Investigation of the human folate gene *MTHFD1L*: polymorphisms and disease risk.

Stefano Minguzzi

PhD 2013

Investigation of the human folate gene *MTHFD1L*: polymorphisms and disease risk.

By

Stefano Minguzzi M.Sc.

Supervisor: Anne Parle-McDermott

School of Biotechnology, Dublin City University

A thesis submitted for the degree

of Doctor of Philosophy

September 2013

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of PhD is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: _____ (Candidate) ID No.: 59122714 Date: _____

Abbreviations

- **AA** Amino Acid
- **AdoMet** Adenosylmethionine
- **ACN** Acetonitrile
- **ACTB** Actin Beta
- **APS** Ammonium Persulfate
- **BCA** Bicinchoninic Acid
- **bp** Base Pair
- **BSA** Bovine Serum Albumin
- **cDNA** Complementary Deoxyribonucleic Acid
- **CH_2-THF** 5,10-methylene-THF
- **CH^+-THF** 5,10-methenyl-THF
- **CLP** Cleft Lip with or without cleft Palate
- **CPO** Cleft Palate Only
- **CVD** Cardiovascular Disease
- **DHFR** Dihydrofolate Reductase
- **DHFRL1** Dihydrofolate Reductase like-1
- **DIP** Deletion Insertion Polymorphism
- **DMEM** Dulbecco's Modified Eagle Medium
- **DNA** Deoxyribonucleic Acid
- **dNTPs** Deoxy Nucleotide Phosphate
- **FBS** Fetal Bovine Serum
- **GAPDH** Glyceraldehyde 3-phosphate dehydrogenase
- **GUS** Beta-Glucuronidase
- **GWAS** Genome-wide Association Study
- **Hcy** Homocysteine
- **HEK** Human Embryonic Kidney
- **HWE** Hardy-Weinberg Equilibrium
- **IPTG** Isopropyl β-D-1-thiogalactopyranoside

- **LB** Luria Bertani
- **LC-MS** Liquid Chromatography–Mass Spectrometry
- **LD** Linkage Disequilibrium
- **MAT** Methionine Adenosyltransferase
- **MCF7** Michigan Cancer Foundation – 7.
- **miRNA** MicroRNA
- **mRNA** messenger RNA
- **MTHFD1** Methylenetetrahydrofolate Dehydrogenase 1 (cytoplasmic isoform)
- **MTHFD1L** Methylenetetrahydrofolate Dehydrogenase 1-like (mitochondrial isoform)
- **MTHFR** Methylenetetrahydrofolate Reductase
- **MTR** Methionine synthase
- **ncRNA** non-coding RNA
- **NTD** Neural Tube Defects
- **ORF** Open Reading Frame
- **PBS** Phosphate-Buffered Saline
- **PCR** Polymerase Chain Reaction
- **PDH** Pyruvate Dehydrogenase
- **RNA** Ribonucleic Acid
- **rRNA** Ribosomal RNA
- **RT** Reverse Transcriptase
- **RT-PCR** Reverse Transcriptase Polymerase Chain Reaction
- **RT-qPCR** Reverse Transcriptase - Quantitative Polymerase Chain Reaction
- **SAM** S-Adenosylmethionine
- **SHMT1** Serine Hydroxymethyltransferase (cytoplasmic isoform)
- **SHMT2** Serine Hydroxymethyltransferase (mitochondrial isoform)
- **SNP** Singular Nucleotide Polymorphism
- **TBE** Tris Borate EDTA

- **TBT** TATA Binding Protein
- **TFA** Trifluoroacetic Acid
- **THF** Tetrahydrofolate
- **TYMS** Thymidylate Synthase
- **UPL** Universal Probe Library
- **UTR** Untranslated Region

Index

Acknowledgments

First of all, I would like to thank my supervisor Anne Parle-McDermott for picking up my CV back in 2008 to undertake the Leonardo internship that became this PhD one year later. Anne always supported me throughout my project and made me grow professionally. It has been a real pleasure to work in her lab where I always felt free to purse my own ideas.

I also thank the Irish Research Council for founding my PhD and for rewarding my project.

I would like to give a special thanks to my family. Thank you to Bianca for being on my side for all these years. Thanks to my parents Andrea and Daniela and my brother Alessandro for all of their love and support.

Finally, a big thank and a symbolic hug to all the people that walked along with me during this journey, all my lab mates all my collaborators and all the friends that I met in Ireland during these four years. It has been a sincere pleasure to live in this great country for four years and I will carry wonderful memories of this experience with me.

P.S. Thank you Alan for proofreading these acknowledgments.

Abstract

Investigation of the human folate gene *MTHFD1L*: polymorphisms and disease risk.

Stefano Minguzzi

The *MTHFD1L* gene encodes the mitochondrial monofunctional enzyme with proven 10-formyltetrahydrofolate synthetase activity. *MTHFD1L* expression is upregulated in human colon cancer and breast cancer, and high levels of this protein are correlated with growth rate of human cancer cell lines. Genetic association studies demonstrated that *MTHFD1L* polymorphisms are associated with coronary artery disease, Alzheimer's disease and neural tube defects (NTDs). Moreover, *MTHFD1L* knockout mice develop NTDs with 100% penetrance. This project involves the characterization of *MTHFD1L* with a particular focus on genetic polymorphisms and disease risk. *MTHFD1L* polymorphisms have been studied in relation to cleft disease and Melting Curve Analysis was developed for genotyping the Deletion Insertion Polymorphism rs3832406. MicroRNA-9 and 197 were proven to downregulate MTHFD1L at mRNA and protein level. The impact of polymorphism rs7646 on miR-197 binding suggests a mechanistic explanation of the previous association of this SNP with NTD risk. Sequence analysis was performed to discover and characterize *MTHFD1* and *MTHFD1L* similar sequences. *MTHFD1L* expression was proven to be affected by a 24 hour exposure to cycloleucine in lymphoblast. Finally, HEK293 stable cell lines with either MTHFD1L downregulated or overexpressed have been generated and the resulting proteomes have been assessed by Mass Spec.

CHAPTER 1

Introduction

1.1 Introduction

1.1.1 Overview

Nutritional genomics is a branch of science that studies the relationship between the human genome, nutrition and health. In particular, it studies how individual differences in genes influence the body's response to diet and nutrition and how these genetic variations (called polymorphisms) play a role as risk factors of disease. Folate is an essential nutrient necessary for many cellular functions and folate deficiency has been correlated with numerous diseases. Genes that are involved in the transport and metabolism of this vitamin are currently considered the main candidates for disease association.

This thesis consists of the characterization of the folate-related gene methylenetetrahydrofolate dehydrogenase 1-like (*MTHFD1L*) with a particular focus on genetic polymorphisms and disease risk. Our laboratory previously investigated the association between *MTHFD1L* polymorphisms and Neural Tube Defects (NTDs) in the Irish population (Parle-McDermott *et al*. 2009). Following up on this study, *MTHFD1L* genetic variations have been considered in relation to another birth defect, Cleft disease. The impact of these genetic variations on microRNA regulation has also been investigated. The results of these studies have been corroborated by proteomic analysis on samples with different MTHFD1L expression levels. The other aspects covered by this manuscript are the role of alternative splicing on MTHFD1L expression and the potential function of additional homologous sequences.

This introductive first chapter starts off by explaining the concept of nutritional genomics and illustrating the folate-mediated one-carbon metabolism along with the genes involved in it. It continues by presenting the literature on the MTHFD gene family and their disease association and concludes with the aims and objectives of this thesis.

1.1.2 Nutritional genomics

Nutritional genomics is a highly innovative and fast-growing science that studies how dietary components interact with genes and how these genes are related to metabolism. This discipline includes nutrigenomics, which studies how nutrients interact with the genome, proteome and metabolome; and nutrigenetics, which studies the interaction between diet and health as a consequence of genetic variations with implications to susceptible populations (Debusk *et al*. 2005, Subbiah 2007).

The completion of the Human Genome Project and the improvement of genetic tools to aid genome-wide association study (GWAS) have had a great impact on the development of these disciplines. A GWAS is where many thousands of genetic variations, called single-nucleotide polymorphisms (SNPs), are genotyped in a large number of individuals and tested for association with disease risk (Ott *et al.* 2011). Thanks to these extensive studies it is becoming evident that specific SNPs (and other polymorphisms) can affect the function of genes and alter specific metabolic pathways. Dietary factors might differentially interact with these SNPs to increase or decrease disease risk. An elegant example of this is a diet–SNP interaction involving the methylenetetrahydrofolate reductase (*MTHFR*) gene, one of the main genes involved in folate metabolism (see Figure 1.1). The common 677C>T SNP of this gene encodes for an amino acid substitution that generates a thermolabile and less active protein (Frost *et al*. 1995). This results in reduced capacity to use folate to remethylate homocysteine to methionine in individuals that are homozygote for the T allele. TT individuals, therefore, have elevated homocysteine levels, which are a risk factor for a number of common diseases including cardiovascular disease (CVD). In this example diet plays an important role in interacting with genotype, because the amount of folate (substrate of MTHFR) and riboflavin (a cofactor of MTHFR) modify the activity of MTHFR and subsequently increase or decrease the disease risk (Parle-McDermott *et al*. 2006a).

1.1.3 Folate

Folate and its synthetic form folic acid (also known as vitamin B_9), are essential nutrients in the human diet providing 1-carbon units for DNA synthesis and methylation reactions within the cell. Folate belongs to a group of water-soluble B vitamins and it was isolated for the first time in 1941 by Mitchell *et al*. from spinach leaves. The term folate comes from the Latin word *folium*, which means leaf, in fact leafy vegetables such as spinach, asparagus, turnip greens have high amount of folates. Other natural folate rich foods are yeast extract spread, liver, legumes, sunflower seeds and peanuts (*USDA National Nutrient Database for Standard Reference, Release 24.*)

Folate is especially important during periods of rapid cell division and it is required to produce healthy red blood cells and to prevent anemia (Iyer and Tomar 2009). Folates consist of a 2-amino-4hydroxy-pteridine (pterin) group conjugated by a methylene group to ρ-aminobenzoic acid, which in turn is linked to one or more glutamate units (Figure 1.2) (Damaraju *et al*. 2008). Even though the pteridine ring can be produced by mammalian cells, inability to couple it to ρ-aminobenzoic acid makes folate an essential nutrient in mammals (Birn 2006). This vitamin, which is naturally present in certain types of food, is hydrolyzed to monoglutamates by folate conjugase in the brush border of the small intestine prior to absorption into the portal vein (Fitzpatrick 2003, Iyer and Tomar 2009). Monoglutamate form of folate, including synthetic folic acid, which is the fully oxidized and most stable variant, can be easily transported across cell membranes and circulated in blood serum (Damaraju *et al*. 2008, Fox and Stover 2009). The absorption of monoglutamate is approximately twice as efficient as polyglutamate form and as a consequence, in the human body the bioavailability of the former is significantly greater than the latter (Iyer and Tomar 2009).

A modest lack of dietary folate leads to suboptimal folate levels that appear to play an important role in the pathogenesis of several disorders, including CVDs, neural tube defects (NTDs) (Chapter 4, Section 4.1.2) and other congenital defects like Cleft (Chapter 3, Section 3.1.2), thromboembolitic processes, Alzheimer's disease, osteoporosis,

increased risk of breast and colorectal cancer (Daly *et al*. 1995, Kim 2004). The large volume of evidence for a protective effect of periconceptional folic acid supplementation against NTDs led to mandatory folic acid fortification in several countries like the United States, Canada and Chile (Food and Drug Administration 1996, Health Canada 1997, Freire *et al*. 2000). In the European Union, where folic acid fortification is not mandatory, the Daily Recommended Intake for adults (DRI) is 400 μg/d, with this requirement increasing to 600 μg/d during pregnancy (FAO/WHO 2002, IOM 2004, McPartlin *et al*. 1993).

Apart from nutritional intake, polymorphisms or variations within specific folate genes are also known to contribute to human disease risk, particularly in the context of a poor nutritional diet. As in the example above, genes that are involved in the transport and metabolism of this vitamin are currently considered the main candidates for disease association (Subbiah 2007).

1.1.4 Folate-Mediated One-Carbon metabolism

Folate is involved in both the activation of single carbon units as well as in their oxidation and reduction. Figure 1.1 summarizes the main pathways of folate-mediated one-carbon (1C) metabolism that, in eukaryotes, is highly compartmentalized in the cytoplasm, mitochondria and the nucleus. Cytoplasmic pathway is responsible for the synthesis of purines and for converting homocysteine to methionine. The mitochondrial pathway is required for the synthesis of formylated methionyl-tRNA, the catabolism of choline, purines and histidine, and the *interconversion* of glycine and serine (Fox and Stover 2008). Although *de novo* thymidylate synthesis was thought to occur in cytoplasm, recent studies suggested that this pathway predominantly takes place in the nucleus (Anderson *et al.* 2012) and the mitochondria (Anderson *et al*. 2011, McEntee *et al.* 2011). The cytoplasmic and mitochondrial compartments are connected by transport of 1C donors such as formate, serine and glycine across the mitochondrial membrane (Momb *et al.* 2013). Since 1C metabolism is mainly unidirectional in flow (clockwise in Figure 1.1), mitochondria are also the primary source of one-carbon units, converting serine to formate (Tibbetts and

Appling 2010). In fact, it appears that under most conditions, the majority of 1C units for cytoplasmic reactions are derived from mitochondrial formate. Metabolic experiments carried out in mouse embryonic fibroblasts showed that more than 75% of 1C units that enter the cytoplasmic methyl cycle derive from mitochondria (Pike *et al.* 2010).

As explained above, folate is transported across cell membranes and in serum as monoglutamate form. Once folate is inside the cell folylpoly-gamma-glutamate synthetase (FPGS) catalyzes the addition of poly-γ-glutamate tails to the ρ-aminobenzoic acid (Shane 1995, Moran 1999). The number of added glutamate residues varies from one cell type to another, but in the majority of eukaryotic cells, the predominant forms are penta- and hexaglutamate (Tibbetts and Appling 2010). The polyglutamate tail prevents efflux of folate cofactors from the cell and intracellular organelles, and it increases their binding affinity for many folate-dependent enzymes (Schirch and Strong 1989).

The active form of folate is tetrahydrofolate (THF) which is the carrier of one-carbon units on the N5 and/or N10. To produce this active cofactor, folate is first reduced to 7,8-dihydrofolate and then to 5,6,7,8-tetrahydrofolate. The last step is catalyzed by dihydrofolate reductase (DHFR) in the presence of NADPH (reaction 13). This enzyme is present in both the cytoplasm and the nucleus and recent studies identified a homologous protein, DHFRL1 (dihydrofolate reductase like-1), which is localized to the mitochondria (Anderson *et al*. 2011, McEntee *et al*. 2011). In Figure 1.1 reaction number 1, 2, 3 and 4 are both in the cytoplasmic and mitochondrial compartments. In reaction 4, 1C unit is transfered to THF by serine hydroxymethyltransferase (SHMT) generating 5,10-methylene-THF (CH_2-THF) and glycine. In mammals, there are two SHMT isozyme that are encoded by distinct genes. *SHMT1* encodes a cytoplasmic/nuclear isozyme (SHMT1), while *SHMT2* encodes the mitochondrial (SHMT2) isoform and the cytoplasmic/nuclear (SHMT2α) isoform through alternative promoter use (Garrow *et al*. 1993, Anderson and Stover 2009, Anderson *et al.* 2012). In the cytoplasm CH_2-THF is required for homocysteine (Hcy) remethylation when converted to 5-methyl-THF by methylenetetrahydrofolate reductase (MTHFR) (reaction 6). 5-methyl-THF is the main circulating form of folate and it acts as methyl donor in the remethylation of Hcy to

methionine carried out by Methionine Synthase (reaction 7) (Thomas and Fenech 2008). Through the Methyl cycle (top left in Figure 1.1) methionine is converted to S-adenosylmethionine (AdoMet) which is the principal methyl donor, necessary for the methylation of DNA, RNA, histones, lipids and small molecules (MacFarlane *et al.* 2011).

Depending on the cell needs CH_2-THF may also be oxidized to 10-formyl-THF or used as cofactor for *de novo* thymidylate biosynthesis (TYMS) (reaction 10) which occurs in the nucleus (Anderson *et al.* 2012) and mitochondria (Anderson *et al*. 2011, McEntee *et al.* 2011). CH_2-THF is oxidized to 10-formyl-THF in two reactions catalyzed by NADP-dependent CH_2-THF dehydrogenase (reaction 3) and 5,10-methenyl-THF (CH^+-THF) cyclohydrolase (reaction 2). At this stage, cytoplasmic 10-formyl-THF may be used for *de novo* purine biosynthesis, a 10-reaction pathway particularly important during rapid cell division (Fox and Stover 2008). On the other hand, mitochondrial 10-formyl-THF may either be converted to CO_2 or serve as formyl donor to synthesize formyl-methionyl-tRNA (reaction 12), necessary as initiator for mitochondrial protein synthesis (Christensen and Mackenzie 2008). There is one other important fate of 10-formyl-THF in both mitochondria and cytoplasm: the synthesis of formate and THF catalyzed by 10-formyl-THF synthetase (reaction 1). While cytoplasmic reactions 1,2 and 3 are performed by the trifunctional enzyme 10-formyl-THF synthase (MTHFD1), its mitochondrial homolog MTHFD1-like (MTHFD1L) catalyzes only reaction 1 in mammals. The other two reactions are carried out by MTHFD2 and MTHFD2L (Belanger *et al*. 1989, Di Pietro *et al*. 2002, Bolusani *et al*. 2011).

While the synthesis of cytosine and the *de novo* purines occurs in the cytoplasm, the folate-dependent *de novo* synthesis of thymidylate occurs in mitochondria (Anderson *et al*. 2011, McEntee *et al*. 2011) and in the nucleus (Chen *et al*. 2010 and Anderson et *al.* 2012). Recent findings better elucidated the role of 1-C metabolism inside the nucleus. It has been shown that TYMS, DHFR and SHMT1 localize to the nucleus only during S and G_2/M phases enabling nuclear *de novo* thymidylate synthesis during DNA replication and repair (Anderson *et al.* 2012). These enzymes, which normally localize to cytoplasm (TYMS also to mitochondria), are transported to the nucleus upon post-translational modification with

the small ubiquitin-like modifier (SUMO) (Anderson *et al.* 2007 and Anderson *et al.* 2009). TYMS, DHFR and SHMT1 form a thymidylate biosynthesis complex associated with the nuclear lamina and the DNA replication machinery. SHMT1 proved to have an important scaffold function for complex formation (Anderson *et al.* 2012). MTHFD1 protein, which was believed to be present only in cytoplasm, was also identified as component of the nuclear thymidylate biosynthesis complex (Anderson *et al.* 2012).

1.1.5 MTHFD genes

As explained before, the reversible conversion of 10-formyl-THF and ADP to formate, THF and ATP is carried out by the trifunctional enzyme MTHFD1, which requires monovalent cations (NH_4^+, Mg_2^+, K^+ or Rb^+) to achieve maximal activity (Appling 1991, Fox and Stover 2008). This 935-aa protein exists in mammals as a homodimer of about 100 kDa subunits. Each subunit consists of a ~30 kDa N-terminal domain containing CH_2-THF dehydrogenase and 5,10-methenyl-THF cyclohydrolase activities, and a ~70 kDa C-terminal domain containing 10-formyl-THF synthetase activity (Figure 1.3) (Tibbetts and Appling 2010). Leaphart *et al*. (2002) demonstrated that MTHFD1 is inhibited by THF and purines, while it is transcriptionally upregulated when increased DNA synthesis is required (Christensen and MacKenzie 2006). In mammals MTHFD1 is ubiquitously expressed in embryos and adults with the highest expression in kidney and liver (Di Pietro *et al.* 2004, MacFarlane *et al.* 2009).

The enzyme is encoded by the *MTHFD1* gene that spans 72 kilobase pairs on chromosome 14 at 14q24 and it is spliced into a 3,466-base transcript sequence composed of 28 exons. *MTHFD1* is an essential gene as its homozygous knock-out is lethal in mouse embryos, probably due to inadequate *de novo* purine synthesis. Heterozygous disruption of *MTHFD1* results in viable mice with lower hepatic AdoMet levels which is consistent with formate serving as a 1C source for cellular methylation reactions (MacFarlane *et al.* 2009). On the other hand MTHFD1$^{+/-}$ mice exhibited decreased levels of uracil misincorporation into liver nuclear DNA, indicating enhanced *de novo* thymidylate synthesis and suggesting a competition between SHMT1 and MTHFD1 for a limited pool of THF (MacFarlane *et al.* 2009). Another study also showed that azoxymethane-induced colon cancer in *MTHFD1*$^{+/-}$

mice increased tumour incidence 2.5-fold (MacFarlane *et al.* 2010). In a recent study, Christensen *et al.* (2013) generated a mouse model in which the MTHFD1 synthetase activity was inactivated without affecting protein expression or the other activities of this enzyme. Mthfd1S$^{-/-}$ mouse embryos died shortly after 10.5 days gestation, and showed a delayed or abnormal development. On the other hand, in Mthfd1S$^{+/-}$ heterozygote mice *de novo* purine synthesis was impaired and the proportion of 10-formylTHF in the plasma and liver was reduced. The authors, therefore, suggested that synthetase deficiency may lead to pregnancy complications through decreased purine synthesis and reduced cellular proliferation (Christensen *et al.* 2013). Indeed, MTHFD1 has been widely investigated in relation to birth malformations. A common variant of MTHFD1, 1958G→A, results in an arginine to glutamine substitution (R653Q) in the active site of the enzyme. MTHFD1 1958G→A has been associated with increased risk of NTDs (Brody *et al*. 2002, Parle-McDermott *et al*. 2006b, De Marco *et al*. 2006), cleft lip and palate (Mills *et al*. 2008), congenital heart defects (Christensen *et al.* 2009), placental abruption (Parle-McDermott *et al*. 2005a), unexplained second-trimester pregnancy loss (Parle-McDermott *et al*. 2005b) and intrauterine growth restriction (Furness *et al.* 2008). In 2012, Beaudin *et al*. generated another Mthfd1 knock-out mouse model that resulted in impaired fetal growth. *MTHFD1*$^{+/-}$ mice exhibit the same range of birth malformations mentioned above.

A mitochondrial isozyme of MTHFD1 was first identified in *Saccharomyces cerevisiae* by Paukert and coworkers in 1977 and subsequently cloned and characterized by Shannon and Rabinowitz (1986 and 1988). In yeast both cytoplasmic and mitochondrial isozymes are trifunctional showing dehydrogenase, cyclohydrolase and synthetase activities (Christensen and MacKenzie 2008). Differently from yeast, mammalian mitochondria do not exhibit a unique trifunctional enzyme. Mammalian mitochondria contain a bifunctional NAD-dependent CH^2-THF dehydrogenase/CH^+ cyclohydrolase encoded by *MTHFD2* gene (Christensen and MacKenzie 2006). This essential gene is located on chromosome 2 at 2p31.1 and it is believed to have evolved from trifunctional *MTHFD1* through gene duplication. It subsequently lost the synthetase domain and changed the redox cofactor specificity from $NADP^+$ to NAD^+ (Patel *et al.* 2002). Homozygous knockout of Mthfd2 is embryonic lethal in mice. Embryos develop to about E15.5 and display no evident

developmental abnormalities, but they are smaller and paler than WT and heterozygous littermates (Di Pietro *et al.* 2002). In mice, *MTHFD2* expression has been detected in all the embryonic tissues as well as in cancer and transformed cell lines, but it was not found in any adult tissue (Christensen *et al.* 2008). This lack of CH^2-THF dehydrogenase and CH^+ cyclohydrolase activities has left a gap in the mitochondrial 1C metabolism of adult mammals until *MTHFD2L* gene was recently discovered by Bolusani and colleagues (2011). They demonstrated that MTHFD2L is ubiquitously expressed in adult tissues of humans and rodents and it carries out the same 2-step reaction of MTHFD2 (although the cyclohydrolase step still has to be confirmed). In humans *MTHFD2L* is localized on chromosome 4 at 4q13.3 and it contains 9 exons encoding a protein of 347 aa. Like its homolog MTHFD2, MTHFD2L possesses only a cyclohydrolase/dehydrogenase domain that is similar to N-terminal domain of MTHFD1 (Figure 1.3).

The monofunctional enzyme that completes the mitochondrial 1C pathway and possesses 10-formyl-THF synthetase activity is encoded by the *MTHFD1L* gene. MTHFD1L protein retains the size and the two structural domains of its cytoplasmic homolog but due to non-conservative mutation of critical residues, both the dehydrogenase and cyclohydrolase activities are not functional (Figure 1.3) (Christensen and MacKenzie 2008). In fact MTHFD1L shares about 61% amino acid identity with MTHFD1, but the sequence identity in the N-terminal domain (33%) is substantially lower than in the C-terminal synthetase domain (77%). In particular, four critical mutated amino acids (aa) (K93, R206, G211 and R283) are believed to be responsible for the lack of activity of the N-terminal domain (Christensen *et al.* 2005). Differently from its cytoplasmic homolog, MTHFD1L also contains 62 N-terminal residues with a predicted mitochondrial targeting sequence and a cleavage site between residues 31 and 32. The function of the next 31 aa including an unusual sequence of 9 consecutive glycine residues, is still unknown. MTHFD1L also lacks 12 aa near the junction between the two domains (Prasannan *et al*. 2003).

MTHFD1L acts as peripheral membrane protein, associated with the matrix side of the mitochondrial inner membrane (Prasannan *et al*. 2008). MTHFD1L is more efficient with a longer THF polyglutamate chain length as substrate, similar to other enzymes of the same

family. From monoglutamate to pentaglutamate the K_m for THF decreases at least 100-fold and the K_m for ATP and formate reacts in a similar way (Tibbets and Appling 2010). Like MTHFD1 the mitochondrial isozyme is expressed ubiquitously in mammalian cells, however the expression pattern is different; high levels of MTHFD1L were detected in the thymus, spleen, brain and placenta while in liver and kidney were relatively low (Prasannan *et al.* 2003, Walkup *et al.* 2005, Prasannan *et al.* 2009). Sugiura et al. (2004) also reported the highest levels of MTHFD1L expression in ovary, lung and thymus, whereas the expression levels were the lowest in white blood cells, muscle and lymphocytes. The *MTHFD1L* gene is also ubiquitously expressed throughout the embryo during all stages of mammalian embryogenesis, but with localized regions of higher expression along the neural tube, the brain, craniofacial structures, the tail bud and the limb buds (Pike *et al.* 2010).

1.1.6 MTHFD1L and disease

The relevance of *MTHFD1L* is increasing given its identification in genome wide association screens as being associated with several diseases. In 2007, a GWAS carried out by Samani et al. found an association between coronary artery disease risk and many intronic *MTHFD1L* polymorphisms in the European population. The lead SNP (rs6922269) is present in intron 9 and its risk allele (A) has a prevalence of approximately 25% in the European population, with the coronary artery disease risk increased by 23% per copy (95% CI, 15 to 33). Haplotype analysis showed that only the two haplotypes carrying the A allele tended to be more frequent in cases than in controls, confirming the increased odds ratio for coronary artery disease with the A allele in the single locus analysis (Samani *et al*. 2007). These findings were later confirmed by an independent GWAS performed in the German population (Bressler *et al*. 2010).

Another GWAS demonstrated that SNP rs11754661, in *MTHFD1L* intron 6, and other intronic SNPs are associated with Alzheimer's disease risk (Naj *et al*. 2010). This finding was confirmed by two independent studies in the in the Chinese Han population (Ren *et al.* 2011, Ma *et al.* 2012). As the 1C metabolism is important for homocysteine conversion to

methionine a potential mechanism that can explain Alzheimer's disease connection with MTHFD1L is hyperhomocysteinaemia. Elevated plasma homocysteine levels have been implicated in Alzheimer's disease (Van Dam *et al.* 2009, Morris 2003) and in other neurodegenerative disease (Herrmann *et al.* 2002). Decreased levels of MTHFD1L protein have been reported in the hippocampus in a mouse model of Alzheimer's disease using a proteomic approach (Martin *et al.* 2008). Homocysteic acid, derived from homocysteine and methionine, is elevated in these mice and treatment with anti-homocysteic acid antibodies reduced amyloid presence and inhibited cognitive decline in these animals (Hasegawa *et al.* 2010). B6-deficient diets lead to further increases in homocysteic acid in these mice.

Moreover, Sugiura *et al*. (2004) have shown that *MTHFD1L* and SHMT2 are upregulated in human colon adenocarcinoma. The expression level of *MTHFD1L* mRNA was increased to some degree (1.45-fold) in benign adenoma tissues but the increase was higher (2.38-fold) in colon adenocarcinomas. In addition to colon cancer, the increased expression of *MTHFD1L* was denoted in other tumours such as stomach (1.97-fold) or pancreatic cancer (1.44-fold) (Sugiura *et al*. 2004). The authors also demonstrated that the overexpression of MTHFD1L stimulated the cell growth and the colony formation of HEK293 cells. The authors suggested that mitochondrial folate genes could be connected to cancer through the c-myc oncogene, which is a crucial transcription factor in the control of cell proliferation, differentiation, and apoptosis (Sugiura *et al*. 2004). Both SHMT2 (Nikiforov *et al*. 2002) and MTHFD1L (Tavtigian *et al*. 1994, Sugiura *et al*. 2004) seemed to be responsive to the expression of c-myc oncogene in mammalian cells. A myc consensus sequence (CACGTG) is also present in the first intron of both SHMT2 and MTHFD1L genes.

Consistently with Sugiura *et al*. findings, it has been recently demonstrated that high expression levels of MTHFD1L, along with other mitochondrial folate proteins like SHMT2 and MTHFD2, were significantly correlated with the growth rates of human cancer cell lines (Jain *et al*. 2012, Tomita *et al.* 2012). Interestingly, the authors did not find such correlation with their cytosolic homologues SHMT1 and MTHFD1. It has been

shown that breast cancer patients with above-median expression levels of MTHFD1L, SHMT2 and MTHFD2, but not their cytosolic homologues, had higher mortality (Jain *et al*. 2012). The authors also found a strong association between growth rate of tumour cells and demand of glycine which is produced from serine by SHMT enzymes. Intriguingly, in contrast to the cancer cell lines, they found that rapidly proliferating nontransformed cells release rather than consume glycine (Jain *et al*. 2012). These data are relevant for cancer therapeutics as it may be possible to tackle aggressive cancer cells specifically by inhibiting glycine influx, synthesis, or consumption (Jain *et al*. 2012). Altogether, these findings match with the view that changes in the mitochondrial metabolism play an important role in cancer cell proliferation.

1.1.7 MTHFD1L and NTDs

In a recent publication, *Mthfd1l* knockout mice were analysed and the loss of the protein was found to be lethal to developing embryos, causing fetal growth restriction and aberrant neural tube phenotypes with 100% penetrance (Momb *et al.* 2013). All the *Mthfd1l* knockout embryos developed a clear NTD phenotype failing to close the neural tube (exencephaly or craniorachischisis) or displaying a wavy neural tube with a small, aberrantly formed head (Momb *et al.* 2013). Maternal supplementation with sodium formate significantly decreased the incidence of NTDs and partially rescued the growth defects in mouse embryos lacking *Mthfd1l.* Although there are other folate related mouse models that show NTDs, this *Mthfd1l* knockout mouse model is fully penetrant and does not require a folate-deficient diet to cause this phenotype (Momb *et al.* 2013).

Moreover, in humans MTHFD1L variants have been shown to be associated with risk of NTDs (Parle-McDermott *et al*. 2009). *MTHFD1L* gene spans 236 kilobase pairs on human Chr 6 at 6q25.1 and consists of 28 exons plus one alternative exon. To date four different *MTHFD1L* transcript variants have been identified in humans: transcript variant number 1 (3490 bases) and 3 (3174 bp) share the same coding region, but the latter has a shorter 5′UTR. Transcript variant 2 (3437 bp) has the same exact sequence of transcript 1 with the

only exception of a lacking codon at the end of exon 6. Instead transcript variant 4 (1072 bp) is a shorter isoform which lacks enzymatic activity (Prasannan *et al*. 2003) This short transcript is caused by a premature stop codon in the alternatively spliced exon 8a and just upstream of this exon there is a polypyrimidine tract which is a well known cis-acting factor required for splicing (Sawicka *et al*. 2008) (Figure 1.4). Within this polypyrimidine tract there is the Deletion/Insertion Polymorphism (DIP) rs3832406 composed of several ATT repeats: Allele 1 has 7 ATT repeats, allele 2 has 8 repeats and allele 3 has 9 repeats.

Our laboratory previously investigated the association between *MTHFD1L* polymorphisms and Neural Tube Defects (NTDs) in the Irish population (Parle-McDermott *et al*. 2009). Common genetic variations and low maternal folate levels are important risk factors for this kind of birth malformation. It has been reported that the Allele 1 of the DIP and numerous Single Nucleotide Polymorphisms (SNPs) of *MTHFD1L* are associated with the risk of NTDs and suggested that the DIP could affect the alternative splicing of the gene. In fact, it has been demonstrated that in Coriell lymphoblast cells the ratio between Long (variants 1, 2 and 3) and Short (variant 4) transcripts is higher in cells homozygous for allele 1 than in cells homozygous for allele 2 (Figure 1.5).

Figure 1.6 shows the Linkage Disequilibrium (LD) plot of pairwise values of D' for 119 *MTHFD1L* genetic markers. Brackets (a, b) mark two regions in which single markers were found to be associated with NTD risk. These regions represent two different signals of disease association and DIP rs3832406 is included in region ''a''. All the polymorphisms associated with NTD risk in region "a" were intronic, so the best candidate for driving the haplotype risk of this region is the DIP. As a consequence, it has been proposed that the DIP polymorphism is the direct disease causing variant within the associated LD region "a" by affecting alternative splicing of the gene. In this scenario, higher MTHFD1L protein levels would be associated with an increased NTD risk (Parle-McDermott *et al*. 2009). However, how the other polymorphisms that independently increase NTD-risk impact on MTHFD1L remain to be elucidated.

Moreover, unpublished results from our laboratory showed that *MTHFD1L* Long transcript is significantly upregulated in HEK293 cells cultured in the absence of folate (Figure 1.7). These findings suggest that a high level of *MTHFD1L* Long transcript might be an indicator of folate deficiency.

1.1.8 Summary

Nutritional genomics aims to identify genetic susceptibility to diseases related to diet factors. Folate, or vitamin B9, is an essential nutrient in our diet. Folate deficiency and polymorphisms within folate-related enzymes have been extensively associated with a number of common diseases and disorders, particularly, NTDs. The MTHFD gene family play an important role in folate-dependent 1C metabolism. MTHFD1L, in particular, is responsible for supplying the cytoplasm with 1C units in the form of formate. This key role puts MTHFD1L under the spotlight for its possible involvement in human disease. Indeed, several studies have reported its upregulation in cancer cells and its association with coronary artery disease, Alzheimer's disease and NTD risk. Despite these studies, the mechanism of MTHFD1L's association with disease is not yet known and many aspects of this essential gene remain to be elucidated. Taking the cue from previous studies, this project further investigates the role of MTHFD1L and its genetic variants in a bid to better understand its involvement in disease.

1.1.9 Aims and objectives

This project involves the characterization of *MTHFD1L* with a particular focus on genetic polymorphisms and the role of *MTHFD1L* levels in disease risk. The project will also involve an analysis of the potential function of additional homologous sequences. The key questions/aims that are being addressed are:

1. Development of a new genotyping assay for the technically challenging *MTHFD1L* DIP polymorphism rs3832406.
2. Is MTHFD1L a genetic risk factor for cleft disease in Ireland?

3. Do NTD-associated polymorphisms in the 3′UTR of *MTHFD1L* have a functional impact on miRNA regulation?
4. Investigation of why altered expression of MTHFD1L has a role to play in disease risk.
5. What is the functional role of other human MTHFD-like sequences?

The above aims will be met by achieving the following objectives:

1) Development of a Melting Curve genotyping assay for the triallelic DIP rs3832406 using Melting Curve analysis on the Roche Lightcycler 480 instrument.
2) Genotyping of key *MTHFD1L* polymorphisms in an Irish Cleft cohort and subsequent statistical analysis to test for an association.
3) Investigation of the NTD risk associated polymorphisms in the 3′ UTR of *MTHFD1L* for a potential impact on micro RNA (miRNA) regulation. This objective will involve the development of luciferase constructs and the utilisation of a cell culture model.
4) The role of altered MTHFD1L expression in disease will be assessed by further investigation of the Long and Short transcripts of *MTHFD1L* and by a proteomic analysis. Assessment of the impact of folate metabolism disruption on the relative ratio of the Long and Short transcripts of *MTHFD1L* will be assessed using a cycloleucine cell culture model and RT-qPCR. A proteomic analysis on cells that have been manipulated to inhibit the expression of MTHFD1L and to overexpress it will be carried out to provide an understanding of how the expression level of MTHFD1L is significant in disease.
5) Analysis of genomic sequences similar to *MTHFD1* and/or *MTHFD1L* to find any evidence of expression and to determine if these sequences have a functional relationship with their parent genes. Expression of candidate sequences will be assessed by carefully designed RT-qPCR assays with subsequent cloning and expression for functional characterisation.

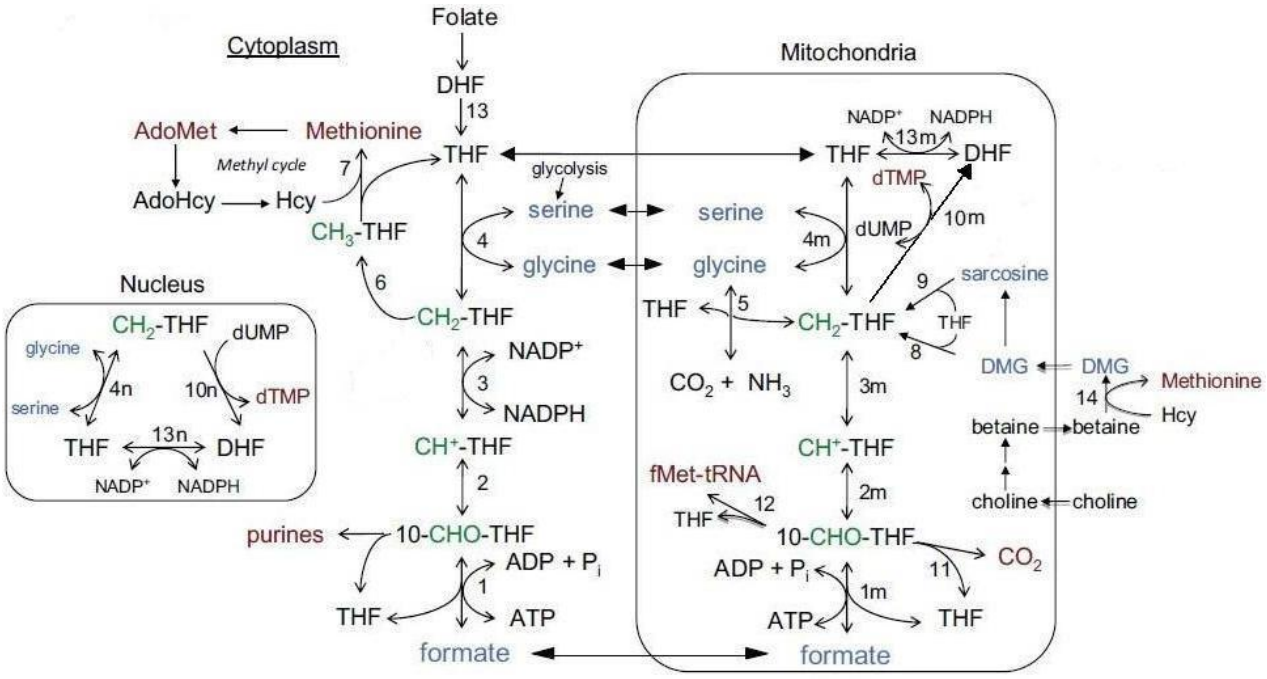


Figure 1.1 Compartmentalization of mammalian one-carbon metabolism (adapted from Tibbetts and Appling 2010).

End products of one-carbon metabolism are in red. One-carbon donors are in blue. Activated one-carbon units carried by tetrahydrofolate (THF) are in green. Reactions 1–4 and 13 are in both the cytoplasmic and mitochondrial (m) compartments. Reactions 4, 10 and 13 are also present in the nucleus (n). Reactions **1**, **2**, and **3**: 10-formyl-THF synthetase, 5,10-methenyl-THF (CH^+-THF) cyclohydrolase, and 5,10-methylene-THF (CH_2-THF) dehydrogenase, respectively, are catalyzed by trifunctional C1-THF synthase in the cytoplasm (MTHFD1). In mammalian mitochondria, reaction **1m** is catalyzed by monofunctional MTHFD1L and reactions **2m** and **3m** by bifunctional MTHFD2 or MTHFD2L. The other reactions are catalyzed by the following enzymes: **4**, **4n** and **4m**, serine hydroxymethyltransferase (SHMT); **5**, glycine cleavage system; **6**, 5,10-methylene-THF reductase; **7**, methionine synthase; **8**, dimethylglycine (DMG) dehydrogenase; **9**, sarcosine dehydrogenase; **10n** and **10m**, thymidylate synthase (TYMS); **11**, 10-formyl-THF dehydrogenase (only the mitochondrial activity of this enzyme is shown, but it has been reported in both compartments in mammals); **12**, methionyl-tRNA formyltransferase; **13**, **13n and 13m**, dihydrofolate (DHF) reductase; **14**, betaine-homocysteine

methyltransferase. AdoHcy, S-adenosylhomocysteine; AdoMet, S-adenosylmethionine; Hcy, homocysteine.

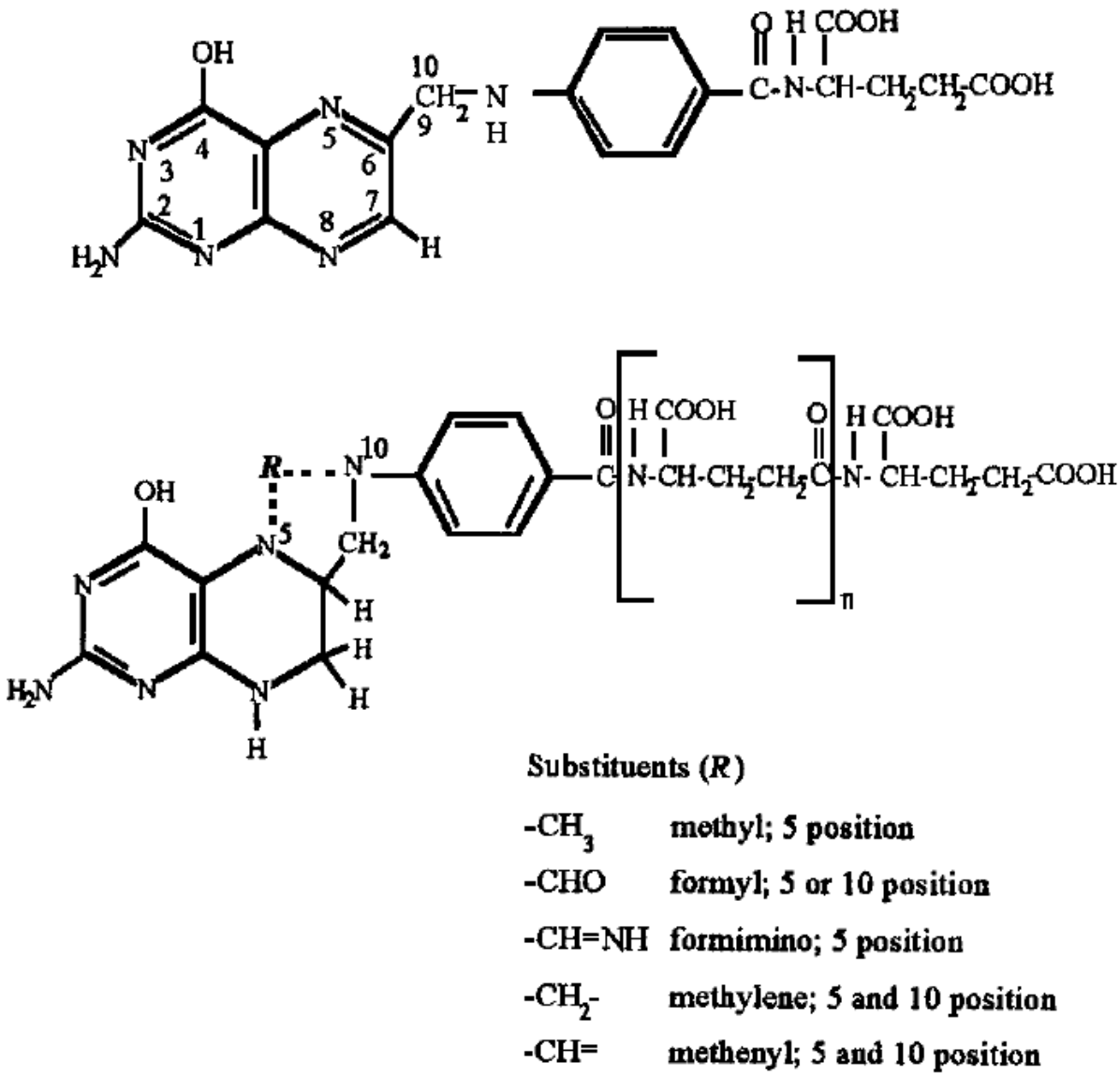


Figure 1.2 Folic acid and folate (modified from Forssén *et al*. 2000).

Structures of synthetic folic acid (up) and natural folates including reduced one-carbon substituted forms of polyglutamates (down).

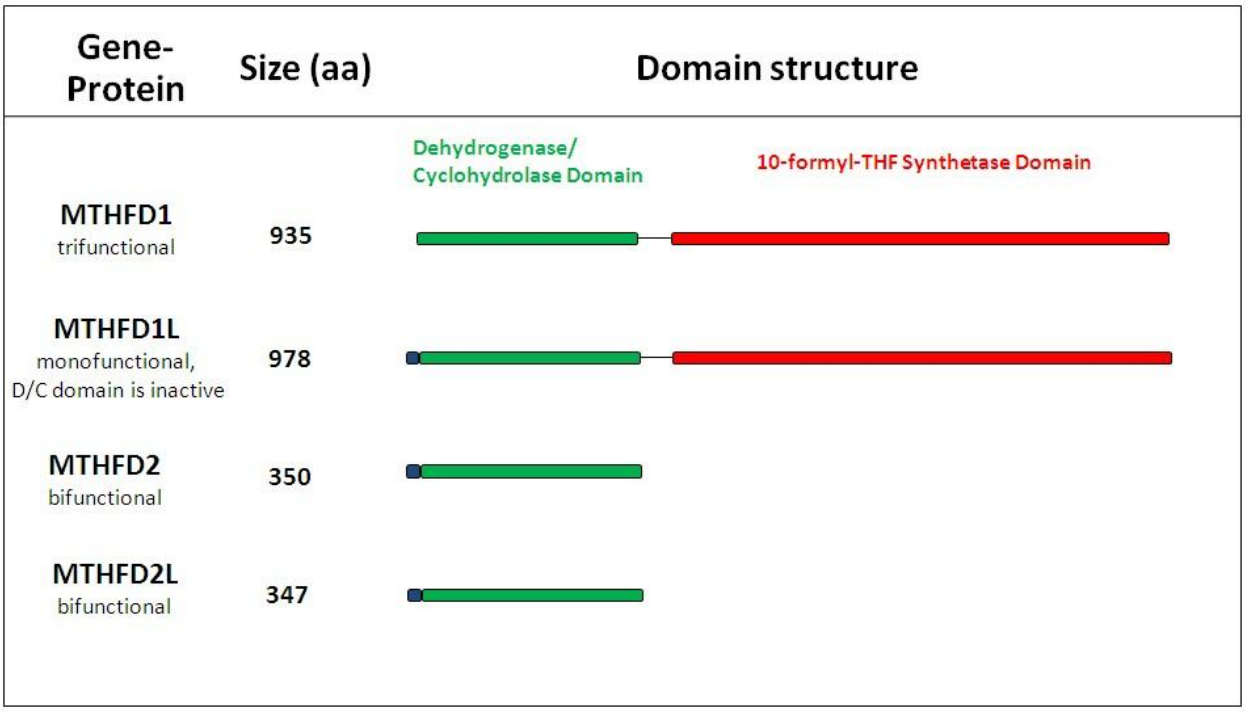


Figure 1.3 Domain structures of mammalian one-carbon interconverting enzymes (figure drawn with Microsoft PowerPoint).

MTHFD1 expresses a cytoplasmic trifunctional enzyme of 935aa composed of two domains, a ~30 kDa N-terminal D/C domain (*green*), which contains NADP-dependent CH_2-THF dehydrogenase and CH+-THF cyclohydrolase activities, connected to a ~70 kDa C-terminal domain (*red*), which contains the 10-formyl-THF synthetase activity. *MTHFD1L* encodes a mitochondrial protein of 978aa with the same two-domain structure, but with an N-terminal mitochondrial targeting sequence (*blue*). The D/C domain of *MTHFD1L* is not functional. *MTHFD2* and *MTHFD2L* both encode bifunctional NAD-dependent CH_2-THF dehydrogenase/CH+-THF cyclohydrolase and present an N-terminal mitochondrial targeting sequence.

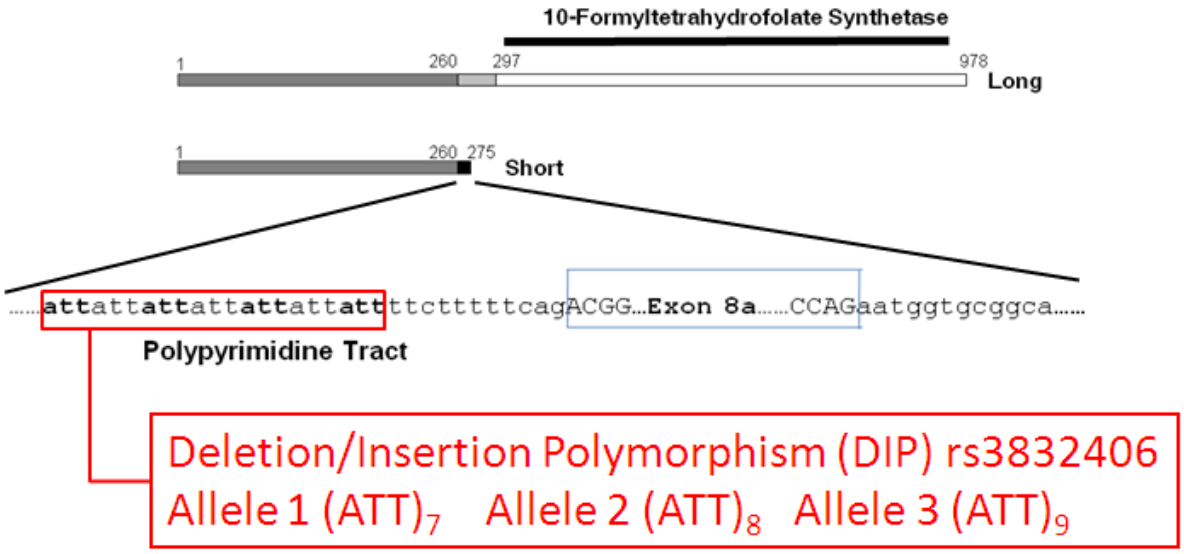


Figure 1.4 *MTHFD1L* Short and Long isoforms and DIP rs3832406 position (adapted from Parle McDermott *et al.* 2009).

The potential proteins resulting from translation of the *MTHFD1L* Long and Short mRNAs. The longer transcript encodes a monofunctional mitochondrial protein with a synthetase domain localized to the C-terminal end (black bar). The shorter transcript is caused by the inclusion of the alternative exon 8a which contains a premature stop codon. The shorter protein lacks a catalytic domain. DIP rs3832406 position is highlighted in red.

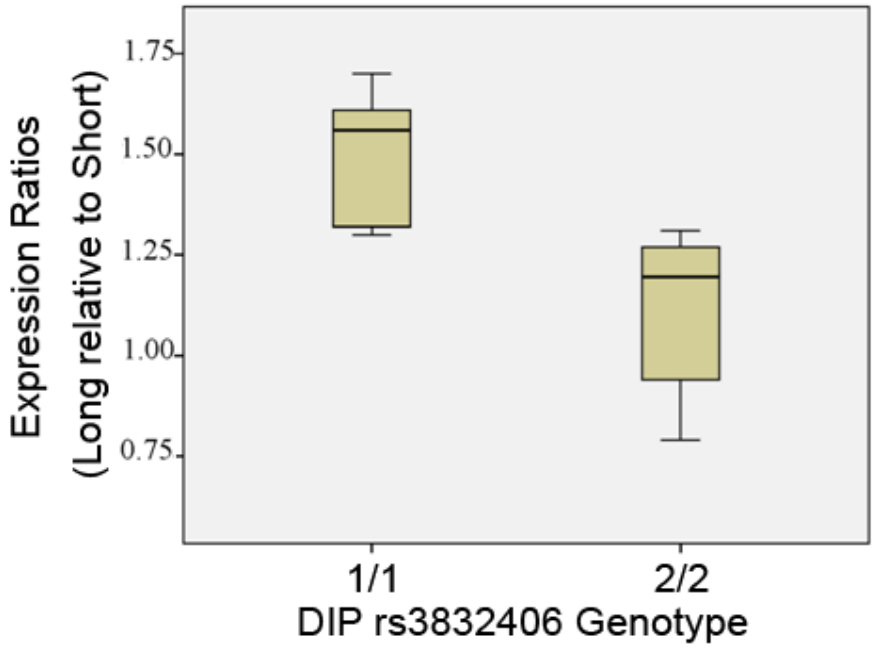


Figure 1.5 The ratio of MTHFD1L Long and Short transcripts differs by rs3832406 genotype in Coriell lymphoblast cell lines (Parle-McDermott *et al*. 2009).

Homozygotes for Allele 1 have a significantly greater proportion (mean is 35% more) of the Long transcript relative to the Short transcript compared to homozygotes for Allele 2 as assessed by Mann-Whitney U test (two tailed; P=0.006).

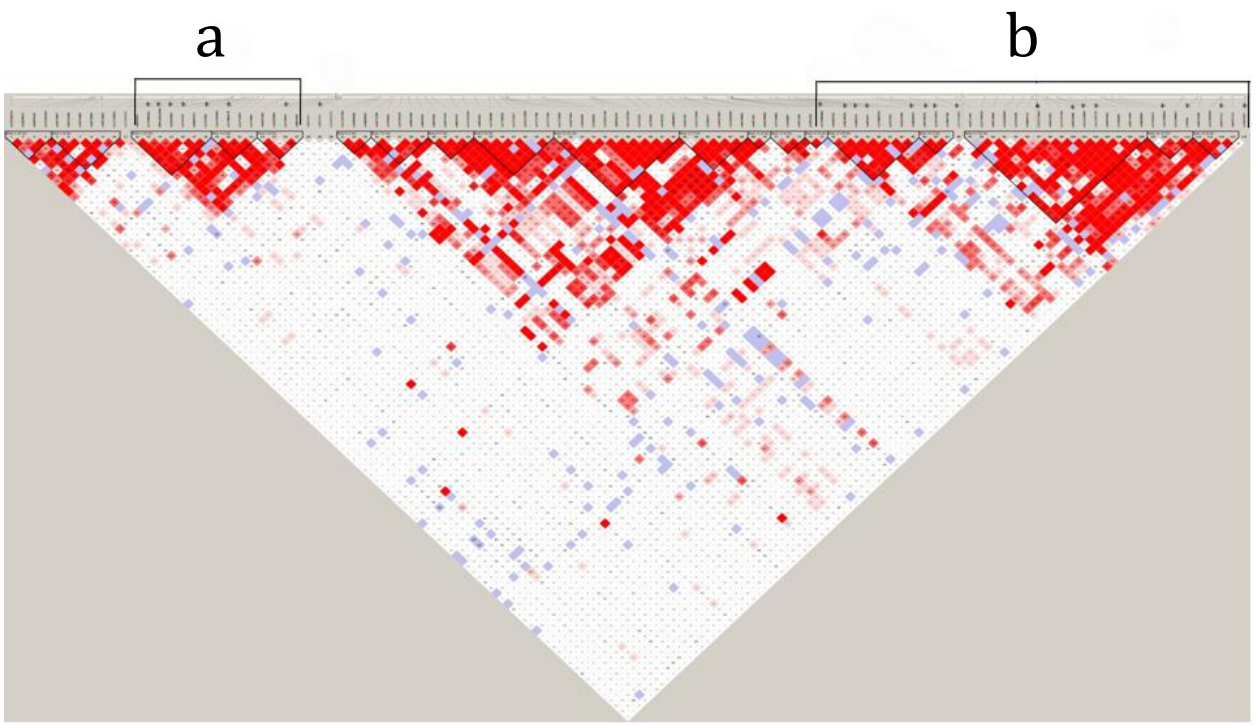


Figure 1.6 Linkage Disequilibrium (LD) plot of pairwise values of D' for 119 markers within and adjacent to MTHFD1L (Modified from Parle-McDermott *et al*. 2009).

The LD plot was constructed using genotype data from 338 Irish controls. Brackets (a, b) mark the two regions in which single markers were found to be associated with NTD risk. Region ''a'' contains the rs3832406 DIP. Each individual polymorphism that showed an association with NTD are indicated by an asterisk (*).

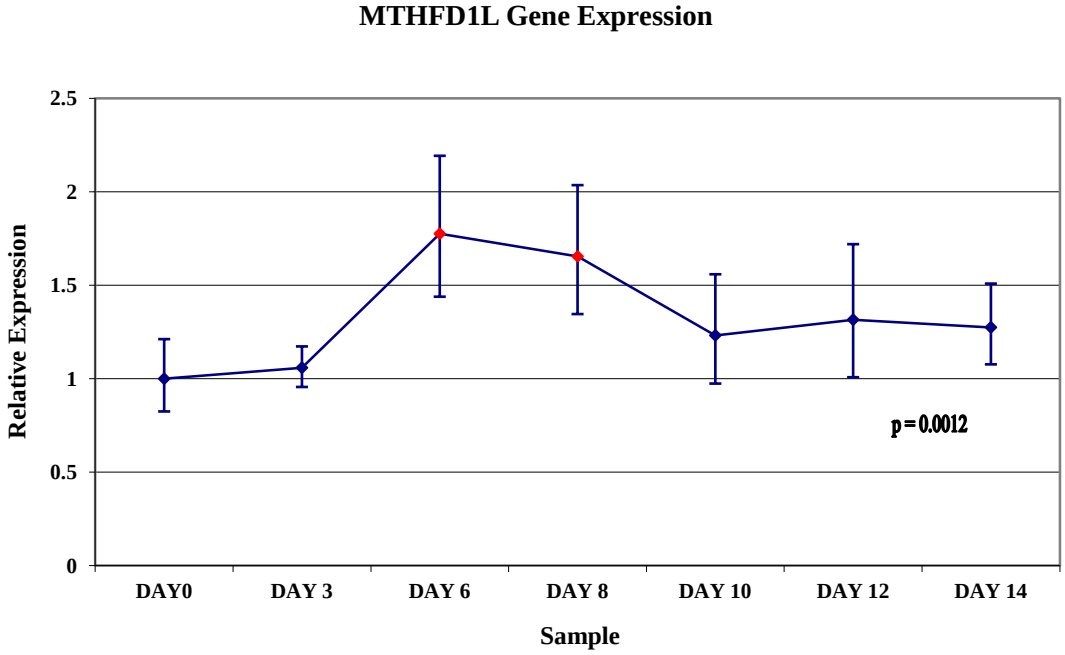


Figure 1.7 *MTHFD1L* gene expression was compared between samples from seven different time-points in a folate depletion time course analysis in HEK293 cells (experiment performed by Nicola Carroll).

The Day 0 sample is the calibrator for the relative quantification of gene expression. A significant difference in *MTHFD1L* mRNA levels between different samples was identified by one-way ANOVA, *p*-value = 0.0012. A post-hoc test determined samples at Day 6 and Day 8 (red data points) were those contributing to this difference when compared to the Day 0 sample.

CHAPTER 2

Material and Methods

2.1 Materials

2.1.1 Reagents

Ambion (Applied Biosystems): Placental RNA (Cat. No. AM7950), mir9 and mir197 Pre-miR™ precursor molecules (Cat. no. AM17100), negative control #1 Pre-miR™ precursor molecule (Cat. no. AM17110), miR-9, miR-197 and RNU6B TaqMan® assays (Cat. no. 4427975), anti-9 and anti-197 Anti-miR™ miRNA Precursor (Cat. no. AM17005), Anti-miR™ miRNA Inhibitor Negative Control #1 (Cat. no. AM17010), Taqman RT-qPCR assay (Cat. no. 4331182) for DHFR (Hs00758822_s1), TYMS (Hs00426591_m1), SHMT1 (Hs00244618_m1), MTHFD1 (Hs01068268_g1) .

Agilent Technology: QuikChange II Site-Directed Mutagenesis Kit (Cat. no. 200523).

Biological Resource Centre National Institute of Technology and Evaluation, Chiba, Japan: AK097152 and AK074198 sequences in pME18SFL3 vector (Appendix A).

Bioline: BioScript (Cat. no. BIO-27036), ISOLATE Plasmid Mini Kit (Cat. no. BIO-52026), ISOLATE RNA Mini Kit (Cat. no. BIO-52043), Oligo dT (Cat. no. BIO-38029), Random Hexamers (Cat. no. BIO-38028), SensiFast™ Probe One Step Kit (Cat. no. BIO-76001).

Biosera: Fetal Bovine Serum (FBS) (Cat no S1900).

Bio-Rad: Quick Start Bradford Protein Assay Kit (Cat. no. 500-0201).

Coriell Institute: human variation panel – Caucasian panel of 100 genomic DNA (Cat. no. HD100CAU), African Americans panel of 100 genomic DNA (Cat. no. HD100AA). Coriell lymphoblast cells: 17158 (Cat. no. GM17158), 17229 (Cat. no. GM17229), 17240 (Cat. no. GM17240), 17291(Cat. no. GM17291), 17189 (Cat. no. GM17189), 17205 (Cat. no. GM17205), 17220 (Cat. no. GM17220).

GE Healthcare: ExoSap-IT (Cat. no. US78200).

Greiner Bio-One: 75cm^3 flasks (Cat. no. 658175CI), 25cm^3 flasks (Cat. no. 690175CI), Cell Scrappers (Cat. no. 541070G), 6-well plates (Cat. no.657160), 24-well plate (Cat. no. 662160).

Gibco: Dulbecco's Modified Eagle Medium (DMEM) 1X (Cat. no. 41965), Trypan Blue Stain 0.4% (Cat. no. 15250-061). RPMI (Roswell Park Memorial Institute) 1640 (Cat. no. 12633-012).

Invitrogen: Gateway® pDEST™15 Vector (Cat. no. 11802-014) (Appendix D), Gateway® BP Clonase® II Enzyme mix (Cat. no. 11789-020), Gateway® LR Clonase® II Enzyme mix (Cat. no. 11791-020), Lipofectamine 2000 reagent (Cat. no. 11668500), SuperScript II (Cat. no. 18064), NuPAGE® 3-8% Tris-Acetate Gels 1.0 mm/12 well (Cat. no. EA03752BOX), NuPAGE® Tris-Acetate SDS Running Buffer (20X) (Cat. no. LA0041), HiMark™ Pre-Stained HMW Protein Standard (Cat. no. LC5699), iBlotR Gel Transfer Device (Cat. no. IB1001), iBlotR Gel Transfer Stacks, Regular Nitrocellulose (Cat. no. IB3010-01), NuPAGE® 4X LSD sample buffer (Cat. no. NP0007), pDONR™211 (Cat. no. 12536-017) (Appendix B), pcDNA™3.2-DEST(Cat. no. 12489-019) (Appendix C).

Lennox: Puromycin Dihydrochloride (Cat. no. CA2856.0010).

MitoSciences: monoclonal antibody anti PDH subunit E1 alpha (Cat. no. MSP03).

New England Biolabs: 100bp Ladder (Cat. no N3231L), Dnase I (Cat. no. M0303S).

Promega: ProteaseMAX™ Surfactant (Cat. no. V2071), Trypsin Gold, Mass Spectrometry Grade (Cat. no. V5280), Endoproteinase Lys-C Sequencing grade (Cat. no. V1071).

Qiagen: IPA Ingenuity software (Cat. no. IPA-NUL-00002), QIAamp DNA Blood Mini Kit (Cat. no 51106), RNeasy Mini Spin Columns (Cat. no. 74104), Qproteome Mitochondria Isolation Kit (Cat. no. 37612).

Roche: LightCycler 480 Genotyping Master (Cat. no. 04707524001).

Sigma-Aldrich: Acetonitrile (Cat. no. 271004), Agar (Cat. no. A5306), Ampicillian (Cat. no. A0166), Ammonium Bicarbonate (Cat. no. 09830)Agarose (Cat. no. A9539), CHAPS hydrate (Cat. no. C9426), Cycloleucine (Cat. no. A48105), DTT (Cat. no. D0632), Formic Acid (Cat. no. F0507), G418 disulfate salt (Cat. no. A1720), Hexadimethrine bromide (Cat. no. H9268-5G), Iodoacetamide (Cat. no. I1149), Isopropyl β-D-1-thiogalactopyranoside (IPTG) (Cat. no. I6758), Magnesium Chloride (Cat. no. M8787), NaCl (Cat. no. S7653), Orange G (Cat. no. O3756), Penicillin-Streptomycin (Cat. no. P4333), RNA Loading Buffer (Cat. no. R1386), Sodium Pyruvate (Cat. no. S8636), Taq

Polymerase 5U/µl (D4545), Trypsin-EDTA (Cat. no. T4049), Tryptone (Cat. no. T7293), Yeast Extract (Cat. no. Y1333), 10X PCR buffer without Magnesium Chloride (Cat. no. P2317), Non-target shRNA (Cat. no. SHC002) (Appendix E), MTHFD1L shRNA (SHCLNV-NM_015440) (Appendix F, Appendix G), Thiourea (Cat. no. T8656), Trifluoroacetic Acid (TFA) (Cat. no. T6508), Tris (Cat. no. T1503) UREA (Cat. no. U5378).

Switch Gear Genomics: LightSwitch Luciferase Assay System (Cat. no. LS010), *MTHFD1L* 3′UTR reporter construct (Cat. no. S890005).

Thermo Scientific: SuperSignal® West Femto Trial Kit (Cat. no. 34094), Restore™ PLUS Western Blot Stripping Buffer (Cat. no. 46430), BCA Protein Assay Reagent A (Cat. no. 23228) and Reagent B (Cat. no. 23224), Immuno Pure Goat Anti-Mouse IgG, (H+L) Peroxidase conjugated (Cat. no. IG1082801), Immuno Pure Goat Anti-Rabbit IgG, (H+L) Peroxidase conjugated (Cat. no. IE10656916), Pep-clean C18 spin columns (Cat. no. 89870), PepMap C18, 75 µm ID × 500 mm, 3 µm particle and 100 Å pore size (Cat. no. 164570), PepMap C18, 300 µm ID × 5 cm, 5 µm particle size, 100 Å pore size (Thermo Fisher Cat. no. 160292), protease inhibitors cocktail (Cat. no. BPE9706-1).

2.1.2 Solutions

Detergent free buffer for protein resuspension: 6M Urea, 2M thiourea, 10mM tris, pH 8.0

LB Agar: 15g of agar per litre of LB Broth

LB Broth: 2.5g Tryptone, 2.5g NaCl, 1.25g Yeast Extract, 250ml Deionised H_2O

Orange G (10X): 0.1 g Orange G, 20 g Sucrose, 50 mL H2O

PBS (10X): 80g NaCl, 2g KCl, 14.4g Na_2HPO_4, 2.4g KH_2PO_4, 1L H_2O (pH7.4)

Mammalian Protein Extraction Buffer: 50mM Tris, pH 7.8, 1mM EDTA, 0.1% v/v Triton X-100, 10mM β-mercapthoethanol, 0.5% v/v protease inhibitors cocktail, 0.07% v/v ß-mercaptoethanol.

Protein Extraction Buffer used for Mass Spec experiments: 7 M urea, 2 M thiourea, 30 mM Tris, 4% CHAPS, pH 8.5

TBE (10X): 48.44gTris HCl, 12.37g Boric Acid, 1.5g EDTA, 500ml H_2O

TBS (10X): 24.23g Tris HCl, 80.06g NaCl, 1L H_2O (pH 7.6)

TBS-T: TBS + 1% Tween

2.1.3 Cell Lines

HEK 293 (Human Embryonic Kidney): ATCC (Cat. no. CRL-1573)

MCF7 (Breast cancer cell line): ATCC (Cat. no. HTB-22)

MAX Efficiency® DH5Alpha (*Escherichia coli*): Invitrogen (Cat. no. 12034013)

2.2 Methods

2.2.1 Tissue culture methods

2.2.1.1 HEK 293 and MCF-7 cell culture

Human Embryonic Kidney 293 (HEK 293) and human breast adenocarcinoma (MCF-7) cell lines were cultured in Dulbecco's modified Eagle medium (DMEM) (Gibco) supplemented with 10% (v/v) fetal bovine serum and 1% (v/v) Penicillin-streptomycin solution (PenStrep) in a 5% CO_2 incubator at 37°C. Cells were usually cultured in 12 ml of media 75cm^3 flasks. Cell passages were performed by incubating with 3ml of Trypsin-EDTA (Sigma-Aldrich) at 37^oC for 5 minutes after washing cells with phosphate-buffered saline (PBS). Trypsin was then in-activated by adding twice the volume of DMEM. Cells were collected by centrifugation at 500 g for 5 minutes.

2.2.1.2 Coriell lymphoblast cell culture

Coriell lymphoblast cells were cultured in suspension in RPMI (Roswell Park Memorial Institute) 1640 supplemented with 10% (v/v) fetal bovine serum and 1% (v/v) Penicillin-streptomycin solution (PenStrep) in a 5% CO_2 incubator at 37°C. Cells were usually cultured in 12ml of media in 75cm^3 flasks. Cell passages were performed by centrifugation at 500 x g for 5 minutes. Cells were then resuspended in new media.

2.2.1.3 Cell counts

Cells were collected by centrifugation and pellet was resuspended in 1ml of complete media. Cell suspension (100µl) was then added to 400µl of trypan blue in a 1:5 dilution and left at R.T. for 5 minutes. The solution was then loaded to a chamber of the hemocytometer and observed under a microscope at 100x magnification. Viable cells were counted in the four corners (each corner has 16 squares). The same procedure was repeated for the other chamber of the hemocytometer. The coefficient of variation was calculated. The average cell count was then calculated and multiplied by the dilution factor and then by 10^4.

2.2.1.4 Cell transfection

Cells were counted and seeded in a 6 well plate at a density of 1-2 x 10^5 cells/ml in complete media without antibiotic and incubated at 37^{o}C for about 24 hours until 70%-90% confluency was obtained. The media was removed from the cells and replaced with 1ml of incomplete media (without FBS). For each well, 2-4µg of plasmid DNA or miRNA precursor/anti-miR inhibitor at a final concentration of 50 nM were mixed with 10ul of Lipofectamine 2000 transfection reagent and added to 1ml of incomplete media. Transfecting solution was then incubated for 30 minutes at R.T. and finally added to cells. Two negative controls were also performed: the first adding only 1ml of incomplete media and the second adding 1ml of incomplete media and 10ul of Lipofectamine 2000 (-CTR). Cells were incubated in a 5% CO_2 incubator at 37°C for 24-48 or 72 hours.

2.2.2 DH5α transformation

AK097152 and AK074198 clones containing ncRNA sequences (Chapter 5, Section 5.2.3) were transformed in DH5α competent cells by heat shock method. About 100 ng of plasmid were added to a 50 µl vial of DH5α competent cells. Cells were incubated on ice for 30 minutes and then heat shocked for 30 seconds at 42°C without shaking. Vial was immediately transferred to ice and 250 µl of room temperature S.O.C. medium was added. Vial was shaken horizontally at 37°C for 1 hour at 225 rpm in a shaking incubator. Cells were then spread on a LB agar plate 100 µg/ml ampicillin. After overnight incubation at

37ºC single colonies were selected and grown in liquid LB containing 100 µg/ml ampicillin.

2.2.3 Plasmid Mini Prep

Bacteria were cultured overnight and then they were centrifuged at 1500 x g for 10 minutes to pellet cells. The supernatant was removed and the pellet was resuspended in lysis buffer from ISOLATE Plasmid Mini Kit (Bioline). The plasmid DNA was then isolated as per manufacturer's protocol.

2.2.4 RNA Extraction and Reverse Transcription

Culture cells were centrifuged and the cell pellet was washed in 1ml of PBS buffer. RNA was extracted using either Qiagen or Bioline RNA extraction kit. Quantity and quality of RNA samples were analyzed by Nanodrop (Thermo Scientific) measuring at A260 and by gel electrophoresis methods (Chapter 2, Section 2.2.8). RNA samples (5-20 µg) were treated with 2-4 units of DNase I (NEB) at 37°C for 10 minutes. Upon addition of 1 µl of EDTA to a final concentration of 5 mM Dnase was heat inactivated at 75°C for 10 minutes. Samples were then reverse transcribed to cDNA using Bioscript (Bioline) reverse transcriptase and a mix of random Hexamers (Bioline) and Oligo dT primers (Bioline) as follows. RNA (2-4 µg) was mixed with 0.2 µg of random Hexamers and 0.5 µg of Oligo dT primers and incubated for 5 minutes at 70 °C then chilled on ice for at least 1 minute. Samples were then mixed with 1X RT buffer, 0.2 mM of each dNTP and 200U of Bioscript reverse transcriptase in a final volume of 20µl and incubated at 25°C for 10 minutes, 42°C for 30 minutes and 85°C for 5 minutes. A control reaction (-RT), without the addition of reverse transcriptase, was carried out in parallel for each sample to act as a control to ensure genomic DNA was adequately removed. All the cDNAs obtained were used as template in the PCR and qPCR reactions described below.

2.2.5 DIP rs35337982 validation assay

A PCR assay was designed to amplify 282 bp containing DIP rs35337982 in twenty different genomic DNA samples. Ten random samples were selected from Coriell Caucasian panel and ten from cleft study samples (Chapter 3, Section 3.2.1). PCR methods are described below. PCR samples were sequenced and then analyzed by BLAST (an example of sequenced sample is reported in Appendix S).

2.2.6 PCR

Unless indicated differently all PCR assays were carried out using 0.2 µM of each primer, 0.2 mM of each dNTP, 1.5 mM of $MgCl_2$, 1X PCR buffer without $MgCl_2$, 1U of Taq Polymerase and 1µl of cDNA in a final volume of 50 µl at the following conditions: 5 minutes pre-incubation at 95ºC, amplification 35 cycles of 30 seconds at 95ºC, 30 seconds at 55ºC, 1 minute at 72 ºC, final elongation for 5 minutes at 72 ºC. Primers were purchased from either Eurofins or IDT. All the primers used for RT-PCR analysis in Chapter 5 were designed to have at least 5-base mismatches with parental sequences. Primer positions are shown in Appendix Z, AA, and BB while primer sequences are shown in Table 2.1.

2.2.7 Genomic DNA contamination test

All cDNA products were also tested for genomic contamination by an intron spanning PCR assay designed to amplify the MTHFD1 gene. Resulting bands are 232 bp for cDNA amplification and 330 bp for gDNA amplification. The difference in size is due to the presence of an intron in the genomic sequence and its absence in the spliced reverse transcribed mRNA.

