NEXT GENERATION SEQUENCING ANALYSIS OF THE CYP2D6 GENE LOCUS

By

Tyler Joseph Brenneman

Submitted in partial fulfillment of the

requirements for Departmental Honors in

the Department of Biology

Texas Christian University

Fort Worth, Texas

May 2, 2014

NEXT GENERATION SEQUENCING ANALYSIS OF THE CYP2D6 GENE LOCUS

Project Approved:

Supervising Professor: Phil Hartman, Ph.D.

Department of Biology

Ray Drenner, Ph.D.

Department of Biology

Ronald Pitcock, Ph.D.

John V. Roach Honors College

ABSTRACT

The CYP2D6 enzyme is very active in the metabolism of many therapeutic drugs currently on the market. A highly polymorphic enzyme with four phenotypic classes of metabolic speed, CYP2D6 is a gene that would be very beneficial if a treating physician knew the allele variant present in each patient. Having this information in treatment can improve therapeutic effects while preventing toxic levels of a drug from building up in a patient's blood because the dosage can be tailored to the individual patient's metabolic rate of the drug. Sanger Sequencing is currently the Gold Standard for being able to call genotypic variants of this highly polymorphic gene, however, time constraints associated with this method information prevent its therapeutic use. New Next Generation Sequencing technologies have greatly increased the speed of obtaining sequencing data and could be useful in moving to a more personalized medical treatment. If the Next Generation Sequencing technology is accurately able to call the genotype of the alleles present in each patient, this information could be used to alter dosages and plan treatment for patients around the world. From this study, it was determined that the Next Generation Sequencing technology was able to accurately call the correct phenotypic class for each of the 12 patients. With this information, we hope to be able to move to a personalized medicine to treat patients in a shorter period of time.

ACKNOWLEDGMENTS

I would like to start by thanking Dr. Andrea Gaedigk at Children's Mercy Hospital in Kansas City, MO for providing me with the opportunity to complete this research in the clinical pharmacology lab. Through her extensive work on p450 enzymes and on the CYP2D6 gene in particular, she has become a world leader in the field and taught me everything I know about the gene. Her hours spent mentoring and helping me understand the concepts in the lab were truly what made this experience so great. I would also like to thank Liliane Ndjountche who helped me learn many new lab techniques and answered my questions so kindly every day. I am also very appreciative of Greyson Twist who was the brains behind much of the data interpretation and assay setups. Also from Children's Mercy I want to thank Dr. Ralph Kauffman who put me in touch with Dr. Gaedigk to help get me placed in the lab as well as the Whole Genome Team and Genomics lab that provided me with the Next Gen Sequencing data that was crucial to this study. From TCU I want to thank Dr. Phil Hartman for mentoring me throughout this project starting with his taking the time to allow me to pitch this external research and use the study for my senior thesis. The patience and dedication that he has shown throughout this process made it a very enjoyable task to accomplish. Finally, I am grateful for Dr. Ray Drenner and Dr. Ron Pitcock for taking the time to read and edit this thesis and provide crucial input to help make it a polished final project. Thank you all again for your guidance and assistance in making this project a successful one

TABLE OF CONTENTS

INTRODUCTION

Cytochrome P450s (CYPs) are a super family of enzymes that are responsible for metabolism of many exogenous and endogenous compounds within mammals, including humans (Niwa et al, 2011). While interesting in that they have a wide range of substrates, it is their ability to metabolize therapeutic drugs that makes the CYPs so heavily studied. The CYP family has over 6000 known members; however, it is seven specific isoforms that account for the metabolism of over 90% of drugs that are currently approved for clinical use on the market (Marechal, 2008). The specific CYP2D6 isoform makes up between 2-9% of all of the p450 enzymes that reside in human livers. Despite this low percentage, CYP2D6 metabolizes 20-30% of the therapeutic drugs on the market. These include important agents such as analgesics, antidepressants, β-blockers, antipsychotics and antiarrhythmics, among many others (Niwa, et al, 2011). Many of these types of drugs have the potential to have very profound impacts in patients, especially children.

Even though the CYP2D6 isoform of the p450 enzyme makes up a small percentage, it deserves extensive study due to the highly polymorphic nature of the gene that exhibits itself in numerous genotypes and phenotypes (Marechal, 2008). This allelic variance results in three major families of activity, which are active, reduced function and non-functional. These translate into four separate phenotypic families: poor metabolizers, intermediate metabolizers, extensive metabolizers and ultra-rapid metabolizers (Gaedigk, et al, 2007). As of January 21, 2014 there were 155 allelic variants of the CYP2D6 gene locus (http://www.cypalleles.ki.se/cyp2d6.htm).

Since the variability in CYP2D6 metabolic activity is so large, dosages of therapeutic drug dosages have the very real possibility of being either insufficient or occurring at toxic levels in the blood. In fact, according to the Centers for Disease Control and Prevention, there were 30,006 unintentional drug overdose deaths. Looking at prescription drug deaths, 75% of the over 22,000 deaths involved opioid analgesics, which are therapeutic drugs directly metabolized by CYP2D6 (Drug Overdose in the United States: Fact Sheet). These numbers are rising as drugs are becoming more available, diverse and affordable. The negative effects of unintentional overdoses can be more efficiently combated if a more personalized form of a medical treatment were to become widely available.

Upon first analysis, it might appear that the best way to prescribe a personalized drug dose that matches a given patient's genetically controlled rate of drug breakdown would be to measure enzyme levels directly. However, methodologies to directly analyze genotype will likely provide a more direct, less costly, and more accurate determination (Shendure et al, 2004). In addition, in order to truly understand the fundamental basis of phenotypes, genotyping and sequencing of a genomic locus is necessary (Cosart et al, 2011). For over two decades, DNA sequencing has been completed almost solely through capillary-based, semi-automated implementations of Sanger Sequencing, and it is considered to be the gold standard of sequencing. Sanger sequencing can be done using more than one method, but in this study, primers were used to flank the specific region of CYP2D6 and were amplified with PCR. By using fluorescent dideoxynucleotide triphosphates (ddNTPs), which randomly insert as DNA is proliferating, an extremely high-resolution electrophoretic separation of the DNA allows the sequence to be read

(Shendure, 2008). This capillary process essentially works like gel electrophoresis where the shorter strands of DNA migrate more quickly than the larger strands. This separation allows the machine to read the fluorescently labeled ddNTP at the end of the strand and make a nucleotide call for that specific location on the DNA.

Sanger sequencing has been continually improved and studied since the method was invented over three decades ago. Due to this continuous tweaking and modifying, it is now possible to make nucleotide calls with raw accuracy of up to 99.999% (Shendure, 2008). It has also become much more affordable to sequence DNA as costs per megabase are said to be approximately $0.10 (Shendure, 2012) and continue to plummet at exponential rates. With this accuracy at such a cheap cost it may prompt the question as to why there is even the need to develop alternative methods. That very simple answer lies in time. Sanger sequencing is not costly and is very accurate, but the time it would take to sequence and interpret the sequencing data of one gene locus can be longer than one week. If one had a goal to use sequence information for therapeutic treatment, the results may not come back in time to be useful. Simply put, NGS approaches stand the promise of being orders of magnitude more rapid than Sanger-based methodologies.

