Next generation sequencing to identify genes underlying methylmalonic aciduria

By

Lina Sobhi Abdrabo

Department of Human Genetics

McGill University

Montreal, Quebec Canada

Date

February 2019

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree

of

Masters of Science

Abstract

Inborn errors of cobalamin metabolism give rise to an array of clinical disorders with hematological and neurological manifestations, elevations of methylmalonic acid and/or homocysteine in blood and/or urine. Adenosylcobalamin is required as a cofactor for methylmalonyl-CoA mutase in the production of succinyl CoA from methylmalonyl CoA, and methylcobalamin is required as a cofactor for methionine synthase in the remethylation of homocysteine to methionine. Over the past 40 years, our laboratory has studied fibroblasts from patients with elevated methylmalonic acid, homocysteine, or both, using somatic cell complementation studies to classify patients into different complementation groups. These studies have allowed further investigations for gene discovery. The laboratory has accumulated over 200 cell lines from patients with elevated methylmalonic acid levels in which no genetic cause could be identified. Following analysis of DNA from 188 patients with a next generation sequencing gene panel, there remained 127 patients with elevated methylmalonic acid and no genetic diagnosis. Cultured fibroblasts from 26 of these patients had low [^{14}C] propionate incorporation into macromolecules, possibly reflecting decreased methylmalonyl-CoA mutase function. Whole genome sequencing was performed on genomic DNA extracted from these patients. Copy number variation analysis was employed to explore structural variations such as duplications and deletions that could be causal. In addition, RNA sequencing was performed to detect monoallelic events, splicing defects and to investigate any variants that could have been missed by whole genome sequencing. In 3 different patients, the analysis has revealed pathogenic events in two genes (*PCCA*, *EPCAM*),

and a 17q12 duplication. In an additional patient that was added to the study, there were pathogenic findings in the *TTN* gene. These results do explain part of the phenotype in these respective patients. One patient carries one nonsense variant and an intragenic duplication in *PCCA*, (NM_000282.3:c.1749_1750delGGinsTT). The nonsense variant causes a stop- gained mutation and leads to a stop codon which is involved in propionic academia. The second patient had a homozygous frameshift mutation, c.583dupC; p.Gln195fs, in *EPCAM* that has been reported to be pathogenic by ClinVar for diagnosis of chronic diarrhea. The third patient had a duplication, location 17q12 which corresponds to 17q12 duplication syndrome. These findings do not explain the elevated methylmalonic acid in this cohort of patients. Somatic cell studies and targeted sequencing such as a cobalamin gene panel have been used as the standard methods of diagnosis. Whole genome and RNA-sequencing did not detect any additional genetic defects in cobalamin genes, indicating that somatic cell studies including complementation analysis are reliable for the diagnosis of patients with inborn errors of cobalamin metabolism.

Resume

Les erreurs innées du métabolisme de la cobalamine donnent lieu à une série de troubles cliniques avec des manifestations hématologiques et neurologiques et à des élévations de l'acide méthylmalonique et / ou de l'homocystéine dans le sang et / ou l'urine. L'adénosylcobalamine est nécessaire en tant que cofacteur pour la méthylmalonyl-CoA mutase dans la production de succinyl-CoA à partir de méthylmalonyl-CoA, et la méthylcobalamine en tant que cofacteur pour la méthionine synthase dans la reméthylation de l'homocystéine en méthionine. Au cours des 40 dernières années, notre laboratoire a étudié les fibroblastes de patients présentant une élévation de l'acide méthylmalonique, de l'homocystéine, ou les deux, en utilisant des études de complémentation en cellules somatiques pour classer les patients en différents groupes de complémentation. Ces études ont permis de poursuivre les recherches sur la découverte de gènes. Le laboratoire a accumulé plus de 200 lignées cellulaires de patients présentant des niveaux élevés d'acide méthylmalonique dans lesquels aucune cause génétique n'a pu être identifiée. Après analyse de l'ADN de 188 patients porteurs d'un panel de gènes de séquençage de prochaine génération, il restait 127 patients avec un taux élevé d'acide méthylmalonique et aucun diagnostic génétique. Les fibroblastes en culture de 26 de ces patients présentaient une faible incorporation de propionate [14C] dans les macromolécules, reflétant peut-être une diminution de la fonction méthylmalonyl-CoA mutase. Un séquençage complet du génome a été réalisé sur l'ADN génomique extrait de ces patients. L'analyse de la variation du nombre de copies a été utilisée pour explorer les variations structurelles telles que les duplications et les suppressions pouvant être causales. En outre, le séquençage de l'ARN a été effectué

pour détecter les événements monoalléliques, les défauts d'épissage et pour rechercher les variants que le séquençage du génome entier aurait pu omettre. Chez 3 patients différents notre analyse a révélé événements pathogènes dans deux gènes (*PCCA*, *EPCAM*) et un duplication 17q12. Un autre patient qui a été ajouté à l'étude a également révélé des résultats pathogènes dans le gène *TTN*. Ces résultats expliquent une partie du phénotype chez ces patients. Un patient porte une variante non-sens et une duplication intragénique dans *PCCA*, (NM_000282.3: c.1749_1750delGGinsTT). La variante non-sens provoque une mutation arrêtée et conduit à un codon d'arrêt qui est impliqué dans les universités propioniques. Le deuxième patient avait une mutation homozygote, avec décalage de cadre, c.583dupC; p.Gln195fs, dans *EPCAM*, qui a été signalé comme pathogène par ClinVar pour le diagnostic de la diarrhée chronique. Le troisième patient avait une duplication, emplacement 17q12 qui correspond au syndrome de duplication 17q12. Ces découvertes n’expliquent pas l’acidité élevée de l’acide méthylmalonique dans cette cohorte de patients. Les études sur les cellules somatiques et le séquençage ciblé, comme un panel de gènes de la cobalamine, ont été utilisés comme méthodes de diagnostic standard. Le génome entier et le séquençage d'ARN n'ont pas détecté de défauts génétiques supplémentaires dans les gènes de la cobalamine, ce qui indique que les études sur les cellules somatiques, y compris l'analyse de complémentation, sont fiables pour le diagnostic des patients présentant des erreurs innées du métabolisme de la cobalamine.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ABCC1: ATP-binding cassette subfamily C member 1
ABCD4: ATP-binding cassette subfamily D member 4
AdoCbl: adenosylcobalamin
ADP: adenosine diphosphate
AMN: amnionless
ATP: adenosine triphosphate
ATR: cobalamin adenosyltransferase
bp: basepair
CD320: transcobalamin receptor
cDNA: complementary DNA
CNCbl: cyanocobalamin
CNV: copy number variant
CoA: coenzyme A
Cbl: Cobalamin
Cob(I)alamin: oxidation state of Co atom in cobalamin is +1
Cob(II)alamin: oxidation state of Co atom in cobalamin is +2
Cob(III)alamin: oxidation state of Co atom in cobalamin is +3
CUBN: cubilin
DNA: deoxyribonucleic acid
EMEM: Eagle’s minimum essential medium
ExAc: Exome Aggregation Consortium Database
g: gram
GATK: genome analysis toolki
GDP: guanosine diphosphate
hr: hour
HC: haptocorrin
HCFC1: host cell factor 1
IF: intrinsic factor
IGS: Imerslund-Gräsbeck Syndrome
kb: kilobase
L: litre
LMBD1: limb reduction domain 1
MAE: monoallelic expression site
MCEE: methylmalonyl-CoA epimerase
MCM: methylmalonyl-CoA mutase
MDRP1: multi drug resistance protein
MeCbl: methylcobalamin
MeTHF: methyltetrahydrofolate
mg: milligram
mL: millilitre
mm: millimeter
mM: millimolar
MMA: isolated methylmalonic aciduria
MMAA: methylmalonic aciduria type A protein

MMAB: cobalamin adenosyltransferase
MMACHC: methylmalonic aciduria and homocystinuria cblC type protein
MMADHC: methylmalonic aciduria and homocystinuria cblD type protein
mRNA: messenger RNA
MS: methionine synthase
MSR: methionine synthase reductase
Nl: normal
NGS: next generation sequencing
nmol: nanomole
OHCbl: hydroxycobalamin
PPA: propionic acid
PBS: phosphate buffered saline
PCC: propionyl-CoA carboxylase
PCR: polymerase chain reaction
PEA: pernicious anemia
PPI: proton pump inhibitors
RNA: ribonucleic acid
RNAi: interfering RNA
RT-PCR: reverse transcription polymerase chain reaction
RYGB: Roux-en-Y gastric bypass
SAM: S-adenosyl methionine
SCD: subacute combined degeneration of the spinal cord
SG: sleeve gastrectomy
SIFT: Sorting Intolerant from Tolerant
SNP: single-nucleotide polymorphism
SNV: single-nucleotide variant
SUCL: succinyl-CoA synthetase
SUCLA2: succinyl-CoA synthetase ADP-forming beta subunit
SUCLG1: succinyl-CoA synthetase alpha subunit
SUCLG2: succinyl-CoA synthetase GDP-forming beta subunit
SV: structural variation
TC: transcobalamin
TCbIR: transcobalamin receptor
TCA: trichloroacetic acid
THAP11: THAP domain containing 11
µg: microgram
µM: micrometre
WGS: whole genome sequencing
Wt: wild type
ZES: Zollinger Ellison syndrome
ZNF143: zinc finger protein 143

Acknowledgements

First and foremost, I would like to greatly thank my supervisor Dr. David Rosenblatt for all his support, mentorship and guidance for the past years. His endless support and belief in me have pushed me to reach my potential, and I am extremely grateful for this experience. I would also like to thank my co-supervisor Dr. Jean-Baptiste Rivière, his excellent teaching skills, extensive expertise and guidance has definitely taught me immensely and helped me succeed. Thanks also to Dr. Jacek Majewski for taking time to be a committee member and supporting this project till the end. Nonetheless, I would like to greatly thank Dr. David Watkins for teaching me everything there is cobalamin related, discussing papers and interesting concepts. In addition, I would like to thank Dr. Brian Gilfix for his mentoring and advice all throughout my project during the past two years. I would like to thank Keo Phommarihn endlessly for being so patient with providing patient cell lines, as well as Jocelyne Tossa, and Sina Yak. Thanks to Judith St-Onge for accommodating me and helping me complete crucial experiments. Also, I must thank Joel Lafond Lapalme, our bioinformatician for his incredible work with running the data analysis and providing all the software tools necessary. I cannot begin to thank Sophie Ran Wang for all her remarkable work on the variant filtering process, for everything she taught me and for all her support. Thanks to Mihaela Pupuvac for teaching me all the basics of bench work, thanks to Jordan Chu, Courtney Ells, Camilah Arbabian, Gawa Bidla, and Najmah Almuhssen for everything they taught me, our time in the lab and all their support. I would like to thank the beautiful Linda Chernin Rosenblatt, for her most genuine love, support and for always believing in me, for her elegance and kindness and true patient care. I would like to thank my husband and lifetime partner, Mustafa Al Shakr, for providing me with infinite love and care, allowing the best environment for my studies. Finally, but most importantly I would like to thank my dearest parents, Nasreldin Sobhi and Magda Gad, my brother Mahmood Sobhi, for being my source of joy in life, my backbone, and for raising me the way they did. Their indescribable support during my most stressful times was endless, they showed me they are proud of me every day, and finally they taught me that what truly matters is my happiness wherever it is.

Contributions of Authors

The completion of this research project and the work involved starting with examining clinical files of the accumulated cohort of patients with elevated MMA and unknown genetic diagnosis, extracting necessary medical information, selecting 26 patients, growing these specific cell lines by cell culture, extracting genomic DNA and RNA, ensuring high quality and quantity of genomic material by specific testing, sending for sequencing, ensuring appropriate Compute Canada cloud processing for running the pipeline, overseeing the completion of the data analysis and subsequent variant filtering process, extracting HPO terms for each patient, manual variant and CNV examination from the filtered variant call set for each patient, identifying pathogenic findings in four patients, and finally writing of this thesis was all of my contribution to this project. Joel Lafond Lapalme is the bioinformatician that was responsible for converting FASTQ files that were received after sequencing into variant call sets using the data analysis pipeline, and providing all necessary toolkits for the generation of Figure 7 and 9. Sophie Ran Wang was responsible for the application of certain filtering parameters to the generated variant call set, ensuring successful completion of this process, evaluating the pathogenic findings with me, and generating Figure 8. My co-supervisor, Dr. Jean-Baptiste Rivière was responsible for overseeing the successful completion of this project, reviewing pathogenic findings and editing of this thesis. Dr. David Watkins, has also contributed by reviewing and editing this thesis. Dr. David Rosenblatt, the principal investigator has been involved from the beginning to the end of this project, to ensure successful completion, and contributed by reviewing and editing this thesis.

CHAPTER 1

Introduction to Cobalamin Metabolism

1.0 Cobalamin structure

Cobalamin (Cbl) is a complex vitamin that is essential in human beings for proper cellular metabolism. Cbl structure consists of a corrin ring bound to a central cobalt atom (Figure 1). There are four out of the six coordination sites of the cobalt atom that are provided by the corrin ring. A fifth coordination site is provided by a 5,6-dimethylbenzimidazole base in the lower axial position that is also covalently bound to the corrin ring. Cbl can be in either a “base-on” or “base-off” configuration depending on whether or not the 5,6-dimethylbenzamidizole base is coordinated to the central cobalt atom. The sixth coordination site in the upper axial position may bind with multiple different compounds. A cyanide group occupies this position in cyanocobalamin (CNCbl), which is a synthetic form of Cbl used pharmaceutically. A hydroxyl group occupies this coordination site in hydroxocobalamin (OhCbl). Naturally occurring cobalamins include OHCbl, 5’-deoxyadenosylcobalamin (AdoCbl), methylcobalamin (MeCbl) and glutathionylcobalamin (GSCbl). The upper-axial ligand also distinguishes AdoCbl and MeCbl, the two Cbl derivatives that are active in human metabolism. The central cobalt atom in Cbl has three oxidation states: fully oxidized Co^{+++} (cob(III)alamin), Co^{++} (cob(II)alamin) and the fully reduced Co^{+}(cob(I)alamin). (Watkins and Rosenblatt, 2011)

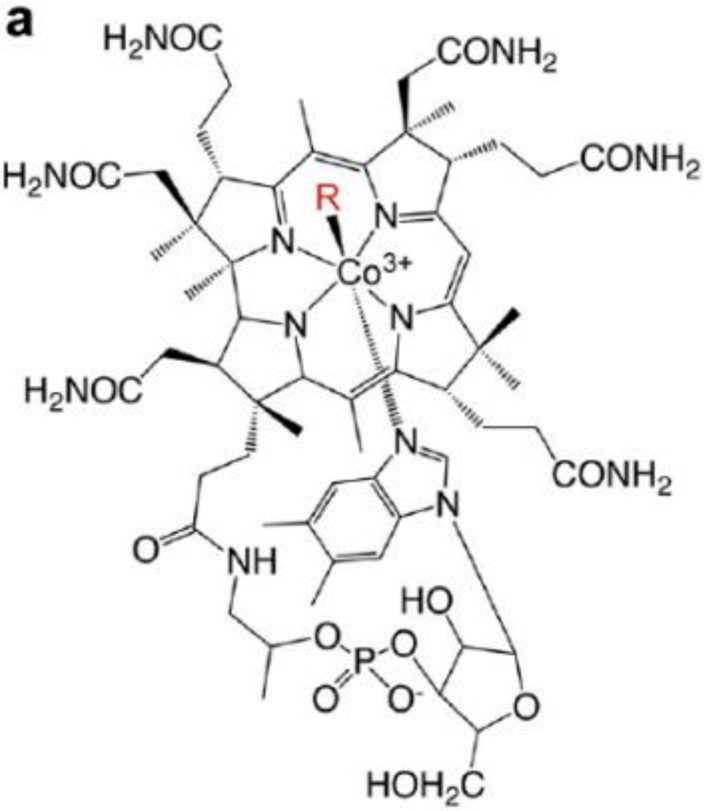


Figure 1. Cobalamin structure Cbl consists of a planar corrin ring bound to a central cobalt atom. Multiple different compounds (Rgroups), which distinguish the different derivatives of Cbl; MeCbl, CNCbl and AdoCbl can bind to the central cobalt. Adapted from Gherasim et al. Originally published in the Journal of Biological Chemistry (Gherasim, Lofgren and Banerjee, 2013) Navigating the B_{12} Road: Assimilation, Delivery, and Disorders of Cobalamin. Journal of Biological Chemistry. 2013; 288:13186-93. © the American Society for Biochemistry and Molecular Biology.

1.1 Cobalamin absorption and transport:

There are several steps involved in Cbl transport and metabolism; these include absorption, transport and intracellular utilization by the Cbl dependent enzymes methylmalonyl-CoA mutase (MCM) and methionine synthase (MS) (Figure 2). Ingested proteins are digested by proteases and hydrochloric acid secreted by the gastric mucosa. Haptocorrin is released from salivary glands and binds to Cbl in the stomach after it dissociates from other dietary proteins. Haptocorrin is digested in the intestine by pancreatic proteases and Cbl binds to a second protein named intrinsic factor (IF). IF is secreted by gastric parietal cells, and together with Cbl forms an IF-Cbl complex that binds to the receptor cubam in the distal ileum. Cubam consists of two proteins, cubilin and amnionless. After binding to cubam, the IF-Cbl complex is internalized by receptor mediated endocytosis. Cbl exits enterocytes and subsequently enters the bloodstream, where it circulates bound to transcobalamin (TC) (Andersen *et al.*, 2010). TC-Cbl can bind the transcobalamin receptor (TCblR) on the cell surface and is engulfed by receptor mediated endocytosis.

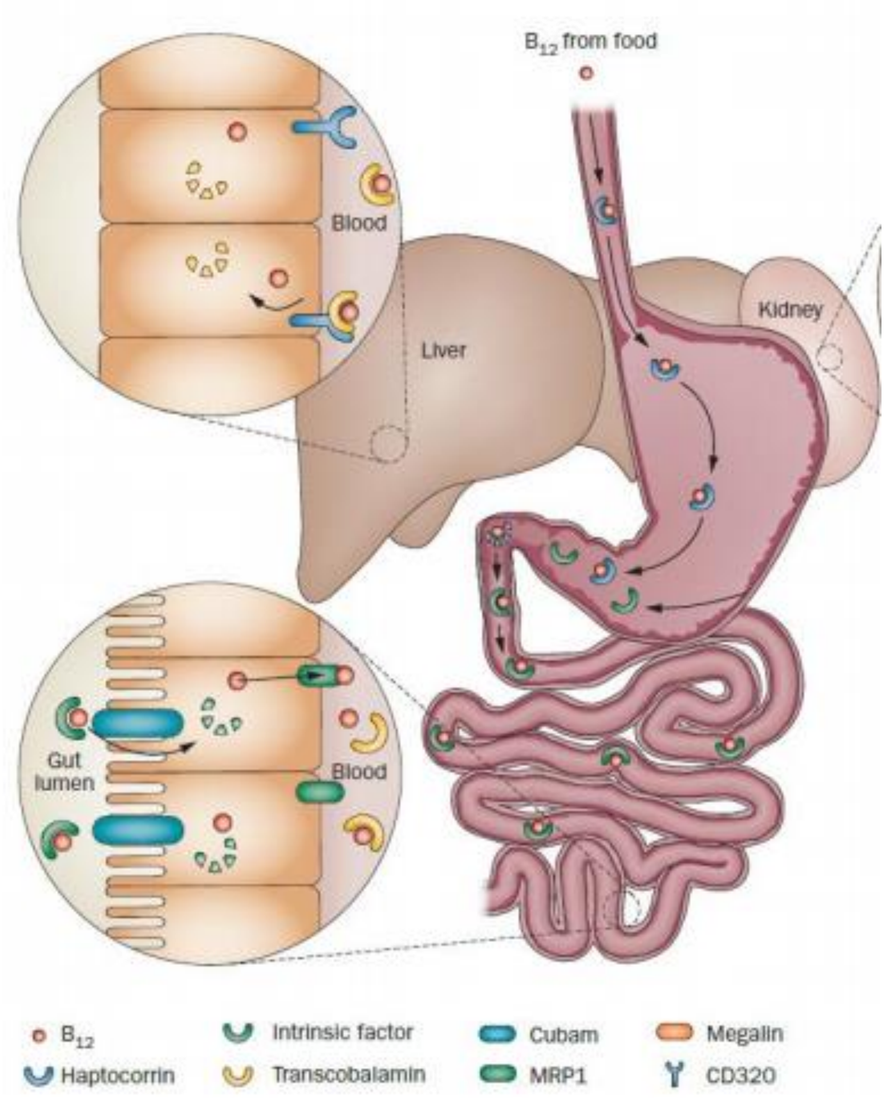


Figure 2. Cobalamin absorption and transport. Cbl ingested from food sources binds to Haptocorrin in the stomach, and passes along the gastrointestinal tract to reach the duodenum. In the duodenum, haptocorrin is digested by proteases and Cbl binds to IF. In the distal ileum the Cbl-IF complex is recognized by cubam, a specific receptor complex and is then taken up by enterocytes. Cbl is then released from IF by lysosomal enzymes inside the cell and free Cbl is exported across the basolateral membrane through a process involving ATP-binding cassette drug transporter (ABCC1), also known as multidrug resistance protein [MRP1]. Cbl associates with TC in the blood which facilitates cellular uptake. Cellular absorption of the TC-Cbl complex is mediated by TCbIR. A process of reabsorption of Cbl from glomerular filtrate in the kidney is facilitated by megalin, a receptor on the apical membrane of proximal tubule cells. Adapted from Nielsen et al. 2012 (Nielsen *et al.*, 2012).

