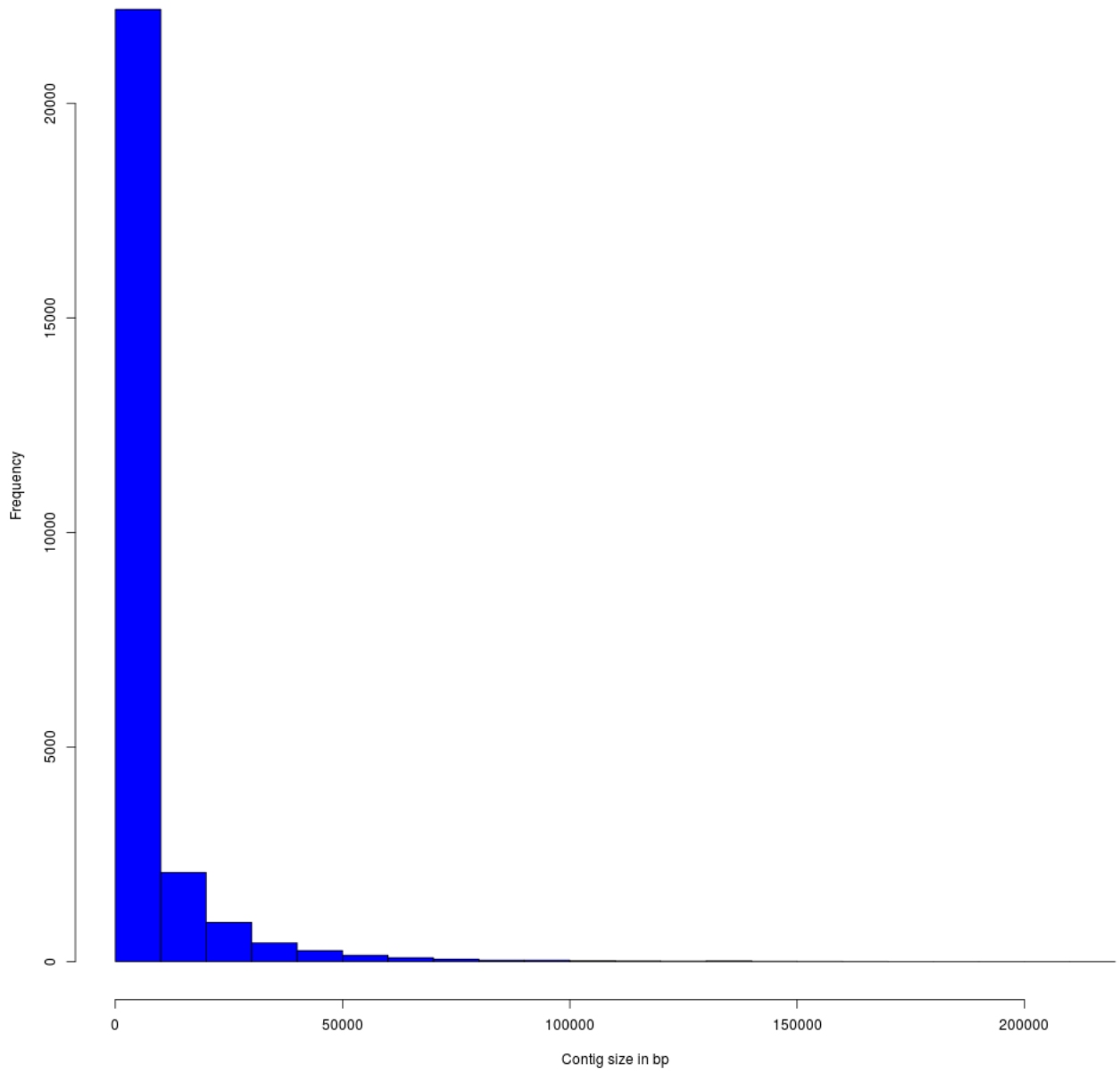
**Table 2.3 Annotated RefSeq genes inside the 4.3Mb engineered region**

Using only unique reference matches, 134,918,666bp of the assembled 129S5/SvEv<sup>Brd</sup> chromosome 4 aligns to C57Bl6/J chromosome 4 reference sequence using 500bp minimal cluster length. Based on the gaps information derived from Sanger mapping, there exist 26,314 gaps with respect to the chromosome 4 reference sequence, with a minimal length of 1 bp, a median of 23 bp, an average size of 101bp, and a maximal value of 50kb. The reported gaps total 2,673,245bp of sequence. As can be seen in Fig. 2.8, the assembled 129S5/SvEv<sup>Brd</sup> chromosome 4 reports the same results in terms of global sequence similarity to C57Bl6/J and no major rearrangements.

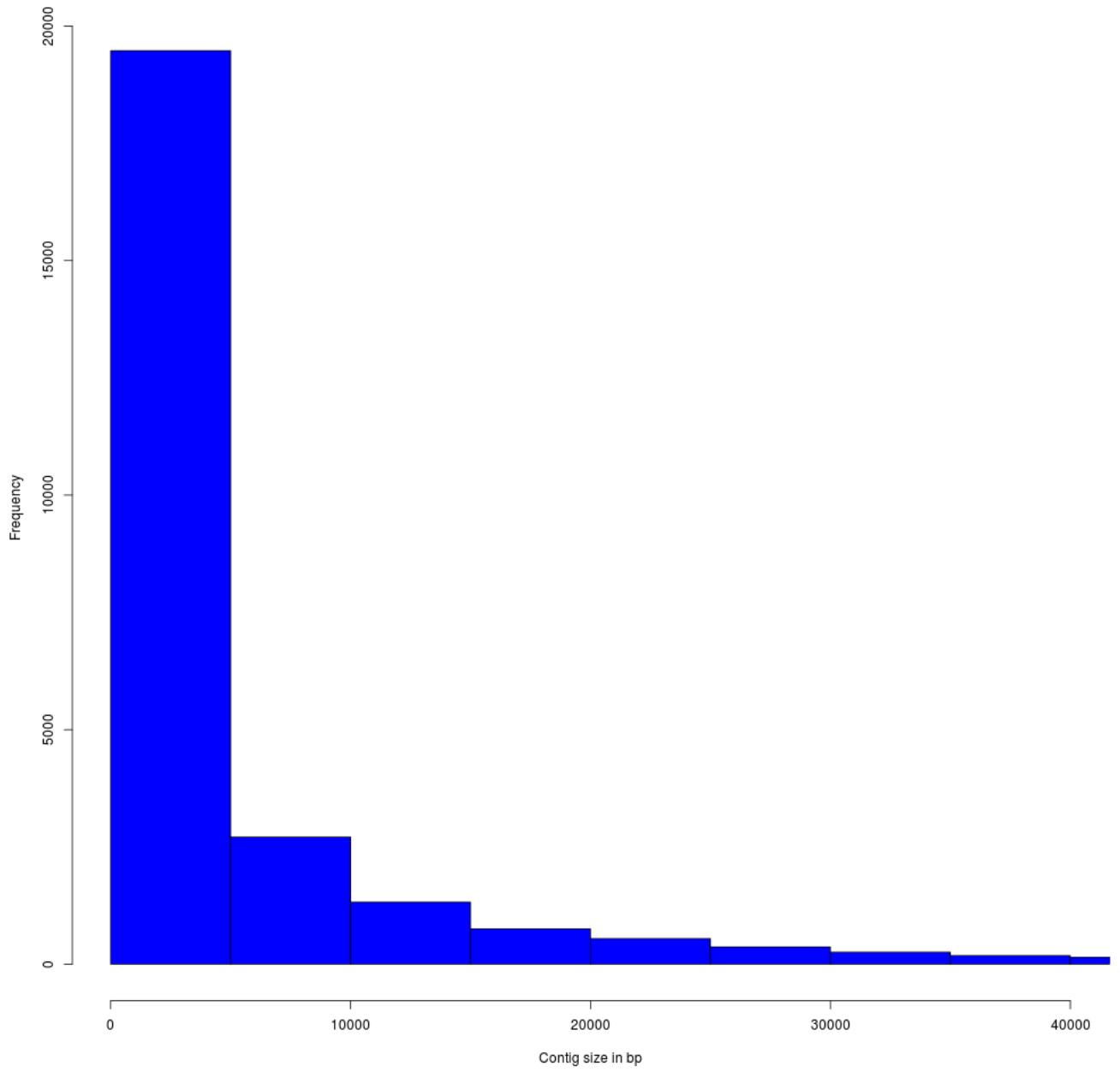
Zooming into the 4E2 region, from bases 147,000,000-155630120bp, there are 702 reported gaps with a minimal size of 1bp (single nucleotide polymorphisms, SNPs), a median size of 7bp, an average of 51bp, and a maximal of 1,313bp, adding up to 36,131bp in total. From the nucmer alignments, we can observe two segments located at ~144Mb and ~145-147Mb along the reference chromosome 4 for which 129S5/SvEv<sup>Brd</sup> sequence presents homology breaks and reverse alignment hits [Fig. 2.9]. These correspond to regions of enriched segmental duplications and simple repeated elements, therefore the lack of proper alignments. Upstream regions show no obvious changes in terms of structural variants (big inversions or deletions), which is optimal for design of molecular experiments for probing chromatin conformation.

Importantly, a total of 323,240 high-confidence SNPs were reported between the chr4 sequences of C57Bl6/J and 129S5/SvEv<sup>Brd</sup>. Of these, 379 are located inside the 4.3Mb engineered region [Fig. 2.10], which proved useful to our design of molecular conformation experiments (see Chapter 4).

A)



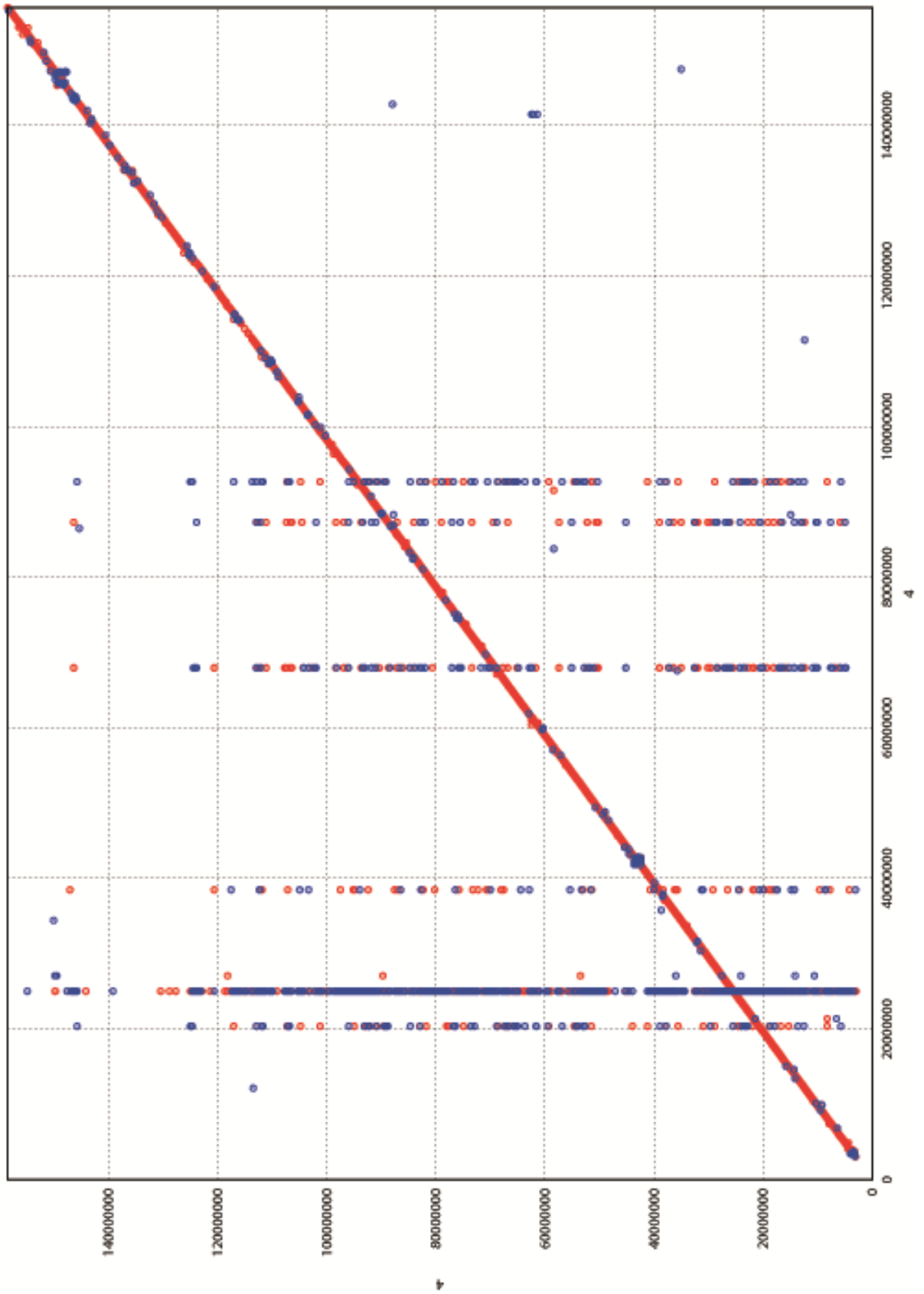
**B)**



**Figure 2.7 Contig size distributions for chromosome 4 sequence of 129S5/SvEv<sup>Brd</sup>**

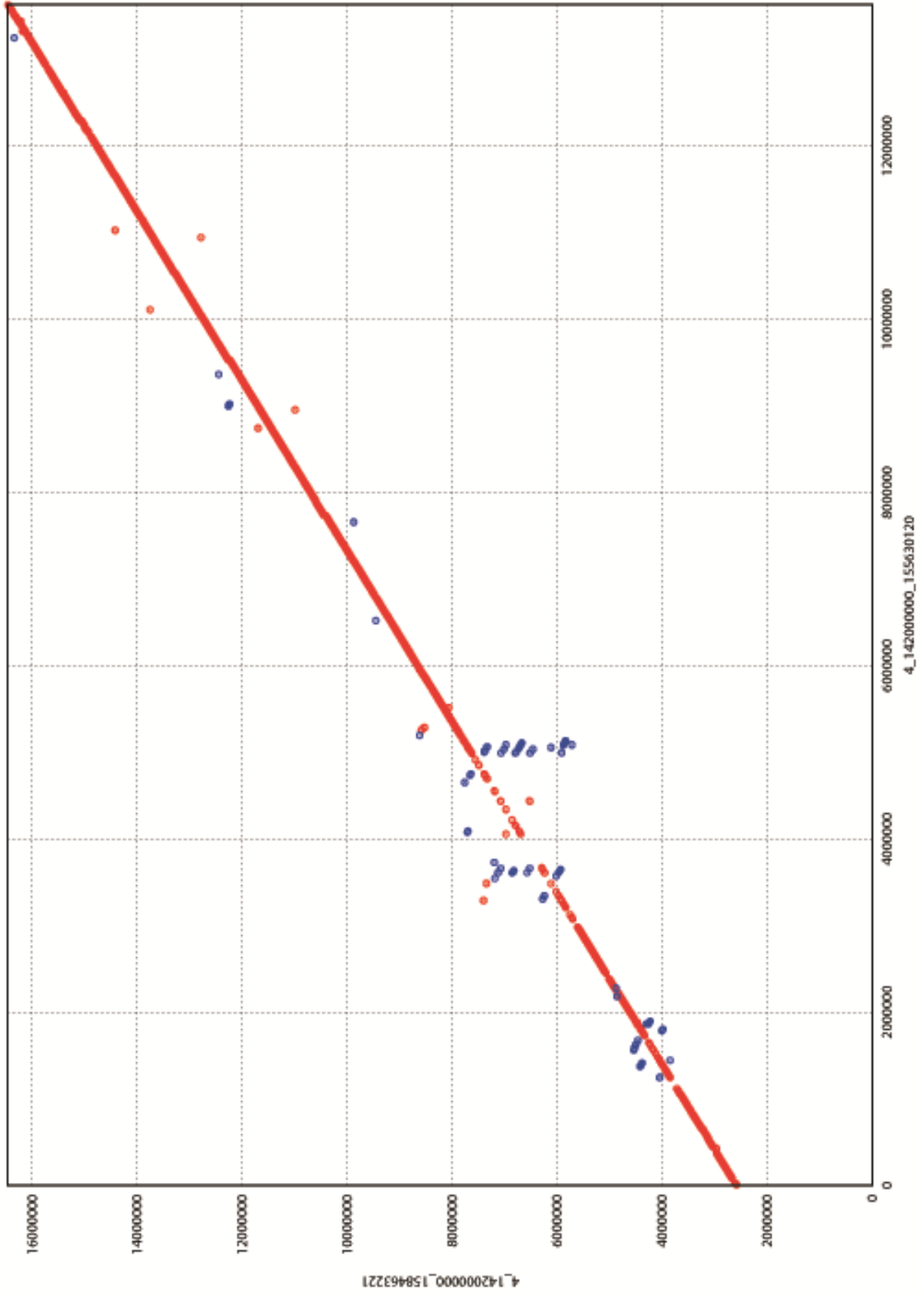
**A)** Overall distribution histogram. **B)** Zoom into contig sizes ranging from 1bp-40Kb.





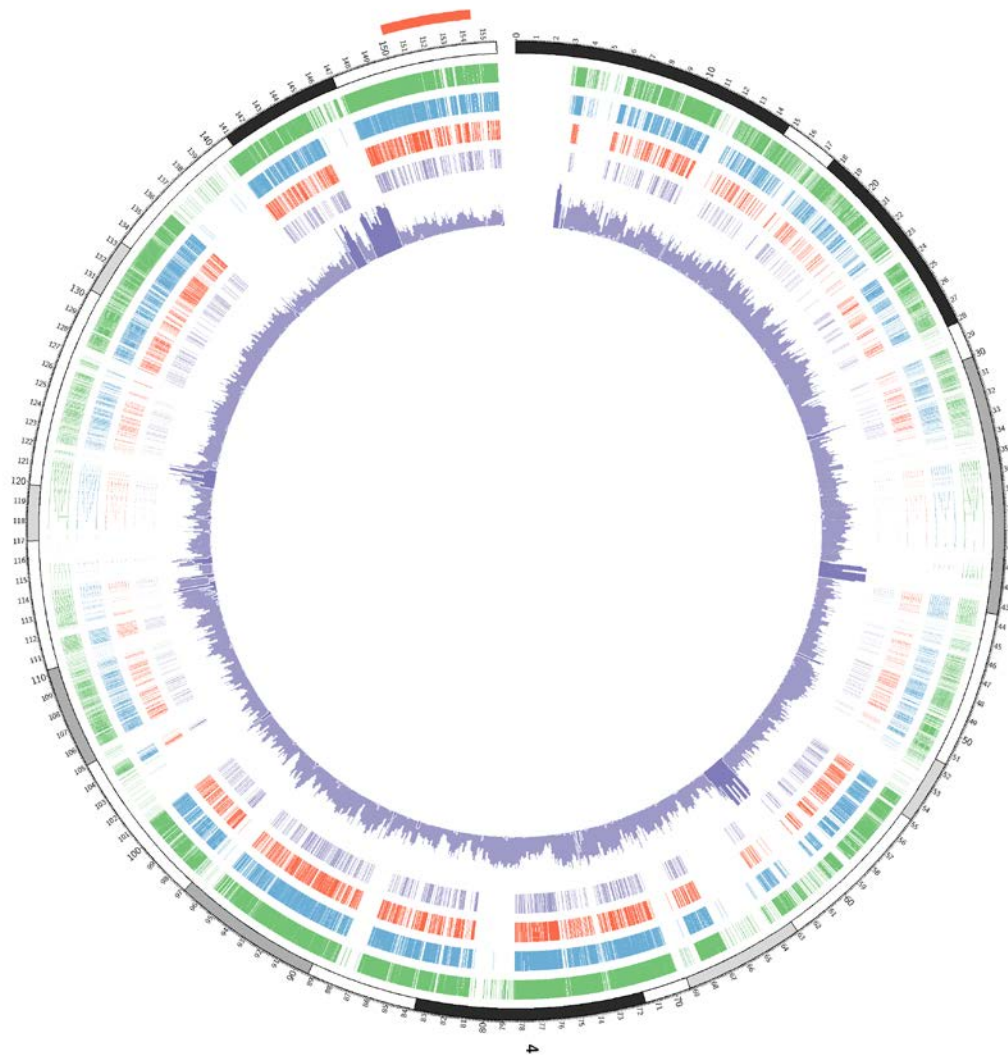
**Figure 2.8 Mummerplot of nucmer aligned 129S5/SvEv<sup>Brd</sup> assembly chromosome 4**

129S5/SvEv<sup>Brd</sup> assembly chromosome 4 is represented on the x axis, while reference C57B16/J chromosome 4 is represented in the y axis. Minimum cluster length of 1000bp. Red is forward aligned sequences while blue indicates reverse orientation. Horizontal lines are extensive regions of high repeat content, such as SDs.



**Figure 2.9 mummerplot of nucmer aligned 129S5/SvEv<sup>Brd</sup> 4E2 to repeat masked reference C57Bl6/J 4E2**

129S5/SvEv<sup>Brd</sup> 4E2 is shown on the *x* axis, while repeat masked reference C57Bl6/J 4E2 is represented in the *y* axis. Minimal cluster length of 500bp. Red is forward aligned sequences while blue indicates reverse orientation.



**Figure 2.10 SNP locations inside the 4E2 region**

Outer red box corresponds to the deletion CNV. Circles going from outside to inside: Green: 129S5/SvEv<sup>Brd</sup> SNPs as detected by Sanger Mouse Sequencing Project. Blue: SNPs as reported by Perlegen sequence for mouse strain 129S1. Red: SNPs from combined Sanger-Perlegen projects that fall inside *HindIII-DpnII* 4C sites (see Chapter 4). Purple: SNPs contained within 4C sites not overlapping RepeatMasked elements. Inner circle histograms: RepeatMasker elements (light color) and SD regions (darker color).

## 2.5 Spectral karyotyping analysis of $df/+^{Bl6}$ and $+^{129}/+^{Bl6}$ MEFs

In order to assess the chromosomal integrity of the cell lines used for this project, in particular those that would be used for 3D DNA FISH (Chapter 3) and PE-4Cseq (Chapter 4) experiments, spectral karyotype (SKY) analysis was performed. The goal was to detect constitutive translocations in chromosome 4 that could potentially affect the interpretation of *cis* chromatin interactions from PE-4Cseq data. Additionally, I wanted to assess the karyotypic variability that both genotypes could display in culture. For this reasons, representative samples were chosen from  $df/+^{Bl6}$  (129S5E71) and  $+^{129}/+^{Bl6}$  (129S5E117) MEF lines, and analyzed at passages >P5 (culture passage number is higher in order to obtain enough cells for metaphase spreads). SKY analysis was performed in collaboration with Hessed Padilla-Nash, from Thomas Ried's laboratory at NIH. No  $dp/+^{Bl6}$  MEFs were analyzed for technical and biological reasons discussed in Chapter 3.

25 metaphase spreads of  $+^{129}/+^{Bl6}$  (129S5E117) MEFs analyzed by SKY revealed that this cell line exhibits aneuploidy and low level of chromosome instability (CIN). Karyotypes are all: 40,XX. 8 cells were classified as normal (diploid, 2n). 8 are near-diploid (+2n), 2 are near-triploid (+3n), and 7 are near-tetraploid (+4n) [see Fig. 2.11 for an example of an abnormal  $+^{129}/+^{Bl6}$  MEF cell with a translocation in the terminal part of chromosome 4]. Overall, 64% of the analyzed MEFs are nearly normal, and no constitutive translocations or other major karyotypic alterations were found for chromosome 4 of these cells, even at passages >P5.

22 metaphase spreads of  $df/+^{Bl6}$  (129S5E71) MEFs analyzed by SKY revealed that this cell line exhibits aneuploidy, and low level of CIN. Karyotypes are all: 40,XX. 8 cells were classified normal (2n), 7 are near-diploid, 3 are near-triploid, 2 are tetraploid, and 1 is

octaploid (8n) [see Fig. 2.12 for an example of an abnormal  $df/+^{Bl6}$  MEF cell with a reciprocal translocation between chromosomes 4 and 8]. The data revealed that 68% of the analyzed MEFs are nearly normal, and no constitutive translocations or other karyotypic alterations were found on chromosome 4.

Quantification of SKY results for both  $df/+^{Bl6}$  and  $+^{129}/+^{Bl6}$  genotypes showed an average of one translocation per 25 cells analyzed between the terminal part of chromosome 4 and the rest of the chromosomes [Supp. Table 2.2A,B]. Chromosome 7 was selected for karyotype translocation frequencies comparison. It harbors the housekeeping gene *Rps13*, which was included for 3D DNA FISH experiments (Chapter 3) and molecular characterizations of chromatin changes (PE-4Cseq, Chapter 4). SKY results for chromosome 7 showed no translocations with other chromosomes in  $+^{129}/+^{Bl6}$  cells, and a single deletion was detected in the 22 analyzed  $df/+^{Bl6}$  MEFs.

Aneuploidies and polyploidies (n=3,4,8) of chromosome 4 are present in 25-40% of  $df/+^{Bl6}$  and  $+^{129}/+^{Bl6}$  MEFs, but their structural integrity is preserved. In all cases analyzed, the 4E2 region was not directly affected by translocations or major sequence deletions/amplifications. These results suggest that, although translocations exist for the terminal part of chromosome 4, these occur at low frequencies (4%) in the studied populations. The individual contributions of such changes to the bulk of data in PE-4Cseq experiments would be diluted when performing *cis* chromatin contact analyses, given that, if not falling inside 4E2, potential translocation contacts would be accounted as inter-chromosomal, and therefore not targeted for the current study (see Chapter 6 for discussion). Additionally, the mapped read counts are bias-corrected and compared between two biological replicates, therefore reducing the contribution of spurious interactions to

chromatin conformation analyses (Chapter 4).

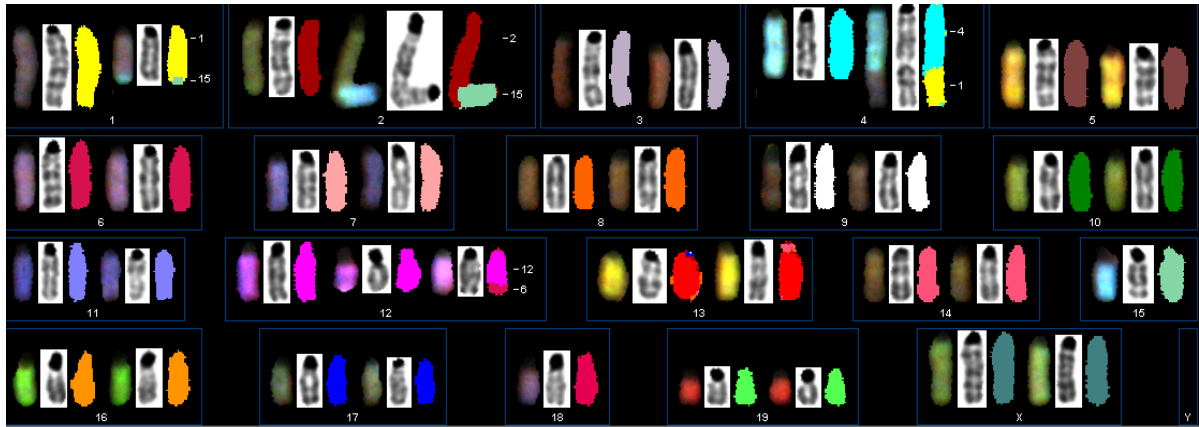
Moreover, the MEFs analyzed by SKY correspond to passages >P5 for both  $df/+^{Bl6}$  and  $+^{129}/+^{Bl6}$  genotypes. These passages were used to derive enough number of cells for metaphase preparations. However, given that all the MEFs used for the 3D DNA FISH (Chapter 3), PE-4C-seq (Chapter 4), and RNA-Seq (Chapter 5), as well as all additional validation experiments were performed in an earlier passage (P4), we can conclude that the number of expected chromosomal alterations would be reduced compared to results obtained for the currently analyzed SKY data.

## 2.6 Discussion of CNV mouse models of 4E2

We proposed the use of  $+^{129}/+^{Bl6}$ ,  $df/+^{Bl6}$ , and  $dp/+^{Bl6}$  MEFs for the study of chromatin organization after the occurrence of CNV changes. Several factors make these models suitable for the design of microscopic and molecular experiments for the study of chromatin architecture.

1. 4E2 is syntenic to human 1p36, where deletions are common and associated with disease phenotypes (Heilstedt *et al.*, 2003; reviewed in Bagchi and Mills, 2008)
2.  $df$  and  $dp$  chromosomes were engineered using 129S5/SvEv<sup>Brd</sup>-derived ES cells, therefore providing enough sequence differences that can be potentially targeted for allele-specific analysis of chromatin conformation (Chapter 4).
3. The specific locations of the deletion and duplication CNVs in 4E2 are known.
4. Phenotypic characterizations had been previously published for  $df/+^{Bl6}$ , and  $dp/+^{Bl6}$  MEFs (Bagchi *et al.*, 2007).





**Figure 2.11 Abnormal karyotype for  $+^{129}/+^{Bl6}$  (129S5E117) MEF cell 24 as revealed by SKY.**

Karyotype: 40,XX,T(1;15),Dic(2;15),T(4;1),+12,Del(12),T(12;6),-18. This cell is near-diploid ( $2n=40$ ) with several numerical and structural aberrations. Chromosomes 1, 4, and 12 are unbalanced translocations, chromosome 2 has formed a dicentric chromosome with chromosome 15, and chromosome 12 has additional chromosome 12 material, and chromosome 18 has only one copy. Cells 10, 22, and 24 also had loss of chromosome 18. Display (RGB) for each chromosome is on left, aligned next to the inverted-DAPI banded chromosome, and classification pseudocolors are on the far right.



**Figure 2.12** Abnormal karyotype for *df/+<sup>Bl6</sup>* (129S5E71) MEF cell 11 as revealed by SKY

Karyotype: 57,XX,-X,+1,-2,-2,-3,-3,T(4;8),+5,-7,-8,T(8;4),-10,-11,+12,+14,+15, +19. This cell is near-triploid ( $3n=60$ ) with several numerical aberrations (gains of 1, 5, 12, 14, 15, and 19; losses of chromosomes X, 2, 3, 7, 8, 10, and 11, and one reciprocal translocation T(4,8) & T(8;4). Classification pseudocolors for each chromosome are on left, aligned next to the inverted-DAPI banded chromosome.

5. Observations derived from  $df/+^{Bl6}$  MEF analyses could be assessed in cases of Monosomy 1p36 in human patients.

Given the importance of knowing the underlying chromosome 4 structure for chromatin organization studies, a SKY analysis was performed on two representative  $+^{129}/+^{Bl6}$  and  $df/+^{Bl6}$  MEF cell lines that would be used for future experiments. Although 1/3 of the analyzed cells show deviations on the overall expected ploidy level for both genotypes, chromosome 4 is diploid in ~60% of the cells in  $+^{129}/+^{Bl6}$ , and ~80% in  $df/+^{Bl6}$ . Because these results were derived from >P5 MEFs, we expect to have a lower ratio of karyotypically abnormal cells in the analyzed P4 populations for 3D DNA FISH (Chapter 3), PE-4C-seq (Chapter 4), and RNA-Seq (Chapter 5) experiments. Even though it is not certain how deviant ratios from the expected 2n chromosome number in nuclei would affect the organization of chromosomal territories and affect intra-chromosomal interactions, we have verified at the karyotypic level that the chromosomes evaluated for this study do not possess clonal aberrations that could bias the *cis* chromatin interactions interpretation of our analyses. The use of biological replicates in all experiments performed will shed more light into the interpretation of chromatin interaction data, and the impact of karyotype abnormalities in the detection of architectural changes.

Chapter 3 will introduce a general assessment of changes in chromatin organization as revealed by 3D DNA FISH, while Chapter 4 will molecularly describe the magnitude of the impact that CNVs can have on chromosome 4 architecture.

### Chapter 3: Microscopic characterization of higher-order chromatin organization in 4E2 CNVs

For decades, DNA *in situ* hybridization has been used by Cytogeneticists for the detection of chromosomal abnormalities that cause disease. First introduced by Joseph Gall and Marie Lou Pardue in the 1960's (Gall and Pardue, 1969), *in situ* hybridization makes use of DNA complementarity to locate positions of labeled DNA probes on chromosomes. This methodology allows the assessment of diverse chromosomal aspects, such as copy number (aneuploidies, interstitial duplications, interstitial deletions), structure (translocations, chromatin folding), chromatin fiber compaction, chromosome/gene positioning, chromosome/gene overlap with diverse nuclear features (nuclear lamina, nuclear bodies), and genomic scaffold assembly, among others. Ever since its introduction, the DNA *in situ* hybridization protocol has undergone several variations and optimizations, including the substitution of radioactively labeled probes by fluorescent labeled ones (FISH) (reviewed in Levsky and Singer, 2003).

Given the relative ease by which FISH can be performed, and the amount of single-cell information one can derive from it, I performed a comprehensive characterization in  $+^{129}/+^{Bl6}$ ,  $df/+^{Bl6}$  and  $dp/+^{Bl6}$  MEFs through the use of 3D DNA FISH. I evaluated chromatin states of the CNV regions and their adjacent sequences up to 60Mb away for the *df*, *dp* and WT chromosomes to understand the long-range effects that copy-number variation can exert on chromosome structure. Such microscopic analyses provided us with rough information on the type of chromatin organization (compact, open) that exists within each CNV and its neighboring regions, and whether there were specific nuclear features associated with these regions (heterochromatin foci overlap, distinctive nuclear localization, etc).

### 3.1 3D DNA FISH of 4E2 and neighboring regions in $+^{129}/+^{Bl6}$ , $df/+^{Bl6}$ and $dp/+^{Bl6}$ MEFs

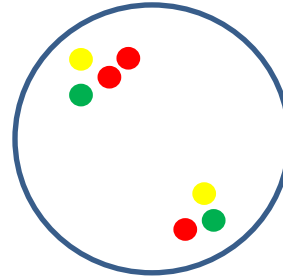
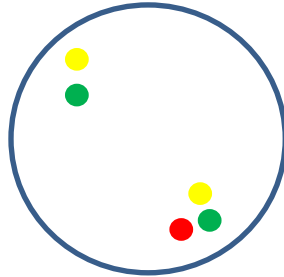
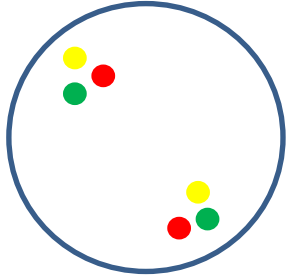
The 4E2 and its neighboring regions were analyzed in  $df/+^{Bl6}$  (129S5E36),  $dp/+^{Bl6}$  (129S5E60) and  $+^{129}/+^{Bl6}$  (129S5E90) P4 MEFs using 3D DNA FISH in fixed cells. In order to derive the most accurate information from these experiments, we used a FISH protocol that had been previously optimized for the preservation of nuclear structure (Solovei and Cremer, 2010). Each FISH experiment consisted of the use of one red (Alexa 594) and one green (Alexa 488) labeled bacterial artificial chromosome (BAC) probe separated by ~500Kb from each other along 4E2. A third Cy5-labeled (Alexa 647) BAC probe was included inside the CNV region, so that each red-green pair measurement was identified as belonging to either the WT ( $+^{129}$  or  $+^{Bl6}$ ),  $df$ , or  $dp$  chromosomes [Fig. 3.1A]. A total of 8 different red-green probe pairs were used, separated by ~5Mbp between each other. An additional control probe set bordering the CNV start and end was included, and two control BAC sets on chromosomes 6 and 7 for assessing chromatin characteristics of the *Gapdh* and *Rps13* genes, respectively [Fig. 3.1B. Table 3.1].




100+ nuclei images were obtained per BAC pair for the  $+^{129}/+^{Bl6}$  and  $df/+^{Bl6}$  genotypes, while the average for  $dp/+^{Bl6}$  MEFs was about 70 nuclei (see discussion at the end of this chapter for further comments into the  $dp/+^{Bl6}$  MEFs). Images were acquired using an Applied Precision DeltaVision Core wide-field fluorescence microscope system (GE Healthcare) with a PlanApo 60× 1.40 numerical aperture objective lens (Olympus America). Image stacks were taken at 0.3µm intervals throughout the entire cell and deconvoluted using Applied Precision softWoRx software version 4.2.1 with default parameters.

A)  $+^{129}/+^{Bl6}$

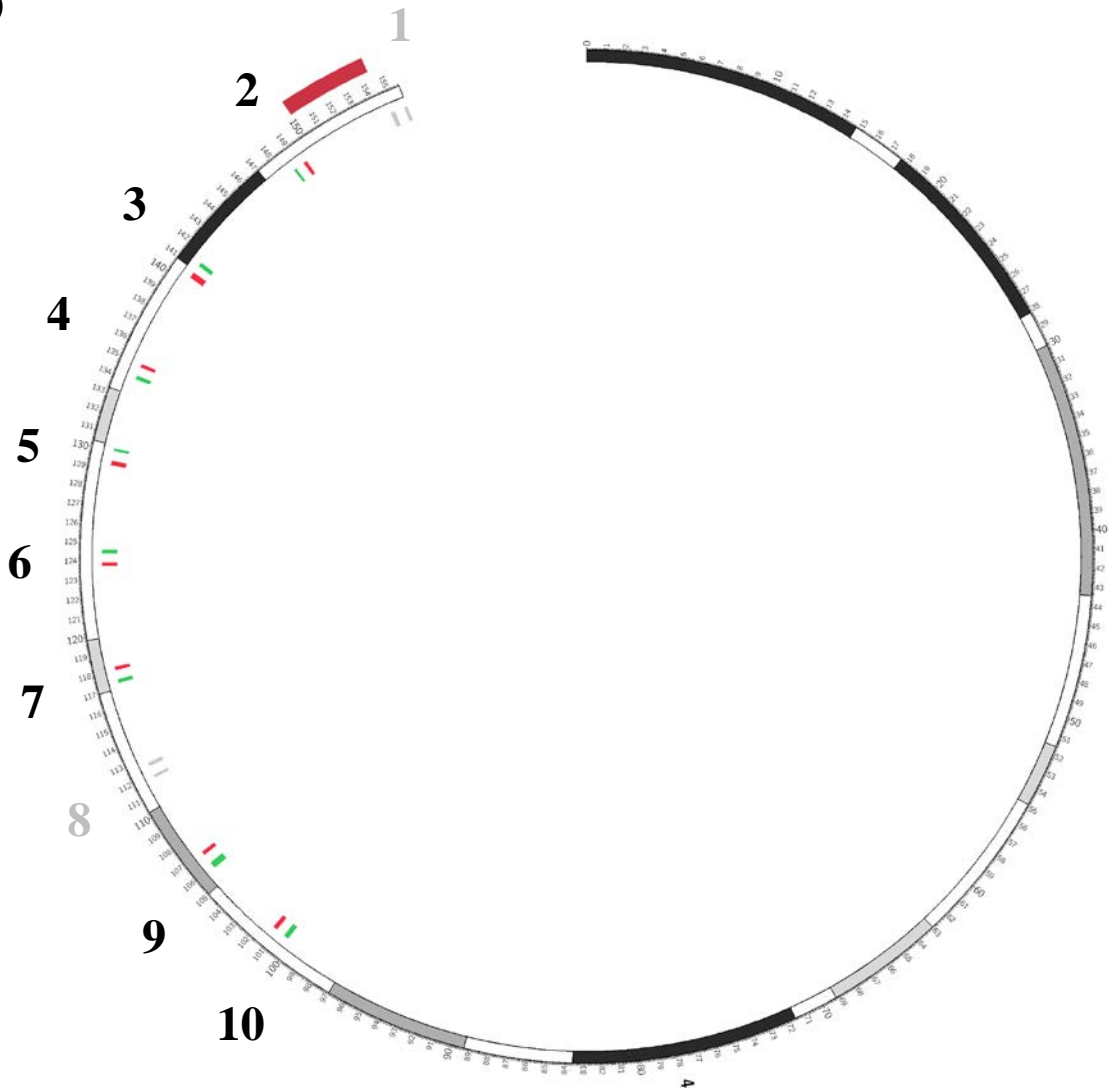
$df/+^{Bl6}$

$dp/+^{Bl6}$



Channel 1   
Channel 2   
Channel 3 

B)



### Figure 3.1 3D DNA FISH experiments and analysis

A) Chromosomal classification based on the expected number of signals per nuclei in channel 3.  $+^{129}/+^{B16}$  cells have 2 channel 3 signals, while  $df/+^{B16}$  and  $dp/+^{B16}$  have 1 and 3, respectively. B) Circular depiction of mouse chr4. CNV region is depicted as the outer red box towards the telomere. BAC pairs used in DNA FISH experiments are labeled as red and green internal boxes marked with numbers. BACs labeled in gray overlap repeats and therefore could not be used. Zoom: the CNV bordering red and green probes were used to assess compaction changes in both *df* and *dp* chromosomes.

BAC set	BAC 1	Chr	Start	End	Size	BAC 2	Chr	Start	End	Size	BACs sep
2	RP24-155J20	4	149,191,899	149,363,905	172,007	RP23-156B13	4	148,550,430	148,681,807	131,378	510,093
3	CH29-36F05	4	141,166,002	141,388,055	222,054	RP23-236F18	4	140,395,870	140,709,751	313,882	456,252
4	RP23-183A2	4	134,810,292	134,993,274	182,983	RP24-347G23	4	134,108,484	134,307,306	198,823	502,987
5	CH29-523J16	4	130,030,674	130,165,237	134,564	RP23-296I2	4	129,222,727	129,484,507	261,781	546,168
6	CH29-15G03	4	124,398,963	124,602,689	203,727	RP23-223D4	4	123,718,139	123,899,231	181,093	499,733
7	RP23-448M11	4	118,006,501	118,193,604	187,104	RP23-284F14	4	117,291,581	117,491,409	199,829	515,093
9	RP23-230A12	4	106,777,315	106,986,786	209,472	RP24-84P2	4	105,885,213	106,233,140	347,928	544,176
10	RP23-147G4	4	101,122,289	101,342,534	220,246	RP23-148M5	4	100,325,609	100,595,797	270,189	526,493
<i>Gapdh</i>	CH29-580O08	6	124,992,733	125,175,262	182,530	CH29-578H14	6	124,315,842	124,498,088	182,247	494,646
<i>Rps13</i>	CH29-545O18	7	116,129,297	116,336,844	207,548	CH29-72O18	7	115,340,405	115,620,681	280,277	508,617
CNV borders	RP24-123J14	4	154,664,416	154,891,026	226,611	RP24-155J20	4	149,191,899	149,363,905	172,007	5,300,512
Probe inside CNV	RP23-114D1	4	153,343,762	153,525,186	181,425						

**Table 3.1 BACs used as probes for the 3D DNA FISH experiments and their corresponding chromosomal location**

BAC1 are always Alexa 594-labeled and BAC2 are Alexa 488-labeled. BAC RP23-114D1 labeled with Alexa 647. Sizes and BAC distances are given in bp.



## 3.2 Development of a dedicated ImageJ plugin for automated analysis of 3D DNA FISH

Given the bulk of FISH images obtained per genotype and BAC set (>5000 total), we sought to analyze the information in a fast, unbiased, and reproducible manner. In collaboration with Nathalie Harder from the group of Karl Rohr at the University of Heidelberg, an ImageJ plugin was developed to analyze the 3D DNA FISH images derived from these experiments.

The plugin, named `Correct_and_Measure_3D.class` (see Computational Methods in Chapter 8), is based on the segmentation of DAPI and FISH signals for the calculation of various biological parameters, which include: measurements of nuclear volume, number of FISH signals per cell per excitation channel, 3D distance separating FISH signals within the same channel and between channels, percentage of FISH signals which overlap heterochromatin foci, FISH signals distances to the nuclear periphery and the nuclear centroid, and automatic classification of channels in the order red (Alexa 594), green (Alexa 488), Cy5 (Alexa 647), and DAPI for the correct assignment of measurements based on CNV genotypes.

**Plugin analysis steps** [see Fig. 3.2 for summary and Fig. 3.3 for an example of data analysis]

### 1. Segmentation of nuclei

Gaussian filtering for noise reduction ( $\sigma=2$ ).

Automatic thresholding based on brightest slice of the stack to avoid bias by noise in

border slices (multi-level Otsu with 3 levels to deal with bright regions).

Hole filling, splitting of cell clusters in 3D and 3D labeling using watershed transform based on Euclidean distance transform.

## 2. Segmentation of heterochromatic regions

- Gaussian filtering ( $\sigma=2$ ) followed by tophat transform (structuring element radius=2, reduces background and emphasizes bright objects of defined size).
- Automatic thresholding based on brightest slice using Renyi entropy (histogram-based).
- Median filtering (radius=1).

## 3. Segmentation of FISH channels

- Gaussian filtering ( $\sigma=1$ ) followed by tophat transform (structuring element radius of 12 to 24, depending on image resolution).
- Automatic thresholding based on brightest slice using Renyi entropy (histogram-based).
- Median filtering (radius=1).
- Splitting of signal clusters in 3D and 3D labeling using watershed transform based on Euclidean distance transform.

#### 4. Optimal signal mapping and quantification of distances

- Goal: determine distances between FISH signals in different channels.
- Problem: multiple signals in both channels, partly true signals, partly non-specific signals, not necessarily the same number.
  - determine optimal mapping of the FISH signals in both channels between which the distances should be determined automatically.
- First approach: if signal number is larger than two, perform clustering into two classes and determine distances between cluster centers.
  - inaccurate if many non-specific signals.
- Current approach: determine best matching by minimizing the mean of squared distances between the signals of the two channels (clustering only if distance is below  $1\mu\text{m}$ ).
  - always finds best overall solution.
  - provides the two min distances and the mean distance between signals of best fit.
- Quantification of distances between FISH signals, of distances to the cell nucleus center and border, and quantification of overlap of FISH signals and heterochromatin regions.

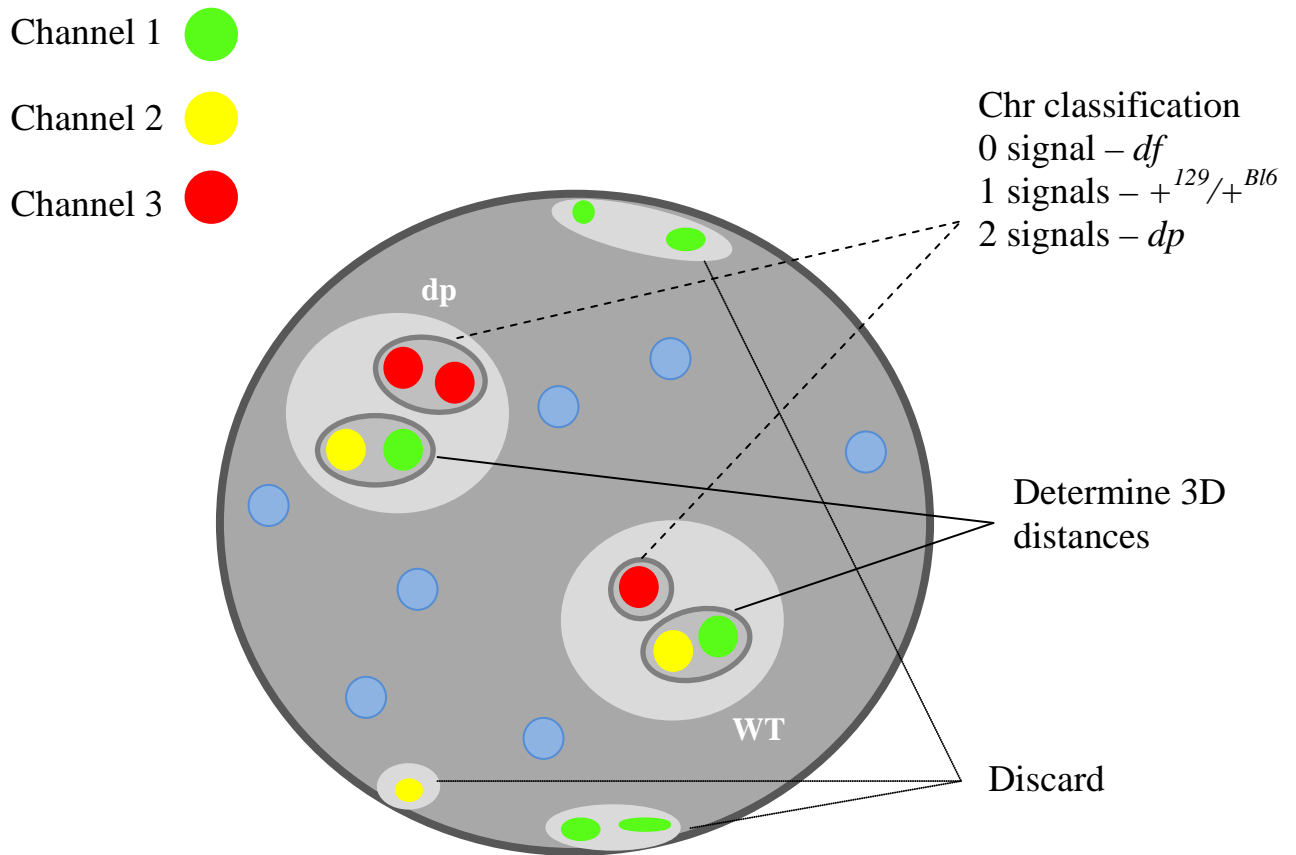
## 5. Classification of chromosome sets according to channel 3

- Measurements are classified to belonging to each chromosome set ( $+^{129}/+^{Bl6}$ ,  $df/+^{Bl6}$ ,  $dp/+^{Bl6}$ ) based on the number of Cy-5 (channel 3) signals detected.

Results reported by the plugin include 2 files (ImageName\_Measurements.txt and ImageName\_ParticleStatistics.txt, where ImageName is the name of the FISH image file analyzed) that describe all the measured parameters (see Computational Methods in Chapter 8 for output description). A summary file of FISH data measurements (Summary.txt) together with a list of all ignored images that did not pass quality filters (CORRUPTED\_FILES.txt) were produced per genotype per BAC set used and analyzed using custom R scripts (see Computational Methods in Chapter 8 for the analysis pipeline).

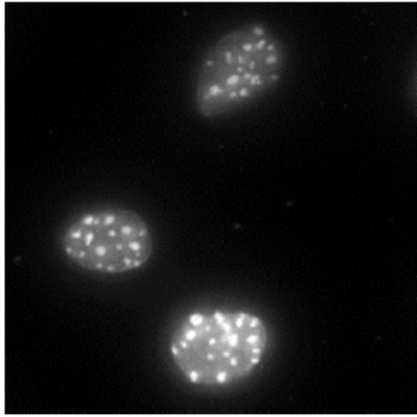
### **3.3 ImageJ plugin results and validation**

In order to assess the performance of our custom-made plugin for the analysis of 3D DNA FISH images, we performed a comparison of manually measured 3D physical distances between red-green BAC probes with their corresponding plugin calculated distances. Physical distances between red-green probes are a measurement of chromatin compaction, a useful parameter for the evaluation of the plugin's performance in image segmentation, signal identification, and chromosome assignment.

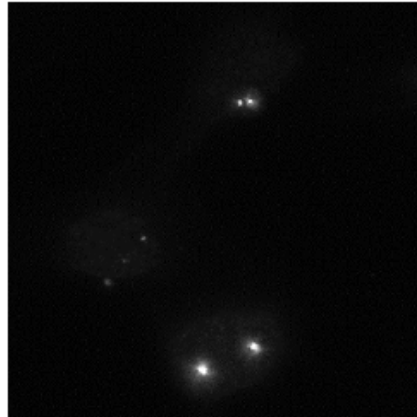


**Figure 3.2 Overview of 3D DNA FISH analysis workflow**

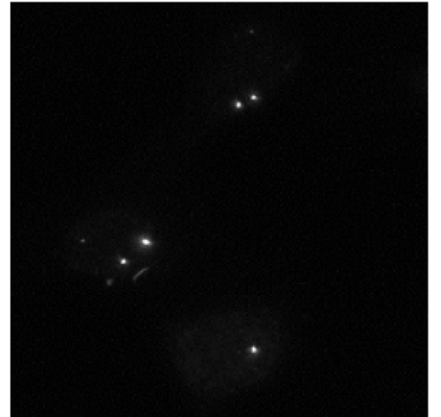
Analysis steps include: 1) 3D segmentation of cell nuclei (outer black line), heterochromatin (inner blue circles), and FISH signals (red, green, and yellow circles). 2) Quantification of basic features (i.e. nuclear volume, BAC signals distances to nuclear periphery and centroid, overlap of BAC signals and heterochromatin, etc). 3) Clustering of signals of different channels. 4) Classification of chromosome sets according to Channel 3 (*dp* and WT in this example).



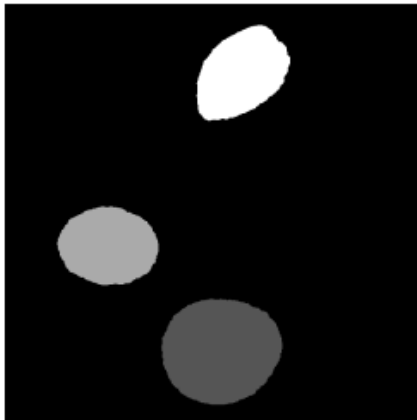
Original image DAPI



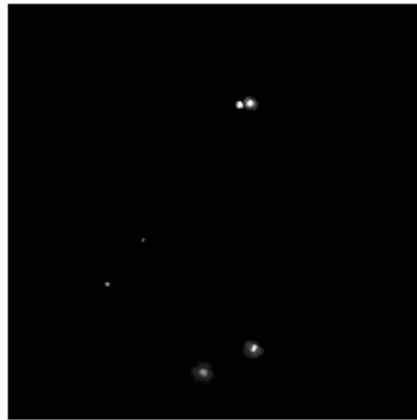
Original image FISH 1



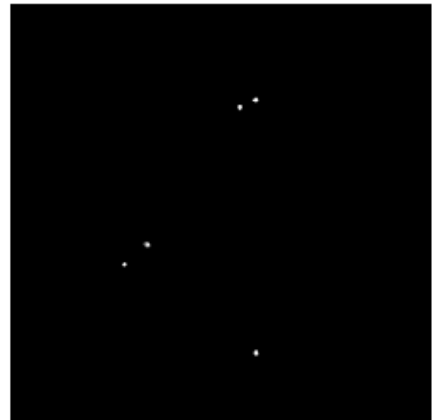
Original image FISH 2



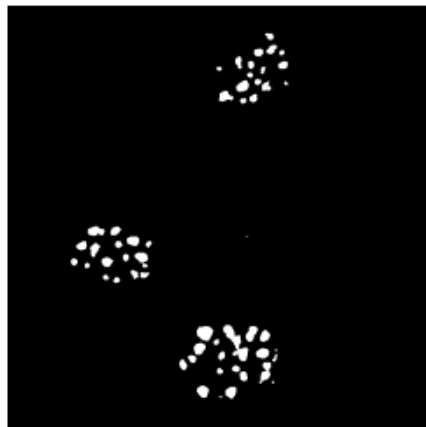
Segmentation result nuclei



Segmentation result FISH 1



Segmentation result FISH 2



Segmentation result  
heterochromatic regions

**Figure 3.3 An example of 3D DNA FISH segmentation results**

Representative example of segmentation for 3  $+^{129}/+^{Bl6}$  MEFs. Notice the agreement between the imaged DAPI and BAC fluorescence signals for channels 1 and 2 and the segmentation results. The bottom cell analyzed possesses only 1 signal of FISH probes on Channel 2. Consequently, only for the top and middle nuclei 255 the inter-channel distance of FISH signals can be determined.

Compared images include  $df/+^{Bl6}$  and  $dp/+^{Bl6}$  data derived from BAC pair 2. As can be observed from Fig. 3.4A and B, both manual and plugin-derived measurements are highly concordant ( $\sim 0.08\mu\text{m}$  absolute average difference, medians range from  $0.05\text{-}0.06\mu\text{m}$ , and mode is  $0.02\mu\text{m}$  for both cases). Chromosomal classification is 100% correct in all cells analyzed.

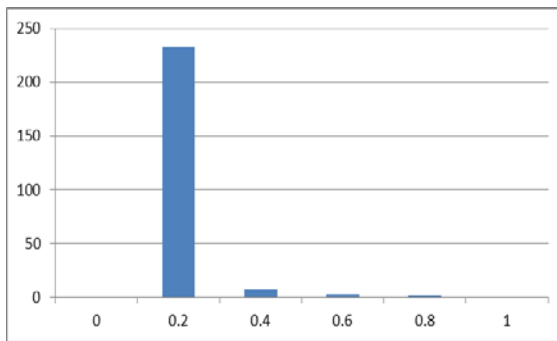
### **3.4 Assessing the reproducibility of results derived from 3D DNA FISH experiments**

In order to estimate the reproducibility of results between biological replicates, we repeated FISH experiments for BAC sets 2 and 10 using different  $+^{129}/+^{Bl6}$  (129S5E88),  $df/+^{Bl6}$  (129S5E71) and  $dp/+^{Bl6}$  (129S5E61) MEF lines. BAC set 2 is  $\sim 1\text{Mb}$  away from the CNV, while BAC set 10 is  $\sim 50\text{Mb}$  away from it, therefore probing two different sequence environments as experimental quality controls.

As can be observed in Table 3.2A and B, the magnitude of absolute differences between medians of measured allelic 3D distances is  $\sim 0.001\text{-}0.06\mu\text{m}$  for both assayed BAC set probes in all three genotypes, arguing for reproducibility of results regardless of the MEF line of origin. Although these observations could be due to stable chromatin conformations for the BAC sets 2 and 10 regions, we decided to not make biological replicates for all BAC sets given the high time consumption of the experiments, and the high degree of reproducibility of the two BAC sets probing different chromosome 4 environments.

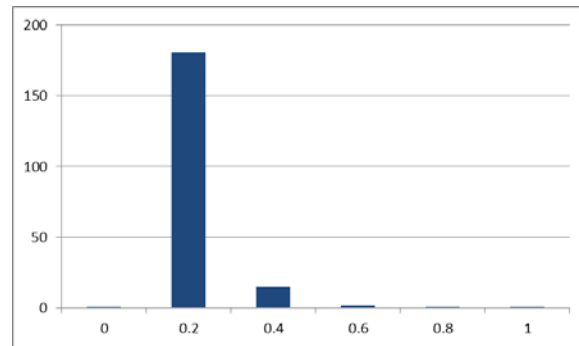


**A)**  $df/+^{Bl6}$  MEFs



min	0
max	0.81
average	0.08
mode	0.02
median	0.047

**B)**  $dp/+^{Bl6}$  MEFs



min	0
max	0.81
average	0.09
mode	0.00
median	0.06

**Figure 3.4 Comparison of plugin vs manual distances of chromatin compaction**

For BAC set #2 for **A)**  $df/+^{Bl6}$  (n=246) and **B)**  $dp/+^{Bl6}$  (n=200) alleles. Histograms summarize absolute value differences between plugin and manual SoftWorx results, which for both cell types fall mostly in the 0-0.2 $\mu$ m bin. Tables provide difference descriptive statistics per sample of the manual-automated measurement differences.

FISH experiments were also performed for BAC set 10 using MEFs in passage 2 and 3 to assess whether chromatin compaction and other measurements varied depending on culture passaging. As can be seen from Table 3.3, the range of differences between passages median compaction measurements is 0.02-0.11 $\mu\text{m}$  for the three genotypes. These observations suggest that at least for BAC set 10, compaction does not change for all genotypes as MEFs increase their passage number.

### **3.5 ImageJ plugin results of 3D DNA FISH of 4E2 and neighboring regions in $+^{129}/+^{Bl6}$ , $df/+^{Bl6}$ and $dp/+^{Bl6}$ MEFs**

All the plugin-analyzed MEFs in the 3D DNA FISH images were filtered based on the number of FISH signals present (excluded cells with  $<2$  alleles per red/green excitation channel) and cell size (excluded cells  $<400\mu\text{m}^3$  and  $>200,000\mu\text{m}^3$ ). In total, 5,236 images were analyzed, and 5,402 cells (10,804 alleles) were included in our analyses after quality filtering. Of these 3,632 cells correspond to the eight chromosome 4 regions analyzed in all 3 genotypes, while the rest belong to the validation experiments described in the previous sections [Table 3.4. Supp. Table 3.1].

