UNIVERSITY OF CALIFORNIA

Los Angeles

# Applications of Next-Generation DNA Sequencing
# to the Identification of Rare Variants in Congenital Disorders
# of the Intestine and Brain

A dissertation submitted in partial satisfaction of the
requirements for the degree of Doctor of Philosophy
in Human Genetics

by

Michael Yourshaw

2014

ABSTRACT OF THE DISSERTATION

# Applications of Next-Generation DNA Sequencing to the Identification of Rare Variants in Congenital Disorders of the Intestine and Brain

by

Michael Yourshaw

Doctor of Philosophy in Human Genetics

University of California, Los Angeles, 2014

Professor Stanley F. Nelson, Chair

High throughput, massively parallel DNA sequencing provides a powerful technology to study the human genome and to identify variations in DNA that cause disease. Sequencing the protein coding region of the genome ('whole-exome sequencing') is a cost effective method to search the part of the genome that is most likely to harbor disease related mutations.

We developed software methods to process sequencing data and to annotate variants with data on genes, function, conservation, expression, diseases, pathways, and protein structure.  We applied whole-exome sequencing to search for the molecular basis of disease in three projects: 1) a cohort of patients with congenital diarrheal disorders (CDDs); 2) a cohort of patients with congenital chronic intestinal pseudo-obstruction (CIPO) or the related disease, megacystis-microcolon-intestinal hypoperistalsis syndrome (MMIH); and 3) four siblings with infantile pontocerebellar hypoplasia and spinal motor neuron degeneration.

We sequenced 45 probands from diverse ethnic backgrounds who were diagnosed with a variety of CDDs of probable, but unknown genetic cause. Patients had been diagnosed with generalized malabsorptive diarrhea, selective nutrient malabsorption, secretory diarrhea, and infantile IBD. We found homozygous or compound heterozygous mutations, 25 of them novel, in genes known to be associated with CDDs in 27 cases (60%). The genes implicated were *ADAM17*, *DGAT1*, *EPCAM*, *IL10RA*, *MALT1*, *MYO5B*, *NEUROG3*, *PCSK1*, *SI*, *SKIV2L*, *SLC26A3*, and *SLC5A*.

With whole-exome sequencing in a cohort of 20 patients with congenital CIPO or MMIH, we identified a subset of 10 cases with potentially damaging de-novo dominant acting mutations at highly conserved loci in the *ACTG2* gene, encoding actin, gamma-enteric smooth muscle precursor, a protein essential to the functioning of muscle cells in the intestinal wall.

By exome sequencing, we discovered rare recessive mutations in *EXOSC3* (encoding exosome component 3) that were responsible for pontocerebellar hypoplasia and spinal motor neuron degeneration in the four probands, and identified identical and additional novel mutations in a large percentage of other children with the same disorder.

In conclusion, we demonstrated that whole-exome sequencing is an effective approach for the identification of casual mutations in that may escape detection with standard practice involving a complex diagnostic workup and targeted gene sequencing.

The dissertation of Michael Yourshaw is approved.

Rita M. Cantor

Lars Dreier

J. Aldons Lusis

Stanley F. Nelson, Committee Chair

University of California, Los Angeles

2014

*Dedicated to*

*my parents*

*Mike and Elizabeth*

*my children*

*Ivan, Erik, Alexis, Amanda, and Christopher*

*and my grandchildren*

*Sarah, Thunder, Steel, Heaven, Jewel, Love, and Mercy*

## TABLE OF CONTENTS

## LIST OF FIGURES

Chapter 2

Chapter4

Chapter 5

Chapter 7

## LIST OF TABLES

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

# ACKNOWLEDGMENTS

Chapter 2 is a manuscript, "Rich Annotation of DNA Sequencing Variants by Leveraging the Ensembl Variant Effect Predictor with Plugins", which has been submitted to Briefings in Bioinformatics and is currently under review. It is presented with the permission of the co-authors, S. Paige Taylor, Aliz R. Rao, **Martín G. Martín**, and **Stanley F. Nelson**.

Chapter 3 is an in preparation first author manuscript, "Whole-Exome Sequencing for the Identification of Casual Mutations in Congenital Diarrheal Disorders". I acknowledge my co-authors, Robert Venick, Pam Brown, Rune Rose Tronstad, David Brumbaugh, Vered Pinsk, Robert Bandsma, Jeff Lewis, John Kerne, Anja Koren Jeverica, Anupama Chawla, Jerry Levine, **Stanley F. Nelson**, and **Martín G. Martín**.

Chapter 4 is a published first author paper presented with permission from the Journal of Pediatric Gastroenterology and Nutrition and the co-authors. Yourshaw M, Solorzano-Vargas RS, Pickett LA, Lindberg I, Wang J, Cortina G, Pawlikowska-Haddal A, Baron H, Venick RS, **Nelson SF**, **Martin MG**. Exome Sequencing Finds a Novel PCSK1 Mutation in a Child With Generalized Malabsorptive Diarrhea and Diabetes Insipidus. Journal of pediatric gastroenterology and nutrition. 2013;57(6):759-67. Epub 2013/11/28. doi: 10.1097/MPG.0b013e3182a8ae6c. PubMed PMID: 24280991.

Chapter 5 is a published paper presented with permission from PLOS ONE and the co-authors. Pickett LA, Yourshaw M, Albornoz V, Chen Z, Solorzano-Vargas RS, **Nelson SF**, **Martin MG**, **Lindberg I**. Functional consequences of a novel variant of PCSK1. PLoS One. 2013;8(1):e55065.

doi: 10.1371/journal.pone.0055065. PubMed PMID: 23383060; PubMed Central PMCID: PMC3557230.

Chapter 6 is an in preparation co-first author manuscript, "Mutations in *ACTG2* are associated with sporadic congenital chronic intestinal pseudo-obstruction". I acknowledge my co-authors, Antia Sicolo (who contributed equally to the work), Robert Venick, Laurie Reyen, Doug Farmer, Jorge Vargas, **Stanley F. Nelson**, and **Martín G. Martín**.

Chapter 7 is a co-first author published paper presented with permission from Nature Genetics and the co-authors. Wan J, Yourshaw M, Mamsa H, Rudnik-Schoneborn S, Menezes MP, Hong JE, Leong DW, Senderek J, Salman MS, Chitayat D, Seeman P, von Moers A, Graul-Neumann L, Kornberg AJ, Castro-Gago M, Sobrido MJ, Sanefuji M, Shieh PB, Salamon N, Kim RC, Vinters HV, Chen Z, Zerres K, Ryan MM, **Nelson SF**, **Jen JC**. Mutations in the RNA exosome component gene EXOSC3 cause pontocerebellar hypoplasia and spinal motor neuron degeneration. Nat Genet. 2012;44(6):704-8. Epub 2012/05/01. doi: 10.1038/ng.2254. PubMed PMID: 22544365; PubMed Central PMCID: PMC3366034.

PIs and project directors are in bold.

I thank my committee members Rita M. Cantor, Lars Drier, Jake Lusis, and Stanley F. Nelson for their guidance during this work, and my collaborators Martín Martín, Joanna Jen, and Berit Kerner for their generosity with interesting and worthwhile projects. I also thank my friends and colleagues for their help and tolerance throughout my research: Vivian Chang (who introduced me to Dominion and Elle), Zugen Chen, Bobby Chin, Michael Clark, Ascia Eskin, Bret Harry, Nils Homer, Janeen Ishii, Hane Lee, Hafsa Mamsa, Barry Merriman, Tove Olafsen, Aliz

Raksi Rao, Sergio Solorzano-Vargas, Cris Sosa, Kevin Squire, Sara Stanford, Sam Strom, Paige Taylor (especially for a memorable World Series game one at Fenway), Traci Toy, Chris Walthers, Jenny Wang, Richard Wang, Jijun Wan, Anna Wu, Jen Zieba, and Judith Zimmerman. I am especially grateful to my colleague and BFF Genevieve Kendall for her inspiration, friendship, advice, and assistance.

Curriculum Vitae

Michael Yourshaw

**Education**

Harvard College AB 1963

Harvard Law School JD 1971

Colorado State University BS 2006

**Employment**

Neaera Consulting Group, LLC, Fort Collins CO 2004-2006

Engineering Computer Consultants, Inc., Fort Collins CO 1999-2004

Wiley, Rein & Fielding, Washington DC 1983-1999

Kirkland & Ellis, Washington DC 1971-1983

United States Air Force 1964-1969

United Fruit Company, Boston MA 1963-1964

George Washington University, Washington DC 1962

United States Army 1960,1961

**Publications**

Wan J, Yourshaw M, Mamsa H, Rudnik-Schoneborn S, Menezes MP, Hong JE, Leong DW, Senderek J, Salman MS, Chitayat D, Seeman P, von Moers A, Graul-Neumann L, Kornberg AJ, Castro-Gago M, Sobrido MJ, Sanefuji M, Shieh PB, Salamon N, Kim RC, Vinters HV, Chen Z, Zerres K, Ryan MM, Nelson SF, Jen JC. Mutations in the RNA exosome component gene EXOSC3 cause pontocerebellar hypoplasia and spinal motor neuron degeneration. Nat Genet. 2012;44(6):704-8. Epub 2012/05/01. doi: 10.1038/ng.2254. PubMed PMID: 22544365; PubMed Central PMCID: PMC3366034.

Yourshaw M, Solorzano-Vargas RS, Pickett LA, Lindberg I, Wang J, Cortina G, Pawlikowska-Haddal A, Baron H, Venick RS, Nelson SF, Martin MG. Exome Sequencing Finds a Novel PCSK1 Mutation in a Child With Generalized Malabsorptive Diarrhea and Diabetes Insipidus. Journal of pediatric gastroenterology and nutrition. 2013;57(6):759-67. Epub 2013/11/28. doi: 10.1097/MPG.0b013e3182a8ae6c. PubMed PMID: 24280991.

Martin MG, Lindberg I, Solorzano-Vargas RS, Wang J, Avitzur Y, Bandsma R, Sokollik C, Lawrence S, Pickett LA, Chen Z, Egritas O, Dalgic B, Albornoz V, de Ridder L, Hulst J, Gok F, Aydogan A, Al-Hussaini A, Gok DE, Yourshaw M, Wu SV, Cortina G, Stanford S, Georgia S. Congenital proprotein convertase 1/3 deficiency causes malabsorptive diarrhea and other endocrinopathies in a pediatric cohort. Gastroenterology. 2013;145(1):138-48. doi: 10.1053/j.gastro.2013.03.048. PubMed PMID: 23562752.

Pickett LA, Yourshaw M, Albornoz V, Chen Z, Solorzano-Vargas RS, Nelson SF, Martin MG, Lindberg I. Functional consequences of a novel variant of PCSK1. PLoS One. 2013;8(1):e55065. doi: 10.1371/journal.pone.0055065. PubMed PMID: 23383060; PubMed Central PMCID: PMC3557230.

Rudnik-Schoneborn S, Senderek J, Jen JC, Houge G, Seeman P, Puchmajerova A, Graul-Neumann L, Seidel U, Korinthenberg R, Kirschner J, Seeger J, Ryan MM, Muntoni F, Steinlin M, Sztriha L, Colomer J, Hubner C, Brockmann K, Van Maldergem L, Schiff M, Holzinger A, Barth P, Reardon W, Yourshaw M, Nelson SF, Eggermann T, Zerres K. Pontocerebellar hypoplasia type 1: clinical spectrum and relevance of EXOSC3 mutations. Neurology. 2013;80(5):438-46. doi: 10.1212/WNL.0b013e31827f0f66. PubMed PMID: 23284067; PubMed Central PMCID: PMC3590055.

Kerner B, Rao AR, Christensen B, Dandekar S, Yourshaw M, Nelson SF. Rare genomic variants link bipolar disorder to CREB regulated intracellular signaling pathways. Frontiers in Psychiatry. 2013. doi: 10.3389/fpsyt.2013.00154.

# CHAPTER ONE

# Introduction

### *1.1 High-throughput exome sequencing and rare Mendelian disorders*

The basic laws of monogenic inheritance were first set forth by Gregor Mendel in 1865 (1). In 1910 Thomas Hunt Morgan discovered that chromosomes were the physical and mechanistic basis of Mendelian inheritance (2) and in 1953 Watson and Crick determined the double helix structure of DNA (3). After these fundamental discoveries a framework for associating genetic mutations with human disease was in place, but DNA could not be reliably sequenced in the laboratory for several decades. In 1974 Fredrick Sanger developed the chain-terminator method of DNA sequencing (4). However, even today one of the most advanced Sanger sequencing machines (the ABI 3730xl) can produce only 2100 kilobases of data per day, or 0.07% of the human genome. Two further developments revolutionized the field of medical genetics: publication of the complete human genome in 2004 (5-7) based on automated Sanger technology; and the development of next generation high-throughput DNA sequencing technologies such as the Solexa/Illumina Genome Analyzer in 2006. In contrast to Sanger sequencing, the Genome Analyzer could sequence a billion bases per run in ten days (8). Today, the widely used Illumina HiSeq 2500 next-generation DNA sequencing platform can generate 100 billon bases in 27 hours, enough to resequenced an entire human genome.

While Sanger sequencing processes a single fragment at a time, high throughput platforms process billions of fragments in parallel. A DNA library is prepared by random fragmentation of genomic DNA into templates, which are then ligated to adapters and PCR amplified (9). The templates are immobilized on a solid support (glass slides or beads) and clonally amplified *in situ* by solid phase amplification, specifically bridge PCR on the Illumina platform, to create billions of colonies (10). The Illumina platform then performs one cycle of synthetic sequencing with fluorescent reversible terminator deoxyribonucleotides for each base position in parallel on all colonies. Images of the slide surface are analyzed to generate high-confidence base calls (11).

The availability of this powerful technology inspired us to tackle daunting problems in medical genetics that hitherto would have been extremely difficult to solve. Specifically, we were interested in gaining a better understanding of the molecular basis of inherited diseases. Genetic diseases are simplistically divided into those that are polygenic, complex, and common, on the one hand, and monogenic, simple, and rare, on the other. Multiple genes acting through a complex network of pathways control complex diseases.  Monogenic diseases are often called 'Mendelian disorders' because they have certain characteristics of the garden peas studied by Mendel: a particular genotype at a locus is both necessary and sufficient for a phenotype to be expressed under the normal range of genetic and environmental backgrounds (12). Humans inherit two copies of the 22 non-sex chromosomes (autosomes), one from each parent, and thus may have the same allele (homozygosity) or two different alleles (heterozygosity) at a given locus. Mendelian patterns of disease inheritance may be dominant, where an affected individual need only inherit a disease causing allele from one parent, or recessive, where both parents must transmit a disease causing allele. The type of chromosome further categorizes Mendelian patterns: the non-sex chromosomes (autosomes) and the X or Y sex chromosomes. Disease can also be inherited in a not strictly Mendelian manner via the maternally transmitted mitochondrial chromosome. The reality is more complex than this simplified description, as a single specific allele, different alleles in the same gene, and different alleles in different genes all give rise to Mendelian inheritance patterns in a given pedigree. Moreover, a genetically caused disease may arise in an individual sporadically as a result of a de novo mutation, typically a dominant acting one.

The reference human genome is a haploid representation of each chromosome, with a relatively complete list of its nucleotides, assembled from a number of individuals chosen primarily for reasons of technical sequencing quality and not intended to represent a typical or perfectly healthy person. High-throughput sequencing will find a large number of variants from the reference in every subject, most of which are phenotypically insignificant. Accordingly, we decided at the outset to explore this new technology by focusing on rare, debilitating, congenital diseases with a distinct phenotype, to empower us to find, in a small sample size, a causative allele from among thousands of false positives. First, a distinct phenotype observed in early childhood is more likely to be caused by inborn genetic variants than by environmental factors acting over a lifetime. Second, a rare incidence of a disorder implies that it is caused by a rare allele, and we could therefore filter out variants that were sufficiently common in the population that they could not possibly cause the disease. Third, individuals affected by a severely debilitating condition in childhood would not generally have survived to reproductive age absent heroic interventions with modern medicine, which would insure that recessive alleles would remain heterozygous in the population and dominant acting alleles would not be found. Fourth, if the pedigree is consistent with autosomal recessive inheritance or dominant acting de novo mutations, sequencing candidate variants in parents and affected or unaffected siblings would allow us to exclude variants that did not segregate with the disease. We made a further simplifying assumption that the disease would be fully penetrant, *i.e.*, a person with the genotype would always manifest the disease. Congenital malabsorptive diarrhea, intestinal pseudo-obstruction, and pontocerebellar hypoplasia satisfied these criteria. A presented in the following chapters, we found homozygous or compound heterozygous mutations exhibiting recessive inheritance and dominant acting de novo heterozygous mutations in these diseases. These disorders may be distinguished from comparatively common diseases, such as obesity, heart disease, type 2 diabetes mellitus, and others, which typically manifest well after infancy. It is hypothesized that the genetic component of these diseases can explained by a combination of common variants in one or more genes, each with a small effect or perhaps by multiple extremely rare variants with a stronger effect. This class of disease has been studied with large case-control associations studies, which search for the difference in allele frequency of polymorphic markers between unrelated groups of affected and unaffected individuals or within families. Association studies may require costly genotyping of thousands of individuals to have sufficient power to detect subtle effects and, like linkage analysis, lack resolution at the nucleotide level.

A practical concern regarding the use of high-throughput sequencing for modest studies with limited budgets is achieving sufficient sequencing depth, *i.e.*, the number of independent fragments that support a genotype call at a given locus. Multiple observations per base are necessary (a minimum of 10-20 to sensitively detect heterozygous variants) to make it likely that both alleles of a heterozygous locus will be observed and also to account for inevitable errors in the sequencing process. Although the per base cost of sequencing is low, there are more than 6 billion bases in a diploid human genome. Thus, the cost of sequencing the genomes of many cases at a depth sufficient to detect rare variants is considerable. Furthermore, because the number of reads varies by orders of magnitude among loci, it is necessary to get a mean coverage >100X to achieve 20X coverage for >90% of targeted bases. For this reason, we initially explored the development of strategies to enrich sequencing libraries for regions of interest in order to minimize the cost of sequencing less informative regions. This was a well-recognized difficulty and commercial reagents became available to meet the demand, obviating the need for us to pursue an in-house solution. A reasonable strategy for the discovery of rare alleles responsible for Mendelian phenotypes is to sequence only the protein coding regions of the genome (the 'exome'), which represent fewer than 2% of all bases (15). Most known genetic causes of Mendelian diseases affect protein-coding regions (16), and there is reason to believe that many rare missense alleles and small insertions and deletions (indels) in the exome have a functional consequence or are damaging (17). Promoters, enhancers, short RNAs, and other regulatory elements outside the exome doubtless govern some disorders, but variants in these regions are comparatively difficult to interpret.

Thus, the exome is an attractive target for an initial sub-genomic screen (18). Another consideration is that regions outside the coding DNA sequences (CCDs) have been noted to perform less efficiently in capture sequence experiments (19). When linkage or homozygosity mapping identify a particular region of interest it is possible to develop a custom probe set to enrich the sequencing library for only that region. Nonetheless, it still may be cost-effective to use a standard exome probe set unless a large number of samples are involved.

A library can be enriched for exome fragments with a capture method that uses biotinylated probes or 'baits' (RNA in the Agilent kit, DNA in the Illumina and Nimblegen kits) to fish targets out of a 'pond' of DNA fragments. In the Agilent process RNA is transcribed from PCR-amplified oligodeoxynucleotides originally synthesized on a microarray, generating sufficient bait for multiple captures at concentrations high enough to drive the hybridization (20). After the library and probes are hybridized, magnetic beads on the probes select probe-exome hybrids for sequencing. In our experience ~76% of bases map on or near a bait and the baited region is enriched 32-fold relative to the remainder of the genomic background.

A high-throughput sequencing instrument has a minimum unit of production determined by how many DNA clusters are processed in parallel. For example, in a typical configuration of an Illumina HiSeq 2500 instrument, two flowcells can run simultaneously, each flowcell being divided into eight lanes. Thus one lane is the minimum platform unit and a single lane can sequence up to 180 million paired end fragments. This is sufficient sequencing capacity to get satisfactory coverage on three or more exomes. Accordingly, we could sequence more samples for almost the same cost if we could multiplex samples in a lane and computationally identify the samples in downstream processing. We developed a method of Hamming code based DNA barcodes that were concatenated to the adapters, but switched to commercially available barcoded adapters when they became available to allow efficient exome sequencing.

### *1.2 Annotation of variants identified by high-throughput exome sequencing*

In almost all of the experiments we performed, the sequencing platform produced paired-end reads of 100 bases from either end of library fragments that are ~700 bases long. Extensive downstream processing is necessary to transform the unmapped raw reads into a useful dataset for variant discovery. The steps included: demultiplexing barcoded reads to separate reads by sample; removing PCR duplicates to prevent overrepresented fragments from biasing allele counts; recalibrating base quality scores to improve accuracy by analyzing the covariation among reported quality score, position within read, dinucleotide, and probability of mismatching the reference genome; mapping the fragments to the GRCh37 human reference genome; calling genotypes; assigning a well-calibrated probability of being true to each variant call in a call set (under a Gaussian mixture model using the variables inbreeding coefficient, quality by depth, mapping quality map sum test, mapping quality, read position rank sum test, and Fisher strand); and homozygosity block identification. These functions are performed by several software packages, including Picard (21), Novoalign (22), the Genome Analysis Toolkit (23, 24), Samtools (25), and PLINK (26, 27). We developed pipeline software that could keep track of case IDs, samples, libraries, machine runs, lanes, barcodes and other experiment-related metadata, coordinate the parallel execution of the programs on a compute cluster, and manage the output files. The pipeline software was written in Python and Scala, is modular, and requires minimal manual intervention once the metadata has been entered. An SQL Server stores metadata and the program results for downstream analysis.

The output of such a sequencing pipeline is a Variant Call Format (VCF) (28, 29) file that succinctly and systematically describes the genomic location, dbSNP ID, reference and alternate alleles, genotype, and other information related to each variant. For an exome, a VCF file typically consists of over 20,000 individual protein coding variants, and >50,000 records to account for the effects of variants on different transcripts of the same gene.

A basic VCF file does not contain most of the information that will be needed by a physician or researcher, such as the transcript and gene that contain the variant, the effect, if any, on protein encoding (synonymous, missense, nonsense) or structure, the likelihood that the variant is damaging, association with diseases or phenotypes, tissue expression data, or phenotypes in model organisms. There are several applications that can add such annotations to a VCF file, each with strengths an weaknesses (30). One characteristic of most of these tools is that they have little or no flexibility to include customized user-defined annotations. Furthermore, while on-line tools, such as SeattleSeq (15), have the advantage of simplicity of use, they may not be appropriate for confidential patient data or proprietary intellectual property.

We developed a custom annotator, which we call 'VAX' (Variant Annotator eXtras) that runs on local servers, is not heavily dependent on an outside researcher for software development and maintenance, and has a simple, modular mechanism for adding new features. We used the Ensembl Variant Effect Predictor (VEP) (31) as an engine. The VEP annotates variants with transcript and protein consequences including estimates of the extent of protein damage as computed by SIFT (32-36), PolyPhen (37-39), and Condel (40) We incorporated additional annotations from datasets such as Online Mendelian Inheritance in Man (OMIM) (41), the Human Gene Mutation Database (HGMD pro, BIOBASE Biological Databases), the Universal Protein Resource (UniProt) (42), KEGG Pathways (43), RefGene (44), the MitoCarta Inventory of Mammalian Mitochondrial Genes (45), Mouse Genome Informatics (MGI) (46)(47) and the Human Protein Atlas (HPA) (47), as well as allele frequencies and statistics on the number of damaging variants per gene.

Several factors were decisive in adopting the Ensembl VEP as the underlying engine. Ensembl, a joint scientific project between the European Bioinformatics Institute and the Wellcome Trust Sanger Institute, provides access to genomic annotation for numerous species stored on a MySQL database that can be accessed programmatically via a Perl application programming interface (API). The database is supported by a large professional organization, is updated regularly, and can be accessed remotely or by downloading a local copy. The Ensembl database and VEP have a large and active user community, which provide excellent and timely advice and support. The VEP is a mature open source Perl script that can be run locally, connected either to the remote Ensembl database or a local copy thereof, or with some limitations used with a local cache.

In addition to its use in our research projects, VAX is used routinely for CLIA/CAP-accredited whole-exome sequencing by the UCLA Clinical Genomics Center, which has processed more than 1000 exomes to date (48). VAX is also used by other researchers at UCLA, for example in a study of bipolar disorder in a family of four affected siblings (14). This study identified variants in genes that encode proteins with significant regulatory roles in the ERK/MAPK and CREB-regulated intracellular signaling pathways and supported the hypothesis that multiple rare, damaging mutations in genes functionally related to a common signaling pathway may contribute to the manifestation of bipolar disorder.

### *1.3 Congenital gastrointestinal disorders*

Congenital gastroenterological disorders may be caused by genetic mutations, environmental factors, or a combination of both. These disorders fall into three broad categories: diarrheal, motility, and obstructive.

Congenital diarrheal disorders (CDDs) are a set of enteropathies caused by inherited or sporadic genetic mutations that generally manifest soon after birth or in early childhood. The presenting symptom is chronic diarrhea that often requires total parenteral nutrition (TPN).

These patients frequently endure a complex and costly diagnostic odyssey that commonly fails to produce a definitive diagnosis (49).

Congenital motility disorders are a heterogenous group of disorders affecting gut neuromuscular function, which typically present with symptoms of vomiting, constipation or diarrhea, and abdominal pain (50). These disorders account for a significant portion of pediatric cases of intestinal failure (51). Hirschsprung disease (aganglionic megacolon) is a complex genetic disease caused by both rare and common mutations in RET and related genes. Linkage analysis, homozygosity mapping, and case-control association studies have all contributed to identifying these genes (52). Non-Hirschsprung cases of congenital intestinal motility failure are grouped under the term 'chronic intestinal pseudo-obstruction' (CIPO) and also include a related disorder, megacystis-microcolon-intestinal hypoperistalsis syndrome (MMIH).

Congenital intestinal obstructive disorders (atresia and stenosis) involve narrowed, blocked or disconnected intestine. Although genetic mutations are believed to be responsible for some congenital intestinal atresias and the disorder is associated with cystic fibrosis and Down syndrome, no genetic cause was discovered without the use of exome sequencing.

In the current research we assessed whole-exome sequencing as a method for identifying casual mutations in CDD patients, and for the discovery of novel genes that cause CIPO.

### *1.4 A pilot study of high-throughput exome sequencing to identify the molecular basis of congenital diarrheal disorders*

Congenital diarrheal disorders (CDDs) are rare diseases with serious, even life-threatening, consequences that impose massive diagnostic and treatment costs as well as great emotional stress on patients and their families. Until recently, little was known of the genetic etiology of these diseases, yet identification of a casual mutation can lead to improved management of the disease and inform research efforts to develop new treatment modalities. Prior to the availability of exome sequencing, CDDs with known genetic causes included secretory chloride diarrhea caused by mutations in the *SLC26A3* gene (53), microvillus inclusion disease caused by mutations in the *MYO5B* gene (54), a syndromic form of congenital secretory sodium diarrhea caused by mutations in the *SPINT2* gene (55), malabsorptive congenital diarrhea caused by mutations in the *NEUROG3* gene (56), congenital tufting enteropathy, caused by mutations in the *EPCAM* gene (57), and early-onset chronic diarrhea, caused by mutations in the *GUCY2C* gene (58).

The discovery of the aforementioned genes involved arduous genetic and molecular studies. In the 1990s genome wide linkage analysis, using 300-500 microsatellite markers, offered a method of identifying disease susceptibility loci. This approach was time-consuming and limited to a resolution of down to about 10Mb, a length, which can harbor over 100 genes (59). To achieve adequate power to find linkage it was necessary to genotype one or more extended families or a very large number of nuclear families. For example, linkage disequilibrium and genetic linkage as determined by this technique in Finnish families indicated that an unknown gene near the cystic fibrosis transmembrane regulator gene (CFTR) was probably associated with secretory chloride diarrhea (60). Subsequently, cloning of the linkage region identified four known genes, two of which were considered to be functionally relevant (61). Finally, segregation of mutations in the *SLC26A3* gene with the disorder in a large number of patients confirmed that such mutations cause the disorder (53). The development of highly parallel genotyping based on arrays with probes for thousands of single nucleotide polymorphisms (SNPs) enabled more efficient genotyping with a 10-fold or better improvement in resolution compared to microsatellite base methods. This technology, applied to extended kindreds, enabled the discovery of mutations in the causative genes for microvillus inclusion disease, syndromic congenital secretory sodium diarrhea, congenital tufting enteropathy, and early-onset chronic diarrhea. A candidate gene resequencing approach, founded on a mouse model, identified mutations in *NEUROG3* as the cause of malabsorptive congenital diarrhea (56).

Despite much progress in uncovering the molecular basis for this class of diseases there remained many patients with strong indications of a genetic etiology for whom no mutation in these genes could be identified. We hypothesized that a large fraction of these cases had undetected mutations in genes known to be associated with CDD, and others would have mutations in discoverable novel genes. An additional working hypothesis was that some casual mutations would be in the protein coding portion of the genome, and the phenotype would be recessive the mutation would have high penetrance. Accordingly, we sought to determine the effectiveness of whole-exome sequencing to identify, in genes that have been reported to be associated with CDD, the molecular causes of the disease in a cohort of patients with CDDs that had defied conventional diagnostic methods.

We chose 45 patients from 38 families for exome sequencing. Inclusion criteria were a diagnosis of congenital diarrhea and probability that the disease had a genetic cause (typically consanguineous parents or affected family members). Patients were excluded if they had a confirmed genetic diagnosis or a clinical presentation already suggesting a mutation in a gene known to cause congenital diarrhea. These patients were of diverse ethnic backgrounds and had clinical presentations of generalized malabsorptive diarrhea, selective nutrient malabsorption, secretory diarrhea, and infantile IBD.

We performed whole-exome sequencing, as described above, validated variants found in candidate genes with Sanger sequencing, and Sanger sequenced available relatives to ensure that genotypes segregated with the phenotype. We originally sequenced seven cases in five families on the ABI SOLiD platform. This method identified a casual mutation in *PCSK1* in one case (62), but failed to find interesting variants in the other four families. The SOLiD data appeared to be quite noisy (about twice the expected number of raw variants), so we resequenced the unsolved cases and all further cases on the Illumina platform. In all, we identified 31 different mutations in 21 families (27 cases) that we considered likely to be the cause of the disease because they were deemed to be damaging and were found in genes known to associated with the phenotype or reported in the literature during the course of the study. We concluded that exome sequencing would be valuable for diagnosis of CDDs in a clinical setting. Importantly, we identified novel candidate genes in many of the remaining cases; work is in progress to confirm these findings by functional studies *in vitro*, in model organisms, and/or in a model of intestinal tissue developed from patient intestinal stem cells or differentiated embryonic stem cell cultures.

### *1.5 PC1/3 deficiency*

Before our study, three reported cases indicated that the gene proprotein convertase subtilisin/kexin type 1 (*PCSK1*), which encodes the neuroendocrine convertase 1 precursor protein (PC1/3), is involved in disorders characterized by abnormal enteroendocrine development or function that manifest in generalized malabsorption (63-65). PC1/3 is a calcium-dependent serine endoprotease essential for the conversion of a variety of prohormones into their bioactive forms. It has a well-defined role of processing proinsulin in β cells of the pancreas (66); it is expressed richly in endocrine cells in the gut, where its function is obscure.

One of the patients in our CDD cohort was a perplexing child with congenital malabsorptive diarrhea and other presumably unrelated clinical problems (62). The patient (of consanguineous parentage) was initially assessed at three weeks of age for recurrent diarrhea and associated metabolic acidosis. At six days of age he was transferred to the intensive care nursery due to poor peripheral perfusion and indirect hyperbilirubinemia. DNA genotyping of CFTR for cystic fibrosis was negative for the 97 mutations most commonly observed. He was hospitalized 21 times before age three, including presentations of hypovolemic shock with profound metabolic acidosis, central venous catheter occlusions, heparin-induced thrombocytopenia, multiple deep venous thrombi, pneumonia, hyperglycemia, and left ventricular dysfunction.

Exome sequencing analysis as described above identified a novel Tyr343Ter mutation in *PCSK1* that terminated the protein within its catalytic domain. This nonsense mutation rendered the gene product undetectable in either cells or secreted into media, probably due to nonsense-mediated decay, and a caused a total lack of enzyme activity. Immunohistochemistry for PC1/3-expressing enteroendocrine cells was negative.

The identification of a mutation in *PCSK1* suggested a specific clinical diagnosis that includes diabetes insipidus (DI) as a component. Indeed, follow up with the patient confirmed the presence of DI, and intranasal desmopressin (DDAVP) improved the patient's condition significantly.

Contemporaneously with the pilot exome study, we determined the clinical features of 13 other children with PC1/3 deficiency caused by *PCSK1* mutations (67). We performed Sanger sequencing analysis of *PCSK1* and measured enzymatic activity of recombinant PC1/3 proteins. We identified a pattern of endocrinopathies that develop in an age-dependent manner. Neonates had severe malabsorptive diarrhea and failure to thrive, required prolonged parenteral nutrition support, and had high mortality. Additional endocrine abnormalities developed as the disease progressed, including diabetes insipidus, growth hormone deficiency, primary hypogonadism, adrenal insufficiency, and hypothyroidism.

We further explored the PC1/3 landscape by searching for potentially consequential variants in the dbSNP (68), 1000 Genomes Project (69), NHLBI (70), and NIEHS (71) datasets. We found that a novel Arg80GLN variant (rs1799904) both exhibits adverse effects on PC1/3 activity and is prevalent in the population at a low level, suggesting that further biochemical and genetic analysis would be warranted to assess its contribution to the risk of metabolic disease within the general population.

### *1.6 Application of high-throughput exome sequencing to identify a probable genetic cause of sporadic chronic intestinal pseudo-obstruction*

In addition to the CDD cohort, we also sequenced a cohort of 20 cases diagnosed with CIPO or MMIH in the hope of finding one or more novel genes responsible for these rare conditions. Chronic intestinal pseudo-obstruction (CIPO) is a heterogenous set of diseases characterized by repetitive episodes or continuous symptoms of intestinal obstruction, in the absence of a lesion that occludes the lumen of the gut (50, 51). A small fraction of cases are secondary to organic, systemic, or metabolic diseases, but the majority are primary and may be myopathic, mesenchymopathic, or neuropathic, depending upon whether predominant abnormalities are found in the enteric nervous system, interstitial cells of Cajal (ICC), or intestinal smooth muscle (72). A related disorder, megacystis-microcolon-intestinal hypoperistalsis syndrome (MMIH), is characterized by constipation and urinary retention, microcolon, giant bladder (megacystis), intestinal hypoperistalis, hydronephrosis, and dilated small bowel (73).

Congenital forms of CIPO are rare and can be life-threatening; congenital CIPO is an important cause of intestinal failure, for which the only treatment may be complete visceral transplantation (74). Congenital CIPO may sometimes be due to prenatal exposure to toxins such as alcohol or narcotics. A handful of familial cases of CIPO have been reported with autosomal dominant (with variable penetrance), autosomal recessive, and X-linked modes of inheritance (75-85). It is well known that mutations in mitochondrial tRNA genes, *POLG* (polymerase (DNA directed), gamma), and *TYMP* (thymidine phosphorylase), which are expressed in the mitochondrion, cause a severe form of CIPO requiring frequent and long-term parenteral nutrition and with frequently fatal digestive and neurologic complications.

Mitochondrial disorders may account for ~19% of CIPO cases (86). Contrawise, it is rare for CIPO to be the principal clinical manifestation of a mitochondrial disorder (87). Primary defects of the mitochondrial oxidative phosphorylation pathway are phenotypically heterogenous, and affecting multiple organs, typically with nervous system and skeletal or ocular muscle dysfunction (88). Mitochondrial neurogastrointestinal encephalomyopathy (MNGIE) is a rare, autosomal recessive syndrome due to the loss of thymidine phosphorylase activity associated with loss-of-function mutations in *TYM*P (89-93). Mutations in *POLG*, the mitochondrial myopathy, epilepsy, lactic acidosis, and strokelike episodes ('MELAS') mutation in the tRNA$^{leu(UUR)}$ gene, or mutations in the tRNA$^{lys}$ gene are sometimes associated with CIPO (85, 94-101). Still, congenital CIPO is usually sporadic and prior to the advent of exome sequencing, no non-mitochondrial gene had been convincingly associated with primary sporadic CIPO.

We hypothesized for this study that the mode of inheritance would be recessive, de novo, or mitochondrial, and that mutation effects would be fully penetrant. Identifying de novo variants with a dominant effect on phenotype typically requires sequencing of parent-child trios to eliminate the large number of potential heterozygous variants. Unfortunately, we did not have parental DNA for most of these patients, but we believed there was a reasonable chance we could narrow down the potential de novo candidates by looking for mutations in the same gene in multiple patients. We further hypothesized that casual variants would be found in the protein coding or splicing regions of genes. Because failure of muscle function in the intestinal wall was believed to be a common cause of CIPO, we were particularly alert for mutations in genes that might affect muscle cell function, such as myosins, actins, and proteins that bind or regulate myosins and actins.

A few genes are present on the small circular mitochondrial DNA but many more are encoded by nuclear DNA and then localize to the mitochondrion. We developed a mitochondrial gene annotation in the VAX program for this study, based on the MitoCarta inventory of 1098 mouse genes encoding proteins with strong support of mitochondrial localization (45). We also developed an SQL query that generates a matrix of genes and cases to graphically display the number and ID of cases mutated in each gene. During the course of our study Lehtonen *et al.* reported that a missense variant in *ACTG*, encoding γ-enteric actin, segregated in a Finnish family with autosomal dominant familial visceral myopathy (FVM), a disorder that is subsumed within the broad definition of CIPO (102). Thus, we were particularly interested in mutations in this gene, and identified nine novel de novo mutations in the CIPO cohort. We also found a potential mitochondrial gene (*POLG*) compound heterozygous mutation of interest in the mitochondrial gene (*POLG*), but have not yet confirmed whether it is causal.

### *1.7 Application of high-throughput exome sequencing to discover an unsuspected gene, EXOSC3, that causes pontocerebellar hypoplasia and spinal motor neuron degeneration*

Another exome sequencing project involved the search for mutations causing a mysterious neuromuscular disorder that affected four siblings in a large family (103). The children were floppy at birth, had ocular motor apraxia, progressive muscle wasting, distal contractures, progressive microcephaly, growth retardation and global developmental delay, and never reached any motor milestone or spoke. Initially we were unable to categorize this condition, but after receiving the autopsy report on one subject, we suspected pontocerebellar hypoplasia (PCH), which is characterized by cerebellar hypoplasia or atrophy, variable pontine atrophy and progressive microcephaly with global developmental delay (104). Pontocerebellar hypoplasia type1 (PCH1) is a distinctive subtype of PCH, characterized by diffuse muscle wasting that is secondary to spinal cord anterior horn cell loss and cerebellar hypoplasia (105-108). Diagnosis of PCH1 is often delayed or never made because the combination of cerebellar and spinal motor neuron degeneration is not commonly recognized, and the presentation of diffuse weakness and devastating brain involvement is atypical of classical proximal spinal muscular atrophy (SMA) (109). The literature contains only a handful of descriptions of case series (110-113) and reports of PCH1 (114-120). Prior to our study, a causative gene had not been identified in the majority of individuals with PCH1. Recessive mutations have been found in *VRK1* (encoding vaccinia-related kinase 1) (121), *RARS2* (encoding mitochondrial arginyl-tRNA synthetase 2) (104) and *TSEN54* (encoding tRNA splicing endonuclease 54) (122) in single individuals with PCH1. In PCH without SMA, *TSEN54* mutations account for most cases of PCH2 and PCH4 (104, 123), and *RARS2* mutations have been found in two families with PCH6 (124, 125).

Array-based identity by descent analysis of the four affected siblings, three healthy siblings, and their parents, highlighted candidate regions in four sub-chromosomal loci with more than 100 candidate genes in total. Exome sequencing (Illumina IIx single end 76 base reads) of the four affected siblings yielded a single candidate variant, g.9:37783990T>G (c.395A>C, p.Asp132Ala) in the *EXOSC3* gene (encoding exosome component 3). The variant was homozygous in all four affected siblings, segregated with the disorder upon Sanger sequencing of unaffected relatives, and was within one of the intervals identical by descent in all affected siblings; the parents were heterozygous for the variant. The variant was at a completely conserved locus.

Exosome component 3, also known as the ribosomal RNA–processing protein 40 (Rrp40), is a core component of the human RNA exosome complex (distinct from exosome vesicles). RNA exosomes are multi-subunit complexes conserved throughout evolution (126) and are emerging as the major cellular machinery for processing, surveillance and turnover of a diverse spectrum of coding and noncoding RNA substrates essential for viability (127). The exosome's nine subunits are arranged in a two-layered ring; the bottom 'hexamer' layer is formed by six subunits. Rrp40 is one of three RNA binding subunits that comprise the 'cap' of the exosome complex (128, 129).

Eight probands with PCH1 out of twelve additional families had homozygous or compound heterozygous mutations in *EXOSC3* and all available parents were heterozygous. The Asp132Ala mutation was present in six of these families, homozygous in three families and compound heterozygous in another three. One case was homozygous for Gly31Ala, another compound heterozygous for Gly31Ala and Trp238Arg, and the remainder compound heterozygous for Asp132Ala plus 99fs*11, Ala139Pro, or intronic c.475–12A>G causing exon skipping. Genotyping the original family and two others revealed an identical short 1 cM region flanking the g.9:37783990T>G locus, suggesting a distant ancestry for the mutation. Interestingly, another founder mutation, Gly31Ala, also seen in two of our cases, was recently identified as a cause of severe PCH1 among the Czech Roma (130).

Knockdown of *exosc3* expression in zebrafish embryos by antisense morpholinos led to a dose-dependent phenotype of a short, curved spine and small brain with poor motility and even death by 3 days post fertilization. Whole-mount *in situ* hybridization showed decreased expression of *atoh1a* (a marker specific for dorsal hindbrain progenitors) in the upper and lower rhombic lips and a lack of expression of *pvalb7*, which is specific for differentiated cerebellar Purkinje neurons (131). The abnormal phenotype from *exosc3*-specific morpholino injection was largely rescued by co-injection with wild-type zebrafish *exosc3* mRNA whereas co-injection with mRNA containing the mutation was ineffective in rescue.

In a companion study, biallelic mutations in *EXOSC3* were detected in 10 of 27 families (37%) (132). The mutation-positive subjects typically presented with normal pregnancy, normal birth measurements, and relative preservation of brainstem and cortical structures. Psychomotor retardation was profound in all patients but lifespan was variable, with 3 subjects surviving beyond the late teens. Abnormal oculomotor function was commonly observed in patients surviving beyond the first year. Major clinical features previously reported in PCH1, including intrauterine abnormalities, postnatal hypoventilation and feeding difficulties, joint contractures, and neonatal death, were rarely observed in mutation-positive infants but were typical among the mutation-negative subjects, indicating that variability in survival and clinical severity is correlated with the genotype.

The same homozygous Asp132Ala mutation as that in our original family was reported in four patients with muscle hypertonia, developmental delay, spinal anterior horn involvement, and prolonged survival, consistent with a milder form of PCH1, suggesting phenotypic variability possibly by caused by protective factors in the genetic background (133). Another recently reported case from Bangladesh with Asp132Ala and a novel Val80Phe mutation suffered from intellectual disability, early onset spasticity, and cerebellar atrophy (134).

### *1.8 Overview of the chapters*

Chapter 1 is this introduction. Chapter 2 describes the methods used in the VAX program for rich annotation of DNA sequencing variants by leveraging the Ensembl Variant Effect Predictor with plugins. This paper was submitted to Briefings in Bioinformatics and is under review. Chapter 3 reports the results of a pilot study to use whole-exome sequencing for the identification of casual mutations in congenital diarrheal disorders. Chapter 4 describes in more detail one of the cases analyzed in the pilot study of Chapter 3, where exome sequencing found a novel *PCSK1* mutation in a child with generalized malabsorptive diarrhea and diabetes insipidus. This paper was published in the Journal of Pediatric Gastroenterology and Nutrition (62). Chapter 5 contains an overview of *PCSK1* variants in the human population and the functional consequences of a novel variant of *PCSK1*. This paper was published in PLoS One (135). Chapter 6 presents evidence from exome sequencing that mutations in *ACTG2* are a probable significant cause of chronic intestinal pseudo-obstruction. Chapter 8 shows the application of whole-exome sequencing in the field of neurology and describes our discovery that mutations in the RNA exosome component gene *EXOSC3* cause pontocerebellar hypoplasia and spinal motor neuron degeneration. This paper was published in Nature Genetics (103). Chapter 10 is the conclusion.

# References

1. Mendel G. Versuche über Plflanzen-hybriden. Des naturforschenden Vereines; 1865; Brünn: Verhandlungen des naturforschenden Vereines in Brünn; 1866.

2. Morgan TH. Sex Limited Inheritance in Drosophila. Science. 1910;32(812):120-2. Epub 1910/07/22. doi: 10.1126/science.32.812.120. PubMed PMID: 17759620.

3. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature. 1953;171(4356):737-8. Epub 1953/04/25. PubMed PMID: 13054692.

4. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463-7. Epub 1977/12/01. PubMed PMID: 271968; PubMed Central PMCID: PMC431765.

5. International Human Genome Sequencing C. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931-45. Epub 2004/10/22. doi: 10.1038/nature03001. PubMed PMID: 15496913.

6. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, International Human Genome Sequencing C. Initial

sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921. Epub 2001/03/10. doi: 10.1038/35057062. PubMed PMID: 11237011.

7. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. The sequence of the human genome. Science. 2001;291(5507):1304-51. Epub 2001/02/22. doi: 10.1126/science.1058040. PubMed PMID: 11181995.

8. Illumina Inc. Solexa Technology 2013 [cited 2013 2013-11-23]. Available from: http://www.illumina.com/technology/solexa_technology.ilmn.

9. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008;26(10):1135-45. doi: 10.1038/nbt1486. PubMed PMID: 18846087.

10. Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res. 2006;34(3):e22. Epub 2006/02/14. doi: 10.1093/nar/gnj023. PubMed PMID: 16473845; PubMed Central PMCID: PMC1363783.

11. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X,

Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara ECM, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456(7218):53-9. Epub 2008/11/07. doi: 10.1038/nature07517. PubMed PMID: 18987734; PubMed Central PMCID: PMC2581791.

12. Strachan T, Read AP, Strachan T. Human molecular genetics. 4th ed. New York: Garland Science; 2011. xxv, 781 p. p.

13. Goodwin FK, Jamison KR. Manic-Depressive Illness. New York: Oxford University Press; 1990.

14. Kerner B, Rao AR, Christensen B, Dandekar S, Yourshaw M, Nelson SF. Rare genomic variants link bipolar disorder to CREB regulated intracellular signaling pathways. Frontiers in Psychiatry. 2013. doi: 10.3389/fpsyt.2013.00154.

15. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009;461(7261):272-6. Epub 2009/08/18. doi: nature08250 [pii]
10.1038/nature08250. PubMed PMID: 19684571.

16. Stenson PD. The Human Gene Mutation Database: 2008 update. Genome Med. 2009;1:13.

17. Kryukov GV, Pennacchio LA, Sunyaev SR. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet. 2007;80(4):727-39. Epub 2007/03/16. doi: 10.1086/513473. PubMed PMID: 17357078; PubMed Central PMCID: PMC1852724.

18. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 2011;12(11):745-55. Epub 2011/09/29. doi: 10.1038/nrg3031. PubMed PMID: 21946919.

19. Bainbridge MN, Wang M, Wu Y, Newsham I, Muzny DM, Jefferies JL, Albert TJ, Burgess DL, Gibbs RA. Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. Genome Biol. 2011;12(7):R68. Epub 2011/07/27. doi: 10.1186/gb-2011-12-7-r68. PubMed PMID: 21787409.

20. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol. 2009;27(2):182-9. Epub 2009/02/03. doi: nbt.1523 [pii]
10.1038/nbt.1523. PubMed PMID: 19182786; PubMed Central PMCID: PMC2663421.

21. Picard 2013 [updated 2013-11-18]. Available from: http://picard.sourceforge.net.

22. Hercus C. Novocraft.com Novocraft 2013. Available from: http://www.novocraft.com/main/index.php.

23. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491-8. Epub 2011/04/12. doi: 10.1038/ng.806. PubMed PMID: 21478889; PubMed Central PMCID: PMC3083463.

24. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-303. Epub 2010/07/21. doi: 10.1101/gr.107524.110. PubMed PMID: 20644199; PubMed Central PMCID: PMC2928508.

25. Li H. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078-9.

26. Purcell S. PLINK 1.07. Available from: http://pngu.mgh.harvard.edu/purcell/plink/.

27. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559-75. Epub 2007/08/19. doi: 10.1086/519795. PubMed PMID: 17701901; PubMed Central PMCID: PMC1950838.

28. 1000 Genomes Project. VCF (Variant Call Format) version 4.1 2013 [updated 2013-10-09]. Available from: http://www.1000genomes.org/wiki/Analysis/Variant Call Format/vcf-variant-call-format-version-41.

29. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156-8. Epub 2011/06/10. doi: 10.1093/bioinformatics/btr330. PubMed PMID: 21653522; PubMed Central PMCID: PMC3137218.

30. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J, Trajanoski Z. A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform. 2013. doi: 10.1093/bib/bbs086. PubMed PMID: 23341494.

31. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010;26(16):2069-70. Epub 2010/06/22. doi: btq330 [pii] 10.1093/bioinformatics/btq330. PubMed PMID: 20562413; PubMed Central PMCID: PMC2916720.

32. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols. 2009;4(7):1073-81. Epub 2009/06/30. doi: 10.1038/nprot.2009.86. PubMed PMID: 19561590.

33. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001;11(5):863-74. Epub 2001/05/05. doi: 10.1101/gr.176601. PubMed PMID: 11337480; PubMed Central PMCID: PMC311071.

34. Ng PC, Henikoff S. Accounting for human polymorphisms predicted to affect protein function. Genome Res. 2002;12(3):436-46. Epub 2002/03/05. doi: 10.1101/gr.212802. PubMed PMID: 11875032; PubMed Central PMCID: PMC155281.

35. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812-4. Epub 2003/06/26. PubMed PMID: 12824425; PubMed Central PMCID: PMC168916.

36. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. Annual review of genomics and human genetics. 2006;7:61-80. Epub 2006/07/11. doi: 10.1146/annurev.genom.7.080505.115630. PubMed PMID: 16824020.

37. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 2002;30(17):3894-900. Epub 2002/08/31. PubMed PMID: 12202775; PubMed Central PMCID: PMC137415.

38. Sunyaev S, Ramensky V, Bork P. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. Trends Genet. 2000;16(5):198-200. Epub 2000/04/27. PubMed PMID: 10782110.

39. Sunyaev S, Ramensky V, Koch I, Lathe W, 3rd, Kondrashov AS, Bork P. Prediction of deleterious human alleles. Hum Mol Genet. 2001;10(6):591-7. Epub 2001/03/07. PubMed PMID: 11230178.

40. Gonzalez-Perez A, Lopez-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet. 2011;88(4):440-9. Epub 2011/04/05. doi: 10.1016/j.ajhg.2011.03.004. PubMed PMID: 21457909; PubMed Central PMCID: PMC3071923.

41. Online Mendelian Inheritance in Man OMIM®. Online Mendelian Inheritance in Man, OMIM® Baltimore, MD: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University; [2013-11-04]. Available from: http://omim.org/.

42. Apweiler R, Martin MJ, O'Donovan C, Magrane M, Alam-Faruque Y, Alpi E, Antunes R, Arganiska J, Casanova EB, Bely B, Bingley M, Bonilla C, Britto R, Bursteinas B, Chan WM, Chavali G, Cibrian-Uhalte E, Da Silva A, De Giorgi M, Dimmer E, Fazzini F, Gane P, Fedotov A, Castro LG, Garmiri P, Hatton-Ellis E, Hieta R, Huntley R, Jacobsen J, Jones R, Legge D, Liu WD, Luo J, MacDougall A, Mutowo P, Nightingale A, Orchard S, Patient S, Pichler K, Poggioli D, Pundir S, Pureza L, Qi GY, Rosanoff S, Sawford T, Sehra H, Turner E, Volynkin V, Wardell T, Watkins X, Zellner H, Corbett M, Donnelly M, van Rensburg P, Goujon M, McWilliam H, Lopez R, Xenarios I, Bougueleret L, Bridge A, Poux S, Redaschi N, Auchincloss A, Axelsen K, Bansal P, Baratin D, Binz PA, Blatter MC, Boeckmann B, Bolleman J, Boutet E, Breuza L, de Castro E, Cerutti L, Coudert E, Cuche B, Doche M, Dornevil D, Duvaud S, Estreicher A, Famiglietti L, Feuermann M, Gasteiger E, Gehant S, Gerritsen V, Gos A, Gruaz-Gumowski N, Hinz U, Hulo C, James J, Jungo F, Keller G, Lara V, Lemercier P, Lew J, Lieberherr D, Martin X, Masson P, Morgat A, Neto T, Paesano S, Pedruzzi I, Pilbout S, Pozzato M, Pruess M, Rivoire C, Roechert B, Schneider M, Sigrist C, Sonesson K, Staehli S, Stutz A, Sundaram S, Tognolli M, Verbregue L, Veuthey AL, Zerara M, Wu CH, Arighi CN, Arminski L, Chen CM, Chen YX, Huang HZ, Kukreja A, Laiho K, McGarvey P, Natale DA, Natarajan TG, Roberts NV, Suzek BE, Vinayaka CR, Wang QH, Wang YQ, Yeh LS, Yerramalla MS, Zhang J, Consortium U. Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Research. 2013;41(D1):D43-D7. doi: Doi 10.1093/Nar/Gks1068. PubMed PMID: WOS:000312893300007.

43. Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. Nucleic Acids Res. 2002;30(1):42-6. PubMed PMID: 11752249; PubMed Central PMCID: PMC99091.

44. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. 2012;40(Database issue):D130-5. doi: 10.1093/nar/gkr1079. PubMed PMID: 22121212; PubMed Central PMCID: PMC3245008.

45. Pagliarini DJ, Calvo SE, Chang B, Sheth SA, Vafai SB, Ong SE, Walford GA, Sugiana C, Boneh A, Chen WK, Hill DE, Vidal M, Evans JG, Thorburn DR, Carr SA, Mootha VK. A mitochondrial protein compendium elucidates complex I disease biology. Cell. 2008;134(1):112-23. doi: 10.1016/j.cell.2008.06.016. PubMed PMID: 18614015; PubMed Central PMCID: PMC2778844.

46. Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE, Mouse Genome Database G. The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. Nucleic Acids Res. 2012;40(Database issue):D881-6. doi: 10.1093/nar/gkr974. PubMed PMID: 22075990; PubMed Central PMCID: PMC3245042.

47. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S, Wernerus H, Bjorling L, Ponten F. Towards a knowledge-based Human Protein Atlas. Nat Biotechnol. 2010;28(12):1248-50. doi: 10.1038/nbt1210-1248. PubMed PMID: 21139605.

48. Lee H, Nelson SF. Rethinking clinical practice: clinical implementation of exome sequencing. Pers Med. 2012;9(8):785-7. doi: Doi 10.2217/Pme.12.101. PubMed PMID: WOS:000311977800001.

49. Terrin G, Tomaiuolo R, Passariello A, Elce A, Amato F, Di Costanzo M, Castaldo G, Canani RB. Congenital diarrheal disorders: an updated diagnostic approach. International journal of molecular sciences. 2012;13(4):4168-85. doi: 10.3390/ijms13044168. PubMed PMID: 22605972; PubMed Central PMCID: PMC3344208.

50. Hyman P, Thapar N. Chronic Intestinal Pseudo-Obstruction. In: Faure C, Di Lorenzo C, Thapar N, editors. Pediatric Neurogastroenterology: Humana Press; 2013. p. 257-70.

51. Antonucci A, Fronzoni L, Cogliandro L, Cogliandro RF, Caputo C, De Giorgio R, Pallotti F, Barbara G, Corinaldesi R, Stanghellini V. Chronic intestinal pseudo-obstruction. World journal of gastroenterology : WJG. 2008;14(19):2953-61. PubMed PMID: 18494042; PubMed Central PMCID: PMC2712158.

52. Alves MM, Sribudiani Y, Brouwer RW, Amiel J, Antinolo G, Borrego S, Ceccherini I, Chakravarti A, Fernandez RM, Garcia-Barcelo MM, Griseri P, Lyonnet S, Tam PK, van Ijcken WF, Eggen BJ, te Meerman GJ, Hofstra RM. Contribution of rare and common variants determine complex diseases-Hirschsprung disease as a model. Developmental biology. 2013;382(1):320-9. Epub 2013/05/28. doi: 10.1016/j.ydbio.2013.05.019. PubMed PMID: 23707863.

53. Hoglund P, Haila S, Socha J, Tomaszewski L, Saarialho-Kere U, Karjalainen-Lindsberg ML, Airola K, Holmberg C, de la Chapelle A, Kere J. Mutations of the Down-regulated in adenoma (DRA) gene cause congenital chloride diarrhoea. Nat Genet. 1996;14(3):316-9. Epub 1996/11/01. doi: 10.1038/ng1196-316. PubMed PMID: 8896562.

54. Muller T, Hess MW, Schiefermeier N, Pfaller K, Ebner HL, Heinz-Erian P, Ponstingl H, Partsch J, Rollinghoff B, Kohler H, Berger T, Lenhartz H, Schlenck B, Houwen RJ, Taylor CJ, Zoller H, Lechner S, Goulet O, Utermann G, Ruemmele FM, Huber LA, Janecke AR. MYO5B mutations cause microvillus inclusion disease and disrupt epithelial cell polarity. Nat Genet. 2008;40(10):1163-5. doi: 10.1038/ng.225. PubMed PMID: 18724368.