Sequencing data will only become more valuable as time and effort are continuously put into genomics research. The Sanger method is rapidly being replaced by Next Generation Sequencing (NGS), a series of technologies, some of which, that rely on sequence data being aligned to templates and referenced to a whole genome alignment (Metzker, 2010). There are a number of different methods that NGS machines employ to sequence DNA as the technology has been a huge focal point in the recent past. In fact, major advancements have allowed for an exponential increase in the amount of data that

is able to be collected. Some machines are able to read over a billion segments per instrument (Metzker, 2010). This huge amount of information has caused advancements that have resulted in treatments for disease, preventative medicine and to help aid in the patient treatment. Through studying the phenotypes of specific genes by using genotyping and sequencing data, we will be able to find gene loci that contribute to variable expression (Shendure, 2004). In the case of CYP2D6, variable expression can lead to various metabolic properties and different therapeutic relief one receives from a certain drug.

The longer a technology is around, the more efficient and cheaper it has the ability to become. When the healthcare spending per capita was amortized over the average lifespan for an individual, it was extrapolated that if one spent just $1,000 for a personalized genome sequence, the information would only have to do $13 of benefit per year for the benefit to outweigh the cost (Shendure, 2004). As the sequence information becomes more and more familiar to researchers and doctors, this information will easily surpass the cost of $13 per year to be beneficial.

While it appears as though it would be completely logical to use NGS for clinical applications, it has not yet been proven as the gold standard in the field. It has been reported that NGS is able to have a high concordance value for calling single nucleotide polymorphisms (SNPs) at greater than 99.5% and having a 2.5% false-positive call rate with novel SNPs (Metzker, 2010). Much of the difficulty comes in AT-rich regions where specific numbers of thymine residues is often mistaken (Ruan et al, 3013). Other issues with NGS have to do with its difficulty sequencing highly complex regions in DNA. One of these highly complex regions is the CYP2D6 region of chromosome 22.

Some of the problems areas on the DNA include small insertions, transposition events, copy-number variants (CNVs), tandem repeat expansions and copy-neutral rearrangements (Shendure, 2012). The CYP2D6 region on chromosome 22 is highly complex and often has CNVs that were taken into account when the DNA was sequenced and compared back to previous genotyping data as well as Sanger sequencing.

As previously stated, CYP2D6 is a highly polymorphic region of the human genome, but through years of study and assay development immense knowledge has been amassed. Researchers across the globe have focused on CYP2D6 due to its high importance in therapeutic drug metabolism for many different drugs on the market. Despite our extensive knowledge of CYP2D6, the use of Sanger sequencing may delay physicians' ability to prescribe appropriate dosages that result from genetic heterogeneity. The purpose of the current project was to determine if the faster Next-Generation Sequencing technology can accurately ascertain the CYP2D6 genotype. This was a particular challenge since the gene locus is highly polymorphic and contains many CNVs, we were unsure if the data would be accurate. We hypothesized that due to the new paired-end sequencing technology available our lab, that the correct variant calls would be able to be made and would compare to the Sanger sequencing gold standard and previous genotyping data.

By using Sanger Sequencing, genotyping data and Next Generation Sequencing data, this project attempted to discover if the new technology is able to call the variant of CYP2D6 allele to the gold standards already present. Samples taken from Children's Mercy Hospital patients and the Coriell Institute were used for sequencing and study. Many of the samples have had previous assays and genotyping studies done on them to

help confirm the variant calls made in this particular study. If we can prove that NGS can accurately call the allelic variants, we hope to see a move to personalized medicine that will adjust medicine dosages to fit the metabolic rate of each individual patient. This would lead to a more effective use of medicine as well as prevent malicious side effects and consequences of the misuse of drugs and even may save the life of children receiving treatment.

MATERIALS AND METHODS

Study Subjects and Samples

The samples for this study were obtained from two separate populations of patients. The first cohort consisted of CMH patients. There were 7 of these patients and their samples were collected anonymously prior to this study. The second cohort consisted of NA and UDT samples, and were obtained from the Coriell Institute. There were a total of 5 of these samples. All of the samples had various studies completed on their genomes.

PCR Reaction forming 6.6 kb XL Fragment

The long PCR reaction was completed using predetermined primers that flanked the CYP2D6 region. The combination of the 5'2D6 primer, which is 5'CCA GAA GGC TTT GCA GGC TTC AG and the 3'2D6 primer, which is 5'ACT GAG, CCC TGG GAG GTA GGT AG made an XL PCR product that is 6.6 kbp in length (Gaedigk, 2002). In each 8 μL reaction there was:

0.025 units KAPA LR DNA Polymerase

1.6 μL 5x PCR buffer

2.66 μL 15% DMSO

0.24 µL 10 mM dNTPs

0.4 µL of each 10 µM primer

1 µL genomic DNA (10-20 ng/µL)

The PCR cycling conditions were as follows:

Denaturation:

Three minutes at 94°C

Cycle Sequence:

35 cycles:

Denaturation: 20s at 94°C

Annealing & Extension combined: 7.5 min at 68°C

Final Extension: 7 min at 68°C

Holding temperature: 10°C (Gaedigk, 2012)

The product was then stored in 4 degrees Celsius refrigerator and then prepared to insure 6.6 kb XL PCR product was produced for Sanger Sequencing.

Preparation of 6.6 kb XL PCR Products

The presumed 6.6 kbp XL PCR Product was run on a 0.7% agarose gel electrophoresis to determine if the sample succeeded in forming the segment. Amplification was determined by the intensity of the 6.6 kbp band and its integrity of the DNA was determined by the location at approximately 6.6 kbp. 1.5 µL of the PCR product was used when the gel electrophoresis was completed. The 6.6 kbp (XL) fragment of sample DNA was then purified through column purification using a GenElute™ PCR Clean-Up Kit (Sigma-Aldrich). The protocol provided with the kit was followed for the purification process. Once purified, the DNA concentration was

determined using a Nanodrop Spectrophotometer. Once the concentration was determined, samples were diluted to approximately 10 ng/μL using DI water. The 10 ng/μL DNA was used for the PCR Sanger sequencing reaction.