PCR reaction were carried out with 0.5 µM of each primer, 0.5 mM of each dNTP, 1.5 mM of $MgCl_2$, 1X PCR buffer without $MgCl_2$, 1U of Taq Polymerase and 1µl of cDNA in a final volume of 50 µl 10 at following conditions: 3 minutes of pre-incubation at 95ºC, amplification 35 cycles of 30 seconds at 95ºC, 1 minute at 58ºC, 1 minute at 72 ºC, final elongation for 10’ at 72 ºC.

Primer sequences are:
MTHFDRQ For 5′-CAC TCC AGT GTT TGT CCA TG -3′
MTHFDRQ Rev 5′-GCA TCT TGA GAG CCC TGA C -3′

2.2.8 Agarose gel

Agarose gels were made by boiling 1% agarose (w/v) in 30ml of 1xTBE. Liquid solution was then cooled to 55°C and mixed with 2µl of 500µg/ml ethidium bromide before pouring into mould. The electrophoresis buffer was also 1xTBE. PCR product (10 µl) was mixed with 2 µl of orange G loading dye and electrophoresised at 90 volts for 30-45 minutes. Gels were visualized under UV light using DNR Mini-Bis Pro Bio-Imaging System. Alternatively, extracted RNA quality was verified through the same procedure after heating it to 65^{o}C in RNA denaturing buffer for 5 minutes.

2.2.9 Sequencing

Samples were cleaned-up for Sanger sequencing as follows: 15 µl of sample were mixed with 5 µl of ExoSap-IT (GE Healthcare) and incubated at 37 °C for 15 minutes followed by enzyme inactivation at 80°C for 15 minutes. About 5-10 ng of PCR products were then sequenced by either Eurofins or Source Bioscience.

2.2.10 RT-qPCR

cDNA was obtained by reverse transcription as described in Chapter 2, Section 2.2.4 . All the assays were designed using the Universal Probe Library (Roche). For primers, probes and their concentrations see Table 2.2. Primers were obtained from Eurofins MWG or Integrated DNA Technologies (IDT) while the probes were from Roche. Taqman® Gene Expression assays were used for the following genes: MTHFD1, DHFR, TYMS and SHMT1 (Ambion). Specific reverse transcription and qPCR analysis of miR-9, miR-197 and RNU6B were conducted using TaqMan microRNA assays (Ambion). In all the experiments samples were run in triplicates. All assays and analysis were performed using

Sensimix™ II Probe Kit (Bioline) on the Roche Lightcycler® 480 instrument at following conditions: 5 minutes pre-incubation at 95ºC, amplification 45 cycles of 30 seconds at 95ºC, 30 seconds at 60ºC, 1 second at 72 ºC with signal acquisition. Results were calculated by applying the comparative E-method to measure the fold change in target gene expression relative to reference gene. Statistical significance was determined by Student's t-test or ANOVA.

2.2.11 Cycloleucine time course

17158 Coriell lymphoblast cell line was used for this experiment which was performed in triplicate. Cells were treated with cycloleucine dissolved in PBS at the following concentrations: 5mM, 7.5mM and 10mM. An untreated negative control was carried out in parallel. RNA was collected at the following time points for each concentration: 24hours, 48hours and 72hours. Cell cultures and RNA extraction were previously performed by Nicola Carroll (Carroll *et al.* 2012).

2.2.12 Invitrogen® Gateway cloning

The vectors used in Chapter 5, Section 5.2.4 and Chapter 7, Section 7.2.3 were generated using the Invitrogen Gateway cloning system (http://www.invitrogen.com/site/us/en/home/Products-and-Services/Applications/Cloning/Gateway-Cloning/Protocols.html). ChrX 3′UTR and ChrX 750 ORF sequences were PCR amplified from HEK293 cDNA (Chapter2, Section 2.2.4) using attB-flanked primers (Table 2.3). A modified version of *MTHFD1L* coding sequence with attB-flanking regions in pMA-RQ vector was purchased from Invitrogen® (Appendix H). *MTHFD1L* coding sequence was modified without changing the resulting aa sequence, in order to optimize the protein expression in mammalian cells.

Entry clones were generated from the above attB-flanked sequences through BP recombination reaction with pDONR™211 (Appendix B) vector following manufacturer's procedures (Invitrogen). A destination vector was then created by performing an LR

recombination reaction (Invitrogen) between the entry vector and the respective destination vectors. ChrX 3′UTR and MTHFD1Lopt sequences were cloned in the mammalian expression vector pcDNA™3.2/V5-DEST (Appendix C), while ChrX 750 ORF sequence was cloned in the *E.Coli* protein expression vector Gateway® pDEST™15 (Appendix D). pcDNA3.2-MTHFD1Lopt (Appendix I, J, K and L), pDEST15-750ORF (Appendix M and N) and pcDNA3.2-ChrX 3′UTR (Appendix O) sequences were completely verified by Sanger sequencing using Source Biosciences service based in Dublin. Primers used for Sanger sequencing are shown in Table 2.4.

2.2.13 Protein Analysis Methods

2.2.13.1 Mitochondrial and cytosolic protein extraction.

Cytosolic and mitochondrial fractions were isolated using Qproteome Mitochondria Isolation kit (Qiagen) following manufacturer's instruction. Protein acetone precipitation was performed to concentrate cytosolic fraction. Four volumes of ice-cold acetone were added to the protein fraction and incubated for 15 min on ice. Sample was centrifuged for 10 min at 12,000 x *g* in a pre-cooled centrifuge at 4°C. Supernatant was discarded and the pellet was washed with 1 ml of 80% acetone. Pellet was then let air dry. Cytosolic proteins were resuspended in 50 µl of H_2O. Mitochondrial fraction obtained following manufacturer's instruction were resuspended in 50 µl of protein extraction buffer and incubated on ice for 10 minutes after vortexing. Proteins were either used directly for SDS Page electrophoresis or stored for brief periods at -20°C.

2.2.13.2 Protein Concentration Estimation

Bicinchoninic acid assay (BCA) was used to determine protein concentration. Standards were prepared using Bovine Serum Albumen (BSA) diluted in sterile H_2O to concentrations of 25 µg/ml, 125 µg/ml, 250 µg/ml, 500 µg/ml, 750 µg/ml, 1000 µg/ml, 1500 µg/ml and 2000 µg/ml. Unknown samples were diluted 1/10 in sterile H_2O. The working reagent was prepared making a solution of reagent A and reagent B (1:50). Ten µl of either standard or sample were added in triplicate on a suitable 96-well plate and 100µl of the working reagent were added to each well. The plate was then incubated at 37^{o}C for

30 minutes. Samples were then measured at 562 nM using Tecan i600 spectrophotometer. BSA standard values were used to make the standard curve from which unknown sample concentrations were then calculated.

2.2.13.3 Sample Preparation, Electrophoresis and iBlot Dry Tranfer.

Samples were diluted with 0.5 volume of NuPAGE® 4X LSD sample buffer and boiled for 5 minutes on a heating block. Samples could then be stored at -20^{o}C until required. Proteins (25 µg) were loaded into precast NuPAGE® 3-8% Tris-Acetate Gel (1.0 mm/12 well) and proteins were separated by the application of a 140 volts current to the gel apparatus filled with NuPAGE® Tris-Acetate SDS Running Buffer. Five µl of HiMark™ Pre-Stained HMW Protein Standard were used to confirm the molecular weight of protein bands. The migration of bromophenol blue was monitored and the current switched off when the dye band reached the bottom of the gel (approximately 90 minutes). The gel was then removed from the apparatus and transfered to nitrocellulose membrane for 10 minutes using iBlotR Gel Transfer Device.

2.2.13.4 Western Blotting for MTHFD1L

The nitrocellulose membrane was stained with ATX Ponceau S red staining solution to ensure successful transfer and then washed several times in dH_2O. The nitrocellulose blotting paper was blocked by incubating in tris buffered saline (TBS) containing 5% non fat dried milk for 1 hour at R.T., with gentle shaking. The membrane was washed twice for 5 minutes in TBS containing 0.05% Tween (TBS-T) and probed with the primary antibody. The primary antibody was a rabbit polyclonal antibody raised against MTHFD1L protein (gifted from Prof. Dean Appling, University of Texas at Austin) and was used at a 1:500 dilution in TBS-T. The membrane was incubated with the primary antibody overnight at 4^{o}C with gentle shaking and then washed 4 times for 5 minutes in TBS-T. An anti-rabbit secondary antibody conjugated to horseradish peroxidase (HRP; anti-rabbit IgG-HRP) was used at a 1:20000 dilution in TBS-T. The membrane was incubated with the secondary antibody at 4^{o}C for 1hour at room temperature and washed 5-6 times for 5 minutes with TBS-T. The substrate Working Solution was prepared mixing the two components of Signal® West Femto kit at a 1:1 ratio. The membrane was then

incubated at room temperature for 5 min in 1 ml of substrate Working Solution before image was taken using Gene Gnome instrument (SynGene).

2.2.13.5 Western Blotting for PDH

After Western blotting for MTHFD1L, the membrane was stripped using Restore™ PLUS Western Blot Stripping Buffer. It was then probed using a mouse monoclonal primary antibody raised against PDH (Pyruvate Dehydrogenase) subunit E1 alpha at a 1:1,000 dilution. The membrane was incubated and washed as before and then probed with the HRP-linked anti-mouse secondary antibody. Again, 1ml of substrate Working Solution was added to membrane and incubated for 5 minutes before image was taken using Gene Genome instrument.

2.2.13.6 Image J density comparison

ImageJ software (http://rsb.info.nih.gov/ij/index.html) was used to compare the density of Western blot bands. A complete tutorial to use this software is present at this link http://lukemiller.org/index.php/2010/11/analyzing-gels-and-western-blots-with-image-j/.

2.2.14 Bioinformatics Resources

2.2.14.1 BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. Nucleotide BLAST was used to find sequences similar to *MTHFD1L* and *MTHFD1*. All the sequences analysis and sequence similarity localization were made using mRNA transcript variant 1 of *MTHFD1L* (accession number: NM_001242767) and MTHFD1 mRNA (accession number: NM_005956) using default parameters. Database research was carried out in Ref seq (RNA entries from NCBI's Reference Sequence project) and EST database that contains sequence data and other information on cDNA sequences, or "Expressed Sequence Tags".

2.2.14.2 ORF FINDER (http://www.ncbi.nlm.nih.gov/projects/gorf/)

The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool that finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database. This tool identifies all open reading frames using the standard or alternative genetic codes. Default parameter values were used for all analyses.

2.2.14.3 CLUSTALW (www.ebi.ac.uk/clustalw)

CLUSTALW is a multiple sequence alignment program for DNA or proteins. Default parameter values were used for all analysis.

2.2.14.4 miRBase Version 19 (http://www.mirbase.org)

Database of published miRNA sequences and annotation. Each entry in the miRBase Sequence database represents a predicted hairpin portion of a miRNA transcript with information on the location and sequence of the mature miRNA sequence.

2.2.14.5 miRANDA (http://www.microrna.org)

miRANDA is a microRNA binding prediction tool that uses an algorithm which incorporates current biological knowledge on target rules and on the use of an up-to-date compendium of mammalian miRNAs. The web resource provides users with functional information about the set of genes that are potentially regulated by a particular miRNA, the implied cooperativity of multiple microRNAs on a particular mRNA and microRNA expression profiles in various tissues.

2.2.14.6 Target Scan (http://www.targetscan.org)

TargetScan predicts biological targets of miRNAs by seeking for the presence of conserved 8mer and 7mer sites that match the seed region of each miRNA. Nonconserved sites can also be predicted. In mammals, predictions are ranked based on the predicted efficacy of targeting as calculated using the context+ scores of the sites. As an option, predictions are ranked by their probability of conserved targeting.

2.2.14.7 RNAfold (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi)

The RNAfold web server (Vienna RNA Package) predicts secondary structures of single stranded RNA or DNA sequences.

2.2.14.8 RNAcofold (http://rna.tbi.univie.ac.at/cgi-bin/RNAcofold.cgi)

The RNAcofold web server will predict secondary structures of single stranded RNA or DNA sequences upon dimer formation.

2.2.14.9 EST database (http://www.ncbi.nlm.nih.gov/nucest/)

EST The EST database is a collection of short single-read transcript sequences from GenBank. These sequences provide a resource to evaluate gene expression, find potential variation, and annotate genes.

2.2.14.10 RefSeq: NCBI Reference Sequence Database (http://www.ncbi.nlm.nih.gov/refseq/)

A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein.

Table 2.1 Primers used for RT-PCR assays.

Assay	Primer Name	Primer Sequence
Chr11seq	CHR11 A For	5′- GTT GTG AAT AGG ATA CAC CTT CAA TGA GG -3′
	CHR11 A Rev	5′- GGC TTT CAT AGG AAT TGA TTC AGG TTC -3′
Chr2seq and Chr9seq	NCBI 2 For	5′- TGC CCA TCT GCA TGG CAA AGA CC -3′
	NCBI 2 Rev	5′- CCC CAC CCG GCA TTC GTA GC -3′
	NCBI ALL For	5′- GCT GGT GCC ACT TGA ACC AAG T -3′
	NCBI 5 For	5′- CCT CGT CCC ACC CCC ATG GT -3′
ChrXseq	CHRX 1 For	5′- AGC AGA CAA CAT TGA ATT ACT CCT CAA -3′
	CHRX 3′UTR For	5′- TGA ATG GAT CAT TCT AAA CAG ATC ATC AT -3′
	CHRX 3′UTR 2nd Rev	5′- AAA ATG ACA TCC GGG AAA ACT CCT A -3′
	CHRX 2 For	5′- AGT ATT CCG GTC TCC ACC CT -3′
	CHRX 2 Rev	5′- TTG AGG AGT AAT TCA ATG TTG TCT GCT -3′
	CHRX 3 For	5′- AAG TCA CAC ACA GAC GGT CCA C -3′
	CHRX 3 Rev	5′- AGG GTG GAG ACC GGA ATA CT -3′
	CHRX 4 Rev	5′- GTG GAC CGT CTG TGT GTG ACT T -3′
	CHRX 5 For	5′- GAC TCA GAC AAT ACA CTG AAG GAG -3′
	CHRX 5 Rev	5′- GAA CGG GCC TCG TCG ATG -3′
	CHRX 6 For	5′- CAT ACG ACG AGG CCC GTT C -3′
	CHRX 6 Rev	5′- CGC ACA CAC ATA AAA AAG TTC TGG -3′
	CHRX 7 For	5′- CCA GAA CTT TTT TAT GTG TGT GCG -3′
DIP rs35337982 validation	A insertion For	5′- GGG ACG CCC ACC CAA GAA -3′
	A insertion Rev	5′- ACA CTG CTG CCC CAG GCA T -3′

Table 2.2 Primers and probes used for qPCR assays.

Assay	Primer / Probe	Name	Primer Sequence	Final Conc.
MTHFD1L Long Isoform	Forward	MTHFD1LuplLoHF	5′ - GAGCTCTGAAGATGCATGGAG - 3′	0.2 µM
	Reverse	MTHFD1LuplLoHR	5′ - TGCTTCTGGAGGTTACAGCA - 3′	0.2 µM
	Probe	#42		
MTHFD1L Short Isoform	Forward	MTHFD1LuplShrtFo	5′ - ACGCCAGCTTCAAAGCAA - 3′	0.2 µM
	Reverse	MTHFD1LuplShrtRe	5′ - TCACAGGAGAATCACTTCAACC - 3′	0.2 µM
	Probe	#13		
MTHFD1L opt*	Forward	MTHFD1LoptFor	5′- CGAGTTCAGCGAGATCCAG -3′	0.2 µM
	Reverse	MTHFD1LoptRev	5′- CTGGGGTCGGTCTTGTTG -3′	0.2 µM
	Probe	#41		
Beta-Glucuronidase (GUS)	Forward	GUSupl(HPLC)F	5′ - ACGCCAGCTTCAAAGCAA - 3′	0.2 µM
	Reverse	GUSupl(HPLC)R	5′ - TCACAGGAGAATCACTTCAACC - 3′	0.2 µM
	Probe	#57		
ncRNAChr9	Forward	NCBI5For	5′- GACTCAGACAATACACTGAAGGAG -3′	0.5 µM
	Reverse	NCBI5Rev	5′- GAACGGGCCTCGTCGATG -3′	0.5 µM
	Probe	#23		
ncRNAChr2	Forward	NCBIALLFor	5′- GCT GGT GCC ACT TGA ACC AAG T -3′	0.5 µM
	Reverse	NCBI5Rev	5′- GAACGGGCCTCGTCGATG -3′	0.5 µM
	Probe	#61		
40S ribosomal protein S13 gene (RPS13)	Forward	RPS13For	5′- GGT TGA AGT TGA CAT CTG ACG A -3′	0.2 µM
	Reverse	RPS13Rev	5′- TGT GCA ACA CCA TGT GAA TCT -3′	0.2 µM
	Probe	#68		
MTR	Forward	MTR upl For	5′- ACAGCAGCATCATGGTCAAG -3′	0.2 µM
	Reverse	MTR upl Rev	5′- TTCATGGAGCTCTTCTGCAA -3′	0.2 µM
	Probe	#53		
DHFRL1	Forward	DHFRL1uplFor	5′ - CGGACCTTAGAAAGTCACACATC- 3′	0.2 µM
	Reverse	DHFRL1uplRev	5′ - TCACAGGAGAATCACTTCAACC - 3′	0.2 µM
	Probe	#89		
GAPDH	Forward	gapdh-for	5′- CTCTGCTCCTCCTGTTCGAC -3′	0.2 µM
	Reverse	gapdh-rev	5′- ACGACCAAATCCGTTGACTC -3′	0.2 µM
	Probe	#60		
TBT (TATA Binding Protein)	Forward	tbt-for	5′- TTGGGTTTTCCAGCTAAGTTCT -3′	0.2 µM
	Reverse	tbt-rev	5′- CCAGGAAATAACTCTGGCTCA -3′	0.2 µM
	Probe	#24		
ACTB (Actin Beta)	Forward	acbt-for	5′- CCAACCGCGAGAAGATGA -3′	0.2 µM
	Reverse	acbt-rev	5′- TCCATCACGATGCCAGTG -3′	0.2 µM
	Probe	#64		

*MTHFD1L sequence optimized for protein expression in mammals (Chapter 2, Section 2.2.12).

Table 2.3 Primers used for Invitrogen® Gateway cloning.

Assay	Primer Name	Primer Sequence
ChrX 3′UTR	CHRX attB 3′UTR For	5′- GGGGACAAGTTTGTACAAAAAAGCAGGCTCC TGAATGGATCATTCTAAACAGATCATCAT -3′
	attB 3′UTR 2nd Rev	5′- GGGGACCACTTTGTACAAGAAAGCTGGGTCCTA AAAATGACATCCGGGAAAACTCCTA -3′
ChrX 750 ORF	attBCHRX750 For Good	5′- GGGGACAAGTTTGTACAAAAAAGCAGGCTCC ATGTGTGTGCGACAGCCTTCTCAG -3′
	attB Long ORF Rev	5′- GGGGACCACTTTGTACAAGAAAGCTGGG TCCTAGAGTGTGATCCGGTCTGCAATGATA -3′

Table 2.4 Sequencing Primers.

Primer Name	Primer Sequence
T7 Promoter For	5′- TAATACGACTCACTATAGGG -3′
T7 Terminator Rev	5′- GCTAGTTATTGCTCAGCG G -3′
MTHFD1Lopt For	5′- CGAGTTCAGCGAGATCCAG -3′
MTHFD1Lopt2 For	5′- GAACGCCCTGAAGCCCGAGAAG -3′
MTHFD1Lopt4 For	5′- TCGCCAATATCGCCCACGGC -3′
NCBI 2 For	5′- TGCCCATCTGCATGGCAAAGA CC -3′
NCBI ALL For	5′- GCTGGTGCCACTTGAACCAAG T -3′
NCBI 5 For	5′- CCTCGTCCCACCCCCATGGT -3′
A insertion For	5′- GGGACGCCCACCCAAGAA -3′
CHRX 1 For	5′- AGCAGACAACATTGAATTACTCCTCAA -3′
CHRX 2 For	5′- AGTATTCCGGTCTCCACCCT -3′
CHRX 3 For	5′- AAGTCACACACAGACGGTCCAC -3′
CHRX 5 For	5′- GACTCAGACAATACACTGAAGGAG -3′
CHRX 6 For	5′- CATACGACGAGGCCCGTTC -3′
CHRX 7 For	5′- CCAGAACTTTTTTATGTGTGTGCG -3′

CHAPTER 3

MTHFD1L polymorphisms and cleft disease

3.1 Introduction

3.1.1 Cleft lip with or without cleft palate

Cleft lip with or without cleft palate (CLP) and cleft palate only (CPO) are common birth defects of complex and heterogeneous aetiology and they are characterized by disruptions of normal facial structure (Figure 3.1) (Dixon *et al*. 2011). While these pathologies can occur in numerous Mendelian or teratogenic syndromes, the isolated, non-syndromic form represents 70% of all CPL cases and 50% of all CPO cases (Jugessur *et al*. 2009, Marazita *et al.* 2002). CLP and CPO, with an average birth occurrence ranging from 1 in 700 to 1 in 1,000 among European populations, are a cause of substantial morbidity and mortality. The frequency differs widely across geographical origin, with a prevalence as high as 1 in 500 in Asian and Native American population, and as low as 1 in 2,500 in African derived populations. These data suggest that the contribution given by individual risk factors may vary across different populations (Mossey *et al*. 2009). Furthermore, also the gender may influence the frequency: CLP have a 2:1 male to female ratio while CPO have around 1:2 male to female ratio. Since CLP and CPO are aetiologically heterogeneous, many efforts have been done in order to identify distinct risk factors spanning genetics and environment. The important role played by genetic factors is indicated by several twin and family studies (Mitchell *et al*. 2002) and by recurrence risk, which is 20-30 times greater than population prevalence (Mangold *et al*. 2010). Nevertheless, CLP and CPO are well known to be associated with several environmental risk factors, like, for example, maternal smoking (Little *et al*. 2002, Shi *et al*. 2008, Butali *et al*. 2013).

3.1.2 Cleft and folic acid

Many studies suggest that folate deficiency before or during pregnancy can be another environmental risk factor (Boyles *et al*. 2008; van Rooij *et al*. 2004, Wehby and Murray 2010, Blanton *et al*. 2011). Folate supplementation in pregnancy has been shown to reduce the recurrence of CLP in families (Chevrier *et al.*, 2007; Wilcox *et al*., 2007; Jia *et al.*, 2011; Kelly *et al*., 2012, Butali *et al*. 2013). Nevertheless this association is still

controversial (Ray *et al*. 2003, Wehby and Cassell 2010) as demonstrated by the mixed results of numerous association studies between cleft occurrence and folate related genes, including *MTHFR* (Shaw *et al*. 1998, Martinelli *et al*. 2001, Blanton *et al*. 2002, Prescott *et al*. 2002, Jugessur *et al*. 2003, Shotelersuk *et al*. 2003, van Rooij *et al*. 2003, Vieira *et al*. 2005, Mills *et al*. 2008, Butali *et al*. 2013), *MTHFD1* (Mostowska *et al*. 2010, Boyles *et al*. 2008, Mills *et al*. 2008), 5,10-methenyltetrahydrofolate synthetase (*MTHFS*) and 5-methyltetrahydrofolate-homocysteine methyltransferase (*MTR*) (Martinelli *et al*. 2006, Mostowska *et al*. 2010, Bufalino *et al*. 2010, Blanton *et al*. 2011).

To our knowledge, so far, genetic variants of the folate related gene *MTHFD1L* have never been tested for cleft association. As explained in Chapter 1, Sections 1.1.6-7, this gene is gaining relevance as a target for disease association studies as more and more investigations are reporting polymorphisms associated with illnesses like coronary artery disease (Samani *et al*. 2007, Bressler *et al*. 2010) and Alzheimer disease (Naj *et al*., 2010). Sugiura *et al*. (2004) have also shown that *MTHFD1L* is upregulated in human colon adenocarcinoma. In a recent publication, it has been also proved that deletion of Mthfd1l causes embryonic lethality and neural tube and craniofacial defects in mice (Momb *et al.* 2013).

Moreover, it has been previously reported that numerous *MTHFD1L* polymorphisms are associated with the risk of NTDs in the Irish population (Chapter 1, Section 1.1.7, Parle-McDermott *et al*. 2009). *MTHFD1L* gene encodes the mitochondrial C1-THF Synthase protein which has a monofunctional 10-formyl-THF synthetase activity lacking the 5,10-methylene-THF dehydrogenase and 5,10-methenyl-THF cyclohydrolase activities typically found in the trifunctional cytoplasmic protein encoded by *MTHFD1* (Prasannan and Appling 2008). The short *MTHFD1L* isoform of 1.1 kb (transcript variant 4), that lack enzymatic activity (Prasannan *et al*. 2003), is caused by the alternatively spliced exon 8a. Just upstream this exon there is a polypyrimidine tract containing the triallelic deletion/insertion polymorphism (DIP) rs3832406 which has been shown to be associated with NTDs (Chapter 1, Section 1.1.7). DIP rs3832406 is composed of several ATT repeats: Allele 1 has 7 ATT repeats, allele 2 has 8 repeats and allele 3 has 9 repeats. The DIP is

within the region "a" of the *MTHFD1L* LD map (Chapter 1, Figure 1.6). In this scenario DIP polymorphism could be the direct disease causing variant within the associated LD region "a" by affecting alternative splicing of the gene (Parle-McDermott *et al*. 2009).

Based on its association with neural tube defects (NTDs) (Parle-McDermott *et al.* 2009), and the previously detected association of its cytoplasmic homologue MTHFD1 in our cleft cohort, in this chapter of the project, we considered the mitochondrial enzyme MTHFD1L to be a prime candidate for consideration for association with cleft. We tested the DIP and another SNP within region "a" for a possible association with cleft disease (Figure 3.2). We genotyped *MTHFD1L* DIP rs3832406 and the most statistically significant NTD-associated SNP from that study i.e., rs17080476, in 981 Irish case-parent trios affected by CPL or CPO. We also adapted a melting curve method capable of genotyping deletion/insertion polymorphisms without the need for capillary electrophoresis. Finally, another *MTHFD1L* DIP rs35337982 was reported in NCBI database, but it has not been validated; here we carried out a brief investigation to validate it.

The results of this chapter have been published:
Minguzzi S, Molloy AM, Peadar K, Mills J, Scott JM, Troendle J, Pangilinan F, Brody L, Parle-McDermott A: Genotyping of a tri-allelic polymorphism by a novel melting curve assay in MTHFD1L: an association study of nonsyndromic Cleft in Ireland. *BMC Med Genet* 2012; doi:10.1186/1471.

3.2 Methods

3.2.1 Subjects and Ethic Statement

Cleft samples were obtained at the Dublin Cleft Centre in Ireland as previously described (Mills *et al*. 2008) from subjects with CPO or CLP along with their mothers and fathers. Not all triad families were complete because samples from all three family members were not always available. A total of 2688 samples including 758 complete triads and 223

incomplete triads were collected for this work. Medical information was provided by the attending surgeons for 87% of the cleft cases.

Following complete description of the study to the subjects, written informed consent was obtained from all participants. Ethical approval was granted by the hospitals, and the Institutional Review Board at NIH. Control samples were collected from a population of 56,049 pregnant women attending the three main maternity hospitals in the Dublin area between 1986 and 1990 as previously described (Mills *et al*. 2008 and Parle-McDermott *et al*. 2006b).

For this study a random sample, (*n* = 1,008), were selected after all mothers who had given birth to a child with a known malformation, including NTD, orofacial cleft, congenital heart defect, Down syndrome, or limb deformity had been removed. Mothers with a previous history of an NTD affected pregnancy were also excluded. Subjects gave consent for all the samples.

3.2.2 Genotyping Methods

Genomic DNA was extracted from blood or buccal swab collected samples using a QIAamp DNA Blood Mini Kit (Qiagen, UK). HybProbe melting curve assays were designed to genotyped DIP rs3832406 and SNP rs17080476 on a LightCycler 480 Real Time PCR machine (Roche) using LightCycler 480 Genotyping Master and are described in more detail below. Genotyping quality was verified by repeat genotyping of at least 10% of samples. The success rates and agreement rates for repeat genotyping were, respectively, 99.2% and 99.1% for DIP rs3832406, 99.7% and 100% for SNP rs17080476.

SNP rs17080476 assay

SNP rs17080476 reagents and analysis conditions are: Forward Primer 5′-GCAACTTTGTTTAGTATGAAAATTTGAT-3′ (4µM), reverse primer 5′-TCTGTCTTCACCCAGCC (2µM), anchor probe 5′-Bodipy630/650-

AAGAGGGGAAAAAAAACCTTTCTCCATTATTCCTA-PHO-3′ (0.4 µM), sensor probe 5′-ATTCATTTCTTTACAGCAGTGGGATTATGAAA-Fluorescein 3′ (0.2 µM), pre-incubation 10 minutes at 95ºC, amplification 45 cycles of 15 seconds at 95ºC, 15 seconds at 56ºC, 15 seconds at 72 ºC, melting curve 1 minute at 95 ºC, 2 seconds at 50 ºC, acquisition ramp till 80 ºC (0,11ºC/s, 5 acquisitions per ºC).

DIP rs3832406 assay

DIP rs3832406 reagents and analysis conditions are: forward primer 5′-AAGCTTCCTGTTACCAC-3′ (4µM), reverse primer 5′-AGGAGAATCACTTCAACC-3′ (2µM), anchor probe: 5′-AGCCCCACGTTTGAATTTTATGTTTTTCCTAAAGT-Fluorescein-3′ (0.2µM). Sensor probe: 5′BODIPY630/650-AGGGAAGATTATTATTATT ATTATTATTATTATTTTCTTTTTCAGACGGA-Phosphate-3′ (0.2µM), pre-incubation 10 minutes at 95 ºC, amplification 45 Cycles of 10 minutes at 95 ºC, 10 seconds at 56 ºC, 10 seconds at 72 ºC, melting curve 1 minute at 95 ºC, 2 seconds at 50 ºC, acquisition ramp till 70 ºC (0,02 ºC/s, 30 acquisitions per ºC).

3.2.3 Statistical Methods

Our primary analysis was made with isolated nonsyndromic cases of CLP and CPO and their parents were included. A secondary analysis was then carried out including nonsyndromic cleft cases with other defects. Hardy-Weinberg equilibrium (HWE) was tested within each subject class (case, mother, father and controls) for SNP and DIP by chi-squared test. Associations with CLP and CPO were tested for SNP and DIP in cases/controls and separately in mothers/controls by logistic regression and odds ratios using either a dominant or recessive disease model. Triads (case, mother, and father) were used to perform the Transmission Disequilibrium Test (TDT) of Spielman *et al*. (1993) for each allele separately. Case and maternal effects were also assessed by log-linear modeling using 2 degrees of freedom.

3.3 Results

3.3.1 Development of a novel assay to genotype DIP rs3832406 by Melting Curve Analysis

A single assay which is able to distinguish 3 alleles contemporaneously was developed taking advantage of the GC-rich regions flanking the DIP (Figure 3.3). As described above, this polymorphism is a repeated "ATT" sequence which has three common alleles, Allele 1 (ATT_7) Allele2 (ATT_8) and Allele 3 (ATT_9). A wide 50-bases sensor probe was designed to perfectly match allele 3 with 9 ATT repeats and its flanking regions, producing a melting temperature (Tm) of 63°C (Figure 3.3c). The same probe, starting the annealing from GC-rich regions, produces a 3-base mismatched bubble on allele 2 and a 6-base mismatched bubble on allele 1 causing a Tm of 60.3°C and 58.8°C respectively (Figure 3.3a-b). As a result the peaks produced by each allele are clearly distinguishable (Figure 3.4). This assay can be easily adopted by any laboratory with a Real-Time PCR machine with melting curve capacity. This is a valid alternative to capillary electrophoresis for the genotyping of multiple allele DIPs.

3.3.2 DIP rs3832406 showed an associated with CLP

CPL and CPO triad (mother, father and affected case) and control samples necessary for this association study were collected from the Irish population. Primary analysis was carried out on triads with isolated cases of CLP and CPO, while in secondary analysis nonsyndromic cases with other defects were included. DIP rs3832406 and SNP rs17080476 were genotyped in 2688 samples including 758 complete triads and in 1008 controls.

SNP and DIP genotype frequencies for all groups are shown in Table 3.1 Genotype distributions in all groups were in HWE. The DIP showed an association with CLP case status based on TDT and log-linear analysis (Table 3.2). Based on TDT Allele 1 is transmitted to the offspring 55% of times (p= 0.017) in isolated CLP cases, carrying an

increased risk of having this disease. When the multiple case families were added to the analysis this value becomes more significant (56% transmission, p= 0.006). Allele 3 which has the lowest frequency, was passed to the offspring only 42.9% of times (p= 0.045) in multiple CLP cases, appearing to have a protective role against the disease. A Log-linear analysis with 2 degrees of freedom based on a recessive disease model confirmed the case association for Allele 1 (p= 0.039) in multiple CLP cases (Table 3.2). However, correction of these significant p-values using permutational adjustment resulted in loss of statistical significance. All the other analysis performed for DIP and SNP in CLP or CPO cases did not exhibit a significant association with the disease (Table 3.3, 3.4 and 3.5).

3.3.3 DIP rs35337982 validation

Deletion/insertion polymorphism rs35337982 is reported on dbSNP Short Genetic Variations database, but it has not been validated. DIP rs35337982 is reported as an insertion of a single adenosine nucleotide in exon 19 of *MTHFD1L* which leads to a premature stop codon in exon 19 due to frameshift. Such a shortened sequence would probably result in an inactive protein. A brief investigation was conducted to validate the presence of DIP rs35337982 in 20 genomic DNA sample (Chapter 2, Section 2.2.5). Samples were amplified by PCR and then sequenced (an example of sequenced sample is reported in Appendix S). None of them showed to have the insertion at the DIP rs35337982 (data not shown).

3.4 Discussion

The *MTHFD1L* gene has received particular attention in recent years owing to its association with coronary artery disease, Alzheimer's disease and NTDs. Our previous study, demonstrated that the *MTHFD1L* rs3832406 DIP is functional by impacting on alternative splicing efficiency (Parle-McDermott *et al*. 2009). We report a new modified melting curve assay to genotype this functionally relevant triallelic *MTHFD1L*

polymorphism without the need for traditional capillary electrophoresis methods. We genotyped rs3832406 DIP and SNP rs17080476 in an Irish cleft cohort in a bid to test for association. Our initial analysis shows that Allele 1 of rs3832406 DIP is associated with CLP cases. Allele 1 as the disease risk allele is in concordance with our previous NTD study which showed a similar association (Parle-McDermott *et al*. 2009).

However, adjustment of our analyses resulted in loss of statistical significance indicating that this variant is simply not associated with CLP or that a larger sample size is required to detect an effect. We did not observe statistical significance with SNP rs17080476 which shares a D' value of 0.56 with DIP rs3832406 and represented the most statistically compelling variant from this genomic region in our NTD study (Parle-McDermott *et al*. 2009). In conclusion, our analysis shows no strong association between specific polymorphisms within the *MTHFD1L* gene and risk of cleft in an Irish cohort. We suggest further screening of rs3832406 DIP in a larger cohort and describe a new assay that will facilitate this. We have demonstrated that the modified Melting Curve Analysis developed for DIP rs3832406 could be a valid alternative to capillary electrophoresis for the genotyping of multiple allele deletion/insertion polymorphisms and can be employed by any laboratory with a Real-Time PCR instrument with melting curve capacity.

Finally we tried to validate the reported DIP rs35337982 of *MTHFD1L*. Twenty genomic DNA samples were genotyped without finding any trace of this DIP. The presence of this potential polymorphism probably leads to an inactive protein. Since MTHFD1L function is essential, such a DIP would exist only in a heterozygous state and it would be likely to be associated with serious dysfunctions of a rare disease. Hence, our sequencing data indicates that this variant is not common in the Irish population, but we cannot rule out that it is a rare variant or whether it is simply the result of a sequencing error. The recent increase of whole genome or exome sequencing of rare disease patients and a future creation of a shared database might answer this open question (Roach *et al*. 2010).

While a second DIP (rs35337982) was not validated within MTHFD1L, DIP rs3832406 has been shown to impact on MTHFD1L enzyme levels and is associated with NTD and potentially with cleft. The molecular mechanism of how it increases risk remains to be elucidated but is explored further in Chapter 7.

Figure 3.1 Examples of Cleft Lip (left) and Cleft Plate (right) (adapted from Dixon *et al*. 2011).

The picture on the left shows a left unilateral cleft lip case, while the picture on the right shows a soft cleft palate only case.

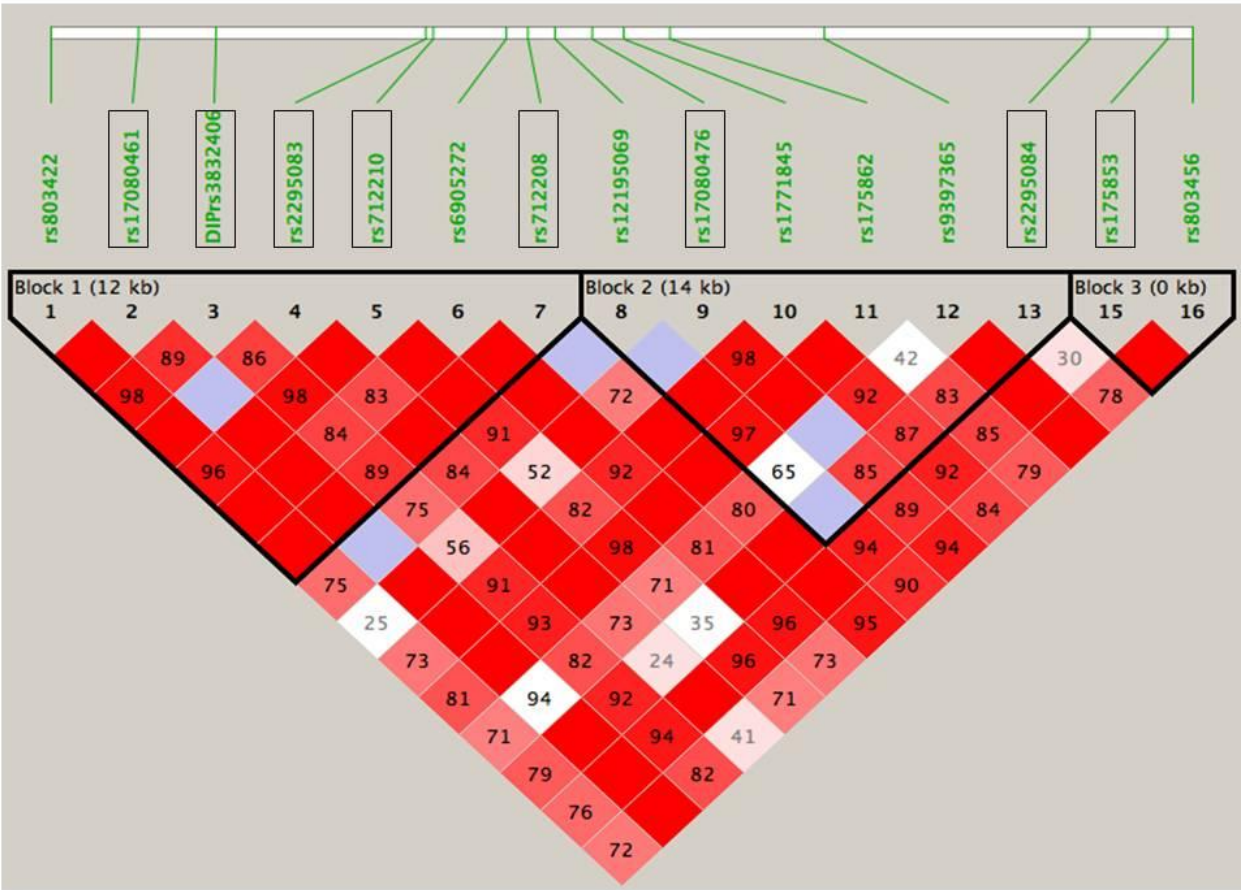


Figure 3.2 Linkage disequilibrium (LD) map of one associated region of the *MTHFD1L* Gene (marked region ‘‘a’’ in Chapter 1, Figure 1.6) (Parle-McDermott *et al*. 2009).

LD plot of pairwise values of D’ for 15 markers within the region spanning introns 7–10 of the *MTHFD1L* gene. Boxed markers were associated with NTD risk. DIP rs3832406 is the number 3, while SNP rs17080476 is the number 9. The DIP is in LD with the SNP with a D’ value of 0.56 and an r^2 value of 0.13. Therefore, although they are in LD, they do not correlate well for one another.

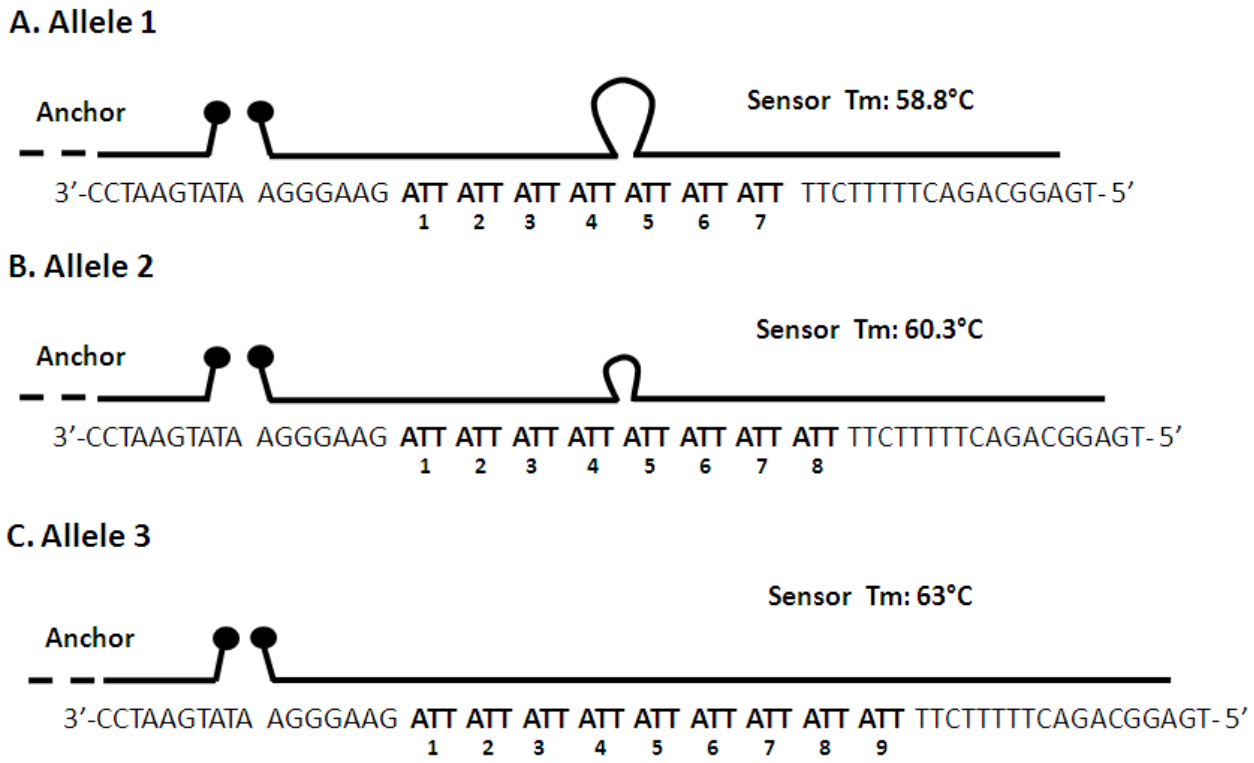


Figure 3.3 Modified Melting Curve Analysis for DIP rs3832406 (figure drawn with Microsoft PowerPoint).

The Sensor probe design for detection of all three alleles of DIP rs3832406 is shown. The Sensor probe is designed to perfectly match the complement of Allele 3.

A. Sensor probe bound to Allele 1 has a Tm of 58.8°C. **B**. Sensor probe bound to Allele 2 has a Tm of 60.3° **C**. Sensor probe bound to Allele 3 has a Tm of 63°C.

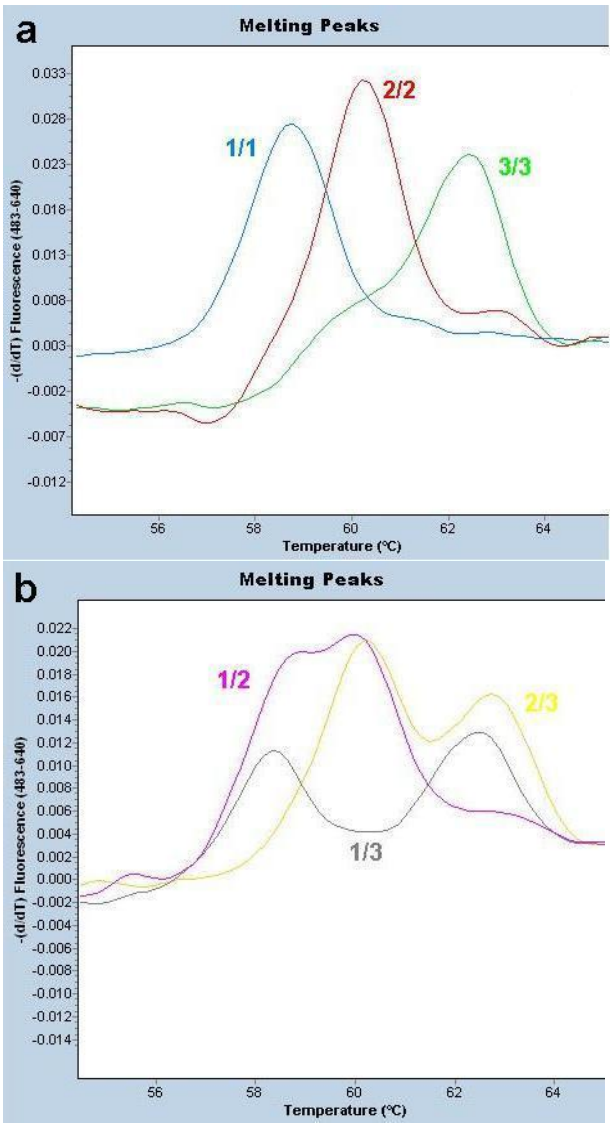


Figure 3.4 Peaks resulting from DIP rs3832406 Melting Curve Analysis.

Figure **a** shows the peaks produced from homozygote samples, while Figure **b** shows the heterozygote sample peaks.

Table 3.1 Genotyping Results in Triads (Cases, Mother and Fathers) and Controls for CLP and CPO (Isolated) or with Other Defects (Multiple).

DIPrs3832406	Isolated defects						Multiple defects							
CPL	Fathers		Mothers		Cases		Fathers		Mothers		Cases		Controls	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
11	162	41.8	187	40.1	209	41.9	174	42.0	203	40.8	231	43.5	419	42.1
12	102	26.3	128	27.5	149	29.9	116	28.0	140	28.1	163	30.7	267	26.8
13	76	19.6	89	19.1	84	16.8	82	19.8	97	19.5	91	17.1	196	19.7
22	18	4.6	24	5.2	18	3.6	18	4.3	24	4.8	18	3.4	40	4.0
23	24	6.2	31	6.7	28	5.6	24	5.8	34	6.8	28	5.3	52	5.2
33	6	1.5	7	1.5	11	2.2	6	1.4	7	1.4	11	2.1	21	2.1
CPO	Fathers		Mothers		Cases		Fathers		Mothers		Cases			
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%		
11	98	37.3	118	38.1	134	41.7	145	40.8	154	36.9	175	41.0		
12	78	29.7	82	26.5	88	27.4	97	27.3	110	26.4	118	27.6		
13	60	22.8	69	22.3	62	19.3	77	21.7	99	23.7	86	20.1		
22	8	3.0	12	3.9	7	2.2	8	2.3	18	4.3	9	2.1		
23	18	6.8	20	6.5	23	7.2	25	7.0	26	6.2	29	6.8		
33	1	0.4	9	2.9	7	2.2	3	0.8	10	2.4	10	2.3		

SNPrs17080476	Isolated defects						Multiple defects							
CPL	Fathers		Mothers		Cases		Fathers		Mothers		Cases		Controls	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
AA	268	68.2	296	63.1	320	63.1	285	67.1	320	62.9	344	62.5	660	65.9
AG	110	28.0	160	34.1	170	33.5	124	29.2	176	34.6	188	34.2	308	30.8
GG	15	3.8	13	2.8	17	3.4	16	3.8	13	2.6	18	3.3	33	3.3

	Isolated defects						Multiple defects					
CPO	Fathers		Mothers		Cases		Fathers		Mothers		Cases	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
AA	173	65.5	205	65.7	214	65.6	234	65.9	274	65.6	288	66.8
AG	86	32.6	93	29.8	100	30.7	115	32.4	127	30.4	125	29.0
GG	5	1.9	14	4.5	12	3.7	6	1.7	17	4.1	18	4.2

Table 3.2 TDT and Log-Linear analysis for DIP rs3832406 in isolated and multiple CPL cases. Significant values are marked in bold.

		TDT					Log-linear *p*-value	
	DIP Allele	Passed		Not Passed		P-value	Dominant	Recessive
		n	%	*n*	%			
Isolated CLP	1	197	55.0	161	45.0	**0.017**	0.075	0.066
	2	126	48.3	135	51.7	0.225	0.581	0.078
	3	91	43.5	118	56.5	0.073	**0.024**	0.368
Multiple CLP	1	219	56.0	172	44.0	**0.006**	**0.034**	**0.039**
	2	135	47.4	150	52.6	0.143	0.398	0.078
	3	96	42.9	128	57.1	**0.045**	**0.017**	0.593

Table 3.3 Logistic Regression Results for SNP rs17080476 and DIP rs3832406. OR = Odds Ratio CI = confidence interval

	Polym	Analysis	Dominant		Recessive		Multiplicative	
			OR (95% CI)	*p*-value	OR (95% CI)	*p*-value	OR (95% CI)	*p*-value
CLP all	DIP 1	Case-CTRL	1.1 (0.8, 1.5)	0.6159	1 (0.8, 1.3)	0.8468	1 (0.9, 1.2)	0.7022
		Mother-CTRL	0.9 (0.6, 1.2)	0.3917	0.9 (0.7, 1.1)	0.4774	0.9 (0.8, 1.1)	0.3476
	DIP 2	Case-CTRL	1.1 (0.9, 1.4)	0.3358	0.8 (0.5, 1.4)	0.4926	1.1 (0.9, 1.3)	0.5534
		Mother-CTRL	1.1 (0.9, 1.4)	0.2363	1.2 (0.7, 2)	0.5077	1.1 (0.9, 1.4)	0.2176
	DIP 3	Case-CTRL	0.9 (0.7, 1.1)	0.1928	1 (0.5, 2)	0.9158	0.9 (0.7, 1.1)	0.2349
		Mother-CTRL	1 (0.8, 1.3)	0.9044	0.7 (0.3, 1.5)	0.3311	1 (0.8, 1.2)	0.8726
	SNP	Case-CTRL	1 (0.6, 1.8)	0.9798	0.9 (0.7, 1.1)	0.1816	0.9 (0.7, 1.1)	0.2494
		Mother-CTRL	1.3 (0.7, 2.5)	0.4286	0.9 (0.7, 1.1)	0.2382	0.9 (0.8, 1.1)	0.4336
CLP iso	DIP 1	Case-CTRL	1 (0.7, 1.4)	0.9697	1 (0.8, 1.2)	0.9333	1 (0.8, 1.2)	0.9358
		Mother-CTRL	0.8 (0.6, 1.2)	0.2857	0.9 (0.7, 1.2)	0.4737	0.9 (0.8, 1.1)	0.2955
	DIP 2	Case-CTRL	1.1 (0.9, 1.4)	0.258	0.9 (0.5, 1.6)	0.6969	1.1 (0.9, 1.3)	0.4046
		Mother-CTRL	1.1 (0.9, 1.4)	0.2396	1.3 (0.8, 2.2)	0.3264	1.1 (0.9, 1.4)	0.1809
	DIP 3	Case-CTRL	0.9 (0.7, 1.1)	0.323	1 (0.5, 2.2)	0.9053	0.9 (0.7, 1.1)	0.3998
		Mother-CTRL	1 (0.8, 1.3)	0.9303	0.7 (0.3, 1.7)	0.4316	1 (0.8, 1.2)	0.8883
	SNP	Case-CTRL	1 (0.5, 1.8)	0.9538	0.9 (0.7, 1.1)	0.2786	0.9 (0.8, 1.1)	0.3381
		Mother-CTRL	1.2 (0.6, 2.3)	0.5906	0.9 (0.7, 1.1)	0.2905	0.9 (0.8, 1.1)	0.4526
CPO all	DIP 1	Case-CTRL	1 (0.7, 1.4)	0.9498	1 (0.8, 1.2)	0.6933	1 (0.8, 1.2)	0.7915
		Mother-CTRL	0.9 (0.6, 1.2)	0.3981	0.8 (0.6, 1)	0.0708	0.9 (0.7, 1)	0.0808
	DIP 2	Case-CTRL	1 (0.8, 1.3)	0.8703	0.5 (0.2, 1.1)	0.0749	1 (0.8, 1.2)	0.6497
		Mother-CTRL	1 (0.8, 1.3)	0.7614	1.1 (0.6, 1.9)	0.798	1 (0.8, 1.3)	0.7284
	DIP 3	Case-CTRL	1.1 (0.9, 1.4)	0.3873	1.1 (0.5, 2.4)	0.7843	1.1 (0.9, 1.4)	0.3963
		Mother-CTRL	1.3 (1, 1.7)	0.0431	1.1 (0.5, 2.4)	0.7368	1.2 (1, 1.5)	0.0576
	SNP	Case-CTRL	0.8 (0.4, 1.4)	0.4111	1 (0.8, 1.3)	0.7454	1 (0.8, 1.2)	1
		Mother-CTRL	0.8 (0.4, 1.5)	0.4739	1 (0.8, 1.3)	0.8894	1 (0.8, 1.2)	0.7186
CPO iso	DIP 1	Case-CTRL	1 (0.7, 1.5)	0.9335	1 (0.8, 1.3)	0.9081	1 (0.8, 1.2)	0.8998
		Mother-CTRL	0.8 (0.6, 1.2)	0.3735	0.8 (0.7, 1.1)	0.2065	0.9 (0.7, 1.1)	0.1723
	DIP 2	Case-CTRL	1 (0.8, 1.3)	0.8253	0.5 (0.2, 1.2)	0.1285	1 (0.8, 1.2)	0.7474
		Mother-CTRL	1 (0.8, 1.3)	0.824	1 (0.5, 1.9)	0.9076	1 (0.8, 1.3)	0.882
	DIP 3	Case-CTRL	1.1 (0.8, 1.4)	0.5704	1 (0.4, 2.5)	0.9392	1.1 (0.8, 1.4)	0.598
		Mother-CTRL	1.2 (0.9, 1.6)	0.1179	1.4 (0.6, 3.1)	0.4179	1.2 (1, 1.6)	0.1044
	SNP	Case-CTRL	0.9 (0.5, 1.7)	0.7393	1 (0.8, 1.3)	0.9235	1 (0.8, 1.2)	0.847
		Mother-CTRL	0.7 (0.4, 1.4)	0.3247	1 (0.8, 1.3)	0.9406	1 (0.8, 1.2)	0.6925

Table 3.4 Spielman Transmission Disequilibrium Test (TDT) and Log-Linear analysis for DIPrs3832406. GRR = genotype relative risk, CI =confidence interval. Significant values are marked in bold.

		DIP rs3832406					
		Spielman TDT		Log Linear Model			
				Dominant		Recessive	
	1 vs 2 or 3	GRR (95% CI)	*p*-value	GRR (95% CI)	*p*-value	GRR (95% CI)	*p*-value
	Case effect	1.3 (1.0 1.6)	**0.01734**	1.5 (1.0 2.2)	0.0754	1.3 (1.0 1.7)	0.0663
	Mom effect			0.9 (0.6 1.4)	0.6663	1.0 (0.7 1.3)	0.8828
CLPiso	**2 vs 1 or 3**						
	Case effect	0.9 (0.7 1.1)	0.2253	0.9 (0.7 1.2)	0.5807	0.5 (0.3 1.1)	0.0776
	Mom effect			1.1 (0.8 1.4)	0.6009	1.3 (0.6 2.4)	0.5059
	3 vs 1 or 2						
	Case effect	0.8 (0.6 1.0)	0.0727	0.7 (0.5 1.0)	**0.0236**	1.5 (0.6 3.4)	0.3677
	Mom effect			0.9 (0.7 1.3)	0.7472	0.8 (0.3 2.7)	0.7633
	1 vs 2 or 3						
	Case effect	1.3 (1.1 1.6)	**0.0062**	1.6 (1.0 2.3)	**0.0337**	1.3 (1.0 1.7)	**0.0393**
	Mom effect			0.9 (0.6 1.3)	0.5228	1.0 (0.8 1.3)	0.9429
CLPall	**2 vs 1 or 3**						
	Case effect	0.8 (0.7 1.1)	0.1432	0.9 (0.7 1.2)	0.3976	0.5 (0.3 1.1)	0.0775
	Mom effect			1.0 (0.8 1.4)	0.7729	1.3 (0.6 1.4)	0.5059
	3 vs 1 or 2						
	Case effect	0.8 (0.6 1.0)	**0.0445**	0.7 (0.5 0.9)	**0.0171**	1.2 (0.6 2.8)	0.5926
	Mom effect			1.0 (0.7 1.3)	0.8751	0.8 (0.3 2.7)	0.7633
	1 vs 2 or 3						
	Case effect	1.1 (0.8 1.4)	0.5399	1.4 (0.9 2.2)	0.1564	1.0 (0.7 1.3)	0.8062
	Mom effect			0.9 (0.5 1.5)	0.5795	1.0 (0.7 1.4)	0.8539
CPOiso	**2 vs 1 or 3**						
	Case effect	0.9 (0.7 1.2)	0.5989	1.0 (0.7 1.4)	0.9325	0.6 (0.2 1.4)	0.2307
	Mom effect			0.9 (0.6 1.3)	0.6016	1.1 (0.4 2.9)	0.8085
	3 vs 1 or 2						
	Case effect	1.0 (0.7 1.3)	0.8083	0.9 (0.7 1.3)	0.7549	1.1 (0.4 2.7)	0.9028
	Mom effect			1.1 (0.8 1.6)	0.5776	6.0 (0.7 49.8)	0.0972
	1 vs 2 or 3						
	Case effect	1.1 (0.9 1.3)	0.4866	1.4 (0.9 2.4)	0.1223	1.0 (0.7 1.2)	0.8326
	Mom effect			0.8 (0.5 1.3)	0.4738	0.8 (0.6 1.1)	0.2436
CPOall	**2 vs 1 or 3**						
	Case effect	1.0 (0.7 1.2)	0.7422	1.0 (0.8 1.4)	0.9119	0.6 (0.3 1.4)	0.2449
	Mom effect			1.0 (0.8 1.40	0.9397	1.8 (0.7 4.2)	0.2067
	3 vs 1 or 2						
	Case effect	0.9 (0.7 1.2)	0.5774	0.9 (0.7 1.2)	0.5666	1.0 (0.4 2.3)	0.9136
	Mom effect			1.1 (0.8 1.5)	0.4295	2.3 (0.6 9.0)	0.2195

Table 3.5 Spielman Transmission Disequilibrium Test (TDT) and Log-Linear analysis for SNP rs17080476. GRR = genotype relative risk, CI =confidence interval. Significant values are marked in bold.

		SNP rs17080476					
		Spielman TDT		Log Linear Model			
				Dominant		Recessive	
		GRR (95% CI)	*p*-value	GRR (95% CI)	*p*-value	GRR (95% CI)	*p*-value
CLPiso	Case effect	0.9 (0.7 1.1)	0.2157	0.8 (0.4 1.5)	0.4495	0.9 (0.6 1.1)	0.2806
	Mom effect			1.4 (0.6 3.2)	0.4164	0.8 (0.6 1.1)	0.1917
CLPall	Case effect	0.8 (0.7 1.1)	0.1907	0.9 (0.5 1.7)	0.8119	0.8 (0.6 1.1)	0.1629
	Mom effect			1.5 (0.7 3.3)	0.3206	0.9 (0.7 1.2)	0.3715
CPOiso	Case effect	0.9 (0.7 1.3)	0.7006	1.0 (0.4 2.1)	0.9183	0.9 (0.7 1.3)	0.6962
	Mom effect			0.5 (0.2 1.3)	0.1438	0.9 (0.6 1.30	0.6548
CPOall	Case effect	1.0 (0.7 1.2)	0.7411	0.7 (0.4 1.4)	0.3272	1.0 (0.8 1.3)	0.9703
	Mom effect			0.4 (0.2 1.1)	0.0825	0.9 (0.7 1.2)	0.4889

CHAPTER 4

MTHFD1L and miRNA regulation

4.1 Introduction

4.1.1 MicroRNA regulation and biogenesis

MicroRNAs (miRNAs) are small endogenous noncoding RNAs (ncRNAs) that can pair to sites in the mRNAs of protein-coding genes to modulate their expression, playing a role in many biological processes. The first miRNA was discovered in yeast in 1993 (Lee *et al*. 1993, Wightman *et al*. 1993) and since that, these ~22-nucleotide regulators have been thoroughly investigated, making them the best characterized group among ncRNAs. To date more than 2000 human miRNAs have been identified with each having up to hundreds of unique target mRNAs (miRBase Version 19, Griffiths-Jones *et al.* 2006, Kozomara *et al*. 2011).

Most microRNA genes originate from intergenic regions or in anti-sense orientation to genes and contain their own promoter (Lee *et al.* 2004). The biogenesis of miRNAs begins with the transcription of primary transcripts (pri-miRNAs) by either RNA polymerase II or RNA polymerase III in the nucleus (Figure 4.1) (Winter et a. 2009). pri-miRNAs may be edited by adenosine deaminases acting on RNA (ADARs) which modifies adenosine (A) into inosine (I) enhancing their processing by Drosha (Kawahara *et al*. 2008). In the nucleus, pri-miRNAs are then processed by the microprocessor complex which recognises their hairpin structure. This complex is formed by the RNase III type endonuclease Drosha and its partner, DGCR8 (DiGeorge syndrome critical region 8). It produces a ~60-nucleotide stem loop called precursor miRNA (pre-miRNA), which is exported to the cytoplasm by exportine 5 (XPO 5) in complex with Ran-GTP61 (Yi *et al*. 2003). Once in the cytoplasm, the pre-miRNA are cleaved by Dicer and TAR RNA-binding protein (TRBP), generating a ~22-nucleotide mature miRNA duplex (Inui *et al*. 2010).

At this stage, one strand of miRNA is usually degraded and the other strand is loaded into the RNA-induced silencing complex (RISC). This complex, which contains Argonaute (Ago) proteins, is the cytoplasmic effector machine of the miRNA pathway. It is guided by the single-stranded miRNA to its final target mRNAs partially complementing its sequence. A pairing between the seed sequence of the miRNA and

the seed match sequence in the 3′UTR of the mRNA is particularly important. The RISC can repress the target mRNA through two main mechanisms that have different variations (Inui *et al*. 2010). It can remove the polyA tail by promoting the activity of deadenylases (such as CCR-NOT), causing mRNA degradation. Alternatively it can block the translation at the initiation step or at the elongation step by causing, for example the ribosome stalling or the repression of initiation factor 4E (EIF4E).

4.1.2 Neural Tube Defects and *MTHFD1L*

Common birth malformations known as neural tube defects (NTDs) are severe craniospinal defects resulting from the failure of the neural tube to close during early embryonic development (Rampersaud E 2006). Low periconceptional maternal folate levels are well known as important risk factors for NTD, even though the mechanisms behind this association remains poorly understood (Kirke 2005). It is claimed that 50-70% of NTDs in humans could be prevented if women ingest a folic acid supplement in the periconceptional period (Wald 1991, Czeizel and Dudas 1992, Zhang *et al.* 2006). Common genetic variation is another important risk factor for NTD, whereby genes involved in the transport and metabolism of folate are prime candidates for genetic association studies.

One of these genes involved in the folate pathway is *MTHFD1L*, which encodes a mitochondrial mono-functional enzyme with a 10-formyltetrahydrofolate synthetase activity (Prasannan *et al.* 2003). This enzyme localizes to mitochondria and catalyzes the synthesis of formate within the highly compartmentalized one-carbon (1C) metabolism. In this crucial metabolism, 1C units are processed and transported through cytoplasm and mitochondria, generating a mostly unidirectional flow that supports important biological functions like *de novo* purine and thymidylate biosynthesis, methionine biosynthesis and amino acid metabolism (Prasannan and Appling 2009). Metabolic experiments carried out in mouse embryonic fibroblasts indicate that more than 75% of 1C units that enter the cytoplasmic methyl cycle derive from mitochondria (Pike *et al.* 2010). It appears that MTHFD1L is the only enzyme in both embryos and adults, (which catalyzes the last step in the production of formate in the mitochondria), which is subsequently exported to the cytoplasm. This role puts MTHFD1L under the spotlight for a possible involvement in human disease including cancer, which has been

investigated in several studies. For instance, *MTHFD1L* expression is upregulated in human colon cancer (Sugiura *et al.* 2004) and breast cancer (Selcuklu *et al.* 2012), and high levels of this protein are correlated with growth rate of human cancer cell lines (Jain *et al.* 2012). In addition, genome-wide association studies demonstrated that *MTHFD1L* polymorphisms are associated with coronary artery disease (Samani *et al.* 2007, Bressler *et al.* 2010) and Alzheimer's disease (Naj *et al.* 2010).

In mouse models MTHFD1L is clearly linked to NTD. Where *Mthfd1l* knockout mice have been analysed, the loss of the protein was found to be lethal to developing embryos, causing fetal growth restriction and aberrant neural tube phenotypes with 100% penetrance (Momb *et al.* 2013). All the *Mthfd1l* knockout embryos developed a clear NTD phenotype failing to close the neural tube (exencephaly or craniorachischisis) or displaying a wavy neural tube with a small, aberrantly formed head (Momb *et al.* 2013). Maternal supplementation with formate significantly decreased the incidence of NTDs and partially rescued the growth defects in mouse embryos lacking *Mthfd1l* (Momb *et al.* 2013).

In humans *MTHFD1L* variants have been shown to be associated with risk of NTDs in an Irish cohort (Parle-McDermott *et al.* 2009). In our previous study, two separate regions (referred to as regions "a" and "b") of the gene exhibited significant association signals suggesting the presence of two independent disease-causing variants (Chapter 1, Section 1.1.7, Figure 1.6). The "a" region at the 5′ end of the gene has already been investigated and a common deletion/insertion polymorphism (DIP) rs3832406 was proposed as a functional variant through affecting the alternative splicing of *MTHFD1L* (Parle-McDermott *et al.* 2009).

In this chapter, we have focused on the "b" region that overlaps the 3′UTR of *MTHFD1L* (Figure 4.2) and have investigated SNP variants of this region that could alter microRNA regulation. MicroRNAs (miRNAs) are important regulators of gene expression levels, involving binding of the miRNAs (predominantly) to target sites in the 3′UTR of targeted genes (Bartel 2004). *MTHFD1L*, along with another folate-related gene *MTHFD2*, has been previously identified as direct targets of miR-9 in MCF-7 breast cancer cells (Selcuklu *et al.* 2012). Using miRNA binding prediction tools, in addition to the previously confirmed miR-9 binding (Selcuklu *et al.* 2012), we

found that miR-197 also has a predicted binding site on MTHFD1L 3′UTR, which overlaps with the miR-9 binding site and the NTD associated SNP rs7646 (Figure 4.3a). In this chapter, we have further examined the binding efficacy of miR-9 and miR-197 to MTHFD1L 3′UTR and the effects on *MTHFD1L* expression levels in different genotypic backgrounds (SNP rs7646 A/G alleles). In addition, other folate related genes have been investigated as possible targets of miR-9 and miR-197.

The majority of the data in this chapter has been submitted for publication to *Human Mutation* and is currently under review:
Minguzzi S, Selcuklu SD, Spillane C, Parle-McDermott A: An NTD-associated polymorphism in the 3′ UTR of MTHFD1L can affect disease risk by altering miRNA binding.

4.2 Materials and methods

4.2.1 Constructs and LightSwitch Luciferase Assay

The original full-length 3′UTR reporter construct for *MTHFD1L* (obtained from Switchgear Genomics) contained the A allele, designated as UTR-A. Four nucleotides were mutated in the seed position of the related miRNA within the MTHFD1L 3′UTR luciferase vector in order to generate a negative control (Figure 4.3b). Mutations were generated by using QuikChange II Site-Directed Mutagenesis Kit (Agilent Technology) following manufacturer's protocol. These negative controls are necessary to demonstrate that the miRNA knocks-out the gene reporter specifically binding the predicted sequence in the 3′UTR. In the same way, the A nucleotide in SNP7646 position of MTHFD1L 3′UTR (UTR_snpG) was replaced with a G nucleotide. These constructs were designated as UTRmut9, UTRmut197 and UTR-G, respectively (Figure 4.3b). All the mutated constructs were generated by Duygu Selcuklu in the Charles Spillane's laboratory in the National University of Ireland, Galway.

HEK293 and MCF7 cells were cultured for one week (two passages) in complete DMEM 1640 in 75cm^3 flasks as per Chapter 2, Section 2.2.1.1. In each well of a 24-well plate 2 x 10^5 cells were seeded in 500 µl of complete DMEM 1640 and they were

grown for 24 hours until they reached a confluency of about 80%. Cells were transiently transfected following pLightSwitch system protocol. Each 3′UTR or mutated construct (100 ng) was co-transfected with 50 nM of Pre-miR™ miRNA precursor (pre-9, pre-197, pre-control) or 50 nM of miRNA inhibitor oligos (anti-9, anti-197, anti-control) (Ambion) in 24-well plates. Pre-miR™ miRNA Precursor Molecules are small, chemically modified double-stranded RNA molecules designed to mimic endogenous mature miRNAs. The negative control was composed of random sequence Pre-miR™ molecules validated to not produce identifiable effects on known mRNAs.

After 48 hours, transfection media was removed and 100 µl of incomplete DMEM were added to each well. In each well, 100 µl of LightSwitch Assay Solution containing lysis buffer and luciferase substrate were added. After shaking, the plate was incubated wrapped with tin foil at R.T. for 30 minutes. Samples were aliquoted in a 96-well plate to make four technical replications of 50 µl each and luminescence was read for 2 seconds in luminometer plate reader (Glomax, Turner Biosystem). Log 2 ratios of targeting/non-targeting miRNA (pre-miR/pre-control or anti-miR/anti-control) luciferase activity for each construct were plotted. Log2 and standard deviation (SD) were calculated for all the normalised values using Microsoft Excel®. Correlations between relative Luciferase activities were compared using one-way ANOVA, with results considered significant if a two-tailed *p*-value was < 0.05.

4.2.2 Cell transfection and RT-qPCR assay

All the transfection experiments were performed as per Chapter 2, Section 2.2.1.4. RT-qPCR assays are described in Chapter 2 Section 2.2.10 and primers and probes are showed in Table 2.2. In Chapter 4, Section 4.3.6, five different endogenous control genes (GAPDH, GUS, ACTB, RPS13 and TBT) were investigated to identify the control gene least affected by experimental conditions and most suitable for real-time PCR normalisation in cDNA samples of interest. Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) endogenous control gene displayed the least amount of variation and was chosen as reference gene.

4.2.3 Thermodynamic Model for the miRNA-Target Interaction

TargetScan 6.2 (http://www.targetscan.org) and MiRanda (http://www.microrna.org) were used to identify the miRNAs predicted to bind MTHFD1L 3′UTR region containing SNP rs7646. These binding prediction tools were also used to identify other possible miR-9 and miR-197 targets among folate related genes. The binding affinity of miR-197 and miR-9 for "A" and "G" alleles of SNP rs7646 was investigated using a parameter-free thermodynamic model (Mathews, Sabina et al. 1999; Wuchty, Fontana et al. 1999). E_{Target} is the energy of dissociated mRNA target region having no interaction with mRNA, $E_{Intermediate}$ represents the energy necessary to make the target region accessible for microRNA binding and $E_{Complex}$ is the energy of the final microRNA-target complex which is paired according to constraints imposed by the seed sequence. All ensemble free energies were computed with RNAfold or RNAcofold (Vienna RNA Package) using partition function folding and default settings. To determine $E_{Intermediate}$ free energy, binding parameters were modified in a way to leave the target nucleotides (and one additional nucleotide upstream and downstream) unpaired (Kertesz *et al.* 2007, Zhang *et al.* 2011).

The target size used for computational analysis had a length of 80 nucleotides consisting of the seed sequences and flanking regions. The analysis was repeated increasing the target size up to 160 nucleotides and results did not change. The activation and binding energies were calculated to estimate the binding affinity. The activation energy, $\Delta E(a)$, is the difference between $E_{Intermediate}$ and $E_{Target,}$ while the binding energy, $\Delta E(b)$, is the difference between $E_{Complex}$ and E_{Target} (Kertesz *et al.* 2007, Zhang *et al.* 2011).

4.3 Results

4.3.1 SNP rs7646 (A/G) overlaps miR-9 and miR-197 predicted binding sites in the MTHFD1L 3′UTR

MTHFD1L 3′UTR SNP rs7646 (A/G) has been previously identified as the only polymorphism (amongst those that were tested in the 3′UTR region) significantly

associated with Neural Tube Defects (NTD) in an Irish cohort (Parle-McDermott *et al.* 2009). SNP rs7646 is a common variant with a frequency of approximately 69% for the A allele and 31% for the G allele in a Caucasian population (HapMap Genome Browser release #28). Mother control logistic regression showed an association between the G allele and NTD risk (OR = 1.5 [1.09-2.06], p=0.0127) (Parle-McDermott *et al.* 2009). Here, using the miRNA target prediction tools, MiRanda and TargetScan, miR-9 and miR-197 were predicted to bind to the MTHFD1L 3′UTR, in the region that also contains SNP rs7646 (Figure 4.3a). miR-9 is the only miRNA predicted by TargetScan to target a 'conserved site for miRNA families broadly conserved among vertebrates' within the MTHFD1L 3′UTR. Although it is identified as a 'poorly conserved site for miRNA families conserved among mammals or vertebrates', miR- 197 is also predicted to target MTHFD1L 3′UTR (TargetScan 6.2) at the bases 121-127 (7mer-1A). The SNP (rs7646) is located adjacent to the 5′ end of the miR-197 seed sequence and 5′ end of the miR-9 binding site (Figure 4.3a).

4.3.2 Minimum Free Energy (MFE) of miR-9 and miR-197 binding to MTHFD1L 3′UTR differs between allelic variants of SNP rs7646 (A/G)

To analyse the binding affinity of miR-9 and miR-197 to *MTHFD1L* 3′ UTR with different genotypes (A or G alleles of SNP rs7646), we generated a parameter-free thermodynamic model (Mathews *et al*. 1999, Wuchty *et al.* 1999). Secondary structure-based free energies of molecules were calculated for each stage of the binding process: E_{Target} represents the energy of dissociated mRNA target region, $E_{Intermediate}$ is the energy necessary to make the target region accessible for miRNA binding and $E_{Complex}$ is the energy of miRNA-target complex. The activation energy, ΔE(a), is represented by the difference between $E_{Intermediate}$ and E_{Target}, while the binding energy, ΔE(b), is the difference between $E_{Complex}$ and E_{Target} (Kertesz, *et al*. 2007; Zhang *et al.* 2011). The miR-9 binding diagram displayed mild differences between SNP rs7646 genotypes, whereby both ΔE (a) and ΔE (b) of the A allele were slightly lower compared to G allele free energies (Figure 4.4a, Table 4.1). In the final microRNA-target complex, SNP rs7646 resides in a loop region, not part of the complementary region (Figure 4.4a, Table 4.1). This prediction suggests that the A allele is unlikely to have a major effect on the miRNA binding. However, the predicted structure highlights that miR- 197 gains an extra match in the presence of G allele (changing from 7mer-1A to 8mer

match) decreasing the binding energy, ΔE (b) from -8.05 kcal/mol (A allele) to -10.79 kcal/mol (G allele) (Figure 4.4b, Table 4.1). This analysis thereby suggests that the G allele generates a stronger binding of miR-197 on the MTHFD1L 3′UTR compared to the A allele.

