HHS Public Access

Author manuscript

Circ Cardiovasc Genet. Author manuscript; available in PMC 2015 April 22.

Published in final edited form as:

Circ Cardiovasc Genet. 2015 April ; 8(2): 343–350. doi:10.1161/CIRCGENETICS.114.000776.

Exome Sequencing in Suspected Monogenic Dyslipidemias

Nathan O. Stitziel, M.D., Ph.D.[1,2,25], **Gina M. Peloso, Ph.D.**[3,4,5,25], **Marianne Abifadel, Ph.D.**[6,7], **Angelo B. Cefalu, M.D., Ph.D.**[8], **Sigrid Fouchier, Ph.D.**[9], **M. Mahdi Motazacker, Ph.D.**[9], **Hayato Tada, M.D., Ph.D.**[3,4], **Daniel B. Larach, B.A.**[10], **Zuhier Awan, M.D.**[11], **Jorge F. Haller, Ph.D.**[12,13,14], **Clive R. Pullinger, Ph.D.**[15], **Mathilde Varret, M.D.**[6], **Jean-Pierre Rabès, M.D., Ph.D.**[6,16], **Davide Noto, M.D., Ph.D.**[8], **Patrizia Tarugi, Ph.D.**[17], **Masa-aki Kawashiri, M.D.**[18], **Atsushi Nohara, M.D.**[19], **Masakazu Yamagishi, M.D.**[18], **Marjorie Risman, M.S.**[10], **Rahul Deo, M.D., Ph.D.**[15], **Isabelle Ruel, Ph.D.**[11], **Jay Shendure, M.D., Ph.D.**[20], **Deborah A. Nickerson, Ph.D.**[20], **James G. Wilson, M.D.**[21], **Stephen S. Rich, Ph.D.**[22], **Namrata Gupta, Ph.D.**[4], **Deborah N. Farlow, Ph.D.**[4], **NHLBI Grand Opportunity Exome Sequencing Project Family Studies Project Team**, **Benjamin M. Neale, Ph.D.**[3,4,23], **Mark J. Daly, Ph.D.**[3,4,23], **John P. Kane, M.D., Ph.D.**[15], **Mason W. Freeman, M.D.**[12,13,14], **Jacques Genest, M.D.**[11], **Daniel J. Rader, M.D.**[10], **Hiroshi Mabuchi, M.D.**[19], **John J.P. Kastelein, M.D., Ph.D.**[24], **G. Kees Hovingh, M.D., Ph.D.**[24], **Maurizio R. Averna, M.D.**[8], **Stacey Gabriel, Ph.D.**[4], **Catherine Boileau, Ph.D.**[6,16], and **Sekar Kathiresan, M.D.**[3,4,5,14,26]

[1]Cardiovascular Division, Department of Medicine, Washington University School of Medicine, St. Louis MO USA [2]Division of Statistical Genomics, Washington University School of Medicine, St. Louis, MO USA [3]Center for Human Genetic Research, Massachusetts General Hospital, Boston MA, 02114, USA [4]Program in Medical and Population Genetics, Broad Institute, Cambridge MA 02142, USA [5]Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, 02114, USA [6]INSERM U698; Université Paris-Diderot, Paris, France [7]Laboratoire de Biochimie, Faculté de Pharmacie et Pôle Technologie Santé, Université Saint-Joseph, Beirut, Lebanon [8]Dipartimento Biomedico di Medicina Interna e Specialistica, Università degli Studi di Palermo, Palermo, Italy [9]Department of Experimental Vascular Medicine, Academic Medical Center, Amsterdam, 1105 AZ, The Netherlands [10]Division of Translational Medicine and Human Genetics, Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA [11]The Research Institute of the McGill University Health Centre, 687 Pine avenue West, Montréal, QC, Canada [12]Lipid Metabolism Unit, Massachusetts General Hospital, Boston, MA, 02114, USA [13]Center for Computational &

[26]To whom correspondence should be addressed: Sekar Kathiresan, M.D., Center for Human Genetic Research and Cardiovascular Research Center, Massachusetts General Hospital, 185 Cambridge Street, CPZN 5.252, Boston, MA 02114, Tel: 617-643-6120, Fax: 617-830-0690, skathiresan1@partners.org.

[25]These authors contributed equally to this work

Disclosures

The authors have no conflicts to disclose.

Author Contributions

NOS and GMP performed the primary analysis. MA, ABC, SF, MM, HT, DBL, ZA, JFH, CP, DN, PT, MR, RD, IR, JK, MWF, JG, DJR, JJPK, GKH, MRA, and CB ascertained phenotypes and collected samples. JS, DAN, JGW, SSR, and SG contributed to the study design and obtained funding. NG and DNF oversaw exome sequencing. BMN and MJD contributed to the design of the polygenic score analysis. NOS, GMP, and SK designed the overall study. NOS, GMP, and SK wrote the manuscript. All authors critically reviewed and approved the manuscript.

Integrative Biology, Massachusetts General Hospital, Boston, MA, 02114, USA [14]Department of Medicine, Harvard Medical School, Boston, MA, 02115, USA [15]Cardiovascular Research Institute, University of California, San Francisco, CA 94158 [16]AP-HP, Hôpital Ambroise Paré, Laboratoire de Biochimie et Génétique Moléculaire, Boulogne-Billancourt; Université Versailles Saint-Quentin-en-Yvelines, UFR de Médecine Paris Ile-de-France Ouest, Guyancourt, France [17]Department of Life Sciences, University of Modena and Reggio Emilia, Modena, Italy [18]Division of Cardiovascular Medicine Kanazawa University Graduate School of Medicine, Kanazawa, Japan [19]Department of Lipidology, Graduate School of Medical Science, Kanazawa University, Kanazawa, Japan [20]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA [21]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, Mississippi, USA [22]Center for Public Health Genomics, University of Virginia, Charlottesville, VA [23]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, 02114, USA [24]Department of Vascular Medicine, Academic Medical Center, Amsterdam, The Netherlands

Abstract

Background—Exome sequencing is a promising tool for gene mapping in Mendelian disorders. We utilized this technique in an attempt to identify novel genes underlying monogenic dyslipidemias.

