2014-11-18

The Three-Dimensional Structure of the Cystic Fibrosis Locus: A Dissertation

Emily M. Smith
University of Massachusetts Medical School

THE THREE-DIMENSIONAL STRUCTURE OF THE CYSTIC FIBROSIS LOCUS

A Dissertation Presented

By

EMILY MALINDA SMITH

Submitted to the Faculty of the

University of Massachusetts Graduate School of Biomedical Sciences, Worcester

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

NOVEMBER 18, 2014

BIOMEDICAL SCIENCES

THE THREE-DIMENSIONAL STRUCTURE OF THE CYSTIC FIBROSIS LOCUS

Dissertation Presented By

Emily Malinda Smith

The signatures of the Dissertation Committee signify completion and approval as to style and content of the Dissertation

[redacted]

Job Dekker, Ph.D., Thesis Advisor

[redacted]

Anthony Imbalzano, Ph.D., Member of Committee

[redacted]

Jeanne Lawrence, Ph.D., Member of Committee

[redacted]

Michael Lee, Ph.D., Member of Committee

[redacted]

Benjamin Raby, M.D., MPH, Member of Committee

The signature of the Chair of the Committee signifies that the written dissertation meets the requirements of the Dissertation Committee

[redacted]

Sean [redacted] Committee

The signature of the Dean of the Graduate School of Biomedical Sciences signifies that the student has met all graduation requirements of the school.

[redacted]

Anthony Carruthers, Ph.D.,
Dean of the Graduate School of Biomedical Sciences

Interdisciplinary Graduate Program

November 18, 2014

ACKNOWLEDGEMENTS

I would like to acknowledge a number of people who helped me personally and professionally while I completed the work for this degree.

Job Dekker is a fantastic PI and mentor. He has a deep knowledge of the field and continues to remain on the cutting edge of experimental techniques. The projects that come from his lab make waves throughout the scientific community. I am grateful for his assistance throughout this process. Without Job's vision my project would not be what it is today. I am a better scientist because of what he has taught me.

The members of the Dekker Lab, past and present, have also contributed to my growth as a scientist and as a person. I would like to thank Nele Gheldof, who began this *CFTR* project while she was in the lab as a postdoc. Thanks to Adrianna Geldart (formerly Miele) for welcoming me into the lab and teaching me many things. Thanks to Ninke VanBerkum for teaching me 3C, and to Amartya Sanyal and Ye Zhan who helped me with 5C. I also thank Bryan Lajoie and Gaurav Jain for all their help with the analysis side of this project – it would not have been done without you and your programming skills. Thanks also to Johan Gibcus, Noam Kaplan, Natasha Naumova, Rachel Patton McCord, Jenn Phillips-Cremins and Jon Belton for your great support as colleagues. You all have made the lab atmosphere a great place to work and through our shared experiences we have all become friends. Who could ask for more?

Additional thanks go out to my extended family at Wesley United Methodist Church, who welcomed Phil and me with open arms. You have prayed for me, celebrated with me and supported me throughout this journey. There are too many names to list. You are all special and I thank God we were able to walk together for a time. To Pastors Vicki Woods, Lisa Bruget-Cass and Shandi Mawokomatanda, who have helped guide me in my spiritual walk these past 7 years, I thank you. Special thanks is needed for all the youth that have welcomed Phil and me as their youth group leaders and mentors along the way – you have made every Wednesday night so exceptional. I continue to learn from you all: Munah, Isaac, Stacy, Elaine, Natasha, Josephine, Abby, Danielle, Will, Eleanor, Rachel, Rebecca, Gabby, Joseph, Kendra, Kristina, Akosua, Olivia and Amelia – you are all such special people. I've loved the trips we've taken, the events that we've hosted, the mission work we've done, the long rides in cars and busses, and the advice about African hair. We are God's people – we are his hands and feet.

Thank you Russell Ulbrich and Kristine Reed for your friendliness, your hospitality, your random dinner invitations, your lovely house and patio, your intriguing conversations and your companionship. It's been so great.

Thank you to my in-laws and my sister in law, Carol, John and Megan Smith. Carol, you've been great to talk to as I go through this process, and you've given me good solid advice. John, thanks for bringing the family to visit – it's always a pleasure. I hope we will soon have a house that you can build all

sorts of fun trinkets for. Meg, I'm glad we've gotten to know each other in the brief visits we've had, and I hope there are many more in the future. I wish you could stamp my thesis for me.

Thanks to my mom and dad, Becky and Tom Wisnieski, for supporting me and paying for me to go to Purdue, even though it was expensive and far away. I know when I decided to come to Massachusetts for graduate school you weren't sure it was the best decision, but you continued to support me. You understood when I couldn't always come home for as long as we both wanted, but we had fun anyway, visiting on Cape Cod and in Vermont too. We made great memories together. Thanks again for loving me like you do. Mom, I love your determination to understand my work and explain it to other people. You taught me to love learning, and more importantly you taught me to love teaching others. Dad, you taught me the importance of good hard work, which has been so important in this process. Mark, you're a great brother. I wish we lived closer so we could hang out as adults, which I hope we'd enjoy just as much as playing together when we were younger. Mary Rosso, you are a wonderful person and a fantastic grandma. To Bruce Rosso and Bill and Sylvia Wisnieski, you will always be in my heart.

Thank you so much Meredith Edwards, my true friend and confidant. It's been really hard living so far away from you. I've cherished our friendship despite the distance. You've got my back forever. I love you.

Phil Smith, you are the best husband I could ever have. You have been so patient with me as I struggled through this PhD. You have always given me

space when I need it. You have put up with my frustrations, my tendency to not cook, my need to escape into literature, and my random desires for creativity. You've never complained. I am delighted we have come through to the other side, and I can't wait to see where life takes us next. No matter what we do, we're going to do it together, and that thought comforts me more than I can say. I love you forever. You are my perfect companion.

ABSTRACT

The three dimensional structure of the human genome is known to play a critical role in gene function and expression. I used chromosome conformation capture (3C) and 3C-carbon copy (5C) techniques to investigate the three-dimensional structure of the cystic fibrosis transmembrane conductance regulator (CFTR) locus. This is an important disease gene that, when mutated, causes cystic fibrosis. 3C experiments identified four distinct looping elements that contact the CFTR gene promoter only in CFTR-expressing cells. Using 5C, I expanded the region of study to a 2.8 Mb region surrounding the CFTR gene. The 5C study shows 7 clear topologically associating domains (TADs) present at the locus, identical in all five cell lines tested, regardless of gene expression status. CFTR and all its known regulatory elements are contained within one TAD, suggesting TADs play a role in constraining promoters to a local search space. The four looping elements identified in the 3C experiment and confirmed in the 5C experiment were then tested for enhancer activity using a luciferase assay, which showed that elements III and IV could act as enhancers. These elements were tested against a library of human transcription factors in a yeast one-hybrid assay to identify potential binding proteins. Element III gave two strong candidates, TCF4 and LEF1. A literature search supported these transcription factors as playing a role in CFTR gene expression. Overall, this work represents a model locus that can be used to test important questions regarding the role of three dimensional looping on gene expression.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS, ABBREVIATIONS OR NOMENCLATURE

ΔF508 – deletion of the amino acid phenylalanine at position 508 of the CFTR protein

3AT – 3-Amino-1,2,4-triazole

3C – Chromosome Conformation Capture

3D – three dimensional

5C – Chromosome Conformation Capture Carbon Copy

ABC – ATP-binding cassette

AD – activation domain

AP-1 – activator protein 1

ATCC – American Type Culture Collection

ATF-1 – activating transcription factor 1 protein

ATP – adenosine triphosphate

BAC – bacterial artificial chromosome

bp – base pair

BPES – blepharophimosis syndrome

BrUTP – 5-bromouridine 5'-triphosphate

BSA – bovine serum albumin

cAMP – cyclic adenosine monophosphate

cDNA – complementary DNA

C/EBP – CCAAT-enhancer binding protein

C. elegans – *Caenorhabditis elegans*

CF – Cystic Fibrosis

CFTR – cystic fibrosis transmembrane conductance regulator gene

CFTR – cystic fibrosis transmembrane conductance regulator protein

ChIP – chromatin immunoprecipitation

Cl- – chloride ion

c-Myc – myelocytomatosis viral oncogene

CRE – cAMP response element

CREB – The cAMP response element binding protein

CRISPR – clustered regularly interspaced short palindromic repeats

CTCF – CCCTC-binding factor

DamID – DNA adenine methyltransferase identification

DHS – DNase 1 hypersensitive site(s)

DMEM – Dulbecco's Modified Eagle's medium

DNA – deoxyribonucleic acid

dNTP – deoxyribonucleotides

DTT – dithiothreitol

EDTA – ethylediamine tetraacetic acid

ENCODE – Encyclopedia of DNA Elements

ELKF – erythroid Krüppel-like Factor

ESRRB – estrogen-related receptor beta

ESRRG – estrogen-related receptor gamma

FBS – fetal bovine serum

FISH – fluorescence *in situ* hybridization

FOG1 – friend of GATA protein 1

FOXA1 – forkhead box protein A1

FOXL2 – forkhead box L2 gene

GATA1 – GATA-binding factor 1

gRNA – guide RNA

GWAS – genome-wide association studies

H_2SO_4 – sulfuric acid

H3 – histone 3

HCL – hydrogen chloride

HIS3 - imidazoleglycerol-phosphate dehydratase

HNF1α – hepatocyte nuclear factor 1α

HoxB – homeobox B gene cluster

HPLC –high performance liquid chromatography

HPRT1 –hypoxanthine phosphoribosyltransferase 1

hTF – human transcription factor

K4 – lysine 4

kb – kilobase (1000 DNA base pairs)

LacZ - β-galactosidase

LAD – lamin-associated domain

LCR – locus control region

LEF1 – lymphoid enhancer-binding factor-1

M – molar

Mb – megabase (1 million DNA base pairs)

Me1 – mono-methylation

Me2 – di-methylation

Me3 – tri-methylation

MEM – minimal essential media Eagle

$MgCl_2$ – magnesium chloride

$MgSO_4$ – magnesium sulfate

Na+ – sodium ion

NaCl – sodium chloride

NEB – New England Biolabs

$(NH_4)_2SO_4$ – ammonium sulfate

NF-κB – nuclear factor kappa-light-chain-enhancer of activated B cells protein

NR5A1 – nuclear receptor subfamily 5, group A, member 1

NR5A2 – nuclear receptor subfamily 5, group A, member 2

PAX6 – paired box protein gene

PBM – protein-binding microarray

PCR – polymerase chain reaction

RNA – ribonucleic acid

RNAi – RNA interference

RPMI – Roswell Park Memorial Institute medium

RT-PCR – reverse transcriptase polymerase chain reaction

SCN- – Thiocyanate

SELEX - systematic evolution of ligands by exponential enrichment

Shh – sonic hedgehog gene

siRNA – small interfering RNA

SNPs – single nucleotide polymorphisms

SP1 – specificity protein 1

ssDNA – salmon sperm DNA

TAD – topologically associating domains

TAL1 – T-cell acute lymphocytic leukemia protein 1

TCF4 (TCF7L2) – transcription factor 4

TE – Tris and EDTA

THRB - thyroid hormone receptor, beta

TF – transcription factor

Tm – melting temperature

Tris – hydroxymethly amino methane-chloride

TSS – transcription start site

URA3 – orotidine 5'-phosphate decarboxylase

YAC – yeast artificial chromosome

Y1H – yeast one-hybrid

YEPD - yeast extract peptone dextrose

YY1 – ying yang 1 protein

X-gal – 5-bromo-4-chloro-3-indolyl-β-d-galactoside

PREFACE

This thesis contains published works in the chapters. Each chapter cites (when necessary) published works contained therin, and details the contribution of the authors. These published works are:

Nele Gheldof, Emily M. Smith, Tomoko M. Tabuchi, Christoph M. Koch, Ian Dunham, John A. Stamatoyannopoulos, and Job Dekker. “Cell-type-specific Long-range Looping Interactions Identify Distant Regulatory Elements of the *CFTR* Gene.” Nucleic Acids Research 38, no. 13 (July 2010): 4325–4336. doi:10.1093/nar/gkq175.

Natalia Naumova, Emily M. Smith, Ye Zhan, and Job Dekker. “Analysis of Long-range Chromatin Interactions Using Chromosome Conformation Capture.” Methods 58, no. 3 (November 2012): 192–203. doi:10.1016/j.ymeth.2012.

Emily M. Smith, Bryan Lajoie, Gaurav Jain and Job Dekker. “Regulation of gene expression around the *CFTR* locus through cell type-specific chromatin looping in the context of invariant topologically associating domains.” Manuscript in preparation.

In Chapter 1, Figure 1.1 is adapted from a slide made by Job Dekker.

In Chapter 3, the data in Figure 3.6 was generated by Nele Gheldof. Emily M Smith created the figure.

CHAPTER I

INTRODUCTION

Cystic Fibrosis – An Uncommonly Common Recessive Disease

"*Wehe dem Kind, das beim Kuß auf die Stirn salzig schmekt, er ist verhext und muss bald sterbe"*– Beware of the child who tastes salty from a kiss on the brow, for it is cursed and soon must die. This medieval German quote refers to the disease Cystic Fibrosis (CF). Thought to be first accurately described in 1595 by Pieter Pauw of the Netherlands, CF is a recessive genetic disorder that disproportionately affects those of European descent (Quinton, 1999). Although the disease itself has been recognized in some form for centuries, only recently have major advances in treatment significantly extended patient life-span. The *CFTR* gene, responsible for causing CF when mutated, was the first human gene to be cloned (Riordan et al., 1989). Upon its discovery, many believed it was only a matter of time until scientists could use gene therapy to insert a healthy copy of the *CFTR* gene into diseased tissues. This goal proved elusive. There is still no cure.

Cystic fibrosis is characterized by a number of clinical phenotypes (listed at the Cystic Fibrosis Foundation, www.cff.org). The most noticeable symptom, as referenced by the quote above, is an increased level of sweat chloride, leading to a salty taste of patient sweat (Kumar et al., 2012). A major symptom is mucus buildup in the airway of an affected individual, leading to difficulty breathing and chronic lung infections. These infections are caused by the pathogens *Haemophilus influenzae, Staphylococcus aureus*, and *Pseudomonas aeruginosa* (Saiman, 2004). Biofilm formation of *P. aeruginosa* is the most

common infection. Many CF patients continually take antibiotics to fight bacterial infections prophylactically. Eventually patients may require a lung transplant from a healthy, non CF individual (Belkin et al., 2006). The condition also produces problems in the gastrointestinal tract. Analogous to mucus buildup in the airway, pancreatic secretions also accumulate, causing a barrier to digestive enzymes entering the gut. This leads to pancreatic insufficiency and a lessened ability to properly digest food (Kumar et al., 2012). Enzymatic supplements are used to treat this problem. In the reproductive tract, congenital absence of the vas deferens is present in many male cases of CF (Chen et al., 2012). Indeed in many infertility cases, including congenital absence of the vas deferens, the patient is found to harbor a mild CF mutation (Du et al., 2014; Jiang et al., 2013; Sharma et al., 2014; Yu et al., 2012). However CF patients can have children using *in vitro* fertilization methods (McCallum et al., 2000).

The Genetic Basis of Cystic Fibrosis

Cystic fibrosis is a single gene recessive disorder. The *CFTR* gene is highly expressed in the epithelial cells of the airway and pancreas, as well as in the colon, sweat glands, liver, and reproductive tissues (Riordan et al., 1989; Trezise and Buchwald, 1991). About 1/25 to 1/30 people of European descent are carriers of a *CFTR* mutant allele, which is a strikingly high number for a recessive disease (Cystic Fibrosis Foundation, 2014). There is a hypothesis that this high carrier frequency is tolerated because of some type of heterozygote

advantage, perhaps similar to the advantage given to heterozygote carriers of the sickle cell anemia allele against malaria (Allison, 1954). Some studies have shown links between mutant *CFTR* and the incidence of cholera, typhoid fever, and lactose tolerance (Gabriel et al., 1994; Modiano et al., 2006; Pier et al., 1998). However, cholera is not limited to Europe, and further work into this theory has not shown any link to cholera resistance (Cuthbert et al., 1995; Högenauer et al., 2000). Additionally, the idea that *CFTR* mutant alleles provide resistance to typhoid fever is also refuted by the prevalence of typhoid fever across the world, while CF is limited to Europe (and by extension the Americas). An increase to lactose tolerance does correlate with the high European prevalence of CF, but further research must be done to prove any correlation. A fourth idea, that CF heterozygosity confers resistance to heavy metal poisoning, coincides with research showing that mutant CFTR protein changes the redox potential inside cells, leading to a greater ability to sequester toxic lead used from the time of the Roman empire (Childers et al., 2007). None of these theories have yet been confirmed.

The location of *CFTR* was discovered to be in a 30 Mb window on chromosome 7 in 1985 by Lap-Chee Tsui (Tsui et al., 1985). Tsui then collaborated with Francis Collins, who came up with a chromosome jumping technique, to more precisely identify where the gene was located (Collins et al., 1987). In 1989, three papers published in Science showed not only the exact location and length of *CFTR*, but identified the length of the CFTR protein and in

which tissues it was expressed, predicted its function, and pinpointed the now famous phenylalanine deletion at position 508 (ΔF508) as the key mutation in over 70% of CF patients (Kerem et al., 1989; Riordan et al., 1989; Rommens et al., 1989). This groundbreaking work was met with excitement about potential treatments, including gene therapy, as a total cure for this disease.

A Defect in CFTR Results in a Multitude of Complications

The *CFTR* gene produces the CFTR protein, which is an ABC family transmembrane chloride ion (Cl-) channel activated by cyclic AMP (cAMP) (Bear et al., 1992; Rich et al., 1990). Normal CFTR allows Cl- to flow down its electrochemical gradient, moving Cl- and thiocyanate (SCN-) ions out of the cell. Sodium (Na+) ions follow, creating an increase in electrolytes in the mucosal layer of the airway and digestive tract. “A number of other cellular functions have been ascribed to CFTR: it down-regulates transepithelial sodium transport, in particular the epithelial sodium channel; it also regulates calcium-activated chloride channels and potassium channels and may serve important functions in exocytosis and the formation molecular complexes in the plasma membrane (Rowe et al., 2005).” When this pump is impaired or missing, the symptoms of CF appear, though there is debate about how exactly the many symptoms of CF are caused by absence of the CFTR transporter. The salty-tasting sweat may be the easiest symptom to explain. Human sweat contains some salt (NaCl) but a majority of it is re-absorbed in the sweat gland duct. In the sweat glands of a CF

patient, there is an imbalance of Cl- and Na+ in the ductal lumen of the gland due to the defect in CFTR. Na+ can be taken into the cell, but Cl- cannot, increasing the negative charge, and thus the amount of NaCl, in the lumen and consequently in the excreted sweat of the patient (Rowe et al., 2005). The heavy mucous coating the airways is thought to be the result of a dehydration of the mucosal layer, resulting in a thicker, sticky mucus that is harder to clear (Brennan, 2008). In the respiratory tract, it has been shown that airway epithelial cells homozygous for the ΔF508 deletion failed to produce thiocyanate, a component of an antimicrobial pathway, contributing to the growth of *S. aureus* and *P. aeruginosa* (Moskwa et al., 2007). There is also evidence that CFTR is responsible for the transport of glutathione, which affects the redox state of the cell. If this redox state is imbalanced, a multitude of downstream effects are observed, including mis-regulation of zinc-dependent enzymes, improper gene regulation, hyper inflammation, altered pH, and imbalance of fatty acids (Childers et al., 2007). Additionally, mutations in *CFTR* can result in congenital bilateral absence of the vas deferens and male infertility, even if CF has not been clinically diagnosed (Yu et al., 2012). The influence of losing one transport channel in the cell is far reaching, with many negative consequences, leading to the complex CF phenotype.

Cystic Fibrosis and the Promise of Gene Therapy

The study of molecular biology holds great promise for the potential treatment of genetic disorders. Notably, cystic fibrosis has been at the forefront of this type of groundbreaking research. *CFTR* is an ideal candidate, since only one good gene copy is needed for a fully functional protein to reach the cell membrane. Over 1500 mutations have been found in the *CFTR* gene (Bobadilla et al., 2002). The most severe mutation, ΔF508, occurs in over 70% of CF cases (Riordan et al., 1989). This mutation causes incorrect folding and subsequent degradation of the CF protein, so it never reaches the cell membrane. Intriguingly, overexpression of this allele causes overloading of the protein degradation pathway, and does allow some of the mutated protein to arrive at the cell membrane, where it is a functional transporter (Cheng et al., 1995). It has also been shown that the ΔF508 mutation can be rescued via transcomplementation of only a fragment of wild-type CFTR, producing a functional channel at the cell membrane (Cormet-Boyaka et al., 2004). This hints at potential therapy – if one can rescue the native allele, even though it is mutated, one can rescue the CF phenotype. Since CF is a single-gene recessive disorder, it has been at the forefront of potential gene therapy treatments for decades. Simply replacing the mutated gene with a normal copy, or overexpressing the commonly mutated version, would cure this disease. This approach was taken on human cell lines, and it was found that inserting *CFTR* cDNA into cells with mutated *CFTR* created a functional CFTR protein (Drumm et al., 1990; Krauss et al., 1992; Rich et al., 1990). A recent study published in late

2013 used CRISPR/Cas9 technology to “correct the *CFTR* locus by homologous recombination in cultured intestinal stem cells of CF patients (Schwank et al., 2013).” CRISPR/Cas9 repaired both mutant alleles in the majority of clones in this study and created functional CFTR protein. The authors were able to grow primary intestinal organoids from the corrected cells in which CFTR function was normal. Though the 2013 paper used modern technology in clonal cells that can grow into an organ ready for transplant, the dream of CF gene therapy is still that – a dream. Correctly targeting all the cells necessary for a cure to work is a huge barrier for gene therapy. Although in theory reaching the airway cells by a nasal spray is relatively simple, getting the corrected gene into cells of the gut and reproductive system is a much larger hurdle. The challenge of creating this type of cure *in vivo*, and not just in the petri dish, remains.

Animal Models for Cystic Fibrosis

Three years after the Cystic Fibrosis gene was cloned, a mouse model was created containing a truncated form of the *CFTR* gene (stopping at exon 10) (Snouwaert et al., 1992). This mouse produced similar symptoms to CF patients including failure to thrive and alteration of mucous glands, however most mice died within 40 days of birth, making long-term studies unfeasible. Since then, 11 different CF mice models have been created (reviewed in Guilbault et al., 2007). Though these mice all represent the CF phenotype to some degree, none of them completely recapitulate the human symptoms, especially in the airway.

Researchers then turned to the pig, an animal that more closely represents human physiology. In 2008 a *CFTR*-/- pig was reported (Rogers et al., 2008) which closely recapitulated the symptoms seen in human CF patients (Stoltz et al., 2010). More recently a homozygous pig for the ΔF508 mutation was produced, and four pigs with this mutation developed lung disease as they aged (Ostedgaard et al., 2011). It has also been shown that transgenically expressing porcine *CFTR* in the intestine of *CFTR*-/- animals partially restored and improved intestinal function, but did not improve pancreatic function or prevent lung disease (Stoltz et al., 2013). Though mouse models have advantages such as low cost and years of data collected, the pig model seems very promising for development and treatment of new CF therapies.

Regulation of the *CFTR* Gene

If gene therapy is to be attempted, proper *CFTR* regulation needs to be understood. The gene is expressed in a tissue-specific manner, yet the promoter region appears to contain no tissue-specific elements and looks like a housekeeping gene (Koh, et al., 1993; Yoshimura et al., 1991). Studies focused on the *CFTR* promoter were done in the early 90s. Binding sites were identified for proteins such as SP1, AP-1, CRE, NF-κB, HNF1α and C/EBP (reviewed in McCarthy and Harris, 2005). Multiple transcription start sites have been identified in different human tissues (reviewed in McCarthy and Harris, 2005). Alternate transcripts may include either exon1a or exon-1a which splice directly to exon2

(Koh et al., 1993; Lewandowska et al., 2010). Additionally, the *CFTR* gene shows some hormonal regulation by both progesterone and estrogen, though the main studies showing this have been done in rats (de Andrade Pinto, Ana C.O. et al., 2007; Gholami et al., 2012; He et al., 2010; McCarthy and Harris, 2005).

DNase 1 hypersensitive sites (DHS) are interesting to investigate in terms of gene regulation. DHS are indicative of open chromatin regions, areas where protein can easily bind to DNA and cause activation or repression of transcription. The Harris lab has done extensive work characterizing DHSs within and around the *CFTR* gene, with much of this work reviewed in McCarthy and Harris (2005). Of important note are the studies examining DHSs in introns 1 and 11 of *CFTR*, one located about 202 kb downstream of the *CFTR* promoter, and two located 80 and 20 kb upstream of the *CFTR* transcription start site (TSS) (Kerschner and Harris, 2012). The upstream DHS at -20 kb was shown to contain CTCF binding sites (Blackledge et al., 2007). The DHS in intron 1 has been shown to increase luciferase expression driven by the *CFTR* promoter and shown to decrease *CFTR* expression in the intestine up to 60% when removed from a YAC reporter (Mogayzel Jr. and Ashlock, 2000; Ott et al., 2009a; Rowntree et al., 2001; Smith et al., 1996). Nucleosome positioning studies have shown a depletion at DHSs in intron 1, 11 and at the downstream 202 kb site in *CFTR* expressing cells (Yigit et al., 2013). The DHSs in introns 1, 11 and 202kb downstream of the gene have been shown to act as *CFTR* enhancer elements (Gheldof et al., 2010; Ott et al., 2009b; Rowntree et al., 2001, chapter 5 of this

thesis). More recently, Zhang et al. described an upstream enhancer located 35 kb from the TSS active only in airway cells that drove luciferase expression of *CFTR* (Zhang et al., 2013). Thus there are a number of known *CFTR* regulatory elements in the 400 kb region of the locus.

We are still working to understand the cell type specific regulation of the *CFTR* gene, to identify the role of enhancer elements located hundreds of kb from the promoter, and to discover transcription factors involved in *CFTR* expression. Though the gene has been known for over 20 years, its precise function and regulation remain unclear. There are still debates as to how the lack of CFTR protein function causes the multiple symptoms of CF. Although treatment options for patients have improved, a complete cure for the disease remains elusive. Many different attempts at gene therapy cures have been attempted, but success remained constrained to the lab. Though extensive work has been done to characterize *CFTR* regulatory elements, the full picture has yet to emerge. This thesis aims to address questions relating to the regulation of *CFTR* with the hope that the knowledge contained herein may one day contribute to treatments and perhaps a total cure for CF.

The Human Genome in Three Dimensions

Now we must change focus, away from an in-depth analysis of CF and the *CFTR* gene to an entirely different topic, the organization of the human genome in three dimensions. There are two popular ways to measure genome

organization: microscopy methods (especially fluorescence *in situ* hybridization (FISH)), and molecular methods ("C" based methods). Each of these methods has their pros and cons.

FISH microscopy is a powerful tool for investigating the position of genomic loci in three dimensions (3D) (Langer-Safer et al., 1982). FISH uses molecular probes that hybridize to target genomic sequences. These probes can be labeled with a variety of fluorescent labels, allowing for identification of multiple targets in one experiment. Multicolor FISH has been used to paint entire chromosomes, identifying chromosome territories within the nucleus and translocations implicated in disease (Bolzer et al., 2005; Speicher et al., 1996). In terms of genome architecture, FISH was used to show that when the *HoxB* cluster becomes active it loops out of its chromosome territory (Chambeyron and Bickmore, 2004). Recently it was shown that the β-globin locus decreased in volume by approximately 40% upon gene activation, presumably because of an increase in chromatin looping within that locus (Corput et al., 2012). The technology to image finer and finer structure is rapidly progressing; however there are still limitations to the amount of detail one can gain through this method (Markaki et al., 2012). Additionally, FISH is challenging to multiplex because of a limited number of color options.

Another way to identify the 3D structure of chromatin in the nucleus is to use "C"-based methodology. Chromosome Conformation Capture (3C) was first described by Dekker et al. in 2002. This technique uses formaldehyde to

crosslink DNA in its 3D nuclear structure. Once chromatin is crosslinked, it is digested with a restriction enzyme. The digested ends are ligated together to form a library of genomic interactions in 3D. Pieces of DNA that may have been far apart in the linear genome can be close together in 3D space, and these fragments become ligated in the 3C library. These new DNA fragments can be analyzed using PCR-based methods. Many other methods based on this approach have been described, including chromosome conformation capture on chip (4C), chromosome conformation capture carbon copy (5C), ChIA-PET, and the genome-wide Hi-C technique (Dostie et al., 2006; Fullwood et al., 2009; Lieberman-Aiden et al., 2009; Simonis et al., 2006). All of these techniques identify which parts of the genome are close together in 3D space with a higher resolution and much higher throughput than current microscopy methods allow.

There are Many Levels of Chromatin Organization

Human DNA can stretch to 2 meters in length, yet it is packaged into a small cell nucleus that averages 10 microns in diameter. Although the genome could be randomly packaged within that 3D space, it is organized within the nucleus on many different levels. The highest level of nuclear organization is the chromosome. Humans have 23 pairs of chromosomes, which have been shown to occupy individual territories in the nucleus (Bolzer et al., 2005; Cremer and Cremer, 2001). Chromosomes of the same size tend to be near each other, with larger and gene-poor chromosomes near the periphery of the nucleus and

smaller and gene-rich chromosomes near the center (Bolzer et al., 2005; Boyle et al., 2001; Croft et al., 1999; Tanabe et al., 2002; Zhang et al., 2012). Regions of chromosomes have been shown to interact with the inner nuclear membrane via lamin-associated domains (LADs) (Guelen et al., 2008). These loci are between 0.1-10 Mb in size and associate with the protein lamin B1 at the inner nuclear membrane. LADs tend to be transcriptionally inactive, consistent with the idea that active genes tend to be located in the middle of the nucleus, not the periphery (Guelen et al., 2008). Indeed, this is the case for the *CFTR* gene. In cell lines that do not express *CFTR*, the gene is preferentially located at the periphery of the nucleus, while in cell lines that do express *CFTR* the gene is localized more to the center of the nucleus (Zink et al., 2004). Another interesting organizational phenomenon seen with the microscope is the presence of sub-nuclear foci. Labeling nascent RNA transcripts with 5-bromouridine 5'-triphosphate (BrUTP) revealed that transcription from RNA polymerase II takes place in punctate speckles throughout the nucleus (Wansink et al., 1993). This observation, and others, led to the idea that polymerases are gathered into particular transcription stations in the nucleus, and actively transcribed genes are brought to these stations (Cook, 1999; Iborra et al., 1996; Osborne et al., 2004). Within each chromosome, active and inactive genes are organized into separate compartments (Lieberman-Aiden et al., 2009; Zhang et al., 2012) referred to as A and B compartments, respectively. These compartments stretch for megabases. A compartments tend to contact other A compartments in 3D, while B

compartments prefer to contact other B compartments (Lieberman-Aiden et al., 2009) (Figure 1.1). The A compartment is defined by containing gene-rich, transcriptionally active chromatin, while the B compartment is defined by containing mostly gene-poor, inactive chromatin (Lieberman-Aiden et al., 2009). It is possible that the detection of A and B compartments by Hi-C mirrors microscopy data of nuclear speckles and other types of sub-nuclear foci. Zooming into the A and B compartments, one finds Topologically Associating Domains (TADs) (Dixon et al., 2012; Nora et al., 2012; Sexton et al., 2012) (Figure 1.1). These structures are usually hundreds of kbs to a few Mbs in length. Looking within a TAD, one will see specific looping interactions, such as those between an enhancer and a promoter, which is the finest resolution we can currently obtain (Figure 1.1). These interactions are less than 1 Mb in distance. Thus, the human genome is organized into a variety of structures, beginning with the largest chromosomal domain and shrinking to the smallest looping domain (Reviewed in Gibcus and Dekker, 2013).

The composition of the human genome into the compartments described above could be random, or it could have important biological meaning. Evidence supports the latter claim. It has been shown that A and B compartments as well as TADs exist in a variety of human cell types, and that TADs are, for the most part, unchanging between the different cell lines tested

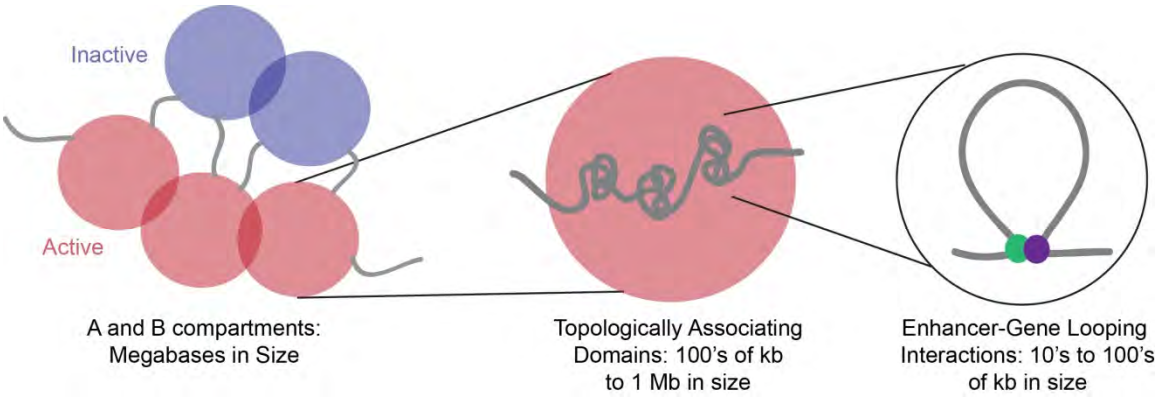


Figure 1.1: A Simplified Model of Levels of Genome Organization (adapted from Job Dekker). From "C" based studies, we know that chromosomes are organized into multiple Mb-sized compartments of active and inactive chromatin, which we refer to as A and B compartments, respectively. Within these compartments, TADs exist at a length scale of a few 100 kbs to about 1 Mb. Within TADs, individual looping structures exist such as enhancer-promoter contacts.

(Dixon et al., 2012; Lieberman-Aiden et al., 2009; Naumova et al., 2013; Phillips-Cremins et al., 2013). TAD boundaries do not change upon the addition of hormones estradiol and progestin, but genes within TADs seem to respond to hormone treatment in a correlated fashion (Dily et al., 2014). TADs have been seen to correlate with large blocks of histone marks (Dily et al., 2014; Nora et al., 2012). TAD boundaries may act as a barrier to histone marks, keeping them from spreading into neighboring TADs. Therefore it seems that the TAD structure may play an important role in gene regulation. Indeed, it has been shown that while TAD boundaries do not shift, interactions within those boundaries do change between mouse embryonic stem cells and differentiated cell types (Nora et al., 2012; Phillips-Cremins et al., 2013). Additionally, random insertions of a *lacZ* reporter throughout the mouse genome showed that the influence of enhancers seems to be constrained to genes within their own TAD (Symmons et al., 2014). There has been a lot of speculation as to what might cause a TAD, particularly what might create a TAD boundary (reviewed in Gibcus and Dekker, 2013; Gorkin et al., 2014). Although CTCF is found to be enriched at TAD boundaries, this mark is widespread throughout the genome and cannot be used to identify a boundary location (Dixon et al., 2012). Additionally, TAD boundaries are enriched for gene promoters, but Dily et al. note that despite this enrichment, gene promoters were in median located >299 kb from TAD boundaries, indicating that this measurement does not clearly mark TAD boundaries either (Dily et al., 2014; Dixon et al., 2012). Some have proposed that looping within the TAD creates the

boundary. However, deleting a boundary causes the neighboring TADs to partially blend together and creates new loops (Nora et al., 2012). Therefore it is more likely that TAD boundaries constrain looping contacts within the TAD; thus TAD boundaries are likely not caused by their internal loop structure. How the boundary regions perform this task, and what molecular structures take part, is still an area of investigation.

3D Looping is a Common Strategy for Gene Regulation

One region that has been well-studied via “C” methods is the β-globin locus, home to 5 globin genes that are regulated based on developmental stages (reviewed in Levings and Bungert, 2002). Gene expression in the locus is controlled by a region called the locus control center (LCR). As cells pass through development, the genes are turned on and off as follows: first the ε-globin gene is expressed in the embryo. Once the liver develops, the two γ-globin genes are turned on, and ε-globin is repressed. After birth, the β-globin (or adult globin) gene is expressed, along with α-globin at a much lower level. It has been shown using 3C that the LCR contacts the different gene promoters at the correct developmental stage to induce expression (Palstra et al., 2003; Tolhuis et al., 2002). When these genes are not expressed, their promoters do not contact the LCR. These 3C experiments match microscopy data of the same locus (Carter et al., 2002). Thus, 3D genome looping plays a critical role in developmental regulation of the globin locus. Another example is Sonic hedgehog gene

expression in mouse limb buds (*Shh*). *Shh* expression is regulated by an enhancer located 1 Mb upstream of its promoter (Lettice et al., 2003). *Shh* loops out of its chromosome territory only when *Shh* expression is active (Amano et al., 2009). When this enhancer is deleted, the mutant mice have no feet (Sagai et al., 2005). Other loci that show similar expression-specific looping interactions are the *CFTR* gene (Gheldof et al., 2010; Ott et al., 2009a, Chapter 3 of this thesis), the c-MYC locus (Wright et al., 2010), the α-globin locus (Baù et al., 2011; Vernimmen et al., 2007) and key developmentally regulated genes (Oct4, Nanog, Nestin, Sox2, Klf4, and Olig1-Olig2) (Phillips-Cremins et al., 2013). These studies show that 3D looping between promoters and gene regulatory elements is a common strategy to control gene expression in eukaryotic cells.

Goals of this Thesis

In order to discover regulatory DNA elements of the *CFTR* gene, I have used molecular biology techniques that examine the 3D conformation of chromatin inside the cell nucleus. These techniques include chromosome conformation capture (3C) and chromosome conformation capture carbon copy (5C) (Dekker et al., 2002; Dostie et al., 2006). In this thesis I describe the experimental methods of 3C and 5C. Then I describe the 3C experiments which show that the *CFTR* locus contains looping interactions specific to cell lines expressing the gene. Next I introduce an expanded experiment which interrogates a 2.8 Mb region surrounding the *CFTR* gene using 5C to search for

more promoter-specific interactions, examine differences in looping between cell types, and investigate the role of TADs in the region. Finally I present evidence that the elements discovered in the previous experiments act as *CFTR*-specific enhancers, and show that they bind to transcription factors that have been implicated in *CFTR* expression.

I find that the *CFTR* locus contacts four main regulatory elements in cells that express the gene. Two of these elements also contact each other in 3D space (elements III and IV). I show that these two of these elements act as enhancers in a luciferase assay. Using the yeast one-hybrid assay, I then identify sets of human transcription factors that may act in forming and maintaining the looping structure and/or regulating *CFTR* gene expression. When I looked at a larger region using 5C, I identified 6 TAD boundaries consistent across a panel of 5 diverse cell types. These TAD boundaries are not dependent on gene expression or the interactions within or between the TADs. All *CFTR* enhancers are contained within one TAD. It appears that TAD boundaries may function to limit search space of gene promoters looking for their appropriate genomic regulatory elements.

CHAPTER II

CHROMOSOME CONFORMATION CAPTURE METHODS

This chapter was partially adapted from Natalia Naumova, Emily M. Smith, Ye Zhan, and Job Dekker. “Analysis of Long-range Chromatin Interactions Using Chromosome Conformation Capture.” Methods 58, no. 3 (November 2012): 192–203. doi:10.1016/j.ymeth.2012.07.022.

Contributions

This paper presents an update to the original chromosome conformation capture method and describes not only the protocol but important analysis and troubleshooting steps. N.N., E.M.S. and Y.Z. are co-first authors. N.N. wrote the troubleshooting section, E.M.S. wrote the analysis section, and Y.Z. wrote the materials and methods section. E.M.S. and N.N. together wrote the introduction, technical overview, and experimental design considerations. E.M.S. and N.N. created the figures. Y.Z. provided help with the figures and advice on all sections of the manuscript. J.D. gave guidance throughout the writing and editing process. All 5C sections were added by E.M.S.

Abstract

Chromosome Conformation Capture, or 3C, is a pioneering method for investigating the three-dimensional structure of chromatin. 3C is used to analyze long-range looping interactions between any pair of selected genomic loci. Most 3C studies focus on defined genomic regions of interest that can be up to several hundred kilobases (kb) in size. 5C, or chromosome conformation capture carbon copy, is a method based on 3C technology. 5C utilizes deep sequencing technology to investigate a larger genomic region (up to a few Megabases (Mb) in size) in depth. The methods have become widely adopted and have been modified to increase throughput to allow unbiased genome-wide analysis. Here we describe the 3C and 5C procedures in detail, including the appropriate use of the technology, the experimental set-up, an optimized protocol and troubleshooting guide, and considerations for data analysis.

Introduction

Chromosomes form intricate three-dimensional structures inside the confined cell nucleus. This organization is thought to play roles in many, if not all, aspects of genome regulation, including gene expression, DNA replication, chromosome transmission and maintenance of genome stability (Cremer and Cremer, 2001; Miele and Dekker, 2008; Misteli and Soutoglou, 2009). Gene expression in particular is profoundly dependent on chromatin folding, where looping interactions facilitate long-range control by distant gene regulatory elements (Dean, 2011; Dekker, 2008a; Miele and Dekker, 2008). Furthermore, at the nuclear level, groups of active genes are found clustered around sub-nuclear structures enriched in transcription and splicing machineries (Fraser and Bickmore, 2007). Similarly, inactive regions of the genome are found in clusters, e.g. around polycomb bodies (Bantignies et al., 2011) and at the nuclear lamina (Guelen et al., 2008).

Chromosome structure and nuclear organization have been studied extensively for over a century, using an expanding array of technologies that allow observation of chromosome folding at increasing resolution. Currently, two types of approaches are being used. First, microscopic studies allow study of chromosome structure and chromatin dynamics in single cells. Recent technologies using GFP-tagged DNA binding proteins allow the visualization of the positions and movements of defined loci inside living cells (Belmont, 2001).

Recent developments in optics and image analysis have increased the resolution with which the relative sub-nuclear positions of loci can be determined. A second category of technologies employs molecular and genomic approaches to obtain information on average chromatin folding for large populations of cells (van Steensel and Dekker, 2010). This set of approaches is based on the Chromosome Conformation Capture technology (3C), developed over a decade ago (Dekker et al., 2002). 3C-based technologies allow the detection of the relative frequency of interaction between any pair of loci in the genome. From these interaction frequencies the folding of chromatin can be inferred. For instance, frequent interactions between two distant genomic loci point to the presence of a chromatin loop (Dekker, 2006). The resolution of 3C is determined by the choice of restriction enzyme, but is usually in the range of several kb, significantly higher than achievable by light microscopy. Resolution of other C-based methods is limited only by deep-sequencing capabilities.

Application of 3C has identified direct and non-random looping interactions between distant parts of the linear genome, including physical contacts between enhancers and their distal target genes (Miele and Dekker, 2008). Further application of 5C has led to the notion that genomes are organized in complex spatial networks via looping interactions that often are cell-type and condition-dependent and directly related to long-range gene control (Nora et al., 2012; Phillips-Cremins et al., 2013; Sanyal et al., 2012).

3C and its offspring of variants including 4C (circular chromosome conformation capture) (Simonis et al., 2006; Zhao et al., 2006), 5C (chromosome conformation capture carbon copy) (Dostie et al., 2006), ChIA-PET (Fullwood et al., 2009) and Hi-C (Lieberman-Aiden et al., 2009), are all based on the same basic principle of capturing and detecting long-range chromatin interactions and have 4 common steps (Figure 2.1A): 1) Chemical cross-linking of chromosomes to covalently link chromatin segments that are in close spatial proximity; 2) Fragmenting the solubilized genome into small pieces, usually by digesting it with a restriction enzyme; 3) ligation of linked DNA fragments under diluted conditions where intra-molecular ligation is strongly favored over inter-molecular events; and 4) detection and quantification of ligation products. The various 3C-based methods differ mostly in how ligation products are detected. In the case of 3C, ligation products are detected one at a time by PCR using locus specific primers (Figure 2.1B). 5C uses a multiplexing approach to increase the number of interactions (ligation products) that are detected in parallel, thereby increasing the throughput of the assay (Figure 2.2) (Hakim and Misteli, 2012; Sanyal et al., 2011, 2012; van Steensel and Dekker, 2010).

The 3C procedure produces a comprehensive library of ligation products representing chromatin interactions throughout the entire genome. However, because interactions in standard 3C experiments are detected one at a time, a typical 3C analysis is usually limited to interrogation of at most hundreds of pair-wise interactions and is focused on the detection of looping interactions in

relatively small regions - from 10kb up to 1Mb (Gheldof et al., 2006, 2010; Tolhuis et al., 2002). 5C expands upon 3C by using multiple probes to detect an increased number of pair-wise interactions in one experiment. 3C and 5C are mainly used in hypothesis-driven experiments, designed based on some prior knowledge such as the genomic locations of functional elements of interest. 5C is more amenable to experiments where less is known about a particular locus, since the design of multiple probes throughout the region allows more of a search mode than 3C.

Whereas subsequent 3C-based methods (4C, 5C and Hi-C) were designed to increase the throughput of interaction detection, 3C has remained a critical technique that is commonly used for fine scale analysis of genomic regions of interest. 3C has been used to study chromatin folding in a range of organisms, ranging from bacteria, yeast and plant to human. In the original study performed on *S. cerevisiae*, 3C was used to measure changes of inter-chromosomal contacts between centromeres and homologous chromosomes during meiosis, and to determine the overall population average three-dimensional conformation of chromosome III (Dekker et al., 2002). Since then, 3C has been applied mostly to study the interactions between genes and distal regulatory elements such as enhancers. The first such study demonstrated physical contacts between the β-globin genes and the Locus Control Region (LCR), which is known to strongly activate these genes (Tolhuis et al., 2002). Many more examples of such looping interactions have now been described,

indicating that chromatin looping is a general mechanism of gene control in higher eukaryotes (Miele and Dekker, 2008; Sanyal et al., 2012). 3C was also used to discover that looping interactions between the LCR and the globin genes require several transcription factors, including GATA1 and EKLF1, and some require the CTCF protein (Drissen et al., 2004; Splinter et al., 2006; Vakoc et al., 2005).

Currently, 3C is mostly used for targeted analysis of loci of interest, to identify long-range interaction between candidate genes and regulatory elements, and to probe how these interactions change upon perturbations such as knockdown of specific chromatin factors thought to mediate chromatin folding. Another emerging application is to link regions identified as playing a role in disease by genome-wide association studies (GWAS) to other genomic loci. GWAS studies often identify regions devoid of genes, but containing putative gene regulatory elements. 3C is now being used to identify potential target genes located around the GWAS region that physically interact with the GWAS regions, or the regulatory elements located within it (Ahmadiyeh et al., 2010; Davison et al., 2012; Wright et al., 2010).

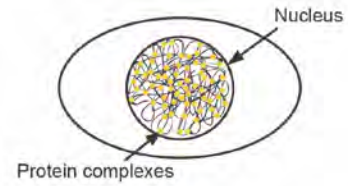
5C is used to gain high-resolution data of a large genomic region that is not yet obtainable by HiC or other deep sequencing methods. The 5C technique allows high-resolution analysis of larger genomic regions on the scale of a few megabases to entire chromosomes (Dostie et al., 2006). 5C has been used to analyze the interactions of transcription start sites with potential regulatory

elements in 14 genomic regions of the ENCODE project (Sanyal et al., 2012). 5C was done on the entire X chromosome in mouse and revealed the presence of Topologically Associating Domains (TADs) along the entire chromosome, and also showed that the boundaries between two of these domains changed when the XIST locus was removed (Nora et al., 2012). It was recently shown that, though TAD structures seem invariant between cell lines, sub-TAD looping interactions are cell-type specific and change during cell differentiation (Phillips-Cremins et al., 2013). The 5C method is currently the ideal technique to look at large genomic regions (or many different regions at the same time) with a resolution of tens of kb. In the future, it is likely that deep sequencing costs will no longer be restrictive and 3C libraries can be directly sequenced. Until that time, 5C remains the go-to technique for directed 3D looping studies.

Here we discuss in detail the principles of the 3C and 5C methods, paying special attention to experiment design and data analysis. We present an updated protocol for performing 3C analysis in mammalian cells and discuss potential pitfalls. The presented protocol can be easily adapted for many other organisms. Further, we build on our years of experience to describe troubleshooting solutions and to identify critical issues related to the planning, execution, and interpretation of the experiments.

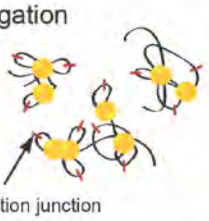
A

1. Crosslinking

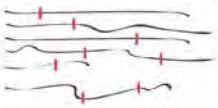


2. Digestion

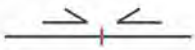
3. Ligation



4. Reverse crosslinks

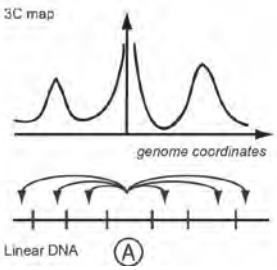


5. PCR detection



6. Gel Quantification

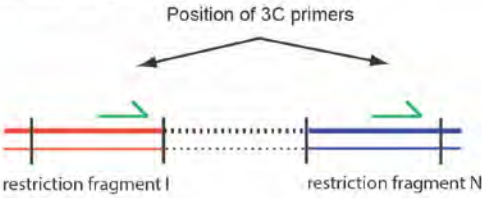
7. Plot 3C profile



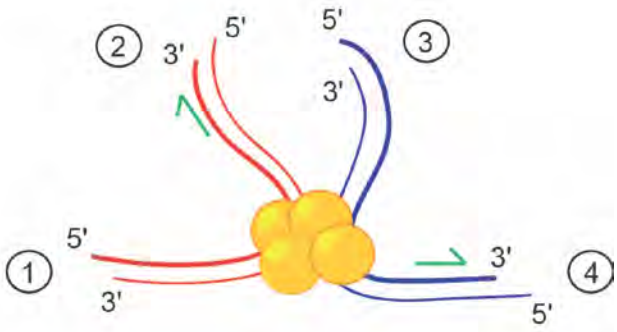
B

Simplified Diagram of Ligation:

1. Linear genome:

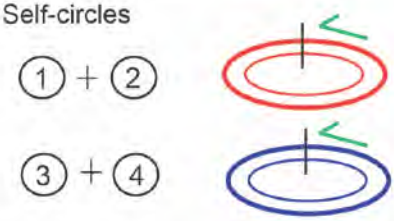


2. Ligation:



Possible Ligation Products:

Self-circles



Interaction between restriction fragments

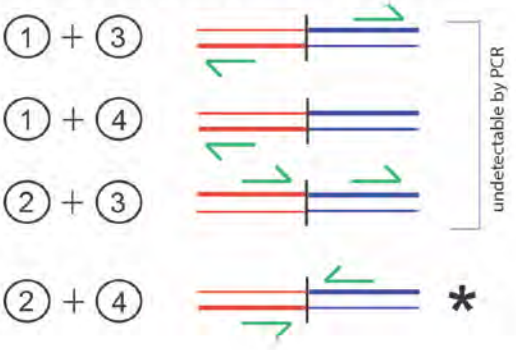


Figure 2.1: Method Overview with Details of Ligation. **A**) An illustration of the 3C method. Genomic DNA is crosslinked (1), capturing three-dimensional interactions inside the cell. After cell lysis and removal of cell membranes, the captured chromatin is solubilized and digested (2), isolating protein-DNA complexes from the chromatin network. The free DNA ends are then ligated together (3) in dilute conditions favoring intra-molecular ligation, creating new DNA junctions representing the proximity of restriction fragments in the fixed sample. After ligation, the crosslinks are reversed (4) and 3C template is purified to eliminate cellular debris. Finally, the ligation products are detected (5) using PCR-based methods. After quantification (6) the results are plotted as a 3C profile (7), revealing interactions between anchor (labeled "A") and all other fragments in the genomic regions, which mirrors 3D spacing in the nucleus. **B**) Possible outcome of ligation reaction between two restriction fragments. As seen in part A, there are many restriction fragments contained within one complex. To further understand the ligation step, we have simplified the reality and show a view where one complex contains only two restriction fragments - red and blue. The 5' and 3' ends are indicated for each strand. Each digested end has been numbered 1-4. Also indicated are the locations where 3C primers have been designed. Note that all the primers are on the "forward" strand, located near the restriction site. There are six possible ligation products that result from this molecule. Two of these produce self circles, which are not of interest. Only one of the remaining four ligation products results in a detectable product - that is when end 2 and end 4 are ligated to each other. This ligation event will bring the primers into the proper orientation to produce a PCR product. None of the other ligation products will be detected.

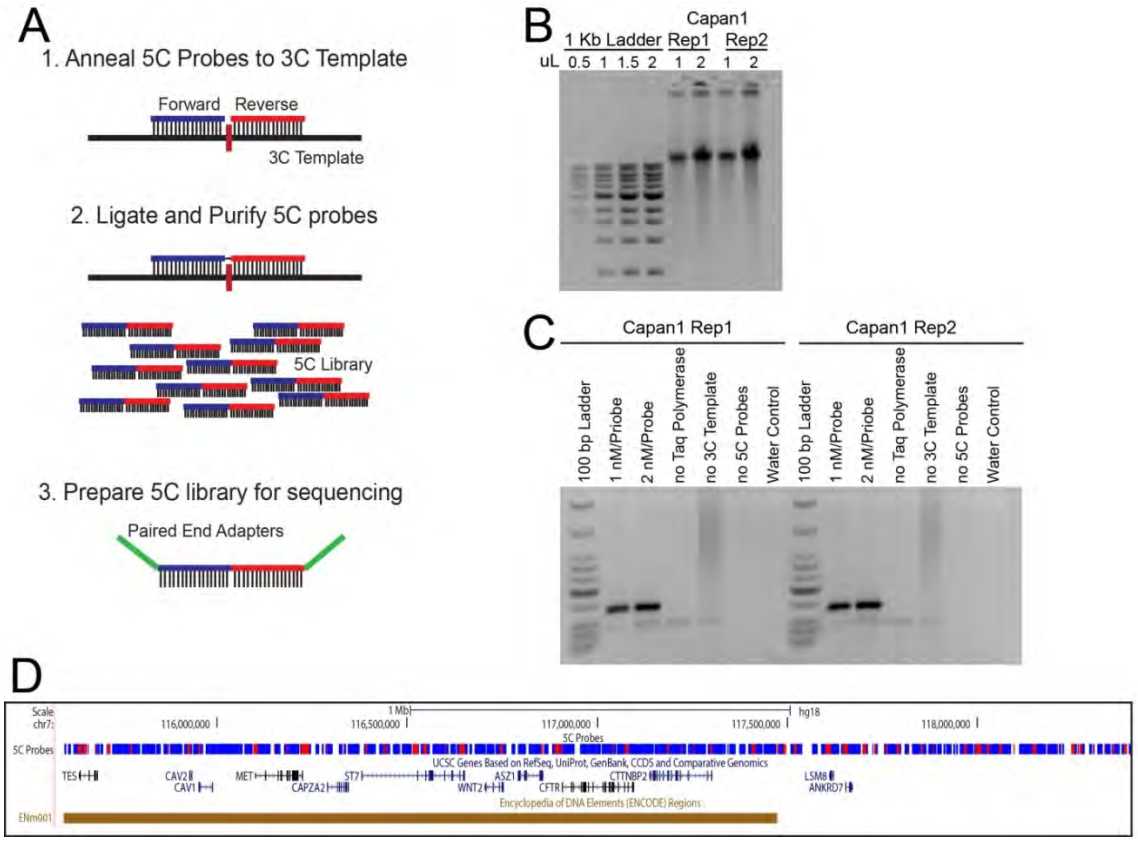


Figure 2.2: Preparation of the 5C Library. A) Method describing the steps creating a 5C library. 1. 5C probes are annealed to a 3C template exactly at the restriction site of the 3C template. 2. 5C probes are nick ligated and the purified from the 3C template. 3. The 5C library is prepared for sequencing by addition of specific sequencing primers. **B**) Sample 3C template used to create a 5C library. Two replicates (reps) of the cell line Capan1 are shown. **C**) 5C libraries created using the Capan1 3C templates. Note the increase in product upon increase in probe concentration. Equally important is the absence of a band in the control lanes, indicating the 5C library is not contaminated. **D**) Snapshot from the USCS genome browser showing the 5C probe design. Blue indicates the presence of a forward-facing probe, and red indicates the presence of a reverse-facing probe. The probe design incorporates the ENCODE region Enm001 with an additional 1 Mb downstream. There are 12 genes contained in this region.

Technical Overview of the Methods

In this section we provide a discussion of the steps of 3C and 5C, paying particular attention to the molecular biology behind the method. The basic steps of 3C are: 1) formaldehyde crosslinking, 2) digestion with a restriction enzyme, 3) intra-molecular ligation, and 4) ligation product detection via PCR-based methods (Figure 2.1A). 5C builds off the 3C template by 1) annealing probes to the 3C template, 2) ligating and purifying the probes, and 3) prepping the 5C library for deep sequencing (Figure 2.2A).

Formaldehyde concentration and fixation time

Formaldehyde concentration and time of fixation affect 3C-based experiments and need to be standardized to facilitate accurate comparison between samples. Fixation conditions can be different for different species and are defined by chromatin properties, presence of a cell wall etc. We crosslink mammalian cells with 1% formaldehyde for 10 minutes at room temperature. Other groups have reported using 2% formaldehyde for 10 minutes (Tolhuis et al., 2002). 3C experiments in Drosophila embryos were successfully performed with fixation with 3% formaldehyde for 30 minutes at 25^{o}C (Comet et al., 2011). Intact yeast cells should be fixed with 3% formaldehyde for 10 minutes (Belton, J-M. and Dekker, J., unpublished), but yeast spheroplasts are fixed with 1% formaldehyde for 10 minutes (Dekker et al., 2002).

Changing fixation conditions affects the amount and density of protein-DNA cross-links, which in turn affects the efficiency of restriction digestion and thus the size of DNA-protein complexes. Inefficient formaldehyde cross-linking, caused by using a low concentration of formaldehyde or a too short of an incubation time, may lead to a failure to capture looping interactions. Over-fixed chromatin will make digestion very inefficient, leading to large DNA fragments and low PCR signals. Optimal conditions can be determined experimentally by performing 3C using a range of formaldehyde concentrations. An appropriate formaldehyde concentration will lead to readily detectable and abundant ligation products for fragments separated only a few kb. When analyzing cells in tissues or organs, it is recommended to first dissociate these materials, e.g. with collagenase, into single cells prior to fixation (e.g. (Montavon et al., 2011)).

Digestion with Restriction Enzymes

The second step of 3C involves digesting the crosslinked chromatin with a restriction enzyme. We recommend using restriction enzymes that recognize and cut 6 bp sites when possible (“6-cutter”) for overnight digestion. The amount of enzyme in the reaction can be increased if higher digestion efficiency is required (at least up to 5-fold).

Protein complexes crosslinked to DNA may block restriction sites and reduce efficiency of restriction digestion. For instance, efficiency of yeast and mammalian chromatin digestion is around 70-75% on average as was measured

by PCR with primers located across restriction sites (Gheldof et al., 2006, 2010; Sanyal et al., 2012; Tolhuis et al., 2002).

Inactive and condensed chromatin is generally less accessible to nucleases than active and open chromatin, and this might confound 3C studies. Importantly, 3C analyses have been found to be not, or only to a very limited extent, affected by this intrinsic difference in chromatin compaction. Several studies directly determined the digestion efficiency in the context of a 3C experiment of actively transcribed accessible loci and of repressed, methylated and condensed loci. Digestion efficiency was found to be unaffected (Gheldof et al., 2006; Kim et al., 2009; Tolhuis et al., 2002), as chromatin digestion in 3C is performed after chromatin is partly denatured in the presence of 0.1% SDS and brief incubation at 37°C, which removes proteins that are not cross-linked from DNA and partly denatures cross-linked proteins. This dramatically increases accessibility of DNA. Efficiency of restriction digestion can be easily determined by PCR using primers on either side of restriction sites (Figure 2.3). We recommend saving a small aliquot of chromatin directly after digestion (and before ligation) for this analysis if desired.

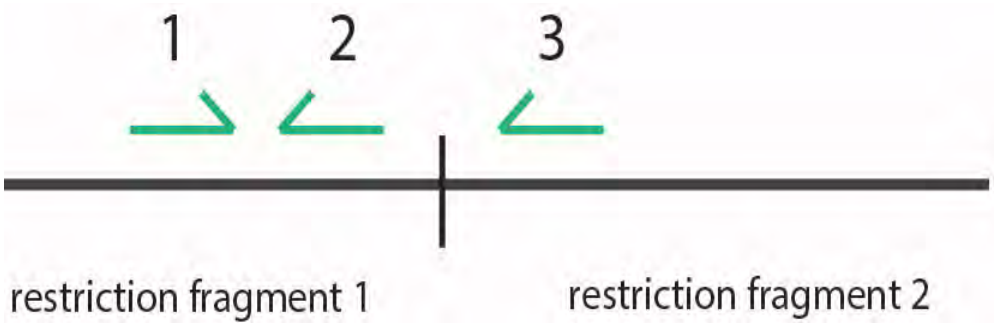


Figure 2.3: **Analysis of Restriction Efficiency of the Chromatin Sample**. Restriction efficiency in a given site is estimated by PCR as a ratio between a PCR product obtained with internal primers 1 and 2 versus primers across the restriction junction 1 and 3. Note that both PCR products should be of similar length to be synthesized simultaneously by the same program. The ratio obtained from digested unligated sample should be corrected by the PCR with the same pairs of primers performed on naked uncut genomic (or BAC) DNA.

Ligation and Reversal of Crosslinking

The third step involves DNA ligation. This step is performed at low DNA concentrations to strongly favor intra-molecular ligation of cross-linked chromatin fragments over background inter-molecular ligation between fragments that are not cross-linked. Intra-molecular ligation is kinetically fast, obviating the need for prolonged ligation times. The ligation time should be kept at a minimum, to avoid increasing the level of background ligations. After ligation, the cross-links are reversed by heating at 65°C in the presence of proteinase K. The 3C ligation product library is then purified and is ready for analysis. The library can also be used as a template for 5C experiments at this stage.

Detection of Ligation Products

The final step of 3C is the detection and quantification of ligation products, representing long-range chromatin interactions. For this step, locus-specific primers are designed to specifically amplify ligation junctions. PCR amplicons are typically around 200 bp in size to facilitate efficient amplification. Both end-point PCR and quantitative real-time PCR (qPCR) have been employed to quantify the abundance of 3C ligation products, with very comparable results (e.g compare regular 3C data for the beta-globin locus described in Tolhuis et al. (Tolhuis et al., 2002) with 3C-qPCR data for the same region described in Splinter et al. (Splinter et al., 2006). In both cases one needs to carefully titrate the amount of 3C ligation product library to ensure amplification and quantification is in the

linear range. In addition, controls should be included to correct for any biases in PCR primer efficiency. These controls are described below.

Annealing 5C probes to the 3C template

The first step in creating a good 5C library is to start with a good 3C template (described above). The 3C template is the basis of the 5C library. Once a good 3C template is obtained, the 5C probes are annealed to the 3C template. 5C probes are designed to anneal exactly at the restriction sites of the ligated 3C template (Figure 2.2A). The 3C template and 5C probes are mixed together and heated to 95˚C for 9 minutes. The temperature is then dropped to 55˚C by 1˚C per minute and maintained overnight.

Ligating and purifying the 5C library

Once the 5C probes are annealed to the 3C template, they are ligated to each other via a nick ligase (Figure 2.2A). The 5C library now consists of ligated probes annealed to the 3C template. This ligated library is amplified via universal PCR. Finally, the 5C library is separated from the 3C template via gel purification.

Preparing the 5C library for sequencing

Specific tails must be added to the 5C library to prepare it for sequencing. First, “A” bases are added to the 5C probes to allow for the attachment of sequencing tails. Second, the sequencing adapters are added (in our case, we

use paired end (PE) adapters for sequencing with the Illumina platform) (Figure 2.2A). The 5C library is gel purified and then amplified before finally being sent for sequencing.

Experimental Design Considerations

When embarking on a 3C or 5C analysis, one needs to carefully plan the design of the experiment. Here we describe important considerations related to selection of the genomic regions for analysis, the inclusion of a control region, the choice of restriction enzyme and the design of PCR primers and 5C probes.

Choosing the region of interest and a control region

The first step in setting up a "C" experiment is to identify a unique region(s) to investigate. For 3C, the size of a region is limited firstly by the desired resolution, determined by the restriction enzyme, and secondly by the amount of PCR reactions one can perform, which is related to the amount of 3C library one can obtain for the cells of interest. The typical size of a region that can be comprehensively studied by 3C ranges from tens to hundreds of kb, although longer-range interactions have been studied (Amano et al., 2009; Wright et al., 2010), as well as interactions between chromosomes (Bacher et al., 2006; Spilianakis et al., 2005; Xu et al., 2007).[1] 5C experimental regions can be much larger, extending from a few megabases to entire chromosomes (such as X and

[1]**Signal to Noise**: 3C signals typically decay with genomic distance. As a result the signal to noise ratio decreases with increased distance between two interrogated loci, which usually limits 3C analysis to regions up to 1Mb. Other 3C-based techniques do not have this limitation because they employ binned data, where the signal is determined by all interactions within the bin and not by a single point, as in 3C.

chromosome 21 (Nora et al., 2012; Sanyal et al., 2012) and unpublished data, Dekker Lab).

