Role of CTCF Poly(ADP-ribosyl)ation in the regulation of cellular functions

Ioanna Pavlaki

A thesis submitted for the degree of Doctor of Philosophy

Department of Biological Sciences

University of Essex

Acknowledgements

Firstly, I would like to express my deepest gratitude to my supervisor Prof Elena Klenova for all her help and sincere support throughout this challenging experience. This endeavour would not have been possible without her constant and valuable guidance.

I would like to thank Dr Metodi Metodiev, my PhD advisor, for his contribution and feedback during supervisory boards and also to place on record the invaluable contribution made by Dr Igor Chernukhin to the bioinformatics analysis of this project.

I take this opportunity to express my sincere thanks to Dr Dawn Farrar for generously sharing her expertise and for providing endless moral support during the challenging times of this project. I would also like to thank Dr France Docquier for all her support and guidance.

I would like to express my deepest appreciation to Prof Alexander Buerkle and the members of his group at the University of Konstanz for welcoming me in their laboratory and sharing their time and knowledge with me.

I would also like to extend a special thanks to Mrs Adele Angel for her precious support and it is a pleasure to thank the current and former members of the EK Molecular Oncology laboratory: Dr Svetlana Gretton, Dr Georgia-Xanthi Kita, Dr Hulkar Mamayusupova, Dr Jay Mani, Krista McHugh, Dr Olayinka Oloko and soon-to-be-doctors Rosie Bryan and Myla Pavlova for all the unforgettable times we shared in the lab.

I cannot forget my friends, whose presence has been motivating and distracting at all the right times and I consider myself lucky to have them. Special thanks from the heart to Onur, Rama, Chrissy, Giota and the flying Lori, Julio, Liven, Iro and Emmh.

Finally, I would like to deeply thank my family in Greece for their unconditional love and support throughout my life but also during the good and the challenging times of this particular journey. I would not have been able to embark on it and complete it without them encouraging me during the difficult times and celebrating with me each accomplishment.

ABSTRACT

CTCF is an evolutionary conserved and ubiquitously expressed protein which regulates a plethora of cellular functions using different molecular mechanisms. However, the role of poly(ADP-ribosyl)ation (PARylation) of CTCF, in particular, two differentially PARylated forms of CTCF (termed CTCF130 and CTCF180) in the regulation of cellular processes is not well understood. We hypothesize that differentially PARylated isoforms of CTCF control different groups of genes (Hypothesis 1) and different functions (Hypothesis 2) in different biological situations. To investigate Hypothesis 1, the 226LDM cells, proliferating (expressing CTCF130 and CTCF180) and arrested (expressing CTCF180) will be used. The high-throughput ChIP-Seq and RNA-Seq assays will be used to obtain and link the CTCF binding and gene expression profiles. This study revealed that out of 2051 CTCF binding sites in proliferating 226LDM cells identified using the polyclonal CTCF antibody (which recognizes both isoforms, CTCF130 and CTCF180), 1009 were associated with differentially expressed transcripts. Among those, 520 were up-regulated while the remaining 489 were down-regulated. From 64 binding sites identified in the arrested cells (CTCF180 only), 26 were associated with genes that were differentially expressed; 16 were up-regulated and 10 were down-regulated. There were 8 common CTCF binding sites between control and treated cells in the up-regulated and 6 in the down-regulated group. Overall, binding of CTCF130 and CTCF180 was associated with differential gene expression thus confirming Hypothesis 1. To test Hypothesis 2, the role of CTCF PARylation in the DNA damage response mechanism was investigated. Following DNA damage, translocation of CTCF into the nucleolus and co-localization with PARP1 was observed. Ectopic expression of a mutant CTCF deficient for PARylation resulted in nucleolar disorganisation and was associated with faster recovery from damage. It also suggests that CTCF PARylation is implicated in DNA damage response as well as for nucleolar stability after DNA damage thus supporting Hypothesis 2.

Abbreviations

ADPr	ADP-ribose
APP-β	Amyloid β Protein Precursor
BP	Base pairs
BPH-1	Benign Prostatic Hyperplasia
ChIP	Chromatin Immunoprecipitation
CTCF	CCCTC-Binding Factor
DBD	DNA Binding Domain
DDR	DNA Damage Response
DE	Differential Expression
DMEM	Dulbecco's Modified Eagles Medium
DSB	DNA Double Strand Breaks
EtBr	Ethidium Bromide
FADU	Fluorimetric Analysis of DNA Unwinding
FBS	Foetal Bovine Serum
FDR	False Discovery Rate
FPKM	Fragments per Kilobase of Exon per Million Fragments Mapped
GO	Gene Ontology
HOXA	Homeobox Gene A Locus
HPC	High Performance Computer
HRP	Horse Radish Peroxidase
HU	Hydroxyurea
IF	Immunofluorescence

IP	Immunoprecipitation
LB	Luria Broth
MDS	Multi-Dimensional Scaling
MEME	Multiple EM for Motif Elicitation
mRNA	Messenger RNA
NAD^+	Nicotinamide Adenine Dinucleotide
ncRNA	Non-Coding RNA
NO	Nocodazole
NGS	Next Generation Sequencing
NuLS	Nucleolar Localization Signal
pADPr	Poly(ADP-ribose)
PARG	Poly(ADP-ribose) glycohydrolase
PARP	Poly(ADP-ribose) polymerase
PARylation	Poly(ADP-ribosyl)ation
PCDH	Protocadherin
PFA	Paraformaldehyde
rDNA	Ribosomal DNA
RPKM	Reads per Kilobase per Million Mapped Reads
rRNA	Ribosomal RNA
SDS	Sodium Dodecyl Sulphate
SNP	Single Nucleotide Polymorphism
SSB	Single Strand Breaks
SUMO	Small Ubiquitin-Like Modifier

TF	Transcription Factors
TRE	Thyroid-Hormone Response Element
tRNA	Transfer RNA
TSS	Transcription Start Site
ZF	Zinc Fingers

Table of contents

List of figures

List of tables

Chapter 1 Introduction

1.1 CTCF: The CCCTC-binding factor

The CCCTC-binding factor (CTCF) is a multifunctional and ubiquitously expressed protein, which remained highly evolutionarily conserved from *Drosophila* to man (Ohlsson et al., 2001, Phillips and Corces, 2009).

It was originally identified by Lobanenkov et al. (1990) as a DNA binding protein, regulating the expression of the chicken *c-myc* oncogene. The protein had a binding site upstream of the start of the transcription site of *c-myc* to one of the CCCTC core sequences, which gave the protein its name. Since its discovery, studies on CTCF revealed that this protein is not only involved in transcriptional control, but that it is a key regulator of many other cellular functions (Holwerda and de Laat, 2013, Ohlsson et al., 2001, Ohlsson et al., 2010, Millau and Gaudreau, 2011).

1.2 CTCF, the gene and the protein

1.2.1 The *CTCF* gene

The gene encoding CTCF is positioned, in humans, on chromosome 16, on locus 16q22.1 (Filippova et al., 1998). This genomic locus has been indicated as a hot-spot region for loss of chromosomal material linked with cancer, including breast and prostate cancers (Filippova et al., 1998, Green et al., 2009, Yeh et al., 2002). The intron-exon organization of the chicken *CTCF* gene was determined by Klenova et al. (1998) while that of the human gene was reported later (Filippova et al., 2002). These studies showed that there are seven exons in the chicken gene and ten in human. The CTCF gene structures were refined in a more recent publication reporting 10 exons in human, chicken, mouse, zebrafish and frog *CTCF*, seven of them encode the zinc fingers (ZF) which constitute the proteins DNA binding domain (Hore et al., 2008).

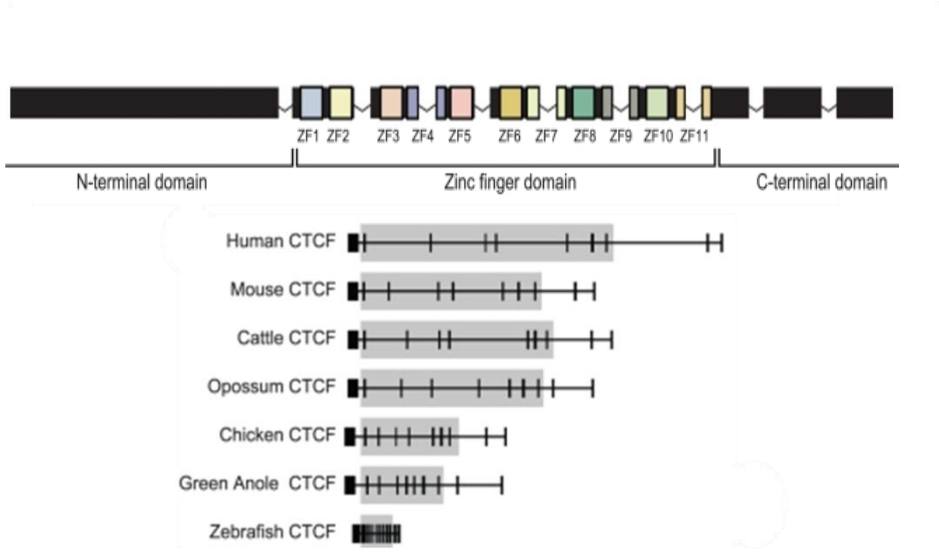


Figure 1-1 Intron-exon structure of the human *CTCF* gene and comparison with other species

Top: The structure of the *CTCF* gene. Bottom: CTCF orthologues from human, mouse, cattle, opossum, chicken, green anole and zebrafish. All vertebrate *CTCF* orthologues possess ten exons (adapted from Hore et al. (2008)).

1.2.2 The CTCF protein

Human CTCF is composed of 727 amino acids and structurally highly conserved among species (Filippova et al., 1996). The protein can be divided into three main domains; the DNA binding domain (DBD) which is comprised of 11 zinc fingers (ZF) (311 amino acids) and the C- and N-termini regions flanking the DBD (150 and 250 amino acids respectively) (figure 1-2) (Filippova et al., 1996).

The zinc finger is a versatile supersecondary protein structure, very commonly found in eukaryotes mediating the interaction between the protein and a DNA or RNA sequence (Matthews and Sunde, 2005). There are different classes of zinc fingers that have different roles and can co-exist within a protein. CTCF is such an example protein with 11 zinc fingers out of which, 10 belong to the most studied C2H2 class and one in the C2HC class (Klenova *et al*, 1993). The CTCF zinc fingers can be employed in different combinations in order for the protein to interact with a wide variety of DNA sequences and protein partners; each of the ZFs can be important for one interaction and completely unnecessary for another. This feature is indicative of the proteins high versatility and it is why it is known as a "multivalent factor" (Filippova *et al*, 1996).

The main ZF domain is flanked by the N- and C-terminus regions, both of which do not appear to be implicated in the binding process, however they reportedly can affect the results of the interaction (Filippova *et al*, 1996).

There are three additional, less significant although not without function, binding motifs within the C-terminal region of CTCF. The first is a KRRGRP-type AT-hook that possibly has a role in DNA binding and protein-protein interactions in chromatin (Ohlsson et al., 2001, Aravind and Landsman, 1998). The second one is a conserved SKKEDSSDSE motif. The third is located between the other two and it is an HS3-domain (Ohlsson et al., 2001).

The molecular weight of CTCF was determined as 82 kDa (Klenova et al., 1993), however it migrates at 130 kDa on SDS-PAGE. This abnormality was studied by Klenova et al.

(1997) who concluded that this was the results of aberrant mobility on the SDS-PAGE due to properties of the C- and N-termini regions.

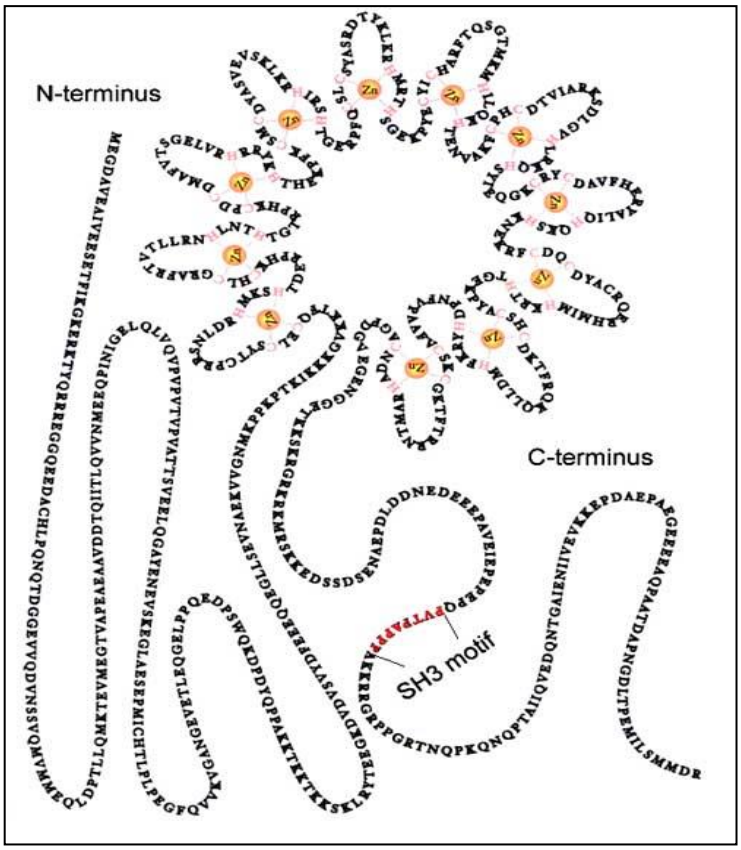


Figure 1-2 Schematic representation of the CTCF protein structure

The main zinc finger domain is flanked by a N-terminus and a C-terminus domain. The zinc fingers are employed in different combinations in order to interact with DNA or a target protein to take place. Apart from the ZF domain another binding domain located in the C-terminus region is the SH3 motif. (Adapted from Ohlsson et al. (2001)).

1.3 **Biological Functions of CTCF**

Since the discovery of CTCF as a transcriptional repressor, a great variety of functions has been attributed to the protein including a role in cell cycle progression and proliferation, genomic imprinting, insulation function and chromatin architecture (Holwerda and de Laat, 2013, Ong and Corces, 2014, Ohlsson et al., 2001).

The list of very important processes in which the protein is involved in is mainly a result of its capacity to engage in multiple protein-protein and DNA-protein interactions and explains why the presence of CTCF is crucial for the viability of the eukaryotic cell (Klenova *et al,* 2002). The most important functions of CTCF are discussed in this section.

1.3.1 Transcription Regulation

Transcription is a fundamental cellular process, involved in the main pathways of cell maintenance, proliferation and death. It is the second step, after DNA replication, in the pathway to protein synthesis. Transcription factors (TF) assume a regulatory role in the cell and their binding to key DNA positions or, equally important the lack of it, decides the expression status of the nearby gene. The balance between over and under-expression of a gene is very delicate and requires a tight and fail-proof control system, otherwise both extremes can result in disease (Libermann and Zerbini, 2006, Spitz and Furlong, 2012).

CTCF was originally identified as a transcriptional repressor of the *c-myc* oncogene in chicken and later it was reported to have the same function for the *c-myc* oncogene in mice as well as humans (Filippova et al., 1998, Ohlsson et al., 2001). Other genes that are negatively regulated by CTCF include *hTERT* (Renaud et al., 2005, Renaud et al., 2007, Choi et al., 2010) and the lysozyme gene, in conjunction with the thyroid-hormone response element (TRE) (Awad et al., 1999). Evidence accumulated highlighting its function as a transcription activator as well. CTCF can activate several genes including the amyloid β protein precursor (*APP-β),* human *p14ARF* and human *PIM-1* (Filippova, 2008, Vostrov and Quitschke, 1997).

CTCF transcription regulation studies showed that the flanking regions of the main binding domain may also be indirectly involved in the transcription factor activity. Specifically, repressor activity is associated with the N-terminal, the C-terminal and the ZF regions, while activation activity is localized in the N-terminal region (Klenova et al., 2002).

1.3.2 Insulator Function

Insulation is a transcription regulation-related feature of chromatin. Insulator elements are either "enhancer blockers", which prevent the successful communication between the enhancer and promoter region, or "chromatin barriers" which prevent the unwanted spread of heterochromatin and facilitate the euchromatic state of chromatin, which is more approachable for transcription DNA. The exact mechanism that underlies the insulation process remains unclear however chromatin loop formation and long-range chromatin interactions frequently appear to be involved in the process. Insulator proteins can mediate chromatin architecture modulation and three-dimensional looping in order to exert their function (Barkess and West, 2012, Van Bortle et al., 2014, Schwartz et al., 2012, Kyrchanova et al., 2013, Brasset and Vaury, 2005).

CTCF is the only known mammalian protein to be involved in insulation and has been implicated in both enhancer blocking and chromatin barrier function (Kim et al., 2007, Xie et al., 2007, Herold et al., 2012). Bell et al. (1999) first implicated CTCF in insulation function by discovering that it binds to the 42 bp insulation element of the chicken β-globin, known to be both necessary and sufficient for enhancer blocking activity in human cells (Ohlsson et al., 2010). Subsequently, more studies emerged highlighting the insulator function of CTCF (Weth et al., 2010, Kim et al., 2011, Zielke et al., 2012, Eggeling et al., 2014, Splinter et al., 2006).

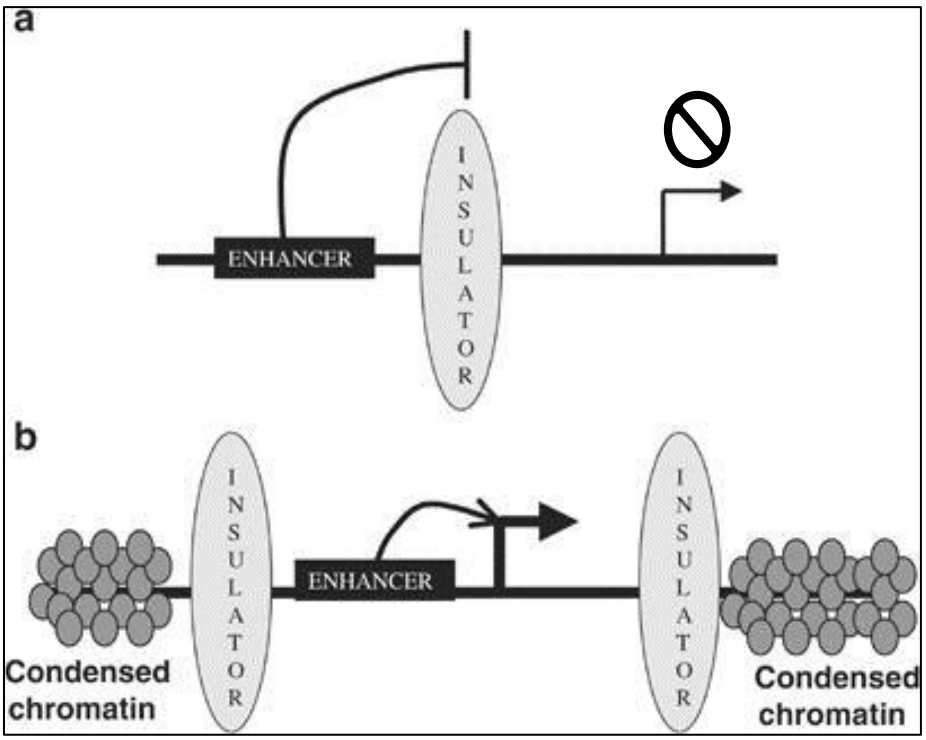


Figure 1-3 Schematic representation of protein insulation function

The insulator can block the communication between an enhancer and a promoter thus preventing gene expression. This type of insulation is called “enhancer blocker” (a). The second type can promote the expression of a gene by preventing the spread of heterochromatin in the gene regional areas (b) (Adapted from Brasset and Vaury (2005))

1.3.3 Genomic Imprinting

The expression of a gene can be restricted to only one of the two parental alleles, commonly through methylation or other modifications that lead to the silencing of the other. This conserved epigenetic regulation method is employed mostly, but not exclusively, in the case of genes that are important during development. To distinguish between the alleles and decide which one will be expressed, an epigenetic mark such as DNA methylation can be found at the Imprinting Control Region (ICR) (Peters, 2014, Barlow and Bartolomei, 2014)

CTCF is involved in this epigenetic inheritance phenomenon and the modulation of the *Igf2*/*H19* genes represents a classic model of genomic imprinting (Lewis and Murrell, 2004, Ideraabdullah et al., 2014, Singh et al., 2012). The methylation state of the ICR of the two genes differs between the two parental alleles and since CTCF binding is sensitive to methylation, it defines whether CTCF will exert its activity as an enhancer blocker or not (Bell and Felsenfeld, 2000, Kanduri et al., 2000, Schoenherr et al., 2003). When the paternal ICR is methylated, CTCF cannot bind to it. In this case the enhancer is free to activate *Igf2*, while in the opposite case, when the ICR is un-methylated, CTCF can bind to it, prevent the *Igf2* enhancer from acting and thus resulting in the expression of *H19* only (figure 1-4) (Reik and Murrell, 2000, Szabo et al., 2004, Yang et al., 2003).

Another imprinting regulation phenomenon that CTCF is involved in is the X chromosome inactivation. In females, the expression of genes in one of the two X chromosomes is under epigenetic silencing control; the choice mechanism is based on the non-coding *Xist* (inactive x-specific transcript) and the antisense *Tsix* (McCarrey et al., 2002). CTCF was implicated with the X chromosome silencing when it was discovered that the promoter of *Xist* has a CTCF binding site and additionally that the binding site of CTCF in the *Tsix* promoter is under methylation control (Pugacheva et al., 2005, Chao et al., 2002, Boumil et al., 2006). More recent studies showed that between *Xist* and *Tsix*, at the X inactivation center (Xic), another CTCF binding site exists (RS14) and that blocking of this site caused aberrant regulation of X chromosome expression (Spencer et al., 2011).

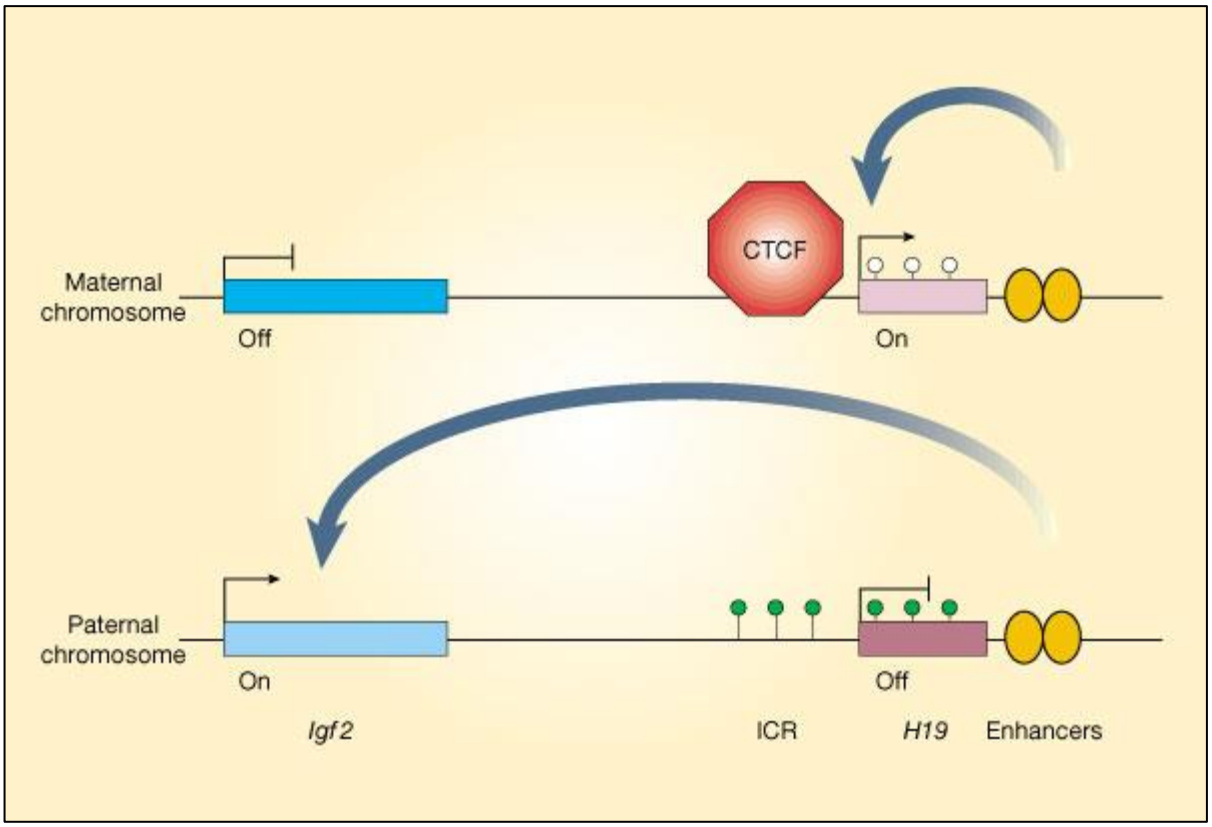


Figure 1-4 Genomic imprinting of the *Igf2/ H19* locus

The expression of *Igf2* and *H19* is under genomic imprinting control exerted by the CTCF insulator function. The methylation status of the imprinting control region (ICR) of *H19* decides whether CTCF can bind to it or not. In the maternal allele, the ICR is not methylated; CTCF can bind to it and disallow the communication between the promoter of *Igf2* and its enhancer. In this case, the *H19* gene is expressed and the *Igf2* is not. On the other hand, on the paternal chromosome the *H19* ICR is methylated and CTCF does not bind to it. In this case *Igf2* is expressed and *H19* is not (adapted from Reik and Murrell (2000)).

1.3.4 Chromatin architecture

The development of new techniques which allow high-throughput analysis of the organization of chromosomes has offered new insight on CTCF function. Evidence from independent studies pinpoints CTCF as an architecture protein able to mediate inter- and intra-chromosomal interactions between distant sites.

MacPherson and Sadowski (2010) obtained biochemical evidence that CTCF induces the formation of "an unusual DNA structure". They concluded that CTCF acts as a looping protein and the zinc finger domain of the protein is sufficient for the looping function to be exerted. Later, Kim et al. (2011) studied the Homeobox gene A locus (*HOXA)* and discovered that CTCF and cohesin act in collaboration to link higher-order chromatin state regulation with transcription function. In particular, they found that CTCF keeps heterochromatin off the *HOXA* locus and acts as a docking system for the cohesin complex.

Sanyal et al. (2012) used the chromosome conformation capture carbon copy (5C) technique to show that 79% of long range interactions taking place in the cell need at least one CTCF site. This high percentage accentuates the importance of the involvement of CTCF in chromatin structure modulation. CTCF can reportedly use the chromatin looping function to tether distinct enhancers to their promoters, regulate alternative mRNA splicing and transcriptional pausing via RNA polymerase II and to regulate the expression of complex gene clusters (Ong and Corces, 2014). In the mammalian brain, the diversity between neurons is achieved through alternatively splicing of the protocadherin A *(PCDH)* gene. The gene cluster is composed of over 50 exons, all of which have different promoters. CTCF and cohesin bind to most of these as well as the distant enhancer HS5-1. The regulation of the isoform expression relies on CTCF binding and facilitating chromatin looping (figure 1-5).

1.4

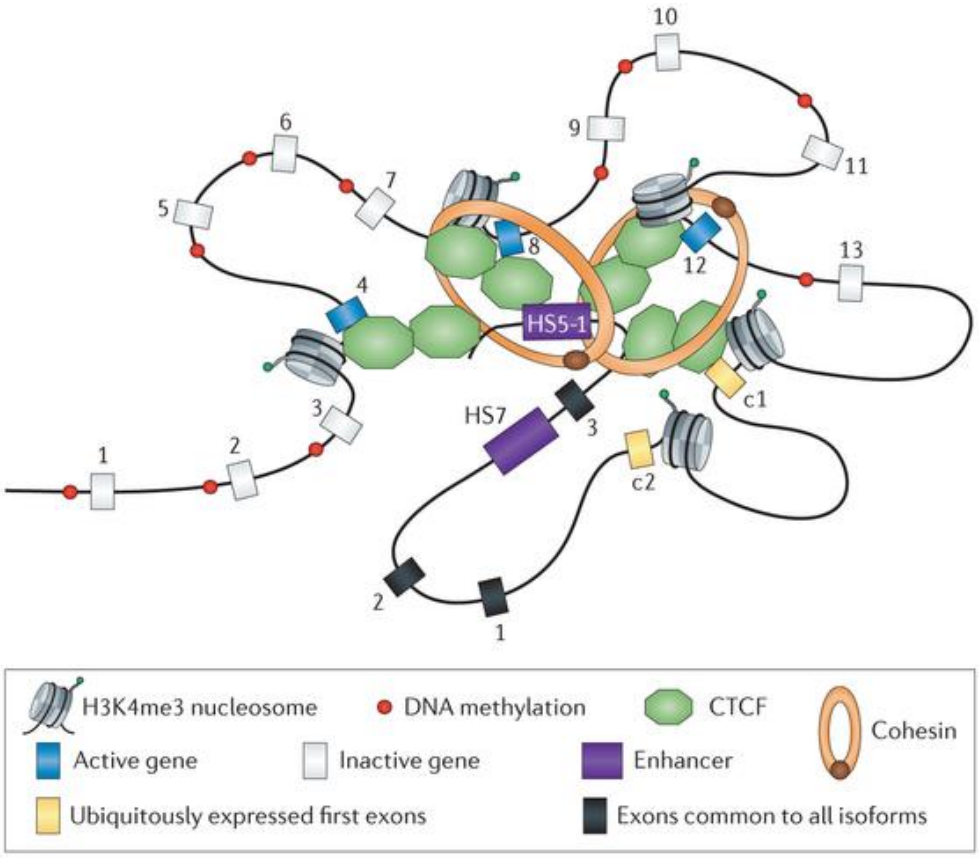


Figure 1-5 CTCF-mediated chromatin looping is necessary for the transcription regulation of the PCDHA gene cluster

The human protocadherin A (*PCDHA*) gene cluster contains 15 exons. Each of these 15 variable exons has its own promoter. Promoter choice and the formation of an active chromatin hub is mediated by CTCF-cohesin DNA looping between the distal HS5-1 enhancer and distinct promoters at the PCDHA gene cluster (adapted from Ong and Corces (2014)).

1.4 CTCF binding sites

Extensive research has been aimed at analyzing CTCF binding sites, however the combinatorial use of the CTCF zinc fingers allow it to bind to a wide variety of sequences hindering the search for a consensus motif for CTCF binding (Filippova et al., 1996).

Despite the complexity of CTCF binding, a 20-mer motif has been identified as highly conserved in vertebrates (Kim et al., 2007) The wide variety of CTCF binding sites includes some which are extensively evolutionarily conserved among different species (Ni et al., 2012), while others are unique, not only between species but also between different cell types of the same species (Chen et al., 2012).

Computational methods had revealed more than 13,000 potential binding sites for CTCF within the human genome (Kim *et al* 2007) and with the employment of more recently developed next generation sequencing (NGS) techniques the number has risen to over 30,000 (Chen et al., 2012). Overall more than 15 million binding sites have been discovered in the genomes of 15 species including human. The majority of CTCF binding sites are within 1000 bp of the transcription start site (TSS) of genes although binding in areas far from any gene has also been reported. Although their function is mainly unknown, they could be utilized during chromatin looping or other CTCF functions (Chen et al., 2012).

1.4.1 Next generation sequencing for the discovery of CTCF binding sites

Transcription factors like CTCF exert their regulatory function mainly through binding to DNA. The value of identifying their genome-wide binding sites has become clear especially over the last decade and many techniques have been developed aiming at the discovery and functional analysis of the binding sites and motifs.

Genome-wide high-throughput techniques like chromatin immunoprecipitation sequencing (ChIP-seq) is used to discover the binding profile of transcription factors and histones (Ouyang et al., 2009). Integration of ChIP-seq with RNA-sequencing, which reveals the

cellular expression profile, can assist with the elucidation of gene regulatory mechanisms and the interplay between genetic and epigenetic regulation (Angelini and Costa, 2014, Gomez-Cabrero et al., 2014)

1.5 CTCF involvement in cellular processes

CTCF is involved in a wide variety of cellular processes by exerting different functions. This is achieved through engaging in protein-DNA or protein-protein interactions (Holwerda and de Laat, 2013). A selection of these functions is discussed in the following subsections.

1.5.1 Cell Cycle

Previous studies have concluded that abnormalities in the expression levels of CTCF have severe effects on the viability of the cell (Heath et al., 2008). Tissue specific inhibition of CTCF expression has a critical result in cell cycle progression, while CTCF knockout mice do not survive beyond early embryonic stages (Filippova et al., 2002). On the other hand, ectopic expression of CTCF causes growth retardation which has been attributed to CTCF's ability to hinder DNA replication (Rasko et al., 2001). These findings indicate that CTCF is necessary for cell cycle progression and survival, but also that its levels have to remain under strict control. The interaction of CTCF with proteins known to participate in the regulation of the cell cycle, such as p21 and p27 further support the involvement of CTCF in the modulation of the cell cycle (Heath et al., 2008, Phillips and Corces, 2009).

1.5.2 Apoptosis

Apoptosis is the term used by Gerschenson and Rotello (1992) to describe the programmed cell death and comes from the Greek word that means "falling off" (Duque-Parra, 2005). Studies on CTCF expression levels in breast cancer cell lines, as well as normal and cancer tissues, revealed that CTCF is up-regulated in cancer cells (Docquier et al., 2005). Surprisingly for a tumour suppressor candidate, this finding suggested that CTCF may have an

anti-apoptotic effect in breast cancer cells, which was confirmed by the detection of high levels of apoptotic markers following the removal of CTCF from those cells (Docquier et al., 2005).

1.5.3 Nucleolar transcription

Guerrero and Maggert (2011) discovered that in *Drosophila*, CTCF binds to the regulatory elements of ribosomal DNA (rDNA) and its absence leads to nucleolar fragmentation and decrease of rDNA silencing. In the repeat array of *Drosophila* rDNA approximately half of the genes are active and the other half are inactive. CTCF binding appears to be involved in the epigenetic regulation of these genes and modulate ribosomal transcription in the nucleolus. Another study using the human HeLa cell line showed that absence of CTCF from the nucleolus affects the organization of the nucleolus and results in up-regulated rDNA transcription (Hernandez-Hernandez et al., 2012). In addition to this, Torrano et al. (2006) discovered that the translocation of CTCF into the nucleolus is associated with growth arrest, apoptosis in MCF-7 cells and differentiation of the leukemia K562 myeloid cells. All in all, evidence implicating CTCF with the architecture and epigenetic regulation of the nucleolus is accumulating. However, the exact mechanisms underlying this implication are not yet fully understood.

1.6 CTCF and post-translational modifications

Many proteins undergo post-translational modifications which sometimes alter their biological functions. CTCF can be reportedly subjected to phosphorylation (Klenova et al., 2001), SUMOylation (MacPherson et al., 2009) or Poly(ADP-ribosyl)ation (PARylation) (Docquier et al., 2009).

1.6.1 Phosphorylation

Phosphorylation describes the addition of a phosphate group to a protein and can cause the activation or de-activation of the protein. Delgado et al. (1999) first studied CTCF phosphorylation during myeloid cell differentiation in leukaemia cell lines. From their experiments it was observed that cells induced to differentiate were more highly phosphorylated, while cells that are in the proliferation process are under-phosphorylated. Also, they discovered that phosphorylation is involved in the modulation of CTCF expression and cell differentiation pathways can dictate the levels of phosphorylated CTCF. Klenova et al. (2001) described that the majority of the phosphorylation sites of CTCF are located on the C-terminus region and to a lesser extent in the ZF domain.

1.6.2 SUMOylation

SUMOylation is a relatively newly discovered modification, which includes the covalent attachment of small ubiquitin-like modifier (SUMO) proteins in a substrate protein (Alontaga et al., 2012). The number of processes that it is implicated in is constantly increasing and includes protein localization (Wang et al., 2008, Saito et al., 2010), DNA damage (Cremona et al., 2012, Sarangi et al., 2015), senescence (Ivanschitz et al., 2013), and transcription (Sun et al., 2013). SUMOylation levels are increased as a result of stressful conditions like hypoxia and this increase in SUMOylated proteins is also sometimes followed by cell senescence (Ivanschitz et al., 2013).

MacPherson et al. (2009) showed that the modification does not interfere with the DNA binding capacity of CTCF while their findings also suggested the SUMOylation promotes the suppressive role of CTCF on the P2 promoter of *myc*. Recently, Wang et al. (2012b) discussed the de-SUMOylation of CTCF as a means to regulate its activity in hypoxic conditions, based on experiments conducted on human corneal epithelial cells

1.6.3 Poly(ADP-ribosyl)ation (PARylation)

PARylation is a phylogenetically ancient mechanism, important for the cellular stability mostly by playing an important role in the defense mechanism against threats to the integrity of the genome. These threats, such as endogenous or exogenous DNA damage, activate PARylation and are at the same time the target of its activity. PARylation is involved in many other cellular functions among which are DNA replication, gene expression and cellular differentiation (Golia et al., 2015, Robert et al., 2013, Burkle, 2000).

PARylation has been widely linked to CTCF. The insulator and transcription factor functions of CTCF are reportedly by PARylation (Yu et al., 2004a). Docquier et al. (2009) showed that PARylation of CTCF results in the formation of the 180kD CTCF isoform (CTCF180) which can be detected both in normal and in cancerous conditions. The CTCF130 on the other hand is hypo- or non-PARylated and appears in many immortalized cell lines and in cancer tissues, including breast tumours.

The PARylation of CTCF is of particular interest for the present study and this modification will be discussed in more detail in the following section.

1.7 PARylation and the PARP polymerases

1.7.1 PARylation

PARylation utilizes cellular nicotinamide adenine dinoucleotide (NAD^+) as a substrate with a reaction that involves the breakage of the glycosylic link between the nicotinamide and a ribose followed by the release of the first along with a proton, and the usage of the ribose as a substrate for the formation of ADPR monomers. This is followed by the polymerization of these monomers and the attachment of the resulting polymers to a glutamic acid residue of a target protein (**Figure 1-6**) (Kim et al., 2005a).

PARylation is catalyzed by enzymes called poly(ADP-ribose)polymerases (PARPs). The enzyme family contains more than 10 proteins which have common structure and/or functions (Nguewa et al., 2005) while the most significant and extensively studied one of the enzymes is PARP1 (Kim et al., 2005a, Kim, 2011). PARP family-characteristic functions include the addition of the ADPR moieties to target proteins, as well as the elongation of these by creating ribosyl-ribosyl bonds. This may result in linear and/or branched polymers of various lengths ranging from only a few to more than one hundred residues (Mendoza-Alvarez et al., 2000). The variety in the size of the polymers can be attributed to their rapid degradation by the PARG enzyme, as well as the nature of the proteins that they are attached to. Not all PARPs can catalyze the whole range of the PARP family-characteristic reactions; for example, the non-hydrolytic cleavage of the protein proximal ADPR moieties is one function restricted to the enzyme ADP-ribosyl protein lyase only (Buerkle, 2008, Oka et al., 1984).

The effect that PARylation exerts on its target proteins can be traced back to the function of the polymers. Structurally, they incorporate features found in polynucleotides and/ or polysaccharides while they are also significantly negatively charged. As a consequence, they affect the structure and the biochemical characteristics of the acceptor protein leading to important changes in its function (Buerkle, 2008, Heeres and Hergenrother, 2007).

Known PARP targets are involved in many cellular functions such as PCNA, topoisomerase I, p53, XRCC1, NF-kβ, CTCF, histone proteins and others (Buerkle, 2008, El-Khamisy et al., 2003, Wesierska-Gadek et al., 2005, Hegedus and Virag, 2014).The variety in the spectrum of targets highlights the importance of this modification for the cellular fate and health.

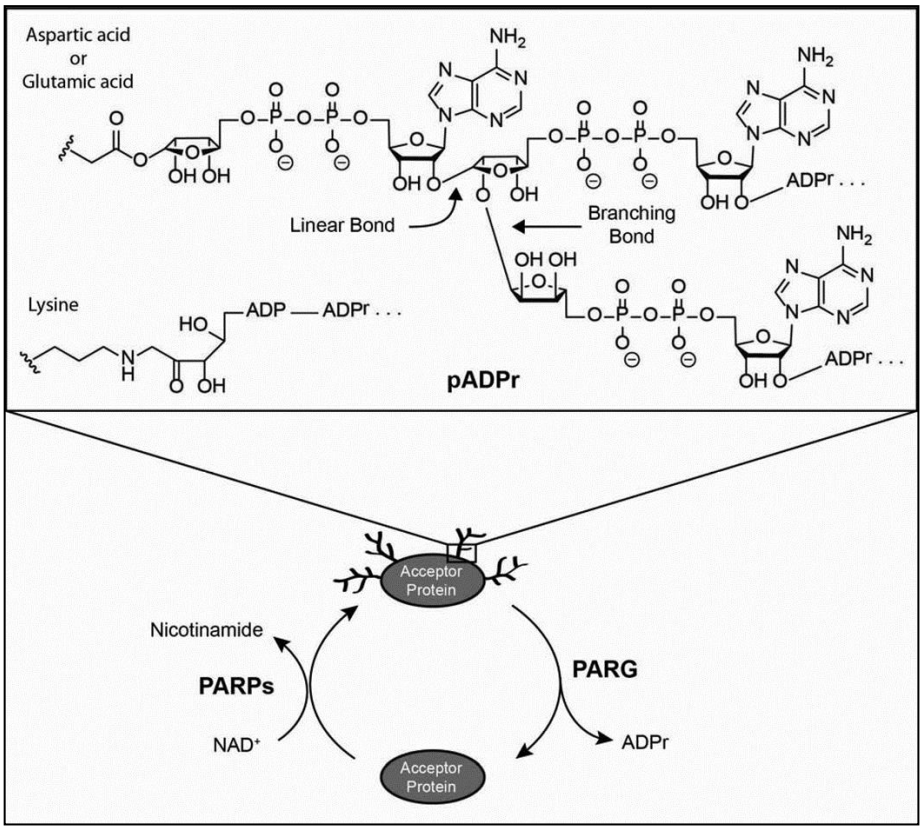


Figure 1-6 The metabolism of poly(ADP-ribose) (pADPr)

PARPs biosynthesize pADPr from NAD^+ while PARG degrades the polymer to ADP-ribose (ADPr). The monomer pADPr can be covalently attached to aspartic acid, glutamic acid, or lysine residues of acceptor proteins and the ADPr units of the polymer are connected in a linear or branched manner (Tan et al. (2012).

1.7.2 PARP polymerases

1.7.2.1 PARP1

PARP1 is the most extensively studied enzyme belonging to the PARP family. The gene coding for the protein is located on chromosome 1, locus 1q42 and the protein can be found mainly throughout the nucleus, as well as the centrosome (Megnin-Chanet et al., 2010). The structure of PARP1 includes three distinct domains that play different parts related to its function (figure 1-7). The first is a binding domain that infers the polymerase activity and which is located on the C-terminal region. It is a highly conserved helix-loop-helix motif, which is considered to be the signature domain of the protein and plays the major role in NAD-binding and synthesis of the PAR polymers (Megnin-Chanet et al., 2010, Kim et al., 2005a). The N-terminal DNA binding domain (DBD) contains two zinc fingers and is responsible and necessary for DNA binding and the recognition of DNA lesions, in a manner that is not dependant on specific DNA sequences (Ko and Ren, 2012). This interaction with DNA results in a conformational change in the molecule that increases its catalytic activity greatly. The zinc fingers are also involved in the interaction of PARP1 with protein partners and with other PARP1 molecules. Between these two domains a central 22kD domain can be found which contains the site which is responsible for auto-modification (Ko and Ren, 2012).

For a long time PARP1 was thought to be the only PARP polymerase, however the existence of more members such as PARP2, the tankyrases, ADP-ribosyl protein lyase is now well documented.

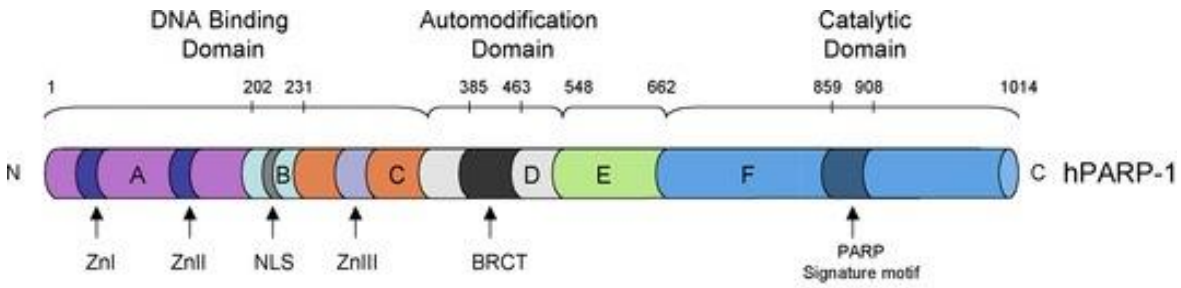


Figure 1-7 . The domain structure of human PARP1

The structure of human PARP1 contains the DNA-binding, the auto-modification and catalytic domains (domains A–F). The PARP signature sequence (in dark blue within the F catalytic domain) is the most conserved between the PARPs. (Zn I and Zn II: Zinc-finger motifs - NLS nuclear localization signal (adapted from Megnin-Chanet et al. (2010)).

1.7.2.2 **PARP2**

PARP2 is the second most studied PARP enzyme and, like PARP1, it is involved in the detection of DNA damage. Whereas PARP1 detects single, double strand breaks and various types of lesions, PARP2 is mostly responsible for recognizing gaps and flap structures (Beck et al., 2014). Until recently PARP1 and PARP2 were considered to be the only member of the PARP family to be involved in DNA damage repair. However, recent studies have also proved that PARP3 can detect double strand DNA breaks (Beck et al., 2014). PARP2 only represents 10% of the activated PARP activity after DNA damage (Buerkle, 2008). One can therefore infer that PARP1 acts as the main contributor in the detection and of the DNA strand breaks and the alarm mechanism that is activated against the instability they can cause to the cell.

PARP2 is a 66kD protein with the coding gene positioned on chromosome 14q11.2. The structure of the enzyme resembles that of PARP1 with a catalytic and a DNA binding domain found in the C- and N- terminal regions respectively. The auto-modification domain is not present and the protein is thought of as a "shorter" version of PARP1. These similarities in structure and function led scientists into believing that PARP2 acted as a "back up" enzyme for PARP1, however evidence today suggests that the enzymes have unique roles (Beck et al., 2014). For example, except for the centromeres, PARP2 can be also found at the telomeres which is feature not shared with PARP1 (Buerkle, 2008).

1.7.2.3 **Tankyrase -1**

The tankyrase-1 gene is located on the 8q23.1 chromosome and codes for a 142kD protein. Tankyrase is another member of the poly(ADP-ribosyl)ation polymerase family carrying the characteristic catalytic domain at the C terminal region. It differs from the previously described PARPs in that the polymers it creates can only be linear and more importantly, it does not contain the N terminal DNA binding domain. Lack of this domain means that the protein

does not have the capacity to detect DNA strand breaks and it is therefore not involved in the pathway which is activated in response to genotoxic stress. The enzyme can be found on the chromosomic telomere region and is involved in the maintenance of telomere length, mainly by interacting with the telomeric repeat binding factor (TRF-1) (Chang et al., 2005, Cook et al., 2002). TRF-1 is part of a nucleoprotein complex that protects telomeres. PARylation by tankyrase-1 exposes the telomere and telomerase can then add telomeric repeats (Lehtio et al., 2008).

1.7.3 PARP inhibitors

PARP inhibitors are used as chemosensitizers and cancer therapeutics, based on the inefficiency of tumour cells to repair DNA double strand breaks (DSB). The rationale behind this strategy is that induction of single strand breaks (SSB) together with blocking of the enzymes that are involved in the repair (PARPs) will result in the accumulation of DBS that the tumour cells cannot repair and thus eventually in cell death (Underhill et al., 2011).

Because of the common structural organization of the PARPs, all of the family members can be inhibited by 3-ammobenzamide which binds to the nicotinamide region of the NAD binding site and prevents the enzyme to bind and use NAD (Buerkle, 2008). Several PARP inhibitors are currently in clinical trials including TIQ-A, olaparib and ABT-888. PARP inhibitors mainly target PARP1 however a lack of strict target specificity is common between all inhibitors (Wahlberg et al., 2012).

1.7.4 Poly(ADP-ribosyl)ation and DNA damage response (DDR)

PARylation is involved in the response mechanism against DNA damage (Malanga and Althaus, 2005, Zhou and Elledge, 2000, Golia et al., 2015, Oliver et al., 1999). Mostly PARP1, but also PARP2, can detect single and double strand DNA breaks leading to an increase in the catalytic activity of both PARP proteins. Recently, PARP3 also emerged as a part of the DDR (Beck et al., 2014). PARP1's affinity for single DNA ends/ lesions renders it a "nick sensor" in the cell (Shall and de Murcia, 2000). When activated, the two PARPs modify themselves and then recruit and PARylate other important proteins. A very important part of their role is this recruitment of DNA damage checkpoint proteins which will later on coordinate the downstream events and decide the fate of the cell. Target proteins of the PARPs all have in common a polymer binding domain comprised of 20-26 amino acids (Buerkle, 2008). Interestingly, in many cases this domain overlaps with other functional domains of the acceptor proteins and thus binding of the polymer causes a disruption of other functions and/or interactions of the target PARylated proteins (Buerkle, 2008), perhaps highlighting the high priority of the repair function.

1.7.5 Poly(ADP-ribose) glycohydrolase (PARG)

PARPs catalyze mono(ADP-ribosyl)ation of the acceptor protein substrate, elongation and branching of the poly(ADP-ribose) chain. The polymers can vary in length of units and they have very short half-life. The polymerization can be reversed by an enzyme called Poly(ADP-ribose) glycohydrolase (PARG) which can cleave the polymer branches (Bonicalzi et al., 2005).

There is only a single gene coding for PARG localized on chromosome 10q11.23 and three different isoforms of the protein can be detected as a result of alternative splicing of the gene transcripts. The gene comprises of 18 exons and 17 introns and from these, the catalytic region is found between introns 9 and 14 (Slade et al., 2011, Hassler et al., 2011). The largest PARG isoform can be found in the nucleus and has a molecular weight of 111kD. The two splice variants are detected in the cytoplasm and lack the first one and two exons resulting in a 102kD and a 99kD isoform respectively. The nuclear localization sequence (NLS) is on exon 1 which explains why the two splice variants are not localized in the nucleus (Meyer-Ficca et al., 2004).

The PARG enzyme is activated as a result of increased concentration of poly(ADP) ribose and has a balancing effect. It acts by hydrolyzing ribose bonds and releasing free ADP-ribose residues from both linear and branched forms of poly(ADP)ribose (Slade et al., 2011). PARG was discovered during the decade of 1970, it is present in all eukaryotic cells apart from yeast and its isoforms are differentially localized in the nucleus (Bonicalzi et al., 2005). Mostly due to the fact that PARG is present in small amounts in the cell, studying it has been proven to be a difficult task. Experiments on directed mutations of the gene in *Drosophila* showed that lack of it prohibited the larvae from developing into adult flies. The small percentage of flies that did progress to adulthood, suffered from neurodegeneration, pointing to the fact that PARylation and, equally importantly the regulation of PARylation, is also involved in the normal function of the neurons (Bonicalzi et al., 2005).

1.8 **Project aims**

CTCF has been pinpointed as a key regulator of many important cellular functions, mainly through engaging in protein-DNA interactions throughout the genome. However, the role of poly(ADP-ribosyl)ation of CTCF in the regulation of cellular processes is not well understood. We hypothesize that differentially PARylated isoforms of CTCF (CTCF130 and CTCF180) control different groups of genes (Hypothesis 1) and different functions (Hypothesis 2) in different biological situations. To investigate Hypothesis 1 the following Objectives will be studied as follows:

I. To identify the binding sites of the CTCF isoforms (CTCF130 and CTCF180) using a cell line model, LDM226, proliferating (expressing CTCF130 and CTCF180) and arrested (expressing only CTCF180). The high-throughput chromatin immunoprecipitation (ChIP-seq) technique will be used to achieve this Objective.

II. To identify gene expression profiles in the above cell line model by employing the high-throughput RNA-Seq technique.

III. To integrate the ChIP-seq output data with the global mRNA expression profile data generated by RNA-sequencing on the same cell model. This integration will provide useful information regarding the potential of CTCF180 as a gene expression regulator.

In the second part of our investigation, we will focus on the involvement of CTCF PARylation in the DNA damage response mechanism and test Hypothesis 2; to achieve this, Objective IV will be studied as follows:

IV. To investigate the role of CTCF PARylation in the context of DNA damage and repair. The cellular expression ratio and localization of CTCF, as well as the effect of PARylation inhibition will be studied in response to induced DNA damage.

Chapter 2 Materials and Methods

2.1 Cell lines and culture techniques

All cell culture experiments were performed in laminar flow type II hoods (Thermo Fisher Scientific, USA) using sterile plastic and glassware. 10% Virkon (DuPont, USA) solutions and 70% ethanol were used to ascertain aseptic conditions. All cell lines were maintained at 37^{o}C in a 5% CO_2 (SANYO, USA) incubator.

2.1.1 Cell lines

The majority of the experiments conducted in this study utilized the cell line 226LDM which derived from normal luminal breast cells. This cell line was immortalized using the viral constructs carrying the modified T antigen, TAg(U19dl89-97) (Cotsiki et al., 2005). Selected experiments were performed using a panel of cell lines derived from non-transformed cells such as the immortalized human Benign Prostatic Hyperplasia (BPH-1) (Hayward et al., 1995) and the human foreskin fibroblast cell line Hs27. A panel of cancer cell lines was also employed including the human epithelial cervical cancer cell line HeLa (Scherer et al., 1953) and the human ductal carcinoma cell line ZR-75.1 (Engel et al., 1978). In addition to these, the human embryonic kidney cell line 293T was also utilized (DuBridge et al., 1987).

2.1.2 Culture media

For culturing 226LDM cells the growth medium DMEM/F-12 (PAA) was supplemented with 5 μg / ml insulin, 1 μg / ml hydrocortisone, 20 ng / ml epidermal growth factor, 20 ng / ml cholera toxin (all from Sigma), 10% fetal bovine serum (FBS) (Biosera), and 50 μg / ml gentamicin (Life Technologies-Invitrogen).

The growth medium used for culturing Hs27, HeLa and 293T cells was the DMEM (Dulbecco's Modified Eagles Medium) (PAA-GE Healthcare) supplemented with 10% of FBS (Biosera) and 50 μg / ml of gentamicin (Life Technologies-Invitrogen).

The ZR-75.1 and BPH-1 cells were maintained in RPMI 1640 medium (PAA-GE Healthcare) supplemented with 10% FBS (Biosera) and 50 μg / ml gentamicin (Life Technologies-Invitrogen).

2.1.3 Cell culture techniques

2.1.3.1 Passaging cells

All cell lines used in this study grow adherently in a monolayer and require passaging when reaching approximately 70-80% confluence. To trypsinize, the spent growth medium was aspirated and 2 ml of EDTA (1X) (SIGMA) was used for washing the cells. To detach cells, 1 ml of warm 10% trypsin/EDTA (SIGMA) was added and the flasks were incubated for 2-10 min incubation at 37^{o}C. Especially for the 226LDM cells, due to being strongly adherent to the flask, 3 ml of trypsin were added instead of 1 ml and the incubation time was raised to 15-20 min. After confirming under the microscope that the cells had detached, the trypsin in the flask was diluted ten times with warm complete culture medium and then the cells were carefully transferred into a centrifuge tube. The cell suspension was centrifuged at 450 g for 5 min and the supernatant was aspirated off. The cell pellet was re-suspended in fresh growth medium and one-tenth of the suspension was transferred in a culture flask for future culture. The remaining cell suspension was used according to the requirements of other experiments.

2.1.3.2 Counting and seeding of cells

Cells from the final cell suspension as described in 2.1.3.1 were counted using a haemocytometer. The concentration of cells per ml was determined by calculating the average of cells contained in the four corners of a haemocytometer and multiplying it by 10^4. Depending on the experiment, the cells were diluted to the required concentration using growth medium and seeded in 6- or 12-well plates or flasks.

2.1.3.3 **Freezing and defrosting cell stocks**

Cells were trypsinised as described in 2.1.3.1, counted and 1 ml of cell suspension, containing approximately 1 x 10^6 cells, was mixed with 1 ml of freezing mix composed of 20% Dimethyl Sulfoxide-DMSO (SIGMA) and 80% FBS. The suspension was transferred to a labelled cryotube wrapped in insulating material to ensure that the drop of temperature will be achieved gradually. The tube was stored in -80^{o}C and after 2-3 days it was moved to the liquid nitrogen storage tank.

To defrost cells, the cryotubes containing cells were removed from the liquid nitrogen tank on dry ice. The following day, the cells were defrosted by placing the vial under warm running water. The defrosted suspension was added to warm complete medium and centrifuged at 200 g for 3 minutes. The pellet was re-suspended in 9 ml of complete medium and transferred to a small flask. The flask was maintained at 37^{o}C and 5% CO_2 until reaching confluence of 70-80%.

2.2 Cell culture treatments

2.2.1 Cell cycle arrest treatment with hydroxyurea and nocodazole

Cells approximately 60-70% confluent were stripped off their spent medium and fresh culture medium containing 100 mM hydroxyurea was added to the flask. The cells were incubated for 24 h at 37^{o}C and CO_2. After the end of the incubation time, the cells were incubated in fresh complete medium for 1hr at 37^{o}C and CO_2. After 1 hr the complete medium was aspirated off and fresh medium supplemented with 500 ng / ml nocodazole was added. After 24 h the detached cells were harvested, counted as described in 2.1.3.2 and used for a variety of assays.

2.2.2 DNA damage

2.2.2.1 DNA damage caused by exposure to hydrogen peroxide (H_2O_2)

Cells were seeded in 12-well plates (commonly 1.5 x 10^5 / well) and grown on cover-slips until reaching approximately 70% confluence. At that time, the spent medium was aspirated off and fresh complete medium containing 200 μM H_2O_2 was added to the wells. Depending on the requirements of the experiment, the treatment was allowed for different time-points varying between 5 and 90 min at all times incubated at 37^{o}C and CO_2. After the completion of the treatment, the cells were fixed with 4% paraformaldehyde for immunostaining experiments or harvested and lysed if they were used in a western blotting experiment.

2.2.2.2 DNA damage caused by exposure to X-irradiation

Cells grown in flasks were trypsinised (as described in 2.1.3.1) and aliquoted to microcentrifuge tubes with a concentration of 6 x10^5 / tube. The cells were placed in an iron ice box and then irradiated using an X-ray generator from CHF Müller (Germany). The dose to be delivered was applied by variation of irradiation time at a fixed dose rate. The irradiation parameters were: 70 keV energy, 1.25-mm aluminium filter, and 9.4 mA current.

2.2.3 Treatment with the PARP inhibitor ABT-888

Cells were seeded in 12-well plates and grown on cover-slips. When they reached 70% confluence the spent medium was removed and fresh medium was added supplemented with 5-10 μM ABT-888. The incubation, at 37^{o}C and CO_2 varied between 2 and 12 h.

2.3 Mammalian cell transfection

Transient transfection is a method of introducing exogenous DNA into the nucleus of eukaryotic cells (Scangos and Ruddle, 1981, Scangos et al., 1981). Plasmid DNA carrying a gene of interest was transfected into mammalian cells for the purposes of this PhD study. Descriptions of the plasmids used in this study are summarized in table 2-1. Two transient transfection methods were used; the calcium phosphate method and the SuperFect transfection reagent method from Qiagen.

2.3.1 DNA transfection using the calcium phosphate method

The calcium phosphate method is based on the formation of micro-precipitates in DNA-calcium mixtures. These adhere to the surface of cells and enhance the DNA uptake capacity of cells (Kingston et al., 2001).

Cells reaching 60-70% confluence were stripped of their spent medium and supplied with growth medium without antibiotic. They were incubated in this medium for 1 h. In the meantime, the transfection solutions were prepared. For each well of a 12-well plate a total of 100 µl of transfection solution was prepared. This composed of 1 µg of plasmid DNA in 50 µl TE buffer (10 mM Tris/HCl pH 8.0 / 1 mM EDTA) and 2 M $CaCl_2$. The mix was carefully added to 50 µl of HBS buffer (Hepes 50 mM / NaCl 280 mM/ sodium phosphate 1.5 mM). The mixture was incubated for 20 min, then transferred drop-wise into the well and incubated for a period of 48 h.

2.3.2 DNA transfection using the SuperFect transfection reagent (Qiagen)

According to the manufacturer, SuperFect consists of activated-dendrimer molecules with a defined spherical architecture. These are branched and have charged ends that attract the negatively charged DNA. The DNA-dendrimer structures adhere to the cells and they enter the cell through non-specific endocytosis. The reagent buffers the pH of the endosome, leading to pH inhibition of endosomal nucleases, which ensures stability of SuperFect–DNA complexes.

Transfection using the SuperFect transfection reagent from Qiagen was used for enhanced transfection efficiency following the manufacturer's instructions.