55. Heinz-Erian P, Muller T, Krabichler B, Schranz M, Becker C, Ruschendorf F, Nurnberg P, Rossier B, Vujic M, Booth IW, Holmberg C, Wijmenga C, Grigelioniene G, Kneepkens CM, Rosipal S, Mistrik M, Kappler M, Michaud L, Doczy LC, Siu VM, Krantz M, Zoller H, Utermann G, Janecke AR. Mutations in SPINT2 cause a syndromic form of congenital sodium diarrhea. Am J Hum Genet. 2009;84(2):188-96. Epub 2009/02/03. doi: 10.1016/j.ajhg.2009.01.004. PubMed PMID: 19185281; PubMed Central PMCID: PMC2668003.

56. Wang J, Cortina G, Wu SV, Tran R, Cho JH, Tsai MJ, Bailey TJ, Jamrich M, Ament ME, Treem WR, Hill ID, Vargas JH, Gershman G, Farmer DG, Reyen L, Martin MG. Mutant neurogenin-3 in congenital malabsorptive diarrhea. N Engl J Med. 2006;355(3):270-80. doi: 10.1056/NEJMoa054288. PubMed PMID: 16855267.

57. Sivagnanam M, Mueller JL, Lee H, Chen Z, Nelson SF, Turner D, Zlotkin SH, Pencharz PB, Ngan BY, Libiger O, Schork NJ, Lavine JE, Taylor S, Newbury RO, Kolodner RD, Hoffman HM. Identification of EpCAM as the gene for congenital tufting enteropathy. Gastroenterology. 2008;135(2):429-37. doi: Doi 10.1053/J.Gastro.2008.05.036. PubMed PMID: ISI:000258439900020.

58. Fiskerstrand T, Arshad N, Haukanes BI, Tronstad RR, Pham KD, Johansson S, Havik B, Tonder SL, Levy SE, Brackman D, Boman H, Biswas KH, Apold J, Hovdenak N, Visweswariah SS,

Knappskog PM. Familial diarrhea syndrome caused by an activating GUCY2C mutation. N Engl J Med. 2012;366(17):1586-95. Epub 2012/03/23. doi: 10.1056/NEJMoa1110132. PubMed PMID: 22436048.

59. Hearne CM, Ghosh S, Todd JA. Microsatellites for Linkage Analysis of Genetic-Traits. Trends in Genetics. 1992;8(8):288-94. doi: Doi 10.1016/0168-9525(92)90137-S. PubMed PMID: WOS:A1992JE96800009.

60. Kere J, Sistonen P, Holmberg C, de la Chapelle A. The gene for congenital chloride diarrhea maps close to but is distinct from the gene for cystic fibrosis transmembrane conductance regulator. Proc Natl Acad Sci U S A. 1993;90(22):10686-9. Epub 1993/11/15. PubMed PMID: 7504277; PubMed Central PMCID: PMC47842.

61. Hoglund P, Haila S, Scherer SW, Tsui LC, Green ED, Weissenbach J, Holmberg C, de la Chapelle A, Kere J. Positional candidate genes for congenital chloride diarrhea suggested by high-resolution physical mapping in chromosome region 7q31. Genome Res. 1996;6(3):202-10. Epub 1996/03/01. PubMed PMID: 8963897.

62. Yourshaw M, Solorzano-Vargas RS, Pickett LA, Lindberg I, Wang J, Cortina G, Pawlikowska-Haddal A, Baron H, Venick RS, Nelson SF, Martin MG. Exome Sequencing Finds a Novel PCSK1 Mutation in a Child With Generalized Malabsorptive Diarrhea and Diabetes Insipidus. Journal of pediatric gastroenterology and nutrition. 2013;57(6):759-67. Epub 2013/11/28. doi: 10.1097/MPG.0b013e3182a8ae6c. PubMed PMID: 24280991.

63. Jackson RS, Creemers JW, Farooqi IS, Raffin-Sanson ML, Varro A, Dockray GJ, Holst JJ, Brubaker PL, Corvol P, Polonsky KS, Ostrega D, Becker KL, Bertagna X, Hutton JC, White A, Dattani MT, Hussain K, Middleton SJ, Nicole TM, Milla PJ, Lindley KJ, O'Rahilly S. Small-intestinal dysfunction accompanies the complex endocrinopathy of human proprotein convertase 1 deficiency. The Journal of clinical investigation. 2003;112(10):1550-60. Epub 2003/11/18. doi: 10.1172/JCI18784. PubMed PMID: 14617756; PubMed Central PMCID: PMC259128.

64. Jackson RS, Creemers JW, Ohagi S, Raffin-Sanson ML, Sanders L, Montague CT, Hutton JC, O'Rahilly S. Obesity and impaired prohormone processing associated with mutations in the human prohormone convertase 1 gene. Nat Genet. 1997;16(3):303-6. Epub 1997/07/01. doi: 10.1038/ng0797-303. PubMed PMID: 9207799.

65. O'Rahilly S, Gray H, Humphreys PJ, Krook A, Polonsky KS, White A, Gibson S, Taylor K, Carr C. Brief report: impaired processing of prohormones associated with abnormalities of glucose homeostasis and adrenal function. N Engl J Med. 1995;333(21):1386-90. Epub 1995/11/23. doi: 10.1056/NEJM199511233332104. PubMed PMID: 7477119.

66. Hoshino AL, Iris. Peptide Biosynthesis: Prohormone Convertases 1/3 and 2. In: Fricker LDD, Lakshmi, editor. Colloquium Series on Neuropeptides. 1 ed: Morgan and Claypool Life Sciences Publishers; 2012.

67. Martin MG, Lindberg I, Solorzano-Vargas RS, Wang J, Avitzur Y, Bandsma R, Sokollik C, Lawrence S, Pickett LA, Chen Z, Egritas O, Dalgic B, Albornoz V, de Ridder L, Hulst J, Gok F, Aydogan A, Al-Hussaini A, Gok DE, Yourshaw M, Wu SV, Cortina G, Stanford S, Georgia S. Congenital proprotein convertase 1/3 deficiency causes malabsorptive diarrhea and other

endocrinopathies in a pediatric cohort. Gastroenterology. 2013;145(1):138-48. doi: 10.1053/j.gastro.2013.03.048. PubMed PMID: 23562752.

68. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308-11. Epub 2000/01/11. PubMed PMID: 11125122; PubMed Central PMCID: PMC29783.

69. Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061-73. Epub 2010/10/29. doi: 10.1038/nature09534. PubMed PMID: 20981092; PubMed Central PMCID: PMC3042601.

70. Exome Variant Server [Internet]. NHLBI Exome Sequencing Project (ESP). 2011 [cited 2011-09-10]. Available from: http://evs.gs.washington.edu/EVS/.

71. NIEHS Exome Variant Server [Internet]. NIEHS Environmental Genome Project. 2011 [cited 2011-09-21]. Available from: http://evs.gs.washington.edu/niehsExome/.

72. Gisser J, Gariepy C. Genetics of Motility Disorder: Gastroesophageal Reflux, Triple A Syndrome, Hirschsprung Disease, and Chronic Intestinal Pseudo-Obstruction. In: Faure C, Di Lorenzo C, Thapar N, editors. Pediatric Neurogastroenterology: Humana Press; 2013. p. 203-16.

73. Berdon WE, Baker DH, Blanc WA, Gay B, Santulli TV, Donovan C. Megacystis-microcolon-intestinal hypoperistalsis syndrome: a new cause of intestinal obstruction in the newborn. Report of radiologic findings in five newborn girls. AJR American journal of roentgenology. 1976;126(5):957-64. Epub 1976/05/01. doi: 10.2214/ajr.126.5.957. PubMed PMID: 178239.

74. Stanghellini V, Cogliandro RF, De Giorgio R, Barbara G, Cremon C, Antonucci A, Fronzoni L, Cogliandro L, Naponelli V, Serra M, Corinaldesi R. Natural history of intestinal failure induced by chronic idiopathic intestinal pseudo-obstruction. Transplantation proceedings. 2010;42(1):15-8. doi: 10.1016/j.transproceed.2009.12.017. PubMed PMID: 20172271.

75. Anuras S, Mitros FA, Nowak TV, Ionasescu VV, Gurll NJ, Christensen J, Green JB. A familial visceral myopathy with external ophthalmoplegia and autosomal recessive transmission. Gastroenterology. 1983;84(2):346-53. Epub 1983/02/01. PubMed PMID: 6687359.

76. Auricchio A, Brancolini V, Casari G, Milla PJ, Smith VV, Devoto M, Ballabio A. The locus for a novel syndromic form of neuronal intestinal pseudoobstruction maps to Xq28. Am J Hum Genet. 1996;58(4):743-8. Epub 1996/04/01. PubMed PMID: 8644737; PubMed Central PMCID: PMC1914695.

77. Faulk DL, Anuras S, Gardner GD, Mitros FA, Summers RW, Christensen J. A familial visceral myopathy. Annals of internal medicine. 1978;89(5 Pt 1):600-6. Epub 1978/11/01. PubMed PMID: 717927.

78. Ionasescu V, Thompson SH, Ionasescu R, Searby C, Anuras S, Christensen J, Mitros F, Hart M, Bosch P. Inherited ophthalmoplegia with intestinal pseudo-obstruction. Journal of the neurological sciences. 1983;59(2):215-28. Epub 1983/05/01. PubMed PMID: 6687898.

79. Mayer EA, Schuffler MD, Rotter JI, Hanna P, Mogard M. Familial visceral neuropathy with autosomal dominant transmission. Gastroenterology. 1986;91(6):1528-35. Epub 1986/12/01. PubMed PMID: 3770377.

80. Patel H, Norman MG, Perry TL, Berry KE. Multiple system atrophy with neuronal intranuclear hyaline inclusions. Report of a case and review of the literature. Journal of the neurological sciences. 1985;67(1):57-65. Epub 1985/01/01. PubMed PMID: 2580060.

81. Roy AD, Bharucha H, Nevin NC, Odling-Smee GW. Idiopathic intestinal pseudo-obstruction: a familial visceral neuropathy. Clin Genet. 1980;18(4):291-7. Epub 1980/10/01. PubMed PMID: 7438508.

82. Schuffler MD, Bird TD, Sumi SM, Cook A. A familial neuronal disease presenting as intestinal pseudoobstruction. Gastroenterology. 1978;75(5):889-98. Epub 1978/11/01. PubMed PMID: 212342.

83. Schuffler MD, Lowe MC, Bill AH. Studies of idiopathic intestinal pseudoobstruction. I. Hereditary hollow visceral myopathy: clinical and pathological studies. Gastroenterology. 1977;73(2):327-38. Epub 1977/08/01. PubMed PMID: 873134.

84. Schuffler MD, Pope CE, 2nd. Studies of idiopathic intestinal pseudoobstruction. II. Hereditary hollow visceral myopathy: family studies. Gastroenterology. 1977;73(2):339-44. Epub 1977/08/01. PubMed PMID: 873135.

85. Van Goethem G, Schwartz M, Lofgren A, Dermaut B, Van Broeckhoven C, Vissing J. Novel POLG mutations in progressive external ophthalmoplegia mimicking mitochondrial neurogastrointestinal encephalomyopathy. Eur J Hum Genet. 2003;11(7):547-9. Epub 2003/06/26. doi: 10.1038/sj.ejhg.5201002. PubMed PMID: 12825077.

86. Amiot A, Tchikviladze M, Joly F, Slama A, Hatem DC, Jardel C, Messing B, Lombes A. Frequency of mitochondrial defects in patients with chronic intestinal pseudo-obstruction. Gastroenterology. 2009;137(1):101-9. Epub 2009/04/07. doi: 10.1053/j.gastro.2009.03.054. PubMed PMID: 19344718.

87. Chinnery PF, Turnbull DM. Clinical features, investigation, and management of patients with defects of mitochondrial DNA. J Neurol Neurosurg Psychiatry. 1997;63(5):559-63. Epub 1998/01/04. PubMed PMID: 9408091; PubMed Central PMCID: PMC2169824.

88. DiMauro S. Mitochondrial diseases. Biochimica et biophysica acta. 2004;1658(1-2):80-8. Epub 2004/07/30. doi: 10.1016/j.bbabio.2004.03.014. PubMed PMID: 15282178.

89. Blondon H, Polivka M, Joly F, Flourie B, Mikol J, Messing B. Digestive smooth muscle mitochondrial myopathy in patients with mitochondrial-neuro-gastro-intestinal encephalomyopathy (MNGIE). Gastroenterologie clinique et biologique. 2005;29(8-9):773-8. Epub 2005/11/19. PubMed PMID: 16294144.

90. Hirano M, Marti R, Spinazzola A, Nishino I, Nishigaki Y. Thymidine phosphorylase deficiency causes MNGIE: an autosomal recessive mitochondrial disorder. Nucleosides, nucleotides & nucleic acids. 2004;23(8-9):1217-25. Epub 2004/12/02. doi: 10.1081/NCN-200027485. PubMed PMID: 15571233.

91. Hirano M, Nishino I, Nishigaki Y, Marti R. Thymidine phosphorylase gene mutations cause mitochondrial neurogastrointestinal encephalomyopathy (MNGIE). Internal medicine. 2006;45(19):1103. Epub 2006/11/02. PubMed PMID: 17077575.

92. Nishino I, Spinazzola A, Hirano M. Thymidine phosphorylase gene mutations in MNGIE, a human mitochondrial disorder. Science. 1999;283(5402):689-92. Epub 1999/01/29. PubMed PMID: 9924029.

93. Slama A, Lacroix C, Plante-Bordeneuve V, Lombes A, Conti M, Reimund JM, Auxenfants E, Crenn P, Laforet P, Joannard A, Seguy D, Pillant H, Joly P, Haut S, Messing B, Said G, Legrand A, Guiochon-Mantel A. Thymidine phosphorylase gene mutations in patients with mitochondrial neurogastrointestinal encephalomyopathy syndrome. Molecular genetics and metabolism. 2005;84(4):326-31. Epub 2005/03/23. doi: 10.1016/j.ymgme.2004.12.004. PubMed PMID: 15781193.

94. Chang TM, Chi CS, Tsai CR, Lee HF, Li MC. Paralytic ileus in MELAS with phenotypic features of MNGIE. Pediatric neurology. 2004;31(5):374-7. Epub 2004/11/03. doi: 10.1016/j.pediatrneurol.2004.05.009. PubMed PMID: 15519124.

95. Chinnery PF, Jones S, Sviland L, Andrews RM, Parsons TJ, Turnbull DM, Bindoff LA. Mitochondrial enteropathy: the primary pathology may not be within the gastrointestinal tract. Gut. 2001;48(1):121-4. Epub 2000/12/15. PubMed PMID: 11115833; PubMed Central PMCID: PMC1728165.

96. Filosto M, Mancuso M, Nishigaki Y, Pancrudo J, Harati Y, Gooch C, Mankodi A, Bayne L, Bonilla E, Shanske S, Hirano M, DiMauro S. Clinical and genetic heterogeneity in progressive external ophthalmoplegia due to mutations in polymerase gamma. Archives of neurology. 2003;60(9):1279-84. Epub 2003/09/17. doi: 10.1001/archneur.60.9.1279. PubMed PMID: 12975295.

97. Garcia-Velasco A, Gomez-Escalonilla C, Guerra-Vales JM, Cabello A, Campos Y, Arenas J. Intestinal pseudo-obstruction and urinary retention: cardinal features of a mitochondrial DNA-related disease. Journal of internal medicine. 2003;253(3):381-5. Epub 2003/02/27. PubMed PMID: 12603507.

98. Li JY, Kong KW, Chang MH, Cheung SC, Lee HC, Pang CY, Wei YH. MELAS syndrome associated with a tandem duplication in the D-loop of mitochondrial DNA. Acta neurologica Scandinavica. 1996;93(6):450-5. Epub 1996/06/01. PubMed PMID: 8836308.

99. Mancuso M, Filosto M, Oh SJ, DiMauro S. A novel polymerase gamma mutation in a family with ophthalmoplegia, neuropathy, and Parkinsonism. Archives of neurology. 2004;61(11):1777-9. Epub 2004/11/10. doi: 10.1001/archneur.61.11.1777. PubMed PMID: 15534189.

100. Tanji K, Gamez J, Cervera C, Mearin F, Ortega A, de la Torre J, Montoya J, Andreu AL, DiMauro S, Bonilla E. The A8344G mutation in mitochondrial DNA associated with stroke-like episodes and gastrointestinal dysfunction. Acta neuropathologica. 2003;105(1):69-75. Epub 2002/12/10. doi: 10.1007/s00401-002-0604-y. PubMed PMID: 12471464.

101. Van Goethem G, Luoma P, Rantamaki M, Al Memar A, Kaakkola S, Hackman P, Krahe R, Lofgren A, Martin JJ, De Jonghe P, Suomalainen A, Udd B, Van Broeckhoven C. POLG mutations in neurodegenerative disorders with ataxia but no muscle involvement. Neurology. 2004;63(7):1251-7. Epub 2004/10/13. PubMed PMID: 15477547.

102. Lehtonen HJ, Sipponen T, Tojkander S, Karikoski R, Jarvinen H, Laing NG, Lappalainen P, Aaltonen LA, Tuupanen S. Segregation of a missense variant in enteric smooth muscle actin gamma-2 with autosomal dominant familial visceral myopathy. Gastroenterology. 2012;143(6):1482-91 e3. doi: 10.1053/j.gastro.2012.08.045. PubMed PMID: 22960657.

103. Wan J, Yourshaw M, Mamsa H, Rudnik-Schoneborn S, Menezes MP, Hong JE, Leong DW, Senderek J, Salman MS, Chitayat D, Seeman P, von Moers A, Graul-Neumann L, Kornberg AJ, Castro-Gago M, Sobrido MJ, Sanefuji M, Shieh PB, Salamon N, Kim RC, Vinters HV, Chen Z, Zerres K, Ryan MM, Nelson SF, Jen JC. Mutations in the RNA exosome component gene EXOSC3 cause pontocerebellar hypoplasia and spinal motor neuron degeneration. Nat Genet. 2012;44(6):704-8. Epub 2012/05/01. doi: 10.1038/ng.2254. PubMed PMID: 22544365; PubMed Central PMCID: PMC3366034.

104. Namavar Y, Barth PG, Poll-The BT, Baas F. Classification, diagnosis and potential mechanisms in pontocerebellar hypoplasia. Orphanet journal of rare diseases. 2011;6:50. Epub 2011/07/14. doi: 10.1186/1750-1172-6-50. PubMed PMID: 21749694; PubMed Central PMCID: PMC3159098.

105. Barth PG. Pontocerebellar hypoplasias. An overview of a group of inherited neurodegenerative disorders with fetal onset. Brain & development. 1993;15(6):411-22. Epub 1993/11/01. PubMed PMID: 8147499.

106. de Leon GA, Grover WD, D'Cruz CA. Amyotrophic cerebellar hypoplasia: a specific form of infantile spinal atrophy. Acta neuropathologica. 1984;63(4):282-6. Epub 1984/01/01. PubMed PMID: 6475488.

107. Goutieres F, Aicardi J, Farkas E. Anterior horn cell disease associated with pontocerebellar hypoplasia in infants. J Neurol Neurosurg Psychiatry. 1977;40(4):370-8. Epub 1977/04/01. PubMed PMID: 874513; PubMed Central PMCID: PMC492704.

108. Norman RM. Cerebellar hypoplasia in Werdnig-Hoffmann disease. Archives of disease in childhood. 1961;36:96-101. Epub 1961/02/01. PubMed PMID: 13729575; PubMed Central PMCID: PMC2012675.

109. Melki J, Lefebvre S, Burglen L, Burlet P, Clermont O, Millasseau P, Reboullet S, Benichou B, Zeviani M, Le Paslier D, et al. De novo and inherited deletions of the 5q13 region in spinal muscular atrophies. Science. 1994;264(5164):1474-7. Epub 1994/06/03. PubMed PMID: 7910982.

110. Gorgen-Pauly U, Sperner J, Reiss I, Gehl HB, Reusche E. Familial pontocerebellar hypoplasia type I with anterior horn cell disease. European journal of paediatric neurology : EJPN : official journal of the European Paediatric Neurology Society. 1999;3(1):33-8. Epub 2000/03/22. doi: 10.1053/ejpn.1999.0177. PubMed PMID: 10727190.

111. Muntoni F, Goodwin F, Sewry C, Cox P, Cowan F, Airaksinen E, Patel S, Ignatius J, Dubowitz V. Clinical spectrum and diagnostic difficulties of infantile ponto-cerebellar hypoplasia type 1. Neuropediatrics. 1999;30(5):243-8. Epub 1999/12/22. doi: 10.1055/s-2007-973498. PubMed PMID: 10598835.

112. Rudnik-Schoneborn S, Sztriha L, Aithala GR, Houge G, Laegreid LM, Seeger J, Huppke M, Wirth B, Zerres K. Extended phenotype of pontocerebellar hypoplasia with infantile spinal muscular atrophy. American journal of medical genetics Part A. 2003;117A(1):10-7. Epub 2003/01/28. doi: 10.1002/ajmg.a.10863. PubMed PMID: 12548734.

113. Ryan MM, Cooke-Yarborough CM, Procopis PG, Ouvrier RA. Anterior horn cell disease and olivopontocerebellar hypoplasia. Pediatric neurology. 2000;23(2):180-4. Epub 2000/10/06. PubMed PMID: 11020648.

114. Chou SM, Gilbert EF, Chun RW, Laxova R, Tuffli GA, Sufit RL, Krassikot N. Infantile olivopontocerebellar atrophy with spinal muscular atrophy (infantile OPCA + SMA). Clinical neuropathology. 1990;9(1):21-32. Epub 1990/01/01. PubMed PMID: 2407400.

115. Gomez-Lado C, Eiris-Punal J, Vazquez-Lopez ME, Castro-Gago M. [Pontocerebellar hypoplasia type I and mitochondrial pathology]. Revista de neurologia. 2007;45(10):639-40. Epub 2007/11/17. PubMed PMID: 18008272.

116. Lev D, Michelson-Kerman M, Vinkler C, Blumkin L, Shalev SA, Lerman-Sagie T. Infantile onset progressive cerebellar atrophy and anterior horn cell degeneration--a late onset variant of PCH-1? European journal of paediatric neurology : EJPN : official journal of the European Paediatric Neurology Society. 2008;12(2):97-101. Epub 2007/08/08. doi: 10.1016/j.ejpn.2007.06.005. PubMed PMID: 17681808.

117. Salman MS, Blaser S, Buncic JR, Westall CA, Heon E, Becker L. Pontocerebellar hypoplasia type 1: new leads for an earlier diagnosis. Journal of child neurology. 2003;18(3):220-5. Epub 2003/05/07. PubMed PMID: 12731647.

118. Sanefuji M, Kira R, Matsumoto K, Gondo K, Torisu H, Kawakami H, Iwaki T, Hara T. Autopsy case of later-onset pontocerebellar hypoplasia type 1: pontine atrophy and pyramidal tract involvement. Journal of child neurology. 2010;25(11):1429-34. Epub 2010/06/19. doi: 10.1177/0883073810372991. PubMed PMID: 20558670.

119. Szabo N, Szabo H, Hortobagyi T, Turi S, Sztriha L. Pontocerebellar hypoplasia type 1. Pediatric neurology. 2008;39(4):286-8. Epub 2008/09/23. doi: 10.1016/j.pediatrneurol.2008.06.017. PubMed PMID: 18805371.

120. Tsao CY, Mendell J, Sahenk Z, Rusin J, Boue D. Hypotonia, weakness, and pontocerebellar hypoplasia in siblings. Seminars in pediatric neurology. 2008;15(4):151-3. Epub 2008/12/17. doi: 10.1016/j.spen.2008.09.001. PubMed PMID: 19073313.

121. Renbaum P, Kellerman E, Jaron R, Geiger D, Segel R, Lee M, King MC, Levy-Lahad E. Spinal muscular atrophy with pontocerebellar hypoplasia is caused by a mutation in the VRK1 gene. Am J Hum Genet. 2009;85(2):281-9. Epub 2009/08/04. doi: 10.1016/j.ajhg.2009.07.006. PubMed PMID: 19646678; PubMed Central PMCID: PMC2725266.

122. Simonati A, Cassandrini D, Bazan D, Santorelli FM. TSEN54 mutation in a child with pontocerebellar hypoplasia type 1. Acta neuropathologica. 2011;121(5):671-3. Epub 2011/04/07. doi: 10.1007/s00401-011-0823-1. PubMed PMID: 21468723.

123. Budde BS, Namavar Y, Barth PG, Poll-The BT, Nurnberg G, Becker C, van Ruissen F, Weterman MA, Fluiter K, te Beek ET, Aronica E, van der Knaap MS, Hohne W, Toliat MR, Crow YJ, Steinling M, Voit T, Roelenso F, Brussel W, Brockmann K, Kyllerman M, Boltshauser E, Hammersen G, Willemsen M, Basel-Vanagaite L, Krageloh-Mann I, de Vries LS, Sztriha L, Muntoni F, Ferrie CD, Battini R, Hennekam RC, Grillo E, Beemer FA, Stoets LM, Wollnik B, Nurnberg P, Baas F. tRNA splicing endonuclease mutations cause pontocerebellar hypoplasia. Nat Genet. 2008;40(9):1113-8. Epub 2008/08/20. doi: 10.1038/ng.204. PubMed PMID: 18711368.

124. Edvardson S, Shaag A, Kolesnikova O, Gomori JM, Tarassov I, Einbinder T, Saada A, Elpeleg O. Deleterious mutation in the mitochondrial arginyl-transfer RNA synthetase gene is associated with pontocerebellar hypoplasia. Am J Hum Genet. 2007;81(4):857-62. Epub 2007/09/12. doi: 10.1086/521227. PubMed PMID: 17847012; PubMed Central PMCID: PMC2227936.

125. Rankin J, Brown R, Dobyns WB, Harington J, Patel J, Quinn M, Brown G. Pontocerebellar hypoplasia type 6: A British case with PEHO-like features. American journal of medical genetics Part A. 2010;152A(8):2079-84. Epub 2010/07/17. doi: 10.1002/ajmg.a.33531. PubMed PMID: 20635367.

126. Mitchell P, Petfalski E, Shevchenko A, Mann M, Tollervey D. The exosome: a conserved eukaryotic RNA processing complex containing multiple 3'-->5' exoribonucleases. Cell. 1997;91(4):457-66. Epub 1997/12/09. PubMed PMID: 9390555.

127. Jensen TH. RNA exosome. Preface. Advances in experimental medicine and biology. 2010;702:v-vi. Epub 2010/01/01. PubMed PMID: 21618869.

128. Chlebowski A, Lubas M, Jensen TH, Dziembowski A. RNA decay machines: the exosome. Biochimica et biophysica acta. 2013;1829(6-7):552-60. Epub 2013/01/29. doi: 10.1016/j.bbagrm.2013.01.006. PubMed PMID: 23352926.

129. Liu Q, Greimann JC, Lima CD. Reconstitution, activities, and structure of the eukaryotic RNA exosome. Cell. 2006;127(6):1223-37. Epub 2006/12/19. doi: 10.1016/j.cell.2006.10.037. PubMed PMID: 17174896.

130. Schwabova J, Brozkova DS, Petrak B, Mojzisova M, Pavlickova K, Haberlova J, Mrazkova L, Hedvicakova P, Hornofova L, Kaluzova M, Fencl F, Krutova M, Zamecnik J, Seeman P. Homozygous EXOSC3 Mutation c.92G-->C, p.G31A is a Founder Mutation Causing Severe Pontocerebellar Hypoplasia Type 1 Among the Czech Roma. Journal of neurogenetics. 2013. Epub 2013/07/26. doi: 10.3109/01677063.2013.814651. PubMed PMID: 23883322.

131. Kani S, Bae YK, Shimizu T, Tanabe K, Satou C, Parsons MJ, Scott E, Higashijima S, Hibi M. Proneural gene-linked neurogenesis in zebrafish cerebellum. Developmental biology. 2010;343(1-2):1-17. Epub 2010/04/15. doi: 10.1016/j.ydbio.2010.03.024. PubMed PMID: 20388506.

132. Rudnik-Schoneborn S, Senderek J, Jen JC, Houge G, Seeman P, Puchmajerova A, Graul-Neumann L, Seidel U, Korinthenberg R, Kirschner J, Seeger J, Ryan MM, Muntoni F, Steinlin M, Sztriha L, Colomer J, Hubner C, Brockmann K, Van Maldergem L, Schiff M, Holzinger A, Barth P, Reardon W, Yourshaw M, Nelson SF, Eggermann T, Zerres K. Pontocerebellar hypoplasia type 1: clinical spectrum and relevance of EXOSC3 mutations. Neurology. 2013;80(5):438-46. doi: 10.1212/WNL.0b013e31827f0f66. PubMed PMID: 23284067; PubMed Central PMCID: PMC3590055.

133. Biancheri R, Cassandrini D, Pinto F, Trovato R, Di Rocco M, Mirabelli-Badenier M, Pedemonte M, Panicucci C, Trucks H, Sander T, Zara F, Rossi A, Striano P, Minetti C, Santorelli FM. EXOSC3 mutations in isolated cerebellar hypoplasia and spinal anterior horn involvement. Journal of neurology. 2013;260(7):1866-70. Epub 2013/04/09. doi: 10.1007/s00415-013-6896-0. PubMed PMID: 23564332.

134. Zanni G, Scotton C, Passarelli C, Fang M, Barresi S, Dallapiccola B, Wu B, Gualandi F, Ferlini A, Bertini E, Wei W. Exome sequencing in a family with intellectual disability, early onset spasticity, and cerebellar atrophy detects a novel mutation in EXOSC3. Neurogenetics. 2013;14(3-4):247-50. Epub 2013/08/27. doi: 10.1007/s10048-013-0371-z. PubMed PMID: 23975261.

135. Pickett LA, Yourshaw M, Albornoz V, Chen Z, Solorzano-Vargas RS, Nelson SF, Martin MG, Lindberg I. Functional consequences of a novel variant of PCSK1. PLoS One. 2013;8(1):e55065. doi: 10.1371/journal.pone.0055065. PubMed PMID: 23383060; PubMed Central PMCID: PMC3557230.

# Chapter Two

## Rich annotation of DNA sequencing variants by leveraging the Ensembl Variant Effect Predictor with plugins

Authors

Michael Yourshaw, BS, JD, Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90025, USA; S. Paige Taylor, BS, Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90025, USA; Aliz R. Rao, MS, Bioinformatics Interdepartmental Program, University of California Los Angeles, Los Angeles, CA 90025, USA; Martín G. Martín, MD, MPP, Department of Pediatrics, Division of Gastroenterology and Nutrition, Mattel Children's Hospital and the David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90095, USA; Stanley F. Nelson, MD, Department of Human Genetics and Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90025, USA

Biographical notes

Michael Yourshaw is a Ph.D. student at the Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles. His research interests include next-generation sequencing and rare Mendelian disorders of the intestine and nervous system.

S. Paige Taylor is a Ph.D. student at the Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles. Her research interests include dissecting the molecular basis of genetic disease, investigating the structure and function of primary cilia, and understanding the regulation of developmental signaling pathways.

Aliz R. Rao is a Ph.D. student at the Bioinformatics Interdepartmental Program at the University of California, Los Angeles. Her research interests include improving variant interpretation and gene prioritization techniques, and complex psychiatric diseases.

Martín G. Martín (MD, MPP) is a Professor in the Department of Pediatrics, Division of Gastroenterology, David Geffen School of Medicine, University of California Los Angeles. His research interests include monogenic forms of Pediatric intestinal failure and intestinal stem cell biology.

Stanley F. Nelson (MD) is a Professor in the Department of Human Genetics and the Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California Los Angeles and co-director of the UCLA Clinical Genomics Center. His research interests include next-generation sequencing and rare Mendelian disorders including Duchenne muscular dystrophy.

Correspondence

Michael Yourshaw, Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90025, USA. Email: myourshaw@ucla.edu; telephone: 310-825-7920; fax: 310-794-5446.

Disclosures

None of the authors have potential financial, professional, or personal conflicts to disclose.

Abstract

High-throughput DNA sequencing has become a mainstay for the discovery of genomic variants that may cause disease or affect phenotype. A next-generation sequencing pipeline typically identifies thousands of variants in each sample. A particular challenge is the annotation of each variant in a way that is useful to downstream consumers of the data, such as clinical sequencing centers or researchers. Such users may require that all data storage and analysis remain on secure local servers to protect patient confidentiality or intellectual property, may have unique and changing needs to draw upon a variety of annotation datasets, and may prefer not to rely on closed source applications beyond their control. Here we describe scalable methods for using the plug-in capability of the Ensembl Variant Effect

Predictor to enrich its basic set of variant annotations with additional data on genes, function, conservation, expression, diseases, pathways, and protein structure, and describe an extensible framework for easily adding additional custom datasets.

Keywords

Ensembl Variant Effect Predictor; plugin; annotation; DNA sequencing; database

Introduction

The recent development of technology to sequence the entire genome of an individual at moderate cost is revolutionizing clinical genetics and greatly accelerating the discovery of new genetic causes of disease (1, 2). Next generation sequencing (NGS) platforms now provide clinical laboratories with the ability to sequence in a single process nearly all of the thousands of genes known to be causal of human Mendelian diseases at a cost comparable to that of sequencing a single disease gene by conventional Sanger sequencing (3). Similarly, researchers can use NGS for an unbiased examination of all genes and regulatory features in order to discover the relationship of unsuspected genes and pathways to diseases or traits with unknown causes (4). In operation, a modern NGS platform typically reads the sequence of over one hundred million short DNA fragments extracted from an individual's blood, saliva, or other tissue. These fragments may have been enriched during library preparation for protein coding regions (the "exome") or for targeted regions, such as those known to be involved in a class of diseases. Mature algorithms have been developed to align these short reads to a reference genome, assign read and mapping quality scores, and genotype loci that vary from the reference. The output of such a sequencing pipeline is a Variant Call Format (VCF) (5, 6) file that succinctly and systematically describes the genomic location, dbSNP ID, reference and alternate alleles, genotype, and other information related to each variant (Figure 1a). Where the entire exome was sequenced, a VCF file typically consists of  over 20,000 individual protein coding variant records.

A basic VCF file does not contain most of the information that will be needed by a physician or researcher, such as the transcript and gene that contain the variant, the effect, if any, on protein encoding (synonymous, missense, nonsense) or structure, the likelihood that the variant is damaging, association with diseases or phenotypes, or tissue expression data. There are several applications that can add such annotations to a VCF file, each with strengths and weaknesses, and these have been reviewed elsewhere (7). One characteristic of most of these tools is that they have little or no flexibility to include customized user-defined annotations. Furthermore, on-line tools, such as SeattleSeq (8), have the advantage of simplicity of use but they may not be appropriate for confidential patient data or proprietary intellectual property.

Here we present an approach for developing a custom annotator that can be run on local servers, is not heavily dependent on an outside single researcher or small group for software development and maintenance, and has a simple, modular mechanism for adding new features. Thus, instead of a stand-alone software package, our goal is to share "how-to" directions for using the plug-in capability of the Ensembl Variant Effect Predictor (VEP) (9) to enrich its basic set of variant annotations with additional data from datasets such as Online Mendelian Inheritance in Man (OMIM), the Human Gene Mutation Database (HGMD pro), the Universal Protein Resource (UniProt), KEGG Pathways, RefSeq, the MitoCarta Inventory of Mammalian Mitochondrial Genes, the Catalogue of Somatic Mutations in Cancer (COSMIC), Mouse Genome Informatics (MGI) and the Human Protein Atlas (HPA).

To satisfy the needs of our laboratory research projects and the initiation of the UCLA Clinical Genomics Center we elected to use the Ensembl database and VEP as the basis of a custom annotator, which we call "VAX" (Variant Annotator Extras). Several factors were decisive in adopting this approach. Ensembl, a joint scientific project between the European Bioinformatics Institute and the Wellcome Trust Sanger Institute, provides access to genomic annotation for numerous species stored on a MySQL database that can be accessed programmatically via a Perl application programming interface (API). The database is supported by a large professional organization, is updated regularly, and can be accessed remotely or by downloading a local copy. The Ensembl database and VEP have a large and active user community, and provide excellent and timely advice and support. The VEP is a mature open source Perl script that can be run locally, connected either to the remote Ensembl database or a local copy thereof, or with some limitations used with a local cache. Without any modifications, VEP produces many useful annotations, including genes affected by the variants, consequence of variants on the protein sequence, minor allele frequencies in the population, and SIFT/PolyPhen scores. VEP plugins are a powerful way to extend, filter, and manipulate the output of the VEP, and form the foundation of our methods for integrating diverse datasets into our VAX annotation pipeline.

With the guidelines we present here, a research laboratory or clinical sequencing center, with access to a modest data processing infrastructure and having easily acquired basic Perl and SQL programming skills, can implement a custom annotation system similar to VAX. The following sections describe 1) installation of the data and programs from Ensembl that are needed to run the Variant Effect Predictor and plugin basics, 2) methods for altering VEP output for downstream entry into a relational database, 3) enriching basic annotations using Ensembl, 4) examples of how to implement several useful annotations from non-Ensembl databases, and 5) additional considerations for variant analysis. Computer code for the modules described herein is in the supplemental file vax_code.tar.gz, an index for which is in vax_code_contents.docx. The code is available under a GNU Public License on an "as-is" basis; users should expect to invest additional effort in adapting this code to their particular systems and needs.

Base installation and plugins

*Ensembl data and VEP.* VAX consists of a locally installed MySQL database system, which hosts the Ensembl database and custom data used by the annotator, local installations of the Ensembl Perl API and VEP, and a library of custom VEP plugins (Figure 1b). We downloaded Ensembl data from the Ensembl FTP site in the form of tab-delimited text files for importing into MySQL (download_ensembl_databases.sh). We installed the Ensembl API and VEP according to instructions on the Ensembl web site [http://www.ensembl.org/info/docs/api/api_installation.html].

*VEP plugin interfaces.* Adding functionality to VEP via plugins is straightforward and should be within the abilities of any researcher who has a working knowledge of the Perl programming language. ProteinSeq (Text box 1) is a simple plugin that, for each variant, adds an annotation of the amino acid sequence of the gene. The 'use base' line tells the plugin to inherit the properties of a base class defined by VEP, allowing the module to interact with VEP via well-defined methods. The 'new' method is called once by VEP to initialize the plugin. One-time code, such as establishing a database connection, would be placed in the 'new' method. The 'version' method returns the version of VEP for which the plugin was designed, and the 'feature_types' method tells VEP only to call the plugin for variants that are within transcripts. The 'get_header_info' method defines the annotation. The 'run' method is where the plugin processes each given variation-allele-feature combination. In the example, $tva gets the transcript variation annotation object from VEP, which contains all information necessary to identify the variant's genomic context. The plugin uses the $tva object to access the related translation object, and then returns the amino acid sequence of the translation to VEP, where the annotation will be output in the Extra column. There are a number of additional examples of plugins available through the Ensembl web site (https://github.com/ensembl-variation/VEP_plugins).

*Database connection.* For convenience and efficiency, we implemented a single plugin, vw.pm, to establish a connection to non-Ensembl MySQL databases, and a non-plugin module, vax.pm, for commonly used functions such as get_unique, which removes duplicates from lists of annotations.

*Cross-references.* External databases may use different gene IDs and number chromosomes differently from Ensembl.  The ensemb_xref2db.pl script builds a table of gene, transcript, and protein cross references, and the faidx_decoy.txt table is useful to interconvert human genome build GRCh37 (1000 Genomes decoy version) and UCSC build hg19 chromosomes.

*Database-friendly output.* By default, VEP places many of its annotations as key-value pairs in the Extra column, for example, the amino acid sequence from the ProteinSeq plugin example and a SIFT score of likelihood of protein damage might be represented as 'ProteinSequence= MEAEESE…SLVRDS;SIFT=tolerated(0.34)'. For use in downstream analysis it may be more convenient to separate data into one column for each annotation, and even to create two columns for an annotation like SIFT: one for the verbal description and one for the numerical score. This approach works well for Excel spreadsheets and is almost essential for relational databases.

*ExtraCols plugin.* VEP outputs many annotations, such as HGVS coding and protein sequence names and the SIFT, PolyPhen, and Condel values, as key-value pairs in the Extra column. The ExtraCols plugin (ExtraCols.pm) adds selected additional columns to the output file for each key, making the values more easily accessible to database queries. The command "use Bio::EnsEMBL::Variation::Utils::VEP qw(@OUTPUT_COLS);" gives the plugin access to VEP's list of columns that will appear in the output, and the get_header_info method demonstrates the technique for adding additional column headers. In the 'new' method, the plugin places data for each Extra annotation into its own column, and also separates text from numbers for the SIFT, PolyPhen, and Condel scores.

*VCFCols plugin.* VEP's output format does not preserve all of the columns originally present in the input VCF file, and represents insertions and deletions in a way that is not directly comparable to the VCF standard use of POS, REF and ALT. Some downstream applications require data in the original VCF form. The VCFCols plugin (VCFCols.pm) modifies the VEP output format to include all input VCF columns. The 'new' method scans the input VCF file to identify the columns and stores their names in $self->{_vcf_cols} for future use by the 'run' method. The get_header_info method adds output columns and the 'run' method places data from the original input data line into these columns.

Additional Ensembl annotations

The real power of plugins starts with the ability to add additional annotations from Ensembl's own rich data collection, as in the ProteinSeq plugin example, above. Two plugins, adapted from the NGS-SNP collection of command-line scripts (10), and a plugin to get gene and variant phenotype data, illustrate this.

*Protein plugin.* The Protein plugin (Protein.pm), derived from NGS-SNP, adds several useful annotations. Protein_Length is helpful for analysis when considering where in the protein a variant falls and the likelihood of protein mutation. Protein_Length_Decrease(%), Protein_Sequence_Lost, Protein_Length_Increase(%), and Protein_Sequence_Gained lend perspective to stop_gained and stop_lost variants. Reference_Splice_Site, and Variant_Splice_Site clarify the effect of mutations in the essential splice site region. The plugin computes each of these from the consequence annotation and the amino acid sequence. The Overlapping_Protein_Domains annotation presents all the domain features annotated in Ensembl's translation object that overlap the variant locus.

*Alignment plugin.* NGS-SNP's annotations for detailed comparisons with orthologous sequences is, to our knowledge, unique among existing variant annotation tools. The Alignment plugin (Alignment.pm) adapts portions of NGS-SNP to function as a VEP plugin, and illustrates the use of the Ensembl comparative genomics (Compara) database, which is not implemented in the basic VEP. The 'new' method accepts additional parameters used by the plugin and establishes connections to the compara database. This plugin calculates three values that are useful for evaluating the conservation of amino acid residues. Alignment_Score_Change is the alignment score for the variant amino acid vs. the orthologous amino acids minus the alignment score for the reference amino acid vs. the orthologous amino acids. C_blosum is a measure of the conservation of the reference amino acid with the aligned amino acids in orthologous sequences using the C_blosum formula given in (11). Context_Conservation is the average percent identity obtained when the region of the reference protein containing the SNP-affected residue is aligned with the orthologous region from other species. Additionally, this plugin generates an alignment of amino acids in orthologous species, ordered by evolutionary distance from humans as calculated from the phylogenetic tree obtained from Ensembl(12). This data is displayed in two compact columns: Amino_Acids_In_Orthologues lists the amino acids and Orthologue_Species lists the species from which sequences were obtained to generate the alignment as well as the three numerical measures.

*Phenotypes plugin.* The Phenotypes plugin (Phenotypes.pm) creates columns for phenotypes associated with a gene or variant locus, cancer associations from COSMIC, and the public HGMD dataset, sourced from Ensembl (a plugin for the commercial HGMD Pro dataset is discussed below).

External Databases

It is possible to access some remote databases directly from within a VEP plugin, but in our experience this presents two difficulties: first, throughput may be slow, and second, annotating multiple whole exome variant calls may place an undue burden on remote servers. Consequently, we elected to create local copies of external datasets, and accepted the burden of performing regular (usually quarterly) updates. Updating can be somewhat automated, limiting the human workload to a few hours per year. To access external data we use the MySQL ISAM engine, also used by Ensembl, for reasons of user familiarity, speed, and cost, but any other database system with a Perl interface can easily interface with VEP plugins using code similar to that in the vw plugin. Using external databases within Ensembl plugins generally requires 1) obtaining data from an external source, 2) preprocessing the data, 3) loading MySQL table(s), and 4) developing a plugin to access the database and produce output.

A simple example of this process is the Mito plugin (Text box 2), which produces a column named MT that contains ‘1’ if the gene is annotated by MitoCarta (13) as being found in mitochondria, otherwise blank. The first step is to download the Human.MitoCarta.xls file from http://www.broadinstitute.org/pubs/MitoCarta/human.mitocarta.html. The second, preprocessing, step is to select from the SYM (gene symbol) column only those genes with a ‘1’ in the MITOCARTA_LIST column (indicating strong support of mitochondrial localization) and remove any duplicates. Step 3 involves creating a MySQL table named ‘mitocarta_gene’ with a single column, ‘mito_gene’, which contains the selected mitochondrial gene symbols. For robust and secure access we also create a stored procedure named ‘get_mitocarta_gene’ that takes a gene symbol as input and returns a list of one or zero matching symbols. Finally, the plugin creates an MT column with its get_header_info method and populates the column for each variant in its ‘run’ method by querying the database, using the database connection established by the generic vw plugin. Note the use of the $line_hash parameter passed from VEP to store data in the output line.

*GeneIDs plugin.* The GeneIDs plugin (GeneIDs.pm) creates columns for the chromosomal strand containing the transcribed gene, The Ensembl permanent gene identifier

(ENSG), a gene description, a RefSeq gene summary, the Entrez gene name, the UniProt KB_AC and ID, Gene Ontology references, and mitochondrial location. With two exceptions, these columns are populated from information in Ensembl. The RefSeq summary column contains brief summary paragraphs for more than 5000 well-understood genes, and is an excellent starting point for an analyst when first confronted with an unfamiliar gene (14). This plugin is implemented by downloading RefSeqGene and refseqgene*.genomic.gbff.gz from the NCBI ftp site (download_refseq_data.sh), converting to a database friendly text file (refgene2db.py), and adding cross-references to Ensembl gene names (refseq_ensg_cross-references.sql). The MT column is created as described above for the Mito plugin.

*OMIM plugin.* The OMIM plugin annotates gene associations with Mendelian disorders from OMIM. (15) To build the dataset locally, download the OMIM data files (download_omim_data.sh), convert to database tables (omim2db.py), and import genemap.txt into MySQL.

*DiseasesPhenotypes plugin.* The DiseasesPhenotypes plugin creates a convenience column intended to contain the union of all annotations of disease and phenotype associations from the HGMD, OMIM, Phenotypes, and UniProt plugins. This plugin provides an easy way for a relational database to scan all of these annotations with one query. It is an example of how one plugin can create a column that will be populated by other plugins.

*HGMD plugin.* The HGMD plugin obtains gene and locus disease associations from the commercial professional version of The Human Gene Mutation Database (HGMD®) (BIOBASE Biological Databases). This plugin requires a local installation of the database as documented in its distribution. Stored procedures for access to the data by a plugin are included in plugin_stored_procs.sql.

*HPA plugin.* The HPA plugin creates annotations of gene expression by tissue, cell type, and subcellular location from The Human Protein Atlas. (16) To install the HPA plugin, download the normal_tissue and subcellular_location tables (download_hpa_data.sh) and import them into MySQL.

*KEGG plugin.* The KEGG plugin annotates gene participation in molecular pathways and interaction networks from the KEGG Pathway database. (17) Due to usage restrictions on the database, the gene_pathways table must be created with a Perl script from the KEGG SOAP server (kegg_pathways.pl). After importing the table into MySQL, create a new table that is indexed by Ensembl gene IDs for use by the plugin (gene_pathways2ensg_kegg.sql).

*MousePhenotypes plugin.* Especially when a gene has no known associated phenotypes in humans, it is important to consider whether there is a phenotype in a model organism. The Mouse Genome Informatics (MGI) database(18) has extensive annotations of phenotypes observed in mice when genes orthologous to human genes are mutated or knocked out. The MousePhenotypes plugin annotates variants with all known mouse phenotypes in equivalent human genes. The plugin requires downloading the  HMD_HumanPhenotype.rpt and VOC_MammalianPhenotype.rpt files (download_mgi_data.sh), cleaning up the format (MGI_mouse_phenotype_files.py), and loading the two tables into MySQL. Many other model organisms have similar human gene/phenotype datasets for which VEP plugins could be developed by similar methods.

*UniProt plugin.* The UniProt plugin creates columns for many of the extensive protein annotations in the UniProt database (19). These include VARIANT, MUTAGEN, SITES, OTHER_OVERLAPPING_FEATURES, ALLERGEN, ALTERNATIVE_PRODUCTS, CATALYTIC_ACTIVITY, CAUTION, COFACTOR, DE, DEVELOPMENTAL_STAGE, DISEASE, DOMAIN, ENZYME_REGULATION, FUNCTION, GeneNames, GO, GO_term, INDUCTION, INTERACTION, KEGG, KEYWORDS, MIM_gene, MIM_phenotype, MISCELLANEOUS, PATHWAY, Pathway_Interaction, PE, POLYMORPHISM, PTM, Reactome, RecName, RefSeq_NM, RefSeq_NP, RNA_EDITING, SEQUENCE_CAUTION, SIMILARITY, SUBCELLULAR_LOCATION, SUBUNIT, TISSUE_SPECIFICITY, UCSC, and WEB_RESOURCE. When available, the plugin will annotate specific protein features that overlap the variant. Install this plugin by downloading the UniProt data in its unique format (download_uniprot_data.sh), convert to database tables (uniprot2db.pl), load into MySQL, and create Ensembl transcript ID indexed tables (uniprot2enst_uniprot.sql).

### Additional considerations

VEP plugin-based annotations can work well for producing output that will be reviewed directly or in an Excel spreadsheet. Even greater analytical power is available if annotated variants are used as input to downstream applications, possibly including relational databases. The ExtraCols and VCFCols plugins enable output formatting that is more conducive to relational database analysis by ensuring that each column of contains a single discrete unit of data. Two other issues with the way the VCF input format provides for genotypes of multiple samples can complicate use of the data in some downstream applications. First, the VCF format requires that each genotyped sample's genotype and related data be stored in one column per sample. Second, because samples may have different alternate alleles at a given locus, the ALT column must contain a list of all observed alleles. Therefore, we preprocess VCF files before running the VEP to create two files: 1) a VCF file without sample genotype columns and with each alternate allele on a separate line (this file serves as input to VEP); 2) a file listing each sample's genotype on a separate line. After annotation by VEP we relate sample genotypes and annotations in a database system. During preprocessing we also split the input into a number of smaller files in order to run VEP in parallel for faster throughput. Although none of these steps are essential to operation of a successful annotation pipeline based on VEP plugins, we include our preprocessing program as an example for those who may be interested (vcf2vax.py).

## Conclusion

A local installation of the Ensembl databases and Perl API provides a robust and flexible framework for annotating DNA sequencing variants from many different data sources using Variant Effect Predictor plugin modules. We have outlined the design and usage of VEP plugins for a number of widely used databases. In addition, modules may be easily designed for incorporating annotations from any external dataset that is kept in a flat file or relational database, such as the Zebrafish Model Organism database (20), and the Rat Genome database (21). We have used the VAX system for the discovery of the causes of rare Mendelian diseases and genes involved in psychiatric disorders. (4, 22, 23). VAX is used routinely for CLIA/CAP-accredited whole exome sequencing by the UCLA Clinical Genomics Center, which has processed more than 1000 exomes to date (24). An example of VAX output, demo.vax, is in the vax_code.tar.gz file.

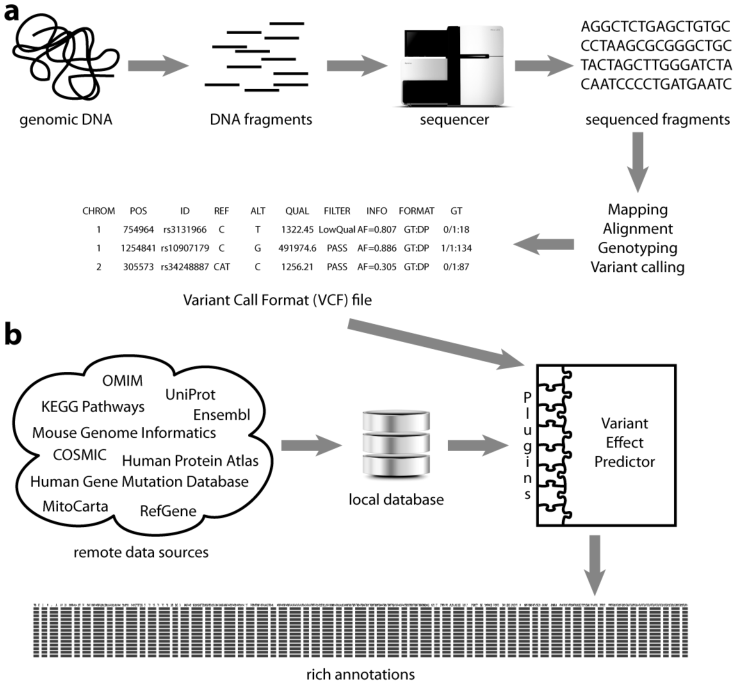
## Key Points

- Richly annotated variants produced by next-generation sequencing are the foundation of modern clinical sequencing and gene discovery research.
- Ensembl Variant Effect Predictor (VEP) plugins provide a robust and flexible framework for annotating DNA sequencing variants.
- VEP plugins are Perl scripts that can use the extensive data in Ensembl, such as comparative genomics and variant annotations.
- Custom VEP plugins can associate variants with data from diverse external sources.
- An annotation pipeline incorporating VEP plugins is within the reach of small laboratories and clinical sequencing centers

## Funding

This work was supported by a grant from the California Institute of Regenerative Medicine (CIRM), RT2-01985 to MGM.

## Figures



| CHROM | POS | ID | REF | ALT | QUAL | FILTER | INFO | FORMAT | GT |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 754964 | rs3131966 | C | T | 1322.45 | LowQual | AF=0.807 | GT:DP | 0/1:18 |
| 1 | 1254841 | rs10907179 | C | G | 491974.6 | PASS | AF=0.886 | GT:DP | 1/1:134 |
| 2 | 305573 | rs34248887 | CAT | C | 1256.21 | PASS | AF=0.305 | GT:DP | 0/1:87 |

Figure 1. Overview of DNA sequencing and annotation. (a) DNA sequencing pipeline. Fragmented genomic DNA is sequenced by a next-generation sequencer and aligned to a reference genome. Each locus is genotyped and variants from the reference are output to a Variant Call Format (VCF) file. (b) Rich variant annotation. Multiple datasets are stored on a local database server. Modular plugins integrated with the Ensembl Variant Effect Predictor (VEP) create an output file with rich annotations of each variant.

```
package ProteinSeq;

use base qw(Bio::EnsEMBL::Variation::Utils::BaseVepPlugin);

sub new {
    my $class = shift;
    my $self = $class->SUPER::new(@_);
    return $self;
}

sub version { return '73'; }

sub feature_types { return ['Transcript']; }

sub get_header_info {
    return { ProteinSeq => "amino acid sequence of transcript's translated protein", };
}

sub run {
    my ($self, $tva) = @_;
    if ( defined $tva->transcript->translation ){
           return { ProteinSeq => $tva->transcript->translation->seq() };
    }
    return {};
}

1;
```

Text box 1. ProteinSeq plugin. This plugin illustrates the methods that a VEP plugin should implement (new, version, feature_types, get_header_info, and run) and demonstrates a simple annotation of the complete amino acid sequence of the protein affected by a variant.

```
package Mito;

use base qw(Bio::EnsEMBL::Variation::Utils::BaseVepPlugin);
use Bio::EnsEMBL::Variation::Utils::VEP qw(@OUTPUT_COLS);
use vw;

sub new {
    my $class = shift;
    my $self = $class->SUPER::new(@_);
    return $self;
}

sub version { return '73'; }

sub feature_types { return ['Transcript']; }

sub get_header_info {
    my @new_output_cols = qw( MT );
    @OUTPUT_COLS = (@OUTPUT_COLS, @new_output_cols);
    return { MT => "annotated as in mitochondrion by MitoCarta", };
}

sub run {
    my ($self, $tva, $line_hash) = @_;
    my $config = $self->{config};
       my $hgnc = $tva->transcript->{_gene_hgnc};
    if (defined $hgnc){
        my $query = "CALL $vw::vw_database.get_mitocarta_gene('$hgnc')";
        my $qh = $vw::vw_conn->prepare($query);
        $qh->execute() or die "Unable to execute $query: $DBI::errstr\n";
        my @row = $qh->fetchrow_array();
        if( defined($row[0]) && $row[0] ne '' ) {
            $line_hash->{MT} = '1';
        }
        else {
            $line_hash->{MT} = '';
        }
    }
    return {};
}

1;

# SQL stored procedure
#     CREATE DEFINER=`sa`@`%` PROCEDURE `get_mitocarta_gene`(hgnc varchar(15))
#     BEGIN
#     SELECT `mitocarta_gene`.`mito_gene`
#     FROM `vw`.`mitocarta_gene`
#     WHERE `mitocarta_gene`.`mito_gene` = hgnc;
#     END
```

Text box 2. Mito plugin. This plugin illustrates the use of data from an external database (the MitoCarta Inventory of Mammalian Mitochondrial Genes) that is stored on a local MySQL server. Consult the vw.pm file in the supplemental vax_code for details of the database connection.