4.3.3 Allelic variants show no significant difference in the miR-9 directed suppression of MTHFD1L

Previously, it has been demonstrated that miR-9 decreased the expression levels of MTHFD1L at both mRNA and protein levels in MCF-7 cells (Selcuklu *et al.* 2012). Ablation of the miR-9 seed region confirmed the direct binding of miR-9 to the MTHFD1L 3′UTR (Selcuklu *et al.* 2012). In this study, we have confirmed the direct targeting and suppression of MTHFD1L by miR-9 in both MCF-7 and HEK293 cell lines (Figure 4.5ab). To test whether the variants of the SNP rs7646 (A/G) affect miR-9 regulation of MTHFD1L levels, we generated MTHFD1L 3′UTR reporter constructs containing each of the variants (A/G) or mutated miR-9 seed region (Figure 4.3b). These constructs are designated as UTR-A, UTR-G and UTRmut9, respectively. All three constructs were analysed for luciferase activity levels in MCF-7 (Figure 4.5a) and HEK293 (Figure 4.5b) cells with overexpression of miR-9 (pre-9) or a control miRNA (pre-ctr) that does not target any known mRNA. The luciferase analysis demonstrated that the expression of 3′ UTR constructs (UTR-A and UTR-G) were suppressed by miR-9, whilst mutation of the miR-9 seed region (UTRmut9) recovered the luciferase levels (Figure 4.5ab), indicating direct binding of miR-9 to this predicted site in the MTHFD1L 3′UTR. Comparison of the MTHFD1L 3′UTR construct that contains the A allele (UTR-A) with the construct with the G allele (UTR-G) suggested that although both could be suppressed by miR-9, there was no significant difference in the luciferase expression levels (Figure 4.5ab). This indicates that miR-9 binding affinity to MTHFD1L 3′UTR is not directly affected by the polymorphism (SNP rs7646) in a manner that affects MTHFD1L expression levels.

4.3.4 MiR-197 directly targets the MTHFD1L 3′UTR with enhanced suppression observed for the G allele

To test whether miR-197 directly binds to the putative site on MTHFD1L 3′UTR and whether the presence of either of the allelic variants of the SNP rs7646 (A/G) affects binding efficacy; luciferase assays were performed in both MCF-7 and HEK293 cells using the MTHFD1L 3′UTR reporter constructs for either of the SNP variants (UTR-A and UTR-G) and also the construct with mutated miR-197 binding site (UTRmut197) (Figure 4.3b). Luciferase activity was suppressed in the wild type construct (UTR-A). However, ablation of the seed region (UTRmut197) recovered the luciferase activity, indicating that miR-197 directly binds to this predicted region in MTHFD1L 3′UTR in a manner that downregulates expression (Figure 4.5cd). In addition, presence of the G allele (UTR-G) displayed a greater level of suppression of the luciferase activity compared to UTR-A, indicating that miR-197 has stronger miRNA:target binding efficacy and therefore stronger suppression effects on the MTHFD1L 3′UTR variant containing the G allele (Figure 4.5cd).

To determine the effect of miR-197 on the expression levels of endogenous MTHFD1L, miR-197 was overexpressed in MCF-7 and HEK293 cell lines. RT-qPCR analysis showed that, in both cell lines, miR-197 downregulated the expression levels of MTHFD1L (Figure 4.5cd), confirming the downregulation effect observed in the luciferase reporter construct assays. In addition, Western blot analysis demonstrated that miR-197 expression downregulated MTHFD1L at the protein level (Figure 4.5cd). Taken together, these results confirm that miR-197 directly regulates *MTHFD1L* gene expression levels at the mRNA and protein levels and that SNP rs7646 has a functional impact on miR-197 binding efficacy.

4.3.5 miR-9 and miR-197 display antagonistic target suppression

To understand the relative efficiencies of miR-9 or miR-197 in the regulation of MTHFD1L, the endogenous expression of miR-9, miR-197 and *MTHFD1L* were measured in MCF-7 and HEK293 cells. The HEK293 cells displayed a higher expression level of miR-9 and miR-197 compared to MCF-7 cells (Figure 4.6). In contrast, MTHFD1L expression was higher in MCF-7 cells compared to HEK293 cells, negatively correlating with miR-9 and miR-197 expression as expected. The endogenous expression levels of miR-197 is several orders of magnitude greater than

miR-9 in both cell lines (Figure 4.7). miR-9 and miR-197 binding sites partially overlap in the MTHFD1L 3′UTR (Figure 4.3a).

To investigate the possibility that endogenous miR-9 and miR-197 may affect each other's binding and consequently the target expression, we performed further luciferase assays in MCF-7 and HEK293 cells with exogenously added miRNA (pre-9 or pre-197) in the presence or absence of miRNA inhibitors to deplete endogenous miR-9 or miR-197. Both cell lines displayed a further pre-9 mediated decrease in luciferase activity when endogenous miR-197 was depleted by anti-197 for either MTHFD1L 3′ UTR allele (Figure 4.8). A similar experiment was performed with miR-197 overexpression (pre-197) in the presence and absence of anti-9 and showed no statistically significant difference for the A allele in either cell line (Figure 4.9ab). On the other hand, the 3′ UTR construct with the G allele showed an opposite effect in miR-9 depleted cells (anti-9); a slightly increased luciferase activity (*p*-value =0.003) in MCF-7 (Figure 4.9c) and a decreased luciferase activity in HEK293 cells (Figure 4.9d).

4.3.6 miR-9 and miR-197 may target other folate genes

Computational analysis was carried out using miRANDA and TargetScan, to predict other possible folate genes regulated by these miRNAs. As a result, miR-9 was predicted by both tools to bind *DHFR*, *SHMT1*, methionine synthase (*MTR*) and *DHFR1L* genes (Chapter 1, Section 1.1.4). The first three genes encode cytoplasmic proteins while the latter encodes a mitochondrial protein. miR-9 binding sequence in *DHFR* and *DHFRL1* is conserved among primates, whereas miR-9 binding sequence in SHMT1 is conserved in primates, rabbits, armadillos and opossums. miR-9 binding sequence in MTR is highly conserved among vertebrates. On the other hand, miR-197 analysis showed that, along with MTHFD1L, *TYMS* is another possible target gene. miR-197 binding sequence in *TYMS* 3′UTR is conserved among humans, macaques and armadillos. Transient transfection experiments in HEK293 and MCF-7 cell lines were performed to verify *in silico* predictions. Transfection with miR-9 in HEK293 cells showed that all the predicted genes are significantly downregulated (Figure 4.10a) while in MCF-7 only MTR showed a significant downregulation (Figure 4.10b). Instead, miR-197 transfection inhibited TYMS expression in both the cell lines (Figure 4.11ab).

4.4 Discussion

Polymorphisms in miRNA target binding sites can affect the binding efficacy of miRNAs. Consequently, altered miRNA binding to the target can potentially alter the expression levels of the target gene, and can contribute to risk factors for the development of certain diseases. Indeed, the association of genetic polymorphisms that alter miRNA binding with disease risk has been reported in a number of studies (Haas *et al.* 2012, Mishra *et al.* 2007, Song *et al.* 2009, Nicoloso *et al*. 2010, Zhang *et al.* 2011). The patho-mechanistic role of miRNAs is increasingly acknowledged (Haas *et al*. 2012). It has also been shown that SNPs are less frequent in miRNAs or their target sites than in other regions of the genome, suggesting some level of purifying selection against mutations in miRNA:target gene binding systems (Chen and Rajewsky 2006, Saunders *et al.* 2007; Yu *et al.* 2007). This negative selection of sequence variations emphasises the crucial role of miRNA regulation in pivotal cellular processes.

miR-9 has been reported as an important regulator in neural development (Yuva-Aydemir *et al.* 2011). In vertebrates, miR-9 modulates the proliferation, migration and differentiation of neural progenitor cells by targeting diverse mRNAs (Yuva-Aydemir *et al.* 2011). Moreover, some evidence suggests that miR-9 expression in post-mitotic neurons plays a role in neurodegenerative diseases (Yuva-Aydemir *et al.* 2011). Altered miR-9 levels have been found in postmortem Alzheimer disease hippocampus and hippocampal cultures (Lukiw 2007, Schonrock *et al.* 2012). Although most current evidence indicates tumour suppressor activity of miR-9 (Laios *et al.* 2008, Lujambio *et al.* 2008) conflicting findings exist (Khew-Goodall and Goodall 2010) and the role of this miRNA in cancer biology is not well understood. Thus, the importance of miR-9 in development and disease is emerging through its ability to regulate different targets in a manner that can be dependent on the cellular context and the developmental stage.

In this NTD focused miRNA study, we identified a novel direct interaction between the MTHFD1L 3′ UTR and miR-197 in MCF-7 and HEK293 cell lines. This is the second miRNA to be identified as a regulator of MTHFD1L, along with miR-9 (Selcuklu *et al.* 2012). Our results demonstrate that both miRNAs specifically downregulate

MTHFD1L at mRNA and protein levels in both cell lines. Using computational modelling, we predicted secondary structures of miR-197 and miR-9 binding to the 3′ UTR containing either an A or G allele of SNP rs7646. We found that the G allele significantly favours miR-197 binding, providing an extra match just beside the seed sequence and contributing to the pairing stability. On the other hand, SNP rs7646 locates within an unmatched loop of the predicted secondary structure and, according to free energy diagrams (*in silico*); it should not affect miR-9 binding. These findings were verified in vitro using the luciferase reporter assays. Our results demonstrate that SNPrs7646 significantly changed miR-197 binding affinity, causing stronger suppression when bound to the G allele of *MTHFD1L* mRNA compared to the A allele, while it did not generate any significant variation for miR-9 binding affinity. miR-9 and miR-197 binding sites partially overlap in the MTHFD1L 3′UTR.

We consider that the two miRNAs are unlikely to simultaneously bind to the 3′UTR due to secondary structure constraints. According to this hypothesis, depleting miR-197 from the cells should leave more *MTHFD1L* mRNAs available for miR-9 and vice-versa. We tested the effects of altered expression levels of miR-9 and miR-197 using miRNA precursors (premiR) and inhibitors (anti-miR). miR-9 overexpression (pre-9) led to more downregulation of the target expression when miR-197 was depleted (anti-197) in all the conditions tested (Figure 4.8), while miR-197 overexpression (pre-197) in combination with miR-9 depletion (anti-9) led to differential outcomes depending on the genotype and the cell line (Figure 4.9). Overall, these data indicate that there is competition between miR-9 and miR-197 for MTHFD1L 3′ UTR binding. However, the differential regulatory effect observed depends on which endogenous miRNA is downregulated, whereby cell type and genotype could affect miRNA binding efficiency and endogenous miRNA abundance. The miR-9 binding site within the 3′ UTR of *MTHFD1L* is a more highly conserved site than the miR-197 site. It is possible that, although miR-197 is much more abundant in both cell lines compared to miR-9 (Figure 4.7), miR-9 may preferentially bind to the 3′UTR of *MTHFD1L* even in the presence of excess miR-197. In other words, miR-9 is a more specific, efficient regulator of *MTHFD1L* and miR-197 can be outcompeted by miR-9 regardless of abundance. miR-197 may increase its binding efficiency in the presence of G allele of the rs7647 SNP as we already showed that this allele appears to increase miR197 binding efficacy (Figure 4.5cd). This can explain the differences we observe based on cell type and *MTHFD1L*

genotype. MCF-7 cells express endogenous miR-9 at a 3-fold lower level than HEK293 cells (Figure 4.6). The increase in MTHFD1L expression in the presence of anti-9 in MCF-7 cells (Figure 4.9ac) indicates that the endogenous miR-9 is inhibited and exogenous miR-197 cannot compensate for this. In HEK293 cells endogenous miR-9 is expressed at a higher level. The A allele shows no change in the presence of anti-9 (Figure 4.9b), suggesting that exogenous miR-197 cannot successfully compete with any residual endogenous miR-9 that has escaped anti-9 inhibition. Further downregulation is observed for the G allele in the presence of anti-9 (Figure 4.9d), as it provides a better binding site for miR-197 and can successfully compete with any residual endogenous miR-9. We also cannot rule out that there are differences in the mechanism of action of cellular miRNAs compared to precursor mimic miRNAs, as a previous study suggested (Haas *et al.* 2012). Our results can suggest that pre-9 and pre-197 directed downregulation of MTHFD1L is affected by the competitive binding behaviour of both miRNAs. Therefore, the overall impact of miRNA regulation of MTHFD1L must be considered in the context of both of these MTHFD1L targeting miRNAs.

Previously, it has been demonstrated that another folate related gene, *MTHFD2*, is downregulated by miR-9 (Selcuklu *et al.* 2012). Like MTHFD1L, MTHFD2 localizes to mitochondria and the two enzymes together contribute to the synthesis of formate that is the main source of 1C-units in the folate pathway. Moreover, our data suggested that miR-9 may target other folate genes including DHFR, DHFRL1, SHMT1 and MTR. Transfection with miR-9 in HEK293 cells showed these four genes are significantly downregulated (Figure 4.10a) while in MCF-7 only MTR showed a significant downregulation (Figure 4.10b). The differences between cell types may be related to the breast cancer phenotype of MCF-7 versus the embryonic kidney cell phenotype of HEK293 cells. Though these data require further validation, they provide a hint of a possible broader regulative role of miR-9 for folate genes. This possible role of miR-9 along with its involvement in neurogenesis further highlights that this miRNA warrants further investigation in relation to NTD and other neuro-developmental diseases. Although our results suggest that the NTD associated SNP rs7646 does not influence the miR-9 mediated regulation of MTHFD1L, our data have relevance for how miR-9 and miR-197 compete.

In the context of functional effects in neurological tissues, miR-197 is commonly upregulated in Alzheimer Disease, brain parenchyma and cerebrospinal fluid (Maes *et al.* 2009). In this study, we have identified and validated MTHFD1L as a direct miR-197 target and demonstrated that the maternal NTD-associated SNP rs7646 affects the miR-197 binding efficacy. The G allele, associated with an increased risk, provides more efficient binding that alters the expression level of MTHFD1L. Recent reports indicate that mthfd1l knockout mice develop NTD with 100% penetrance (Momb *et al.* 2013). Thus, the impact of SNP rs7646 on the binding of miR-197 to the 3′ UTR of *MTHFD1L* provides a mechanistic explanation of the previous association of this SNP with NTD risk (Parle-McDermott *et al.* 2009). Studying the expression pattern of both miR-197 and miR-9 during embryogenesis will improve our understanding of the miRNA-mediated control of MTHFD1L and its relevance to NTD risk. Apart from NTD risk, it is also worth noting the relevance of miR-197 in cancer. miR-197 has been reported to be upregulated in human male breast cancer (Lehmann *et al.* 2010), lung cancer (Yanaihara *et al.* 2006, Zheng *et al.* 2011), and cervical carcinoma (Pereira *et al.* 2010) and it has been shown to modulate the expression of the tumour suppressor gene FUS1 whose expression is lost in a large proportion of lung tumours (Du *et al.* 2009). We previously have showed that an increase in MTHFD1L enzyme levels through a separate associated genetic variant (DIP polymorphism) increases the case risk of NTD (Parle-McDermott *et al.* 2009). The DIP polymorphism within intron 7 of *MTHFD1L* impacts on alternative splicing (Parle-McDermott *et al.* 2009) while SNP rs7646 within the 3′ UTR appears to influence miRNA binding. Both studies highlight that it is the level of MTHFD1L enzyme that is relevant for disease as both an increase (Parle-McDermott *et al.* 2009) or a decrease in MTHFD1L (this study) can be associated with increasing NTD case and maternal risk respectively. This suggests that maintaining MTHFD1L levels within an expression level “window” during development may be a critical factor for reducing NTD risk.

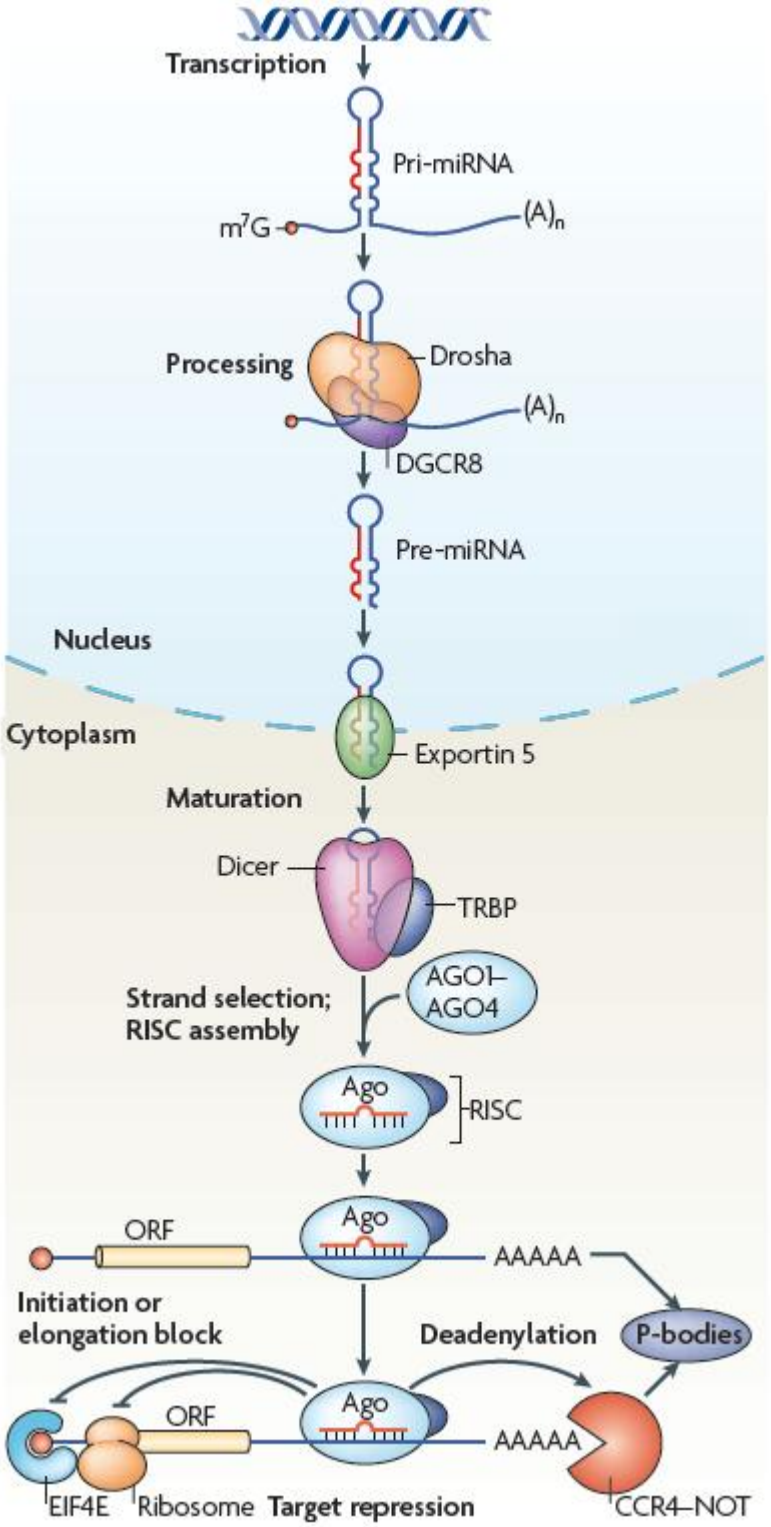


Figure 4.1 miRNA biogenesis and mechanisms of action (modified from Inui *et al*. 2010).

pri-miRNAs are transcribed in the nucleus and processed by the microprocessor complex by the RNase III type endonuclease Drosha and its partner, DGCR8. It

produces a 70-nucleotide stem loop called pre-miRNA, which is exported to the cytoplasm by exportine 5. Once in the cytoplasm, the pre-miRNAs are cleaved by Dicer and TRBP, generating a 20-nucleotide mature miRNA duplex. At this stage, one strand of miRNA is usually degraded and the other strand is loaded into the RISC which contains Ago proteins. The RISC is guided by the single-stranded miRNA to its final target mRNAs partially complementing its sequence. The RISC can repress the target mRNA through two main mechanisms. It can remove polyA tail by promoting the activity of deadenylases (such as CCR4-NOT), causing mRNA degradation in P-bodies. Alternatively it can block the translation at the initiation step or at the elongation step by causing, e.g., the ribosome stalling or the repression of initiation factor 4E (EIF4E).

pri-miRNAs, primary transcripts; DGCR8, DiGeorge syndrome critical region 8; pre-miRNA, precursor miRNA; TRBP, TAR RNA-binding protein, RISC, RNA-induced silencing complex, Ago, Argonaute; P-bodies, Processing bodies; EIF4E, repression of initiation factor 4E.

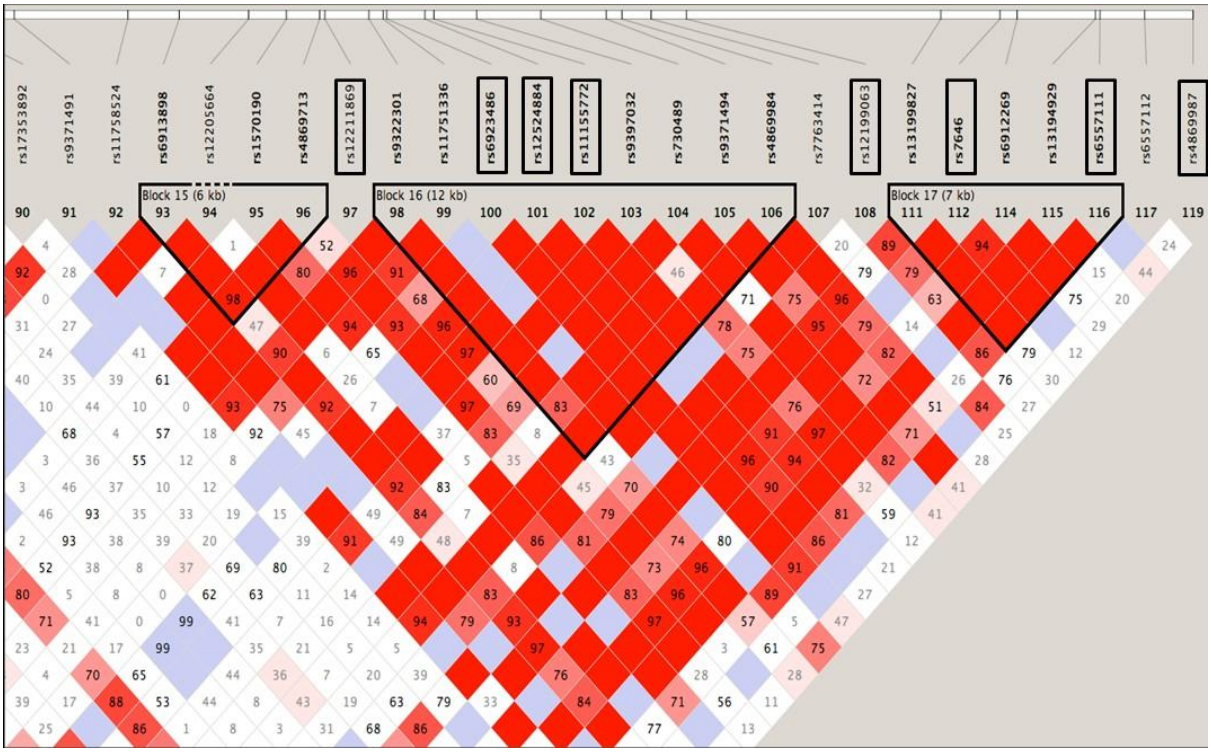


Figure 4.2 Linkage disequilibrium (LD) map of one associated region of the *MTHFD1L* Gene (marked region ‘‘b’’ in Chapter 1, Figure 1.6).

LD plot of pairwise values of D’ for the 26 markers within the region comprised between intron 26 and *MTHFD1L* 3′end. Boxed markers were associated with NTD risk. SNP rs7646 is within the 3′UTR of *MTHFD1L*.

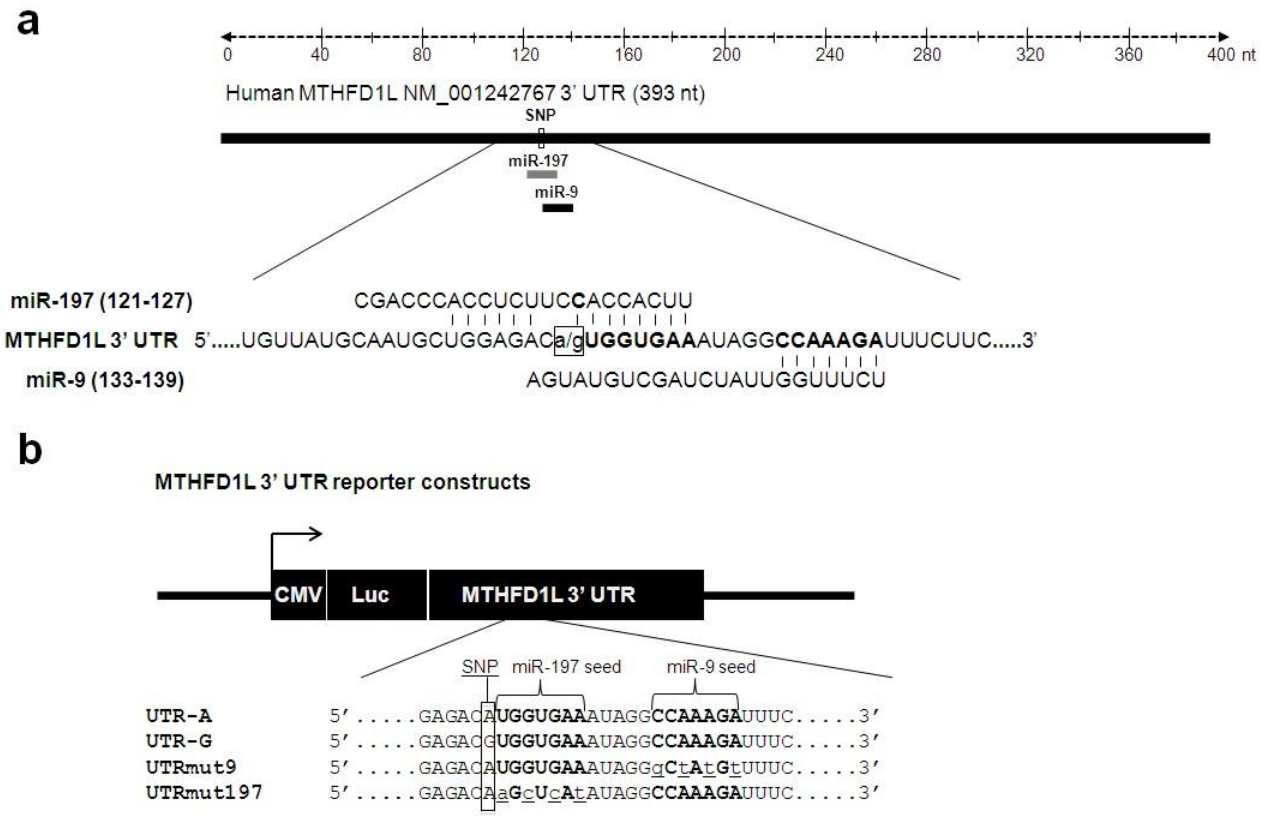


Figure 4.3 Schematic depiction of *MTHFD1L* 3′UTR and miR-9 / miR-197 binding regions (figure drawn with Microsoft PowerPoint).

(a) *MTHFD1L* 3′UTR schematic representation according to NCBI Reference Sequence NM_001242767. Numbering starts from the first nucleotide of the 3′UTR, thus, SNPrs7646 is present in position 120. Seed sequences of miR-197 and miR-9 are in position 121-127 and 133-139, respectively. **(b)** Schematic depiction of the constructs used for luciferase analysis. Sequences underneath show nucleotides that were mutated to generate SNPrs7646 G allele 3′UTR (UTR-G) and negative controls for miR-9 (UTRmut9) and miR-197 (UTRmut197).

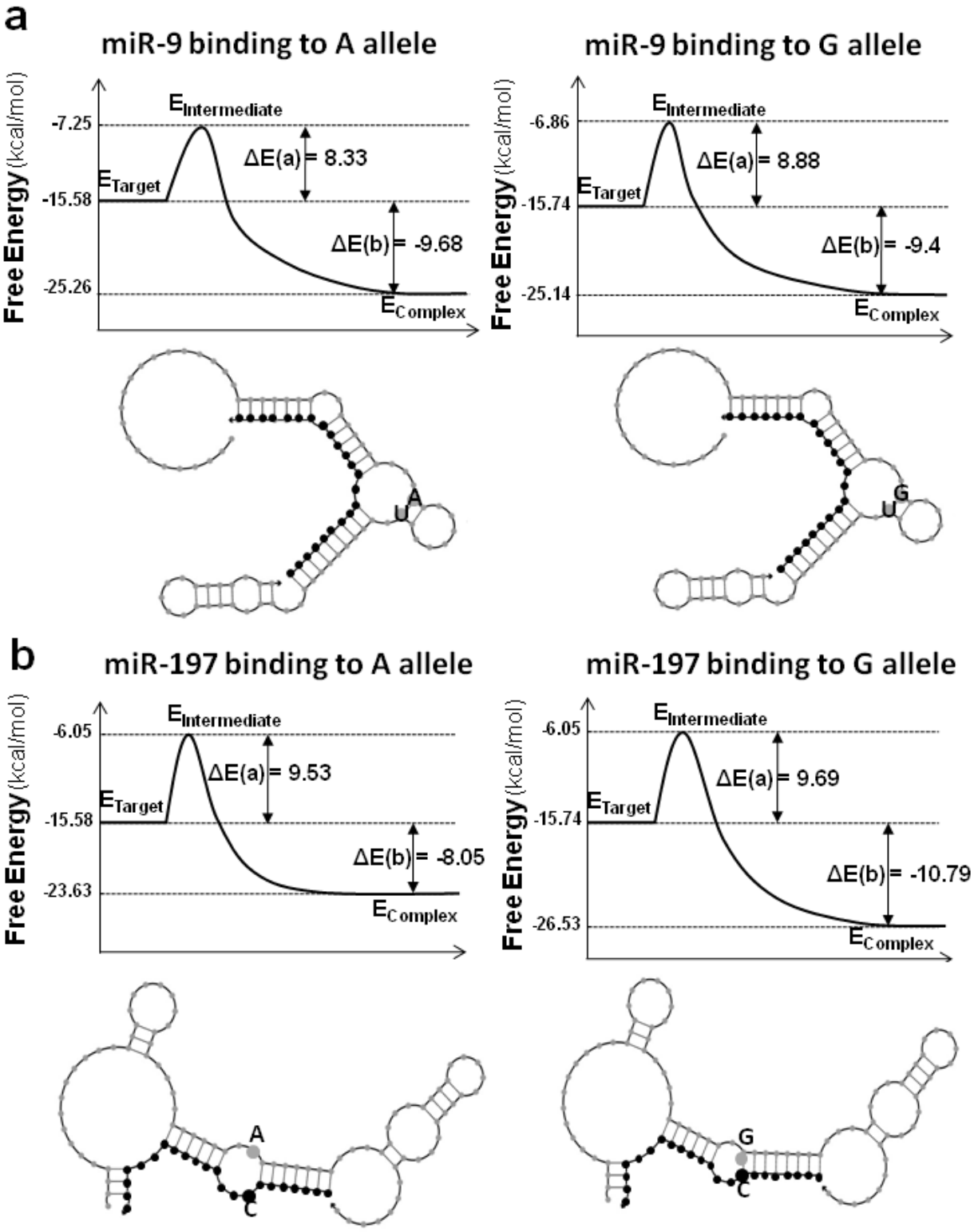


Figure 4.4 Thermodynamic model of miR-9 and miR-197 binding to *MTHFD1L* 3′UTR and the influence of SNPrs7646 (figure drawn with Microsoft PowerPoint).

Secondary structure prediction of **(a)** miR-9 and **(b)** miR-197, binding to *MTHFD1L* 3′UTR harbouring A or G allele and the relative free energy diagrams are shown. Secondary structure-based free energies at different stages of the binding process represented by E_{Target}; the energy of dissociated mRNA target region, $E_{Intermediate}$; the energy necessary to make the target region accessible for miRNA binding, and $E_{Complex}$; the energy of miRNA-target complex. The activation energy, ΔE (a), is given by the difference between $E_{Intermediate}$ and E_{Target}, and the binding energy, ΔE (b), is the difference between $E_{Complex}$ and E_{Target}.

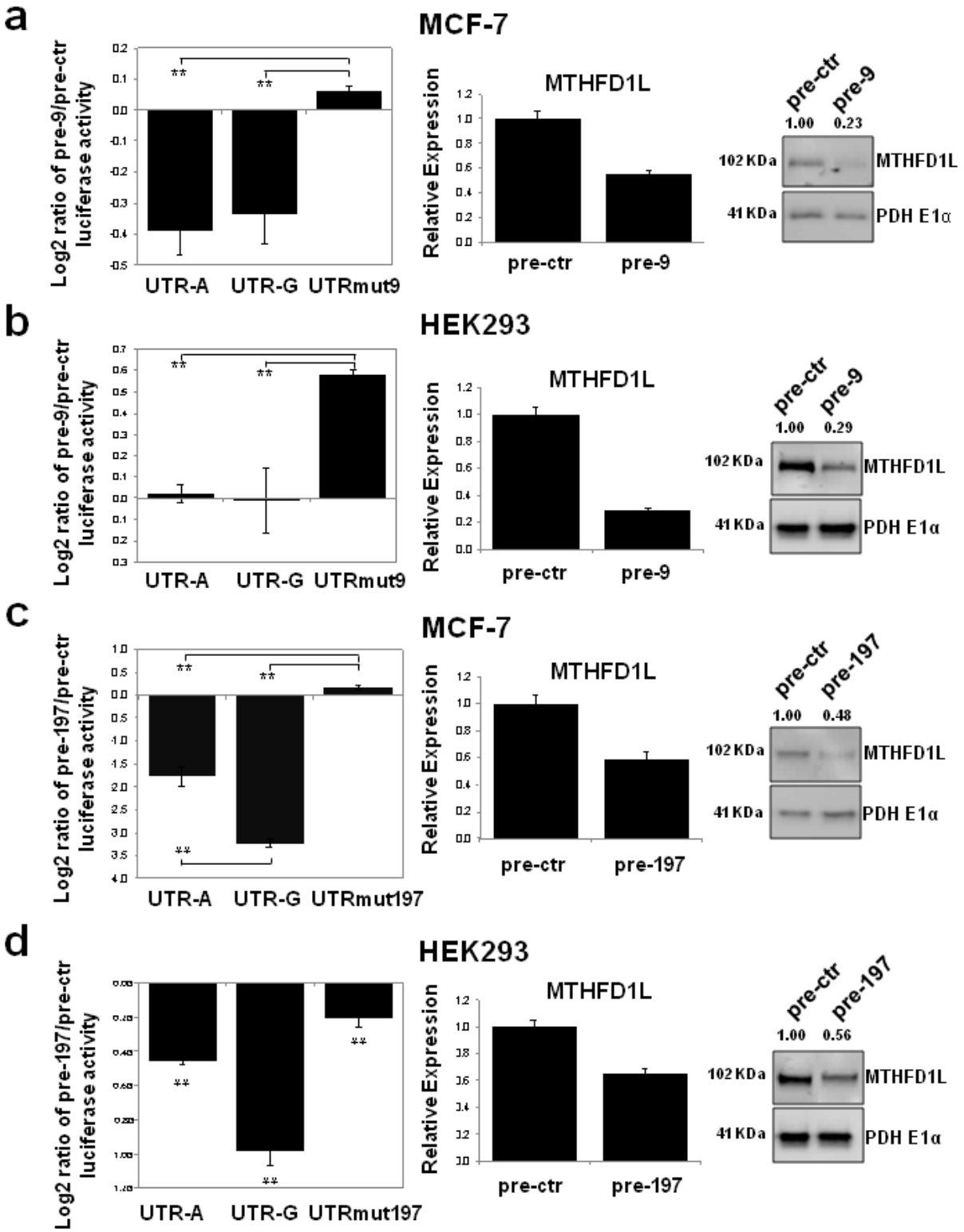


Figure 4.5 MiR-9 and miR-197 targeting effects on MTHFD1LwithSNPrs7646 variants (A or G allele).

Data are presented as luciferase assays, RT-qPCR and Western blot for miR-9 in **(a)** MCF-7 and **(b)** HEK293cells and for miR-197 in **(c)** MCF-7 and **(d)** HEK293 cells. Luciferase analysis compared *MTHFD1L* 3′UTR A allele (UTR-A), G allele (UTR-G) and the mutant controls (UTRmut9 or UTRmut197). Data are represented as Log2 ratio of targeting to non-targeting miRNA luciferase activity (pre-miR to pre-ctr ratio). RT-qPCR analysis of *MTHFD1L* is represented as relative expression in pre-9, pre-197 and pre-ctr transfected cells. All data were analysed by the E (Efficiency)-Method using GUS as an endogenous control for normalization. Statistical significance was calculated using Student's t-test (*$p<0.05$, **$p<0.001$). The Western blot shows MTHFD1L protein in pre-9, pre-197 and pre-ctr transfected cells. Numbers represent relative band intensities measured by densitometry analysis normalized using PDH E1α as an endogenous control for mitochondrial proteins.

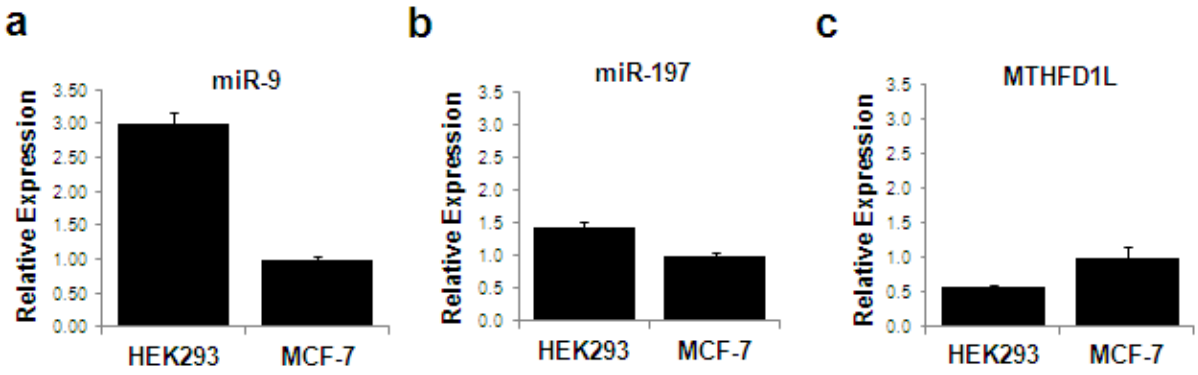


Figure 4.6 Endogenous levels of miR-9, miR197 and *MTHFD1L* in HEK293 and MCF-7 cells.

RT-qPCR analysis of endogenous **(a)** miR-9, **(b)** miR197 and, **(c)** *MTHFD1L* in HEK293 and MCF-7 cells are shown. RNU6B was used to normalize miRNAs expression and GUS was used to normalize MTHFD1L expression. MCF-7 values were used as the calibrator.

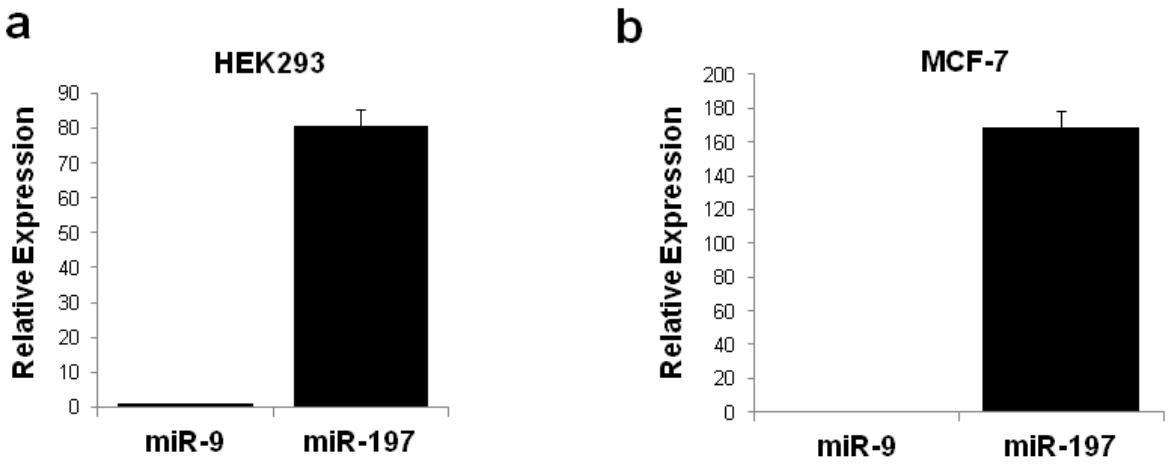


Figure 4.7 Endogenous levels of miR-9 compared to miR-197 in HEK293 and MCF-7 cells.

RT-qPCR analysis of endogenous miR-9, miR-197 in HEK293 **(a)** and MCF-7 **(b)** cells are shown. RNU6B was used to normalize miRNAs expression and miR-9 is the calibrator.

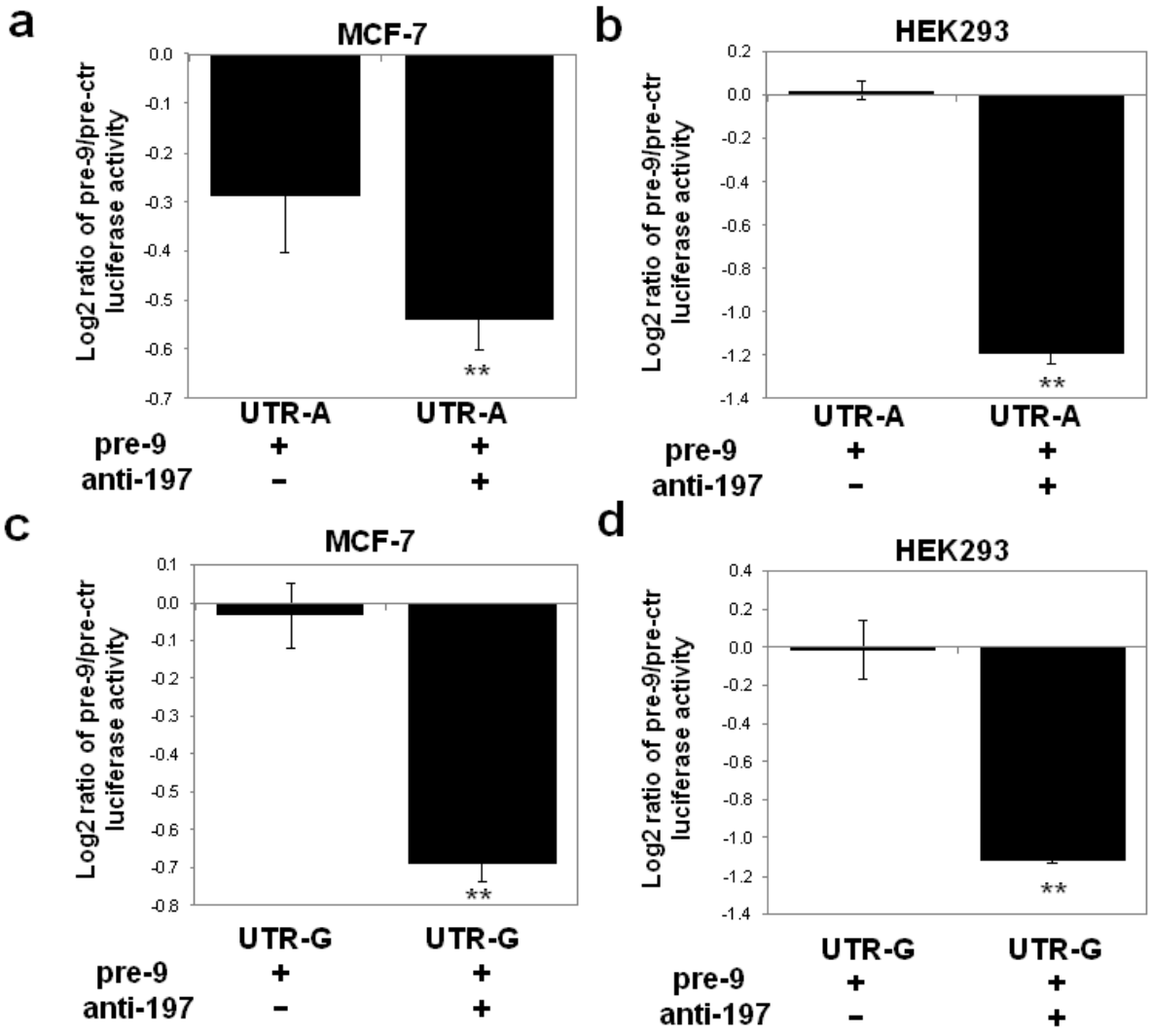


Figure 4.8 Effects of miR-9 overexpression and miR-197 knockdown on *MTHFD1L 3′* UTR with SNPrs7646 variants (A or G allele).

The data show luciferase analysis of **(a)** UTR-A in MCF-7, **(b)** UTR-A in HEK293 cells, **(c)** UTR-G in MCF-7 and **(d)** UTR-G in HEK293 cells with miR-9 overexpression (pre-9) and endogenous miR-197 or miR-197 knockdown (anti-197). Data are represented as Log2 ratio of targeting to non-targeting miRNA luciferase activity. Statistical significance was calculated using Student's t-test (*$p<0.05$, **$p<0.001$).

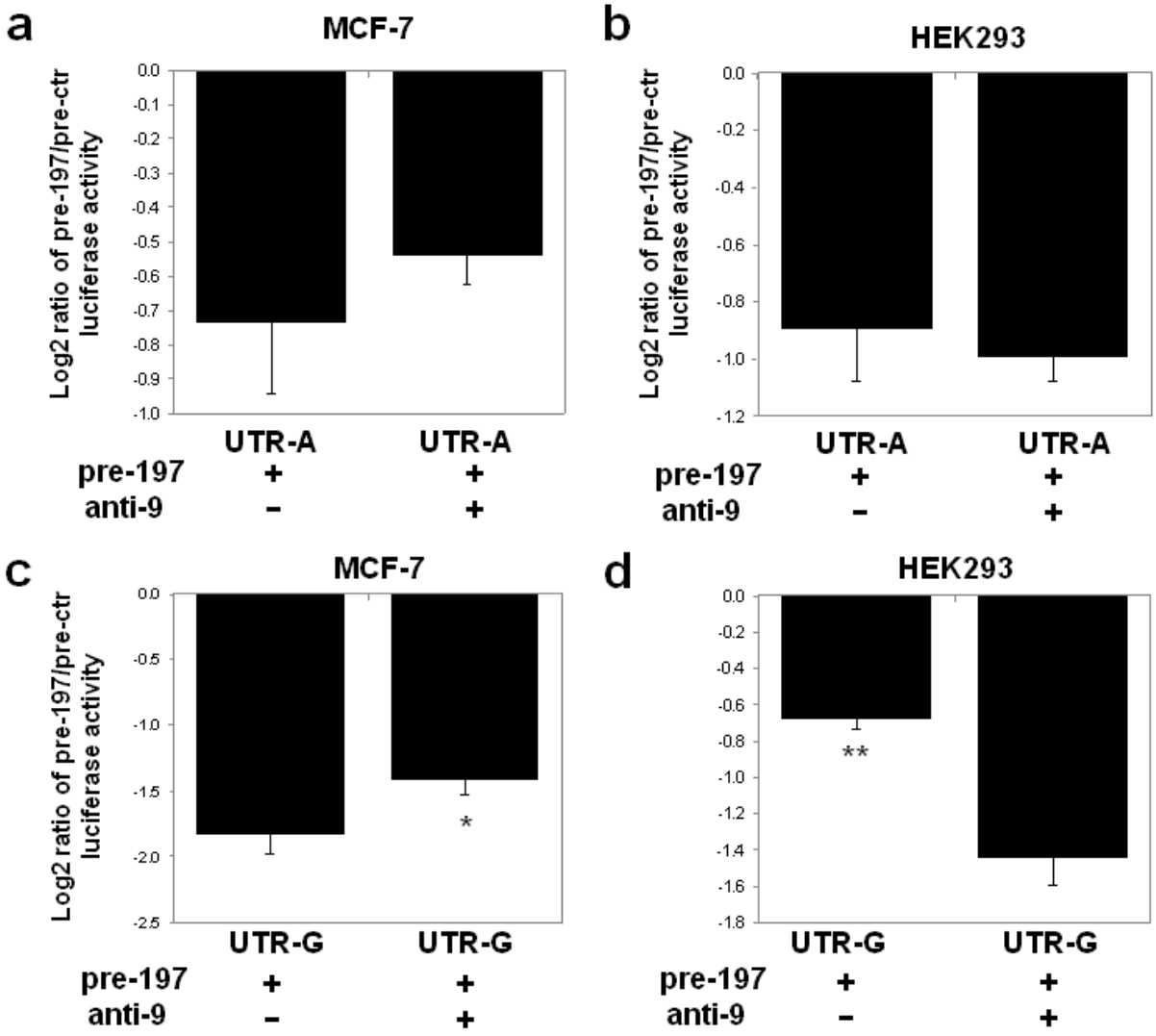


Figure 4.9 Effects of miR-197 overexpression and miR-9 knockdown on *MTHFD1L 3′* UTR withSNPrs7646 variants (A or G allele).

Data show luciferase analysis of **(a)** UTR-A in MCF-7, **(b)** UTR-A in HEK293 cells, **(c)** UTR-G in MCF-7 and **(d)** UTR-G in HEK293 cells with miR-197 overexpression (pre-197) and endogenous miR-9 or miR-9 knockdown (anti-9). Data are represented as Log2 ratio of targeting to non-targeting miRNA luciferase activity. Statistical significance was calculated using Student's t-test (*$p<0.05$, **$p<0.001$).

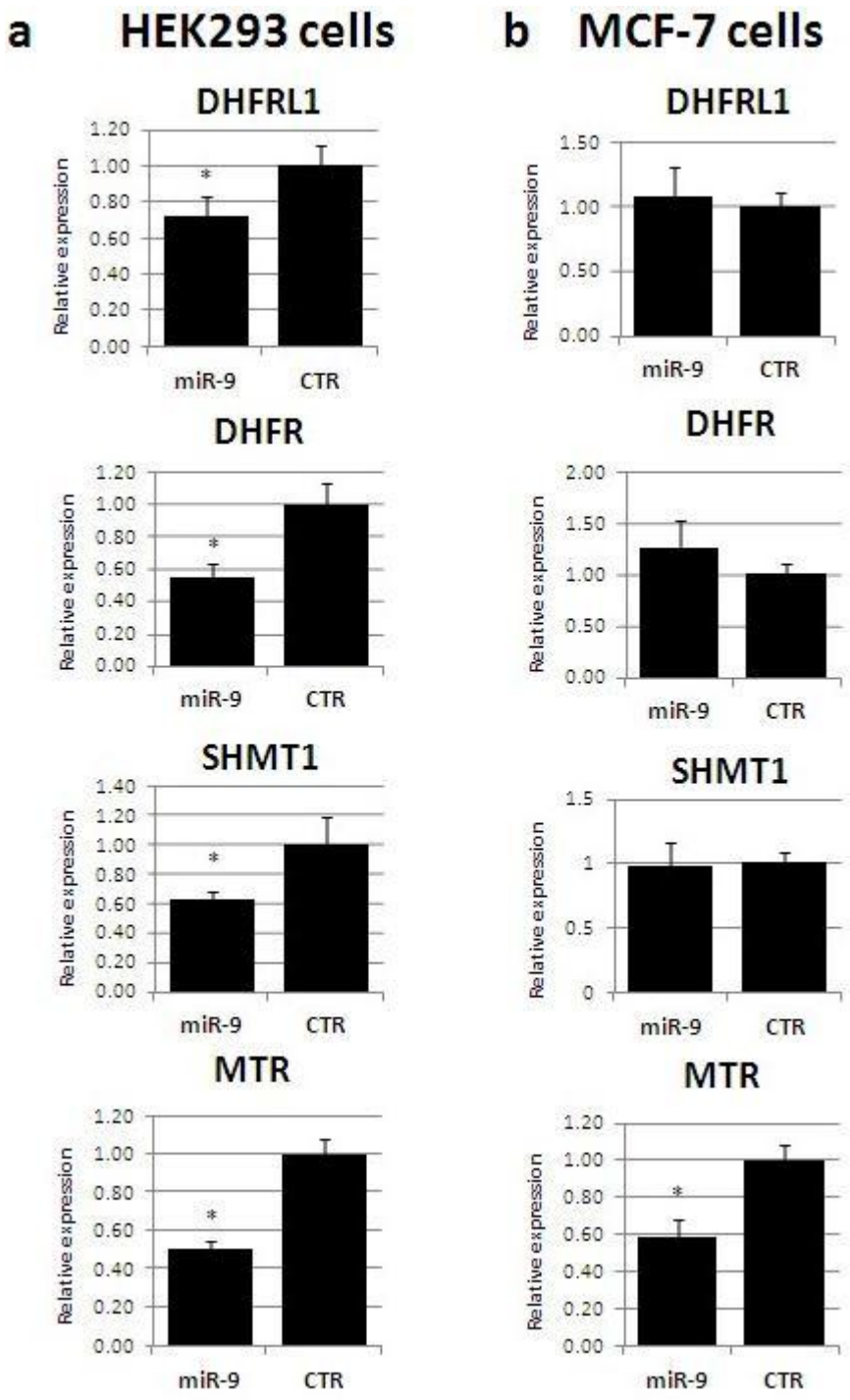


Figure 4.10 Gene expression after transfection with miR-9 precursor.

RT-qPCR analysis of DHFRL1, DHFR, SHMT1 and MTR in HEK293 **(a)** and MCF-7 **(b)** cells. Data are represented as relative expression in pre-9, pre-197 and pre-ctr (CTR) transfected cells. All data were analysed by the E (Efficiency)-Method using GAPDH as an endogenous control for normalization. Statistical significance was calculated using Student's t-test (*$p<0.05$, **$p<0.001$).

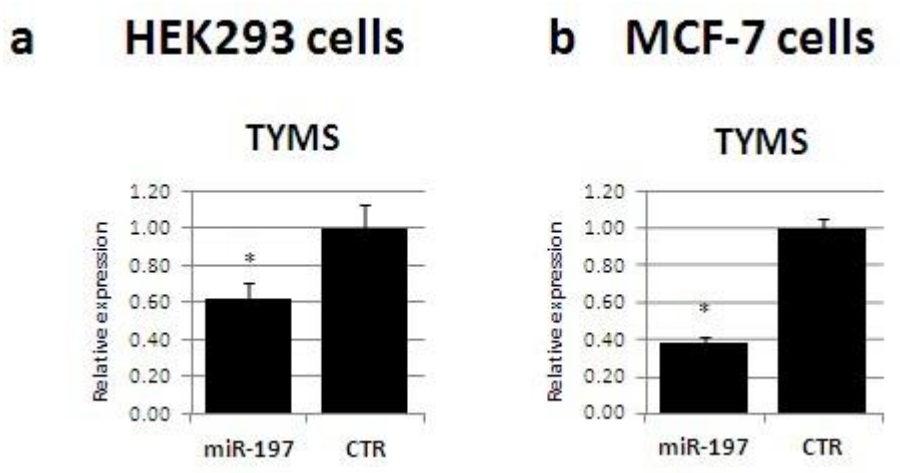


Figure 4.11 Gene expression after transfection with miR-197 precursor.

RT-qPCR analysis of TYMS in HEK293 **(a)** and MCF-7 **(b)** cells. Data are represented as relative expression in pre-9, pre-197 and pre-ctr (CTR) transfected cells. All data were analysed by the E (Efficiency)-Method using GAPDH as an endogenous control for normalization. Statistical significance was calculated using Student's t-test (*$p<0.05$, **$p<0.001$).

Table 4.1 Free Energies of miR-9 and miR-197 binding to *MTHFD1L* 3′UTR with SNP rs7646 alleles (A/G).

		miR-9		miR-197	
Genotype*		**A**	**G**	**A**	**G**
Free Energies	**Target[1]**	-15.58	-15.74	-15.58	-15.74
	Intermediate[2]	-7.25	-6.86	-6.05	-6.05
	Complex[3]	-25.26	-25.14	-23.63	-26.53
	ΔE(a)[4]	8.33	8.88	9.53	9.69
	ΔE(b)[5]	-9.68	-9.4	-8.05	-10.79

* SNP rs7646 alleles (A/G) in the target *MTHFD1L* 3′UTR

[1]Energy of dissociated mRNA target region.

[2]Energy necessary to make the target region accessible for miRNA binding

[3] Energy of miRNA-target complex.

[4] Activation energy is the difference between $E_{Intermediate}$ and E_{Target},

[5] Binding energy is the difference between $E_{Complex}$ and E_{Target}

CHAPTER 5

Investigation of novel MTHFD-like sequences across the human genome

5.1 Introduction

The comparison of genomic sequences within a species or between different species is an important genetic tool, useful for the identification of protein-coding genes, structural motifs, and regulatory elements such as components of promoters and enhancers (Margulies and Birney 2008). Throughout evolution transposable elements have shuffled eukaryotic genomes generating a multitude of new sequences including new genes and pseudogenes (Xing *et al*. 2006). Genome studies have revealed that approximately 50% of the human genome and >70% of genomes of some grass species like maize, consist of transposable element sequences (Wessler 2006).

Pseudogenes are classified as genomic sequences resulting from either retrotransposition or genomic duplication of functional genes. Processed pseudogenes arose from the former event, while non-processed pseudogene derived from the latter and often retain the exon-intron structures of their parental genes (Zhang *et al*. 2003). During their evolutionary history, pseudogenes generally accumulate mutations which cause premature stop codons or frameshifts and consequent loss of functionality. Previous studies have located and annotated human pseudogenes, counting around 20,000 sequences out of which about 8,000 show evidence of processing (Torrents *et al*. 2003). Different studies have also revealed that in mammals most of the genome, including numerous pseudogenes, is transcribed in both sense and antisense direction (Katayama *et al*. 2005, Carninci *et al*. 2005). But less than 2% of the total genome is transcribed into protein-coding RNA, because the large majority of transcripts are noncoding RNAs (ncRNAs) (Cheng and Carmicheal 2010). Over the last decade there has been a near-exponential growth of publications focused on ncRNAs (Taft *et al*. 2010). They have been classified into various categories such as snoRNAs, siRNAs and miRNAs (as discussed in Chapter 4). ncRNAs have many different functions playing an important role in the regulation of a multitude of cellular processes. Recently, it has been demonstrated that expressed pseudogenes and ncRNAs can compete with their ancestral protein coding genes for the same pool of miRNAs (Salmena *et al*. 2011).

The aim of this chapter was to analyze genomic sequences similar to *MTHFD1* and *MTHFD1L* to determine if there is evidence for the expression. Expressed sequences

were then investigated further to discover possible protein coding homologous or regulatory elements such as ncRNAs related to *MTHFD1* and *MTHFD1L*.

5.2 Results

5.2.1 *MTHFD1* and *MTHFD1L* homologous sequence search

Nucleotide BLAST (Human genome database) was used to find sequences similar to *MTHFD1L* and *MTHFD1*. All the sequences analysis and sequence similarity localization was made using mRNA transcript variant 1 of *MTHFD1L* (accession number: NM_001242767) and MTHFD1 mRNA (accession number: NM_005956) using default parameters. Evidence of transcription for the reported DNA sequences was sought in Ref seq and the EST database (Chapter 2, Sections 2.2.14.9-10) using default parameters.

5.2.2 CHR 11 homologue sequence

A sequence on Chromosome 11 (Chr11seq) (location: 11p15.1, 18892534..18895009) is a processed pseudogene which shows about 88% similarity with a region of *MTHFD1L* spanning between exon 21 and 3′UTR (Figure 5.1). This sequence shares 1291 bases with *MTHFD1L* (position 2144-3434) out of which 939 bases come from the coding region. ORF FINDER analysis did not show any clear ORF due to stop codon mutations along the sequence. These findings suggest that no functional protein is encoded from Chr11seq. Since the Chr11seq sequence spans the *MTHFD1L* 3′UTR we investigated a possible expression of a ncRNA. EST and Refseq database analyses (Chapter 2, Sections 2.2.14.9-10) did not report any expression of this sequence; neither did RT-qPCR analysis using placental, HEK293 and lymphoblast RNAs (Chapters 2, Sections 2.2.4 and 2.2.6) (data not shown). *MTHFD1L* and Chr11seq sequence alignment along with the position of the primers used for RT-qPCR are shown in Appendix Z.

5.2.3 CHR2 and CHR9 homologue sequences

5.2.3.1 Sequence analysis

There are non-processed sequences present on both Chr2 (Chr2seq) (location: 2p11.1, 861528..871795) and Chr9 (Chr9seq) (location: 9q12, 663889..679324) derived from a *MTHFD1L* region which includes exon 25-26 and a part of flanking introns (Figure 5.2). Sequence similarity is about 95% with Chr2seq and 94% with Chr9seq, reaching for both 97% in the region corresponding to exon 25-26 (sequence alignment is shown in Appendix BB). An EST database search showed that two transcribed RNAs were reported under NEDO human cDNA sequencing project (Yudate *et al*. 2001). They are reported as ncRNAs because of the lack of a readable ORF. The expressed ncRNAs are spliced sequences, one originating from Chr2seq (AK074198) and the other from Chr9seq (AK097152). Figure 5.3 shows that both ncRNAs share exon 2, 3 and 4 while they have a different exon 1 which derives from an upstream sequence with no similarity to *MTHFD1L*. Exons 2 and 3 exactly match to exon 25 and 26 of *MTHFD1L* whereas exon 4 originates from intron 26 of the gene.

5.2.3.2 RT-PCR assays

Expression of the ncRNAs was assessed by RT-PCR in human placental, HEK 293 and lymphoblast RNA using a variety of different primer sets (Chapter 2, Sections 2.2.4 and 2.2.6) followed by agarose gel electrophoresis. The positions of primers are shown in Appendix BB. The results for the assay that would detect both ncRNAs simultaneously, showed a clear expression of the ncRNAs in placental sample and a very low expression in the lymphoblast sample, whereas no expression was detected in the HEK 293 sample (Figure 5.3b) (for sequencing data see Appendix P). Primers to distinguish each ncRNA were examined in placental RNA (Figure 5.3a) and both ncRNAChr2 and ncRNAChr9 were detected (Figure 5.3c) (for sequencing data see Appendix Q and Appendix R, respectively). ncRNAChr2 and ncRNAChr9 show an interesting combination of exons originated from *MTHFD1L* preceded by an unrelated exon. Since the database evidence of ncRNAs expression was corroborated by these RT-PCR experiments we decided to investigate them as possible *MTHFD1L* regulation elements.

5.2.3.3 ncRNA transfection in HEK 293 cells and the impact on MTHFD1L expression levels

ncRNAChr2 and ncRNAChr9 sequences were obtained in mammalian expression vector pME18SFL3 (Appendix A) from the Biological Resource Centre National Institute of Technology and Evaluation, Chiba, Japan. HEK 293 cells were transfected following the method described in Chapter 2, Section 2.2.1.4. mRNA and proteins were extracted at 3 different time points: 24 hours, 48 hours and 72 hours after transfection. ncRNAChr2 and ncRNAChr9 expressions were determined to verify transfection efficiency (Figures 5.4 and 5.5). No endogenous expression was detected in non-transfected controls as determined by an RT-PCR assay.

Eleven different endogenous control genes were previously investigated by Kirsty O'Brien to identify the control gene least affected by experimental conditions and most suitable for real-time PCR normalisation in cDNA samples of interest. Beta-Glucuronidase (GUS) endogenous control gene displayed the least amount of variation and was chosen as reference gene for the ncRNA transfection. Results were calculated by applying the comparative C_T method ($2^{-\Delta\Delta Ct}$) to measure the fold change in gene expression relative to the reference gene. Transfected samples and negative control with only transfection reagent (-CTR) were normalized to non-transfected sample for each time point. Differences in relative expression ratios were compared using one-way ANOVA, with results considered significant if a two-tailed *p*-value was < 0.05.

The expression of *MTHFD1L* long transcript was detected and results are shown in Figures 5.6, 5.7 and 5.8. *MTHFD1L* expression did not show any significant variation at any time point after the transfection. The same result was obtained for MTHFD1L protein expression detected by Western blot (Figure 5.9). This findings suggest that ncRNAChr2 and ncRNAChr9 does not modulate MTHFD1L expression either at transcriptional or at post-transcriptional level.

5.2.4 CHR X homologue sequence

5.2.4.1 Sequence analysis

This sequence on Chromosome X (ChrXseq) (location: Xp11.21, 57418826..57421885) is a processed pseudogene that shows about 93% similarity with *MTHFD1* (Figure 5.10). The 3005-base similar sequence covers almost the whole *MTHFD1* gene, including its coding region and 3′UTR. MTHFD1 mRNA and ChrXseq sequence alignment along with the position of the primers used for RT-qPCR are shown in Appendix AA. Stop codon mutations cut off the original *MTHFD1* ORF in several parts, the longest of which is 750 bp (black arrows in Figure 5.10). ChrXseq was analyzed to see if any polymorphisms may create these stop codons, but none of the variants overlap them. There is no evidence of ChrXseq expression in EST and Refseq databases.

5.2.4.2 RT-PCR assay

RT-PCR analysis showed an abundant expression of ChrXseq in placental, HEK 293 and lymphoblast RNAs. Red arrows in Figure 5.10 show the position of the RT-PCR amplicons analyzed (Figure 5.11). Negative controls without enzyme in the RT step were performed in order to avoid possible genomic DNA contamination. All cDNA products were also tested for genomic contamination by an intron spanning PCR assay designed on the *MTHFD1* gene (Chapter 2, Section 2.2.7). PCR products obtained were sequenced and no discrepancy with ChrXseq database sequence was found. Sequenced PCR products are shown in Appendix T, U, V, W, X and Y.

5.2.4.3 ChrX 3′UTR overexpression in HEK293

To verify if ChrXseq can modulate the expression of its parental gene *MTHFD1*, part of ChrXseq was cloned into the mammalian expression vector pcDNA™3.2/V5-DEST (Invitrogen) as explained in Chapter 2, Section 2.2.12. A preliminary test was performed using a cloned region from ChrXseq that overlaps MTHFD1 3′UTR (Figure 5.12). This region was selected because of the possible miRNA regulation involvement. pcDNA™3.2-ChrX3′UTR was transiently transfected in a 6-well plate seeded with HEK293 cells as described in Chapter 2, Section 2.2.1.4. pcDNA™3.2/V5-DEST empty vector was employed as a control and RNA was extracted after 24, 48, and 72 hours of transfection (Chapter 2.2.4). RT-qPCR for MTHFD1 was performed using

GUS as reference gene. Samples transfected with CHR3′UTR showed a lower MTHFD1 expression (Figure 5.13). Results were statistically significant for 48 hour (*p*-value=0.031) and 72 hour (*p*-value =0.042) samples.

5.2.4.4 Generation of a protein expression vector with ChrXseq longest ORF

Stop codon mutations cut off the original *MTHFD1* ORF in several parts, the longest of which is 750 bp (black arrows in Figure 5.10 and 5.12). This sequence could encode for a part of the 10-formyl-THF-synthethase domain of MTHFD1, which may be enough to retain enzymatic activity. The 750 ORF sequence was PCR amplified and cloned in *E.coli* protein expression vector Gateway® pDEST™15 (Chapter 2, Section 2.2.12) (Figure 5.12).

5.3 Discussion

Sequence analysis was performed to discover homologous coding genes or regulatory elements like ncRNAs, related to *MTHFD1* and *MTHFD1L*. The first similar sequence that we addressed was CHR11seq. For this sequence we found no evidence of expression either from the database or RT-PCR analysis. We therefore concluded that Chr11seq is probably an inactive sequence derived by transposition from *MTHFD1L*.

We next focused on two ncRNAs expressed from Chr2 and Chr9. In this case evidence of expression was found from both database and RT-PCR analysis but only a significant level of ncRNA was detected and only in the placental samples. A deeper investigation did not show any association between ncRNAs expression and *MTHFD1L* regulation. These findings suggest these ncRNAs might be expressed only in specific tissues and/or at specific stages of development. Despite sequence similarity, their function, if any, appear to be unrelated to *MTHFD1L* regulation.

ChrXseq was then studied as a pseudogene of *MTHFD1*. Despite the lack of annotation in numerous databases, this sequence was found to be highly expressed in different mammalian cell samples. The longest ORF, spans over less than one third of the parental gene. Previous reports demonstrated that expressed pseudogenes and ncRNAs can compete with their ancestral protein coding genes for the same pool of miRNAs

(Salmena *et al*. 2011). Circular RNA (circRNA), competing endogenous RNAs (ceRNAs) and pseudogene decoys share miRNA-response elements (MREs) with many mRNAs (Kosik 2013). To test the possibility that ChrX could act as a pseudogene decoy, the portion of ChrXseq corresponding to *MTHFD1* 3′UTR has been cloned into a mammalian expression vector and transfected in HEK293 cells. Results showed a significant downregulation of *MTHFD1* expression after 48 and 72 hours of expression. These data suggest that the ncRNA expression could be involved in the regulation of its parental gene *MTHFD1*. Nevertheless, this regulation does not seem to occur in a miRNA-regulated fashion. If the ChrX3′UTR sequence had acted like a miRNA decoy it would have competed with MTHFD1 for binding the same pool of miRNAs leading to an upregulation of the gene. A possible mechanism that could explain the inhibitory effect of MTHFD1 is the formation of DNA-RNA triplex. Recent publications are reporting considerable evidence for the existence of nucleic triplexes *in vivo* and their potential participation in a variety of biological processes including chromatin organization, DNA repair, transcriptional regulation and RNA processing (Buske *et al*. 2011). Nonetheless, the hypothesis of DNA-RNA triplex involvement is intriguing but remains to be tested. Further investigation, which will require the cloning of the full length ChrXseq transcript, may help to understand the role of this pseudogene. A protein expression vector containing the longest ORF in ChrXseq has also been generated and will be used to verify if the putative protein retain catalytic activity. In conclusion, this study provides interesting hints for the development of a possible future project.

Figure 5.1 Schematic representation of *MTHFD1L* spliced gene on Chr6.

Arrows show exons within the coding sequence of the gene. Blue line represents corresponding processed pseudogene on Chr11.

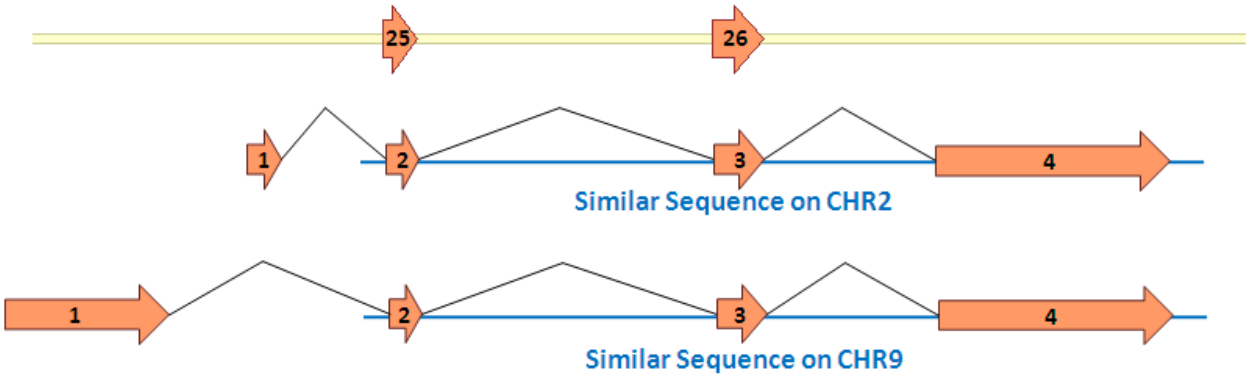


Figure 5.2 Schematic representation of exon 25-26 flanking region of *MTHFD1L* non-spliced gene on Chr6.

Blue lines represent corresponding non-processed pseudogenes on Chr2 and Chr9, while arrows on blue lines represent spliced ncRNAChr2 and ncRNAChr9 respectively.

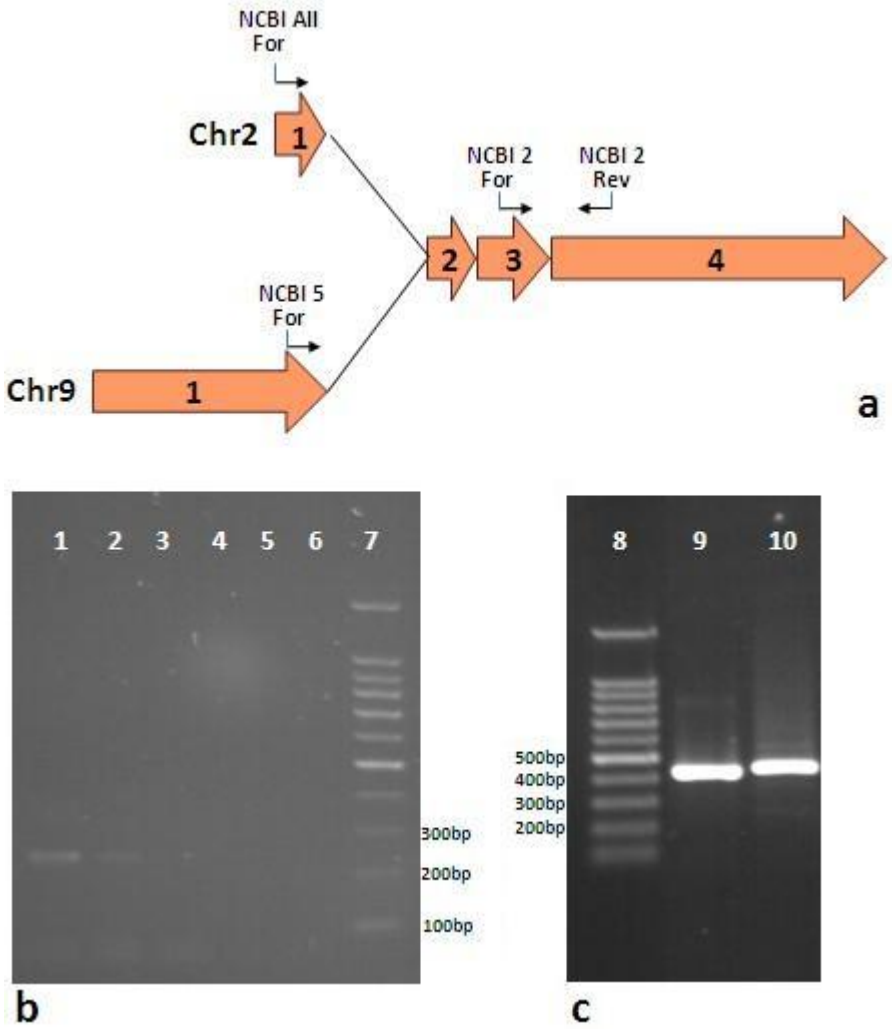


Figure 5.3 RT-PCR analysis of ncRNAChr2 and ncRNAChr9.