Table 2-1 Plasmids used in transient DNA transfection experiments

Plasmid	Description	Source
pEGFP-C1	Enhanced Green Fluorescent Protein	Clontech
pEGFP-CTCFwt	CTCF cDNA cloned in pEGFR-C1 vector	(Farrar et al., 2010)
pEGFP-CTCFmut	PARylation – deficient CTCF cDNA cloned in pEGFR-C1 vector	(Farrar et al., 2011)

2.4 General microbiology techniques

2.4.1 Bacterial culture

All bacterial culture work was carried out in a specific laminar flow hood designated for bacterial work. DH5α, an *Escherichia coli (E. coli)* strain, were the competent cells (Invitrogen) used for plasmid transformation (Hanahan et al., 1991). The bacterial cells were grown in Luria Broth (LB). To prepare 1 liter of LB we used 10 g NaCl, 10 g Bactotryptone and 5 g yeast extract supplemented either with 100 μg / ml of ampicillin (SIGMA) or 50 μg / ml kanamycin (SIGMA).

2.4.2 Amplification of plasmid DNA using bacterial system

2.4.2.1 Transformation of competent cells

Competent bacterial cells (DH5α) were thawed on ice. 0.5 μg of the plasmid DNA was added to 50 μl of the competent cells and the mix was incubated on ice for 30 min. The cells were heat shocked by placing the tubes in a heat-block at 42^{o}C for 20 seconds. Following the heat shock the cells were immediately placed on ice. After 2 min, 950 μl of warm LB medium was added to the cells. The tube was placed in the Stuart Orbital shaker incubator S150 at 225 rpm and 37^{o}C for 1 h. From the cell suspension, 100 μl were spread onto an agar plate containing antibiotic. The remaining cell suspension was centrifuged at 2500 g for 5 min and the pellet was re-suspended in 100 μl of LB. a second agar plate was plated and both plates were incubated overnight at 37^{o}C. Bacterial colonies could be seen in both plates the following day.

2.4.2.2 Inoculation of LB medium for plasmid DNA extraction

From one of the plates prepared as described in 2.4.2.1 a single colony was carefully picked using a sterile tip. The colony was dropped in a culture tube containing 5 ml of LB medium supplemented with antibiotic. The culture was incubated overnight in the Stuart Orbital incubator at 37^{o}C. The following day, 3 ml of the transformed bacterial cells were harvested by centrifugation (5000 g / 3 min) and used for small scale DNA extraction. The remaining 2 ml

were stored at 4^{o}C. Once enzymatic digestion of the isolated DNA confirmed that the plasmid transformation and culture was successful, the stored 2 ml were used for the preparation of culture for plasmid isolation at a bigger scale.

2.4.3 Plasmid DNA isolation from bacterial cells

2.4.3.1 Small scale plasmid DNA isolation

Plasmid DNA was isolated from the small scale culture that was prepared as described in 2.4.2.2. This was achieved using the QIAPrep Spin Miniprep kit (Qiagen) following the manufacturer's instructions. The quality and quantity of the obtained DNA was measured by using the Nanodrop ND-1000 spectrophotometer using the manufacturer's guidelines.

2.4.3.2 Large scale plasmid DNA isolation

For isolation of larger quantities of plasmid DNA, the Endotoxin-free Plasmid Maxi prep kit (Qiagen) was used according to the manufacturer's instructions. The Nanodrop spectrophotometer was used to assess the quality and quantity of the isolated DNA.

2.4.4 DNA quantification and quality control

2.4.4.1 Isolated DNA quality control using restriction enzymes

Enzymatic digestions were done on plasmids prepared as described in 2.4.3. in each case a detailed restriction site map was generated using the online software NEB cutter (Vincze et al., 2003). One microliter of the appropriate restriction enzyme (FastDigest from Thermo Scientific) was used to digest 1 μg of plasmid DNA. In the solution the appropriate buffer solution provided by the manufacturer was also added. The reaction was incubated for 30 min at 37^{o}C before running the digested and undigested DNA in agarose gel.

2.4.4.2 **Agarose gel electrophoresis**

Agarose gel electrophoresis was used to measure the size of the digested DNA (and RNA) according to how it migrates in the agarose matrix in response to electric current. Agarose gel was prepared by 1 g of agarose powder (for 1% gel) added to 100 ml of 1 x TAE (Tris Acetate EDTA) as well as 5 μl of ethidium bromide. The samples, containing a tracking dye (30% glycerol / 0.25% bromophenol blue), were added to the wells of the gel and the gel was run in 1 x TAE (40 mM Tris/ 20 mM acetic acid/ 1 mM EDTA) buffer. Apart from the samples, a DNA ladder (GeneRuler™ DNA ladder, Fermentas) of appropriate size range was also loaded in parallel to help with the estimation of the unknown samples. The samples could be visualized under a UV light system (Alpha Innotech and Fusion Fx7 from Peqlab, Germany) because of the ethidium bromide which is contained in the gel and has the capacity to intercalate into DNA (LePecq and Paoletti, 1967).

2.5 Automated Fluorimetric detection of Alkaline DNA Unwinding (FADU) assay

The automated fluorimetric detection of alkaline DNA unwinding (FADU) assay was used to measure DNA strand breaks in large populations of cells (Moreno-Villanueva et al., 2011). The assay is based on the principles of the original work from Birnboim and Jevcak (1981). The automated FADU modified by Moreno-Villanueva et al. (2009) produces less bias by being conducted by a liquid handling robot in a controlled environment. The assay has increased throughput while requiring smaller number of output cells per assay compared to its predecessor. An overview of the steps involved in FADU can be seen in figure 2-1.

2.5.1 Cell preparation

Cells (6 x 10^5 / well) were seeded in custom-made 96-well plates, treated and allowed repair incubation time according to the experiment requirements. After the completion of the treatment/ repair protocol, the culture medium was aspirated off the well plate and the plate was positioned in the working space of the robotic device and kept in dark at 0^oC.

2.5.2 Lysis

In an automatic manner, dispensing of 70 µl of lysis buffer (9 M urea / 10 mM NaOH / 2.5 mM cyclohexyl-diamine-tetraacetate / 0.1% sodium dodecyl sulphate / at RT) was completed for each well at a rate of 150 µl / s. Incubation with this buffer for 12 min causes the cellular membranes to lyse and the contents of the cells to come into solution.

2.5.3 DNA unwinding

Following cell lysis, 70 µl of alkaline solution (0.425 parts lysis solution in 0.2 M NaOH, pre-cooled to 0°C) was added on top of the cell lysate in such a way as to form a second layer, thus avoiding any mixing with the lysate. To achieve this, the robotic pipettes were placed at a controlled distance from the well-plates and the dispensing speed was lowered to 10 µl / s. The

pH of the cell solution during the 12 min incubation with this buffer reaches 12.5 which causes the unwinding of the DNA double helix in the contents of the samples.

2.5.4 Neutralization

To stop the DNA unwinding after the 12 min of incubation, 140 µl of neutralization solution (14 mM β-mercaptoethanol / 1 M glucose) were added at a rate of 200 µl / s. At this point the temperature was shifted to 22°C.

2.5.5 SybrGreen® addition and fluorescence detection

156 µl of diluted SybrGreen (1:8,333 in H_2O) (from MoBiTec, Germany) were dispensed onto the samples and the robotic pipettor mixed the solution by re-suspending. The fluorescence emitted from the samples was analysed immediately in a fluorescence plate reader at 492 nm excitation / 520 nm emission. As SybrGreen intercalates to double stranded DNA, the level of fluorescence was inversely correlated with single stranded DNA representing DNA damage. Statistical significance was evaluated by using *t*-test.

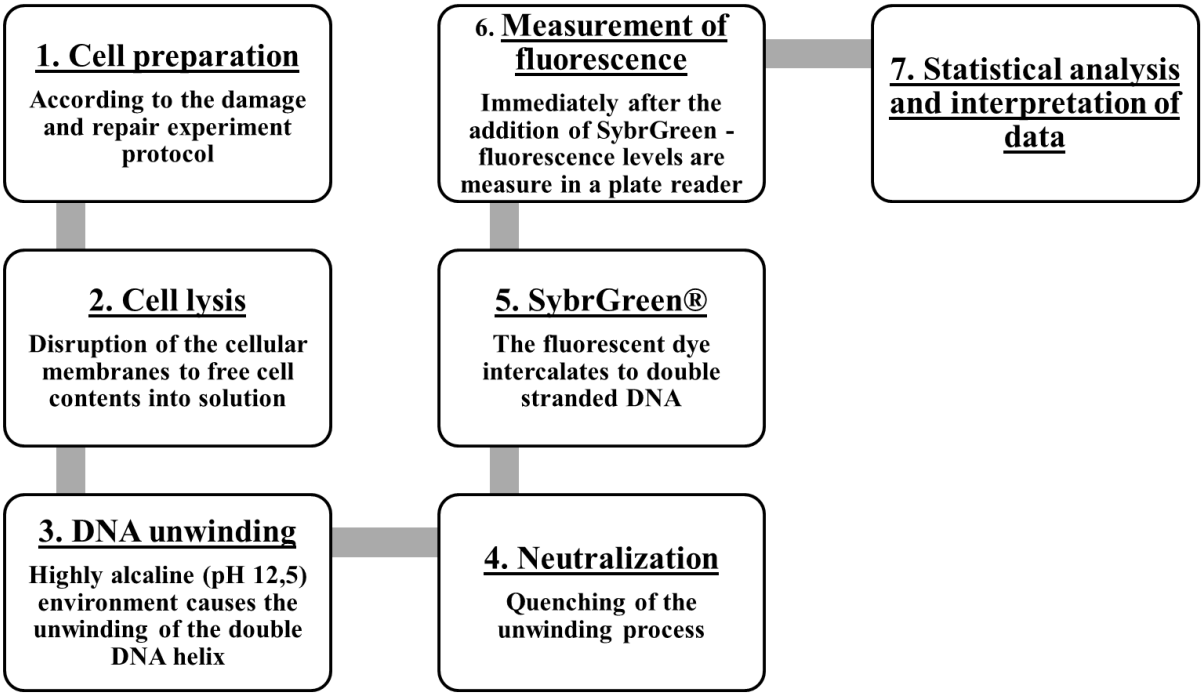


Figure 2-1 Diagrammatic representation of the stapes included in the automated FADU assay

2.6 Total RNA extraction and purification

For RNA experiments, all work was done under a fume hood using sufficient amounts of the RNA decontamination solution RNase-zap (Life Technologies). Total RNA from cell lines was extracted using the TRIsure reagent (Bioline) according to the manufacturer's guidelines.

2.6.1 Extraction of total RNA from cell lines

Cells were cultured in a flask until 70-80% confluent. At that point, the medium was aspirated off and 2 ml of ice-cold PBS were added to the attached cells. The flask was drained again and new PBS was added (1 ml). The cells were scrapped off and collected in a microcentrifuge tube. Centrifugation at 300 g for 5 min was done to obtain a cell pellet. The supernatant was removed and the white-coloured pellet was re-suspended in 1 ml of PBS. The pellet was once again centrifuged and the supernatant removed. 1 ml of TRIsure was added to the pellet and following re-suspension, the solution was incubated for 5 min at RT. 200 μl of chloroform were added and the tube was shaken with force and incubated for 15 min at RT. At this point the solution had separated into three phases, the top aqueous phase, the white phase and the oily bottom layer. The solution was centrifuged at 9,500 g for 15 min at 4^{o} and the separated top aqueous layer was carefully extracted and transferred into a fresh centrifuge tube. The volume of this separated phase was approximately 250 μl. The genetic material was precipitated with 625μl of isopropanol and incubated for 20 min on ice. The solution was centrifuged under the same conditions as previously, the supernatant was removed and the pellet re-suspended in 1 ml of 75% ethanol. The solution was centrifuged for 15 min at 4^{o}C at 6000 g. The ethanol washing step was repeated twice before removing the supernatant completely and air-drying the RNA pellet. The RNA was solubilized in sterile water (40-50 μl) and heated for 10 min at 55^{o}C. The pellet was stored at -80^{o}C.

2.6.2 RNA quantification and quality control

2.6.2.1 RNA quantification using the Nanodrop ND-1000

The Nanodrop ND-1000 was used for the quantification of the RNA samples. The manufacturer's instructions were followed. Nanodrop results are adequate for assessing the concentration of RNA samples but not to establish the integrity of the samples and to identify DNA contamination. For this reason the Bioanalyzer 2100 by (Agilent) is preferable.

2.6.2.2 RNA quality analysis using the Bioanalyzer 2100 (Agilent)

The quality of the RNA samples was assessed using the Bioanalyzer 2100 by Agilent according to the manufacturer's instructions. The assay is based on microfluidics technology and it is an efficient method for testing the quality of RNA samples prior to microarrays and sequencing techniques. The quality of the RNA was assessed from the electropherograms obtained at the end of the analysis. Only RNA of acceptable quality was used for the RNA-sequencing experiments.

2.7 Methods for protein extraction and analysis

2.7.1 Preparation of cell extracts for SDS-PAGE

Cells were trypsinized and centrifuged for 5 min at 400 g. The obtained pellet, was lysed immediately with SDS lysis buffer (0.1 M Tris/HCl pH6.8 / 7 M Urea / 4% SDS, phenol red dye and 10% β-mercaptoethanol -added before use) with a ratio of 20 μl of buffer per 1 x 10^5 cells. The lysate was vortexed and heated at 95^{o}C for 5 min. The prepared samples could be used immediately or stored at -20^{o}C.

2.7.2 Preparation of acrylamide gel, Sodium Dodecyl Sulphate –Polyacrylamide Gel Electrophoresis (SDS-PAGE) and Western Blot

SDS-PAGE is a method commonly used for the separation of proteins by electrophoresis based on their molecular weight (Chrambach and Rodbard, 1971). A porous polyacrylamide gel is used as a matrix and sodium dodecyl sulphate (SDS) is used to denature and negatively charge the proteins of interest. In SDS-PAGE, proteins migrate on the gel according to their size from the negative to the positive pole of the gel tank. Smaller proteins travel faster through the pores of the gel, while larger proteins migrate slower.

In this study an SDS-PAGE apparatus from Bio Rad was used. The composition of the buffers used in these experiments can be seen in table 2-2. The cell lysates were prepared as described in 2.7.1 and together with a pre-stained protein marker they were loaded on a 10% acrylamide gel (the composition of the gels can be seen in table 2-3). After the electrophoresis in running buffer (125 V / 40 mA / 2.5 h) the gel contained all the proteins of the cell lysates separated according to size. To detect and visualize a specific protein of interest SDS-PAGE was followed by Western Blotting (Towbin et al., 1979).

Western Blotting is a technique that follows SDS-PAGE and involves the transfer of all the proteins resolved in an acrylamide gel onto a membrane. This membrane can then be

incubated with a specific antibody which will bind to the protein of interest. This binding can be visualized and the protein size can be compared to the pre-stained marker.

The resolved proteins and marker were transferred to a polyvinylidene difluoride (PVDF) membrane (Millipore, USA). Prior to the transfer, the gel was incubated in running buffer 1% methanol for 15 min and the membrane was hydrated in absolute methanol for 10-15 sec and rinsed with RO H_2O. A stack consisting of 10 Whatman papers soaked in transfer buffer, in the middle of which the gel and membrane were placed, was put on the surface of the blotting apparatus. The semi-dry transfer was set for 2 hours (30 V / 100 mA).

After the transfer, the membrane containing all the proteins was washed in 20% methanol, rinsed with RO H_2O_2 and incubated overnight in blocking buffer, which prevents nonspecific binding. The membrane was then probed with the primary antibody (see table 2-4 for concentrations) for 2 h and then washed thrice with washing buffer. The membrane was incubated with the secondary antibody (prepared in blocking buffer) for 2 h at RT. The membrane was finally washed 3 times for 5 min each in washing buffer. All the secondary antibodies used in our study were conjugated with the horse radish peroxidase (HRP) enzyme which reacts with the substrate (UptiLight) provided in the Enhanced Chemiluminescence kit (ECL) (Interchim). The development of the signal with chemiluminescent peroxidase substrate for blotting was performed in a dark room using a red light and the signal was captured on an autoradiography film (Kodak, Japan). The signal on the film was visualized using the GBX Developer and Fixer (Kodak, Japan). For some of the Western blotting experiments in our study, the signal was visualized using the Fusion FX7 gel documentation system from Peqlab (Germany).

Table 2-2 Composition of buffers used in SDS-PAGE and Western Blotting

Buffer	Composition
SDS Lysis Buffer	0.1 M Tris/HCl pH 6.8, 7 M Urea, 4% SDS, phenol red dye 2 β-mercaptoethanol (just before use)
Resolving Buffer	2 M Tris/HCl(pH 8.9), 0.2% SDS
Stacking Buffer	0.1 M Tris/HCl (pH 6.8), 0.1% SDS, 5% Acrylamide / Bis-acrylamide 30% solution
Running Buffer	0.025 M Tris/HCl, 0.192 M Glycine, 0.1% SDS
Transfer Buffer	20 mM Na_2PO_4, 2% Methanol, 0.05% SDS
Blocking Buffer	0.1% Tween, 5% dried skimmed milk in powder form, 1x PBS
Washing Buffer	0.1% Tween, 1 x PBS

Table 2-3 Composition of resolving and stacking gels used in SDS-PAGE

Composition	Resolving Gel (10% gel) Volume (ml)	Stacking Gel (4% gel) Volume (ml)
Acrylamide / Bis-acrylamide 30%	3.3	-
Resolving buffer / stacking buffer	5	2
APS 10%	0.05	0.01
TEMED	0.02	0.004
ddH_2O	1.7	-

2.7.3 Immunofluorescence (IF) staining on fixed cells

In immunofluorescence staining, which is based on the same principle as immunocytochemistry staining, an antibody is used to detect a specific protein. This antibody is appropriately tagged with a visible dye such as a fluorescent dye (Ramos-Vara, 2005).

IF staining was performed on adherent cells grown on cover slips. The cells were fixed with addition of 4% paraformaldehyde (PFA). After three washes with PBS/glycine (0.1 M), they were incubated for 15 min in boiling citrate buffer (10 mM citric acid, pH 6.0). After incubation for 20 min with permeabilization buffer (0.25% Triton / PBS) the cells were washed thrice with 1 x PBS. The coverslips were placed in petri dishes, circled with hydrophobic marker on the slides and 90 μl of blocking buffer (0.05% Tween, 2% serum, 1% BSA / 1 x PBS) was added. A moist towel was put in the dish to keep the environment humid and the dish was covered and left on slow rocking in 4^{o}C overnight. A 2 h long incubation with primary antibodies in buffer (0.05% Tween, 1% BSA, in 1 x PBS) was followed by three PBS washes. Incubation continued in the dark (covered with foil) with secondary antibodies conjugated with fluorescent dyes (e.g. FITC, TRITC) for 1hr. After 3 more washes with 1 x PBS, the coverslips were left to dry and then they were mounted to microscope slides with DAPI (4’,6-diamidino-2-phenylindole, dilactate) mounting medium.

2.7.3.1 **Microscopy**

The majority of the immunofluorescence staining was observed under the Nikon Ti-Eclipse wide-field microscope which was used to capture the staining images. Staining from selected experiments was observed under the Nikon A1R confocal microscope. The visualisation tool used to view the images was Fusion FX viewer from Nikon.

Table 2-4 Antibodies used in our study and their applications

Primary Antibody / Product Code	Supplier	Application Concentration	Secondary Antibody	Supplier
Rabbit polyclonal, anti-CTCF **07-729, lot # JBC1903613**	Millipore	WB (1:10,000)	Goat anti-rabbit horseradish peroxidase (1:15,000)	Abcam
		IF (1:500)	Goat anti-rabbit fluorescein isothiocyanate (FITC) (1:500)	Millipore
		ChIP, IP (1:250)	-	-
Mouse monoclonal, anti-α tubulin **T6074**	SIGMA	WB (1:5,000)	Goat anti-mouse horseradish peroxidase (1:10,000)	Abcam
Mouse monoclonal, anti-PARP1	Provided by the Buerkle group, Konstanz, Germany	IF (1:400)	Goat anti-mouse tetramethylrhodamine (TRITC) (1:400)	Millipore
Rabbit polyclonal, anti-γH2AX **H5912**	Millipore	IF (1:200)	Goat anti-rabbit fluorescein isothiocyanate (FITC) (1:500)	Millipore
Mouse monoclonal, anti-UBF(F-9) **Sc-13125**	Santa Cruz	IF (1:300)	Goat anti-mouse tetramethylrhodamine (TRITC) (1:400)	Millipore
Rabbit polyclonal, anti-trimethyl-Histone H3 (Lys9) **07-523**	Millipore	ChIP (1:250)	-	-
Mouse monoclonal, anti-CTCF130	Peovided by the Loba-nenkov group, USA	ChIP (1:50)	-	-

2.7.4 Individual protein immunoprecipitation (IP)

This method is used to detect and immunoprecipitate a protein of interest out of a solution containing thousands of proteins using an antibody that specifically recognizes this protein (Kaboord and Perr, 2008).

Cells cultured on flasks were trypsinised according to protocol (2.1.3.1). After centrifugation, the cell pellet was washed twice with 1 x PBS and then lysed by vortexing in BF2 (25 mM Tris/Hepes - pH 8.0, 2 mM EDTA, 0.5% Tween20, 0.5 M NaCl, 1:100 Halt Protease Inhibitors). The lysate was incubated on ice for 15 min and then equal volume of BF1 (25 mM Tris/Hepes - pH 8.0, 2 mM EDTA, 0.5% Tween20, 1:100 Halt Protease Inhibitors) was added.

For Immunoprecipitation, the cell lysate was pre-cleared by incubating 500 µl of the lysate in 50 µl of pre-blocked Protein A/Sepharose beads for 30 minutes at 4^{o}C on a rotor shaker. The sample was then centrifuged at 200 x g for 1 minute at RT and the pre-cleared supernatant was transferred into a fresh centrifuge tube. 50 µl of the sepharose beads were added to the pre-cleared lysate along with the antibody (see table 2-4 for concentrations) and the samples were incubated overnight at 4^{o}C on a rotating wheel. On the following day, the immune-complexes were recovered by centrifugation at low speed for 1 min and the supernatant was removed. The pellet was washed three times with immunoprecipitation buffer (BF1+BF2) and each time the beads were collected with centrifugation at low speed for 1 minute. The sepharose was then lysed in SDS-lysis buffer and analysed by SDS-PAGE and western blot analysis as described in 2.7.2.

2.9 Chromatin Immunoprecipitation (ChIP)

Chromatin immunoprecipitation is a technique that allows the selection of specific protein-DNA complexes using an antibody that recognizes the protein of interest (Solomon et al., 1988, Collas, 2010). ChIP is commonly used in transcription factor research, especially when coupled with massive parallel sequencing (ChIP-seq) (Johnson et al., 2007)

In this study, chromatin immunoprecipitation was used to identify the genome-wide binding sites of the transcription factor CTCF in the 226LDM cell line. Two methods were employed to that end, the manual protocol and a specialized kit.

2.9.1 Manual Chromatin Immunoprecipitation method

2.9.1.1 Cell Sample Preparation for ChIP

Medium was drained off of one plate containing roughly 8 x 10^5 cells and 2 ml of 1 x PBS was added. To crosslink DNA and protein 54.7 μl of formaldehyde (37%) was added to the plate and left for 10 min in RT on a rocker to incubate. To quench the crosslink reaction 0.125 M of glycine was added and the plate was incubated for 5 min. The cells were scraped and transferred into two centrifugation tubes. Centrifugation for 5 min at 1000 x g at 4^{o}C was done to collect the cells as the supernatant was removed. 1 ml of ice cold hypotonic buffer was added in each of the microcentrifuge tubes and the samples were re-suspended a few times to make the final cell suspension. The samples were incubated on ice for 10 min and then span at 2000 x g for 5 min at 4^{o}C. The pellets were collected and re-suspended in 300 μl of LB-A buffer (50 mM Tris/HCL - pH8, 10 mM EDTA, 1%SDS, 1:100 Halt Protease Inhibitor Cocktail 100x Thermo Scientific x10) and incubated for 10 min on ice. Finally the samples were sonicated (3 x 7 min) with the Bioruptor from Diagenode to shear the DNA in fragments of 200-1000 bp size. To confirm the fragment size, samples were run on 1% agarose gel and observed under UV light as described in 2.4.4.2.

2.9.1.2 **Chromatin immunoprecipitation protocol**

After obtaining fragments of the desired size the cell suspension was diluted 10 times using ChIP dilution buffer (0.01% SDS, 1.1%Triton x-100, 1.2 mM EDTA, 16.7 mM Tris-HCL - pH 8.0, 16.7 mM NaCl, 1:100 Protease inhibitors). 10% of the diluted suspension was kept for Input, while the rest was used for antibody incubation and negative control. The Agarose beads were blocked overnight with BSA1% PBS 1x at 4^{o}C with agitation. The next day after washing the beads with PBS three times, antibody (see table 2-4 for concentrations) was added in one of the tubes at the right concentration and the samples were incubated for 1 hr at 4^{o}C on a rotating wheel. They were centrifuged briefly at 250 x g and 4^{o}C. The supernatant containing the unbound DNA was discarded, while the pellet containing the antibody-DNA complexes were washed for 3-5 min on a rotating wheel with 1 ml of the following buffers: 2x Low Salt buffer (0,1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris/HCl - pH 8.0, 150 mM NaCl), 1x High Salt (0.1% SDS, 1.1% Triton X-100, 2 mM EDTA, 20 mM Tris/HCl - pH 8.0, 500 mM NaCl), 2x ChIP dilution buffer (as described above), 3x TE buffer. All the buffers were kept on ice and the samples were centrifuged briefly (250 x g / 1 min / 4^{o}C) between each wash. After the last wash, DNA was eluted with the addition of 200 µl of freshly prepared Elution buffer (1% SDS/ 0.1 M NaHCO3) in each tube. In this step, the sample previously kept as "input", was eluted as well. The suspension was incubated for 10 min at 65^{o}C and then spun at 13,000 x g. The supernatant was carefully transferred to a new tube and the elution step was repeated. After that, 400 µl of eluates were obtained. In each one of the tubes 18 µl of 5 M NaCl was added and the samples were incubated for 4-5 h at 65^{o}C to reverse the DNA-protein crosslinks. Following this incubation step, 10 µl of 0.5 M EDTA, 20 µl of 1 M Tris/HCl - pH 6.8 and 1.5 µl of 14-22 mg/ml Proteinase K were added to the eluates and incubated for 1 hr at 45^{o}C. After the incubation step, 400 µl of phenol/chloroform (Fluka Biochemica) were added to each eluate. After 30 sec of vortexing the samples were centrifuged for 5min at 13,000 x g at 4^{o}C. 1.5 µl of glycogen and 1 ml of 100% ethanol were added to the samples and stored in -20^{o}C overnight. The following day the samples were spun for 20 min at 13,000 x g at 4^{o}C. Without disturbing the visible pellet, the supernatant was discarded and 1ml of 70% ethanol was added to each tube. The samples were spun again under the same conditions as before and the supernatant was discarded. The samples were left to air-dry and then 20 µl of sterile ddH_2O was added. The samples diluted in ddH_2O were stored in -20^{o}C.

2.9.1.3 **Chromatin Immunoprecipitation kit method.**

Several kits are commercially available to assist with the immunoprecipitation of protein-DNA complexes. The kit used for ChIP in our study was the Zymo Spin® DNA kit from (Zymo Research, USA). The samples were prepared, sonicated and handled following the protocol as described by the manufacturer.

2.9.1.4 **Measurement of ChIP DNA concentration using the NanoDrop 3300 fluorospectrophotometer**

The concentration of DNA in ChIP samples was in the pico-scale and a special spectrophotometer was used to establish the concentration of DNA in our samples. This was the NanoDrop 3300 (Thermo Scientific). The Quant-iT™ PicoGreen® ds DNA assay kit was used according to the manufacturer's instructions.

2.10 Next Generation Sequencing (NGS) techniques

Next generation sequencing (NGS) techniques including ChIP- and RNA-seq, also known as high-throughput sequencing, are a rapidly advancing field of genome-wide research, used for genome wide gene expression profiling, transcription factor and histone modification research (Cullum et al., 2011, Johnson et al., 2007).

In our study, we performed ChIP and RNA sequencing. The ChIP samples were generated with the Zymo Spin kit (Zymo Research) as described in 2.9.1.3 and the RNA samples were prepared as described in 2.6.1.

Samples for both experiments were delivered to the genetics laboratory at University College London (UCL), under the supervision of Prof Mike Hubank. The library preparation and sequencing using the sequencing platforms from Illumina was performed at UCL and King's College London.

2.10.1 Analysis of ChIP and RNA sequencing output data

The raw sequenced data from both ChIP and RNA-seq were received in FASTQ format (Cock et al., 2010). The bioinformatics analysis of both datasets was performed using a high performance computer (HPC) and the Galaxy bioinformatics tools organizing platform (Giardine et al., 2005, Goecks et al., 2010).

2.10.1.1 ChIP-sequencing data analysis

The raw sequenced data underwent initial quality control with the FASTQ Quality Trimmer tool from Galaxy. The Novoalign 3.2 tool was used to align the reads to the genome (Ruffalo et al., 2011) and the peak calling was done with the MACS2 algorithm from the Galaxy toolbox (Feng et al., 2012). The entries were filtered for *p* value lower or equal to 0.05, q value lower or equal to 0.4 and then the binding sites further form +/- 1000 from known genes were discarded.

The gene ontology analysis was performed using the Gene Ontology database (Ashburner et al., 2000, Harris et al., 2004), the database for annotation, visualization and integrated discovery (DAVID) (Huang da et al., 2009b) and the online enrichment analysis tool REViGO (Supek et al., 2011). Cytoscape (v3.2.1) was used to visualize and optically manipulate the ontology graphs (Shannon et al., 2003). The *de novo* discovery of binding motifs from the datasets was performed with use of the Multiple Em for Motif Elicitation (MEME) algorithm (Bailey et al., 2006). The visualization of the binding sites on human chromosomes was achieved using MATLAB installed on a workstation with multi-threading enabled (package: bioinfotoolbox MATLAB library v.R2012a).

2.10.1.2 **RNA-sequencing data analysis**

The FASTQ Quality Trimmer tool from Galaxy was used to discard the low quality reads. The Novoalign 3.2 tool was used to align the reads to the reference genome (human) and the raw counts were normalized using the Bam2fpkm tool form Galaxy. To discover the differentially expressed genes in our study and prepare the diagnostics plots we used the DESEQ toolkit. The ensuing lists of genes were filtered for p value (≤ 0.05) and q value (≤ 0.05). Gene ontology analysis as well as visualization of the findings on chromosomes was performed as described 2.10.1.1.

Chapter 3 Genome-wide analysis of CTCF binding in proliferating and arrested 226LDM cells using ChIP-Seq

3.1 Introduction / Background

CTCF was initially discovered as a transcriptional repressor of the c-myc gene in chicken and mammalian cells (Filippova et al., 1996, Rasko et al., 2001). However, in following research studies, it became clear that repression was not the only function that CTCF was involved in. It is now known to be a transcriptional repressor, silencer and an activator, the only insulator known in mammals, the regulator of imprinting and X-chromosome inactivation, and as an architectural protein capable of mediating chromatin structure remodelling (Ohlsson et al., 2001, Holwerda and de Laat, 2013, Ong and Corces, 2014, Burcin et al., 1997, Hancock et al., 2007, Qi et al., 2015).

3.1.1 CTCF binding

All functions attributed to CTCF can be exerted through specific, and at the same time divergent, DNA binding sites. While most DNA binding proteins have a consensus, a sequence they preferably bind to and only very rarely diverge from it, CTCF uses its 11 DNA binding zinc fingers in different combinations that allow it to bind to different motifs (Filippova et al., 1996). Recent studies of genome-wide CTCF distribution in different cells revealed tens of thousands CTCF binding sites in the genomes of these cells (Chen et al., 2012, Xie et al., 2007). Some specific motifs and sites are remarkably evolutionarily conserved among species, from *Drosophila* to man (Ni et al., 2012). However, CTCF binding capacity is dynamic resulting in thousands of unique binding sites even between cell lines originating from the same species. Moreover, it would appear that binding affinity decreases with uniqueness (Chen et al., 2012, Plasschaert et al., 2014).

Mainly, but not exclusively, CTCF binds in the proximity of the transcription start site (TSS) of genes. The transcription of thousands of genes is regulated by CTCF in this fashion,

but studies have also shown, or predicted, CTCF binding in non-coding regions, closed chromatin and areas far away from any gene (Chen et al., 2012). The mode of CTCF activity on the majority of these sites, if such exists, remains unclear.

The overall number of known and predicted binding sites of CTCF across the genomes of 15 different species has been estimated in the region of 15 million. Such extraordinary numbers lead to the creation of the CTCFBSDB (http://insulatordb.uthsc.edu/); a database designed to assist researchers in the field by collecting and making available all the binding sites of CTCF with useful information such as the exact location and experiment that led to its discovery (Ziebarth et al., 2013).

The complexity of CTCF binding has been the focus of many studies, gene-by-gene in the past and genome-wide in the recent years due to the development of next generation sequencing (NGS). Chromatin immunoprecipitation (ChIP) is a method developed to explore DNA binding proteins, such as transcription factors and histones by enriching the fragments to which they bind. Coupled with massively parallel sequencing of the ensuing fragments (ChIP-seq), it is currently leading the way in research aimed to unravel the mechanisms underlying gene regulation.

3.1.2 ChIP-seq technique

ChIP sequencing can be broken down to three major parts: (1) Chromatin immunoprecipitation, (2) sequencing of the isolated fragments and (3) computational analysis of the sequenced data (figure 3-1) (Cullum et al., 2011, Nielsen and Mandrup, 2014).

1. Chromatin Immunoprecipitation: Initially, all protein-DNA binding events are "locked in place" by chemical cross-linking, often using paraformaldehyde. The cells are then disrupted and sonicated, resulting in fragments of DNA ranging between 200-600bp. The fragments of interest – still in complex with the proteins – are immunoprecipitated using a specific ChIP-grade antibody and finally after reversal of the cross-linking, the proteins are removed and the

sample contains the DNA fragments where the protein of interest was binding to at the time of cross-linking.

2. Library preparation and sequencing: Following ChIP, library preparation includes size selection and amplification of the DNA fragments in each sample. The sequencing of the isolated fragment is performed by a high-throughput sequencing platform; the Mi-seq from Illumina and the 454 GS junior from Roche are among the most popular ones. Sequencing platforms have been compared in a number of studies, concluding that they essentially provide corresponding results (Quail et al., 2012, Loman et al., 2012).

3. Bioinformatics: The set of raw sequences corresponding to each fragment are stored in FASTQ format files (Cock et al., 2010) and several steps of computational analysis are required in order for information of biological value to be obtained from them.

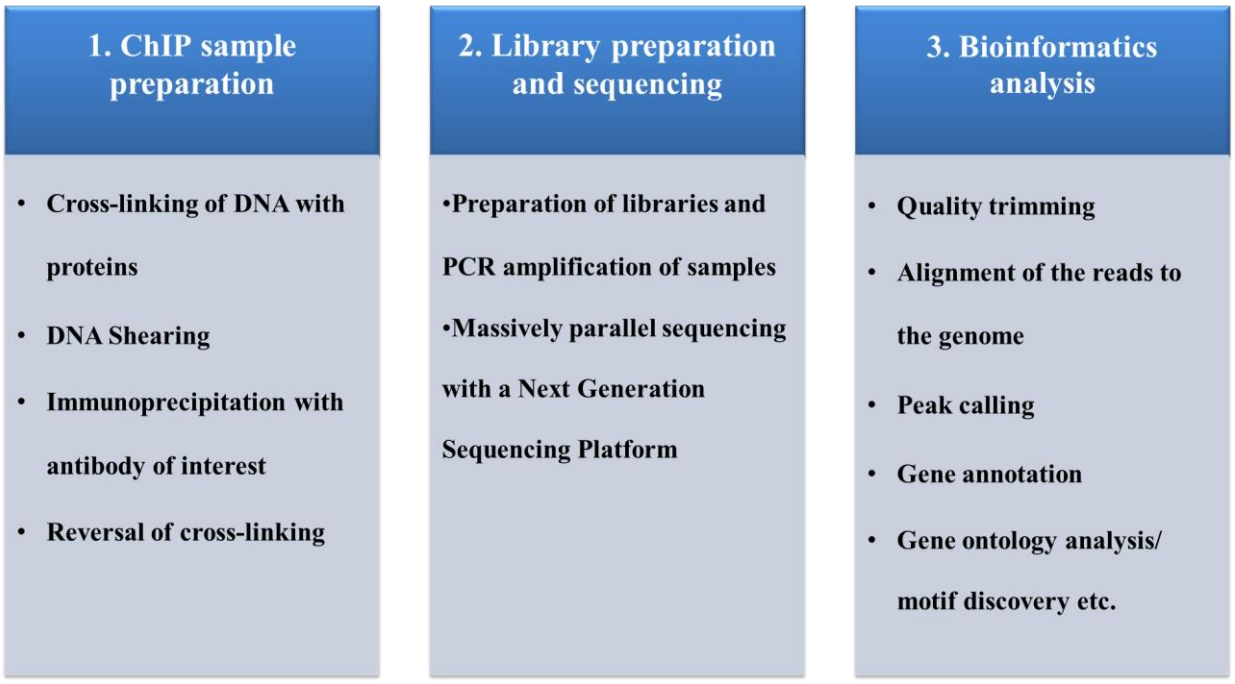


Figure 3-1 Diagrammatic presentation of the steps involved in a ChIP-sequencing experiment

The first step in a ChIP-sequencing experiment requires the immunoprecipitation of DNA fragments that the protein of interest binds to (1). This is achieved by cross-linking DNA with proteins, shearing of the DNA by sonication, selection of the fragments using the antibody of interest and then reversal of the cross-linking. At the end of this step, the sample will contain the DNA fragments that the protein of interest binds to. After chromatin immunoprecipitation, the samples are used to create sequencing libraries and the sequencing takes place in NGS platforms (2). The raw sequenced data must then be analyzed with bioinformatics tools and algorithms to elicit the biologically relevant information according to the experimental aims (3).

Initially, the sequences (reads) are aligned/ mapped to the reference genome of choice, by identifying all the possible locations in the genome that match the reads – allowing only for 1-2 mismatched base-pairs owing to single nucleotide polymorphisms (SNP) or small errors in the sequencing process.

After the reads have been mapped, the elimination of background signal is imperative as several of the previous steps can introduce artefacts, such as non-uniform DNA shearing during sonication or the non-specific binding of proteins other than the one of interest to the antibody used for immunoprecipitation. For this reason, control ChIP samples are usually prepared in advance. Controls can include sheared input DNA (input control) or a ChIP sample created without an antibody (mock control) or by using a non-specific antibody. All the reads found in the control sample are removed from the results in the sample of interest. The rationale behind this is that all the "background reads" are this way discarded while the remaining "true reads" correspond to the specific binding events of interest.

Peak calling software is subsequently employed to scan across the genome and discover the areas where the true reads appear to show enrichment. These enrichment peaks correspond to the binding sites that were isolated during ChIP. Each one of the peaks is associated with the nearest gene and the ensuing list contains all the binding sites discovered for the protein of interest for this experiment.

After these steps are completed, further analysis can be done depending on the experimental design, including comparison between different samples, retrieval of the top-scoring binding motifs, or a study on the ontology relationships between the discovered genes.

3.2 **Experimental Aims**

Currently, all the existing information regarding the binding characteristics of CTCF has been mined from experiments on the main, non-modified CTCF isoform, which has a molecular weight of 130 kD (CTCF130). However, a highly PARylated CTCF isoform has also been

discovered. The molecular weight of this isoform is 180 kD (CTCF180) and it is the only isoform that can be found in healthy breast tissue (Docquier et al., 2009). There is at present no information regarding the binding characteristics of this isoform.

This chapter will be aimed at the identifying the binding sites of CTCF in 226LDM cells, with special focus on the binding of CTCF180 and the genes it is implicated with in contrast with CTCF130. To achieve this, ChIP-seq experiments will be performed in two separate groups of 226LDM cells.

The first group will consist of control, proliferating cells which express both of the isoforms, while the second group will be chemically induced to stop proliferating. This treatment, with hydroxyurea and nocodazole, causes the depletion of CTCF130 and the expression of CTCF180 only.

Both groups of cells will be used in ChIP experiments using a polyclonal antibody recognising all CTCF isoforms as well as a monoclonal antibody recognizing the CTFC130 only. Therefore, the 226LDM cell line model provides us with the unique opportunity to study the binding of this isoform, whereby overcoming the problem with the absence of a specific antibody against the CTCF180.

Following ChIP, library preparation from the isolated DNA fragments will be performed at the University College London (UCL) and the sequencing of the enriched samples will take place at Kings College London using the Mi-seq platform from Illumina. The output data will be analyzed using the bioinformatics software organizing platform Galaxy among other specialized computational tools and algorithms (Giardine et al., 2005).

The study will then be directed to identify the genes associated with either or both of the CTCF isoforms. Special interest will be cast upon CTCF180 binding characteristics as there is currently no known information regarding the function and features of this isoform in the existing literature. The top-scoring motifs for each isoform will be discovered and a comparison

will be run in order to obtain further information regarding the possible differences between the isoforms.

3.3 Results

3.3.1 Cell cycle blocking treatment

226LDM cells in the control condition express both CTCF isoforms, namely CTCF180 and CTCF130. In order to manipulate the cells into expressing the 180kD CTCF isoform only, they were treated with hydroxyurea (HU) and nocodazole (NO), which block the cell cycle on the S and G2-M phase respectively.

Hydroxyurea is a useful tool for synchronizing asynchronous cell populations to the S phase of the mitotic cycle as it completely inhibits DNA replication (Sinclair, 1965, Yarbro, 1992), while nocodazole, in concentrations on the nano-molar scale, is involved in rearranging the dynamics of the mitotic spindle (Jordan et al., 1992, Vasquez et al., 1997).

Optimization of the effects of HU and NO treatment on the cells was performed by adjustments on parameters such as the incubation duration and the concentration of the chemicals. 24 and 48 hours of incubation with each and both of the two reagents were tested, while the concentrations ranged from 100 to 200 mM for HU and 500 to 1000 ng / ml for NO.

The CTCF expression profiles of the cells that were attached to the bottom of the flask as well as the detached cells were analyzed by western blotting (data not shown). This experiment confirmed that the most favourable conditions for this treatment to be effective in the case of 226LDM cells include a 24 hour-long incubation with 100 mM HU followed by 24 hours of incubation with 500 ng / ml of NO as described in materials and methods.

In these conditions, the detached 226LDM cells (>79% viable according to Countess® automated cell counter from Life Sciences, USA) express the CTCF180 isoform while the CTCF130 disappears (figure 3-2). This complies with the parameters of the protocol described by Docquier et al. (2009).

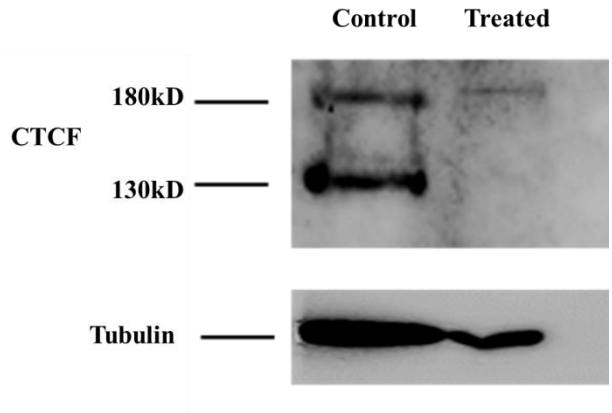


Figure 3-2 Western blotting of the control and nocodazole/hydroxyurea treated 226LDM cells

226LDM cells were cell-cycle arrested by addition of 100 mM hydroxyurea and 500 ng/ml nocodazole in their culturing medium. The treatment was administered as follows: After 24 hours of incubation with hydroxyurea, the cells were incubated with fresh complete culturing medium for 1 hour at 37^{o}C / 5% CO2. Nocodazole was added for another 24 hours after which the detached 226LDM cells were harvested and prepared for western blotting. The treated sample, as well a sample prepared from control (untreated) proliferating cells, were used for western blotting, probed with a polyclonal anti-CTCF antibody found to recognize both CTCF130 and CTCF180. The development of the signal was done with the UptiLight™ chemiluminescence substrate. Tubulin was used as a loading control.

3.3.2 Immunoprecipitation of CTCF180

As there is currently no available antibody specifically recognising CTCF180 and a ChIP experiment focusing on this isoform has not been performed before, a series of optimization experiments were conducted to assess whether the immunoprecipitation of this protein would be possible. Firstly, prior to ChIP, a protein immunoprecipitation (IP) experiment was performed using the treated cells. An IP involves the retrieval of a specific protein from a solution using an antibody raised against this protein. Both IP and ChIP techniques are based on the capacity of a protein to be immunoprecipitated by an antibody and therefore if one of the two experiments is successful, this can be an indication regarding the other.

The polyclonal antibody was used to immunoprecipitate CTCF from lysed 226LDM cells (figure 3-3). This experiment revealed that the CTCF180 can be successfully immunoprecipitated by the polyclonal antibody.

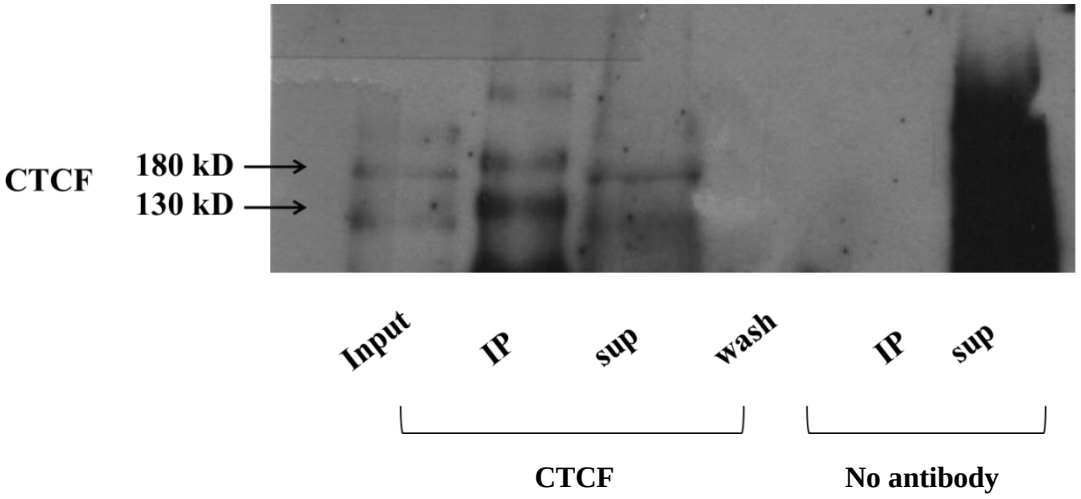


Figure 3-3 Immunoprecipitation of CTCF in 226LDM cells using a polyclonal antibody

226LDM cells were used in a single protein immunoprecipitation experiment to assess whether CTCF180 can be precipitated. The two isoforms are shown on the graph with black arrows. A no-antibody sample was used as control for the experiment. The visualization of the signal was performed using UptiLight™.

3.3.3 ChIP sample preparation and Shearing

The chromatin immunoprecipitation of CTCF from control and treated 226LDM cells was initially attempted following the protocol as described by Farrar et al. (2010), however the resulting DNA yield was insufficient for sequencing. To obtain the required amount of DNA we conducted the ChIP experiments using the Zymo-Spin™ ChIP Kit from Zymo Research (USA), following the manufacturer's protocol.

According to the experimental design of this study, a total of 16 ChIP samples were prepared (see table 3-1). For each of the two conditions, control and treatment, there were 4 ChIP samples produced in duplicates: (1) immunoprecipitated with the CTCF polyclonal antibody, (2) with the CTCF monoclonal antibody, (3) the histone (closed chromatin) antibody and (4) with no-antibody.

The control (untreated) ChIP samples were obtained from two biological replicates, each equally distributed into 4 ChIP samples. For treated samples, due to low yield of cells after each treatment, several biological replicates were pooled together into two groups. From each of these groups, 4 ChIP samples were produced.

The ChIP reactions of the 16 samples were completed with two experiments; 8 samples were prepared each time, and their duplicates were prepared in the second attempt (from table 3-1: samples 1, 3, 5, 7, 9, 11, 13 and 15 were included in ChIP experiment #1, while 2, 4, 6, 8, 10, 12, 14 and 16 in ChIP experiment no.2).

Prior each of the two immunoprecipitation experiments, the samples were fragmented by sonication in order to obtain DNA lengths ranging between 200-400 bp (figure 3-4).

The concentration of the DNA samples generated by ChIP was measured using the NanoDrop™ ND 3300 fluorospectrophotometer (Desjardins and Conklin, 2010) and the Quant-iT™ PicoGreen® fluorescent dye kit from Invitrogen.

Table 3-1 Experimental design of the ChIP-Seq experiments using control and treated 226LDM cells. The antibodies and their recognition molecules are indicated

	Sample number	Antibody	Recognizes
Control	1	CTCF polyclonal	CTCF130+CTCF180
	2		
	3	CTCF monoclonal	CTCF130
	4		
	5	Histone H3K9me3	Histone modification associated with closed chromatin
	6		
	7	No Antibody	-
	8		
Treated	9	CTCF polyclonal	CTCF130+CTCF180
	10		
	11	CTCF monoclonal	CTCF130
	12		
	13	Histone H3K9me3	Histone modification associated with closed chromatin
	14		
	15	No Antibody	-
	16		

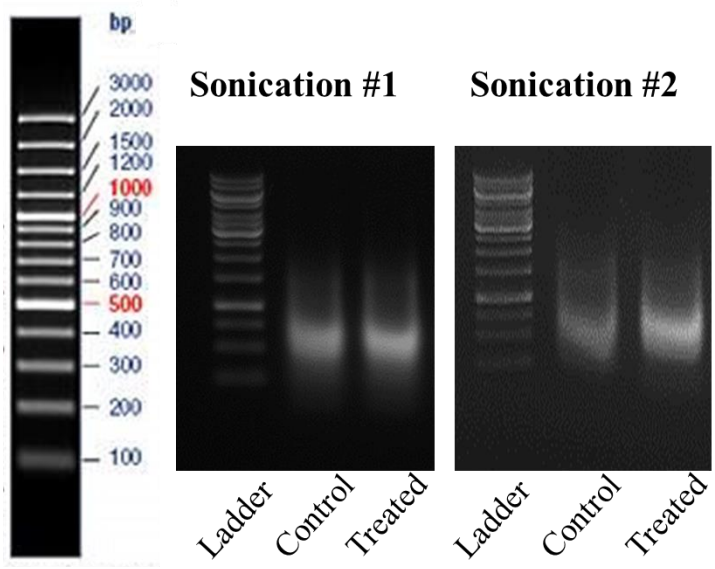


Figure 3-4 Fragmentation of cross-linked chromatin in control and treated 226LDM samples

Control and treated 226LDM cells were prepared as described previously. After cross-linking and prior to chromatin immunoprecipitation, the chromatin from both groups of cells was fragmented by sonication using the Bioruptor® from Diagenode. The samples were run on a 1% agarose gel to confirm that the lengths were between 200-400 bp. The GeneRuler™ 100 bp ladder from Thermo Scientific was used to estimate the size of the DNA fragments.

3.3.4 Library preparation and sequencing

The obtained DNA samples were delivered to UCL for the preparation of the libraries and subsequently the sequencing of the enriched DNA fragments was performed at Kings College London using the HI-seq platform from Illumina.

3.3.5 Computational analysis

The raw sequenced data was acquired from UCL stored in FASTQ format files. The data underwent the initial quality control with the FAST Q quality trimmer tool from Galaxy and then the Novoalign 3.2 tool was used to align the reads to the human genome (Ruffalo et al., 2011).

For this analysis, the values of the ChIP samples for inactive chromatin were used as controls to remove the background noise. All reads that were common between the closed chromatin samples and the CTCF antibody samples were discarded, aiming to eliminate artefacts stemming from the shearing and other processes of the sample preparation, which would be common between them.

After removing the background reads, peak calling was performed with the MACS2 algorithm (Feng et al., 2012). The ensuing lists of peaks for each sample were filtered to discard the non-significant ones (only entries with p value ≤ 0.05 and q value >0.4 were kept) and the binding sites that were within 1000 bp upstream or downstream of a gene were used associated with their closest gene. At the end of this process, the genes associated with the binding sites for each of the samples in the control and treated cells were discovered and listed in tables.

3.3.6 Analysis of CTCF binding sites in 226LDM cells

The binding sites from the ChIP-Seq samples obtained from the control and treated 226LDM cells using the monoclonal and polyclonal CTCF antibodies were then retrieved and analyzed. From the generated tables it was possible to compare the binding events between control and treated cells.

With the polyclonal antibody, 2051 sites were found in control 226LDM cells and only 64 in the treated cells. Given the experimental design of this study, we suppose that the binding of the polyclonal CTCF in the treated cells can be attributed to CTCF180.

Out of the 64 binding sites of CTCF180 in treated cells, 41 are common between the control and treated cell datasets for this antibody. This means that, judging from the binding of the polyclonal antibody, CTCF lost 2002 binding sites in the treated cells, retaining only 41 and gaining 23 new sites.

Presumably, these 41 common sites would be bound by CTCF180 in the control cells as well, however further experiments would be required to prove this. The remaining 23 sites appear to be uniquely bound by CTCF180 in treated cells, with the exception of two genes (RBM15B and FXYD3) which are bound by CTCF130 in control cells.

To produce a more stringent list of true targets, the false discovery rate (FDR)-adjusted p value (Q value) was used. The Q value implies the percentage of significant tests that would give a false positive (Benjamini and Hochberg, 1995). A Q value threshold of <0.4 was used.

In table 3-2 and table 3-3 the transcript and gene IDs associated with each binding event discovered with the polyclonal CTCF in control cells are shown in order of declining Q value. Next, in table 3-4 the binding sites in treated cells is shown in color-coded format according to whether this site is uniquely bound by the polyclonal antibody in treated cells (white entries), or if it is also bound by CTCF in control cells (blue entries).

The samples belonging to each of the colour-coded groups are shown next (table 3-5) featuring the description of each gene, when this is available. The descriptions were retrieved from the RefSeq project database (Pruitt et al., 2014) and the GeneCards® online database (Safran et al., 2010).

Table 3-2 Top 50 binding sites of the polyclonal CTCF antibody in control 226LDM cells

Transcript ID	Gene Symbol	Chromosome	Strand	Distance from the TSS	Q value
1. MEAF6:NM_022756	MEAF6	chr1	-	0	0,309780
2. GRM4:NM_001256809	GRM4	chr6	-	503	0,310226
3. EGR1:NM_001964	EGR1	chr5	+	116	0,310449
4. KCNA2:NM_001204269	KCNA2	chr1	-	467	0,311346
5. EPHA61:NM_001278301	EPHA6	chr3	+	82	0,311496
6. EPHA6:NM_001080448	EPHA6	chr3	+	82	0,311496
7. PLEKHG5:NM_001042665	PLEKHG5	chr1	-	0	0,311646
8. CNOT11:NM_017546	CNOT11	chr2	+	0	0,312022
9. SIRT5:NM_012241	SIRT5	chr6	+	-664	0,312399
10. STAG2:NM_001042749	STAG2	chrX	+	0	0,312475
11. STAG2:NM_001282418	STAG2	chrX	+	0	0,312475
12. OCRL:NM_000276	OCRL	chrX	+	0	0,312929
13. OCRL:NM_001587	OCRL	chrX	+	0	0,312929
14. FIBP:NM_004214	FIBP	chr11	-	0	0,313308
15. FIBP:NM_198897	FIBP	chr11	-	0	0,313308
16. SPG11:NM_001160227	SPG11	chr15	-	0	0,313460
17. SPG11:NM_025137	SPG11	chr15	-	0	0,313460
18. IDE:NM_004969	IDE	chr10	-	0	0,313764
19. C15orf40:NM_001160113	C15orf40	chr15	-	530	0,314145
20. C15orf40:NM_001160114	C15orf40	chr15	-	530	0,314145
21. C15orf40:NM_001160115	C15orf40	chr15	-	530	0,314145
22. C15orf40:NM_001160116	C15orf40	chr15	-	530	0,314145
23. C15orf40:NM_144597	C15orf40	chr15	-	530	0,314145
24. ARL6IP1:NM_015161	ARL6IP1	chr16	-	0	0,314988
25. CD300LG:NM_001168322	CD300LG	chr17	+	0	0,315295
26. CD300LG:NM_001168323	CD300LG	chr17	+	0	0,315295
27. CD300LG:NM_001168324	CD300LG	chr17	+	0	0,315295
28. CD300LG:NM_145273	CD300LG	chr17	+	0	0,315295
29. ACSS2:NM_001242393	ACSS2	chr20	+	0	0,315680
30. HOXC5:NM_018953	HOXC5	chr12	+	0	0,315757
31. FAM83H:NM_198488	FAM83H	chr8	-	156	0,315835
32. ADRA2A:NM_000681	ADRA2A	chr10	+	0	0,316221
33. ZNF398:NM_020781	ZNF398	chr7	+	0	0,316608
34. ZNF425:NM_001001661	ZNF425	chr7	-	0	0,316608
35. DDT:NM_001084392	DDT	chr22	-	0	0,316686
36. SAMD5:NM_001030060	SAMD5	chr6	+	802	0,318404
37. MTUS1:NM_020749	MTUS1	chr8	-	-906	0,318561
38. MYO7A:NM_000260	MYO7A	chr11	+	0	0,318718

39. MYO7A:NM_001127179	MYO7A	chr11	+	0	0,318718
40. MYO7A:NM_001127180	MYO7A	chr11	+	0	0,318718
41. NUTM1:NM_001284292	NUTM1	chr15	+	0	0,319191
42. NUTM1:NM_001284293	NUTM1	chr15	+	0	0,319191
43. NUTM1:NM_175741	NUTM1	chr15	+	0	0,319191
44. NOP10:NM_018648	NOP10	chr15	-	0	0,319191
45. LPGAT1:NM_014873	LPGAT1	chr1	-	0	0,320379
46. MESP1:NM_018670	MESP1	chr15	-	327	0,320538
47. TBX1:NM_005992	TBX1	chr22	+	0	0,320936
48. TBX1:NM_080646	TBX1	chr22	+	0	0,320936
49. TBX1:NM_080647	TBX1	chr22	+	0	0,320936
50. NARFL:NM_022493	NARFL	chr16	-	0	0,321576

Table 3-3 Top 50 genes associated with CTCF in control 226LDM cells with descriptions

Gene Symbol	Description
1. MEAF6	This gene encodes a nuclear protein involved in transcriptional activation. The encoded protein may form a component of several different histone acetyltransferase complexes. There is a pseudogene for this gene on chromosome 2. Alternative splicing results in multiple transcript variants.
2. GRM4	L-glutamate is the major excitatory neurotransmitter in the central nervous system and activates both ionotropic and metabotropic glutamate receptors. Glutamatergic neurotransmission is involved in most aspects of normal brain function and can be perturbed in many neuropathologic conditions. The metabotropic glutamate receptors are a family of G protein-coupled receptors, which have been divided into 3 groups on the basis of sequence homology, putative signal transduction mechanisms, and pharmacologic properties. Group I includes GRM1 and GRM5 and these receptors have been shown to activate phospholipase C. Group II includes GRM2 and GRM3 while Group III includes GRM4, GRM6, GRM7 and GRM8. Group II and III receptors are linked to the inhibition of the cyclic AMP cascade but differ in their agonist selectivities. Several transcript variants encoding different isoforms have been found for this gene.
3. EGR1	The protein encoded by this gene belongs to the EGR family of C2H2-type zinc-finger proteins. It is a nuclear protein and functions as a transcriptional regulator. The products of target genes it activates are required for differentiation and mitogenesis. Studies suggest this is a cancer suppressor gene.
4. KCNA2	Potassium channels represent the most complex class of voltage-gated ion channels from both functional and structural standpoints. Their diverse functions include regulating neurotransmitter release, heart rate, insulin secretion, neuronal excitability, epithelial electrolyte transport, smooth muscle contraction, and cell volume. Four sequence-related potassium channel genes - shaker, shaw, shab, and shal - have been identified in Drosophila, and each has been shown to have human homolog(s). This gene encodes a member of the potassium channel, voltage-gated, shaker-related subfamily. This member contains six membrane-spanning domains with a shaker-type repeat in the fourth segment. It belongs to the delayed rectifier class, members of which allow nerve cells to efficiently repolarize following an action potential. The coding region of this gene is intronless, and the gene is clustered with genes KCNA3 and KCNA10 on chromosome 1.
5. EPHA6	EPHA6 (EPH Receptor A6) is a Protein Coding gene. Among its related pathways are EPHA forward signaling and G-protein signaling_RhoA regulation pathway. GO annotations related to this gene include *ephrin receptor activity*. An important paralog of this gene is EPHB3.
6. PLEKHG5	This gene encodes a protein that activates the nuclear factor kappa B (NFKB1) signaling pathway. Mutations in this gene are associated with autosomal recessive distal spinal muscular atrophy. Multiple transcript variants encoding different isoforms have been found for this gene.
7. CNOT11	CNOT11 (CCR4-NOT Transcription Complex, Subunit 11) is a Protein Coding gene. Among its related pathways are Gene Expression and Deadenylation-dependent mRNA decay.
8. SIRT5	This gene encodes a member of the sirtuin family of proteins, homologs to the yeast Sir2 protein. Members of the sirtuin family are characterized by a sirtuin core domain and grouped into four classes. The functions of human sirtuins have not yet been determined; however, yeast sirtuin proteins are known to regulate epigenetic gene silencing and suppress recombination of rDNA. Studies suggest that the human sirtuins may function as intracellular regulatory proteins with mono-ADP-ribosyltransferase activity. The protein encoded by this gene is included in class III of the sirtuin family. Alternative splicing of this gene results in multiple transcript variants.
9. STAG2	The protein encoded by this gene is a subunit of the cohesin complex, which regulates the separation of sister chromatids during cell division. Targeted inactivation of this gene results in chromatid cohesion defects and aneuploidy, suggesting that genetic disruption of cohesin is a cause of aneuploidy in human

	cancer. Alternatively spliced transcript variants encoding different isoforms have been found for this gene.
10. OCRL	This gene encodes a phosphatase enzyme that is involved in actin polymerization and is found in the trans-Golgi network. Mutations in this gene cause oculocerebrorenal syndrome of Lowe and also Dent disease.
11. FIBP	Acidic fibroblast growth factor is mitogenic for a variety of different cell types and acts by stimulating mitogenesis or inducing morphological changes and differentiation. The FIBP protein is an intracellular protein that binds selectively to acidic fibroblast growth factor (aFGF). It is postulated that FIBP may be involved in the mitogenic action of aFGF. Two transcript variants encoding different isoforms have been found for this gene.
12. SPG11	The protein encoded by this gene is a potential transmembrane protein that is phosphorylated upon DNA damage. Defects in this gene are a cause of spastic paraplegia type 11 (SPG11). Multiple transcript variants encoding different isoforms have been found for this gene.
13. IDE	This gene encodes a zinc metallopeptidase that degrades intracellular insulin, and thereby terminates insulins activity, as well as participating in intercellular peptide signalling by degrading diverse peptides such as glucagon, amylin, bradykinin, and kallidin. The preferential affinity of this enzyme for insulin results in insulin-mediated inhibition of the degradation of other peptides such as beta-amyloid. Deficiencies in this protein's function are associated with Alzheimer's disease and type 2 diabetes mellitus but mutations in this gene have not been shown to be causitive for these diseases. This protein localizes primarily to the cytoplasm but in some cell types localizes to the extracellular space, cell membrane, peroxisome, and mitochondrion. Alternative splicing results in multiple transcript variants encoding distinct isoforms. Additional transcript variants have been described but have not been experimentally verified.
14. C15orf40	-No known function-
15. ARL6IP1	ARL6IP1 (ADP-Ribosylation Factor-Like 6 Interacting Protein 1) is a Protein Coding gene. Diseases associated with ARL6IP1 include spastic paraplegia 61, autosomal recessive. Among its related pathways are PAK Pathway and Phospholipase-C Pathway.
16. CD300LG	Members of the CD300 (see MIM 606786)-like (CD300L) family, such as CD300LG, are widely expressed on hematopoietic cells. All CD300L proteins are type I cell surface glycoproteins that contain a single immunoglobulin (Ig) V-like domain.[supplied by OMIM, Mar 2008]
17. ACSS2	This gene encodes a cytosolic enzyme that catalyzes the activation of acetate for use in lipid synthesis and energy generation. The protein acts as a monomer and produces acetyl-CoA from acetate in a reaction that requires ATP. Expression of this gene is regulated by sterol regulatory element-binding proteins, transcription factors that activate genes required for the synthesis of cholesterol and unsaturated fatty acids. Alternative splicing results in multiple transcript variants.
18. HOXC5	This gene belongs to the homeobox family of genes. The homeobox genes encode a highly conserved family of transcription factors that play an important role in morphogenesis in all multicellular organisms. Mammals possess four similar homeobox gene clusters, HOXA, HOXB, HOXC and HOXD, which are located on different chromosomes and consist of 9 to 11 genes arranged in tandem. This gene, HOXC5, is one of several homeobox HOXC genes located in a cluster on chromosome 12. Three genes, HOXC5, HOXC4 and HOXC6, share a 5' non-coding

	exon. Transcripts may include the shared exon spliced to the gene-specific exons, or they may include only the gene-specific exons. Two alternatively spliced variants have been described for HOXC5. The transcript variant which includes the shared exon apparently doesn't encode a protein. The protein-coding transcript variant contains gene-specific exons only.
19. FAM83H	The protein encoded by this gene plays an important role in the structural development and calcification of tooth enamel. Defects in this gene are a cause of amelogenesis imperfecta type 3 (AI3).
20. ADRA2A	Alpha-2-adrenergic receptors are members of the G protein-coupled receptor superfamily. They include 3 highly homologous subtypes: alpha2A, alpha2B, and alpha2C. These receptors have a critical role in regulating neurotransmitter release from sympathetic nerves and from adrenergic neurons in the central nervous system. Studies in mouse revealed that both the alpha2A and alpha2C subtypes were required for normal presynaptic control of transmitter release from sympathetic nerves in the heart and from central noradrenergic neurons; the alpha2A subtype inhibited transmitter release at high stimulation frequencies, whereas the alpha2C subtype modulated neurotransmission at lower levels of nerve activity. This gene encodes alpha2A subtype and it contains no introns in either its coding or untranslated sequences.
21. ZNF398	This gene encodes a member of the Kruppel family of C2H2-type zinc-finger transcription factor proteins. The encoded protein acts as a transcriptional activator. Two transcript variants encoding distinct isoforms have been identified for this gene. Other transcript variants have been described, but their full length sequence has not been determined.
22. ZNF425	ZNF425 (Zinc Finger Protein 425) is a Protein Coding gene. Among its related pathways are Gene Expression and Gene Expression. An important paralog of this gene is ZNF786. The encoded protein acts as a transcriptional repressor.
23. DDT	D-dopachrome tautomerase converts D-dopachrome into 5,6-dihydroxyindole. The DDT gene is related to the migration inhibitory factor (MIF) in terms of sequence, enzyme activity, and gene structure. DDT and MIF are closely linked on chromosome 22.
24. SAMD5	-No known function-
25. MTUS1	This gene encodes a protein which contains a C-terminal domain able to interact with the angiotension II (AT2) receptor and a large coiled-coil region allowing dimerization. Multiple alternatively spliced transcript variants encoding different isoforms have been found for this gene. One of the transcript variants has been shown to encode a mitochondrial protein that acts as a tumor suppressor and partcipates in AT2 signaling pathways. Other variants may encode nuclear or transmembrane proteins but it has not been determined whether they also participate in AT2 signaling pathways.
26. MYO7A	This gene is a member of the myosin gene family. Myosins are mechanochemical proteins characterized by the presence of a motor domain, an actin-binding domain, a neck domain that interacts with other proteins, and a tail domain that serves as an anchor. This gene encodes an unconventional myosin with a very short tail. Defects in this gene are associated with the mouse shaker-1 phenotype and the human Usher syndrome 1B which are characterized by deafness, reduced vestibular function, and (in human) retinal degeneration. Alternative splicing results in multiple transcript variants.