Methods and Results—We performed exome sequencing on 213 selected family members from 41 kindreds with suspected Mendelian inheritance of extreme levels of low-density lipoprotein (LDL) cholesterol (after candidate gene sequencing excluded known genetic causes for high LDL cholesterol families) or high-density lipoprotein (HDL) cholesterol. We used standard analytic approaches to identify candidate variants and also assigned a polygenic score to each individual in order to account for their burden of common genetic variants known to influence lipid levels. In nine families, we identified likely pathogenic variants in known lipid genes (*ABCA1*, *APOB*, *APOE*, *LDLR*, *LIPA*, and *PCSK9)*; however, we were unable to identify obvious genetic etiologies in the remaining 32 families despite follow-up analyses. We identified three factors that limited novel gene discovery: (1) imperfect sequencing coverage across the exome hid potentially causal variants; (2) large numbers of shared rare alleles within families obfuscated causal variant identification; and (3) individuals from 15% of families carried a significant burden of common lipid-related alleles, suggesting complex inheritance can masquerade as monogenic disease.

Conclusions—We identified the genetic basis of disease in nine of 41 families; however, none of these represented novel gene discoveries. Our results highlight the promise and limitations of exome sequencing as a discovery technique in suspected monogenic dyslipidemias. Considering the confounders identified may inform the design of future exome sequencing studies.

Keywords

genetics; human; DNA sequencing; Exome sequencing; lipids; Mendelian Genetics

Introduction

"Exome" sequencing refers to the use of next generation sequencing (NGS) technology[1] to sequence all protein-coding regions of the genome. This approach has emerged as a promising tool for gene discovery in families with suspected monogenic disorders[2] with some reports suggesting a success rate in excess of 50%[3]. Identifying the genetic basis underlying monogenic forms of dyslipidemia has revealed insights into human biology[4] and spurred the development of novel therapeutics[5]. In an attempt to map novel dyslipidemia genes, we performed exome sequencing on 213 selected family members from 41 kindreds with suspected Mendelian inheritance of extreme levels of low-density lipoprotein cholesterol (LDL-C) or high-density lipoprotein cholesterol (HDL-C). To enrich for novel gene discoveries, we excluded probands from high LDL-C families that had mutations in genes known to cause monogenic hypercholesterolemia.

Methods

Subject Recruitment

Forty-one families of European ancestry with suspected Mendelian inheritance of extreme LDL-C or HDL-C levels were recruited from eight different centers across North America and Europe. The pedigrees of these 41 families are shown in Supplementary Figure 1. Families A1–A9 were recruited as part of the French National Research Network on Hypercholesterolemia that includes clinicians from 11 different cities in France. Probands were selected if they met the following criteria: total and LDL-C levels above the 95th percentile when compared with a sex- and age-matched French population[6], triglyceride level below 1.5 mmol/L, and presumed autosomal dominant transmission of hypercholesterolemia in the family. Family A10 was recruited from the Preventive Cardiology/Lipid Clinic of the McGill University Health Centre. Affected individuals had LDL-C concentration exceeding the 95th percentile for age- and gender-matched subjects, a plasma triglyceride concentration less than 1.0 mmol/L, and no known secondary causes of hypercholesterolemia. Families A11–A14 were recruited from the Lipid Clinic at the Academic Medical Center, University of Amsterdam, the Netherlands based on a clinical diagnosis of familial hypercholesterolemia in the proband. LDL-C levels exceeding the 95th percentile when adjusted for age and gender defined affected family members. Families A15–A20 were recruited from the Lipid Clinic of the University Hospital of Palermo. The Simon-Broome Register criteria were used to clinically diagnose heterozygous autosomal dominant hypercholesterolemia after excluding secondary hypercholesterolemia. In family A20, a pathogenic mutation in *LDLR* (c.2390-1G/A) was discovered previously but displayed incomplete penetrance (Supplementary Figure 1; A20, individuals shaded in black) and was not present in the individual with the highest level of LDL-C (the proband III:1 did not carry the LDLR mutation and had LDL-C = 455 mg/dL in addition to a history of myocardial infarction at the age of 35). Two other subjects (Supplementary Figure 1; A20, individuals shaded in blue) also showed high LDL-C and did not have the *LDLR* mutation. Based on a review of the pedigrees (Supplementary Figure 1) an autosomal dominant mode of inheritance was presumed for Families A1–A20 with the exception of A13 in which an autosomal recessive mode of inheritance was presumed.

Families B1 and B2 were recruited from the University Hospital of Palermo and Modena-Reggio Emilia. Families B3–B12 were recruited from the Washington University Lipid Research Clinic. Affected individuals in these families were identified due to an LDL-C level corresponding to the bottom 5th percentile when adjusted for age, ethnicity, and gender. The proband (subject III:A) in family B13 was referred to the MGH Lipid Metabolism Unit due to a LDL-C value of 25 mg/dL. She was noted to have 4 family members with LDL-C values less than 47 mg/dL. An autosomal recessive mode of inheritance was presumed for family B1 while an autosomal dominant mode of inheritance was presumed for B2–B13 based on the pedigrees (Supplementary Figure 1).

Family C1 was recruited as part of the Genomic Resource in Arteriosclerosis and Metabolic Disease at the Cardiovascular Research Institute of the University of California, San Francisco. The clinical diagnosis of familial hypoalphalipoproteinemia was based on levels of HDL-C below the 5th percentile for five individuals, and below the 10th percentile for one individual, when adjusted for age, sex, and the known inverse relationship between TG and HDL-C. Family C2 was recruited from the Preventive Cardiology/Lipid Clinic of the McGill University Health Centre. Families C3 and C4 were recruited from the outpatient clinic for Vascular Medicine at the Academic Medical Center, University of Amsterdam, the Netherlands. Affected individuals from families C2–C4 had HDL-C concentration below the 5th percentile for age- and gender-matched subjects, a plasma triglyceride concentration less than 1 mmol/L, and no known secondary causes of hypoalphalipoproteinemia. An autosomal dominant mode of inheritance was presumed for families C1–C4 based on the pedigrees (Supplementary Figure 1).