When one intends to compare the folding of a locus of interest in different cell types, or under different conditions, it is important to choose a separate control region. This region needs to be selected based on prior knowledge that suggests that the region is similarly organized in the selected cell types or conditions. 3C interactions determined throughout this region are assumed to be the same and thus can be used as an internal data set to quantitatively normalize the 3C data obtained for the region of interest in the different cell types or conditions. We advise the use of gene-poor regions (or so called gene deserts), although a locus with house-keeping genes has also been successfully used (Gheldof et al., 2006). 5C analysis benefits from the inclusion of a “gene desert” region as well, though it is not always necessary.

Choosing an appropriate restriction enzyme

Restriction enzymes are used to digest crosslinked chromatin. Once the chromatin is digested, it is ligated to create a 3C template. The choice of restriction enzyme to use in a “C” experiment is highly dependent on the goal of the experiment and the region selected for analysis. Several points should be kept in mind when selecting a restriction enzyme.

Desired Resolution

The resolution at which interactions can be mapped is primarily determined by the size of the restriction fragments and thus the choice of the restriction enzyme. We recommend using a restriction enzyme that recognizes a 6 base-pair sequence cut site, such as EcoRI or HindIII. Such enzymes will cut the genome approximately once every 4 kb (although a wide variety of fragment sizes ranging from 100s of base pairs to 10s of kb will be obtained), resulting in around 1 million restriction fragments in the human genome. In some cases, higher resolution is desired. For instance, after initial 3C analysis with a "6-cutter" enzyme, one might want to map the location of an interacting element more precisely. "Fine-mapping" can be achieved by using a "4-cutter" restriction enzyme, which cuts on average every 256 base pairs, giving approximately 16,000,000 fragments of the human genome.[2]

Proper Position of Cut Sites

It is desirable to choose a restriction enzyme which has a more or less equal spacing of cut sites across the analyzed region. Fragments that are too short or too long should be excluded from the primer design, as they can introduce biases to the data.[3] Thus, fragments less than 1 kb or greater than 10

---

[2]**Desired Resolution of Restriction Enzymes.** The complexity of the 3C library (i.e. number of potentially formed ligation products) is determined by the restriction enzyme, and can impact the reliability of PCR detection and quantification of individual ligation product. The complexity of the library obtained with a "4-cutter" restriction enzyme will be greatly increased as compared to a library generated with a "6-cutter".

[3]**Exclusion of Restriction Fragments:** We have found that very large, and very small fragments can sometime yield aberrant interaction frequencies. This might be due to differences in intra-

kb should be excluded when using a “6-cutter.” In addition, when prior knowledge of positions of putative interacting elements is available, e.g. by the presence of histone modifications indicative of the presence of an enhancer or promoter, one can select a restriction enzyme that cuts the region in appropriate fragments that separate these elements from flanking regions, thus leaving elements of interest intact.

Digestion Efficiency

Not all restriction enzymes perform equally well in 3C. The reason for this is that digestion is performed in sub-optimal buffer conditions containing considerable concentrations of detergents. We have found that EcoRI, HindIII, BglII, XhoI, AciI, and BsrGI digest cross-linked chromatin efficiently, typically reaching 70% of digestion of each restriction site (although the region selected for analysis can change the digestion efficiency). Enzymes that produce staggered ends are recommended, as these ends are more efficiently ligated. Enzymes that generate blunt ends can be used as well, but ligation efficiency is somewhat reduced.

3C Primer Design

---

molecular ligation efficiency for very long DNA fragments. Therefore, we recommend avoiding, if at all possible, very long (>10Kb) restriction fragments.

In a 3C experiment any pair of interacting loci can lead to formation of six different ligation products (Figure 2.1B). Two of the resulting products are self-circles, which occur when a restriction fragment is ligated to itself. The other four combinations occur when two different restriction fragments are ligated to each other in various orientations. In a typical 3C experiment primers are designed to detect only one of the four ligation products between the two fragments. In order to detect a ligation product between two different fragments, PCR primers should be placed in an orientation indicated by the asterisk in Figure 2.1B. In this section we describe primer design and common physical properties of 3C primers.

Primer design in the region of interest

3C primers are designed for all restriction fragments of interest. For correct interpretation of the data it is important to not only interrogate interactions between pairs of loci of interest, but to obtain a more comprehensive interaction profile throughout the region. In general this profile will show an inverse relationship between interaction frequency and genomic distance (Figure 2.4). A looping interaction is then inferred when a peak on top of this overall profile is observed (Dekker, 2006; Sanyal et al., 2012).

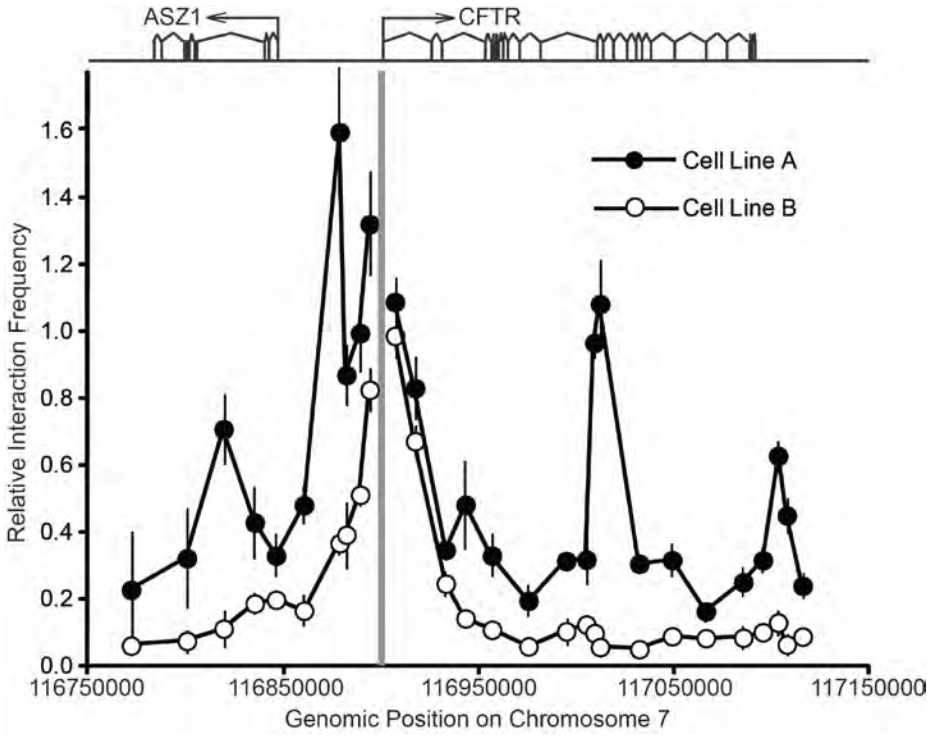


Figure 2.4: Representation of Final 3C Data. Figure adapted from Gheldof et al. Figure 1 (Gheldof et al., 2010). Plotted here is sample 3C data from Gheldof et al., 2010 for two cell lines. Above the graph is a schematic of the two genes covered in this data set. Each data point represents the average of three technical replicates and three normalized biological replicates. The cell lines were normalized to each other before they were plotted on the same graph. The data is plotted as relative interaction frequency versus genomic distance. The grey bar represents the anchor fragment. Note that as the genomic distance increases, the interaction frequency of both cell lines generally decreases. In cell line B, for example, this relationship is clear throughout the whole profile. Note four areas of high interaction frequency in cell line A. These four regions stand out as peaks, and represent locations in the genome that are close to the anchor point in three-dimensional space.

An example of a 3C experiment determines whether a given genomic element, for instance a gene promoter, is engaged in a long-range interaction with one or more distally located elements, such as enhancers. When the positions of these distal elements are not known, one designs 3C primers for the promoter and all restriction fragments throughout the region under study. If the location of putative distal elements is known, one designs primers for the corresponding fragments, but also for a number of flanking fragments located in between the promoter and the distal elements to obtain a larger interaction profile. If flanking regions are excluded, one cannot conclude which fragment contains the actual point of looping contact.

Primer design in the control region

As described above a control region is included in 3C studies to allow for comparison of 3C data obtained in different cell types or conditions. We recommend designing several primers throughout the region, so a control 3C interaction profile can be obtained that covers a similar genomic distance as that obtained in the region of interest. 3C primers are designed for at least 10 different restriction fragments spaced at various distances.

Direction of the primers

We strongly encourage researchers to use a unidirectional 3C primer design. In such a design all primers are oriented in the same direction, on the

same DNA strand, along the chromosomal region of interest. All pairs of primers will amplify ligation products that are the result of head-to-head ligation of the corresponding restriction fragments. A unidirectional primer design is important because it avoids amplification of non-informative ligation products (Figure 2.5). First, when primers are used that point away from each other in the linear genome one runs the risk of accidentally amplifying a self-ligated partial digestion product. Second, primers for two directly adjacent restriction fragments that point towards each other in the linear genome sequence will produce a PCR product even when the restriction site between them was never cut in the 3C experiment.

Physical properties of the primers

To increase specificity of the primers we recommend designing long primers with high melting temperature (on average the Tm is 90°C); the length of 3C primers is 28-30bp with a GC content of ~50%, preferably carrying a single G or C nucleotide on the 3' end. We have found that the use of rather long primers is especially important for complex genomes, where short 20bp primers do not provide necessary specificity and efficiency. Primers are designed ~80-150bp away from the restriction cut site so that the predicted amplicon will be between 160 and 300bp in size. We recommend checking the uniqueness of each primer.[4]

[4]**Checking Primer Uniqueness:** We recommend using both BLAST and BLAT for checking uniqueness of 3C primers as those programs have different algorithms of searching for a match in a genome. BLAT works much quicker, however BLAST gives more comprehensive results. It is also possible to check if primers have strong secondary structures (hairpins) and form stable

5C Probe Design

Once the region(s) of interest are selected, one must design probes. There are two general types of probe design: an alternating scheme, where the investigator is not searching for particular interactions but is interested in the general properties of the region, and a targeted scheme, where the investigator is looking for specific interactions in the region. An alternating scheme places forward and reverse probes throughout the region in an alternating pattern. This scheme produces a data set with complete coverage of the region and is useful for identifying unknown looping interactions. An example of a targeted scheme is shown in Figure 2.2D . Here, reverse (anchor) probes are located on every transcription start site (TSS) in the region, with forwards situated throughout. This probe scheme is attempting to discover which genomic elements loop to gene promoters and act as regulators.

Probe design in the control region

We suggest the use of an alternating scheme for probes located in the control region. An alternating probe design placed forward probes and reverse probes in alternating restriction fragments. This produces an unbiased view of the region and provides plenty of reads to use during normalization procedures.

---

homodimers. Primers should also be checked for formation of heterodimers with anchor primer. Free online tools such as IDT oligo analyzer (http://www.idtdna.com/analyzer/Applications/OligoAnalyzer/) can be used for this analysis.

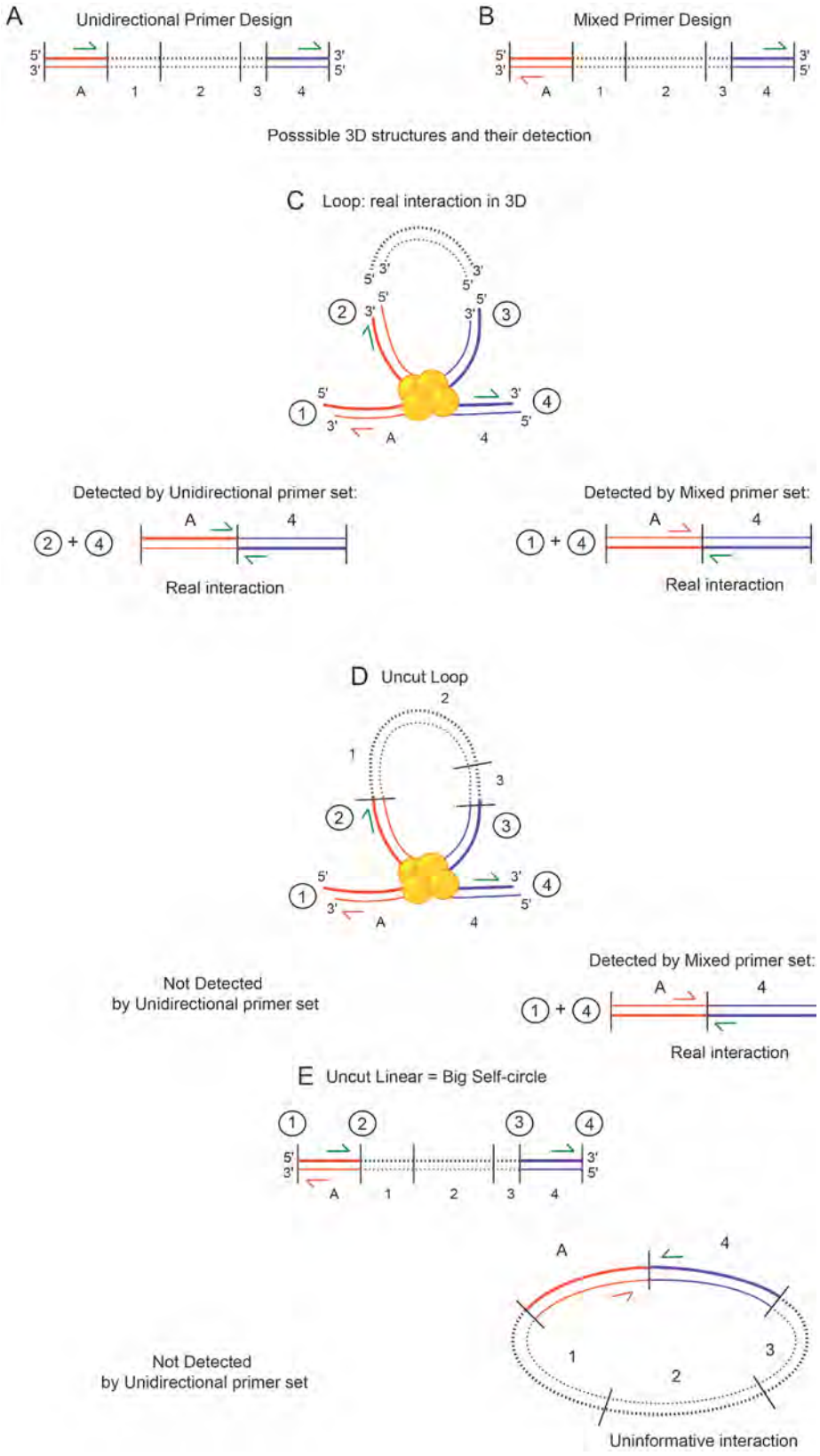

A
Unidirectional Primer Design
5'
3'
3'
5'
A
1
2
3
4
B
Mixed Primer Design
Posssible 3D structures and their detection
C
Loop: real interaction in 3D
Detected by Unidirectional primer set:
Real interaction
Detected by Mixed primer set:
Real interaction
D
Uncut Loop
Not Detected
by Unidirectional primer set
Detected by Mixed primer set:
Real interaction
E
Uncut Linear = Big Self-circle
Not Detected
by Unidirectional primer set
Uninformative interaction

Figure 2.5: Unidirectional vs Mixed 3C Primer Design and Types of Interactions Detected in 3C Library. A) Unidirectional primer design: all 3C primers are designed to the same strand of DNA; **B**) Mixed primer design: some primers in the set are designed to the opposite strand of DNA. Unidirectional 3C primers will only detect real looping interactions when distant regions of the linear genome are brought together by protein complexes **C**). In the shown example 3C molecules resulted from cut and ligation between end 2 of the anchor fragment A and end 4 of restriction fragment 4 will be detected by PCR. Note that such a ligation product can be formed only when anchor A and fragment 4 are crosslinked together. In the case when digestion is inefficient, undigested loops **D**) or long multi restriction fragment self-circles **E**) can be formed. These structures will be detected only when using a mixed primer design when 3C primers in the anchor (**A**) and a fragment on the other end of the DNA molecule (here restriction fragment 4) are designed to the opposite strands. Both structures will not be detected with unidirectional 3C primers.

However, an alternating scheme can eat up reads, so instead of placing a forward and reverse prove on every restriction fragment, we suggest skipping one or two fragments in between probes to reduce the amount of reads in this region.

Physical properties of the probes

5C probes can be designed at the My5C website (http://my5C.umassmed.edu) (Lajoie et al., 2009). Probes are 40 bp in length. If a shorter sequence is sufficient to obtain a predicted annealing temperature of 65 °C, random sequences can be added to make a total of 40 bases. All of the 5C probes have an extension of universal tail sequences at the 5′ end for forward 5C probes and at the 3′ end for reverse 5C probes. DNA sequence of the universal tails of forward probes is 5′-CCTCTCTATGGGCAGTCGGTGAT-3′; DNA sequence for the universal tails of reverse probes is 5′-AGAGAATGAGGAACCCGGGGCAG-3′.

Creating a Control Library for 3C

3C employs PCR with locus-specific primers, which may amplify their target ligation product with different efficiencies, even when great care is taken to design primers. It is therefore critical to correct for this differential primer efficiency. This can be done by PCR analysis of a control library that contains all interrogated ligation products in equimolar amounts. Any differences in PCR

product formation obtained by pairs of 3C primers with this control library as template can then be used to estimate primer pair efficiency.

A control library is prepared by digesting and randomly ligating non-crosslinked purified DNA. For small genomes (yeast, bacteria, fly) the control library can be generated using purified genomic DNA. For larger genomes, such as mouse or human, a genome-wide random control library is too complex to allow reliable detection of individual ligation products. For these organisms the control library can be made from one or more bacterial artificial chromosomes (BACs) that span the genomic region(s) investigated by 3C, including the control region (e.g. (Dostie et al., 2006)). When multiple BACs are used they should be selected so that they display minimal overlap while simultaneously keeping the number and size of gaps to a minimum. This ensures minimal over- or under-representation of genomic regions in the control library. If BACs are unavailable for a genomic region, they might be substituted with fosmids, cosmids or even plasmids. The control library is then generated by mixing the clones in equimolar amounts and digesting the DNA followed by random intermolecular ligation.

Determination of Interaction Frequencies

It is important to first determine the optimal amount of 3C library to use in each PCR reaction. This amount has to be found experimentally for each 3C library in a titration experiment (Figure 2.6). To build a titration curve, a series of PCR reactions with a single pair of 3C primers and different amounts of input 3C

template must be done. Table 1.1 gives examples of pairs of 3C primers which were successfully used in our lab for the titration of human and mouse libraries. We recommend selecting the library concentration from the middle of a linear region of an amplification curve in order to avoid both saturation of a signal (at high library concentrations) and loss of a signal (at low library concentration). Next PCR reactions are performed with each primer pair, using both the 3C ligation product library and the control ligation product library as a template. The relative interaction frequency of a pair of loci is then calculated by dividing the amount of PCR product obtained with the 3C ligation product library by the amount of PCR product obtained with the control library (see data analysis). By calculating this ratio one effectively normalizes for differences in primer efficiency. Since the control library is used to normalize for primer efficiency, reactions for each primer pair should be performed in both templates simultaneously, to reduce PCR variation as much as possible. Given that the control library contains all ligation products in equimolar amounts, all primer pairs should yield similar, though not identical, amounts of PCR products. When a pair of primers fails to amplify any product, or a product with the wrong size, these primers should be discarded and new primers should be designed (Figure 2.7).

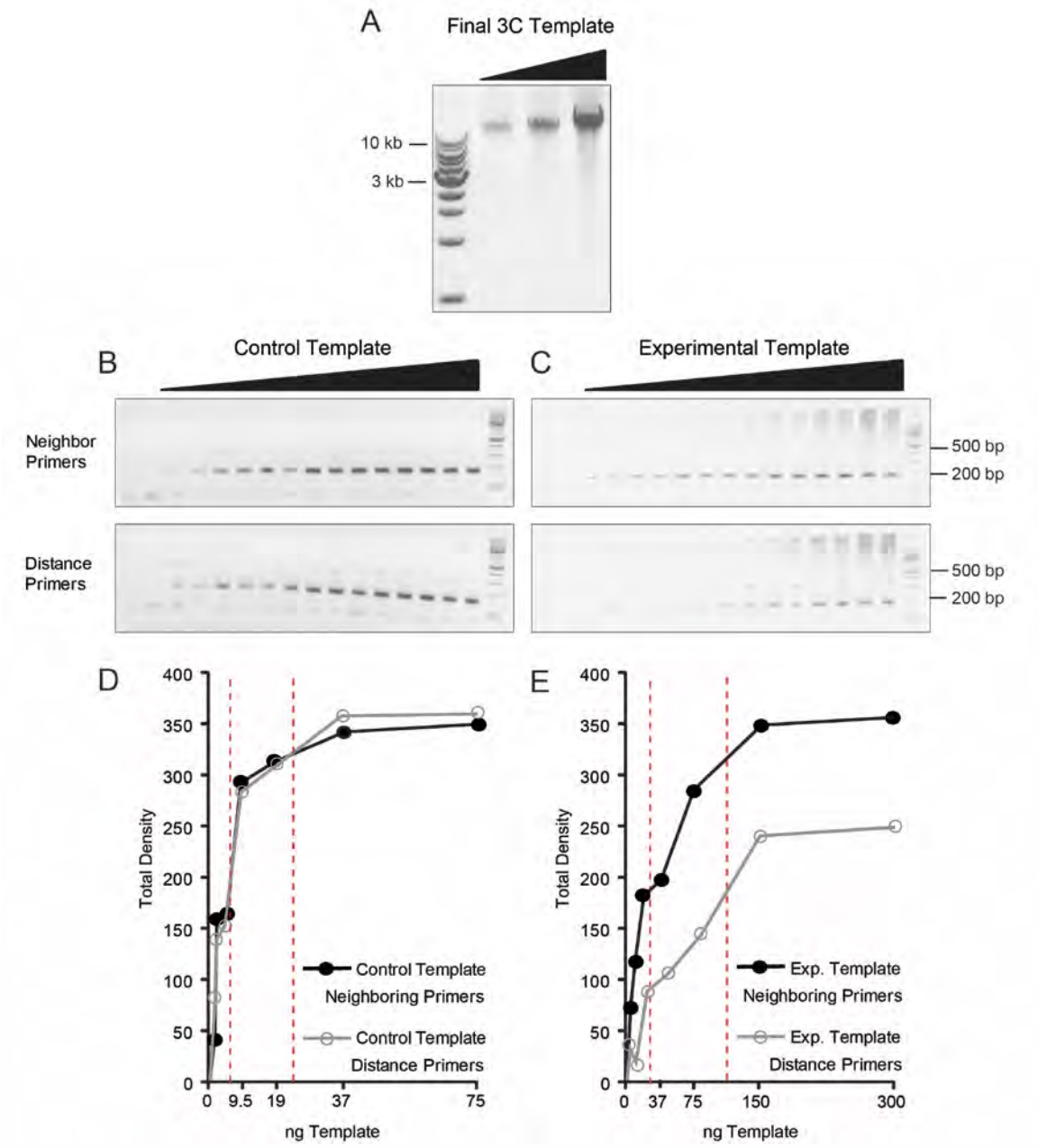

A
Final 3C Template
10 kb
3 kb
B
Control Template
C
Experimental Template
Neighbor Primers
Distance Primers
500 bp
200 bp
500 bp
200 bp
D
E
Total Density
ng Template
Control Template Neighboring Primers
Control Template Distance Primers
Exp. Template Neighboring Primers
Exp. Template Distance Primers
0
9.5
19
37
75
150
300
50
100
150
200
250
300
350
400

Figure 2.6: Examples of Final 3C Library and Titrations. A) An example of a completed 3C template run on a 0.8% agarose gel in increasing concentration. The main band of the template should run above 10 kb. There should be a minimal amount of smear underneath the band. If no band is seen, the template cannot be used for downstream analysis. **B, C**) Titrations of a control template (BAC) and an experimental template (mammalian cell line). Each band is present in duplicate. The upper gel represents ligation products from neighboring restriction fragments. These fragments are close together in both linear and three-dimensional distance, thus giving a strong signal in both the control and experimental templates, even at low template concentration. The lower gel represents ligation products from distal restriction fragments. These fragments are far apart in the linear genome and far apart in three-dimensional distance as well. The control template shows no difference between the neighboring and distal fragments, indicating there is no distance bias. In contrast, the experimental library shows a marked decrease of interaction frequency using the distance primers, which indicates that those fragments are indeed far apart in both linear and three-dimensional space. **D,E**) Graphs showing the quantification of gel bands from **B** and **C**. As described above, there is no difference in gel band intensity in the control template, while there is a difference in the experimental template. The dashed red lines indicate the shoulder of the graphs. When conducting the final 3C experiment, the working concentration in each PCR reaction should be chosen based on the concentrations in this area of the graph. Using too high or too low of a concentration will skew results.

Human				
EcoRI	Distant	57 kb	Runx1-14	AAAGTTCTCACGCACCGACTGAACACTCCA
			Runx1-18	GCCTGGGGTGTACCCACTTTCTATGGATAT
	Distant	37 Kb	GD8	GCAGCAAAGCAAACCAAAAGAACAACAGG
			GD14	GTCGCCGTTGCCTTTGCAGTTTACAGTG
	Neighbors	-	GD6	GTTTAAGACCCTCAGTATACTAGTCATAGAAGG
			GD7	GATGCCATTTCTTATCTTGTCTTGGCAGGTC
HindIII	Neighbors	-	AHF64	GCATGCATTAGCCTCTGCTGTTCTCTGAAATC
			AHF66	CTGTCCAAGTACATTCCTGTTCACAAACCC
Mouse				
HindIII	Neighbors	-	mGAPDH_3	TATCAAGGGTGCCCGTCACCTTCAGCTTTC
			mGAPDH_4	GGGCTTTTATAGCACGGTTATAAAGTGG

Table 2.1: List of Example 3C Primers for Library Titrations.

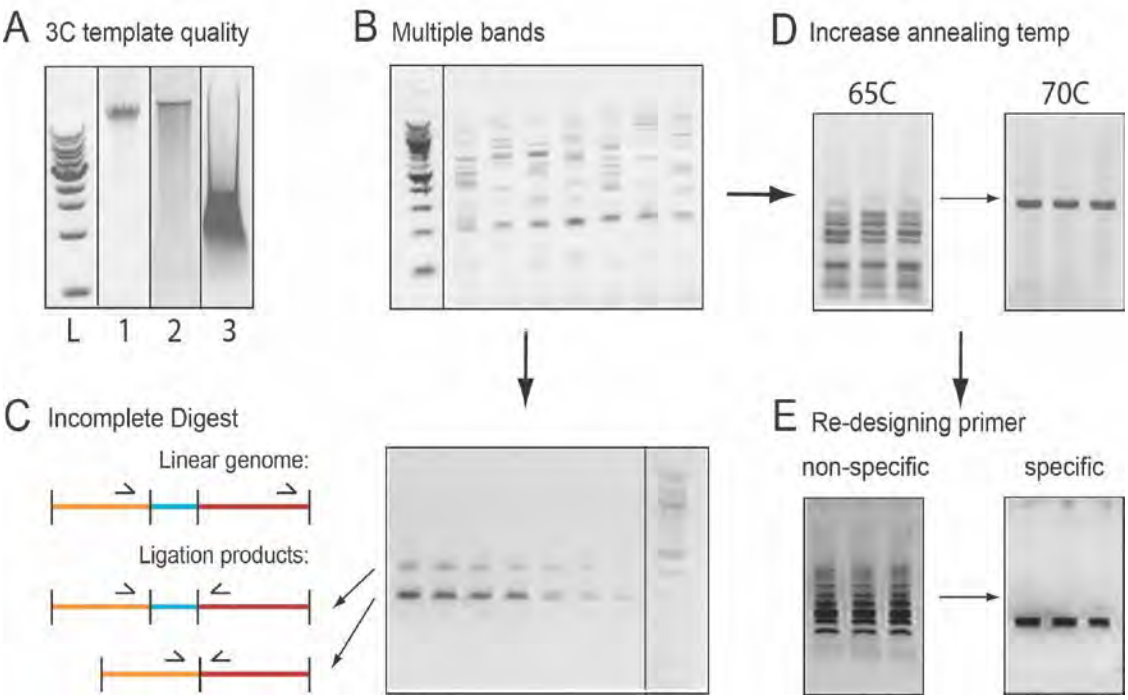


Figure 2.7: Examples of Common 3C Problems. A) Quality of 3C template. Typically, an excellent 3C template runs as one tight band above 10kb in size on 0.8% agarose gel (Lane 1). A mediocre 3C template (Lane 2) looks like a tight band with a noticeable smear running down the gel. Usually smeary templates like this are successfully used in 3C experiments. Lane 3 shows a completely degraded template, which should be excluded from downstream analysis. A 1kb ladder is shown in Lane L. **B-E**): Multiple bands problem and possible solutions. **B**) – Different pairs of 3C primers amplify several bands from a 3C template. There might be several reasons for that, which should be checked one by one. The first possibility is incomplete digestion of the chromatin (**C**). The figure represents the titration of a 3C template with same pair of primers (see the diagram on the left). The incomplete digestion results in two ligation products, with and without the blue restriction fragment, which are both detected by the same pair of 3C primers. If the size of the blue restriction fragment is relatively small, both ligation products could be amplified in the same PCR reaction. Usually severe incomplete digest, resulting in multiple extra bands of about same intensity as the "main" band (**B**), is seen in 3C templates made with a fine cutter restriction enzyme (one that recognized a 4-bp cut site). In this case one should consider re-making the 3C library with an increased amount of enzyme and elongated restriction time. The case presented in **C** is legitimate for further analysis. Next, if the size of the additional bands cannot be explained by incomplete digest, one should consider non-specific primer annealing. Note that even when a pair of primers amplifies one band of expected size from a (BAC)

control template, the more complex 3C template, made from the whole genome, might have additional sites for primer annealing. Simply increasing the annealing temperature in the PCR reaction can resolve the issue (**D**). If that's not the case, the primer pair (or at least the non-anchoring 3C primer) should be re-designed (**E**).

3C Experimental Protocol

Reagents required

Buffers

10X Ligation buffer
500 mM Tris-HCl pH7.5
100 mM MgCl2
100 mM Dithiothreitol
Autoclaved water
--- Store 20ml and 1ml aliquots at -80°C

1X Lysis buffer
10 mM Tris-HCl pH8.0
10 mM NaCl
0.2% Igepal CA-630 (NP40)
Autoclaved water
--- Store at 4°C

10X PCR buffer
600 mM Tris, adjust to pH 8.9 with H2SO4,
180 mM (NH4)2SO4
--- Store 1ml aliquots at -20°C

Reagents

Saturated phenol, pH 6.6±0.2 *bring pH to 8.0

Chloroform

Formaldehyde, 37% by weight

ATP

Proteinase K (Fungal)

Protease inhibitors

T4 DNA ligase

Amicon® Ultra – 0.5ml 30K

Creating a BAC control library (for use with complex genomes)

Digestion of BAC genomic DNA:

1. Digest the DNA in the following reaction overnight at 37°C with rotation

   10µg BAC DNA[5]
   10X restriction buffer
   10 mg/ml BSA (100X)
   Restriction enzyme (up to 10% of total reaction volume)
   Water (up to 500µl)

2. After digestion, add an equal volume of phenol (pH 8.0)/ chloroform (1:1), vortex for 30 seconds
3. Spin at 18,000 x *g* in a microfuge at room temperature for 5 minutes
4. Transfer the upper (aqueous) phase to a new 2 ml tube and add 1/10 volume of 3M sodium acetate, pH 5.2, vortex briefly
5. Add 2.5 volume of ice-cold 100% ethanol and invert the tube gently
6. Incubate at -20°C for at least 15 minutes
7. Spin at 18,000 x *g* in a microfuge at 4°C for 20 minutes
8. Wash the pellet in 1ml of 70% ethanol
9. Spin at 18,000 x g in a microfuge at 4°C for 15 minutes
9. Remove all supernatant
10. Briefly air-dry the pellet

[5] If using more than one BAC, combine them in equimolar amounts before digestion. Concentration of each BAC clone should be determined using qPCR with primers that recognize the common BAC vector backbone.

11. Resuspend the pellet in 44µl water and incubate at 37ºC for 15 minutes. Reserve 1µl of digestion product to run as a control on the final analytical gel.

BAC DNA ligation

12. Prepare the ligation reaction as follows and incubate at 16°C overnight

Digested BAC DNA	43µl
5X T4 ligase buffer	12µl
T4 DNA ligase	5µl
Final total volume	60µl

13. After overnight ligation, incubate at 65°C for 15 minutes to inactivate the ligase

Purification of BAC genomic DNA control template

14. Add 140µl of water to the ligation to make the final volume 200µl
15. Add an equal volume of phenol (pH 8.0)/chloroform (1:1), vortex 30 seconds and then spin at 18,000 x *g* in a microfuge at room temperature for 5 minutes
16. Transfer the upper (aqueous) phase to a fresh 1.7 ml tube
17. Repeat phenol/ chloroform extraction once
18. Transfer the upper phase to a fresh 1.7 ml tube
19. Add equal volume of chloroform, vortex 30 seconds, then spin at 18,000 x *g* at room temperature for 5 minutes
20. Transfer the upper phase to a new 1.7 ml tube
21. Add 1/10 volume of 3M sodium acetate, pH 5.2, briefly vortex

22. Add 2.5 volumes of ice-cold 100% ethanol and invert the tube several times
23. Incubate at -20ºC for at least 15 minutes
24. Spin at 18,000 x *g* in a microfuge at 4°C for 20 minutes
25. Wash the pellet in 1 ml of 70% ethanol, spin at 18,000 x *g* in a microfuge at 4°C for 15 minutes
26. Briefly air-dry the pellet
27. Resuspend the pellet in 100µl of 1X TE, pH 8.0
28. Incubate at 37°C for 15 minutes
29. Store the 3C control template at -20°C. This library can be stored for a few years.

Creating a 3C library from mammalian cells

Crosslinking of cells

1. Centrifuge 1 x 10^8 cells at 300 x *g* for 10 minutes[6]
2. Resuspend the cell pellet in 45 ml of fresh medium, mix by pipetting

[6] In cases, when primary cells [38], [39], [40] or siRNA treated cells [38] are being analyzed by 3C, getting 10^8 cells might be very difficult. Several groups has successfully applied 3C to 2 – 10 x 10^6 cells and in some of those studies the original protocol has been modified and we advice to refer to the original studies. In general, while working with primary tissues it is necessary to break down tissues by applying collagenase to single-cell suspension before cross-linking [41]. Time of lysis can be increased for up to 2h as well as treatment with SDS prior restriction can be more severe [39]. At last, one might want to use qPCR to quantify 3C signal [41].When number of cells is not an issue, we strongly recommend starting with large number of cells, so one would have enough material for making repeats in case it is needed.

3. Add 1.25 ml of 37% formaldehyde to obtain 1% final concentration and pipette to mix
4. Incubate at room temperature for 10 minutes on a shaker
5. Add 2.5 ml of 2.5M glycine to stop crosslinking
6. Incubate at room temperature for 5 minutes and then on ice for 15 minutes
7. Spin down at 4°C 800 x *g* for 10 minutes and remove as much supernatant as possible[7]
8. Add 2 ml of cold Lysis buffer and 200 µl of 10X Protease Inhibitors to each pellet and resuspend it well on ice
9. Incubate on ice for at least 15 minutes to let the cells swell
10. Lyse the cells on ice using a Dounce homogenizer (pestle A). Use 10 slow strokes, then let the cells rest for 1 minute on ice. Complete lysis by applying 10 more strokes
11. Transfer to 2 microcentrifuge tubes, spin at 2,200 x *g* in a microfuge for 5 minutes
12. Wash each pellet with 1 ml of 1X NEB buffer 2 (or the buffer appropriate to your restriction enzyme, however do not use NEB4 as we have found it is not compatible with 3C buffers) and spin down at 2,200 x g in a microfuge for 5 minutes
13. Repeat Step 12 once
14. Resuspend all the pellets from the same sample in 1 ml of 1X NEB buffer

[7] The experiment can stop at this point. Store the pellet at -80°C by incubating the pellet on dry ice for 20 minutes and then store at -80°C for up to at least two years.

Digestion

15. Distribute the suspension over 20 1.7 ml microcentrifuge tubes, 50 µl per tube. (Optional) save 10µl of lysate for the chromatin integrity check.
16. (Optional) Chromatin integrity control. Take 10 µl of lysed cells from the previous step. Add 50 µl of 1x NEB and 10 µl of Proteinase K (10mg/ml). Incubate for 30 min at 65°C. Purify DNA by single phenol-chloroform extraction without ethanol precipitation. Add 5 µl of RNAseA (10mg/ml) to the aqueous phase and incubate for 15 min at 37°C. Check quality of the sample by running it on 0.7% agarose gel. The sample is good if DNA is either stuck in the well or runs as a single high molecular weight band (>23 kb).
17. Add 312 µl of corresponding 1X NEB buffer per tube.
18. Add 38 µl of 1% SDS per tube and mix well by pipetting up and down, avoiding air bubbles
19. Incubate at 65°C for exactly 10 minutes, then place on ice immediately[8]
20. Add 44 µl of 10% Triton X-100 to each tube to quench SDS, mix well, avoiding air bubbles
21. Add 400 U of the restriction enzyme (up to 10% total reaction volume) of choice to each tube, mix well
22. Incubate at 37°C overnight with rotation

Ligation and reverse crosslink

[8] This step removes proteins that are not crosslinked to the DNA.

23. Add 86 µl of 10% SDS to each tube, mix carefully to avoid making air bubbles
24. Incubate at 65°C for 30 min to inactivate the enzyme. Place tubes on ice immediately after incubation
25. Prepare the ligation cocktail master mix as follows (except ligase) and distribute 7.61 ml of cocktail to each of 20 pre-chilled 15 ml conical tube

Ligation cocktail	per reaction
10% Triton X-100	745 µl
10X Ligation buffer	745 µl
10 mg/ ml BSA	80 µl
100 mM ATP	80 µl
Milli-Q water	5960 µl

26. Transfer each digestion product from Step 23 into each of the 15 ml conical tubes with ligation cocktail. Keep on ice.
27. Add 50 U (10 uL) of T4 DNA ligase (Invitrogen) to each tube
28. Mix by inverting the tubes several times and spin briefly
29. Incubate at 16°C for 2 hours
30. Add 50 µl of fresh 10 mg/ ml proteinase K solution per tube
31. Mix by inverting the tubes several times and spin briefly
32. Incubate at 65°C for 4 hours
33. Add 50 µl of 10 mg/ ml fresh proteinase K per tube and continue incubating at 65°C overnight[9].

[9] Alternatively, one can incubate overnight after the first proteinase K treatment, add the second proteinase K aliquot the next morning and incubate at 65°C for 2 hours

DNA Purification

34. Transfer and combine reactions into 50 ml conical tubes, 20 ml per tube
35. Add equal volume Phenol per tube, vortex for 2 minutes and spin at 3,000 x *g* for 10 minutes
36. Transfer the aqueous phase to fresh 50 ml conical tubes
37. Add an equal volume of Phenol (pH 8.0)/ Chloroform (1:1) to each tube, vortex for 1 minute and spin at 3,000 x *g* for 10 minutes
38. Pool aqueous phases and distribute equally in new 50 ml conical tubes
39. Add 1x TE buffer, pH 8.0 to a total volume of 50 ml per tube to prevent DTT from precipitating.
40. Transfer each 50 ml pool to a 250-ml screw-cap centrifuge bottle that is suitable for high-speed spinning
41. Add 1/10 volume of 3M sodium acetate, pH 5.2 per bottle, vortex briefly
42. Add 2.5 volumes of ice-cold 100% ethanol per bottle, mix well by inverting the bottles several times
43. Incubate at -80°C at least 60 minutes or overnight
44. Spin at 12,000 x *g* for 20 minutes at 4°C (Make sure the mixture has thawed completely on ice before spinning down)
45. Dissolve each pellet in 450 µl 1x TE buffer, pH 8.0 and transfer to a 15 ml conical tube

46. Wash the bottle with 450 µl of 1x TE buffer, pH 8.0 and transfer to the same tubes as in step 44. *If dissolving is difficult, incubate samples at 65 °C no more than 10 minutes*
47. Add an equal volume of phenol (pH 8.0) to each tube and vortex for 1 minute; then spin at room temperature at 3,000 x *g* for 5 minutes
48. Transfer the upper aqueous phases to fresh 15 mL tubes
49. Add equal volume of phenol (pH 8.0)/ chloroform (1:1), vortex 1 minute and spin at 3,000 x *g* for 5 minutes

    If the interfaces are clear, go to step 49; otherwise repeat step 48 once
50. Transfer the aqueous phases into clean 1.7 ml centrifuge tubes (about 400µl per tube)
51. Add 1/10 volume of 3M sodium acetate, pH 5.2, vortex briefly
52. Add 2.5 volumes of ice-cold 100% ethanol, invert gently
53. Incubate at -80°C at least 1 hour
54. Spin at 4°C at full speed in a microfuge for 20 minutes
55. Remove all supernatant and briefly air dry the pellets
56. Dissolve each pellet in 500 µl of 1X TE buffer, pH 8.0
57. Insert an Amicon ultra-0.5 30K device into one provided microcentrifuge tube
58. Transfer DNA sample to the Amicon ultra filter device (up to 500 µl) and cap it
59. Place the capped devices into the rotor of a microfuge, aligning the cap strap towards the center of the rotor

60. Spin at 18,000 x *g* for 5 minutes and discard the flow-through from the collecting tubes
61. Add 450 µl of 1X TE to each filter device, spin at 18,000 x *g* for 5 minutes and discard the flow-through
62. Repeat wash step 3 times
63. After the final wash, place the filter device upside down in a new provided collecting tube
64. Place the tubes in centrifuge, aligning the cap strap towards the center of the rotor, spin at 1,000 x *g* for 5 minutes
65. Transfer all collected DNA from the same sample to a 1.7ml centrifuge tube
66. Add 1X TE to each sample to make a total volume of 1ml
67. Add 1 µl of 10 mg/ ml of DNase-free, RNase A and incubate at 37°C for 15 minutes
68. Load 1 µl and 2 µl of 10X diluted template on 0.8% agarose/ 0.5X TBE gel, along with a sample of a 1Kb DNA ladder to check quality and quantity of the template
69. Store the 3C template up to 2 years at -20°C (kept in aliquots)

Determination of quantity and quality of 3C Libraries

Before embarking on quantitative analysis of the 3C library, one first has to determine the amount of 3C library to use in each PCR reaction. To do this, one can perform a titration experiment, as shown in Figure 2.2. Both the BAC

control template and the experimental 3C library should be titrated using a serial two-fold dilution series beginning with 240 ng of 3C template and 25 -50ng of BAC control template.[10] We routinely use two different primer pairs for the control region for this analysis. The first primer pair interrogates a short-range interaction (i.e. a pair of restriction fragments separated by only a few thousand bp in the genome). The second primer pair is chosen to interrogate a longer-range interaction (i.e. a pair of restriction fragments separated by tens of thousands of bp). We suggest performing each PCR reaction in duplicate. PCR products are run on a 2% agarose gel and quantified using a standard gel imaging set up. A water control should be included.

The amount of PCR product is then quantified and plotted versus the amount of input DNA. The resulting titration curve should plateau to a flat shoulder, as shown in Figure 2.6 D and E. The concentration of 3C template to use in 3C experiments should be taken from the linear slope of the graph to ensure that one will not over-or under-saturate signals from the 3C library. The PCR reaction is assembled as follows:

Titration Reaction:

10x PCR Buffer	2.5 µl
50 mM MgSO4	2 µl
20 mM dNTPs	0.2 µl
80 µM Primer1	0.125 µl
80 µM Primer2	0.125 µl

[10]**BAC Dilution Series:** The starting amount of a control template depends on the BAC composition of the template: the more complex the control library is, i.e. the more possible interactions it is covering, the more DNA is needed in a PCR reaction. In our experience, using BAC clones covering up to a megabase of genomic DNA, starting the dilution series with 50-70 ng of the template works well.

Taq Polymerase	0.2 µl
Diluted template	4 µl
dH2O	15.85 µl
Total Volume	25 µl

PCR conditions:
1. 95 C 5 minutes
2. 95 C 30 seconds
3. 65 C 30 seconds
4. 72 C 30 seconds
5. 95 C 30 seconds
6. 65 C 30 seconds
7. 72 C 8 minutes
8. 10 C forever

Repeat steps 2-4 34 times.

Generating 3C Data

After choosing the appropriate concentration of both the control and the experimental 3C library from the titration analysis, one can start to determine interaction frequencies between pairs of loci. To do so, one uses pairs of primers for restriction fragments of interest to perform semi-quantitative PCR on each of the two templates. Each PCR reaction is performed in triplicate. The PCR conditions are identical to the ones used to titrate the 3C and control libraries. PCR products are run on a 2% agarose gel and the amount of PCR product is quantitated using a standard gel quantification set up. 3C products can also be quantified using qPCR, with very similar results (Hagège et al., 2007).

5C Experimental Protocol

Reagents Required

Reagents

10x NEB4

Isopropanol

1 mg/mL Salmon Sperm DNA

Taq Ligase

Taq Ligase Buffer

Amplitaq Gold

Supplied with 25 mM Deoxyribonucleotides (dNTPs)

25 mM $MgCl_2$ (Magnesuim Chloride)

PCR buffer

Pfu Ultra II

Supplied with 25 mM Deoxyribonucleotides (dNTPs)

25 mM $MgCl_2$ (Magnesuim Chloride)

PCR buffer

Taq Polymerase

10 mM Adenosine Triphosphate (ATP)

T4 DNA Ligase

T4 DNA Ligase buffer

PE Adapter Oligo Mix

Kits

QIAquick PCR Purification Kit (Qiagen)

QIAquick Gel Extraction Kit (Qiagen)

MinElute PCR Purification Kit (Qiagen)

Library Quantification Kit, Illumina Genome Analyzer v2 (KAPA)

Primers

PE adapter oligo mix

Phosphorylated Universal Primers (Standard Desalting):

FWD: 5' - /5Phos/CCTCTCTATGGGCAGTCGGTGAT – 3'

REV: 5' – /5Phos/CTGCCCCGGGTTCCTCATTCTCT – 3'

PE Primers (HPLC Purification):

1.0: 5' –

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC

CGATCT – 3'

2.0: 5' –

CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCT

CTTCCGATCT – 3'

5C Library Titration

Before beginning the 5C protocol, the 3C library must be titrated with the 5C probes to determine the appropriate number of genome copies and primer

concentration to use in the downstream steps. To determine the number of human genome copies, we use the following numbers:

1 Dalton = 1.65×10^{-24} gram

1 base pair = 650 Daltons

Thus, 1 base pair = 1.0725×10^{-12} gram

And, 3 billion base pairs (1 genome copy) = 3.2×10^{-12} gram = 3.2 picograms

Therefore, 100,000 genome copies = 320 nanograms

We recommend starting the titration at 100,000 genome copies and doubling the amount (200,000, 400,000, etc.). We also recommend beginning the probe titration at 1 fmol/probe (1 nM probe). In our hands most libraries are made using 1 or 2 fmol/probe. The following reaction is used for titration:

5C Annealing Reaction[11]

10x NEB4	2 uL
1 mg/mL ssDNA	filler for controls (replaces 3C library)
3C Template	X Genome copies
5C Probes	X fmol/probe
dH_20	up to 20 uL
Incubate:	95˚ for 9 minutes
	Drop temperature to 55˚ at a rate of 0.1˚/second
	55˚ overnight

5C Ligation Reaction[12]

[11] It is important to include 3 controls whenever the annealing reaction is performed: no Taq ligase, no 3C library, and no 5C probes. The no Taq ligase control is necessary for the ligation reaction but must be set up at this step.

Taq Ligase buffer	2 uL
Taq ligase	0.25 uL
dH_20	17.75 uL

Incubate: 55˚ 1 hour
75˚ 10 minutes
4˚ 10 minutes

5C Universal PCR Reaction

5C ligated template	6 uL
10x PCR buffer	2.5 uL
25 mM $MgCl_2$	1.8 uL
25 mM dNTP	0.2 uL
Amplitaq Gold	0.225 uL
80 uM Forward Primer	0.5 uL
80 uM Reverse Primer	0.5 uL
dH_20	13.275 uL

A water control should always be included with this reaction.

Incubate 95˚ 9 minutes
95˚ 30 seconds
65˚ 30 seconds
72˚ 30 seconds
Repeat 27x[13]
72˚ 7 minutes
10˚ forever

After this step, run the products out on a gel (Figure 2.2C). Choose the appropriate number of genome copies and probe concentration. In our hands 400,000 genome copies and 2 fmol/probe are generally successful. If there are any bands in the control lanes, the library has been contaminated and should not be used in downstream steps (Figure 2.8). See the troubleshooting section for tips on avoiding and solving contamination issues.

[12] Remember to include a no Taq ligase control. This reaction is added directly to the annealing reaction.
[13] This cycle number can also be modified, though 30 cycles is the maximum we recommend.

Creating the 5C Library

To create the 5C library, repeat the reactions as outlined in the Titration section. Performing 10 annealing reactions is enough to produce sufficient product for downstream steps. After the ligation step, we suggest taking at least 4 reactions from each ligation product for the universal PCR, resulting in at least 40 universal PCR reactions plus the water control. After the universal PCR is run, combine all the reactions and then run a gel to ensure there is no contamination in the control lanes (Figure 2.8). If the control lanes are clean, it is time to move on to the next part of the protocol.

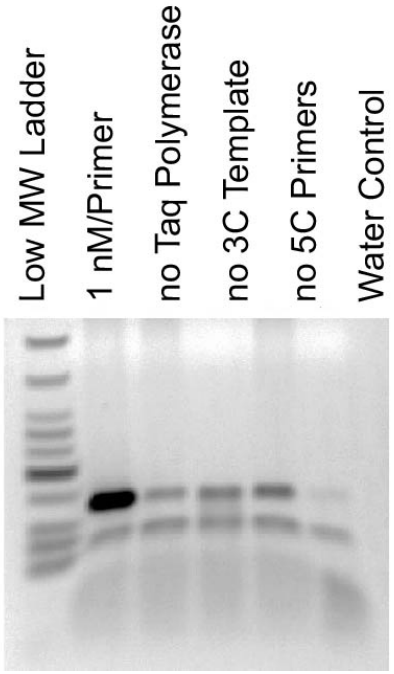


Figure 2.8: 5C library with Contamination in the Control Lanes. Lane 1 shows a 5C library with an expected size of 138 bp. Lanes 2-5 are controls, and should not show a band at 138 bp. The lower band in all lanes represents primer dimers. This library should be re-made.

Preparing the 5C Library for Sequencing

To prepare the 5C library for sequencing, we first concentrate the library and then run it all out on a 1.8% agarose gel.

Concentrate with QIAquick column

1) Add 5 volumes buffer PB to 1 volume 5C library
2) Spin for 30 seconds at 6000 rpm, then for 30 seconds at full speed
3) Discard the flow-through, and repeat step 2
4) Wash with 750 mL buffer PE
5) Spin for 30 seconds at full speed, discard the flow-through, and place the column in a new tube
6) Spin for 1 minute full speed and place the column in a collection tube
7) Incubate the column with 50 uL hot buffer EB[14] for 3 minutes
8) Spin 1 minute full speed
9) Add 100 uL EB buffer to the eluate and proceed to gel purification

Gel Purification

10) Pour a long 1.8% agarose gel using broad combs
11) Load the entire 5C concentrated library alongside a low MW ladder
12) Run the gel at 4˚ at 200 volts for 90 minutes

Gel Extraction

[14] Heating the EB buffer at 65˚ before use enhances the amount of DNA retrieved from the column.

13) Carefully cut out the band of interest from the gel (the band will run at 135 bp)
14) Weigh the gel slice, then chop into fine pieces to aid in melting
15) Add 3x the gel slice weight of buffer QG
16) Melt the gel slice by vortexing and hand heat (DO NOT place the gel slice at high temperature in a heating block or water bath, this will lower the yield).
17) Add 1x the gel slice weight of isopropanol
18) Transfer the liquid to a QIAquick column and spin for 30 seconds at 6000 rpm, then for 1 minute at full speed
19) Discard the flow-through
20) Wash the column with 500 mL buffer QG and spin 1 minute full speed
21) Discard the flow-through
22) Wash the column with 750 mL buffer PE and spin at 1 minute full speed
23) Discard the flow-through
24) Repeat step 22 and 23 once
25) Spin 1 minute full speed in a new tube
26) Place the column in a collection tube and incubate with 31 µL hot buffer EB for 3 minutes
27) Spin for 2 minutes full speed

Addition of "A" bases

28) Add directly to the eluate: 3.5 uL 10x Taq buffer

1.2 uL Taq polymerase

1.2 uL 10 mM ATP

29) Incubate the reaction for 45 minutes at 72˚

Purify and Concentrate with MinElute column

29) Add 175 µL buffer PB to the above reaction

30) Spin for 30 seconds at 6000 rpm, then 1 minute full speed

31) Discard flow-through

32) Wash with 750 µL buffer PE, spin for 1 minute full speed

33) Discard flow-through

34) Place column in new tube, spin 1 minute full speed

35) Place column in collection tube and add 31 µL hot buffer EB

36) Incubate for 3 minutes, then spin for 2 minutes full speed to elute the DNA.

Ligation of PE Adapters

37) Add to the above eluate: 10 µL 5x T4 DNA Ligase buffer

4 uL T4 DNA ligase

7 uL PE adapter oligo mix

38) Incubate the reaction for 2 hours at room temperature

Gel Purification

39) Pour a long 1.8% agarose gel using broad combs

40) Load the entire 5C concentrated library alongside a low MW ladder

41) Run the gel at 4˚ at 200 volts for 90 minutes

Gel Extraction

42) Carefully cut out the band of interest from the gel (the band will run at 135 bp)

43) Weigh the gel slice, then chop into fine pieces to aid in melting

44) Add 3x the gel slice weight of buffer QG

45) Melt the gel slice by vortexing and hand heat (DO NOT place the gel slice at high temperature in a heating block or water bath, this will lower the yield).

46) Add 1x the gel slice weight of isopropanol

47) Transfer the liquid to a QIAquick column and spin for 30 seconds at 6000 rpm, then for 1 minute at full speed

48) Discard the flow-through

49) Wash the column with 500 mL buffer QG and spin 1 minute full speed

50) Discard the flow-through

51) Wash the column with 750 mL buffer PE and spin at 1 minute full speed

52) Discard the flow-through

53) Repeat step 51 and 52 once

54) Spin 1 minute full speed in a new tube

55) Place the column in a collection tube and incubate with 20 µL hot buffer EB for 3 minutes

56) Spin for 1 minute full speed

57) Repeat elution

Amplification of the 5C library

Amplification Reaction

5C Library	2 µL
10x Pfu Ultra II buffer	5 µL
25 mM dNTP	0.5 µL
25 µM PE Primer 1	0.7 µL
25 µM PE Primer 2	0.7 µL
Pfu Ultra II	1 µL
dH_20	31.8 µL

Incubate: 95˚ 2 minutes
95˚ 20 seconds
65˚ 20 seconds
72˚ 15 seconds
} Repeat 15-17x (16 cycles is recommended)
72˚ 3 minutes
10˚ forever

Gel Purification

58) Pour a long 1.8% agarose gel using broad combs

59) Load the entire 5C concentrated library alongside a low MW ladder

60)Run the gel at 4˚ at 200 volts for 90 minutes

Gel Extraction

61)Carefully cut out the band of interest from the gel (the band will run at 135 bp)

62) Weigh the gel slice, then chop into fine pieces to aid in melting

63) Add 3x the gel slice weight of buffer QG

64) Melt the gel slice by vortexing and hand heat (DO NOT place the gel slice at high temperature in a heating block or water bath, this will lower the yield).
65) Add 1x the gel slice weight of isopropanol
66) Transfer the liquid to a QIAquick column and spin for 30 seconds at 6000 rpm, then for 1 minute at full speed
67) Discard the flow-through
68) Wash the column with 500 mL buffer QG and spin 1 minute full speed
69) Discard the flow-through
70) Wash the column with 750 mL buffer PE and spin at 1 minute full speed
71) Discard the flow-through
72) Repeat step 70 and 71 once
73) Spin 1 minute full speed in a new tube
74) Place the column in a collection tube and incubate with 15 µL hot buffer EB for 3 minutes
75) Spin for 1 minute full speed
76) Repeat elution
77) Run a small 2% agarose gel to check the final library – it should now be just above 250 bp in size.

Library Quantification

We use the KAPA library quantification kit to accurately assess the amount of 5C library before we send the samples to be sequenced. The KAPA kit provides DNA standards and primers for qPCR analysis of DNA concentration in the 5C library.

Sequencing 5C libraries

At this point the 5C libraries are ready to be sequenced. With the scheme presented here, libraries are sequenced on the Illumina GAIIx machine using 50 bp paired end reads. To sequence 5C libraries using other platforms, simply change the PE adapters to fit the platform of choice. Some sequencing facilities will perform sample preparation for you – in this case one should send the 5C library before addition of PE adapters, just after the universal PCR step.

Analysis of 3C Data

A typical 3C experiment includes the analysis of three biological replicates of the 3C and control library. Further, each interaction frequency of interest is determined by three PCR reactions (technical replicates) using each of the three 3C and control libraries. Ideally, all 3C reactions should be prepared with the same PCR master mix and run simultaneously in the same PCR block, but practically it is not always possible if the experiment covers a large region. To minimize experimental noise we recommend that PCR replicates for the 3C library and the control library are performed in parallel and run side-by-side. We use LabWorks software (version 4.0, BioImaging Systems) to analyze the intensity of each band minus the background on an agarose gel. Then one calculates the average of the three technical replicates. Thus for each biological replicate one obtains an average value for the interaction frequency of each pair of loci. Finally, the three datasets obtained with the three biological replicates are normalized to each other so that they are all on the same scale. This allows the data from different replicates and different conditions to be directly compared. Below we present an illustration of how 3C data can be calculated and how 3C datasets obtained for different cells or conditions can be quantitatively compared.

Example of a 3C Analysis

We describe a 3C analysis of a gene in two cell lines, A and B. A BAC-based control library was also generated. In this example, cell line A expresses the gene of interest while cell line B does not. In this analysis interactions between a single anchor restriction fragment, containing the gene promoter, and 20 flanking restriction fragments were determined to generate a 3C interaction profile. In addition, a control region (ENCODE region Enr313) was analyzed to obtain a set of interactions that are assumed to be identical in cell line A and B.

Calculation of Interaction Frequencies for Each Biological Replicate

The first step in analyzing a 3C experiment is to average the technical replicates. Simply average the technical replicates and find the standard deviation for each primer pair. In order to control for primer efficiency, divide each averaged technical replicate by its corresponding averaged control template value. In this example, for each pair of primers the averaged value of three PCR reactions performed on a given 3C library for each cell line (A or B) is divided by the corresponding average of three PCR reaction performed in the BAC control library. For example, the interaction frequency for a given primer pair is (Equation 1):

$$\frac{(A_1 + A_2 + A_3)}{(CL_1 + CL_2 + CL_3)}$$

where the values A_1, A_2 and A_3 represent three technical replicates for that primer pair in cell line A, and CL_1, CL_2 and CL_3 represent three technical replicates for the same primer pair in the control library. This value is the interaction frequency of a pair of loci for a given biological replicate. In order to calculate the standard deviation of each interaction frequency, use the following formula (Equation 2):

$$StDev = sqrt\left[\left(\frac{StDev_{Control\ Library}}{Avg_{Control\ Library}}\right)^2 + \left(\frac{StDev_{Exp.\ Library}}{Avg_{Exp.\ Library}}\right)^2\right] * Interaction\ Frequency$$

where StDev is the standard deviation, Avg is the average, Exp. Library is the experimental 3C library value, and Interaction Frequency is the value calculated using Equation 1.

Using this approach the average interaction frequency, and the standard deviation, for each pair of loci is determined for each of the three biological replicate 3C libraries.

Combining Biological Replicates

The results from individual biological replicates can be directly compared in separate graphs (to determine whether the same peaks occur in each replicate) or the biological replicates can be averaged together. To calculate the

combined standard deviation of all biological replicates, use the following formula (Equation 3):

$$StDev = sqrt\left[\left(\frac{StDev_{B.Rep1}}{Avg_{B.Rep1}}\right)^2 + \left(\frac{StDev_{B.Rep2}}{Avg_{B.Rep2}}\right)^2 + \left(\frac{StDev_{B.Rep3}}{Avg_{B.Rep3}}\right)^2\right] * Avg_{All\ Reps}$$

where StDev is the standard deviation, Avg is the average, B is biological, and Rep is replicate.[15]

Normalizing 3C Data Obtained from Different Experiments

In order to allow direct quantitative comparison between two 3C datasets, e.g. two different biological replicates and/or data obtained with two different cell lines, they must be first normalized to each other. This normalization is done using the interaction frequencies measured within the control genomic region, which was selected based on the assumption that it has the same conformation in both cell lines. In our example we have analyzed two cell lines (A and B), and have three biological replicates. Here we provide an example of how these

[15]**Standard Deviation versus Standard Error of the Mean**. We suggest that when plotting the average of biological replicates to also display the standard deviation (SD) of each data point and not the standard error of the mean (SEM). The standard error of the mean reflects the certainty with which the average can be estimated. The SEM incorporates the number of measurements taken because the more measurements made, the more likely it is that the correct value has been found. This is a valid error to plot while examining an individual biological replicate. A large SEM indicates that the value of a give data point is very uncertain, and it may be necessary to perform additional technical replicates to increase the precision of measuring that specific value. However, it is more informative to indicate the SD while comparing biological replicates, since the SD will better reflect potentially relevant variation between samples.

datasets can be compared. We will normalize the data for cell line B to the data for cell line A, for each of the three biological replicates separately, so that three independent cell line comparisons are obtained.

First, the interaction frequencies are calculated for each biological replicate for cell line A and B, as described above, for each of the pairs of loci in the control region. Next, a normalization factor is calculated to normalize the data for one experiment to the other, e.g. to normalize the data obtained with cell line B to the data obtained with cell line A. For this, determine the log ratio for each interaction frequency in the control region (Equation 4):

$$\log \frac{\sum A_1 + A_2 + A_3}{\sum B_1 + B_2 + B_3}$$

where A_1, A_2 and A_3 are the normalized interaction frequencies of one primer pair in the first cell line, and B_1, B_2 and B_3 are the normalized interaction frequencies of one primer pair in the second cell line. Next, the average of these log ratios is calculated. The normalization factor is then found by taking the inverse log of this average. Finally, the entire dataset that was in the denominator in the calculation, in this case, cell line B biological replicate 1, is multiplied by the normalization factor. Each normalized interaction frequency is multiplied by the normalization factor. This will bring the two datasets to the same scale, and they can now be plotted on the same graph. This analysis is then repeated for replicates 2 and 3 individually. One can subsequently average the normalized biological replicates

together to obtain a final data set for each cell line, and these final data sets can be plotted on one graph.

Find Peaks of High Interaction Frequency

3C interaction frequencies are typically plotted versus genomic position with respect to the anchor point (Figure 2.4). In general interaction frequencies will decrease rapidly with increasing genomic distance. A specific looping interaction can be inferred when a peak is observed on top of this overall 3C profile. Visual inspection of 3C profiles has been used to identify such looping interactions. To obtain further support for a looping interaction additional analyses are essential, e.g. 3C analysis of cells or conditions where the looping interaction is absent. Figure 2.4 also illustrates the importance of obtaining a larger 3C profile so that a local background in non-specific 3C interactions is obtained. In the absence of this baseline estimation it is not possible to identify peaks in 3C interaction profiles and can lead to misinterpretation of individual 3C signals. In some cases it is possible to apply ANOVA statistic test for peak calling on the 3C profile, but usually there are not enough data points to perform in-depth analysis (McCord et al., 2011).

If a long-range interaction is inferred, it may be necessary to perform further experiments to validate the finding. For instance one can analyze the genomic region using a different restriction enzyme to confirm the looping interaction. Comparison of 3C data to other types of data sets, such as histone

modification patterns or DNase hypersensitive sites, and analysis of the looping interaction across cell types that do or do not express the gene of interest can help further define the functional elements involved and the role of the interaction in gene expression. Final experimental confirmation of the looping interaction and the DNA elements involved can be obtained by deleting or mutating the interacting regions and/or knocking down transcription factors that may mediate the interaction.

Analysis of 5C Data

5C experiments involve a number of steps that can locally differ in efficiency, thereby introducing biases in efficiency of detection of pairs of interactions. These biases could be due to differences in the efficiency of crosslinking, the efficiency of restriction digestion (related to crosslinking efficiency), the efficiency of ligation (related to fragment size), the efficiency of 5C probes (related to annealing and PCR amplification) and finally the efficiency of DNA sequencing (related to base composition). All of these potential biases—several of which are common to other approaches such as chromatin immunoprecipitation (for example, crosslinking efficiency, PCR amplification, base-composition-dependent sequencing efficiency)—will have an impact on the overall efficiency with which long-range interactions for a given locus (restriction fragment) can be detected.

Analysis of 5C data can be accomplished via a published suite of tools called my5C (http://my5C.umassmed.edu) (Lajoie et al., 2009). We use an updated version of the novoalign mapping algorithm to map our reads (V2.07.11 http://novocraft.com). 5C data sets are scanned for misbehaving primers (primers that are blowouts or severely underperforming), which are then removed from downstream analysis. The data sets are also scanned for individual blowout points between a particular forward-reverse probe pair, termed singletons. These blowout points are also removed from downstream analysis.