Analysis of results reported by the plugin allowed us to identify the BAC probes bordering the 4.3Mb deletion as the ones displaying the largest (average of  $\sim 0.4\mu\text{m}$ ) chromatin compaction differences compared to the  $+^{129}/+^{Bl6}$  and  $dp/+^{Bl6}$  values [Fig. 3.5]. This was an expected result given the reduction in the size of the chromatin fiber after the deletion of intervening sequence, allowing the probes to become neighbors along the

chromosome sequence. Intriguingly, the 4.3Mb duplication does not seem to affect compaction distances between these probes. This argues for a possible looping of the duplicated region out of the preferred chromatin conformation state, but this idea was not further tested given the slow growth and phenotypic abnormalities (senescence) of  $dp/+^{Bl6}$  MEFs which prevented us from obtaining enough materials for the experiments (more on  $dp/+^{Bl6}$  MEFs CNV conformation is discussed in the end of this chapter).

The compaction differences between the *Gapdh* and *Rps13* BAC sets was always  $<0.2\mu\text{m}$  in all 3 genotypes [Supp. Table 3.2].  $0.2\mu\text{m}$  is the microscope's resolution limit, therefore compaction measurements between the genotypes in these control regions is not significantly different [Fig. 3.6. Supp. Table 3.2]. The data also revealed that 2 out of the 8 regions surrounding the CNV had distinct compaction distributions in  $dp/+^{Bl6}$  MEFs (regions 4, and 7. Kolmogorov-Smirnov two-sided test  $p<0.05$ ). Both regions display a difference  $>0.2\mu\text{m}$  in the third quartile distribution of values between  $dp/+^{Bl6}$  MEFs and  $+^{129}/+^{Bl6}$  and  $df/+^{Bl6}$  [Fig. 3.7. Supp. Table 3.2].

It was immediately noted from the FISH analysis that  $dp/+^{Bl6}$  MEF nuclei tend to have larger volumes ( $\sim 260\mu\text{m}^3$  difference) compared to  $+^{129}/+^{Bl6}$ , while  $df/+^{Bl6}$  MEFs tend to have smaller nuclei ( $\sim 100\mu\text{m}^3$  difference) compared to  $+^{129}/+^{Bl6}$  [Kolmogorov-Smirnov two-sided test,  $p<0.05$ . Fig. 3.8A,B]. However, there is a weak-to-low ( $df/+^{Bl6}$ ) and non-meaningful ( $dp/+^{Bl6}$ ,  $+^{129}/+^{Bl6}$ ) correlations between the chromatin compaction of the regions and the nuclei volumes of all genotypes as revealed by the Spearman rank correlation coefficient ( $+^{129}/+^{Bl6}$  rho = 0.13;  $df/+^{Bl6}$  rho = 0.22;  $dp/+^{Bl6}$  rho = 0.058). Therefore, the changes in compaction observed for BAC sets 4 and 7 cannot be explained by an increase in nuclear volume.

A)

Sample	Allele	Min	1st Qu	Median	Averag	3rd Qu	Max
1st Rep	WT1	0.15	0.38	0.50	0.56	0.70	1.93
	WT2	0.15	0.37	0.50	0.56	0.73	1.47
	Average	0.15	0.37	0.50	0.56	0.72	1.70
2nd Rep	WT1	0.08	0.42	0.57	0.63	0.74	1.93
	WT2	0.07	0.39	0.55	0.64	0.78	2.48
	Average	0.07	0.41	0.56	0.63	0.76	2.21
	Differen	0.07	0.03	0.06	0.07	0.04	0.51

1st Rep	dfBl6	0.06	0.39	0.54	0.59	0.71	1.54
	df129	0.07	0.41	0.55	0.58	0.73	1.34
	Average	0.06	0.40	0.54	0.59	0.72	1.44
2nd Rep	dfBl6	0.09	0.39	0.56	0.61	0.76	2.24
	df129	0.01	0.37	0.53	0.59	0.73	2.06
	Average	0.05	0.38	0.55	0.60	0.74	2.15
	Differen	0.02	0.02	0.00	0.02	0.03	0.71

1st Rep	dpBl6	0.12	0.50	0.62	0.71	0.90	2.02
	dp129	0.07	0.47	0.63	0.64	0.78	1.51
	Average	0.10	0.48	0.63	0.68	0.84	1.76
2nd Rep	dpBl6	0.07	0.42	0.57	0.66	0.77	3.17
	dp129	0.11	0.44	0.58	0.63	0.80	1.54
	Average	0.09	0.43	0.57	0.64	0.78	2.36
	Differen	0.01	0.05	0.05	0.03	0.06	0.60

B)

Sample	Allele	Min	1st Qu	Median	Averag	3rd Qu	Max
1st Rep	WT1	0.14	0.32	0.50	0.58	0.73	3.02
	WT2	0.10	0.34	0.48	0.52	0.63	1.58
	Average	0.12	0.33	0.49	0.55	0.68	2.30
2nd Rep	WT1	0.10	0.34	0.48	0.53	0.64	1.99
	WT2	0.08	0.34	0.51	0.55	0.67	1.62
	Average	0.09	0.34	0.49	0.54	0.65	1.80
	Differenc	0.03	0.01	0.00	0.01	0.02	0.50

1st Rep	dfBl6	0.13	0.36	0.52	0.59	0.74	1.83
	df129	0.02	0.35	0.53	0.63	0.73	4.73
	Average	0.08	0.36	0.52	0.61	0.73	3.28
2nd Rep	dfBl6	0.10	0.34	0.48	0.62	0.69	4.04
	df129	0.18	0.36	0.50	0.57	0.71	2.44
	Average	0.14	0.35	0.49	0.59	0.70	3.24
	Differenc	0.06	0.00	0.03	0.02	0.03	0.04

1st Rep	dpBl6	0.15	0.41	0.54	0.62	0.74	1.79
	dp129	0.06	0.37	0.56	0.62	0.76	2.47
	Average	0.10	0.39	0.55	0.62	0.75	2.13
2nd Rep	dpBl6	0.06	0.38	0.56	0.63	0.72	3.07
	dp129	0.05	0.40	0.57	0.63	0.79	1.40
	Average	0.05	0.39	0.57	0.63	0.75	2.23
	Differenc	0.05	0.00	0.02	0.01	0.00	0.10

**Table 3.2 Descriptive statistics of compaction measurements between 3D DNA FISH of two biological replicates**

BAC sets analyzed include **A)** 2 and **B)** 10. In the case of the  $+^{129}/+^{Bl6}$  genotype, alleles are denoted as WT1 and WT2, as we cannot distinguish the wild-type chromosome 4 from the 129S5/SvEv<sup>Brd</sup> and C57Bl6/J strains. Therefore, WT1 and WT2 can be a mixture of 129S5/SvEv<sup>Brd</sup> and C57Bl6/J chromosomes (additional permutation tests on the data yield the same significance results, data not shown). For the  $df/+^{Bl6}$  and  $dp/+^{Bl6}$  genotypes, WT denotes the  $+^{Bl6}$  chromosome. Note the agreement between overall descriptive statistic values. Values marked as 0.00 are 0.001, rounded up.

Sample	Allele	Min	1st Qu	Median	Averag	3rd Qu	Max
P4	WT1	0.14	0.32	0.50	0.58	0.73	3.02
	WT2	0.10	0.34	0.48	0.52	0.63	1.58
	Averag	0.12	0.33	0.49	0.55	0.68	2.30
P3	WT1	0.17	0.37	0.48	0.53	0.67	1.34
	WT2	0.12	0.31	0.42	0.49	0.58	1.81
	Averag	0.14	0.34	0.45	0.51	0.63	1.58
P2	WT1	0.08	0.31	0.48	0.52	0.68	1.55
	WT2	0.08	0.34	0.45	0.53	0.64	2.50
	Averag	0.08	0.32	0.47	0.53	0.66	2.03
Differenc	P4-P3	0.02	0.01	0.04	0.04	0.05	0.73
	P4-P2	0.04	0.01	0.02	0.02	0.02	0.28
	P3-P2	0.06	0.02	0.01	0.02	0.03	0.45

Sample	Allele	Min	1st Qu	Median	Averag	3rd Qu	Max
P4	dfBI6	0.13	0.36	0.52	0.59	0.74	1.83
	df129	0.02	0.35	0.53	0.63	0.73	4.73
	Averag	0.08	0.36	0.52	0.61	0.73	3.28
P3	dfBI6	0.08	0.33	0.46	0.54	0.66	2.33
	df129	0.14	0.33	0.43	0.50	0.59	2.44
	Averag	0.11	0.33	0.45	0.52	0.62	2.39
P2	dfBI6	0.08	0.33	0.44	0.52	0.61	4.42
	df129	0.08	0.29	0.39	0.51	0.59	3.68
	Averag	0.08	0.31	0.41	0.52	0.60	4.05
Differenc	P4-P3	0.03	0.02	0.08	0.09	0.11	0.89
	P4-P2	0.00	0.05	0.11	0.09	0.13	0.77
	P3-P2	0.03	0.02	0.03	0.00	0.02	1.66

Sample	Allele	Min	1st Qu	Median	Averag	3rd Qu	Max
P4	dpBl6	0.15	0.41	0.54	0.62	0.74	1.79
	dp129	0.06	0.37	0.56	0.62	0.76	2.47
	Averag	0.10	0.39	0.55	0.62	0.75	2.13
P3	dpBl6	0.09	0.36	0.48	0.55	0.66	1.56
	dp129	0.06	0.38	0.50	0.70	0.69	4.67
	Averag	0.08	0.37	0.49	0.62	0.68	3.11
P2	dpBl6	0.09	0.40	0.51	0.65	0.72	4.74
	dp129	0.20	0.42	0.57	0.81	0.79	4.62
	Averag	0.14	0.41	0.54	0.73	0.76	4.68
Differenc	P4-P3	0.03	0.02	0.06	0.00	0.07	0.99
	P4-P2	0.04	0.02	0.01	0.11	0.01	2.55
	P3-P2	0.07	0.04	0.05	0.10	0.08	1.57

**Table 3.3 Descriptive statistics of compaction measurements between 3D DNA FISH of 3 different MEF passages**

Absolute differences between descriptive statistics of averaged compaction measurements for cells in passages 2, 3 and 4 (P2, P3, and P4) for BAC set 10 in  $+^{129/+^{Bl6}}$ ,  $df/+^{Bl6}$ ,  $dp/+^{Bl6}$  MEFs. In the case of the  $+^{129/+^{Bl6}}$  genotype, alleles are denoted as WT1 and WT2, as we cannot distinguish the WT chromosomes 4 from  $129S5/SvEv^{Brd}$  and C57Bl6/J. Therefore, WT1 and WT2 can be a mixture of  $129S5/SvEv^{Brd}$  and C57Bl6/J alleles (additional permutation tests on the data give the same significance results, data not shown). For the  $df/+^{Bl6}$  and  $dp/+^{Bl6}$  genotypes, WT denotes the  $+^{Bl6}$  chromosome. Note the small differences between overall descriptive statistic values. Values marked as 0.00 are 0.001, rounded up.

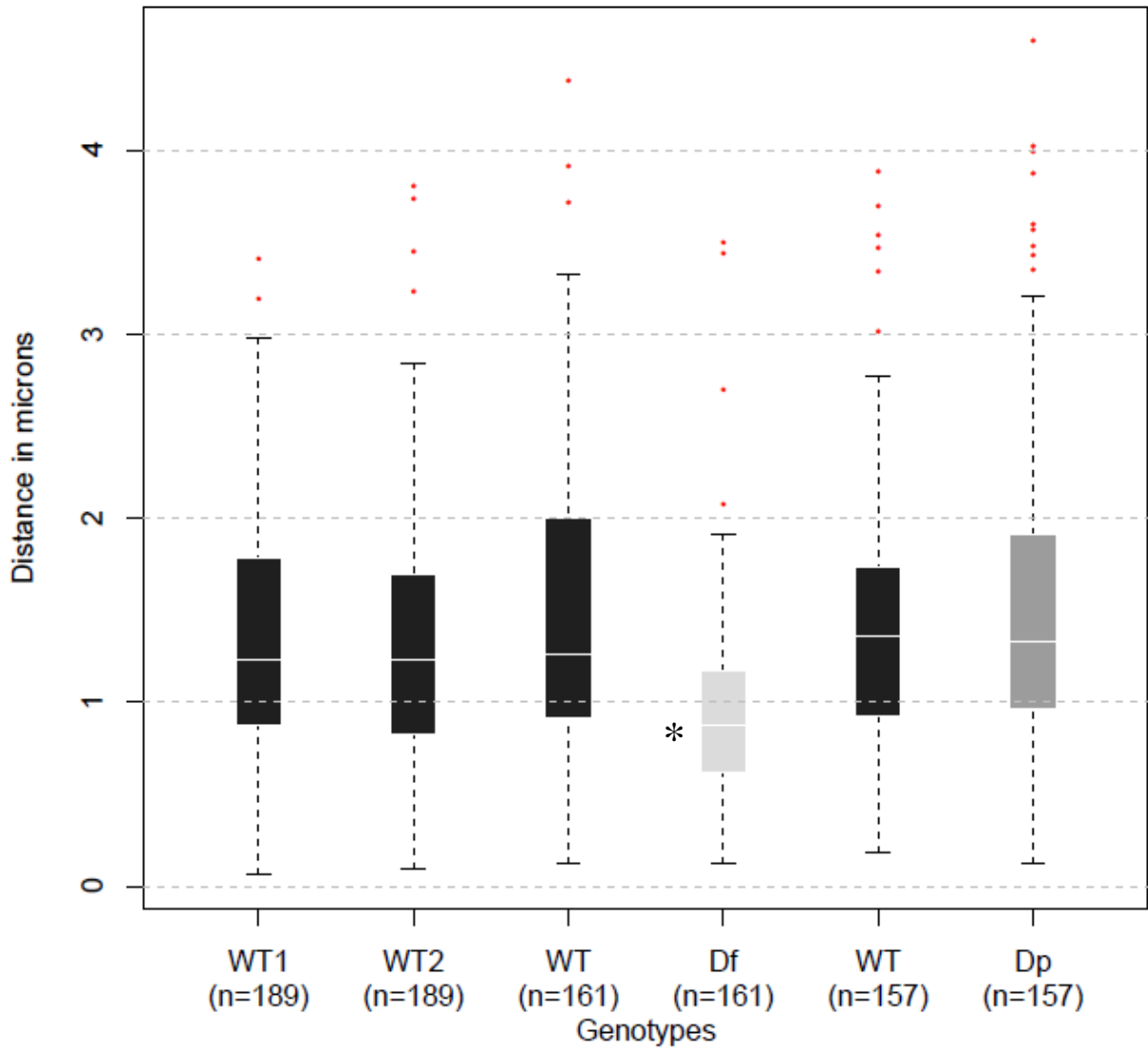


<b>BAC set</b>	<b>WT</b>	<b>df/+</b>	<b>dp/+</b>
<b>2</b>	116	142	62
<b>3</b>	168	125	73
<b>4</b>	116	123	76
<b>5</b>	141	119	77
<b>6</b>	95	107	67
<b>7</b>	139	108	51
<b>9</b>	110	104	81
<b>10</b>	144	106	70
<b>CNV borders</b>	189	161	157
<b>Gapdh</b>	111	119	88
<b>Rps13</b>	133	94	60

**Table 3.4 Summary of total cells included in the present 3D DNA FISH analysis per genotype and BAC set.**

*Gapdh* gene is present in chromosome 6, while Rps13 is located in chromosome 7.

### Calculated Distance Between FISH Probes



**Figure 3.5 Chromatin compaction differences between probes bordering the deletion CNV in *df* chromosomes**

Notice the significant difference between the distances separating probes in *df* chromosomes (marked with an asterisk in Df, light grey bar) and the rest of the WT chromosomes in  $+^{129}/+^{Bl6}$ , *df*/ $+^{Bl6}$ , and *dp*/ $+^{Bl6}$  MEFs. Interestingly, *dp* chromosomes do not show significant changes in distances separating both BACs, despite the presence of an additional 4.3Mb segment. This is probably related to the appearance of a new 3D arrangement adopted by the *dp* chromosome in *dp*/ $+^{Bl6}$  MEFs (see the discussion at the end of this chapter).

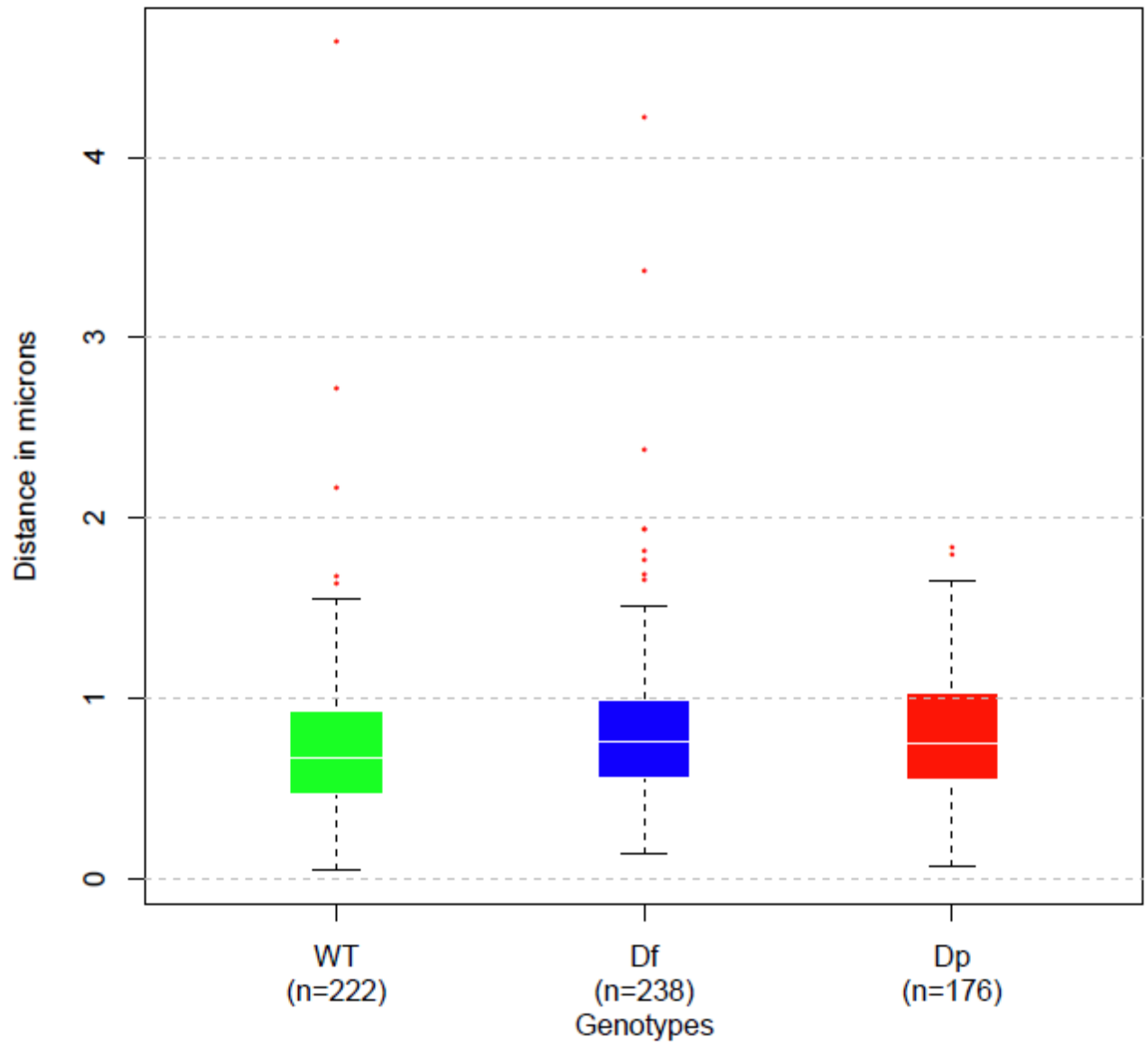
Additionally, the plugin data revealed that 4 out of the 8 BAC regions have difference of 10% or more in ratios of FISH signals overlapping heterochromatin foci in  $df/+^{Bl6}$  or  $dp/+^{Bl6}$  MEFs compared to  $+^{129}/+^{Bl6}$  (regions 4,5,9,10) [Table 3.5]. In terms of nuclear positioning, only BAC region 4 on the  $dp$  chromosome seems to change its preferential location towards being more peripheral (10-16% change compared to its homologous  $+^{Bl6}$  and chromosome) [Supp. Table 3.3A,B,C].

### 3.6 Discussion of 3D DNA FISH results

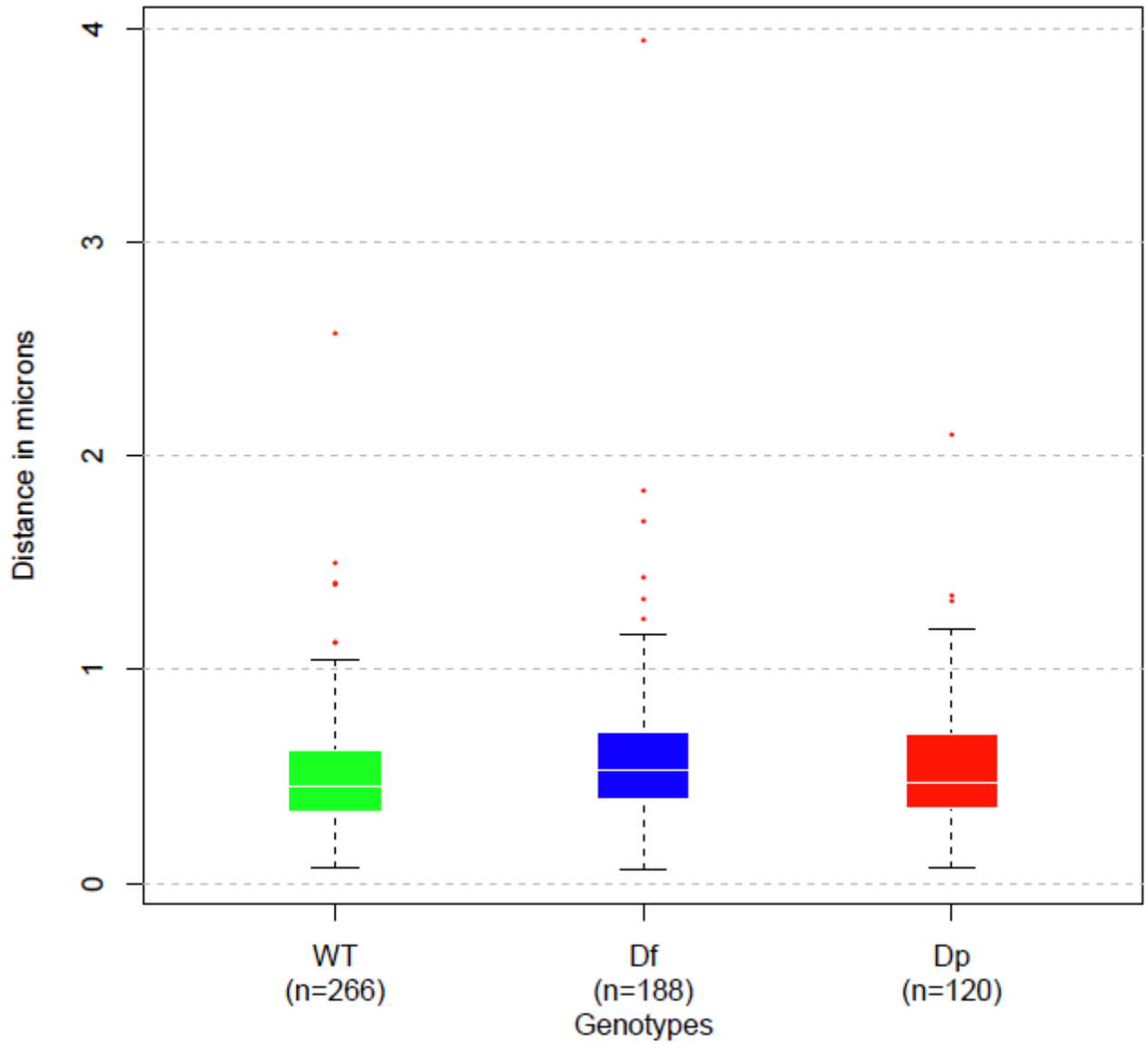
Completion of the first aim of this project allowed me to assess basic chromatin properties of the mouse 4E region in its WT and after the occurrence of deletion and duplication CNVs. 3D DNA FISH probes designed 0.5Mbs apart from each other have an average distance of 0.6 $\mu$ m for  $+^{129}/+^{Bl6}$  and  $df/+^{Bl6}$  MEFs, and 0.76 $\mu$ m for  $dp/+^{Bl6}$  MEFs for all regions analyzed. While  $dp/+^{Bl6}$  nuclei tend to have larger volumes compared to  $+^{129}/+^{Bl6}$  and  $df/+^{Bl6}$  nuclei, and  $df/+^{Bl6}$  nuclei tend to be smaller compared to  $+^{129}/+^{Bl6}$  and  $dp/+^{Bl6}$ , there are no significant correlations between measured chromatin compaction and nuclear volumes. We were subsequently able to make direct comparisons between results without the inclusion of genotype-dependent measuring biases.

Careful analysis of FISH data revealed that several regions analyzed with the different BAC probe sets display small chromatin compaction differences. However, these fall below the microscopic resolution range (<200nm) and were therefore excluded from further analysis. Interestingly, BAC region 4 was the only neighboring CNV site that displayed changes in chromatin compaction in  $dp/+^{Bl6}$  MEFs >200nm resolution limit.

A)



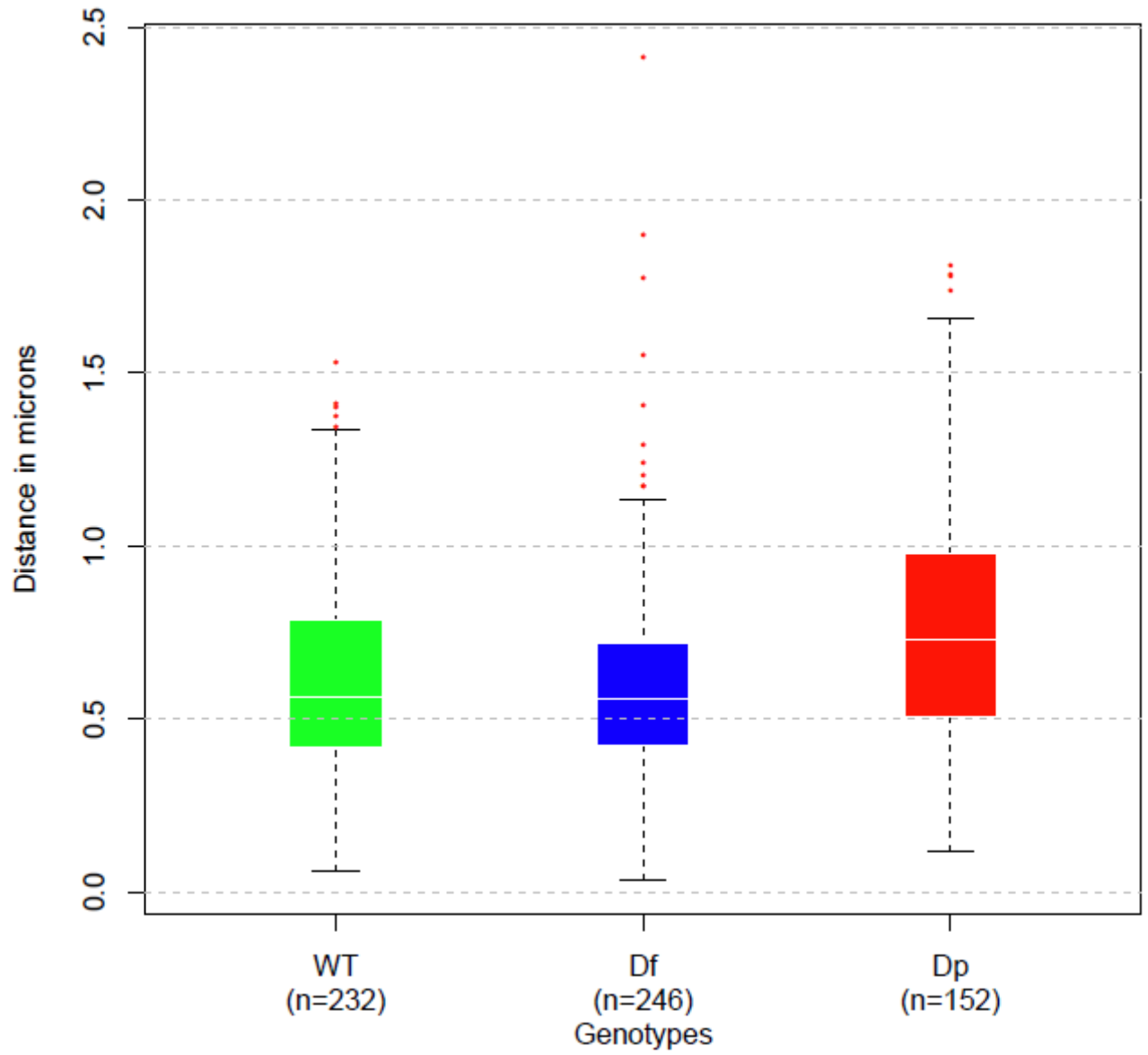
**B)**



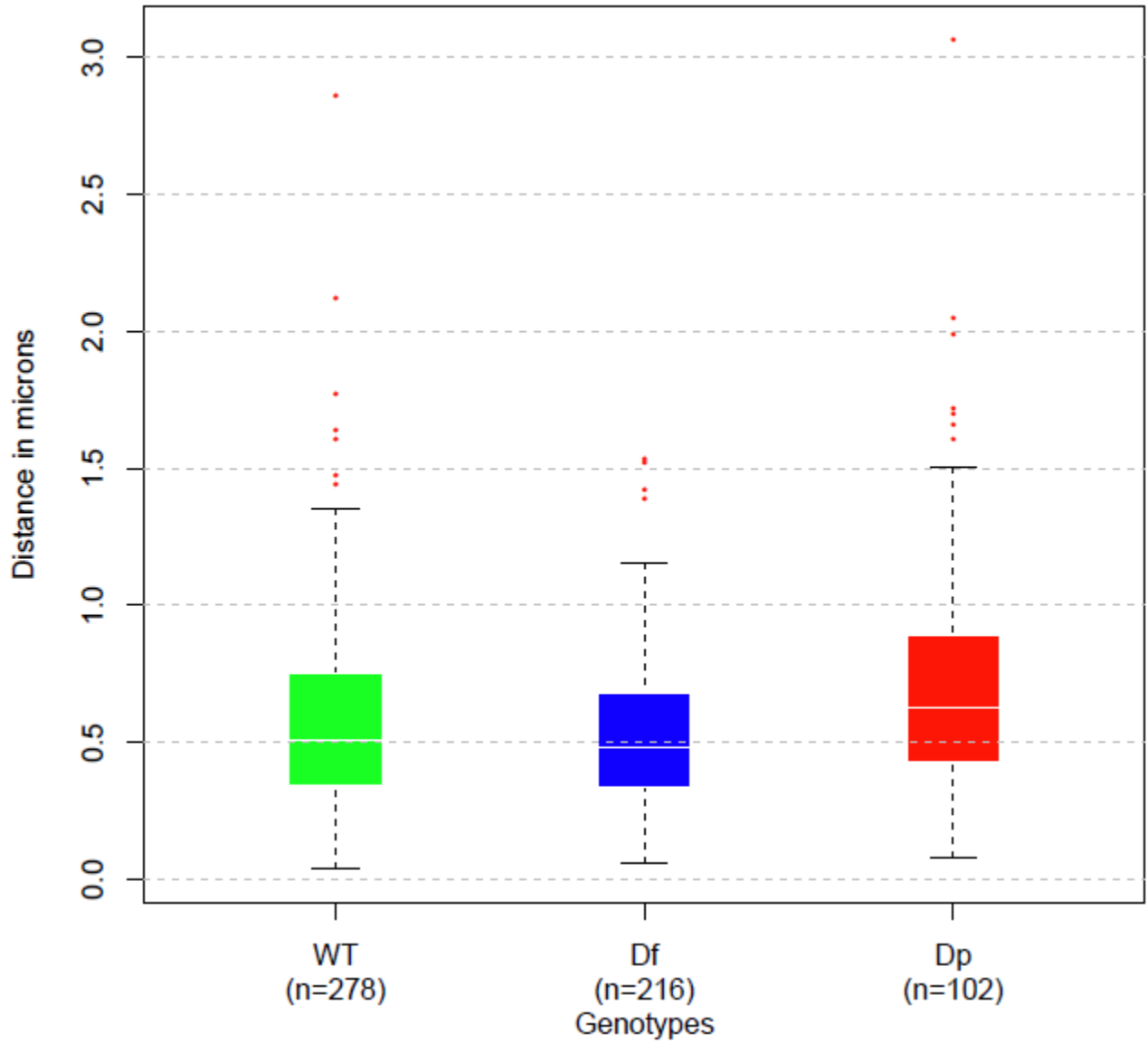
**Figure 3.6 Chromatin compaction distributions of BAC sets in control regions**

**A)** *Gapdh* and **B)** *Rps13*. Aggregate allelic values are displayed per genotype. WT =  $+^{129}/+^{Bl6}$ , Df =  $df/+^{Bl6}$ , and Dp =  $dp/+^{Bl6}$ .

A)



**B)**

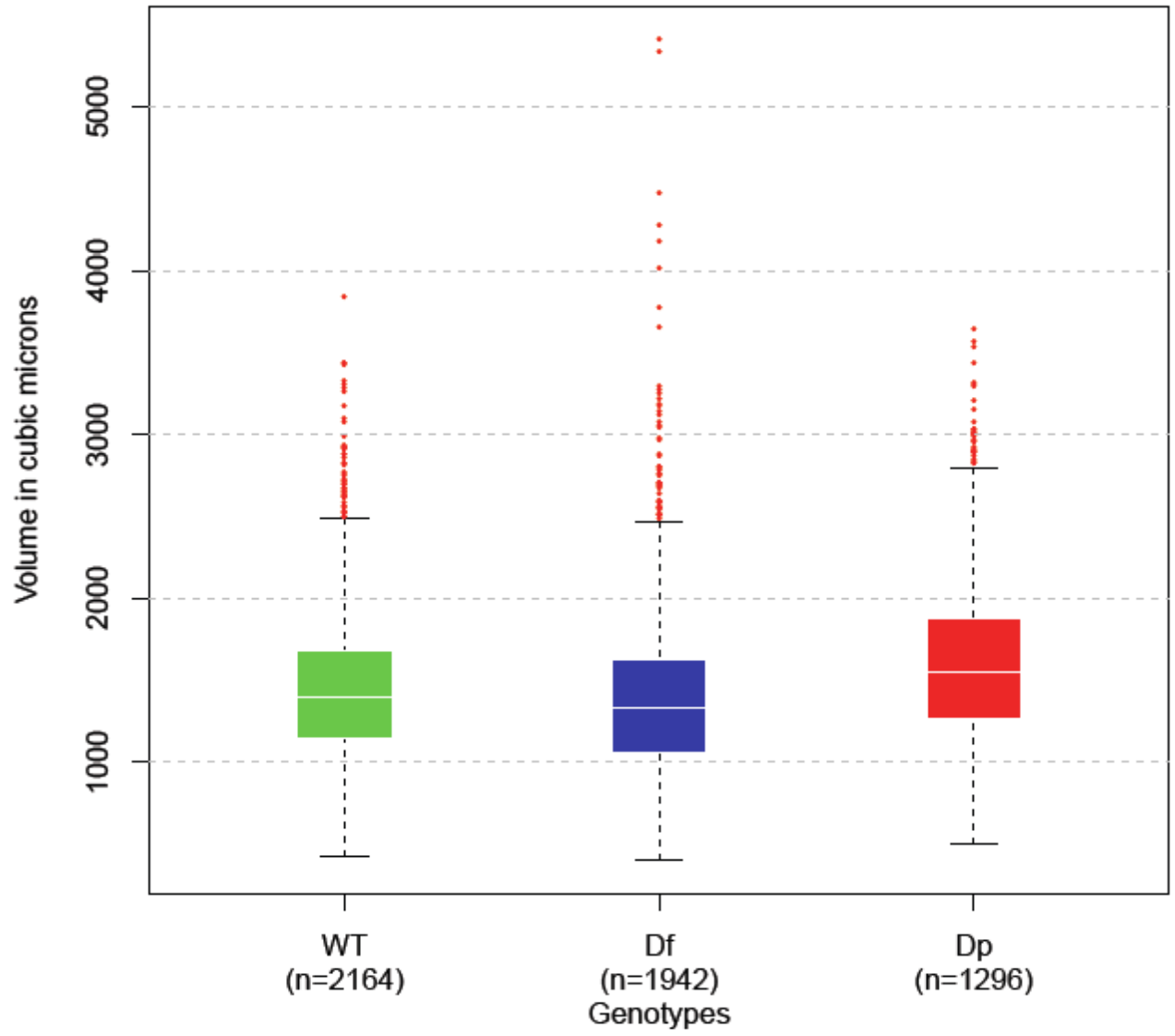


**Figure 3.7 Chromatin compaction distributions of BAC sets 4 and 7**

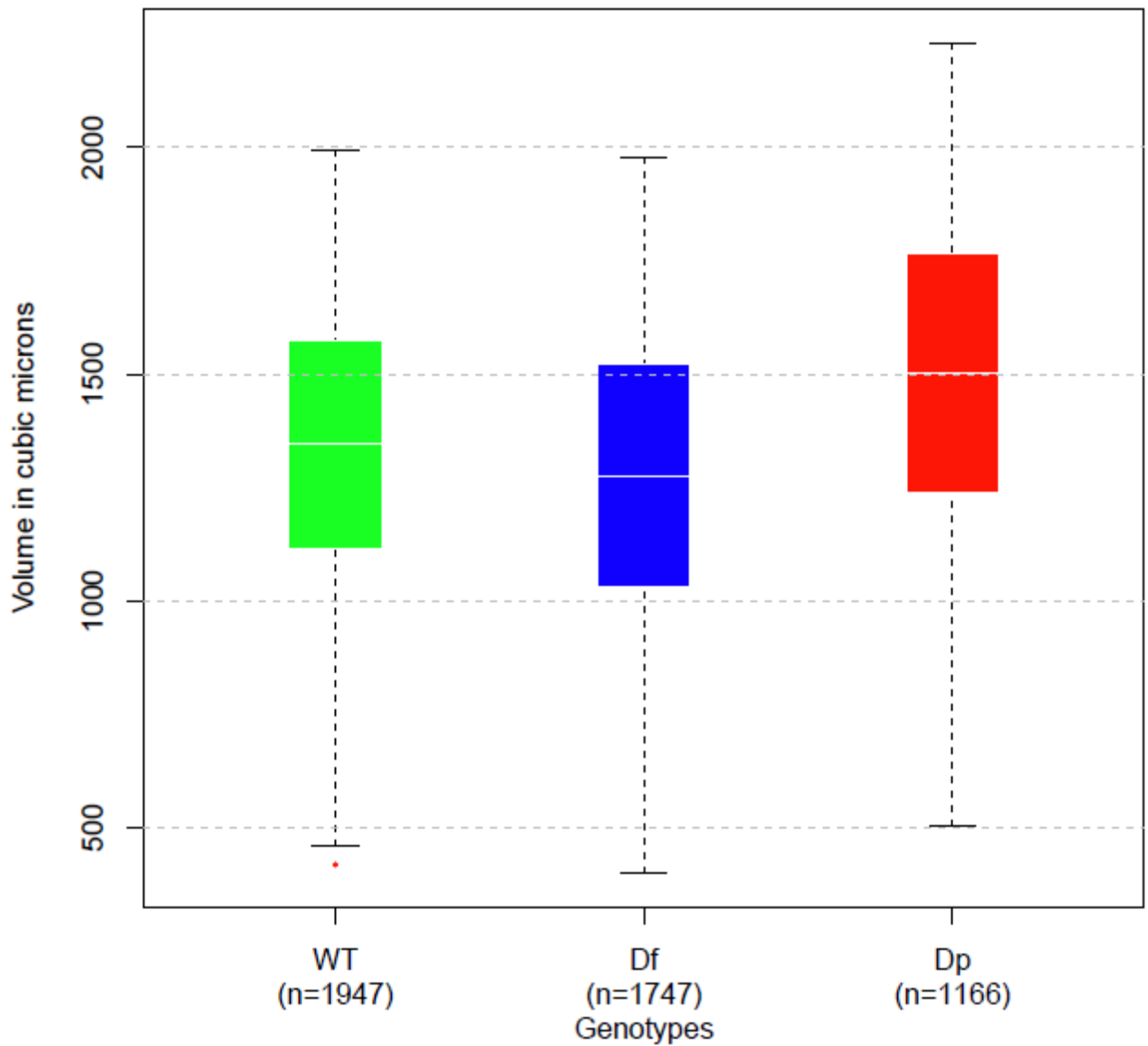
A) 4 and B) 7. Aggregate allelic values are displayed per genotype. WT =  $+^{129}/+^{Bl6}$ , Df =  $df/+^{Bl6}$ , and Dp =  $dp/+^{Bl6}$ .



A)



**B)**



**Figure 3.8 Nuclei volume differences between the analyzed MEFs**

A) all data gathered, and B).1-9 quantiles for clearer volume visualization. Notice how  $dp/+^{Bl6}$  MEFs tend to have larger nuclear volume compared to  $+^{129}/+^{Bl6}$  and  $df/+^{Bl6}$  cells.

	WT			df/+			dp/+		
BAC set	Chann1	Chann2	Chann3	Chann1	Chann2	Chann3	Chann1	Chann2	Chann3
2	16.4	15.5	11.2	9.9	13.0	7.7	8.1	10.5	8.1
3	10.7	13.1	7.4	13.2	14.8	13.6	8.9	7.5	7.8
4	22.0	22.8	15.5	15.0	11.8	8.1	9.2	13.8	4.4
5	17.0	20.6	12.8	12.6	16.4	9.2	8.4	10.4	10.4
6	10.5	7.4	8.9	12.1	12.1	7.5	9.0	7.5	7.0
7	14.0	13.3	14.0	11.6	13.0	9.3	15.7	17.6	10.5
9	20.0	14.1	13.2	8.2	12.5	9.6	7.4	9.3	7.0
10	13.2	19.8	12.8	9.4	19.3	34.9	4.3	19.3	11.4
11	9.9	14.9	11.3	6.7	9.7	11.8	9.7	9.1	9.5
12	13.5	9.8	9.4	10.1	8.5	11.7	7.5	11.7	8.9
13	15.6	12.9	11.7	12.2	12.2	9.9	8.3	12.2	8.7
14	11.3	16.3	12.7	11.0	15.2	10.5	5.4	8.5	6.7
15	14.6	27.2	18.3	15.0	19.2	16.7	10.7	21.3	8.9
16	16.1	23.0	14.1	18.8	28.1	14.8	10.2	22.2	12.1
17	10.6	18.5	10.3	11.2	13.4	10.6	9.2	11.8	9.3

**Table 3.5 Heterochromatin overlap ratios per channel per BAC set and genotype**

Marked in yellow are the signals which exceeded a 10% difference compared to +<sup>129</sup>/+<sup>Bl6</sup> heterochromatin overlap ratios.

BAC region 4 is located ~16Mb away from the CNV start, and presented different compaction, heterochromatin overlap, and nuclear positioning values compared to  $+^{129}/+^{Bl6}$  and  $df/+^{Bl6}$  cells. There are two genes inside the probes that border this region: *Runx3* and *Clic4*. *Runx3* belongs to the runt domain family of transcription factors that act as master regulators of gene expression in major developmental pathways. It has been shown to be involved in neurogenesis of the dorsal root ganglia, T-cell differentiation and tumorigenesis of gastric epithelium (Bae and Choi, 2004). On the other hand, *Clic4* is an intracellular chloride ion channel protein expressed in many tissues, and exhibits an intracellular vesicular pattern in Panc-1 cells (pancreatic cancer cells). Interestingly,  $dp/+^{Bl6}$  MEFs enter a senescence program after passage 5, and have bigger nuclear volumes compared to  $+^{129}/+^{Bl6}$  at passage 4, which could be related to the tumor suppressor activity of *Runx3* and the fundamental process of cell volume regulation by chloride channels (*Clic4*). However, RNA-Seq results (Chapter 5), did not reveal any statistically significant expression change for these genes in  $dp/+^{Bl6}$  MEFs compared to  $+^{129}/+^{Bl6}$ .

As expected, chromatin compaction measurements differ the most between the *df* and the  $+^{129}$  and  $+^{Bl6}$  chromosomes when using probes bordering the deletion CNV. Nevertheless, it was unforeseen that duplication of the same region did not alter significantly the distances separating these probes. We hypothesize this might be due a possible looping of the duplicated region out of the preferred chromosomal conformation state, but this idea was not further tested after the decision to not continue with the study of  $dp/+^{Bl6}$  MEFs.

$dp/+^{Bl6}$  MEFs were excluded from the study after performing the 3D DNA FISH experiments and a couple of molecular chromatin studies (Chapter 4).  $dp/+^{Bl6}$  MEFs grow slowly and in reduced numbers compared to  $+^{129}/+^{Bl6}$  and  $df/+^{Bl6}$  MEFs, and we observed a

high ratio of 2 instead of 3 classification signals for the identification of the *dp* chromosome. These numbers scaled to up to 50% in all counted cells on FISH experiments (data not shown), and a subsequent observation of a deviant ratio of counted 4C captures for viewpoint 154.9 for the  $+^{Bl6}$  chromosome in *dp/+<sup>Bl6</sup>* MEFs (Chapter 4). These observations suggest a possible loss of the duplicated fragment. With this in mind, the biological relevance of the changes observed in transcription and chromatin conformation for the *dp* chromosome could be confounded by this loss of the duplicated region, as both  $+^{129/+<sup>Bl6</sup>}$  and *dp/+<sup>Bl6</sup>* MEFs would be present in the culture. From now on, all of the analyses presented will focus on  $+^{129/+<sup>Bl6</sup>}$  and *df/+<sup>Bl6</sup>* MEFs comparisons unless otherwise specified.

In summary, 3D DNA FISH studies showed no gross changes in chromatin architecture for the 4E analyzed regions at the Mb length scale, except for the rearranged deleted segment in chromosome *df*. While chromatin conformation changes could occur below the 200nm fluorescence microscope resolution limit, or in other regions not covered by our BAC probe sets, these observations point out to the existence of obvious specific chromatin organization changes mostly present surrounding the deletion CNV in the *df* chromosome. The extent of these changes and their transcriptional impact will be presented and discussed in Chapter 4.

## **Chapter 4: Molecular characterization of higher-order chromatin organization in a 4E2 deletion CNV**

The initial microscopic characterization of chromatin in the 4E2 region in its WT and deletion states pointed out the existence of regional changes in conformation after the occurrence of a deletion CNV in *df* chromosomes. Additionally, potential long-range distance alterations were highlighted, which needed to be assessed at a higher resolution to overcome fluorescence microscopy detection limits.

In order to evaluate chromatin interactions at a genome-wide scale for the 4E2 region, we used an allele-specific chromosome conformation capture strategy (PE-4Cseq. Holwerda *et al.*, 2013; de Wit *et al.*, 2013) to characterize the higher-order chromatin architecture of 4E2 in its heterozygous WT state ( $+^{129}/+^{Bl6}$ ) and upon a 4.3Mb deletion ( $df/+^{Bl6}$ ). We chose to analyze the heterozygous deletion CNV genotype as Monosomy 1p36 patients are heterozygous for this region (Heilstedt *et al.*, 2003), and heterozygous deletions in 1p36 are associated with cancer progression/maintenance (Bagchi and Mills, 2008, and references therein). Additionally, heterozygote genotypes offer the advantage to study phenomena such as transcriptional dosage compensation, which could be related to changes in chromatin organization.

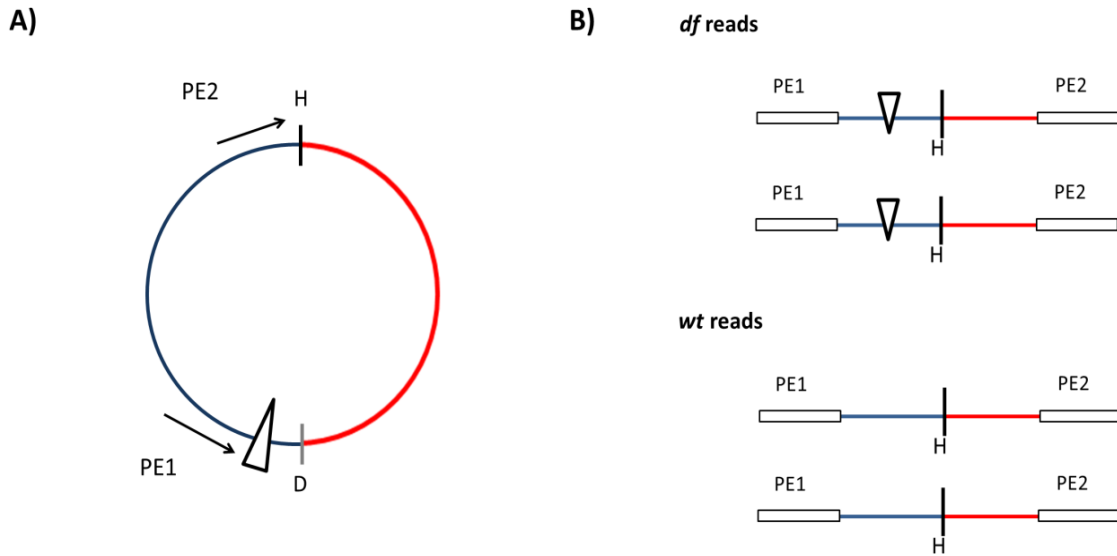
### **4.1. PE-4Cseq measurement of 4E2 chromatin contacts**

PE-4Cseq, explained in detail in Chapter 1, performs a second restriction digestion on the ligated 3C template using a frequent cutter enzyme (i.e. 4bp recognition sequence).

Products are subsequently ligated under dilute conditions to generate small DNA template circles [See Chapter 1, Fig. 1.4A,B,C]. Performing an inverse PCR reaction using primers targeting a specific genomic region (“viewpoint”), interacting sequences (“captures”) can be amplified and their identities determined by the use of DNA PE sequencing. PE sequencing allows the distinction of allelic origin based on the amplification of a genotyping SNP with one of the reads [Fig. 4.1A]. This methodology provides higher resolution and ability to detect intra-chromosomal (*cis*) as well as inter-chromosomal (*trans*) interactions in an allele-specific manner.

The positions of *HindIII-DpnII* restriction fragments along chromosome 4 were queried and only those which overlapped high-confidence SNPs between the C57BL6/J and 129S5/SvEv<sup>Brd</sup> genomes (Keane *et al.*, 2011) were considered for potential viewpoints [Supp. Table 4.1]. A total of 12 viewpoints spanning 4E2 were selected, and SNP presence validated by Sanger sequencing [Fig. 4.2. Supp. Fig. 4.1. Supp. Table 4.2]. Two viewpoints span ~83Mb upstream of the deletion CNV start, eight are located inside the deletion, and two cover ~1Mb downstream of the deletion end. Additionally, we amplified two control viewpoints: one in chromosome 4, located ~83Mb away from the deletion start, and a second one in chromosome 7, covering the *Rps13* housekeeping gene.

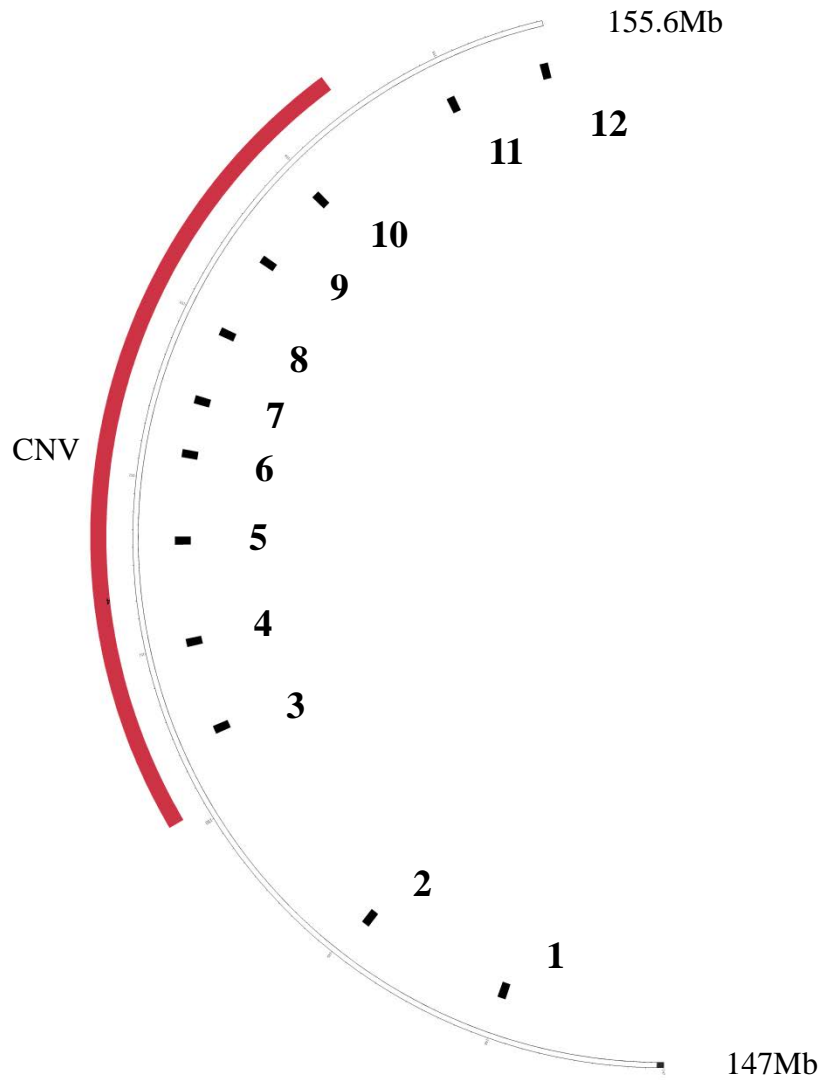
Each viewpoint was amplified from two biological replicates of *df/+<sup>Bl6</sup>* (129S5E71, 129S5E98) and *+<sup>129</sup>/+<sup>Bl6</sup>* (129S5E117, 129S5E118) using barcoded primers with Illumina PE sequencing adaptors [Supp. Table 4.3]. The use of barcodes included in the primers allowed us to pool one entire set of *df/+<sup>Bl6</sup>* viewpoints with another from *+<sup>129</sup>/+<sup>Bl6</sup>* MEFs to minimize technical biases when performing data comparisons.



**Figure 4.1 Allelic assignments of chromatin interactions by PE-4Cseq**

**A)** Amplification of a genotyping SNP and its corresponding capture. **B)** Amplified reads can be categorized as 129S5/SvEv<sup>Brd</sup> or C57Bl6/J-derived based on the identity of the SNP amplified in the first read of this example.





**Figure 4.2 Circular depiction of mouse of region 147-155.6Mb from chromosome 4**

CNV region is depicted as the outer red box towards the telomere. The most internal black lines correspond to 4C viewpoints used for this study, numbered as used in the subsequent analyses.

Library pooling and quantification was performed by isolating amplified viewpoint PCR products with AMPure beads (.8-.9X concentration) to eliminate primer dimer contamination. Product size quantifications were performed on Bioanalyzer DNA chip 1000, and molar concentrations calculated with KAPA (KAPA Library Quant kit, Kapa Biosystems) qPCR reactions of each viewpoint. A total of two Illumina HiSeq 200 PE 100 lanes were run for both replicates, and reads further processed for analysis as discussed below.

## 4.2. PE-4Cseq data filtering and read mapping

Each Illumina HiSeq 200 PE 100 lane with amplified viewpoints from *df/+<sup>Bl6</sup>* and *+<sup>129</sup>/<sup>Bl6</sup>* 4C templates was further processed with custom-made perl scripts (see Chapter 7 & 8 for details). The script *split\_fastq\_withqual.pl* identifies reads belonging to each viewpoint based on the *HindIII/DpnII* primer reading sequence and barcodes, and the script *split\_4C\_snp\_withqual.pl* identifies the genotyping SNP and categorizes each read's allelic origin. We obtained an average of ~1M reads per allelic viewpoint for the first biological replicate, and ~2.2M reads per viewpoint for the second biological replicate [Supp. Table 4.4A,B]. Capture reads were trimmed to 30bp to have the highest quality bases for downstream analyses.

Reads were subsequently mapped to a reduced *HindIII* mouse mm9 database which contains 150bp of sequence upstream and downstream *HindIII* cutting sites. This strategy has been used in previously published 4C studies (Simonis *et al.*, 2006; Noordermeer *et al.*, 2008, 2011; Splinter *et al.*, 2011; van de Werken *et al.*, 2012; Holwerda *et al.*, 2013; de Wit *et al.*, 2013), given that ligations will always occur on *HindIII* restriction sites and primers

directly flank the restriction/ligation regions. Alignments were performed using bowtie with up to 3 mismatches accepted per read to account for SNPs in the 129S5/SvEv<sup>Brd</sup> sequence, and only uniquely mapped reads were taken into account for further analysis. An example of the distribution of reads for the first biological replicate is shown in Fig. 4.3.

### **4.3. Quantitative analysis of PE-4Cseq data**

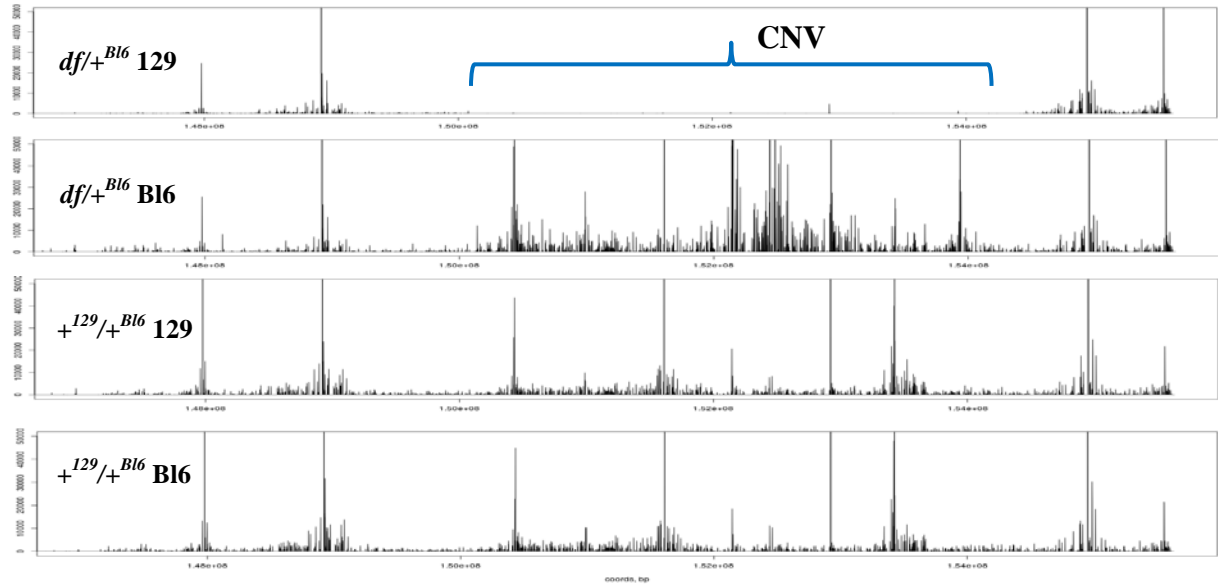
One of the most important aspects in the development of this project concerns the quantitative analysis of 4E2 PE-4Cseq data. Several sources of bias have been previously reported for 3C-derived methodologies. For example, differences in restriction digestion and ligation efficiencies, primer amplifications, PCR biases, viewpoints and captures GC content, sequencing biases, and restriction fragment sizes, all contribute to variations between the final numbers of reads obtained per viewpoint per sample. These biases (plus several other unaccounted factors), make quantitative comparisons between samples quite challenging.

Previously published 4C studies analyzed data using statistical procedures (Simonis *et al.*, 2006; Splinter *et al.*, 2011; de Wit *et al.*, 2008; van de Werken *et al.*, 2012). These include running window approaches to smoothen the data, whose output results are robust indicators of domain interactions (Simonis *et al.*, 2006). The most recent modification to the running window approach includes data binarization (i.e. assigning a value of one to restriction fragments with mapped reads, and zero to non-captured ones). The binarization is performed on the data in order to diminish the impact of biases on 4C data interpretation and comparison. Typical background windows sizes are 3001 fragments in length, and running windows are 100 fragments. For each window, one-tailed binomial tests can be applied,

testing if the number of occurrences is greater than expected based on background window values. Obtained p-values are typically used to build “domainograms,” which display the regions of interaction with a specific viewpoint that surpassed probability background values. While statistical strategies have been successfully used for the analysis of previously published 4C-Seq and PE-4Cseq data, our project faced a much more subtle and important question to be addressed: after the occurrence of a deletion CNV, are the observed changes in chromatin interactions derived from the shortening of the chromatin fiber itself, or due to genuine formation/disappearance of chromatin contacts?