Supplementary Materials

Contents of vax_code.txt

| File | Description |
|---|---|
| Alignment.pm | Alignment plugin Perl module |
| DiseasesPhenotypes.pm | DiseasesPhenotypes plugin Perl module |
| demo.vax | Sample vax annotated output |
| download_ensembl_databases.sh | Shell script to download Ensembl databases from FTP site |
| download_hpa_data.sh | Shell script to download tables from Human Protein Atlas |
| download_hpa_data.sh | Shell script to download mouse phenotype data from MGI |
| download_omim_data.sh | Shell script to download OMIM data |
| download_refseq_data.sh | Shell script to download RefSeq data |
| download_uniprot_data.sh | Shell script to download UniProt data |
| ensembl_database_install_server.sh | Install downloaded Ensembl databases on MySQL server |
| ensembl_xref2db.pl | Perl script to create Ensembl cross-reference table for genes, transcripts, and proteins |
| ExtraCols.pm | ExtraCols plugin Perl module |
| faidx_decoy.txt | Database table to interconvert human genome build GRCh37 (1000 Genomes decoy version) and UCSC build hg19 chromosomes |
| gene_pathways2ensg_kegg.sql | T-SQL code to index KEGG pathways by Ensembl gene IDs |
| GeneIDs.pm | GeneIDs plugin Perl module |
| genemap2ensg_omin.sql | T-SQL code to index OMIM gememap by Ensembl gene IDs |
| HGMD.pm | HGMD plugin Perl module |
| HPA.pm | HPA plugin Perl module |
| kegg_pathways.pl | Perl script to download KEGG pathway data from SOAP server |
| KEGG.pm | KEGG plugin Perl module |
| LICENSE.txt | GNU General Public License |
| MGI_mouse_phenotype_files.py | Python script to convert downloaded Mouse Genome Informatics phenotype data to database tables |
| Mito_example.pm | Mito example plugin Perl module |
| MousePhenotypes.pm | Mouse Phenotypes plugin Perl module |
| my.py | Common python functions used by vcf2vax.py |
| OMIM.pm | OMIM plugin Perl module |
| omim2db.py | Python script to convert downloaded OMIM data to database tables |
| Phenotypes.pm | Phenotypes plugin Perl module |

| File | Description |
|---|---|
| plugin_stored_procs.sql | MySQL stored procedures to interface plugins with database |
| Protein.pm | Protein plugin Perl module |
| ProteinSeq_example.pm | ProteinSeq example plugin Perl module |
| refgene2db.py | Python script to convert downloaded RefGene data to RefSeq gene summary database table |
| refseq2refseq_gene_summary.sql | T-SQL code to index RefSeq gene summary table by Ensembl gene IDs |
| UniProt.pm | UniProt plugin Perl module |
| uniprot2db.pl | Perl script to download UniProt data and convert to database tables |
| uniprot2enst_uniprot.sql | T-SQL code to index UniProt tables by Ensembl transcript IDs |
| vax_code_contents.docx | This document |
| vax.pm | vax common functions Perl module |
| vcf2vax.py | Python script to pre-process VCF files and run the Variant Effect Predictor in parallel on a compute cluster |
| VCFCols.pm | VCFCols plugin Perl module |
| vw.pm | MySql connector plugin Perl module |

These items are contained in the supplementary file vax_code.txt.

# References

1. Gonzaga-Jauregui C, Lupski JR, Gibbs RA. Human genome sequencing in health and disease. Annual review of medicine. 2012;63:35-61. doi: 10.1146/annurev-med-051010-162644. PubMed PMID: 22248320; PubMed Central PMCID: PMC3656720.

2. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008;26(10):1135-45. doi: 10.1038/nbt1486. PubMed PMID: 18846087.

3. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 2011;12(11):745-55. Epub 2011/09/29. doi: 10.1038/nrg3031. PubMed PMID: 21946919.

4. Wan J, Yourshaw M, Mamsa H, Rudnik-Schoneborn S, Menezes MP, Hong JE, Leong DW, Senderek J, Salman MS, Chitayat D, Seeman P, von Moers A, Graul-Neumann L, Kornberg AJ, Castro-Gago M, Sobrido MJ, Sanefuji M, Shieh PB, Salamon N, Kim RC, Vinters HV, Chen Z, Zerres K, Ryan MM, Nelson SF, Jen JC. Mutations in the RNA exosome component gene EXOSC3 cause pontocerebellar hypoplasia and spinal motor neuron degeneration. Nat Genet. 2012;44(6):704-8. Epub 2012/05/01. doi: 10.1038/ng.2254. PubMed PMID: 22544365; PubMed Central PMCID: PMC3366034.

5. 1000 Genomes Project. VCF (Variant Call Format) version 4.1 2013 [updated 2013-10-09]. Available from: http://www.1000genomes.org/wiki/Analysis/Variant Call Format/vcf-variant-call-format-version-41.

6. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156-8. Epub 2011/06/10. doi: 10.1093/bioinformatics/btr330. PubMed PMID: 21653522; PubMed Central PMCID: PMC3137218.

7. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J, Trajanoski Z. A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform. 2013. doi: 10.1093/bib/bbs086. PubMed PMID: 23341494.

8. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009;461(7261):272-6. Epub 2009/08/18. doi: nature08250 [pii]
10.1038/nature08250. PubMed PMID: 19684571.

9. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010;26(16):2069-70. Epub 2010/06/22. doi: btq330 [pii]
10.1093/bioinformatics/btq330. PubMed PMID: 20562413; PubMed Central PMCID: PMC2916720.

10. Grant JR, Arantes AS, Liao X, Stothard P. In-depth annotation of SNPs arising from resequencing projects using NGS-SNP. Bioinformatics. 2011;27(16):2300-1. Epub 2011/06/24.

doi: 10.1093/bioinformatics/btr372. PubMed PMID: 21697123; PubMed Central PMCID: PMC3150039.

11. Kowarsch A, Fuchs A, Frishman D, Pagel P. Correlated mutations: a hallmark of phenotypic amino acid substitutions. PLoS computational biology. 2010;6(9). doi: 10.1371/journal.pcbi.1000923. PubMed PMID: 20862353; PubMed Central PMCID: PMC2940720.

12. Ensembl. Ensembl/UCSC phylogenetic tree [2013-07-26]. Available from: http://tinyurl.com/ensembltree.

13. Pagliarini DJ, Calvo SE, Chang B, Sheth SA, Vafai SB, Ong SE, Walford GA, Sugiana C, Boneh A, Chen WK, Hill DE, Vidal M, Evans JG, Thorburn DR, Carr SA, Mootha VK. A mitochondrial protein compendium elucidates complex I disease biology. Cell. 2008;134(1):112-23. doi: 10.1016/j.cell.2008.06.016. PubMed PMID: 18614015; PubMed Central PMCID: PMC2778844.

14. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. 2012;40(Database issue):D130-5. doi: 10.1093/nar/gkr1079. PubMed PMID: 22121212; PubMed Central PMCID: PMC3245008.

15. Online Mendelian Inheritance in Man OMIM®. Online Mendelian Inheritance in Man, OMIM® Baltimore, MD: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University; [2013-11-04]. Available from: http://omim.org/.

16. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S, Wernerus H, Bjorling L, Ponten F. Towards a knowledge-based Human Protein Atlas. Nat Biotechnol. 2010;28(12):1248-50. doi: 10.1038/nbt1210-1248. PubMed PMID: 21139605.

17. Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. Nucleic Acids Res. 2002;30(1):42-6. PubMed PMID: 11752249; PubMed Central PMCID: PMC99091.

18. Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE, Mouse Genome Database G. The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. Nucleic Acids Res. 2012;40(Database issue):D881-6. doi: 10.1093/nar/gkr974. PubMed PMID: 22075990; PubMed Central PMCID: PMC3245042.

19. Apweiler R, Martin MJ, O'Donovan C, Magrane M, Alam-Faruque Y, Alpi E, Antunes R, Arganiska J, Casanova EB, Bely B, Bingley M, Bonilla C, Britto R, Bursteinas B, Chan WM, Chavali G, Cibrian-Uhalte E, Da Silva A, De Giorgi M, Dimmer E, Fazzini F, Gane P, Fedotov A, Castro LG, Garmiri P, Hatton-Ellis E, Hieta R, Huntley R, Jacobsen J, Jones R, Legge D, Liu WD, Luo J, MacDougall A, Mutowo P, Nightingale A, Orchard S, Patient S, Pichler K, Poggioli D, Pundir S, Pureza L, Qi GY, Rosanoff S, Sawford T, Sehra H, Turner E, Volynkin V, Wardell T, Watkins X, Zellner H, Corbett M, Donnelly M, van Rensburg P, Goujon M, McWilliam H, Lopez R, Xenarios I, Bougueleret L, Bridge A, Poux S, Redaschi N, Auchincloss A, Axelsen K, Bansal P, Baratin D, Binz PA, Blatter MC, Boeckmann B, Bolleman J, Boutet E, Breuza L, de Castro E, Cerutti L, Coudert E, Cuche B, Doche M, Dornevil D, Duvaud S, Estreicher A, Famiglietti L, Feuermann M, Gasteiger E,

Gehant S, Gerritsen V, Gos A, Gruaz-Gumowski N, Hinz U, Hulo C, James J, Jungo F, Keller G, Lara V, Lemercier P, Lew J, Lieberherr D, Martin X, Masson P, Morgat A, Neto T, Paesano S, Pedruzzi I, Pilbout S, Pozzato M, Pruess M, Rivoire C, Roechert B, Schneider M, Sigrist C, Sonesson K, Staehli S, Stutz A, Sundaram S, Tognolli M, Verbregue L, Veuthey AL, Zerara M, Wu CH, Arighi CN, Arminski L, Chen CM, Chen YX, Huang HZ, Kukreja A, Laiho K, McGarvey P, Natale DA, Natarajan TG, Roberts NV, Suzek BE, Vinayaka CR, Wang QH, Wang YQ, Yeh LS, Yerramalla MS, Zhang J, Consortium U. Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Research. 2013;41(D1):D43-D7. doi: Doi 10.1093/Nar/Gks1068. PubMed PMID: WOS:000312893300007.

20. Howe DG, Bradford YM, Conlin T, Eagle AE, Fashena D, Frazer K, Knight J, Mani P, Martin R, Moxon SA, Paddock H, Pich C, Ramachandran S, Ruef BJ, Ruzicka L, Schaper K, Shao X, Singer A, Sprunger B, Van Slyke CE, Westerfield M. ZFIN, the Zebrafish Model Organism Database: increased support for mutants and transgenics. Nucleic Acids Res. 2013;41(Database issue):D854-60. doi: 10.1093/nar/gks938. PubMed PMID: 23074187; PubMed Central PMCID: PMC3531097.

21. Laulederkind SJ, Hayman GT, Wang SJ, Smith JR, Lowry TF, Nigam R, Petri V, de Pons J, Dwinell MR, Shimoyama M, Munzenmaier DH, Worthey EA, Jacob HJ. The Rat Genome Database 2013--data, tools and users. Brief Bioinform. 2013;14(4):520-6. doi: 10.1093/bib/bbt007. PubMed PMID: 23434633; PubMed Central PMCID: PMC3713714.

22. Kerner B, Rao AR, Christensen B, Dandekar S, Yourshaw M, Nelson SF. Rare genomic variants link bipolar disorder to CREB regulated intracellular signaling pathways. Frontiers in Psychiatry. 2013. doi: 10.3389/fpsyt.2013.00154.

23. Yourshaw M, Solorzano-Vargas RS, Pickett LA, Lindberg I, Wang J, Cortina G, Pawlikowska-Haddal A, Baron H, Venick RS, Nelson SF. Exome Sequencing Finds a Novel PCSK1 Mutation in a Child With Generalized Malabsorptive Diarrhea and Diabetes Insipidus. Journal of pediatric gastroenterology and nutrition. 2013. doi: 10.1097/MPG.0b013e3182a8ae6c.

24. Lee H, Nelson SF. Rethinking clinical practice: clinical implementation of exome sequencing. Pers Med. 2012;9(8):785-7. doi: Doi 10.2217/Pme.12.101. PubMed PMID: WOS:000311977800001.

CHAPTER THREE

# Whole-exome sequencing for the identification of casual mutations in congenital diarrheal disorders

## Abstract

*Background.* Congenital diarrheal disorders (CDDs) are rare diseases with serious, even life-threatening, consequences. Whole-exome sequencing enables an unbiased search for genetic variants that cause the disease.

*Methods.* We developed a bioinformatic pipeline and analytical methods for whole-exome sequencing to discover variants in the protein coding region of genes associated with CDDs.

*Results.* We sequenced 45 probands from diverse ethnic backgrounds who were diagnosed with a variety of CDDs of probable, but unknown genetic cause, and searched for damaging mutations in genes known to be associated with CDDs. Patients had been diagnosed with generalized malabsorptive diarrhea (*n*=33), selective nutrient malabsorption (*n*=5), secretory diarrhea (*n*=3), and infantile IBD (*n*=4). Although most cases had thorough diagnostic workups, sometimes including sequencing of suspected genes, none had a convincing genetic finding. Surprisingly, we found a mutation in a known gene, or a gene that was reported during the course of the study, that is likely to cause disease in 27 cases (60%). The genes implicated were *ADAM17*, *DGAT1*, *EPCAM*, *IL10RA*, *MALT1*, *MYO5B*, *NEUROG3*, *PCSK1*, *SI*, *SKIV2L*, *SLC26A3*, and *SLC5A1*. While six of the mutations were previously reported, 25 of them were novel.

*Conclusions.* Whole-exome sequencing is an effective approach for the identification of casual mutations in known CDD genes that may escape detection with standard practice involving a complex diagnostic workup and targeted gene sequencing.

## Introduction

Congenital diarrheal disorders (CDDs) are rare diseases with serious, even life-threatening, consequences that impose massive diagnostic and treatment costs as well as great emotional stress on patients and their families. These patients frequently endure a complex and costly diagnostic odyssey that often fails to produce a definitive diagnosis (1). Affected families may be unaware of appropriate treatments and prognosis (2) and lack information on the recurrence risk in subsequent pregnancies. Until recently, little was known of the genetic etiology of these diseases, yet identification of a casual mutation can lead to improved management of the disease and inform research efforts to develop new treatment modalities. Importantly, identifying new genes provides a better understanding of the molecular pathways and mechanisms responsible for the specific disease and other phenotypes governed by the same pathways. For example, mutations in the gene *PCSK1* (encoding neuroendocrine convertase 1 precursor) can cause a severe form of generalized malabsorptive diarrhea, while milder variants are associated with obesity (3).

Although rare, these diseases impose serious emotional and economic burdens on families and society. Children with CDDs often require long-term parenteral nutrition in order to absorb sufficient nutrients to sustain normal growth and development. Parenteral nutrition, however, does not address the primary cause of the disorder and its prolonged use is associated with numerous complications, including loss of central venous access, sepsis, intestinal failure associated liver disease, poor health related quality of life, multivisceral transplantation, and death. Economically, the expenditure of home parenteral nutrition alone is estimated in the range of $100,000-$150,000 per year, and the cost of transplantation is ~$1.5-1.9 million for the first year alone (4).

CDDs are a group of enteropathies caused by inherited or sporadic genetic mutations that generally manifest soon after birth or in early childhood. The primary presenting symptom is chronic diarrhea that often requires total parenteral nutrition. Mutations in number of genes are known to cause CDDs, but many cases, even those that harbor a mutation in a known gene, do not receive a genetic diagnosis.

The discovery of most of the known genes involved arduous genetic and molecular studies. In the 1990s genome-wide linkage analysis, using 300-500 microsatellite markers, offered a method of identifying disease susceptibility loci. This approach was time-consuming and limited to a resolution of down to about 10Mb, a length, which can harbor several genes (5). To achieve adequate power it was necessary to genotype one or more extended families or a very large number of nuclear families. For example, linkage disequilibrium as well as genetic linkage, as determined by this technique in Finnish families, indicated that an unknown gene near the cystic fibrosis transmembrane regulator gene (CFTR) was probably associated with secretory chloride diarrhea (6). Subsequently, cloning of the linkage region identified four known genes, two of which were considered to be functionally relevant (7). Finally, segregation of mutations in the SLC26A3 gene with the disorder in a large number of patients confirmed that such mutations cause the disorder (8). The development of highly parallel genotyping based on arrays with probes for thousands of single nucleotide polymorphisms (SNPs) enabled more efficient genotyping with a 10-fold or better improvement in resolution (9) This technology, applied to extended kindreds, enabled the discovery of mutations of the causative genes for microvillus inclusion disease (10), syndromic congenital secretory sodium diarrhea (11), congenital tufting enteropathy (12), and early-onset chronic diarrhea (13). A candidate gene resequencing approach, founded on a mouse model, identified mutations in NEUROG3 as the cause of malabsorptive congenital diarrhea (14).

The paradigm of earlier techniques for mutation discovery required first, a narrowing down the genome by application of candidate gene resequencing, linkage analysis, homozygosity mapping, or case-control association studies, followed by a search for causative mutations within a small part of the genome. Specific clinical tests would then be developed to assess for the presence of a known mutation. The advent of affordable technology to sequence a human genome has rapidly changed physicians' and researchers' approach across a wide range of congenital disorders. It is now possible to sequence all the protein-coding DNA (the "exome") of a patient at a cost that approaches the price of clinical sequencing of a single gene by traditional methods (15). For example, the widely used Illumina 2500 next-generation DNA sequencing platform can sequence an entire genome in 27 hours for under $10,000. Thus, next-generation sequencing technology enables researchers and clinicians to examine almost all mutations in the whole genome, or the protein encoding portion thereof, at a cost much lower than that of previous locus-specific methods. In particular, whole-exome sequencing is a transformative technology that may alter the clinician's approach to the evaluation of CDD patients (16). Accordingly, we sought to determine the effectiveness of whole-exome sequencing to identify, in genes that have been reported to be associated with CDD, the molecular causes of the disease in a cohort of patients with congenital gastrointestinal disorders that had defied conventional diagnostic methods.

## Methods

*Subjects.* Samples for exome sequencing were identified from the UCLA Pediatric Diarrhea Database, which includes samples referred for clinical diagnosis or research since 2004, and was approved by our institutional review board. Inclusion criteria for the database were a history of chronic (>2 mo) severe diarrhea during childhood (<18 y), although subjects with various causes of short-bowel syndrome, inflammatory bowel disease (but not infantile IBD), celiac disease, and pancreatic insufficiency were excluded. The database contains more than 172 kindreds comprising 194 children with chronic diarrhea, of which 163 cases were classified as congenital in origin. Approximately 25 of the subjects were identified with various forms of selective malabsorptive diarrhea and 133 were classified with the generalized form of malabsorption.

We chose 45 patients from 38 families for exome sequencing. Inclusion criteria were a diagnosis of congenital diarrhea and probability that the disease had a genetic cause (consanguineous parents or affected family members). Patients were excluded if they had a confirmed genetic diagnosis or a clinical presentation strongly suggesting a mutation in a gene known to cause congenital diarrhea. These patients were of diverse ethnic backgrounds and had clinical presentations of generalized malabsorptive diarrhea (n=33), selective nutrient malabsorption (n=5), secretory diarrhea (n=3), and infantile IBD (n=4) [Table 1].

*Whole-exome sequencing.* Over the course of this study we evaluated several methodologies for sequencing library preparation and sequencing platform. Genomic DNA from probands, and in some cases family members, was either fragmented by sonication and ligated to Illumina bar-coded adapters or fragmented and ligated in a single step with Illumina engineered transposases, and then in either case, the fragments were amplified by PCR. Fragments were then enriched for the protein coding portion of the genome by hybridization to probes from either the Agilent SureSelect XT Human All Exon 50Mb, Illumina TruSeq Exome, or Illumina Nextera ExpandedExome enrichment kits. Different kits were used for assessment of coverage and sample preparation efficiency. The exome-enriched library was sequenced for 100x100 paired end reads on an Illumina Genome Analyzer 2000 or 2500 platform to a mean coverage depth of 116X, with 85% of RefGene CDS and essential splice sites having at least 20X coverage [Table 2] (Data not shown for subject 45, who was sequenced on an Applied Biosystems SOLiD 4 System instrument as described previously (2)).

Variants that were predicted to be damaging in genes known to be associated with CDD were validated by Sanger sequencing. Sanger sequencing of relatives from whom DNA was available confirmed segregation of the variant allele with the disorder.

*Data analysis.* We converted sequenced reads from the native bcl files to the FastQ format by the Illumina bcl2fastq program. We processed the FastQ files to create aligned bam files with in-house pipeline software. Briefly, we aligned the reads to build GRCh37 of the human genome (17) with Novoalign (http://www.novocraft.com) to obtain a mean of 120 million uniquely aligned 100x100 paired end reads per sample after removing PCR duplicates [Table 2]. We recalibrated base quality scores to improve accuracy by analyzing the covariation among reported quality score, position within read, dinucleotide, and probability of mismatching the reference genome using the Genome Analysis Toolkit (GATK) (18, 19). We used the GATK Unified Genotyper and Haplotype Caller tools to genotype single nucleotide variants and indels and recalibrated variant quality score recalibration with the GATK Variant Quality Score Recalibrator to assign probabilities to each variant call. We obtained the variant consequences on transcripts and proteins with the Ensembl Variant Effect Predictor (VEP) (19), and estimated the extent of protein damage with SIFT (20-24), PolyPhen (25-27), and Condel (28). We further annotated variants with additional data from Online Mendelian Inheritance in Man (OMIM) (29), the Human Gene Mutation Database (HGMD pro, BIOBASE Biological Databases), the Universal Protein Resource (UniProt) (30), KEGG Pathways (31), RefGene (32), the MitoCarta Inventory of Mammalian Mitochondrial Genes (33), Mouse Genome Informatics (MGI) (34)  and the Human Protein Atlas (HPA) (35) using in-house plugins for the VEP.

We hypothesized for this study that the mode of inheritance would be recessive (homozygous or compound heterozygous) and mutation effects would be fully penetrant. Further hypothesizing that casual variants would be found in the protein coding or splicing regions of genes, we filtered to include only splice acceptor or donor, stop gained or lost, frameshift, initiator codon, inframe insertion or deletion, missense, or splice region variants (Ensembl consequence rank >13). We also removed variants with a minor allele frequency >0.5% in the combined 1000 Genomes (36) and NHLBI (37) datasets to remove variants that, given the rare incidence of CDDs, are too frequent in the population to cause these disorders. In addition, we removed variants that were observed in more than 2 (homozygous model) or 8 (compound heterozygous model) samples from ~150 unaffected control exomes, to eliminate false positives caused by technical artifacts. We ranked variants for likelihood of damage using multiple factors. We deemed a splice acceptor or donor variant, stop gained, and frameshift, to be probably damaging. We prioritized missense single nucleotide variants (SNVs) by SIFT, PolyPhen, and Condel predictions, and variants in the splice site region near the acceptor and donor by GERP conservation scores. In the case of compound heterozygous variants we assigned the priority of the second ranked variant. In families where we had sequenced members other than the proband, we filtered out all variants that were inconsistent with fully penetrant Mendelian inheritance. Finally, for this phase of the study, we selected variants in genes that were known or suspected of being involved in congenital diarrheal disorders [Table 3] and had reported mutations that cause a phenotype consistent with that of the patient.

## Results

After filtering by consequence rank, allele frequency, and controls and eliminating variants that did not segregate with the disease, the probands had 1-237 variants in all genes and a mean of two filtered variants in one of the genes known or suspected to be associated with CDDs (range 0-5) [Table 4]. After applying bioinformatic metrics of protein damage and conservation, reviewing the literature, and comparing the expected phenotype of a mutation in a gene to the proband's presenting phenotype, we identified 31 different mutations in 21 families (27 cases) that we considered to be the likely cause of the disease [Table 5]. The genes (and number of families) implicated were ADAM17 (*n*=1), DGAT1 (*n*=2), EPCAM (*n*=3), IL10RA (*n*=1), MALT1 (*n*=1), MYO5B (*n*=2), NEUROG3 (*n*=3), PCSK1 (*n*=2), SI (*n*=3), SKIV2L (*n*=1), SLC26A3 (*n*=1), and SLC5A1 (*n*=1). Six of these variants were previously known to cause a CDD, but the other 25 are novel. All of these mutations had a recessive mode of inheritance; mutations in a gene were homozygous in 13 probands and compound heterozygous in the remaining 9 (one proband had mutations in two genes). The variant consequences included large deletion (*n*=1), splice donor (*n*=2), stop gained (*n*=4), frameshift (*n*=5), missense (*n*=17), and splice region (*n*=2). In summary, we identified a mutation highly likely to cause the patient's phenotype in 60% of the probands that we sequenced.

Discussion

Whole-exome sequencing identified a probable molecular explanation for the disorder in a majority of cases, thus suggesting the value of this approach for diagnosis of CDDs in a clinical setting. Certified clinical sequencing of the genes and in some cases clinical follow up to confirm genotype/phenotype correlations would be needed to confirm these findings but each phenotype, as recorded in the UCLA Pediatric Diarrhea Database, is consistent with the reported phenotypes associated with mutations in the identified casual genes. For example, loss of function mutations in the metallopeptidase domain 17 gene (ADAM17) cause neonatal-onset inflammatory skin and bowel disease with diarrhea (38). A rare splice site mutation in diacylglycerol O-acyltransferase (DGAT1) has been linked to CDD (39). The epithelial cell adhesion molecule (EPCAM) is the casual gene for congenital tufting enteropathy (12). Mutations in the interlukin-10 receptor, alpha gene (IL10RA) cause a form of early onset inflammatory bowel disease (40). The mucosa associated lymphoid tissue lymphoma translocation gene (MALT1), which has been implicated in combined immunodeficiency (41), can also present with generalized malabsorptive diarrhea. Microvillus inclusion disease is caused by mutations in the myosin VB (MYO5B) gene (10). Congenital malabsorptive diarrhea is caused by mutations in the transcription factor neurogenin 3 (NEUROG3) (14). Proprotein convertase 1/3 (PC1/3) deficiency, an autosomal-recessive disorder caused by rare mutations in the proprotein convertase subtilisin/kexin type 1 (PCSK1) gene frequently cause generalized malabsorptive diarrhea in childhood (42). Congenital sucrase-isomaltase deficiency is a selective nutrient malabsorption disorder caused by mutations in the sucrase-isomaltase (alpha-glucosidase) (SI) gene (43). The superkiller viralicidic activity 2-like (S. cerevisiae) gene (SKIV2L) causes trichohepatoenteric syndrome-2, a severe disease characterized by intrauterine growth retardation, facial dysmorphism, hair abnormalities, intractable diarrhea, and immunodeficiency (44). Congenital chloride diarrhea is characterized by excretion of large amounts of watery stool containing high levels of chloride; it is caused by mutations in solute carrier family 26, member 3 (SLC26A3) (8). An intestinal monosaccharide transporter deficiency known as glucose/galactose malabsorption is caused by mutations in solute carrier family 5 (sodium/glucose cotransporter), member 1 (SLC5A1) (45).

Infants presenting with CDD often endure multiple hospitalizations, a batteries of indirect, redundant, and expensive tests, and the life threatening risks of PN, as well as even more serious procedures such as multivisceral transplantation. Whole-exome sequencing, now becoming widely available from certified diagnostic centers (46, 47), is a revolutionary technology that that should be considered early in a clinician’s evaluation of patients with CDD.

However, our study also demonstrates that much remains unknown about the genetic etiology of CDDs, as 40% of the probands we sequenced did not yield a clear molecular finding, although many interesting and novel candidate genes have been identified. There are several factors contributing to this “genetic dark matter”. Most importantly, the cellular pathways involved in the development and function of the intestine are not fully understood and many genes are yet to be identified that contribute to the risk of CDDs. The present study generated a number of candidate genes that are now the subject of active research.

A second consideration is that many de novo cases of CDD without a family history of disease or evidence of consanguinity are likely be caused by sporadic mutations with dominant effect and possibly found in novel genes. Before exome sequencing, it was technically difficult to identify this class of mutation in a single case, and identifying a variant in an autosomal dominant mode of inheritance would usually require that the variant did not have a serious effect on reproductive fitness and was, therefore, present in a large kindred whose DNA was available to researchers. Accordingly, we are aware of only two genes reported before exome sequencing was introduced where mutations cause a CDD, namely, protease, serine, 1 (Trypsin 1) (PRSS1) which causes hereditary pancreatitis (48), and autoimmune regulator (AIRE), which apparently caused autoimmune polyendocrinopathy syndrome type I with a dominant pattern of inheritance in a single family (49). Exome sequencing of a single proband usually returns too many heterozygous candidate variants (mean 315 in our study) to be practically studied for their effect on function. However, exome sequencing of both parents and an affected child allows de novo mutations to be identified from a few candidates (50). A third source of genetic dark matter is the 98% of the human genome that is not in the exome, which includes promoters, enhancers, short RNAs, and other regulatory elements. Identifying and characterizing these features is an active area of research, but at this time the size of the genome to be sequenced and the lack of data on these features limits our ability to determine how they affect phenotypes.

Because some of the cases in this study were in consanguineous families, we considered the approach of using a high density microarray to locate regions of homozygosity in the patients' genome, then developing a set of custom capture probes to select that region for deep sequencing, or to sequence all exons in the region with the traditional Sanger method. However this would require a different array for each family and was unlikely to be cost effective. Furthermore, we wished to develop and evaluate  an unbiased analytical pipeline that could be used for many other genetic patterns in addition to homozygosity arising from inbreeding.

Our success in identifying a majority of the probable causative mutations by screening with a relatively short list of known genes raises the question of whether it would be more efficient to perform targeted exome sequencing for these genes instead of whole-exome sequencing. Comparing the costs of these two options is rather complex and depends on variable such as type of sequencing platform (high throughput or low throughput), caseload and batching factors, and the relative cost of a standard probe set versus a custom panel. Another factor to consider is that new genes will be discovered and a targeted panel will become obsolescent. Finally, whole-exome sequencing has the potential to generate novel candidate genes as a side effect, and with proper consenting and safeguards, this data can usefully be employed in a research setting.

In sum, we believe that with present technology, and for several years to come, whole-exome sequencing is an effective approach for the identification of casual mutations in known CDD genes that may escape detection with standard practice involving a complex diagnostic workup and targeted gene sequencing. We envision that portions of the standard metabolic panels and other urine, blood and radiographic tests will be used less frequently once exome sequencing becomes fully implemented into clinical practice. More directed phenotypic evaluation will be possible after a molecular basis for the gene is known, rather than the shotgun approaches typically employed currently for children with rare genetic conditions.

Tables

| Family | Probands | Others | Population | Phenotype |
|---|---|---|---|---|
| 1 | 1 | | | GMD |
| 2 | 1 | | | diarrhea; fistulizing disease |
| 7 | 1 | | Mexican in US | GMD |
| 45 | 1 | | Hispanic in US | GMD |
| 46 | 1 | f,m | Pakistani in US | GMD |
| 52 | 1 | m,2 | Pakistani in US | GMD, no secretory cells |
| 54 | 1 | | Arab in Turkey | GMD, bile acid malabsorption |
| 72 | 1 | | Caucasian | secretory diarrhea |
| 73 | 1 | | Caucasian | GMD |
| 81 | 1 | f,m,1 | Irish | GMD |
| 82 | 1 | | Caucasian | GGM |
| 93 | 1 | | Arab in Turkey | GMD, bile acid malabsorption |
| 108 | 1 | | Austrian | GGM-like |
| 119 | 2 | f,m | | GMD |
| 125 | 1 | | Caucasian | GGM-like |
| 128 | 1 | | Caucasian | GGM-like |
| 133 | 1 | f,m | Caucasian | secretory diarrhea |
| 137 | 1 | f,m | | GMD |
| 138 | 1 | | Arab | GMD |
| 141 | 1 | | Mexican | GMD, bile acid malabsorption |
| 145 | 1 | | Arab | GMD, APO-like |
| 148 | 1 | | Spaniard | GMD, no paneth cells, diabetes |
| 149 | 1 | | Arab | diarrhea; constantly inflamed colon |
| 154 | 1 | f,m | Italian/Asian | GMD, tufting |
| 158 | 2 | | Bedouin in Israel | GMD |
| 159 | 2 | 1 | Bedouin in Israel | GMD, Fanconi syndrome |
| 160 | 1 | 1 | Bedouin in Israel | GMD, Fanconi syndrome, rickets |
| 161 | 1 | | Bedouin in Israel | GMD |
| 162 | 1 | | Bedouin in Israel | GGM |
| 165 | 1 | | Arab in Netherlands | GMD, dematitis skin lesions |
| 171 | 1 | | Bedouin in Israel | GMD, tufting |
| 172 | 2 | | Bedouin in Israel | diarrhea; fistulizing disease |
| 173 | 1 | f,m | Caucasian | GMD |
| 174 | 1 | f,m | Hondudas | GMD |
| 180 | 1 | f,m | | secretory diarrhea |
| 181 | 2 | | | GMD |
| 1819 | 2 | | Mexican in US | GMD |
| 9899 | 2 | f,m,1 | Mexican in US | malabsorptive diarrhea |

**Table 1. Subjects.** Children afffected with a congenital diarrhea disorder that were sequenced in this study. Probands: number of affected siblings; Others: other relatives sequenced (father, mother, number of unaffected siblings); Population: self-reported ethnic background, when available; Phenotype: presenting diagnosis prior to sequencing, GMD generalized malabsorptive diarrhea, GGM glucose galactose malabsorption.

| id | capture kit | reads (M) | coverage | fraction bases ≥20X |
|---|---|---|---|---|
| 1 | Illumina | 140 | 127 | 0.90 |
| 2 | Illumina | 147 | 128 | 0.90 |
| 7 | Illumina | 71 | 75 | 0.89 |
| 46 | Illumina | 124 | 109 | 0.88 |
| 52 | Illumina | 96 | 103 | 0.90 |
| 54 | Illumina | 19 | 24 | 0.49 |
| 72 | Agilent | 123 | 136 | 0.85 |
| 73 | Illumina | 70 | 78 | 0.87 |
| 81 | Illumina | 182 | 198 | 0.92 |
| 82 | Agilent | 187 | 221 | 0.87 |
| 93 | Illumina | 16 | 18 | 0.27 |
| 108 | Agilent | 189 | 221 | 0.91 |
| 119 | Illumina | 140 | 96 | 0.86 |
| 119C | Illumina | 137 | 96 | 0.86 |
| 125 | Agilent | 92 | 98 | 0.83 |
| 128 | Agilent | 129 | 146 | 0.86 |
| 133 | Agilent | 121 | 63 | 0.84 |
| 137 | Illumina | 85 | 71 | 0.84 |
| 138 | Agilent | 155 | 190 | 0.90 |
| 141 | Agilent | 133 | 92 | 0.89 |
| 145 | Agilent | 89 | 56 | 0.87 |
| 148 | Agilent | 155 | 185 | 0.91 |
| 149 | Agilent | 79 | 52 | 0.85 |
| 154A | Agilent | 94 | 81 | 0.80 |
| 158 | Agilent | 86 | 104 | 0.86 |
| 158A | Agilent | 57 | 60 | 0.76 |
| 159 | Agilent | 83 | 100 | 0.86 |
| 159A | Agilent | 134 | 147 | 0.86 |
| 160 | Agilent | 138 | 153 | 0.88 |
| 161 | Agilent | 199 | 232 | 0.92 |
| 162 | Agilent | 187 | 210 | 0.92 |
| 165 | Agilent | 86 | 58 | 0.86 |
| 171 | Illumina | 42 | 53 | 0.79 |
| 172 | Illumina | 200 | 194 | 0.93 |
| 172A | Illumina | 165 | 180 | 0.93 |
| 173 | Illumina | 98 | 70 | 0.80 |
| 174 | Illumina | 138 | 101 | 0.85 |
| 180 | Illumina | 128 | 86 | 0.84 |
| 181 | Illumina | 146 | 104 | 0.87 |
| 181A | Illumina | 127 | 85 | 0.84 |
| 1819 | Illumina | 102 | 109 | 0.91 |

| id | capture kit | reads (M) | coverage | fraction bases ≥20X |
|---|---|---|---|---|
| 1819 | Illumina | 66 | 72 | 0.89 |
| 9899 | Illumina | 106 | 130 | 0.91 |
| 9899 | Illumina | 201 | 206 | 0.93 |
| **mean** | | 120 | 116 | 0.85 |
| **s.d.** | | 46 | 56 | 0.11 |

**Table 2. Sequencing metrics.** Agilent: SureSelect XT Human All Exon 50Mb; Illumina: TruSeq/NexteraRapidCapture ExpandedExome; reads: the number of unique reads that pass the sequencer's quality filters and are aligned with mapping score > 0 to the reference genomecoverage: the mean coverage of all baits in the capture kit; % bases ≥20X: percentage of all RefGene exon and essential splice site bases acheiving 20X or greater coverage.

| **gene** | **name** | **OMIM** |
|---|---|---|
| ADAM17 | ADAM metallopeptidase domain 17 | Inflammatory skin and bowel disease, neonatal |
| AIRE | autoimmune regulator | Autoimmune polyendocrinopathy syndrome , type I, with or without reversible metaphyseal dysplasia |
| APOB | apolipoprotein B (including Ag(x) antigen) | Hypobetalipoproteinemia; Hypobetalipoproteinemia, normotriglyceridemic; Hypercholesterolemia, due to ligand-defective apo B |
| DGAT1 | diacylglycerol O-acyltransferase 1 | |
| EPCAM | epithelial cell adhesion molecule | Diarrhea 5, with tufting enteropathy, congenital; Colorectal cancer, hereditary nonpolyposis, type 8 |
| FOXP3 | forkhead box P3 | Immunodysregulation, polyendocrinopathy, and enteropathy, X-linked; Diabetes mellitus, type I, susceptibility to |
| GUCY2C | guanylate cyclase 2C (heat stable enterotoxin receptor) | Diarrhea 6; Meconium ileus |
| HPS1 | Hermansky-Pudlak syndrome 1 | Hermansky-Pudlak syndrome 1 |
| IL10RA | interleukin 10 receptor, alpha | Inflammatory bowel disease 28, early onset, autosomal recessive |
| LCT | lactase | Lactase deficiency, congenital |
| MGAM | maltase-glucoamylase (alpha-glucosidase) | |
| MLL2 | myeloid/lymphoid or mixed-lineage leukemia 2 | Kabuki syndrome 1 |
| MPI | mannose phosphate isomerase | Congenital disorder of glycosylation, type Ib |
| MTTP | microsomal triglyceride transfer protein | |
| MYO5B | myosin VB | Microvillus inclusion disease |
| NEUROG3 | neurogenin 3 | Diarrhea 4, malabsorptive, congenital |
| PCSK1 | proprotein convertase subtilisin/kexin type 1 | Obesity with impaired prohormone processing; Obesity, susceptibility to, BMIQ12 |
| PNLIP | pancreatic lipase | Pancreatic lipase deficiency |
| RFX6 | regulatory factor X, 6 | Martinez-Frias syndrome |
| SAR1B | SAR1 homolog B (S. cerevisiae) | Chylomicron retention disease |
| SBDS | Shwachman-Bodian-Diamond syndrome | Shwachman-Bodian-Diamond syndrome |
| SI | sucrase-isomaltase (alpha-glucosidase) | Sucrase-isomaltase deficiency, congenital |

| gene | name | OMIM |
|---|---|---|
| SKIV2L | superkiller viralicidic activity 2-like (S. cerevisiae) | Trichohepatoenteric syndrome 2 |
| SLC10A2 | solute carrier family 10 (sodium/bile acid cotransporter family), member 2 | Bile acid malabsorption, primary |
| SLC26A3 | solute carrier family 26, member 3 | Chloride diarrhea, congenital, Finnish type; ?Colon cancer |
| SLC2A2 | solute carrier family 2 (facilitated glucose transporter), member 2 | Diabetes mellitus, noninsulin-dependent; Fanconi-Bickel syndrome |
| SLC2A5 | solute carrier family 2 (facilitated glucose/fructose transporter), member 5 | |
| SLC39A4 | solute carrier family 39 (zinc transporter), member 4 | Acrodermatitis enteropathica |
| SLC46A1 | solute carrier family 46 (folate transporter), member 1 | Folate malabsorption, hereditary |
| SLC5A1 | solute carrier family 5 (sodium/glucose cotransporter), member 1 | Glucose/galactose malabsorption |
| SLC7A7 | solute carrier family 7 (amino acid transporter light chain, y+L system), member 7 | Lysinuric protein intolerance |
| SPINT2 | serine peptidase inhibitor, Kunitz type, 2 | Diarrhea 3, secretory sodium, congenital, syndromic |
| TCN2 | transcobalamin II | Transcobalamin II deficiency |
| TMPRSS15 | transmembrane protease, serine 15 | |
| TREH | trehalase (brush-border membrane glycoprotein) | Trehalase deficiency |
| TTC37 | tetratricopeptide repeat domain 37 | Trichohepatoenteric syndrome 1 |
| UBR1 | ubiquitin protein ligase E3 component n-recognin 1 | Johanson-Blizzard syndrome |

**Table 3.** Genes. Genes known or suspected to be associated with congenital diarrheal disorders. OMIM: disease annotation of gene in OMIM.

| id | filtered variants | filtered SNVs | filtered indels | known gene variants | known gene snvs | known gene indels |
|---|---|---|---|---|---|---|
| 1 | 127 | 79 | 48 | 4 | 4 | 0 |
| 2 | 77 | 55 | 22 | 1 | 0 | 0 |
| 7 | 76 | 52 | 24 | 1 | 0 | 0 |
| 46 | 34 | 28 | 6 | 1 | 1 | 0 |
| 52 | 39 | 32 | 7 | 1 | 0 | 0 |
| 54 | 133 | 68 | 65 | 1 | 0 | 0 |
| 72 | 104 | 85 | 19 | 4 | 4 | 0 |
| 73 | 70 | 50 | 20 | 1 | 0 | 0 |
| 81 | 9 | 5 | 4 | 1 | 0 | 0 |
| 82 | 116 | 97 | 19 | 2 | 2 | 1 |
| 93 | 119 | 77 | 42 | 1 | 0 | 0 |
| 108 | 56 | 41 | 15 | 1 | 0 | 0 |
| 119 | 2 | 2 | 0 | 2 | 2 | 0 |
| 119C | 2 | 2 | 0 | 2 | 2 | 0 |
| 125 | 121 | 95 | 26 | 4 | 4 | 2 |
| 128 | 195 | 176 | 19 | 2 | 2 | 0 |
| 133 | 16 | 11 | 5 | 1 | 0 | 0 |
| 137 | 13 | 6 | 7 | 1 | 0 | 0 |
| 138 | 155 | 127 | 28 | 2 | 2 | 0 |
| 141 | 134 | 87 | 47 | 2 | 2 | 1 |
| 145 | 237 | 182 | 55 | 2 | 2 | 0 |
| 148 | 165 | 121 | 44 | 1 | 0 | 0 |
| 149 | 152 | 114 | 38 | 1 | 0 | 0 |
| 154A | 7 | 3 | 4 | 1 | 0 | 0 |
| 158 | 58 | 44 | 14 | 1 | 1 | 1 |
| 158A | 37 | 31 | 6 | 1 | 1 | 1 |
| 159 | 40 | 33 | 7 | 1 | 1 | 0 |
| 159A | 35 | 30 | 5 | 1 | 1 | 0 |
| 160 | 103 | 82 | 21 | 1 | 0 | 0 |
| 161 | 169 | 127 | 42 | 5 | 5 | 0 |
| 162 | 153 | 121 | 32 | 1 | 1 | 0 |
| 165 | 151 | 100 | 51 | 2 | 2 | 1 |
| 171 | 175 | 112 | 63 | 5 | 5 | 0 |
| 172 | 54 | 44 | 10 | 2 | 2 | 0 |
| 172A | 40 | 33 | 7 | 2 | 2 | 0 |
| 173 | 12 | 6 | 6 | 1 | 0 | 0 |
| 174 | 17 | 11 | 6 | 1 | 1 | 0 |
| 180 | 12 | 9 | 3 | 4 | 4 | 0 |
| 181 | 33 | 27 | 6 | 1 | 0 | 0 |
| 181A | 23 | 20 | 3 | 1 | 0 | 0 |
| 1819 | 4 | 3 | 1 | 2 | 2 | 1 |

| id | filtered variants | filtered SNVs | filtered indels | known gene variants | known gene snvs | known gene indels |
|---|---|---|---|---|---|---|
| 1819 | 4 | 3 | 1 | 2 | 2 | 1 |
| 9899 | 1 | 1 | 0 | 2 | 0 | 0 |
| 9899 | 1 | 1 | 0 | 0 | 0 | 0 |
| **mean** | 75 | 55 | 19 | 2 | 1 | 0 |
| **stdev** | 65 | 49 | 19 | 1 | 1 | 0 |
| **min** | 1 | 1 | 0 | 0 | 0 | 0 |
| **max** | 237 | 182 | 65 | 5 | 5 | 2 |

**Table 4.** Variant counts. Filtered: homozygous or compound heterozygous variants that satisfied consequence, allele frequency, exome control , and segregation constraints; known gene: filtered variants that were in genes known or suspected to be associated with CCDs.

| ID | Gene | GT | cDNA | protein | conseq | SIFT | PolyPhen | Condel | GERP | HGMD |
|---|---|---|---|---|---|---|---|---|---|---|
| 45 | PCSK1 | 1/1 | 1029C>G | Tyr343Ter | stop gained | | | | -9.280 | n |
| 46 | EPCAM | 1/1 | 377G>A | Arg126Lys | missense | D | D | D | | n |
| 72 | MYO5B | 0/1 | 3698G>T | Ser1233Ile | missense | t | b | n | -0.826 | n |
| | | 0/1 | 4240G>A | Glu1414Lys | missense | D | D | D | 4.65 | n |
| 82 | SI | 0/1 | 5234T>G | Phe1745Cys | missense | D | D | D | 4.14 | * |
| | | 0/1 | 635+2dupT | | splice region | | | | 2.555 | n |
| 119 | SKIV2L | 0/1 | 3188G>C | Arg1063Pro | missense | D | D | D | 3.2 | n |
| | | 0/1 | 3629T>C | Leu1210Pro | missense | D | D | D | 4.3 | n |
| 125 | SI | 0/1 | 3218G>A | Gly1073Asp | missense | D | D | D | 4.65 | * |
| | | 0/1 | 834_837delAACA | Gln278HisfsTer18 | frameshift | | | | -0.064 | n |
| 128 | SI | 0/1 | 1730T>G | Val577Gly | missense | D | D | D | 3.83 | * |
| | | 0/1 | 5234T>G | Phe1745Cys | missense | D | D | D | 4.14 | * |
| 138 | DGAT1 | 1/1 | chr8:g.145541784_145541915del | | deletion* | * this deletion spans the splice acceptor and part of the exon (novel) | | | | |
| 141 | SLC26A3 | 0/1 | 1177G>T | Gly393Trp | missense | D | D | D | 0.907 | n |
| | | 0/1 | 1939delC | His647ThrfsTer18 | frameshift | | | | 4.92 | n |
| 158 | EPCAM | 1/1 | 578delT | Ile193MetfsTer17 | frameshift | | | | | n |
| 159 | NEUROG3 | 1/1 | 410A>G | Gln137Arg | missense | D | benign | n | 4.54 | n |
| 160 | NEUROG3 | 1/1 | 556G>C | Gly186Arg | missense | D | D | D | 3.82 | n |
| 161 | NEUROG3 | 1/1 | 410A>G | Gln137Arg | missense | D | benign | n | 4.54 | n |
| 162 | SLC5A1 | 1/1 | 947T>C | Leu316Pro | missense | D | D | D | 3.91 | n |
| 165 | ADAM17 | 1/1 | 308dupA | Asn103LysfsTer20 | frameshift | | | | -1.705 | n |
| 171 | EPCAM | 1/1 | 227C>G | Ser76Ter | stop gained | | | | | n |
| | PCSK1 | 0/1 | 1069A>G | Ser357Gly | missense | D | pD | D | 4.64 | n |
| | | 0/1 | c.1213C>T | Arg405Ter | stop gained | | | | -1.39 | n |
| 172 | IL10RA | 1/1 | 537G>A | | splice region | | | | 3.8 | n |
| 173 | | | | | | | | | | |
| 174 | DGAT1 | 1/1 | 751+2T>C | | splice donor | | | | 3.67 | * |
| 180 | MYO5B | 0/1 | 1744G>C | Ala582Pro | missense | D | b | n | 2.45 | n |
| | | 0/1 | 4036C>T | Gln1346Ter | stop gained | | | | 3.71 | n |
| 181 | MALT1 | 1/1 | 550G>T | Asp184Tyr | missense | D | D | D | 3.51 | n |
| 1819 | EPCAM | 0/1 | 491+1G>A | | splice donor | | | | | * |
| | | 0/1 | 538delT | Phe180LeufsTer30 | frameshift | | | | | n |

**Table 5. Subjects and sequencing results.** Gene in which a damaging mutation is considered highly likely to be causitive of the disorder. GT: genotype, 0/1=homozygous, 1/1=compound heterozygous; cDNA: effect of mutation on transcript; protein: effect of the mutation on protein; conseq: classification of mutation's effect; SIFT: prediction of effect of mutation on protein by SIFT (D deleterious, t tolerated); PolyPhen: prediction of effect of mutation on protein by PolyPhen (D probably damaging, pD possibly damaging, b benign); Condel: consensus of SIFT and PolyPhen predictions as calculated by Condel (D deleterious, n neutral); GERP: GERP conservation score; HGMD: variant annotated in the HGMD database (*=present in HGMD, n=no HGMD annotation).

# References

1. Terrin G, Tomaiuolo R, Passariello A, Elce A, Amato F, Di Costanzo M, Castaldo G, Canani RB. Congenital diarrheal disorders: an updated diagnostic approach. International journal of molecular sciences. 2012;13(4):4168-85. doi: 10.3390/ijms13044168. PubMed PMID: 22605972; PubMed Central PMCID: PMC3344208.

2. Yourshaw M, Solorzano-Vargas RS, Pickett LA, Lindberg I, Wang J, Cortina G, Pawlikowska-Haddal A, Baron H, Venick RS, Nelson SF. Exome Sequencing Finds a Novel PCSK1 Mutation in a Child With Generalized Malabsorptive Diarrhea and Diabetes Insipidus. Journal of pediatric gastroenterology and nutrition. 2013. doi: 10.1097/MPG.0b013e3182a8ae6c.

3. Jackson RS, Creemers JW, Ohagi S, Raffin-Sanson ML, Sanders L, Montague CT, Hutton JC, O'Rahilly S. Obesity and impaired prohormone processing associated with mutations in the human prohormone convertase 1 gene. Nat Genet. 1997;16(3):303-6. Epub 1997/07/01. doi: 10.1038/ng0797-303. PubMed PMID: 9207799.

4. Spencer AU, Kovacevich D, McKinney-Barnett M, Hair D, Canham J, Maksym C, Teitelbaum DH. Pediatric short-bowel syndrome: the cost of comprehensive care. The American journal of clinical nutrition. 2008;88(6):1552-9. Epub 2008/12/10. doi: 10.3945/ajcn.2008.26007. PubMed PMID: 19064515.

5. Hearne CM, Ghosh S, Todd JA. Microsatellites for Linkage Analysis of Genetic-Traits. Trends in Genetics. 1992;8(8):288-94. doi: Doi 10.1016/0168-9525(92)90137-S. PubMed PMID: WOS:A1992JE96800009.

6. Kere J, Sistonen P, Holmberg C, de la Chapelle A. The gene for congenital chloride diarrhea maps close to but is distinct from the gene for cystic fibrosis transmembrane conductance regulator. Proc Natl Acad Sci U S A. 1993;90(22):10686-9. Epub 1993/11/15. PubMed PMID: 7504277; PubMed Central PMCID: PMC47842.

7. Hoglund P, Haila S, Scherer SW, Tsui LC, Green ED, Weissenbach J, Holmberg C, de la Chapelle A, Kere J. Positional candidate genes for congenital chloride diarrhea suggested by high-resolution physical mapping in chromosome region 7q31. Genome Res. 1996;6(3):202-10. Epub 1996/03/01. PubMed PMID: 8963897.

8. Hoglund P, Haila S, Socha J, Tomaszewski L, Saarialho-Kere U, Karjalainen-Lindsberg ML, Airola K, Holmberg C, de la Chapelle A, Kere J. Mutations of the Down-regulated in adenoma (DRA) gene cause congenital chloride diarrhoea. Nat Genet. 1996;14(3):316-9. Epub 1996/11/01. doi: 10.1038/ng1196-316. PubMed PMID: 8896562.

9. Sellick GS, Longman C, Tolmie J, Newbury-Ecob R, Geenhalgh L, Hughes S, Whiteford M, Garrett C, Houlston RS. Genomewide linkage searches for Mendelian disease loci can be efficiently conducted using high-density SNP genotyping arrays. Nucleic Acids Res. 2004;32(20):e164. Epub 2004/11/25. doi: 10.1093/nar/gnh163. PubMed PMID: 15561999; PubMed Central PMCID: PMC534642.

10. Muller T, Hess MW, Schiefermeier N, Pfaller K, Ebner HL, Heinz-Erian P, Ponstingl H, Partsch J, Rollinghoff B, Kohler H, Berger T, Lenhartz H, Schlenck B, Houwen RJ, Taylor CJ, Zoller H, Lechner S, Goulet O, Utermann G, Ruemmele FM, Huber LA, Janecke AR. MYO5B

mutations cause microvillus inclusion disease and disrupt epithelial cell polarity. Nat Genet. 2008;40(10):1163-5. doi: 10.1038/ng.225. PubMed PMID: 18724368.

11. Heinz-Erian P, Muller T, Krabichler B, Schranz M, Becker C, Ruschendorf F, Nurnberg P, Rossier B, Vujic M, Booth IW, Holmberg C, Wijmenga C, Grigelioniene G, Kneepkens CM, Rosipal S, Mistrik M, Kappler M, Michaud L, Doczy LC, Siu VM, Krantz M, Zoller H, Utermann G, Janecke AR. Mutations in SPINT2 cause a syndromic form of congenital sodium diarrhea. Am J Hum Genet. 2009;84(2):188-96. Epub 2009/02/03. doi: 10.1016/j.ajhg.2009.01.004. PubMed PMID: 19185281; PubMed Central PMCID: PMC2668003.

12. Sivagnanam M, Mueller JL, Lee H, Chen Z, Nelson SF, Turner D, Zlotkin SH, Pencharz PB, Ngan BY, Libiger O, Schork NJ, Lavine JE, Taylor S, Newbury RO, Kolodner RD, Hoffman HM. Identification of EpCAM as the gene for congenital tufting enteropathy. Gastroenterology. 2008;135(2):429-37. doi: Doi 10.1053/J.Gastro.2008.05.036. PubMed PMID: ISI:000258439900020.

13. Fiskerstrand T, Arshad N, Haukanes BI, Tronstad RR, Pham KD, Johansson S, Havik B, Tonder SL, Levy SE, Brackman D, Boman H, Biswas KH, Apold J, Hovdenak N, Visweswariah SS, Knappskog PM. Familial diarrhea syndrome caused by an activating GUCY2C mutation. N Engl J Med. 2012;366(17):1586-95. Epub 2012/03/23. doi: 10.1056/NEJMoa1110132. PubMed PMID: 22436048.

14. Wang J, Cortina G, Wu SV, Tran R, Cho JH, Tsai MJ, Bailey TJ, Jamrich M, Ament ME, Treem WR, Hill ID, Vargas JH, Gershman G, Farmer DG, Reyen L, Martin MG. Mutant neurogenin-3 in congenital malabsorptive diarrhea. N Engl J Med. 2006;355(3):270-80. doi: 10.1056/NEJMoa054288. PubMed PMID: 16855267.

15. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 2011;12(11):745-55. Epub 2011/09/29. doi: 10.1038/nrg3031. PubMed PMID: 21946919.

16. Maxmen A. Exome sequencing deciphers rare diseases. Cell. 2011;144(5):635-7. Epub 2011/03/08. doi: 10.1016/j.cell.2011.02.033. PubMed PMID: 21376225.

17. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931-45. Epub 2004/10/22. doi: 10.1038/nature03001. PubMed PMID: 15496913.

18. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491-8. Epub 2011/04/12. doi: 10.1038/ng.806. PubMed PMID: 21478889; PubMed Central PMCID: PMC3083463.

19. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-303. Epub 2010/07/21. doi: 10.1101/gr.107524.110. PubMed PMID: 20644199; PubMed Central PMCID: PMC2928508.

20. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols. 2009;4(7):1073-81. Epub 2009/06/30. doi: 10.1038/nprot.2009.86. PubMed PMID: 19561590.

21. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001;11(5):863-74. Epub 2001/05/05. doi: 10.1101/gr.176601. PubMed PMID: 11337480; PubMed Central PMCID: PMC311071.

22. Ng PC, Henikoff S. Accounting for human polymorphisms predicted to affect protein function. Genome Res. 2002;12(3):436-46. Epub 2002/03/05. doi: 10.1101/gr.212802. PubMed PMID: 11875032; PubMed Central PMCID: PMC155281.

23. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812-4. Epub 2003/06/26. PubMed PMID: 12824425; PubMed Central PMCID: PMC168916.

24. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. Annual review of genomics and human genetics. 2006;7:61-80. Epub 2006/07/11. doi: 10.1146/annurev.genom.7.080505.115630. PubMed PMID: 16824020.

25. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 2002;30(17):3894-900. Epub 2002/08/31. PubMed PMID: 12202775; PubMed Central PMCID: PMC137415.

26. Sunyaev S, Ramensky V, Bork P. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. Trends Genet. 2000;16(5):198-200. Epub 2000/04/27. PubMed PMID: 10782110.

27. Sunyaev S, Ramensky V, Koch I, Lathe W, 3rd, Kondrashov AS, Bork P. Prediction of deleterious human alleles. Hum Mol Genet. 2001;10(6):591-7. Epub 2001/03/07. PubMed PMID: 11230178.

28. Gonzalez-Perez A, Lopez-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet. 2011;88(4):440-9. Epub 2011/04/05. doi: 10.1016/j.ajhg.2011.03.004. PubMed PMID: 21457909; PubMed Central PMCID: PMC3071923.

29. Online Mendelian Inheritance in Man OMIM®. Online Mendelian Inheritance in Man, OMIM® Baltimore, MD: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University; [2013-11-04]. Available from: http://omim.org/.