The data sets are then coverage corrected in order to allow equal visibility to all probes. Once the biases are corrected, the data is run through a peak-calling pipeline to identify regions of high interaction frequency that are present in both replicates.

Probe filtering – Cis-Purge

Not all probes are represented equally in our 5C dataset due to over- and under-performance in the assay. As the first step in our data correction pipeline, we remove probes that perform significantly differently than the overall set. This correction is achieved using a Loess smoothing algorithm. A global average Loess is calculated for each data set. This global average Loess is compared to individual probe Loess plots. If the individual Loess is more or less than .85 of the scaled Z-score distance from the average global Loess, the probe is flagged as problematic. If a probe is flagged as problematic in more than 40% of the data, it is removed from downstream analysis.

Singleton removal

In some 5C datasets there are instances when the interaction between two probes was higher than any of its neighboring interactions by an order or magnitude or more. These blowout reactions should be removed from the dataset to avoid problems downstream during peak calling. We suggest placing a high threshold on the entire data set and removing any data points with interaction frequencies above that threshold.

Coverage correction

Once the outlier probes and interactions are removed from the 5C data set, it is important to normalize each probe in order to compare them to each other. Generally, one uses trans space for this type of correction, though it is possible to correct a 5C experiment using only cis data (Smith et al in preparation, chapter 4 of this thesis). An average Loess value is computed for the trans dataset. Separately, a Loess value was calculated for each individual probe in the combined dataset. Then a factor is generated by which each probe should be lifted or lowered in order to match the average Loess profile from the combined dataset.

$$\frac{Value - Average}{Standard\ Deviation}$$

The calculated factors were applied to the individual, read-normalized datasets to produce the final coverage corrected datasets.

Peak calling

To detect significant looping interactions over background we developed an in-house '5C peak calling' algorithm (Lajoie et al., 2009; Sanyal et al., 2012). We call peaks in each 5C biological replicate separately and then take only the peaks that intersect across replicates as our final list of significant looping interactions.

Troubleshooting

Degraded Template

Degradation of a 3C template can be observed when the DNA is run on an agarose gel (Figure 2.7A). In our experience this degradation occurs early in the 3C protocol, often at the step where cells are lysed. This may be due to contaminating nucleases. The quality of a cell pellet can be checked in the first steps of the 3C protocol. If the 3C template is degraded we recommend replacing all plastics and buffers before redoing the experiment.

Linear Range of PCR Amplification of 3C Template Cannot be Determined With Titration PCR

This problem is most likely the result of high salt concentrations in the 3C library preparation. The use of Amicon columns typically removes most salt. However, if this problem is observed, the 3C template can be re-purified with phenol/chloroform extraction, ethanol precipitated and washed again on Amicon columns.

Titration PCR on 3C Template is Very Inefficient for All Pairs of Primers While Primers Work Fine on the Control Template

For large genomes, e.g. from mammals, PCR amplification of ligation produces using the control library is much more efficient than for the experimental 3C library. To optimize PCR amplification of ligation products with

the 3C library it may be important to further optimize the PCR conditions, including the time and temperature of annealing and the concentration of magnesium ions in the PCR buffer. Several PCR primer pairs for human and mouse, which have been used in our lab and reproducibly given good titration curves, are listed in Table 2.1

Poor PCR amplification of ligation products with the 3C library can also be the result of inefficient digestion and/or ligation during the 3C procedure. Restriction efficiency can be estimated by taking an aliquot of chromatin right after digestion in the 3C protocol. DNA is then purified and analyzed by PCR with primers designed to amplify a genomic region containing a restriction site. An equal amount of DNA purified from an undigested chromatin sample should be used as a control. Digestion efficiency in the 3C protocol is then defined as the ratio of the amount of PCR product obtained with the 3C DNA divided by the amount obtained with genomic DNA. 3C digestion is considered successful when a more than 70% reduction of PCR product is observed for the digested 3C template compared to the undigested genomic DNA template.

PCR Amplification of Ligation Products Leads to Multiple Bands on the Gel

There are might be several reasons for multiple bands in a 3C PCR (Figure 2.7B-E). We recommend approaching this problem step by step; finding and eliminating each possible reason (Figure 2.7B-E). First of all, multiple amplified DNA fragments can be the result of the incomplete digestion of the

cross-linked chromatin, as is typical in 3C experiments (Figure 2.7C). This problem might be especially severe when frequently cutting enzymes are used, as the average size of restriction fragments (i.e. potential insert in the ligation product) is small and hence non-canonical ligation products can be amplified efficiently. If the extra bands are very prominent, one may consider re-doing 3C template with increased concentration of restriction enzyme and extended incubation time. Multiple bands might also be the result of non-specific annealing of PCR primers. In that case, we recommend first to modify PCR condition (increase annealing temperature, Figure 2.7D) and if that does not help to redesign the primers (Figure 2.7E).

Contamination in the 5C Template

The issue of contamination occurs frequently in the lab. To prevent this, we strongly encourage aliquoting each reagent before beginning the experiment. If possible, aliquot into small sizes so that after one use the tube can be thrown away. It is also important to use filter tips and clean pipettes while creating the 5C library. Cleaning working surfaces and pipettes with a DNase solution before the creation of each library is helpful. If contamination is discovered after the universal PCR (Figure 2.8), the 5C library must be re-made using clean reagents. As an alternative, one can attempt to lower the number of cycles during the universal PCR step, though is not a recommended solution.

Low PE Adapter Ligation Efficiency

If a low amount of ligation product is obtained, it is possible to increase the starting concentration of the library. Additionally, ensure that a fresh aliquot of dATP is used in the addition of the “A” base step, as multiple freeze-thaw cycles negatively impact its effectiveness.

CHAPTER III

IDENTIFYING FUNCTIONAL LONG-RANGE INTERACTIONS OF THE CYSTIC FIBROSIS GENE

This chapter is adapted from Nele Gheldof, Emily M. Smith, Tomoko M. Tabuchi, Christoph M. Koch, Ian Dunham, John A. Stamatoyannopoulos, and Job Dekker. “Cell-type-specific Long-range Looping Interactions Identify Distant Regulatory Elements of the CFTR Gene.” Nucleic Acids Research 38, no. 13 (July 2010): 4325–4336. doi:10.1093/nar/gkq175.

Contributions

This paper describes the identification of four genomic regions that contact the *CFTR* promoter in three-dimensional space, and identifies two of those regions as potential enhancer elements. N.G., E.M.S., and T.M.T. performed the experiments. C.M.K., I.D., and J.A.S. provided DHS and methylation data. J.D. and E.M.S. prepared the manuscript.

Abstract

The identification of regulatory elements and subsequent identification of their target genes is complicated by the fact that regulatory elements often act over large genomic distances. Identification of long-range acting elements is particularly important in the case of disease genes as mutations in these elements can result in human disease. It is becoming increasingly clear that long-range control of gene expression is facilitated by chromatin looping interactions. These interactions can be captured and analyzed by chromosome conformation capture (3C). Here we employed 3C as a discovery tool for *ab initio* identification of long-range regulatory elements that control the cystic fibrosis transmembrane conductance regulator gene, *CFTR*. We identified four elements in a 460 kb region around the *CFTR* locus that loop specifically to the *CFTR* promoter exclusively in *CFTR* expressing cells. The elements are located 21 and 80 Kb upstream and 108 and 202 kb downstream of the *CFTR* promoter. These elements contain DNaseI hypersensitive sites and histone modification patterns characteristic of enhancers. We show that two of these elements are enhancers using a luciferase assay. Additionally we investigate what proteins may be binding these elements using a yeast one-hybrid approach. One promising factor discovered is TCF4, binding to the enhancer element 108 kb downstream of the promoter.

Introduction

Appropriate spatial and temporal control of gene expression depends on regulatory input from *cis*-acting elements such as promoters, enhancers and repressors. Identification of these elements is challenging because they can be located far from their target gene, sometimes up to several Mb (Dekker, 2008a; Kleinjan and van Heyningen, 2005; West and Fraser, 2005). Detailed knowledge of such distant regulatory elements and their mechanism of action will greatly contribute to basic understanding of gene expression.

Abundant evidence suggests that human disease can be caused by mutations that affect distant regulatory elements, while leaving the disease gene itself intact. Examples include Aniridia, caused by loss of distant regulatory elements of the *PAX6* gene, and blepharophimosis syndrome (BPES) that can be caused by deletion of regulatory elements located > 600 kb from *FOXL2* gene (D'haene et al., 2009; Kleinjan et al., 2001).

Regulatory elements are often characterized by the presence of DNase I hypersensitive sites (DHS), which can mark the position where transcription factors are bound to DNA. Other chromatin features found at distant regulatory elements are increased levels of H3K4me1 and histone acetylation (Heintzman et al., 2007, 2009). In addition, these sequences are often conserved across species (Maston et al., 2006; Visel et al., 2007). All these features can be used to identify putative functional elements and these powerful strategies are currently

widely applied (e.g. (Birney et al., 2007; Visel et al., 2007)). However, these analyses do not immediately reveal the target genes of these regulatory elements.

Regulatory elements can directly associate with target promoters through chromatin looping (Dekker, 2008; Kleinjan and van Heyningen, 2005; de Laat and Grosveld, 2003; West and Fraser, 2005). These looping interactions can be detected using chromosome conformation capture, or 3C (Dekker et al., 2002). The insight that regulatory elements physically associate with promoters provides a methodology to discover novel regulatory elements by performing systematic 3C analyses to search for genomic elements that are found to interact with a specific promoter. Here we tested the feasibility of such an approach by analysis of the cystic fibrosis transmembrane conductance regulator (*CFTR*) locus. Identification of extragenic regulatory elements for this locus (and other disease related loci) is especially important because 1) they could be screened for mutations in patients with no known mutations in the *CFTR* gene itself, and thus aid in proper diagnosis, and 2) they could be included in gene therapy constructs (to recapitulate endogenous *CFTR* regulation). Finally, identification of *CFTR* regulatory elements will provide basic insights into the mechanisms that control expression of *CFTR*, which could also lead to new approaches to boost or manipulate *CFTR* expression in patients.

The *CFTR* gene, when both alleles are mutated, causes cystic fibrosis. It contains a promoter that has many characteristics of a housekeeping gene

including potential binding sites for SP1 (Yoshimura et al., 1991). Also present is a critical CCAAT-like element, shown to bind C/EBP (Pittman et al., 1995), implicating cAMP as a possible regulator. Supporting a role for cAMP in *CFTR* regulation is data showing that cAMP activation of protein kinase A can regulate basal *CFTR* expression (McDonald et al., 1995) and the discovery that CREB and ATF-1 bind the *CFTR* promoter in a cAMP-responsive manner (Matthews and McKnight, 1996). A YY1 element has also been identified that, when mutated, significantly increases the expression of *CFTR* (Romey et al., 2000). However, despite the abundance of regulatory elements in the *CFTR* promoter, it is clear that additional elements are required for the complex spatial- temporal expression pattern of the *CFTR* gene (McCarthy and Harris, 2005). Indeed, work from the Harris laboratory has identified additional putative regulatory elements located within introns of the gene, as well up-and downstream of the locus. For instance, HNF1α has been found to interact with a putative regulatory element in introns 1, 10, 17a and 20 and over-expression of this protein results in increased *CFTR* mRNA levels (Mouchel et al., 2004; Ott et al., 2009a).

Here we applied a systematic 3C analysis to a 460 kb chromosomal region surrounding the *CFTR* transcription start site (TSS). We examined different cell lines that either express or do not express the gene in order to identify regulatory elements that function specifically in *CFTR*-expressing cells. Our approach was validated by identification of previously discovered regulatory elements, e.g. an element located 202 kb downstream of the promoter that

coincides with a DNAseI hypersensitive site (referred to as DHS4574+15.3 (Nuthall et al., 1999a)) and that has been found to be able to act as an insulator (Blackledge et al., 2009). Importantly, we discovered an additional regulatory element: one located within intron 11 (108 kb downstream of the TSS) that interacts with the *CFTR* TSS exclusively in cells that express the gene. Additional 3C analyses allowed us to define the locations of long-range acting elements at ~1 kb resolution. In *CFTR* expressing cells, these elements contain the characteristic features of known regulatory elements, such as the presence of a DHS and specific histone modification patterns. Interestingly, we find that regulatory elements also interact with each other.

We suspect that architectural proteins as well as transcription factors (TFs) are responsible for creating and maintaining looping interactions. It has been shown that the architectural proteins CTCF, cohesion and Mediator complexes regulate looping interactions, and that knocking down these proteins can change loops (Phillips-Cremins et al., 2013; Sanyal et al., 2012; Splinter et al., 2006). In the well-studied β-globin locus, the TFs ELKF, GATA1, FOG1 and TAL1 are critical for proper looping formation of the LCR with target genes (Drissen et al., 2004; Vakoc et al., 2005; Yun et al., 2014). In order to identify if the looping elements are influencing *CFTR* expression, we used a luciferase assay to measure the effect each element has on gene expression. Additionally, we placed the active elements in a yeast one-hybrid assay (Deplancke et al.,

2004; Reece-Hoyes et al., 2011) to discover which human transcription factors may bind to these DNA fragments .

These studies identify novel *CFTR* regulatory elements and provide insights into combinatorial control of the gene. We show that both upstream elements (+108 and +202) can act as enhancers in a luciferase assay and identified a minimal enhancer region contained within these looping regions. We identify high-confidence transcription factors that may play a role in forming or maintaining the 3D contacts between the upstream elements and the *CFTR* promoter. These results provide strong evidence that 3C-based approaches provide tools for *ab initio* discovery of regulatory elements and their target genes.

Materials and Methods

Cell Culture

We used six cell lines for the 3C analyses: three *CFTR*-expressing cell lines Caco2, HT29 and HeLa S3, compared to three non-*CFTR* expressing cell lines GM06990, K562 and HepG2. The GM06990 cell line was obtained from Coriell Cell Repositories, and the Caco2, HT29, HeLa S3, K562 and HepG2 cell lines from the American Type Culture Collection (ATCC). All cell lines were grown at 37°C in 5% CO_2 in medium containing 1% penicillin-streptomycin. Caco2 cells were grown in MEM alpha medium supplemented with 20% fetal bovine serum (FBS), HT29 cells in DMEM medium with 10% FBS, HeLa S3 in F12K medium with 10% FBS, GM06990 and K562 cells in RPMI medium with 10% FBS and HepG2 in MEM alpha medium with 10% FBS. Suspension cells were harvested at log-phase and monolayer cells were grown to 95% confluency before harvesting for RT-PCR and 3C analysis.

RT-PCR

Total RNA from the six cell lines was isolated using the RNeasy Mini Kit (Qiagen). *CFTR* RNA transcript levels were analyzed using *Power* SYBR® Green RNA-to-C_T *1-Step* Kit (Applied Biosystems) on a *StepOnePlus*™ Real-Time PCR System (Applied Biosystems). *HPRT1* was used to normalize the data. Primer sequences are available in Appendix I.

3C Analysis

The 3C analysis was performed using EcoRI and BsrGI as described previously (Dostie and Dekker, 2007; Gheldof et al., 2006). We generated a control library for each enzyme using BAC clones obtained from Invitrogen and the Children's Hospital Oakland Research Institute (CHORI). We used three minimally overlapping BAC clones spanning the investigated 450 kb *CFTR* locus: RP11-35E12, RP11-450L14 and CTD-2034E23, as well as the BAC clone RP11-197K24 from a gene desert region (ENCODE region Enr313) as described in Dostie et al. (Dostie and Dekker, 2007). We normalized the data of each experiment using the interaction frequencies measured in this gene desert region to allow direct comparison of data obtained from the cell line in which *CFTR* is expressed versus the cell line in which *CFTR* is not expressed. Broad 3C analysis of Caco2 versus GM06990 was performed in 4 independent experiments, HT29 versus K562 in 2, and HeLa S3 versus HepG2 in 1, and each experiment was quantified at least in triplicate. 3C fine mapping of elements III and IV was performed with BsrGI in the cell lines Caco2 and GM06990, and 3 independent experiments were performed. Primer sequences and 3C data are available in Appendix I.

DNAseI Mapping and Analysis of Histone Modification Patterns in the ENCODE Regions

DNAseI hypersensitive site data and histone modification data were generated as previously described (Dorschner et al., 2004; Koch et al., 2007). Data are available at http://genome.ucsc.edu/ENCODE/pilot.html, where detailed descriptions of the methods can also be found.

Generation of Reporter Constructs

The pGL3 basic vector (Promega) was transformed into a Gateway-compatible destination vector by inserting the R4R2 cassette into the MluI restriction site according to the MultiSite Gateway Cloning manual (Invitrogen). Insertion of the cassette in the correct orientation to drive transcription of the luciferase gene was verified by sequencing. The R4R2 cassette was a generous gift by the Walhout lab. A 1.7-kb fragment spanning the CFTR basal promoter (from – 1,691 to – 35 bp from the ATG translation start site) was amplified with primers with B1B2 tails and cloned into the destination vector.
Potential enhancer elements were amplified with primers containing B4B1R tails and each of these elements was cloned upstream of the promoter fragment. As the 4 EcoRI DNA looping elements as identified by 3C were on average 4 kb in size, we split each element in smaller overlapping segments of 1.1-1.8 kb to allow positioning of the regulatory element more precisely. Elements I and IV were split into 3 fragments and Element III was split into 2. Additionally, elements III and IV were made without their DHS, and fragments were produced that contained only the DHS. Genomic coordinates (hg18, chromosome 7) of each

fragment are listed in Table 3.1. As negative control, the pGL3-basic vector-promoter alone construct was made by cloning a short 51-bp fragment with B4B1R tails located upstream of the CFTR promoter insert (from –1,825 to –1,774 bp from the ATG translation start site; coordinates: 116905560-116905611). Inclusion of this 51 bp fragment in the promoter-only construct, but not in the element-promoter constructs did not affect the interpretation of the enhancing effects of the elements, as shown by analysis of a pair of promoter and promoter-element IV constructs that both lack this sequence yet gave identical levels of promoter activation by element IV (data not shown). As a positive control, we generated a construct containing a fragment of intron 1 of the CFTR gene. This fragment was previously shown to have significant enhancer activity (Smith et al., 1996). The intron 1 fragment was amplified using primers IA1R and TSR8 as described in Smith et al. (1996) and containing the B4B1R tails. We also generated a construct containing the segments with the highest enhancer activity from both element III and IV as a fusion. To this end, we performed a fusion-PCR (Heckman and Pease, 2007) of element IIIb and IVc, and cloned this fused fragment upstream of the CFTR promoter insert. Fragments IIIb and IVc were further characterized by analyzing the location of a DHS within each of these regions. Fragment IIIb-DHS was created by truncating the IIIb fragment so it would not contain the DHS. IIIbDHS is the portion of IIIb that contains the DHS. IIIbDHS and IIIb-DHS constructs have a 50bp overlap.

Element ID	Beginning	End
Promoter	116905694	116907350
Ia	116822691	116824500
Ib	116824484	116825593
Ic	116825438	116827097
II	116883223	116885151
Intron1	116916750	116917500
IIIa	117013018	117014528
IIIb	117014436	117016269
III-DHS	117014436	117015582
IIIDHS	117015552	117016269
IVa	117105596	117107448
IVb	117107337	117109457
IVc	117109425	117110818
IVd	117109425	117111668
IV-DHSa	117109425	117110125
IV-DHSb	117110685	117111668
IVDHS	117110095	117110715

Table 3.1: Luciferase Fragment Locations. Table with the genomic coordinates (Chr7, Hg18) of the elements cloned into the luciferase assay. IVd is an expanded version of element IVc. IVd was used in the second assay in place of IVc, but the addition of this extra DNA did not affect the luciferase results. IV-DHSa and IV-DHSb were combined to create IVd-DHS

IVd-DHS was created by combining the regions of IVd that lack the DHS, while IVdDHS is the portion of IVd that contains the DHS.

Transfection and Luciferase Assays

The first round of transient transfections were performed on two monolayer cell lines, the CFTR-expressing Caco2 and non-CFTR expressing HepG2 cell line grown in 96-well plates to 70-80% confluency. In all transfection experiments the pGL3 constructs were cotransfected with 1/10 the amount of DNA of pRL-TK as a transfection control, using Effectene as transfection reagent, as indicated in the manufacturer´s manual (Qiagen). Luciferase assays were carried out using the dual luciferase kit (Promega) on a 96-well plate reader. Each transfection experiment was carried out at least five times with individual constructs being assayed in duplicate in each experiment. For the second round of testing, Attractene was used as a transfection reagent according to the manufacturer's protocol (Qiagen 2008 cat#301005) modified for 96-well plates. 300 ng of each DNA construct was added to Caco2 cells and 250 ng of each DNA construct to HepG2 cells in serum-free and antibiotic-free media. 0.75 µl Attractene was used for Caco2 cells and 0.375 µl Attractene for HepG2 cells. Each transfection reaction was performed in quadruplicate with at least three biological replicates tested for each construct.

48 hours after transfection, the cells were assayed using the Dual-Glo Luciferase Assay System (Promega 2009 Cat#E2920). Measurements were

made on a Perkin Elmer Victor3 plate reader using Wallac 1420 Manager software. Results are expressed as relative luciferase activity, with the pGL3-promoter construct activity equal to 1. One-tailed t-tests were performed in Excel to test significance of the increased luciferase activity.

Yeast One-hybrid

The Y1H protocol was adapted from Deplancke et al., 2004, 2006; and Reece-Hoyes et al., 2011. Gateway cloning was used to create constructs for testing. Elements from the luciferase assay were cloned into two destination vectors, pMW2 (containing the HIS gene) and pMW3 (containing *LacZ*) (Deplancke et al., 2004). Clones were digested with XhoI and NcoI, respectively, to check for the correct insert size. Clones were transformed into the yeast strain YM4271 (a gift from the Walhout Lab) and grown at 30˚C for five days. Colonies were re-streaked onto a fresh Sc-UH plate and grown at 30˚C overnight, and then replica plated onto plates containing 10, 20 and 40 mM 3AT-UH and one YEPD plate with a nitrocellulose filter. After growth at 30˚C for two days, the YEPD plate with a nitrocellulose filter was used to check *LacZ* self-activation of each bait strain. This was done by a β-galactose overlay assay. The 3AT-UH plates were used to check self-activation of the *HIS3* gene by the inserted bait DNA. These plates were checked for growth over a period of 7 days. Selected clones tested as slightly positive in the *LacZ* self-activation test and grew on 10

mM but not on 40 mM 3AT-UH plates. Yeast colony PCR was used to check for the correct insertion in yeast using the following primers:

pMW2: M13FW: 5′ - GTAAAAGCACGGCCAGT - 3′

HIS293RV: 5′ - GGGACCACCCTTTAAAGAGA - 3′

pMW3: 1HIFW: 5′ - GTTCGGAGATTACCGAATCAA - 3′

LacZ592RV: 5′ - ATGCGCTCAGGTCAAATTCAGA - 3′

The yeast one-hybrid experiment was performed using a Rotor HAD robot (Singer). Yeast carrying the bait elements were mated with the human transcription factor array (Reece-Hoyes et al., 2011) on YEPD plates. After one night of growth at 30˚C, the colonies were transferred to –UHT plates. After two days of growth at 30˚C, the colonies were transferred to plates containing 5 mM 3AT and 160 mg/L X-Gal. The colonies were assayed for growth and color by eye every day or every other day up to 7 days post-transfer. All Y1H results can be found in Appendix I.

Results

Identification of long-range interactions between the *CFTR* promoter and distant elements

Long-range looping interactions can be detected using chromosome conformation capture (3C) (Dekker et al., 2002). 3C is a widely used method and the procedure has previously been described in detail and in chapter 2 of this thesis (Dekker et al., 2002; Simonis et al., 2007; Splinter et al., 2003). Briefly, 3C employs formaldehyde cross-linking to capture physically interacting chromatin segments. Cross-linked chromatin is then solubilized, digested and intra-molecularly ligated so that pairs of interacting genomic elements are converted into unique ligation products. Ligation products are then detected and quantified by semi-quantitative PCR (see Materials and Methods and chapter 2 of this thesis).

Here we used 3C as a discovery tool for *ab initio* identification of distant regulatory elements that interact specifically with the active *CFTR* promoter. We performed 3C with Caco2 cells, which express high levels of *CFTR*, and with GM06990 lymphoblastoid cells, which express very low levels of *CFTR* (Figure 3.1), to identify genomic elements that physically associate with the *CFTR* promoter. We used a PCR primer located in the EcoRI fragment that contains the *CFTR* transcription start site (TSS) and paired it with primers in restriction fragments throughout a 460 kb region surrounding the promoter (primer

sequences available in Appendix I). This experimental setup will detect ligation (and thus long-range interaction) of the promoter fragment with any of the other restriction fragments. The results are shown in Figure 3.2A (3C data available in Appendix I).

In GM06990 cells, we observe that the promoter fragment interacts most strongly with nearby restriction fragments and that interaction frequencies decrease precipitously for fragments located farther away. This inverse relationship between interaction frequency and genomic distance is expected for a flexible chromatin fiber in the absence of any specific long-range looping interactions (Dekker, 2006, 2008b; Dekker et al., 2002; Gheldof et al., 2006; Rippe, 2001). Specific long-range looping interactions would result in peaks of interaction frequency super-imposed upon this background of interactions. We conclude that in GM06990 cells, the promoter does not engage in specific interactions with elements located in the surrounding 460 kb (Figure 3.2A).

In Caco2 cells, we also observe frequent interactions between the promoter fragment and nearby restriction fragments. In addition, we observe frequent interactions with several other restriction fragments, as evidenced by local peaks in interaction frequency above the background of non-specific associations. Two interacting restriction fragments are located 20 kb and 80 kb upstream of the TSS; one is located within the *CFTR* gene 108 kb downstream of the promoter in intron 11 and a fourth fragment is located 202 kb downstream of the TSS (15 kb downstream of the 3’ end of the gene).

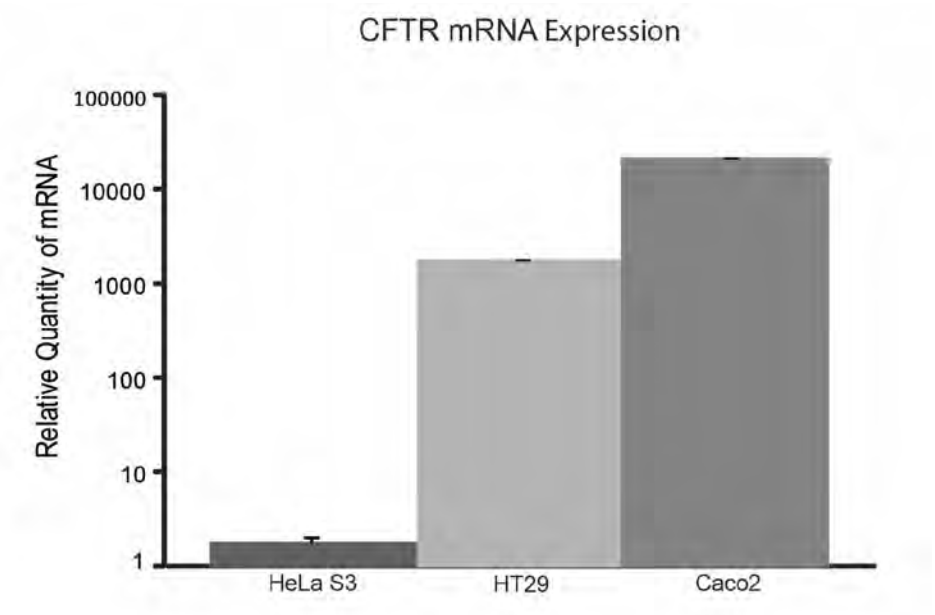


Figure 3.1: Expression Level of *CFTR* as Detected by RT-PCR. All six cell lines were analyzed at the *CFTR* exon2-exon3 junction and normalized using *HPRT1* (primer sequences are given Appendix I). HeLa S3, HT29 and Caco2 cell lines express *CFTR*. K562 and HepG2 cells did not produce detectable levels of *CFTR* product. GM06990 cells expressed a low level of *CFTR* mRNA (here set as 1). HeLa S3 expressed *CFTR* 2X higher than GM06990, while both HT29 and Caco2 show high levels of *CFTR* mRNA expression (1822 and 22606 fold higher than the level in GM06990, respectively).

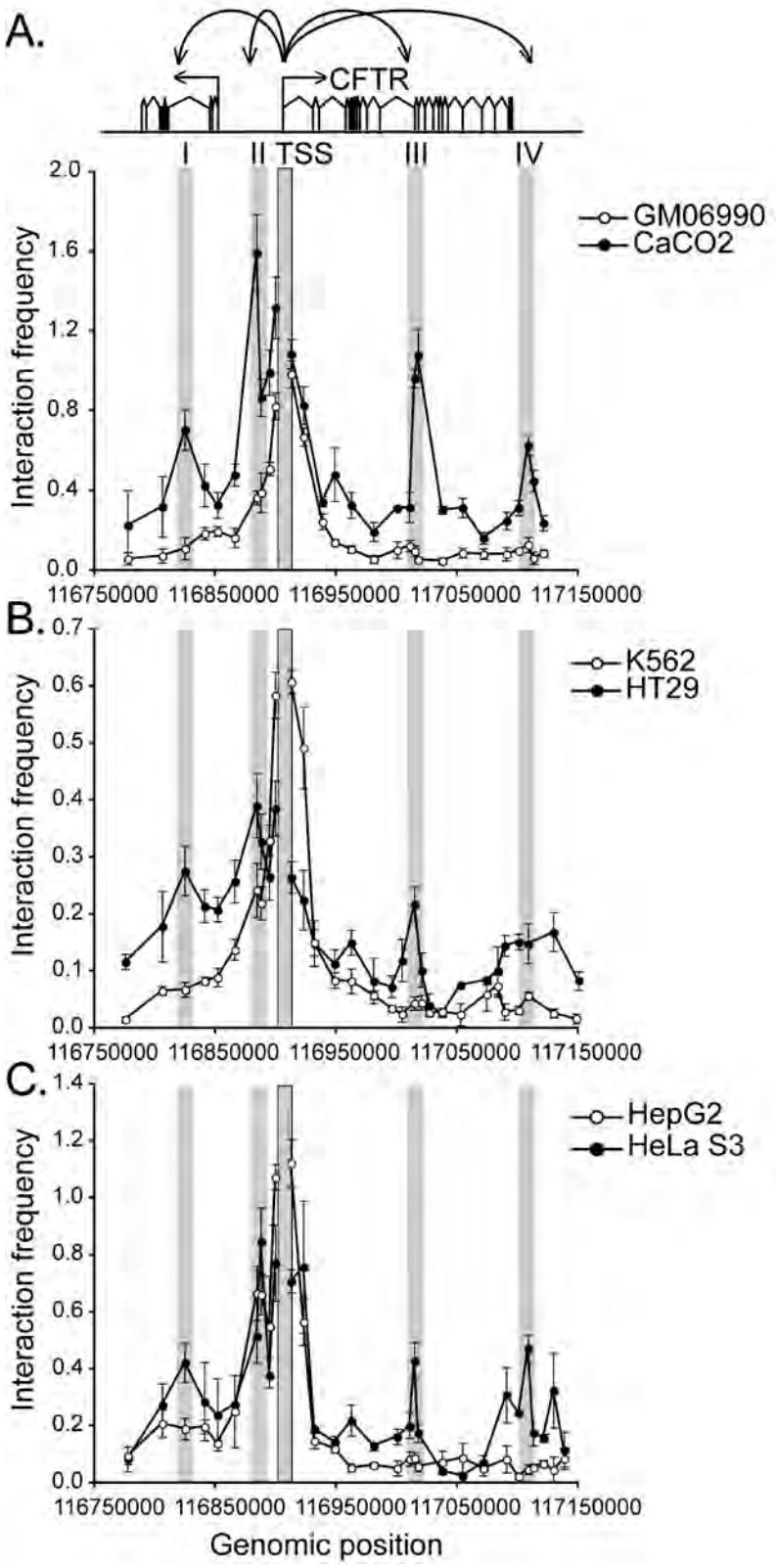

A.
CFTR
I
II
TSS
III
IV
Interaction frequency
GM06990
CaCO2
B.
K562
HT29
C.
HepG2
HeLa S3
116750000
116850000
116950000
117050000
117150000
Genomic position

Figure 3.2: Chromosome Conformation Capture (3C) Reveals Four Elements that Interact with the *CFTR* Transcription Start Site. All experiments used EcoRI to digest the chromatin. Interaction frequencies between the *CFTR* transcription start site and a 460 kb surrounding region are determined by 3C. Primer sequences and 3C data are found in Appendix I. **A**) 3C profile of Caco2 cells (black circles), which express the *CFTR* gene, and GM06990 cells (open circles), which do not express the gene. **B**) 3C profile of HT29 cells (black circles), which express the *CFTR* gene, and K562 cells (open circles) which do not express the gene. **C**) 3C profile of HeLa S3 cells (black circles) which express the *CFTR* gene, and HepG2 cells (white circles), which do not express the gene.

To confirm these results we analyzed two additional pairs of cell lines, each composed of one cell line that expresses *CFTR* and one that does not. First, we chose HT29 cells, which express high levels of *CFTR* (Sood et al., 1992), and K562 cells, which do not express *CFTR* (as determined by RT-PCR, Figure 3.1). We again performed 3C to analyze long-range interactions with the *CFTR* promoter fragment (Figure 3.2B). We observe that in both cell lines the promoter fragment frequently interacts with neighboring fragments, as expected, but these interactions are less frequent in HT29 cells than in K562 cells. We have previously observed that interaction frequencies around active promoters can be reduced, possibly as the result of local chromatin decondensation (Gheldof et al., 2006). Importantly, we found no evidence for specific long-range looping interactions in K562 cells, but we again observe local peaks in interaction frequency for multiple restriction fragments in *CFTR*-expressing HT29 cells. These interacting restriction fragments are the same as those observed in Caco2 cells (Figure 3.2A). In addition, we observe a region encompassing several EcoRI restriction fragments around the 3' end of the *CFTR* gene that all display increased interaction frequencies as compared to K562 cells, suggesting the presence of multiple interaction elements in that region.

Second, we analyzed HeLa S3 cells that express *CFTR* at a low level (as determined by RT-PCR, Figure 3.1) and HepG2 cells that do not express *CFTR*. In HeLa S3 cells we again detect specific long-range interactions between the *CFTR* promoter fragment and a set of distant fragments, whereas no specific

long-range interactions were detected in HepG2 cells (Figure 3.2C). All interacting fragments correspond to the same set of fragments detected in Caco2 and HT29 cells, except one: in HeLa S3 cells the promoter fragment does not frequently interact with the fragment located at -21 kb. Instead the promoter fragment associates more frequently with the restriction fragment immediately adjacent to this fragment, located at -17 kb. In HepG2 this -17 kb fragment also interacts with the promoter fragment somewhat more frequently than neighboring fragments. The two adjacent fragments located at -17 and -21 kb may contain two separate elements, or a single interacting region that displays cell type specific differences in the precise point of contact within the promoter. Finally, as in HT29 cells, we observe that in HeLa S3 cells the promoter fragment frequently interacts with several restriction fragments near and beyond the 3' end of the gene.

Multiple long-range interactions are observed between the *CFTR* promoter fragment and several restriction fragments upstream and downstream of the gene as well as within an intron of the gene. Moreover, these interactions are observed only in *CFTR* expressing cells (Caco2, HT29 and HeLa S3) and not in non-expressing cells (K562 and HepG2). Interestingly, GM06990 cells do not display these long-range interactions, yet express *CFTR* at a very low, but detectable, level. Thus, the correlation between *CFTR* expression and looping between the promoter and these distal elements is not absolute (see discussion). Analysis of digestion efficiency confirmed that these frequent interactions are not

due to more efficient digestion of the corresponding restriction sites, consistent with many other studies that showed that digestion efficiency is not a major contributor to variation in 3C interaction frequencies (not shown, (Dekker, 2007; Gheldof et al., 2006; Miele et al., 2009; Tolhuis et al., 2002)).

For further analyses, we focused on loci that are consistently found to interact with the *CFTR* promoter in all three *CFTR* expressing cell lines. These looping elements are located at -80, -21/-17, +108 and +202 kb from the TSS (highlighted by grey bars in Figure 3.2). From here on we refer to these looping elements with Roman numerals (I through IV, with element II containing the two adjacent restriction fragments located 21 and 17 kb upstream of the TSS). It is important to point out that other interacting elements may be present at other locations in and around the *CFTR* locus, as suggested by additional cell-line specific peaks, e.g. around the 3' end of the gene in HT29 cells. Consistently, one of these restriction fragments contains an element (located 6.8 kb downstream of the gene) that has recently been shown to bind CTCF and to also interact with the *CFTR* promoter in primary epididymis cells (Blackledge et al., 2009).

Two of the four identified elements (elements I and II) have been previously identified. Element I may play a structural role but does not appear to be critical for *CFTR* regulation as constructs lacking this element faithfully reproduce the appropriate expression pattern of *CFTR* in mouse (Nuthall et al., 1999a). Element II has previously been shown to play a role in *CFTR* regulation

as deletion of this element reduces *CFTR* expression (Nuthall et al., 1999a). We note that the ENCODE consortium recently identified sites that are bound by the CTCF protein in the restriction fragment directly adjacent to element II in Caco2 cells (John Stamatoyannopoulos, unpublished). The Harris lab also demonstrated CTCF binding to this element in Caco2 and Calu3 cells (Blackledge et al., 2007). This is interesting because CTCF has been directly implicated in mediating long-range looping interactions (Kurukuti et al., 2006; Phillips and Corces, 2009; Splinter et al., 2006). Indeed, knocking down CTCF in Caco2 cells reduces looping of element II to the *CFTR* promoter and increases *CFTR* gene expression (Gosalia et al., 2014). In addition, DNaseI hypersensitive sites, indicating the presence of putative regulatory elements, have been identified in or directly adjacent to both elements I and II (Smith et al., 1995). These data validate our 3C-based strategy for regulatory element discovery.

Long-range Interactions Between Looping Elements

Long-range regulatory elements not only interact with promoters, but in some cases have also been found to interact with other regulatory elements, e.g. in the case of the beta-globin locus (Dostie and Dekker, 2007; Splinter et al., 2006; Tolhuis et al., 2002). To determine whether any long-range interactions occur between the looping elements we identified above, we again employed 3C analysis in Caco2 and HeLa S3 cells.

First we performed 3C analyses anchored on elements I, II (anchored on the restriction fragment located 21 kb upstream of the TSS), III and IV in Caco2 and GM06990 cells (Figure 3.3). As expected, we find that these elements interact frequently with nearby restriction fragments in both cell lines. In addition, these elements interact frequently with the *CFTR* promoter fragment in Caco2 cells, but not in GM06990 cells. Elements I and II do appear to interact with elements III and IV, as minor peaks at the corresponding locations are visible (Figure 3.3A and 3.3C), although these interactions are considerably less frequent than their interaction with the promoter fragment. In Caco2 cells elements III and IV are found to interact prominently with each other, as strong peaks of interaction frequencies are readily detected between the two restriction fragments (Figure 3.3E and 3.3G). In GM06990 cells we did not detect any obvious peaks in the 3C interaction profiles, suggesting that there are no specific looping interactions between these elements in that cell line. We note that interactions obtained with all 4 anchor elements are generally higher in Caco2 cells as compared to GM06990 cells throughout the region, which is as predicted for a locus that is more compact as a result of long-range looping interactions (Rippe, 2001). We conclude that in Caco2 cells elements III and IV not only interact with the *CFTR* promoter fragment but also associate with each other, whereas elements I and II interact mainly with the promoter and less frequently with the other looping elements.

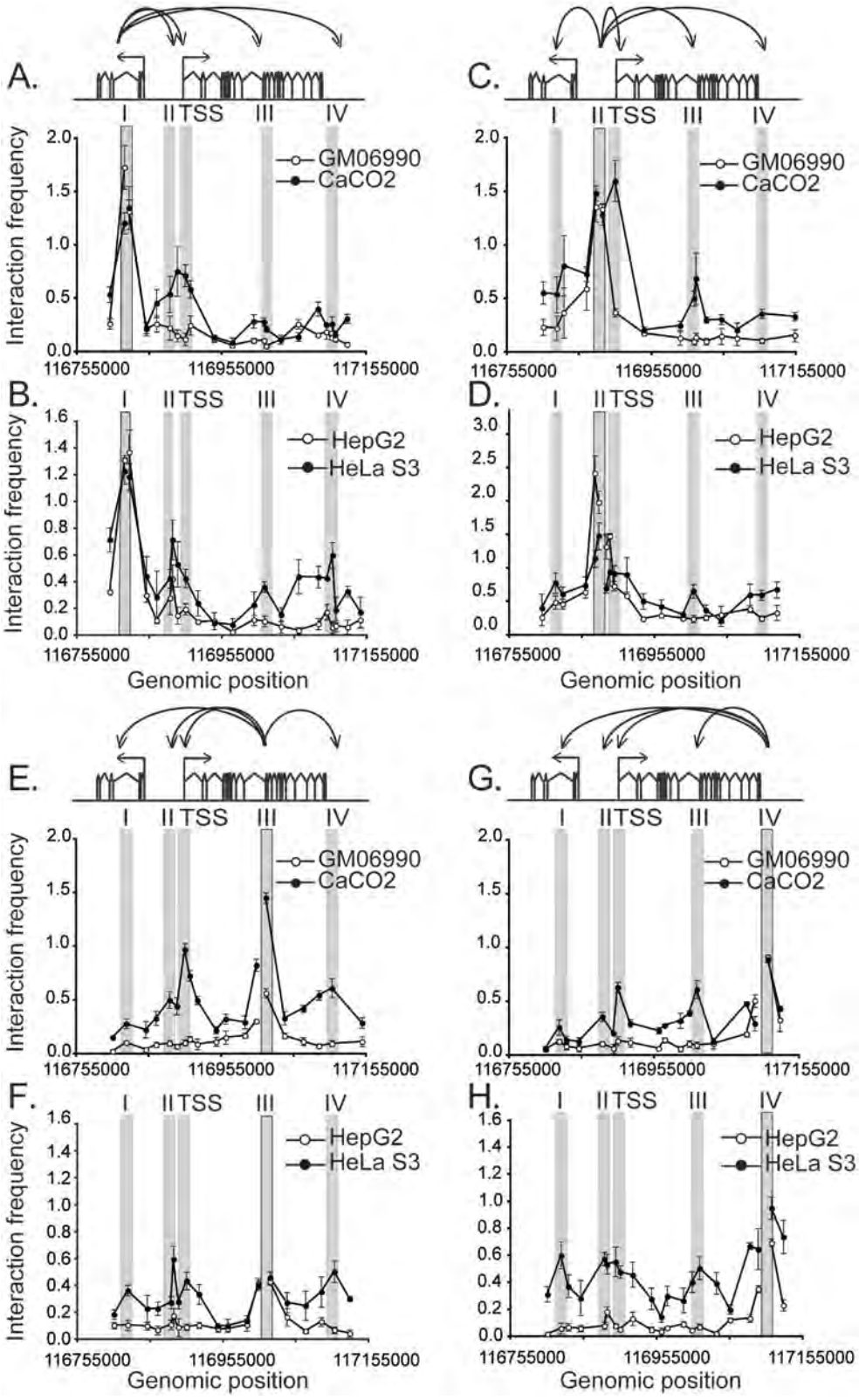

A.
B.
C.
D.
E.
F.
G.
H.
I
II
TSS
III
IV
GM06990
CaCO2
HepG2
HeLa S3
Interaction frequency
Genomic position
116755000
116955000
117155000

Figure 3.3: 3C Interaction Profiles for Each of the Four Elements. All experiments used EcoRI to digest the chromatin. 3C data can be found in Appendix I. **A**) The anchor point is set at Element I in Caco2 and GM06990 cells, showing interactions between Element I and Element II, the transcription start site (TSS), and Element III and IV in Caco2 cells, with no notable interactions in GM06990 cells. **B**) The anchor point is set at Element I in HeLa S3 and HepG2 cells, showing interactions between Element I and Element II, III and IV in HeLa S3 cells, with no notable interactions in HepG2 cells. **C**) The anchor point is set at Element II in Caco2 and GM06990 cells, showing interactions between Element II and Element I, TSS, Element III and IV in Caco2 cells, with no notable interactions in GM cells. **D**) The anchor point is set at Element II in HeLa S3 and HepG2 cells, showing interactions between Element II and Element I, III and IV in HeLa S3 cells, with no notable interactions in HepG2 cells. **E**) The anchor point is set at Element III in Caco2 and GM cells, showing interactions between Element III and Element I, II, TSS and element IV in Caco2 cells, with no notable interactions in GM cells. **F**) The anchor point is set at Element III in HeLa S3 and HepG2 cells, showing interactions between Element III and Element I, II, TSS, and Element IV in HeLa S3 cells, with no notable interactions in HepG2 cells. **G**) The anchor point is set at Element IV in Caco2 and GM cells, showing interactions between Element IV and Element I, II, TSS, and Element III in Caco2 cells, with no notable interactions in GM cells. **H**) The anchor point is set at Element IV in HeLa S3 and HepG2 cells, showing interactions between Element IV and Element I, II, TSS, and Element III in HeLa S2 cells, with no notable interactions in HepG2 cells.

We repeated the 3C analysis in HeLa S3 and K562 cells with the exception that for these cells the anchor for element II was on the restriction fragment located at -17 kb that we found to interact more frequently with the promoter fragment than the -21 kb restriction fragment (Figure 3.2C). We again observe that in HeLa S3 cells, but not in HepG2 cells, elements III and IV interact prominently with each other, as compared to background levels at other positions in the locus (Figure 3.3F and 3.3H). Elements I and II also display local peaks of interaction at the locations of the other looping elements (Figure 3.3B and 3.3D). The interaction between element I and IV is particularly prominent in HeLa S3 cells, whereas this interaction is much less pronounced in Caco2 cells.

We conclude that in *CFTR* expressing cells, elements I-IV not only interact with the promoter fragment but also with each other. In addition, the interaction frequency of these elements is cell line specific. In HeLa S3 cells all elements interact prominently with each other. In Caco2 cells only elements III and IV interact prominently with each other, and they do so with a frequency that is comparable to that of interactions between these elements and the promoter fragment.

Fine Mapping of Looping Elements III and IV

We did not analyze elements I and II further, as they have been characterized before (Blackledge et al., 2007; Gosalia et al., 2014; Nuthall et al.,

1999a). Rather we focused on elements III and IV in Caco2 cells as they could represent novel regulatory elements that control the *CFTR* gene.

The above 3C analysis identified individual EcoRI restriction fragments that displayed the most prominent interactions with the promoter fragment. The size of these restriction fragments averages 4 kb, but the elements that mediate these interactions are most likely considerably smaller. In order to determine the position of the putative elements at higher resolution, we repeated the 3C analysis with another restriction enzyme, BsrGI. We chose this enzyme because it cuts the selected EcoRI fragments into smaller pieces. We then analyzed the interactions between the BsrGI fragment containing the TSS and several BsrGI fragments located around the positions of elements III and IV (Figure 3.4). We find a local peak of interaction between the promoter fragment and a 1066 bp fragment containing element III, located in the 3' end of the original 3252 bp EcoRI fragment (Figure 3.4A, 3.5A). We also see a peak between the promoter fragment and a 1560 bp fragment, located near the 3' end of the original 6069 bp EcoRI fragment containing element IV (Figure 3.4B, 3.5B). These interactions were not observed in GM06990 cells. These results confirm the presence of looping elements at these locations, and further define the positions of these elements. Based on the two independent 3C analyses performed with EcoRI and BsrGI we conclude that element III is contained within an 806 bp BsrGI-EcoRI fragment and that element IV is contained within the 1560 bp BsrGI restriction fragment (Figure 3.5).

Above we identified interactions between EcoRI fragments containing elements III and IV. We wanted to know if the interactions between elements III and IV are mediated by the same elements as the interactions between the promoter fragment and element III and IV, respectively, or whether these distinct interactions involved other elements contained within the EcoRI restriction fragments. Therefore, we tested whether the same BsrGI fragments that interact with the promoter fragment also interact with each other. 3C analysis confirmed that the BsrGI fragment containing element III interacts most prominently will BsrGI fragment that contains element IV (Figure 3.4C). These results suggest that the same elements that interact with the promoter also interact with each other.

Additional Evidence for the Presence of Functional Elements in the Looping Elements III and IV

The use of two enzymes in our 3C analysis allowed us to define the minimal interacting and functional regions containing elements III and IV. In Figure 3.5 we show, in modified UCSC genome browser shots, the EcoRI and BsrGI fragments that interacted most prominently with the promoter fragment. The minimal region of overlap between these fragments defines the positions of elements III and IV (indicated by vertical lines).

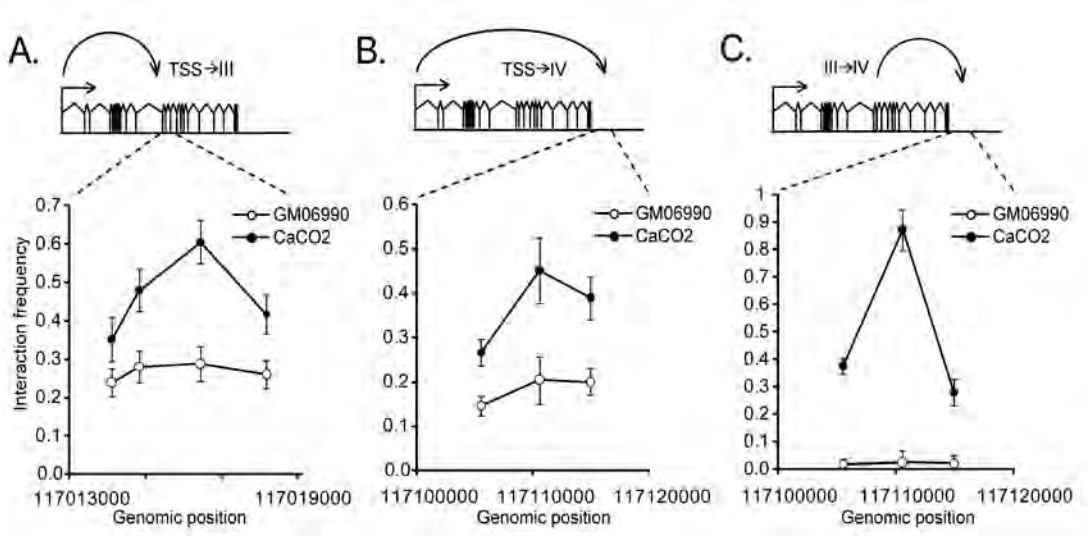


Figure 3.4: Fine Mapping of the Looping Interactions Between the *CFTR* Promoter and Elements III and IV. 3C primers and data can be found in Appendix I. **A-C**) 3C profiles using BsrGI to digest the chromatin focused on Element III (**A**), anchored at the TSS. The interaction peak corresponds to a 1066 bp fragment. **B**) 3C focused on Element IV, anchored at the TSS. The interaction peak corresponds to a 1560 bp fragment. **C**) 3C focused on Elements IV anchored at Element III. The interaction peak corresponds to a same 1560 bp fragment that interacts with the *CFTR* TSS.

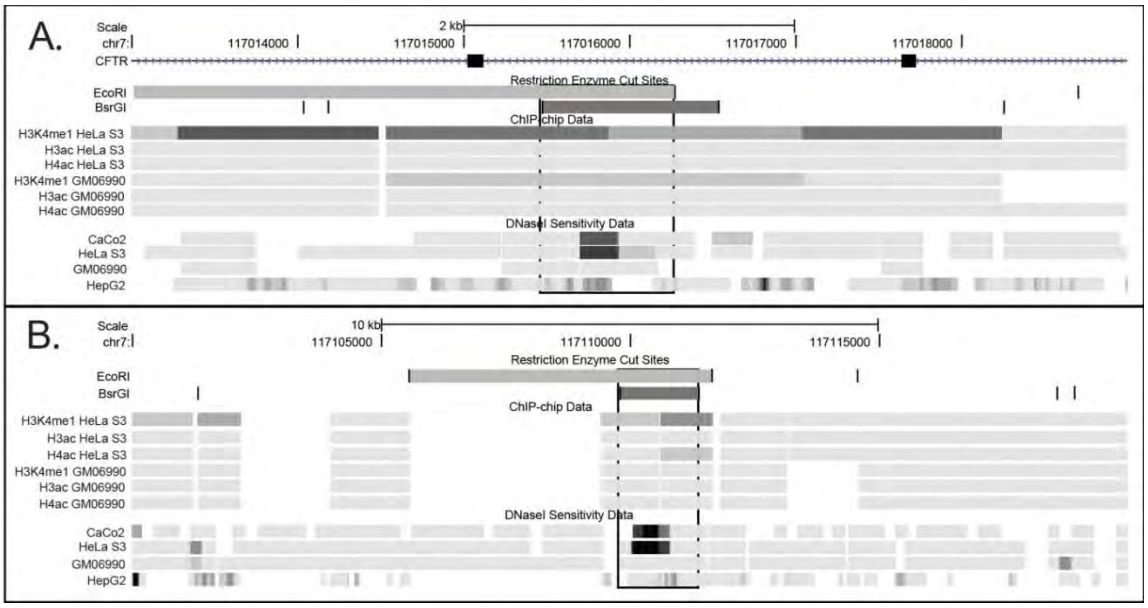


Figure 3.5: Data from the ENCODE Consortium in the UCSC Genome Browser hg18 [http://genome.ucsc.edu/ENCODE/]. Restriction fragments that show 3C interaction peaks are outlined with grey boxes. A boxed outline indicates the narrowed-down functional region. Histone methylation, acetylation and DHS data are shown, with "positive hits" shown as darker grey regions. **A**) Region 117013000-117019000 on chromosome 7, which contains element III. **B**) Region 117100000-117120000 on chromosome 7, which contains element IV.

To obtain further evidence that the regions we identified here contain functional elements, we determined whether they display chromatin features that indicate the presence of regulatory elements. One hallmark of functional elements is the presence of DNaseI hypersensitive sites, corresponding to nucleosome free regions bound by regulatory protein complexes. In addition, regulatory elements, e.g. enhancers, have been found to display characteristic histone modifications including increased levels of acetylation and high levels of histone H3 monomethylation at the fourth lysine residue (H3K4Me1). The *CFTR* locus was selected for study by the ENCODE pilot project and as part of this consortium we generated extensive data related to chromatin structure and modification (publicly available http://genome.ucsc.edu/ENCODE/). Figure 3.5 shows our ENCODE data representing DNaseI hypersensitivity (in Caco2, GM06990, and HeLa S3 cells) and a variety of histone modifications (as detected in GM06990 and HeLa S3 cells) for the genomic regions containing elements III and IV. Strikingly, the minimal regions containing elements III and IV (indicated by vertical lines in Figure 3.5) coincide precisely with the presence of DNaseI hypersensitive sites. These hypersensitive are found in *CFTR* expressing cells (Caco2 and HeLa S3), but not in GM06990 or HepG2 cells, which do not express *CFTR*. Also in HeLa S3 cells the regions around elements III and IV are enriched in H3K4Me1 and acetylated H4. Some H3K4Me1 is also found at element III in GM06990 cells. These comparisons strongly suggest that elements III and IV

contain functional elements, and indicate that our 3C-based element discovery approach identified *bona fide CFTR* regulatory elements.

Two Looping Elements Act as *CFTR* Enhancers

Data from the 3C experiments showed four elements looping to the *CFTR* promoter. Indeed we saw that elements III and IV contained DHS specific to cells expressing *CFTR*. We wondered if any of these four elements acted as an enhancer of *CFTR* expression. To test this, we performed a luciferase assay with all four looping elements. We used the *CFTR* promoter, consisting of 100 bp upstream of the transcription start site (TSS) and 1556 bp downstream of the TSS (hg18 chr7 116905696-116907350, total size 1656 bp), Gateway cloned into the pGL3 vector immediately upstream of the luciferase gene. Different genomic fragments (Table 3.1) were Gateway cloned directly upstream of the promoter, creating the following construct: Element – *CFTR* Promoter – Luciferase. The constructs were transfected into both Caco2 and HepG2 cells.

The results of the first assay can be seen in Figure 3.6. Figure 3.6A-E shows the genomic locations of all four elements. We split each element fragment (defined as the EcoRI fragment that was found to interact with the *CFTR* promoter fragment) into smaller fragments of 1.1–1.8 kb to allow location of the position of a potential regulatory element more precisely (Figure 3.6): element I into three (Ia, Ib and Ic), element III into two (IIIa and IIIb) and element IV into three fragments (IVa, IVb and IVc). As a positive control we used

a known enhancer located in intron 1 (Smith et al., 1996). Luciferase activity was measured in *CFTR*-expressing Caco2 cells and non-*CFTR*-expressing HepG2 cells and normalized to the level detected with the *CFTR* promoter alone. As shown in Figure 3.6F, we find that the known enhancer (intron 1) activates the *CFTR* promoter in both Caco2 cells and HepG2 cells, although to a significantly higher level in HepG2 cells. The strong activity of the intron 1 enhancer in HepG2 cells was unexpected, given that this enhancer has been reported to be specific to intestinal cells (Rowntree et al., 2001; Smith et al., 1996). Interestingly, we found that the DHS present at this enhancer in intestinal cells is also prominently present in HepG2 cells (J.A.S., unpublished results), suggesting that this enhancer may be active in these cells. Additionally, we note that this element loops to the *CFTR* promoter in both Caco2 and GM06990 cells (Figure 3.7), indicating that it may not be cell type-specific. Elements I and II have been previously described as having little effect on increasing *CFTR* expression (Smith et al., 1995). Accordingly, we did not see any enhancer activity from either of these elements in the assay. Furthermore, only one fragment (IVc), encompassing the distal portion of the EcoRI fragment containing element IV, modestly activates the *CFTR* promoter to a similar extent in both cell types. Importantly, fragment IVc overlaps the BsrGI restriction fragment that most prominently interacted with the promoter fragment, allowing us to further narrow down the element's position to a 1002 bp locus (Figure 3.6E). We noted that when element IIIb and IVc were both cloned into the luciferase vector they

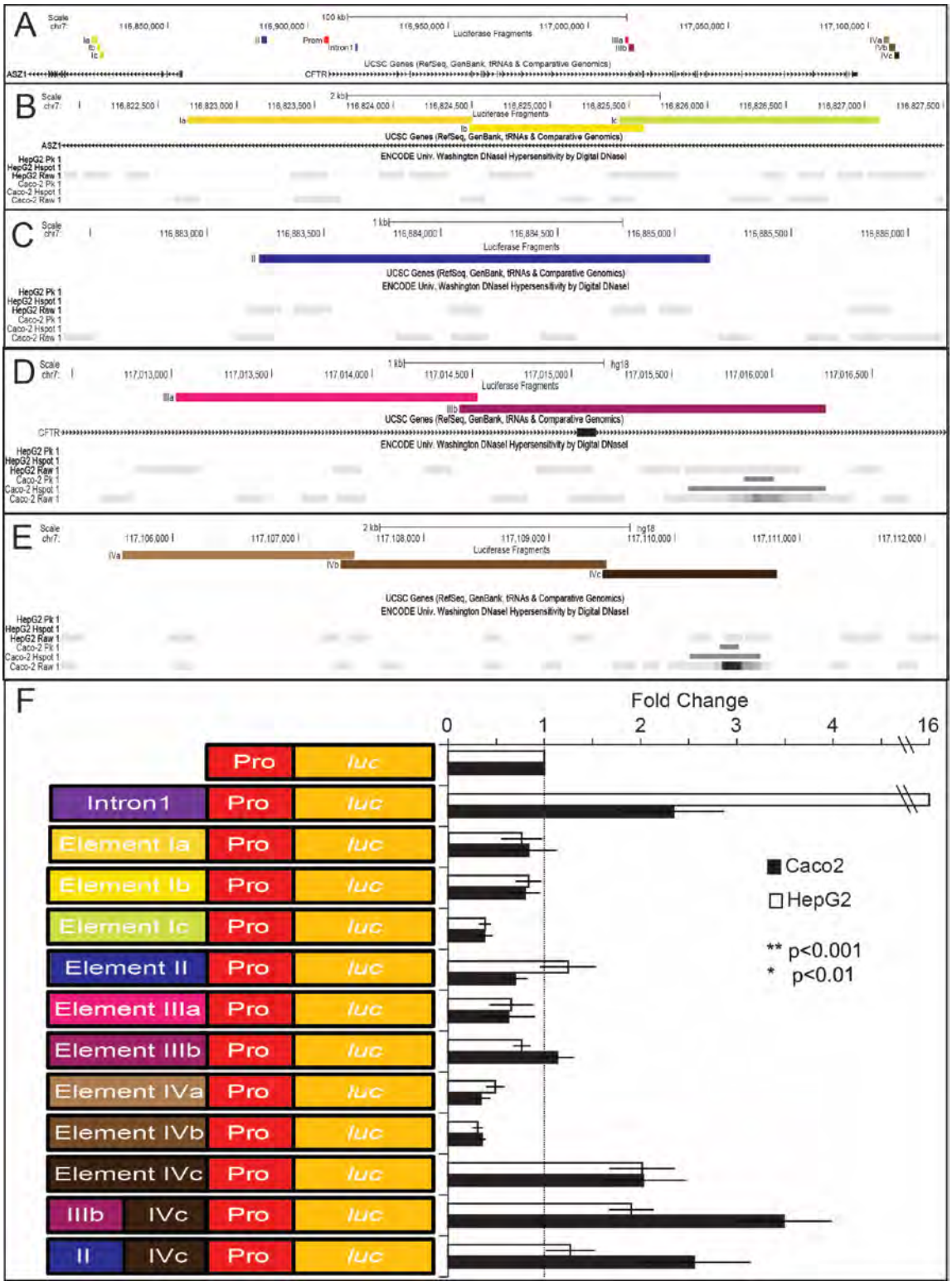

A
B
C
D
E
F
Luciferase Fragments
UCSC Genes (RefSeq, GenBank, tRNAs & Comparative Genomics)
ENCODE Univ. Washington DNaseI Hypersensitivity by Digital DNaseI
Fold Change
0
1
2
3
4
16
Pro
luc
Intron1
Element Ia
Element Ib
Element Ic
Element II
Element IIIa
Element IIIb
Element IVa
Element IVb
Element IVc
IIIb IVc
II IVc
Caco2
HepG2
** p<0.001
* p<0.01

Figure 3.6: Luciferase Assay Shows Elements IIIb and IVc Act as Additive *CFTR* Enhancers A-E) Data showing the luciferase fragments in the USCS Genome Browser. DNase1 Hypersensitivity (DHS) data is given for HepG2 and Caco2 cells. **A**) The locations of all fragments used in the luciferase assay. **B**) Zoom in to element I. The yellow rectangles show the regions cloned into the luciferase vector directly upstream of the luciferase gene. **C**) Zoom in to element II. **D**) Zoom in to element III. **E**) Zoom in to element IV. **F**) Results from the luciferase assay. Results were normalized so the level of expression in the construct containing the promoter alone is set to 1. A one-tailed T test was used to determine p-values. Error bars represent the standard error of the mean.

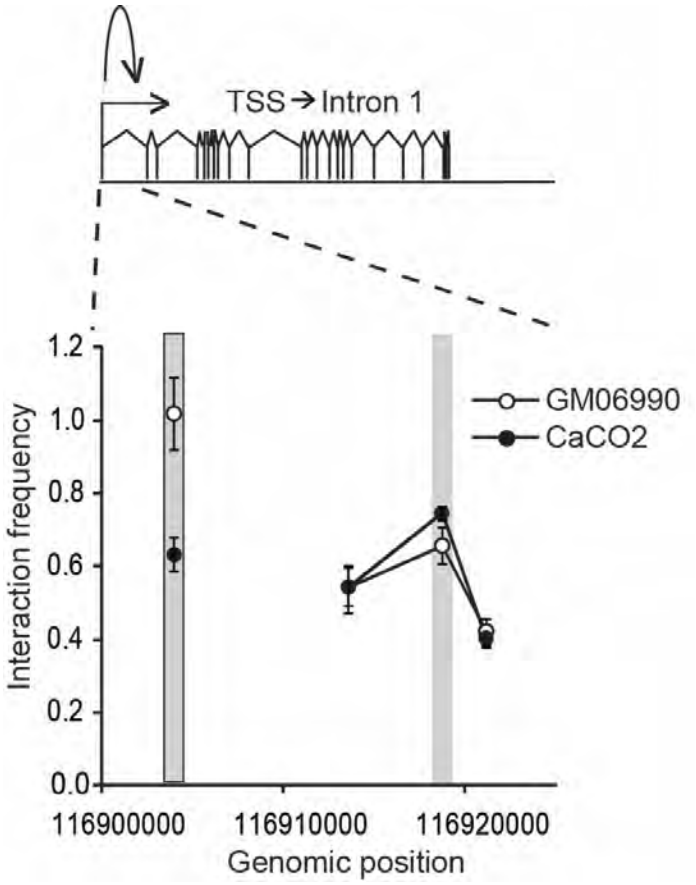


Figure 3.7: An Enhancer Element in Intron 1 Loops to the *CFTR* Promoter. BsrGI was used to digest the chromatin. The grey bar outlined in black represents the promoter anchor point. The grey bar with no outline shows the interaction peak of the fragment containing the enhancer in intron 1. In both GM06990 (open circles) and Caco2 (closed circles) the enhancer elements interacts with the promoter, though only Caco2 cells express *CFTR*, while GM06990 cells do not.

enhanced activity more than when they were cloned in separately. Additionally, both fragments contain a DHS specific to Caco2 cells (Nuthall et al., 1999b; Sabo et al., 2004, 2006). Using 3C, we have shown that these two regions contained enough information to maintain the looping contact when cut with BsrGI. Thus, we investigated these two elements further.