27. NUTM1	NUTM1 (NUT Midline Carcinoma, Family Member 1) is a Protein Coding gene. Among its related pathways are Chromatin Regulation / Acetylation. An important paralog of this gene isNUTM2B.
28. NOP10	This gene is a member of the H/ACA snoRNPs (small nucleolar ribonucleoproteins) gene family. snoRNPs are involved in various aspects of rRNA processing and modification and have been classified into two families: C/D and H/ACA. The H/ACA snoRNPs also include the DKC1, NOLA1 and NOLA2 proteins. These four H/ACA snoRNP proteins localize to the dense fibrillar components of nucleoli and to coiled (Cajal) bodies in the nucleus. Both 18S rRNA production and rRNA pseudouridylation are impaired if any one of the four proteins is depleted. The four H/ACA snoRNP proteins are also components of the telomerase complex. This gene encodes a protein related to Saccharomyces cerevisiae Nop10p.
29. LPGAT1	Acyl-CoA:lysophosphatidylglycerol (LPG) acyltransferase catalyzes the reacylation of LPG to phosphatidylglycerol, a membrane phospholipid that is an important precursor for the synthesis of cardiolipin [supplied by OMIM, Mar 2008]
30. MESP1	MESP1 (Mesoderm Posterior Basic Helix-Loop-Helix Transcription Factor 1) is a Protein Coding gene. Diseases associated with MESP1 include spondylocostal dysostosis. Among its related pathways are Cardiac Progenitor Differentiation. GO annotations related to this gene include *sequence-specific DNA binding transcription factor activity* and *protein dimerization activity*. An important paralog of this gene is MSGN1.
31. TBX1	This gene is a member of a phylogenetically conserved family of genes that share a common DNA-binding domain, the T-box. T-box genes encode transcription factors involved in the regulation of developmental processes. This gene product shares 98% amino acid sequence identity with the mouse ortholog. DiGeorge syndrome (DGS)/velocardiofacial syndrome (VCFS), a common congenital disorder characterized by neural-crest-related developmental defects, has been associated with deletions of chromosome 22q11.2, where this gene has been mapped. Studies using mouse models of DiGeorge syndrome suggest a major role for this gene in the molecular etiology of DGS/VCFS. Several alternatively spliced transcript variants encoding different isoforms have been described for this gene.
32. NARFL	NARFL (Nuclear Prelamin A Recognition Factor-Like) is a Protein Coding gene. Diseases associated with NARFL include pertussis. Among its related pathways are Metabolism andCytosolic Iron-sulfur Cluster Assembly. GO annotations related to this gene include *4 iron, 4 sulfur cluster binding*. An important paralog of this gene is NARF.
33. NTRK2	This gene encodes a member of the neurotrophic tyrosine receptor kinase (NTRK) family. This kinase is a membrane-bound receptor that, upon neurotrophin binding, phosphorylates itself and members of the MAPK pathway. Signalling through this kinase leads to cell differentiation. Mutations in this gene have been associated with obesity and mood disorders. Alternative splicing results in multiple transcript variants.
34. KCTD15	KCTD15 (Potassium Channel Tetramerization Domain Containing 15) is a Protein Coding gene. Among its related pathways are Activation of cAMP-Dependent PKA and Activation of cAMP-Dependent PKA. An important paralog of this gene is KCTD11. During embryonic development, interferes with neural crest formation (By similarity). Inhibits AP2 transcriptional activity by interaction with its activation domain.

35. SNX19	SNX19 (Sorting Nexin 19) is a Protein Coding gene. GO annotations related to this gene include *phosphatidylinositol binding*.
36. MYOM3	MYOM3 (Myomesin 3) is a Protein Coding gene. GO annotations related to this gene include *protein homodimerization activity*. An important paralog of this gene is MYBPC2.
37. OLFM1	This gene product shares extensive sequence similarity with the rat neuronal olfactomedin-related ER localized protein. While the exact function of the encoded protein is not known, its abundant expression in brain suggests that it may have an essential role in nerve tissue. Several alternatively spliced transcripts encoding different isoforms have been found for this gene.
38. PVRL1	This gene encodes an adhesion protein that plays a role in the organization of adherens junctions and tight junctions in epithelial and endothelial cells. The protein is a calcium(2+)-independent cell-cell adhesion molecule that belongs to the immunoglobulin superfamily and has 3 extracellular immunoglobulin-like loops, a single transmembrane domain (in some isoforms), and a cytoplasmic region. This protein acts as a receptor for glycoprotein D (gD) of herpes simplex viruses 1 and 2 (HSV-1, HSV-2), and pseudorabies virus (PRV) and mediates viral entry into epithelial and neuronal cells. Mutations in this gene cause cleft lip and palate/ectodermal dysplasia 1 syndrome (CLPED1) as well as non-syndromic cleft lip with or without cleft palate (CL/P). Alternative splicing results in multiple transcript variants encoding proteins with distinct C-termini.
39. CLUL1	-No known function-
40. GNAI2	The protein encoded by this gene is an alpha subunit of guanine nucleotide binding proteins (G proteins). The encoded protein contains the guanine nucleotide binding site and is involved in the hormonal regulation of adenylate cyclase. Several transcript variants encoding different isoforms have been found for this gene.
41. TMTC2	-No known function-
42. KMT2A	This gene encodes a transcriptional coactivator that plays an essential role in regulating gene expression during early development and hematopoiesis. The encoded protein contains multiple conserved functional domains. One of these domains, the SET domain, is responsible for its histone H3 lysine 4 (H3K4) methyltransferase activity which mediates chromatin modifications associated with epigenetic transcriptional activation. This protein is processed by the enzyme Taspase 1 into two fragments, MLL-C and MLL-N. These fragments reassociate and further assemble into different multiprotein complexes that regulate the transcription of specific target genes, including many of the HOX genes. Multiple chromosomal translocations involving this gene are the cause of certain acute lymphoid leukemias and acute myeloid leukemias. Alternate splicing results in multiple transcript variants.
43. PPIF	The protein encoded by this gene is a member of the peptidyl-prolyl cis-trans isomerase (PPIase) family. PPIases catalyze the cis-trans isomerization of proline imidic peptide bonds in oligopeptides and accelerate the folding of proteins. This protein is part of the mitochondrial permeability transition pore in the inner mitochondrial membrane. Activation of this pore is thought to be involved in the induction of apoptotic and necrotic cell death.
44. PIGO	This gene encodes a protein that is involved in glycosylphosphatidylinositol (GPI)-anchor biosynthesis. The GPI-anchor is a glycolipid which contains three mannose molecules in its core backbone. The GPI-anchor is found on many blood cells and

	serves to anchor proteins to the cell surface. This protein is involved in the transfer of ethanolaminephosphate (EtNP) to the third mannose in GPI. At least three alternatively spliced transcripts encoding two distinct isoforms have been found for this gene.
45. PEX5	The product of this gene binds to the C-terminal PTS1-type tripeptide peroxisomal targeting signal (SKL-type) and plays an essential role in peroxisomal protein import. Peroxins (PEXs) are proteins that are essential for the assembly of functional peroxisomes. The peroxisome biogenesis disorders (PBDs) are a group of genetically heterogeneous autosomal recessive, lethal diseases characterized by multiple defects in peroxisome function. The peroxisomal biogenesis disorders are a heterogeneous group with at least 14 complementation groups and with more than 1 phenotype being observed in cases falling into particular complementation groups. Although the clinical features of PBD patients vary, cells from all PBD patients exhibit a defect in the import of one or more classes of peroxisomal matrix proteins into the organelle. Defects in this gene are a cause of neonatal adrenoleukodystrophy (NALD), a cause of Zellweger syndrome (ZWS) as well as may be a cause of infantile Refsum disease (IRD). Alternatively spliced transcript variants encoding different isoforms have been identified.
46. LRP10	LRP10 (Low Density Lipoprotein Receptor-Related Protein 10) is a Protein Coding gene. Among its related pathways are Signaling by GPCR and Disease. An important paralog of this gene is LRP3.
47. SMIM13	-No known function-
48. RAB3IP	RAB3IP (RAB3A Interacting Protein) is a Protein Coding gene. GO annotations related to this gene include *Rab guanyl-nucleotide exchange factor activity*. An important paralog of this gene is RAB3IL1.
49. TTLL12	-No known function-
50. SEC24B	The protein encoded by this gene is a member of the SEC24 subfamily of the SEC23/SEC24 family, which is involved in vesicle trafficking. The encoded protein has similarity to yeast Sec24p component of COPII. COPII is the coat protein complex responsible for vesicle budding from the ER. The role of this gene product is implicated in the shaping of the vesicle, and also in cargo selection and concentration. Two transcript variants encoding different isoforms have been found for this gene.

Table 3-4 Binding sites of the polyclonal CTCF antibody in treated 226LDM cells: white entry, site is uniquely bound by the polyclonal antibody in treated cells; blue entries, the site is also bound by CTCF in control cells

Transcript ID	Gene Symbol	Chromosome	Strand	Distance from the TSS	Q value
S100A13:NM_001024213	S100A13	chr1	-	0	0,382065
S100A13:NM_001024212	S100A13	chr1	-	0	0,384789
S100A1:NM_006271	S100A1	chr1	+	0	0,384789
CISD1:NM_018464	CISD1	chr10	+	-615	0,384789
SGPL1:NM_003901	SGPL1	chr10	+	0	0,386166
TCTN3:NM_001143973	TCTN3	chr10	-	0	0,393198

Table 3-5 Genes associated uniquely with CTCF180 in treated 226LDM cells and genes associated with CTCF180 in treated 226LDM cells but also discovered in control cells; with descriptions

Gene Symbol	Description
CISD1	This gene encodes a protein with a CDGSH iron-sulfur domain and has been shown to bind a redox-active [2Fe-2S] cluster. The encoded protein has been localized to the outer membrane of mitochondria and is thought to play a role in regulation of oxidation. Genes encoding similar proteins are located on chromosomes 4 and 17, and a pseudogene of this gene is located on chromosome 2.
S100A13	The protein encoded by this gene is a member of the S100 family of proteins containing 2 EF-hand calcium-binding motifs. S100 proteins are localized in the cytoplasm and/or nucleus of a wide range of cells, and involved in the regulation of a number of cellular processes such as cell cycle progression and differentiation. S100 genes include at least 13 members which are located as a cluster on chromosome 1q21. This protein is widely expressed in various types of tissues with a high expression level in thyroid gland. In smooth muscle cells, this protein co-expresses with other family members in the nucleus and in stress fibers, suggesting diverse functions in signal transduction. Multiple alternatively spliced transcript variants encoding the same protein have been found for this gene.
S100A1	The protein encoded by this gene is a member of the S100 family of proteins containing 2 EF-hand calcium-binding motifs. This protein may function in stimulation of Ca2+-induced Ca2+ release, inhibition of microtubule assembly, and inhibition of protein kinase C-mediated phosphorylation. Reduced expression of this protein has been implicatedin cardiomyopathies.
SGPL1	-No known function-
TCTN3	This gene encodes a member of the tectonic gene family which functions in Hedgehog signal transduction and development of the neural tube. Mutations in this gene have been associated with Orofaciodigital Syndrome IV and Joubert Syndrom 18. Alternatively spliced transcript variants encoding multiple isoforms have been observed for this gene.

3.3.6.1 **Gene Ontology analysis**

To further compare the characteristics of the CTCF occupancies in samples obtained with the polyclonal CTCF antibodies (“pan”) in control and treated cells and in order to gain an understanding of the biological functions of CTCF in each case, a gene ontology analysis was performed.

Gene Ontology (GO) is based on universal annotation of genes with one or more GO terms according to the processes that it is involved. The GO terms can be specific or broad and they are structured in a parent-child manner, portraying the relationships between functions (Ashburner et al., 2000, Harris et al., 2004, Gene Ontology, 2008).

Gene Ontology annotations for the binding sites of the polyclonal CTCF in control and treated cells were retrieved form the gene ontology project database, the enrichment analysis was performed by DAVID functional annotation tool (Huang da et al., 2009b, Huang da et al., 2009a). Based on this analysis, in control cells the genes bound by CTCF are mostly involved in cellular protein localization, intracellular transport, membrane organization and neuron differentiation and development. On the other hand, the genes bound in the control cells are involved in cell cycle and differentiation (S100A13) and the hedgehog pathway (TCTN3). The set of proteins involved in the hedgehog pathway, which is mostly set in motion during embryogenesis, have also been linked with cancer and in particular with basal cell carcinomas (Rubin and de Sauvage, 2006, Von Hoff et al., 2009).

3.3.6.2 **Motif search**

One more method of analysing the characteristics of CTCF180 binding involves the analysis of the motifs to which it binds. The motifs corresponding to the top scoring peaks from the ChIP-seq analysis were entered into the motif finding algorithm MEME (Multiple EM for Motif Elicitation) (Bailey et al., 2006). The algorithm identifies shared motifs in a dataset of sequences and with statistical modelling it produces visuals of the motifs along with occurrence number and e-value. E-value expresses the "expected frequency" of random alignments of the same width as our motif that would generate the same or higher entropy score. Similarly with p-value, a motif scored with a low e-value is predicted more confidently than one with a higher e-value (Nagarajan et al., 2006).

A comparison between the lists of top 10 scoring motifs representing the binding events taking place within 1000 bp upstream or downstream of gene promoters in control and treated cells reveals divergent binding characteristics for CTCF in control and treated cells.

The most conserved and well documented consensus motif of CTCF is listed first in the control cells but fourth in the treated. There are many factors that can affect this list, including motif enrichment due to interaction with another protein or complex instead of direct binding to DNA. However the change in the list and the introduction of new motifs in it is an additional indication that the CTCF180 binding profile can be divergent from that of CTCF130.

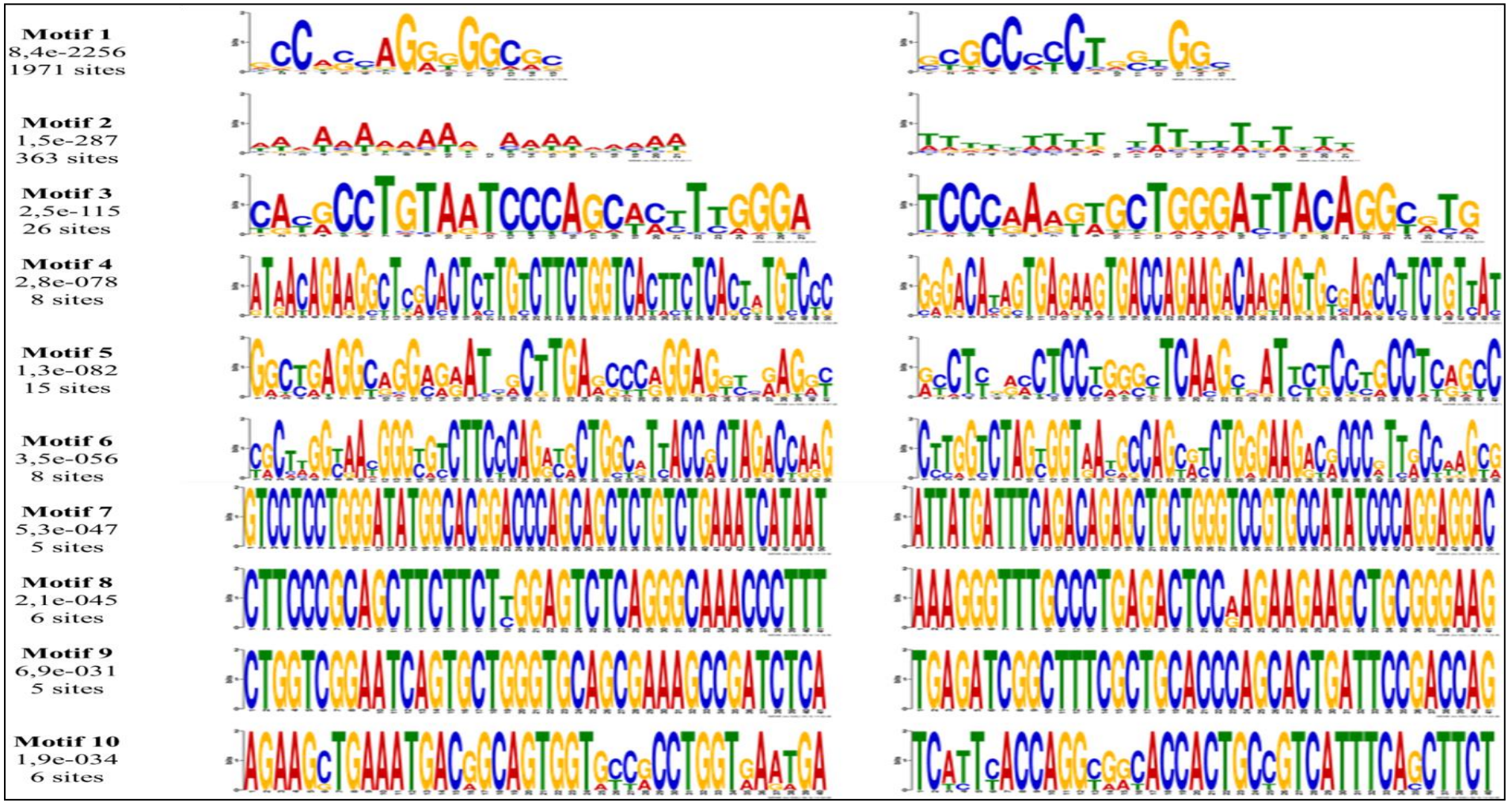


Figure 3-5 Top 10 binding motifs of CTCF discovered by ChIP with the polyclonal CTCF antibody in control 226LDM cells

Control 226LDM cells were used to conduct a ChIP-seq experiment using the polyclonal CTCF antibody which recognizes both CTCF130 and CTCF180 isoforms. Following bioinformatics analysis the top peaks representing the binding events were entered in the MEME algorithm for *de novo* discovery of binding motifs. The top ten discovered motifs and the frequency of their occurrence in our datasets are listed above in order of decreasing e-value. The reverse complement motifs are shown beside each motif.

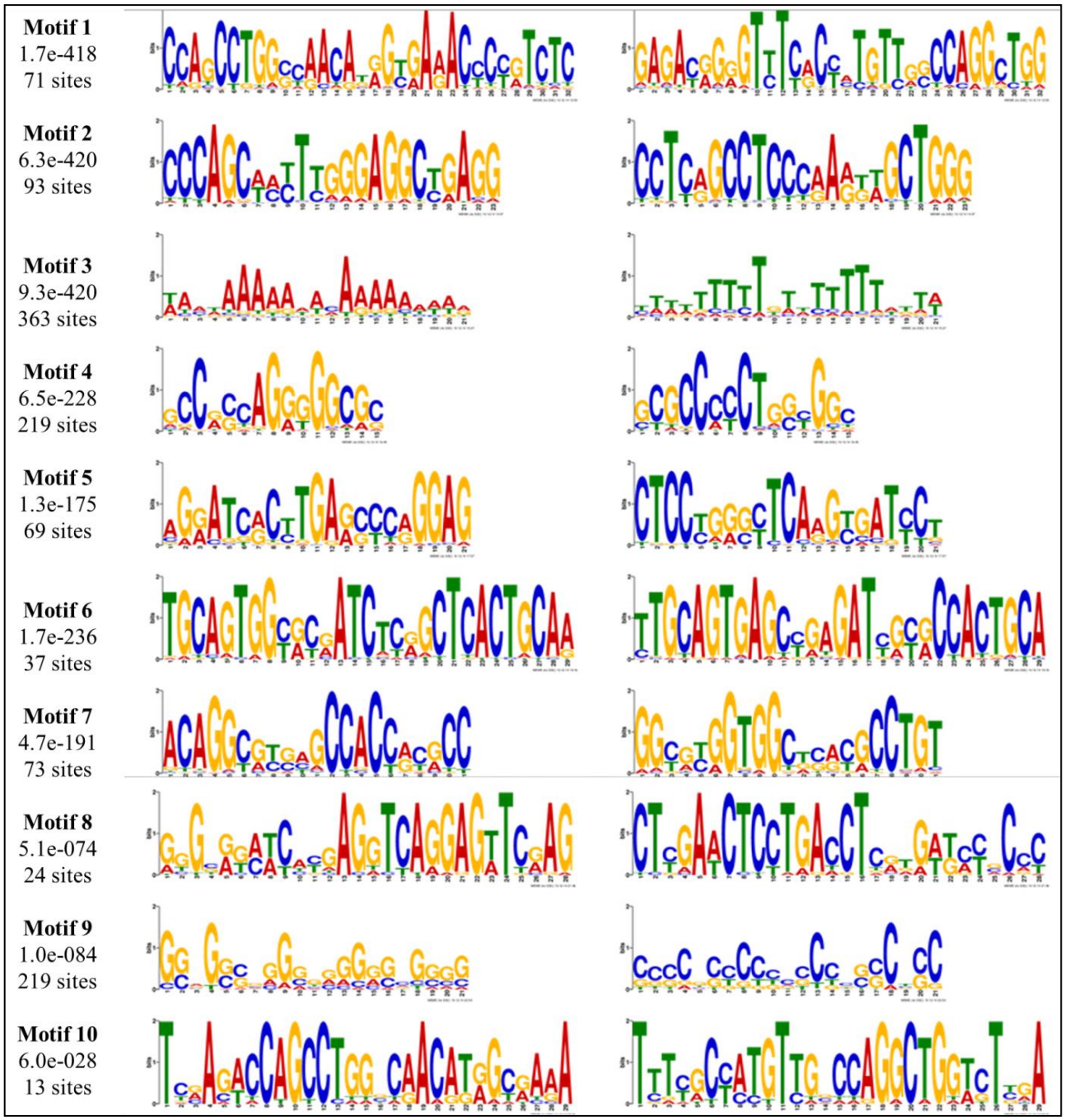


Figure 3-6 Top 10 binding motifs of CTCF discovered by ChIP with the polyclonal CTCF antibody in treated 226LDM cells

226LDM cells treated with hydroxyurea and nocodazole were used to conduct a ChIP-seq experiment using the polyclonal CTCF antibody which recognizes both CTCF130 and CTCF180 isoforms. Following bioinformatics analysis the top peaks representing the binding events were entered in the MEME algorithm for *de novo* discovery of binding motifs. The top ten discovered motifs and the frequency of their occurrence in our datasets are listed above in order of decreasing e-value. The reverse complement motifs are shown beside each motif.

3.4 **Discussion**

The aim of this chapter was to identify and analyze the binding sites of CTCF in 226LDM cells, with special focus on the binding of CTCF180. To achieve this, a cell line model was generated comprised of two populations of 226LDM cells. The first consisted of proliferating cells (control), which normally express both known CTCF isoforms, namely CTCF130 and CTCF180. The second group was exposed to controlled concentrations of hydroxyurea and nocodazole leading to a cell-cycle block at the G2-M stage. This treatment causes the disappearance of CTCF130 in 226LDM cells, whereas CTCF180 remains.

In the absence of a specific anti-CTCF180 antibody, this model makes the study of the binding for this isoform possible for the first time. Two ChIP-grade CTCF antibodies were at our disposal for this experiment; a polyclonal CTCF antibody recognizing both CTCF130 and CTCF180 and a monoclonal antibody recognizing CTCF130 only.

In our initial experiments, CTCF180 was successfully immunoprecipitated with the polyclonal antibody, which indicated that this antibody could be used for ChIP assays. The ChIP-DNA fragments obtained with this antibody (precipitating CTCF130 and CTCF180) and also the monoclonal antibody (precipitating CTCF130) were sequenced by a next generation sequencing methods and, subsequently, a bioinformatics analysis was performed using the raw output data. The analysis generated lists of the annotated regions associated with the CTCF binding sites discovered in each sample.

Taking into account the design of these experiments, the binding events discovered in the treated cells by the polyclonal antibody are presumably attributed to CTCF180. This list, although short, is the first evidence of CTCF180 binding and can be a stepping stone to future research of the binding profile of this isoform and its characteristics, for example in normal breast tissues where it is present as a sole isoform (Docquier et al., 2009).

It should be noted that experimental validation of these findings has not been conducted in this study due to time limitations. It will be a necessary step in the follow up investigation to ensure the accuracy of the data. Such validation would require real-time PCR assays using the ChIP materials from the control and treated cells obtained as described in this Chapter to confirm specific binding of CTCF180 and CTCF130 to specific targets identified in the ChIP-Seq analyses.

Moreover, for this study the ChIP background-reads were eliminated using the closed chromatin reads as control at the peak calling stage of the analysis. At this stage the decision was made to limit the discovered data to open chromatin binding sites, however it is known that CTCF can bind to chromatin in closed conformation, and therefore analysis including these regions should be performed. This could be accomplished using the no-antibody ChIP samples as controls. Comparing the outputs from the two pipelines would strengthen the robustness of the binding sites in open chromatin and provide new information about the closed chromatin binding events.

Furthermore, one of the most important factors contributing to ChIP-seq efficiency and quality is the ChIP-grade antibodies used to immunoprecipitate the protein of interest. In our investigation two ChIP-grade antibodies were used, a polyclonal and a monoclonal. The difference in the performance of the two antibodies was greater than originally expected; with the polyclonal over 2000 binding sites where discovered in the control cells, while with the monoclonal the number was lower, at 280. Such great number variation renders the comparison between the results from the two antibodies inadequately informative. Another antibody, specific for recognizing CTCF130, could solve this issue. Without a doubt, the production of a CTCF180-specific antibody would also enhance the current experimental framework.

It is worth highlighting that although we hypothesize that the binding sites obtained by the polyclonal antibody in treated cells are linked to CTCF180, these may not be the same sites that CTCF180 would bind to in control cells. Given the exposure to treatment we anticipate that the need for gene regulation changes and therefore the binding sites as well as the motifs would be affected accordingly.

Additionally, another possibility must be considered, that during ChIP some of the CTCF180 molecules could lose their PARylation mark. PARylation is known to be a dynamic procedure and the circumstances surrounding the switch between isoforms are currently obscure. This would account for the small number of binding sites in treated cells discovered by the monoclonal CTCF antibody.

What is more, one of the limitations that come with ChIP experiments is that it is not possible to distinguish whether a DNA fragment is immunoprecipitated due to direct contact of the protein to the binding site or if the protein of interest is there as part of a complex with another protein or proteins. In the latter situation, a binding site would be discovered in the ChIP sample of our protein of interest; however it would not be a true binding site for it.

All these limitations emphasize the argument for data validation; however they do not decrease the value of this study as the first attempt in exploring CTCF180 DNA binding.

The wealth of data that has been generated by this experiment has been vast and many more steps could be made to exhaust all avenues of research. Firstly, the discovered binding sites could be compared to those retrieved from preceding studies to determine whether any previously unknown sites were discovered in our experiments. Similarly, this could be applied to the discovered binding motifs as well. The motif search could be expanded to regions outside the promoters as well to assess whether these change. This could help to associate specific motifs with transcription on regulation or other functions.

Finally, the binding of CTCF alone does not provide any information about how or whether it regulates a gene. A combination of ChIP-seq with RNA-seq can provide more information about whether the binding of CTCF or, equally the loss of binding, affects the expression levels of a gene from one condition to the other.

In the following chapter, the isolation and sequencing of total RNA from control and treated 226LDM cells will be discussed and eventually, in chapter 5, an intersection of the two experiments will be explored.

Chapter 4 Genome-wide gene expression analysis in proliferating and arrested 226LDM cells using RNA-Seq

4.1 Introduction / Background

RNA is an important molecule which performs a wide range of molecular functions in various biological systems. The full set of RNA transcripts of a cell is composed of, in order of contribution, ribosomal (rRNA), transfer (tRNA), messenger (mRNA), as well as intronic and non-coding RNA (ncRNA) (Lindberg and Lundeberg, 2010). The field of study focused on the exploration of these transcripts and how they change under certain circumstances or developmental changes is called transcriptomics (Hine, 2008).

The fundamental target of transcriptomics is to identify and characterize the RNA contents of cells in various developmental stages and physiological/ pathological conditions, and ultimately to understand their function.

More and more evidence of the treasure of information that could be mined from RNA transcripts was discovered with the development of experimental techniques such as real-time PCR and, subsequently, with the development of gene expression microarrays (Valasek and Repa, 2005, Brown and Botstein, 1999).

However, it was the recent development of massively parallel sequencing platforms that admittedly revolutionized the field and became the springboard for unprecedented progress. RNA sequencing (RNA-seq) is a Next Generation Sequencing (NGS) method used to study the complexity of the transcriptome and it offers a better chance of understanding the role of the its constituents as well as their role in development and disease (Wang et al., 2009, Marguerat and Bahler, 2010, Mutz et al., 2013). RNA-seq has a clear advantage over its predecessors as it holds the benefit of being more practical (inexpensive, rapid, high-throughput) while at the same time

offering higher sensitivity, quantitative results and the option of different types of analyses (Marioni et al., 2008, Mutz et al., 2013).

Published reviews give a comprehensive description of the steps that are required in order to complete an RNA sequencing experiment (Wang et al., 2009, Mutz et al., 2013). The three most important segments include, firstly, the RNA isolation and library preparation; secondly, the sequencing of the samples by a specialized platform and, finally, the bioinformatics analysis of the output data (figure 4-1).

Initially for the library preparation, the isolated RNA samples are fragmented and converted to cDNA. After ligation to NGS-specific adapters, the samples can be amplified by PCR prior to sequencing on a high-throughput platform such as Illumina. The sequencing event generates millions of short reads from both ends of each cDNA fragment (for the "paired-end" method).

As discussed by Voelkerding et al. (2009), Liu et al. (2012) and (Fox et al., 2014), the main companies that supply NGS platforms presently are Illumina, Roche, PacBio and Life Sciences. Despite engineering differences and variations in the chemical compositions of consumables, they follow a comparable paradigm, in that short and clonally amplified DNA is sequenced in a massively paralleled manner.

The datasets that are generated in an RNA-seq experiment are extremely large and complex and the choice of the appropriate analysis methods is crucial to correctly interpret the data. Although several methods and variations have been devised and a number of software tools have been created to accommodate them, the main steps involved in the procedure remain the same.

The raw sequencing data come in the form of FASTQ files that include the nucleotide sequence of each short read and quality score for each position. These reads are mapped on the

reference transcriptome and using the appropriate tools, the mapped data are counted and normalized and the gene expression is measured, leading to a list of genes with associated p-values and fold changes. The majority of studies using the RNA-seq approach aim to explore the differential expression (DE) of genes between wild type and mutant or untreated and treated conditions. These lists can then be studied in parallel with ChIP-seq or other proteomics data to get a biological insight on the results (Oshlack et al., 2010, Marguerat and Bahler, 2010).

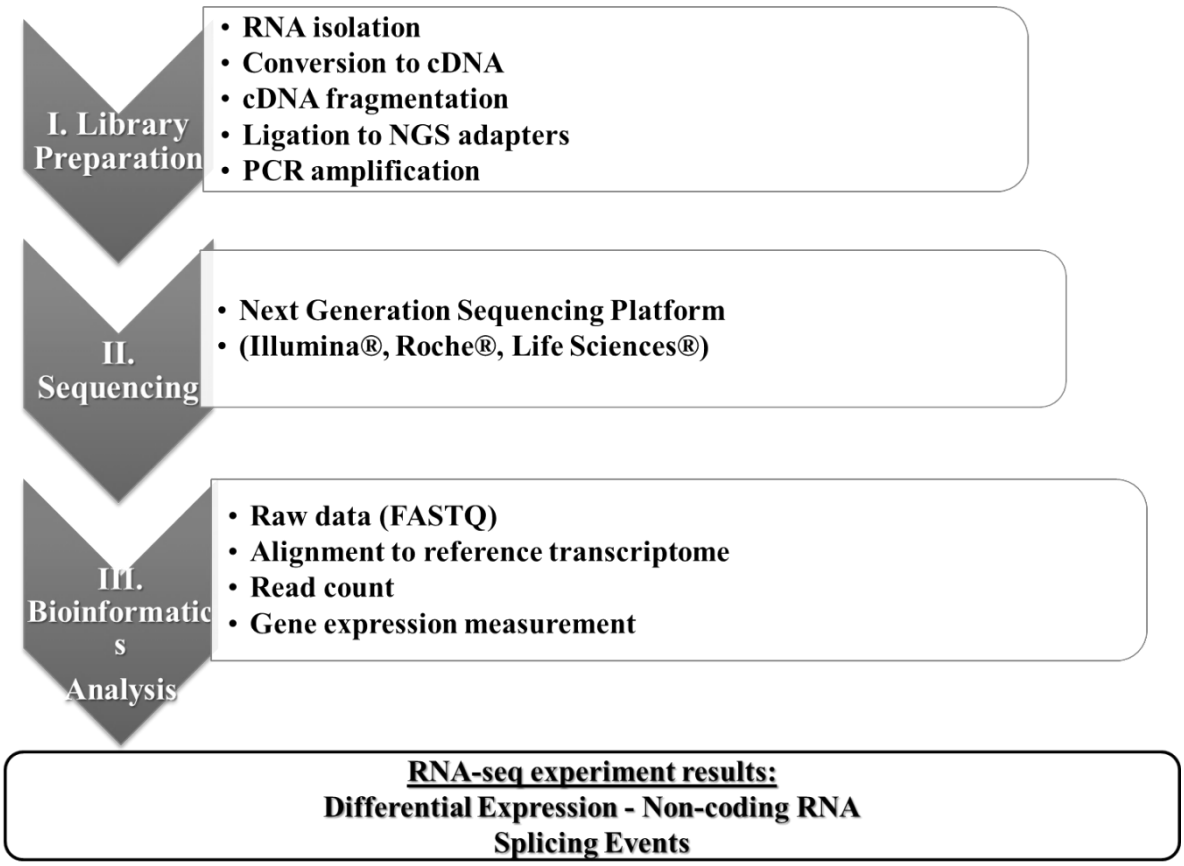


Figure 4-1 Overview of the steps involved in an RNA sequence experiment
An RNA-seq experiment consists of three major steps. Firstly, the library preparation from isolated purified RNA. Then, the sequencing takes place with aid from specialized sequencing platforms and lastly, the raw output data is analyzed with bioinformatics tools. The analysis can focus on events such as differential expression, non-coding RNA and splicing events.

Differential expression between two biological systems has been the main focus of most RNA-seq studies as it offers the opportunity to gain further insight on the interconnection between the transcriptome and phenotype. However, there is so much more information that could be mined from a specially designed RNA-seq experiment, including transcript splicing, epigenetic modifications, chromatin structure variation and evolutionary development (Voelkerding et al., 2009). RNA-seq has emerged also as a means of detecting molecular mutations or mutation patterns that could be responsible for initiating cancer or other diseases (Costa et al., 2013, Wei et al., 2014).

Recent explorations of the genome using modern technologies have advanced our knowledge and understanding of non-coding RNA (ncRNA) and its biological role. NcRNA includes sequences of transcribed RNA that are not translated into proteins and their function remains largely unspecified. These transcripts can be divided in two categories based on an arbitrary size threshold (200 nucleotides), with the shorter ones, especially miRNA, capturing most of the interest due to its implication in numerous biological processes. However, long non-coding RNA is also emerging as an important factor in gene expression and more (Huntzinger and Izaurralde, 2011, Nagano and Fraser, 2011).

4.2 Experimental Aims

The aim of this chapter is to explore the gene expression profile of 226LDM cells and to compare the differences in expression between two groups, namely control 226LDM cells and cell cycle arrested cells. The primary aim is to obtain a list of the ranked genes whose expression is affected by the cell-cycle cessation. The secondary aim of this Chapter is to record the long non-coding RNA sequences that are affected by this change and which could be later on exploited in further studies.

To block cell cycle, the 226LDM cells will be treated as described in chapter 2, with the addition of hydroxyurea (Hu) and nocodazole (No). RNA will be then isolated from both groups of cells (i.e. control and treated) and, after the quality control, the RNA samples in triplicates- will be sequenced at the University College London (UCL).

The bioinformatics analysis of the raw data will be conducted using a high performance computer (HPC) and the Galaxy server, which is a bioinformatics software management system in a click-and-point environment. Following the analysis pipeline, the raw output files will be aligned and mapped to the human genome, the read counts will be normalized and the DESEQ package will be employed to measure differential expression (DE) (Anders and Huber, 2010). From the obtained results, the top up-regulated and down-regulated genes will be identified and Gene Ontology analysis on them will be performed. Furthermore, the analysis will be repeated focusing on the long non-coding RNA transcripts.

This DE study is interesting on its own due to the richness of data that comes with it as well as the fact that, to our knowledge, 226LDM cells have never been used in RNA-seq experiments previously. In the context of the current study, these data will be linked to the ChIP-seq data described in chapter 3.

4.3 Results

4.3.1 Cell cycle blocking treatment

In these experiments, 226LDM cells were initially grown in culture flasks and treated with 100 mM hydroxyurea and 500ng / ml nocodazole, as described in chapter 2. After treatment, the detached treated cells as well as control 226LDM cells were harvested for sample preparation. Three biological replicates were prepared for each condition to ensure the reliability of the results interpretation. The success of the treatment for each replicate was confirmed by western blotting where control cells exhibited expression of both CTCF isoforms, while all treated samples showed no presence of CTCF130 (figure 4-2).

4.3.2 RNA isolation and quality control

The isolation of total RNA from the 6 samples was done following the protocol from Trizol®. The samples were run on a 1% agarose gel stained with ethidium bromide (EtBr) in order to test the quality of the RNA. Two distinct bands representing the 28S and 18S rRNAs were visible indicating that the integrity of the RNA was high (figure 4-3).

The concentration and purity of the samples were measured with the NanoDrop ND-1000 spectrophotometer. The 260/280 (RNA/ protein) absorbance ratios varied between 1.86 and 1.96 while they were between 1.64 and 1.89 for the 260/230 (RNA/ contaminants) (see table 4-1 for values). Subsequently, integrity control was performed using the Agilent 2100 Bioanalyzer where clear and distinct 18S and 28S bands representing ribosomal RNA were present in all the samples (figure 4-4). Upon completion of the quality control the samples were delivered to UCL for library preparation and sequencing.

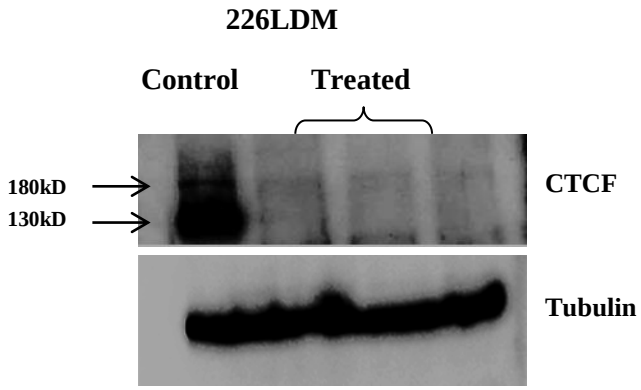


Figure 4-2 Western blotting using lysates from 226LDM cells after treatment with Hydroxyurea and Nocodazole

226LDM cells were treated with 100 mM hydroxyurea for 24 hours and then 500 ng /ml nocodazole for another 24 hours. Following treatment the detached cells were harvested and used to prepare samples for western blotting. A polyclonal antibody recognizing both CTCF isoforms was used for the experiment. Each isoform can be seen on the image with arrows. The visualization of the signal was done with use of the Fusion FX7 and tubulin was used as a loading control.

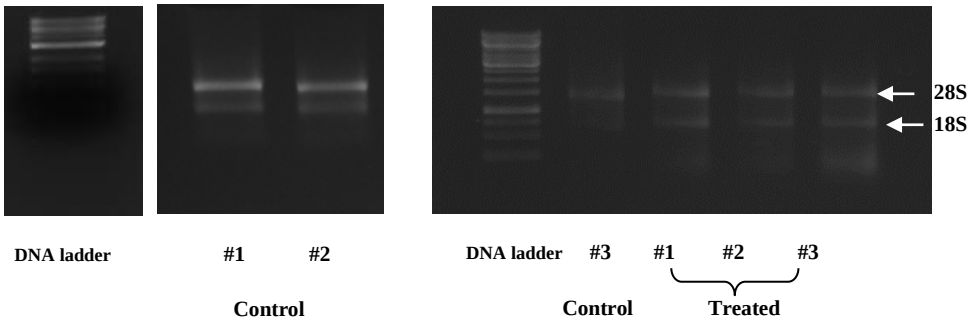


Figure 4-3 Assessment of RNA integrity on samples prepared from 226LDM control and treated cells

Total RNA extracted from control and treated 226LDM cells, both prepared in triplicates, was run on an agarose gels stained with ethidium bromide (EtBr) in order to assess its integrity. The bands representing the 28S and 18S ribosomal RNAs are shown with arrows.

Table 4-1 Concentrations and absorbance ratios of RNA samples measured with the NanoDrop ND-1000.

Sample Number	Concentration	Provided to UCL (in 20µl of H_2O)	260/280	260/230
1	163.8 ng/µl	1 µg	1.91	1.82
2	164.7 ng/µl	1 µg	1.96	1.88
3	117.8 ng/µl	1.5 µg	1.95	1.69
4	63.4 ng/µl	1 µg	1.86	1.64
5	173.2 ng/µl	1.5 µg	1.89	1.91
6	131.6 ng/µl	1.5 µg	1.95	1.82

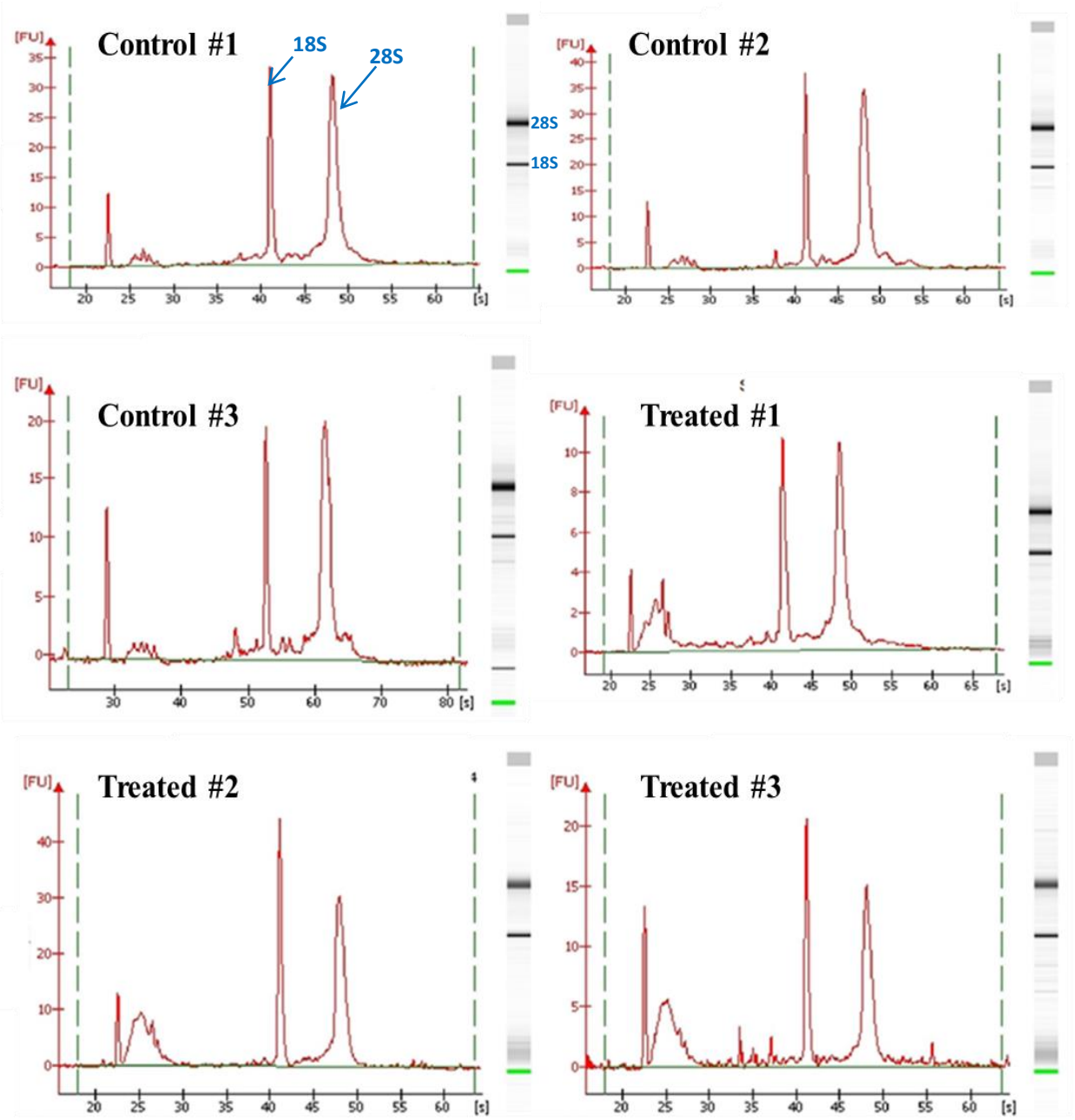


Figure 4-4 Quality control of RNA isolated from control and treated 226LDM cells

Total RNA was isolated from control and treated 226LDM cells, each group in triplicates. To confirm the integrity of the content, the samples were run on a microfluidic chip using the Agilent Bioanalyzer 2100. The 28S and 18S ribosomal subunits are represented by peaks on the electropherograph and by bands on the gel placed on the right side of each graph.

4.3.3 Sequencing data analysis

The sequencing of the samples took place at UCL using the Illumina Hiseq platform. The raw sequence files were received in FASTQ format which is the most common format used in NGS (Cock et al., 2010). It contains the output reads from the sequencer which need to be mapped in the reference genome which in this study is the human genome.

Initially, the FASTQ quality trimmer was used to discard the low quality reads that did not pass the quality threshold and then the reads were aligned to the human genome using the robust aligning tool Novoalign (Ruffalo et al., 2011) and a high performance computer (HPC). The sequence aligning data were stored in BAM file format using SAMtools (Li et al., 2009).

The next step of the procedure involved the raw reads count (RPKM) and normalization of the reads (FPKM) using the Bam2fpkm tool in galaxy. Multi-dimensional scaling (MDS) (Borg and Groenen, 2005) was employed to generate plots from both the raw and normalized data in order to assess the level of similarity between samples (figure 4-5). The samples of each group appeared clustered together (figure 4.5.A) and the in the normalized data they are grouped even closer which fits to the expectation from the experimental design (figure 4.5.B).

Using the raw count reads, the differential gene expression between the two sample groups was investigated using the DESEQ tool kit. With this package, diagnostics plots were created to visualize the characteristics of the analysis, such as a histogram portraying the distribution of *p* values (figure 4-6) and two heatmaps; one to show the expression variance of the 100 most differentially expressed genes by *p* value and one portraying the similarities between the samples according to Euclidean distance (figure 4-7). Furthermore, a volcano plot was generated with MATLAB combining the magnitude of expression change (log2 fold change) with the significance (*q* value) (figure 4-8).

The genes whose transcription was affected by the treatment were discovered and analysed. From the set of results, the data correlated with a p-value higher than 0.05 were discarded as potentially not-significant and the remaining entries were sorted according to fold-change of expression. The list was further filtered by *q* value (<0.4). The top 50 transcripts (at this point, different transcripts of the same gene represent will represent a different entry), as well as the top 50 genes that were up~ and down-regulated are shown in tables 4.2-4.3 and 4.4-4.5 respectively. Unless referenced otherwise on the table, the gene descriptions were provided by the RefSeq Gene project (Pruitt et al., 2005, Pruitt et al., 2014). In these tables the gene symbol along with the log2 fold change and *q* value can be seen.

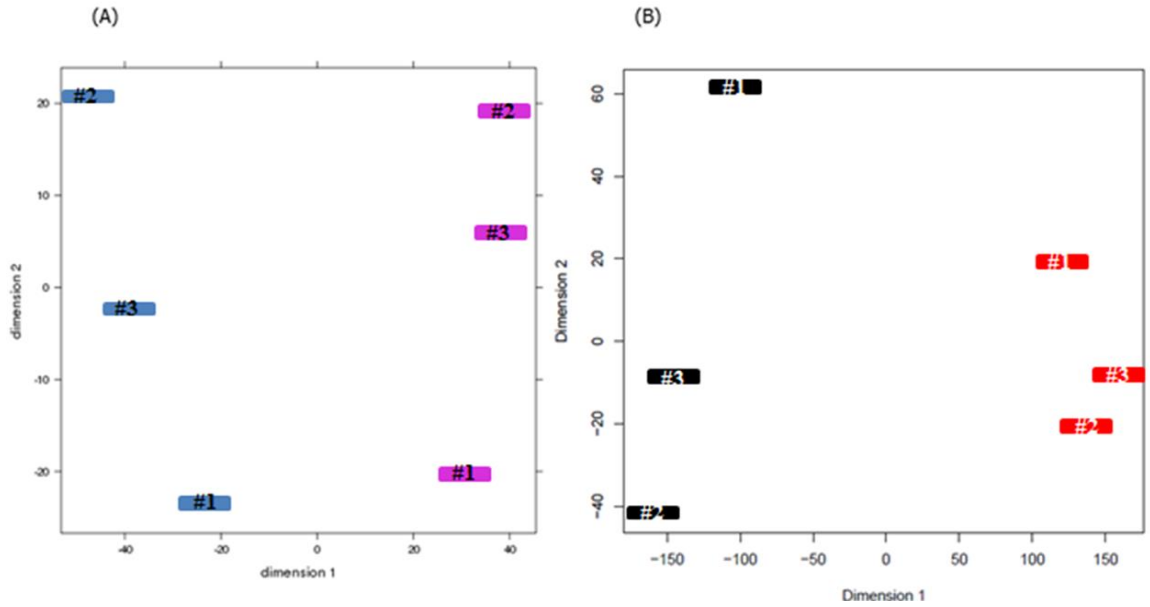


Figure 4-5 MDS plots of raw and normalized count data showing the similarities between control and treated samples

Multidimensional scaling (MDS) plots were generated from the raw (A) and normalized (B) count reads using the DESEQ analysis package on Galaxy. MDS plots are used to visualize the level of similarity between samples of a dataset. In the raw counts plot the control samples are shown in blue color and the treated samples are shown in pink. In the MDS plot of normalized counts the control samples are shown in black color while the treated samples are shown in red.

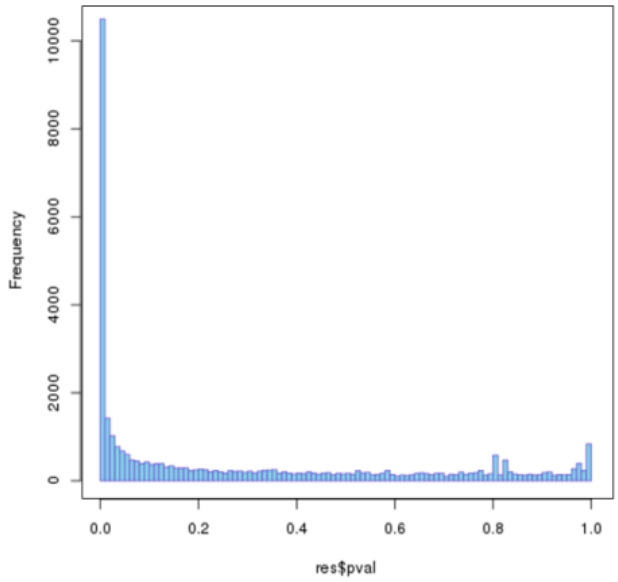


Figure 4-6 Histogram of P values generated from the DESEQ package

A histogram of P values was generated by the DESEQ package to assess the quality of the samples. The histogram is used to visualize the distribution of p-values within certain intervals and to assess whether problems are present with the analysis.

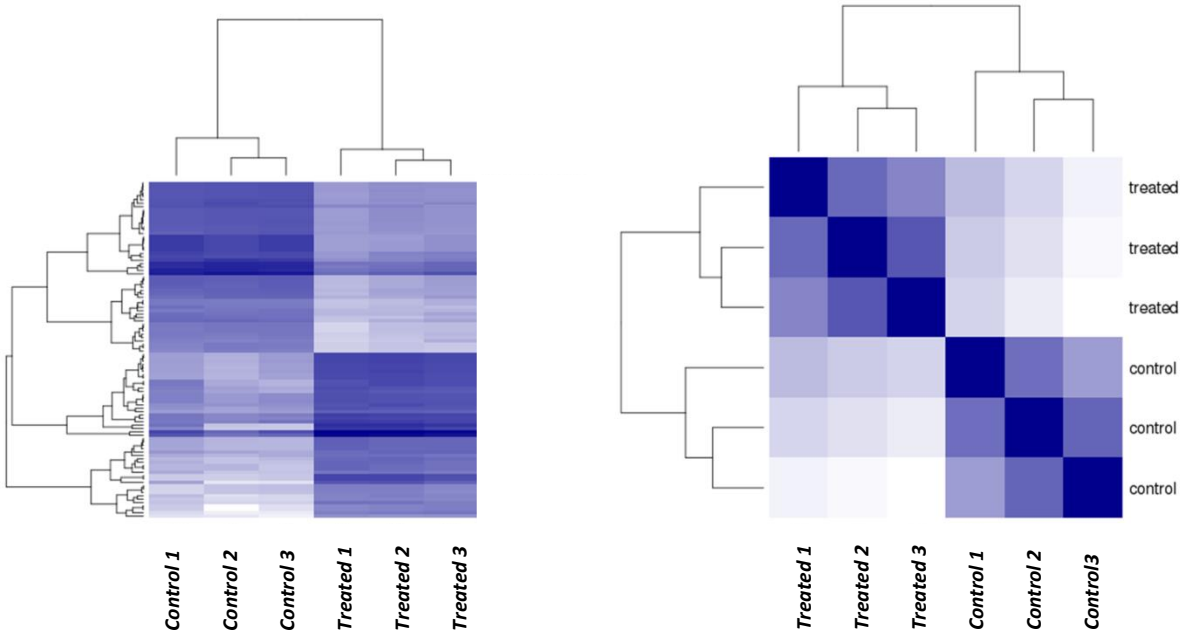


Figure 4-7 Heatmaps portraying the differential expression of the most affected genes and the clustering between the control and treated samples

The DESEQ package was employed to generate heatmaps, which are two-dimensional grids depicting data from a count table. Left: A gene expression heatmap shows the expression variance between control and treated genes for the top 100 genes/transcripts affected by the treatment and sorted by *p* value. Right: The Euclidean distances between the samples as calculated from the variance-stabilizing transformation of the count data can be seen in the sample heatmap. The dendrogram at the sides demonstrate the overall similarity between samples.