Numerous studies have shown that chromatin is a dynamic structure that undergoes diffusive motion within the nucleus (Marshall *et al.*, 1997; Tumbar and Belmont, 2001; reviewed in Spector, 2003). Chromatin fibers can be modeled as polymers (i.e. “beads-on-a-string” configuration, with nucleosomes as beads, and linker DNA as the string) whose dynamics can be described and predicted by equations. Polymer physics establishes that a defined chromatin region will interact with its surrounding sequences in a way that is proportional to the separation of both sites and the flexibility of the intervening chromatin sequence (reviewed in Mirny, 2011). These contacts are derived from the entropy produced by arranging the chromatin fiber into specific conformations, and therefore constitute a background state of interactions for any region across the genome.

In quantifying the effects of a deletion on *cis* chromatin interactions, we must discriminate the contribution of the shortening of the chromatin fiber from genuine changes in chromatin interactions, such as switching/maintenance of promoter contacts with regulatory elements, the occurrence of new contacts determined by architectural protein binding, altered chromatin tethering effects, etc. [Fig. 4.4].



**Figure 4.3** Raw mapped reads for the  $df/+^{Bl6}$  (129S5E71) and  $+^{129}/+^{Bl6}$  (129S5E117) first biological PE-4Cseq replicates.

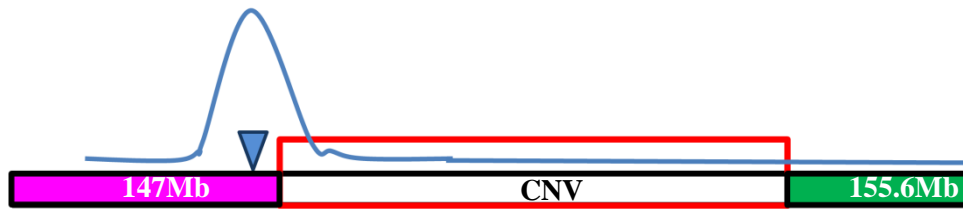
Display spans 147-155.6 Mb in chromosome 4. Viewpoints can be seen as peaks of mapped reads. Notice the reduction of reportable reads in chromosome  $df$  for regions 150-154.4 (CNV region, marked in blue), consistent with its sequence deletion. Reported mapped reads inside these regions correspond to sequences where there were sequencing errors. We calculated a  $<0.05\%$  sequencing error rate for both PE-4Cseq lanes based on numbers derived from these viewpoints, which made the specified bp be considered as one of our genotyping SNPs.

In order to address this question, we developed a new methodology, grounded on polymer physics, for bias reduction, data normalization, and differential analysis of contact probability signal across multiple 4C viewpoints and genotypes. This pipeline, skillfully developed and implemented by Swagatam Mukhopadhyay, a physicist at CSHL, corrects for data biases already discussed in the literature within the context of Hi-C (Yaffe and Tanay, 2011), and others specific to PE-4Cseq. Moreover, it only reports genuine changes in chromatin interactions by comparisons to background contact probability profiles calculated from PE-4Cseq data. The use of this modeling approach allowed quantitative viewpoint comparisons to resolve differentially interacting regions across  $+^{129}/+^{Bl6}$  and *df* chromosomes.

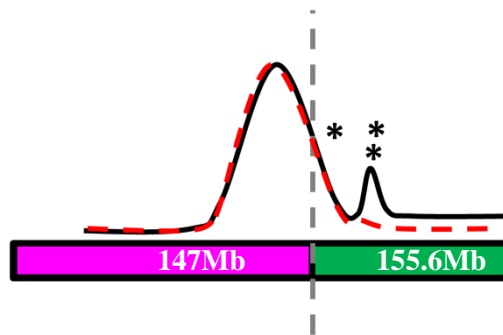
#### **4.3.1. Bias correction and data normalization across PE-4Cseq multi-viewpoints**

We developed a pipeline for the correction of biases in PE-4Cseq data. Biases are removed without modeling them individually by their technical or biological sources. There are two classes of biases in 3C-derived data. The first class has already been reported in Hi-C data analysis (Yaffe and Tanay, 2011), and includes restriction fragment lengths, fragment GC content, primer efficiencies, etc. The second class of bias is specific to multi-viewpoint 4C experiments. The sample preparation and viewpoint amplification steps for sequencing introduce unknown overall biases. The ideal output of any chromatin conformation capture method is the contact probability profile of the chromatin polymer. Such a profile can be used to assess the statistically significant differential signals between genotypes.

A)



B)



#### **Figure 4.4 Genuine and physical chromatin contact changes**

**A)** Schematic representation of background chromatin interactions for a viewpoint bordering the CNV. Most of the interactions are expected to be surrounding the viewpoint given the smaller distance separating them (blue line). The studied 4.3Mb CNV is shown as the empty red rectangle, separating two different domains of the chromosome (pink and green rectangles). Notice how interactions further downstream of the viewpoint have a low contact probability. **B)** Upon deletion, the viewpoint presents a new interaction profile (black curve) where new interactions appear (marked with asterisks). However, only the region marked with a double asterisk would be considered as a genuine change in chromatin interactions, as overlay of the previous WT background profile (red dashed curve) shows that the newly joined region (green rectangle) is simply following the expected WT background contact probability for the viewpoint.



In earlier work (Yaffe and Tanay, 2011; Fullwood *et al.*, 2009) the background of non-specific chromatin interactions arising from polymer entropy has been treated as yet another noise source. We argue that this background contains valuable signatures of local chromatin compaction in genotypes. Moreover, the significance of specific interactions can be estimated once the non-specific interactions of a viewpoint are determined. The expected non-specific interactions are generated by our polymer model—the null hypothesis being that all interactions are entropic in origin. Without such a model, it is impossible to tease apart the genuine and physical (i.e. chromatin fiber shortening) contributions of large-scale deletions.

We introduce a generalized Gaussian model for the chromatin polymer. In this model, 4C fragments  $(i, i + 1)$ , which are neighbors along the chromatin, are connected by Gaussian springs with spring constants  $k_{i,i+1}$ . The spring constants enjoy a general (not necessarily random-walk-like) scaling with their separation in DNA base-pairs,  $k_{i,i+1} \sim s_{i,i+1}^{2\nu/3}$ , where  $\nu$  is the scaling exponent for the contact probability,  $P_{ij} \sim s_{ij}^\nu$  (des Cloizeaux and Jannink, 2010) (for extensive details of the model see Chapter 7).

The polymer model dictates the form of the expected contact probability  $P_{IJ}$  between any two viewpoint regions  $I$  and  $J$ , parameterized by their genomic separation  $s_{IJ}$ . We first normalize the capture data for each experiment by the product of viewpoint and fragment lengths, and call this the *biased contact probability* (BCP). The average of this BCP in the genomic region corresponding to the viewpoints is assessed for each capture experiment, and is denoted by  $F_{IJ}$  for the pair  $(I, J)$ . We model  $F_{IJ} = C_I C_J K_I P_{IJ}$  where  $C_I$  and  $C_J$  are the overall bias factors corresponding to the viewpoint sequences, and  $K_I$  is the overall bias factor for the capture experiment of viewpoint  $I$ . Similarly, for the experiment corresponding to viewpoint  $J$ , the normalized capture data is  $F_{JI} = C_I C_J K_J P_{IJ}$ ; note that only the experiment bias factor  $K_J$

is distinct. We solve the linear (in logarithm space) set of equations to compute the bias factors  $C$  and  $K$  from  $F_{IJ}$  and  $F_{JI}$ . The unbiased estimate of the contact probability, determined up to an overall scale, is  $P_{IJ}$  and should be symmetric in  $I$  and  $J$ . In Fig. 4.5 we show that just by bias correction for neighboring viewpoints  $I$  and  $J$ , biases for all other pairs are significantly reduced and  $P_{IJ}$  is very close to being symmetric, whereas  $F_{IJ}$  is clearly not. Testing the algorithm with simulated data reproducibly recovered original  $P_{IJ}$  values even after the inclusion of significant noise biases [Supp. Figure 4.2].

### 4.3.2 Identification of differentially interacting regions in the *df* chromosome

The unbiased contact probability is obtained from BCP by normalizing with respect to  $C_I$  and  $K_I$ . The raw contact probability  $P_{I\alpha}$  between viewpoint  $I$  and fragment  $\alpha$  is obtained by  $P_{I\alpha} = \frac{F_{I\alpha}}{C_I K_I}$ , where  $F_{I\alpha}$  is the BCP. Note that there may be sequence-specific biases associated with the fragments, but there is no systematic way to correct for them in PE-4Cseq because, unlike in Hi-C, interactions between all fragments is not measured. However, because our focus is on deducing differential signals for the same chromosomal region in two genotypes, such fragment-related biases do not confound our analysis.

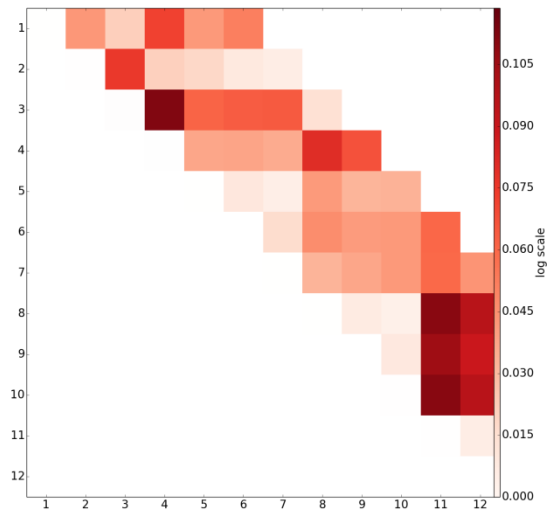
We smooth the raw contact probabilities using a Gaussian kernel with a standard deviation of 20 Kb, a length scale in the upper range of viewpoint sizes ( $\sim 0.5$ -20Kb). In Fig. 4.6A we show the comparison of contact probability profiles for all viewpoints outside of the deletion region, paired in panels for the homologous chromosomes in *df/+<sup>Bl6</sup> (del<sup>129</sup>)* and *+<sup>129</sup>/+<sup>Bl6</sup> (wt<sup>129</sup>)* genotypes. The vertical lines highlight regions where the differential signal between *del<sup>129</sup>* and *wt<sup>129</sup>* is stronger than 10%, color coded by difference-value and sign. Notice the overall similarity in contact probability patterns between *del<sup>129</sup>* and *wt<sup>129</sup>*

chromosomes, which is only altered at specific sites after the occurrence of the 4.3Mb deletion. In Fig. 4.6B an alternative visualization ('rainbow plot') of the differential signal surrounding the deletion region is presented. Each arc is color coded in the same fashion as Fig. 4.3A and represents a long-range interaction that changed in the deletion chromosome.

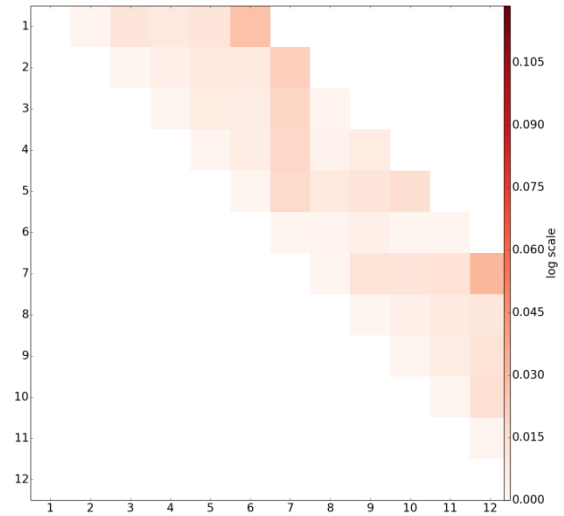
There are a total of 608 combined regions detected as differentially contacting viewpoints 1, 2, 11, and 12. These regions add up to ~35Mb of chromosome 4, meaning that almost ~22% of chromosome 4 sequence WT contact probabilities are affected after the occurrence of the 4.3Mb deletion CNV. The changes observed for viewpoint 1 mostly involve increases in contact probabilities with surrounding sequences (64% of total detected regions), while viewpoints 2 and 11 show decreases in interactions between 60% and 95% of the total detected regions. Viewpoint 12 displays equal levels of increases and decreases in interactions [Table 4.1]. The differentially interacting regions are scattered along chromosome 4, however we detected clusters of these regions neighboring the deletion up to 40Mb upstream of the CNV start. This observation suggests that there could be an underlying chromatin property which makes these regions display high contact probabilities. Mouse bands 4C7-E2 are gene rich, which suggests that one such genomic property could be the transcriptional status of the region (see Chapter 5).

The identified differentially interacting regions for chromosome *del*<sup>129</sup> constitute one level of chromatin organization, focused in the Kb scale. We can unequivocally assign the start and end of such regions with the developed PE-4Cseq analysis pipeline, and concentrate on their validation (see section 4.5). Not only did we detect regions with contact probability changes, but we also uncovered higher-order Mb scale chromatin compaction differences for the region downstream of the deletion CNV, towards the telomeric end.

**A)**



**B)**



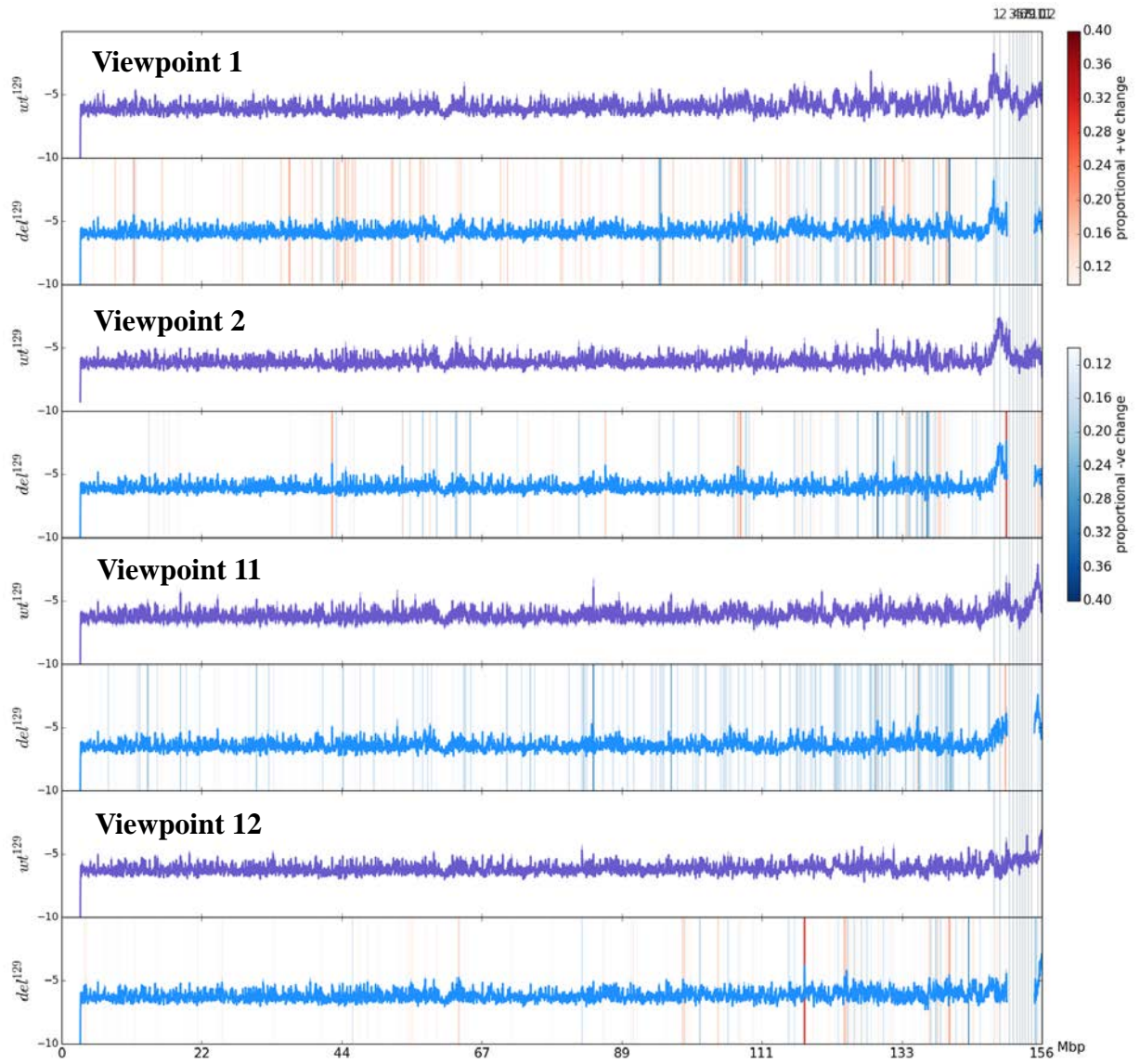
**Figure 4.5 Bias-correction for +<sup>Bl6</sup> chromosome from +<sup>129</sup>/<sup>+Bl6</sup> for all viewpoints denoted by viewpoint index in  $x$  and  $y$  axis**

**A)** The heatmap is of relative asymmetry  $\frac{|F_{IJ} - F_{JI}|}{F_{IJ} + F_{JI}}$  in BCP  $F_{IJ}$ . **B)** The relative asymmetry

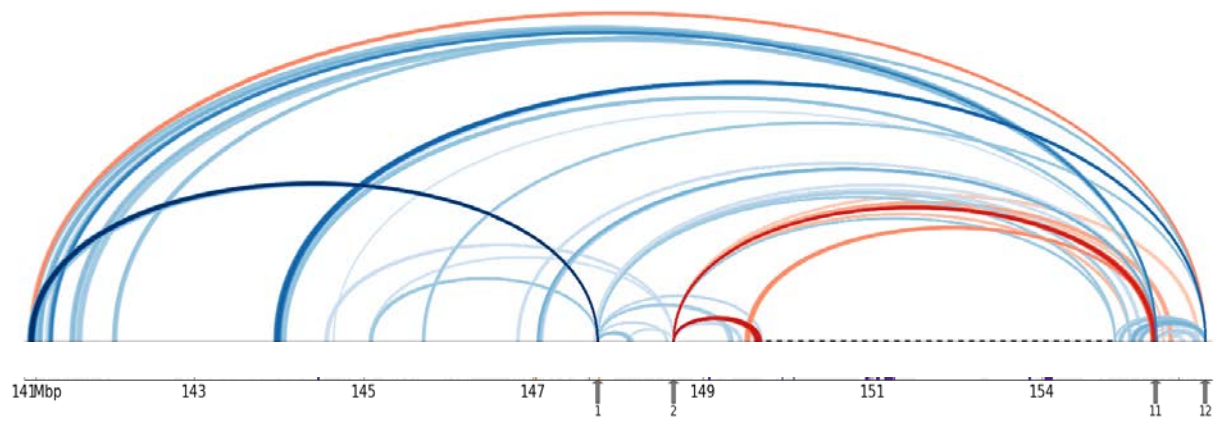
$\frac{|P_{IJ} - P_{JI}|}{P_{IJ} + P_{JI}}$  on the same scale for  $P_{IJ}$  obtained after bias-correction. Notice the reduction in both

row and column-wise biases and in the net asymmetry between viewpoints.

A)



**B)**



**Figure 4.6 Contact probability profiles for the *del*<sup>129</sup> and *wt*<sup>129</sup> in chromosome 4**

**A)** Comparison of contact probability profiles for the *del*<sup>129</sup> and *wt*<sup>129</sup> for chromosome 4. Each horizontal panel corresponds to the contact probability profiles per chromosome (blue for *del*<sup>129</sup> and purple for *wt*<sup>129</sup>) derived from PE-4Cseq data paired for viewpoints 1, 2, 11, 12 (bordering the deletion). Two biological replicates are used to assess the error profile shown as a band around the contact probability histograms. Shown in red are regions whose contact probability in *del*<sup>129</sup> chromosome is increased >10% compared to *wt*<sup>129</sup>. Shown in blue are regions whose contact probability in *del*<sup>129</sup> chromosome is decreased >10% compared to *wt*<sup>129</sup>. Notice how the majority of the changes observed in *del*<sup>129</sup> concentrate adjacently to the CNV position. Interestingly, changes in contact probabilities for viewpoints 1 and 11 extend further upstream chromosome 4, arguing for the existence of long-range effects in chromatin interactions. **B)** The differentially interacting regions on sequence 141-155.6Mb of *del*<sup>129</sup> chromosome are summarized in the rainbow plot for viewpoints bordering the deletion (grey arrows). Arcs represent long-range interactions color-coded by their strength as in Fig. 4.6A. The dashed line corresponds to the deletion region. Notice the appearance of increased contact probabilities between regions bordering the CNV after deletion.

Region	Total regs	Median	No. increase	Median	% of tot regs	No. decrease	Median	% of tot regs
Df-WT-129DiffRegs_chr4_vp_1	299	0.12	190	0.12	64	109	0.12	36
Df-WT-129DiffRegs_chr4_vp_2	186	0.12	74	0.12	40	112	0.12	60
Df-WT-129DiffRegs_chr4_vp_11	324	0.12	16	0.12	5	308	0.12	95
Df-WT-129DiffRegs_chr4_vp_12	164	0.12	79	0.12	48	85	0.11	52

**Table 4.1 Summary of median magnitude of change, direction, and number of *del*<sup>129</sup> differentially interacting regions for viewpoints 1, 2, 11, and 12.**

Notice how viewpoint 1 displays an increase in contact probabilities with surrounding sequences (64% of total detected regions), while viewpoints 2 and 11 show mostly a decrease in interactions (60% for viewpoint 2, and 95% for viewpoint 11). Viewpoint 12 has both increase and decrease in interactions in approximately the same magnitude (~50%).



#### 4.4. Changes in local chromatin compaction in the deletion chromosome

The contact probability of chromatin locally varies owing to its more extended or compact states, correlated with gene expression, epigenetic marks, loop domains, etc. (Lieberman-Aiden *et al.*, 2009; Dixon *et al.*, 2012). The contact probability between fragments at intermediate ranges of separation (10Kb-10Mb) is expected to fall off as a power law of separating length. The mean size of our *HindIII* fragments (3Kb) is the highest possible resolution of our contact probability measurements. The average power law exponent observed across the genome at similar resolutions is not given by random-walk scaling in the case of mammalian cells (Lieberman-Aiden *et al.*, 2009). Locally, however, we observe that there is considerable variability from the average scaling reported, also seen in Hi-C data heatmaps (Lieberman-Aiden *et al.*, 2009; Dixon *et al.*, 2012).

We characterized local compaction by the scaling of the contact probability in the 100Kb range. We fit a smoothing spline to the logarithm of contact probability against the logarithm of genomic separation; the slope of this curve at 100Kb is our local scaling exponent  $\nu_{\mathbf{i}}$  for viewpoint  $\mathbf{i}$  and our local measure of compaction. The local exponent  $\nu_{\mathbf{i}}$  effectively captures the changes in local compaction bracketing the deletion region for both genotypes [Fig. 4.7].

Statistically significant changes in  $\nu_{\mathbf{i}}$  are observed in the *del*<sup>129</sup> chromosome; prominently, viewpoint 11 and viewpoint 12 (both located towards the telomere end) have smaller  $\nu_{\mathbf{i}}$  values compared to *wt*<sup>129</sup>. This observation is interpreted as viewpoints 11 and 12 being *less* compact than expected from the behavior of the new neighboring regions on deletion. Viewpoints 1 and 2 are also smaller in *del*<sup>129</sup> chromosome compared to *del*<sup>129</sup>, and display higher levels of variability [Fig. 4.7].

One of the possible explanations for the appearance of such an extended chromatin state after the occurrence of the deletion is that tethering points exist within these regions. These may cause the telomeric end and adjacent upstream CNV sequences to remain in their original preferred positions, therefore stretching the intervening sequence after deletion. Such tethering points may well be constituted by LADs. Introduced in Chapter 1, LADs are lamina-associated domains, important features of nuclear architecture and genomic regulation (Pickersgill *et al.*, 2006; Guelen *et al.*, 2008; Peric-Hupkes *et al.*, 2010). It could be possible that LAD regions exist bordering the CNV regions, therefore constituting chromatin tethering points. Analysis of published LAD positions (Wu and Yao, 2013) identified in 3T3 MEFs (Todaro and Green, 1963) revealed the presence of only LAD-free regions bordering the CNV (147-150Mb and 154.4-155.6Mb on chromosome 4) [Supp. Table 4.5], suggesting that LADs may not serve as structural tethering points. However, inside the CNV there exist 10 LADs within a 1.3Mb segment whose combined lengths add up to 1Mb. This represents 25% of the CNV sequence. It is therefore possible that one major tethering point exists within the CNV sequence itself, which after deletion affects the surrounding sequences and allows for chromatin to be looser, and therefore, less compact. Further investigation into these observations is discussed at the end of this chapter.

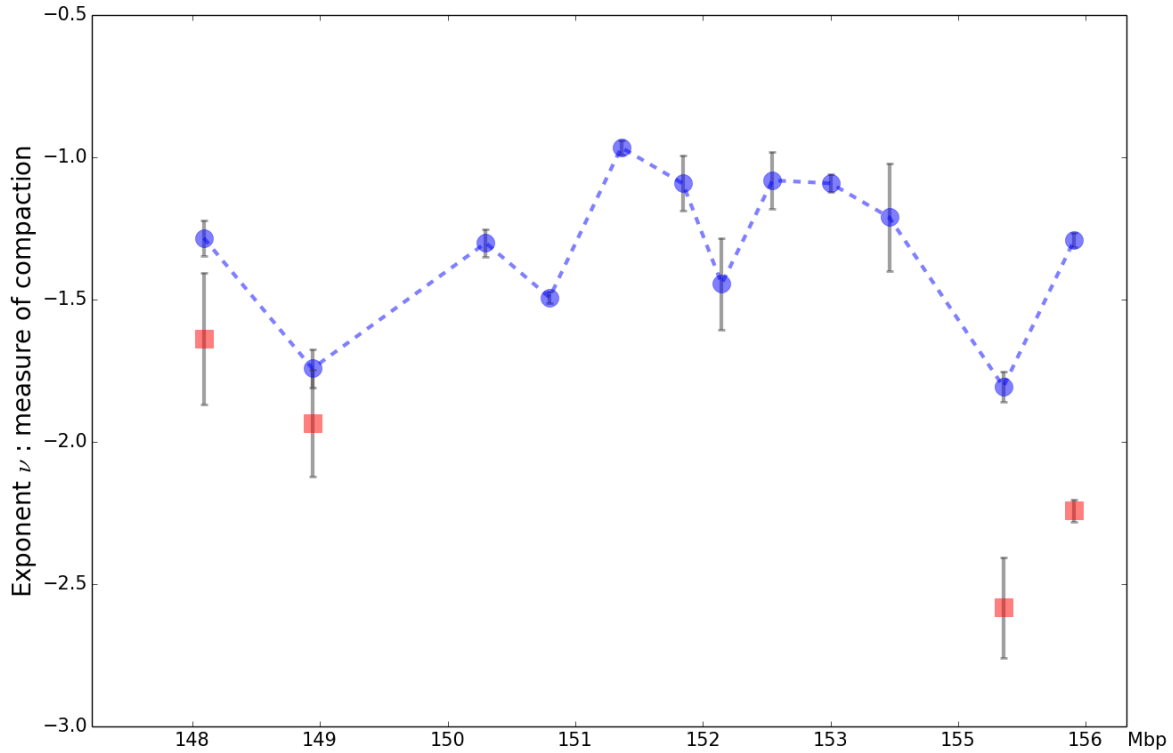
#### **4.5. Validation of changes in *del*<sup>129</sup> chromatin interactions by 3D DNA FISH**

After the identification of chromatin decompaction caused by the 4.3Mb deletion, and the dissection of differentially interacting regions at the Kb scale in the *del*<sup>129</sup> chromosome, I sought to validate PE-4Cseq results and ascertain the existence of regions displaying changes

in chromatin compaction as well as chromatin interactions.

I designed five 3D DNA FISH experiments, all of which have two probes ('query probes') on one or either side of the deletion region and one within it ('deletion probe'), the same strategy described in Chapter 3 for the analysis of chromatin organization changes in regions upstream of the CNV. The deletion probe is identical for all experiments and is used to distinguish the deletion chromosome, whereas the query probes are either upstream or downstream of the deletion [Table 4.2]. Four of the FISH experiments test changes in contact probabilities and compaction observed in PE-4Cseq data (BACsets 1-4). The fifth experiment (BACset 5) tests a specific interacting pair we observe in PE-4Cseq data, with a known enrichment in CTCF and Smc1 binding sites for one of the interacting regions [Supp. Table 4.6A,B]. Each experiment consisted in the acquisition of query probe pair distances for the  $df/+^{Bl6}$  and  $+^{129}/+^{Bl6}$  chromosomes in over a hundred cells. Images were processed using `Correct_and_Measure_3D.class`, the ImageJ plugin introduced in Chapter 3.

Similar to the strategy used for the analysis of PE-4Cseq data, the Gaussian polymer model dictates a distribution of query probe distances parameterized by the spring constant  $k$  for the effective spring connecting them. First of all, we observe that the model closely approximates the measured distribution of distances for all probes, demonstrating the validity of using this approach for quantitative data comparisons, even when using different units (i.e. contact frequencies measured in PE-4Cseq versus separation distances measured in 3D DNA FISH). Secondly, we performed a least-square fit to obtain the query probe pair's  $k$  values for comparison between  $df/+^{Bl6}$  and  $+^{129}/+^{Bl6}$  MEFs.



**Figure 4.7** Calculated  $\nu$  per viewpoint for  $del^{I29}$  vs. the average of  $wt^{Bl6}$ ,  $wt^{I29}$  and  $del^{Bl6}$

$del^{I29}$  (red squares).  $wt^{Bl6}$ ,  $wt^{I29}$  and  $del^{Bl6}$  (blue circles). Error bars determined from the two available biological replicates. Values of  $\nu < -3/2$  corresponds to less compact states than expected from Gaussian behavior, whereas  $\nu > -3/2$  corresponds to more compact states. Notice the *decrease* in overall compaction for  $del^{I29}$  and most significantly at its telomeric end. Such a behavior would be expected from tethering induced stretching of the chromosome regions flanking the deletion, or after the deletion of a major tethering point inside the CNV.

Experiment	Query probe 1	Start	End	Query probe 2	Start	End	Deletion probe	Start	End
BACset1	RP24-63D19	154,825,506	155,033,518	RP24-206E18	155,204,543	155,462,367	RP24-448A23	151,566,375	151,757,716
BACset2	RP24-325H12	155,378,045	155,565,687	RP23-159J19	154,571,627	154,757,237	RP24-448A23	151,566,375	151,757,716
BACset3	RP24-391E9	148,848,345	149,048,056	RP24-395H15	149,893,457	150,052,922	RP24-448A23	151,566,375	151,757,716
BACset4	RP24-391E9	148,848,345	149,048,056	RP24-63D19	154,825,506	155,033,518	RP24-448A23	151,566,375	151,757,716
BACset5	RP24-63D19	154,825,506	155,033,518	RP23-298E4	154,415,469	154,671,718	RP24-448A23	151,566,375	151,757,716

**Table 4.2 BACS used for selected PE-4Cseq and chromatin decompaction regions**

Location of BACs used in 3D DNA FISH experiments for the validation of *del*<sup>I29</sup> PE-4Cseq differentially interacting regions and chromatin decompaction

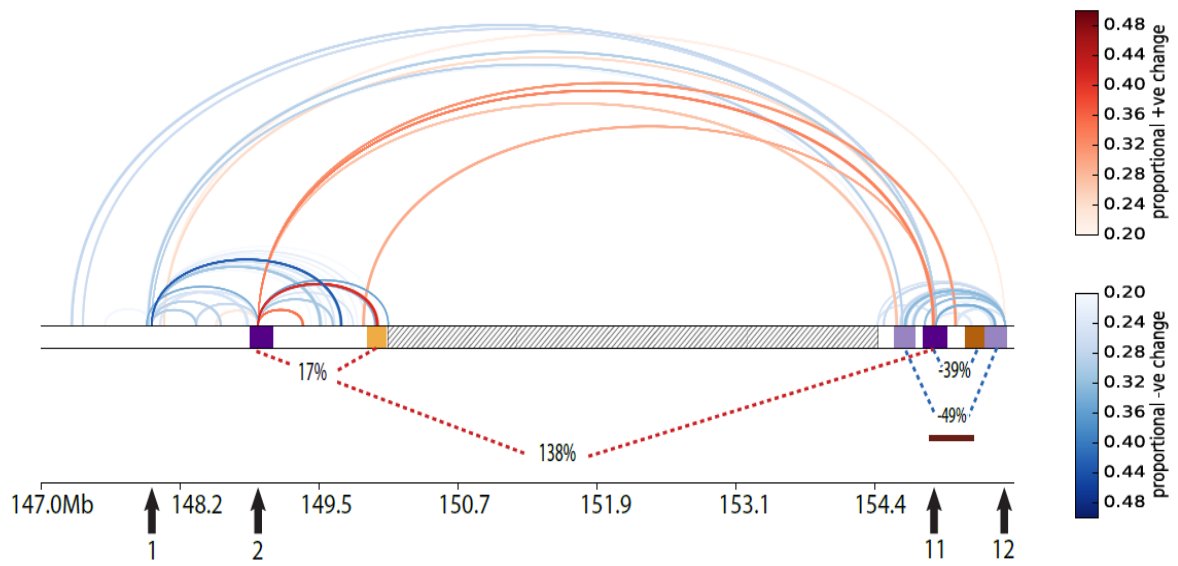
In Fig. 4.8A, we zoom into the deletion region, and show the ‘rainbow plot’ for changes in interaction observed in PE-4Cseq, and the corresponding changes in  $k$  values derived from FISH experiments. As can be seen in this figure, there is good agreement in both the direction and scale of the chromatin interaction changes, therefore validating our PE-4Cseq analysis framework and demonstrating for the first time a significant correlation between the magnitude of interaction changes detected in PE-4Cseq and 3D DNA FISH data.

Our control FISH experiment displayed a much narrower distribution of distances between query probes in both *del*<sup>129</sup> and *wt*<sup>129</sup> chromosomes, compared to the expectation for similar genomic separation, indicative of a more pronounced interaction [Fig. 4.8C,D]. This is in agreement with the peak in contact probability observed in PE-4Cseq data [Fig. 4.8B], validating the analysis pipeline’s prediction. Interestingly, the enrichment of CTCF and cohesin subunit Smc1 binding compared to the other BAC probe regions used, suggests that this stable contact may be mediated by CTCF and cohesin protein binding.

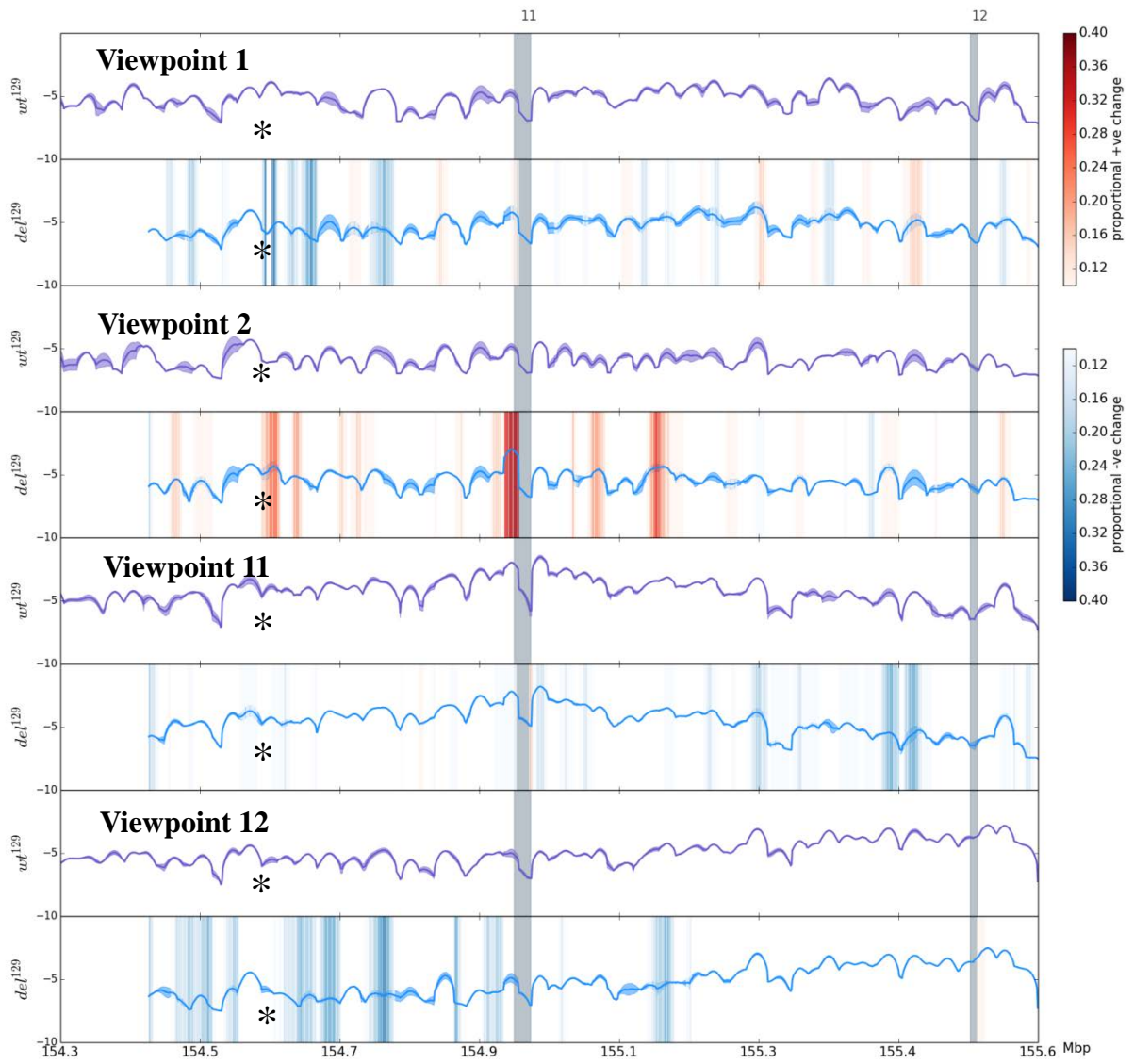
#### **4.6. Protein binding sites inside PE-4Cseq differentially contacting regions**

Diverse 3C and Hi-C studies have implicated proteins such as CTCF, cohesin, and Mediator as structural determinants of the three-dimensional organization of the mammalian genome. CTCF and cohesin have been shown to be boundary proteins between TADs, while Mediator plays important roles in loop formation for correct gene activation (Kagey *et al.*, 2010; Nora *et al.*, 2012; Phillips-Cremins *et al.*, 2013; Seitan *et al.*, 2013; Sofueva *et al.*, 2013; Zuin *et al.*, 2013).

A)

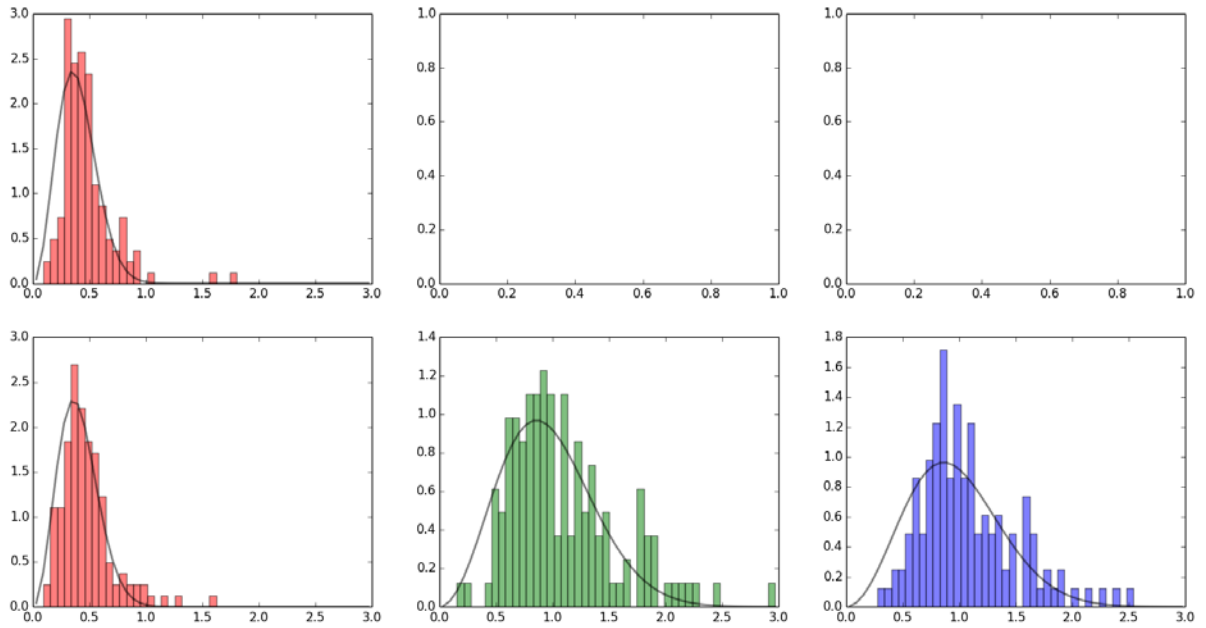


B)

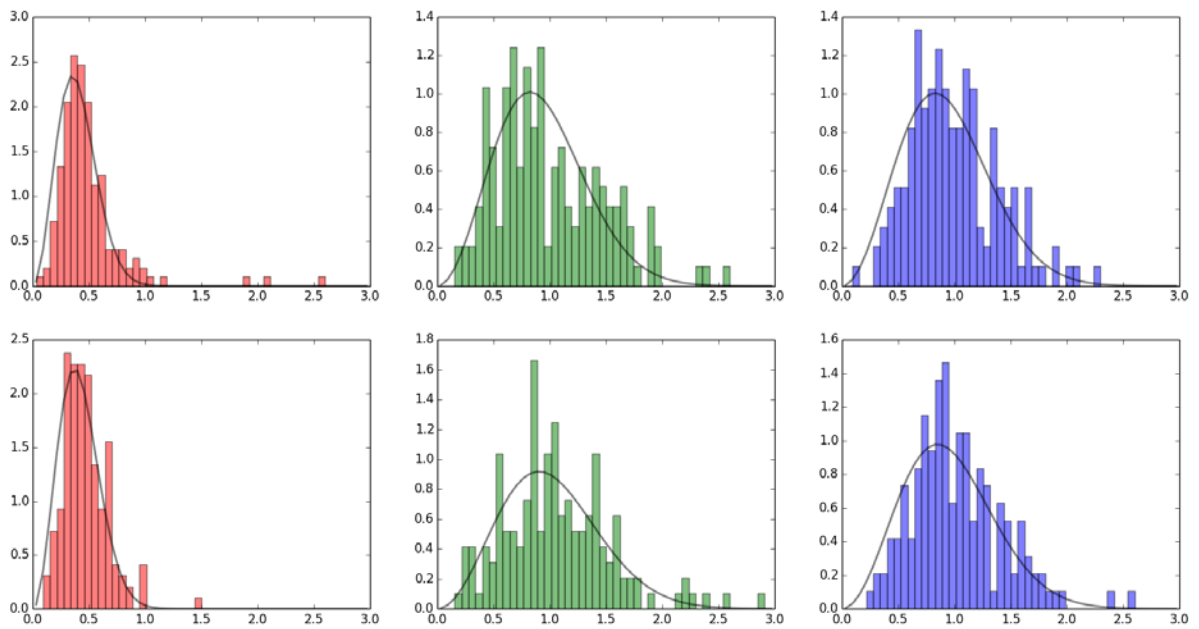




**C)**



**D)**



### Figure 4.8 *del*<sup>129</sup> 3D DNA FISH validations

**A)** Rainbow plot for the 147-155.6Mb region on *del*<sup>129</sup> chromosome. Shown in red links are regions whose contact probabilities in *del*<sup>129</sup> chromosome are at least 10% higher compared to *wt*<sup>129</sup>. Shown in blue links are regions whose contact probabilities in *del*<sup>129</sup> are at least 10% smaller compared to *wt*<sup>129</sup>. Viewpoint positions are shown by grey arrows. The dashed box of the chromosome bar is the deletion region; the four distinct colored boxes are the four query probe pairs. The dashed triangles represent query probe interaction changes as detected from 3D DNA FISH experiments. Enrichment is marked in red and depletion in blue colors, with associated proportional changes in interaction reported as percentages (positive for increases in interaction and negative for decreases in interaction). Notice the agreement between the red and blue links and dashed triangles in both PE-4Cseq and FISH interactions.

**B)** PE-4Cseq interaction profile for the highly interacting pair composed of viewpoint 11 and region 154.4-154.6Mb (marked with asterisks under each profile). This interaction was assessed by BACset2, with their corresponding distance distribution shown in **C)**. Notice the narrower distribution of distances between query probes (red histogram) in both *del*<sup>129</sup> chromosomes and **D)** *wt*<sup>129</sup> compared to distances measured between deletion probe and query probes (green and blue histograms), indicative of a highly frequent interaction between these two regions.

In order to further characterize the differential interactions of *del<sup>129</sup>* chromosomes, I queried the CTCF/Mediator/cohesin binding sites falling into these regions. I used a previously published CTCF/Mediator/cohesin binding dataset derived from MEFs, as described in Kagey *et al.*, 2010. Overlapping regions between the binding sites of these proteins in chromosome 4 and the detected differentially interacting regions in *del<sup>129</sup>* are summarized in Table 4.3. As can be seen from this table, there are a significant number of *del<sup>129</sup>* differentially interacting regions containing structural protein binding sites. Of interest, CTCF and cohesin subunit Smc1 have the highest total numbers of regions covered, between 30-60%. On the contrary, Mediator subunits 1 and 12 overlap with differentially interacting regions is 10% or smaller for all viewpoints analyzed.

To assess whether the CTCF/Smc1 overlap ratio was significant for the differentially interacting regions, I computed the probability of exceeding the number of protein binding sites in these regions against randomly chosen sequences of the same size as the differentially interacting regions in *del<sup>129</sup>*. I performed this task using a Monte Carlo simulation with 1,000 repeats (see Chapter 7 & 8 for details). As can be seen in Table 4.4, associated p-values for CTCF, Med1, and Smc1 binding to *del<sup>129</sup>* differentially interacting regions is <0.001, making these results highly significant. We can therefore say that there is an enrichment in CTCF, Med1, and Smc1 binding to *del<sup>129</sup>* differentially interacting regions in chromosome 4.

These observations suggest that the alterations observed in the contact probability of these regions could be mediated by these structural proteins, whose binding may be affected by differentially expressed transcription factors or chromatin remodeling genes in *df/+<sup>Bl6</sup>* MEFs (see discussion at the end of this chapter).

<b>Region</b>	<b>diff sites</b>	<b>dif site bp</b>	<b>no. sites CTCF</b>	<b>%</b>	<b>bp sites CTCF</b>	<b>%</b>	<b>no. CTCF</b>	<b>%</b>
chr4_dfwt_1	291	12,691,291	82	28	4,619,082	36	124	11
chr4_dfwt_2	183	8,303,183	74	40	4,566,074	55	114	10
chr4_dfwt_11	318	17,195,318	108	34	8,160,108	47	175	16
chr4_dfwt_12	163	6,316,163	50	31	2,842,050	45	78	7
chr4_dfwt_all	608	34,976,608	202	33	17,337,202	50	361	33
Total CTCF sites	1091							
<b>Region</b>	<b>diff sites</b>	<b>dif site bp</b>	<b>no. sites Med1</b>	<b>%</b>	<b>bp sites Med1</b>	<b>%</b>	<b>no. Med1</b>	<b>%</b>
chr4_dfwt_1	291	12,691,291	19	7	1,121,019	9	31	9
chr4_dfwt_2	183	8,303,183	17	9	1,154,017	14	26	8
chr4_dfwt_11	318	17,195,318	31	10	2,323,031	14	56	17
chr4_dfwt_12	163	6,316,163	14	9	620,014	10	21	6
chr4_dfwt_all	608	34,976,608	60	10	5,240,060	15	107	32
Total Med1 sites	332							
<b>Region</b>	<b>diff sites</b>	<b>dif site bp</b>	<b>no. sites Med12</b>	<b>%</b>	<b>bp sites Med12</b>	<b>%</b>	<b>no. Med12</b>	<b>%</b>
chr4_dfwt_1	291	12,691,291	11	4	609,011	5	15	9
chr4_dfwt_2	183	8,303,183	7	4	634,007	8	9	5
chr4_dfwt_11	318	17,195,318	14	4	864,014	5	19	11
chr4_dfwt_12	163	6,316,163	7	4	338,007	5	8	5
chr4_dfwt_all	608	34,976,608	33	5	2,492,033	7	44	26
Total Med12 sites	171							
<b>Region</b>	<b>diff sites</b>	<b>dif site bp</b>	<b>no. sites Smc1</b>	<b>%</b>	<b>bp sites Smc1</b>	<b>%</b>	<b>no. Smc1</b>	<b>%</b>
chr4_dfwt_1	291	12,691,291	116	40	6,611,116	52	239	12
chr4_dfwt_2	183	8,303,183	103	56	6,178,103	74	225	11
chr4_dfwt_11	318	17,195,318	156	49	11,037,156	64	357	17
chr4_dfwt_12	163	6,316,163	64	39	3,435,064	54	146	7
chr4_dfwt_all	608	34,976,608	277	46	22,419,277	64	722	35
Total Smc1 sites	2076							

**Table 4.3 Summary of *del*<sup>129</sup> differentially interacting regions overlap with CTCF, Mediator, and cohesin binding sites.**

Column 1, *Region*, refers to the viewpoint assessed. Column 2, *diffsites*, refers to *del*<sup>129</sup> differentially interacting regions. Column 4, *no. sites* and feature name corresponds to the number of differentially interacting regions that contain the specified genomic feature. Column 6, *bp sites* feature, presents the sum of differentially interacting regions bp which contain the specified feature. Column 8, *no. features*, indicates the number of features included inside the differentially interacting regions. Percentages in columns 5, 7, and 9 are calculated based on the total number of regions or features in the preceding column.

Region	diff sites	dif site bp	no. CTCF	%	MC simulation	p-val
chr4_dfwt_all	608	34,976,608	361	33	0	0
Total CTCF sites	1091					

Region	diff sites	dif site bp	no. Med1	%	MC simulation	p-val
chr4_dfwt_all	608	34,976,608	107	32	20	0.02
Total Med1 sites	332					

Region	diff sites	dif site bp	no. Med12	%	MC simulation	p-val
chr4_dfwt_all	608	34,976,608	44	26	312	0.312
Total Med12 sites	171					

Region	diff sites	dif site bp	no. Smc1	%	MC simulation	p-val
chr4_dfwt_all	608	34,976,608	722	35	0	0
Total Smc1 sites	2076					

**Table 4.4 Summary of Monte Carlo simulations for assessing statistical significance of protein binding overlaps for *del*<sup>129</sup> differentially interacting regions**

Column 6, *MC simulation*, summarizes the number of Monte Carlo simulations in which the number of overlapped protein binding sites with randomly generated regions exceeded that of the observed values (column 4). *p-val* expressed the probability of having such overlaps occurring just by chance. Notice the significant p-values obtained for CTCF, Med1, and Smc1 binding (p-val < 0.001, rounded down to zero in table).

## 4.7. PE-4Cseq results for the *del<sup>Bl6</sup>* chromosome

When comparing the contact probabilities of *df/+<sup>Bl6</sup>* and *+<sup>129</sup>/+<sup>Bl6</sup>* genotypes, each chromosome was compared against its WT homologue. Therefore, *del<sup>129</sup>* comparisons were performed against *wt<sup>129</sup>* chromosomes, and results discussed in the previous sections. We also performed the reciprocal *del<sup>Bl6</sup>-wt<sup>Bl6</sup>* comparison in order to assess the level of interaction changes of the WT chromosome 4 in *df/+<sup>Bl6</sup>* MEFs. The idea behind this comparison lies in the fact that numerous gene expression changes occur in *df/+<sup>Bl6</sup>* compared to *+<sup>129</sup>/+<sup>Bl6</sup>* MEFs, and that transcriptional differences are highly correlated for the C57Bl6/J and 129S5/SvEv<sup>Brd</sup> alleles (Chapter 5). Transcription factors as well as other proteins involved in chromatin condensation are affected, potentially having an impact in chromatin interactions in both C57Bl6/J and 129S5/SvEv<sup>Brd</sup> chromosomes. Assessing the degree of contact probability changes in *del<sup>Bl6</sup>* and comparing it to *del<sup>129</sup>* may give an insight into the fraction of differential interactions which could be potentially attributed to changes in gene expression and those which may be exclusive to the *cis* positional effects of the deletion CNV.

### 4.7.1. Contact probability changes for viewpoints surrounding the deletion coordinates

Analysis of viewpoints 1, 2, 11, and 12 revealed a total of 594 regions as differentially interacting between *del<sup>Bl6</sup>* and *wt<sup>Bl6</sup>* chromosomes, with a minimal size of 20Kb and at least 10% contact probability difference with respect to WT. The regions add up to

~27.5Mb of sequence, approximately 17.7% of chromosome 4 sequence, and 4.3% smaller than *del<sup>l29</sup>* regions. The changes observed for these viewpoints mostly involve slight reductions in contact probabilities (>60% of total detected regions) [Table 4.5. See Fig. 4.9A for a rainbow plot of the most terminal part of chromosome 4]. However, slight increases in interaction probabilities are also observed [Fig. 4.9B]. No obvious changes exist in terms of chromatin compaction [Fig. 4.10], suggesting no major impact on higher-order structure for this chromosome.

Similar to *del<sup>l29</sup>*, 25-50% of *del<sup>Bl6</sup>* differentially interacting regions overlap CTCF and cohesin binding sites, with Mediator overlapping <11% of these regions [Table 4.6]. I detected a statistically significant enrichment of CTCF and Smc1 only for *del<sup>Bl6</sup>* differentially interacting regions (p-val < 0.001) [Table 4.7].

There are 352 *del<sup>l29</sup>* differentially interacting regions that intersect *del<sup>Bl6</sup>*-derived ones. They have a mean size of ~34Kb, and their sequences add up to ~12Mb (~7.7% chromosome 4). After excluding *del<sup>Bl6</sup>*-derived segments from the dataset, there are 659 unique *del<sup>l29</sup>* differentially interacting regions with a mean size of 35Kb, covering ~23Mb (~15% chromosome 4). Accordingly, there exist 521 differentially interacting regions that are unique to *del<sup>Bl6</sup>*, with a mean size of 30Kb and covering ~15Mb of sequence (~10% chromosome 4). These regions are summarized in Table 4.8, as well as their CTCF, cohesin, and Mediator overlaps. Interestingly, CTCF and cohesin overlap percentages are still high (20-40%) compared to Mediator (<7%), and overlap enrichment is significant for CTCF and Smc1 only [Table 4.9]. Figure 4.11 summarizes positions for these regions and protein binding overlaps is for the 147-155.6Mb segment of chromosome 4. The whole chromosome view is shown in Supp. Fig. 4.3.

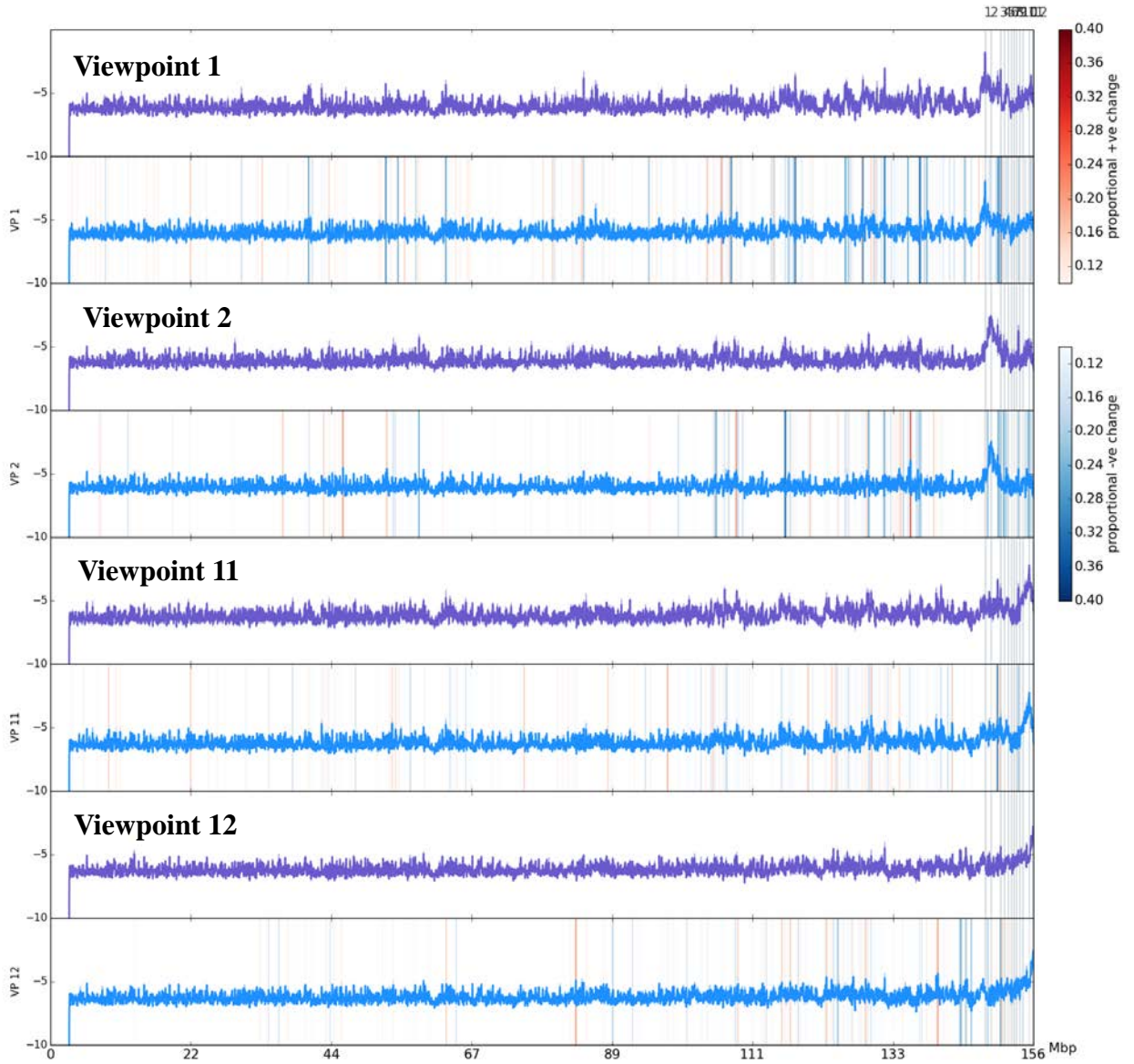


Region	Total regs	Median	No. increase	Median	% of tot regs	No. decrease	Median	% of tot regs
Df-WT-C57DiffRegs_chr4_vp_1	287	0.12	111	0.11	39	176	0.12	61
Df-WT-C57DiffRegs_chr4_vp_2	203	0.12	64	0.12	32	139	0.12	68
Df-WT-C57DiffRegs_chr4_vp_3	460	0.12	243	0.12	53	217	0.12	47
Df-WT-C57DiffRegs_chr4_vp_4	347	0.13	174	0.14	50	168	0.12	48
Df-WT-C57DiffRegs_chr4_vp_5	348	0.12	220	0.12	63	128	0.12	37
Df-WT-C57DiffRegs_chr4_vp_6	653	0.13	139	0.12	21	514	0.13	79
Df-WT-C57DiffRegs_chr4_vp_7	636	0.14	164	0.17	26	471	0.14	74
Df-WT-C57DiffRegs_chr4_vp_8	318	0.13	121	0.13	38	196	0.13	62
Df-WT-C57DiffRegs_chr4_vp_9	511	0.12	239	0.12	47	272	0.13	53
Df-WT-C57DiffRegs_chr4_vp_10	167	0.12	51	0.12	31	115	0.12	69
Df-WT-C57DiffRegs_chr4_vp_11	199	0.12	80	0.11	40	119	0.12	60
Df-WT-C57DiffRegs_chr4_vp_12	160	0.11	63	0.11	39	96	0.12	60

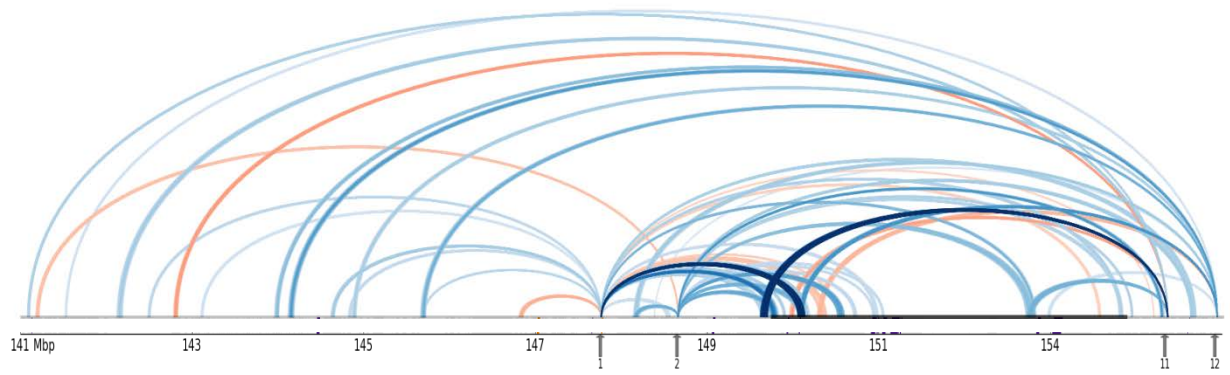
**Table 4.5 Summary of median magnitude of change, direction, and number of *del<sup>Bl6</sup>* differentially interacting regions for viewpoints 1-12**

Notice how only viewpoint 5 displays an increase in contact probabilities (63% of total regions). Viewpoints 3, 4, and 9 show a ~50% split between regions with an increase and decrease in contact probabilities, while the rest of the viewpoints show a variable range in the decrease in interactions (60-79% of total regions per viewpoint).