PCR Sanger Sequence Reaction

The 6.6kbp (XL) fragments that were obtained using the specific CPY2D6 p450 primers described above were then multiplied using the sequencing PCR reaction in order to use the ABI3739xl DNA Analyzer for sequencing data. The sequencing reaction was set up using 96-well plates (Midsci, St Louis, MO, USA) and 2-3 samples were run per plate per trial. First, the master mix was made for each individual DNA sample, which did not contain the primers for the sequencing reaction. This master mix consisted of enough solution to run 28 different primers (15 forward, 13 reverse) plus some excess solution for measurement errors. The master mix was divided into 7.23 μL aliquots in each well for 28 wells. The total master mix volume per DNA sample was 227.5 μL (82.85 μL H2O; 90.72 μL 5X Buffer; 10.08 μL BD; 12.60 μL 100% DMSA; 31.50 μL DNA). Once distributed, each reaction consisted of:

2.63 μL H_20

2.88 μL 5X Buffer

0.32 μL Binding Dye

0.40 μL DMSO (100%)

1.00 μL DNA (10 ng/μL)

0.77 μL Primer

The 15 forward primers and 13 reverse primers were used to cover the entire 6.6 kbp (XL) fragment and their accurate lengths ranged from approximately 200-800 base pairs. The forward primers used were 5'int28F, 5'E1seq, 5'intron1F, 5'971, 5'E2seq, 5'2d6_3F, 5'1846, 5'E3seq, 5'E4_ins/3, 5'E5seq/3, 5'RT-11F, 5'E6seq, 5'E7seq/2, 5'RT-30F and 5'RT-16F. The reverse primers used were 3'int8, 3'I2Eco, 3'intron1R, 3'E2seq/2, 3'2D6_copy_6R, 3'RT-24R, 3'2061, 3'RT-7R, 3'E5seq, 3'intron6_seq, 3'RT-43R, 3'E8seq and 3'E9seq/2. All of the primers were provided based on their effectiveness determined from previous studies.

The PCR was run on a MJ Research Dyad PCR Dual Block machine and the PCR conditions were:

Denaturation:

Two minutes at 95°C

Cycle Sequence:

Twenty-two cycles:

Denaturation: 10s at 96°C

Annealing: 5s at 50°C

Extension: 4min at 60°C

Holding temperature: 4°C

The 96-well plate was then spun down to have the DNA collect at the bottom of each well so the capillary electrophoresis would be effective. The 96-well plates were then put in the ABI3730xl DNA Analyzer and 96-well capillary electrophoresis was completed. The data was imported into Sequencher (Gene Codes Corporation) for DNA nucleotide calls.

DNA Sequence Determination using Sequencher Program

Sequencher (Gene Codes Corporation) was the software that the capillary electrophoresis sequencing data utilized to call the nucleotides on the DNA fragments. Once the data was imported it was lined up against the AY545216 sequence, which is a known CYP2D6*1 allelic variant. This AY545216 sequence is used as a baseline by which to call single nucleotide polymorphisms (SNPs) that differ from the *1 genotype. Hg19, the most up to date complete human genome at the time is a *2 genotype so it was not used for SNP calls.

Once the sequences were lined up against the AY545216 sequence, the confidence of the program for calling a specific nucleotide was denoted by the sequence having a white background. On more blue, or uncertain nucleotide calls, the multiple fragments were assessed and a nucleotide call was made. A compilation of SNPs was then compared to industry-accepted genotypes to determine the specific genotype of the patient. Each of the chromosomes was designated a genotype for each sample.

Copy Number Detection using Multiplex PCR Amplification

The copy number detection assay was used to determine the number of *CYP 2D6* alleles present in a given sample. PCR was run using 96-well plates (MidSci, St Louis, MO, USA) that were placed in an Eppendorf Mastercycler® ep Gradient S instrument. The polymerase chain reaction was composed of the following reagents: 0.4 units KAPA2G™ Fast HotStart DNA Polymerase (KAPA Biosystems), 2.4 µl supplied 5x PCR buffer (1.5x final concentration following the recommendation of the KAPA2G Fast HotStart application note for multiplex PCR V2.08), 0.4 µl 100% DMSO (dimethyl sulfoxide), 0.16- µl 10-mM dNTPs and 1 µl gDNA (5-20 ng/µl). The specific regions of

the *CYP 2D6* gene locus amplified were found in intron 6 and exon 1. These assays contained a 0.5 μM concentration of one set of primers. The PCR conditions for the MPA assay were as follows (Gaedigk, 2012):

Denaturation:

Two minutes at 95°C

Cycle Sequence:

Twenty-two cycles:

Denaturation: 15s at 95°C

Annealing: 30s at 60°C

Extension: 10s at 72°C (a final extension of 3 min at 72°C)

Holding temperature: 10°C

After the PCR was completed, 1 μl of the product was transferred into a new 96-well plate containing 0.2 μl of GeneScan® 500(-250) LIZ® size standard and 6.8 μl of HiDi (Applied Biosystems). This solution was incubated at 95°C for 5 min and cooled to 4°C for at least 2 min. The ABI 3730 DNA analyzer was then used to separate out the PCR products and the software GeneMapper® Version 4.0 (Applied Biosystems).

Next Gen Genome and Exome Sequencing

The DNA that was prepared for Whole Genome Sequencing was completed using the Illumina TruSeq sample preparation (Illumina). The process was initiated with 500 ng of DNA, which was sheared with Covaris S2 Biodisruptor, end-repaired, A-tailed, and adaptor-ligated. The preparation was completed without the use of any polymerase chain reactions (PCR). The libraries were then purified with SPRI beads. Next, PCR was used to quantify the DNA. Once the amount of DNA was determined, a 0.1 M NaOH solution

was used to denature the sample. It was then diluted with the use of a 2.8 pM buffer solution. Once the DNA was prepared, it was loaded into flowcells in the Illumina HiSeq 2500 instruments. These instruments, used in rapid run mode, automatically sequenced the DNA by cluster generation, followed by 2 x 100 cycle sequencing reads separated by paired-end turnaround.

The DNA that was prepared for exome sequencing was done in a different way. Illumina TruSeq protocol was used to prepare the DNA to be selected for a custom gene panel. 20,477 eighty-nucleotide probes for 8366 genomic regions were used to select for 8366 genomic regions that consisted of exons and intron-exon borders (Saunders, 2012). Among the regions selected was the *CYP 2D6* p450 gene locus. The sequencing was then completed on HiSeq 2000 instruments with TruSeq v3 reagents. Exons present in the samples were enriched twice through Illumina protocols to a depth of >8 GB of singleton 100-bp reads per sample.

The sequencing performed did not comply with the routine diagnostic tests defined by CLIA as the sequencing was performed for research purposes.

Next Gen Sequencing analysis

The sequences attained were referenced to Hg19, the most updated whole genome sequences at the time of experimentation. The interpretations were completed using CASAVA 1.8.2 (Illumina) software using gapped ELAND alignment. FASTQ used base-call quality scores initially called the base pair variant. This information was then compressed into binary code and annotated by RUNES, a variant characterization software (Saunders, 2012). Once compressed, the information was imported into Integrated Genomic Viewer (IGV) to view the data. The information compiled in this

format was compared to Sanger sequenced data completed at Children's Mercy to compare variant calls of the *CYP 2D6* p450 gene locus.