(a) Scheme of ncRNAs and relative position of primers (figure drawn with Microsoft PowerPoint). **(b)** cDNAs obtained from human placental (1), lymphoblast (2) and HEK 293 (3) RNAs and relative –RT controls (4-6) were amplyfied with NCBI 2For/2Rev primer, and run onto an agarose gel along with 100bp ladder (7). A band of 244 bases was clearly detected in the placental cDNA and lymphoblast cDNA (for sequencing data see Appendix P). The placental sample band was the strongest in intensity. **(c)** Placental cDNA amplified by using the following primers: NCBI 5For/2Rev (9), NCBI AllFor/2Rev (10), producing respectively 458 and 468; 100bp ladder (8) (for sequencing data see Appendix R and Appendix Q, respectively).

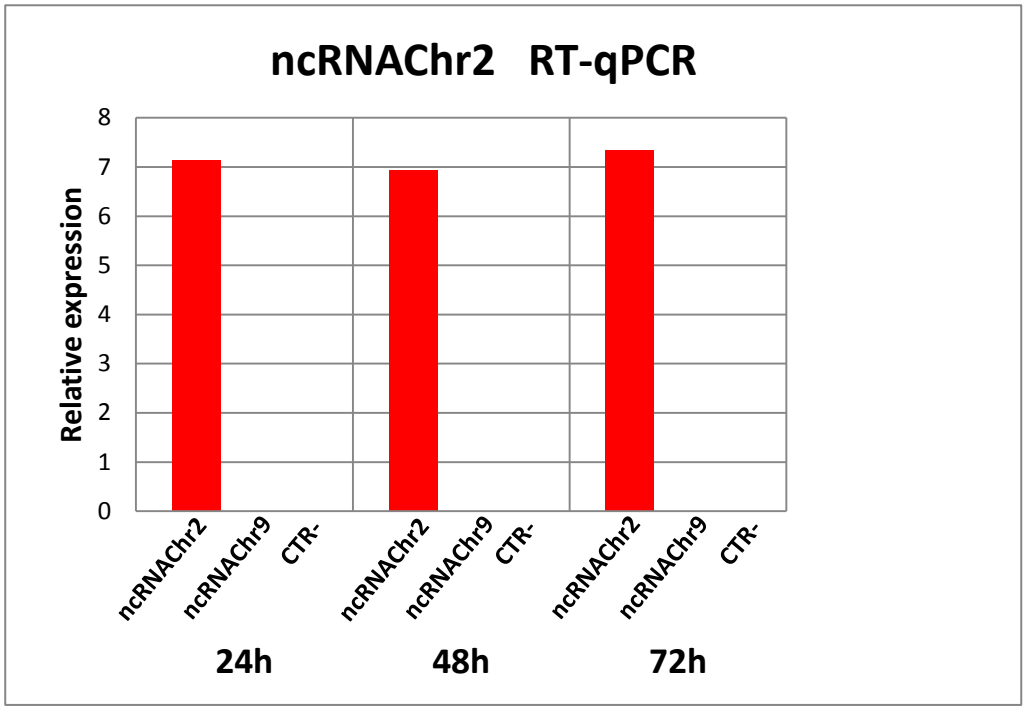


Figure 5.4 RT-qPCR analysis for ncRNAChr2.

ncRNAChr2 expression was determined to verify transfection efficiency at 24, 48 and 72 hours. Relative expression was calculated using GUS as reference gene. Samples which are not transfected with ncRNAChr2 did not show any expression.

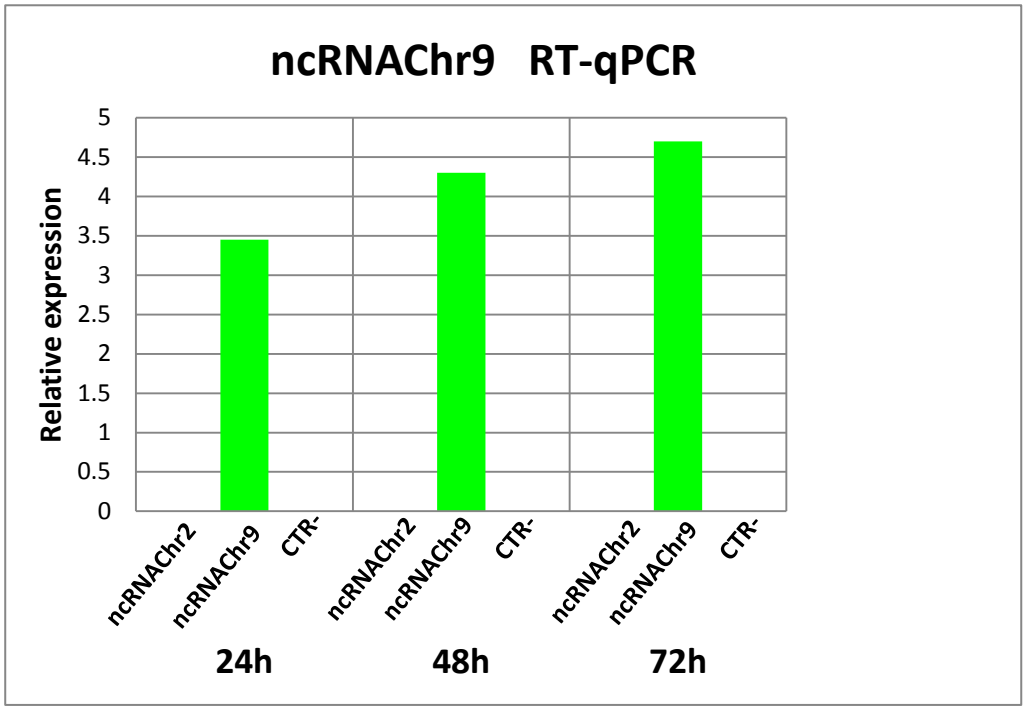


Figure 5.5 RT-qPCR analysis for ncRNAChr9.

ncRNAChr9 expression was determined to verify transfection efficiency at the different time points. Relative expression was calculated using GUS as reference gene. No expression was detected in samples not transfected with ncRNAChr9.

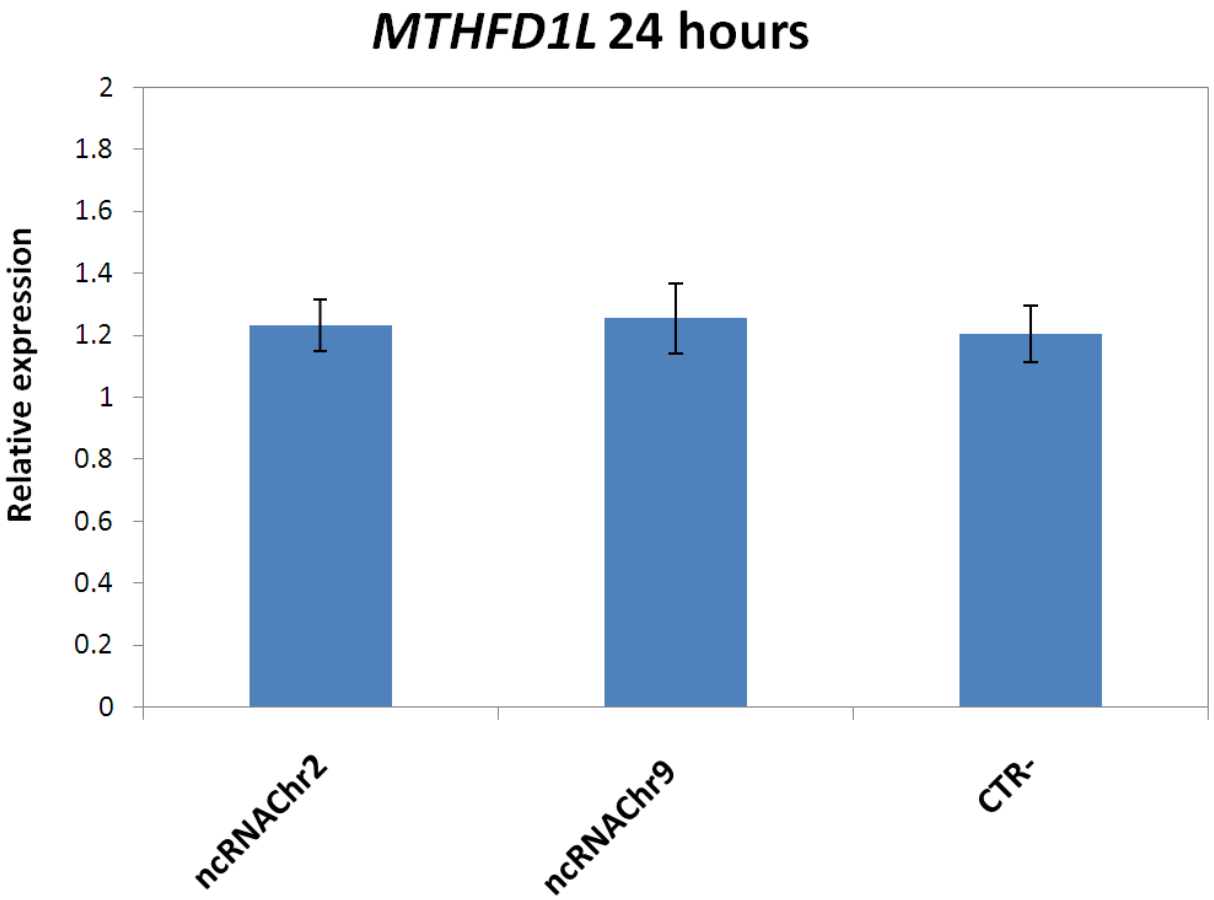


Figure 5.6 Relative *MTHFD1L* gene expression after 24 hours of transfection.
MTHFD1L gene expression was compared between transfected samples and negative control. Results were calculated by applying the comparative C_T method ($2^{-\Delta\Delta Ct}$) to measure the fold change in *MTHFD1L* expression relative to Beta-Glucuronidase (GUS) reference gene (Chapter 2, Section 2.2.10). Transfected samples and negative control with only transfection reagent (-CTR) were normalized to non-transfected sample. Differences in relative expression ratios were compared using one-way ANOVA, with results considered significant if a two-tailed *p*-value was < 0.05. *MTHFD1L* expression did not show any significant change after 24 hours of transfection.

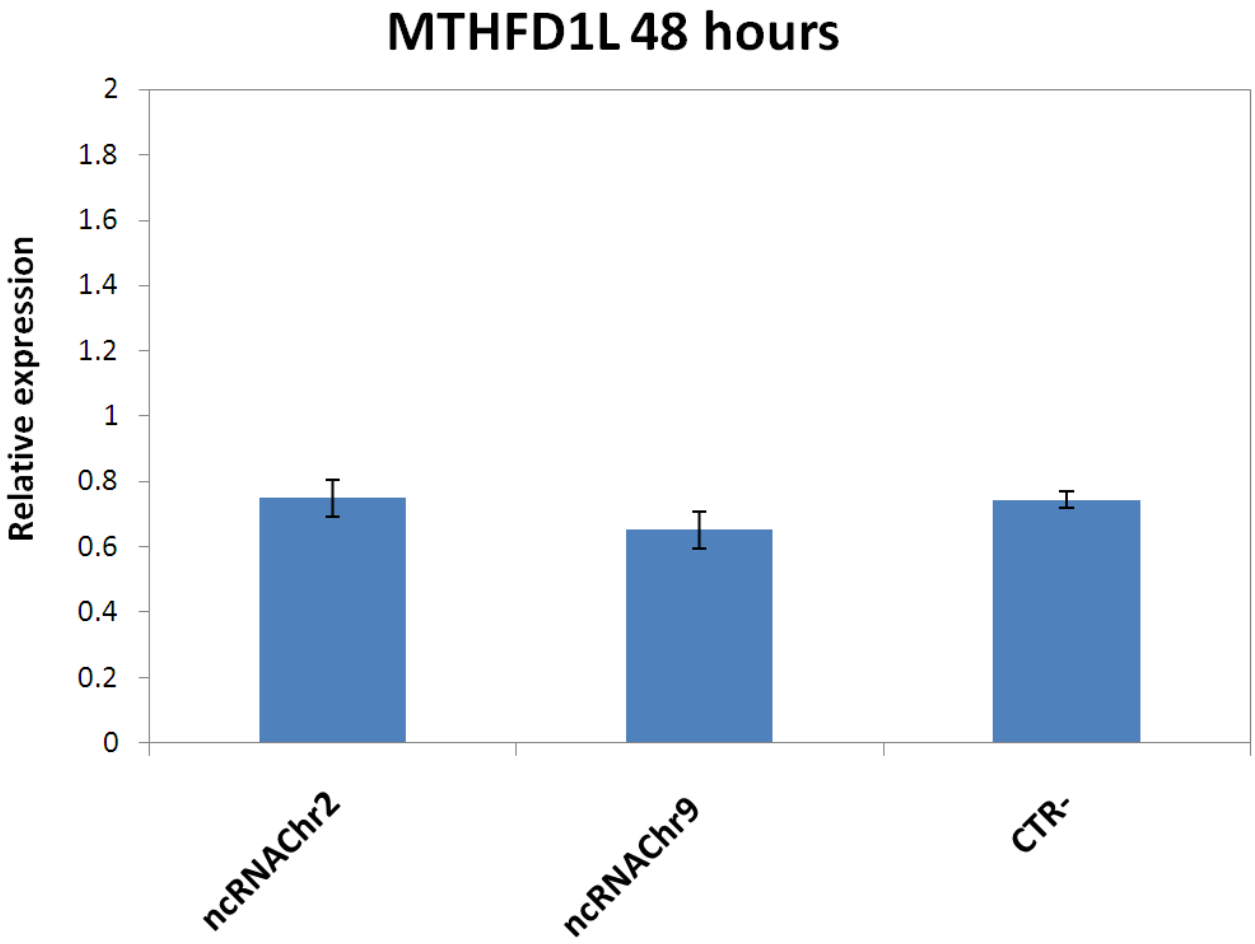


Figure 5.7 Relative *MTHFD1L* gene expression after 48 hours of transfection. *MTHFD1L* gene expression was compared between transfected samples and negative control. Results were calculated as for the 24 hour time point. *MTHFD1L* expression was not affected by ncRNA transfection after 48 hours of transfection.

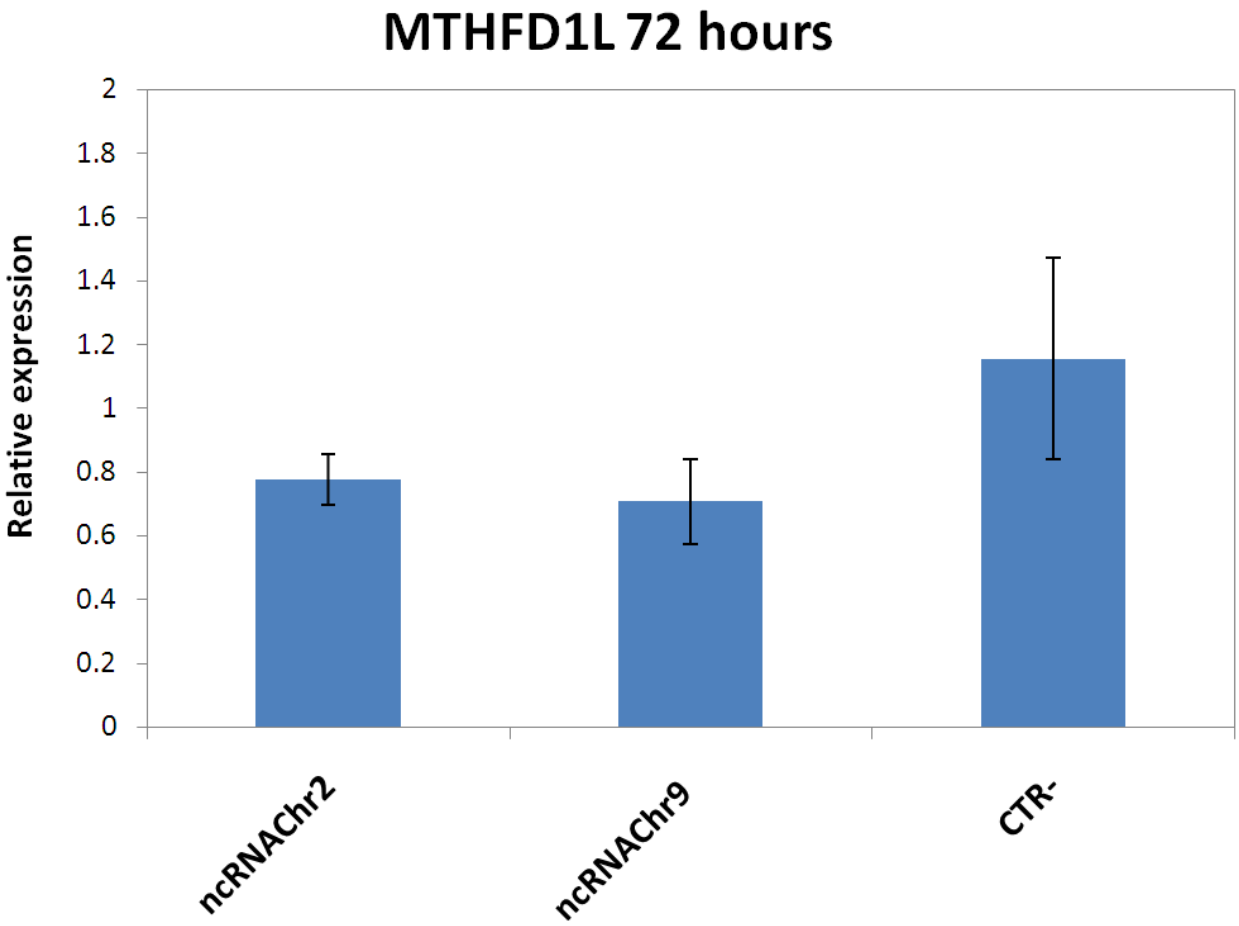


Figure 5.8 Relative *MTHFD1L* gene expression after 72 hours of transfection.
MTHFD1L gene expression was compared between transfected samples and negative control. Results were calculated as for the 24 hour and 48 hour time points. Again, *MTHFD1L* expression did not show any significant change after 72 hours of transfection.

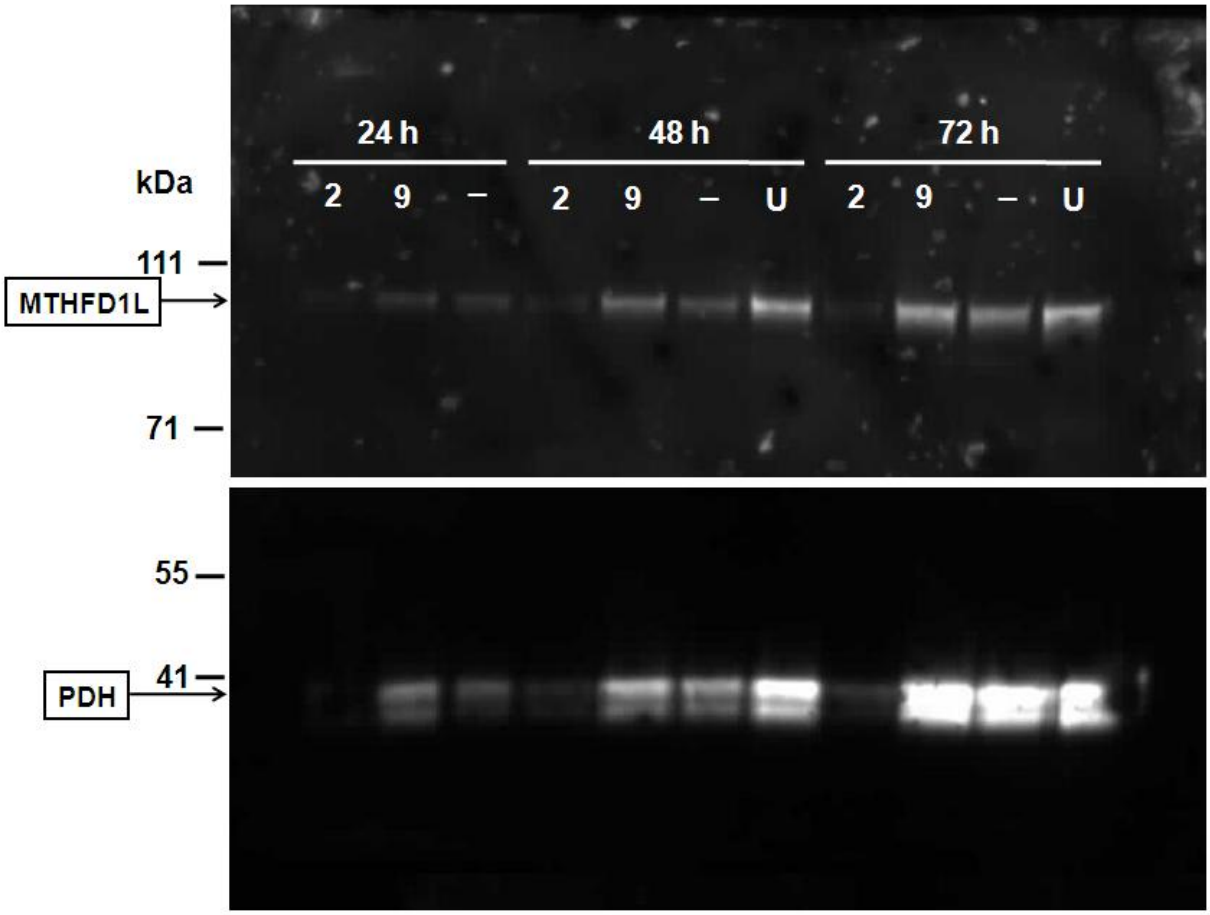


Figure 5.9 Western Blot for MTHFD1L expression at 24, 48 and 72 hour time points.

Mitochondrial proteins were extracted from HEK 293 cells after 24, 48 and 72 hours of transfection (Chapter 2, Section 2.2.1.4). Samples were analysed by Western blotting and probed with antibodies against both MTHFD1L and PDH subunit E1 alpha used as reference protein and mitochondrial marker. MTHFD1L expected size is 102 kDa after cleavage of the mitochondrial targeting sequence, while PDH subunit E1 alpha expected size is 41 kDa. ncRNA tranfected samples did not show a variation in MTHFD1L expression compared to negative control samples and normalised to PDH. ncRNAChr2 samples showed a general lower expression of both MTHFD1L and PDH. This can be explained by an higher mortality of ncRNAChr2 cell samples probably due to an high amount of bacterial endotoxins derived from plasmid Minipreps. **2**: ncRNAChr2, **9**: ncRNAChr9, **–**: negative control (only transfection reagent), **U**: untransfected negative control.

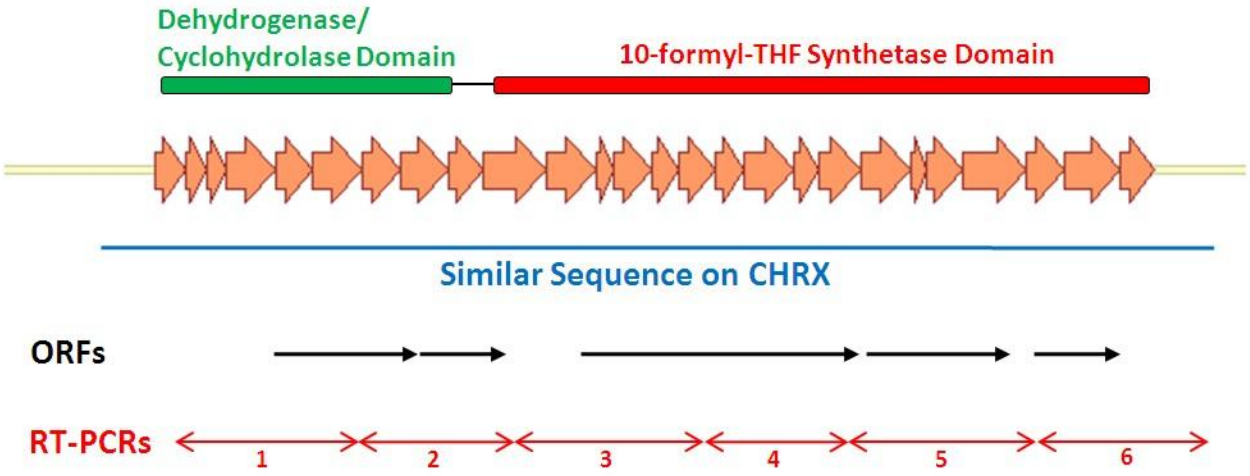


Figure 5.10 Schematic representation of *MTHFD1* spliced gene on Chr14 (figure drawn with Microsoft PowerPoint).

Thick pink arrows show exons within the coding sequence of the gene flanked by 5′ and 3′UTR. Above exons there are corresponding catalytic domains of MTHFD1 protein. Blue line represents corresponding processed pseudogene on ChrX, while black arrows underneath show ORFs along ChrXseq. Red arrows show RT-PCR amplicons obtained using following primer pairs: 1, ChrX 5For/5Rev (471 bases); 2, CHRX 6For/6Rev (431 bases); 3, CHRX 7For/4Rev (475 bases); 4, ChrX 3For/3Rev (418 bases); 5, ChrX 2For/2Rev (474 bases); 6, CHRX 1For/ 3′UTR 2nd Rev (453 bases).

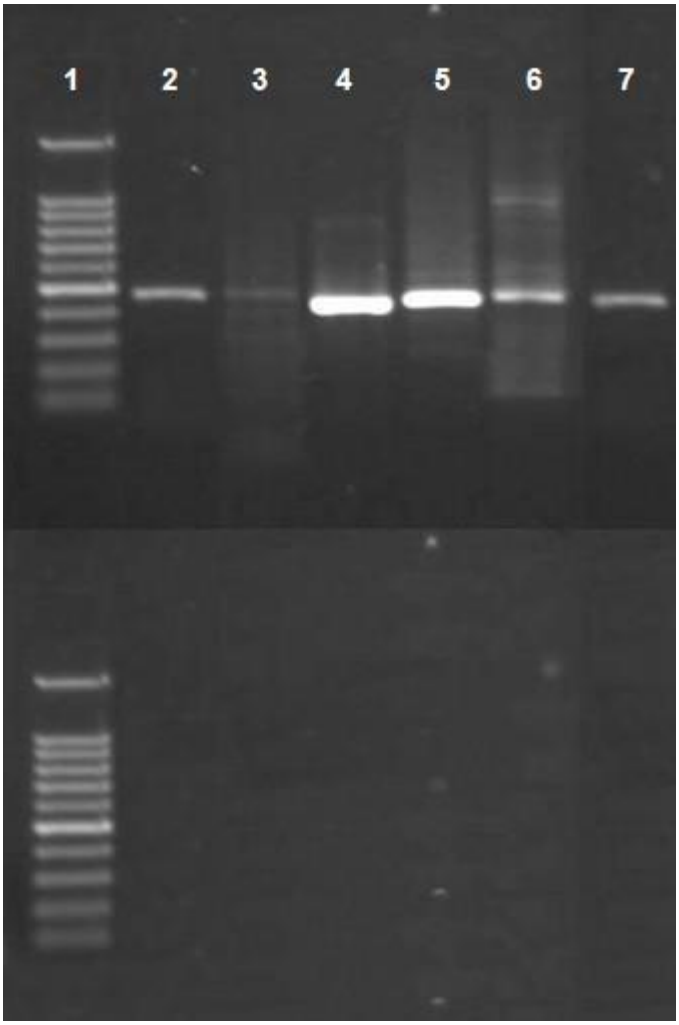


Figure 5.11 PCR products from Placental cDNA samples.

100bp ladder (**1**), CHRX 7For/4Rev (475 bases) (**2**), ChrX 5For/5Rev (471 bases) (**3**), ChrX 3For/3Rev (418 bases) (**4**), CHRX 6For/6Rev (431 bases) (**5**), ChrX 2For/2Rev (474 bases) (**6**), CHRX 1For/3′UTR 2nd Rev (453 bases) (**7**). Bands were sequenced and data are reported in Appendix T, U, V, W, X and Y, respectively.

The lower part of the figure shows the relative –RT controls which are all negative.

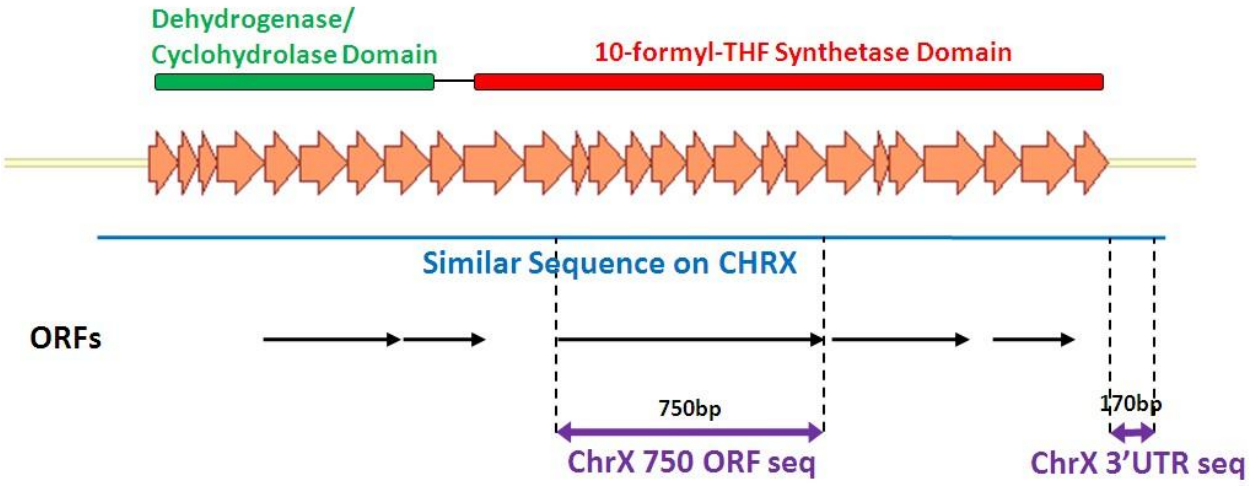


Figure 5.12 Schematic representation of *MTHFD1* spliced gene on Chr14 (figure drawn with Microsoft PowerPoint).

Thick pink arrows show exons within the coding sequence of the gene flanked by 5′ and 3′UTR. Above exons there are corresponding catalytic domains of MTHFD1 protein. Blue line represents corresponding processed gene on ChrX, while black arrows underneath show ORFs along ChrXseq. Purple arrows represent ChrX sequences that were selected for recombinant cloning. ChrX 750 ORF seq (750 bases) was cloned in vector Gateway® pDEST™15 vector (Appendix D), while ChrX 3′UTR seq was cloned in pcDNA™3.2/V5-DEST vector (Appendix C).

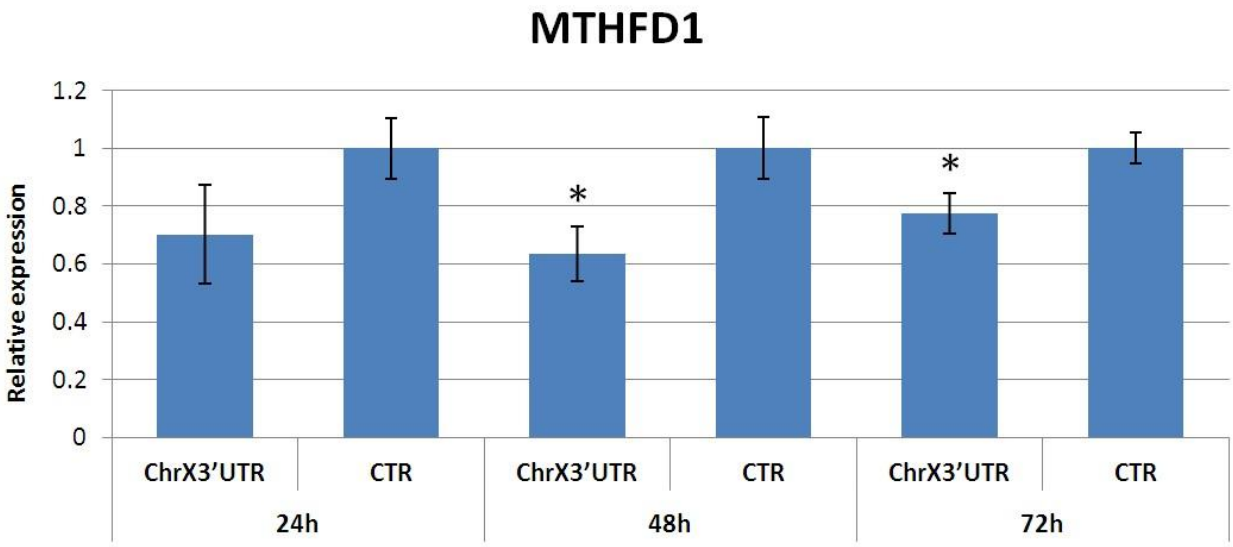


Figure 5.13 MTHFD1 expression after 24, 48 and 72 hour transfection ChrX3′UTR.

Data are represented as relative expression of ChrX3′UTR and CTR (empty vector) samples for 3 time points. All data were analysed by the E (Efficiency)-Method using GUS as an endogenous control for normalization. Statistical significance was calculated using Student's t-test (*$p<0.05$).

CHAPTER 6

MTHFD1L expression after cycloleucine exposure

6.1 Introduction

Cycloleucine is a non-metabolisable amino acid formed by cyclization of leucine. This molecule is an inhibitor of Methionine Adenosyltransferase (MAT) which is responsible for the production of S-Adenosylmethionine (SAM) (Figure 6.1). The exposure of cells to cycloleucine decreases the cellular SAM level, which is the major methyl donor (Zhuge and Cederbaum 2007, Tchantchou and Shea 2008). The donation of a methyl group during the methylation reaction converts SAM to S-adenosylhomocysteine (SAH) under the catalysis of S-adenosylmethionine methyltransferase (Finkelstein, 1998). SAH is subsequently hydrolyzed to homocysteine and the methyl cycle is completed by the remethylation of homocysteine to methionine.

As explained in Chapter 1, Section 1.1.4, methyl cycle forms part of 1C metabolism. Cycloleucine inhibition of MAT and consequent disruption of the methyl cycle may have different effects on 1C metabolism regulation and the regulation of its associated genes. Our group recently investigated the molecular response to 1C metabolism inhibition by exposing human lymphoblast cell lines to the cycloleucine (Carroll *et al*. 2012). Microarray analysis was carried out to examine the expression of 47,000 transcripts.

Taking advantage of this previous study, here, we specifically tested the expression of Short and Long *MTHFD1L* transcripts in response to cycloleucine exposure. As described in Chapter 1, Section 1.1.7, the Long transcript expresses an active protein, while the Short transcript, which is caused by the alternatively spliced exon 8A, is enzymatically inactive. Our previous study suggested that high Long/Short transcript ratio could be associated with the risk of NTDs and that this alternative splicing could have a role in the regulation of *MTHFD1L* (Parle-McDermott *et al*. 2009). Moreover, differences in the alternative splicing patterns are recurrent among folate-related genes (Duan *et al*. 2009). The aim of this chapter is to verify if a perturbation of the folate pathway by using cycloleucine could affect *MTHFD1L* regulation and if this regulation occurs through the mechanism of alternative splicing.

6.2 Results

6.2.1 *MTHFD1L* Long isoform expression after cycloleucine exposure

Cycloleucine time course was carried out as described in Chapter 2, Section 2.2.11. *MTHFD1L* Long transcript, was detected in Coriell lymphoblast cells after exposure to 5mM, 7.5mM and 10mM of cycloleucine for 24, 48 and 72 hours (Carroll *et al*. 2012). Five different endogenous control genes were investigated to identify the control gene most suitable for real-time PCR normalisation in cDNA samples of interest. The 40S Ribosomal Protein S13 gene (*RPS13*) displayed the least amount of variation and was selected as reference gene. The mean and standard deviation (SD) were calculated for treated and non-treated samples for each time point using Microsoft Excel®. Results were calculated by applying the E-method to measure the fold change in gene expression relative to the reference gene. Treated samples were then normalised to non-treated samples of each specific time point. Differences in relative expression ratios were compared using one-way ANOVA, with results considered significant if a two-tailed *p*-value was < 0.05.

Gene expression of MTHFD1L long isoform was significantly increased after 24 hours of exposure in samples treated with 7.5mM and 10mM of cycloleucine (*p*-value 0.001 and 0.004 respectively) (Figures 6.2 and 6.3). In 5mM/24hours treated samples the upregulation effect is still detectable, but no longer significant (*p*-value = 0.103) (Figure 6.4). In all the treated samples *MTHFD1L* Long transcript expression returns to a level comparable to negative control after 48 and 72 hours.

6.2.2 *MTHFD1L* Short isoform expression after cycloleucine exposure

The MTHFD1L Short isoform expression showed a similar but opposite trend, compared to the Long isoform as expected. Samples exposed to all the concentrations of cycloleucine showed a significant inhibition of *MTHFD1L* Short transcript expression at 24 hour time point (Figures 6.5, 6.6 and 6.7). Consistently with Long isoform result, Short mRNA levels return to normal expression after 48 and 72 hours of exposure.

6.3 Discussion

Results show that *MTHFD1L* transcript level responds to cycloleucine exposure in Coriell lymphoblast cells. This effect is limited in time as we detected a change in gene expression only after 24 hours of exposure. The upregulation of the Long transcript coincides with a downregulation of the Short transcript suggesting that the regulation operates through their alternative splicing. An increase of Long/Short transcript ratio results in an increment of active MTHFD1L protein in mitochondria. These findings continue to highlight the key role of alternative splicing in the *MTHFD1L* gene. The regulation of alternative splicing depends on the interaction between splicing factors and regulatory elements in the pre-mRNA, (Kornblihtt 2006). The methyl cycle forms part of 1C metabolism and the MTHFD1L enzyme is a key supplier of the 1C flow between the mitochondria and the cytoplasm. Cycloleucine inhibition of MAT and consequent disruption of the methyl cycle appears to lead to cascade effects resulting in *MTHFD1L* upregulation through the altered activity of a yet to be identified alternative splicing factor. For example, the blocking of the methyl cycle might lead to a higher availability of 1-C units for other reactions in the folate metabolic pathway and this can result in an overexpression of proteins necessary for 1-C metabolism such as MTHFD1L.

As mentioned in Chapter 1, Section 1.1.7, we showed that *MTHFD1L* Long transcript is significantly upregulated in HEK293 cells cultured in the absence of folate. This cycloleucine time course experiment confirms our hypothesis that a high level of *MTHFD1L* Long transcript might be a marker for a folate pathway inhibition and, as a consequence, for folate deficiency. These data, in combination with our previous work on NTDs, further emphasise the importance of the amount of the functional MTHFD1L enzyme to support 1C metabolism in the mitochondria and the flow of 1Cs between the mitochondria and the cytoplasm. The level of MTHFD1L enzyme responds to disruption of folate metabolism and this appears to operate by alternative splicing (this chapter) and possibly miRNA regulation (Chapter 4).

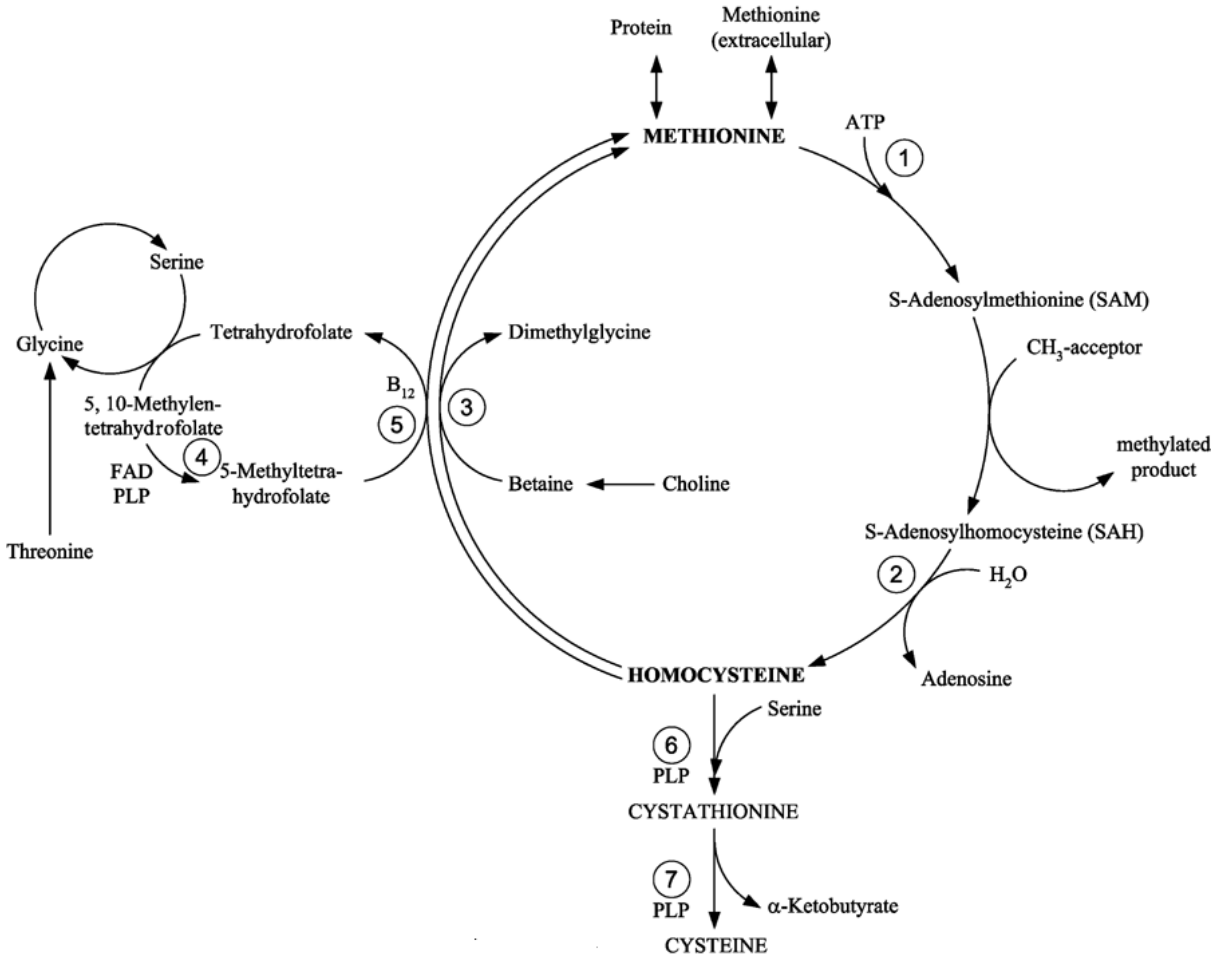


Figure 6.1 Pathways of methionine and homocysteine (modified from Ditscheid *et al.* 2005).

This figure shows two metabolic pathways involving methionine and homocysteine: remethylation and transsulphuration. The remethylation of homocysteine into methionine requires 5-methyltetrahydrofolate for methyl group donation and cobalamine as coenzyme. 5-Methyltetrahydrofolate is formed from serine. Tetrahydrofolate, FAD and pyridoxalphosphate act as coenzymes. Serine, itself, is generated by glycine. Alternatively, betaine may act as methyl group donor. The second pathway results in the formation of cysteine via cystathionine. Cysteine plays an important role as a precursor of glutathione and taurine. This transsulphuration pathway requires serine as cofactor and pyridoxalphosphate as coenzyme.

1:methionine-adenosyl-transferase, **2**:*S*-adenosyl-homocysteine-hydrolase, **3**:betaine-homocysteine-methyltransferase, **4**:methyltetrahydrofolate-reductase, **5**:5-methyltetrahydrofolate-methyl-transferase, **6**:cystathionine-beta-synthase, **7**:cystathionine-γ-lyase; FAD=flavine adenine dinucleotide, PLP:pyridoxalphosphate.

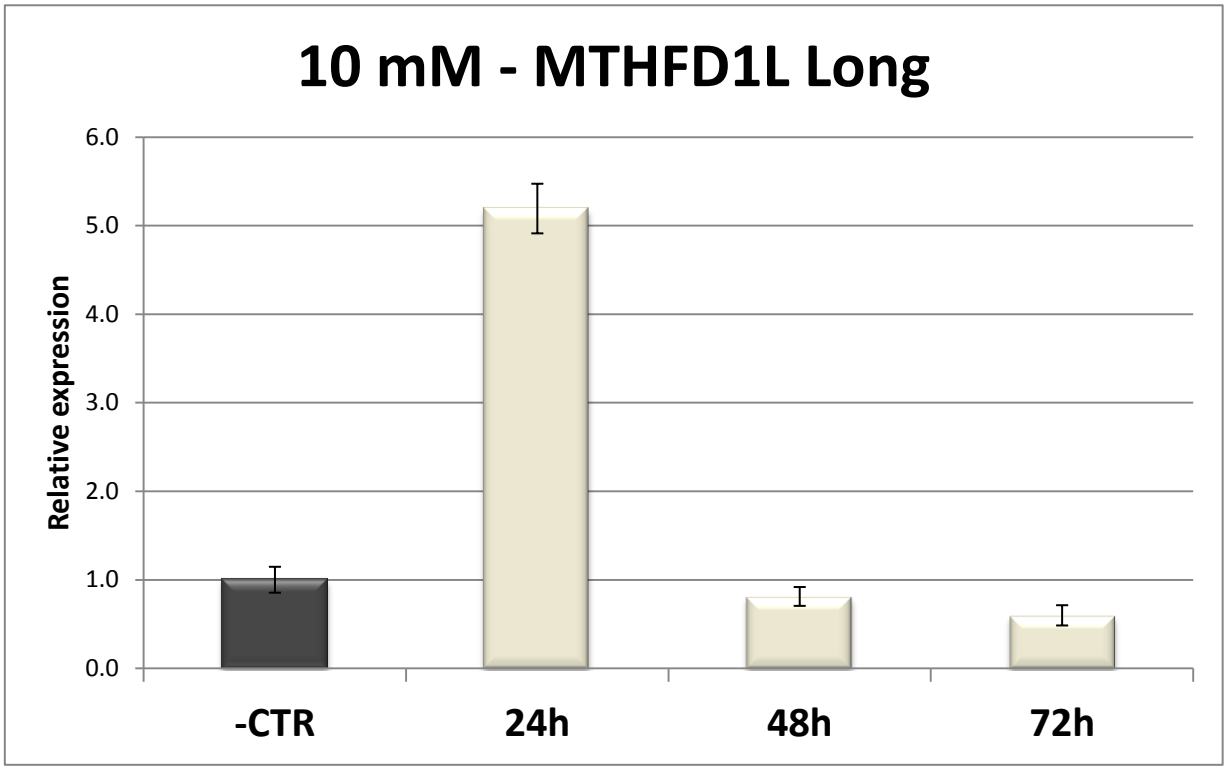


Figure 6.2 Expression of *MTHFD1L* mRNA Long isoform after exposure to 10mM of cycloleucine.

The gene relative expression was calculated using *RPS13* as reference gene. Treated samples were normalised to negative control of each specific time point. 24 hour of exposure showed a significant ~5 fold increase of gene expression (*p*-value = 0.001) that returns to normal level after 48 and 72 hours (*p*-value = 0.35 and 0.076 respectively).

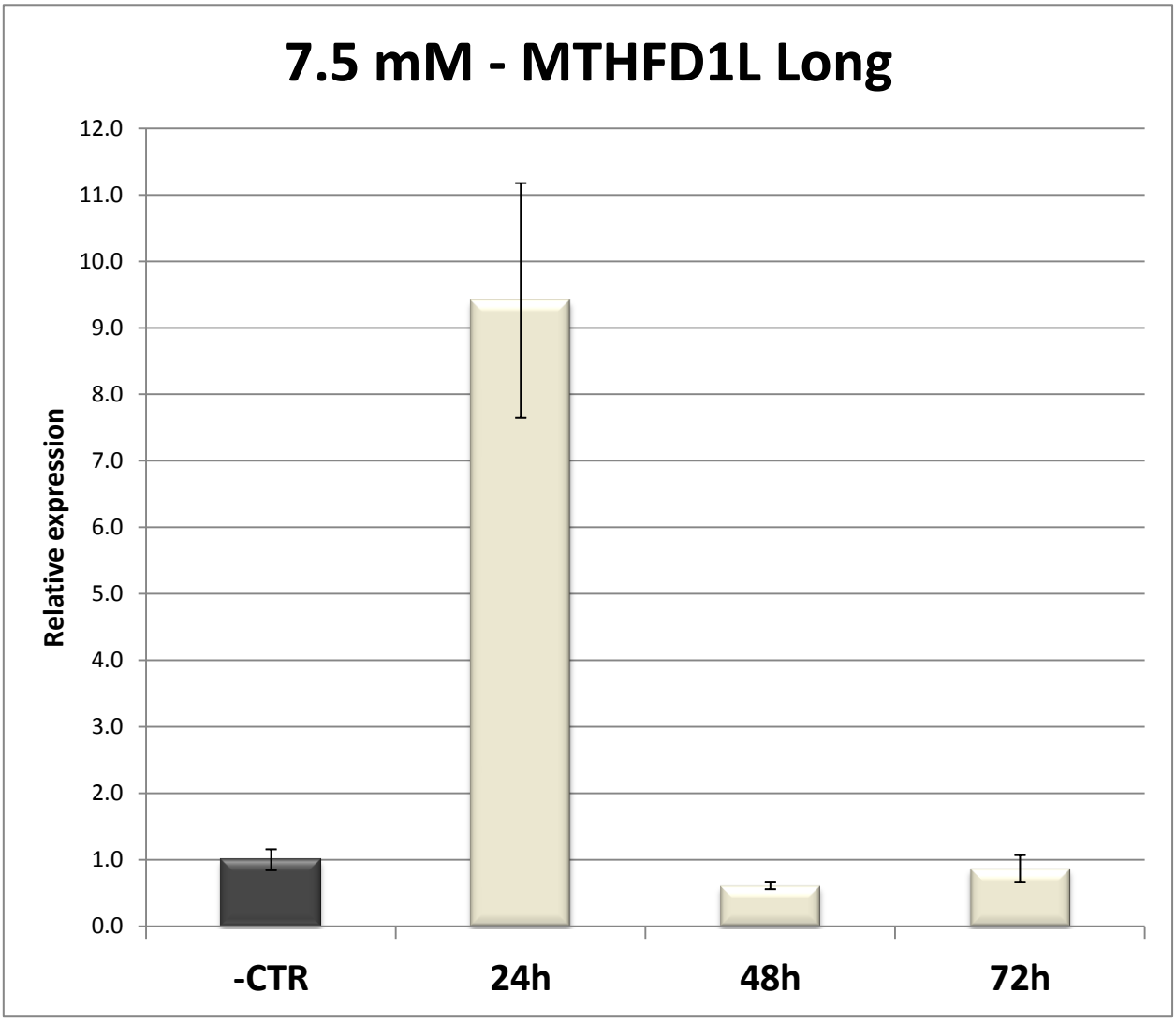


Figure 6.3 Expression of *MTHFD1L* mRNA Long isoform after exposure to 7.5mM of cycloleucine.

The gene relative expression was calculated as above. Again, 24 hour exposure point showed a significant ~9-fold overexpression of the gene (*p*-value = 0.004). After 48 and 72 hours of exposure did not show any significant change in *MTHFD1L* expression.

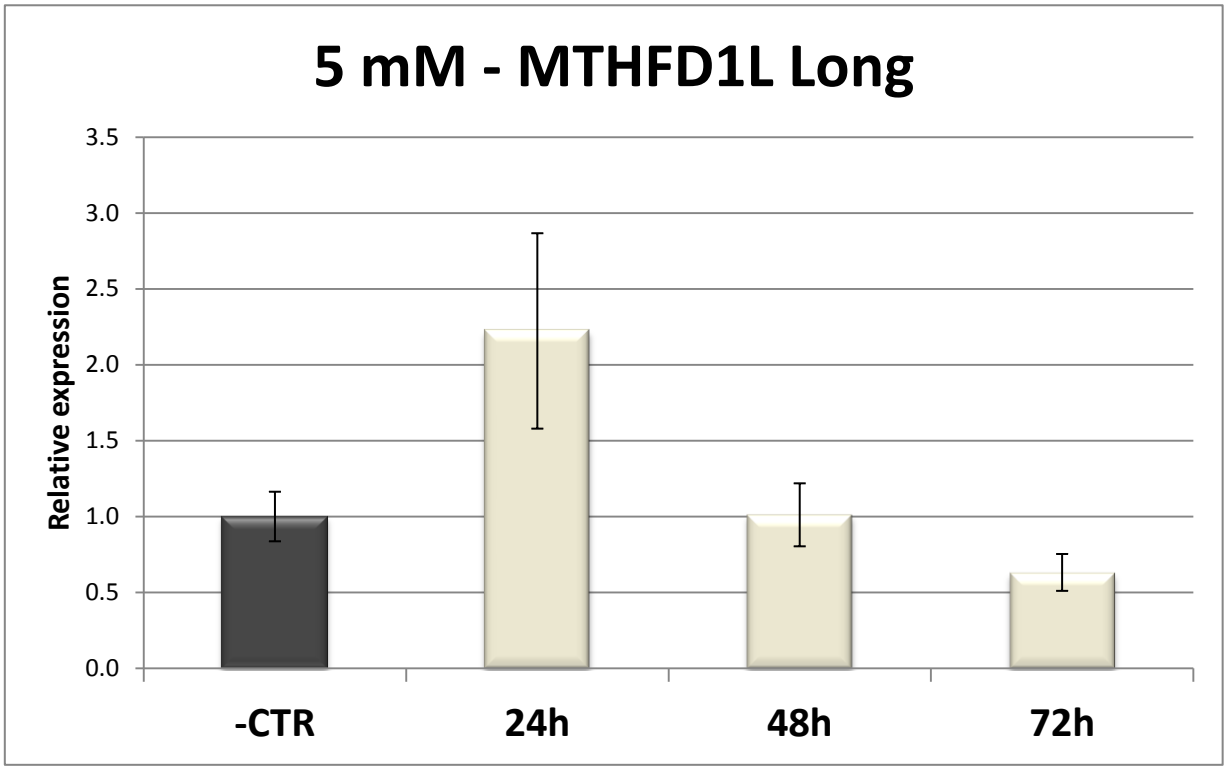


Figure 6.4 Expression of *MTHFD1L* mRNA Long isoform after exposure to 5mM of cycloleucine.

The gene relative expression was calculated as above. Exposition to 5mM of cycloleucine showed a smaller change of *MTHFD1L* expression (~2 folds) with a *p*-value no longer significant (p=0.103). After 48 and 72 hours of exposure did not show any significant change in *MTHFD1L* expression (p-value = 0.217 and 0.175 respectively).

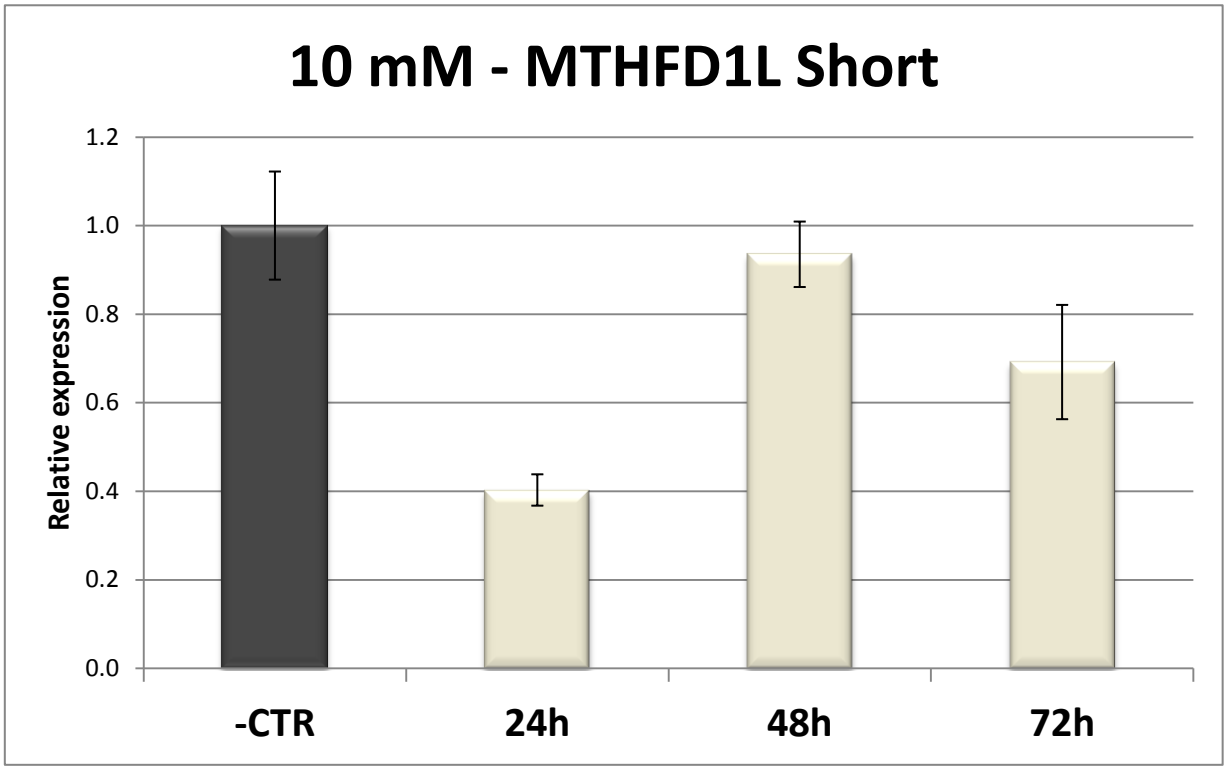


Figure 6.5 Expression of *MTHFD1L* mRNA Short isoform after exposure to 10mM of cycloleucine.

The gene relative expression was calculated using *RPS13* as reference gene. Treated samples were normalised to negative control of each specific time point. Differently from *MTHFD1L* Long isoform, the Short transcript showed a significant downregulation of ~2.5 folds after 24 hours of exposition to cycloleucine (*p*-value = 0.002). After 48 and 72 hours of exposure no more significant changes in gene expression are detected.

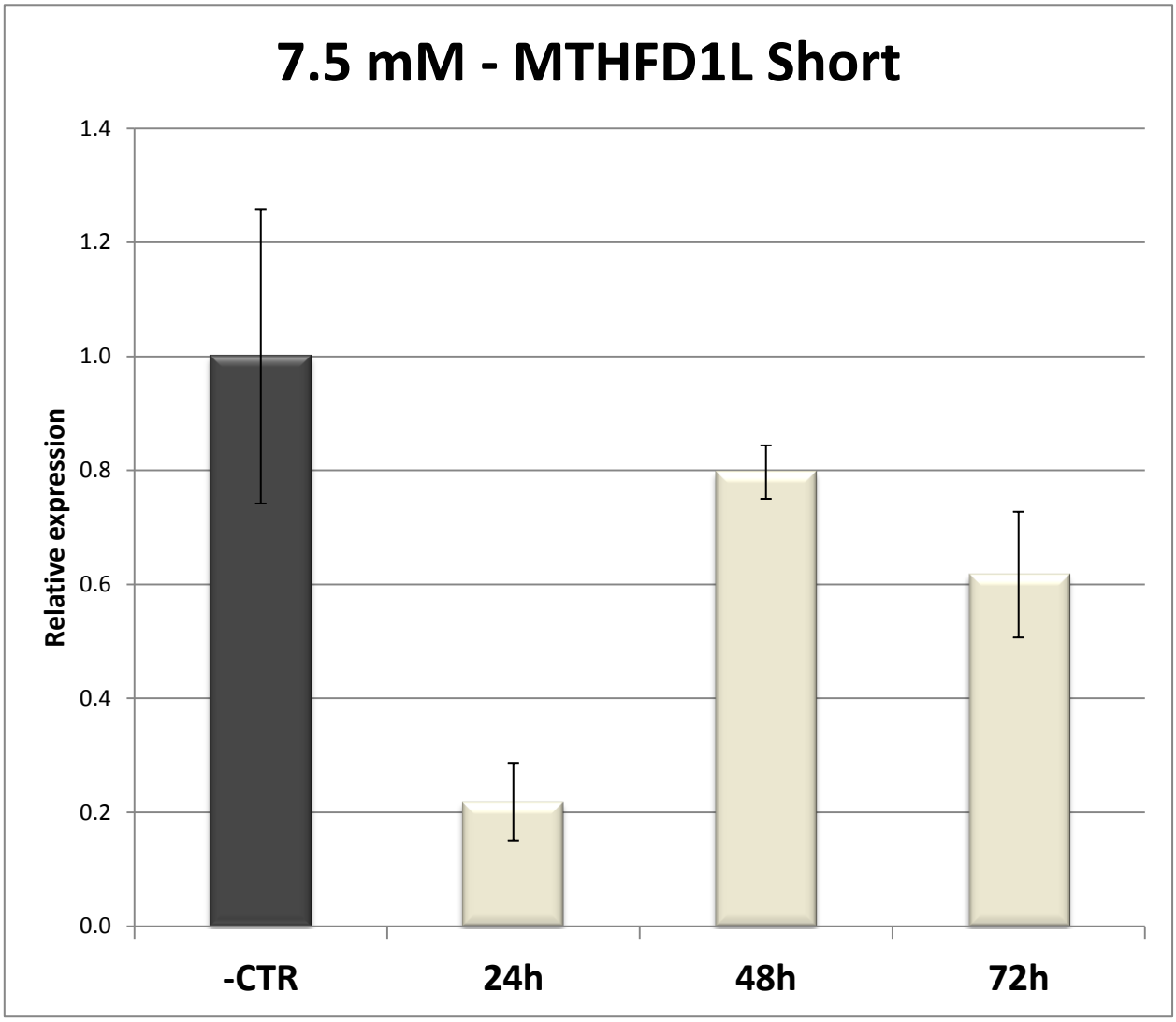


Figure 6.6 Expression of *MTHFD1L* mRNA Short isoform after exposure to 7.5mM of cycloleucine.

The gene relative expression was calculated as above. Again, *MTHFD1L* Short transcript showed a significant downregulation of ~5 folds after 24 hours of exposure to cycloleucine (*p*-value = 0.004). After 48 and 72 hours of exposure *MTHFD1L* Short isoform expression returns to levels comparable to negative control.

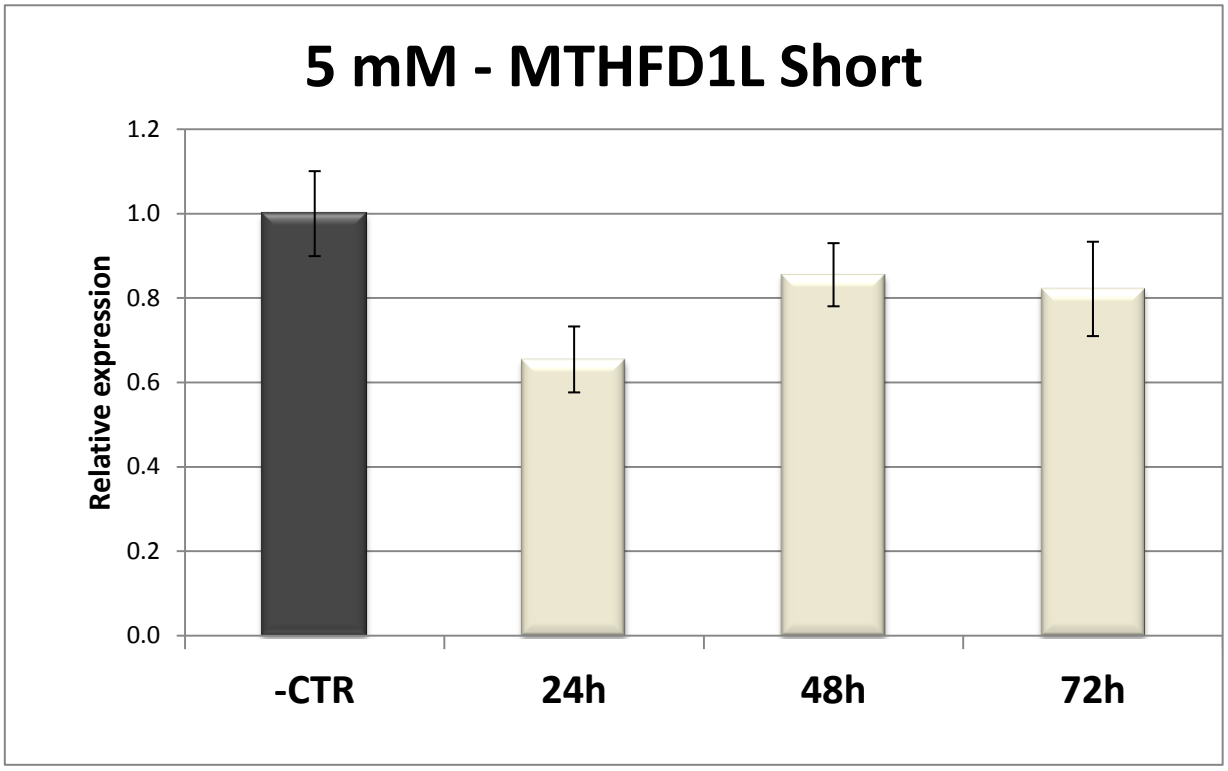


Figure 6.7 Expression of *MTHFD1L* mRNA Short isoform after exposure to 7.5mM of cycloleucine.

The gene relative expression was calculated as above. Similarly to *MTHFD1L* Long transcript, the exposition to 5mM of cycloleucine showed a lighter effect on the Short transcript expression after 24 hours of exposition. In this case the expression is inhibited of ~1.5 folds showing a slightly significant *p*-value ($p = 0.036$). After 24 and 48 hours of exposure the gene expression level returns to negative control level.

CHAPTER 7

Proteomic analysis after perturbation of MTHFD1L expression

7.1 Introduction

The term "proteomics" generally refers to the large-scale analysis of proteins and it is often specifically used for protein purification and mass spectrometry. The opportunities offered by proteomic tools to broaden our knowledge of disease mechanisms are formidable. Particularly promising areas of research include the development of novel biomarkers for diagnosis, the identification of new targets for therapeutics and the study of altered protein expression (Hanash 2003).

As explained in Chapter 1, Sections 1.1.6 and 1.1.7, perturbations of MTHFD1L expression levels seem to be associated with disease. High expression levels of MTHFD1L were found in human colon adenocarcinoma (Sugiura *et al*. 2004) and significantly correlated with the growth rates of human cancer cell lines (Jain *et al*. 2012). In the latter study, it has been also demonstrated that breast cancer patients with above-median expression levels of MTHFD1L, SHMT2 and MTHFD2, but not their cytosolic homologues, displayed higher mortality (Jain *et al*. 2012). In addition, disease association studies (Chapter 1, Section 1.1.7) performed in our laboratory suggested that higher MTHFD1L protein levels could be associated with an increased NTD risk (Parle-McDermott *et al*. 2009).

On the other hand, the loss of the protein in *Mthfd1l* knockout mice was found to be lethal to developing embryos, causing fetal growth restriction and NTD phenotypes with 100% penetrance (Momb *et al.* 2013). In Chapter 4, we also showed that the impact of SNP rs7646 on the binding of miR-197 to the 3′UTR of *MTHFD1L* provides a mechanistic explanation of the previous association of this SNP with NTD risk (Parle-McDermott *et al*. 2009) i.e., the 'G' allele provides a better binding site for miR-197. The immediate assumption is that the 'G' allele would lead to a decrease in embryonic MTHFD1L levels. However, the actual expression of miR-197 and miR-9 during embryogenesis and their competition for the binding of MTHFD1L 3′UTR require further investigation to prove this assumption. What is clear from the data in the literature and in this thesis is that the MTHFD1L functional enzyme levels are of particular significance given that increased or decreased levels are consistently observed across a number of human diseases. Maintaining MTHFD1L levels within an

expression level "window", especially during cell proliferation, may be crucial for the maintenance of healthy cells.

In this chapter, the role of an altered MTHFD1L expression in disease will be assessed by a proteomic analysis. This will be performed on cells that have been manipulated to 'knock-down' expression of MTHFD1L and to overexpress it providing an understanding of how the expression level of MTHFD1L is significant in disease. Changes in protein expression of such cell lines will be assessed by means of label-free liquid chromatography–mass Spectrometry (LC-MS).

7.2 Material and methods

7.2.1 *MTHFD1L* knockdown by shRNA Lentiviral transduction and generation of stable HEK293 cell lines

MTHFD1L knockdown in HEK293 cells was performed using the MISSION^{TM} TRC lentiviral shRNA vectors. Five different constructs, targeting different regions of MTHFD1L, and a non-target shRNA control were purchased from Sigma-Aldrich. Two different kind of Isopropyl β-D-1-thiogalactopyranoside (IPTG) inducible vectors were used for the experiment. pLKO_IPTG_3xLacO vector (Appendix G) contains three lac operon sequences (two in the U6 promoter and one 3′ of the promoter) affording a tight gene regulation. Whereas, the pLKO_IPTG_1xLacO (Appendix F) vector contains a single lac operon sequence in the U6 promoter, which allows a stronger shRNA expression, but looser control of the promoter when it is not induced. Constructs and vectors are reported in Table 7.1. Constructs number 3 and 5 were already validated for knocking down *MTHFD1L* by Sigma.

HEK 293 cells were cultured as explained in Chapter 2, Section 2.2.1.1. Cells were counted and 1.6 x 10^4 cells were seeded in a 96 well plate in 120 µl of complete media and incubated at 37^{o}C for about 20 hours until ~70% confluency was obtained. Media was removed from wells and 110 µl media and Hexadimethrine bromide, at a final concentration of 8 mg/ml, were added to each well to enhance transduction. The plate was gently swirled to mix. A range of volume (2, 5, 10, and 15 µl) of each lentiviral construct was added to appropriate wells. The plate was gently swirled and incubated at

37°C for about 20 hours. Then the media containing viral particles replaced with 120 µl of fresh media particles and the plate was incubated at 37°C for 24 hours. Media was then replaced with fresh media containing 5 µg/ml of puromycin and resistant cells were selected for 1 week. Media with 5 µg/ml puromycin was also tested in non-transduced cells and all died within 4 days. Resistant transducted cells were then cultured for 2 weeks in 5 µg/ml puromycin media until 75cm^3 flasks were obtained and used for making backup frozen stocks and for preliminary downregulation test.

7.2.2 Preliminary shRNA downregulation test

A preliminary test was run to determine the best concentration of IPTG to use and to select two shRNA constructs for the proteomic analysis. A 100% confluent 75cm^3 flask for each of the 6 shRNA was split in 4 flasks and cells were grown in 10 ml of 5 µg/ml puromycin DMEM with 0 µM, 10 µM, 50µM or 100 µM of IPTG. To confirm that non-target shRNA does not affect *MTHFD1L* expression, normal non-transduced HEK293 cells were also grown in DMEM with 100 µM of IPTG. After 3 days cells were collected and *MTHFD1L* expression was tested. For RNA extraction and RT-qPCR analysis see Chapter 2, Sections 2.2.4 and 2.2.10, respectively.

7.2.3 Generation of HEK293 stable cell line overexpressing MTHFD1L

A modified version of *MTHFD1L* coding sequence with attB-flanked in pMA-RQ vector was purchased from Invitrogen®. *MTHFD1L* coding sequence in this clone was modified without changing the resulting amino acid sequence in order to optimize the protein expression in mammalian cells. The sequence was cloned in the mammalian expression vector pcDNA™3.2/V5-DEST (Appendix C) using Invitrogen® Gateway cloning system as explained in Chapter 2, Section 2.2.12. HEK293 cells were transfected in a 6-well plate with pcDNA3.2-MTHFD1Lopt plasmid and pcDNA™3.2/V5-DEST empty vector (used negative control for the experiment) as described in Chapter 2, Section 2.2.1.4. After 3 days of transfection, media was replaced with fresh media containing 500 µg/ml of G418 antibiotic and resistant cells were selected for 2 weeks. Media with 500 µg/ml of G418 antibiotic was also tested in non-transfected cells, that all died within 10 days. Stable transfected cells were then

cultured for 2 weeks in 500 µg/ml G418 until 75cm^3 flasks were obtained and used for making backup frozen stocks and for single clone selection.

7.2.4 Single clone selection

Single clone cell lines were generated using three different shRNAS: 1(-CTR), 3 and 7 for knock-down and pcDNA3.2-MTHFD1Lopt and pcDNA™3.2/V5-DEST empty vector for overexpression. Serial dilutions 1:5 were made with each cell line in a 96-well plate and incubated for 24 hours at 37^{o}C. Clones were screened by microscopy and wells containing single cells were marked. At least 3 single clones for each cell line were then subcultured from the 96-well plate into larger vessels following this order: 24-well plate, 6-well plate, 25cm^3 flask and 75cm^3. The process required 30-40 days depending on the cell line. Two clones for each cell line were grown in 75cm^3 flasks for 5 days in media with 200 µM IPTG. Cell pellets were then collected and tested for *MTHFD1L* expression by RT-qPCR as per Chapter 2 Sections 2.2.4 and 2.2.10. A specific RT-qPCR UPL assay was designed for MTHFD1Lopt sequence (Chapter 2, Section 2.2.10).

7.2.5 MTHFD1L overexpression and downregulation experiments

Single clone cell lines named shRNA1-1 (CTR), shRNA3-3 and shRNA7-4 were selected for the downregulation experiment. A 100% confluent 75cm^3 flask for each cell line was split into five 75cm^3 flasks and cells were grown for 5 days in 5 µg/ml puromycin media with 200 µM IPTG. Cells were collected and washed 5 times in 1ml of 1x PBS buffer and cell pellets were stored at -80°C until required for use.

Similarly, single clone cell lines named 1L-1 and CTR-1 were selected for the overexpression experiment. A 100% confluent 75cm^3 flask for each cell line was split in five 75cm^3 flasks and cells were grown for 5 days in 500 µg/ml G418 media. Cells were collected and washed 5 times in 1ml of 1x PBS buffer and cell pellets were stored at -80°C until required for use. One cell pellet was used for RNA extraction and RT-qPCR detection of *MTHFD1L* mRNA (Chapter 2, Section 2.2.10). A second cell pellet was used for mitochondrial protein extraction (Chapter 2, Section 2.2.13.1) and

Western Blotting using anti MTHFD1L anti PDH antibodies (Chapter 2, Sections 2.2.13.4-5). The three remaining cell pellets were used for label-free LC-MS analysis as reported below.

7.2.6 Sample preparation for label-free LC-MS analysis

Sample preparation and label-free and LC-MS analysis were performed by our collaborators Michael Henry and Paula Meleady in the National Institute for Cellular Biology in Dublin. Cell pellets were lysed with 300uL of lysis buffer consisting 300uL of lysis buffer (7 M urea, 2M thiourea, 30 mM Tris, 4% CHAPS, pH 8.5), and then homogenised by carefully passing the samples through a 20 gauge needle 5 times (avoid sample frothing). Samples were left on a shaker for 1 hr at room temp to allow extraction to take place, and then centrifuged at approximately 10000 g for 15 minutes at 4°C to remove insoluble material. The supernatant was removed and protein concentrations determined using the thiourea-compatible Quick Start Bradford assay (Bio-Rad) and stored at -80°C until required for use.

To remove any trace of detergent, 100µg of sample were cleaned using the Ready Prep 2-D clean up kit (Bio-Rad). The protein pellet was resuspended in detergent free buffer of 6M Urea, 2M thiourea, 10mM tris, pH 8.0 and protein concentration was determined again using Quick Start Bradford assay (Bio-Rad). Protein concentration was then adjusted to 2µg/µL using the same buffer. Proteins were then digested in the following way. Ten micrograms of protein sample (5µl) were added to 35.4 µL of 50 mM ammonium bicarbonate and 0.5µl of 1% Protease MAX™ Surfactant (Promega).