A)

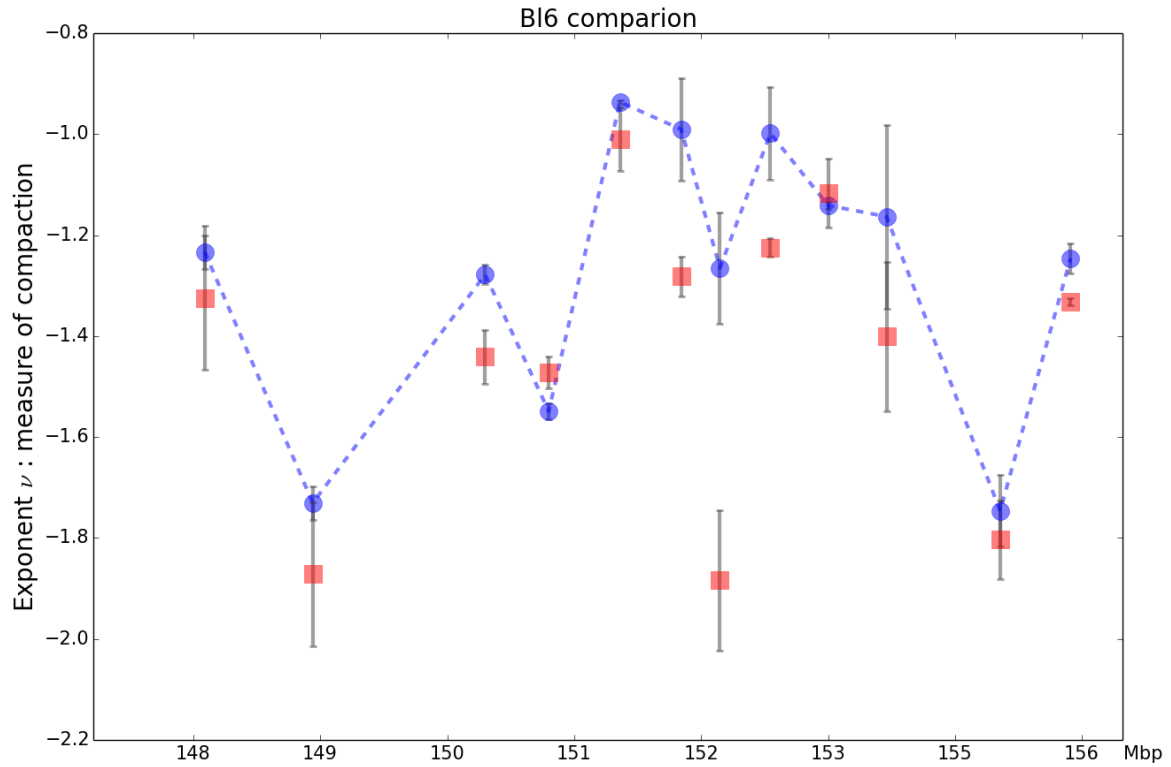


**B)**



**Figure 4.9 Contact probability profiles for the  $del^{Bl6}$  and  $wt^{Bl6}$  for chromosome 4**

**A)** Comparison of contact probability profiles for the  $del^{Bl6}$  and  $wt^{Bl6}$  for chromosome 4 sequence. Each horizontal panel corresponds to the contact probability profiles per chromosome (blue for  $del^{Bl6}$  and purple for  $wt^{Bl6}$ ) derived from PE-4Cseq data paired for viewpoints 1, 2, 11, 12, bordering the deletion. Two biological replicates are used to assess the error profile shown as a band around the contact probability histograms. Shown in red and blue are regions whose contact probabilities in  $del^{Bl6}$  chromosome are at least 10% higher/smaller compared to  $wt^{Bl6}$ , respectively. Note how the majority of the changes observed in  $del^{Bl6}$  concentrate surrounding the deletion, up to 40Mb upstream, similar to  $del^{129}$ . **B)** Rainbow plot for the 141-155.6Mb region on  $del^{Bl6}$  chromosome. Red/blue links are regions whose contact probabilities in  $del^{Bl6}$  chromosome are at least 10% higher/smaller compared to  $wt^{Bl6}$ , respectively. Viewpoints positions are shown by grey arrows. The dark line of the chromosome bar is the deletion region. Contrary to  $del^{129}$ , the majority of the changes in contact probabilities are decreases in interaction, especially upstream of the deletion start, on viewpoints 1 and 2.



**Figure 4.10** Calculated  $\nu$  per viewpoint for  $del^{Bl6}$  vs.  $wt^{Bl6}$

$del^{Bl6}$  (red squares).  $wt^{Bl6}$  (blue circles). Error bars determined from the two available biological replicates. Notice there are no major differences between compaction values for both chromosomes, except for viewpoints 152.1, 152.4, and 152.9, where one of the  $+^{129}/+^{Bl6}$  biological replicates had fewer reads compared to  $df/+^{Bl6}$ .

Region	diff sites	dif site bp	no. sites CTCF	%	bp sites CTCF	%	no. CTCF	%
chr4_dfw_t_bl6_1	281	12,014,281	89	32	5,732,089	48	148	14
chr4_dfw_t_bl6_2	200	8,052,200	72	36	3,831,072	48	103	9
chr4_dfw_t_bl6_11	196	7,923,196	49	25	2,797,049	35	73	7
chr4_dfw_t_bl6_12	160	5,522,160	44	28	2,332,044	42	64	6
chr4_dfw_t_bl6_all	594	27,579,594	189	32	13,147,189	48	312	29
Total CTCF sites	1091							
Region	diff sites	dif site bp	no. sites Med1	%	bp sites Med1	%	no. Med1	%
chr4_dfw_t_bl6_1	281	12,014,281	30	11	1,898,030	16	42	13
chr4_dfw_t_bl6_2	200	8,052,200	10	5	521,010	6	13	4
chr4_dfw_t_bl6_11	196	7,923,196	12	6	814,012	10	15	5
chr4_dfw_t_bl6_12	160	5,522,160	4	3	191,004	3	4	1
chr4_dfw_t_bl6_all	594	27,579,594	45	8	3526045	13	63	19
Total Med1 sites	332							
Region	diff sites	dif site bp	no. sites Med12	%	bp sites Med12	%	no. Med12	%
chr4_dfw_t_bl6_1	281	12,014,281	12	4	718,012	6	14	8
chr4_dfw_t_bl6_2	200	8,052,200	2	1	52,002	1	2	1
chr4_dfw_t_bl6_11	196	7,923,196	1	1	81,001	1	1	1
chr4_dfw_t_bl6_12	160	5,522,160	4	3	159,004	3	4	2
chr4_dfw_t_bl6_all	594	27,579,594	19	3	1,284,019	5	21	12
Total Med12 sites	171							
Region	diff sites	dif site bp	no. sites Smc1	%	bp sites Smc1	%	no. Smc1	%
chr4_dfw_t_bl6_1	281	12,014,281	128	46	7,761,128	65	279	13
chr4_dfw_t_bl6_2	200	8,052,200	91	46	4,671,091	58	187	9
chr4_dfw_t_bl6_11	196	7,923,196	72	37	4,118,072	52	135	7
chr4_dfw_t_bl6_12	160	5,522,160	47	29	2,339,047	42	81	4
chr4_dfw_t_bl6_all	594	27,579,594	247	42	16,497,247	60	556	27
Total Smc1 sites	2076							

**Table 4.6 Summary of *del<sup>Bl6</sup>* differentially interacting regions overlap for viewpoints 1, 2, 11, and 12 with CTCF, Mediator, and Smc1 binding sites**

Column identities are as described in Table 4.3.

Region	diff sites	dif site bp	no. CTCF	%	MC simulation	p-val
chr4_dfwt_bl6_all	594	27,579,594	312	29	0	0
Total CTCF sites	1091					

Region	diff sites	dif site bp	no. Med1	%	MC simulation	p-val
chr4_dfwt_bl6_all	594	27,579,594	63	19	376	0.376
Total Med1 sites	332					

Region	diff sites	dif site bp	no. Med12	%	MC simulation	p-val
chr4_dfwt_bl6_all	594	27,579,594	21	12	905	0.905
Total Med12 sites	171					

Region	diff sites	dif site bp	no. Smc1	%	MC simulation	p-val
chr4_dfwt_bl6_all	594	27,579,594	556	27	0	0
Total Smc1 sites	2076					

**Table 4.7 Summary of Monte Carlo simulations for assessing statistical significance of protein binding overlaps for *del<sup>Bl6</sup>* differentially interacting regions for viewpoints 1, 2, 11, and 12**

Column identities are as described in Table 4.4. Notice the significant p-values obtained for CTCF and Smc1 binding (p-val < 0.001, rounded down to zero in table).

Region	diff sites	dif site bp	no. sites CTCF	%	bp sites CTCF	%	no. CTCF	%
bl6_minus_df_intpiece_diffinterregs	521	15,654,521	119	23	5,334,119	34	167	15
bl6_thatintersect_df_intpiece_diffinterregs	352	11,925,352	101	29	5,239,101	44	145	13
df_minus_bl6_intpiece_diffinterregs	659	23,051,659	147	22	8,351,147	36	216	20
df_thatintersect_bl6_intpiece_diffinterregs	352	11,925,352	101	29	5,239,101	44	145	13
Total CTCF sites	1091							
Region	diff sites	dif site bp	no. sites Med1	%	bp sites Med1	%	no. Med1	%
bl6_minus_df_intpiece_diffinterregs	521	15,654,521	23	4	1,202,023	8	28	8
bl6_thatintersect_df_intpiece_diffinterregs	352	11,925,352	25	7	1,353,025	11	35	11
df_minus_bl6_intpiece_diffinterregs	659	23,051,659	44	7	2,287,044	10	72	22
df_thatintersect_bl6_intpiece_diffinterregs	352	11,925,352	25	7	1,353,025	11	35	11
Total Med1 sites	332							
Region	diff sites	dif site bp	no. sites Med12	%	bp sites Med12	%	no. Med12	%
bl6_minus_df_intpiece_diffinterregs	521	15,654,521	10	2	391,010	2	10	6
bl6_thatintersect_df_intpiece_diffinterregs	352	11,925,352	10	3	516,010	4	11	6
df_minus_bl6_intpiece_diffinterregs	659	23,051,659	23	3	1,130,023	5	33	19
df_thatintersect_bl6_intpiece_diffinterregs	352	11,925,352	10	3	516,010	4	11	6
Total Med12 sites	171							
Region	diff sites	dif site bp	no. sites Smc1	%	bp sites Smc1	%	no. Smc1	%
bl6_minus_df_intpiece_diffinterregs	521	15,654,521	160	31	7,150,160	46	282	14
bl6_thatintersect_df_intpiece_diffinterregs	352	11,925,352	143	41	7,069,143	59	275	13
df_minus_bl6_intpiece_diffinterregs	659	23,051,659	217	33	11,652,217	51	448	22
df_thatintersect_bl6_intpiece_diffinterregs	352	11,925,352	143	41	7,069,143	59	275	13
Total Smc1 sites	2076							

**Table 4.8 Summary of unique and overlapping  $del^{129}$  and  $del^{Bl6}$  differentially interacting regions for viewpoints 1, 2, 11, and 12**

Region names in column one describe differentially regions: bl6\_minus\_df = unique  $del^{Bl6}$ ; df\_minus\_bl6 = unique to  $del^{129}$ ; bl6\_thatintersect\_df/df\_thatintersect\_bl6 = shared between  $del^{129}$  and  $del^{Bl6}$ . Overlap with CTCF, Mediator, and Smc1 binding sites and their corresponding percentages and sequence coverage are also shown. Column identities are as described in Table 4.3.



Region	diff sites	dif site bp	no. CTCF	%	MC simulation	p-val
bl6_minus_df_intpiece_diffinterregs	521	15,654,521	167	15	0	0
df_minus_bl6_intpiece_diffinterregs	659	23,051,659	216	20	0	0
df_thatintersect_bl6_intpiece_diffinterregs	352	11,925,352	145	13	0	0
Total CTCF sites	1091					
Region	diff sites	dif site bp	no. Med1	%	MC simulation	p-val
bl6_minus_df_intpiece_diffinterregs	521	15,654,521	28	8	735	0.735
df_minus_bl6_intpiece_diffinterregs	659	23,051,659	72	22	23	0.023
df_thatintersect_bl6_intpiece_diffinterregs	352	11,925,352	35	11	139	0.139
Total Med1 sites	332					
Region	diff sites	dif site bp	no. Med12	%	MC simulation	p-val
bl6_minus_df_intpiece_diffinterregs	521	15,654,521	10	6	946	0.946
df_minus_bl6_intpiece_diffinterregs	659	23,051,659	33	19	176	0.176
df_thatintersect_bl6_intpiece_diffinterregs	352	11,925,352	11	6	660	0.66
Total Med12 sites	171					
Region	diff sites	dif site bp	no. Smc1	%	MC simulation	p-val
bl6_minus_df_intpiece_diffinterregs	521	15,654,521	282	14	0	0
df_minus_bl6_intpiece_diffinterregs	659	23,051,659	448	22	0	0
df_thatintersect_bl6_intpiece_diffinterregs	352	11,925,352	275	13	0	0
Total Smc1 sites	2076					

**Table 4.9 Summary of Monte Carlo simulations for assessing statistical significance of protein binding overlaps for unique and overlapping  $del^{129}$  and  $del^{Bl6}$  differentially interacting regions for viewpoints 1, 2, 11, and 12**

Column identities are as described in Table 4.4. Region names are as described in Table 4.8. *bl6\_thatintersect\_df* region was omitted from this table as results are equivalent to *df\_thatintersect\_bl6*. Notice the significant p-values obtained for CTCF and Smc1 binding (p-val < 0.001, rounded down to zero in table).



**Figure 4.11 Summary of  $del^{I29}$  and  $del^{Bl6}$ , as well as unique and overlapping  $del^{I29}$  and  $del^{Bl6}$  differentially interacting regions are shown for region 147-155.6Mb of mouse chromosome 4**

CTCF, Mediator, and cohesin protein binding sites are also shown, as well as positions of RefSeq genes.

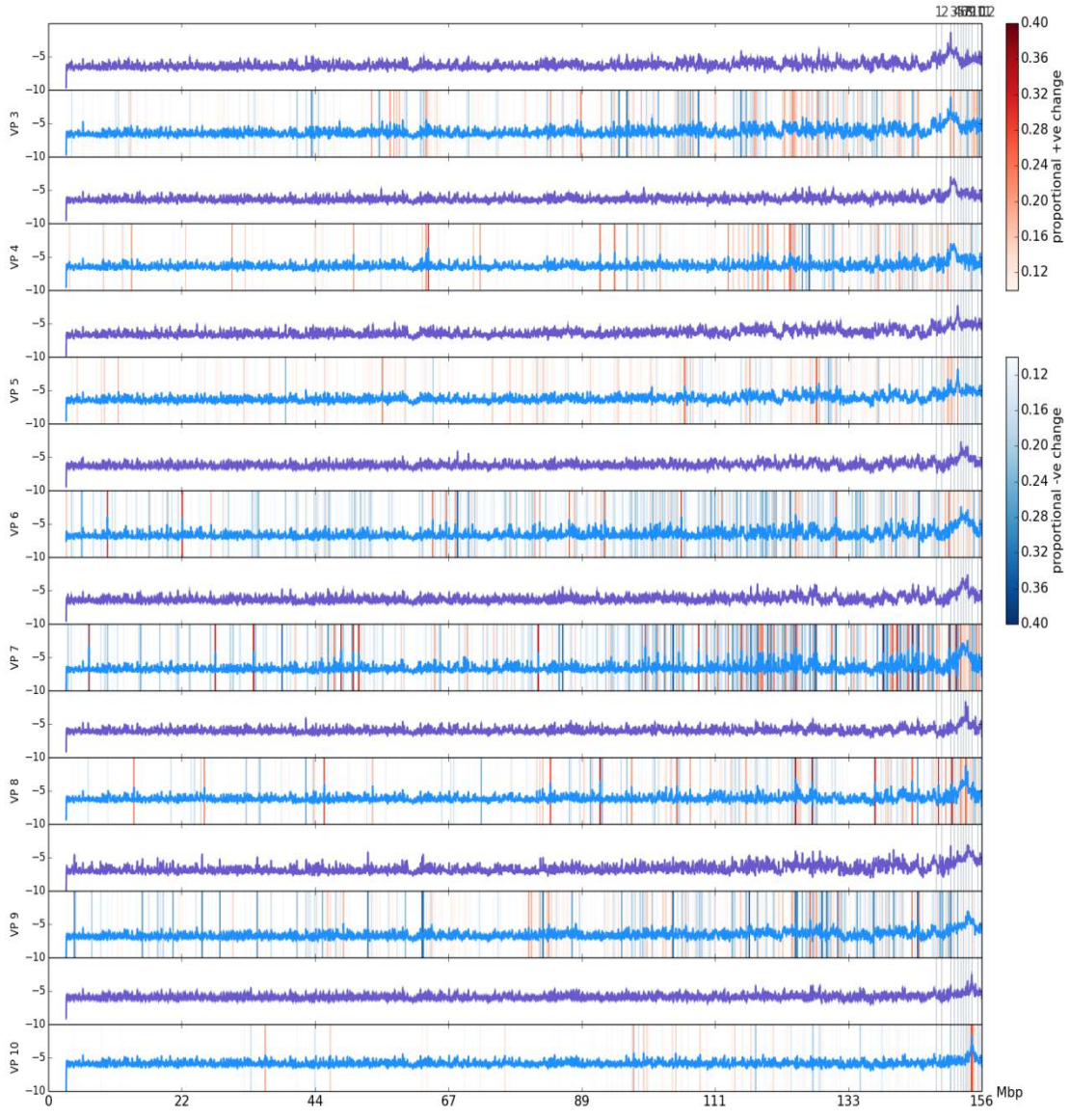
#### **4.7.2. Contact probability changes for viewpoints inside deletion CNV coordinates**

Viewpoints 3-10, inside the deletion CNV, can only be assessed in *del*<sup>Bl6</sup> chromosomes in *df/+*<sup>Bl6</sup> MEFs, and compared against contact probability profiles of in *wt*<sup>Bl6</sup> chromosomes of +<sup>129</sup>/<sub>+</sub><sup>Bl6</sup>. The decision to assess chromatin organization for this region was motivated for the study of transcriptional (“dosage”) compensation. Extensively studied examples of dosage compensation include X chromosome inactivation (reviewed in Schulz and Heard, 2013) and genomic imprinting (reviewed in Bartolomei and Ferguson-Smith, 2011; McAnally and Yampolsky, 2010). While these examples mostly constitute a decrease in gene expression of specific genes or chromosomes, transcriptional upregulation has also been observed. Previous studies have identified genes whose heterozygous KO mouse models show similar mRNA levels to their WT states (Wheway *et al.*, 2013; Homma *et al.*, 2006). Recently, a comprehensive characterization of monoallelic gene expression in ES and NPC cells showed transcriptional compensation for up to 8% of the identified monoallelic genes, with possible important biological consequences (Eckersley-Maslin *et al.*, 2014). Several mechanisms could act in response to the loss or silencing of one gene copy to increase the transcriptional output of the active remaining allele. Transcription factor accessibility and/or specific network regulation feedback loop processes have been suggested (Eckersley-Maslin *et al.*, 2014 and references therein). Given the extensive networks of promoter-regulatory elements associations detected for selected regions in the human and mouse genomes (Fullwood *et al.*, 2009; Sanyal *et al.*, 2012; Li *et al.*, 2012; Kieffer-Kwon *et al.*, 2013), one could also hypothesize that the switching of promoters to more “active”

regulatory elements or chromatin conformations could be a component of the transcriptional compensation mechanism. The *df/+<sup>Bl6</sup>* genotype thus provides an excellent model for the study of chromatin organization and its potential influence on transcriptional compensation, should this happen for the C57Bl6/J alleles in the *del<sup>Bl6</sup>* chromosome of *df/+<sup>Bl6</sup>* MEFs (Chapter 5).

PE-4Cseq pipeline analyses revealed a total of 1,112 regions as differentially interacting between *del<sup>Bl6</sup>* and *wt<sup>Bl6</sup>* chromosomes [Fig. 4.12], with a minimal size of 20Kb and at least 10% contact probability difference with respect to WT. The regions add up to ~92Mb of sequence, approximately 60% of chromosome 4 sequence, and 3X the detected *del<sup>l29</sup>* regions coverage. Viewpoints 3, 4, and 9 have an almost equal number of regions having a decrease and increase in contact probabilities, viewpoint 5 shows a higher number of regions with an increase in chromatin contacts, while viewpoints 6, 7, 8, and 10 show mostly decreases in contact probabilities [Table 4.5]. The differentially interacting regions are scattered along chromosome 4, however most of the regions concentrate up to 40Mb upstream of the CNV start, similar to *del<sup>l29</sup>* regions [Fig. 4.11]. As shown in Figure 4.10, no obvious changes exist in terms of chromatin compaction.

20-40% of *del<sup>Bl6</sup>* differentially interacting regions inside the CNV overlap CTCF and cohesin binding sites, with Mediator overlapping <10% of these regions [Table 4.10]. I detected a statistically significant enrichment of CTCF with these differentially interacting regions (p-val < 0.001) [Table 4.11]. Smc1 binding enrichment was significant for only viewpoints 3, 4, 5, 9, and 10, while Med1 and Med12 binding was not shown to be enriched for any of the viewpoints [Table 4.11].



**Figure 4.12 Comparison of contact probability profiles for the  $del^{Bl6}$  and  $wt^{Bl6}$  for chromosome 4 sequence**

Each horizontal panel corresponds to the contact probability profiles per chromosome (blue for  $del^{Bl6}$  and purple for  $wt^{Bl6}$ ) derived from PE-4Cseq data paired for viewpoints 3-10, inside the deletion coordinates. Two biological replicates are used to assess the error profile shown as a band around the contact probability histograms. Shown in red and blue are regions whose contact probabilities in  $del^{Bl6}$  chromosome are at least 10% higher/smaller compared to  $wt^{Bl6}$ , respectively. Notice how the majority of the changes observed in  $del^{Bl6}$  concentrate surrounding the deletion, up to 40Mb upstream, similar to  $del^{129}$ .

Region	diff sites	dif site bp	no. sites CTCF	%	bp sites CTCF	%	no. CTCF	%
chr4_dfwtBl6_3	455	19,050,455	132	29	8,192,132	43	209	19
chr4_dfwtBl6_4	337	12,330,337	89	26	5,227,089	42	129	12
chr4_dfwtBl6_5	344	13,336,344	82	24	4,492,082	34	134	12
chr4_dfwtBl6_6	643	36,054,643	176	27	14,212,176	39	318	29
chr4_dfwtBl6_7	625	36,953,625	177	28	15,100,177	41	310	28
chr4_dfwtBl6_8	314	16,229,314	73	23	5,405,073	33	135	12
chr4_dfwtBl6_9	505	22,735,505	139	28	8,920,139	39	236	22
chr4_dfwtBl6_10	161	5,654,161	36	22	2,432,036	43	55	5
chr4_dfwtBl6_all	1112	91,905,112	306	28	47,443,306	52	774	71
Total CTCF sites	1091							

Region	diff sites	dif site bp	no. sites Med1	%	bp sites Med1	%	no. Med1	%
chr4_dfwtBl6_3	455	19,050,455	30	7	2,156,030	11	48	14
chr4_dfwtBl6_4	337	12,330,337	14	4	875,014	7	26	8
chr4_dfwtBl6_5	344	13,336,344	25	7	1,391,025	10	30	9
chr4_dfwtBl6_6	643	36,054,643	48	7	3,723,048	10	77	23
chr4_dfwtBl6_7	625	36,953,625	48	8	4,164,048	11	91	27
chr4_dfwtBl6_8	314	16,229,314	17	5	1,252,017	8	28	8
chr4_dfwtBl6_9	505	22,735,505	41	8	2,341,041	10	67	20
chr4_dfwtBl6_10	161	5,654,161	8	5	352,008	6	9	3
chr4_dfwtBl6_all	1112	91,905,112	108	10	18,419,108	20	208	63
Total Med1 sites	332							

Region	diff sites	dif site bp	no. sites Med12	%	bp sites Med12	%	no. Med12	%
chr4_dfwtBl6_3	455	19,050,455	15	3	1,188,015	6	22	13
chr4_dfwtBl6_4	337	12,330,337	6	2	358,006	3	9	5
chr4_dfwtBl6_5	344	13,336,344	14	4	847,014	6	17	10
chr4_dfwtBl6_6	643	36,054,643	27	4	2,239,027	6	40	23
chr4_dfwtBl6_7	625	36,953,625	27	4	2,402,027	7	36	21
chr4_dfwtBl6_8	314	16,229,314	7	2	665,007	4	9	5
chr4_dfwtBl6_9	505	22,735,505	18	4	1,139,018	5	24	14
chr4_dfwtBl6_10	161	5,654,161	3	2	170,003	3	4	2
chr4_dfwtBl6_all	1112	91,905,112	63	6	11,017,063	12	93	54
Total Med12 sites	171							

Region	diff sites	dif site bp	no. sites Smc1	%	bp sites Smc1	%	no. Smc1	%
chr4_dfwtBl6_3	455	19,050,455	186	41	10,672,186	56	388	19
chr4_dfwtBl6_4	337	12,330,337	110	33	6,319,110	51	226	11
chr4_dfwtBl6_5	344	13,336,344	129	38	6,651,129	50	267	13
chr4_dfwtBl6_6	643	36,054,643	239	37	18,167,239	50	530	26
chr4_dfwtBl6_7	625	36,953,625	236	38	18,865,236	51	536	26
chr4_dfwtBl6_8	314	16,229,314	100	32	7,072,100	44	239	12
chr4_dfwtBl6_9	505	22,735,505	197	39	11,932,197	52	432	21
chr4_dfwtBl6_10	161	5,654,161	51	32	3,107,051	55	101	5
chr4_dfwtBl6_all	1112	91,905,112	414	37	57,018,414	62	1425	69
Total Smc1 sites	2076							



**Table 4.10 Summary of *del<sup>B16</sup>* differentially interacting regions overlap for viewpoints 3-10 with CTCF, Mediator, and Smc1 binding sites**

Column identities are as described in Table 4.3.

Region	diff sites	dif site bp	no. CTCF	MC simulations	p-val
chr4_dfwtBl6_3	455	19,050,455	209	0	0
chr4_dfwtBl6_4	337	12,330,337	129	1	0.001
chr4_dfwtBl6_5	344	13,336,344	134	3	0.003
chr4_dfwtBl6_6	643	36,054,643	318	1	0.001
chr4_dfwtBl6_7	625	36,953,625	310	7	0.007
chr4_dfwtBl6_8	314	16,229,314	135	49	0.049
chr4_dfwtBl6_9	505	22,735,505	236	0	0
chr4_dfwtBl6_10	161	5,654,161	55	39	0.039
chr4_dfwtBl6_all	1112	91,905,112	774	0	0
Total CTCF sites	1091				

Region	diff sites	dif site bp	no. Med1	MC simulations	p-val
chr4_dfwtBl6_3	455	19,050,455	48	253	0.253
chr4_dfwtBl6_4	337	12,330,337	26	526	0.526
chr4_dfwtBl6_5	344	13,336,344	30	401	0.401
chr4_dfwtBl6_6	643	36,054,643	77	476	0.476
chr4_dfwtBl6_7	625	36,953,625	91	210	0.21
chr4_dfwtBl6_8	314	16,229,314	28	758	0.758
chr4_dfwtBl6_9	505	22,735,505	67	52	0.052
chr4_dfwtBl6_10	161	5,654,161	9	702	0.702
chr4_dfwtBl6_all	1112	91,905,112	208	325	0.325
Total Med1 sites	332				

Region	diff sites	dif site bp	no. Med12	MC simulations	p-val
chr4_dfwtBl6_3	455	19,050,455	22	434	0.434
chr4_dfwtBl6_4	337	12,330,337	9	846	0.846
chr4_dfwtBl6_5	344	13,336,344	17	355	0.355
chr4_dfwtBl6_6	643	36,054,643	40	493	0.493
chr4_dfwtBl6_7	625	36,953,625	36	699	0.699
chr4_dfwtBl6_8	314	16,229,314	9	968	0.968
chr4_dfwtBl6_9	505	22,735,505	24	567	0.567
chr4_dfwtBl6_10	161	5,654,161	4	759	0.759
chr4_dfwtBl6_all	1112	91,905,112	93	718	0.718
Total Med12 sites	171				

Region	diff sites	dif site bp	no. Smc1	MC simulations	p-val
chr4_dfwtBl6_3	455	19,050,455	388	0	0
chr4_dfwtBl6_4	337	12,330,337	226	3	0.003
chr4_dfwtBl6_5	344	13,336,344	267	0	0
chr4_dfwtBl6_6	643	36,054,643	530	76	0.076
chr4_dfwtBl6_7	625	36,953,625	536	92	0.092
chr4_dfwtBl6_8	314	16,229,314	239	126	0.126
chr4_dfwtBl6_9	505	22,735,505	432	0	0
chr4_dfwtBl6_10	161	5,654,161	101	35	0.035
chr4_dfwtBl6_all	1112	91,905,112	1425	3	0.003
Total Smc1 sites	2076				

**Table 4.11 Summary of Monte Carlo simulations for assessing statistical significance of protein binding overlaps for *del<sup>Bl6</sup>* differentially interacting regions for viewpoints 3-10**

Column identities are as described in Table 4.4. Notice the significant p-values obtained for all CTCF overlaps for all viewpoints, and Smc1 binding to viewpoints 3, 4, 5, 9, and 10 (p-val < 0.001, rounded down to zero in table).

## 4.8 PE-4Cseq results summary and discussion

We developed a quantitative framework for the analysis of multi-viewpoint PE-4Cseq data. Our method corrects for biases typically found in 4Cseq experiments, including those intrinsic to the amplification and sequencing steps of the 4C protocol. Through the use of modeling based on polymer physics, we are able to extract the contact probability signal for viewpoints along the studied chromosome in *cis*, allowing for quantitative comparison between WT and *df* chromosomes.

The use of such new analysis methodology allowed us to detect measurable changes in chromatin interactions across the sequence of chromosome 4 upon the occurrence of a 4.3Mb deletion CNV in the 4E2 region in mouse. Up to 22% of chromosome 4 sequence displays changes in contact probabilities between the *del*<sup>129</sup> and *wt*<sup>129</sup> chromosomes. Several long-range interactions across the deletion region were augmented at levels higher than expected purely from their altered genomic proximity. I verified a select few of these changes through 3D DNA FISH experiments. Notably, a strong agreement between the change trends for both experimental modalities (PE-4Cseq and 3D DNA FISH) was found, giving us confidence to trust the results reported by the new PE-4Cseq analysis pipeline. To our knowledge, this is the first time a quantitative agreement between a C technique and 3D DNA FISH is reported.

Notably, the CNV caused an overall reduction in compaction in the *df* chromosome, especially at its telomeric end. Hypothetically, CNV-neighboring regions may harbor tethering points which could cause the intervening chromatin to extend upon the occurrence of the 4.3Mb deletion. No major LAD associations were found on regions surrounding the CNV, however, a major 1Mb segment encompassing numerous LADs is contained within the

CNV, potentially serving as a tethering point of the 4E2 band. Subsequent experiments using BAC probes inside this segment could be used to study whether associations with the nuclear periphery or other nuclear feature exist for this region, and other further upstream LAD sequences.

Very interestingly, *del*<sup>129</sup> differentially interacting regions are enriched for CTCF, Med1, and Smc1 protein binding, suggesting that regions whose chromatin interactions are altered could potentially be controlled by changes in these proteins transcription or upstream binding regulators. This hypothesis is further discussed in Chapter 5, RNA-seq analysis of *df/+<sup>Bl6</sup>* and *+<sup>129</sup>/+<sup>Bl6</sup>* MEFs. While performing CTCF, Med1, and Smc1 ChIP-seq experiments using *df/+<sup>Bl6</sup>* and *+<sup>129</sup>/+<sup>Bl6</sup>* MEFs would be the ideal experiment to determine changes in protein binding directly affecting the underlying chromatin architecture of 4E2, we are facing the challenge of growing enough number of cells required for the ChIP-seq protocol. This is due to the fact that the cells used in this study are primary cultures used at P4. At this passage number, the 3D DNA FISH, PE-4C-seq, and RNA-seq experiments, used most of the available material. Subsequent culture passages to expand the population were carried out. However, after P8 both *df/+<sup>Bl6</sup>* and *+<sup>129</sup>/+<sup>Bl6</sup>* MEFs halt growth and undergo apoptosis. This expansion has been repeated at least 3 times with the same results. For now, more details into CTCF, Med1, and Smc1 gene expression will be presented in Chapter 5. A final discussion and evaluation of how much useful information ChIP-seq experiments could provide to this project will be presented in Chapter 6, to determine whether it is worth pursuing these new set of experiments.

I observed a high degree of overlap of differentially interacting regions between *del*<sup>129</sup> and *del*<sup>Bl6</sup> chromosomes after comparisons with their corresponding WT homologues. Up to

~33% of the  $del^{129}$  regions are shared with  $del^{Bl6}$ , while  $del^{Bl6}$  shares ~50% of its differentially interacting regions with  $del^{129}$ . The high overlap ratios between differentially interacting regions of  $del^{129}$  and  $del^{Bl6}$  suggests global mechanisms of chromatin architecture regulation which are common to both homologous chromosomes. This hypothesis is further strengthened by the observation that up to ~92Mb of chromosome 4 sequences is included for differentially interacting regions detected for viewpoints 3-10 inside the CNV in  $del^{Bl6}$  chromosomes. A straightforward hypothesis is that shared chromatin interaction differences are due to global transcriptional changes, given the observed high correlation between fold changes of C57Bl6/J and 129S5/SvEv<sup>Brd</sup> alleles (Chapter 5). Specific TFs, chromatin remodelers, or other proteins targeting these regions should be investigated, in order to elucidate what common mechanism ties changes in chromatin contacts for the  $del^{129}$  and  $del^{Bl6}$  chromosomes. However, testing such hypothesis is a challenge, given the hundreds of genes and altered cellular pathways in  $df/+^{Bl6}$  MEFs (Chapter 5). This hypothesis requires further investigation and will be discussed in Chapter 6.

Even after the exclusion of  $del^{Bl6}$  differential interactions from the  $del^{129}$  dataset, there remain 659 unique  $del^{129}$  differentially interacting regions covering ~23Mb (~15%) of chromosome 4. If we conservatively assume these changes are not caused by common regulatory signals for both chromosomes, we can assume that ~23Mbs of chromosome 4 sequence change contact probabilities simply by the shortening of 4.3Mb of the chromatin fiber. This is a considerable proportion of the chromosome, and the changes could be even higher. Our PE-4Cseq data only uses 4 viewpoints surrounding the deletion (1, 2, 11, 12) to assess changes in conformation. We are therefore not considering other regions of variation which are not reported by these viewpoints, either because no chromatin interactions exist

between these regions, or because of technical limitations. These numbers give us an idea of the profound impact that CNVs can have not only on gene expression, but also on chromosome organization, which may in turn feedback into functional outputs (see Chapter 6 for an extensive discussion on the topic).

Because the newly developed PE-4Cseq analysis methodology uses multi-viewpoint data, the viewpoint located ~83Mb away from the CNV start and the one covering *Rps13* on chromosome 7 were not analyzed. Additionally, we had performed amplification and sequencing of viewpoints 148.9 (viewpoint 1) and 154.9 (viewpoint 11) on *Chd5* KO/+ and *dp/+<sup>Bl6</sup>* MEFs. *dp/+<sup>Bl6</sup>* MEF data revealed biases in read number for the 154.9 viewpoint on *del<sup>129</sup>* compared to *del<sup>Bl6</sup>*, suggesting some form of recombination for this region. This observation, together with the 3D DNA FISH data presented in Chapter 3, prompted me to discard further studies on *dp/+<sup>Bl6</sup>* MEFs, given the several technical challenges as well as the uncertainty of analyzing *bona fide* duplication CNVs.

This chapter presented the magnitude of the chromatin interaction changes arising upon the occurrence of a 4.3Mb deletion in mouse 4E2. The putative functional implications of such changes will be presented in the next chapter, the RNA-seq analysis of *df/+<sup>Bl6</sup>* and *+<sup>129</sup>/+<sup>Bl6</sup>* MEFs

## Chapter 5: Gene expression characterization of $df/+^{Bl6}$ and $+^{129}/+^{Bl6}$ MEFs

With the development of 3C and 3C-based technologies, a clearer picture of the impact of chromatin interactions on gene expression emerged. Initial 3C studies confirmed the presence of long-range interactions at the extensively studied  $\beta$ -globin locus in mouse (Tolhuis *et al.*, 2002), while later discoveries include the long-range interactions established by the  $\alpha$ -globin locus (Vernimmen *et al.*, 2007), the TH2 locus (Spilianakis and Flavell, 2004), and the Igf2 locus (Murrell *et al.*, 2004) with their regulatory elements. In addition to single gene studies, regional and genomic analyses have also been performed. For example, marked chromatin re-organization at the subMb scale was observed in various regions during ES cell to neural progenitor cell (NPC) differentiation (Phillips-Cremins *et al.*, 2013). A study performed on a human breast cancer cell line revealed the presence of specific long-range chromatin interactions formed between estrogen receptor  $\alpha$  binding sites and up-regulated genes upon oestrogen treatment (Fullwood *et al.*, 2009). More recently, the comprehensive analysis of ENCODE pilot project regions through 5C revealed complex networks of long-range interactions between promoters and distal regulatory elements (Sanyal *et al.*, 2012). These examples, selected from numerous published reports correlating chromatin architecture with transcriptional outputs, underscore the importance of chromatin interactions in quantitatively and temporally controlling gene expression.

Because of its role as a mode of transcriptional control, disruption of chromatin interactions due to CNVs can have functional implications by altering expression patterns of distal and neighboring genes. Genome-wide studies performed in HapMap cell lines revealed widespread genetic associations of CNVs and gene expression changes in *cis* over large genomic distances (Stranger *et al.*, 2007). A recurrent DNA deletion in human chromosome 7



causing Williams-Beuren syndrome not only induced expression changes for the genes in the aneuploid segment, but also altered expression of diploid genes lying near the breakpoints and up to 6.5Mb away, which are thought to play functional roles in the disease pathology (Merla *et al.*, 2006). A different analysis performed on a mouse model for Smith-Magenis and Potocki-Lupski syndromes also detected altered expression of genes outside of the rearranged segment in chromosome 11, extending over half a Mb from the rearranged segment (Ricard *et al.*, 2010). In mice, hundreds of CNVs show significant associations with expression profiles, constituting up to ~30% of strain-specific transcriptional variation in hematopoietic stem and progenitor cells. Notably, most of such associations occur between CNVs and genes mapping outside of the rearranged sequence (Cahan *et al.*, 2009). All of these observations have led to the hypothesis that CNVs have a complex effect on gene transcription that might involve altered chromatin structure.

Physically, CNVs could alter TAD structures by deletion of boundary regions, potentially joining differentially regulated regions along the chromosome (i.e. more active vs more silent). CNVs could also affect preferential associations between gene promoters and regulatory elements, either by deletion or re-positioning along the chromatin fiber. Such scenarios could have an impact in the transcriptional activity of the affected genes.

To assess the potential functional impact of differentially interacting regions detected in the *del*<sup>I29</sup> and *del*<sup>Bl6</sup> chromosome, I explored the relationship between these and gene expression profiles for *df*/<sup>Bl6</sup> and <sup>I29</sup>/<sup>Bl6</sup> MEFs.

## 5.1 Combined and allele-specific RNA-Seq analysis of $df/+^{Bl6}$ and $+^{129}/+^{Bl6}$ MEFs

RNA from seven independent MEF lines was isolated ( $+^{129}/+^{Bl6}$ : 129S5E88, 129S5E90, 129S5E95;  $df/+^{Bl6}$ : 129S5E36, 129S5E56, 129S5E71, 129S5E98). PolyA+ RNA was prepared and used to develop stranded libraries for PE 100 sequencing on the Illumina HiSeq platform (see Chapter 7 & 8 for experimental details). 3  $dp/+^{Bl6}$  samples (129S5E32, 129S5E60, 129S5E61) were also analyzed, but given the scope solely on  $df/+^{Bl6}$  and  $+^{129}/+^{Bl6}$  chromatin interaction comparisons, the report of RNA-Seq analyses of such samples is delegated to Supplemental Table 5.1.

Currently, there are few publicly available pipelines for the allele-specific analysis of RNA-Seq data (Rozowsky *et al.*, 2011; Turro *et al.*, 2011; Pandey *et al.*, 2013). However, their use is convoluted, and often tailored to the analysis of well-annotated human data. Because of the need to assign allele-specific values to differentially expressed (DE) genes between  $df/+^{Bl6}$  and  $+^{129}/+^{Bl6}$  MEFs, I decided to establish a collaboration with Emily Wong, from the group of Paul Flicek at the EMBL/EBI, given her expertise in haplotype-specific genomic analyses.

The allele-specific RNA-Seq analysis pipeline starts by separately aligning reads derived for each sample to both the C57BL/6J and 129S5/SvEv<sup>Brd</sup> transcriptomes [Table 5.1]. The C57BL/6J transcriptome was used as downloaded from Ensembl gene set version 72 (assembly version: mm10). For all all graphs, correlations, and overlap analyses presented in this chapter, gene locations were mapped to the mm9 genome using UCSC LiftOver. The 129S5/SvEv<sup>Brd</sup> transcriptome was constructed by modifying the C57BL/6J transcriptome using SNPs and indels calls from Keane *et al.*, 2011. Where multiple transcripts exist for a

gene, only the longest transcript was selected as representative for the gene in the transcriptome.

Alignments were performed using the GSNAP algorithm with the parameter of no mismatches (-m 0) (Wu and Nacu, 2010). As expected, no reads were detected inside the deletion CNV for 129S5/SvEv<sup>Brd</sup> alleles in *df/+<sup>Bl6</sup>* MEF samples [Supp. Fig. 5.1A,B,C]. Reads were filtered to keep only those with one best mapping location. To obtain estimates of expression values, only those reads aligning at a gene location were counted if both reads of a PE set were mapped to the same gene. To avoid biological interpretation from mapping noise, genes with less than 10 reads mapping to each allele were excluded if this occurred across genotypes.

Differential expression analyses were performed using the R Bioconductor package – DESeq (Anders and Huber, 2010), using an FDR cut-off of 0.05. A combined (non-allele-specific) differential expression analyses was performed (pairwise between WT and deletion) using counts summed from both alleles. Allele-specific analyses were performed only using reads that mapped to the transcriptome of each strain and compared in a pairwise manner between *+<sup>129</sup>/+<sup>Bl6</sup>* and *df/+<sup>Bl6</sup>* samples. To account for the allelic mapping biases that are a result of more reads mapping to the C57BL/6J transcriptome, we tested for changes in the proportion of reads mapping to each allele between deletion and WT genotypes, on a gene by gene basis, to determine whether similar degree of changes to expression levels occurred between alleles. Counts were normalized using DESeq and tests were done using the R function, `prop.test`, using median counts across replicates and p-values were adjusted for multiple testing in R using the `fdr` method (adjusted p-value cut-off = 0.01).

Name	ReadsP1	ReadsP2	Align_B16 (3_mm)	Align_129 (0_mm)	Align_B16 (0_mm)	Count129 (SNPs)	CountB16 (SNPs)
df36	254,991,068	254,991,068	105,325,376	90,607,701	90,837,633	3,565,724	6,333,456
df56	175,000,776	175,000,776	70,553,894	61,824,531	61,984,940	2,336,369	4,156,336
df71	243,001,260	243,001,260	100,558,258	85,608,934	85,797,168	3,219,026	5,668,421
df98	616,025,948	616,025,948	231,143,538	218,401,667	218,892,531	8,208,931	14,385,113
wt88	426,804,972	426,804,972	178,458,610	152,213,251	152,603,867	5,926,046	10,422,893
wt90	348,462,600	348,462,600	145,418,622	125,723,669	126,054,862	4,890,585	8,587,854
wt95	341,077,428	341,077,428	141,488,075	122,691,344	123,009,936	4,792,645	8,451,820

**Table 5.1 RNA-Seq mapping stats per sample**

Shown are the total number of reads obtained, and the number of aligned reads to each transcriptome.

RNA-Seq experiments of three  $+^{129}/+^{Bl6}$  and four  $df/+^{Bl6}$  MEF lines revealed 1,345 combined DE genes between both genotypes [Table 5.2] [Fig. 5.1A,B] [Supp. Table 5.2]. 118 of those genes are located in chromosome 4, 31 fall within the 4E2 region, and 28 inside the CNV. 59% of the 1345 genes show an increase in expression in  $df/+^{Bl6}$  MEFs (0.9 log2fold change average), and the remaining 41% show a decrease in expression (0.97 log2fold change average). Enrichment analysis WEB-based GENE SeT AnaLysis Toolkit (WebGestalt) with hypergeometric tests and Bonferroni corrections (Zhang *et al.*, 2005) revealed 4E2 and 4E as the cytogenetic bands with the most significant DE clustering locations in  $df/+^{Bl6}$  MEFs ( $p=1.48e-11$ ,  $p=3.26e-08$ , respectively). Interestingly, the 4E2D region, directly upstream of 4E, was the second most significant DE clustering location in the genome ( $p=2.27e-08$ ).

At the allelic level, 257 129S5/SvEv<sup>Brd</sup> alleles were DE, 39 of them in chromosome 4 (24 in 4E2, and 22 inside the deletion CNV) [Supp. Table 5.3]. ~52% of these genes show an increase in expression in  $df/+^{Bl6}$  MEFs compared to  $+^{129}/+^{Bl6}$  (1.12 log2fold average), while the remaining ~48% decrease their expression with an average of 2.3 in the log2fold scale. On the other hand, 326 C57BL/6J alleles were DE, with 39 in chromosome 4 (17 in 4E2, 12 lying inside CNV) [Supp. Table 5.4]. ~56% of these genes show an increase in expression in  $df/+^{Bl6}$  MEFs (1.1 log2fold average), while the remaining ~44% decrease their expression (1 log2fold average).

Both allelic sets cluster in the 4E2 and 4E region ( $p<0.001$ ) [Fig. 5.1A,B]. Of these, 189 genes are mis-regulated at both alleles, and 27 are in chromosome 4 [Supp. Table 5.5]. Interestingly, DE genes are strongly correlated between their allelic fold change ( $\rho=0.95$ ,  $p=2.2e-16$ ), which indicate *trans* effects on transcription, where mRNA levels are regulated similarly between the alleles [Fig. 5.2]. Not surprisingly, the exception to this phenomenon

are the deleted alleles located inside the CNV region, which exhibit a higher fold change for the 129S5/SvEv<sup>Brd</sup> allele only.

In order to validate levels of expression change detected in RNA-Seq, a set of 9 genes within the CNV region was randomly selected for C57Bl6/J allele-specific and total combined RNA qPCR experimental validations [Table 5.3] [Supp. Table 5.6 for list of primers used] [Supp. Fig. 5.2 for primer validation reactions]. Geometric mean of expression values for CycloB, Gapdh, and Pabpc1 were used as normalization controls (Chapter 8 for details on qPCR reactions). Overall, the qPCR results reflect the decrease in expression for these genes as detected in the combined RNA-Seq DE analysis [Fig. 5.3A]. However, qPCR amplifications using C57Bl6/J allele-specific primers for Gpr153, Klhl21, and Phf13 genes showed no statistically significant changes between expression of C57Bl6/J alleles in *df/+<sup>Bl6</sup>* compared to *+<sup>129</sup>/+<sup>Bl6</sup>* [Fig. 5.3B]. The later observation could be derived from the differences in sensitivity between both techniques when evaluating fold changes in expression. For example, the 3 assessed genes have less than 1 fold change in expression value according to RNA-Seq C57Bl6/J allelic results, which could explain the lack of qPCR detection differences.

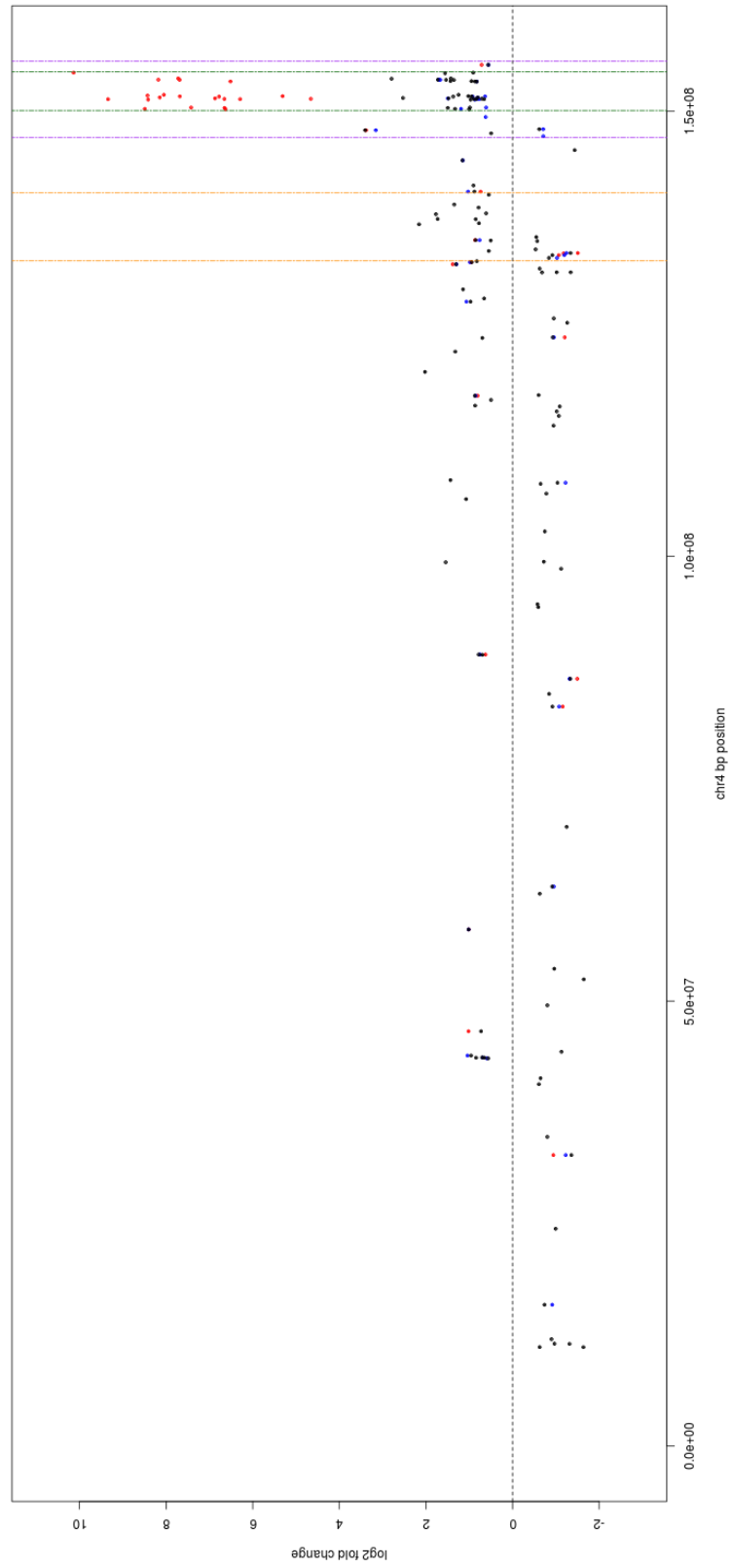
Mis-regulated KEGG pathways for the combined as well as allelic DE genes include cell cycle ( $p=3.74e-42$ ), DNA replication ( $p=3.48e-24$ ), metabolic pathways ( $p=1.09e-17$ ), homologous recombination ( $p=2.37e-16$ ), among others [Supp. Table 5.7]. This is in agreement with previous findings made in *df/+<sup>Bl6</sup>* MEFs for the identification of tumor suppressors (Bagchi *et al*, 2007), and with identified enriched phenotypes which span tumorigenesis ( $p=1.15e-07$ ), growth/size phenotype ( $p=2.13e-08$ ), abnormal cell physiology ( $p=5.76e-22$ ), as well as mortality/aging ( $p=2.44e-32$ ), among others [Supp. Table 5.8].

Gene set	Genome	chr4	4E2	CNV
DE_129	257	39	24	22
DE_B16	326	39	17	12
DE_combined	1345	118	31	28

**Table 5.2 DE summary for  $df/+^{B16}$  and  $+^{129}/+^{B16}$  MEF RNA-Seq data.**

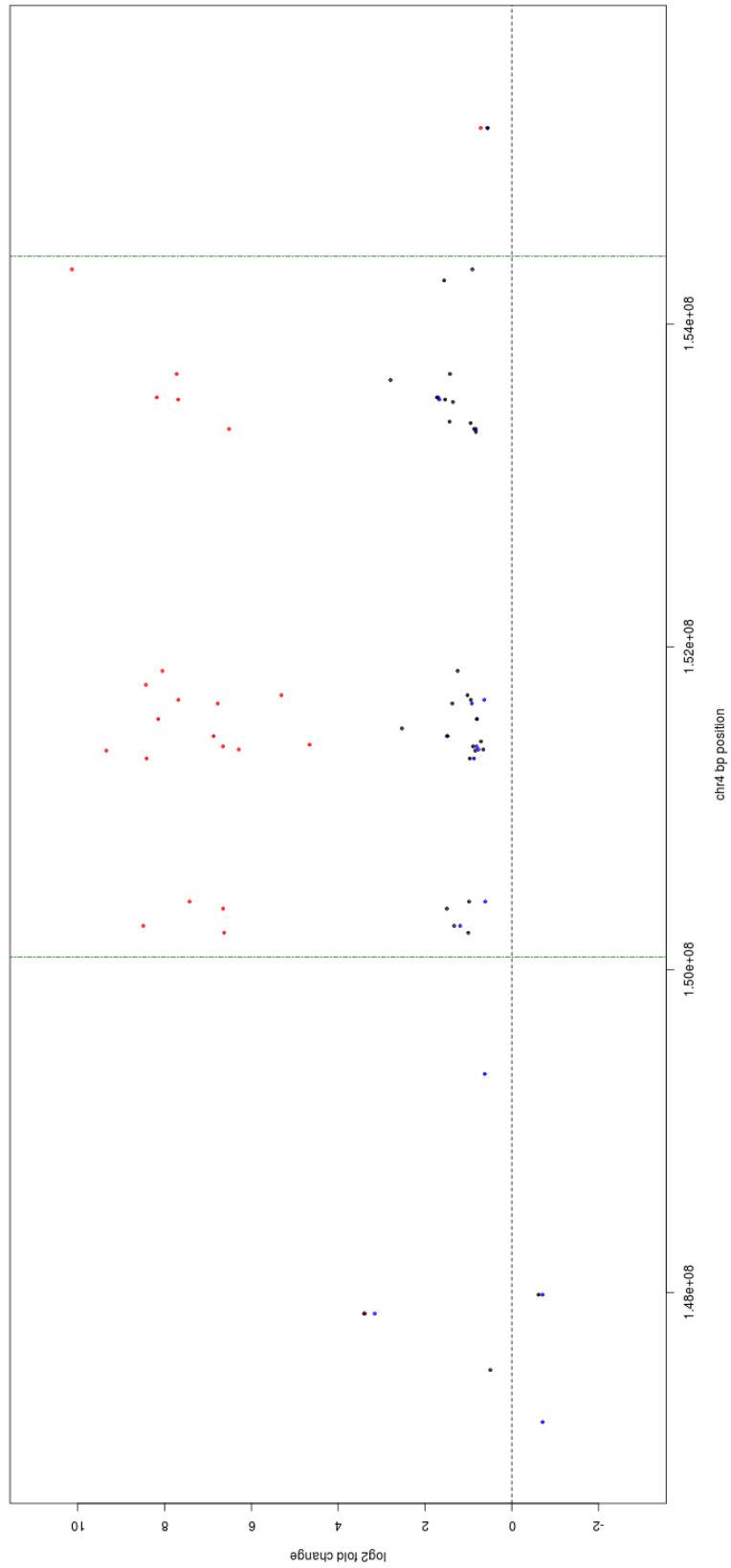
Shown are combined and allele-specific DE number of genes for the whole genome, and within chromosome 4, 4E2, and CNV coordinates.

A)



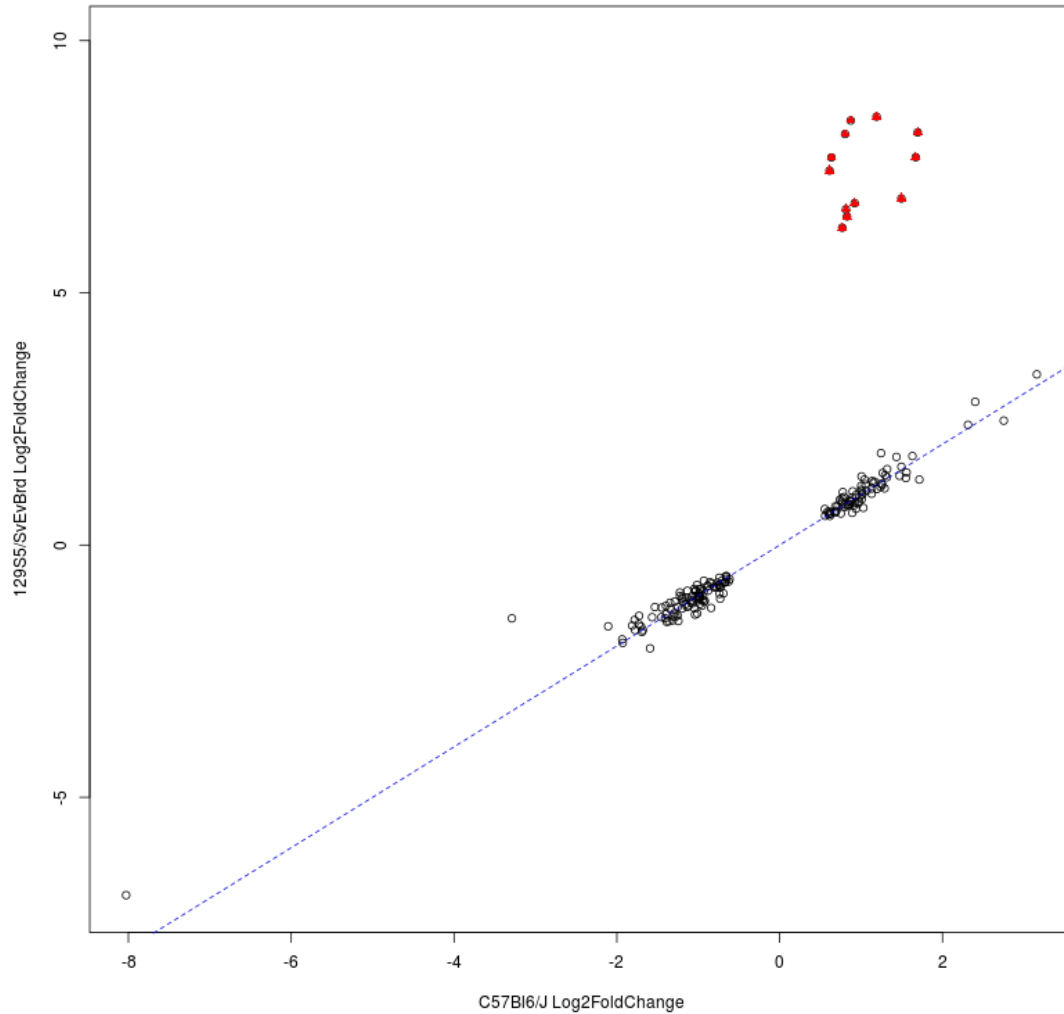


**B)**



### Figure 5.1 Chromosome 4 depictions of DE genes

**A)** Chromosome 4 view of DE genes, for both combined (black), and allele-specific results (129S5/SvEv<sup>Brd</sup> in red and C57Bl6/J in blue). Vertical lines indicate positions of band 4E2D3 (orange), 4E2 (purple), and CNV (dark green). Notice the clustering of DE genes in all 3 categories nearby the deletion CNV region. Horizontal dashed black line corresponds to a zero log<sub>2</sub>Fold change in gene expression. Positive log<sub>2</sub>Fold changes indicate higher expression in +<sup>129</sup>/+<sup>Bl6</sup>, while negative log<sub>2</sub>Fold changes indicate increased expression in *df*/+<sup>Bl6</sup> MEFs. **B)** Zoom into 4E region for the same features discussed in A). Notice the decrease in gene expression for genes located inside the deletion CNV (upper y scale points shown in red).