Using the Next Gen Compiled Data

Once the data was imported into IGV and interpreted, a Microsoft Excel® Spreadsheet was used to assess the data. The Next Gen sequences were aligned to Hg19, the most up to date human genome sequence template. Hg19, however, is a *CYP2D6*2* genotype and for the *CYP2D6* gene, allelic variants are always calls in reference to *CYP2D6*1* variants. Working through each base pair, the Next Gen sequences were converted back to sequences that could then be assessed using the *CYP2D6*1* genotype. All Sanger Sequencing was assessed in reference to *CYP2D6*1* (AY545216) sequence alignment. In conclusion, the *CYP2D6*1* (AY545216) sequence was employed to reference all sequence calls.

RESULTS

Sample Preparation

For all 12 samples, after the PCR reaction to form the 6.6 kbp XL PCR fragment was completed, the samples were run on a 0.7% agarose gel to check for DNA integrity. Often times, due to mistakes in making the gel or a failed PCR reaction, the gel did not run correctly. Through going back and concentrating on master mix procedures and gel procedures correct gels could be run for every single sample. An example of a correct sample is shown in Figure 1.

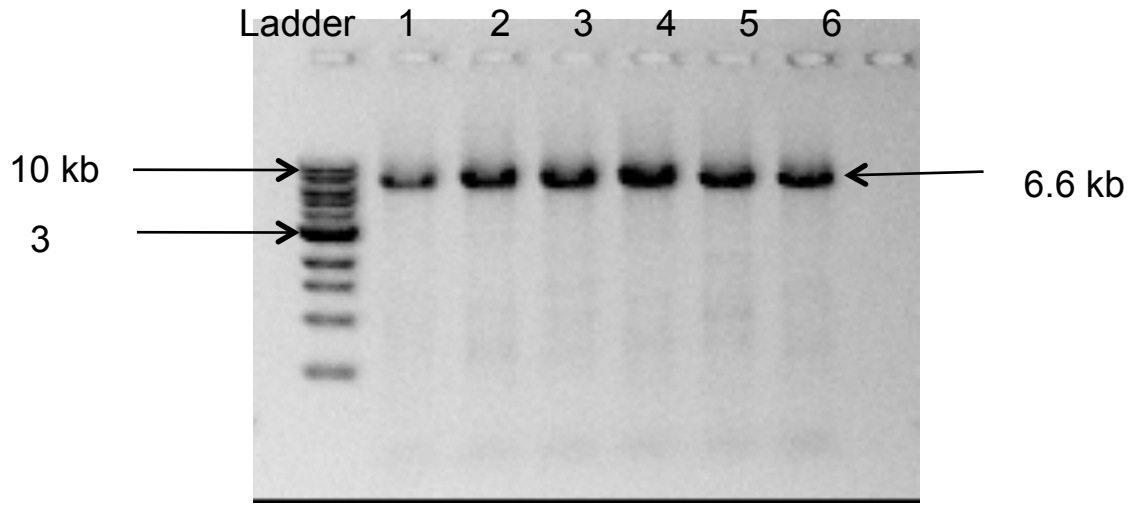

Figure 1. Example of successful production of 6.6 kbp XL PCR fragment using gel electrophoresis. The lane marked ladder contained size standards. Lanes 1-6 contained six different samples. Neg was a negative control in which PCR was run in the absence of sample DNA.

Nanodrop Instruments were then used to determine the amount of DNA present in the sample. Samples generally ranged from a 10-200 ng/µL concentration of DNA that were present. Regardless of the amount of DNA present, they were all diluted to a concentration of 10 ng/µL for sequencing purposes.

Sanger Sequencing

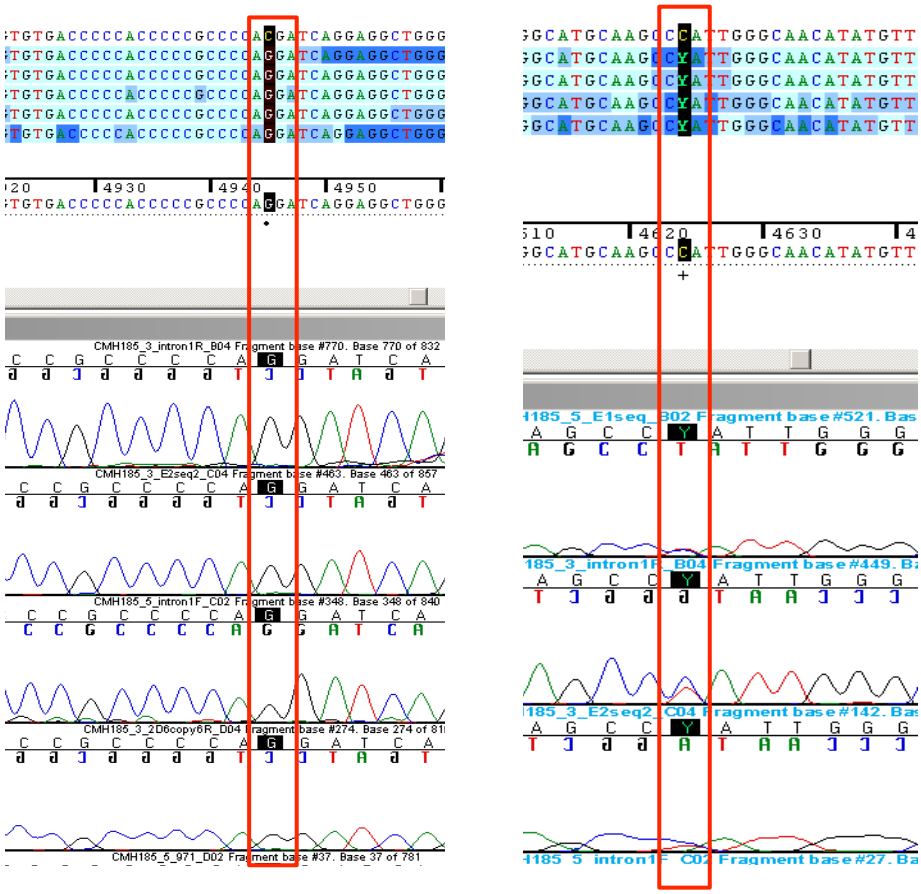

Figure 2: Examples of a Sequencher homozygous SNP (left) and a heterozygous SNP (right).

Results compiled from Sanger Sequencing data were imported into Sequencher, a DNA analyzing technology to compare the DNA to AY545216, a *1 variant of the CYP2D6 allele. The software was in most cases able to call the nucleotide for each position on the gene. In cases of adenine, guanine, cytosine and thymine, letters A, G, C and T were assigned, respectively. Heterozygous SNPs were typically called with substitute letters. When the software was unable to make a nucleotide call due to too much background noise, the data was assessed and assigned a nucleotide call. In the Sequencher program, the closer to white the background of a letter was, the more accuracy this nucleotide call could be made with. Places of darker blue correspond to areas with a lot of background

noise or of a place with a heterozygous SNP. Examples of homozygous and heterozygous SNPs are found in Figure 2.