1.2 Intracellular processing of cobalamin:

Cbl dissociates from TC and enters the cytoplasm from the lysosome with the aid of ABCD4 and LMBD1 proteins. In the cytoplasm, MeCbl is required as a cofactor for MS which, with methionine synthase reductase (MSR) catalyzes the conversion of homocysteine to methionine. In the mitochondrion, Cbl is converted to AdoCbl by cobalamin adenosyltransferase (MMAB), and transferred to MCM with the assistance of methylmalonic aciduria type A protein (MMAA). This function has been suggested based on bacterial homolog studies of MMAA. MMAA also transfers AdoCbl from MMAB to MCM, as well as plays a role in protection of MCM-bound AdoCbl from degradation and removal and replacement of damaged AdoCbl (Gherasim, Lofgren and Banerjee, 2013). MMAB functions as an ATP:cob(I)alamin adenosyltransferase (ATR) that transfers 5′-deoxyadenosyl from ATP to Cbl forming AdoCbl and delivers it to MCM (Froese et al. 2010). MCM catalyzes the conversion of methylmalonyl-CoA to succinyl-CoA. Decreased MTR activity results in accumulation of homocysteine in blood and urine; decreased MCM activity results in accumulation of methylmalonic acid in blood and urine. The Cbl pathway overlaps with the folate pathway where the methylation of homocysteine to methionine occurs, another simultaneous reaction takes place and MS catalyzes the reaction of converting methyltetrahydrofolate (MeTHF) to tetrahydrofolate (THF) (Watkins and Rosenblatt, 2011).The human genes responsible for the intracellular conversion of Cbl to its cofactors have been identified (Figure 3), and the corresponding mutations with biochemical phenotypes have been classified into distinct complementation groups. These complementation groups will be further discussed in detail later.

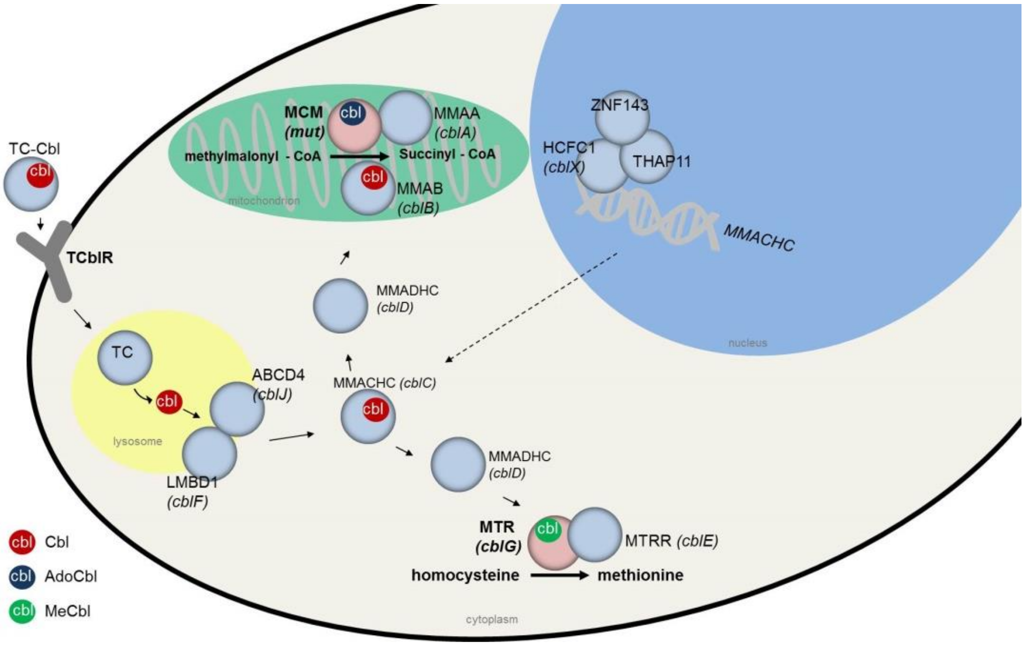


Figure 3. Intracellular cobalamin processing: All the genes implicated in the intracellular processing of Cbl are labelled and their corresponding Cbl disorder. After binding of the Cbl-TC complex on the TC receptor, following endocytosis, TC dissociates from Cbl in the lysosome. Free Cbl then exits the lysosome with the help of two proteins LMBD1 and ABCD4. Upon exiting the lysosome, Cbl is processed into its coenzyme derivatives, AdoCbl and MeCbl. The MMACHC (methylmalonic aciduria and homocystinuria cblC type) protein interacts with Cbl with various upper-axial ligands, converting them to a common intermediate that can be used for synthesis of AdoCbl and MeCbl. Pupavac M, 2017, Next generation sequencing to discover genes for Mendelian disorders, Thesis Dissertation, McGill University.

THAP11 (THAP domain containing 11), HCFC1 (host cell factor 1), and ZNF143 (zinc finger protein 143) are transcription factors that have been recently identified as modulators of MMACHC expression (Yu *et al.*, 2013). HCFC1 acts as a scaffold protein that facilitates the regulatory roles of the THAP11 and ZNF143 protein (Michaud *et al.*, 2013). Clinically-significant decrease of MMACHC protein levels have been shown to be a result of mutations affecting the HCFC1, THAP11 and ZNF143 genes (Quintana *et al.*, 2017). MMADHC (methylmalonic aciduria and homocystinuria cblD type) protein enables the distribution of Cbl to either the mitochondria or the cytoplasm for synthesis of AdoCbl or MeCbl, respectively. Cbl that remains in the cytoplasm for MeCbl synthesis is delivered to MS, which catalyzes a methyl transfer from MeTHF to homocysteine to form methionine and THF in a two-step reaction. In the mitochondrion, the successful synthesis of AdoCbl and delivery to MCM requires two proteins, encoded by the MMAA and MMAB genes. ATR encoded by the MMAB gene, is responsible for the generation of AdoCbl from cob(II)alamin. Cob(II)alamin binds to ATR in the "base-off" configuration, which favors reduction to the cob(I)alamin state by 100-fold compared to cob(II)alamin in the "base-on" configuration (Yamanishi, Vlasie and Banerjee, 2005). The ATR-coupled protein that catalyzes this reduction remains to be identified. ATR subsequently catalyzes adenosylation of cob(I)alamin to form AdoCbl. ATR is also directly involved in the transfer of AdoCbl to MCM in a process that is believed to be dependent on the MMAA gene product (Padovani *et al.*, 2008).

1.3 Dietary requirements for Cobalamin:

Derivatives of Cbl are required for human intermediary metabolism: AdoCbl is required for activity of MCM, the mitochondrial enzyme that catalyzes conversion of methylmalonylCoA, generated during catabolism of branched-chain amino acids, odd-chain fatty acids and cholesterol, to succinylCoA that can be metabolized by the Krebs cycle; and MeCbl is required for activity of MS. The daily recommended dose of Cbl is 2.4 µg per day, for both males and females above 18 years of age, this amount is increased in pregnant or lactating women for optimal health (Institute of Medicine (US) Standing Committee on the Scientific Evaluation of Dietary Reference Intakes and its Panel on Folate, 1998).

1. 4 Cobalamin Deficiency:

Cbl deficiency occurs due to a variety of reasons such as impaired Cbl absorption, decreased Cbl intake, or due to the autosomal recessive inborn errors of Cbl metabolism. It presents with a wide range of medical disorders affecting most importantly the hematological and neurological systems.

The reference range of serum Cbl is from 200- 1000 pg/ml (150 – 750 pmol/L) and Cbl deficiency is defined as a value of less than 200 pg/ml (150 pmol/L). (Qiu *et al.*, 2006). Nighty to ninety-five percent of patients with a Cbl deficiency that have hematological and/or neurological signs and symptoms will have a Cbl- serum level lower than 200 pg/ml. An estimate of 5% to 10% of patients have Cbl-serum levels in the range of 200-300 pm/ml. In addition to a small percentage of 0.1% to 1% have Cbl-serum levels greater than 300pg/ml.

The hematological presentation of Cbl deficiency ranges from the incidental finding of asymptomatic anemia, to severe symptoms that manifest as dyspnea on exertion and fatigue or those related to congestive heart failure. Neuropsychiatric findings can include myelopathy, neuropathy, dementia and, less frequently, optic nerve atrophy, which may precede the hematologic signs. In addition, subacute combined degeneration of the spinal cord (SCD) may occur which refers to the degeneration of the posterior and lateral columns of the spinal cord, and is characterized by symmetric dysesthesia, disturbance of position sense and spastic paraparesis or tetraparesis (Hemmer *et al.*, 1998). Cbl deficiency is one of the most common causes of SCD in addition to vitamin E deficiency and copper deficiency.

1.4.1 Hematological Manifestations:

Cbl and folate deficiency cause megaloblastic anemia, which is the first clinical presentation in IF deficiency. It is suggested that the process of nucleic acid metabolism is impaired, resulting in nuclear-cytoplasmic dyssynchrony, reduced number of cell divisions in the bone marrow, and nuclear abnormalities in both myeloid and erythroid precursors. The increase of the mean corpuscular volume is the earliest manifestation of megaloblastosis. Red blood cells appear larger than normal, and some of them may lose the central pale area (Briani *et al.*, 2013a). During adolescence or adulthood, clinical manifestations start to manifest in some cases where there is partial IF deficiency.

1.4.2 Neurological Manifestations:

Cbl dependant coenzymes in the MCM reaction are necessary for myelin synthesis. Therefore, Cbl deficiency results in defective myelin synthesis, leading to central and peripheral nervous system dysfunctions. The lack of AdoCbl that functions as a cofactor for the conversion of methylmalonyl-CoA to succinyl-CoA leads to accumulation of methylmalonyl-CoA, causing a decrease in normal myelin synthesis and incorporation of abnormal fatty acids into neuronal lipids (Briani et al. 2013). Another opinion contradicts this hypothesis and strongly suggests that the neurological damage occurs more due to the methyl group dependant reactions. An old study hypothesized that impaired MS activity leads to deficiencies of both methionine and S-adenosylmethionine (SAM) (Scott et al. 1981). Since SAM is a key intermediary in methylation reactions, SAM deficiency could impair methylation reactions in myelin. Therefore, methyl group deficiency could result in demyelination and clinical neuropathy (Jack Metz, 1993).

Neuropsychiatric symptoms such as peripheral neuropathy, developmental delay, ataxia, dementia and in rare cases optic atrophy in elderly patients, may precede hematologic manifestations and could be the presenting indicator of a Cbl deficiency.

Impairment of position sense in SCD occurs due to the involvement of the posterior and lateral columns of the spinal cord in the cervical and upper dorsal region. Distal and symmetrical sensory impairment of the lower limbs is usually the first abnormality presenting, which is frequently associated with ataxia (Briani et al. 2013).

1.5 Impaired cobalamin absorption causes:

Impaired absorption of Cbl could occur due to a variety of medical conditions such as pernicious anemia, gastrectomy, Zollinger Ellison Syndrome, and fish tapeworm. Pernicious anemia (PEA) previously defined as lethal, is now a rare condition that causes improper absorption of Cbl and subsequently impaired red blood cell formation and anemia. PEA is a type of chronic atrophic gastritis, that is macroscopically detected by thinning and loss of gastric mucosal folds. Two types of chronic atrophic gastritis are classified; type A is the autoimmune gastritis and type B the non-autoimmune, which is believed to be associated with *Hellicobacter pylori* infection. Type A gastritis is associated with PEA, autoantibodies to gastric parietal cells and to intrinsic factor, achlorhydria, low serum pepsinogen I concentrations, and high serum gastrin concentrations, the latter resulting from hyperplasia of gastrin-producing cells (Brzezinski A., 1997). Since IF plays a crucial role in the absorption of Cbl, during this autoimmune disorder, IF deficiency causes the malabsorption of Cbl. The clinical presentation of this disorder includes the hematological and neurological manifestations that occur with Cbl deficiency, as well as gastrointestinal manifestations that may include atrophic glossitis, diarrhea and malabsorption. The disease is now treated with intramuscular injections of Cbl (Lahner and Annibale, 2009).

A sleeve gastrectomy (SG) and the Roux-en-Y gastric bypass (RYGB) are types of bariatric surgery performed as successful treatments for morbid obesity. Recently, these types of surgeries have been successful at greatly lowering body mass index and decreasing comorbidities that occur with obesity. With the increasing popularity of such procedures, they have become affordable and commonly requested by patients. However, the serious nutritional defects that arise with bariatric surgery are not sufficiently addressed. Vitamin and mineral deficiencies such as Cbl and iron deficiency are common, and require patients to take supplements to prevent these deficiencies. Cbl deficiency may occur due to the decreased digestion of protein bound Cbl due to decrease acid and pepsin, limited mixing of nutrients with pancreatic secretions causes the incomplete release of Cbl from R proteins, and decreased availability of intrinsic factor. The small pouch constructed from the gastric cardia in the RYGB procedure does not contain secreted gastric acid, and food-bound Cbl is maldigested and subsequently malabsorbed. There is evidence that SG had better outcomes with respect to postoperative Cbl deficiency than RYGB. (Kwon *et al.*, 2014). Oral supplementation of vitamin B_{12} has shown to successfully correct deficiencies in 81% of cases (Brolin and Leung, 1999) (Gorman *et al.*, 1998). Multivitamin supplementation alone that contained 10 μg CNCbl, did not prevent Cbl deficiency in bariatric patients, however, intramuscular CNCbl supplementation did correct the deficiency in 91% of patients (Gehrer *et al.*, 2010).

Zollinger Ellison Syndrome (ZES) is characterized by refractory and severe peptic disease that is caused by a gastrinoma; a type of a neuroendocrine tumor that causes acidic hypersecretion of gastrin. Since this disease cause acidic hypersecretion and hypergastrinemia, peptic ulcer disease, esophagitis, and esophageal strictures may occur. Clinical symptoms include heartburn, abdominal pain, weight loss, malabsorption and diarrhea. The appropriate treatment with antacid agents such as proton pump inhibitors (PPI) is essential to decrease the acidic hypersecretion and improve symptoms. The long-term use of PPI leads to decrease serum Cbl levels, which is potentially the most important side effect. It is unclear however, if long term PPI use will cause a significant Cbl deficiency (Gibril and Jensen, 2004).

The fish tapeworm, or more specifically *Diphyllobothrium latum* is the largest intestinal helminths that can infect people and can grow up to 30 feet long. Most infections are asymptomatic; however, complications include intestinal obstruction, gall bladder disease and Cbl deficiency. Microscopic examination of stool samples identifies eggs or segments of the tapeworm and this concludes a diagnosis. The parasite-mediated dissociation of the Cbl-IF complex within the gut lumen makes Cbl unavailable to the host, this occurs during prolonged or heavy *D.Latum* infections. (Scholz *et al.*, 2009). Safe and effective medications are available to treat *Diphyllobothrium*. Eating raw or undercooked fish, usually from the Northern Hemisphere (Europe, newly independent states of the Former Soviet Union, North America, Asia) causes infections, however cases have been reported in Uganda and Chile. Fish infected with *Diphyllobothrium* larvae may be transported to and consumed in any area of the world. (www.cdc.gov/parasites/diphyllobothrium/index.html)

1.6 Decreased cobalamin intake:

Cbl is ingested and acquired from animal food sources such as meat and dairy and is almost always bound to proteins. Since Cbl is found only in animal food products, strict vegans or vegetarians are at risk of developing a deficiency if a supplement is not added to their diets (Stabler and Allen, 2004). A vegan diet has shown to typically present with lower plasma Cbl concentrations, higher prevalence of Cbl deficiency, and higher concentrations of plasma homocysteine. Besides the neurological side effects of Cbl deficiency, children following a vegan diet can develop apathy and failure to thrive, and megaloblastic anemia at all ages (Craig, 2009).

1.7 Inborn errors of cobalamin metabolism:

Inborn errors affecting Cbl absorption (inherited intrinsic factor deficiency, Imerslund–Gräsbeck syndrome) and transport (TC deficiency) have been described. A series of inborn errors of intracellular Cbl metabolism, designated *cblA-cblG*, *cblJ*, and *cblX*, have been differentiated by complementation analysis. Cbl disorders that give rise to isolated methylmalonic acidemia are recognized as *cblA*, *cblB*, *cblD* variant 2, while the ones responsible for isolated homocystinuria are *cblD* variant 1, *cblE*, *cblG* and finally for combined methylmalonic acidemia and homocystinuria the identified disorders are *cblC*, classic *cblD*, *cblF*, *cblJ* (Watkins and Rosenblatt, 2011). These disorders are rare diseases that are inherited in an autosomal recessive manner, except *cblX*, which is an X-linked recessive disease. *cblX* is caused by mutations in *HCFC1* (Pupavac *et al.*, 2016). Mutations in *SUCLG1*, *SUCLG2*, *MCEE,* and *ACSF3* have recently been identified in patients with mild MMA (Chu *et al.*, 2016).

1.7.1 Combined methylmalonic aciduria and homocysteinuria

These disorders are categorized into complementation groups that affect both key enzymatic reactions that occur intracellularly and are facilitated by MCM and MS respectively. Here we discuss these different complementation groups and their subsequent disorders.

1.*cblF*

Defects in the *LMBDR1* gene lead to a loss of *LMDB1* function, and patients having these defects belong to the *cblF* complementation group. *LMDB1* provides instructions for synthesizing the LMBD1 protein, that is available in the lysosomal membrane and together with the help of ABCD4 transports Cbl into the cytoplasm. Therefore, defects in *LMBDR1* lead to defective Cbl transport into the cytoplasm and hence free Cbl accumulation in the lysosome. As a result, synthesis of both AdoCbl and MeCbl is inadequate and activity of both Cbl-dependent enzymes is low. Patients with the *cblF* disorder have elevations in both methylmalonic acid and homocysteine since both key enzymes are deficient. The majority of the *cblF* patients harbor the c.1056delG frameshift mutation in *LMBRD1* (Rutsch *et al.*, 2009). Failure to thrive, developmental delay, anemia, neutropenia, facial abnormalities, and congenital heart defects are some of the clinical findings that may be associated with the disorder, however, clinical findings tend to be variable (Gailus *et al.*, 2010). Recently, different frameshift mutations leading to loss of function of both *LMBRD1* alleles were detected in five patients (Rutsch *et al.*, 2011).

2.*cblJ*

Mutations affecting the *ABCD4* gene give rise to the *cblJ* disorder, which is characterized by accumulation of free Cbl in the lysosome, similar to the *cblF* disease. The first two patients identified with this disorder suffered from hypotonia, poor feeding, macrocytic anemia, bone marrow suppression and cardiac defects (Coelho *et al.*, 2012). Clinical and cellular phenotypes of the few *cblJ* patients that have been identified show a degree of variability (Kim *et al.*, 2012a). Studies of cultured *cblJ* fibroblasts found apparently normal AdoCbl and moderately reduced MeCbl, and identified it as *cblJ* by WES. (Kim *et al.*, 2012b). The onset of symptoms in early childhood has been reported which included macrocytic anemia, MMA, and hyperhomocysteinemia (Kim *et al.*, 2012b). A case of progressive hyperpigmentation has been reported in a Taiwanese patient who presented with symptoms in childhood. (Takeichi *et al.*, 2015a)

3.cblC

The most common inborn error of Cbl metabolism is the *cblC* type, and this disorder is secondary to mutations in the *MMACHC* gene, the product of which is involved in the processing of Cbl for cofactor synthesis. The MMACHC protein has been shown to have the ability decyanate CNCbl to cob(II)alamin *in vitro* to in the presence of a reductase and to dealkylate Cbls containing C2–C6 alkanes, adenosyl or methyl as the upper axial ligand *in vivo* (D S Froese *et al.*, 2010). Defects in the MMACHC protein result in the abnormal synthesis of MeCbl and AdoCbl leading to reduced activity of both Cbl-dependent enzymes. Patients having this disorder present with elevated MMA and homocystinuria which can lead to a wide range of symptoms including failure to thrive, feeding difficulties, and hematological, neurological, ophthalmological, and dermatological abnormalities (Lerner-Ellis *et al.*, 2009). A recent study that has investigated the clinical and biochemical differences between neonatal and early-onset presentations, has found no major differences between these groups and combined them as an infatile-onset group. The clinical manifestations of this group predominantly included hypotonia, lethargy, feeding problems and developmental delay, while late-onset patients presented with psychiatric/behaviour problems and myelopathy (Fischer *et al.*, 2014). Elevated homocysteine concentrations and defective methyl group metabolism may contribute to disease-related complications, however, as previously mentioned that a wide presenation of clinical manifestations could be detected in this disorder. It has been shown that the characteristic macular and retinal degeneration that are seen in affected patients appear to be unique to *cblC* disease. (Carrillo-Carrasco and Venditti, 2012).

4.*cblD*

Patients with the *cblD* complementation group were previously thought to present with combined elevated MMA and homocystinuria. Mutations first discovered in *MMADHC* have that the MMADHC protein has sequence homology with a bacterial ATP-binding cassette transporter and contains a putative Cbl binding motif and a putative mitochondrial targeting sequence (Coelho *et al.*, 2008). Recent investigations have further characterized this disorder into *cblD*-variant 1, and *cblD*-variant 2, which vary from the combined MMA and homocystinuria presentation (Suormala *et al.*, 2004). Decreased synthesis of MeCbl and isolotaed homocystinuria occur due to mutations affecting a conserved C-terminus region (p.D246-L259) of the *MMADHC* gene, categorizing this as the *cblD*-variant 1 disorder. Combined MMA and homocystinuria arise due to nonsense mutations located across the C-terminus (p.Y140-R250. The *cblD*-variant 2 disorder arises due to mutations in the N-terminus of the *MMADHC* gene resulting in impaired AdoCbl synthesis and isolated MMA (Jusufi *et al.*, 2014). Depending on the underlying genetic defect, clinical manifestations of the *cblD* disorder vary among patients. These manifestations may include neurological impairment, metabolic acidosis, muscle weakness, vomiting, developmental delay, and megaloblastic anemia.