**Figure 5.2 High degree of correlation between log<sub>2</sub>FoldChange DE values between 129S5/SvEv<sup>Brd</sup> and C57Bl6/J alleles in *df/+<sup>Bl6</sup>* and *+<sup>129</sup>/+<sup>Bl6</sup>* MEFs**

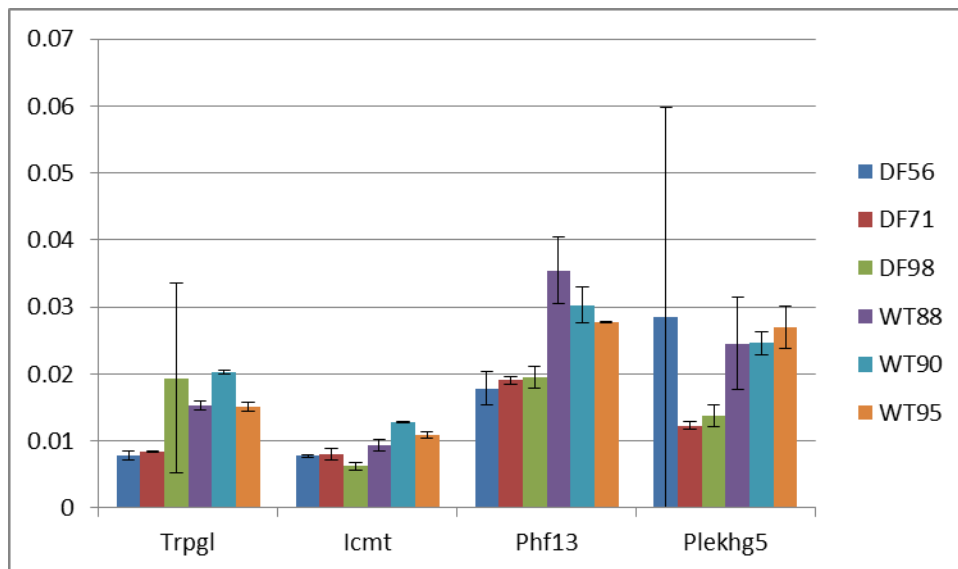
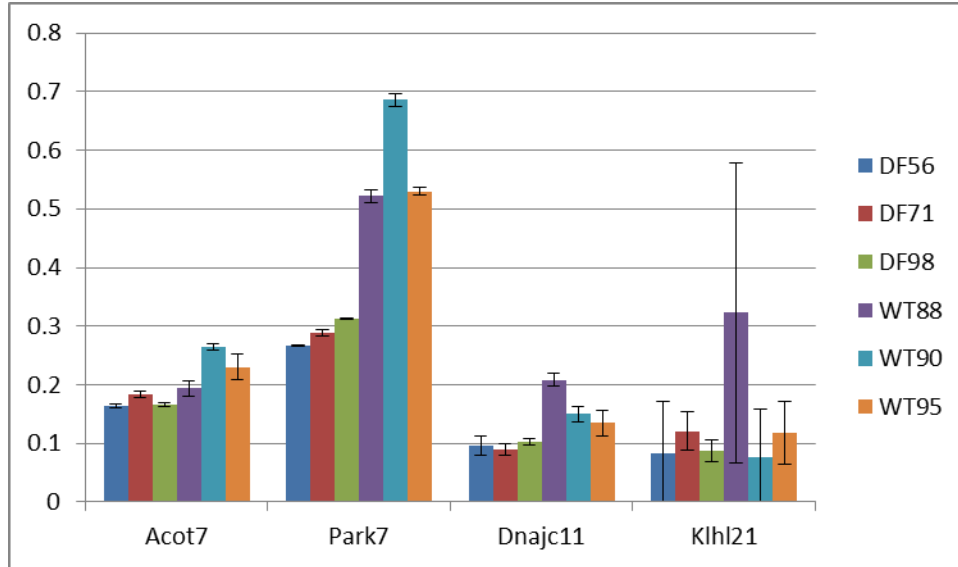
Genes shown in red fall within the CNV sequence, and therefore display a different behavior compared to the rest of alleles. Observe the high degree of correlation between allelic fold changes when compared to a perfect correlation score (1, blue dashed line). Outlier gene in the -8 value of x axis corresponds to gene ENSMUSG00000027596 (MGI name “a”, an agouti-signaling protein precursor) located in chromosome 2. a overexpression is expected in *df/+<sup>Bl6</sup>* given that it is a selection transgene integrated into the engineered *df* chromosome.

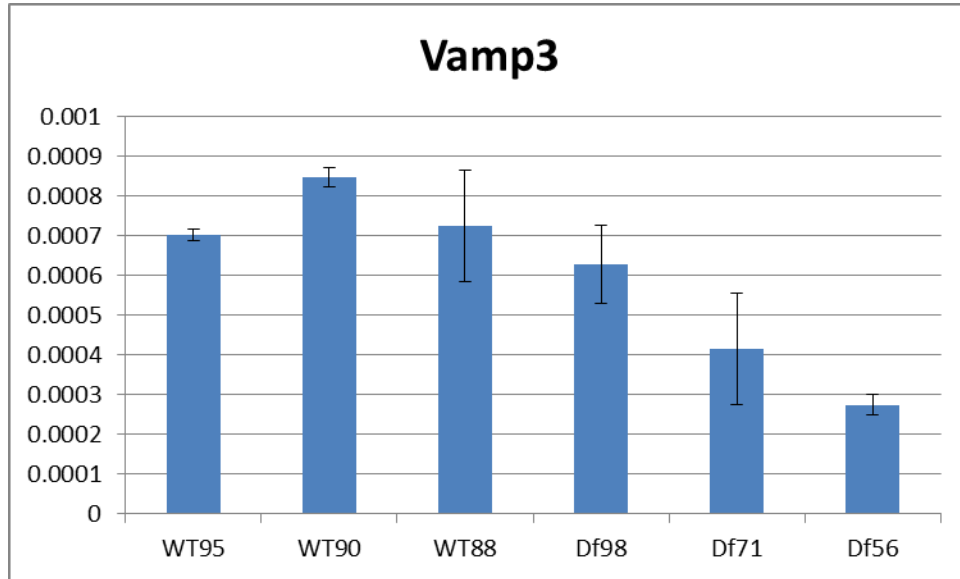
Ensembl Gene ID	Chr	Gene Start	Gene End	Strand	MGI symbol	log2fold_comb	log2fold_B16	MGI Description
ENSMUSG00000028964	4	150271242	150288546	-1	Park7	1.33	1.19	Parkinson disease (autosomal recessive, early onset) 7
ENSMUSG00000028955	4	150421409	150432072	-1	Vamp3	0.99	0.61	vesicle-associated membrane protein 3
ENSMUSG00000039768	4	151307800	151356246	1	Dnalc11	0.97	0.87	DnaJ (Hsp40) homolog, subfamily C, member 11
ENSMUSG00000047777	4	151363742	151370367	-1	Phf13	0.66	0.77	PHD finger protein 13
ENSMUSG00000073700	4	151382912	151391789	1	Klhl21	0.89	0.81	kelch-like 21 (Drosophila)
ENSMUSG00000039713	4	151446607	151489509	1	Plekhg5	1.48	1.50	pleckstrin homology domain containing, family G (with RhoGef domain) member 5
ENSMUSG00000028937	4	151552243	151645964	1	Acot7	0.81	0.80	acyl-CoA thioesterase 7
ENSMUSG00000042804	4	151648341	151659446	1	Gpr153	1.37	0.92	G protein-coupled receptor 153
ENSMUSG00000039662	4	151671336	151681230	1	Icmt	0.94	0.64	isoprenylcysteine carboxyl methyltransferase
ENSMUSG00000029030	4	153531594	153534775	-1	Tprgl	1.54	1.67	transformation related protein 63 regulated like
ENSMUSG000000057751	4	153544839	153649822	1	Megf6	1.73	1.70	multiple EGF-like-domains 6

**Table 5.3 Selected genes for RNA-Seq validations with their corresponding gene expression**

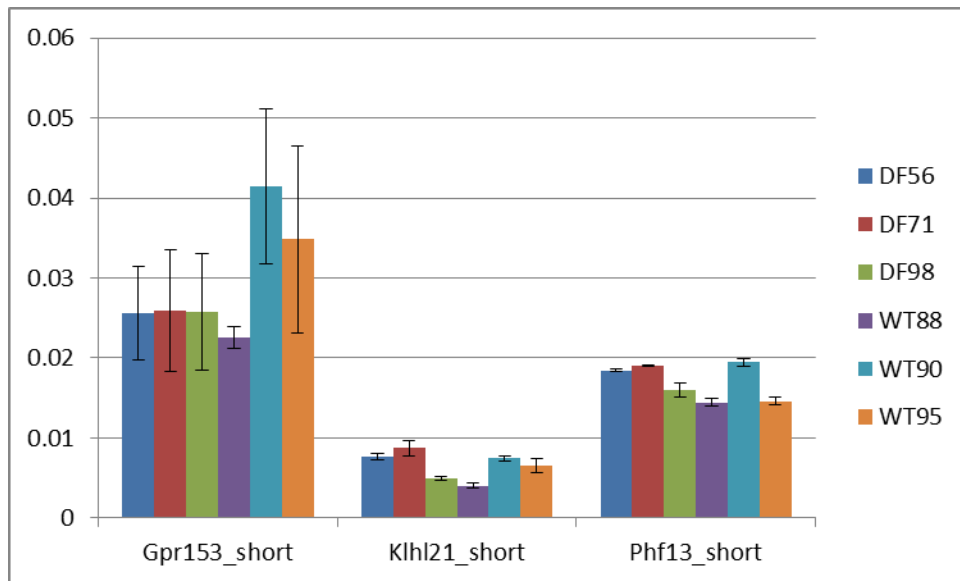
Changes expressed in log<sub>2</sub> scale in DE C57Bl6/J and DE combined analyses

A)





**B)**



**Figure 5.3 qPCR validations for expression values for 9 genes inside the deletion CNV**

**A)** qPCR CT values for 129S5E56, 129S5E56, and 129S5E56  $df/+^{Bl6}$  MEFs against 129S5E88, 129S5E90, and 129S5E96  $+^{129}/+^{Bl6}$  MEFs, using primers assessing transcripts derived from both C57Bl6/J and 129S5/SvEv<sup>Brd</sup> alleles. **B)** qPCR CT values for the same  $df/+^{Bl6}$  and  $+^{129}/+^{Bl6}$  MEFs as in A), using primers assessing transcripts derived from C57Bl6/J alleles only. Notice there is no significant difference between expression values for these genes in  $df/+^{Bl6}$  compared to  $+^{129}/+^{Bl6}$ .



## 5.2 Enriched DE content within *del*<sup>129</sup> and *del*<sup>Bl6</sup> differentially interacting regions

Upon completion of the RNA-Seq analysis, I was able to investigate the associations between DE genes and differentially interacting regions in *df/+<sup>Bl6</sup>* and *+<sup>129</sup>/+<sup>Bl6</sup>* MEFs. ~26% of the DE 129S5/SvEv<sup>Brd</sup> alleles fall inside *del*<sup>129</sup> regions with altered contact probabilities (presented in Chapter 4), while 44% of DE combined genes are contained within these regions [Table 5.4]. When compared against the whole annotated gene set for chromosome 4, 37% of these genes fall within differentially interacting regions. MC simulations were performed to assess the significance of DE genes and differentially interacting region overlaps. The number of DE combined genes falling inside *del*<sup>129</sup> differentially interacting regions is highly significant when compared against associations on randomly selected regions (p-val < 0.001). DE 129S5/SvEv<sup>Brd</sup> associations barely exceed the 0.05 p-val limit (p-val 0.06), probably due to a decrease in statistical power given the smaller number of 129S5/SvEv<sup>Brd</sup> DE alleles. Interestingly, there is also a significant enrichment in overlaps between chromosome 4 annotated genes and differentially interacting regions [Table 5.5]. *del*<sup>Bl6</sup> differentially interacting regions also display strong enrichment with gene content. 59% of DE C57Bl6/J alleles and 54% of DE combined genes are contained within *del*<sup>Bl6</sup> [Table 5.6]. 29% of chromosome 4 annotated genes fall within these regions. The associated overlaps are highly significant, as determined by MC simulations (p < 0.001) [Table 5.7]. Except for the *del*<sup>129</sup> and *del*<sup>Bl6</sup> intersection overlaps with DE 129S5/SvEv<sup>Brd</sup> and C57Bl6/J alleles, as well as unique *del*<sup>Bl6</sup> differentially interacting regions, the rest of the unique and shared differentially interacting regions for *del*<sup>129</sup> and *del*<sup>Bl6</sup> displayed associated enrichment with DE and total gene overlaps [Table 5.8. Table 5.9]. An example of *del*<sup>129</sup> differentially

interacting regions and DE gene positions is shown in Figure 5.4, focusing on the 141-155.6Mb segment of chromosome 4. An alternative visualization of these features and its integration with structural protein binding regions is shown in Figure 5.5 for 147-155.6Mb of *del<sup>I29</sup>* and *del<sup>Bl6</sup>* [whole chromosome view in Supp. Fig. 5.4].

It is clear that several DE genes fall inside *del<sup>I29</sup>* and *del<sup>Bl6</sup>* differentially interacting regions, and that these overlaps are highly significant when compared to overlaps of randomly chosen regions. Chromatin contact changes could potentially impact gene expression by altering the patterns of associations between gene promoters and their regulatory elements. In an effort to understand more about the nature of these associations in our PE-4Cseq and RNA-Seq datasets, I explored the enrichment of enhancer elements falling within *del<sup>I29</sup>* and *del<sup>Bl6</sup>* differentially interacting regions, as well as the patterns of correlation between DE genes and differential interactions (increase/decrease in interactions and expression). I downloaded the available histone H3 lysine 27 acetylation (H3K27ac) and histone H3 lysine 4 monomethylation (H3K4me1) ChIP datasets produced by the ENCODE project on C57Bl6 MEFs ([www.encodeproject.org](http://www.encodeproject.org)). H3K4me1 is a mark for poised enhancers, while H3K27ac is associated with active enhancers (Creyghton *et al.*, 2010; Rada-Iglesias *et al.*, 2011). Using MC simulations, I discovered a significant enrichment of overlaps between *del<sup>I29</sup>* and *del<sup>Bl6</sup>* differentially interacting regions and H3K4me1 and H3K27ac marks (p-val < 0.001) [Table 5.10, Table 5.11]. However, no obvious associations exist between the magnitudes of DE log2fold and differential interaction changes (Spearman rank correlation test, p-val >0.05) [Supp. Table 5.9A,B]. I also did not observe significant correlations between the direction of the contact probability change (increase/decrease) and the direction and magnitude of the DE changes [Supp. Table 5.9,A,B].

Region	diff sites	dif site bp	no. sites DE 129	%	bp sites DE 129	%	no. DE 129	%
chr4_dfwt_1	291	12,691,291	4	1	1,232,022	10	4	10
chr4_dfwt_2	183	8,303,183	2	1	1,076,023	13	2	5
chr4_dfwt_11	318	17,195,318	5	2	1,905,028	11	6	15
chr4_dfwt_12	163	6,316,163	1	1	524,011	8	1	3
chr4_dfwt_all	608	34,976,608	9	1	4,889,058	14	10	26
Total DE 129S5 genes	39							

Region	diff sites	dif site bp	no. sites DE com	%	bp sites DE com	%	no. DE com	%
chr4_dfwt_1	291	12,691,291	22	8	349,004	3	20	17
chr4_dfwt_2	183	8,303,183	23	13	193,002	2	21	18
chr4_dfwt_11	318	17,195,318	28	9	362,005	2	29	25
chr4_dfwt_12	163	6,316,163	11	7	66,001	1	13	11
chr4_dfwt_all	608	34,976,608	58	10	1,025,009	3	52	44
Total DE comb genes	118							

Region	diff sites	dif site bp	no. sites genes	%	bp sites genes	%	no. genes	%
chr4_dfwt_1	291	12,691,291	232	80	10,824,232	85	478	16
chr4_dfwt_2	183	8,303,183	154	84	7,358,154	89	358	12
chr4_dfwt_11	318	17,195,318	252	79	14,523,252	84	590	20
chr4_dfwt_12	163	6,316,163	124	76	5,399,124	85	256	8
chr4_dfwt_all	608	34,976,608	469	77	30,045,469	86	1104	37
Total chr4 genes	3014							

**Table 5.4 *del*<sup>129</sup> differentially interacting regions overlap with DE 129S5/SvEv<sup>Brd</sup> alleles, combined genes, and total annotated genes in chromosome 4**

Column 1, *Region*, refers to the analyzed viewpoints (1, 2, 11, and 12, as well as their combined lengths -all-). Column 2, *diff sites*, reports the number of differentially interacting regions detected for the corresponding viewpoint, while column 3 reports the bp size of all regions combined. Column 4 reports the number of differentially interacting regions overlapping DE genes (129S5/SvEv<sup>Brd</sup>, combined) or the whole annotated gene set for chromosome 4. Column 8, *no. feature*, reports the number of DE genes (129S5/SvEv<sup>Brd</sup>, combined) or the whole annotated gene set for chromosome 4 that overlap with differentially interacting regions. Columns 5, 7, and 9 report associated % values regarding their previous column numbers.

Region	diff sites	dif site bp	no. DE 129	%	MC simulation	p-val
chr4_dfwt_all	608	34,976,608	10	26	62	0.062
Total DE 129S5 genes	39					

Region	diff sites	dif site bp	no. DE com	%	MC simulation	p-val
chr4_dfwt_all	608	34,976,608	61	52	0	0
Total DE comb genes	118					

Region	diff sites	dif site bp	no. genes	%	MC simulation	p-val
chr4_dfwt_all	608	34,976,608	1240	41	0	0
Total chr4 genes	3014					

**Table 5.5 MC simulations to assess the significance of *del*<sup>129</sup> and DE and total annotated genes overlap**

Number of features in column 4 may differ from Table 5.4 as we count total number of intersections between both datasets for this analysis. The CNV coordinates were excluded from analyses to avoid biases in the selection of random differentially interacting regions. p-values of 0 correspond to values <0.001, rounded down.

Region	diff sites	dif site bp	no. sites DE Bl6	%	bp sites DE Bl6	%	no. DE Bl6	%
chr4_dfw_t_bl6_1	281	12,014,281	12	4	1,031,012	9	14	36
chr4_dfw_t_bl6_2	200	8,052,200	9	5	495,009	6	9	23
chr4_dfw_t_bl6_11	196	7,923,196	7	4	434,007	5	7	18
chr4_dfw_t_bl6_12	160	5,522,160	7	4	325,007	6	7	18
chr4_dfw_t_bl6_all	594	27,579,594	21	4	1,860,021	7	23	59
Total DE C57Bl6 genes	39							

Region	diff sites	dif site bp	no. sites DE com	%	bp sites DE com	%	no. DE com	%
chr4_dfw_t_bl6_1	281	12,014,281	32	11	2,061,032	17	38	32
chr4_dfw_t_bl6_2	200	8,052,200	24	12	1,162,024	14	25	21
chr4_dfw_t_bl6_11	196	7,923,196	16	8	1,052,016	13	17	14
chr4_dfw_t_bl6_12	160	5,522,160	13	8	552,013	10	16	14
chr4_dfw_t_bl6_all	594	27,579,594	57	10	4,515,057	16	64	54
Total DE comb genes	118							

Region	diff sites	dif site bp	no. sites genes	%	bp sites genes	%	no. genes	%
chr4_dfw_t_bl6_1	281	12,014,281	222	79	10,236,222	85	445	15
chr4_dfw_t_bl6_2	200	8,052,200	148	74	6,691,148	83	274	9
chr4_dfw_t_bl6_11	196	7,923,196	146	74	6,459,146	82	309	10
chr4_dfw_t_bl6_12	160	5,522,160	120	75	4,750,120	86	204	7
chr4_dfw_t_bl6_all	594	27,579,594	441	74	23,318,441	85	876	29
Total chr4 genes	3014							

**Table 5.6 *del<sup>Bl6</sup>* differentially interacting regions overlap with DE C57Bl6/J alleles, combined genes, and total annotated genes in chromosome 4**

Column notations are as described in Table 5.4 in this chapter.

Region	diff sites	dif site bp	no. DE B16	%	MC simulation	p-val
chr4_dfwt_bl6_all	594	27,579,594	23	59	1	0.001
Total DE C57B16 genes	39					

Region	diff sites	dif site bp	no. DE com	%	MC simulation	p-val
chr4_dfwt_bl6_all	594	27,579,594	69	12	0	0
Total DE comb genes	118					

Region	diff sites	dif site bp	no. genes	%	MC simulation	p-val
chr4_dfwt_bl6_all	594	27,579,594	995	168	0	0
Total chr4 genes	3014					

**Table 5.7 MC simulations to assess the significance of *del<sup>B16</sup>* and DE/total annotated genes overlap**

Number of features in column 4 may differ from Table 5.6 as we count total number of intersections between both datasets for this analysis. p-values of 0 correspond to values <0.001, rounded down.

Region	diff sites	dif site bp	no. sites DE 129	%	bp sites DE 129	%	no. DE 129	%
bl6_minus_df_intpiece_diffinterregs	521	15,654,521	16	3	1,029,016	7	21	54
bl6_thatintersect_df_intpiece_diffinteregs	352	11,925,352	4	1	343,004	3	4	10
df_minus_bl6_intpiece_diffinterregs	659	23,051,659	8	1	383,008	2	8	21
df_thatintersect_bl6_intpiece_diffinteregs	352	11,925,352	4	1	343,004	3	4	10
Total DE 129S5 genes	39							

Region	diff sites	dif site bp	no. sites DE Bl6	%	bp sites DE Bl6	%	no. DE Bl6	%
bl6_minus_df_intpiece_diffinterregs	521	15,654,521	15	3	991,015	6	17	44
bl6_thatintersect_df_intpiece_diffinteregs	352	11,925,352	7	2	406,007	3	7	18
df_minus_bl6_intpiece_diffinterregs	659	23,051,659	14	2	854,014	4	12	31
df_thatintersect_bl6_intpiece_diffinteregs	352	11,925,352	7	2	406,007	3	7	18
Total DE C57Bl6 genes	39							

Region	diff sites	dif site bp	no. sites DE com	%	bp sites DE com	%	no. DE com	%
bl6_minus_df_intpiece_diffinterregs	521	15,654,521	44	8	1,996,044	13	52	44
bl6_thatintersect_df_intpiece_diffinteregs	352	11,925,352	27	8	1,161,027	10	26	22
df_minus_bl6_intpiece_diffinterregs	659	23,051,659	42	6	2,232,042	10	37	31
df_thatintersect_bl6_intpiece_diffinteregs	352	11,925,352	27	8	1,161,027	10	26	22
Total DE comb genes	118							

Region	diff sites	dif site bp	no. sites genes	%	bp sites genes	%	no. genes	%
bl6_minus_df_intpiece_diffinterregs	521	15,654,521	368	71	12,334,368	79	510	17
bl6_thatintersect_df_intpiece_diffinteregs	352	11,925,352	273	78	10,385,273	87	515	17
df_minus_bl6_intpiece_diffinterregs	659	23,051,659	474	72	18,818,474	82	765	25
df_thatintersect_bl6_intpiece_diffinteregs	352	11,925,352	273	78	10,385,273	87	515	17
Total chr4 genes	3014							

**Table 5.8 Unique and shared  $del^{129}$  and  $del^{Bl6}$  differentially interacting regions overlap for viewpoints 1, 2, 11, and 12 with DE C57Bl6/J alleles, DE 129S5/SvEv<sup>Brd</sup> alleles, combined genes, and total annotated genes in chromosome 4**

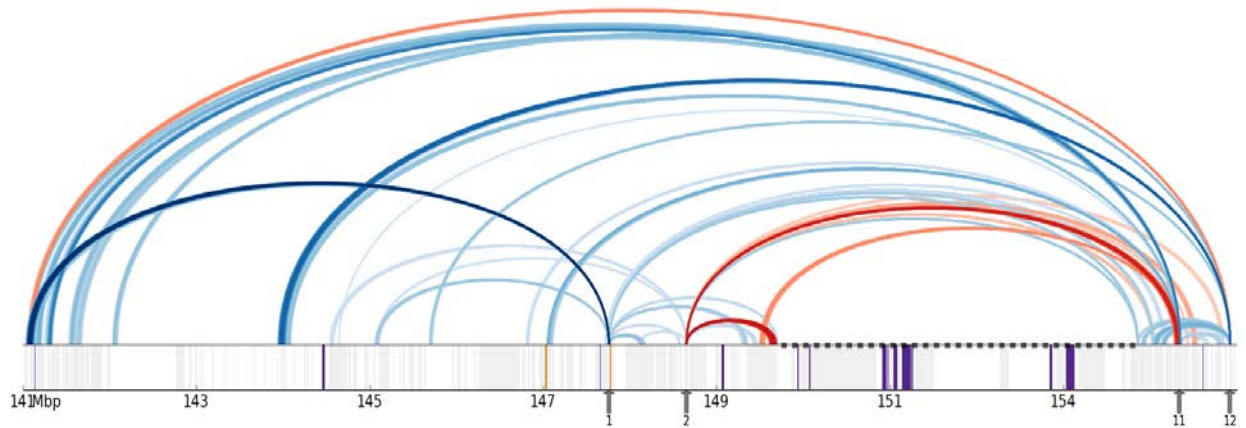
Column notations are as described in Table 5.4 in this chapter.

Region	diff sites	dif site bp	no. DE 129	%	MC simulation	p-val
bl6_minus_df_intpiece_diffinterregs	521	15,654,521	22	56	2	0.00
df_minus_bl6_intpiece_diffinterregs	659	23,051,659	9	23	30	0.03
df_thatintersect_bl6_intpiece_diffinterregs	352	11,925,352	4	10	807	0.81
Total DE 129S5 genes	39					
Region	diff sites	dif site bp	no. DE Bl6	%	MC simulation	p-val
bl6_minus_df_intpiece_diffinterregs	521	15,654,521	17	44	6	0.01
df_minus_bl6_intpiece_diffinterregs	659	23,051,659	14	36	25	0.03
df_thatintersect_bl6_intpiece_diffinterregs	352	11,925,352	7	18	336	0.34
Total DE C57Bl6 genes	39					
Region	diff sites	dif site bp	no. DE com	%	MC simulation	p-val
bl6_minus_df_intpiece_diffinterregs	521	15,654,521	55	47	0	0.00
df_minus_bl6_intpiece_diffinterregs	659	23,051,659	43	36	4	0.00
df_thatintersect_bl6_intpiece_diffinterregs	352	11,925,352	29	25	24	0.02
Total DE comb genes	118					
Region	diff sites	dif site bp	no. genes	%	MC simulation	p-val
bl6_minus_df_intpiece_diffinterregs	521	15,654,521	588	20	82	0.08
df_minus_bl6_intpiece_diffinterregs	659	23,051,659	894	30	2	0.00
df_thatintersect_bl6_intpiece_diffinterregs	352	11,925,352	586	19	0	0.00
Total chr4 genes	3014					

**Table 5.9 MC simulations to assess the significance of unique and shared *del*<sup>129</sup> and *del*<sup>Bl6</sup> differentially interacting regions for viewpoints 1, 2, 11, and 12 and DE/total annotated genes overlap**

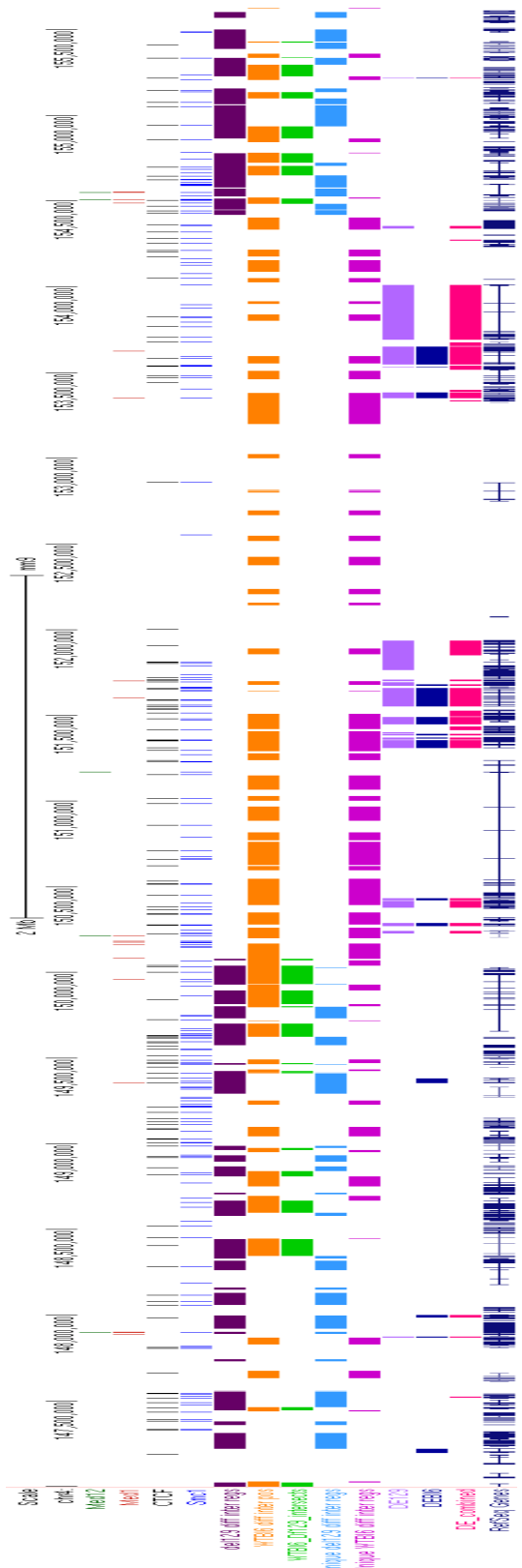
Number of features in column 4 may differ from Table 5.8 as we count total number of intersections between both datasets for this analysis. p-values of 0 in table are <0.001, rounded down.





**Figure 5.4 Rainbow plot of the differential signal surrounding the deletion region in *del<sup>129</sup>***

Each arc represents a long-range interaction that changed in the deletion chromosome (red/blue, increase/decrease in interaction of at least 10%, respectively). The chromosome panel shows in grey all annotated genes in this region, and the DE combined genes from RNA-Seq analysis is purple (down-regulated) and orange (up-regulated) in *df/+<sup>Bl6</sup>*. The dashed line corresponds to the deletion region.



**Figure 5.5 Graph of CTCF, Med1, Med12, Smc1 protein binding sites along 147-155.Mb of chromosome 4**

Middle rows in the figure display *del<sup>I29</sup>*, *del<sup>Bl6</sup>*, unique *del<sup>I29</sup>*, unique *del<sup>Bl6</sup>*, and shared *del<sup>I29</sup>-del<sup>Bl6</sup>* differentially interacting regions. DE genes positions for 129S5/SvEv<sup>Brd</sup>, C57Bl6/J, and combined analyses, as well as RefSeq genes, are shown in the bottom rows.

Region	diff sites	dif site bp	no. sites H3K27ac	%	no. H3K27ac	%
chr4_dfw_t_1	291	12,691,291	283	97	34063	12
chr4_dfw_t_2	183	8,303,183	173	95	26765	9
chr4_dfw_t_11	318	17,195,318	311	98	47901	17
chr4_dfw_t_12	163	6,316,163	156	96	17474	6
chr4_dfw_t_all	608	34,976,608	580	95	95759	34
Total H3K27ac sites	285273					
Region	diff sites	dif site bp	no. sites H3K4me1	%	no. H3K4me1	%
chr4_dfw_t_1	291	12,691,291	284	98	52404	12
chr4_dfw_t_2	183	8,303,183	177	97	40565	9
chr4_dfw_t_11	318	17,195,318	312	98	71479	16
chr4_dfw_t_12	163	6,316,163	156	96	26544	6
chr4_dfw_t_all	608	34,976,608	585	96	144764	33
Total H3K4me1 sites	439514					
Region	diff sites	dif site bp	no. sites H3K27ac	%	no. H3K27ac	%
chr4_dfw_t_bl6_1	281	12,014,281	272	97	35944	13
chr4_dfw_t_bl6_2	200	8,052,200	191	96	24400	9
chr4_dfw_t_bl6_11	196	7,923,196	190	97	20946	7
chr4_dfw_t_bl6_12	160	5,522,160	148	93	12693	4
chr4_dfw_t_bl6_all	594	27,579,594	563	95	74502	26
Total H3K27ac sites	285273					
Region	diff sites	dif site bp	no. sites H3K4me1	%	no. H3K4me1	%
chr4_dfw_t_bl6_1	281	12,014,281	272	97	53174	12
chr4_dfw_t_bl6_2	200	8,052,200	191	96	37550	9
chr4_dfw_t_bl6_11	196	7,923,196	190	97	32196	7
chr4_dfw_t_bl6_12	160	5,522,160	148	93	21441	5
chr4_dfw_t_bl6_all	594	27,579,594	561	94	114517	26
Total H3K4me1 sites	439514					
Region	diff sites	dif site bp	no. sites H3K27ac	%	no. H3K27ac	%
bl6_minus_df_intpiece_diffinterregs	521	15,654,521	484	93	37534	13
bl6_thatintersect_df_intpiece_diffinterregs	352	11,925,352	342	97	37184	13
df_minus_bl6_intpiece_diffinterregs	659	23,051,659	625	95	58888	21
df_thatintersect_bl6_intpiece_diffinterregs	352	11,925,352	342	97	37184	13
Total H3K27ac sites	285273			0		
Region	diff sites	dif site bp	no. sites H3K4me1	%	no. H3K4me1	%
bl6_minus_df_intpiece_diffinterregs	521	15,654,521	484	93	59007	13
bl6_thatintersect_df_intpiece_diffinterregs	352	11,925,352	346	98	55817	13
df_minus_bl6_intpiece_diffinterregs	659	23,051,659	633	96	89378	20
df_thatintersect_bl6_intpiece_diffinterregs	352	11,925,352	346	98	55817	13
Total H3K4me1 sites	439514					

**Table 5.10 *del*<sup>L29</sup> and *del*<sup>B16</sup>, as well as unique and shared *del*<sup>L29</sup> and *del*<sup>B16</sup> differentially interacting regions overlap with H3K27ac and H3K4me1 marks**

Column 1, *Region*, refers to the analyzed viewpoints (1, 2, 11, and 12, as well as their combined lengths -all-). Column 2, *diff sites*, reports the number of differentially interacting regions detected for the corresponding viewpoint, while column 3 reports the bp size of all regions combined. Column 4 reports the number of differentially interacting regions overlapping H3K27ac and H3K4me1 marks in chromosome 4. Column 6, *no. feature*, reports the number of H3K27ac and H3K4me1 marks in chromosome 4 that overlap with differentially interacting regions. Columns 5 and 7 report associated % values regarding their previous column numbers.

Region	diff sites	no. H3K27ac	%	MC simulation	p-val
chr4_dfwt_1	291	34063	12	0	0
chr4_dfwt_2	183	26765	9	0	0
chr4_dfwt_11	318	47901	17	0	0
chr4_dfwt_12	163	17474	6	0	0
chr4_dfwt_all	608	95759	34	0	0
Total H3K27ac sites chr4	285273				
Region	diff sites	no. H3K4me1	%	MC simulation	p-val
chr4_dfwt_1	291	52404	12	0	0
chr4_dfwt_2	183	40565	9	0	0
chr4_dfwt_11	318	71479	16	0	0
chr4_dfwt_12	163	26544	6	0	0
chr4_dfwt_all	608	144764	33	0	0
Total H3K4me1 sites chr4	439514				
Region	diff sites	no. H3K27ac	%	MC simulation	p-val
chr4_dfwt_bl6_1	281	35944	13	0	0
chr4_dfwt_bl6_2	200	24400	9	0	0
chr4_dfwt_bl6_11	196	20946	7	0	0
chr4_dfwt_bl6_12	160	12693	4	3	0.003
chr4_dfwt_bl6_all	594	74502	26	0	0
Total H3K27ac sites chr4	285273				
Region	diff sites	no. H3K4me1	%	MC simulation	p-val
chr4_dfwt_bl6_1	281	53174	12	0	0
chr4_dfwt_bl6_2	200	37550	9	0	0
chr4_dfwt_bl6_11	196	32196	7	0	0
chr4_dfwt_bl6_12	160	21441	5	0	0
chr4_dfwt_bl6_all	594	114517	26	0	0
Total H3K4me1 sites chr4	439514				
Region	diff sites	no. H3K27ac	%	MC simulation	p-val
bl6_minus_df_intpiece_diffinterregs	521	37534	13	0	0
bl6_thatintersect_df_intpiece_diffinterregs	352	37184	13	0	0
df_minus_bl6_intpiece_diffinterregs	659	58888	21	0	0
df_thatintersect_bl6_intpiece_diffinterregs	352	37184	13	0	0
Total H3K27ac sites chr4	285273		0		0
Region	diff sites	no. H3K4me1	%	MC simulation	p-val
bl6_minus_df_intpiece_diffinterregs	521	59007	13	0	0
bl6_thatintersect_df_intpiece_diffinterregs	352	55817	13	0	0
df_minus_bl6_intpiece_diffinterregs	659	89378	20	0	0
df_thatintersect_bl6_intpiece_diffinterregs	352	55817	13	0	0
Total H3K4me1 sites chr4	439514		0		

**Table 5.11 MC simulations to assess the significance of  $del^{I29}$  and  $del^{Bl6}$ , as well as unique and shared  $del^{I29}$  and  $del^{Bl6}$  differentially interacting regions for viewpoints 1, 2, 11, and 12 with H3K27ac and H3K4me1 marks genes overlap**

Column notations are as described in Table 5.8. p-values  $<0.001$  are shown in the table as zero.

### 5.3 Enriched DE content within *del*<sup>Bl6</sup> differentially interacting regions inside the CNV

The concept of transcriptional compensation was introduced in Chapter 4, in which the heterozygous active copy of one allele can increase its transcriptional output to levels similar to homozygous WT. Of special interest is the evaluation of transcriptional dosage compensation events for the C57Bl6/J genes that fall within CNV coordinates. After the deletion of their 129S5/SvEv<sup>Brd</sup> homologues, C57Bl6/J alleles in the *del*<sup>Bl6</sup> chromosome could potentially increase their levels of expression to compensate for the loss of the 129S5/SvEv<sup>Brd</sup> alleles.

Comparisons between normalized read counts for these CNV-contained genes in *df/+*<sup>Bl6</sup> and *+<sup>129</sup>/+*<sup>Bl6</sup> MEFs showed a generalized decrease in expression in *df/+*<sup>Bl6</sup> MEFs, with ratios ranging from 0.1-0.7 values compared to *+<sup>129</sup>/+*<sup>Bl6</sup> (average ~43%) [Table 5.12]. Dosage compensation events would require at least a 0.9-1 ratio in expression differences. Therefore, clear events of dosage compensation for CNV-contained genes after deletion are absent in the RNA-Seq datasets.

Although transcriptional compensation events were not detected in *df/+*<sup>Bl6</sup> MEFs, differentially interacting regions for viewpoints located inside the CNV in *del*<sup>Bl6</sup> overlap 79% of DE C57Bl6/J alleles, 85% of DE combined genes, and 67% of chromosome 4 annotated genes [Table 5.13]. However, there is no selective enrichment for overlaps between these regions and DE C57Bl6/J alleles ( $p > 0.05$ ) [Table 5.14]. A few viewpoints show enrichment for the overlap with DE combined genes and total annotated genes, but the overall covered regions do not show ratios above random expected levels of overlap ( $p > 0.05$ ) [Table 5.14].



Ensembl Gene ID	df1	df2	df3	df4	average df	wt1	wt2	wt3	average wt	ratio
ENSMUSG00000029032	7.5	10.3	1.1	42.8	15.4	116.4	127.7	76.3	106.8	0.1
ENSMUSG00000058183	0.0	4.4	0.0	2.3	1.7	9.3	11.6	13.4	11.4	0.1
ENSMUSG00000029055	3.2	1.5	5.3	14.0	6.0	38.0	27.1	44.8	36.6	0.2
ENSMUSG00000028943	35.3	42.7	11.7	86.9	44.1	239.6	262.3	263.5	255.2	0.2
ENSMUSG00000057751	167.7	361.1	774.5	830.7	533.5	1696.0	1788.4	1807.0	1763.8	0.3
ENSMUSG00000029059	129.3	107.6	87.9	66.2	97.7	284.4	288.7	293.4	288.8	0.3
ENSMUSG00000029030	3306.5	2961.2	2815.2	2871.3	2988.5	8452.4	8880.2	8737.6	8690.1	0.3
ENSMUSG00000028957	325.8	319.8	450.3	456.1	388.0	1048.7	1266.8	973.1	1096.2	0.4
ENSMUSG00000039713	684.8	490.8	920.7	930.7	756.8	1850.3	2539.1	1951.7	2113.7	0.4
ENSMUSG00000078350	253.2	458.4	363.4	929.8	501.2	1368.6	1499.0	1196.5	1354.7	0.4
ENSMUSG00000039410	456.2	387.7	543.5	734.4	530.4	1313.8	1547.0	1415.2	1425.3	0.4
ENSMUSG00000085069	0.0	0.0	20.1	21.2	10.3	21.8	33.3	27.5	27.5	0.4
ENSMUSG00000042333	20.3	16.2	18.0	5.0	14.9	38.0	37.9	41.7	39.2	0.4
ENSMUSG00000058498	67.3	28.0	9.5	14.9	29.9	77.8	68.9	87.3	78.0	0.4
ENSMUSG00000042804	2877.0	2636.9	3499.6	3140.5	3038.5	7452.9	8434.5	7735.3	7874.2	0.4
ENSMUSG00000029029	644.2	458.4	583.8	557.0	560.8	1425.9	1420.1	1467.1	1437.7	0.4
ENSMUSG00000028950	9.6	7.4	7.4	13.1	9.4	29.3	22.4	19.7	23.8	0.4
ENSMUSG00000028964	3417.6	3483.0	3103.3	3606.1	3402.5	8471.1	8586.2	8621.1	8559.5	0.4
ENSMUSG00000039577	351.5	316.9	253.2	302.1	305.9	728.8	680.2	776.4	728.5	0.4
ENSMUSG00000014592	11.8	19.2	13.8	22.1	16.7	39.2	36.4	40.1	38.6	0.4
ENSMUSG00000028931	272.4	364.1	988.5	182.4	451.8	995.2	1098.1	812.6	968.6	0.5
ENSMUSG00000028936	10693.1	8243.8	9740.2	8530.1	9301.8	19413.6	18567.6	18811.7	18931.0	0.5
ENSMUSG00000028936	10693.1	8243.8	9740.2	8530.1	9301.8	19413.6	18567.6	18811.7	18931.0	0.5
ENSMUSG00000028967	7765.8	6566.5	5332.6	10117.7	7445.6	16208.4	13314.5	15347.2	14956.7	0.5
ENSMUSG00000028955	4223.2	4119.7	3817.5	4576.9	4184.3	9016.3	7614.2	8260.8	8297.1	0.5
ENSMUSG00000039768	4800.1	4653.3	4440.5	4198.6	4523.1	8918.0	8677.5	8998.7	8864.7	0.5
ENSMUSG00000029028	1840.8	1672.9	1872.2	1798.8	1796.2	3421.2	3520.4	3485.7	3475.8	0.5
ENSMUSG00000039662	6387.6	4644.4	4353.6	3684.5	4767.5	9456.9	8730.1	9379.5	9188.8	0.5
ENSMUSG00000039838	9.6	4.4	9.5	11.3	8.7	20.5	17.8	11.8	16.7	0.5
ENSMUSG00000029056	1231.8	936.0	816.9	863.1	961.9	1775.6	1822.5	1832.2	1810.1	0.5
ENSMUSG00000073700	3107.8	2073.9	1806.5	1461.1	2112.3	3818.2	3659.7	4329.1	3935.7	0.5
ENSMUSG00000039523	2454.0	2551.4	2416.8	2362.5	2446.2	4714.5	4226.1	4477.0	4472.5	0.5
ENSMUSG00000039759	385.7	361.1	340.1	300.3	346.8	610.5	674.0	582.9	622.5	0.6
ENSMUSG00000029027	182.7	163.6	138.8	143.6	157.2	289.4	284.0	268.3	280.6	0.6
ENSMUSG00000047613	746.8	658.9	641.0	647.0	673.4	1195.6	1232.8	1165.1	1197.8	0.6
ENSMUSG00000028937	9983.7	7126.6	5400.4	6360.8	7217.9	13036.8	11039.3	13860.4	12645.5	0.6
ENSMUSG00000050545	63.0	110.5	66.7	89.6	82.5	145.0	141.6	140.8	142.5	0.6
ENSMUSG00000028948	1676.2	1737.8	1617.9	1402.5	1608.6	2665.6	2590.2	2659.7	2638.5	0.6
ENSMUSG00000047777	2225.4	2147.6	2399.8	2074.3	2211.8	3411.2	3873.2	3199.4	3494.6	0.6
ENSMUSG00000028952	272.4	281.5	318.9	241.8	278.7	447.5	403.2	395.7	415.5	0.7

**Table 5.12 Normalized reads counts for  $df/+^{Bl6}$  and  $+^{129}/+^{Bl6}$  MEFs and their associated  $df/+^{Bl6}$  over  $+^{129}/+^{Bl6}$  ratios**

Region	diff sites	dif site bp	no. sites DE B16	%	bp sites DE B16	%	no. DE B16	%
chr4_dfwt_3	455	19,050,455	9	2	445,009	2	8	21
chr4_dfwt_4	337	12,330,337	10	3	575,010	5	10	26
chr4_dfwt_5	344	13,336,344	9	3	528,009	4	10	26
chr4_dfwt_6	643	36,054,643	14	2	1,274,014	4	15	38
chr4_dfwt_7	625	36,953,625	13	2	1,153,013	3	17	44
chr4_dfwt_8	314	16,229,314	7	2	403,007	2	10	26
chr4_dfwt_9	505	22,735,505	14	3	920,014	4	14	36
chr4_dfwt_10	161	5,654,161	7	4	402,007	7	7	18
chr4_dfwt_all	1112	91,905,112	26	2	5,645,026	6	31	79
Total DE B16 genes	39							
Region	diff sites	dif site bp	no. sites DE com	%	bp sites DE com	%	no. DE com	%
chr4_dfwt_3	455	19,050,455	31	7	1,564,031	8	28	24
chr4_dfwt_4	337	12,330,337	26	8	1,399,026	11	26	22
chr4_dfwt_5	344	13,336,344	26	8	1,369,026	10	30	25
chr4_dfwt_6	643	36,054,643	52	8	3,779,052	10	51	43
chr4_dfwt_7	625	36,953,625	44	7	3,219,044	9	55	47
chr4_dfwt_8	314	16,229,314	23	7	1,360,023	8	26	22
chr4_dfwt_9	505	22,735,505	43	9	2,668,043	12	46	39
chr4_dfwt_10	161	5,654,161	16	10	1,226,016	22	20	17
chr4_dfwt_all	1112	91,905,112	82	7	14,337,082	16	100	85
Total DE comb genes	118							
Region	diff sites	dif site bp	no. sites genes	%	bp sites genes	%	no. genes	%
chr4_dfwt_3	455	19,050,455	362	80	16,586,362	87	671	22
chr4_dfwt_4	337	12,330,337	225	67	10,212,225	83	458	15
chr4_dfwt_5	344	13,336,344	259	75	10,805,259	81	440	15
chr4_dfwt_6	643	36,054,643	483	75	30,042,483	83	950	32
chr4_dfwt_7	625	36,953,625	446	71	30,916,446	84	963	32
chr4_dfwt_8	314	16,229,314	222	71	12,614,222	78	416	14
chr4_dfwt_9	505	22,735,505	391	77	19,242,391	85	725	24
chr4_dfwt_10	161	5,654,161	106	66	4,982,106	88	196	7
chr4_dfwt_all	1112	91,905,112	734	66	81,130,734	88	2011	67
Total chr4 genes	3014							

**Table 5.13 *del<sup>B16</sup>* differentially interacting regions overlap for viewpoints 3-10 with DE C57B16/J alleles, DE combined genes, and total annotated genes in chromosome 4**

Column notations are as described in Table 5.4 in this chapter.

Region	diff sites	dif site bp	no. DE Bl6	MC simulations	p-val
chr4_dfwtBl6_3	455	19,050,455	9	405	0.405
chr4_dfwtBl6_4	337	12,330,337	11	37	0.037
chr4_dfwtBl6_5	344	13,336,344	10	70	0.07
chr4_dfwtBl6_6	643	36,054,643	16	309	0.309
chr4_dfwtBl6_7	625	36,953,625	18	161	0.161
chr4_dfwtBl6_8	314	16,229,314	10	108	0.108
chr4_dfwtBl6_9	505	22,735,505	14	90	0.09
chr4_dfwtBl6_10	161	5,654,161	8	9	0.009
chr4_dfwtBl6_all	1112	91,905,112	34	309	0.309
Total DE Bl6 genes	39				
Region	diff sites	dif site bp	no. DE com	MC simulations	p-val
chr4_dfwtBl6_3	455	19,050,455	33	111	0.111
chr4_dfwtBl6_4	337	12,330,337	30	6	0.006
chr4_dfwtBl6_5	344	13,336,344	30	13	0.013
chr4_dfwtBl6_6	643	36,054,643	62	12	0.012
chr4_dfwtBl6_7	625	36,953,625	61	22	0.022
chr4_dfwtBl6_8	314	16,229,314	29	52	0.052
chr4_dfwtBl6_9	505	22,735,505	49	1	0.001
chr4_dfwtBl6_10	161	5,654,161	21	1	0.001
chr4_dfwtBl6_all	1112	91,905,112	115	75	0.075
Total DE comb genes	118				
Region	diff sites	dif site bp	no. genes	MC simulations	p-val
chr4_dfwtBl6_3	455	19,050,455	735	0	0
chr4_dfwtBl6_4	337	12,330,337	485	8	0.008
chr4_dfwtBl6_5	344	13,336,344	489	15	0.015
chr4_dfwtBl6_6	643	36,054,643	1083	33	0.033
chr4_dfwtBl6_7	625	36,953,625	1077	59	0.059
chr4_dfwtBl6_8	314	16,229,314	459	425	0.425
chr4_dfwtBl6_9	505	22,735,505	806	3	0.003
chr4_dfwtBl6_10	161	5,654,161	207	110	0.11
chr4_dfwtBl6_all	1112	91,905,112	2290	357	0.357
Total chr4 genes	3014				

**Table 5.14 MC simulations to assess the significance of *del<sup>Bl6</sup>* differentially interacting regions for viewpoints 3-10 and their overlaps with DE C57Bl6/J alleles, DE combined genes, and total annotated genes in chromosome 4**

Number of features in column 4 may differ from Table 5.11 as we count total number of intersections between both datasets for this analysis. p-vals of 0 in table are <0.001, rounded down.

## 5.4 DE $df/+^{Bl6}$ genes and Monosomy 1p36

The  $df/+^{Bl6}$  mouse genotype is homologous to the heterozygous Monosomy 1p36 deletions in human. Such deletions frequently occur *de novo* and tend to have different sizes and positions (Redon *et al.*, 2005; Heilstedt *et al.*, 2003; Rosenfeld *et al.*, 2010; reviewed in Zaveri *et al.*, 2014). Because different genes are affected upon deletion, Monosomy 1p36 clinical features are varied, and include developmental delay, growth abnormalities (microcephaly, obesity), craniofacial dysmorphism (deep set eyes, midface hypoplasia, ear asymmetry, pointed chin, orofacial clefting, prominent forehead), hearing loss, and variable ophthalmological anomalies (reviewed in Slavotinek, Shaffer, and Shapira, 1999). Cardiovascular and cardiomyopathy malformations have also been reported (reviewed in Zaveri *et al.*, 2014). Very interestingly, a case of two patients presenting similar clinical features and different deletion sizes and positions was reported (Redon *et al.*, 2005), which suggests that Monosomy 1p36 could be a syndrome where deletions, besides altering gene dosage, could have positional effects.

To further explore this hypothesis, I examined the associations between Monosomy 1p36 candidate genes, and their corresponding changes in gene expression and chromatin interaction data in  $df/+^{Bl6}$  MEFs. A list of candidate genes associated with different Monosomy 1p36 phenotypes is shown in Table 5.15, together with their corresponding mouse homologues, their RNA-Seq derived expression in  $df/+^{Bl6}$  MEFs, and their overlaps with  $del^{129}$  differentially interacting regions. With the exception of *Prdm16* and *Pdprn*, all Monosomy 1p36 candidate genes fall within differentially interacting  $del^{129}$  regions in mouse. Moreover, gene *Ece1* (endothelin-converting enzyme 1), outside of the deletion CNV, shows a decrease in expression in  $df/+^{Bl6}$  MEFs, potentially constituting an example of the

positional effects that the deletion could exert upon neighboring gene expression.

To know whether a decrease in *ECE1* gene expression is observed after the occurrence of a deletion in 1p36 in humans, I decided to perform an RT-qPCR analysis of *ECE1* mRNA levels in limoblastoid cell lines derived from Monosomy 1p36 patients [Supp. Table 6.1]. As can be seen in Fig.5.6A,B, *ECE1* mRNA levels in Monosomy 1p36 derived cell lines are consistently lower compared to the karyotypically normal controls, therefore reproducing the observations made in mouse of reduced *Ece1* expression after the occurrence of a 4.3Mb deletion. *ECE1* has been suggested to be a strong candidate gene involved in the generation of cardiovascular defects in Monosomy 1p36 patients. Evidence for this role came from a single patient with a heterozygous loss-of-function mutation in *ECE1* (Hofstra *et al.*, 1999). The patient displayed patent ductus arteriosus, a small subaortic ventricular septal defect, and a small atrial septal defect. Very interestingly, heart defects are also observed in *Ece1*-null mice (Yanagisawa *et al.*, 1998, 1998). Based on this evidence, it has been suggested that haploinsufficiency of *ECE1* could potentially be involved in the generation of cardiovascular malformations in Monosomy 1p36 patients (Zaveri *et al.*, 2014).

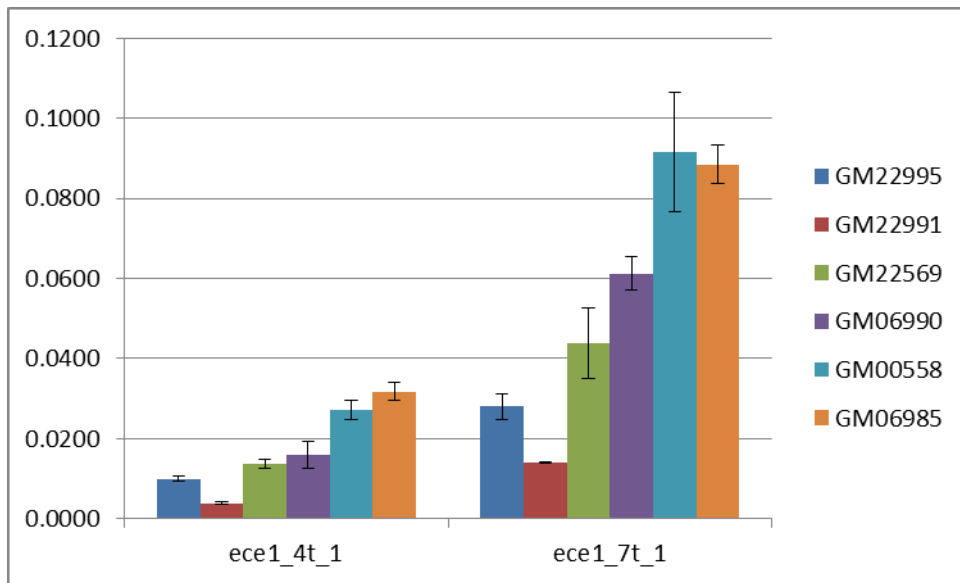
One can hypothesize that the reduced *ECE1* gene expression in Monosomy 1p36 cell lines and our mouse model could be one of the positional effects that CNVs can have, with putative roles in disease phenotypes. However, whether the reduction in *Ece1* expression in mouse and humans is a product of altered chromatin structure arising after CNV occurrence, will have to be further investigated.