30. Apweiler R, Martin MJ, O'Donovan C, Magrane M, Alam-Faruque Y, Alpi E, Antunes R, Arganiska J, Casanova EB, Bely B, Bingley M, Bonilla C, Britto R, Bursteinas B, Chan WM, Chavali G, Cibrian-Uhalte E, Da Silva A, De Giorgi M, Dimmer E, Fazzini F, Gane P, Fedotov A, Castro LG, Garmiri P, Hatton-Ellis E, Hieta R, Huntley R, Jacobsen J, Jones R, Legge D, Liu WD, Luo J, MacDougall A, Mutowo P, Nightingale A, Orchard S, Patient S, Pichler K, Poggioli D, Pundir S, Pureza L, Qi GY, Rosanoff S, Sawford T, Sehra H, Turner E, Volynkin V, Wardell T, Watkins X, Zellner H, Corbett M, Donnelly M, van Rensburg P, Goujon M, McWilliam H, Lopez R, Xenarios I, Bougueleret L, Bridge A, Poux S, Redaschi N, Auchincloss A, Axelsen K, Bansal P, Baratin D, Binz PA, Blatter MC, Boeckmann B, Bolleman J, Boutet E, Breuza L, de Castro E, Cerutti L, Coudert E,

Cuche B, Doche M, Dornevil D, Duvaud S, Estreicher A, Famiglietti L, Feuermann M, Gasteiger E, Gehant S, Gerritsen V, Gos A, Gruaz-Gumowski N, Hinz U, Hulo C, James J, Jungo F, Keller G, Lara V, Lemercier P, Lew J, Lieberherr D, Martin X, Masson P, Morgat A, Neto T, Paesano S, Pedruzzi I, Pilbout S, Pozzato M, Pruess M, Rivoire C, Roechert B, Schneider M, Sigrist C, Sonesson K, Staehli S, Stutz A, Sundaram S, Tognolli M, Verbregue L, Veuthey AL, Zerara M, Wu CH, Arighi CN, Arminski L, Chen CM, Chen YX, Huang HZ, Kukreja A, Laiho K, McGarvey P, Natale DA, Natarajan TG, Roberts NV, Suzek BE, Vinayaka CR, Wang QH, Wang YQ, Yeh LS, Yerramalla MS, Zhang J, Consortium U. Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Research. 2013;41(D1):D43-D7. doi: Doi 10.1093/Nar/Gks1068. PubMed PMID: WOS:000312893300007.

31. Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. Nucleic Acids Res. 2002;30(1):42-6. PubMed PMID: 11752249; PubMed Central PMCID: PMC99091.

32. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. 2012;40(Database issue):D130-5. doi: 10.1093/nar/gkr1079. PubMed PMID: 22121212; PubMed Central PMCID: PMC3245008.

33. Pagliarini DJ, Calvo SE, Chang B, Sheth SA, Vafai SB, Ong SE, Walford GA, Sugiana C, Boneh A, Chen WK, Hill DE, Vidal M, Evans JG, Thorburn DR, Carr SA, Mootha VK. A mitochondrial protein compendium elucidates complex I disease biology. Cell. 2008;134(1):112-23. doi: 10.1016/j.cell.2008.06.016. PubMed PMID: 18614015; PubMed Central PMCID: PMC2778844.

34. Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE, Mouse Genome Database G. The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. Nucleic Acids Res. 2012;40(Database issue):D881-6. doi: 10.1093/nar/gkr974. PubMed PMID: 22075990; PubMed Central PMCID: PMC3245042.

35. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S, Wernerus H, Bjorling L, Ponten F. Towards a knowledge-based Human Protein Atlas. Nat Biotechnol. 2010;28(12):1248-50. doi: 10.1038/nbt1210-1248. PubMed PMID: 21139605.

36. 1000 Genomes Project. VCF (Variant Call Format) version 4.1 2013 [updated 2013-10-09]. Available from: http://www.1000genomes.org/wiki/Analysis/Variant Call Format/vcf-variant-call-format-version-41.

37. Exome Variant Server [Internet]. NHLBI Exome Sequencing Project (ESP). 2011 [cited 2011-09-10]. Available from: http://evs.gs.washington.edu/EVS/.

38. Blaydon DC, Biancheri P, Di WL, Plagnol V, Cabral RM, Brooke MA, van Heel DA, Ruschendorf F, Toynbee M, Walne A, O'Toole EA, Martin JE, Lindley K, Vulliamy T, Abrams DJ, MacDonald TT, Harper JI, Kelsell DP. Inflammatory skin and bowel disease linked to ADAM17 deletion. N Engl J Med. 2011;365(16):1502-8. Epub 2011/10/21. doi: 10.1056/NEJMoa1100721. PubMed PMID: 22010916.

39. Haas JT, Winter HS, Lim E, Kirby A, Blumenstiel B, DeFelice M, Gabriel S, Jalas C, Branski D, Grueter CA, Toporovski MS, Walther TC, Daly MJ, Farese RV, Jr. DGAT1 mutation is linked to a congenital diarrheal disorder. The Journal of clinical investigation. 2012;122(12):4680-4. doi: 10.1172/JCI64873. PubMed PMID: 23114594; PubMed Central PMCID: PMC3533555.

40. Glocker EO, Kotlarz D, Boztug K, Gertz EM, Schaffer AA, Noyan F, Perro M, Diestelhorst J, Allroth A, Murugan D, Hatscher N, Pfeifer D, Sykora KW, Sauer M, Kreipe H, Lacher M, Nustede R, Woellner C, Baumann U, Salzer U, Koletzko S, Shah N, Segal AW, Sauerbrey A, Buderus S, Snapper SB, Grimbacher B, Klein C. Inflammatory bowel disease and mutations affecting the interleukin-10 receptor. N Engl J Med. 2009;361(21):2033-45. Epub 2009/11/06. doi: 10.1056/NEJMoa0907206. PubMed PMID: 19890111; PubMed Central PMCID: PMC2787406.

41. Jabara HH, Ohsumi T, Chou J, Massaad MJ, Benson H, Megarbane A, Chouery E, Mikhael R, Gorka O, Gewies A, Portales P, Nakayama T, Hosokawa H, Revy P, Herrod H, Le Deist F, Lefranc G, Ruland J, Geha RS. A homozygous mucosa-associated lymphoid tissue 1 (MALT1) mutation in a family with combined immunodeficiency. The Journal of allergy and clinical immunology. 2013;132(1):151-8. Epub 2013/06/04. doi: 10.1016/j.jaci.2013.04.047. PubMed PMID: 23727036; PubMed Central PMCID: PMC3700575.

42. Martin MG, Lindberg I, Solorzano-Vargas RS, Wang J, Avitzur Y, Bandsma R, Sokollik C, Lawrence S, Pickett LA, Chen Z, Egritas O, Dalgic B, Albornoz V, de Ridder L, Hulst J, Gok F, Aydogan A, Al-Hussaini A, Gok DE, Yourshaw M, Wu SV, Cortina G, Stanford S, Georgia S. Congenital proprotein convertase 1/3 deficiency causes malabsorptive diarrhea and other endocrinopathies in a pediatric cohort. Gastroenterology. 2013;145(1):138-48. doi: 10.1053/j.gastro.2013.03.048. PubMed PMID: 23562752.

43. Ouwendijk J, Moolenaar CE, Peters WJ, Hollenberg CP, Ginsel LA, Fransen JA, Naim HY. Congenital sucrase-isomaltase deficiency. Identification of a glutamine to proline substitution that leads to a transport block of sucrase-isomaltase in a pre-Golgi compartment. The Journal of clinical investigation. 1996;97(3):633-41. Epub 1996/02/01. doi: 10.1172/JCI118459. PubMed PMID: 8609217; PubMed Central PMCID: PMC507098.

44. Fabre A, Charroux B, Martinez-Vinson C, Roquelaure B, Odul E, Sayar E, Smith H, Colomb V, Andre N, Hugot JP, Goulet O, Lacoste C, Sarles J, Royet J, Levy N, Badens C. SKIV2L mutations cause syndromic diarrhea, or trichohepatoenteric syndrome. Am J Hum Genet. 2012;90(4):689-92. Epub 2012/03/27. doi: 10.1016/j.ajhg.2012.02.009. PubMed PMID: 22444670; PubMed Central PMCID: PMC3322239.

45. Turk E, Zabel B, Mundlos S, Dyer J, Wright EM. Glucose/galactose malabsorption caused by a defect in the Na+/glucose cotransporter. Nature. 1991;350(6316):354-6. Epub 1991/03/28. doi: 10.1038/350354a0. PubMed PMID: 2008213.

46. Lee H, Nelson SF. Rethinking clinical practice: clinical implementation of exome sequencing. Pers Med. 2012;9(8):785-7. doi: Doi 10.2217/Pme.12.101. PubMed PMID: WOS:000311977800001.

47. Yang J, Duan S, Zhong R, Yin J, Pu J, Ke J, Lu X, Zou L, Zhang H, Zhu Z, Wang D, Xiao H, Guo A, Xia J, Miao X, Tang S, Wang G. Exome sequencing identified NRG3 as a novel susceptible

gene of Hirschsprung's disease in a Chinese population. Molecular neurobiology. 2013;47(3):957-66. doi: 10.1007/s12035-012-8392-4. PubMed PMID: 23315268.

48. Whitcomb DC, Preston RA, Aston CE, Sossenheimer MJ, Barua PS, Zhang Y, Wong-Chong A, White GJ, Wood PG, Gates LK, Jr., Ulrich C, Martin SP, Post JC, Ehrlich GD. A gene for hereditary pancreatitis maps to chromosome 7q35. Gastroenterology. 1996;110(6):1975-80. Epub 1996/06/01. PubMed PMID: 8964426.

49. Cetani F, Barbesino G, Borsari S, Pardi E, Cianferotti L, Pinchera A, Marcocci C. A novel mutation of the autoimmune regulator gene in an Italian kindred with autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy, acting in a dominant fashion and strongly cosegregating with hypothyroid autoimmune thyroiditis. The Journal of clinical endocrinology and metabolism. 2001;86(10):4747-52. Epub 2001/10/16. PubMed PMID: 11600535.

50. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. The next-generation sequencing revolution and its impact on genomics. Cell. 2013;155(1):27-38. Epub 2013/10/01. doi: 10.1016/j.cell.2013.09.006. PubMed PMID: 24074859.

# CHAPTER FOUR

## Exome sequencing finds a novel *PCSK1* mutation in a child with generalized malabsorptive diarrhea and diabetes insipidus

Original Article: Gastroenterology

# Exome Sequencing Finds a Novel *PCSK1* Mutation in a Child With Generalized Malabsorptive Diarrhea and Diabetes Insipidus

*Michael Yourshaw, †R. Sergio Solorzano-Vargas, ‡Lindsay A. Pickett, ‡Iris Lindberg, †Jiafang Wang, ¶Galen Cortina, #Anna Pawlikowska-Haddal, **Howard Baron, †Robert S. Venick, ‡‡Stanley F. Nelson, and †Martín G. Martín

## ABSTRACT

**Objectives:** Congenital diarrhea disorders are a group of genetically diverse and typically autosomal recessive disorders that have yet to be well characterized phenotypically or molecularly. Diagnostic assessments are generally limited to nutritional challenges and histologic evaluation, and many subjects eventually require a prolonged course of intravenous nutrition. Here we describe next-generation sequencing techniques to investigate a child with perplexing congenital malabsorptive diarrhea and other presumably unrelated clinical problems; this method provides an alternative approach to molecular diagnosis.

**Methods:** We screened the diploid genome of an affected individual, using exome sequencing, for uncommon variants that have observed protein-coding consequences. We assessed the functional activity of the mutant protein, as well as its lack of expression using immunohistochemistry.

**Results:** Among several rare variants detected was a homozygous nonsense mutation in the catalytic domain of the *proprotein convertase subtilisin/kexin type 1* gene. The mutation abolishes prohormone convertase 1/3 endoprotease activity as well as expression in the intestine. These primary genetic findings prompted a careful endocrine reevaluation of the child at 4.5 years of age, and multiple significant problems were subsequently identified consistent with the known phenotypic consequences of *proprotein convertase subtilisin/kexin type 1 (PCSK1)* gene mutations. Based on the molecular diagnosis, alternate medical and dietary management was implemented for diabetes insipidus, polyphagia, and micropenis.

**Conclusions:** Whole-exome sequencing provides a powerful diagnostic tool to clinicians managing rare genetic disorders with multiple perplexing clinical manifestations.

**Key Words:** enteroendocrine cell, *Neurogenin-3*, PC1/3, proprotein convertases

(*JPGN* 2013;57: 759–767)

Congenital diarrheal disorders are uncommon yet frequently devastating chronic conditions that are secondary to a diverse group of autosomal recessive mutations. They can be classified into those that selectively impair the transport or hydrolysis of single nutrients or electrolytes and those that attenuate the assimilation of all forms of nutrients (1). They may also be grouped as either malabsorptive or secretory in nature, and by an array of histologic features, including changes within the enterocytes or their migration along the crypt–villus axis.

Regardless, children presenting shortly after birth with severe diarrhea—given their rarity and heterogeneity—are frequently misdiagnosed. Failure to diagnose such patients quickly and accurately can undermine their tenuous hold on life. When correctly diagnosed, subjects with impairment of selective nutrient assimilation generally do well on a lifelong nutrient-specific restricted diet; however, those presenting with malabsorption of multiple forms of nutrients (proteins, carbohydrates, and fats) generally have an adverse clinical course that includes lifelong or prolonged total intravenous (parenteral) nutrition, and/or intestinal and occasionally concomitant liver transplantation (1). Although these disorders are frequently fatal without proper dietary and nutritional modifications, present state-of-the-art therapeutic modalities are primitive and are associated with extremely significant morbidity and mortality, as well as daunting medical care costs (2).

As a group, generalized malabsorptive diarrheal disorders are frequently idiopathic, and their physiologic basis is poorly understood (3). These limitations serve as the impetus to a general search for the molecular basis of these disorders, thus propelling the use of recently feasible pluripotent and somatic stem cell technologies to discover alternative therapeutic approaches (4).

Received April 13, 2013; accepted August 7, 2013.

From the *Department of Human Genetics, David Geffen School of Medicine, the †Department of Pediatrics, Division of Gastroenterology and Nutrition, Mattel Children's Hospital and David Geffen School of Medicine, University of California, Los Angeles, the ‡Department of Anatomy and Neurobiology, University of Maryland-Baltimore County, Baltimore, MD, the ¶Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, the #Department of Pediatrics, Division of Endocrinology, Mattel Children's Hospital and David Geffen School of Medicine, University of California, Los Angeles, the **Department of Pediatrics, University of Nevada School of Medicine, Las Vegas, and the ‡‡Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles.

Address correspondence and reprint requests to Martín G. Martín, MD, MPP, Department of Pediatrics, Division of Gastroenterology and Nutrition, Mattel Children's Hospital and the David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095 (e-mail: mgmartin@mednet.ucla.edu).

Supplemental digital content is available for this article. Direct URL citations appear in the printed text, and links to the digital files are provided in the HTML text of this article on the journal's Web site *(www.jpgn.org)*.

This work was supported by grants from the National Institute of Diabetes and Digestive and Kidney Diseases (#DK083762), and the California Institute of Regenerative Medicine RT2-01985, to M.G.M. and DA05084 to I.L.

The authors report no conflicts of interest.

DOI: 10.1097/MPG.0b013e3182a8ae6c

In some cases, the histology and nutrient absorption characteristics may point to the possibility of dysfunction in known genes (*Neurogenin-3* [*NEUROG3*], *SGLT1*, *EPCAM*, *MYO5B*, *SPINT2*, *TTC37*, *SKIV2L*, *ADAM17*), which can then be directly sequenced to identify likely causative mutations (5–11). Often, however, clinical evidence may be insufficient to implicate known genes or sequencing of candidate genes fails to reveal damaging mutations resulting from genetic heterogeneity. These aspects, as is typical of all rare disorders, greatly impede efficient and timely diagnosis. Recent advances in sequencing technology now make it possible to sequence the coding portion and essential splice sites of approximately 95% of all protein-coding bases of all known genes (the ''exome'') at a cost comparable to clinical sequencing tests of a single gene (12,13). Thus, an unbiased scan of the exome can discover known and novel mutations in known genes and also mutations in hitherto unsuspected genes in a manner that efficiently directs clinical care. Here we show an example of the use of whole-exome sequencing to identify the causative mutation in a child with congential diarrhea.

Prohormone convertase 1/3 (PC1/3) is a calcium-dependent serine endoprotease that converts proinsulin and other prohormones into active forms (14). PC1/3 is highly expressed in the small intestine. PC1/3 deficiency, resulting from the mutations in the *proprotein convertase subtilisin/kexin type 1* (*PCSK1*) gene, can prevent enteroendocrine cells from producing functional hormones and cause generalized malabsorption and a variety of systemic endocrinopathies that develop in an age-dependent fashion (15). The mechanism by which PC1/3 deficiency causes malabsorption is not well understood, but it may be that a novel peptide, or multiple redundant peptides, processed by PC1/3 enhance nutrient assimilation.

## METHODS

### DNA Sequencing

We prepared an exon-enriched sequencing library following Agilent Technologies (Santa Clara, CA) SureSelect Target Enrichment System for the Applied Biosystems (Foster City, CA) SOLiD System protocol (version 1.5.1). Briefly, genomic DNA extracted from patient saliva with an Oragene DNA Collection kit (DNA-Genotek, Kanata, Canada) was ultrasonically sheared (Covaris, Woburn, MA) into ~125 bp fragments. After fragment end repair, ligation of adapters, and gel-size selection for ~175 bp product, the library was nick translated and amplified by 12 polymerase chain reaction (PCR) cycles. The library was hybridized in solution to RNA probes from the 3-Mb SureSelect All Human Exon kit (Agilent G3361), covering 1.22% of the human genome containing the exons of the Consensus CDS genes (16). The exon-enriched library was selected by magnetic bead separation, further amplified by 12 cycles of PCR, and clonally amplified on beads by emulsion PCR (Applied Biosystems SOLiD 4 System Templated Bead Preparation Guide [March 2010]). Fragment sequencing by ligation was performed on a SOLiD 4 System (Applied Biosystems SOLiD 4 System Instrument Operation Guide [March 2010]), which yielded ~108 million 50-base reads.

### Bioinformatics Analysis

We aligned the sequenced reads to build GRCh37 of the human genome (17) with Novoalign (*http://www.novocraft.com*) to obtain ~31 million uniquely aligned 50-base reads (~1.5G bases) after removing PCR duplicates. ~1.2G bases were within the targeted exome with a mean coverage of ~33X at each targeted position, achieving at least 10X coverage of 67% of the coding sequence of annotated protein-coding genes per Ensembl (Wellcome Trust Sanger Institute, Hinxton, UK) (18). Base quality scores were recalibrated to improve accuracy by analyzing the covariation among reported quality score, position within read, dinucleotide, and probability of mismatching the reference genome using the Genome Analysis Toolkit (GATK) (19,20). The GATK Unified Genotyper was used to genotype single-nucleotide variants (SNVs) and indels. The GATK variant quality score recalibration module was used to assign probabilities to each variant call. In addition to variants that passed quality checks, for improved sensitivity, we retained for downstream analysis a tranche of variants that had a likelihood $\leq$1% of being false-positives. Variant consequences were determined by the Ensembl Variant Effect Predictor (20), and the extent of protein damage was estimated with SIFT (Sorting Intolerant From Tolerant) (21–25), PolyPhen (Polymorphism Phenotyping) (26–28), and Condel (Consensus Deleteriousness) (29).

### Development of Wild-Type and Mutant PC1/3 Expression Vectors

Human PC1/3 complement DNA (cDNA) was generated from RNA isolated from human pancreatic carcinoid (BON) cells. The wild-type PC1/3 was amplified by using an oligonucleotide that contained a *Kpn*I site, and 20 nucleotides of PC1/3 from the translational start site. A 3′ oligonucleotide contained an *Xho*I site, and a FLAG amino acid (DYKDDDDK) sequence introduced at the stop site of the wild-type and Y343X mutant cDNA. The wild-type cDNA was subcloned into the *Kpn*I and *Xho*I restriction-digested pcDNA3 vector and clones were screened. A single clone, designated WT PC1/3, was used to generate the Y343X mutant clone, and the entire clone was sequenced to verify that only the targeted variant was altered.

### Transient Transfection of Expression Vectors

HEK293 cells at a density of $2 \times 10^5$ cells per well in 24-well plates were transfected with plasmids encoding WT PC1/3 or Y343X in triplicate wells. Cells were transfected with 200 ng of plasmid DNA per well using Lipofectamine Reagent (Invitrogen, Carlsbad, CA). Five hours posttransfection, 1 mL of growth medium was added to each well and incubation continued for an additional 24 hours. Postincubation, cells were washed with phosphate-buffered saline and 0.3 mL of Opti-MEM (Invitrogen) containing 100-μg/mL bovine aprotinin was added to each well. Cells were incubated for an additional 24 hours before conditioned medium and cells were harvested. Conditioned medium was analyzed first by enzyme assay; both cells and medium were then subjected to sodium dodecyl sulfate-polyacrylamide gel electrophoresis followed by Western blotting. Expression in both types of samples was assessed using primary antiserum against the amino terminus of mature PC1/3 followed by horseradish peroxidase–coupled secondary antiserum (30). Immunoreactive protein visualization was accomplished using the SuperSignal West Femto Maximum Sensitivity Substrate kit (Thermo Scientific, Rockford, IL).

### Enzyme Assay

Enzymatic activity of secreted recombinant PC1/3 proteins present in conditioned medium obtained from transiently transfected HEK293 cells was measured in triplicate 50-μL reactions in a 96-well polypropylene plate containing 25 μL of conditioned medium, 200 μmol/L substrate (pyr-RTKR-amc [7-amino-4-methylcoumarin]), 100 mmol/L of sodium acetate, pH 5.5, 2 mmol/L of $CaCl_2$, 0.1% of Brij 35, and a protease inhibitor cocktail (final concentrations 1 μmol/L of pepstatin, 0.28 mmol/L of tosyl phenylalanyl chloromethylketone, 10 μmol/L of E-64, and 0.14 mmol/L of Nα-tosyl-Lys-chloromethylketone). Reaction mixtures were incubated at 37°C and fluorescence measurements (380 nm excitation, 460 emission) were taken under kinetic conditions every 20 seconds for 1 hour in a Fluoroskan fluorometer (Thermo Scientific). Maximum rates were calculated.

## Histology and Immunohistochemistry

Small bowel and colonic mucosal biopsies were stained with hematoxylin and eosin and immunohistochemistry (IHC) with antiserum to chromogranin A (CGA) as previously described (31). Briefly for IHC, sections were cut at 2 to 3 μm, deparaffinized, and endogenous peroxidase activity was quenched using 0.5% hydrogen peroxide. Heat-induced epitope retrieval was performed. Slides were placed on a Dako Autostainer and then incubated sequentially in primary anti-CGA (DakoCytomation, Campbellfield, Australia) for 30 minutes in rabbit anti-mouse immunoglobulin, followed by Envision+ (DakoCytomation). Diaminobenzidine and hydrogen peroxide were used as the substrates for the peroxidase enzyme. Similarly, PCSK1 staining was performed manually using a human anti-PCSK1 mouse monoclonal antibody (Novus/Biologicals, Littleton, CO); this antiserum is directed against the C-terminus and would be expected to be lost in a C-terminal truncation mutant. Antigen retrieval was performed in citrate buffer pH 6.0 at 95°C for 30 minutes in a steamer. Tissue was stained with the primary antibody (1:250) at room temperature for 45 minutes. Secondary antibody was used as obtained directly from the company and samples incubated at room temperature for 45 minutes (DakoCytomation). Diaminobenzidine incubation time was 10 minutes at room temperature.

# RESULTS

## Clinical History Before Exome Sequencing

The patient was initially assessed at 3 weeks of age for recurrent diarrhea and associated metabolic acidosis. He was a 41-weeks' gestational age male infant born to a 17-year-old gravida 1 female and a biological father said to be the mother's father's first cousin. There was no history of substance abuse, prenatal infections, or other complications during the pregnancy. The baby was born at the appropriate size for gestational age with normal Apgar scores. At 6 days of age, he was transferred to the intensive care nursery because of poor peripheral perfusion and indirect hyperbilirubinemia, with an initial bicarbonate level of 8 and an anion gap of 13. Other liver enzymes (aspartate aminotransferase, alanine transaminase, and alkaline phosphatase) were normal, and urine was trace positive for reducing substances. The infant was treated for possible sepsis with antibiotics. Blood and urine cultures were subsequently negative.

Because of the initial presentation, metabolic laboratory values were sent twice. These included urine organic acids, serum amino acids, acylcarnitine profile, lactate and pyruvate as well as serum ammonia levels. These tests were normal. He was treated in the interim with carnitine supplementation. Two newborn state metabolic screens were normal for congenital adrenal hyperplasia, hypothyroidism, and disorders of amino acid, organic acid, and fatty acid oxidation. When feedings were stopped, the diarrhea ceased, but feeds were subsequently resumed using a standard milk protein–based formula.

Because of the extremely early onset of the diarrhea, congenital disorders were considered and stool was evaluated for reducing substances, pH, qualitative fat, and elastase-1 level, as well as for white and red blood cells. These were normal or negative. DNA genotyping of *CFTR* for cystic fibrosis was negative for the 97 mutations most commonly observed. Serum immunoreactive trypsinogen, serum $\alpha_1$-antitrypsin levels, phenotyping, and an ultrasound of the liver and gallbladder were normal.

Nephrology consultation for possible renal tubular acidosis resulted in brief treatments with oral bicarbonate replacement, but this was stopped when it was believed that no renal tubular issue was present. The patient received amoxicillin prophylaxis for 1 urinary tract infection during his neonatal intensive care unit course and grade 2 bilateral vesicoureteral reflux on a voiding cystourethrogram study.

Both upper and lower gastrointestinal (GI) endoscopies revealed no gross or microscopic abnormalities in duodenal, gastric, and colonic biopsies, and electron microscopy of small bowel biopsies showed no ultrastructural abnormalities. Disaccharidase levels were normal (more detailed information can be found at *http://links.lww.com/MPG/A254* and *http://links.lww.com/MPG/A255*).

The patient had several bouts of acute-onset acidosis requiring several boluses of sodium bicarbonate and fluids. He was discharged home at 3 months of age. Upon discharge, he was placed on Elecare (amino acid–based formula, Abbott Nutrition) and gaining weight adequately despite multiple interruptions of his feeding schedule for intolerance and diarrhea and multiple stool tests were positive for *Clostridium difficile* toxin. Even on a casein-hydrolyzed formula, he had gross blood in his stool, which dissipated on Elecare. His medications upon discharge included amoxicillin for urinary tract infection prophylaxis, multivitamin with iron, and metronidazole to complete a course for the stool *C difficile*.

Five weeks after discharge, he presented to a different hospital with a reported 3-day history of diarrhea and was found to be in hypovolemic shock with profound metabolic acidosis and an initial bicarbonate level of only 4.1. His serum sodium level was 163, and his chloride was 138. Before hydration, it was noted that he had decreased 420 g in weight from his neonatal intensive care unit discharge weight of 4620 g. He was transferred to another children's medical center for admission, had his first central venous (Broviac) catheter placed, and started receiving total parenteral nutrition. At 6 months of age, he had significant failure to thrive with length <5%, and a weight of 5.1 kg ($Z$ −3.75). During this prolonged hospitalization, he was transferred to UCLA Medical Center for 1 month for additional evaluation and was returned to the transferring facility, where he remained hospitalized for an additional 3 weeks. Among various tests that were performed, the serum pancreatic polypeptide level was extremely elevated (>1600 pg/mL, normal <519) and serotonin was low (34 ng/mL, normal range 50–220); however, serum substance P (540 pg/mL, normal <1780), chromogranin A (26.2 ng/mL, normal <36.4), and vasoactive intestinal peptide (28.6 pg/mL, normal <50) levels were normal for age. A proinsulin level was, unfortunately, not obtained at that time.

He was readmitted to local community hospitals 19 times during the subsequent 31 months. Nine of these were emergency department visits, 7 were inpatient stays, and 3 were simply outpatient contacts for testing. He was subsequently placed into foster care because it was believed that many of his admissions were because of inadequate care of his central venous line by his biological parents or because of lack of appropriate outpatient follow-up.

He subsequently had multiple problems with central venous catheter occlusions and was diagnosed as having heparin-induced thrombocytopenia and later as having a plasminogen inhibitor deficiency, which were believed to result in multiple deep venous thrombi, for which he was treated with enoxaparin and later warfarin. Given these significant thrombotic events, his central venous catheter was removed and a percutaneous gastrostomy tube was placed to aid in the transition from parenteral to enteral nutrition support. Repeat upper endoscopy at that time revealed mild chronic gastritis and lactase deficiency on tissue analysis for disaccharidase levels (lactase activity 1.4, normal 24.5 ± 8.0). An upper GI and small bowel follow-through x-ray were normal, including normal transit time.

There was also an admission for pneumonia and respiratory distress. During that admission, he was noted to exhibit excessive thirst and hyperglycemia, with glucose levels running in the high 100s. He was hypokalemic and acidotic, requiring intravenous bicarbonate infusions as well as baking soda enterally. He also had evidence of left ventricular dysfunction requiring Lasix, enalapril, Aldactone, and potassium supplementation. As part of his evaluation for heart failure, fluorescence in situ hybridization studies for Williams syndrome were negative.

## Sequencing/Bioinformatics Results

The initial dataset contained 21,804 nonreference variants (20,129 SNVs) and 1675 small insertions/deletions [indels]) (Table 1). In addition, another 17,172 SNVs were in the 1% false-positive tranche, meaning there was a 1% likelihood that the actual genotype at a given locus was wild-type. This unusually large number of false-positive tranche alleles appeared to be an artifact of the SOLiD platform's quality score assignment algorithm as well as the fact that we had only a small number (n = 7) of sequencing experiments from the same platform available for analysis by the GATK Unified Genotyper's Gaussian mixture model. We used custom data analysis software, based in part on the Ensembl Variant Effect Predictor (20) and next-generation sequencing-single nucleotide polymorphism (32) and implemented on a Microsoft SQL Server database system, to identify potentially causative alleles. We limited the search to variants within the coding region and flanking intronic essential splice site of protein-coding genes in the Ensembl dataset. Under the hypothesis that the disorder was rare, and therefore the causative allele(s) would not be common, we filtered out variants that were in the dbSNP (33), 1000 Genomes (34), National Heart, Lung, and Blood Institute (NHLBI) (35), or National Institute of Environmental Health Sciences (NIEHS) (36) datasets with an allele frequency ≥0.01 in any population. We also excluded variants observed in 74 locally sequenced exomes from unrelated individuals. Six of these exomes were sequenced on the SOLiD platform and were particularly useful to remove systemic bias and false-positives. We thus reduced the number of variants to 1043 SNVs and 11 indels that were sufficiently rare in multiple populations to be consistent with a rare disorder. Analysis of the consequences on protein-coding transcripts for significant adverse effects (nonsense, missense, or essential splice site mutations) reduced the candidate variant list to 467 SNVs and 3 indels.

Under a recessive model, we searched the autosomes and sex chromosomes in this set of variants for homozygous and compound heterozygous mutations (the latter defined as 2 variants in the same transcript). We required that the variants be likely to have a deleterious effect on protein structure as predicted by at least 1 of SIFT (21–25), PolyPhen (26–28), or Condel (29)). Initially, we limited the search to variants that fully passed all of the data quality filters, but this resulted in no homozygous variants being selected, a surprising finding given the stated consanguineous parental relationship. By including the 1% false-positive tranche homozygous variants, we identified a single homozygous variant in *PCSK1* and 6 compound heterozygous variants in 3 other genes (Table 2). The *PCSK1* variant was within a 7.5-Mb homozygous interval, identified by 49 polymorphous dbSNP markers, consistent with inbreeding.

The Tyr343X mutation in *PCSK1* was highlighted by the fact that the protein would be truncated by a premature stop codon within its catalytic domain and by *PCSK1*'s bioinformatics annotation as the only 1 of the 4 genes having a known association with human disorders.

Defects in *PCSK1* are the cause of PC1/3 deficiency (MIM:600955), which had previously been identified in 3 subjects and is characterized by obesity, hypogonadism, hypoadrenalism, and reactive hypoglycemia, as well as significant small-intestinal absorptive dysfunction (37–40) (Fig. 1C). Given the subject's history of diarrhea, we confirmed the presence of the mutation in 71 sequenced fragments (Fig. 1A) and by Sanger sequencing (Fig. 1B), and assessed further whether the mutation of the *PCSK1* gene could alter the protein function and account for the subject's medical problems.

## Functional Analysis and In Vitro Assessment

The Y343X mutation eliminates the final 410 amino acids of the protein, which includes the entire C-terminal and

TABLE 1. Exome sequencing statistics

| | Called | | | Filtered | | | Raw | | |
|---|---|---|---|---|---|---|---|---|---|
| | All | Known | Novel | All | Known | Novel | All | Known | Novel |
| Called loci | 10,2961 | 85,413 | 17,548 | 82,876 | 19,608 | 63,268 | 18,5837 | 10,5021 | 80,816 |
| Ref loci | 81,157 | 66,133 | 15,024 | 61,104 | 16,843 | 44,261 | 14,2261 | 82,976 | 59,285 |
| Variant loci | 21,804 | 19,280 | 2524 | 21,772 | 2765 | 19,007 | 43,576 | 22,045 | 21,531 |
| SNVs | 20,129 | 18,291 | 1838 | 21,095 | 2639 | 18,456 | 41,224 | 20,930 | 20,294 |
| Insertions | 775 | 465 | 310 | 148 | 30 | 118 | 923 | 495 | 428 |
| Deletions | 900 | 524 | 376 | 529 | 96 | 433 | 1429 | 620 | 809 |
| Hets | 13,172 | 10,863 | 2309 | 21,330 | 2389 | 18,941 | 34,502 | 13,252 | 21,250 |
| Hom ref | 57,506 | 46,193 | 11,313 | 42,576 | 8661 | 33,915 | 10,0082 | 54,854 | 45,228 |
| Hom var | 8632 | 8417 | 215 | 442 | 376 | 66 | 9074 | 8793 | 281 |
| Het/hom ratio | 1.53 | 1.29 | 10.74 | 48.26 | 6.35 | 286.98 | 3.8 | 1.51 | 75.62 |
| Ti/Tv ratio | 2.64 | 3.11 | 0.69 | 0.54 | 1.27 | 0.48 | 1.15 | 2.73 | 0.49 |

Metrics for variants identified by exome sequencing. Called = variants that passed all quality control filters; Filtered = variants that did not pass all QC, including variants with a 1% likelihood that the actual genotype is wild-type; Hets = heterozygous genotypes; Hom var = homozygous nonreference genotypes; Hom ref = homozygous reference genotypes (wt); Known/novel = variants in/not in dbSNP137; Ref loci = loci that match GRCh37 reference genome; SNVs = single nucleotide polymorphisms; Ti/Tv ratio = ratio of transition (purine to purine or pyrimidine to pyrimidine) to transversion (purine to/from pyrimidine) variants; Variant loci = loci that differ from the reference.

TABLE 2. Genes with homozygous or compound heterozygous variants likely to affect protein

| Model | Gene | Variant_human_g1k_v37 | Variant_cDNA | Variant_protein | OMIM_disorder |
|---|---|---|---|---|---|
| Hom | *PCSK1* | chr5:95746544G>C | ENST00000311106.3:c.1029C>G | ENSP00000308024.2:p.Tyr343X | 600955: Obesity with impaired prohormone processing |
| 2-Het | *NOL9* | chr1:6601892A>G | ENST00000377705.5:c.1073T>C | ENSP00000366934.5:p.Leu358Pro | |
| | | chr1:6601893G>C | ENST00000377705.5:c.1072C>G | ENSP00000366934.5:p.Leu358Val | |
| 2-Het | *BAMBI* | chr10:28971098A>C | ENST00000375533.3:c.551A>C | ENSP00000364683.3:p.Gln184Pro | |
| | | chr10:28971100G>T | ENST00000375533.3:c.553G>T | ENSP00000364683.3:p.Asp185Tyr | |
| 2-Het | *SPTBN4* | chr19:40996047T>A | ENST00000352632.2:c.387T>A | ENSP00000263373.2:p.Phe129Leu | |
| | | chr19:41025432G>A | ENST00000352632.2:c.3028G>A | ENSP00000263373.2:p.Val1010Met | |

Seven rare (allele frequency ≤0.01) homozygous (hom) or compound heterozygous (2-het) mutations were found in the patient and absent from 74 control exomes.

P domains and a portion of the catalytic domain (41). Loss of even 1 residue of the P domain is known to block enzyme expression (42). In vitro assessment allows for confirmation of the deleterious effect of the Y343X mutation on PC1/3 catalytic activity (30). The nonsense mutation rendered the *Y343X* gene product undetectable in either cells or media, most likely because of rapid intracellular degradation (Fig. 2A). As expected, the absence of detectable Y343X PC1/3 protein in the conditioned medium resulted in a total lack of enzyme activity (Fig. 2B).

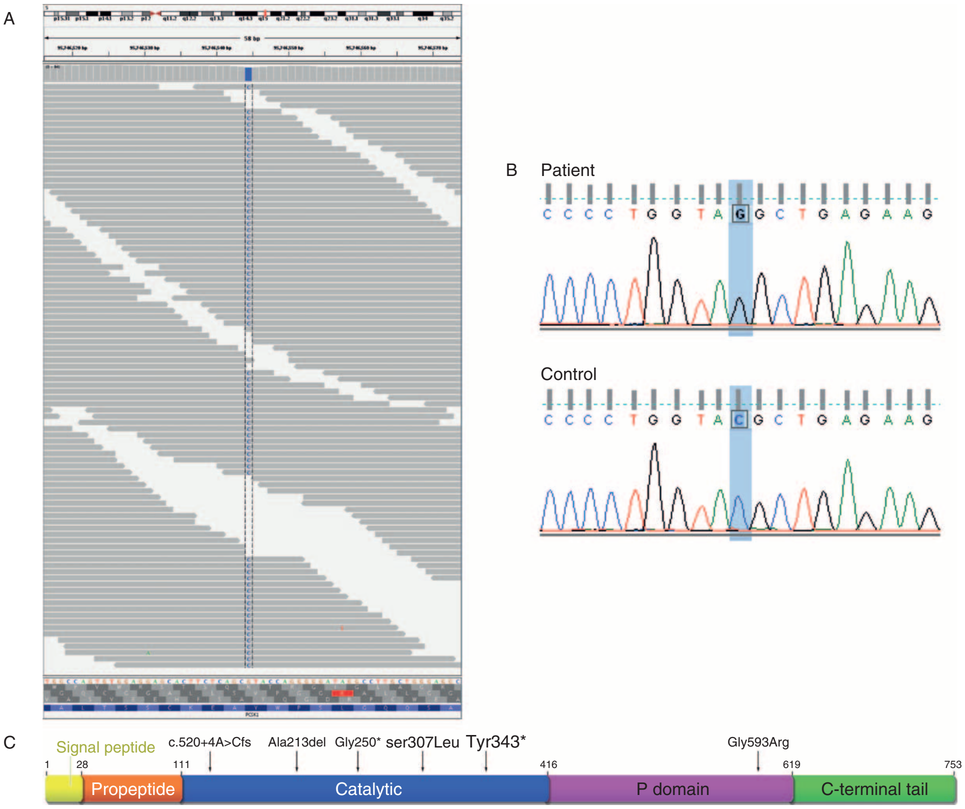


**FIGURE 1.** Exome sequencing results. A, Aligned pileup of sequenced fragments at chr5:95746544 with 30 forward strand reads and 41 reverse strand reads of the C variant and 2 low-quality reads of the reference G allele on fragment 3′ ends. Visualized by the Integrative Genomics Viewer (39). B, Sanger sequencing validation of variant. C, Structure of PC1/3 showing locations of previously identified mutations and the novel Y343X. Adapted from (38) and (40).

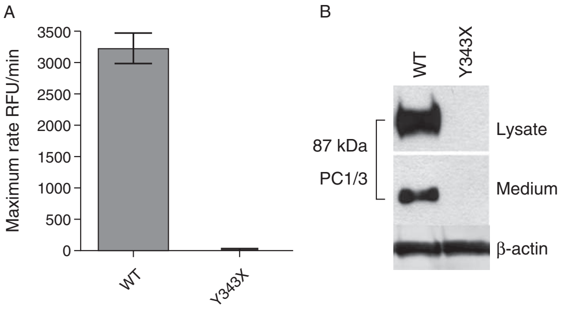


**FIGURE 2.** Functional characterization and visualization of wild-type (WT) prohormone convertase 1/3 (PC1/3) complementary DNA and PC1/3 containing the Y343X nonsense mutation. HEK293 cells were transiently transfected with empty pcDNA3 (not shown), WT PC1/3, and Y343X PC1/3. A, Enzymatic activity of secreted recombinant PC1/3 proteins in conditioned medium was compared by measuring maximum cleavage rates using the fluorogenic substrate pyr-RTKR-amc during a 1-hour kinetic assay. Three replicates per condition were assayed in triplicate and are shown as the mean ± standard deviation, $P < 0.0017$ (2-tailed). B, Western blot of cell lysates and media from transfected HEK293 cells, using amino terminal–directed PC1/3 primary antiserum for detection of recombinant PC1/3 proteins. The data shown represent 1 of 3 independent experiments. β-actin shows equivalent loading of cell extracts.

## Histologic Assessment

The small and large bowel mucosa from the subject was histologically normal in all respects except for the loss of PC1/3-positive enteroendocrine cells (Fig. 3). By hematoxylin and eosin staining, the architecture, immune cell complement, and epithelium were indistinguishable from normal mucosa (Fig. 3A). By IHC, the appearance and number of chromogranin A–positive enteroendocrine cells were normal (Fig. 3B); however, IHC for PC1/3-expressing enteroendocrine cells was negative relative to wild-type controls in the small (Fig. 3C, D) and large bowel (data not shown). PC1/3 IHC should decorate a subset of enteroendocrine cells in colonic and small bowel mucosa.

## Clinical History Following Genetic Testing

A more comprehensive and focused evaluation was prompted by identification of the genetic mutation. The adoptive parents were contacted to obtain an update of the subject's clinical status to determine whether known clinical manifestation of PC1/3 deficiency described in the other probands was observed in this child (15). The subject had been managed at several local community facilities and was not seen at the major referral institution for >4 years. The family reported evidence suggestive of polydipsia, polyuria, enuresis, and polyphagia, but he had not been evaluated for diabetes insipidus. The subject was urgently assessed locally and found to have an undetectable serum vasopressin level. Intranasal desmopressin (DDAVP) improved significantly the severity of the polydipsia and enuretic episodes.

An endocrine analysis was performed at the referral center when the subject was 4 years, 5 months old. Before starting DDAVP, he had significant polyuria and polydipsia, drinking approximately 4 L of water per day. When parents restricted water, he would drink from the toilet, fish tank, or outdoor faucet. He experienced temper tantrums when water intake was limited. He also had evidence of severe pica and attempted to eat intravenous tubing, wood from his crib, and paper products. He was toilet trained, but still went through 1 bag of diapers per day. His appetite was described as excessive, eating more than his teenage siblings combined, and his parents keeping the refrigerator and pantry locked. His examination was otherwise normal, with the exception of a small penis: approximately 2.8 cm in stretched length (mean 5.7 cm, Z of −2.5 is 3.5 cm). Blood tests, including insulin, were generally normal except for dramatically elevated proinsulin, slightly elevated thyroid-stimulating hormone, and low serum insulin-like growth factor 1 and insulin-like growth factor–binding protein-3 (more detailed information can be found at *http://links.lww.com/MPG/A254* and *http://links.lww.com/MPG/A255*).

## DISCUSSION

Our laboratory began a systematic assessment of several distinct kindreds with various putatively genetic forms of congenital diarrhea, using exome sequencing, with the hope of identifying the molecular basis of the disorders. This article describes our analysis of a single patient, and nicely illustrates how this research technique was used to understand a challenging patient with an inherited disorder and multiple medical problems, and led to a more definitive phenotypic workup that greatly altered clinical management.

Our finding in this single case confirms the value and efficiency of exome sequencing as a primary diagnostic mode to identify mutations of genes associated with rare clinical conditions. This case pointedly illustrates how multiple hospitalizations and a barrage of various indirect, redundant, and expensive tests are frequently required—and sometimes fail—to establish medical diagnoses. Whole-exome sequencing is a transformative technology that should alter the clinician's approach to the evaluation of such patients (43). It is conceivable that portions of the standard metabolic panels and other urine, blood, and radiographic tests will be used less frequently once exome sequencing becomes fully implemented into clinical practice. More directed phenotypic evaluation will be possible rather than the shotgun approaches typically used in children with rare genetic conditions. This phenotypic evaluation is called a ''diagnostic odyssey'' and is often years in duration, to the disadvantage of the patient and at great cost to family and society.

In a relatively short period of time, high-throughput sequencing has solved the mystery of numerous novel inherited disorders and has been used to identify common variants associated with various primary tumors (44,45). As costs continue to decline, this

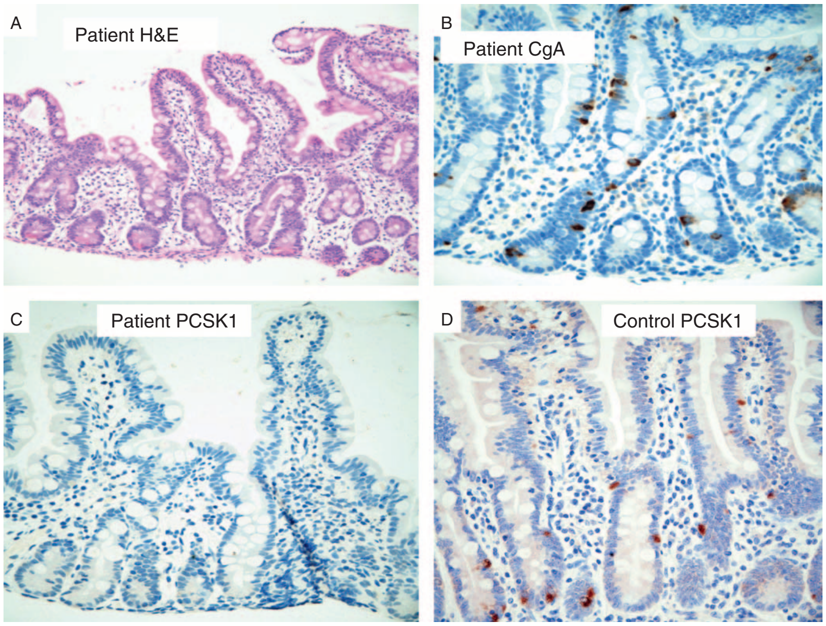


**FIGURE 3.** Absence of prohormone convertase 1/3 (PC1/3)-positive enteroendocrine cells in small bowel mucosa. A, Hematoxylin and eosin (H&E)–stained proprotein convertase subtilisin/kexin type 1 mutant showing normal morphology; original magnification ×200. B, Chromogranin A immunohistochemistry of mutant tissue showing a normal pattern of enteroendocrine cells; original magnification ×400. C, PC1/3 immunohistochemistry showing complete absence of PC1/3-positive enteroendocrine cells in the mutant tissue; original magnification ×400. D, PC1/3 immunohistochemistry showing normal PC1/3 enteroendocrine cells in a healthy control; original magnification ×400.

technology promises to identify the genetic basis of numerous clinical diagnoses, especially those, such as congenital diarrhea, that can be causally related to a large number of possible genes, thereby bypassing the traditional approach of targeted candidate gene sequencing (46).

Given the possibility of inbreeding, we considered the approach of using a high-density microarray to locate regions of homozygosity in the patient's genome, then developing a set of custom-capture probes to select that region for deep sequencing, or to sequence all of the exons in the region with the traditional Sanger method. We ultimately decided on whole-exome sequencing as being more cost-effective. Furthermore, without knowledge that other family members were affected, we did not have a high previous likelihood that the mutation would be in a large region of homozygosity that could be detected by a microarray. Even though we may hypothesize that this child's disease was caused by a homozygous mutation, there was essentially no additional cost or time needed to analyze the data for both homozygous and compound heterozygous mutations. Finally, we wished to develop an unbiased analytical pipeline that could be used for many other genetic patterns.

Others have reported on the use of next-generation sequencing to identify mutations in a gene (*SLC26A3*) known to be associated with chronic diarrhea; however, unlike our approach, sequencing was limited to regions that were homozygous by consanguineous descent (46a). The general efficiency of whole-exome sequencing favors the generation of whole-exome data on virtually all such patients as the most efficient and thorough means of genetic evaluation.

Severe mutations of the *PCSK1* gene are rare. The Y343X mutation is novel and was not identified in the 7252 chromosomes that comprise the publicly available 1000 Genomes, NHLBI, and NIEHS datasets, suggesting an allele frequency of $<0.00014$ and an inferred incidence of homozygous individuals in the population of $<1$ in 52 million, which is consistent with the rarity of PC1/3 deficiency disorders. This mutation broadens the phenotypic consequences of PC1/3 deficiency disorders. The first *PCSK1*-mutant proband was a middle-aged woman who was evaluated for postprandial hypoglycemia and was found to have obesity, hypogonadotropic hypogonadism, hypoadrenalism, and elevated proinsulin levels (47,48). A second case of PC1/3 deficiency was established in an infant with generalized malabsorptive diarrhea who expired at 18 months of age (37). A third proband was described as a 6-year-old boy with a diarrheal condition that resembled the previous case, and intestinal biopsies from both children were described as a persistent enteropathy with patchy villous atrophy (38). In contrast, the subject reported here had no histopathological evidence of enteropathy and had perfectly normal crypt-villus axis without a pathologic inflammatory component (Fig. 3). We have recently presented findings on additional cases with *PCSK1* mutations that we identified by Sanger sequencing (15). Retrospective questioning of the primary proband confirmed similar diarrheal symptoms that were certainly worse during early childhood (37). In our patient, as in the other 3 patients, the proinsulin level was significantly elevated. These data would suggest that serum proinsulin levels and sequencing of the *PCSK1* gene could be used to establish the diagnosis; however, given the breadth of phenotypic presentations, molecular diagnosis is likely to remain challenging without implementation of broader approaches such as whole-exome sequencing.

To develop a sense of the mutational load of rare variants on *PCSK1* in the whole population, we examined all 913 *PCSK1* variants reported in dbSNP and the 1000 Genomes, NHLBI, and NIEHS datasets, and found 47 variants causing nonsynonymous codons or more serious consequences and having minor allele frequencies $\leq$0.01 (supplementary Table S1, *http://links.lww.com/MPG/A255*). Interestingly, common variants in the coding region of *PCSK1* are also associated with common forms of obesity (N221D, Q665E, S690T) and type 2 diabetes mellitus (Q665E, S690T) (49,50). A recent study suggests that PC1/3 deficiency is dependent on the dosage of *PCSK1*, and rare heterozygous mutations can cause obesity (51). From the population data, rare protein-altering mutations will be homozygous or compound heterozygous, resulting in substantial loss of PC1/3 activity in ~86 individuals per million, and these individuals would be predicted to be at risk for a life-threatening PC1/3 deficiency. Additionally, an estimated 1 in 20 individuals may harbor a modest PC1/3 deficiency, which may contribute to PC1/3 deficiency–related obesity.

Exome sequencing will typically generate upwards of 20,000 candidate variants from the reference genome in a given individual. A challenge for diagnosis by exome sequencing is to filter out variants that cannot possibly cause the disease in question. Alleles, such as synonymous or intronic variants, can be eliminated with slight risk that they are false-negatives, as can alleles that are too frequent in the population to be consistent with the incidence of the disorder. Beyond that, prediction of the functional consequences of novel mutations remains a daunting task.

Enteric anendocrinosis is another inherited intestinal endocrinopathy that has clinical features that resemble the early stages of PC1/3 deficiency, including a generalized form of malabsorption (MIM:610370) (5). Homozygous mutations of *NEUROG3* were described in 3 probands, the intestines of which were devoid of enteroendocrine cells, yet had an otherwise normal-appearing intestine. *NEUROG3* is a basic helix-loop-helix transcriptional factor that is necessary and sufficient to drive endocrine cell development in the pancreas and intestine (5). Although 2 of the 3 cases in the initial report did not develop insulin-dependent diabetes mellitus until preadolescent age, 2 recent cases describe the onset of diabetes during the neonatal period (52). Unlike patients with PC1/3 deficiency, children with enteric anendocrinosis do not appear to develop hypothalamic, pituitary, adrenal, thyroid, or gonadal insufficiencies.

Establishing the precise diagnosis of a congenital diarrheal condition requires an intestinal biopsy and a thoughtful approach to dietary challenges. A differential diagnosis of this condition, when presenting with seemingly histologically normal intestinal mucosa, is mostly limited to specific defects of nutrient assimilation (digestive enzymes or transport proteins) or enteroendocrinopathies. Enteroendocrinopathies are histologically subtle, are generally only discovered when specifically sought, and require immunohistochemical confirmation (31). Exome sequencing will certainly be used in the coming years to identify the inherited basis of novel diarrheal disorders and will likely be the standard of practice for genotype testing of established disorders. In our case, exome sequencing provided a diagnosis that resulted in immediate changes in patient care and an improved ability to predict clinical progression, based on previous cases of PC1/3 deficiency.

## REFERENCES

1. Martin MG, Wright EM. Congenital intestinal transport defects. In: Walker WA, Goulet O, Kliegman RM, eds. *Pediatric Gastrointestinal Disease.* 4th ed. Hamilton, Canada: BC Decker; 2004:898–921.
2. Fishbein TM. Intestinal transplantation. *N Engl J Med* 2009;361:998–1008.
3. Binder HJ. Causes of chronic diarrhea. *N Engl J Med* 2006;355:236–9.
4. Spence JR, Mayhew CN, Rankin SA, et al. Directed differentiation of human pluripotent stem cells into intestinal tissue in vitro. *Nature* 2011;470:105–9.
5. Wang J, Cortina G, Wu SV, et al. Mutant neurogenin-3 in congenital malabsorptive diarrhea. *N Engl J Med* 2006;355:270–80.
6. Martin MG, Turk E, Lostao MP, et al. Defects in Na+ glucose cotransporter (SGLT1) trafficking and function cause glucose-galactose malabsorption. *Nat Genet* 1996;12:216–20.
7. Sivagnanam M, Mueller JL, Lee H, et al. Identification of EpCAM as the gene for congenital tufting enteropathy. *Gastroenterology* 2008;135: 429–37.
8. Muller T, Hess MW, Schiefermeier N, et al. MYO5B mutations cause microvillus inclusion disease and disrupt epithelial cell polarity. *Nat Genet* 2008;40:1163–5.
9. Hartley JL, Zachos NC, Dawood B, et al. Mutations in TTC37 cause trichohepatoenteric syndrome (phenotypic diarrhea of infancy). *Gastroenterology* 2010;138:2388–982398 e1–2.
10. Heinz-Erian P, Muller T, Krabichler B, et al. Mutations in SPINT2 cause a syndromic form of congenital sodium diarrhea. *Am J Hum Genet* 2009;84:188–96.
11. Blaydon DC, Biancheri P, Di WL, et al. Inflammatory skin and bowel disease linked to ADAM17 deletion. *N Engl J Med* 2011;365:1502–8.
12. Shendure J. Next-generation human genetics. *Genome Biol* 2011;12: 408.
13. Singleton AB. Exome sequencing: a transformative technology. *Lancet Neurol* 2011;10:942–6.
14. Hoshino A, Lindberg I. Peptide biosynthesis: prohormone convertases 1/3 and 2. In: Fricker LD, Devi L, eds. *Colloquium Series on Neuropeptides.* 1st ed. San Rafael, CA: Morgan & Claypool Life Sciences Publishers; 2004.
15. Martin MG, Lindberg I, Solorzano-Vargas RS, et al. Congenital proprotein convertase 1/3 deficiency causes malabsorptive diarrhea and other endocrinopathies in a pediatric cohort. *Gastroenterology* 2013;145: 138–48.
16. Pruitt KD, Harrow J, Harte RA, et al. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res* 2009;19:1316–23.
17. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* 2004;431:931–45.
18. Flicek P, Amode MR, Barrell D, et al. Ensembl 2011. *Nucleic Acids Res* 2011;39:D800–806.
19. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;43:491–8.
20. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–303.
21. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009;4:1073–81.
22. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. *Genome Res* 2001;11:863–74.
23. Ng PC, Henikoff S. Accounting for human polymorphisms predicted to affect protein function. *Genome Res* 2002;12:436–46.
24. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* 2003;31:3812–4.
25. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet* 2006;7:61–80.
26. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* 2002;30:3894–900.
27. Sunyaev S, Ramensky V, Bork P. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet* 2000;16:198–200.
28. Sunyaev S, Ramensky V, Koch I, et al. Prediction of deleterious human alleles. *Hum Mol Genet* 2001;10:591–7.
29. Gonzalez-Perez A, Lopez-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet* 2011;88:440–9.