5.*cblX*

Mutations affecting the Kelch domain of the *HCFC1* gene are associated with the *cblX* disorder, which is an X-linked inborn error of Cbl metabolism (Yu *et al.*, 2013). The *HCFC1* gene is a member of the host cell factor family, that encodes a protein containing five Kelch repeats, as well as a fibronectin-like motif and six HCF repeats. Each repeat contains a highly specific cleavage signal. An N-terminal chain and the matching C-terminal chain are a result of the proteolytic cleavage at one of the six possible sites at the nuclear coactivator. The final form of this protein consists of noncovalently bound N- and C-terminal chains that plays a role in gene regulation. Although *cblX* presents phenotypically like the *cblC* disorder, complementation of fibroblasts from *cblX* patients does not occur with fibroblast from *cblC* patients. A male patient had been assigned to the *cblC* complementation group without mutations in *MMACHC* lead to the discovery of the *cblX* disorder. A missense mutation affecting the N-terminal Kelch domain of the *HCFC1* gene was identified by exome sequencing. Additionaly, the examination of 14 male patients who had been assigned to the *cblC* complementation class without mutations in *MMACHC,* lead to the discovery of *HCFC1* mutations in these patients. The investigation of *MMACHC* mRNA and protein showed decreased levels in patients' cells, proving that *HCFC1* plays a regulatory role in *MMACHC* expression. Clinical manifestations are severe and include neurological symptoms, intractable epilepsy and cognitive impairment in all affected patients. Numerous *cblX* patients have normal levels of plasma homocysteine, in contrast to the biochemical phenotype of early-onset *cblC* disease, showing that *cblX* disease appears to be more variable biochemicaly than *cblC* (Yu *et al.*, 2013).

6. *THAP11*

Mutations in *THAP11* in a single patient have been found to be associated with Cbl disorders by interacting with *HCFC1*.The *HCFC1*/*THAP11* complex binds to regulatory elements controlling the expression of *MMACHC* as well as a diverse array of downstream targets. This results in biochemical and other phenotypes similar to those observed in patients with *cblX* (Quintana *et al.*, 2017).

1.7.2 Isolated homocystinuria

1. *cblG*

The *cblG* disorder is caused by mutations in the *MTR* gene leading to defective MS, subsequently leading to the impaired remethylation of homocysteine to form methionine. In addition to isolated homocystinuria, patients also typically present with megaloblastic anemia, cognitive impairment, developmental delay, mental retardation, and decreased visual acuity (Huemer *et al.*, 2015). Muscular hypotonia appears to be more frequently reported in *cblG* patients, however, the *cblG* and the *cblE* disorders' clinical phenotype are quite similar (Huemer *et al.*, 2015).

2.*cblE*

The *cblE* disorder is caused by pathogenic mutations affecting the *MTRR* gene, encoding MSR, lead to reduced MS activity. Cloning of a cDNA corresponding to the MSR reducing system was shown to be required for the maintenance of MS in a functional state (Leclerc *et al.*, 1998). Early in life, this disorder is characterized by isolated homocystinuria and presents with megaloblastic anemia, failure to thrive, and neurological impairment (Huemer *et al.*, 2015). A deep intronic mutation (c.903+469T>C) that results in inclusion of a pseudoexon via creation of an exonic splicing enhancer site is the most frequent mutation associated with the *cblE* disorder (Homolova *et al.*, 2010).

1.7.3 Isolated Methylmalonic aciduria

1.*mut*

Isolated MMA in 60% of cases is caused by a defect in the Cbl-dependent enzyme MCM, the *mut* type of MMA (Manoli, Sloan and Venditti, 1993). The clinical presentation of patients affected with *mut* MMA typically include acute metabolic distress in the neonatal period, failure to thrive, recurrent vomiting, mild microcephaly, lethargy, hypotonia, dehydration, respiratory distress, hyperammonemia, and ketoacidosis (Hörster *et al.*, 2007a). Patients alternate between periods of relative health intersected with intermittent acute metabolic crises that may lead to death (Manoli, Sloan and Venditti, 1993). The identification of two cellular subtypes of the disorder have been achieved using early enzymatic studies performed on fibroblasts derived from *mut* patients to evaluate MCM activity (Mellman *et al.*, 1977). Virtually undetectable MCM activity in cell lines at basal and high concentrations of AdoCbl, were designated *mut*$^{0.}$ Cells that had an increased K_m for AdoCbl and showed measurable activity of MCM, ranging from 1-50% of wild-type, were designated *mut*$^{-}$ (Willard and Rosenberg, 1980). The ability to incorporate [^{14}C]propionate into acid-precipitable protein in fibroblasts from both *mut*^{0} and *mut* patients was decreased indicating defective metabolism of propionyl-CoA. The improvement of [^{14}C]propionate incorporation towards the reference range has been shown in *mut* cell cultures with the supplementation of OHCbl, while *mut*^{0} cultures are nonresponsive. Approximately 78% of *mut* patients belong to the *mut*^{0} subtype while the remaining 22% are *mut*$^{-}$ (Manoli, Sloan and Venditti, 1993). Underlying genetic defects and mutations in *MUT* have allowed the *mut* subtype classification as either mut^{0} or *mut*$^{-}$. The *mut*^{t0} and *mut*$^{-}$ phenotypes are difficult to distinguish, since there is very low residual enzymatic activity in *mut*$^{-}$ associated mutations (Janata, Kogekar and Fenton, 1997). Genetic analysis of the *MUT* gene and/or somatic cell studies allow for the diagnosis of *mut* MMA. To distinguish *mut* disorder from the other forms of MMA, somatic cell complementation analysis is reliable. However the existence of interallelic complementation within the *mut* phenotype complicates the process (Raff *et al.*, 1991).

2.*cblA*

Mutations in the *MMAA* gene are responsible for the *cblA* type of MMA where interrupted synthesis of AdoCbl leads to impaired function of MCM (Dobson *et al.*, 2002). When cells from *cblA* patients are supplied with exogenous CNCbl, decreased ability to synthesize AdoCbl is found. The common c.433C>T (p.R145*) mutation in *MMAA* is associated with European ancestry and accounts for nearly 50% of pathogenic alleles (Lerner‐Ellis *et al.*, 2004). Clinical presentations of *cblA* patients occur during the newborn period or early-childhood with severe, life-threatening metabolic acidosis secondary to elevated levels of MMA. Therapy with OHCbl therapy in *cblA* patients has shown to be responsive. Plasma MMA levels are significantly reduced following Cbl treatment in these patients (Hörster *et al.*, 2007b).

3.*cblB*

A severe form of isolated MMA is the *cblB* disorder that is associated with pathogenic mutations affecting the *MMAB* gene leading to dysfunctional ATR protein. Reduced ATR affinity for substrate and AdoCbl in identified mutant genotypes has been shown by functional analyses of *cblB*-causing mutations (Brasil *et al.*, 2015a). Complete or partial loss of enzymatic activity can occur in other cases and both early and late-onset forms of the disorder have been described. In the severe early-onset disease, patients suffer from symptoms of neonatal ketoacidosis, failure to thrive, and encephalomyopathy. The late-onset form is usually diagnosed in infancy and has a milder phenotype with less pronounced neurological impairment (Hörster *et al.*, 2007c). Fibroblasts derived from *cblB* patients have shown mitochondrial dysfunction, in which increased levels of reactive oxygen species and reduced energy production occur (Brasil *et al.*, 2015b). A definitive treatment for MMA *cblB* type does not exist, however, pharmacological doses of OHCbl can be administered. Only about 40% of patients experience a positive biochemical response.

1.8. Cobalamin-dependent propionyl-CoA metabolism

During the catabolism of certain branched-chain amino acids (isoleucine, threonine, and valine), methionine, odd-chain fatty acids, and cholesterol, propionyl-CoA is a common intermediate that is produced. In the Kreb's cycle, propionyl-CoA is channeled into intermediate succinyl-CoA through a process involving multiple nuclear-encoded mitochondrial enzymes in the Cbl-dependent pathway of propionyl-CoA metabolism. The conversion of propionyl-CoA to D-methylmalonyl-CoA is mediated by Propionyl-CoA carboxylase (PCC). PCC is a protein consisting of an α-subunit, encoded by the *PCCA* gene, and a β-subunit, encoded by the *PCCB* gene. Mutation in either *PCCA* or *PCCB* lead to propionic acidemia, a metabolic disorder characterized by episodic vomiting, lethargy and ketosis, neutropenia, periodic thrombocytopenia, hypogammaglobulinemia, developmental retardation, and protein intolerance. The second enzymatic reaction in Cbl-dependent propionyl-CoA metabolism is catalyzed by methylmalonyl-CoA epimerase (racemase) (*MCEE*). Based upon the finding that *MCEE* and *MUT* reside on the same operon in prokaryotic genomes, this suggested that they function in a synchronized manner, thus led to the hypothesis that the *MCEE* gene was involved in the propionyl-CoA metabolic pathway. This hypothesis has been confirmed by showing the interconversion of D- to L-methylmalonyl-CoA involved *MCEE* and has been demonstrated by biochemical studies (Bobik and Rasche, 2001). However, a mild reduction in propionyl-CoA metabolic activity has been shown by performing *MCEE* knockdown using small interfering RNA (siRNA) in HeLa cells, suggesting nonenzymatic conversion of D- to L-methylmalonyl-CoA may contribute to epimerization (Dobson *et al.*, 2006). In a recent study, pathogenic variations in *MCEE* have been identified in ten patients with MMA of unknown origin. Nine patients were homozygous for the known nonsense variation p.Arg47* (c.139C > T), and one for the novel missense variation p.Ile53Arg (c.158T > G). Further experiments were established to have a clearer picture of the molecular basis of *MCEE* deficiency, such as the work done by Heuberger (2019) found the following:

> A p.Ile53Arg was mapped, and two previously described pathogenic variations p.Lys60Gln and p.Arg143Cys, onto a 1.8 Å structure of wild type (wt) human MCEE. This revealed potential dimeric assembly disruption by p.Ile53Arg, but no clear defects from p.Lys60Gln or p.Arg143Cys were found. The structure of MCEE-Arg143Cys to 1.9 Å was solved and found significant disruption of two important loop structures, potentially impacting surface features as well as the active site pocket. Functional analysis of MCEE-Ile53Arg expressed in a bacterial recombinant system as well as patient-derived fibroblasts revealed nearly undetectable soluble protein levels, defective globular protein behavior, and using a newly developed assay, lack of enzymatic activity - consistent with misfolded protein findings. By contrast, soluble protein levels, unfolding characteristics and activity of MCEE-Lys60Gln were comparable to wt, leaving unclear how this variation may cause disease. MCEE-Arg143Cys was detectable at comparable levels to wt MCEE, but had slightly altered unfolding kinetics and greatly reduced activity. (Heuberger *et al.*, 2019)

Cbl-dependent MCM binds the L-isomer of methylmalonyl-CoA which catalyzes the rearrangement to succinyl-CoA. In the Krebs cycle, the reversible synthesis of succinate from succinyl-CoAis catalyzed by succinylCoA synthetase (SUCL). The reverse reaction produces succinyl-CoA for heme synthesis and activation of ketone bodies. SUCL consists of an α-subunit encoded by *SUCLG1* and a β- 27 subunit encoded by either the ADP-forming *SUCLA2* gene or the GDP-forming *SUCLG2* gene (Johnson *et al.*, 1998). Expression of *SUCLA2* have been identified exclusively in neurons in the cerebral cortex by localization studies in human brains (Dobolyi, Ostergaard, *et al.*, 2015) while expression of *SUCLG2* appears to be in the vasculature of the brain (Dobolyi, Bagó, *et al.*, 2015). Expression of both *SUCLA2* and *SUCLG2* was entirely absent in glial cells, suggesting the participation of alternate pathways such as the GABA shunt and ketone body metabolism in these cells (Dobolyi, Bagó, *et al.*, 2015).

1.9. Cobalamin-independent propionyl-CoA metabolism

The utilization of a β-oxidation-like pathway to prevent the toxic accumulation of propionyl-CoA has been shown in certain organisms that are non-Cbl dependent, such as plants, fungi and *Candida albicans* (Otzen *et al.*, 2014). A metabolic pathway has been recently characterized in *C. elegans* that involves the degradation of propionyl-CoA and is non-Cbl dependent (Watson *et al.*, 2016). A hypothesis was made that this alternate pathway provides *C. elegans* with the ability to metabolize propionyl-CoA in accordance with Cbl availability. Watson et al. have shown that *C. elegans* respond to elevated levels of propionyl-CoA by transcriptional activation of a Cbl-independent propionate shunt, when under conditions simulating dietary Cbl deficiency, or in genetic conditions that mimic propionyl-CoA metabolism disorders. Additionally, they show that Cbl transcriptionally represses the genes involved in this pathway (Watson *et al.*, 2016).

1.10 MMA in non-cobalamin disorders

As discussed previously, MMA accumulates when the conversion of methylmalonyl CoA to succinyl CoA fails. The long-term consequences of elevated MMA concern neurologic damage and terminal kidney failure. A recent study explored a *MUT* knockdown in a neuroblastoma cell line. Affected cellular pathways by *MUT* deficiency were examined through a quantitative proteomics method on a cellular model of *MUT* knockdown. A consistent reduction of the *MUT* protein expression was obtained in the neuroblastoma cell line (SH-SY5Y) by using small-interfering RNA (siRNA) directed against a *MUT* transcript (MUT siRNA). The *MUT* absence did not affect the cell viability and apoptotic process in SH-SY5Y (Costanzo *et al.*, 2018).

1.11. NGS technologies for gene discovery

The identification of somatic or germline DNA variants with important links to cancer, neurobiological disorders and other complicated diseases, has been for years the focus of molecular scientists and translational investigators. Quantitative PCR, Sanger sequencing (capillary electrophoresis sequencing), and microarray technology all have critical roles in genetics. In parallel, next-generation sequencing (NGS) has been revolutionizing biological sciences, for which its output has only been doubling every year since the issue of HiSeq X® Ten in 2014. Not only has it quickly revolutionized science, but its quick drop in costs made NGS more accessible and central to various questions and answers made by scientists. With NGS, it is possible to examine complete human genome through one experiment, as well as sequencing up to tens of thousands of genomes a year. While the drop in price for whole-genome sequencing (WGS) is a thrilling transformation for science, the reduction in cost and the increasing simplicity of other NGS methods, such as targeted resequencing, have made the benefits of NGS available to the wider research community. With a development rate on the rise, targeted resequencing is proving to be a powerful tool for somatic and germline variant findings.

Furthermore, targeted sequencing allows a sequencing panel, a specific set of genes or targeted region, are isolated and enriched. This allows NGS to be more efficient and cost-effective. With targeted sequencing, various advantages come in play, such as deep sequencing (sequencing at much higher coverage levels), which allows greater confidence over Sanger sequencing for calling variants or low-frequency alleles in a given region of interest (Rivas *et al.*, 2011),(Jamuar *et al.*, 2014). If speed is required, targeted resequencing can also provide fast processing, due to a higher multiplexing capacity, lower data analysis requirements, and the capacity to sequence tens to thousands of targets in a single experiment. Targeted resequencing such a whole exome sequencing can reveal variants, such as low-frequency variants that would be more expensive or more challenging to identify with PCR or Sanger sequencing. The ability to detect low-frequency variants can enable identification of novel functional variants, facilitate biomarker discovery, or lead to the identification of clinically relevant targets for translational research. Whole exome sequencing is particularly useful for the discovery of somatic mutations in complex samples such as cancerous tumors mixed with germline DNA. Researchers can focus on regions of the genome most relevant to their interests, that being for cancer studies, microbial genomics, agrogenomics or molecular epidemiology.

Sanger sequencing (∼1 kb sequence reads) and Roche 454 sequencing (up to 800 bp) have been not as frequently used. Moreover, there has been a clear trend over the years towards short read technologies such as Illumina HiSeq (at present typically 150 bp) and SOLiD (typically 50 bp). Whole-genome sequencing, particularly of long-insert size libraries, requires high-quality, intact, non-degraded DNA at a sufficient amount (Takeichi *et al.*, 2015b). Whole-genome sequencing has been made accessible using the HiSeq X Ten System, with incredible speed and throughput. A comprehensive catalog of human variation, forge population-based references, significant discoveries, and advances in the understanding of biology and human genetic disease has all been made possible using the HiSeq X Ten System. (Illumina Inc. 2016). According to Cirulli & Goldstein, 2010, whole-genome sequencing will provide the best means of identifying rare causal variants by proposing the resequencing the genomes of individuals with extreme phenotypes and resequencing the genomes of individuals with a familial disease. They also propose that genome sequencing will identify rare variants with large effects on many diseases and traits in the coming years. The knowledge that could potentially be gained about these traits, such as the type of mutation and the gene that influences each trait, could provide information for new drug targets.

RNA-Seq offers many advantages over previous methods such as qPCR and gene expression (GEX) arrays, it can detect both known and novel findings, enabling analysis of the transcriptome without the limitation of prior knowledge, unlike both qPCR and GEX arrays.

The study of the transcriptome has been revolutionized using NGS technologies. Simpler and easier workflows have been achieved with the advances in transcriptome studies, from the steps of library preparation to data analysis with highly accurate results. The aim of utilizing RNA-seq with genome sequencing in this research project was to increase the diagnostic yield among the undiagnosed cohort of patients, confirm any pathogenic findings and explore their effect at the level of the RNA, and detect any missed variants by genome sequencing.

1.12. World effort for rare disease discovery

Over the past decades, there has been increased interest in rare disease, and more light has been shed on the genes underlying many rare genetic diseases. Many organizations around the world have put in extra effort to identify these genes. FORGE and Care4rare have been the major Canadian leaders in this area.

1.12.1 FORGE Canada

Finding of Rare Genetic Disease Genes (FORGE) in Canada is a program developed by Genome Canada to assist in gene discovery of rare pediatric disorders by creating a network of Canadian doctors and scientists and giving them access to NGS technology for their patients. The aim of this national collaboration was to be able to rapidly identify many genes responsible for genetic disorders that affect children. The Canadian Pediatric Genetic Disorders Sequencing (CPGDS) Consortium has had 150 members and brought together doctors from all genetics centers across Canada, internationally-recognized Canadian scientists with expertise in finding genes, and teams from the three Genome Canada Science and Technology (GC S&T) Innovation Centers (Montreal, Toronto, Vancouver), which have already set up the new sequencing technology. The CPGDS Consortium assisted doctors to identify patients with rare pediatric diseases. The Consortium has had members from all the medical genetic clinics in Canada. Therefore, for any given disorder, children and families from across Canada were enrolled into the program. Even for very rare conditions, disease-causing genes will be found. The goal of the Consortium included genome sequencing of patients to identify disease-causing genetic changes, that has allowed the setup of a national data coordination center to streamline and improve existing large-scale sequence analysis tools. This has improved the ability to distinguish genetic changes that cause disease from ones that are normal variants. (www.Care4rare.ca)

The consortium has allowed for rapid gene discovery of rare childhood-onset disorders, with immediate and long-term health benefits for Canadian families. The discoveries have led to genetic tests that allowed earlier and more precise diagnoses. The improvement of diagnoses allowed Canadian health care teams to reduce or prevent patient complications, develop tailored treatment. Of the 264 disorders studied during the FORGE project, definitively disease-causing mutations were identified to explain 120 disorders, and for 26 disorders, highly likely disease-causing variants were identified in novel candidate genes; 118 remain unexplained. These 146 disorders represented 67 novel genes and 95 known genes. (Beaulieu *et al.*, 2014). This data represents the added value of sequencing technologies in contrast to using gene panels or exome sequencing alone.

1.12.2 Care4Rare

After the foundation of FORGE and the discovery of rare genetic disorders, the need to manage and care for these disorders was imminent. Therefore, in 2013 Care4Rare was established by the same founders of FORGE as a more developed continuation of the project. There are about 7,000 rare genetic diseases in Canada that impact more than one million Canadians and their families. Two thirds of these diseases cause significant disability, three quarters affect children, more than half lead to early death and almost all have no targeted treatment. Further more, more than third of these diseases remain unsolved genetically.

Care4Rare is a pan-Canadian collaborative team of clinicians, bioinformaticians, scientists, and researchers, focused on improving the care of rare disease patients in Canada and around the world. This program is led out of the Children's Hospital of Eastern Ontario (CHEO) Research Institute in Ottawa, Canada. Care4Rare includes 21 academic sites across the country, and is recognized internationally as a pioneer in the field of genomics and personalized medicine. The use of state-of-the-art genetic technology such as sequencing has helped identify new rare disease genes for patients. More than 200 physicians and 100 scientists work together to advance rare disease research as part of three Care4Rare programs: C4R – SOLVE, Genomics4RD, and RareConnect. (www.care4rare.ca)

1.13. Hypothesis and Objective of Research Project

The aim of this thesis is to apply state of the art Next Generation Sequencing (NGS) technologies to investigate genetic cause in a cohort of patients with isolated elevated methylmalonic acid (MMA) in which no diagnosis could be confirmed by somatic cell complementation studies and by testing with a specialized Cbl gene panel. This cohort of patients presented with elevated levels of MMA in blood and/or urine and cultured fibroblasts from patient cell lines were submitted to the vitamin B_{12} diagnostic laboratory at McGill. In addition to elevated MMA, they presented with a wide array of clinical disorders, ranging from a mild clinical picture to a very severe one, including an unexplained sudden death in one patient. Any patient with a suspected Cbl disorder undergoes functional cell studies in the laboratory in an attempt to achieve a diagnosis. In some cases, a definite diagnosis has not been possible. Over the years, there has been an accumulation of such patients. In order to select the best candidates for this research project, we examined the [^{14}C] propionate incorporation levels without hydroxy into macromolecules in this cohort and selected 26 patients that had incorporation levels lower than 5.57nmol/mg protein/18 h (Reference range: 10.8 ± 3.7nmol/mg protein/18 h). This functional test indirectly assesses the activity of MCM, a key enzyme in intracellular Cbl processing inside the mitochondrion. When MCM malfunctions, the conversion of methylmalonyl-coA to succinyl-coA does not occur and leads to elevations of MMA in blood and/or urine. In such cases, a suspicion of a Cbl disorder is made and further somatic cell studies and genetic testing is done to attain a diagnosis. The [^{14}C] propionate incorporation without hydroxy test is performed by measuring incorporation of label from [^{14}C] propionate into trichloroacetic acid-precipitable material in cultured fibroblasts. These 26 patients reflected the lowest levels of [^{14}C] propionate incorporation without hydroxy test, however, following complementation analysis and Cbl distributions we could not obtain a biochemical diagnosis. These patients cell lines were further tested using the Cbl gene panel at Baylor Miraca Extended Genetics that includes 24 Cbl related genes and yet no genetic diagnosis was found. An additional patient has been added for testing using WGS and RNA-sequencing because of elevated MMA and low propionate incorporation levels (5.37 nmol/mg protein/18h). This patient however, has not been tested previously with the extended Cbl gene panel. Thanks to the pronounced advances in NGS technologies, with now a respectable decrease in sequencing costs and a rapid time frame, it was feasible to perform WGS and RNA-Sequencing on these patients and analyze the results, aiming to find novel genes and/or causal variants that could have been missed by classic diagnostic methods.