Phenotype	Human gene name	Ensembl Mouse ID	chr	Gene start	Gene end	DiffReg Start	DiffReg End	Viewpoints	Direction of change	DE	log2foldchange
Cardiovascular	WASF2	ENSMUSG00000028868	chr4	132,686,420	132,755,671	132,694,000 132,754,000	132,737,000 132,770,000	1 1	1	no	
Cardiovascular	LUZP1	ENSMUSG00000001089	chr4	136,025,676	136,110,695	136,018,000	136,148,000	2,11	-1	no	
Cardiovascular	HSPG2	ENSMUSG00000028763	chr4	137,024,684	137,126,545	137,021,000 137,081,000	137,055,000 137,132,000	12 1,2	1,-1	no	
Cardiovascular	ECE1	ENSMUSG000000057530	chr4	137,418,152	137,521,144	137,363,000	137,444,000	2	-1	yes	0.78
Cardiovascular	CLCNKA	ENSMUSG000000033770	chr4	140,940,525	140,954,639	140,870,000	141,251,000	1x2,2x4,11x3,12x3	1,-1	no	
Cardiovascular	SPEN	ENSMUSG000000040761	chr4	141,023,805	141,094,512	140,870,000	141,251,000	1x2,2x4,11x3,12x3	1,-1	no	
Cardiovascular	PDPN	ENSMUSG000000028583	chr4	142,857,334	142,889,467					no	
Cardiomyopathy	MASP2	ENSMUSG000000028979	chr4	147,976,663	147,989,608	147,921,000	147,998,000	11	-1	no	
Cardiomyopathy	UBE4B	ENSMUSG000000028960	chr4	148,702,525	148,800,858	148,706,000	148,716,000	1	-1	no	
Cardiovascular,Cardiomyopathy	RERE	ENSMUSG000000039852	chr4	149,655,755	149,996,075	149,576,000 149,733,000 149,815,000	149,700,000 149,803,000 149,895,000	1,2,12 2,12 1,11	1,-1	no	
Cardiomyopathy	PRDM16	ENSMUSG000000039410	chr4	153,690,234	154,010,982			inside CNV		yes	1.43
Cardiovascular,Cardiomyopathy	SKI	ENSMUSG000000029050	chr4	154,528,184	154,596,701	154,526,000	154,573,000	11	1,-1	no	
Dysmorphism, Neurologic	PRKCZ	ENSMUSG000000029053	chr4	154,634,238	154,735,470	154,578,000	154,780,000	1x2,2,11x2,12		no	
Cardiomyopathy, Neurologic	GABRD	ENSMUSG000000029054	chr4	154,759,089	154,772,221	154,578,000	154,780,000	1x2,2,11x2,12	1,-1	no	
Seizures	GABRD	ENSMUSG000000029054	chr4	154,759,089	154,772,221	154,578,000	154,780,000	1x2,2,11x2,12	1,-1	no	
Cardiovascular	DVL1	ENSMUSG000000029071	chr4	155,221,511	155,233,412	155,227,000	155,336,000	11	-1	no	

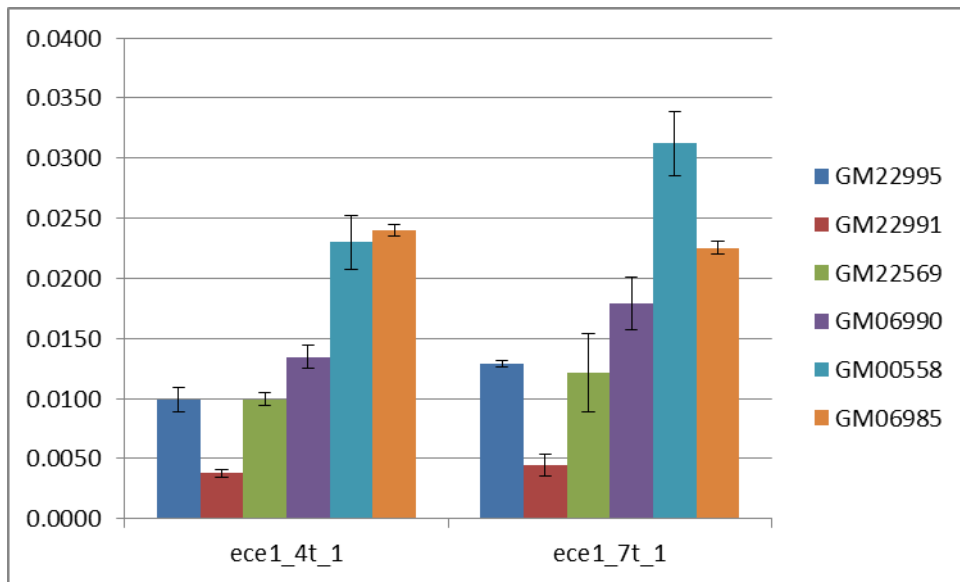
**Table 5.15 Candidate genes associated with different Monosomy 1p36 phenotypes**

Their corresponding mouse homologues are shown in column 3, together with their chromosomal positions (columns 5,6). Their overlaps with *del*<sup>129</sup> differentially interacting regions are displayed in columns 7-9, as well as the direction of change of the contact probabilities (1= increase, -1=decrease. Both compared to *wt*<sup>129</sup>). RNA-Seq derived expression in *df/+*<sup>Bl6</sup> MEFs is shown in columns 11,12. Mouse gene coordinates are expressed in mm9 assembly, while human is GRCh38.

A)



B)





**Figure 5.6 Graph of ECE1 mRNA levels in Monosomy 1p36 derived cell lines and normal karyotypic controls**

*y* axis measures CT values as measured by RT-qPCR. *x* axis represents samples. First 3 columns represent Monosomy 1p36 derived lymphoblastoid cell lines per primer pair used, next 3 columns represent controls. Controls: GM06990, GM00558, GM06985. Monosomy cell lines used: GM22995, GM22991, GM22569. A) Repeat #1 of RT-qPCR reaction. B) Repeat #2 of RT-qPCR reaction. Note the agreement between results of both experiments. Primers used assessing 4 transcripts: F: 5' AGTACAGCAACTACAGCGT 3', R: 5' TTCTGGTAAGCCCGATAGG 3'. 7 transcripts: F: 5' CCTATTGTGGTCTATGACAAGGA 3'. R: 5' GTTGTTGAGCAGGCATCTG 3'.

## 5.5 Summary of RNA-Seq characterizations of $df/+^{Bl6}$ and $+^{129}/+^{Bl6}$ MEFs

A significant number of gene expression changes occur for  $df/+^{Bl6}$  compared to  $+^{129}/+^{Bl6}$  MEFs. Of the 5495 expressed genes that passed our filtering criteria, 1345 were significantly differentially expressed between  $df/+^{Bl6}$  and  $+^{129}/+^{Bl6}$  genotypes across both alleles. 796 genes were higher expressed and 549 genes were lower expressed compared to  $+^{129}/+^{Bl6}$ . 28 of the 51 annotated genes in the deleted region were DE.

No dosage compensation events were detected for the genes contained within the CNV in  $df/+^{Bl6}$ , and although 79% of DE C57Bl6/J alleles overlap differentially interacting regions for viewpoints located inside the CNV in  $del^{Bl6}$ , there is no selective enrichment for overlaps between these regions and DE C57Bl6/J alleles and DE combined genes ( $p > 0.05$ ). This observation suggests that functional contributions from  $del^{Bl6}$  differentially interacting regions in the transcriptional regulation of these genes may be limited or non-existent.

After accounting for strain-specific mapping bias, 189 DE genes do not show unequal changes in expression levels between alleles, suggesting they are typically coordinately regulated (75% of 129S5/SvEv<sup>Brd</sup> alleles, and ~58% of C57Bl6/J). The genes that did showed unequal allelic levels of expression change included genes located inside the deleted region on chromosome 4 ( $FDR < 0.01$ ). The other genes that were not located in the manipulated region were regulated in an undefined and unbalanced allele-specific manner. These genes did not localize to specific regions of the genome and were not significantly enriched for functional terms.

Despite manipulation of only one haplotype (129S5/SvEv<sup>Brd</sup>), DE genes were strongly correlated between their allelic fold change values. This is indicative of *trans*

effects, where mRNA levels are regulated similarly between the alleles through the induction of a transcription factor, chromatin remodeler, or other modulators in the nuclear environment. This observation, together with the shared ~12Mb of differentially interacting regions between *del<sup>I29</sup>* and *del<sup>Bl6</sup>* chromosomes, suggests the participation of a *trans* mechanism which may affect both features. In fact, RNA-Seq results have shown the existence of significant enrichment of DE combined genes contained within differentially interacting regions in both *del<sup>I29</sup>* and *del<sup>Bl6</sup>*. 44% of DE combined genes are contained within *del<sup>I29</sup>* regions, while 54% of DE combined genes are contained within *del<sup>Bl6</sup>*. These overlap ratios are highly significant (MC simulations,  $p < 0.05$ ). In addition, both *del<sup>I29</sup>* and *del<sup>Bl6</sup>* showed enrichment in overlaps with H3K4me1 and H3K27ac, histone marks associated with poised and active enhancers, respectively, which may explain some of the changes in gene expression (for example, by altering preferred promoter-enhancer interactions). However, no obvious associations exist between the magnitude and direction (increase/decrease) of DE log2fold and differential interaction changes (Spearman rank correlation test,  $p\text{-val} > 0.05$ ). These observations suggest that although the deletion CNV modifies the local chromatin structure, especially the contacts established by enhancer and other regulatory elements, and that these may have local effects on gene transcription, *trans* effects may be largely responsible for regulating quantitative expression differences given the extensive DE genes present not only in the *df* (*del<sup>I29</sup>*) chromosome, but also in its wild-type copy (*del<sup>Bl6</sup>*).

Very interestingly, CTCF gene expression is increased in *df/+<sup>Bl6</sup>* MEFs (0.5 log2fold change) compared to *+<sup>I29</sup>/+<sup>Bl6</sup>*. Similarly, Gene Ontology (GO) analyses into cellular function for *df/+<sup>Bl6</sup>* MEFs revealed 26 genes associated with “condensed nuclear chromosome” [Supp. Table 5.10, 5.11]. Genes such as centromere protein E (*CenpE*),

regulator of chromosome condensation 1 (*Rcc1*), structural maintenance of chromosomes 3 (*Smc3*), among others, are higher expressed in *df/+<sup>Bl6</sup>* MEFs, with possible important consequences in chromosome architecture in this genotype (see Chapter 6 for an extensive discussion on the topic).

Given the long list of DE *df/+<sup>Bl6</sup>* genes involved in different aspects of chromosome architecture, identification of the mechanism leading to changes in chromatin interactions after deletion may be a difficult challenge. The complex organizational state of a chromosome may depend not only on a certain class of proteins enriched at differentially interacting regions, but could be an interplay of diverse components. Therefore, even after the observation of altered gene expression in several genes falling within or flanking differentially interacting regions, teasing out the association between architectural and transcriptional signatures still requires further investigation. This will be particularly important for the study of potential effects of CNVs on the long-range control of gene expression, such as the one observed for *Ece1* gene in our mouse datasets. The reproducibility of mouse *Ece1* downregulation in human Monosomy 1p36 cell lines points to future exciting new studies combining chromatin architecture and gene expression, with important consequences in disease studies.

## Chapter 6: Conclusion and Perspectives

### 6.1. Summary

Identification of allele-specific chromatin interactions was performed for 12 PE-4Cseq viewpoints within and around a 4E2 4.3Mb deletion in heterozygote ( $df/+^{Bl6}$ ) and WT ( $+^{I29}/+^{Bl6}$ ) MEFs. A quantitative framework for the analysis of multi-viewpoint PE-4Cseq data was developed, which allowed the detection of changes in chromatin interactions at levels higher than expected purely from their altered genomic proximity (i.e. shortening of the chromatin fiber).

Up to 22% of chromosome 4 sequence display changes in contact probabilities between the deletion ( $del^{I29}$ ) and WT ( $wt^{I29}$ ) chromosomes. Several long-range interactions across the deletion region were augmented, while a marked chromatin decompaction was detected towards the telomeric end of chromosome 4 (downstream of the deletion). 4 differentially interacting regions plus a constitutive control interaction were verified through 3D DNA FISH experiments, where a strong agreement was observed with the change trends detected by PE-4Cseq.

Interestingly, a high degree of overlap exists in differentially interacting regions between  $del^{I29}$  and the WT copy of chromosome 4 in  $df/+^{Bl6}$  MEFs ( $del^{Bl6}$ ). Up to ~33% of the  $del^{I29}$  regions are shared with  $del^{Bl6}$ , while  $del^{Bl6}$  shares ~50% of its differentially interacting regions with  $del^{I29}$ . Both  $del^{I29}$  and  $del^{Bl6}$  differentially interacting regions are enriched for CTCF and Smc1 protein binding, suggesting that shared changes in chromatin interactions altered after deletion could be controlled by changes in CTCF/cohesin transcription, upstream binding regulators, transcription factors, chromatin remodelers, or other proteins.

A significant number of gene expression changes occur for  $df/+^{Bl6}$  compared to  $+^{129}/+^{Bl6}$  MEFs. 1345 genes were significantly DE between  $df/+^{Bl6}$  and  $+^{129}/+^{Bl6}$  genotypes across 129S5/SvEv<sup>Brd</sup> and C57Bl6/J alleles. 28 of the 51 annotated genes in the deleted region were DE. No dosage compensation events were detected for the genes contained within the CNV in  $df/+^{Bl6}$  MEFs. DE genes were enriched in GO terms related to cell cycle (P=1.87e-86), cell and nuclear division (P=< 5.27e-62), DNA replication (P=1.50e-44), and chromosome organization (P=2.30e-26), among others [Supp. Table 5.11]. 189 DE genes showed a high correlation in expression level changes between 129S5/SvEv<sup>Brd</sup> and C57Bl6/J alleles, suggesting they are typically coordinately regulated by a *trans* mechanism.

There was a significant enrichment of DE combined genes contained within differentially interacting regions in both  $del^{129}$  and  $del^{Bl6}$ . 44% of DE combined genes are contained within  $del^{129}$  regions, while 54% of DE combined genes are contained within  $del^{Bl6}$ . These overlap ratios are highly significant (p<0.05).  $del^{129}$  and  $del^{Bl6}$  differentially interacting regions showed enrichment in overlaps with H3K4me1 (poised) and H3K27ac (active) histone marks associated with enhancers, putatively altering preferred promoter-enhancer interactions and causing local changes in gene expression. However, no obvious associations exist between the magnitude and direction (increase/decrease) of DE log2fold and differential interaction changes (Spearman rank correlation test, p >0.05).

The characterization of chromatin interactions upon the occurrence of a 4.3Mb deletion in mouse chromosome 4 revealed yet another aspect of the impact of CNVs present in the genome, that is, their potential impact on chromosome organization. Although many of the  $del^{129}$  chromatin interaction changes could be explained by *trans* mechanisms affecting both chromosome 4 copies, there exist 659 regions (~23Mb) not shared with  $del^{Bl6}$ , pointing

to possible direct effects of CNVs in the underlying chromosome architecture by the alteration of the length of the chromatin fiber and its physical units of organization. Given the extensive number of changes in contact probabilities along chromosome 4, proteins such as chromatin remodelers, architectural proteins, or transcription factors are hypothesized to be involved in the generation of such changes. Such hypothesis would require further validations and the use of a comprehensive C technique for the evaluation of the impact of CNVs on chromatin structure.

## 6.1. Discussion

CNVs are known to affect gene expression in *cis* over large genomic distances (Stranger *et al.*, 2007; Merla *et al.*, 2006; Ricard *et al.*, 2010). These observations led to the hypothesis that CNVs have a complex effect on gene transcription that might involve altered long-range chromatin organization.

Theoretically, the mere duplication or deletion of a chromatin segment could disrupt associations between gene promoters and enhancers, disturb the positioning of regulatory element networks, or fuse differentially regulated chromatin regions [see Fig. 1.8 in Chapter 1]. All of these events, triggered by a CNV, could have many important functional and pathological implications.

Understanding changes in chromatin architecture upon copy-number variation is important to expand on the current knowledge of chromosome conformation, its alteration upon sequence disruption, and its functional impact on cellular transcriptional status. For this reason, the purpose of my thesis research was to characterize in detail the higher-order chromatin organization of a genomic region in its diploid state and upon the occurrence of

CNVs.

The selection of a region for this analysis was of prime importance. One can argue that the impact of a CNV on the underlying chromatin organization of a genomic region may vary depending on the region analyzed. For example, changes in chromosome conformation in a transcriptionally silent region will differ from an actively transcribed one. I therefore concentrated in the analysis of a genomic segment associated with recurrent recombination.

For this purpose, I selected the mouse 4E2 region for CNV-chromatin organization studies. Mouse 4E2 band is syntenic to human 1p36. 1p36 deletions are a relatively common chromosome abnormality (Heilstedt *et al.*, 2003; Bagchi and Mills, 2008, and references therein), often present in a wide variety of cancers (reviewed in Bagchi and Mills, 2008), and often result in a mental retardation syndrome known as “Monosomy 1p36” (reviewed in Slavotinek, Shaffer, and Shapira, 1999). Given that Monosomy 1p36 patients are heterozygous for this region (Heilstedt *et al.*, 2003), and heterozygous deletions in 1p36 are associated with cancer progression/maintenance (Bagchi and Mills, 2008, and references therein), there is a compelling need for the correct identification of the altered chromosome from its WT homologue to study CNVs in a functionally relevant scenario.

A 4E2 4.3Mb deletion and duplication (*df/dp*) mouse strain had been previously engineered in 129S5/SvEv<sup>Brd</sup>-derived ES cells for the study of 1p36 tumor suppressors (Bagchi *et al.*, 2007). Such a model provided the best available material for the study of CNVs and chromatin organization, given the information on the precise location of the CNVs, the previous phenotypic characterizations for the heterozygous progeny of the engineered chromosomes (Bagchi *et al.*, 2007), and the availability of thousands of genotyping SNPs which could distinguish the CNV chromosomes from WT C57Bl6/J in



$df/+^{Bl6}$  and  $+^{129}/+^{Bl6}$  F1 MEFs.  $dp/+^{Bl6}$  MEFs were not further studied for numerous technical reasons as well as the potential inclusion of cells within the population which had lost the duplication after recombination, therefore emulating WT cell behaviors (Chapters 2 and 3).

PE-4Cseq was selected as the technique for the analysis of allele-specific chromatin interactions. I had originally planned the use of the 5C technique for the study of chromosome conformation upon the occurrence of CNVs. The 5C technique is based on the ligation of primers bordering interacting segments in 3C templates (Dostie *et al.*, 2006), and it was the first method used to reveal the TAD organization in mammalian cells (Baù *et al.*, 2011; Nora *et al.*, 2012). However,  $df/df$  and  $dp/dp$  genotypes are lethal (Bagchi *et al.*, 2007). Because 5C is not able to detect SNPs given its contact amplification strategy, the idea was dropped, and PE-4Cseq used instead. This methodology, modified from the standard 4C-Seq technique, uses PE sequencing for the amplification of the interacting partners of a region of interest, together with a genotyping SNP (Holwerda *et al.*, 2013; de Wit *et al.*, 2013) (see Chapter 1). Although PE-4Cseq does not provide the contact probability matrices for all restriction fragments present in a specific region (and therefore does not give information into the specific TAD folding), PE-4Cseq data extends to the whole chromosome in *cis*, which allows the evaluation of long-range chromatin interactions.

The study of chromatin organization upon CNVs offered a different and unique challenge in terms of PE-4Cseq data analysis. This challenge lies in the fact of teasing apart genuine changes in chromatin contacts from those derived from background probability profiles [see Fig. 4 in Chapter 4]. To address this problem, a new 4Cseq analysis approach was developed by Swagatam Mukhopadhyay, CSHL, for the differential analysis of contact probability signal across multiple PE-4Cseq viewpoints and genotypes.

This pipeline, grounded on polymer physics, corrects for several data biases common to 3C-derived methodologies and others specific to PE-4Cseq. It reports genuine changes in chromatin interactions by comparisons to background contact probability profiles calculated from PE-4Cseq data. The use of this modeling approach allowed quantitative viewpoint comparisons to resolve differentially interacting regions across chromosomes from *df/+<sup>Bl6</sup>* and *+<sup>129</sup>/+<sup>Bl6</sup>* MEFs.

Up to 22% of chromosome 4 sequence display changes in contact probabilities between the *del<sup>129</sup>* and *wt<sup>129</sup>* chromosomes (Chapter 4). Several long-range interactions across the deletion region were augmented at levels higher than expected purely from their altered genomic proximity, while the deletion downstream regions showed a generalized decrease in interactions with their surrounding sequences [Fig. 4.6A,B in Chapter 4]. I verified four of these changes through 3D DNA FISH experiments. Notably, a strong agreement between the change trends for both experimental modalities was found [Fig. 4.8A in Chapter 4], constituting the first time such a correlation is ever shown between PE-4Cseq and 3D DNA FISH data. The validation of results reported by the new PE-4Cseq analysis pipeline was an important step for the advance of this project and the analysis of functional correlations.

Notably, the CNV caused an overall reduction in compaction downstream of the deletion in the *del<sup>129</sup>* chromosome, towards its telomeric end. Decompaction of this region would be caused by a higher transcriptional output from the genes contained within this segment, which may cause the chromatin to be in a more open state. However, no increase in gene expression was detected for this region in *df/+<sup>Bl6</sup>* MEFs [Fig. 5.1A,B in Chapter 5], invalidating this hypothesis. Another possible explanation for the observed decompaction

would be that CNV-neighboring regions harbor tethering points which could cause the intervening chromatin to extend upon the occurrence of the 4.3Mb deletion. Such tethering points may well be constituted by LADs, lamina-associated domains, important features of nuclear architecture and genomic regulation (Pickersgill *et al.*, 2006; Guelen *et al.*, 2008; Peric-Hupkes *et al.*, 2010). No major LAD associations were found on regions surrounding the CNV (Wu and Yao, 2013). However, a major 1Mb segment encompassing numerous LADs is contained within the CNV, potentially serving as a tethering point of the 4E2 band. Subsequent experiments using BAC probes inside this LAD-rich segment could be used to study whether associations with the nuclear periphery or other nuclear features exist for this region, and also address whether further upstream LAD sequences participate in the changes in compaction. Additional studies may target specific nuclear bodies and heterochromatin foci serving as tethering points in addition to LADs.

Interestingly, there was a high degree of overlap between  $del^{I29}$  and  $del^{Bl6}$  differentially interacting regions. Up to ~33% of the  $del^{I29}$  regions are shared with  $del^{Bl6}$ , while  $del^{Bl6}$  shares ~50% of its regions with  $del^{I29}$ . This is equivalent to ~12Mb of shared differentially interacting regions, constituting ~7.7% of chromosome 4 length. After excluding  $del^{Bl6}$ -derived segments from the dataset, there are 659 unique  $del^{I29}$  differentially interacting regions with a mean size of 35Kb, covering ~23Mb (~15%) of chromosome 4. Accordingly, there exist 521 differentially interacting regions that are unique to  $del^{Bl6}$ , with a mean size of 30Kb and covering ~15Mb of sequence (~10% of chromosome 4 length). The high overlap ratio between  $del^{I29}$  and  $del^{Bl6}$  PE-4Cseq data suggests global mechanisms of chromatin architecture regulation which are common to both homologous chromosomes.

This hypothesis is further strengthened by the observed changes in gene expression.

Allele-specific RNA-seq analysis of  $df/+^{Bl6}$  and  $+^{129}/+^{Bl6}$  MEFs revealed a high degree of correlation between 129S5/SvEv<sup>Brd</sup> and C57Bl6/J alleles [Fig. 5.2 in chapter 5]. This is indicative of *trans* effects, where mRNA levels are regulated similarly between the alleles through the induction of a transcription factor, chromatin remodeler, or other modulators in the nuclear environment. Overall, there is an enrichment of DE combined genes falling inside  $del^{129}$  and  $del^{Bl6}$  shared and unique differentially interacting regions (Chapter 5). 44% of DE combined genes are contained within  $del^{129}$  regions, 54% of DE combined genes are contained within  $del^{Bl6}$ , and 22% of DE combined genes are shared by both  $del^{129}$  and  $del^{Bl6}$ . Interestingly,  $del^{129}$  and  $del^{Bl6}$  differentially interacting regions are enriched in overlaps with H3K4me1 and H3K27ac, histone marks associated with poised and active enhancers, respectively. The possible alteration of preferred promoter-enhancer interactions in the *df* chromosome could explain some of its observed changes in gene expression. However, further investigation will need to be performed to assess whether this hypothesis is true. In my current analysis, no obvious associations exist between the magnitudes of DE log2fold and differential interaction changes. Therefore, even after the observation of altered gene expression in several genes falling within or flanking differentially interacting regions, teasing out the association between architectural and transcriptional signatures still requires further investigation.

Both  $del^{129}$  and  $del^{Bl6}$  shared and unique differentially interacting regions are enriched for CTCF and Smc1 protein binding, suggesting that shared changes in chromatin interactions altered after deletion could be controlled by changes in the transcription of these architectural proteins, or possibly by upstream binding regulators, transcription factors, chromatin remodelers, or other proteins.

Very interestingly, CTCF gene expression is increased in  $df/+^{Bl6}$  MEFs (0.5 log2fold change) compared to  $+^{129}/+^{Bl6}$ . Similarly, Gene Ontology (GO) analyses into cellular function for  $df/+^{Bl6}$  MEFs revealed 26 genes associated with “condensed nuclear chromosome” [Supp. Table 5.10, 5.11]. Genes such as *CenpE* (located in mouse chr3), essential for the maintenance of chromosomal stability through efficient stabilization of microtubule capture at kinetochores (Schaar *et al.*, 1997; Wood *et al.*, 1997; Yao *et al.*, 2000), *Rcc1* (located in mouse chr4), involved in the regulation of onset of chromosome condensation in the S phase (reviewed in Hadjebi *et al.*, 2008), *Smc3* (in mouse chr19), a component of the multimeric cohesin complex that holds together sister chromatids during mitosis and define TAD boundaries (Guacci *et al.* 1997; Michaelis *et al.* 1997; Losada *et al.* 1998; Kagey *et al.*, 2010; Nora *et al.*, 2012; Phillips-Cremins *et al.*, 2013; Seitan *et al.*, 2013; Sofueva *et al.*, 2013; Zuin *et al.*, 2013), among others, are higher expressed in  $df/+^{Bl6}$  MEFs.

The increased expression of these proteins in  $df/+^{Bl6}$  MEFs could have important consequences in chromosome architecture for this genotype. In an attempt to study the effects of chromatin remodelers prior to the availability of RNA-Seq data, I had previously derived 4C templates from *Chd5* KO/ $+^{Bl6}$  MEFs (provided by Alea Mills, CSHL). *Chd5* is a chromatin remodeler located inside the CNV region (Quan and Yusufzai, 2014; Li *et al.*, 2014). 2 PE-4Cseq viewpoints (148.9 and 154.9) were amplified from a single biological replicate. However, *Chd5* DE is not readily detected in our datasets, probably due to its filtering given the low number of reads obtained in both  $df/+^{Bl6}$  and  $+^{129}/+^{Bl6}$  MEFs. Additionally, the developed 4Cseq analysis pipeline requires at least 4 viewpoints for the quantitative analysis of contact probabilities. However, after the detection of DE for numerous genes associated with different aspects of chromosome conformation, elucidating a

mechanism by which these changes occur would need the study of not a single chromatin remodeler, but of the several proteins whose transcription was affected after the occurrence of the 4.3Mb deletion in chromosome 4. The complex organizational state of a chromosome may depend not only on a certain class of proteins enriched at differentially interacting regions, but could be an interplay among diverse components.

It is important to note that all of the presented information in this thesis regarding *del*<sup>129</sup> differentially interacting regions is derived from the analysis of only 4 viewpoints surrounding the deletion (Chapter 4). Only 4 viewpoints were enough to show that up to 22% of chromosome 4 sequence have altered contact probabilities with these selected regions, pointing to the existence of further changes in chromosome conformation, possibly related to the DE of architectural and chromosome segregation/structural proteins.

However, the existence of 659 *del*<sup>129</sup> differentially interacting regions (~23Mb) not shared with *del*<sup>B16</sup> points to possible direct effects of CNVs in the underlying *cis* chromosome architecture by the alteration of the chromatin fiber organizational units (i.e. TADs). In mouse ES cells, the 4E2 region harbors 8 TADs, while cortex data shows the existence of 3 TAD structures (Dixon *et al.*, 2012). The 4.3Mb deletion directly falls within TADs in both cell types, suggesting that TAD fusion could also happen in MEFs. It is not yet known whether the fusion of two TADs alters internal TAD structures in a way that it produces a new chromatin contact arrangement, or whether intermingling and contact differences occur only within a certain fraction of the new boundary. PE-4Cseq data derived in this project showed the local and long-range effects that this particular 4.3Mb deletion CNV caused along chromosome 4, yet knowledge on internal TAD structure is missing given PE-4Cseq different approach at evaluating chromosome conformation. Future experiments using the Hi-

C technique will be necessary to further evaluate the effects of CNVs on higher order chromatin organization, such as alteration of TAD structures.

## 6.2. Perspectives and future directions

The presented thesis work focused on the analysis of 4 PE-4Cseq viewpoints surrounding a 4.3Mb deletion in mouse chromosome 4. A comprehensive view of the changes in chromosome structure for *df/+<sup>Bl6</sup>* MEFs, or about any other heterozygote state CNV, would be provided by performing a genome-wide chromatin contact analysis such as Hi-C (Lieberman-Aiden *et al.*, 2009), together with ChIP-Seq analysis of the major architectural proteins (CTCF, cohesin) and other chromatin organization candidates (such as transcription factors and chromatin remodelers), and an allele-specific RNA-Seq characterization. In theory, the Hi-C technique is able to perform allele-specific assignments of chromosome conformation given its sequencing-based detection of interacting segments (Lieberman-Aiden *et al.*, 2009). However, the success of such assignments heavily depends on the number of SNPs distinguishing each chromosome that fall nearby restriction enzyme cutting sites. One can hypothesize that a full distinction between homologous chromosomes, in mouse, would not be possible given the presence of highly conserved genes which have not undergone high rates of mutational changes, for example, the *Oct4* gene (Medvedev *et al.*, 2008, and references therein).

Upon the improvement of allele-specific detection of chromosome conformation in a genome-wide manner, integration of all 3 genome-wide techniques in study models such as the ones used for this project would provide a comprehensive measure of the global impact a CNV can have not only in gene expression, but also in chromatin organization in *cis* and

*trans*. Although PE-4Cseq data derived for this project also gave a measure of *trans* interactions for the analyzed viewpoints, the analysis solely focused in *cis* (intra-chromosomal) changes in chromatin contacts given the straightforward interpretation and analysis of such regions after the development of the polymer physics analysis pipeline. Modeling inter-chromosomal interactions requires additional assumptions about the nuclear distributions and random chromosome collisions, for which I do not have the required data. The inclusion of such information is out of the scope of this project, however, after the corresponding publication of the *cis* interaction results, special emphasis will be placed on the discovery of interesting inter-chromosomal interactions which would have a potential functional impact, especially for cancer and Monosomy 1p36 research (see below).

An important analysis which would help elucidate the contribution of DE genes to the changes observed in chromatin interactions in the *del*<sup>129</sup> chromosome is one based on systems biology. Inside the 4.3Mb deletion CNV there exist 51 annotated RefSeq genes. The targeted study of the networks in which these genes participate could give a list of potentially affected genes upon deletion of this region. Such analyses can provide some insight into the *cis* gene expression changes not explained by the affected gene networks from the CNV deletion. In fact, such analyses are being performed in other types of RNA-Seq and CNV datasets by the laboratory of Dana Pe'er at the Departments of Biological Sciences and Computer Science in Columbia University. The data produced in this project not only offers RNA-Seq information, but the changes in chromatin contacts for several regions along chromosome 4. A collaboration with her lab could be established, or the data made publicly available upon publication of this research for the analysis by interested systems biology groups.



Currently, I have focused on the study of the detected mouse chromatin interaction and gene expression changes in human Monosomy 1p36 samples. The  $df/+^{B16}$  mouse genotype is homologous to the heterozygous Monosomy 1p36 deletions in human. Such deletions frequently occur *de novo* and tend to have different sizes and positions (Redon *et al.*, 2005; Heilstedt *et al.*, 2003; Rosenfeld *et al.*, 2010; reviewed in Zaveri *et al.*, 2014). Very interestingly, a case of two patients presenting similar clinical features and different deletion sizes and positions was reported (Redon *et al.*, 2005), which suggests that Monosomy 1p36 could be a syndrome where deletions, besides altering gene dosage, could have positional effects.

I observed that, with the exception of 2 genes, all homologous Monosomy 1p36 candidate genes fall within  $del^{129}$  differentially interacting regions in mouse. Moreover, the gene *Ece1*, outside of the deletion CNV, shows a decrease in expression in  $df/+^{B16}$  MEFs, which I was able to see also in Monosomy 1p36 human cell lines. *Ece1* changes in gene expression therefore potentially constitute an example of the positional effects that the deletion could exert upon neighboring gene expression.

Ongoing experiments in Monosomy 1p36 lymphoblast cell lines include 3D DNA FISH experiments to survey for changes in chromatin contacts between the *ECE1* gene region and the corresponding sequences of mouse viewpoints 2 and 12, which showed a 30-40% decrease in contact probabilities. Additionally, I will test if the observed decompaction phenotype downstream of the deletion is observed in human cells. Such an experiment would provide an answer as to whether the decompaction is produced by changes in tethering points, given that the syntenic region in humans is not located towards the telomere, but upstream of the equivalent deletion coordinates in chromosome 1. While further studies are

needed to assess the participation on CNVs positional effects in the generation of Monosomy 1p36 phenotypes (and its distinction from multigenic traits that may give rise to this disease's clinical features), the future translation of observations from mouse to human cases of chromosome deletions in their corresponding syntenic regions would be one of the most exciting results, given the potential implications for human disease studies.

This project provided one of the initial studies of chromatin architecture and copy-number variation. Further integrative studies will expand our understanding of changes in chromatin architecture upon recombination, and the intrinsic interplay between gene expression and the determination of chromosome structure, not only for this model, but for any studied CNV in their heterozygote and homozygote states.

## Chapter 7: Experimental methods

### 7.1 Generation of F1 +<sup>129</sup>/+<sup>Bl6</sup> and *df*/+<sup>Bl6</sup> Embryos

The *df/dp* mouse strain described in Bagchi *et al.*, 2007, was re-established by injecting the chromosomally-engineered *D4Mit190-D4Mit51* *df/dp* ES cell line into C57BL/6J blastocysts and put into surrogate mothers. F1 progeny from segregating chimeras X C57BL/6J crosses were genotyped by PCR on tail-derived DNA. *df* genotyping was performed using primer pairs #2137 (*df* FWD): 5' – CCTCATGGACTAATTATGGAC – 3' and #2138 (*df* REV): 5' – CCAGTTTCACTAATGACACA – 3', using the following PCR conditions: 94°C for 4 minutes, followed by 40 cycles of 30 s at 94°C, 1:15min at 53°C, and 2:30min at 70°C, and a final cycle of 5 min at 70°C. 25µl genotyping reactions were made using 2.5 µl 10X PCR buffer, 0.5 µl of dNTPs (10mM, New England BioLabs), 1.25 µl of each primer (10µM), 1 µl of DNA (20-50ng), 1.25 µl of DMSO, 17.1 µl of dH<sub>2</sub>O, and 0.15 µl of TaqPolymerase (AmpliTaq DNA polymerase, Applied Biosystems, 5U/µl). PCR product is approximately 2.2Kb in size. *dp* genotyping was performed using primer pairs #1991 (*dp* FWD): 5' – CGGTAGAATTTTCGAGGTCGCTAG - 3' and #1992 (*dp*REV): 5' – GCCCAAGCTGATCCGGAACCC – 3', using the following PCR conditions: 94°C for 4 minutes, followed by 40 cycles of 30 s at 94°C, 1 min at 63°C, and 2:30min at 70°C, and a final cycle of 5 min at 70°C. 25µl genotyping reactions were made using 2.5 µl 10X PCR buffer, 0.5 µl of dNTPs (10mM), 1.25 µl of each primer (10µM), 1 µl of DNA (20-50ng), 18.35 µl of dH<sub>2</sub>O, and 0.15 µl of TaqPolymerase (AmpliTaq DNA polymerase, Applied Biosystems, 5U/µl). PCR product is approximately 800bp in size.

## 7.2 MEF Preparation and Cell Culture

Embryos of *df/dp* male chimeras X C57BL/6J crosses were dissected at 13.5 days after plug observation. Heads were removed for DNA extraction and genotyping, and the bodies minced by passing through syringes with 18G 1/2 and 20G 1/2 needles and plated on 10cm dishes previously coated with 0.1% gelatin and using for culture Dulbecco's Modified Eagle Medium (DMEM) High Glucose (4.5g/L) supplemented with 10% Fetal Bovine Serum (FBS) (v/v), 50U/ml Penicillin G, and 100µg/ml Streptomycin sulfate. Cells were incubated on 5% CO<sub>2</sub> at 37°C, and passaged every 2-3 days depending on confluency and growth rate. All experiments were subsequently performed on MEF plates 10 hours after reaching confluency at passage 4.

## 7.3 3D DNA FISH

3D DNA FISH was performed as described in Solovei and Cremer, 2010. In summary, ~60% confluent MEF #1.5 22mm acid free coverslips were prepared by fixing cells in 4% PFA/PBS for 10min at RT. During the last minute, 2 drops of 0.5% Tx100/PBS were added. Coverslips were then washed 3 times in PBS for 5 min at RT. Nuclei were permeabilized by incubating coverslips in 0.5% Tx100/PBS for 10 min at RT, washed in PBS 3 times for 5 min, and incubated with 0.1mg/ml RNaseA/PBS for 30 min at 37°C. After washing coverslips 3 times in PBS for 5 min at RT, these were transferred twice to coplin jars with freshly-made 20% Glycerol/PBS. Coverslips were incubated overnight at 4°C in 20% Glycerol/PBS, and subsequently submerged into liquid nitrogen, frozen and thawed for a total of 5 times, soaking with 20% Glycerol/PBS between each freeze/thaw cycle. Cells were

washed 3 times in PBS for 5 min at RT, briefly rinsed in 0.1N HCl, and then incubated in fresh 0.1N HCl for 10 mins. Finally, coverslips were washed 3 times in PBS for 5 min at RT, equilibrated in 2x SSC for 5 min, and incubated in 50% Formamide/2x SSC for 30 min. Prepared coverslips were stored at 4°C until further used.

Hybridization mixes were prepared by nick translating isolated BAC DNA. Reactions include: x  $\mu$ l (2 $\mu$ g) of maxi-prep BAC DNA, 22-x  $\mu$ l nuclease free water, 2.5 $\mu$ l 0.2mM labeled dUTP (green: Alexa 488, red: Alexa 594; Cy-5: Alexa 647, Life Technologies), 5 $\mu$ l 0.1mM dTTP (Roche), 10 $\mu$ l dNTP mix (0.1mM, New England BioLabs), 5 $\mu$ l 10X nick translation buffer, and 5 $\mu$ l nick translation enzyme (Abbott Molecular Inc.). Reactions were incubated at 15°C for 10hrs, heat inactivated at 70°C for 10 min, and cooled down to 4°C. Reactions were transferred to 1.5ml tubes and mixed with 1 $\mu$ l of 0.5M EDTA, 1 $\mu$ l of linear acrylamide (Ambion), 5  $\mu$ l of 3M NaOAc (pH 5.2), and 125  $\mu$ l of 100% EtOH (-20°C cold), and incubated overnight at -20°C. Samples were then centrifuged at 20,000g for 1 hr at 4°C. At this point the colored pellet should be visible. Pellets were cleaned by adding 1ml of 75% EtOH, centrifuged at 14,000RPM for 5 min, repeating the last two steps, drying the pellet in 37°C incubator for 15 min, and dissolving in 50  $\mu$ l of DEPC-treated water by vortexing at 37°C for 1 hr. Hybridization mixes were made by combining 3 $\mu$ l nick-translated probe with 5  $\mu$ l mouse Cot1 DNA, 5  $\mu$ l yeast tRNA, and 5  $\mu$ l ssDNA, and lyophilized in Speed-Vac for ~20min. Hybridization buffer was made using 4XSSC, 20% dextran sulfate, and dH<sub>2</sub>O, mixed together and kept in the heating block at 37°C. Lyophilized probes were resuspended in 10 $\mu$ l formamide (Ambion) and kept shaking in the heating block at 37°C for at least 30min. 10 $\mu$ l of hybridization buffer were added to the 10 $\mu$ l fluorescent probes, and the mix was loaded onto clean glass slides. The prepared coverslips were mounted cell-side down

onto the hybridization mix, sealed with rubber cement, and kept in the dark until dry. Sealed slides were put onto 75°C heat block for exactly 3min, and hybridized overnight at 42°C in humid chamber. Post-hybridization washes include: twice in 50% formamide/2x SSC for 10min at 42°C (water bath), twice in 2x SSC for 10min at 42°C (shaking), twice in 1x SSC for 10min at 42°C (shaking). Coverslips were equilibrated in 4x SSC for 3min at RT, stained with DAPI/4x SSC for 3min, rinsed in 4x SSC, and mounted on clean microscope slides. Coverslips were sealed with nail polish, and imaged using an Applied Precision DeltaVision Core wide-field fluorescence microscope system (GE Healthcare, Issaquah, WA) equipped with a PlanApo 60x 1.40 numerical aperture objective lens (Olympus America).

#### **7.4 MEF Karyotyping**

Spectral karyotyping (SKY) analysis of all mouse chromosomes was performed on the 129S5E71 and 129S5E117 MEF samples using the protocol described in Padilla-Nash et al, 2006. The SKY protocol is composed of various steps, including the initial preparation of metaphase chromosome spreads, slide pre-treatment and probe denaturation, probe hybridization, detection, image acquisition, and final analysis. Due to the length of the protocol, the various critical steps needed, and the troubleshooting advice provided, the reader is referred to the Padilla-Nash *et al.*, 2006 protocol for careful in-depth knowledge of the procedure performed for this study.

## 7.5 4C Template Preparation

4C templates for *df/+<sup>Bl6</sup>* (MEF lines 129S5E71 and 129S5E98) and *+<sup>129</sup>/+<sup>Bl6</sup>* (MEF lines 129S5E117 and 129S5E118) were prepared as described in Splinter et al, 2012. Briefly, 1x10<sup>7</sup> MEFs were cross-linked for 10 min at RT using 2% formaldehyde (Calbiochem) and 10% FCS in PBS (pH 7.4). 10 ml reactions were transferred to ice and added 1.425ml of 1M glycine, followed by centrifugation for 8 min at 225g at 4°C. Supernatant was subsequently removed and the resulting cell pellet resuspended in 500 µl of ice cold nuclei buffer (10mM Tris pH 7.6, 10mM NaCl, 2mM MgCl<sub>2</sub>, dH<sub>2</sub>O) containing protease inhibitors (Roche) for 10 min on ice. An equal volume of nuclei buffer/0.5% NP-40 was added to the tube and incubated for 5 min on ice. Samples were vortexed for 10 s and centrifuged for 1 min at 1,000g and 4°C. Cells were washed once in nuclei buffer/0.5% NP-40 containing protease inhibitors and centrifuged for 1 min at 1,000g and 4°C.

Pellets were resuspended in 450µl dH<sub>2</sub>O and 60µl 10X restriction buffer (buffer 2 supplied with *Hind*III enzyme, New England BioLabs), incubated 1 hr with 15µl 10% SDS shaking at 900RPM at 37°C, and followed by an additional 1 hr incubation with 75µl 20% Triton X-100. 5 µl aliquots were taken as undigested controls and stored at 4°C. Samples were subsequently digested by adding 800U of *Hind*III (New England BioLabs) and incubating overnight at 37°C while shaking. 5 µl aliquots were taken as digested controls and de-crosslinked by incubation with 10 µl Proteinase K (10mg/ml, Roche) in 90 µl of 10 mM Tris (pH 7.5) at 65 °C for 1 h. Digestion efficiencies were estimated based on the pattern of smear of the undigested and digested controls by running 20 µl of decrosslinked sample on a 0.6% agarose gel. If digestion was sufficient, *Hind*III was inactivated by incubating the sample for 20 min at 65°C (shaking gently). The digested nuclei were transferred to a 50 ml

falcon tube and mixed with 5.7ml dH<sub>2</sub>O, 700 µl 10X Ligase Buffer, and 50U T4 Ligase (Roche), and incubated overnight at 16°C. Ligation efficiency was determined by taking 100 µl of ligation reaction and incubating 1 hr at 65°C with 5 µl Proteinase K (10 mg/ml). When run in a 0.6% agarose gel, ligated DNA should appear as a single upper band similar to the undigested control. If ligation occurred, DNA crosslinks were reversed by adding 30 µl of 10 mg/ml Proteinase K and incubation at 65 °C overnight. Subsequently, 15 µl of 20 mg/ml PureLink RNase A (Invitrogen) was added and the reactions incubated for 45 min at 37 °C, followed by phenol extraction and DNA purification as described in Splinter et al, 2012. The DNA pellet was dissolved in 150 µl of 10 mM Tris (pH 7.5), and digested overnight with 50U *DpnII* (New England BioLabs) at 37°C while shaking. An aliquot of 5 µl was taken from the *DpnII* reaction and mixed with 95 µl of 10 mM Tris (pH 7.5), and 20 µl loaded into a 0.6% agarose gel to assess digestion efficiency. If sufficient digestion was achieved, *DpnII* was heat inactivated by incubating 20 minutes at 65°C, and DNA was ligated at low concentrations (12.1ml dH<sub>2</sub>O, 1.3ml 10X ligation buffer, 100U T4 DNA Ligase) overnight at 16°C. DNA was phenol extracted and ethanol precipitated with glycogen (Roche) as a carrier. The resulting 4C templates were purified using QIAquick PCR purification kit columns (Qiagen), dissolved in 10 mM Tris (pH 7.5), and stored at -20°C.

## **7.6 PE-4CSeq Viewpoint Amplifications, Sequencing, and Reads Mapping**

Inverse 4C amplification primers were designed per viewpoint following standard rules for PCR primer design, and checking alignment uniqueness to the desired fragment as compared to the rest of the genome. Primers used in this study are listed in Supplemental



Table 4.3. Additionally, amplification primers for all viewpoints were added the PE1 and PE2 Illumina paired-end primers plus a 1-2 nucleotide barcode in their 5' ends for HiSeq PEx100 sequencing.

Each of the 14 viewpoints was amplified from the available *df/+<sup>Bl6</sup>* and *+<sup>129</sup>/+<sup>Bl6</sup>* 4C templates in reactions using 3.2µg 4C template, 16µl dNTP (10mM, New England BioLabs), 24µl reading primer PE1 of a 1µg/µl primer stock, 24µl reading primer PE2 of a 1µg/µl primer stock, 11.2 µl Expand Long Template polymerase (Roche), 80µl 10X PCR buffer 1 (supplied with polymerase), and dH<sub>2</sub>O until completing 800µl total. This volume is then mixed and separated into 16x50µl PCR reactions, and run using the following program: 94 °C for 2 min, followed by 30 cycles of 15 s at 94 °C, 1 min at 55 °C and 3 min at 68 °C, and one final step of 5min at 68 °C. PCR reactions were subsequently collected and pooled together, and purified using the High Pure PCR Product Purification Kit (Roche) for viewpoints amplified on 129S5E71 and 129S5E117 4C templates, or using AMPure beads (Beckman Coulter) with a 0.9X volume ratio for viewpoints amplified from 129S5E98 and 129S5E118 4C templates. Equimolar amounts of isolated captured viewpoints were pooled together using the KK4824 kit to correct for insert size lengths (Kapa Biosystems). Pooled libraries were sequenced using two lanes of HiSeq PE100.

Obtained reads were separated using custom perl scripts based on the sample and genotyping SNP on PE1 reads. PE2 reads were trimmed to 30bp to have the highest quality bases for captures mapping. Alignments were performed using bowtie against a reduced database of sequences bordering *HindIII* restriction sites in mm9. Up to 3 mismatches were accepted per read to account for SNPs in the 129S5/SvEv<sup>Brd</sup> sequence, and only uniquely mapped reads were taken into account for the 4C data analysis.

## 7.7 Polymer Physics Analysis of PE-4CSeq data

### 7.7.1 Model for bias correction

Our goal in interpreting 4C data is two-fold: Firstly, to recover as closely as possible the underlying contact probability indirectly measured in the 4C experiments. Doing so requires correcting for experimental biases and translating the counts of viewpoint-interaction-partners (called “captures” from here on) in each experiment into estimates of physical contact probabilities. Secondly, learning the local chromatin compactness and signatures of conformational changes arising from large-scale chromatin deletion. Both of these goals require constructing a null model of chromatin. The null model sets the expectation on random non-specific contacts as a function of genomic separation. In turn, the statistical significance of specific interactions is judged against the profile of such non-specific interactions.

The typical separation between viewpoints in our experiments range from 300Kb to 1.5Mb. The persistence length—the length of polymer beyond which it is floppy and behaves like a random walk— is roughly 2.5Kb-3.5Kb. Therefore, the length-scales of separation of viewpoints is much larger than then persistence length— polymer physics dictates a scaling form for contact probability at such length-scales,

$$(1) \quad P_{ij} \sim N_{ij}^{\nu}$$

Where  $P_{ij}$  is the probability of contact between viewpoint  $i$  and  $j$ ,  $N_{ij}$  is their separation along the chromatin, and  $\nu$  is the scaling exponent. For example, for a non-

interacting Gaussian polymer (3D random walk)  $\nu = -3/2$ . Measurements of this exponent in Hi-C data in the hundred Kb to hundred Mb range of length-scales has yielded an exponent of approximately  $\nu = -1$  for mammalian cells, implying that the chromatin polymer in such cells is more compact than a 3D random walk. The compaction of chromatin is not uniform along the genome—the exponent can depend on the viewpoint position and is a signature of local chromatin compaction. We do observe such dependence in our data, see Results. Large-scale genomic deletions affect chromatin compaction as measured by this exponent. We propose a null model of the polymer where each viewpoint is an effective unit connected by springs. The probability of separation of two effective units is assumed to be a Gaussian function. Nevertheless, the spring constants connecting these units enjoy the observed non-Gaussian scaling with respect to genomic separation. To be specific, the probability of  $M$  such effective units denoted by the set  $\{i\}$  to be at positions  $\{x_i\}$  is given by

$$(2) \quad \mathcal{P}(\{\mathbf{x}_i\}) = \prod_{i=1}^{M-1} \left( \frac{\kappa_{i,i+1}}{2\pi} \right)^{3/2} \text{Exp} \left[ -\frac{\kappa_{i,i+1} (\mathbf{x}_i - \mathbf{x}_{i+1})^2}{2} \right]$$

Where  $\kappa_{i,i+1}$  is the local “spring constant” and  $P(\{x_i\})$  is the probability density of conformation  $\{x_i\}$ . For random  $-1$  walk,  $\kappa_{i,i+1}$  scales as  $N_{i,i+1}$  where  $N_{i,i+1}$  is the genomic separation between neighboring units  $i$  and  $i+1$ . Equivalently,  $-3/2$  the probability of contact scales as  $N_{i,i+1}$ , yielding the Gaussian value for exponent  $\nu$ . In our model however, each  $\kappa_{i,i+1}$  is allowed a regional scaling exponent. In the polymer literature, such a model is called the Gaussian approximation to a non-Gaussian polymer. The contact probability between the effective units of the polymer obtained from such a Gaussian approximation is

not exact; the error has been discussed extensively in the polymer physics literature in the context of self-avoiding polymers, which are less compact compared to a random walk. The error is owing to rather subtle reasons, but roughly the approximation underestimates the number of conformations of a self-avoiding polymer. For a polymer typically more compact than a random-walk, the model overestimates the number of polymer conformations. However, the level of noise in the estimation of the regional  $v$  and the statistical uncertainties in long-range measurements far outweighs the error introduced by such a Gaussian approximation. In fact, no “multi-C” datasets currently warrant a more complex model. The advantage of the Gaussian approximation is that the model becomes exactly computable. In previous works, the dependence of the capture data on the fragment lengths was classified as a bias. We consider it to be a genuine effect; fragments of the chromatin are expected to have number of potential contact points in proportion to their lengths. Therefore, we normalize the capture data by the product of the viewpoint and fragment lengths. For each viewpoint, we compute the local scaling  $v_i$  from a spline fit (in log-log space) of the normalized capture data against their genomic distances. The smoothed spline is observed to be roughly linear in the 10Kb to 1 Mb range. This normalization and fit yields our biased contact frequency  $F_{ij}$  between viewpoint  $i$  and  $j$ . We now discuss our modeling of biases. Our bias model attempts to be general by accounting for both known and unknown bias sources. To this end, we assign each viewpoint fragment an unknown bias factor  $C_i$ . The PE-4CSeq capture data for each viewpoint is assigned another bias factor  $K_i$ . Therefore, the observed contact frequency  $F_{ij}$  is modeled as

$$(3) \quad F_{ij} = C_i C_j K_i P_{ij}$$

where  $P_{ij}$  is the true and yet unknown contact probability. Note that  $P_{ij} = P_{ji}$  by definition but observed  $F_{ij}$  is typically not equal to  $F_{ji}$  owing to different biases in each viewpoint capture data. Note that  $P_{ij}$  in the scaling regime, see (1), is only known up to an overall constant prefactor which cannot be determined from the data alone. Similarly,  $F_i$  and  $K_i$  are also determined in their ratios. However, these indeterminacies pose no problem in bias correction. The bias corrected capture data is qualitatively comparable across experiments but their units of measure are not meaningful.

The contact probability of distant fragments is determined by the Gaussian model given by (2). The variance of (2) separation between neighbors is

$$\sigma_{i,i+1}^2 \equiv 1/k_{i,i+1}$$

therefore, the net variance of spring constant between distant fragments  $i$  and  $j$  in is given by

$$\sum_{m=i}^{m=j-1} \sigma_{m,m+1}^2$$

yielding

$$(4) \quad P_{ij} = \frac{1}{\left(2\pi \sum_{m=i}^{m=j-1} \frac{1}{k_{m,m+1}}\right)^{3/2}}$$

In general, the bias correction algorithm needs to learn both the biases  $C$ 's and  $K$ 's, and the local spring constants  $k_{m,m+1}$ . In PE-4CSeq experiments, successive viewpoints which are of the order of megabasepair apart, therefore (3), the local variation of spring

constants in the intervening region cannot be estimated. In Hi-C data, the bias and the spring constants can be learned simultaneously because all mutual contacts of neighboring fragments are recorded. In the present context, we assume that spring constants  $K_{i,i+1} = 1/N^{-2/3}$ , consistent with  $P_{ij} \sim N_{ij}$  observed in mammalian cells. This is the  $P_{ij}$  is use to learn the  $C$ 's and  $K$ 's. A linear set of equations in logarithm space in the unknowns  $\log C_i$  and  $\log K_i$ 's and the knowns  $\log F_{ij}$ ,  $\log F_{ji}$  and  $\log P_{ij}$  is solved by least square method to compute the  $C$ 's and  $K$ 's. We perform bias correction only for nearest neighbors. Nevertheless, the bias is reduced for all neighbors, as judged by reduced asymmetry in bias-corrected capture data for  $(i, j)$  and  $(j, i)$  pair of viewpoints. The bias correction method should be robust to noise in the capture. Such noise may be modeled as Poisson process. We show robustness of our method by analyzing recovery on simulated  $P_{ij}$  corrupted by Poisson noise and multiplicative bias [Supp. Fig. 4.2A,B,C,D].

### 7.7.1 Comparison of bias-corrected capture data

The bias-corrected capture data has a resolution of the typical fragment sizes (1-10Kbp). It is a noisy signal along the genome reflecting the underlying contact probability, per unit length, of fragment-viewpoint pairs. Though the uncorrected capture data profiles are widely different in the deletion and the WT strain, after bias correction we observe that the smoothed capture data profiles are nearly indistinguishable in large section of the chromosome in the two strains. This adds confidence in our method and allows us to report region specific differences of  $>10\%$  in capture data in the two strains. In order compare the profiles of wild type and deletion capture data, we smoothen the signal by a Gaussian kernel of widths of 20Kb for region specific comparisons.

## 7.8 Allele-specific RNA-Sequencing and Analysis

RNA from seven independent primary MEF lines was isolated ( $+^{129}/+^{Bl6}$ : 129S5E88, 129S5E90, 129S5E95;  $df/+^{Bl6}$ : 129S5E36, 129S5E56, 129S5E71, 129S5E98) using TRIzol reagent (Ambion) and polyA+ RNA was isolated (Oligotex kit; QIAGEN). Stranded libraries were prepared using a protocol adapted from Parkhomchuk *et al.*, 2009, for paired-end sequencing on the Illumina HiSeq platform. PEx100 reads were separately aligned to both the C57BL/6J and 129S5/SvEv<sup>Brd</sup> transcriptomes. We used the C57BL/6J transcriptome as downloaded from Ensembl gene set version 72 (assembly version: mm10). The 129S5/SvEv<sup>Brd</sup> transcriptome was constructed by modifying the C57BL/6J transcriptome using SNPs and indels calls from Keane *et al.*, 2011. Where multiple transcripts exist for a gene, we selected the longest transcript as the representative transcript for the gene in the transcriptome.

We used the GSNAP alignment algorithm with the parameter of no mismatches (-m 0) (Wu and Nacu, 2010). Reads were filtered to keep only those with one best mapping location. To obtain estimates of expression values, we only counted those reads aligning at a gene location if both reads of a paired-end set were mapped to the same gene. To avoid biological interpretation from mapping noise, we excluded genes with less than 10 reads mapping to each allele if this occurs across genotypes. Differential expression analyses were performed using the R Bioconductor package – DESeq (Anders and Huber, 2010), using an FDR cut-off of 0.05. We performed non-allele-specific differential expression analyses (pairwise between WT and treatments) using counts summed from both alleles. Allele-specific analysis were performed only using reads that mapped to the transcriptome of each

strain and compared in a pairwise manner, that is, between  $+^{129}/+^{Bl6}$  samples (C57BL/6J x 129S5/SvEv<sup>Brd</sup>) and  $df/+^{Bl6}$  samples. To account for the allelic mapping biases that is a result of more reads mapping to the C57BL/6J transcriptome, we tested for changes in the proportion of reads mapping to each allele between treatment and WT groups, on a gene by gene basis, to determine whether similar degree of changes to expression levels occurred between alleles. Counts were normalized using DESeq and tests were done using the R function, `prop.test`, using median counts across replicates and p-values were adjusted for multiple testing in R using the `fdr` method (adjusted p-value cut-off = 0.01). The software GREAT was used for functional term enrichment analysis with single gene associations (McLean et al., 2010) as well as WEB-based GENE SeT AnaLysis Toolkit with hypergeometric tests and Bonferroni corrections (Zhang, Kirov, Snoddy. 2005). Locations were mapped to the mm9 genome for correlation testing using UCSC LiftOver.



## **Chapter 8: Extended Materials and Methods**

### **8.1 Protocols, Buffers, and Cell Culture media recipes**

#### **8.1.1 Mouse Tail DNA Isolation**

- In a ventilated hood, cut a small piece of the end of the tail from the mouse into a sterile 1.5ml epptube. Use a sterile scalpel for each tail, and gloves.
- Add 500  $\mu$ l lysis buffer + 5  $\mu$ l Proteinase K stock solution (stock = 40 $\mu$ g/ $\mu$ l) to each tail tube and mix homogeneously.
- Incubate overnight in heating block at 56°C with shaking (500RPM)
- Vortex the lysate and spin 10 min max speed in an Eppendorf microcentrifuge. Mark new epptubes while spinning.
- Transfer the supernatant to a new 1.5 ml epptube.
- Take one sample at a time and add 500  $\mu$ l isopropanol (RT) and rock the tube until DNA precipitates.
- Spin tubes 10 min max speed in an Eppendorf microcentrifuge.
- Discard supernatant and wash pellet twice with 70% ethanol and once with 99% ethanol.
- Let the DNA pellet air dry for 10 mins.
- Add 300  $\mu$ l sterile 1xTE or DNase-free water and let the DNA dissolve overnight at 4°C.
- Store DNA at -20°C.

Lysis buffer

<i>Final concentration</i>	<i>Stock</i>	<i>250ml</i>	<i>25 ml</i>
50mM Tris pH 7.5	1M Tris pH 7.5	12.5 ml	1.25 ml
0.1 M EDTA pH 8.0	0.5 M EDTA	50 ml	5 ml
0.1 M NaCl	5 M NaCl	5 ml	0.5 ml
1% SDS	10% SDS	25 ml	2.5 ml
H <sub>2</sub> O		157.5 ml	15.75 ml

Proteinase K

Roche Proteinase K PCR grade, catalogue number 0311 587 9001

Stock solution 40µg/µl. Dissolve proteinase K in DEPC-treated water.

### 8.1.2 *df* PCR genotyping

Model: *df* – 4.3Mb deletion in mouse chromosome 4

Source: A. Bagchi (Bagchi *et al.*, 2007).

#### *PCR reaction:*

10X PCR buffer	2.5 $\mu$ l
dNTPs (10mM)	0.5 $\mu$ l
#2137 primer (10 $\mu$ M)	1.25 $\mu$ l
#2138 primer (10 $\mu$ M)	1.25 $\mu$ l
DNA (20-50ng)	1 $\mu$ l
DMSO	1.25 $\mu$ l
dH <sub>2</sub> O	17.1 $\mu$ l
TaqPolymerase	0.15 $\mu$ l

#### *Cycles:*

94°C - 4:00

94°C - 0:30 ---

53°C - 1:15 | 40X

70°C - 2:30 ---

70°C - 5:00

4°C - infinite

*Primers:*

#2137 (df FWD): 5' – CCTCATGGACTAATTATGGAC - 3'

#2138 (df REV): 5' – CCAGTTTCACTAATGACACA - 3'

PCR product is approximately 2.2Kb in size. DNA polymerase used: Applied Biosystems

AmpliTaq DNA Polymerase + Mg<sup>2+</sup>, catalogue number N8080-152

### 8.1.3 *dp* PCR genotyping

Model: *dp* – 4.3Mb duplication in mouse chromosome 4

Source: A. Bagchi (Bagchi *et al.*, 2007).