The probands in families D1 and D2 were ascertained from a general patient population in the center of The Netherlands and were selected based on having HDL-C levels above the 99th percentile after adjusting for age and gender[7]. Family members with HDL-C levels above the 95th percentile for age- and gender-matched subjects were considered affected. Families D3 and D4 were recruited at the Perelman School of Medicine at the University of Pennsylvania as part of a study enrolling individuals with HDL-C levels above the 75th percentile for age-, race-, and gender-matched subjects. Spouses and blood relatives of affected individuals were also recruited. An autosomal dominant mode of inheritance was presumed for families D1–D4 based on the pedigrees (Supplementary Figure 1).

Causal mutations in *LDLR*, *APOB*, and *PCSK9* were excluded in the probands of families A1–A20 as previously described[8, 9]. In addition, causal mutations in *LDLRAP1* were excluded in the proband from family A13 in which an autosomal recessive mode of inheritance was presumed. Candidate gene sequencing was not performed in the other families.

Replication in Japanese Families

The families shown in Supplementary Figure 6 (Families A-D) were recruited from Kanazawa University Hospital in Kanazawa, Japan. The probands in Families A, C, and D were identified due to high LDL-C values and tendinous xanthomas. An off-treatment LDL-C value was not available for the proband in Family B; her LDL-C value was normal, however she was on intensive lipid-lowering therapy and was noted to have tendinous

xanthomas. Affected relatives in Families A-D had LDL-C values exceeding 200mg/dL. An autosomal dominant mode of inheritance was presumed for all four families based on the pedigrees (Supplementary Figure 6). All individuals in Families A-D were of self-described Japanese ancestry.

Exome Sequencing

Subsets of samples from each family were selected for exome sequencing based on DNA availability, presence of informed consent allowing for genetic studies, and prioritization of phenotypic extremes. These selected samples underwent exome sequencing at the Broad Institute. The IRB at the Broad Institute and all participating sites approved the study protocols and all individuals who were selected for sequencing provided informed consent. Randomly sheared genomic DNA was used as input for library construction and in-solution hybrid selection to enrich for exomic DNA as previously described[10]. In all samples except seven, 33Mb of genomic sequence was defined as the "exome" and targeted using the Whole Exome Agilent 1.1 plus boosters preparation kit (Agilent Technologies, Santa Clara, CA, USA). The remaining seven samples underwent hybrid selection using a prior version of the Agilent whole exome preparation kit that targeted 28.6 Mb of genomic sequence. Exome-enriched DNA for each sample was then sequenced on an Illumina GA-II sequencer using 75-base pair paired-end reads. Samples were sequenced with a goal of achieving at least 20-fold coverage in at least 80% of targeted bases. Samples with less than 80% of targeted bases covered by ≥ 20 sequencing reads were not used for the primary analysis.

The Burroughs-Wheeler Alignment algorithm[11] was used to map raw sequence reads to the human reference sequence (UCSC build HG19). The Genome Analysis Toolkit (GATK version 2)[12] and SAMtools[13] were used to locally realign reads, recalibrate individual base qualities, and flag duplicate sequencing reads for removal. The GATK UnifiedGenotyper (UG) was then used to identify single nucleotide variants and small insertions and deletions in the exome target definition specific for each sample along with up to 50 flanking intronic bases. The UG was used in multisample mode and samples were grouped into three batches keeping related samples together when possible. The GATK Variant Score Recalibration tool was used to update the quality score of the identified variants and SnpEff[14] was used to predict the functional consequences of each variant. The population allele frequency of each variant was estimated using the National Heart Lung and Blood Institute's Exome Sequencing Project (ESP) Exome Variant Server (http://evs.gs.washington.edu/EVS).

Analysis of Exome Sequencing Data

To exclude genetic variation unlikely to be causal for the extreme phenotype in the affected individuals from the families, we employed a heuristic analytical process commonly used in the analysis of exome sequencing studies[15]. Starting with the total number of variants shared by individuals from the family, we excluded variation that did not fit the expected pattern of inheritance based on examining the pedigree. Next, we excluded common genetic variation, defined as ≥ 1% frequency in the population (given the extreme excess of very rare alleles in the human population[16], the exact choice of this threshold – i.e. 1% or 0.01% – has little practical impact since most rare alleles within families are well below this threshold). On the assumption that a specific causal variant could not be responsible for both high and low lipid

levels, we excluded variation present in the affected individuals of the opposite extreme. Finally, we excluded silent and non-genic variation as most Mendelian syndromes are caused by coding or splice site mutations that alter the protein sequence[17]. The remaining single nucleotide variants and short insertions or deletions were considered candidates. When possible, from this list of candidates we attempted to identify variants demonstrating co-segregation with the phenotype in the extended kindred. We considered candidate mutations as causal if (1) the mutation was identified in prior publications as causal for the same phenotype; (2) if the mutation was novel but in a gene known to cause the phenotype and functionally similar to causal mutations in that gene (i.e. a novel nonsense mutation occurring in a gene in which other nonsense mutations have been shown to be causal), or (3) if the mutation was novel and occurred in a novel gene but demonstrated co-segregation with the phenotype in the extended kindred. The 95% confidence intervals (CI) surrounding the success estimates were estimated from the binomial distribution.

Polygenic score analysis

To determine the likelihood that polygenic inheritance could explain the extreme lipid phenotype in some families, individuals with sufficient DNA (n=130) were genotyped on the Illumina HumanExome Beadchip v1.0 according to the manufacturer's recommended protocol. This genotyping array includes the SNPs reported in the Global Lipids Genetic Consortium (GLGC) meta-analysis of genome-wide association studies of plasma lipid levels[18]. Of the 102 SNPs reported in GLGC Table 1, we successfully genotyped 87 SNPs plus 4 proxies ($r^2 > 0.9$ with the GLGC SNP). Using all 91 SNPs (all SNPs were used for each lipid trait since some of the SNPs are associated with more than one lipid fraction), we built baseline polygenic models for the LDL-C and HDL-C phenotypes in 9,134 subjects not taking lipid-lowering medications from the Ottawa Heart Study[19], PROCARDIS[20] and the Malmo Diet and Cancer Study[21] to obtain estimated regression coefficients. Next, we used the estimated coefficients to obtain a predicted lipid level for each individual in our study based on these 91 SNPs. This predicted lipid level was the population mean plus the sum of the individual's observed genotypes weighted by the estimated coefficients. We calculated a residualized phenotype for each individual by subtracting the observed lipid level from the predicted lipid level based on the common SNPs. The observed lipid levels were either obtained off treatment or adjusted for lipid-lowering treatment by dividing the observed value by 0.7. Externally standardized residuals were created to assess the statistical significance of each individual's residualized phenotype. We used a threshold of ≤ 0.01 (a Bonferroni correction for the average number of individuals sequenced in each family) to define a significant residualized score.