Reduction was performed by adding DTT to a final concentration of 5 mM, at 56 °C for 20 min, and allowed to cool to room temperature for approximately 20 min. Samples were alkylated by adding Iodoacetamide 15mM to a final concentration and then incubated for 15 minutes in the dark at room temperature. Digestion with sequence grade Lys-C (Promega) was carried out at a ratio of 1:20 (w/w) Lys-C:Protein at 37 °C for 6 h, followed by a second digestion with sequence grade Trypsin (Promega) at a ratio of 1:25 (w/w) Trypsin:Protein overnight at 37 °C. To stop the digestion, 2µL of a 50% TFA/50% water solution was added. Samples were then cleaned up using Pep-

clean C18 spin columns (Thermo Fisher Scientific), dried under a vacuum and stored at −80 °C. Prior to mass spectrometry analysis dried peptides were resuspended in 50 µL of 0.1% trifluoroacetic acid (TFA) in 2% acetonitrile (ACN), vortexed and sonicated to ensure an even suspension.

7.2.7 Label-free LC-MS quantitative profiling

Nano LC–MS/MS analysis was carried out using an Ultimate 3000 nano LC system (Dionex) coupled to a hybrid linear ion trap/Orbitrap mass spectrometer (LTQ Orbitrap XL; Thermo Fisher Scientific). Five microlitres of digest were loaded onto a C18 trap column (C18 PepMap, 300 µm ID × 5 cm, 5 µm particle size, 100 Å pore size; Thermo Fisher) and desalted for 10 min using a flow rate of 25 µL/min in 0.1% TFA. The trap column was then switched online with the analytical column (PepMap C18, 75 µm ID × 500 mm, 3 µm particle and 100 Å pore size; Thermo Fisher) and peptides were eluted with the following binary gradients of solvent A and B: 0–25% solvent B in 240 minutes and 25–50% solvent B in a further 60 minutes, where solvent A consisted of 2% ACN and 0.1% formic acid in water and solvent B consisted of 80% ACN and 0.08% formic acid in water. Column flow rate was set to 300 nL/min. Data were acquired with Xcalibur software, version 2.0.7 (Thermo Fisher Scientific). The mass spectrometer was operated in data-dependent mode and externally calibrated. Survey MS scans were acquired in the Orbitrap in the 300–2000 m/z range with the resolution set to a value of 60,000 at m/z 400. Up to seven of the most intense ions (1+, 2+ and 3+) per scan were CID fragmented in the linear ion trap. A dynamic exclusion window was applied within 40 s. All tandem mass spectra were collected using a normalised collision energy of 35%, an isolation window of 3 m/z, and one microscan.

7.2.8 LC–MS Data analysis

Label-free LC–MS analysis was carried out using Progenesis label-free LC–MS software version 3.1 (NonLinear Dynamics), essentially as recommended by the manufacturer (see www.nonlinear.com for further background to alignment, normalisation, calculation of peptide abundance, etc.). The software processed the raw data in two steps. Firstly each sample run was subjected to alignment which involved

aligning the data based on the LC retention time of each sample; this allows for any drift in retention time giving an adjusted retention time for all runs in the analysis. The sample run that yielded most features (i.e. peptide ions) was used as the reference run, to which retention time of all of the other runs were aligned and peak intensities were normalised. The Progenesis peptide quantification algorithm calculates peptide abundance as the sum of the peak areas within its isotope boundaries. Each abundance value is then transformed to a normalised abundance value by applying a global scaling factor. Protein abundance was calculated as the sum of the abundances of all peptide ions which have been identified as coming from the same protein. A number of criteria were used to filter the data before exporting the MS/MS output files to MASCOT (www.matrixscience.com) for protein identification; peptide features with ANOVA < 0.05 between experimental groups, mass peaks (features) with charge states from +1 to +3, and greater than 3 isotopes per peptide. All MS/MS spectra were exported from Progenesis software as a MASCOT generic file (mgf) and used for peptide identification with MASCOT (version 2.2) searched against the UniProtKB–SwissProt database (taxonomy, Mammals). The search parameters used were as follows: peptide mass tolerance set to 20 ppm, MS/MS mass tolerance set at 0.5 Da; up to two missed cleavages were allowed, carbamidomethylation set as a fixed modification and methionine oxidation set as a variable modification. Only peptides with ion scores of 30 and above were considered and re-imported back into Progenesis LC–MS software for further analysis. An ANOVA *p*-value of ≤0.05 between experimental groups was used as cut-off. Protein function and network analysis was performed using IPA Ingenuity software (Qiagen).

7.3 Results

7.3.1 Preliminary shRNA downregulation test

MTHFD1L is an essential gene (Momb *et al*. 2013) and a severe downregulation could cause cell death, hence, IPTG inducible vectors were used to allow a certain control on *MTHFD1L* downregulation. Out of five shRNA constructs, we selected the two constructs that showed the best *MTHFD1L* downregulation. Two different shRNA

constructs were employed to combine the final proteomic data for the purpose of excluding possible off-target effects.

Thus, a preliminary test was performed to determine the best concentration of IPTG and to select two shRNA constructs for the proteomic analysis. RT-qPCR results are shown in Figure 7.1. Normal non-transduced HEK293 cells and the non-target shRNA1 control showed a similar value confirming that the latter does not affect *MTHFD1L* expression. shRNA4 did not change *MTHFD1L* mRNA level, while shRNA5 and shRNA6 showed a moderate downregulation. shRNA3 and shRNA7 produced the best downregulation effect, ~80% and ~75% respectively, and therefore they were selected for the proteomic analysis. As expected, increasing the IPTG dosage leads to a better *MTHFD1L* downregulation. Both the inducible vectors (3xLacO and 1xLacO) showed a limited leakage with no IPTG induction (~20%) and a tight control of the regulative effect.

7.3.2 Single clone selection for downregulation and overexpression experiments

Due to the random integration of vectors into the host genome, varying levels of target gene knockdown may be seen with different resistant colonies. Testing a number of colonies allows the optimal degree of knockdown to be determined. Two clones for each cell line were then tested for *MTHFD1L* expression by RT-qPCR. Results are shown in Figure 7.2. shRNA 7-4 and shRNA 3-3 single clones which presented the best downregulation levels were selected for the proteomic analysis (Figure 7.2a and b). Regarding the upregulation experiment, 1L-1 single clone was selected for the proteomic analysis as it showed the highest expression of MTHFD1Lopt sequence (Figure 7.2c).

7.3.3 MTHFD1L overexpression and downregulation experiments

MTHFD1L Overexpression Experiment

Prior to carry out LC-MS analysis, MTHFD1L expression level was verified by RT-qPCR and Western Blot (Figure 7.3a). RT-qPCR confirmed a high expression level of MTHFD1Lopt sequence (Figure 7.3a) while Western Blot showed a 2.89 fold overexpression of MTHFD1L protein. In the MTHFD1L overexpression experiment, LC-MS data analyzed with Progenesis3.1 software reported a total of 92 significant

proteins of which 17 upregulated (Table 7.2) and 75 downregulated (Table 7.3). MTHFD1L was first in the list of upregulated proteins (7.04 fold change) proving the effectiveness of the experiment.

MTHFD1L Knockdown Experiment

MTHFD1L downregulation was also confirmed by RT-qPCR and Western Blot (Figure 7.3b). LC-MS analysis resulted in 83 significant proteins for shRNA3-3 sample and 79 for shRNA7-4 sample, while the two samples shared 50 proteins. MTHFD1L protein showed a 31.62 fold downregulation in shRNA3-3 and a 10.51 fold downregulation in shRNA7-4 sample confirming RT-qPCR and Western Blot results. Sample datasets were combined and all the proteins listed in only one group (62) or displaying opposite regulation (2) were discarded. Among the resulting 48 proteins that were shared between the two lines, 18 were upregulated (Table 7.4) and 30 were downregulated (Table 7.5).

Six proteins were present in both MTHFD1L overexpression and knockdown lists (Table 7.6) and showed opposite trend of expression as expected. Specifically, all these proteins showed an opposite regulation compared with MTHFD1L expression. Many of the proteins affected by MTHFD1L expression are involved in nucleic acid metabolism and DNA replication, recombination and repair (Table 7.7). Almost all these proteins seem to follow a clear expression trend, they are downregulated when MTHFD1L level is high and upregulated when MTHFD1L level is low. Protein networks have been drawn using IPA Ingenuity software and they are shown in Figures 7.4 and 7.5. Four proteins, whose expression is associated with DNA damage prevention, are an exception to this trend (CRYZ, HSPB1, PRKDC, and SMARCA1). According to the literature an inhibition of these proteins should increase DNA damage.

IPA analysis also showed that MTHFD1L overexpression data significantly match with an increased signal of cell death (necrosis, apoptosis, apoptosis of tumour cell line, neural cell death and apoptosis of neurons). In particular the downregulation of 7 proteins (PRDX3, PRDX2, MSH2, XRCC5, HSD17B10, XRCC6 and NEFL) is associated with an increased apoptotic signal in neurons. Among these proteins there is neurofilament, light polypeptide (NEFL) that was one of the proteins most affected by

MTHFD1L expression. It showed a 27.88 fold downregulation when MTHFD1L levels was high and a 9.29 fold upregulation when MTHFD1L level was low.

7.4 Discussion

In this chapter, the role of altered MTHFD1L expression in disease was assessed by a proteomic analysis. Such analysis has been accurately designed to produce different datasets that can be compared to each other in order to strengthen the final outcome. Single clone cell lines were sub-cultured to guarantee a stable MTHD1L expression among cell population. In addition to this, two different inducible shRNA constructs have been employed for protein inhibition. MTHFD1L is essential for cell viability, therefore it was important to downregulate it enough to induce appreciable proteomic changes, but not too much to utterly damage the cell stability. shRNA3-3 and shRNA7-4 datasets were combined and 64 proteins were omitted as appearing in only one dataset or showing opposite regulation. These omitted proteins presumably were the result of off-target shRNA regulation. MTHFD1L downregulation and upregulation results were compared with the aim to identify proteomic patterns directly related to MTHFD1L expression.

Interestingly, many proteins affected by MTHFD1L expression are involved in nucleic acid metabolism and DNA replication, recombination and repair (Table 7.7). Almost all these proteins are downregulated when MTHFD1L is overexpressed while they are upregulated when MTHFD1L is inhibited. An example of proteins displaying this expression trend is given by the DNA replication licensing factor MCM proteins. They form a complex (MCM2-7) with replicative helicase activity essential for 'once per cell cycle' DNA replication initiation and elongation in eukaryotic cells (Tsuji *et al.* 2006). Another protein in this list (Table 7.7) is the folate related TYMS, which ubiquitously localizes to the nucleus, cytoplasm and mitochondria, and it is responsible for the synthesis of thymidylate (Anderson *et al.* 2012). The enzymes that constitute the *de novo* thymidylate pathway include TYMS, DHFR and SHMT1. These enzymes, along with other proteins like MTHFD1, form a nuclear complex that is associated with the nuclear lamina during S and G_2/M phases (Anderson *et al.* 2012). This complex provides thymidylate for DNA replication and repair, and it is important to prevent

uracil accumulation in the nuclear DNA (MacFarlane *et al.* 2011). SHMT1 serves as a scaffold protein that is essential for complex formation. In 2012 Anderson *et al.* investigated further the nature of this nuclear complex seeking for the involvement of other proteins. By using co-immunoprecipitation experiments and tandem affinity purification they identified a number of proteins associated with SHMT1 and the nuclear complex. Intriguingly, in addition to TYMS, three of these proteins (PCNA, LMNA and LMNB1) also appeared in our proteomic experiment and they followed the same regulation trend, opposite to MTHFD1L expression direction (Table 7.7). Proliferating cell nuclear antigen (PCNA), which acts as a processivity factor for DNA replication (Shumaker *et al*. 2008), was previously identified as a SHMT interacting partner (Woeller *et al.* 2007, Li *et al.* 2004). PCNA was shown to co-precipitate with SHMT1, TYMS and DHFR in nuclear extracts but not in cytoplasmic extracts (Anderson *et al.* 2012). Both Lamin A/C (LMNA) and Lamin B1 (LMNB1) also co-precipitate with SHMT1. LMNB1, which is a component of the nuclear lamina, exhibited co-localization with TYMS, DHFR and SHMT1 in filament-like linear structures during the S and G_2/M phase of the cell cycle. LMNB1 directly binds SHMT1 anchoring the complex to the nuclear lamina (Anderson *et al.* 2012).

Our findings suggest that MTHFD1L levels might be inversely proportional to the levels of the proteins involved in the nuclear thymidylate biosynthesis complex level. The stability of this complex is important to prevent uracil misincorporation into genomic DNA (MacFarlane *et al.* 2011). Thymidine nucleotides differ from the other nucleotides in that they are not crucial during DNA replication. DNA polymerases typically fail to distinguish between dTTP and dUTP during DNA replication and repair and can incorporate dUTP into DNA when dTTP is limiting (Stover *et al*. 2012). Moreover, while the synthesis of cytosine and purine nucleosides occurs in the cytoplasm, the salvage and folate-dependent *de novo* synthesis of thymidylate, catalyzed by thymidine kinase (TMPK) and TYMS respectively, occurs in the nucleus at sites of DNA synthesis (Chen *et al*. 2010 and Anderson et *al.* 2012). Cellular dTTP pools are normally maintained at very low levels and when they are depleted or expanded DNA integrity and human health can be affected (Samsonoff *et al.* 1997). There is strong evidence that dTTP is synthesized "on-site and on-demand", and a diminished dTTP synthesis results in dU accumulation in DNA, causing genomic instability through futile cycles of DNA synthesis and repair (Blount *et al* 1997).

Elevated uracil misincorporation and chromosome breaks contribute to the increased risk of cancer and cognitive defects associated with folate deficiency in humans (Blount *et al* 1997). Indeed, the inhibition of SHMT1, which is an essential scaffold for the formation of the thymidylate biosynthesis complex, leads to increased uracil misincorporation in DNA, increased risk of intestinal cancer (MacFarlane *et al.* 2011), and NTDs (Beaudin *et al.* 2011) in mice. In our analysis, MTHFD1L overexpression seems to inhibit the presence of the proteins belonging to the nuclear thymidylate biosynthesis complex. This should result in an increased uracil accumulation in the DNA. Thus, this mechanism provides an explanation to the association of increased MTHFD1L levels with NTDs (Parle-McDermott *et al.* 2009) and cancer (Sugiura *et al*. 2004, Jain *et al*. 2012). In agreement with this mechanism, also proteins involved in the repair of the double-strand breaks (DSBs) showed to be downregulated by MTHFD1L overexpression. The XRCC5 and XRCC6 proteins form a dimer that is involved in stabilizing broken DNA ends and bringing them together (Tuteja *et al.* 1994). ATP-dependent RNA helicase DDX1 and DDX11 also play a role in the RNA clearance DSBs, thereby facilitating the template-guided repair of transcriptionally active regions of the genome (Li *et al.* 2008). Defects in the cellular response to DSBs underpin many human diseases, including cancer and neurodegeneration (McKinnon *et al.* 2007).

The effect of MTHFD1L expression on the nuclear thymidylate biosynthesis complex and other proteins involved in DNA replication and repair is probably linked to the formate production within mitochondria. It appears that the majority of 1C units for cytoplasmic processes are derived from mitochondrial formate produced by MTHFD1L (Momb *et al*. 2013). Moreover, the identification of MTHFD1 (the cytoplasmic paralogue of MTHFD1L) as a component of the nuclear thymidylate biosynthesis complex indicates that formate can be an important source of nuclear methylene-THF for *de novo* thymidylate biosynthesis (Anderson *et al*. 2012). Another point to consider is the toxicity of formate to respiration *via* inhibition of cytochrome *c* oxidase, which was proven in an early study (Nicholls 1975). More recently, Kapur *et al*. (2007) reported that a formate concentration as low as 1 mM induced neurotoxicity in rat hippocampal brain slices. Here, we corroborated those findings by showing that cytochrome *c* oxidase subunit 7A2 (COX7A2) is downregulated by MTHFD1L overexpression (Table 7.3). Thus, an impaired production of formate may affect the cell at different levels. MTHFD1L, along with the mitochondrial proteins SHMT2 and

MTHFD2, were correlated with the growth rates of human cancer cell lines (Jain *et al*. 2012). Thus, in healthy cells, an increasing formate production may be a signal for uncontrolled cell proliferation that triggers defensive mechanisms like the inhibition of nuclear thymidylate biosynthesis and DNA replication and repair. Augmented uracil misincorporation and DNA breaks may eventually lead to apoptosis.

This possible explanation is in concordance with the increased signal of cell death reported upon MTHFD1L overexpression. Specifically, the downregulation of PRDX2 (Boulos *et al.*2007), PRDX3 (Hattori *et al.* 2003), MSH2 (Francisconi *et al.* 2006), XRCC5, XRCC6 (Gu *et al.* 2000), HSD17B10 (Lustbader *et al.* 2004) and NEFL (Lee *et al.* 1994, Cluskey *et al*. 2001) have been associated with increased apoptosis in neurons.

Among these proteins NEFL showed the strongest response to MTHFD1L upregulation and overexpression. Neurofilament medium polypeptide (NEFM) was also highly downregulated (5.19 fold change) by the overexpression of MTHFD1L. Neurofilaments comprise the axoskeleton and functionally maintain neuronal calibre (Lalonde *et al*. 2003). They may also play a role in intracellular transport to axons and dendrites. In mouse, depletion of NEFL causes massive selective motor neuron death (Lee *et al.* 1994, Cluskey *et al*. 2001). These proteins are commonly used as a biomarker of neuronal damage and they have been proposed to be involved in the onset and progression of amyotrophic lateral sclerosis, mainly characterized by motor neuron atrophy and paresis (Lalonde *et al*. 2003). Another protein associated with neurodegenerative disease is the Aβ-binding alcohol dehydrogenase HSD17B10, which has been proven to directly link Aß to mitochondrial toxicity in Alzheimer's Disease (AD) (Lustbader *et al.* 2004). In mice, XRCC5 and XRCC6 deficiency results in dramatically increased death of neurons in developing embryos and adults (Gu *et al.* 2000). Thus, our protein expression results might provide a link for the former genetic association of *MTHFD1L* with AD (Naj *et al*., 2010) and NTD risk (Parle-McDermott *et al*. 2009).

SMARCA1, CRYZ, HSPB1, and PRKDC, which are associated with DNA damage prevention, showed to be downregulated by MTHFD1L knockdown. SMARCA1 protein, though present in the control sample, was totally depleted in shRNA3-3 and

shRNA7-4 samples. This protein is annotated as probable global transcription activator of SNF2L1 that is a chromatin remodeling gene expressed in diverse tissues, cancers, and derived cell lines, contributes to the chromatin remodeling complex and facilitates transcription (Ye *et al*. 2009). A reduction of SNF2L1 protein level induced growth inhibition, DNA damage and phosphorylation of checkpoint proteins and marked apoptosis. This effect was detected in cancer cells but not in normal cells (Ye *et al*. 2009). SMARCA1 may be also involved in brain development by regulating En-1 and En-2. Quinone oxidoreductase CRYZ has been shown to decrease oxidative damage of genomic DNA in mouse mammary gland (Krishnamurthy *et al*. 2012). Heat shock protein beta-1 (HSPB1) has been shown to provide neuroprotection against ischemic injury (Stetler *et al*. 2012). Its phosphorylation is required to decrease DNA damage in primary rat cortical neurons (Stetler *et al*. 2012). Inactivation of PPKDC in human cells increases damage of DNA enhanced by ionizing radiation (Dziegielewski *et al.* 2002). Taken together, these available data suggest that the downregulation of MTHFD1L and the consequent decrease of formate may cause DNA damage. This hypothesis is supported by the evidence of RNRR1 upregulation in MTHFD1L knocked-down samples (Table 7.7). Ribonucleoside-diphosphate reductase (RNRR1 protein, *RRM1* gene) catalysis the production of dADP, dGDP, dCDP and dUDP from their corresponding NDPs, whereas the synthesis of dTDP is unique because it requires TMPK (Hu *et al.* 2012). RNRR1 is recruited to sites of DNA damage and without the coordination of TMPK can lead to dUTP incorporation during repair (Hu *et al.* 2012).

In summary, we demonstrated that MTHFD1L levels affect the nuclear thymidylate biosynthesis complex and other proteins involved in DNA replication and repair. We proposed that an increased formate production may be a signal for uncontrolled cell proliferation that triggers defensive mechanisms like the inhibition of nuclear thymidylate biosynthesis and DNA replication and repair. On the other hand, a downregulation of MTHFD1L has been linked with an increased DNA damage.

In conclusion, these data suggest that nucleotide biosynthesis and DNA replication and repair are sensitive to changes in the expression level of mitochondrial MTHFD1L, which is likely to correspond to a change in formate production from the mitochondria. Formate supply may act as a ‘sensor’ to ensure that there is a balanced supply of nucleotides for DNA replication and repair and to prevent uracil misincorporation. This

findings further highlight the importance of MTHFD1L levels and, consequently, of formate abundance inside mitochondria, the cytoplasm and the nucleus.

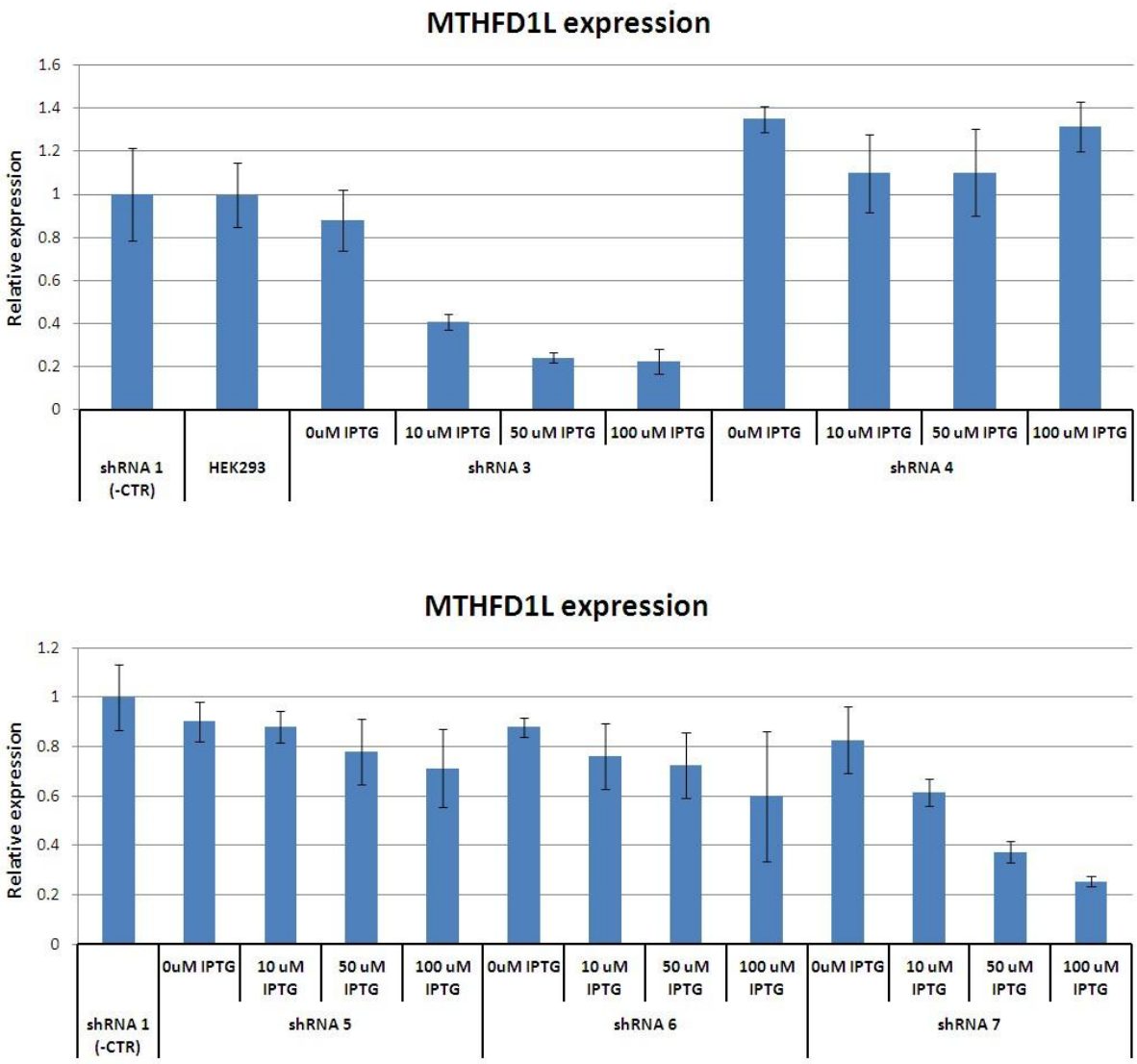


Figure 7.1 Relative *MTHFD1L* gene expression in shRNA transduced cell lines. *MTHFD1L* gene expression was analysed in shRNA transduced samples and compared to the negative control. shRNA 3 and shRNA 7 samples produced the highest *MTHFD1L* downregulation. Relative expression was calculated using GUS as reference gene and normalized against shRNA1 (-CTR) sample.

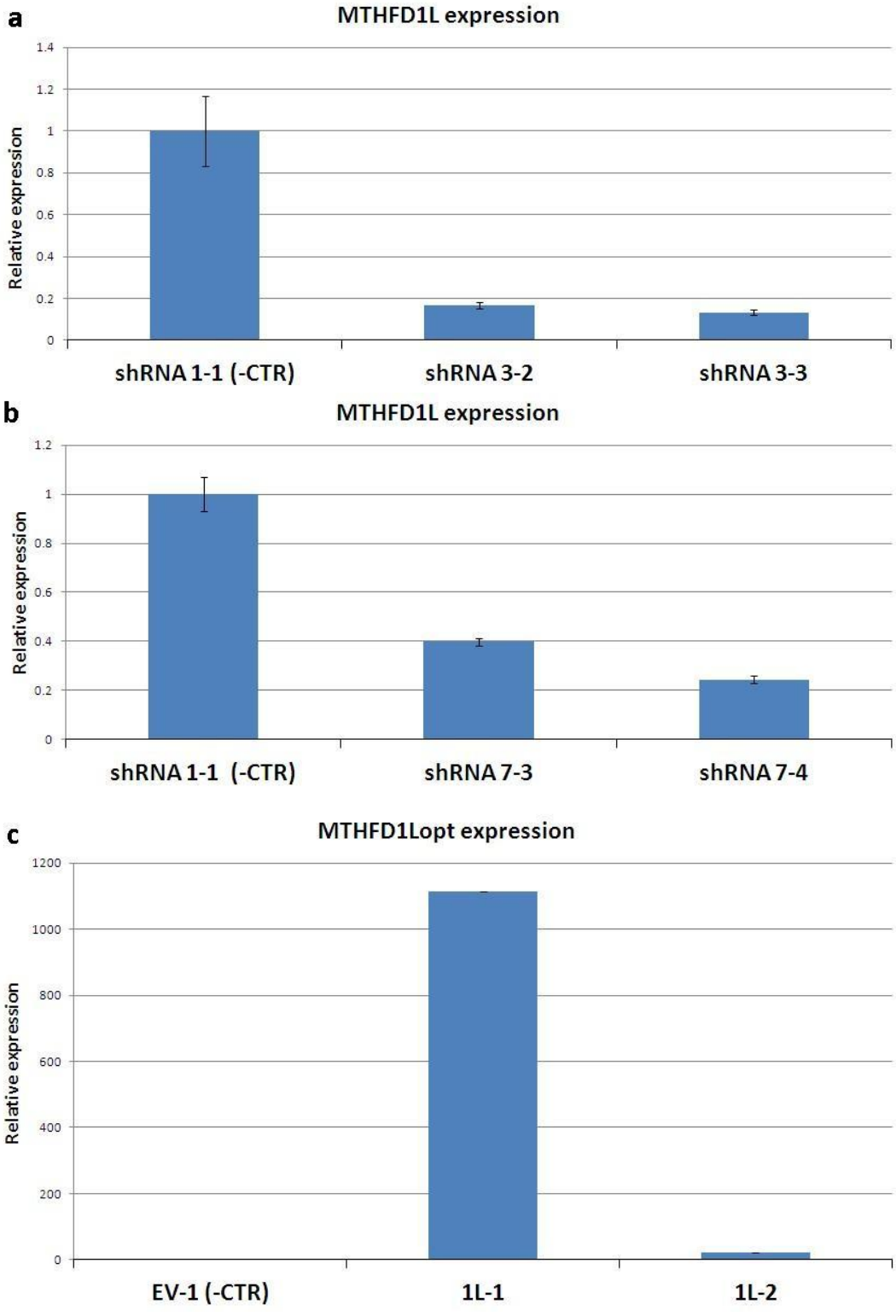


Figure 7.2 Single clone selection for proteomic analysis.

MTHFD1L gene expression was analysed in single clone samples and compared to the respective negative control sample. **(a)** shRNA 3-2 and shRNA3-3 showed similar levels of downregulation and the latter was selected for the proteomic experiment. **(b)**

shRNA 7-4 presented a better downregulation than shRNA 7-3, thus it was selected for the proteomic experiment. **(c)** 1L-1 upregulation was about 100 times higher than 1L-2 one. Therefore, 1L single clone was selected for the proteomic analysis. Relative expression was calculated using GUS as reference gene and normalized against the respective negative control sample.

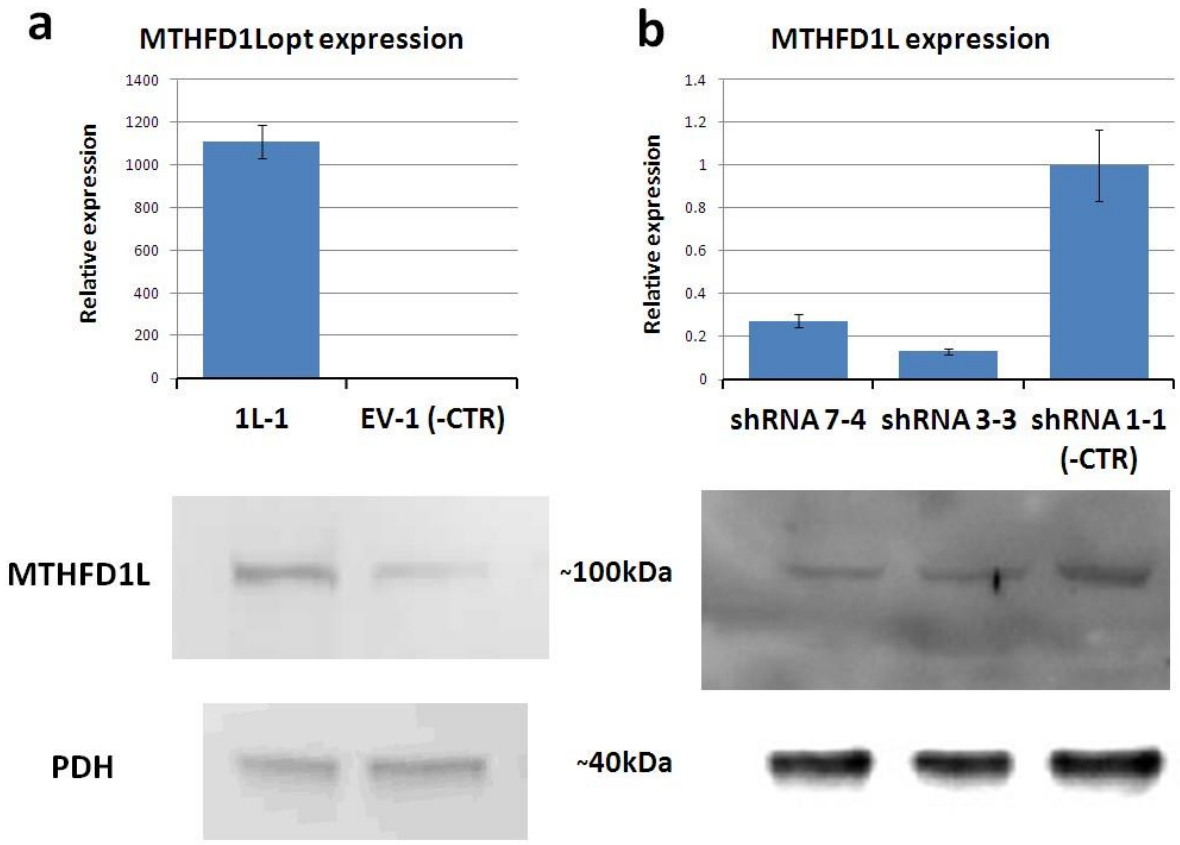


Figure 7.3 Validation of MTHFD1L expression for proteomic analysis.

MTHFD1L expression was analysed at mRNA and protein level in samples for overexpression **(a)** and downregulation **(b)** experiments. Relative expression was calculated using GUS as reference gene and normalized against shRNA1 (-CTR) sample. In Western blotting analysis samples were probed with antibodies against both MTHFD1L and PDH subunit E1 alpha used as reference protein and mitochondrial marker. MTHFD1L expected size is 102 kDa after cleavage of the mitochondrial targeting sequence, while PDH subunit E1 alpha expected size is 41 kDa.

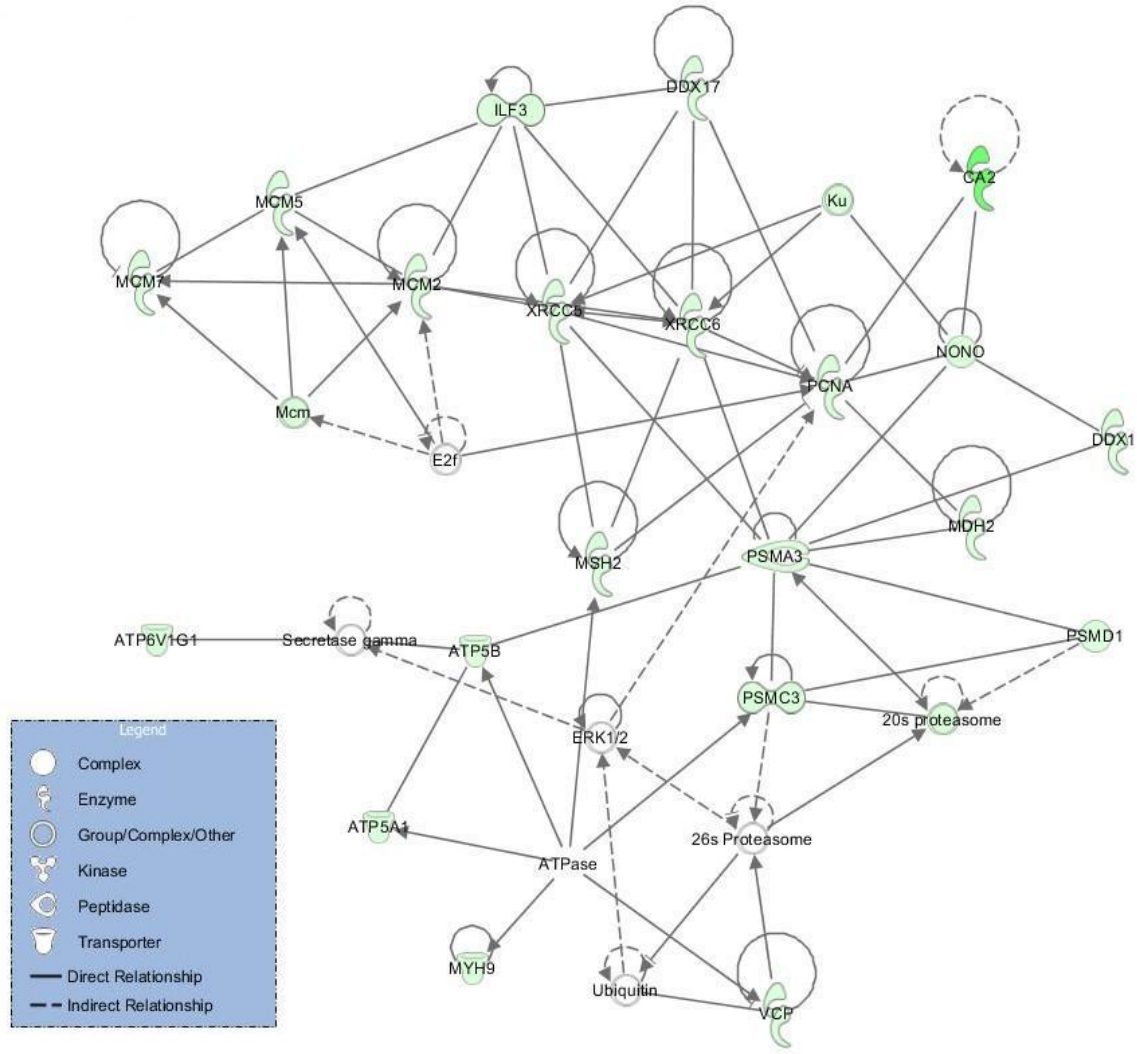


Figure 7.4 Protein network designed from MTHFD1L overexpression data.

Protein network was drawn using IPA Ingenuity software with proteins classified as involved in nucleic acid metabolism, small molecule biochemistry and DNA replication, recombination and repair. Colour legend: light blue = mild downregulation, green = strong downregulation, pink = mild upregulation, red = strong upregulation. Uncoloured proteins are not affected by MTHFD1L expression and they were used only to complete the network.

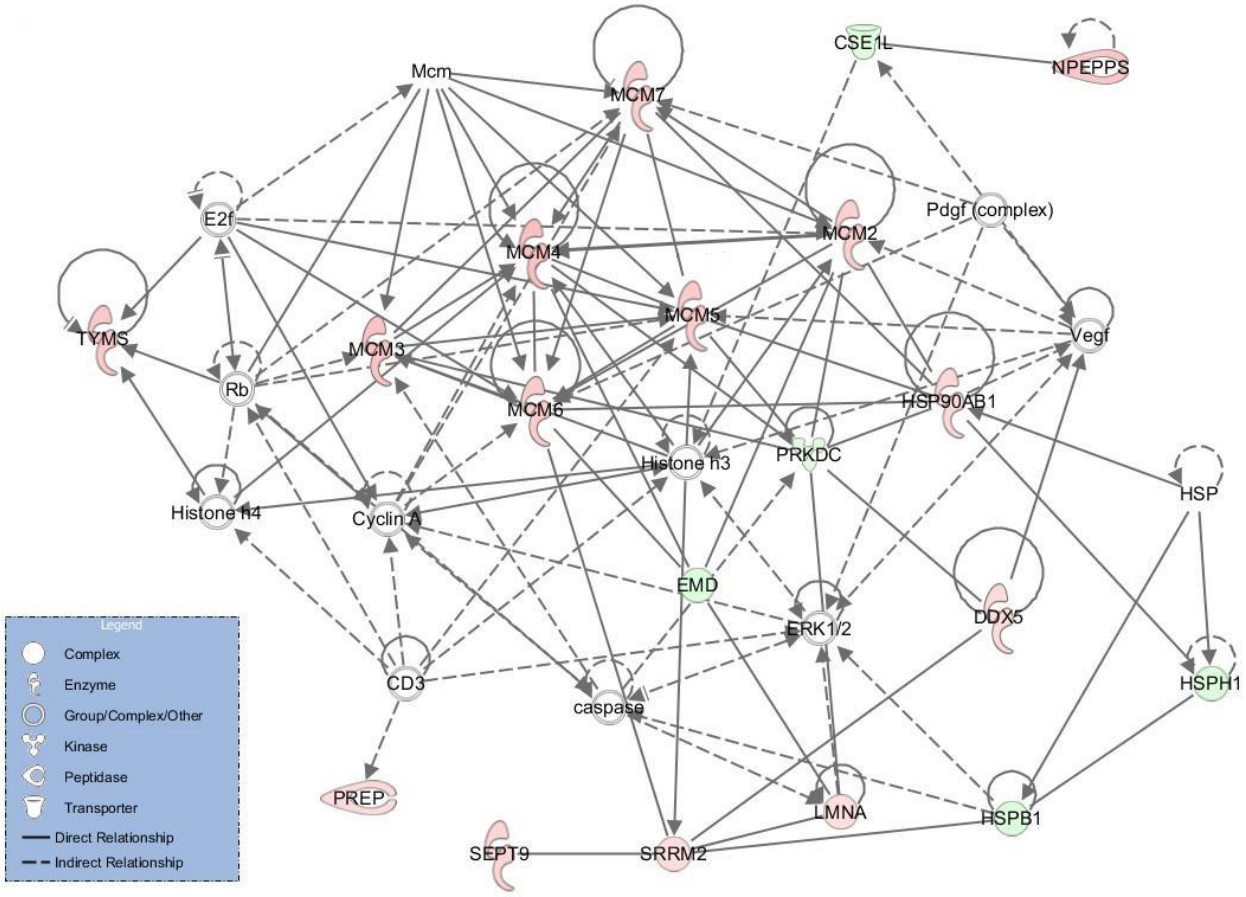


Figure 7.5 Protein network designed from MTHFD1L downregulation data.

Protein network was drawn using IPA Ingenuity software with proteins classified as involved in Developmental disorders, Hereditary disorders and DNA replication, recombination and repair. Colour legend: light blue = mild downregulation, green = strong downregulation, pink = mild upregulation, red = strong upregulation. Uncoloured proteins are not affected by MTHFD1L expression and they were used only to complete the network.

Table 7.1 shRNA constructs used in MTHFD1L downregulation experiment.

Name[a]	Cat. no.	TRC number	Target	Vector	Validated[b]	Vector map
shRNA 1 (-CTR)	SHC002V		None[c]	pLK0.1	Yes	Appendix E
shRNA 3	SHCLNV-NM_015440	TRCN0000045400	MTHFD1L	pLKO_IPTG_1xLacO	Yes	Appendix F
shRNA 4	SHCLNV-NM_015440	TRCN0000045398	MTHFD1L	pLKO_IPTG_3xLacO	No	Appendix G
shRNA 5	SHCLNV-NM_015440	TRCN0000045399	MTHFD1L	pLKO_IPTG_3xLacO	Yes	Appendix G
shRNA 6	SHCLNV-NM_015440	TRCN0000217958	MTHFD1L	pLKO_IPTG_1xLacO	No	Appendix F
shRNA 7	SHCLNV-NM_015440	TRCN0000229790	MTHFD1L	pLKO_IPTG_3xLacO	No	Appendix G

[a] Name used in this study

[b] Some shRNAs have already been validated by Sigma-Alright

[c] The negative control shRNA is design to not target any known mammalian gene.

Table 7.2 List of proteins with increased expression following overexpression of MTHFD1L in HEK293 cells.

Accession No.[a]	Gene Name[b]	Protein Description	Pepti-des[c]	Score[d]	Anova *p*-value	Fold change
Q6UB35	MTHFD1L	Monofunctional C1-tetrahydrofolate synthase, mitochondrial	6	238.52	1.11E-05	**7.04**
P13797	PLS3	Plastin-3	3	147.45	0.01	**2.55**
Q9Y617	PSAT1	Phosphoserine aminotransferase	3	139.75	0.01	**1.57**
P08107	HSPA1A	Heat shock 70 kDa protein 1A/1B	2	102.14	9.43E-05	**3.65**
P50225	SULT1A1	Sulfotransferase 1A1	2	68.05	9.41E-05	**3.18**
P22695	UQCRC2	Cytochrome b-c1 complex subunit 2, mitochondrial	1	63.01	0.02	**1.32**
O00410	IPO5	Importin-5	1	61.01	0.02	**1.19**
O95373	IPO7	Importin-7	1	53.84	0.02	**1.62**
P41252	IARS	Isoleucine--tRNA ligase, cytoplasmic	1	53.79	0.03	**1.27**
P39687	ANP32A	Acidic leucine-rich nuclear phosphoprotein 32 family member A	1	46.77	0.02	**1.63**
O43175	PHGDH	D-3-phosphoglycerate dehydrogenase	1	42.33	0.01	**1.25**
P08195	SLC3A2	4F2 cell-surface antigen heavy chain	1	41.43	0.03	**2.19**
P50991	CCT4	T-complex protein 1 subunit delta	1	36.66	0.05	**1.22**
P09661	SNRPA1	U2 small nuclear ribonucleoprotein A'	1	36.33	0.03	**1.19**
Q5QNW6	HIST2H2BF	Histone H2B type 2-F	1	36.22	0.05	**2.91**
P00338	LDHA	L-lactate dehydrogenase A chain	1	33.91	0.04	**1.47**
Q99832	CCT7	T-complex protein 1 subunit eta	1	33.55	0.05	**1.24**

[a] Uniprot accession number from MASCOT search of UniProtKB-SwissProt, taxonomy Homo sapiens.

[b] Official recommended gene name taken from UniProtKB-SwissProt.

[c] Number of peptides used to identify the protein.

[d] The Mascot Score is a statistical score for how well the experimental data match the database sequence. It corresponds to the summed score, calculated with an algorithm, for the individual peptides.

Table 7.3 List of proteins with decreased expression following overexpression of MTHFD1L in HEK293 cells.

Accession No.[a]	Gene Name[b]	Protein Description	Pepti-des[c]	Score[d]	Anova *p*- value	Fold change
P07197	NEFM	Neurofilament medium polypeptide	4	187.29	1.38E-05	**5.19**
P12004	PCNA	Proliferating cell nuclear antigen	3	166.36	3.32E-04	**1.56**
P04181	OAT	Ornithine aminotransferase, mitochondrial	2	132.83	1.21E-03	**1.39**
P12236	SLC25A6	ADP/ATP translocase 3	2	127.02	1.20E-03	**1.71**
P00918	CA2	Carbonic anhydrase 2	3	108.65	9.60E-08	**17.91**
P13796	LCP1	Plastin-2	2	99.6	5.07E-04	**1.48**
P49327	FASN	Fatty acid synthase	2	93.3	0.02	**1.28**
Q12906	ILF3	Interleukin enhancer-binding factor 3	2	89.26	0.02	**1.32**
P08670	VIM	Vimentin	2	81.72	0.01	**1.54**
P37802	TAGLN2	Transgelin-2	1	79.73	0.01	**1.29**
Q9NVI7	ATAD3A	ATPase family AAA domain-containing protein 3A	2	75.3	7.28E-04	**1.32**
P00387	CYB5R3	NADH-cytochrome b5 reductase 3	1	73	0.03	**1.55**
P37837	TALDO1	Transaldolase	1	72.41	0.05	**1.53**
Q99714	HSD17B10	3-hydroxyacyl-CoA dehydrogenase type-2	1	71.4	0.02	**1.43**
Q15102	PAFAH1B3	Platelet-activating factor acetylhydrolase IB subunit gamma	1	70.39	9.22E-04	**1.31**
P62318	SNRPD3	Small nuclear ribonucleoprotein Sm D3	1	69.43	1.73E-03	**1.24**
Q16658	FSCN1	Fascin	2	69.2	0.01	**1.52**
Q99623	PHB2	Prohibitin-2	1	67.77	0.02	**1.44**
P33992	MCM5	DNA replication licensing factor MCM5	1	65.91	0.04	**1.89**
P07196	NEFL	Neurofilament light polypeptide	2	64.7	6.03E-05	**27.88**
P25705	ATP5A1	ATP synthase subunit alpha, mitochondrial	1	62.67	0.05	**1.27**
P06576	ATP5B	ATP synthase subunit beta, mitochondrial	1	61.71	0.05	**1.49**
P20700	LMNB1	Lamin-B1	1	61.17	0.05	**1.54**
Q96CX2	KCTD12	BTB/POZ domain-containing protein KCTD12	1	59.69	4.60E-07	**12.10**
P49736	MCM2	DNA replication licensing factor MCM2	1	58.75	0.01	**1.92**
Q13162	PRDX4	Peroxiredoxin-4	1	57.14	0.01	**1.33**
P40926	MDH2	Malate dehydrogenase, mitochondrial	1	55.73	0.01	**1.33**
A8MWD9		Small nuclear ribonucleoprotein G-like protein	1	55.65	0.04	**1.74**
A6NHG4	DDTL	D-dopachrome decarboxylase-like protein	1	50.41	0.03	**2.43**
Q7Z4W1	DCXR	L-xylulose reductase	1	50.35	3.34E-03	**1.51**
Q9BVA1	TUBB2B	Tubulin beta-2B chain	1	50.13	7.63E-04	**1.77**
Q04760	GLO1	Lactoylglutathione lyase	1	49.27	0.02	**1.31**

P51649	ALDH5A1	Succinate-semialdehyde dehydrogenase, mitochondrial	1	47.57	2.13E-03	**1.36**
Q13838	DDX39B	Spliceosome RNA helicase DDX39B	1	46.94	0.02	**1.43**
P11142	HSPA8	Heat shock cognate 71 kDa protein	1	46.86	0.03	**1.49**
P12956	XRCC6	X-ray repair cross-complementing protein 6	1	45.17	0.03	**1.29**
P09211	GSTP1	Glutathione S-transferase P	1	44.65	0.03	**1.21**
Q13310	PABPC4	Polyadenylate-binding protein 4	1	43.58	0.02	**1.27**
Q15233	NONO	Non-POU domain-containing octamer-binding protein	1	42.99	0.05	**1.64**
Q13263	TRIM28	Transcription intermediary factor 1-beta	1	42.95	1.07E-03	**1.67**
P15311	EZR	Ezrin	1	42.42	0.02	**1.28**
P31939	ATIC	Bifunctional purine biosynthesis protein PURH	1	41.57	0.03	**1.26**
P00505	GOT2	Aspartate aminotransferase, mitochondrial	1	41.23	0.02	**1.36**
Q92499	DDX1	ATP-dependent RNA helicase DDX1	1	41.21	0.03	**1.39**
P14406	COX7A2	Cytochrome c oxidase subunit 7A2, mitochondrial	1	41.02	0.02	**2.13**
Q9BPW8	NIPSNAP1	Protein NipSnap homolog 1	1	40.92	0.01	**1.59**
Q99460	PSMD1	26S proteasome non-ATPase regulatory subunit 1	1	40.57	0.04	**1.41**
Q14152	EIF3A	Eukaryotic translation initiation factor 3 subunit A	1	39.86	0.01	**1.13**
P17980	PSMC3	26S protease regulatory subunit 6A	1	39.62	0.02	**1.41**
P33993	MCM7	DNA replication licensing factor MCM7	1	38.96	0.02	**1.52**
Q9Y277	VDAC3	Voltage-dependent anion-selective channel protein 3	1	37.74	2.92E-04	**1.85**
P14678	SNRPB	Small nuclear ribonucleoprotein-associated proteins B and B'	1	37.45	0.01	**1.33**
P32119	PRDX2	Peroxiredoxin-2	1	37.1	0.01	**1.59**
P30048	PRDX3	Thioredoxin-dependent peroxide reductase, mitochondrial	1	36.79	0.01	**1.30**
P46060	RANGAP1	Ran GTPase-activating protein 1	1	36.71	0.01	**1.32**
P07910	HNRNPC	Heterogeneous nuclear ribonucleoproteins C1/C2	1	36.13	0.02	**1.64**
Q9HC38	GLOD4	Glyoxalase domain-containing protein 4	1	35.45	0.03	**1.45**
P14625	HSP90B1	Endoplasmin OS=Homo sapiens	1	35.35	2.64E-04	**1.31**
P25788	PSMA3	Proteasome subunit alpha type-3	1	35.06	0.03	**1.30**
Q12931	TRAP1	Heat shock protein 75 kDa, mitochondrial	1	34.18	0.02	**1.39**
P31150	GDI1	Rab GDP dissociation inhibitor alpha	1	34.12	0.03	**1.70**
Q6ZU15	SEPT14	Septin-14	1	33.74	2.08E-03	**1.26**
Q92841	DDX17	Probable ATP-dependent RNA helicase DDX17	1	33.52	0.02	**1.39**
P55072	VCP	Transitional endoplasmic reticulum ATPase	1	33.33	0.03	**1.84**
P13010	XRCC5	X-ray repair cross-complementing protein 5	1	33.07	4.45E-03	**2.19**

P14868	DARS	Aspartate--tRNA ligase, cytoplasmic	1	33	0.05	**1.69**
P45880	VDAC2	Voltage-dependent anion-selective channel protein 2	1	32.88	0.01	**1.55**
P35579	MYH9	Myosin-9	1	32.64	0.04	**1.41**
P14174	MIF	Macrophage migration inhibitory factor	1	32.58	0.01	**1.61**
O95336	PGLS	6-phosphogluconolactonase	1	32.38	0.02	**1.47**
O43390	PGLS	Heterogeneous nuclear ribonucleoprotein R	1	32.02	0.03	**1.25**
Q2TAY7	SMU1	WD40 repeat-containing protein SMU1	1	31.33	0.02	**1.32**
P62249	RPS16	40S ribosomal protein S16	1	31.14	0.04	**1.40**
O75348	ATP6V1G1	V-type proton ATPase subunit G 1	1	30.7	0.02	**1.68**
P43246	MSH2	DNA mismatch repair protein Msh2	1	30.42	0.01	**1.49**

[a] Uniprot accession number from MASCOT search of UniProtKB-SwissProt, taxonomy Homo sapiens.

[b] Official recommended gene name taken from UniProtKB-SwissProt (for A8MWD9 there is no official gene name available so it has not been included).

[c] Number of peptides used to identify the protein.

[d] The Mascot Score is a statistical score for how well the experimental data match the database sequence. It corresponds to the summed score, calculated with an algorithm, for the individual peptides.

Table 7.4 List of proteins with increased expression following downregulation of MTHFD1L in HEK293 cells.

				shRNA 3-3				shRNA 4-7			
Accession No.[a]	**Gene Name[b]**	**Protein Description**	**Average Fold Change[c]**	**Pepti-des[d]**	**Score[e]**	**Anova *p*- value**	**Fold change**	**Pepti-des[d]**	**Score[e]**	**Anova *p*- value**	**Fold change**
P07196	NEFL	Neurofilament light polypeptide	**9.29**	6	339.63	8.37E-08	**17.23**	6	339.63	0.03	**1.36**
P55786	NPEPPS	Puromycin-sensitive aminopeptidase	**2.50**	5	235.19	5.78E-04	**2.75**	5	235.19	0.01	**2.24**
O75131	CPNE3	Copine-3	**2.02**	3	180.52	8.13E-05	**1.98**	3	180.52	2.08E-03	**2.06**
P33993	MCM7	DNA replication licensing factor MCM7	**2.20**	2	150.99	1.34E-03	**2.10**	2	150.99	2.70E-03	**2.30**
P25205	MCM3	DNA replication licensing factor MCM3	**2.69**	2	120.25	0.01	**2.12**	2	120.25	3.64E-03	**3.27**
P31939	ATIC	Bifunctional purine biosynthesis protein PURH	**1.81**	2	117.67	1.30E-03	**2.01**	2	117.67	0.01	**1.60**
P17844	DDX5	Probable ATP-dependent RNA helicase DDX5	**1.56**	2	111.99	9.14E-04	**1.83**	2	111.99	0.01	**1.30**
Q14566	MCM6	DNA replication licensing factor MCM6	**2.31**	2	102.86	0.01	**1.52**	2	102.86	5.99E-04	**3.10**
P02545	LMNA	Prelamin-A/C	**1.42**	1	94.36	0.01	**1.16**	1	94.36	0.02	**1.67**
Q92945	KHSRP	Far upstream element-binding protein 2	**2.28**	1	84.23	2.57E-03	**2.15**	1	84.23	4.39E-04	**2.41**
P51649	ALDH5A1	Succinate-semialdehyde dehydrogenase, mitochondrial	**1.94**	1	81.28	4.23E-03	**2.33**	1	81.28	0.02	**1.55**
P33992	MCM5	DNA replication licensing factor MCM5	**2.29**	1	74.25	0.01	**2.38**	1	74.25	4.52E-03	**2.20**
Q9H2H8	PPIL3	Peptidyl-prolyl cis-trans isomerase-like 3	**4.78**	1	71.88	4.43E-03	**4.16**	1	71.88	2.58E-03	**5.39**
P62993	GRB2	Growth factor receptor-bound protein 2	**2.01**	1	68.8	0.03	**1.92**	1	68.8	0.02	**2.09**
Q8TBC5	ZSCAN18	Zinc finger and SCAN domain-containing protein 18	**3.66**	1	67.6	0.03	**2.82**	1	67.6	4.91E-03	**4.49**
Q15084	PDIA6	Protein disulfide-isomerase A6	**1.83**	1	67.24	4.74E-03	**1.98**	1	67.24	0.01	**1.69**

O75436	VPS26A	Vacuolar protein sorting-associated protein 26A	**1.41**	1	66.59	3.49E-03	**1.46**	1	66.59	0.02	**1.36**
P23921	RRM1	Ribonucleoside-diphosphate reductase large subunit	**1.45**	1	65.83	0.03	**1.59**	1	65.83	0.02	**1.31**
Q9UQ35	SRRM2	Serine/arginine repetitive matrix protein 2	**1.50**	1	63.08	1.91E-03	**1.40**	1	63.08	0.04	**1.59**
P30520	ADSS	Adenylosuccinate synthetase isozyme 2	**1.51**	1	62.2	0.03	**1.48**	1	62.2	0.02	**1.53**
P48147	PREP	Prolyl endopeptidase	**1.63**	1	61.38	1.11E-03	**1.61**	1	61.38	0.03	**1.64**
O60664	PLIN3	Perilipin-3 OS=Homo sapiens	**2.02**	1	59.64	0.01	**2.25**	1	59.64	0.01	**1.79**
O00233	PSMD9	26S proteasome non-ATPase regulatory subunit 9	**1.66**	1	59.5	0.02	**1.62**	1	59.5	0.03	**1.69**
P08238	HSP90AB1	Heat shock protein HSP 90-beta	**1.70**	1	54.2	0.02	**1.95**	1	54.2	0.02	**1.46**
P49736	MCM2	DNA replication licensing factor MCM2	**2.12**	1	45.84	0.01	**1.90**	1	45.84	0.01	**2.33**
P30419	NMT1	Glycylpeptide N-tetradecanoyltransferase 1	**1.43**	1	41.96	0.02	**1.59**	1	41.96	0.05	**1.27**
Q9UHD8	SEPT9	Septin-9	**1.84**	1	41.65	1.49E-03	**1.73**	1	41.65	0.02	**1.94**
P33991	MCM4	DNA replication licensing factor MCM4	**2.89**	1	37.19	3.93E-03	**3.14**	1	37.19	0.01	**2.64**
O43633	CHMP2A	Charged multivesicular body protein 2a	**3.69**	1	32.85	4.43E-03	**4.28**	1	32.85	8.54E-04	**3.11**
P04818	TYMS	Thymidylate synthase	**2.40**	1	31.82	0.01	**2.28**	1	31.82	0.01	**2.52**

[a] Uniprot accession number from MASCOT search of UniProtKB-SwissProt, taxonomy Homo sapiens.

[b] Official recommended gene name taken from UniProtKB-SwissProt.

[c] Average fold change calculated form shRNA3-3 and shRNA4-7 values.

[d] Number of peptides used to identify the protein.

[e] The Mascot Score is a statistical score for how well the experimental data match the database sequence. It corresponds to the summed score, calculated with an algorithm, for the individual peptides.

Table 7.5 List of proteins with decreased expression following downregulation of MTHFD1L in HEK293 cells.

				shRNA 3-3				shRNA 4-7			
Accession No.[a]	**Gene Name[b]**	**Protein Description**	**Average Fold Change[c]**	**Pepti-des[d]**	**Score[e]**	**Anova *p*- value**	**Fold change**	**Pepti-des[d]**	**Score[e]**	**Anova *p*- value**	**Fold change**
P26639	TARS	Threonine--tRNA ligase, cytoplasmic	**1.58**	4	166.70	4.81E-05	**1.70**	4	166.70	2.09E-04	**1.46**
Q9P2J5	LARS	Leucine--tRNA ligase, cytoplasmic	**1.60**	3	134.07	0.01	**1.67**	3	134.07	0.02	**1.53**
Q14677	CLINT1	Clathrin interactor 1	**1.38**	1	75.53	0.04	**1.26**	1	75.53	0.02	**1.49**
P04792	HSPB1	Heat shock protein beta-1	**1.99**	1	69.13	0.02	**2.44**	1	69.13	0.03	**1.53**
Q08257	CRYZ	Quinone oxidoreductase	**1.51**	1	68.71	0.02	**1.54**	1	68.71	0.04	**1.48**
P78527	PRKDC	DNA-dependent protein kinase catalytic subunit	**1.98**	2	68.04	3.07E-03	**1.81**	2	68.04	0.00	**2.15**
P30825	SLC7A1	High affinity cationic amino acid transporter 1	**3.16**	1	66.99	0.03	**3.22**	1	66.99	0.02	**3.09**
P55809	OXCT1	Succinyl-CoA:3-ketoacid coenzyme A transferase 1, mitochondrial	**2.41**	1	62.33	0.05	**2.71**	1	62.33	0.03	**2.11**
Q6UB35	MTHFD1L	Monofunctional C1-tetrahydrofolate synthase, mitochondrial	**21.06**	1	58.66	5.65E-05	**31.62**	1	58.66	4.83E-04	**10.51**
Q13428	TCOF1	Treacle protein	**1.77**	1	57.69	4.53E-04	**1.73**	1	57.69	1.17E-04	**1.80**
O43592	XPOT	Exportin-T	**2.61**	1	56.94	0.01	**2.47**	1	56.94	4.09E-03	**2.75**
P50402	EMD	Emerin	**1.39**	1	53.16	3.88E-03	**1.38**	1	53.16	0.02	**1.39**
P28370	SMARCA1	Probable global transcription activator SNF2L1	**Infinity**	1	52.50	8.14E-04	**Infinity**	1	52.50	0.01	**Infinity**
Q8TEQ6	GEMIN5	Gem-associated protein 5	**1.82**	1	46.87	0.05	**1.61**	1	46.87	0.01	**2.04**
P55060	CSE1L	Exportin-2	**1.75**	1	41.01	0.04	**1.53**	1	41.01	0.04	**1.96**
Q92598	HSPH1	Heat shock protein 105 kDa	**1.88**	1	40.55	0.01	**1.77**	1	40.55	0.01	**1.99**

Q5TDH0	DDI2	Protein DDI1 homolog 2	**5.98**	1	38.34	0.03	**4.20**	1	38.34	0.01	**7.76**
Q15149	PLEC	Plectin	**2.20**	1	36.53	0.02	**2.79**	1	36.53	0.02	**1.62**

[a] Uniprot accession number from MASCOT search of UniProtKB-SwissProt, taxonomy Homo sapiens.

[b] Official recommended gene name taken from UniProtKB-SwissProt.

[c] Average fold change calculated form shRNA3-3 and shRNA4-7 values.

[d] Number of peptides used to identify the protein.

[e] The Mascot Score is a statistical score for how well the experimental data match the database sequence. It corresponds to the summed score, calculated with an algorithm, for the individual peptides.

Table 7.6 List of proteins presents in both MTHFD1L overexpression and downregulation groups.

Accession No.	Gene Name	Protein Description	Function	MTHFD1L Overexpression (Fold change)	MTHFD1L Downregulation (Fold change)
P49736	MCM2	DNA replication licensing factor MCM2	unwinding DNA	**-1.92**	**2.12**
P33992	MCM5	DNA replication licensing factor MCM5	unwinding DNA	**-1.89**	**2.29**
P33993	MCM7	DNA replication licensing factor MCM7	unwinding DNA	**-1.52**	**2.20**
P31939	ATIC	Bifunctional purine biosynthesis protein PURH	metabolism of nucleotide acid component or derivate	**-1.26**	**1.81**
P07196	NEFL	Neurofilament light polypeptide	maintenance of neuronal caliber	**-27.88**	**9.29**
P51649	ALDH5A1	Succinate-semialdehyde dehydrogenase, mitochondrial	Catalyzes one step in the degradation of the inhibitory neurotransmitter gamma-aminobutyric acid (GABA)	**-1.36**	**1.94**

Table 7.7 List of proteins involved in nucleic acid metabolism and DNA replication, recombination and repair.

Gene Name	Protein Description	Functions annotation	MTHFD1L Overexpression (Fold change)	MTHFD1L Downregulation (Fold change)
ATP5A1	ATP synthase subunit alpha, mitochondrial	metabolism of purine nucletide	-1.27	-
ATP5B	ATP synthase subunit beta, mitochondrial	metabolism of purine nucletide	-1.49	-
ATP6V1G1	V-type proton ATPase subunit G 1	metabolism of purine nucletide	-1.68	-
HSPA8	Heat shock cognate 71 kDa protein	metabolism of purine nucletide	-1.49	-
MSH2	DNA mismatch repair protein Msh2	metabolism of purine nucletide	-1.49	-
MYH9	Myosin-9	metabolism of purine nucletide	-1.41	-
PSMC3	26S protease regulatory subunit 6A	metabolism of purine nucletide	-1.41	-
VCP	Transitional endoplasmic reticulum ATPase	metabolism of purine nucletide	-1.84	-
DDX17	Probable ATP-dependent RNA helicase DDX17	double-stranded DNA break repair	-1.39	-
XRCC5	X-ray repair cross-complementing protein 5	double-stranded DNA break repair	-2.19	-
XRCC6	X-ray repair cross-complementing protein 6	double-stranded DNA break repair	-1.29	-
DDX1	ATP-dependent RNA helicase DDX1	double-stranded DNA break repair	-1.39	-
PCNA	Proliferating cell nuclear antigen	processivity factor for DNA replication	-1.56	-
TRIM28	Transcription intermediary factor 1-beta	repair of DNA	-1.67	-
VIM	Vimentin	morphology of nuclear matrix	-1.54	-
GSTP1	Glutathione S-transferase P	joining of DNA	-1.21	-
HNRNPC	Heterogeneous nuclear ribonucleoproteins C1/C2	annealing of RNA	-1.64	-
HSD17B10	3-hydroxyacyl-CoA dehydrogenase type-2	metabolism of DNA	-1.43	-
LMNB1	Lamin-B1	interaction with chromatine	-1.54	-
DCXR	L-xylulose reductase	metabolism of purine nucletides	-1.51	-
MYH9	Myosin-9	catabolism of ATP	-1.41	-
LDHA	L-lactate dehydrogenase A chain	concentration of ATP	1.47	-
FASN	Fatty acid synthase	metabolism of nucleotide	-1.28	-
TALDO1	Transaldolase	metabolism of nucleotide	-1.53	-
PGLS	6-phosphogluconolactonase	pentose shunt of D-glucose	-1.47	-
VDAC3	Voltage-dependent anion-selective channel	transport of adenin	-1.85	-

	protein 3			
ATIC	Bifunctional purine biosynthesis protein PURH	metabolism of nucleotide acid component or derivate	**-1.26**	**1.808694803**
MCM2	DNA replication licensing factor MCM2	unwinding DNA	**-1.92**	**2.118732976**
MCM5	DNA replication licensing factor MCM5	unwinding DNA	**-1.89**	**2.289314331**
MCM7	DNA replication licensing factor MCM7	unwinding DNA	**-1.52**	**2.200014921**
MCM3	DNA replication licensing factor MCM3	unwinding DNA	-	**2.694198068**
MCM4	DNA replication licensing factor MCM4	unwinding DNA	-	**2.892855431**
MCM6	DNA replication licensing factor MCM6	unwinding DNA	-	**2.308368053**
ADSS	Adenylosuccinate synthetase isozyme 2	purine Nucleotides De Novo Biosynthesis II	-	**1.506368023**
LMNA	Prelamin-A/C	fragmentation of heterochromatin DNA	-	**1.415791614**
RRM1	Ribonucleoside-diphosphate reductase large subunit	depletion of mitocondrial DNA	-	**1.450368774**
CHMP2A	Charged multivesicular body protein 2a	segregration of chromosomes	-	**3.694267782**
TYMS	Thymidylate synthase	thymidylate synthesis	-	**2.400240441**
CRYZ	Quinone oxidoreductase	DNA damage	-	**-1.507881316**
HSPB1	Heat shock protein beta-1	DNA damage	-	**-1.986877817**
PRKDC	DNA-dependent protein kinase catalytic subunit	DNA damage	-	**-1.97900253**
SMARCA1	Probable global transcription activator SNF2L1	DNA damage	-	**-Infinity**

CHAPTER 8

Conclusion

and

Future Work

8.1 Conclusion

This project is focused on the characterization of the folate-related gene *MTHFD1L* starting from its previous association with NTD risk (Parle-McDermott *et al*. 2009). Figure 8.1 depicts a schematic representation of the project that started from the NTD associated polymorphisms clustered in two regions of *MTHFD1L*. Region "a" inspired the cleft risk association study and the cycloleucine experiment. SNP rs7646 within region "b" led to the miRNA and ncRNA investigation. The final protein expression experiment was performed to corroborate our results and provide a better insight into MTHFD1L's role in disease.

In summary, *MTHFD1L* polymorphisms have been studied in relation to cleft disease. TDT analysis showed an association between DIP rs3832406 and CLP cases. However, adjustment of our analyses resulted in loss of statistical significance indicating that this variant is simply not associated with CLP or that a larger sample size is required to detect an effect. We suggest further screening of rs3832406 DIP in a larger cohort and describe a new assay that will facilitate this. We also demonstrated that the modified Melting Curve Analysis developed for DIP rs3832406 is a valid alternative to capillary electrophoresis for the genotyping of multiple allele deletion/insertion polymorphisms.

We identified a novel direct interaction between the *MTHFD1L* 3′ UTR and miR-197 in MCF-7 and HEK293 cell lines. This is the second miRNA to be identified as a regulator of MTHFD1L, along with miR-9 (Selcuklu *et al.* 2012). We verified that miR-197 downregulates MTHFD1L and that SNPrs7646 within *MTHFD1L* 3′UTR significantly changes miR-197 binding affinity. This provides a functional explanation as to why we previously found SNP rs7646 to be associated with NTDs in an Irish population (Parle-McDermott *et al*. 2009) (Figure 8.1). Our results also suggested that pre-9 and pre-197 directed downregulation of MTHFD1L could be affected by the competitive binding behaviour of both miRNAs. Moreover, our data also indicates that miR-9 may target other folate genes including DHFR, DHFRL1, SHMT1 and MTR. This putative role of miR-9 along with its involvement in neurogenesis further highlights that this miRNA warrants

further investigation in relation to NTD and other neuro-developmental diseases. miRNA regulation provides an elegant mechanism to co-ordinately regulate enzymes from the same pathway and the association of rs7636 with NTDs emphasises the importance of this regulation during development.

Sequence analysis was performed to discover and characterize *MTHFD1* and *MTHFD1L* similar sequences. A ncRNA transcribed from CHRX was shown to downregulate *MTHFD1* mRNA and the formation of DNA-RNA triplex could have a role in this effect. This part of the project brought interesting cues for future work.

We also demonstrated that *MTHFD1L* expression is affected by a 24-hour exposure to cycloleucine in lymphoblast cells. Results suggest that this regulation operates through the alternative splicing of the gene (Figure 8.1). These findings further support our hypothesis of the pivotal role played by alternative splicing in controlling the level of MTHFD1L.

Finally, a proteomic analysis demonstrated that MTHFD1L levels affect the nuclear thymidylate biosynthesis complex and other proteins involved in DNA replication and repair. We proposed that an increased formate production may be a signal for uncontrolled cell proliferation that triggers defensive mechanisms like the inhibition of nuclear thymidylate biosynthesis and DNA replication and repair. On the other hand, a downregulation of MTHFD1L has been linked with an increased DNA damage. These data suggest that nucleotide biosynthesis and DNA replication and repair are sensitive to changes in formate production from the mitochondria. Formate supply may act as a 'sensor' to ensure that there is a balanced supply of nucleotides for DNA replication and repair and to prevent uracil misincorporation.

Formate is a pivotal component of 1C metabolism and, despite its importance, our knowledge of its direct involvement with disease is still limited. A 7-fold increase in formate was detected in the serum and urine of vitamin B12 deficient rats and a similar increase predicted for folate deficiency (Lamarre *et al.* 2012). Furthermore, increased formate levels have been identified in advance stage oesophageal cancer compared to

normal mucosae (Wang *et al*. 2013). Another key factor to understand the role of formate is the mechanism that transports this metabolite through the cellular compartments. Little is known about formate transport across the mitochondrial inner membrane. Formate is known to enter intact mitochondria, as suggested by swelling assays. It also effluxes rapidly from mitochondria, both *in vivo* (in yeast) and *in vitro* (Tibbets and Appling 2010). High concentrations of formate in mitochondria is likely to be toxic to mitochondrial respiration *via* inhibition of cytochrome c oxidase (Nicholls 1975, Kapur *et al.* 2007, this study) suggesting that formate transport is carrier mediated. Therefore, to better understand the role of formate in disease it is fundamental to increase our comprehension on how formate is transported and compartmentalised.

Altogether, our findings suggest that maintaining MTHFD1L and, consequently, formate levels within a determined range might be crucial for the health of the cell. This prompts further investigation into how this signal, *via* subcellular formate concentration, is translated into a wide cellular response. The study of broad cellular regulators like c-myc oncogene and E2F1 transcription factor may provide an answer to this question. Indeed, bioinformatic analysis on the proteomic data showed a significant match between c-myc and E2F1 targeted proteins and the protein dataset resulting from MTHFD1L inhibition (IPA software) (Figure 8.2). The c-myc oncogene, which is a crucial transcription factor in the control of cell proliferation, differentiation, and apoptosis, was proposed as an MTHFD1L regulator by Sugiura *et al*. (2004). Both mitochondrial proteins SHMT2 (Nikiforov *et al*. 2002) and MTHFD1L (Tavtigian *et al*. 1994, Sugiura *et al*. 2004) seem to be responsive to the expression of c-myc in mammalian cells. A c-myc consensus sequence (CACGTG) is also present in the first intron of both *SHMT2* and *MTHFD1L* genes. Evidence of c-myc regulation of the cytosolic MTHFD1 has also been reported (Mao *et al.* 2003, Carroll *et al.* 2009). Thus, studying the impact of formate levels in different cellular compartments on c-myc and E2F1 regulation may explain how MTHFD1L expression changes result in a wide cellular response. This could add another piece to the puzzle of understanding folate-mediated 1C metabolism and its relationship to disease.

In conclusion, this study investigated MTHFD1L common genetic variants in relation to cleft disease and how polymorphisms can affect regulation of MTHFD1L levels and the subsequent risk of disease. The impact of alternative splicing on MTHFD1L expression and the potential function of additional homologous sequences has also been investigated. These data have been corroborated by proteomic analysis on samples with different MTHFD1L expression levels. On the tenth anniversary of the first characterisation of MTHFD1L (Prasannan *et al*. 2003), these findings provide relevant insights into the role played by MTHFD1L inside the cell and pave the way for future investigations on the signalling effect of formate.

8.2 Future work

8.2.1 Transcription factors responsive to formate levels

The future work will address the open issues mentioned above. Bioinformatic analysis will be performed to select possible candidates for the role of a formate transporter, taking advantage of our proteomic data and other folate related analysis. In a similar way, database analysis will focus on potential transcription factors and regulators, such as c-myc and E2F1. Depending on the results of *in silico* analysis, we will decide which experiments to carry out in the wet lab.