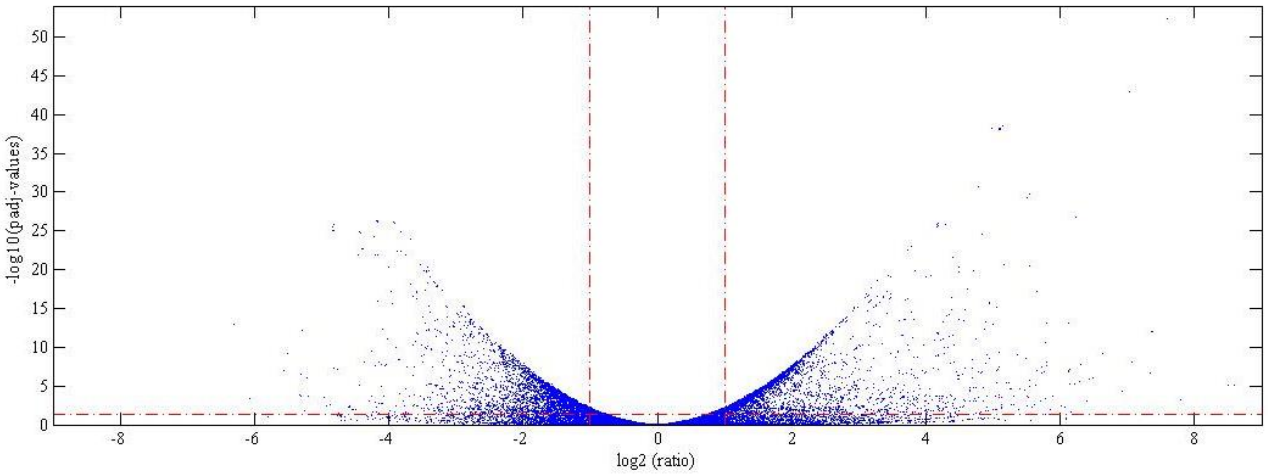


Figure 4-8 Volcano plot summarizing the significant and highly deregulated genes

In this scatterplot, genes that are highly deregulated will appear to the upper left and right on the outside of the pair of threshold vertical red dotted lines. Significant changes according to Q value are shown towards the higher top of the plot above the horizontal red dotted line. The graph was produced using MATLAB.

Table 4-2 Top 50 up-regulated gene transcripts, sorted according to fold change, as a result of cell-cycle block treatment in 226LDM cells, discovered with the DESEQ package

Transcript ID	Gene ID	log2 FoldChange	Q-value
1. KRT23:NM_015515	KRT23	8.599638	7.85E-06
2. KRT23:NM_001282433	KRT23	8.495657	7.75E-06
3. CHP2:NM_022097	CHP2	7.78822	0.000613
4. C15orf48:NM_032413	C15orf48	7.600153	5.03E-53
5. C15orf48:NM_197955	C15orf48	7.592132	5.03E-53
6. MUC4:NM_004532	MUC4	7.36881	9.71E-13
7. MUC4:NM_138297	MUC4	7.362573	1.10E-12
8. MUC4:NM_018406	MUC4	7.335745	1.78E-07
9. S100A7A:NM_176823	S100A7A	7.073195	7.06E-09
10. KRT34:NM_021013	KRT34	7.027633	1.18E-43
11. CKB:NM_001823	CKB	6.923193	5.60E-05
12. KPNA7:NM_001145715	KPNA7	6.631406	6.36E-10
13. GNGT2:NM_001198755	GNGT2	6.485113	8.02E-07
14. GNGT2:NM_031498	GNGT2	6.485113	8.02E-07
15. GNGT2:NM_001198754	GNGT2	6.433824	1.29E-06
16. GNGT2:NM_001198756	GNGT2	6.424219	1.47E-06
17. CIB3:NM_054113	CIB3	6.398279	0.001035
18. LCN2:NM_005564	LCN2	6.31412	1.78E-08
19. SPDEF:NM_001252294	SPDEF	6.289303	3.12E-06
20. CRCT1:NM_019060	CRCT1	6.230484	1.51E-27
21. CNFN:NM_032488	CNFN	6.213195	5.81E-06
22. DES:NM_001927	DES	6.167984	0.026839
23. SPDEF:NM_012391	SPDEF	6.150613	4.85E-06
24. IGF2:NM_001291861	IGF2	6.140172	6.55E-08
25. IGF2:NM_000612	IGF2	6.127075	1.69E-07
26. S100A7:NM_002963	S100A7	6.122937	8.40E-14
27. IGF2:NM_001127598	IGF2	6.117659	1.09E-07
28. IGF2:NM_001291862	IGF2	6.117659	1.09E-07
29. IGF2:NM_001007139	IGF2	6.117659	1.09E-07
30. CIB3:NM_001300922	CIB3	6.077545	0.004625
31. MUC16:NM_024690	MUC16	6.010402	0.003392
32. RAB33A:NM_004794	RAB33A	5.89684	0.000178
33. LOC388282:NM_001278081	LOC388282	5.843758	0.011923
34. MUC20:NM_001291833	MUC20	5.803648	7.42E-09
35. MUC20:NM_020790	MUC20	5.799892	2.13E-08
36. S100A12:NM_005621	S100A12	5.792389	7.50E-14
37. MGARP:NM_032623	MGARP	5.784396	2.95E-11
38. TCHHL1:NM_001008536	TCHHL1	5.772711	0.000382
39. SPANXB1:NM_032461	SPANXB1	5.765314	1.64E-07

40. SPANXB1.1:NM_032461.1	SPANXB1.1	5.733483	2.43E-07
41. MUC20:NM_152673	MUC20	5.730657	3.34E-08
42. LYPD2:NM_205545	LYPD2	5.709685	0.000902
43. MUC20:NM_001282506	MUC20	5.658244	1.01E-07
44. SPRR2A:NM_005988	SPRR2A	5.648773	7.49E-18
45. PRAP1:NM_001145201	PRAP1	5.638928	8.83E-06
46. DEFB103B:NM_018661	DEFB103B	5.622059	0.002648
47. DEFB103A:NM_001081551	DEFB103A	5.622059	0.002648
48. RASGEF1C:NM_175062	RASGEF1C	5.571098	0.004904
49. CARD17:NM_001007232	CARD17	5.549296	1.83E-06
50. SPRR2B:NM_001017418	SPRR2B	5.54848	3.31E-21

Table 4-3. Top 50 up-regulated genes (and their descriptions) due to treatment of 226LDM cells with HU and NO resulting in cell-cycle block. The DESEQ package was used to generate this data

Gene ID	Description
1. KRT23	The protein encoded by this gene is a member of the keratin family. The keratins are intermediate filament proteins responsible for the structural integrity of epithelial cells and are subdivided into cytokeratins and hair keratins. The type I cytokeratins consist of acidic proteins which are arranged in pairs of heterotypic keratin chains. The type I cytokeratin genes are clustered in a region of chromosome 17q12-q21. Alternative splicing results in multiple transcript variants.
2. CHP2	This gene product is a small calcium-binding protein that regulates cell pH by controlling plasma membrane-type Na+/H+ exchange activity. This protein shares sequence similarity with calcineurin B and can bind to and stimulate the protein phosphatase activity of calcineurin A (CnA) and functions in the calcineurin/NFAT (nuclear factor of activated T cells) signaling pathway. Another member of the CHP subfamily, Calcineurin B homologous protein 1, is located on Chromosome 15 and is an inhibitor of calcineurin activity and has a genetic phenotype associated with Parkinson's Disease (OMIM:606988). This gene was initially identified as a tumor-associated antigen and was previously referred to as Hepatocellular carcinoma-associated antigen 520.
3. C15orf48	This gene was first identified in a study of human esophageal squamous cell carcinoma tissues. Levels of both the message and protein are reduced in carcinoma samples. In adult human tissues, this gene is expressed in the the esophagus, stomach, small intestine, colon and placenta. Alternatively spliced transcript variants that encode the same protein have been identified.
4. MUC4	The major constituents of mucus, the viscous secretion that covers epithelial surfaces such as those in the trachea, colon, and cervix, are highly glycosylated proteins called mucins. These glycoproteins play important roles in the protection of the epithelial cells and have been implicated in epithelial renewal and differentiation. This gene encodes an integral membrane glycoprotein found on the cell surface, although secreted isoforms may exist. At least two dozen transcript variants of this gene have been found, although for many of them the full-length transcript has not been determined or they are found only in tumor tissues. This gene contains a region in the coding sequence which has a variable number (>100) of 48 nt tandem repeats.
5. S100A7A	-No known function-
6. KRT34	The protein encoded by this gene is a member of the keratin gene family. As a type I hair keratin, it is an acidic protein which heterodimerizes with type II keratins to form hair and nails. The type I hair keratins are clustered in a region of chromosome 17q12-q21 and have the same direction of transcription.
7. CKB	The protein encoded by this gene is a cytoplasmic enzyme involved in energy homeostasis. The encoded protein reversibly catalyzes the transfer of phosphate between ATP and various phosphogens such as creatine phosphate. It acts as a homodimer in brain as well as in other tissues, and as a heterodimer with a similar muscle isozyme in heart. The encoded protein is a member of the ATP:guanido phosphotransferase protein family. A pseudogene of this gene has been characterized.
8.	-No known function-
9. GNGT2	Phototransduction in rod and cone photoreceptors is regulated by groups of signaling proteins. The encoded protein is thought to play a crucial role in cone phototransduction. It belongs to the G protein gamma family and localized specifically in cones. Several transcript variants encoding the same protein have been found for this gene.
10. CIB3	This gene product shares a high degree of sequence similarity with DNA-dependent protein kinase catalytic subunit-interacting protein 2 in human and mouse, and like them may bind the catalytic subunit of DNA-dependent protein kinases. The exact function of this gene is not known. Alternative splicing results in multiple transcript variants.
11. LCN2	-No known function-
12. SPDEF	The protein encoded by this gene belongs to the ETS family of transcription factors. It is highly expressed in the prostate epithelial cells, and functions as an androgen-independent transactivator of prostate-specific antigen (PSA) promoter. Higher expression of this protein

	has also been reported in brain, breast, lung and ovarian tumors, compared to the corresponding normal tissues, and it shows better tumor-association than other cancer-associated molecules, making it a more suitable target for developing specific cancer therapies. Alternatively spliced transcript variants encoding different isoforms have been found for this gene.
13. DES	This gene encodes a muscle-specific class III intermediate filament. Homopolymers of this protein form a stable intracytoplasmic filamentous network connecting myofibrils to each other and to the plasma membrane. Mutations in this gene are associated with desmin-related myopathy, a familial cardiac and skeletal myopathy (CSM), and with distal myopathies.
14. IGF2	This gene encodes a member of the insulin family of polypeptide growth factors, which are involved in development and growth. It is an imprinted gene, expressed only from the paternal allele, and epigenetic changes at this locus are associated with Wilms tumour, Beckwith-Wiedemann syndrome, rhabdomyosarcoma, and Silver-Russell syndrome. A read-through INS-IGF2 gene exists, whose 5' region overlaps the INS gene and the 3' region overlaps this gene. Alternatively spliced transcript variants encoding different isoforms have been found for this gene.
15. S100A7	The protein encoded by this gene is a member of the S100 family of proteins containing 2 EF-hand calcium-binding motifs. S100 proteins are localized in the cytoplasm and/or nucleus of a wide range of cells, and involved in the regulation of a number of cellular processes such as cell cycle progression and differentiation. S100 genes include at least 13 members which are located as a cluster on chromosome 1q21. This protein differs from the other S100 proteins of known structure in its lack of calcium binding ability in one EF-hand at the N-terminus. The protein is overexpressed in hyperproliferative skin diseases, exhibits antimicrobial activities against bacteria and induces immunomodulatory activities.
16. MUC16	-No known function-
17. RAB33A	The protein encoded by this gene belongs to the small GTPase superfamily, Rab family. It is GTP-binding protein and may be involved in vesicle transport.
18. MUC20	This gene encodes a member of the mucin protein family. Mucins are high molecular weight glycoproteins secreted by many epithelial tissues to form an insoluble mucous barrier. The C-terminus of this family member associates with the multifunctional docking site of the MET proto-oncogene and suppresses activation of some downstream MET signaling cascades. The protein features a mucin tandem repeat domain that varies between two and six copies in most individuals. Multiple variants encoding different isoforms have been found for this gene. A related pseudogene, which is also located on chromosome 3, has been identified.
19. S100A12	The protein encoded by this gene is a member of the S100 family of proteins containing 2 EF-hand calcium-binding motifs. S100 proteins are localized in the cytoplasm and/or nucleus of a wide range of cells, and involved in the regulation of a number of cellular processes such as cell cycle progression and differentiation. S100 genes include at least 13 members which are located as a cluster on chromosome 1q21. This protein is proposed to be involved in specific calcium-dependent signal transduction pathways and its regulatory effect on cytoskeletal components may modulate various neutrophil activities. The protein includes an antimicrobial peptide which has antibacterial activity.
20. TCHHL1	This gene belongs to the S100 fused-type protein (SFTP) gene family, and is located in a cluster of SFTP genes on chromosome 1q21. Several members of this family have been implicated in the development of complex skin disorders. This gene is evolutionarily conserved; its expression appears to be hair-specific and spatially restricted within the distal inner root sheath of the hair follicle. It thus may have an important role in hair morphogenesis.
21. SPANXB1	Temporally regulated transcription and translation of several testis-specific genes is required to initiate the series of molecular and morphological changes in the male germ cell lineage necessary for the formation of mature spermatozoa. This gene is a member of the SPANX family of cancer/testis-associated genes, which are located in a cluster on chromosome X. The SPANX genes encode differentially expressed testis-specific proteins that localize to various subcellular compartments. This particular family member contains an additional 18 nucleotides in its coding region compared to the other family members in the same gene cluster. This family member is also subject to gene copy number variation. Although the protein encoded by this gene contains consensus nuclear localization signals, the major site for subcellular localization of expressed protein is in the cytoplasmic droplets of ejaculated spermatozoa. This protein provides a biochemical marker for studying the unique structures in spermatazoa, while attempting to further define its role in spermatogenesis.

22. SPRR2A	-No known function-
23. PRAP1	-No known function-
24. DEFB103B	Defensins form a family of microbicidal and cytotoxic peptides made by neutrophils. Members of the defensin family are highly similar in protein sequence. This gene encodes defensin, beta 103, which has broad spectrum antimicrobial activity and may play an important role in innate epithelial defense.
25. DEFB103A	Defensins form a family of microbicidal and cytotoxic peptides made by neutrophils. Members of the defensin family are highly similar in protein sequence. This gene encodes defensin, beta 103, an antibiotic peptide which is induced by bacteria and interferon gamma, and which displays antimicrobial activity against S. aureus, S. pyogenes, P. aeruginosa, E. coli, and C. albicans.
26. BCL2A1	This gene encodes a member of the BCL-2 protein family. The proteins of this family form hetero- or homodimers and act as anti- and pro-apoptotic regulators that are involved in a wide variety of cellular activities such as embryonic development, homeostasis and tumorigenesis. The protein encoded by this gene is able to reduce the release of pro-apoptotic cytochrome c from mitochondria and block caspase activation. This gene is a direct transcription target of NF-kappa B in response to inflammatory mediators, and is up-regulated by different extracellular signals, such as granulocyte-macrophage colony-stimulating factor (GM-CSF), CD40, phorbol ester and inflammatory cytokine TNF and IL-1, which suggests a cytoprotective function that is essential for lymphocyte activation as well as cell survival. Alternatively spliced transcript variants encoding different isoforms have been found for this gene.
27. VNN3	This gene is the central gene in a cluster of three vanin genes on chromosome 6q23-q24. Extensive alternative splicing has been described; the two most common variants are represented as RefSeqs.
28. FXYD2	This gene encodes a member of the FXYD family of transmembrane proteins. This particular protein encodes the sodium/potassium-transporting ATPase subunit gamma. Mutations in this gene have been associated with Renal Hypomagnesemia-2. Alternatively spliced transcript variants have been described. Read-through transcripts have been observed between this locus and the upstream FXYD domain-containing ion transport regulator 6 (FXYD6, GeneID 53826) locus.
29. TMEM71	-No known function-
30. KRT75	This gene is a member of the type II keratin family clustered on the long arm of chromosome 12. Type I and type II keratins heteropolymerize to form intermediate-sized filaments in the cytoplasm of epithelial cells. This gene is expressed in the companion layer, upper germinative matrix region of the hair follicle, and medulla of the hair shaft. The encoded protein plays an essential role in hair and nail formation. Variations in this gene have been associated with the hair disorders pseudofolliculitis barbae (PFB) and loose anagen hair syndrome (LAHS).
31. VTCN1	This gene encodes a protein belonging to the B7 costimulatory protein family. Proteins in this family are present on the surface of antigen-presenting cells and interact with ligand bound to receptors on the surface of T cells. Studies have shown that high levels of the encoded protein has been correlated with tumor progression. A pseudogene of this gene is located on chromosome 20. Multiple transcript variants encoding different isoforms have been found for this gene.
32. MYL9	Myosin, a structural component of muscle, consists of two heavy chains and four light chains. The protein encoded by this gene is a myosin light chain that may regulate muscle contraction by modulating the ATPase activity of myosin heads. The encoded protein binds calcium and is activated by myosin light chain kinase. Two transcript variants encoding different isoforms have been found for this gene.
33. LAMA4	Laminins, a family of extracellular matrix glycoproteins, are the major noncollagenous constituent of basement membranes. They have been implicated in a wide variety of biological processes including cell adhesion, differentiation, migration, signaling, neurite outgrowth and metastasis. Laminins are composed of 3 non identical chains: laminin alpha, beta and gamma (formerly A, B1, and B2, respectively) and they form a cruciform structure consisting of 3 short arms, each formed by a different chain, and a long arm composed of all 3 chains. Each laminin chain is a multidomain protein encoded by a distinct gene. Several isoforms of each chain have been described. Different alpha, beta and gamma chain isomers combine to give rise to different heterotrimeric laminin isoforms which are designated by

	Arabic numerals in the order of their discovery, i.e. alpha1beta1gamma1 heterotrimer is laminin 1. The biological functions of the different chains and trimer molecules are largely unknown, but some of the chains have been shown to differ with respect to their tissue distribution, presumably reflecting diverse functions in vivo. This gene encodes the alpha chain isoform laminin, alpha 4. The domain structure of alpha 4 is similar to that of alpha 3, both of which resemble truncated versions of alpha 1 and alpha 2, in that approximately 1,200 residues at the N-terminus (domains IV, V and VI) have been lost. Laminin, alpha 4 contains the C-terminal G domain which distinguishes all alpha chains from the beta and gamma chains. The RNA analysis from adult and fetal tissues revealed developmental regulation of expression, however, the exact function of laminin, alpha 4 is not known. Tissue-specific utilization of alternative polyA-signal has been described in literature. Alternative splicing results in multiple transcript variants encoding distinct isoforms.
34. UPK2	This gene encodes one of the proteins of the highly conserved urothelium-specific integral membrane proteins of the asymmetric unit membrane which forms urothelium apical plaques in mammals. The asymmetric unit membrane is believed to strengthen the urothelium by preventing cell rupture during bladder distention. The encoded protein is expressed in the peripheral blood of bladder cancer patients with transitional cell carcinomas.
35. IL32	This gene encodes a member of the cytokine family. The protein contains a tyrosine sulfation site, 3 potential N-myristoylation sites, multiple putative phosphorylation sites, and an RGD cell-attachment sequence. Expression of this protein is increased after the activation of T-cells by mitogens or the activation of NK cells by IL-2. This protein induces the production of TNFalpha from macrophage cells. Alternate transcriptional splice variants, encoding different isoforms, have been characterized.
36. CLIC3	Chloride channels are a diverse group of proteins that regulate fundamental cellular processes including stabilization of cell membrane potential, transepithelial transport, maintenance of intracellular pH, and regulation of cell volume. Chloride intracellular channel 3 is a member of the p64 family and is predominantly localized in the nucleus and stimulates chloride ion channel activity. In addition, this protein may participate in cellular growth control, based on its association with ERK7, a member of the MAP kinase family.
37. MT1M	This gene encodes a member of the metallothionein superfamily, type 1 family. Metallothioneins have a high content of cysteine residues that bind various heavy metals. These genes are transcriptionally regulated by both heavy metals and glucocorticoids.
38. CDH5	This gene is a classical cadherin from the cadherin superfamily and is located in a six-cadherin cluster in a region on the long arm of chromosome 16 that is involved in loss of heterozygosity events in breast and prostate cancer. The encoded protein is a calcium-dependent cell-cell adhesion glycoprotein comprised of five extracellular cadherin repeats, a transmembrane region and a highly conserved cytoplasmic tail. Functioning as a classic cadherin by imparting to cells the ability to adhere in a homophilic manner, the protein may play an important role in endothelial cell biology through control of the cohesion and organization of the intercellular junctions. An alternative splice variant has been described but its full length sequence has not been determined.
39. CCL28	This antimicrobial gene belongs to the subfamily of small cytokine CC genes. Cytokines are a family of secreted proteins involved in immunoregulatory and inflammatory processes. The CC cytokines are proteins characterized by two adjacent cysteines. The cytokine encoded by this gene displays chemotactic activity for resting CD4 or CD8 T cells and eosinophils. The product of this gene binds to chemokine receptors CCR3 and CCR10. This chemokine may play a role in the physiology of extracutaneous epithelial tissues, including diverse mucosal organs. Multiple transcript variants encoding two different isoforms have been found for this gene.
40. KRT8	This gene is a member of the type II keratin family clustered on the long arm of chromosome 12. Type I and type II keratins heteropolymerize to form intermediate-sized filaments in the cytoplasm of epithelial cells. The product of this gene typically dimerizes with keratin 18 to form an intermediate filament in simple single-layered epithelial cells. This protein plays a role in maintaining cellular structural integrity and also functions in signal transduction and cellular differentiation. Mutations in this gene cause cryptogenic cirrhosis. Alternatively spliced transcript variants have been found for this gene.
41. FXYD6-FXYD2	This locus represents naturally occurring read-through transcription between the neighboring FXYD domain-containing ion transport regulator 6 (GeneID 53826) and sodium/potassium-transporting ATPase subunit gamma (GeneID 486) genes on chromosome 11. One read-through transcript produces a fusion protein that shares sequence identity with each individual

	gene product, while another read-through transcript encodes a protein that has a distinct C-terminus and only shares sequence identity with the upstream locus (GeneID 53826).
42. COL15A1	This gene encodes the alpha chain of type XV collagen, a member of the FACIT collagen family (fibril-associated collagens with interrupted helices). Type XV collagen has a wide tissue distribution but the strongest expression is localized to basement membrane zones so it may function to adhere basement membranes to underlying connective tissue stroma. The proteolytically produced C-terminal fragment of type XV collagen is restin, a potentially antiangiogenic protein that is closely related to endostatin. Mouse studies have shown that collagen XV deficiency is associated with muscle and microvessel deterioration.
43. SPRR2E	This gene encodes a member of a family of small proline-rich proteins clustered in the epidermal differentiation complex on chromosome 1q21. The encoded protein, along with other family members, is a component of the cornified cell envelope that forms beneath the plasma membrane in terminally differentiated stratified squamous epithelia. This envelope serves as a barrier against extracellular and environmental factors. The seven SPRR2 genes (A-G) appear to have been homogenized by gene conversion compared to others in the cluster that exhibit greater differences in protein structure.
44. C2orf72	-No known function-
45. KRT33A	This gene encodes a member of the keratin gene family. This gene is one of multiple type I hair keratin genes that are clustered in a region of chromosome 17q12-q21 and have the same direction of transcription. As a type I hair keratin, the encoded protein is an acidic protein which heterodimerizes with type II keratins to form hair and nails. There are two isoforms of this protein, encoded by two separate genes, keratin 33A and keratin 33B.
46. TNNC2	Troponin (Tn), a key protein complex in the regulation of striated muscle contraction, is composed of 3 subunits. The Tn-I subunit inhibits actomyosin ATPase, the Tn-T subunit binds tropomyosin and Tn-C, while the Tn-C subunit binds calcium and overcomes the inhibitory action of the troponin complex on actin filaments. The protein encoded by this gene is the Tn-C subunit.
47. NLRP10	Members of the NALP protein family typically contain a NACHT domain, a NACHT-associated domain (NAD), a C-terminal leucine-rich repeat (LRR) region, and an N-terminal pyrin domain (PYD). The protein encoded by this gene belongs to the NALP protein family despite lacking the LRR region. This protein likely plays a regulatory role in the innate immune system. The protein belongs to the signal-induced multiprotein complex, the inflammasome, that activates the pro-inflammatory caspases, caspase-1 and caspase-5. Other experiments indicate that this gene acts as a multifunctional negative regulator of inflammation and apoptosis.
48. SPRR2D	-No known function-
49. CXorf49	-No known function-
50. CXorf49B	-No known function-

Table 4-4. Top 50 down-regulated gene transcripts, sorted according to fold change, resulting from cell-cycle block treatment in 226LDM cells, discovered with the DESEQ package

Transcript ID	Gene ID	log2 FoldChange	Q-value
1. IFNE:NM_176891	IFNE	-6.30539	1.17E-13
2. KCNJ10:NM_002241	KCNJ10	-6.07528	0.000371
3. CD37:NM_001774	CD37	-5.78961	0.105869
4. CD37:NM_001040031	CD37	-5.78961	0.105869
5. CHST4:NM_005769	CHST4	-5.56221	9.82E-08
6. SBSPON:NM_153225	SBSPON	-5.51535	5.88E-10
7. SLC25A27:NM_004277	SLC25A27	-5.32327	0.000171
8. SLC25A27:NM_001204051	SLC25A27	-5.30305	0.000171
9. CHST4:NM_001166395	CHST4	-5.30287	3.30E-07
10. ADAMTS15:NM_139055	ADAMTS15	-5.29835	6.33E-13
11. PGLYRP1:NM_005091	PGLYRP1	-5.2588	0.013616
12. MARCH1:NM_001166373	MARCH1	-5.21713	9.58E-07
13. MARCH1:NM_017923	MARCH1	-5.21713	9.58E-07
14. SLC25A27:NM_001204052	SLC25A27	-5.19801	0.000195
15. FAM184B:NM_015688	FAM184B	-4.97151	0.000322
16. OCLM:NM_022375	OCLM	-4.92508	0.003207
17. SLITRK5:NM_015567	SLITRK5	-4.86939	0.002609
18. SLC29A1:NM_001078175	SLC29A1	-4.83037	1.05E-25
19. SLC29A1:NM_001078176	SLC29A1	-4.82799	3.14E-26
20. SLC29A1:NM_001078174	SLC29A1	-4.82574	1.49E-26
21. SLC29A1:NM_004955	SLC29A1	-4.82569	1.49E-26
22. SLC29A1:NM_001078177	SLC29A1	-4.82535	8.54E-26
23. SLC16A7:NM_004731	SLC16A7	-4.81848	4.69E-08
24. FRMPD2:NM_001042512	FRMPD2	-4.76662	0.060652
25. SLC16A7:NM_001270623	SLC16A7	-4.75744	7.07E-08
26. SLC16A7:NM_001270622	SLC16A7	-4.75744	7.07E-08
27. HOXD9:NM_014213	HOXD9	-4.74498	0.033306
28. SLC47A2:NM_152908	SLC47A2	-4.73102	0.078496
29. DSG1:NM_001942	DSG1	-4.71024	0.322558
30. SLC47A2:NM_001256663	SLC47A2	-4.70516	0.079138
31. SLC47A2:NM_001099646	SLC47A2	-4.70051	0.079915
32. MXRA5:NM_015419	MXRA5	-4.65616	0.337272
33. SLC34A1:NM_001167579	SLC34A1	-4.64595	0.082399
34. GPRASP1:NM_001184727	GPRASP1	-4.60355	0.00573
35. GPRASP1:NM_014710	GPRASP1	-4.58758	0.006117
36. GPRASP1:NM_001099411	GPRASP1	-4.58758	0.006117
37. GPRASP1:NM_001099410	GPRASP1	-4.58758	0.006117
38. SPATA25:NM_080608	SPATA25	-4.57957	0.003438

39. METTL7A:NM_014033	METTL7A	-4.56485	0.182034
40. CYP3A5:NM_001291829	CYP3A5	-4.53891	0.009357
41. CYP3A5:NM_001291830	CYP3A5	-4.52157	0.010025
42. CYP3A5:NM_000777	CYP3A5	-4.48625	0.011512
43. CHRNG:NM_005199	CHRNG	-4.46742	0.249945
44. ASIC1:NM_001256830	ASIC1	-4.44766	1.13E-22
45. PPARGC1B:NM_001172699	PPARGC1B	-4.4433	1.54E-25
46. PNKD:NM_022572	PNKD	-4.43	7.13E-06
47. PPARGC1B:NM_001172698	PPARGC1B	-4.42352	2.15E-25
48. PPARGC1B:NM_133263	PPARGC1B	-4.41747	2.15E-25
49. ASIC1:NM_001095	ASIC1	-4.40213	2.22E-23
50. ASIC1:NM_020039	ASIC1	-4.38977	2.47E-23

Table 4-5. Top 50 down-regulated genes (and their descriptions) due to treatment of 226LDM cells with HU and NO resulting in cell-cycle block. The DESEQ package was used to generate this data

Gene ID	Description
1. IFNE	-No known function-
2. KCNJ10	This gene encodes a member of the inward rectifier-type potassium channel family, characterized by having a greater tendency to allow potassium to flow into, rather than out of, a cell. The encoded protein may form a heterodimer with another potassium channel protein and may be responsible for the potassium buffering action of glial cells in the brain. Mutations in this gene have been associated with seizure susceptibility of common idiopathic generalized epilepsy syndromes.
3. CD37	The protein encoded by this gene is a member of the transmembrane 4 superfamily, also known as the tetraspanin family. Most of these members are cell-surface proteins that are characterized by the presence of four hydrophobic domains. The proteins mediate signal transduction events that play a role in the regulation of cell development, activation, growth and motility. This encoded protein is a cell surface glycoprotein that is known to complex with integrins and other transmembrane 4 superfamily proteins. It may play a role in T-cell-B-cell interactions. Alternate splicing results in multiple transcript variants encoding different isoforms.
4. CHST4	This gene encodes an N-acetylglucosamine 6-O sulfotransferase. The encoded enzyme transfers sulfate from 3'phosphoadenosine 5'phospho-sulfate to the 6-hydroxyl group of N-acetylglucosamine on glycoproteins. This protein is localized to the Golgi and is involved in the modification of glycan structures on ligands of the lymphocyte homing receptor L-selectin. Alternate splicing in the 5' UTR results in multiple transcript variants that encode the same protein.
5. SBSPON	-No known function-
6. SLC25A27	Mitochondrial uncoupling proteins (UCP) are members of the larger family of mitochondrial anion carrier proteins (MACP). UCPs separate oxidative phosphorylation from ATP synthesis with energy dissipated as heat, also referred to as the mitochondrial proton leak. UCPs facilitate the transfer of anions from the inner to the outer mitochondrial membrane and the return transfer of protons from the outer to the inner mitochondrial membrane. They also reduce the mitochondrial membrane potential in mammalian cells. Tissue specificity occurs for the different UCPs and the exact methods of how UCPs transfer H+/OH- are not known. UCPs contain the three homologous protein domains of MACPs. Transcripts of this gene are only detected in brain tissue and are specifically modulated by various environmental conditions. Alternative splicing results in multiple transcript variants.
7. ADAMTS15	This gene encodes a member of the ADAMTS (a disintegrin and metalloproteinase with thrombospondin motifs) protein family. ADAMTS family members share several distinct protein modules, including a propeptide region, a metalloproteinase domain, a disintegrin-like domain, and a thrombospondin type 1 (TS) motif. Individual members of this family differ in the number of C-terminal TS motifs, and some have unique C-terminal domains. The protein encoded by this gene has a high sequence similarity to the proteins encoded by ADAMTS1 and ADAMTS8.
8. MARCH1	MARCH1 is a member of the MARCH family of membrane-bound E3 ubiquitin ligases (EC 6.3.2.19). MARCH proteins add ubiquitin (see MIM 191339) to target lysines in substrate proteins, thereby signaling their vesicular transport between membrane compartments. MARCH1 downregulates the surface expression of major histocompatibility complex (MHC) class II molecules (see MIM 142880) and other glycoproteins by directing them to the late endosomal/lysosomal compartment [supplied by OMIM, Mar 2010].
9. OCLM	The protein encoded by this gene is induced by cyclic mechanical stretching in trabecular cells of the eye and it is also expressed in retina. This protein may play a role in trabecular meshwork function and the development of glaucoma.
10. SLITRK5	Members of the SLITRK family, such as SLITRK5, are integral membrane proteins with 2 N-terminal leucine-rich repeat (LRR) domains similar to those of SLIT proteins (see SLIT1; MIM 603742). Most SLITRKs, including SLITRK5, also have C-terminal

	regions that share homology with neurotrophin receptors (see NTRK1; MIM 191315). SLITRKs are expressed predominantly in neural tissues and have neurite-modulating activity [supplied by OMIM, Mar 2008].
11. SLC29A1	This gene is a member of the equilibrative nucleoside transporter family. The gene encodes a transmembrane glycoprotein that localizes to the plasma and mitochondrial membranes and mediates the cellular uptake of nucleosides from the surrounding medium. The protein is categorized as an equilibrative (as opposed to concentrative) transporter that is sensitive to inhibition by nitrobenzylthioinosine (NBMPR). Nucleoside transporters are required for nucleotide synthesis in cells that lack de novo nucleoside synthesis pathways, and are also necessary for the uptake of cytotoxic nucleosides used for cancer and viral chemotherapies. Multiple alternatively spliced variants, encoding the same protein, have been found for this gene.
12. SLC16A7	This gene is a member of the monocarboxylate transporter family. Members in this family transport metabolites, such as lactate, pyruvate, and ketone bodies. The protein encoded by this gene catalyzes the proton-linked transport of monocarboxylates and has the highest affinity for pyruvate. This protein has been reported to be more highly expressed in prostate and colorectal cancer specimens when compared to control specimens. Alternative splicing results in multiple transcript variants.
13. FRMPD2	This gene encodes a peripheral membrane protein and is located in a region of chromosome 10q that contains a segmental duplication. This copy of the gene is full-length and is in the telomeric duplicated region. Two other more centromerically proximal copies of the gene are partial and may represent pseudogenes. This full-length gene appears to function in the establishment and maintenance of cell polarization. The protein is recruited to cell-cell junctions in an E-cadherin-dependent manner, and is selectively localized at the basolateral membrane in polarized epithelial cells. Alternative splicing results in multiple transcript variants.
14. HOXD9	This gene belongs to the homeobox family of genes. The homeobox genes encode a highly conserved family of transcription factors that play an important role in morphogenesis in all multicellular organisms. Mammals possess four similar homeobox gene clusters, HOXA, HOXB, HOXC and HOXD, located on different chromosomes, consisting of 9 to 11 genes arranged in tandem. This gene is one of several homeobox HOXD genes located at 2q31-2q37 chromosome regions. Deletions that removed the entire HOXD gene cluster or 5' end of this cluster have been associated with severe limb and genital abnormalities. The exact role of this gene has not been determined.
15. SLC47A2	This gene encodes a protein belonging to a family of transporters involved in excretion of toxic electrolytes, both endogenous and exogenous, through urine and bile. This transporter family shares homology with the bacterial MATE (multidrug and toxin extrusion) protein family responsible for drug resistance. This gene is one of two members of the MATE transporter family located near each other on chromosome 17. Alternatively spliced transcript variants encoding different isoforms have been identified for this gene.
16. DSG1	Desmosomes are cell-cell junctions between epithelial, myocardial and certain other cell types. Desmoglein 1 is a calcium-binding transmembrane glycoprotein component of desmosomes in vertebrate epithelial cells. Currently, three desmoglein subfamily members have been identified and all are members of the cadherin cell adhesion molecule superfamily. These desmoglein gene family members are located in a cluster on chromosome 18. The protein encoded by this gene has been identified as the autoantigen of the autoimmune skin blistering disease pemphigus foliaceus.
17. MXRA5	This gene encodes one of the matrix-remodelling associated proteins. This protein contains 7 leucine-rich repeats and 12 immunoglobulin-like C2-type domains related to perlecan. This gene has a pseudogene on chromosome Y. [provided by RefSeq, Mar 2010].
18. SLC34A1	This gene encodes a member of the type II sodium-phosphate cotransporter family. Mutations in this gene are associated with hypophosphatemia nephrolithiasis/osteoporosis 1. Alternative splicing results in multiple transcript variants.
19. GPRASP1	This gene encodes a member of the GPRASP (G protein-coupled receptor associated sorting protein) family. The protein may modulate lysosomal sorting and functional down-regulation of a variety of G-protein coupled receptors. It targets receptors for

	degradation in lysosomes. The receptors interacting with this sorting protein include D2 dopamine receptor (DRD2), delta opioid receptor (OPRD1), beta-2 adrenergic receptor (ADRB2), D4 dopamine receptor (DRD4) and cannabinoid 1 receptor (CB1R). Multiple alternatively spliced transcript variants encoding the same protein have been identified.
20. SPATA25	-No known function-
21. METTL7A	-No known function-
22. CYP3A5	This gene encodes a member of the cytochrome P450 superfamily of enzymes. The cytochrome P450 proteins are monooxygenases which catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids. The encoded protein metabolizes drugs as well as the steroid hormones testosterone and progesterone. This gene is part of a cluster of cytochrome P450 genes on chromosome 7q21.1. Two pseudogenes of this gene have been identified within this cluster on chromosome 7. Expression of this gene is widely variable among populations, and a single nucleotide polymorphism that affects transcript splicing has been associated with susceptibility to hypertensions. Alternative splicing results in multiple transcript variants.
23. CHRNG	The mammalian muscle-type acetylcholine receptor is a transmembrane pentameric glycoprotein with two alpha subunits, one beta, one delta, and one epsilon (in adult skeletal muscle) or gamma (in fetal and denervated muscle) subunit. This gene, which encodes the gamma subunit, is expressed prior to the thirty-third week of gestation in humans. The gamma subunit of the acetylcholine receptor plays a role in neuromuscular organogenesis and ligand binding and disruption of gamma subunit expression prevents the correct localization of the receptor in cell membranes. Mutations in this gene cause Escobar syndrome and a lethal form of multiple pterygium syndrome. Muscle-type acetylcholine receptor is the major antigen in the autoimmune disease myasthenia gravis.
24. ASIC1	This gene encodes a member of the acid-sensing ion channel (ASIC) family of proteins, which are part of the degenerin/epithelial sodium channel (DEG/ENaC) superfamily. Members of the ASIC family are sensitive to amiloride and function in neurotransmission. The encoded proteins function in learning, pain transduction, touch sensation, and development of memory and fear. Alternatively spliced transcript variants have been described.
25. PPARGC1B	The protein encoded by this gene stimulates the activity of several transcription factors and nuclear receptors, including estrogen receptor alpha, nuclear respiratory factor 1, and glucocorticoid receptor. The encoded protein may be involved in fat oxidation, non-oxidative glucose metabolism, and the regulation of energy expenditure. This protein is downregulated in prediabetic and type 2 diabetes mellitus patients. Certain allelic variations in this gene increase the risk of the development of obesity. Three transcript variants encoding different isoforms have been found for this gene.
26. PNKD	This gene is thought to play a role in the regulation of myofibrillogenesis. Mutations in this gene have been associated with the movement disorder paroxysmal non-kinesigenic dyskinesia. Alternative splicing results in multiple transcript variants.
27. SLFNL1	-No known function-
28. CLEC7A	This gene encodes a member of the C-type lectin/C-type lectin-like domain (CTL/CTLD) superfamily. The encoded glycoprotein is a small type II membrane receptor with an extracellular C-type lectin-like domain fold and a cytoplasmic domain with an immunoreceptor tyrosine-based activation motif. It functions as a pattern-recognition receptor that recognizes a variety of beta-1,3-linked and beta-1,6-linked glucans from fungi and plants, and in this way plays a role in innate immune response. Alternate transcriptional splice variants, encoding different isoforms, have been characterized. This gene is closely linked to other CTL/CTLD superfamily members on chromosome 12p13 in the natural killer gene complex region.
29. HOXD10	This gene is a member of the Abd-B homeobox family and encodes a protein with a homeobox DNA-binding domain. It is included in a cluster of homeobox D genes located on chromosome 2. The encoded nuclear protein functions as a sequence-specific transcription factor that is expressed in the developing limb buds and is involved in differentiation and limb development. Mutations in this gene have been

	associated with Wilm's tumor and congenital vertical talus (also known as 'rocker-bottom foot' deformity or congenital convex pes valgus) and/or a foot deformity resembling that seen in Charcot-Marie-Tooth disease.
30. RNF128	The protein encoded by this gene is a type I transmembrane protein that localizes to the endocytic pathway. This protein contains a RING zinc-finger motif and has been shown to possess E3 ubiquitin ligase activity. Expression of this gene in retrovirally transduced T cell hybridoma significantly inhibits activation-induced IL2 and IL4 cytokine production. Induced expression of this gene was observed in anergic CD4(+) T cells, which suggested a role in the induction of anergic phenotype. Alternatively spliced transcript variants encoding distinct isoforms have been reported.
31. BGN	The protein encoded by this gene is a small cellular or pericellular matrix proteoglycan that is closely related in structure to two other small proteoglycans, decorin and fibromodulin. The encoded protein and decorin are thought to be the result of a gene duplication. Decorin contains one attached glycosaminoglycan chain, while this protein probably contains two chains. For this reason, this protein is called biglycan. This protein plays a role in assembly of collagen fibrils and muscle regeneration. It interacts with several proteins involved in muscular dystrophy, including alpha-dystroglycan, alpha- and gamma-sarcoglycan and collagen VI, and it is critical for the assembly of the dystrophin-associated protein complex.
32. PTPRZ1	This gene encodes a member of the receptor protein tyrosine phosphatase family. Expression of this gene is restricted to the central nervous system (CNS), and it may be involved in the regulation of specific developmental processes in the CNS. Alternatively spliced transcript variants encoding different isoforms have been described for this gene.
33. KCNA7	Potassium channels represent the most complex class of voltage-gated ion channels from both functional and structural standpoints. Their diverse functions include regulating neurotransmitter release, heart rate, insulin secretion, neuronal excitability, epithelial electrolyte transport, smooth muscle contraction, and cell volume. Four sequence-related potassium channel genes - shaker, shaw, shab, and shal - have been identified in Drosophila, and each has been shown to have human homolog(s). This gene encodes a member of the potassium channel, voltage-gated, shaker-related subfamily. This member contains six membrane-spanning domains with a shaker-type repeat in the fourth segment. The gene is expressed preferentially in skeletal muscle, heart and kidney. It is a candidate gene for inherited cardiac disorders.
34. CLDN10	This gene encodes a member of the claudin family. Claudins are integral membrane proteins and components of tight junction strands. Tight junction strands serve as a physical barrier to prevent solutes and water from passing freely through the paracellular space between epithelial or endothelial cell sheets, and also play critical roles in maintaining cell polarity and signal transductions. The expression level of this gene is associated with recurrence of primary hepatocellular carcinoma. Six alternatively spliced transcript variants encoding different isoforms have been reported, but the transcript sequences of some variants are not determined.
35. TLL1	This gene encodes an astacin-like, zinc-dependent, metalloprotease that belongs to the peptidase M12A family. This protease processes procollagen C-propeptides, such as chordin, pro-biglycan and pro-lysyl oxidase. Studies in mice suggest that this gene plays multiple roles in the development of mammalian heart, and is essential for the formation of the interventricular septum. Allelic variants of this gene are associated with atrial septal defect type 6. Alternatively spliced transcript variants encoding different isoforms have been found for this gene.
36. DLX1	This gene encodes a member of a homeobox transcription factor gene family similiar to the Drosophila distal-lessgene. The encoded protein is localized to the nucleus where it may function as a transcriptional regulator of signals from multiple TGF-beta superfamily members. The encoded protein may play a role in the control of craniofacial patterning and the differentiation and survival of inhibitory neurons in the forebrain. This gene is located in a tail-to-tail configuration with another member of the family on the long arm of chromosome 2. Alternatively spliced transcript variants encoding different isoforms have been described.
37. ZNF540	-No known function-
38. ATP8A1	The P-type adenosinetriphosphatases (P-type ATPases) are a family of proteins which use the free energy of ATP hydrolysis to drive uphill transport of ions across

	membranes. Several subfamilies of P-type ATPases have been identified. One subfamily catalyzes transport of heavy metal ions. Another subfamily transports non-heavy metal ions (NMHI). The protein encoded by this gene is a member of the third subfamily of P-type ATPases and acts to transport amphipaths, such as phosphatidylserine. Two transcript variants encoding different isoforms have been found for this gene.
39. MRGPRX3	This gene encodes a member of the mas-related/sensory neuron specific subfamily of G protein coupled receptors. The encoded protein may be involved in sensory neuron regulation and in the modulation of pain.
40. MFSD2A	-No known function-
41. EPB41L4A	Members of the band 4.1 protein superfamily, including EPB41L4A, are thought to regulate the interaction between the cytoskeleton and plasma membrane [supplied by OMIM, Jul 2008].
42. KRT1	The protein encoded by this gene is a member of the keratin gene family. The type II cytokeratins consist of basic or neutral proteins which are arranged in pairs of heterotypic keratin chains coexpressed during differentiation of simple and stratified epithelial tissues. This type II cytokeratin is specifically expressed in the spinous and granular layers of the epidermis with family member KRT10 and mutations in these genes have been associated with bullous congenital ichthyosiform erythroderma. The type II cytokeratins are clustered in a region of chromosome 12q12-q13.
43. FLRT2	This gene encodes a member of the fibronectin leucine rich transmembrane protein (FLRT) family. FLRT family members may function in cell adhesion and/or receptor signalling. Their protein structures resemble small leucine-rich proteoglycans found in the extracellular matrix.
44. ATP7B	This gene is a member of the P-type cation transport ATPase family and encodes a protein with several membrane-spanning domains, an ATPase consensus sequence, a hinge domain, a phosphorylation site, and at least 2 putative copper-binding sites. This protein functions as a monomer, exporting copper out of the cells, such as the efflux of hepatic copper into the bile. Alternate transcriptional splice variants, encoding different isoforms with distinct cellular localizations, have been characterized. Mutations in this gene have been associated with Wilson disease (WD).
45. C2orf88	-No known function-
46. GPR63	This gene encodes a G protein-coupled receptor. Multiple alternatively spliced variants, encoding the same protein, have been identified.
47. MAGEE1	This gene encodes an alpha-dystrobrevin-associated MAGE (melanoma-associated antigen) protein, which is a member of the MAGE family. The protein contains a nuclear localization signal in the N-terminus, 30 12-amino acid repeats beginning at nt 60 with the consensus sequence ASEGPSTSVLPT, and two MAGE domains in the C-terminus. It may play a signaling role in brain, muscle, and peripheral nerve. This gene is located on X chromosome in a region containing loci linked to mental retardation.
48. INHBC	This gene encodes the beta C chain of inhibin, a member of the TGF-beta superfamily. This subunit forms heterodimers with beta A and beta B subunits. Inhibins and activins, also members of the TGF-beta superfamily, are hormones with opposing actions and are involved in hypothalamic, pituitary, and gonadal hormone secretion, as well as growth and differentiation of various cell types.
49. LHX4	This gene encodes a member of a large protein family which contains the LIM domain, a unique cysteine-rich zinc-binding domain. The encoded protein is a transcription factor involved in the control of differentiation and development of the pituitary gland. Mutations in this gene cause combined pituitary hormone deficiency 4.
50. TMC2	This gene is considered a member of a gene family predicted to encode transmembrane proteins. The specific function of this gene is unknown; however, expression in the inner ear suggests that it may be crucial for normal auditory function. Mutations in this gene may underlie hereditary disorders of balance and hearing.

4.3.4 Gene Ontology Enrichment Analysis

The obtained lists of significantly differentially expressed genes were used to perform Gene Ontology analysis in order to visualize the common functional relationships between them. Gene Ontology annotation terms (GO terms) form a universally agreed vocabulary assigned to genes according to their attributes or functions. Multiple GO terms can be assigned to each gene depending on the current knowledge of the functions it is involved. REViGO is a web server that can manage long lists of gene GO terms by removing redundant terms and visualizing the remaining ones based on semantic similarity (Supek et al., 2011). This tool, along with the Cytoscape V3.2.1 software program (Shannon et al., 2003), was used to perform enrichment analysis and visualize the relationships on all the up-regulated and down-regulated genes from this study (figure 4-9 and 4-10 respectively).

The nodes and edges presented in GO graphs summarize some of the biological processes that the differentially expressed genes are known to participate in. Genes involved in the immune response, cell adhesion, regulation of cell activation and gene expression, are among others that get more actively transcribed in the treated cells. At the same time, genes mostly involved in cell signaling pathways, embryo morphogenesis, ion and transmembrane transport seem to be de-activated during this cell cycle block.

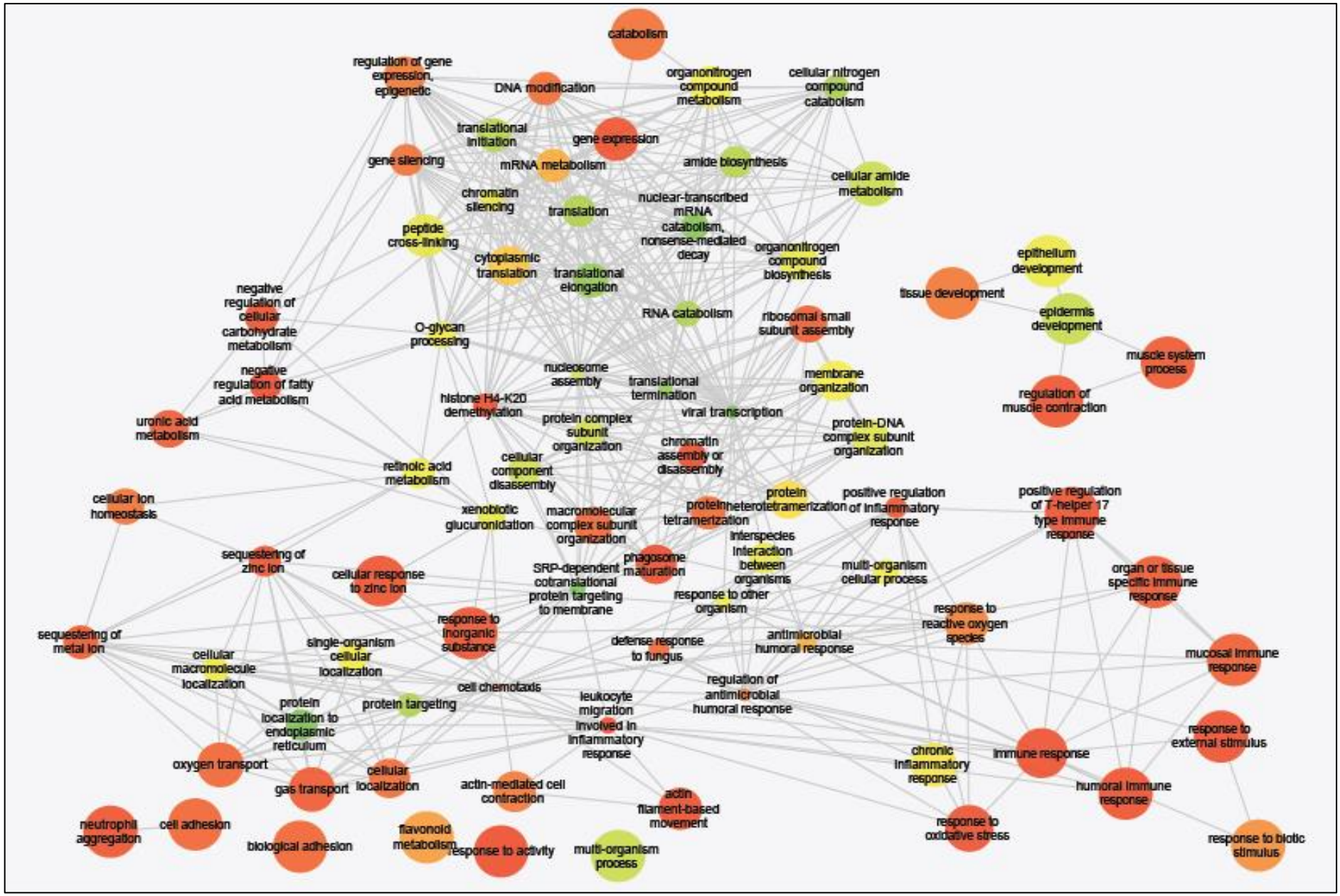

catabolism
regulation of gene expression, epigenetic
DNA modification
organonitrogen compound metabolism
cellular nitrogen compound catabolism
translational initiation
gene expression
gene silencing
mRNA metabolism
amide biosynthesis
cellular amide metabolism
chromatin silencing
translation
nuclear-transcribed mRNA catabolism, nonsense-mediated decay
peptide cross-linking
cytoplasmic translation
translational elongation
organonitrogen compound biosynthesis
epithelium development
tissue development
epidermis development
negative regulation of cellular carbohydrate metabolism
RNA catabolism
ribosomal small subunit assembly
O-glycan processing
muscle system process
nucleosome assembly
translational termination
membrane organization
negative regulation of fatty acid metabolism
histone H4-K20 demethylation
viral transcription
regulation of muscle contraction
uronic acid metabolism
protein complex subunit organization
protein-DNA complex subunit organization
chromatin assembly or disassembly
cellular component disassembly
retinoic acid metabolism
cellular ion homeostasis
protein heterotetramerization
protein tetramerization
positive regulation of inflammatory response
positive regulation of T-helper 17 type immune response
xenobiotic glucuronidation
macromolecular complex subunit organization
interspecies interaction between organisms
sequestering of zinc ion
phagosome maturation
multi-organism cellular process
organ or tissue specific immune response
cellular response to zinc ion
SRP-dependent cotranslational protein targeting to membrane
response to other organism
response to reactive oxygen species
response to inorganic substance
sequestering of metal ion
defense response to fungus
antimicrobial humoral response
mucosal immune response
single-organism cellular localization
cellular macromolecule localization
regulation of antimicrobial humoral response
cell chemotaxis
protein targeting
leukocyte migration involved in inflammatory response
protein localization to endoplasmic reticulum
response to external stimulus
chronic inflammatory response
immune response
oxygen transport
cellular localization
gas transport
actin-mediated cell contraction
actin filament-based movement
humoral immune response
neutrophil aggregation
cell adhesion
response to oxidative stress
flavonoid metabolism
response to activity
multi-organism process
biological adhesion
response to biotic stimulus

Figure 4-9 Gene Ontology analysis for the genes with up-regulated expression in cell-cycle blocked 226LDM cells treated with hydroxyurea and nocodazole

Differential expression analysis was performed on control and cell-cycle blocked 225LDM cells with the DESEQ package. A gene ontology analysis was performed with the ensuing ranked list of significantly up-regulated genes. GO annotation terms were assigned to each gene and statistical analysis and clustering of these terms was performed by the REViGO web server. The ontology relationships between the up-regulated genes are shown in the graph with nodes representing biological processes connected in a parent-child manner with edges. The size of the nodes represents the log_size of terms falling into the biological process described in the node label. The color of each node varies according to the value (from red to green, red representing the lowest q value). The editing of the graph was performed with the CytoScape

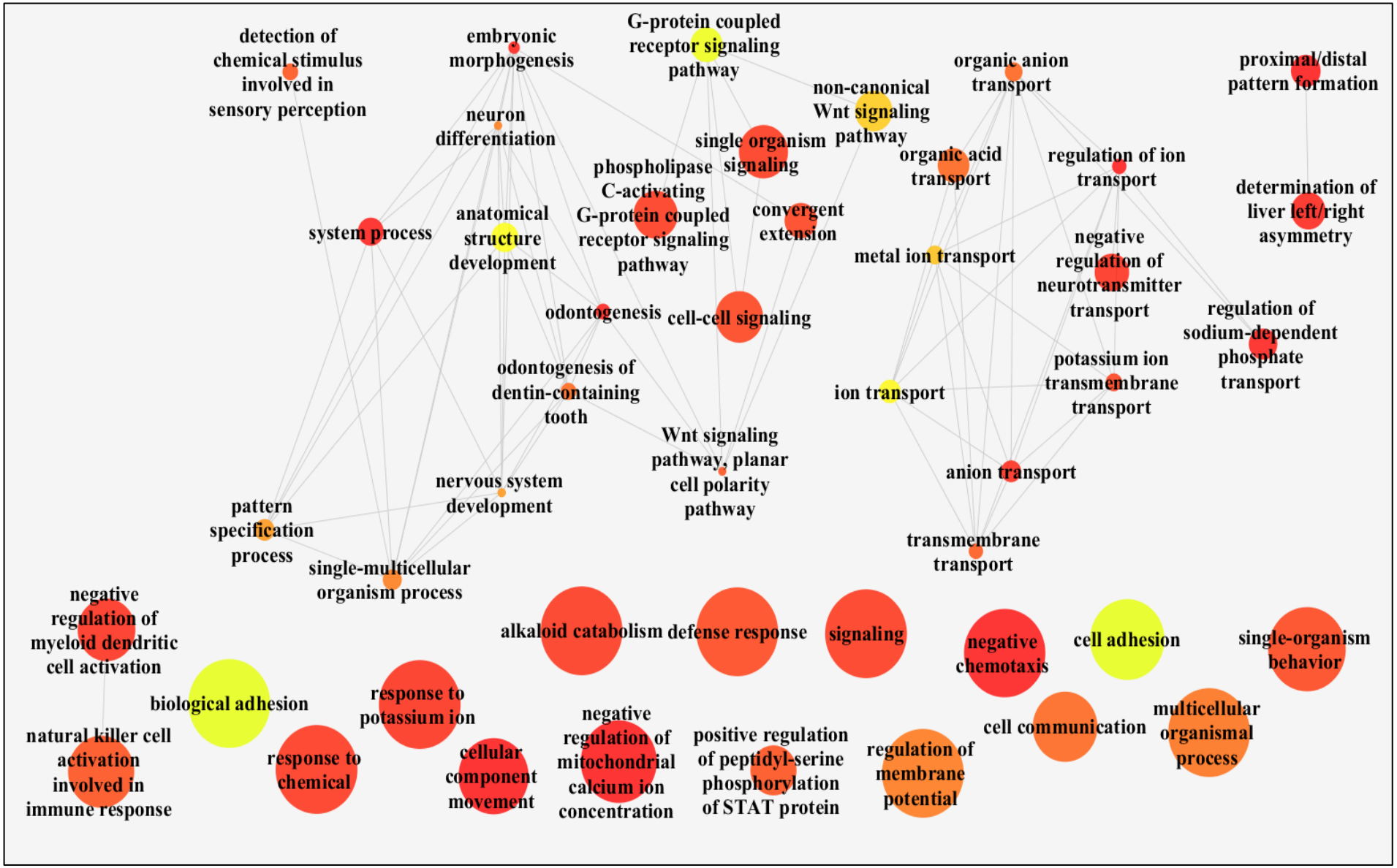

detection of chemical stimulus involved in sensory perception
embryonic morphogenesis
G-protein coupled receptor signaling pathway
organic anion transport
proximal/distal pattern formation
non-canonical Wnt signaling pathway
neuron differentiation
single organism signaling
organic acid transport
regulation of ion transport
phospholipase C-activating G-protein coupled receptor signaling pathway
determination of liver left/right asymmetry
system process
anatomical structure development
convergent extension
metal ion transport
negative regulation of neurotransmitter transport
odontogenesis
cell-cell signaling
regulation of sodium-dependent phosphate transport
potassium ion transmembrane transport
odontogenesis of dentin-containing tooth
ion transport
Wnt signaling pathway, planar cell polarity pathway
anion transport
nervous system development
pattern specification process
transmembrane transport
single-multicellular organism process
negative regulation of myeloid dendritic cell activation
alkaloid catabolism
defense response
signaling
negative chemotaxis
cell adhesion
single-organism behavior
biological adhesion
response to potassium ion
negative regulation of mitochondrial calcium ion concentration
positive regulation of peptidyl-serine phosphorylation of STAT protein
cell communication
multicellular organismal process
natural killer cell activation involved in immune response
response to chemical
cellular component movement
regulation of membrane potential

Figure 4-10 Gene Ontology analysis for the genes with down-regulated expression in cell-cycle arrested 226LDM cells treated with hydroxyurea and nocodazole

Differential expression analysis was performed on control and cell-cycle blocked 225LDM cells with the DESEQ package. Gene Ontology analysis was performed with the ensuing ranked list of significantly down-regulated genes. GO annotation terms were assigned to each gene and statistical analysis and clustering of these terms was performed by the REViGO web server. The relationships between the up-regulated genes are shown in the graph with nodes representing biological processes connected in a parent-child manner with edges. The size of the nodes represents the log_size of terms falling into the biological process described in the node label. The color of each node varies according to the q-value (from red to green, red representing the lowest *q* value). The editing of the graph was performed with the CytoScape V3.2.1 software.

4.3.5 Non-coding RNA

The function of non-coding RNAs remains to a great extent unknown; however their expression patterns have been reported to be dynamic and potentially purposeful. Considering this, further to the expression analysis on protein coding genes, we also explored the expression patterns of non-coding genes and identified significantly differentially expressed genes between the control and treated 226LDM cells.