Results

We performed exome sequencing on 213 selected individuals from the 41 families with suspected monogenic inheritance of extreme lipid levels, with a median of 4 individuals selected per kindred. On average, the mean coverage of targeted bases for each individual was 103. We identified an average of 12,544 nonsynonymous single nucleotide variants and 802 insertion/deletions per individual. Within each kindred, we used standard approaches to identify candidate variants[15] (Supplementary Figure 2). In five families (12%), we identified

likely pathogenic variants (Table 1) in genes previously proven to cause monogenic dyslipidemias ("known lipid genes", Supplementary Table 1). We also identified the genetic etiology in three families after follow-up analysis of their candidate variants (Supplementary Figure 1, Families A1[26], A10[30], and A13[27]) and one after considering the effect of common genetic variants (described below), bringing the total to nine (22%; 95% CI [9.3%,34.7%]) (see Table 1 and Supplementary Table 2 for details).

In the remaining 32 families however, the number of candidate variants ranged from 0–287, without obvious genetic etiologies despite follow-up analyses. We sought to understand potential reasons for the lack of novel gene discovery and identified three main confounders: 1) an inability to identify potentially causal variants due to imperfect sequencing coverage; 2) an inability to identify the causal variant among hundreds of shared variants within families; and 3) an inability to identify the effect of complex genetics using exome sequencing.

To successfully discover a causal variant, the variant must first be identified. We find that despite high average coverage across the exome (on average 89% of targeted bases are covered with ≥ 20 reads; see Figure 1A), a small but substantial portion of the exome is poorly covered across all affected individuals. Across the known lipid genes we find affected individuals have, on average, 3.7% of targeted bases covered with ≤ 10 sequencing reads (Figure 1a), a sequencing depth that provides 99% confidence of observing a rare allele at least twice. At these positions there is a chance we would fail to identify a variant in the affected individual with shallow sequencing coverage and these positions would then be removed from consideration under the assumption of complete penetrance without phenocopies.

Family A1 (Supplementary Figure 1) illustrates this problem. In this family, affected individuals were identified to harbor a pathogenic *APOE* deletion[26]. Initially, the pathogenic deletion (p.Leu167del) was only identified in two of three affected individuals using exome sequencing and was thus removed from further consideration. When orthogonal methods (linkage analysis and sequencing under linked peaks) identified p.Leu167del as a candidate, we performed Sanger sequencing to confirm the presence of the deletion in all affected individuals. This mutation occurs in the last exon of *APOE* which is difficult to capture and sequence with NGS[31] and has the lowest coverage and highest GC content of the known lipid genes (Supplementary Tables 3 and 4). The individual in Family A1 initially misclassified by NGS only had one sequencing read at that position of the genome. Across the remainder of the exome, we find affected individuals have, on average, 6.4% of targeted bases covered with ≤10 sequencing reads (Figure 1a). This effect appears to be independent of overall sequencing depth (Supplemental Figure 3) suggesting it cannot be solved simply by sequencing to deeper overall coverage. It has been previously suggested that exome sequencing fails to identify the genetic basis of some strongly inherited conditions due to causal non-coding mutations[32]; another explanation could be that the causal variant is present in the coding region but hitherto unidentified.

Second, we find a confounding effect from the many rare alleles within families that also segregate with the phenotype by chance. This is highlighted by examining the total number

of candidate variants even in families harboring pathogenic variants in a known lipid gene. In these families, between 2–346 additional variants remain candidates at the end of the analysis and would be considered potentially causal if a pathogenic variant had not been identified. This is similar to the number of variants remaining in families without known genetic causes (range 0–287; see Figure 1b), highlighting the vast amount of very rare variation "private" to families that segregates with the disease phenotype merely by chance.

Third, we also find that the effect of complex genetics in families with suspected monogenic dyslipidemias can be substantial. Both LDL-C and HDL-C levels are influenced by multiple common genetic loci[18]; we[18] and others[33] have previously demonstrated that polygenic inheritance may be sufficient to explain extreme lipid phenotypes. To address this possibility in the families sequenced in the present study, we genotyped a set of common genetic variants robustly associated with lipid traits in genome-wide association studies[18]. Using these common variants, we created a polygenic score and calculated a residualized phenotypic z-score, effectively assigning a level of statistical significance to each individual's lipid level after correcting for that individual's burden of common lipid-related alleles (Supplementary Figure 4).

As a proof-of-principle, we find highly significant scores for individuals from families A9 and B2, in which pathogenic *PCSK9* and *APOB* mutations, respectively, perfectly segregate with disease status, indicating that common genetic factors are not sufficient to explain the phenotype in these families (Supplementary Table 5). In contrast, of the families without a readily apparent genetic answer, we find six (15% of all families) where either all affected individuals have non-significant scores or only one affected individual retains a significant score, suggesting that the burden of common alleles is sufficient to largely explain the extreme phenotype in these families (Supplementary Table 6 and Supplementary Figure 5).

We also find this approach can help refine phenotypic definitions within families. In Family A7, an initial analysis using clinically-defined affection status (Figure 2a) yielded 17 candidate variants; additional analyses were unable to identify a causal variant from this list. Using the polygenic score, we found individuals III-1 and III-2 had LDL-C levels that were largely explained by a burden of common variants whereas individuals II-3 and III-5 had highly significant residualized scores (Supplementary Table 7). An analysis using these updated phenotypes identified a candidate variant in *LDLR* (p.E228K, also known as FH Modena, previously shown to be pathogenic[28]) that was subsequently confirmed to perfectly segregate with extreme LDL-C levels in the extended kindred (Figure 2b).