CHAPTER 2

Materials and Methods

2.1 Patient Selection criteria:

The selection process for the research cohort started with a larger cohort of 222 patient cell lines. The patient's lines have been collected over many years, and some of the cell lines are over twenty years old. The patients had variable degrees of elevation of MMA, in addition to other variable clinical symptoms. The cohort is very heterogeneous in terms of clinical phenotype and disease severity; some patients are mild and while others clinical presentation is more severe. The elevated MMA was either detected by new born screening or as part of a diagnostic work-up later on in life. Then they were referred to the laboratory for somatic cell studies to obtain a diagnosis. 188 of these patients were successfully sequenced for the twenty genes in the Cbl metabolism panel and severe MTHFR deficiency by Massively Parallel Sequencing at Baylor Miraca genetics as part of previous research projects. (Pupavac *et al.*, 2016) Out of this group, 61 patients received a genetic diagnosis, leaving a remainder of 127 patients yet without a genetic cause identified.

An application of a systematic approach was performed by using a selection criterion to identify strongest potential candidates out of the 127 patients. We used [C^{14}]-propionate incorporation w/o OH levels of less than 5.5 nmol/mg protein/18 h as the cut off value for selection. (Reference range: 10.8 ± 3.7nmol/mg protein/18 h). This test is run in the laboratory and measures the decreased incorporation of label from propionate into acidperceptible material, this reflects decreased MCM function. As mentioned previously, MCM is a key enzyme in intracellular Cbl metabolism and when not functioning properly, leads to elevations of MMA. Patients were then stratified based on the levels of propionate incorporation, proposing that patients with the lowest values are more likely to be associated with a Cbl disorder.

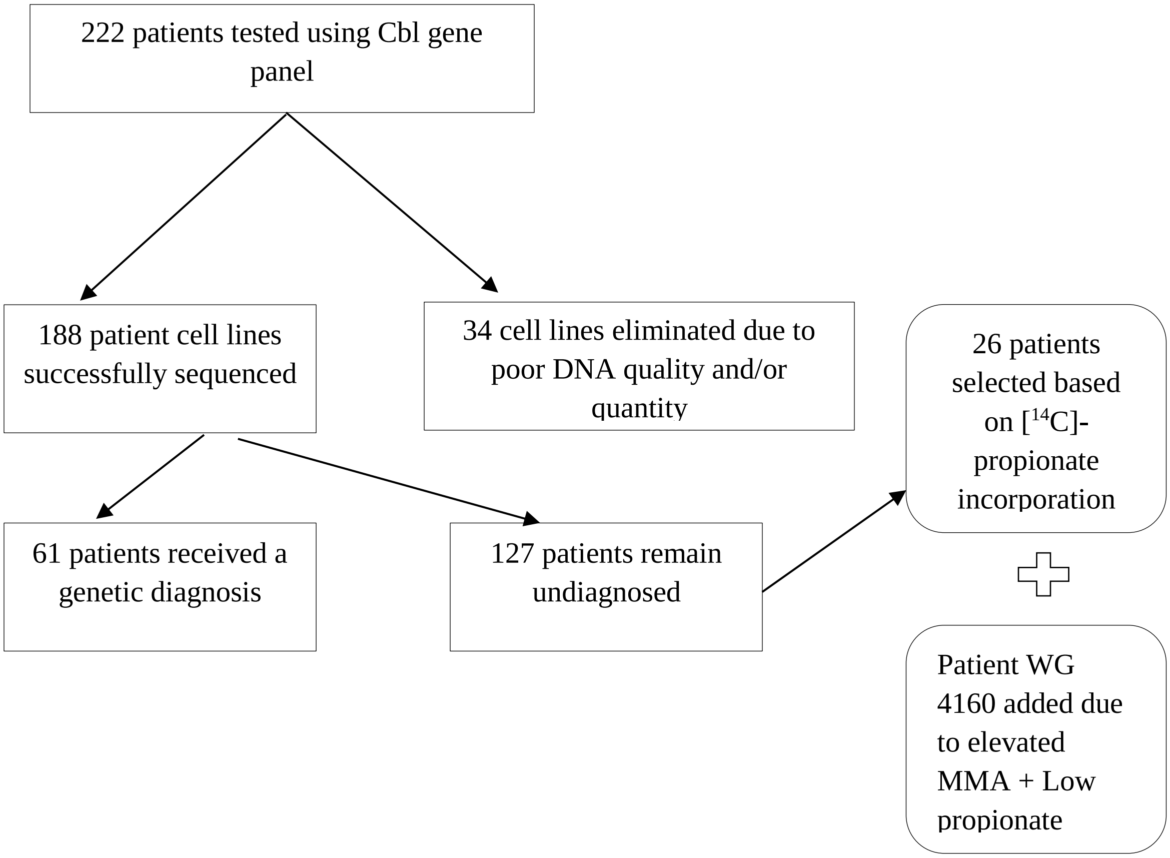

222 patients tested using Cbl gene panel
188 patient cell lines successfully sequenced
34 cell lines eliminated due to poor DNA quality and/or quantity
26 patients selected based on [^{14}C]-propionate incorporation
61 patients received a genetic diagnosis
127 patients remain undiagnosed
Patient WG 4160 added due to elevated MMA + Low propionate

Figure 4. **Patient selection process**. 222 patient cell lines were originally sent to be tested with a Cbl gene panel. Out of the 127 undiagnosed patients, 26 patients were selected due to low propionate incorporation levels. Patient WG 4160 was added to the study due to elevated MMA and low propionate incorporation levels; this patient was not tested by the Baylor panel.

2.2 Patients' clinical description:

2.2.1 Mild decrease in propionate incorporation levels:

This category includes 12 patients in which their propionate incorporation levels without hydroxycobalamin (w/o OH) were ranging from **3.7-5.5** nmol/mg protein/18 h. The youngest patient in this category is 6 days old that had an acute life-threatening event at home and eventually died, testing then showed that this patient has elevated levels of MMA. The oldest patient in this category is 17 months old, that had mild developmental delay, low plasma carnitine, and seizure like episodes. The clinical symptoms of all the patients in this group vary from mild to severe, with signs of developmental delay, failure to thrive, seizure like activity, or acidosis. Further detailed explanation of each patient in this category is explained in Table 1.

An additional patient, WG 4160 that had not been tested by the Cbl gene panel at Baylor was added to the study due to elevated MMA and low propionate incorporation level of 5.3 nmol/mg protein/18h. WG 4160 came to medical attention to rule out a Cbl disorder. This patient had acute irreversible cardiomyopathy and high C3- carnitines in addition to the elevated MMA. This patient had been tested on an extensive panel for dilated cardiomyopathy-related genes that did not include *TTN*, as the gene had not yet been discovered at the time.

Table 1. Clinical information for the 12 patients in which propionate incorporation levels w/o OH were ranging from 3.7-5.5 nmol/mg protein/18 h.

WG	Sex	Age	Clinical Findings
2316	F	6 days	Five-day old female infant with an acute life-threatening event at home eventually expired. Central nervous system and liver findings.
2368	F	7.5 mo	MMA diagnosis. Stiffness of the body, rolling of the eyeballs, and also jerking movements of the upper and lower extremities initially at six weeks of age. Several episodes of seizure activity. Shallow breathing, cyanosis and tachycardia.
2389	F	6.5 mo	Stiffness on left side; failure to thrive, dental delay, dysmorphic. Elevated MMA.
2701	F	7 mo	Methylmalonic aciduria. FFT, high MMA on quantitative AA.
2731	N/A	N/A	Elevated MMA. Autism. The family are vegans. No hematological abnormalities.
2740	M	6.5 mo	Failure to thrive. Gall stones. Mild elevation of ammonia. Modest increase of MMA and EMA on urine organic acids.
3092	M	9.5 mo	Elevated methylmalonic acid, blood and urine. Hypotonia, gross motor delays - onset in infancy. Mixed developmental delay, mild dysmorphic features.
3221	F	48 days	History of elevated MMA. History of slow weight gain. History of irritability, gastroesophageal reflux, and failure to thrive.
3357	M	17 months	Mild developmental delay, low plasma carnitine, seizure-like episodes.
4131	M	1 mo	Elevated glycine, elevated C3 on NBS, elevated MMA.
4142	M	11 mo	Persistent elevated MMA on NBS.
4190	M	4 mo	Chronic diarrhea, failure to thrive, metabolic acidosis, and macrocytic anemia

WG is the identification number assigned to each patient sample received, followed by sex, age in months/days, ethnicity and clinical data that was documented on each patient's file. FFT: failure to thrive, NBS: new born screening, EMA: ethylmalonic acid.

2.2.2 Moderate decrease in propionate incorporation levels:

This category includes 9 patients in which propionate incorporation w/o OH ranged from **2.7- 3.7** nmol/mg protein/18 h. The youngest patient in this category is 5 months old that reported mild methylmalonic aciduria, possibly B_{12} responsive. This apparently healthy infant had a metabolic evaluation because of the sudden unexplained death of his brother at 16 months of age. He has unexplained elevations of methylmalonic acid in serum and methylmalonate on urine organic acid analysis. The oldest patient in this category is a 13-year-old who had moderate mental retardation of unknown etiology and nonspecific neonatal problems. Further clinical information about each patient in this category is explained in Table 2.

Table 2. Clinical information of the 9 patients in which their propionate incorporation levels w/o OH were ranging from 2.7-3.7 nmol/mg protein/18 h.

WG	Sex	Age	Clinical picture
2575	M	7 mo	Failure to thrive, weight loss at 4 months of age. Hypotonia noted 2 months.
2716	M	19 mo	Hypotonia and recurrent illnesses associated with an acetone-like odor to his breath. He has had small amounts of methylmalonic acid in his urine on a couple of occasions. Low muscle tone. Doing well.
2718	F	1 year	MMA variant. Dysgenesis corpus collosum, developmental delay. Sudden collapse with acute infection at age 1 year with elevated lactate. Severe developmental problems. Neutropenia. Acute lactic acidemia.
2727	F	11 mo	Abnormal methylmalonic acid excretion. Sleeping a lot (had to be woken up for feeding). After 1mg subcutaneous implementation of B_{12} for 5 days, patient was much more active and alert.
2837	M	5 mo	Mild methylmalonic aciduria, possibly B_{12} responsive. Apparently healthy infant who had a metabolic evaluation because of the sudden unexplained death of his brother at 16 months of age. He has unexplained elevations of methylmalonic acid in serum and methylmalonate on urine organic acid analysis.
3086	M	9 mo	Probable benign MMA. Developmental delay. Seizure and developmental delay.
2686	M	7 mo	Seizures. Mild lactic acidosis, elevated MMA at 45 to 70 uM (plasma) and elevated urine MMA at 3-12 mM [normal 0 - 0.4]
2823	F	13 mo	Developmental delay. Hypotonia, failure to thrive, methylmalonic aciduria. Jaundiced during newborn period. Colicky - improved after mother removed lactose from her own diet. Hypotonic. Unusual movements including scissoring of her legs and fisting of hands.
2324	F	13 years	Moderate mental retardation. Unknown etiology. History of nonspecific neonatal problems. Organic acids were normal.

WG is the identification number assigned to each patient sample received, followed by sex, age in months/days, ethnicity and clinical data that was documented on each patient's file. FFT: failure to thrive, NBS: newborn screening.

2.2.3 Severe decrease in propionate incorporation levels:

This category includes 5 patients in which their propionate incorporation w/o OH were less than **2.7** nmol/mg protein/18 h, which are the lowest levels reported. The youngest patient is 4 months old had developmental delay, movement disorder, infantile spasms and elevated MMA in urine. The oldest patient was 9 months old and reported MMA, normal vitamin B^{12} levels and joint laxity. She also had gastroesophageal reflux and presented with an acute life-threatening respiratory event at 3 months of age.

Table 3. Clinical information of the 5 patients in which their propionate incorporation levels w/o OH were less than 2.7 nmol/mg protein/18 h.

WG	Sex	Age	Clinical picture
2436	F	8 mo	Myoclonic seizures, abnormal EEG which is intermittent with developmental delay.
2625	M	8 mo	B_{12} responsive methylmalonic acidemia. Failure to feed well for 1st month, then hyperammonemia, encephalopathy responsive to dialysis. Had neutropenia, thrombocytopenia and acidosis. Propionic acidemia. First urine contained small amount of methylmalonic acid in addition to the metabolites indicative of propionic acidemia.
3099	F	9 mo	MMA. Normal vitamin B_{12} levels and joint laxity. She has GE reflux and presented with an acute life-threatening respiratory event at 3 months of age.
3162	M	7 mo	Spinal cord inflammation NOS; elevated MMA. Minor motor vehicle accident, subsequently developed pneumonia and swelling around spinal cord. Work-up revealed elevated methylmalonic acid 901nmol/L (nl 73-271), repeat 281 ng/mL (nl 0 - 105)
3023	F	4 mo	Developmental delay, movement disorder. Infantile spasms. Elevated MMA in urine.

WG is the identification number assigned to each patient sample received, followed by sex, age in months/days, ethnicity and clinical data that was documented on each patient's file.GE: gastroesophageal, EEG: electroencephalogram, Spinal cord inflammation NOS: unspecified site of spinal cord injury without evidence of spinal bone injury, nl: normal.

2.3 Patients and cell culture:

All patient cell lines had been sent to the Vitamin B_{12} Clinical Research Laboratory (Department of Medical Genetics, McGill University Health Centre) to rule out an inborn error of Cbl metabolism. These samples are obtained by skin punch biopsy sent to the laboratory from multiple countries all over the world, there is a submission form and a consent form required for each cell line. Mycoplasma-negative primary fibroblast lines derived from patients were obtained from the cell bank at the McGill University Health Centre. Frozen vials of patients' fibroblasts were thawed to room temperature and grown in T75 flasks. Cells were cultured in Eagle's minimum essential medium (EMEM) containing Earle's salts, L-glutamine, and high glucose (Wisent Inc, Saint-Jean-Baptiste, Quebec), supplemented with non-essential amino acids, 0.11g/L sodium pyruvate, 0.01g/L ferric nitrate, 5% fetal bovine serum (Wisent Inc, Saint-Jean-Baptiste, Quebec), and 5% bovine calf serum (Wisent Inc, Saint-Jean-Baptiste, Quebec). These cultures of the patients' cell lines were routinely fed twice a week and incubated at 37°C in 5% CO_2.

2.4 Cell Harvesting

After maintaining cell cultures, cell harvesting was done to later on extract genomic DNA and RNA. All solutions and equipment that come in contact with the cells must be sterile. The use of proper sterile technique was ensured and work was performed in a laminar flow hood. The spent cell culture media was removed and discarded from the culture vessels, then cells were washed using 5 ml of PBS solution. The wash solution was gently added to the side of the vessel opposite the attached cell layer to avoid disturbing the cell layer, and the vessel was rocked back and forth several times. The wash step removes any traces of serum, calcium, and magnesium that would inhibit the action of the dissociation reagent. The wash solution is then removed and discarded from the culture vessel. This is then followed by adding 2 ml of the pre-warmed dissociation reagent, trypsin to the side of the flask. Enough reagent was used to cover the cell layer (approximately 0.5 mL per 10 cm^2). The container is gently rocked to get complete coverage of the cell layer. The culture vessel is then incubated at 37° C in 5 % CO^2 for 8 minutes. The cells are then visualized under the microscope for detachment, if cells are less than 90% detached, the incubation time is increased a few more minutes. When $\geq$ 90% of the cells have detached, 8 ml of the pre-warmed EDTA growth medium were added into the culture vessel and repeated resuspension is performed by pipetting over the cell layer surface several times to ensure mixing of the detached cells with the culture medium. The cells are then transferred to a 15-mL conical tube and centrifuge then at 200 × g for 5 minutes. The supernatant is discarded and the cell pellet is then stored for use later on.

2.5 DNA extraction

Cell pellets that were prepared during cell harvesting were stored at -20°C or immediately used for genomic DNA extraction. DNA extraction was completed from patient fibroblasts using the FlexiGene DNA kit (Qiagen, Canada) according the manufacturer's instructions. Concentration and purity of DNA were assessed using the BioDrop µLITE spectrophotometer and extracts were stored at -20°C. An additional step to ensure sufficient DNA quantity and quality was performed using the Qubit Fluorometer (Thermo Fischer).

2.6 Qubit Fluorometer

The Invitrogen Qubit 4 Fluorometer is designed to quickly and specifically quantitate DNA. The Qubit assays utilize target-selective dyes that emit fluorescence when bound to DNA, RNA or protein. Unlike UV absorbance, which can overestimate sample concentrations due to contaminants in the sample such as salts, solvents, detergents, proteins, free nucleotides. Qubit fluorescence is also much more sensitive than UV absorbance, and the system is able to accurately measure dilute samples with significantly less noise. (Thermo Fischer).

2.7 RNA extraction

Cell lines for RNA sequencing studies were grown in T75 tissue culture flasks until confluent. Once cells were inspected and looked at optimal conditions, they were harvested, pelleted and dissolved in 1mL of QIAzol Lysis Reagent and immediately frozen at -80ºC. The suspensions were used to extract total RNA using the miRNAeasy Mini Kit (Qiagen). Quality was assessed using the BioDrop µLITE spectrophotometer and quantified. The miRNeasy Micro Kit combines phenol/guanidine-based lysis of samples and silica-membrane–based purification of total RNA. QIAzol Lysis Reagent, included in the kit, is a monophasic solution of phenol and guanidine thiocyanate, designed to facilitate lysis of tissues, to inhibit RNases, and also to remove most of the cellular DNA and proteins from the lysate by organic extraction. Cells or tissue samples are homogenized in QIAzol Lysis Reagent. After addition of chloroform, the homogenate is separated into aqueous and organic phases by centrifugation. RNA partitions to the upper, aqueous phase, while DNA partitions to the interphase and proteins to the lower, organic phase or the interphase. The upper, aqueous phase is extracted, and ethanol is added to provide appropriate binding conditions for all RNA molecules from approximately 18 nucleotides (nt) upwards. The sample is then applied to the RNeasy MinElute spin column, where the total RNA binds to the membrane and phenol and other contaminants are efficiently washed away. High-quality RNA is then eluted in a small volume of RNase-free water.

2.8 Whole genome sequencing

WGS was performed on 500 ng to 1.5 μg of genomic DNA using a PCR-free protocol. These libraries were sequenced on the Illumina HiSeq X Ten sequencers with 151-bp paired-end reads and a mean coverage of >30X, using the TruSeq PCR-Free DNA Kit and HiSeq X Reagent Kit v2.5. A workflow was applied to identify a spectrum of variant types, including single nucleotide variants (SNVs), indels and structural variants (SVs) (Li and Durbin, 2009). Pre-processed sequences were mapped to the reference genome (UCSC hg19) using the Burrow-Wheeler Aligner (BWA v.0.7.12) (Li and Durbin, 2009). Using the Picard tools v.2.1.0 duplicate reads are removed (http://broadinstitute.github.io/picard/). The Genome Analysis Toolkit (GATK v.3.5) Realigner Target Creator / IndelRealigner tools are used to remove artifacts in the vicinity of insertion/deletions known to occur in the initial alignment process (McKenna *et al.*, 2010). Following re-alignment, base quality scores are recalibrated using the GATK BaseRecalibrator tool and the GATK HaplotypeCaller to identify variant positions (DePristo *et al.*, 2011). The Variant Quality Score Recalibration (VQSR) process is used to further filter and output the most confident variant set. Finally, the variant set is annotated with population allele frequencies (derived from in-house controls, the 1000 Genomes database (Clarke *et al.*, 2012) EVS , ExAC v.0.3) (Karczewski *et al.*, 2017), conservation scores (GERP and PhastCons),(Siepel *et al.*, 2005)] and pathogenicity scores (SIFT (Kumar, Henikoff and Ng, 2009), PolyPhen2 (Adzhubei *et al.*, 2010)]and Combined Annotation Dependent Depletion [CADD v.1.3]) (Kircher *et al.*, 2014) using Variant Effect Predictor (VEP v.83) and in-house tools and control databases, to determine the effect of variants on genes, transcripts, and protein sequence, as well as regulatory regions. In addition to identifying SNVs and indels, PopSV (Monlong *et al.*, 2018) was used to detect SVs: deletions and duplication events. The SVs are curated by filtering most common SVs, by overlapping called SVs to previous in-house control SV database, and finally annotating with gene information. Those databases are invaluable in filtering out false positive and common results, as well as interpreting genetic variants in the light of available biological data. The concept was that candidate genes identified by WGS, with either a homozygous variant or compound heterozygous variants (or a hemizygous variant, for genes on the X chromosome) in one or more patients, will be selected for further study on the assumption that inheritance is autosomal recessive or X-linked, in line with other genes causing metabolic disorders. Priority will be given to genes that are identified in more than one patient, on the assumption that these are more likely to be genuine causal genes.