In the second round of luciferase assays, we fragmented elements IIIb and IVc, creating constructs containing the entire element, constructs containing just the DHS, and constructs with everything except the DHS (Figure 3.8). We then assayed these constructs in the luciferase assay as described (see Materials and Methods). As seen in Figure 3.8F, when element IIIb or IVc lose their DHS, the enhancer activity is abolished. The DHS region alone is enough to create enhanced luciferase expression (though not as strongly as the entire element). When both these minimal DHS elements are combined, an increase in enhancer activity is observed in both Caco2 and HepG2 cell lines. This result agrees with our 3C data which show elements III and IV in close three-dimensional proximity to each other and to the promoter, suggesting they may act in concert to regulate *CFTR*. Using both the DHS and luciferase data, combined with our earlier studies done with 3C, we conclude that the minimal area needed to enhance *CFTR* expression coincides with the DHS sites within each fragment. Thus, we define the minimal enhancer of element III to be 717 bp (chr7 hg18:117015552-117016269) and the minimal enhancer of element IV to

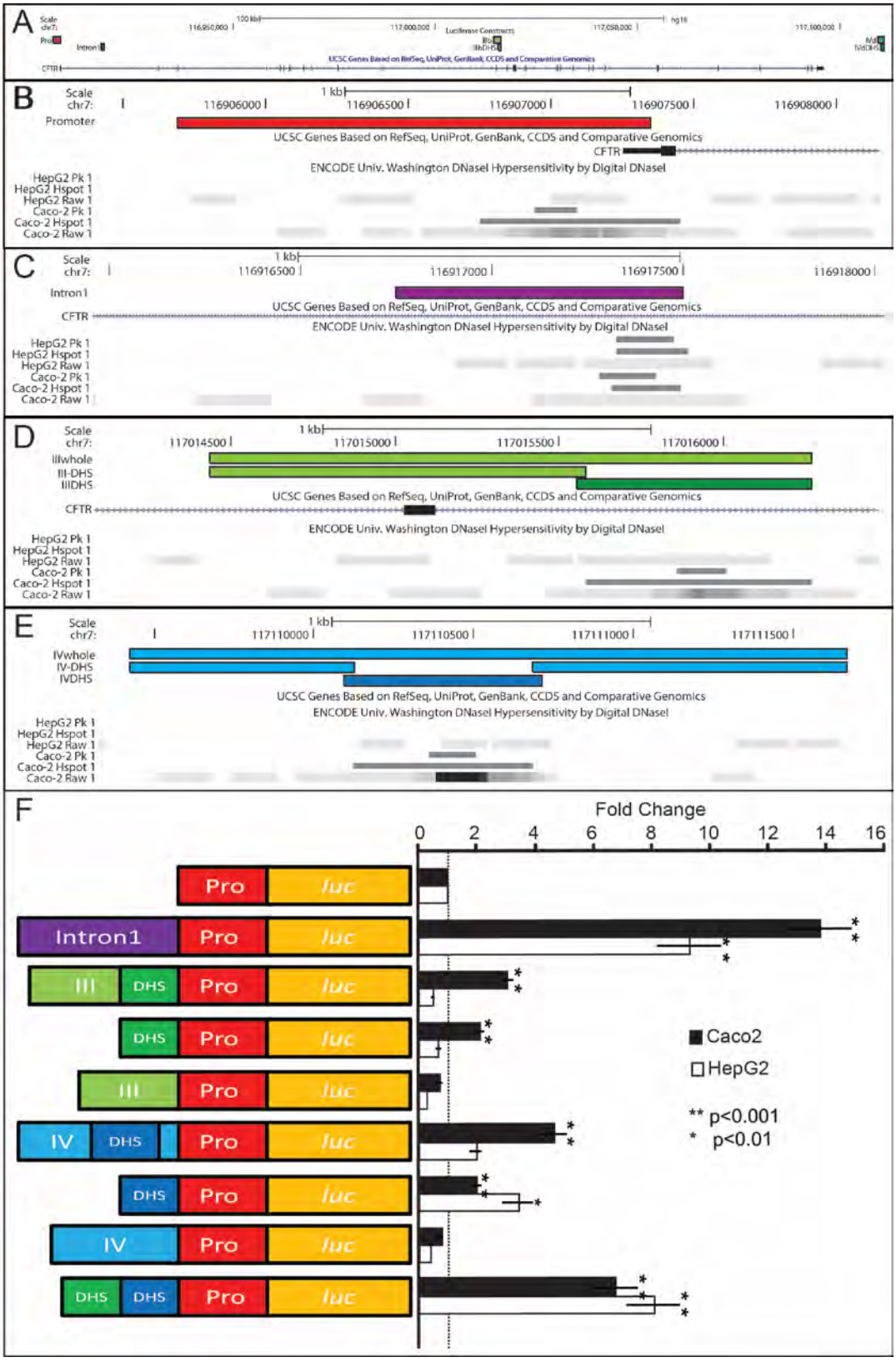

A
B
Scale
chr7:
116906000
116906500
116907000
116907500
116908000
1 kb
Promoter
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
CFTR
ENCODE Univ. Washington DNaseI Hypersensitivity by Digital DNaseI
HepG2 Pk 1
HepG2 Hspot 1
HepG2 Raw 1
Caco-2 Pk 1
Caco-2 Hspot 1
Caco-2 Raw 1
C
116916500
116917000
116917500
116918000
Intron1
D
117014500
117015000
117015500
117016000
IIIwhole
III-DHS
IIIDHS
E
117110000
117110500
117111000
117111500
IVwhole
IV-DHS
IVDHS
F
Fold Change
0
2
4
6
8
10
12
14
16
Pro
luc
Intron1
III
DHS
IV
Caco2
HepG2
** p<0.001
* p<0.01

Figure 3.8: DHS Within Elements III and IV are Necessary and Sufficient for Enhancer Function. A-E Data showing the luciferase fragments in the USCS Genome Browser. DNase1 Hypersensitivity (DHS) data is given for HepG2 and Caco2 cells. **A**) The *CFTR* gene is shown with the locations of fragments used in the luciferase assay. **B**) Zoom in to the *CFTR* promoter region. The red rectangle shows the region cloned into the luciferase vector directly upstream of the luciferase gene. **C**) Zoom in to the positive control region Intron1. **D**) Zoom in to element III. "IIIwhole" is the minimal looping region identified by 3C. "III-DHS" is the minimal looping region without the DHS. "IIIDHS" is just the DHS site as defined in Caco2 cells. **E**) Zoom in to element IV. "IVwhole" is the minimal looping region identified by 3C. "IV-DHS" is the minimal looping region without the DHS – these two fragments were combined. "IVDHS" is just the DHS site as defined in Caco2 cells. **F**) Results from the luciferase assay. Results were normalized so the level of expression in the construct containing the promoter alone is set to 1. A one-tailed T test was used to determine p-values. Error bars represent the standard error of the mean.

be 620 bp (chr7 hg18:117110095-117110715). We used these minimal enhancers to identify which transcription factors may bind to the DNA.

Identifying Transcription Factors that Bind to the Minimal Enhancer Elements

After identifying the minimal regions that enhanced *CFTR* expression, we wanted to know what transcription factors might be mediating both the enhancer effect and/or the three-dimensional contacts. Experimental technologies for identifying protein-DNA interactions include ChIP, DamID, Protein-binding microarrays (PBMs), and SELEX. However, these technologies require prior knowledge of the proteins of interest. We wanted to discover unknown proteins that bind to a known DNA sequence. For this type of experiment, the best approach is to use the Yeast One-hybrid technique (Y1H) (Deplancke et al., 2004, 2006; Reece-Hoyes et al., 2011) (Figure 3.9). This technique uses elements of interest as “bait” and a library of human transcription factors as “prey.” Yeast contain inserts of the bait in two different genomic locations (HIS3 and URA3) to decrease the amount of false positives (Figure 3.9A). The yeast is then mated to a library of yeast expressing one human transcription factor each. The yeast are plated on selective media and assayed for both growth and color (Figure 3.9B, C). We did not screen elements I and II because they do not show DHS or enhancer activity in our cell lines of interest. The baits that covered the

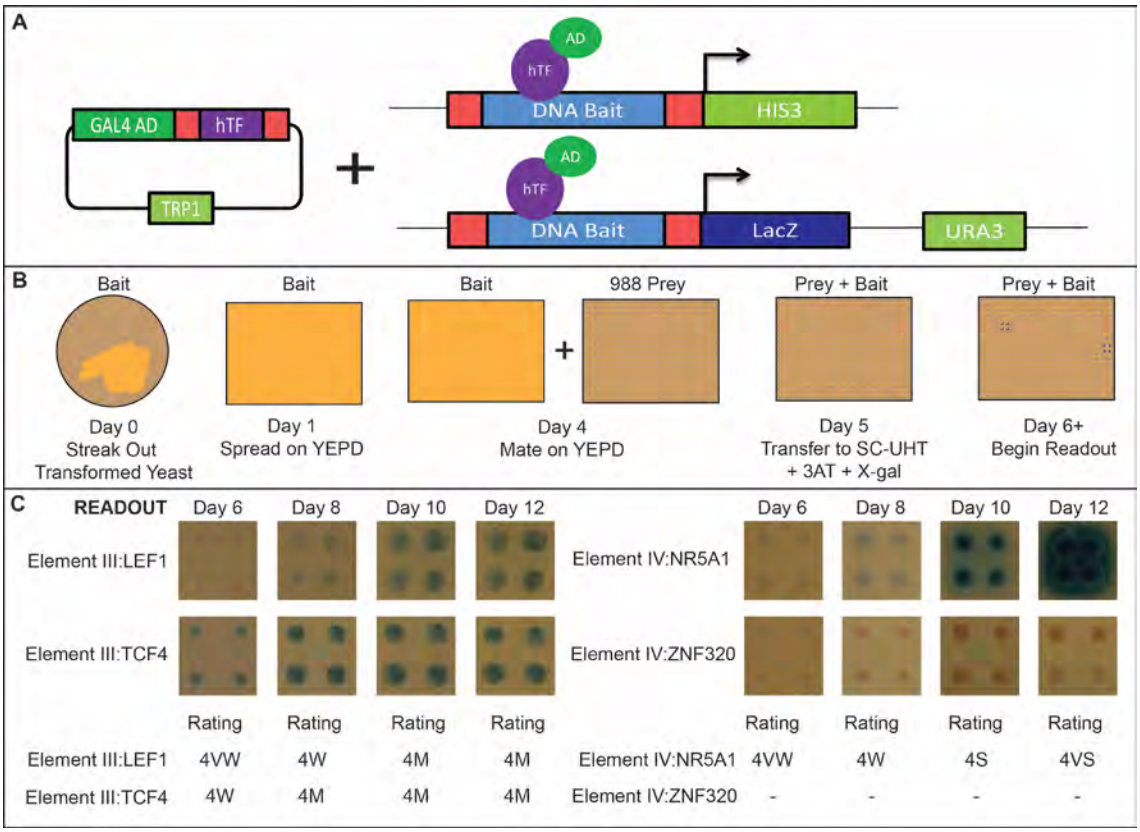


Figure 3.9: The Experimental Procedure and Sample Readout for the Yeast one-hybrid (Y1H) Assay. A) Depiction of the molecular biology used in the assay. The yeast strain YM4271 contains two insertions of the same DNA bait sequence – one located at the HIS3 locus and another, containing the LacZ gene, located at the URA3 locus. A plasmid containing the human transcription factor (hTF) of interest cloned downstream of the GAL4 activation domain (AD) and containing the TRP1 gene for selection is transformed into the yeast strain. Readout comes from growth on Sc-UHT media + 3 amino-trizole (3AT) + X-gal. **B**) The timeline of the experiment. On Day 0, the yeast containing the bait DNA sequence is plated on YEPD media. The next day the yeast is transferred to square plates with YEPD media for use with the robot. On Day 4, the bait strain is mated with the human array, which contained 1,536 clones. On Day 5 the mated yeast are transferred to selective media, and readout begins on Day 6. **C**) Samples of Y1H readout. Each bait-prey mating is plated in quadruplicate, allowing for high confidence when scoring the assay. Yeast must grow and turn blue in 2 out of the 4 locations to be called positive. The growth and color intensity were rated by eye for 7 days following plating on selective media. Three positive examples and one negative example are shown for comparison.

DHSs of elements III and IV were identical to the fragments IIIDHS and IVDHS from the luciferase assay.

Yeast one-hybrid Identified High-confidence Transcription Factors

Screening element III resulted in two high-confidence hits: TCF4 (also called TCF7L2) and LEF1. This result was surprising and interesting. TCF4 and LEF1 act together to bind specific DNA elements and are known downstream activators of the WNT signaling pathway (Beildeck et al., 2009; Hatzis et al., 2008; Korinek et al., 1998). Gratifyingly, additional evidence exists confirming these TFs as valid interactions. ChIP-seq has identified TCF4 binding to both element III and element IV in intestinal cells (Mokry et al., 2010). Additionally, a 2007 paper by Paul et al. showed that RNAi against TCF4 in Caco2 cells reduced expression of *CFTR* by half (Paul et al., 2007). With this evidence, we are quite confident that TCF4 and LEF1 do indeed play a role in regulating *CFTR* gene expression.

Screening element IV resulted in five high-confidence hits: ESRRB, ESRRG, THRB, NR5A1 and NR5A2. The ESRRB protein has similarity to the estrogen receptor though its function is unknown, and ESRRG is known to act as a transcription factor in the absence of a bound ligand. THRB is a nuclear hormone receptor for triiodothyronine. NR5A1 is involved in sex determination, and NR5A2 acts as a tissue-specific transcription factor in pancreas and liver. As

of yet there is no additional evidence relating these proteins to *CFTR* gene expression.

Discussion

3C Identified Four Elements That Loop to the *CFTR* Gene Promoter

We have employed 3C to identify long-range regulatory elements that may be involved in controlling the *CFTR* gene. We identified 4 elements that interact with the *CFTR* promoter specifically in *CFTR*-expressing cells. These looping elements appear to be gene regulatory elements based on the following observations. First, two of the elements had been previously identified as putative regulatory elements and one of these elements (element II) has been found to directly affect *CFTR* expression (Gosalia et al., 2014; Nuthall et al., 1999a; Smith et al., 1996). Second, these elements contain DNaseI hypersensitive sites and, in the case of elements III and IV patterns of histone modifications in *CFTR* expressing cells that have previously been found to be predictive of distal regulatory elements (Heintzman et al., 2007, 2009). Third, we show that elements III and IV contain DHS that act as enhancers in a luciferase assay.

Our studies did not identify all previously discovered regulatory elements around the *CFTR* locus. There are several reasons for this. First, we did not analyze all restriction fragments throughout the 460 Kb surrounding the *CFTR* promoter. For instance, we did not analyze interactions between the *CFTR* promoter and a known enhancer of the *CFTR* gene that is located 10 kb downstream of the promoter in intron 1 when we used EcoRI. The Harris lab has

recently shown that this element interacts with and regulates the *CFTR* promoter (Ott et al., 2009a). We have been able to confirm this looping interaction in our 3C experiments using BsrGI (Figure 3.7). Second, we have focused on only those elements that consistently interact with the *CFTR* promoter in *CFTR* expressing cell lines (Caco2, HT29, HeLa S3). We did find that in some cell lines (e.g. HT29) there are additional elements that appear to frequently interact with the *CFTR* promoter, for instance a region just downstream of the gene. Consistently, the Harris lab has shown that that region contains additional elements that loop to the *CFTR* promoter in primary epididymis cells (Blackledge et al., 2009). Thus, we have been able to confirm previously detected regulatory elements (elements I and II, as well as the known enhancer in intron 1), thereby validating our approach, and we have discovered novel *CFTR* regulatory elements.

Elements III and IV can act as *CFTR* Enhancers

In this study we focused on elements III and IV. Further 3C experiments allowed us to narrow the regions of interaction of these elements down to ~ 1kb. In addition, these elements coincide precisely with the presence of DNaseI hypersensitive sites present in *CFTR* expressing cells and with regions enriched in histone modifications associated with enhancers (Heintzman et al., 2007). Combined, these independent lines of research provide strong evidence that true *CFTR* regulatory elements were identified.

We find that element IV consistently enhances luciferase expression. Element III has weak enhancer activity. However these elements combined display an additive enhancer effect. Additionally, we were able to identify the minimal region necessary for enhancement by deleting the DHS region from the constructs and noting that the enhancer effect disappears (Figure 3.8). When the DHS regions were placed in the construct alone or together, these minimal elements maintained enhancer activity. DHS's are known to be marks of open, active chromatin (Birney et al., 2007). Thus it is not surprising that the minimal enhancer elements contain DHS. Locating the minimal enhancer elements also allowed us to define the smallest element to introduce into the yeast one-hybrid system to reduce off-target effects and false positive results.

The Presence of Looping Elements Does Not Correlate with Gene Expression

The long-range looping interactions described here are clearly correlated with expression of the *CFTR* gene. However, we note that the frequency of looping is not quantitatively related to the level of *CFTR* expression. The frequency of interaction between the *CFTR* promoter region and elements III and IV is quite comparable in Caco2 cells and HeLa S3 cells, despite the fact that Caco2 cells express the gene at a much higher level (Figure 3.1). This indicates that these long-range interactions are not sufficient for high levels of expression. One possible explanation is that in Caco2 cells additional transcription factors

and/or co-regulators bind these elements to further activate the gene. We also point out that the long-range looping interactions between the *CFTR* promoter and elements I – IV are not required for a low level of expression, as GM06990 cells express very low, but detectable, levels of *CFTR* while none of these long-range interactions is detected. It is possible that this low level is simply due to basal promoter activity or is the result of long-range interactions with other distal regulatory element. For instance, we find that in GM06990 cells the promoter is interacting with the enhancer element in intron 1 (Figure 3.7), and we note that intron 1 activates *CFTR* promoter expression in HepG2 cells (Figures 3.6 and 3.8). Thus, the 3D contact with intron 1 in GM06990 cells may be enough to promoter *CFTR* gene expression at a very low level.

The four elements we uncovered in our 3C analysis interact with the *CFTR* promoter in all three *CFTR* expressing cells. These elements also all interact with each other, suggesting they form a single cluster of interacting chromatin segments. We do note that the frequency of interaction between pairs of elements varies (Figure 3.3), and is also somewhat different between cell lines. Thus, although the overall conformation of the locus is comparable in Caco2 and HeLa S3 cells, there may be slight differences in the frequencies with which long-range interaction occur. Alternately, HeLa S3 cells may be in a poised conformation, ready to express *CFTR* at a higher level upon some stimulus.

High-confidence Human Transcriptions Factors Bind to *CFTR* Enhancers

We tested the elements found to be enhancers in a yeast one-hybrid assay (Deplancke et al., 2004, 2006; Reece-Hoyes et al., 2011). This assay contained clones for 988 human transcription factors. We used the minimal enhancer regions identified in the luciferase assay to act as the bait for the hTFs. Any hTF that had a positive hit in all three biological references was placed into our high-confidence category. For element III we identified two high-confidence hTFs: LEF1 and TCF4. For element III we identified five high-confidence hTFs: ESRRB, ESRRG, THRB, NR5A1 and NR5A2.

We identified some transcription factors as false positives, including ZIC1 and ZIC3. These TFs gave extremely high readouts with many of the tested baits. It is hard to identify false negatives in the Y1H, though they will indeed occur. FOXA1 has been shown by ChIP to be bind element III specifically in Caco2 cells (Yigit et al., 2013). FOXA1 is included in the Y1H array at position 01_A01 but the clone's sequence was incorrect. Therefore we did not actually test FOXA1 in the Y1H, though the hTF name is included on the list of hTFs available in the assay.

Once we obtained our high-confidence hits, we did a literature search to see if there was any previous evidence linking the hTFs with *CFTR* gene expression. We were able to find literature supporting TCF4 and LEF1 binding to element III (Mokry et al., 2010). Interestingly, Mokry et al found TCF4 binding to both element III and IV, but our yeast one-hybrid screen only found TCF4 binding to element III. We know from our 3C experiments that the enhancer in element III

and the enhancer in element IV are all connected in three-dimensional space. Therefore, it could be that the TCF4 found at element IV by ChIP resulted from its close proximity to element III. Alternatively, theY1H might not have picked up TCF4 binding at element IV.

Identification of extragenic elements that affect expression of *CFTR* will not only provide basic insights into spatio-temporal regulation of this important gene, it may also be important for genetic diagnosis of Cystic Fibrosis. A significant number of patients with Cystic Fibrosis symptoms do not appear to carry mutations in the *CFTR* exons or promoter, suggesting that extra-genic or intronic mutations may be present, e.g. in long-range acting gene regulatory elements. In the absence of information of the positions of distant regulatory elements it is not feasible to screen for mutations in a very large genomic region. Thus, identification of *CFTR* regulatory elements, as we have described here, provides new targets for mutation screening. Further, gene therapy approaches for Cystic Fibrosis could benefit from knowledge of gene regulatory elements by including such elements in *CFTR* gene targeting constructs.

We have shown that 3C technology can be used to discover novel regulatory elements throughout gene loci. A variety of 3C adaptations have recently been developed that allow large-scale detection of chromatin looping interactions (Dostie et al., 2006; Lieberman-Aiden et al., 2009; Simonis et al., 2006; Zhao et al., 2006). Using these technologies, it may become possible to map the regulatory elements that control genes throughout the genome.

CHAPTER IV

THREE-DIMENSIONAL STUDY OF THE CYSTIC FIBROSIS LOCUS REVEALS CELL-TYPE SPECIFIC CHROMATIN LOOPING AND WELL-DEFINED TAD STRUCTURE

This chapter is adapted from the paper "Regulation of gene expression around the *CFTR* locus through cell type-specific chromatin looping in the context of invariant topologically associating domains" by E.M. Smith, B. Lajoie, G. Jain and J. Dekker, in preparation.

Contributions

This paper examines a 2.8 Mb region surrounding the *CFTR* gene and shows well-defined TAD structure in the region. It analyzes gene expression and looping interactions within the region. E.M.S. and B.L. designed the 5C experiment. E.M.S. performed the experiments. E.M.S., B.L. and G.J. performed data analysis. E.M.S. and J.D. wrote the manuscript. J.D. provided guidance and feedback throughout the project.

Abstract

Three-dimensional genome structure plays an important role in gene expression and regulation. Chromosomes are organized into active and inactive compartments, while genes are regulated by specific looping interactions between promoters and regulatory elements. In between these two levels of structure, topologically associating domains (TADs) have recently been described. TADs occur throughout the genome range in size from a few hundred kilobases to around a megabase. Major questions include TADs relationship to long-range looping interactions between genes and regulatory elements. Here we examine chromosome conformation of a 2.8 Mb region on human chromosome 7 surrounding the cystic fibrosis gene (*CFTR*) in a panel of different cell types. We find the same set of 6 TAD boundaries present in all cell types studied, irrespective of gene expression status or internal looping. In contrast the internal organization of TADs is cell line-specific. This is particularly exemplified by the *CFTR* promoter that interacts with distinct sets of cell type-specific regulatory elements located within the same TAD. Interestingly, interactions between TAD are also highly cell type-specific and often occur in clusters at and around TAD boundaries. These data shed new light on the roles of TAD boundaries in constraining as well as mediating long-range looping interactions and gene regulation.

Introduction

The three-dimensional (3D) structure of chromosomes is thought to play a critical role in gene regulation. At the nuclear level, individual chromosomes occupy their own territories within the cell nucleus (Bolzer et al., 2005; Cremer and Cremer, 2001), with some intermingling where they touch (Branco and Pombo, 2006). Larger chromosomes tend to be positioned more peripherally, while the smaller chromosomes are preferentially located near other small chromosomes in the center of the nucleus (Boyle et al., 2001, 2011; Croft et al., 1999). Chromosomes themselves are compartmentalized so that active (open) and inactive (closed) chromatin domains are spatially separated. In Hi-C data this is apparent by the detection of A and B compartments (Lieberman-Aiden et al., 2009): large (several Megabases (Mb)) chromatin domains that alternate along the length of the chromosomes. A compartments represent active regions of chromosomes as assessed by gene expression and the presence of chromatin features such as DNase1 sensitivity and the presence of active histone modifications (H3K4Me3, H3K27Ac). B compartments typically display less or no transcription and are composed of closed chromatin (Lieberman-Aiden et al., 2009; Zhang et al., 2012). These compartments tend to cluster with themselves, i.e. A compartments interact more frequently with other A compartments and less frequently with B compartments, and vice versa. A compartments may represent transcription factories where active genes cluster together, while inactive

chromatin is constrained to repressed sites in the nucleus such as the nuclear periphery (Cook, 1999; Guelen et al., 2008; Iborra et al., 1996; Wansink et al., 1993).

At a considerably smaller scale, chromatin organization plays a direct role in regulation of gene expression through looping interactions between gene promoters and their distal regulatory elements, including enhancers and CTCF-bound insulator-like elements. Such locus-specific looping interactions mostly occur on a scale of a few kilobases (kb) to 1 Mb (Amano et al., 2009; Baù et al., 2011; Gheldof et al., 2010; Lettice et al., 2003; Murrell et al., 2004; Ott et al., 2009b; Palstra et al., 2003; Phillips-Cremins et al., 2013; Sagai et al., 2005; Sanyal et al., 2012; Tolhuis et al., 2002; Vernimmen et al., 2007; Wright et al., 2010). This is consistent with genetic analyses and functional studies that show that most regulatory elements act on a length scale of several hundred kb (Kleinjan and van Heyningen, 2005).

Recently, a new feature of chromatin organization was described at intermediate length scales: topologically associating domains (TADs) (Dixon et al., 2012; Nora et al., 2012; Sexton et al., 2012). TADs are defined as contiguous chromatin domains that display relatively high levels of self-association, separated by boundaries. Loci located in adjacent TADs interact much less frequently, suggesting that TAD boundaries act as physical insulators. TADs range in size from a few hundred kb to a few Mb, and are found across cell types and across species (Dixon et al., 2012; Nora et al., 2012). TAD boundaries seem

to correlate with CTCF sites, consistent with an insulation-type role (Dixon et al., 2012; Phillips-Cremins et al., 2013). However, not all TAD boundaries occur at a CTCF site, indicating that the presence of a CTCF site is not sufficient for formation of a TAD boundary. Deletion of a TAD boundary region causes the two neighboring TADs to partially intermingle (Nora et al., 2012). TADs seem to be invariant between the small set of cell types studied to date (Dixon et al., 2012; Nora et al., 2012) although differences in the internal organization of TADs has been observed between different cell lines (Phillips-Cremins et al., 2013). TADs are also structurally modulated during the cell cycle, when they disappear in mitosis and reform in early G1 (Naumova et al., 2013).

Several lines of evidence indicate that TADs are important for appropriate regulation of gene expression. First, genes located within a TAD show, in some cases, increased correlation in their expression during cell differentiation, as compared to genes located in different TADs (Nora et al., 2012). Second, domains of histone modifications and lamin association, both features related to the expression status of genes, can correlate with TADs (Dily et al., 2014; Nora et al., 2012). Third, using an enhancer sensor approach functional domains were identified that represent target regions of enhancers. These domains correlated well with TADs, suggesting that regulatory elements can act on entire TADs as a structural unit (Symmons et al., 2014). Further, looping interactions between promoters and enhancers occur frequently within TADs and less frequently between TADs (Jin et al., 2013).

A major question is whether TADs are defined by their boundaries only or whether their formation is determined or facilitated by looping interactions between loci, e.g. between promoters and enhancers, located within them (Giorgetti et al., 2014). Thus, a key question is whether TADs act upstream of chromatin looping or whether TADs are, at least in part, driven by looping interactions within them. Another question is whether and how TADs interact with each other, and whether specific elements are involved.

In order to examine the relationship between TAD structure and looping interactions in more depth, we analyzed a 2.8 Mb region on chromosome 7 containing the cystic fibrosis transmembrane conductance regulator gene (*CFTR*). We and others have previously used 3C to identify several regulatory elements located up to 200 kb from the CFTR promoter that directly loop and interact with the promoter and with each other (Gheldof et al., 2010; Ott et al., 2009). We expand our original study to include 11 more genes and a 700 kb gene-poor region. We use a variety of cell types to understand how the 3D structure of the locus may contribute to *CFTR* gene regulation. We find that 7 TADs exist in this region, and that all the known regulatory elements of *CFTR* are contained within one TAD. We find that all cell lines maintain the same TAD structure, regardless of gene expression within the TAD. We provide detailed information about the types of loops that occur within and between TADs.

Materials and Methods

Cell Culture

Five cell lines were used in the experiments in this chapter. All cell lines were grown with antibiotic (1% penicillin–streptomycin). GM12878 lymphoblastoid cells (Coriell Cell Repositories) were grown in RPMI 1640 medium supplemented with 2mM L-glutamine and 15% fetal bovine serum (FBS). HepG2 hepatocellular carcinoma cells (ATCC) were grown in MEMα with 10% FBS. Caco2 colorectal adenocarcinoma cells (ATCC) were grown in MEMα with 20% FBS. Calu3 lung adenocarcinoma cells (ATCC) were grown in ATCC-formulated E-MEM with 10% FBS. Capan1 pancreas adenocarcinoma cells (ATCC) were grown in IMDM with 20% FBS. Cell densities were maintained as recommended using Accutase (Life) to detach adherent cells from plates.

Chromosome Conformation Capture Carbon Copy (5C)

5C experiment

5C was carried out as previously described (Sanyal et al., 2012). We investigated a 2.8 Mb region on Chromosome 7 (hg18 chr7:115797757-118405450) containing the ENCODE region ENm001. The 5C experiment was designed to interrogate looping interactions between HindIII fragments containing transcription start sites (TSSs) and any other HindIII restriction fragment (distal fragments) in the target region. Libraries were generated for five cell lines:

Caco2, Calu3, Capan1, GM12878 and HepG2, with two biological replicates for each line.

5C probe design

5C probes were designed at HindIII restriction sites (AAGCTT) using 5C primer design tools previously developed and made available online at My5C website (http://my5C.umassmed.edu) (Lajoie et al., 2009). Probes were designed based on the ENCODE manual region 1 design (Sanyal et al., 2012) with additional probes placed throughout the region when appropriate. We also added probes in a gene desert region approximately 700 kb in length. All probe locations can be found in Table 4.1. Probe settings were: U-BLAST, 3; S-BLAST, 100; 15-MER, 3,000; MIN_FSIZE, 250; MAX_FSIZE, 20,000; OPT_TM, 65; OPT_PSIZE, 40.

Generation of 5C libraries

3C was performed with HindIII restriction enzyme as previously described (Naumova et al., 2012) for Caco2, Calu3, Capan1, GM12878 and HepG2 cells separately with two biological replicates for each cell line. The 3C libraries were then interrogated by 5C. The region was analyzed by pooling all probes for a final concentration of 0.5 fmol μl^{-1}. In total, 75 reverse probes and 605 forward probes were pooled for a possible 44,770 interactions in the *CFTR* region (Probes can be found in Appendix I). 5C was performed as described (Sanyal et al., 2012) with the following changes: 10 ligation reactions were performed for each 5C

library, each containing an amount of 3C template that represents 400,000 genome equivalents and 2 fmol of each primer.

5C read mapping

Sequencing data was obtained from an Illumina GAIIx machine and was processed by a custom pipeline to map and assemble 5C interactions, as previously described (Lajoie et al., 2009; Sanyal et al., 2012). We used an updated version of the novoalign mapping algorithm (V2.07.11 http://novocraft.com). Statistics regarding the 5C library quality, mapping efficiency, etc. can be found in Table 4.1. A table summarizing the read depth of each 5C library can be found in Table 4.2. Pearson correlation coefficients between the biological replicates can be found in Table 4.3.

5C bias correction

5C experiments involve a number of steps that can differ in efficiency, thereby introducing biases in efficiency of detection of pairs of interactions. These biases could be due to differences in the efficiency of crosslinking, the efficiency of restriction digestion (related to crosslinking efficiency), the efficiency of ligation (related to fragment size), the efficiency of 5C probes (related to annealing and PCR amplification) and finally the efficiency of DNA sequencing (related to base composition). All of these potential biases—several of which are common to other approaches such as chromatin immunoprecipitation (for example, crosslinking efficiency, PCR amplification, base-composition-dependent

Mapping	**Caco2 R1**		**Caco2 R2**	
# numRawReads	13,734,053		13,763,949	
# numQCFailedReads	0	0.00%	0	0.00%
#numQCPassedReads	13,734,053	100.00%	13,763,949	100.00%
# side1Mapped	13,215,901	96.20%	13,121,771	95.30%
# side2Mapped	13,184,787	96.00%	13,040,901	94.70%
# noSideMapped	263,970	1.90%	326,612	2.40%
# oneSideMapped	539,478	3.90%	712,002	5.20%
# bothSideMapped	12,930,605	94.10%	12,725,335	92.50%
# errorPairs		0.00%	0	0.00%
# invalidPairs	447	0.00%	843	0.00%
# validPairs	12,930,158	94.10%	12,724,492	92.40%

Mapping	**Calu3 R1**		**Calu3 R2**	
# numRawReads	19,323,745		26,078,596	0.00%
# numQCFailedReads	0	0.00%	0	100.00%
#numQCPassedReads	19,323,745	100.00%	26,078,596	98.80%
# side1Mapped	18,669,930	96.60%	25,754,552	98.30%
# side2Mapped	18,640,276	96.50%	25,625,059	0.30%
# noSideMapped	333,334	1.70%	80,343	2.40%
# oneSideMapped	670,616	3.50%	616,895	97.30%
# bothSideMapped	18,319,795	94.80%	25,381,358	0.00%
# errorPairs	8	0.00%	0	0.00%
# invalidPairs	11,169	0.10%	877	97.30%
# validPairs	18,308,618	94.70%	25,380,481	

Mapping	**Capan1 R1**		**Capan1 R2**	
# numRawReads	18,671,787		28,715,154	
# numQCFailedReads	0	0.00%	0	0.00%
#numQCPassedReads	18,671,787	100.00%	28,715,154	100.00%
# side1Mapped	18,433,875	98.70%	28,281,242	98.50%
# side2Mapped	18,283,457	97.90%	27,887,919	97.10%
# noSideMapped	56,948	0.30%	99,813	0.30%
# oneSideMapped	512,346	2.70%	1,061,521	3.70%
# bothSideMapped	18,102,493	97.00%	27,553,820	96.00%
# errorPairs	1	0.00%	0	0.00%
# invalidPairs	1,254	0.00%	5,937	0.00%
# validPairs	18,101,238	96.90%	27,547,883	95.90%

Mapping	**GM12878 R1**		**GM12878 R2**	
# numRawReads	33,025,556		29,352,987	
# numQCFailedReads	0	0	0	0.00%
#numQCPassedReads	33,025,556	100	29,352,987	100.00%
# side1Mapped	12,811,429	38.8	28,698,758	97.80%
# side2Mapped	12,876,439	39	28,526,444	97.20%
# noSideMapped	19,462,222	58.9	205,409	0.70%
# oneSideMapped	1,438,800	4.4	1,069,954	3.60%
# bothSideMapped	12,124,534	36.7	28,077,624	95.70%
# errorPairs	3	0	37	0.00%
# invalidPairs	20,454	0.1	65,809	0.20%
# validPairs	12,104,077	36.7	28,011,778	95.40%

Mapping	**HepG2 R1**		**HepG2 R2**	
# numRawReads	33,762,419		27,522,416	
# numQCFailedReads	0	0.00%	0	0.00%
#numQCPassedReads	33,762,419	100.00%	27,522,416	100.00%
# side1Mapped	32,908,047	97.50%	26,989,444	98.10%
# side2Mapped	32,301,383	95.70%	26,840,431	97.50%
# noSideMapped	254,633	0.80%	155,116	0.60%
# oneSideMapped	1,806,142	5.30%	904,725	3.30%
# bothSideMapped	31,701,644	93.90%	26,462,575	96.10%
# errorPairs	3	0.00%	24	0.00%
# invalidPairs	20,263	0.10%	36,035	0.10%
# validPairs	31,681,378	93.80%	26,426,516	96.00%

Table 4.1 Mapping Statistics of All Libraries Used in this Study.

	Raw Data		*cis*-purged Data	
	#interactions	#reads	#interactions	#reads
Caco2-R1	**50,986**	**12,930,158**	**24,492**	**9,240,611**
Caco2-R2	**50,970**	**12,724,492**	**27,090**	**9,783,411**
Calu3-R1	**66,029**	**18,308,618**	**36,812**	**12,073,332**
Calu3-R2	**148,785**	**25,380,481**	**39,883**	**15,507,757**
Capan1-R1	**126,462**	**18,101,240**	**38,046**	**6,041,896**
Capan1-R2	**161,499**	**27,547,883**	**39,308**	**8,722,529**
GM12878-R1	**88,198**	**24,408,480**	**37,880**	**9,907,568**
GM12878-R2	**190,199**	**28,011,778**	**40,190**	**15,253,503**
HepG2-R1	**161,365**	**31,681,378**	**39,568**	**24,630,673**
HepG2-R2	**147,911**	**26,426,516**	**39,909**	**16,945,637**
	Singleton-removed Data		Coverage-corrected Data	
	#interactions	#reads	#interactions	#reads
Caco2-R1	**24,449**	**7,265,532**	**20,351**	**7,354,017**
Caco2-R2	**27,046**	**9,453,039**	**22,278**	**8,248,685**
Calu3-R1	**36,768**	**11,970,277**	**32,886**	**8,222,536**
Calu3-R2	**39,908**	**15,378,124**	**36,733**	**8,522,545**
Capan1-R1	**38,002**	**5,760,920**	**35,337**	**8,853,211**
Capan1-R2	**39,264**	**8,380,681**	**36,799**	**8,897,244**
GM12878-R1	**37,836**	**9,687,772**	**34,869**	**10,734,618**
GM12878-R2	**40,146**	**14,800,666**	**39,392**	**9,711,216**
HepG2-R1	**39,524**	**24,508,802**	**35,515**	**11,181,660**
HepG2-R2	**39,865**	**16,718,446**	**35,701**	**9,771,604**

Table 4.2 Read Counts for Each Correction Step.

	Caco2 R1	Caco2 R2	Calu3 R1	Calu3 R2	Capan1 R1	Capan1 R2	GM12878 R1	GM12878 R2	HepG2 R1	HepG2 R2
Caco2 R1	1									
Caco2 R2	0.738	1								
Calu3 R1	0.617	0.754	1							
Calu3 R2	0.645	0.816	0.911	1						
Capan1 R1	0.617	0.751	0.792	0.792	1					
Capan1 R2	0.594	0.752	0.802	0.796	0.896	1				
GM12878 R1	0.518	0.731	0.651	0.669	0.707	0.726	1			
GM12878 R2	0.503	0.677	0.642	0.656	0.647	0.632	0.799	1		
HepG2 R1	0.565	0.798	0.736	0.747	0.779	0.755	0.825	0.816	1	
HepG2 R2	0.447	0.589	0.551	0.564	0.569	0.539	0.705	0.897	0.762	1

Table 4.3: Pearson Correlation of All Replicates.

sequencing efficiency)—will have an impact on the overall efficiency with which long-range interactions for a given locus (restriction fragment) can be detected. We implemented the following steps to estimate and correct for such technical biases.

Probe filtering – Cis-Purge

Not all probes are represented equally in our 5C dataset due to over- and under-performance in the assay. As the first step in our data correction pipeline, we remove probes that perform significantly differently than the overall set. The relative performance of each probe is determined as follows. First, a global average relationship between interaction frequency and genomic distance is calculated using Loess smoothing for each data set. Interaction profiles anchored on each probe across the 2.8 Mb region is then compared to this global average. If the individual Loess is more or less than .85 of the scaled Z-score distance (a measurement of the number of standard deviations a data point is from the mean) from the average global Loess, the probe is flagged as problematic. If a probe is flagged as problematic in more than 40% of the datasets, it is removed from downstream analysis. Using this threshold we removed 34 probes from downstream analysis (Table 4.4).

Singleton removal

As we examined the 5C datasets, we noticed several instances when the interaction between two probes was higher than its neighboring interactions by

Trim Amount: 0.85

Flagged Probe	#datasets in which this probe is flagged
5C_2410_EMS03_FOR_102	10
5C_2410_EMS03_FOR_140	10
5C_2410_EMS03_FOR_349	10
5C_2410_EMS03_FOR_429	10
5C_2410_EMS03_FOR_773	10
5C_2410_EMS03_FOR_165	10
5C_2410_EMS03_FOR_2	10
5C_2410_EMS03_REV_10	9
5C_2410_EMS03_FOR_75	9
5C_2410_EMS03_FOR_74	9
5C_2410_EMS03_REV_13	8
5C_2410_EMS03_FOR_8	8
5C_2410_EMS03_FOR_51	8
5C_2410_EMS03_FOR_54	7
5C_2410_EMS03_FOR_197	7
5C_2410_EMS03_FOR_762	7
5C_2410_EMS03_REV_111	7
5C_2410_EMS03_FOR_228	7
5C_2410_EMS03_FOR_389	7
5C_2410_EMS03_FOR_350	7
5C_2410_EMS03_REV_523	6
5C_2410_EMS03_FOR_407	6
5C_2410_EMS03_FOR_1	6
5C_2410_EMS03_FOR_57	5
5C_2410_EMS03_FOR_129	5
5C_2410_EMS03_FOR_677	5
5C_2410_EMS03_FOR_117	5
5C_2410_EMS03_FOR_248	4
5C_2410_EMS03_FOR_246	4
5C_2410_EMS03_FOR_864	4
5C_2410_EMS03_FOR_283	4
5C_2410_EMS03_FOR_607	4
5C_2410_EMS03_FOR_658	4
5C_2410_EMS03_FOR_298	4

Table 4.4: Probes Removed in the Probe Filtering Step of the Correction Method

an order or magnitude or more. Although there were not many of these blowouts, we removed them from the dataset to avoid problems downstream during peak calling. Thus, we removed any interaction that had a Z-score of 12 or more, resulting in the removal of 44 individual interactions from downstream analysis (Table 4.5). To calculate the Z-score of a given data point, we used the following equation:

$$Z = (x - \mu)/\sigma$$

where the Z-score (Z) is the raw score (x) minus the population mean (μ) divided by the standard deviation (σ). This measurement was calculated for the given distance of the raw score, meaning the average and standard deviation were calculated only for points separated by the same genomic distance as the data point (x).

Coverage correction

Once the outlier probes and interactions are removed from the 5C data set, the profiles of each probe were normalized so that they could be quantitatively compared to each other. Here, we applied a slightly different method compared to the one we used before (Sanyal et al., 2012) based only on local (*cis*) chromatin interaction data (within the *CFTR* region) because that data is the most reliably detected. First, all 5C datasets were read normalized (each interaction value was divided by the number of reads obtained for that dataset).

z-score 12

Probe-Probe interaction	#datasets in which this interaction is flagged
5C_2410_EMS03_FOR_124_5C_2410_EMS03_REV_17	10
5C_2410_EMS03_FOR_128_5C_2410_EMS03_REV_17	10
5C_2410_EMS03_FOR_464_5C_2410_EMS03_REV_405	10
5C_2410_EMS03_FOR_126_5C_2410_EMS03_REV_17	9
5C_2410_EMS03_FOR_128_5C_2410_EMS03_REV_203	9
5C_2410_EMS03_FOR_21_5C_2410_EMS03_REV_203	9
5C_2410_EMS03_FOR_339_5C_2410_EMS03_REV_898	9
5C_2410_EMS03_FOR_605_5C_2410_EMS03_REV_777	9
5C_2410_EMS03_FOR_7_5C_2410_EMS03_REV_405	9
5C_2410_EMS03_FOR_817_5C_2410_EMS03_REV_405	9
5C_2410_EMS03_FOR_90_5C_2410_EMS03_REV_613	9
5C_2410_EMS03_FOR_93_5C_2410_EMS03_REV_405	9
5C_2410_EMS03_FOR_467_5C_2410_EMS03_REV_719	8
5C_2410_EMS03_FOR_548_5C_2410_EMS03_REV_405	8
5C_2410_EMS03_FOR_579_5C_2410_EMS03_REV_72	8
5C_2410_EMS03_FOR_125_5C_2410_EMS03_REV_17	7
5C_2410_EMS03_FOR_19_5C_2410_EMS03_REV_203	7
5C_2410_EMS03_FOR_214_5C_2410_EMS03_REV_761	7
5C_2410_EMS03_FOR_71_5C_2410_EMS03_REV_761	7
5C_2410_EMS03_FOR_87_5C_2410_EMS03_REV_782	7
5C_2410_EMS03_FOR_105_5C_2410_EMS03_REV_405	6
5C_2410_EMS03_FOR_208_5C_2410_EMS03_REV_711	6
5C_2410_EMS03_FOR_503_5C_2410_EMS03_REV_420	6
5C_2410_EMS03_FOR_125_5C_2410_EMS03_REV_665	5
5C_2410_EMS03_FOR_132_5C_2410_EMS03_REV_17	5
5C_2410_EMS03_FOR_139_5C_2410_EMS03_REV_405	5
5C_2410_EMS03_FOR_154_5C_2410_EMS03_REV_405	5
5C_2410_EMS03_FOR_641_5C_2410_EMS03_REV_420	5
5C_2410_EMS03_FOR_68_5C_2410_EMS03_REV_594	5
5C_2410_EMS03_FOR_98_5C_2410_EMS03_REV_405	5
5C_2410_EMS03_FOR_12_5C_2410_EMS03_REV_203	4
5C_2410_EMS03_FOR_136_5C_2410_EMS03_REV_405	4
5C_2410_EMS03_FOR_237_5C_2410_EMS03_REV_665	4
5C_2410_EMS03_FOR_394_5C_2410_EMS03_REV_405	4
5C_2410_EMS03_FOR_5_5C_2410_EMS03_REV_808	4

5C_2410_EMS03_FOR_502_5C_2410_EMS03_REV_420	4
5C_2410_EMS03_FOR_503_5C_2410_EMS03_REV_421	4
5C_2410_EMS03_FOR_597_5C_2410_EMS03_REV_666	4
5C_2410_EMS03_FOR_626_5C_2410_EMS03_REV_420	4
5C_2410_EMS03_FOR_63_5C_2410_EMS03_REV_405	4
5C_2410_EMS03_FOR_641_5C_2410_EMS03_REV_421	4
5C_2410_EMS03_FOR_672_5C_2410_EMS03_REV_203	4
5C_2410_EMS03_FOR_751_5C_2410_EMS03_REV_665	4
5C_2410_EMS03_FOR_883_5C_2410_EMS03_REV_405	4

Table 4.5: Individual Interactions Removed in the Singleton Removal Step of the Correction Method. We used a z-score threshold of 12 as our cutoff. This cutoff maintained interactions we saw from 3C experiments (Gheldof et al. 2010, Ott et al., 2009) but still eliminated points displaying interactions orders of magnitude larger than their neighbors.

Second, to determine differential performance of probes due to non-biological technical biases (see above) we combined all 10 5C datasets. Next, a global average relationship between interaction frequency and genomic distance was calculated using Loess smoothing, and the interaction profile detected with each probe was compared to this average. In the absence of any technical detection bias we assume that the overall domain-wide profile of each probe is similar to the dataset-wide average profile. For each probe we then calculated a correction factor by which the profile of that probe should be lifted or lowered in order to match the average Loess profile from the combined dataset. The calculated factors for each probe were then applied to correct each of the 10 individual, read-normalized datasets to produce the final bias corrected datasets. By combining the datasets before correction, we reduce the risk of overcorrecting certain probes that were truly giving high biological signals in one cell line. When the datasets were combined, this high signal was averaged with the other lower signals, giving that interaction a less stringent correction. If correction factors were calculated for each dataset separately, the interaction would be penalized for giving such a high interaction and corrected too harshly. Data corrections did not change the overall structure of the data. In fact, analysis of specific probes from raw and corrected data show only few differences in their profiles, indicating that probes display only minor differences in detection efficiency.

Insulation Index (TAD calling)

To define regions of our dataset that contain TAD boundaries we calculated an insulation score along the locus. This method is based on the concept that TAD boundaries act as physical insulators that prevent or inhibit interaction across them. First, 5C data was binned at 100 kb with a 10x step size. Next, we calculated for each bin the combined number of observed interactions across it by summing all interactions between loci located up to 250 Kb upstream with loci located up to 250 Kb downstream of the bin. This sum was then calculated for each bin along the 2.8 Mb locus, and then divided by the average sum of all bins to obtain insulation scores. Insulation scores are then plotted along the locus to obtain an insulation profile. Local minima in the insulation profiles were detected by identifying the bins with the lowest insulation scores in a local 500 kb window. We set the midpoint of this low-value bin as the boundary.

Peak calling

To detect statistically significant looping interactions at the restriction fragment level we applied a '5C peak calling' algorithm as described before (Lajoie et al., 2009; Sanyal et al., 2012) with the following modifications. We called peaks on three different sub-sets of the data – all the data, intraTAD data and interTAD data. Peaks were defined as signals that are significantly higher than expected. Expected values were calculated as follows: For peak calling of the complete dataset we calculated the average interaction frequency for each genomic distance using Loess smoothing (alpha value 0.01). This provides a

weighted average and a weighted standard deviation at each genomic distance. For peak calling within or between TADs separately, we calculated the average interaction frequency for each genomic distance using Loess smoothing using only intraTAD or interTAD data respectively (alpha value 0.01). We assume the large majority of interactions are not significant looping interactions and therefore we interpret this weighted average as the expected 5C signal for a given genomic distance. Observed 5C signals are then transformed into a Z-score by calculating the (observed-expected/standard deviation) as described earlier, where the observed value is the detected 5C signal for a specific interaction, expected is the calculated weighted average of 5C signals for a specific genomic distance and standard deviation is the calculated weighted standard deviation of 5C signals for a specific genomic distance. Once the Z-scores have been calculated, their distribution is fit to a Weibull distribution. P-values are then mapped to each Z-score and then also transformed into q-values for FDR analysis. The ‘qvalue’ package from R was used to compute the q-values for the given set of p-values determined from the fit to the Weibull distribution. We used a stringent FDR threshold of 0.001%. We detected all known looping interactions in the *CFTR* locus in the appropriate cell lines, consistent with previous 3C studies (Gheldof et al., 2010; Ott et al., 2009). We called peaks in each 5C biological replicate separately and then took only the peaks that intersect across replicates as our final list of significant looping interactions. Using this cutoff, we

saw a comparable number of peaks as in previous 5C studies (Sanyal et al., 2012).

qRT-PCR

Gene expression levels were determined with qRT-PCR. Three technical replicates and three biological replicates were performed for each cell line. Gene expression levels were analyzed using a StepOnePlus instrument (Applied Biosystems) with the Power SYBR Green RNA-to-Ct 1-step kit (Life Technologies). Results were normalized to HPRT as an internal control. Any results with a Ct value higher than 34 were considered “not expressed.” RT-PCR primers were designed in neighboring exons using the Primer3 tool (http://fokker.wi.mit.edu/primer3/input.htm). Primers were tested for effectiveness by checking their titration ability and whether they gave a single melt curve. Primers used for this experiment can be found in Appendix I.

Results

Generation of 5C Chromatin Interaction Profiles in 5 Cell Lines

To determine the relationship between the location of TADs and the presence of chromatin looping interactions between genes and regulatory elements we applied chromosome conformation capture carbon copy, or 5C (Dostie et al., 2006). 5C is particularly suited for such analysis as it allows simultaneous high-resolution (single restriction fragment) detection of looping interactions (Sanyal et al., 2012) and analysis of large chromosomal domains to identify TADs (Nora et al., 2012). We applied 5C to analyze the conformation of a 2.8 Mb domain on human chromosome 7 (Figure 4.1A). This region was chosen because it is centered on the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene where we and others previously identified several cell-type specific looping interactions between the *CFTR* promoter and distal enhancers and CTCF-bound elements (Gheldof 2010, Ott 2009). This domain contains the ENCODE manual region ENm001 (Sanyal et al., 2012), is overall quite rich in gene content but also contains a large 700 kb gene-poor domain. Additionally, it contains several TADs as defined by a lower-resolution Hi-C analysis (Dixon 2012).

We used a 5C probe set (Figure 4.1A) based on a previously published design (Sanyal et al., 2012) which placed reverse probes on gene promoters and forward probes on the remaining restriction fragments in the region. This design allows analysis of long-range interactions between gene promoters and

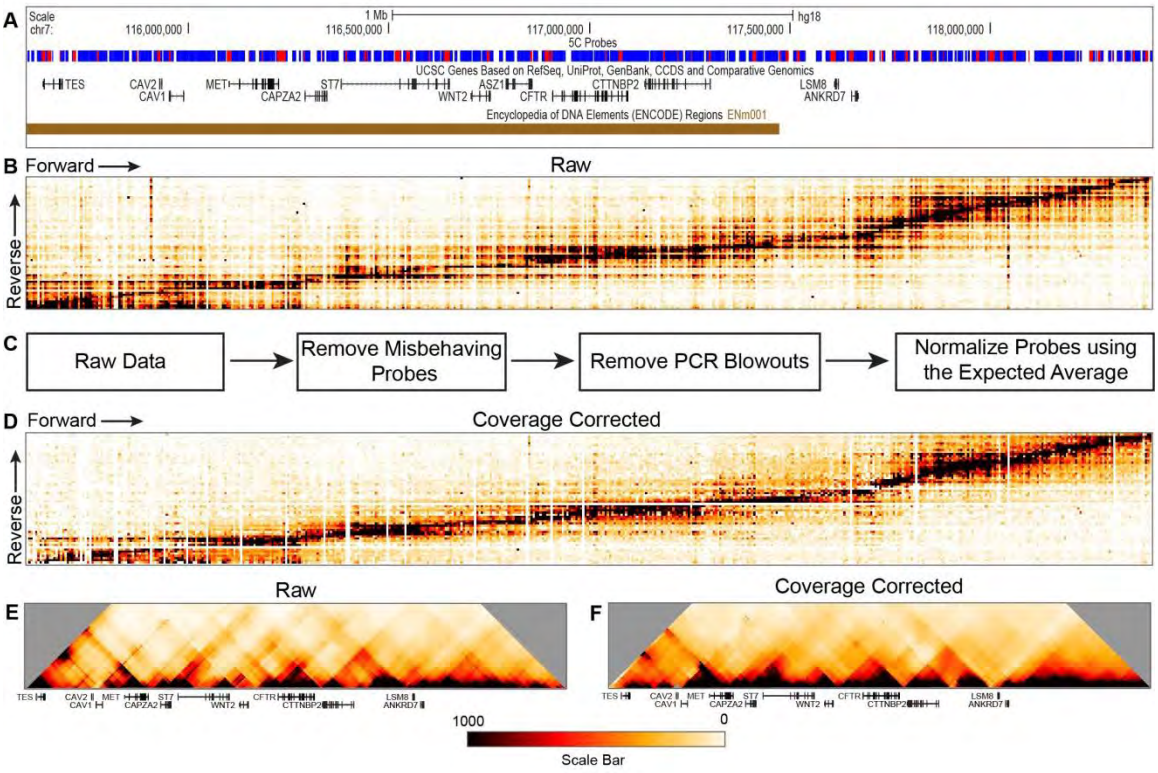


Figure 4.1: Walk-through of the Modifications Made to the 5C Data when it is Run Through our Pipeline. A) 5C region showing probe design. Red fragments = reverse probes. Blue fragments = forward probes. **B**) Raw 5C data plotted as all reverse primers vs. all forward primers. Color scale: white to orange to red to black, where white represents no interaction frequency and black represents the highest interaction frequency. It is apparent that some probes are misbehaving in the raw data (stripes on the heatmap). **C**) Steps in our data correction pipeline. Misbehaving probes are removed first (cis-purge), followed by removal of PCR blowouts (singleton removal). The final step normalizes all probes to each other (coverage correction). For full details see Methods. **D**) 5C data after coverage correction. White stripes running through the heatmap indicate misbehaving primers that were removed in the correction steps. **E**) Binned heatmap of raw data, before running it through the pipeline. **F**) Binned final 5C data after it has been run through the pipeline. Data is binned in 100 kb windows with a 10x step, meaning each square represents 10 kb.

surrounding chromatin, e.g. distal enhancers. We expanded the published probe set for ENm001 (Sanyal et al., 2012) to include additional forward and reverse probes on previously unused fragments and to incorporate the additional 700 kb gene-poor region. The final design contains 74 reverse probes and 605 forward probes for a possible 44,770 interrogated interactions.

We selected a panel of cell types to study, including three cell lines known to express the *CFTR* gene (Caco2, Calu3 and Capan1). These cell lines are derived from different locations in the body: Caco2 are colon-derived colorectal adenocarcinoma cells; Calu3 are lung-derived adenocarcinoma cells; Capan1 are pancreas-derived adenocarcinoma cells. It is likely that distinct cell type-specific enhancers may be involved in the expression of *CFTR* in these diverse tissues. We also included two cell lines that do not express the gene: the lymphoblastoid cell line GM12878 and the liver-derived hepatocellular carcinoma cell line HepG2 .

5C data is displayed as a heatmap, where an increase in color intensity corresponds with an increase in interaction frequency between two regions of the genome. Figure 4.1B shows raw data obtained from GM12878 cells. Reverse probes are plotted as rows and forward probes as columns. Each intersection between a reverse and forward probe represents a measured interaction frequency between two genomic loci. As expected, neighboring genomic regions interact with each other quite frequently, creating a black “diagonal” through the

middle of the heatmap. As the distance between fragments increases, the interaction frequency predictably decreases (Dekker, 2006; Dekker et al., 2002).

When examining the raw data (Figure 4.1B), it is clear that some probes over- or under-perform in the assay, creating horizontal and vertical stripes across the heatmap (dark stripes for over-performing probes and white stripes for under-performing probes). We removed data obtained with probes that performed aberrantly (either too high or too low, see Figure 4.1C, Methods). In addition, some probe pairs detect specific interactions that produce extremely strong signals (orders of magnitude greater than their neighbors), which in our experience are probably due to PCR amplification steps in the 5C protocol. We removed individual interactions when deemed an outlier (Figure 4.1C, Methods). Finally, we corrected the data for any remaining minor variations in probe efficiency, as we did previously (Sanyal et al., 2012), but with some modifications (see Methods). The corrected data is displayed in Figure 4.1D. Figure 4.1E and F show the same data as Figure 4.1B and D, but here the heatmaps are binned, displaying the 5C region versus itself (all heatmaps in this paper are binned in 100 kb bins, 10x step, unless otherwise noted). Pearson correlation analysis showed that replicates of the same cell line are highly correlated and also tend to be more correlated with each other than with replicates from different cell lines (Table 4.3).

Identification of Topologically Associating Domains (TADs)

Topologically Associating Domains (TADs) are consecutive regions hundreds of kb in size where loci located within the TAD associate and mix more frequently with each other than with loci located in adjacent TADs (Dixon et al., 2012; Nora et al., 2012). Visual inspection of the binned 5C interaction maps indicates that TAD structures are readily detected as triangles of strong self-association along the region (Figures 4.1E, 4.1F and 4.2). We employed a straightforward approach to quantify the pattern of TAD signals along the locus. The approach is based on the observation that TAD boundaries represent loci across which only very few long-range chromatin associations occur, i.e. between loci located upstream and downstream of the boundary. We quantified the relative frequency of interactions occurring across each bin throughout the 2.8 Mb region. We refer to this number as the insulation score of a genomic location. We then plotted these scores along the region to obtain an "insulation profile" for each cell line (Figure 4.2A-E, Figure 4.3, see Methods). Minima in the insulation profile represent TAD boundaries since these locations are sites where interactions across them occur at a low frequency. Insulation profiles and the locations of TAD boundaries were not dependent on the size of the window used for calculation of the insulation score (Figure 4.3). Figure 4.2 shows binned 5C interaction maps for all cell lines with their insulation profiles plotted below. Strikingly, although the amplitude of the signal varies, the insulation profiles of all 5 cell lines are overall very similar, with TAD boundaries present at

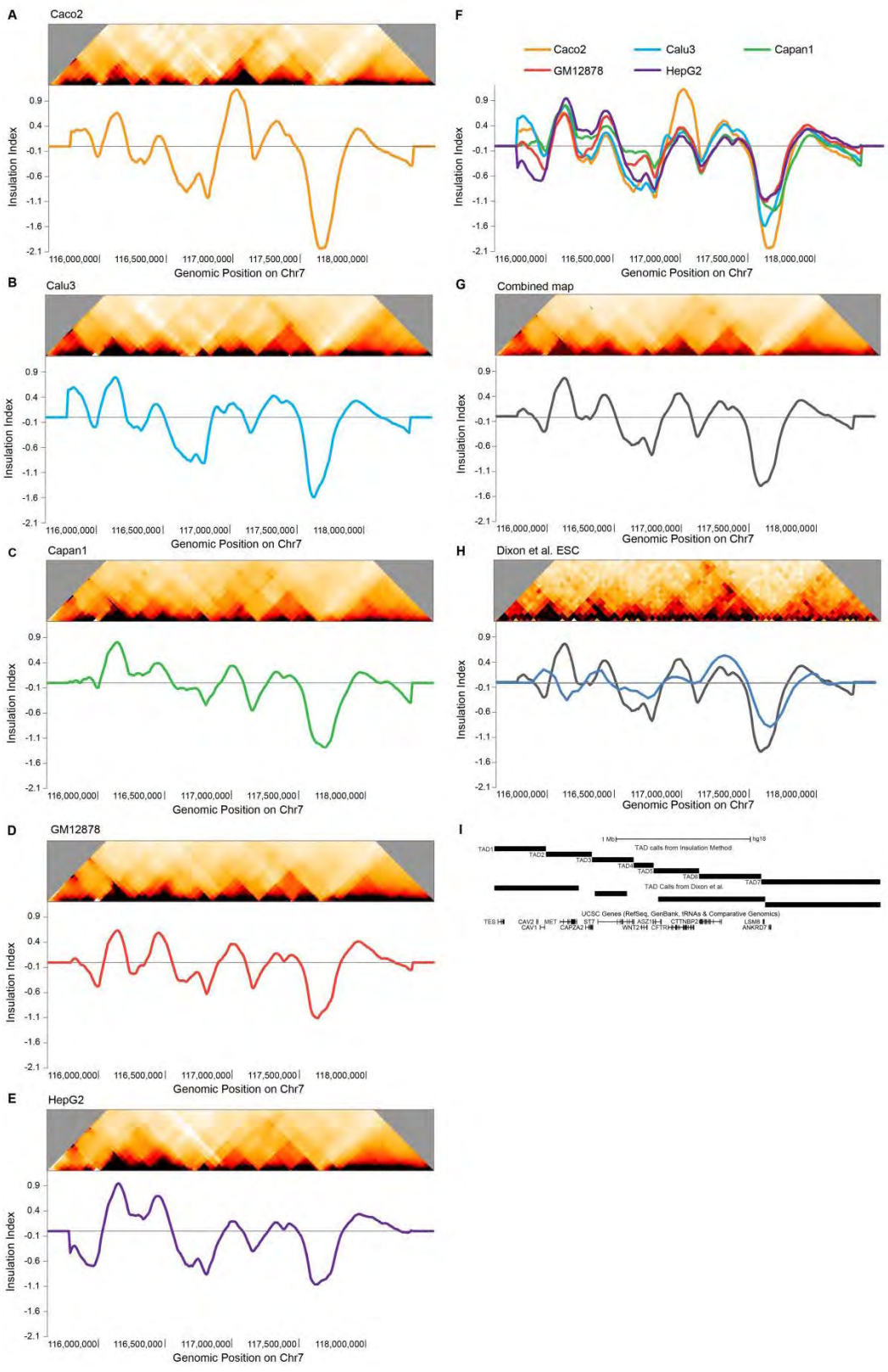
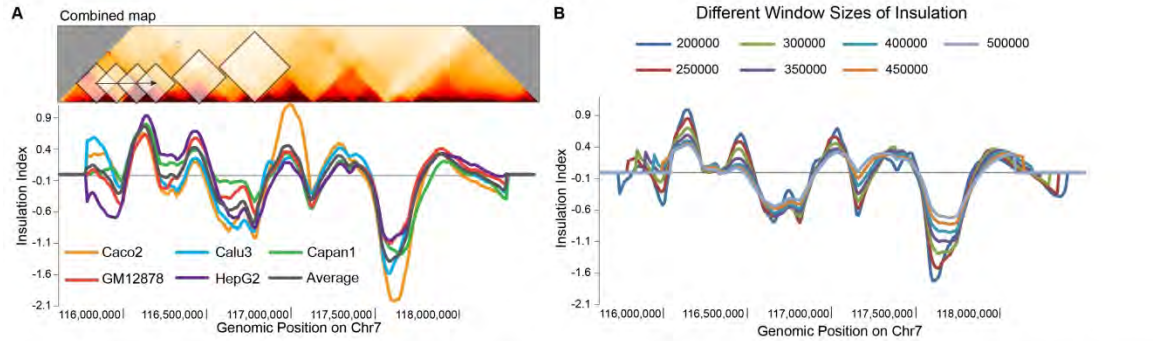

A
Caco2
B
Calu3
C
Capan1
D
GM12878
E
HepG2
F
Caco2
Calu3
Capan1
GM12878
HepG2
G
Combined map
H
Dixon et al. ESC
I
Insulation Index
0.9
0.4
-0.1
-0.6
-1.1
-1.6
-2.1
116,000,000
116,500,000
117,000,000
117,500,000
118,000,000
Genomic Position on Chr7
1 Mb
hg18
TAD calls from Insulation Method
TAD1
TAD2
TAD3
TAD4
TAD5
TAD6
TAD7
TAD Calls from Dixon et al.
UCSC Genes (RefSeq, GenBank, tRNAs & Comparative Genomics)
TES
CAV2
CAV1
MET
STT
CAPZA2
WNT2
CFTR
CTTNBP2
LSM6
ANKRD7

Figure 4.2: Calling TADs on the 5C data using an Insulation Index. A-E) Heatmaps and corresponding insulation indexes for each cell line: Caco2, Calu3, Capan1, GM12878 and HepG2. Dips on the graph represent boundaries between two TADs. **F**) The insulation indexes for all 5 cell lines plotted in one graph. **G**) The heatmap is a combined map of all our data. Below is the average insulation index of all the data. **H**) Our insulation index method run on human embryonic stem cell Hi-C data from Dixon et al., 2012. This heatmap is binned as in the Dixon et al. paper with 40 kb bins. Its color scale is from 0 to 35. **I**) The TAD calls in the 5C region based on our insulation index compared with the TAD calls made with the method in Dixon et al., 2012.

the same locations (Figure 4.2F, Figure 4.3). Further, we compared our 5C data to previously published Hi-C data in human embryonic stem cell lines (ESC) (Dixon et al., 2012) (Figure 4.2H). The overall TAD organization in ESCs as detected by Hi-C and quantified using our insulation score approach is again very similar to the organization we detected by 5C (Figure 4.2H). We note that there are some differences (e.g. the region around position 116,250,000). These could represent real differences in TAD boundary positions, or experiment dependent variations in interaction patterns, e.g. due to the lower resolution of the Hi-C data.