30. Vindrola O, Lindberg I. Biosynthesis of the prohormone convertase mPC1 in AtT-20 cells. *Mol Endocrinol* 1992;6:1088–94.
31. Cortina G, Smart CN, Farmer DG, et al. Enteroendocrine cell dysgenesis and malabsorption, a histopathologic and immunohistochemical characterization. *Hum Pathol* 2007;38:570–80.
32. Grant JR, Arantes AS, Liao X, et al. In-depth annotation of SNPs arising from resequencing projects using NGS-SNP. *Bioinformatics* 2011;27: 2300–1.
33. National Center for Biotechnology Information NLoM. Database of Single Nucleotide Polymorphisms (dbSNP Build ID: 134). Vol dbSNP134. Bethesda, MD: National Center for Biotechnology Information, National Library of Medicine.
34. The 1000 Genomes Project ConsortiumA map of human genome variation from population-scale sequencing. *Nature* 2010;467:1061–73.
35. NHLBI Exome Sequencing Project. Exome Variant Server. Volume September. Seattle, WA: NHLBI Exome Sequencing Project; 2011.
36. NIEHS Environmental Genome Project. NIEHS Exome Variant Server. Volume September. Seattle, WA: NIEHS Environmental Genome Project; 2011.
37. Jackson RS. Small-intestinal dysfunction accompanies the complex endocrinopathy of human proprotein convertase 1 deficiency. *J Clin Invest* 2003;112:1550–60.
38. Farooqi IS, Volders K, Stanhope R, et al. Hyperphagia and early-onset obesity due to a novel homozygous missense mutation in prohormone convertase 1/3. *J Clin Endocrinol Metab* 2007;92:3369–73.
39. Robinson JT, Thorvaldsdottir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol* 2011;29:24–6.
40. Steiner DF. The proprotein convertases. *Curr Opin Chem Biol* 1998;2:31–9.
41. Ueda K, Lipkind GM, Zhou A, et al. Mutational analysis of predicted interactions between the catalytic and P domains of prohormone convertase 3 (PC3/PC1). *Proc Natl Acad Sci U S A* 2003;100:5622–7.
42. Zhou A, Martin S, Lipkind G, et al. Regulatory roles of the P domain of the subtilisin-like prohormone convertases. *J Biol Chem* 1998;273: 11107–14.
43. Maxmen A. Exome sequencing deciphers rare diseases. *Cell* 2011;144: 635–7.
44. Ng SB, Buckingham KJ, Lee C, et al. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* 2010;42:30–5.
45. Stransky N, Egloff AM, Tward AD, et al. The mutational landscape of head and neck squamous cell carcinoma. *Science* 2011;333:1157–60.
46. Lifton RP. Individual genomes on the horizon. *N Engl J Med* 2010;362:1235–6.
46a. Choi M, Scholl UI, Ji W, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A* 2009;106:19096–101.
47. Jackson RS, Creemers JW, Ohagi S, et al. Obesity and impaired prohormone processing associated with mutations in the human prohormone convertase 1 gene. *Nat Genet* 1997;16:303–6.
48. O'Rahilly S, Gray H, Humphreys PJ, et al. Brief report: impaired processing of prohormones associated with abnormalities of glucose homeostasis and adrenal function. *N Engl J Med* 1995;333: 1386–90.
49. Benzinou M, Creemers JW, Choquet H, et al. Common nonsynonymous variants in PCSK1 confer risk of obesity. *Nat Genet* 2008;40:943–5.
50. Strawbridge RJ, Dupuis J, Prokopenko I, et al. Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. *Diabetes* 2011;60:2624–34.
51. Creemers JWM, Choquet H, Stijnen P, et al. Heterozygous mutations causing partial prohormone convertase 1 deficiency contribute to human obesity. *Diabetes* 2012;61:383–90.
52. Gradwohl G, Dierich A, LeMeur M, et al. Neurogenin3 is required for the development of the four endocrine cell lineages of the pancreas. *Proc Natl Acad Sci U S A* 2000;97:1607–11.

**Table S1. Potentially harmful mutations in PCSK1 found in public datasets**

| Variant_human_g1k_v37 | ID | Consequence | Variant_cDNA | Variant_protein | SIFT | PolyPhen | Condel | UniProt annotation | Alleles | Alt_alleles | MAF | hom/M | het/M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| chr5:95728710T>G | rs147016634 | non_synonymous_codon | ENST00000311106.3:c.2257A>C | ENSP00000308024.2:p.Asn753His | deleterious | possibly_damaging | deleterious |  | 7026 | 3 | 0.00043 | 0.18 | 854 |
| chr5:95728736G>A | rs138164746 | non_synonymous_codon | ENST00000311106.3:c.2231C>T | ENSP00000308024.2:p.Ala744Val | deleterious | probably_damaging | deleterious |  | 4872 | 1 | 0.00021 | 0.04 | 410 |
| chr5:95728748C>T | rs140941383 | non_synonymous_codon | ENST00000311106.3:c.2219G>A | ENSP00000308024.2:p.Arg740Gln | deleterious | probably_damaging | deleterious |  | 4866 | 1 | 0.00021 | 0.04 | 411 |
| chr5:95728797A>G | rs113376374 | non_synonymous_codon | ENST00000311106.3:c.2170T>C | ENSP00000308024.2:p.Tyr724His | deleterious | probably_damaging | deleterious |  | 0 | 0 | <0.00014 | <0.02 | <276 |
| chr5:95728809G>C | . | non_synonymous_codon | ENST00000311106.3:c.2158C>G | ENSP00000308024.2:p.Leu720Val | tolerated | benign | neutral |  | 4872 | 1 | 0.00021 | 0.04 | 410 |
| chr5:95728862T>G | . | non_synonymous_codon | ENST00000311106.3:c.2105A>C | ENSP00000308024.2:p.Glu702Ala | deleterious | benign | deleterious |  | 2184 | 1 | 0.00046 | 0.21 | 915 |
| chr5:95728863C>T | rs140899352 | non_synonymous_codon | ENST00000311106.3:c.2104G>A | ENSP00000308024.2:p.Glu702Lys | tolerated | benign | deleterious |  | 7060 | 5 | 0.00071 | 0.50 | 1415 |
| chr5:95728877T>G | rs138433207 | non_synonymous_codon | ENST00000311106.3:c.2090A>C | ENSP00000308024.2:p.Tyr697Ser | tolerated | benign | deleterious |  | 4876 | 1 | 0.00021 | 0.04 | 410 |
| chr5:95728982G>T | rs148807505 | non_synonymous_codon | ENST00000311106.3:c.1985C>A | ENSP00000308024.2:p.Ala662Asp | deleterious | benign | deleterious |  | 4872 | 1 | 0.00021 | 0.04 | 410 |
| chr5:95729007G>A | . | non_synonymous_codon | ENST00000311106.3:c.1960C>T | ENSP00000308024.2:p.Arg654Trp | deleterious | possibly_damaging | deleterious |  | 2184 | 1 | 0.00046 | 0.21 | 915 |
| chr5:95729039G>C | rs142453906 | non_synonymous_codon | ENST00000311106.3:c.1928C>G | ENSP00000308024.2:p.Ser643Cys | tolerated | benign | deleterious |  | 4062 | 1 | 0.00025 | 0.06 | 492 |
| chr5:95729048G>A | . | non_synonymous_codon | ENST00000311106.3:c.1919C>T | ENSP00000308024.2:p.Thr640Ile | tolerated | benign | deleterious |  | 2184 | 1 | 0.00046 | 0.21 | 915 |
| chr5:95729049T>C | rs139453594 | non_synonymous_codon | ENST00000311106.3:c.1918A>G | ENSP00000308024.2:p.Thr640Ala | tolerated | benign | neutral |  | 6074 | 11 | 0.00181 | 3.28 | 3615 |
| chr5:95730597C>G | rs144324144 | non_synonymous_codon | ENST00000311106.3:c.1855G>C | ENSP00000308024.2:p.Gly619Arg | tolerated | possibly_damaging | deleterious |  | 4878 | 1 | 0.00021 | 0.04 | 410 |
| chr5:95730719A>G | rs145196120 | non_synonymous_codon | ENST00000311106.3:c.1733T>C | ENSP00000308024.2:p.Ile578Thr | tolerated | benign | deleterious |  | 4878 | 1 | 0.00021 | 0.04 | 410 |
| chr5:95734621C>T | rs149124467 | non_synonymous_codon | ENST00000311106.3:c.1550G>A | ENSP00000308024.2:p.Arg517Gln | tolerated | probably_damaging | deleterious |  | 4876 | 1 | 0.00021 | 0.04 | 410 |
| chr5:95734724C>T | rs137852821 | non_synonymous_codon | ENST00000311106.3:c.1447G>A | ENSP00000308024.2:p.Gly483Arg | tolerated | benign | deleterious | 483:483:G -> R (in PC1 deficiency; prevents processing of pro-PCSK1 and leads to its retention in the endoplasmic reticulum) | 0 | 0 | <0.00014 | <0.02 | <276 |
| chr5:95734741C>A | rs80319394 | splice_acceptor_variant | ENST00000311106.3:c.1431-1G>T | . | . | . | . |  | 0 | 0 | <0.00014 | <0.02 | <276 |
| chr5:95735697T>C | rs112274762 | non_synonymous_codon | ENST00000311106.3:c.1390A>G | ENSP00000308024.2:p.Lys464Glu | deleterious | possibly_damaging | deleterious |  | 0 | 0 | <0.00014 | <0.02 | <276 |
| chr5:95735700C>T | rs143174906 | non_synonymous_codon | ENST00000311106.3:c.1387G>A | ENSP00000308024.2:p.Glu463Lys | tolerated | benign | deleterious |  | 7060 | 7 | 0.00099 | 0.98 | 1981 |
| chr5:95735703G>T | rs151257336 | non_synonymous_codon | ENST00000311106.3:c.1384C>A | ENSP00000308024.2:p.Pro462Thr | deleterious | probably_damaging | deleterious |  | 4876 | 1 | 0.00021 | 0.04 | 410 |
| chr5:95735742G>C | rs140481124 | non_synonymous_codon | ENST00000311106.3:c.1345C>G | ENSP00000308024.2:p.Leu449Val | deleterious | probably_damaging | deleterious |  | 4874 | 1 | 0.00021 | 0.04 | 410 |
| chr5:95746504T>C | rs1050622 | non_synonymous_codon | ENST00000311106.3:c.1069A>G | ENSP00000308024.2:p.Ser357Gly | deleterious | probably_damaging | deleterious |  | 0 | 0 | <0.00014 | <0.02 | <276 |
| chr5:95746543C>T | . | non_synonymous_codon | ENST00000311106.3:c.1030G>A | ENSP00000308024.2:p.Ala344Thr | tolerated | benign | neutral |  | 2184 | 1 | 0.00046 | 0.21 | 915 |
| chr5:95746638C>T | rs138879299 | non_synonymous_codon | ENST00000311106.3:c.935G>A | ENSP00000308024.2:p.Arg312His | deleterious | possibly_damaging | deleterious |  | 4878 | 1 | 0.00021 | 0.04 | 410 |
| chr5:95746653G>A | rs137852824 | non_synonymous_codon | ENST00000311106.3:c.920C>T | ENSP00000308024.2:p.Ser307Leu | deleterious | probably_damaging | deleterious | 307:307:S -> L (in PC1 deficiency; in vitro the mutation markedly impairs the catalytic activity of the enzyme; however intracellular trafficking of this mutant enzyme appears normal; retains some autocatalytic activity even though it is completely inactive on other substrates) | 0 | 0 | <0.00014 | <0.02 | <276 |
| chr5:95746663C>T | rs148617898 | non_synonymous_codon | ENST00000311106.3:c.910G>A | ENSP00000308024.2:p.Val304Ile | tolerated | probably_damaging | deleterious |  | 4878 | 3 | 0.00062 | 0.38 | 1229 |
| chr5:95748035T>C | . | non_synonymous_codon | ENST00000311106.3:c.869A>G | ENSP00000308024.2:p.Tyr290Cys | tolerated | probably_damaging | deleterious |  | 2184 | 1 | 0.00046 | 0.21 | 915 |
| chr5:95748068C>G | rs142673134 | non_synonymous_codon | ENST00000311106.3:c.836G>C | ENSP00000308024.2:p.Gly279Ala | deleterious | probably_damaging | deleterious |  | 4878 | 1 | 0.00021 | 0.04 | 410 |
| chr5:95748122A>G | rs139602265 | non_synonymous_codon | ENST00000311106.3:c.782T>C | ENSP00000308024.2:p.Val261Ala | deleterious | probably_damaging | deleterious |  | 4878 | 1 | 0.00021 | 0.04 | 410 |
| chr5:95748156C>A | rs137852822 | stop_gained | ENST00000311106.3:c.748G>T | ENSP00000308024.2:p.Glu250X | . | . | . |  | 0 | 0 | <0.00014 | <0.02 | <276 |
| chr5:95748188C>T | rs147140869 | non_synonymous_codon | ENST00000311106.3:c.716G>A | ENSP00000308024.2:p.Arg239Lys | deleterious | probably_damaging | deleterious |  | 0 | 0 | <0.00014 | <0.02 | <276 |
| chr5:95751742A>G | . | non_synonymous_codon | ENST00000311106.3:c.704T>C | ENSP00000308024.2:p.Val235Ala | deleterious | probably_damaging | deleterious |  | 2184 | 1 | 0.00046 | 0.21 | 915 |
| chr5:95751745T>A | rs145127903 | non_synonymous_codon | ENST00000311106.3:c.701A>T | ENSP00000308024.2:p.Lys234Ile | deleterious | probably_damaging | deleterious |  | 4876 | 1 | 0.00021 | 0.04 | 410 |
| chr5:95751796G>A | . | non_synonymous_codon | ENST00000311106.3:c.650C>T | ENSP00000308024.2:p.Ala217Val | deleterious | probably_damaging | deleterious |  | 2184 | 1 | 0.00046 | 0.21 | 915 |
| chr5:95751806CTGC>C | rs137852823 | inframe_codon_loss | ENST00000311106.3:c.637_639delGCA | ENSP00000308024.2:p.Ala213del | . | . | . | 213:213:Missing (in PC1 deficiency) | 0 | 0 | <0.00014 | <0.02 | <276 |
| chr5:95759019A>G | rs145592525 | non_synonymous_codon | ENST00000311106.3:c.541T>C | ENSP00000308024.2:p.Tyr181His | deleterious | probably_damaging | deleterious |  | 4854 | 1 | 0.00021 | 0.04 | 412 |
| chr5:95759036G>A | rs140520429 | non_synonymous_codon | ENST00000311106.3:c.524C>T | ENSP00000308024.2:p.Thr175Met | deleterious | probably_damaging | deleterious |  | 7048 | 4 | 0.00057 | 0.32 | 1134 |
| chr5:95759098T>G | rs145659863 | non_synonymous_codon | ENST00000508626.1:c.321A>C | ENSP00000421600.1:p.Lys107Asn | tolerated | benign | neutral |  | 4876 | 1 | 0.00021 | 0.04 | 410 |
| chr5:95759151T>G | . | non_synonymous_codon | ENST00000311106.3:c.409A>C | ENSP00000308024.2:p.Met137Leu | tolerated | benign | deleterious |  | 4876 | 1 | 0.00021 | 0.04 | 410 |
| chr5:95759156G>A | . | non_synonymous_codon | ENST00000311106.3:c.404C>T | ENSP00000308024.2:p.Thr135Ile | deleterious | benign | deleterious |  | 4876 | 1 | 0.00021 | 0.04 | 410 |
| chr5:95761545C>T | rs146545244 | non_synonymous_codon | ENST00000311106.3:c.375G>A | ENSP00000308024.2:p.Met125Ile | tolerated | benign | deleterious |  | 4876 | 3 | 0.00062 | 0.38 | 1230 |
| chr5:95761576T>G | . | non_synonymous_codon | ENST00000311106.3:c.344A>C | ENSP00000308024.2:p.Asp115Ala | tolerated | benign | deleterious |  | 2184 | 1 | 0.00046 | 0.21 | 915 |
| chr5:95764963C>T | rs1799904 | non_synonymous_codon | ENST00000311106.3:c.239G>A | ENSP00000308024.2:p.Arg80Gln | tolerated | possibly_damaging | deleterious |  | 2184 | 19 | 0.00870 | 76 | 17248 |
| chr5:95764968C>A | rs148354360 | non_synonymous_codon | ENST00000311106.3:c.234G>T | ENSP00000308024.2:p.Arg78Ser | deleterious | probably_damaging | deleterious |  | 4872 | 1 | 0.00021 | 0.04 | 410 |
| chr5:95767810T>C | rs144071994 | non_synonymous_codon | ENST00000508626.1:c.16A>G | ENSP00000421600.1:p.Ile6Val | tolerated | benign | deleterious |  | 2184 | 2 | 0.00092 | 0.84 | 1830 |
| chr5:95768682A>G | . | non_synonymous_codon | ENST00000311106.3:c.65T>C | ENSP00000308024.2:p.Leu22Pro | tolerated | benign | deleterious |  | 2374 | 2 | 0.00084 | 0.71 | 1684 |

**Table S1. Potentially harmful mutations in PCSK1 found in public datasets**. Data from dbSNP134, 1000 Genomes, and the NHLBI and NIEHS exome projects was scanned for variants in PCSK1 that were predicted by the Ensembl Variant Effect Predictor to have an adverse consequence on the transcript. Alleles: total number of alleles observed in 1000 Genomes, NLHBI, and NIEHS; Alt_alleles: total number of observed non-reference alleles; MAF: minor allele frequency; hom/M: estimate of incidence of homozygous individuals per 1 million population, assuming Hardy-Weinberg equilibrium; het/M estimate of incidence of heterozygous individuals per million.

# CHAPTER FIVE

## Functional consequences of a novel variant of *PCSK1*

# Functional Consequences of a Novel Variant of *PCSK1*

**Lindsay A. Pickett[3], Michael Yourshaw[1], Valeria Albornoz[3], Zijun Chen[2], R. Sergio Solorzano-Vargas[2], Stanley F. Nelson[1,4], Martín G. Martín[2], Iris Lindberg[3]***

**1** Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, California, United States of America, **2** Division of Gastroenterology and Nutrition, Department of Pediatrics, Mattel Children's Hospital and David Geffen School of Medicine, University of California Los Angeles, Los Angeles, California, United States of America, **3** Department of Anatomy and Neurobiology, University of Maryland-Baltimore, Baltimore, Maryland, United States of America, **4** Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California Los Angeles, Los Angeles California, United States of America

## Abstract

***Background:*** Common single nucleotide polymorphisms (SNPs) in proprotein convertase subtilisin/kexin type 1 with modest effects on PC1/3 *in vitro* have been associated with obesity in five genome-wide association studies and with diabetes in one genome-wide association study. We here present a novel SNP and compare its biosynthesis, secretion and catalytic activity to wild-type enzyme and to SNPs that have been linked to obesity.

***Methodology/Principal Findings:*** A novel PC1/3 variant introducing an Arg to Gln amino acid substitution at residue 80 (within the secondary cleavage site of the prodomain) (rs1799904) was studied. This novel variant was selected for analysis from the 1000 Genomes sequencing project based on its predicted deleterious effect on enzyme function and its comparatively more frequent allele frequency. The actual existence of the R80Q (rs1799904) variant was verified by Sanger sequencing. The effects of this novel variant on the biosynthesis, secretion, and catalytic activity were determined; the previously-described obesity risk SNPs N221D (rs6232), Q665E/S690T (rs6234/rs6235), and the Q665E and S690T SNPs (analyzed separately) were included for comparative purposes. The novel R80Q (rs1799904) variant described in this study resulted in significantly detrimental effects on both the maturation and *in vitro* catalytic activity of PC1/3.

***Conclusion/Significance:*** Our findings that this novel R80Q (rs1799904) variant both exhibits adverse effects on PC1/3 activity and is prevalent in the population suggests that further biochemical and genetic analysis to assess its contribution to the risk of metabolic disease within the general population is warranted.

**Citation:** Pickett LA, Yourshaw M, Albornoz V, Chen Z, Solorzano-Vargas RS, et al. (2013) Functional Consequences of a Novel Variant of *PCSK1*. PLoS ONE 8(1): e55065. doi:10.1371/journal.pone.0055065

**Editor:** Ludmila Prokunina-Olsson, National Cancer Institute, National Institutes of Health, United States of America

**Received** May 29, 2012; **Accepted** December 22, 2012; **Published** January 28, 2013

**Funding:** This work was supported by grants from the National Institute of Diabetes and Digestive and Kidney Diseases (#DK083762) and the California Institute of Regenerative Medicine (CIRM), RT2-01985 to MGM; and a grant from the National Institute of Drug Abuse (#DA05084) to IL. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: ilind001@umaryland.edu

## Introduction

Prohormone convertase 1/3 is a calcium-dependent serine endoprotease essential for the conversion of a variety of prohormones and neuropeptide precursors to their bioactive forms. Human prohormone convertase 1/3 (PC1/3) is encoded by the gene *PCSK1*, which is located on chromosome 5 and is comprised of 14 exons [1]. PC1/3 is expressed in a subset of endocrine and neuroendocrine tissues, cells equipped with a regulated secretory pathway. During transit through the secretory pathway, PC1/3 is first synthesized in the endoplasmic reticulum (ER) as an inactive 94 kDa zymogen composed of an N-terminal signal peptide, a prodomain which serves as an intramolecular chaperone and inhibitor; a catalytic domain which accomplishes substrate hydrolysis; a P (homo B) domain which contributes to enzymatic properties; and a carboxyl-terminal (CT) domain which, when removed by partial or complete *in trans* proteolytic processing, results in a much more active, but also less stable, enzymatic form (reviewed in [2] (**Figure 1**). PC1/3 is abundantly expressed in the arcuate and paraventricular nuclei of the hypothalamus [3,4], tissues that are known to mediate satiety and hunger signals [5]. Substrates of PC1/3, such as proinsulin, proglucagon, proghrelin, agouti-related protein, pro-neuropeptide Y, provasopressin and proopiomelanocortin are responsible for the regulation of absorption, metabolism and acquisition (appetite) of nutrients [6,7,8,9,10,11,12,13,14].

Deficiencies in PC1/3 frequently lead to imbalances in prohormone processing that result in an array of metabolic phenotypes, previously investigated both in mouse models and in humans. Three human subjects have been described with an autosomal recessive disorder (MIM:600955) associated with severe mutations of PC1/3 resulting in early-onset obesity, hyperphagia, hypoadrenalism, reactive hypoglycemia, malabsorptive diarrhea, and hypogonadism [15,16,17]. Interestingly, the PC1/3 null mouse model, unlike the PC1/3-deficient human, is not obese. Although of normal weight at birth, PC1/3 null mice have a high post-natal mortality rate, and those that do survive have a significant reduction in body mass as compared to wild-type animals by the age of 6 weeks. The stunted growth of PC1/3 null mice is believed to be due at least in part to reduced processing of

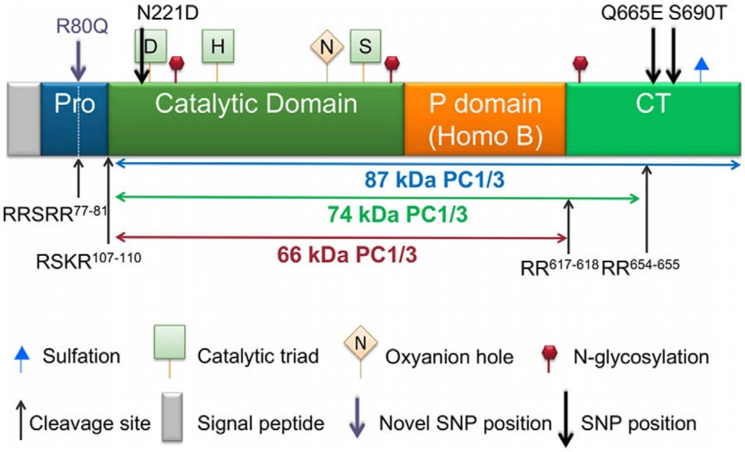


**Figure 1. Domain structure and SNP locations within preproPC1/3.** The upward arrows indicate the cleavage sites required for PC1/3 maturation. The downward arrows indicate locations of previously described (black) and novel (purple) SNP. The dashed line between the pro and catalytic domains represents a primary cleavage site (occurring in the ER) that is required for activation. The dashed line in the middle of the prodomain indicates the secondary cleavage site (likely cleaved in the trans-Golgi network). The P or Homo B domain following the catalytic domain is important for the stabilization of the catalytic domain, as well as determining various enzymatic properties. The C-terminal domain plays a role in efficient routing of PC1/3 to the secretory granules, and contributes to substrate specificity as well as to specific activity and stability.
doi:10.1371/journal.pone.0055065.g001

growth hormone releasing hormone (GHRH) and thus reduced circulating levels of growth hormone (GH) [8]. In addition to a reduction in GHRH, the levels of several key neuroendocrine peptides such as ACTH, insulin and glucagon-like peptides-1 and -2 are reduced in these animals due to lack of precursor processing by PC1/3 [8].

While the PC1/3 null mouse is not obese, a mouse model of obesity has been generated via introduction of a missense mutation in *PCSK1* at amino acid position 222, near the calcium-binding pocket in the catalytic domain. This hypomorph mutation resulted in obesity, hyperphagia and increased metabolic efficiency due to decreased autocatalytic maturation of the enzyme to smaller molecular weight forms [18]. Three common SNPs in *PCSK1* have been identified and associated with obesity. All three SNPs (included in this study for comparison) exhibit moderate effects on catalytic activity *in vitro* and on natural substrate processing in rat pituitary tumor cells [19,20]. Two of the three non-deleterious SNPs (S690T [rs6235] and Q665E [rs6234]) have been associated with diabetes-related traits [20,21,22].

In the work presented below, the novel variant NP_000430.3:p.Arg80Gln (R80Q; rs1799904), identified and functionally evaluated for the first time here, was compared with previously described SNPs associated with obesity and/or diabetes (N221D [rs6232], Q665E/S690T [rs6234/rs6235], Q665E [rs6234], and S690T [rs6235]) for potentially deleterious effects on the biosynthesis, secretion and catalytic activity of PC1/3. Our data suggest that this novel R80Q variant (rs1799904) deserves further analysis to assess its genetic association with metabolic diseases such as obesity and diabetes.

## Materials and Methods

### Databases used and protein structure/function analysis methods

Alleles that varied from the human reference genome build GRCh37 [23] were obtained from the dbSNP [24], 1000 Genomes [25], NHLBI [26], and NIEHS [27] datasets and were merged into a custom SQL database. dbSNP data were compiled from various sources, with allele frequencies available only for a subset of variants. The 1000 Genomes dataset was based on both low coverage whole genome and higher coverage exome sequencing of 1092 individuals. The NHLBI and NIEHS data were obtained from exome sequencing of 6500 and 95 individuals respectively. Population allele frequencies were calculated using the combined datasets wherever allele counts were present. Variations in *PCSK1* (chr5:95726119-95769847) were identified and analyzed with the Ensembl Variant Effect Predictor version 2.6 [28] and Ensembl database homo_sapiens_variation_68_37 [29] to determine the effect of the variant on the transcript. Non-synonymous codon substitutions were analyzed using the SIFT [23,30,31,32,33], PolyPhen [34,35,36], and Condel [37] models to estimate the variant's probable impact on protein structure and function.

### Sanger sequencing of genomic DNA

Genomic DNA from individuals homozygous for two SNPs of interest was isolated from EBV-infected B cells by the Coriell Institute and sent to us for sequencing. The HG00596 DNA sample containing rs1799904 (p.R80Q; g.5:95764963C>T; c.239G>A) was obtained from a southern Han Chinese female, while the N586Tfsx4-containing (g.5:95730696TC>T; c.1755delG) DNA sample, HG00350, was obtained from a Finnish female. The primers used for sequencing bidirectionally were:

Exon 2 (510 bp):
(F) CTCAACCAATTCAACCCAATC;
(R) CCCGTGACACAAGTTTACCTATG; and
Exon 13 (545 bp):
(F) CAGCTTTCCAAGAACACATCC;
(R) CCATGTTTGACTTATTTCCTGC

### Expression vector construction/mutagenesis

Flag-tagged human PC1/3, a kind gift of J. W. Creemers [20] was modified by site-directed mutagenesis using the Stratagene

**Table 1.** Potentially consequential variant alleles in *PCSK1*.

| Pos | ID | REF | ALT | Rank | cDNA | Protein | Effect | MAF | Samples | Het | Hom |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 5:95768682 | rs201377789 | A | G | 12 | 65T>C | Leu22Pro | | 0.00033 | 7690 | 5 | 0 |
| 5:95764976 | | G | A | 12 | 226C>T | Pro76Ser | | 0.00008 | 6501 | 1 | 0 |
| 5:95764968 | rs148354360 | C | A | 12 | 234G>T | Arg78Ser | SpC | 0.00015 | 6501 | 2 | 0 |
| **5:95764963** | **rs1799904** | **C** | **T** | **12** | **239G>A** | **Arg80Gln** | **p** | **0.00870** | **1092** | **17** | **1** |
| 5:95761576 | rs200893367 | T | G | 12 | 344A>C | Asp115Ala | S | 0.00046 | 1092 | 1 | 0 |
| 5:95761546 | | A | T | 12 | 374T>A | Met125Lys | | 0.00008 | 6503 | 1 | 0 |
| 5:95761545 | rs146545244 | C | T | 12 | 375G>A | Met125Ile | | 0.00038 | 6503 | 5 | 0 |
| 5:95759156 | | G | A | 12 | 404C>T | Thr135Ile | PC | 0.00008 | 6503 | 1 | 0 |
| 5:95759151 | | T | G | 12 | 409A>C | Met137Leu | | 0.00008 | 6503 | 1 | 0 |
| 5:95759098 | rs145659863 | T | G | 12 | 462A>C | Lys154Asn | S | 0.00008 | 6503 | 1 | 0 |
| 5:95759093 | | A | G | 12 | 467T>C | Ile156Thr | SpC | 0.00008 | 6503 | 1 | 0 |
| 5:95759090 | rs200462856 | G | A | 12 | 470C>T | Thr157Met | SPC | 0.00015 | 6503 | 2 | 0 |
| 5:95759036 | rs140520429 | G | A | 12 | 524C>T | Thr175Met | SPC | 0.00026 | 7595 | 4 | 0 |
| 5:95759019 | rs145592525 | A | G | 12 | 541T>C | Tyr181His | SPC | 0.00038 | 6503 | 5 | 0 |
| 5:95757611 | | G | A | 12 | 593C>T | Pro198Leu | SPC | 0.00008 | 6503 | 1 | 0 |
| 5:95751796 | rs202203086 | G | A | 12 | 650C>T | Ala217Val | SPC | 0.00013 | 7595 | 2 | 0 |
| *5:95751785* | *rs6232* | *T* | *C* | *12* | *661A>G* | *Asn221Asp* | | *0.03289* | *8254* | *523* | *10* |
| 5:95751745 | rs145127903 | T | A | 12 | 701A>T | Lys234Ile | SPC | 0.00008 | 6503 | 1 | 0 |
| 5:95751742 | rs183045011 | A | G | 12 | 704T>C | Val235Ala | SPC | 0.00046 | 1092 | 1 | 0 |
| 5:95748134 | | T | C | 12 | 770A>G | Asn257Ser | S | 0.00008 | 6503 | 1 | 0 |
| 5:95748123 | | C | T | 12 | 781G>A | Val261Met | SPC | 0.00008 | 6503 | 1 | 0 |
| 5:95748122 | rs139602265 | A | G | 12 | 782T>C | Val261Ala | SPC | 0.00008 | 6503 | 1 | 0 |
| 5:95748068 | rs142673134 | C | G | 12 | 836G>C | Gly279Ala | SPC | 0.00008 | 6503 | 1 | 0 |
| 5:95748035 | rs193214131 | T | C | 12 | 869A>G | Tyr290Cys | PC | 0.00026 | 7595 | 4 | 0 |
| 5:95746664 | | G | C | 12 | 909C>G | Phe303Leu | SPC | 0.00008 | 6503 | 1 | 0 |
| 5:95746663 | rs148617898 | C | T | 12 | 910G>A | Val304Ile | pC | 0.00038 | 6503 | 5 | 0 |
| 5:95746638 | rs138879299 | C | T | 12 | 935G>A | Arg312His | S | 0.00008 | 6503 | 1 | 0 |
| 5:95746543 | rs189927028 | C | T | 12 | 1030G>A | Ala344Thr | P | 0.00046 | 1092 | 1 | 0 |
| 5:95744026 | | G | A | 12 | 1097C>T | Thr366Met | S | 0.00008 | 6503 | 1 | 0 |
| 5:95735742 | rs140481124 | G | C | 12 | 1345C>G | Leu449Val | SPC | 0.00008 | 6503 | 1 | 0 |
| 5:95735724 | | G | T | 12 | 1363C>A | Pro455Thr | SpC | 0.00008 | 6503 | 1 | 0 |
| 5:95735703 | rs151257336 | G | T | 12 | 1384C>A | Pro462Thr | SpC | 0.00008 | 6503 | 1 | 0 |
| 5:95735700 | rs143174906 | C | T | 12 | 1387G>A | Glu463Lys | | 0.00059 | 7595 | 9 | 0 |
| 5:95734621 | rs149124467 | C | T | 12 | 1550G>A | Arg517Gln | pC | 0.00015 | 6503 | 2 | 0 |
| 5:95734610 | | G | A | 12 | 1561C>T | Leu521Phe | SpC | 0.00008 | 6503 | 1 | 0 |
| 5:95734581 | | A | G | 3 | 1588+2T>C | | - | 0.00015 | 6503 | 2 | 0 |
| 5:95730719 | rs145196120 | A | G | 12 | 1733T>C | Ile578Thr | | 0.00008 | 6503 | 1 | 0 |
| 5:95730638 | | C | T | 12 | 1814G>A | Arg605His | SPC | 0.00008 | 6503 | 1 | 0 |
| 5:95730629 | | G | A | 12 | 1823C>T | Thr608Met | SpC | 0.00015 | 6503 | 2 | 0 |
| 5:95730597 | rs144324144 | C | G | 12 | 1855G>C | Gly619Arg | P | 0.00008 | 6503 | 1 | 0 |
| 5:95730576 | | G | T | 12 | 1876C>A | Pro626Thr | | 0.00008 | 6503 | 1 | 0 |
| 5:95729049 | rs139453594 | T | C | 12 | 1918A>G | Thr640Ala | | 0.00145 | 7595 | 22 | 0 |
| 5:95729048 | rs193015519 | G | A | 12 | 1919C>T | Thr640Ile | | 0.00013 | 7595 | 2 | 0 |
| 5:95729039 | rs142453906 | G | C | 12 | 1928C>G | Ser643Cys | | 0.00008 | 6502 | 1 | 0 |
| 5:95729007 | rs200614230 | G | A | 12 | 1960C>T | Arg654Trp | S | 0.00046 | 1092 | 1 | 0 |
| 5:95728982 | rs148807505 | G | T | 12 | 1985C>A | Ala662Asp | S | 0.00008 | 6503 | 1 | 0 |
| *5:95728974* | *rs6234* | *G* | *C* | *12* | *1993C>G* | *Gln665Glu* | | *0.24962* | *7900* | *2988* | *478* |
| *5:95728898* | *rs6235* | *C* | *G* | *12* | *2069G>C* | *Ser690Thr* | | *0.23747* | *7900* | *2852* | *450* |
| 5:95728877 | rs138433207 | T | G | 12 | 2090A>C | Tyr697Ser | | 0.00008 | 6503 | 1 | 0 |

**Table 1.** Cont.

| Pos | ID | REF | ALT | Rank | cDNA | Protein | Effect | MAF | Samples | Het | Hom |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 5:95728863 | rs140899352 | C | T | 12 | 2104G>A | Glu702Lys |  | 0.00039 | 7595 | 6 | 0 |
| 5:95728862 | rs188666266 | T | G | 12 | 2105A>C | Glu702Ala | S | 0.00046 | 1092 | 1 | 0 |
| 5:95728749 |  | G | A | 12 | 2218C>T | Arg740Trp | SPC | 0.00008 | 6503 | 1 | 0 |
| 5:95728748 | rs140941383 | C | T | 12 | 2219G>A | Arg740Gln | SPC | 0.00008 | 6503 | 1 | 0 |
| 5:95728710 | rs147016634 | T | G | 12 | 2257A>C | Asn753His | SC | 0.00046 | 7595 | 7 | 0 |

The R80Q (rs1799904) variant that differed from the human reference genome and was predicted to have a potentially consequential effect on the transcript was selected from the dbSNP 137, 1000 Genomes, NHLBI, and NIEHS public datasets. *Pos*: genomic position in GRCh37; *ID*: dbSNP 137 rs ID; *REF*: reference allele; *ALT*: alternate allele (variant); *Rank*: 3 splice_donor_variant, 12 missense_variant; *cDNA*: position and consequence of variant in cDNA of canonical NM_000439.4 transcript; *Protein*: position and consequence of variant in NP_000430.3 protein; *Effect*: computational prediction of effect on protein structure or function ("S" predicted deleterious by SIFT, "P" or "p" predicted probably or possibly damaging by PolyPhen, "C", predicted deleterious by Condel from a consensus of SIFT and PolyPhen, "-" no prediction); *MAF*: minor allele frequency across all populations; *Samples*: total number of individuals genotyped; *Het*: number of individuals heterozygous for the variant allele; *Hom*: number of individuals homozygous for the variant allele. Known, common variants are listed in italics, and the rare novel variant is shown in bold.
doi:10.1371/journal.pone.0055065.t001

QuikChange method [38] to introduce the mutations shown in **Figure 1**. All mutations were verified by sequencing of the entire PC1/3 cDNA insert.

## Transient transfection of PC1/3 variants

To assess the biosynthesis and secretion profiles of PC1/3 variants in a cell line that does not express endogenous PC1/3, Ad-293 (Stratagene) HEK cells, plated at a density of $2\times10^5$ cells per well in 24-well plates, were transfected with plasmids encoding either wild-type or variant PC1/3s in triplicate wells. Cells were transfected with 200 ng of plasmid DNA per well using Lipofectamine (Invitrogen, Carlsbad, CA). To assess effects in a regulated neuroendocrine cell line (also lacking expression of endogenous PC1/3), Neuro-2A cells (ATCC, cat. No. CCL-131) were transfected in triplicate with the same protocol using Lipofectamine 2000 (Invitrogen, Carlsbad, CA). For both cell lines, five hours post-transfection, 1 ml of growth medium was added to each well and incubation continued for an additional 24 h. Cells were then washed with PBS and 0.3 ml of Opti-MEM (Invitrogen, Carlsbad, CA) containing 100 ug/ml bovine aprotinin (Desert Biologicals) was added to each well. Cells were incubated for an additional 18–24 h before conditioned medium and cells were harvested. Conditioned medium was analyzed first by enzyme assay; both cells and medium samples (for HEK cells) and medium samples (for Neuro- 2A cells) were then subjected to SDS-PAGE followed by Western blotting using primary antiserum against the amino terminus of mature mouse PC1/3 [39]. Mouse monoclonal anti-ß-actin antiserum (Sigma-Aldrich, St. Louis, MO) was used to assess cellular actin levels as a loading control. Western blots were then probed with horseradish peroxidase-coupled secondary antiserum. Visualization of immunoreactive protein was accomplished using the SuperSignal West Femto Maximum Sensitivity Substrate kit (Thermo Scientific, Rockford, IL).

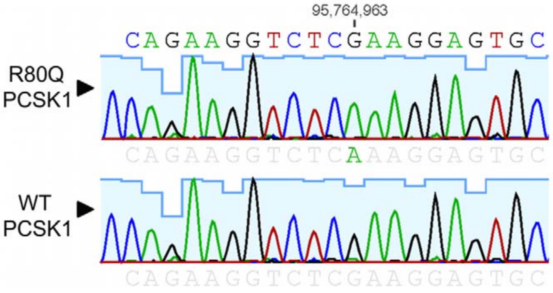


**Figure 2. Direct Sanger sequencing of genomic DNA from a subject bearing the Arg80Gln variant.**
doi:10.1371/journal.pone.0055065.g002

## Enzyme assay

Enzymatic activity of secreted recombinant PC1/3 proteins present in conditioned medium obtained from transiently transfected HEK293 cells was measured in triplicate 50 ul reactions in a 96-well polypropylene plate containing 25 ul of conditioned medium and final concentrations of 200 uM substrate (pyr-ERTKR-amc [7-amino-4-methlcoumarin]), 100 mM sodium acetate, pH 5.5, 2 mM $CaCl_2$, 0.1% Brij 35, and a protease inhibitor cocktail (final concentrations: 1 uM pepstatin, 0.28 mM TPCK, 10 uM E-64, and 0.14 mM TLCK). Reaction mixtures were incubated at 37°C and fluorescence measurements (380 nm excitation, 460 emission) were taken under kinetic conditions every 20 seconds for 1 h in a SpectraMax M2 Microplate Reader. Maximum rates were obtained from the linear portion of the kinetic measurement curves. Specific activities of PC1/3 proteins in the conditioned medium were determined by dividing maximum rates by band intensities of total secreted immunoreactive protein, each determined in triplicate, and quantified with an Alphaimager 3300 (Alpha Innotech Corporation, San Leandro, CA) imaging system.

# Results

## Analysis of public databases; structure-function analysis

A total of 1020 allelic variants (data not shown) within the *PCSK1* gene were found in the public databases, of which 54 were potentially consequential splice site or missense variants (**Table 1**). Thirty-seven non-synonymous substitutions were predicted to be possibly or probably deleterious by at least one model (SIFT, PolyPhen, or Condel, where Condel represents a consensus modeling program). Two of the three previously described variants are common, with MAFs of 23.7% for S690T (rs6235) and 25.0% for Q665E (rs6234), whereas the N221D SNP (rs6232) is less common (MAF = 3.3%). None of these three variants were predicted to be deleterious using SIFT, PolyPhen, or Condel. In contrast, the novel variants that were predicted as "possibly" or "probably" deleterious were unique to one sample or were observed with very low frequency (minor allele frequencies (MAFs) of 0.008%–0.87%. In addition we considered a frameshift variant

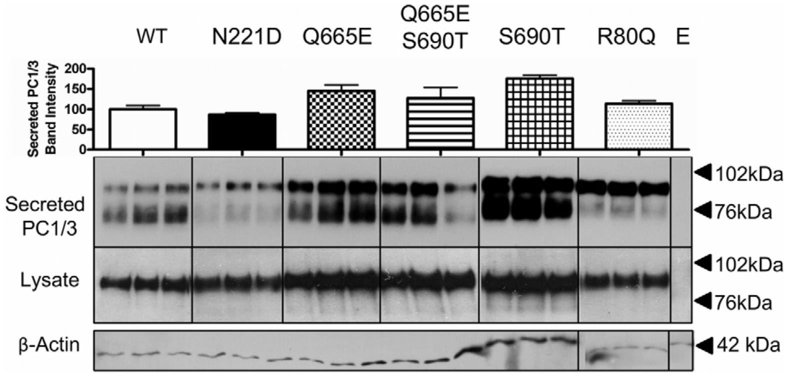


**Figure 3. Western blotting of wild-type and variant PC1/3 proteins expressed in HEK cells.** HEK cells were transiently transfected with empty pcDNA3 (E); pcDNA3 encoding either wild-type PC1/3; or PC1/3 proteins bearing the mutations under study. Western blots of cell lysates and media from transfected HEK cells were probed with amino-terminally directed PC1/3 primary antiserum for detection of recombinant proteins. The data shown represent 1 of 3 independent experiments performed in triplicate. Total secreted immunoreactive band intensity values, obtained through densitometry analysis and used to calculate specific activity for each variant, are represented above the Western blot and shown as the mean ± S.D.

doi:10.1371/journal.pone.0055065.g003

N586TfX4 (g.5:95730696), which exhibited an unusually large MAF of 6.1% in a previous release of the 1000 Genomes data. We selected the most common novel variant R80Q (rs1799904; MAF = 0.87%), and N586TfsX4 for genomic sequencing and potential functional studies, comparing them with already described common variants of PC1/3.

## SNP validation by sequencing

Genomic DNA from two individuals homozygous for the most common variants was obtained from the Coriell Institute and subjected to Sanger sequencing. The DNA sample containing rs1799904 (R80Q (g.5:95764963C>T; c.239G>A) was found to be homozygous for the R80Q mutation in exon 2 (**Figure 2**), while the N586Tfsx4-containing SNP (g.5:95730696TC>T; c.1755delG) was determined to be a false positive (*i.e.* no frameshift mutation was found in in exon 13) (data not shown).

## Secretion and biosynthesis of PC1/3 variants

In order to assess whether the novel variant R80Q (rs1799904) affected the biosynthesis or secretion of PC1/3, expression vectors encoding wild-type and variant *PCSK1*s were transiently transfected into HEK and/or Neuro-2a cells (both lines lack expression of endogenous PC1/3). PC1/3 proteins containing the previously described S690T/Q665E (rs6234/rs6235) pair, as well as the individual S690T and Q665E SNPs, did not exhibit significantly altered expression and secretion patterns as compared to wild-type PC1/3. The N221D (rs6232) substitution resulted in reduced secretion and cleaved forms of PC1/3 in the medium (**Figure 3**). The secretion profile of the R80Q (rs1799904) substitution differed

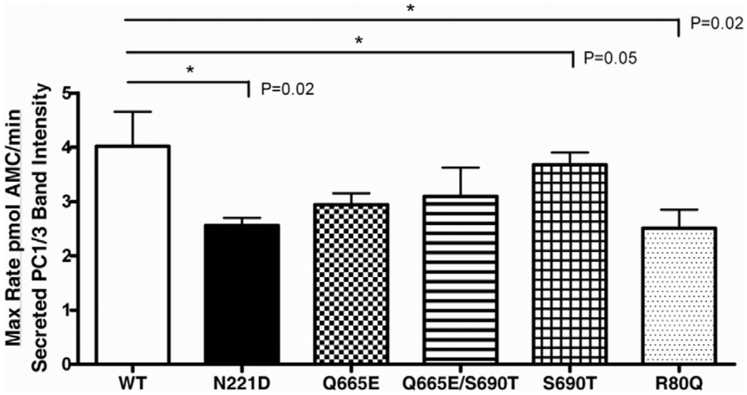


**Figure 4. Specific activities of wild-type and variant PC1/3 proteins, expressed in HEK cells.** Enzymatic activities of secreted recombinant PC1/3 proteins in conditioned medium of transfected HEK cells were compared by measuring maximum cleavage rates using the fluorogenic substrate pyr-ERTKR-amc during a 1 h kinetic assay. Three replicates per transfection condition were assayed in triplicate, and maximum rates were divided by band intensity of immunoreactive protein in the spent medium of the same wells from which activity data were derived. Specific activity values are shown as the mean ± S.D (n = 3). Data represent one of 3 independent experiments performed in triplicate.

doi:10.1371/journal.pone.0055065.g004

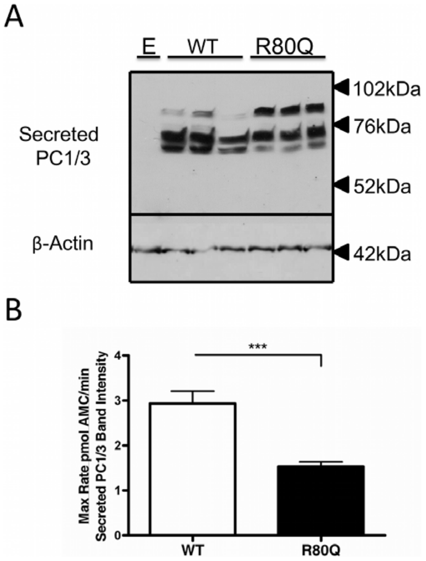


**Figure 5. Western blotting of wild-type and novel R80Q (rs1799904) variant PC1/3s, expressed in Neuro-2A cells. Panel A:** Neuro-2a cells were transiently transfected with equal amounts of empty pcDNA3 (E), or pcDNA3 encoding wild-type PC1/3 or the novel variant R80Q (rs1799904) PC1/3. Western blots of media were probed using amino-terminally directed PC1/3 primary antiserum. The data shown represent one of 3 independent experiments performed in triplicate. **Panel B: Specific activities of wild-type PC1/3 and the R80Q PC1/3 variant.** Enzymatic activities of secreted recombinant PC1/3 proteins in conditioned medium were compared by measuring maximum cleavage rates using the fluorogenic substrate pyr-ERTKR-amc during a 1 h kinetic assay. Three replicates per transfection condition were assayed in triplicate, and maximum rates were divided by band intensity of immunoreactive protein in the spent medium of the same wells from which the activity data were derived. Specific activity values are shown as the mean ± S.D. Data represent one of 3 independent experiments, each performed in triplicate.
doi:10.1371/journal.pone.0055065.g005

from wild-type PC1/3, in that the 74 and 66 kDa lower molecular weight forms of PC1/3 were absent from the medium (in HEK cell experiments) or reduced (in Neuro-2a cell experiments), although the total level of secreted PC1/3 was not reduced.

### Catalytic activity of PC1/3 variants

To determine the impact of these variations on PC1/3 catalytic activity, conditioned medium of HEK cells transfected with either empty vector, variant PC1/3s, or wild-type PC1/3 was subjected to a fluorogenic assay. Maximum rates of fluorogenic substrate cleavage were normalized using the band intensities of secreted PC1/3s in order to determine the specific activity of each variant relative to wild-type PC1/3. The S690T/Q665E (rs6234/rs6235) and S690T (rs6234) amino acid substitutions did not significantly alter specific activity (95% confidence level; p>0.13). The Q665E substitution alone resulted in a small but significant 27% decrease in specific activity as compared to wild-type (p = 0.05). The N221D (rs6232) substitution decreased specific activity by 36% (p = 0.02), and the R80Q variation resulted in a 38% decrease (p = 0.02) (**Figure 4**). When expressed in Neuro-2a cells, the R80Q (rs1799904) variant resulted in a 42–48% decrease (p<0.0001) in activity as compared to wild-type PC1/3 (**Figure 5**).

## Discussion

In studies of European populations, *PCSK1* represents the third most important gene contributing to extreme obesity [40]. Functional studies of certain SNPs associated with obesity that impose modest or no significant effects on PC1/3 function *in vitro* have supported the idea that even slight variations in PC1/3 activity can predispose an individual to higher risk of obesity [20]. Individuals who are compound heterozygotes or are homozygous for rare severe deleterious mutations in *PCSK1* suffer from multi-dimensional disease states, including small intestinal dysfunction, hyperphagia and obesity [15,16,17]. Even heterozygous mutations which result in functional enzymatic changes have been linked to obesity, despite the presence of a normal allele [40]. The mechanism by which modest deficiencies in PC1/3 activity can lead to such profound phenotypes when present on a single allele remains unknown. A closer look into the complex biochemistry of commonly found variations of this enzyme may provide answers to these questions. In this work, we have analyzed public databases for other less common and rare deleterious variants and identified the variant R80Q (rs1799904), and have compared the effects of this variant to those of known polymorphisms.

Consistent with previous studies [19,20], we found that the amino acid substitutions S690T/Q665E (rs6234/rs6235) did not significantly alter the specific activity or biosynthesis and secretion of PC1/3 in HEK cells. The Q665E substitution alone did result in a slight decrease in specific activity as compared to wild-type enzyme, and may represent the more detrimental of the two mutations (S690T/Q665E), which were previously identified as a paired SNP associated with a higher risk of developing obesity and diabetes [19,20,21]. In our hands, the N221D (rs6232) substitution decreased specific activity by a somewhat greater extent than previously reported, possibly due to differences in enzyme assay methods [20].

However, of all of the variants we analyzed in HEK cells, the novel R80Q (rs1799904) variant exhibited the most detrimental effects on PC1/3 maturation and specific activity. This variant yielded an 87 kDa product in the conditioned medium that did not undergo further carboxy-terminal processing to the more active 74 and 66 kDa forms, resulting in an enzyme with significantly lower specific activity, similar to the more common obesity risk N221D (rs6232) variant. This novel R80Q variant exhibited an even more pronounced decrease in specific activity when expressed in a cell line containing a regulated secretory pathway (Neuro-2a), where wild-type PC1/3 is likely able to achieve greater specific activity through more complete maturation to its lower molecular weight forms within regulated secretory vesicles. The lower molecular weight forms of PC1/3 exhibit a different substrate specificity than full-length 87 kDa PC1/3 [41]; this could be an important mechanism for SNPs to exert functional effects. Another possible functional consequence of altering the profile of active species is a change in enzyme stability, since carboxy-terminally truncated species are known to be more labile than the 87 kDa form (reviewed in [2]). Since the C-terminal region of PC1/3 has been implicated in targeting of this enzyme to secretory granules [42,43], altered C-terminal processing may also result in changes in enzyme distribution. Further studies using immunocytochemistry in transfected Neuro- 2A cells will shed additional light on this question.

The proPC1/3 maturation process begins with the autocatalytic intramolecular cleavage of the pro-domain in the ER at the primary cleavage site, RSKR$^{107-110}$ [44,45]. This cleavage yields an 87 kDa form of PC1/3 that, by analogy with the related enzyme furin [46] likely remains associated with its own prodomain through non-covalent interactions until its arrival at the trans-Golgi network. Although this has not yet been strictly demonstrated for PC1/3, the PC1/3 prodomain most likely assists in the folding of the catalytic domain and in enzyme inhibition during secretory pathway transport [47,48,49,50,51,52]. If prodomain processing of PC1/3 occurs similarly to that of furin, trans-Golgi network protonation of a histidine in the vicinity of the secondary cleavage site (RRSRR$^{77-81}$) then results in secondary site cleavage at R$^{81}$, followed by dissociation of prodomain fragments from PC1/3 [53,54]. The inhibitory role of the prodomain is of particular interest to this study when we consider the location of the R80Q (rs1799904) substitution within the secondary cleavage site of the prodomain (**Figure 1**). Independent studies have shown that alteration of mouse proPC1/3 prodomain residues either within or surrounding cleavage motifs can affect propeptide processing; the *in vitro* proteolytic conversion of an R80A mutant propeptide (the same residue as the R80Q variant studied here) by wild-type enzyme was impaired compared to wild-type propeptide [44]. Given this finding, our lack of identification of propeptide-bearing R80Q PC1/3 is puzzling. We have previously found that a portion of newly synthesized proPC1/3 is subjected to endoplasmic reticulum- associated degradation [52]; this might represent the fate of this molecular species. Collectively, these data support the idea that residues within the secondary cleavage site, including the novel variant studied here, contribute to the proper processing of proPC1/3.

The novel R80Q (rs1799904) variant (MAF = 0.87%) is about one-third as common as the N221D (rs6232) SNP (MAF = 3.3%). Although less common, the R80Q variant should be subjected to further analysis to evaluate its influence on insulin sensitivity, proinsulin conversion and the risk of developing obesity, similarly to the effect of the N221D (rs6232) SNP [20,22]. We note that 119 individuals in the public datasets have other, less common and rare variants of *PCSK1*, most of which are predicted to have some detrimental effect on protein function. This mutational burden on the population is not trivial and may also play a role in susceptibility to obesity or other disorders. The importance of rare variants in common disorders is not clear at present, but advances in massively parallel sequencing and computational analysis may soon shed additional light on this question.

In conclusion, we show that the novel *PCSK1* variant R80Q (rs1799904) exhibits deleterious effects on PC1/3 maturation. This PC1/3 variant exhibits decreased catalytic activity as compared to wild-type PC1/3 and to previously described obesity risk SNPs; therefore, it may contribute to a higher risk of metabolic disease in the general population. Our results suggest that further study of less common and rare variations in *PCSK1* from both biochemical and genetic standpoints will be useful in elucidating the mechanisms by which variant PC1/3s contribute to metabolic diseases such as obesity and diabetes.

## Author Contributions

Conceived and designed the experiments: MY MGM IL. Performed the experiments: LAP MY VA ZC RSS. Analyzed the data: LAP VA MY ZC RSS SN MGM IL. Contributed reagents/materials/analysis tools: MGM IL. Wrote the paper: LAP MY MGM IL.

## References

1. Seidah NG, Mattei MG, Gaspar L, Benjannet S, Mbikay M, et al. (1991) Chromosomal assignments of the genes for neuroendocrine convertase PC1 (NEC1) to human 5q15-21, neuroendocrine convertase PC2 (NEC2) to human 20p11.1-11.2, and furin (mouse 7[D1-E2] region). Genomics 11: 103–107.
2. Hoshino A, Lindberg I (2012) Peptide Biosynthesis: Prohormone Convertases 1/3 and 2. In: Fricker LD, Devi L, editors: Morgan & Claypool Life Sciences Publishers.
3. Schafer MK, Day R, Cullinan WE, Chretien M, Seidah NG, et al. (1993) Gene expression of prohormone and proprotein convertases in the rat CNS: a comparative in situ hybridization analysis. J Neurosci 13: 1258–1279.
4. Dong W, Seidel B, Marcinkiewicz M, Chretien M, Seidah NG, et al. (1997) Cellular localization of the prohormone convertases in the hypothalamic paraventricular and supraoptic nuclei: selective regulation of PC1 in corticotrophin-releasing hormone parvocellular neurons mediated by glucocorticoids. J Neurosci 17: 563–575.
5. Wynne K, Stanley S, McGowan B, Bloom S (2005) Appetite control. J Endocrinol 184: 291–318.
6. Smeekens S, Montag AG, Thomas G, Albiges-Rizo C, Carrol R, et al. (1992) Proinsulin processing by the subtilisin-related proprotein convertases furin, PC2, and PC3. Proc Natl Acad Sci USA 89: 8822–8826.
7. Rouille Y, Kantengwa S, Irminger JC, Halban PA (1997) Role of the prohormone convertases in the processing of proglucagon to glucagon-like peptides. J Biol Chem 72: 32810–32816.
8. Zhu X, Zhou A, Dey A, Norrbom C, Carroll R, et al. (2002) Disruption of PC1/3 expression in mice causes dwarfism and multiple neuroendocrine peptide processing defects. Proc Natl Acad Sci USA 99: 10293–10298.
9. Zhu X, Cao Y, Voogd K, Steiner DF (2006) On the processing of proghrelin to ghrelin. J Biol Chem 281: 38867–38870.
10. Creemers JW, Pritchard LE, Gyte A, Le Rouzic P, Meulemans S, et al. (2006) Agouti-related protein is posttranslationally cleaved by proprotein convertase 1 to generate agouti-related protein (AGRP)83-132: interaction between AGRP83-132 and melanocortin receptors cannot be influenced by syndecan-3. Endocrinology 147: 1621–1631.
11. Brakch N, Rist B, Beck-Sickinger AG, Goenaga J, Wittek R, et al. (1997) Role of prohormone convertases in pro-neuropeptide Y processing: coexpression and in vitro kinetic investigations. Biochemistry 36: 16309–16320.
12. Coates LC, Birch NP (1998) Differential cleavage of provasopressin by the major molecular forms of SPC3. J Neurochem 70: 1670–1678.
13. Benjannet S, Rondeau N, Day R, Chretien M, Seidah NG (1991) PC1 and PC2 are proprotein convertases capable of cleaving proopiomelanocortin at distinct pairs of basic residues. Proc Natl Acad Sci U S A 88: 3564–3568.
14. Thomas L, Leduc R, Thorne BA, Smeekens SP, Steiner D, et al. (1991) Kex2-like endoproteases PC2 and PC3 accurately cleave a model prohormone in mammalian cells: evidence for a common core of neuroendocrine processing enzymes. Proc Natl Acad Sci USA 88: 5297–5301.
15. Jackson RS, Creemers JW, Farooqi IS, Raffin-Sanson ML, Varro A, et al. (2003) Small-intestinal dysfunction accompanies the complex endocrinopathy of human proprotein convertase 1 deficiency. J Clin Invest 112: 1550–1560.
16. Jackson RS, Creemers JW, Ohagi S, Raffin-Sanson ML, Sanders L, et al. (1997) Obesity and impaired prohormone processing associated with mutations in the human prohormone convertase 1 gene. Nat Genet 16: 303–306.
17. Farooqi IS, Volders K, Stanhope R, Heuschkel R, White A, et al. (2007) Hyperphagia and early-onset obesity due to a novel homozygous missense mutation in prohormone convertase 1/3. J Clin Endocrinol Metab 92: 3369–3373.
18. Lloyd DJ, Bohan S, Gekakis N (2006) Obesity, hyperphagia and increased metabolic efficiency in Pc1 mutant mice. Hum Mol Genet 15: 1884–1893.
19. Mbikay M, Sirois F, Nkongolo KK, Basak A, Chretien M (2011) Effects of rs6234/rs6235 and rs6232/rs6234/rs6235 PCSK1 single-nucleotide polymorphism clusters on proprotein convertase 1/3 biosynthesis and activity. Mol Genet Metab 104: 682–687.
20. Benzinou M, Creemers JW, Choquet H, Lobbens S, Dina C, et al. (2008) Common nonsynonymous variants in PCSK1 confer risk of obesity. Nat Genet 40: 943–945.
21. Strawbridge RJ, Dupuis J, Prokopenko I, Barker A, Ahlqvist E, et al. (2011) Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. Diabetes 60: 2624–2634.
22. Heni M, Haupt A, Schafer SA, Ketterer C, Thamer C, et al. (2010) Association of obesity risk SNPs in PCSK1 with insulin sensitivity and proinsulin conversion. BMC Med Genet 11: 86.
23. (2004) Finishing the euchromatic sequence of the human genome. Nature 431: 931–945.
24. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29: 308–311.
25. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.
26. NHLBI Exome Sequencing Project (2011) Exome Variant Server. Seattle WA: NHLBI Exome Sequencing Project (ESP).
27. NIEHS Environmental Genome Project (2011) NIEHS Exome Variant Server. Seattle WA: NIEHS Environmental Genome Project.

28. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297–1303.
29. Flicek P, Amode MR, Barrell D, Beal K, Brent S, et al. (2011) Ensembl 2011. Nucleic Acids Res 39: D800–806.
30. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols 4: 1073–1082.
31. Ng PC, Henikoff S (2002) Accounting for human polymorphisms predicted to affect protein function. Genome Research 12: 436–446.
32. Ng CP, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Research 31: 3812–3814.
33. Ng PC, Henikoff S (2006) Predicting the effects of amino acid substitutions on protein function. Annual Review of Genomics and Human Genetics 7: 61–80.
34. Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30: 3894–3900.
35. Sunyaev S, Ramensky V, Bork P (2000) Towards a structural basis of human non-synonymous single nucleotide polymorphisms. Trends Genet 16: 198–200.
36. Sunyaev S, Ramensky V, Koch I, Lathe W, 3rd, Kondrashov AS, et al. (2001) Prediction of deleterious human alleles. Human Molecular Genetics 10: 591–597.
37. Gonzalez-Perez A, Lopez-Bigas N (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. American Journal of Human Genetics 88: 440–449.
38. Braman J, Papworth C, Greener A (1996) Site-directed mutagenesis using double-stranded plasmid DNA templates. Methods Mol Biol 57: 31–44.
39. Vindrola O, Lindberg I (1992) Biosynthesis of the prohormone convertase mPC1 in AtT-20 cells. Mol Endocrinol 6: 1088–1094.
40. Creemers JW, Choquet H, Stijnen P, Vatin V, Pigeyre M, et al. (2012) Heterozygous mutations causing partial prohormone convertase 1 deficiency contribute to human obesity. Diabetes 61: 383–390.
41. Zhou Y, Rovere C, Kitabgi P, Lindberg I (1995) Mutational analysis of PC1 (SPC3) in PC12 cells. 66-kDa PC1 is fully functional. J Biol Chem 270: 24702–24706.
42. Zhou A, Mains RE (1994) Endoproteolytic processing of proopiomelanocortin and prohormone convertases 1 and 2 in neuroendocrine cells overexpressing prohormone convertases 1 or 2. J Biol Chem 269: 17440–17447.
43. Bernard N, Kitabgi P, Rovere-Jovene C (2003) The Arg617–Arg618 cleavage site in the C-terminal domain of PC1 plays a major role in the processing and targeting of the enzyme within the regulated secretory pathway. J Neurochem 85: 1592–1603.
44. Rabah N, Gauthier D, Wilkes BC, Gauthier DJ, Lazure C (2006) Single amino acid substitution in the PC1/3 propeptide can induce significant modifications of its inhibitory profile toward its cognate enzyme. J Biol Chem 281: 7556–7567.
45. Goodman LJ, Gorman CM (1994) Autoproteolytic activation of the mouse prohormone convertase mPC1. Biochem Biophys Res Commun 201: 795–804.
46. Anderson ED, VanSlyke JK, Thulin CD, Jean F, Thomas G (1997) Activation of the furin endoprotease is a multiple-step process: requirements for acidification and internal propeptide cleavage. EMBO J 16: 1508–1518.
47. Creemers JW, Vey M, Schafer W, Ayoubi TA, Roebroek AJ, et al. (1995) Endoproteolytic cleavage of its propeptide is a prerequisite for efficient transport of furin out of the endoplasmic reticulum. J Biol Chem 270: 2695–2702.
48. Mains RE, Milgram SL, Keutman HT, Eipper BA (1995) The NH2-terminal proregion of peptidylglycine a-amidating monooxygenase facilitates the secretion of soluble proteins. Mol Endocrinol 9: 3–13.
49. Apletalina EV, Juliano MA, Juliano L, Lindberg I (2000) Structure-function analysis of the 7B2 CT peptide. Biochem Biophys Res Commun 267: 940–942.
50. Muller L, Cameron A, Fortenberry Y, Apletalina EV, Lindberg I (2000) Processing and sorting of the prohormone convertase 2 propeptide. J Biol Chem 275: 39213–39222.
51. Bissonnette L, Charest G, Longpre JM, Lavigne P, Leduc R (2004) Identification of furin pro-region determinants involved in folding and activation. Biochem J 379: 757–763.
52. Lee SN, Prodhomme E, Lindberg I (2004) Prohormone convertase 1 (PC1) processing and sorting: effect of PC1 propeptide and proSAAS. J Endocrinol 182: 353–364.
53. Tangrea MA, Bryan PN, Sari N, Orban J (2002) Solution structure of the pro-hormone convertase 1 pro-domain from Mus musculus. J Mol Biol 320: 801–812.
54. Benjannet S, Rondeau N, Paquet L, Boudreault A, Lazure C, et al. (1993) Comparative biosynthesis, covalent post-translational modifications and efficiency of prosegment cleavage of the prohormone convertases PC1 and PC2: glycosylation, sulphation and identification of the intracellular site of prosegment cleavage of PC1 and PC2. Biochem J 294 (Pt 3): 735–743.

## CHAPTER SIX

# Mutations in *ACTG2* are associated with sporadic congenital chronic intestinal pseudo-obstruction and megacystis-microcolon-intestinal hypoperistalsis syndrome

## ABSTRACT

Chronic intestinal pseudo-obstruction (CIPO) is a serious motility dysfunction syndrome characterized by symptoms of intestinal obstruction in the absence of any mechanical blockage. It is a major cause of intestinal failure. With whole-exome sequencing in a cohort of 20 patients with congenital CIPO or MMIH, we identified a subset of 10 cases with potentially damaging de-novo mutations at highly conserved loci in the *ACTG2* gene. In light of a recent finding that a mutation in *ACTG2* caused familial visceral myopathy in a Finnish family, we conclude that *ACTG2* also governs a significant proportion of cases of sporadic congenital CIPO.

## INTRODUCTION

Chronic intestinal pseudo-obstruction (CIPO) is a heterogenous set of diseases characterized by repetitive episodes or continuous symptoms of intestinal obstruction, in the absence of a lesion that occludes the lumen of the gut (1, 2). A small fraction of cases are secondary to organic, systemic, or metabolic diseases, but the majority are primary and may be myopathic, mesenchymopathic, or neuropathic, depending upon whether predominant abnormalities are found in the enteric nervous system, Interstitial Cells of Cajal (ICC), or intestinal smooth muscle (3). A related disorder, megacystis-microcolon-intestinal hypoperistalsis syndrome (MMIH), is characterized by constipation and urinary retention, microcolon, giant bladder (megacystis), intestinal hypoperistalis, hydronephrosis, and dilated small bowel (4).

Congenital forms of CIPO are rare and can be life-threatening; congenital CIPO is an important cause of intestinal failure, for which the only treatment may be complete visceral transplantation (5). Congenital CIPO may sometimes be due to prenatal exposure to toxins such as alcohol or narcotics. A handful of familial cases of CIPO have been reported with autosomal dominant (with variable penetrance), autosomal recessive, and X-linked modes of inheritance (6-16). It is well known that mutations in mitochondrial tRNA genes, *POLG* (polymerase (DNA directed), gamma), and *TYMP* (thymidine phosphorylase), which are expressed in the mitochondrion, cause a severe form of CIPO requiring frequent and long-term parenteral nutrition and with frequently fatal digestive and neurologic complications. Mitochondrial disorders may account for ~19% of CIPO cases (17). Contrawise, it is rare for CIPO to be the principal clinical manifestation of a mitochondrial disorder (18). Primary defects of the mitochondrial oxidative phosphorylation pathway are phenotypically heterogenous, and affecting multiple organs, typically with nervous system and skeletal or ocular muscle dysfunction (19). Mitochondrial neurogastrointestinal encephalomyopathy (MNGIE) is a rare, autosomal recessive syndrome due to the loss of thymidine phosphorylase activity associated with loss-of-function mutations in *TYM*P (20-24). Mutations in *POLG*, the mitochondrial myopathy, epilepsy, lactic acidosis, and strokelike episodes ('MELAS') mutation in the tRNA$^{leu(UUR)}$ gene (*MT-TL1*), or mutations in the tRNA$^{lys}$ gene (*MT-TK*) are sometimes associated with CIPO (16, 25-32). Still, congenital CIPO is usually sporadic and prior to the advent of exome sequencing, no non-mitochondrial gene had been convincingly associated with primary sporadic CIPO.

## METHODS

Samples were selected from 20 patients seen at UCLA, and were approved by our institutional review board. We chose 18 patients for exome sequencing; we sequenced only one of monozygotic twins in one family and a second patient entered the study after exome sequencing had concluded. Inclusion criteria were a diagnosis of chronic severe CIPO or MMIH. These patients were of diverse ethnic backgrounds and had clinical presentations of myopathic CIPO (*n*=6), neuropathic CIPO (*n*=5), idiopathic CIPO (*n*=3), and megacystis-microcolon-intestinal hypoperistalsis syndrome (MMIH) (*n*=4). The phenotype of several patients was complex with syndromic presentations of developmental delay (*n*=1), severe neurological problems (*n*=1), or microcephaly and mental retardation (*n*=2 siblings) [Table 1].

Genomic DNA from probands, and in some cases unaffected family members, was either fragmented by sonication and ligated to Illumina bar-coded adapters or fragmented and ligated in a single step with Illumina engineered transposases, and then in either case, the fragments were amplified by PCR. Fragments were then enriched for the protein coding portion of the genome by hybridization to probes from either the Agilent SureSelect XT Human All Exon 50Mb, Illumina TruSeq Exome, or Illumina Nextera Expanded Exome enrichment kits. The exome-enriched library was sequenced for 100x100 paired end reads on an Illumina Genome Analyzer 2000 or 2500 platform to a mean coverage depth of 97X, with 84% of RefGene CDS and essential splice sites having at least 20X coverage. Sanger sequencing confirmed candidate variants and segregation of the variant allele with the disorder in relatives from whom DNA was available.

We converted sequenced reads from the native bcl files to the FastQ format by the Illumina bcl2fastq program. We processed the FastQ files to create aligned bam files with in-house pipeline software. Briefly, we aligned the reads to build GRCh37 of the human genome (33) with Novoalign (http://www.novocraft.com) to obtain a mean of 93 million uniquely aligned 100x100 paired end reads per sample after removing PCR duplicates. We recalibrated base quality scores to improve accuracy by analyzing the covariation among reported quality score, position within read, dinucleotide, and probability of mismatching the reference genome using the Genome Analysis Toolkit (GATK) (34, 35). We used the GATK Unified Genotyper and Haplotype Caller tools to genotype single nucleotide variants and indels and recalibrated variant quality score recalibration with the GATK Variant Quality Score Recalibrator to assign probabilities to each variant call. We obtained the variant consequences on transcripts and proteins with the Ensembl Variant Effect

Predictor (VEP) (35), estimated the extent of protein damage with SIFT (36-40), PolyPhen (41-43), and Condel (44), and computed GERP conservation scores (45). We further annotated variants with additional data from Online Mendelian Inheritance in Man (OMIM) (46), the Human Gene Mutation Database (HGMD pro, BIOBASE Biological Databases), the Universal Protein Resource (UniProt) (47), KEGG Pathways (48), RefGene (49), the MitoCarta Inventory of Mammalian Mitochondrial Genes (50), Mouse Genome Informatics (MGI) (51) and the Human Protein Atlas (HPA) (52) using in-house plugins for the VEP.

We hypothesized for this study that the mode of inheritance would be recessive, de novo, or mitochondrial, with a dominant effect on phenotype and that mutation effects would be fully penetrant. Further hypothesizing that casual variants would be found in the protein coding or splicing regions of genes, we filtered to include only splice acceptor or donor, stop gained or lost, frameshift, initiator codon, inframe insertion or deletion, missense, or splice region variants. We removed variants that, given the rare incidence of CIPO, are too frequent in the population to cause this disorder; under the recessive model we removed variants with a minor allele frequency (MAF) >0.5% in the combined 1000 Genomes (53) and NHLBI (54) datasets and with a MAF of >0.1% under the de-novo model. In addition, we removed variants that were observed in more than 2 (homozygous and de-novo models) or 8 (compound heterozygous model) samples from ~150 unaffected control exomes, to eliminate false positives caused by technical artifacts. We ranked variants for likelihood of damage using multiple factors. We deemed a splice acceptor or donor variant, stop gained, and frameshift, to be probably damaging. We prioritized missense single nucleotide variants (SNVs) by SIFT, PolyPhen, and Condel predictions, and variants in the splice site region near the acceptor and donor by GERP conservation scores. In the case of compound heterozygous variants we assigned the priority of the second ranked variant. In families where we had sequenced members other than the proband, we filtered out all variants that were inconsistent with fully penetrant Mendelian inheritance under the recessive model or present in any unaffected relative under the de-novo model. We hypothesized that there would be genotypic heterogeneity among the cases, so we looked for cases that shared predicted damaging mutations in the same gene.

## RESULTS

The probands had a mean of ~93,000 variations from the GRCh37 reference genome before filtering. After filtering by consequence rank, allele frequency, and controls and eliminating variants that did not segregate with the disease, the probands had a mean of 85 variants consistent with either the recessive or de-novo models. Looking for genes that had de-novo mutations in multiple cases, our attention focused on *ACTG2* (encoding actin, gamma-enteric smooth muscle precursor), in which there were seven different de-novo mutations in seven cases and no controls. *ACTG2* was of interest because of actin's involvement in muscle function (55) and particularly because γ-enteric actin is the dominant isoform of actin expressed in the jejunum (56). We performed Sanger sequencing on *ACTG2* exons in all cases and found two additional mutations, one that had been missed in exome sequencing and one in a case added to the study after exome sequencing was complete, as well as confirming the presence of a mutation in the unsequenced twin [Tables 2 & 3]. Each of the mutations was predicted to be deleterious by the Condel consensus statistic based on PolyPhen and SIFT and all were at highly conserved loci with GERP scores ≥3.59. All variants were confirmed by Sanger sequencing, and were not found in unaffected relatives. No variant was found in ~150 control exomes, nor in the 1000 Genomes or NHLBI datasets.

DISCUSSION

During the course of our study Lehtonen *et al.* reported that a missense variant (Arg148Ser) in *ACTG* segregated in a Finnish family with autosomal dominant familial visceral myopathy (FVM), a disorder that is subsumed within the broad definition of CIPO (57). Although none of the mutations we identified overlapped with the FVM mutations, our results along with the FVM finding, make a strong case that mutations in γ-enteric actin have a profound effect on intestinal motility and may be a predominant cause of sporadic and familial intestinal failure due to pseudo-obstruction.

Nonetheless, *ACTG* mutations explain only 50% of the cases we examined, suggesting that CIPO must be genetically heterogeneous. Moreover, we found *ACTG2* mutations in patients initially suspected to be myopathic, neuropathic, or having MIMH syndrome. This may be explained in part by the challenges in diagnosing this class of disorders and in part by phenotypic heterogeneity. There may also be identifiable genotype-phenotype correlations, which may become apparent when a larger number of cases are genotyped. Interestingly, none of the cases in our study with syndromic CIPO had detectable mutations in *ACTG2*.

A large study ($n$=115) observed absent or partial staining for smooth muscle α-actin in the jejunal circular muscle layer of in 24% of CIPO patients (58). This report, which found no differences in staining of γ-actin, may indicate a different molecular basis for another class of patients than those with γ- actin mutations. It is also possible that the near sequence identity of α- and γ-actin made it difficult to distinguish the isoforms, even though the antibodies used purported to be specific. Another study also found α-actin deficiencies in CIPO patient tissue (59), but that finding was not replicated (60).

The molecular mechanism by which disrupted actin affects enteric smooth muscle cells remains to be discovered. Now that ten different probably disease-causing *ACTG2* mutations have been identified it will be interesting to discover how each of these variants affects the cell.

Mutations in other actin isoforms at homologous loci to those reported here can cause disease. The *ACTA2* Arg258Cys mutation, homologous to *ACTG2* Arg257Cys, causes thoracic aneurisms and dissections (61); *ACTA1* Arg258His causes nemaline myopathy (62); and *ACTA2* Arg258His causes thoracic aortic disease & strokes (63). Several other loci when mutated in *ACTA1* or *ACTA2* to different amino acids than those reported here also cause disease [Table 3] (64-68). This suggests the possibility of common molecular mechanisms at work, which manifest variously in different tissues where other actin isoforms are expressed. However, determining the these mechanisms will be challenging (69).

Tables

| ID | Sex | Population | Predicted Disease | Sequenced |
|---|---|---|---|---|
| 1A | F | EA/Arab | CIPO | p,f,m |
| 2A | F | EA | CIPO, myopathic | p |
| 3A | F | MA | CIPO | p,f,m |
| 4A | M | MA | CIPO, myopathic | p |
| 5A1 | F | EA | CIPO; mycocephaly; MR | p,s |
| 5A2 | f | EA | CIPO; mycocephaly; MR | p,s |
| 6A | M | AA | CIPO | p |
| 7A | F | AA | MMIH | p,m |
| 8A | F | EA | CIPO, neuropathic | p |
| 9A | M | EA | CIPO, myopathic | p,m |
| 10A | M | MA | CIPO, myopathic; developmental delay | p,m |
| 12A | M | MA | CIPO, neuropathic | p |
| 13A | F | EA/Asian | CIPO, neuropathic; severe neurologic problems | p,f,m |
| 14A | M | EA | CIPO, neuropathic | p |
| 15A | F | EA | CIPO, myopathic | p |
| 16A | F | MA | MMIH | p |
| 17A | M | MA | CIPO, neuropathic | p,f,m |
| 19A1 | M | MA | MMIH | S |
| 19A2 | M | MA | MMIH | p |
| 23A | M | EA | CIPO, myopathic | S |

**Table 1. Subjects.** Population: EA, Eurpean American; MA, Mexican American. Phenotype: CIPO, chronic intestinal pseudo-obstruction; MMIH, megacystis-microcolon-intestinal hypoperistalsis syndrome; MR, mental retardation. Sequenced: p,proband; f, father; m, mother; s, affected sibling; S, Sanger sequencing only.

| ID | Age (yrs) | GI Symptoms | Sugeries | Disease Status | Addtional Information | Pathology Findings |
|---|---|---|---|---|---|---|
| 2A | 36 | Abdominal pain, portal hypertension, bleeding, jaundice, edema | OLT/SBT/PT | Regular diet? Post transplant course complicated by high output, neurogenic bladder with hydronephrosis, multiple line infections, and bladder infections. | Neurogenic bladder with hydronephrosis | Spotty degeneration of muscle fibers. Disorganization of muscle fibers. Nerves are prominent and ganlions are present. |
| 4A | 23 | Abdominal distention, high ostomy outputs | Colectomy, jejunostomy, GT | TPN support | | ileostomy specimens: no decriptions available |
| 7A | 7 | FTT, TPN dependent with IFALD, chronic cholestasis, unable to tolerate feeds via GT. | GJ tube, ileostomy, subtotal gastrectomy, h/o vesicostomy now closed, splenectomy and OLT/SBT/PT | Full enteral feeds | neurogenic bladder requiring catheterization | Muscularis propria present,  no vacuolization. External layer of muscularis propria slightly thinned out, but not hypoplastic. Ganglion cells are present. |
| 9A | 10 | Abdominal distention with increased outputs. Multiple line infections. | Malrotation s/p repair, Nissen/GT, total colectomy and ileostomy with multiple revisions, broviac placement with multiple line infections, lap cholecystectomy, h/o vesicostomy | TPN support, increased outputs, multiple line infections, bladder involvement with multiple UTIs. | Neurogenic bladder with bilateral hydronephrosis | Distal ileum: Ganglion cells present within the submucosa and muscularis propria |

| ID | Age (yrs) | GI Symptoms | Sugeries | Disease Status | Addtional Information | Pathology Findings |
|---|---|---|---|---|---|---|
| 14A | 8 | Retained meconium with constipation and abdominal distention. Then developed persistent diarrhea, previously able to tolerate enteral feeds, but signif decreased since enteritis in 2010. | Nissen/GT, colectomy with ileoanal pull through, tempoary gastric stimulator | TPN support, bladder involvment, temporary gastric stimulator without positive clinical response | Mother was 36yo G5P2. Older sibling was born46XY with ambiguous genitalia, hypoplastic left kidney, adrenal hypoplasia, normal facies and passed away at 6weeks of age. | none available |
| 15A | 32 | Abdominal distention, constipation, cholelithiasis | Total colectomy with ileostomy, cholecystecomy , hernia repair, GJ tube | TPN support | History of atonic bladder at birth requiring catheterizations in the neonatal period. | **Small intestine:** thinning of muscularis propria, vacuolar degeneration of muscle c/w myopathy. **Colon:** vaculolar degeneration of myentericplexus cells and mild atrophy of muscularis. **Liver:** mild centrilobular steatosis. **Electron microscopy:** smooth muscle coat. |
| 16A | 28 | Abdominal distention, nausea, diarrhea, large GT outputs | Ex-lap at birth, repair of malrotation, GT, vesicostomy. | TPN support, depression, chronic pain management issues. | Neurogenic bladder. H/o AD manometry at Cedars by Dr Hyman (result?). Hemolytic anemia with chronic pancytopenia. | consistent with visceral myopathy. Ganglion cells present. |
| 19A 1 | 6 | TPN dependent with intestinal failure s/p transplant, complicated by graft rejection with total enterectomy, adenovirus infection, and PTLD diffuse large B cell. | Repair of malrotation at birth, GT, multivisceral transplant (liver, colon, SB, pancreas) now s/p subtotal enterectomy | Listed for re-transplant | ex twin | IHC: ckit positive with interstitial cells of cajal present. SB: thin walled |

| ID | Age (yrs) | GI Symptoms | Sugeries | Disease Status | Addtional Information | Pathology Findings |
|---|---|---|---|---|---|---|
| 19A 2 | 6 | TPN dependent with IFALD, listed for mulativisceral transplant. | Listed for SBT/OLT. H/o vesicostomy , malrotatin s/p repair, GJ tube. | TPN support, bladder involvement. Currently listed for multivisceral transplant | ex twin | Appendix: mild architectural dissarray with ganglion cells present |
| 23A | 16 | diffuse abdominal pain and intermittent vomiting with episodes of constipation and recurrent bacterial overgrowth | Colostomy with revision, hiatal hernia repair | Tolerating feeds, colostomy functioning well, bladder involvement requiring catherterizatio n. | Ex 31 wk premie,neurogenic bladder requires catherterization daily. H/o intrauterine bladder stent surgery at 24 weeks. Father s/p pancreatic transplant. | none available |

**Table 2. Patients with ACTG2 mutations.**

| **ID** | ***ACTG2* Mutation** | **Condel Prediction** | **GERP Score** | **Other Mutations** |
|---|---|---|---|---|
| 2A | Gln138Lys | deleterious | 4.65 | *ACTA1: Gln139His Nemaline myopathy (PMID 18461503)* |
| 4A | Gly37Ser | deleterious | 4.57 | *ACTA1: Gly38Ala Nemaline myopathy (PMID 19562689); ACTA2: Gly38Arg Thoracic aortic aneurysms and dissections (PMID 21937134)* |
| 7A | Arg178Ser | deleterious | 4.65 | |
| 9A | Arg211Cys | deleterious | 4.65 | |
| 14A | Arg257Cys | deleterious | 4.65 | *ACTA2*: Arg258Cys Thoracic aortic aneurysms and dissections (PMID 17994018) |
| 15A | Arg257His | deleterious | 4.65 | *ACTA1*: Arg258His Nemaline myopathy (PMID 10508519);<br>*ACTA2*: Arg258His<br>Thoracic aortic disease & strokes (PMID 19409525) |
| 16A | Arg40Cys | deleterious | 3.59 | *ACTA1: Arg41Term Nemaline myopathy (PMID 12921789)* |
| 19A1 | Gly198Asp | deleterious | 4.65 | *ACTA1: Gly199Ser Nemaline myopathy (PMID 15236405)* |
| 19A2 | Gly198Asp | deleterious | | *ACTA1: Gly199Ser Nemaline myopathy (PMID 15236405)* |
| 23A | Arg40His | deleterious | 4.65 | *ACTA1: Arg41Term Nemaline myopathy (PMID 12921789)* |

**Table 3. Sequencing results.** Condel: Consnesus of PolyPhen and SIFT protein damage predictions. GERP: Conservation score. Other mutations: Disease-causing mutations reported in homologues of *ACTG2*; those in italics are at the same locus as the *ACTG2* mutation, but with a different amino acid substitution.

# References

1. Antonucci A, Fronzoni L, Cogliandro L, Cogliandro RF, Caputo C, De Giorgio R, Pallotti F, Barbara G, Corinaldesi R, Stanghellini V. Chronic intestinal pseudo-obstruction. World journal of gastroenterology : WJG. 2008;14(19):2953-61. PubMed PMID: 18494042; PubMed Central PMCID: PMC2712158.

2. Hyman P, Thapar N. Chronic Intestinal Pseudo-Obstruction. In: Faure C, Di Lorenzo C, Thapar N, editors. Pediatric Neurogastroenterology: Humana Press; 2013. p. 257-70.

3. Gisser J, Gariepy C. Genetics of Motility Disorder: Gastroesophageal Reflux, Triple A Syndrome, Hirschsprung Disease, and Chronic Intestinal Pseudo-Obstruction. In: Faure C, Di Lorenzo C, Thapar N, editors. Pediatric Neurogastroenterology: Humana Press; 2013. p. 203-16.

4. Berdon WE, Baker DH, Blanc WA, Gay B, Santulli TV, Donovan C. Megacystis-microcolon-intestinal hypoperistalsis syndrome: a new cause of intestinal obstruction in the newborn. Report of radiologic findings in five newborn girls. AJR American journal of roentgenology. 1976;126(5):957-64. Epub 1976/05/01. doi: 10.2214/ajr.126.5.957. PubMed PMID: 178239.

5. Stanghellini V, Cogliandro RF, De Giorgio R, Barbara G, Cremon C, Antonucci A, Fronzoni L, Cogliandro L, Naponelli V, Serra M, Corinaldesi R. Natural history of intestinal failure induced by chronic idiopathic intestinal pseudo-obstruction. Transplantation proceedings. 2010;42(1):15-8. doi: 10.1016/j.transproceed.2009.12.017. PubMed PMID: 20172271.

6. Anuras S, Mitros FA, Nowak TV, Ionasescu VV, Gurll NJ, Christensen J, Green JB. A familial visceral myopathy with external ophthalmoplegia and autosomal recessive transmission. Gastroenterology. 1983;84(2):346-53. Epub 1983/02/01. PubMed PMID: 6687359.

7. Auricchio A, Brancolini V, Casari G, Milla PJ, Smith VV, Devoto M, Ballabio A. The locus for a novel syndromic form of neuronal intestinal pseudoobstruction maps to Xq28. Am J Hum Genet. 1996;58(4):743-8. Epub 1996/04/01. PubMed PMID: 8644737; PubMed Central PMCID: PMC1914695.

8. Faulk DL, Anuras S, Gardner GD, Mitros FA, Summers RW, Christensen J. A familial visceral myopathy. Annals of internal medicine. 1978;89(5 Pt 1):600-6. Epub 1978/11/01. PubMed PMID: 717927.

9. Ionasescu V, Thompson SH, Ionasescu R, Searby C, Anuras S, Christensen J, Mitros F, Hart M, Bosch P. Inherited ophthalmoplegia with intestinal pseudo-obstruction. Journal of the neurological sciences. 1983;59(2):215-28. Epub 1983/05/01. PubMed PMID: 6687898.

10. Mayer EA, Schuffler MD, Rotter JI, Hanna P, Mogard M. Familial visceral neuropathy with autosomal dominant transmission. Gastroenterology. 1986;91(6):1528-35. Epub 1986/12/01. PubMed PMID: 3770377.

11. Patel H, Norman MG, Perry TL, Berry KE. Multiple system atrophy with neuronal intranuclear hyaline inclusions. Report of a case and review of the literature. Journal of the neurological sciences. 1985;67(1):57-65. Epub 1985/01/01. PubMed PMID: 2580060.

12. Roy AD, Bharucha H, Nevin NC, Odling-Smee GW. Idiopathic intestinal pseudo-obstruction: a familial visceral neuropathy. Clin Genet. 1980;18(4):291-7. Epub 1980/10/01. PubMed PMID: 7438508.

13. Schuffler MD, Bird TD, Sumi SM, Cook A. A familial neuronal disease presenting as intestinal pseudoobstruction. Gastroenterology. 1978;75(5):889-98. Epub 1978/11/01. PubMed PMID: 212342.

14. Schuffler MD, Lowe MC, Bill AH. Studies of idiopathic intestinal pseudoobstruction. I. Hereditary hollow visceral myopathy: clinical and pathological studies. Gastroenterology. 1977;73(2):327-38. Epub 1977/08/01. PubMed PMID: 873134.

15. Schuffler MD, Pope CE, 2nd. Studies of idiopathic intestinal pseudoobstruction. II. Hereditary hollow visceral myopathy: family studies. Gastroenterology. 1977;73(2):339-44. Epub 1977/08/01. PubMed PMID: 873135.

16. Van Goethem G, Schwartz M, Lofgren A, Dermaut B, Van Broeckhoven C, Vissing J. Novel POLG mutations in progressive external ophthalmoplegia mimicking mitochondrial neurogastrointestinal encephalomyopathy. Eur J Hum Genet. 2003;11(7):547-9. Epub 2003/06/26. doi: 10.1038/sj.ejhg.5201002. PubMed PMID: 12825077.

17. Amiot A, Tchikviladze M, Joly F, Slama A, Hatem DC, Jardel C, Messing B, Lombes A. Frequency of mitochondrial defects in patients with chronic intestinal pseudo-obstruction. Gastroenterology. 2009;137(1):101-9. Epub 2009/04/07. doi: 10.1053/j.gastro.2009.03.054. PubMed PMID: 19344718.

18. Chinnery PF, Turnbull DM. Clinical features, investigation, and management of patients with defects of mitochondrial DNA. J Neurol Neurosurg Psychiatry. 1997;63(5):559-63. Epub 1998/01/04. PubMed PMID: 9408091; PubMed Central PMCID: PMC2169824.

19. DiMauro S. Mitochondrial diseases. Biochimica et biophysica acta. 2004;1658(1-2):80-8. Epub 2004/07/30. doi: 10.1016/j.bbabio.2004.03.014. PubMed PMID: 15282178.

20. Blondon H, Polivka M, Joly F, Flourie B, Mikol J, Messing B. Digestive smooth muscle mitochondrial myopathy in patients with mitochondrial-neuro-gastro-intestinal encephalomyopathy (MNGIE). Gastroenterologie clinique et biologique. 2005;29(8-9):773-8. Epub 2005/11/19. PubMed PMID: 16294144.

21. Hirano M, Marti R, Spinazzola A, Nishino I, Nishigaki Y. Thymidine phosphorylase deficiency causes MNGIE: an autosomal recessive mitochondrial disorder. Nucleosides, nucleotides & nucleic acids. 2004;23(8-9):1217-25. Epub 2004/12/02. doi: 10.1081/NCN-200027485. PubMed PMID: 15571233.

22. Hirano M, Nishino I, Nishigaki Y, Marti R. Thymidine phosphorylase gene mutations cause mitochondrial neurogastrointestinal encephalomyopathy (MNGIE). Internal medicine. 2006;45(19):1103. Epub 2006/11/02. PubMed PMID: 17077575.

23. Nishino I, Spinazzola A, Hirano M. Thymidine phosphorylase gene mutations in MNGIE, a human mitochondrial disorder. Science. 1999;283(5402):689-92. Epub 1999/01/29. PubMed PMID: 9924029.

24. Slama A, Lacroix C, Plante-Bordeneuve V, Lombes A, Conti M, Reimund JM, Auxenfants E, Crenn P, Laforet P, Joannard A, Seguy D, Pillant H, Joly P, Haut S, Messing B, Said G, Legrand A, Guiochon-Mantel A. Thymidine phosphorylase gene mutations in patients with mitochondrial neurogastrointestinal encephalomyopathy syndrome. Molecular genetics and metabolism. 2005;84(4):326-31. Epub 2005/03/23. doi: 10.1016/j.ymgme.2004.12.004. PubMed PMID: 15781193.

25. Chang TM, Chi CS, Tsai CR, Lee HF, Li MC. Paralytic ileus in MELAS with phenotypic features of MNGIE. Pediatric neurology. 2004;31(5):374-7. Epub 2004/11/03. doi: 10.1016/j.pediatrneurol.2004.05.009. PubMed PMID: 15519124.

26. Chinnery PF, Jones S, Sviland L, Andrews RM, Parsons TJ, Turnbull DM, Bindoff LA. Mitochondrial enteropathy: the primary pathology may not be within the gastrointestinal tract. Gut. 2001;48(1):121-4. Epub 2000/12/15. PubMed PMID: 11115833; PubMed Central PMCID: PMC1728165.

27. Filosto M, Mancuso M, Nishigaki Y, Pancrudo J, Harati Y, Gooch C, Mankodi A, Bayne L, Bonilla E, Shanske S, Hirano M, DiMauro S. Clinical and genetic heterogeneity in progressive external ophthalmoplegia due to mutations in polymerase gamma. Archives of neurology. 2003;60(9):1279-84. Epub 2003/09/17. doi: 10.1001/archneur.60.9.1279. PubMed PMID: 12975295.

28. Garcia-Velasco A, Gomez-Escalonilla C, Guerra-Vales JM, Cabello A, Campos Y, Arenas J. Intestinal pseudo-obstruction and urinary retention: cardinal features of a mitochondrial DNA-related disease. Journal of internal medicine. 2003;253(3):381-5. Epub 2003/02/27. PubMed PMID: 12603507.

29. Li JY, Kong KW, Chang MH, Cheung SC, Lee HC, Pang CY, Wei YH. MELAS syndrome associated with a tandem duplication in the D-loop of mitochondrial DNA. Acta neurologica Scandinavica. 1996;93(6):450-5. Epub 1996/06/01. PubMed PMID: 8836308.

30. Mancuso M, Filosto M, Oh SJ, DiMauro S. A novel polymerase gamma mutation in a family with ophthalmoplegia, neuropathy, and Parkinsonism. Archives of neurology. 2004;61(11):1777-9. Epub 2004/11/10. doi: 10.1001/archneur.61.11.1777. PubMed PMID: 15534189.

31. Tanji K, Gamez J, Cervera C, Mearin F, Ortega A, de la Torre J, Montoya J, Andreu AL, DiMauro S, Bonilla E. The A8344G mutation in mitochondrial DNA associated with stroke-like episodes and gastrointestinal dysfunction. Acta neuropathologica. 2003;105(1):69-75. Epub 2002/12/10. doi: 10.1007/s00401-002-0604-y. PubMed PMID: 12471464.

32. Van Goethem G, Luoma P, Rantamaki M, Al Memar A, Kaakkola S, Hackman P, Krahe R, Lofgren A, Martin JJ, De Jonghe P, Suomalainen A, Udd B, Van Broeckhoven C. POLG mutations in neurodegenerative disorders with ataxia but no muscle involvement. Neurology. 2004;63(7):1251-7. Epub 2004/10/13. PubMed PMID: 15477547.

33. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931-45. Epub 2004/10/22. doi: 10.1038/nature03001. PubMed PMID: 15496913.

34. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491-8. Epub 2011/04/12. doi: 10.1038/ng.806. PubMed PMID: 21478889; PubMed Central PMCID: PMC3083463.

35. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-303. Epub 2010/07/21. doi: 10.1101/gr.107524.110. PubMed PMID: 20644199; PubMed Central PMCID: PMC2928508.

36. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols. 2009;4(7):1073-81. Epub 2009/06/30. doi: 10.1038/nprot.2009.86. PubMed PMID: 19561590.

37. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001;11(5):863-74. Epub 2001/05/05. doi: 10.1101/gr.176601. PubMed PMID: 11337480; PubMed Central PMCID: PMC311071.

38. Ng PC, Henikoff S. Accounting for human polymorphisms predicted to affect protein function. Genome Res. 2002;12(3):436-46. Epub 2002/03/05. doi: 10.1101/gr.212802. PubMed PMID: 11875032; PubMed Central PMCID: PMC155281.

39. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812-4. Epub 2003/06/26. PubMed PMID: 12824425; PubMed Central PMCID: PMC168916.

40. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. Annual review of genomics and human genetics. 2006;7:61-80. Epub 2006/07/11. doi: 10.1146/annurev.genom.7.080505.115630. PubMed PMID: 16824020.

41. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 2002;30(17):3894-900. Epub 2002/08/31. PubMed PMID: 12202775; PubMed Central PMCID: PMC137415.

42. Sunyaev S, Ramensky V, Bork P. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. Trends Genet. 2000;16(5):198-200. Epub 2000/04/27. PubMed PMID: 10782110.

43. Sunyaev S, Ramensky V, Koch I, Lathe W, 3rd, Kondrashov AS, Bork P. Prediction of deleterious human alleles. Hum Mol Genet. 2001;10(6):591-7. Epub 2001/03/07. PubMed PMID: 11230178.

44. Gonzalez-Perez A, Lopez-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet.

2011;88(4):440-9. Epub 2011/04/05. doi: 10.1016/j.ajhg.2011.03.004. PubMed PMID: 21457909; PubMed Central PMCID: PMC3071923.

45. Cooper GM, Brudno M, Program NCS, Green ED, Batzoglou S, Sidow A. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. Genome Res. 2003;13(5):813-20. Epub 2003/05/03. doi: 10.1101/gr.1064503. PubMed PMID: 12727901; PubMed Central PMCID: PMC430923.

46. Online Mendelian Inheritance in Man OMIM®. Online Mendelian Inheritance in Man, OMIM® Baltimore, MD: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University; [2013-11-04]. Available from: http://omim.org/.

47. Apweiler R, Martin MJ, O'Donovan C, Magrane M, Alam-Faruque Y, Alpi E, Antunes R, Arganiska J, Casanova EB, Bely B, Bingley M, Bonilla C, Britto R, Bursteinas B, Chan WM, Chavali G, Cibrian-Uhalte E, Da Silva A, De Giorgi M, Dimmer E, Fazzini F, Gane P, Fedotov A, Castro LG, Garmiri P, Hatton-Ellis E, Hieta R, Huntley R, Jacobsen J, Jones R, Legge D, Liu WD, Luo J, MacDougall A, Mutowo P, Nightingale A, Orchard S, Patient S, Pichler K, Poggioli D, Pundir S, Pureza L, Qi GY, Rosanoff S, Sawford T, Sehra H, Turner E, Volynkin V, Wardell T, Watkins X, Zellner H, Corbett M, Donnelly M, van Rensburg P, Goujon M, McWilliam H, Lopez R, Xenarios I, Bougueleret L, Bridge A, Poux S, Redaschi N, Auchincloss A, Axelsen K, Bansal P, Baratin D, Binz PA, Blatter MC, Boeckmann B, Bolleman J, Boutet E, Breuza L, de Castro E, Cerutti L, Coudert E, Cuche B, Doche M, Dornevil D, Duvaud S, Estreicher A, Famiglietti L, Feuermann M, Gasteiger E, Gehant S, Gerritsen V, Gos A, Gruaz-Gumowski N, Hinz U, Hulo C, James J, Jungo F, Keller G, Lara V, Lemercier P, Lew J, Lieberherr D, Martin X, Masson P, Morgat A, Neto T, Paesano S, Pedruzzi I, Pilbout S, Pozzato M, Pruess M, Rivoire C, Roechert B, Schneider M, Sigrist C, Sonesson K, Staehli S, Stutz A, Sundaram S, Tognolli M, Verbregue L, Veuthey AL, Zerara M, Wu CH, Arighi CN, Arminski L, Chen CM, Chen YX, Huang HZ, Kukreja A, Laiho K, McGarvey P, Natale DA, Natarajan TG, Roberts NV, Suzek BE, Vinayaka CR, Wang QH, Wang YQ, Yeh LS, Yerramalla MS, Zhang J, Consortium U. Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Research. 2013;41(D1):D43-D7. doi: Doi 10.1093/Nar/Gks1068. PubMed PMID: WOS:000312893300007.

48. Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. Nucleic Acids Res. 2002;30(1):42-6. PubMed PMID: 11752249; PubMed Central PMCID: PMC99091.

49. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. 2012;40(Database issue):D130-5. doi: 10.1093/nar/gkr1079. PubMed PMID: 22121212; PubMed Central PMCID: PMC3245008.

50. Pagliarini DJ, Calvo SE, Chang B, Sheth SA, Vafai SB, Ong SE, Walford GA, Sugiana C, Boneh A, Chen WK, Hill DE, Vidal M, Evans JG, Thorburn DR, Carr SA, Mootha VK. A mitochondrial protein compendium elucidates complex I disease biology. Cell. 2008;134(1):112-23. doi: 10.1016/j.cell.2008.06.016. PubMed PMID: 18614015; PubMed Central PMCID: PMC2778844.

51. Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE, Mouse Genome Database G. The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the

laboratory mouse. Nucleic Acids Res. 2012;40(Database issue):D881-6. doi: 10.1093/nar/gkr974. PubMed PMID: 22075990; PubMed Central PMCID: PMC3245042.

52. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S, Wernerus H, Bjorling L, Ponten F. Towards a knowledge-based Human Protein Atlas. Nat Biotechnol. 2010;28(12):1248-50. doi: 10.1038/nbt1210-1248. PubMed PMID: 21139605.

53. 1000 Genomes Project. VCF (Variant Call Format) version 4.1 2013 [updated 2013-10-09]. Available from: http://www.1000genomes.org/wiki/Analysis/Variant Call Format/vcf-variant-call-format-version-41.

54. Exome Variant Server [Internet]. NHLBI Exome Sequencing Project (ESP). 2011 [cited 2011-09-10]. Available from: http://evs.gs.washington.edu/EVS/.

55. Dominguez R, Holmes KC. Actin structure and function. Annual review of biophysics. 2011;40:169-86. doi: 10.1146/annurev-biophys-042910-155359. PubMed PMID: 21314430; PubMed Central PMCID: PMC3130349.

56. Fatigati V, Murphy RA. Actin and tropomyosin variants in smooth muscles. Dependence on tissue type. The Journal of biological chemistry. 1984;259(23):14383-8. Epub 1984/12/10. PubMed PMID: 6501298.

57. Lehtonen HJ, Sipponen T, Tojkander S, Karikoski R, Jarvinen H, Laing NG, Lappalainen P, Aaltonen LA, Tuupanen S. Segregation of a missense variant in enteric smooth muscle actin gamma-2 with autosomal dominant familial visceral myopathy. Gastroenterology. 2012;143(6):1482-91 e3. doi: 10.1053/j.gastro.2012.08.045. PubMed PMID: 22960657.

58. Knowles CH, Silk DB, Darzi A, Veress B, Feakins R, Raimundo AH, Crompton T, Browning EC, Lindberg G, Martin JE. Deranged smooth muscle alpha-actin as a biomarker of intestinal pseudo-obstruction: a controlled multinational case series. Gut. 2004;53(11):1583-9. doi: 10.1136/gut.2003.037275. PubMed PMID: 15479676; PubMed Central PMCID: PMC1774262.

59. Smith VV, Lake BD, Kamm MA, Nicholls RJ. Intestinal pseudo-obstruction with deficient smooth muscle alpha-actin. Histopathology. 1992;21(6):535-42. Epub 1992/12/01. PubMed PMID: 1468752.

60. Gamba E, Carr NJ, Bateman AC. Deficient alpha smooth muscle actin expression as a cause of intestinal pseudo-obstruction: fact or fiction? Journal of clinical pathology. 2004;57(11):1168-71. doi: 10.1136/jcp.2003.015297. PubMed PMID: 15509678; PubMed Central PMCID: PMC1770492.

61. Guo DC, Pannu H, Tran-Fadulu V, Papke CL, Yu RK, Avidan N, Bourgeois S, Estrera AL, Safi HJ, Sparks E, Amor D, Ades L, McConnell V, Willoughby CE, Abuelo D, Willing M, Lewis RA, Kim DH, Scherer S, Tung PP, Ahn C, Buja LM, Raman CS, Shete SS, Milewicz DM. Mutations in smooth muscle alpha-actin (ACTA2) lead to thoracic aortic aneurysms and dissections. Nat Genet. 2007;39(12):1488-93. Epub 2007/11/13. doi: 10.1038/ng.2007.6. PubMed PMID: 17994018.

62. Nowak KJ, Wattanasirichaigoon D, Goebel HH, Wilce M, Pelin K, Donner K, Jacob RL, Hubner C, Oexle K, Anderson JR, Verity CM, North KN, Iannaccone ST, Muller CR, Nurnberg P, Muntoni F, Sewry C, Hughes I, Sutphen R, Lacson AG, Swoboda KJ, Vigneron J, Wallgren-Pettersson C, Beggs AH, Laing NG. Mutations in the skeletal muscle alpha-actin gene in patients with actin myopathy and nemaline myopathy. Nat Genet. 1999;23(2):208-12. doi: 10.1038/13837. PubMed PMID: 10508519.

63. Guo DC, Papke CL, Tran-Fadulu V, Regalado ES, Avidan N, Johnson RJ, Kim DH, Pannu H, Willing MC, Sparks E, Pyeritz RE, Singh MN, Dalman RL, Grotta JC, Marian AJ, Boerwinkle EA, Frazier LQ, LeMaire SA, Coselli JS, Estrera AL, Safi HJ, Veeraraghavan S, Muzny DM, Wheeler DA, Willerson JT, Yu RK, Shete SS, Scherer SE, Raman CS, Buja LM, Milewicz DM. Mutations in smooth muscle alpha-actin (ACTA2) cause coronary artery disease, stroke, and Moyamoya disease, along with thoracic aortic disease. Am J Hum Genet. 2009;84(5):617-27. Epub 2009/05/05. doi: 10.1016/j.ajhg.2009.04.007. PubMed PMID: 19409525; PubMed Central PMCID: PMC2680995.

64. Koy A, Ilkovski B, Laing N, North K, Weis J, Neuen-Jacob E, Mayatepek E, Voit T. Nemaline myopathy with exclusively intranuclear rods and a novel mutation in ACTA1 (Q139H). Neuropediatrics. 2007;38(6):282-6. Epub 2008/05/08. doi: 10.1055/s-2008-1065356. PubMed PMID: 18461503.

65. Laing NG, Dye DE, Wallgren-Pettersson C, Richard G, Monnier N, Lillis S, Winder TL, Lochmuller H, Graziano C, Mitrani-Rosenbaum S, Twomey D, Sparrow JC, Beggs AH, Nowak KJ. Mutations and polymorphisms of the skeletal muscle alpha-actin gene (ACTA1). Hum Mutat. 2009;30(9):1267-77. Epub 2009/06/30. doi: 10.1002/humu.21059. PubMed PMID: 19562689; PubMed Central PMCID: PMC2784950.

66. Renard M, Callewaert B, Baetens M, Campens L, MacDermot K, Fryns JP, Bonduelle M, Dietz HC, Gaspar IM, Cavaco D, Stattin EL, Schrander-Stumpel C, Coucke P, Loeys B, De Paepe A, De Backer J. Novel MYH11 and ACTA2 mutations reveal a role for enhanced TGFbeta signaling in FTAAD. International journal of cardiology. 2013;165(2):314-21. Epub 2011/09/23. doi: 10.1016/j.ijcard.2011.08.079. PubMed PMID: 21937134; PubMed Central PMCID: PMC3253210.

67. Sparrow JC, Nowak KJ, Durling HJ, Beggs AH, Wallgren-Pettersson C, Romero N, Nonaka I, Laing NG. Muscle disease caused by mutations in the skeletal muscle alpha-actin gene (ACTA1). Neuromuscular disorders : NMD. 2003;13(7-8):519-31. Epub 2003/08/19. PubMed PMID: 12921789.

68. Agrawal PB, Strickland CD, Midgett C, Morales A, Newburger DE, Poulos MA, Tomczak KK, Ryan MM, Iannaccone ST, Crawford TO, Laing NG, Beggs AH. Heterogeneity of nemaline myopathy cases with skeletal muscle alpha-actin gene mutations. Ann Neurol. 2004;56(1):86-96. Epub 2004/07/06. doi: 10.1002/ana.20157. PubMed PMID: 15236405.

69. Rubenstein PA, Mayer EA. Familial visceral myopathies: from symptom-based syndromes to actin-related diseases. Gastroenterology. 2012;143(6):1420-3. doi: 10.1053/j.gastro.2012.10.031. PubMed PMID: 23085350.

# Chapter Seven

# Mutations in the RNA exosome component gene *EXOSC3* cause pontocerebellar hypoplasia and spinal motor neuron degeneration

# Mutations in the RNA exosome component gene *EXOSC3* cause pontocerebellar hypoplasia and spinal motor neuron degeneration

Jijun Wan[1,24], Michael Yourshaw[2,24], Hafsa Mamsa[1], Sabine Rudnik-Schöneborn[3], Manoj P Menezes[4], Ji Eun Hong[1], Derek W Leong[1,23], Jan Senderek[3,5], Michael S Salman[6,7], David Chitayat[8,9], Pavel Seeman[10], Arpad von Moers[11], Luitgard Graul-Neumann[12], Andrew J Kornberg[13], Manuel Castro-Gago[14], María-Jesús Sobrido[15,16], Masafumi Sanefuji[17], Perry B Shieh[1], Noriko Salamon[18], Ronald C Kim[19,20], Harry V Vinters[1,21], Zugen Chen[2], Klaus Zerres[3], Monique M Ryan[13], Stanley F Nelson[2,21,22] & Joanna C Jen[1]

**RNA exosomes are multi-subunit complexes conserved throughout evolution[1] and are emerging as the major cellular machinery for processing, surveillance and turnover of a diverse spectrum of coding and noncoding RNA substrates essential for viability[2]. By exome sequencing, we discovered recessive mutations in *EXOSC3* (encoding exosome component 3) in four siblings with infantile spinal motor neuron disease, cerebellar atrophy, progressive microcephaly and profound global developmental delay, consistent with pontocerebellar hypoplasia type 1 (PCH1; MIM 607596)[3–6]. We identified mutations in *EXOSC3* in an additional 8 of 12 families with PCH1. Morpholino knockdown of *exosc3* in zebrafish embryos caused embryonic maldevelopment, resulting in small brain size and poor motility, reminiscent of human clinical features, and these defects were largely rescued by co-injection with wild-type but not mutant *exosc3* mRNA. These findings represent the first example of an RNA exosome core component gene that is responsible for a human disease and further implicate dysregulation of RNA processing in cerebellar and spinal motor neuron maldevelopment and degeneration.**

Pontocerebellar hypoplasia (PCH) is a clinically and genetically heterogeneous group of autosomal recessive disorders characterized by cerebellar hypoplasia or atrophy, variable pontine atrophy and progressive microcephaly with global developmental delay[7]. PCH1 is a distinctive subtype of PCH, characterized by diffuse muscle wasting that is secondary to spinal cord anterior horn cell loss and cerebellar hypoplasia[3–6]. Diagnosis with PCH1 is often delayed or never made because the combination of cerebellar and spinal motor neuron degeneration is not commonly recognized, and the presentation of diffuse weakness and devastating brain involvement is atypical of classical proximal spinal muscular atrophy (SMA)[8]. The literature contains only a handful of descriptions of case series[9–12] and reports of PCH1 (refs. 13–19). The causative gene has not been identified in the majority of individuals with PCH1. Recessive mutations have been found in *VRK1* (encoding vaccinia-related kinase 1)[20], *RARS2* (encoding mitochondrial arginyl-tRNA synthetase 2)[21] and *TSEN54* (encoding tRNA splicing endonuclease 54)[22] in single individuals with PCH1. In PCH without SMA, *TSEN54* mutations account for most cases of PCH2 and PCH4 (refs. 21,23), and *RARS2* mutations have been found in two families with PCH6 (refs. 24,25).

[1]Department of Neurology, University of California, Los Angeles, California, USA. [2]Department of Human Genetics, University of California, Los Angeles, California, USA. [3]Institute of Human Genetics, Medical Faculty, University Hospital Rheinisch Westfälische Technische Hochschule (RWTH) Aachen, Aachen, Germany. [4]Institute for Neuroscience and Muscle Research, Children's Hospital at Westmead, Westmead, New South Wales, Australia. [5]Institute of Neuropathology, Medical Faculty, University Hospital RWTH Aachen, Aachen, Germany. [6]Section of Pediatric Neurology, Children's Hospital, Winnipeg, Manitoba, Canada. [7]Department of Pediatrics and Child Health, University of Manitoba, Winnipeg, Manitoba, Canada. [8]The Prenatal Diagnosis and Medical Genetics Program, Department of Obstetrics and Gynecology, Mount Sinai Hospital, University of Toronto, Toronto, Ontario, Canada. [9]Division of Clinical and Metabolic Genetics, The Hospital for Sick Children, University of Toronto, Toronto, Ontario, Canada. [10]Department of Child Neurology, DNA Laboratory, 2nd School of Medicine, Charles University Prague and University Hospital Motol, Prague, Czech Republic. [11]Department of Pediatrics, Deutsches Rotes Kreuz (DRK) Kliniken Berlin Westend, Berlin, Germany. [12]Institute of Medical and Human Genetics, Charité Universitätsmedizin, Berlin, Germany. [13]Royal Children's Hospital, Murdoch Childrens Research Institute, University of Melbourne, Melbourne, Queensland, Australia. [14]Servicio de Neuropediatría, Departamento de Pediatría, Hospital Clínico Universitario, Facultad de Medicina, Universidad de Santiago de Compostela, Santiago de Compostela, Spain. [15]Fundación Pública Galega de Medicina Xenómica, Clinical Hospital of Santiago de Compostela, Servicio Galego de Saúde (SERGAS), Santiago de Compostela, Spain. [16]Centre for Biomedical Network Research on Rare Diseases (CIBERER), Institute of Health Carlos III, Barcelona, Spain. [17]Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan. [18]Department of Radiology, University of California, Los Angeles, California, USA. [19]Department of Pathology, University of California, Irvine, California, USA. [20]Department of Neurology, University of California, Irvine, California, USA. [21]Department of Pathology & Laboratory Medicine, University of California, Los Angeles, California, USA. [22]Department of Psychiatry, University of California, Los Angeles, California, USA. [23]Present address: Georgetown University School of Medicine, Washington, DC, USA. [24]These authors contributed equally to this work. Correspondence should be addressed to J.C.J. (jjen@ucla.edu).

Received 15 December 2011; accepted 22 March 2012; published online 29 April 2012; doi:10.1038/ng.2254

We identified one family (family 1) in which four children were floppy at birth, had ocular motor apraxia, progressive muscle wasting, distal contractures, progressive microcephaly, growth retardation and global developmental delay, and never reached any motor milestone or spoke. Although normal in size at birth, in all four children, head circumference, height and weight dropped to below the 5th percentile by the age of 7–10 months. Magnetic resonance imaging (MRI) showed marked cerebellar atrophy with prominent sulci and decreased volume of folia (**Fig. 1a**–**d**) compared to age- and gender-matched normal individuals (**Fig. 1e**–**h**). In the affected individuals, the brainstem and the cerebral cortex appear normal in configuration but are small. Electromyography showed neurogenic motor changes in an 18-year-old subject (**Fig. 1i**), which were exemplified by a single fast-firing (25 Hz) wave complex that was polyphasic (crossing the baseline multiple times) and unstable. The high-frequency firing in the absence of other complexes suggests a loss of axons, and unstable polyphasic units are manifestations of reinnervation in response to denervation. In a 9-year-old subject with PCH1, we observed borderline neurogenic motor changes (**Fig. 1j**), with a normal recruitment pattern but occasional large-amplitude motor unit action potentials (~4.5 mV) suggestive of reinnervation, compared to a normal individual (**Fig. 1k**), who showed multiple distinct wave complexes of normal amplitude (200–400 μV) that represent preserved motor axons without injury. Nerve conduction studies showed motor responses with severely reduced amplitude but normal sensory responses in the affected individuals (**Supplementary Table 1**). Furthermore, when the oldest child in family 1 died at age 18 years after a respiratory infection, the autopsy revealed a severe loss of cerebellar (**Fig. 1l**,**m**) and spinal (**Fig. 1n**) motor neurons compared to a control individual (**Fig. 1o**–**q**). These clinical features are most consistent with PCH1.