The raw sequenced RNA data, in FASTQ format, were aligned and mapped on the whole human genome and a pipeline was developed to screen for non-coding genes. Diagnostics plots and quality control was performed as described previously using the DESEQ package and the quality of the samples and the analysis was confirmed (figure 4-11, **Error! Reference source not found.**-12 and figure 4-12).

From the analysis, the differentially expressed genes were filtered to discard those that have a Q value higher than 0.05 and the remaining ones were divided in up-regulated and down-regulated according to their fold change in between the two physiological conditions. The top 50 up~ and down-regulated non-coding genes are presented in table 4-6 and table 4-7 respectively. In these tables the gene symbol along with the fold change and *q* value can be seen.

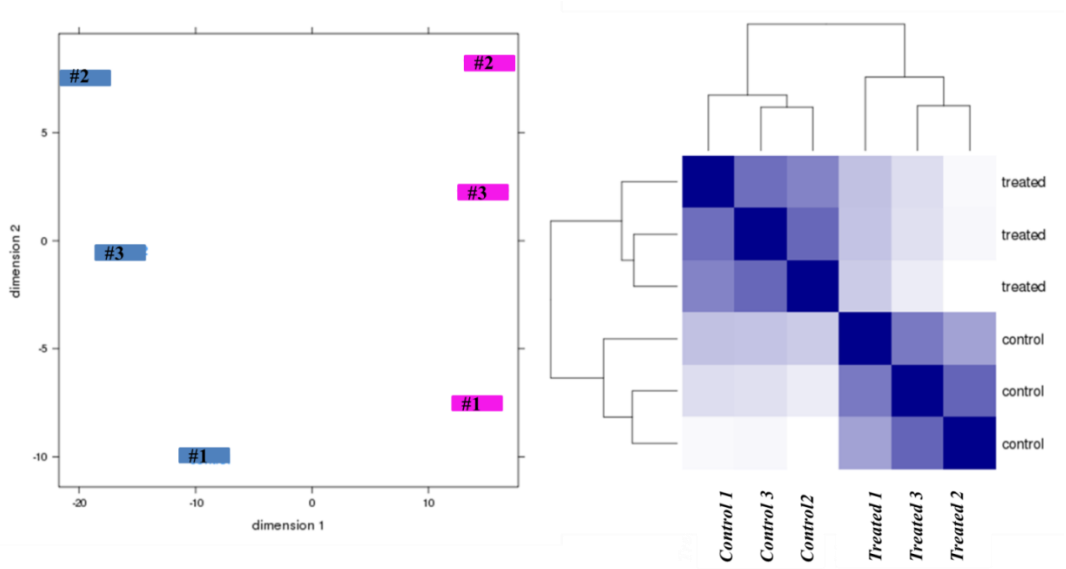


Figure 4-11 Non-coding RNA sample distances visualized in two-dimensional graphs by the DESEQ package

The similarities between the control and treated samples are pictured in a multi-dimensional scaling (MDS) plot and a sample-distance heatmap. Left: In MDS plot, the level of similarity between samples of our dataset can be seen. The control samples are shown in blue color while the treated samples are shown in pink. Right: The Euclidean distances between the samples can be seen in the sample heatmap. The dendrogram at the sides demonstrate the overall similarity between samples.

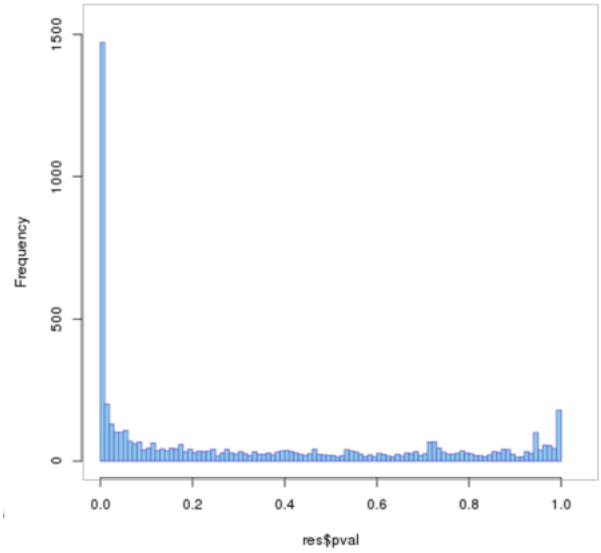


Figure 4-12 **Histogram of P values generated from the DESEQ package**

A histogram of P values was generated by the DESEQ package to assess the quality of the samples. The histogram is used to visualize the distribution of *p* values within certain intervals and to assess whether problems are present with the analysis.

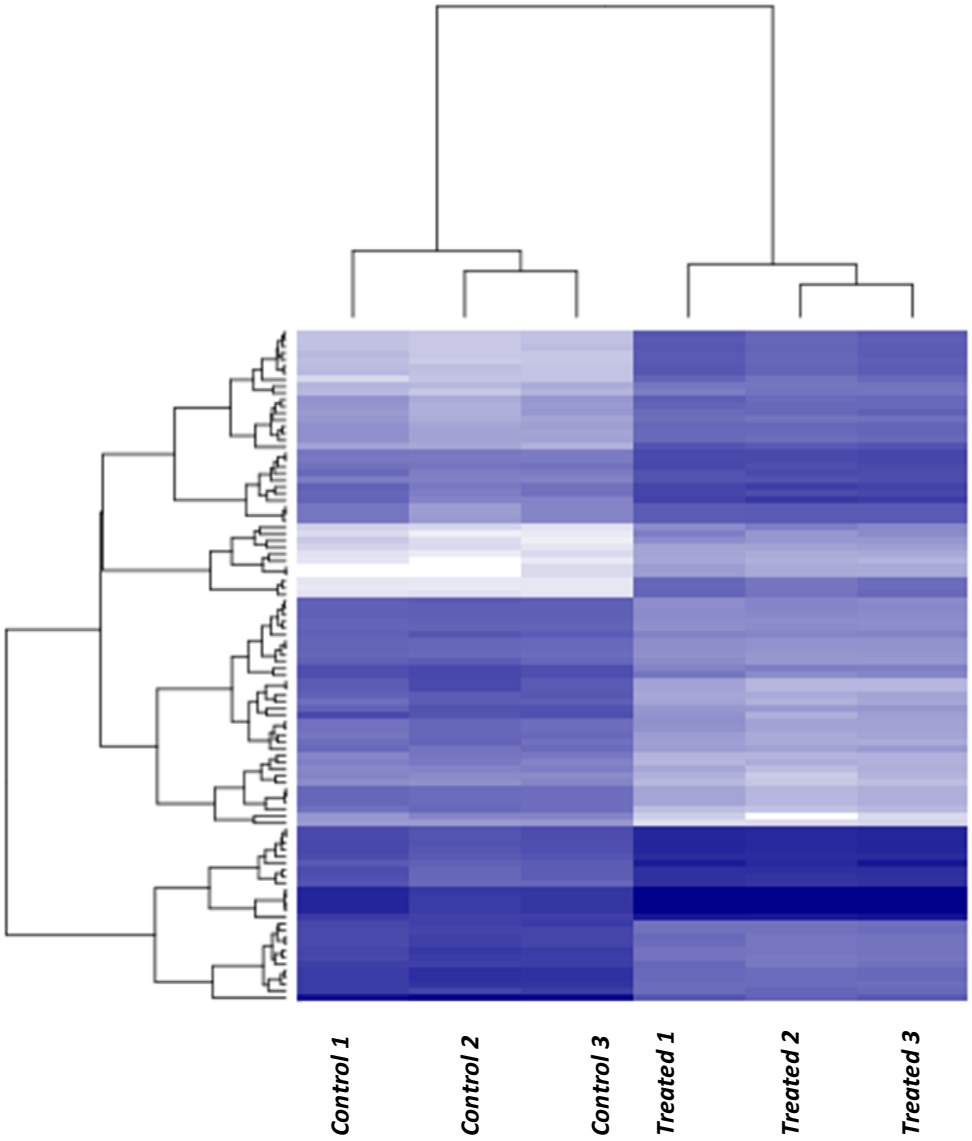


Figure 4-12 Heatmap of differential expression of non-coding genes in control and treated 226LDM cells

The DESEQ package was used to generate the expression heatmap for non-coding transcripts. The expression divergence of the top 100 genes/transcripts affected by the treatment in 226LDM cells can be seen in the heatmap according to *p* value. A clustering dendrogram representing the clustering of the column data (samples) can be seen on top of the heatmap.

Table 4-6 Top 50 up-regulated non-coding gene transcripts, sorted according to fold change, as a consequence of cell cycle arrest in 226LDM cells, discovered using the DESEQ package

Transcript ID	log2 FoldChange	p-value adjusted
1. RNU6-2.9:NR_125730.9	8.089406202	3.16E-31
2. RNU6-2:NR_125730	7.757098858	1.05E-29
3. RNU6-2.2:NR_125730.2	7.462141616	9.97E-27
4. KRTAP5-AS1:NR_021489	6.584861497	5.95E-05
5. TMCO5B:NR_046005	6.321300035	0.0018814
6. RNVU1-19:NR_104086	6.316590435	0.0088546
7. LINC01405:NR_036513	6.103343714	0.00015
8. INS-IGF2:NR_003512	6.093284077	1.80E-08
9. SPRR2C:NR_003062	5.771581929	4.56E-06
10. ABHD11-AS1:NR_026690	5.753429704	6.32E-06
11. RNU6-2.5:NR_125730.5	5.558473821	1.00E-24
12. RNU6-2.4:NR_125730.4	5.552410981	1.40E-33
13. TEX26-AS1:NR_038287	5.544292915	0.0061583
14. RNU6-2.10:NR_125730.10	5.537704365	9.04E-27
15. RNU6-2.8:NR_125730.8	5.506411466	4.49E-28
16. VTCN1:NR_045603	5.435773714	0.0033258
17. RNU6-2.1:NR_125730.1	5.422723069	3.16E-31
18. VTCN1:NR_045604	5.421724439	0.0030636
19. RNU6-2.7:NR_125730.7	5.371384241	4.05E-37
20. RNU6-2.6:NR_125730.6	5.358272526	1.57E-31
21. RNU6-2.3:NR_125730.3	5.35144416	4.94E-19
22. WFDC21P:NR_030732	5.12062436	0.0001172
23. LINC01269:NR_125769	5.118798495	0.0001259
24. C1QTNF1-AS1:NR_040018	5.103587415	2.06E-12
25. C1QTNF1-AS1:NR_040019	5.103587415	2.06E-12
26. RNU6ATAC:NR_023344	4.968268105	3.30E-10
27. SERPINB9P1:NR_033851	4.953280428	0.0001971
28. APOBEC3B-AS1:NR_104187	4.920520368	4.75E-13
29. LINC00273:NR_038368	4.897432426	0.0009009
30. GLIPR2:NR_104639	4.788283528	1.33E-06
31. GLIPR2:NR_104640	4.786723099	1.51E-06
32. GLIPR2:NR_104638	4.77872	1.46E-06
33. GLIPR2:NR_104637	4.776676104	1.72E-06
34. ANO1-AS2:NR_103835	4.690662392	6.55E-07
35. ANKRD30BL:NR_027020	4.651558571	0.0003838
36. FMOD:NR_103757	4.650916487	0.023637
37. GLIPR2:NR_104641	4.585937318	5.73E-07
38. CHRNA4:NR_046317	4.443940678	0.0443368

39. LINC01133:NR_038849	4.39171289	1.24E-29
40. C5orf66-AS1:NR_105050	4.382870174	0.0014902
41. C5orf66-AS1:NR_105049	4.382870174	0.0014902
42. KRT19P2:NR_036685	4.36634295	0.001897
43. PDE6G:NR_026872	4.316991767	0.0129551
44. FAM25BP:NR_104039	4.167544243	1.56E-15
45. LINC01540:NR_110429	4.129800733	0.0062991
46. LINC01540:NR_110428	4.09481155	0.0075615
47. LINC00592:NR_027358	4.064977222	5.09E-09
48. CIB2:NR_125435	3.879924246	0.0003211
49. HSH2D:NR_111903	3.732336682	8.43E-19
50. LINC00160:NR_024351	3.732053845	4.12E-05

Table 4-7 Top 50 down-regulated long non-coding gene transcripts, sorted according to fold change, as a result of cell-cycle arrest in 226LDM cells, discovered with the DESEQ package

Transcript ID	log2 FoldChange	p-value adjusted
1. MLLT10P1:NR_045115	-6.296682603	1.80E-05
2. FAM106A:NR_026809	-5.798489612	0.0013112
3. KIZ-AS1:NR_109956	-5.515303113	0.0616641
4. LINC00173:NR_027345	-5.296498623	0.0231589
5. HCG8:NR_103542	-5.118264081	6.27E-10
6. LINC01355:NR_110616	-5.110507631	0.0001508
7. MANEA-AS1:NR_104136	-5.086470455	0.0243794
8. MRPL23-AS1:NR_024471	-4.997112262	0.1149861
9. CYP3A5:NR_033807	-4.979831552	0.0001357
10. SLC16A7:NR_073055	-4.76189461	2.61E-10
11. SLC16A7:NR_073056	-4.761369187	2.61E-10
12. FRMPD2B.1:NR_033172.1	-4.699310133	0.0667126
13. ZNF571-AS1:NR_038247	-4.672797079	0.0248735
14. ZNF571-AS1:NR_038248	-4.617578806	0.0296082
15. ZNF571-AS1:NR_038249	-4.617578806	0.0296082
16. LINC00173:NR_027346	-4.588640859	0.0950188
17. SCARNA27:NR_003703	-4.559327696	0.0053897
18. LINC00852:NR_026829	-4.539455161	1.62E-05
19. TSIX:NR_003255	-4.461121648	2.39E-15
20. XIST:NR_001564	-4.423481431	1.93E-08
21. GRTP1-AS1:NR_120385	-4.375893521	0.0732026
22. ASIC1:NR_046389	-4.358679274	2.32E-25
23. FRMPD2B:NR_033172	-4.340154087	0.0256721
24. RPS15AP10:NR_026768	-4.289058273	0.0150915
25. LINC00479:NR_027272	-4.259274125	0.0001975
26. NPTN-IT1:NR_103844	-4.235113165	5.73E-07
27. DPRXP4:NR_002221	-4.199329154	0.0075905
28. LINC01179:NR_121676	-4.184252885	0.000336
29. FLJ44511:NR_033963	-4.17057588	0.1460655
30. MFSD2A:NR_109896	-4.154319272	6.62E-28
31. LINC01004:NR_039981	-4.10359801	0.0004323
32. INE1:NR_024616	-4.102383616	3.25E-12
33. LINC01252:NR_033890	-4.101614637	0.0931621
34. HCG27.2:NR_026791.2	-4.06847	0.0204587
35. PWAR5:NR_022008	-4.046436951	0.032033
36. LINC00086:NR_024359	-4.04028287	0.0869138
37. STAM-AS1:NR_110370	-4.012747071	0.0008809
38. DKFZP434I0714:NR_033797	-3.970170784	4.13E-05

39. LHX4-AS1:NR_037642	-3.960061996	1.38E-24
40. RPL23AP64:NR_003040	-3.955597229	6.18E-05
41. LINC00663:NR_026956	-3.940774443	0.006069
42. HERC2P7:NR_036470	-3.925136331	8.03E-07
43. HERC2P4:NR_109773	-3.92286214	3.58E-06
44. N4BP2L2-IT2:NR_026928	-3.910954469	1.13E-07
45. MTVR2:NR_027025	-3.87489278	0.1974436
46. MBL1P:NR_002724	-3.81830226	3.21E-05
47. ZNF154:NR_110974	-3.759754608	8.23E-06
48. LINC00965:NR_027000	-3.741198478	0.0554387
49. USP32P2:NR_003554	-3.723082332	5.28E-10
50. SLC7A11-AS1:NR_038380	-3.715296713	0.0475209
51. IL12RB2:NR_047584	-3.692018342	0.0121721

4.4 Discussion

The objective of this chapter was to identify gene expression changes that took place in 226LDM cells that were chemically induced to stop proliferating. This was accomplished by extracting, sequencing, analysing and comparing total RNA from two separate groups of cells: the first consisted of proliferating cells and the second population was treated with hydroxyurea and nocodazole to block cell cycle. Cells belonging to the latter group expressed only the CTCF180 isoform instead of both known CTCF isoforms.

The extracted RNA was used in a massively parallel sequencing experiment and the output data were explored using an analysis pipeline based on bioinformatics tools and software. With the RNA-seq experiment it was possible to acquire virtually complete snapshots of the transcriptomes in the two cell populations. This allowed us to perform a comparison analysis and discover the pathways that are potentially activated or deactivated as a result of the treatment. Thousands of coding and non-coding genes were found to be significantly over~ or under-expressed in the treated cells.

From this investigation it was possible to retrieve lists of differentially expressed ranked genes and to visualize the functional relationships of the biological processes that they are involved in by exploiting the Gene Ontology project annotations and tools. In both groups, namely up and down-regulated genes, the actual results generally met the expectations. For example, among the up-regulated genes in arrested cells there are those involved in processes such as cell adhesion, translation, immune response and gene expression. On the other hand, down-regulated genes are mostly involved in cell signaling pathways, ion transport, morphogenesis and polarity as well as cell adhesion. The number of the affected processes involved in RNA metabolism is particularly remarkable, including RNA processing and catabolism. The biological interpretation of this observation may be that in the general

requirement in the RNA-mediated processes is reduced in the arrested cells, which are less metabolically active. In the treated cells, pathways that are involved in response to stress and chemicals is activated and at the same time it appears that the immune response is also noticeably stimulated. In conclusion, the changes identified in the transcriptomes of the cells appear to reflect two different biological situations (healthy proliferating *vs* arrested/stressed cells).

It should be noted that the GO annotation is a dynamic event and numerous terms can change; they can be assigned to different genes or even be discontinued. Moreover, many of the genes from the list of the differentially expressed genes could currently not be assigned a GO term at all, in which case they were not included in the GO analysis.

Differential expression of non-coding RNA transcripts was also reported in this study. Although the identities of significantly affected non-coding transcripts were obtained, their functions remain still largely unknown. Known information regarding non-coding RNA is stored in online databases, most of which focus on miRNA. Databases specifically for long non-coding RNA, most prominently NONCODE (Xie et al., 2014), collect information from the existing literature and other available databases regarding expression profiles of known lncRNAs, potential links to disease and prediction of function. The challenges of lncRNA annotation and the need for more universal and consistently accessible information for long non-coding RNAs have been highlighted in previous studies (Paschoal et al., 2012, Pauli et al., 2015).

In summary, in this Chapter the RNA profiles of proliferating and arrested 226LDM cells have been generated for the first time. A huge volume of novel data has been obtained and duly analyzed using a number of bioinformatics tools, which in turn allowed us to identify gene expression patterns characteristic for two populations of 226LDM cells.

With regards to both, coding and non-coding RNA, an inherent limitation of RNA-seq is that it can provide answers regarding gene expression without information on to how these changes occurred and/ or in which order it happened. The underlying molecular processes which are involved in the regulation of gene expression cannot be elucidated using the RNA-seq technique alone; a combination of RNA-seq experiments with other experimental techniques, such as ChIP-seq and real-time PCR, will be important to link the expression profiles to the function of particular transcription factors.

More specifically, in our study the particular focus will be on a transcription factor, CTCF. We will investigate whether differential expression of the genes in two populations of 226LDM cells is associated with the occupancies of the promoters of these genes by CTCF130/CTCF180. This will be achieved by intersecting the RNA-seq data with the CTCF ChIP-seq data in proliferating and arrested 226LDM cells; it is described in the following Chapter 5.

Chapter 5 Integration of the ChIP-Seq and RNA-Seq data obtained from the populations of proliferating and arrested 226LDM cells.

5.1 Introduction

The recent development of next generation high-throughput genomic technologies has led to an unprecedented progress in genomic research. Techniques such as chromatin immunoprecipitation sequencing (ChIP-seq), RNA sequencing and genome-wide association studies (GWAS) are becoming increasingly accessible and have paved the way for new research opportunities. Many studies have been published using high throughput techniques and at the same time a variety of bioinformatics tools has been developed to aid the analysis of the large output datasets (Cullum et al., 2011).

It remains challenging to elucidate the interplay between genetics and epigenetics and the combination of next generation techniques allows for more complex research questions like this to be asked and more informative answers to be generated (Hawkins et al., 2010, Angelini and Costa, 2014).

The integration of ChIP and RNA-seq can provide the information about a possible association between the binding of a transcription factor and the effects of this binding on gene expression. Other studies have used the combination of these techniques to predict changes in expression according to differential transcription factor binding (Ouyang et al., 2009).

More research is needed to accurately and quantitatively understand the existing, complex genome-wide networks however these techniques are a considerable step forward by elucidating the roles of key players in main events.

5.2 Experimental Aims

For the experiments conducted in the two previous chapters, the 226LDM cell model was used in which the control population expressed both CTCF isoforms (CTCF130, CTCF180) and the treated expressed the CTCF180 only. The treatment, with hydroxyurea and nocodazole, caused cell-cycle arrest in the G2-M phase and induced a switch in the CTCF isoforms.

In chapter 3, the polyclonal and the monoclonal CTCF antibody were used in ChIP-seq that revealed the CTCF binding sites across the genomes of control and treated cells. In the following chapter 4, the global mRNA expression profiles of the two conditions were compared.

The aim of this chapter is to integrate the results from the results from these chapters. We hypothesize that there is correlation between CTCF binding at specific sites and changes in the expression of the associated genes. The investigation will focus on genes which:

1. Were bound by CTCF in the control and/or treated cells and
2. Were differentially expressed between the control and treated 226LDM cells

The analysis will be performed by intersecting the datasets produced from each of the two experiments and the account of the up and down-regulated CTCF targets will be given.

5.3 **Results**

In this study, the differentially expressed transcripts identified in the control and treated 226LDM cells by RNA-Seq (Chapter 4) were linked to the binding of CTCF isoforms revealed by ChIP-Seq (Chapter 3) (p-value filtered results). The CTCF binding sites within 1000bp of the transcription start site (TSS) of genes were selected for these analyses (p value threshold 0.05). Intersection of these data showed that multiple links appear to exist between changes in gene expression and the loss or gain of CTCF binding, as illustrated in the Venn diagram (figure 5-1).

Out of 2051 CTCF binding sites in control 226LDM cells identified using the polyclonal CTCF antibody (which recognizes both isoforms, CTCF130 and CTCF180), 1009 were associated with differentially expressed transcripts. Among those, 520 were up-regulated while the remaining 489 were down-regulated.

From 64 binding sites identified in the treated cells (CTCF180 only), 26 were associated with genes that were differentially expressed; 16 were up-regulated and 10 were down-regulated. There were 8 common CTCF binding sites between control and treated cells in the up-regulated and 6 in the down-regulated group (figure 5-1).

Aiming to evidence more robustly the true targets, the more stringent Q value was used to generate the lists of targets shown in this chapter. Using a *q* value threshold of 0.4, we found 180 targets bound by CTCF whose expression changed in the treated cells. In particular, 83 targets were up-regulated (2 of them remained bound in the treated condition) while 97 were down-regulated (3 remained bound in treated cells).

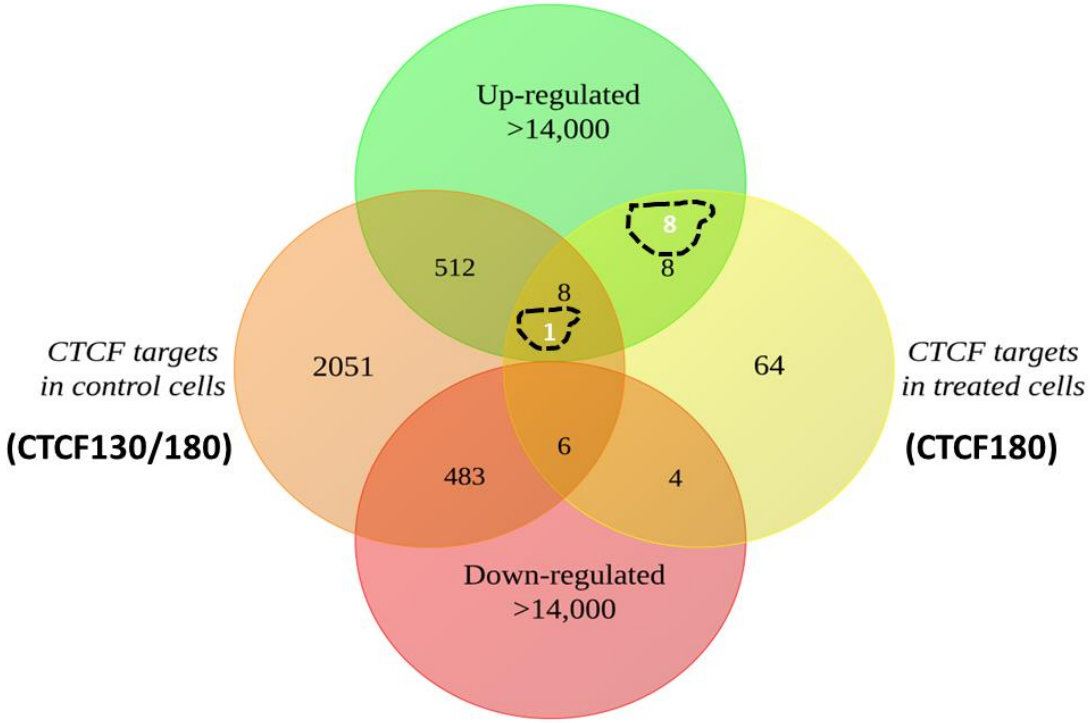


Figure 5-1 Venn Diagram presenting the differentially expressed CTCF binding targets in control and treated 226LDM cells

ChIP and RNA sequencing was conducted in control and treated 226LDM cells. Intersection of the sequencing results revealed the binding targets of CTCF (CTCF130/180 in control and CTCF180 in treated cells, identified with the polyclonal antibody), which were differentially expressed in the treated cells. Number of CTCF130 binding sites in control cells is indicated as a white number in the dashed shape in the intersections. (See main text for detailed explanation).

5.3.1 CTCF association with up-regulated genes in 226LDM cells

By intersecting ChIP and RNA-seq data, filtered by Q value, we discovered that 85 genes occupied by CTCF in control 226LDM cells were significantly up-regulated in treated cells. Among these genes, 2 were also targets both in control and treated cells (table 5.1).

The remaining 83 binding sites were associated with increased expression in the treated cells coupled with the loss of CTCF binding. The top 50 transcripts that displayed increased expression are listed in table 5.2 in order of the decreasing fold change. Based on this and pending experimental validation, their expression is potentially influenced directly or indirectly by CTCF binding.

Table 5-1 Up-regulated targets bound by CTCF in both control and in treated 226LDM cells

Transcript ID	Gene Symbol	Fold Change
1. S100A13:NM_001024213	S100A13	6.677046
2. S100A13:NM_001024212	S100A13	6.660637

Table 5-2 Top 50 targets associated with CTCF whose expression was up-regulated in treated 226LDM cells and binding was lost

Transcript ID	Gene Symbol	Log2 FoldChange
1. UPK2:NM_006760	UPK2	5,162216
2. LCN15:NM_203347	LCN15	3,42951
3. HPD:NM_001171993	HPD	2,654225
4. TTC39A:NM_001297666	TTC39A	2,504616
5. TTC39A:NM_001297667	TTC39A	2,440188
6. STARD10:NM_006645	STARD10	2,407431
7. TP53I3:NM_001206802	TP53I3	2,257445
8. MYOM3:NM_152372	MYOM3	2,145771
9. MZT2A:NM_001085365	MZT2A	2,07671
10. UNC13D:NM_199242	UNC13D	2,070984
11. AKR1B1:NM_001628	AKR1B1	2,061382
12. PSG4:NM_001276495	PSG4	2,032634
13. PSG4:NM_213633	PSG4	2,030582
14. PSG4:NM_002780	PSG4	2,029551
15. RPL39L:NM_052969	RPL39L	1,926199
16. CLTB:NM_001834	CLTB	1,891354
17. CLTB:NM_007097	CLTB	1,890508
18. UQCC2:NM_032340	UQCC2	1,846334
19. KLC3:NM_177417	KLC3	1,780857
20. RPL7:NM_000971	RPL7	1,764561
21. SLC27A5:NM_012254	SLC27A5	1,746649
22. CEACAM1:NM_001712	CEACAM1	1,698591
23. CEACAM1:NM_001024912	CEACAM1	1,697494
24. CEACAM1:NM_001205344	CEACAM1	1,697494
25. C15orf40:NM_001160115	C15orf40	1,679434
26. C15orf40:NM_001160116	C15orf40	1,667749
27. CEACAM1:NM_001184813	CEACAM1	1,654582
28. CEACAM1:NM_001184816	CEACAM1	1,653375
29. CEACAM1:NM_001184815	CEACAM1	1,652473
30. SMAGP:NM_001033873	SMAGP	1,650975
31. HPCAL1:NM_002149	HPCAL1	1,629145
32. MESP1:NM_018670	MESP1	1,625055
33. C14orf2:NM_004894	C14orf2	1,625011

34. C14orf2:NM_001127393	C14orf2	1,62355
35. HPCAL1:NM_134421	HPCAL1	1,606879
36. MRPL23:NM_021134	MRPL23	1,598766
37. ISY1-RAB43:NM_001204890	ISY1-RAB43	1,510335
38. ZNF425:NM_001001661	ZNF425	1,48103
39. C15orf40:NM_001160113	C15orf40	1,46562
40. MRPL38:NM_032478	MRPL38	1,43757
41. NOP10:NM_018648	NOP10	1,423719
42. NABP2:NM_024068	NABP2	1,412168
43. C15orf40:NM_001160114	C15orf40	1,401653
44. NME1-NME2:NM_001018136	NME1-NME2	1,400934
45. C10orf54:NM_022153	C10orf54	1,379281
46. C15orf40:NM_144597	C15orf40	1,349093
47. C11orf1:NM_022761	C11orf1	1,343673
48. SFN:NM_006142	SFN	1,312197
49. RPL26L1:NM_016093	RPL26L1	1,227189
50. FIBP:NM_004214	FIBP	1,222373
51. FIBP:NM_198897	FIBP	1,217752

5.3.2 CTCF association with down-regulated genes in 226LDM cells

The intersection of ChIP and RNA-seq (Q value filtered) revealed that 97 targets were down-regulated in the treated cells. Out of these 97 transcripts, 3 were also associated with CTCF180 in treated cells (table 5-3). The remaining 94 were not associated with CTCF in treated cells; top 50 of these transcripts are shown in table 5.4, ranked according to fold change.

Table 5-3 List of down-regulated targets bound by CTCF both in control and treated 226LDM cells

Transcript ID	Gene Symbol	Log2FoldChange
1. SGPL1:NM_003901	SGPL1	-2,3565
2. TCTN3:NM_001143973	TCTN3	-1,87127
3. S100A1:NM_006271	S100A1	-1.0968

Table 5-4 Top 50 CTCF targets whose expression was down-regulated in treated 226LDM cells and binding was lost

Transcript ID	Gene Symbol	Log2 FoldChange
1. ARL6IP1:NM_015161	ARL6IP1	-3,92177
2. SLC6A15:NM_018057	SLC6A15	-3,43998
3. KCNIP2:NM_173194	KCNIP2	-3,34652
4. LTB4R2:NM_001164692	LTB4R2	-3,03998
5. JRK:NM_003724	JRK	-3,0288
6. LTB4R2:NM_019839	LTB4R2	-3,02425
7. SLC6A15:NM_001146335	SLC6A15	-2,91716
8. SLC6A15:NM_182767	SLC6A15	-2,90977
9. MAT2A:NM_005911	MAT2A	-2,87756
10. JRK:NM_001077527	JRK	-2,83656
11. GSTA4:NM_001512	GSTA4	-2,79356
12. JRK:NM_001279352	JRK	-2,63189
13. SAMD5:NM_001030060	SAMD5	-2,43252
14. COL27A1:NM_032888	COL27A1	-2,33791
15. KIAA1919:NM_153369	KIAA1919	-2,19629
16. KMT2A:NM_001197104	KMT2A	-2,07487
17. KMT2A:NM_005933	KMT2A	-2,07487
18. DST:NM_001144770	DST	-2,06823
19. DST:NM_183380	DST	-2,06722
20. XYLB:NM_005108	XYLB	-1,94564
21. XXYLT1:NM_152531	XXYLT1	-1,90918
22. LPGAT1:NM_014873	LPGAT1	-1,8628
23. GDPD1:NM_001165993	GDPD1	-1,74502

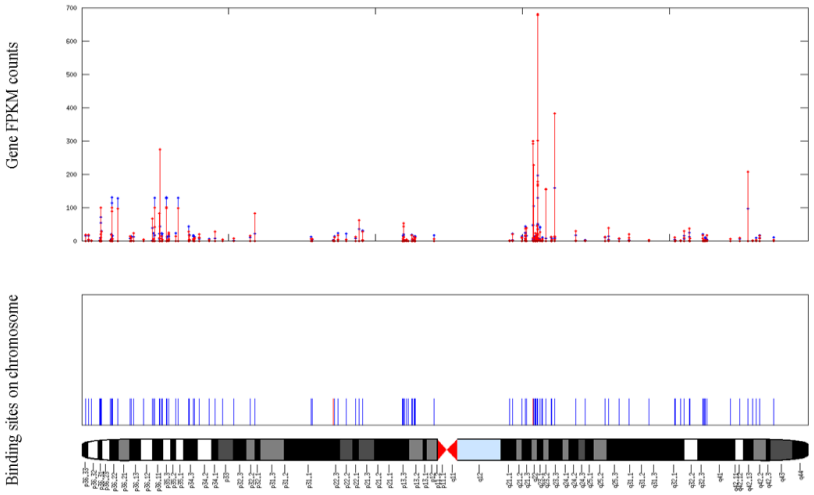
24. LRIG2:NM_014813	LRIG2	-1,73306
25. ABCC5:NM_001023587	ABCC5	-1,6847
26. PIGO:NM_032634	PIGO	-1,65188
27. GLTPD2:NM_001014985	GLTPD2	-1,64126
28. PIGO:NM_001201484	PIGO	-1,63659
29. PIGO:NM_152850	PIGO	-1,63423
30. ARRDC3:NM_020801	ARRDC3	-1,58151
31. ZNF783:NM_001195220	ZNF783	-1,54794
32. SYNM:NM_015286	SYNM	-1,534
33. SYNM:NM_145728	SYNM	-1,51835
34. SLC3A2:NM_001013251	SLC3A2	-1,48721
35. PVRL1:NM_002855	PVRL1	-1,47781
36. PGBD2:NM_170725	PGBD2	-1,43834
37. TMTC2:NM_152588	TMTC2	-1,42514
38. TM2D3:NM_025141	TM2D3	-1,40488
39. ABCC5:NM_005688	ABCC5	-1,40468
40. CREB3L2:NM_194071	CREB3L2	-1,40402
41. EGR1:NM_001964	EGR1	-1,40144
42. TM2D3:NM_078474	TM2D3	-1,40017
43. PTPRU:NM_005704	PTPRU	-1,37153
44. PTPRU:NM_001195001	PTPRU	-1,37126
45. PTPRU:NM_133178	PTPRU	-1,37126
46. PTPRU:NM_133177	PTPRU	-1,37119
47. SLC29A3:NM_001174098	SLC29A3	-1,33866
48. SLC29A3:NM_018344	SLC29A3	-1,33866
49. RBM4B:NM_001286135	RBM4B	-1,29116
50. GEMIN4:NM_015721	GEMIN4	-1,28311

5.3.3 The CTCF binding sites associated with the alteration in gene expression in control and treated cells are distributed in a non-uniformed manner in all chromosomes

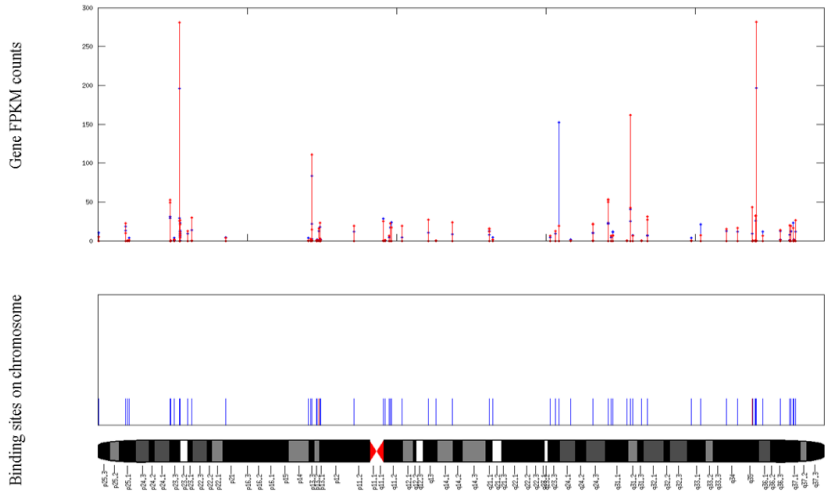
The results described previously suggest that in our model of proliferating and cell-cycle blocked 226LDM cells, the binding of CTCF can affect expression of the genes associated with it; the same applies for the loss of CTCF binding or change in the PARylation status.

Interestingly, there are chromosomes where this causal relationship appears to be more pronounced, such as chromosome 17 or 20. On other chromosomes (e.g. chromosome 7 and 11) this effect is less prominent. These spatial arrangements and clustering of binding sites, which may be unique for our model and dynamically change depending on the treatment that the cells undergo, are presented in figure 5.2. The graphs were generated using MATLAB.

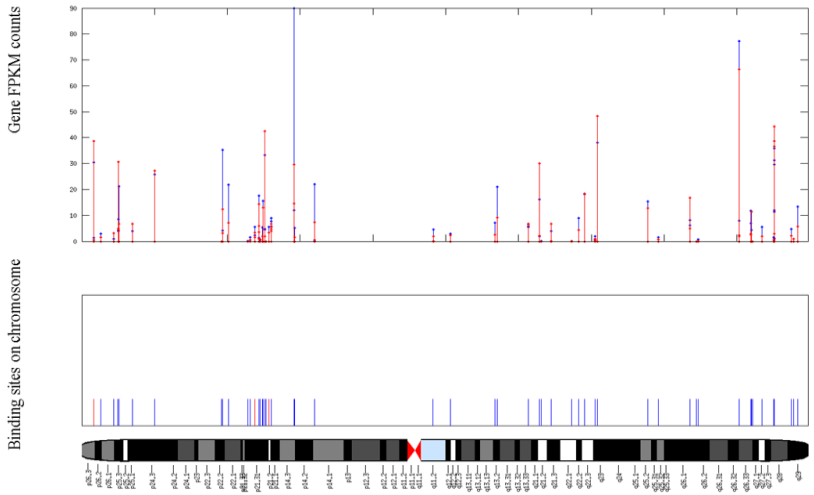
Gene expression on chromosome 1



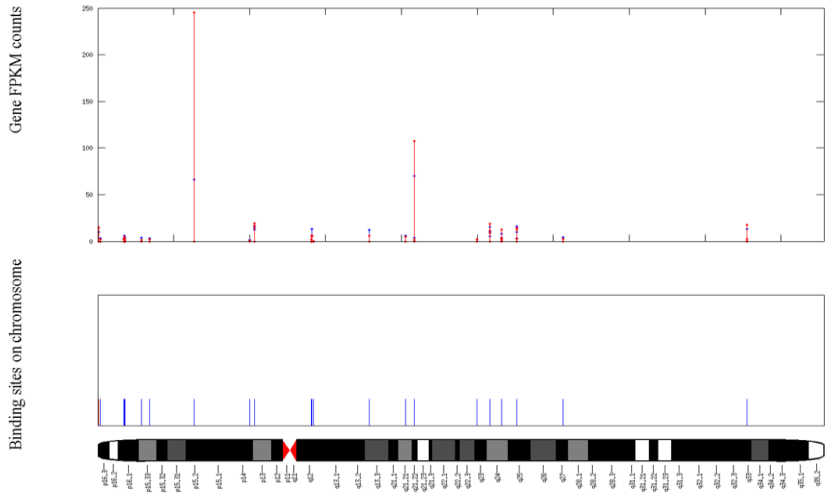
Gene expression on chromosome 2



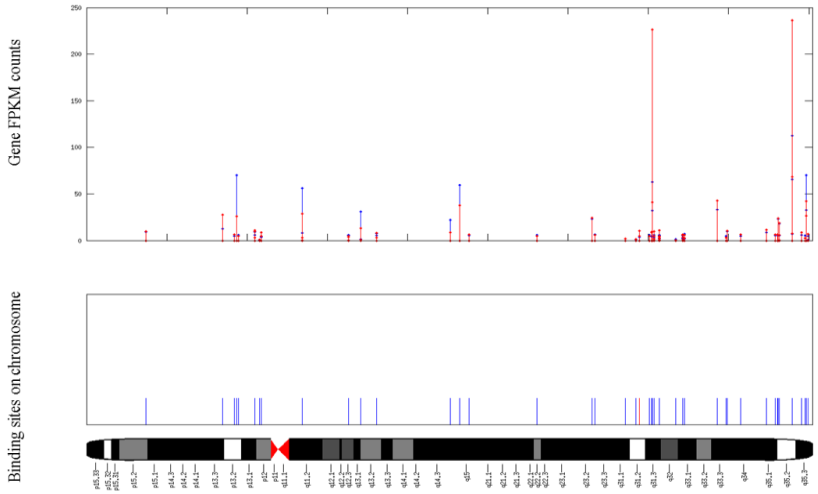
Gene expression on chromosome 3



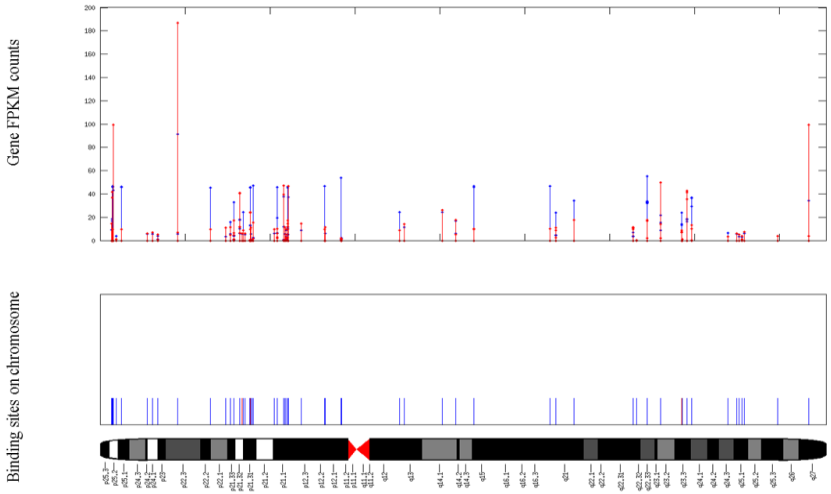
Gene expression on chromosome 4



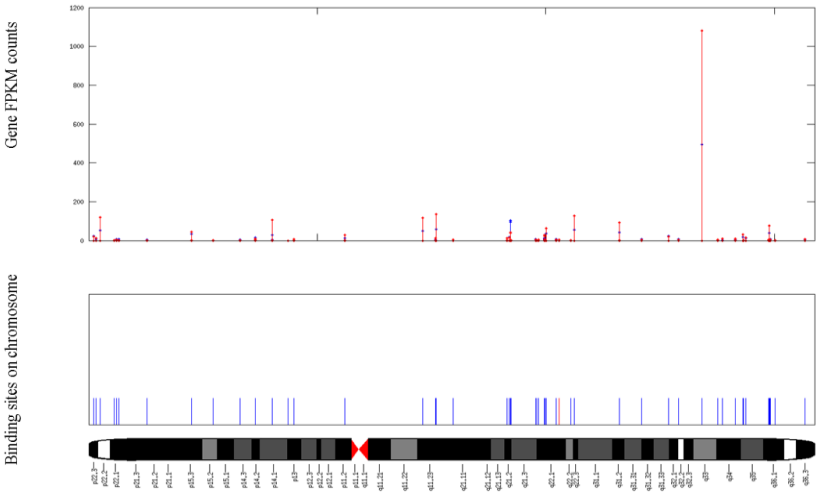
Gene expression on chromosome 5



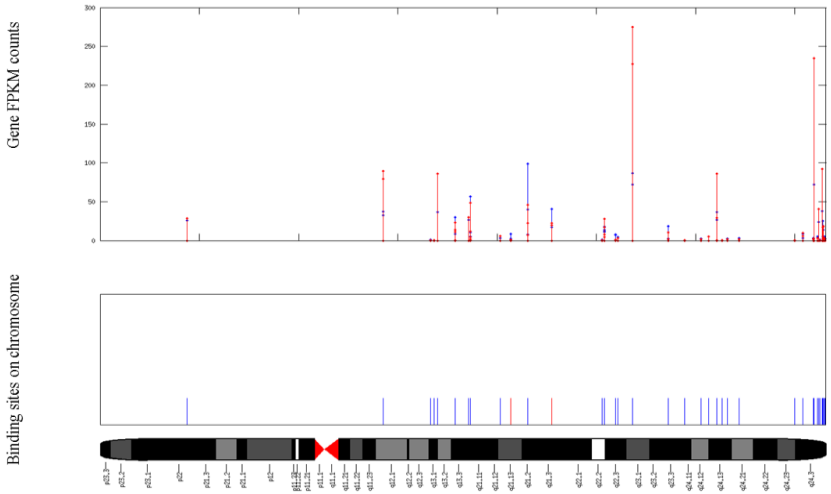
Gene expression on chromosome 6

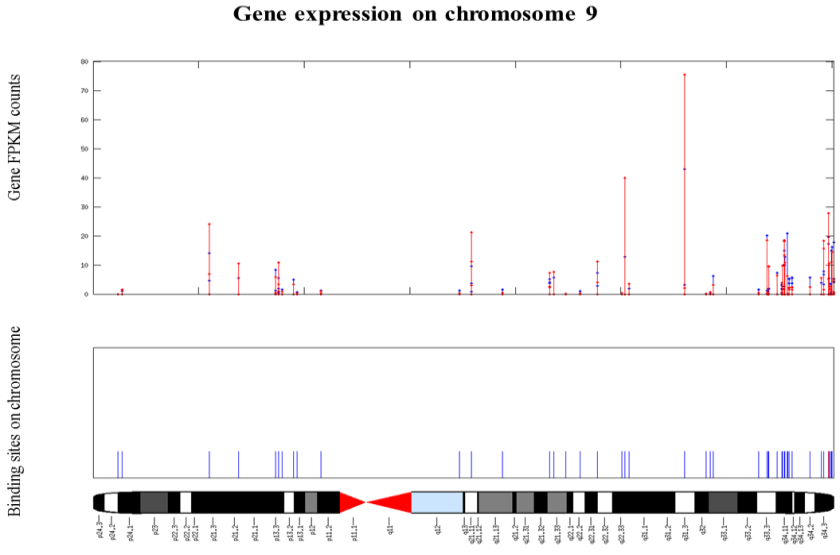


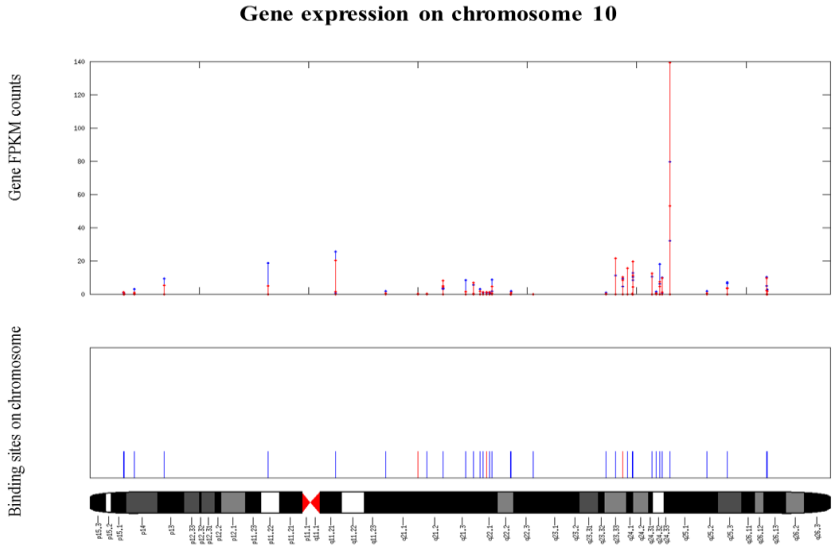
Gene expression on chromosome 7

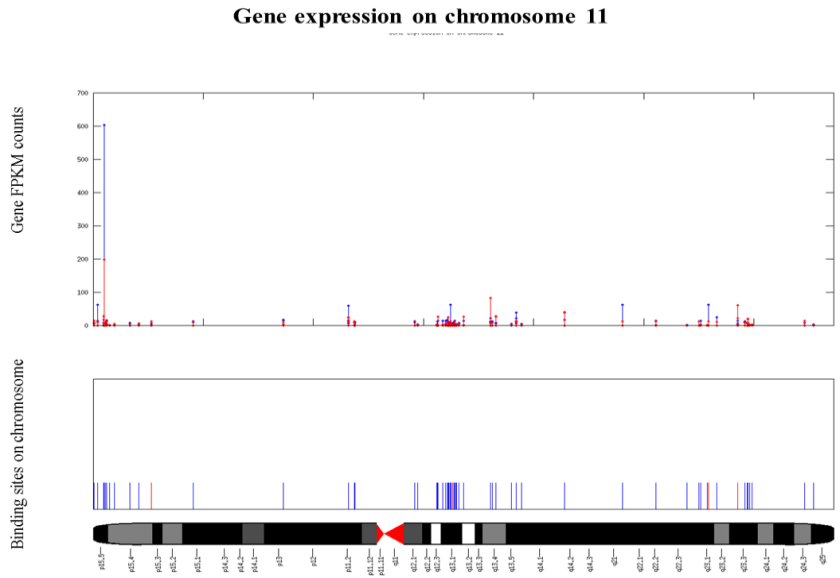


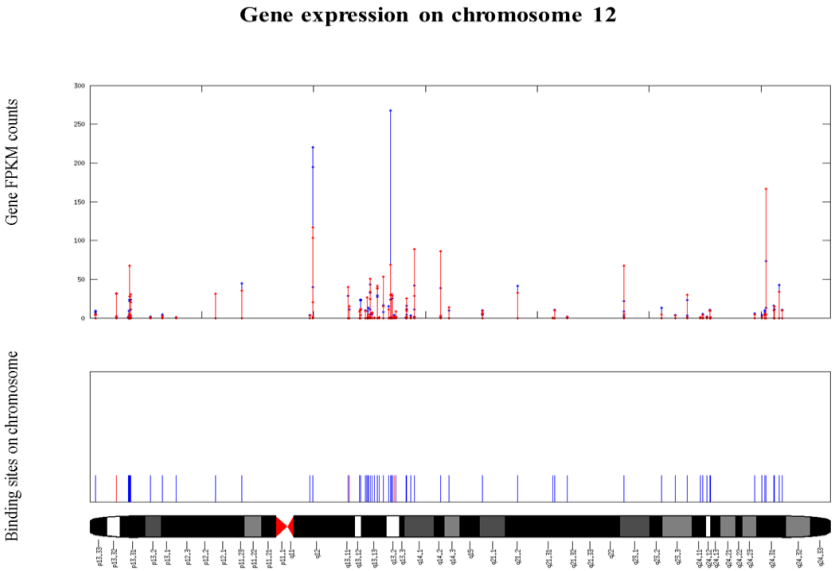
Gene expression on chromosome 8



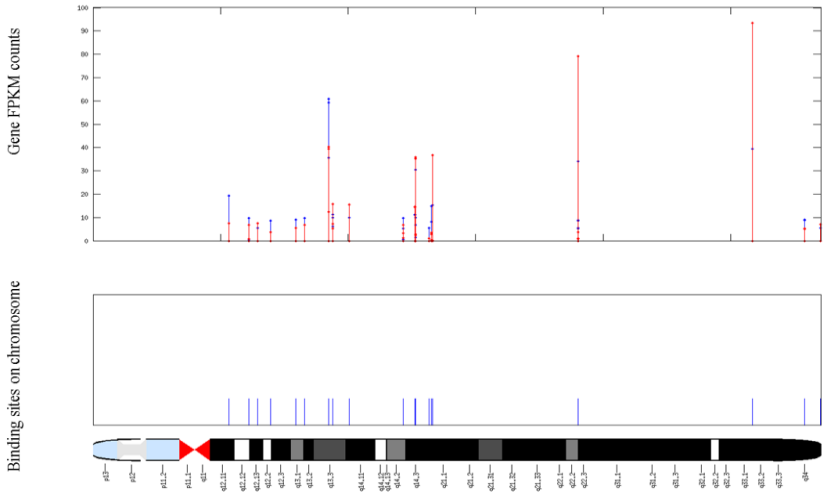

Gene expression on chromosome 9
Gene FPKM counts
Binding sites on chromosome


Gene expression on chromosome 10
Gene FPKM counts
Binding sites on chromosome

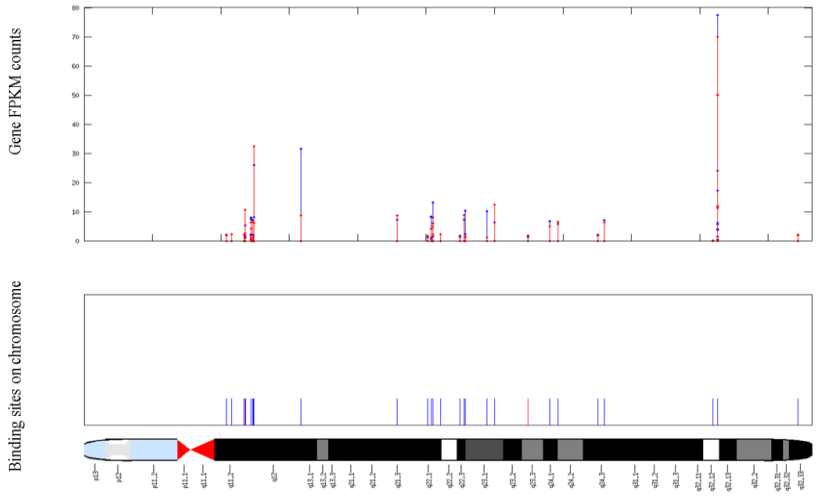

Gene expression on chromosome 11
Gene FPKM counts
Binding sites on chromosome


Gene expression on chromosome 12
Gene FPKM counts
Binding sites on chromosome

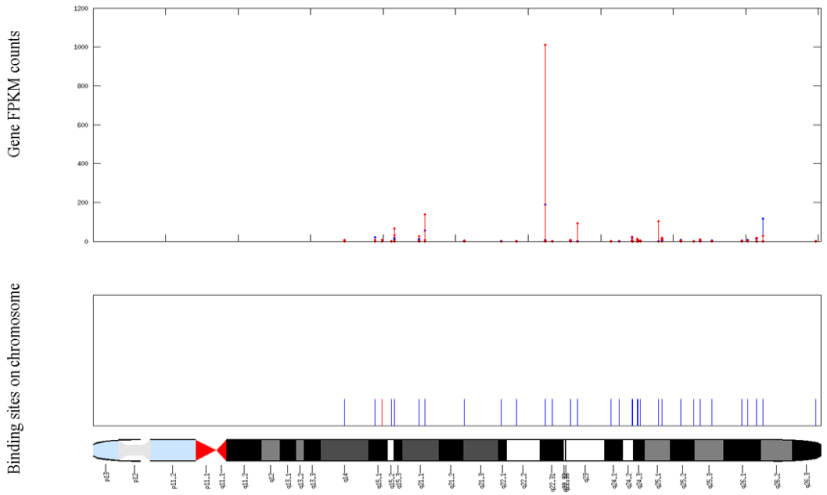
Gene expression on chromosome 13



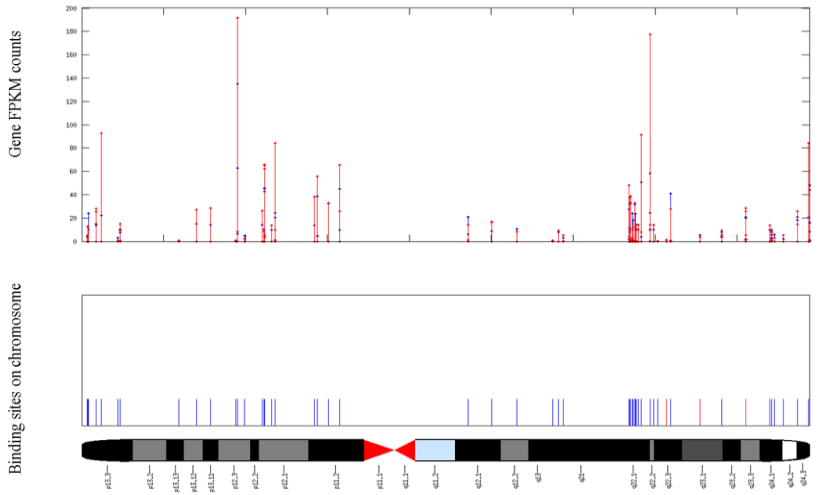
Gene expression on chromosome 14



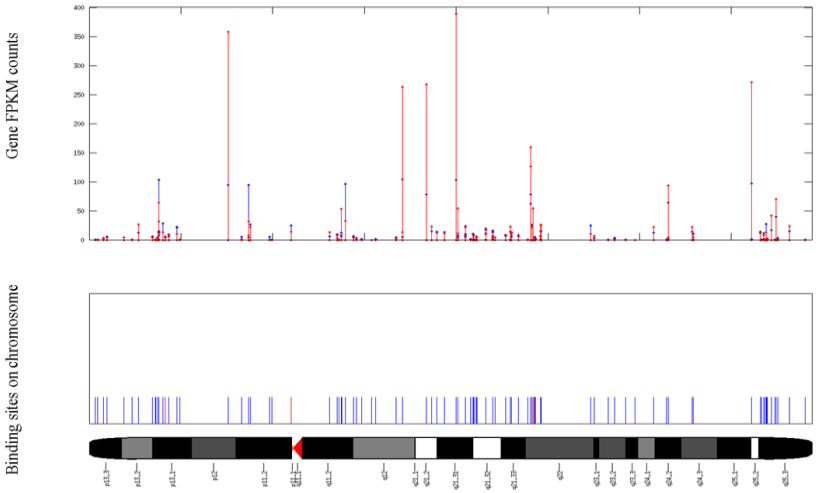
Gene expression on chromosome 15



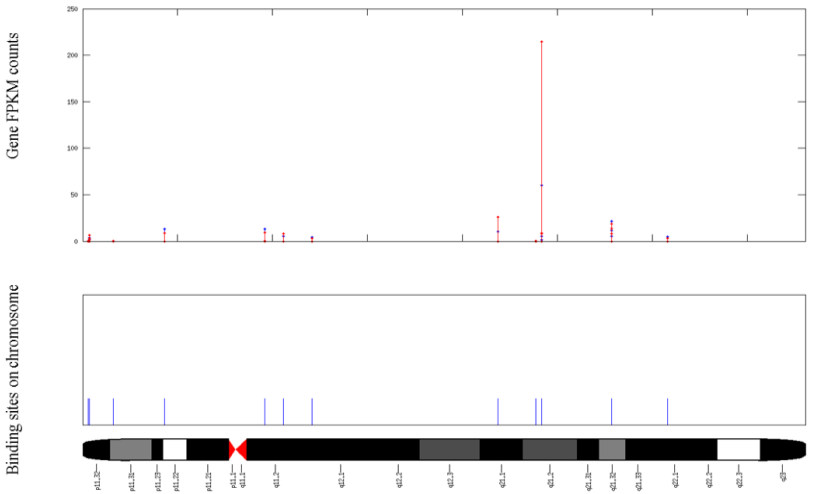
Gene expression on chromosome 16



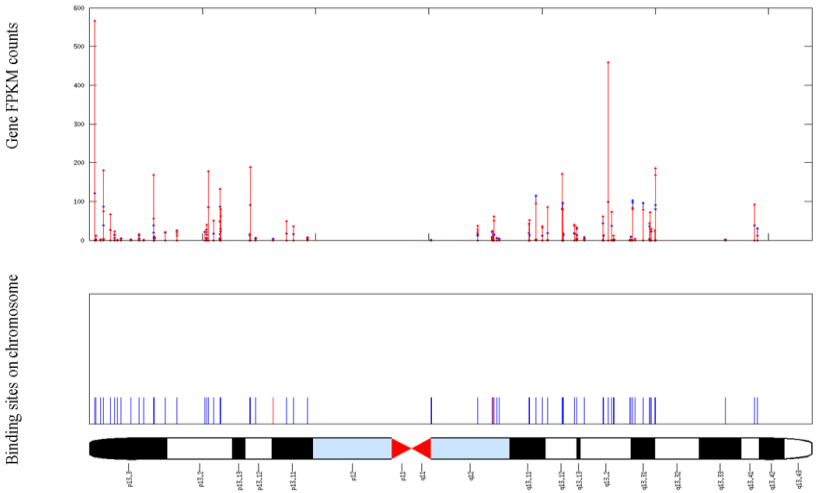
Gene expression on chromosome 17



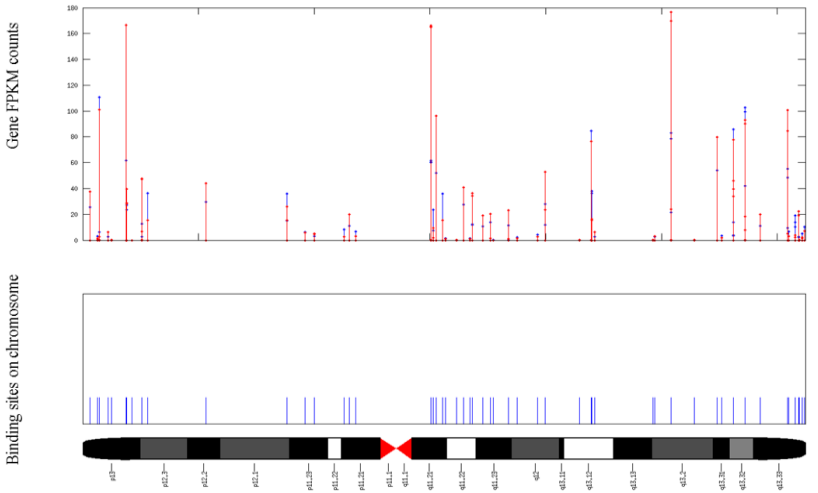
Gene expression on chromosome 18



Gene expression on chromosome 19



Gene expression on chromosome 20



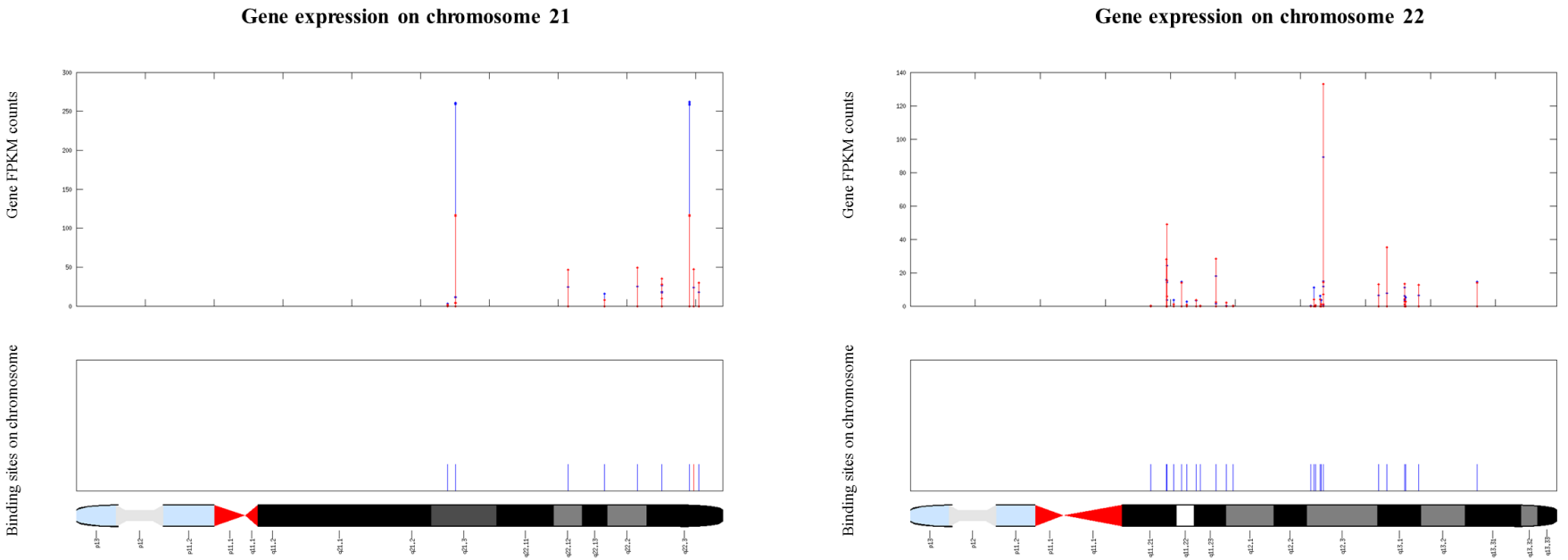


Figure 5-2 Diagrammatic representation of CTCF binding events on chromosomes and the effect on mRNA expression

ChIP and RNA sequencing experiments were conducted in a cell line model generated from control and treated 226LDM cells. The treated cells underwent a hydroxyurea and nocodazole treatment which led to cell-cycle blockage and a switch to the expression of CTCF180 only, in comparison to control cells which express both known CTCF isoforms, namely CTCF130 and CTCF180. The binding events discovered with the polyclonal CTCF antibody (which recognizes both isoforms) are shown on top of the chromosomes. Blue lines represent the binding targets in control cells while red lines represent the binding in treated cells. For each binding, the expression level (in the form of FPKM counts) is shown; blue represents the expression in control cells while red represents the expression in treated cells. The graphs were prepared using MATLAB.