#### *PCR reaction:*

10X PCR buffer	2.5 µl
dNTPs (10mM)	0.5 µl
#1991 primer (10µM)	1.25 µl
#1992 primer (10µM)	1.25 µl
DNA (10ng)	1 µl
dH <sub>2</sub> O	18.35 µl
TaqPolymerase	0.15 µl

#### *Cycles:*

94°C - 4:00

94°C - 0:30 ---

63°C - 1:00 | 40X

70°C - 2:30 ---

70°C - 5:00

4°C - infinite

#### *Primers:*

#1991 (*dp* FWD): 5' – CGGTAGAATTTTCGAGGTCGCTAG - 3'

#1992 (dpREV): 5' – GCCCAAGCTGATCCGGAACCC - 3'

PCR product is approximately 800bp in size. DNA polymerase used: Applied Biosystems

AmpliTaq DNA Polymerase + Mg<sup>2+</sup>, catalogue number N8080-152

### 8.1.4 IACUC Standard Procedure – Mouse Embryonic Fibroblasts (MEFs)

- *Anesthesia and Tail Biopsy*

- Mice should be weighed and one of the following anesthetic agents used
  - Tail biopsy Avertin (2.5%) @ with 0.015-0.017 ml/gm body weight
    - Isoflurane (Drop Method) – contact vet staff
    - Ketamine (80-120 mg/kg) and Xylazine (5 mg/kg), IP
    - Pentobarbital (50 mg/kg), IP
- Anesthesia is required for tail biopsy of mice older than 3 weeks and for all retro-orbital bleeding.
- Preparation of Avertin
- A solution of 100% Avertin is prepared by mixing 10 g of tribromoethyl alcohol with 10 ml of tertiary amyl alcohol (Sigma). Dilute 10 ml of this solution to 2.5 % in 390 ml isotonic saline (PBS), then sterilize by filtration ( 0.2 filter ) and aliquot into a series of sterile snap cap tubes. The 2.5 % stock solution is stored wrapped in foil (to protect from the light) at 4 ° C.
- The proper dose of Avertin may vary with different preparations and should be re-determined each time a new 2.5 % stock is made by conducting a dose response experiment. Briefly, inject a set of age matched mice with either 0.01, 0.015, 0.017, 0.02, or 0.025 ml/g of the new stock, monitoring completeness of anesthesia and absence of subsequent adverse side effects. The optimal dose typically proves to be around 0.015-0.017 ml/g body weight.
- When diluting the alcohol mixture with some commercially available complex

phosphate buffered saline solutions, precipitation of the tribromoethyl alcohol may occur. This is due to the presence of calcium and/or magnesium in the PBS. To avoid this, check that the PBS you use is simple sodium phosphate buffered saline (0.8 %) and does not contain calcium and/or magnesium or use the following Tris buffered saline solution:

0.8 % sodium chloride

1mM Tris, pH 7.4

0.25 mM EDTA

- If there is crystallization or a change in color of the Avertin, it must not be used.
- The use of any other anesthetic agents must be identified in the IACUC application.
- *Anesthesia Monitoring*
  - During the tail biopsy procedures the following parameters must be monitored at a minimum of 5-10 minute intervals:
    - Respiratory rate
    - Response to noxious stimulus
    - Spontaneous movement



- *Anesthesia Recovery Monitoring*

- During recovery from anesthesia, the following clinical parameters must be monitored at a minimum of 15 minute intervals until the animal is ambulatory.
  - Respiratory rate
  - Movement
  - Ability to maintain sternal recumbancy
  - It is estimated that animals will recover within 30-60 minutes postoperatively.
  - To protect the animal from hypothermia they should never be placed on metal surfaces – place animals on a water re-circulating heating blanket or wrap them in a towel (while still allowing visible monitoring) to conserve body temperature. Thermal packs can also be used.

- *Use of Aseptic Surgical Techniques*

- All instruments must be pre-sterilized by acceptable methods, including steam sterilization, Cidex™ cold sterilization or by the use of a glass bead sterilizer. Instruments must be re-sterilized between animals. When performing surgery on more than one animal, effective sterilization can be best achieved by using either a glass bead sterilizer or by pre-sterilization of multiple sets of instruments. Cidex™ cold sterilization requires 10 hours of contact time to be effective. Dipping instruments in 70% alcohol between surgeries does not achieve sterility (>30 hrs of contact time

- required) and is not an acceptable method.
- The surgical site must be covered with a sterile drape or sterile, clear surgical adhesive material. The size of the drape should be adjusted to the size of the animal so that aseptic techniques can be maintained and the animal properly monitored.
- *Tail Biopsy*
    - Transgenic founders or progeny may be identified by analysis of genomic DNA obtained from a tail biopsy. Sufficient DNA for PCR, Southern, and dot blot analysis can be obtained from a 5-10 mm fragment of the distal portion of the tail. The tail biopsy can be obtained by a trained investigator from mice under 3 weeks of age without anesthesia. If the mouse is older than 3 weeks or a larger section of tail is required, an appropriate anesthetic agent should be used.
    - Weigh and anesthetize the mouse.
    - When the animal is sufficiently anesthetized, remove 5-10 mm. of the tip of the tail using a new scalpel blade.
    - Hemostasis can be achieved using a sterile gauze pad to apply direct pressure to the wound.
  - *MOUSE EMBRYO FIBROBLASTS*
    - Male mice are housed one per cage.
    - Females ovulate once every 4-5 days, 3-5 hours after the onset of the dark cycle.

- Natural matings are set up by examining the female , if she is in estrus she can be placed with the male.
  - The next morning females are checked for plugs.
  - 13 days after observing a mucus plug, the pregnant mice are sacrificed by CO<sub>2</sub> asphyxiation.
  - Using sterile technique, extract the two uterine horns containing the embryo. Cut the uterus between each embryo to divide them into individual segments. Examine the embryo. Clamp down the neck with tweezers and cut off the head with scalpel or sharp scissors. Dissect out red organs (e.g., heart, liver, kidney).
  - Transfer the torso to a 6-cm plate with 1.0 ml trypsin. Mince the embryos with fine scissors or scalpel to approximately 1 mm<sup>3</sup> pieces. Incubate the plates in 37C incubator for 45 minutes. After the incubation, add 5 ml of growth medium into each plate, pipette up and down 15 times and transfer to flasks. This is recorded as passage 1. Split the cells once and freeze down the cells in passage 2. Usually each embryo can give out 15-18 vials of MEF cells.
  - The purpose of this protocol is to get the MEF cells from embryo
- *EARLY ENDPOINTS*
    - If animals are experiencing weight loss (15% initial body weight), have wound infections that are non-responsive to therapeutic intervention or have major surgical dehiscence, they should be immediately euthanized.

### 8.1.5 PI Staining of fixed whole cells

Protocol from: Cells: A Laboratory Manual. Volume 1: Culture and Biochemical Analysis of Cells. 1998. David Spector.

- Isolate cells and transfer to a 15ml conical tube. Check that there is a single-cell suspension. Centrifuge at 1000g for 5 minutes. Remove supernatant.
- Wash cells two times in PBS without calcium or magnesium. At last wash, count the total number of cells and record the number on the tube.
- Resuspend the pellet in approximately 500  $\mu$ l of PBS.  
*It is important that this be a good single-cell suspension at this point or cells will be fixed as clumps.*
- Add 5 ml cold ethanol. Fix at 4°C overnight.  
*Add ethanol very slowly while vortexing to prevent clumping. Cells can remain in fixative up to 3 weeks before staining.*
- Take 5 million cells into a 15 ml conical tube. Centrifuge at 1000g. Remove ethanol.
- Vortex pellet. Wash two times in 5 ml of PBS + 1% BSA or calf serum. Ethanol-fixed cells are difficult to pellet. This can be overcome by the addition of BSA or serum to the wash medium.
- Resuspend the pelleted cells in 800  $\mu$ l of PBS containing 1% BSA or 1% calf serum.
- Add 100  $\mu$ l of 10X PI solution (500 $\mu$ g/ml PI [Sigma] in 3.8x10<sup>-2</sup> M sodium citrate, pH 7.0).  
*Caution: PI (Propidium Iodide) is harmful if swallowed, inhaled, or absorbed through the skin. It is irritating to the eyes, skin, mucous membranes, and upper*

*respiratory tract. It is mutagenic and possibly carcinogenic. Wear gloves, safety glasses, and protective clothing, and always work with extreme care in the chemical hood.*

*The PI solution can be stored at room temperature wrapped in aluminum foil.*

- Add 100  $\mu$ l of boiled RNase A (10mg/ml prepared in 10mM Tris-HCl, pH 7.5), and incubate at 37°C for 30 minutes.

*If not used immediately, samples should be stored and protected from light at 4°C.*

- Analyze the fixed samples by flow cytometry.

### **8.1.6 MEF medium**

Dulbecco's Modified Eagle Medium (DMEM) High Glucose (4.5g/L) supplemented with:

10% Fetal Bovine Serum (FBS) (v/v)

50U/ml Penicillin G

100µg/ml Streptomycin sulfate

### **8.1.7 Trypsin**

0.125% Trypsin (Gibco 15090-038)

1mM EDTA

in HEPES-buffered saline

Sterile filter and store at -20°C.

### **8.1.8 HEPES-buffered saline**

Per liter

7.07 g NaCl

0.4 g KCl

0.043 g Na<sub>2</sub>HPO<sub>4</sub>

1.0 g D-glucose

4.77 g HEPES

Combine ingredients and bring volume up to 1 liter with dH<sub>2</sub>O. pH to 7.3 with NaOH. Sterile

filter and store at 4°C.

### **8.1.9 Phosphate Buffered Saline (PBS)**

Per liter:

7.07 g NaCl

0.4 g KCl

0.06 g KH<sub>2</sub>PO<sub>4</sub>

Combine ingredients and bring volume up to 1 liter with dH<sub>2</sub>O.

### 8.1.10 MEF Culture and Splitting

Recommended seeding is to split 1 frozen stock tube into 2x10cm dishes previously coated with 0.1% gelatin one hour before use. Cells are passaged every 2-3 days depending on confluency and growth rate. 10-cm dishes are typically split 1:3 or 1:4.

To split a 10-cm dish:

- Coat 10-cm plates with 0.1% gelatin one hour before use
- Wash the cells twice with PBS on the plates
- Incubate ~3 minutes in 1 ml of trypsin at 37°C
- Resuspend in 9 ml of warm MEF medium
- Transfer to a 15 ml Falcon tube, and centrifuge 3-5 minutes at 1500RPM at room temperature
- Discard the supernatant and resuspend the pellet homogeneously in 10 ml of MEF medium
- Remove gelatin excess from plates
- Seed 2.5-3.3 ml of the suspension to each 10-cm plate, and complete volume to 15 ml medium for each plate
- Rock back and forth and sideways each plate to distribute MEFs along the plate
- Incubate on 5% CO<sub>2</sub> at 37°C



### 8.1.11 Nick Translation Protocol

Reagents from Nick Translation Kit (Abbott Molecular Cat. 32-801300 or homemade reagents). Labeled dUTPs are Alexa dyes from Life Technologies.

1. Make reaction mixture:

22-x $\mu$ l water  
x $\mu$ l DNA (2 $\mu$ g total)  
2.5 $\mu$ l 0.2mM labeled dUTP  
5 $\mu$ l 0.1mM dTTP  
10 $\mu$ l dNTP mix  
5 $\mu$ l 10x nick translation buffer

2. Mix well, add 5 $\mu$ l nick translation enzyme

3. PCR reaction:

15°C 10 hours  
70°C 10 min  
hold at 4°C

4. Transfer to 1.5ml eppendorf and add:

1 $\mu$ l 0.5M EDTA  
1 $\mu$ l linear acrylamide  
5 $\mu$ l 3M NaOAc  
125 $\mu$ l 100% EtOH (ice cold)

5. Precipitate at -20°C overnight or -80°C 2 hours

6. Centrifuge max speed 1 hr 4°C

7. Remove supernatant
8. Wash 2X with 1ml 75% EtOH
9. Air dry pellet 15 min 37°C incubator
10. Resuspend 50ul H<sub>2</sub>O. Place on shaking 37°C heat block to completely resuspended.
11. Run 5ul on 2% agarose gel. Smear pattern should be between 50-400nt.
12. Store at -20°C in dark. Use 3-5ul per FISH reaction

### 8.1.12 3D DNA FISH

From: Solovei and Cremer, *Methods in Mol. Biology*, vol. 659 (Solovei and Cremer, 2010).

Coverslips were prepared one day before formaldehyde fixation. ~200,000 cells were seeded into gelatinized 10mm glass acid-free coverslips in 6-well plates. Plates were incubated under standard MEF conditions (37°C and 5% CO<sub>2</sub>). Coverslip confluency used was ~60-70% per experiment.

#### *Fixation of cells:*

- Rinse coverslips with PBS.
- Fix cells in 4% PFA/PBS, 10min, RT. During the last minute, add 2 drops of 0.5% Tx100/PBS.
- Wash coverslips in PBS, 5min, RT with 3 changes.

#### *Permeabilization of nuclei:*

- Incubate coverslips in 0.5% Tx100/PBS, 10min, RT.
- Wash in PBS, 5min, RT with 3 changes.
- Incubate with 0.1mg/ml RNaseA / PBS, 30min, 37°C.
- Wash in PBS, 5min, RT with 3 changes.
- Transfer coverslips to coplin jar with freshly-made 20% Glycerol/PBS.
- After a few moments, transfer to a new jar with fresh 20% Glycerol/PBS. Incubate for AT LEAST 1h, RT. [better: overnight at 4°C].
- Submerge coverslip in liquid nitrogen and wait until completely frozen. Place coverslip cell-side up on a paper towel. When glycerol is thawed, briefly soak in 20% Glycerol/PBS again. Repeat freeze-thaw cycle for a total of 5 times. [be careful not to

break the frozen coverslip]

- Wash cells in PBS, 5min, RT with 3 changes.
- Briefly rinse coverslips in 0.1N HCl, then incubate in fresh 0.1N HCl, 5min, RT.  
[time in HCl depends on cell type and can be extended to 10min]
- Wash cells in PBS, 5min, RT with 3 changes.
- Equilibrate cells in 2x SSC, 5min, RT. [IMPORTANT: pH of diluted SSC needs to be adjusted to less than pH 7.5 to preserve nuclear morphology!]
- Equilibrate cells in 50% Formamide / 2x SSC, 30min, RT.

*Hybridization:*

- Load probe / competitor / hybridization mix on a clean glass slide.
- Pull coverslip out of formamide solution, quickly drain excess formamide and place cell-side down on hybridization mix [DO NOT allow cells to dry!].
- Seal with rubber cement and keep in the dark at RT until cement is dried.
- Place slides onto 75°C heat block for EXACTLY 3min.
- Hybridize at 42°C overnight in humid chamber [better: 2 days].

*Post-hybridization washes:*

- Wash coverslips twice in 50% formamide / 2x SSC, 10min, 42°C (water bath).
- Wash twice in 2x SSC, 10min, 42°C (shaking).
- Wash twice in 1x SSC, 10min, 42°C (shaking).

Alternatively for higher stringency: wash twice in 0.1x SSC, 5min, 60°C (water bath).

**B)** Equilibrate in 4x SSC, 3min, RT.

**C)** Stain with DAPI / 4x SSC, 3min, RT.

**D)** Rinse in 4x SSC and mount on clean microscope slide. Seal with nail polish.

#### Buffers / Reagents:

20x SSC pH7.0, BioRad (cat. nr. 161-0775)

dilute to 4x, 2x, 1x (optional:) 0.1x in dH<sub>2</sub>O. IMPORTANT: adjust pH to 7.0-7.5 with 1N HCl. (1-2 drops per 250ml).

PureLink RNaseA, Invitrogen (cat. nr. 12091-021)

Ultra-Pure Glycerol, Invitrogen (cat. nr. 15514-011)

Deionized Formamide, Ambion (cat. nr. AM9342)

### 8.1.13 RNA isolation

Total RNA was isolated using Trizol reagent (Ambion).

1. Wash 1x10cm MEF plates with 10ml PBS
2. Resuspend 1x10cm MEF plate in 1ml Trizol (*samples can either stored at -80°C or processed immediately*)
3. Add 0.2ml of chloroform
4. Shake vigorously, and then incubate for 2-3 minutes at room temperature
5. Centrifuged at 12,000g for 15 minutes at 4°C
6. Transfer the upper aqueous phase containing RNA to a new tube
7. Add 0.5ml isopropanol, gently mix
8. Incubate at room temperature for 10 minutes
9. Centrifuged at 12,000g for 10 minutes at 4°C
10. Discard supernatant, and wash RNA pellet once in 1ml 75% ethanol
11. Air dry for 10 minutes
12. Resuspend in 10-30µl nuclease free water.
13. Incubate at 60°C for 5 minutes to completely dissolve RNA pellet
14. Transfer immediately to ice to measure RNA concentrations

RNA concentration was measured using nanodrop, only samples with OD<sub>260/280</sub> and OD<sub>260/230</sub> ratios above 1.6 were used for subsequent experiments.

## 8.1.14 RNA Sequencing Library Preparation

### Long RNA-seq protocol (paired-end stranded library)

Start with 10ug of total RNA.

PolyA+ isolation (Qiagen Oligotex kit)

\* use DEPC-treated water

\* preheat Oligotex suspension to 37 °C, mix by vortexing then keep at room temp

\* heat water bath or heating block to 70 °C, heat 400µl of buffer OEB per sample

\* ensure that buffer OBB does not have precipitates by prewarming at 37 °C for 10 min then place at room temperature.

\* perform all steps at room temperature unless otherwise indicated.

\* all centrifugation steps should be performed in a microcentrifuge tube at max speed (14,000g to 18,000g)

1. Pipet 10ug total RNA into an RNase-free 1.5ml microcentrifuge tube and adjust the volume of water to 250µl.
2. Add 250µl buffer OBB, 15µl oligotex suspension. Mix thoroughly by vortexing or flicking the tube
3. Incubate 3 min at 70°C to disrupt secondary structure
4. Remove sample from waterbath/heating block and place at room temperature for 12 min to allow hybridization between oligo dT30 and polyA tails
5. Centrifuge 2 min at 14,000-18,000g, room temperature. Collect and save the supernatant (polyA minus fraction). It doesn't matter if not all of the supernatant is

- collected.
6. Resuspend the pellet in 1ml buffer OW2 by pipetting. Make sure pellet is completely resuspended. Centrifuge 12,000g 2 min. Carefully remove supernatant.
  7. Wash again in 1ml buffer OW2. Be careful when removing supernatant, often it is necessary to remove all but ~100µl, spin down again and then remove the rest.
  8. Add 100µl preheated buffer OEB (70°C). Resuspend by pipetting, place back at 70°C for 10 seconds before centrifuging 2 min 12,000g room temp.
  9. Transfer supernatant containing polyA+ RNA to new microcentrifuge.
  10. Resuspend again with 100µl preheated buffer OEB. Add the supernatant to the polyA+ fraction.
  11. For second round of polyA+ purification repeat steps 1-10. Otherwise continue to ethanol precipitation.
  12. Spin polyA+ RNA in spin filter column for 1 min at 18,000g to remove any remaining oligotex suspension from the polyA+ RNA. Transfer flowthrough to a new tube as the Ambion tubes don't close very well.
  13. Add 1µl glycoblu, 1/10V 3M sodium acetate pH5.5, 3V 100% EtOH. Incubate - 70°C for at least 30 min
  14. Centrifuge 30 min 4°C 15,000g
  15. Wash 1x with 70% EtOH (-20°C), remove EtOH.
  16. Either airdry or put in speedvac for 4 min to remove residual EtOH which can interfere with subsequent reactions.
  17. Resuspend pellet in 10µl H<sub>2</sub>O on ice for 5 min.



## **Ribominus treatment**

\*use 10 $\mu$ l of polyA+ RNA from previous step or <10ug of total RNA.

\*set a waterbath or heat block to 70°C

1. Add to 1-10ug of RNA, 10 $\mu$ l of ribominus probe, 100 $\mu$ l hybridization buffer
2. Incubate at 70°C for 5 min to denature the RNA.
3. Cool sample slowly over 30 min by placing tube in 37°C heat block to allow sequence specific hybridization
4. Prepare beads during the incubation:
  - a. Vortex ribominus beads thoroughly, pipet 750 $\mu$ l into a sterile 1.5ml tube
  - b. Place on magnet for 1 min, remove supernatant.
  - c. Add 750 $\mu$ l sterile DEPC water, vortex, place on magnet, discard supernatant
  - d. Repeat wash with 750 $\mu$ l water
  - e. Resuspend in 750 $\mu$ l hybridization buffer and transfer 250 $\mu$ l to a new tube
  - f. Place the tube with 500 $\mu$ l on magnet for 1 min, remove supernatant and resuspend in 200 $\mu$ l hybridization buffer
  - g. Keep both tubes at 37°C until needed
5. After 37°C incubation, transfer ~120 $\mu$ l RNA-probe sample to the prepared ribominus beads (200 $\mu$ l beads). Mix well
6. Incubate 37°C for 15 min, gently mix occasionally
7. Briefly centrifuge, place on magnet for 1 min. **DO NOT DISCARD SUPERNATANT AS THIS CONTAINS THE RNA!**
8. Place the tube with 250 $\mu$ l beads on magnet 1 min, remove supernatant

9. Transfer ribominus RNA from the first tube to the second tube of beads. Mix well by pipetting.
10. Incubate 37°C for 15 min, gently mix occasionally
11. Place tube on magnetic separator for 1 min, transfer the supernatant containing ribominus RNA to a small filter column and spin at max speed for 2 min to remove any residual magnetic particles
12. Transfer flow through to a new tube
13. Add 1/10V 3M sodium acetate pH5.5, 3V 100% EtOH (glycobblue from polyA purification will still be present). Incubate -70 for at least 30 min
14. Centrifuge 30 min 4°C 15,000g
15. Wash 1x with 70% EtOH (-20°C), remove EtOH
16. Either airdry or put in speedvac for 4 min to remove residual EtOH which can interfere with subsequent reactions.
17. Resuspend pellet in 4µl H<sub>2</sub>O on ice for 5 min.

### **cDNA-1st strand synthesis**

\*Add all of the polyA+ ribominus RNA from 10ug of total RNA.

\*if have 2 or more samples make up mastermixes for all steps

1. To 4µl of RNA add:

1.6µl random primers (50ng/µl, Invitrogen)

2µl polydT20 (50uM, Invitrogen)

1µl NIST spike-ins

2. Start PCR program:

98°C 2 min

70°C 5 min

0.1deg/sec to 15°C

PAUSE

3. As soon as 15°C is reached (after ~15 min), add:

4µl Superscript III 1st strand buffer (5X, Invitrogen)

1µl 0.1M MgCl<sub>2</sub> (diluted from 1M MgCl<sub>2</sub> Ambion stock)

1µl 10mM dNTPs (Invitrogen)

2µl 0.1M DTT

1µl RNase Inhibitor (Ambion 20U/µl)

0.5µl H<sub>2</sub>O

4. The reaction total should be 17.9µl

5. Resume PCR program:

15°C 30 min

PAUSE

6. After 30 min at 15°C, pause program and add:

1.0µl actinomycin-D (120ng/µl in 10mM Tris pH7.6, dilute from 1mg/ml stock before use)

1.1µl superscript III enzyme (Invitrogen)

7. The reaction total should be 20µl

8. Resume PCR program (approx 1 hour 40 min)

0.1deg/seec to 25°C

25°C 10 min

0.1deg/sec to 42°C

42°C 45 min

0.1 deg/sec to 50°C

50°C 15 min

75°C 15 min

4°C hold

9. Bring total reaction volume to 100µl with H<sub>2</sub>O. Add 5 volumes of buffer PB

10. Add to minelute Qiagen spin column.

11. Centrifuge 1 min 10,000g

12. Wash column 1x with buffer PE

13. Centrifuge 1 min 10,000g

14. Remove flow through, centrifuge 1 min 12,000g

15. Add 16µl of EB to column, sit 1 min at room temp, spin 12,000g.

16. Elute again with 15µl EB. Pool sample (~30µl).

## **2nd strand synthesis**

\* add enzymes last in order listed in protocol to prevent RNase H activity before DNAPol is present.

\* prepare reaction on ice

1. Prepare 2nd strand mix:

2µl 5x first strand buffer (Invitrogen)

15µl 5x second strand buffer (Invitrogen)  
0.5µl 0.1M MgCl<sub>2</sub>  
1µl 0.1 M DTT  
2µl dUNTP mix (10mM each of dATP, dCTP, dGTP, dUTP)  
0.5µl *E. coli* DNA ligase (10U/µl)  
2µl *E. coli* DNA polymerase I (10U/µl)  
0.5µl RNase H (2U/µl)  
21.5µl RNase free H<sub>2</sub>O

2. Add 45µl second strand mix to 30µl of purified 1st strand reaction, bringing total reaction volume to 75µl
3. Incubate 2 hours at 16 °C, hold at 4°C in PCR machine
4. Bring total reaction volume to 100µl with H<sub>2</sub>O. Add 5 volumes of buffer PB
5. Add to minelute Qiagen spin column.
6. Centrifuge 1 min 10,000g
7. Wash column 1x with buffer PE
8. Centrifuge 1 min 10,000g
9. Remove flow through, centrifuge 1 min 12,000g
10. Add 26µl of EB to column, sit 1 min at room temp, spin 12,000g.
11. Elute again with 25µl EB. Pool sample (~50µl).
12. Save 1.5µl to run on Bioanalyzer DNA high-sensitivity chip (pre-fragmentation)

### **Fragmentation of ds cDNA using Covaris**

\* If machine is off: switch machine on, ensure chambers are filled with autoclaved DI

water. Run degas program prior to fragmenting samples (~30 min)

1. Transfer 50 $\mu$ l sample to covaris microtube using a pipette
2. Place in machine by snapping into place
3. Run program 'degas100ulsnapcap60sec'
4. Sonication takes 60 seconds
5. Run 1 $\mu$ l on DNA high-sensitivity chip (post-fragmentation). Fragmentation size should have a peak at 200-300.

### **End-Repair cDNA**

48 $\mu$ l sample

27 $\mu$ l H<sub>2</sub>O

10 $\mu$ l T4 DNA ligase buffer with 10mM ATP

4 $\mu$ l 10mM dNTP mix

5 $\mu$ l T4 DNA polymerase 3U/ $\mu$ l(NEB M0203)

1 $\mu$ l Klenow DNA polymerase 5U/ $\mu$ l(NEB M0210)

5 $\mu$ l T4 PNK 10U/ $\mu$ l(NEB M0201)

100 $\mu$ l

Incubate room temperature 30 min

Add 500 $\mu$ l PB, clean-up using Qiagen minelute columns. Elute 2 x 16 $\mu$ l

### **Addition of single A base**

32µl eluted cDNA  
5µl NEB buffer 2  
10µl 1mM dATP  
3µl Klenow fragment 3' to 5' exo -5U/µl (NEB M0212)  
50µl

Incubate 37°C, 30 min

Bring volume to 100µl with 50µl H<sub>2</sub>O, Add 500µl PB, minelute columns, Elute 1 x 19µl

### **Adapter Ligation**

19µl eluted cDNA  
25µl 2x Rapid DNA ligase buffer (Enzymatics B101)  
1µl Illumina Paired-End adapter oligo mix  
5µl DNA T4 ligase (Enzymatics 600U/µl)  
50µl

Incubate room temperature 30 min

Bring volume to 100µl with 50µl H<sub>2</sub>O, Add 500µl PB, minelute columns. Elute 1 x 15µl

### **UNG treatment**

15 $\mu$ l eluted cDNA

1.7 $\mu$ l 500mM KCl

1 $\mu$ l UNG (Roche N808-0096)

Incubate 37°C 15 min, 95°C 10 min. Hold on ice.

### **Gel purification**

Add 10 $\mu$ l of loading dye to 17.7 $\mu$ l UNG treated sample.

Run on 2% ultra-pure agarose gel for 2 hours at 90V. Use 100bp ladder and have a spare lane between samples.

Cut out 200bp band and another band at about 250bp. It is normal not to see anything on the gel, cut out gel anyway. Freeze larger slice.

Weigh out gel slice (~120g). Add 3V buffer QG, dissolve 15-20 min at 55°C.

Add 1V isopropanol. Load onto minelute column, spin through. Wash 1x 0.5ml buffer QG, 1x 0.75ml buffer PE. Dry spin x1. Eulte 2 x 15 $\mu$ l buffer EB



## PCR amplification

Use 15 $\mu$ l of eluted cDNA from gel purification. Save other 15 $\mu$ l in case PCR does not work.

15 $\mu$ l eluted cDNA

1 $\mu$ l PE primer 1.0 (100 $\mu$ M HPLC purified)

1 $\mu$ l PE primer 2.0 (100 $\mu$ M HPLC purified)

50 $\mu$ l 2x HF Phusion Mix (Finzymes)

33 $\mu$ l H<sub>2</sub>O (incase need to add more or less cDNA, can adjust this amount)

100 $\mu$ l

Cycle conditions:

98°C 1 min

98°C 10s

60°C 30s 18 cycles

72°C 30s

72°C 5 min

hold at 4°C

Add 500 $\mu$ l PB, minelute clean-up, elute 1x15 $\mu$ l

## **Gel Purification**

Add 10µl of loading dye to 15µl eluted PCR product

Run on 1% ultra-pure agarose gel for 2 hours at 90V. Use 100bp ladder and have a spare lane between samples.

Cut out band about 100pb larger than cDNA band.

Gel purify as above, elute 2x 25µl EB.

Dilute to 100µl with 50µl H<sub>2</sub>O. Add 10µl Sodium acetate, 330µl EtOH. Precipitate 30 min at -80°C or overnight at -20. Centrifuge 30 min at max speed. Wash 1x in 70% EtOH. Airdry or speedvac for 4 min. Resuspend in 25µl H<sub>2</sub>O.

## **Library Quantitation**

Run Agilent DNA high sensitivity chip. Run 2 dilutions of sample at 1:20 and 1:30. At least one of the dilutions should be between max and half max of loading peak height. Calculate the peak size (should be consistent between dilutions) and take the average of the concentration.

Dilute library to 10nM.

Send 25µl or half of library, whichever is less, to sequencers. Keep remaining library as backup.

**Reagents required – separate stocks of everything to prevent contamination!!!**

Oligotex mRNA midi kit (12 reactions)	Qiagen Cat # 70042
Glycoblue (300 reactions)	Ambion Cat # AM9515
Ribominus kit (8 reactions)	Invitrogen Cat # A10837-08
Superscript III RT (2,000U)	Invitrogen Cat # 18080-093
Random primers	Invitrogen Cat # 48190-011
Oligo-dT20 primers	Invitrogen Cat # 18418-020
NIST spike-ins	from Gingeras lab
RNAse Inhibitor	Ambion Cat # AM2690
1M MgCl <sub>2</sub>	Ambion Cat # AM9530G
10mM Tris-HCl pH7.6	Sigma Cat # T2444-100mL
Actinomycin-D (5mg)	Invitrogen 11805-017
5x second strand buffer	Invitrogen 10812-014
dUTP	Roche #11934554001
dNTPs	Roche # 11969064001
<i>E. coli</i> DNA ligase	Invitrogen Cat # 18052-019 \$39
<i>E. coli</i> DNA polymerase I	Invitrogen Cat # 18010-017 \$99.25
RNAse H	Invitrogen Cat # 18021-014 \$128
Bioanalyser high-sensitivity DNA chips	Agilent Cat # 5067-4626 \$453

Bioanalyser RNA nano chip reagents	Agilent Cat # 5067-1512 \$362
Covaris microtube snap-cap (25 tubes)	Covaris Cat #520045 \$125
T4 DNA ligase buffer with 10mM ATP	
T4 DNA polymerase (3U/ $\mu$ l)	NEB Cat # M2030
Klenow DNA polymerase (5U/ $\mu$ l)	NEB Cat # M0210
T4 PNK (10U/ $\mu$ l)	NEB Cat # M0201
NEB buffer 2	NEB
Klenow fragment 3' to 5' exo – (5U/ $\mu$ l)	NEB Cat # M0212
2x Rapid Ligation Buffer	Enzymatics B101
T4 DNA ligase (600U/ $\mu$ l)	Enzymatics 12 2012
Illumina Paired-end adapter Oligo Mix Illumina – got aliquot from Gingeras Lab	
Uracil N-Glycosylase (UNG) AmpErase	ABI N8080096
PE primer 1.0	Order HPLC purified from IDT
PE primer 2.0	Order HPLC purified from ID
2x HF phusion mix	Finzymes

### **8.1.15 cDNA synthesis**

1. cDNA synthesis is performed using 1µg total RNA
2. Add 1µl DNase I (Invitrogen), bring up volume to 15µl adding nuclease free water, and incubate 15 minutes at 25°C
3. Inactivated DNase I by adding 1µl EDTA and
4. Heat reaction to 70°C for 10 minutes.
5. Perform reverse transcriptase reaction using TaqMan RT reagents from Applied Biosystems (#N808-024), using random hexamer primers and a reaction time of 30 minutes at 48°C.

### **8.1.16 Quantitative RT-PCR**

Quantitative RT-PCR was performed using 2µl of cDNA and using primers amplifying a maximum of 300bp [Supp. Table 5.6]. SYBR green reagents were used for the reactions (Applied Biosciences). 3 biological and 3 technical replicates were used in each experiment, and values normalized to the geometric mean of at least 3 separate housekeeping genes (Chapter 5). Data was analyzed and graphed using Excel (Microsoft).

### 8.1.17 PE-4CSeq Protocol

#### *Collection of cells*

Tissue culture:

-Suspension cells: proceed to step 1.

-Adherent cells can both be formaldehyde treated (step 2-3) and scraped from the culture dish, or first collected, using e.g. trypsin, before proceeding to step 1.

- Primary tissue: For efficient fixation, a (viable) single cell preparation of the tissue of interest is required. To facilitate this process the use of collagenase and/or a cell strainer is advised, but incubation conditions have to be optimized empirically. For reference: a 14.5dpc fetal brain is dispersed by 0.00625% collagenase treatment in 250µl 10%FCS/PBS for 45min at 37deg, followed by the use of a 40µm cell strainer (BD Falcon #352340). For the disruption of tissues containing mainly non-adherend cells, like 14.5dpc fetal liver (red blood cells) or thymus (T-cells), collagenase treatment can be omitted, while including the use of the cell strainer.

#### *Fixation and cell lysis*

All the steps in this protocol are optimized for using  $1 \times 10^7$  cells.

1. Count cells and centrifuge 5 min, 280g at RT.
2. Discard the supernatant and resuspend the pellet in 10ml 2%formaldehyde/PBS/10% FCS.

3. Incubate tubes for 10 minutes at RT while tumbling.
4. Add 1.425ml 1M glycine (final concentration 0.125M), mix and put tubes immediately on ice to quench the cross-linking reaction. Directly proceed to step 5.
5. Centrifuge 8 min, 400g at 4°C and remove all the supernatant.
6. Resuspend pellet in 1 ml cold lysis buffer (50mM Tris-HCl pH7.5, 150mM NaCl, 5mM EDTA, 0.5% NP-40, 1% TX-100 and 1X Complete protease inhibitors (Roche #11245200) and incubate 10 minutes on ice.
7. Determine the efficiency of cell lysis: Mix 3µl of cells with 3µl of Methyl Green-Pyronin staining (Sigma #HT70116) on a microscope slide and overlay with a coverslip. Assess the lysis efficiency using a microscope. Cytoplasm stains pink and the nuclei stains blue/green. When cell lysis is incomplete, douncing can be applied to increase efficiency. *Note: cell lysis is an important step in the protocol, as failure of lysis can hamper digestion efficiency. Lysis conditions should be optimized based on the cell type used. For MEFs and ES cells, the Nuclei Isolation Protocol from Paola Vagnarelli, described in this section, has been tested and proven useful.*
8. Centrifuge 5 min, 750g at 4°C and carefully remove all supernatant. At this point nuclei can be stored for later use (proceed to step 9) or the protocol is continued directly (proceed to step 10).
9. Storing the nuclei at -80°C:
  - 9.1. Resuspend nuclei pellets in lysis buffer and transfer to a 1.5ml safe lock tube.
  - 9.2. Centrifuge 2 min, 540g at 4°C.
  - 9.3. Remove the supernatant, freeze the pellet in liquid nitrogen and store at -

80°C.

10. Resuspend the pellet in 450µl Milli-Q and continue with step 11.

### ***Digestion***

11. Add 60µl of 10X restriction buffer B (supplied with *HindIII*).

*Note: It is preferable to resuspend the nuclei pellet with pre-mixed 450µl Milli-Q + 60µl of 10X restriction buffer B at RT.*

12. Place the tube at 37°C and add 15 µl 10% SDS.

*Note: always use freshly prepared 10% SDS solutions. Old materials compromise the efficiency of digestion and ligation in subsequent steps.*

13. Incubate 1hr at 37°C while shaking at 900 RPM using an Eppendorf Thermomixer.

14. Add 75µl 20% Triton X-100.

*Note: always use freshly prepared 20% Triton X-100 solutions. Old materials compromise the efficiency of digestion and ligation in subsequent steps.*

- Incubate 1hr at 37°C while shaking at 900 RPM.
- Take a 5µl aliquot of the sample as the “Undigested control” and store at 4°C until used in step 21.
- Add 200U *HindIII* (Roche #11274040001); incubate 4 hrs at 37°C while shaking at 900 RPM.
- Add 400U *HindIII*; incubate O/N at 37°C while shaking at 900 RPM.
- Add 200U *HindIII*; incubate 4 hrs at 37°C while shaking at 900 RPM.
- Take a 5µl aliquot of the sample as the “Digested control”.
- Determination of the digestion efficiency:



- 21.1. Add 90µl 10mM Tris-HCl pH 7.5 to the 5µl samples from step 16 and 20.
- 21.2. Add 5µl Prot K (10 mg/ml Roche #03115836001) and incubate for 4 hours at 65°C.
- 21.3. Add 100µl Phenol-Chloroform (Sigma) to the samples and mix vigorously.
- 21.4. Spin for 10 min, 16400g at RT.
- 21.5. Transfer water phase to a clean tube and load ~ 20µl on a 0.6% agarose gel.  
  
Alternatively, Q-PCR analysis can be used for more precise determination digestion efficiency using multiple primer sets spanning a restriction site. This step is highly recommended when 4C is applied for the first time.
- 21.6. If digestion is OK proceed with step 22, otherwise repeat step 18, 20 and 21.

### ***Ligation***

- Heat-inactivate the restriction enzyme by incubating 20 min. at 65°C and continue with step 23. Alternatively, when the restriction enzyme is not sensitive to heat inactivation, e.g. *Bgl*III, continue with step 22.1.
- 22.1. Add 80µl 10% SDS and incubate 30 min. at 65°C.
  - 22.2. Transfer the sample to a 50ml Falcon tube and add 5.4ml Milli-Q
  - 22.3. Add 700µl 10X Ligase buffer (10X: 660mM Tris-HCl pH 7.5, 50mM MgCl<sub>2</sub>, 10mM, DTT, 10mM ATP)
  - 22.4. Add 375µl 20% TX-100 and incubate 1hr 37°C
  - 22.5. Continue with step 26.

*Note: When facing ligation problems, these may be due to problems in nuclear accessibility. One solution is to perform these alternative heat inactivation steps for these samples, even if*

*the enzyme is heat-sensitive. The higher concentrations of detergents will improve nuclear lamina breakage and therefore improve ligation efficiencies. Be careful to assess template quality in the end!*

- Transfer the sample to a 50ml Falcon tube.
- Add 5.7ml Milli-Q.
- Add 700µl 10X Ligase buffer (see step 22.3).
- Add 50U T4 DNA Ligase (Roche, #799009), mix by swirling and incubate O/N at 16°C.
- Take a 100µl aliquot of the sample as the “Ligation control”.
- Determine ligation efficiency:

28.1. Add 5µl Prot K (10mg/ml) and incubate for 4 hours at 65°C.

28.2. Add 100µl Phenol-Chloroform to the sample and mix vigorously.

28.3. Spin 10 min, 16400g at RT.

28.4. Transfer water phase to a clean tube and load ~ 20µl on a 0.6% agarose gel next to the ‘digestion control’ from step 20.

28.5. If ligation is OK, proceed with step 29. If not, add fresh ATP (final concentration of 1mM) and repeat step 26-28.

#### ***Reverse cross-linking and precipitation***

- Add 30µl Prot K (10mg/ml) and reverse cross-link O/N at 65°C.
- Add 30µl RNase A (10mg/ml, Roche #10109169001) and incubate 45 minutes at 37°C.

- Add 7ml Phenol-Chloroform, mix vigorously.
- Centrifuge 15 min, 4800RPM at RT.
- Transfer the aqueous phase to a new 50ml Falcon tube and add:
  - 750µl 2M NaAC pH 5.6
  - 7µl Glycogen (20mg/ml, Roche #10901393001)
  - 17.5ml 100% EtOH.

Increasing the volume twice before precipitation (partially) prevents the co-precipitation of DTT from the ligase buffer and therefore results in a sample with higher purity.

- Mix and incubate at -80°C until the sample is frozen solid.
- Spin 30 min, 9500RPM at 4°C.
- Remove the supernatant and add 10 ml cold 70% ethanol.
- Centrifuge 15 min, 3270g at 4°C.
- Remove the supernatant and briefly dry the pellet at RT.
- Dissolve the pellet in 150µl 10mM Tris-HCl pH 7.5 at 37°C.

*Note: To completely dissolve pellet, you can incubate for 5 mins at 50°C or 65°C.*

- Continue with step 41 or store sample at -20°C.

### ***Second Digestion***

- To 150µl 3C sample (~1x10<sup>7</sup> cells) add:
  - 50µl 10X *DpnII* restriction buffer
  - Milli-Q to 500µl
  - 50U *DpnII* (New England Biolabs #R0543S)

- Incubate O/N at 37°C.
- Take a 5µl aliquot of the sample as the “Digestion control”.
- Determine digestion efficiency:

44.1. Add 95µl 10mM Tris-HCl pH 7.5 to the 5µl sample from step 43.

44.2. Load ~20µl on a 0.6% agarose gel next to the ‘ligation control’ from step 28.

44.3. If digestion is OK, proceed with step 45. If not, add fresh restriction enzyme and repeat step 42-44. Alternatively the sample can be re-purified to facilitate efficient digestion.

### ***Second Ligation and purification***

- Inactivate enzyme by incubating at 65°C for 25 minutes and continue with step 46. If not heat sensitive, the restriction enzyme can be inactivated by sample purification.

Continue with step 45.1.

45.1. Add 500µl Phenol-Chloroform and mix vigorously

45.2. Spin 10 min, 16400g at RT.

45.3. Transfer the aqueous phase to a fresh tube and add 50µl 2M NaAc pH 5.6 and 950µl 100% EtOH

45.4. Incubate at -80°C until completely frozen

45.5. Spin 20min 16400g at 4°C

45.6. Remove supernatant and add 150µl cold 70% ethanol.

45.7. Spin 10min 16400g at 4°C

45.8. Resuspend the pellet in 500µl 10mM Tris-HCl pH 7.5

- Transfer sample to a 50ml tube and add:
  - 12.1ml Milli-Q
  - 1.4 ml 10X Ligation buffer (see step 22.3)
  - 100U T4 DNA Ligase
- Ligate O/N at 16°C.
- Add 14ml Phenol-Chloroform, mix vigorously.

48.1. Centrifuge 15 min, 4800RPM at RT.

- Transfer aqueous phase to a new Falcon 50ml tube. Add: 1.4ml 2M NaAC pH 5.6, 14µl Glycogen (1mg/ml) and 35ml 100% EtOH. Mix well. Store at –80°C until completely frozen.
- Spin 45 min, 8346g at 4°C
- Remove the supernatant and add 10ml cold 70% ethanol.
- Spin 15 min, 3270g at 4°C.
- Remove the supernatant and briefly dry the pellet at RT.
- Dissolve the pellet in 150µl 10mM Tris-HCl pH 7.5 at 37°C.
- Purify samples with the QIAquick PCR purification kit (Qiagen #28104)

Use 3 columns per sample; binding capacity is 10µg DNA per column.

Elute columns with 50µl 10mM Tris-HCl pH 7.5 and pool samples.

- Measure concentration using the Nanodrop spectrophotometer and run a serial dilution of 0.125, 0.25, 0.5 and 1µl sample on a 2% agarose gel in order to estimate the concentration compared to a reference sample, e.g. phage-λ DNA.
- The 4C template is now finished and can be stored at -20°C or continued with directly in step 58.

## **PCR**

- Determine linear range of amplification by performing a PCR using template dilutions of 12.5, 25, 50 and 100ng 4C template. A typical 25µl PCR reaction consist of:
  - 2.5µl 10X PCR buffer 1(supplied with the Expand Long Template Polymerase)
  - 0.5µl dNTP (10mM)
  - 35pmol forward primer (1.5µl of a 1/7 dilution from a 1µg/µl 20nt primer stock)
  - 35pmol reverse primer (1.5µl of a 1/7 dilution from a 1µg/µl 20nt primer stock)
  - 0.35µl Expand Long Template Polymerase (Roche #11759060001)
  - X µl Milli-Q to a total volume of 25µl

A typical 4C-PCR program: 2' 94 °C; 10" 94 °C; 1' 55 °C; 3' 68 °C; 29x repeat; 5' 68 °C; 12°C. The concentration of primers used in a 4C-PCR is typically three times higher than a regular PCR as this often facilitates the efficiency of amplification.

- Separate 15µl PCR product on a 1.5% agarose gel and quantify to asses linear amplification and template quality.
- Determine the functionality of the adaptor primers by comparing them with the 'short' primers from step 58. Note the volume of the adaptor primers is corrected for their length difference by using 4.5µl and 3µl of a 1/7 diluted 1µg/µl stock solution of the ~75nt reading primer and the ~40nt reverse primer. The adaptor primers should cause a shift in PCR product length which should be visible when separated and

compared on a 1.5% agarose gel.

- When satisfied about the quality and quantity of the PCR product generated using the adaptor primers, the high complexity PCR is performed.
  - 80µl 10X PCR buffer 1
  - 16µl dNTP (10mM)
  - 1.12nmol 75nt reading primer (24µl reading primer of a 1µg/µl 75nt primer stock)
  - 1.12nmol 40nt reverse primer (16µl reverse primer of a 1µg/µl 40nt primer stock)
  - typically 3.2µg 4C template
  - 11.2µl Expand Long Template polymerase
  - Milli-Q water till 800ul total

Mix and separate into 16 reactions of 50µl before running the PCR

- Collect and pool the 16 reactions. Purify the sample using the High Pure PCR Product Purification Kit (Roche #11732676001), which effectively separates between the non-used adaptor primers (~75nt) and the PCR product (>120nt). Use minimal two columns per 16 reactions.

*Note: It is better to separate PCR products using AMPure beads. Users should optimize their beads concentration depending on the batch of beads available in their lab. A protocol is presented in this section which aids in preparing these tests. Typically, with new beads, a ratio of .85-.9X to sample volume effectively separates PCR products <175bp, which could potentially represent adaptor ligations. These artifacts should be minimized to retrieve as many informative reads from sequencing as possible.*

- Determine sample quantity and purity using the Nanodrop-spectrophotometer. Typically the yield resides between 10 and 20µg with A260/A280 ~1.85 and

A260/A230 >1.5. Sample purity is important to control to prevent complications during the sequencing procedure. If absorption ratios deviate re-purification is advised.

- Quality is determined by separation of 300ng purified PCR product on a 1.5% agarose gel.
- Combine 4C PCR products of different experiments in preferred ratios for sequencing.

*Note: For quantitative and equimolar pooling of different 4CSeq viewpoints, the preferred method is to use the KK4824 kit to correct for insert size lengths (Kapa Biosystems). BioAnalyzer is NOT recommended for this purpose.*



### **8.1.18 ATP, 100mM solution**

1 g ATP (adenosine triphosphate)

12 ml H<sub>2</sub>O

Adjust pH to 7.0 with 4M NaOH

Adjust volume to 16.7ml with H<sub>2</sub>O

Store in aliquots at -20°C

### **8.1.19 10x Ligation buffer**

- 660 mM Tris pH 7.5 26.4 ml 1 M Stock

- 50 mM MgCl<sub>2</sub> (Sigma: M2670) 2 ml 1M Stock

- 10 mM DTT (Sigma: 43816) 0.4 ml 1M Stock

- Aliquot per 2 ml and store @ -20°C

Add to 40 ml H<sub>2</sub>O.

### 8.1.20 Nuclei Isolation

Source: Paola Vegnarelli

1. Harvest cells and spin down 1,300rpm, 3min.
2. Wash cells twice in PBS.
3. Resuspend cells in ice-cold Nuclei Buffer (5 $\mu$ l / 10<sup>5</sup> vc) containing protease inhibitors and RNase Inhibitor. [Avoid cell clumps, but don't vortex!]
4. Incubate on ice (hypotonic swelling), 10 min.
5. Add equal volume Nuclei Buffer / 0.5% NP-40 [up to 1% depending on cell type]
6. Incubate on ice, 5min.
7. Vortex 10sec.
8. Centrifuge 1,000g, 1min, 4°C.
9. Wash nuclei once in Nuclei Buffer / 0.5% NP-40 containing protease inhibitors and RNase Inhibitor.
10. Centrifuge 1,000g, 1min, 4°C.

The nuclei can be directly lysed for protein or RNA extraction in 1x protein sample buffer or TRIZol, respectively. Alternatively, nuclei can be snap-frozen and stored at -80°C.

Buffers / Reagents:

Nuclei Buffer

Stock	final concentration	volume
1M Tris pH 7.6	10mM	100 $\mu$ l
5M NaCl	10mM	20 $\mu$ l
1M MgCl <sub>2</sub>	2mM	20 $\mu$ l
dH <sub>2</sub> O		to 10ml

10% NP-40 solution in dH<sub>2</sub>O (protect from light and store at 4°C)

Protease inhibitor cocktail; P8340 (Sigma): 1:100

Anti-RNase; AM-2690 (Ambion): 40U/ $\mu$ l

### 8.1.21 Ampure XP Protocol 0.9x:

For removal of adapters, nucleotides, etc sizes < 175 bp.

1. Shake AMPure XP Beads to resuspend
2. Add 0.9x AMPure XP Beads to sample and mix thoroughly via pipetting up and down 10x
  - 2.1. 20  $\mu$ L library + 18  $\mu$ l of resuspended AMPure XP Beads
3. Incubate at RT for 5 minutes
4. Place Beads in magnetic stand and let settle for 2 min
5. Remove Supernatant (but keep, just in case...)
6. Wash twice with 100  $\mu$ l with fresh 70% Ethanol
  - 6.1. Allow Ethanol wash to incubate for 30 seconds to 2 min each
7. Remove Ethanol. Let dry for 4 mins.
8. Remove the sample from the magnetic stand
9. Quickly centrifuge (< 5 seconds @ < 2000xg) to get all of the beads to the bottom of the tube
10. Elute with small volume of EB (20  $\mu$ l). Pipette up and down 10 times or more until dissolved.
11. Allow to incubate for 5 minutes
12. Place back into magnetic stand
13. Let settle for 2 minutes
14. Remove cleaned sample

## 8.2 Computational Methods

### 8.2.1 3D DNA FISH analysis by `Correct_and_Measure_3D.class` ImageJ plugin

The 3D DNA FISH analysis by `Correct_and_Measure_3D.class` is an ImageJ plugin. It automatically analyzes 3D DNA FISH z-stacks files and searches for 4 different channels (488, 594, 647, plus DAPI). The segmentation and image analysis is performed as described in Chapter 2. The plugin can be accessed and downloaded from the CSHL bnbdev server:

`/sonas-`

`hs/spector/nlsas/data/czepeda/Paper/3D_DNA_FISH/Correct_and_Measure_3D_v6.class`

The plugin produces 2 files for each FISH image analyzed:

`imagename_Measurements.txt`

`imagename_ParticleStatistics.txt.`

Where `imagename` is the name of the FISH file. Descriptive headers are included in each table. Additionally, a summary of all measurements for the image folder analyzed is made, `Summary.txt`, which filters results and excludes cells with deviant number of signals.

For this summary file to be made, FISH image folders should be formatted in the form:

`Aim1b_expnumber_date_genotype`

For example, `Aim1b_22_073114_Df` refers to Aim1b experiment no. 22, performed on the

31st of July of 2014 for  $df/+^{Bl6}$  cells. WT refers to  $+^{129}/+^{Bl6}$  MEFs.

A final file, reporting all distances measures between all 3 probe channels, is produced for each analyzed image folder, with the name

Dist\_Ch1Ch2\_Ch3.txt

## 8.2.2 Custom R, Bash, and Perl scripts for the analysis of Correct\_and\_Measure\_3D.class ImageJ plugin

The result files from the Correct\_and\_Measure\_3D.class are analyzed through the use of custom made R, Bash, and Perl scripts. The scripts can be accessed, run, or downloaded from the CSHL bnbdev server:

```
/sonas-hs/spector/nlsas/data/czepeda/Paper/3D_DNA_FISH/FISH_scripts
```

The script loop\_and\_cat\_files\_Jan2013.sh automatically runs a battery of Perl table formatting scripts, and submits the output to dedicated R scripts which run detailed statistical analyses on folders of  $df/+^{Bl6}$  and  $+^{129}/+^{Bl6}$  data.

Scripts run include:

```
parse_aim1b_measurements_Df.pl
```

```
parse_aim1b_measurements_Dp.pl
```

```
parse_aim1b_measurements_WT.pl
```

```
Aim1_b_stats_Jan2013.r
```

```
Aim1_b_heterochrstats_Jan2013.r
```

```
Aim1_b_perifcentstats_Jan2013.r
```

```
measur_permutations_Jan2013.r
```

```
aim1b_newstats_nucvol_filtered_Jan2013.r
```

Script descriptions, input/output formats, and additional comments are included in the body of the script itself, and read through any .txt reader.

The script `getFISH_3probedist.sh` automatically runs Perl scripts which analyzes the distances separating the 3 probe channels and outputs the data as tables for a specified folder.

Scripts run include:

`obtain_3probedistances_WT.pl`

`obtain_3probedistances_Df.pl`

Additional questions or comments regarding the scripts and how to run them should be addressed to [czepeda@cshl.edu](mailto:czepeda@cshl.edu). Additional inquiries regarding the plugin, should be addressed to Nathalie Harder [N.Harder@dkfz-heidelberg.de](mailto:N.Harder@dkfz-heidelberg.de)



### 8.2.3 PE-4Cseq reads analysis pipeline

PE-4Cseq fastq files were filtered first based on the viewpoint of origin. The script `split_fastq_withqual.pl` performs this task for all of the viewpoints analyzed in this thesis. Each viewpoint should be run separately, and each line of the analyzed viewpoint selected in the script to obtain the desired reads. The outputs are viewpoint filtered fastq files for PE1 and PE2. The script's input/output formats and general overview are included in its text body.

The filtered viewpoint reads are further processed with the `split_4C_snp_withqual.pl` script. This script takes each viewpoint's reads and separates them based on allelic origin, either 129S5/SvEvBrd (129), or C57Bl6/J (bl6). Fastq files obtained from this program can be used for subsequent mapping.

Both scripts can be accessed, run, or downloaded from the CSHL bnbdev server:

`/sonas-hs/spector/nlsas/data/czepeda/Paper/4Cseq_scripts`

#### **8.2.4 Monte Carlo Simulations for CTCF, Smc1, Med1, and Med12 data**

To assess whether the CTCF/Smc1 overlap ratio was significant for the differentially interacting regions, I computed the probability of exceeding the number of protein binding sites in these regions against randomly chosen sequences of the same size as the differentially interacting regions analyzed. I performed this task using in a Monte Carlo simulation with 1,000 repeats using the bedtools suite (Quinlan AR., Hall, I. 2010).

The intersections between all datasets and the differentially interacting regions is performed by the script:

```
intersections_data.sh
```

The results from the intersections are then used to establish the observed values against which simulation will be compared.

The BEDtools Shuffle program will choose a new location for each of the original differentially interacting regions while preserving its size in chromosome 4. The script montecarlo.sh prints out how many intersections were observed for each of 1000 shuffles making use of this BEDtools program. A p-value was derived by counting the number of times that the number of shuffled intersections exceeds the observed intersections. If 0, then p-val is less than 0.001.

Scripts:

intersections\_data.sh

montecarlo.sh

CTCF\_MEF\_enriched\_regions\_mm9\_noheader.bed

Smc1\_MEF\_enriched\_regions\_mm9\_noheader.bed

Med1\_MEF\_enriched\_regions\_mm9\_noheader.bed

Med12\_MEF\_enriched\_regions\_mm9\_noheader.bed

mm9.chr.sizes

cnv\_coords.bed

DE129\_chr4.bed

DEB16\_chr4.bed

DEcombined\_chr4.bed

All scripts can be accessed, run, or downloaded from the CSHL bnbdev server:

[/sonas-hs/spector/nlsas/data/czepeda/Paper/montecarlo\\_scripts](#)

All the differentially interacting data can be accessed or downloaded from the CSHL bnbdev

server:

[/sonas-hs/spector/nlsas/data/czepeda/Paper/4C\\_data](#)

## References

- Ahmed, S. *et al.* (2010). DNA zip codes control an ancient mechanism for gene targeting to the nuclear periphery. *Nature Cell Biol*, 12:111–118.
- Aitman, T. J. *et al.* (2006). Copy number polymorphism in *Fcgr3* predisposes to glomerulonephritis in rats and humans. *Nature*, 439:851–855.
- Alkan C., Coe BP., Eichler EE. (2011). Genome structural variation discovery and genotyping. *Nat Rev Genet*, 12(5):363-76.
- Alkan, C. *et al.* (2009). Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genet*, 41:1061–1067.
- Anders S, Huber W. (2010). Differential expression analysis for sequence count data. *Genome Biol*, 11(10):R106.
- Andrulis, E.D., Neiman, A.M., Zappulla, D.C., Sternglanz, R. (1998). Perinuclear localization of chromatin facilitates transcriptional silencing. *Nature*, 394:592–595.
- Arndt AK, Schafer S, Drenckhahn JD, Sabeh MK, Plovie ER, Caliebe A, Klopocki E, Musso G, Werdich AA, Kalwa H, Heinig M, Padera RF, Wassilew K, Bluhm J, Harnack C, Martitz J, Barton PJ, Greutmann M, Berger F, Hubner N, Siebert R, Kramer HH, Cook SA, MacRae CA, Klaassen S. (2013). Fine mapping of the 1p36 deletion syndrome identifies mutation of PRDM16 as a cause of cardiomyopathy. *Am J Hum Genet*, 93(1):67-77.
- Avery, OT., MacLeod, CM., McCarty, M. (1944). Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III. *Journal of Experimental Medicine*, 79(2):137–158.

Bae, S., Choi, J. (2004) Tumor suppressor activity of RUNX3. *Oncogene*. 23, 4336–4340.

Bagchi A, Papazoglu C, Wu Y, Capurso D, Brodt M, Francis D, Bredel M, Vogel H, Mills AA. (2007). CHD5 is a tumor suppressor at human 1p36. *Cell*, 128:459-75.

Bagchi, A. & Mills, A.A. (2008). The quest for the 1p36 tumor suppressor. *Cancer Res*, 68:2551–2556.

Bantignies F, Roure V, Comet I, Leblanc B, Schuettengruber B, Bonnet J, Tixier V, Mas A, Cavalli G. (2011). Polycomb-dependent regulatory contacts between distant Hox loci in *Drosophila*. *Cell*, 144(2):214-26.

Barbieri M, Scialdone A, Piccolo A, Chiariello AM, di Lanno C, Prisco A, Pombo A, Nicodemi M. (2013). Polymer models of chromatin organization. *Front Genet*, 4:113.

Bartolomei, M.S., and Ferguson-Smith, A.C. (2011). Mammalian genomic imprinting. *Cold Spring Harb. Perspect. Biol*, 3:a002592.

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

Belmont AS, Bruce K. (1994). Visualization of G1 chromosomes: a folded, twisted, supercoiled chromonema model of interphase chromatid structure. *J Cell Biol*, 127(2):287-302.

Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, Sboner A, Esgueva R, Pflueger D, Sougnez C, Onofrio R, Carter SL, Park K, Habegger L, Ambrogio L, Fennell T, Parkin M, *et al.* (2010). The genomic complexity of primary human prostate cancer. *Nature*, 470(7333):214-20.

Blakeslee, AF. (1922). Variation in *Datura* due to changes in chromosome number. *Amer*

*Naturalist*, 56 16-31.

Boisvert FM, van Koningsbruggen S, Navascués J, Lamond AI. (2007). The multifunctional nucleolus. *Nat Rev Mol Cell Biol*, 8:574–585.

Bolzer, A., *et al.* (2005). Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol*, 3:e157.

Bongiorno-Borbone L., De Cola A., Vernole P., Finos .L, Barcaroli D., Knight RA., Melino G, De Laurenzi V. (2008). FLASH and NPAT positive but not Coilin positive Cajal Bodies correlate with cell ploidy. *Cell Cycle*, 7(15):2357-67.

Boveri T. (1909). Die Blastomerenkerne von *Ascaris megalcephala* und die Theorie der Chromosomenindividualität. *Arch Zellforsch*, 3:181-268.

Boveri, M. (1903). Über Mitosen bei einseitiger Chromosomenbindung. *Jenaische Zeitschrift für Naturwissenschaft*, 37:401–443.

Branco MR, Pombo A. (2006). Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *K*, 4(5):e138.