2.9 RNA sequencing

RNA was isolated from cell lysates using the miRNAeasy Mini Kit (Qiagen) and. Library preparation, library QC testing and RNA-seq was performed at the McGill University and Genome Quebec Innovation Center. The library preparation was completed using the NEB mRNA stranded library preparation kit according to the supplier's protocol, followed by QC testing using the Bioanalyzer. 1 µg of RNA was poly(A) selected, fragmented and reverse transcribed with the Elute, Prime and Fragment Mix (Illumina). End repair, A-tailing, adaptor ligation and library enrichment were performed as described in the Low Throughput protocol of the TruSeq RNA Sample Prep Guide (Illumina). RNA libraries were assessed for quality and quantity with the Agilent 2100 BioAnalyzer and the Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies). RNA libraries were sequenced as 100 bp paired-end runs on an Illumina HiSeq4000 platform. The resulting call sets were then analyzed according to certain strategies to complement the existing genome sequence data and add more diagnostic value to this cohort of patients. The strategies included were to detect monoallelic expression sites (MAE), detect any variants missed by WGS, and detect splicing defects. For the detection of MAE, the call sets of WGS and RNA-seq were compared to identify potential MAE sites. A potential MAE site would classify as (i)one when the read depth ≥10 in RNA- seq data, (ii)has a proportion of alternative allele reads in RNA sequencing more than the proportion of alternative allele reads in WGS, (iii) in RNA-seq data, the proportion of alternative allele reads ≥ 2/3. The criteria were intentionally designed to detect as many potential MAE sites as possible; heterozygous or homozygous genotypes that were called by GATK (Genome Analysis Toolkit) were ignored.

To explore the second strategy, the call sets of WGS and RNA-seq were compared to identify variants. WGS variants were filtered based on quality metrics and allele frequency as described in the variant filtering section for WGS data. In contrast, the full call set of RNA-seq data was used with the application of the same allele frequency filtering as for WGS (except that G5A tag is not present in the RNA sequencing vcf file). A ≥ 10x coverage was required and ≥3 alternative allele reads in RNA-seq data. For variants unique to RNA sequencing, variants were required to be observed in <5 samples, otherwise too common to be interesting. Additionally, the third strategy was the detection of splicing defects using leafcutter, a software used for analysis that will be explained in detail in segment 2.17.6. On average around 34 aberrant splicing events were detected per sample (defined as p.adj <=0.05), ranging from 2 to 255 events in individual samples.

2.10 Data analysis pipeline design

WGS and RNA-Seq were both completed at the McGill University and Genome Quebec Innovation Centre, and results were received in uBAM file format (unaligned, raw sequencing reads). These files were then processed through a customized data analysis pipeline that is further discussed in detail in the below segments, to create variant call set files that were filtered based on certain parameters for manual examination.

2.11 Application of the GATK best practices:

The Genome Analysis toolkit (GATK) was developed in the Data Sciences Platform at the Broad Institute. This toolkit offers an extensive selection of tools with a primary focus on variant discovery and genotyping. It is equipped with a powerful processing engine and high-performance computing features allowing it to process projects of any size. In terms of analyzing WGS and RNA-Seq raw reads, the GATK best practices has been one of the most reliable and successful systems to apply, hence it was ideal to use in this project. After the process of high throughput sequencing is complete, the sequences are received in the form of raw reads, uBAM files. These files are then run under a set of instructions and key principles in the data analysis application steps, this results in generating appropriately filtered variant callset.

2.12 Analysis Phase:

The best practices guidelines are tailored to different applications according to the research project requirements and end results. The following analysis phases are applied to achieve an efficient workflow:

2.13 Data Pre-processing:

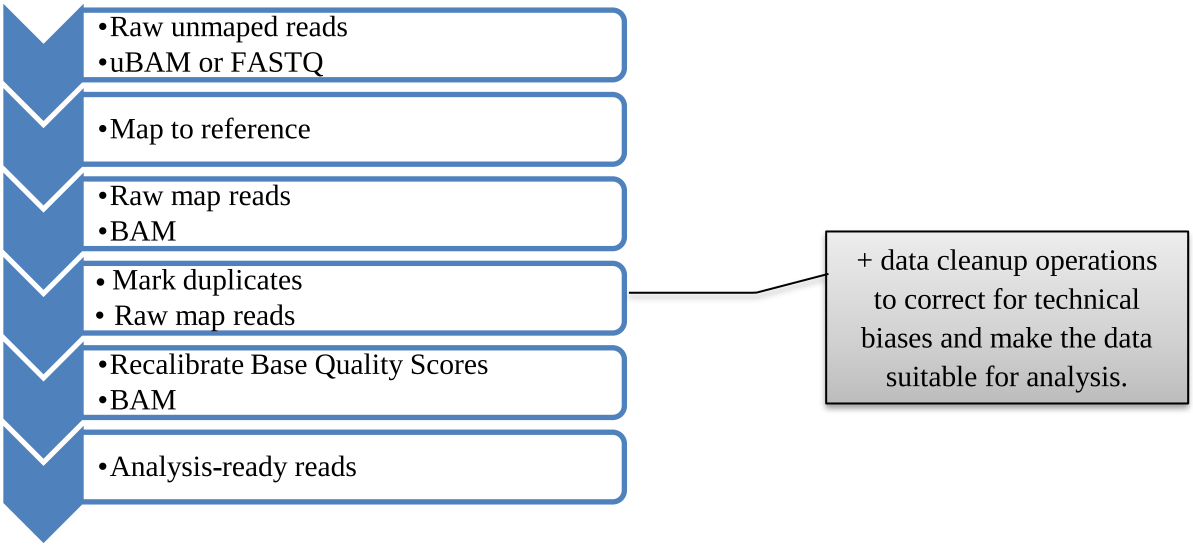
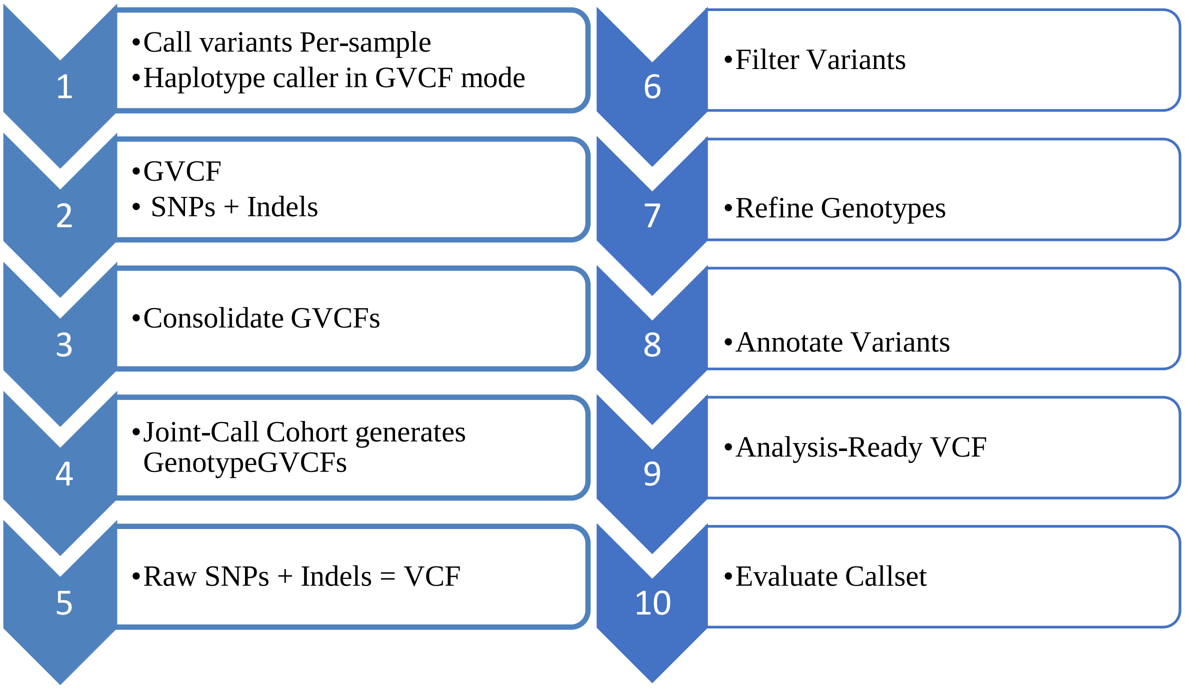

•Raw unmaped reads
•uBAM or FASTQ
•Map to reference
•Raw map reads
•BAM
• Mark duplicates
• Raw map reads
•Recalibrate Base Quality Scores
•BAM
•Analysis-ready reads
+ data cleanup operations to correct for technical biases and make the data suitable for analysis.

Figure 5. Outline of sequencing data pre-processing. This figure summarizes the steps involved in data pre-processing. Raw unmapped reads in the form of uBAM or FASTQ are mapped to the reference genome. The duplicated regions are then marked, and base quality scores are recalibrated. This then results in analysis ready reads, with additional steps that include data clean up and corrections for technical bias. These steps are adapted from the Broad Institute best practices guidelines. (https://software.broadinstitute.org/gatk/best-practices/workflow?id=11165)

2.14 Variant discovery:

This involves identifying genomic variation in one or more individuals in comparison to the reference genome and applying filtering methods appropriate to the experimental design. The output generated is in VCF format. The first step begins by calling variants per sample in order to produce a file in GVCF format. Then consolidate GVCFs from multiple samples into a Genomics DB datastore. This is followed by joint genotyping, and finally, applying filtering to produce the final multisample callset with the desired balance of precision and sensitivity.


1
- Call variants Per-sample
- Haplotype caller in GVCF mode

2
- GVCF
- SNPs + Indels

3
- Consolidate GVCFs

4
- Joint-Call Cohort generates GenotypeGVCFs

5
- Raw SNPs + Indels = VCF

6
- Filter Variants

7
- Refine Genotypes

8
- Annotate Variants

9
- Analysis-Ready VCF

10
- Evaluate Callset

Figure 6. Overview of identifying variants This figure shows the steps involved to call variants per sample using Haplotype caller in GVCF mode. (SNPs and Indels) in one or more individuals to produce a joint callset in VCF format. These steps are adapted from the Broad Institute best practices guidelines. (https://software.broadinstitute.org/gatk/best-practices/workflow?id=11145)

2.15 Call variants per-sample: HaplotypeCaller (in GVCF mode)

The HaplotypeCaller is one of the most popular variant caller that is capable of calling SNPs and indels concurrently via local de-novo assembly of haplotypes in an active region. Simply, whenever a region is showing signs of variation the program discards existing mapping information and completely reassembles the reads in that region. This is beneficial since it allows the program to more accurately call regions that are traditionally difficult to call, for example when they contain dissimilar types of variants in approximate regions. However, HaplotypeCaller may be very efficient at detecting rare variants, the execution time could be tremendous and very time consuming.

2.16 Data analysis

2.16.1 Processing of WGS data

BAM files received from the McGill University and Genome Quebec Innovation Center after sequencing were converted to FASTQ files using Sam2fastq tool. The mapping process was done using BWA MEM. BWA is a software package for mapping DNA sequences against a large reference genome, such as the human genome. The BWA-MEM algorithm is designed for longer Illumina sequences reads ranged from 70bp to a few megabases. BWA-MEM has features such that support long reads and chimeric alignment and is generally recommended as faster and more accurate tool. BWA-MEM also has better performance than BWA-backtrack for 70-100bp Illumina reads. We marked duplicate reads using Picard followed by recalibration of base quality scores. Variant calling was done by haplotypecaller and variant annotation by SnpEff. SnpEff is a variant annotation and effect prediction tool. It annotates and predicts the effects of variants on genes such as amino acid changes.

QC metrics were generated regarding mapping, duplicate regions and coverage by using samtools flagstat and bedtools genomecov.

Sample gender was inferred using WGS data, and compared against the gender information on our records. To infer sample gender, common SNVs were used (5-95% in gnomAD; in dbSNP) on X chromosome. The number and proportion of heterozygous and homozygous genotypes were calculated for each sample. The female samples were expected to have a lot more heterozygous genotypes, while the male samples were expected to be hemizygous at most SNV positions.

Pairwise IBD (identity-by-descent) estimation using plink (http://pngu.mgh.harvard.edu/purcell/plink/) (Purcell *et al.*, 2007) were performed to find pairs of individuals who look too similar to each other more than we would expect by chance in a random sample. This would help detect sample contaminations, swaps and duplications. A multidimensional scaling (MDS) analysis was also performed using plink. This is to identify a cluster of relatively homogenous samples within our cohort and detect any outlier.

To ensure that patient DNA and RNA data were identity-matched, we compared variants identified in WGS and RNA-seq data. The analysis was first limited to synonymous variants, which are in dbSNP (build 151) and >5% in gnomAD. For each sample, the number of such variants observed in WGS, the number of such variants observed in RNA-sequencing and the number of such variants which are present in both WGS and RNA sequencing were counted. The percentage of WGS variants captured by RNA sequencing, as well as the percentage of RNA sequencing variants captured by WGS was then calculated. This analysis was also done with common missense variants, rare synonymous variants, and rare missense variants.

2.16.2 Variant filtering

The process of variant filtering is applied to generate the most reliable variant call set that could lead to identifying causal and/or pathogenic variants. Variants (SNVs and INDELs) were filtered based on quality metrics, allele frequency, function prediction scores and clinical relevance of variants. More specifically, for *quality metrics*, a variant is required to satisfy all criteria that include ≥ 62.5% of the samples to have called genotypes, an allele count ≤ 5 among all samples, in the sample of interest ≥ 10x coverage and ≥3 alternative allele reads.

For allele frequency, a variant is required to satisfy all criteria that include; an overall allele frequency in gnomAD < 0.5%, African allele frequency in gnomAD < 1%, European (Non-Finnish) allele frequency in gnomAD < 1%, East Asian allele frequency in gnomAD < 1%, number of homozygotes in gnomAD < 5, and not more than 5% minor allele frequency in each and all populations.

For function prediction and clinical relevance of variants, a variant is required to satisfy at least one of the criteria that include; putative impact estimated by SnpEff to be HIGH or MODERATE, PHRED-scaled CADD score ≥ 15 ,ReMM score ≥ 0.9 ,dbscSNV scores equal to ada_score ≥ 0.6 or rf_score ≥ 0.6, SPIDEX scores: |dpsi_max_tissue| ≥5 or |dpsi_zscore| ≥2 ,SIFT score < 0.05 , POLYPHEN2 score > 0.15, and a clinical significance reported in ClinVar as pathogenic, likely pathogenic, risk factor, association, drug response, or uncertain significance. Variation is interrogated in a clinical diagnostic assay (CDA tag provided by dbSNP).

Briefly, CADD scores the deleteriousness of SNVs and indels across the entire genome; ReMM score predicts function of non-coding variants (SNVs and indels); dbscSNV scores and SPIDEX scores are for variants within splicing regions; SIFT score and POLYPHEN2 score are for non-synonymous variants.

2.16. 3 Gene ranking

Variant scores

Variants passing the filtering as described above, are scored based on their function prediction and clinical relevance. If there are multiple variants in one gene, these scores are then added to generate a final score.

Phenotype matching scores

Primary phenotype matching scores are based on matching disease phenotypes caused by a gene with the patient’s phenotype.

First, the annotation file was downloaded from HPO, which links genes to diseases and phenotypes (in HPO terms). Second, a similarity score was calculated between the patient’s phenotypes (in HPO terms) and a particular disease (annotated in HPO terms as well).

2.16.4 CNV analysis

A copy number variation (CNV) analysis was added to detect duplications or deletions as a method to discover any causality of genetic disease in our cohort. CNV generated through duplication or deletion events that affect one or more loci are widespread in the human genomes and are often associated with functional consequences that may include changes in gene expression levels or fusion of genes (Gorfine *et al.*, 2015) Genome-wide association studies indicate that some disease phenotypes and physiological pathways might be impacted by CNV in a small number of characterized genomic regions. PopSV is the CNV detection method used for this part of the analysis, this toolkit allows for structural variation detection from high-throughput sequencing. It uses abnormal read depth in comparison to population samples as signals.

Using this data, several different files were generated based on different parameters of CNV detection. CNVs based on a size larger than 200kb were first explored. This was followed by any CNVs that overlap with haploinsufficient gene, with a pLI score of greater than 0.9, proposing that these genes will not tolerate such aberrations. The pLI score is the probability that a given gene falls into the Haploinsufficient category, therefore is extremely intolerant of loss-of-function variation. Genes with high pLI scores (pLI **≥ 0.9**) are extremely LoF intolerant, whereby genes with low pLI scores (pLI ≤ 0.1) are LoF tolerant.

Furthermore, a list of CNVs that overlap with genes that harbor rare, potentially functional and/or pathogenic SNVs and/or indels was added and examined. And finally, one for homozygous deletions.

2.16.5 Detection of variants captured by RNA sequencing but missed by WGS

The call sets for each sample of WGS and RNA sequencing were compared to identify variants called by RNA sequencing but missed by WGS. For WGS, we filtered variants based on quality metrics and allele frequency, as described in the variant prioritization section. For RNA sequencing, the same allele frequency was applied for filtering as WGS; a ≥10x depth coverage was required and ≥3 alternative allele reads in RNA sequencing data. In addition, for variants unique to RNA sequencing, the variant was required to be a coding region variant, and observed in <5 samples in RNA sequencing data, otherwise this predicts the variant is too common to be interesting and more likely to be artifact.

2.16.6 Detection of splicing defects

The LeafCutter66 software was utilized to detect aberrant splicing. Each patient was tested against all others. To adjust LeafCutter to the rare disease setting, the parameters were modified to detect rare clusters, capture local gene fusion events and to detect junctions unique to a patient (minclureads ¼ 30; maxintronlen ¼ 500,000; mincluratio ¼ 1e-5) Furthermore, one sample was tested against all other samples (min_samples_per_group ¼ 1; min_samples_per_intron ¼ 1). The resulting P values were corrected for multiple testing using a family-wise error rate approach.

2.16.7 Variant inspection

After the intricate process of filtering variants using the described parameters, a variant call set was generated with the potentially best candidate variants. Since many filtering methods were applied, on three main types of analysis; genome, RNA seq and CNV, the generated results were each examined for each of these categories separately for each patient. Manual inspection of all generated lists was performed. Priority was given to detect any variants in Cbl related genes, if none potentially pathogenic variants were found, any variant that has been reported by ClinVar as pathogenic or likely pathogenic would be examined, by evaluating the gene and disease involved, MAF, and how relative is this disease to the patient’s phenotype. Followed by inspecting any variants detected that are in disease-associated genes according to the databases used such as OMIM and Orphanet, regardless of their reported pathogenicity. Finally, coding region variants in non-disease associated genes were also examined by exploring any interactions with Cbl disorder according to gene function.

Results

Chapter 3

3.1.1 Variant Review:

The lists of generated variant call sets were carefully examined according to the ACMG (American College of Medical Genetics) guideline for variant prioritization. CNVs were manually inspected using DECIPHER (Database of genomic variation and phenotype in humans using Ensembl resources) genome. A systematic approach was applied by starting with ClinVar significance, allele frequency databases, diseases association, and gene function.

3.1.1 a) Overview of variants passing filtering

Table 4 provides an overview of the large number of variants detected after applying the parameters for variant prioritization for each sample.

Table 4. Number and type of variants passing filtering

WG	Missense	Nonsense	Frameshift	Splice	Total coding
2316	351	10	21	48	572
2368	353	8	13	51	542
2389	471	14	21	60	752
2837	331	10	23	38	520
3162	359	9	23	40	561
3023	605	13	19	73	911
4131	500	16	35	79	857
4142	459	11	23	48	726

2625	802	13	42	132	1343
3099	348	6	21	51	552
2575	370	6	20	39	564
2324	389	11	30	45	620
2436	320	6	19	51	540
2701	412	8	112	62	758
4190	561	9	30	80	896
3357	386	9	26	50	619
2686	347	11	14	44	552
3086	324	14	25	34	530
2727	386	5	20	39	585
2718	662	17	38	71	1058
3092	314	3	17	55	545
2740	346	10	27	34	550
2716	360	13	17	52	593
2823	519	14	31	68	817
2731	333	11	16	43	523
3221	427	10	26	43	679

Total coding variants includes missense, nonsense, frameshift, synonymous, splice, inframe deletion/insertion and a few other variant categories. Splice variants as defined here refer to sequence variants in which a change has occurred within the region of the splice site, within 1-3 bases of the exon or 1-8 bases of the intron.

3.1.1 b) Inspecting variants passing filtering in four categories:

Variants passing filtering were inspected in four categories, to find potentially causal ones, meaning according to variant classification guidelines by the ACMG are considered pathogenic or likely pathogenic. Candidate genes were classified as any detected variants in known Cbl genes. Variants that were detected previously in ClinVar with supporting evidence are detected and gathered under the reported column. Variants that have been previously detected in OMIM or Orphanet and were associated with disease were reported under the disease associated column. Coding region variants that cause a protein change but have not been associated previously with any disease, were reported as non-disease associated coding genes. Variants that were reported to be pathogenic, or likely pathogenic, with a very low minor allele frequency, or have never been seen before in gnomAD would make the best candidates. Specifically, if these variants were reported in disease causing genes in which the disease corresponds to the patient's phenotype. Or, if they were coding region but non-disease-causing variants and the variant is predicted to have strong evidence of pathogenicity according to ACMG guidelines.

Table 5. Number of variants reported in each category

WG	Candidate Genes	Reported by ClinVar	Disease-associated Genes	Non-disease associated Genes
2316	3	91	441	456

2368	2	69	410	431
2389	2	78	605	594
2837	2	67	434	405
3162	1	68	455	441
3023	4	85	755	745
4131	1	79	766	687
4142	3	75	599	586
2625	9	124	1191	1088
3099	0	69	465	438
2575	2	69	421	466
2324	1	60	420	499
2436	2	61	407	443
2701	2	65	544	595
4190	9	107	740	707
3357	3	78	442	487
2686	2	69	455	448
3086	0	75	448	426
2727	1	57	427	472
2718	5	73	887	869
3092	2	61	393	443
2740	3	71	473	426
2716	2	58	449	478
2823	5	80	588	675

2731	2	72	389	422
3221	1	79	515	560

Each sample had the corresponding number of variants detected in each category, in which all were carefully examined to find potential candidate variants. Variants under **Disease-associated genes** include both coding and non-coding region variants. Variants under **Non-disease associated genes** include only coding region variants.

3.2 Whole Genome Sequencing Results

In the 26 patients that had been studied on the gene panel, we confirmed nine heterozygous variants of unknown significance that were previously detected by the gene panel in known Cbl genes; however, we did not identify a second mutation in any affected patient. These results are in Table 6. No novel putative pathogenic variants were discovered in any known Cbl genes. In four patients, candidate variants that explain the non-metabolic clinical phenotype were identified in other genes; these will be in discussed in detail in section 3.2.