Next Gen Sequencing

Next Gen Sequencing Results were provided in a Microsoft Excel document that had the list of the entire genome of each individual. The specific locus of the gene on Chromosome 22 was identified and mapped the SNPs were found with relation to Hg19. Hg19 is a *2 variant of the CYP2D6 allele so it was necessary to first map this back to AY545216 in order to determine if the SNPs present in each patient were different from a *1 allele, not a *2.

Sequence Compilation

All sequencing results were compiled onto a Microsoft Excel document that had already been started with previous sequencing data (Tables 1A and 1B for the first six patients; Tables 2A and 2B for the next six patients). Each patient has 3-4 rows, and the table is blocked by darkened lines that delineate a specific patient. The top line represents the Next Gen sequencing data. The second line represents both strands compiled as one from the Sanger Sequencing data. If there are no more rows after the second, then the patient was homozygous for the particular allele. If there are two more rows associated with the patient, however, then each individual allele was broken down into its own row. This was done using patterns of which SNPs are associated with each other from past data and documented literature.

Blacked out boxes represent the particular SNP that varies from AY545216. Yellowed out boxes were not sequenced within the 6.6 kbp XL PCR fragment. Grey boxes with the letter H in them represent that the patient was heterozygous at that

particular SNP. If there is a G in any box, it was confirmed by genotyping assays completed in the lab either during the study or at a previous time. At the bottom of the document, there is a template that shows where each SNP is located as far as the intron, exon, upstream or downstream region of the gene. The allele and subject are located in columns on each end of the sequencing data and it gives information that was deduced from the nomenclature website as well as previous literature on the CYP2D6 gene.

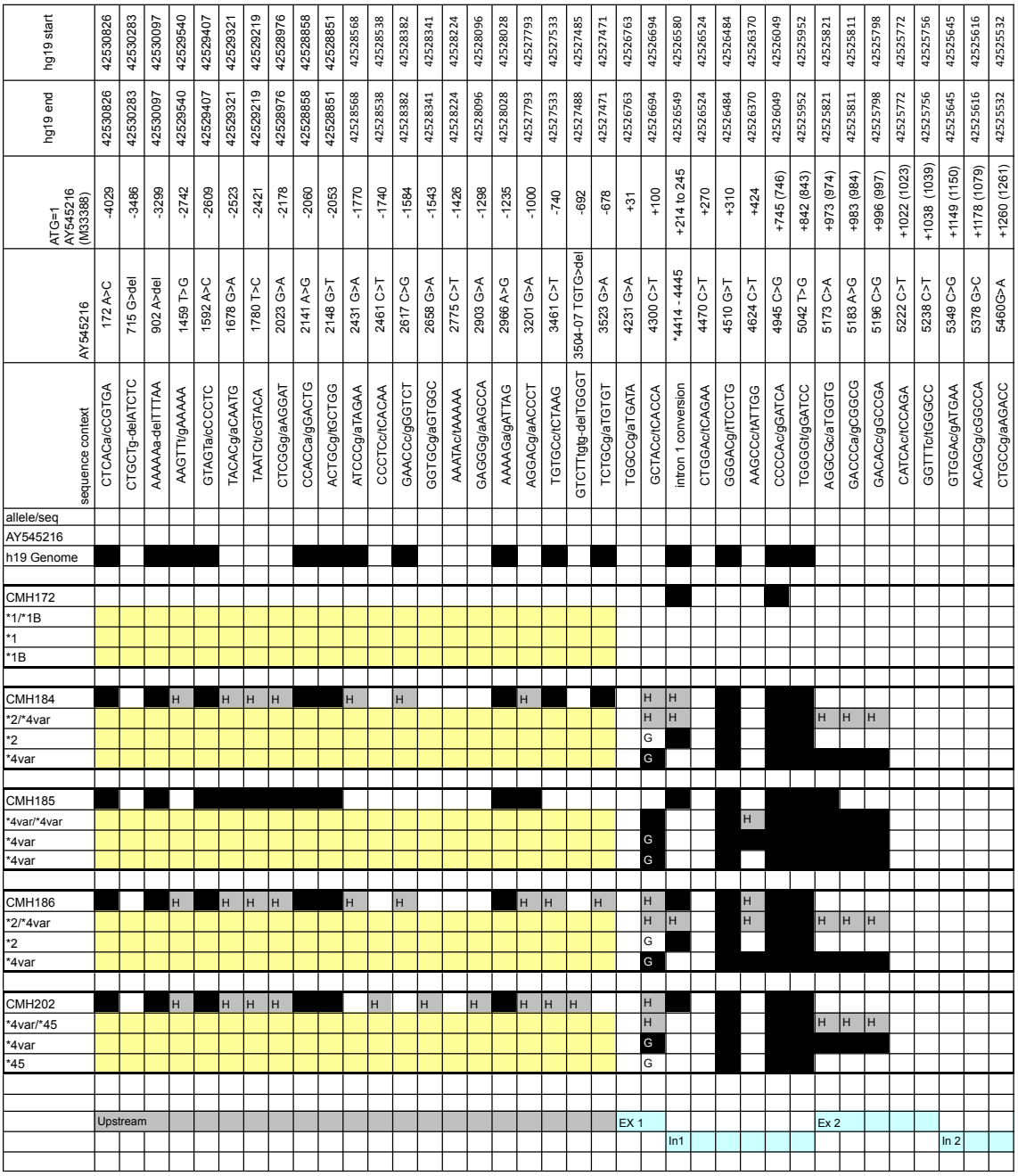

Table 1A: Sequencing Results from patients CMH182, CMH184, CMH185, CMH186, CMH202.