These analyses show that the overall TAD organization is conserved across 5 differentiated human cell lines. Some differences in TAD positions may be present between ESCs and differentiated cells. These results confirm and extend earlier observations that TADs are similar, but not always identical, in different cell types (Dixon et al. 2012, Nora et al. 2012). Since the insulation profiles are extremely similar between cell lines (Figure 4.2F, Figure 4.3), we created a consensus insulation profile by calculating the average profile across all 5 differentiated cell lines to use for downstream analysis (Figure 4.2G). 6 TAD boundaries are present in the regions, defining 7 TADs.

Previous studies have shown that, in general, TAD boundaries are enriched in gene promoters and CTCF sites (Dixon et al., 2012). We examined our TAD boundary regions to identify what chromatin features are present. We



C

Pearson: Insulation	Caco2	Calu3	Capan1	GM12878	HepG2	Average
Caco2	1	0.912	0.798	0.844	0.568	0.933
Calu3		1	0.819	0.866	0.745	0.944
Capan1			1	0.876	0.811	0.926
GM12878				1	0.890	0.958
HepG2					1	0.876
Average						1

D

Pearson: Windows	200000	250000	300000	350000	400000	450000	500000
200000	1	0.982	0.953	0.925	0.896	0.877	0.865
250000		1	0.989	0.968	0.942	0.922	0.903
300000			1	0.991	0.972	0.952	0.930
350000				1	0.992	0.976	0.955
400000					1	0.993	0.978
450000						1	0.995
500000							1

Figure 4.3: Insulation Index of All Cell Lines and Different Window Search Spaces. A) The insulation index method is shown in the diamonds slid along the the heatmap. The sum of each diamond is plotted as a point in the insulation index. Below the heatmap is plotted the insulation index for all cell lines and the average, as in Figure 4.2. **B**) Pearson correlations between the insulation indexes plotted above. **C**) Insulation indexes run on the combined heatmap using different window sizes. As the window size increases the index smooths out but maintains the same peaks and dips. **D**) Pearson correlation of the different window sizes.

noticed that only a few of our TAD boundaries were close to a gene promoter. Indeed, the boundary between TADs 1 and 2 is 9291 bp from the 3´ end of the CAV1 gene. The closest promoter is 45 kb away. Additionally, the boundary between TADs 2 and 3 occurs 8812 bp from the end of the CAPZA2 gene, and the boundary between TADs 3 and 4 occurs 3440 bp from the end of the ST7 gene. Also the boundary between TADs 4 and 5 occurs within the ASZ1 gene, 7245 bp from its termination site but 57 kb from its promoter. The boundary between TADs 5 and 6 is very close to the end of the CTTNBP2 gene (185 bp). The boundary between TADs 6 and 7 is the only one that is closer to a gene promoter than a gene end, located 13.5 kb from the LSM8 promoter. Although there is a global enrichment of TAD boundaries occurring near gene promoters, we conclude that this is not the case for every TAD boundary, suggesting that gene promoters do not determine TAD boundaries. We do find that most boundaries are located very close to CTCF sites, and that many of these sites are at the 3´ end of genes. However, CTCF sites are found within TADs as well, indicating that CTCF is not sufficient for boundary formation, consistent with previous reports (Bortle et al., 2014; Dixon et al., 2012; Nora et al., 2012; Zuin et al., 2014).

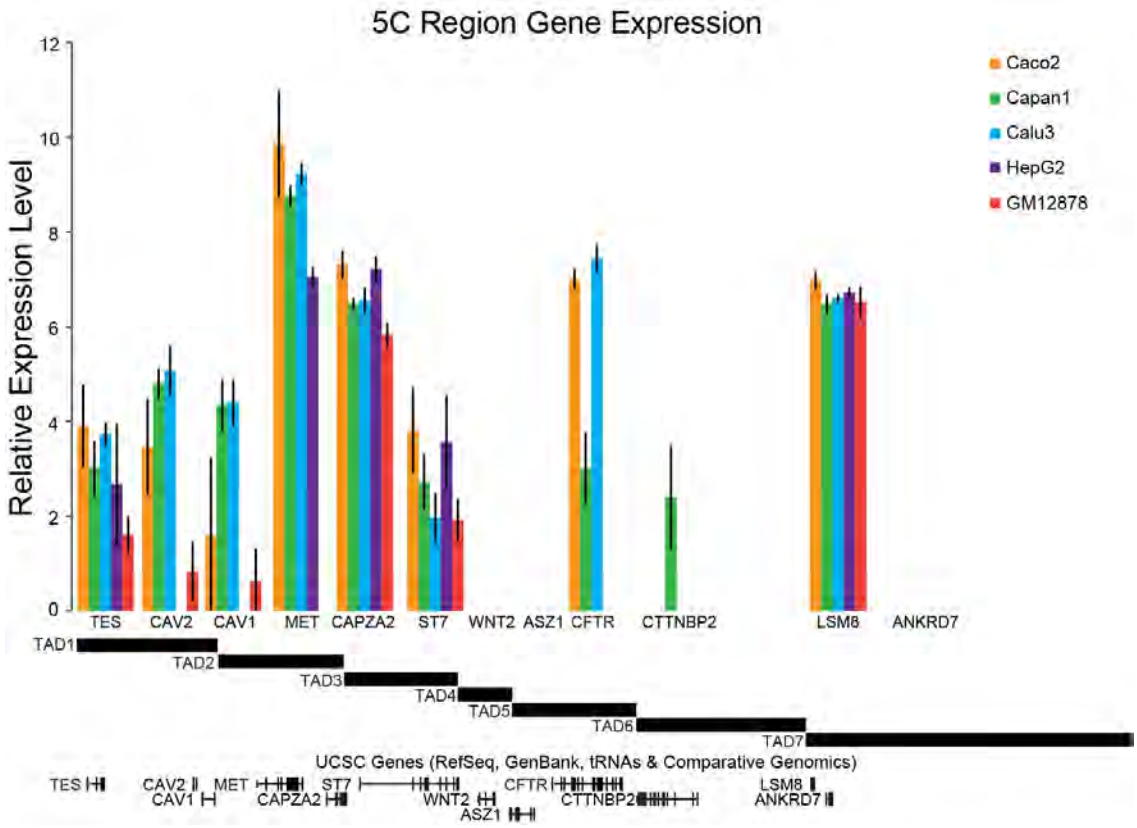


Figure 4.4: Gene Expression Within the 5C Region is Not Related to TAD Structure. Relative gene expression is plotted for the 5 cell types studied. The TAD structure of the locus is shown below the graph for reference. Genes promoters with different expression patterns can be located within the same TAD. Primers can be found in Appendix I.

TAD positions are not affected by cell-type specific gene expression

To investigate the relationship between TAD boundaries and gene expression we measured the expression level of all genes in the 2.8 Mb regions in the 5 cell lines (Figure 4.4). Interestingly, we find that several TADs are entirely transcriptionally silent in some cell lines, while displaying transcription of at least one gene in others. Yet, TAD boundaries are the same whether or not the TAD contains an expressed gene. This is particularly well illustrated in GM12878 and HepG2 cells where TADs 4, 5 and 6 are transcriptionally inactive, but the same set of TAD boundaries that separate them are present in cells that do express genes located in these TADs. On the other hand, TADs 1 and 2 have at least one gene active in each cell line but the precise set of genes that is active differs between different cell lines, and TAD boundaries are invariant as well. Together, these observations indicate that TAD boundaries occur irrespective of gene transcription and are not determined by the expression status of genes located within TADs.

Identification of long-range looping interactions

Next we set out to identify specific and statistically significant long-range interactions such as promoter-enhancer contacts throughout the 2.8 Mb domain. Previously, we developed and applied a statistical methodology to identify pair-

wise interactions in a 5C dataset that occur significantly more frequently than expected (Phillips-Cremins et al., 2013; Sanyal et al., 2012). The approach first calculates the expected baseline contact frequency of pairs of loci dependent on their genomic site separation. Then, individual interactions between pairs of loci are identified that are significantly above this baseline (referred to as looping interactions). Here we further refined this approach by taking into consideration the presence of TADs. Specifically, we calculated the expected baseline interactions separately for interactions occurring within and between TADs. For comparison we also performed our peak-calling on the entire dataset ignoring TADs, as we did previously (Figure 4.5F, Methods). Using a FDR of 0.001% the two different approaches give similar but not identical sets of statistically significant looping interactions, indicating that the presence of TADs has some impact on our ability to detect significant signals (Figure 4.5). We use the set of significant interactions detected when explicitly taking TADs in consideration, and observed in two independent biological replicates, for our analyses.

Loci located within a TAD interact more frequently than pairs of loci located in different TADs (Dixon et al., 2012; Hou et al., 2012; Jin et al., 2013; Nora et al., 2012). This can be visualized for our 5C data set by plotting all intra-TAD and inter-TAD interactions as a function of their genomic site separation (Figure 4.6). This plot shows that at a given genomic distance, intra-TADs interactions occur more frequently than inter-TAD interactions. As a result the interaction frequency of statistically significant looping interactions within TADs

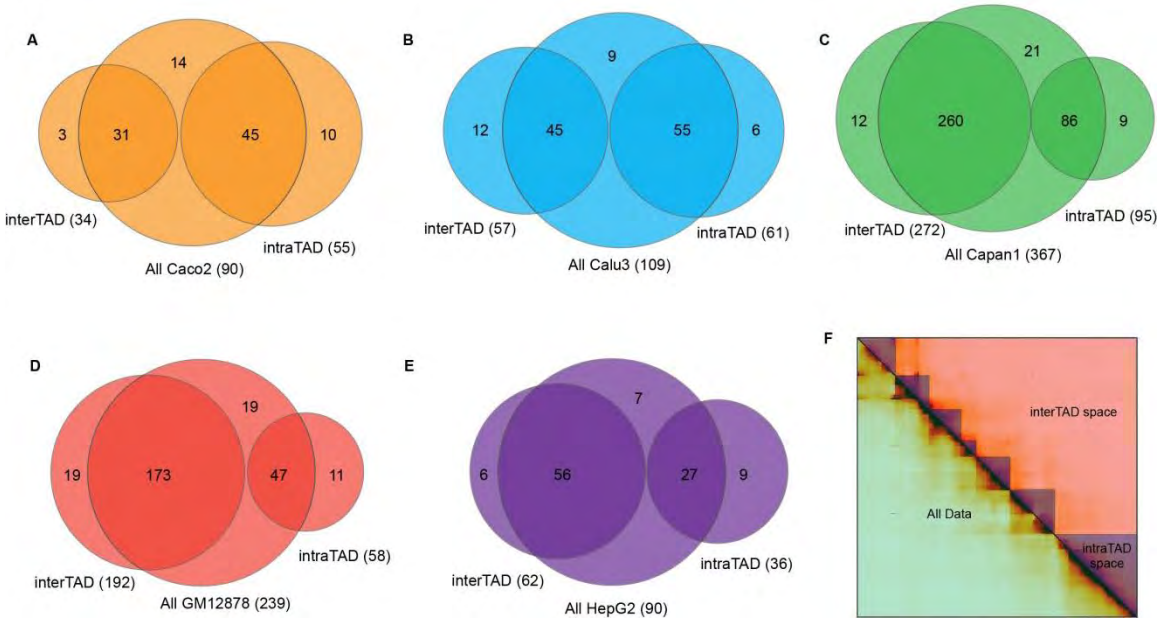


Figure 4.5: Peak Calling Overlap of the Entire Data versus IntraTAD and InterTAD spaces. A) Peak calling on Caco2 cells. **B**) Peak calling on Calu3 cells. **C**) Peak calling on Capan1 cells. **D**) Peak calling on GM12878 cells. **E**) Peak calling on HepG2 cells. **F**) Heatmap displaying the divisions between the peak calling space. The green section of this map indicates peaks called on the entire data without taking into account TAD structure. Alternately, we called peaks on only intraTAD structure (blue triangles) and interTAD structure (red space).

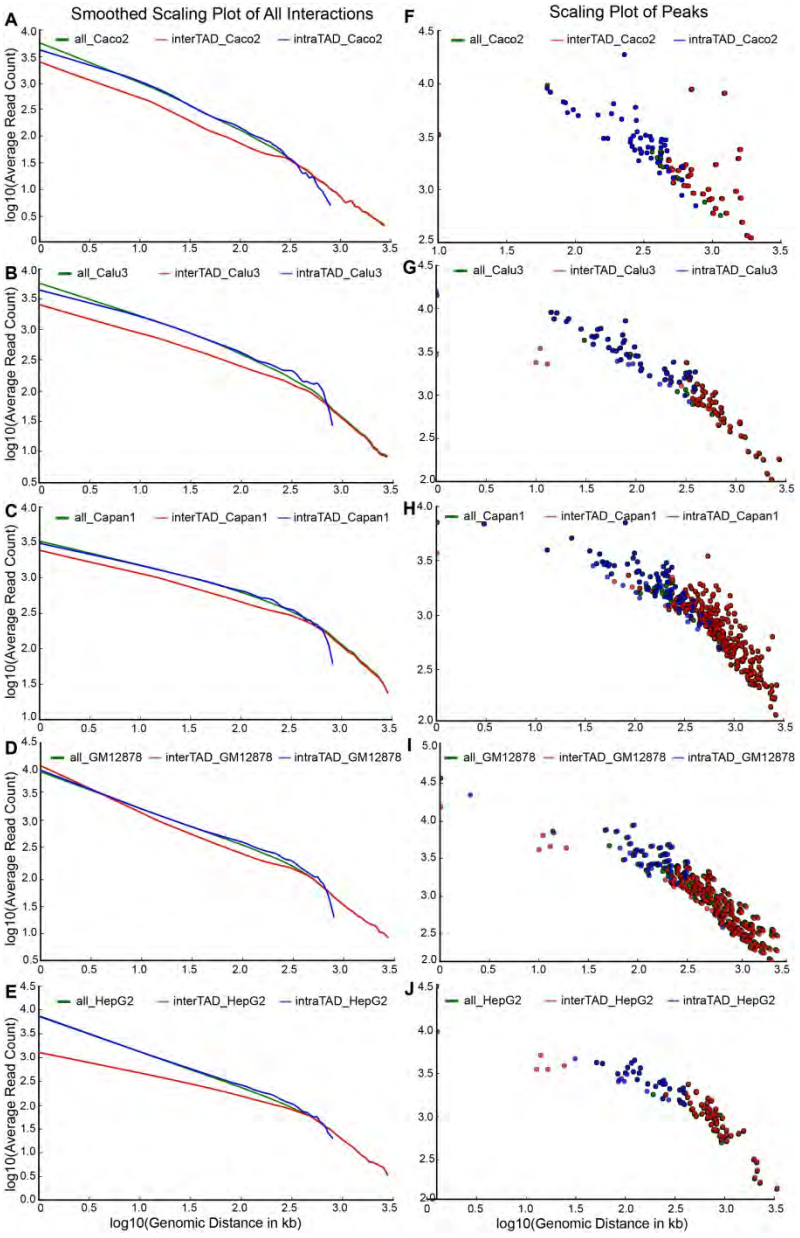


Figure 4.6: Scaling Plots for All Cell Lines - Interactions and Peaks. A-E) Smoothed (10%) read count for all interactions. **F-J**) Non-smoothed read count for all significant interactions called as peaks. Green: Interactions/Peaks from the entire dataset. Red: Interactions/Peaks from the interTAD space. Blue: Interactions/Peaks from the intraTAD space.

is also much higher than the interaction frequency of inter-TAD looping interactions, even when they involve loci separated by the same genome distance (Figure 4.6F-J). Generally, the frequency of statistically significant looping interactions between TADs is much lower than those that occur within TADs because they involve loci that are separated by much larger genomic distances. This implies that intra-TAD interactions occur in a larger fraction of cells, whereas inter-TAD interactions are considerably less frequent in the cell population. Finally, the relative number of statistically significant intra-TAD looping interactions (305 out of 34372 interrogated interactions (0.88%)) detected is much higher than the relative number of significant inter-TAD looping interactions (617 out of 112723 interrogated interactions (0.54%)). This is consistent with previous reports which showed that looping interactions between gene regulatory elements occurs more frequently within TADs (Dixon et al., 2012; Jin et al., 2013). We conclude that intra-TAD looping interactions display higher contact frequencies, even when corrected for genomic distance, implying these interactions occur in a larger fraction of cells and occur more often than inter-TAD looping interactions.

When we compared looping interactions among the 5 cell lines we find that the majority of interactions are cell type-specific with a minority of interactions observed in 2 or more cell lines, and with only a handful interactions observed in all 5 cell lines. This holds for interactions within TADs and for interactions that occur between TADs (Figure 4.7A, B). This is consistent with

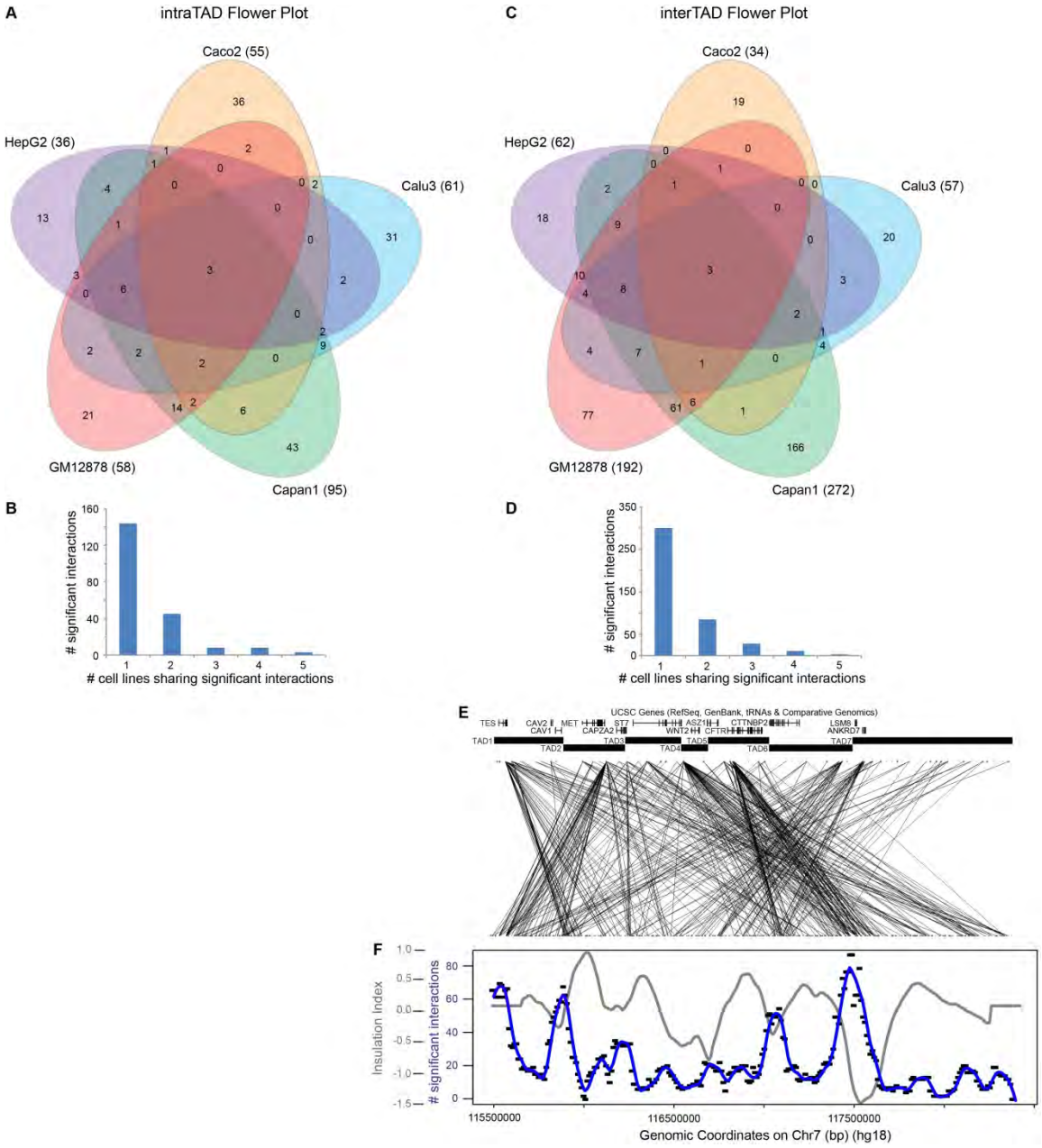


Figure 4.7: Interactions Within or Between TADs are Mostly Cell-Type Specific. A) Flower plot showing all intraTAD interactions. **B**) The number of cell lines that share significant intraTAD interactions. **C**) Flower plot showing all interTAD interactions. **D**) The number of cell lines that share significant interTAD interactions. **E**) Webplot displaying all interTAD interactions, with TADs and gene locations across the top. **F**) The blue plot shows interTAD interactions, while the grey plot shows the insulation index as in Figure 4.2.

other 5C analysis (Phillips-Cremins et al., 2013; Sanyal et al., 2012). Thus, looping interactions are highly tissue-specific, while TAD organization is largely cell type invariant.

A Single TAD Constrains All Known *CFTR* Promoter-Enhancer Loops

We focused our analysis on long-range interactions of the *CFTR* gene. Previous studies have identified a number of putative regulatory elements within and flanking the *CFTR* gene. These were identified as DNase 1 hypersensitive sites (DHS), often representing nucleosome free DNA sequences that can interact with transcription factors. Functional characterization using reporter assays showed that several of these DHSs are putative cell-type specific enhancers (reviewed in McCarthy and Harris, 2005; Ott and Harris, 2011) (Gheldof et al., 2010; Ott et al., 2009; Zhang et al., 2013, chapter 5 of this thesis). These elements are spread out over several hundred kb making it difficult to predict what gene(s) they regulate. 3C studies revealed that several of these elements directly loop to the *CFTR* promoter and to each other specifically in cells that express *CFTR*, but not (or much less frequently) in cells that do not express *CFTR* (Gheldof et al., 2010; Ott et al., 2009). The *CFTR* promoter was found to interact with four elements, which we call elements I, II, III and IV. Elements I and II are CTCF-bound sites located ~21 and ~80 kb upstream of the gene in Caco2 cells (Gheldof et al., 2010, chapter 3 of this thesis). Elements III

and IV, located in intron 11 and +202 kb downstream from the TSS, act as enhancer elements when tested in a luciferase assay and contain chromatin marks typically associated with enhancer activity (Gheldof et al., 2010; Ott et al., 2009, chapter 3 of this thesis). These studies strongly suggest that these elements regulate the *CFTR* gene.

The 5C analysis presented here reproduced most of the previously identified looping interactions between the *CFTR* promoter and distal regulatory elements. First, in all three *CFTR* expressing cells (Caco2, Capan1, Calu3) we detected significant interaction frequencies between the *CFTR* promoter and the known *CFTR* enhancer located within intron 11, element III (108 Kb downstream) (Figure 4.8G-K). In cells that do not express *CFTR* this interaction is either strongly reduced (GM12878), or not significant (HepG2). Furthermore, we find that the *CFTR* promoter engages in additional cell type-specific looping interactions. First, in Caco2 cells the *CFTR* promoter interacts with a region just downstream of the gene (202 kb downstream of the promoter) that contains a known enhancer, consistent with previous 3C studies performed with this cell line (Gheldof et al,. 2010; Ott et al., 2009, chapter 3 of this thesis). This interaction is not observed in Calu3 and Capan1 cells, or in the non-expressing HepG2 cells, but it is significant in non-expressing GM12878 cells (Figure 4.8G-K) (see discussion). Second, in Calu3 and Caco2 cells, but not in Capan1 cells, the *CFTR* promoter also engages in looping interactions with sites upstream of the promoter (Figure 4.8G-I). In both Caco2 and Calu3 cell lines the promoter loops

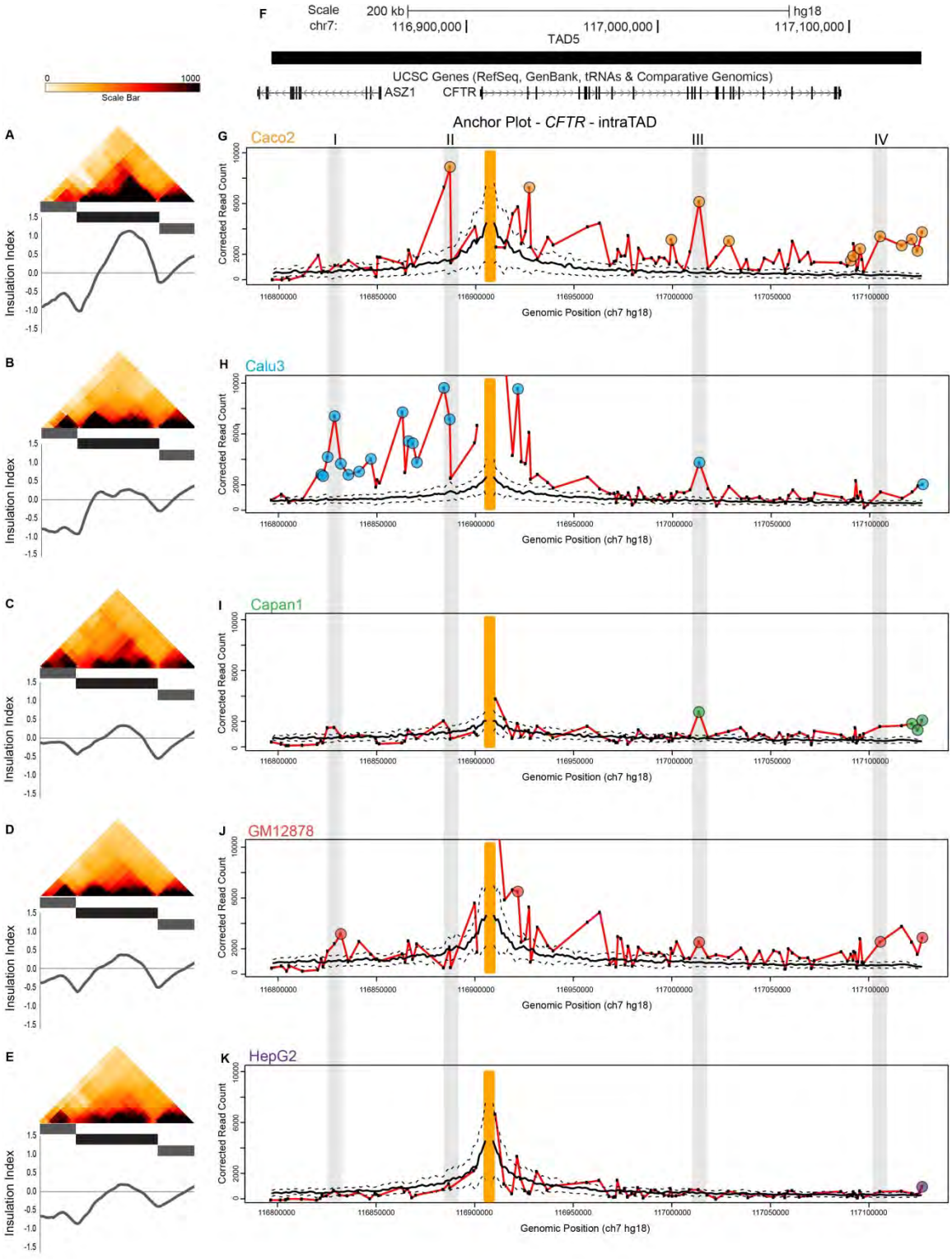

F
Scale
chr7:
200 kb
hg18
116,900,000
117,000,000
117,100,000
TAD5
0
1000
Scale Bar
UCSC Genes (RefSeq, GenBank, tRNAs & Comparative Genomics)
ASZ1
CFTR
Anchor Plot - CFTR - intraTAD
A
B
C
D
E
G Caco2
H Calu3
I Capan1
J GM12878
K HepG2
I
II
III
IV
Insulation Index
Corrected Read Count
Genomic Position (ch7 hg18)
116800000
116850000
116900000
116950000
117000000
117050000
117100000

Figure 4.8: 3C-style Plots of intraTAD Looping Interactions of the *CFTR* Promoter. A-E) Heatmaps showing TAD 5, the TAD that houses the *CFTR* gene. Insulation plots of the region are below, with dips in the plots indicating TAD boundaries. **F**) Image from the UCSC Genome Browser depicting the genes inside TAD 5. **G-K**) 3C-style plots for probe REV_421 located at the *CFTR* promoter. Grey lines indicate elements I-IV identified by our 3C experiments (Gheldof et al. 2010, chapter 3 of this thesis). Colored balls depict interactions that are significantly higher than background (peaks).

to a site located ~21 kb upstream that binds the CTCF protein, consistent with 3C experiments (Figure 4.8G and H) (Gheldof et al., 2010; Ott et al., 2009, chapter 3 of this thesis). In addition, in Calu3 cells the promoter loops to several additional sites located further upstream. We note that the entire region up to around 100 kb upstream is interacting frequently, and statistically significantly, with the *CFTR* promoter. However, three peaks in the interaction profile stand out as the highest peaks in the region. The other weaker, but significant, interactions probably represent indirect interactions that are brought in relative close proximity of the *CFTR* promoter as a result of the other prominent looping interactions present in the region (see discussion). One prominent peak occurs at and around a site located ~80 kb upstream that corresponds to a CTCF-bound element. We previously found this site to weakly interact with the *CFTR* promoter in Caco2 cells (Gheldof et al., 2010, chapter 3 of this thesis) but our current 5C analysis detects this interaction in Calu3 cells, not Caco2. Interestingly, another prominent looping interaction that is specific for lung-derived Calu3 cells involves a site located ~35 kb upstream of the promoter. This region contains a previously identified lung-specific enhancer (Zhang et al., 2013). From these analyses we conclude that the *CFTR* promoter engages in several long-range looping interactions with enhancers that are active in the corresponding cell line. The promoter also interacts with several distal CTCF-bound elements. The role of cell-type specific interactions with otherwise tissue invariant CTCF sites is currently not known (see discussion). Interestingly, we note that in all cell lines

studied the CFTR promoter is interacting with the right boundary of its TAD. As described below, TAD boundaries often engage in long-range interactions, including with elements located in other TADs. The relevance of these interactions is currently not known. Finally, we note that the *ASZ1* gene, located in the same TAD as *CFTR,* is inactive in all cell lines and does not engage in any significant long-range looping interactions (Figure 4.9). This further shows that the distal elements that interact with the *CFTR* promoter do not interact with other (inactive) promoters in the domain.

Of additional interest is the impact of these diverse intraTAD interactions on the boundaries of TAD5. Figure 4.9A-E show a zoomed in view of TAD5, with the insulation score of this zoomed-in region plotted below. TAD5 is indicated as the black box, while the neighboring TADs 4 and 6 are indicated by the grey boxes. The numerous significant interactions present inside the *CFTR* gene in Caco2 cells (Figure 4.9M) are indicated in the 5C heatmap as a dark triangle located to the right in TAD5 (Figure 4.9A). The numerous significant interactions present between the *CFTR* promoter and the upstream region in Calu3 cells are indicated by the black triangle located in the left of TAD5 (Figure 4.9B). The relatively flat 3C profiles in Capan1, GM12878 and HepG2 cells are represented by lack of intraTAD structure in their corresponding heatmaps (Figure 4.9C-E). Although the looping patterns within TAD5 differ between cell lines, the boundary regions of this TAD are clearly defined in all cell lines. This indicates that intraTAD looping does not affect TAD boundaries.

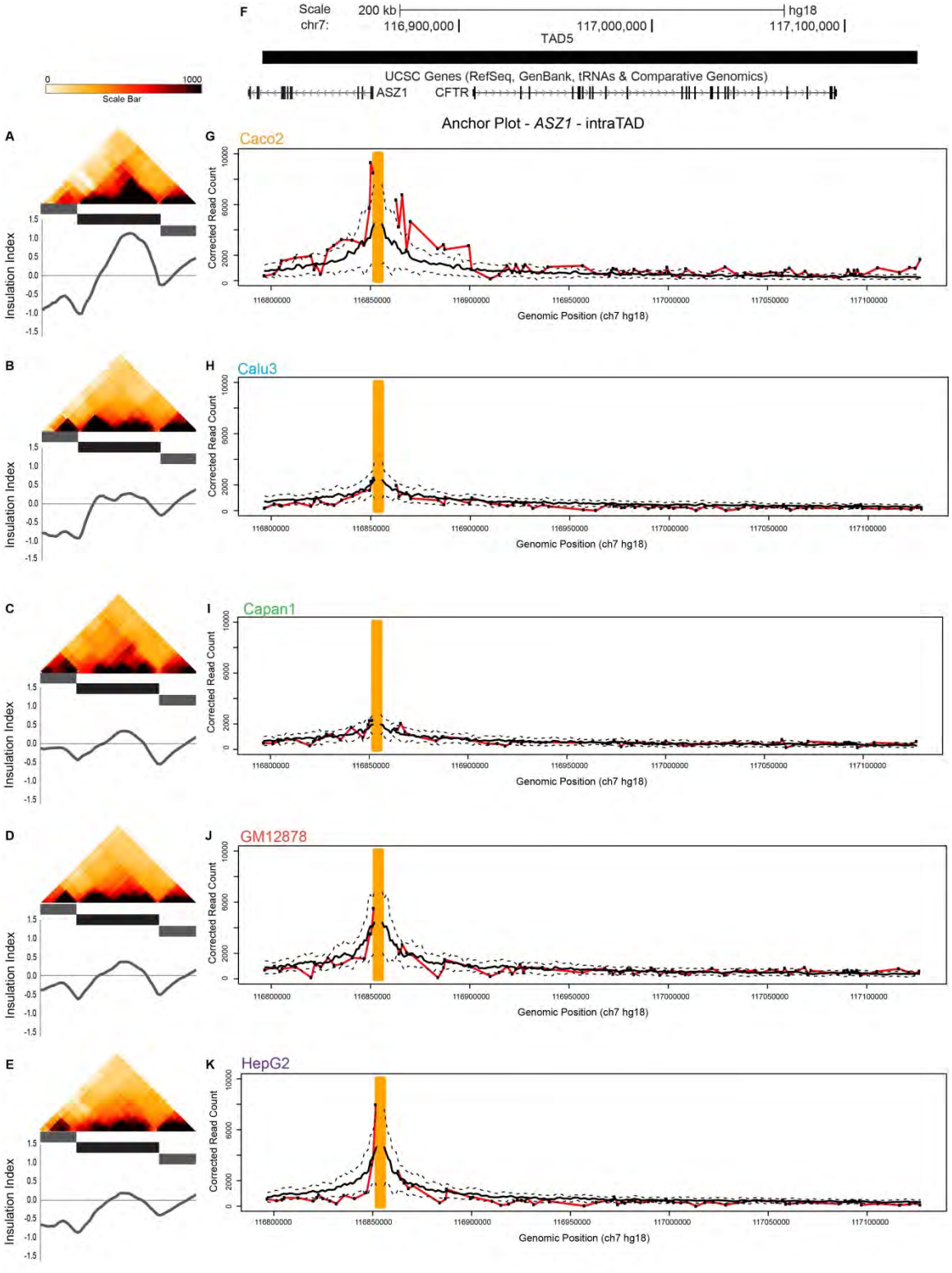

F
Scale
chr7:
200 kb
116,900,000
117,000,000
117,100,000
hg18
TAD5
0
1000
Scale Bar
UCSC Genes (RefSeq, GenBank, tRNAs & Comparative Genomics)
ASZ1
CFTR
Anchor Plot - ASZ1 - intraTAD
A
B
C
D
E
G
Caco2
H
Calu3
I
Capan1
J
GM12878
K
HepG2
Insulation Index
Corrected Read Count
Genomic Position (ch7 hg18)

Figure 4.9: 3C-style Plots of intraTAD Looping Interactions of the *ASZ1* Promoter. A-E) Heatmaps showing TAD 5, the TAD that houses the *CFTR* gene and some of the *ASZ1* gene. Insulation plots of the region are below, with dips in the plots indicating TAD boundaries. **F**) Image from the UCSC Genome Browser depicting the genes inside TAD 5. **G-K**) 3C-style plots for probe REV_404 located at the *ASZ1* promoter.

Long-range Looping Interactions Within Other TADs in the Region

We also detected statistically significant long-range looping interactions within the other TADs. We find that active gene promoters are more likely to be involved in long-range interactions than promoters that are not expressed, consistent with our earlier findings (Figure 4.10) (Sanyal et al., 2012). For instance the *CTTNBP2* gene is only expressed in Capan1 cells, and it is engaged in long-range looping interactions only in that cell type. There are exceptions to this trend. For example, the *CAV2* gene interacts with several elements throughout its TAD in Calu3 and Capan1 cells, which both express the gene. No looping interactions are detected other cell lines which express the gene, Caco2 and HepG2. This may be due to technical reasons (e.g. no 5C probe was included for a restriction fragment that contains Caco2 or HepG2-specific elements, or the interaction is simply missed due to false-negatives). Alternatively, in some cell types no distal looping interactions are required for gene activation. Based on previous analyses from our lab and those of others (Jin et al., 2013; Phillips-Cremins et al., 2013; Sanyal et al., 2012), it is most likely that the looping interactions in these TADs involve gene regulatory elements, e.g. enhancers, and architectural elements such as CTCF-bound elements.

Our analysis also identified significant long-range looping interactions between loci located in different TADs. As mentioned above, these interactions tend to be much longer range (more then several hundred kb), and much lower in

contact frequency (Figure 4.6). Interestingly, these interactions were highly cell type-specific, as were the interactions within TADs (Figure 4.7C, D). Strikingly, when we examined which loci engage in such inter-TAD interactions, we found a strong correlation with TAD boundaries (Figure 4.7E, F). Thus, while the positions of TAD boundaries are invariant, loci located near them engage in highly cell type-specific, but rather weak, long-range interactions with other TADs.

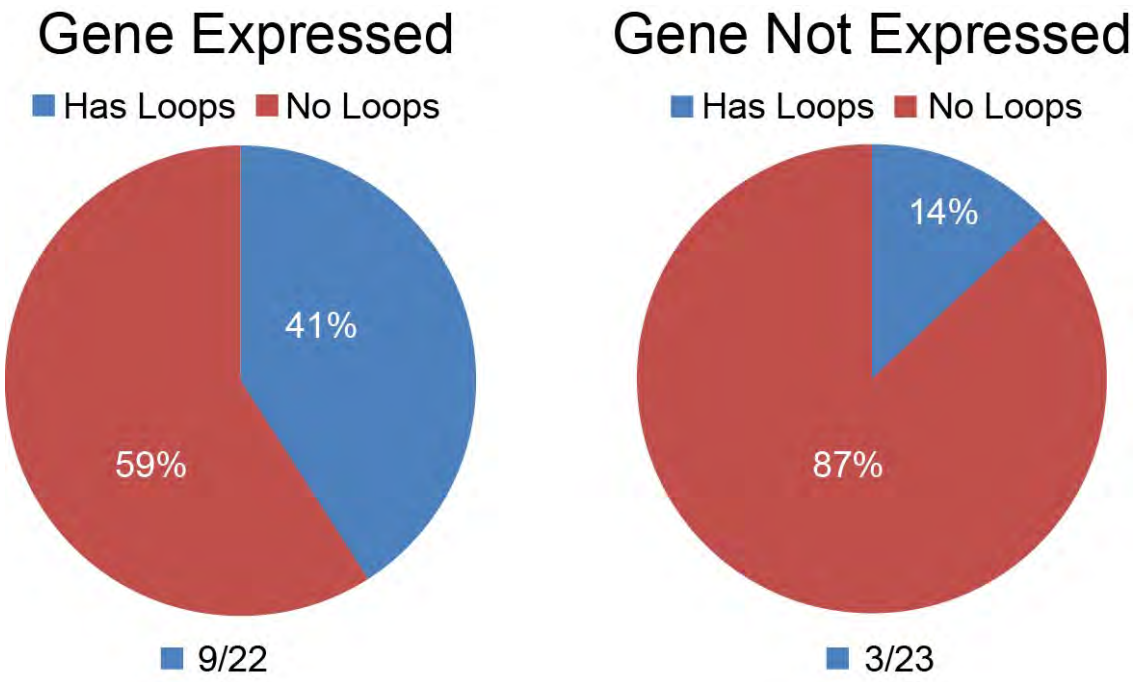


Figure 4.10: Genes that are Not Expressed are Unlikely to Have Looping Interactions. In the 2.8 Mb region we examined, genes that are not expressed are much less likely to have significant looping interactions between elements and their gene promoters. *This analysis does not include the *ST7* gene because we do not have a probe at that gene promoter.

Discussion

We present a study of a 2.8 Mb region on chromosome 7 of the human genome surrounding the cystic fibrosis locus. This study expanded our previous 3C study centered on the cystic fibrosis transmembrane conductance regulator gene (*CFTR*), which identified four DNA elements interacting with the *CFTR* promoter in cells that express the gene (Gheldof et al., 2010; chapter 3 of this thesis). We wanted to learn more about potential *CFTR* regulatory elements. Our 5C experiment included 5 human cell lines: a colorectal adenocarcinoma (Caco2), a lung adenocarcinoma (Calu3), a pancreatic adenocarcinoma (Capan1), a lymphoblastoid (GM12878) and liver hepatocyte (HepG2). Among these cells, Caco2, Calu3 and Capan1 express high levels of *CFTR*, while GM12878 and HepG2 do not express *CFTR*.

TADs Exist Regardless of Gene Expression Status or intraTAD Looping Elements

An obvious feature of our data set is the clearly defined topologically associating domains (TADs) present in all cell lines we studied. TADs have been proposed to be unchanging between cell types, and we provide more evidence that this is indeed the case (Figure 4.2). Importantly, we note that TADs are present independently of gene expression. Throughout the region there are 12 genes, some of which are differentially expressed between the cell lines in our investigation. For example, in HepG2 and GM12878 cells, a string of four genes

(*WNT2, ASZ1, CFTR* and *CTTNBP2*) spanning three TADs are not expressed. The division between TADs 4, 5 and 6 are still clearly demarcated in both these cell lines. Thus TADs are not maintained or determined by gene expression. We see that a majority (5/6) of our TAD boundaries are located very close to the 3´ end of genes, in contrast to earlier claims that TAD boundaries are marked by gene promoters (Dily et al., 2014; Dixon et al., 2012). However, we note that in a recent study Dily et al. find a genome-wide enrichment of gene promoters at TADs, yet they observed that in median gene promoters were located >200 kb from the boundaries (Dily et al., 2014). Therefore we caution against accepting that a genome-wide rule applies to every, or perhaps even a significant number of, TAD boundaries. We note that the specific loops found within TADs do not seem to affect their structure or location. We see that the majority of significant loops that occur in our data are unique to one cell line (Figure 4.7A, B), though TADs are invariant between cell lines. We see an example of this feature when we look within TAD5 (Figure 4.8), which shows that although looping interactions inside that TAD differ between the 5 cell lines, the TAD boundaries remain constant. Thus it appears that TADs are not defined by the interactions within them.

Actively Transcribed Genes Do Not Always Have Looping Interactions

We have shown that expressed genes have significantly more looping interactions than non-expressed genes (Figure 4.10). It is clearly not the case

that all expressed genes have looping interactions. An interesting observation is that some expressed genes have looping interactions in some cell lines but not in others. For example, *CAV2* shows looping to its promoter in Calu3 and Capan1 lines, but not in Caco2, GM12878 or HepG2 lines, even though it is expressed in all five lines. This is similar to the *MET* gene, which is expressed in all cell lines studied except GM12878, and shows looping interactions in all except Caco2 cells. Perhaps in the case of *MET* the probe was less efficient in Caco2 cells, and there may be looping interactions occurring that we did not observe. In the case of *CAV2* however, it is unclear why some cell lines display looping while others do not. One example of a gene not being expressed but containing a significant looping interaction is the *CFTR* promoter contacting the TAD boundary region in HepG2 cells (Figure 4.8K). This interaction seems likely to be a structural interaction, as it is also present in the other cell lines (Figure 4.8).

Interactions Between TADs

We detected statistically significant interactions between loci located in different TADs. These interactions are as cell type-specific as the looping interactions within TADs. However, these interactions also differ in at least two ways from looping interactions within TADs. First, looping interactions between TADs are much weaker (fewer reads) than looping interactions within TADs (Figure 4.6). This can mean that they occur in fewer cells in the population (Gibcus and Dekker, 2013). Second, they often involve loci located at or near

TAD boundaries (Figure 4.7E, F), whereas looping interactions within TADs can occur throughout these domains. The functional relevance of these interactions is currently not known.

Loci located near TAD boundaries often engage in looping interactions with loci located in different TADs (Figure 4.7E, F). These interactions tend to be much longer-range than interaction within TADs, and consequently are at least an order of magnitude weaker and probably more stochastic in the cell population (Figure 4.6) (Gibcus and Dekker, 2013). Interestingly, these interactions are highly cell type-specific. The roles, if any, of these interactions in gene regulation are currently not known. Possibly they play an architectural role. For instance, we previously proposed that cell type invariant TADs assemble into higher order structures such as A- and B- compartments that are cell type-specific and related to chromatin status (Gibcus and Dekker, 2013).

A Number of Different Looping Strategies Are Present at the *CFTR* Locus

The *CFTR* gene is contained within one TAD (number 5). TAD 5 also contains all the known *CFTR* regulatory elements (reviewed in McCarthy and Harris, 2005; Ott and Harris, 2011; Gheldof et al., 2010; Ott et al., 2009; Zhang et al., 2013, chapter 5 of this thesis). We see that, in Caco2 cells, the *CFTR* promoter interacts with elements II, III and IV as identified previously in our 3C experiment (Gheldof et al., 2010; chapter 3 of this thesis). We do not see the *CFTR* promoter looping to element I in Caco2 cells in 5C, which we saw by 3C

(Gheldof et al., 2010; chapter 3 of this thesis). This is because we used a stringent FDR cutoff when calling significant peaks – at a lower FDR threshold we do see the interaction between the *CFTR* promoter and element I in Caco2 cells, but we also dramatically increase the number of interactions called significant. We chose to use a more stringent FDR to decrease false positives rather than capture this interaction in Caco2 cells.

Interestingly, we see that Cau3 cells interact with the region upstream of the *CFTR* gene almost exclusively, only contacting element III within the gene body. A known lung-specific enhancer is located 35 kb upstream of the *CFTR* promoter (Zhang et al., 2013). Unfortunately our probe design does not cover the exact location of this enhancer, but we see a very strong signal from its neighboring fragment (the second very high peak in the upstream region, Figure 4.7H). We infer that this strong peak occurs because the *CFTR* promoter is contacting its neighboring fragment just as (or more) frequently. Additionally we see that the *CFTR* promoter in Cau3 cells interacts very frequently with elements I and II identified in our 3C study (though we did not use a lung cell line in the 3C work) (Gheldof et al., 2010, chapter 3 of this thesis). These elements are known to contain CTCF binding cites (Sabo et al., 2004). These elements may be active enhancers in Calu3 cells, or they may play another role, perhaps structurally. Further tests need to be done to confirm the function of these elements in lung cells.

All cell lines show at least one interaction with the 3´ end of the TAD. This is also corroborated by our 3C study, which showed this whole region interacted with the *CFTR* promoter in HeLa S3 and HT-29 cells (Gheldof et al., 2010, chapter 3 of this thesis). This may be a functional interaction (as is the element IV-*CFTR* promoter interaction in Caco2 cells) or it may be a structural interaction – or both.

In this chapter, we show that all the known *CFTR* regulatory elements are contained within the same TAD, and that different cell lines display diverse patterns of promoter-element looping within that TAD. We show that TADs are present regardless of gene expression status. We also describe differences between loops that occur inside TADs versus loops that occur between TADs. We propose that TAD boundaries play two important roles: first they constrain frequent looping interactions between gene promoters and gene regulatory elements within TADs. Second, they are involved in weaker longer-range interactions with other TADs, possibly leading to formation of higher order chromatin architectures. Overall, this study expands our knowledge of the *CFTR* gene and TAD organization.

CHAPTER V

CONCLUSION AND FUTURE DIRECTIONS

This thesis studies the relationship between three-dimensional structure of chromatin and gene regulation, focusing on the cystic fibrosis transmembrane conductance regulator (*CFTR*) locus as a model. Using the chromosome conformation capture technologies presented herein, this work identifies 3D looping between the *CFTR* gene promoter and a set of genomic elements, two of which act as enhancers. This work detects a common topologically associating domain structure of the *CFTR* locus in a variety of cell types, regardless of the expression status of the genes within the domains or the looping contacts they create. Possible transcription factors that may mediate *CFTR* expression enhancement were identified. This work has contributed to our understanding of 3D genome structure, how this structure is related to cell type-specific gene regulation and contributed new information for the field of Cystic Fibrosis research.

Implications for Genome-Wide Association Studies (GWAS)

GWAS are used to identify mutations associated with disease. In many GWAS, mutations, or single nucleotide polymorphisms (SNPs) are often found in non-genic regions and thus are hard to assign to a gene. 3C technology has already begun helping properly assigning SNPs to target genes (Ahmadiyeh et al., 2010; French et al., 2013; Harismendy et al., 2011; Visser et al., 2012; Zhou et al., 2012). Knowledge of TAD structure may also help GWAS identify target genes by limiting search space to within the SNP-containing TAD. Additionally,

GWAS authors can also use existing Hi-C and 5C studies to see if a looping interaction is present between their SNP location and a gene of interest. However, as gene-element looping can be very cell type-specific, researchers should include their own follow-up 3C studies in relevant cell lines. Techniques such as C technology will continue to help in the search for disease-causing mutations and should be more widely adapted in the field.

Implications for *CFTR* Mutation Screens

All children born in the United States are screened for a number of diseases at the time of birth, including Cystic Fibrosis. Babies are screened for a digestive chemical produced in the pancreas, and if the newborn has a high amount of this chemical, the child may have CF. At this point mutation screens are carried out. The child will also be screened if a parent is a known carrier of a CF allele, or if there is a family history of CF. The current screen checks for the most common 23 mutations and was put into use in 2004 (American College of Obstetricians and Gynecologists, 2011). Generally it is cost-prohibitive and unnecessary to screen for more mutations, but in some cases a CF mutation is not found in the patient. In addition to screening for further mutations, our research suggests that including known *CFTR* enhancers may reveal novel CF-causing mutations, as it is known that deleting or mutating an enhancer can result in complete loss of gene function (Sagai et al., 2005).

Elements III and IV are *CFTR* –specific Enhancers

Elements III and IV were shown to contact the *CFTR* promoter in three independent 3C experiments (Gheldof et al., 2010; Ott et al., 2009b, Chapter 3 of this thesis). These elements overlapped DNase1 hypersensitive sties identified by the Harris lab (Nuthall et al., 1999b; Smith et al., 2000). Additionally, these sites contained H3K4 acetylation and H3K4me3 in cells that express the *CFTR* gene, which imply enhancer-type elements in the human genome (ENCODE consortium). Since these sites loop to the *CFTR* promoter, they were tested as potential regulatory elements of *CFTR*. We investigated the regulatory potential of these elements and found that in a luciferase assay elements III and IV do indeed drive expression of a luciferase gene under the control of the *CFTR* promoter. We placed the element of interest directly upstream of the *CFTR* promoter. The classical definition of an enhancer is an element that will enhance gene expression regardless of its genomic position or its orientation. Elements III and IV were found to enhance *CFTR*-driven luciferase expression in a different assay, where the elements were placed behind the luciferase gene (Ott et al., 2009b), meeting the first requirement of the classical enhancer definition. However, the function of these elements when they are in the reverse orientation has yet to be tested. Given the corroborating genomic evidence discussed in chapter 3 and the success of these elements to enhance *CFTR* expression in two independently performed assays I am confident in classifying them as enhancer elements.

Elements I and II May Be *CFTR* Repressors

Although we discovered four elements looping to the *CFTR* promoter (Gheldof et al., 2010; Ott et al., 2009b, chapter 3 of this thesis), only two of them displayed enhancer activity in the luciferase assay, as discussed above and in chapter 3 of this thesis. Elements I and II did not show any increase in *CFTR* expression, even though they form 3D contacts with the promoter. ChIP-seq tracks from the ENCODE consortium and data from the Harris lab show that elements I and II bind to CTCF in Caco2, GM12878 and HepG2 cells (Blackledge et al., 2007, 2009; Sabo et al., 2004). Interestingly, elements I and II were also identified by the Harris lab as containing DHS sites, and element II was shown to be important for *CFTR* expression of a yeast artificial chromosome (Nuthall et al., 1999a; Smith et al., 1995). Element II has been shown to act as an enhancer-blocker (Blackledge et al., 2007). A recent study has shown that knocking down CTCF and RAD21 (cohesin) using siRNA decreases 3D interactions between element II and the *CFTR* promoter, resulting in a 2.5-fold increase in *CFTR* gene expression (Gosalia et al., 2014). Thus it appears that element II is acting as a moderator of *CFTR* by constraining gene expression. Additionally, element Ic may be interpreted as a repressive element from our luciferase assays (chapter 3 of this thesis). Thus elements I and II may play an important regulatory role in *CFTR* gene expression.

What is a TAD?

TADs are regions of the genome where interactions occur preferentially within the same TAD (Dixon et al., 2012; Nora et al., 2012). TADs are defined by clear boundary regions that show depletion for interactions occurring between the regions on either side of the boundary. TADs have been shown to be similar across cell lines and species (Dixon et al., 2012; Nora et al., 2012; Sexton et al., 2012). In our study we found that the TAD structure in our 2.8 Mb region was extremely similar between the 5 cell lines studied, in agreement with the published studies.

TADs are not defined by the looping that occurs within them. We clearly show that different intraTAD loops exist in the 5 cell lines we studied, yet TAD boundaries are similar across these 5 cell lines. TADs are also not defined by gene expression, as we show that cell lines with diverse gene expression status maintain the same TAD boundaries. I believe that TADs are best defined by their boundaries, however, what constitutes a TAD boundary is not yet clear.

It has been reported that TAD boundaries are enriched for CTCF sites and for gene promoters (Dixon et al., 2012). This analysis used TAD boundary calls from a genome-wide Hi-C experiment, and their TAD calling method placed a boundary with +/- 20 kb accuracy. Therefore, the enrichment of CTCF and gene promoters occurs in a 40 kb window, within which a TAD boundary is located. In our 5C experiment, we placed each boundary at the midpoint of a 10 kb bin. We noted that CTCF was not enriched in our boundary regions compared to the rest

of the 2.8 Mb dataset, and that our boundary regions tended to be enriched for the ends of genes instead of gene promoters. CTCF sites also occur in locations without a boundary, indicating that CTCF alone is not sufficient for boundary formation. Additional evidence comes from Zuin et al. who show that TAD boundaries are unchanged following knockdown of CTCF or RAD21 (Zuin et al., 2014). Though genome-wide studies allow a general claim to be made about these boundary areas, individual boundaries may not meet those assertions. It is clear that more research is needed to identify what creates and maintains TAD boundaries.

Possible TAD Function in the Human Genome

It has been suggested that TADs may correlate with gene expression, i.e. that genes within a TAD have similar expression status (Dily et al., 2014; Nora et al., 2012). This seems to be a reasonable conclusion if TADs constrain regulatory elements – the same regulatory elements would be present in one TAD and thus could regulate all the genes within. Indeed, Dily et al. showed that genes responsive to hormones resided in TADs that contained other genes similarly responsive to the hormone (i.e. in a TAD with one gene strongly induced by the presence of a hormone, other genes in that TAD were also upregulated, and in TADs where one gene was strongly repressed by a hormone other genes in that TAD were also repressed). They saw this effect in about 20% of TADs. Though this effect may be present for hormonal response, our dataset shows

that one gene in a TAD can be highly expressed while another gene located in the same TAD is not expressed at all. As mentioned earlier, the *CFTR* enhancers do not interact with the *ASZ1* gene promoter located in the same TAD. We conclude that TADs do not impose a similar regulation pattern over the genes contained within them.

TADs could play an important role in the process of gene promoters contacting their appropriate regulatory elements by creating a smaller genomic search space. We note that all known *CFTR* regulatory elements are contained within one TAD. A promising hypothesis for TAD function may be that promoters are limited to the area within their TAD when they search for their target regulatory elements. Interestingly, in our dataset we observe that gene promoters are not limited to interactions within their own TAD. For example, the *MET* gene promoter, which is expressed in Caco2, Calu3, Capan1 and HepG2 cells, shows a combined 25 out of 55 interactions occurring in the interTAD space. Similarly, the two fragments at the promoter of *CFTR* have 41/80 and 31/71 interTAD interactions, while the promoter fragment of *CTTNBP2* has 11/14 interactions in the interTAD space. Thus, TAD boundaries are not absolute barriers, and gene promoters do show interactions beyond their local TAD. However, interactions in the interTAD space are overall much weaker than the interactions in the intraTAD space. The interTAD interactions could be important for structural interactions rather than gene regulation. TADs may provide an important environment for

proper gene regulation. Experiments proposed below will help address the role of TADs in the human genome.

CFTR As A Model Locus for Further Investigations into TAD Structure and Gene Regulation

The contents of this thesis have set up a model locus for further study of the impact of 3D genome structure on gene expression. Our 5C analysis has identified specific TAD boundaries in a 2.8 Mb region of the genome surrounding the *CFTR* gene. We have shown that the entire *CFTR* gene and all its known regulatory elements are contained within one TAD (Smith et al. in preparation, chapter 4 of this thesis), and that known enhancers make physical contact with the *CFTR* promoter (Gheldof et al., 2010; Ott et al., 2009b, Smith et al. in preparation, chapter 3 and 4 of this thesis). Many questions remain in the field about how TADs function and about how looping elements form and are maintained, and how they contribute to gene regulation. One technology that will help answer these questions is the recently described CRISPR/Cas9 endonuclease mechanism (reviewed in Hsu et al., 2014). CRISPR/Cas9 is a molecular machine discovered in bacterial genomes, used as a defense against viruses. It has been co-opted by molecular biologists in order to edit the human genome. Cas9 is an endonuclease that makes a double strand break at its target site (determined by the CRISPR sequence). Human cells can be transfected with Cas9 and given CRISPR sequences in the form of guide RNAs (gRNA) that

guide the Cas9 enzyme to a targeted location, creating a double strand break at a precise genomic location. This system provides a pwerful tool for modifying the human genome. In our case, the CRISPR/Cas9 system can be used for many follow-up experiments.

In our model locus, there are three known enhancers that interact with the *CFTR* promoter: and element in intron1 and elements III and IV (Gheldof et al., 2010; Ott et al., 2009b). Using CRISPR/Cas9, it is possible to delete one or both of these elements from the genome and assay both gene expression and looping. I hypothesize that both *CFTR* gene expression and looping would be disrupted, with gene expression decreasing and looping to these elements abolished. It is also possible to swap these two elements. In this case I hypothesize that *CFTR* gene expression and looping would be maintained, since both elements are present within the TAD boundaries. It would be very interesting to swap one or both of the *CFTR* enhancers with other known enhancer elements. I hypothesize that *CFTR* expression would change (correlating with the strength of the new enhancer) and that looping would occur between the new enhancer and the *CFTR* promoter. Additionally one could delete the two known enhancers (elements III and IV) and put one or both of them back into the locus at a different position. In this case I hypothesize that *CFTR* gene expression would be maintained at a similar level to pre-swap, but that new looping interactions would appear between the promoter and the new locations of the elements.

Another set of interesting experiments that can be facilitated by the CRISPR/Cas9 system include moving the known *CFTR* enhancers into a neighboring TAD, without disrupting the TAD boundary. I predict that in this instance, *CFTR* gene expression would decrease and the promoter would no longer loop to these enhancers, since they are now located in a different TAD. This assumes that our idea that TADs create a small search space for gene promoters to find their target enhancers is true. It is possible that the *CFTR* gene promoter would loop into the neighboring TAD to contact its enhancers. This would force us to adjust our current TAD model, since the TAD boundary would not be preventing the gene promoter from searching neighboring TADs to find its enhancers.

This brings us to the TADs themselves. CRISPR/Cas9 can also be used to aid in our understanding of what TAD function is in the genome. TADs seem to be defined by their boundary elements. CTCF is enriched at these boundaries, though this is not true for every boundary case (Dixon et al., 2012). At this time how a TAD boundary is created and maintained is still unknown. It would be interesting to entirely delete one or more boundaries in our locus and, using 5C, look to see what happens to the TADs. I predict that the two TADs next to the deleted boundary would mix into one larger TAD, as this has occurred before (Nora et al., 2012). Next, I would try to place a known boundary region inside a stable TAD. I hypothesize that placing a boundary region into a TAD would create a new boundary, cutting the current TAD into two smaller TADs. It would

be very interesting to place a boundary like this between a gene promoter and its known regulatory element and test to see if the looping interaction between these two gene loci is maintained or disrupted when a boundary is placed in between them.

These types of experiments are natural extensions of the results in this thesis. They would allow us to investigate the details of enhancer function and TAD structure. The locus presented in this thesis is a great model system to test these ideas. It contains 12 genes, one that is clinically relevant (*CFTR)*, some that are variably expressed, and some that are off in most cell types and some that are on in most cell types. The more knowledge we gain about the *CFTR* gene and its surrounding DNA will only benefit the current therapies available to Cystic Fibrosis patients. Indeed, I submit the idea that proper gene therapy, if ever realized, will require more than just the *CFTR* gene, but will need to include its proper regulatory elements. This work provides a baseline for future experiments that will allow researchers to learn more about the proper three-dimensional genomic structure of the genome and of the *CFTR* gene.

APPENDIX I

EXPERIMENTAL DATA AND PRIMER TABLES

Table A.1: Primer Sequences and Chromosome Positions Used in the 3C Experiments.