Table 6. Patients with heterozygous variants previously reported by cobalamin gene panel.

WG#	Sex	Age at biopsy (years)	Gene	Mutation Call	Amino Acid Change	Propionate uptake w/o OHCbl
2436	F	8 mo	GIF	c.435_437delGAA	p.K145_N146delinsN	2.7
2625	M	8 mo	CUBN	c.9986C>CT	p.S3329SL	0.77
2701	F	7 mo	MTHFR	c.1333C>CT	p.R445RW	5.2
2731			ACSF3	c.854C>CT	p.P285PL	4.20
2837	M	5 mo	ACSF3	c.1672C>CT	p.R558RW	3.2
3086	M	9 mo	CUBN	c.2594G>AG(p.S865NS)	p.S865NS	3

			CUBN	c.6469A>AG(p.N2157ND)	p.N2157ND	
3357	M	17 mo	HCFC1	c.4442C>T(p.T1481M)	p.T1481M	4.8
4131	M	1 mo	CD320	c.262_264delGAG	p.E88del	5.5
			CD320	c.658G>AG	p.G220RG	
3023	F	4 mo	ChrX- HCFC1	c.4475C>CT	1492P>PL	0.97

DNA nucleotide +1 is the A of the ATG translation initiation codon in the reference sequence (MUT: NM_000255.3, CUBN: NM_001081.3). Prop. Inc. = [^{14}C]-labeled propionate incorporation without OHCbl; values are given in nmol/mg protein/18h.

3.3. Patients with identified findings:

3.3.1WG2625

WGS analysis revealed this patient to be a compound heterozygote carrying one nonsense variant **c.1749_1750delGGinsTT** (NM_000282.3) causing a **stop-gained mutation**, leading to a stop codon and one intragenic duplication, a **9 kb** duplication that covers exon 21 of *PCCA.* This variant has not been previously seen in population databases, the duplication is detected in trans with the pathogenic variant, as well as, the patient's phenotype fits with the gene etiology. The 9kb duplication is shown in Figure 7. Using the sequencing results, exploration of the ratios of the variant allele in WGS and RNA-seq shows nonsense mediated decay. Detecting impact of the duplication at RNA-level has shown the duplication of exon 21 in *PCCA.*

Figure 7. PopSV breakpoint finder of CNV spanning the location of ch.13: 101,097,001 to 101,106,000 in sample WG2625

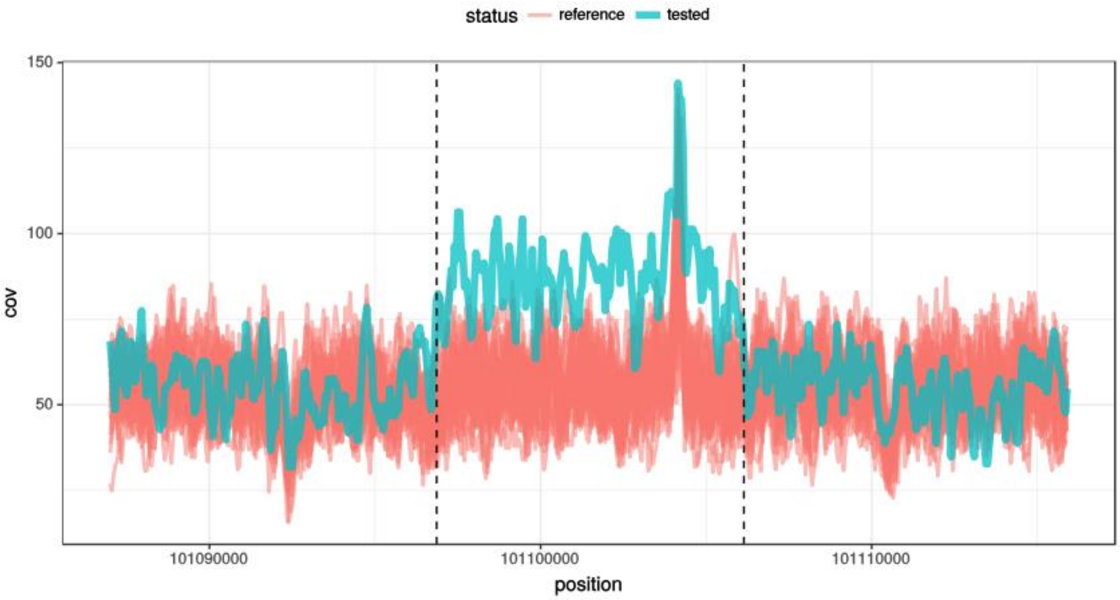


Figure 7. This computational method accurately identifies the structural variation event and delineates the breakpoints from the massive amounts of reads generated by the NGS experiment. The lines in red shows mapped reads from the reference genomes, in comparison to the one blue representing the reads from the WG2625 sample. Figure generated by Joel Lafond Lapalme.

3.3.2 WG 4190

WGS analysis revealed a homozygous frameshift mutation, **c.583dupC;p.Gln195fs**, in *EPCAM* that has been reported to be pathogenic by ClinVar. Since it is a homozygous frameshift mutation, this will drastically alter the coded amino acid, that classifies it as having very strong evidence of pathogenicity, as well as pathogenic variants in *EPCAM* have been previously reported in the same disease the patient presented with. RNA-seq was explored to evaluate the effect of the mutation, and unfortunately there are very few reads since *EPCAM* is almost unexpressed in fibroblasts, which is the cell line type used for this project. In Figure 9 *EPCAM* expression is further evaluated in different tissues.

Figure 8. Expression of *EPCAM* in different tissues

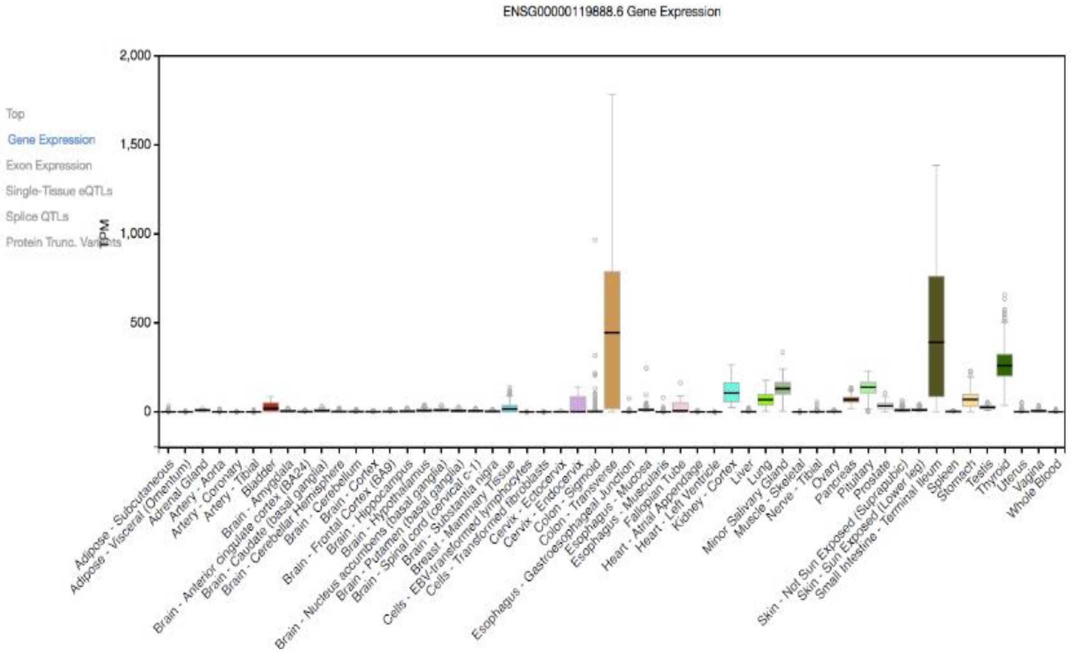


Figure 8. The Genotype-Tissue Expression (GTEx) tool was used to study tissue-specific gene expression. Expression of *EPCAM* is shown above in all tissues, such as skin, vessels, brain, blood, and gastrointestinal tract. It shows no expression in skin tissues, and samples used for sequencing were cultured fibroblasts, therefore we cannot detect any expression levels. *EPCAM* is almost solely expressed in the colon and terminal ileum, hence, examining the RNA-seq dataset would have been more useful using cells from colon biopsy. Figure generated by Sophie Ran Wang. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

3.3.3 WG 4160

This patient was added to the study cohort due to elevated levels of MMA and decreased propionate incorporation levels, and has not been tested using the extended Cbl gene panel. WGS analysis revealed a frameshift variant in *TTN* gene,

c.75138_75141delAGAA; **p.Lys25046fs**, that has been reported to be likely pathogenic by ClinVar, since it creates a premature termination of the protein. It has not been observed in gnomAD previously, hence the MAF is 0. The patient's phenotype fits with *TTN* being the causal gene. The variant is also found in a highly conserved region, suggesting that this variation is causing a change that is not well tolerated.

3.3.4 WG 2716

WGS and CNV analysis revealed a duplication, location 17q12 which corresponds to 17q12 duplication syndrome. The location of the duplicated region is shown in Figure 9.

Figure 9. PopSV breakpoint finder of CNV spanning the location of ch.17: 34,815,001-36,250,000 in sample WG2716

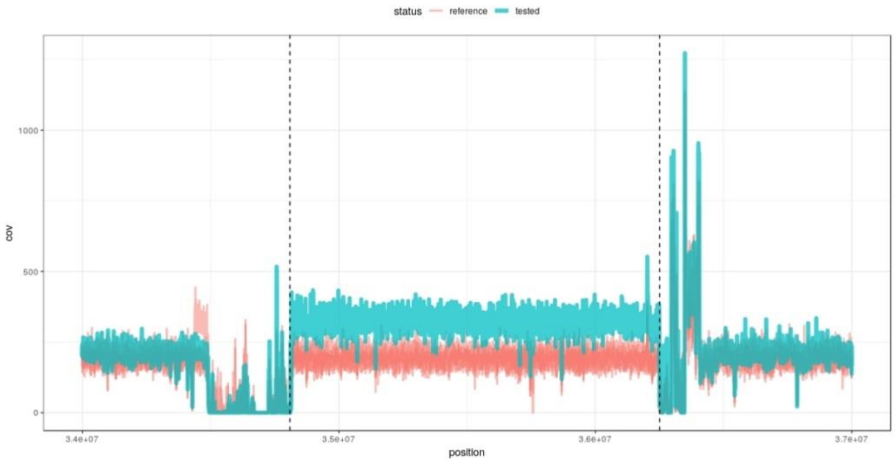


Figure 9. This computational method accurately identifies the structural variation event and delineates the breakpoints from the massive amounts of reads generated by the NGS experiment. The lines in red shows mapped reads from the reference samples, in comparison to the one blue representing the reads from the WG2716 sample, which show a duplicated number of reads compared to the reference genome. Figure generated by Joel Lafond Lapalme.

Discussion

Chapter 4

In this study, twenty-six patients who had been undiagnosed using somatic cell complementation analysis and testing using a panel comprising the 24 genes involved in Cbl metabolism and absorption were further investigated by WGS and RNA-seq .In 13 patients, no variants of interest were found. In nine patients, heterozygous variants of unknown significance (VUS) in Cbl genes that has been previously reported by the gene panel were confirmed. Of interest 3 of the 26 patients had novel findings, presumably unrelated to Cbl metabolism that could explain part of their reported phenotype. This diagnostic yield of 3/26 or 11.5 % is comparable to what is commonly seen for WGS studies. However, factors such as disease type, cohort size, and additional parental genetic information do have a better impact on diagnosis (Stavropoulos *et al.*, 2016). An additional novel finding was detected in WG 4160 that was added to the study.

4.1. Value of WGS, RNA-Seq and CNV analysis versus Conventional testing

Complementation analysis and gene panel analysis are both clinical tests that are used for achieving diagnosis of inherited Cbl disorders. In this study, the aim of applying NGS applications such as WGS and RNA-seq was to discover added diagnostic value of such methods specifically in undiagnosed patients. In this cohort of patients, elevated MMA and low [^{14}C]-labeled propionate incorporation levels are both considered biochemical features involved in Cbl disorders. Propionate incorporation has been shown to add value in determining the pathogenicity of sequencing findings by assessing whether the identified variants disturb functional enzymatic activity. Hence, the hypothesis is that, since conventional testing has not detected any genetic Cbl disorder, WGS and RNA-seq could lead to novel gene discovery and diagnosis. The clinical presentations of these patients were very variable and limited, and since the original medical files were sent many years ago it was difficult to receive any more medical information that could be beneficial in the analysis, specifically when using HPO terms. It has been shown that conventional methods of somatic cell studies and gene panel testing are sufficient for diagnosis of inborn errors of Cbl. It is proposed that the elevated MMA could be an incidental finding in these patients, or yet could be elevated due to reasons nonrelated to Cbl disorders. It could be possible that the MMA levels were temporarily elevated and not permanently, therefore eliminating a Cbl related cause. For future studies, this could be greatly taken into consideration. However, the non-Cbl related novel findings detected in 3 patients could not have been possibly detected using the gene panel, and shows the added benefit of using WGS to achieve a clinical diagnosis. The role of somatic cell studies in the diagnosis of inborn errors of Cbl metabolism has still shown to be crucial. The sensitivity and specificity that have been previously shown (Chu *et al.*, 2016) for gene panel analysis in detecting mutations in Cbl genes suggests that it is a reasonable first-line diagnostic approach. Complementation analysis can be useful as an independent test in the diagnosis of Cbl disorders and has been used as a confirmation test after finding causal mutations. The primary advantage of complementation analysis is to distinguish the different forms of MMA and assign patients to one of the known complementation groups. However, complementation studies are not necessary to make the diagnosis if the genetic defect of a patient is already known and propionate incorporation has been found to be low.

4.2. Novel findings

4.2.1 WG2625

Patient WG2625 that reported to have the *PCCA* pathogenic variant, had B_{12} responsive methylmalonic acidemia, failure to feed well for 1st month, neutropenia, thrombocytopenia and ketotic hyperglycinemia. His first urine contained both elevation of MMA and PPA (propionic acid). *PCCA* encodes for the propionyl-CoA carboxylase alpha subunit and is causal for propionic acidemia. Propionyl-CoA is an important intermediate in the metabolism of several amino acids and is also produced by oxidation of odd-numbered fatty acids. Propionyl-CoA carboxylase (PCC), comprised of alpha and beta subunits, catalyzes the first step in the catabolism of propionyl-CoA. PCC is composed of 2 nonidentical subunits, alpha and beta. The alpha subunit is encoded by the *PCCA* gene and the beta subunit by the *PCCB* gene. Cell lines from patients with propionic acidemia with mutations in the *PCCA* gene fall into complementation group pccA.

Propionic acidemia is an autosomal recessive disorder, clinically presenting with episodic vomiting, lethargy and ketosis, neutropenia, periodic thrombocytopenia, hypogammaglobulinemia, developmental retardation, and intolerance to protein. The number of known pathogenic variants in *PCCA* are 40 according to ClinVar. Using the Integrative Genomic Visualization (IGV) for interactive exploration of large genomic datasets including our NGS data, the WGS dataset is compared to the RNA-seq dataset of sample WG2625. Based on the ratios of the variant allele in WGS and RNA-seq, the mutation in *PCCA* probably leads to nonsense mediated decay of the RNA.

Paired-end NGS technique sequences both ends of each DNA fragment with library insert sizes specific to a given library preparation method and size selection procedure, leading to two paired reads to be generated at an approximately known distance in the sample genome. A signature of a discordant read-pair is formed when the mapping span and/or orientation of the read-pairs crossing the breakpoint are inconsistent with the reference genome. Specifically, both reads of the pair can be mapped to the reference genome, but they may map to different chromosomes or different orientations, or their coordinates may not agree with the insert size.(Liu *et al.*, 2015)

In cell lines from 2 patients with type I propionic acidemia, an Arg288-to-ter mutation leading to truncation of the *PCCA* molecule. The underlying mutation, a C-to-T transition at nucleotide 862, was present in homozygous form in 1 patient and in heterozygous form in the second. (Campeau *et al.*, 1999)

4.2.2 WG 4190

Patient WG 4190 had chronic diarrhea, failure to thrive, metabolic acidosis, and macrocytic anemia. He was sent to rule out *cblC* diagnosis, and WGS revealed a homozygous frameshift mutation in *EPCAM*. The *EPCAM* gene provides instructions for making a protein known as epithelial cellular adhesion molecule (EpCAM). There is reported evidence that diarrhea-5 with congenital tufting enteropathy (DIAR5) is caused by homozygous or compound heterozygous mutation in the *EPCAM* gene. This finding fits the patient’s phenotype of chronic diarrhea.

Studies on 6 consanguineous families from Kuwait and 1 family from Qatar, in which affected individuals had severe neonatal diarrhea with total dependence on parenteral nutrition as well as typical tufts on intestinal biopsy have been described. Sequencing of the *EPCAM* gene revealed that patients from 5 of the families were homozygous for the 498insC mutation, whereas the proband from 1 of the Kuwaiti families was homozygous for a splice site mutation.(Salomon *et al.*, 2011)

4.2.3. WG4160

Although this patient had not been previously tested by the Cbl gene panel, it was added as part of study. WGS would have detected any findings in Cbl related genes. This patient WG 4160 came to medical attention to rule out a Cbl disorder due to elevated MMA. This patient had acute irreversible cardiomyopathy, high C3- carnitines in addition to the elevated MMA. Propionate incorporation levels measured in this patient were 5.3 nmol/mg protein/ 18 hr. This patient has been previously tested on an extensive panel for dilated cardiomyopathy related genes that did not include *TTN*, the gene was not yet discovered. There is evidence that autosomal dominant dilated cardiomyopathy-1G (CMD1G) is caused by heterozygous mutation in the titin gene on chromosome 2q31.

The *TTN* gene provides instructions for making a massive protein called titin. This protein plays a crucial role in skeletal muscles and cardiac muscles. Slightly different isoforms of titin are created in different muscle tissues. There are multiple mutations recorded in *TTN* causing either cardiomyopathy, or muscular dystrophy. NGS was used to analyze the *TTN* gene in 203 individuals with dilated cardiomyopathy, 231 with hypertrophic cardiomyopathy (CMH), and

249 controls. The frequency of *TTN* mutations was significantly higher among individuals with CMD (27%) than among those with CMH (1%) or controls (3%). In CMD families, *TTN* mutations cosegregated with dilated cardiomyopathy, with highly observed penetrance (greater than 95%) after the age of 40 years (Herman *et al.*, 2012).

4.2.4 WG 2716

Patient WG 2716 reported to have hypotonia, head lag and recurrent illnesses associated with an acetone-like odor to his breath. He also suffered from low muscle tone, severe eczema, mild developmental delay, difficulty in gaining weight, and increased MMA in blood and urine intermittently. Propionate incorporation levels measured were 3.2 nmol/mg protein/ 18 hr17q12 is a recurrent duplication, diagnosed by detection of a 1.4-Mb submicroscopic heterozygous duplication and causes developmental delay and occasionally congenital anomalies. This syndrome has an autosomal dominant mode of inheritance, and varies in the clinical severity. Signs and symptoms related to 17q12 duplications vary even among members of the same family. Some individuals have no apparent signs or symptoms, and mild features. Other individuals can have intellectual disability, delayed development, and a wide range of physical abnormalities. The largest CNV case-control study to that time, comprising 15,749 International Standards for Cytogenomic Arrays cases and 10,118 published controls, focusing on recurrent deletions and duplications involving 14 copy number variant regions has shown evidence of the 17q12 duplication syndrome. When this study was compared with controls, 14 deletions and 7 duplications were significantly overrepresented in cases, providing a clinical diagnosis as pathogenic. The 17q12 duplication was identified in 21 cases and 4 controls for a p value of 0.022 and a frequency of 1 in 750 cases.

Conclusion and Summary

Chapter 5

In the work presented in this thesis, a combination of WGS, RNA-Seq and CNV analysis has been used to detect any genetic defects that could be responsible for the isolated elevated MMA that was seen in these twenty-six undiagnosed patients. The use of WGS has been encouraged as the new genetic diagnostic tool, and the total diagnostic yield in this study was 3/26 or 11.5 %; taking into consideration the small cohort size, and limited phenotypic information on the patients, this percentage is fair compared to other studies that used the same methodologies. A recent pediatric study evaluated the use of genome sequencing vs. WES and found that WGS detected diagnostic variants in 41% of cases. WGS has also captured all molecular diagnoses found by conventional testing methods in this study. The 18 new diagnoses made with WGS included structural and non-exonic sequence variants not detectable with whole-exome sequencing, and confirmed recent disease associations with the certain genes. (Lionel *et al.*, 2018). Another study discusses another benefit of WGS is by providing genome-wide read coverage, the reliable detection of CNVs is allowed, which can contribute substantially to disease burden. This study also suggest problematic issues with WES such as insufficient exome coverage and GC content sensitivity, in contrast to WGS which is proposed to surpass these problems (Meienberg *et al.*, 2016). In the research project discussed in this thesis, the use of WGS has certainly allowed the identification of CNV detection in two patients and have further explained their genetic findings. An extensive study on neurodevelopmental disorders compared the effectiveness of WES and WGS in one hundred families with 119 children. Forty-five percent received molecular diagnoses. An accelerated sequencing modality, rapid WGS, yielded diagnoses in 73% of families with acutely ill children (11 of 15). Forty percent of families with children with nonacute NDD, followed in ambulatory care clinics (34 of 85), received diagnoses: 33 by WES and 1 by staged WES then WGS.(Soden *et al.*, 2014).