No PCH1-associated genes were known when this study began. We performed a genome scan of the four affected siblings, three healthy siblings and their parents, which narrowed the candidate regions to four subchromosomal loci with more than 100 candidate genes in total (**Supplementary Fig. 1**). To identify the gene underlying PCH1, we captured the exomes from the four affected siblings using the SureSelect Human All Exon kit (Agilent Technology, G3362) and performed sequencing on a Genome Analyzer IIx (Illumina). This analysis yielded one candidate variant fulfilling the requirement of being a rare biallelic variant within the intervals identical by descent in all the affected individuals: g.9:37783990T>G in the *EXOSC3* gene (c.395A>C, Ensembl ENST00000327304.4; encoding p.Asp132Ala, ENSP00000323046.4). We did not observe variants in *VRK1*, *RARS2* or *TSEN54* that have previously been reported in PCH1.

There are multiple alternatively spliced forms of *EXOSC3*, with the longest reading frame spanning 4 exons over 5,119 bases (NM_016042.2) and encoding a 275–amino acid protein, human exosome component 3 (EXOSC3), also known as the ribosomal RNA–processing protein 40 (RRP40) (NP_057126.2). EXOSC3 is a core component of the human RNA exosome complex (distinct from exosome vesicles) that is present in the cytoplasm and the nucleus and especially enriched in the nucleolus[26]. The N-terminal (NT) domain and putative RNA-binding S1 and KH domains are evolutionarily conserved (**Fig. 2**).

We confirmed genotype-phenotype co-segregation in family 1 by Sanger sequencing. To validate the association between *EXOSC3* mutations

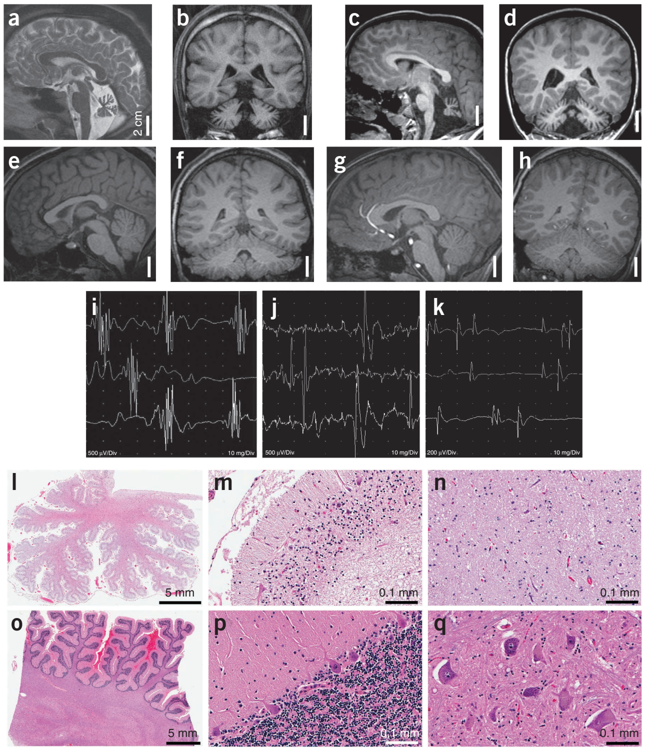


**Figure 1** Neuroimaging, neuromuscular and pathological features in family 1. (**a**–**h**) Sagittal T2- (**a**) and coronal T1-weighted (**b**) images from the oldest surviving sibling, who was 18 years old at the time of the study, showing the presence of all cerebellar lobules, yet with marked atrophy, compared to T1-weighted sagittal (**e**) and coronal (**f**) images from an age-matched, normal male. Sagittal (**c**) and coronal (**d**) T1-weighted images from the youngest surviving sibling, who was 9 years old at the time of the study, showing cerebellar volume loss comparable to that seen in **a**,**b**, in contrast to sagittal (**g**) and coronal (**h**) T1-weighted images from an age-matched, normal male. (**i**–**k**) Needle electromyography (EMG) tracings in the left triceps muscle of the 18-year-old sibling (**i**) and right vastus lateralis muscle of the 9-year-old sibling (**j**) showed neurogenic changes compared to normal EMG tracings in the right biceps muscle of a healthy adult male (**k**). (**l**–**q**) Brain autopsy of the subject who died at age 18 years (**l**) shows profound cerebellar atrophy compared to control (**o**), with dysmorphic Purkyně (also known as Purkinje) cells and loss of granule cells seen at higher magnification (**m**) compared to in control (**p**). Diffuse loss of motor neurons was observed in the anterior horn of the spinal cord of the affected subject (**n**) compared to the control with normal-appearing spinal motor neurons (**q**).

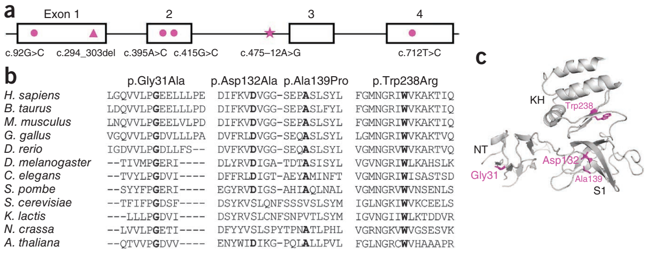

a
Exon 1
2
3
4
c.92G>C
c.294_303del
c.395A>C
c.415G>C
c.475–12A>G
c.712T>C
b
p.Gly31Ala
p.Asp132Ala
p.Ala139Pro
p.Trp238Arg
H. sapiens LGQVVLPGEELLLPE DIFKVDVGG-SEPASLSYL FGMNGRIWVKAKTIQ
B. taurus LDQVVLPGEELLLPD DIFKVDVGG-SEPASLFYL FGMNGRIWVKAKTIQ
M. musculus LNQVVLPGEELVLPD DIFKVDVGG-SEPASLSYL FGMNGRIWVKAKTIQ
G. gallus VGQVVLPGDVLLLPA DVFRLDVGG-SEQASLSYL LGMNGRIWVKAKTVQ
D. rerio IGDVVLPGDLLFS-- DVFKVDVGG-SEQASLSYL VGMNGRVWVKARTVQ
D. melanogaster --TIVMPGERI---- DLYRVDIGA-TDTASISYL VGVNGRIWLKAHSLK
C. elegans --TVYLPGDVI---- DFFRLDIGT-AEYAMINFT VGMNGRIWISASTSD
S. pombe --SYYFPGERI---- EGYRVDIGS-AHIAQLNAL VGMNGRVWVNSENLS
S. cerevisiae --TFIFPGDSF---- DSYKVSLQNFSSSVSLSYM IGLNGKIWVKCEELS
K. lactis ---LLLPGDVI---- DSYRVSLCNFSNPVTLSYM IGVNGIIWLKTDDVR
N. crassa --LVVLPGETI---- DFYYVSLSPYTPNATLPHL VGRNGKVWVGSESVK
A. thaliana --QTVVPGDVV---- ENYWIDIKG-PQLALLPVL FGLNGRCWVHAAAPR
c
KH
Trp238
NT
Gly31
Asp132
Ala139
S1

**Figure 2** *EXOSC3* mutations in PCH1. (**a**) Genomic structure of *EXOSC3*, with the four exons indicated by open boxes and mutations highlighted in magenta. Circle, missense mutation; triangle, deletion; star, splice-site mutation. (**b**) Alignment of protein sequences encoded by orthologs in humans and other eukaryotic organisms, including vertebrate, insect, plant and yeast species, showing that the mutated amino acids (highlighted in bold) are conserved. (**c**) Schematic of the locations of the mutated amino acids (highlighted in magenta) in the EXOSC3 protein, with the conserved NT, S1 and KH domains indicated (Protein Data Bank (PDB) 2NN6).

and PCH1, we sequenced all exons and flanking introns of *EXOSC3* (**Supplementary Table 2**) in the probands from 12 additional families with PCH1. Eight probands had recessive mutations in the gene (**Fig. 2** and **Table 1**). All available samples from the parents of the affected subjects were heterozygous for the mutations. None of the mutations were found in Turkish ($n = 94$), Czech ($n = 96$) or North American (principally of northern and western European ancestry) ($n = 189$) control individuals. A more recent review of databases, including the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project, showed that the mutation encoding the p.Asp132Ala alteration has been observed in 6 of 4,870 exomes, with an estimated allele frequency of 0.0012. None of the other variants has been previously reported.

The p.Asp132Ala alteration was present in seven of the nine mutation-positive families (**Fig. 2** and **Table 1**). This change affects a highly conserved amino-acid residue in the putative RNA-binding S1 domain; the crystal structure suggests that Asp132 may be important for intersubunit interaction within the exosome complex[27]. We genotyped the probands of families 1–3 who were homozygous for mutations causing the p.Asp132Ala alteration to find identical haplotypes in a 1-cM region flanking the mutation locus, which would suggest an ancestral origin for the mutation (**Supplementary Table 3**).

We found three additional missense mutations. Two mutations, encoding p.Gly31Ala and p.Trp238Arg alterations, were present in family 4, with the parents being identified as carriers. The mutation encoding p.Gly31Ala was homozygous in the affected subject in family 9. Strictly conserved from yeast to humans, the Gly31 residue in the N-terminal domain seems to be involved in intersubunit interaction[27], whereas the Trp238 residue is in the putative RNA-binding KH domain[27]. In family 8, the affected subject harbored a mutation encoding p.Asp132Ala in *trans* with another missense mutation in the S1 domain, c.415G>C, encoding p.Ala139Pro (**Fig. 2** and **Supplementary Note**).

We identified one frameshift mutation, a deletion spanning ten nucleotides in family 5, which is predicted to result in premature termination of the protein; the faulty transcript may be subject to nonsense-mediated mRNA degradation. *In silico* analysis of the intronic mutation c.475–12A>G in family 6 suggested that it may introduce a new splice site just upstream of the normal splice acceptor for exon 3. RT-PCR in expression studies primarily showed skipping of exon 3 (shifting the reading frame) and evidence of aberrant splicing (which incorporated 11 nucleotides upstream of the normal splice site), with a minority of transcripts having normal splicing (**Supplementary Fig. 2** and **Supplementary Table 2**).

The fact that biallelic missense, frameshift and splice-site mutations all led to the same clinical manifestations suggests that the alleles may be null or hypomorphic. Because all components of the exosome are essential for viability[1], it is unlikely that individuals with PCH1 harbor biallelic null mutations; it is more likely that the missense mutations are hypomorphic, and the frameshift mutations could be null. *In silico* analyses predicted detrimental consequences from the missense mutations (**Supplementary Table 4**). The standard marker for impaired exosome function has long been an abnormal accumulation of unprocessed rRNA[1], which we did not observe in fibroblasts from the subject in family 1 (**Supplementary Fig. 3**), suggesting that the impact of the homozygous mutations in *EXOSC3* encoding p.Asp132Ala may be more nuanced and subtle than a complete elimination of exosome function.

**Table 1 Ancestry and *EXOSC3* mutations in subjects with PCH1**

| Family | Subjects | Ancestry | Age (death) | Nucleotide change | Amino-acid change | Ref. |
|---|---|---|---|---|---|---|
| 1 | 4 male | American, European | (18 years), 18 years, 16 years, 9 years | c.395A>C, homozygous | p.Asp132Ala | – |
| 2 | 1 female | Canadian, Cuban | (40 months) | c.395A>C, homozygous | p.Asp132Ala | 14 |
| 3[a] | 1 female, 1 male | German, Turkish | 20 years, 16 years | c.395A>C, homozygous | p.Asp132Ala | – |
| 4 | 1 male, 1 female | Czech | (8 months), (8 months) | c.92G>C, c.712T>C | p.Gly31Ala, p.Trp238Arg | – |
| 5 | 1 male | New Caledonian | (Seen at 3 months) | c.294_303del; c.395A>C | p.99fs*11; p.Asp132Ala | 11 |
| 6 | 1 female | Australian | (26 months) | c.395A>C; c.475-12A>G | p.Asp132Ala; exon 3 skipping, aberrant splicing | – |
| 7[a] | 1 male | Australian, Turkish | (3 years) | c.395A>C, homozygous | p.Asp132Ala | – |
| 8 | 1 male | Australian | (11 months) | c.395A>C; c.415G>C | p.Asp132Ala; p.Ala139Pro | 11 |
| 9 | 1 male | Czech | (17 months) | c.92G>C homozygous | p.Gly31Ala | – |
| 10 | 1 male | Spanish | 10 years | – | – | 15 |
| 11 | 1 male | Japanese | (15 years) | – | – | 19 |
| 12 | 1 male, 1 female | Australian | (9 months) | – | – | – |
| 13 | 1 female | Australian | (20 months) | – | – | – |

Mutations in *EXOSC3* were identified by exome sequencing in affected subjects in family 1 and further investigated in DNA samples from families 2–13 by targeted sequencing.
[a]Parental consanguinity.

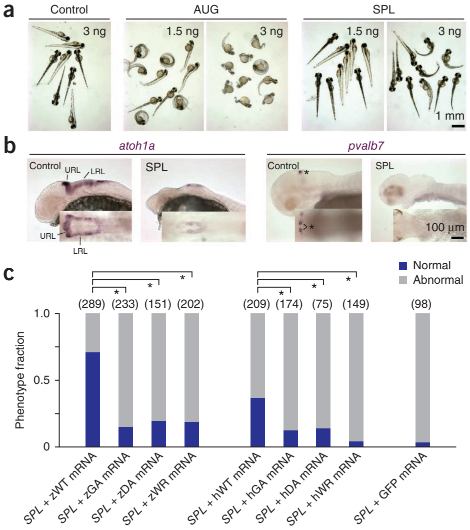


**Figure 3** Knockdown of *exosc3* in zebrafish embryos disrupts normal development. (**a**) Zebrafish embryos injected with *exosc3*-specific antisense morpholinos AUG (directed against the start codon) or SPL (directed against the splice-donor site for exon 2) compared to those injected with nonspecific control. (**b**) Whole-mount *in situ* hybridization in embryos injected with SPL morpholinos in lateral view (inset, dorsal view with rostral to the left) showing diminished expression of dorsal hindbrain progenitor–specific marker *atoh1a* and cerebellar-specific marker *pvalb7* compared to embryos injected with nonspecific control. URL, upper rhombic lip; LRL, lower rhombic lip; *, distinct clusters of differentiated Purkinje cells in embryos 3 d.p.f. (**c**) Survival data from embryos 3 d.p.f. co-injected with 3 ng of SPL and with 240 pg of human *EXOSC3* or zebrafish *exosc3* mRNA versus GFP mRNA as control from three separate experiments. z, zebrafish; h, human; WT, wild-type *EXOSC3* or *exosc3*; GA, mutant mRNA encoding p.Gly31Ala in human or p.Gly20Ala in zebrafish; DA, mutant mRNA encoding p.Asp132Ala in human or p.Asp102Ala in zebrafish; WR, mutant mRNA encoding p.Trp238Arg in human or p.Trp208Arg in zebrafish. Embryos were classified as normal (blue) or abnormal (gray; including embryos that were mildly abnormal, severely abnormal or dead). *$P < 0.0001$, two-tailed Pearson's $\chi$-squared test.

To further examine the functional effects of the mutations, we knocked down endogenous *exosc3* expression (NM_001029961.1) in zebrafish embryos by *exosc3*-specific antisense morpholino injection (**Fig. 3**, **Supplementary Fig. 4** and **Supplementary Table 2**). Zebrafish embryos injected with antisense morpholinos directed against the start codon or the splice-donor site of exon 2 of *exosc3* led to a dose-dependent phenotype of a short, curved spine and small brain with poor motility and even death by 3 days post fertilization (d.p.f.) compared to embryos injected with nonspecific, control morpholinos (**Fig. 3a**).

The observation of shrunken or collapsed hindbrain in embryos injected with morpholinos targeting the splice-donor site prompted us to further investigate hindbrain-specific cells. Whole-mount *in situ* hybridization showed decreased expression of *atoh1a* (a marker specific for dorsal hindbrain progenitors)[28] by 1 d.p.f. in the upper and lower rhombic lips in embryos injected with morpholinos to the splice-donor site compared to the normal pattern of robust expression of *atoh1a* in hindbrain progenitors in the control-injected embryos[28] (**Fig. 3b**). Whole-mount *in situ* hybridization further showed a lack of expression of *pvalb7*, which is specific for differentiated cerebellar Purkinje neurons[28], by 3 d.p.f. in embryos injected with morpholinos to the splice-donor site compared to normal expression in distinct clusters of differentiated Purkinje cells in embryos injected with control morpholinos (**Fig. 3b**).

The abnormal phenotype from *exosc3*-specific morpholino injection was largely rescued by co-injection with wild-type zebrafish *exosc3* mRNA (**Fig. 3c** and **Supplementary Table 5**), suggesting that the detrimental effects of the antisense morpholinos were specific to *exosc3* knockdown. Co-injection with wild-type human *EXOSC3* mRNA, which shares 67% identity with the zebrafish ortholog, was less effective in rescue. Co-injection with zebrafish or human mRNA containing the mutations was ineffective in rescue, suggesting that the mutations disrupted the normal function of EXOSC3 (**Fig. 3c** and **Supplementary Table 5**). Survival data of embryos 1–3 d.p.f. are stratified and summarized (**Supplementary Table 5**).

We have discovered disease-causing mutations in a gene encoding the exosome component EXOSC3 leading to PCH1 with combined cerebellar and spinal motor neuron degeneration of infantile onset. There is clinical heterogeneity. Affected individuals in families 1 and 3 do not present with primary hypoventilation and have survived beyond infancy and early childhood, which is exceptionally unusual for classical PCH1 (refs. 7,23). Furthermore, in families 1 and 2, autopsy showed profound cerebellar atrophy and variable involvement of the pons and inferior olives, suggesting a degenerative process in addition to a developmental disorder. Additional studies will facilitate endophenotype stratification of PCH1. There is clear genetic heterogeneity in PCH1, as some affected individuals do not harbor mutations in any known PCH1-associated genes.

RNA exosomes are the principal enzymes that process and degrade RNA. The bulk of the human genome is transcribed to produce an extraordinary diversity of RNA[29]. The versatility and specificity of the exosome regulate the activity and maintain the fidelity of gene expression. Although exosomes are immunogenic in some individuals with polymyositis-scleroderma[30,31] or chronic myelogenous leukemia[32,33], the findings in this report are the first to our knowledge to establish a pathogenic role for exosome core component mutations in human disease. Despite a growing effort to examine exosome function and subunit contribution, its substrates have not been fully characterized in humans or in lower animals, and the specific contribution of each component is incompletely understood. The discovery of naturally occurring mutations in exosome component genes provides a valuable opportunity to define subunit contribution to exosome function. Our findings suggest that normal function of the EXOSC3 component is essential for the survival of cerebellar and spinal motor neurons. Of note, RNA dysregulation is increasingly understood to be important in the etiology of motor and cerebellar degeneration. RNA processing defects are implicated in *SMN1* deficiency in SMA[8]. Mutations in RNA- and/or DNA-binding proteins[34–37] and pathogenic repeat expansions generating RNA that is likely toxic[38,39] cause amyotrophic lateral sclerosis (ALS), an adult-onset motor neuron disease. Gain of function of RNA from noncoding repeat expansions was recently proposed to cause combined spinocerebellar and brainstem motor neuron degeneration of late onset in SCA36 (ref. 40). Dysregulation of tRNA processing underlies other subtypes of PCH[21,23,24]. Elucidation of the pathological mechanism underlying PCH1 may lead to new insights regarding RNA processing in the development and survival of cerebellar and spinal motor neurons.

**URLs.** NHLBI Exome Sequencing Project, http://evs.gs.washington.edu/EVS/; Novocraft, http://www.novocraft.com; Ensembl, http://www.ensembl.org/; dbSNP132, http://www.ncbi.nlm.nih.gov/projects/SNP/; 1000 Genome Project, http://www.1000genomes.org/.

## METHODS

Methods and any associated references are available in the online version of the paper.

*Note: Supplementary information is available in the online version of the paper.*

ACKNOWLEDGMENTS

The authors are grateful for the generosity of the families who participated in this study. The sequencing and analytical work were supported by the Bioinformatics and Genomics Core of the UCLA Muscular Dystrophy Core Center (US National Institutes of Health (NIH)/National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS) P30AR057230) within the Center for Duchenne Muscular Dystrophy. We appreciate fibroblasts from three subjects from the Eunice Kennedy Shriver National Institute of Child Health & Human Development (NICHD) Brain and Tissue Bank for Developmental Disorders. We acknowledge the contribution of clinical data by P. Carpenter (University of California, Irvine). The work was supported by the NIH/National Eye Institute (NEI) (R01 EY015311 and NINDS R01 NS064183 to J.C.J.), Deutsche Forschungsgemeinschaft (Ru-746/1-1 to S.R.-S. and K.Z.), IZKF Aachen (N5-4 to S.R.-S. and J.S.), the Australian National Health and Medical Research Council (NHMRC) Centre for Research Excellence (M.M.R.), the Internal Grant Agency of the Ministry of Health of the Czech Republic (NS 10552-3 to P.S.) and Xunta de Galicia–Plan Galego de InvestigaCión, Desenvolvemento e Innovación Tecnolóxica (INCITE) (10PXIB9101280PR to M.-J.S.).

### AUTHOR CONTRIBUTIONS

S.F.N. and J.C.J. designed the study. M.Y., Z.C. and S.F.N. analyzed data from SNP genotyping and exome sequencing. S.R.-S., M.P.M., J.S., M.S.S., D.C., P.S., A.v.M., L.G.-N., A.J.K., M.C.-G., M.-J.S., M.S., P.B.S., N.S., R.C.K., H.V.V., K.Z. and M.M.R. provided and analyzed clinical material from subjects. H.M., J.E.H. and D.W.L. performed and analyzed data from Sanger sequencing of subjects and controls. H.M. and J.E.H. generated constructs and performed molecular genetics studies. J.W. performed and analyzed the functional studies in zebrafish. All authors contributed to the manuscript written by J.C.J.

### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2254.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Mitchell, P., Petfalski, E., Shevchenko, A., Mann, M. & Tollervey, D. The exosome: a conserved eukaryotic RNA processing complex containing multiple 3′→5′ exoribonucleases. *Cell* **91**, 457–466 (1997).
2. Jensen, T.H. RNA exosome. *Adv. Exp. Med. Biol.* **702** (2010).
3. Norman, R.M. Cerebellar hypoplasia in Werdnig-Hoffmann disease. *Arch. Dis. Child.* **36**, 96–101 (1961).
4. Goutières, F., Aicardi, J. & Farkas, E. Anterior horn cell disease associated with pontocerebellar hypoplasia in infants. *J. Neurol. Neurosurg. Psychiatry* **40**, 370–378 (1977).
5. de Leén, G.A., Grover, W.D. & D'Cruz, C.A. Amyotrophic cerebellar hypoplasia: a specific form of infantile spinal atrophy. *Acta Neuropathol.* **63**, 282–286 (1984).
6. Barth, P.G. Pontocerebellar hypoplasias. An overview of a group of inherited neurodegenerative disorders with fetal onset. *Brain Dev.* **15**, 411–422 (1993).
7. Namavar, Y., Barth, P.G., Poll-The, B.T. & Baas, F. Classification, diagnosis and potential mechanisms in pontocerebellar hypoplasia. *Orphanet J. Rare Dis.* **6**, 50 (2011).
8. Melki, J. *et al. De novo* and inherited deletions of the 5q13 region in spinal muscular atrophies. *Science* **264**, 1474–1477 (1994).
9. Görgen-Pauly, U., Sperner, J., Reiss, I., Gehl, H.B. & Reusche, E. Familial pontocerebellar hypoplasia type I with anterior horn cell disease. *Eur. J. Paediatr. Neurol.* **3**, 33–38 (1999).
10. Muntoni, F. *et al.* Clinical spectrum and diagnostic difficulties of infantile pontocerebellar hypoplasia type 1. *Neuropediatrics* **30**, 243–248 (1999).
11. Ryan, M.M., Cooke-Yarborough, C.M., Procopis, P.G. & Ouvrier, R.A. Anterior horn cell disease and olivopontocerebellar hypoplasia. *Pediatr. Neurol.* **23**, 180–184 (2000).
12. Rudnik-Schöneborn, S. *et al.* Extended phenotype of pontocerebellar hypoplasia with infantile spinal muscular atrophy. *Am. J. Med. Genet. A* **117A**, 10–17 (2003).
13. Chou, S.M. *et al.* Infantile olivopontocerebellar atrophy with spinal muscular atrophy (infantile OPCA + SMA). *Clin. Neuropathol.* **9**, 21–32 (1990).
14. Salman, M.S. *et al.* Pontocerebellar hypoplasia type 1: new leads for an earlier diagnosis. *J. Child Neurol.* **18**, 220–225 (2003).
15. Gómez-Lado, C., Eiris-Punal, J., Vazquez-Lopez, M.E. & Castro-Gago, M. Pontocerebellar hypoplasia type I and mitochondrial pathology. *Rev. Neurol.* **45**, 639–640 (2007).
16. Lev, D. *et al.* Infantile onset progressive cerebellar atrophy and anterior horn cell degeneration—a late onset variant of PCH-1? *Eur. J. Paediatr. Neurol.* **12**, 97–101 (2008).
17. Szabó, N., Szabo, H., Hortobagyi, T., Turi, S. & Sztriha, L. Pontocerebellar hypoplasia type 1. *Pediatr. Neurol.* **39**, 286–288 (2008).
18. Tsao, C.Y., Mendell, J., Sahenk, Z., Rusin, J. & Boue, D. Hypotonia, weakness, and pontocerebellar hypoplasia in siblings. *Semin. Pediatr. Neurol.* **15**, 151–153 (2008).
19. Sanefuji, M. *et al.* Autopsy case of later-onset pontocerebellar hypoplasia type 1: pontine atrophy and pyramidal tract involvement. *J. Child Neurol.* **25**, 1429–1434 (2010).
20. Renbaum, P. *et al.* Spinal muscular atrophy with pontocerebellar hypoplasia is caused by a mutation in the *VRK1* gene. *Am. J. Hum. Genet.* **85**, 281–289 (2009).
21. Namavar, Y. *et al.* Clinical, neuroradiological and genetic findings in pontocerebellar hypoplasia. *Brain* **134**, 143–156 (2011).
22. Simonati, A., Cassandrini, D., Bazan, D. & Santorelli, F.M. *TSEN54* mutation in a child with pontocerebellar hypoplasia type 1. *Acta Neuropathol.* **121**, 671–673 (2011).
23. Budde, B.S. *et al.* tRNA splicing endonuclease mutations cause pontocerebellar hypoplasia. *Nat. Genet.* **40**, 1113–1118 (2008).
24. Edvardson, S. *et al.* Deleterious mutation in the mitochondrial arginyl-transfer RNA synthetase gene is associated with pontocerebellar hypoplasia. *Am. J. Hum. Genet.* **81**, 857–862 (2007).
25. Rankin, J. *et al.* Pontocerebellar hypoplasia type 6: a British case with PEHO-like features. *Am. J. Med. Genet. A* **152A**, 2079–2084 (2010).
26. Brouwer, R. *et al.* Three novel components of the human exosome. *J. Biol. Chem.* **276**, 6177–6184 (2001).
27. Liu, Q., Greimann, J.C. & Lima, C.D. Reconstitution, activities, and structure of the eukaryotic RNA exosome. *Cell* **127**, 1223–1237 (2006).
28. Kani, S. *et al.* Proneural gene–linked neurogenesis in zebrafish cerebellum. *Dev. Biol.* **343**, 1–17 (2010).
29. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
30. Wolfe, J.F., Adelstein, E. & Sharp, G.C. Antinuclear antibody with distinct specificity for polymyositis. *J. Clin. Invest.* **59**, 176–178 (1977).
31. Allmang, C. *et al.* The yeast exosome and human PM-Scl are related complexes of 3′→5′ exonucleases. *Genes Dev.* **13**, 2148–2158 (1999).
32. Yang, X.F. *et al.* CML28 is a broadly immunogenic antigen, which is overexpressed in tumor cells. *Cancer Res.* **62**, 5517–5522 (2002).
33. Xie, L.H. *et al.* Activation of cytotoxic T lymphocytes against CML28-bearing tumors by dendritic cells transduced with a recombinant adeno-associated virus encoding the *CML28* gene. *Cancer Immunol. Immunother.* **57**, 1029–1038 (2008).
34. Kabashi, E. *et al. TARDBP* mutations in individuals with sporadic and familial amyotrophic lateral sclerosis. *Nat. Genet.* **40**, 572–574 (2008).
35. Sreedharan, J. *et al.* TDP-43 mutations in familial and sporadic amyotrophic lateral sclerosis. *Science* **319**, 1668–1672 (2008).
36. Kwiatkowski, T.J. Jr. *et al.* Mutations in the *FUS/TLS* gene on chromosome 16 cause familial amyotrophic lateral sclerosis. *Science* **323**, 1205–1208 (2009).
37. Vance, C. *et al.* Mutations in FUS, an RNA processing protein, cause familial amyotrophic lateral sclerosis type 6. *Science* **323**, 1208–1211 (2009).
38. DeJesus-Hernandez, M. *et al.* Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p–linked FTD and ALS. *Neuron* **72**, 245–256 (2011).
39. Renton, A.E. *et al.* A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21–linked ALS-FTD. *Neuron* **72**, 257–268 (2011).
40. Kobayashi, H. *et al.* Expansion of intronic GGCCTG hexanucleotide repeat in *NOP56* causes SCA36, a type of spinocerebellar ataxia accompanied by motor neuron involvement. *Am. J. Hum. Genet.* **89**, 121–130 (2011).

## ONLINE METHODS

**Clinical characterization.** PCH1 was diagnosed in 13 unrelated families from around the world, with documented congenital combined cerebellar and spinal motor neuron disease. Almost all affected subjects were hypotonic from birth. All had spontaneous breathing. Neurogenic muscle atrophy with spinal motor neuron disease was confirmed by EMG, muscle biopsy or autopsy. Many subjects developed progressive microcephaly, with prominent cerebellar atrophy and variable involvement of the brainstem, as determined by MRI or autopsy.

**Genetic analysis.** DNA was extracted from peripheral blood, with consent from all participants and their legal guardians, using standard methods. The study was approved by the University of California, Los Angeles (UCLA) Institutional Review Board. For exome sequencing, each library produced approximately 28 million single-end 76-bp reads. The mean coverage of bases in the target exomes was 23×. Raw reads that passed Illumina's quality filters were aligned to the reference human genome Build 37 with Novoalign from Novocraft. The GATK UnifiedGenotyper was used to call single-nucleotide variants and indels[41,42]. Each case had ~15,000 variants not present in the GRCh37 reference human genome, amounting to 19,098 total variants in the 4 cases. We limited the search to variants within the coding regions or flanking intronic essential splice sites of protein-coding genes in the Ensembl data set. Under the hypothesis that the disorder was rare and the causative allele(s) would therefore not be common, we filtered out variants that were in dbSNP132 (refs. 43,44) and the 1000 Genomes Project[45], leaving ~400 variants in each case and a total of 699 variants. Under a recessive model, we searched for homozygous variants and compound heterozygous variants (defined as 2 variants in the same transcript) that were shared by all 4 cases (15 and 10 variants, respectively). To compensate for bias in our own analytical system, we then filtered out variants that we had identified in 25 exomes from unrelated, unaffected subjects. Sanger sequencing for further validation was performed using standard protocols (**Supplementary Table 2**).

**Zebrafish morpholino injection and *in situ* hybridization.** Cloning, mutagenesis and *in vitro* mRNA synthesis were performed using standard protocols (**Supplementary Table 2**). Zebrafish embryos were provided by the UCLA Zebrafish Core Facility, and procedures were approved by the UCLA Animal Research Committee. Fish were maintained at 28 °C using a 14-h light/10-h dark cycle and bred to obtain embryos. Morpholino oligonucleotides (Gene Tools) were designed to block translation initiation or the splice-donor site of exon 2 of zebrafish *exosc3* pre-mRNA (**Supplementary Table 2**). We obtained a standard control from Gene Tools. We used fine glass needles and a microinjector to perform injections of embryos at the one-cell stage. The injection volume ranged from 0.5 to 2.0 nl at a concentration of 3 ng/nl. Embryos were incubated in E3 medium (5 mM NaCl, 0.17 mM KCl, 0.33 mM $CaCl_2$, 0.33 mM $MgSO_4$ and 0.00001% methylene blue; Sigma) at 28 °C.

Whole-mount *in situ* hybridization was performed, and the expression of specific genes was detected using an alkaline phosphatase–conjugated antibody against digoxigenin (DIG) and a chromogenic substrate, as described previously[46].

**Riboprobe generation.** To generate the antisense probe, full-length zebrafish *exosc3* cDNA was digested with EcoRI and NotI and ligated into pCR-Blunt II-TOPO (Invitrogen) with T4 DNA ligase. The sense probe was transcribed using full-length zebrafish *exosc3* in pcDNA3.1/Zeo(+) (Invitrogen).

Full-length zebrafish *pvalb7* (clone 7087368, Thermo Scientific Open Biosystems) was digested with EcoRI and NotI and ligated into pCR-Blunt II-TOPO with T4 DNA ligase. The sense probe was transcribed *in vitro* from the SP6 promoter, and the antisense was transcribed from the T7 promoter.

Full-length zebrafish *atoh1a* (clone 7428977; Thermo Scientific Open Biosystems) was digested with EcoRI and NotI and ligated into pcDNA3.1/Zeo(+) to generate the sense probe and pcDNA3.1/Zeo(−) (Invitrogen) for the antisense probe.

RNA probes for *in situ* hybridization were generated with the mMESSAGE mMACHINE SP6 or mMESSAGE mMACHINE T7 Ultra kits (Ambion), substituting the DIG RNA labeling Mix (Roche Applied Science) for the NTP mix supplied in the kit. All probes were analyzed by denaturing agarose gel electrophoresis and quantified by NanoDrop.

**Assessing functional impact on *EXOSC3* splicing.** We obtained full-length *EXOSC3* constructs (all exons and intervening introns) by PCR amplification using Phusion High-Fidelity DNA Polymerase (NEB; **Supplementary Table 2**) from genomic DNA of a normal subject and the proband of family 6 (harboring the c.475–12A>G mutation). Gel-extracted amplicons were cloned into pcDNA3.1/Zeo(+) (linearized with BamHI and XhoI) by using the In-Fusion HD Cloning System (Clonetech). The full-length clones were confirmed by bidirectional Sanger sequencing.

The wild-type or mutant full-length constructs were introduced into HeLa cells by transfection with Lipofectamine 2000 (Invitrogen). Two days after transfection, RNA was extracted using TRIzol Reagent (Invitrogen). cDNA was generated using the Transcriptor First Strand cDNA Synthesis kit (Roche Applied Science) and the BGH reverse primer–binding site located upstream of the BGH polyadenylation signal on pcDNA3.1/Zeo(+) (**Supplementary Fig. 2**). Reverse transcription from the BGHr site ensured that cDNA was synthesized exclusively from exogenous mRNA and eliminated transcription from endogenous *EXOSC3*. cDNA (1 µl) was PCR amplified with Phusion High-Fidelity DNA Polymerase and the hEXOSC3-c.423f and hEXOSC3-c.670r primers (**Supplementary Fig. 2** and **Supplementary Table 2**). RT-PCR products were resolved on a 12% polyacrylamide gel and visualized by GelRed (Biotium). We observed multiple products for both wild-type and mutant constructs. To ascertain the identity of each product, amplicons were cloned into pCR-Blunt II-TOPO and directly sequenced.

**Cloning, mutagenesis and *in vitro* transcription.** Full-length human *EXOSC3* cDNA (clone 3346075; Thermo Scientific Open Biosystems) was digested with EcoRI and XhoI and ligated into pcDNA3.1/Zeo(+)with T4 DNA ligase. Full-length zebrafish *exosc3* (clone 7282897; Thermo Scientific Open Biosystems) was digested with EcoRI and NotI and ligated into pcDNA3.1/Zeo(+) with T4 DNA ligase. Specific missense mutations were introduced into the wild-type cDNA constructs with the QuikChange II XL Site-Directed Mutagenesis kit (Agilent Technologies; **Supplementary Table 2**). All clones were fully sequenced bidirectionally by Sanger sequencing. *In vitro* transcription was performed with the mMESSAGE mMACHINE T7 Ultra kit. Resulting mRNA was analyzed by denaturing agarose gel electrophoresis and quantified by NanoDrop.

**Cell culture, siRNA transfection and RNA extraction.** All medium components were purchased from Invitrogen. HeLa cells were cultured in DMEM supplemented with 10% FBS at 37 °C and 5% $CO_2$. Cells were transfected every 48 h over 6 d with 60 nM siRNA duplexes (sense: 5′-CACGCACAGUACUAGGUCATT-3′) with Lipofectamine 2000 or left untreated, as detailed previously[47]. Cells were lysed in TRIzol Reagent, and RNA was extracted from the aqueous phase and protein from the remaining organic phase.

Fibroblasts were grown in DMEM supplemented with 20% FBS, 100 nM MEM Non-Essential Amino Acids, 1 mM GlutaMAX, 100 U penicillin and 100 µg streptomycin at 37 °C and 5% $CO_2$. Total RNA was extracted with TRIzol Reagent.

**RNA blots.** RNA blotting was performed following standard protocols. RNA integrity was analyzed by denaturing agarose gel electrophoresis, and RNA was quantified by NanoDrop. Total RNA (4 µg) was resolved on a 6% denaturing polyacrylamide gel and transferred to positively charged nylon membrane (Roche). RNA was cross-linked to the membrane, and blots were hybridized with DIG-labeled, locked nucleic acid (LNA)-modified antisense oligonucleotide probe to 5.8S rRNA (5′-CGAAGTG**TCGATGAT**CAAT-3×DIG-3′; LNA in bold). Bound probes were detected with an alkaline phosphatase–conjugated antibody to DIG (1:10,000 dilution) and CSPD chemiluminescence (Roche Applied Science).

**Protein blots.** Protein concentration was determined using the Micro BCA Protein Assay kit (Pierce). Lysates (10 µg) were separated by SDS-PAGE and transferred to a nitrocellulose membrane. Protein blotting was performed following standard protocols with mouse monoclonal antibody to EXOSC3 (Santa Cruz Biotechnology, sc-166568; 1:400 dilution) and with secondary horseradish peroxidase–conjugated horse antibody to mouse IgG (Vector Laboratories, PI-2000; 1:5,000 dilution). Blots were subsequently stripped and reprobed with chicken antibody to GAPDH (Millipore, ab2302; 1:1,000 dilution) and secondary horseradish peroxidase–conjugated goat antibody to chicken IgY (Abcam, ab97150; 1:5,000 dilution). Bound antibodies were visualized with Amersham ECL Plus Western Blotting Detection Reagents.

41. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
42. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
43. Biesecker, L.G. *et al.* The ClinSeq Project: piloting large-scale genome sequencing for research in genomic medicine. *Genome Res.* **19**, 1665–1674 (2009).
44. Bhagwat, M. Searching NCBI's dbSNP database. *Curr. Protoc. Bioinformatics* **Chapter 1**, Unit 1.19 (2010).
45. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
46. Thisse, C. & Thisse, B. High-resolution *in situ* hybridization to whole-mount zebrafish embryos. *Nat. Protoc.* **3**, 59–69 (2008).
47. Preker, P. *et al.* RNA exosome depletion reveals transcription upstream of active human promoters. *Science* **322**, 1851–1854 (2008).

*These authors contributed equally.
[†]Current address: Georgetown University School of Medicine, Washington D.C., U.S.A.

Corresponding Author:
J. C. Jen
UCLA Neurology
710 Westwood Plaza
Los Angeles, CA 90095-1769
TEL 310 825 3731
email jjen@ucla.edu

**Supplementary Note**

***Ascertainment of biallelic mutations in families 5 and 8 by long-range PCR***

Since parental DNA was not available for these two families, we determined that the mutations were in trans by long range sequencing. In Family 5, we amplified in the patient the genomic sequence spanning exons 1 and 2 (**Supplementary Table 2**) with Phusion High-Fidelity DNA Polymerase (NEB) and cloned the gel-extracted 1344 bp product into pCR-Blunt II-TOPO (Invitrogen) for transformation of competent bacterial cells. The two mutations were found in separate clones, as confirmed by direct sequencing of purified plasmids.

For Family 8, we PCR amplified the entire gene genomic DNA with Phusion High-Fidelity DNA Polymerase (NEB) and primers hEXOSC3-ATGf and hEXOSC3-Rcr (**Supplementary Table 2**). The gel extracted amplicon was cloned into pcDNA3.1/Zeo(+) (Invitrogen) with In-Fusion® HD Cloning System (Clonetech) and transformed into competent bacterial cells. Plasmids from multiple colonies were sequenced to ascertain that the two mutations resided in different clones.

**SUPPLEMENTARY INFORMATION**

# Mutations in the RNA exosome component gene *EXOSC3* cause pontocerebellar hypoplasia and spinal motor neuron degeneration

Jijun Wan*[1], Michael Yourshaw*[2], Hafsa Mamsa[1], Sabine Rudnik-Schöneborn[3], Manoj P. Menezes[4], Ji Eun Hong[1], Derek W. Leong[1†], Jan Senderek[3,5], Michael S. Salman[6], David Chitayat[7,8], Pavel Seeman[9], Arpad von Moers[10], Luitgard Graul-Neumann[11], Andrew J. Kornberg[12], Manuel Castro-Gago[13], María-Jesús Sobrido[14,15], Masafumi Sanefuji[16], Perry B. Shieh[1], Noriko Salamon[17], Ronald C. Kim[18, 19], Harry V. Vinters[1,20], Zugen Chen[2], Klaus Zerres[3], Monique M. Ryan[12], Stanley F. Nelson[2, 20, 21], & Joanna C. Jen[1].

[1]Department of Neurology, University of California, Los Angeles, U.S.A.
[2]Department of Human Genetics, University of California, Los Angeles, U.S.A.
[3]Institute of Human Genetics, Medical Faculty, University Hospital Rheinisch Westfälische Technische Hochschule (RWTH) Aachen, Germany
[4] Institute for Neuroscience and Muscle Research, Children's Hospital at Westmead, Westmead, Australia.
[5]Institute of Neuropathology, Medical Faculty, University Hospital RWTH Aachen, Germany
[6]Section of Pediatric Neurology, Children's Hospital & Department of Pediatrics and Child Health, University of Manitoba, Winnipeg, Manitoba, Canada
[7]Mount Sinai Hospital, The Prenatal Diagnosis and Medical Genetics Program, Department of Obstetrics and Gynecology, University of Toronto, Toronto, Ontario, Canada;
[8]The Hospital for Sick Children, Division of Clinical and Metabolic Genetics, Toronto, Ontario, Canada
[9]Department of Child Neurology, DNA Laboratory, 2nd School of Medicine, Charles University Prague and University Hospital Motol, the Czech Republic
[10]Department of Pediatrics, DRK-Kliniken Westend, Berlin, Germany
[11]Institute of Medical and Human Genetics, Charité Universitätsmedizin, Berlin, Germany
[12] Royal Children's Hospital, Murdoch Childrens Research Institute, University of Melbourne, Melbourne, Australia
[13] Servicio de Neuropediatría, Departamento de Pediatría, Hospital Clínico Universitario, Facultad de Medicina, Universidad de Santiago de Compostela, Santiago de Compostela, Spain.
[14] Fundación Pública Galega de Medicina Xenómica, Clinical Hospital of Santiago de Compostela, Servicio Galego de Saúde (SERGAS), Santiago de Compostela, Spain. [15]Center for Network Research on Rare Disorders (CIBERER), Institute of Health Carlos III, Barcelona, Spain
[16]Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan
[17]Department of Radiology, University of California, Los Angeles, U.S.A.
[18]Department of Pathology, University of California, Irvine, U.S.A.
[19]Department of Neurology, University of California, Irvine, U.S.A.
[20] Department of Pathology & Laboratory Medicine, University of California, Los Angeles, U.S.A.
[21] Department of Psychiatry, University of California, Los Angeles, U.S.A.

**Supplementary Table 1. Nerve conduction studies.** Nerve conduction studies in the **a.** oldest (18-year-old) and **b.** youngest (9-year-old) surviving patients in Family 1 showed motor responses with severely reduced amplitudes but normal sensory responses. The velocities were calculated based on onset latencies.

**a.** Nerve conduction study on the oldest surviving patient in Family 1

| | | | |
|---|---|---|---|
| **Sensory Studies** | | | |
| | Peak Latency | Amplitude | Velocity |
| Left Radial | 1.2 msec | 28.3 µV | 58 m/s |
| Left Sural | 2.0 msec | 11.7 µV | 50 m/s |
| **Motor Studies** | | | |
| | Distal Latency | Amplitude | Velocity |
| Left Ulnar | 3.0 msec | 1.0 mV | 47 m/s |
| Left Tibial | 4.8 msec | 0.3 mV | 37 m/s |

**b.** Nerve conduction study on the youngest surviving patient in Family 1

| | | | |
|---|---|---|---|
| **Sensory Studies** | | | |
| | Peak Latency | Amplitude | Velocity |
| Left Median | 1.8 msec | 48.7 µV | 61 m/s |
| **Motor Studies** | | | |
| | Distal Latency | Amplitude | Velocity |
| Left Median | 3.1 msec | 1.5 mV | 42 m/s |

## Supplementary Table 2. Oligonucleotides and PCR conditions for molecular genetic studies.

| Oligonucleotide Name | Direction | Oligonucleotide Sequence 5'-->3' | Tm | Anneal Temp | Amplicon Size (bp) |
|---|---|---|---|---|---|
| *Primers and PCR conditions for EXOSC3 mutation screening (lower case M13 sequence)* | | | | | |
| Exon 1 | Forward | tgtaaaacgacggccagtACGGCCATCAAGCTTCATAAAC | 54.9 | 67-60 | 539 |
| | Reverse | caggaaacagctatgaccCTCTTCTTTTGGGAGGTCTTCT | 50.4 | | |
| Exon 2 | Forward | tgtaaaacgacggccagtGGGGTGCCTAAGAGATAATGGAG | 55 | 68-55 | 441 |
| | Reverse | caggaaacagctatgaccGATAGCCTTCTGGATATGTGAGTGTTC | 55.7 | | |
| Exon 3 | Forward | tgtaaaacgacggccagtTCCCCAAGACTCAACTCCAAAG | 54.8 | 67-60 | 539 |
| | Reverse | caggaaacagctatgaccATCAGCCCACCAGAAACTACACAG | 56.2 | | |
| Exon 4 | Forward | tgtaaaacgacggccagtTGGAAGAAAGGAGGCAGCAAATG | 59.3 | 67-60 | 515 |
| | Reverse | caggaaacagctatgaccCACAAAAGCGTGGGTGAAAAC | 54.6 | | |
| *Primers and PCR conditions for In-Fusion HD cloning (lower case sequence required for cloning)* | | | | | |
| hEXOSC3-ATGf | Forward | taccgagctcggatccATGGCCGAACCTGCGTCTGTC | 61 | 72-67 | 4708 |
| hEXOSC3-Rcr | Reverse | gccctctagactcgagTTCCTCTGGTGAACCTGGCTTACTG | 60 | | |
| *Primers and PCR conditions for RT-PCR* | | | | | |
| hExosc3-c.423f | Forward | AGCCAGCTTCTTTGTCTTACTTGTC | 54.5 | 67-60 | |
| hExosc3-c.670r | Reverse | GTTTTCCCACTTCCTGTATGATTTC | 54.2 | | |
| *Primers for mutagenesis of zebrafish exosc3* | | | | | |
| zfExosc3-D102A | Forward | CTGGAGACGTCTTCAAAGTGGCCGTTGGAGGAAGTGAGC | 77.6 | | |
| | Reverse | GCTCACTTCCTCCAACGGCCACTTTGAAGACGTCTCCAG | 77.6 | | |
| zfExosc3-W208R | Forward | GCATGAACGGCAGAGTGCGGGTGAAGGCCAGAACCGTC | 82.5 | | |
| | Reverse | GACGGTTCTGGCCTTCACCCGCACTCTGCCGTTCATGC | 82.5 | | |
| zfExosc3-G20A | Forward | GGAGATGTGGTTCTTCCAGCCGACCTGCTGTTCTCCTTCAG | 78.4 | | |
| | Reverse | CTGAAGGAGAACAGCAGGTCGGCTGGAAGAACCACATCTCC | 78.4 | | |
| *Primers for mutagenesis of human EXOSC3* | | | | | |
| hEXOSC3-G31A | Forward | GTCAGGTGGTGCTCCCGGCTGAGGAGCTGCTCCTGCCG | 84.5 | | |
| | Reverse | CGGCAGGAGCAGCTCCTCAGCCGGGAGCACCACCTGAC | 84.5 | | |
| hEXOSC3-D132A | Forward | GAGATATATTCAAAGTTGCTGTTGGAGGGAGTGAG | 80 | | |
| | Reverse | CTCACTCCCTCCAACAGCAACTTTGAATATATCTC | 80 | | |
| hEXOSC3-W238R | Forward | TTGGAATGAATGGAAGAATACGGGTTAAGGCAAAAACCATC | 73.1 | | |
| | Reverse | GATGGTTTTTGCCTTAACCCGTATTCTTCCATTCATTCCAA | 73.1 | | |
| *Morpholino oligonucleotides for zebrafish exosc3* | | | | | |
| AUG MO | | TCCATGATGGAGGAGCGGAAAACAC | | | |
| SPL MO | | CCTCTTACCTCAGTTACAATTTATA | | | |

**Supplementary Table 3. Shared haplotypes spanning *EXOSC3* in Families 1-3.** Affected individuals from Families 1-3 with homozygous D132A mutations share approximately 1 cM of homozygosity with identical haplotypes around D132A (chr9:37751044-38395492 hg18 coordinates). This haplotype was not observed in 126 controls genotyped on the same platform. The findings are consistent with source of the mutation being a remote common ancestor.

| chr | pos_b36 | rsID | cM | FAM1 | FAM2 | FAM3 |
|---|---|---|---|---|---|---|
| 9 | 37751044 | rs1409145 | 62.0880 | BB | BB | BB |
| 9 | 37771455 | rs3827515 | 62.1171 | BB | BB | BB |
| 9 | 37772561 | rs7029518 | 62.1187 | BB | BB | BB |
| 9 | 37772708 | rs13294227 | 62.1189 | AA | AA | AA |
| 9 | 37772967 | rs10973542 | 62.1193 | BB | BB | BB |
| 9 | 37773990 |  | 62.1270 | 11 | 11 | 11 |
| 9 | 37779033 | rs10814621 | 62.1279 | BB | BB | BB |
| 9 | 37804051 | rs10814625 | 62.1636 | BB | BB | BB |
| 9 | 37831248 | rs10973580 | 62.2035 | AA | AA | AA |
| 9 | 37847067 | rs12000384 | 62.2267 | BB | BB | BB |
| 9 | 37848861 | rs16934508 | 62.2293 | AA | AA | AA |
| 9 | 37879753 | rs41436845 | 62.2749 | AA | AA | AA |
| 9 | 37882000 | rs16934574 | 62.2783 | BB | BB | BB |
| 9 | 37888768 | rs16934581 | 62.2885 | BB | BB | BB |
| 9 | 37891883 | rs4878724 | 62.2932 | AA | AA | AA |
| 9 | 37926503 | rs12002323 | 62.3456 | AA | AA | AA |
| 9 | 37947229 | rs7048063 | 62.3769 | AA | AA | AA |
| 9 | 37979893 | rs2243893 | 62.4270 | AA | AA | AA |
| 9 | 38014458 | rs2890783 | 62.4800 | BB | BB | BB |
| 9 | 38069516 | rs1999095 | 62.5736 | AA | AA | AA |
| 9 | 38080221 | rs4878183 | 62.5919 | BB | BB | BB |
| 9 | 38094144 | rs10973666 | 62.6150 | BB | BB | BB |
| 9 | 38109900 | rs7033592 | 62.6407 | BB | BB | BB |
| 9 | 38131337 | rs2890776 | 62.6757 | AA | AA | AA |
| 9 | 38132145 | rs1928239 | 62.6770 | AA | AA | AA |
| 9 | 38141190 | rs1001959 | 62.6918 | BB | BB | BB |
| 9 | 38141268 | rs10973683 | 62.6919 | BB | BB | BB |
| 9 | 38177709 | rs10973695 | 62.7516 | AA | AA | AA |
| 9 | 38178524 | rs7034598 | 62.7529 | AA | AA | AA |
| 9 | 38192630 | rs2585669 | 62.7760 | BB | BB | BB |
| 9 | 38193508 | rs2810740 | 62.7775 | AA | AA | AA |
| 9 | 38196362 | rs2053556 | 62.7822 | BB | BB | BB |
| 9 | 38283902 | rs341474 | 62.9442 | AA | AA | AA |
| 9 | 38299503 | rs1885491 | 62.9740 | AA | AA | AA |
| 9 | 38320114 | rs16935064 | 63.0096 | BB | BB | -- |
| 9 | 38352713 | rs1022770 | 63.0659 | BB | BB | BB |
| 9 | 38364977 | rs2181139 | 63.0870 | AA | AA | AA |
| 9 | 38373549 | rs17451412 | 63.1025 | BB | BB | -- |
| 9 | 38394292 | rs12336048 | 63.1406 | AA | AA | AA |
| 9 | 38395492 | rs4878203 | 63.1428 | BB | BB | BB |

**Supplementary Table 4. *In silico* predictions of pathogenicity.** Missense *EXOSC3* mutations were assessed by various algorithms of phylogenetic conservation and functional impact. PhastCons (PHylogenetic Analysis with Space/Time models; evolutionary conservation) http://compgen.bscb.cornell.edu/phast/ spans between 0 and 1, with 1 being the most highly conserved.

GERP (Genomic Evolutionary Rate Profiling) http://mendel.stanford.edu/SidowLab/downloads/gerp/index.html is designated between -11.6 to maximum conservation at 5.82.

Grantham scores, which categorize codon replacements into classes of increasing chemical dissimilarity, are designated conservative (0-50), moderately conservative (51-100), moderately radical (101-150), or radical (≥151).[1]

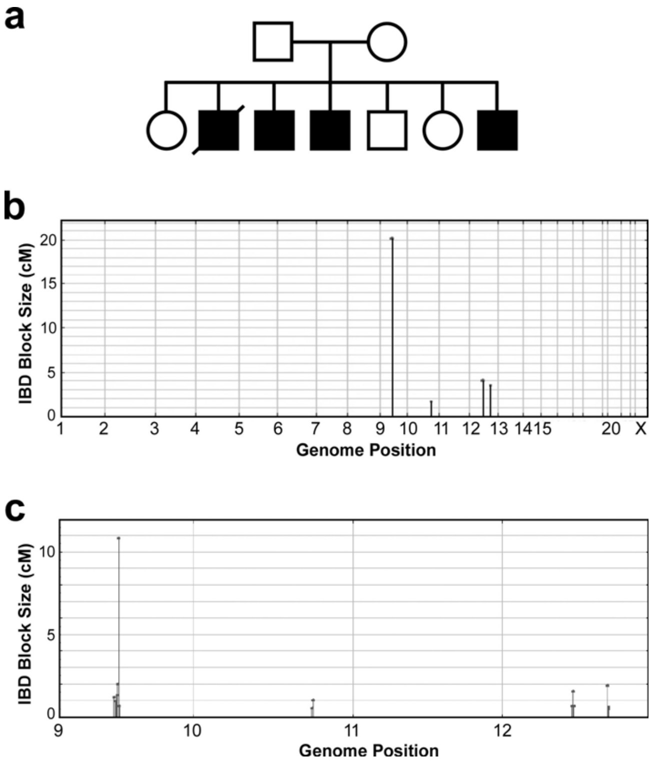
PolyPhen (Polymorphism Phenotype) scores are designated probably damaging (≥2.00), possibly damaging (1.50-1.99), potentially damaging (1.25-1.49), borderline (1.00-1.24), or benign (0.00-0.99).[2]

SIFT (Sorting Intolerant from Tolerant; http://sift.jcvi.org/ ) scores are designated damaging (<0.05) or not.