Brenner, S., Jacob, F., Meselson, M. (1961) An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature*, 190:576–581.

Bridges, C. B. (1936). The Bar "gene": A duplication. *Science*, 83, 210–211.

Brown, C. R., Kennedy, C. J., Delmar, V. A., Forbes, D. J. & Silver, P. A. (2008). Global histone acetylation induces functional genomic reorganization at mammalian nuclear pore complexes. *Genes Dev*, 22:627–639.

Cabal, G. G. *et al.* (2006). SAGA interacting factors confine sub-diffusion of transcribed genes to the nuclear envelope. *Nature*, 441:770–773.

Cahan P, Li Y, Izumi M, Graubert TA. (2009). The impact of copy number variation on local

gene expression in mouse hematopoietic stem and progenitor cells. *Nat Genet*, 41:430–437.

Callan, H. G., Tomlin, S. G. (1950). Experimental studies on amphibian oocyte nuclei. I. Investigation of the structure of the nuclear membrane by means of the electron microscope. *Proc. R. Soc. Lond. B Biol. Sci*, 137:367–378.

Campbell PJ, Yachida S, Mudie LJ, Stephens PJ, Pleasance ED, Stebbings LA, Morsberger LA, Latimer C, McLaren S, Lin ML, McBride DJ, Varela I, Nik-Zainal SA, Leroy C, Jia M, Menzies A, Butler AP, Teague JW, Griffin CA, Burton J, Swerdlow H, Quail MA, Stratton MR, Iacobuzio-Donahue C, Futreal PA. (2010). The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature*, 467(7319):1109-13.

Carter D., Chakalova L., Osborne CS., Dai YF., Fraser P. (2002). Long-range chromatin regulatory interactions in vivo. *Nat Genet*, 32(4):623-6.

Chen D, Huang S. (2001). Nucleolar components involved in ribosome biogenesis cycle between the nucleolus and nucleoplasm in interphase cells. *J Cell Biol*, 153(1):169-76.

Chen WK, Swartz JD, Rush LJ, Alvarez CE. (2009). Mapping DNA structural variation in dogs. *Genome Res*, 19:500–509.

Chen, K. *et al.* (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods* 6, 677–681.

Chuang, C.H. & Belmont, A.S. (2007). Moving chromatin within the interphase nucleus-controlled transitions? *Semin. Cell Dev. Biol*, 18:698–706.

Coco R., Penchaszadeh V.B. (1982). Cytogenetic findings in 200 children with mental retardation and multiple congenital anomalies of unknown cause. *Am. J. Med. Genet*, 12:155–173.

Cohen SN., Chang AC., Boyer HW., Helling RB. (1973). Construction of biologically

functional bacterial plasmids *in vitro*. *Proc Natl Acad Sci*, 70(11):3240-4.

Conrad D.F., Andrews T.D., Carter N.P., Hurles M.E., Pritchard J.K. (2006). A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet*, 38:75–81.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AW, Robson S, Stirrups K, Valsesia A, Walter K, Wei J; Wellcome Trust Case Control Consortium, Tyler-Smith C, Carter NP, Lee C, Scherer SW, Hurles ME. (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704-12.

Cooper G.M., Coe B.P., Girirajan S., Rosenfeld J.A., *et al.* (2011). A copy number variation morbidity map of developmental delay. *Nat. Genet*, 43:838-846.

Craddock, N., *et al.* (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 464:713–720.

Cremer T, Cremer C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet*, 2:292–301.

Cremer T, Cremer C. (2006). Rise, fall and resurrection of chromosome territories: a historical perspective. Part II. Fall and resurrection of chromosome territories during the 1950s to 1980s. Part III. Chromosome territories and the functional nuclear architecture: experiments and models from the 1990s to the present. *Eur J Histochem*, 50(4):223-72.

Cremer T, Cremer M. (2010). Chromosome territories. *Cold Spring Harb Perspect Biol*, 2(3):a003889.

Cremer, T., Cremer, C., Baumann, H., Luedtke, E. K., Sperling, K., Teuber, V., and Zorn, C. (1982). Rabl's model of the interphase chromosome arrangement tested in Chinese hamster



cells by premature chromosome condensation and laser-UV-microbeam experiments. *Human Genetics*, 60(1):46–56.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA, Young RA, Jaenisch R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A*. 107(50):21931-6.

Croft, J. A. *et al.* (1999). Differences in the localization and morphology of chromosomes in the human nucleus. *J. Cell Biol*, 145:1119–1131.

Davies HG, Murray AB, Walmsley ME. (1974). Electron-microscope observations on the organization of the nucleus in chicken erythrocytes and a superunit thread hypothesis for chromosome structure. *J Cell Sci*, 16:261–299

de Cid, R., *et al.* (2009). Deletion of the late cornified envelope *LCE3B* and *LCE3C* genes as a susceptibility factor for psoriasis. *Nature Genet*, 41:211–215.

de Laat W, Duboule D. (2013). Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature*, 502(7472):499-506.

de Wit E, Bouwman BA, Zhu Y, Klous P, Splinter E, Verstegen MJ, Krijger PH, Festuccia N, Nora EP, Welling M, Heard E, Geijsen N, Poot RA, Chambers I, de Laat W. (2013). The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature*, 501(7466):227-31.

DeBolt S. (2010). Copy number variation shapes genome diversity in Arabidopsis over immediate family generational scales. *Genome Biol Evol*, 2:441-53.

Dekker J. (2006). The three 'C' s of chromosome conformation capture: controls, controls, controls. *Nat Methods*, 3(1):17-21.

- Dekker J., Rippe K., Dekker M., Kleckner N. (2002). Capturing chromosome conformation. *Science*, 295:1306-11.
- Dellaire G., Bazett-Jones DP. (2007). Beyond repair foci: subnuclear domains and the cellular response to DNA damage. *Cell Cycle*, 6(15):1864-72.
- des Cloizeaux, J., Jannink, G. (2010). Polymers in Solution: Their Modelling and Structure. *Oxford Classic Texts in the Physical Sciences*.
- Dieppois, G., Iglesias, N. & Stutz, F. (2006). Cotranscriptional recruitment to the mRNA export receptor Mex67p contributes to nuclear pore anchoring of activated genes. *Mol Cell Biol*, 26:7858–7870.
- Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376-80.
- Dopman EB, Hartl DL. (2007). A portrait of copy-number polymorphism in *Drosophila melanogaster*. *Proc Natl Acad Sci USA*, 104(50):19920-5.
- Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C, Green RD, Dekker J. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res*, 16(10):1299-309.
- Dousset T, Wang C, Verheggen C, Chen D, Hernandez-Verdun D, Huang S. (2000). Initiation of nucleolar assembly is independent of RNA polymerase I transcription. *Mol Biol Cell*, 11(8):2705-17.
- Draker R, Cheung P. (2009). Transcriptional and epigenetic functions of histone variant H2A.Z. *Biochem Cell Biol*, 87(1):19-25.

Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau CA, Noble WS. (2010). A three-dimensional model of the yeast genome. *Nature*, 465(7296):363-7.

Dundr M, Misteli T. (2010). Biogenesis of nuclear bodies. *Cold Spring Harb Perspect Biol*, 2(12):a000711.

Eberharter A, Becker PB. (2004). ATP-dependent nucleosome remodelling: factors and functions. *J Cell Sci*, 117(17):3707-11.

Eckersley-Maslin MA, Thybert D, Bergmann JH, Marioni JC, Flicek P, Spector DL. (2014). Random monoallelic gene expression increases upon embryonic stem cell differentiation. *Dev Cell*, 28(4):351-65.

Edwards J.H., Harnden D.G., Cameron A.H., Crosse V.M., Wolff O.H. (1960). A new trisomic syndrome. *Lancet*, 1:787-790.

Egan CM, Sridhar S, Wigler M, Hall IM. (2007). Recurrent DNA copy number variation in the laboratory mouse. *Nat Genet*, 39(11):1384-9.

Engelhardt, P. & Pusa, K. (1972). Nuclear pore complexes: “press-stud” elements of chromosomes in pairing and control. *Nature New Biol*, 240:163-166.

Felsenfeld G, Groudine M. (2003). Controlling the double helix. *Nature*, 421(6921):448-53.

Ferraiuolo MA, Rousseau M, Miyamoto C, Shenker S, Wang XQ, Nadler M, Blanchette M, Dostie J. (2010). The three-dimensional architecture of Hox cluster silencing. *Nucleic Acids Res*, 38(21):7472-84.

Ferreira, J, G Paoella, C Ramos, AI Lamond. (1997). Spatial organization of large-scale chromatin domains in the nucleus: a magnified view of single chromosome territories. *J Cell Biol*, 139:1597-1610.

- Feuerbach, F., Galy, V., Trelles-Sticken, E., Fromont-Racine, M., Jacquier, A., Gilson, E., Olivo-Marin, J.C., Scherthan, H., Nehrbass, U. (2002). Nuclear architecture and spatial positioning help establish transcriptional states of telomeres in yeast. *Nat. Cell Biol*, 4:214–221.
- Feuk, L., *et al.* (2006). Structural variation in the human genome. *Nature Reviews Genetics*, 7:85–97.
- Finch JT, Klug A. (1976) Solenoidal model for superstructure in chromatin. *Proc Natl Acad Sci USA*, 73:1897-1901.
- Flemming, W. (1877). Beobachtungen über die Beschaffenheit des Zell-kerns. *Archiv für mikroskopische Anatomie*, 13:693-717.
- Flemming, W. (1878). Zur Kenntniss der Zelle und ihrer Theilungs-Erscheinungen. *Schriften des Naturwissenschaftlichen Vereins für Schleswig-Holstein*, 3:23–27.
- Fontanesi L, Martelli PL, Beretti F, Riggio V, Dall'Olio S, Colombo M, Casadio R, Russo V, Portolano, B. (2010). An initial comparative map of copy number variations in the goat (*Capra hircus*) genome. *BMC Genomics*, 11:639.
- Fox, MH, DJ Arndt-Jovin, TM Jovin, PH Baumann, M Robert-Nicoud. (1991). Spatial and temporal distribution of DNA replication sites localized by immunofluorescence and confocal microscopy in mouse fibroblasts. *J Cell Sci*, 99:247–253.
- Franklin, R. & Gosling, R. G. (1953). Molecular configuration in sodium thymonucleate. *Nature*, 171, 740–741.
- Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, *et al.* (2009). An oestrogen-receptor- $\alpha$ -bound human chromatin interactome. *Nature*, 462(7269):58-64.

- Fussner E, Ching RW, Bazett-Jones DP. (2011). Living without 30nm chromatin fibers. *Trends Biochem Sci*, 36(1):1-6.
- Gall JG. (1966). Chromosome fibers studied by a spreading technique. *Chromosoma*, 20(2):221-33.
- Gall, J. G., Pardue, M. L. (1969). Formation and detection of RNA-DNA hybrid molecules in cytological preparations. *Proc Natl Acad Sci USA*, 63:378–383.
- Gant, T. M., and Wilson, K. L. (1997). Nuclear assembly. *Annu. Rev. Cell Dev. Biol*, 13:669–695.
- Gavrilov AA, Golov AK, Razin SV. (2013). Actual ligation frequencies in the chromosome conformation capture procedure. *PLoS One*, 8(3):e60403.
- Gazave E, *et al.* (2011). Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. *Genome Res*, 21(10):1626–1639.
- Gheldof N, Smith EM, Tabuchi TM, Koch CM, Dunham I, Stamatoyannopoulos JA, Dekker J. (2010). Cell-type-specific long-range looping interactions identify distant regulatory elements of the CFTR gene. *Nucleic Acids Res*, 38(13):4325-36.
- Gibcus JH., Dekker J. (2013). The hierarchy of the 3D genome. *Mol Cell*, 49(5):773-82.
- Gilman SR, Iossifov I, Levy D, Ronemus M, Wigler M, Vitkup D. (2011). Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron*, 70(5):898-907.
- Gokcumen O, *et al.* (2011). Refinement of primate copy number variation hotspots identifies candidate genomic regions evolving under positive selection. *Genome Biol*, 12(5):R52.
- Gokcumen O, Tischler V, Tica J, Zhu Q, Iskow RC, Lee E, Fritz MH, Langdon A, Stütz AM, Pavlidis P, Benes V, Mills RE, Park PJ, Lee C, Korbelt JO. (2013). Primate genome

architecture influences structural variation mechanisms and functional consequences. *Proc Natl Acad Sci USA*, 110(39):15764-9.

Goldman RD, Gruenbaum Y, Moir RD, Shumaker DK, Spann TP. (2002). Nuclear lamins: building blocks of nuclear architecture. *Genes Dev*, 16(5):533-47.

Gonzalez, E., *et al.* (2005). The Influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*, 307:1434–1440.

Graubert TA, Cahan P, Edwin D, Selzer RR, Richmond TA, Eis PS, Shannon WD, Li X, McLeod HL, Cheverud JM, Ley TJ. (2007). A high-resolution map of segmental DNA copy number variation in the mouse genome. *PLoS Genet*, 5;3(1):e3.

Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A., Stevens C., Edkins S., O'Meara S., Vastrik I., Schmidt EE., *et al.* (2007). Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153-8.

Greil, F., Moorman, C., van Steensel, B. (2006). DamID: mapping of in vivo protein-genome interactions using tethered DNA adenine methyltransferase. *Methods Enzymol*, 410:342–359.

Grigoryev SA, Woodcock CL. (2012). Chromatin organization - the 30 nm fiber. *Exp Cell Res*, 318(12):1448-55.

Grigoryev SA, Woodcock CL. (2012). Chromatin organization - the 30 nm fiber. *Exp Cell Res*, 318(12):1448-55.

Groot P.C., Mager W.H., Frants R.F. (1991). Interpretation of polymorphic DNA patterns in the human  $\alpha$ -amylase multigene family. *Genomics*, 10:779–785.

Gruenbaum Y, Margalit A, Goldman RD, Shumaker DK, Wilson KL. (2005). The nuclear lamina comes of age. *Nat Rev Mol Cell Biol*, 6(1):21-31.

Guacci V., Koshland D., Strunnikov A. (1997) A direct link between sister chromatid

cohesion and chromosome condensation revealed through the analysis of MCD1 in *S. cerevisiae*. *Cell*, 91:47–57.

Guan, X. Y., Trent, J. M., and Meltzer, P. S. (1993). Generation of band-specific painting probes from a single microdissected chromosome. *Human Molecular Genetics*, 2(8):1117–1121.

Guelen, L., *et al.* (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453:948–951.

Guryev V, Saar K, Adamovic T, Verheul M, van Heesch SA, Cook S, Pravenec M, Aitman T, Jacob H, Shull JD, Hubner N, Cuppen E. (2008). Distribution and functional impact of DNA copy number variation in the rat. *Nat Genet*, 40(5):538-45.

Hadjebi O, Casas-Terradellas E, Garcia-Gonzalo FR, Rosa JL. (2008). The RCC1 superfamily: from genes, to function, to disease. *Biochim Biophys Acta*, 1783(8):1467-79.

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

Hakim O, Sung MH, Voss TC, Splinter E, John S, Sabo PJ, Thurman RE, Stamatoyannopoulos JA, de Laat W, Hager GL. (2011). Diverse gene reprogramming events occur in the same spatial clusters of distal regulatory elements. *Genome Res*, 21(5):697-706.

Hanahan D, Weinberg RA. (2011). Hallmarks of cancer: the next generation. *Cell*, 144(5):646-74.

Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CW, Ye C, Ping JL, Mulawadi F, Wong E, Sheng J, Zhang Y, Poh T, Chan CS, Kunarso G, Shahab A, Bourque G, Cacheux-Rataboul V, Sung WK, Ruan Y, Wei CL. (2011). CTCF-mediated functional chromatin

interactome in pluripotent cells. *Nat Genet*, 43(7):630-8.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. (2009). Mechanisms of change in gene copy number. *Nat Rev Genet*, 10(8):551-64.

Heilstedt HA, Ballif BC, Howard LA, Kashork CD, Shaffer LG. (2003). Population data suggest that deletions of 1p36 are a relatively common chromosome abnormality. *Clin Genet*, 64(4):310-6.

Hillmer AM, Yao F, Inaki K, Lee WH, Ariyaratne PN, Teo AS, Woo XY, Zhang Z, Zhao H, Ukil L, Chen JP, Zhu F, So JB, Salto-Tellez M, Poh WT, Zawack KF, Nagarajan N, Gao S, Li G, Kumar V, Lim HP, Sia YY, Chan CS, Leong ST, Neo SC, Choi PS, Thoreau H, Tan PB, Shahab A, Ruan X, Bergh J, Hall P, Cacheux-Rataboul V, Wei CL, Yeoh KG, Sung WK, Bourque G, Liu ET, Ruan Y. (2011). Comprehensive long-span paired-end-tag mapping reveals characteristic patterns of structural variations in epithelial cancer genomes. *Genome Res*, 21(5):665-75.

Hinds D.A., Kloek A.P., Jen M., Chen X., Frazer K.A. (2006). Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat. Genet*, 38:82–85.

Hofstra RM, Valdenaire O, Arch E, Osinga J, Kroes H, Löffler BM, Hamosh A, Meijers C, Buys CH. (1999). A loss-of-function mutation in the endothelin-converting enzyme 1 (ECE-1) associated with Hirschsprung disease, cardiac defects, and autonomic dysfunction. *Am J Hum Genet*, 64(1):304-8.

Hollox E.J., Armour J.A.L., Barber J.C.K. (2003). Extensive normal copy number variation of a  $\beta$ -Defensin antimicrobial-gene cluster. *Am. J. Hum. Genet*, 73:591–600.

Holwerda SJ, van de Werken HJ, Ribeiro de Almeida C, Bergen IM, de Bruijn MJ, Verstegen MJ, Simonis M, Splinter E, Wijchers PJ, Hendriks RW, de Laat W. (2013). Allelic exclusion



of the immunoglobulin heavy chain locus is independent of its nuclear localization in mature B cells. *Nucleic Acids Res*, 41(14):6905-16.

Homma, S., Iwasaki, M., Shelton, G.D., Engvall, E., Reed, J.C., and Takayama, S. (2006). BAG3 deficiency results in fulminant myopathy and early lethality. *Am. J. Pathol*, 169:761–773.

Horike S, Cai S, Miyano M, Cheng JF, Kohwi-Shigematsu T. (2005). Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome. *Nat Genet*, 37(1):31-40.

Hormozdiari, F., Alkan, C., Eichler, EE., Sahinalp, SC. (2009). Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res*, 19:1270–1278.

Hosak L. (2013). New findings in the genetics of schizophrenia. *World J Psychiatry*, 3(3):57-61.

Hou, C., Li, L., Qin, Z.S., and Corces, V.G. (2012). Gene density, transcription, and insulators contribute to the partition of the *Drosophila* genome into physical domains. *Mol. Cell*, 48:471–484.

Hu Y, Kireev I, Plutz M, Ashourian N, Belmont AS. (2009). Large-scale chromatin structure of inducible genes: transcription on a condensed, linear template. *J Cell Biol*, 185(1):87-100.

Huang B., Bates M., Zhuang X. (2009). Super-resolution fluorescence microscopy. *Annu Rev Biochem*, 78:993-1016.

Huang RC, Bonner J. (1962). Histone, a suppressor of chromosomal RNA synthesis. *Proc. Natl. Acad. Sci*, 48:1216–22.

Iafrate, A. J., *et al.* (2004). Detection of large-scale variation in the human genome. *Nature*

*Genetics*, 36:949–951.

International HapMap 3 Consortium, Altshuler DM, Gibbs RA, *et al.* (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52-8.

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature* 409(6822):860-921.

Ishii, K., Arib, G., Lin, C., Van Houwe, G. & Laemmli, U. K. (2002). Chromatin boundaries in budding yeast: the nuclear pore connection. *Cell*, 109:551–562.

Iskow RC, Gokcumen O, Abyzov A, Malukiewicz J, Zhu Q, Sukumar AT, Pai AA, Mills RE, Habegger L, Cusanovich DA, Rubel MA, Perry GH, Gerstein M, Stone AC, Gilad Y, Lee C. (2011). Regulatory element copy number differences shape primate expression profiles. *Proc Natl Acad Sci USA*, 109(31):12656-61.

Jackson DA, Pombo A. (1998). Replicon clusters are stable units of chromosome structure: Evidence that nuclear organization contributes to the efficient activation and propagation of S phase in human cells. *J Cell Biol*, 140:1285–1295.

Jackson DA., Symons RH., Berg P. (1972). Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of *Escherichia coli*. *Proc Natl Acad Sci*, 69(10): 2904-9.

Jacobs P.A., Baikie A.G., Court Brown W.M., Strong J.A. (1959). The somatic chromosomes in mongolism. *Lancet*, 1:710.

Jacobs P.A., Browne C., Gregson N., Joyce C., White H. (1992). Estimates of the frequency of chromosome abnormalities detectable in unselected newborns using moderate levels of banding. *J. Med. Genet*, 29:103–108.

Jacobs P.A., Matsuura J.S., Mayer M., Newlands I.M. (1978). A cytogenetic survey of an institution for the mentally retarded: I. Chromosome abnormalities. *Clin. Genet*, 13:37–60.

Jakobsson, J., *et al.* (2006). Large differences in testosterone excretion in Korean and Swedish men are strongly associated with a UDP-glucuronosyl transferase 2B17 polymorphism. *J. Clin. Endocrinol. Metab*, 91:687–693.

Jenuwein T, Allis C. (2001). Translating the histone code. *Science*, 293(5532):1074–80.

Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen CA, Schmitt AD, Espinoza CA, Ren B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, 503(7475):290-4.

Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, Ebmeier CC, Goossens J, Rahl PB, Levine SS, Taatjes DJ, Dekker J, Young RA. (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314):430-5.

Kaplan N, Dekker J. (2013). High-throughput genome scaffolding from in vivo DNA interaction frequency. *Nat Biotechnol*, 31(12):1143-7.

Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, Furlotte NA, Eskin E, Nellåker C, Whitley H, Cleak J, Janowitz D, Hernandez-Pliego P, Edwards A, Belgard TG, Oliver PL, McIntyre RE, Bhomra A, Nicod J, Gan X, Yuan W, van der Weyden L, Steward CA, Bala S, Stalker J, Mott R, Durbin R, Jackson IJ, Czechanski A, Guerra-Assunção JA, Donahue LR, Reinholdt LG, Payseur BA, Ponting CP, Birney E, Flint J, Adams DJ. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, 477(7364):289-94.

Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen J, Harmanci A, Das J, Abyzov A, Balasubramanian S, Beal K, Chakravarty D, Challis

D, Chen Y, Clarke D, Clarke L, Cunningham F, Evani US, Flicek P, Fragoza R, Garrison E, Gibbs R, Gümüs ZH, Herrero J, Kitabayashi N, Kong Y, Lage K, Liliashvili V, Lipkin SM, MacArthur DG, Marth G, Muzny D, Pers TH, Ritchie GR, Rosenfeld JA, Sisu C, Wei X, Wilson M, Xue Y, Yu F; 1000 Genomes Project Consortium, Dermitzakis ET, Yu H, Rubin MA, Tyler-Smith C, Gerstein M. (2013). Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science*, 342(6154):1235587.

Kieffer-Kwon KR, Tang Z, Mathe E, *et al.* (2013). Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. *Cell*, 155(7):1507-20.

Kill, IR, JM Bridger, KHS Campbell, G Maldonado-Codina, CJ Hutchison. (1991). The timing of the formation and usage of replicase clusters in S-phase nuclei of human diploid fibroblasts. *J Cell Sci*, 100:869–876.

Kim BJ, Zaveri HP, Shchelochkov OA, Yu Z, Hernández-García A, Seymour ML, Oghalai JS, Pereira FA, Stockton DW, Justice MJ, Lee B, Scott DA. (2013). An allelic series of mice reveals a role for RERE in the development of multiple organs affected in chromosome 1p36 deletions. *PLoS One*, 8(2):e57460.

Kind J, Pagie L, Ortabozkoyun H, Boyle S, de Vries SS, Janssen H, Amendola M, Nolen LD, Bickmore WA, van Steensel B. (2013). Single-cell dynamics of genome-nuclear lamina interactions. *Cell*, 153(1):178-92.

Kizilyaprak C, Spehner D, Devys D, Schultz P. (2010). In vivo chromatin organization of mouse rod photoreceptors correlates with histone modifications. *PLoS One*, 5(6):e11039.

Korbel, J. O. *et al.* (2007). Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318:420–426.

Kornberg, R. (1974). Chromatin structure: a repeating unit of histones and DNA. *Science*, 184:868–871.

Kruhlak MJ, Lever MA, Fischle W, Verdin E, Bazett-Jones DP, Hendzel MJ. (2000). Reduced mobility of the alternate splicing factor (ASF) through the nucleoplasm and steady state speckle compartments. *J Cell Biol*, 10;150(1):41-51.

Krull, S. *et al.* (2010). Protein Tpr is required for establishing nuclear pore-associated zones of heterochromatin exclusion. *EMBO J*, 29:1659–1673.

Kumaran RI, Spector DL. (2008). A genetic locus targeted to the nuclear periphery in living cells maintains its transcriptional competence. *J Cell Biol*, 180(1):51-65.

Lallemand-Breitenbach, V, de Thé, H. (2010). PML Nuclear Bodies. *Cold Spring Harb Perspect Biol*, 2(5):a000661.

Lamond AI, Earnshaw WC. (1998). Structure and function in the nucleus. *Science*, 280(5363):547-53.

Lanzuolo C., Roue V., Dekker J., *et al.* (2007). Polycomb response elements mediate the formation of chromosome higher-order structures in the bithorax complex. *Nat Cell Biol*, 9:1167–74.

Laybourn PJ, Kadonaga JT. (1991). Role of nucleosomal cores and histone H1 in regulation of transcription by RNA polymerase II. *Science*, 254:238–45.

Lee AS, Gutiérrez-Arcelus M, Perry GH, Vallender EJ, Johnson WE, Miller GM, Korbel JO, Lee C. (2008). Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. *Hum Mol Genet*, 17(8):1127-36.

Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, Yue P, Zhang Y, Pant KP, Bhatt D, Ha C, Johnson S, Kennemer MI, Mohan S, Nazarenko I, Watanabe C, Sparks AB, Shames DS,

Gentleman R, de Sauvage FJ, Stern H, Pandita A, Ballinger DG, Drmanac R, Modrusan Z, Seshagiri S, Zhang Z. (2010). The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature*, 465(7297):473-7.

Levsky JM, Singer RH. (2003). Fluorescence in situ hybridization: past, present and future. *J Cell Sci*, 116(14):2833-8.

Levy D, Ronemus M, Yamrom B, Lee YH, Leotta A, Kendall J, Marks S, Lakshmi B, Pai D, Ye K, Buja A, Krieger A, Yoon S, Troge J, Rodgers L, Iossifov I, Wigler M. (2011). Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron*, 70(5):886-97.

Li G, Ruan X, Auerbach RK, *et al.* (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 148(1-2):84-98.

Li W, Wu J, Kim SY, Zhao M, Hearn SA, Zhang MQ, Meistrich ML, Mills AA. (2014). Chd5 orchestrates chromatin remodelling during sperm development. *Nat Commun*, 13;5:3812.

Li, S.P. *et al.* (2001). Genome-wide analyses on loss of heterozygosity in hepatocellular carcinoma in Southern China. *J. Hepatol*, 34:840–849.

Liang J, Lacroix L, Gamot A, Cuddapah S, Queille S, Lhoumaud P, Lepetit P, Martin PG, Vogelmann J, Court F, Hennion M, Micas G, Urbach S, Bouchez O, Nöllmann M, Zhao K, Emberly E, Cuvier O. (2014). Chromatin Immunoprecipitation Indirect Peaks Highlight Long-Range Interactions of Insulator Proteins and Pol II Pausing. *Mol Cell*, 53(4):672-81.

Lieberman-Aiden E, *et al.* (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289-93.

Light, W. H., Brickner, D. G., Brand, V. R. & Brickner, J. H. (2010). Interaction of a DNA zip code with the nuclear pore complex promotes H2A.Z incorporation and INO1 transcriptional

memory. *Mol. Cell*, 40:112–125.

Liu, Z. & Garrard, W.T. (2005). Long-range interactions between three transcriptional enhancers, active  $\kappa$  gene promoters, and a 3' boundary sequence spanning 46 kilobases. *Mol. Cell. Biol*, 25:3220–3231.

Losada A., Hirano M., Hirano T. (1998). Identification of *Xenopus* SMC protein complexes required for sister chromatid cohesion. *Genes & Dev*, 12:1986–1997.

Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389:251–260.

Lupski J.R. (1998). Genomic disorders: Structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet*, 14:417–422.

Luthra, R. *et al.* (2007). Actively transcribed GAL genes can be physically linked to the nuclear pore by the SAGA chromatin modifying complex. *J. Biol. Chem*, 282:3042–3049.

Ma H, Samarabandu J, Devdhar RS, Acharya R, Cheng PC, Meng C, Berezney R. (1998). Spatial and temporal dynamics of DNA replication sites in mammalian cells. *J Cell Biol*, 143:1415–1425.

Machyna M, Heyn P, Neugebauer KM. (2013). Cajal bodies: where form meets function. *Wiley Interdiscip Rev RNA*, 4(1):17-34.

Maillet, L., Boscheron, C., Gotta, M., Marcand, S., Gilson, E., Gasser, S.M. (1996). Evidence for silencing compartments within the yeast nucleus: A role for telomere proximity and Sir protein concentration in silencer-mediated repression. *Genes & Dev*, 10:1796–1811.

Malik HS, Henikoff S. (2003). Phylogenomics of the nucleosome. *Nat Struct Biol*, 10:882-891.

Mao YS, Sunwoo H, Zhang B, Spector DL. (2011). Direct visualization of the co-

transcriptional assembly of a nuclear body by noncoding RNAs. *Nat Cell Biol*, 13(1):95-101.

Mao YS, Zhang B, Spector DL. (2011). Biogenesis and function of nuclear bodies. *Trends Genet*, 27(8):295-306.

Marcand, S., Buck, S.W., Moretti, P., Gilson, E., Shore, D. (1996). Silencing of genes at nontelomeric sites in yeast is controlled by sequestration of silencing factors at telomeres by Rap 1 protein. *Genes & Dev*, 10:1297–1309.

Marshall WF, Straight A, Marko JF, Swedlow J, Dernburg A, Belmont A, Murray AW, Agard DA, Sedat JW. (1997). Interphase chromosomes undergo constrained diffusional motion in living cells. *Curr Biol*, 7(12):930-9.

McAnally, A.A., and Yampolsky, L.Y. (2010). Widespread transcriptional autosomal dosage compensation in *Drosophila* correlates with gene expression level. *Genome Biol. Evol*, 2:44–52.

McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, Altshuler DM; International HapMap Consortium. (2006). Common deletion polymorphisms in the human genome. *Nat. Genet*, 38:86–92.

McCarroll, SA., *et al.* (2008). Deletion polymorphism upstream of *IRGM* associated with altered *IRGM* expression and Crohn's disease. *Nature Genet*, 40:1107–1112.

McCarthy, SE., *et al.* (2009). Microduplications of 16p11.2 are associated with schizophrenia. *Nature Genet*, 41:1223–1227.

McHale LK, Haun WJ, Xu WW, Bhaskar PB, Anderson JE, Hyten DL, Gerhardt DJ, Jeddeloh JA, Stupar RM. (2012). Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiol*, 159(4):1295-308.

McKeon, F. (1991). Nuclear lamin proteins: Domains required for nuclear targeting,



assembly, and cell-cycle-regulated dynamics. *Curr. Opin. Cell Biol*, 3:82–86

McKernan, K. J. *et al.* (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res*, 19:1527–1541.

McLean CY, *et al.* (2011) Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature*, 471:216–219.

Medvedev SP, Shevchenko AI, Elisaphenko EA, Nesterova TB, Brockdorff N, Zakian SM. (2008). Structure and expression pattern of Oct4 gene are conserved in vole *Microtus rossiaemeridionalis*. *BMC Genomics*, 9:162.

Mendjan, S., Taipale, M., Kind, J., *et al.* (2006). Nuclear pore components are involved in the transcriptional regulation of dosage compensation in *Drosophila*. *Mol. Cell*, 21:811–823.

Merla G, Howald C, Henrichsen CN, Lyle R, Wyss C, Zobot MT, Antonarakis SE, Raymond A. (2006). Submicroscopic deletion in patients with Williams-Beuren syndrome influences expression levels of the nonhemizygous flanking genes. *Am J Hum Genet*, 79, 332-41.

Meselson, M., Stahl, FW. (1958). The Replication of DNA in *Escherichia coli*. *PNAS*, 44:671–82.

Meyerson M, Gabriel S, Getz G. (2010). Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet*, 11(10):685-96.

Michaelis C., Ciosk R., Nasmyth K. (1997). Cohesins: Chromosomal proteins that prevent premature separation of sister chromatids. *Cell*, 91:35–45.

Midorikawa, Y., *et al.* (2009). Allelic imbalances and homozygous deletion on 8p23.2 for stepwise progression of hepatocarcinogenesis. *Hepatology*, 49:513–522.

Mills RE, Walter K, Stewart C, *et al.*, 1000 Genomes Project. (2011). Mapping copy number

variation by population-scale genome sequencing. *Nature*, 470(7332):59-65.

Mirny LA. (2011). The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res*, 19(1):37-51.

Moissiard G, Cokus SJ, Cary J, Feng S, Billi AC, Stroud H, Husmann D, Zhan Y, Lajoie BR, McCord RP, Hale CJ, Feng W, Michaels SD, Frand AR, Pellegrini M, Dekker J, Kim JK, Jacobsen SE. (2012). MORC family ATPases required for heterochromatin condensation and gene silencing. *Science*, 336(6087):1448-51.

Morgan, T. H., Sturtevant, A. H., Muller, H. J. & Bridges, C. B. (1915). The Mechanism of Mendelian Heredity. *Henry Holt and Company*.

Morse RH. (1989). Nucleosomes inhibit both transcriptional initiation and elongation by RNA polymerase III in vitro. *EMBO J*, 8:2343–51.

Muñoz-Amatriaín M, Eichten SR, Wicker T, Richmond TA, Mascher M, Steuernagel B, Scholz U, Ariyadasa R, Spannagl M, Nussbaumer T, Mayer KF, Taudien S, Platzer M, Jeddloh JA, Springer NM, Muehlbauer GJ, Stein N. (2013). Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. *Genome Biol*, 14(6):R58.

Murrell A., Heeson S., Reik W. (2004). Interaction between differentially methylated regions partitions the imprinted genes *Igf2* and *H19* into parent-specific chromatin loops. *Nat Genet*, 36:889–93.

Murrell, A., Heeson, S. & Reik, W. (2004). Interaction between differentially methylated regions partitions the imprinted genes *Igf2* and *H19* into parent-specific chromatin loops. *Nat. Genet*, 36:889–893.

Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A,

- Fraser P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59-64.
- Naumova N, Imakaev M, Fudenberg G, Zhan Y, Lajoie BR, Mirny LA, Dekker J. (2013). Organization of the mitotic chromosome. *Science*, 342(6161):948-53.
- Noordermeer D, Branco MR, Splinter E, Klous P, van Ijcken W, Swagemakers S, Koutsourakis M, van der Spek P, Pombo A, de Laat W. (2008). Transcription and chromatin organization of a housekeeping gene cluster containing an integrated  $\beta$ -globin locus control region. *PLoS Genet*, 4(3):e1000016.
- Noordermeer D, de Wit E, Klous P, van de Werken H, Simonis M, Lopez-Jones M, Eussen B, de Klein A, Singer RH, de Laat W. (2011). Variegated gene expression caused by cell-specific long-range DNA interactions. *Nat Cell Biol*, 13(8):944-51.
- Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J, Gribnau J, Barillot E, Blüthgen N, Dekker J, Heard E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485(7398):381-5.
- Oakes M, Aris JP, Brockenbrough JS, Wai H, Vu L, Nomura M. (1998). Mutational analysis of the structure and localization of the nucleolus in the yeast *Saccharomyces cerevisiae*. *J Cell Biol*, 143(1):23-34.
- Olins, AL., Olins, DE. (1974). Spheroid chromatin units ([upsilon] bodies). *Science*, 183:330–332.
- Olson MO, Dundr M. (2005). The moving parts of the nucleolus. *Histochem*, 123:203–216.
- O'Neill TE, Roberge M, Bradbury EM. (1992). Nucleosome arrays inhibit both initiation and elongation of transcripts by bacteriophage T7 RNA polymerase. *J. Mol. Biol*, 223:67–78.

Ong CT, Van Bortle K, Ramos E, Corces VG. (2013). Poly(ADP-ribosylation) regulates insulator function and intrachromosomal interactions in *Drosophila*. *Cell*, 155(1):148-59.

Orozco LD, *et al.* (2009) Copy number variation influences gene expression and metabolic traits in mice. *Hum Mol Genet*, 18(21):4118-29.

Oudet, P., Gross-Bellard, M. & Chambon, P. (1975). Electron microscopic and biochemical evidence that chromatin structure is a repeating unit. *Cell*, 4:281–300.

Padilla-Nash HM, Hathcock K, McNeil NE, Mack D, Hoepfner D, Ravin R, Knutsen T, Yonescu R, Wangsa D, Dorritie K, Barenboim L, Hu Y, Ried T. (2012). Spontaneous transformation of murine epithelial cells requires the early acquisition of specific chromosomal aneuploidies and genomic imbalances. *Genes Chromosomes Cancer*, 51(4):353-74.

Palstra RJ, Simonis M, Klous P, Brassat E, Eijkelkamp B, de Laat W. (2008). Maintenance of long-range DNA interactions after inhibition of ongoing RNA polymerase II transcription. *PLoS One*, 3(2):e1661.

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.

Pandey RV, Franssen SU, Futschik A, Schlötterer C. (2013) Allelic imbalance metre (Allim), a new tool for measuring allele-specific gene expression with RNA-seq data. *Mol Ecol Resour*, 13(4):740-745.

Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H., and Soldatov, A. (2009). Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Research*, 37(18):e123.

Patau K., Smith D.W., Therman E., Inhorn S.L., Wagner H.P. (1960). Multiple congenital anomaly caused by an extra autosome. *Lancet*, 1:790–793.

Peric-Hupkes, D., *et al.* (2010). Molecular maps of the reorganization of genome— nuclear lamina interactions during differentiation. *Mol. Cell*, 38:603–613.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, Carter NP, Lee C, Stone AC. (2007). Diet and the evolution of human amylase gene copy number variation. *Nat Genet*, 39(10):1256-60.

Perry GH, *et al.* (2008). Copy number variation and evolution in humans and chimpanzees. *Genome Res*, 18(11):1698–1710.

Perry GH, Tchinda J, McGrath SD, Zhang J, Picker SR, Cáceres AM, Iafrate AJ, Tyler-Smith C, Scherer SW, Eichler EE, Stone AC, Lee C. (2006). Hotspots for copy number variation in chimpanzees and humans. *Proc Natl Acad Sci USA*, 103(21):8006-11.

Perry GH, Yang F, Marques-Bonet T, Murphy C, Fitzgerald T, Lee AS, Hyland C, Stone AC, Hurles ME, Tyler-Smith C, Eichler EE, Carter NP, Lee C, Redon R. (2008). Copy number variation and evolution in humans and chimpanzees. *Genome Res*, 18(11):1698-710.

Phair RD, Misteli T. (2000). High mobility of proteins in the mammalian cell nucleus. *Nature*, 404: 604–609.

Phillips-Cremins, J.E., *et al.* (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, 153:1281–1295.

Pickersgill, H., *et al.* (2006). Characterization of the *Drosophila melanogaster* genome at the nuclear lamina. *Nat. Genet*, 38:1005–1014.

Pinto, D. *et al.* (2010). Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, 466:368–372.

Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, Varela I, Lin ML, Ordóñez GR, Bignell GR, Ye K, Alipaz J, Bauer MJ, Beare D, Butler A, Carter RJ, Chen L, *et al.* (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463(7278):191-6.

Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, Lin ML, Beare D, Lau KW, Greenman C, Varela I, Nik-Zainal S, Davies HR, Ordoñez GR, Mudie LJ, Latimer C, Edkins S, *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*. 2010 Jan 14;463(7278):184-90.

Quan J, Yusufzai T. (2014). The Tumor Suppressor Chromodomain Helicase DNA-binding Protein 5 (CHD5) Remodels Nucleosomes by Unwrapping. *J Biol Chem*, 289(30):20717-20726.

Quénet D, Dalal Y. (2012). The CENP-A nucleosome: a dynamic structure and role at the centromere. *Chromosome Res*, 20(5):465-79.

Quinlan AR, Hall IM. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841-2.

Rabl C. (1885). Über Zellteilung. *Morph Jb*, 10:214-330.

Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*. 470(7333):279-83.

Rae, PM, WW Franke. (1972). The interphase distribution of satellite DNA-containing heterochromatin in mouse nuclei. *Chromosoma*, 39:443-456.

Raices M, D'Angelo MA. (2012). Nuclear pore complex composition: a new regulator of tissue-specific and developmental functions. *Nat Rev Mol Cell Biol*, 13(11):687-99.

Ramayo Caldas Y, Castello A, Pena RN, Alves E, Mercade A, Souza CA, Fernandez AI, Perez-Enciso M, Folch JM. (2010). Copy number variation in the porcine genome inferred from a 60k SNP BeadChip. *BMC Genomics*, 11:593.

Redon R, *et al.* (2006). Global variation in copy number in the human genome. *Nature*, 444(7118):444-54.

Redon R, Rio M, Gregory SG, Cooper RA, Fiegler H, Sanlaville D, Banerjee R, Scott C, Carr P, Langford C, Cormier-Daire V, Munnich A, Carter NP, Colleaux L. (2005). Tiling path resolution mapping of constitutional 1p36 deletions by array-CGH: contiguous gene deletion or "deletion with positional effect" syndrome? *J Med Genet*, 42:166-71.

Rees E, Walters JT, Chambert KD, O'Dushlaine C, Szatkiewicz J, Richards AL, Georgieva L, Mahoney-Davies G, Legge SE, Moran JL, Genovese G, Levinson D, Morris DW, Cormican P, Kendler KS, O'Neill FA, Riley B, Gill M, Corvin A; Wellcome Trust Case Control Consortium, Sklar P, Hultman C, Pato C, Pato M, Sullivan PF, Gejman PV, McCarroll SA, O'Donovan MC, Owen MJ, Kirov G. (2013). CNV analysis in a large schizophrenia sample implicates deletions at 16p12.1 and SLC1A1 and duplications at 1p36.33 and CGNL1. *Hum Mol Genet*, 23(6):1669-76.

Ricard G, Molina J, Chrast J, Gu W, Gheldof N, Pradervand S, Schütz F, Young JI, Lupski JR, Reymond A, Walz K. (2010). Phenotypic consequences of copy number variation: insights from Smith-Magenis and Potocki-Lupski syndrome mouse models. *PLoS Biol*, 8:e1000543.

Rosenfeld JA, Crolla JA, Tomkins S, Bader P, Morrow B, Gorski J, Troxell R, Forster-Gibson C, Cilliers D, Hislop RG, Lamb A, Torchia B, Ballif BC, Shaffer LG. (2010). Refinement of causative genes in monosomy 1p36 through clinical and molecular

cytogenetic characterization of small interstitial deletions. *Am J Med Genet A*, 152A:1951-9.

Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, Harmanci A, Leng J, Bjornson R, Kong Y, Kitabayashi N, Bhardwaj N, Rubin M, Snyder M, Gerstein M. (2011). AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol*, 7:522.

Ryba T, Hiratani I, Lu J, Itoh M, Kulik M, Zhang J, Schulz TC, Robins AJ, Dalton S, Gilbert DM. (2010). Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res*, 20(6):761-70.

S. Kurtz, A. Phillippy, A.L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S.L. Salzberg. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, 5:R12.

Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, *et al.* (2011). Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron*, 70(5):863-85.

Sanyal A, Lajoie BR, Jain G, Dekker J. (2012). The long-range interaction landscape of gene promoters. *Nature*, 489(7414):109-13.

Sasaki H, Matsui Y. (2008). Epigenetic events in mammalian germ-cell development: reprogramming and beyond. *Nat Rev Genet*, 9(2):129-40.

Schaar BT, Chan GK, Maddox P, Salmon ED, Yen TJ. (1997). CENP-E function at kinetochores is essential for chromosome alignment. *J Cell Biol*, 139(6):1373-82.

Schlattl A, Anders S, Waszak SM, Huber W, Korb J. (2011). Relating CNVs to transcriptome data at fine resolution: Assessment of the effect of variant size, type, and overlap with functional regions. *Genome Res*, 21(12):2004–2013.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L,



Graves TA, *et al.* (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science*, 326:1112–1115.

Schulz, EG., Heard, E. (2013). Role and control of X chromosome dosage in mammalian development. *Curr. Opin. Genet. Dev*, 23:109–115.

Schuster-Böckler B, Conrad D, Bateman A. (2010). Dosage sensitivity shapes the evolution of copy-number varied regions. *PLoS One*, 5(3):e9474.

Sebat, J. *et al.* (2007). Strong association of de novo copy number mutations with autism. *Science*, 316:445–449.

Sebat, J., *et al.* (2004). Large-scale copy number polymorphism in the human genome. *Science*, 305:525–528.

Seitan, V.C., *et al.* (2013). Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. *Genome Res*, 23:2066–2077.

Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. (2012). Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell*, 148(3):458-72.

Sharp A.J., Locke D.P., McGrath S.D., Cheng Z., *et al.* (2005). Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet*, 77:78–88.

She X, Cheng Z, Zöllner S, Church DM, Eichler EE. (2008). Mouse segmental duplication and copy number variation. *Nat Genet*, 40(7):909-14.

Shi, J. *et al.* (2013). Role of SWI/SNF in acute leukemia maintenance and enhancer-mediated Myc regulation. *Genes Dev*, 15;27(24):2648-62.

Shopland, L.S. *et al.* (2006). Folding and organization of a contiguous chromosome region according to the gene distribution pattern in primary genomic sequence. *J. Cell Biol*, 174:27–

38.

Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet*, 38(11):1348-54.

Sims RJ 3<sup>rd</sup>., Reinberg D. (2008). Is there a code embedded in proteins that is based on post-translational modifications? *Nat. Rev. Mol. Cell Biol*, 9:815–20.

Slavotinek A, Shaffer LG, Shapira SK. (1999). Monosomy 1p36. *J Med Genet*, 36(9):657-63.

Snaar S, Wiesmeijer K, Jochemsen AG, Tanke HJ, Dirks RW. (2000). Mutational analysis of fibrillarin and its mobility in living human cells. *J Cell Biol*, 151(3):653-62.

Sofueva, S., *et al.* (2013). Cohesin-mediated interactions organize chromosomal domain architecture. *EMBO J*, 32:3119–3129.

Solovei I, Cremer M. (2010). 3D-FISH on cultured cells combined with immunostaining. *Methods Mol Biol*, 659:117-26.

Song F, Chen P, Sun D, Wang M, Dong L, Liang D, Xu RM, Zhu P, Li G. (2011). Cryo-EM study of the chromatin fiber reveals a double helix twisted by tetranucleosomal units. *Science*, 344(6182):376-80.

Soutoglou, E., Misteli, T. (2007). Mobility and immobility of chromatin in transcription and genome stability. *Curr. Opin. Genet. Dev*, 17:435–442.

Sparvoli E, Levi M, Rossi E. (1994). Replicon clusters may form structurally stable complexes of chromatin and chromosomes. *J Cell Sci*, 107:3097–3103.

Spector DL, Lamond AI. (2011). Nuclear speckles. *Cold Spring Harb Perspect Biol*, 1;3(2).

Spector DL, Schrier WH, Busch H. (1983). Immunoelectron microscopic localization of snRNPs. *Biol Cell*, 49:1–10.

- Spector DL. (1993). Macromolecular domains within the cell nucleus. *Annu Rev Cell Biol*, 9:265-315.
- Spector, D.L. (2003). The dynamics of chromosome organization and gene regulation. *Annu. Rev. Biochem*, 72: 573–608
- Spilianakis, C.G. & Flavell, R.A. (2004). Long-range intrachromosomal interactions in the T helper type 2 cytokine locus. *Nat. Immunol*, 5:1017–1027.
- Splinter E, de Wit E, Nora EP, Klous P, van de Werken HJ, Zhu Y, Kaaij LJ, van Ijcken W, Gribnau J, Heard E, de Laat W. (2011). The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes Dev*, 25(13):1371-83.
- Stack, SM., Brown, DB., Dewey, WC. (1977). Visualization of interphase chromosomes. *Journal of Cell Science*, 26:281–299.
- Stankiewicz P., Lupski J.R. (2002). Genome architecture, rearrangements and genomic disorders. *Trends Genet*, 18:74–82.
- Stavenhagen, J.B., Zakian, V.A. (1994). Internal tracts of telomeric DNA act as silencers in *Saccharomyces cerevisiae*. *Genes & Dev*, 8:1411–1422.
- Stefansson, H., *et al.* (2008). Large recurrent microdeletions associated with schizophrenia. *Nature*, 455:232–236.
- Stephens PJ, McBride DJ, Lin ML, Varela I, Pleasance ED, Simpson JT, Stebbings LA, Leroy C, Edkins S, *et al.* (2009). Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, 462(7276):1005-10.
- Strahl BD, Allis CD. (2000). The language of covalent histone modifications. *Nature*, 403:41–45.
- Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de

Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavaré S, Deloukas P, Hurles ME, Dermitzakis ET. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315(5813):848-53.

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J; 1000 Genomes Project, Eichler EE. (2010). Diversity of human copy number variation and multicopy genes. *Science*, 330(6004):641-6.

Sutton, W. S. (1902). On the morphology of the chromosome group in *Brachystola magna*. *Biological Bulletin*, 4:24–39.

Swift H. (1959). Studies on nuclear fine structure. *Brookhaven Symp Biol*, 12:134–152.

Szenker E, Ray-Gallet D, Almouzni G. (2011). The double face of the histone variant H3.3. *Cell Res*, 21(3):421-34.

Taddei, A. *et al.* (2006). Nuclear pore association confers optimal expression levels for an inducible yeast gene. *Nature*, 441:774–778.

The 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073.

The International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature*, 437:1299-1320.

Thoma F, Koller T. (1977). Influence of histone H1 on chromatin structure. *Cell*, 12(1):101–7.

Thompson, J.S., Johnson, L.M., Grunstein, M. (1994). Specific repression of the yeast silent mating locus HMR by an adjacent telomere. *Mol. Cell. Biol*, 14:446–455.

Tiwari VK, Cope L, McGarvey KM, Ohm JE, Baylin SB. (2008). A novel 6C assay uncovers Polycomb-mediated higher order chromatin conformations. *Genome Res*, 18(7):1171-9.

Todaro GJ, Green H. (1963). Quantitative studies of the growth of mouse embryo cells in culture and their development into established lines. *J. Cell Biol*, 17:299-313.

Tolhuis B, Blom M, Kerkhoven RM, Pagie L, Teunissen H, Nieuwland M, Simonis M, de Laat W, van Lohuizen M, van Steensel B. (2011). Interactions among Polycomb domains are guided by chromosome architecture. *PLoS Genet*, 7(3):e1001343.

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.

Trask B.J., Friedman C., Martin-Gallardo A., Rowen L., Akinbami C., Blankenship J., Collins C., Giorgi D., Iandonato S., Johnson F., et al. (1998). Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. *Hum. Mol. Genet*, 7:13–26.

Tumbar T, Belmont AS. (2001). Interphase movements of a DNA chromosome region modulated by VP16 transcriptional activator. *Nat Cell Biol*, 3(2):134-9.

Turro E, Su SY, Gonçalves Â, Coin LJ, Richardson S, Lewin A. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol*. 2011;12(2):R13.

Tuzun E., Sharp A.J., Bailey J.A., Kaul R., et al. (2005). Fine-scale structural variation of the human genome. *Nat. Genet*, 37:727–732.

Umbarger MA, Toro E, Wright MA, Porreca GJ, Baù D, Hong SH, Fero MJ, Zhu LJ, Marti-Renom MA, McAdams HH, Shapiro L, Dekker J, Church GM. (2011). The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation. *Mol Cell*, 44(2):252-64.

van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, Dekker

J, Lander ES. (2010). Hi-C: a method to study the three-dimensional architecture of genomes. *J Vis Exp*, 6;(39).

van de Werken HJ, Landan G, Holwerda SJ, Hoichman M, Klous P, Chachik R, Splinter E, Valdes-Quezada C, Oz Y, Bouwman BA, Verstegen MJ, de Wit E, Tanay A, de Laat W. (2012). Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nat Methods*, 9(10):969-72

van de Werken HJ, de Vree PJ, Splinter E, Holwerda SJ, Klous P, de Wit E, de Laat W. (2012). 4C technology: protocols and data analysis. *Methods Enzymol*, 513:89-112.

van Steensel B, Henikoff S. (2000). Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat Biotechnol*, 18(4):424-8.

Vaquerizas, J. M. *et al.* (2010). Nuclear pore proteins Nup153 and Megator define transcriptionally active regions in the Drosophila genome. *PLoS Genet*. 6(2):e1000846.

Vergnes L, Péterfy M, Bergo MO, Young SG, Reue K. (2004). Lamin B1 is required for mouse development and nuclear integrity. *Proc Natl Acad Sci USA*, 101(28):10428-33.

Vernimmen D, De Gobbi M, Sloane-Stanley JA, Wood WG, Higgs DR. (2007). Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. *EMBO J*, 26(8):2041-51.

Vogel, M.J., Peric-Hupkes, D. & van Steensel, B. (2007). Detection of in vivo protein-DNA interactions using DamID in mammalian cells. *Nat. Protoc*, 2:1467–1478.

Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, Lajoie BR, Protacio A, Flynn RA, Gupta RA, *et al.* (2011). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature*, 472(7341):120-4.

Watson ML. (1959). Further observations on the nuclear envelope of the animal cell. *J*

*Biophys Biochem Cytol*, 6:147–156.

Watson, JD., Crick, F., (1953). A structure for deoxyribose nucleic acid. *Nature*, 171(4356): 737–738.

Weidtkamp-Peters S, Lenser T, Negorev D, Gerstner N, Hofmann TG, Schwanz G, Hoischen C, Maul G, Dittrich P, Hemmerich P. (2008). Dynamics of component exchange at PML nuclear bodies. *J Cell Sci*, 121(16):2731-43.

Weischenfeldt J., Symmons O., Spitz F., Korbel JO. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet*, 14(2):125-38.

Wente SR, Rout MP. (2010). The nuclear pore complex and nuclear transport. *Cold Spring Harb Perspect Biol*, 2(10):a000562.

Wheway, G., Abdelhamed, Z., Natarajan, S., Toomes, C., Inglehearn, C., and Johnson, C.A. (2013). Aberrant Wnt signalling and cellular over-proliferation in a novel mouse model of Meckel-Gruber syndrome. *Dev. Biol*, 377:55–66.

Wilkins, M. H. F., Stokes, A. R., Wilson, H. R. (1953). Molecular structure of deoxypentose nucleic acids. *Nature*, 171:738–740.

Wood KW, Sakowicz R, Goldstein LS, Cleveland DW. (1997). CENP-E is a plus end-directed kinetochore motor required for metaphase chromosome alignment. *Cell*, 91(3):357-66.

Woodcock CL, Frado LL, Rattner JB. (1984). The higher-order structure of chromatin: evidence for a helical ribbon arrangement. *J Cell Biol*, 99:42-52

Worcel A, Strogatz S, Riley D. (1981). Structure of chromatin and the linking number of DNA. *Proc Natl Acad Sci USA*, 78(3):1461-5.

Workman JL. (2006). Nucleosome displacement in transcription. *Genes Dev*, 20:2009–17.

Worman HJ. (2012). Nuclear lamins and laminopathies. *J Pathol*, 226(2):316-25.

Wu C, Bassett A, Travers A. (2007). A variable topology for the 30-nm chromatin fibre. *EMBO Rep*, 8(12):1129-34.

Wu F, Yao J. (2013). Spatial compartmentalization at the nuclear periphery characterized by genome-wide mapping. *BMC Genomics*, 14:591.

Wu TD, Nacu S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873-81.

Yaffe E, Tanay A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet*, 43(11):1059-65.

Yanagisawa H, Hammer RE, Richardson JA, Williams SC, Clouthier DE, Yanagisawa M. (1998). Role of Endothelin-1/Endothelin-A receptor-mediated signaling pathway in the aortic arch patterning in mice. *J Clin Invest*, 102(1):22-33.

Yanagisawa H, Yanagisawa M, Kapur RP, Richardson JA, Williams SC, Clouthier DE, de Wit D, Emoto N, Hammer RE. (1998). Dual genetic pathways of endothelin-mediated intercellular signaling revealed by targeted disruption of endothelin converting enzyme-1 gene. *Development*, 125:825–836.

Yang L, Luquette LJ, Gehlenborg N, Xi R, Haseley PS, Hsieh CH, Zhang C, Ren X, Protopopov A, Chin L, Kucherlapati R, Lee C, Park PJ. (2013). Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell*, 153(4):919-29.

Yang, TL., *et al.* (2008). Genome-wide copy-number-variation study identified a susceptibility gene, *UGT2B17*, for osteoporosis. *Am. J. Hum. Genet*, 83:663–674.

Yao X, Abrieu A, Zheng Y, Sullivan KF, Cleveland DW. (2000). CENP-E forms a link between attachment of spindle microtubules to kinetochores and the mitotic checkpoint. *Nat*



*Cell Biol*, 2(8):484-91.

Zaveri HP, Beck TF, Hernández-García A, Shelly KE, Montgomery T, van Haeringen A, Anderlid BM, Patel C, Goel H, Houge G, Morrow BE, Cheung SW, Lalani SR, Scott DA. (2014). Identification of critical regions and candidate genes for cardiovascular malformations and cardiomyopathy associated with deletions of chromosome 1p36. *PLoS One*, 9(1):e85600.

Zhang B, Kirov S, Snoddy J. (2005). WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res*, 1(33):W741-8.

Zhang H, Zhai Y, Hu Z, Wu C, Qian J, Jia W, Ma F, Huang W, Yu L, Yue W, Wang Z, Li P, Zhang Y, Liang R, Wei Z, Cui Y, Xie W, Cai M, Yu X, Yuan Y, Xia X, Zhang X, Yang H, Qiu W, Yang J, Gong F, Chen M, Shen H, Lin D, Zeng YX, He F, Zhou G. (2010). Genome-wide association study identifies 1p36.22 as a new susceptibility locus for hepatocellular carcinoma in chronic hepatitis B virus carriers. *Nat Genet*, 42(9):755-8.

Zhang Y., McCord RP., Ho YJ., Lajoie BR., Hildebrand DG., Simon AC., Becker MS., Alt FW., Dekker J. (2012). Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell*, 148(5):908-21.

Zhao Z, Tavoosidana G, Sjolinder M, Gondor A, Mariano P, Wang S, Kanduri C, Lezcano M, Sandhu KS, Singh U, *et al.*. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet*, 38(11):1341-7.

Zhu X, Zhang Y, Wang J, Yang JF, Yang YF, Tan ZP. (2013). 576 kb deletion in 1p36.33-p36.32 containing SKI is associated with limb malformation, congenital heart disease and epilepsy. *Gene*, 528(2):352-5.

Zink D, Bornfleth H, Visser A, Cremer C, Cremer T. (1999). Organization of early and late replicating DNA in human chromosome territories. *Exp Cell Res*, 247:176–188.

Zuela N, Bar DZ, Gruenbaum Y. (2012). Lamins in development, tissue maintenance and stress. *EMBO Rep*, 13(12):1070-8.

Zuin, J., *et al.* (2013). Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc. Natl. Acad. Sci. USA*, 111:996–1001.

## Author contributions

Cinthya Zepeda and David Spector conceived the study, designed the experiments, and wrote results.

Cinthya Zepeda performed the experiments and contributed in the analysis of 3D DNA FISH, PE-4Cseq, and RNA-Seq data.

Swagatam Mukhopadhyay developed the quantitative PE-4Cseq pipeline for the analysis of chromosome conformation capture data, analyzed PE-4Cseq viewpoints, and wrote results.

Nathalie Harder developed the Image plugin for the analysis of 3D DNA FISH images and analyzed data.

Hesed Padilla-Nash performed the spectral karyotyping analysis of deletion and wild type MEF lines.

Erik Splinter taught me the 4C-seq technique, and provided comments on viewpoint design. Elzo de Wit and Wouter de Laat provided feedback with PE-4Cseq experiments analysis and interpretation.

Emilie Wong performed the allele-specific analysis of RNA-Seq data.

Melanie Eckersley-Maslin prepared the RNA-Seq libraries.

Alea Mills shared the *df/dp* chromosomally engineered mouse models.