We attempted to extend these findings to a non-European population and found similar results in families of Japanese descent with extreme LDL-C levels (Supplementary Figure 6). In this analysis a pathogenic variant in *LDLR* was found in one family (Supplementary Table 8) while there was no clear molecular etiology in the remaining three, resulting in a 25% success rate. We found similar levels of imperfect sequence coverage and numbers of rare variants segregating within the family (Supplementary Table 9) compared to the families of European descent.

Discussion

The present study summarizes our experience using exome sequencing to map novel genes in families with a suspected Mendelian dyslipdemia. After sequencing the exome in 213 individuals across 41 kindreds, we find this technique identifies a likely causal variant in 22% of cases. Three of these families harbor causal mutations in known lipid genes that were excluded by candidate sequencing prior to entry into this study, reconfirming previous results that candidate gene sequencing can fail to identify causal mutations in candidate genes that are subsequently identified via NGS[34]. Notably, we did not identify any novel monogenic dyslipidemia genes. From the remainder of the families in our study, we identify evidence of polygenic inheritance in 15%. We are, however, currently unable to define the genetic basis in the remaining 63% of families (Figure 3).

Several conclusions emerge from these results. First, from this empiric evaluation, we find the yield of exome sequencing as a tool for novel gene mapping to be modest. Multiple reports detailing the success of this technique have been published since 2009[2]; however, these reports may be susceptible to the well-known bias to publish positive results. Overall, we are unaware of reports detailing the overall success rate of exome sequencing. Our study highlights the “real-world” challenges in using this technique for mapping novel genes in Mendelian disorders and appears to reflect the collective experience in monogenic dyslipidemias as evidenced by the current literature. We are unaware of a Mendelian dyslipidemia gene other than *ANGPTL3*[35] that has been identified using exome sequencing.

Second, we identified several technical issues that have not been previously highlighted that adversely impacted our ability to discover novel genes. An underlying assumption when mapping genes with NGS is that all potentially causal variants will be identified. Our study reveals some of the limitations of exome sequencing which may be useful to address in future design of research or clinical sequencing studies.

Third, we find a substantial concerted effect of common lipid-related alleles that appears to result in extreme phenotypes in some individuals and families. A similar effect has been shown previously for samples drawn from the extremes of the population distribution of plasma lipids[18] and in a significant proportion of mutation-negative probands who underwent clinical genetic testing for familial hypercholesterolemia[33]. While only approximately 10% of the total phenotypic variance in lipid traits is explained by the common variants genotyped in our study[18], by removing this variance we were able to create a more refined phenotype and enrich for genetic factors not present in the set of common variants. We show that incorporating this information may inform genetic mapping as we have demonstrated above in Family A7.

Our study has several limitations. Exome sequencing has limited reliability to identify large structural variants and while we did consider small insertions and deletions, this technique is less reliable in identifying these compared with single nucleotide substitutions. Reliable incorporation of these forms of genetic variation might identify additional candidates[36]. We also did not consider genome-wide linkage data due to the small size of the pedigrees and incorporating such information has the potential to reduce the number of candidates[37]. From

a technical perspective, it is possible that other exome capture reagents or sequencing platforms (including longer reads) may result in more complete coverage. Finally, it is important to note this experience may not necessarily generalize to other phenotypes. For example, one might expect the yield to be higher for studying extreme syndromic phenotypes less susceptible to the influences of polygenic inheritance and environmental factors.

Addressing the three problematic areas outlined above has the potential to improve the success of gene mapping. Whole genome sequencing (WGS) could be used to remove the bias of target definition and hybrid selection inherent in exome sequencing, although we recognize certain portions of the genome will remain recalcitrant to NGS technology[38]. The process for identifying causal rare alleles within families can be improved as larger population-based sequencing studies are performed and as better high-throughput functional assays are developed. Sequencing more distantly related relatives can decrease the total number of shared alleles; however, investigators typically resort to exome sequencing in small pedigrees for which linkage analysis has been intractable. Finally, as we identify additional alleles contributing to complex phenotypic traits, we can use these findings to inform family-based genetic studies by both selecting families without significant polygenic inheritance and refining phenotypic definitions within families.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank all probands and family members who consented for participation in this study. We thank Gustav Schonfeld, M.D., for his work in recruiting families B3–B12 from the Washington University Lipid Research Clinic. We would like to thank The French Research Network on ADH for participating in family recruitment, especially: Dr. Michel Farnier (Dijon), Dr. Gérald Luc (Lille), Pr. Philippe Moulin, Dr. Laurence Perrot (Lyon), Pr. Michel Krempf, Dr. Yassine Zaïr (Nantes), Pr. Eric Bruckert, Dr. Valérie Carreau (Paris), Dr. Laurence Collet (Saint- Etienne).

Funding sources

NOS is supported, in part, by a career development award (K08HL114642) from the National Heart Lung and Blood Institute (NHLBI) and also by The Foundation for Barnes-Jewish Hospital. GMP is supported by award number T32HL007208 from the NHLBI. SK is supported by a Research Scholar award from the Massachusetts General Hospital (MGH), the Howard Goodman Fellowship from MGH, the Donovan Family Foundation, R01 HL107816, and a grant from Fondation Leducq. We thank the National Heart, Lung, and Blood Institute GO Exome Sequencing Project (ESP) Family Study Project Team for supporting the exome sequencing. We also thank the ESP component studies including the Lung Cohorts Sequencing Project (HL-102923), the WHI Sequencing Project (HL-102924), the Heart Cohorts Sequencing Project (HL-103010), the Broad Institute Sequencing Project (HL-102925), the Northwest Genomics Center Sequencing Project (HL-102926), and the Family Studies Project Team. This work was also supported by grants from PHRC (AOM06024) and ANR (ANR-05-PCOD-017, ANR-06-MRAR-038, ANR-08-GENO-002-01). M.A is supported by grants from Conseil de la Recherche de l'Université Saint-Joseph (Beirut, Lebanon). JJPK is a recipient of the Lifetime Achievement Award of the Dutch Heart Foundation (2010T082). GKH is a recipient of a NWO Veni grant (project number 91612122), a grant from the Netherlands CardioVascular Research Initiative (CVON2011-19; Genius) and the European Union (Resolve: FP7-305707; TransCard: FP7-603091-2). MM is supported by a grant from Fondation LeDucq (2009–2014).