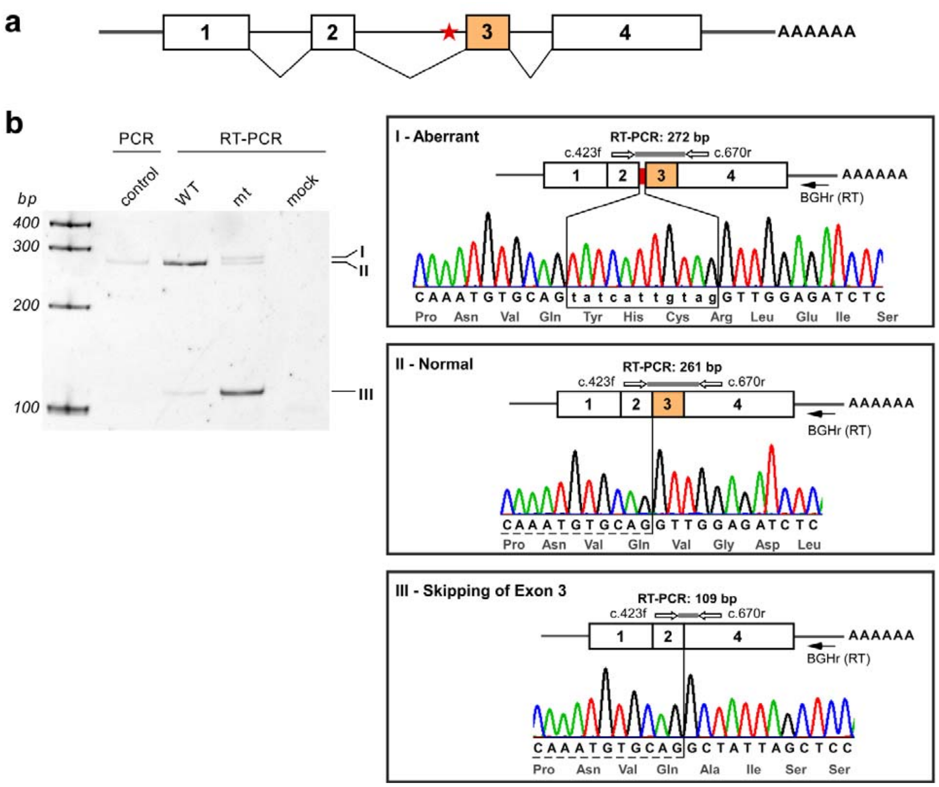
| Mutation | PhastCons | GERP | Grantham | PolyPhen | SIFT |
|---|---|---|---|---|---|
| G31A | 1 | 5.27 | 60 | 1.35 (potentially damaging) | 0.01 (damaging) |
| D132A | 1 | 5.34 | 126 | 2.493 (probably damaging) | 0.02 (damaging) |
| A139P | 1 | 5.34 | 27 | 1.985 (possibly damaging) | 0 (damaging) |
| W238R | 1 | 5.66 | 101 | 4.142 (probably damaging) | 0 (damaging) |

**Supplementary Table 5. Survival assays in zebrafish embryos.** Survival assays were performed in antisense *exosc3*-specific splice morpholino-injected zebrafish embryos that were co-injected with *in vitro* transcribed wildtype or mutant *exosc3* mRNA, with GFP as control. Embryos were scored as normal or abnormal (mildly abnormal, severely abnormal, and dead), as stratified in **Supplementary Figure 4** for embryos 1 dpf. For embryos 2 or 3 dpf (shown in **Figure 3**), those with straight spine of normal length and normal brain were scored as "normal"; those with curved spine and small brain but still mobile were scored as "mildly abnormal"; and those that were severely malformed without movement were scored as "severely abnormal".

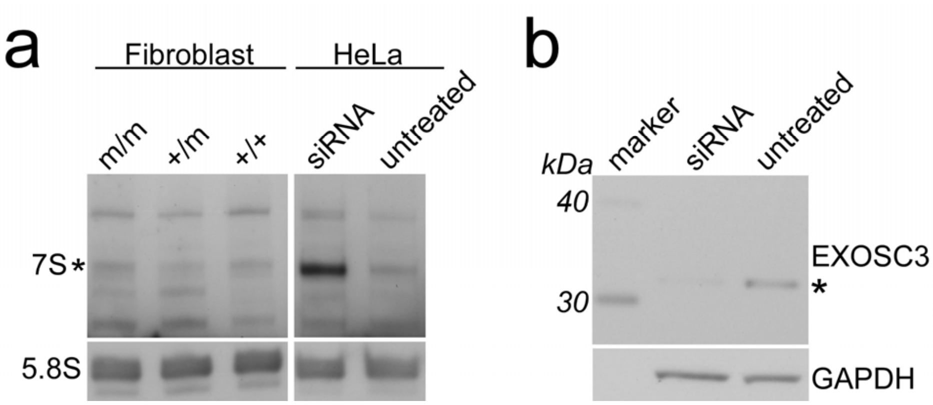
|  | **dpf** | **normal** | **mildly abnormal** | **severely abnormal** | **dead** |
|---|---|---|---|---|---|
| SPL MO + zWT mRNA | 1 | 248 | 5 | 3 | 33 |
|  | 2 | 240 | 9 | 1 | 39 |
|  | 3 | 206 | 39 | 5 | 39 |
| SPL MO + zDA mRNA | 1 | 56 | 37 | 25 | 33 |
|  | 2 | 50 | 32 | 32 | 37 |
|  | 3 | 35 | 38 | 21 | 57 |
| SPL MO + zGA mRNA | 1 | 69 | 110 | 27 | 27 |
|  | 2 | 65 | 48 | 84 | 36 |
|  | 3 | 32 | 57 | 72 | 72 |
| SPL MO + zWR mRNA | 1 | 68 | 50 | 45 | 39 |
|  | 2 | 73 | 30 | 46 | 53 |
|  | 3 | 44 | 45 | 37 | 76 |
| SPL MO + hWT mRNA | 1 | 129 | 24 | 27 | 29 |
|  | 2 | 112 | 35 | 17 | 45 |
|  | 3 | 94 | 31 | 18 | 66 |
| SPL MO + hDA mRNA | 1 | 14 | 12 | 29 | 20 |
|  | 2 | 18 | 6 | 22 | 29 |
|  | 3 | 11 | 6 | 11 | 47 |
| SPL MO + hGA mRNA | 1 | 38 | 43 | 53 | 40 |
|  | 2 | 36 | 23 | 65 | 50 |
|  | 3 | 23 | 20 | 37 | 94 |
| SPL MO + hWR mRNA | 1 | 12 | 32 | 49 | 56 |
|  | 2 | 11 | 22 | 56 | 60 |
|  | 3 | 7 | 13 | 30 | 99 |
| SPL MO + GFP mRNA | 1 | 13 | 55 | 3 | 27 |
|  | 2 | 6 | 55 | 5 | 32 |
|  | 3 | 6 | 47 | 6 | 39 |



**Supplementary Figure 1. Genome-wide SNP Genotyping & Linkage Analysis. a.** Pedigree of Family 1 with four affected siblings, three unaffected siblings, and their parents. Genotyping on Affymetrix 250K NspI mapping array was performed on DNA from the four affected siblings, three unaffected siblings, and their parents. Copy number analysis was used to look for large deletions or duplications shared by the affected individuals; none were found. Pedigree-free IBD mapping was performed to search for intervals that were compatible with a common extended haplotype among all the affected individuals, and not shared by unaffected family members. **b.** For the most general compound recessive model with exclusion of the unaffecteds, we searched for identical inheritance blocks to find an aggregate of 15 Mb candidate region in 4 loci on three chromosomes 1) where all four affecteds inherited the same allele from the mother and the father and 2) that these inheritance blocks were shared by the affected but not by unaffected siblings. **c.** The most limited recessive model under this analysis, requiring a single founder allele and exclusion of unaffected, identified 15 blocks >0.5 cM that were identical and homozygous in all four affected and not identical in any affected and unaffected. These loci fit the model of an old founder ~10-20 generations back, with no recent inbreeding. This analysis excluded the possibility of an X-linked disorder.

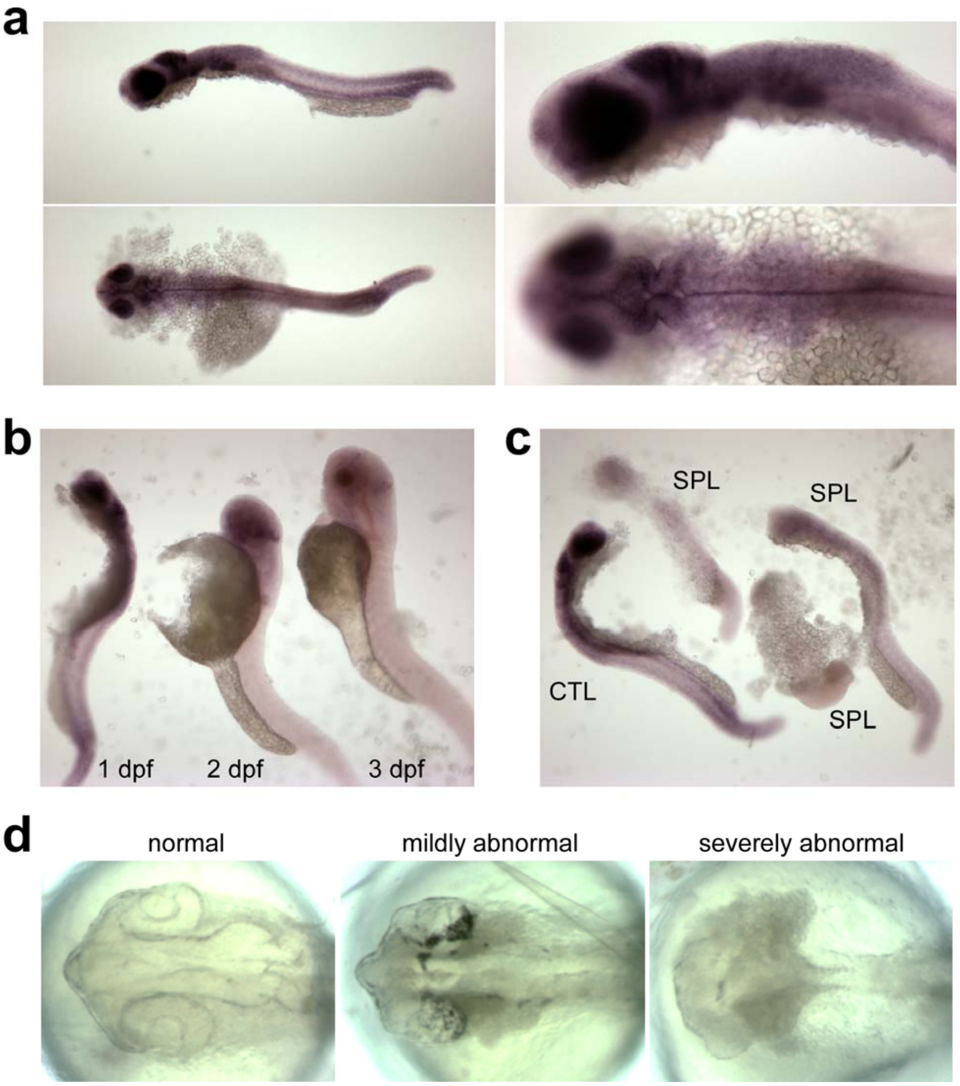

a
1
2
3
4
AAAAAA
b
PCR
RT-PCR
control
WT
mt
mock
bp
400
300
200
100
I
II
III
I - Aberrant
RT-PCR: 272 bp
c.423f
c.670r
BGHr (RT)
C A A A T G T G C A G t a t c a t t g t a g G T T G G A G A T C T C
Pro Asn Val Gln Tyr His Cys Arg Leu Glu Ile Ser
II - Normal
RT-PCR: 261 bp
C A A A T G T G C A G G T T G G A G A T C T C
Pro Asn Val Gln Val Gly Asp Leu
III - Skipping of Exon 3
RT-PCR: 109 bp
C A A A T G T G C A G G C T A T T A G C T C C
Pro Asn Val Gln Ala Ile Ser Ser

**Supplementary Figure 2. Aberrant splicing from the intronic mutation c.475-12A>G. a.** Diagrammatic representation of the four exons and the location of the intronic mutation (represented by the star in red). **b.** RT-PCR products run on a 12% polyacrylamide gel. Transcript-specific RT-PCR yielded two products from the cells transfected with the wild type construct: a major product **II** (88% by densitometry)**,** which was of the same size as the PCR product from a full-length EXOSC3 cDNA clone (control), and a minor band of smaller size **III** (12% by densitometry). Sanger sequencing confirmed normal splicing of intron 2 (**II-Normal**) and skipping of exon 3 (**III-Skipping of Exon 3**) and shifting the open reading frame. Transcript-specific RT-PCR yielded three products from cells transfected with the mutant clone harboring the intronic variant. The major band (**III**) was missing exon 3 (85% by densitometry), as confirmed by sequencing. There were two minor bands of larger sizes. The lower band (**II**) was of the same size as the control (6% by densitometry) and confirmed by sequencing. The upper band (**I**) was the product of aberrant splicing using the newly introduced splice acceptor site (9% by densitometry), with the incorporation of 11 additional nucleotides upstream from the normal splice site to shift the open reading frame, as demonstrated by sequencing (**I-Aberrant**).



**Supplementary Figure 3. rRNA processing. a.** Northern blots demonstrating no marked accumulation of unprocessed 5.8S rRNA (*7S, as seen in siRNA-treated HeLa cells) in a patient or his parent compared to an unaffected control. 4μg RNA extracted from patient-derived fibroblasts or HeLa was loaded to each lane on 6% denaturing PAGE and detected with DIG-labeled probe for 5.8S rRNA. m/m-homozygous for mutant D132A; +/m-heterozygous for D132A; +/+-normal control homozygous wildtype; siRNA- HeLa cells treated with EXOSC3-specific siRNA; untreated- HeLa cells not exposed to siRNA. **b.** Western blot demonstrating diminished EXOSC3 in HeLa cells treated with EXOSC3-specific siRNA. 10μg protein extracted from cells treated or untreated with siRNA probed first with EXOSC3-specific antibodies (Santa Cruz Biotechnology), stripped, then probed with GAPDH-specific antibodies (Millipore).

Nature Genetics: doi:10.1038/ng.2254



**Supplementary Figure 4. *Exosc3* expression in zebrafish embryos by whole-mount *in situ* hybridization. a.** Control morpholino-injected zebrafish embryos 1 dpf probed with *Exosc3*-specific antisense riboprobes demonstrating diffuse expression with especially strong signal intensity in the brain and eyes, in lateral and dorsal views under different magnifications. **b.** There is a progressive decrease in the expression of exosc3 in zebrafish embryos 1 dpf compared to those at 2 dpf and 3 dpf. The yolk sacs were partially removed to not obscure the visualization of the embryos. **c.** Zebrafish embryos 1 dpf injected with antisense morpholinos targeting the splice site of exosc3 (SPL) led to a marked but variable decrease in the expression of exosc3 compared to those injected with control morpholinos (CTL). There was higher expression of exosc3 in the embryo on the far right compared to the other two SPL-injected embryos. **d.** Light microscopic examination in dorsal view of unstained live chorion-enveloped zebrafish embryos 1 dpf revealed variable phenotypes. Embryos with smaller brain and eyes are scored as mildly abnormal. Those embryos with very small brain and malformed eyes as well as a very small hindbrain and thin spinal cord are scored as severely abnormal.

## References

1. Li WH, Wu CI, Luo CC. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J Mol Evol* **21**, 58–71(1984).
2. Xi T, Jones IM, Mohrenweiser HW. Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function. *Genomics* **83**, 970–979 (2004).

# CHAPTER EIGHT

# Conclusions

*General conclusions*

Eric Lander, first author of the 2001 paper that presented the initial sequencing of the human genome (1), more recently noted that "[t]he human genome has had a certain tendency to incite passion and excess" (2). The work presented here justifies a certain passion for unlocking the secrets of genetic diseases and suggests caution against excessive expectations. At the outset we sought to employ a new and powerful technology to understand the molecular basis of severe congenital disorders of the intestine and brain. In this endeavor we had success: we pinpointed many unsuspected mutations interfering with intestinal absorption in genes that already were known to medicine; we identified the probable, and possibly predominant, genetic cause of sporadic congenital intestinal motility disorders; and we discovered a surprising link between RNA processing and cerebellar and spinal motor neuron survival. In a few instances, this knowledge can immediately translate into improved disease management; in all, it can bring closure to patients and families, provide guidance for pregnancy planning, and can open doors for further basic research on the underlying mechanisms and pathways and for the development of new treatments.

During the course of this project the field of high throughput sequencing matured rapidly. Cost per base plummeted and read lengths and read quality improved as a result of technical improvements in sequencing platforms (3). Software developers have produced tools for read quality assessment (4-12), mapping and alignment of short reads (13-24), genotyping and variant identification(25-42), and variant annotation(43-103). The 1000 Genomes project and NHLBI released exome sequencing results for data from thousands of individuals (104, 105). Nonetheless, many challenges remain for the discovery of rare variants that cause Mendelian diseases. Foremost among these is the difficulty of obtaining sufficient numbers of samples from consented and well phenotyped individuals. Power to detect a single casual gene for a Mendelian disorder rises dramatically when two unrelated samples are sequenced. Multi-center projects such as a proposed study of intestinal failure can address this problem. Another practical concern is storing the large and growing amount of aligned data; hardware, maintenance, and physical infrastructure are costly, and benefit less from economies of scale and innovation than has the generation of sequencing data.

Existing tools reliably detect single nucleotide variants with acceptable false positive rates. The detection of small indels is somewhat less reliable, but software tools such as the GATK Haplotype Caller are improving in both sensitivity and specificity.  A number of tools attempt to call larger scale duplications and deletions from exome data (106-114), but there is little consistency among them in our experience. Exon capture variability may prove to introduce too much noise for this technique to become reliable. Complex rearrangements can be identified with paired end reads of whole genome data (115-131), but most such events cannot be detected with exome sequencing.

As this research progressed, we became aware of a number of limitations and pitfalls. Most importantly, at the outset we underestimated the likelihood that the diseases we studied might be due to sporadic de novo mutations. The prevailing medical genetics literature has doubtless underestimated the contribution of de novo mutations to disease burden because of the difficulty of identifying such mutations in the absence of *a priori* reasons to investigate a particular genomic region. Most discoveries had been made in extended families showing autosomal dominant or recessive, X-linked, or mitochondrial inheritance patterns. Indeed, in the congenital diarrhea disorders (CDDs) pilot study, all of the mutations we identified have a recessive effect on phenotype. But this was unsurprising precisely because we looked for mutations in genes reported in the literature, which typically had been found by a family linkage methodology. Our identification of de novo dominant acting mutations in half of the chronic intestinal pseudo-obstruction (CIPO) cohort highlighted the possible importance of sporadic mutations more generally in families with no prior history of disease. And in the CIPO cases, publication of a milder autosomal dominant familial form of CIPO helped us weed out several other candidate genes in which members of the CIPO cohort shared mutations.

The estimated baseline rate of germline mutation in the human genome is currently believed to be 1.18 x $10^{-8}$ per nucleotide or 74 de novo mutations per generation of which ~1.19 mutations per generation affect the coding region and ~78% of these (~0.9 mutations) are predicted to alter the protein sequence (132-137). The rate is highly variable within and between families (138), and increases with paternal age (139). Although there may only be a few true potentially damaging de novo variants in each exome, there will be many seeming possibilities in the exome sequencing results of a single individual because of the large number of inherited heterozygous variants, which can be filtered out by sequencing trios (both parents and affected child) and excluding any variants that are seen in either parent. False positives are a more intractable problem. These can be due to false positive variant calls in the proband or by false negative variant calls in the parents causing the variant in the proband falsely seem to be de novo. We sequenced only a few trios in the CDD (*n*=8) and CIPO (*n*=4) studies. Coincidentally none of the CIPO trios had *ACTG2* mutations. Moreover, even with filtering it can be difficult to screen out false positives without loosing sensitivity. For example, using our standard filters for allele frequency, annotation, and segregation filters, we observed a mean of 39 putative variants per proband across all trios, but with stricter mapping quality (QUAL>=500) and allele balance (40-60%) filters, as might be used for clinical sequencing, the mean number of de novos mutation calls was just two. Although this approximates the 'true' number of de novo germline variants, if we had applied strict filtering to the CIPO cohort, we would not have found the *ACTG2* mutations in three of the cases. Further work may be needed to find the right balance between sensitivity and specificity in variant detection, and the balance may be different in a clinical context where avoiding false positives is critical, versus a research context where sensitivity to novel findings may take precedence.

We also failed to find with exome sequencing an *ACTG2* mutation in one of the cases that we later confirmed by Sanger sequencing. This may have been due to the relatively poor coverage of this sample (*mean*=76X, 86%>=20X), illustrating that variability of coverage is another pitfall of whole-exome sequencing. Moreover, we used two different modules from the Genome Analysis Toolkit (GATK) for genotyping, Unified Genotyper and Haplotype Caller. The latter is now recommended by the GATK's developers for projects such as the present work; unlike the Unified Genotyper, the Haplotype Caller performs a local de novo assembly to resolve misaligned reads and is said to be more accurate and more sensitive to indels. In our experience, the Haplotype caller succeeds in calling more and longer indels. Nonetheless, with respect to the de novo mutations in the CIPO cases, both methods found five of the variants, but only the Unified Genotyper selected two others, yet Sanger sequencing confirmed all of these.

Even more challenging than the technical issues, is the process of determining whether novel mutations, especially missense or splice regulatory region mutations, are truly harmful. One report found that human genomes contain over 1000 apparently loss-of-function variants, of which ~100 (~20 homozygous) were deemed genuine after strict filtering for false positives; many genes appeared to be tolerant of mutations that would be predicted to be severely damaging (140). Much work remains to be done to algorithmically predict which variants are genuinely damaging and likely to be responsible for disease. Ideally, one would want to demonstrate that a variant causes the expected phenotype *in vivo*, but this can be costly and time consuming. For example, we were able to confirm the effect of the *EXOSC3* mutation by experiments in a model organism. In this case, with four affected siblings, whole-exome sequencing produced only one candidate that needed to be confirmed. More typically, we find a number of mutations that are predicted to be damaging and the effort involved in narrowing down the candidates may be daunting. Many researchers are now attempting to develop improved bioinformatic methods for predicting whether a given variant is likely to cause harm to the organism. We expect some improvements in this area, but it is unlikely to become a panacea.

Finding the same mutation, or other damaging mutations in the same gene, as we did with CIPO, is quite powerful. The difficulty, however, is that the class of disorders we study are quite rare, limiting the number of cases available for study. One estimate suggests that exome sequencing of two to five unrelated subjects is necessary for adequate power to identify rare dominant-acting variants in casual genes (141). Furthermore, despite sharing common syndromes, the diseases are genetically and phenotypically heterogeneous. A large multi-center program to sequence and rigorously phenotype many subjects, such as one we have proposed elsewhere for intestinal failure, is a promising approach.

It also will be necessary to develop high-throughput screening methods for each class of disease. We are currently developing methods to model the intestine in cell cultures derived from embryonic stem cells, induced pluripotent stem cells reprogrammed from patient fibroblasts, and intestinal progenitor stem cells from patient biopsy tissue. We are developing methods for overexpression of mutant and wild-type alleles in these systems, and evaluating the resulting phenotypes. In the near future, methods for specifically editing the genome, such as TALENs or CRISPR/Cas systems (142) will permit more precise recapitulation of the disease state and the development of methods to repair patient-derived organoids looking towards future therapeutic modalities. High throughput assays with model organisms could be a powerful method for quickly and inexpensively screening candidate mutations. For example, libraries of gene knockdown morpholinos for use with zebrafish, or siRNA for use with *C. elegans*, can be used to run many experiments in parallel. Still, mapping human diseases onto other organisms is not straightforward, and a major screening program will involve much effort and expense.

*Specific conclusions*

Our VAX method, which uses of local installation of the Ensembl databases and Perl API, provides a robust and flexible framework for annotating DNA sequencing variants from many different data sources using Variant Effect Predictor plugin modules. We have outlined the design and usage of VEP plugins for a number of widely used databases. In addition, modules may be easily designed for incorporating annotations from any external dataset that is kept in a flat file or relational database, such as the Zebrafish Model Organism database (143), and the Rat Genome database (144).  We have used the VAX system for the discovery of the causes of rare Mendelian diseases and genes involved in psychiatric disorders. (145-147). VAX is used routinely for CLIA/CAP-accredited whole exome sequencing by the UCLA Clinical Genomics Center, which has processed more than 1000 exomes to date (148).

The CDD pilot project successfully identified a probable molecular explanation for the disorder in a majority of cases, thus suggesting the value of this approach for diagnosis of CDDs in a clinical setting. The case of the child with a PCSK1 mutation, illustrates how timely use of whole-exome sequencing could reduce medical costs and patient misery. However, our study also demonstrates that much remains unknown about the genetic etiology of CDDs, as 40% of the probands we sequenced did not yield a clear molecular finding, although many interesting and novel candidate genes have been identified. There are several factors contributing to this “genetic dark matter”. Most importantly, the cellular pathways involved in the development and function of the intestine are not fully understood and many genes are yet to be identified that contribute to the risk of CDDs. The present study generated a number of candidate genes that are now the subject of active research. As discussed above, de novo mutations, which the study was not well powered to identify, may account for some of the unsolved cases. A third source of genetic dark matter is the 98% of the human genome that is not in the exome, which includes promoters, enhancers, short RNAs, and other regulatory elements. Identifying and characterizing these features is an active area of research, but at this time the size of the genome to be sequenced and the lack of data on these features limits our ability to determine how they affect phenotypes.

Finding that familial visceral myopathy in one family and a subset of CIPO cases appear to have a common genetic cause resolves a longstanding mystery. Our finding at this point is preliminary and it will be necessary in future work to demonstrate at the cellular or organism level that each mutation causes a disordered phenotype. Beyond that, much remains to be discovered about the actual molecular mechanism that causes a point mutation in one form of actin to have a dominant effect. It could involve disruption of actin filament polymerizability or stability, altered or blocked interactions with actin regulatory proteins, changes to actin/myosin interactions, or weakening of actin filament structure (149).

Finally, the finding that the RNA exosome seems to play a unique role in the survival of cerebellar and spinal motor neurons ties in with work on understanding how RNA regulation is involved in motor and cerebellar degeneration.

# References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, International Human Genome Sequencing C. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921. Epub 2001/03/10. doi: 10.1038/35057062. PubMed PMID: 11237011.

2. Lander ES. Initial impact of the sequencing of the human genome. Nature. 2011;470(7333):187-97. Epub 2011/02/11. doi: 10.1038/nature09792. PubMed PMID: 21307931.

3. Mardis ER. Next-generation sequencing platforms. Annual review of analytical chemistry. 2013;6:287-303. Epub 2013/04/09. doi: 10.1146/annurev-anchem-062012-092628. PubMed PMID: 23560931.

4. Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A, Galaxy T. Manipulation of FASTQ data with Galaxy. Bioinformatics. 2010;26(14):1783-5. Epub 2010/06/22. doi: 10.1093/bioinformatics/btq281. PubMed PMID: 20562416; PubMed Central PMCID: PMC2894519.

5. Cibulskis K, McKenna A, Fennell T, Banks E, DePristo M, Getz G. ContEst: estimating cross-contamination of human samples in next-generation sequencing data. Bioinformatics. 2011;27(18):2601-2. Epub 2011/08/02. doi: 10.1093/bioinformatics/btr446. PubMed PMID: 21803805; PubMed Central PMCID: PMC3167057.

6. Cox MP, Peterson DA, Biggs PJ. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics. 2010;11:485. Epub 2010/09/30. doi: 10.1186/1471-2105-11-485. PubMed PMID: 20875133; PubMed Central PMCID: PMC2956736.

7. Dai M, Thompson RC, Maher C, Contreras-Galindo R, Kaplan MH, Markovitz DM, Omenn G, Meng F. NGSQC: cross-platform quality analysis pipeline for deep sequencing data. BMC Genomics. 2010;11 Suppl 4:S7. Epub 2010/12/22. doi: 10.1186/1471-2164-11-S4-S7. PubMed PMID: 21143816; PubMed Central PMCID: PMC3005923.

8. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. 2008;36(16):e105. Epub 2008/07/29. doi: 10.1093/nar/gkn425. PubMed PMID: 18660515; PubMed Central PMCID: PMC2532726.

9. Dolan PC, Denver DR. TileQC: a system for tile-based quality control of Solexa data. BMC Bioinformatics. 2008;9:250. Epub 2008/05/30. doi: 10.1186/1471-2105-9-250. PubMed PMID: 18507856; PubMed Central PMCID: PMC2443380.

10. Martinez-Alcantara A, Ballesteros E, Feng C, Rojas M, Koshinsky H, Fofanov VY, Havlak P, Fofanov Y. PIQA: pipeline for Illumina G1 genome analyzer data quality assessment. Bioinformatics. 2009;25(18):2438-9. Epub 2009/07/16. doi: 10.1093/bioinformatics/btp429. PubMed PMID: 19602525; PubMed Central PMCID: PMC2735671.

11. Planet E, Attolini CS, Reina O, Flores O, Rossell D. htSeqTools: high-throughput sequencing quality control, processing and visualization in R. Bioinformatics. 2012;28(4):589-90. Epub 2011/12/27. doi: 10.1093/bioinformatics/btr700. PubMed PMID: 22199381.

12. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;27(6):863-4. Epub 2011/02/01. doi: 10.1093/bioinformatics/btr026. PubMed PMID: 21278185; PubMed Central PMCID: PMC3051327.

13. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, Sahinalp SC, Gibbs RA, Eichler EE. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet. 2009;41(10):1061-7. Epub 2009/09/01. doi: 10.1038/ng.437. PubMed PMID: 19718026; PubMed Central PMCID: PMC2875196.

14. Galinsky VL. YOABS: yet other aligner of biological sequences--an efficient linearly scaling nucleotide aligner. Bioinformatics. 2012;28(8):1070-7. Epub 2012/03/10. doi: 10.1093/bioinformatics/bts102. PubMed PMID: 22402614.

15. Homer N, Merriman B, Nelson SF. BFAST: an alignment tool for large scale genome resequencing. PLoS One. 2009;4(11):e7767. Epub 2009/11/13. doi:

10.1371/journal.pone.0007767. PubMed PMID: 19907642; PubMed Central PMCID: PMC2770639.

16. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357-9. Epub 2012/03/06. doi: 10.1038/nmeth.1923. PubMed PMID: 22388286; PubMed Central PMCID: PMC3322381.

17. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25. Epub 2009/03/06. doi: 10.1186/gb-2009-10-3-r25. PubMed PMID: 19261174; PubMed Central PMCID: PMC2690996.

18. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60. Epub 2009/05/20. doi: 10.1093/bioinformatics/btp324. PubMed PMID: 19451168; PubMed Central PMCID: PMC2705234.

19. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589-95. Epub 2010/01/19. doi: 10.1093/bioinformatics/btp698. PubMed PMID: 20080505; PubMed Central PMCID: PMC2828108.

20. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008;18(11):1851-8. Epub 2008/08/21. doi: 10.1101/gr.078212.108. PubMed PMID: 18714091; PubMed Central PMCID: PMC2577856.

21. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009;25(15):1966-7. Epub 2009/06/06. doi: 10.1093/bioinformatics/btp336. PubMed PMID: 19497933.

22. Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 2011;21(6):936-9. Epub 2010/10/29. doi: 10.1101/gr.111120.110. PubMed PMID: 20980556; PubMed Central PMCID: PMC3106326.

23. Ning Z, Cox AJ, Mullikin JC. SSAHA: a fast search method for large DNA databases. Genome Res. 2001;11(10):1725-9. Epub 2001/10/10. doi: 10.1101/gr.194201. PubMed PMID: 11591649; PubMed Central PMCID: PMC311141.

24. Yu X, Guda K, Willis J, Veigl M, Wang Z, Markowitz S, Adams MD, Sun S. How do alignment programs perform on sequencing data with varying qualities and from repetitive regions? BioData mining. 2012;5(1):6. Epub 2012/06/20. doi: 10.1186/1756-0381-5-6. PubMed PMID: 22709551; PubMed Central PMCID: PMC3414812.

25. Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R. Dindel: accurate indel calls from short-read data. Genome Res. 2011;21(6):961-73.

26. Altmann A, Weber P, Quast C, Rex-Haffner M, Binder EB, Muller-Myhsok B. vipR: variant identification in pooled DNA using R. Bioinformatics. 2011;27(13):77-84.

27. Bansal VK, Misra MC, Krishna A, Kumar S, Garg P, Khan RN, Loli A, Jindal V. Pancreatic hydatid cyst masquerading as cystic neoplasm of pancreas. Trop Gastroenterol. 2010;31(4):335-7.

28. Browning BL, Yu Z. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. Am J Hum Genet. 2009;85(6):847-61.

29. Challis D, Yu J, Evani US, Jackson AR, Paithankar S, Coarfa C, Milosavljevic A, Gibbs RA, Yu F. An integrative variant analysis suite for whole exome next-generation sequencing data. BMC Bioinformatics. 2012;13:8-.

30. Chen K, McLellan MD, Ding L, Wendl MC, Kasai Y, Wilson RK, Mardis ER. PolyScan: an automatic indel and SNP detection approach to the analysis of human resequencing data. Genome Res. 2007;17(5):659-66.

31. Dalca AV, Rumble SM, Levy S, Brudno M. VARiD: a variation detection framework for color-space and letter-space platforms. Bioinformatics. 2010;26(12):343-9.

32. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491-8.

33. Edmonson MN, Zhang J, Yan C, Finney RP, Meerzaman DM, Buetow KH. Bambino: a variant detector and alignment viewer for next-generation sequencing data in the SAM/BAM format. Bioinformatics. 2011;27(6):865-6.

34. Goya R, Sun MGF, Morin RD, Leung G, Ha G, Wiegand KC, Senz J, Crisan A, Marra MA, Hirst M, Huntsman D, Murphy KP, Aparicio S, Shah SP. SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. Bioinformatics. 2010;26(6):730-6.

35. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5(6).

36. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet. 2012;44(2):226-32.

37. Le SQ, Durbin R. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. Genome Res. 2011;21(6):952-60.

38. Lee S, Hormozdiari F, Alkan C, Brudno M. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. Nat Methods. 2009;6(7):473-4.

39. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.

40. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010;34(8):816-34.

41. Malhis N, Jones SJM. High quality SNP calling using Illumina data at shallow coverage. Bioinformatics. 2010;26(8):1029-35.

42. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, Boucher G, Ripke S, Ellinghaus D, Burtt N, Fennell T, Kirby A, Latiano A, Goyette P, Green T, Halfvarson J, Haritunians T, Korn JM, Kuruvilla F, Lagace C, Neale B, Lo KS, Schumm P, Torkvist L, Dubinsky MC, Brant SR, Silverberg MS, Duerr RH, Altshuler D, Gabriel S, Lettre G, Franke A, D'Amato M, McGovern DPB, Cho JH, Rioux JD, Xavier RJ, Daly MJ. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet. 2011;43(11):1066-73.

43. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-9.

44. Asthana S, Roytberg M, Stamatoyannopoulos J, Sunyaev S. Analysis of sequence conservation at nucleotide resolution. PLoS computational biology. 2007;3(12).

45. Banerji S, Cibulskis K, Rangel-Escareno C, Brown KK, Carter SL, Frederick AM, Lawrence MS, Sivachenko AY, Sougnez C, Zou L, Cortes ML, Fernandez-Lopez JC, Peng S, Ardlie KG, Auclair D, Bautista-Pina V, Duke F, Francis J, Jung J, Maffuz-Aziz A, Onofrio RC, Parkin M, Pho NH, Quintanar-Jurado V, Ramos AH, Rebollar-Vega R, Rodriguez-Cuevas S, Romero-Cordoba SL, Schumacher SE, Stransky N, Thompson KM, Uribe-Figueroa L, Baselga J, Beroukhim R, Polyak K, Sgroi DC, Richardson AL, Jimenez-Sanchez G, Lander ES, Gabriel SB, Garraway LA, Golub TR, Melendez-Zajgla J, Toker A, Getz G, Hidalgo-Miranda A, Meyerson M. Sequence analysis of mutations and translocations across breast cancer subtypes. Nature. 2012;486(7403):405-9.

46. Bao H, Guo H, Wang J, Zhou R, Lu X, Shi S. MapView: visualization of short reads alignment on a desktop computer. Bioinformatics. 2009;25(12):1554-5.

47. Bromberg Y, Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res. 2007;35(11):3823-35.

48. Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat. 2009;30(8):1237-44.

49. Capriotti E, Arbiza L, Casadio R, Dopazo J, Dopazo H, Marti-Renom MA. Use of estimated evolutionary strength at the codon level improves the prediction of disease-related protein mutations in humans. Hum Mutat. 2008;29(1):198-204.

50. Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics. 2006;22(22):2729-34.

51. Capriotti E, Fariselli P, Calabrese R, Casadio R. Predicting protein stability changes from sequences using support vector machines. Bioinformatics. 2005;21 Suppl 2:54-8.

52. Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR. ESEfinder: A web resource to identify exonic splicing enhancers. Nucleic Acids Res. 2003;31(13):3568-71.

53. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, Beroukhim R, Pellman D, Levine DA, Lander ES, Meyerson M, Getz G. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012;30(5):413-21.

54. Chelala C, Khan A, Lemoine NR. SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. Bioinformatics. 2009;25(5):655-61.

55. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, Lu X. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front Genet. 2012;3:35-.

56. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS computational biology. 2010;6(12).

57. Fairbrother WG, Yeh R-F, Sharp PA, Burge CB. Predictive identification of exonic splicing enhancers in human genes. Science. 2002;297(5583):1007-13.

58. Freimuth RR, Stormo GD, McLeod HL. PolyMAPr: programs for polymorphism database mining, annotation, and functional analysis. Hum Mutat. 2005;25(2):110-7.

59. Gamazon ER, Zhang W, Konkashbaev A, Duan S, Kistner EO, Nicolae DL, Dolan ME, Cox NJ. SCAN: SNP and copy number annotation. Bioinformatics. 2010;26(2):259-62.

60. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. 2009;25(12):54-62.

61. Ge D, Ruzzo EK, Shianna KV, He M, Pelak K, Heinzen EL, Need AC, Cirulli ET, Maia JM, Dickson SP, Zhu M, Singh A, Allen AS, Goldstein DB. SVA: software for annotating and visualizing sequenced human genomes. Bioinformatics. 2011;27(14):1998-2000.

62. Grant JR, Arantes AS, Liao X, Stothard P. In-depth annotation of SNPs arising from resequencing projects using NGS-SNP. Bioinformatics. 2011;27(16):2300-1.

63. Grover D, Woodfield AS, Verma R, Zandi PP, Levinson DF, Potash JB. QuickSNP: an automated web server for selection of tagSNPs. Nucleic Acids Res. 2007;35(Web Server issue):115-20.

64. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. Journal of molecular biology. 2002;320(2):369-87.

65. Han A, Kang HJ, Cho Y, Lee S, Kim YJ, Gong S. SNP@Domain: a web resource of single nucleotide polymorphisms (SNPs) within protein domain structures and sequences. Nucleic Acids Res. 2006;34(Web Server issue):642-4.

66. Hemminger BM, Saelim B, Sullivan PF. TAMAL: an integrated approach to choosing SNPs for genetic studies of human complex traits. Bioinformatics. 2006;22(5):626-7.

67. Hubisz MJ, Pollard KS, Siepel A. PHAST and RPHAST: phylogenetic analysis with space/time models. Brief Bioinform. 2011;12(1):41-51.

68. Jegga AG, Chen J, Gowrisankar S, Deshmukh MA, Gudivada R, Kong S, Kaimal V, Aronow BJ. GenomeTrafac: a whole genome resource for the detection of transcription factor binding site clusters associated with conventional and microRNA encoding genes conserved between mouse and human gene orthologs. Nucleic Acids Res. 2007;35(Database issue):116-21.

69. Kang HJ, Choi KO, Kim B-D, Kim S, Kim YJ. FESD: a Functional Element SNPs Database in human. Nucleic Acids Res. 2005;33(Database issue):518-22.

70. Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, Eswar N, Haussler D, Sali A. LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics. 2005;21(12):2814-20.

71. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols. 2009;4(7):1073-81.

72. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics. 2009;25(21):2744-50.

73. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat. 2011;32(8):894-9.

74. Makarov V, O'Grady T, Cai G, Lihm J, Buxbaum JD, Yoon S. AnnTools: a comprehensive and versatile annotation toolkit for genomic variants. Bioinformatics. 2012;28(5):724-5.

75. Masso M, Vaisman II. AUTO-MUTE: web-based tools for predicting stability changes in proteins due to single amino acid replacements. Protein Eng Des Sel. 2010;23(8):683-7.

76. Mathe E, Olivier M, Kato S, Ishioka C, Hainaut P, Tavtigian SV. Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. Nucleic Acids Res. 2006;34(5):1317-25.

77. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010;26(16):2069-70.

78. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12(4).

79. Parthiban V, Gromiha MM, Schomburg D. CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Res. 2006;34(Web Server issue):239-42.

80. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39(17).

81. Riva A, Kohane IS. A SNP-centric database for the investigation of the human genome. BMC Bioinformatics. 2004;5:33-.

82. Saccone SF, Quan J, Mehta G, Bolze R, Thomas P, Deelman E, Tischfield JA, Rice JP. New tools and methods for direct programmatic access to the dbSNP relational database. Nucleic Acids Res. 2011;39(Database issue):901-7.

83. Schaefer C, Meier A, Rost B, Bromberg Y. SNPdbe: constructing an nsSNP functional impacts database. Bioinformatics. 2012;28(4):601-2.

84. Schmitt AO, Assmus J, Bortfeldt RH, Brockmann GA. CandiSNPer: a web tool for the identification of candidate SNPs for causal variants. Bioinformatics. 2010;26(7):969-70.

85. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010;7(8):575-6.

86. Shetty AC, Athri P, Mondal K, Horner VL, Steinberg KM, Patel V, Caspary T, Cutler DJ, Zwick ME. SeqAnt: a web service to rapidly identify and annotate DNA sequence variations. BMC Bioinformatics. 2010;11:471-.

87. Stitziel NO, Binkowski TA, Tseng YY, Kasif S, Liang J. topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. Nucleic Acids Res. 2004;32(Database issue):520-2.

88. Stone EA, Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res. 2005;15(7):978-86.

89. Stoyanovich J, Pe'er I. MutaGeneSys: estimating individual disease susceptibility based on genome-wide SNP array data. Bioinformatics. 2008;24(3):440-2.

90. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 2003;13(9):2129-41.

91. Tian J, Wu N, Guo X, Guo J, Zhang J, Fan Y. Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines. BMC Bioinformatics. 2007;8:450-.

92. Uzun A, Leslin CM, Abyzov A, Ilyin V. Structure SNP (StSNP): a web server for mapping and modeling nsSNPs on protein structures with linkage to metabolic pathways. Nucleic Acids Res. 2007;35(Web Server issue):384-92.

93. Venselaar H, Te Beek TAH, Kuipers RKP, Hekkelman ML, Vriend G. Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. BMC Bioinformatics. 2010;11:548-.

94. Wainreb G, Ashkenazy H, Bromberg Y, Starovolsky-Shitrit A, Haliloglu T, Ruppin E, Avraham KB, Rost B, Ben-Tal N. MuD: an interactive web server for the prediction of non-neutral substitutions using protein structural data. Nucleic Acids Res. 2010;38(Web Server issue):523-8.

95. Wang J, Ronaghi M, Chong SS, Lee CGL. pfSNP: An integrated potentially functional SNP resource that facilitates hypotheses generation through knowledge syntheses. Hum Mutat. 2011;32(1):19-24.

96. Wang L, Liu S, Niu T, Xu X. SNPHunter: a bioinformatic software for single nucleotide polymorphism data acquisition and management. BMC Bioinformatics. 2005;6:60-.

97. Wang P, Dai M, Xuan W, McEachin RC, Jackson AU, Scott LJ, Athey B, Watson SJ, Meng F. SNP Function Portal: a web database for exploring the function implication of SNP alleles. Bioinformatics. 2006;22(14):523-9.

98. Wong WC, Kim D, Carter H, Diekhans M, Ryan MC, Karchin R. CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. Bioinformatics. 2011;27(15):2147-8.

99. Xu H, Gregory SG, Hauser ER, Stenger JE, Pericak-Vance MA, Vance JM, Zuchner S, Hauser MA. SNPselector: a web tool for selecting SNPs for genetic association studies. Bioinformatics. 2005;21(22):4181-6.

100. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009;25(21):2865-71.

101. Yeo G, Hoon S, Venkatesh B, Burge CB. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. Proc Natl Acad Sci U S A. 2004;101(44):15700-5.

102. Yue P, Melamud E, Moult J. SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics. 2006;7:166-.

103. Zhang XHF, Kangsamaksin T, Chao MSP, Banerjee JK, Chasin LA. Exon inclusion is dependent on predictable exonic splicing enhancers. Molecular and cellular biology. 2005;25(16):7323-32.

104. Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061-73. Epub 2010/10/29. doi: 10.1038/nature09534. PubMed PMID: 20981092; PubMed Central PMCID: PMC3042601.

105. Exome Variant Server [Internet]. NHLBI Exome Sequencing Project (ESP). 2011 [cited 2011-09-10]. Available from: http://evs.gs.washington.edu/EVS/.

106. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011;21(6):974-84.

107. Chiang DY, Getz G, Jaffe DB, O'Kelly MJT, Zhao X, Carter SL, Russ C, Nusbaum C, Meyerson M, Lander ES. High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat Methods. 2009;6(1):99-9103.

108. Ivakhno S, Royce T, Cox AJ, Evers DJ, Cheetham RK, Tavare S. CNAseg--a novel framework for identification of copy number changes in cancer from second-generation sequencing data. Bioinformatics. 2010;26(24):3051-8.

109. Li J, Lupat R, Amarasinghe KC, Thompson ER, Doyle MA, Ryland GL, Tothill RW, Halgamuge SK, Campbell IG, Gorringe KL. CONTRA: copy number analysis for targeted resequencing. Bioinformatics. 2012;28(10):1307-13.

110. Medvedev P, Fiume M, Dzamba M, Smith T, Brudno M. Detecting copy number variation with mated short reads. Genome Res. 2010;20(11):1613-22.

111. Sathirapongsasuti JF, Lee H, Horst BAJ, Brunner G, Cochran AJ, Binder S, Quackenbush J, Nelson SF. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics. 2011;27(19):2648-54.

112. Waszak SM, Hasin Y, Zichner T, Olender T, Keydar I, Khen M, Stutz AM, Schlattl A, Lancet D, Korbel JO. Systematic inference of copy-number genotypes from personal genome sequencing data reveals extensive olfactory receptor gene content diversity. PLoS computational biology. 2010;6(11).

113. Xie C, Tammi MT. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinformatics. 2009;10:80-.

114. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. 2009;19(9):1586-92.

115. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, Ding L, Mardis ER. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 2009;6(9):677-81.

116. Clark MJ, Homer N, O'Connor BD, Chen Z, Eskin A, Lee H, Merriman B, Nelson SF. U87MG decoded: the genomic sequence of a cytogenetically aberrant human cancer cell line. PLoS Genet. 2010;6(1).

117. Ge H, Liu K, Juan T, Fang F, Newman M, Hoeck W. FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. Bioinformatics. 2011;27(14):1922-8.

118. Ha G, Roth A, Lai D, Bashashati A, Ding J, Goya R, Giuliany R, Rosner J, Oloumi A, Shumansky K, Chin S-F, Turashvili G, Hirst M, Caldas C, Marra MA, Aparicio S, Shah SP. Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. Genome Res. 2012;22(10):1995-2007.

119. Hormozdiari F, Alkan C, Ventura M, Hajirasouliha I, Malig M, Hach F, Yorukoglu D, Dao P, Bakhshi M, Sahinalp SC, Eichler EE. Alu repeat discovery and characterization within human genomes. Genome Res. 2011;21(6):840-9.

120. Karakoc E, Alkan C, O'Roak BJ, Dennis MY, Vives L, Mark K, Rieder MJ, Nickerson DA, Eichler EE. Detection of structural variants and indels within exome data. Nat Methods. 2012;9(2):176-8.

121. Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z, Snyder M, Gerstein MB. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol. 2009;10(2).

122. Lam HYK, Mu XJ, Stutz AM, Tanzer A, Cayting PD, Snyder M, Kim PM, Korbel JO, Gerstein MB. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. Nat Biotechnol. 2010;28(1):47-55.

123. Marschall T, Costa IG, Canzar S, Bauer M, Klau GW, Schliep A, Schonhuth A. CLEVER: clique-enumerating variant finder. Bioinformatics. 2012;28(22):2875-82.

124. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, Chinwalla A, Conrad DF, Fu Y, Grubert F, Hajirasouliha I, Hormozdiari F, Iakoucheva LM, Iqbal Z, Kang S, Kidd JM, Konkel MK, Korn J, Khurana E, Kural D, Lam HYK, Leng J, Li R, Li Y, Lin C-Y, Luo R, Mu XJ, Nemesh J, Peckham HE, Rausch T, Scally A, Shi X, Stromberg MP, Stutz AM, Urban AE, Walker JA, Wu J, Zhang Y, Zhang ZD, Batzer MA, Ding L, Marth GT, McVean G, Sebat J, Snyder M, Wang J, Ye K, Eichler EE, Gerstein MB, Hurles ME, Lee C, McCarroll SA, Korbel JO. Mapping copy number variation by population-scale genome sequencing. Nature. 2011;470(7332):59-65.

125. Quinlan AR, Clark RA, Sokolova S, Leibowitz ML, Zhang Y, Hurles ME, Mell JC, Hall IM. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. Genome Res. 2010;20(5):623-35.

126. Sindi SS, Onal S, Peng LC, Wu H-T, Raphael BJ. An integrative probabilistic model for identification of structural variation in sequencing data. Genome Biol. 2012;13(3).

127. Sun R, Love MI, Zemojtel T, Emde A-K, Chung H-R, Vingron M, Haas SA. Breakpointer: using local mapping artifacts to support sequence breakpoint discovery from single-end reads. Bioinformatics. 2012;28(7):1024-5.

128. Suzuki S, Yasuda T, Shiraishi Y, Miyano S, Nagasaki M. ClipCrop: a tool for detecting structural variations with single-base resolution using soft-clipping information. BMC Bioinformatics. 2011;12 Suppl 14.

129. Wang J, Mullighan CG, Easton J, Roberts S, Heatley SL, Ma J, Rusch MC, Chen K, Harris CC, Ding L, Holmfeldt L, Payne-Turner D, Fan X, Wei L, Zhao D, Obenauer JC, Naeve C, Mardis ER, Wilson RK, Downing JR, Zhang J. CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nat Methods. 2011;8(8):652-4.

130. Wong K, Keane TM, Stalker J, Adams DJ. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. Genome Biol. 2010;11(12).

131. Zeitouni B, Boeva V, Janoueix-Lerosey I, Loeillet S, Legoix-ne P, Nicolas A, Delattre O, Barillot E. SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. Bioinformatics. 2010;26(15):1895-6.

132. de Ligt J, Veltman JA, Vissers LE. Point mutations as a source of de novo genetic disease. Current opinion in genetics & development. 2013;23(3):257-63. Epub 2013/03/05. doi: 10.1016/j.gde.2013.01.007. PubMed PMID: 23453690.

133. Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J, Yamrom B, Lee YH, Narzisi G, Leotta A, Kendall J, Grabowska E, Ma B, Marks S, Rodgers L, Stepansky A, Troge J, Andrews P, Bekritsky M, Pradhan K, Ghiban E, Kramer M, Parla J, Demeter R, Fulton LL, Fulton RS, Magrini VJ, Ye K, Darnell JC, Darnell RB, Mardis ER, Wilson RK, Schatz MC, McCombie WR, Wigler M. De novo gene disruptions in children on the autistic spectrum. Neuron. 2012;74(2):285-99. Epub 2012/05/01. doi: 10.1016/j.neuron.2012.04.009. PubMed PMID: 22542183; PubMed Central PMCID: PMC3619976.

134. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin CF, Stevens C, Wang LS, Makarov V, Polak P, Yoon S, Maguire J, Crawford EL, Campbell NG, Geller ET, Valladares O, Schafer C, Liu H, Zhao T, Cai G, Lihm J, Dannenfelser R, Jabado O, Peralta Z, Nagaswamy U, Muzny D, Reid JG, Newsham I, Wu Y, Lewis L, Han Y, Voight BF, Lim E, Rossin E, Kirby A, Flannick J, Fromer M, Shakir K, Fennell T, Garimella K, Banks E, Poplin R, Gabriel S, DePristo M, Wimbish JR, Boone BE, Levy SE, Betancur C, Sunyaev S, Boerwinkle E, Buxbaum JD, Cook EH, Jr., Devlin B, Gibbs RA, Roeder K, Schellenberg GD, Sutcliffe JS, Daly MJ. Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature. 2012;485(7397):242-5. Epub 2012/04/13. doi: 10.1038/nature11011. PubMed PMID: 22495311; PubMed Central PMCID: PMC3613847.

135. O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S, Karakoc E, Mackenzie AP, Ng SB, Baker C, Rieder MJ, Nickerson DA, Bernier R, Fisher SE, Shendure J, Eichler EE. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat Genet. 2011;43(6):585-9. Epub 2011/05/17. doi: 10.1038/ng.835. PubMed PMID: 21572417; PubMed Central PMCID: PMC3115696.

136. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL, Walker MF, Ober GT, Teran NA, Song Y, El-Fishawy P, Murtha RC, Choi M, Overton JD, Bjornson RD, Carriero NJ, Meyer KA, Bilguvar K, Mane SM, Sestan N, Lifton RP, Gunel M, Roeder K, Geschwind DH, Devlin B, State MW. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature. 2012;485(7397):237-41. Epub 2012/04/13. doi: 10.1038/nature10945. PubMed PMID: 22495306; PubMed Central PMCID: PMC3667984.

137. Veltman JA, Brunner HG. De novo mutations in human genetic disease. Nat Rev Genet. 2012;13(8):565-75. Epub 2012/07/19. doi: 10.1038/nrg3241. PubMed PMID: 22805709.

138. Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella KV, Zilversmit M, Cartwright R, Rouleau GA, Daly M, Stone EA, Hurles ME, Awadalla P, Genomes P. Variation in genome-wide mutation rates within and between human families. Nat Genet. 2011;43(7):712-4. Epub 2011/06/15. doi: 10.1038/ng.862. PubMed PMID: 21666693; PubMed Central PMCID: PMC3322360.

139. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, Wong WS, Sigurdsson G, Walters GB, Steinberg S, Helgason H, Thorleifsson G, Gudbjartsson DF, Helgason A, Magnusson OT, Thorsteinsdottir U, Stefansson K. Rate of de novo mutations and the importance of father's age to disease risk.

Nature. 2012;488(7412):471-5. Epub 2012/08/24. doi: 10.1038/nature11396. PubMed PMID: 22914163; PubMed Central PMCID: PMC3548427.

140. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang ZD, Conrad DF, Lunter G, Zheng H, Ayub Q, DePristo MA, Banks E, Hu M, Handsaker RE, Rosenfeld JA, Fromer M, Jin M, Mu XJ, Khurana E, Ye K, Kay M, Saunders GI, Suner MM, Hunt T, Barnes IH, Amid C, Carvalho-Silva DR, Bignell AH, Snow C, Yngvadottir B, Bumpstead S, Cooper DN, Xue Y, Romero IG, Genomes Project C, Wang J, Li Y, Gibbs RA, McCarroll SA, Dermitzakis ET, Pritchard JK, Barrett JC, Harrow J, Hurles ME, Gerstein MB, Tyler-Smith C. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012;335(6070):823-8. Epub 2012/02/22. doi: 10.1126/science.1215040. PubMed PMID: 22344438; PubMed Central PMCID: PMC3299548.

141. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 2011;12(11):745-55. Epub 2011/09/29. doi: 10.1038/nrg3031. PubMed PMID: 21946919.

142. Gaj T, Gersbach CA, Barbas CF, 3rd. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends in biotechnology. 2013. doi: 10.1016/j.tibtech.2013.04.004. PubMed PMID: 23664777.

143. Howe DG, Bradford YM, Conlin T, Eagle AE, Fashena D, Frazer K, Knight J, Mani P, Martin R, Moxon SA, Paddock H, Pich C, Ramachandran S, Ruef BJ, Ruzicka L, Schaper K, Shao X, Singer A, Sprunger B, Van Slyke CE, Westerfield M. ZFIN, the Zebrafish Model Organism Database: increased support for mutants and transgenics. Nucleic Acids Res. 2013;41(Database issue):D854-60. doi: 10.1093/nar/gks938. PubMed PMID: 23074187; PubMed Central PMCID: PMC3531097.

144. Laulederkind SJ, Hayman GT, Wang SJ, Smith JR, Lowry TF, Nigam R, Petri V, de Pons J, Dwinell MR, Shimoyama M, Munzenmaier DH, Worthey EA, Jacob HJ. The Rat Genome Database 2013--data, tools and users. Brief Bioinform. 2013;14(4):520-6. doi: 10.1093/bib/bbt007. PubMed PMID: 23434633; PubMed Central PMCID: PMC3713714.

145. Kerner B, Rao AR, Christensen B, Dandekar S, Yourshaw M, Nelson SF. Rare genomic variants link bipolar disorder to CREB regulated intracellular signaling pathways. Frontiers in Psychiatry. 2013. doi: 10.3389/fpsyt.2013.00154.

146. Wan J, Yourshaw M, Mamsa H, Rudnik-Schoneborn S, Menezes MP, Hong JE, Leong DW, Senderek J, Salman MS, Chitayat D, Seeman P, von Moers A, Graul-Neumann L, Kornberg AJ, Castro-Gago M, Sobrido MJ, Sanefuji M, Shieh PB, Salamon N, Kim RC, Vinters HV, Chen Z, Zerres K, Ryan MM, Nelson SF, Jen JC. Mutations in the RNA exosome component gene EXOSC3 cause pontocerebellar hypoplasia and spinal motor neuron degeneration. Nat Genet. 2012;44(6):704-8. Epub 2012/05/01. doi: 10.1038/ng.2254. PubMed PMID: 22544365; PubMed Central PMCID: PMC3366034.

147. Yourshaw M, Solorzano-Vargas RS, Pickett LA, Lindberg I, Wang J, Cortina G, Pawlikowska-Haddal A, Baron H, Venick RS, Nelson SF. Exome Sequencing Finds a Novel PCSK1 Mutation in a Child With Generalized Malabsorptive Diarrhea and Diabetes Insipidus. Journal of pediatric gastroenterology and nutrition. 2013. doi: 10.1097/MPG.0b013e3182a8ae6c.

148. Lee H, Nelson SF. Rethinking clinical practice: clinical implementation of exome sequencing. Pers Med. 2012;9(8):785-7. doi: Doi 10.2217/Pme.12.101. PubMed PMID: WOS:000311977800001.

149. Rubenstein PA, Mayer EA. Familial visceral myopathies: from symptom-based syndromes to actin-related diseases. Gastroenterology. 2012;143(6):1420-3. doi: 10.1053/j.gastro.2012.10.031. PubMed PMID: 23085350.