5.4 Discussion

Previously, a cell line model was generated using control 226LDM cells together with a population that underwent treatment with hydroxyurea and nocodazole. This treatment which induces cell cycle arrest also causes a transition in CTCF PARylation status in these cells. Control cells express both known isoforms, CTCF130 and CTCF180, while in the treated cells, only CTCF180 is detected.

This cell model was used in ChIP-seq experiments that were aimed to identify binding sites of CTCF in both cell populations. Using the polyclonal CTCF antibodies, sites bound to both CTCF130 and CTCF180 isoforms of CTCF could be selected; the monoclonal antibody was used to precipitate sites bound to CTCF130 (described in Chapter 3). The same two cell populations were employed to generate global RNA profiles (described in Chapter 4).

In this chapter, we presented the intersection of the two datasets. Combining these next generation sequencing assays is a very important step in extracting biological value from the rich and complex data that each of them generates. This allows to link binding of a particular transcription factor (CTCF in this study) to gene expression.

By comparing the ChIP-Seq and RNA-Seq data (filtered by p-value), 1009 CTCF sites discovered in control cells, approximately half of the total binding sites, were associated with differential gene expression in treated cells. More specifically, over 500 were up-regulated and just under 500 were down-regulated as a result of losing CTCF binding. With the exception of only 14 targets (8 for the up- and 6 for the down-regulated) all targets lost binding of CTCF in treated cells.

In treated cells, the number of discovered CTCF binding events was considerably smaller that in control cells (64 targets compared to the 2051 in the control). From these, 26 were found significantly affected by the treatment and 14 of them were commonly bound in control cells (8 of the up-regulated and 6 of the down-regulated targets).

Notably among the sites associated with up-regulated genes in treated cells, there were 8 binding sites (all associated with the same gene) that were also bound by the CTCF130 in the control cells. This gene, FXYD3, codes for a protein that belongs to a small family of proteins which contain an FXYD sequence and may regulate the function of ion-pumps and ion-channels (Pruitt et al., 2014). The protein has also been shown to demonstrate interesting yet varying expression changes in various cancers including gastric adenocarcinoma, lung and breast cancer (Okudela et al., 2009, Yamamoto et al., 2009, Zhu et al., 2010).

On the other hand, 10 down-regulated targets were discovered and four of these were not present in any other of the datasets; AKAP5, SCAF11, SCAP and DUSP11.

AKAP5 encodes the A kinase anchor protein 5, member of a family of proteins which facilitate signal transduction by guiding enzymes in the proximity of their substrates, mostly known to do so for protein-kinase-A. Members of this family are involved in various functions including cell cycle and cancer progression (Esseltine and Scott, 2013, Saini et al., 2013).

SCAF11, also known as SIP1, is reportedly involved in mRNA splicing events (Zhang and Wu, 1998). SCAP is a protein with a function as a regulator of cholesterol biosynthesis (Sun et al., 2007). DUSP11 is a target of the p53 transcription factor, it is reportedly activated after DNA damage and is also involved in RNA splicing events and cell growth (Caprara et al., 2009).

Following further filtering of the target lists by using the more stringent Q value, the number of hits was reduced as expected. We found 83 up-regulated targets, out of which only 2 were still binding in the treated cells. This means that, for the remaining 81, loss of CTCF binding lead to increased expression. On the other hand, 97 targets were down-regulated; in the majority of cases (94) the change seems to be associated with loss of CTCF binding.

The genes described above and those associated with the rest of the binding sites discussed in this chapter are involved in a wide spectrum of functions. Due to time restrictions it was not possible to analyze in depth all the pathways that the genes are involved in. For the same reason,

the validation of our findings by molecular methods in the laboratory was not performed. Taking into consideration the underlying limitations of both sequencing techniques this validation will be necessary. Pending experimental validation, our study suggests that CTCF binding can be essential for the regulation of key genes involved in a variety of functions linked to different biological situations.

Chapter 6 CTCF PARylation and DNA Damage Response

6.1 Introduction

6.1.1 DNA Damage Response (DDR)

Cells need to maintain the integrity of their genetic information and to pass this information to the daughter cells in an accurate and controlled manner. This genetic inheritance can be compromised by damage in the DNA of the mother cells caused by errors in the internal replication machinery, the cellular environment or by external events and agents. Events of this nature are not a rarity and therefore it is important to have prompt and efficient mechanisms in place to correct the errors in DNA or eliminate cells in case of substantial damage. The pathways that are activated as a reaction to damage are collectively referred to as the DNA Damage Response (DDR) (for reviews see Zhou and Elledge (2000) and Ljungman (2010)).

The front runner in the protection system against damage is a multi-component signaling network, responsible for the detection of cellular structure abnormalities by using specific checkpoints. The cellular responses after the initial detection of the damage are common between all eukaryotic cells and include cell cycle arrest and/or block of the DNA replication process (Pieper et al., 1999, Zhou and Elledge, 2000, Jackson and Bartek, 2009).

6.1.2 PARylation is involved in DNA Damage Response

PARylation caused mainly by PARP1, but also PARP2, is one of the first events following a lesion in DNA. PARP1 acts as a sensitive “nick sensor” and at the same time as a recruiter of “repairmen” to the disrupted area (Malanga and Althaus, 2005, Kim et al., 2005a). Upon discovering a lesion, the activated PARP1 modifies itself and recruits important repair factors such as XRCC1, DNA ligase III and others (El-Khamisy et al., 2003, Malanga and Althaus, 2005).

The importance of PARylation in DDR has been proved in experiments where PARP1-knockdown cells exhibit defective repair capacity and hyper-sensitivity to damaging agents (Pieper et al., 1999, Malanga and Althaus, 2005).

6.1.3 CTCF isoforms and DNA Damage Response

A growing body of evidence support the hypothesis that CTCF could be involved in DDR. Firstly, CTCF has been a documented factor in the regulation of cellular processes such as cell growth, proliferation and cell cycle progression (Heath et al., 2008), all of which are affected during DDR. Furthermore, CTCF has been linked with key proteins, including p53 and cohesin, which are involved in DNA damage repair (Ohlsson et al., 2001, Millau and Gaudreau, 2011).

Of particular relevance for this study is the fact that CTCF has been linked with PARP1 and can be PARylated (Yu et al., 2004b, Farrar et al., 2010). An investigation by Guastafierro et al. (2008) revealed another link between the two proteins by showing that CTCF can activate PARP1 auto-modification. The potential significance of this modification of CTCF in DDR is the main focus of this study.

6.2 Experimental Aims

The main hypothesis which this study will address is that the PARylation of CTCF could be important or required in the context of DDR. The underlying assumption implies that the activated PARP pathway is responsible for the recruitment and PARylation of CTCF at the damaged sites. Supposing that this is the case, CTCF could then function in cellular defense against instability and damage.

The first objective of the work described in this chapter is to establish that DNA damage affects CTCF with regards to “how much” and “where” it is in the cell. It will be pursued by determining the expression (western blot) and localization patterns (immunofluorescence

staining) of CTCF following the induction of small, non-lethal lesions to cells treated with the DNA damaging agent hydrogen peroxide (H_2O_2).

The second objective of the work is to investigate the importance of CTCF PARylation during DDR. To achieve this, two experimental tools will be used: (1) the mutant variant of CTCF deficient for PARylation and (2) PARP inhibitors.

In the first case, EGFP plasmids containing the PARylation-deficient CTCF constructs, previously produced by Farrar et al. (2011) will be transfected into cells and the response to damage will be compared to that of cells transfected with the wild-type CTCF. In the second case, the PARP inhibitor ABT-888 will be used to block PARylation activity (Wahlberg et al., 2012) and the results of this inhibition will be monitored in control as well as in H_2O_2-treated cells. The states of CTCF and PARP1 will be monitored by Immunofluorescent staining (IF) and microscopy.

The immunostaining experiments can help with the interpretation of events that take place in individual cells or a small population of cells. To investigate the effects of the treatments in total cell populations, another technique will be employed, namely the automated fluorimetric analysis of DNA unwinding (FADU) assay. This automated technique, discussed in more detail in chapter 1, offers the option of robust, high throughput detection and quantification of DNA damage and repair (Moreno-Villanueva et al., 2011). In our study, FADU will be used to determine the extent of DNA damage in cells treated with a damaging agent when the PARylation of CTCF is blocked through either of the aforementioned methods.

The cell line to be used in the majority of the experiments described in this chapter is 226LDM, derived from normal breast luminal cells and immortalized by viral constructs carrying the modified T antigen TAg (U19dl89-97) and hTERT (Docquier et al., 2009). This cell line was chosen because of its closeness to the healthy normal cell *in vivo*. For purpose of

comprehensiveness and in order to explore whether the nature of the results was global, a further panel of selected normal and cancer cell lines was used for some of the experiments.

6.3 Results

In a series of experiments described below, CTCF cellular distribution was investigated in 226LDM cells treated with low concentrations of H_2O_2, which is a known DNA damaging/oxidizing agent (Imlay et al., 1988, Henle and Linn, 1997). The concentration of H_2O_2 for these experiments was optimized at 200μM; this amount of the reagent on the micro- and milimolar scale is effective in generating small DNA lesions that trigger the activation of a repair mechanism without causing damage serious enough to trigger apoptotic cell death (Nakamura et al., 2003).

6.3.1 The localization pattern of CTCF in 226LDM cells changes after 30 minutes of treatment with H_2O_2, while its expression remains unaltered

In the first set of experiments, the 226LDM cells, seeded in wells of a 12-well plate, were treated with 200 μM of H_2O_2 and samples of these cells were harvested at a range of different time-points. Alongside, non-treated cell samples were used as a control to verify that cell cycle stage/events are not interfering with the interpretation of the results. The cells were prepared for Western Blotting (WB) and Immunofluorescence (IF) staining experiments in order to demonstrate the CTCF expression and localization patterns, respectively.

Western blotting experiments on the treated cells did not portray significant changes in the expression of CTCF or the ratio between the two isoforms in the course of time (figure 6-1). This implies that if CTCF is involved in the damage response mechanism, as hypothesized, then this would occur most likely by utilizing the existing molecules rather than producing new ones.

Although WB revealed no obvious variance in CTCF abundance following the treatment, a change was detected in the localization and distribution patterns of CTCF after 30 minutes of exposure to the damaging agent. As shown in figure 6-2, in control cells and in cells after 15

min and 60 min of treatment CTCF is characterized by generally diffuse distribution in the nucleoplasm, with no considerable co-localization between CTCF and PARP1. However, after 30 min CTCF is detected in the nucleoli as confirmed by the nucleolar marker UBF (Goenechea et al., 1992) (examples are indicated with white arrows). At this time point, extensive overlap between CTCF and PARP1 can be observed (examples are indicated with yellow arrows); this co-localization disappears after 60 min of treatment. Control cells exhibited the same localization for all time-points (data not shown).

The re-localization of CTCF from nucleoplasm into nucleoli had been reported previously shown to be associated with growth inhibition, cell differentiation and apoptosis. Interestingly, CTCF inhibited nucleolar transcription through mechanisms involving CTCF PARylation (Torrano et al., 2006).

To obtain more detailed visual information of these events, the stained cells were also viewed using confocal microscopy. As shown in figure 6-3, both CTCF and PARP1 are dispersed throughout the nuclei of the untreated 226LDM cells, although their distribution patterns differ, whereby CTCF appears diffused in the nucleus and PARP1 is more concentrated on specific areas. However, the distribution pattern of CTCF and PARP1 changes in 226LDM cells after 30 min with H_2O_2, revealing considerable co-localization between these proteins (figure 6-4).

PARP1 (and PARP2) have been reported to accumulate in nucleoli in a complex with B23 (Meder et al., 2005) Moreover, the interactions have been documented between PARP-1 and B23, CTCF and B23 (Yusufzai et al., 2004) between CTCF and PARP1 (Guastafierro et al., 2008) CTCF and UBF (van de Nobelen et al., 2010) . It is therefore possible CTCF may be a part of a large multifunctional protein complex involved in the regulation of nucleolar

function(s) dependent on PARPs. These aspects will be further addressed in the “Discussion” section of this Chapter.

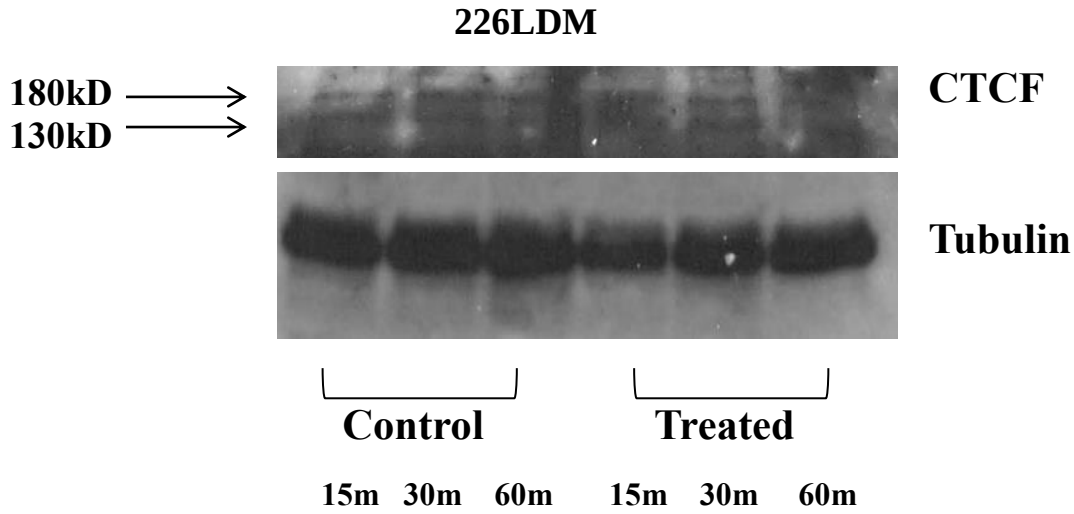


Figure 6-1 Analysis of CTCF expression in non-treated 226LDM cells and in cells treated with H_2O_2

226LDM cells seeded in wells were treated with 200 μM of H_2O_2. After 15, 30 and 60 minutes cells were harvested and samples were prepared for WB. For the same time-points, samples were prepared from control non-treated cells. For the blotting experiment, the polyclonal anti-CTCF antibody which recognizes both of the CTCF isoforms, 130kD and 180kD, was used. Tubulin was used as a loading control.

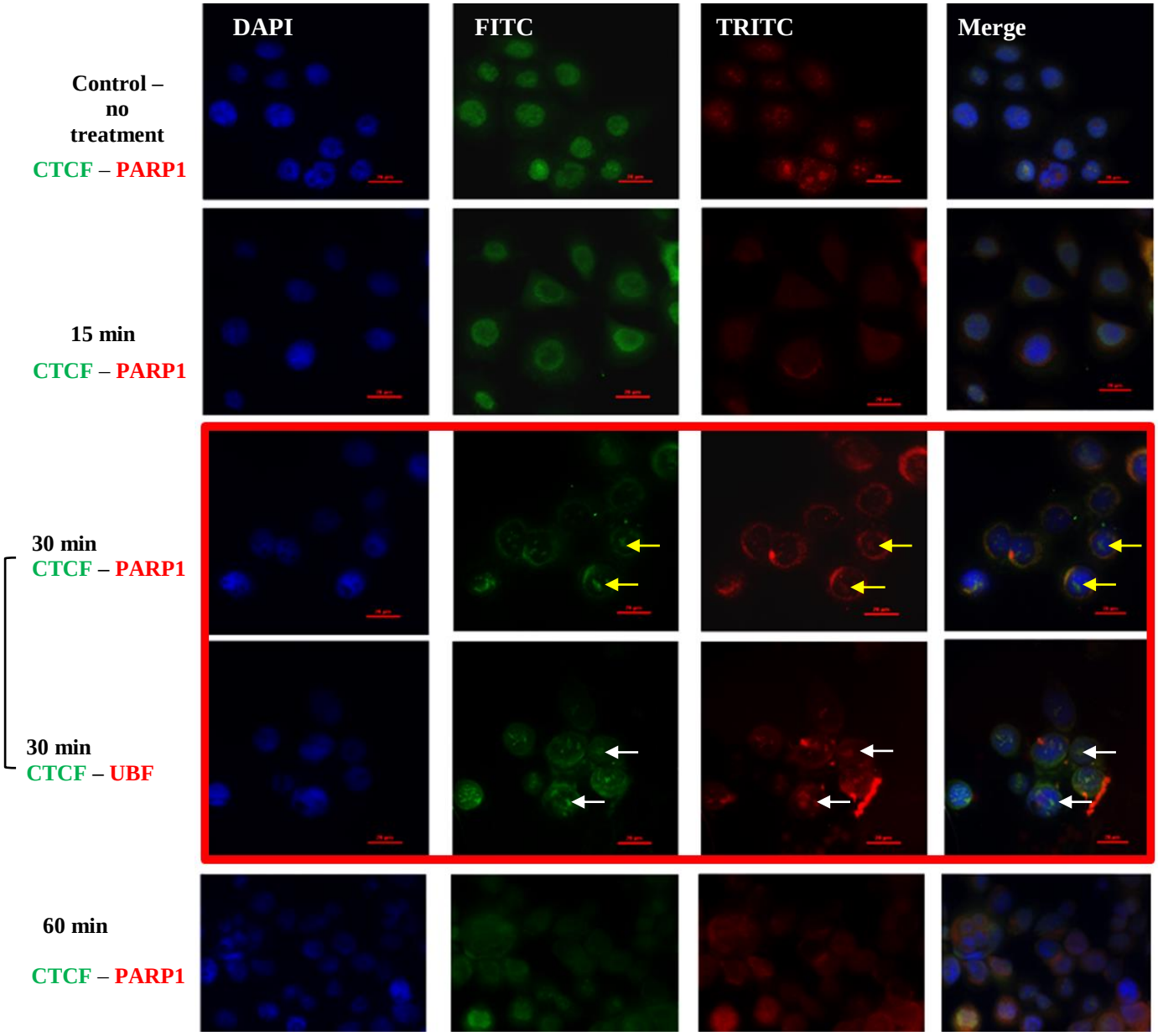


Figure 6-2 Widefield microscopy of 226LDM cells immunofluorescently stained with the anti-CTCF, anti-PARP1 and anti-UBF antibodies

226LDM cells were cultured on glass cover-slips placed in the wells of 12 well-plates, treated withH_2O_2, fixed and prepared for IF staining experiments. *Green (FITC),* staining with the anti-CTCF antibody. R*ed (TRITC),* staining with the anti-PARP1 or –UBF antibodies (as indicated). Nuclei were visualized with 4′,6-diamidino-2-phenylindole (DAPI) staining (*blue*). *Right,* merge of the green and red channels. CTCF co-localization with PARP1 is indicated with yellow arrows and with UBF with white arrows*).* Images were taken at 60x magnification using the Nikon Ti-Eclipse widefield microscope. Scale bar shows 20 μm.

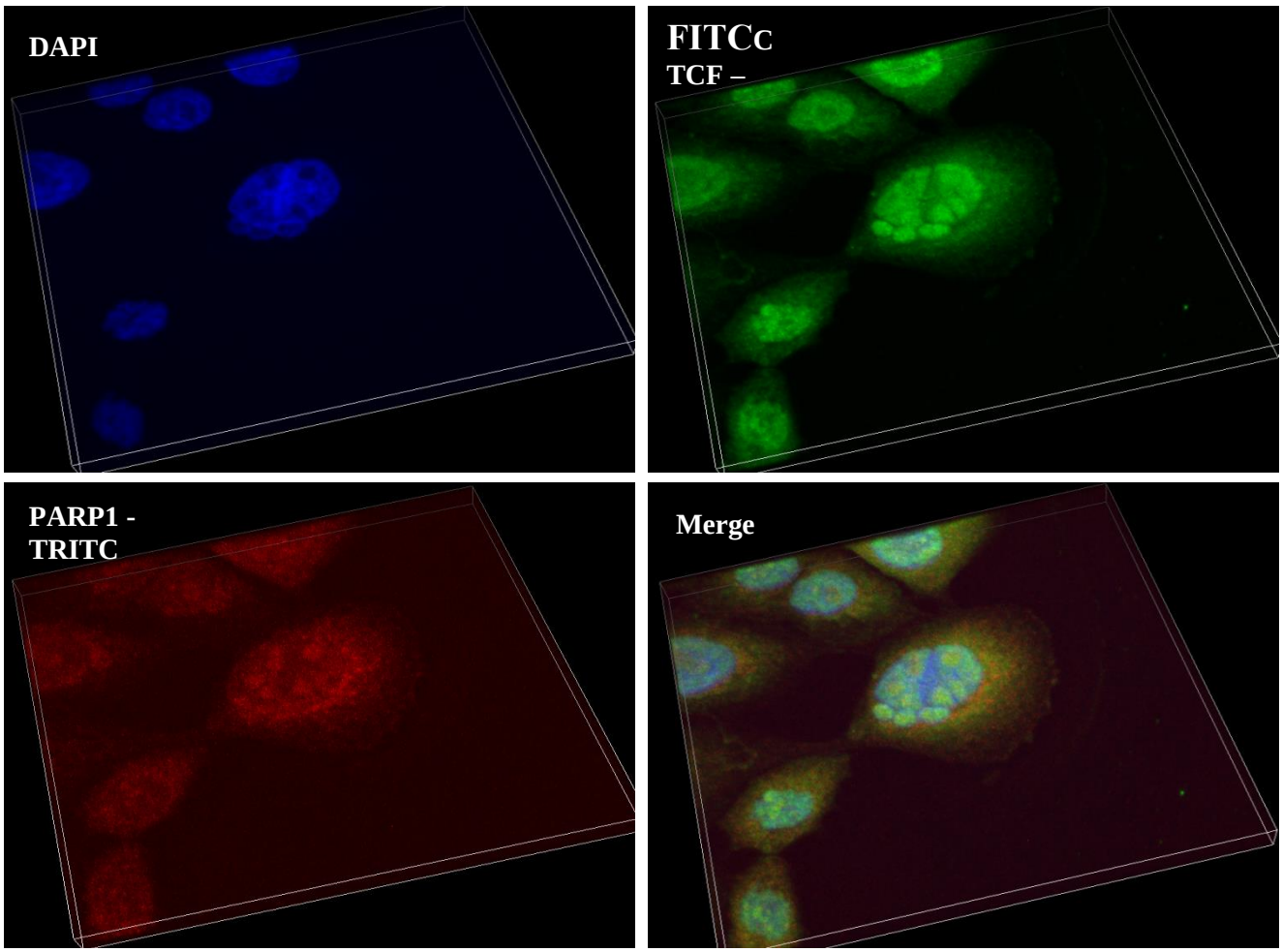


Figure 6-3 Analysis of untreated 226LDM cells, stained with the anti-CTCF and anti-PARP1 antibodies, using confocal microscopy

The 226LDM cells grown under normal conditions without treatment were prepared as described in figure 6.2. *Green (FITC),* staining with the anti-CTCF antibody. R*ed (TRITC),* staining with the anti-PARP1 antibody. Nuclei were visualized with 4′,6-diamidino-2-phenylindole (DAPI) staining (*blue*). Bottom *right,* merge of the green and red channels. Images were taken at 60×magnification using the Nikon A1Rconfocal microscope.

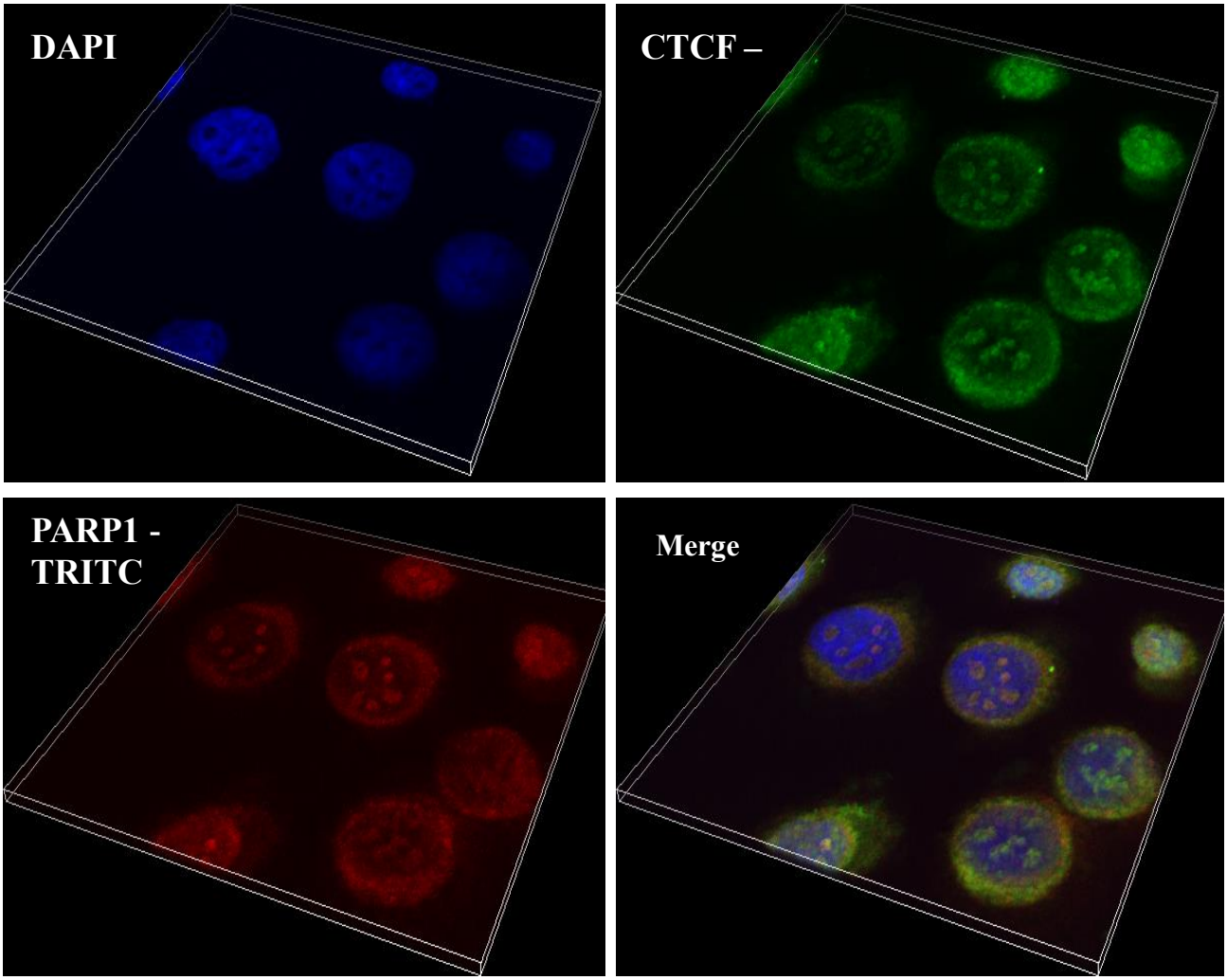


Figure 6-4 Analysis of 226LDM cells after 30 minutes of treatment with H_2O_2 and stained with the anti-CTCF and anti-PARP1 antibodies, using confocal microscopy

The 226LDM cells grown under normal conditions without treatment were prepared as described in figure 6.2. *Green (FITC),* staining with the anti-CTCF antibody. R*ed (TRITC)* , staining with the anti-PARP1 antibody. Nuclei were visualized with 4′,6-diamidino-2-phenylindole (DAPI) staining (*blue*). Bottom right, merge of the green and red channels. Images were taken at 60×magnification using the Nikon A1Rconfocal microscope.

6.3.2 CTCF is detected in the nucleoli co-localizing with PARP1 in response to treatment with H_2O_2 in a panel of normal-immortalized cell lines, whereas these features were not observed in cancer cell lines.

From the experimental results described in the previous section, it can be inferred that in 226LDM cells treated with H_2O_2 there is an obvious change in the localization from nucleoplasm to nucleoli rather than in the amount of CTCF, and in these cells CTCF co-localizes with PARP1(figures 6.1-6.2). To establish whether this phenomenon was cell line specific or if this is a global event, the experiment was conducted using a panel of cell lines with different characteristics.

Firstly, experiments were carried out using ZR-75.1 (breast cancer cells) (figure 6-5), HeLa (cervical cancer cells) (figure 6-6) and 293T cells (transformed human kidney embryonic cells) (figure 6-7). In all cases, the cells were grown on coverslips, treated with 200 μM H_2O_2, fixed and stained following the same protocol as for the 226LDM cells in the previous section. Several post –treatment time-points were selected, however the distribution patterns of CTCF and PARP1 did not resemble those in 226LDM cells at any of these points.

On the other hand, in the experiments conducted with Hs27 (normal foreskin fibroblast) and BPH-1 (benign prostate hyperplasia) cells, 5 min post-treatment, CTCF appeared in the nucleoli, co-localizing with PARP1. 226LDM cells (figures 6.8-6.9).This can be explained by the fact that the immortalized 226LDM, Hs27 and BPH-1 cells were generated from primary cells and have characteristics of normal cells. The earlier response time may reflect different origin of these cells.

From these experiments it can be concluded that, in response to treatment with H_2O_2, CTCF is detected in the nucleoli co-localizing with PARP1, in a panel of normal-immortalized cell lines, however these features were not present in cancer cell lines.

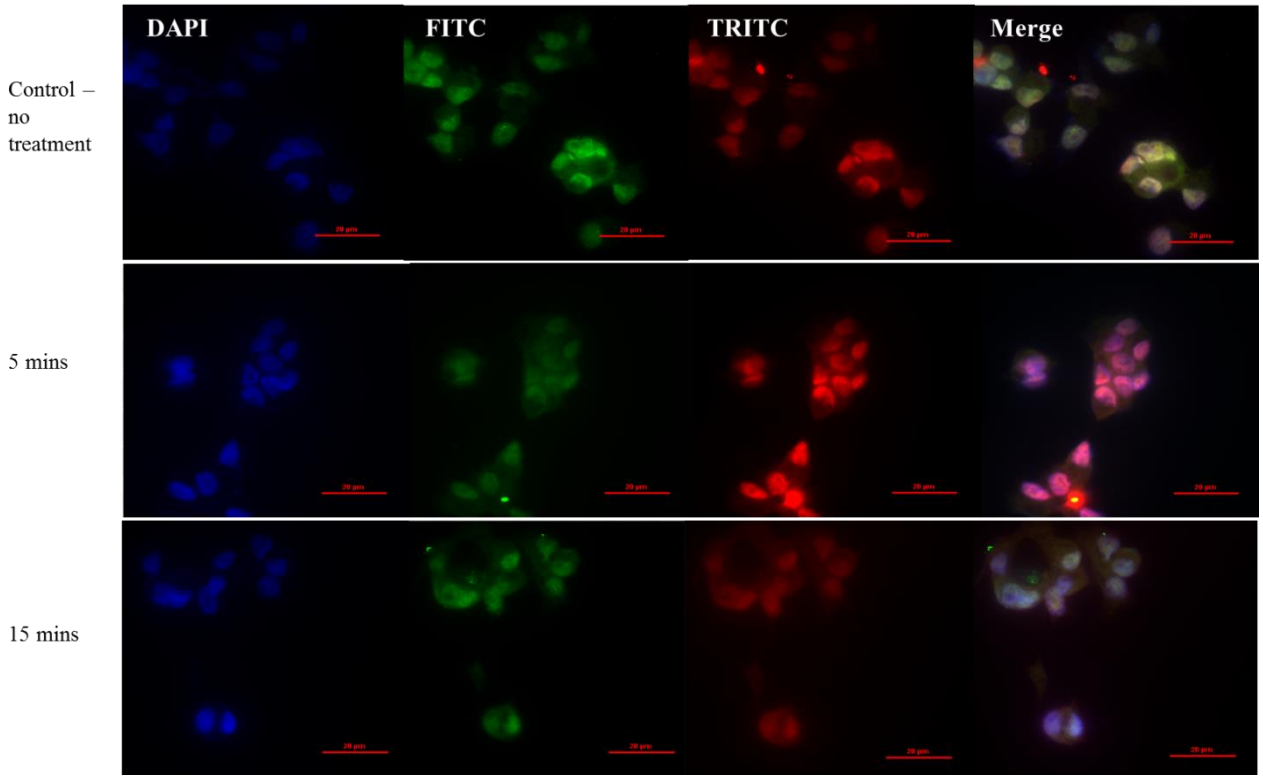


Figure 6-5 Widefield microscopy of ZR-75.1 cells after treatment with H_2O_2 immunofluorescently stained with the anti-CTCF and anti-PARP antibodies

ZR-75.1 cells were cultured on glass coverslips, placed in the wells of a 12 well-plates, treated with H_2O_2 fixed and prepared for IF staining experiments. *Green (FITC),* staining with the anti-CTCF antibody. R*ed (TRITC),* staining with the anti-PARP1 antibody. Nuclei were visualized with 4′,6-diamidino-2-phenylindole (DAPI) staining (*blue*). *Right*, merge of the green and red channels Images were taken at ×60 magnification using the Nikon Ti-Eclipse widefield microscope. Scale bar shows 20 μm.

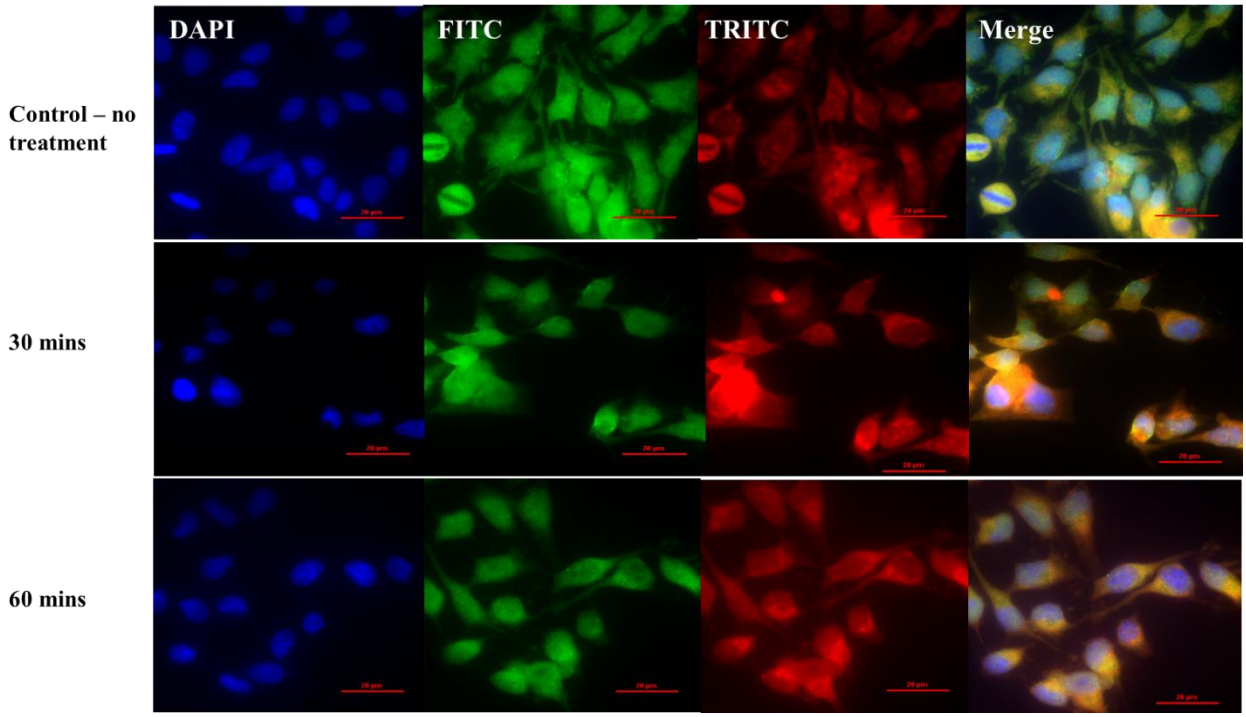


Figure 6-6 Widefield microscopy of HeLa cells after treatment with H_2O_2 immunofluorescently stained with the anti-CTCF and anti-PARP antibodies

HeLa cells were cultured on glass cover-slips placed in the wells of a 12 well-plates, treated with H_2O_2 fixed and prepared for IF staining experiments. *Green (FITC)*, staining with the anti-CTCF antibody. R*ed (TRITC)* , staining with the anti-PARP1 antibody. Nuclei were visualized with 4′,6-diamidino-2-phenylindole (DAPI) staining (*blue*). *Right*, merge of the green and red channels. Images were taken at ×60 magnification using the Nikon Ti-Eclipse widefield microscope. Scale bar shows 20 μm.

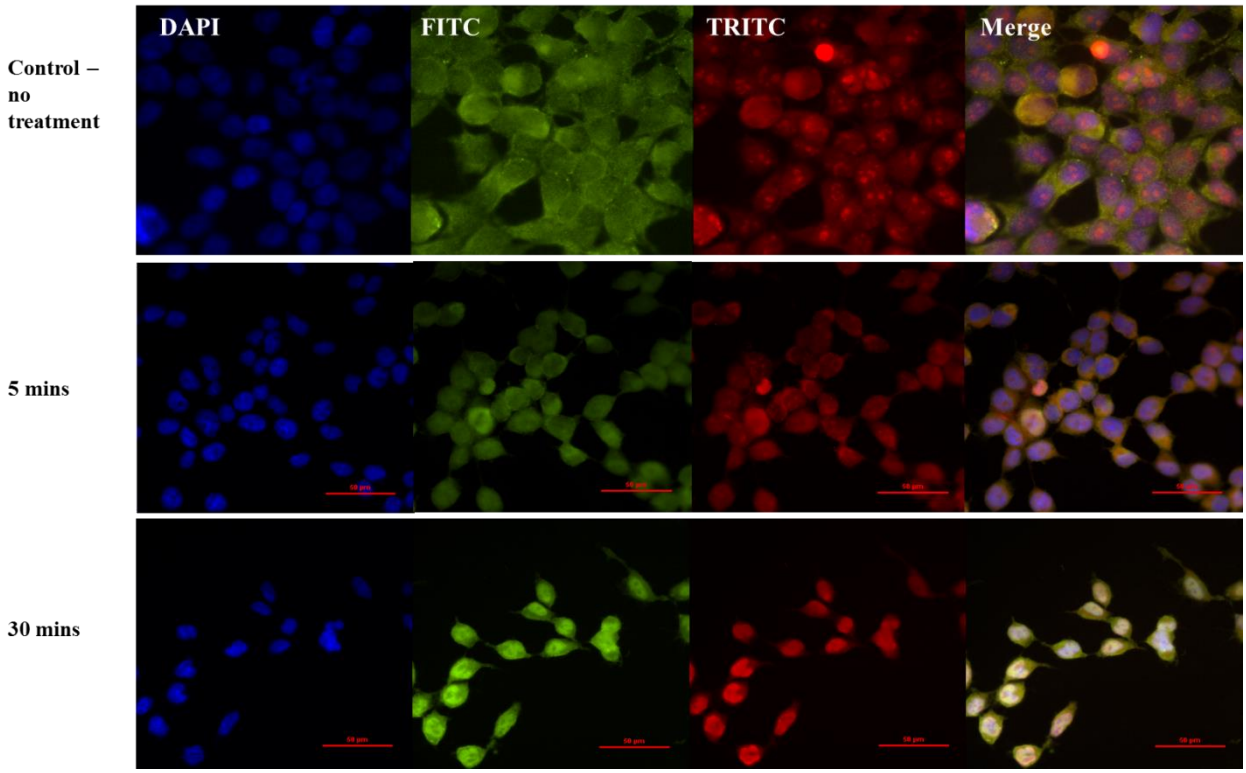


Figure 6-7 Widefield microscopy of 293T cells after treatment with H_2O_2 immunofluorescently stained with the anti-CTCF and anti-PARP antibodies

293T cells were cultured on glass cover-slips placed in the wells of a 12 well-plates, treated with H_2O_2 fixed and prepared for IF staining experiments. *Green (FITC)*, staining with the anti-CTCF antibody. R*ed (TRITC)*, staining with the anti-PARP1 antibody. Nuclei were visualized with 4′,6-diamidino-2-phenylindole (DAPI) staining (*blue*). *Right*, merge of the green and red channels. Images were taken at ×60 magnification using the Nikon Ti-Eclipse widefield microscope. Scale bar shows 20 μm.

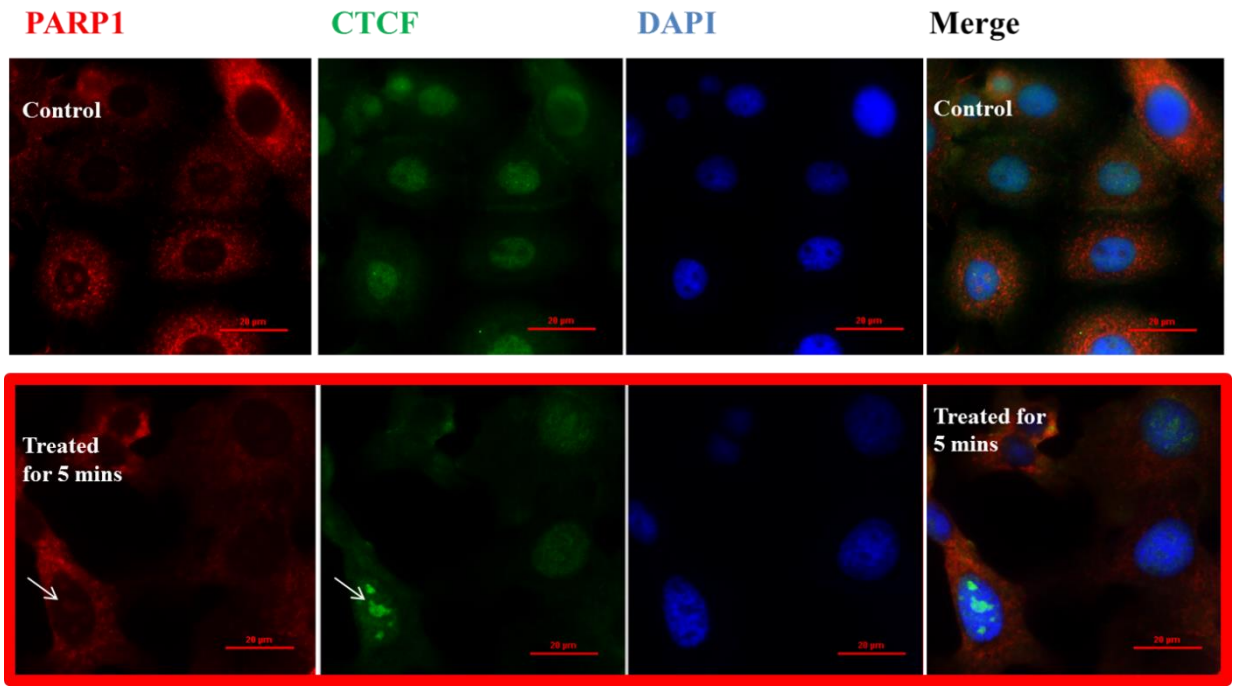


Figure 6-8 Immunofluorescence staining on BPH-1 cells after treatment with H_2O_2 using anti-CTCF and anti-PARP1 antibodies viewed by widefield microscopy

BPH-1 cells cultured, treated with H_2O_2**,** fixed and stained following the same protocol as for the cells in figure 6.1. Several time-points were tested and compared to the healthy situation. At 5 minutes after treatment CTCF and PARP1 appear to co-localize inside the nucleus (shown with white arrows). Images were taken at ×60 magnification using the Nikon Ti-Eclipse widefield microscope. Scale bar shows 20 μm.

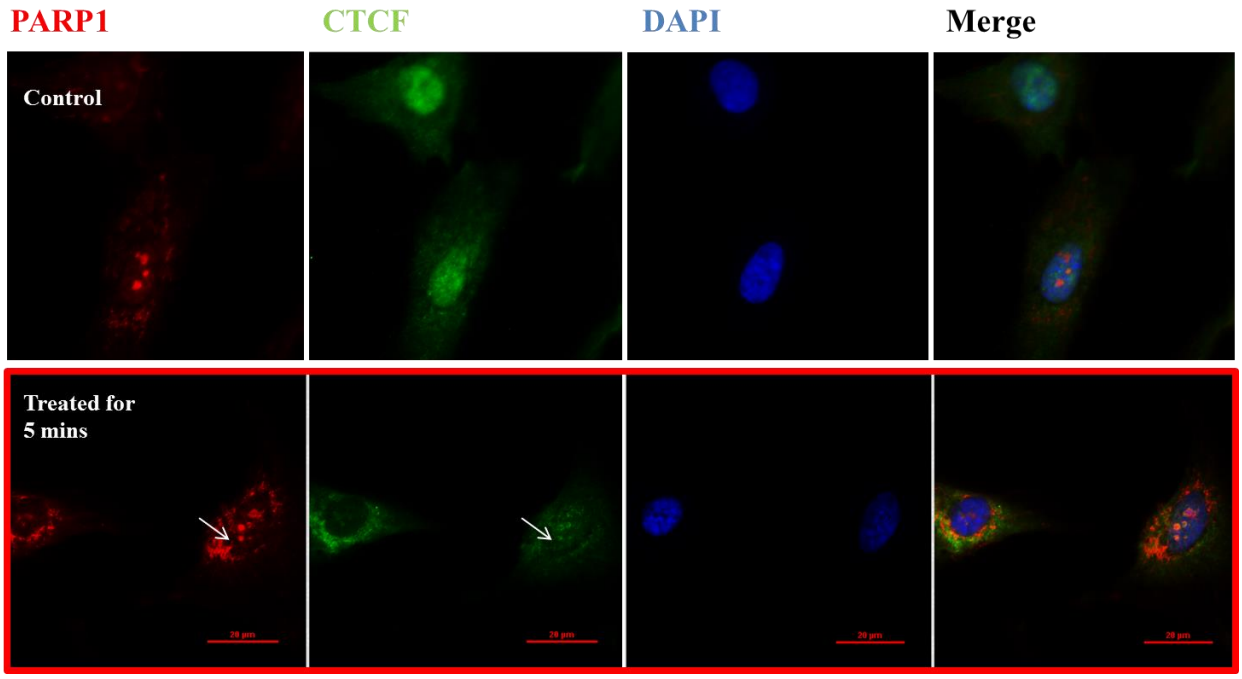


Figure 6-9 Immunofluorescence staining on Hs27 cells after treatment with H_2O_2 using anti-CTCF and anti-PARP1 antibodies viewed by widefield microscopy

Hs27 cells cultured, treated with H_2O_2**,** fixed and stained following the same protocol as for the cells in figure 6.1. Several time-points were tested and compared to the healthy situation. At 5 minutes after treatment CTCF and PARP1 appear to co-localize inside the nucleus (shown with white arrows). Images were taken at ×60 magnification using the Nikon Ti-Eclipse widefield microscope. Scale bar shows 20 μm.

6.3.3 PARylation of CTCF is important in its involvement in DDR

The observations described in the previous sections indicated that CTCF and PARP1 may be involved in the regulation of the DDR pathways. This section will further explore the relationship between CTCF and PARP1, in particular the importance of CTCF PARylation, in the context of DDR.

To achieve this, two experimental approaches will be utilized: (1) transient transfections using the PARylation-deficient CTCF vector and (2) application of the PARP inhibitor, ABT-888. The effects of the aforementioned methods will be studied in cells after DNA damage. The experiments described in this section were all conducted on 226LDM cells.

6.3.3.1 Exogenous PARylation-deficient CTCF does not co-localize with PARP1 nor UBF in the event of DNA damage and appears to affect the formation of the nucleolus

Two types of EGFP-CTCF hybrid plasmids generated in our laboratory (Farrar et al., 2010) were used for the transfection experiments: the wild-type CTCF and the PARylation-deficient CTCF mutant. The transfected 226LDM cells grown on coverslips were transfected with these plasmids and then subjected to the DNA damage experiment with H_2O_2 as described in the previous section. The IF staining images, obtained with the aid of widefield microscopy, revealed that in the cells with exogenous wild-type CTCF a similar localization pattern could be observed as in the cells with endogenous CTCF (figure 6-10). On the other hand, in cells transfected with the PARylation-deficient CTCF mutant no nucleolar co-localization of the exogenous EGFP-tagged CTCF with PARP1, and also with UBF, could be observed at any time-point. (figure 6-11). Furthermore, Remarkably, UBF and PARP1 appeared to lose their own normal spatial pattern in the cells transfected with the mutant CTCF, compared with non-transfected cells. These observations indicate that PARylation of CTCF may play a role in the stabilization of the nucleoli structure, at least in the context of DNA damage.

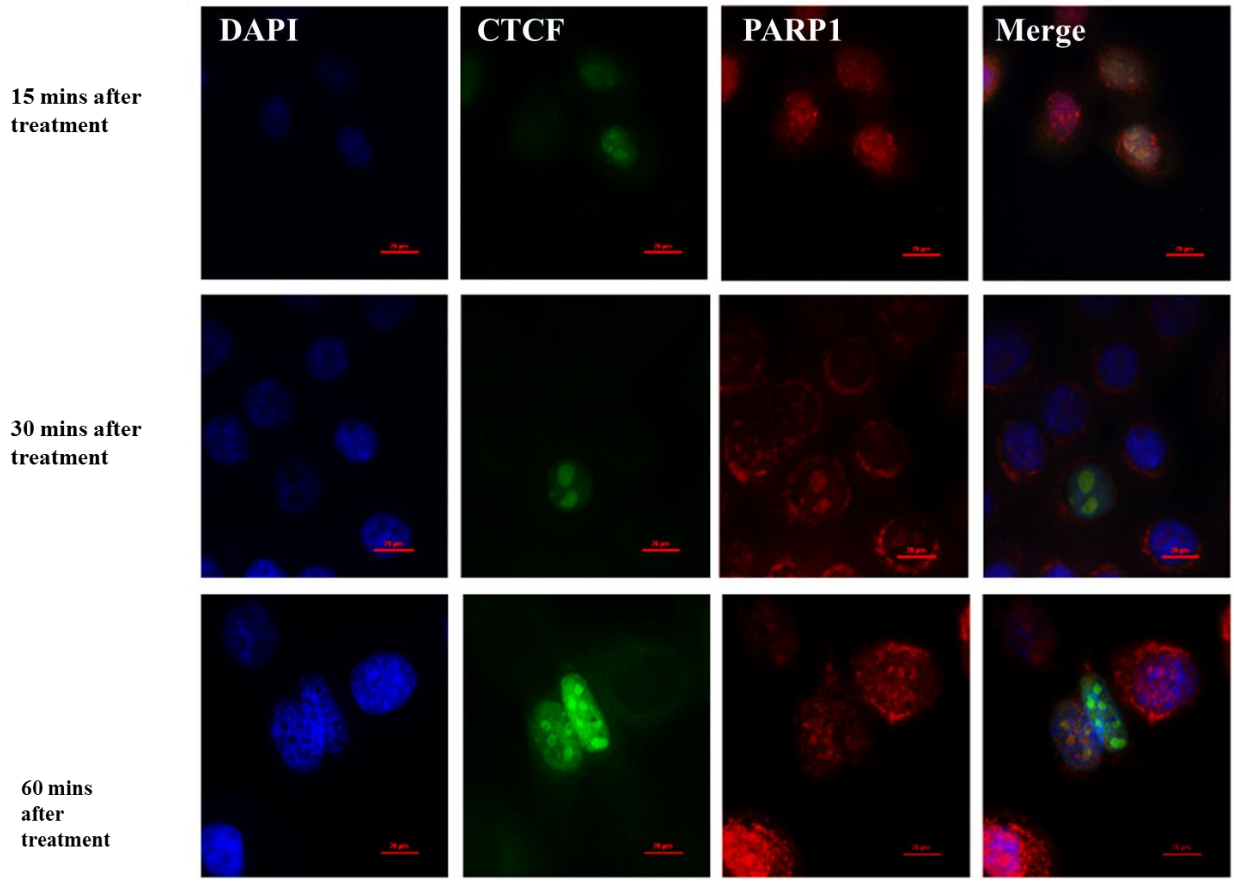


Figure 6-10 Imaging of 226LDM cells, transfected with the exogenous, EGFP-tagged, wild-type CTCF following treatment with H_2O_2 and viewed by widefield microscopy

226LDM cells grown on glass coverslips in wells of a 12-well plate were transfected with plasmids carrying the wild-type EGFP-CTCF constructs. The cells were then treated with 200 μM H_2O_2 for various time-points, fixed and stained using the monoclonal anti-PARP1 antibody. Exogenous wild-type CTCF (green – EGFP) is localized in the nucleus and co-localizes with PARP1 (red – TRITC) following DNA damage. Images were taken at ×60 magnification using the Nikon Ti-Eclipse widefield microscope. Scale bar shows 20 μm.

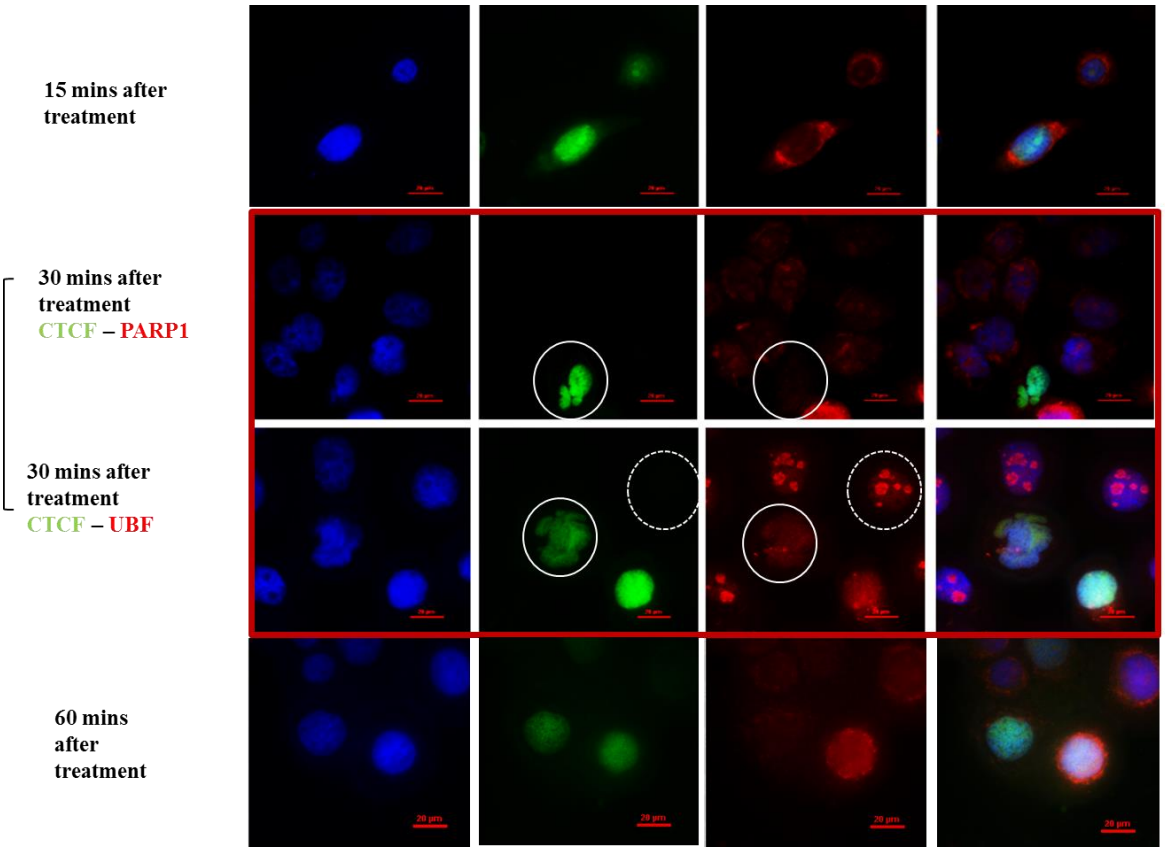


Figure 6-11 Imaging 226LDM cells, transfected with the EGFP-tagged, PARylation-deficient CTCF mutant following treatment with H_2O_2 and viewed by widefield microscopy

226LDM cells were grown on glass coverslips, transfected with the plasmid expressing the EGFP-CTCF mutant deficient for PARylation. The cells were treated with 200 μM H_2O_2 for various time-points, fixed and stained using a monoclonal anti-PARP1 antibody. Note that the diffuse localization of exogenous CTCF (green – EGFP) in nucleoplasm and absence of co-localization with PARP1 (red – TRITC) or UBF (red – TRITC) following DNA damage. The normal distribution pattern of UBF and PARP1 seen in untransfected cells (dashed line circles) is compromised in transfected cells (solid line circles). Nuclei were visualized with 4′,6-diamidino-2-phenylindole (DAPI) staining (*blue*). *Right*, merge of the green and red channels. Images were taken at ×60 magnification using the Nikon Ti-Eclipse widefield microscope. Scale bar shows 20 μm.