References

1. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008; 26:1135–1145. [PubMed: 18846087]
2. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome sequencing as a tool for mendelian disease gene discovery. Nat Rev Genet. 2011; 12:745–755. [PubMed: 21946919]
3. Gilissen C, Hoischen A, Brunner HG, Veltman JA. Unlocking mendelian disease using exome sequencing. Genome Biol. 2011; 12:228. [PubMed: 21920049]
4. Brown MS, Goldstein JL. A receptor-mediated pathway for cholesterol homeostasis. Science. 1986; 232:34–47. [PubMed: 3513311]
5. Abifadel M, Varret M, Rabes JP, Allard D, Ouguerram K, Devillers M, et al. Mutations in *PCSK9* cause autosomal dominant hypercholesterolemia. Nat Genet. 2003; 34:154–156. [PubMed: 12730697]
6. Siest G, Visvikis S, Herbeth B, Gueguen R, Vincent-Viry M, Sass C, et al. Objectives, design and recruitment of a familial and longitudinal cohort for studying gene-environment interactions in the field of cardiovascular risk: The stanislas cohort. Clinical chemistry and laboratory medicine : CCLM / FESCC. 1998; 36:35–42. [PubMed: 9594084]
7. Motazacker MM, Peter J, Treskes M, Shoulders CC, Kuivenhoven JA, Hovingh GK. Evidence of a polygenic origin of extreme high-density lipoprotein cholesterol levels. Arterioscler Thromb Vasc Biol. 2013; 33:1521–1528. [PubMed: 23685560]
8. Fouchier SW, Kastelein JJ, Defesche JC. Update of the molecular basis of familial hypercholesterolemia in the netherlands. Human mutation. 2005; 26:550–556. [PubMed: 16250003]
9. Marduel M, Carrie A, Sassolas A, Devillers M, Carreau V, Di Filippo M, et al. Molecular spectrum of autosomal dominant hypercholesterolemia in France. Human mutation. 2010; 31:E1811–1824. [PubMed: 20809525]
10. Fisher S, Barry A, Abreu J, Minie B, Nolan J, Delorey TM, et al. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. Genome Biol. 2011; 12:R1. [PubMed: 21205303]
11. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]
12. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011; 43:491–498. [PubMed: 21478889]
13. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and samtools. Bioinformatics. 2009; 25:2078–2079. [PubMed: 19505943]
14. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012; 6:80–92. [PubMed: 22728672]
15. Li MX, Gui HS, Kwan JS, Bao SY, Sham PC. A comprehensive framework for prioritizing variants in exome sequencing studies of mendelian diseases. Nucleic Acids Res. 2012; 40:e53. [PubMed: 22241780]
16. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012; 337:64–69. [PubMed: 22604720]
17. Online Mendelian Inheritance in Man OMIM®. Mckusick-Nathans Institute of Genetic Medicine, Johns Hopkins University;
18. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010; 466:707–713. [PubMed: 20686565]
19. Davies RW, Wells GA, Stewart AF, Erdmann J, Shah SH, Ferguson JF, et al. A genome-wide association study for coronary artery disease identifies a novel susceptibility locus in the major histocompatibility complex. Circ Cardiovasc Genet. 2012; 5:217–225. [PubMed: 22319020]

20. Barlera S, Chiodini BD, Franzosi MG, Tognoni G. PROCARDIS: A current approach to the study of the genetics of myocardial infarct. Italian heart journal Supplement : official journal of the Italian Federation of Cardiology. 2001; 2:997–1004. [PubMed: 11675837]
21. Berglund G, Elmstahl S, Janzon L, Larsson SA. The Malmo Diet and Cancer study. Design and feasibility. Journal of internal medicine. 1993; 233:45–51. [PubMed: 8429286]
22. den Dunnen JT, Antonarakis SE. Mutation nomenclature extensions and suggestions to describe complex mutations: A discussion. Human mutation. 2000; 15:7–12. [PubMed: 10612815]
23. Cefalu AB, Pirruccello JP, Noto D, Gabriel S, Valenti V, Gupta N, et al. A novel *APOB* mutation identified by exome sequencing cosegregates with steatosis, liver cancer, and hypocholesterolemia. Arterioscler Thromb Vasc Biol. 2013; 33:2021–2025. [PubMed: 23723369]
24. Bodzioch M, Orso E, Klucken J, Langmann T, Bottcher A, Diederich W, et al. The gene encoding ATP-binding cassette transporter 1 is mutated in Tangier disease. Nat Genet. 1999; 22:347–351. [PubMed: 10431237]
25. Guo Z, Inazu A, Yu W, Suzumura T, Okamoto M, Nohara A, et al. Double deletions and missense mutations in the first nucleotide-binding fold of the ATP-binding cassette transporter a1 (*ABCA1*) gene in Japanese patients with Tangier disease. Journal of human genetics. 2002; 47:325–329. [PubMed: 12111381]
26. Marduel M, Ouguerram K, Serre V, Bonnefont-Rousselot D, Marques-Pinheiro A, Erik Berge K, et al. Description of a large family with autosomal dominant hypercholesterolemia associated with the *APOE* p.Leu167del mutation. Human mutation. 2013; 34:83–87. [PubMed: 22949395]
27. Stitziel NO, Fouchier SW, Sjouke B, Peloso GM, Moscoso AM, Auer PL, et al. Exome sequencing and directed clinical phenotyping diagnose cholesterol ester storage disease presenting as autosomal recessive hypercholesterolemia. Arterioscler Thromb Vasc Biol. 2013; 33:2909–2914. [PubMed: 24072694]
28. Hobbs HH, Brown MS, Goldstein JL. Molecular genetics of the LDL receptor gene in familial hypercholesterolemia. Human mutation. 1992; 1:445–466. [PubMed: 1301956]
29. Leigh SE, Foster AH, Whittall RA, Hubbart CS, Humphries SE. Update and analysis of the university college london low density lipoprotein receptor familial hypercholesterolemia database. Ann Hum Genet. 2008; 72:485–498. [PubMed: 18325082]
30. Awan Z, Choi HY, Stitziel N, Ruel I, Bamimore MA, Husa R, et al. *APOE* p.Leu167del mutation in Familial Hypercholesterolemia. Atherosclerosis. 2013; 231:218–222. [PubMed: 24267230]
31. Parla JS, Iossifov I, Grabill I, Spector MS, Kramer M, McCombie WR. A comparative analysis of exome capture. Genome Biol. 2011; 12:R97. [PubMed: 21958622]
32. Fairfield H, Gilbert GJ, Barter M, Corrigan RR, Curtain M, Ding Y, et al. Mutation discovery in mice by whole exome sequencing. Genome Biol. 2011; 12:R86. [PubMed: 21917142]
33. Talmud PJ, Shah S, Whittall R, Futema M, Howard P, Cooper JA, et al. Use of low-density lipoprotein cholesterol gene score to distinguish patients with polygenic and monogenic familial hypercholesterolaemia: A case-control study. Lancet. 2013; 381:1293–1301. [PubMed: 23433573]
34. Futema M, Plagnol V, Whittall RA, Neil HA, Humphries SE, et al. Simon Broome Register G. Use of targeted exome sequencing as a diagnostic tool for familial hypercholesterolaemia. J Med Genet. 2012; 49:644–649. [PubMed: 23054246]
35. Musunuru K, Pirruccello JP, Do R, Peloso GM, Guiducci C, Sougnez C, et al. Exome sequencing, *ANGPTL3* mutations, and familial combined hypolipidemia. N Engl J Med. 2010; 363:2220–2227. [PubMed: 20942659]
36. Norton N, Li D, Rieder MJ, Siegfried JD, Rampersaud E, Zuchner S, et al. Genome-wide studies of copy number variation and exome sequencing identify rare variants in *BAG3* as a cause of dilated cardiomyopathy. Am J Hum Genet. 2011; 88:273–282. [PubMed: 21353195]
37. Yamaguchi T, Hosomichi K, Narita A, Shirota T, Tomoyasu Y, Maki K, et al. Exome resequencing combined with linkage analysis identifies novel *PTH1R* variants in primary failure of tooth eruption in Japanese. Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research. 2011; 26:1655–1661.
38. Kirby A, Gnirke A, Jaffe DB, Baresova V, Pochet N, Blumenstiel B, et al. Mutations causing Medullary Cystic Kidney Disease Type 1 lie in a large VNTR in *MUC1* missed by massively parallel sequencing. Nat Genet. 2013; 45:299–303. [PubMed: 23396133]