8.2.2 Putative thymidylate biosynthesis complex in mitochondria

In the nucleus, SHMT1, TYMS, DHFR, MTHFD1 and other proteins form a thymidylate biosynthesis complex associated with the nuclear lamina (Anderson *et al.* 2012, Chapter 7, Section 7.4). Similarly, we will investigate a possible formation of a multi enzymatic complex in the mitochondria by using co-immunoprecipitation experiments and tandem affinity purification. For these experiments we will employ monoclonal antibodies for MTHFD1L and DHFRL1 and we will carry out LC-MS analysis to identify the proteins belonging to the complex. DHFRL1 was a formerly annotated pseudogene and, only recently, was proven to be a functional protein by our laboratory, in a publication that I

made a major contribution to (McEntee, Minguzzi *et al.* 2011). Taking advantage of the work already accomplished, we will generate single cell clones with DHFRL1 knocked-down and overexpressed. Proteomic analysis will be carried out in the same way as for MTHFD1L. Mitochondrial and nuclear protein extracted from DHFRL1 and MTHFD1L may be also specifically screened using LC-MS analysis. LC-MC instrument may be also used for metabolic analysis using the above-mentioned samples.

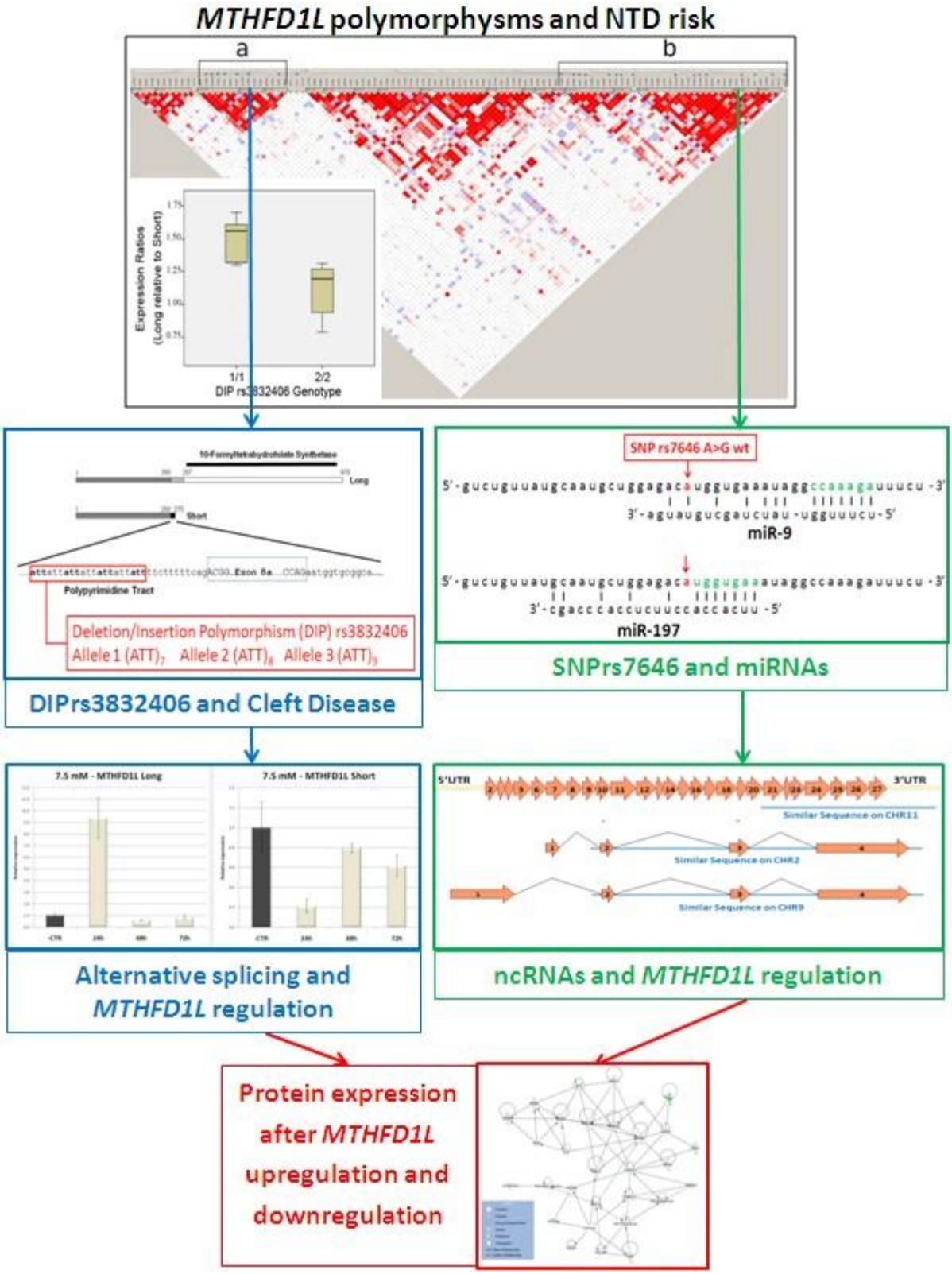

MTHFD1L polymorphysms and NTD risk
a
b
Expression Ratios (Long relative to Short)
1/1
2/2
DIP rs3832406 Genotype
Long
Short
Polypyrimidine Tract
Deletion/Insertion Polymorphism (DIP) rs3832406
Allele 1 (ATT)7 Allele 2 (ATT)8 Allele 3 (ATT)9
DIPrs3832406 and Cleft Disease
SNP rs7646 A>G wt
miR-9
miR-197
SNPrs7646 and miRNAs
7.5 mM - MTHFD1L Long
7.5 mM - MTHFD1L Short
Alternative splicing and MTHFD1L regulation
5'UTR
3'UTR
Similar Sequence on CHR11
Similar Sequence on CHR2
Similar Sequence on CHR9
ncRNAs and MTHFD1L regulation
Protein expression after MTHFD1L upregulation and downregulation

Figure 8.1 Schematic representation of the whole project (figure drawn with Microsoft PowerPoint).

The scheme represents the project starting from the previous NTD risk association study (Parle-McDermott *et al*. 2009). DIP rs3832406 within region "a" inspired the cleft risk association study and the cycloleucine experiment. SNP rs7646 within region "b" led to the miRNA and ncRNA investigation. The final protein expression experiment was performed to corroborate our results and provide a better insight of MTHFD1L in relation to disease.

Upstream Regulator	Molecule Type	Predicted Activation State	Activation z-score	Notes	p-value of overlap	Target molecules in dataset
MYC	transcription regulator		-1.811		1.43E-13	↓ATAD3A/ATAD3B, ↓DDX39B ...all 24
E2F1	transcription regulator	Inhibited	-2.041	bias	2.91E-12	↓CA2, ↑CCT4, ↓DDX39B, ↑... ...all 16
TP53	transcription regulator		-1.200		4.24E-10	↓COX7A2, ↓CYB5R3, ↓EZR ...all 24
tanespimycin	chemical drug		0.213	bias	9.69E-10	↓HNRNPC, ↓HSP90B1, ↓HS... ...all 9
CD 437	chemical drug	Activated	2.714	bias	1.94E-09	↓ATP5A1, ↓ATP5B, ↓COX... ...all 11
INSR	kinase		-1.010	bias	2.39E-09	↑ANP32A, ↓ATP5A1, ↓ATP5B...all 13
5-fluorouracil	chemical drug		0.847	bias	3.37E-09	↓ATP5B, ↑CCT4, ↓GOT2, ↑... ...all 11
MAPT	kinase				1.43E-08	↑ANP32A, ↓ATP5A1, ↓ATP5B...all 11
CD3	complex	Activated	2.145	bias	1.34E-07	↓HNRNPR, ↓HSP90B1, ↓H... ...all 15
1,2-dithiol-3-thione	chemical reagent		-1.640	bias	4.06E-07	↑CCT7, ↓DDX39B, ↓GSTP1, ...all 9
NFE2L2	transcription regulator		-1.711	bias	7.15E-06	↑CCT7, ↓DDX39B, ↓GSTP1 ...all 10
L-serine	chemical - endogenous mar				7.25E-06	↓PCNA, ↑PHGDH, ↑PSAT1 ...all 3
TNFSF8	cytokine				1.41E-05	↓MSH2, ↓XRCC5, ↓XRCC6 ...all 3
BMS-690514	chemical drug	Activated	2.000	bias	1.42E-05	↓HSPA8, ↓MCM5, ↓MCM7, ↑... ...all 4
HTT	transcription regulator				1.42E-05	↑ANP32A, ↓ATP5A1, ↓ATP5B...all 13
SNCA	other				1.69E-05	↓EZR, ↓GLO1, ↓HSP90B1, ↑... ...all 5
APP	other				1.97E-05	↑ANP32A, ↓ATP5A1, ↓ATP5B...all 13
PSEN1	peptidase				4.80E-05	↑ANP32A, ↓ATP5A1, ↓ATP5B ...all 9
p70 S6k	group				5.71E-05	↓FASN, ↓PABPC4, ↓VIM ...all 3
CUL4B	other				5.71E-05	↓DARS, ↓GLO1, ↓PRDX3 ...all 3
FOS	transcription regulator		-0.955		6.33E-05	↓CA2, ↓EZR, ↓GSTP1, ↓... ...all 10
sirolimus	chemical drug		1.767	bias	7.16E-05	↓FASN, ↓GSTP1, ↓HSP90B1, ...all 9
Ins1	other		-0.794	bias	9.12E-05	↓COX7A2, ↓FASN, ↓GSTP1, ...all 8
mono-(2-ethylhexyl)phthalate	chemical toxicant		-1.664	bias	9.43E-05	↓ATP5B, ↓COX7A2, ↓FASN, ...all 6
PLN	transporter				1.26E-04	↓EZR, ↓HSP90B1, ↓VCP, ↓... ...all 4
methylprednisolone	chemical drug		-0.592		1.34E-04	↑ANP32A, ↓ATIC, ↓CYB5R3 ...all 10

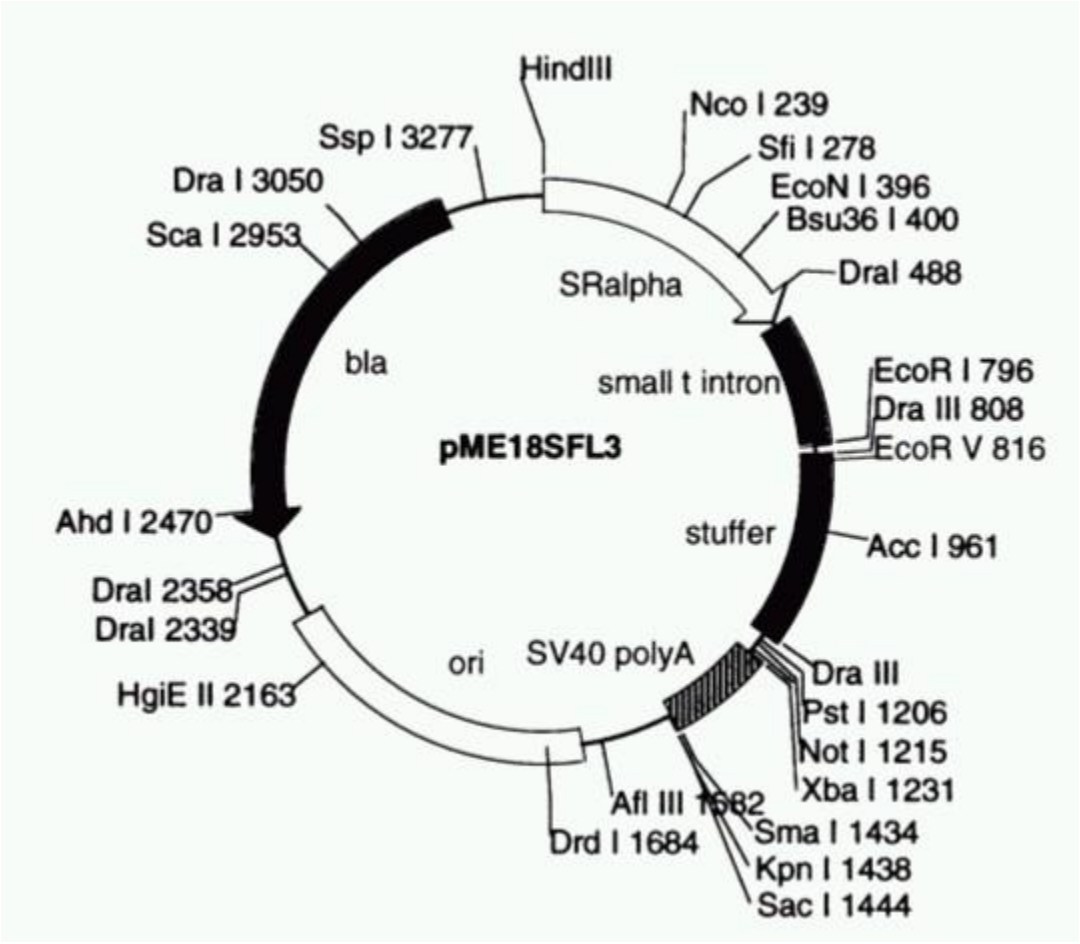
Figure 8.2 Bioinformatic analysis on the MTHFD1L overexpression dataset using IPA software.

The table shows a list of cellular regulators predicted to regulate the proteins affected by MTHFD1L overexpression. The regulators are listed in order of significance.

Appendices

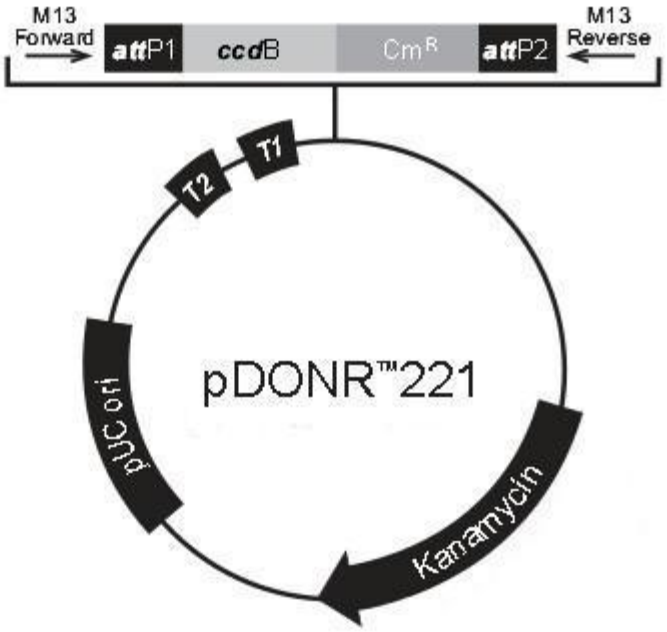
Appendix A

Vector map for pME1SFL3. The image was taken from Invitrogen and it is available at http://www.nbrc.nite.go.jp/vector_FIG.html



Appendix B

Vector map for pDONR™221. The image was taken from Invitrogen and it is available at http://tools.invitrogen.com/content/sfs/vectors/pdonr221_pdonrzeo_map.pdf

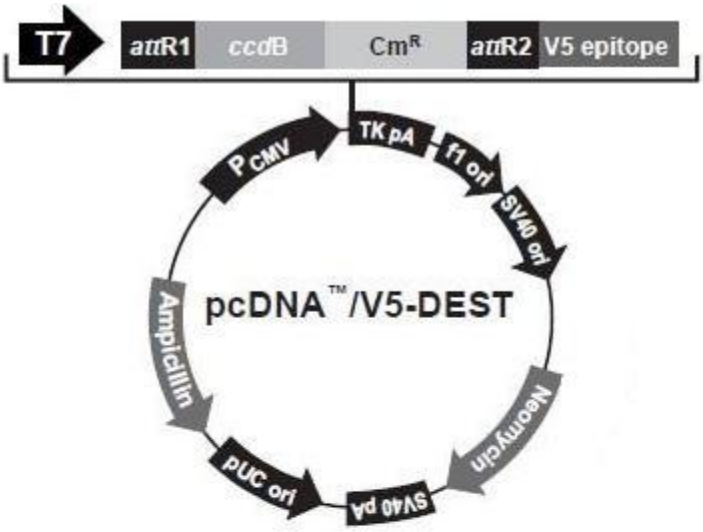


Comments for:	pDONR™221 4761 nucleotides
*rrn*B T2 transcription termination sequence (c):	268-295
rrnB T1 transcription termination sequence (c):	427-470
M13 Forward (-20) priming site:	537-552
*att*P1:	570-801
*ccd*B gene (c):	1197-1502
Chloramphenicol resistance gene (c):	1825-2505
*att*P2 (c):	2753-2984
M13 Reverse priming site:	3026-3042
Kanamycin resistance gene:	3155-3964
EM7 promoter (c):	—
Zeocin resistance gene (c):	—
pUC origin:	4085-4758

(c) = complementary strand

Appendix C

Vector map for pcDNA™3.2/V5-DEST. The image was taken from Invitrogen and it is available at http://tools.invitrogen.com/content/sfs/vectors/pcdnav5dest_map.pdf

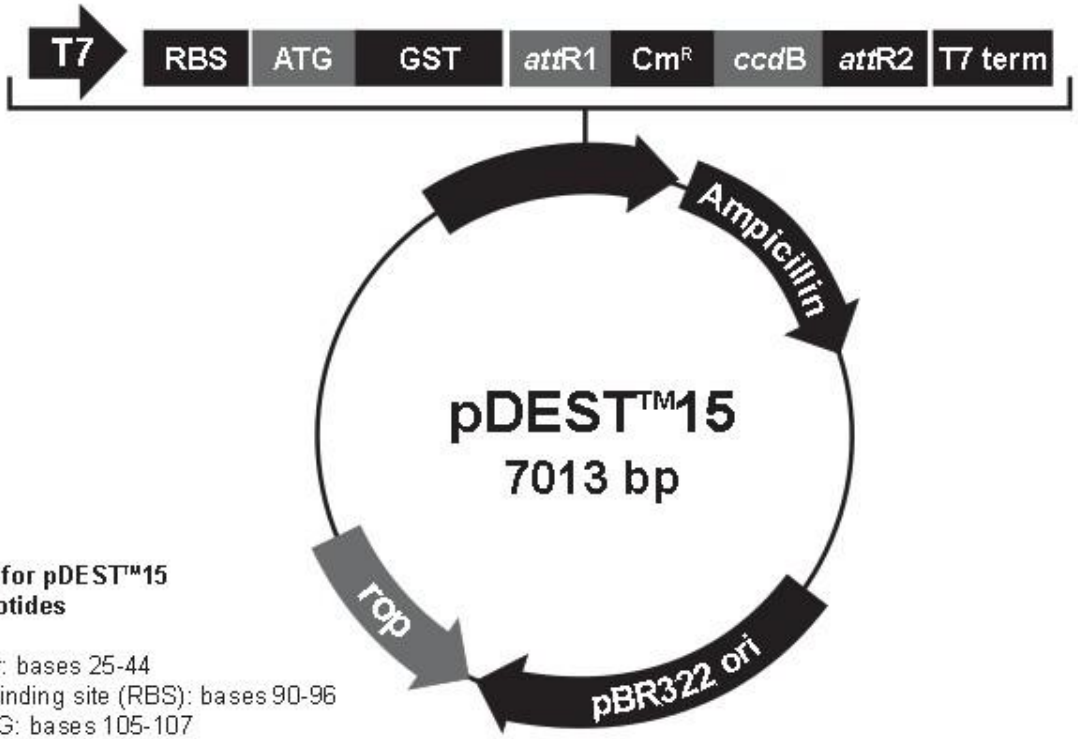


Comments for:	pcDNA™3.2/V5-DEST 7711 nucleotides
CMV promoter:	232-819
T7 promoter/priming site:	863-882
*att*R1 site:	911-1035
*ccd*B gene (c):	1464-1769
Chloramphenicol resistance gene (c):	2111-2770
*att*R2 site:	3051-3175
V5 epitope:	3201-3242
V5 reverse priming site:	3210-3230
TK polyadenylation signal:	3269-3540
f1 origin:	3576-4004
SV40 early promoter and origin:	4031-4339
Neomycin resistance gene:	4414-5208
SV40 early polyadenylation signal:	5384-5514
pUC origin (c):	5897-6570
Ampicillin (*bla*) resistance gene (c):	6715-7575
bla promoter (c):	7576-7674

(c) = complementary strand

Appendix D

Vector map for Gateway® pDEST™15 Vector. The image was taken from Invitrogen and it is available at http://tools.invitrogen.com/content/sfs/vectors/pdest15_map.pdf



Comments for pDEST™15
7013 nucleotides

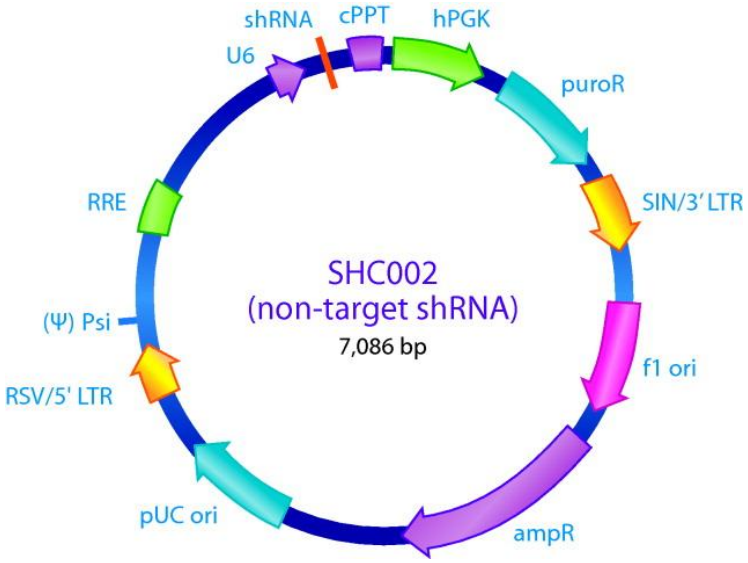
T7 promoter: bases 25-44
Ribosome binding site (RBS): bases 90-96
Initiation ATG: bases 105-107
GST tag: bases 108-776
*att*R1: bases 792-916
Chloramphenicol resistance gene (Cm^R): bases 1025-1684
*ccd*B gene: bases 2026-2331
*att*R2: bases 2372-2496
T7 transcription termination region: bases 2518-2646
bla promoter: bases 3134-3232
Ampicillin (*bla*) resistance gene: bases 3233-4093
pBR322 origin: bases 4238-4911
ROP ORF: bases 5282-5473 (C)

C=complementary strand

Appendix E

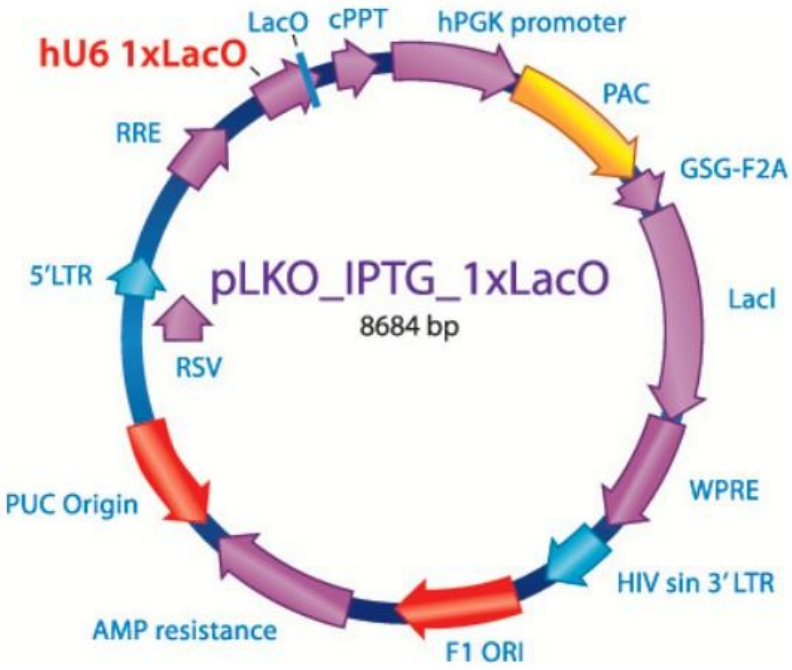
Map for non target shRNA in pLK0.1 vector. The image was taken from Sigma and it is available at

http://www.sigmaaldrich.com/catalog/product/sigma/shc002v?lang=en&region=IE



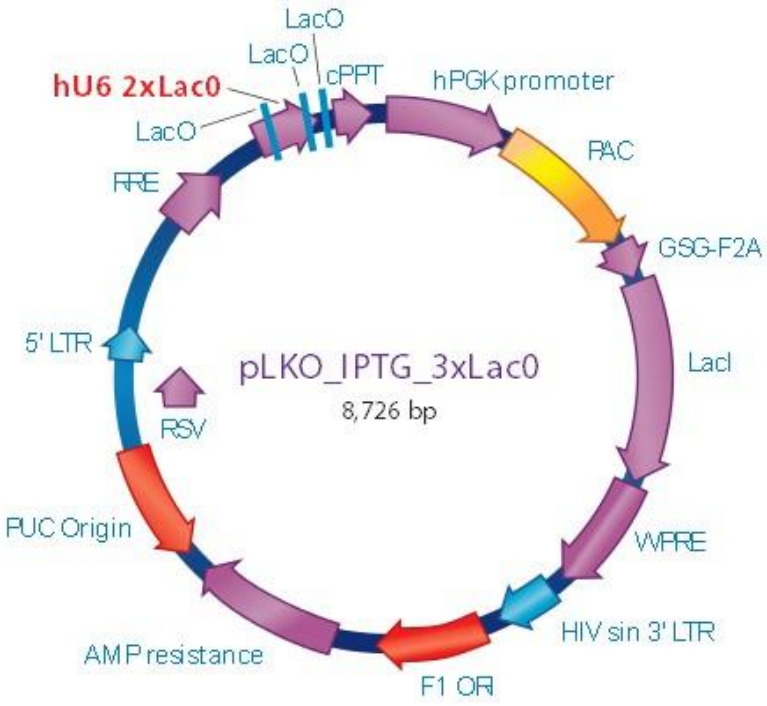
Appendix F

Vector map for non target pLK_IPTG_1xLac0 vector. The image was taken from Sigma and it is available at http://www.sigmaaldrich.com/content/dam/sigma-aldrich/docs/Sigma-Aldrich/Vector/1/plko_iptg_1xlaco.pdf



Appendix G

Vector map for non target pLK_IPTG_3xLac0 vector. The image was taken from Sigma and it is available at http://www.sigmaaldrich.com/content/dam/sigma-aldrich/docs/Sigma-Aldrich/Vector/1/plko_iptg_3xlaco.pdf



Appendix H

Nucleotide sequence of MTHFD1L optimized for protein expression in mammalian cells.

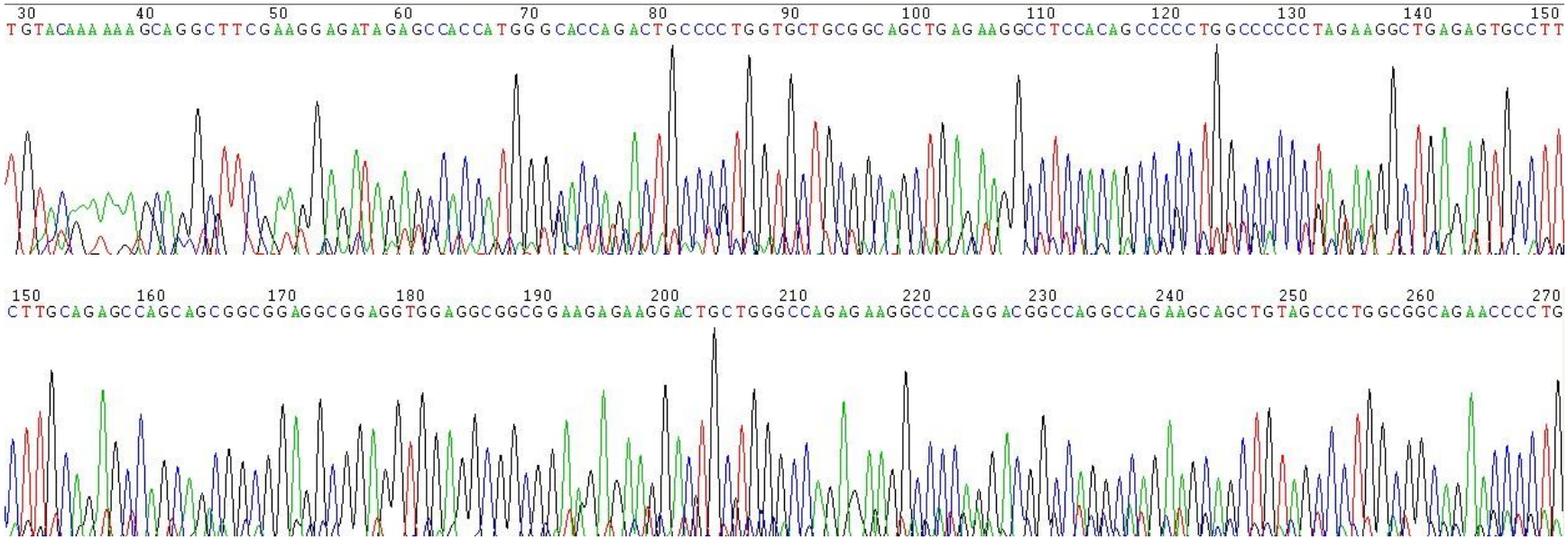
The sequence, which is cloned in pMA-RQ vector, was purchased from Invitrogen® (Chapter 2, Section 2.2.12). In bold there are the start codon and the stop codon of protein coding sequence. attB regions, necessary for Gateway recombination, are highlighted in yellow. Shine-Dalgarno and Kozak consensus sequence are highlighted in green and blue respectively. Inclusion of the Shine-Dalgarno and Kozak consensus sequence allows protein expression in both E. coli and mammalian cells.

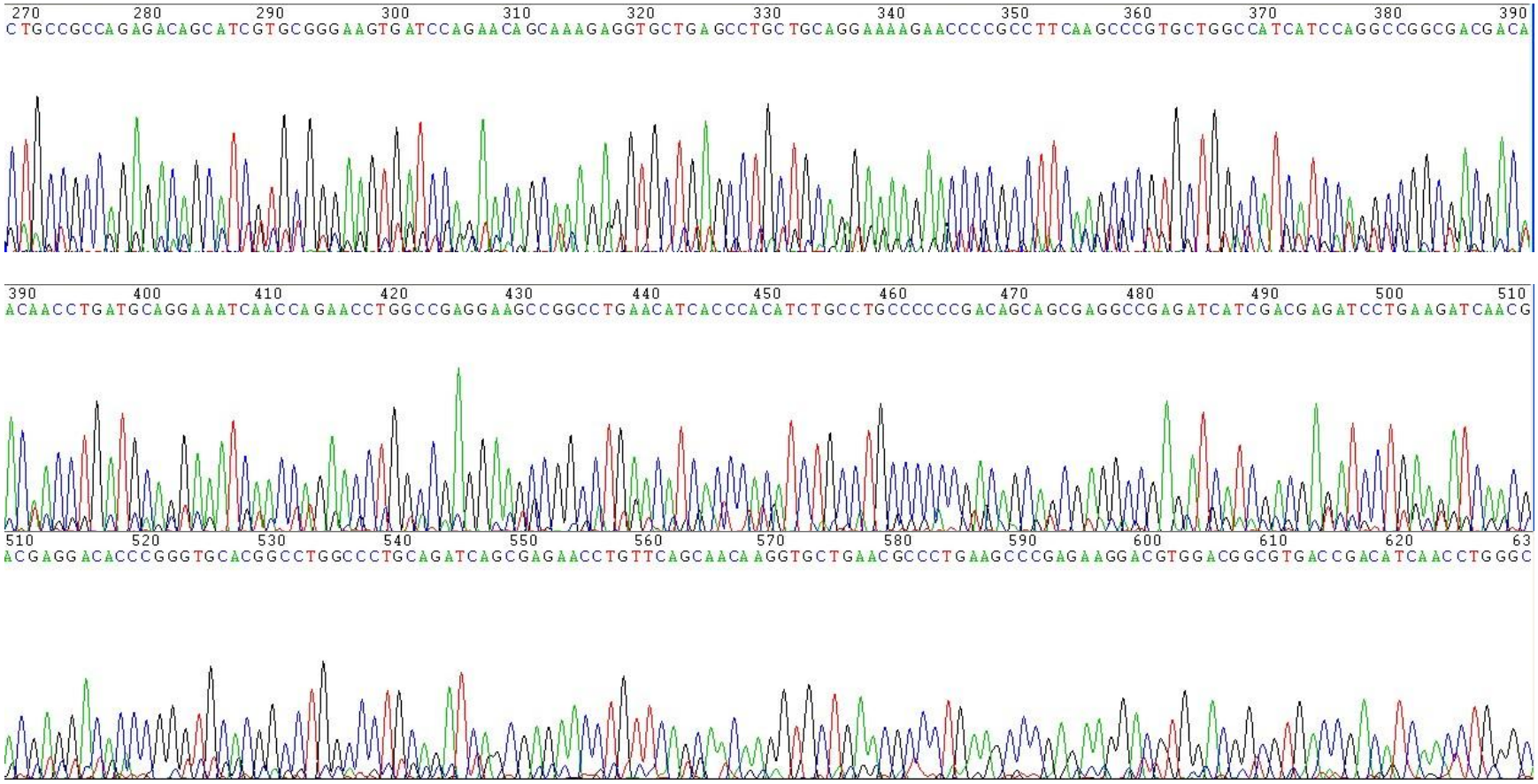
5′-

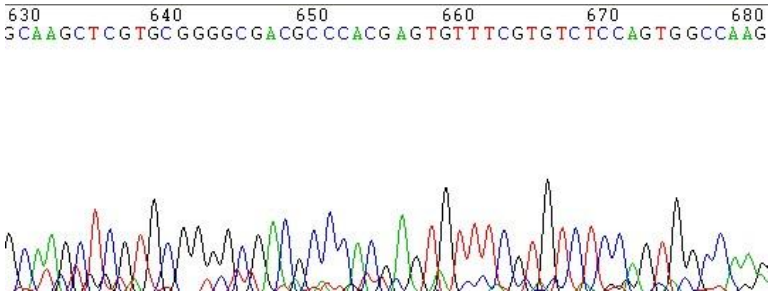
gagctctctagaggggacaagtttgtacaaaaaagcaggcttcgaaggagatagagccacc**atg**ggcaccagactgcccctggtgctgcggcagctgaga
aggcctccacagccccctggccccctagaaggctgagagtgccttgcagagccagcagcggcggaggcggaggtggaggcggcggaagagaaggac
tgctgggccagagaaggccccaggacggccaggccagaagcagctgtagccctggcggcagaaccccctgccgccagagacagcatcgtgcgggaagt
gatccagaacagcaaagaggtgctgagcctgctgcaggaaaagaaccccgccttcaagcccgtgctggccatcatccaggccggcgacgacaacctgatg
caggaaatcaaccagaacctggccgaggaagccggcctgaacatcacccacatctgcctgccccccgacagcagcgaggccgagatcatcgacgagatc
ctgaagatcaacgaggacacccgggtgcacggcctggccctgcagatcagcgagaacctgttcagcaacaaggtgctgaacgccctgaagcccgagaag
gacgtggacggcgtgaccgacatcaacctgggcaagctcgtgcggggcgacgcccacgagtgtttcgtgtctccagtggccaaggccgtgatcgagctgct
ggaaaagtccggcgtgaacctggacggcaagaaaatcctggtcgtgggagcccacggcagcctggaagccgccctgcagtgcctgttccagcggaaggg
cagcatgaccatgagcatccagtggaaaacccggcagctgcagagcaagctgcacgaggccgacatcgtggtgctgggctccccccaagcccgaggaaat
ccccctgacctggatccagcccggcaccaccgtgctgaactgcagccacgactttctgagcggcaaagtgggctgtggcagccccagaatccacttcggcg
gcctgatcgaagaggacgacgtgatcctgctggccgctgccctgagaatccagaacatggtgtccagcggcagacggtggctgagagagcagcagcaca
gaagatggcggctgcactgcctgaagctgcagcccctgagccccgtgcccagcgacatcgagatcagcagaggccagacccctaaggccgtggatgtgct
ggccaaagagatcggactgctggccgacgagatcgagatctacggcaagagcaaggccaaagtgcggctgagcgtgctggaacggctgaaggaccagg
ccgacggcaaatacgtgctggtggccggcatcacccccacccctctgggcgagggcaagagcaccgtgaccatcggactggtgcaggccctgaccgccc
acctgaacgtgaacagcttcgcctgcctgcggcagcccagccagggccctacatttggagtgaagggcggagccgctggcggcggatatgcccaggtcat
ccccatggaagagttcaacctgcacctgaccggcgacatccacgccatcaccgccgccaacaatctgctggctgccgccatcgacacccggatcctgcacg
agaacacccagaccgacaaggccctgtacaaccggctggtgcccctggtcaacggcgtgcgcgagttcagcgagatccagctggcccggctgaagaagc
tgggcatcaacaagaccgaccccagcaccctgaccgaggaagaggtgtccaagttcgcccggctggacatcgacccctccaccatcacctggcagcgggt
gctggataccaacgaccggttcctgcggaagatcaccatcggccagggcaacaccgagaagggccactaccggcaggcccagttcgatatcgccgtggcc
agcgagatcatggccgtgctggcactgaccgacagcctggccgacatgaaggcccggctgggcagaatggtggtggcctccgataagagcggccagccc
gtgaccgccgacgatctgggagtgacaggcgccctgaccgtgctgatgaaggacgccatcaagcccaatctgatgcagaccctggaaggcacccccgtgtt
cgtgcacgccggacccttcgccaatatcgcccacggcaacagcagcgtgctggccgataagattgccctgaagctcgtgggcgaagagggcttcgtggtca
ccgaggccggctttggcgccgacatcggcatggaaaagttcttcaacatcaagtgccgggccagcggcctggtgcccaacgtggtggtgctggtcgccaca
gtgcgggccctgaagatgcacggcggaggcccatctgtgacagccggcgtgcccctgaagaaagagtacaccgaagagaacatccagctggtggctgac
ggctgctgcaacctgcagaagcagatccagatcacccagctgttcggcgtgcccgtggtggtggctctgaacgtgttcaagaccgataccagagccgagatc
gacctcgtgtgcgagctggccaagcgggctggcgccttcgatgccgtgccctgctatcattggagcgtgggcggcaagggcagcgtggacctggccagag
ccgtcagagaggccgccagcaagcggagccggttccagttcctgtacgacgtgcaggtccccatcgtggacaagatccggacaatcgcccaggccgtgta
cggcgccaaggacatcgagctgagccccgaggcccaggccaagatcgacagatacacccagcagggcttcggcaacctgcccatctgcatggccaagac
ccacctgagcctgagccaccagcccgacaagaaaggcgtgccccgggacttcatcctgcccatcagtgacgtgcgggccagcatcggagccggcttcatct
acccccctcgtgggcaccatgagcaccatgcccggcctgcccaccggccctgcttctacgacatcgacctggacaccgagacagagcaggtcaaaggcct
gttc**taa**gacccagctttcttgtacaaagtggtcccctctagaggtacc -3′

Appendix I

Sanger sequencing result for pcDNA3.2- MTHFD1Lopt using T7 For primer (Table 2.4).

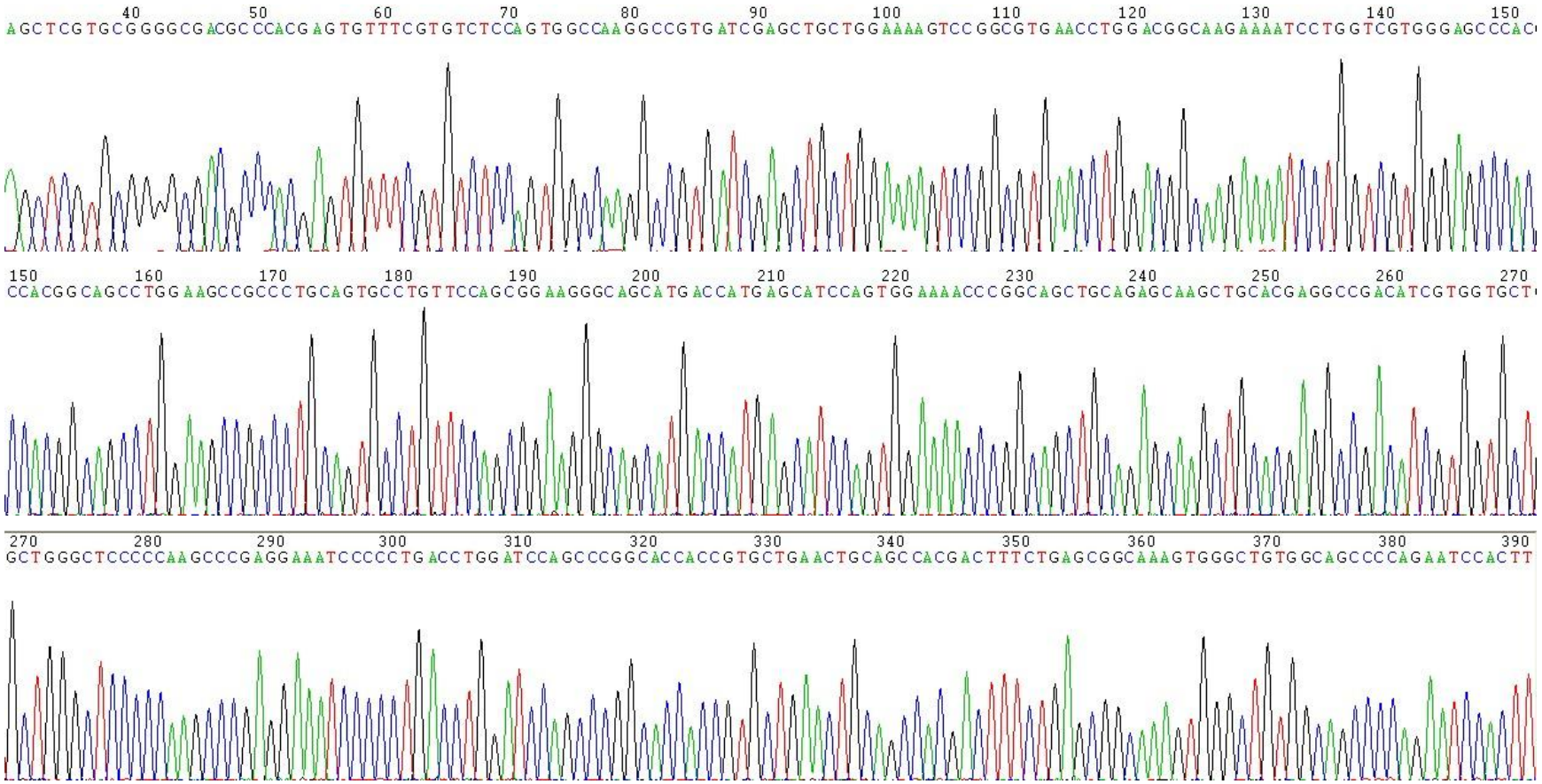


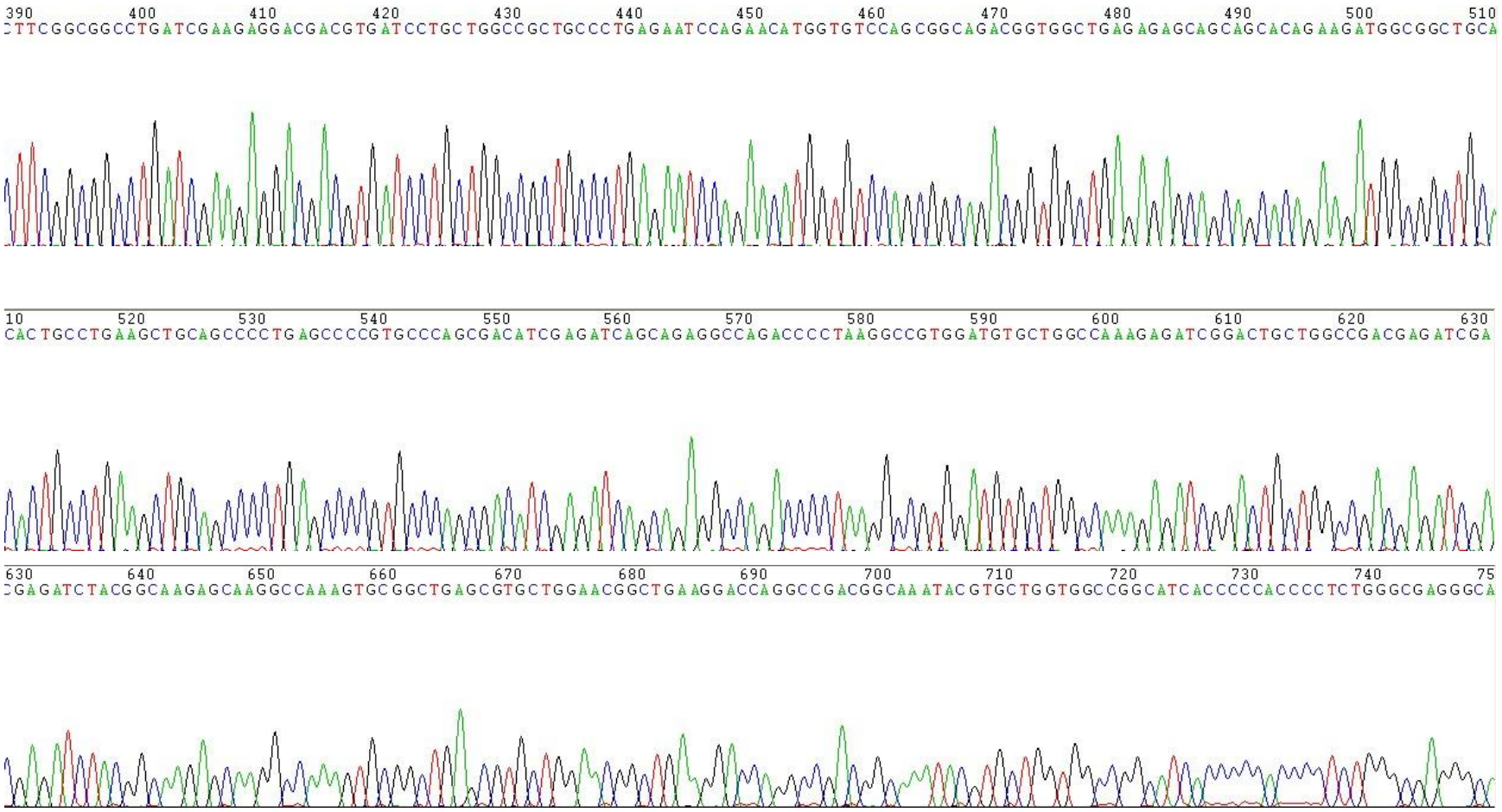

270 280 290 300 310 320 330 340 350 360 370 380 390
CTGCCGCCAGAGACAGCATCGTGCGGGAAGTGATCCAGAACAGCAAAGAGGTGCTGAGCCTGCTGCAGGAAAAGAACCCCGCCTTCAAGCCCGTGCTGGCCATCATCCAGGCCGGCGACGACA
390 400 410 420 430 440 450 460 470 480 490 500 510
ACAACCTGATGCAGGAAATCAACCAGAACCTGGCCGAGGAAGCCGGCCTGAACATCACCCACATCTGCCTGCCCCCCGACAGCAGCGAGGCCGAGATCATCGACGAGATCCTGAAGATCAACG
510 520 530 540 550 560 570 580 590 600 610 620 63
ACGAGGACACCCGGGTGCACGGCCTGGCCCTGCAGATCAGCGAGAACCTGTTCAGCAACAAGGTGCTGAACGCCCTGAAGCCCGAGAAGGACGTGGACGGCGTGACCGACATCAACCTGGGC


630
640
650
660
670
680
GCAAGCTCGTGCGGGGCGACGCCCACGAGTGTTTCGTGTCTCCAGTGGCCAAG

Appendix J

Sanger sequencing result for pcDNA3.2- MTHFD1Lopt using MTHFD1Lopt2 For primer (Table 2.4).

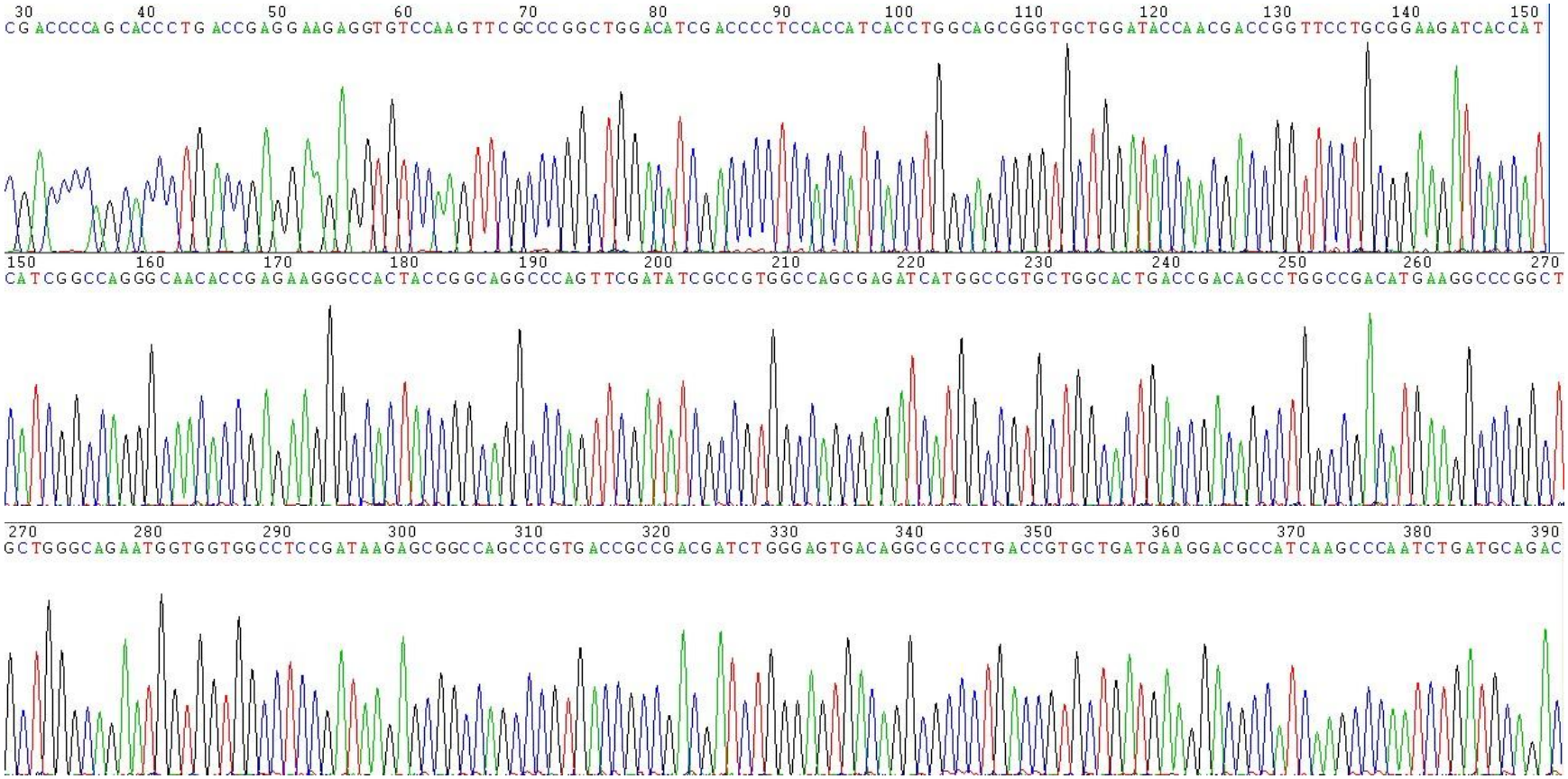


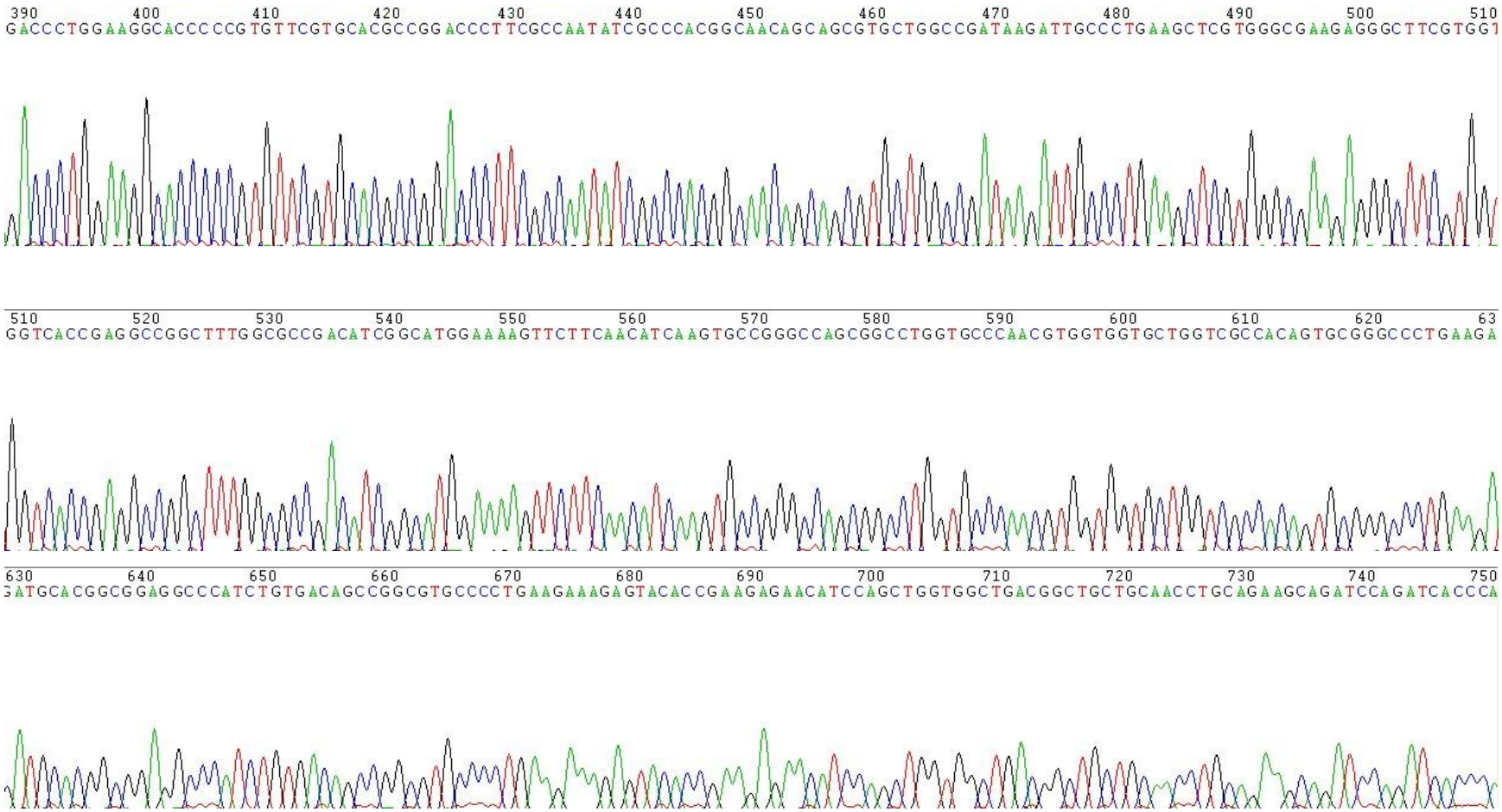

390 400 410 420 430 440 450 460 470 480 490 500 510
CTTCGGCGGCCTGATCGAAGAGGACGACGTGATCCTGCTGGCCGCTGCCCTGAGAATCCAGAACATGGTGTCCAGCGGCAGACGGTGGCTGAGAGAGCAGCAGCACAGAAGATGGCGGCTGCA
10 520 530 540 550 560 570 580 590 600 610 620 630
CACTGCCTGAAGCTGCAGCCCCTGAGCCCCGTGCCCAGCGACATCGAGATCAGCAGAGGCCAGACCCCTAAGGCCGTGGATGTGCTGGCCAAAGAGATCGGACTGCTGGCCGACGAGATCGA
630 640 650 660 670 680 690 700 710 720 730 740 75
CGAGATCTACGGCAAGAGCAAGGCCAAAGTGCGGCTGAGCGTGCTGGAACGGCTGAAGGACCAGGCCGACGGCAAAATACGTGCTGGTGGCCGGCATCACCCCCACCCCTCTGGGCGAGGGCA

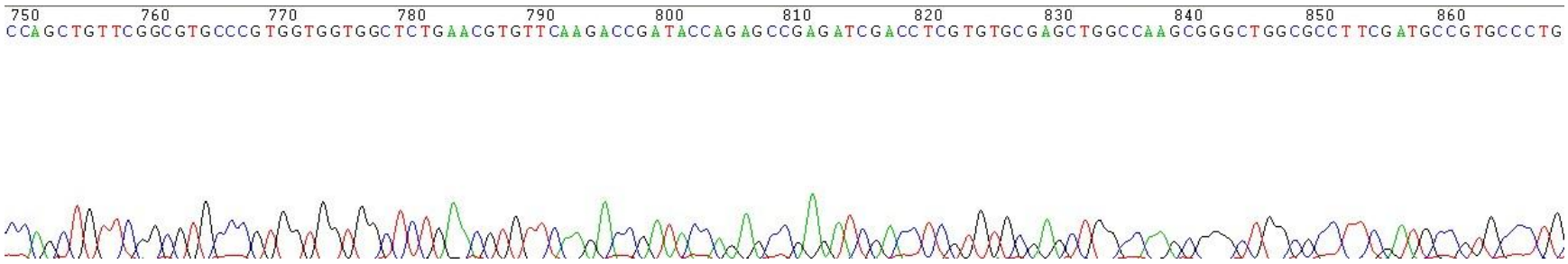

750 760 770 780 790 800 810 820 830 840 850 860
CAAGAGCACCGTGACCATCGGACTGGTGCAGGCCCTGACCGCCCACCTGAACGTGAACAGCTTCGCCTGCCTGCGGCAGCCCAGCCAGGGCCCTACATTTGGAGTGAAGGGCGGAGCCGC
870 880 890 900 910 920 930 940 950 960 970 980
CGCTGGCGGCGGATATGCCCAGGTCATCCCCATGGAAGAGTTCAACCTGCACCTGACCGGCGACATCCACGCCATCACCGCCGCCAACAATCTGCTGGCTGCCGCCATCGACACCCC

Appendix K

Sanger sequencing result for pcDNA3.2- MTHFD1Lopt using MTHFD1Lopt For primer (Table 2.4).

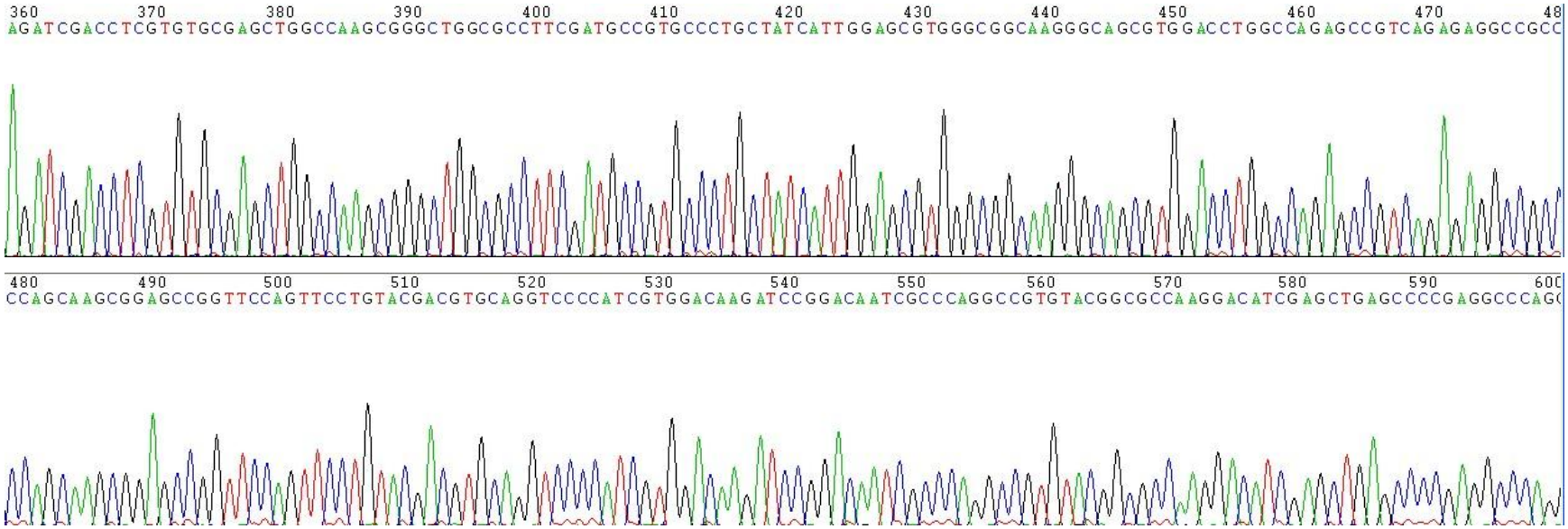


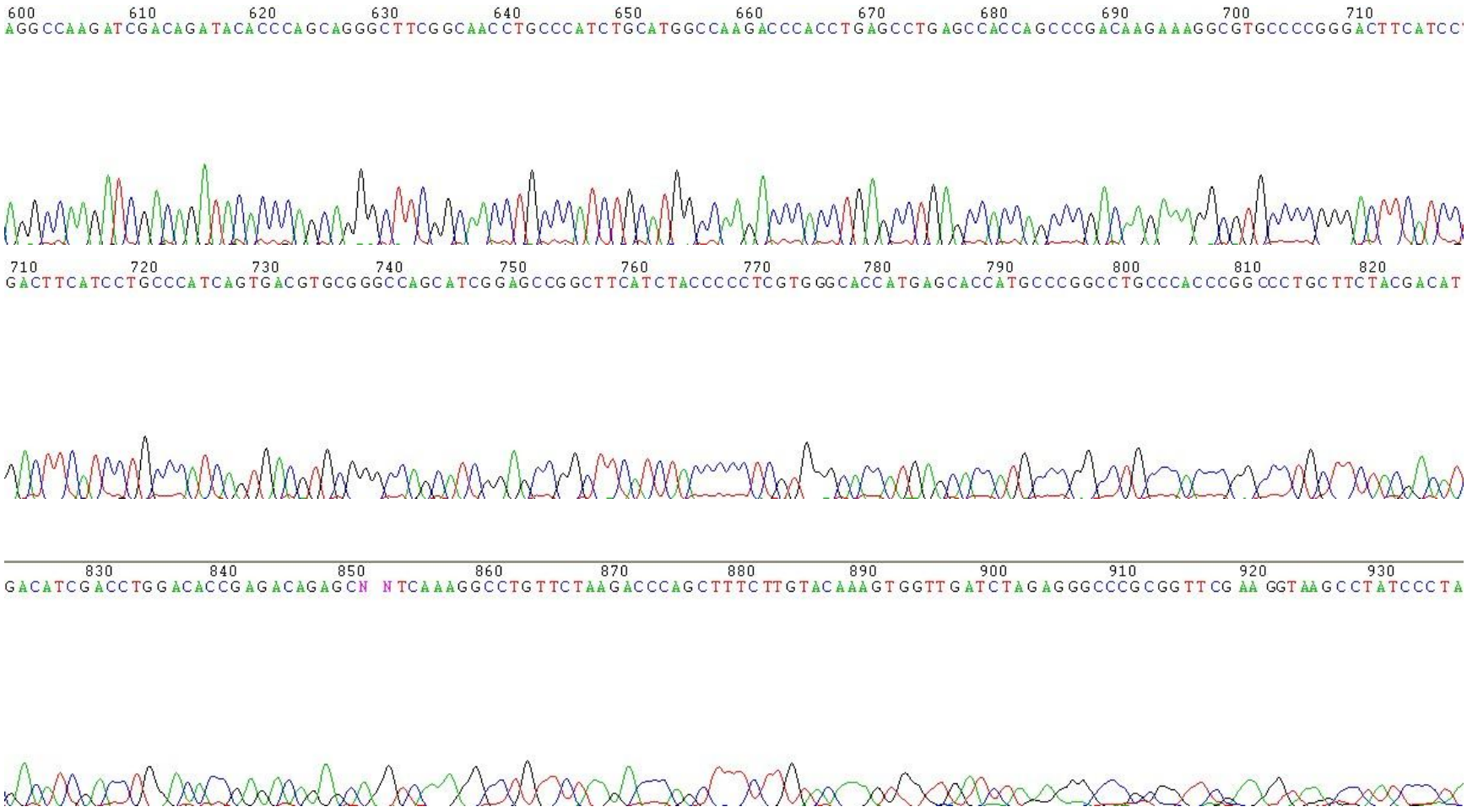

390 400 410 420 430 440 450 460 470 480 490 500 510
GACCCTGGAAGGCACCCCCGTGTTCGTGCACGCCGGACCCTTCGCCAATATCGCCCACGGCAACAGCAGCGTGCTGGCCGATAAGATTGCCCTGAAGCTCGTGGGCGAAGAGGGCTTCGTGG
510 520 530 540 550 560 570 580 590 600 610 620 63
GGTCACCGAGGCCGGCTTTGGCGCCGACATCGGCATGGAAAAGTTCTTCAACATCAAGTGCCGGGCCAGCGGCCTGGTGCCCAACGTGGTGGTGCTGGTCGCCACAGTGCGGGCCCTGAAGA
630 640 650 660 670 680 690 700 710 720 730 740 750
GATGCACGGCGGAGGCCCATCTGTGACAGCCGGCGTGCCCCTGAAGAAAGAGTACACCGAAGAGAACATCCAGCTGGTGGCTGACGGCTGCTGCAACCCTGCAGAAGCAGATCCAGATCACCCA

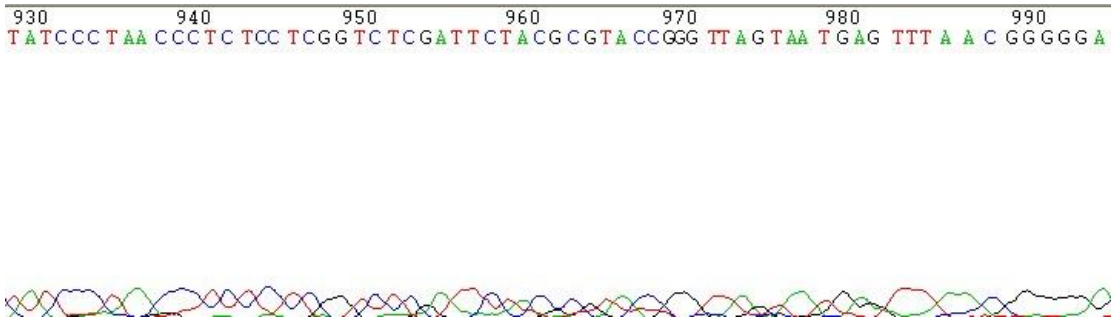

750 760 770 780 790 800 810 820 830 840 850 860
CCAGCTGTTCGGCGTGCCCGTGGTGGTGGCTCTGAACGTGTTCAAGACCGATACCAGAGCCGAGATCGACCTCGTGTGCGAGCTGGCCAAGCGGGCTGGCGCCTTCGATGCCGTGCCCTG

Appendix L

Sanger sequencing result for pcDNA3.2- MTHFD1Lopt using MTHFD1Lopt4 For primer (Table 2.4).

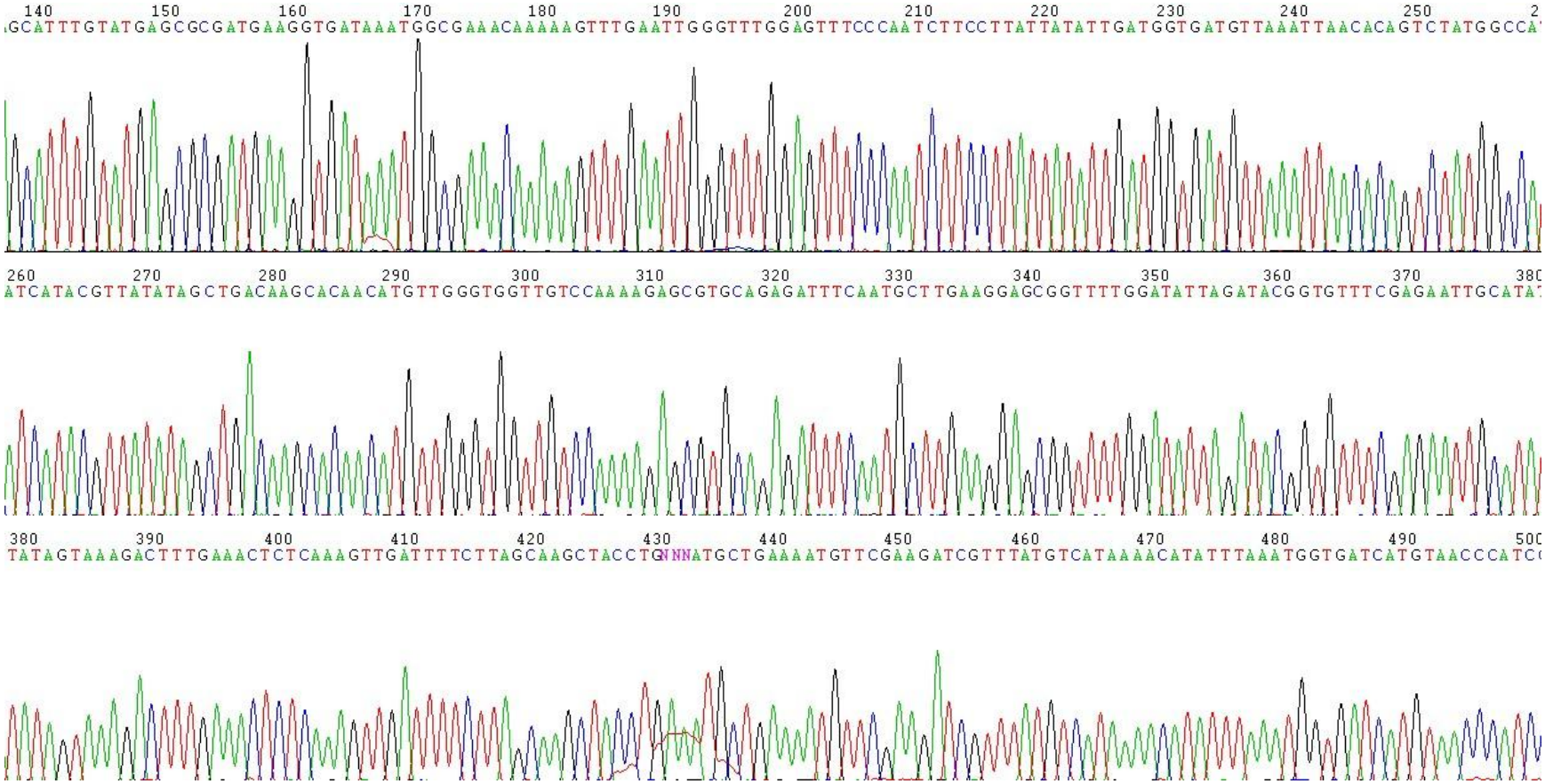


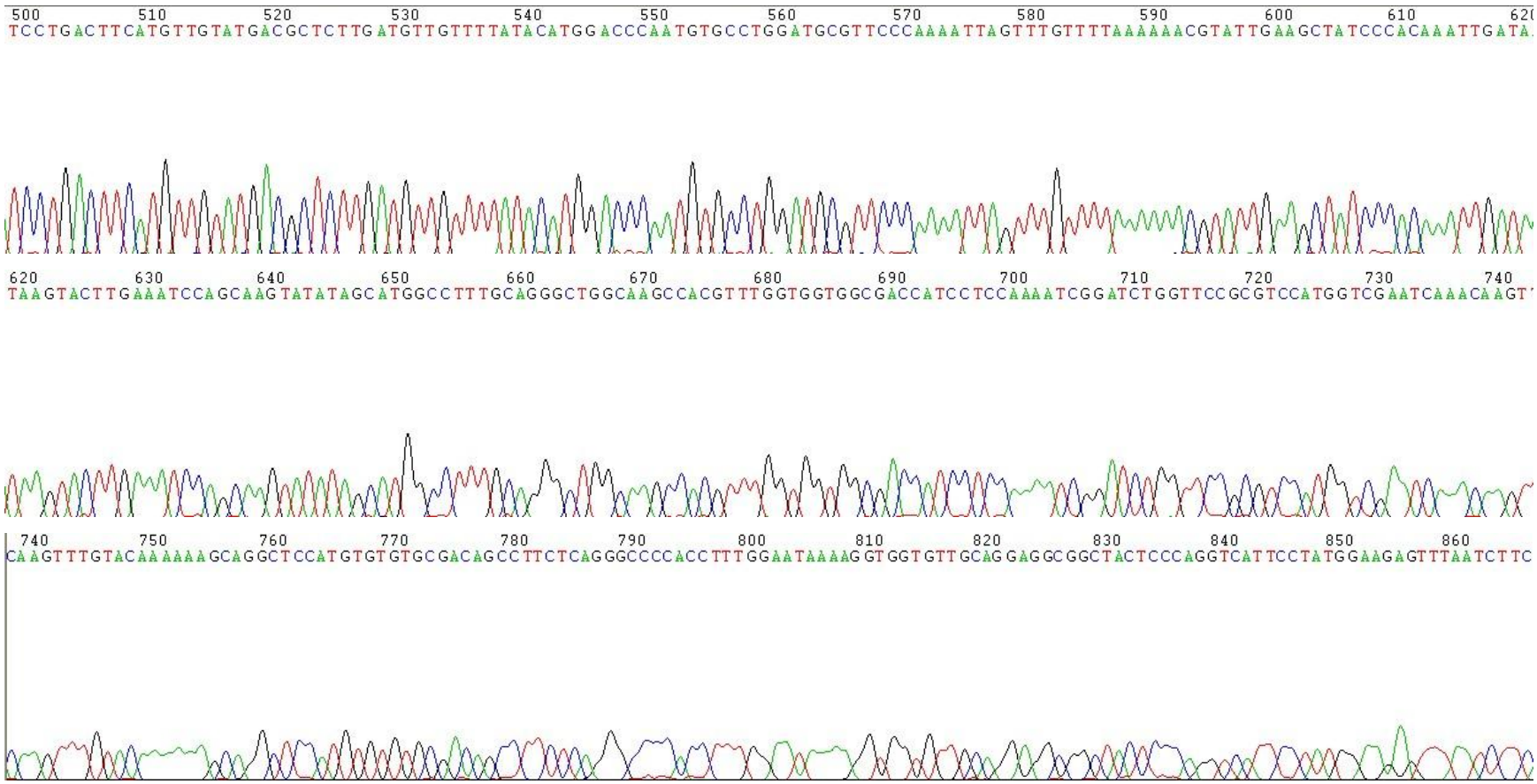

600 610 620 630 640 650 660 670 680 690 700 710
AGGCCAAGATCGACAGATACACCCAGCAGGGCTTCGGCAACCTGCCCATCTGCATGGCCAAGACCCACCTGAGCCTGAGCCACCAGCCCCGACAAGAAAGGCGTGCCCCGGGACTTCATCC
710 720 730 740 750 760 770 780 790 800 810 820
GACTTCATCCTGCCCATCAGTGACGTGCGGGCCAGCATCGGAGCCGGCTTCATCTACCCCCTCGTGGGCACCATGAGCACCATGCCCGGCCTGCCCACCCGGCCCTGCTTCTACGACAT
830 840 850 860 870 880 890 900 910 920 930
GACATCGACCTGGACACCGAGACAGAGCN NTCAAAGGCCTGTTCTAAGACCCAGCTTTCTTGTACAAAGTGGTTGATCTAGAGGGCCCGCGGTTCG AA GGTAAGCCTATCCCTA


930
940
950
960
970
980
990
TATCCCTAACCCTCTCCTCGGTCTCGATTCTACGCGTACCGGGTTAGTAATGAGTTTAACGGGGGA

Appendix M

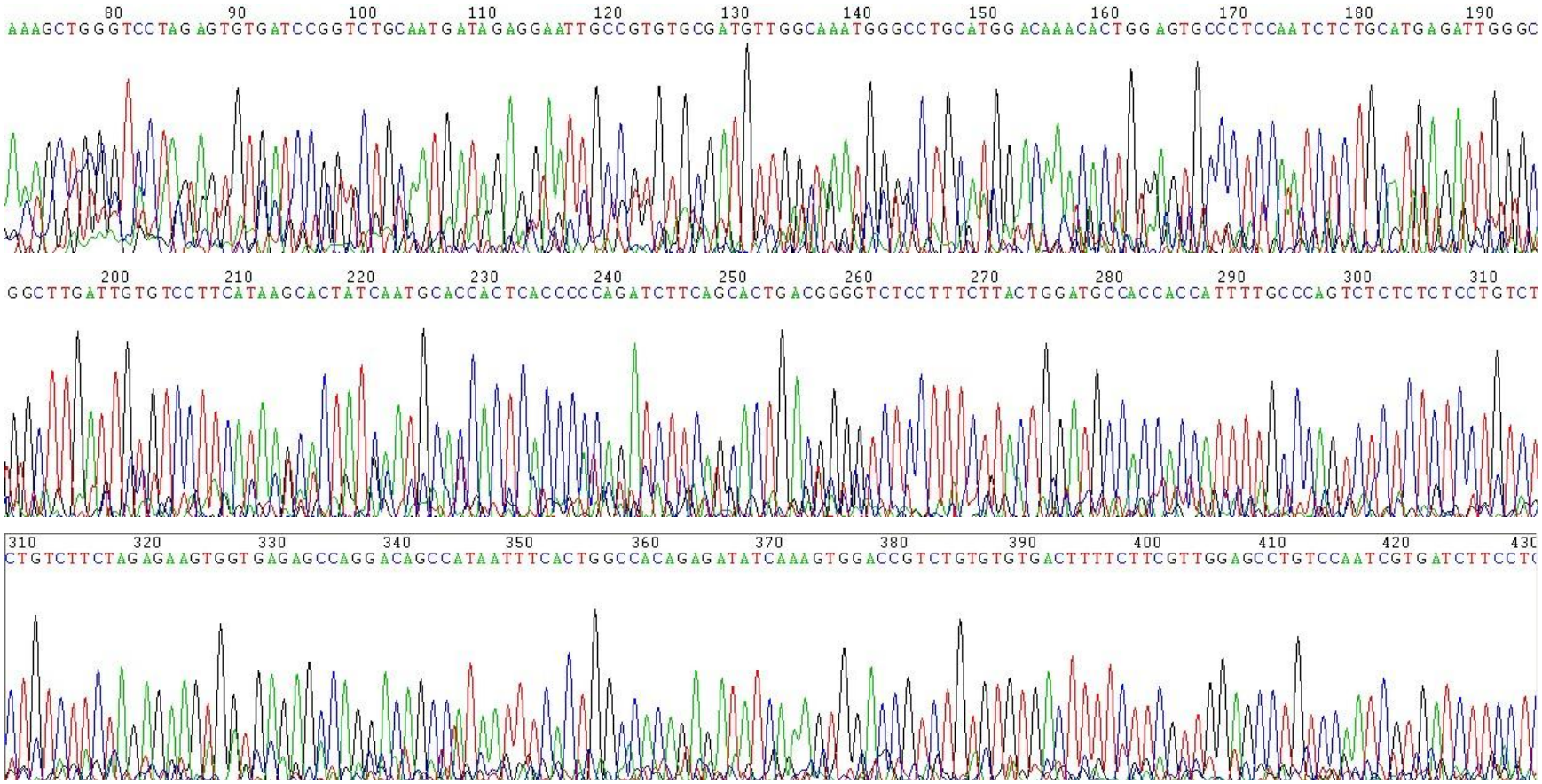
Sanger sequencing result for pDEST15-750 ORF using T7 Promoter For primer (Table 2.4).