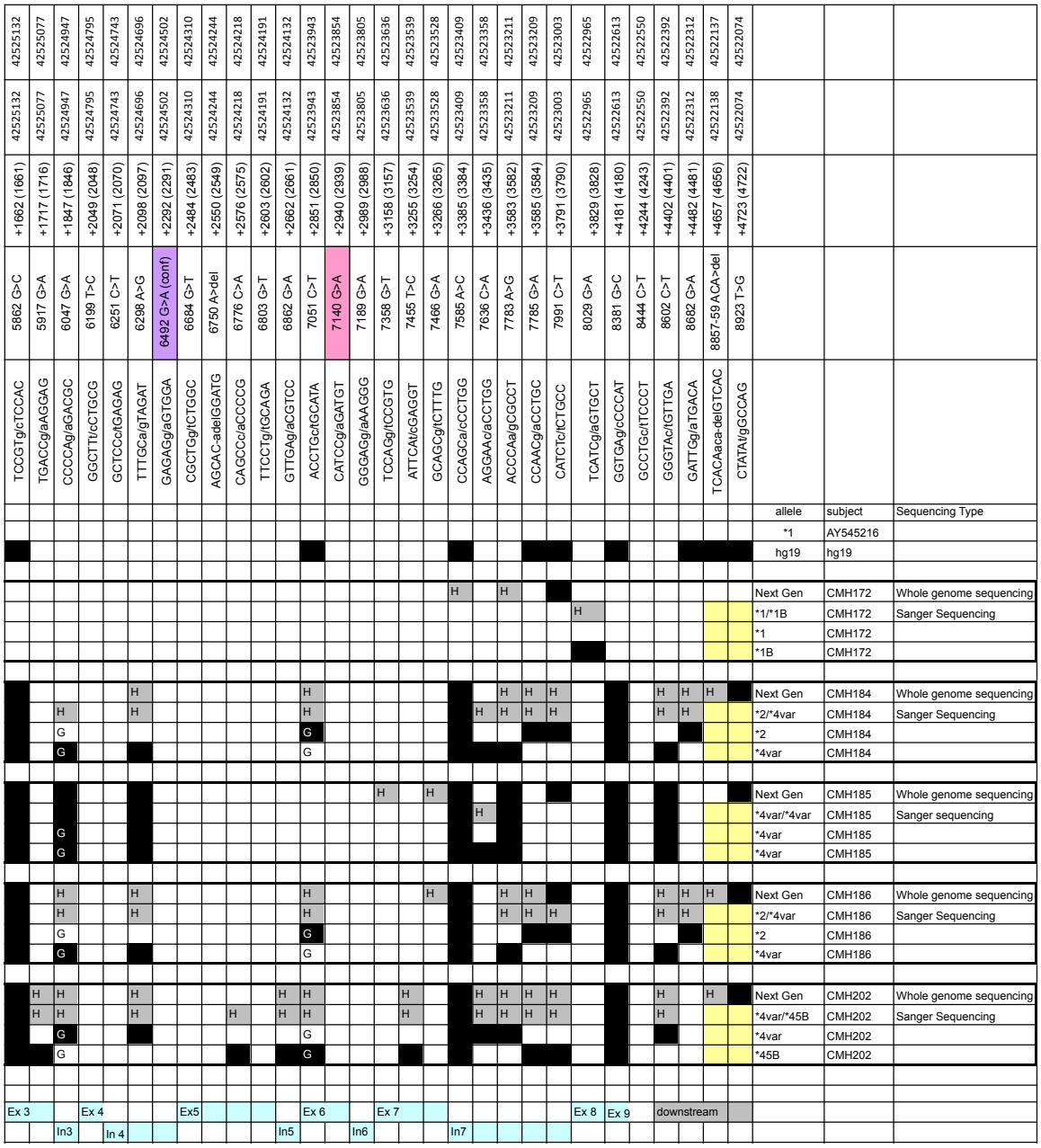

Table 1B: Sequencing results from patients CMH182, CMH184, CMH185, CMH186, CMH202.

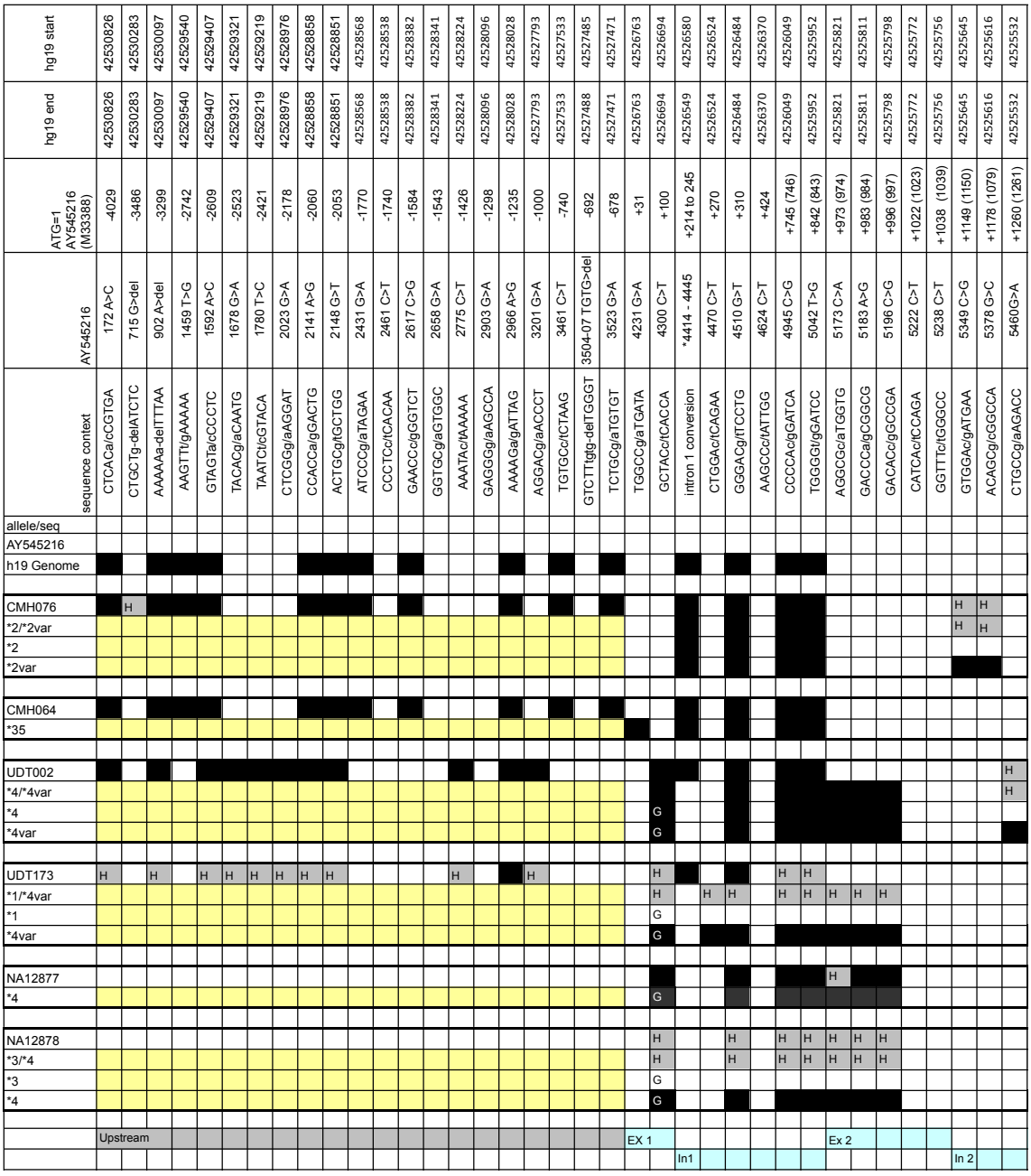

Table 2A: Sequencing results from patients CMH076, CMH064, UDT002, UDT173, NA12887, NA12878.