Elevations of MMA and homocysteine, either alone or in combination, have been the reason for referring patients with suspected Cbl disorders for many years. Clinically, measurement of total homocysteine and MMA are used as primary tests when a suspicion of Cbl deficiency is made. However, the cause of elevation of MMA in the 26 original patients and the additional added patient remains unclear, and is probably a question that will need to be answered in the future. Exploring the possible links between MMA and mutations in the genes involved in our findings would be very interesting to explore in future studies and may shed light on whether there are other metabolic pathways involved in the processing of MMA.For 9 patients, NGS analysis confirmed the VUS identified previously by the Cbl gene panel. Novel findings in 3 patients haven been detected and explain part of the patient's phenotype, but do not explain the elevated MMA. After selecting patients according to [^{14}C] propionate incorporation levels, it has not been shown to affect results. Moreover, it can be stated that somatic cell studies performed in the laboratory are reliable methods in diagnosing patients with typical phenotypic picture of a Cbl disorder. In atypical patients, somatic cell studies can only rule out a Cbl diagnosis, and not add much more. Further studies are needed to evaluate the reason behind the elevation of MMA in this cohort of patients, and if there are any external factors affecting these levels. Finally, 13 patients with apparent defects in propionyl-CoA metabolism remain undiagnosed, WGS and RNA-seq has not been fruitful in these cases. Whole genome and RNA-sequencing did not identify novel Cbl genes in our research project, confirming that somatic cell studies including complementation analysis remain clinically useful.

Bibliography

Chapter 6

Adzhubei, I.A. et al., 2010. A method and server for predicting damaging missense mutations. Nature Methods, 7(4), pp.248–249.

Andersen, C.B.F. et al., 2010. Structural basis for receptor recognition of vitamin-B12–intrinsic factor complexes. Nature, 464(7287), pp.445–448.

Beaulieu, C.L. et al., 2014. FORGE Canada Consortium: Outcomes of a 2-Year National Rare-Disease Gene-Discovery Project. The American Journal of Human Genetics, 94(6), pp.809–817.

Bobik, T.A. & Rasche, M.E., 2001. Identification of the human methylmalonyl-CoA racemase gene based on the analysis of prokaryotic gene arrangements: implications for decoding the human genome §*. § Florida Agricultural Experiment Station Journal Series

Brasil, S. et al., 2015. Methylmalonic aciduria cblB type: characterization of two novel mutations and mitochondrial dysfunction studies. Clinical Genetics, 87(6), pp.576–581.

Briani, C. et al., 2013a. Cobalamin deficiency: clinical picture and radiological findings. Nutrients, 5(11), pp.4521–39.

Brolin, R.E. & Leung, M., 1999. Survey of Vitamin and Mineral Supplementation after Gastric Bypass and Biliopancreatic Diversion for Morbid Obesity

Brzezinski A., 1997. Review article: Mechanisms of Disease - Melatonin in Humans. The new England Journal of Medicine, 336(3), pp.186–195.

Campeau, E. et al., 1999. Detection of a normally rare transcript in propionic acidemia patients with mRNA destabilizing mutations in the PCCA gene. Human Molecular Genetics, 8(1), pp.107–113.

Carrillo-Carrasco, N. & Venditti, C.P., 2012. Combined methylmalonic acidemia and homocystinuria, cblC type. II. Complications, pathophysiology, and outcomes. Journal of Inherited Metabolic Disease, 35(1), pp.103–114.

Chu, J. et al., 2016. Next generation sequencing of patients with mut methylmalonic aciduria: Validation of somatic cell studies and identification of 16 novel mutations. Molecular Genetics and Metabolism, 118(4), pp.264–271.

Cirulli, E.T. & Goldstein, D.B., 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nature Reviews Genetics, 11(6), pp.415–425.

Clarke, L. et al., 2012. The 1000 Genomes Pproject: Data management and community access. Nature Methods.

Coelho, D. et al., 2008. Gene Identification for the cblD Defect of Vitamin B 12 Metabolism. New England Journal of Medicine, 358(14), pp.1454–1464.

Coelho, D. et al., 2012. Mutations in ABCD4 cause a new inborn error of vitamin B12 metabolism. Nature Genetics, 44(10), pp.1152–1155.

Costanzo, M. et al., 2018. Label-Free Quantitative Proteomics in a Methylmalonyl-CoA Mutase-Silenced Neuroblastoma Cell Line. International Journal of Molecular Sciences, 19(11), p.3580.

Craig, W.J., 2009. Health effects of vegan diets. The American Journal of Clinical Nutrition, 89(5), p.1627S–1633S.

DePristo, M.A. et al., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics, 43(5), pp.491–498.

Dobolyi, A., Ostergaard, E., et al., 2015. Exclusive neuronal expression of SUCLA2 in the human brain. Brain Structure and Function, 220(1), pp.135–151.

Dobolyi, A., Bagó, A.G., et al., 2015. Localization of SUCLA2 and SUCLG2 subunits of succinyl CoA ligase within the cerebral cortex suggests the absence of matrix substrate-level phosphorylation in glial cells of the human brain. Journal of Bioenergetics and Biomembranes, 47(1–2), pp.33–41.

Dobson, C.M. et al., 2006. Homozygous nonsense mutation in the MCEE gene and siRNA suppression of methylmalonyl-CoA epimerase expression: A novel cause of mild methylmalonic aciduria. Molecular Genetics and Metabolism, 88(4), pp.327–333.

Dobson, C.M. et al., 2002. Identification of the gene responsible for the cblA complementation group of vitamin B12-responsive methylmalonic acidemia based on analysis of prokaryotic gene arrangements. Proceedings of the National Academy of Sciences of the United States of America, 99(24), pp.15554–9.

Fischer, S. et al., 2014. Clinical presentation and outcome in a series of 88 patients with the cblC defect. Journal of Inherited Metabolic Disease, 37(5), pp.831–840.

Froese, D.S. et al., 2010. Structures of the human GTPase MMAA and vitamin B12-dependent methylmalonyl-CoA mutase and insight into their complex formation. The Journal of biological chemistry, 285(49), pp.38204–13.

Froese, D.S. et al., 2010. Thermolability of mutant MMACHC protein in the vitamin B12-responsive cblC disorder. Molecular genetics and metabolism, 100(1), pp.29–36.

Gailus, S. et al., 2010. A novel mutation in LMBRD1 causes the cblF defect of vitamin B 12 metabolism in a Turkish patient Cologne Excellence Cluster on Cellular Stress Response in Aging-associated Diseases. J Inherit Metab Dis, 33, pp.17–24.

Gehrer, S. et al., 2010. Fewer Nutrient Deficiencies After Laparoscopic Sleeve Gastrectomy (LSG) than After Laparoscopic Roux-Y-Gastric Bypass (LRYGB)—a Prospective Study. Obesity Surgery, 20(4), pp.447–453.

Gherasim, C., Lofgren, M. & Banerjee, R., 2013. Navigating the B 12 road: assimilation, delivery and disorders of cobalamin Running Title: B 12 Trafficking in Mammals

Gibril, F. & Jensen, R.T., 2004. Zollinger-Ellison Syndrome Revisited: Diagnosis, Biologic Markers, Associated Inherited Disorders, and Acid Hypersecretion. Current Gastroenterology Reports, 6, pp.454–463.

Gorfine, M. et al., 2015. Function of Cancer Associated Genes Revealed by Modern Univariate and Multivariate Association Tests L. Chen, ed. Plos one, 10(5), p.e0126544.

Gorman, R.C., Are Vitamin B12 and Folate Deficiency Clinically Important After Roux-en-Y Gastric Bypass ? , pp.436–437.

Hemmer, B. et al., 1998. Subacute combined degeneration: clinical, electrophysiological, and magnetic resonance imaging findings.

Herman, D.S. et al., 2012. Truncations of Titin Causing Dilated Cardiomyopathy. New England Journal of Medicine, 366(7), pp.619–628.

Heuberger, K. et al., 2019. Genetic, structural, and functional analysis of pathogenic variations causing methylmalonyl-CoA epimerase deficiency. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease.

Homolova, K. et al., 2010. The deep intronic c.903+469T&gt;C mutation in the MTRR gene creates an SF2/ASF binding exonic splicing enhancer, which leads to pseudoexon activation and causes the cblE type of homocystinuria. Human Mutation, 31(4), pp.437–444.

Hörster, F. et al., 2007a. Long-Term Outcome in Methylmalonic Acidurias Is Influenced by the Underlying Defect (mut0, mut−, cblA, cblB). Pediatric Research, 62(2), pp.225–230.

Huemer, M. et al., 2015. Clinical onset and course, response to treatment and outcome in 24 patients with the cblE or cblG remethylation defect complemented by genetic and in vitro enzyme study data. Journal of Inherited Metabolic Disease, 38(5), pp.957–967.

Institute of Medicine (US) Standing Committee on the Scientific Evaluation of Dietary Reference Intakes and its Panel on Folate, O.B.V. and C., 1998. Dietary Reference Intakes for Thiamin, Riboflavin, Niacin, Vitamin B6, Folate, Vitamin B12, Pantothenic Acid, Biotin, and Choline, National Academies Press (US).

Jack Metz, 1993. Pathogenesis of Cobalamin Neuropathy: Deficiency of Nervous System S-Adenosylmethionine? Nutrition Reviews.

Jamuar, S.S. et al., 2014. Somatic mutations in cerebral cortical malformations. The New England journal of medicine, 371(8), pp.733–43.

Janata, J., Kogekar, N. & Fenton, W.A., 1997. Expression and kinetic characterization of methylmalonyl-CoA mutase from patients with the mut- phenotype: evidence for naturally occurring interallelic complementation. Human Molecular Genetics, 6(9), pp.1457–1464.

Johnson, J.D. et al., 1998. Genetic evidence for the expression of ATP- and GTP-specific succinyl-CoA synthetases in multicellular eucaryotes. The Journal of biological chemistry, 273(42), pp.27580–6.

Jusufi, J. et al., 2014. Characterization of functional domains of the cblD (MMADHC) gene product. Journal of Inherited Metabolic Disease, 37(5), pp.841–849.

Karczewski, K.J. et al., 2017. The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Research, 45(D1), pp.D840–D845.

Kim, J.C. et al., 2012a. Late onset of symptoms in an atypical patient with the cblJ inborn error of vitamin B12 metabolism: Diagnosis and novel mutation revealed by exome sequencing. Molecular Genetics and Metabolism, 107(4), pp.664–668.

Kim, J.C. et al., 2012b. Late onset of symptoms in an atypical patient with the cblJ inborn error of vitamin B12 metabolism: Diagnosis and novel mutation revealed by exome sequencing. Molecular Genetics and Metabolism, 107(4), pp.664–668.

Kircher, M. et al., 2014. A general framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics, 46(3), pp.310–315.

Kumar, P., Henikoff, S. & Ng, P.C., 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols, 4(7), pp.1073–1081.

Kwon, Y. et al., 2014. Anemia, iron and vitamin B12 deficiencies after sleeve gastrectomy compared to Roux-en-Y gastric bypass: a meta-analysis. Surgery for Obesity and Related Diseases, 10(4), pp.589–597.

Lahner, E. & Annibale, B., 2009. Pernicious anemia: new insights from a gastroenterological point of view. World journal of gastroenterology, 15(41), pp.5121–8.

Leclerc, D. et al., 1998. Cloning and mapping of a cDNA for methionine synthase reductase, a flavoprotein defective in patients with homocystinuria,

Lerner-Ellis, J.P. et al., 2009. Spectrum of mutations in MMACHC , allelic expression, and evidence for genotypeâ€“phenotype correlations. Human Mutation, 30(7), pp.1072–1081.

Lerner‐Ellis, J.P. et al., 2004. Mutations in the MMAA gene in patients with the cblA disorder of vitamin B12 metabolism. Human Mutation, 24(6), pp.509‐516.

Li, H. & Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14), pp.1754–1760.

Lionel, A.C. et al., 2018. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. Genetics in Medicine, 20(4), pp.435–443.

Liu, B. et al., 2015. Structural variation discovery in the cancer genome using next generation sequencing: computational solutions and perspectives. Oncotarget, 6(8), pp.5477–89.

Manoli, I., Sloan, J.L. & Venditti, C.P., 1993. Isolated Methylmalonic Acidemia, University of Washington, Seattle.

McKenna, A. et al., 2010. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research.

Meienberg, J. et al., 2016. Clinical sequencing: is WGS the better WES? Human Genetics, 135(3), pp.359–362.

Mellman, I.S. et al., 1977. Intracellular binding of radioactive hydroxocobalamin to cobalamin-dependent apoenzymes in rat liver* (subcellular fractionation/cobalamin metabolism/vitamin B12/binding proteins/human genetic disorder),

Michaud, J. et al., 2013. HCFC1 is a common component of active human CpG-island promoters and coincides with ZNF143, THAP11, YY1, and GABP transcription factor occupancy. Genome research, 23(6), pp.907–16.

Monlong, J. et al., 2018. Global characterization of copy number variants in epilepsy patients from whole genome sequencing S. Petrou, ed. PLOS Genetics, 14(4), p.e1007285.

Nielsen, M.J. et al., 2012. Vitamin B12 transport from food to the body's cells—a sophisticated, multistep pathway. Nature Reviews Gastroenterology & Hepatology, 9(6), pp.345–354.

Otzen, C. et al., 2014. Candida albicans utilises a modified β-oxidation pathway for the degradation of toxic propionyl-CoA*.

Padovani, D. et al., 2008. Adenosyltransferase tailors and delivers coenzyme B12. Nature Chemical Biology, 4(3), pp.194–196.

Pupavac, M. et al., 2016. Added value of next generation gene panel analysis for patients with elevated methylmalonic acid and no clinical diagnosis following functional studies of vitamin B12 metabolism. Molecular Genetics and Metabolism, 117(3), pp.363–368.

Qiu, A. et al., 2006. Identification of an Intestinal Folate Transporter and the Molecular Basis for Hereditary Folate Malabsorption. Cell, 127(5), pp.917–928.

Quintana, A.M. et al., 2017. Mutations in THAP11 cause an inborn error of cobalamin metabolism and developmental abnormalities. Human Molecular Genetics, 26(15), pp.2838–2849.

Raff, M.L. et al., 1991. Genetic characterization of a MUT locus mutation discriminating heterogeneity in mut0and mut-methylmalonic aciduria by interallelic complementation. Journal of Clinical Investigation, 87(1), pp.203–207.

Rivas, M.A. et al., 2011. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nature genetics, 43(11), pp.1066–73.

Rutsch, F. et al., 2009. Identification of a putative lysosomal cobalamin exporter altered in the cblF defect of vitamin B12 metabolism. Nature Genetics, 41(2), pp.234–239.

Rutsch, F. et al., 2011. LMBRD1: the gene for the cblF defect of vitamin B12 metabolism. Journal of Inherited Metabolic Disease, 34(1), pp.121–126.

Salomon, J. et al., 2011. A founder effect at the EPCAM locus in Congenital Tufting Enteropathy in the Arabic Gulf. European Journal of Medical Genetics, 54(3), pp.319–322.

Scholz, T. et al., 2009. Update on the human broad tapeworm (genus diphyllobothrium), including clinical relevance. Clinical microbiology reviews, 22(1), p.146–60,

SCOTT, J.M., 1981. Pathogenesis of subacute combined degeneration: a result of methyl group deficiency

Siepel, A. et al., 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome research, 15(8), pp.1034–50.

Soden, S.E. et al., 2014. Effectiveness of exome and genome sequencing guided by acuity of illness for diagnosis of neurodevelopmental disorders. Science Translational Medicine, 6(265), p.265ra168-265ra168.

Stabler, S.P. & Allen, R.H., 2004. Vitamin B12 deficiency as a worldwide problem. Annual Review of Nutrition, 24(1), pp.299–326.

Stavropoulos, D.J. et al., 2016. Whole-genome sequencing expands diagnostic utility and improves clinical management in paediatric medicine. npj Genomic Medicine,

Suormala, T. et al., 2004. The cbld defect causes either isolated or combined deficiency of methylc adenosylcobalamin synthesis downloaded from, JBC.

Takeichi, T. et al., 2015. Progressive hyperpigmentation in a Taiwanese child due to an inborn error of vitamin B12 metabolism (cblJ). British Journal of Dermatology, 172(4), pp.1111–1115.

Watkins, D. & Rosenblatt, D.S., 2011. Inborn errors of cobalamin absorption and metabolism. American Journal of Medical Genetics Part C: Seminars in Medical Genetics, 157(1), pp.33–44.

Watson, E. et al., 2016. Metabolic network rewiring of propionate flux compensates vitamin B12 deficiency in C. elegans. eLife, 5.

Willard, H.F. & Rosenberg, L.E., 1980. Inherited Methylmalonyl CoA Mutase Apoenzyme Deficiency in Human Fibroblasts : evidence for allelic heterogeneity , genetic compounds , and codominant expression

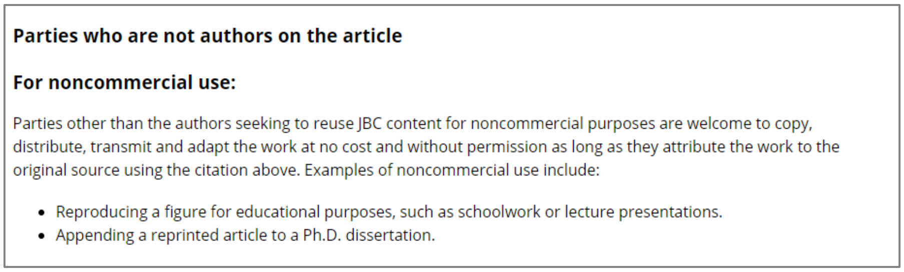
Yamanishi, M., Vlasie, M. & Banerjee, R., 2005. Adenosyltransferase: an enzyme and an escort for coenzyme B12? Trends in Biochemical Sciences, 30(6), pp.304–308.

Yu, H.-C. et al., 2013. An X-Linked Cobalamin Disorder Caused by Mutations in Transcriptional Coregulator HCFC1. The American Journal of Human Genetics, 93, pp.506–514.

Appendix

Copyright permissions for published figures

1- Permission for using Figure 1, editorial policies from the Journal of Biological Chemistry:



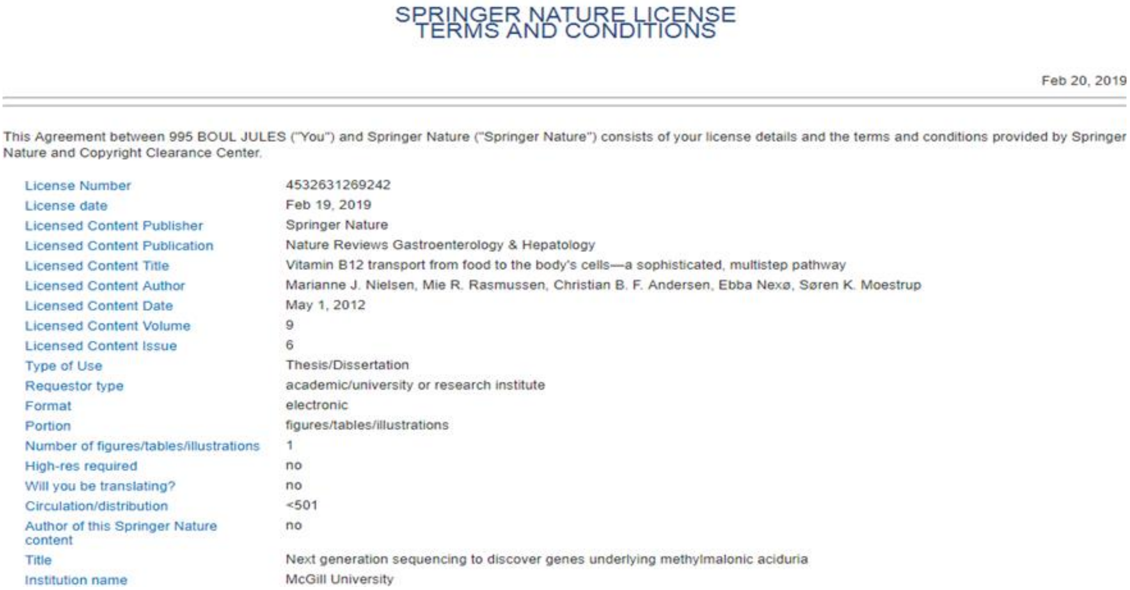
Parties who are not authors on the article

For noncommercial use:

Parties other than the authors seeking to reuse JBC content for noncommercial purposes are welcome to copy, distribute, transmit and adapt the work at no cost and without permission as long as they attribute the work to the original source using the citation above. Examples of noncommercial use include:

- Reproducing a figure for educational purposes, such as schoolwork or lecture presentations.
- Appending a reprinted article to a Ph.D. dissertation.

2- Permission for using Figure 2, license number: 4532631269242, obtained from Springer Nature.