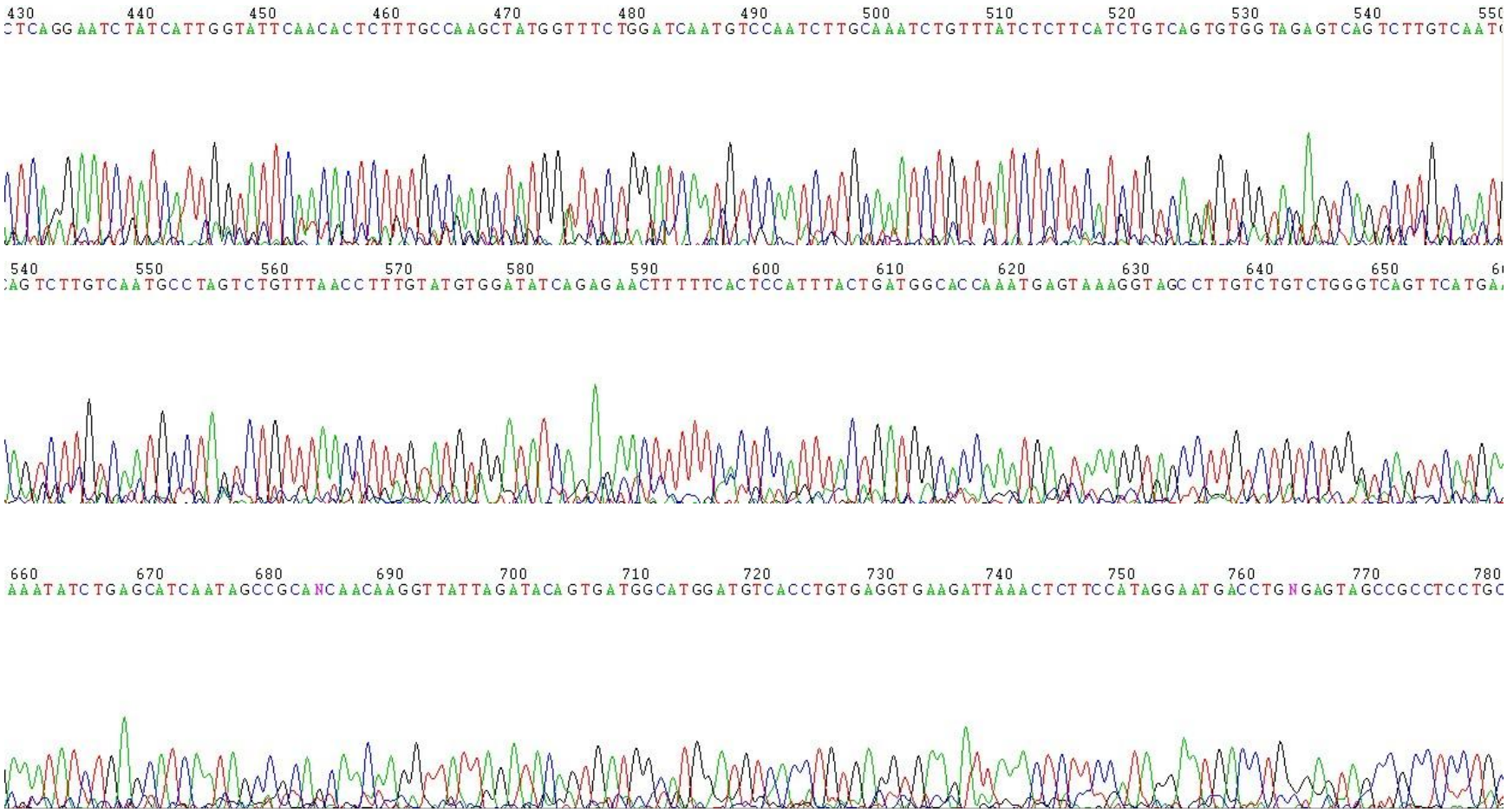



500 510 520 530 540 550 560 570 580 590 600 610
TCCTGACTTCATGTTGTATGACGCTCTTGATGTTGTTTTATACATGGACCCAATGTGCCTGGATGCGTTCCCAAAATTAGTTTGTTTTAAAAAAACGTATTGAAGCTATCCCACAAATTGATA
620 630 640 650 660 670 680 690 700 710 720 730 740
TAAGTACTTGAAATCCAGCAAGTATATAGCATGGCCTTTGCAGGGCTGGCAAGCCACGTTTGGTGGTGGCGACCATCCTCCAAAATCGGATCTGGTTCCGCGTCCATGGTCGAATCAAACAAGT
740 750 760 770 780 790 800 810 820 830 840 850 860
CAAGTTTGTACAAAAAAGCAGGCTCCATGTGTGTGCGACAGCCTTCTCAGGGCCCCACCTTTGGAATAAAAGGTGGTGTTGCAGGAGGCGGCTACTCCCAGGTCATTCCTATGGAAGAGTTTAATCTTC

Appendix N

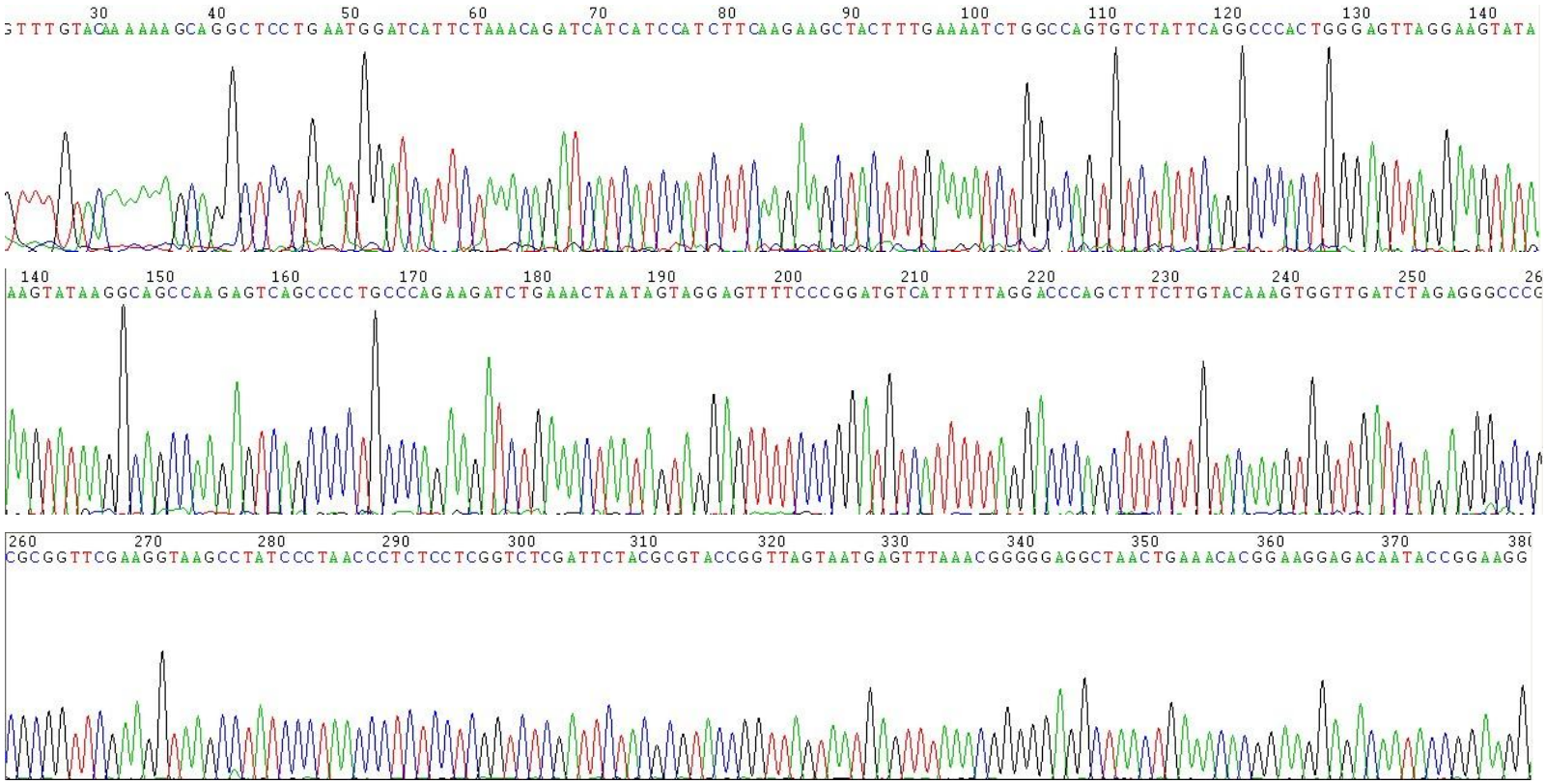
Sanger sequencing result for pDEST15-750 ORF using T7 Terminator Rev primer (Table 2.4).

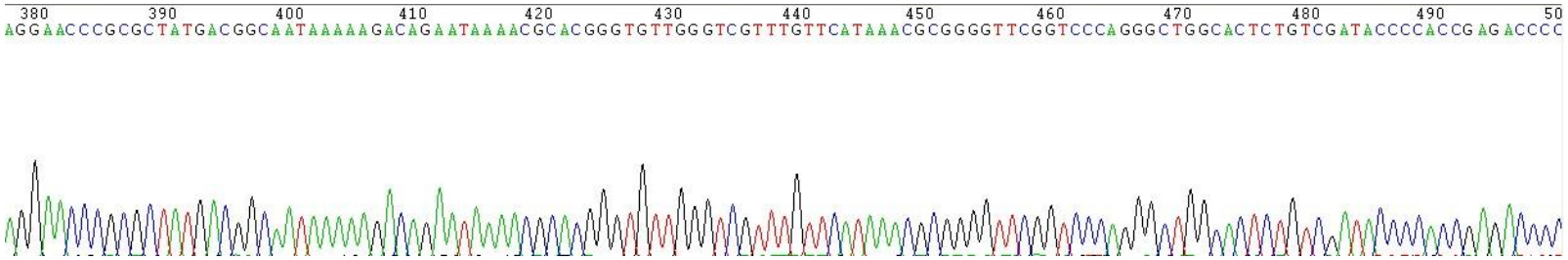



430 440 450 460 470 480 490 500 510 520 530 540
CTCAGGAATCTATCATTGGTATTCAACACTCTTTGCCAAGCTATGGTTTCTGGATCAATGTCCAATCTTGCAAATCTGTTTATCTCTTCATCTGTCAGTGTGGTAGAGTCAGTCTTGTCAATC
540 550 560 570 580 590 600 610 620 630 640 650
AGTCTTGTCAATGCCTAGTCTGTTTAACCTTTGTATGTGGATATCAGAGAACTTTTTCACTCCATTTACTGATGGCACCAAATGAGTAAAGGTAGCCTTGTCTGTCTGGGTCAGTTCATGA
660 670 680 690 700 710 720 730 740 750 760 770 780
AAATATCTGAGCATCAATAGCCGCANCAACAAGGTTATTAGATACAGTGATGGCATGGATGTCACCTGTGAGGTGAAGATTAAACTCTTCCATAGGAATGACCTGNGAGTAGCCGCCTCCTGC

Appendix O

Sanger sequencing result for pcDNA3.2-CHRX 3′UTR using T7 Promoter For primer (Table 2.4).

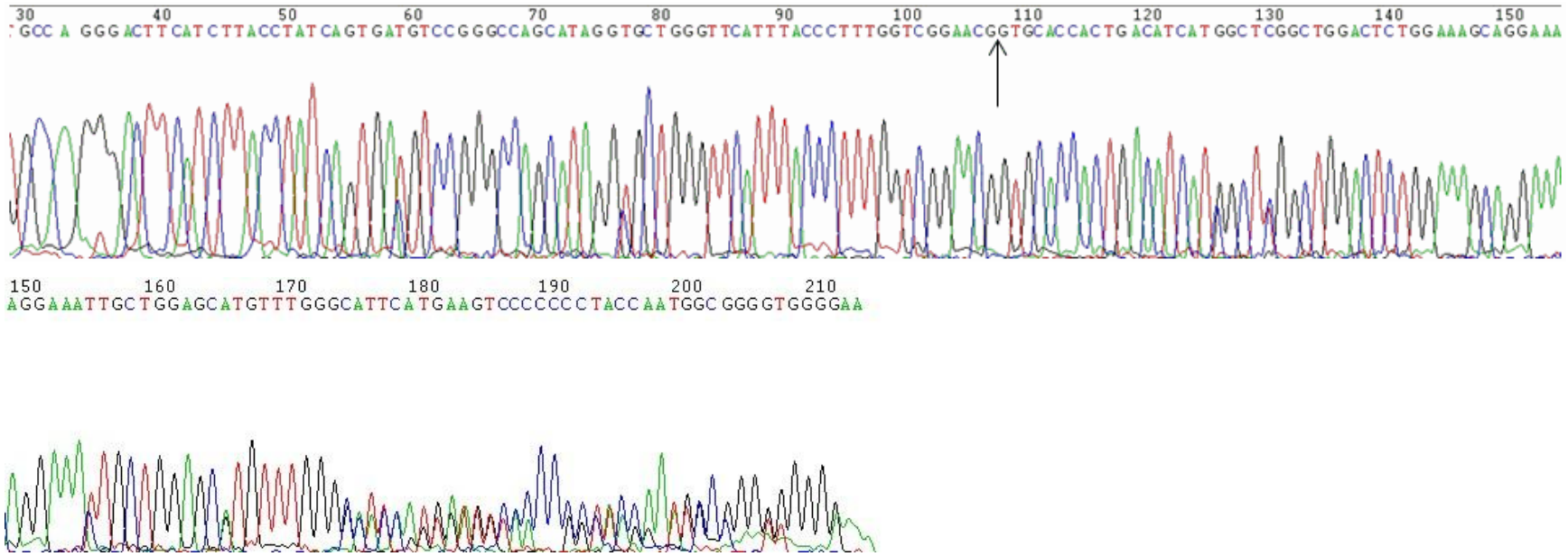



380 390 400 410 420 430 440 450 460 470 480 490 50
AGGAACCCGCGCTATGACGGCAATAAAAAGACAGAATAAAACGCACGGGTGTTGGGTCGTTTGTTCATAAACGCGGGGTTCGGTCCCAGGGCTGGCACTCTGTCGATACCCCACCGAGACCCC

Appendix P

Sanger sequencing result for NCBI 2 For/Rev PCR product using NCBI 2 For primer (Table 2.4).

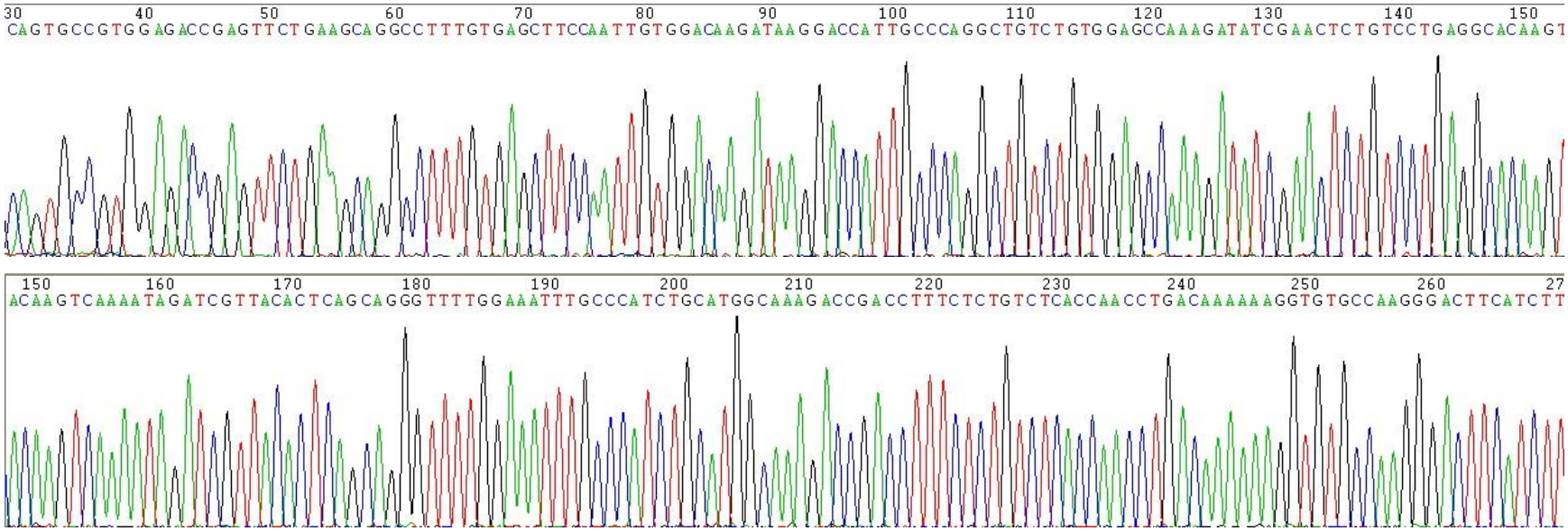
The amplified sequence spans exon 3 and 4 which are present in both ncRNAChr2 and ncRNAChr9 (Chapter 5, Figure 5.3a). The arrow marks the hedge between exon 3 and 4.

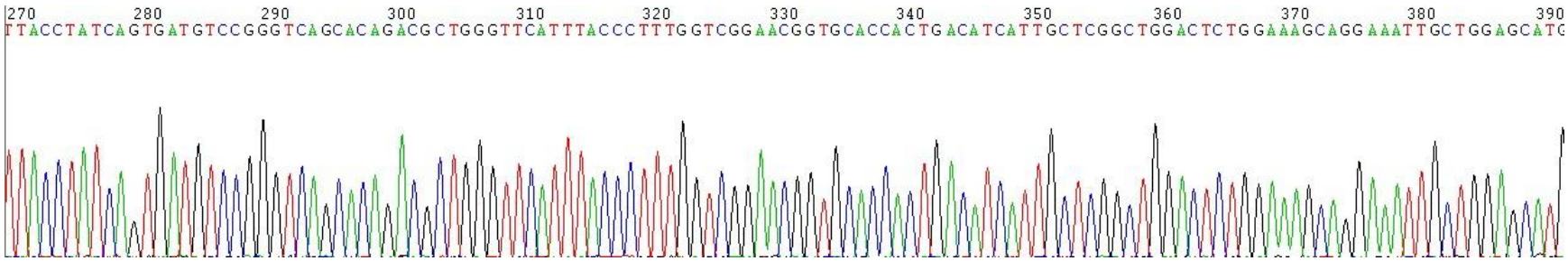


Appendix Q

Sanger sequencing result for ncRNAChr2 using NCBI All For primer (Table 2.4).

The sequence was amplified using NCBI All For/ NCBI2 Rev primers (Chapter 5, Figure 5.3).

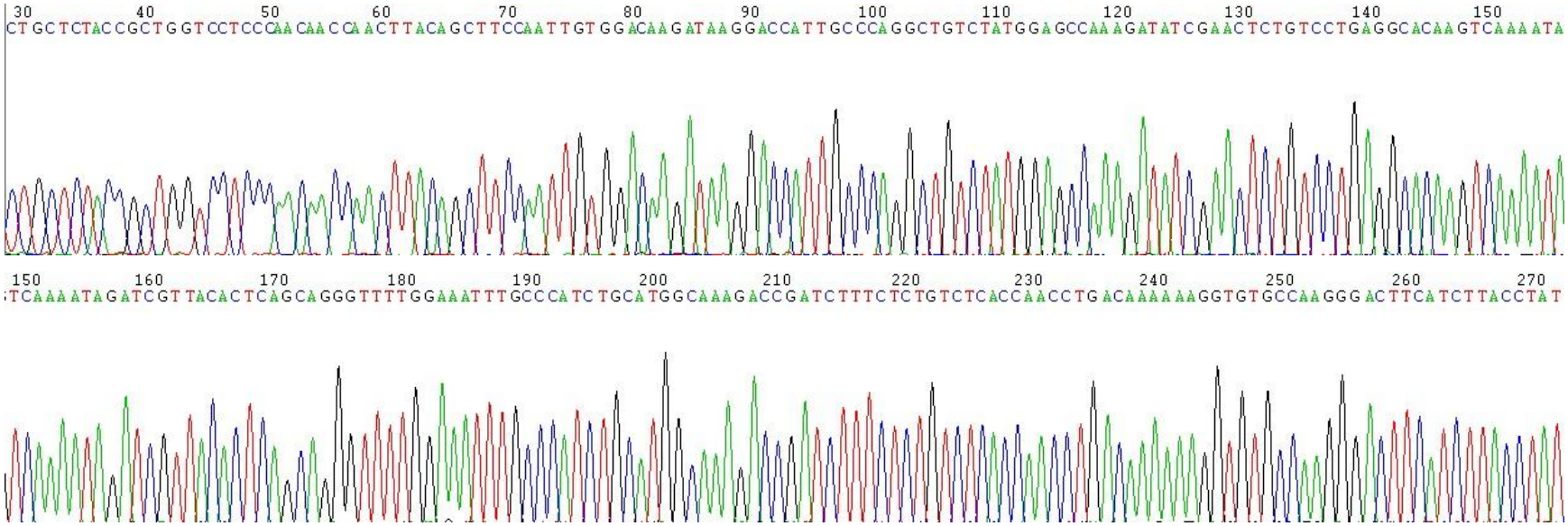


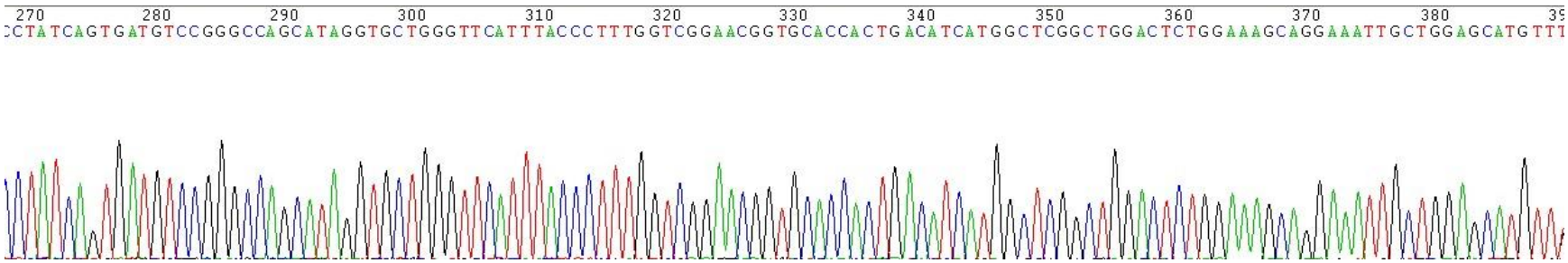

270 280 290 300 310 320 330 340 350 360 370 380 390
TTACCTATCAGTGATGTCCGGGTCAGCACAGACGCTGGGTTCATTTACCCTTTGGTCGGAACGGTGCACCACTGACATCATTGCTCGGCTGGACTCTGGAAAGCAGGAAATTGCTGGAGCATG

Appendix R

Sanger sequencing result for ncRNAChr9 using NCBI 5 For primer (Table 2.4).

The sequence was amplified using NCBI 5 For/ NCBI2 Rev primers (Chapter 5, Figure 5.3).



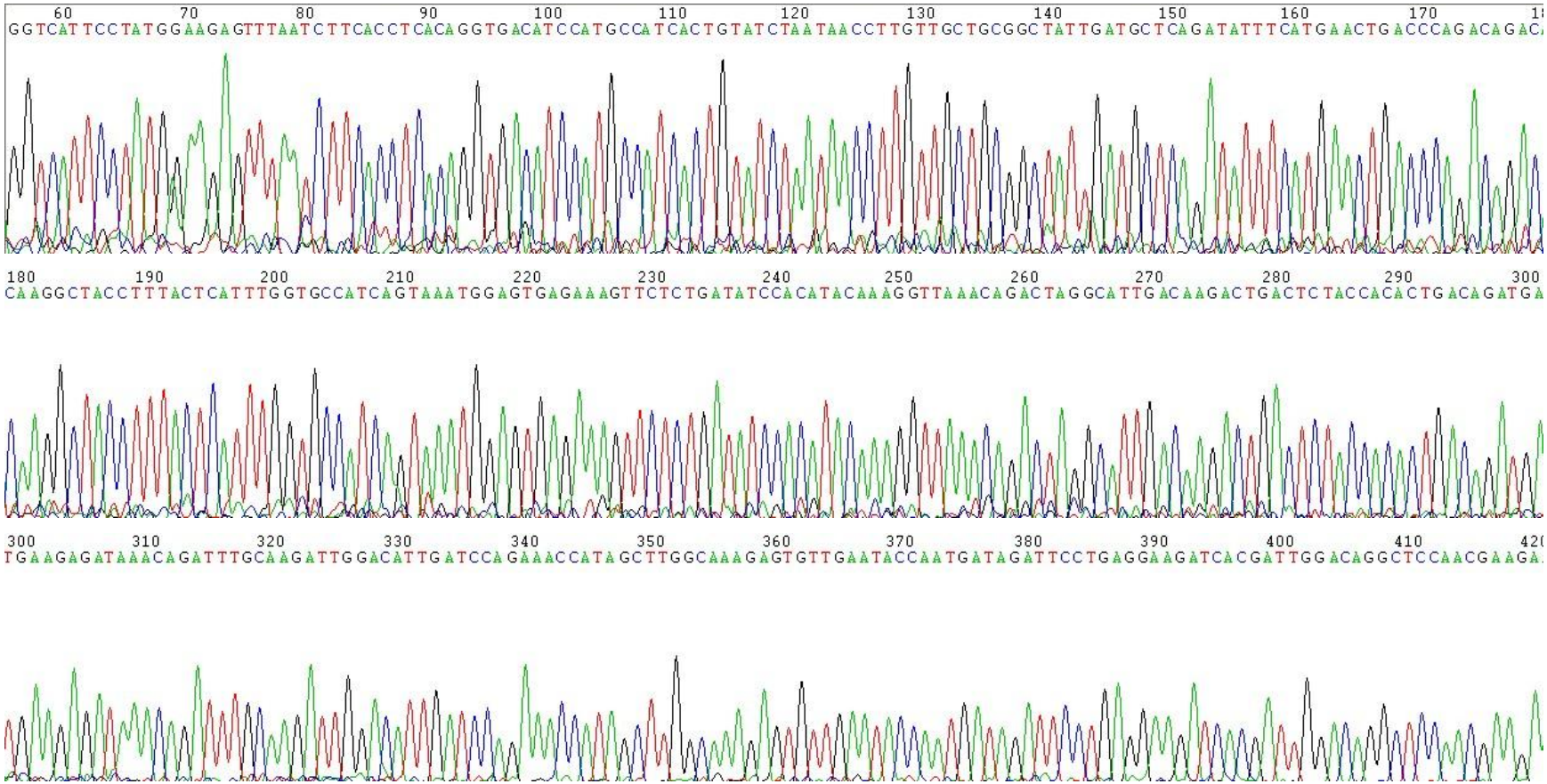

270 280 290 300 310 320 330 340 350 360 370 380
CCTATCAGTGATGTCCGGGCCAGCATAGGTGCTGGGTTCATTTACCCTTTGGTCGGAACGGTGCACCACTGACATCATGGCTCGGCTGGACTCTGGAAAGCAGGAAATTGCTGGAGCATGTTT

Appendix S

Sanger sequencing result for DIP rs35337982 validation (Chapter 3, Section 3.3.3) using A insertion For For primer (Table 2.4). The arrow indicates the position of the hypothetical polymorphism, namely the insertion of an extra 'A'.

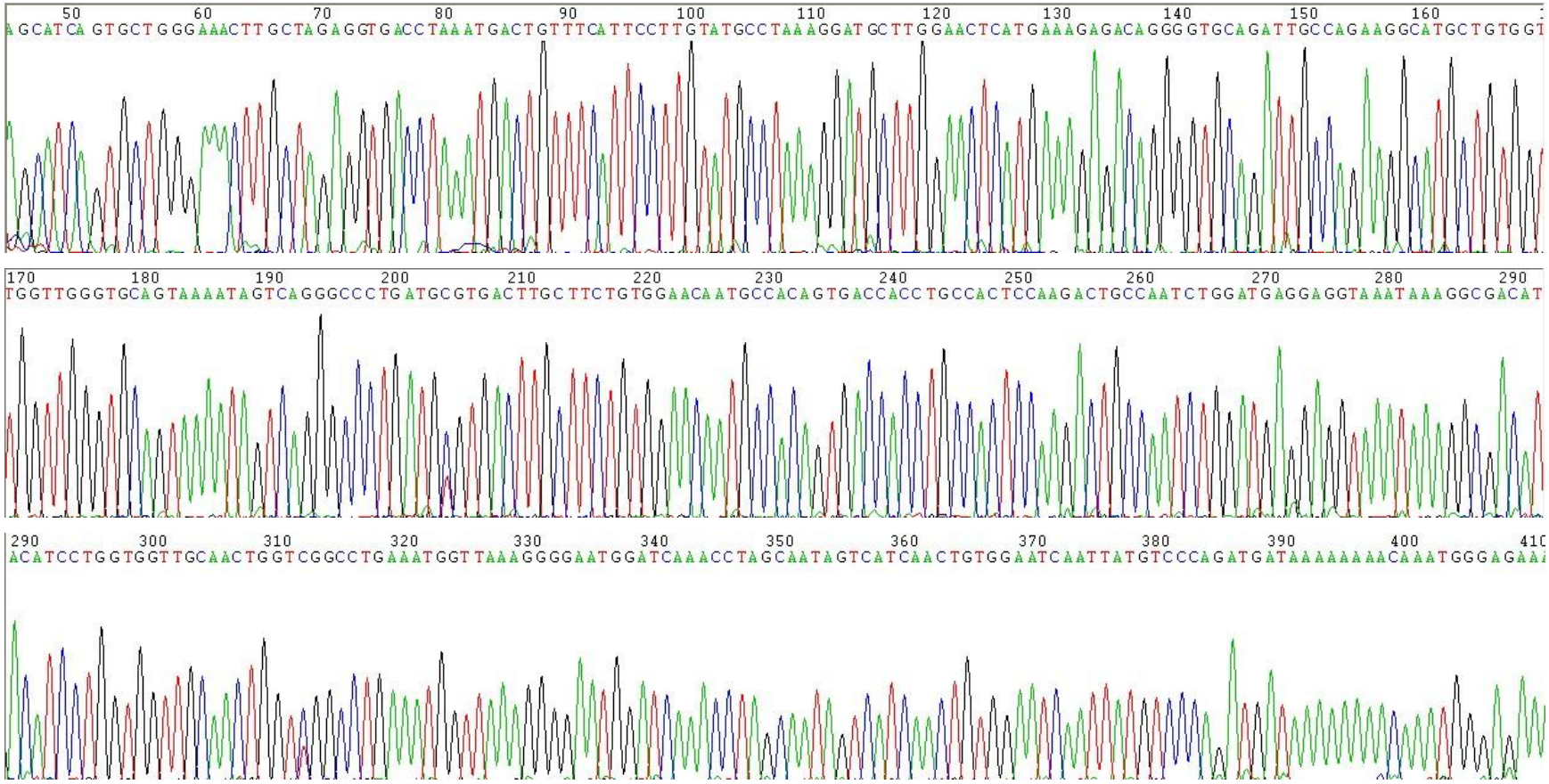
Appendix T

Sanger sequencing result for CHRX 7For/4Rev band (Chapter 5, Figure 5.11) using CHRX 7For primer (Table 2.4).



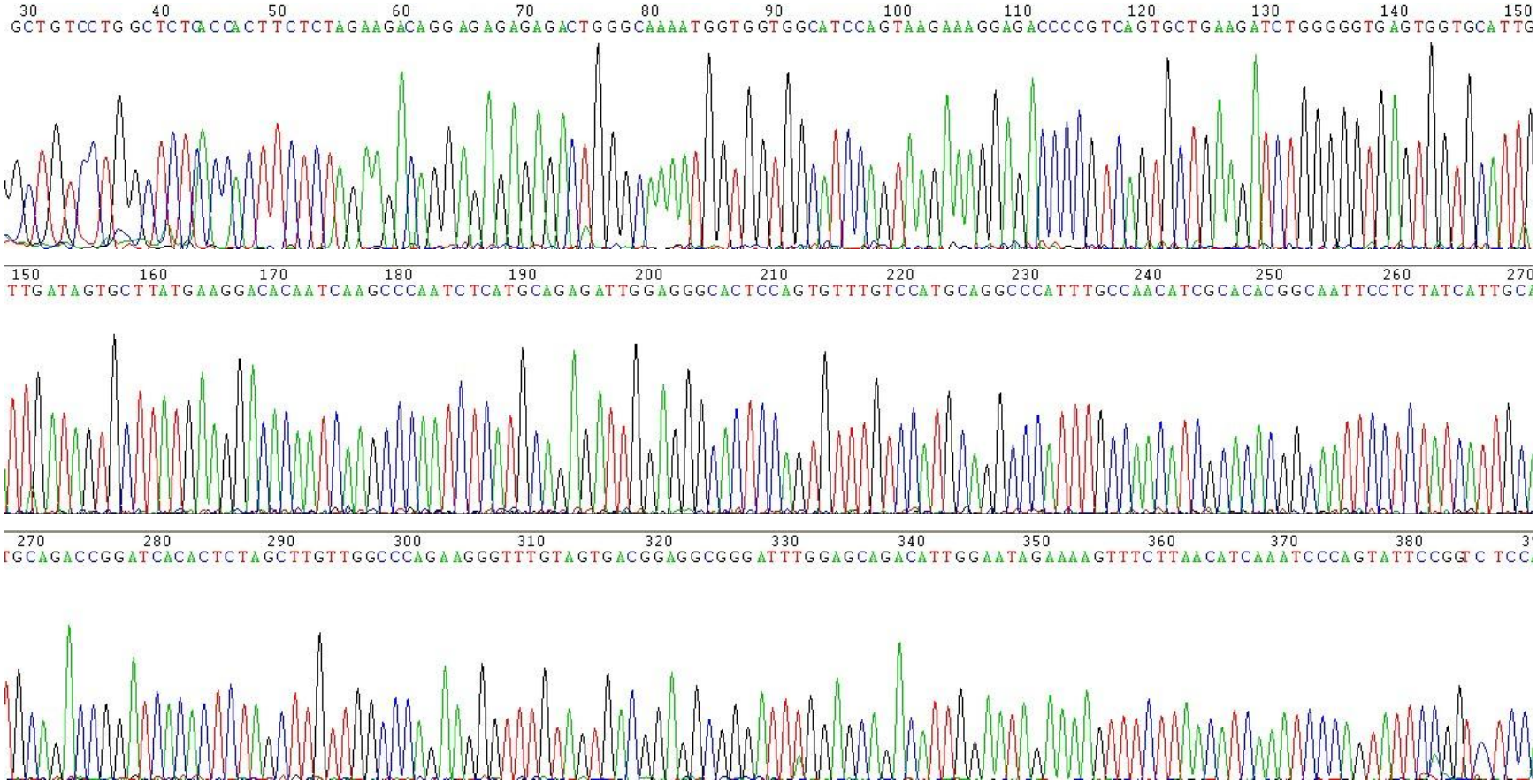
Appendix U

Sanger sequencing result for CHRX 5For/5Rev band (Chapter 5, Figure 5.11) using CHRX 5For primer (Table 2.4).



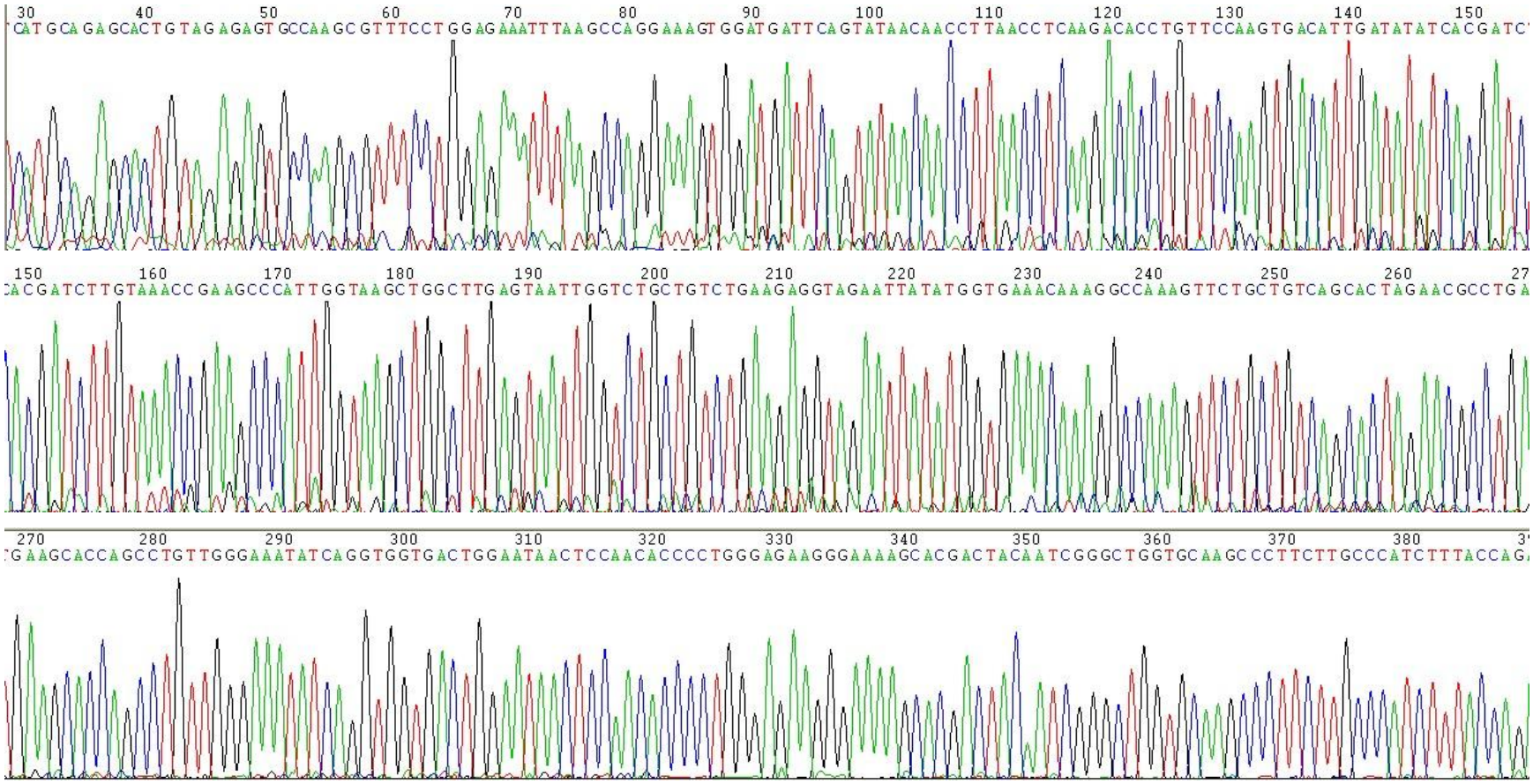
Appendix V

Sanger sequencing result for CHRX 3For/3Rev band (Chapter 5, Figure 5.11) using CHRX 3For primer (Table 2.4).



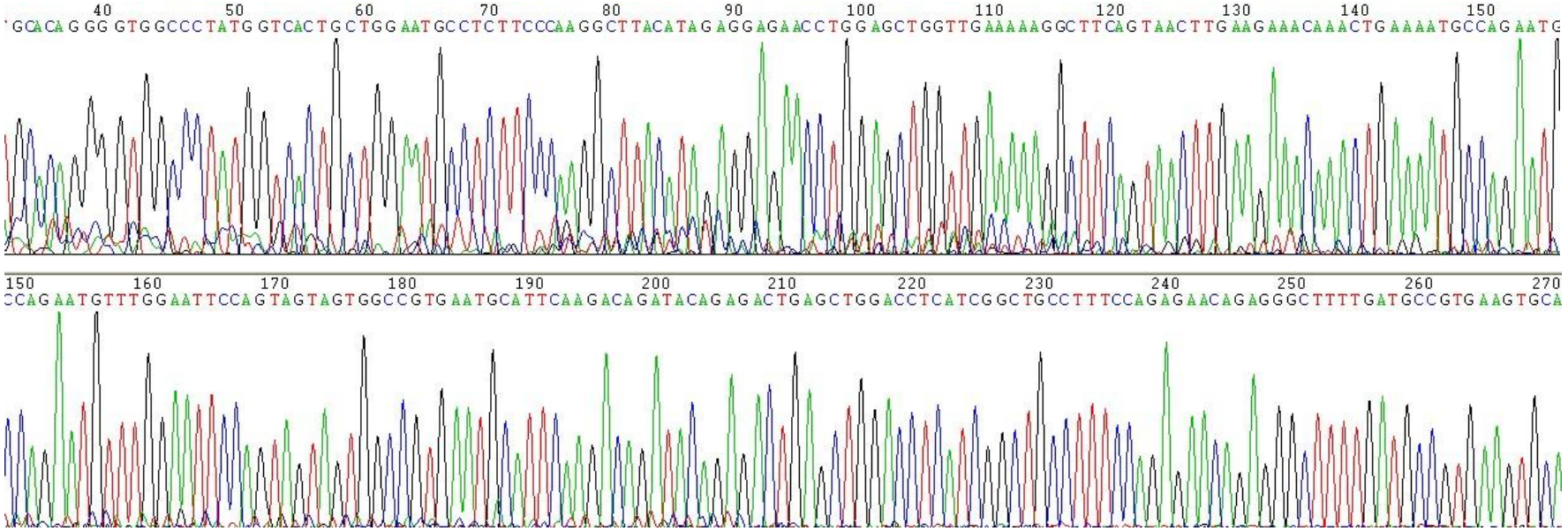
Appendix W

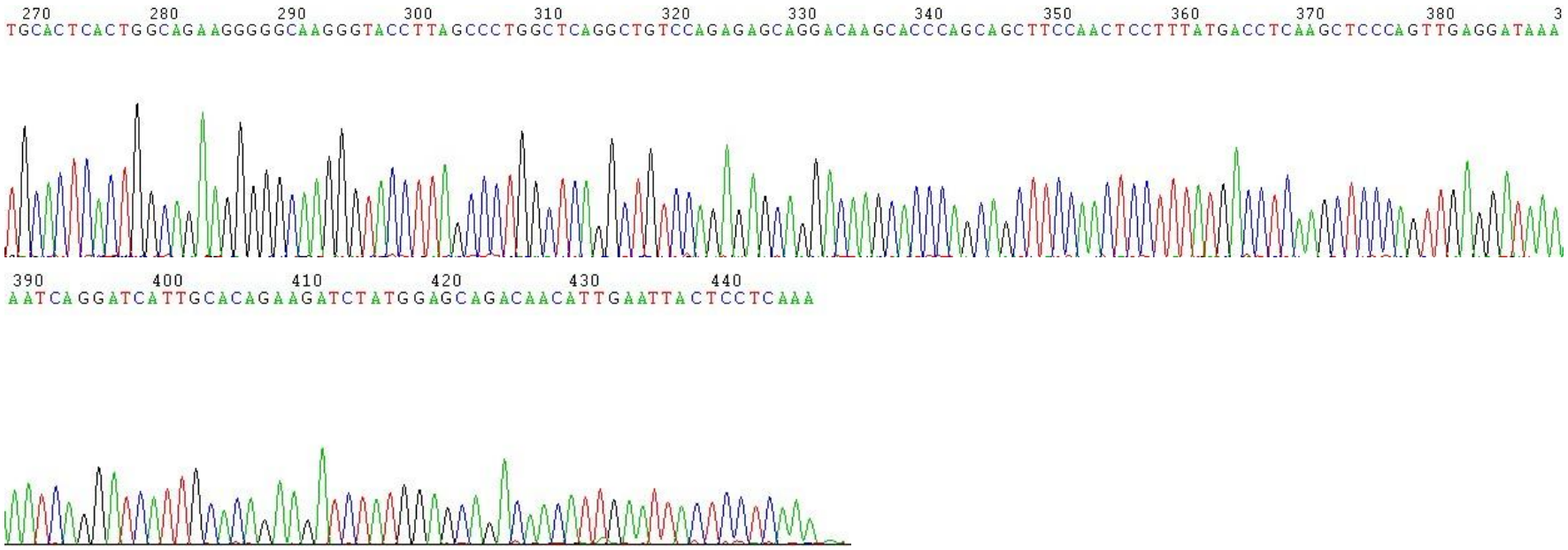
Sanger sequencing result for CHRX 6For/6Rev band (Chapter 5, Figure 5.11) using CHRX 6For primer (Table 2.4).



Appendix X

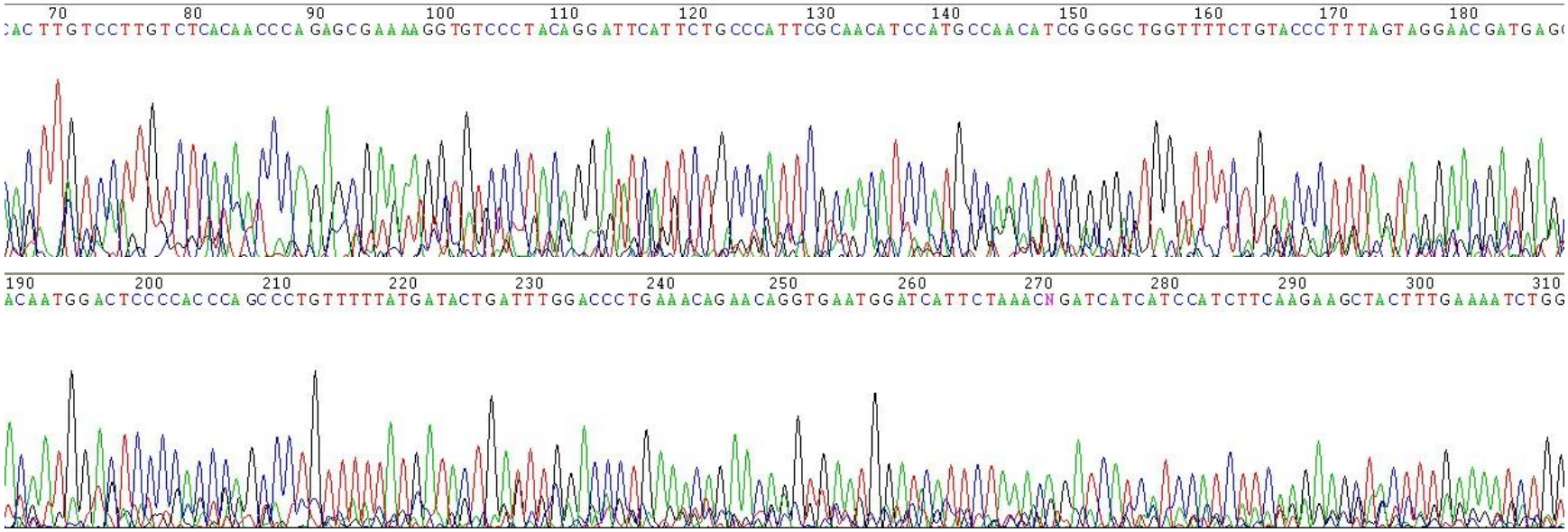
Sanger sequencing result for CHRX 2For/2Rev band (Chapter 5, Figure 5.11) using CHRX 2For primer (Table 2.4).

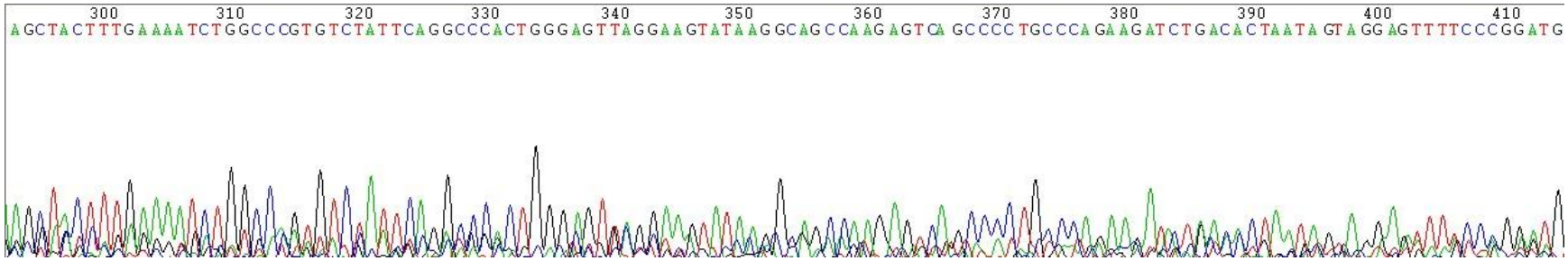



270 280 290 300 310 320 330 340 350 360 370 380 3
TGCACTCACTGGCAGAAGGGGGCAAGGGTACCTTAGCCCTGGCTCAGGCTGTCCAGAGAGCAGGACAAGCACCCAGCAGCTTCCAACTCCTTTATGACCTCAAGCTCCCAGTTGAGGATAAA
390 400 410 420 430 440
AATCAGGATCATTGCACAGAAGATCTATGGAGCAGACAACATTGAATTACTCCTCAAA

Appendix Y

Sanger sequencing result for CHRX 1For/3′UTR 2nd Rev band (Chapter 5, Figure 5.11) using CHRX 1For primer (Table 2.4).




300 310 320 330 340 350 360 370 380 390 400 410
AGCTACTTTGAAAATCTGGCCCGTGTCTATTCAGGCCCACTGGGAGTTAGGAAGTATAAGGCAGCCAAGAGTCA GCCCCTGCCCAGAAGATCTGACACTAATAGTAGGAGTTTTCCCGGATG

Appendix Z

MTHFD1L and Chr11seq sequence alignment made using CLUSTALW (Chapter 2, Section 2.2.14.3).

Position of the primer CHR11 AFor and CHR11 ARev are highlighted in yellow and green, respectively.

```
SeqA Name           Len(nt)  SeqB   Name           Len(nt)  Score
=====================================================================
1    CHR11           1287     2    MTHFD1L          1291     88
=====================================================================


CHR11           CTGATGCAGACCCTGGAATGGACACCTGTATTCATGCACGTGGGCCCTTTTGCTAACATC 60
MTHFD1L         CTGATGCAGACCCTGGAAGGGACACCTGTGTTCGTGCATGCGGGCCCTTTTGCTAACATT 60
                ****************** ********** *** **** * ******************

CHR11           ACCCATGGCAATTCTTCAGTGTTGGCTGACAAAATTGCCCTGAAACTGGTTGGTGAAGAA 120
MTHFD1L         GCTCACGGCAACTCTTCAGTGTTGGCTGATAAAATTGCCCTGAAACTGGTTGGTGAAGAA 120
                 * ** ***** ***************** ******************************

CHR11           GGATTTGTAGTGACTGAAGCTGGCTTTGGTGCTGACATTGAAATGGAGAAATTTTTCAAC 180
MTHFD1L         GGATTTGTAGTGACCGAAGCTGGCTTTGGTGCTGACATCGGAATGGAGAAATTCTTCAAC 180
                ************** ********************** * ************ ******

CHR11           A---AGTGTCAAGCTTCCAGCTTGGTGCCCAGTGTGGTTGTGTTGGTGGCAACGGTGCAA 237
MTHFD1L         ATCAAGTGCCGAGCTTCCGGCTTGGTGCCCAACGTGGTTGTGTTAGTGGCAACGGTGCGA 240
                *   **** * ******* ************  *********** ************* *

CHR11           GCTCTGAAGATGCACAGTGGCGGGCCAAGTGTAACTGCTGGTGTTCCTCTGAAGAAAGAA 297
MTHFD1L         GCTCTGAAGATGCATGGAGGCGGGCCAAGTGTAACGGCTGGTGTTCCTCTTAAGAAAGAA 300
                **************  * ***************** ************* *********

CHR11           CATAAAGAAAAGAACATCCAGTTGGTGGCAGATGGCTGCCGTAACCTCTAGCAGCAAATT 357
MTHFD1L         TATACAGAGGAGAACATCCAGCTGGTGGCAGACGGCTGCTGTAACCTCCAGAAGCAAATT 360
                 *** ***  *********** ********** ****** ******** ** ********

CHR11           CAGATTGCTCAGCTCTTTGGGGTTCCCATTATGGTGACTCTGAAGGTCTTCAAGATCCAC 417
MTHFD1L         CAGATCACTCAGCTCTTTGGGGTTCCCGTTGTGGTGGCTCTGAATGTCTTCAAGACCGAC 420
                *****  ******************** ** ***** ******* ********** * **

CHR11           ACCGGTGCTGAGATTGACTTGGTGTGTGAGCTCACAAAACGGGCTGGTGCCTTTGATGCA 477
MTHFD1L         ACCCGCGCTGAGATTGACTTGGTGTGTGAGCTTGCAAAGCGGGCTGGTGCCTTTGATGCA 480
                *** * ***************************  **** ********************

CHR11           GTCCCTTGCTATCACTGGTCAGTTGGTAGAAAACGGTTG-CGGACTTGGCTTGGGCTTTG 536
MTHFD1L         GTCCCCTGCTATCACTGGTCCGTTGGTGGAAAAGGATCGGTGGACTTGGCTCGGGCTGTG 540
                ***** ************** ****** ***** * * *  ********** ***** **

CHR11           AGAGAAGCTGCAAGTAAAAGAAGTTGATTCCATTTCCCGTATGATGTTCAGCTTCCAATT 596
MTHFD1L         AGAGAGGCTGCGAGTAAAAGAAGCCGATTCCAGTTCCTGTATGATGTTCAGGTTCCAATT 600
                ***** ***** ************  ******* **** ************* ********

CHR11           GTGGAAAAGATAAGAACCATGGCCCTGGCTGTCTATGGAGCCAACGATATCCAGCTCTCT 656
MTHFD1L         GTGGACAAGATAAGGACCATTGCTCAGGCTGTCTATGGAGCCAAAGATATTGAACTCTCT 660
                ***** ******** ***** ** * ****************** *****  * ******

CHR11           CCTGAGGCACAAACCAAAATAGATTGTTAACACTCAACAGGGTTTCAGAAATTTGCCCAT 716
```

```
MTHFD1L         CCTGAGGCACAAGCCAAAATAGATCGTTA-CACTCAACAGGGTTTTGGAAATTTGCCCAT 719
                ************ *********** **** ***************  *************

CHR11           CTGCATGGCAAGGACCCACCTTTCTCTGTCTCACCAACCTGCCAAAAAAT-TGTGCCAAG 775
MTHFD1L         CTGCATGGCAAAGACCCACCTTTCTCTATCTCACCAACCTGACAAAAAAGGTGTGCCAAG 779
                *********** *************** ************ *******  *********

CHR11           GGACTTCATTGTACCTCTCAGTGATGTCAAGGCCAGCATAGGTGCTGGGTTCATTGGCCC 835
MTHFD1L         GGACTTCATCTTACCTATCAGTGACGTCCGGGCCAGCATAGGCGCTGGGTTCATTTACCC 839
                *********  ***** ******* ***  ************ ************  ***

CHR11           TTTGGTTAGAACAATGAGCATCGTGCCAGGACTGCC-ACTTGGCCCTGCTTTCATGACAT 894
MTHFD1L         TTTGGTCGGAACGATGAGCACCATGCCAGGACTGCCCACCCGGCCCTGCTTTTATGACAT 899
                ******  **** ******* * ************ **  *********** *******

CHR11           AGACCTTGATATGGAAACAGAACAAATTAAAGCCTTGTTCTAAGTGAACAAGGCTTTCAT 954
MTHFD1L         AGATCTTGATACCGAAACAGAACAAGTTAAAGGCTTGTTCTAAGTGGACAAGGCTCTCAC 959
                *** *******  ************ ****** ************ ******** ***

CHR11           AGGAATTGATTCAGGTTCCTGAAACAGACTCCTCCTTGCCTTTTTGCTGCAGTTAGAGAA 1014
MTHFD1L         AGGACCCGATGCAGACTCCTGAAACAGACTACTCTTTGCCTTTTTGCTGCAGTTGGAGAA 1019
                ****   *** ***  ************** *** ****************** *****

CHR11           GAAACTGAATTTTAAAAATGTCCTTTATGCAATGTTGGAGAAATGGTGAAATAGGTCAAG 1074
MTHFD1L         GAAACTGAATTTGAAAAATGTCTGTTATGCAATGCTGGAGACGTGGTGAAATAGGCCAAA 1079
                ************ *********  ********** ******  ************ ***

CHR11           GATTTCTTCCTCATTGAAGGTGTATCCTATTCACAACAGAGTATAGTGTTCAGCAAAAGG 1134
MTHFD1L         GATTTCTTCTTCGTTCAAGATGAATTCTGTTCACAGTGGAGTATGGTGTTCGGCAAAAGG 1139
                ********* ** ** *** ** ** ** ******   ****** ****** ********

CHR11           ACCTCCACCAAGTCTGAAAAAAAATAATTTACTTAAGTTTCAATAGAGTTTACATTATTT 1194
MTHFD1L         ACCTCCACCAAGACTGAAAGAAACTAATTTATTTCTGTTTCTGTGGAGTTTCCATTATTT 1199
                ************ ****** *** ******* **  *****  * ****** ********

CHR11           CTACTGCTTACACTTTAGAATGTTAATTTTATGGG-ACTAAGGGATTAAAAGAGTGTGAA 1253
MTHFD1L         CTACTGCTTACACTTTAGAATGTTTATTTTATGGGGACTAAGGGATTA--GGAGTGTGAA 1257
                ********************** ********** ************   *********

CHR11           CTACAAGGTAACATTTTCCACTCTTGAGTTTTCT 1287
MTHFD1L         CTAAAAGGTAACATTTTCCACTCTCAAGTTTTCT 1291
                *** ********************  ********
```

Appendix AA

MTHFD1 mRNA and ChrXseq sequence alignment performed using CLUSTALW (Chapter 2, Section 2.2.14.3).

Position of the primers are highlight as follows: ChrX 5For (gray), ChrX 6For/ ChrX 5Rev (purple), ChrX 7For/ ChrX 6Rev (red), ChrX 3For/ ChrX 4Rev (light blue), ChrX 2For/ ChrX 3Rev (green), ChrX 1For/ ChrX 2Rev (yellow), ChrX 3′UTR For (gold), ChrX 3′UTR 2nd Rev (underlined). All the reverse primers were designed on the antisense strand.

```
CLUSTAL 2.1 Multiple Sequence Alignments

Sequence 1: MTHFD1      3466 bp
Sequence 2: ChrXseq      3005 bp

MTHFD1          AATTACGGCCGGATTCCGGAGTCCTTTCCAGCTCCCTCTTCGGCCGGGTTTCCCGCCGAA
ChrXseq         ------------------------------------------------------------

MTHFD1          TACAAAGGCGCACTGTGAACTGGCTCTTTCTTTCCGCCAATCATTTCCGCCAGCCATTCA
ChrXseq         ------------------------------------------------------------

MTHFD1          TCACCGATTTTCTTCATCTTCCCCTCCCTCTTCCGTCCCGCAGTCCCCGACCTGTTAGCT
ChrXseq         ------------------------------------------------------------

MTHFD1          CTCGGTTAGTTAAGGGACTCGGGTCCTTCCGAACTGCGCATGCGCCACCGCGTCTGCAGG
ChrXseq         ------------------------------------------------------------

MTHFD1          GGGAGAAGCGGGCAGGGGCGCAGGCGCAGTAGTGTGATCCCCTGGCCAGTCCCTAAGCAC
ChrXseq         ------------------------------------------------------------

MTHFD1          GTGGGTTGGGTTGTCCTGCTTGGCTGCGGAGGGAGTGGAACCTCGATATTGGTGGTGTCC
ChrXseq         ----------------------------------------GATATTGGTGGTGTCC
                                                          ****************

MTHFD1          ATCGTGGGCAGCGGACTAATAAAGGCCATGGCGCCAGCAGAAATCCTGAACGGGAAGGAG
ChrXseq         ATCGTGGGCAGCAGACTAATAAAGGCCATGGCACCAGCAGAAATCCTGAACAGGAAGGAG
                ************.******************.*****************.********

MTHFD1          ATCTCCGCGCAAATAAGGGCGAGACTGAAAAATCAAGTCACTCAGTTGAAGGAGCAAGTA
ChrXseq         ATCTCCGCACAAATAAGGGCGAGACTGAAGAATCAAGTCACTCAGTTGAAG-GACAAGTA
                ********.******************.********************  ..*******

MTHFD1          CCTGGTTTCACACCACGCCTGGCAATATTACAGGTTGGCAACAGAGATGATTCCAATCTT
ChrXseq         CCTGGTTTCACACCAGGCCTGGCAATATTACAGGTTGGCAACAGAGATGATTCCAATCTT
                *************** ******************************************

MTHFD1          TATATAAATGTGAAGCTGAAGGCTGCTGAAGAGATTGGGATCAAAGCCACTCACATTAAG
ChrXseq         TATATAATG-TGAAGCTGAAGGCTGCTGAAGAGATTGGGATGAAAGCCACTGACATTAAG
                *******:  ****************************** ********* ********

MTHFD1          TTACCAAGAACAACCACAGAATCTGAGGTGATGAAGTACATTACATCTTTGAATGAAGAC
ChrXseq         TTACCAAGAACAACCACAGAATCTGAGGTGATCAAGTACATCATATCTTTGAATGAAGAC
```

```
                ********************************* ******** * ****************

MTHFD1          TCTACTGTACATGGGTTCTTAGTGCAGCTACCTTTAGATTCAGAGAATTCCATTAACACT
ChrXseq         TCTACTGTACACGGGTTCGTAGTGCAGCTATCTCTAGACTCAGACAAT-------ACACT
                ********** ****** ********** ** **** ***** ***       *****

MTHFD1          GAAGAAGTGATCAATGCTATTGCACCCGAGAAGGATGTGGATGGATTGACTAGCATCAAT
ChrXseq         GAAGGAGTGATCAATGCTATTGCACCTGAAAAGGATGTGGATGGATTGACTAGCATCAGT
                ****.********************* **.****************************.*

MTHFD1          GCTGGGAAACTTGCTAGAGGTGACCTCAATGACTGTTTCATTCCTTGTACGCCTAAGGGA
ChrXseq         GCTGGGAAACTTGCTAGAGGTGACCTAAATGACTGTTTCATTCCTTGTATGCCTAAAGGA
                *************************.********************** ******.***

MTHFD1          TGCTTGGAACTCATCAAAGAGACAGGGGTGCCGATTGCCGGAAGGCATGCTGTGGTGGTT
ChrXseq         TGCTTGGAACTCATGAAAGAGACAGGGGTGCAGATTGCCAGAAGGCATGCTGTGGTGGTT
                ************** ****************.*******.********************

MTHFD1          GGGCGCAGTAAAATAGTTGGGGCCCCGATGCATGACTTGCTTCTGTGGAACAATGCCACA
ChrXseq         GGGTGCAGTAAAATAGTCAGGGCCCTGATGCGTGACTTGCTTCTGTGGAACAATGCCACA
                *** ************* .****** *****.****************************

MTHFD1          GTGACCACCTGCCACTCCAAGACTGCCCATCTGGATGAGGAGGTAAATAAAGGTGACATC
ChrXseq         GTGACCACCTGCCACTCCAAGACTGCCAATCTGGATGAGGAGGTAAATAAAGGCGACATC
                ****************************.************************ ******

MTHFD1          CTGGTGGTTGCAACTGGTCAGCCTGAAATGGTTAAAGGGGAGTGGATCAAACCTGGGGCA
ChrXseq         CTGGTGGTTGCAACTGGTCGGCCTGAAATGGTTAAAGGGGAATGGATCAAACCTAG--CA
                ******************.**********************.************.*  **

MTHFD1          ATAGTCATCGACTGTGGAATCAATTATGTCCCAGAT-GATAAAAAACCAAATGGGAGAAA
ChrXseq         ATAGTCATCAACTGTGGAATCAATTATGTCCCAGATGATAAAAAAAACAAATGGGAGAAA
                *********.************************** .::******.**************

MTHFD1          AGTTGTGGGTGATGTGGCATACGACGAGGCCAAAGAGAGGGCGAGCTTCATCACTCCTGT
ChrXseq         AGTTGTGGGTGATGTGGCATACGACGAG-------------------------GCCCGT
                ****************************                          ** **

MTHFD1          TCCTGGCGGCGTAGGGCCCATGACAGTTGCAATGCTCATGCAGAGCACAGTAGAGAGTGC
ChrXseq         TCCTGGTGGCTTAGGGCTCATGACGGTTGCAATGCTCATGCAGAGCACTGTAGAGAGTGC
                ****** *** ****** *******.************************:************

MTHFD1          CAAGCGTTTCCTGGAGAAATTTAAGCCAGGAAAGTGGATGATTCAGTATAACAACCTTAA
ChrXseq         CAAGCGTTTCCTGGAGAAATTTAAGCCAGGAAAGTGGATGATTCAGTATAACAACCTTAA
                ************************************************************

MTHFD1          CCTCAAGACACCTGTTCCAAGTGACATTGATATATCACGATCTTGTAAACCGAAGCCCAT
ChrXseq         CCTCAAGACACCTGTTCCAAGTGACATTGATATATCACGATCTTGTAAACCGAAGCCCAT
                ************************************************************

MTHFD1          TGGTAAGCTGGCTCGAGAAATTGGTCTGCTGTCTGAAGAGGTAGAATTATATGGTGAAAC
ChrXseq         TGGTAAGCTGGCTTGAGTAATTGGTCTGCTGTCTGAAGAGGTAGAATTATATGGTGAAAC
                ************* ***:********************************************

MTHFD1          AAAGGCCAAAGTTCTGCTGTCAGCACTAGAACGCCTGAAGCACCGGCCTGATGGGAAATA
ChrXseq         AAAGGCCAAAGTTCTGCTGTCAGCACTAGAACGCCTGAAGCACCAGCCTGTTGGGAAATA
                ******************************************.*****:**********

MTHFD1          CGTGGTGGTGACTGGAATAACTCCAACACCCCTGGGAGAAGGGAAAAGCACAACTACAAT
ChrXseq         TCAGGTGGTGACTGGAATAACTCCAACACCCCTGGGAGAAGGGAAAAGCACGACTACAAT
                  :************************************************.********

MTHFD1          CGGGCTAGTGCAAGCCCTTGGTGCCCATCTCTACCAGAATGTCTTTGCGTGTGTGCGACA
```

```
ChrXseq         CGGGCTGGTGCAAGCCCTTCTTGCCCATCTTTACCAGAACTTTTTTATGTGTGTGCGACA
                ******.************  ********* ********  * ***. ************

MTHFD1          GCCTTCTCAGGGCCCCACCTTTGGAATAAAAGGTGGCGCTGCAGGAGGCGGCTACTCCCA
ChrXseq         GCCTTCTCAGGGCCCCACCTTTGGAATAAAAGGTGGTGTTGCAGGAGGCGGCTACTCCCA
                ************************************ * **********************

MTHFD1          GGTCATTCCTATGGAAGAGTTTAATCTCCACCTCACAGGTGACATCCATGCCATCACTGC
ChrXseq         GGTCATTCCTATGGAAGAGTTTAATCTTCACCTCACAGGTGACATCCATGCCATCACTGT
                *************************** ******************************

MTHFD1          AGCTAATAACCTCGTTGCTGCGGCCATTGATGCTCGGATATTTCATGAACTGACCCAGAC
ChrXseq         ATCTAATAACCTTGTTGCTGCGGCTATTGATGCTCAGATATTTCATGAACTGACCCAGAC
                * ********** *********** **********.************************

MTHFD1          AGACAAGGCTCTCTTTAATCGTTTGGTGCCATCAGTAAATGGAGTGAGAAGGTTCTCTGA
ChrXseq         AGACAAGGCTACCTTTACTCATTTGGTGCCATCAGTAAATGGAGTGAGAAAGTTCTCTGA
                **********. *****.**.*****************************.**********

MTHFD1          CATCCAAATCCGAAGGTTAAAGAGACTAGGCATTGAAAAGACTGACCCTACCACACTGAC
ChrXseq         TATCCACATACAAAGGTTAAACAGACTAGGCATTGACAAGACTGACTCTACCACACTGAC
                 *****.**.*.********* **************.********* **************

MTHFD1          AGATGAAGAGATAAACAGATTTGCAAGATTGGACATTGATCCAGAAACCATAACTTGGCA
ChrXseq         AGATGAAGAGATAAACAGATTTGCAAGATTGGACATTGATCCAGAAACCATAGCTTGGCA
                ******************************************************.********

MTHFD1          AAGAGTGTTGGATACCAATGATAGATTCCTGAGGAAGATCACGATTGGACAGGCTCCAAC
ChrXseq         AAGAGTGTTGAATACCAATGATAGATTCCTGAGGAAGATCACGATTGGACAGGCTCCAAC
                **********.*************************************************

MTHFD1          GGAGAAGGGTCACACACGGACGGCCCAGTTTGATATCTCTGTGGCCAGTGAAATTATGGC
ChrXseq         GAAGAAAAGTCACACACAGACGGTCCACTTTGATATCTCTGTGGCCAGTGAAATTATGGC
                *.****..*********.***** *** ********************************

MTHFD1          TGTCCTGGCTCTCACCACTTCTCTAGAAGACATGAGAGAGAGACTGGGCAAAATGGTGGT
ChrXseq         TGTCCTGGCTCTCACCACTTCTCTAGAAGACAGGAGAGAGAGACTGGGCAAAATGGTGGT
                ******************************** ***************************

MTHFD1          GGCATCCAGTAAGAAAGGAGAGCCCGTCAGTGCCGAAGATCTGGGGGTGAGTGGTGCACT
ChrXseq         GGCATCCAGTAAGAAAGGAGACCCCGTCAGTGCTGAAGATCTGGGGGTGAGTGGTGCATT
                ********************** *********** ************************ *

MTHFD1          GACAGTGCTTATGAAGGACGCAATCAAGCCCAATCTCATGCAGACACTGGAGGGCACTCC
ChrXseq         GATAGTGCTTATGAAGGACACAATCAAGCCCAATCTCATGCAGAGATTGGAGGGCACTCC
                ** ****************.************************ * **************

MTHFD1          AGTGTTTGTCCATGCTGGCCCGTTTGCCAACATCGCACATGGCAATTCCTCCATCATTGC
ChrXseq         AGTGTTTGTCCATGCAGGCCCATTTGCCATCATCGCACACGGCAATTCCTCTATCATTGC
                ***************:******.*******:********* *********** ********

MTHFD1          AGACCGGATCGCACTCAAGCTTGTTGGCCCAGAAGGGTTTGTAGTGACGGAAGCAGGATT
ChrXseq         AGACCGGATCACACTCTAGCTTGTTGGCCCAGAAGGGTTTGTAGTGACGGAGGCGGGATT
                **********.*****:************************************.**.*****

MTHFD1          TGGAGCAGACATTGGAATGGAAAAGTTTTTTAACATCAAATGCCGGTATTCCGGCCTCTG
ChrXseq         TGGAGCAGACATTGGAATAGAAAAGTTTCTTAACATCAAATCCCAGTATTCCGGTCTCCA
                ******************.********* ************ **.********* *** .

MTHFD1          CCCCCACGTGGTGGTGCTTGTTGCCACTGTCAGGGCTCTCAAGATGCACGGGGGCGGCCC
ChrXseq         CCCTCACGTGGTGGTGCTTGTTGCCACTGTCAGCGCTCTAAAGATGCACAGGGGTGGCCC
                *** ***************************** *****.*********.**** *****
```

```
MTHFD1          CACGGTCACTGCTGGACTGCCTCTTCCCAAGGCTTACATACAGGAGAACCTGGAGCTGGT
ChrXseq         TATGGTCACTGCTGGAATGCCTCTTCCCAAGGCTTACATAGAGGAGAACCTGGAGCTGGT
                 * ************.********************** ******************

MTHFD1          TGAAAAAGGCTTCAGTAACTTGAAGAAACAAATTGAAAATGCCAGAATGTTTGGAATTCC
ChrXseq         TGAAAAAGGCTTCAGTAACTTGAAGAAACAAACTGAAAATGCCAGAATGTTTGGAATTCC
                ******************************** ***************************

MTHFD1          AGTAGTAGTGGCCGTGAATGCATTCAAGACGGATACAGAGTCTGAGCTGGACCTCATCAG
ChrXseq         AGTAGTAGTGGCCGTGAATGCATTCAAGACAGATACAGAGACTGAGCTGGACCTCATCGG
                ******************************.*********:*****************.*

MTHFD1          CCGCCTTTCCAGAGAACATGGGGCTTTTGATGCCGTGAAGTGCACTCACTGGGCAGAAGG
ChrXseq         CTGCCTTTCCAGAGAACAGAGGGCTTTTGATGCCGTGAAGTGCACTCACTGG-CAGAAGG
                * ***************** .********************************* *******

MTHFD1          GGGCAAGGGTGCCTTAGCCCTGGCTCAGGCCGTCCAGAGAGCAGCACAAGCACCCAGCAG
ChrXseq         GGGCAAGGGTACCTTAGCCCTGGCTCAGGCTGTCCAGAGAGCAGGACAAGCACCCAGCAG
                **********.****************** ************* ***************

MTHFD1          CTTCCAGCTCCTTTATGACCTCAAGCTCCCAGTTGAGGATAAAATCAGGATCATTGCACA
ChrXseq         CTTCCAACTCCTTTATGACCTCAAGCTCCCAGTTGAGGATAAAATCAGGATCATTGCACA
                ******.****************************************************

MTHFD1          GAAGATCTATGGAGCAGATGACATTGAATTACTTCCCGAAGCTCAACACAAAGCTGAAGT
ChrXseq         GAAGATCTATGGAGCAGACAACATTGAATTACTCCTCAAAGCTCAACACAAAGCTGAAGT
                ****************** .************* * *.**********************

MTHFD1          CTACACGAAGCAGGGCTTTGGGAATCTCCCCATCTGCATGGCTAAAACACACTTGTCTTT
ChrXseq         CTACACAAAGCAGGGCTTTGGGAATCTCCCCATCTGCATGGCCAAAACACACTTGTCCTT
                ******.************************************ ************** **

MTHFD1          GTCTCACAACCCAGAGCAAAAAGGTGTCCCTACAGGCTTCATTCTGCCCATTCGCGACAT
ChrXseq         GTCTCACAACCCAGAGCGAAAAGGTGTCCCTACAGGATTCATTCTGCCCATTCGCAACAT
                *****************.******************.******************.****

MTHFD1          CCGCGCCAGCGTTGGGGCTGGTTTTCTGTACCCCTTAGTAGGAACGATGAGCACAATGCC
ChrXseq         CCATGCCAGCATCGGGGCTGGTTTTCTGTACCCTTTAGTAGGAACGATGAGCACAAT---
                **. ******.* ******************** ***********************

MTHFD1          TGGACTCCCCACCCGGCCCTGTTTTTATGATATTGATTTGGACCCTGAAACAGAACAGGT
ChrXseq         -GGACTCCCCACCCAGCCCTGTTTTTATGATACTGATTTGGACCCTGAAACAGAACAGGT
                 *************.****************** ****************************

MTHFD1          GAATGGATTATTCTAAACAGATCACCATCCATCTTCAAGAAGCTACTTTGAAAGTCTGGC
ChrXseq         GAATGGATCATTCTAAACAGATCATCATCCATCTTCAAGAAGCTACTTTGAAAATCTGGC
                ******** *************** ****************************.******

MTHFD1          CAGTGTCTATTCAGGCCCACTGGGAGTTAGGAAGTATAAGTAAGCCAAGAGAAGTCAGCC
ChrXseq         CAGTGTCTATTCAGGCCCACTGGGAGTTAGGAAGTATAAGGCAGCCAAGA---GTCAGCC
                ************************************** .********   ********

MTHFD1          CCTGCCCAGAAGATCTGAAACTAATAGTAGGAGTTTCCCCAGAAGTCATTTTCAGCCTTA
ChrXseq         CCTGCCCAGAAGATCTGAAACTAATAGTAGGAGTTTTCCCGGATGTCATTTTCAGCCTTA
                ********************************** ***.**:*****************

MTHFD1          ATTCTCATCATGTATAAATTAACATAAATCATGCATGTCTGTTTACTTTAGTGACGTTCC
ChrXseq         ATTCTCATCGTGTATAAATTAACATAAATCATGC--------------------------
                *********.************************

MTHFD1          ACAGAATAAAAGGAAACAAGTTTGCCATCAAAAAAAAAAAAAAAAAA
ChrXseq         -------------------------------------------------
```

Appendix BB

CLUSTALW alignment (Chapter 2, Section 2.2.14.3) of the following sequences: Chr9seq, Chr2seq, part of MTHFD1L.