6.3.4 Cells treated with the PARP inhibitor ABT-888 repair DNA damage slower / less efficiently than control cells

To confirm the general importance of PARP activity in the cellular response to DNA damage the ABT-888 PARP inhibitor, which has been reported to block the activity of PARP1-PARP4 (Wahlberg et al., 2012), was employed for our experiments. The extent of DNA damage was determined through the abundance of the modified histone γ-H2AX, which represents a phosphorylated H2AX histone and is a known marker for DNA damage (Kuo and Yang, 2008, Sharma et al., 2012).

In these experiments, 226LDM cells grown on coverslips were treated with 5μM ABT-888 overnight and fresh inhibitor was added to the cells together with H_2O_2. Cells, with and without ABT-888, were then fixed and stained with the anti γ-H2AX antibody at different time-points.

The IF staining images, obtained with widefield microscopy, revealed that in cells where PARPs were inhibited the levels of γ-H2AX following treatment with H_2O_2 were still considerable after 60 min, whereas in control cells staining almost disappeared at this time point (figure 6-12). This result can be interpreted as slower and/ or defective function of the DNA repair machinery.

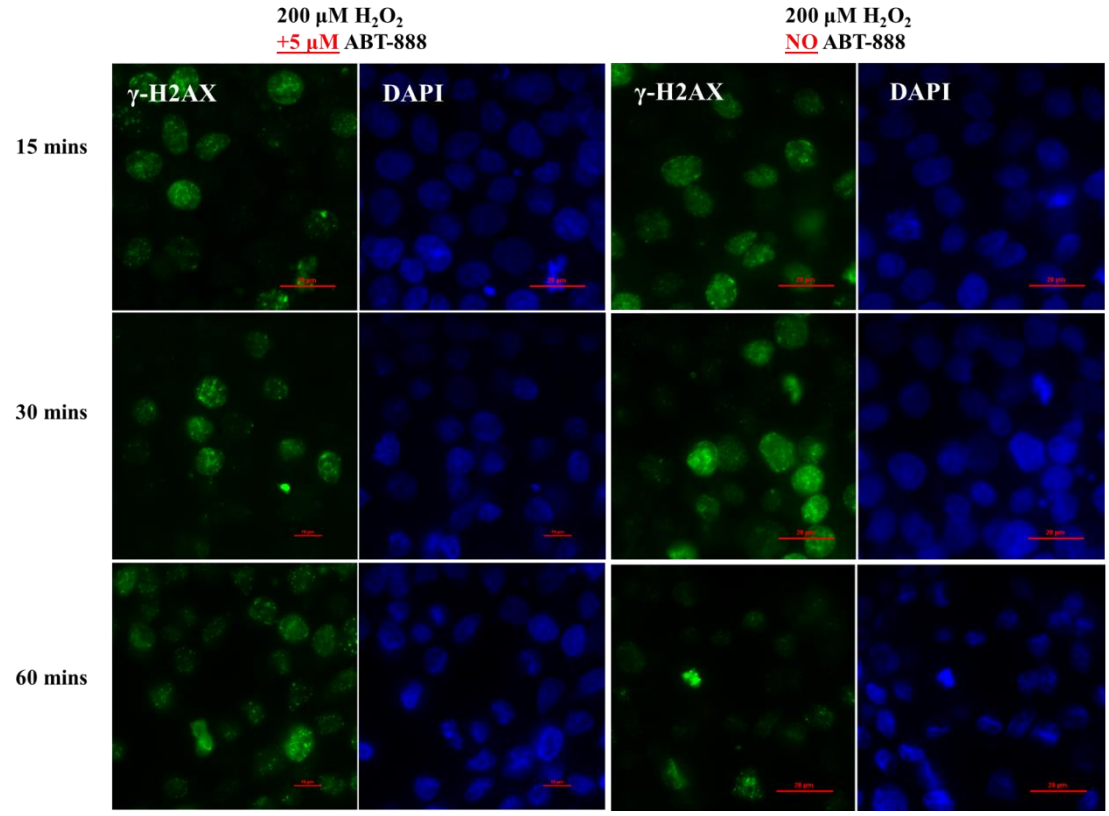


Figure 6-12 Distribution of γ-H2AX, a marker for DNA damage, in untreated 226LDM cells and cells treated with 5 µM ABT-888 following H_2O_2-induced DNA damage: widefield microscopy

226LDM cells cultured on coverslips in 12-well plates were exposed overnight to 5 µM ABT-888 or left untreated to be used as control. To induce DNA damage, 200 µM H_2O_2 were applied to all cells fresh ABT-888 was added at the same time in the relevant wells). *Green (FITC)*, staining with the anti- γ-H2AX antibody. Nuclei were visualized with 4′,6-diamidino-2-phenylindole (DAPI) staining (*blue*). Images were taken at ×60 magnification using the Nikon Ti-Eclipse widefield microscope. Scale bar =20 µm.

6.3.5 High throughput measurement of DNA Damage - FADU assay

In the previous section, through use of the γ-H2AX antibody the visualization of the repair progress was possible in the single cells/small populations of cells. The employment of the FADU assay (described in Section 2.5) enabled the monitoring of the repair process at the total cell population level.

6.3.5.1 Overview of the DNA damage repair pattern in cell lines

In the first series of FADU experiments, 226LDM cells were seeded in nontransparent FADU well plates (1×10^6 cells/ well – 4 replicates for each time point). Once the attachment of cells to the surface of the wells was confirmed, the plates were left overnight in the CO_2 incubator. On the following day, the cells were exposed to 200 μM H_2O_2 for a wide range of time points. The plates were then placed in the FADU machine and at the end of the experiment the fluorescence emitted from the wells was measured in a plate reader. The fluorescence in this case represents double-stranded DNA and therefore the level of fluorescence is inversely proportional to the damage within the cell population. Two controls were used in each FADU experiment; T0 included undamaged cells that did not undergo the alkaline unwinding step and P0 contained undamaged cells.

The results obtained from this initial experiment showed the repair pattern of 226LDM displaying a gradual repair in the course of 60 minutes from the introduction of DNA damage. This pattern is in agreement with the general existing model (Moreno-Villanueva et al., 2011, Moreno-Villanueva et al., 2009) (figure 6-13 – A). The T0 control confirmed that the cells at the time of these experiments were viable and healthy. The same protocol, as described above for 226LDM cells, was used in the experiment with breast cancer cells, ZR-75.1. A response pattern similar to the 226LDM cells was observed in this case (figure 6-13 – B). It should be noted that in both experiments the fluorescence levels at 60 minutes were lower than the P0 control (non-

damaging conditions), therefore in our next experiments the final time point was extended to 90 minutes to ensure maximal repair.

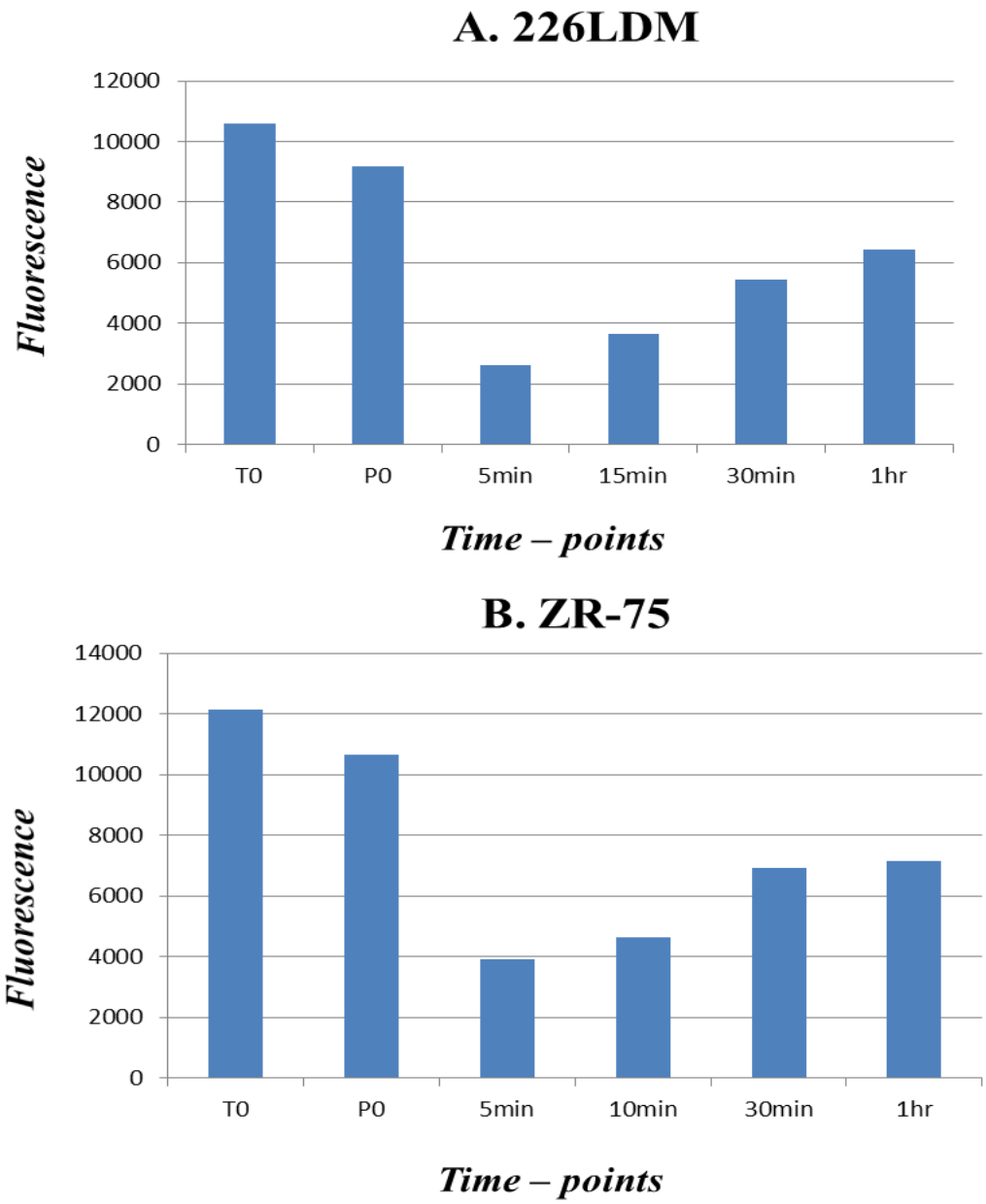


Figure 6-13 Repair pattern from H_2O_2 -induced damage in 226LDM and ZR-75.1 cells as recorded using the FADU assay

226LDM (A) and ZR-75.1 (B) cells were seeded in FADU well plates in a concentration of 1 x 10^6/ well and grown overnight. On the following day, 200 µM H_2O_2 was introduced to the cells for various time points. Subsequently the plates were placed in the FADU robot and after the end of the assay the emitted fluorescence was measured in a plate reader. The results were plotted in a graph in which high fluorescence measurements represent lower damage levels. T0 and P0 are controls. T0 represents the level of SybrGreen fluorescence in cells that did not undergo the alkaline unwinding step. P0 represents the level of SybrGreen fluorescence obtained in undamaged cells.

6.3.5.2 **Introduction of PARylation-deficient CTCF in 226LDM cells affects their repair pattern**

Following the experiments that produced the repair pattern in 226LDM cells, we asked how PARylation of CTCF affects this process. To investigate this, 226LDM cells were seeded in FADU well plates and then transfected with the plasmids expressing the wild-type CTCF and the mutant variant of CTCF deficient for PARylation. Cells transfected with the empty vector were used as control. The transfected cells were treated with 200 μM H_2O_2 at time points 15, 30, 45 and 90 minutes followed by the FADU assay.

The results of this experiment show that a similar pattern in all cell samples, namely a gradual repair which reached the plateau at ~ 90 minutes from the introduction of DNA damage (figure 6.15). The fluorescence levels were similar in all transfected cells and comparable with the control (P0, non-damaged cells). However, the rate of the repair is increased in in the cells expressing the mutant CTCF deficient for PARylation. Although the exact events taking place within the cells and leading to this outcome are not yet known to us, it could be hypothesized that in these cells an alternative, more rapid, repair pathway is activated.

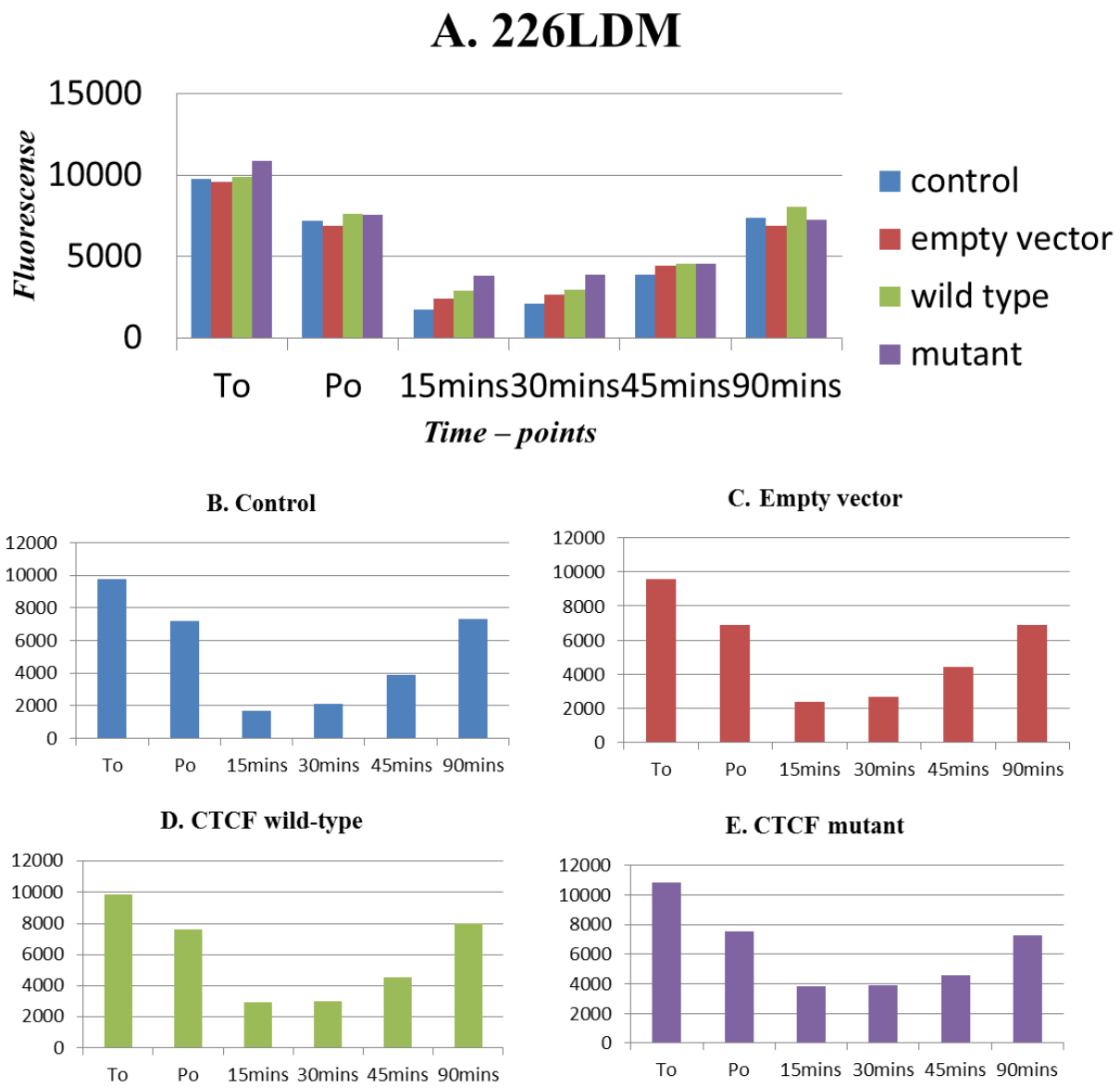


Figure 6-14 DNA damage repair patterns in control and transfected 226LDM cells

226LDM cells were seeded in wells of FADU plates with a concentration of $1x10^6$ cells per well. Four replicates were prepared for each time-point. On the following day the cells were transfected with Empty Vector, CTCF wild-type and CTCF mutant (deficient for PARylation). Control represent non-transfected cells. After treatment with 200µM H_2O_2 at the indicated time points, cells were used in the FADU assay and the fluorescence results were obtained. Panel A represents the combined results of all experiments, whereas panels C-E represent results of transfections with individual plasmids and control. T0 and P0 are controls. T0 represents the level of SybrGreen fluorescence in cells that did not undergo the alkaline unwinding step. P0 represents the level of SybrGreen fluorescence obtained in undamaged cells.

6.3.5.3 **The processes of DNA repair of the H_2O_2-induced DNA damage in 226LDM is slower in cells treated with the PARP inhibitor ABT-888 than in untreated cells.**

The importance of the general PARylation activity in the DNA repair processes was discussed earlier, and the experiments shown in figure 6.13 demonstrated the reduced rate of repair in cells treated with the PARP inhibitor, ABT-888. To confirm these observations, the effects of the ABT-888 on DNA repair in 226LDM cells were investigated using the FADU assay. For these experiments, 226LDM cells were seeded in wells of a FADU plate as described in previous sections, the cells were exposed to 5 μM ABT-888 overnight and then the following day 200 μM H_2O_2 was added; control cells were not treated with ABT-888. The findings suggest that since the fluorescence levels increase more rapidly in control cells than in cells treated with ABT-888, the effectiveness and/or the rate of repair is effected by PARP inhibition (figure 6-15 - A). The damage was also expressed in the Gray unit which is an irradiation absorbance unit, used to allow comparisons between damaging agents (Barton, 1995). Conversely, the DNA damage is considerably higher in PARP-inhibited cells (figure 6.16 - B).

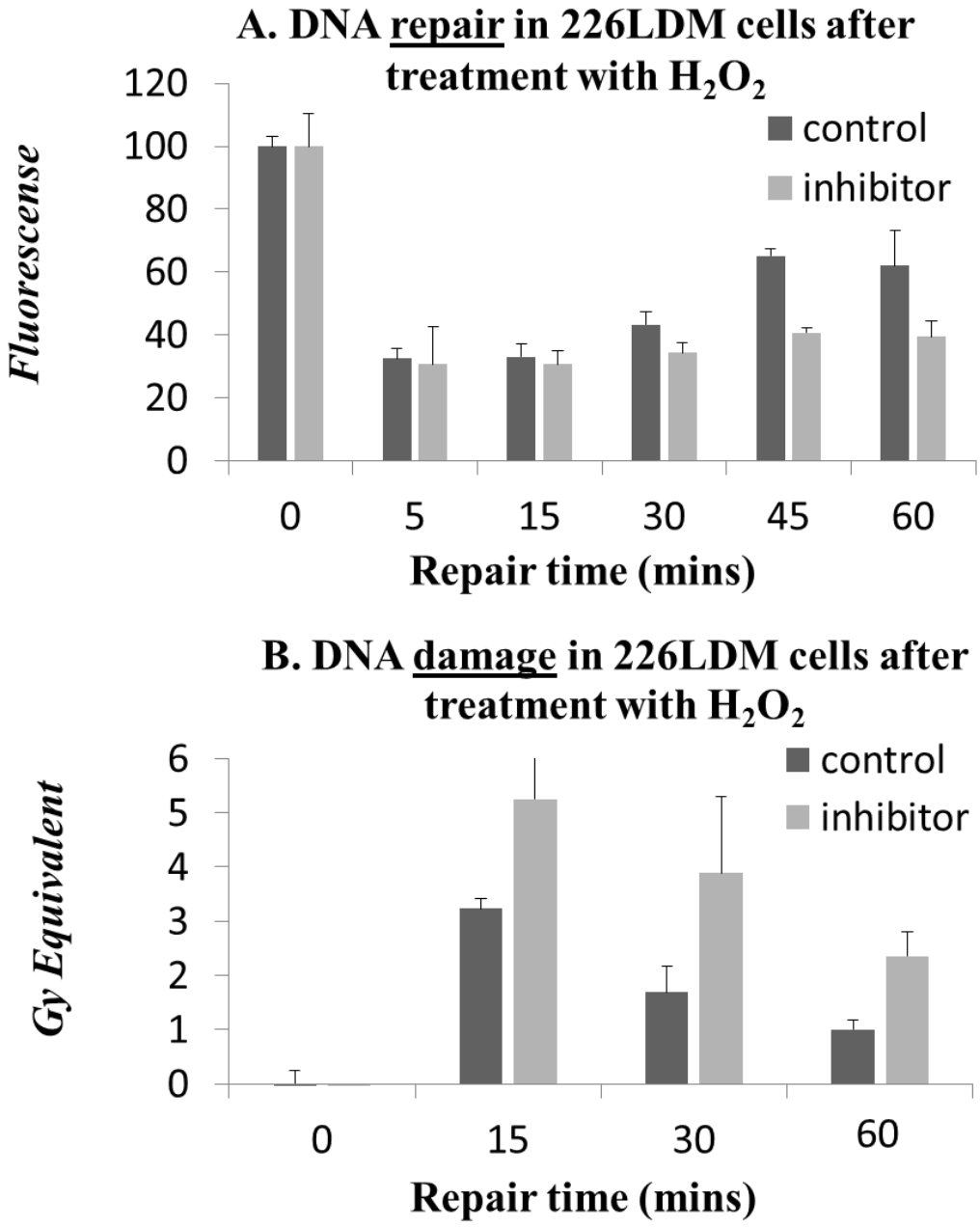


Figure 6-15 The DNA repair (A) and damage (B) patterns in 226LDM cells treated with the PARP inhibitor ABT-888 and H_2O_2 and analyzed using the FADU assay

226LDM cells were seeded in FADU plates and treated with 5 μM ABT-888 overnight; control cells were not treated. DNA damage was induced on the following day by addition of 200 μM H_2O_2 followed by FADU assay at the indicated time points (0-60 min). In (B) the results from selected time points are used to show the damage expressed in relative absorbance of irradiation units (Gy).

6.3.5.1 **The processes of DNA repair of the X-Ray -induced DNA damage in 226LDM are slower in cells treated with the PARP inhibitor ABT-888 than in untreated cells.**

Many difference factors can cause DNA damage in cells. The rationale for selection of a chemical agent such as hydrogen peroxide (H_2O_2) for this study was the reagent availability, reproducibility and relatively low hazard. Ionizing X-radiation is another known factor which has been linked to DNA damage (Bradley and Erickson, 1981). We therefore investigated whether ionizing X-radiation would lead to similar effects earlier described in 226LDM treated with H_2O_2. In these experiments, 226LDM cells were plated in FADU plates, irradiated and analyzed by the FADU assay. As shown in figure 6-16 –A, the low fluorescence levels that are present in the first time-points after damage increase with time until the repair is almost complete. The pattern appears to be similar to that resulting from the exposure to hydrogen peroxide (figures 6.14 and 6.15).

The PARP inhibitor ABT-888 was then employed to confirm that the role of PARylation activity in DDR does not depend on the damage factor. In these experiments, 226LDM cells were exposed to PARP inhibitor (5 μM ABT-888) overnight. The untreated cells were used as control. On the following day, cells were transferred in micro-centrifuge tubes and irradiated on ice with 3.8 Gy using an X-ray generator. Cells were then transferred in a water bath for various time-points at 37^{o}C to repair the damage (the longest time length was 90 minutes), and analyzed by the FADU assay. As shown in (figure 6-16 – B), the fluorescence level the repair patterns are similar to those seen after hydrogen peroxide damage. Indeed, the repair seems to take place quicker in cells that have not been affected by a PARP inhibitor (figure 6-16 – B).

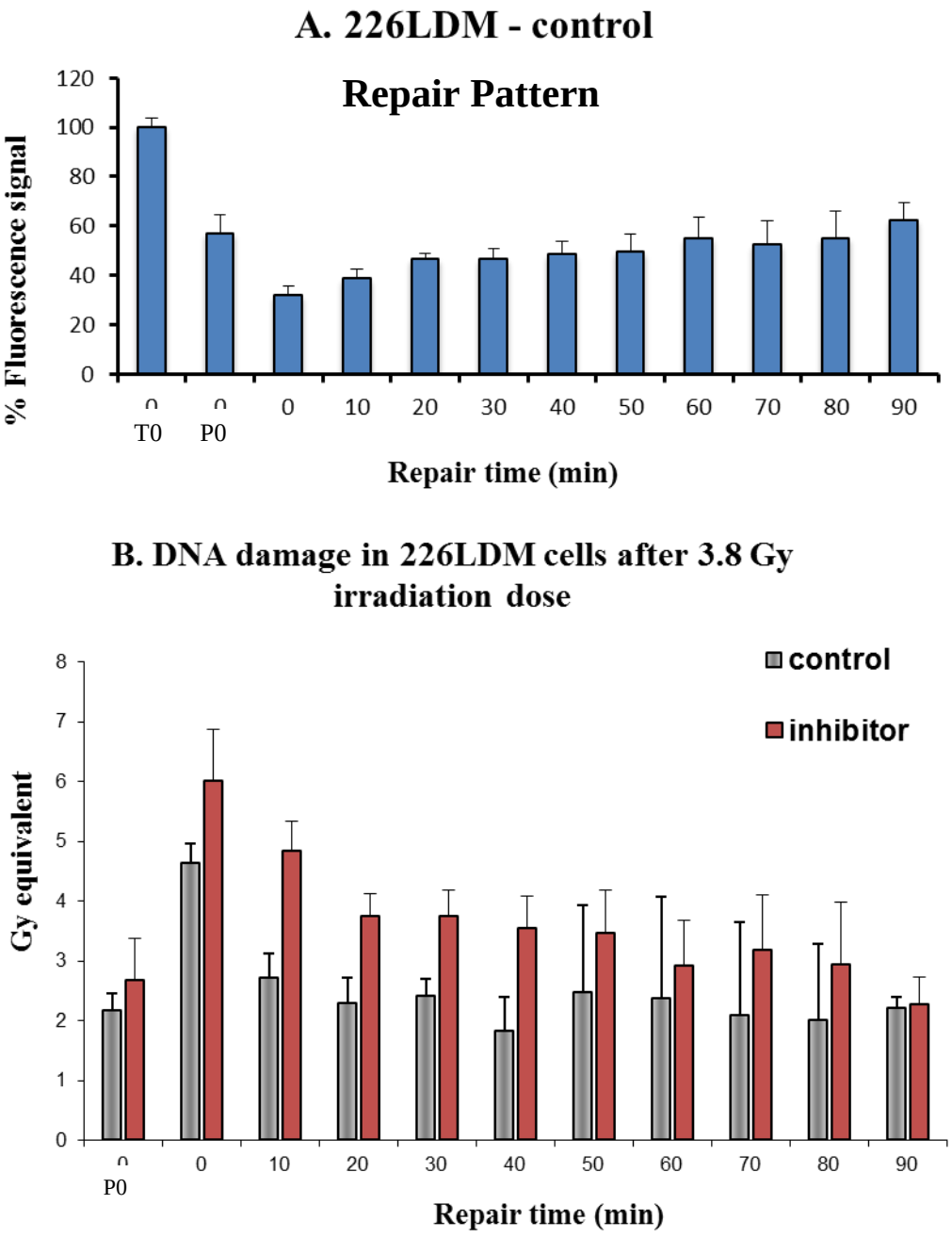


Figure 6-16 The DNA repair (A) and damage (B) patterns in 226LDM cells treated with the PARP inhibitor ABT-888 and ionizing X-radiation and analyzed using the FADU assay

226LDM cells ($6x10^5$/ sample) were exposed to PARP inhibitor (5 μM ABT-888) overnight and untreated cells were used as control. On the following day, cells were transferred in micro-centrifuge tubes and irradiated on ice with 3.8 Gy using an X-ray generator. The cells were then placed in a water bath at 37^{o}C and exposed at several time points as indicated. Following repair, the samples underwent the FADU assay and the results were obtained from the fluorescence reader. The patterns of recovery from X-ray irradiation show low fluorescence (increased damage) right after irradiation which increases (decreased damage) with time (A). The exposure of cells to the PARP inhibitor (5 μM ABT-888) decreases the speed of recovery (B).

6.4 Discussion

The sources of DNA damage are diverse and the occurrence of damaging effects is frequent, therefore, various repair mechanisms evolved in the cells in order to survive. The robust DNA damage response and repair mechanisms are very important in the maintenance of the genomic integrity, and defective repair is associated with disease (cancer in particular), cell death and ageing (Wei et al., 2007, Ljungman, 2010, Kirkwood, 2005, Hoeijmakers, 2009).

Despite extensive studies in this area, the DNA response pathways and the proteins participating in them have not been yet fully defined. The PARylation reactions performed by Poly(ADP-ribose) polymerases (PARPs), have been found to be one of the earliest post-translational modification at the sites of the DNA single and double-strand breaks, resulting in changes local chromatin structure and the accumulation of several DDR response proteins(Robert et al., 2013, Golia et al., 2015).

CTCF is an important regulator of many cellular functions and a key genome organizer (Phillips and Corces, 2009, Ohlsson et al., 2010, Handoko et al., 2011). However, the role of CTCF in DDR has not been studied. The aim of this chapter was to investigate whether CTCF is involved in the DDR and a possible role of CTCF PARylation in this process. Our experimental findings support the existence of such a link and pave the way for further research on the subject.

The 226LDM breast luminal cells were used for the majority of our experiments due to their non-cancerous nature and their closeness to the “normal” condition. Two CTCF isoforms, CTCF-130 (non-PARylated or hypo-PARylated) and CTCF-180 (hyper-PARylated), were detected in undamaged proliferating cells (Figure 6.1), in agreement with a previously published reports (Docquier et al., 2005), (Docquier et al., 2009).

The amount of CTCF-130 and CTCF-180 and the ratio between them have not changed after DNA damage by H_2O_2 thus implying that the PARylation status of CTCF remained

unaltered. It is however possible that the amount of hypo-PARylated CTCF, co-migrating with non-PARylated CTCF, increased; such difference is not possible to resolve using Western analysis and further experiments (e.g. immunoprecipitations) (Farrar et al., 2010) will be required to confirm hypo-PARylation. Since the overall levels of CTCF protein have not changed after DNA damage, we concluded that if CTCF plays a role in DDR, it involves the existing CTCF protein molecules.

In our further experiments we observed the appearance of CTCF in the nucleoli of 226LDM cells and, furthermore, its co-localization with PARP-1 in these structures after DNA damage (figure 6-2). Translocation of CTCF from nucleoplasm into nucleoli of cells following growth inhibition, cell differentiation and apoptosis was reported previously (Torrano et al., 2006). The findings described in this and another publication (van de Nobelen et al., 2010) link CTCF to nucleolar functions, in particular transcription from rDNA. In the latter study CTCF was found to interact with UBF, a transcription factor involved in RNA polymerase I-mediated ribosomal (r)RNA transcription. Furthermore, CTCF binding sites were mapped to a site upstream of the rDNA spacer promoter (van de Nobelen et al., 2010). It should be noted that CTCF function appears to differ depending on particular biological situations. In ES and MEFs CTCF supports nucleolar transcription (van de Nobelen et al., 2010), however in differentiated cells such as rat neuron –like UR61 nucleolar transcription is inhibited by CTCF (Torrano et al., 2006). Interestingly, inhibition of PARPs by 3-ABA restores nucleolar transcription in UR61 cells transfected with CTCF, which indicates that PARylation may regulate CTCF function in the nucleoli (Torrano et al., 2006). Another notable difference is the timescale of the observed events; the changes in localization are detected within minutes in 226LDM cells after treatment with H_2O_2, whereas it is hours or days in other systems. This observation suggests that such rapid events may not require RNS and protein synthesis and rely on post-translational

modification (such as PARylation) and is supported by the equal levels of CTCF-130 and CTCF-180 in damaged cells.

It is conceivable that in the undamaged 226LDM cells CTCF and PARP1 are present at lower levels in the nucleoli, and the exposure to DNA damaging agents such as H_2O_2 leads to activation of PARP1, PARylation of CTCF and translocation of both proteins to the nucleoli where CTCF modification may be sustained by PARP1. It remains to be determined whether nucleolar transcription is affected by DNA damage in these cells, and if CTCF and PARylation of CTCF are important for this process.

To examine whether nucleolar translocation of CTCF and PARP1 after DNA damage was a general characteristic or restricted to 226LDM cells, a wider panel of cell lines which included transformed cells (293T, HeLa and ZR-75.1) and "normal" cells (Hs-27 and BPH-1). These experiments revealed that, in response to treatment with H_2O_2, CTCF was detected in the nucleoli co-localizing with PARP1 in a panel of normal-immortalized cell lines and not in cancer cell lines.

A possible interpretation of the results in cancer cells lines is that the activated pathways in these cells do not involve CTCF. On the other hand, it is also plausible that the higher tolerance of cancer cells to damage would require drastically higher H_2O_2 concentrations to initiate the repair processes, or that the events take place on a completely different time-scale. Indeed, it should be taken into consideration that the time point where these features could be observed in cancer cell lines were simply missed, although a range of time-points and H_2O_2 concentrations were used for all cell lines. Additional time points may need to be used in future experiments to explore this further. The exact damage repair mechanism operational in cancer cells, although of great interest, remains beyond the scope of this study.

In the previous study, the nucleolar localization signal (NuLS) was found to be within the DNA binding domain of CTCF, in fact, it appeared that there may be more than one NuLS in this domain (Torrano et al., 2006). However, the full length CTCF was necessary for the biological function of CTCF. It remains to be investigated whether these observations are applicable in the context of the DDR, or whether a different NuLS(s) may be responsible for the nucleolar localization. A possible relationship between the NuLS(s) and the PARylation sites within the N-terminal domain of CTCF, which may be important for the regulated transfer will also need to be addressed.

Focusing on 226LDM cells, we showed through staining with γ-H2AX (figure 6-12) but also for the first time by using the automated FADU assay (figure 6-15), that PARP inhibition leads to a slower DNA damage repair capacity. These observations was confirmed when another DNA damaging agent (X –ray radiation) was used (figure 6-16). These results are supported by other studies demonstrating the correlation between inhibition of PARP and DDR (Durkacz et al., 1981a, Durkacz et al., 1981b, Villani et al., 2013, Curtin, 2012).

To examine the effects of CTCF PARylation in DDR in 226LDM cells, we used a mutant CTCF deficient for PARylation (Farrar et al., 2010). In cells expressing the wild type CTCF tagged with EGFP, the exogenous CTCF behaved similarly to the endogenous CTCF and was observed in nucleoli, together with PARP1, after 30 min post-treatment with H_2O_2 (figure 6-10). However, in the cells containing the mutant CTCF the nucleoli structures collapsed following exposure to H_2O_2 (figure 6-17). This finding supports previous observations that CTCF is important for nucleolar stability and transcription (Torrano et al., 2006, Guerrero and Maggert, 2011, Huang et al., 2013). CTCF PARylation may also be important in the establishment and maintenance of nucleolar architecture, where appropriate short and long-range DNA interactions are important for proper nucleolar function. The disappearance of such interactions and also global changes in nuclear (Handoko et al., 2011, Guastafierro et al., 2013,

Soto-Reyes and Recillas-Targa, 2010) and nucleolar (Hernandez-Hernandez et al., 2012, van de Nobelen et al., 2010) organization have been reported in CTCF-depleted cells, thus confirming the role of CTCF in these processes. Interestingly, counter-intuitively to the assumption that the repair effectiveness might decline in the presence of CTCF deficient for PARylation, the FADU assay revealed that in the total cell population the rate of the DNA repair processes is more rapid than in cells with the wild-type CTCF. However, it should be acknowledged that the PARylation-dependent DDR pathways are very complex, and the observed inhibition of DDR in cells treated with a general PARP inhibitor represent the outcome integrated from multiple and probably opposing pathways. The introduction of the exogenous CTCF deficient for PARylation into 226LDM cells may disrupt the regulation of the DDR resulting in the situation when other pathways become predominant and PARylation of CTCF is no longer important or necessary. The alternative pathways used in these cells may be quicker however the accuracy may be compromised. Although this proposition is hypothetical, the link between decreased CTCF PARylation and breast tumourigenesis has been reported (Docquier et al., 2009). It is conceivable that CTCF PARylation may be important for regulation of DDR in normal breast (and other) cells, however in the process of tumour evolution CTCF PARylation decreases and the CTCF-dependent DDR pathway become deregulated. Due to time limitations, no further FADU experiments were conducted; they will be considered in the future experiments to test this hypothesis.

In summary, our research has shown that in normal/immortalized cells in response to DNA damage CTCF is translocated to the nucleoli. CTCF PARylation is important in this process because the use of the PARylation-deficient CTCF mutant results in destabilization of nucleoli. CTCF PARylation is therefore likely to play a regulatory role in DDR in normal cells, whereas in cancer cells the alternative, CTCF PARylation-independent mechanisms may be in

operation. Future study is needed to prove the exact nature of the role of CTCF in the context or damage.

Chapter 7 General Discussion and Future work

7.1 CTCF Poly(ADP-ribosyl)ation

CTCF is a multifunctional protein involved in many cellular processes and pathways. CTCF was originally thought to be as a transcription repressor due to its role in the negative regulation of c-myc gene (Klenova et al., 1993, Filippova et al., 1996). However, more and more evidence accumulated attributing several other roles to CTCF, including transcriptional activator and silencer, insulator, mediator of long range chromatin interactions and others (Ohlsson et al., 2001, Holwerda and de Laat, 2013).

As a transcription factor, CTCF can regulate specific genes directly by binding to DNA, but it has also been found to act indirectly by recruiting other transcription factors to their binding sites or disallowing them from binding (Kim et al., 2007, Chernukhin et al., 2007). Apart from DNA it also interacts with various proteins (Holwerda and de Laat, 2013).

In this study, we examined the involvement of CTCF in cellular processes on two different levels. Firstly, we investigated CTCF occupancies in the DNA using a genome-wide technique, ChIP-Seq. Secondly, we looked into CTCF role in one particular cellular process, namely DNA damage response. In both parts of our investigation, the main focus was set on the properties of a post-translationally modified isoform of CTCF; this isoform is highly poly(ADP-ribosyl)ated and has a molecular weight of 180kDa (CTCF180) (Yu et al., 2004b, Docquier et al., 2009). Currently, the function of this isoform and the mechanism underlying why and when the switch between isoforms occurs remain unknown.

A study using breast tissues revealed that the PARylated CTCF isoform is the only one present in healthy breast tissues as compared to cancer tissues where both CTCF130 and CTCF180 can be detected (Docquier et al., 2009). This finding, together with the known involvement of PARylation in numerous regulatory procedures (Kim et al., 2005b, Burkle, 2005)

provided an indication that CTCF180 could have a divergent role in cellular processes and motivated us to research this potential further.

7.2 Next generation sequencing for the analysis of CTCF180 binding targets

We used the next generation sequencing (NGS) methods to identify genome wide binding targets of the two CTCF isoforms and to explore the effects of these binding events in control and proliferating cells. The 226LDM cell line was used to generate a model that allowed us to study CTCF180. The cells normally express both isoforms, however under controlled exposure to treatment they can be manipulated into expressing CTCF180 only (figure 3.2). There is currently no available antibody able to specifically detect CTCF180, thus this model provided us with a unique opportunity to develop an experimental framework to study this isoform.

Initially, the analysis of the ChIP sequencing results confirmed for the first time that CTCF180 has binding targets, paving the way for further research into the specifics of this binding in different conditions, cell lines or tissues.

CTCF180 binding in treated cells was limited in numbers, especially when filtered by the Q value. Upon closer inspection, some of the targets interestingly appear to be involved in cell cycle regulation or to be linked to cancer. As an example, S100A13 is involved in cell cycle progression and differentiation. The target is bound by CTCF both in the control and in the treated cells. The protein encoded by this gene is known to play a role in the regulation of the Fibroblast Growth Factor (FGF) (Mouta Carreira et al., 1998) and also has been linked with tumourigenesis (Landriscina et al., 2006).

Another gene bound by CTCF in the cell cycle blocked cells is *TCTN3*. The protein encoded by it belongs to the family of tectonic proteins and is part of the hedgehog pathway. In principle, this pathway is important during embryogenesis; however its members are also thought to play a role in cancer, in particular basal cell carcinoma.

Furthermore, the function of SGPL1, which is bound by CTCF in both conditions, is not yet clear however a study into its antineoplastic potential was previously conducted (Brizuela et al., 2012). From the experiments described in this study, it is not clear which isoform was binding to the targets in the control cells. It is possible that initially the CTCF130 isoform was binding and CTCF180 replaced it after treatment affecting the expression. It could also be the case that CTCF180 is bound on both occasions and the change in expression is the result of an event that cannot be accounted for within the scope of this experiment. Further experiments would be necessary to confirm either hypothesis.

Evidence on the role of PARP1 and PARylation on cell cycle regulation has been provided by Ciccarone et al. (2014), (2015) who showed that PARP1 regulates the expression of TET1. The protein encoded by this gene regulates global methylation events, which affect nucleosome positioning and CTCF binding (Teif et al., 2014), and more specifically hydroxymethylation which is also crucial in tumorigenesis. Since methylation is an important factor affecting CTCF binding (Wang et al., 2012a) and PARylation can affect methylation events (Ciccarone et al., 2015, Nalabothula et al., 2015), it is conceivable that during the cell cycle block caused by the HU and NO treatment, methylation status alterations were involved in changing the binding site profile of CTCF as discovered with ChIP-seq.

Following the analysis of the ChIP-seq results, the global mRNA expression levels were compared between the control and treated cells. The RNA-seq experiment revealed that the cell cycle blocking treatment had a significant effect in the expression levels of a wide range of genes. Thousands of genes were up-regulated and thousands were down-regulated. Gene Ontology analysis on the affected genes showed an extensive up-regulation of genes that are part of the immune response pathway such as *DEFB103B, NLRP10, MARCO and IL32*. Also, genes coding for ribosomal proteins and metabolism as well as transcription regulation were also affected.

In the down-regulated entries, genes involved in cell adhesion were included. This was partially expected as the successful cell treatment lead to the physical detachment of the cells from the surface of the culture flask. As an example, many members of the protocadherin alpha cluster, which are linked to adhesion, were affected; CTCF is known to be involved in the expression of these genes and reduced CTCF levels due to treatment can lead to reduced expression of the genes. Not surprisingly given the experimental design, genes involved in cell proliferation were mostly down-regulated including several members of the E2F family of transcription factors, including E2F1, E2F5 and E2F7, which are crucial in cell cycle progression (Ren et al., 2002)

The integration of the ChIP-seq output with global mRNA expression data obtained from RNA-seq further revealed that CTCF binding or loss of it may be responsible for changes in the expression of a target, leading to expression up-regulation or down-regulation.

It should be acknowledged that the number of sites occupied by CTCF180 and associated with gene expression is small. This may be due to re-localization of at least some of the CTCF180 molecules into the cytoplasm after cell cycle arrest. Such CTCF distribution pattern had previously been reported in normal breast tissues where only CTCF180 is detected (Docquier et al., 2005, Docquier et al., 2009). Our preliminary experiments support this hypothesis (figure 7.1). The presence of smaller number of CTCF sites in the genome of the arrested cells may indicate that, individually, these sites organize and regulate larger chromatin domains; this hypothesis will need to be tested in the future.

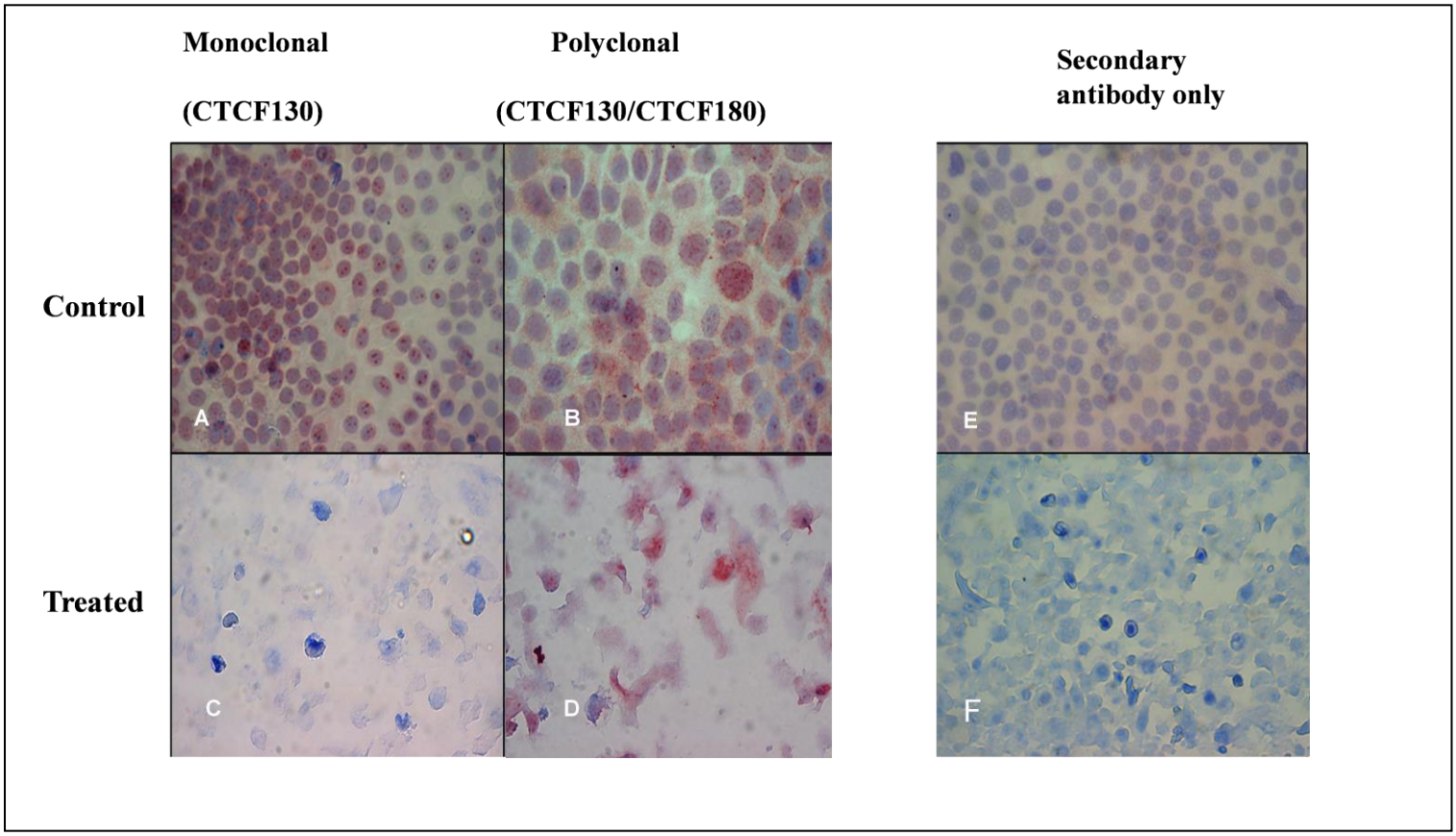


Figure 7.1 Expression profile of CTCF on aging 226LDM and HeLa cells
CTCF Immunocytochemistry stainig using control (untreated) and treated 226LDM cells. With the monoclonal CTCF antibody that only recognizes the CTCF130 isoform, the signal appears more concentrated in the nuclear area (a). With the polyclonal antibody that recognizes both isoforms the cytoplasmic stainig can be detected (b). In treated cells, the monoclonal antibody did not stain any cells (c) while with the polyclonal antibody the staining was observed (d). No primary antibody control cells were used to test the specificity of the staining. There is no staining present in this case, which confirms the specificity (e) (f).

At this point it should be noted that despite obvious advantages, there are inherent disadvantages and bias within NGS techniques, and the interpretation of their output using bioinformatics (Shendure and Ji, 2008, Treangen and Salzberg, 2012). In view of these limitations it is essential for any findings produced using these techniques to also be experimentally validated using other techniques such as real-time PCR.

Moreover, in our data analysis pipeline, closed chromatin ChIP samples were utilized as a non-specific binding events control against the CTCF ChIP samples. This strategy allowed us to focus on the open chromatin binding sites of CTCF, however an insight on the global binding pattern is needed especially since CTCF can reportedly bind to regions far away from the regulatory regions of known genes (Chen et al., 2012).

Further computational analysis of our data which could not be completed within the time limits of this study will generate additional valuable information. A comparison analysis between our and other studies could highlight similarities between the datasets and potentially reveal novel binding sites and motifs. The integration of data from other assays, such as chromosome conformation capture (3C), investigating long range interactions could elucidate the potential role of CTCF180 in chromatin looping and architecture (Dekker et al., 2002, Naumova et al., 2012, Merelli et al., 2015).

Finally, although the 226LDM cell line model was of great value for our experimental design as it represents the closest a cell line can be to the normal healthy breast cells, nonetheless it cannot match the clinical relevance or the wealth of information and that could be mined from experimenting with tissues.

In conclusion, we believe that this study constitutes an encouraging first step towards the exploration of CTCF180 binding and its potential as a regulator of gene expression.

7.3 CTCF PARylation is involved in regulation of DNA damage response pathways

In the second part of our investigation we concentrated specifically on the DNA damage response (DDR) mechanism and explored the involvement of CTCF PARylation in DDR regulation.

Although the DDR pathways are not currently fully understood, it is known that PARylation and the PARPs play an important role in DDR regulation (Oliver et al., 1999, El-Khamisy et al., 2003, Robert et al., 2013). In addition, studies have revealed links between CTCF and PARP1 (Zampieri et al., 2012, Caiafa and Zlatanova, 2009). These findings support the hypothesis that CTCF could be a part of the DDR in a PARylation-dependent manner.

To address this hypothesis we conducted experiments on 226LDM cells. The closeness of this cell line to normal cells was one of the factors for selection of these cells. Our results showed that controlled concentrations of a DNA damaging agent (H_2O_2) did not affect the expression ratio between the isoforms (figure 6.1), although the localization of the protein changed (figure 6.2). Following DNA damage, CTCF translocated into the nucleolus, where it co-localized extensively with PARP1 (figure 6.2). Interestingly, such a translocation event has been reported previously in the context of nucleolar transcription (Torrano et al., 2006, van de Nobelen et al., 2010). Whether the translocation observed in our experiments is related or not to nucleolar transcription remains unknown at the present time and further experiments would be needed to answer this question.

Subsequently, we proved that general inhibition of PARylation, caused by exposure to the ABT-888 inhibitor, has a limiting effect on the cellular repair process both by immunohistochemistry (figure 6.12) and for the first time with the automated FADU assay (figure 6.15). The observed hindering in repair is in total agreement with previous research linking PARylation and DDR (Durkacz et al., 1981a, Chevanne et al., 2010). The effect of the

inhibition was a delayed repair rather than a lack of repair activity which is consistent with previous research (Allinson et al., 2003).

Inhibition of CTCF PARylation on the other hand, did not appear to limit the speed of repair. On the contrary, FADU experiments showed that cells induced to express a mutant CTCF deficient for PARylation, recovered from damage faster as compared to the control population. A closer look inside these transfected cells however, revealed what seemed as a catastrophic effect on the nucleolus. Pending further experiments which could elucidate the implication of the discovered events further, we suggest that the DDR pathway involves the PARylation of CTCF but that it can be circumvent when this is not a possibility.

In accordance with our findings that portray a link between CTCF, PARylation and DDR, Izhar et al. (2015) showed that over 30 transcription factors are recruited at DNA damage sites in a PARP1-related manner, including CTCF. According to this study, several proteins were observed to re-localize to the sites of damage caused by UV irradiation shortly after the damage occurs. This recruitment is on many occasions completely dependent on PARP1, while the role of PARP1 and the TFs at the location includes chromatin remodelling. One of the roles of PARP1 in damage sites is to render them it accessible for the repair apparatus (Parsons et al., 2005, Izhar et al., 2015). Given the known capacity of CTCF to act as architectural protein (Ong and Corces, 2014, Xu et al., 2014) it is conceivable that CTCF could be recruited to assist with keeping chromatin accessible at the damage site.

Closely linked with DDR and cell cycle progression is the study of cellular senescence (Lou and Chen, 2006, Cichowski and Hahn, 2008). Senescence is the state of permanent cell-cycle arrest that cells are programmed to go into after a certain number of proliferation cycles or in response to accumulating stress (Campisi and d'Adda di Fagagna, 2007, Campisi, 2013). It is the field where PARylation holds a key role and there is potential for CTCF to be involved as well.

Initial experiments revealed that indeed, cells going into senescence can display a clear switch in the isoform expression (figure 7.). As it was the case with the DDR mechanism, CTCF PARylation appears to have a divergent significance in healthy compared to cancer cells. In cells resembling healthy tissues (226LDM) the expression ratio between CTCF130 and CTCF180 was decreasing with time, while in cancer cell lines (HeLa) the prevalence of CTCF130 did not appear to decrease in the time-course of our experiment. This finding implies that CTCF PARylation is important in the replicative senescence pathways and, given the links between the DDR and senescence, the argument that CTCF PARylation could play an active part in both pathways appears valid.

To continue the above studies it will be important to employ a cell system in which senescence can be induced synchronously. In this system, based on human lung fibroblasts, the telomere function can be blocked triggering senescence(van Steensel et al., 1998). These cells (termed T19) have been obtained from Prof T. von Zglinicki and tested in our laboratory. They will be used in future experiments to study the role of CTCF and CTCF PARylation in the processes of ageing and DNA damage repair.

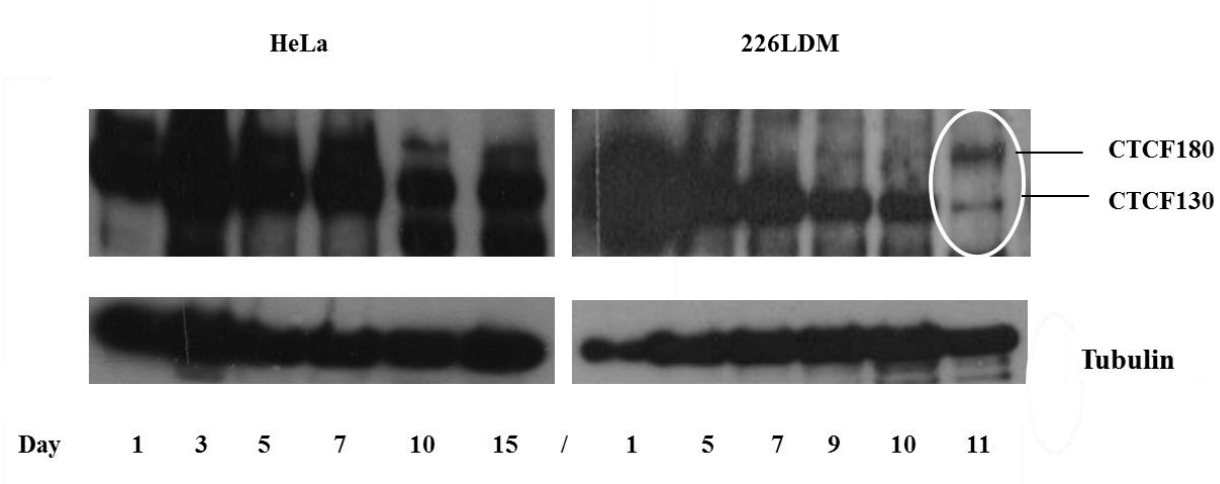


Figure 7.2 Expression profile of CTCF on aging 226LDM and HeLa cells

226LDM and HeLa cells were grown in wells of a 12-well plate and allowed to grow for consecutive days without the addition of new nutrients. Each day the contents of a well were harvested and prepared for a western blot experiment. The CTCF polyclonal antibody was used in the experiment. Tubulin was used as a loading control and the signal was visualized with the UptiLight™ chemiluminescence substrate.

7.4 Concluding Remarks

CTCF is a major regulator in a plethora of cellular processes and this has made it appealing and at the same time complex to study. The full range of the pathways that it is involved in remains realistically far from elucidated. But especially when it comes to its PARylated isoform the view is even more unclear.

Our investigation into DNA damage response suggests that CTCF PARylation is involved in the paths that are activated as part of the cellular defence to damage, particularly for non-cancer cells. Moreover, our genome-wide study provided novel insights into CTCF180 binding calling for further research that would highlight new areas of interest for this isoform.

Our initial hypotheses were that differentially PARylated isoforms of CTCF control different groups of genes (Hypothesis 1) and different functions (Hypothesis 2) in different biological situations. Our experimental findings supports Hypothesis 1 since the overall binding of CTCF130 and CTCF180 was associated with differential gene expression. Our findings also support Hypothesis 2 because CTCF PARylation is involved in DNA damage response and important for nucleolar stability following DNA damage.

Reference List

ALLINSON, S. L., DIANOVA, II & DIANOV, G. L. 2003. Poly(ADP-ribose) polymerase in base excision repair: always engaged, but not essential for DNA damage processing. *Acta Biochim Pol,* 50**,** 169-79.

ALONTAGA, A. Y., BOBKOVA, E. & CHEN, Y. 2012. Biochemical analysis of protein SUMOylation. *Curr Protoc Mol Biol,* Chapter 10**,** Unit10 29.

ANDERS, S. & HUBER, W. 2010. Differential expression analysis for sequence count data. *Genome Biol,* 11**,** R106.

ANGELINI, C. & COSTA, V. 2014. Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data: statistical solutions to biological problems. *Front Cell Dev Biol,* 2**,** 51.

ARAVIND, L. & LANDSMAN, D. 1998. AT-hook motifs identified in a wide variety of DNA-binding proteins. *Nucleic Acids Res,* 26**,** 4413-21.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. & SHERLOCK, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* 25**,** 25-9.

AWAD, T. A., BIGLER, J., ULMER, J. E., HU, Y. J., MOORE, J. M., LUTZ, M., NEIMAN, P. E., COLLINS, S. J., RENKAWITZ, R., LOBANENKOV, V. V. & FILIPPOVA, G. N. 1999. Negative transcriptional regulation mediated by thyroid hormone response element 144 requires binding of the multivalent factor CTCF to a novel target DNA sequence. *J Biol Chem,* 274**,** 27092-8.

BAILEY, T. L., WILLIAMS, N., MISLEH, C. & LI, W. W. 2006. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res,* 34**,** W369-73.

BARKESS, G. & WEST, A. G. 2012. Chromatin insulator elements: establishing barriers to set heterochromatin boundaries. *Epigenomics,* 4**,** 67-80.

BARLOW, D. P. & BARTOLOMEI, M. S. 2014. Genomic imprinting in mammals. *Cold Spring Harb Perspect Biol,* 6.

BARTON, M. 1995. Tables of equivalent dose in 2 Gy fractions: a simple application of the linear quadratic formula. *Int J Radiat Oncol Biol Phys,* 31**,** 371-8.

BECK, C., ROBERT, I., REINA-SAN-MARTIN, B., SCHREIBER, V. & DANTZER, F. 2014. Poly(ADP-ribose) polymerases in double-strand break repair: focus on PARP1, PARP2 and PARP3. *Exp Cell Res,* 329**,** 18-25.

BELL, A. C. & FELSENFELD, G. 2000. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature,* 405**,** 482-5.

BELL, A. C., WEST, A. G. & FELSENFELD, G. 1999. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell,* 98**,** 387-96.

BENJAMINI, Y. & HOCHBERG, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological),* 57**,** 289-300.

BIRNBOIM, H. C. & JEVCAK, J. J. 1981. Fluorometric Method for Rapid Detection of DNA Strand Breaks in Human White Blood Cells Produced by Low Doses of Radiation. *Cancer Research,* 41**,** 1889-1892.

BONICALZI, M. E., HAINCE, J. F., DROIT, A. & POIRIER, G. G. 2005. Regulation of poly(ADP-ribose) metabolism by poly(ADP-ribose) glycohydrolase: where and when? *Cell Mol Life Sci,* 62**,** 739-50.

BORG, I. & GROENEN, P. J. F. 2005. *Modern multidimensional scaling : theory and applications,* New York ; London, Springer.

BOUMIL, R. M., OGAWA, Y., SUN, B. K., HUYNH, K. D. & LEE, J. T. 2006. Differential methylation of Xite and CTCF sites in Tsix mirrors the pattern of X-inactivation choice in mice. *Mol Cell Biol,* 26**,** 2109-17.

BRADLEY, M. O. & ERICKSON, L. C. 1981. Comparison of the effects of hydrogen peroxide and x-ray irradiation on toxicity, mutation, and DNA damage/repair in mammalian cells (V-79). *Biochim Biophys Acta,* 654**,** 135-41.

BRASSET, E. & VAURY, C. 2005. Insulators are fundamental components of the eukaryotic genomes. *Heredity (Edinb),* 94**,** 571-6.

BRIZUELA, L., ADER, I., MAZEROLLES, C., BOCQUET, M., MALAVAUD, B. & CUVILLIER, O. 2012. First evidence of sphingosine 1-phosphate lyase protein expression and activity downregulation in human neoplasm: implication for resistance to therapeutics in prostate cancer. *Mol Cancer Ther,* 11**,** 1841-51.

BROWN, P. O. & BOTSTEIN, D. 1999. Exploring the new world of the genome with DNA microarrays. *Nat Genet,* 21**,** 33-7.

BUERKLE, A. 2008. *Poly(ADP-Ribosyl)ation*, Springer US.

BURCIN, M., ARNOLD, R., LUTZ, M., KAISER, B., RUNGE, D., LOTTSPEICH, F., FILIPPOVA, G. N., LOBANENKOV, V. V. & RENKAWITZ, R. 1997. Negative protein 1, which is required for function of the chicken lysozyme gene silencer in conjunction with hormone receptors, is identical to the multivalent zinc finger repressor CTCF. *Mol Cell Biol,* 17**,** 1281-8.