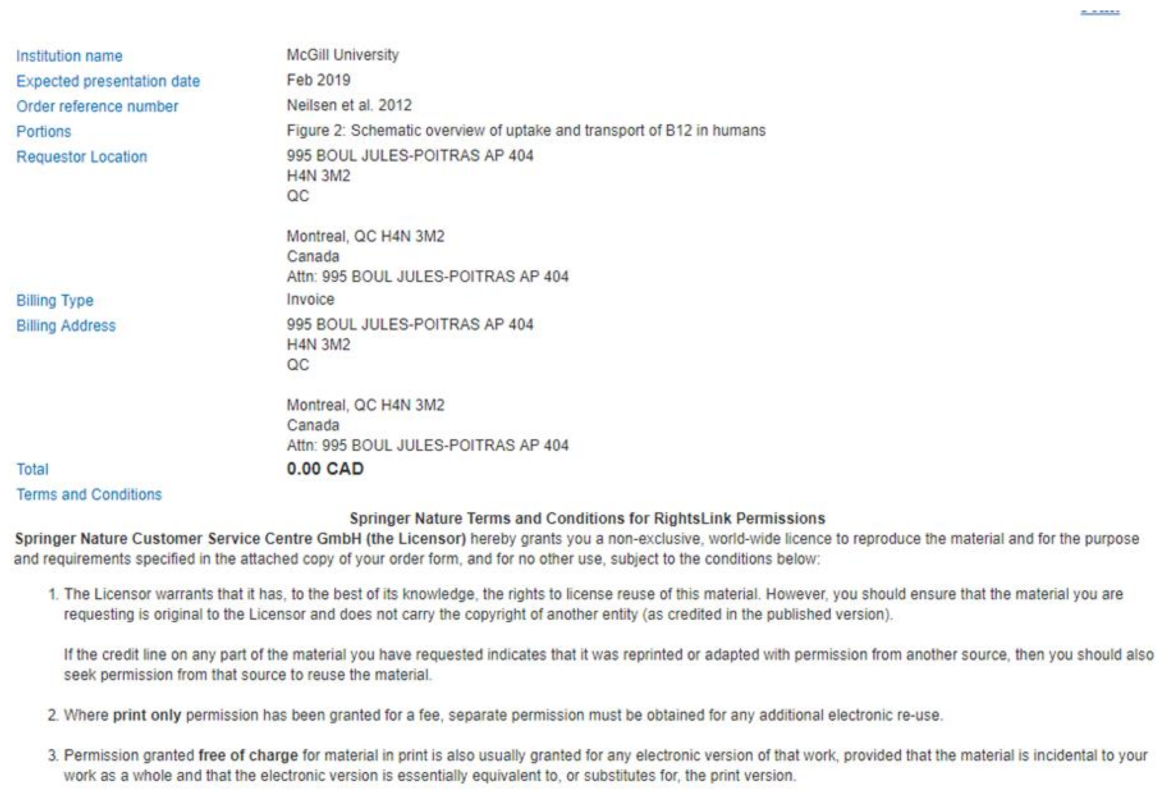


SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Feb 20, 2019

This Agreement between 995 BOUL JULES ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4532631269242
License date	Feb 19, 2019
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Nature Reviews Gastroenterology & Hepatology
Licensed Content Title	Vitamin B12 transport from food to the body's cells—a sophisticated, multistep pathway
Licensed Content Author	Marianne J. Nielsen, Mie R. Rasmussen, Christian B. F. Andersen, Ebba Nexø, Søren K. Moestrup
Licensed Content Date	May 1, 2012
Licensed Content Volume	9
Licensed Content Issue	6
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Will you be translating?	no
Circulation/distribution	<501
Author of this Springer Nature content	no
Title	Next generation sequencing to discover genes underlying methylmalonic aciduria
Institution name	McGill University



Institution name	McGill University
Expected presentation date	Feb 2019
Order reference number	Neilsen et al. 2012
Portions	Figure 2: Schematic overview of uptake and transport of B12 in humans
Requestor Location	995 BOUL JULES-POITRAS AP 404 H4N 3M2 QC Montreal, QC H4N 3M2 Canada Attn: 995 BOUL JULES-POITRAS AP 404
Billing Type	Invoice
Billing Address	995 BOUL JULES-POITRAS AP 404 H4N 3M2 QC Montreal, QC H4N 3M2 Canada Attn: 995 BOUL JULES-POITRAS AP 404
Total	**0.00 CAD**
Terms and Conditions	

Springer Nature Terms and Conditions for RightsLink Permissions

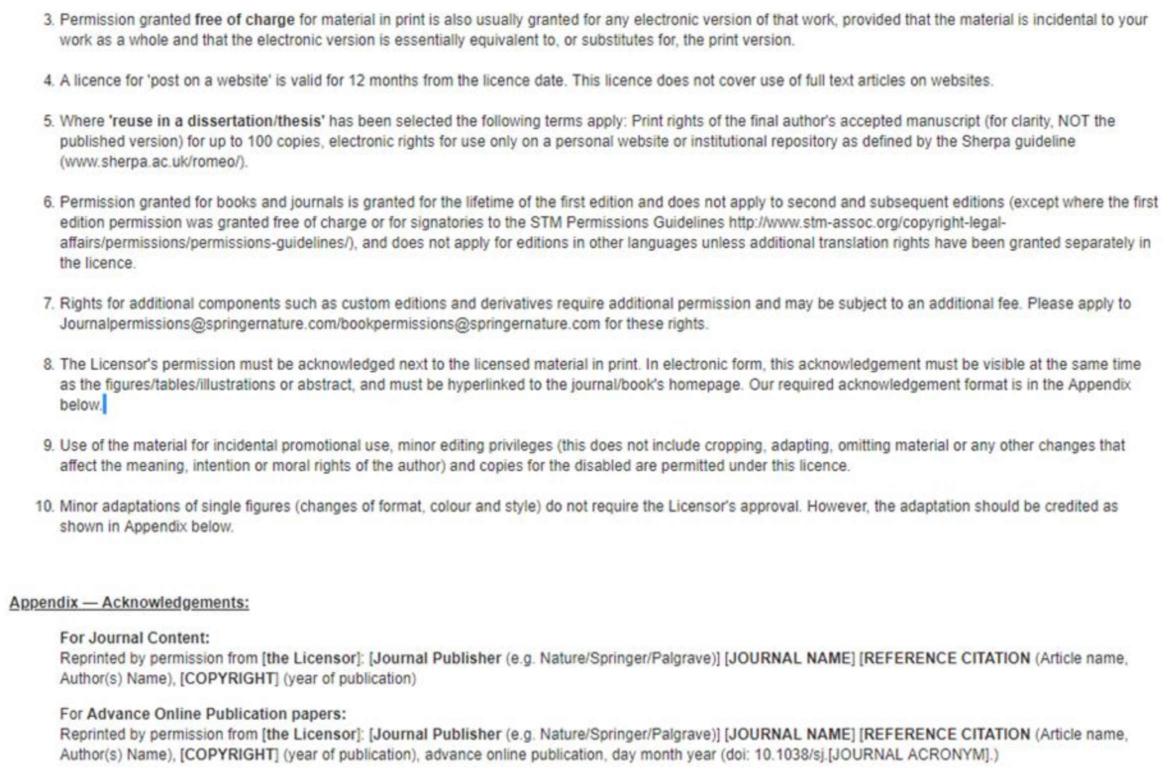
Springer Nature Customer Service Centre GmbH (the Licensor) hereby grants you a non-exclusive, world-wide licence to reproduce the material and for the purpose and requirements specified in the attached copy of your order form, and for no other use, subject to the conditions below:

1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of this material. However, you should ensure that the material you are requesting is original to the Licensor and does not carry the copyright of another entity (as credited in the published version).

   If the credit line on any part of the material you have requested indicates that it was reprinted or adapted with permission from another source, then you should also seek permission from that source to reuse the material.

2. Where **print only** permission has been granted for a fee, separate permission must be obtained for any additional electronic re-use.

3. Permission granted **free of charge** for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to your work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.



3. Permission granted **free of charge** for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to your work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.

4. A licence for 'post on a website' is valid for 12 months from the licence date. This licence does not cover use of full text articles on websites.

5. Where **'reuse in a dissertation/thesis'** has been selected the following terms apply: Print rights of the final author's accepted manuscript (for clarity, NOT the published version) for up to 100 copies, electronic rights for use only on a personal website or institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/).

6. Permission granted for books and journals is granted for the lifetime of the first edition and does not apply to second and subsequent editions (except where the first edition permission was granted free of charge or for signatories to the STM Permissions Guidelines http://www.stm-assoc.org/copyright-legal-affairs/permissions/permissions-guidelines/), and does not apply for editions in other languages unless additional translation rights have been granted separately in the licence.

7. Rights for additional components such as custom editions and derivatives require additional permission and may be subject to an additional fee. Please apply to Journalpermissions@springernature.com/bookpermissions@springernature.com for these rights.

8. The Licensor's permission must be acknowledged next to the licensed material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract, and must be hyperlinked to the journal/book's homepage. Our required acknowledgement format is in the Appendix below.

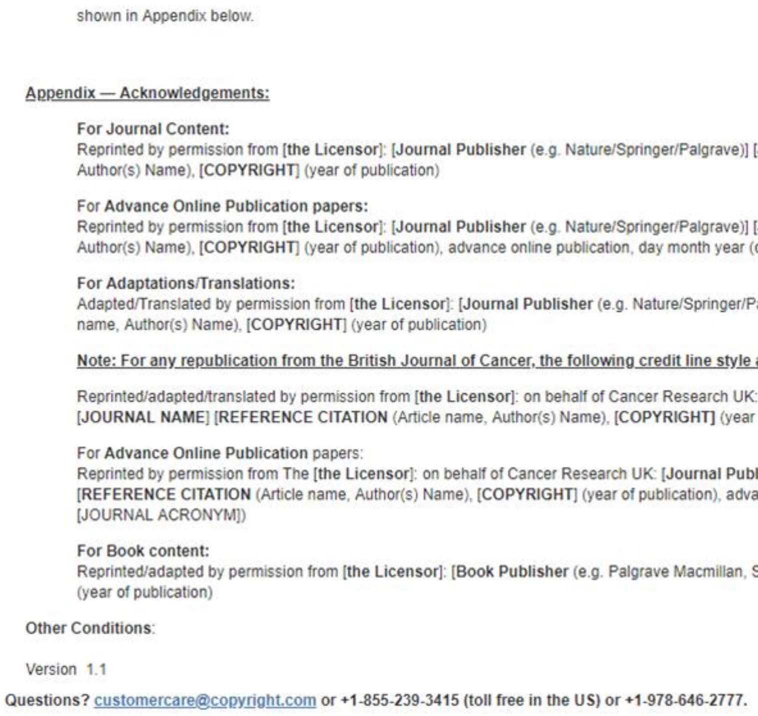
9. Use of the material for incidental promotional use, minor editing privileges (this does not include cropping, adapting, omitting material or any other changes that affect the meaning, intention or moral rights of the author) and copies for the disabled are permitted under this licence.

10. Minor adaptations of single figures (changes of format, colour and style) do not require the Licensor's approval. However, the adaptation should be credited as shown in Appendix below.

Appendix — Acknowledgements:

For Journal Content:
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication papers:**
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].)



shown in Appendix below.

Appendix — Acknowledgements:

For Journal Content:
Reprinted by permission from **[the Licensor]**: **[Journal Publisher** (e.g. Nature/Springer/Palgrave)] **[JOURNAL NAME] [REFERENCE CITATION** (Article name, Author(s) Name), **[COPYRIGHT]** (year of publication)

For **Advance Online Publication papers:**
Reprinted by permission from **[the Licensor]**: **[Journal Publisher** (e.g. Nature/Springer/Palgrave)] **[JOURNAL NAME] [REFERENCE CITATION** (Article name, Author(s) Name), **[COPYRIGHT]** (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].)

For Adaptations/Translations:
Adapted/Translated by permission from **[the Licensor]**: **[Journal Publisher** (e.g. Nature/Springer/Palgrave)] **[JOURNAL NAME] [REFERENCE CITATION** (Article name, Author(s) Name), **[COPYRIGHT]** (year of publication)

Note: For any republication from the British Journal of Cancer, the following credit line style applies:

Reprinted/adapted/translated by permission from **[the Licensor]**: on behalf of Cancer Research UK: : **[Journal Publisher** (e.g. Nature/Springer/Palgrave)] **[JOURNAL NAME] [REFERENCE CITATION** (Article name, Author(s) Name), **[COPYRIGHT]** (year of publication)

For **Advance Online Publication** papers:
Reprinted by permission from The **[the Licensor]**: on behalf of Cancer Research UK: **[Journal Publisher** (e.g. Nature/Springer/Palgrave)] **[JOURNAL NAME] [REFERENCE CITATION** (Article name, Author(s) Name), **[COPYRIGHT]** (year of publication), advance online publication, day month year (doi: 10.1038/sj. [JOURNAL ACRONYM])

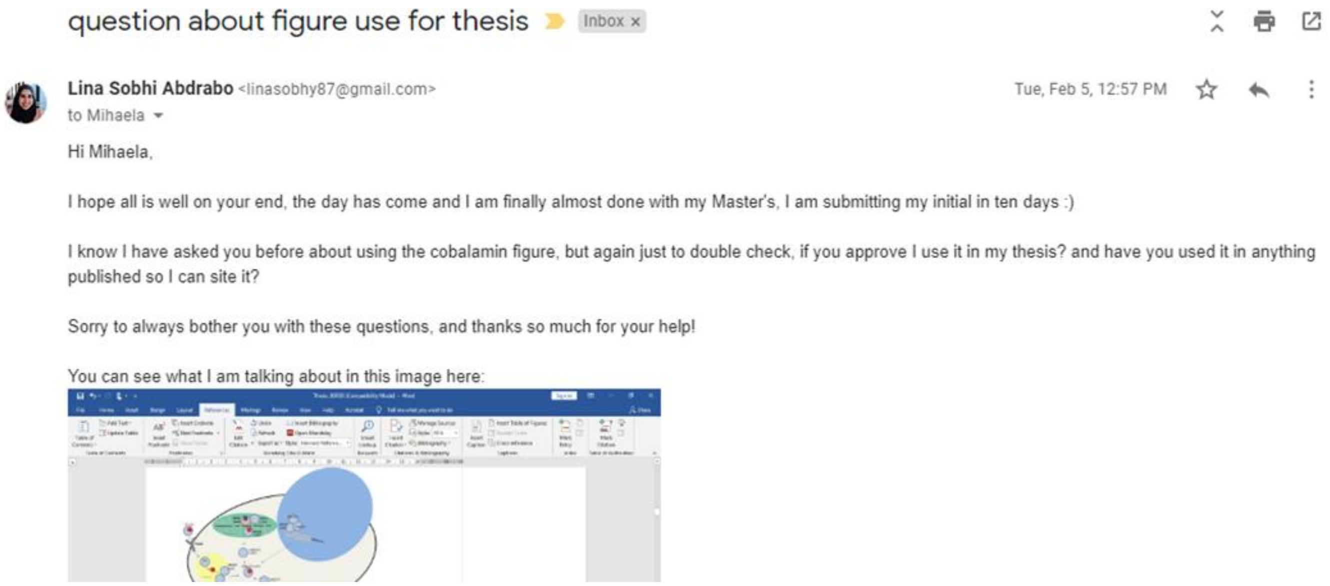
For Book content:
Reprinted/adapted by permission from **[the Licensor]**: **[Book Publisher** (e.g. Palgrave Macmillan, Springer etc) **[Book Title]** by **[Book author**(s)] **[COPYRIGHT]** (year of publication)

Other Conditions:

Version 1.1

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

1- Permission to use Figure 3, from Author, obtained via email:



question about figure use for thesis

Lina Sobhi Abdrabo <linasobhy87@gmail.com>
to Mihaela

Tue, Feb 5, 12:57 PM

Hi Mihaela,

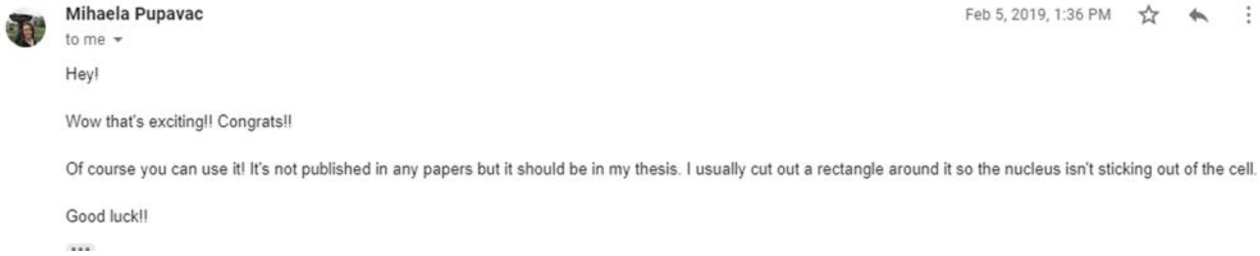
I hope all is well on your end, the day has come and I am finally almost done with my Master's, I am submitting my initial in ten days :)

I know I have asked you before about using the cobalamin figure, but again just to double check, if you approve I use it in my thesis? and have you used it in anything published so I can site it?

Sorry to always bother you with these questions, and thanks so much for your help!

You can see what I am talking about in this image here:

Figure 3. Intracellular cobalamin processing. All the genes implicated in the intracellular processing of cobalamin are labelled and their corresponding Cbl disorder. After binding of the Cbl-TC complex on the TC receptor and facilitated endocytosis, TC dissociates from Cbl in the lysosome. Free Cbl then exits the lysosome with the help of two proteins LMBRD1 and ABCD4. Upon exiting the lysosome, Cbl is processed into its coenzyme derivatives, AdoCbl and MeCbl. The MMACHC (methylmalonic aciduria and homocystinuria cblC type) protein interacts with cobalamin with various upper-axial ligands, converting them to a common intermediate that can be used the synthesis of AdoCbl and MeCbl. Adapted from Mihaela Pupavac with permission.



Mihaela Pupavac
to me

Feb 5, 2019, 1:36 PM

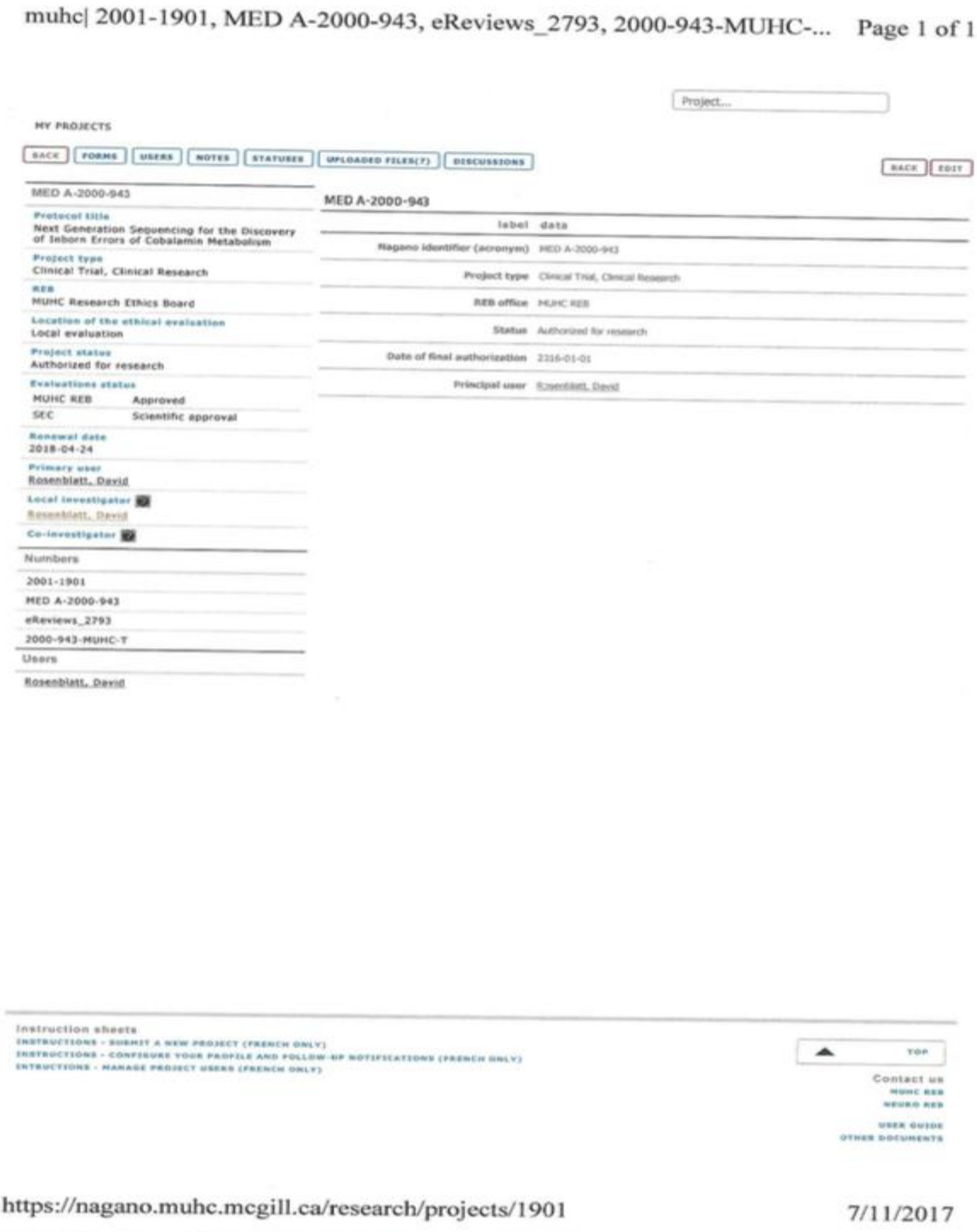
Hey!

Wow that's exciting!! Congrats!!

Of course you can use it! It's not published in any papers but it should be in my thesis. I usually cut out a rectangle around it so the nucleus isn't sticking out of the cell.

Good luck!!

Ethics and Related Certificates



muhc| 2001-1901, MED A-2000-943, eReviews_2793, 2000-943-MUHC-... Page 1 of 1

MED A-2000-943

Protocol title
Next Generation Sequencing for the Discovery of Inborn Errors of Cobalamin Metabolism

Project type
Clinical Trial, Clinical Research

REB
MUHC Research Ethics Board

Location of the ethical evaluation
Local evaluation

Project status
Authorized for research

Evaluations status

MUHC REB	Approved
SEC	Scientific approval

Renewal date
2018-04-24

Primary user
Rosenblatt, David

Local investigator
Rosenblatt, David

Co-investigator

Numbers

2001-1901
MED A-2000-943
eReviews_2793
2000-943-MUHC-T

Users

Rosenblatt, David

MED A-2000-943

label	data
Nagano identifier (acronym)	MED A-2000-943
Project type	Clinical Trial, Clinical Research
REB office	MUHC REB
Status	Authorized for research
Date of final authorization	2316-01-01
Principal user	Rosenblatt, David

Instruction sheets
INSTRUCTIONS - SUBMIT A NEW PROJECT (FRENCH ONLY)
INSTRUCTIONS - CONFIGURE YOUR PROFILE AND FOLLOW-UP NOTIFICATIONS (FRENCH ONLY)
INTRUCTIONS - MANAGE PROJECT USERS (FRENCH ONLY)
TOP
Contact us
MUHC REB
NEURO REB
USER GUIDE
OTHER DOCUMENTS

https://nagano.muhc.mcgill.ca/research/projects/1901 7/11/2017