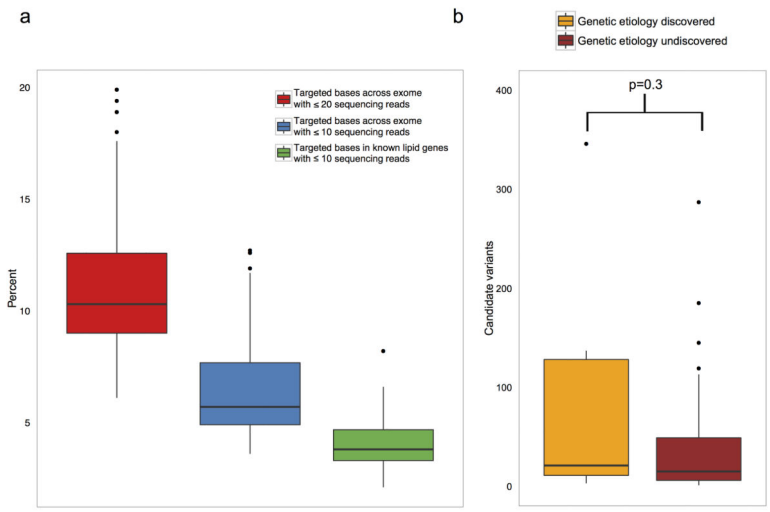

Figure 1. Selected metrics from exome sequencing analysis

(a) Percent targeted bases across the exome definition supported by ≤ 20 sequencing reads (Red) or ≤ 10 sequencing reads (Blue) or across genes previously identified to cause monogenic dyslipidemia supported by ≤ 10 sequencing reads (Green). (b) Number of candidate variants after analysis for families with a suspected pathogenic variant in a gene known to cause monogenic dyslipidemia (orange) compared with families without known cause (brown). P refers to the p-value from the Kolmogorov–Smirnov test in testing for differences between the distributions.

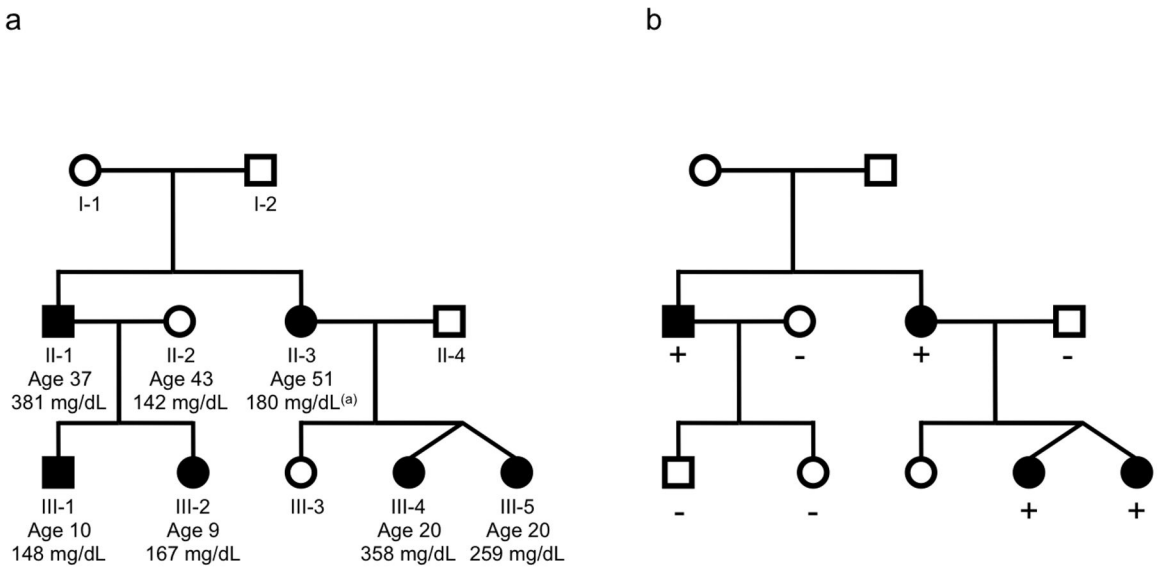

Figure 2. Pedigree of Family A7, demonstrating the utility of refining phenotypes based on burden of common alleles

(A) Initial pedigree defining affected individuals (shaded) by LDL-C level adjusted for age and gender. (B) Updated pedigree based on residualized phenotype score (see text) where individuals III-1 and III-2 are classified as unaffected. *LDLR* p.E228K carrier status is indicated with + (heterozygous) or - (wild type). The superscript (a) indicates the LDL-C level was obtained while the individual was on lipid-lowering medication therapy.