3C primers

Name	Sequence (5´-3´)	EcoRI fragment position (HG18) start (Chromosome 7)	end (Chromosome 7)
EcoRI CFTR primers			
CFTR324	GATCTCTCCTTTCACTGAGCACCATAGG	116773744	116777008
CFTR325	GGCAACAGAAACTAACTGTGAGAGGGGCTA	116777009	116778018
CFTR327	TGTGTAGCACAGGTGCCTGACATATAGC	116778535	116793159
CFTR330	GAACCTAGGAGTTTGAGACCAACTGTGG	116797877	116804745
CFTR331	CCAACTTTGTCTCTTCAGATAGCTCTCC	116804746	116807423
CFTR337	TCCCATCTGTTTCTTTCTTTGGCGAGCG	116820538	116822605
CFTR338	GCTACATGTAAAGTGTCCAGAACAGGCG	116822606	116827113
CFTR339	GGATTAGCTACAGAAGAGAATTAATGGCCTGG	116827114	116828613
CFTR343	CAGTAGACCATTACACCCAACCATTGCC	116833484	116835882
CFTR348	GAACTGGACTGGCAAGAGACACAATGGA	116837533	116844091
CFTR350	TCCTGGACCAATGAAACATCTCCGATGG	116849034	116854362
CFTR352	TCTAGGACATCTACACTTCATAGCGGGG	116849034	116854362
CFTR355	ACGTCTGGTGGGTTATGGTGTGCTCCTA	116863884	116868023
CFTR357	GGGCATCCACATAGGAAAGGAAGAAGTC	116877428	116880943
CFTR359	GCCCTAGGTGGGTCTAAAAGTTCCAGTT	116883208	116885279
CFTR360	AACCCCATTGTCTCAGCCACAAAACTCC	116885280	116890156
CFTR362	GGTTTTTCCCCCACCCCATTTTAGAGCA	116891302	116898038
CFTR363	GAACTGGGAATTGGCTGTCCTTTTGACC	116898039	116901211
CFTR368	CACATACAGGTGCCATGCCTGCATATAG	116903489	116907872
CFTR369	CCCAGGATCATGTTCGGATCAATCTAGTGTGC	116907873	116917169
CFTR371	AGAAACCATGTGTGAGACAACCATGTTGAG	116920524	116925167
CFTR374	CCACCTTGTTCTAAGCACAGCAATGAGC	116927536	116936141
CFTR375	CTTTGCAGGTCCACTTCTATCTGCTGGG	116936142	116940935
CFTR378	GGCCCATGCCACCCAAATATTATGTAAGC	116943215	116947747
CFTR379	TTCCCACCAGGAATGCAATCACTTGAGC	116947748	116949548
CFTR384	CAGCACGTGCATTGAACCCATAAGAACC	116954209	116959011
CFTR386	GGTTCTAGGAGACTCAGTGAAAGTTGATACAC	116959563	116964868
CFTR388	GAAGACCACTATGCTAAAGCTTCCCACG	116965514	116975353
CFTR390	CTCTTTCAAAGCTGTAGACTGTGACTCAGC	116975463	116986175
CFTR391	GGTAATATTGTTCCCATGAGCCTTTCTTGGAG	116986176	116990697
CFTR393	ACCACTTCTCCTGCACTTGACAATACCG	116991712	116992706

CFTR395	GGAGGATAAGACTCAAACAGGTAGTGGG	116994136	116997094
CFTR396	GCCCTCTGTTTGTAGAATGATGCTTGCC	116997095	116998150
CFTR397	AGGCATGAAGAGATTCTTTCCCAAGGGG	116998151	117002217
CFTR400	ATTGACCCTTGGACAATGTGGGGGTTAG	117003291	117005125
CFTR402	TGGATCAGAGGAACCAAACAGAAAAGCC	117005136	117007103
CFTR405	ATCAACCCCACAGTAACTGTCTCG	117008018	117013017
CFTR406	GACAGGAAACTAACCCTAACTGAGCACC	117013018	117016269
CFTR407	CTCAGTTCTATGTTGGATGTCTAGAAGCC	117016270	117018702
CFTR410	TAGTCCCCACAACTACCCTGGAAGGTAT	117019552	117021424
CFTR412	CAGCTGTGATCTTGGCTTTCTTGTGAGG	117023248	117029897
CFTR413	AGGTAGGTGGGTTGAGGCCAAATTATGC	117029898	117033264
CFTR415	GACAGTGAATACTGGCAGAGCAAATGCC	117035191	117039940
CFTR417	GATTCCCATCATAGTGCTGTCACCCATG	117042373	117044539
CFTR420	TATCTGGGAAGGATTGAGCCACAGGATC	117051477	117053896
CFTR421	GCGATCTGTGAGCCGAGTCTTTAAGTTC	117053897	117054925
CFTR422	GTTTCACAGACCTTCATTTGCCTGAGCC	117054926	117057582
CFTR425	TCCATCCATTCTTCCATCTGGCTATCCC	117062277	117068221
CFTR426	TACCACAGGTGAGCAAAAGGACTTAGCC	117068222	117070058
CFTR427	CTCAACCTGGCGATTGCTAACTTGGAGG	117070059	117073275
CFTR428	GCAGAGACACAGATCCAAACCATACCAC	117073276	117075081
CFTR432	GCCTCCTGTTACTCATCTTTGAATCTGGGG	117078729	117088313
CFTR433	TAGGAAGATCCCTGGGGTGTTTGTATGG	117088314	117090094
CFTR434	TGGCCTTCATACACTCAGGAGTACTTGC	117090095	117091392
CFTR435	TCAGGTAGTACCTTTCTCAACCACCTTCTC	117091393	117093441
CFTR437	ATCAAGGGTACACTGCCTTCTCAACTCC	117093828	117096070
CFTR438	ACTGAAGCCTGTCAACAGGGACAAGAAC	117096071	117105570
CFTR439	GGGTTTCATAGTGCTGCACTCAGTTC	117105571	117111634
CFTR440	GTTTAGGCTTGCTAGGAGGACAGCTAGG	117111635	117114566
CFTR441	TTGGCATAGATAAGTCACTCACCTGAGG	117114567	117127743
CFTR442	TGCAGGATTCTGATTCCATTCCCTCAGC	117127744	117131345
CFTR444	TGCCTTATGAGTGGCTGTAACGTTCAGC	117145780	117148471
CFTR446	TGTGTACACCTGATTCACCTGGAAAGGC	117149228	117151618
CFTR448	GTAATAGTGGCAAAGCCAGGGCTAGAGT	117152182	117156149
CFTR449	GGGGACAGGACATGTTGAGTTTGTTACG	117156149	117159202
CFTR452	TCAGGGCTTATGGAGAGGTGAAAGAAGG	117161794	117166672

EcoRI Gene desert primers		EcoRI fragment position (HG18) start (Chromosom	end (Chromosom

		e 16)	e 16)
GD2	AGCTTCACCTCTCAAACTACAGGACTGG	60845578	60849115
GD3	GTATACTCAGTTGAGCAGCCCATGACAC	60849116	60851238
GD5	GTTCTCTGTCTTATAATTATGCTACAAGAATGAG G	60851792	60855935
GD6	GTTTAAGACCCTCAGTATACTAGTCATAGAAGG	60855936	60858531
GD7	GATGCCATTTCTTATCTTGTCTTGGCAGGTC	60858532	60866968
GD8	GCAGCAAAGCAAACCAAAAGAACAACAGG	60866969	60873873
GD9	GTGTCATGGAATCAAAGGTGAGTGAGGG	60873874	60876657
GD11	CTTAGGACCCTGTGACTCCTGTTTTCTCC	60877690	60880026
GD17	GGCTGGCTGAGGTCATTCATGCAATCTT	60894707	60905483
GD18	CCATTCCATCATACACCCTCATCTCACTGCC	60905484	60908417

BsrGI CFTR primers		BsrGI fragment position (HG18) start (Chromosom e 7)	end (chromosome 7)
EMS1	CTGAATATCATGTCGCAATTAGTGCATC	116903317	116904928
EMS2	TGTTAACAGCCAAGTGCTACCTAATTT	116904940	116910993
EMS3	TTTACTCATAGCTCGTGACAATGCAGGC	117109806	117111366
EMS19	GTGAATATGCTAATGCTGGCCTTAATTC	117096897	117101305
EMS18	ATCCATCACTGGCTACTTGGTTTGCTTC	117092945	117096897
EMS20	CAACAAAAACACCCACCGTGAGGTGATG	117111366	117118556
EMS21	CACCTTTGTTGTCTACTGAGCAGAGGTAGC	117118556	117118931
EMS27	GAAACTTTCCGTAAAATTATCGATTCCAG	117101305	117109806
EMS33	TTTCCACTATACTCTGTCATGTTCAGAG	116903167	116903316
EMS34	CACCTTAAAGCAAGTACGCATGATAAAG	116910993	116916818
EMS35	TTGGGGAAAGTCGATTCAAGGCAGTAAC	116916819	116921623
EMS21a	GATACACCTTTGTTGTCTACTGAGCAGAG	117118556	117118930
EMS36	GCAGGATACCAAGTCTATCTTAGCACCATA	117003597	117005294
EMS37	CAATAGTACATCTGGCAAGGTCATGCAAAC	117005422	117008602
EMS38	ACCTATCTAACTCTTCGCATTCTTGAAGTC	117014031	117014185
EMS39	TCTAGGATTCTGCATAATACTGGACAGAGA	117014186	117015473
EMS40	GTGTAATGTGTTGTCCAGTTTTGGATGATG	117016422	117016451
EMS41	GCTTAGACTACTCATCTACCTCAATACTTC	117018151	117018180
EMS42	CTAAACTCAAGTCCCATGCTACCTTCAGAG	117025603	117025632
EMS43	ATTAACACTCCTCTAGTTAGAACAAAGAGG	117028400	117028429
EMS44	CCCAGTTAATAGATATTTCGATTGTTCC	116921624	116923928

BsrGI Gene desert primers		BsrGI fragment position (HG18) start (Chromosom	end (chromosome 16)

		e 16)	
EMS13	CTTTAGATAATGCCGATCTACTTTCTGGG	60835929	60839882
EMS14	TTTCCCCGTGTGACTATGAAGGATACAG	60839882	60844701
EMS15	TGCACTCCGGCCAAGGTAGCAGAACAAGA	60844701	60852673
EMS16	GGTGGTAAGTCCTGCAACCATTTAGTGC	60867202	60871330
EMS17	TCATACTGATGACCCTAACCCTGTGCAG	60898436	60900384

Table A.2: All 3C Interaction Values. Start and end coordinates, chr7 hg18. Fragments named based on EcoR1 restriction site number in the region, except for the BsrGI fine mapping experiment. Values listed are the averages if at least three independent values, with SE = standard error of those values.

3C experimental data

Anchor at Promoter: Fragment 368

Start	End	name	GM06990	Caco2	SE GM06990	SE Caco2
116773744	116777008	EcoRI-324				
116777009	116778018	EcoRI-325	0.059028	0.2255	0.029706	0.1762
116804746	116807423	EcoRI-331	0.072963	0.3203	0.035137	0.1519
116822606	116827113	EcoRI-338	0.109829	0.7058	0.05593	0.1032
116837533	116844091	EcoRI-348	0.184689	0.4268	0.031965	0.1075
116849034	116854362	EcoRI-350	0.196216	0.3278	0.024463	0.0653
116863884	116868023	EcoRI-355	0.163491	0.4788	0.048081	0.055
116883208	116885279	EcoRI-359	0.365195	1.5935	0.035409	0.1925
116885280	116890156	EcoRI-360	0.390279	0.8674	0.099406	0.0938
116891302	116898038	EcoRI-362	0.509712	0.9929	0.03367	0.1138
116898039	116901211	EcoRI-363	0.823164	1.3186	0.067132	0.155
116907873	116917169	EcoRI-369	0.984552	1.0857	0.066975	0.0738
116920524	116925167	EcoRI-371	0.669593	0.8284	0.04563	0.0941
116927536	116936141	EcoRI-374				
116936142	116940935	EcoRI-375	0.24437	0.3449	0.040186	0.0259
116943215	116947747	EcoRI-378				
116947748	116949548	EcoRI-379	0.1404	0.4799	0.01642	0.1339
116959563	116964868	EcoRI-386	0.106288	0.3279	0.017809	0.0648
116975463	116986175	EcoRI-390	0.05664	0.1933	0.019519	0.047
116994136	116997094	EcoRI-395				
116998151	117002217	EcoRI-397	0.101649	0.312	0.041325	0.0055
117003291	117005125	EcoRI-400				
117008018	117013017	EcoRI-405	0.121977	0.3161	0.027904	0.0762
117013018	117016269	EcoRI-406	0.09631	0.9639	0.027054	0.046
117016270	117018702	EcoRI-407	0.054558	1.0794	0.009394	0.1323
117019552	117021424	EcoRI-410				
117023248	117029897	EcoRI-412				
117035191	117039940	EcoRI-415	0.048333	0.3034	0.016351	0.0198
117051477	117053896	EcoRI-420				
117053897	117054925	EcoRI-421	0.086966	0.3144	0.024326	0.0479
117070059	117073275	EcoRI-427	0.081519	0.1614	0.026643	0.0295
117073276	117075081	EcoRI-428				

117078729	117088313	EcoRI-432				
117088314	117090094	EcoRI-433				
117090095	117091392	EcoRI-434	0.0824	0.2484	0.034428	0.044
117096071	117105570	EcoRI-438	0.098919	0.3143	0.009577	0.0377
117105571	117111634	EcoRI-439	0.127753	0.626	0.037908	0.0472
117111635	117114566	EcoRI-440	0.062175	0.448	0.030215	0.0541
117114567	117127743	EcoRI-441	0.085828	0.2377	0.018547	0.0388
117127744	117131345	EcoRI-442				
117145780	117148471	EcoRI-444				
117149228	117151618	EcoRI-446				

Start	End	name	HepG2	HeLa	SE HepG2	SE HeLa
116773744	116777008	EcoRI-324				
116777009	116778018	EcoRI-325	0.095055	0.0849	0.030743	0.0461
116804746	116807423	EcoRI-331	0.208439	0.2737	0.04823	0.0709
116822606	116827113	EcoRI-338	0.189723	0.422	0.037485	0.0683
116837533	116844091	EcoRI-348	0.197211	0.2846	0.024337	0.1353
116849034	116854362	EcoRI-350	0.138971	0.2379	0.009357	0.1245
116863884	116868023	EcoRI-355	0.251463	0.2756	0.126408	0.0287
116883208	116885279	EcoRI-359	0.665637	0.5145	0.095394	0.0946
116885280	116890156	EcoRI-360	0.661475	0.8481	0.070451	0.1166
116891302	116898038	EcoRI-362	0.549455	0.3763	0.174991	0.0434
116898039	116901211	EcoRI-363	1.073387	0.771	0.045413	0.1322
116907873	116917169	EcoRI-369	1.122547	0.7069	0.08207	0.0397
116920524	116925167	EcoRI-371	0.564628	0.7585	0.081609	0.2329
116927536	116936141	EcoRI-374	0.147572	0.1892	0.027432	0.0078
116936142	116940935	EcoRI-375				
116943215	116947747	EcoRI-378				
116947748	116949548	EcoRI-379	0.120832	0.1446	0.015455	0.0186
116959563	116964868	EcoRI-386	0.052881	0.2178	0.01475	0.0535
116975463	116986175	EcoRI-390	0.06216	0.1312	0.007956	0.0179
116994136	116997094	EcoRI-395				
116998151	117002217	EcoRI-397	0.051556	0.1624	0.026291	0.0277
117003291	117005125	EcoRI-400				
117008018	117013017	EcoRI-405	0.084211	0.1991	0.02476	0.0421
117013018	117016269	EcoRI-406	0.087058	0.4297	0.020314	0.0657
117016270	117018702	EcoRI-407	0.057757	0.1754	0.018444	0.0319
117019552	117021424	EcoRI-410				
117023248	117029897	EcoRI-412				
117035191	117039940	EcoRI-415	0.073816	0.0422	0.038378	0.0121

117051477	117053896	EcoRI-420				
117053897	117054925	EcoRI-421	0.088373	0.0243	0.051973	0.0246
117070059	117073275	EcoRI-427	0.050296	0.0727	0.026232	0.0247
117073276	117075081	EcoRI-428				
117078729	117088313	EcoRI-432				
117088314	117090094	EcoRI-433				
117090095	117091392	EcoRI-434	0.082921	0.31	0.047573	0.0971
117096071	117105570	EcoRI-438	0.023515	0.2455	0.012048	0.0086
117105571	117111634	EcoRI-439	0.046716	0.4706	0.016642	0.0444
117111635	117114566	EcoRI-440	0.054599	0.1758	0.011503	0.046
117114567	117127743	EcoRI-441	0.067805	0.1572	0.011872	0.014
117127744	117131345	EcoRI-442	0.045559	0.3247	0.045105	0.1318
117145780	117148471	EcoRI-444	0.085247	0.1151	0.037509	0.0586
117149228	117151618	EcoRI-446				

Start	**End**	**name**	**K562**	**HT29**	**SE K562**	**SE HT29**
116773744	116777008	EcoRI-324	0.015124	0.1162	0.004882	0.0128
116777009	116778018	EcoRI-325				
116804746	116807423	EcoRI-331	0.065392	0.1781	0.008163	0.0621
116822606	116827113	EcoRI-338	0.067476	0.2755	0.013022	0.043
116837533	116844091	EcoRI-348	0.082678	0.2135	0.006645	0.028
116849034	116854362	EcoRI-350	0.088788	0.2075	0.016058	0.021
116863884	116868023	EcoRI-355	0.137671	0.2569	0.018188	0.0374
116883208	116885279	EcoRI-359	0.242225	0.3901	0.047365	0.0559
116885280	116890156	EcoRI-360	0.219342	0.3265	0.028884	0.0473
116891302	116898038	EcoRI-362	0.329156	0.2653	0.027368	0.0407
116898039	116901211	EcoRI-363	0.583964	0.3852	0.040811	0.0478
116907873	116917169	EcoRI-369	0.607507	0.2643	0.020956	0.0271
116920524	116925167	EcoRI-371	0.492266	0.2242	0.072069	0.0513
116927536	116936141	EcoRI-374	0.150291	0.1484	0.022919	0.0398
116936142	116940935	EcoRI-375				
116943215	116947747	EcoRI-378	0.08382	0.1123	0.013717	0.0242
116947748	116949548	EcoRI-379				
116959563	116964868	EcoRI-386				
116975463	116986175	EcoRI-390	0.057818	0.0821	0.007832	0.0388
116994136	116997094	EcoRI-395	0.034532	0.0718	0.004406	0.0181
116998151	117002217	EcoRI-397				
117003291	117005125	EcoRI-400	0.024274	0.1183	0.013262	0.0365
117008018	117013017	EcoRI-405				
117013018	117016269	EcoRI-406	0.043408	0.2176	0.011537	0.0299

117016270	117018702	EcoRI-407				
117019552	117021424	EcoRI-410	0.044562	0.1012	0.012435	0.0289
117023248	117029897	EcoRI-412	0.026932	0.0397	0.002326	0.0198
117035191	117039940	EcoRI-415	0.028748	0.026	0.006036	0.0067
117051477	117053896	EcoRI-420	0.023855	0.0759	0.019714	0.0041

Anchor at Element I - Fragment 338

Start	End	name	GM06990	Caco2	SE GM06990	SE Caco2
116797877	116804745	EcoRI-330	0.254685	0.5284	0.049502	0.0725
116820538	116822605	EcoRI-337	1.722173	1.1985	0.204827	0.0996
116827114	116828613	EcoRI-339	1.308083	1.3419	0.110867	0.2041
116849034	116854362	EcoRI-350	0.206251	0.2153	0.065428	0.0521
116863884	116868023	EcoRI-355	0.252983	0.4468	0.073971	0.1321
116883208	116885279	EcoRI-359	0.217338	0.5313	0.112178	0.1689
116885280	116890156	EcoRI-360				
116891302	116898038	EcoRI-362	0.142764	0.7456	0.057846	0.2309
116903489	116907872	EcoRI-368	0.109829	0.7058	0.05593	0.1032
116907873	116917169	EcoRI-369	0.240473	0.5779	0.089852	0.083
116920524	116925167	EcoRI-371				
116943215	116947747	EcoRI-378	0.111934	0.1292	0.028472	0.0185
116965514	116975353	EcoRI-388	0.053488	0.0798	0.004437	0.0443
116998151	117002217	EcoRI-397	0.097606	0.277	0.020365	0.0638
117013018	117016269	EcoRI-406	0.098859	0.2723	0.012458	0.0386
117016270	117018702	EcoRI-407	0.045716	0.2055	0.018914	0.0283
117035191	117039940	EcoRI-415	0.103985	0.1072	0.011148	0.0409
117061386	117062276	EcoRI-424	0.251107	0.132	0.043936	0.0407
117088314	117090094	EcoRI-433	0.150655	0.3941	0.010713	0.0664
117096071	117105570	EcoRI-438	0.177773	0.2453	0.066427	0.0058
117105571	117111634	EcoRI-439	0.123496	0.2536	0.017319	0.0683
117111635	117114566	EcoRI-440	0.114591	0.1514	0.030705	0.0357
117127744	117131345	EcoRI-442	0.062213	0.3032	0.017502	0.0416
117145780	117148471	EcoRI-444				

Start	End	name	HepG2	HeLa	SE HepG2	SE HeLa
116797877	116804745	EcoRI-330	0.319813	0.7106	0.016533	0.0894
116820538	116822605	EcoRI-337	1.305474	1.2247	0.033585	0.0979
116827114	116828613	EcoRI-339	1.367229	1.1811	0.132693	0.099
116849034	116854362	EcoRI-350	0.294874	0.4376	0.04853	0.1478
116863884	116868023	EcoRI-355	0.107873	0.2804	0.021437	0.1965

116883208	116885279	EcoRI-359	0.274132	0.4182	0.119795	0.0695
116885280	116890156	EcoRI-360	0.417959	0.7101	0.099026	0.1489
116891302	116898038	EcoRI-362	0.143237	0.5268	0.059961	0.1703
116903489	116907872	EcoRI-368	0.189723	0.422	0.037485	0.0683
116907873	116917169	EcoRI-369				
116920524	116925167	EcoRI-371	0.096782	0.2346	0.014691	0.0955
116943215	116947747	EcoRI-378	0.100889	0.0893	0.058148	0.018
116965514	116975353	EcoRI-388	0.020868	0.0748	0.020197	0.0471
116998151	117002217	EcoRI-397	0.111503	0.2199	0.037132	0.1023
117013018	117016269	EcoRI-406	0.100502	0.3567	0.035413	0.0402
117016270	117018702	EcoRI-407				
117035191	117039940	EcoRI-415	0.062414	0.1494	0.046212	0.054
117061386	117062276	EcoRI-424	0.032764	0.435	0.016115	0.1258
117088314	117090094	EcoRI-433	0.079868	0.431	0.038075	0.0818
117096071	117105570	EcoRI-438	0.169819	0.4213	0.053504	0.0083
117105571	117111634	EcoRI-439	0.053686	0.5933	0.039055	0.0969
117111635	117114566	EcoRI-440	0.066429	0.1837	0.039418	0.0966
117127744	117131345	EcoRI-442	0.057082	0.3243	0.048285	0.0399
117145780	117148471	EcoRI-444	0.109981	0.1697	0.054748	0.112

Anchor at Element II - Fragment 359

Start	End	name	GM06990	Caco2	SE GM06990	SE Caco2
116804746	116807423	EcoRI-331	0.2277	0.5477	0.071372	0.1042
116822606	116827113	EcoRI-338	0.217338	0.5313	0.112178	0.1689
116833484	116835882	EcoRI-343	0.361827	0.8007	0.231007	0.2829
116863884	116868023	EcoRI-355	0.584795	0.7231	0.194489	0.07
116877428	116880943	EcoRI-357	1.350006	1.4773	0.134305	0.0715
116883208	116885279	EcoRI-359				
116885280	116890156	EcoRI-360	1.327891	1.2762	0.051307	0.0978
116891302	116898038	EcoRI-362				
116898039	116901211	EcoRI-363				
116903489	116907872	EcoRI-368	0.365195	1.5935	0.035409	0.1925
116907873	116917169	EcoRI-369				
116920524	116925167	EcoRI-371	0.518776	0.8335	0	0.2548
116943215	116947747	EcoRI-378	0.16953	0.1997	0.025429	0.03
116965514	116975353	EcoRI-388				
116994136	116997094	EcoRI-395	0.125006	0.2409	0.060316	0.0416
116998151	117002217	EcoRI-397				
117013018	117016269	EcoRI-406	0.092062	0.4933	0.027538	0.0705
117016270	117018702	EcoRI-407	0.132119	0.6796	0.035192	0.2415

117029898	117033264	EcoRI-413	0.093554	0.2941	0.028888	0.0276
117051477	117053896	EcoRI-420	0.142561	0.2979	0.071592	0.0463
117073276	117075081	EcoRI-428	0.120912	0.2003	0.060801	0.0661
117091393	117093441	EcoRI-435				
117096071	117105570	EcoRI-438				
117105571	117111634	EcoRI-439	0.101915	0.3518	0.017523	0.0424
117127744	117131345	EcoRI-442				
117152182	117156149	EcoRI-448	0.148889	0.3261	0.055404	0.039

Start	End	name	HepG2	HeLa	SE HepG2	SE HeLa
116804746	116807423	EcoRI-331	0.194641	0.3358	0.122792	0.2063
116822606	116827113	EcoRI-338	0.417959	0.7101	0.099026	0.1489
116833484	116835882	EcoRI-343	0.392633	0.5442	0.073409	0.104
116863884	116868023	EcoRI-355	0.570234	0.6693	0.085609	0.1318
116877428	116880943	EcoRI-357	2.346159	1.0749	0.262381	0.1245
116883208	116885279	EcoRI-359	1.908843	1.4083	0.166004	0.1967
116885280	116890156	EcoRI-360				
116891302	116898038	EcoRI-362	1.250003	0.631	0.185834	0.0635
116898039	116901211	EcoRI-363	1.405198	0.8714	0.036471	0.1931
116903489	116907872	EcoRI-368	0.661475	0.8481	0.070451	0.1166
116907873	116917169	EcoRI-369				
116920524	116925167	EcoRI-371				
116943215	116947747	EcoRI-378	0.170959	0.4377	0	0.1118
116965514	116975353	EcoRI-388	0.223925	0.3565	0.004055	0.1007
116994136	116997094	EcoRI-395				
116998151	117002217	EcoRI-397	0.178905	0.2337	0.020779	0.0407
117013018	117016269	EcoRI-406	0.170655	0.59	0.054152	0.0969
117016270	117018702	EcoRI-407				
117029898	117033264	EcoRI-413	0.191264	0.3003	0.037448	0.0813
117051477	117053896	EcoRI-420	0.233349	0.1456	0.126907	0.0761
117073276	117075081	EcoRI-428				
117091393	117093441	EcoRI-435	0.318457	0.5282	0.052486	0.1651
117096071	117105570	EcoRI-438				
117105571	117111634	EcoRI-439	0.176969	0.528	0.038786	0.0716
117127744	117131345	EcoRI-442	0.262463	0.6111	0.114092	0.1122
117152182	117156149	EcoRI-448				

Start	End	name	K562	HT29	SE K562	SE HT29
116804746	116807423	EcoRI-331	0.332786	0.5915	0.175332	0.3216

116822606	116827113	EcoRI-338	0.274132	0.4182	0.119795	0.0695
116833484	116835882	EcoRI-343	0.313688	0.4168	0.296745	0.1149
116863884	116868023	EcoRI-355	0.76115	0.542	0.221117	0.1692
116877428	116880943	EcoRI-357	1.737913	1.4452	0.434226	0.341
116883208	116885279	EcoRI-359				
116885280	116890156	EcoRI-360	1.908843	1.4083	0.166004	0.1967
116891302	116898038	EcoRI-362				
116898039	116901211	EcoRI-363				
116903489	116907872	EcoRI-368	0.665637	0.5145	0.095394	0.0946
116907873	116917169	EcoRI-369	0.260161	0.2956	0.104427	0.1092
116920524	116925167	EcoRI-371				
116943215	116947747	EcoRI-378	0.360015	0.3916	0.097008	0.0047
116965514	116975353	EcoRI-388	0.15566	0.0445	0.048838	0.0436
116994136	116997094	EcoRI-395	0.073374	0.1175	0.037159	0.0293
116998151	117002217	EcoRI-397				
117013018	117016269	EcoRI-406	0.105428	0.2689	0.033887	0.0454
117016270	117018702	EcoRI-407				
117029898	117033264	EcoRI-413	0.037639	0.3288	0.022237	0.0389
117051477	117053896	EcoRI-420	0.03192	0.078	0.031526	0.0128
117073276	117075081	EcoRI-428	0.135132	0.3435	0.046305	0.0673
117091393	117093441	EcoRI-435	0.128634	0.2702	0.048733	0.1034
117096071	117105570	EcoRI-438	0.351526	0.1406	0.029004	
117105571	117111634	EcoRI-439	0.567412	0.0255	0.046177	
117127744	117131345	EcoRI-442	0.1352	0.6263	0.035918	0.0872
117152182	117156149	EcoRI-448	0.155954	0.2316	0.042735	0.1566

Anchor at Element III - Fragment 406

Start	End	name	GM06990	Caco2	SE GM06990	SE Caco2
116804746	116807423	EcoRI-331	0.017301	0.1466	0.012728	0.0166
116822606	116827113	EcoRI-338	0.098859	0.2723	0.012458	0.0386
116849034	116854362	EcoRI-350	0.034164	0.2178	0.015033	0.0707
116863884	116868023	EcoRI-355	0.07663	0.3239	0.025211	0.0588
116883208	116885279	EcoRI-359	0.092062	0.4933	0.027538	0.0705
116885280	116890156	EcoRI-360				
116891302	116898038	EcoRI-362	0.06332	0.4369	0.01918	0.0754
116903489	116907872	EcoRI-368	0.09631	0.9639	0.027054	0.046
116907873	116917169	EcoRI-369	0.129613	0.7198	0.026181	0.043
116920524	116925167	EcoRI-371	0.082935	0.4924	0.048947	0.0323
116947748	116949548	EcoRI-379	0.105267	0.2142	0.041681	0.0262
116959563	116964868	EcoRI-386	0.14584	0.3186	0.062296	0.0385

116986176	116990697	EcoRI-391	0.164215	0.2877	0.024379	0.0532
117003291	117005125	EcoRI-400	0.299104	0.8186	0.006406	0.0514
117016270	117018702	EcoRI-407	0.558641	1.4461	0.04279	0.042
117019552	117021424	EcoRI-410				
117042373	117044539	EcoRI-417	0.160777	0.3252	0.027017	0.0544
117068222	117070058	EcoRI-426	0.108055	0.4172	0.036531	0.023
117090095	117091392	EcoRI-434	0.067027	0.5391	0.019846	0.0384
117105571	117111634	EcoRI-439	0.088219	0.6045	0.02756	0.0831
117127744	117131345	EcoRI-442				
117149228	117151618	EcoRI-446	0.107094	0.281	0.045659	0.0422

Start	End	name	HepG2	HeLa	SE HepG2	SE HeLa
116804746	116807423	EcoRI-331	0.099695	0.1798	0.021341	0.0356
116822606	116827113	EcoRI-338	0.100502	0.3567	0.035413	0.0402
116849034	116854362	EcoRI-350	0.093923	0.2227	0.026431	0.0993
116863884	116868023	EcoRI-355	0.064584	0.2234	0.029959	0.0491
116883208	116885279	EcoRI-359	0.105428	0.2689	0.033887	0.0454
116885280	116890156	EcoRI-360	0.170655	0.59	0.054152	0.0969
116891302	116898038	EcoRI-362	0.083668	0.2735	0.064624	0.0519
116903489	116907872	EcoRI-368	0.087058	0.4297	0.020314	0.0657
116907873	116917169	EcoRI-369				
116920524	116925167	EcoRI-371	0.101026	0.3301	0.019326	0.0734
116947748	116949548	EcoRI-379	0.073574	0.0972	0.015759	0.0037
116959563	116964868	EcoRI-386	0.073332	0.0958	0.014309	0.0483
116986176	116990697	EcoRI-391	0.102106	0.132	0.041825	0.0385
117003291	117005125	EcoRI-400	0.409836	0.3966	0.040836	0.0375
117016270	117018702	EcoRI-407				
117019552	117021424	EcoRI-410	0.423273	0.4545	0.016615	0.0432
117042373	117044539	EcoRI-417	0.16042	0.268	0.063396	0.0743
117068222	117070058	EcoRI-426	0.055721	0.246	0.010634	0.1085
117090095	117091392	EcoRI-434	0.125978	0.3476	0.03103	0.113
117105571	117111634	EcoRI-439	0.064345	0.4983	0.02095	0.0801
117127744	117131345	EcoRI-442	0.039965	0.297	0.026499	0.0178
117149228	117151618	EcoRI-446				

Anchor at Element IV - Fragment 439

Start	End	name	GM06990	Caco2	SE GM06990	SE Caco2
116804746	116807423	EcoRI-331	0.049524	0.0578	0.024889	0.0009
116822606	116827113	EcoRI-338	0.123496	0.2536	0.017319	0.0683

116833484	116835882	EcoRI-343	0.077888	0.1371	0.02814	0.0361
116849034	116854362	EcoRI-350	0.057388	0.121	0.027801	0.0341
116883208	116885279	EcoRI-359	0.101915	0.3518	0.017523	0.0424
116885280	116890156	EcoRI-360				
116898039	116901211	EcoRI-363	0.05697	0.1997	0.041046	0.0111
116903489	116907872	EcoRI-368	0.127753	0.626	0.037908	0.0472
116920524	116925167	EcoRI-371	0.114751	0.292	0.039124	0.034
116947748	116949548	EcoRI-379				
116959563	116964868	EcoRI-386	0.058143	0.2227	0.010776	0.0273
116965514	116975353	EcoRI-388	0.137084	0.2715	0.018633	0.0126
116991712	116992706	EcoRI-393	0.054297	0.3156	0.022269	0.0704
117003291	117005125	EcoRI-400	0.101937	0.3829	0.02802	0.0251
117013018	117016269	EcoRI-406	0.088219	0.6045	0.02756	0.0831
117035191	117039940	EcoRI-415	0.102279	0.1227	0.046829	0.0307
117054926	117057582	EcoRI-422				
117078729	117088313	EcoRI-432	0.191792	0.4739	0.01247	0.0166
117093828	117096070	EcoRI-437	0.500427	0.2866	0.056936	0.0667
117111635	117114566	EcoRI-440	0.908109	0.883	0.016413	0.0255
117127744	117131345	EcoRI-442	0.323102	0.4276	0.101611	0.0245

Start	**End**	**name**	**HepG2**	**HeLa**	**SE HepG2**	**SE HeLa**
116804746	116807423	EcoRI-331	0.013254	0.3049	0.009316	0.0539
116822606	116827113	EcoRI-338	0.053686	0.5933	0.039055	0.0969
116833484	116835882	EcoRI-343	0.062249	0.3652	0.027459	0.0814
116849034	116854362	EcoRI-350	0.053271	0.2769	0.02409	0.1268
116883208	116885279	EcoRI-359	0.073231	0.5674	0.025453	0.0462
116885280	116890156	EcoRI-360	0.176969	0.528	0.038786	0.0716
116898039	116901211	EcoRI-363	0.072794	0.5461	0.013026	0.1069
116903489	116907872	EcoRI-368	0.046716	0.4706	0.016642	0.0444
116920524	116925167	EcoRI-371	0.125269	0.4505	0.048806	0.0904
116947748	116949548	EcoRI-379	0.043544	0.2709	0.021189	0.0765
116959563	116964868	EcoRI-386	0.03143	0.1385	0.02705	0.0375
116965514	116975353	EcoRI-388	0.057233	0.2927	0.006505	0.0703
116991712	116992706	EcoRI-393	0.084916	0.2593	0.019332	0.0882
117003291	117005125	EcoRI-400	0.040701	0.3956	0.010032	0.0743
117013018	117016269	EcoRI-406	0.064345	0.4983	0.02095	0.0801
117035191	117039940	EcoRI-415	0.017003	0.3872	0.008242	0.0782
117054926	117057582	EcoRI-422	0.11625	0.193	0.013081	0.0327
117078729	117088313	EcoRI-432	0.129096	0.6628	0.02938	0.0222
117093828	117096070	EcoRI-437	0.345992	0.6406	0.025566	0.1492

117111635	117114566	EcoRI-440	0.688112	0.9445	0.028563	0.0789
117127744	117131345	EcoRI-442	0.225562	0.7307	0.041095	0.12

BsrGI Fine Mapping

Start	End	name	GM06990	Caco2	SE GM06990	SE Caco2
Anchor at Promoter, looking at Element III						
117014031	117014185	EMS38	0.24	0.3513	0.037574	0.0574
117014186	117015473	EMS39	0.28093	0.4803	0.041359	0.054
117016422	117016451	EMS40	0.288757	0.606	0.04583	0.0556
117018151	117018180	EMS41	0.261154	0.4173	0.035647	0.0512
Anchor at Promoter, looking at Element IV						
117101305	117109806	EMS27	0.146282	0.2666	0.021951	0.031
117109806	117111366	EMS3	0.205144	0.4522	0.053743	0.0734
117111366	117118556	EMS20	0.201013	0.3901	0.029906	0.0484
Anchor at Element III, looking at Element IV						
117101305	117109806	EMS27	0.114342	0.3771	0.017601	0.0294
117109806	117111366	EMS3	0.117414	0.874	0.023627	0.0895
117111366	117118556	EMS20	0.076762	0.2799	0.020966	0.0275

Table A.3: Primers Used in qRT-PCR. Gene names are listed in italics with their abbreviation in bold lettering. F and R refer to forward and reverse primers. Sequences for each primer are listed followed by the exon number and genome location (hg18).

Primer Name	Primer Sequence	Exon #	Genome Coordinates (hg18)
TES: *testin isoform 1*			
TES_F_A	GCCCCTTGTTTAAAATGCAA	2	115661854-115661873
TES_R_A	TGCTCAAGAGGACATCATGC	3	115676346-115676365
CAV2: *caveolin 2 isoform a and b*			
CAV2_F_A	GGCTCAACTCGCATCTCAAG	1	115927202-115927221
CAV2_R_A	CAGGAACACCGTCAGGAACT	2	115927656-115927675
CAV1: *caveolin 1*			
CAV1_F_A	GAGCTGAGCGAGAAGCAAGT	2	115953887-115953906
CAV1_R_A	CAAATGCCGTCAAAACTGTG	3	115986275-115986294
MET: *met proto-oncogene isoform a precursor*			
MET_F_B	CCAATGACCTGCTGAAATTG	11	116197041-116197060
MET_R_B	CTTTTCCAAGGACGGTTGAA	12	116198805-116198824
CAPZA2: *capping protein (actin filament) muscle Z-line*			
CAPZA2_F_A	GAAGGAGGCAACTGATCCAA	5	116331553-116331572
CAPZA2_R_A	GCTTGGAACTGATGGCTTTC	6	116333601-116333620
ST7: *suppression of tumorigenicity 7 isoform b*			
ST7_F_A	TTCCAGTAACGGGGACTCAG	3	116546967-116546986
ST7_R_A	TGGATTTCGCCATACTTTGC	4	116557086-116557105
WNT2: *wingless-type MMTV integration site family*			
WNT2_F_B	GTGGATGCAAAGGAAAGGAA	3	116742416-116742435
WNT2_R_B	AGCCAGCATGTCCTGAGAGT	4	116725090-116725109
ASZ1: *ankyrin repeat, SAM and basic leucine zipper*			
ASZ1_F_B	CACGTCAGGGTCATAAA	6	116812115-116812137
ASZ1_R_B	GCTGTTGAAGTTTTCCTTCCA	7	116810346-116810366
CFTR: *cystic fibrosis transmembrane conductance regulator*			
CFTR2-3F	CCCTTCTGTTGATTCTGCTG	2	116931609-116931628
CFTR2-3R	AAGGGCATTAATGAGTTTAGGA	3	116936357-116936378
CTTNBP2: *cortactin binding protein 2*			
CTTNBP2_F_C	AAAATGGCTTCACACCCTTG	6	117210181-117210200
CTTNBP2_R_C	TGTCTGTCCTCCATCAGCAG	7	117207818-117207837
LSM8: *U6 snRNA-associated Sm-like protein LSm8*			
LSM8_F_A	CAGCTCTTCACAGGGGGTAG	3	117615633-117615652
LSM8_R_A	CTGCTCGAATATTCCCCAAA	4	117619247-117619000

ANKRD7: *ankyrin repeat domain 7 isoform b*

ANKRD7_F_A	ACCTTTGCACCTAGCCTGTG	2	117661724-117661743
ANKRD7_R_A	ATCTGGGTCTGCACCAAAGT	3	117662034-117662053

HPRT1: *Chromosome X: hypoxanthine phosphoribosyltransferase 1*

HPRT2-3F	TGAGGATTTGGAAAGGGTGT	2	133435114-133435133
HPRT2-3R	TAATCCAGCAGGTCAGCAAA	3	133436965-133436984

Table A.4: Yeast One-hybrid Bait:Prey Interactions. This table lists all observed Y1H bait:prey interactions. III DHS and IV DHS refer to the minimal enhancer elements found in the luciferase assay. PFA-E represent the promoter fragments A,B,C,D and E that were used in the assay. At least three integrations were tested for each bait. ”Round” refers to which experiment tested which bait. “Coordinate” shows the location on the Y1H plate gave the positive signal. This coordinate was then cross-referenced with the master list to determine which hTF gave the positive signal. Each bait:prey interaction was scored by eye on the indicated days. The number refers to the number of yeast that gave a positive result (from 0-4, since there were 4 technical replicates on each plate). The letter refers to the strength of the positive interaction. VW= very weak, W = weak, M = medium, S = strong and VS = very strong.

Bait Name	Integrant	Round	Coordinate	TF Name	Day1	Day3/4	Day5	Day6/7
III DHS	1	1	3_H11	TCF7L2	4W	4M	4M	4S
III DHS	1	1	7_H02	LEF1	4W	4M	4M	4S
III DHS	1	1	12_B12	LEF1	4VW	4M	4M	4M
III DHS	1	1	2_C03	SMAD9	-	4VW	4VW	-
III DHS	2	1	3_H11	TCF7L2	4VW	4W	4W	4W
III DHS	2	1	7_H02	LEF1	4VW	4W	4W	4W
III DHS	2	1	12_B12	LEF1	4VW	4W	4W	4W
III DHS	2	1	15_C12	GRHL1	-	2VW	2VW	2VW
III DHS	4	2	3_H11	TCF7L2	4VW	4W	4M	4M
III DHS	4	2	6_H02	ZNF366	4VW	-	-	-
III DHS	4	2	7_H02	LEF1	-	4M	4S	4S
III DHS	4	2	12_B12	LEF1	-	4W	4W	4W
IV DHS	1	1	1_A02	ESRRB	-	4S	4VS	4VS
IV DHS	1	1	2_B06	THRB	4VW	4M	4M	4S
IV DHS	1	1	5_A09	ESRRG	-	4VS	4VS	4VS
IV DHS	1	1	7_D01	NR5A1	4VW	4S	4VS	4VS
IV DHS	1	1	8_F12	NR5A2	3W	4W	4S	4S
IV DHS	1	1	10_E03	MIXED (ERROR)	2VW	2VS	2VS	2VS
IV DHS	1	1	14_B09	GCM1	-	-	4VW	4VW
IV DHS	1	1	14_F02	ZIC3	-	4W	4M	4VS
IV DHS	1	1	19_A08	NR5A2	-	3W	3W	3M
IV DHS	1	1	19_B10	ZIC1	-	-	2W	2S

IV DHS	1	1	19_B11	GCM1	-	-	4VW	4VS
IV DHS	1	1	19_D07	DMTF1	-	-	4W	4W
IV DHS	1	1	19_E01	ZIC3	4W	4VS	4VS	4VS
IV DHS	2	2	1_A02	ESRRB				
IV DHS	2	2	2_B06	THRB				
IV DHS	2	2	5_A09	ESRRG				
IV DHS	2	2	7_D01	NR5A1				
IV DHS	2	2	8_F12	NR5A2				
IV DHS	2	2	14_F02	ZIC3	-	-	4VW	4W
IV DHS	2	2	19_A08	NR5A2	-	4VW	4W	4W
IV DHS	2	2	19_B10	ZIC1	-	3VW	3W	3W
IV DHS	2	2	19_E01	ZIC3	-	4W	4M	4W
IV DHS	3	3	01_A02	ESRRB	-	4W	4S	4VS
IV DHS	3	3	01_A03	FOXC1	-	3S	3VS	3VS
IV DHS	3	3	02_B06	THRB	-	4VW	-	4VW
IV DHS	3	3	05_A09	ESRRG	-	4W	4M	4S
IV DHS	3	3	05_F06	ZNF471	-	-	2VW	3VW
IV DHS	3	3	08_F12	NR5A2	-	4VW	-	4W
IV DHS	3	3	10_E03	MIXED (ERROR)	-	4M	4VS	4VS
IV DHS	3	3	10_C07	TEAD2	-	-	4W	4W
IV DHS	3	3	11_C03	TEAD2	-	-	4W	4W
IV DHS	3	3	14_F02	ZIC3	-	4W	4W	4M
IV DHS	3	3	15_D07	REL	-	3S	3VS	3VS
IV DHS	3	3	19_A08	NR5A2	-	3W	3W	3W
IV DHS	3	3	19_E01	ZIC3	-	4VS	4VS	4VS
IV DHS	3	3	19_H08		-	-	-	2VW
PFA	1	3			-	-	-	-
PFA	2	3	15_H04	GRHL2	-	-	3M	3M
PFA	2	3	19_E01	ZIC3	-	-	4M	4M
PFA	3	3	14_F02	ZIC3	-	-	4VW	4W
PFA	3	3	15_H04	GRHL2	-	-	3W	3W
PFA	3	3	19_E01	ZIC3	-	-	4W	4W
PFB	1	2	10_E03	MIXED (ERROR)	-	2W	2S	2S
PFB	1	2	19_E01	ZIC3	-	4W	4M	4S
PFB	2	3	01_A02	ESRRB	-	-	-	3VW
PFB	2	3	03_F01	ZBTB7B	-	-	-	4VW
PFB	2	3	04_A01	SPIB	-	-	-	3VW

PFB	2	3	05_A09	ESRRG	-	-	4VW	4VW
PFB	2	3	05_C01	DMBX1	-	-	2VW	3W
PFB	2	3	05_F01	RXRG	-	4VW	4W	4W
PFB	2	3	06_A12	NKX2-2	-	-	-	4VW
PFB	2	3	10_E03	MIXED(ERR0R)	-	2W	2W	2M
PFB	2	3	14_C12	ZBTB10	-	-	-	4VW
PFB	2	3	14_F02	ZIC3	-	4VW	-	-
PFB	2	3	15_D07	REL	-	4VW	-	4W
PFB	2	3	19_E01	ZIC3	4W	4VS	4VS	4VS
PFB	3	3	01_F05	LBX2	-	-	4VW	4VW
PFB	3	3	02_G12	MAX	-	-	2VW	-
PFB	3	3	04_A07	TEAD4	-	-	4VW	-
PFB	3	3	05_A09	ESRRG	-	-	4VW	-
PFB	3	3	05_D03	DLX4	-	-	3VW	4VW
PFB	3	3	05_F01	RXRG	-	4W	4W	4W
PFB	3	3	05_F06	ZNF471	-	4W	4W	4M
PFB	3	3	10_E03	MIXED(ERR0R)	-	4W	2VS	2VS
PFB	3	3	10_E08	NR1I2	-	4W	4S	4M
PFB	3	3	11_G10	MAFG	-	-	4VW	-
PFB	3	3	13_D06	KLF3	-	4VW	4W	4W
PFB	3	3	14_F02	ZIC3	-	4S	4VS	4VS
PFB	3	3	15_D07	REL	-	4S	4VS	4VS
PFB	3	3	19_A04	ZNF710	-	4VW	4W	4VW
PFB	3	3	19_B11	GCM1	-	-	4M	4M
PFB	3	3	19_E01	ZIC3	4W	4VS	4VS	4VS
PFB	3	3	20_D02		-	-	4W	4VW
PFC	1	2	10_C06	SMAD4	4VW	4S	4M	4S
PFC	1	2	10_C07	TEAD2	-	-	4W	4W
PFC	1	2	10_E03	MIXED (ERROR)	-	3VS	3VS	3VS
PFC	1	2	11_C03	TEAD2	-	-	4W	4W
PFC	1	2	12_A11	EBF1	-	4VW	-	4VW
PFC	1	2	14_B09	GCM1	4VW	4W	4W	4M
PFC	1	2	14_F02	ZIC3	-	4W	4M	4S
PFC	1	2	15_F01	GRHL2	-	-	2M	2S
PFC	1	2	15_H04	GRHL2	-	4M	4M	4S
PFC	1	2	19_A03	TFEB	-	-	-	4W
PFC	1	2	19_B10	ZIC1	-	3M	3M	3W
PFC	1	2	19_B11	GCM1	-	4VW	4W	4W
PFC	1	2	19_E01	ZIC3	4VW	4VS	4VS	4VS

PFC	2	2	1_A08	HES5	-	4W	4W	4W
PFC	2	2	1_A09	LHX1	-	4W	4W	4W
PFC	2	2	3_C10	ZDHHC7	-	4VW	4W	4W
PFC	2	2	4_C09	TFAM	-	4VW	4W	4W
PFC	2	2	5_A07	SATB2	-	4W	4M	4M
PFC	2	2	5_B12	ELK4	-	4VW	4M	3M
PFC	2	2	5_E05	ZNF207	-	-	4VW	4W
PFC	2	2	5_H12	SNAI1	-	4VW	4M	4M
PFC	2	2	6_C09	RFX4	-	4W	4S	4S
PFC	2	2	7_H03	KLF17	-	-	3W	3W
PFC	2	2	9_H10	HEY2	-	4VW	-	4W
PFC	2	2	10_C06	SMAD4	4W	4S	4S	4S
PFC	2	2	10_C07	TEAD2	-	4VW	4M	4M
PFC	2	2	10_E08	NR1I2				
PFC	2	2	11_C03	TEAD2	-	-	4M	4W
PFC	2	2	12_A11	EBF1	-	4VW	4W	4W
PFC	2	2	13_D04	TEAD2	-	-	4VW	4W
PFC	2	2	14_B09	GCM1	4VW	4W	4M	4M
PFC	2	2	14_F02	ZIC3	-	4S	4VS	4VS
PFC	2	2	15_F01	GRHL2	-	2W	2VS	2S
PFC	2	2	15_H04	GRHL2	-	3VW	3M	3M
PFC	2	2	19_B10	ZIC1	-	-	2W	2W
PFC	2	2	19_B11	GCM1	-	-	-	4VW
PFC	2	2	19_E01	ZIC3	4VW	4S	4VS	4S
PFC	3	3	01_A04	DLX5	-	-	-	4VW
PFC	3	3	01_A07	IRX5	-	-	4W	4VW
PFC	3	3	01_A08	HES5	-	4VW	4W	4W
PFC	3	3	01_A09	LHX1	-	4VW	4W	4W
PFC	3	3	01_C10	BHLHB2	-	-	-	4VW
PFC	3	3	02_H07	GABPA	-	-	-	4VW
PFC	3	3	03_A02	ERF	-	-	4W	-
PFC	3	3	03_A11	ZNF18	-	-	4VW	-
PFC	3	3	03_C10	ZDHHC7	-	-	4VW	4VW
PFC	3	3	03_F01	ZBTB7B	-	-	4VW	4VW
PFC	3	3	04_A01	SPIB	-	-	4W	4W
PFC	3	3	04_C09	TFAM	-	-	-	4VW
PFC	3	3	04_H04	PRRX1	-	3VW	4VW	4W
PFC	3	3	05_A07	SATB2	-	4VW	4VS	4VS
PFC	3	3	05_A09	ESRRG	-	-	4W	4VW
PFC	3	3	05_B07	MXD1	-	-	4VW	4VW

PFC	3	3	05_F02	USF2	-	-	4VW	4VW
PFC	3	3	05_G01	PAX5	-	-	4W	4VW
PFC	3	3	06_C09	RFX4	-	4W	4VS	4VS
PFC	3	3	06_G01	ZNF221	-	-	-	4VW
PFC	3	3	07_G01	ALX1	-	-	4W	-
PFC	3	3	07_H03	KLF17	-	-	4W	4W
PFC	3	3	10_B10	WT1	-	-	4W	4W
PFC	3	3	10_C06	SMAD4	4VW	4M	4VS	4VS
PFC	3	3	10_C07	TEAD2	-	-	4W	4M
PFC	3	3	10_E03	MIXED (ERROR)	3VW	4S	-	-
PFC	3	3	11_C03	TEAD2	-	-	4W	4M
PFC	3	3	14_B09	GCM1	4VW	4W	4M	4M
PFC	3	3	14_C12	ZBTB10	-	-	-	4VW
PFC	3	3	14_F02	ZIC3	2VW	4W	4VS	4VS
PFC	3	3	15_F01	GRHL2	-	4VW	4VS	4VS
PFC	3	3	15_H04	GRHL2	-	2W	3S	3VS
PFC	3	3	17_D12	RUVBL1	-	-	-	4VW
PFC	3	3	17_H10	TGIF1	-	-	-	4VW
PFC	3	3	19_A03	TFEB	-	-	4W	4W
PFC	3	3	19_B10	ZIC1	3VW	4W	4VS	3VS
PFC	3	3	19_B11	GCM1	4VW	4W	4S	4S
PFC	3	3	19_D11	TFAP4	-	-	-	4W
PFC	3	3	19_E01	ZIC3	4M	4VS	4VS	4VS
PFC	3	3	19_E05	ZNF3	-	3VS	-	-
PFC	3	3	19_G02	NR1H4	-	-	-	4VW
PFC	3	3	20_B06		-	-	4VW	-
PFC	3	3	20_B08		-	-	4W	-
PFD	1	3	01_A04	DLX5	-	4VW	-	-
PFD	1	3	01_A05	POU4F1	-	4VW	-	-
PFD	1	3	01_A06	LHX5	-	4VW	-	-
PFD	1	3	01_A07	IRX5	-	4VW	-	-
PFD	1	3	01_B07	CDX2	-	4VW	-	-
PFD	1	3	01_B12	HOXD9	-	4VW	4W	-
PFD	1	3	01_C06	SOX21	-	4M	-	4M
PFD	1	3	01_E03	NR2F2	-	4M	4S	4S
PFD	1	3	01_F07	TGIF2LX	-	-	-	4W
PFD	1	3	02_C06	THAP10	-	-	4M	-
PFD	1	3	02_D07	MYOG	-	4W	4W	4W

PFD	1	3	02_D08	SPIC	-	4W	4W	4W
PFD	1	3	02_E08	NFIA	-	4W	4M	4M
PFD	1	3	02_G02	ZNF581	-	4W	4M	4M
PFD	1	3	02_H03	ZNF670	-	4W	4M	-
PFD	1	3	02_H07	GABPA	-	-	-	4W
PFD	1	3	03_B02	MAZ	-	4M	4S	4S
PFD	1	3	03_A05	ZNF20	-	4VW	-	-
PFD	1	3	03_A11	ZNF18	-	4VW	-	-
PFD	1	3	03_C08	SOX6	-	4W	4VW	4VW
PFD	1	3	03_F01	ZBTB7B	-	4W	-	-
PFD	1	3	04_A01	SPIB	-	4VW	-	4M
PFD	1	3	04_F01	MEIS3	-	-	4M	-
PFD	1	3	05_A07	SATB2	-	4S	4VS	4VS
PFD	1	3	05_B04	POU4F3	-	2S	4W	4M
PFD	1	3	05_C08	FOXA3	-	4W	4VW	4W
PFD	1	3	05_C10	PLAGL1	-	4W	-	4W
PFD	1	3	05_D05	ZNF398	-	-	4W	4W
PFD	1	3	05_E04	CEBPB	-	4W	4W	4W
PFD	1	3	05_G01	PAX5	-	4M	4W	-
PFD	1	3	06_B06	HOXC8	-	4W	4VW	-
PFD	1	3	06_C09	RFX4	-	-	-	4W
PFD	1	3	10_B08	KLF4	-	4M	4M	4M
PFD	1	3	10_B10	WT1	-	4S	4S	4S
PFD	1	3	10_D04	RFX2	-	4W	4M	4VW
PFD	1	3	10_D09	PLAGL1	-	4M	4M	4M
PFD	1	3	10_E03	MIXED	-	4M	4M	4S
PFD	1	3	10_G06	PURA	-	4VW	4W	4W
PFD	1	3	11_D04	MYCL1	-	4W	-	-
PFD	1	3	12_D04	KLF4	-	-	4M	4VW
PFD	1	3	13_D06	KLF3	4W	4M	4S	4S
PFD	1	3	14_C12	ZBTB10	4M	4VS	4VS	4VS
PFD	1	3	14_E06	SP3	4W	4M	4S	4S
PFD	1	3	14_H05	SP4	4W	4M	4M	4M
PFD	1	3	15_A10	NFATC4	-	-	-	4W
PFD	1	3	15_B11	DLX1	-	4W	4W	4M
PFD	1	3	15_B12	SPI1	-	4W	4W	4VW
PFD	1	3	15_D03	HP1BP3	-	-	4M	-
PFD	1	3	15_F10	ZNF653	-	-	4VW	-
PFD	1	3	16_B10	RBMS1	-	4W	4W	4M
PFD	1	3	16_D03	DAZAP1	-	4M	-	4S

PFD	1	3	16_E11	AKAP10	-	-	4VW	-
PFD	1	3	16_E12	DDX43	4W	-	-	-
PFD	1	3	17_B10	PTPMT1	-	4W	4VW	4VW
PFD	1	3	18_B02	RBMS2	-	4S	4S	4S
PFD	1	3	18_D05	PDE6H	-	4M	4W	4S
PFD	1	3	19_B04	SIX3	-	4W	4W	4VW
PFD	1	3	19_B07	HOXA6	-	4W	4VW	-
PFD	1	3	19_B08	POU1F1	-	4W	4VW	-
PFD	1	3	19_B09	ELF2	-	-	4VW	-
PFD	1	3	19_B11	GCM1	-	4W	4VW	4VW
PFD	1	3	19_E01	ZIC3	4S	4VS	4VS	4VS
PFD	2	3	01_A04	DLX5	-	4W	-	-
PFD	2	3	01_A05	POU4F1	-	4W	-	-
PFD	2	3	01_A06	LHX5	-	4W	-	-
PFD	2	3	01_A07	IRX5	-	4W	-	-
PFD	2	3	01_B07	CDX2	-	-	4W	-
PFD	2	3	01_B12	HOXD9	-	4M	-	-
PFD	2	3	01_C04	FOXH1	-	4M	4M	4W
PFD	2	3	01_C06	SOX21	-	4M	4M	4S
PFD	2	3	01_E03	NR2F2	-	4M	4S	4S
PFD	2	3	01_F07	TGIF2LX	-	-	4W	-
PFD	2	3	02_B10	ZNF174	-	4M	-	-
PFD	2	3	02_C07	ZNF10	-	-	-	4W
PFD	2	3	02_D07	MYOG	-	4W	4W	-
PFD	2	3	02_D08	SPIC	-	4M	4W	-
PFD	2	3	02_D09	ELF5	-	-	-	4S
PFD	2	3	02_E08	NFIA	-	4M	4S	4S
PFD	2	3	02_G02	ZNF581	-	-	4W	-
PFD	2	3	02_H03	ZNF670	-	4M	4W	-
PFD	2	3	02_H07	GABPA	-	-	4W	-
PFD	2	3	03_B02	MAZ	4W	4M	4S	4S
PFD	2	3	03_B10	ZNF322A	-	4M	-	-
PFD	2	3	03_C08	SOX6	-	-	-	4W
PFD	2	3	03_F01	ZBTB7B	-	4M	-	-
PFD	2	3	04_H04	PRRX1	3W	3M	-	-
PFD	2	3	05_A07	SATB2	-	4W	-	4S
PFD	2	3	05_A09	ESRRG	-	4W	-	-
PFD	2	3	05_B04	POU4F3	-	4M	-	-
PFD	2	3	05_C08	FOXA3	-	-	4W	4W
PFD	2	3	05_D03	DLX4	4VW	4M	4W	4S

PFD	2	3	05_D05	ZNF398	4VW	4M	4W	4S
PFD	2	3	05_E04	CEBPB	-	-	4W	4W
PFD	2	3	05_F07	ZNF16	-	4M	-	4W
PFD	2	3	05_G01	PAX5	-	4W	-	-
PFD	2	3	07_F12	CAMTA2	-	4M	-	-
PFD	2	3	08_E09	FOXD4L6	-	-	4W	-
PFD	2	3	08_H03	ZBTB10	-	4W	-	-
PFD	2	3	10_B08	KLF4	4VW	4M	4S	4M
PFD	2	3	10_B10	WT1	4VW	4M	4S	4VS
PFD	2	3	10_D04	RFX2	-	4M	4S	4M
PFD	2	3	10_D09	PLAGL1	-	4M	4S	4S
PFD	2	3	10_F02	KLF7	4VW	4M	4S	4M
PFD	2	3	12_C01	WT1	4VW	-	4S	4S
PFD	2	3	12_D04	KLF4	-	4M	4S	4M
PFD	2	3	12_E02	KLF12	-	-	4S	4M
PFD	2	3	12_E03	ZBTB25	-	4M	-	-
PFD	2	3	13_D06	KLF3	4VW	4M	4M	4S
PFD	2	3	13_G03	KLF4	-	4VW	-	-
PFD	2	3	14_C12	ZBTB10	4VW	4S	4VS	4VS
PFD	2	3	14_E06	SP3	4VW	4M	4M	4S
PFD	2	3	14_H05	SP4	4VW	4M	4M	4S
PFD	2	3	15_B12	SPI1	-	4VW	-	-
PFD	2	3	16_B10	RBMS1	-	4M	4M	4W
PFD	2	3	16_D03	DAZAP1	-	4W	4W	4W
PFD	2	3	16_E12	DDX43	4VW	-	-	-
PFD	2	3	18_B02	RBMS2	-	4M	4M	4S
PFD	2	3	18_B05	BAX	-	4W	-	-
PFD	2	3	19_B04	SIX6	-	4M	4VW	4M
PFD	2	3	19_B07	HOXA6	-	4W	4VW	4W
PFD	2	3	19_B08	POU1F1	-	4W	4VW	4W
PFD	2	3	19_B09	ELF2	-	4W	4VW	4W
PFD	2	3	19_B11	GCM1	4W	-	4VW	-
PFD	2	3	19_D05	ZNF707	-	-	-	4W
PFD	2	3	19_E01	ZIC3	4W	4VS	4VS	4VS
PFD	2	3	19_G08		4VW	-	-	-
PFD	2	3	19_H08		-	-	4VW	4W
PFD	2	3	20_B01		-	4M	4W	-
PFD	2	3	20_D03		4VW	-	-	-
PFD	2	3	20_D06		-	-	-	4W
PFD	3	3	01_B07	CDX2	-	4VW	-	-

PFD	3	3	01_D05	ZNF498	-	4VW	-	-
PFD	3	3	01_E03	NR2F2	-	4S	4VS	4S
PFD	3	3	01_E09	STAT4	-	4W	-	-
PFD	3	3	01_F07	TGIF2LX	-	4M	-	4W
PFD	3	3	02_D07	MYOG	-	4W	-	-
PFD	3	3	02_D08	SPIC	-	4W	-	-
PFD	3	3	02_E08	NF1A	-	4S	4S	4S
PFD	3	3	02_G02	ZNF581	-	4W	4M	-
PFD	3	3	02_H07	GABPA	-	4M	4M	-
PFD	3	3	03_B02	MAZ	-	4S	4M	4S
PFD	3	3	03_D06	NFKB1	-	4VW	-	-
PFD	3	3	03_F01	ZBTB7B	-	4W	-	-
PFD	3	3	04_D05	HNF4A	-	4VW	-	-
PFD	3	3	04_H04	PRRX1	4VW	-	-	-
PFD	3	3	05_A07	SATB2	-	4VS	4VS	4VS
PFD	3	3	05_A09	ESRRG	-	4M	4S	4M
PFD	3	3	05_B04	POU4F3	-	4W	4M	4M
PFD	3	3	05_C08	FOXA3	-	-	4W	-
PFD	3	3	05_D03	DLX4	-	4W	4W	4W
PFD	3	3	05_D09	USF1	-	-	-	4M
PFD	3	3	05_E04	CEBPB	-	4W	4W	4W
PFD	3	3	05_G01	PAX5	-	4M	4M	4M
PFD	3	3	10_B08	KLF4	4VW	4M	4S	4M
PFD	3	3	10_B10	WT1	4VW	4M	4S	4S
PFD	3	3	10_C01	ZBTB1	4VW	-	-	-
PFD	3	3	10_D02	ZNF205	-	-	-	4W
PFD	3	3	10_D04	RFX2	-	4W	4M	-
PFD	3	3	10_D09	PLAGL1	-	4M	4M	4M
PFD	3	3	10_E03	MIXED	-	4S	4S	4S
PFD	3	3	10_F02	KLF7	4VW	4M	-	4M
PFD	3	3	12_C01	WT1	-	4S	-	4S
PFD	3	3	12_D04	KLF4	-	4W	4M	4W
PFD	3	3	12_E02	KLF12	-	-	4W	-
PFD	3	3	12_G01	HEY2	-	-	4W	-
PFD	3	3	13_D06	KLF3	4VW	4S	4VS	4VS
PFD	3	3	14_C12	ZBTB10	4VW	4M	4VS	4S
PFD	3	3	14_E06	SP3	4VW	4S	4VS	4VS
PFD	3	3	14_H05	SP4	4VW	4M	4M	4S
PFD	3	3	16_D03	DAZAP1	-	4M	4M	4S
PFD	3	3	16_E12	DDX43	4VW	-	-	-

PFD	3	3	18_B02	RBMS2	-	-	4S	4M
PFD	3	3	18_B05	BAX	-	-	-	4M
PFD	3	3	18_D05	PDE6H	-	-	4M	-
PFD	3	3	19_A03	TFEB	-	4W	4W	-
PFD	3	3	19_B04	SIX6	-	4W	4M	4W
PFD	3	3	19_B07	HOXA6	-	4W	4M	4W
PFD	3	3	19_B08	POU1F1	-	4W	4M	4W
PFD	3	3	19_B09	ELF2	-	4W	4M	4W
PFD	3	3	19_D05	ZNF707	-	4W	-	-
PFD	3	3	19_E01	ZIC3	4M	4VS	4VS	4VS
PFD	3	3	19_F01	IKZF3	-	4VW	4W	4W
PFD	3	3	19_H08		-	4W	4W	4W
PFD	3	3	20_B01		-	-	4S	4M
PFD	3	3	20_D06		-	4M	4M	-
PFE	1	3	-					
PFE	2	3	-					
PFE	3	3	-					

Table A.5: Name, Location and Sequence of all 5C Probes. Probe names are listed on the left with their chromosomal positon (hg18). The sequence of each probe is given on the right.