BURKLE, A. 2000. Poly(ADP-ribosyl)ation: a posttranslational protein modification linked with genome protection and mammalian longevity. *Biogerontology,* 1**,** 41-6.

BURKLE, A. 2005. Poly(ADP-ribose). The most elaborate metabolite of NAD+. *FEBS J,* 272**,** 4576-89.

CAIAFA, P. & ZLATANOVA, J. 2009. CCCTC-binding factor meets poly(ADP-ribose) polymerase-1. *J Cell Physiol,* 219**,** 265-70.

CAMPISI, J. 2013. Aging, cellular senescence, and cancer. *Annu Rev Physiol,* 75**,** 685-705.

CAMPISI, J. & D'ADDA DI FAGAGNA, F. 2007. Cellular senescence: when bad things happen to good cells. *Nat Rev Mol Cell Biol,* 8**,** 729-40.

CAPRARA, G., ZAMPONI, R., MELIXETIAN, M. & HELIN, K. 2009. Isolation and characterization of DUSP11, a novel p53 target gene. *J Cell Mol Med,* 13**,** 2158-70.

CHANG, P., COUGHLIN, M. & MITCHISON, T. J. 2005. Tankyrase-1 polymerization of poly(ADP-ribose) is required for spindle structure and function. *Nat Cell Biol,* 7**,** 1133-9.

CHAO, W., HUYNH, K. D., SPENCER, R. J., DAVIDOW, L. S. & LEE, J. T. 2002. CTCF, a candidate trans-acting factor for X-inactivation choice. *Science,* 295**,** 345-7.

CHEN, H., TIAN, Y., SHU, W., BO, X. & WANG, S. 2012. Comprehensive identification and annotation of cell type-specific and ubiquitous CTCF-binding sites in the human genome. *PLoS One,* 7**,** e41374.

CHERNUKHIN, I., SHAMSUDDIN, S., KANG, S. Y., BERGSTROM, R., KWON, Y. W., YU, W., WHITEHEAD, J., MUKHOPADHYAY, R., DOCQUIER, F., FARRAR, D., MORRISON, I., VIGNERON, M., WU, S. Y., CHIANG, C. M., LOUKINOV, D., LOBANENKOV, V., OHLSSON, R. & KLENOVA, E. 2007. CTCF interacts with and recruits the largest subunit of RNA polymerase II to CTCF target sites genome-wide. *Mol Cell Biol,* 27**,** 1631-48.

CHEVANNE, M., ZAMPIERI, M., CALDINI, R., RIZZO, A., CICCARONE, F., CATIZONE, A., D'ANGELO, C., GUASTAFIERRO, T., BIROCCIO, A., REALE, A., ZUPI, G. & CAIAFA, P. 2010. Inhibition of PARP activity by PJ-34 leads to growth impairment and cell death associated with aberrant mitotic pattern and nucleolar actin accumulation in M14 melanoma cell line. *J Cell Physiol,* 222**,** 401-10.

CHOI, J. H., MIN, N. Y., PARK, J., KIM, J. H., PARK, S. H., KO, Y. J., KANG, Y., MOON, Y. J., RHEE, S., HAM, S. W., PARK, A. J. & LEE, K. H. 2010. TSA-induced DNMT1 down-regulation represses hTERT expression via recruiting CTCF into demethylated core promoter region of hTERT in HCT116. *Biochem Biophys Res Commun,* 391**,** 449-54.

CHRAMBACH, A. & RODBARD, D. 1971. Polyacrylamide gel electrophoresis. *Science,* 172**,** 440-51.

CICCARONE, F., VALENTINI, E., BACALINI, M. G., ZAMPIERI, M., CALABRESE, R., GUASTAFIERRO, T., MARIANO, G., REALE, A., FRANCESCHI, C. & CAIAFA, P. 2014. Poly(ADP-ribosyl)ation is involved in the epigenetic control of TET1 gene transcription. *Oncotarget,* 5**,** 10356-67.

CICCARONE, F., VALENTINI, E., ZAMPIERI, M. & CAIAFA, P. 2015. 5mC-hydroxylase activity is influenced by the PARylation of TET1 enzyme. *Oncotarget,* 6**,** 24333-47.

CICHOWSKI, K. & HAHN, W. C. 2008. Unexpected pieces to the senescence puzzle. *Cell,* 133**,** 958-61.

COCK, P. J., FIELDS, C. J., GOTO, N., HEUER, M. L. & RICE, P. M. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res,* 38**,** 1767-71.

COLLAS, P. 2010. The current state of chromatin immunoprecipitation. *Mol Biotechnol,* 45**,** 87-100.
COOK, B. D., DYNEK, J. N., CHANG, W., SHOSTAK, G. & SMITH, S. 2002. Role for the related poly(ADP-Ribose) polymerases tankyrase 1 and 2 at human telomeres. *Mol Cell Biol,* 22**,** 332-42.
COSTA, V., APRILE, M., ESPOSITO, R. & CICCODICOLA, A. 2013. RNA-Seq and human complex diseases: recent accomplishments and future perspectives. *Eur J Hum Genet,* 21**,** 134-42.
COTSIKI, M., GJOERUP, O., JAT, P. & ROBERTS, T. 2005. Immortalisation of mammalian cells and therapeutic applications of said cells. Google Patents.
CREMONA, C. A., SARANGI, P. & ZHAO, X. 2012. Sumoylation and the DNA damage response. *Biomolecules,* 2**,** 376-88.
CULLUM, R., ALDER, O. & HOODLESS, P. A. 2011. The next generation: using new sequencing technologies to analyse gene regulation. *Respirology,* 16**,** 210-22.
CURTIN, N. J. 2012. DNA repair dysregulation from cancer driver to therapeutic target. *Nat Rev Cancer,* 12**,** 801-17.
DEKKER, J., RIPPE, K., DEKKER, M. & KLECKNER, N. 2002. Capturing chromosome conformation. *Science,* 295**,** 1306-11.
DELGADO, M. D., CHERNUKHIN, I. V., BIGAS, A., KLENOVA, E. M. & LEON, J. 1999. Differential expression and phosphorylation of CTCF, a c-myc transcriptional regulator, during differentiation of human myeloid cells. *FEBS Lett,* 444**,** 5-10.
DESJARDINS, P. & CONKLIN, D. 2010. NanoDrop microvolume quantitation of nucleic acids. *J Vis Exp*.
DOCQUIER, F., FARRAR, D., D'ARCY, V., CHERNUKHIN, I., ROBINSON, A., LOUKINOV, D., VATOLIN, S., PACK, S., MACKAY, A., HARRIS, R., DORRICOTT, H., O'HARE, M., LOBANENKOV, V. & KLENOVA, E. 2005. Heightened expression of CTCF in breast cancer cells is associated with resistance to apoptosis. *Cancer Res,* 65**,** 5112 - 5122.
DOCQUIER, F., KITA, G. X., FARRAR, D., JAT, P., O'HARE, M., CHERNUKHIN, I., GRETTON, S., MANDAL, A., ALLDRIDGE, L. & KLENOVA, E. 2009. Decreased poly(ADP-ribosyl)ation of CTCF, a transcription factor, is associated with breast cancer phenotype and cell proliferation. *Clin Cancer Res,* 15**,** 5762-71.
DUBRIDGE, R. B., TANG, P., HSIA, H. C., LEONG, P. M., MILLER, J. H. & CALOS, M. P. 1987. Analysis of mutation in human cells by using an Epstein-Barr virus shuttle system. *Mol Cell Biol,* 7**,** 379-87.
DUQUE-PARRA, J. E. 2005. Note on the origin and history of the term "apoptosis". *Anat Rec B New Anat,* 283**,** 2-4.
DURKACZ, B. W., IRWIN, J. & SHALL, S. 1981a. Inhibition of (ADP-ribose)n biosynthesis retards DNA repair but does not inhibit DNA repair synthesis. *Biochemical and Biophysical Research Communications,* 101**,** 1433-1441.
DURKACZ, B. W., SHALL, S. & IRWIN, J. 1981b. The effect of inhibition of (ADP-ribose)n biosynthesis on DNA repair assayed by the nucleoid technique. *European Journal Of Biochemistry / FEBS,* 121**,** 65-69.
EGGELING, R., GOHR, A., KEILWAGEN, J., MOHR, M., POSCH, S., SMITH, A. D. & GROSSE, I. 2014. On the value of intra-motif dependencies of human insulator protein CTCF. *PLoS One,* 9**,** e85629.
EL-KHAMISY, S. F., MASUTANI, M., SUZUKI, H. & CALDECOTT, K. W. 2003. A requirement for PARP-1 for the assembly or stability of XRCC1 nuclear foci at sites of oxidative DNA damage. *Nucleic Acids Res,* 31**,** 5526-33.
ENGEL, L. W., YOUNG, N. A., TRALKA, T. S., LIPPMAN, M. E., O'BRIEN, S. J. & JOYCE, M. J. 1978. Establishment and characterization of three new continuous cell lines derived from human breast carcinomas. *Cancer Res,* 38**,** 3352-64.
ESSELTINE, J. L. & SCOTT, J. D. 2013. AKAP signaling complexes: pointing towards the next generation of therapeutic targets? *Trends Pharmacol Sci,* 34**,** 648-55.
FARRAR, D., CHERNUKHIN, I. & KLENOVA, E. 2011. Generation of Poly(ADP-ribosyl)ation Deficient Mutants of the Transcription Factor, CTCF. *In:* TULIN, A. V. (ed.) *Poly(ADP-ribose) Polymerase.* Humana Press.
FARRAR, D., RAI, S., CHERNUKHIN, I., JAGODIC, M., ITO, Y., YAMMINE, S., OHLSSON, R., MURRELL, A. & KLENOVA, E. 2010. Mutational Analysis of the Poly(ADP-Ribosyl)ation Sites of the Transcription

Factor CTCF Provides an Insight into the Mechanism of Its Regulation by Poly(ADP-Ribosyl)ation. *Molecular and Cellular Biology,* 30**,** 1199-1216.

FENG, J., LIU, T., QIN, B., ZHANG, Y. & LIU, X. S. 2012. Identifying ChIP-seq enrichment using MACS. *Nat Protoc,* 7**,** 1728-40.

FILIPPOVA, G. N. 2008. Genetics and epigenetics of the multifunctional protein CTCF. *Curr Top Dev Biol,* 80**,** 337-60.

FILIPPOVA, G. N., FAGERLIE, S., KLENOVA, E. M., MYERS, C., DEHNER, Y., GOODWIN, G., NEIMAN, P. E., COLLINS, S. J. & LOBANENKOV, V. V. 1996. An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Mol Cell Biol,* 16**,** 2802-13.

FILIPPOVA, G. N., LINDBLOM, A., MEINCKE, L. J., KLENOVA, E. M., NEIMAN, P. E., COLLINS, S. J., DOGGETT, N. A. & LOBANENKOV, V. V. 1998. A widely expressed transcription factor with multiple DNA sequence specificity, CTCF, is localized at chromosome segment 16q22.1 within one of the smallest regions of overlap for common deletions in breast and prostate cancers. *Genes Chromosomes Cancer,* 22**,** 26-36.

FILIPPOVA, G. N., QI, C. F., ULMER, J. E., MOORE, J. M., WARD, M. D., HU, Y. J., LOUKINOV, D. I., PUGACHEVA, E. M., KLENOVA, E. M., GRUNDY, P. E., FEINBERG, A. P., CLETON-JANSEN, A. M., MOERLAND, E. W., CORNELISSE, C. J., SUZUKI, H., KOMIYA, A., LINDBLOM, A., DORION-BONNET, F., NEIMAN, P. E., MORSE, H. C., 3RD, COLLINS, S. J. & LOBANENKOV, V. V. 2002. Tumor-associated zinc finger mutations in the CTCF transcription factor selectively alter tts DNA-binding specificity. *Cancer Res,* 62**,** 48-52.

FOX, E. J., REID-BAYLISS, K. S., EMOND, M. J. & LOEB, L. A. 2014. Accuracy of Next Generation Sequencing Platforms. *Next Gener Seq Appl,* 1.

GENE ONTOLOGY, C. 2008. The Gene Ontology project in 2008. *Nucleic Acids Res,* 36**,** D440-4.

GERSCHENSON, L. E. & ROTELLO, R. J. 1992. Apoptosis: a different type of cell death. *FASEB J,* 6**,** 2450-5.

GIARDINE, B., RIEMER, C., HARDISON, R. C., BURHANS, R., ELNITSKI, L., SHAH, P., ZHANG, Y., BLANKENBERG, D., ALBERT, I., TAYLOR, J., MILLER, W., KENT, W. J. & NEKRUTENKO, A. 2005. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res,* 15**,** 1451-5.

GOECKS, J., NEKRUTENKO, A., TAYLOR, J. & GALAXY, T. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol,* 11**,** R86.

GOENECHEA, L. G., RENDON, M. C., IGLESIAS, C. & VALDIVIA, M. M. 1992. Immunostaining of nucleolus organizers in mammalian cells by a human autoantibody against the polymerase I transcription factor UBF. *Cell Mol Biol (Noisy-le-grand),* 38**,** 841-51.

GOLIA, B., SINGH, H. R. & TIMINSZKY, G. 2015. Poly-ADP-ribosylation signaling during DNA damage repair. *Front Biosci (Landmark Ed),* 20**,** 440-57.

GOMEZ-CABRERO, D., ABUGESSAISA, I., MAIER, D., TESCHENDORFF, A., MERKENSCHLAGER, M., GISEL, A., BALLESTAR, E., BONGCAM-RUDLOFF, E., CONESA, A. & TEGNER, J. 2014. Data integration in the era of omics: current and future challenges. *BMC Syst Biol,* 8 Suppl 2**,** I1.

GREEN, A. R., KRIVINSKAS, S., YOUNG, P., RAKHA, E. A., PAISH, E. C., POWE, D. G. & ELLIS, I. O. 2009. Loss of expression of chromosome 16q genes DPEP1 and CTCF in lobular carcinoma in situ of the breast. *Breast Cancer Res Treat,* 113**,** 59-66.

GUASTAFIERRO, T., CATIZONE, A., CALABRESE, R., ZAMPIERI, M., MARTELLA, O., BACALINI, M. G., REALE, A., DI GIROLAMO, M., MICCHELI, M., FARRAR, D., KLENOVA, E., CICCARONE, F. & CAIAFA, P. 2013. ADP-ribose polymer depletion leads to nuclear Ctcf re-localization and chromatin rearrangement(1). *Biochem J,* 449**,** 623-30.

GUASTAFIERRO, T., CECCHINELLI, B., ZAMPIERI, M., REALE, A., RIGGIO, G., STHANDIER, O., ZUPI, G., CALABRESE, L. & CAIAFA, P. 2008. CCCTC-binding factor activates PARP-1 affecting DNA methylation machinery. *J Biol Chem,* 283**,** 21873-80.

GUERRERO, P. A. & MAGGERT, K. A. 2011. The CCCTC-binding factor (CTCF) of Drosophila contributes to the regulation of the ribosomal DNA and nucleolar stability. *PLoS One,* 6**,** e16401.

HANAHAN, D., JESSEE, J. & BLOOM, F. R. 1991. Plasmid transformation of Escherichia coli and other bacteria. *Methods Enzymol,* 204**,** 63-113.

HANCOCK, A. L., BROWN, K. W., MOORWOOD, K., MOON, H., HOLMGREN, C., MARDIKAR, S. H., DALLOSSO, A. R., KLENOVA, E., LOUKINOV, D., OHLSSON, R., LOBANENKOV, V. V. & MALIK, K. 2007. A CTCF-binding silencer regulates the imprinted genes AWT1 and WT1-AS and exhibits sequential epigenetic defects during Wilms' tumourigenesis. *Hum Mol Genet,* 16**,** 343-54.

HANDOKO, L., XU, H., LI, G., NGAN, C. Y., CHEW, E., SCHNAPP, M., LEE, C. W., YE, C., PING, J. L., MULAWADI, F., WONG, E., SHENG, J., ZHANG, Y., POH, T., CHAN, C. S., KUNARSO, G., SHAHAB, A., BOURQUE, G., CACHEUX-RATABOUL, V., SUNG, W. K., RUAN, Y. & WEI, C. L. 2011. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet,* 43**,** 630-8.

HARRIS, M. A., CLARK, J., IRELAND, A., LOMAX, J., ASHBURNER, M., FOULGER, R., EILBECK, K., LEWIS, S., MARSHALL, B., MUNGALL, C., RICHTER, J., RUBIN, G. M., BLAKE, J. A., BULT, C., DOLAN, M., DRABKIN, H., EPPIG, J. T., HILL, D. P., NI, L., RINGWALD, M., BALAKRISHNAN, R., CHERRY, J. M., CHRISTIE, K. R., COSTANZO, M. C., DWIGHT, S. S., ENGEL, S., FISK, D. G., HIRSCHMAN, J. E., HONG, E. L., NASH, R. S., SETHURAMAN, A., THEESFELD, C. L., BOTSTEIN, D., DOLINSKI, K., FEIERBACH, B., BERARDINI, T., MUNDODI, S., RHEE, S. Y., APWEILER, R., BARRELL, D., CAMON, E., DIMMER, E., LEE, V., CHISHOLM, R., GAUDET, P., KIBBE, W., KISHORE, R., SCHWARZ, E. M., STERNBERG, P., GWINN, M., HANNICK, L., WORTMAN, J., BERRIMAN, M., WOOD, V., DE LA CRUZ, N., TONELLATO, P., JAISWAL, P., SEIGFRIED, T., WHITE, R. & GENE ONTOLOGY, C. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res,* 32**,** D258-61.

HASSLER, M., JANKEVICIUS, G. & LADURNER, A. G. 2011. PARG: a macrodomain in disguise. *Structure,* 19**,** 1351-3.

HAWKINS, R. D., HON, G. C. & REN, B. 2010. Next-generation genomics: an integrative approach. *Nat Rev Genet,* 11**,** 476-86.

HAYWARD, S. W., DAHIYA, R., CUNHA, G. R., BARTEK, J., DESHPANDE, N. & NARAYAN, P. 1995. Establishment and characterization of an immortalized but non-transformed human prostate epithelial cell line: BPH-1. *In Vitro Cell Dev Biol Anim,* 31**,** 14-24.

HEATH, H., RIBEIRO DE ALMEIDA, C., SLEUTELS, F., DINGJAN, G., VAN DE NOBELEN, S., JONKERS, I., LING, K. W., GRIBNAU, J., RENKAWITZ, R., GROSVELD, F., HENDRIKS, R. W. & GALJART, N. 2008. CTCF regulates cell cycle progression of alphabeta T cells in the thymus. *EMBO J,* 27**,** 2839-50.

HEERES, J. T. & HERGENROTHER, P. J. 2007. Poly(ADP-ribose) makes a date with death. *Curr Opin Chem Biol,* 11**,** 644-53.

HEGEDUS, C. & VIRAG, L. 2014. Inputs and outputs of poly(ADP-ribosyl)ation: Relevance to oxidative stress. *Redox Biol,* 2C**,** 978-982.

HENLE, E. S. & LINN, S. 1997. Formation, Prevention, and Repair of DNA Damage by Iron/Hydrogen Peroxide. *Journal of Biological Chemistry,* 272**,** 19095-19098.

HERNANDEZ-HERNANDEZ, A., SOTO-REYES, E., ORTIZ, R., ARRIAGA-CANON, C., ECHEVERRIA-MARTINEZ, O. M., VAZQUEZ-NIN, G. H. & RECILLAS-TARGA, F. 2012. Changes of the nucleolus architecture in absence of the nuclear factor CTCF. *Cytogenet Genome Res,* 136**,** 89-96.

HEROLD, M., BARTKUHN, M. & RENKAWITZ, R. 2012. CTCF: insights into insulator function during development. *Development,* 139**,** 1045-57.

HINE, R. 2008. *A dictionary of biology,* Oxford, Oxford University Press.

HOEIJMAKERS, J. H. 2009. DNA damage, aging, and cancer. *N Engl J Med,* 361**,** 1475-85.

HOLWERDA, S. J. & DE LAAT, W. 2013. CTCF: the protein, the binding partners, the binding sites and their chromatin loops. *Philos Trans R Soc Lond B Biol Sci,* 368**,** 20120369.

HORE, T. A., DEAKIN, J. E. & MARSHALL GRAVES, J. A. 2008. The evolution of epigenetic regulators CTCF and BORIS/CTCFL in amniotes. *PLoS Genet,* 4**,** e1000169.

HUANG DA, W., SHERMAN, B. T. & LEMPICKI, R. A. 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res,* 37**,** 1-13.

HUANG DA, W., SHERMAN, B. T. & LEMPICKI, R. A. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc,* 4**,** 44-57.

HUANG, K., JIA, J., WU, C., YAO, M., LI, M., JIN, J., JIANG, C., CAI, Y., PEI, D., PAN, G. & YAO, H. 2013. Ribosomal RNA gene transcription mediated by the master genome regulator protein CCCTC-binding factor (CTCF) is negatively regulated by the condensin complex. *J Biol Chem,* 288**,** 26067-77.

HUNTZINGER, E. & IZAURRALDE, E. 2011. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat Rev Genet,* 12**,** 99-110.

IDERAABDULLAH, F. Y., THORVALDSEN, J. L., MYERS, J. A. & BARTOLOMEI, M. S. 2014. Tissue-specific insulator function at H19/Igf2 revealed by deletions at the imprinting control region. *Hum Mol Genet,* 23**,** 6246-59.

IMLAY, J. A., CHIN, S. M. & LINN, S. 1988. Toxic DNA damage by hydrogen peroxide through the Fenton reaction in vivo and in vitro. *Science,* 240**,** 640-2.

IVANSCHITZ, L., DE THE, H. & LE BRAS, M. 2013. PML, SUMOylation, and Senescence. *Front Oncol,* 3**,** 171.

IZHAR, L., ADAMSON, B., CICCIA, A., LEWIS, J., PONTANO-VAITES, L., LENG, Y., LIANG, A. C., WESTBROOK, T. F., HARPER, J. W. & ELLEDGE, S. J. 2015. A Systematic Analysis of Factors Localized to Damaged Chromatin Reveals PARP-Dependent Recruitment of Transcription Factors. *Cell Rep,* 11**,** 1486-500.

JACKSON, S. P. & BARTEK, J. 2009. The DNA-damage response in human biology and disease. *Nature,* 461**,** 1071-8.

JOHNSON, D. S., MORTAZAVI, A., MYERS, R. M. & WOLD, B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science,* 316**,** 1497-502.

JORDAN, M. A., THROWER, D. & WILSON, L. 1992. Effects of vinblastine, podophyllotoxin and nocodazole on mitotic spindles. Implications for the role of microtubule dynamics in mitosis. *J Cell Sci,* 102 (Pt 3)**,** 401-16.

KABOORD, B. & PERR, M. 2008. Isolation of proteins and protein complexes by immunoprecipitation. *Methods Mol Biol,* 424**,** 349-64.

KANDURI, C., PANT, V., LOUKINOV, D., PUGACHEVA, E., QI, C. F., WOLFFE, A., OHLSSON, R. & LOBANENKOV, V. V. 2000. Functional association of CTCF with the insulator upstream of the H19 gene is parent of origin-specific and methylation-sensitive. *Curr Biol,* 10**,** 853-6.

KIM, M. Y. 2011. Regulation of chromatin structure by PARP-1. *Methods Mol Biol,* 780**,** 227-36.

KIM, M. Y., ZHANG, T. & KRAUS, W. L. 2005a. Poly(ADP-ribosyl)ation by PARP-1: 'PAR-laying' NAD+ into a nuclear signal. *Genes Dev,* 19**,** 1951-67.

KIM, M. Y., ZHANG, T. & KRAUS, W. L. 2005b. Poly(ADP-ribosyl)ation by PARP-1: `PAR-laying' NAD+ into a nuclear signal. *Genes & Development,* 19**,** 1951-1967.

KIM, T. H., ABDULLAEV, Z. K., SMITH, A. D., CHING, K. A., LOUKINOV, D. I., GREEN, R. D., ZHANG, M. Q., LOBANENKOV, V. V. & REN, B. 2007. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell,* 128**,** 1231-45.

KIM, Y. J., CECCHINI, K. R. & KIM, T. H. 2011. Conserved, developmentally regulated mechanism couples chromosomal looping and heterochromatin barrier activity at the homeobox gene A locus. *Proc Natl Acad Sci U S A,* 108**,** 7391-6.

KINGSTON, R. E., CHEN, C. A. & OKAYAMA, H. 2001. Calcium phosphate transfection. *Curr Protoc Immunol,* Chapter 10**,** Unit 10 13.

KIRKWOOD, T. B. 2005. Understanding the odd science of aging. *Cell,* 120**,** 437-47.

KLENOVA, E. M., CHERNUKHIN, I. V., EL-KADY, A., LEE, R. E., PUGACHEVA, E. M., LOUKINOV, D. I., GOODWIN, G. H., DELGADO, D., FILIPPOVA, G. N., LEON, J., MORSE, H. C., 3RD, NEIMAN, P. E. & LOBANENKOV, V. V. 2001. Functional phosphorylation sites in the C-terminal region of the multivalent multifunctional transcriptional factor CTCF. *Mol Cell Biol,* 21**,** 2221-34.

KLENOVA, E. M., FAGERLIE, S., FILIPPOVA, G. N., KRETZNER, L., GOODWIN, G. H., LORING, G., NEIMAN, P. E. & LOBANENKOV, V. V. 1998. Characterization of the chicken CTCF genomic locus, and initial study of the cell cycle-regulated promoter of the gene. *J Biol Chem,* 273**,** 26571-9.

KLENOVA, E. M., MORSE, H. C., 3RD, OHLSSON, R. & LOBANENKOV, V. V. 2002. The novel BORIS + CTCF gene family is uniquely involved in the epigenetics of normal biology and cancer. *Semin Cancer Biol,* 12**,** 399-414.

KLENOVA, E. M., NICOLAS, R. H., PATERSON, H. F., CARNE, A. F., HEATH, C. M., GOODWIN, G. H., NEIMAN, P. E. & LOBANENKOV, V. V. 1993. CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms. *Mol Cell Biol,* 13**,** 7612-24.

KLENOVA, E. M., NICOLAS, R. H., U, S., CARNE, A. F., LEE, R. E., LOBANENKOV, V. V. & GOODWIN, G. H. 1997. Molecular weight abnormalities of the CTCF transcription factor: CTCF migrates aberrantly in SDS-PAGE and the size of the expressed protein is affected by the UTRs and sequences within the coding region of the CTCF gene. *Nucleic Acids Res,* 25**,** 466-74.

KO, H. L. & REN, E. C. 2012. Functional Aspects of PARP1 in DNA Repair and Transcription. *Biomolecules,* 2**,** 524-48.

KUO, L. J. & YANG, L. X. 2008. Gamma-H2AX - a novel biomarker for DNA double-strand breaks. *In Vivo,* 22**,** 305-9.

KYRCHANOVA, O., MAKSIMENKO, O., STAKHOV, V., IVLIEVA, T., PARSHIKOV, A., STUDITSKY, V. M. & GEORGIEV, P. 2013. Effective blocking of the white enhancer requires cooperation between two main mechanisms suggested for the insulator function. *PLoS Genet,* 9**,** e1003606.

LANDRISCINA, M., SCHINZARI, G., DI LEONARDO, G., QUIRINO, M., CASSANO, A., D'ARGENTO, E., LAURIOLA, L., SCERRATI, M., PRUDOVSKY, I. & BARONE, C. 2006. S100A13, a new marker of angiogenesis in human astrocytic gliomas. *J Neurooncol,* 80**,** 251-9.

LEHTIO, L., COLLINS, R., VAN DEN BERG, S., JOHANSSON, A., DAHLGREN, L. G., HAMMARSTROM, M., HELLEDAY, T., HOLMBERG-SCHIAVONE, L., KARLBERG, T. & WEIGELT, J. 2008. Zinc binding catalytic domain of human tankyrase 1. *J Mol Biol,* 379**,** 136-45.

LEPECQ, J. B. & PAOLETTI, C. 1967. A fluorescent complex between ethidium bromide and nucleic acids. Physical-chemical characterization. *J Mol Biol,* 27**,** 87-106.

LEWIS, A. & MURRELL, A. 2004. Genomic imprinting: CTCF protects the boundaries. *Curr Biol,* 14**,** R284-6.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. & GENOME PROJECT DATA PROCESSING, S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-9.

LIBERMANN, T. A. & ZERBINI, L. F. 2006. Targeting transcription factors for cancer gene therapy. *Curr Gene Ther,* 6**,** 17-33.

LINDBERG, J. & LUNDEBERG, J. 2010. The plasticity of the mammalian transcriptome. *Genomics,* 95**,** 1-6.

LIU, L., LI, Y., LI, S., HU, N., HE, Y., PONG, R., LIN, D., LU, L. & LAW, M. 2012. Comparison of next-generation sequencing systems. *J Biomed Biotechnol,* 2012**,** 251364.

LJUNGMAN, M. 2010. The DNA damage response--repair or despair? *Environ Mol Mutagen,* 51**,** 879-89.

LOBANENKOV, V. V., NICOLAS, R. H., ADLER, V. V., PATERSON, H., KLENOVA, E. M., POLOTSKAJA, A. V. & GOODWIN, G. H. 1990. A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. *Oncogene,* 5**,** 1743-53.

LOMAN, N. J., MISRA, R. V., DALLMAN, T. J., CONSTANTINIDOU, C., GHARBIA, S. E., WAIN, J. & PALLEN, M. J. 2012. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol,* 30**,** 434-9.

LOU, Z. & CHEN, J. 2006. Cellular senescence and DNA repair. *Exp Cell Res,* 312**,** 2641-6.

MACPHERSON, M. J., BEATTY, L. G., ZHOU, W., DU, M. & SADOWSKI, P. D. 2009. The CTCF insulator protein is posttranslationally modified by SUMO. *Mol Cell Biol,* 29**,** 714-25.

MACPHERSON, M. J. & SADOWSKI, P. D. 2010. The CTCF insulator protein forms an unusual DNA structure. *BMC Mol Biol,* 11**,** 101.

MALANGA, M. & ALTHAUS, F. R. 2005. The role of poly(ADP-ribose) in the DNA damage signaling network. *Biochem Cell Biol,* 83**,** 354-64.

MARGUERAT, S. & BAHLER, J. 2010. RNA-seq: from technology to biology. *Cell Mol Life Sci,* 67**,** 569-79.

MARIONI, J. C., MASON, C. E., MANE, S. M., STEPHENS, M. & GILAD, Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res,* 18**,** 1509-17.

MCCARREY, J. R., WATSON, C., ATENCIO, J., OSTERMEIER, G. C., MARAHRENS, Y., JAENISCH, R. & KRAWETZ, S. A. 2002. X-chromosome inactivation during spermatogenesis is regulated by an Xist/Tsix-independent mechanism in the mouse. *Genesis,* 34**,** 257-66.

MEDER, V. S., BOEGLIN, M., DE MURCIA, G. & SCHREIBER, V. 2005. PARP-1 and PARP-2 interact with nucleophosmin/B23 and accumulate in transcriptionally active nucleoli. *J Cell Sci,* 118**,** 211-22.

MEGNIN-CHANET, F., BOLLET, M. A. & HALL, J. 2010. Targeting poly(ADP-ribose) polymerase activity for cancer therapy. *Cell Mol Life Sci,* 67**,** 3649-62.

MENDOZA-ALVAREZ, H., CHAVEZ-BUENO, S. & ALVAREZ-GONZALEZ, R. 2000. Chain length analysis of ADP-ribose polymers generated by poly(ADP-ribose) polymerase (PARP) as a function of beta-NAD+ and enzyme concentrations. *IUBMB Life,* 50**,** 145-9.

MERELLI, I., TORDINI, F., DROCCO, M., ALDINUCCI, M., LIO, P. & MILANESI, L. 2015. Integrating multi-omic features exploiting Chromosome Conformation Capture data. *Front Genet,* 6**,** 40.

MEYER-FICCA, M. L., MEYER, R. G., COYLE, D. L., JACOBSON, E. L. & JACOBSON, M. K. 2004. Human poly(ADP-ribose) glycohydrolase is expressed in alternative splice variants yielding isoforms that localize to different cell compartments. *Exp Cell Res,* 297**,** 521-32.

MILLAU, J. F. & GAUDREAU, L. 2011. CTCF, cohesin, and histone variants: connecting the genome. *Biochem Cell Biol,* 89**,** 505-13.

MORENO-VILLANUEVA, M., ELTZE, T., DRESSLER, D., BERNHARDT, J., HIRSCH, C., WICK, P., VON SCHEVEN, G., LEX, K. & BURKLE, A. 2011. The automated FADU-assay, a potential high-throughput in vitro method for early screening of DNA breakage. *ALTEX,* 28**,** 295-303.

MORENO-VILLANUEVA, M., PFEIFFER, R., SINDLINGER, T., LEAKE, A., MULLER, M., KIRKWOOD, T. B. & BURKLE, A. 2009. A modified and automated version of the 'Fluorimetric Detection of Alkaline DNA Unwinding' method to quantify formation and repair of DNA strand breaks. *BMC Biotechnol,* 9**,** 39.

MOUTA CARREIRA, C., LAVALLEE, T. M., TARANTINI, F., JACKSON, A., LATHROP, J. T., HAMPTON, B., BURGESS, W. H. & MACIAG, T. 1998. S100A13 is involved in the regulation of fibroblast growth factor-1 and p40 synaptotagmin-1 release in vitro. *J Biol Chem,* 273**,** 22224-31.

MUTZ, K. O., HEILKENBRINKER, A., LONNE, M., WALTER, J. G. & STAHL, F. 2013. Transcriptome analysis using next-generation sequencing. *Curr Opin Biotechnol,* 24**,** 22-30.

NAGANO, T. & FRASER, P. 2011. No-nonsense functions for long noncoding RNAs. *Cell,* 145**,** 178-81.

NAGARAJAN, N., NG, P. & KEICH, U. 2006. Refining motif finders with E-value calculations. *RECOMB on Regulatory Genomics,* 73.

NAKAMURA, J., PURVIS, E. R. & SWENBERG, J. A. 2003. Micromolar concentrations of hydrogen peroxide induce oxidative DNA lesions more efficiently than millimolar concentrations in mammalian cells. *Nucleic Acids Res,* 31**,** 1790-5.

NALABOTHULA, N., AL-JUMAILY, T., ETELEEB, A. M., FLIGHT, R. M., XIAORONG, S., MOSELEY, H., ROUCHKA, E. C. & FONDUFE-MITTENDORF, Y. N. 2015. Genome-Wide Profiling of PARP1 Reveals an Interplay with Gene Regulatory Regions and DNA Methylation. *PLoS One,* 10**,** e0135410.

NAUMOVA, N., SMITH, E. M., ZHAN, Y. & DEKKER, J. 2012. Analysis of long-range chromatin interactions using Chromosome Conformation Capture. *Methods,* 58**,** 192-203.

NGUEWA, P. A., FUERTES, M. A., VALLADARES, B., ALONSO, C. & PEREZ, J. M. 2005. Poly(ADP-ribose) polymerases: homology, structural domains and functions. Novel therapeutical applications. *Prog Biophys Mol Biol,* 88**,** 143-72.

NI, X., ZHANG, Y. E., NEGRE, N., CHEN, S., LONG, M. & WHITE, K. P. 2012. Adaptive evolution and the birth of CTCF binding sites in the Drosophila genome. *PLoS Biol,* 10**,** e1001420.

NIELSEN, R. & MANDRUP, S. 2014. Genome-Wide Profiling of Transcription Factor Binding and Epigenetic Marks in Adipocytes by ChIP-seq. *In:* MACDOUGALD, O. A. (ed.) *Methods of Adipose Tissue Biology.* Elsevier Science.

OHLSSON, R., LOBANENKOV, V. & KLENOVA, E. 2010. Does CTCF mediate between nuclear organization and gene expression? *Bioessays,* 32**,** 37-50.

OHLSSON, R., RENKAWITZ, R. & LOBANENKOV, V. 2001. CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet,* 17**,** 520-7.

OKA, J., UEDA, K., HAYAISHI, O., KOMURA, H. & NAKANISHI, K. 1984. ADP-ribosyl protein lyase. Purification, properties, and identification of the product. *J Biol Chem,* 259**,** 986-95.

OKUDELA, K., YAZAWA, T., ISHII, J., WOO, T., MITSUI, H., BUNAI, T., SAKAEDA, M., SHIMOYAMADA, H., SATO, H., TAJIRI, M., OGAWA, N., MASUDA, M., SUGIMURA, H. & KITAMURA, H. 2009. Down-regulation of FXYD3 expression in human lung cancers: its mechanism and potential role in carcinogenesis. *Am J Pathol,* 175**,** 2646-56.

OLIVER, F. J., MENISSIER-DE MURCIA, J. & DE MURCIA, G. 1999. Poly(ADP-ribose) polymerase in the cellular response to DNA damage, apoptosis, and disease. *Am J Hum Genet,* 64**,** 1282-8.

ONG, C. T. & CORCES, V. G. 2014. CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet,* 15**,** 234-46.

OSHLACK, A., ROBINSON, M. D. & YOUNG, M. D. 2010. From RNA-seq reads to differential expression results. *Genome Biol,* 11**,** 220.

OUYANG, Z., ZHOU, Q. & WONG, W. H. 2009. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc Natl Acad Sci U S A,* 106**,** 21521-6.

PARSONS, J. L., DIANOVA, II, ALLINSON, S. L. & DIANOV, G. L. 2005. Poly(ADP-ribose) polymerase-1 protects excessive DNA strand breaks from deterioration during repair in human cell extracts. *FEBS J,* 272**,** 2012-21.

PASCHOAL, A. R., MARACAJA-COUTINHO, V., SETUBAL, J. C., SIMOES, Z. L., VERJOVSKI-ALMEIDA, S. & DURHAM, A. M. 2012. Non-coding transcription characterization and annotation: a guide and web resource for non-coding RNA databases. *RNA Biol,* 9**,** 274-82.

PAULI, A., VALEN, E. & SCHIER, A. F. 2015. Identifying (non-)coding RNAs and small peptides: challenges and opportunities. *Bioessays,* 37**,** 103-12.

PETERS, J. 2014. The role of genomic imprinting in biology and disease: an expanding view. *Nat Rev Genet,* 15**,** 517-30.

PHILLIPS, J. E. & CORCES, V. G. 2009. CTCF: master weaver of the genome. *Cell,* 137**,** 1194-211.

PIEPER, A. A., VERMA, A., ZHANG, J. & SNYDER, S. H. 1999. Poly (ADP-ribose) polymerase, nitric oxide and cell death. *Trends Pharmacol Sci,* 20**,** 171-81.

PLASSCHAERT, R. N., VIGNEAU, S., TEMPERA, I., GUPTA, R., MAKSIMOSKA, J., EVERETT, L., DAVULURI, R., MAMORSTEIN, R., LIEBERMAN, P. M., SCHULTZ, D., HANNENHALLI, S. & BARTOLOMEI, M. S. 2014. CTCF binding site sequence differences are associated with unique regulatory and functional trends during embryonic stem cell differentiation. *Nucleic Acids Res,* 42**,** 774-89.

PRUITT, K. D., BROWN, G. R., HIATT, S. M., THIBAUD-NISSEN, F., ASTASHYN, A., ERMOLAEVA, O., FARRELL, C. M., HART, J., LANDRUM, M. J., MCGARVEY, K. M., MURPHY, M. R., O'LEARY, N. A., PUJAR, S., RAJPUT, B., RANGWALA, S. H., RIDDICK, L. D., SHKEDA, A., SUN, H., TAMEZ, P., TULLY, R. E., WALLIN, C., WEBB, D., WEBER, J., WU, W., DICUCCIO, M., KITTS, P., MAGLOTT, D. R., MURPHY, T. D. & OSTELL, J. M. 2014. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res,* 42**,** D756-63.

PRUITT, K. D., TATUSOVA, T. & MAGLOTT, D. R. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res,* 33**,** D501-4.

PUGACHEVA, E. M., TIWARI, V. K., ABDULLAEV, Z., VOSTROV, A. A., FLANAGAN, P. T., QUITSCHKE, W. W., LOUKINOV, D. I., OHLSSON, R. & LOBANENKOV, V. V. 2005. Familial cases of point mutations in the XIST promoter reveal a correlation between CTCF binding and pre-emptive choices of X chromosome inactivation. *Hum Mol Genet,* 14**,** 953-65.

QI, H., LIU, M., EMERY, D. W. & STAMATOYANNOPOULOS, G. 2015. Functional validation of a constitutive autonomous silencer element. *PLoS One,* 10**,** e0124588.

QUAIL, M. A., SMITH, M., COUPLAND, P., OTTO, T. D., HARRIS, S. R., CONNOR, T. R., BERTONI, A., SWERDLOW, H. P. & GU, Y. 2012. A tale of three next generation sequencing platforms:

comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics,* 13**,** 341.

RAMOS-VARA, J. A. 2005. Technical aspects of immunohistochemistry. *Vet Pathol,* 42**,** 405-26.

RASKO, J. E., KLENOVA, E. M., LEON, J., FILIPPOVA, G. N., LOUKINOV, D. I., VATOLIN, S., ROBINSON, A. F., HU, Y. J., ULMER, J., WARD, M. D., PUGACHEVA, E. M., NEIMAN, P. E., MORSE, H. C., 3RD, COLLINS, S. J. & LOBANENKOV, V. V. 2001. Cell growth inhibition by the multifunctional multivalent zinc-finger factor CTCF. *Cancer Res,* 61**,** 6002-7.

REIK, W. & MURRELL, A. 2000. Genomic imprinting. Silence across the border. *Nature,* 405**,** 408-9.

REN, B., CAM, H., TAKAHASHI, Y., VOLKERT, T., TERRAGNI, J., YOUNG, R. A. & DYNLACHT, B. D. 2002. E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes Dev,* 16**,** 245-56.

RENAUD, S., LOUKINOV, D., ABDULLAEV, Z., GUILLERET, I., BOSMAN, F. T., LOBANENKOV, V. & BENHATTAR, J. 2007. Dual role of DNA methylation inside and outside of CTCF-binding regions in the transcriptional regulation of the telomerase hTERT gene. *Nucleic Acids Res,* 35**,** 1245-56.

RENAUD, S., LOUKINOV, D., BOSMAN, F. T., LOBANENKOV, V. & BENHATTAR, J. 2005. CTCF binds the proximal exonic region of hTERT and inhibits its transcription. *Nucleic Acids Res,* 33**,** 6850-60.

ROBERT, I., KARICHEVA, O., REINA SAN MARTIN, B., SCHREIBER, V. & DANTZER, F. 2013. Functional aspects of PARylation in induced and programmed DNA repair processes: preserving genome integrity and modulating physiological events. *Mol Aspects Med,* 34**,** 1138-52.

RUBIN, L. L. & DE SAUVAGE, F. J. 2006. Targeting the Hedgehog pathway in cancer. *Nat Rev Drug Discov,* 5**,** 1026-33.

RUFFALO, M., LAFRAMBOISE, T. & KOYUTURK, M. 2011. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics,* 27**,** 2790-6.

SAFRAN, M., DALAH, I., ALEXANDER, J., ROSEN, N., INY STEIN, T., SHMOISH, M., NATIV, N., BAHIR, I., DONIGER, T., KRUG, H., SIROTA-MADI, A., OLENDER, T., GOLAN, Y., STELZER, G., HAREL, A. & LANCET, D. 2010. GeneCards Version 3: the human gene integrator. *Database (Oxford),* 2010**,** baq020.

SAINI, S., JAGADISH, N., GUPTA, A., BHATNAGAR, A. & SURI, A. 2013. A novel cancer testis antigen, A-kinase anchor protein 4 (AKAP4) is a potential biomarker for breast cancer. *PLoS One,* 8**,** e57095.

SAITO, K., KAGAWA, W., SUZUKI, T., SUZUKI, H., YOKOYAMA, S., SAITOH, H., TASHIRO, S., DOHMAE, N. & KURUMIZAKA, H. 2010. The putative nuclear localization signal of the human RAD52 protein is a potential sumoylation site. *J Biochem,* 147**,** 833-42.

SANYAL, A., LAJOIE, B. R., JAIN, G. & DEKKER, J. 2012. The long-range interaction landscape of gene promoters. *Nature,* 489**,** 109-13.

SARANGI, P., STEINACHER, R., ALTMANNOVA, V., FU, Q., PAULL, T. T., KREJCI, L., WHITBY, M. C. & ZHAO, X. 2015. Sumoylation influences DNA break repair partly by increasing the solubility of a conserved end resection protein. *PLoS Genet,* 11**,** e1004899.

SCANGOS, G. & RUDDLE, F. H. 1981. Mechanisms and applications of DNA-mediated gene transfer in mammalian cells - a review. *Gene,* 14**,** 1-10.

SCANGOS, G. A., HUTTNER, K. M., JURICEK, D. K. & RUDDLE, F. H. 1981. Deoxyribonucleic acid-mediated gene transfer in mammalian cells: molecular analysis of unstable transformants and their progression to stability. *Mol Cell Biol,* 1**,** 111-20.

SCHERER, W. F., SYVERTON, J. T. & GEY, G. O. 1953. STUDIES ON THE PROPAGATION IN VITRO OF POLIOMYELITIS VIRUSES : IV. VIRAL MULTIPLICATION IN A STABLE STRAIN OF HUMAN MALIGNANT EPITHELIAL CELLS (STRAIN HELA) DERIVED FROM AN EPIDERMOID CARCINOMA OF THE CERVIX. *The Journal of Experimental Medicine,* 97**,** 695-710.

SCHOENHERR, C. J., LEVORSE, J. M. & TILGHMAN, S. M. 2003. CTCF maintains differential methylation at the Igf2/H19 locus. *Nat Genet,* 33**,** 66-9.

SCHWARTZ, Y. B., LINDER-BASSO, D., KHARCHENKO, P. V., TOLSTORUKOV, M. Y., KIM, M., LI, H. B., GORCHAKOV, A. A., MINODA, A., SHANOWER, G., ALEKSEYENKO, A. A., RIDDLE, N. C., JUNG, Y. L., GU, T., PLACHETKA, A., ELGIN, S. C., KURODA, M. I., PARK, P. J., SAVITSKY, M., KARPEN, G. H.

& PIRROTTA, V. 2012. Nature and function of insulator protein binding sites in the Drosophila genome. *Genome Res,* 22**,** 2188-98.

SHALL, S. & DE MURCIA, G. 2000. Poly(ADP-ribose) polymerase-1: what have we learned from the deficient mouse model? *Mutat Res,* 460**,** 1-15.

SHANNON, P., MARKIEL, A., OZIER, O., BALIGA, N. S., WANG, J. T., RAMAGE, D., AMIN, N., SCHWIKOWSKI, B. & IDEKER, T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res,* 13**,** 2498-504.

SHARMA, A., SINGH, K. & ALMASAN, A. 2012. Histone H2AX Phosphorylation: A Marker for DNA Damage. *In:* BJERGBÆK, L. (ed.) *DNA Repair Protocols.* Humana Press.

SHENDURE, J. & JI, H. 2008. Next-generation DNA sequencing. *Nat Biotechnol,* 26**,** 1135-45.

SINCLAIR, W. K. 1965. Hydroxyurea: differential lethal effects on cultured mammalian cells during the cell cycle. *Science,* 150**,** 1729-31.

SINGH, P., LEE, D. H. & SZABO, P. E. 2012. More than insulator: multiple roles of CTCF at the H19-Igf2 imprinted domain. *Front Genet,* 3**,** 214.

SLADE, D., DUNSTAN, M. S., BARKAUSKAITE, E., WESTON, R., LAFITE, P., DIXON, N., AHEL, M., LEYS, D. & AHEL, I. 2011. The structure and catalytic mechanism of a poly(ADP-ribose) glycohydrolase. *Nature,* 477**,** 616-20.

SOLOMON, M. J., LARSEN, P. L. & VARSHAVSKY, A. 1988. Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell,* 53**,** 937-47.

SOTO-REYES, E. & RECILLAS-TARGA, F. 2010. Epigenetic regulation of the human p53 gene promoter by the CTCF transcription factor in transformed cell lines. *Oncogene,* 29**,** 2217-27.

SPENCER, R. J., DEL ROSARIO, B. C., PINTER, S. F., LESSING, D., SADREYEV, R. I. & LEE, J. T. 2011. A boundary element between Tsix and Xist binds the chromatin insulator Ctcf and contributes to initiation of X-chromosome inactivation. *Genetics,* 189**,** 441-54.

SPITZ, F. & FURLONG, E. E. 2012. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet,* 13**,** 613-26.

SPLINTER, E., HEATH, H., KOOREN, J., PALSTRA, R.-J., KLOUS, P., GROSVELD, F., GALJART, N. & DE LAAT, W. 2006. CTCF mediates long-range chromatin looping and local histone modification in the β-globin locus. *Genes & Development,* 20**,** 2349-2354.

SUN, L., LI, H., CHEN, J., DEHENNAUT, V., ZHAO, Y., YANG, Y., IWASAKI, Y., KAHN-PERLES, B., LEPRINCE, D., CHEN, Q., SHEN, A. & XU, Y. 2013. A SUMOylation-dependent pathway regulates SIRT1 transcription and lung cancer metastasis. *J Natl Cancer Inst,* 105**,** 887-98.

SUN, L. P., SEEMANN, J., GOLDSTEIN, J. L. & BROWN, M. S. 2007. Sterol-regulated transport of SREBPs from endoplasmic reticulum to Golgi: Insig renders sorting signal in Scap inaccessible to COPII proteins. *Proc Natl Acad Sci U S A,* 104**,** 6519-26.

SUPEK, F., BOSNJAK, M., SKUNCA, N. & SMUC, T. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One,* 6**,** e21800.

SZABO, P. E., TANG, S. H., SILVA, F. J., TSARK, W. M. & MANN, J. R. 2004. Role of CTCF binding sites in the Igf2/H19 imprinting control region. *Mol Cell Biol,* 24**,** 4791-800.

TAN, E. S., KRUKENBERG, K. A. & MITCHISON, T. J. 2012. Large-scale preparation and characterization of poly(ADP-ribose) and defined length polymers. *Anal Biochem,* 428**,** 126-36.

TEIF, V. B., BESHNOVA, D. A., VAINSHTEIN, Y., MARTH, C., MALLM, J. P., HOFER, T. & RIPPE, K. 2014. Nucleosome repositioning links DNA (de)methylation and differential CTCF binding during stem cell development. *Genome Res,* 24**,** 1285-95.

TORRANO, V., NAVASCUES, J., DOCQUIER, F., ZHANG, R., BURKE, L. J., CHERNUKHIN, I., FARRAR, D., LEON, J., BERCIANO, M. T., RENKAWITZ, R., KLENOVA, E., LAFARGA, M. & DELGADO, M. D. 2006. Targeting of CTCF to the nucleolus inhibits nucleolar transcription through a poly(ADP-ribosyl)ation-dependent mechanism. *J Cell Sci,* 119**,** 1746-59.

TOWBIN, H., STAEHELIN, T. & GORDON, J. 1979. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proc Natl Acad Sci U S A,* 76**,** 4350-4.

TREANGEN, T. J. & SALZBERG, S. L. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet,* 13**,** 36-46.

UNDERHILL, C., TOULMONDE, M. & BONNEFOI, H. 2011. A review of PARP inhibitors: from bench to bedside. *Ann Oncol,* 22**,** 268-79.

VALASEK, M. A. & REPA, J. J. 2005. The power of real-time PCR. *Adv Physiol Educ,* 29**,** 151-9.

VAN BORTLE, K., NICHOLS, M. H., LI, L., ONG, C. T., TAKENAKA, N., QIN, Z. S. & CORCES, V. G. 2014. Insulator function and topological domain border strength scale with architectural protein occupancy. *Genome Biol,* 15**,** R82.

VAN DE NOBELEN, S., ROSA-GARRIDO, M., LEERS, J., HEATH, H., SOOCHIT, W., JOOSEN, L., JONKERS, I., DEMMERS, J., VAN DER REIJDEN, M., TORRANO, V., GROSVELD, F., DELGADO, M. D., RENKAWITZ, R., GALJART, N. & SLEUTELS, F. 2010. CTCF regulates the local epigenetic state of ribosomal DNA repeats. *Epigenetics Chromatin,* 3**,** 19.

VAN STEENSEL, B., SMOGORZEWSKA, A. & DE LANGE, T. 1998. TRF2 protects human telomeres from end-to-end fusions. *Cell,* 92**,** 401-13.

VASQUEZ, R. J., HOWELL, B., YVON, A. M., WADSWORTH, P. & CASSIMERIS, L. 1997. Nanomolar concentrations of nocodazole alter microtubule dynamic instability in vivo and in vitro. *Mol Biol Cell,* 8**,** 973-85.

VILLANI, P., FRESEGNA, A. M., RANALDI, R., ELEUTERI, P., PARIS, L., PACCHIEROTTI, F. & CORDELLI, E. 2013. X-ray induced DNA damage and repair in germ cells of PARP1(-/-) male mice. *Int J Mol Sci,* 14**,** 18078-92.

VINCZE, T., POSFAI, J. & ROBERTS, R. J. 2003. NEBcutter: A program to cleave DNA with restriction enzymes. *Nucleic Acids Res,* 31**,** 3688-91.

VOELKERDING, K. V., DAMES, S. A. & DURTSCHI, J. D. 2009. Next-generation sequencing: from basic research to diagnostics. *Clin Chem,* 55**,** 641-58.

VON HOFF, D. D., LORUSSO, P. M., RUDIN, C. M., REDDY, J. C., YAUCH, R. L., TIBES, R., WEISS, G. J., BORAD, M. J., HANN, C. L., BRAHMER, J. R., MACKEY, H. M., LUM, B. L., DARBONNE, W. C., MARSTERS, J. C., JR., DE SAUVAGE, F. J. & LOW, J. A. 2009. Inhibition of the hedgehog pathway in advanced basal-cell carcinoma. *N Engl J Med,* 361**,** 1164-72.

VOSTROV, A. A. & QUITSCHKE, W. W. 1997. The zinc finger protein CTCF binds to the APBbeta domain of the amyloid beta-protein precursor promoter. Evidence for a role in transcriptional activation. *J Biol Chem,* 272**,** 33353-9.

WAHLBERG, E., KARLBERG, T., KOUZNETSOVA, E., MARKOVA, N., MACCHIARULO, A., THORSELL, A. G., POL, E., FROSTELL, A., EKBLAD, T., ONCU, D., KULL, B., ROBERTSON, G. M., PELLICCIARI, R., SCHULER, H. & WEIGELT, J. 2012. Family-wide chemical profiling and structural analysis of PARP and tankyrase inhibitors. *Nat Biotechnol,* 30**,** 283-8.

WANG, H., MAURANO, M. T., QU, H., VARLEY, K. E., GERTZ, J., PAULI, F., LEE, K., CANFIELD, T., WEAVER, M., SANDSTROM, R., THURMAN, R. E., KAUL, R., MYERS, R. M. & STAMATOYANNOPOULOS, J. A. 2012a. Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res,* 22**,** 1680-8.

WANG, J., WANG, Y. & LU, L. 2012b. De-SUMOylation of CCCTC binding factor (CTCF) in hypoxic stress-induced human corneal epithelial cells. *J Biol Chem,* 287**,** 12469-79.

WANG, Y. T., CHUANG, J. Y., SHEN, M. R., YANG, W. B., CHANG, W. C. & HUNG, J. J. 2008. Sumoylation of specificity protein 1 augments its degradation by changing the localization and increasing the specificity protein 1 proteolytic process. *J Mol Biol,* 380**,** 869-85.

WANG, Z., GERSTEIN, M. & SNYDER, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet,* 10**,** 57-63.

WEI, I. H., SHI, Y., JIANG, H., KUMAR-SINHA, C. & CHINNAIYAN, A. M. 2014. RNA-Seq accurately identifies cancer biomarker signatures to distinguish tissue of origin. *Neoplasia,* 16**,** 918-27.

WEI, Q., LI, L. D. & CHEN, D. D. 2007. *DNA repair, genetic instability, and cancer,* [Hackensack] N.J. ; London, World Scientific.

WESIERSKA-GADEK, J., RANFTLER, C. & SCHMID, G. 2005. Physiological ageing: role of p53 and PARP-1 tumor suppressors in the regulation of terminal senescence. *J Physiol Pharmacol,* 56 Suppl 2**,** 77-88.

WETH, O., WETH, C., BARTKUHN, M., LEERS, J., UHLE, F. & RENKAWITZ, R. 2010. Modular insulators: genome wide search for composite CTCF/thyroid hormone receptor binding sites. *PLoS One,* 5**,** e10119.

XIE, C., YUAN, J., LI, H., LI, M., ZHAO, G., BU, D., ZHU, W., WU, W., CHEN, R. & ZHAO, Y. 2014. NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic Acids Res,* 42**,** D98-103.

XIE, X., MIKKELSEN, T. S., GNIRKE, A., LINDBLAD-TOH, K., KELLIS, M. & LANDER, E. S. 2007. Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci U S A,* 104**,** 7145-50.

XU, M., ZHAO, G. N., LV, X., LIU, G., WANG, L. Y., HAO, D. L., WANG, J., LIU, D. P. & LIANG, C. C. 2014. CTCF controls HOXA cluster silencing and mediates PRC2-repressive higher-order chromatin structure in NT2/D1 cells. *Mol Cell Biol,* 34**,** 3867-79.

YAMAMOTO, H., OKUMURA, K., TOSHIMA, S., MUKAISHO, K., SUGIHARA, H., HATTORI, T., KATO, M. & ASANO, S. 2009. FXYD3 protein involved in tumor cell proliferation is overproduced in human breast cancer tissues. *Biol Pharm Bull,* 32**,** 1148-54.

YANG, Y., HU, J. F., ULANER, G. A., LI, T., YAO, X., VU, T. H. & HOFFMAN, A. R. 2003. Epigenetic regulation of Igf2/H19 imprinting at CTCF insulator binding sites. *J Cell Biochem,* 90**,** 1038-55.

YARBRO, J. W. 1992. Mechanism of action of hydroxyurea. *Semin Oncol,* 19**,** 1-10.

YEH, A., WEI, M., GOLUB, S. B., YAMASHIRO, D. J., MURTY, V. V. & TYCKO, B. 2002. Chromosome arm 16q in Wilms tumors: unbalanced chromosomal translocations, loss of heterozygosity, and assessment of the CTCF gene. *Genes Chromosomes Cancer,* 35**,** 156-63.

YU, W., GINJALA, V., PANT, V., CHERNUKHIN, I., WHITEHEAD, J., DOCQUIER, F., FARRAR, D., TAVOOSIDANA, G., MUKHOPADHYAY, R., KANDURI, C., OSHIMURA, M., FEINBERG, A. P., LOBANENKOV, V., KLENOVA, E. & OHLSSON, R. 2004a. Poly(ADP-ribosyl)ation regulates CTCF-dependent chromatin insulation. *Nat Genet,* 36**,** 1105-10.

YU, W., GINJALA, V., PANT, V., CHERNUKHIN, I., WHITEHEAD, J., DOCQUIER, F., FARRAR, D., TAVOOSIDANA, G., MUKHOPADHYAY, R., KANDURI, C., OSHIMURA, M., FEINBERG, A. P., LOBANENKOV, V., KLENOVA, E. & OHLSSON, R. 2004b. Poly(ADP-ribosyl)ation regulates CTCF-dependent chromatin insulation. *Nat Genet,* 36**,** 1105-1110.

YUSUFZAI, T. M., TAGAMI, H., NAKATANI, Y. & FELSENFELD, G. 2004. CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. *Mol Cell,* 13**,** 291-8.

ZAMPIERI, M., GUASTAFIERRO, T., CALABRESE, R., CICCARONE, F., BACALINI, M. G., REALE, A., PERILLI, M., PASSANANTI, C. & CAIAFA, P. 2012. ADP-ribose polymers localized on Ctcf-Parp1-Dnmt1 complex prevent methylation of Ctcf target sites. *Biochem J,* 441**,** 645-52.

ZHANG, W. J. & WU, J. Y. 1998. Sip1, a novel RS domain-containing protein essential for pre-mRNA splicing. *Mol Cell Biol,* 18**,** 676-84.

ZHOU, B. B. & ELLEDGE, S. J. 2000. The DNA damage response: putting checkpoints in perspective. *Nature,* 408**,** 433-9.

ZHU, Z. L., ZHAO, Z. R., ZHANG, Y., YANG, Y. H., WANG, Z. M., CUI, D. S., WANG, M. W., KLEEFF, J., KAYED, H., YAN, B. Y. & SUN, X. F. 2010. Expression and significance of FXYD-3 protein in gastric adenocarcinoma. *Dis Markers,* 28**,** 63-9.

ZIEBARTH, J. D., BHATTACHARYA, A. & CUI, Y. 2013. CTCFBSDB 2.0: a database for CTCF-binding sites and genome organization. *Nucleic Acids Res,* 41**,** D188-94.

ZIELKE, K., FULL, F., TEUFERT, N., SCHMIDT, M., MULLER-FLECKENSTEIN, I., ALBERTER, B. & ENSSER, A. 2012. The insulator protein CTCF binding sites in the orf73/LANA promoter region of herpesvirus saimiri are involved in conferring episomal stability in latently infected human T cells. *J Virol,* 86**,** 1862-73.