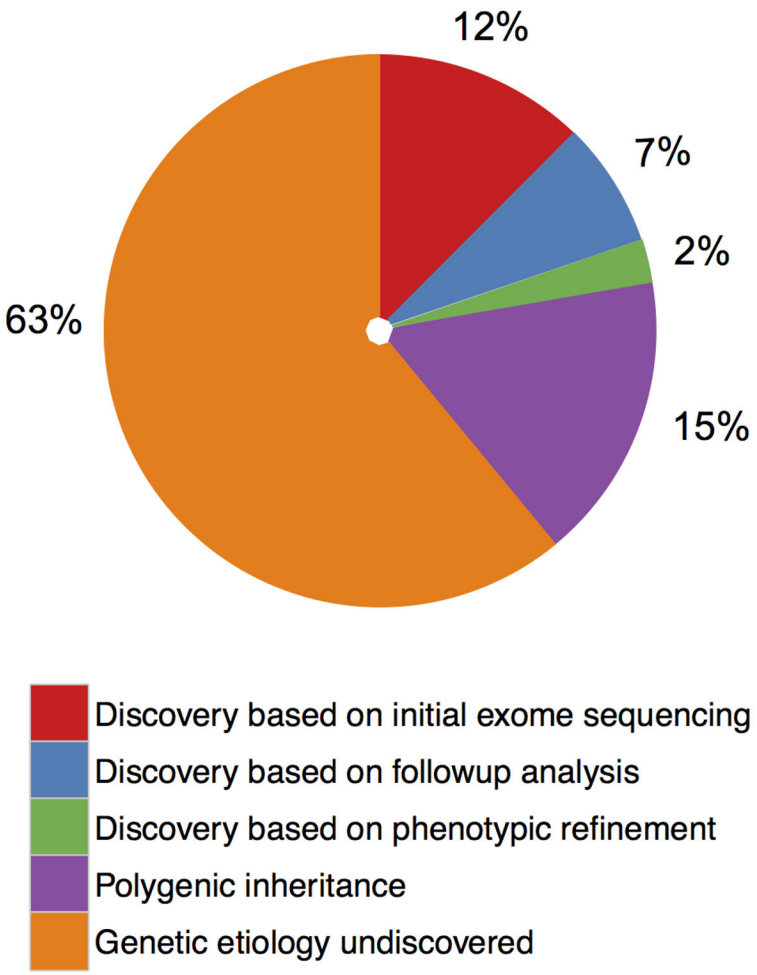

Figure 3. Discovery rates from exome sequencing

The distribution of final discovery status for the 41 families with suspected monogenic dyslipidemias that underwent exome sequencing is shown with approximate percentages.

Table 1

Genetic etiologies identified from exome sequencing.

Family (Trait)	Gene	Genomic position*	Reference allele	Alternate allele	Effect †	Notes
Genetic etiology discovered during initial exome sequencing analysis						
A4 (high LDL)	*APOB*	2:21229554	C	T	p.A3396T	‡
A9 (high LDL)	*PCSK9*	1:55509689	T	A	p.S127R	§
B2 (low LDL)	*APOB*	2:21233022	T	A	p.K2240*	||
B13 (low LDL)	*APOB*	2:21229005	-	G	p.T3579Hfs*34	#
C1 (low HDL)	*ABCA1*	9:107553287	T	C	p.N1948S	**
Genetic etiology discovered from follow-up analysis						
A1 (high LDL)	*APOE*	19:45412048	CTC	-	p.L167del	††
A10 (high LDL)	*APOE*	19:45412048	CTC	-	p.L167del	††
A13 (high LDL)	*LIPA*	10:90982268	C	T	Disruption of donor splice site	‡‡
Genetic etiology discovered based on phenotypic refinement within the family						
A7 (high LDL)	*LDLR*	19:11216264	G	A	p.E228K	§§

*Genomic position lists chromosome and position in hg19 coordinates.

†Effect refers to the predicted protein change using proposed nomenclature[22] based on the cDNA sequence with the ATG initiation codon numbered p.1. The following reference sequences were used: *ABCA1*: NM_005502.3, *APOB*: NM_000384.2, *APOE*: NM_000041.2, *LDLR*: NM_000527.4, *PCSK9*: NM_174936.3, *LIPA*: NM_000235.2.

‡To our knowledge this mutation has not been previously identified as causing autosomal dominant hypercholesterolemia (ADH). However, this mutation occurs at a highly conserved position within the highly conserved LDLR-binding domain of ApoB where other missense mutations causing ADH have been identified. A threonine for alanine substitution at this position is computationally predicted to be damaging.

§This mutation has been previously identified as causing ADH in other families[5].

||A report detailing this mutation has been previously published[23].

#To our knowledge this mutation has not been previously identified as causing familial hypobetalipoproteinemia (FHBL) however this is expected type of causal mutation for FHBL, inducing a premature truncation of ApoB.

**To our knowledge this mutation has not been previously identified as causing Tangier disease. However, this mutation is in the second conserved nucleotide-binding domain (ATP binding cassette) within the Walker A/P-loop of *ABCA1*. It affects the amino acid residue position corresponding to the equivalent Asparagine in the 1st nucleotide binding domain which is known to be causal in Tangier disease[24,25]. The sequences are GHNGAGKTTTM (domain 1) and GVNGAGKSSTF (domain 2) (the mutated residue is underlined).

††A report of this mutation causing ADH has been previously published[26].

[#] A report detailing this mutation has been previously published[27].

[§§] This mutation, also known as FH-Jerusalem, has been previously identified as causing familial hypercholesterolemia[28,29].