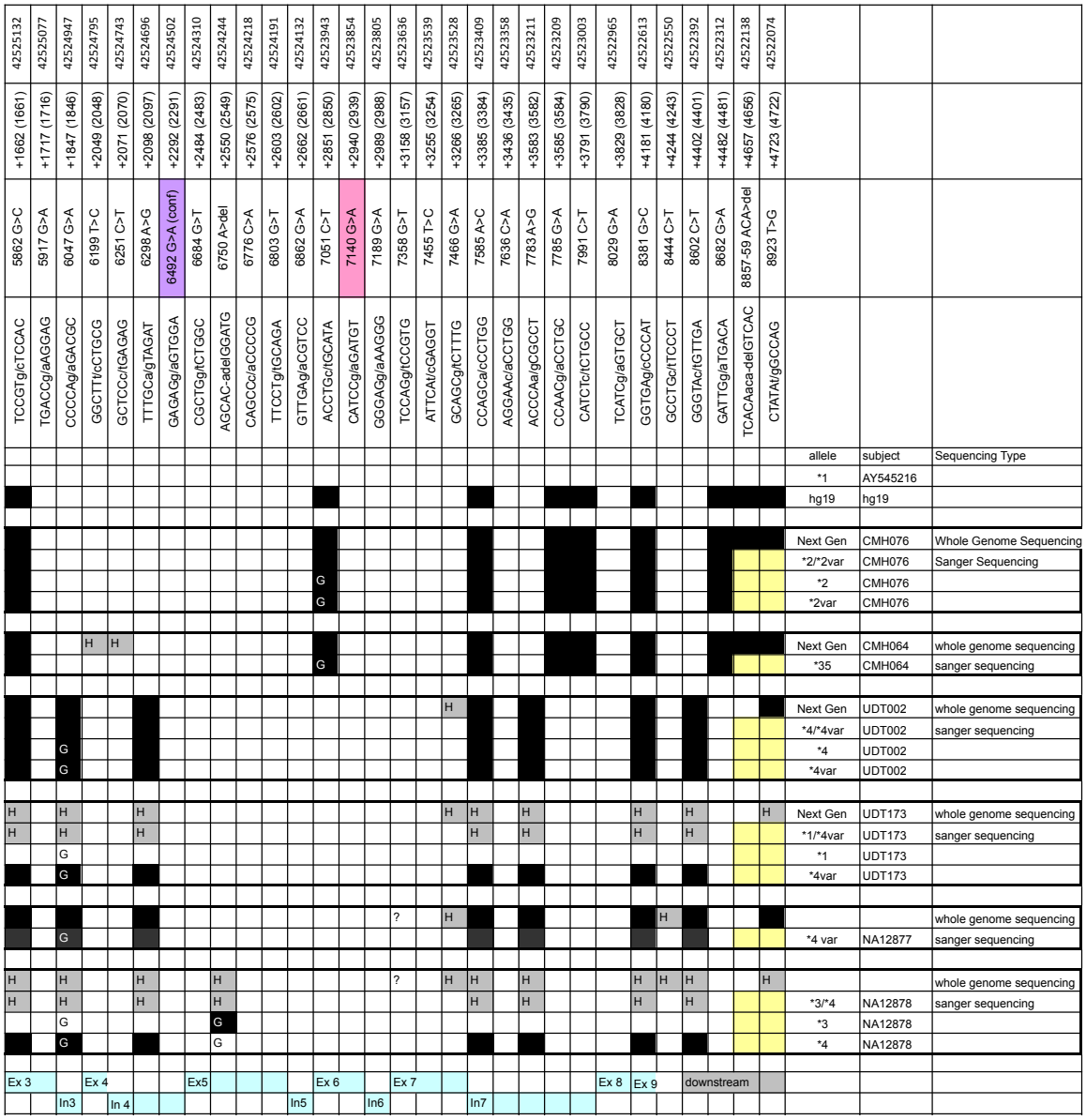

Table 2B: Sequencing results from patients CMH076, CMH064, UDT002, UDT173, NA12887, NA12878.

DISCUSSION

The goal of this project was to determine if the new Next Generation Sequencing technology was able to accurately call SNPs in the highly polymorphic CYP2D6 gene in order to determine the correct genotype for the patient with a high degree of confidence. If able to do so, this genotype information could be used to determine the patient's

phenotype and adjust dosages of therapeutic medication in a short period of time. This information could help progress treatment to personalized medicine that could help better treat patients by dosing more effectively as well as prevent toxic levels of certain drugs from accumulating in the blood.

Despite a learning curve associated with laboratory techniques early on in the study, there were rarely problems with the generation of the 6.6kbp XL PCR fragment. Since Dr. Gaedigk had mastered this technique through many years of research and study on the CYP2D6 gene, the protocols worked very well. The clean up of this PCR product as well as the dilution to approximately 10 ng/μL concentration of DNA went off nearly without a hitch. The Copy Number Variation (CNV) multiplex PCR method was completed working closely with Greyson Twist, who developed this technique and had mastered the protocol. Therefore, primarily Greyson dealt with problems dealing with errors in the CNV assay as it was relatively new and he was the expert.

A major issue was to insure that forward and reverse primers covered the entire PCR product, as there are two strands to DNA, such that sequence data could be obtained for the entirety of both strands. Early on 12 forward and 11 reverse primers were used to cover the entire CYP2D6 gene. However, these early trials ended up showing gaps in the coverage of the CYP2D6 sequence. Therefore, it was decided that three more forward primers and two more reverse primers should be added to completely cover the gene. Some sections of the CYP2D6 gene are extremely similar to regions of the CYP2D7 gene, therefore sequences can be lost or the coverage could be erroneous. This similarity between the two genes is a major reason why there was a question as to whether or not Next Gen Sequencing technology would be able to accurately sequence the CYP2D6

gene. With the paired-end method, the technique used by our NGS machines, it was a concern that sequences would be lost from 2D6 to 2D7 and visa versa.

Compiling the sequencing data for the Sanger sequencing consisted of importing the sequence reads into the Sequencher software and systematically going down the entire gene to determine specific nucleotide calls. Often times background noise would inhibit the software from being able to make a nucleotide call, but the sequence information was still very clear. Also, the software occasionally wasn't be able to recognize a heterozygous SNP because the dominance of one strand over the other. In cases such as these, nucleotide calls were manually entered into the Sequencher program. To make these calls, haplotype patterns were used from previous studies and the nomenclature from the CYP2D6 website. Once the entire gene was covered both forward and reverse, the data was manually translated to the Microsoft Excel document in order to simplify the comparison process. While accurate, this was time consuming and therefore would not be ideal in a clinical setting.

Next Generation sequence data came with its own set of nuances as the Hg19 genome is a *2 allelic variant of the CYP2D6 gene. This means that there are several SNPs that are different from the *1 allele that is used for comparison in the AY545216 strand. The difficulty in this was that SNPs provided from the Next Gen data were all with regard to the *2 Hg19 strand where they needed to be to the *1 AY545216 strand. Therefore it was necessary to work backwards to compare the Hg19 strand with the AY545216 strand and then make the SNP calls with regard to the AY545216 strand. Despite this extra work, all of the data were eventually translated onto the Microsoft

Excel document, at which time the data could be assessed to determine the accuracy of the Next Gen Technology was.