5C Primer List

PRIMER_NAME, LOCATION (hg18)	SEQUENCE
FOR_1 chr7:115597757-115598282	GTCTGGATAAGGAGAAAGTTAATTTACATACAGCACAAAG
FOR_2 chr7:115598283-115604576	ATGTTTGCCTTTATCAGTGACGGATATCAACATCATGAAG
FOR_5 chr7:115608801-115614962	GGGCAAATGATAGGCTAACCTGTAAACTTAAAAGGAGAAG
FOR_7 chr7:115623541-115625042	ACTCAAAGTGCCTTGCGATACTATTTGAAAGTTTATCAAG
FOR_8 chr7:115625043-115633971	AACTGGCACCAGCTTCAGCTAATTGAATCAGAAGTTTAAG
FOR_11 chr7:115644335-115646444	ATTATTTTATCAAGGAAACACCTCCTATAGAATAAAGAAG
FOR_12 chr7:115646445-115648831	GATCAGTAGAATGTGGTAGAAGTGGTGTGTGAATCCTAAG
FOR_15 chr7:115660029-115660112	TTTGAAAATTGGTAATATGGACATGGCAAAGATGATAAAG
FOR_19 chr7:115681634-115684212	ATGTGTCATAGGTAAGACAGGTAAAATTACCACTTCAAAG
FOR_20 chr7:115684213-115685035	TCTTTCCATAGCTTTTCAAATGTGAAATCATTTTTGGAAG
FOR_21 chr7:115685036-115687308	TGATTCATTCTCAAGGAATTTCCAACACATAGCGTTTAAG
FOR_22 chr7:115687309-115691270	TCTGGTAATGAGATCCATGACCCTTCCACACCCCCTCAAG
FOR_24chr7:115692343-115693067	TGGGGATAGCATGGCAGACACTGGCTAGATGATCTCTAAG
FOR_25 chr7:115693068-115699683	AATAGGGAAAAAATCACCATGTCTCTCCTGCTACGGAAAG
FOR_27 chr7:115711032-115711862	GAAATTGTGGTATGTCCTTGATCCTGGAGCATTCAGGAAG
FOR_32 chr7:115725060-115730982	GAGCAAGCAAATGAAGCTGAATTCAATATGGATTATAAAG
FOR_33 chr7:115730983-115733617	TCACGCAATCAGGGCAGGCAGTGAAAGAGAATTTCAGAAG
FOR_35 chr7:115734616-115736008	AAATTTTTGAGCAGTTATGTGATATAGTGAATACATTAAG
FOR_36 chr7:115736009-115736220	TCAATAAATCTTGTAGTTTTCTTGGACCAGAGGAAATAAG
FOR_37 chr7:115736221-115739597	TTTTCGTTATTTCACAATCTTTATCTTGATCATCAGGAAG
FOR_38 chr7:115739598-115745973	CAGGAGGAAACCTTGAAAAAGGGGTACATGTTAGGGAAAG
FOR_39 chr7:115745974-115750038	ACATCATAAGGAATGTGGCTTTTCAGGGTGAGATAAGAAG
FOR_41 chr7:115750990-115751652	TGACTATTGTAGCTAGAGAGGAGAAGCCAATGCCTGTAAG
FOR_42 chr7:115751653-115754768	TTGCTTATCTCAGTAAATTGTTAACATGTAGAAATGAAAG
FOR_43 chr7:115754769-115756853	GTAGAAAAGATATTAGTGAAGATAGACTAGCCTTGAGAAG
FOR_45 chr7:115758299-115759208	TAGAAAGATAGTTGCGCTGATACAAAGAAAATACAAGAAG
FOR_47 chr7:115760111-115764976	CTAATTAAATTGTTTGAGATAGATCTGAAGACAGCAGAAG
FOR_48 chr7:115764977-115766732	CAGTGCTTTATGTTATTGGGATTAAAAGAATTAGTTAAAG
FOR_49 chr7:115766733-115766920	TTAAATACTAACACTCAAACCTGAAAGTTGTCTTTAAAAG
FOR_50 chr7:115766921-115767784	AGCACAGCAATTAAATGGAATTCCACATTTTTAAGGAAAG
FOR_51 chr7:115767785-115770527	TCAAGTCCAAGAAGTCATTGCCAAGTCCAATGTCATAAAG
FOR_52 chr7:115770528-115771572	GGTGAAAGGACATGCTTGTGTTTTGATCTCAGAGCAAAAG
FOR_54 chr7:115775576-115782157	GCTAATGTGAGGGAGCAAGACTTGTCAGGTTCTGACTAAG
FOR_55 chr7:115782158-115787377	ACTTATATGATTTATGAGTTTTGCGTATTGCTTAGGAAAG
FOR_56 chr7:115787378-115792901	AGCTGCAAAATCAGGGAAACCATTAGTAACAACGTGAAAG

FOR_57 chr7:115792902-115797499	GTCTCTCTGGCTGTTAGGATCAGTTATGCTGGGGTTGAAG
FOR_60 chr7:115802318-115805492	GCATAATATCAGATTTAGGTAGAGATTAAGGAGATAAAAG
FOR_61 chr7:115805493-115806404	TTTTCTCATGAGTGTCTAGAAACCAGTCATGTGATTAAAG
FOR_62 chr7:115806405-115807508	GGGACTCTGAGTAATCTAAAATAAGAGCTGGCAACAGAAG
FOR_63 chr7:115807509-115807867	AGCTAATGGCACATTTGTACTTTTCCATTTTTATGTCAAG
FOR_65 chr7:115808565-115809866	TTTTAAACTAGAAGAAACTAAGATAATGCAAATTAACAAG
FOR_66 chr7:115809867-115810396	TCCTCAATGCTATGTAGCAGTGTCTAGCACATAGGAAAAG
FOR_67 chr7:115810397-115811132	AGGTGACTTCAGTCATGTCTTTTAATCAGGAAAACAAAAG
FOR_68 chr7:115811133-115812539	GATCCTTTAAAAATTCTTCACAAGAGGAGTCAAAGGAAAG
FOR_69 chr7:115812540-115813432	TTACAGTTGTTCATTCTCCAGCATCTTGATCTTGACAAAG
FOR_70 chr7:115813433-115819798	CAGATGACAAGTGGAAGGATTGACTTTGAATGGGAAGAAG
FOR_71 chr7:115819799-115835164	GTATGAACACTTAAAGTTTTTGAAATTGCCTTCCAGAAAG
FOR_73 chr7:115835642-115836064	AACTCCCTAGATTTGCAAAATGCTTATACCCCAGAGGAAG
FOR_74 chr7:115836065-115836107	TTCCTGGAGATGTGATAATTTCTTTTTTAGGATAATAAAG
FOR_75 chr7:115836108-115836457	GCCTTCTACTACCACCCGAATACCGCACTTTTCCGACAAG
FOR_78 chr7:115839804-115843700	TAGATCATACTAAGCAGTACTGTTCAGAATATTAACAAAG
FOR_79 chr7:115843701-115851768	TTTTTCTATTATCCGCATCTCTAAGATGGGGAAATTAAAG
FOR_81 chr7:115858762-115865543	AGACAGGAGGAGATATTTGTAGGCACTCCTTTGAAACAAG
FOR_83 chr7:115866339-115870647	AGTAAAACAATGCCTTTTCCAAGTTTTAGTATTTGGAAAG
FOR_86 chr7:115875358-115877543	GGTCTGGAGTGGAAGAATGGACTGTGTTTTTCGTAAAAAG
FOR_87 chr7:115877544-115878831	ATTCGTGGCCAGCACAGCTCCCATAGACATGGCACACAAG
FOR_88 chr7:115878832-115881206	TTATTTTAAGCTGCAATAAGCCTACTCTTCCCACCAGAAG
FOR_89 chr7:115881207-115893184	AGCCAAAGATGTGTCAGTTGAGTTATTCATTTATAGTAAG
FOR_90 chr7:115893185-115899194	AAGCATTGTGTGCATCTATCCTGTTTGGGGTCCATTTAAG
FOR_91 chr7:115899195-115905376	AAAATTTATAACAATATATTTCCTCTTAAATGTGAGTAAG
FOR_92 chr7:115905377-115905816	AATAGTTCCCTTAAACTTTTCAAGACTTTCCTTCACCAAG
FOR_93 chr7:115905817-115908286	TTGCTAGTTTTTAAATTGTGTTCCCAGTCCTTTTCTCAAG
FOR_96 chr7:115912494-115917040	AAAAAGAAGATTTTCTTTAAGTTTATGTAGTTAGACTAAG
FOR_97 chr7:115917041-115918624	TGAGATGTCAAAGTTGTCAGCATGATATATCAATTTGAAG
FOR_98 chr7:115918625-115922291	ACTTATGCCAGAAACAGCAGGTCCTTCCAAATTTCTCAAG
FOR_100 chr7:115929440-115929873	GTGGACAGACCTTACAGAATTGCAAACCCATTTGGGCAAG
FOR_101 chr7:115929874-115931946	AGCAGCCTACTCCCTTATGCCCACCCAGGTAATAATAAAG
FOR_102 chr7:115931947-115932837	TTTTGCGCCAGCATCTACTTTAGGCAAGGACCTCAGGAAG
FOR_105 chr7:115937791-115944343	CCACTTGACCATAATTTTTCTGAAAATTGCTAAGCTCAAG
FOR_106 chr7:115944344-115945834	TACCAACTGTTTAATGGAAATTTTAATATACTGTATAAAG
FOR_107 chr7:115945835-115946603	AAAAAATTGCACTTTCAACATAGGATTTGCTACTCAAAAG
FOR_108 chr7:115946604-115948246	ATTTAAAATACTCAAAGTGGAAAAGCCCAATCCACAGAAG
FOR_109 chr7:115948247-115950557	AGGAAGGAAGACCCTAATTATTTGTTTACTCATAAATAAG
FOR_113 chr7:115957048-115959437	TTAAGGCAAGGTGCACCAAGTACCCTTGAATTATGAAAAG
FOR_114 chr7:115959438-115962331	AGTGCCCTTCAGGCTACTTGGGCTTAAAGCCAACTGCAAG

FOR_117 chr7:115970061-115980026	ATGCCATATGCCTCTGTCTCTTTTCATAAGGAAGACAAAG
FOR_118 chr7:115980027-115983146	CTAATTACTGCTGTGTAAATGGAGGTAGAAAGAAACTAAG
FOR_119 chr7:115983147-115987971	ACACTGAGTCGTACAGAAAGCTGCCTGGTATATCCAAAAG
FOR_120 chr7:115987972-115988758	CCAATGTTTTATATGTCATCTAATTTGAGCCCCCAGCAAG
FOR_121 chr7:115988759-115991418	ACTCGTTTCTGAGCCCTAAGCCTTTCTCCTCTTCATGAAG
FOR_122 chr7:115991419-115997316	ATCATTTTTAATTGTTCAATCATCCACATAGCTACCCAAG
FOR_123 chr7:115997317-116000429	CCAGAACAGTTTCTTTAGATCTTCTTATTTCTCTTTCAAG
FOR_124 chr7:116000430-116004316	GTTCTGACATGGCACCAAGAGGAGAACCATCCAGACAAAG
FOR_125 chr7:116004317-116009839	CACATGTGGGAGACACAGGAGATGTTTCTCAAGGGAAAAG
FOR_126 chr7:116009840-116014022	AGTCCCTAGCAGTCGCCATTCTAATTTCTGTTTCTATAAG
FOR_127 chr7:116014023-116014342	GAATCTGATGGTTCTACAAGGGAGAGTTCTCCTACACAAG
FOR_128 chr7:116014343-116018855	CTGTTTCATTTTCTGGTTTATTGGGTGAACATATTCCAAG
FOR_129 chr7:116018856-116019939	ATGAACGTTCCTGGAGCTAAGGGCAGAGTGAAGAGGGAAG
FOR_130 chr7:116019940-116021980	GTTAGAAGCATATATTCAGGATTTCCAGCTTCAATGGAAG
FOR_131 chr7:116021981-116028864	TCATGTTAATTCTTGCATGATTTATGTGTTATGATTAAAG
FOR_132 chr7:116028865-116030916	CCGTTTGGAGGCCTTACAGAAGTCATCTCACCTCTCTAAG
FOR_133 chr7:116030917-116033015	TTTTAAATTTTCTATGAGGAATCTGCATCTTTTGAAAAAG
FOR_134 chr7:116033016-116046180	TCCCCTCTGAGCTATTTTCTCAGCTACTCTCCAATTAAAG
FOR_135 chr7:116046181-116047724	TTATAAATATTTCTCCTCTGCCAGCCAGCCCAAAGTGAAG
FOR_136 chr7:116047725-116052329	GGCATCACAGTAGCCTGCAGTTTGCTTACTACTCTCCAAG
FOR_137 chr7:116052330-116055533	AAAGCAGTGGGCATATGGGAGGAATTATTAAGGTAATAAG
FOR_138 chr7:116055534-116058179	ATCATTCAGGGTCCATTTTCCATGCATTGCTTATGCAAAG
FOR_139 chr7:116058180-116058542	CGTACTCTCCCTTTTCCACCAGTCATTTCAGAGTCTCAAG
FOR_140 chr7:116058543-116058738	CCTTTCATTTTGTAAAAGAAGAAACCGGGGCCCAGAGAAG
FOR_141 chr7:116058739-116060423	TGCACATACACAACAAATAACTATACATAAGCGGTGCAAG
FOR_142 chr7:116060424-116062230	ATATTTTTGAACATTGCAACCCATTTCCTAATTCAGGAAG
FOR_144 chr7:116062714-116066773	ATATTTGTTAAACTTCCCTATTCGAGATTTTGTTATTAAG
FOR_147 chr7:116071401-116071748	GAGTTTTCATTTCACTTACTATATGTGCTTTATAACTAAG
FOR_148 chr7:116071749-116079800	GTTAGGGTATAAGTGGAACCCTCTAGGTCTGGTATGAAAG
FOR_149 chr7:116079801-116084239	TTAACTCATGTAACATCTTTCAGCAGAAAATCTCAAGAAG
FOR_150 chr7:116084240-116084622	AATTGGAAAGATTTTCATCAGATAGACTTTTAACACCAAG
FOR_153 chr7:116087402-116089954	CTGCTCTGAGACACCCCAGGGAGAGGACCAGTGGTGGAAG
FOR_154 chr7:116089955-116091794	GATTCAAGCTGTGTCCCAGCAGTCTCTGCCTTCTATCAAG
FOR_155 chr7:116091795-116092881	CAGCGTACCACAACAGATGGACCAGACACTTCAGGGAAAG
FOR_158 chr7:116108496-116108907	AAATACGAATTTAGACAGTGATATAAAGATTCAAGGAAAG
FOR_159 chr7:116108908-116116288	AACATTAATTATAACACATGCAAATCATTCCTGTACAAAG
FOR_160 chr7:116116289-116116599	AATGCTATGCAGTTTTGCAATACATATATACCACAGAAAG
FOR_161 chr7:116116600-116126267	AGTTTTTGTTTTCCTTCTATGTAAAAGTCCAGTTGGGAAG
FOR_162 chr7:116126268-116127854	GTGACATTTCAGATTATATTAAAGTTATTTCCCATATAAG
FOR_163 chr7:116127855-116134507	AGAAAAATGTCTAGCAATTGCATGCTAATTTAGACCAAAG

FOR_164 chr7:116134508-116135577	AGGCTGTCCTGAGGAGGTGAAAGCTAACCTTTATATAAAG
FOR_165 chr7:116135578-116135668	ACTAAATATATAAACTAATAGTCATTACTTATTTCACAAG
FOR_166 chr7:116135669-116138207	AGTAAATATTAATTATTTGTGTGATCTTTAGATATGAAAG
FOR_167 chr7:116138208-116138430	TTGGCATGGCTTCATTTTACCTTTCTTGCTATTAGCAAAG
FOR_168 chr7:116138431-116138839	TTCAAAATAAAGCCTGGTTTCATTCATAGTAAATATGAAG
FOR_169 chr7:116138840-116140816	GGAGGAAAGGCACCAAGTACGTGCTTGGTACAGAGAGAAG
FOR_173 chr7:116151992-116154098	CCTGAGACTTTTCCTGCAGGCCTTTGGGAAGCAACTGAAG
FOR_174 chr7:116154099-116155205	GTCATATTTGCACAATGCTTTACATCTGAGATACTGAAAG
FOR_175 chr7:116155206-116157096	AAACTCTTTTAACTTCCCCAATTTCTGGAAAAACAAAAAG
FOR_176 chr7:116157097-116159349	TCTGTTAAATGTGTGGTCTCTTTTTATTAAGTTCTTGAAG
FOR_177 chr7:116159350-116160026	TCAGACAATAACCTCTAGTAAACCAACTGAGAATGATAAG
FOR_178 chr7:116160027-116160694	ATCGTCCATTACCTCACCAATCCAGTTTTGTATGAATAAG
FOR_179 chr7:116160695-116162927	CGTTTTTGTGAACATCTGGCAGAGACACAGCTGAAGGAAG
FOR_180 chr7:116162928-116164732	GGGATGGTATGGAGTGGCTCTTGGGAGCTCAGAGTTGAAG
FOR_181 chr7:116164733-116165014	TGTTATGAAAGTACAAACATTTGGTGCTTTTCAGACAAAG
FOR_182 chr7:116165015-116165638	AAAATTGGGTAACCAAAATCTGGAGAATTTGTAATAGAAG
FOR_186 chr7:116176540-116180390	GTGTAATGGATTTTTATGGGTAAAATGGCTATCATAAAAG
FOR_187 chr7:116180391-116180693	CATTTCCTTTCTCAATTGTCACAAAGCTGCTGAGAAGAAG
FOR_188 chr7:116180694-116183535	GTGCTAGTTCTACTGTATTCTACATGTACCCTTCTACAAG
FOR_190 chr7:116185922-116186482	TGTGTAATTCGGTGAATAAAATACAAATGCATTTCTTAAG
FOR_191 chr7:116186483-116187729	GTCACATTAAAGTGGTAGTTTTAGTTTAGACCTTAGAAAG
FOR_192 chr7:116187730-116191665	CATCTGCTTCTTTAATTAGTTTTAGAAAGGGAGAAATAAG
FOR_193 chr7:116191666-116201664	CATCACTGCACAGTGGAGAATTTACCACTAATGTGAAAAG
FOR_194 chr7:116201665-116201971	TGATAAAGAGAGAAATGTCTAAGGAAATGAGGGTAAAAAG
FOR_196 chr7:116203379-116205630	TAACCAGCTCTACTACAAACTTTATATGGGACTTAATAAG
FOR_197 chr7:116205631-116206264	ATTTCATTCGAAATGAGACTCATGTAAGTTGACTGCCAAG
FOR_198 chr7:116206265-116212750	AGGTGTATTCGGGCAGGAGCAGGACCAAGGTGGTGACAAG
FOR_199 chr7:116212751-116215727	TGAAATTTTGGATGTTTAGGGCAAGTGTTGGGATTCTAAG
FOR_201 chr7:116217701-116220386	GCCTCCATATTGTGAATCAGCTTCTCCTTTCTGTTCTAAG
FOR_202 chr7:116220387-116221965	TTGCCGAGCTGTCTGGTGTGGATTTTCCTTCTGCTCCAAG
FOR_204 chr7:116241645-116247085	TTTCTATCCCCCTTCACCAGTCTTCCTTCTTTCTGCTAAG
FOR_208 chr7:116259833-116263562	CAATGACTTAACTGCCTGCCAGAGATAGGAGAGCATCAAG
FOR_209 chr7:116263563-116268283	GGTTCTCAAATCTAGTTATACTTTAGGATCACCTGAGAAG
FOR_211 chr7:116276679-116279948	TTATTTTACAAAGCACCTTTCATGCATTATTTCATGGAAG
FOR_213 chr7:116284006-116284370	AAATAAAAGTGTCGTTTAATCAATATGCATTCCAACAAAG
FOR_214 chr7:116284371-116284829	AGAATTTTAAAGGATAGGTAGCAAACAACATTCCAGAAAG
FOR_215 chr7:116284830-116288190	GCAGCTTTGATTACATGTTTAATCTCTCATAAACTTCAAG
FOR_217 chr7:116290572-116291465	GAGTGCACAAACATATTTAAAATTGAGGCGTGCTAAAAAG
FOR_218 chr7:116291466-116291670	TAGAAGATGCCTTATTATCTGGCACTCTCAGGGCACTAAG
FOR_220 chr7:116292370-116293242	TTAGGTTCTAAATTTTTGTCTTGATGATTTTAGAATAAAG

FOR_221 chr7:116293243-116298412	TGGTTAAAGCGAGATGGGATAAGCCCTTTAGGACAAAAAG
FOR_223 chr7:116299307-116300309	CGGTCTGGAATAGAGGGTGGGAATCTGTATTTTTTCCAAG
FOR_224 chr7:116300310-116301217	ACACAGTCAAGGATTGCAAGGCAGAAAGATACATGTGAAG
FOR_225 chr7:116301218-116304290	CTTCCCCTGGACCTCCGCAAGGGCTGGCCGGGACCCAAAG
FOR_227 chr7:116320019-116322321	TTTATGAGTAAATGACCACTTGCTCCTAAATATTTTAAAG
FOR_228 chr7:116322322-116322448	GGAAACATACAGCGTTTTTGAACAATACTTTCAAAGTAAG
FOR_229 chr7:116322449-116323266	TTACCTATCATTGAGGCCATCCAAAACTACGTAACAAAAG
FOR_230 chr7:116323267-116325245	CTTCTTTAGAAGTACTTCTGTGTTCTCCTTTCTCAAAAAG
FOR_231 chr7:116325246-116326755	GAACATAAGAGAAAGTTCTTAGTGGCTAATCATGTCAAAG
FOR_232 chr7:116326756-116326857	TAGGCTAGAAGGTGACTTGTAACTGAAGGGAAATTTTAAG
FOR_233 chr7:116326858-116327883	CAATTTAAAATGATGGTAAAAGTTACAGATTAGGAATAAG
FOR_237 chr7:116334861-116338483	TTACAGCAGTTTTAAAATGGTGCCCCTTGTAGGACCAAAG
FOR_240 chr7:116344062-116344149	TTGCCTTTGAGTGTCATGTTGATGCTCAAAATGTTTCAAG
FOR_241 chr7:116344150-116345302	TAAAATCACTGTAATTAATTAGTTTGATTAGAGCACAAAG
FOR_242 chr7:116345303-116346713	TTTGATGATGGTGAACTAAATACTAGATCTAATTGACAAG
FOR_243 chr7:116346714-116350123	AGACATATCCAAATTTGACATTCTGAGTATTATGCAAAAG
FOR_244 chr7:116350124-116353124	AGAAAAAGATAGATATCATTTTATGAGACTTATTTATAAG
FOR_245 chr7:116353125-116355838	GACGTGCAATCTCATCTGAGGCTTGGATTCCTCTTCCAAG
FOR_246 chr7:116355839-116361566	TCAGTTCTCATGGTCACAAAATGGCGGTGATAGCTCCAAG
FOR_248 chr7:116361593-116367841	CAAAATGTATGTTAATCAACTGTGTATGTTATTGGCAAAG
FOR_249 chr7:116367842-116372955	CTAGGAGTAAATGCTCTGTGTCAATAGAAACAAATTTAAG
FOR_250 chr7:116372956-116374893	TATACATGCAGAATTATATCTTAGAGAACTTCTCCCCAAG
FOR_252 chr7:116386571-116391563	GACAGGTAATAGTAGAATGCTTACAGTTCTCTTTGTTAAG
FOR_255 chr7:116396053-116396912	CTTCAGTTCGCTTTGCACAGGCTGCACTTAAGGAATGAAG
FOR_257 chr7:116397833-116406683	TACAGCAGTCTTCTTGTCTTATGCTGTATCTTTCAATAAG
FOR_258 chr7:116406684-116412712	CACATAATTTCCTCTTGATTTTTCAATATGAACAGAAAAG
FOR_259 chr7:116412713-116418208	AACATTCATACAGAAAAGAGCAGATATTCTAAGTTTAAAG
FOR_265 chr7:116427183-116427933	CTTCTGTAGTCTATCAAAGTAATTGCATTATTAGGAGAAG
FOR_266 chr7:116427934-116428247	TCTCTGAAAAACCCAGACATCTGGTAGACATTCTCTGAAG
FOR_267 chr7:116428248-116433502	TTGAAAGAACAGTGTCCATATCTTATTGGAAATAGTAAAG
FOR_271 chr7:116434982-116436704	TAAAAACCACACAGGAAGGTCTGTGGTTTGCCTAAAGAAG
FOR_272 chr7:116436705-116438260	AGGCATATGTCTCAGCCCCTGAGCTGTCCTCTATGAAAAG
FOR_273 chr7:116438261-116438320	CCTCTCAATCAAGTCCTGGAAGAATGGTCTCCTTCAGAAG
FOR_275 chr7:116444568-116447115	GATTGAGTGTACAATCCTTGATCTGTCAATCTACCATAAG
FOR_277 chr7:116449020-116451967	ATGTGGAGGGGGAAAATAAATCACCAGAGAACCAACAAAG
FOR_278 chr7:116451968-116453989	GTACCGTGTACTACCAGGGTTTAACCATCTACCATCTAAG
FOR_280 chr7:116462009-116463443	TTTAATAAATCTCCCGAATTGTTTGGTACTGTTTGGCAAG
FOR_281 chr7:116463444-116464263	ATTCCGCTTACAGAATTCAGCAGGTGTGCTTCCTAATAAG
FOR_282 chr7:116464264-116464575	TTTTATGTTATGCCCTTAATGCACAGCCCTACCACAAAAG
FOR_283 chr7:116464576-116467915	TGCTTCATAGTAATTCTTTAGGGGATGGTCATGGGTAAAG

FOR_284 chr7:116467916-116470897	GAGTTTCCTATATTGTTTAATTGCCTTTAACCAATGGAAG
FOR_285 chr7:116470898-116479290	GTTTTTATTCCCATTCTCTTGCTTCTTACCAAAATAAAAG
FOR_286 chr7:116479291-116480731	CGAACAGGAGGACGGGGATCTTTGCCCTCCTCCTTTGAAG
FOR_289 chr7:116485626-116489121	ATATTGGTTGTCACTTGGAAAATTAATATGTAGGATGAAG
FOR_290 chr7:116489122-116490630	AAGCCCAGGACAAAGATTAAATGTTTTCACAATGATAAAG
FOR_291 chr7:116490631-116492029	CGTCTGTACATTGCCTGGTCTGGGCTTCTTTAAGGAAAAG
FOR_292 chr7:116492030-116496463	ATATATTTTGTATTTGAAAATCTGCTGGTAATTTCAAAAG
FOR_293 chr7:116496464-116505959	ATAGATGCTTTGCATCAGGGTCATTTCTACTATTATAAAG
FOR_294 chr7:116505960-116511717	CGCAACTGGAAATACCTGTGAACACATGCTTGGGTACAAG
FOR_298 chr7:116528256-116528326	AGAGCATAAATGGACCTAGAATAATGGTGGATTTCAGAAG
FOR_299 chr7:116528327-116532214	CTACAGAGTGATTACCCAATATCTCAGTGGGGTTAAGAAG
FOR_302 chr7:116535371-116541890	GTCTACTTTAATTCTTAAGAAGCACTTAGCCAGTTTCAAG
FOR_303 chr7:116541891-116541946	TTTCCCATGGGTTTAATGGTGTTGGTGTGGAATTACTAAG
FOR_304 chr7:116541947-116542198	ATTTTTGTTCCCTGCAGGTAAATATGAGACAAGTTTGAAG
FOR_306 chr7:116552834-116554112	TAAACATTATTCTTTAGGTTCTTTGTCTCTCACAAAGAAG
FOR_307 chr7:116554113-116555471	TTATTTTGTGCTAGATACAGATCTTGCAATATTCATGAAG
FOR_309 chr7:116557409-116558709	CAGGTAAATTTGGGATATAAGTACACATGTTAAGCAAAAG
FOR_312 chr7:116566112-116567132	AAAAGGAGATTTCTGCCGAGATATGAACTTGCTCTCAAAG
FOR_314 chr7:116569718-116575669	TATTTAACTGGTTTATTGTTGTCTAAATTTTATAATCAAG
FOR_315 chr7:116575670-116576527	CGGCTTTCTGGGACCTCCCACTGAGCCATTTAAATACAAG
FOR_316 chr7:116576528-116577105	CTTGGCTAAATGTTAACATTTAAAACCAAGCACCAAAAAG
FOR_317 chr7:116577106-116577644	GTTGTCCACTCTTTTCATGCAGTTACCCATATTCCATAAG
FOR_318 chr7:116577645-116578660	AATCTGGACTCACACTGGATTTCTTTTTGAGTGTGCCAAG
FOR_319 chr7:116578661-116581134	GGTAAGATGTCTACTGGAGTAGTAGAACTCTTGCAATAAG
FOR_320 chr7:116581135-116582676	GTGTAAGGCATTTAGATTTCATTGTAAGTACAATGGGAAG
FOR_321 chr7:116582677-116583219	GTTACTCATTCCCACCTCAGAAACTTTACACTGCTGGAAG
FOR_322 chr7:116583220-116585607	CACTGAAAACATTTTTGTTGGGCTGTGCATAAATGCAAAG
FOR_324 chr7:116585641-116586244	ATGATACAGGCCAAATTCTATAAACTTTAATTAGCCCAAG
FOR_325 chr7:116586245-116586892	TGATAGAAAACAGGCCTTTATCATATGCTTTCATTCAAAG
FOR_326 chr7:116586893-116595480	ATTGGTGATTTGTTTTTGGACACAGTATACTCTGTCCAAG
FOR_327 chr7:116595481-116595603	ATGCGGTCTCACTGAGTTTTAGTAAGACTCAGTTTTTAAG
FOR_328 chr7:116595604-116597168	GCACTGTGTGGCTGTTTTTCACTTTCACCCCATCTTTAAG
FOR_329 chr7:116597169-116599916	CTAAAATCTTACTCATGGTTCAGCTAAACAGTGGTCCAAG
FOR_331 chr7:116602593-116607025	TCGGCTAATTGGGGGTAGGGGGACTGGATTAAAATCCAAG
FOR_332 chr7:116607026-116607805	AATAATACCATGGATTTAATTCTTACCGAAAAGGAAGAAG
FOR_334 chr7:116614251-116616617	CTTATGTGTTCCTACTTTCTTTTTCTCCAACAGATATAAG
FOR_335 chr7:116616618-116621728	GGATGGCCCTTCACCACCCCCATTAAGAGTCATTAGGAAG
FOR_336 chr7:116621729-116626310	AATGTTTCCAGATTCTATTAAGTTTTCGCTCTTTCCAAAG
FOR_337 chr7:116626311-116628522	ATATAAGTTTCTTTGAATTGAGTCATATTTTGTTCCAAAG
FOR_339 chr7:116630906-116635745	GCGGAACACTTCCCTACTCTTCCTGCTTTGAGGCCAGAAG

FOR_340 chr7:116635746-116635878	AGTGGGGCATGCACACCAACCGTGCCAATGGGAGAGGAAG
FOR_341 chr7:116635879-116637096	GTCCTGCTTTTTTATTTCAGTACCTACTAGAAATGAAAAG
FOR_343 chr7:116652179-116652539	GAGGACAGGACTCAACTCGCCTCATTTCATGTTTTCAAAG
FOR_344 chr7:116652540-116653236	AGGACTAACTCTTCCTAGAATGTTTAGTTACAAAGATAAG
FOR_345 chr7:116653237-116654311	TCGGAAAAGGGAGCATTTATGGACATCAACCAGGATGAAG
FOR_346 chr7:116654312-116658005	CTGAACCAAAGTGATCATTTGTCCTCACCACCTTTAAAAG
FOR_349 chr7:116662874-116666084	CGAAATTTCGTGAGGCGCCCCCACCTCCGACAGCCACAAG
FOR_350 chr7:116666085-116670645	TACATCAGTTTTCTGCCACAGGAGAGAAAACTGAACTAAG
FOR_353 chr7:116678790-116681537	TAGTGAATAGGATATGTATTCCTATAAAATTCTAGTAAAG
FOR_354 chr7:116681538-116702196	TGTCATGGCCTGGAAAAAGCCACTGCCCTTGATCACCAAG
FOR_355 chr7:116702197-116707011	AACGTTTGAGTTCCTAGAGACTCTGTATCCTTATAAGAAG
FOR_357 chr7:116707989-116711010	GAGAGAGAGGAGCAGACAAAGCTATGACCATTGACAAAAG
FOR_358 chr7:116711011-116712084	TAAGCATTTTGCCTGCATTGTCTAATTTAATTCTCATAAG
FOR_359 chr7:116712085-116715789	AAAGATATTTACCAGTACAAAATAGGTGTCATGGGGAAAG
FOR_360 chr7:116715790-116716759	TGCGCCTGAATATGAGGAAAAATTCCATTGAACACTTAAG
FOR_361 chr7:116716760-116719057	GTTCTTGCCATAATAATATCTTCCTTAATTGGTTTTTAAG
FOR_362 chr7:116719058-116719473	TAAGGGGCTTCAAGGCAACCTAACCAAAATGTGGGTGAAG
FOR_363 chr7:116719474-116719968	TTTCTCTAGAGGCTGACTTGGTAACTGCTGTCAGTAAAAG
FOR_364 chr7:116719969-116725975	TCTTCCTTGAGAACAGGTGGGCCTGAAACTTCCAGGGAAG
FOR_365 chr7:116725976-116729875	GTATAACCTGGTGATTCCACTCTGGAAACATCTGCACAAG
FOR_367 chr7:116732656-116733772	TGAGTACTTGCTGCCTCTTGCTAATGTGGGATACCGAAAG
FOR_368 chr7:116733773-116734834	TATGGGGCTCATGTTCCTTGGATTCAGTGAGTTCTATAAG
FOR_370 chr7:116737156-116744332	AAAATACCTTTTCCATAGTTACCCCATTTATCTTCTAAAG
FOR_371 chr7:116744333-116744401	ATGTTTAGAGAGATCACATAGAGGAGGAAATTCCCAAAAG
FOR_373 chr7:116752148-116760886	AGAAAAACAACTTGGGAGATTTACATTATGAGATATCAAG
FOR_378 chr7:116773804-116775029	CAAGACAGCATCAGCTGCCATAGTATAAAGAGGATATAAG
FOR_379 chr7:116775030-116776390	CTAACCTAAATTTTCAGCATCCTATAACGTTAACTTAAAG
FOR_380 chr7:116776391-116783913	TGAATTGTATGCTTTCAATGAATGAATTATCTCAAGAAAG
FOR_383 chr7:116791984-116792346	TCCCTTCAATTATTATTGGTAAACTCTGAAATAATATAAG
FOR_384 chr7:116792347-116793757	AATGTATAATCTATCCTTCAAATCAGATATTTCACAAAAG
FOR_385 chr7:116793758-116799657	AAATATGATTGGAAGTTACTGTGTAACTTTTGAACAGAAG
FOR_386 chr7:116799658-116803602	ATCCTATGACCATTAAATAAATCAAATTGTAATTTAAAAG
FOR_387 chr7:116803603-116804982	TACCACAATCTCAGAAAGGTTTTTTGTAGACATAGAAAAG
FOR_388 chr7:116804983-116805860	ATGATTGTGAGGATGGAGAGCAAACTGGATCACTCATAAG
FOR_389 chr7:116805861-116811442	ACTTGATCCTTGCTTAGAAGTTACAGACTTGTGGCTAAAG
FOR_390 chr7:116811443-116813217	ATACCACCAGTACTGCTTTGAAAACGAAAAGAAATGAAAG
FOR_392 chr7:116819026-116821165	TTTTTCTCTTAGTTTCACTCCCACCCACTCCATCTGAAAG
FOR_393 chr7:116821166-116822556	ATAACTTTATCAAAATATTTTAATAGCTTTGTATCTAAAG
FOR_394 chr7:116822557-116822904	TTAGACCTTTCACATTTTTAAATCAAATTGTAATCTCAAG
FOR_395 chr7:116822905-116827129	GCCGTAAATATACTTAAACCATTGAATTCTACACTTTAAG

FOR_396 chr7:116827130-116830002	TGGGCCAAATGGGGAATAAAAGGACTAATACCCTATAAAG
FOR_397 chr7:116830003-116833414	TATCAGAGTATGCTTGAGTTAATGCTGTCGGGCATAGAAG
FOR_398 chr7:116833415-116837623	AAATAAAAGTTCAGTTTCAGAAGTTAGCAAGGTCATGAAG
FOR_399 chr7:116837624-116844404	ATGCTACTTTACACATGGAAACTGAGACTTCAATCTCAAG
FOR_400 chr7:116844405-116849537	AATCATTTAAGTTATCTAAAACATCTTACCCTTCTCAAAG
FOR_401 chr7:116849538-116849668	AGTTGGAATCTACACTAATGCCTGTCAATATAAAAACAAG
FOR_402 chr7:116849669-116850742	CTCTAACAAATCTTCCTCTTTGAACTGCTTTGATTGAAAG
FOR_403 chr7:116850743-116852285	GATATAATAATCCTCATTTTAAAACGGGCAGAAATTGAAG
FOR_407 chr7:116859106-116862886	CTCAGAATGAGCTATGATTCATTTACAGGTAAGTCACAAG
FOR_408 chr7:116862887-116863142	GGCTCTTCTTCCTAGTTCCTGAGTTTTTTTAACTTTCAAG
FOR_409 chr7:116863143-116865693	ACCTGGGTACATCTGTCTCTTTTTATCAGGAAAGCCAAAG
FOR_410 chr7:116865694-116866419	GTGATTATTTGCATTTGATAGTTGGGAATATGAGATCAAG
FOR_411 chr7:116866420-116870064	TTATTAAACTGTAATGAGCTAAAGAATAGATTTTCTGAAG
FOR_412 chr7:116870065-116870540	GGTCGAAGATACCTATAAGAAACATACAAGCCTTGTTAAG
FOR_414 chr7:116881628-116886437	TTTTGTATAAAAGTGTGCTTTTCCGGGATGTTGTTTGAAG
FOR_415 chr7:116886438-116887480	GACAAAAATGAACTTCCTCAGGTCACCTAATGAATACAAG
FOR_416 chr7:116887481-116887586	AGCAAAAAGGAAATTCTTCCAATTAGGCAGAGTGATGAAG
FOR_418 chr7:116899363-116900042	GGAATCACCTAGCACCTAAGGATGGGTGAGTAAAGACAAG
FOR_419 chr7:116900043-116901864	AATTTAATCTGTGTTATGAATTCCCTTGTTTCTAGAAAAG
FOR_422 chr7:116908132-116912912	ATTGTGTCAGCAAAGAGCTTAGGCTTCTTTTAGCAAGAAG
FOR_423 chr7:116912913-116916998	ATAACAGGTTTAAATATGTGCAAGGAGATGTCAGCAGAAG
FOR_424 chr7:116916999-116920846	AAGCAATCTGAGGATGCTGTCCAATCCTGCCTCTCTCAAG
FOR_426 chr7:116920882-116922275	TGATGACGTTTTGACAGTTGCACAAGTTTCTTTCTTTAAG
FOR_427 chr7:116922276-116924466	TAAAATTCTGACTCTGCCCTTGACAAATTAACTAATTAAG
FOR_428 chr7:116924467-116926564	TTAATAAAAGGCAAAAGTCAGATCGAAAAATGAGTATAAG
FOR_429 chr7:116926565-116926608	AATTATATGACTTTAGGAGGATATGTTATGAAAATCAAAG
FOR_430 chr7:116926609-116927976	GGCTTAAGATGACTCTGTGGTCATTGTTGGGCAGCATAAG
FOR_431 chr7:116927977-116928226	TAACATAACTTATTAAAATGGGCACAGATAACACAGGAAG
FOR_432 chr7:116928227-116934991	AAAGGTAATTAGGCTTTATAGAAGGGCTCTCATTAGCAAG
FOR_433 chr7:116934992-116938677	ATGTATGGAGAAAGACTAAATTGTTCTTAAACTTCTCAAG
FOR_434 chr7:116938678-116940500	AGCGAGGATGAAGACCTACCCTCTCAGTGCAGTATGGAAG
FOR_439 chr7:116954445-116959635	TGTTATCATAATGATAATAGCTCAGGATGTGCATGACAAG
FOR_440 chr7:116959636-116966785	TCTACTGAGACTCATTCAGTCAGTATCTCTAAGAGGCAAG
FOR_441 chr7:116966786-116969901	AATAAAAAGAAGTGGAGGAATATTAAATGCAATATAAAAG
FOR_442 chr7:116969902-116971187	TGTACTTAGAACTTATATTTTGTAGTGACTTCTTTAAAAG
FOR_443 chr7:116971188-116971935	CAGTAGAACAGGGGACCAGCAATTATTCTTCCTCCAGAAG
FOR_444 chr7:116971936-116972824	TGTCCGTGTACCTCAAGGTTCTTGTTCATCACTTCCCAAG
FOR_445 chr7:116972825-116975278	TAGGACTGTTTACTTTGATTTGAAGACCACTATGCTAAAG
FOR_446 chr7:116975279-116977027	TTTGTAGAAATGTGAGAAAAAGTTTAGTGGTCTCTTGAAG
FOR_447 chr7:116977028-116978335	TCTTACCCACTCAATACAATTTGGTAATTTGTATCAGAAG

FOR_448 chr7:116978336-116981205	CCTAAGAGTCTTCCAAGTAGCAGGTGAAGCAAGTGCAAAG
FOR_449 chr7:116981206-116981669	ATATGAGAAAAGTCACAGAGGAGTCAAAGATGATTCCAAG
FOR_450 chr7:116981670-116984191	ACTTTTCTTCCTGATTAGGTTTCTACTAAAACCAAACAAG
FOR_451 chr7:116984192-116987118	TGGATCAATTAATAAAACACATGACCTATGCTTTAAGAAG
FOR_453 chr7:116988885-116990395	TGATAGTGAGGAGCAGAATTAAGGTGGACGTGATTAGAAG
FOR_454 chr7:116990396-116993153	AAAATACTTGGGAGGGAAATTGGCAAGAAGTATGAAAAAG
FOR_455 chr7:116993154-116993843	CCTTGAACTACCTGTAATTTTCTCCTGATAGGTAATTAAG
FOR_456 chr7:116993844-116998434	AATTCCCTGGTGGACAGTCAGCAAAGGGGAAGATGAGAAG
FOR_457 chr7:116998435-116998790	GCCAATTCTTAGACCAGTTGAATCAAAATCATCTAAAAAG
FOR_458 chr7:116998791-117000144	GCATTACCTCTAAAGTTCATGTTTCCTTTACCCAACCAAG
FOR_459 chr7:117000145-117000426	ATATCATTGCCCTTTGTATGTGCAAATCAGAGTTAATAAG
FOR_460 chr7:117000427-117004204	TCCACTCACTGCTTGGAACATATTGCCTGCAATAATTAAG
FOR_461 chr7:117004205-117007965	TCACATAGCATCTAATAAACGTATTTATGAACTACAAAAG
FOR_462 chr7:117007966-117010107	CCTCATTTTCTTAAAGCATTTTCTCCTTTTCCTCAACAAG
FOR_464 chr7:117011453-117015792	AAACAATTATTGGTGACTCTGATGTCAAATGTTTCTCAAG
FOR_465 chr7:117015793-117020511	TGGGTTTCTGAAAAGATTTCTTTTAGAGAATGTATATAAG
FOR_466 chr7:117020512-117024397	TCTCGGTGTGAAAGTGGCAACCTGCTGGGGTCTGTAGAAG
FOR_467 chr7:117024398-117033085	CTACCAGTTTGGCTGAAACAGGATATTTGAGCAGAGGAAG
FOR_468 chr7:117033086-117036940	ATATGCTCCGAAAGGTTAGATAGCTTGCCTAAATGACAAG
FOR_469 chr7:117036941-117038746	ATCATATCTATTCAAAGAATGGCACCAGTGTGAAAAAAAG
FOR_470 chr7:117038747-117038904	TCACAGGCAGGAGTCCAATTTTCACTCATCTTGTTACAAG
FOR_472 chr7:117042650-117046025	CCTTTCCAGGGTCTCAGGAGACTATTCAAACAGGACAAAG
FOR_473 chr7:117046026-117047652	TGAAACAGTTTGATGAGAAGTTTTCTAAGGTTCCTTCAAG
FOR_475 chr7:117051319-117051809	CTGCACCCTACTGCTCCCATGTCATGCAAGGCACAGGAAG
FOR_476 chr7:117051810-117057401	CACTGAAACCTGAGGCAGATTTATTGTGGCAATAACAAAG
FOR_477 chr7:117057402-117057519	TCTTTATTCTTGAAAAGAGCCCTCCCTCCAAATTTACAAG
FOR_478 chr7:117057520-117057672	AAGAGAAGGTTTTATATAGGAAGTGGCATTTAGAATGAAG
FOR_479 chr7:117057673-117060421	AATGGAGTTGGGAAGGACTGACATCTTGACAATGTTGAAG
FOR_480 chr7:117060422-117061837	TTTAATTTGATCTTCTTTTGCTAGTTATCCAAGGTGGAAG
FOR_482 chr7:117062193-117067435	TTAGAAAACACTGACTTAGATTTAGGGTATGTCTTAAAAG
FOR_483 chr7:117067436-117070752	CTCCAGGTAGACTGACTATAAATCTAGACTGAAAAAAAAG
FOR_484 chr7:117070753-117070878	TTTTAATTTATTTGTCTAGACTTTGTCATGTTGCCAGAAG
FOR_485 chr7:117070879-117073682	AGAAAGTGATAATTAAGGCGGTTCATAAAGGGTCATAAAG
FOR_487 chr7:117083344-117088524	ACTATGTTGACTTTTCTCTAATGAAGATGATTCTAAAAAG
FOR_488 chr7:117088525-117089649	ATAACACAGAGCTGAAGAGCACGTTTTTACCCTCAAGAAG
FOR_489 chr7:117089650-117091888	AAGGTAGTGGGGGTAGAGGGATTGGTATGAAAAACATAAG
FOR_490 chr7:117091889-117092016	TTACGGGCTCAGATCTGTGATAGAACAGTTTCCTGGGAAG
FOR_491 chr7:117092017-117092396	TTTCATACAAACTCTTCCCCCTTGTCAACACATGATGAAG
FOR_492 chr7:117092397-117093667	TCCTTACTGAAGGCTTGTTATTACCATGATTATCACTAAG
FOR_493 chr7:117093668-117093837	TTAAACTCAGGCACATCCTGACCAATTTGGAATTCTTAAG

FOR_494 chr7:117093838-117094852	TTAAGTAATGATAACTGGAAACTTCAGCGGTTTATATAAG
FOR_495 chr7:117094853-117096018	ACTTGTCATGTTTCATCAGTCTCTCACTCCAATTTCTAAG
FOR_496 chr7:117096019-117097027	AGATTTATATTCTGCAAGTGTCCATCTATTTCTTTTAAAG
FOR_497 chr7:117097028-117097560	GCGTGGACCCCAAGCTCACTATGCCAAAGGGCAAGTTAAG
FOR_498 chr7:117097561-117113770	GAACAAAAGGAAGGCAATATTGAGTAATTCAGCTCTGAAG
FOR_499 chr7:117113771-117119446	GCTACAAGCCCACTATACAATATTGTTGTGATCTGATAAG
FOR_500 chr7:117119447-117124005	AACTAGTTGAGTTATAACCTTGGGTAGGAAATTACATAAG
FOR_501 chr7:117124006-117125194	TGAGAACACTTTTTCCATTCTCCATAAGTCTAACTCTAAG
FOR_502 chr7:117125195-117128742	TTTTGTGTCCTAAGACATCTTGTCTAATCTGAGGTAAAAG
FOR_503 chr7:117128743-117144189	TATTGCCCCACGCCCAGTGCCGCTAGGTGATCAGATAAAG
FOR_504 chr7:117144190-117149507	CAGGAAAGAAATGGGCTGTCTCTTTATCAAATTTGATAAG
FOR_506 chr7:117150627-117150962	AAGATACAGACGTGTTACATCTTCTTTGTTAGAATGAAAG
FOR_507 chr7:117150963-117153567	AAACAAAACCACATCTCCCAGGTTAGGAATGTTTTCAAAG
FOR_508 chr7:117153568-117155028	AAACAGTAAGTACTGAACATATGTCTCAATAAAGGGAAAG
FOR_509 chr7:117155029-117158580	ATTGTTAAATTTTGCTAGATTTATACACAGAAAGATAAAG
FOR_510 chr7:117158581-117160639	CCCACTCTCAAACGCAATGGAAAAGAGAGTGAACCTGAAG
FOR_511 chr7:117160640-117162523	TGGTACTCTGGGAAACAGAAAGATTAGAATATTACAGAAG
FOR_513 chr7:117162744-117166350	CCAAAAAGGGTCAAACAGCACAGCTTTCTTTGGGGGAAAG
FOR_514 chr7:117166351-117170501	GACCTTGTTCATTGCACTCCAGTTTTCTTAGCTCTCTAAG
FOR_515 chr7:117170502-117173345	CGGGCAGAGTGCAAGCTGATGTACTATGTAGTCTTGCAAG
FOR_516 chr7:117173346-117174358	AAACGCGTTATATCTGCCCAGCACACATGCTTGGGAGAAG
FOR_517 chr7:117174359-117177986	CTGACTGTTACTCTATCAGAACCGGGTCGAACATGCAAAG
FOR_518 chr7:117177987-117184316	TAAGTAAAAGTTATCATAGATCATTGCAATCCTCAGAAAG
FOR_521 chr7:117185609-117186934	ACCTTTCTTTTCCAATACGGTGCTGTTGAAAGTACACAAG
FOR_522 chr7:117186935-117187620	GATACGAAGTTTTTCAGCACCGGCAGCTGCCACAAGTAAG
FOR_524 chr7:117190233-117190464	TATTTATTGTTCATTTTATCTCTTAATGTTAGACTATAAG
FOR_525 chr7:117190465-117196776	ATGACCTGGCCCTGTTCAGGTTAGAGGGATCTCTCTAAAG
FOR_526 chr7:117196777-117200917	TAAGCTCTTAGAAAATAGGAGTAGGCAGCTGAGAATAAAG
FOR_527 chr7:117200918-117204998	AAGAATTTCCATGAGCTGGTATTCTATGGTACATAAGAAG
FOR_528 chr7:117204999-117205660	GAGTGTTTTGTGACTGCAGAGGGATACTGATCATTATAAG
FOR_529 chr7:117205661-117214375	ATCTTCACTAACATTTCCAATGTTGTCTAGCATAGACAAG
FOR_530 chr7:117214376-117215135	CTTGGGGCTCTAACTTAGTCATAATTTTTATAGATAAAAG
FOR_531 chr7:117215136-117216509	TTGACAAAATGTTCCTCAGTCATAGATCCTGGTTTTAAAG
FOR_532 chr7:117216510-117220056	GGATGTTCTTGTGCTCTTGCTCAAGGGCTTGTAGCTGAAG
FOR_534 chr7:117222053-117223390	AGTACTCACTTCCCCTCTTCATCCTATCTCCTCCTTTAAG
FOR_535 chr7:117223391-117228413	CACTGATGTAATAAACAAGTAAACAGGGGAGAAAGAAAAG
FOR_536 chr7:117228414-117229567	TAGTAATCACTTGAATATGAGGAAGGATTTCTTTTAAAAG
FOR_539 chr7:117230846-117234904	TGTCCAGTAGTTGCTTTCCTGGTCTTCTGTGTAGTCTAAG
FOR_541 chr7:117239628-117242516	CAATTAAAATGTGCTTTTCGGTCTACAGTTTTCTATCAAG
FOR_542 chr7:117242517-117244976	ACATTATAAAACACTACCAATTTCAAAACATACTGACAAG

FOR_544 chr7:117244985-117246173	AATTATGGCTAATATAATAAAGAATTTTTAAGGGTCAAAG
FOR_545 chr7:117246174-117247010	ATTTTCTGTGCAATGCTGAGAATGTCAGAGAAGAGGTAAG
FOR_547 chr7:117248239-117253247	CTCTGGATCCAAAATGGGAAGCCACGTGCCAGTGCCAAAG
FOR_548 chr7:117253248-117253903	ACTTACTGAAGTTCTAAATCATATCCTTTGTTTTCTCAAG
FOR_550 chr7:117254794-117259229	TAATTCTATCTGTGCTGCCTTCTCCACTTCCTTTTCTAAG
FOR_551 chr7:117259230-117259873	TCCGATACTACCACTCATCTCCTCCTTTCTCCCTGTGAAG
FOR_552 chr7:117259874-117278412	CTATTTGCTGTCACCAGTTAGAAAATAATGCACTCTTAAG
FOR_553 chr7:117278413-117279850	ATCCTATTAGAATCATACCCTAATTACAGAAAAGCACAAG
FOR_554 chr7:117279851-117280370	TGAAGTGCCGGACATATAACGAAAACCTAACAATATTAAG
FOR_555 chr7:117280371-117284325	TACGCAGATTTTGAATAGCCAATTAAATTTAATTCAAAAG
FOR_556 chr7:117284326-117284782	TTCAATGATGTATGCAATTTTTAATAACTTGAAATCAAAG
FOR_557 chr7:117284783-117285361	AAATAAAGAAATCACAATTCTCAATTTTAAACCAAAAAAG
FOR_559 chr7:117291311-117293839	TTTAAAATCCAGCAGGTTGTGCCTATTTTCCATCAGTAAG
FOR_560 chr7:117293840-117295506	GCAAGCAGACATTCAAATGTAATAACCATCCTTTAAGAAG
FOR_561 chr7:117295507-117295997	TACCAACAGAAAAGACAAATCTTTTTCCATGTTACATAAG
FOR_565 chr7:117301977-117310829	GACAGGCTCCACCTCACCACAGATCTCCAAGAAGGAGAAG
FOR_566 chr7:117310830-117311855	TGTCGAGATTCACAGCCAGCCCATCATGAATTGTGCAAAG
FOR_567 chr7:117311856-117316255	CTGCGGAAGTGAAGTATCCTTGGTGTCCCTACTGACAAAG
FOR_568 chr7:117316256-117316630	TCTCTAATTCTCACCTAATTCTTTTGATAATTCATGAAAG
FOR_569 chr7:117316631-117320129	GTCCTTAAACACTTTACCCCACACCCAGAGCCCAGACAAG
FOR_570 chr7:117320130-117320250	ATCCCCCACTTTTCTCTCCTTTATGAAGAGGTCATTTAAG
FOR_571 chr7:117320251-117324194	ATTGGTATGGGAAGTAGCCTGGGAATCAGGATTTTTAAAG
FOR_572 chr7:117324195-117324983	AATTCAGGATGATAAAAGCAAAATTATAAGTGTGTAGAAG
FOR_573 chr7:117324984-117329171	CATGAGAAGGCACTCTGTATAAACAAAACATCATCCAAAG
FOR_574 chr7:117329172-117329423	ACCTGAGGGGCTCCTGGAAAGAATAAAAGTTTTCTAGAAG
FOR_575 chr7:117329424-117332188	GCAAAGACCTTTTATACTTCTTTCTGTTTTCATGATCAAG
FOR_576 chr7:117332189-117334886	CATTTTTCAAGGATTGTTTAGGAACTATCTGTCTGTGAAG
FOR_577 chr7:117334887-117336857	CTGGCAATTTATGACATCGAATTGTGAAGTCTTAGAAAAG
FOR_579 chr7:117342270-117342424	GTGTTTAATTGGCTCTTCTATTAGTTTCATAAGATAGAAG
FOR_582 chr7:117347944-117356272	TTAGTATATGTGCTTGCTATGTAAATATTCAGTGAGCAAG
FOR_583 chr7:117356273-117356691	ACTGTGGATCTTTTTCTCAGATGAAGACACAGAGCAGAAG
FOR_584 chr7:117356692-117364472	CCTACCTTTTCATGATATACTTACTATACAGATCTCAAAG
FOR_586 chr7:117373712-117382856	TCTAATATTGTTTTGAACTTCAATATATTTTCCCACCAAG
FOR_587 chr7:117382857-117389863	CATAGTAACGCATGGCAGGAGCAGTCAGAACTCGGCCAAG
FOR_590 chr7:117397094-117404417	AGGCCAAATCTCATCAAGGTGATAAGCTCTGTTCTGAAAG
FOR_591 chr7:117404418-117404731	TGCTAGTGCTTGACTATTGTTCTGTTTTCTCCTTCCTAAG
FOR_592 chr7:117404732-117404962	CTTATTGAGCAGCTGCTGTGTGCTAAAAAAAGTGTTCAAG
FOR_593 chr7:117404963-117407209	TTTTAAGCTATTGTCATGGCCATTTCATCTTTGGGCCAAG
FOR_595 chr7:117415206-117418643	ATTGATTTGAGTATGCCTTTTTGGGAGTTGAGAGTTTAAG
FOR_596 chr7:117418644-117427061	TCCTTCTAAAGATGGAAAGATTCTTCCTTTGATGATTAAG

FOR_597 chr7:117427062-117428039	ACTAATTAGTCTGGCTCCGTGTACTTGAGAAGCTGGAAAG
FOR_598 chr7:117428040-117428313	GCATTTTTCATACTTGTTTCCGTAGCTGACTGCATACAAG
FOR_599 chr7:117428314-117428672	TCCGTCCCAAAAGGATGTCAGCTGAGGATACCTGGGTAAG
FOR_600 chr7:117428673-117433066	GCCATTACAACTTTAAAAAGTTTTCTGTTCAACAGGTAAG
FOR_601 chr7:117433067-117434754	GCCTGGGAATGTAATAAGGAGAATTTAAATCAGCTAAAAG
FOR_603 chr7:117438608-117451059	CACAACATGTAACTCTTTGGAAGATACCTGTTAGTATAAG
FOR_604 chr7:117451060-117451350	ACAAATATTCATATACTGCACAATCTTTGAAGGTGAGAAG
FOR_605 chr7:117451351-117457028	ACAGCATTCATACTGTTTTGAACCCTTTTCACATAAAAAG
FOR_606 chr7:117457029-117462764	GTCTTGGCCAAGTCAATGGGGACAGAAATAAAGAGATAAG
FOR_607 chr7:117462765-117462875	TGTAACTTTGGGATGGGGGTGAAGCTAGGGGTGAGGAAAG
FOR_608 chr7:117462876-117463473	TGGTTCTCATCTTGGCTGCATGTTGAAATCACCTAGAAAG
FOR_609 chr7:117463474-117469725	TTATTTATACTGTTTATTGTCTTTTCCACTATAATGCAAG
FOR_615 chr7:117476669-117480790	AGTGATTCTCATGGTATTATGTAAGATCACCACCAATAAG
FOR_616 chr7:117480791-117485009	TATGTAAGTGGGTATTGCATACCACTAAGTCTTGGGAAAG
FOR_617 chr7:117485010-117489451	TAACTGGTGATTGTTAGTATAAGGAATACTCTTGGGAAAG
FOR_623 chr7:117508289-117518070	TGGGAACATGACATTTTAGAATGTTTAACAGTCAAAAAAG
FOR_625 chr7:117524320-117532227	CATATTAACAAAATAGCGAATCATTATTTGGATGTATAAG
FOR_626 chr7:117532228-117537156	CTATCATGAGCCCACTTCATACCTAGGCTTCATGGTAAAG
FOR_627 chr7:117537157-117537712	TCTGTGTTGCATGTTCTGTTTTTAGTCTCAGCATCTAAAG
FOR_638 chr7:117565102-117572561	AGAACACTACAAAATAGGTGCTCAATAAGCAAATACCAAG
FOR_640 chr7:117572736-117574054	GTTAGAGGTTAAAGGCAGGAAAAAAAGAGCTAATACCAAG
FOR_641 chr7:117574055-117580157	GGTGTCTCTCTAGCCCTTATTTGCCAAGACCAAAATCAAG
FOR_645 chr7:117589977-117593454	CTCTTAATGATTTGTAAAAAAAGGTTATATTTATCTTAAG
FOR_646 chr7:117593455-117601132	TCCTCCCTTTGAACACCACTGTCAAACAATCAAACTGAAG
FOR_650 chr7:117614663-117615273	TTTGTCTTAGTTTATACCTTTAGCTGCTGGTCTACATAAG
FOR_651 chr7:117615274-117620404	TTTGTTTTTATGAAAGTGCTTAAAAATAGTAGTTGGCAAG
FOR_652 chr7:117620405-117623511	ACACCTATCGCCAGGGAACAGTATTCTTGTGCTTGTGAAG
FOR_653 chr7:117623512-117624942	TTATTTTAAAAAATAGCTCTCTGAATCTGACTTGGATAAG
FOR_654 chr7:117624943-117625814	GTAGCCCGGTAAACTGGTGTTTGAAGGCAATATATAAAAG
FOR_655 chr7:117625815-117627691	CACAAAAACCTCTCAGCCCTACACAGCATCTTTAAGAAAG
FOR_656 chr7:117627692-117633539	TAAAAAATGAGAAATTATGGTCTAAAATTCCCTGTATAAG
FOR_658 chr7:117634323-117636076	CGTACCATCTGGAGATGCAACTCATATCAACATGTTCAAG
FOR_660 chr7:117639998-117642099	CTGGTCCCGCGCAGGACCCTGAAAGTCTCGTCCCAGTAAG
FOR_661 chr7:117642100-117643094	TGAAACGAGGAGATGCATTCAAACCCTAGCAGTTAAGAAG
FOR_662 chr7:117643095-117645597	ATGGTTTAAAGATATACTCAGAACATAGATCATGTAGAAG
FOR_663 chr7:117645598-117648003	CAGTTTTCGTTATGTAAATACTCCCATTTCTGAGTGTAAG
FOR_669 chr7:117659169-117660333	AATTTCATCCTTAAGCCAAACTGTAAATGTGAACAGAAAG
FOR_670 chr7:117660334-117664128	ATAAGAAGCTGTCAGTGGTGAACCACCATGTTTAGTAAAG
FOR_671 chr7:117664129-117665277	TGTTTCTTTTATGTATGATTTTTTTCCATCACTTAAAAAG
FOR_672 chr7:117665278-117665764	GGAAAGTGTGTATTTGCCCCTCTTTGCAGTGGGAACTAAG

FOR_673 chr7:117665765-117667072	ATTTTCACCAGTCTGTTGACAGAGTTGGCTCAAGAGGAAG
FOR_674 chr7:117667073-117676995	GTGTATCTAATCAGTAAACACATAAACTCTTGCTGTAAAG
FOR_675 chr7:117676996-117681530	TCCTTGTGTATTTTCATGTTTCTTAAAAGTTAGTAAGAAG
FOR_676 chr7:117681531-117683117	CTTCCTTAATTTTAGAGGTATAGTCAGTGAGAAAATAAAG
FOR_677 chr7:117683118-117691760	GACAATATGTGAATAACCAGCAAACTCTAGCAACAATAAG
FOR_678 chr7:117691761-117694222	TGTGGAGATTTATAGGTGGATTCATTTTCCAGAGCTTAAG
FOR_681 chr7:117708238-117711933	ACATTCCTTATAAGTAGTGACCAACATGTTGTAACAGAAG
FOR_685 chr7:117724096-117725341	CTCTCAGGAAACAATGTGACTTGGATCTTTGAATCCAAAG
FOR_688 chr7:117733948-117739166	TCCTTGAACAGTTTCTGGCATTTAGTAGGGGAACAATAAG
FOR_692 chr7:117754462-117755196	ACCCTAGAAAGTAGTTGAGGCTCCAAGAAAATCTAAGAAG
FOR_693 chr7:117755197-117758934	CCAAACTGAGGAACTTTTTACAAAATACCTGACAGTTAAG
FOR_694 chr7:117758935-117759925	AGAGCTGTCAAAATTTTTAAAACTTGAAAAGGACTATAAG
FOR_696 chr7:117763482-117775069	AGTTGTTGATCCTGAGGTGTTTGCCCAAGAAGAGGAAAAG
FOR_697 chr7:117775070-117780876	ACTAAAAGTATCTGATACAAGCCTCAATCAATTTATGAAG
FOR_698 chr7:117780877-117785468	CACGTTCAAGTGTGTGAATAGCAGCTGCCTGAGTACAAAG
FOR_699 chr7:117785469-117787223	TGAGTGTATTGAAACAATTATGACTATTTCATAGGTGAAG
FOR_702 chr7:117795166-117795912	GAGTAACAGCAGCAATCAAGATGAATCATAATTACCCAAG
FOR_703 chr7:117795913-117796992	CAATAACAGGTACATGAGACCCAATATGCAGAGGTGAAAG
FOR_705 chr7:117797013-117800251	AATCTGTTTCTGCTTGTGACAGTAAATCTAGAACTGGAAG
FOR_707 chr7:117809275-117810127	TGTCATTTGAATTGCTGAGAAAAGAAATTCATTCTACAAG
FOR_708 chr7:117810128-117813248	TGCATCATTTTACATGTGCTTGTCAAGTATATAGTTTAAG
FOR_709 chr7:117813249-117815722	TGCCATCAGGAATGTTTGAATATACCATTTTCACCATAAG
FOR_710 chr7:117815723-117817378	GTTGAACTGGGTAGGTACCAAAAAGACTGTCTGTCACAAG
FOR_713 chr7:117827230-117831742	AGCCACATTTTGGGTCTTATTTCAAACTAGTAACACAAAG
FOR_714 chr7:117831743-117833241	CAAACTCATGAGGATATTTAACCATGAAAGTTTCCAGAAG
FOR_716 chr7:117833369-117845957	GACCAAGGTAGTAGAGCTGTTTTAAGATAACAAATAAAAG
FOR_717 chr7:117845958-117848439	CACAGAGTAGTTGTGGATCCCAGTGGTTGAATGGTAAAAG
FOR_718 chr7:117848440-117851061	ATTCACTAGGAATTATGAAAACAAAATCATCCATCTAAAG
FOR_720 chr7:117856727-117863427	CAAACTCGAAGAAACAAAGAGCCTGAGAGCTACACTCAAG
FOR_721 chr7:117863428-117865641	CTCATTATCCTCAAGACACGTAATCATTGGATTTTCCAAG
FOR_724 chr7:117869609-117870489	GTGTATTAATAGAAATTGTATATTTTTGTTGTGTCCCAAG
FOR_725 chr7:117870490-117873429	TTGTTACAATACCAAAATATATTATTCATAGACTCACAAG
FOR_726 chr7:117873430-117875326	CTGCATTAGGTCCTAAAAAGAGAGTTAGCCTGTCTTTAAG
FOR_728 chr7:117875455-117875879	GAGTCGGACTCTTGGCTTTCAACATGCTTTCCTGACCAAG
FOR_729 chr7:117875880-117883039	TATAGTTAACAGCTTATTATAGAAACTGTATAAAAAGAAG
FOR_731 chr7:117883765-117892970	GTTATAGTAAAACAGAAATCAGGCTGGTTTCAGCCTGAAG
FOR_733 chr7:117893193-117904369	TAGTTTACGAGCCACTGAAACTATTTCCAGGGCGAAGAAG
FOR_734 chr7:117904370-117904798	GGTAATTTATTTCACAATGTGGGAAATTGTATATGGAAAG
FOR_735 chr7:117904799-117905776	TTGGACCCCTTTGTGGATAAAGCAAGGGAGTCAAGCCAAG
FOR_737 chr7:117909651-117913178	AAATAGCATTTTGAATACTTATAATGATTTTATAAGTAAG