Chr9seq and Chr2seq were aligned with the part of MTHFD1L with similar sequence that correspond to exon 25, exon 26 and a part of flanking intron 26 (see Chapter 5, Figure 5.2). Position of the primers are highlight as follows: NCBI 5For (gray), NCBI All For (light blue), NCBI 2Rev (green), NCBI 2For (yellow). NCBI AllFor/Rev primers were designed to amplify both Chr2seq and Chr9seq but not MTHFD1L mRNA as NCBI 2Rev sits on MTHFD1L intronic region (underlined). NCBI All For/2Rev primers specifically amplify Chr2seq while NCBI 5For/2Rev primers specifically amplify Chr9seq (Chapter 5, Figure 5.3a).

```
CLUSTAL 2.1 Multiple Sequence Alignments

Sequence 1: Chr9seq     2574 bp
Sequence 2: Chr2seq     2028 bp
Sequence 3: MTHFD1L     1937 bp
```

```
Chr9seq         AGTGCTCCTGCTGACCTTCCCTGTCTGCAGCCCAGCTTTGGACTGTGGAGATGCCCTCAG
Chr2seq         ------------------------------------------------------------
MTHFD1L         ------------------------------------------------------------

Chr9seq         CCTCTACTCAGTGGCTCCAGCCTCATATTTGCGTGCTATACCTTTAACCTCAAGTTCTCT
Chr2seq         ------------------------------------------------------------
MTHFD1L         ------------------------------------------------------------

Chr9seq         TGCTGCTAACTTAAGCTGTGACTGTCCAGGACCTGCACACACACATACATGCACGCATGC
Chr2seq         ------------------------------------------------------------
MTHFD1L         ------------------------------------------------------------

Chr9seq         ACACACACACACACACACACACACACACACACACAGAAGGTGAACACACAGGATGAGGCC
Chr2seq         ------------------------------------------------------------
MTHFD1L         ------------------------------------------------------------

Chr9seq         CTGAAAACTCATTTCACCCGTGTACCTCACCCAAAAATTCCCCTTTGTACTCCTATCTTT
Chr2seq         ------------------------------------------------------------
MTHFD1L         ------------------------------------------------------------

Chr9seq         ATCTCTAACTGACAGCATGAAGGTAGCCCGAAGTCCCCTCACCCTGTCTGTCCTCCTGAT
Chr2seq         ------------------------------------------------------------
MTHFD1L         ------------------------------------------------------------

Chr9seq         TCTCTAGCCCAGTTCCCTCTAGCTCCAGCACAGACTCTTATCTCCAGCCCTGACTGCTGA
Chr2seq         ------------------------------------------------------------
MTHFD1L         ------------------------------------------------------------
```


Chr9seq AACCCTCCAGCCCCAGCTCCAGCCCCTCAGGAGGGGACTGCGTGGGAGGAATGACTCCAG
Chr2seq ----------
MTHFD1L ----------
Chr9seq CAAGCCGGCAGCTGACCCGTCTCTGGCCTGCATGCCTCCCCCTGTGGTCCACTGCCGGTG
Chr2seq ----------
MTHFD1L ----------
Chr9seq GTCTTCAGCTCTCTCCTCGTCCCACCCCC-ATGGTCTGCTGACAGCCCAGCCCAGTGCAT
Chr2seq ----GCTGGTGCCACTTGAACCAAGTTCAGATGCACTGCCCAAAATAACATCCCTCGT--
MTHFD1L ----------
Chr9seq CCCTGCCATCTCTGCTCTACCGCTGGTCCTCCCAA-CAACCAACTTACAGCTTCCAATTG
Chr2seq -CCGGCCAGCAGTGCCGTGGAGACCGAGTTCTGAAGCAGGCCTTTGTGAGCTTCCAATTG
MTHFD1L ----------GTTCCAATTG
****
Chr9seq TGGACAAGATAAGGACCATTGCCCAGGCTGTCTATGGAGCCAAAGATATCGAACTCTGTC
Chr2seq TGGACAAGATAAGGACCATTGCCCAGGCTGTCTGTGGAGCCAAAGATATCGAACTCTGTC
MTHFD1L TGGACAAGATAAGGACCATTGCTCAGGCTGTCTATGGAGCCAAAGATATTGAACTCTCTC
********** ********** *************** ******* **
Chr9seq CTGAGGCACAAGTCAAAATAGATCGTTACACTCAGCAGGGTTTTGGAAATTTGCCCATCT
Chr2seq CTGAGGCACAAGTCAAAATAGATCGTTACACTCAGCAGGGTTTTGGAAATTTGCCCATCT
MTHFD1L CTGAGGCACAAGCCAAAATAGATCGTTACACTCAACAGGGTTTTGGAAATTTGCCCATCT
************ ********** ********** ********** **********
Chr9seq GCATGGCAAAGACCGATCTTTCTCTGTCTCACCAACCTGACAAAAAAGGTGTGCCAAGGG
Chr2seq GCATGGCAAAGACCGACCTTTCTCTGTCTCACCAACCTGACAAAAAAGGTGTGCCAAGGG
MTHFD1L GCATGGCAAAGACCCACCTTTCTCTATCTCACCAACCTGACAAAAAAGGTGTGCCAAGGG
************** * ******** **********
Chr9seq ACTTCATCTTACCTATCAGTGATGTCCGGGCCAGCATAGGTGCTGGGTTCATTTACCCTT
Chr2seq ACTTCATCTTACCTATCAGTGATGTCCGGGTCAGCACAGACGCTGGGTTCATTTACCCTT
MTHFD1L ACTTCATCTTACCTATCAGTGACGTCCGGGCCAGCATAGGCGCTGGGTTCATTTACCCTT
********** ******* ***** ** **********
Chr9seq TGGTCGGAACGGTGCACCACTGACATCATGGCTCGGCTGGACTCTGGAAAGCAGGAAATT
Chr2seq TGGTCGGAACGGTGCACCACTGACATCATTGCTCGGCTGGACTCTGGAAAGCAGGAAATT
MTHFD1L TGGTCGGAACGGTGCACCACTGACATCATCGCTTGGCTGGACTCTGGAAAGCAGGAAACT
****************************** *** ********** *
Chr9seq GCTGGAGCATGTTTGGG-CATCAATGATGTCACCCGCTACGAATGCCGGGTGGGGTTACT
Chr2seq GCTGGAGCATGTTTGGGGCATCAGTGATGTCACCCGCTACGAATGCCGGGTGGGGTTACT
MTHFD1L GCTGGAGCGTGTTTGGGGCATCAGTGATGTCACCCGCTACAAATGCTGGGTGGGGTTACT
******** ******** ***** ********** ***** **********
Chr9seq GGACTGAGGGCCATGAAAAAAGAAAGCTGCGGCCCTGCCCCGTGGACTCACTGCCATTCT
Chr2seq GGACTGAGGGCCATGAAAAAAGAAAGCTGCGGCCCTGCCCCGTGGACTCACTGCCATTCT
MTHFD1L GGACTGAGGGCCATGAAAAAAGAAAGCTGCGGCCCCGGCCCGTGGACTCACTGCCATTCT
********** * **********
Chr9seq GTTCTTTCTCACGTTTCAGTTCAGCTCACTTTGTTTTCCGTCCCCCTCCCTCTACTGTAA
Chr2seq GTTCTTTCTCACGTTTCAGTTCAGCTCACTTTGTTTTCCGTCCCCCTCCCCCTACTGTAA
MTHFD1L GCTCTTTCTCACGTTTCAGTTCAGCTCGGTTTGTTTTCCGTCCCCCTCCCTCTACTGTAA
* ********** ********** *********
Chr9seq AAGTATGTACTCATTGTTTCATTTCCTGCTGGATCTGATTCAGGTCAGTGATGAAGTTCT

```
Chr2seq         AATTATGTACTCAGTGTTTCATTTCCTGCTGGATCTGATTCAGGTCAGTGATGAAGTTCT
MTHFD1L         AAGTATGTACTCAGTGTTTCGTTTCCTGCTGGATCTGATTCAGGTCAGTGATGAAGTTAT
                ** ********** ****** ************************************ *

Chr9seq         TTGCGAGCTCAGTGGGGAGGCGTTTCCTCCCTTTCATACCGGCCCTGCTTAGGGATGCAT
Chr2seq         TTGTGAGCTCAGTGGGGAGGCGTTTCCTCCCCTTCATACCGGCCCTGCTTAGGGATGCAT
MTHFD1L         TTGTGAACTCAGTGGGGAGGTGTTTCCTCCCCTTTGTACCGGCCCTACTTAGGGATGCAC
                *** ** ************* ********** **  ********** ************

Chr9seq         GGTTGGTTATATACTTGCTTCACTCGGTTAGGCATGGAGGAGAATTCATTCAGACCTCAT
Chr2seq         GGTTGGTTATATACTTGCTTCACTCGGTTAGGCATGGAGGAGAATTCATTCAGACCTCAT
MTHFD1L         GGTTGGTTATACGCTTGCTTCACTCGGTTAGGCATGGAGGAGAATTCATTCAGCCCTCAT
                **********  ************************************** ******

Chr9seq         AGGTTTAAATCAAATGCATGACCCTTCACATTTTCCAGAGATTTA--GACAGTTACAGTG
Chr2seq         AGGTTTAAATCAAATGCATGACCCTTCAGATTTTCCAGAGATTTA--GACAGTTACAGTG
MTHFD1L         AGGTTTAAATCAAATGCATGACCCTTCACATTTTCCAGAGATTTACGGCCAGTTACAATG
                **************************** ***************  * ******** **

Chr9seq         AGACAATTAAACATTCACCCCCC----AAGCTGCACTTGGAGATGTGTAAGCAGTAATGT
Chr2seq         AGACAATTAAACATTCACCCCCC----AAGCTGCACTTGGAGATGTGTAAGCAGTAATGT
MTHFD1L         AGACAATTAAACATTCACCCCCCCCGCCAAGCTGCACTTGGAGATGTGTAAGCAGTAATGT
                ***********************    **********************************

Chr9seq         AGTCATGCGCTCCCCATGATATGAGAGTGTGCAGAATGCCACTCCCCACAGCAGGTCCAC
Chr2seq         AGTCATGCGCTCCCCATGATATGAGAGTGTGCAGAATGCCACTCCCCACAGCAGGTCCAC
MTHFD1L         AGTCATGCGCTCCCCATGATGTGAGAGCGTGCAGAACGCCACTCCCCCCAGCAGGTCCGC
                ******************** ****** ******** ********** ********** *

Chr9seq         CAGGACCTCCTCAGGGAGCATCTTGCCTACAAAATCACCCATCCCTTTTCCCCCACTTTC
Chr2seq         CAGGACCTCCTCAGGGAGCATCTTGCCTACAAAATCACCCATCCCTTTTTCCCCACTCTC
MTHFD1L         CAGGACCTCCGCAGGGAGCATCTTGCCTGCAAAATCACACATCCCTTTTTCCCC-CTCTC
                ********** ***************** ********* ********** **** ** **

Chr9seq         TTCACTTTCCCTACG---TTTCCACTCCCACACCCACAAATAAACAAAGAAACTTTGAGC
Chr2seq         TTCACTTTCCCTACG---TTTCCACTCCCACACCCACAAATAAACAAAGAAGCTTTGAGC
MTHFD1L         TCCACTTTCCCTACTACTTTTCCACTCCCACACCCACAAATAAACAAAGAAACCTTGAGC
                * ************    ********************************* * ******

Chr9seq         TTATTTTACAAAAGTCTTTGAAATGGCCCTCCTCTCTCCGATGCTCCTGGGCCCTGGTTT
Chr2seq         TTATTTTACAAAAGTGTTTGAAATGGCCCTCCTCTCTCTGATGCTCCTGGGCCCTGGTTT
MTHFD1L         TTATTTTACAAAAGTCTTTGAAATGGCCCTCGTCTCTCCAATGCTCCTGGGTCCTGGTTT
                *************** *************** ******  ************ ********

Chr9seq         GGCCCCATCTGTGCAGTCTCTCTGATGAGTGTGAAATAAGCCAGGTTTGGCTCATGCTCT
Chr2seq         GGCCCCATCTGTGCAGTCTCTCTGATGAGTGTGAAATAAGCCAGGTTTGGCTCATGCTCT
MTHFD1L         GGCCCCATCTGTGCAATCTCTCTGATGAGTGTGAAATAAGCCAGGTTTGGCTCATGCTCT
                *************** ******************************************

Chr9seq         CCCTGTGAGCTTGCTCTCCCCGTGAGCCTGCCCTCTCCATCTGGCCTCACTTTGTTTGTG
Chr2seq         CCCTGTGAGCTTGCTCTCCCCGTGAGACTGCCCTCTCCATCTGGCCTCACTTTGTTTGTG
MTHFD1L         GCCGGTGAGCTTGCCCTCCCCATGAGCCTGTCCTCTCCATCTGGCCTCACTTTGTTTGTG
                 ** ********** ****** **** *** ****************************

Chr9seq         TCCCTGGGTCCTTTGTCTCTGATCCTGGGATCCTGGTGGTTTCCTCCTCCCCGCCCAGCT
Chr2seq         TCCCTGTGTCCTTTATCTCTGATCCTGGGATCCTGGTGGTTTCCTCCTCCCCGCCCAGCT
MTHFD1L         TCCCTGTGTCCTTTGTCTCCCATCCTGGGATCCTGGTGGTTTCCTCCTCCCCTCCCAGCA
                ****** ******* ****  ****************************** ******

Chr9seq         TGGGTCCTTTTCCCCAGGGTTCTTCCCTATCTCTACATCAGAATTTCCTGCTTTCTCCCA
Chr2seq         TGGGTACTTTTCCCCAGGGTTCTTCCCTATCTCTACATCAGAATTTCCTGCTTTCTCCCA
MTHFD1L         TGAGTCCTTTTCCCCAGGGTTCTTCCCTATCTCTACATCAGAATTTCCTGCTTTCTCCCA
                ** ** ******************************************************
```

```
Chr9seq         AATATGCATTTCCCTGGCCCAGGCATCCATTGCTTCCTCTCATCGTGAGGTCCCTGCAGC
Chr2seq         GATATGCATTTCCCTGGCCCAGGCATCCATTACTTCCTCTCATCGTGAGGTCCCTGCAGC
MTHFD1L         AATATGCATTTCCCTGGCCCAGGCATCCACTGCTTCCTCTCATCGTGAGGTCCCTGCCGC
                 *************************** * ************************ **

Chr9seq         AGACTGCCAATGGTGTTGTGGCCATGCCTTCCTCCACAACCCAGGGAAAGCTACATGTGT
Chr2seq         AGGCTGCCGATGGTGTTGTGGCCATGCCTTCCTCCACAACCCAAGGAAAGCTACATGTGT
MTHFD1L         AGGCTGCTGATGGTATTGTGGCCGTGCCTTCCTCCACAACCCAGGGAAAGCTACGTGTGT
                ** ****  ***** ******** ******************** ********** *****

Chr9seq         GTCTGTCCCATAGGGAAGGAGTCATCCCTGTCTCTTCAGTGTGGCATGTTCAGGAGGAAG
Chr2seq         GTCTGTCCCATAGGGAAGGAGTCATCCCTGTCTCTTCAGTGTGGCATGTTCAGGAGGAAG
MTHFD1L         GTCTGTCCCATAGGGAAGGAGTCATCCCTGTCTCCTCAGTGTGGTGTGTTCAGGAGGAAG
                ********************************** *********  ***************

Chr9seq         GAAAGTAACCAGCGTCATTCTCAATTTCCGGAAACTGTTTATCATATTGACAAGAAGAAG
Chr2seq         GAAAGTAACCAGCGTCATTCTCAATTTCCGGAAACTGTTTATCATATTGACAAGAAGAAG
MTHFD1L         GAAAGTAACCAGCGTCGTTCTCAATTTCCAGAAACTGTTTATCATATTGACAAGAAGAAG
                *************** ************ ******************************

Chr9seq         AAAGCTTTGCGACTCATGAGAATGATGTTTTCCTCTCTGGTACATAAGGTTATGTAGGTC
Chr2seq         AAAGCTTTGTGACTCATGAGAATGATGTTTTCCTCTCTGGTACATAAGGTTATGTAGGTA
MTHFD1L         AAAGCTTTGTGACTCATGAGAATGATGTTTTCCTCTCTGGTACATAAGGTTATCTAGGTC
                ********* ****************************************** *****

Chr9seq         CAATCCATT-GTGTAGAAGATCTTTTCTCCCTTAATGGGATGTACACTTATTTTTAGCAC
Chr2seq         CAATCCATTTGTGTAGAGGATCTTTTCTCCCTTAATGTGATGTACAC--ATTTTTAGCAC
MTHFD1L         CAATCCATTTGTGTAGAGGATCTTCTCTCCCCTAGTGTGATATGCACTTATTTTTAGCAC
                ********* ******* ****** ****** ** ** *** * ***  ***********

Chr9seq         AAGTATAAAACTACTTTAAATGAAGTCAGCCTCAGCCAGGGAAATATGCTGAGTAATAAT
Chr2seq         AAGTATAAAACTACTTTAAATGAAATCAGCCTCAGCCAGGGAAATATGCTGAGTAATAAT
MTHFD1L         AAGTATAAAACTACTTTAAATGAAATCAGCCTCAGCCAGGGAAATATGCTGAGTAATAAT
                ********************** ***********************************

Chr9seq         GTTGCCAGGTACTATACCACTGAGTTGAGTTTGCAATTCACTGCTATTAATCCCTGCATG
Chr2seq         GTTGCCAGGTACTATACCACTGAGTTGAGTTTGCAATTCACTGCTATTAATCCCTGTGTG
MTHFD1L         GTTGCCAGGTACTATACCACTGAGTTGAGTTTGTAATTCACTGCTATTAATCCCTGTGTA
                ********************************** ********************  *

Chr9seq         TTAGTTCTGAATTTTTACTCTTTGCATACGTAGAAAAAATGGTGTTTCTCTTCAGAGTCA
Chr2seq         TTAGTTCTGAATTTTTACTCTTTGCATACGTAGAAAAA-TGGTGTTTCTCTTCAGAGTCA
MTHFD1L         TTAGTTCTGATTTTTTACTCTTTGCATACATAGAAAAA-TGGTGTTTCTCTTCAGAGTCA
                ********** ****************** ******** **********************

Chr9seq         AGGAGGGAAAAAA-GAAAAGTTAAAAGACACTTATAACACTTTTGTGTCCACCCCTAAAA
Chr2seq         AGGAGGGAAAAAA-GAAAAGATAAAAGACGCTTATAACACCTTTGTGTCCACCCCTAAAA
MTHFD1L         AGGAGGGAAAAAAAGAAAAGATAAAAGATGCTTATAACACTTTTGTGTCCACCCCTAAAA
                ************* ****** *******  ********** ********************

Chr9seq         TCAGCATATTGATCTACTATTTTT-CTAGGTATTGATGGAATATTGACTCATATACTTTC
Chr2seq         TCAGCATATTGATCTATTACTTTT-CTAGGTATTGATGGAATATTGACTCATATACTTTC
MTHFD1L         TCAGCATATTGATCTATTATTTTTTCTAGGTATTGATGAGATGTTGACTCATATACTTTC
                *************** ** **** *************  ** ******************

Chr9seq         ATACAACTG
Chr2seq         ATAAAACTG
MTHFD1L         ATACAACTG
                *** *****
```

References

- Anderson DD, Woeller CF, Chiang EP, Shane B and Stover PJ: Serine hydroxymethyltransferase anchors de novo thymidylate synthesis pathway to nuclear lamina for DNA synthesis. *J Biol Chem* 2012; **287**:7051-7062.
- Anderson DD, Quintero CM and Stover PJ: Identification of a de novo thymidylate biosynthesis pathway in mammalian mitochondria. *Proc Natl Acad Sci USA* 2011; **108**:15163-151638.
- Anderson DD and Stover PJ: SHMT1 and SHMT2 are functionally redundant in nuclear de novo thymidylate biosynthesis. PLoS One 2009; **4**:e5839.
- Anderson DD, Woeller CF and Stover PJ: Small ubiquitin-like modifier-1 (SUMO-1) modification of thymidylate synthase and dihydrofolate reductase. *Clin Chem Lab Med* 2007; **45**:1760-1763.
- Appling DR: Compartmentalization of folate-mediated one-carbon metabolism in eukaryotes. *FASEB J* 1991; **5**:2645-2651.
- Bartel DP: MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 2004; **116**:281-297.
- Beaudin AE, Perry CA, Stabler SP, Allen RH and Stover PJ: Maternal Mthfd1 disruption impairs fetal growth but does not cause neural tube defects in mice. *Am J Clin Nutr* 2012; **95**:882-891.
- Belanger C and MacKenzie RE: Isolation and characterization of cDNA clones encoding the murine NAD-dependent methylenetetrahydrofolate dehydrogenase-methenyltetrahydrofolate cyclohydrolase. *J Biol Chem* 1989; **264**:4837-4843.
- Birn H: The kidney in vitamin $_{B12}$ and folate homeostasis: characterization of receptors for tubular uptake of vitamins and carrier proteins. *Am J Physiol Renal Physiol* 2006; **291**:F22-F36.
- Blanton SH, Henry RR, Yuan Q, Mulliken JB, Stal S, Finnell RH and Hecht JT: Folate Pathway and Nonsyndromic Cleft Lip and Palate. *Birth Defects Res A Clin Mol Teratol* 2011; **91**:50-60.
- Blanton SH, Patel S, Hecht JT and Mulliken JB: MTHFR is not a risk factor in the development of isolated nonsyndromic cleft lip and palate. *Am J Med Genet* 2002; **110**:404-405.

- Blount BC, Mack MM, Wehr CM, MacGregor JT, Hiatt RA, Wang G, Wickramasinghe SN, Everson RB and Ames BN: Folate deficiency causes uracil misincorporation into human DNA and chromosome breakage: Implications for cancer and neuronal damage. *Proc Natl Acad Sci USA* 1997; **94**:3290-3295.
- Bolusani S, Young BA, Cole NA, Tibbetts AS, Momb J, Bryant JD, Solmonson A and Appling DR: Mammalian *MTHFD2L* Encodes a Mitochondrial methylenetetrahydrofolate dehydrogenase isozyme expressed in adult tissues. *J Biol Chem* 2011; **286**: 5166-5174.
- Boulos S, Meloni BP, Arthur PG, Bojarski C and Knuckey NW: Peroxiredoxin 2 overexpression protects cortical neuronal cultures from ischemic and oxidative injury but not glutamate excitotoxicity, whereas Cu/Zn superoxide dismutase 1 overexpression protects only against oxidative injury. *J Neurosci Res* 2007; **85**:3089-3097.
- Boyles AL, Wilcox AJ, Taylor JA, Fredriksen A, Ueland PM, Drevon CA, Vollset SE and Lie RT: Folate and one-carbon metabolism gene polymorphisms and their associations with oral facial clefts. *Am J Med Genet A* 2008; **146**:440-449.
- Bressler J, Folsom AR, Couper DJ, Volcik KA and Boerwinkle E: Genetic Variants Identified in a European Genome-Wide Association Study That Were Found to Predict Incident Coronary Heart Disease in the Atherosclerosis Risk in Communities Study. *Am J Epidemiol* 2010; **171**:14-23.
- Brody LC, Conley M, Cox C, Kirke PN, McKeever MP, Mills JL, Molloy AM, O'Leary VB, Parle-McDermott A, Scott JM and Swanson DA: A polymorphism, R653Q, in the trifunctional enzyme methylenetetrahydrofolate dehydrogenase/methenyltetrahydrofolate cyclohydrolase/formyltetrahydrofolate synthetase is a maternal genetic risk factor for neural tube defects: report of the Birth Defects Research Group. *Am J Hum Genet* 2002; **71**:1207-1215.
- Bufalino A, Ribeiro Paranaíba LM, Nascimento de Aquino S, Martelli-Júnior H, Oliveira Swerts MS and Coletta RD: Maternal polymorphisms in folic acid metabolic genes are associated with nonsyndromic cleft lip and/or palate in the Brazilian population. *Birth Defects Res A Clin Mol Teratol* 2010; **88**: 980-986.

- Buske FA, Mattick JS and Baile TL: Potential in vivo roles of nucleic acid triple-helices. *RNA Biol* 2011; **8**:427-439.
- Butali A, Little J, Chevrier C, Cordier S, Steegers-Theunissen R, Jugessur A, Oladugba B and Mossey PA: Folic Acid Supplementation Use and the MTHFR C677T Polymorphism in Orofacial Clefts Etiology: An Individual Participant Data Pooled-Analysis. *Birth Defects Res A Clin Mol Teratol* 2013; doi:10.1002/ bdra.23133.
- Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest AR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impiombato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, Chalk AM, Chiu KP, Choudhary V, Christoffels A, Clutterbuck DR, Crowe ML, Dalla E, Dalrymple BP, de Bono B, Della Gatta G, di Bernardo D, Down T, Engstrom P, Fagiolini M, Faulkner G, Fletcher CF, Fukushima T, Furuno M, Futaki S, Gariboldi M, Georgii-Hemming P, Gingeras TR, Gojobori T, Green RE, Gustincich S, Harbers M, Hayashi Y, Hensch TK, Hirokawa N, Hill D, Huminiecki L, Iacono M, Ikeo K, Iwama A, Ishikawa T, Jakt M, Kanapin A, Katoh M, Kawasawa Y, Kelso J, Kitamura H, Kitano H, Kollias G, Krishnan SP, Kruger A, Kummerfeld SK, Kurochkin IV, Lareau LF, Lazarevic D, Lipovich L, Liu J, Liuni S, McWilliam S, Madan Babu M, Madera M, Marchionni L, Matsuda H, Matsuzawa S, Miki H, Mignone F, Miyake S, Morris K, Mottagui-Tabar S, Mulder N, Nakano N, Nakauchi H, Ng P, Nilsson R, Nishiguchi S, Nishikawa S, Nori F, Ohara O, Okazaki Y, Orlando V, Pang KC, Pavan WJ, Pavesi G, Pesole G, Petrovsky N, Piazza S, Reed J, Reid JF, Ring BZ, Ringwald M, Rost B, Ruan Y, Salzberg SL, Sandelin A, Schneider C, Schönbach C, Sekiguchi K, Semple CA, Seno S, Sessa L, Sheng Y, Shibata Y, Shimada H, Shimada K, Silva D, Sinclair B, Sperling S, Stupka E, Sugiura K, Sultana R, Takenaka Y, Taki K, Tammoja K, Tan SL, Tang S, Taylor MS, Tegner J, Teichmann SA, Ueda HR, van Nimwegen E, Verardo R, Wei CL, Yagi K, Yamanishi H, Zabarovsky E, Zhu S, Zimmer A, Hide W, Bult C, Grimmond SM, Teasdale RD, Liu ET, Brusic V, Quackenbush J, Wahlestedt C, Mattick JS, Hume DA, Kai C, Sasaki D, Tomaru Y, Fukuda S, Kanamori-Katayama M, Suzuki M, Aoki J, Arakawa T, Iida J, Imamura K, Itoh M, Kato T, Kawaji H, Kawagashira N, Kawashima T, Kojima M,

Kondo S, Konno H, Nakano K, Ninomiya N, Nishio T, Okada M, Plessy C, Shibata K, Shiraki T, Suzuki S, Tagami M, Waki K, Watahiki A, Okamura-Oho Y, Suzuki H, Kawai J and Hayashizaki Y; FANTOM Consortium; RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group): The transcriptional landscape of the mammalian genome. *Science* 2005**; 309**:1559-1563.

- Carroll N, Hughes L, McEntee G and Parle-McDermott A: Investigation of the molecular response to folate metabolism inhibition. *J Nutr Bioch* 2012; **23**:1531-1536.
- Carroll N, Pangilinan F, Molloy AM, Troendle J, Mills JL, Kirke PN, Brody LC, Scott JM and Parle-McDermott A: Analysis of the *MTHFD1* promoter and risk of neural tube defects. *Hum Genet* 2009; **125**:247-256.
- Chen K and Rajewsky N: Natural selection on human microRNA binding sites inferred from SNP data. *Nat Genet* 2006; **38**:1452-1456.
- Chen LL and Carmicheal G: Long noncoding RNAs in mammalian cells: what, where, and why? *Wiley Interdiscip Rev RNA* 2010; **1**:2-21.
- Chen YL, Eriksson S and Chang ZF: Regulation and Functional Contribution of Thymidine Kinase 1 in Repair of DNA Damage. *Biol. Chem* 2010; **285**: 27327-27335.
- Chevrier C, Perret C, Bahuau M, Zhu H, Nelva A, Herman C, Francannet C, Robert-Gnansia E, Finnell RH, Cordier S: Fetal and maternal MTHFR C677T genotype, maternal folate intake and the risk of nonsyndromic oral clefts. *Am J Med Genet A* 2007; **143**:248-257.
- Christensen KE, Deng L, Leung KY, Arning E, Bottiglieri T, Malysheva OV, Caudill MA, Krupenko NI, Greene ND, Jerome-Majewska L, Mackenzie RE and Rozen R: A novel mouse model for genetic variation in 10-formyltetrahydrofolate synthetase exhibits disturbed purine synthesis with impacts on pregnancy and embryonic development. *Hum Mol Genet* 2013; doi:10.1093/hmg/ddt223.
- Christensen KE, Rohlicek CV, Andelfinger GU, Michaud J, Bigras JL, Richter A, Mackenzie RE and Rozen R: The MTHFD1 p.Arg653Gln variant alters enzyme function and increases risk for congenital heart defects. *Hum Mutat* 2009; **30**:212-220.

- Christensen KE and Mackenzie RE: Mitochondrial methylenetetrahydrofolate dehydrogenase, methenyltetrahydrofolate cyclohydrolase, and formyltetrahydrofolate synthetases. *Vitam Horm* 2008; **79**:393-410.
- Christensen KE and MacKenzie RE: Mitochondrial one-carbon metabolism is adapted to the specific needs of yeast, plants and mammals. *Bioessays* 2006; **28**:595-605.
- Christensen KE, Patel H, Kuzmanov U, Mejia NR and MacKenzie RE: Disruption of the *Mthfd1* gene reveals a monofunctional 10-formyltetrahydrofolate synthetase in mammalian mitochondria. *J Biol Chem* 2005; **280**:7597-7602.
- Cluskey S and Ramsden DB: Mechanisms of neurodegeneration in amyotrophic lateral sclerosis. *Mol Pathol* 2001; **54**:386-392.
- Czeizel AE and Dudas I: Prevention of the first occurrence of neural-tube defects by periconceptional vitamin supplementation. *N Engl J Med* 1992; **327**:1832-1835.
- Daly LE, Kirke PN, Molloy A, Weir DG and Scott JM: Folate levels and neural tube defects. *J Am Med Assoc* 1995; **247**:1698-1702.
- Damaraju VL, Cass CE and Sawyer MB: Renal Conservation of folates: Role of Folate Transport Proteins. In Folic Acid and Folates , ed. G Litwack, pp. 185-198. San Diego 2008.
- De Marco P, Merello E, Calevo MG, Mascelli S, Raso A, Cama A and Capra V: Evaluation of a methylenetetrahydrofolate-dehydrogenase 1958G>A polymorphism for neural tube defect risk. *J Hum Genet* 2006; **51**:98-103.
- Debusk RM, Fogarty CP, Ordovas JM and Kornman KS: Nutritional genomics in practice: where do we begin? *J Amer Diet Assoc* 2005; **105**:589-598.
- Di Pietro E, Wang XL and MacKenzie RE: The expression of mitochondrial methylenetetrahydrofolate dehydrogenase-cyclohydrolase supports a role in rapid cell growth. *Biochim Biophys Acta* 2004; **1674**:78-84.
- Di Pietro E, Sirois J, Tremblay ML and MacKenzie RE: Mitochondrial NAD-dependent methylenetetrahydrofolate dehydrogenase-methenyltetrahydrofolate cyclohydrolase is essential for embryonic development. *Mol Cell Biol* 2002; **22**:4158-4166.

- Ditscheid B, Fünfstück R, Busch M, Schubert R, Gerth J and Jahreis G: Effect of L-methionine supplementation on plasma homocysteine and other free amino acids: a placebo controlled double-blind cross-over study. *Eur J Clin Nutr* 2005; **59**:768-775.
- Dixon M, Marazita M, Beaty T and Murray JC: Cleft lip and palate: Understanding genetic and environmental influences. *Nat Rev Genet* 2011; **12**:167-178.
- Du L, Schageman JJ, Subauste MC, Saber B, Hammond SM, Prudkin L, Wistuba II, Ji L, Roth JA, Minna JD and Pertsemlidis A: miR-93, miR-98, and miR-197 regulate expression of tumor suppressor gene FUS1. *Mol Cancer Res* 2009; **7**:1234-1243.
- Duan S, Huang RS, Zhang W, Mi S, Bleibel WK, Kistner EO, Cox NJ and Dolan ME: Expression and alternative splicing of folate pathway genes in HapMap lymphoblastoid cell lines. *Pharmacogenomics* 2009; **10**:549-563.
- Dziegielewski J and Beerman TA: Cellular responses to the DNA strand-scission enediyne C-1027 can be independent of ATM, ATR, and DNA-PK kinases. *J Biol Chem* 2002; **277**:20549-20554.
- FAO/WHO Food and Agriculture Organization of the United Nations (FAO) and World Health Organization (WHO). 2002; Human vitamin and mineral requirements: report of a joint FAO/WHO expert consultation. Rome, Italy: FAO. 290 p.
- Finkelstein JD: The metabolism of homocysteine: pathways and regulation. *Eur J Pediatric* 1998; **157**:40-44.
- Fitzpatrick A: Folate (Folic acid): implications for health and disease. *AGRO Food Industry hi-tech* 2003; **14**:45-52.
- Food and Drug Administration. Food standards: amendment of standards of identity for enriched grain products to require addition of folic acid. Final rule. 21 CFR Parts 136, 137, and 139. *Fed Regist* 1996; **61**:8781-8807.
- Forssén KM, Jägerstad MI, Wigertz K and Witthöft CM: Folates and Dairy Products: A Critical Update. *J Am Coll Nutr* 2000; **19**:100-110.
- Fox JT and Stover PJ: Folate-mediated one-carbon metabolism. In Folic Acid and Folates , ed. G Litwack, pp. 1-44. San Diego 2008.
- Francisconi S, Codenotti M, Ferrari Toninelli G, Uberti D and Memo M: Mitochondrial dysfunction and increased sensitivity to excitotoxicity in mice deficient in DNA mismatch repair. *J Neurochem* 2006; **98**:223-233.

- Freire WB, Hertrampf E and Cortes F: Effect of folic acid fortification in Chile: preliminary results. *Eur J Pediatr Surg* 2000; **10**:42-43.
- Frost P, Blom HJ, Milos R, Goyette P, Sheppard CA, Matthews RG, Boers GJ, den Heijer M, Kluijtmans LA, van den Heuvel LP and Rozen R: A candidate genetic risk factor for vascular disease: A common mutation in methylenetetrahydrofolate reductase. *Nat Genet* 1995; **10**:111-113.
- Furness DL, Fenech MF, Khong YT, Romero R and Dekker GA: One-carbon metabolism enzyme polymorphisms and uteroplacental insufficiency. *Am J Obstet Gynecol* 2008; **199**:276.e1-8.
- Garrow TA, Brenner AA, Whitehead VM, Chen XN, Duncan RG, Korenberg JR and Shane B: Cloning of human cDNAs encoding mitochondrial and cytosolic serine hydroxymethyltransferases and chromosomal localization. *J Biol Chem* 1993; **268**:11910-11916.
- Griffiths-Jones S, Grocock RJ, van Dongen, Bateman A and Enright AJ: miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* 2006; **34**:D140-144.
- Gu Y, Sekiguchi J, Gao Y, Frank K, Ferguson D, Hasty P, Chun J and Alt FW: Defective embryonic neurogenesis in Ku-deficient but not DNA-dependent protein kinase catalytic subunit-deficient mice. *Proc Natl Acad Sci USA* 2000; **97**:2668-2673.
- Haas U, Sczakiel G and Laufer SD: MicroRNA-mediated regulation of gene expression is affected by disease-associated SNPs within the 3′-UTR via altered RNA structure. *RNA Biol* 2012; **9**:924-937.
- Hanash S: Disease proteomics. *Nature* 2003; **422**:226-232.
- Hasegawa T, Mikoda N, Kitazawa M and LaFerla FM: Treatment of Alzheimer's disease with anti-homocysteic acid antibody in 3xTg-AD male mice. *PLoS One* 2010; **5**:e8593.
- Hattori F, Murayama N, Noshita T and Oikawa S: Mitochondrial peroxiredoxin-3 protects hippocampal neurons from excitotoxic injury in vivo. *J Neurochem* 2003; **86**:860-868.
- Health Canada. Regulations amending the Food and Drug Regulations (1066). *Can Gaz Part 1* 1997; **131**:3702-3737.

- Herrmann W and Knapp JP: Hyperhomocysteinemia: a new risk factor for degenerative diseases. *Clin Lab* 2002; **48**:471-481.
- Hu CH, Yeh MT, Tsao N, Chen CW, Gao QZ, Chang CY, Lee MH, Fang JM, Sheu SY, Lin CJ, Tseng MC, Chen YJ and Chang ZF: Tumor Cells Require Thymidylate Kinase to Prevent dUTP Incorporation during DNA Repair. *Cancer Cell* 2012; doi:10.1016/j.ccr.2012.04.038.
- Inui M, Martello G and Piccolo S: microRNA control of signal Transduction. *Nat Rev Mol Cell Biol* 2010; **11**:252-263.
- IOM Inst. of Medicine 2004. In: dietary reference intakes (DRIs): recommended intakes for individuals, vitamins. Food and Nutrition Board, Inst. Of Medicine. Washington, D.C.: Natl. Academy Press.
- Iyer R and Tomar SR: Folate: a functional food constituent. *J Food Sci* 2009; **74**:114-122.
- Jain M, Nilsson R, Sharma S, Kitami T, Souza AL, Kafri R, Kirschner MW, Clish CB and Mootha VK: Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation. *Science* 2012; **336**:1040-1044.
- Jia ZL, Shi B, Chen CH, Shi JY, Wu J and Xu X: Maternal malnutrition, environmental exposure during pregnancy and the risk of non-syndromic orofacial clefts. *Oral Dis* 2011; **17**:584–589.
- Jugessur A, Farlie PG and Kilpatrick N: The genetics of isolated orofacial clefts: from genotypes to subphenotypes. *Oral Diseases* 2009; **15**: 437-453.
- Jugessur A, Wilcox AJ, Lie RT, Murray JC, Taylor JA, Ulvik A, Drevon CA, Vindenes HA and Abyholm FE: Exploring the effects of methylenetetrahydrofolate reductase gene variants C677T and A1298C on the risk of orofacial clefts in 261 Norwegian case-parent triads. *Am J Epidemiol* 2003; **157**:1083-1091.
- Kapur BM, Vandenbroucke AC, Adamchik Y, Lehotay DC and Carlen PL: Formic acid, a novel metabolite of chronic ethanol abuse, causes neurotoxicity, which is prevented by folic acid. *Alcohol Clin Exp Res* 2007; **31**:2114-2120.
- Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, Suzuki H, Carninci P, Hayashizaki Y, Wells C, Frith M, Ravasi T, Pang KC, Hallinan J, Mattick J, Hume DA, Lipovich L, Batalov S, Engström

PG, Mizuno Y, Faghihi MA, Sandelin A, Chalk AM, Mottagui-Tabar S, Liang Z, Lenhard B and Wahlestedt C; RIKEN Genome Exploration Research Group; Genome Science Group (Genome Network Project Core Group); FANTOM Consortium: Antisense transcription in the mammalian transcriptome. *Science* 2005, **309**: 1564-1566.

- Kawahara Y, Megraw M, Kreider E, Iizasa H, Valente L, Hatzigeorgiou AG and Nishikura K: Frequency and fate of microRNA editing in human brain. *Nucl Acids Res* 2008; **36**:5270-5280.
- Kelly D, O'Dowd T and Reulbach U: Use of folic acid supplements and risk of cleft lip and palate in infants: a population-based cohort study. *Br J Gen Pract* 2012; **62**:466-472.
- Kertesz M, Iovino N, Unnerstall U, Gaul U and Segal E: The role of site accessibility in microRNA target recognition. *Nat Genet* 2007; **39**:1278-1284.
- Khew-Goodall Y and Goodall GJ: Myc-modulated miR-9 makes more metastases. *Nat Cell Biol* 2010; **12**: 209-211.
- Kim YI: Will mandatory folic acid fortification prevent or promote cancer? *Am J Clin Nutr* 2004; **80**:1123-1128.
- Kirke PN and Scott JM. 2005. Pre-conception nutrition and prevention of neural tube defects. In: Sadler MJ, Strain JJ, Caballero B, editors. Encyclopaedia of human nutrition. New York: Oxford Academic Press, p 15-27.
- Kornblihtt AR: Chromatin, transcript elongation and alternative splicing. *Nat Struct Mol Biol* 2006; **13**:5-7.
- Kosik SK: Circles reshape the RNA world The versatility of RNA seems limitless. *Nature* 2013; doi:10.1038/nature11956.
- Kozomara A and Griffiths-Jones S: miRBase: integrating microRNA annotation and deep-sequencing. *Nucl Acids Res* 2011; **39**:D152-D157.
- Krishnamurthy N, Hu Y, Siedlak S, Doughman YQ, Watanabe M and Montano MM: Induction of quinone reductase by tamoxifen or DPN protects against mammary tumorigenesis. *FASEB J* 2012; **26**:3993-4002.

- Laios A, O'Toole S, Flavin R, Martin C, Kelly L, Ring M, Finn SP, Barrett C, Loda M, Gleeson N, D'Arcy T, McGuinness E, Sheils O, Sheppard B and O' Leary J: Potential role of miR-9 and miR-223 in recurrent ovarian cancer. *Mol Cancer* 2008; **7**:35.
- Lalonde R and Strazielle C: Neurobehavioral characteristics of mice with modified intermediate filament genes. *Rev Neurosci* 2003; **14**:369-385.
- Lamarre SG, Molloy AM, Reinke SN, Sykes BD, Brosnan ME and Brosnan JT: Formate can differentiate between hyperhomocysteinemia due to impaired remethylation and impaired transsulfuration. *Am J Physiol Endocrinol Metab* 2012; **302**:E61-E67.
- Leaphart, AB, Spencer HT and Lovell CR: Site directed mutagenesis of a putative catalytic and formyl phosphate binding site and substrate inhibition of ^{N10}-formyltetrahydrofolate synthetase. *Archives of Biochemistry and Biophysics* 2002; **408**:137-143.
- Lee MK, Marszalek JR and Cleveland DW: A mutant neurofilament subunit causes massive, selective motor neuron death: implications for the pathogenesis of human motor neuron disease. *Neuron* 1994; **13**:975-988.
- Lee RC, Feinbaum RL and Ambros V: The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 1993; **75**:843-854.
- Lee Y, Kim M, Han, Yeom KH, Lee S, Baek SH and Kim VN: MicroRNA genes are transcribed by RNA polymerase II. *EMBO J* 2004; **23**:4051-4060.
- Lehmann U, Streichert T, Otto B, Albat C, Hasemeier B, Christgen H, Schipper E, Hille U, Kreipe HH and Länger F.: Identification of differentially expressed microRNAs in human male breast cancer. *BMC Cancer* 2010; **10**:109.
- Li L, Monckton EA and Godbout R: A role for DEAD box 1 at DNA double-strand breaks. *Mol Cell Biol* 2008; **28**:6413-6425.
- Li S, Armstrong CM, Bertin, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, Goldberg DS, Li N, Martinez M, Rual JF, Lamesch P, Xu L, Tewari M, Wong SL, Zhang LV, Berriz GF, Jacotot L, Vaglio P, Reboul J, Hirozane-Kishikawa T, Li Q, Gabel HW, Elewa A, Baumgartner B, Rose DJ, Yu H, Bosak S, Sequerra R, Fraser A, Mango SE, Saxton WM, Strome S, Van Den Heuvel S, Piano F, Vandenhaute J, Sardet C, Gerstein M, Doucette-Stamm L, Gunsalus KC, Harper JW,

Cusick ME, Roth FP, Hill DE and Vidal M: A map of the interactome network of the metazoan *C. elegans*. *Science* 2004; **303**:540-543.

- Little J, Cardy A and Munger RG: Tobacco smoking and oral clefts: a meta-analysis. *Bull World Health Organ* 2002; **82**:213-218.
- Lujambio A, Calin GA, Villanueva A, Ropero S, Sánchez-Céspedes M, Blanco D and Montuenga LM: A microRNA DNA methylation signature for human cancer metastasis. *Proc Natl Acad Sci USA* 2008; **105**:13556-13561.
- Lukiw WJ: Micro-RNA speciation in fetal, adult and Alzheimer's disease hippocampus. *Neuroreport* 2007; **18**:297-300.
- Lustbader JW, Cirilli M, Lin C, Xu HW, Takuma K, Wang N, Caspersen C, Chen X, Pollak S, Chaney M, Trinchese F, Liu S, Gunn-Moore F, Lue LF, Walker DG, Kuppusamy P, Zewier ZL, Arancio O, Stern D, Yan SS and Wu H: ABAD directly links Abeta to mitochondrial toxicity in Alzheimer's disease. Science 2004; **304**:448-452.
- Ma L, Young J, Prabhala H, Pan E, Mestdagh P, Muth D, Teruya-Feldstein J, Reinhardt F, Onder TT, Valastyan S, Westermann F, Speleman F, Vandesompele J and Weinberg RA: miR-9, a MYC/MYCN-activated microRNA, regulates E-cadherin and cancer metastasis. *Nat Cell Biol* 2010; **12**:247-256.
- Ma XY, Yu JT, Wu ZC, Liu QY, Wang HF, Wang W and Tan L: Replication of the MTHFD1L gene association with late-onset Alzheimer's disease in a Northern Han Chinese population. *J Alzheimers Dis* 2012; **29**:521-525.
- MacFarlane AJ, Perry CA, Girnary HH, Gao D, Allen RH, Stabler SP, Shane B and Stover PJ: Mthfd1 is an essential gene in mice and alters biomarkers of impaired one-carbon metabolism. *J Biol Chem* 2009; **284**:1533-1539.
- MacFarlane AJ, Perry CA, McEntee MF, Lin DM and Stover PJ: Mthfd1 is a modifier of chemically induced intestinal carcinogenesis. *Carcinogenesis* 2011; **32**:427-433.
- Maes OC, Chertkow HM, Wang E and Schipper HM: MicroRNA: Implications for Alzheimer Disease and other Human CNS Disorders. *Curr Genomics* 2009; **10**:154-168.
- Mangold E, Ludwig KU, Birnbaum S, Baluardo C, Ferrian M, Herms S, Reutter H, de Assis NA, Chawa TA, Mattheisen M, Steffens M, Barth S, Kluck N, Paul A, Becker J,

Lauster C, Schmidt G, Braumann B, Scheer M, Reich RH, Hemprich A, Pötzsch S, Blaumeiser B, Moebus S, Krawczak M, Schreiber S, Meitinger T, Wichmann HE, Steegers-Theunissen RP, Kramer FJ, Cichon S, Propping P, Wienker TF, Knapp M, Rubini M, Mossey PA, Hoffmann P and Nöthen MM: Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nat Genet* 2010; **42**:24-26.

- Mao DY, Watson JD, Yan PS, Barsyte-Lovejoy D, Khosravi F, Wong WW, Farnham PJ, Huang TH and Penn LZ.: Analysis of Myc bound loci identified by CpG island arrays shows that Max is essential for Myc-dependent repression. *Curr Biol* 2003; **13**:882-886.
- Marazita ML, Field LL, Cooper ME, Tobias R, Maher BS, Peanchitlertkajorn S and Liu YE: Nonsyndromic cleft lip with or without cleft palate in China: assessment of candidate regions. *Cleft Palate Craniofac J* 2002; **39**:149-156.
- Margulies EH and Birney E: Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes. *Nat Rev Genet* 2008; **9**:303-313.
- Martin B, Brenneman R, Becker KG, Cole RN and Maudsley S: iTRAQ analysis of complex proteome alterations in 3xTgAD Alzheimer's mice: understanding the interface between physiology and disease. *PLoS One* 2008; **3**:e2750.
- Martinelli M, Scapoli L, Palmieri A, Pezzetti F, Baciliero U, Padula E, Carinci P, Morselli PG and Carinci F: Study of four genes belonging to the folate pathway: transcobalamin 2 is involved in the onset of non-syndromic cleft lip with or without cleft palate. *Hum Mutat* 2006; **27**:294.
- Martinelli M, Scapoli L, Pezzetti F, Carinci F, Carinci P, Stabellini G, Bisceglia L, Gombos F and Tognon M: C677T variant form at the MTHFR gene and CL/P: a risk factor for mothers? *Am J Med Genet* 2001; **98**:357-360.
- Mathews DH, Sabina J, Zuker M and Turner DH: Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* 1999; **288**:911-940.
- McEntee G, Minguzzi S, O'Brien K, Ben Larbi N, Loscher C, O'Fágáin C and Parle-McDermott A: The former annotated human pseudogene dihydrofolate reductase-like 1

(DHFRL1) is expressed and functional. *Proc Natl Acad Sci USA* 2011; **108**:15157-15162.

- McPartlin J, Halligan A, Scott JM, Darling M and Weir DG: Accelerated folate breakdown in pregnancy. *Lancet* 1993; **341**:148-149.
- Mills JL, Molloy AM, Parle-McDermott A, Troendle JF, Brody LC, Conley MR, Cox C, Pangilinan F, Orr DJ, Earley M, McKiernan E, Lynn EC, Doyle A, Scott JM and Kirke PN: Folate-related gene polymorphisms as risk factors for cleft lip and cleft palate. *Birth Defects Res A Clin Mol Teratol* 2008; **82**:636–643.
- Minguzzi S, Molloy AM, Peadar K, Mills J, Scott JM, Troendle J, Pangilinan F and Brody L, Parle-McDermott A: Genotyping of a tri-allelic polymorphism by a novel melting curve assay in MTHFD1L: an association study of nonsyndromic Cleft in Ireland. *BMC Med Genet* 2012; doi:10.1186/1471.
- Mishra PJ, Humeniuk R, Mishra PJ, Longo-Sorbello GS, Banerjee D and Bertino JR: A miR-24 microRNA binding-site polymorphism in dihydrofolate reductase gene leads to methotrexate resistance. *Proc Natl Acad Sci USA* 2007; **104**:13513-13518.
- Mitchell HK, Snell EE and Williams RJ: The concentration of folic acid. *J Am Chem Soc* 1941; **63**:2284.
- Mitchell LE: Twin Studies in Oral Cleft Research; In Cleft Lip and Palate: From Origin to Treatment 2002; Wyszynski DF (ed). Oxford University Press.
- Momb J, Lewandowski JP, Bryant JD, Fitch R, Surman DR, Vokes SA and Appling DR: Deletion of Mthfd1l causes embryonic lethality and neural tube and craniofacial defects in mice. *Proc Natl Acad Sci USA* 2013; **110**:549-554.
- Moran RG: Roles of folylpoly-gamma-glutamate synthetase in therapeutics with tetrahydrofolate antimetabolites: an overview. *Semin Oncol* 1999; **26**:24-32.
- Morris MS: Homocysteine and Alzheimer's disease. Lancet Neurol 2003; **2**:425-428.
- Mossey P, Little J, Munger RG, Dixon MJ and Shaw WC:. Cleft lip and palate. *Lancet* 2009; **374**:1773-1785.
- Mostowska A, Hozyasz KK, Wojcicki P, Dziegelewska M and Jagodzinski PP: Associations of folate and choline metabolism gene polymorphisms with orofacial clefts. *J Med Genet* 2010; **47**: 809-815.

- Naj AC, Beecham GW, Martin ER, Gallins PJ, Powell EH, Konidari I, Whitehead PL, Cai G, Haroutunian V, Scott WK, Vance JM, Slifer MA, Gwirtsman HE, Gilbert JR, Haines JL, Buxbaum JD and Pericak-Vance MA: Dementia Revealed: Novel Chromosome 6 Locus for Late-Onset Alzheimer Disease Provides Genetic Evidence for Folate-Pathway Abnormalities. *PLoS Genet* 2010; **6**: e1001130.
- Nicholls P: Formate as an inhibitor of cytochrome c oxidase. *Biochem Biophys Res Commun* 1975; **67**:610-616.
- Nicoloso MS, Sun H, Spizzo R, Kim H, Wickramasinghe P, Shimizu M, Wojcik SE, Ferdin J, Kunej T, Xiao L, Manoukian S, Secreto G, Ravagnani F, Wang X, Radice P, Croce CM, Davuluri RV and Calin GA: Single-nucleotide polymorphisms inside microRNA target sites influence tumor susceptibility. *Cancer Res* 2010; **70**:2789-2798.
- Nikiforov MA, Chandriani S, O'Connell B, Petrenko O, Kotenko I, Beavis A, Sedivy JM and Cole MD: A functional screen for Myc-responsive genes reveals serine hydroxymethyltransferase, a major source of the one-carbon unit for cell metabolism. *Mol Cell Biol* 2002; **22**:5793-5800.
- Ott J, Kamatani Y and Lathrop M: Family-based designs for genome-wide association studies. *Nat Rev Genet* 2011; **12**:465-474.
- Parle-McDermott A, Pangilinan F, O'Brien KK, Mills JL, Magee AM, Troendle J, Sutton M, Scott JM, Kirke PN, Molloy AM and Brody LC: A common variant in MTHFD1L is associated with neural tube defects and mRNA splicing efficiency. *Hum Mutat* 2009; **30**:1650-1656.
- Parle-McDermott A, Mills JL, Molloy AM, Carroll N, Kirke PN, Cox C, Conley MR, Pangilinan FJ, Brody LC and Scott JM: The MTHFR 1298CC and 677TT genotypes have opposite associations with red cell folate levels. *Mol Genet Metab* 2006a; **88**:290-294.
- Parle-McDermott A, Kirke PN, Mills JL, Molloy AM, Cox C, O'Leary VB, Pangilinan F, Conley M, Cleary L, Brody LC and Scott JM: Confirmation of the R653Q polymorphism of the trifunctional C1-synthase enzyme as a maternal risk for neural tube defects in the Irish population. *Eur J Hum Genet* 2006b; **14**:768-772.
- Parle-McDermott A, Mills JL, Kirke PN, Cox C, Signore CC, Kirke S, Molloy AM, O'Leary VB, Pangilinan FJ, O'Herlihy C, Brody LC and Scott JM: MTHFD1 R653Q

polymorphism is a maternal genetic risk factor for severe abruptio placentae. *Am J Med Genet A* 2005a; **132**:365–368.

- Parle-McDermott A, Pangilinan F, Mills JL, Signore CC, Molloy AM, Cotter A, Conley M, Cox C, Kirke PN and Scott JM, Brody LC: A polymorphism in the MTHFD1 gene increases a mother's risk of having an unexplained second trimester pregnancy loss. *Mol Hum Reprod* 2005b; **11**:477-480.
- Patel H, Christensen KE, Mejia N and MacKenzie RE: Mammalian mitochondrial methylenetetrahydrofolate dehydrogenase-cyclohydrolase derived from a trifunctional methylenetetrahydrofolate dehydrogenase-cyclohydrolase-synthetase. *Arch Biochem Biophys* 2002; **403**:145-148.
- Paukert JL, Williams GR and Rabinowitz JC: Formyl-methenyl-methylenetetrahydrofolate synthetase (combined); correlation of enzymatic activities with limited proteolytic degradation of the protein from yeast. *Biochem Biophys Res Commun* 1977; **77**:147-154.
- Pereira PM, Marques JP, Soares AR, Carreto L and Santos MA: MicroRNA expression variability in human cervical tissues. *PLoS One* 2010; **5**:e11780.
- Pike ST, Rajendra R, Artzt K and Appling DR: Mitochondrial C1-tetrahydrofolate synthase (MTHFD1L) supports the flow of mitochondrial one-carbon units into the methyl cycle in embryos. *J Biol Chem* 2010; **285**:4612-4620.
- Prasannan P and Appling DR: Human mitochondrial C1-tetrahydrofolate synthase: Submitochondrial localization of the full-length enzyme and characterization of a short isoforms. *Arch Biochem Biophys* 2009; **481**:86-93.
- Prasannan P, Pike S, Peng K, Shane B and Appling DR: Human mitochondrial C1-tetrahydrofolate synthase: gene structure, tissue distribution of the mRNA, and immunolocalization in Chinese hamster ovary calls. *J Biol Chem* 2003; **278**:43178-43187.
- Prescott NJ, Winter RM and Malcolm S: Maternal MTHFR genotype contributes to the risk of non-syndromic cleft lip and palate *J Med Genet* 2002; **39**:368-369.
- Rampersaud E ME and Speer MC: Nonsyndromic neural tube defects: Genetic basis and genetic investigations. 2006 New York, Oxford Univ Press.

- Ray JG, Meier C, Vermeulen MJ, Wyatt PR and Cole DE: Association between folic acid food fortification and congenital orofacial clefts. *J Pediatr* 2003; **143**:805-807.
- Ren RJ, Wang LL, Fang R, Liu LH, Wang Y, Tang HD, Deng YL, Xu W, Wang G and Chen SD: The MTHFD1L gene rs11754661 marker is associated with susceptibility to Alzheimer's disease in the Chinese Han population. *J Neurol Sci* 2011; doi:10.1016/j.jns.2011.06.036.
- Roach JC, Glusman G, Smit AFA, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L and Galas DJ: Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. *Science* 2010; **5978**:636-639.
- Salmena L, Poliseno L, Tay Y, Kats L and Pandolfi PP: A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* 2011; **146**:353-358.
- Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Hengstenberg C, Mangino M, Mayer B, Dixon RJ, Meitinger T, Braund P, Wichmann HE, Barrett JH, König IR, Stevens SE, Szymczak S, Tregouet DA, Iles MM, Pahlke F, Pollard H, Lieb W, Cambien F, Fischer M, Ouwehand W, Blankenberg S, Balmforth AJ, Baessler A, Ball SG, Strom TM, Braenne I, Gieger C, Deloukas P, Tobin MD, Ziegler A, Thompson JR and Schunkert H; WTCCC and the Cardiogenics Consortium: Genomewide association analysis of coronary artery disease. *N Engl J Med* 2007; **357**:443-453.
- Samsonoff J, Reston M, McKee B, O'Connor B, Galivan J, Maley G and Maley F: Intracellular Location of Thymidylate Synthase and Its State of Phosphorylation. *J Biol Chem* 1997; **272**:13281-13285.
- Saunders MA, Liang H and Li WH: Human polymorphism at microRNAs and microRNA target sites. *Proc Natl Acad Sci USA* 2007; **104**:3300-3305.
- Sawicka K, Bushell M, Spriggs KA and Willis AE: Polypyrimidine-tract-binding protein: a multifunctional RNA-binding protein. *Biochem Soc Trans* 2008; **36**:641-647.
- Schirch V and Strong WB: Interaction of folylpolyglutamates with enzymes in one-carbon metabolism. *Arch Biochem Biophys* 1989; **189**:283-290.
- Schonrock N, Humphreys DT, Preiss T and Gotz J: Target gene repression mediated by miRNAs miR-181c and miR-9 both of which are down-regulated by amyloid-beta. *J Mol Neurosci* 2012; **46**:324-335.

- Selcuklu SD, Donoghue MT, Rehmet K, de Souza Gomes M, Fort A, Kovvuru P, Muniyappa MK, Kerin MJ, Enright AJ and Spillane C: MicroRNA-9 inhibition of cell proliferation and identification of novel miR-9 targets by transcriptome profiling in breast cancer cells. *J Biol Chem* 2012; **287**:29516-29528.
- Shane B: Folate chemistry and metabolism. In Folate in Health and Disease 1995, Bailey L.B. eds, pp.1-22. Marcel Dekker New York, NY.
- Shannon KW and Rabinowitz JC: Isolation and characterization of the Saccharomyces cerevisiae MIS1 gene encoding mitochondrial C1- tetrahydrofolate synthase. *J Biol Chem* 1988; **263**:7717-7725.
- Shannon KW and Rabinowitz JC: Purification and characterization of a mitochondrial isozyme of C1-tetrahydrofolate synthase from Saccharomyces cerevisiae. *J Biol Chem* 1986; **261**:12266-12271.
- Shaw GM, Rozen R, Finnell RH, Todoroff K and Lammer EJ: Infant C677T mutation in MTHFR, maternal periconceptional vitamin use, and cleft lip. *Am J Med Genet* 1998; **80**:196-198.
- Shi M, Wehby GL and Murray JC: Review on genetic variants and maternal smoking in the etiology of oral clefts and other birth defects. *Birth Defects Res C Embryo Today* 2008; **84**:16-29.
- Shotelersuk V, Ittiwut C, Siriwan P and Angspatt A: Maternal 677CT/1298AC genotype of the MTHFR gene as a risk factor for cleft lip. *J Med Genet* 2003; **40**:e64.
- Shumaker DK, Solimando L, Sengupta K, Shimi T, Adam SA, Grunwald A, Strelkov SV, Aebi U, Cardoso MC and Goldman RD: The highly conserved nuclear lamin Ig-fold binds to PCNA. Its role in DNA replication. *J Cell Biol* 2008; **181**:269-280.
- Song F, Zheng H, Liu B, Wei S, Dai H, Zhang L, Calin GA, Hao X, Wei Q, Zhang W and Chen K: An miR-502-binding site single-nucleotide polymorphism in the 3′-untranslated region of the SET8 gene is associated with early age of breast cancer onset. *Clin Cancer Res* 2009; **15**:6292-6300.
- Spielman S, McGinnis E and Ewens J: Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993; **52**:506-516.

- Stetler RA, Gao Y, Zhang L, Wei S, Dai H, Zhang L, Calin GA, Hao X, Wei Q, Zhang W and Chen K: Phosphorylation of HSP27 by protein kinase D is essential for mediating neuroprotection against ischemic neuronal injury. *Neurosci* 2012; **32**:2667-2682.
- Stover PJ and Weiss RS: Sensitizing Cancer Cells: Is It Really All about U? *Cancer Cell* 2012; **22**:3-4.
- Subbiah MTR: Nutrigenetics and nutraceuticals: the next wave riding on personalized medicine. *Transl Res* 2007; **149**:55-61.
- Sugiura T, Nagano Y, Inoue T and Hirotani K: A novel mitochondrial C1-tetrahydrofolate synthetase is upregulated in human colon adenocarcinoma. *Biochem Biophys Res Commun* 2004; **315**:204-211.
- Taft RJ, Pang KC, Mercer TR, Dinger M and Mattick JS: Non-coding RNAs: regulators of disease. *J Pathol* 2010; **220**:126-139.
- Tavtigian SV, Zabludoff SD and Wold BJ: Cloning of mid-G1 serum response genes and identification of a subset regulated by conditional myc expression. *Mol Biol Cell* 1994; **5**:375-388.
- Tchantchou F and Shea TB, Folate deprivation, the methionine cycle, and Alzheimer's disease. In Folic Acid and Folates, ed. G Litwack, pp. 84-94. San Diego 2008.
- Thomas P and Fenech M: Chapter 13 methylenetetrahydrofolate reductase, common polymorphisms, and relation to disease. *Vitam Horm* 2008; **79**:375-392.
- Tibbetts AS and Appling RA: Compartmentalization of Mammalian Folate Mediated One-Carbon Metabolism. *Annu Rev Nutr* 2010; **30**:57-81.
- Tomita M and Kami K: Cancer. Systems biology, metabolomics, and cancer metabolism. *Science* 2012; **336**:990-991.
- Torrents D, Suyama M, Zdobnov E and Bork P: A genome-wide survey of human pseudogenes. *Genome Res* 2003; **13**:2559-2567.
- Tsuji T, Ficarro SB, Jiang W: Essential role of phosphorylation of MCM2 by Cdc7/Dbf4 in the initiation of DNA replication in mammalian cells. *Mol Biol Cell* 2006; **17**:4459-4472.

- Tuteja N, Tuteja R, Ochem, Taneja P, Huang NW, Simoncsits A, Susic S, Rahman K, Marusic L and Chen J: Human DNA helicase II: a novel DNA unwinding enzyme identified as the Ku autoantigen. *EMBO J* 1994; **13**:4991-5001.
- USDA. U.S. Department of Agriculture, Agricultural Research Service. 2011. National Nutrient Database for Standard Reference, Release 24. Nutrient Data Laboratory Home Page, http://www.ars.usda.gov/ba/bhnrc/ndl
- Van Dam F and Van Gool WA: Hyperhomocysteinemia and Alzheimer's disease: A systematic review. *Arch Gerontol Geriatr* 2009; **48**:425-430.
- Van Rooij IA, Ocke MC, Straatman H, Zielhuis GA, Merkus HM and Steegers-Theunissen RP: Periconceptional folate intake by supplement and food reduces the risk of nonsyndromic cleft lip with or without cleft palate. *Prev Med* 2004; **39**:689-694.
- Van Rooij IA, Vermeij-Keers C, Kluijtmans LA, Ocké MC, Zielhuis GA, Goorhuis-Brouwer SM, van der Biezen JJ, Kuijpers-Jagtman AM and Steegers-Theunissen RP: Does the interaction between maternal folate intake and the methylenetetrahydrofolate reductase polymorphisms affect the risk of cleft lip with or without cleft palate? *Am J Epidemiol* 2003; **157**:583-591.
- Vieira AR, Murray JC, Trembath D, Orioli IM, Castilla EE, Cooper ME, Marazita ML, Lennon-Graham F and Speer M: Studies of reduced folate carrier 1 (RFC1) A80G and 5,10-methylenetetrahydrofolate reductase (MTHFR) C677T polymorphisms with neural tube and orofacial cleft defects. *Am J Med Genet A* 2005; **135**:220-223.
- Wald N: Prevention of neural-tube defects: results of the Medical Research Council Vitamin Study. *Lancet* 1991; **338**:131-137.
- Walkup AS and Appling DR: Enzymatic characterization of human mitochondrial C1-tetrahydrofolate synthase. *Arch Biochem Biophys* 2005; **442**:196-205.
- Wang L, Chen J, Chen L, Deng P, Bu Q, Xiang P, Li M, Lu W, Xu Y, Lin H, Wu T, Wang H, Hu J, Shao X, Cen X and Zhao YL: H-NMR based metabonomic profiling of human esophageal cancer tissue. *Molecular Cancer* 2013, **12**:25.
- Wehby G and Cassell CH: The impact of orofacial clefts on quality of life and healthcare use and costs. *Oral Dis* 2010; **16**:3-10.
- Wehby GL and Murray JC: Folic acid and orofacial clefts: a review of the evidence. *Oral Dis* 2010; **16**:11-19.

- Wessler SR: Transposable elements and the evolution of eukaryotic genomes. *Proc Natl Acad Sci USA* 2006; **103**:17600-17601.
- Wightman B, Ha I and Ruvkun G: Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell* 1993; **75**:855-862.
- Wilcox A, Lie RT, Solvoll K, McConnaughey DR, Abyholm F, Vindenes H, Vollset SE and Drevon CA: Folic acid supplements and risk of facial clefts: national population based case-control study. *BMJ* 2007; **334**:464.
- Winter J, Jung S, Keller S, Gregory RI and Diederichs S: Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nat Cell Biol* 2009; **11**:228-234.
- Woeller CF, Anderson DD, Szebenyi DM, and Stover PJ: Evidence for small ubiquitin-like modifier-dependent nuclear import of the thymidylate biosynthesis pathway. *J Biol Chem 2007*; **282**:17623-17631.
- Wuchty S, Fontana W, Hofacker IL and Schuster P: Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* 1999; **49**:145-165.
- Xing J, Wang H, Belancio VP, Cordaux R, Deininger PL and Batzer MA: Emergence of primate genes by retrotransposon mediated sequence transduction. *Proc Natl Acad Sci USA* 2006; **103**:17608-17613.
- Yanaihara N, Caplen N, Bowman E, Seike M, Kumamoto K, Yi M, Stephens RM, Okamoto A, Yokota J, Tanaka T, Calin GA, Liu CG, Croce CM and Harris CC: Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. *Cancer Cell* 2006; **9**:189-198.
- Ye Y, Xiao Y, Wang Wet, Wang Q, Yearsley K, Wani AA, Yan Q, Gao JX, Shetuni BS and Barsky SH: Inhibition of expression of the chromatin remodeling gene, SNF2L, selectively leads to DNA damage, growth inhibition, and cancer cell death. *Mol Cancer Res* 2009; **7**:1984-1999.
- Yi R, Qin Y, Macara I G and Cullen BR: Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev* 2003; **17**:3011–3016.
- Yu Z, Li Z, Jolicoeur N, Zhang L, Fortin Y, Wang E, Wu M and Shen SH: Aberrant allele frequencies of the SNPs located in microRNA target sites are potentially associated with human cancers. *Nucleic Acids Res* 2007; **35**:4535-4541.

- Yudate HT, Suwa M and Irie R: HUNT: launch of a full-length cDNA database from the Helix Research Institute. *Nucleic Acids Res* 2001; **29**:185-188
- Yuva-Aydemir Y, Simkin A, Gascon E and Gao FB: MicroRNA-9: functional evolution of a conserved small regulatory RNA. *RNA Biol* 2011; **8**:557-564.
- Zhang L, Liu Y, Song F, Zheng H, Hu L, Lu H, Liu P, Hao X, Zhang W and Chen K: Functional SNP in the microRNA-367 binding site in the 3′UTR of the calcium channel ryanodine receptor gene 3 (RYR3) affects breast cancer risk and calcification. *Proc Natl Acad Sci USA* 2011; **108**:13653-13658.
- Zhang L, Ren A, Li Z, Hao L, Tian Y and Li Z: Folate concentrations and folic acid supplementation among women in their first trimester of pregnancy in a rural area with a high prevalence of neural tube defects in Shanxi, China. *Birth Defects Res A Clin Mol Teratol* 2006; **76**:461-466.
- Zhang Z, Harrison PM, Liu Y and Gerstein M: Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res* 2003; **13**:2541–2558.
- Zheng D, Haddadin S, Wang Y, Gu LQ, Perry MC, Freter CE and Wang MX: Plasma microRNAs as novel biomarkers for early detection of lung cancer. *Int J Clin Exp Pathol* 2011; **4**:575-586.
- Zhuge J and Cederbaum AI: Depletion of *S*-adenosyl-L-methionine with cycloleucine potentiates cytochrome P450 2E1 toxicity in primary rat hepatocytes Cederbaum. *Arch Biochem Biophys* 2007; **466**:177-185.