In Table 3, the various SNPs found in each method are put into three separate categories. In column two, SNPs only found in Next Gen were tallied, in column three, SNPs only found in Sanger Sequencing were tallied; and, in column four, SNPs shared by both methods were tallied. It is apparent that there are discrepancies with the data as the Next Gen and Sanger results do not identically align.

Sample ID	# SNPs only in Next Gen	# SNPs only in Sanger	# SNPs shared
CMH064	2	1	11
CMH076	0	0	13
CMH172	5	1	0
CMH184	0	5	14
CMH185	4	5	11
CMH186	3	5	15
CMH202	1	4	18
UDT002	1	3	12
UDT173	3	5	10
NA12877	2	0	15
NA12878	2	0	15
NA12882	1	1	15

Table 3: Comparison of SNPs found in Next Gen and Sanger Sequencing.

Discrepancies in the data can be as a result of many things. In some cases the Next Gen technology did not distinguish between a heterozygous and homozygous SNP, in which case it was considered an error, but one that is easily recognizable. In other cases one of the methods was incorrect due to a complicated region of the gene. While these results seem to point to the fact that Next Gen technology would be unable to accurately call the genotype of the gene, this was not the case for the most part.

When looking at haplotypes that are present in certain allelic variants, it is a group of SNPs that are most significant in making the call that corresponds to a particular phenotype of the CYP2D6 gene. Therefore, when the SNPs were compared to the nomenclature of past literature, the following genotype calls were made. This information is found in Table 4.

Sample ID	Genotyping by Next Gen Method	Genotyping by Sanger Method	Genotyping Analysis
CMH064	***2/*2**	***35/*35**	***2/*5**
CMH076	*2/*2	*2/*2	*2/*2
CMH172	*1/*1	*1/*1	*1/*1
CMH184	*2/*4	*2/*4	*2/*4
CMH185	***4/*4**	***4/*4**	***4/*68+*4**
CMH186	***2/*4**	***2/*4**	***2/*68+*4**
CMH202	*4/*45	*4/*45	*4/*45
UDT002	***4/*4**	***4/*4**	***4/*68+*4**
UDT173	***1/*4**	***1/*4**	***1/*68+*4**
NA12877	*4/*4	*4/*4	*4/*4
NA12878	*3/*4	*3/*4	*3/*4
NA12882	*4/*4	*4/*4	*4/*4

Table 4: Allele assignment summary for Next Gen, Sanger and Genotyping assays.

Rows that are bolded are patient calls that were incorrect with regards to the genotype call that was made. Still, the data would appear to point to the fact that Next Generation technology was unable to call the correct genotype for every patient. The information that is the most important, however, is if the phenotype matches for each method, as this is the crucial clinical significance. When looking at the phenotypes of the calls made, they all match for every single method. The alleles with a *68 present represents a hybrid 2D6/2D7 gene duplication that does not effect the overall metabolic

rate of an individual. This can be explained by the similarity between the 2D6 and 2D7 regions that were lost by both the Sanger and the Next Gen technologies. Since this tandem is only found through genotyping and does not have an effect on the phenotype of the individual, it was not a cause for concern. Therefore, it was proven that the each patient was placed in the correct metabolic family as far as normal function, reduced function and non-function of the CYP2D6 gene.

Moving forward, seven more patients DNA was assessed using all three methods of study and every single sample was placed into the correct phenotypic group. This conclusion is promising as we are one step closer to moving towards a personalized approach to treatment using the Next Generation Sequencing technology. Further studies will continue to test more complex variants of the CYP2D6 gene to see if every allelic variant will be able to be accurately placed before any clinical result will come from our promising results.

REFERENCES

Cosart, Ted, et al. “Exome-Wide DNA Capture and Next Generation Sequencing in Domestic And Wild Species.”*BMC Genomics* 12.(2011): 347. *MEDLINE with Full Text.*Web. 11. Feb. 2014.

Gaedigk, Andrea. "CYP2D6, SULT1A1 and UGT2B17 copy number variation quantitative detection by multiplex PCR." *Pharmacogenomics*. 13.1 (2012): 91-111. Print.

Gaedigk, Andrea., et al. “Cytochrome P4502D6 (CYP2D6) Gene Locus Heterogeneity: Characterization of Gene Duplication Events”. *Clinical Pharmacology and Therapeutics.* 81.2 (2007). 242-251.

Gaedigk, Andrea. “Unique CYP2D6 activity distribution and genotype-phenotype discordance in black Americans.” *Clinical Pharmacology and Therapeutics*. 72.1 (2002) 76-89.

Marechal, J-D, et al. “Insights Into Drug Metabolism By Cytochromes P450 From Modeling Studies Of CYP2D6-Drug Interactions.” *British Journal Of Pharmacology* 153 Suppl 1.(2008): S82-S89. *MEDLINE with Full Text.* Web. 5 Feb, 2014.

Metzker, Michael L. “Sequencing Technologies – The Next Generation.” *Nature Reviews. Genetics* 11.1 (2010): 31-46. *MEDLINE with Full Text.* Web. 12 Feb. 2014.

Niwa, Toshiro, Norie Murayama, and Hiroshi Yamazaki. “Comparison Of Cytochrome P450 2D6 And Variants In Terms Of Drug Oxidation Rates and Substrate

Inhibition." *Current Drug Metabolism* 12.5 (2011): 412-435. *MEDLINE*. Web. 5 Feb. 2014.

Ruan, Jue et al. "Pseudo-Sanger Sequencing: Massively Parallel Production of Long and Near Error-Free Reads Using NGS Technology." *BMC Genomics* 14.(2013): 711. Web.11 Feb. 2014.

Saunders, Carol Jean. "Rapid Whole-Genome Sequencing for Genetic Disease Diagnosis in Neonatal Intensive Care Units." *Science Translational Medicine*. 4.154 (2012): n. page. Print.

Shendure, Jay, et al. "Advanced Sequencing Technologies: Methods and Goals." *Nature Reviews. Genetics* 5.5(2004): 335-344. *MEDLINE with Full Text.* Web. 12 Feb. 2014.

Shendure, Jay and Erez Lieberman Aiden. " The Expanding Scope of DNA Sequencing." *Nature Biotechnology* 30.11 (2012): 1084-1094. Web. 11 Feb. 2014.

Shendure, Jay, and Hanlee Ji. "Next-Generation DNA Sequencing." *Nature Biotechnology* 26.10 (2008): 1135-1145. *MEDLINE with Full Text.* Web. 11 Feb. 2014.

United States. Centers for Disease Control and Prevention. *Drug Overdose in the United States: Fact Sheet*. 2013. Web. <http://www.cdc.gov/homeandrecreationalsafety/overdose/facts.html>.