FOR_739 chr7:117914160-117915342	GATTTTGGAAAAGAATAAACATCAAGTGTTTGTAACAAAG
FOR_741 chr7:117918022-117919004	AGCTATTATACCATGAGGAAAAGCAACATGTTGAGCTAAG
FOR_742 chr7:117919005-117919940	GTTTACCCAACACTACTTCCTGTCTTGAAGACACGAGAAG
FOR_743 chr7:117919941-117924925	GATAAATAGAATTTTTGCCATGTCTGCAAGTTCTATAAAG
FOR_744 chr7:117924926-117927813	CTGGGATATGTGGTAAGAGCATGTTTAGTAGCATAAGAAG
FOR_745 chr7:117927814-117930607	TAAAATTAAAGGTGTCTCTGTGCTTAGCGAATTTATAAAG
FOR_746 chr7:117930608-117931069	TAATCTAATTAAGTAACAGTTCACAGAAGAAGACATCAAG
FOR_748 chr7:117936326-117941888	TTTGATGGATCCATTGCATTCTCTTAAAATGTAGTAGAAG
FOR_751 chr7:117943862-117946051	AAAGATTTTCTGATTGGCTATTTGGTTGAAAAGAGTTAAG
FOR_752 chr7:117946052-117947632	ATTACAATTAAAAGCAAAATCTCTTCCCAACTCAGAAAAG
FOR_756 chr7:117953828-117960889	TTAGAATAAAGACCTTTTGATTGATCATATTTGATTAAAG
FOR_759 chr7:117961105-117963456	GCCTCTCACACACACAACCTGGATTTCTAGCAAATATAAG
FOR_760 chr7:117963457-117965130	AATATAAAATATTTTACTATGACTTGTGTTTAAGAAAAAG
FOR_762 chr7:117973388-117988981	AAAAAAGTACACCTGTGTAGGCTACTTACTGTGAATGAAG
FOR_765 chr7:117992159-117992442	TAATAACAAAATTGAGAGACATTATACTTTAGTTTCAAAG
FOR_766 chr7:117992443-117993890	TGGAAAAAAATTTATTTGTTTCATGAAAGTTAAATATAAG
FOR_767 chr7:117993891-117994842	AGGTATATACTTCTAATCTTACATAATTTTTTACTCTAAG
FOR_768 chr7:117994843-117997059	AATTGCTTTTTCATTTTGTCGATTTTTTTTTGTACAGAAG
FOR_770 chr7:118009421-118009816	ACACAAAAATGATGGAAGAGGGAATCCACACTCTGGGAAG
FOR_771 chr7:118009817-118012221	AAATATTATTTACTCTGTTTTTCATATGAGAACACTAAAG
FOR_773 chr7:118012383-118015755	GTTCCGGACACCGAAGTGCTTTCTGGAGTGGCAACCCAAG
FOR_774 chr7:118015756-118017020	ATCAAAAGAGTAACTTAAAAATCTAGAATGAGAATTTAAG
FOR_775 chr7:118017021-118022505	TTCACTTTATCTCCTAAATTGCATTTTGCCTAGAAAAAAG
FOR_776 chr7:118022506-118023306	AGTACAAACAAGTAAAAAGCCAAACTGCTTCCACTTGAAG
FOR_778 chr7:118032257-118037031	AGGGAGAGACGAGCAATGCCATCATACAGTCTCCAGGAAG
FOR_779 chr7:118037032-118041118	AATTATATTATTCCAGGGCACTGAACGTTTACTTACAAAG
FOR_780 chr7:118041119-118044577	CACCCATGAGAGCAGTAAAAAGTGAGTTCAGTTATTTAAG
FOR_781 chr7:118044578-118057842	AATAAGCACAGAATCAAACTTATTTTAGACTTTTTTTAAG
FOR_783 chr7:118065172-118071565	TAGGGCCTGTGGACTTCTTATCTCAGAATCATCTTAAAAG
FOR_784 chr7:118071566-118078458	TAGAGAGAGACTTCCACTGTACAGCTGAAACAGGCATAAG
FOR_785 chr7:118078459-118079156	CCTTTCCTTCTGAAAGAACAGAATGGGCCTGTGCAGAAAG
FOR_786 chr7:118079157-118080510	TATTCACAACTATTATTTACTAAATTAGCCTCTTCACAAG
FOR_788 chr7:118080600-118082900	GACACAGGAATTAATTATAAAGAAGTTACGAATGGCCAAG
FOR_789 chr7:118082901-118086695	GGGCTGAGATCATGCAGCTTATAAGTGAAAAAAGATAAAG
FOR_790 chr7:118086696-118087830	GTTTCTTATAGATCAGATGTTAGAGTCCTTTGTAATAAAG
FOR_795 chr7:118110542-118116592	ATGGAAGTTAGAGTCTTTCACATGTTGGCAGAAATCTAAG
FOR_796 chr7:118116593-118118958	TTCGCTGTATTTGGAAATCGGGTTGGGATTTGAGTTAAAG
FOR_798 chr7:118118979-118122757	ATGTATTATGAAAAGGCATTCCAAGTTTTTTACAGGTAAG
FOR_799 chr7:118122758-118123465	GTATATGCTGCATAAACCTGAGCTGTTAACCAGTCTCAAG
FOR_801 chr7:118123896-118127179	ATCCCTAGTGAGTCCTATTTCAGATGTCTGACTCCAAAAG

FOR_802 chr7:118127180-118132591	ATGCCATTCATAAGTGGCTTGAATCTCAACCTTTCAGAAG
FOR_803 chr7:118132592-118133675	ATAATTGTACATAAAGTGATTCTTATCATGGTGAATAAAG
FOR_804 chr7:118133676-118134163	TGTAGTAAGGAGTACTTGTCTTTCTTGCTGCTCTATTAAG
FOR_805 chr7:118134164-118143444	TTTGTTAATATATACTGAAGCTGACCTTTATTGACAAAAG
FOR_806 chr7:118143445-118145173	CCCTTGGAAAAATCTGGAAGGAATTCACAGAGGAAGTAAG
FOR_807 chr7:118145174-118145507	TCAGGAAACTAACTATTTTCATGTTTCTACATGGTCAAAG
FOR_809 chr7:118153941-118170303	TCCTGTCTTCATTGGGGAAGGAGACACAAGACCCACTAAG
FOR_810 chr7:118170304-118172110	AGGGGTACGTGGAACGGAAGGGCCTTGCTGCAGGCTGAAG
FOR_811 chr7:118172111-118180123	GTTCAAATAACCCAAATCGAAATTGAACTTCACATTGAAG
FOR_817 chr7:118185016-118188949	ATGTGCTATTGTGCTGATACAATCATATCTTGTTCTCAAG
FOR_818 chr7:118188950-118192901	GAAATCACATGTTACAACTAGTAGATGAACTAAAATAAAG
FOR_820 chr7:118192926-118198234	AATCTTAAAATTGTTTTAATGAGTGAGATAAACTTAAAAG
FOR_821 chr7:118198235-118202713	ATTTATTTTACAATTATAAACTCAATTTAAAGTTTGAAAG
FOR_822 chr7:118202714-118208267	CCTCCATAAATGCATAATCCTTCTTATTTAGATTATTAAG
FOR_823 chr7:118208268-118210988	GGCAAGTGTACTATCTAAAATGATTTCTCGTTTCAATAAG
FOR_826 chr7:118215209-118221703	ATTTGGGGAATACTGTAAAAGGAAATCAAGCAAAAAGAAG
FOR_828 chr7:118222405-118222956	ACTCGTCCCCGGTTATCTCAGGGGCCTGGGAGAAGAGAAG
FOR_829 chr7:118222957-118224989	ACTAATCTGTGTTCCACTATCTGACTGTCTGACAAAGAAG
FOR_830 chr7:118224990-118225664	CCCTGACAGAGTTTTAAATACAGACAGAGCAGTGGAAAAG
FOR_832 chr7:118225695-118226169	TCCATTGAGACATTGCATTTCATGATAGAAAATTCTCAAG
FOR_833 chr7:118226170-118227462	TTTCGGGACAGGCCCCCTCTTAAGAAATTACTGGGTGAAG
FOR_841 chr7:118256172-118259200	GATAATCTCAAAGACAACCACACGTATACACATACTAAAG
FOR_842 chr7:118259201-118259827	GTAGGGTATTGTTAAGCACAATTAGGGTATTGTTAGGAAG
FOR_844 chr7:118262899-118266249	AATAATATCTCCCTATTGTTCATAATGTCTCCCTATTAAG
FOR_848 chr7:118275081-118276446	ATTAGAAAGGTGTTATGAGCCTTCTTCCAGACATACCAAG
FOR_849 chr7:118276447-118278095	GTGATCTAAGTCAGAGAGAGGATCAAGTCAGGACACCAAG
FOR_850 chr7:118278096-118280167	TCATTGCAACAGAAAATTGTTCCAAACTGTTATGATAAAG
FOR_851 chr7:118280168-118281466	CACATTTTTTAAAGTCGTCACTGTATTGTTGCAATAGAAG
FOR_859 chr7:118295522-118298023	ACCGGATTGTTTGGCTCAGATGAAGATTGGGTGTGTGAAG
FOR_860 chr7:118298024-118299632	TCATTACTTATATAATATTTAAACGATTAGCCATATAAAG
FOR_864 chr7:118305250-118305693	AAATATATAAATATTGTTTACTAAAGAAAACATAGTCAAG
FOR_871 chr7:118316092-118319933	TTATTTAGGTGATGAAAAATTAAAGTTTTTGCAAGTGAAG
FOR_872 chr7:118319934-118326087	AGCGGGGAGAGGTCAGAGCCACCCAACCCAGTTGCCAAAG
FOR_873 chr7:118326088-118326428	ATTAAAATATCACTGCTACTAGGCAGATAACAAATGAAAG
FOR_874 chr7:118326429-118330577	CATTTGAGTTCTTCCCAGTTTTGAACTATTACAAGTAAAG
FOR_877 chr7:118336293-118337454	GGTCCATTTCCAGAACTCCAACAAGGTCATCTCTTCAAAG
FOR_879 chr7:118339972-118341522	CCTTTAAGCCTCAAATTGATTACATAAATCTTGGAGGAAG
FOR_880 chr7:118341523-118342346	TTTATCAAAGATCACATTGTTAGGAAGTATTCAGATCAAG
FOR_883 chr7:118342476-118345933	TTAGAACACTGTTATTATCCTCAGAAGAGAAATCTACAAG
FOR_884 chr7:118345934-118348238	GTTCTGAGGATTGTGATAAGTGCTCTCTTGGATTCTGAAG

FOR_885 chr7:118348239-118349166	AAATTATTCTCATCTAGAGATTTGTATTGATATGCAGAAG
FOR_887 chr7:118349240-118351772	TAAGTCTTCTGACTGGAAAACTGAGTGTCCTTTTTCAAAG
FOR_890 chr7:118353269-118355818	TATTTATTCTAAGGCCACAGGACAAATTTCTGTCTGGAAG
FOR_891 chr7:118355819-118356080	TGGCAATGTTAGGCCACTTTTTCTGCAGGACTTTGTAAAG
FOR_893 chr7:118360340-118371174	TATTATAGTCAGGAGTAGAGATGAGTGGTAGTAGAGAAAG
FOR_894 chr7:118371175-118374774	AAAAAAACATTTAAGATGCTTTGCATTTCAGGGAGGAAAG
FOR_896 chr7:118380120-118382070	GGTATCAGTTTTAATATCTCCTGTTTTGTTTCGAATTAAG
FOR_897 chr7:118382071-118385396	GTTCTCAGCATCAGCTGCAATAATAGAAGGGGGACATAAG
FOR_900 chr7:118399267-118401405	AGATCAGAACGAGCTGTAATGATGAACAGAAACCAAAAAG
FOR_901 chr7:118401406-118405450	TTATCAGGAAAATGTAAAGTCTTGGCCTGGAGAGAAAAAG
REV_10 chr7:115634609-115644334	CTTAAGTTCAGTGTCATTCTTGTTCACAACTATAACCCCA
REV_13 chr7:115648832-115658204	CTTCGCTGAAGGAGAGGTGCATGAATTACAATGGAAAGTC
REV_17 chr7:115678230-115679502	CTTGAACTTATGCTATCACTTAGTTTAGTCTAAATCAATC
REV_28 chr7:115711863-115716752	CTTATTAATATGATAAATAATATATCTTAAGCGTAATGAA
REV_29 chr7:115716753-115718520	CTTCACTGCCTAGAGTTTGGGAAGGGTGACACAATAATAC
REV_46 chr7:115759209-115760110	CTTTATTTTTATCACCACCTGACAAATATTATTTCATTTT
REV_53 chr7:115771573-115775575	CTTTATAATCCATGTCTTTTTAATAATTATTCTATAACAA
REV_72 chr7:115835165-115835641	CTTGTGTTCATCTTGAGCCAATAAAAACTACAAAAAGGAA
REV_84 chr7:115870648-115874160	CTTATACAAATTTTTTGTGTATGTGGGGAGAGAGGGAGTG
REV_99 chr7:115922292-115929439	CTTTGGGAAGCCTAGCTTTGTTGACTATAAATATGGTACT
REV_111 chr7:115951445-115956484	CTTTGGGTGCTTTTAAGAAGCACCCAAAATTATTGTCATG
REV_143 chr7:116062231-116062713	CTTAGGCTACCGCGCGGGACTAACAAAAACATAAAAATCA
REV_157 chr7:116094748-116108495	CTTTCTCTGCTTCTGTGACATATTCCTGCTATTCTTCCTG
REV_172 chr7:116150033-116151991	CTTCCCTTGCTGATTTTTTTTCCGATTCCAAGGAGTAAAA
REV_189 chr7:116183536-116185921	CTTCTACTTACCTAATAAAAGATTTGGTTGGATGAATTTC
REV_203 chr7:116221966-116241644	CTTTATGTCAGGGGCTCACAGACACATTTGGTGTGCCTTA
REV_216 chr7:116288191-116290571	CTTACTGCTAGAGCCACGAAAGGAAAGAGAGAAGGACAGT
REV_236 chr7:116331386-116334860	CTTATTGTAACAGCTCCAATGTATTATAAAGAAGAACGTA
REV_253 chr7:116391564-116395299	CTTATTCATTTAGACCAGGAGTGGAGCTCCTCATGCTAGG
REV_274 chr7:116438321-116444567	CTTGTATCTTAGTGAGTGAGACAGAGAAACAAACACGTGA
REV_276 chr7:116447116-116449019	CTTCTACAGTGAAACAAAAGTAATTAACCTTTTTAAAAAG
REV_296 chr7:116515432-116528143	CTTTTCTAAAACCTTTTTGCCTTAAAGGTCATTTAATTCT
REV_305 chr7:116542199-116552833	CTTTAACTGTTTCATAACTTCAAAATGTTATTTTGAAACA
REV_330 chr7:116599917-116602592	CTTGTAAGAACTTTTGCAACGGCTTCTAAAAGAAGGTTCT
REV_342 chr7:116637097-116652178	CTTCCACCTGAACAGCTGTGATGGCACACCATCCCGATGT
REV_372 chr7:116744402-116752147	CTTTCGTTTATTCCCCTGCTTCACGGGGGAAGCGGGGCAG
REV_404 chr7:116852286-116855373	CTTCTAAATCTAAGGTGAGATCTAAGATCACATGAAATGT
REV_405 chr7:116855374-116855697	CTTTCAAATAATAGTGCATGGCTATTATCTTGAGATACAT
REV_417 chr7:116887587-116899362	CTTATAGTGTTCCATCTCTGTGCCAGGCAGTGGAGAGTTT
REV_420 chr7:116901865-116906597	CTTTATTAGTTTCAGGTTTAGGTGAGTGAACTCCAAGGGT

REV_421 chr7:116906598-116908131	CTTCTGTGGCTCATCACATTAGGCTATGAATACAATTTAT
REV_452 chr7:116987119-116988884	CTTATTTAGGCACAAAAAACCCTATTATATTATTTCATTT
REV_471 chr7:117038905-117042649	CTTCTTTTCATCAAAAGACTCTATAAGCCAAGCCACTGAC
REV_486 chr7:117073683-117083343	CTTCTGTCCCCCTGCCAAAGTCATATGTTGTTTAGGATTG
REV_520 chr7:117185297-117185608	CTTTGGAAGAACCCTGAAACAGTTTTAGATTATGGACAGC
REV_523 chr7:117187621-117190232	CTTTAGCACGAGTTGGACAGAGGCCTGAGGTAGGTGGTTA
REV_533 chr7:117220057-117222052	CTTGAGGAATAATCTATATTCCTGATTATCTATTTATGGG
REV_549 chr7:117253904-117254793	CTTAATATTGGGAATTTTTTTTCTGTTATGAGTTGAAATG
REV_564 chr7:117298280-117301976	CTTTGATCTGTAACCCTTGGACTGAACATTCTAGTCATTC
REV_594 chr7:117407210-117415205	CTTCAAAAAAAAAGTCACTTGGGGCAGTCATTAAAATACA
REV_613 chr7:117472370-117476601	CTTCCTCTAGTTGGAGGCATGTCTTCATGTCTGGGTTTAC
REV_618 chr7:117489452-117495550	CTTTCATTTAGCAATCATCAGTATCCACTAAGCACCATGT
REV_624 chr7:117518071-117524319	CTTGTTACTATAAGCAAATGAGAACGATTATAAGAAAAAA
REV_647 chr7:117601133-117608227	CTTATTCCAATACGGTGTTTTGTTCATGGCCATTTTTTAT
REV_648 chr7:117608228-117612532	CTTCTGTTTTTGTCTTTGTGATAATTACATGATTAAAAAA
REV_649 chr7:117612533-117614662	CTTAGTATTGAATAAAAGAAGACATGGGTCTAGCATAACA
REV_665 chr7:117648602-117650010	CTTTGATGTCATCCTTGACTCTTCTCTTTCTCAAAGGCGT
REV_666 chr7:117650011-117652129	CTTATTCATGGCGGTGCAGGCTGCTCGGAAAGGGAGTCTG
REV_667 chr7:117652130-117653051	CTTCTTCTTCAAAAAATTTAAAAGTTACTTGTTTTGATGT
REV_668 chr7:117653052-117659168	CTTAGTACTGGTTTCTAGGTTTTCAGGCTAACCGATGAGT
REV_680 chr7:117698217-117708237	CTTCTATTTGATCAGCTTTCCTGCAGATAGCTATAAACAT
REV_687 chr7:117729698-117733947	CTTATTTGTATAACTATAAATTTAGAATGAACTAACTTTC
REV_695 chr7:117759926-117763481	CTTAACCCAAAGGAATGAATTTATAACAAAAGAAATCCAG
REV_700 chr7:117787224-117795088	CTTTTTAATAAATCTGGATTTAACCAACTACCACTTAGGT
REV_711 chr7:117817379-117817857	CTTGTCATTATATTCTCAAGTTTCTGGGTAAAGGCCATAT
REV_719 chr7:117851062-117856726	CTTGGCAGCAAGAACAATTTTAAGATGACCCTGGGGTCTC
REV_730 chr7:117883040-117883764	CTTACTTATAAGATTCAAAACCCATATTGTAATCAACTTT
REV_738 chr7:117913179-117914159	CTTAAACAAGTAATGCTCATTTATTGTATGTAAATAGATT
REV_750 chr7:117942852-117943861	CTTGGAAGCATCTACAGCAGCCTTAGCAGTCCGTAACAAC
REV_761 chr7:117965131-117973387	CTTTACAGTAAGGTTAATGTGTAAATGTGCTAAATTAAAG
REV_769 chr7:117997060-118009420	CTTCTAAACTTCCTGTAATCCAGGCCACACCGGTTTTGTT
REV_777 chr7:118023307-118032256	CTTGGCACAGCAATAGAAACAATCAAATGAGTGAAAAGGG
REV_782 chr7:118057843-118065171	CTTAGGCAGATCAGCATGCTGTCAGGATGCACCTGTACTA
REV_792 chr7:118094746-118096542	CTTTGTTCTCCAAGAATCTTTTTTCTATACATATTTTAGA
REV_800 chr7:118123466-118123895	CTTAAGACATCTTTTGCTGGAAAGTAAAAAAGTGCCCAAG
REV_808 chr7:118145508-118153940	CTTTGGAAATTAGTGCATAGAACGACAAAGCTGAAAAGGT
REV_816 chr7:118182448-118185015	CTTTCAAAATAACAGTTTCATTCAGCTTTCTGATAGGTCA
REV_825 chr7:118212782-118215208	CTTATTTCTTAGTAAATGGCACCCATTAATTACTTAATGT
REV_840 chr7:118246300-118256171	CTTTGTTACTGCAGATAGAGATAGGTAATGGCTTTCCCTG
REV_847 chr7:118268800-118275080	CTTCAAGAAGGAAGCCTTCATATAAGGAAAAATGATGTTA

REV_863 chr7:118303374-118305249	CTTTTTTCTAAGACCTTGAAATCAAAAGCAATTCACTATA
REV_876 chr7:118335356-118336292	CTTATACTACTGAGAGACAAGATTGTTTAATTTTCAAATG
REV_892 chr7:118356081-118360339	CTTAGACAAGGCATTTTCCAAGAACTCGATGTATTAAGCC
REV_898 chr7:118385397-118394195	CTTGGAGGTTTTTCTAGGCTAGAAAAAAAGAGAAATGCTC

BIBLIOGRAPHY

Ahmadiyeh, N., Pomerantz, M.M., Grisanzio, C., Herman, P., Jia, L., Almendro, V., He, H.H., Brown, M., Liu, X.S., Davis, M., et al. (2010). 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. Proc. Natl. Acad. Sci. *107*, 9742–9746.

Allison, A.C. (1954). Protection Afforded by Sickle-cell Trait Against Subtertian Malarial Infection. Br. Med. J. *1*, 290–294.

Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H., and Shiroishi, T. (2009). Chromosomal Dynamics at the Shh Locus: Limb Bud-Specific Differential Regulation of Competence and Active Transcription. Dev. Cell *16*, 47–57.

American College of Obstetricians and Gynecologists (2011). Update on Carrier Screening for Cystic Fibrosis.

Bacher, C.P., Guggiari, M., Brors, B., Augui, S., Clerc, P., Avner, P., Eils, R., and Heard, E. (2006). Transient colocalization of X-inactivation centres accompanies the initiation of X inactivation. Nat. Cell Biol. *8*, 293–299.

Bantignies, F., Roure, V., Comet, I., Leblanc, B., Schuettengruber, B., Bonnet, J., Tixier, V., Mas, A., and Cavalli, G. (2011). Polycomb-Dependent Regulatory Contacts between Distant Hox Loci in Drosophila. Cell *144*, 214–226.

Baù, D., Sanyal, A., Lajoie, B.R., Capriotti, E., Byron, M., Lawrence, J.B., Dekker, J., and Marti-Renom, M.A. (2011). The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. Nat. Struct. Mol. Biol. *18*, 107–114.

Bear, C.E., Li, C., Kartner, N., Bridges, R.J., Jensen, T.J., Ramjeesingh, M., and Riordan, J.R. (1992). Purification and functional reconstitution of the cystic fibrosis transmembrane conductance regulator (CFTR). Cell *68*, 809–818.

Beildeck, M.E., Islam, M., Shah, S., Welsh, J., and Byers, S.W. (2009). Control of TCF-4 Expression by VDR and Vitamin D in the Mouse Mammary Gland and Colorectal Cancer Cell Lines. PLoS ONE *4*, e7872.

Belkin, R.A., Henig, N.R., Singer, L.G., Chaparro, C., Rubenstein, R.C., Xie, S.X., Yee, J.Y., Kotloff, R.M., Lipson, D.A., and Bunin, G.R. (2006). Risk Factors for Death of Patients with Cystic Fibrosis Awaiting Lung Transplantation. Am. J. Respir. Crit. Care Med. *173*, 659–666.

Belmont, A.S. (2001). Visualizing chromosome dynamics with GFP. Trends Cell Biol. *11*, 250–257.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799–816.

Blackledge, N.P., Carter, E.J., Evans, J.R., Lawson, V., Rowntree, R.K., and Harris, A. (2007). CTCF mediates insulator function at the CFTR locus. Biochem. J. *408*, 267–275.

Blackledge, N.P., Ott, C.J., Gillen, A.E., and Harris, A. (2009). An insulator element 3' to the CFTR gene binds CTCF and reveals an active chromatin hub in primary cells. Nucleic Acids Res. *37*, 1086–1094.

Bobadilla, J.L., Macek, M., Fine, J.P., and Farrell, P.M. (2002). Cystic fibrosis: A worldwide analysis of CFTR mutations—correlation with incidence data and application to screening. Hum. Mutat. *19*, 575–606.

Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Müller, S., Eils, R., Cremer, C., Speicher, M.R., et al. (2005). Three-Dimensional Maps of All Chromosomes in Human Male Fibroblast Nuclei and Prometaphase Rosettes. PLoS Biol *3*, e157.

Bortle, K.V., Nichols, M.H., Li, L., Ong, C.-T., Takenaka, N., Qin, Z.S., and Corces, V.G. (2014). Insulator function and topological domain border strength scale with architectural protein occupancy. Genome Biol. *15*, R82.

Boyle, S., Gilchrist, S., Bridger, J.M., Mahy, N.L., Ellis, J.A., and Bickmore, W.A. (2001). The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. Hum. Mol. Genet. *10*, 211–219.

Boyle, S., Rodesch, M.J., Halvensleben, H.A., Jeddeloh, J.A., and Bickmore, W.A. (2011). Fluorescence in situ hybridization with high-complexity repeat-free oligonucleotide probes generated by massively parallel synthesis. Chromosome Res. *19*, 901–909.

Branco, M.R., and Pombo, A. (2006). Intermingling of Chromosome Territories in Interphase Suggests Role in Translocations and Transcription-Dependent Associations. PLoS Biol *4*, e138.

Brennan, S. (2008). Innate immune activation and cystic fibrosis. Paediatr. Respir. Rev. *9*, 271–280.

Carter, D., Chakalova, L., Osborne, C.S., Dai, Y., and Fraser, P. (2002). Long-range chromatin regulatory interactions in vivo. Nat. Genet. *32*, 623–626.

Chambeyron, S., and Bickmore, W.A. (2004). Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. Genes Dev. *18*, 1119–1130.

Chen, H., Ruan, Y.C., Xu, W.M., Chen, J., and Chan, H.C. (2012). Regulation of male fertility by CFTR and implications in male infertility. Hum. Reprod. Update *18*, 703–713.

Cheng, S.H., Fang, S.L., Zabner, J., Marshall, J., Piraino, S., Schiavi, S.C., Jefferson, D.M., Welsh, M.J., and Smith, A.E. (1995). Functional activation of the cystic fibrosis trafficking mutant delta F508-CFTR by overexpression. Am. J. Physiol. - Lung Cell. Mol. Physiol. *268*, L615–L624.

Childers, M., Eckel, G., Himmel, A., and Caldwell, J. (2007). A new model of cystic fibrosis pathology: Lack of transport of glutathione and its thiocyanate conjugates. Med. Hypotheses *68*, 101–112.

Collins, F.S., Drumm, M.L., Cole, J.L., Lockwood, W.K., Vande Woude, G.F., and Iannuzzi, M.C. (1987). Construction of a general human chromosome jumping library, with application to cystic fibrosis. Science *235*, 1046–1049.

Comet, I., Schuettengruber, B., Sexton, T., and Cavalli, G. (2011). A chromatin insulator driving three-dimensional Polycomb response element (PRE) contacts and Polycomb association with the chromatin fiber. Proc. Natl. Acad. Sci. *108*, 2294–2299.

Cook, P.R. (1999). The Organization of Replication and Transcription. Science *284*, 1790–1795.

Cormet-Boyaka, E., Jablonsky, M., Naren, A.P., Jackson, P.L., Muccio, D.D., and Kirk, K.L. (2004). Rescuing cystic fibrosis transmembrane conductance regulator (CFTR)-processing mutants by transcomplementation. Proc. Natl. Acad. Sci. U. S. A. *101*, 8221–8226.

Corput, M.P.C. van de, Boer, E. de, Knoch, T.A., Cappellen, W.A. van, Quintanilla, A., Ferrand, L., and Grosveld, F.G. (2012). Super-resolution imaging reveals three-dimensional folding dynamics of the β-globin locus upon gene activation. J. Cell Sci. *125*, 4630–4639.

Cremer, T., and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat. Rev. Genet. *2*, 292–301.

Croft, J.A., Bridger, J.M., Boyle, S., Perry, P., Teague, P., and Bickmore, W.A. (1999). Differences in the Localization and Morphology of Chromosomes in the Human Nucleus. J. Cell Biol. *145*, 1119–1131.

Cuthbert, A.W., Halstead, J., Ratcliff, R., Colledge, W.H., and Evans, M.J. (1995). The genetic advantage hypothesis in cystic fibrosis heterozygotes: a murine study. J. Physiol. *482*, 449–454.

Cystic Fibrosis Foundation (2014). Cystic Fibrosis Foundation.

Davison, L.J., Wallace, C., Cooper, J.D., Cope, N.F., Wilson, N.K., Smyth, D.J., Howson, J.M.M., Saleh, N., Al-Jeffery, A., Angus, K.L., et al. (2012). Long-range DNA looping and

gene expression analyses identify DEXI as an autoimmune disease candidate gene. Hum. Mol. Genet. *21*, 322–333.

Dean, A. (2011). In the loop: long range chromatin interactions and gene regulation. Brief. Funct. Genomics *10*, 3–10.

de Andrade Pinto, Ana C.O., Barbosa, Carolina M.L., Ornellas, Debora S., Novaira, Horacio J., de Souza-Menezes, Jackson, Ortiga-Carvalho, Tania M., Fong, Peying, and Morales, Marcelo M. (2007). Thyroid Hormones Stimulate Renal Expression of CFTR. Cell. Physiol. Biochem. *20*, 83–90.

Dekker, J. (2006). The three “C” s of chromosome conformation capture: controls, controls, controls. Nat. Methods *3*, 17–21.

Dekker, J. (2007). GC- and AT-rich chromatin domains differ in conformation and histone modification status and are differentially modulated by Rpd3p. Genome Biol. *8*, R116.

Dekker, J. (2008a). Gene Regulation in the Third Dimension. Science *319*, 1793–1794.

Dekker, J. (2008b). Mapping in Vivo Chromatin Interactions in Yeast Suggests an Extended Chromatin Fiber with Regional Variation in Compaction. J. Biol. Chem. *283*, 34532–34540.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing Chromosome Conformation. Science *295*, 1306–1311.

Deplancke, B., Dupuy, D., Vidal, M., and Walhout, A.J.M. (2004). A Gateway-Compatible Yeast One-Hybrid System. Genome Res. *14*, 2093–2101.

Deplancke, B., Mukhopadhyay, A., Ao, W., Elewa, A.M., Grove, C.A., Martinez, N.J., Sequerra, R., Doucette-Stamm, L., Reece-Hoyes, J.S., Hope, I.A., et al. (2006). A Gene-Centered C. elegans Protein-DNA Interaction Network. Cell *125*, 1193–1205.

D’haene, B., Attanasio, C., Beysen, D., Dostie, J., Lemire, E., Bouchard, P., Field, M., Jones, K., Lorenz, B., Menten, B., et al. (2009). Disease-Causing 7.4 kb Cis-Regulatory Deletion Disrupting Conserved Non-Coding Sequences and Their Interaction with the FOXL2 Promotor: Implications for Mutation Screening. PLoS Genet *5*, e1000522.

Dily, F.L., Baù, D., Pohl, A., Vicent, G.P., Serra, F., Soronellas, D., Castellano, G., Wright, R.H.G., Ballare, C., Filion, G., et al. (2014). Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. Genes Dev. *28*, 2151–2162.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Dorschner, M.O., Hawrylycz, M., Humbert, R., Wallace, J.C., Shafer, A., Kawamoto, J., Mack, J., Hall, R., Goldy, J., Sabo, P.J., et al. (2004). High-throughput localization of functional elements by quantitative chromatin profiling. Nat. Methods *1*, 219–225.

Dostie, J., and Dekker, J. (2007). Mapping networks of physical interactions between genomic elements using 5C technology. Nat. Protoc. *2*, 988–1002.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., et al. (2006). Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. Genome Res. *16*, 1299–1309.

Drissen, R., Palstra, R.-J., Gillemans, N., Splinter, E., Grosveld, F., Philipsen, S., and Laat, W. de (2004). The active spatial organization of the β-globin locus requires the transcription factor EKLF. Genes Dev. *18*, 2485–2490.

Drumm, M.L., Pope, H.A., Cliff, W.H., Rommens, J.M., Marvin, S.A., Tsui, L.C., Collins, F.S., Frizzell, R.A., and Wilson, J.M. (1990). Correction of the cystic fibrosis defect in vitro by retrovirus-mediated gene transfer. Cell *62*, 1227–1233.

Du, Q., Li, Z., Pan, Y., Liu, X., Pan, B., and Wu, B. (2014). The CFTR M470V, Intron 8 Poly-T, and 8 TG-Repeats Detection in Chinese Males with Congenital Bilateral Absence of the Vas Deferens. BioMed Res. Int. *2014*, 689185.

Fraser, P., and Bickmore, W. (2007). Nuclear organization of the genome and the potential for gene regulation. Nature *447*, 413–417.

French, J.D., Ghoussaini, M., Edwards, S.L., Meyer, K.B., Michailidou, K., Ahmed, S., Khan, S., Maranian, M.J., O'Reilly, M., Hillman, K.M., et al. (2013). Functional Variants at the 11q13 Risk Locus for Breast Cancer Regulate Cyclin D1 Expression through Long-Range Enhancers. Am. J. Hum. Genet. *92*, 489–503.

Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., et al. (2009). An oestrogen-receptor-α-bound human chromatin interactome. Nature *462*, 58–64.

Gabriel, S.E., Brigman, K.N., Koller, B.H., Boucher, R.C., and Stutts, M.J. (1994). Cystic fibrosis heterozygote resistance to cholera toxin in the cystic fibrosis mouse model. Science *266*, 107–109.

Gheldof, N., Tabuchi, T.M., and Dekker, J. (2006). The active FMR1 promoter is associated with a large domain of altered chromatin conformation with embedded local histone modifications. Proc. Natl. Acad. Sci. *103*, 12463–12468.

Gheldof, N., Smith, E.M., Tabuchi, T.M., Koch, C.M., Dunham, I., Stamatoyannopoulos, J.A., and Dekker, J. (2010). Cell-type-specific long-range looping interactions identify distant regulatory elements of the CFTR gene. Nucleic Acids Res. *38*, 4325–4336.

Gholami, K., Muniandy, S., and Salleh, N. (2012). Progesterone Downregulates Oestrogen-Induced Expression of CFTR and SLC26A6 Proteins and mRNA in Rats’ Uteri. BioMed Res. Int. *2012*, e596084.

Gibcus, J.H., and Dekker, J. (2013). The Hierarchy of the 3D Genome. Mol. Cell *49*, 773–782.

Giorgetti, L., Galupa, R., Nora, E.P., Piolot, T., Lam, F., Dekker, J., Tiana, G., and Heard, E. (2014). Predictive Polymer Modeling Reveals Coupled Fluctuations in Chromosome Conformation and Transcription. Cell *157*, 950–963.

Gorkin, D.U., Leung, D., and Ren, B. (2014). The 3D Genome in Transcriptional Regulation and Pluripotency. Cell Stem Cell *14*, 762–775.

Gosalia, N., Neems, D., Kerschner, J.L., Kosak, S.T., and Harris, A. (2014). Architectural proteins CTCF and cohesin have distinct roles in modulating the higher order structure and expression of the CFTR locus. Nucleic Acids Res. gku648.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W., et al. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature *453*, 948–951.

Guilbault, C., Saeed, Z., Downey, G.P., and Radzioch, D. (2007). Cystic Fibrosis Mouse Models. Am. J. Respir. Cell Mol. Biol. *36*, 1–7.

Hagège, H., Klous, P., Braem, C., Splinter, E., Dekker, J., Cathala, G., de Laat, W., and Forné, T. (2007). Quantitative analysis of chromosome conformation capture assays (3C-qPCR). Nat. Protoc. *2*, 1722–1733.

Hakim, O., and Misteli, T. (2012). SnapShot: Chromosome Conformation Capture. Cell *148*, 1068–1068.e2.

Harismendy, O., Notani, D., Song, X., Rahim, N.G., Tanasa, B., Heintzman, N., Ren, B., Fu, X.-D., Topol, E.J., Rosenfeld, M.G., et al. (2011). 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. Nature *470*, 264–268.

Hatzis, P., Flier, L.G. van der, Driel, M.A. van, Guryev, V., Nielsen, F., Denissov, S., Nijman, I.J., Koster, J., Santo, E.E., Welboren, W., et al. (2008). Genome-Wide Pattern of TCF7L2/TCF4 Chromatin Occupancy in Colorectal Cancer Cells. Mol. Cell. Biol. *28*, 2732–2744.

He, Q., Chen, H., Wong, C.H.Y., Tsang, L.L., and Chan, H.C. (2010). Regulatory mechanism underlying cyclic changes in mouse uterine bicarbonate secretion: role of estrogen. Reproduction *140*, 903–910.

Heckman, K.L., and Pease, L.R. (2007). Gene splicing and mutagenesis by PCR-driven overlap extension. Nat. Protoc. *2*, 924–932.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. *39*, 311–318.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108–112.

Högenauer, C., Santa Ana, C.A., Porter, J.L., Millard, M., Gelfand, A., Rosenblatt, R.L., Prestidge, C.B., and Fordtran, J.S. (2000). Active Intestinal Chloride Secretion in Human Carriers of Cystic Fibrosis Mutations: An Evaluation of the Hypothesis That Heterozygotes Have Subnormal Active Intestinal Chloride Secretion. Am. J. Hum. Genet. *67*, 1422–1427.

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.

Hsu, P.D., Lander, E.S., and Zhang, F. (2014). Development and Applications of CRISPR-Cas9 for Genome Engineering. Cell *157*, 1262–1278.

Iborra, F.J., Pombo, A., Jackson, D.A., and Cook, P.R. (1996). Active RNA polymerases are localized within discrete transcription “factories’ in human nuclei. J. Cell Sci. *109*, 1427–1436.

Jiang, L.-Y., Shan, J.-J., Tong, X.-M., Zhu, H.-Y., Yang, L.-Y., Zheng, Q., Luo, Y., Shi, Q.-X., and Zhang, S.-Y. (2013). Cystic fibrosis transmembrane conductance regulator is correlated closely with sperm progressive motility and normal morphology in healthy and fertile men with normal sperm parameters. Andrologia.

Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.-A., Schmitt, A.D., Espinoza, C.A., and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature *advance online publication*.

Kerem, B., Rommens, J.M., Buchanan, J.A., Markiewicz, D., Cox, T.K., Chakravarti, A., Buchwald, M., and Tsui, L.C. (1989). Identification of the cystic fibrosis gene: genetic analysis. Science *245*, 1073–1080.

Kerschner, J.L., and Harris, A. (2012). Transcriptional networks driving enhancer function in the CFTR gene. Biochem. J. *446*, 203–212.

Kim, C.G., Chung, I.-Y., Lim, Y., Lee, Y.H., and Shin, S.Y. (2011). A Tcf/Lef element within the enhancer region of the human NANOG gene plays a role in promoter activation. Biochem. Biophys. Res. Commun. *410*, 637–642.

Kim, S.-I., Bultman, S.J., Kiefer, C.M., Dean, A., and Bresnick, E.H. (2009). BRG1 requirement for long-range interaction of a locus control region with a downstream promoter. Proc. Natl. Acad. Sci. pnas.0806420106.

Kleinjan, D.A., and van Heyningen, V. (2005). Long-Range Control of Gene Expression: Emerging Mechanisms and Disruption in Disease. Am. J. Hum. Genet. *76*, 8–32.

Kleinjan, D.A., Seawright, A., Schedl, A., Quinlan, R.A., Danes, S., and Heyningen, V. van (2001). Aniridia-associated translocations, DNase hypersensitivity, sequence comparison and transgenic analysis redefine the functional domain of PAX6. Hum. Mol. Genet. *10*, 2049–2059.

Koch, C.M., Andrews, R.M., Flicek, P., Dillon, S.C., Karaöz, U., Clelland, G.K., Wilcox, S., Beare, D.M., Fowler, J.C., Couttet, P., et al. (2007). The landscape of histone modifications across 1% of the human genome in five human cell lines. Genome Res. *17*, 691–707.

Koh, James, first, Francis S. Collins, last, and Thomas J. Sferra (1993). Characterization of the Cystic Fibrosis Transmembrane Conductance Regulator Promoter Region. J. Biol. Chem. *268*, 15912–15921.

Korinek, V., Barker, N., Willert, K., Molenaar, M., Roose, J., Wagenaar, G., Markman, M., Lamers, W., Destree, O., and Clevers, H. (1998). Two Members of the Tcf Family Implicated in Wnt/β-Catenin Signaling during Embryogenesis in the Mouse. Mol. Cell. Biol. *18*, 1248–1256.

Krauss, R.D., Bubien, J.K., Drumm, M.L., Zheng, T., Peiper, S.C., Collins, F.S., Kirk, K.L., Frizzell, R.A., and Rado, T.A. (1992). Transfection of wild-type CFTR into cystic fibrosis lymphocytes restores chloride conductance at G1 of the cell cycle. EMBO J. *11*, 875–883.

Kurukuti, S., Tiwari, V.K., Tavoosidana, G., Pugacheva, E., Murrell, A., Zhao, Z., Lobanenkov, V., Reik, W., and Ohlsson, R. (2006). CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. Proc. Natl. Acad. Sci. *103*, 10684–10689.

Laat, W. de, and Grosveld, F. (2003). Spatial organization of gene expression: the active chromatin hub. Chromosome Res. *11*, 447–459.

Lajoie, B.R., van Berkum, N.L., Sanyal, A., and Dekker, J. (2009). My5C: web tools for chromosome conformation capture studies. Nat. Methods *6*, 690–691.

Langer-Safer, P.R., Levine, M., and Ward, D.C. (1982). Immunological method for mapping genes on Drosophila polytene chromosomes. Proc. Natl. Acad. Sci. *79*, 4381–4385.

Lettice, L.A., Heaney, S.J.H., Purdie, L.A., Li, L., Beer, P. de, Oostra, B.A., Goode, D., Elgar, G., Hill, R.E., and Graaff, E. de (2003). A long-range Shh enhancer regulates

expression in the developing limb and fin and is associated with preaxial polydactyly. Hum. Mol. Genet. *12*, 1725–1735.

Levings, P.P., and Bungert, J. (2002). The human β-globin locus control region. Eur. J. Biochem. *269*, 1589–1599.

Lewandowska, M.A., Costa, F.F., Bischof, J.M., Williams, S.H., Soares, M.B., and Harris, A. (2010). Multiple Mechanisms Influence Regulation of the Cystic Fibrosis Transmembrane Conductance Regulator Gene Promoter. Am. J. Respir. Cell Mol. Biol. *43*, 334–341.

Lieberman-Aiden, E., Berkum, N.L. van, Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science *326*, 289–293.

Markaki, Y., Smeets, D., Fiedler, S., Schmid, V.J., Schermelleh, L., Cremer, T., and Cremer, M. (2012). The potential of 3D-FISH and super-resolution structured illumination microscopy for studies of 3D nuclear architecture. BioEssays *34*, 412–426.

Maston, G.A., Evans, S.K., and Green, M.R. (2006). Transcriptional Regulatory Elements in the Human Genome. Annu. Rev. Genomics Hum. Genet. *7*, 29–59.

Matthews, R.P., and McKnight, G.S. (1996). Characterization of the cAMP Response Element of the Cystic Fibrosis Transmembrane Conductance Regulator Gene Promoter. J. Biol. Chem. *271*, 31869–31877.

McCallum, T.J., Milunsky, J.M., Cunningham, D.L., Harris, D.H., Maher, T.A., and Oates, R.D. (2000). Fertility in men with cystic fibrosis*: An update on current surgical practices and outcomes. Chest *118*, 1059–1062.

McCarthy, V.A., and Harris, A. (2005). The CFTR gene and regulation of its expression. Pediatr. Pulmonol. *40*, 1–8.

McCord, R.P., Zhou, V.W., Yuh, T., and Bulyk, M.L. (2011). Distant cis-regulatory elements in human skeletal muscle differentiation. Genomics *98*, 401–411.

McDonald, R.A., Matthews, R.P., Idzerda, R.L., and McKnight, G.S. (1995). Basal expression of the cystic fibrosis transmembrane conductance regulator gene is dependent on protein kinase A activity. Proc. Natl. Acad. Sci. U. S. A. *92*, 7560–7564.

Miele, A., and Dekker, J. (2008). Long-range chromosomal interactions and gene regulation. Mol. Biosyst. *4*, 1046–1057.

Miele, A., Bystricky, K., and Dekker, J. (2009). Yeast Silent Mating Type Loci Form Heterochromatic Clusters through Silencer Protein-Dependent Long-Range Interactions. PLoS Genet *5*, e1000478.

Misteli, T., and Soutoglou, E. (2009). The emerging role of nuclear architecture in DNA repair and genome maintenance. Nat. Rev. Mol. Cell Biol. *10*, 243–254.

Modiano, G., Ciminelli, B.M., and Pignatti, P.F. (2006). Cystic Fibrosis: Cystic fibrosis and lactase persistence: a possible correlation. Eur. J. Hum. Genet. *15*, 255–259.

Mogayzel Jr., P.J., and Ashlock, M.A. (2000). CFTR Intron 1 Increases Luciferase Expression Driven by CFTR 5′-Flanking DNA in a Yeast Artificial Chromosome. Genomics *64*, 211–215.

Mokry, M., Hatzis, P., de Bruijn, E., Koster, J., Versteeg, R., Schuijers, J., van de Wetering, M., Guryev, V., Clevers, H., and Cuppen, E. (2010). Efficient Double Fragmentation ChIP-seq Provides Nucleotide Resolution Protein-DNA Binding Profiles. PLoS ONE *5*, e15092.

Montavon, T., Soshnikova, N., Mascrez, B., Joye, E., Thevenet, L., Splinter, E., de Laat, W., Spitz, F., and Duboule, D. (2011). A Regulatory Archipelago Controls Hox Genes Transcription in Digits. Cell *147*, 1132–1145.

Moskwa, P., Lorentzen, D., Excoffon, K.J.D.A., Zabner, J., McCray, P.B., Nauseef, W.M., Dupuy, C., and Bánfi, B. (2007). A Novel Host Defense System of Airways Is Defective in Cystic Fibrosis. Am. J. Respir. Crit. Care Med. *175*, 174–183.

Mouchel, N., Henstra, S.A., McCarthy, V.A., Williams, S.H., Phylactides, M., and Harris, A. (2004). HNF1alpha is involved in tissue-specific regulation of CFTR gene expression. Biochem. J. *378*, 909–918.

Murrell, A., Heeson, S., and 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.

Naumova, N., Imakaev, M., Fudenberg, G., Zhan, Y., Lajoie, B.R., Mirny, L.A., and Dekker, J. (2013). Organization of the Mitotic Chromosome. Science 1236083.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.

Nuthall, H.N., Vassaux, G., Huxley, C., and Harris, A. (1999a). Analysis of a DNase I hypersensitive site located −20.9 kb upstream of the CFTR gene. Eur. J. Biochem. *266*, 431–443.

Nuthall, H.N., Moulin, D.S., Huxley, C., and Harris, A. (1999b). Analysis of DNase-I-hypersensitive sites at the 3’ end of the cystic fibrosis transmembrane conductance regulator gene (CFTR). Biochem. J. *341*, 601–611.

Osborne, C.S., Chakalova, L., Brown, K.E., Carter, D., Horton, A., Debrand, E., Goyenechea, B., Mitchell, J.A., Lopes, S., Reik, W., et al. (2004). Active genes

dynamically colocalize to shared sites of ongoing transcription. Nat. Genet. *36*, 1065–1071.

Ostedgaard, L.S., Meyerholz, D.K., Chen, J.-H., Pezzulo, A.A., Karp, P.H., Rokhlina, T., Ernst, S.E., Hanfland, R.A., Reznikov, L.R., Ludwig, P.S., et al. (2011). The ΔF508 Mutation Causes CFTR Misprocessing and Cystic Fibrosis–Like Disease in Pigs. Sci. Transl. Med. *3*, 74ra24–ra74ra24.

Ott, C.J., Suszko, M., Blackledge, N.P., Wright, J.E., Crawford, G.E., and Harris, A. (2009a). A complex intronic enhancer regulates expression of the CFTR gene by direct interaction with the promoter. J. Cell. Mol. Med. *13*, 680–692.

Ott, C.J., Blackledge, N.P., Kerschner, J.L., Leir, S.-H., Crawford, G.E., Cotton, C.U., and Harris, A. (2009b). Intronic enhancers coordinate epithelial-specific looping of the active CFTR locus. Proc. Natl. Acad. Sci. U. S. A. *106*, 19934–19939.

Palstra, R.-J., Tolhuis, B., Splinter, E., Nijmeijer, R., Grosveld, F., and de Laat, W. (2003). The β-globin nuclear compartment in development and erythroid differentiation. Nat. Genet. *35*, 190–194.

Paul, T., Li, S., Khurana, S., Leleiko, N.S., and Walsh, M.J. (2007). The epigenetic signature of CFTR expression is co-ordinated via chromatin acetylation through a complex intronic element. Biochem. J. *408*, 317.

Phillips, J.E., and Corces, V.G. (2009). CTCF: Master Weaver of the Genome. Cell *137*, 1194–1211.

Phillips-Cremins, J.E., Sauria, M.E.G., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S.K., Ong, C.-T., Hookway, T.A., Guo, C., Sun, Y., et al. (2013). Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage Commitment. Cell *153*, 1281–1295.

Pier, G.B., Grout, M., Zaidi, T., Meluleni, G., Mueschenborn, S.S., Banting, G., Ratcliff, R., Evans, M.J., and Colledge, W.H. (1998). Salmonella typhi uses CFTR to enter intestinal epithelial cells. Nature *393*, 79–82.

Pittman, N., Shue, G., Neal, S.L., and Walsh, M.J. (1995). Transcription of Cystic Fibrosis Transmembrane Conductance Regulator Requires a CCAAT-like Element for both Basal and cAMP-mediated Regulation. J. Biol. Chem. *270*, 28848–28857.

Quinton, P.M. (1999). Physiological Basis of Cystic Fibrosis: A Historical Perspective. Physiol. Rev. *79*, S3–S22.

Reece-Hoyes, J.S., Barutcu, A.R., McCord, R.P., Jeong, J.S., Jiang, L., MacWilliams, A., Yang, X., Salehi-Ashtiani, K., Hill, D.E., Blackshaw, S., et al. (2011). Yeast one-hybrid assays for gene-centered human gene regulatory network mapping. Nat. Methods *8*, 1050–1052.

Rich, D.P., Anderson, M.P., Gregory, R.J., Cheng, S.H., Paul, S., Jefferson, D.M., McCann, J.D., Klinger, K.W., Smith, A.E., and Welsh, M.J. (1990). Expression of cystic fibrosis transmembrane conductance regulator corrects defective chloride channel regulation in cystic fibrosis airway epithelial cells. Nature *347*, 358–363.

Riordan, J.R., Rommens, J.M., Kerem, B., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., Chou, J.L., et al. (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science *245*, 1066–1073.

Rippe, K. (2001). Making contacts on a nucleic acid polymer. Trends Biochem. Sci. *26*, 733–740.

Rogers, C.S., Stoltz, D.A., Meyerholz, D.K., Ostedgaard, L.S., Rokhlina, T., Taft, P.J., Rogan, M.P., Pezzulo, A.A., Karp, P.H., Itani, O.A., et al. (2008). Disruption of the CFTR Gene Produces a Model of Cystic Fibrosis in Newborn Pigs. Science *321*, 1837–1841.

Romey, M.-C., Pallares-Ruiz, N., Mange, A., Mettling, C., Peytavi, R., Demaille, J., and Claustres, M. (2000). A Naturally Occurring Sequence Variation That Creates a YY1 Element Is Associated with Increased Cystic Fibrosis Transmembrane Conductance Regulator Gene Expression. J. Biol. Chem. *275*, 3561–3567.

Rommens, J.M., Iannuzzi, M.C., Kerem, B., Drumm, M.L., Melmer, G., Dean, M., Rozmahel, R., Cole, J.L., Kennedy, D., Hidaka, N., et al. (1989). Identification of the cystic fibrosis gene: chromosome walking and jumping. Science *245*, 1059–1065.

Rowe, S.M., Miller, S., and Sorscher, E.J. (2005). Cystic Fibrosis. N. Engl. J. Med. *352*, 1992–2001.

Rowntree, R.K., Vassaux, G., McDowell, T.L., Howe, S., McGuigan, A., Phylactides, M., Huxley, C., and Harris, A. (2001). An element in intron 1 of the CFTR gene augments intestinal expression in vivo. Hum. Mol. Genet. *10*, 1455–1464.

Sabo, P.J., Hawrylycz, M., Wallace, J.C., Humbert, R., Yu, M., Shafer, A., Kawamoto, J., Hall, R., Mack, J., Dorschner, M.O., et al. (2004). Discovery of functional noncoding elements by digital analysis of chromatin structure. Proc. Natl. Acad. Sci. U. S. A. *101*, 16837–16842.

Sabo, P.J., Kuehn, M.S., Thurman, R., Johnson, B.E., Johnson, E.M., Cao, H., Yu, M., Rosenzweig, E., Goldy, J., Haydock, A., et al. (2006). Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. Nat. Methods *3*, 511–518.

Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M., and Shiroishi, T. (2005). Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. Development *132*, 797–803.

Saiman, L. (2004). Microbiology of early CF lung disease. Paediatr. Respir. Rev. *5, Supplement 1*, S367–S369.

Sanyal, A., Baù, D., Martí-Renom, M.A., and Dekker, J. (2011). Chromatin globules: a common motif of higher order chromosome structure? Curr. Opin. Cell Biol. *23*, 325–331.

Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109–113.

Schwank, G., Koo, B.-K., Sasselli, V., Dekkers, J.F., Heo, I., Demircan, T., Sasaki, N., Boymans, S., Cuppen, E., van der Ent, C.K., et al. (2013). Functional Repair of CFTR by CRISPR/Cas9 in Intestinal Stem Cell Organoids of Cystic Fibrosis Patients. Cell Stem Cell *13*, 653–658.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. Cell *148*, 458–472.

Sharma, H., Mavuduru, R.S., Singh, S.K., and Prasad, R. (2014). Increased frequency of CFTR gene mutations identified in Indian infertile men with non-CBAVD obstructive azoospermia and spermatogenic failure. Gene *548*, 43–47.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). Nat. Genet. *38*, 1348–1354.

Simonis, M., Kooren, J., and de Laat, W. (2007). An evaluation of 3C-based methods to capture DNA interactions. Nat. Methods *4*, 895–901.

Smith, A.N., Wardle, C.J.C., and Harris, A. (1995). Characterization of DNAse I Hypersensitive Sites in the 120-kb 5′ to the CFTR Gene. Biochem. Biophys. Res. Commun. *211*, 274–281.

Smith, A.N., Barth, M.L., McDowell, T.L., Moulin, D.S., Nuthall, H.N., Hollingsworth, M.A., and Harris, A. (1996). A Regulatory Element in Intron 1 of the Cystic Fibrosis Transmembrane Conductance Regulator Gene. J. Biol. Chem. *271*, 9947–9954.

Smith, D.J., Nuthall, H.N., Majetti, M.E., and Harris, A. (2000). Multiple Potential Intragenic Regulatory Elements in the CFTR Gene. Genomics *64*, 90–96.

Snouwaert, J.N., Brigman, K.K., Latour, A.M., Malouf, N.N., Boucher, R.C., Smithies, O., and Koller, B.H. (1992). An Animal Model for Cystic Fibrosis Made by Gene Targeting. Science *257*, 1083–1088.

Sood, R., Bear, C., Auerbach, W., Reyes, E., Jensen, T., Kartner, N., Riordan, J.R., and Buchwald, M. (1992). Regulation of CFTR expression and function during differentiation of intestinal epithelial cells. EMBO J. *11*, 2487–2494.

Speicher, M.R., Ballard, S.G., and Ward, D.C. (1996). Karyotyping human chromosomes by combinatorial multi-fluor FISH. Nat. Genet. *12*, 368–375.

Spilianakis, C.G., Lalioti, M.D., Town, T., Lee, G.R., and Flavell, R.A. (2005). Interchromosomal associations between alternatively expressed loci. Nature *435*, 637–645.

Splinter, E., Grosveld, F., and Laat, W. de (2003). 3C Technology: Analyzing the Spatial Organization of Genomic Loci In Vivo. In Methods in Enzymology, C. David Allis and Carl Wu, ed. (Academic Press), pp. 493–507.

Splinter, E., Heath, H., Kooren, J., Palstra, R.-J., Klous, P., Grosveld, F., Galjart, N., and Laat, W. de (2006). CTCF mediates long-range chromatin looping and local histone modification in the β-globin locus. Genes Dev. *20*, 2349–2354.

Van Steensel, B., and Dekker, J. (2010). Genomics tools for unraveling chromosome architecture. Nat. Biotechnol. *28*, 1089–1095.

Stoltz, D.A., Meyerholz, D.K., Pezzulo, A.A., Ramachandran, S., Rogan, M.P., Davis, G.J., Hanfland, R.A., Wohlford-Lenane, C., Dohrn, C.L., Bartlett, J.A., et al. (2010). Cystic Fibrosis Pigs Develop Lung Disease and Exhibit Defective Bacterial Eradication at Birth. Sci. Transl. Med. *2*, 29ra31–29ra31.

Stoltz, D.A., Rokhlina, T., Ernst, S.E., Pezzulo, A.A., Ostedgaard, L.S., Karp, P.H., Samuel, M.S., Reznikov, L.R., Rector, M.V., Gansemer, N.D., et al. (2013). Intestinal CFTR expression alleviates meconium ileus in cystic fibrosis pigs. J. Clin. Invest.

Symmons, O., Uslu, V.V., Tsujimura, T., Ruf, S., Nassari, S., Schwarzer, W., Ettwiller, L., and Spitz, F. (2014). Functional and topological characteristics of mammalian regulatory domains. Genome Res. *24*, 390–400.

Tanabe, H., Müller, S., Neusser, M., Hase, J. von, Calcagno, E., Cremer, M., Solovei, I., Cremer, C., and Cremer, T. (2002). Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates. Proc. Natl. Acad. Sci. *99*, 4424–4429.

Tolhuis, B., Palstra, R.-J., Splinter, E., Grosveld, F., and de Laat, W. (2002). Looping and Interaction between Hypersensitive Sites in the Active β-globin Locus. Mol. Cell *10*, 1453–1465.

Trezise, A.E., and Buchwald, M. (1991). In vivo cell-specific expression of the cystic fibrosis transmembrane conductance regulator. Nature *353*, 434–437.

Tsui, L.C., Buchwald, M., Barker, D., Braman, J.C., Knowlton, R., Schumm, J.W., Eiberg, H., Mohr, J., Kennedy, D., Plavsic, N., et al. (1985). Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. Science *230*, 1054–1057.

Vakoc, C.R., Letting, D.L., Gheldof, N., Sawado, T., Bender, M.A., Groudine, M., Weiss, M.J., Dekker, J., and Blobel, G.A. (2005). Proximity among Distant Regulatory Elements at the β-Globin Locus Requires GATA-1 and FOG-1. Mol. Cell *17*, 453–462.

Vernimmen, D., Gobbi, M.D., Sloane-Stanley, J.A., Wood, W.G., and Higgs, D.R. (2007). Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. EMBO J. *26*, 2041–2051.

Vinay Kumar, Abul K. Abbas, and Jon C. Aster (2012). Robbins Basic Pathology (Elsevier Health Sciences).

Visel, A., Bristow, J., and Pennacchio, L.A. (2007). Enhancer identification through comparative genomics. Semin. Cell Dev. Biol. *18*, 140–152.

Visser, M., Kayser, M., and Palstra, R.-J. (2012). HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. Genome Res. *22*, 446–455.

Wansink, D.G., Schul, W., Kraan, I. van der, Steensel, B. van, Driel, R. van, and Jong, L. de (1993). Fluorescent labeling of nascent RNA reveals transcription by RNA polymerase II in domains scattered throughout the nucleus. J. Cell Biol. *122*, 283–293.

West, A.G., and Fraser, P. (2005). Remote control of gene transcription. Hum. Mol. Genet. *14*, R101–R111.

Wright, J.B., Brown, S.J., and Cole, M.D. (2010). Upregulation of c-MYC in cis through a Large Chromatin Loop Linked to a Cancer Risk-Associated Single-Nucleotide Polymorphism in Colorectal Cancer Cells. Mol. Cell. Biol. *30*, 1411–1420.

Xu, N., Donohoe, M.E., Silva, S.S., and Lee, J.T. (2007). Evidence that homologous X-chromosome pairing requires transcription and Ctcf protein. Nat. Genet. *39*, 1390–1396.

Yigit, E., Bischof, J.M., Zhang, Z., Ott, C.J., Kerschner, J.L., Leir, S.-H., Buitrago-Delgado, E., Zhang, Q., Wang, J.-P.Z., Widom, J., et al. (2013). Nucleosome mapping across the CFTR locus identifies novel regulatory factors. Nucleic Acids Res. *41*, 2857–2868.

Yoshimura, K., Nakamura, H., Trapnell, B.C., Dalemans, W., Pavirani, A., Lecocq, J.P., and Crystal, R.G. (1991). The cystic fibrosis gene has a “housekeeping”-type promoter and is expressed at low levels in cells of epithelial origin. J. Biol. Chem. *266*, 9140–9144.

Yu, J., Chen, Z., Ni, Y., and Li, Z. (2012). CFTR mutations in men with congenital bilateral absence of the vas deferens (CBAVD): a systemic review and meta-analysis. Hum. Reprod. *27*, 25–35.

Yun, W.J., Kim, Y.W., Kang, Y., Lee, J., Dean, A., and Kim, A. (2014). The hematopoietic regulator TAL1 is required for chromatin looping between the β-globin LCR and human γ-globin genes to activate transcription. Nucleic Acids Res. gku072.

Zhang, Y., McCord, R.P., Ho, Y.-J., Lajoie, B.R., Hildebrand, D.G., Simon, A.C., Becker, M.S., Alt, F.W., and Dekker, J. (2012). Spatial Organization of the Mouse Genome and Its Role in Recurrent Chromosomal Translocations. Cell *148*, 908–921.

Zhang, Z., Leir, S.-H., and Harris, A. (2013). Immune mediators regulate CFTR expression through a bifunctional airway-selective enhancer. Mol. Cell. Biol. *33*, 2843–2853.

Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Singh Sandhu, K., Singh, U., et al. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat. Genet. *38*, 1341–1347.

Zhou, X., Baron, R.M., Hardin, M., Cho, M.H., Zielinski, J., Hawrylkiewicz, I., Sliwinski, P., Hersh, C.P., Mancini, J.D., Lu, K., et al. (2012). Identification of a chronic obstructive pulmonary disease genetic determinant that regulates HHIP. Hum. Mol. Genet. *21*, 1325–1335.

Zink, D., Amaral, M.D., Englmann, A., Lang, S., Clarke, L.A., Rudolph, C., Alt, F., Luther, K., Braz, C., Sadoni, N., et al. (2004). Transcription-dependent spatial arrangements of CFTR and adjacent genes in human cell nuclei. J. Cell Biol. *166*, 815–825.

Zuin, J., Dixon, J.R., Reijden, M.I.J.A. van der, Ye, Z., Kolovos, P., Brouwer, R.W.W., Corput, M.P.C. van de, Werken, H.J.G. van de, Knoch, T.A., IJcken, W.F.J. van, et al. (2014). Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. Proc. Natl. Acad. Sci. *111*, 996–1001.